BỘ THÔNG TIN VÀ TRUYỀN THÔNG

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN THỊ THU GIANG

NGHIÊN CỨU ỨNG DỤNG VÀ ĐỀ XUẤT CÁC PHƯƠNG PHÁP TÍNH TOÁN ĐỂ DỰ ĐOÁN ĐÁP ỨNG THUỐC TRONG ĐIỀU TRỊ BỆNH

Chuyên ngành: **Hệ thống thông tin**

Mã số: **9.48.01.04**

LUẬN ÁN TIẾN SĨ KỸ THUẬT

Hà Nội - 2024

BỘ THÔNG TIN VÀ TRUYỀN THÔNG
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

LUẬN ÁN TIẾN SĨ KỸ THUẬT

NGHIÊN CỨU ỨNG DỤNG VÀ ĐỀ XUẤT CÁC PHƯƠNG PHÁP TÍNH TOÁN ĐỂ DỰ ĐOÁN ĐÁP ỨNG THUỐC TRONG ĐIỀU TRỊ BỆNH

NGHIÊN CỨU SINH: NGUYỄN THỊ THU GIANG

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS.TS. LÊ ĐỨC HẬU
PGS.TS. NGUYỄN TRỌNG KHÁNH

HÀ NỘI - 2024

LỜI CAM ĐOAN

Tôi cam đoan rằng luận án Tiến sĩ: *"Nghiên cứu ứng dụng và đề xuất các phương pháp tính toán để dự đoán đáp ứng thuốc trong điều trị bệnh"* là công trình nghiên cứu của riêng tôi dưới sự hướng dẫn của PGS.TS Lê Đức Hậu và PGS.TS. Nguyễn Trọng Khánh, trừ những kiến thức, nội dung tham khảo từ các tài liệu đã được trích dẫn theo quy định.

Các giải pháp đề xuất được trình bày ra trong luận án đều trung thực và khách quan. Một số đã được công bố trên các tạp chí và kỷ yếu hội thảo khoa học chuyên ngành, được liệt kê theo danh mục các công trình đã công bố của tác giả ở phần cuối luận án. Các phần còn lại chưa được công bố ở bất kỳ công trình nào khác.

Hà Nội, ngày 20 tháng 01 năm 2024

Tác giả luận án

Nguyễn Thị Thu Giang

LỜI CẢM ƠN

Lời đầu tiên, tôi xin được gửi lời cảm ơn chân thành tới Ban Giám đốc Học viện, Khoa đào tạo Sau Đại học cùng các Thầy Cô đã tận tình giảng dạy, hướng dẫn, tạo mọi điều kiện thuận lợi giúp tôi trong suốt quá trình học tập và nghiên cứu tại Học viện.

Tôi xin gửi lời cảm ơn sâu sắc tới PGS.TS Lê Đức Hậu và PGS.TS. Nguyễn Trọng Khánh, đã luôn tận tình hướng dẫn, động viên, truyền cảm hứng nghiên cứu và năng lượng tích cực cho tôi trong suốt quá trình theo đuổi con đường học thuật. Tầm nhìn và định hướng chuyên môn sâu của các Thầy giúp tôi đạt được những kết quả trong nghiên cứu khoa học.

Tôi xin gửi lời cảm ơn tới các cộng sự, chuyên gia phân tích dữ liệu Nguyễn Thanh Tuấn, Vũ Đức Hòa, đã nhiệt tình hỗ trợ và chia sẻ kinh nghiệm quý báu trong học thuật và phân tích, xử lý dữ liệu lớn.

Tôi cũng xin gửi lời cảm ơn các đồng nghiệp, bạn bè đã luôn tin tưởng, tạo điều kiện thuận lợi và chia sẻ với tôi trong công tác chuyên môn.

Cuối cùng, Tôi xin được dành sự yêu thương, lòng biết ơn tới gia đình, người thân đã luôn quan tâm, động viên, đồng hành cùng tôi trong suốt chặng đường dài.

Xin chân thành cảm ơn!

Hà Nội, ngày 20 tháng 01 năm 2024

DANH MỤC CÁC TỪ VIẾT TẮT

Thuật ngữ	Diễn giải tiếng Anh	Diễn giải tiếng Việt
Acc	Accuracy	Độ chính xác
Antagonistic	Antagonistic	Tương kháng thuốc
AUC	Area Under the Curve	Diện tích dưới đường cong nằm dưới ROC
CCLE	Cancer Cell Line Encyclopedia	Nguồn dữ liệu Bách khoa toàn thư về dòng tế bào ung thư
CCp	Pearson correlation coefficient	Hệ số tương quan Pearson
Cell line	Cell line	Dòng tế bào
CNA	Copy Number Alterations	Biến thể số lượng bản sao
CNN	Convolutional Neural Network	Mạng nơ-ron đồ thị tích chập
CNN1D	CNN1D	Mạng nơ-ron tích chập 1 chiều
CNV	Copy Number Variations	Biến thể số lượng bản sao
DNA	Deoxyribonucleic Acid	Phân tử mang thông tin di truyền
DRP	Drug response prediction	Dự đoán đáp ứng thuốc
DSP	Drug synergy prediction	Dự đoán đáp ứng đa thuốc
Epigenomics	Epigenomics	Hệ di truyền biểu sinh
F1-score	F1-score	Điểm đánh giá trung bình điều hòa của precision và recall
FC	Fully connected	Lớp kết nối đầy đủ
GAT	Graph attention network	Mạng nơ-ron đồ thị cơ chế chú ý
GCN	Graph convolution network	Mạng nơ-ron tích chập đồ thị
GDSC	Genomics of Drug Sensitivity in Cancer	Nguồn dữ liệu nghiên cứu đáp ứng thuốc trong điều trị ung thư
GE	Gene Expression	Dữ liệu biểu hiệu gen
Genome	Genome	Bộ gen
Genomic aberration	Genomic aberration	Đột biến gen
Genomics	Genomics	Hệ gen
GIN	Graph isomorphism network	Mạng nơ-ron đồ thị đẳng cấu
GNN	Graph Neural Network	Mạng nơ-ron đồ thị
IC_{50}	The half maximal inhibitory concentration	Nồng độ ức chế tối đa một nửa của thuốc

Thuật ngữ	Diễn giải tiếng Anh	Diễn giải tiếng Việt
LOOV	Leave-One-Out Cross-Validation	Đánh giá trên mỗi mẫu thử
METH	DNA Methylation	Dữ liệu methyl hóa
MLP	Multiple Layer Perceptrion	Mạng nơ-ron đa tầng lớp
mRNA	Messenger RNA	RNA thông tin
MSE	Mean Squared Error	Sai số bình phương trung bình
MUT	Mutation	Dữ liệu đột biến gen
NCI-60	National Cancer Institute	Nguồn dữ liệu sàng lọc 60 dòng tế bào ung thư ở người
Pooling layer	Pooling layer	Lớp tổng hợp
PPI	Protein - Protein Interaction	Tương tác protein
PRE	Precision	Độ chuẩn
Proteomics	Proteomics	Hệ protein
R	Resitance	Kháng thuốc
Recall	Recall	Độ hồi nhớ
ReLU	Rectified Linear Unit	Hàm kích hoạt ReLU
RMSE	Root Mean Squared Error	Sai số bình phương trung bình gốc
RNA	Ribonucleic acid	Axit nucleic
ROC	Receiver Operating Characteristics	Đường cong ROC
S	Sensitivity	Đáp ứng thuốc
SMILES	Simplified Molecular Input Line Entry System	Chuỗi ký hiệu hóa học của phân tử thuốc
Synergistic	Synergistic	Tương hợp thuốc
Transcriptome	Transcriptome	Bộ phiên mã
Transcriptomics	Transcriptomics	Hệ phiên mã

DANH MỤC CÁC KÝ HIỆU

Ký hiệu	Diễn giải tiếng việt	Trang
$G = (V, E)$	Đồ thị G với tập các nút V và một tập các cạnh E	27
u, v	Đỉnh u, v của đồ thị	27
$h_u^{(l)}$	Vec-tơ đặc trưng của đỉnh u sau khi qua lớp tích chập thứ l	29
$m_{N(u)}^{(l)}$	Thông điệp dựa trên các thông tin hàng xóm của nút u tại lớp thứ l	29
z_u	Vec-tơ embedding của u qua mạng nơ-ron đồ thị	30
$e_{u,\mathrm{v}}$	Cạnh $e_{u,\mathrm{v}} \in E$	30
$X \in R^{N\,x\,D}$	Ma trận đặc trưng của đồ thị, với n là số đỉnh, d là số chiều của một vec-tơ đặc trưng đỉnh	31
A	Ma trận kề biểu diễn kết nối giữa các đỉnh	31
A^T	Ma trận chuyển vị của A	31
W	Ma trận trọng số	31
D	Ma trận bậc của ma trận A, $D_{i,j} = \sum_{j=1}^{n} A_{i,j}$	31
$\widetilde{D}$	Ma trận bậc đã được chuẩn hóa	31
$\alpha(Wh_i, Wh_j)$	α là một cơ chế chú ý (attention), giữa các cặp nút (i, j)	31
σ	Hàm kích hoạt	32
$\alpha_{i,j}$	Hệ số attention	32
$(d_{i,}\ ,\ d_j\ )$	Cặp tương tác thuốc	79
c_n	Cell line n	79
$a_{i,n,j}$	Hệ số chú ý (attention) của thuốc $d_{i,}$ trong cặp thuốc $(d_{i,}\ ,\ d_j)$ tác động trên dòng tế bào $c_{n,}$	79
$\hat{y}$i, j, n	Vec-tơ tổng hợp tương tác cặp thuốc $(d_{i,}\ ,\ d_j)$ trên dòng tế bào $c_{n,}$	79
R_k (u)	Tập các nút ở bán kính k từ nút "u"	92
s(R_k (u))	tập các bậc của các nút trong R_k (u)	92
$f_{\mathrm{k}}(\mathrm{u}, \mathrm{v})$	Khoảng cách cấu trúc tương đồng giữa u và v xét trong trong vùng lân cận k-hop	92
W_k(u,v)	Trọng số cạnh giữa mỗi cặp đỉnh (u,v)	93
Γ_k (u)	Số cạnh của lớp k mà có trọng số lớn hơn trọng số trung bình của các cạnh mà nút u tương tác với đỉnh khác trong lớp k	93
Z_k(u)	Hệ số chuẩn hóa nút u trong lớp k	93
p_k (u,v)	Xác suất để chọn đến một nút "v" bất kỳ ở lớp k	93
Dj	Vec-tơ biểu diễn của thuốc i	94
Cj	Vec-tơ biểu diễn à dòng tế bào j	94

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG

MỤC LỤC

PHẦN MỞ ĐẦU

1. Đặt vấn đề

Trong những năm gần đây, y học chính xác đang là một xu hướng rất được quan tâm nghiên cứu nhằm mục đích hỗ trợ và tìm ra các phương pháp điều trị tốt nhất cho từng bệnh nhân dựa trên đặc trưng sinh học, phong cách sống, môi trường và nền tảng di truyền của họ. Y học chính xác thực hiện phân tích, đánh giá trên từng cá nhân hoặc nhóm bệnh nhân từ đó đưa ra phác đồ điều trị, chăm sóc sức khỏe theo từng giai đoạn như: chẩn đoán, dự phòng, điều trị sao cho phù hợp nhất với từng bệnh nhân hoặc nhóm bệnh nhân [1]. Việc dự đoán chính xác được khả năng đáp ứng của từng bệnh nhân đối với các phương pháp điều trị mang lại nhiều ý nghĩa tích cực trong y học chính xác. Các bác sĩ có thể dựa vào kết quả dự đoán này để đưa ra quyết định, lựa chọn phương pháp điều trị hiện có sao cho hiệu quả và ít tác dụng phụ nhất. Với sự phát triển của công nghệ hiện nay, các hệ thống dự đoán cho phép lựa chọn và theo dõi thử nghiệm lâm sàng trên bệnh nhân thông minh hơn [2], [3].

Mỗi người bệnh có đặc trưng sinh học khác nhau, có khả năng đáp ứng với từng thuốc điều trị khác nhau. Sự đáp ứng không đồng nhất này gây khó khăn trong quá trình điều trị. Việc phát triển thuốc mới là rất tốn kém và mất thời gian, trong khi đó thực tế điều trị có thể có nhiều trường hợp một loại thuốc không chỉ chữa được cho một loại bệnh mà có thể đáp ứng được cho một vài bệnh khác; hay việc kết hợp nhiều thuốc với nhau có thể làm ngăn chặn sự kháng thuốc và tăng khả năng đáp ứng điều trị. Do đó việc dự đoán đáp ứng thuốc trong điều trị là vấn đề quan trọng trong y học chính xác.

Các phương pháp dự đoán đáp ứng thuốc hiện nay thường áp dụng các mô hình tính toán để khai thác, phân tích các dữ liệu y sinh học (như dữ liệu biểu hiện gen, đột biến gen, dữ liệu thuốc, dữ liệu đáp ứng thuốc…), tìm ra mối liên hệ giữa chúng và dự đoán khả năng đáp ứng của thuốc cho người bệnh [4]. Việc khai phá dữ liệu này không chỉ tìm ra được mối quan hệ quan trọng giữa các đặc trưng sinh học người bệnh, giữa thuốc với bệnh mà còn có thể dự đoán khả năng đáp ứng thuốc cho từng bệnh, cũng như dự đoán khả năng đáp ứng thuốc cho các thuốc mới hoặc bệnh mới.

Hai bài toán quan trọng trong dự đoán đáp ứng thuốc hiện nay là dự đoán đáp ứng đơn thuốc (monotherapy) và dự đoán đáp ứng đa thuốc hay kết hợp thuốc (combination therapy). Trong đó, điều trị bằng liệu trình đơn thuốc là dùng một loại thuốc duy nhất để điều trị bệnh. Sau một thời gian đáp ứng ban đầu, hiệu quả của các liệu trình điều trị đơn thuốc (ví dụ: thuốc chống ung thư) thường giảm do sự tồn tại của các cơ chế kháng thuốc nội tại mắc phải. Để khắc phục tình trạng này, liệu trình phổ biến là kết hợp thuốc nhằm làm tăng hiệu quả điều trị mà không cần tăng liều lượng thuốc [5].

Trong nghiên cứu tiền lâm sàng, dòng tế bào (cell line) được coi như một bệnh nhân nhân tạo, mang đầy dủ hầu hết đặc điểm sinh học của người bệnh. Với khả năng dễ triển khai nghiên cứu thử nghiệm với số lượng lớn bệnh nhân nhân tạo này cùng với các sự ra đời của công nghệ thông lượng cao đã tạo ra lượng lớn dữ liệu -omics về các dòng tế bào. Các dữ liệu này là nguồn dữ liệu quan trọng trong các nghiên cứu tiền lâm, tạo điều kiện cho việc dự đoán và chuẩn đoán hướng điều trị tốt hơn. Do đó bài toán dự đoán đáp ứng thuốc thường tập trung vào dự đoán cho dòng tế bào.

2. Lý do chọn đề tài

Các mô hình tính toán dự đoán của đáp ứng thuốc đóng góp tích cực vào nghiên cứu tiền lâm sàng [6], [7], giúp các bác sĩ có thể ra quyết định điều trị nhanh chóng và chính xác hơn. Nhiều công trình nghiên cứu đã được công bố và ngày càng thu hút lượng lớn các nhà nghiên cứu y sinh tính toán [8] tham gia và đề xuất các phương pháp mới.

Một loạt các phương pháp tính toán dựa trên mô hình thống kê, học máy từ hồi quy tuyến tính, máy học vec-tơ hỗ trợ (SVM) đến các mô hình rừng ngẫu nhiên (RF), học đa tác vụ (multi-task learning) được đề xuất mang lại hiệu quả đáng kể trong việc dự đoán đáp ứng đơn thuốc [9], [10], [11], [12] hay các đáp ứng đa thuốc như [13], [14], [15], [16]. Tuy nhiên, các giải pháp này còn nhiều hạn chế như bộ dữ liệu còn nhỏ, không có cách tiếp cận nào có thể vượt trội hơn hẳn so với các phương pháp khác trên các tập dữ liệu khác nhau và trên các loại thuốc khác nhau. Với công nghệ thông lượng cao giải trình tự DNA, lượng lớn dữ liệu hệ gen được tạo ra cũng làm thúc đẩy nghiên cứu các phương pháp tính toán để khai thác sâu và rộng các dữ liệu sinh học cho dự đoán đáp ứng thuốc. Các loại thuốc và dòng tế bào thường được biểu

diễn ở dạng nhiều chiều, ví dụ: dữ liệu –omics của hàng chục nghìn gen được tạo ra cho mỗi dòng tế bào hay các phân tử hóa học của thuốc cũng được biểu diễn bằng lượng lớn các đặc trưng hóa học khác nhau. Trong khi đó, kích thước mẫu nhỏ do số dòng tế bào và thuốc được thử nghiệm còn hạn chế. Do đó, các phương pháp học máy thường phải đối mặt với thách thức "n nhỏ, p lớn" và dẫn đến hạn chế về hiệu năng dự đoán của chúng [17], [18].

Một vài năm gần đây, các mô hình học sâu với khả năng tính toán mạnh mẽ có thể học các biểu diễn trực tiếp từ các dữ liệu đầu mà không cần trích chọn đặc trưng trước khi huấn luyện cũng đang là một giải pháp tiềm năng cho bài toán này [19], [20], [21], [22], [23], [24], [25], [26]. So với các mô hình học máy truyền thống, các mô hình học sâu này cho thấy vượt trội. Tuy nhiên các mô hình này còn một số hạn chế như: (1) chưa tích hợp các đặc trưng phân tử hóa học của thuốc, hoặc có tích hợp nhưng thuốc được biểu diễn dưới dạng đơn giản như chuỗi hoặc ảnh mà chưa phải dạng biểu diễn tự nhiên hơn như dạng dữ liệu đồ thị - dạng biểu diễn có khả năng mang nhiều thông tin hơn; (2) chưa tích hợp đa dạng các dữ liệu đặc trưng sinh học bệnh (multi - omics); (3) chưa áp dụng các phương pháp tính toán tiên tiến, phù hợp hơn để học các biểu diễn thuốc và dữ liệu sinh học để cải thiện hiệu năng mô hình dự đoán.

Do đó, luận án tập trung vào việc nghiên cứu và đề xuất các giải pháp dự đoán đáp ứng thuốc trong điều trị bệnh nhằm giải quyết các vấn đề còn hạn chế trên. Với đề tài này, luận án tiến hành nghiên cứu tổng quan lý thuyết y sinh học, các phương thức xử lý, biểu diễn dữ liệu thuốc và dòng tế bào, các phương pháp tính toán tiên tiến, tích hợp dữ liệu ứng dụng vào bài toán dự đoán đáp ứng thuốc đơn thuốc và dự đoán đáp ứng đa thuốc. Từ đó, đề xuất các giải pháp tính toán để nâng cao hiệu năng dự đoán kết hợp thuốc trong điều trị bệnh. Tên đề tài luận án như sau:

Tên tiếng Việt là: ***"Nghiên cứu và đề xuất các phương pháp tính toán để dự đoán đáp ứng thuốc trong điều trị bệnh"***

Tên tiếng Anh là: ***"Research and propose computational methods to predict drug response in the treatment"***

3. Mục tiêu nghiên cứu

Mục tiêu tổng quát

Mục tiêu của đề tài là nghiên cứu các vấn đề liên quan đến dự đoán đáp ứng thuốc hiện nay và đề xuất một số giải pháp tính toán để tăng hiệu năng dự đoán đáp ứng thuốc trong điều trị bệnh. Bằng cách khai thác các bộ dữ liệu y sinh học công khai được cập nhật mới nhất, tiến hành chuẩn hóa và biểu diễn dữ liệu phù hợp với các giải pháp tính toán tiên tiến. Từ đó xây dựng mô hình dự đoán dự đoán đáp ứng đơn thuốc và dự đoán đáp ứng đa thuốc, góp phần nâng cao hiệu quả dự đoán điều trị trong y học chính xác.

Các mục tiêu cụ thể

Tổng hợp các kiến thức nền tảng về dữ liệu y sinh học, khảo sát, phân tích các phương pháp tính toán, các phương thức đánh giá mô hình dự đoán từ đó đề xuất các giải pháp bài toán dự đoán đáp ứng thuốc:

- Đề xuất giải pháp học dữ liệu biểu diễn thuốc dưới dạng đồ thị và tích hợp dữ liệu biểu diễn dữ liệu hệ gen dòng tế bào, để dự đoán đáp ứng đơn thuốc cho các dòng tế bào.

- Đề xuất giải pháp tích hợp dữ liệu biểu diễn thuốc dưới dạng đồ thị và đa dữ liệu -omics khác nhau như dữ liệu biểu hiện gen, methyl hóa của dòng tế bào để dự đoán đáp ứng đơn thuốc cho các dòng tế bào.

- Đề xuất giải pháp tích hợp dữ liệu biểu diễn thuốc dưới dạng đồ thị và đa dữ liệu -omics khác nhau của dòng tế bào để tổng hợp thông tin kết hợp các cặp thuốc và dòng tế bào để dự đoán đáp ứng đa thuốc cho các dòng tế bào.

- Đề xuất giải pháp tích hợp đa dữ liệu -omics với dữ liệu mạng tương tác protein (interactomics) để cải thiện dự đoán đáp ứng đa thuốc cho các dòng tế bào.

4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu:

Dựa trên các mục tiêu nghiên cứu trên, đối tượng nghiên cứu của luận án như sau:

- Các yếu tố và đặc trưng của dữ liệu sinh học dòng tế bào bệnh như: biểu hiện gen (gene expression), đột biến (mutation) và biến thể số lượng bản sao (copy number

alterations), methyl hóa (methylation). Các dữ liệu biểu diễn đặc trưng cấu tạo phân tử thuốc và đáp ứng thuốc với các dòng tế bào bệnh.

- Các mô hình học máy, học sâu, các phương pháp tính toán xử lý dữ liệu, các giải pháp tính toán nâng cao hiệu quả dự đoán cho cho bài toán dự đoán đáp ứng thuốc trong điều trị bệnh.

- Các phương pháp tích hợp dữ liệu (tích hợp đa dạng các dữ liệu sinh học, tích hợp dữ liệu biểu diễn thuốc) trong bài toán dự đáp ứng đơn thuốc /dự đoán kết hợp thuốc trong điều trị bệnh.

- Các phương pháp tính toán đánh giá hiệu năng và tối ưu hóa tham số mô hình dự đoán.

Phạm vi nghiên cứu.

- Các phương pháp tính toán, các mô hình học sâu nhằm nâng cao hiệu năng cho bài toán dự đoán đáp ứng thuốc đơn thuốc và đa thuốc.

- Các bộ dữ liệu công khai về dòng tế bào bệnh, đáp ứng thuốc cho các dòng tế bào bệnh này.

- Các phương pháp chuẩn hóa dữ liệu, mô hình học các biểu diễn dữ liệu cho dữ liệu sinh học của dòng tế bào bệnh; mô hình học các biểu đồ thị phân tử thuốc

- Các phương pháp đánh giá hiệu năng cho bài toán hồi quy và bài toán phân lớp, các kịch bản thử nghiệm để đánh giá trên bộ dữ liệu chung, kịch bản thử nghiệm để dự đoán đáp ứng thuốc của thuốc mới và kịch bản thử nghiệm dự đoán đáp ứng thuốc cho dòng tế bào mới.

5. Phương pháp nghiên cứu

- *Phương pháp luận:* Khảo sát các nghiên cứu liên quan, tổng hợp, phân tích và hệ thống hóa cơ sở lý thuyết, các vấn đề còn tồn tại và định hướng các giải pháp, đề xuất các phương pháp tính toán áp dụng nhằm nâng cao hiệu năng dự đoán đáp ứng thuốc trong điều trị bệnh.
- *Triển khai các mô hình đề xuất bằng mô hình tính toán, mô phỏng, thực nghiệm:* Thực hiện triển khai những kết quả nghiên cứu vào thực tiễn để kiểm định kết quả nghiên cứu lý thuyết.

- *Đánh giá hiệu năng mô hình đề xuất bằng các chỉ số đánh giá tương ứng:* Đo kết quả thực nghiệm, phân tích, so sánh với các nghiên cứu trước đây, tìm các dấu ấn sinh học có ý nghĩa trong nghiên cứu lâm sàng.

6. Những đóng góp chính của luận án

Phân tích các phương pháp dự đoán đáp ứng thuốc hiện tại, đánh giá ưu nhược các mô hình tiên tiến hiện nay, đề xuất hai giải pháp cho bài toán dự đoán đáp ứng thuốc cho điều trị đơn thuốc (monotherapy) và hai giải pháp cho bài toán dự đoán kết hợp thuốc (combination therapy):

- Giải pháp tích hợp dữ liệu trong dự đoán đáp ứng đơn thuốc.

Đóng góp thứ nhất là đề xuất giải pháp học dữ liệu biểu diễn đồ thị của phân tử thuốc – GraphDRP: Đề xuất này đã áp dụng cách biểu diễn dữ liệu thuốc dưới dạng đồ thị, sử dụng các phương pháp tính toán dựa trên mạng nơ-ron đồ thị (GNN) để học các biểu diễn dữ liệu này từ đó cải thiện hiệu năng dự đoán so với các phương pháp không tích hợp dữ liệu đồ thị phân tử thuốc. Trong số các mô hình GNN được áp dụng, giải pháp đề xuất cũng xác định được mô hình học dữ liệu đồ thị phân tử thuốc hiệu quả nhất.

Đóng góp thứ hai là đề xuất giải pháp tích hợp đa dữ liệu -omics và dữ liệu biểu diễn đồ thị phân tử thuốc -GraOmicDRP: Đề xuất đã tiếp tục cải thiện hiệu năng dự đoán đáp ứng đơn thuốc cho các dòng tế bào, bằng cách áp dụng mô hình học dữ liệu biểu diễn dạng đồ thị phân tử thuốc tích hợp với dữ liệu đa -omics của dòng tế bào. Các giải pháp tích hợp đa dữ liệu -omics cho thấy hiệu quả hơn giải pháp tích hợp đơn -omics, và vượt trội hơn so với các phương pháp tích hợp đa -omics khác nhưng không sử dụng dữ liệu biểu diễn thuốc dưới dạng đồ thị phân tử. Đồng thời giải pháp đề xuất cũng chỉ ra được loại dữ liệu -omics có ý nghĩa cho mô hình dự đoán.

- Giải pháp tích hợp dữ liệu trong dự đoán đáp ứng đa thuốc.

Đóng góp thứ ba là đề xuất giải pháp học biểu diễn đồ thị phân tử thuốc và tích hợp đa dữ liệu -omics để dự đoán đáp ứng đa thuốc - GraOmicSynergy: Đây là đề xuất học các biểu diễn của cặp thuốc dưới dạng đồ thị phân tử và tổng hợp thông tin biểu diễn cặp thuốc thử nghiệm trên các dòng tế bào thông qua cơ chế chú ý. Dữ liệu biểu diễn dòng tế bào cũng được tổng hợp từ mô hình học biểu diễn đa dữ liệu -omics. Giải pháp đề xuất đã cải thiện khả năng dự đoán so với các mô hình khác không sử

dụng biểu diễn đồ thị phân tử thuốc cũng như so với mô hình có sử dụng dữ liệu đồ thị phân tử thuốc nhưng chưa tích hợp đa dữ liệu -omics.

Đóng góp thứ tư là đề xuất giải pháp tích hợp đa dữ liệu -omics và mạng sinh học - AE-XGBSynergy. Đề xuất này tích hợp đa dữ liệu -omics của dòng tế bào, kết hợp với dữ liệu biểu diễn thuốc và dòng tế bào được trích xuất thông qua thông tin cấu trúc mạng tương tác protein (PPI) để dự đoán phân loại đáp ứng đa thuốc. Trong đó, dữ liệu biểu diễn dòng tế bào được trích xuất thông qua bộ mã hóa (AE), những biểu diễn cặp thuốc và dòng tế bào được đưa vào bộ phân loại để dự đoán phân loại đáp ứng đa thuốc. AE-XGBSynergy đã cho thấy hiệu năng vượt trội hơn so với một mô hình dự đoán chỉ có thông tin cấu trúc mạng PPI và không tích hợp dữ liệu -omics của dòng tế bào.

7. Cấu trúc của luận án

Ngoài phần mở đầu, mục lục, kết luận và tài liệu tham khảo, phần nội dung chính của luận án được chia thành 3 chương như sau:

Chương 1: Trình bày tổng quan các kiến thức nền tảng về dữ liệu y sinh học, các giải pháp nghiên cứu dự đoán đáp ứng thuốc gần đây. Trong đó làm rõ các khái niệm liên quan đến các loại dữ liệu sinh học (-omics) của các dòng tế bào (cell lines) như dữ liệu di truyền (geniomics), dữ liệu biểu hiện gen (transcriptomics), dữ liệu methyl hóa (epigenomics); các khái niệm về thuốc và đáp ứng thuốc cho điều trị đơn thuốc (monotherapy) và đáp ứng đa thuốc (combination therapy), đồng thời các kiểu dữ liệu biểu diễn phân tử thuốc (SMILES). Chương 1 cũng giới thiệu các nguồn dữ liệu sẵn có làm cơ sở cho các mô hình thực nghiệm. Bên cạnh đó, chương 1 cũng trình bày tổng quan một số phương pháp tính toán theo mô hình học sâu, mạng nơ-ron đồ thị; các phương pháp đánh giá hiệu năng của các mô hình dự đoán, từ đó đưa ra hướng tiếp cận và đề xuất các giải pháp có thể triển khai nhằm cải thiện hiệu năng dự đoán.

Chương 2: Trình bày giải pháp tích hợp dữ liệu trong dự đoán đơn thuốc với hai đề xuất cho bài toán này dựa trên mô hình mạng nơ-ron đồ thị. Cụ thể, đề xuất 1 là giải pháp học biểu diễn dữ liệu dạng đồ thị phân tử thuốc – GraphDRP. Đây là giải pháp biểu diễn đặc trưng phân tử thuốc dưới dạng đồ thị, áp dụng các mô hình mạng nơ-ron đồ thị khác nhau để học các đặc trưng ẩn của các phân tử thuốc đó đồng thời sử dụng các lớp mạng nơ-ron tích chập 1 chiều để học các biểu diễn đặc trưng của dữ

liệu gen di truyền (genomics) của các dòng tế bào để dự đoán đáp ứng thuốc cho dòng tế bào. Với đề xuất 2 – GraOmicDRP là một giải pháp tích hợp đa dạng các nguồn dữ liệu dữ liệu -omics của các dòng tế bào được học qua mạng nơ-ron tích chập với dữ liệu biểu diễn dữ liệu đồ thị được học qua mạng nơ-ron đồ thị để dự đoán đáp ứng thuốc cho các dòng tế bào. Kết quả nghiên cứu được so sánh đánh giá trên các kịch bản khác nhau và cho thấy hiệu quả vượt trội so với các phương pháp tiên tiến khác đã được khảo sát tại thời điểm đề xuất và được trình bày trong các công trình công bố số 1 và 2.

Chương 3: Trình bày giải pháp tích hợp dữ liệu trong dự đoán đáp ứng đa thuốc cho các dòng tế bào với hai đề xuất. Cụ thể, đề xuất 3 – GraOmicSynergy, là giải pháp dự đoán đáp ứng đa thuốc bằng cách tích hợp nhiều dữ liệu -omics của các dòng tế bào với dữ liệu biểu diễn đồ thị phân tử của các cặp thuốc được trích xuất thông qua mạng nơ-ron đồ thị đẳng cấu. Trong đó dữ liệu biểu diễn cặp thuốc tương tác với dòng tế bào được tổng hợp thông qua một mô-đun cơ chế chú ý đồng thời dữ liệu biểu diễn dòng tế bào được tổng hợp từ nhiều nguồn -omics khác nhau tạo thành một vec-tơ biểu diễn duy nhất qua các khối mạng nơ-ron tích chập 1 chiều. Chương 3 cũng trình bày tiếp tục giải pháp tích hợp đa dữ liệu -omics khác là đề xuất giải pháp 4 - AE-XGBSynergy. Trong đó AE-XGBSynergy thực hiện tích hợp nhiều dữ liệu -omics với thông tin cấu trúc mạng tương tác protein-protein (PPI) để dự đoán phân loại sự tương hợp và tương kháng giữa các cặp thuốc đối với các dòng tế bào. Kết quả nghiên cứu được so sánh đánh giá trên các kịch bản khác nhau và cho thấy hiệu quả vượt trội so với các phương pháp tiên tiến khác đã được khảo sát tại thời điểm đề xuất và được trình bày trong các công trình công bố số 5 và số 4.

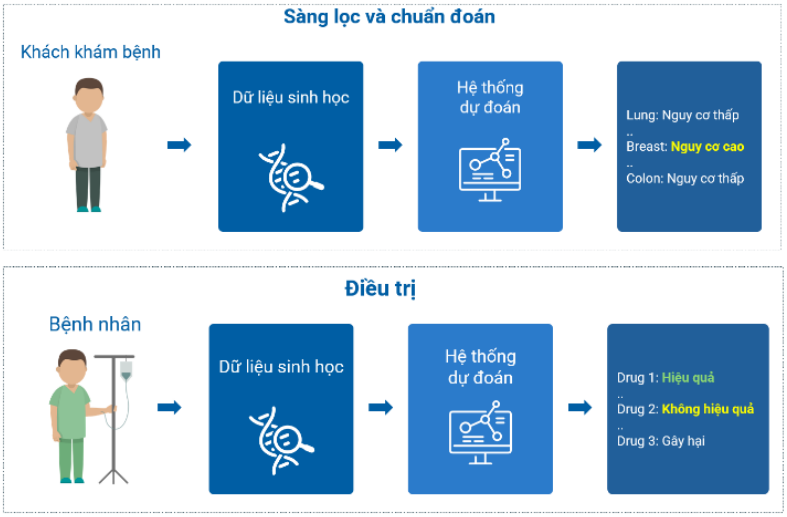
Các kết quả đạt được, các định hướng nghiên cứu tiếp theo của luận án cũng như các công trình nghiên cứu đã được công bố của tác giả được trình bày trong phần kết luận và kiến nghị của luận án.

PHẦN NỘI DUNG

CHƯƠNG 1 – TỔNG QUAN VỀ ĐÁP ỨNG THUỐC VÀ DỰ ĐOÁN ĐÁP ỨNG THUỐC

1.1. GIỚI THIỆU CHUNG

Cho đến gần đây, các phương pháp điều trị vẫn thường được thực hiện theo phương thức "one-size-fits-all" (điều trị đồng loạt, đại trà), mà không dựa trên các phân tích cụ thể về đặc điểm sinh học người bệnh. Điều này dẫn đến giảm hiệu quả điều trị thuốc, bởi có thể có người đáp ứng thấp, có người đáp ứng cao và không đáp ứng gì thậm chí có tác dụng phụ trong quá trình điều trị. Với sự phát triển nhanh chóng của công nghệ các hệ thống dự đoán sàng lọc và chẩn đoán bệnh giúp xác định bệnh chính xác hơn từ đó hệ thống dự đoán cũng cung cấp phương thức xác định loại thuốc có khả năng đáp ứng tốt nhất cho người bệnh [27] . Hình 1.1 minh họa hệ thống dự đoán tổng quát cho việc sàng lọc, chẩn đoán và điều trị bệnh.

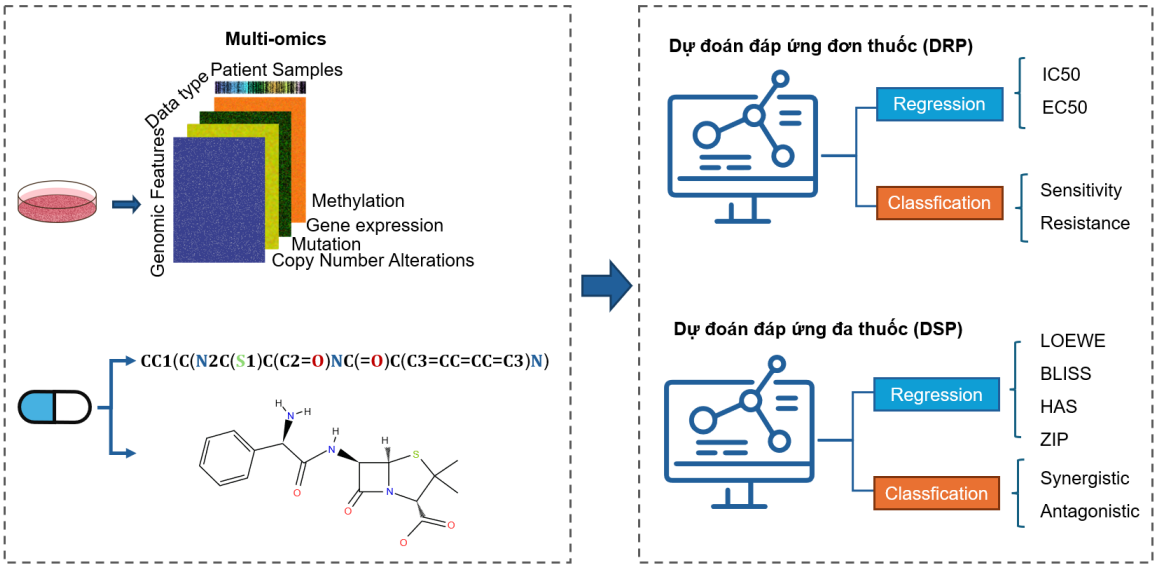


Hình 1.1. Hệ thống tổng quan cho dự đoán đáp ứng thuốc

Với mục tiêu của y học chính xác là xác định được phương thức điều trị chính xác cho từng bệnh nhân dựa trên đặc điểm sinh học của họ thì các phương pháp điều trị chính xác thường xem xét đến việc phân tích các dữ liệu về cấu trúc gen của bệnh nhân, các đặc trưng phân tử thuốc để đưa ra các quyết định điều trị tương ứng. Các mô hình thống kê truyền thống và các phương pháp tiếp cận máy học đã được sử dụng để xây dựng mô hình dự đoán phân loại đáp ứng trong môi trường lâm sàng

[28] và tiền lâm sàng [29]. Khi các mô hình dự đoán tăng độ phức tạp, số lượng quan sát cần thiết để huấn luyện các mô hình này cũng tăng lên. Trong khi dữ liệu sinh học và kết quả lâm sàng có thể sử dụng là nguồn dữ liệu phù hợp nhất để phát triển hệ thống dự đoán đáp ứng thuốc trong điều trị lâm sàng lại thường bị giới hạn về kích thước (nhỏ), chi phí thử nghiệm cao và các hạn chế và quy định phức tạp. Ngoài ra, về bản chất tự nhiên của thử nghiệm, việc thử nghiệm nhiều phương án điều trị cho cùng một bệnh nhân là không khả thi.

Việc dự đoán đáp ứng thuốc đòi hỏi các công cụ tính toán hiệu quả và số lượng mẫu đáng kể. Hiện nay, công nghệ sàng lọc thông lượng cao đang đóng góp một số lượng lớn dữ liệu sinh học về các dòng tế bào và bệnh nhân, từ đó giúp các nhà nghiên cứu xây dựng mô hình dự đoán để xác định đúng thuốc và liều thuốc hiệu quả hơn. Hai bài toán quan trọng của dự đoán đáp ứng thuốc là dự đoán đáp ứng thuốc đơn thuốc (monotherapy) và dự đoán đáp ứng đa thuốc (combination therapy) đang thu hút nhiều lượng lớn cộng đồng nghiên cứu quan tâm và đề xuất giải pháp.



Hình 1.2. Các mô hình đoán đáp ứng thuốc hiện nay

Hình 1.2 mô tả tổng quan các mô hình dự đoán đáp ứng thuốc, trong đó dữ liệu đầu vào là dữ liệu -omics biểu diễn các loại dữ liệu khác nhau của tế bào, thuốc được sàng lọc thử nghiệm khả năng đáp ứng thuốc được biểu diễn thành các dạng dữ liệu khác nhau của phân tử thuốc. Tất cả được đưa vào các mô hình dự đoán tương ứng để xác định mức giá trị đáp ứng hoặc phân loại mức độ đáp ứng khác nhau. Kết hợp thuốc được triển khai do những đặc điểm sinh học không đồng nhất của các dòng tế

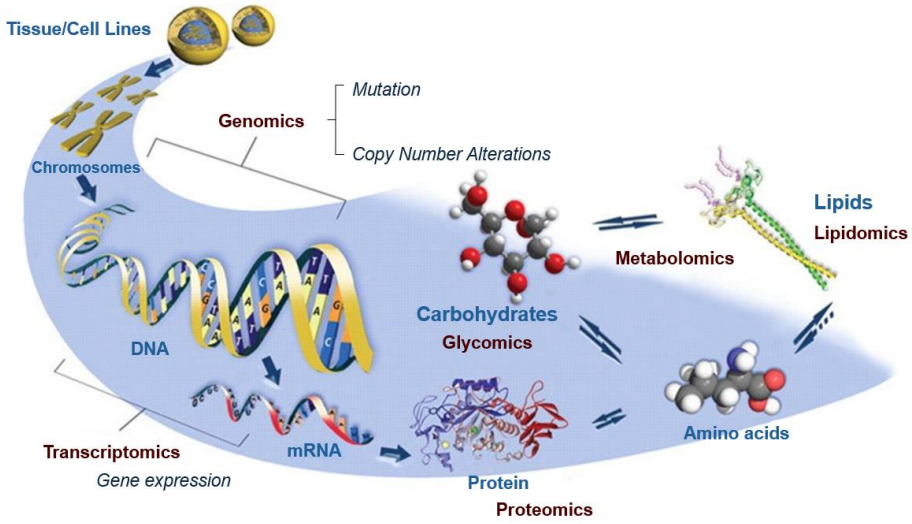
bào và sự kháng thuốc mắc phải, các liệu trình đơn trị liệu có thể không hiệu quả, cần có sự kết hợp bởi hai hay nhiều thuốc khác nhau cho một liệu trình điều trị. Do đó dự đoán liệu trình kết hợp thuốc cũng đang ngày càng được chú ý trong nghiên cứu tiền lâm sàng và lâm sàng.

Các mô hình dự đoán phần lớn dựa trên các dữ liệu về dòng tế bào (cell lines) hoặc mô ghép (xenografts) hơn là dữ liệu trên bệnh nhân do chi phí thấp, linh hoạt, dễ thử nghiệm. Các nghiên cứu có thể thực hiện thử nghiệm với một loại thuốc hay kết hợp nhiều thuốc trên các dòng tế bào hay kết hợp song song với thử nghiệm lâm sàng [29]. Mặc dù các mô hình dự đoán tiền lâm sàng còn có khoảng cách với điều trị thực tế nhưng chúng cung cấp những thông tin quan trọng nhằm định hướng điều trị chính xác hơn.

1.2. TỔNG QUAN VỀ DỮ LIỆU -OMICS VÀ ĐÁP ỨNG THUỐC

1.2.1. Dữ liệu -omics

Dữ liệu -omics được hiểu là dữ liệu được tạo ra từ các công nghệ giải trình tự thông lượng cao được sử dụng để nghiên cứu cấu trúc, tổ chức sinh học khác nhau của một sinh vật, chẳng hạn như bộ gen (tất cả các vật liệu di truyền), bộ phiên mã (tất cả các phân tử RNA), bộ protein (tất cả các protein).



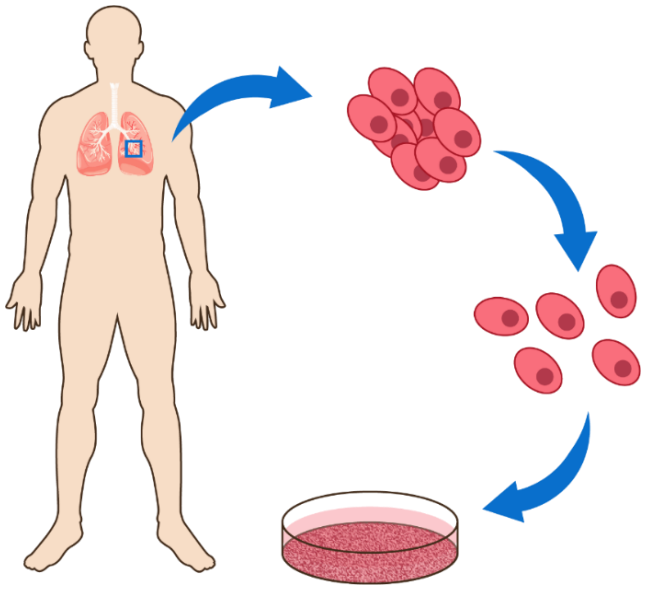
Hình 1.3. Cơ chế sinh học và các dạng dữ liệu -omics của tế bào [30]

Dữ liệu -omics thường được sử dụng trong hệ thống sinh học và gen chức năng để nghiên cứu mối quan hệ giữa các phân tử khác nhau và cách chúng tương tác có ảnh hưởng đến chức năng tổng thể của tế bào, mô và sinh vật [31]. Dữ liệu -omics có

thể phức tạp, nhiều chiều, nhiễu và đòi hỏi các phương pháp và công cụ tính toán chuyên dụng để phân tích và giải thích. Hàng loạt các công nghệ "-omics" như genomics (gen di truyền) transciptomics (phiên mã), epigenomics (di truyền biểu sinh), interactomics (dữ liệu mạng tương tác) ra đời cho phép khám phá bộ gen, bộ phiên mã, dữ liệu mạng tương tác rộng hơn, đồng thời cung cấp các thông tin để phát hiện mục tiêu (target), đặc tính dược lý học, độc tính và khả năng an toàn của thuốc. Từ đó có thể xây dựng mô hình sàng lọc, chuẩn đoán và chăm sóc sức khỏe cá nhân.

1.2.1.1. Dòng tế bào

Các dòng tế bào (cell lines) là các khối tế bào bệnh sống được nuôi cấy trong trong phòng thí nghiệm, mang đầy đủ đặc trưng sinh học bệnh. Để tạo dòng tế bào, các mảnh từ khối u của bệnh nhân được đưa vào môi trường nuôi cấy tế bào trong một tủ ấm đặc biệt sau đó được theo dõi thường xuyên. Các tế bào sẽ tiếp tục sinh sản tạo ra một nguồn tế bào liên tục để nghiên cứu. Trong nghiên cứu ung thư, tập hợp các dòng tế bào có nguồn gốc từ khối u thường được sử dụng làm mô hình nghiên cứu vì chúng mang hàng trăm đến hàng nghìn biến đổi gen trong khối u mà chúng được tạo ra từ đó. Các dòng tế bào ung thư được sử dụng rộng rãi trong các nghiên cứu dược lý học và đáp ứng thuốc [32].



Hình 1.4. Minh họa nuôi cấy tế bào ung thư trong phòng thí nghiệm

1.2.1.2. Đột biến gen và biến thể số lượng bản sao

Có nhiều yếu tố ảnh hưởng đến tình trạng sức khỏe và bệnh tật, trong đó nền tảng di truyền của mỗi cá nhân là một yếu tố quyết định quan trọng. Việc kiểm tra cấu trúc di truyền này là điều quan trọng lớn đối với việc xác định các đột biến hoặc các biến thể riêng lẻ làm cơ sở cho việc xác định tình trạng sức khỏe và bệnh tật. Nhờ công nghệ giải trình tự gen thông lượng cao, thông tin về hệ gen được tạo ra với số lượng ngày càng tăng đã cho phép chuyển đổi từ các nghiên cứu tập trung vào các gen riêng lẻ sang so sánh bộ gen của toàn bộ quần thể. Có nhiều đột biến tồn tại trong bộ gen, trong đó phần lớn là lành tính; một số đột biến có tính chất bảo vệ, mang lại lợi thế chống lại một số điều kiện, nhưng cũng có một số khác có thể có hại. Các đột biến này phát triển ngày càng tăng với một tình trạng (nhóm các đột biến có thâm nhập) hoặc trực tiếp gây ra bệnh (một hoặc một số đột biến có khả năng xâm nhập cao) [33]. Các dữ liệu về đột biến như đột biến gen (MUT) và biến thể số lượng bản sao (CNA) cung cấp các thông tin quan trọng trong việc nghiên cứu các dấu ấn sinh học dự đoán bệnh.

1.2.1.3. Biểu hiện gen

Biểu hiện gen (Gene expression - GE) là quá trình truyền thông tin di truyền trong một gen vào cấu trúc đang có trong tế bào sống, tính trạng tương ứng được tạo thành có thể quan sát được ở kiểu hình. Dữ liệu biểu hiện gen này cung cấp thông tin cơ bản để hiểu rõ hơn về quá trình chuyển hóa tế bào và mô, đồng thời đánh giá những thay đổi trong quá trình phiên mã có ảnh hưởng đến sức khỏe và bệnh tật như thế nào. Trong đó, hệ phiên mã (transcriptome) cung cấp các thông tin cần thiết cho việc giải thích chức năng của hệ gen và khám phá thành phần phân tử của các tế bào và mô. Ngày nay, việc giải trình tự hệ phiên mã thông lượng cao (microarray và RNA-seq) cho phép: (1) lập danh mục tất cả các loại phiên mã; (2) để xác định cấu trúc phiên mã của gen, gồm vị trí khởi đầu sao chép, đầu 5’ và 3’, kiểu cắt nối và những biến đổi sau dịch mã; (3) định lượng sự thay đổi mức độ biểu hiện của mỗi bản phiên mã trong quá trình phát triển và dưới các điều kiện khác nhau. Qua đó lượng lớn dữ liệu biểu hiện gen được cung cấp cho nghiên cứu và ứng dụng trong điều trị bệnh.

1.2.1.4. Methyl hóa DNA

Di truyền biểu sinh (Epigenomic – METH) đề cập đến những thay đổi di truyền trong biểu hiện gen mà không có bất kỳ thay đổi nào trong trình tự DNA. Epigenomics đề cập đến việc phân tích các thay đổi methyl hóa trên toàn bộ bộ gen, cho biết thông tin di truyền ngoài trình tự DNA có thể ảnh hưởng đến chức năng của gen. Điều hòa biểu sinh có thể được bổ sung bởi năm cơ chế khác nhau: methyl hóa DNA, biến đổi sau dịch mã của histone, các biến thể của histone [34], can thiệp RNA, và tổ chức nhân. Dữ liệu methyl hóa (METH) là thông số bộ gen linh hoạt phổ biến nhất cho thấy sự thay đổi chức năng bộ gen dưới tác động ngoại sinh.

1.2.1.5. Mạng tương tác protein

Protein là các đại phân tử trong nhân tế bào, đóng vai trò quan trọng nhiều nhiệm vụ bao gồm làm những enzym xúc tác cho các phản ứng hóa học, vận chuyển chất dinh dưỡng, duy trì và phát triển mô. Các protein không hoạt động độc lập mà thường kết hợp với các protein tạo nên một mạng liên kết sinh học phức tạp gọi là mạng tương tác protein (PPIs). Những tương tác này tác động tới hoạt động phát triển của tế bào, do đó những bất thường trong mạng tương tác protein là nguyên nhân của một số bệnh (ví dụ: bệnh Alzheimer). Dữ liệu mạng tương tác (interactomics) có thể giúp khám phá các đặc tính sinh học của protein hoặc phát triển thuốc nhắm mục tiêu (drug-target) [35].

1.1.2. Thuốc

1.1.2.1. Đáp ứng thuốc

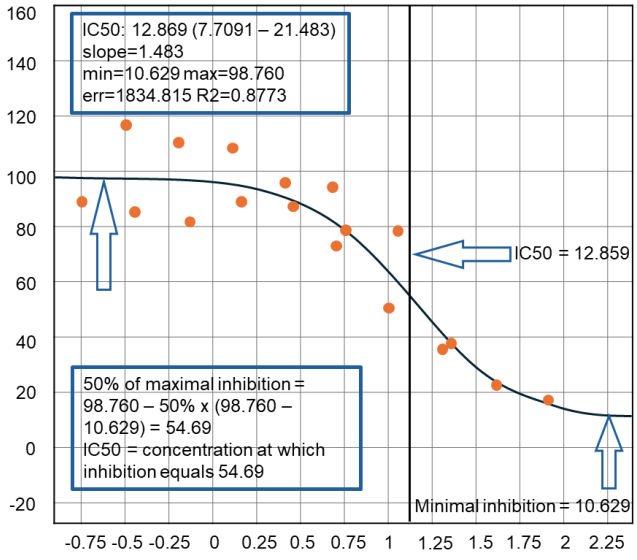
Thuốc là hợp chất hóa học được cấu tạo bởi các nguyên tử và tương tác giữa chúng, thuốc gây ra sự thay đổi trong sinh lý hoặc tâm lý của sinh vật khi được tiêu thụ. Hiện nay có rất nhiều loại thuốc được nghiên cứu đơn lẻ cũng như kết hợp các cặp thuốc để điều trị các bệnh khác nhau.

Đáp ứng thuốc là kết quả của quá trình tương tác giữa thuốc với các thành phần của tế bào trong cơ thể, tạo nên những đáp ứng của các tổ chức đối với thuốc. Thuốc thường có tác dụng tăng cường hoặc gây ức chế một hoặc một vài chức năng nào đó của cơ thể chứ không tạo ra chức năng mới [4]. Đối với phương thức điều trị đơn thuốc, đáp ứng thuốc được hiểu là phép đo khả năng của một thuốc trong việc ức

chế chức năng sinh học của tế bào bệnh, trong khi đối với phương thức điều trị kết hợp thuốc thì nó được hiểu là khả năng kết hợp hai hay nhiều thuốc trong việc ức chế chức năng sinh học tế bào bệnh đó. Đáp ứng thuốc có thể bị ảnh hưởng bởi một số yếu tố bao gồm chế độ ăn uống, bệnh đi kèm, tuổi tác, cân nặng, tương tác thuốc - thuốc và di truyền. Biến thể di truyền riêng lẻ trong các gen quan trọng liên quan đến chuyển hóa, vận chuyển hoặc mục tiêu thuốc (drug target) có thể góp phần vào nguy cơ xảy ra các tác dụng ngoài ý muốn hoặc thất bại trong điều trị.

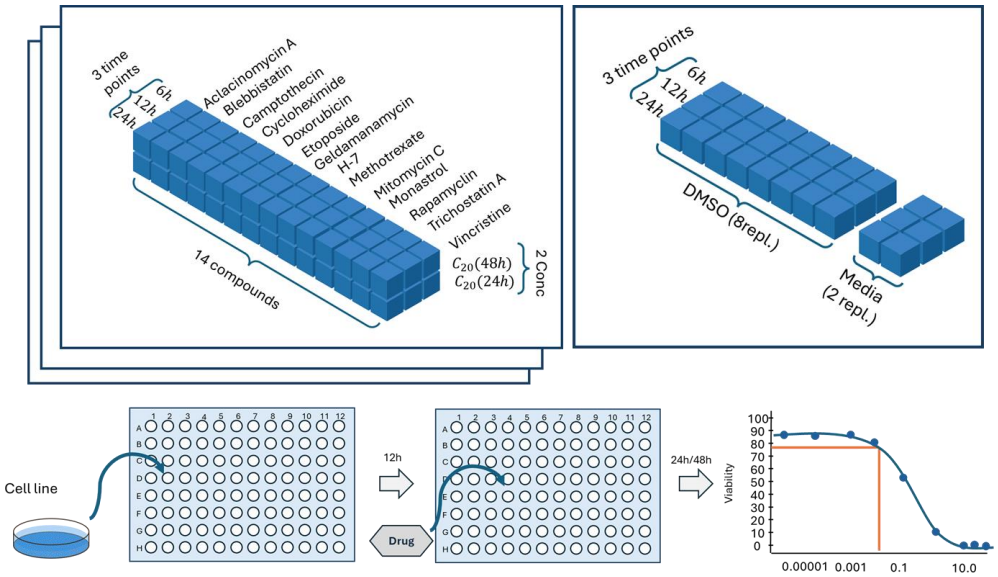
1.1.2.2. Phép đo đáp ứng thuốc

Nồng độ ức chế tối đa một nửa (IC_{50}) là phép đo về khả năng của một thuốc trong việc ức chế chức năng sinh học cụ thể, hay là mức độ cần thiết của thuốc để ức chế một quá trình sinh học hoặc thành phần sinh học nhất định bằng 50%. Các thành phần sinh học đó có thể là enzyme hoặc tế bào. Do đó, đáp ứng thuốc của dòng tế bào được định lượng dựa trên nồng độ của thuốc và tỷ lệ sống của dòng tế bào.



Hình 1.5. Phép đo đáp ứng thuốc - IC_{50}

Việc thử nghiệm lâm sàng trên bệnh nhân và động vật thường tốn kém, mất nhiều thời gian. Do đó đáp ứng thuốc thường được thử nghiệm trên các dòng tế bào, như là dữ liệu nghiên cứu tiền lâm sàng quan trọng. Hình 1.6 minh họa cho việc đo đáp ứng thuốc: Các khay dòng tế bào ung thư mẫu được đổ thuốc ở các thời điểm khác nhau: ví dụ 6 giờ, 12 giờ và 24 giờ, sau đó tiến hành theo dõi đo nồng độ đáp ứng thuốc.

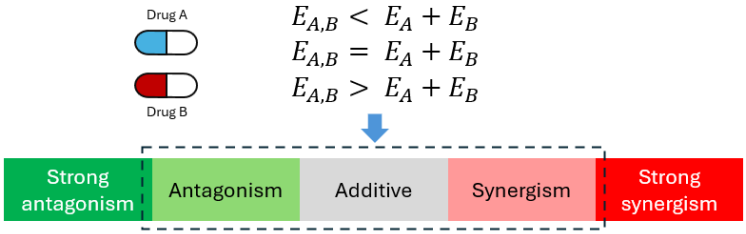


Hình 1.6. Ví dụ minh họa quá trình đo đáp ứng thuốc IC_{50} [36]

Ngoài IC_{50}, một số độ đo khác cũng được sử dụng để đo độ đáp ứng thuốc như: AUC, EC_{50}, GC_{50}. Trong đó AUC (Area Under the Curve) là diện tích vùng dưới đường cong đồ thị, biểu diễn sự biến thiên của nồng độ thuốc trong máu theo thời gian, biểu diễn tượng trưng cho lượng thuốc vào được đại tuần hoàn ở dạng còn hoạt tính sau một khoảng thời gian; EC_{50} đại diện cho liều hoặc nồng độ trong huyết tương cần thiết để đạt được 50% hiệu quả tối đa; GC_{50} (ức chế sinh trưởng 50%) là nồng độ thuốc làm giảm một nửa tốc độ tăng trưởng.

1.1.2.3. Kết hợp thuốc

Trong mô hình tương tác thuốc và độc dược học, vấn đề thường được quan tâm là tác dụng tương hợp hoặc tương kháng (synergistic hoặc antagonistic) giữa các hợp chất sinh học. Khi kết hợp hai hoặc nhiều hợp chất, hiệu ứng tổng hợp của chúng có thể lớn hơn nhiều so với các hiệu ứng riêng lẻ. Tác dụng kết hợp như vậy cũng có thể làm giảm độc tính bằng cách cho phép sử dụng liều thấp hơn của một trong hai loại thuốc để đạt được hiệu quả tương tự. Điều này có thể ứng dụng trong nhiều lĩnh vực đặc biệt là hóa trị liệu [37]. Hình 1.7 minh họa mức độ đáp ứng khi kết hợp thuốc A với thuốc B trong điều trị. Nếu coi E_A, E_B là mức độ ảnh hưởng điều trị của thuốc A và thuốc B lên dòng tế bào thì E_{AB} là mức độ ảnh hưởng điều trị kết hợp thuốc A và thuốc B lên dòng tế bào đó. E_{AB} có thể phân loại ở các ngưỡng khác nhau như: rất tương kháng (strong antagonism), tương kháng (antagonism), tăng ảnh hưởng (additive), tương hợp (synergism), rất tương hợp (strong synergism)



Hình 1.7. Mức độ đáp ứng đa thuốc

Có nhiều mô hình định lượng kết hợp thuốc được đề xuất tuy nhiên có 4 mô hình phổ biến nhất [38] là LOEWE (Loewe Additivity), BLISS (Bliss Independence), HAS (Highest Single Agent), ZIP (Zero Interaction Potency). Cụ thể các phương pháp tính toán như sau:

HSA

Mô hình HSA là một trong những mô hình tham chiếu đơn giản nhất, mô hình này cho biết hiệu quả kết hợp dự kiến là mức tối đa của các đáp ứng thuốc đơn lẻ ở các nồng độ tương ứng. Do đó, giá trị kết hợp SHSA được định nghĩa là

$$S_{HSA} = E_{A,B,\dots,N} - max(E_A, E_b, \dots, E_N) \quad (1.1)$$

Với $E_{A,B,\dots,N}N$ là tác dụng kết hợp giữa N thuốc và $E_A, E_b, \dots, E_N$ là đáp ứng của các thuốc riêng lẻ.

Bliss

Bliss giả định một quy trình ngẫu nhiên trong đó hai loại thuốc phát huy tác dụng của chúng một cách độc lập và hiệu ứng kết hợp dự kiến có thể được tính toán dựa trên xác suất của các sự kiện độc lập. Do đó, giá trị kết hợp thuốc Bliss, được định nghĩa là:

$$S_{Bliss} = E_{A,B,\dots,N} - 100(1 - \left(1 - \frac{E_A}{100}\right)\left(1 - \frac{E_B}{100}\right) \dots \left(1 - \frac{E_N}{100}\right)) \quad (1.2)$$

Với $1 - \frac{E_A}{100}$ là xác suất thuốc *A,B,...,N* không ức chế đích,

$(1 - \left(1 - \frac{E_A}{100}\right)\left(1 - \frac{E_B}{100}\right) \ldots \left(1 - \frac{E_N}{100}\right))$, cho biết ít nhất một thuốc đáp ứng mục tiêu.

Loewe

Coi hai thuốc độc lập tuyến tính trong đóng góp vào khả năng kết hợp hai thuốc. Về mặt toán học, nếu liều thuốc a' của thuốc A tạo ra tác dụng tương tự như liều b' của thuốc B, thì bất kỳ sự kết hợp liều nào khác như (a, b) cho một tỷ lệ tương ứng, đảm bảo

$$\frac{a}{a'} + \frac{b}{b'} = 1 \qquad (1.3)$$

Trong đó, a' và b' có thể được thay thế bằng nghịch đảo của log-logistic là $f_a^{-1}(E)$ và $f_b^{-1}(E)$ tương ứng cho đáp ứng E.

Do vậy, với liều đáp ứng x_A, x_B, ..., x_N của thuốc *A,B,...,N*, Loewe định lượng kết hợp thuốc S_{Loewe} như sau:

$$S_{Loewe} = E_{A,B,\ldots,N} - E_{Loewe} \qquad (1.4)$$

Với E_{Loewe} thỏa mãn: $\sum_{k=A,B,\ldots,N} (f_k^{-1}\left(\frac{x_k}{E_{Loewe}}\right)=1$

Zip

Zip giả định rằng hai loại thuốc không tương tác dự kiến sẽ gây ra những thay đổi tối thiểu trong đường cong liều lượng-phản ứng của chúng. ZIP được định lượng qua kết hợp nhiều loại thuốc đo được với đáp ứng $E_{A,B,\ldots,N}$

$$S_{Zip} = EE_{A,B,\ldots,N} - (\sum \overline{E} - \sum \frac{\overline{E}_A \times \overline{E}_B}{100} + \sum \frac{\overline{E}_A \times \overline{E}_B \times \overline{E}_C}{1000} - \cdots \pm \sum \frac{\overline{E}_A \times \overline{E}_B \times \ldots \times \overline{E}_N}{10^{N-1}} \qquad (1.5)$$

Mặc dù các mô hình có cách tính chỉ số kết hợp thuốc khác nhau nhưng cho đến nay, không có khẳng định nào về mô hình tốt nhất hoặc phương pháp hay nhất về cách xác định độ kết hợp thuốc tốt nhất. Tuy vậy, Loewe được sử dụng rộng rãi hơn cả trong các phương pháp dự đoán kết hợp thuốc.

1.1.2.4. Dữ liệu biểu diễn thuốc

1.1.2.4.1. SMILES

SMILES (Simplified Molecular Input Line Entry System) [39] là hệ thống ký hiệu hóa học đơn giản hóa mô tả các nguyên tử và liên kết giữa các nguyên tử trong phân tử theo cách ngắn gọn cho phép nhà hóa học, người dùng biểu diễn cấu trúc hóa học theo các quy tắc cú pháp cơ bản. Trong dạng biểu diễn SMILES, các nguyên tử, số liên kết, mạch vòng, rẽ nhánh,... được biểu diễn bằng ký tự đặc trưng.

Một số các nguyên tắc cơ bản của chuỗi SMILES như sau:

- Nguyên tử (Atoms): được thể hiện bằng các ký hiệu nguyên tử của chúng. Ngoài ra, các nguyên tử kim loại được biểu diễn bằng các ký hiệu trong dấu ngoặc vuông, ví dụ: Clo [Cl].

- Liên kết (Bond)- liên kết đơn, đôi và ba lần lượt được biểu diễn bằng các ký hiệu -, = và #. Liên kết đơn là mặc định và do đó không cần phải chỉ định.

- Tính thơm (Aromaticity) - Trong khi các ký hiệu nguyên tử thường được sử dụng trong chữ hoa, chẳng hạn như C, O, S và N; để chỉ định các nguyên tử thơm, các ký hiệu chữ thường được sử dụng thay thế, chẳng hạn như c, o, s và n. Đôi khi các liên kết ẩn trong các vòng (xen kẽ = và -) cũng được sử dụng để mô tả các nguyên tử thơm như C1=CC=CC=C1.

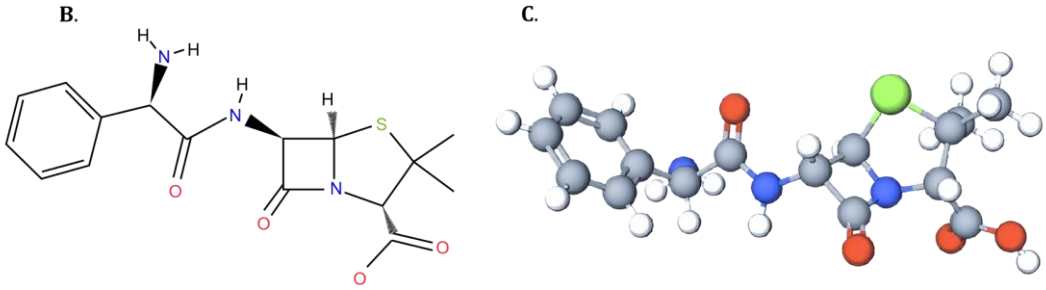
- Mạch vòng (Ring) - SMILES cho phép người dùng xác định cấu trúc vòng bằng cách sử dụng các số để xác định nguyên tử mở và đóng vòng. Ví dụ: trong C1CCCCC1, carbon đầu tiên có số “1” liên kết bằng một liên kết đơn với carbon cuối cùng cũng có số “1”. Cấu trúc thu được là xiclohexan.

- Các nhánh (Branches) - được chỉ định bằng cách đặt chúng trong dấu ngoặc đơn và có thể được lồng vào nhau hoặc sắp xếp. Ví dụ, 2-Propanol được đại diện bởi CC(O)C.

Các cấu trúc hóa học của thuốc có thể được biểu diễn ở các dạng khác nhau như cấu trúc dữ liệu một chiều (1D), hai chiều (2D) và ba chiều (3D). Hình 1.8 minh họa các dạng dữ liệu biểu diễn khác nhau thuốc ampicillin. Dạng 1D của thuốc thường mã hóa các thông tin như số nguyên tử, số liên kết, khối lượng phân tử dưới dạng biểu diễn chuỗi ký tự hóa học. Các dạng biểu diễn 2D được mô tả dưới dạng đồ thị, trong đó các nguyên tử được biểu diễn dưới dạng các nút của đồ thị trong khi các liên

kết và nhánh được biểu diễn dưới dạng liên kết (cạnh) của đồ thị. Ưu điểm của dạng này biểu diễn rõ ràng chứa tương đối đủ thông tin và tăng khả năng học cho máy học. Dữ liệu phân tử biểu diễn 3D chứa thông tin về tọa độ của các nguyên tử và liên kết, thông tin biểu diễn dạng 3D thường không có sẵn cho tất cả các hợp chất.

A. CC1(C(N2C(S1)C(C2 = O)NC(= O)C(C3 = CC = CC = C3)N)C(= O)O)C



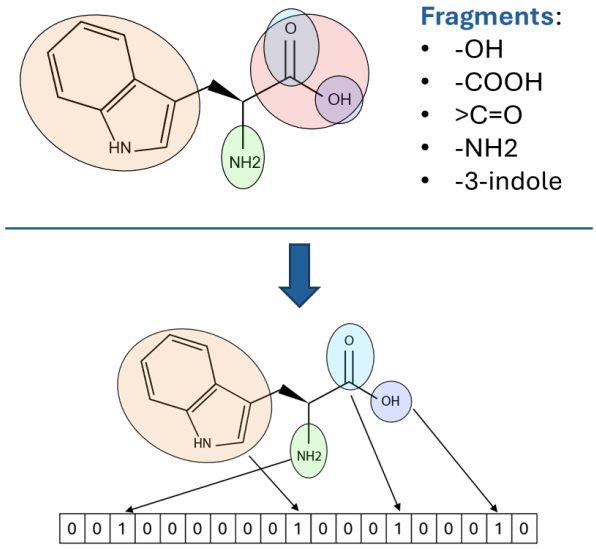
Hình 1.8. Các dạng biểu diễn cấu trúc hóa học của phân tử thuốc

Chuỗi SMILES được sử dụng rộng rãi trong việc lưu trữ và trao đổi dữ liệu cấu trúc hoá học. Từ đó có thể biểu diễn các chuỗi hóa học khác nhau thành các dạng dữ liệu có thể được sử dụng trong các chương trình máy tính. Các bài toán học máy về dự đoán khai phá đặc trưng thuốc thường biểu diễn phân tử thuốc dưới dạng các biểu diễn phân tử khác nhau như dạng chuỗi SMILES hoặc dạng đồ thị. Một chuỗi SMILES này tiếp tục có thể được chuyển đổi thành một số định dạng khác nhau như dấu vân tay phân tử (molecular fingerprint hay fingerprint), mã hóa dưới dạng one-hot (one-hot encoding) hoặc word embedding. Trong khi đó, dạng đồ thị của phân tử thuốc biểu diễn như một ma trận kề kết hợp với ma trận thuộc tính nguyên tử (node) và ma trận đặc trưng cạnh (bond) được sử dụng trực tiếp hoặc biến đổi thành graph embedding cho các tác vụ của mô hình khai phá thuốc.

1.1.2.4.2. Fingerprints

Fingerprints (FP) là kỹ thuật biểu diễn thuốc được sử dụng rộng rãi cho nhiều mô hình khai phá thuốc [40]. Chúng được tổng hợp từ chuỗi SMILES được định nghĩa trước hoặc thông qua các phương pháp tiếp cận dựa trên băm (hash-based approaches). Kiểu dữ liệu cho fingerprints là dạng one-hot vec-tơ. Kiểu dữ liệu này có nhược điểm là cần dựa trên các quy tắc được định nghĩa một cách thủ công trước nên có thể không biểu diễn hết các khía cạnh quan trọng của thuốc và chúng thường

có số chiều lớn (ví dụ: 881, 1024). Hình 1.9 minh họa ví dụ việc phân đoạn phân tử hóa học thành các dấu hiệu riêng và mã hóa dạng one-hot.



Hình 1.9. Biểu diễn thuốc theo Fingerprint

1.1.3. Nguồn dữ liệu y sinh học

Cơ sở dữ liệu về thuốc

ChEMBL [41] là một cơ sở dữ liệu được truy cập rộng rãi khác lưu trữ thông tin về các đặc tính hóa học, mục tiêu protein (protein targets) và hoạt tính sinh học của 1,9 triệu hợp chất. Ngoài ra một loạt các dự án cung cấp nguồn dữ liệu liên quan đến thông tin hóa học, dược lý và dược phẩm cho hơn 500.000 thuốc và các sản phẩm của thuốc như CheMBL [42], ZINC [43], KEGG [34] là cơ sở dữ liệu hóa học quan trọng có giá trị cho các nghiên cứu khám phá thuốc.

Nguồn dữ liệu y sinh học cho dòng tế bào

Hai trong số các dự án lớn cung cấp nguồn tài nguyên công khai quan trọng nhất để nghiên cứu đáp ứng thuốc là Bách khoa toàn thư về dòng tế bào ung thư (CCLE - Cancer Cell Line Encyclopedia) [44] và Genomics of Drug Sensitivity in Cancer (GDSC) [45], cung cấp dữ liệu -omics và dữ liệu đáp ứng thuốc với các dòng tế bào ung thư (Bảng 1.1). Trong khi CCLE chứa dữ liệu về đột biến (mutation), các biến thể số lượng bản sao của gen (copy number variant, CNV/CNA) và dữ liệu biểu hiện gen (gene expression, GE) từ hơn 1000 dòng tế bào từ 36 khối u với hơn 11.000 thực nghiệm trên 24 loại thuốc chống ung thư trên hơn 500 dòng tế bào thì GDSC

cung cấp dữ liệu khoảng 75 nghìn thử nghiệm đã được tiến hành trên 256 loại thuốc chống ung thư cho hơn 1 nghìn dòng tế bào từ các loại ung thư khác nhau. Ngoài ra NCI-60 [46], Cổng thông tin đáp ứng điều trị ung thư - CTRP (Cancer Therapeutics Response Portal), hay cổng thông tin chính phủ Mỹ về các thử nghiệm lâm sàng trong ung thư - GAO (Government Accountability Office) là các nguồn dữ liệu tốt cho các nghiên cứu tiền lâm sàng và lâm sàng.

Bảng 1.1. Nguồn dữ liệu -omics cho dòng tế bào

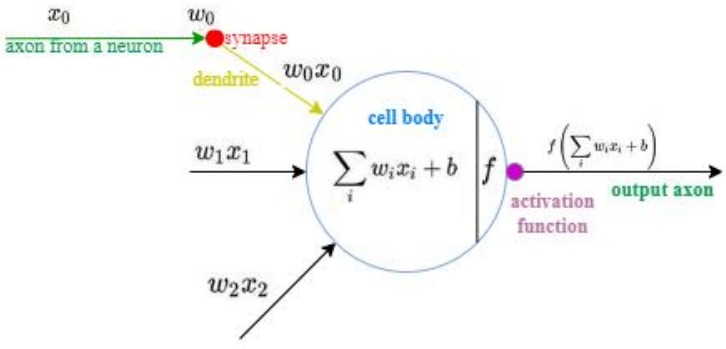
	CCLE	**GDSC**	**NCI-60**
# Dòng tế bào	>1000	>1000	60
# Thuốc	24	266	>15.000
# Thử nghiệm	>11.000	> 75.000	> 100.000
#-omics	MUT, CAN, GE	MUT, CAN, GE, METH	MUT, CAN, GE, METH,
# Mô	36	>15	9

1.3. TỔNG QUAN VỀ CÁC PHƯƠNG PHÁP DỰ ĐOÁN ĐÁP ỨNG THUỐC

1.3.1. Mô hình học sâu

1.3.1.1. Mạng nơ-ron nhân tạo

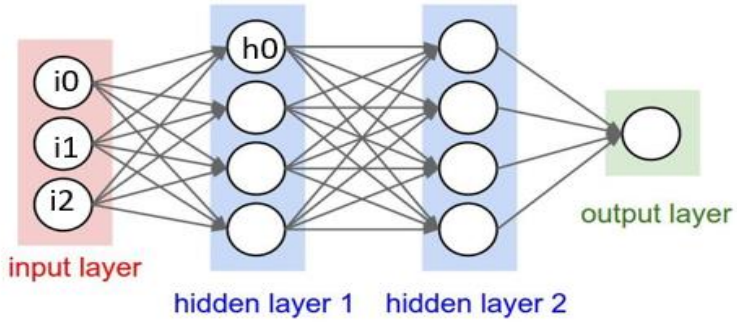
Mạng nơ-ron nhân tạo (ANN: Artificial Neural Networks) là một khối các đơn vị tính toán (nơ-ron) kết nối với nhau mô phỏng hoạt động như các nơ-ron thần kinh của bộ não con người. Mỗi nơ-ron nhân tạo nhận tín hiệu từ các nơ-ron được kết nối, sau đó xử lý chúng và gửi tín hiệu đến các nơ-ron được kết nối khác. Đầu ra của mỗi nơ-ron được tính thông qua một số hàm phi tuyến tính của tổng đầu vào của nó, được gọi là hàm kích hoạt. Cường độ tín hiệu ở mỗi kết nối được xác định bằng trọng số (w), trọng số này sẽ điều chỉnh trong quá trình họ (Hình 1.10).



Hình 1.10. Nơ-ron nhân tạo

Kiến trúc của mạng nơ-ron

Thông thường, các nơ-ron được tập hợp thành các lớp. Các lớp khác nhau có thể thực hiện các phép biến đổi khác nhau trên đầu vào của chúng. Tín hiệu truyền từ lớp đầu tiên (lớp đầu vào) đến lớp cuối cùng (lớp đầu ra), có thể đi qua nhiều lớp trung gian (lớp ẩn). Với mạng nơ-ron thông thường, lớp kết nối đầy đủ hay Fully Connected (FC) là kiến trúc hay được sử dụng nhất (Hình 1.11). Một mạng thường được gọi là mạng nơ-ron sâu nếu nó có ít nhất hai lớp ẩn.



Hình 1.11. Mạng nơ-ron kết nối đầy đủ với các lớp ẩn

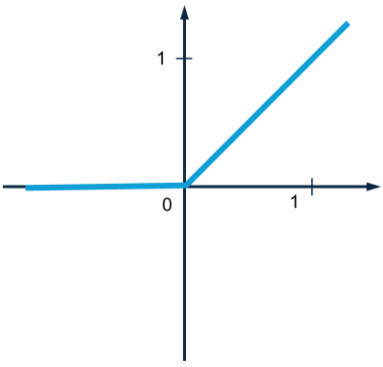
Các hàm kích hoạt phổ biến

Hàm kích hoạt đóng vai trò quan trọng, là thành phần phi tuyến tại đầu ra của các nơ-ron. Một số hàm kích hoạt hay được sử dụng hiện nay như: ReLU, LeakyReLU.

Hàm ReLU

Hàm ReLU (Rectified Linear Unit) chỉ lọc các giá trị với ngưỡng là 0 (Hình 1.12). Công thức của hàm ReLU là:

$$f(x) = max\ (0, x) \tag{1.6}$$



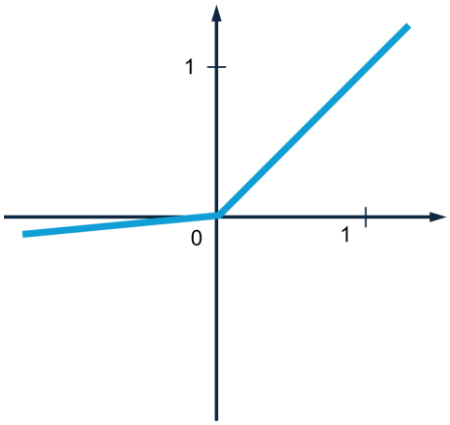
Hình 1.12. Hàm ReLU

Hàm LeakyReLU

Leaky ReLU là một biến thể của ReLU, cùng với ReLU được sử dụng rộng rãi gần đây. Thay vì trả về giá trị 0 như ReLU với các đầu vào ≤ 0 thì Leaky ReLU tạo ra một đường xiên có độ dốc nhỏ (0.01) (Hình 1.13). Leaky ReLU được định nghĩa như sau:

$$f(x) = \begin{cases} x & x > 0 \\ 0.01x & x \leq 0 \end{cases} \quad (1.7)$$

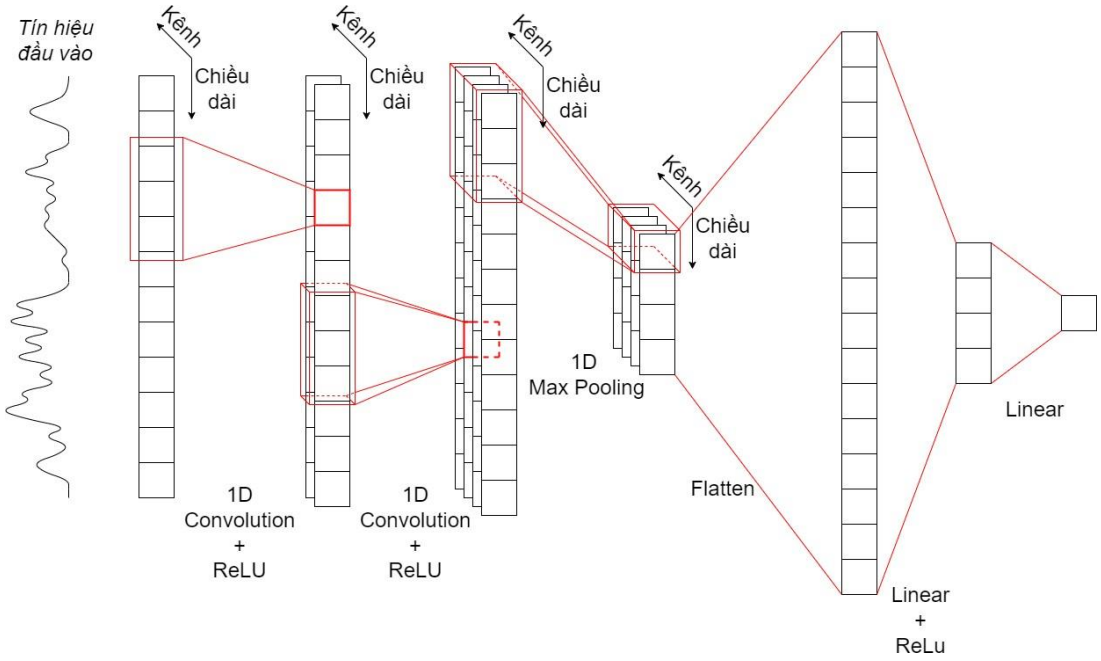
Trong nhiều trường hợp Leaky ReLU được đánh giá hiệu quả hơn ReLU.



Hình 1.13. Hàm Leaky ReLU

1.3.1.2. Mạng nơ-ron tích chập

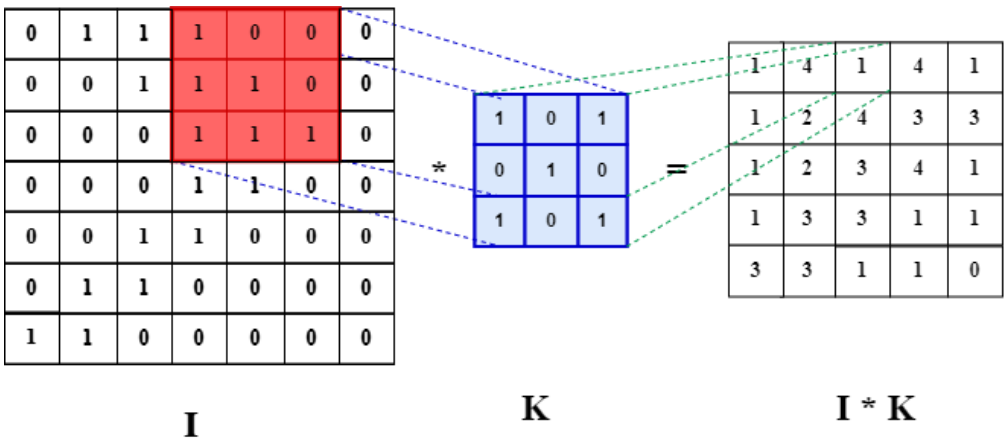
Mạng nơ-ron tích chập (CNN: Convolutional Neural Network) là một trong những mô hình học sâu tiên tiến ứng dụng trong các bài toán thị giác máy tính (computer vision) và nhiều lĩnh vực học máy khác nhau [47]. Thông qua cơ chế tích chập (convolution), mạng nơ-ron tích chập được hình thành từ các lớp liên kết với nhau, kết quả tích chập trước là đầu vào cho lớp sau. Trong khi mạng CNN2D, CNN3D thường được sử dụng để trích xuất các đặc tính không gian của dữ liệu như ảnh 2D, ảnh 3D, video... thì mạng nơ-ron tích chập 1-chiều (CNN1D) thường nhận đầu vào là những dữ liệu 1-chiều, ví dụ như tín hiệu trên miền thời gian, văn bản, tín hiệu sinh học v.v… Về cơ bản, những dữ liệu đó ở dạng ma trận số, có hai chiều là độ dài và độ sâu, hay còn gọi là kênh. Mỗi khối tích chập 1-chiều (1D Convolution) bao gồm nhiều bộ lọc. Phép tích chập được thực hiện giữa mỗi bộ lọc với ma trận số đầu vào. Kết quả đầu ra là một ma trận số mới với số lượng kênh bằng với số lượng bộ lọc. Cuối cùng, ma trận số được cho qua một hàm phi tuyến (ví dụ, ReLU).



Hình 1.14. Mô hình mạng nơ-ron tích chập 1-chiều CNN-1D

Hình 1.14 minh hoạ một kiến trúc mạng nơ-ron 1-chiều, với hai lớp tích chập một chiều, sau mỗi lớp tích chập, số lượng kênh tăng gấp đôi, sau mỗi lớp pooling, chiều dài giảm đi ba lần.

Lớp tích chập - Convolutional Layer

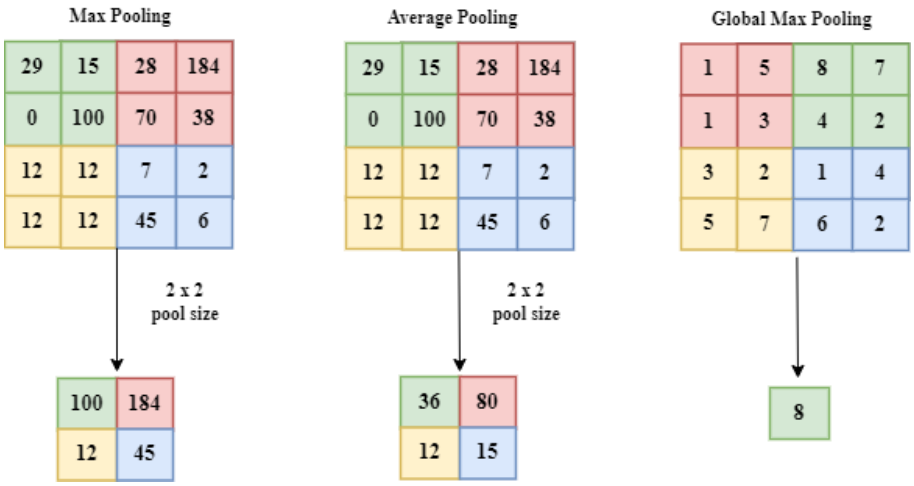


Hình 1.15. Phép toán tích chập

Mục đích của lớp tích chập (Hình 1.15) là biến đổi cục bộ các thuộc tính của các vec-tơ đầu vào, đồng thời làm thay đổi chiều không gian, tạo ra các góc nhìn (kênh) sâu hơn, phức tạp hơn về dữ liệu.

Lớp tổng hợp - Pooling layer

Lớp tổng hợp (pooling layer) cũng gần giống như lớp tích chập, nhưng thay vì tham số hóa, ta sẽ định nghĩa sẵn cách thức tổng hợp các giá trị của nó. Mục đích chính của lớp tổng hợp là để giảm kích cỡ tensor đầu vào và tổng hợp thông tin, thường được sử dụng sau lớp tích chập. Có một số kỹ thuật pooling như: max pooling, average pooling và global max pooling là những dạng pooling đặc biệt, trong đó giá trị lớn nhất và giá trị trung bình cũng như giá trị toàn cục lớn nhất được lấy ra tương ứng (Hình 1.16).



Hình 1.16. Một số kiểu pooling

Lớp liên kết đầy đủ - FC layer

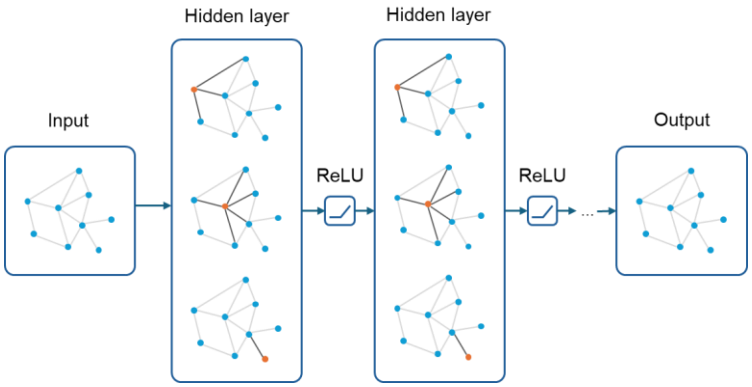
Lớp liên kết đầy đủ để nhận các đặc trưng của dữ liệu đầu vào đã được làm phẳng, mỗi đầu vào đó được kết nối đến tất cả các nơ-ron. Trong mô hình mạng CNN, các lớp FC này được sử dụng như lớp cuối của mô hình để tối ưu hóa mục tiêu mạng.

1.3.1.3. Mạng nơ-ron đồ thị

Cấu trúc dữ liệu đồ thị

Đồ thị là một loại cấu trúc dữ liệu mô hình hóa một tập hợp các đối tượng (các nút - nodes) và các mối quan hệ của chúng (các cạnh - edges). Gần đây, các nghiên cứu về phân tích dữ liệu đồ thị theo các phương pháp học sâu ngày càng nhận được nhiều sự quan tâm do có khả năng học các biểu diễn tốt của đồ thị (Hình 1.17). Dữ liệu dạng đồ thị có thể được áp dụng cho việc biểu diễn một lượng lớn các thông tin tính toán trong nhiều lĩnh vực khác nhau, bao gồm các lĩnh vực khoa học xã hội đặc

biệt được sử dụng biểu diễn dữ liệu phân tử hợp chất hóa học trong các bài toán khai phá thuốc.



Hình 1.17. Mô hình mạng nơ-ron đồ thị

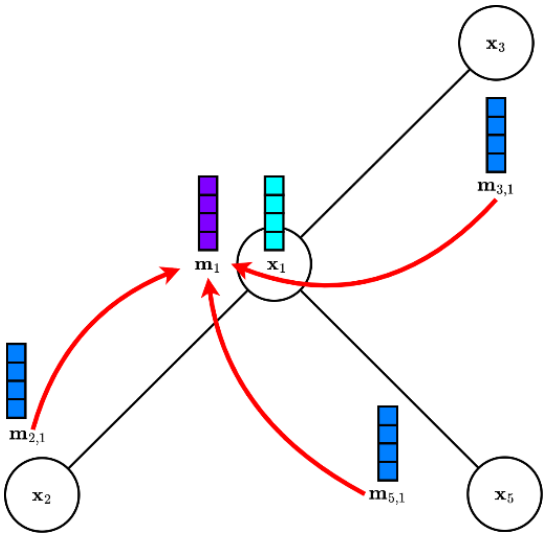
Một đồ thị $G = (V, E)$ được định nghĩa bởi tập các nút V và một tập các cạnh E giữa các nút đó. Một cạnh đi từ nút $u \epsilon V$ đến nút $v \epsilon V$ được ký hiệu là $(u, v) \epsilon E$.

Trong nhiều trường hợp người ta chỉ quan tâm tới các đồ thị đơn giản ở đó các cạnh giữa các cặp nút là vô hướng, ví dụ: $(u, v) \epsilon E \leftrightarrow (v, u) \epsilon E$.

Cách thức biểu diễn đồ thị là thông qua một ma trận kề (adjacency matrix) $A \epsilon R^{|V| \times |V|}$. Để biểu diễn ma trận kề, người ta sắp xếp các nút trong đồ thị theo thứ tự hàng và cột, các cạnh được biểu diễn như các thực thể trong ma trận đó: $A[u, v] = 1$ nếu $(v, u) \epsilon E$ và ngược lại thì $A[u, v] = 0$. Nếu đồ thị chỉ bao gồm các cạnh vô hướng, ta sẽ có một ma trận đối xứng (symmetric matrix), nếu là đồ thị có hướng có nghĩa là các cạnh có ý nghĩa quan trọng thì ma trận kề có thể không đối xứng. Một số dữ liệu đồ thị mà các cạnh có trọng số thì các phần tử trong ma trận này là một số thực không phải là dạng $[0/1]$. Ví dụ như đồ thị phân tử thuốc khi quan tâm đến đặc trưng cạnh, hoặc đồ thị tương tác PPI (protein-protein interaction) thì các cạnh có trọng số là giá trị liên kết giữa các nút tương tác.

Phương thức truyền thông điệp

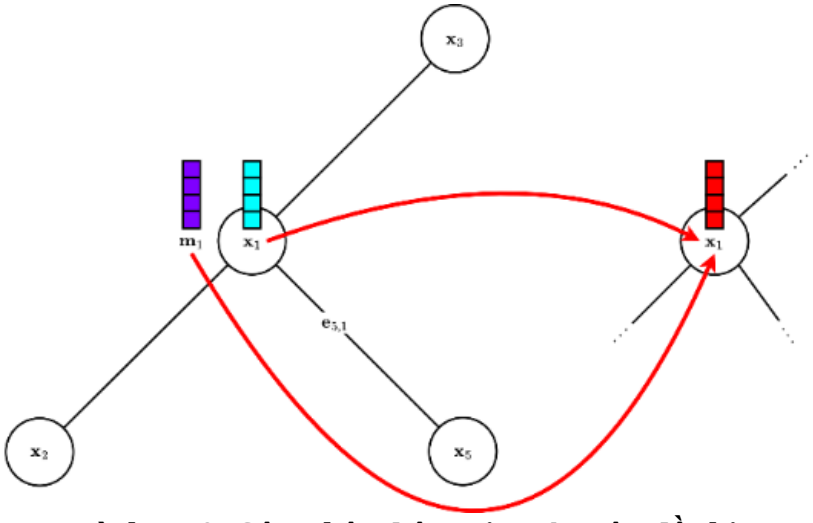
Tạo lập và kết tập thông điệp: Mạng nơ-ron đồ thị (GNN: Graph neural network) có thể học các thông điệp (message) của nút u và các nút trong vùng lân cận $N(u)$ của nó. Để tổng hợp thông điệp cho một nút u và các nút láng giềng v của nó, GNN tham gia vào phương thức kết tập và truyền thông điệp qua lớp tiếp theo của mạng nơ-ron (message passing).



Hình 1.18. Kết tập thông tin trên đồ thị

Đối với mỗi lớp GNN, truyền thông điệp (message passing) được định nghĩa là một tiến trình của việc thu thập/tạo lập các đặc trưng nút của các hàng xóm, tổng hợp, và truyền (passing) chúng tới nút nguồn (Hình 1.18). Tiến trình này được lặp đồng thời cho tất cả các nút trong đồ thị. Bằng cách này, tất cả các hàng xóm được tham gia vào tổng hợp thông tin cho nút.

Cập nhật đỉnh đồ thị: Trong GNN, mỗi nút tổng hợp các vec-tơ đặc trưng của các nút lân cận để tính toán vec-tơ đặc trưng mới của nó [48]. Sau k lần lặp lại của tổng hợp, một nút được biểu diễn bằng vec-tơ đặc trưng đã biến đổi của nó (m_i), vec-tơ này tổng hợp các thông tin cấu trúc trong vùng lân cận k-hop của nút. Sau đó, cập nhật thông tin biểu diễn nút trên của toàn bộ đồ thị. Ví dụ thông điệp từ các nút hàng xóm x_2, x_3, x_5 được truyền và tổng hợp tại nút x (Hình 1.19).



Hình 1.19. Cập nhật thông tin nút trên đồ thị

Quá trình trên được tổng kết bằng mô hình toán học như sau: Trong mỗi lần lặp truyền thông điệp trong GNN, một đặc trưng ẩn $h_u^{(k)}$ tương ứng của mỗi nút $u \epsilon V$ được cập nhật theo thông tin được kết tập từ các nút hàng xóm $v \epsilon N(u)$. Các thông điệp được cập nhật theo công thức sau:

$$h_u^{(l+1)} = UPDATE^l\left(h_u^{(l)}, AGGREGATE^{(l)}\left(\{h_v^{(l)}, \forall\, v \in N(u)\}\right)\right)$$

$$= UPDATE^l\left(h_u^{(l)}, m_{N(u)}^{(l)}\right) \quad (1.8)$$

Trong đó $UPDATE$ và $AGGREGATE$ là các hàm khả vi, $m_{N(u)}$là thông điệp (message) được kết tập từ các hàng xóm $N(u)$ của nút u.

Tại lớp thứ k của GNN, hàm $AGGREGATE$ tổng hợp các đầu vào của nút u và sinh ra thông điệp $m_{N(u)}^{(l)}$dựa trên các thông tin hàng xóm được kết tập của nó.

Hàm $UPDATE$ sau đó kết hợp thông điệp $m_{N(u)}^{(l)}$ với đặc trưng biểu diễn trước đó của nút u $h_u^{(l-1)}$ để sinh ra vec-tơ đặc trưng $h_u^{(l)}$.

Khởi tạo ban đầu $l = 0$, để thiết lập các đặc trưng đầu vào cho tất cả các nút (ví dụ: $h_u^{(o)} = x_u, \forall\, u \in V$).

Sau khi l lớp GNN truyền thông điệp, ta có đầu ra của lớp cuối cùng như là các embedding cho mỗi nút (ví dụ: $z_u = h_u^{(l)}, \forall\, u \in V$)).

Việc các mô hình mạng nơ-ron đồ thị thực hiện tính toán trên dữ liệu đồ thị được mô hình hoá như sự trao đổi thông tin hay thông điệp giữa các đỉnh trong đồ thị.

Dựa trên hàm kết tập và cập nhật khác nhau có các mô hình đồ thị khác nhau. Ví dụ như: các mô hình mạng nơ-ron đồ thị tích chập (GCN: Graph Convolutional Network) [49], mô hình mạng nơ-ron đồ thị cơ chế chú ý (GAT: Graph Attention Network) [50], mạng nơ-ron đồ thị đẳng cấu (GIN: Graph Isomorphism Network) [51]. Các mô hình mạng nơ-ron đồ thị tích chập này, mạng nơ-ron đồ thị sâu cấu tạo từ sự xếp chồng của nhiều lớp, gọi là các lớp tích chập đồ thị. Đồ thị đầu ra của lớp phía trước lại trở thành đồ thị đầu vào của lớp phía sau. Một lớp tích chập đồ thị thực hiện các phép tính

toán làm biến đổi các đặc trưng đỉnh và đầu ra là tập các đặc trưng đỉnh đã biến đổi này. Ký hiệu $h_u^{(l)}$, là vec-tơ đặc trưng của đỉnh u sau khi qua lớp tích chập đồ thị thứ l.

1.3.1.4. Mạng nơ-ron tích chập đồ thị

Mạng nơ-ron tích chập đồ thị (Graph convolutional network – GCN) là một biến thể của mạng nơ-ron đồ thị GNN, sử dụng cơ chế tích chập đồ thị để truyền thông tin qua các đỉnh và cạnh trong đồ thị từ đó tổng hợp thông tin từ hàng xóm của mỗi đỉnh. GCN thực hiện kết tập các thông tin đặc trưng của các đỉnh và cấu trúc đồ thị của chúng để có thể thực hiện tác vụ khác nhau như phân loại hay dự đoán trên đồ thị [49].

Mỗi lớp tích chập đồ thị của GCN xác định:

$$h_u^{(l+1)} = \sigma(W^l \sum_{v \in N_u \cup \{u\}} \frac{h_u^{(l)}}{\sqrt{|N_u||N_v|}}) \quad (1.9)$$

Trong đó, W^l là ma trận trọng số có thể học của lớp l, $\sigma(.)$ là một hàm kích hoạt ví dụ như $ReLU(.) = \max(0,.)$, $|N_u| = I + \sum_{v \in N(u)} e_{u,v}$, $e_{u,v}$ là trọng số cạnh đồ thị vô hướng.

E là tập các cạnh được biểu diễn bởi hai ma trận, ma trận thuộc tính các đỉnh (X), ma trận kề biểu diễn kết nối các đỉnh (A). $X \in \mathbb{R}^{NxF}$ thể hiện cho N nút, mỗi nút được biểu diễn bởi vec-tơ F chiều, $A \in \mathbb{R}^{N \times N}$ làm ma trận vuông biểu diễn kết nối có hay không giữa hai nút. Ban đầu, các lớp tích chập đồ thị sẽ cập nhật theo công thức sau:

$$X_{new} = AXW \quad (1.10)$$

Trong đó, X là ma trận thuộc tính các đỉnh, A là ma trận kề biểu diễn kết nối giữa các đỉnh $W \in \mathbb{R}^{N \times C}$, với C là số chiều đầu ra ta muốn sau hàm biến đổi không gian. Tuy nhiên, mô hình này có hai nhược điểm. Thứ nhất, một nút được cập nhật không dựa trên thuộc tính của chính nó. Thứ hai, sẽ có sự bất lợi cho những nút có quá ít hay quá nhiều các kết nối. Vì vậy theo [49], đã có cải tiến công thức cho GCN, việc cập nhật sẽ được định nghĩa:

$$\widetilde{D}^{\frac{1}{2}}\tilde{A}\widetilde{D}^{\frac{1}{2}}XW \tag{1.11}$$

Trong đó, $\widetilde{D}$ là là ma trận bậc đã được chuẩn hóa, $\tilde{A}$ là ma trận kề đã cho thêm kết nối với chính nó. Tiếp theo, để có một phép biến đổi phi tuyến, áp dụng hàm ReLU để biến đổi các giá trị sau lớp tích chập đồ thị.

1.3.1.5. Mạng nơ-ron đồ thị cơ chế chú ý

Cơ chế chú ý (attention) được sử dụng rộng rãi trong nhiều bài toán học sâu, khi không thể tự định nghĩa các trọng số kết nối giữa hai nút thì dữ liệu sẽ định nghĩa điều đó. Trong [50] GAT là sự kết hợp của một mạng nơ-ron đồ thị và một lớp chú ý. Việc triển khai lớp chú ý trong mạng nơ-ron đồ thị giúp tăng cường cơ chế chú ý, tập trung vào các thông tin quan trọng từ dữ liệu thay vì tập trung vào toàn bộ dữ liệu.

Cơ chế chú ý thực hiện tính toán trọng số cho các đỉnh lân cận của một đỉnh cụ thể: lớp GAT lấy ra một nút, tiếp theo dùng phép biến đổi tuyến tính với trọng số W, sau đó sẽ tính toán mối liên kết giữa các cặp nút (i, j) thông qua công thức:

$$\alpha\left(Wh_i, Wh_j\right) \tag{1.12}$$

Trong đó α là một cơ chế attention được định nghĩa sẵn, nó có thể là phép nhân vec-tơ, hay một phép biến đổi nào đó để đầu ra thuộc không gian R.

Các nút sẽ được cập nhật theo công thức sau:

$$h_i^{(l+1)} = \alpha_{i,j}\Theta h_i^{(l)} + \sum_{j\in N(i)} \Theta h_j^{(l)}\alpha_{i,j} \tag{1.13}$$

Trong đó $h_i^{(l+1)}$ là giá trị sau cập nhật của h_i, N(i) là những nút lân cận của h_i, Θ *tương tự như* W là trong số để biến đổi không gian, $\alpha_{i,j}$ là hệ số attention được định nghĩa:

$$\alpha_{i,j} = \frac{\exp\left(LeakyReLU\left(a^T[\Theta h_i^{(l)} || \Theta h_j^{(l)}]\right)\right)}{\sum_{k\in N(i)\cup\{i\}} \exp\left(LeakyReLU\left(a^T[\Theta h_i^{(l)} || \Theta h_j^{(l)}]\right)\right)} \tag{1.14}$$

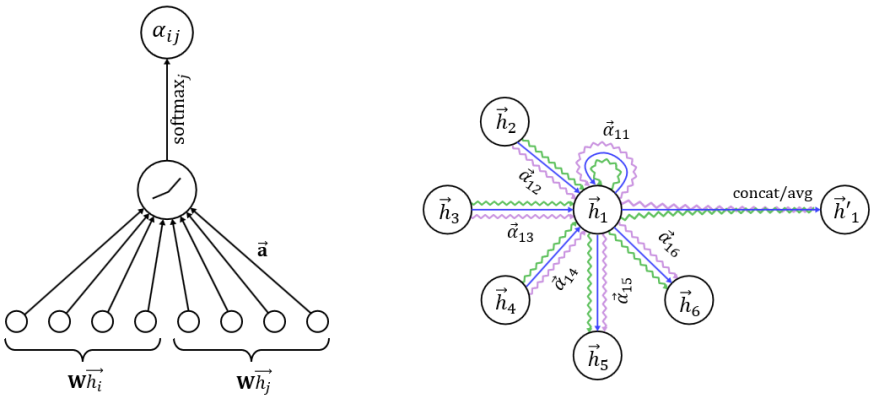
Trong đó, $\Theta \in \mathbb{R}^{F^{(l+1)}\times F^{(l)}}$ và $a \in \mathbb{R}^{2F^{(l+1)}}$ là vec-tơ trọng số học được tại lớp thứ l, $LeakyReLU$ là hàm kích hoạt, $\alpha_{i,j}$ là hệ số chú ý xác định mức độ quan trọng

của nút i so với nút j như thế nào. Các hệ số chú ý được chuẩn hóa như công thức trên để giúp chúng dễ dàng so sánh giữa các nút khác nhau.

Với một cơ chế multi-head GAT layer (Hình 1.20), lớp GAT cuối cùng, đầu ra từ mỗi đầu chú ý (attention head) được tính trung bình trước khi áp dụng hàm kích hoạt (Hình 1.19). Về mặt hình thức, lớp GAT cuối cùng có thể được viết là có thể được biểu diễn như sau:

$$h_i^{(l+1)} = \sigma\left(\frac{1}{K}\sum_{1}^{K}\sum_{j\in N(i)} \Theta h_j^{(l)} h_j\right) \tag{1.15}$$

Trong đó K là số attention heads, σ là hàm kích hoạt, $\Theta h_j^{(l)}$ là ma trận tham số có thể huấn luyện được cho attention head thứ l.

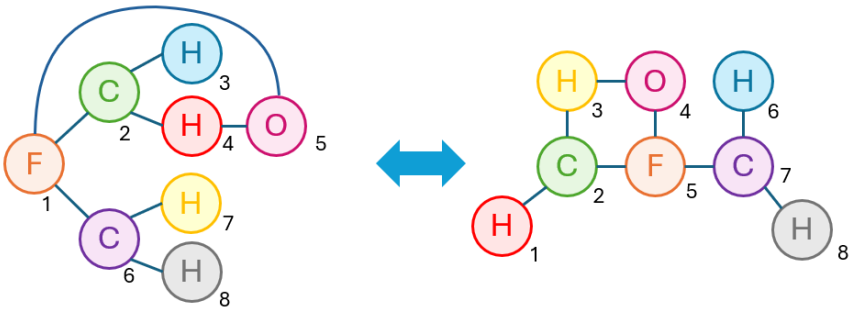


Hình 1.20. Cơ chế attention và multi-head attention [50]

Mạng nơ-ron tích chập đồ thị (GCN) có thể coi là trường hợp đặc biệt của mạng nơ-ron chú ý đồ thị (GAT) khi các hệ số chú ý không có khả năng học nhưng được gán bằng với các trọng số của cạnh $w_{i,j}$.

1.3.1.6. Mạng nơ-ron đồ thị đẳng cấu

Hai đồ thị là đẳng cấu nếu tồn tại một hoán vị chỉ số giữa các đỉnh và bảo toàn các đỉnh kề nhau, thể hiện mối quan hệ tương quan về cấu trúc đồ thị mà không quan tâm đến các đặc trưng nút (Hình 1.21).



Đồ thị phân tử 1 Đồ thị phân tử 2

Hình 1.21. Đồ thị đẳng cấu

Mạng nơ-ron đồ thị đẳng cấu (GIN) có thể học các các biểu diễn đồ thị khác nhau sử dụng mô thức truyền thông điệp có hiệu quả mạnh mẽ [51]. GIN cho phép phân biệt các đồ thị không đẳng cấu với nhau, hay có thể phân biệt cấu trúc đồ thị khác nhau. Sau khi mô hình đã được huấn luyện, GIN có thể được sử dụng để xác định tính đẳng cấu hay tương đồng giữa các đồ thị. Tính tương đồng có thể được đo lường bằng cách so sánh đặc trưng của các đỉnh tương ứng trong các đồ thị.

Sự khác biệt chính giữa các mô hình mạng nơ-ron đồ thị là lược đồ kết tập trong phương thức truyền thông điệp. Trong mạng nơ-ron đồ thị đẳng cấu (GIN), mỗi đỉnh đồ thị được biểu diễn bằng một vec-tơ đặc trưng, các đỉnh sẽ được cập nhật theo hàm:

$$h_i' = MLP^{(l)}\left((1+\epsilon)\cdot h_i^{(l)} + \sum_{j\in N(i)} h_j^{(l)}\right) \quad (1.16)$$

Trong đó giá ϵ là một giá trị được định nghĩa sẵn, trong luận án sử dụng giá trị mặc định bằng 0, N(i) là các lân cận của nút i, $h_i^{(l)}$ biểu diễn đặc trưng của đỉnh i sau l bước tổng hợp, $MLP^{(l)}$ là mạng nơ-ron đa tầng được sử dụng để tổng hợp và định nghĩa chiều không gian đầu ra của các nút.

GIN sử dụng một cơ chế cập nhật đồ thị để tính toán vec-tơ đặc trưng mới cho mỗi đỉnh dựa trên đặc trưng của đỉnh và các đỉnh lân cận. MLP là một mạng nơ-ron đa tầng (Multilayer Perceptron) được áp dụng cho mỗi đỉnh.

1.3.2. Các phương pháp dự đoán đáp ứng thuốc hiện nay

Hiện nay, một số mô hình nghiên cứu phổ biến thường liên quan đến bài toán dự đoán đáp ứng thuốc cho từng thuốc đơn (monotherapy/single drug) và dự đoán cho đáp ứng đa thuốc hay kết hợp thuốc (combination therapy/ drug pair). Hướng tiếp cận dự đoán đáp ứng thuốc cho thuốc đơn thường dự đoán giá trị đáp ứng thuốc IC_{50} hoặc phân loại đáp ứng và kháng thuốc trong khi hướng dự đoán kết hợp thuốc sẽ dự đoán giá trị kết hợp của các thuốc (synergy scores) hoặc phân loại khả năng kết hợp hoặc không kết hợp của các thuốc.

Các mô hình tính toán dự đoán áp dụng cho cả hai hướng tiếp cận này đều có thể triển khai từ các mô hình học máy truyền thống (ML) đến các mô hình học sâu (DL) và mô hình học dựa trên mạng (network-based learning). Đối với bài toán dự đoán đáp ứng thuốc hiện nay thì hầu hết các mô hình tính toán đều dựa trên mô hình học có giám sát. Các bước chính để xây dựng các mô hình dự đoán có thể tóm tắt theo các bước dưới đây (Hình 1.22):

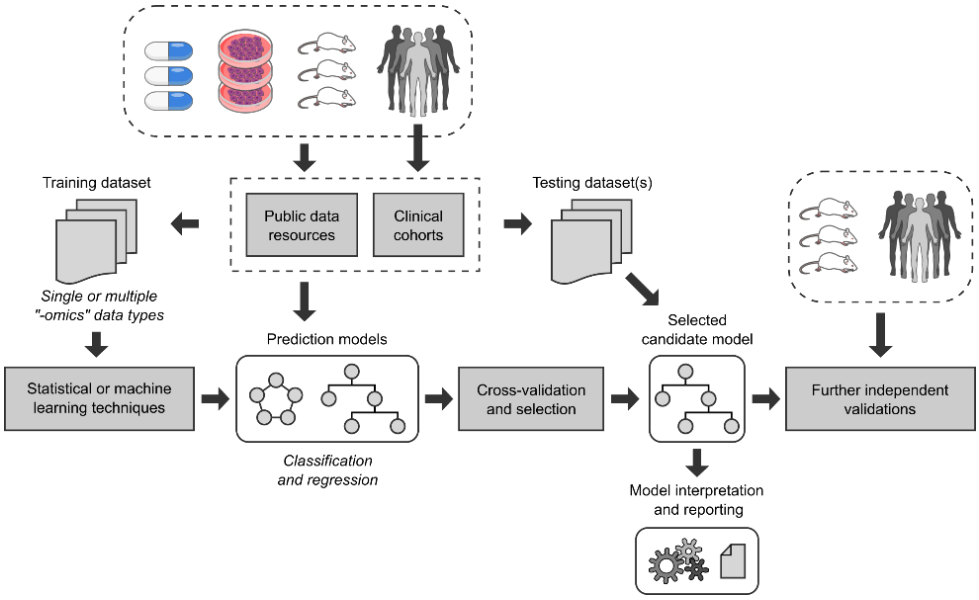
- Dữ liệu thu thập được từ các nguồn nghiên cứu dựa trên các dòng tế bào, động vật hoặc con người được lưu trữ trong các kho dữ liệu khác nhau, bao gồm cả dữ liệu công khai và dữ liệu bản quyền. Các tập dữ liệu này bao gồm thông tin về đáp ứng thuốc tương ứng.

- Tập dữ liệu thu được sau đó được sử dụng làm tập dữ liệu huấn luyện mô hình dự đoán, nó có thể chứa một hoặc nhiều loại dữ liệu '-omics' khác nhau như: biểu hiện gen, đột biến gen.

- Dữ liệu này được sử dụng làm đầu vào cho các mô hình thống kê hoặc học máy cũng như các phương pháp tính toán dự đoán khác nhau.

- Hiệu năng dự đoán của các mô hình thường được đánh giá bằng các kỹ thuật lấy mẫu đánh giá chéo (cross-validation) hoặc (LOOV). Các mô hình có triển vọng nhất được lựa chọn và đánh giá bằng cách sử dụng các tập dữ liệu thử nghiệm, không được sử dụng trong giai đoạn huấn luyện.

- Mô hình và các dự đoán sau đó có thể được diễn giải xác nhận độc lập bằng các dữ liệu liên quan đến lâm sàng để tiếp tục thu hẹp khoảng cách giữa phòng thí nghiệm và phòng khám.



Hình 1.22. Mô hình tính toán dự đoán đáp ứng thuốc

1.3.2.1. Phương pháp dự đoán đáp ứng thuốc cho đơn thuốc

Các mô hình dự đoán đáp ứng thuốc đơn thuốc (DRP) hiện nay chủ yếu dựa trên mô hình học có giám sát mà phần lớn các phương pháp này được thực hiện theo phương pháp hồi quy tuyến tính (dữ liệu đáp ứng thuốc được coi là liên tục) và phân loại (dữ liệu đáp ứng thuốc được coi là rời rạc). Các mô hình hồi quy thực hiện việc tính toán/ước lượng giá trị độ nhạy thuốc thường dựa trên các phép đo AUC/IC_{50} [52]. Trong khi đó, các mô hình phân lớp thường đưa ra dự đoán về độ nhạy theo các mức đáp ứng được xác định trước, ví dụ xác định độ nhạy cao/thấp hoặc xác định độ nhạy và độ kháng thuốc [53].

Một loạt các kỹ thuật dựa trên các phương pháp học máy truyền thống đã được đề xuất [9], [10], [54], [55]. Tuy nhiên không có giải pháp chung và không có phương pháp thực sự vượt trội hơn so với phương pháp khác, hiệu năng phụ thuộc vào tập dữ liệu huấn luyện (ví dụ CCLE) và đo độ nhạy của thuốc được sử dụng (ví dụ IC_{50} hoặc AUC). Các mô hình học có giám sát tích hợp các kỹ thuật khác nhau như học đa tác vụ (multitask learning), thúc đẩy khả năng dự đoán đáp ứng thuốc tốt [56]. Tuy nhiên, độ chính xác của các mô hình học máy như vậy phụ thuộc rất nhiều vào chất lượng của dữ liệu huấn luyện từ bộ dữ liệu mẫu và độ lớn của kích thước mẫu. Các hướng nghiên cứu dựa trên mạng (network based approaches) cho kết quả khả quan khi xem xét các đặc tính -omics được biểu diễn trong các mạng gen/protein hoặc trong các mạng tương đồng giữa các dòng tế bào [53], [57]. Các hướng này có

khả năng thu thập được thông tin về các mối quan hệ giữa các đối tượng trong mạng sinh học, tuy nhiên khó có thể dự đoán cho các thuốc hoặc bệnh mới.

Hiện nay, công nghệ giải trình tự thông lượng cao, nguồn dữ liệu ngày càng lớn thì việc kết hợp các dữ liệu khác như hình ảnh y sinh, hồ sơ miễn dịch tạo ra lượng lớn dữ liệu. Khi không gian đặc tính đủ lớn các mô hình học máy sâu có thể học các hàm phi tuyến phức tạp. Do đó có thể trích xuất các đặc trưng ẩn và tăng độ chính xác của mô hình dự đoán [58]. Trong bài toán dự đoán đáp ứng thuốc, các mô hình học sâu có khả năng học các biểu diễn của thuốc, các dữ liệu -omics một cách đầy đủ các thông tin đầu vào mà không cần trích chọn đặc trưng trước khi huấn luyện [21], [22], [40], [59], [60].

Các nghiên cứu đầu tiên áp dụng mô hình học sâu cho DRP đều là các hướng nghiên cứu không sử dụng dữ liệu biểu diễn thuốc hoặc sử dụng dạng biểu diễn chuỗi hoặc ảnh mà chưa biểu diễn phân tử thuốc dưới dạng đồ thị. Các phương pháp này thường áp dụng mô hình CNN hoặc MLP để trích xuất các đặc trưng phân tử thuốc và dòng tế bào, như tCNNs [21], CDRscan [40], DeepDSC [61]. Trong đó tCNNs [21] là phương pháp tiên tiến đã sử dụng hai lớp mạng nơ-ron đồ thị tích chập 1chiều để học các biểu diễn của dòng tế bào thuốc và thuốc với đầu vào là chuỗi SMILES dưới dạng ký tự. Phương pháp tCNNs xây dựng tập từ điển cho dữ liệu chuỗi ký tự SMILES của thuốc, mã hoá thuốc dưới dạng chuỗi one-hot, sau đó sử dụng mạng tích chập 1-chiều để trích xuất đặc trưng thuốc. Tuy nhiên cũng giống các phương pháp trên, trong tCNNs thuốc thường được biểu diễn dưới dạng chuỗi các nguyên tử hoá học, chưa biểu diễn được các liên kết hóa học giữa các nguyên tử với nhau nên chưa thể hiện được tính tự nhiên của biểu hiện cấu trúc phân tử hóa học (dạng graph), do đó có thể mất đi cấu trúc thông tin của thuốc. Đây là các phương pháp không sử dụng dữ liệu biểu diễn thuốc dưới dạng đồ thị, luận án coi là các phương pháp “no-graph”. Hai năm gần đây, một số phương pháp “no-graph” đã được đề xuất [25], [26]. Các phương pháp này sử dụng thông tin tương tác thuốc – cell line và xây dựng các mạng tương đồng của thuốc, mạng tương đồng của cell line từ đó áp dụng mạng nơ-ron đồ thị tích chập để để học các biểu diễn của cell line và thuốc cho dự đoán đáp ứng thuốc. Các nghiên cứu này tuy có thể tận dụng được các thông tin tương tác thuốc – cell line trong quá trình xây dựng mạng, nhưng không sử dụng dữ liệu thuốc dưới dạng đồ thị

phân tử thuốc, do đó có thể chưa học hết các đặc trưng của thuốc; đồng thời gặp thách thức khi dự đoán đáp ứng cho các dòng tế bào hay các thuốc mới. Một số phương pháp học sâu khác áp dụng biểu diễn dữ liệu thuốc dạng "graph" đã được đề xuất mang lại hiệu quả dự đoán tốt hơn như kết quả hai đề xuất nghiên cứu trong luận án hoặc một số nghiên cứu cải tiến mô hình học transformer như [62], [60], [63] cho thấy hướng nghiên cứu tiềm năng trong dự đoán đáp ứng thuốc.

1.3.2.2. Phương pháp dự đoán đáp ứng thuốc cho đa thuốc

Mặc dù thuốc 'nhắm mục tiêu' (targeted drugs) đã đạt được những tiến bộ trong điều trị bệnh nhân ung thư, nhưng lợi ích lâm sàng của chúng bị hạn chế rất nhiều do khả năng kháng thuốc tự nhiên mắc phải của tế bào ung thư [64]. Cơ chế nội sinh (endogenous mechanism) của kháng thuốc nằm trong sự dẫn truyền tín hiệu bù trừ và sự trao đổi chéo (crosstalk among pathways) giữa các con đường sinh học phát sinh từ quá trình tiến hóa lâu dài [65]. Phương pháp điều trị bằng thuốc 'một mục tiêu' thường dẫn đến việc kích hoạt đường truyền tín hiệu bù trừ duy trì sự phát triển và tồn tại của các tế bào khối u [66]. Ngược lại, kết hợp thuốc đã cho thấy những lợi thế to lớn trong việc khắc phục tình trạng kháng thuốc và nâng cao hiệu quả điều trị trong điều trị ung thư và do đó ngày càng thu hút sự chú ý của các nhà nghiên cứu và doanh nghiệp dược phẩm [67]. Xu hướng chuyển từ mô hình mục tiêu đơn lẻ sang đa mục tiêu và kết hợp trong khám phá thuốc nhưng hầu hết chúng được phát hiện bằng kinh nghiệm lâm sàng hoặc tình cờ phát hiện được.

Một số phương pháp tiếp cận dựa trên máy học để dự đoán kết hợp thuốc đã được đề xuất, bao gồm các mô hình truyền thống như hồi quy tuyến tính, máy vec-tơ hỗ trợ (SVM) [13], [14], mô hình mạng nơ-ron [68], đến các phương pháp học máy bao gồm các phương pháp rừng ngẫu nhiên và Naïve Bayes [15], [16]. Một số cách tiếp cận dựa trên mạng (network-based approaches) [69], [70], [71], đã tạo ra các mạng tương đồng, giữa các dòng tế bào và giữa các loại thuốc một cách độc lập, với mỗi dòng tế bào, giá trị đáp ứng dự đoán được xác định dựa trên đáp ứng đã biết và các nút hàng xóm của cả mạng drug và dòng tế bào, sau đó các giá trị dự đoán được đưa ra bởi mô hình trọng số.

Gần đây, học sâu cũng ngày càng trở thành một giải pháp phổ biến trong các nghiên cứu khám phá thuốc. Mô hình học sâu gần đây cũng được áp dụng triển khai

cho dự đoán đáp ứng đa thuốc cho thấy hiệu năng dự đoán tốt hơn nhiều so với các phương pháp học máy truyền thống [60], [61], [62], [63]. DeepSynergy [60] sử dụng thông tin hóa học của thuốc và đặc điểm gen của bệnh để dự đoán các cặp thuốc có tác dụng kết hợp thuốc. Tuy nhiên trong phương pháp này, dữ liệu thuốc mới biểu diễn dữ liệu fingerprint, chưa biểu diễn dạng đồ thị và chưa tích hợp dữ liệu trong dự đoán. Dựa trên thành công của một số nghiên cứu áp dụng "graph" trong dự đoán đáp ứng đơn thuốc, một vài các đề xuất dự đoán đáp ứng đa thuốc[72], [73] đã áp dụng graph trong việc học các dữ liệu đồ thị phân tử thuốc cho thấy hiệu quả tiềm năng của dự đoán.

Công nghệ sàng lọc thông lượng cao (HTS) hiện nay, có thể thử nghiệm đồng thời đáp ứng đa thuốc đối với hàng trăm dòng tế bào ung thư [55], [74]. Số lượng các cặp thuốc được thử nghiệm còn khá hạn chế, tuy nhiên cũng đã tăng lên đáng kể trong những năm gần đây. O'Neil và các cộng sự [75] đã công bố một nghiên cứu về kết hợp thuốc thông lượng cao, bao gồm hơn 20.000 mẫu về kết hợp thuốc. Công ty dược phẩm nổi tiếng AstraZeneca [11] cũng công bố thực nghiệm phối hợp cặp thuốc mới nhất, bao gồm 11.576 thực nghiệm của 910 cách kết hợp thuốc với 85 dòng tế bào ung thư với các đặc điểm bộ gen. DrugCombDB [76] có hơn 6.000.000 đáp ứng liều lượng thuốc định lượng, từ đó họ tính toán nhiều chỉ số kết hợp để xác định kết hợp hoặc không kết hợp thuốc. Những cơ sở dữ liệu cao này đã tạo điều kiện thuận lợi cho việc phát triển các phương pháp tính toán để dự đoán sự kết hợp thuốc hiệp đồng.

1.3.2.3. Phương pháp tích hợp dữ liệu

Dữ liệu y sinh thường có số chiều lớn nhưng thưa thớt. Điều này trái ngược với các bộ dữ liệu lớn trong các lĩnh vực khác, chẳng hạn như mạng xã hội, thị giác máy tính và ngôn ngữ tự nhiên, thường chứa một số lượng lớn các mẫu chất lượng cao. Một nghiên cứu liên kết toàn bộ bộ gen điển hình (GWAS) [77] nghiên cứu kiểu gen xác định hàng trăm nghìn đa hình đơn nucleotide cho mỗi cá nhân. Tuy nhiên, những dữ liệu này thường chỉ có thể được thu thập cho một số lượng tương đối nhỏ các cá thể với một kiểu hình cụ thể. Hơn nữa, tính chất thưa thớt của những dữ liệu này, tức là mỗi tính đa hình chỉ xuất hiện ở một số lượng nhỏ của tất cả các cá thể, tạo ra một thách thức bổ sung cho các ứng dụng phân tích dữ liệu tiếp sau của quá trình sinh học. Nếu không tích hợp các loại dữ liệu khác nhau, chẳng hạn như thông tin mạng

phân tử hoặc pathway [78], [79], dữ liệu GWAS khó có thể xác định các mẫu có ý nghĩa liên quan đến kiểu hình cần quan tâm.

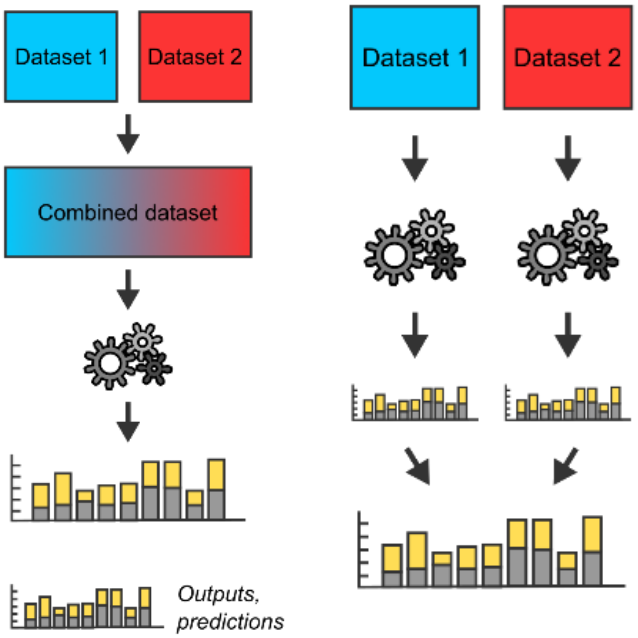
Một vấn đề khác quan trọng phát sinh từ bản chất đó là sự không đầy đủ và thiên lệch (incomplete and biased) một cách tự nhiên của dữ liệu y sinh [80]. Ví dụ, thông tin về những hợp chất hóa học nào tương tác với những gen nào là rất hạn chế chỉ có vài nghìn gen được nghiên cứu, trong toàn bộ hệ gen. Hơn nữa, số lượng các thuốc ức chế mỗi gen rất không đồng đều, số lượng được thử nghiệm cũng rất hạn chế [81], với nhiều gen không được xác định đóng vai trò quan trọng trong hoạt động của thuốc [82].

Ngoài ra, dữ liệu y sinh được tổ chức theo mức độ khác nhau và trải dài theo mức độ phân tử, tế bào, mô, cơ quan, bệnh nhân và quần thể [83] và cũng bao gồm một phổ rộng về thang thời gian và loài phát triển. Rõ ràng, hiểu biết đầy đủ về sinh học đòi hỏi mô hình đa cấp độ, từ việc mô tả các chi tiết nguyên tử của phân tử đến các đặc tính nổi bật của quần thể sinh vật. Hơn nữa, khi kết quả y sinh thay đổi theo thời gian, các phương pháp học máy tích hợp kết quả cần phải tính đến những tác động này. Ví dụ đối với tế bào ung thư, sau quá trình điều trị với một đơn thuốc và tác động của môi trường xung quanh như vi khuẩn và vi rút phát triển nhanh chóng tạo ra sự kháng thuốc, có thể dẫn đến hiệu năng kém trong việc dự đoán hiệu quả và độc tính của thuốc [84]. Do đó cần phải tiến hành khảo sát phân tích kết hợp thêm dữ liệu mới để dự đoán điều chỉnh lại thuốc hoặc kết hợp các thuốc với nhau để tạo hiệu quả trong quá trình điều trị.

Với sự ra đời của công nghệ giải trình tự, nguồn dữ liệu “multi-omics” [85], cung cấp thông tin về các phân tử sinh học từ các lớp khác nhau có thể hiểu được cấu trúc sinh học phức tạp một cách có hệ thống và toàn diện [86], từ đó thúc đẩy nghiên cứu các phương pháp tích hợp mức hệ thống với các dữ liệu -omics này cho các bài toán y học và sinh học hiện nay.

Mỗi mô hình dự đoán có những chiến lược lựa chọn dữ liệu đầu vào khác nhau. Một số mô hình tích hợp các bộ dữ liệu khác nhau, ví dụ: chỉ dùng dữ liệu GE hoặc kết hợp GE với dữ liệu CNV. Một số phân tích so sánh cho thấy dữ liệu GE là quan trọng nhất và các mô hình tích hợp dữ liệu này có thể làm tăng hiệu năng dự đoán

đáp ứng thuốc [12], [13]. Tuy nhiên, trong nghiên cứu dự đoán đáp ứng thuốc cho bệnh đa u tủy (multiple myeloma), Amin et al. [89] cho rằng chỉ riêng dữ liệu về biểu hiện gen là không đủ để dự đoán đáp ứng cho một số thuốc. Hiện nay có các hướng chính để tích hợp dữ liệu gồm: tích hợp sớm (early integration), tích hợp muộn (late integration) (Hình 1.23).



Hình 1.23. Các hướng tiếp cận tích hợp dữ liệu

1.3.2.3.1. Mô hình tích hợp sớm

Đây là phương pháp đơn giản, kết hợp tập các dữ liệu từ các nguồn khác nhau ở mức độ dữ liệu thô hoặc tiền xử lý trước khi đưa vào xử lý và dự đoán [69]. Nghiên cứu [15], [90] sử dụng các phương pháp multi-kenel learning biểu diễn liên kết của nhiều tập dữ liệu. Về mặt lý thuyết, nó có thể tổng hợp tốt các đặc trưng dữ liệu bởi vì mô hình có thể xác định các mối quan hệ giữa các đặc trưng miễn là các tập dữ liệu riêng lẻ không được thu gọn trước khi tạo mô hình. Hướng tích hợp sớm này trước hết kết hợp các ma trận dữ liệu -omics sau đó áp dụng các thuật toán phân cụm hiện tại cho các single-omics trên ma trận kết hợp. Tuy nhiên hướng này có một số nhược điểm:

- Trước hết, nếu dữ liệu không được chuẩn hóa (normalization) thì chúng có thể tạo ra nhiều trọng số cho -omics với nhiều đặc trưng hơn.

- Thứ hai là không xem xét đến các dữ liệu phân bố khác nhau trong các dữ liệu -omics khác nhau.

- Cuối cùng là nó làm tăng chiều dữ liệu (số đặc tính) mà đây là một thách thức ngay cả trong một số tập dữ liệu đơn (single-omics).

1.3.2.3.2. Mô hình tích hợp muộn

Trong tích hợp muộn, các đặc trưng được học riêng cho từng loại dữ liệu -omics và sau đó các đặc trưng này được tích hợp vào một biểu diễn thống nhất để sử dụng làm đầu vào cho bộ phân loại hoặc bộ hồi quy. Đây là mô hình tích hợp trong có hai mức, mức mô hình đầu tiên được xây dựng cho việc học các đặc trưng từ các bộ dữ liệu độc lập. Mức mô hình thứ hai là kết hợp các đặc trưng đã được học biểu diễn trước đó đưa vào mô hình dự đoán [20],

Ưu điểm của phương pháp này là mô hình hoạt động với một phân phối duy nhất của mỗi dữ liệu omics. Phương pháp này có thể sử dụng chuẩn hóa đơn dữ liệu -omics cho từng loại dữ liệu và nó không làm tăng kích thước của không gian đầu vào.

1.3.3. Phương pháp đánh giá hiệu năng dự đoán

Khi đánh giá hiệu quả dự đoán của mô hình đáp ứng thuốc, các phương pháp đánh giá thường được đề xuất theo một chiến lược phù hợp để đảm bảo rằng mô hình có tính tổng quát hóa với cả các trường hợp dự đoán cho thuốc mới và dòng tế bào mới. Dự đoán đáp ứng thuốc cho thuốc mới (Blind-Drug hay leave-drug-out) là cách phân chia bộ dữ liệu thử nghiệm sao cho một hoặc nhiều thuốc trong tập kiểm tra không nằm trong tập huấn luyện. Kết quả của mô hình là xác thực chính xác hơn cách thức mô hình sẽ hoạt động khi đưa một thuốc mới vào quá trình dự đoán. Tương tự như vậy đối với dự đoán đáp ứng thuốc cho dòng tế bào mới (Blind-Cellline hay leave-Cellline-out). Nhiều mô hình đã sử dụng một số hình thức đánh giá xác thực chéo (k-fold cross-validation) để đánh giá hiệu năng và tính khái quát của mô hình. Tuy nhiên không phải tất cả các mô hình đều sử dụng hình thức đánh giá này. Khi dữ liệu đủ lớn việc phân chia bộ dữ liệu thử nghiệm có thể phân chia một tỷ lệ nhất định như tCNNs (80:10:10) [21].

Việc lựa chọn chỉ số đánh giá (evaluation metrics) phù hợp để đánh giá và so sánh các mô hình dự đoán đáp ứng thuốc khác nhau cũng rất quan trọng và tùy thuộc

vào từng loại mô hình dự đoán. Trong mô hình hồi quy, một số chỉ số thường được sử dụng như sai số bình phương trung bình/sai số bình phương trung bình gốc (MSE/RMSE), độ tương quan Pearson (CCp). Trong khi các mô hình phân loại thường dùng các chỉ số như Accuracy, Precision, Recall, F1-score, ROC, AUC, …để đánh giá hiệu năng dự đoán. Các chỉ số đánh giá có thể cho thấy nhiều ý nghĩa khác nhau, trong bất kỳ trường hợp nào, việc sử dụng nhiều chỉ số đánh giá khác nhau có thể cung cấp những thông tin hữu ích về mô hình dự đoán. Một số chỉ số được tính toán cụ thể như sau:

RMSE

RMSE **(**Root Mean Squared Error) là độ đo phổ biến cho các bài toán hồi quy. Chỉ số RMSE cho biết lỗi trung bình bình phương của các giá trị dự đoán và giá trị quan sát được. giá trị RMSE càng thấp càng thể hiện tính hiệu quả của mô hình, có thể dùng làm hàm mất mát (loss function) trong mô hình dự đoán.

$$RMSE = \sqrt{\frac{1}{n}\sum_{1}^{n}(o_i - y_i)^2} \tag{1.17}$$

Trong đó o_i và y_i là các giá trị (đáp ứng thuốc) quan sát được và dự đoán được của mẫu thứ i, n là số mẫu cần đo.

Hệ số tương quan Pearson

Hệ số tương quan Pearson (Pearson correlation coefficient) đánh giá độ tương quan giữa giá trị quan sát được và giá trị dự đoán được.

$$CCp = \frac{\sum_{1}^{n}(o_i - y_i)^2}{\sigma_O \sigma_Y} \tag{1.18}$$

Độ chính xác (Accuracy)

Độ chính xác là đơn vị đo cơ bản nhất, xác định tỷ lệ phân loại đúng của mô hình, Acc càng cao cho thấy mô hình nhìn chung phân loại đúng các mẫu.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}.100\% \tag{1.19}$$

Trong đó:

- TP (True Positive - nhãn đúng và được phân loại là đúng)
- FP (False Positive - nhãn sai nhưng được phân loại là đúng)
- FN (False Negative - nhãn đúng nhưng đươc phân loại là sai)
- TN (True Negative - nhãn sai và được phân loại là sai)

Hầu hết các phép đo trong bài toán phân loại đều tính toán dựa trên 4 thông số này.

Độ chuẩn (Precision – PRE)

Độ chuẩn giúp xác định độ chuẩn xác của phân loại đúng (positive) thực sự trên tổng số dự đoán là đúng.

$$precision = \frac{TP}{TP + FP} \tag{1.20}$$

Độ hồi nhớ (Recall)

Độ hồi nhớ giúp đo khả năng hồi nhớ của mô hình khi gặp đối tượng nhãn đúng (positive) hay tỉ lệ dự đoán positive mà mô hình dự đoán được trong số các mẫu positive trên thực tế

$$recall = \frac{TP}{TP + FN}.100\% \tag{1.21}$$

F1-score

F1-score là tổng hợp của Precision và Recall. Với bài toán phân loại đa nhãn, mỗi nhãn sẽ cho ra 1 bộ (Precision, Recall), tương ứng với một giá trị F1 tương ứng.

$$\begin{aligned} F1 &= \frac{1}{\frac{1}{precision} + \frac{1}{recall}} \\ F1 &= \frac{2.precision.recall}{precision + recall} \\ F1 &= \frac{TP}{TP + \frac{1}{2}(FP + FN)} \end{aligned} \tag{1.22}$$

Trong các mô hình dự đoán, FP và FN là hai lỗi cần đồng thời giảm thiểu khi xây dựng mô hình do đó, hai thang đo Recall và Precision được tính toán đồng thời và có thể được thay thế bằng cách tính giá trị F1-score.

1.3.4. Một số phân tích và định hướng nghiên cứu

Dự đoán đáp ứng thuốc trong điều trị bệnh là một vấn đề tương đối mới tại Việt Nam hiện nay. Dựa trên nguồn dữ liệu công khai, một số ít nghiên cứu trong nước được đề xuất như [12], [91], phần lớn các đề xuất là các tác giả nước ngoài với đa dạng các phương pháp tính toán khác nhau và đã đạt hiệu năng tiềm năng cho dự đoán đáp ứng thuốc trong điều trị bệnh. Tuy nhiên còn một số thách thức trong các nghiên cứu được khảo sát hiện nay như sau:

- Các mô hình học máy truyền thống phù hợp với các bài toán có bộ dữ liệu nhỏ, và trích chọn các. Công trình nghiên cứu 3 (trình bày trong hội nghị KSE2020) của luận án, đã cho thấy điều đó. Thực tế, đề xuất này đã xây dựng một mô hình transfer learning thực hiện việc khảo sát, đánh giá hiệu năng của bảy mô hình học máy được xây dựng trên bộ dữ liệu dòng tế bào và sau đó áp dụng dự đoán đáp ứng thuốc cho bệnh nhân. So sánh hiệu năng dự đoán giữa các phương pháp học máy bao gồm các mô hình hồi quy tuyến tính, Lasso, Ridge và Elastic Net) và các phương pháp phân loại Random Forest (rf); Random Forest ranger (rf_ranger), Support Vector Machine (SVM) cho thấy rằng hiệu năng dự đoán tùy vào mỗi phương pháp và tùy thuộc vào bộ dữ liệu. Đồng thời các mô hình được xây dựng trên các dòng tế bào ung thư không thể hoạt động tốt trên mọi tập dữ liệu bệnh nhân ung thư cụ thể. Điều này có thể do sự không đồng nhất trong các phép đo đáp ứng thuốc của bệnh nhân ở các bộ dữ liệu khác nhau và các nguyên nhân khách quan khác.

- Các mô hình học sâu với các ưu điểm của nó có thể học được các tín hiệu dự đoán từ các nguồn dữ liệu lớn này mà không cần trích chọn đặc trưng đang là một xu hướng cải thiện đáng kể về cả độ chính xác và hiệu năng của phương pháp. Tuy nhiên dữ liệu biểu diễn thuốc một cách tự nhiên - dưới dạng đồ thị phân tử chưa được áp dụng hoặc được được áp dụng một và thử nghiệm trên các kịch bản cách khác nhau. Ngoài ra các mô hình chưa tích hợp đa dạng các loại dữ liệu -omics để học được nhiều nhất các đặc trưng của dòng tế bào.

Do vậy, luận án tập trung theo các hướng chính như sau:

- Nghiên cứu giải pháp biểu diễn dữ liệu thuốc một cách tự nhiên hơn dưới dạng đồ thị phân tử, áp dụng mô hình học các biểu diễn đồ thị tiên tiến để dự đoán đáp ứng thuốc.

- Nghiên cứu giải pháp tích hợp đa dạng các dữ liệu -omics khác nhau để cải tiến hiệu năng mô hình dự đoán đáp ứng thuốc.

- Nghiên cứu giải pháp tích hợp dữ liệu cho bài toán dự đoán đáp ứng đa thuốc.

- Nghiên cứu giải pháp tích hợp đa dữ liệu -omics với dữ liệu cấu trúc mạng tương tác protein để dự đoán đáp ứng đa thuốc.

1.4. KẾT LUẬN CHƯƠNG

Trong chương này, luận án đã trình bày tổng quan cơ sở lý thuyết về dữ liệu y sinh học và các phương pháp toán dựa trên mô hình học sâu, mô hình mạng nơ-ron đồ thị và các biến thể; các phương pháp tích hợp dữ liệu. Luận án đồng thời tổng hợp các phương pháp tính toán, đã được đề xuất cho hai bài toán quan trọng của đáp ứng thuốc là dự đoán đáp ứng thuốc cho đơn thuốc và dự đoán đáp ứng đa thuốc.

Các phân tích cơ bản về các phương pháp tiên tiến hiện nay, các vấn đề còn tồn tại và các hướng nghiên cứu có thể tiếp cận có thể giải quyết các vấn đề còn tồn tại của các nghiên cứu trước đây đã được trình bày. Với sự đa dạng về dữ liệu -omics cao hiện nay cũng là điều kiện tiềm năng để luận án đề xuất các giải pháp tính toán nhằm nâng cao hiệu năng dự đoán đáp ứng thuốc.

CHƯƠNG 2 – GIẢI PHÁP TÍCH HỢP DỮ LIỆU TRONG DỰ ĐOÁN ĐÁP ỨNG ĐƠN THUỐC

2.1. GIỚI THIỆU CHUNG

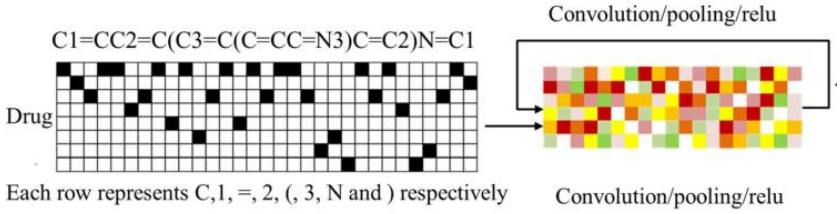
Dự đoán đáp ứng thuốc là một vấn đề cần nghiên cứu quan trọng trong y học chính xác hiện nay. Đã có nhiều phương pháp dự đoán dựa trên học máy, đặc biệt gần đây là các phương pháp dựa trên học sâu đã được đề xuất và mang lại các kết quả tiềm năng. Tuy nhiên, những phương pháp này thường mô hình hóa dữ liệu biểu diễn thông tin thuốc dưới dạng cấu trúc đơn giản như dạng chuỗi phân tử hóa học, dạng ảnh hoặc dấu vân tay (fingerprint). Các cách biểu diễn này chưa phải là cách biểu diễn tự nhiên của cấu trúc hóa học các phân tử thuốc như việc biểu diễn cho các dạng mạch vòng, mạch nhánh, số liên kết, đồng hình... trong phân tử hóa học. Ngoài ra, sự phát triển liên tục của các công nghệ thông lượng cao, làm tăng số lượng dữ liệu –omics khác nhau cũng là thách thức không nhỏ trong việc tích hợp dữ liệu cho bài toán dự đoán.

Trong chương này, luận án đề xuất hai giải pháp để dự đoán đáp ứng đơn thuốc: (1) giải pháp học các biểu diễn đồ thị phân tử thuốc dựa trên một số mô hình mạng nơ-ron đồ thị để dự đoán đáp ứng thuốc; (2) GraOmicDRP – tích hợp đa dữ liệu -omics và dữ liệu biểu diễn đồ thị phân tử thuốc để dự đoán đáp. Trong đó, GraphDRP, thuốc được biểu diễn dưới dạng tự nhiên hơn bằng đồ thị phân tử hóa học với các đỉnh là các nguyên tố hóa học, cạnh là liên kết giữa các nguyên tử đó. Các đặc trưng ẩn của phân tử thuốc được học thông qua mạng nơ-ron đồ thị. Trong khi đó các dòng tế bào được mô tả dưới dạng các vec-tơ nhị phân biểu diễn thông tin đột biến gen (MUT) và biến thể số lượng bản sao (CNA). Các đặc trưng biểu diễn cho thuốc và dòng tế bào đã được học thông qua các lớp tích chập, sau đó được kết hợp thành các biểu diễn đặc trưng cho từng cặp dòng tế bào - thuốc. Đề xuất GraOmicDRP dựa trên mô hình tích hợp muộn là mô hình cho thấy hiệu năng tiềm năng trong quá trình dự đoán. Mô hình này áp dụng cách biểu diễn thông tin dữ liệu thuốc dưới dạng đồ thị phân tử như đề xuất GraphDRP kết hợp với các cách kết hợp dữ liệu -omics khác nhau để trích xuất đặc trưng của các dòng tế bào làm tăng cường thông tin có ý nghĩa của các cặp thuốc – dòng tế bào trong quá trình dự đoán.

Cả hai đề xuất trên được triển khai bằng các thực nghiệm cụ thể và cho ra hiệu năng dư đoán tốt tốt hơn các phương pháp tính toán dự đoán tiên tiến. Kết quả được trình bày trong các công trình nghiên cứu đã được công bố số 1 và số 2.

2.2. CÁC NGHIÊN CỨU LIÊN QUAN

Các mô hình học sâu áp dụng cho bài toán dự đoán đáp ứng đơn thuốc được đề xuất gần đây cho thấy có khả năng học các đặc trưng ẩn của thuốc và, dữ liệu -omics tốt hơn các mô hình học máy truyền thống [22], [21], [59]. Các hướng này thường sử dụng dữ liệu biểu diễn thuốc dạng ảnh hoặc chuỗi mà chưa tiếp cận hướng biểu diễn dữ liệu đồ thị phân tử thuốc. Các phương pháp này cũng thường áp dụng mô hình CNN hoặc MLP để trích xuất các đặc trưng phân tử thuốc và dòng tế bào, như tCNNs [21], CDRscan [40], DeepDSC [61]. So sánh với cách tiếp cận theo học máy cổ điển các mô hình học sâu đã cho thấy kết quả vượt trội hơn nhiều. Trong khi các phương pháp dựa trên CNN những năm gần đây đã đạt được thành công trong thị giác máy tính [92], [93] và xử lý ngôn ngữ tự nhiên [94], [95], thì tCNNs [21], là phương pháp tiên tiến đầu tiên áp dụng mô hình CNN để học các biểu diễn dữ liệu cho cả thuốc và dòng tế bào. tCNNs xây dựng tập từ điển cho dữ liệu chuỗi ký tự trong chuỗi SMILES của thuốc, mỗi thuốc được biểu diễn dưới dạng ma trận nhị phân (one-hot), trong đó mỗi hàng một vec-tơ nhị phân biểu diễn vec-tơ đặc trưng cho mỗi ký tự (ví dụ: C, 1, =, (, …). Sau đó mạng nơ-ron tích chập 1-chiều (CNN1D) được áp dụng để trích xuất đặc trưng biểu diễn thuốc. CNN1D cũng được áp dụng để trích xuất đặc trưng của mỗi vec-tơ một chiều biểu diễn dữ liệu cho dòng tế bào (Hình 2.1). Mô hình này đã cho thấy hiệu năng vượt trội hơn các phương pháp trước đó, tuy nhiên việc biểu diễn các ký tự hay các nguyên tố hóa học trong phân tử thuốc theo các vec-tơ nhị phân này chưa cho thấy được mối liên kết giữa các nguyên tử; thiếu thứ tự liên kết giữa chúng trong phân tử thuốc. Do đó tCNNs chưa biểu diễn được dạng cấu trúc hình học đầy đủ của phân tử, từ đó có thể làm mất đi thông tin cấu trúc của thuốc.



Hình 2.1. Biểu diễn thuốc trong mô hình tCNNs[21]

Trong khi đó, mạng nơ-ron đồ thị (GNN) đang được áp dụng và mang lại những kết quả đáng chú ý trong nhiều lĩnh vực, đặc biệt đạt được các kết quả khả quan cho các nghiên cứu liên quan đến khai phá thuốc nói chung cũng như tác vụ dự đoán đáp ứng thuốc nói riêng. Ví dụ, GraphDTA [96], dự đoán ái lực thuốc nhắm mục tiêu (drug-target affinity), trong đó thuốc được biểu diễn dưới dạng đồ thị, các mô hình GNN được áp dụng cho việc học các biểu diễn thuốc cũng hiệu năng tốt nhất so với các phương pháp dựa trên h[60], [62]ọc sâu khác biểu diễn thuốc dưới dạng chuỗi ký tự. Một số đề xuất gần đây như sử dụng cơ chế transformer để tăng cường học các biểu diễn dữ liệu cho dự đoán đáp ứng thuốc. Trong đó GraTransDRP [62] kế thừa từ hiệu quả của việc áp dụng đồ thị để biểu diễn dữ liệu thuốc giống như đề xuất 1 – GraphDRP và đề xuất 2 – GraOmicDRP (được trình bày cụ thể trong phần tiếp theo), mô hình bổ sung lớp transformer trong các khối GNN để tăng cường học các biểu diễn phân tử thuốc. Cơ chế transformer đã cho thấy tiềm năng dự đoán tuy nhiên đòi hỏi hạ tầng tính toán đủ mạnh và tối ưu dữ liệu hơn, cụ thể trong GraTransDRP, bộ dữ liệu GE đã thực hiện giảm chiều dữ liệu (17,737 thành 1000) có thể sẽ không học được hết các đặc trưng ẩn của GE, trong khi nhiều nghiên cứu đã chứng minh GE là dữ liệu được chứng minh mang ý nghĩa cho dự đoán.

Bên cạnh đó, một số mô hình tích hợp đa dữ liệu -omics đã được đề xuất để dự đoán đáp ứng thuốc như DeepDR [97], MOLI [20]. Cụ thể, DeepDR là mô hình hồi quy sử dụng sử dụng dữ liệu đột biến gen và biểu hiện gen của tập dữ liệu TCGA làm đầu vào tiền huấn luyện (pre-train) trên hai bộ atutoencoder (AE) để trích xuất biểu diễn các dữ liệu -omics, tiếp đó mô hình gắn kết phần mã hóa (encoder) của các AE này với khối dự đoán để huấn luyện và thử nghiệm dự đoán IC_{50} trên các dữ liệu GE và MUT của bộ dữ liệu CCLE. Trong khi đó MOLI là mô hình phân loại dự đoán đáp ứng từng thuốc cụ thể, mô hình này tích hợp dữ liệu đa -omics (MUT, CNA và GE) của các dòng tế bào. Cả hai mô hình tiên tiến này đều là mô hình tích hợp muộn đa dữ liệu -omics. Tuy nhiên cả hai phương pháp này chưa sử dụng dữ liệu biểu diễn thuốc cho mô hình dự đoán đáp ứng và việc tích hợp -omics cũng chưa đa dạng các kết hợp dữ liệu -omics khác (ví dụ: dữ liệu methyl hóa).

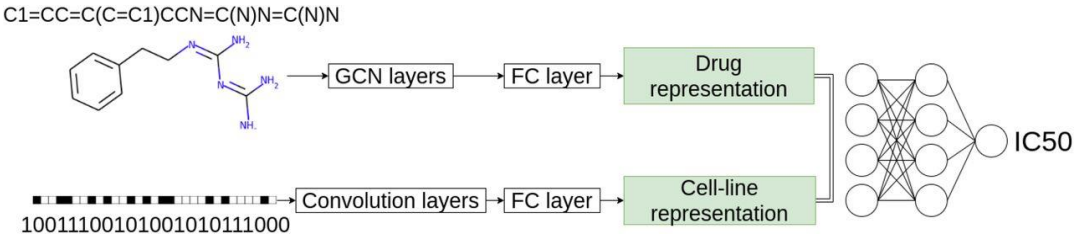
Do đó, trong chương này, luận án trình bày hai đề xuất GraphDRP, GraOmicDRP để áp dụng cách biểu diễn dữ liệu thuốc dạng đồ thị và tích hợp đa dữ

liệu -omics để cải tiến hiệu năng dự đoán đáp ứng đơn thuốc. Các giải pháp đề xuất được đánh giá hiệu năng và so sánh với phương pháp học sâu tiên tiến trên như tCNNs [21], DeepDR [97], MOLI [20]. Kết quả cho thấy hiệu quả rõ rệt của đề xuất trên các kịch bản khác nhau như dự đoán hỗn hợp (Mixed), dự đoán đáp ứng cho thuốc mới (Blind-Drug) và dự đoán cho dòng tế bào mới (Blind-Cellline).

2.3. ĐỀ XUẤT GIẢI PHÁP HỌC DỮ LIỆU BIỂU DIỄN ĐỒ THỊ CỦA PHÂN TỬ THUỐC - GraphDRP

2.3.1. Phương pháp

Mô hình đề xuất được minh họa như trong Hình 2.2. Dữ liệu đầu vào bao gồm thông tin hóa học của thuốc và đặc điểm di truyền bộ gen của các dòng tế bào bao gồm đột biến (MUT) và biến thể số lượng bản sao (CNA).



Hình 2.2. Mô hình đề xuất dự đoán đáp ứng đơn thuốc - GraphDRP

Các đặc trưng phân tử thuốc được tổng hợp từ các thông tin biểu diễn dạng chuỗi SMILES [39] chuyển đổi thành dữ liệu dạng đồ thị dựa trên mã nguồn mở RDKit [98], đưa vào mô hình huấn luyện mô hình. Các đặc trưng các nguyên tử mô tả một nút trong đồ thị xây dựng từ DeepChem [99]. Mỗi nút chứa năm loại đặc điểm nguyên tử hóa học: ký hiệu nguyên tử (atom symbol), độ nguyên tử (atom degree) được tính bằng số láng giềng liên kết và Hydro, tổng số Hydro, giá trị ngầm định (implicit value) của nguyên tử và nguyên tử có thơm hay không. Các đặc trưng nguyên tử này tạo thành một vec-tơ đặc trưng nhị phân. Nếu tồn tại một liên kết giữa một cặp nguyên tử, một cạnh được thiết lập. Kết quả là, một đồ thị với các nút được phân bổ đã được xây dựng cho mỗi chuỗi SMILES đầu vào (Hình 2.4). Tiếp theo mạng nơ-ron đồ thị, một lớp được kết nối đầy đủ (lớp FC) cũng được sử dụng để chuyển đổi kết quả thành 128 chiều. Với các đặc tính quan trọng và khả năng học các biểu diễn đồ thị khác nhau của mạng nơ-ron đồ thị, nghiên cứu triển khai một số thực nghiệm trên một số mô hình mạng nơ-ron đồ thị tiên tiến như: GCN [100], GAT

[101], GIN [102]. Ngoài ra, GCN [100] có điểm yếu là không xử lý được các đỉnh có mối quan hệ phi tuyến tính, và không có khả năng học trọng số đối với các đỉnh hàng xóm khác nhau. Do vậy nghiên cứu cũng tiến hành thử nghiệm với mô hình kết hợp GAT-GCN để xem xét khả năng kết hợp ưu điểm của cơ chế chú ý trong GAT để tổng hợp biểu diễn đỉnh gốc dựa trên hệ số chú ý của các đỉnh láng giềng và cơ chế tích chập trên đồ thị (GCN) để tổng hợp thông tin từ hàng xóm của mỗi đỉnh của đồ thị trong việc dự đoán đáp ứng thuốc.

Trong các mô hình học sâu, mạng nơ-ron tích chập một chiều (CNN1D) thường được sử dụng để giảm kích thước của đối tượng đầu vào và đưa ra dự đoán tốt, do đó CNN1D thường được dùng để học các đặc trưng ẩn từ các đặc trưng ban đầu của bộ gen. Các đặc trưng bộ gen của các dòng tế bào được thể hiện bằng mã hóa dạng các one-hot vec-tơ (vec-tơ nhị phân). Qua các lớp tích chập một chiều, các đặc trưng được làm phẳng thành vec-tơ 128 chiều của biểu diễn dòng tế bào.

Cuối cùng, kết hợp vec-tơ biểu diễn thuốc và vec-tơ biểu diễn cho dòng tế bào tạo thành vec-tơ biểu diễn cặp tương tác thuốc – dòng tế bào (drug-cell line) 256 chiều. Vec-tơ này tiếp tục đưa vào khối dự đoán là mạng kết nối đầy đủ với số nút lần lượt là 1024 và 256, để dự đoán đáp ứng thuốc cho dòng tế bào.

Bộ dữ liệu

Các dự án sàng lọc độ nhạy thuốc đối với các dòng tế bào quy mô lớn như CCLE và GDSC đã tạo ra không chỉ -omics mà còn cả dữ liệu đáp ứng thuốc đối với thuốc chống ung thư trên hàng nghìn dòng tế bào. Các dự án này cung cấp dữ liệu quan trọng là dữ liệu -omics của hệ gen cho biết gen đột biến (MUT) hoặc biến thể số lượng bản sao (CNV) trong bộ gen. Dữ liệu về đáp ứng thuốc (IC_{50}) cho biết mức độ tính toán hiệu quả của thuốc trong việc ức chế sự sống và phát triển của các dòng tế bào ung thư. Trong đó, GDSC là cơ sở dữ liệu lớn nhất về độ nhạy của thuốc đối với các dòng tế bào ung thư với hàng trăm loại thuốc được thử nghiệm trên hơn một nghìn dòng tế bào trong cơ sở dữ liệu. Do đó, nghiên cứu đã chọn GDSC phiên bản 6.0 (https://www.cancerrxgene.org/) với 250 loại thuốc, 1.074 dòng tế bào làm bộ dữ liệu chuẩn cho nghiên cứu này.

Bộ dữ liệu thực nghiệm được trích xuất dựa trên dữ liệu hệ gen (genomics) bao gồm:

- 990 dòng tế bào ung thư từ 13 mô ung thư (tissues), và 56 loại ung thư cụ thể. Đại đa số mỗi dòng tế bào có 735 đặc trưng mã hóa biến đổi gen gồm mã hóa đột biến gen (MUT) và biến thể số lượng bản sao (CNV). Tuy nhiên có 42 dòng tế bào có số đặc trưng ít hơn 735, nên trong nghiên cứu không dùng các dòng tế bào này vào tập dữ liệu thử nghiệm.

- Bộ dữ liệu gồm 223 thuốc, mỗi thuốc biểu diễn dưới dạng một chuỗi ký tự hóa học theo chuẩn CanonicalSMILES. Tập giá trị đáp ứng thuốc (IC_{50}) của 250 thuốc và 1027 dòng tế bào tương ứng.

Tổng hợp các dữ liệu trên thu được bộ thử nghiệm gồm: 948 dòng tế bào, 223 thuốc, 172,114 cặp tương tác được thử nghiệm giữa thuốc và dòng tế bào, chiếm 81.4% tổng số cặp thuốc – dòng tế bào.

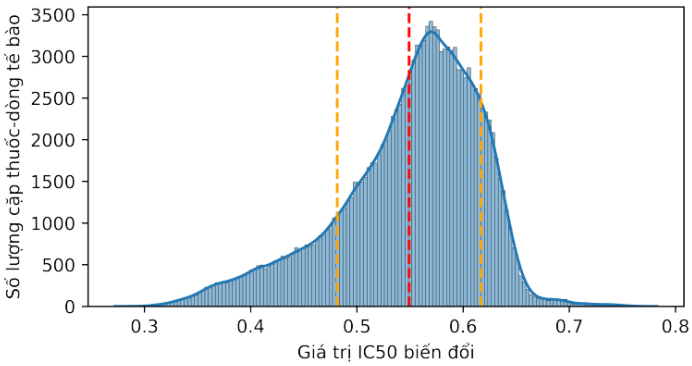
Tiền xử lý dữ liệu:

Biểu diễn dữ liệu dòng tế bào:

- Mỗi dòng tế bào được mô tả bằng một vec-tơ nhị phân có kích thước 735, trong đó 1 hoặc 0 cho biết liệu một dòng tế bào có hay không có biểu hiện sai lệch gen tương ứng.

- Các giá trị đáp ứng thuốc (IC_{50}) là dữ liệu liên tục trong khoảng từ (-10) đến (+12), được chuẩn hóa về khoảng (0, 1) theo công thức:

$$IC_{50} = \frac{1}{1 + e^{-\frac{IC_{50}}{10}}} \qquad (25)$$



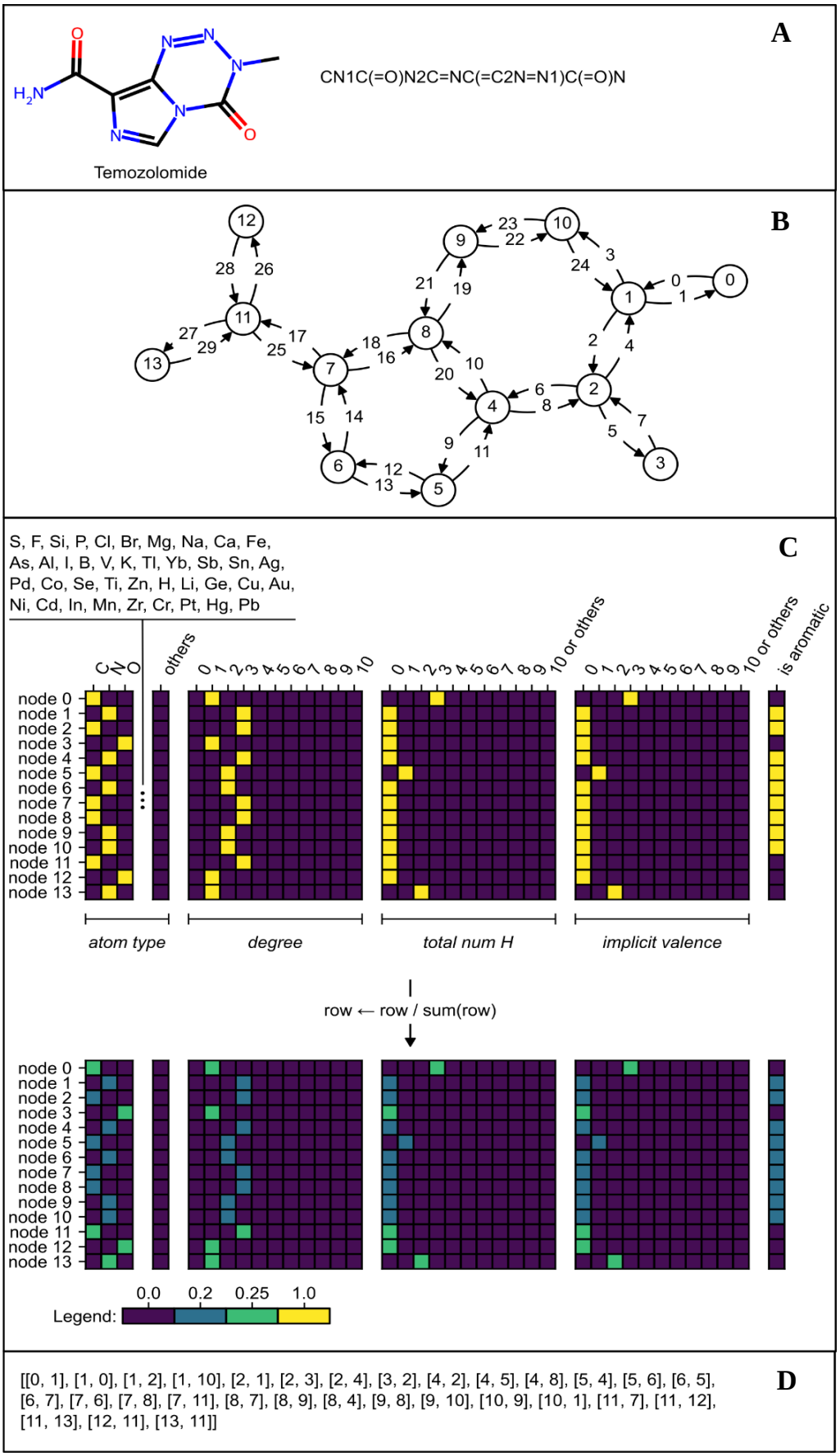
Hình 2.3. Biểu đồ phân phối giá trị IC_{50}

Biểu diễn dữ liệu phân tử thuốc:

Trong nghiên cứu này, xét năm đặc trưng đỉnh phân tử gồm: tên các nguyên tố hóa học, tổng số Hydro (H) ngầm định và công khai, tổng số liên kết ngầm định, nguyên tử có là thơm hay không. Với mỗi phân tử thuốc, chuỗi SMILES sẽ được biểu diễn dưới dạng đồ thị phân tử dựa vào các liên kết giữa các nguyên tử thành phần và các đặc trưng đỉnh của chúng. Trong đồ thị phân tử, ngoại trừ nguyên tử Hydro (H) được coi là ẩn, và không tính vào như một đỉnh của đồ thị thì các đỉnh là các nguyên tử. Nếu tồn tại liên kết giữa các nguyên tử, thì cạnh tương ứng được tạo thành. Tổng hợp các biểu diễn đặc trưng nguyên tử, các đỉnh của đồ thị được biểu diễn bởi vec-tơ đặc trưng đỉnh (nguyên tử) dạng one-hot 78 chiều (Bảng 2.1). Vec-tơ này được chuẩn hóa bằng cách tính tương quan giữa đặc trưng và tổng các đặc trưng của đỉnh đó. Hình 3.2 mô tả quá trình mã hóa dữ liệu phân tử thuốc từ chuỗi SMILES dạng chuỗi (string) thành dữ liệu biểu diễn dạng đồ thị phân tử (graph). Trong đó (A) biểu diễn chuỗi SMILES của phân tử thuốc (ví dụ: Temozolomide); (B) đồ thị vô hướng và thứ tự duyệt đỉnh của đồ thị; (C) mã hóa one-hot các đặc trưng đỉnh của đồ thị phân tử; (D) danh sách các cạnh của đồ thị, hình thành ma trận kề tương ứng. Mỗi đồ thị phân tử thuốc được biểu diễn dưới dạng một đồ thị (danh sách kề Hình 2.4 C) và thuộc tính ở mỗi đỉnh là vec-tơ đặc trưng đỉnh 78 chiều (Hình 2.4 D).

Bảng 2.1. Danh sách các thuộc tính của phân tử thuốc

Tên đặc trưng	Mã hóa đặc trưng	Số chiều
Atom	Mã hóa one-hot cho các nguyên tố hóa học	44
Degree	Mã hóa one-hot cho bậc của nguyên tố hóa học	11
TotalNumHs	Mã hóa one-hot cho tổng số Hydro tường minh và ngầm định của nguyên tố hóa học	11
ImplicitValence	Mã hóa giá trị số lượng liên kết ngầm định của nguyên tố hóa học	11
Aromatic	Mã hóa one-hot nguyên tố hóa học có là thơm hay không thơm	1
Tổng		78



Hình 2.4. Smiles-to-Graph của phân tử thuốc

2.3.2. Kịch bản thử nghiệm

Để đánh giá hiệu năng của mô hình đề xuất, nghiên cứu thực hiện ba thực nghiệm như: so sánh hiệu năng dự đoán đáp ứng thuốc của các cặp thuốc - dòng tế bào chưa biết (Mixed); so sánh hiệu năng dự đoán đáp ứng thuốc cho dòng tế bào chưa biết (Blind-Cellline) và so sánh hiệu năng dự đoán đáp ứng thuốc cho thuốc chưa biết (Blind-Drug) đồng thời điều tra sự đóng góp của đột biến gen đối với đáp ứng thuốc. Nghiên cứu cũng đánh giá, so sánh hiệu năng dự đoán đáp ứng thuốc đối với nghiên cứu tiên tiến gần nhất tCNNs. Một số mô hình mạng nơ-ron đồ thị tích chập gồm GCN, GIN, GAT, GCN-GAT đã được cài đặt thử nghiệm để đánh giá khả năng học các biểu diễn của thuốc cho bài toán này. Siêu tham số mô hình ban đầu, được chọn dựa trên nghiên cứu trước, sau đó, nghiên cứu đã điều chỉnh rất nhiều thông số như learning rate, batch-size để mô hình đạt được hiệu năng tốt nhất có. Để so sánh với các nghiên cứu trước đây, các mô hình này (tCNNs) được cài đặt và chạy lại, đo hiệu năng và so sánh với giải pháp đề xuất.

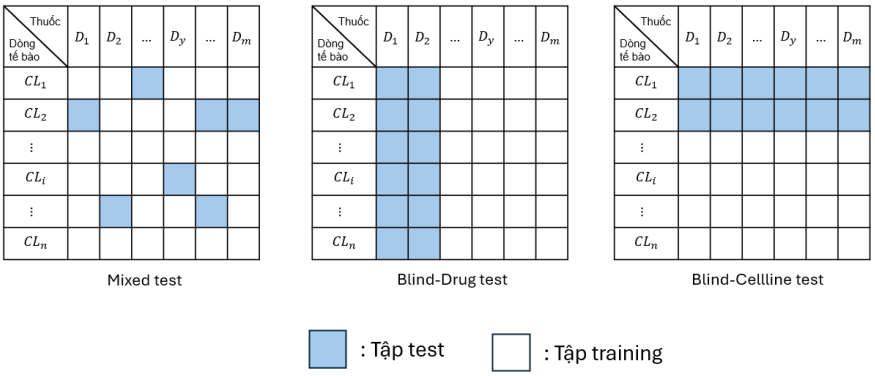
Kịch bản Mixed

Thực nghiệm này đã đánh giá hiệu năng dự đoán của các mô hình trên các thuốc - dòng tế bào đã biết (đã được thử nghiệm). Trong tất cả 211.404 cặp dòng tế bào thuốc có thể, GDSC cung cấp đáp ứng cho 172.114 cặp. Để duy trì tính tổng quát và tránh overfitting, dữ liệu được trộn ngẫu nhiên trước khi chia thành các tập dữ liệu huấn luyện, kiểm tra và đánh giá. Các cặp thuốc – dòng tế bào đã biết được chia theo tỉ lệ 80% là tập huấn luyện, 10% là tập đánh giá và 10% là tập kiểm tra.

Kịch bản Blind-Drug

Trong thực nghiệm Mixed, một loại thuốc hoặc một dòng tế bào có thể xuất hiện trong cả tập huấn luyện và thử nghiệm. Tuy nhiên, đôi khi chúng ta cần dự đoán đáp ứng của một loại thuốc mới/dòng tế bào mới, chẳng hạn như một loại thuốc mới được phát minh hay có một dòng tế bào bệnh mới cần nghiên cứu dự đoán đáp ứng thuốc. Việc đó sẽ rất có ý nghĩa trong định hướng nghiên cứu tiền lâm sàng và lâm sàng. Do đó dự đoán đáp ứng của các loại thuốc/dòng tế bào mới sẽ là thách thức lớn hơn. Với Blind-Drug, thuốc mới nằm trong bộ dữ liệu thử nghiệm sẽ không tồn tại trong bộ dữ liệu huấn luyện. Theo đó 90% (201/223) thuốc, và giá trị IC_{50} của chúng được chọn ngẫu nhiên trong giai đoạn huấn luyện và đánh giá với tỷ lệ 80% cho tập

huấn luyện và 10% thuốc cho tập đánh giá. Bộ dữ liệu thử nghiệm sẽ là 10% (22/223) thuốc còn lại.



Hình 2.5. Phân chia các tập dữ liệu theo các kịch bản thử nghiệm

Kịch bản Blind-Cellline

Tương tự với Blind-Drug, với việc dự đoán cho một dòng tế bào mới Blind-Cellline, các dòng tế bào mới cần dự đoán sẽ không có trong bộ dữ liệu huấn luyện. Vì vậy, nghiên cứu phân chia tổng cộng 90% (853/948) dòng tế bào được chọn ngẫu nhiên và giá trị IC_{50} của chúng được giữ cho giai đoạn huấn luyện. Các dòng tế bào còn lại, 10% (95/948), được sử dụng làm bộ thử nghiệm. Minh họa phép chia tập dữ liệu theo các kịch bản trên được minh họa như Hình 2.5.

Phép đo hiệu năng mô hình

Để đánh giá hiệu năng của các mô hình dự đoán, luận án sử dụng hai bộ tiêu chí đánh giá là lỗi trung bình bình phương gốc (Root Mean Squared Error - RMSE) và hệ số tương quan Pearson (Pearson correlation coefficient – CCp).

2.3.3. Cài đặt mô hình

Trước hết, mô hình thử nghiệm được cài đặt dựa trên việc tham khảo cấu hình từ các mô hình nghiên cứu đã được đề xuất trước đó. Tiếp theo, các tham số mô hình, số lớp mạng được điều chỉnh, chạy thử nghiệm và tinh tỉnh sau mỗi lần thử nghiệm để tối ưu hóa hiệu năng của mô hình. Cụ thể các mô hình đề xuất áp dụng được triển khai như sau:

- Trong mô hình dựa trên GCN nghiên cứu thực hiện cài đặt ba lớp GCN liên tiếp và hàm ReLU được áp dụng sau mỗi lớp. Một lớp global max pooling được thêm

ngay sau GCN cuối cùng lớp để tìm học các biểu diễn của toàn bộ đồ thị và tổng hợp vec-tơ biểu diễn thuốc.

- Trong mô hình dựa trên GAT, nghiên cứu thực hiện với hai lớp GAT sử dụng hàm kích hoạt ReLU và lớp global max pooling để tổng hợp biểu diễn đồ thị. Cụ thể, đối với lớp GAT đầu tiên, nghiên cứu sử dụng multi-head-attentions với 10 head và số lượng đặc trưng đầu ra bằng với số lượng đặc trưng đầu vào. Đầu ra của lớp GAT thứ hai là 128 chiều.

- Trong mô hình dựa trên GIN, năm lớp GIN liên tiếp được sử dụng để học các biểu diễn thuốc. Sau mỗi lớp, đồ thị được chuẩn hóa bằng lớp BatchNorm, hàm kích hoạt ReLU để học các đặc trưng thông qua biến đổi phi tuyến. Tương tự như kiến trúc GAT, một lớp tổng hợp gobal-max-pooling để tổng hợp đặc trưng toàn bộ đồ thị thành một vec-tơ biểu diễn đồ thị.

- Một mô hình GCN khác là sự kết hợp của GAT và GCN được đề xuất để học tối đa các đặc trưng của đồ thị. Trước tiên lớp GAT để học cách kết hợp các nút theo cơ chế attention, do đó, các đặc trưng nút được tổng hợp ở mức cao. Sau đó, các lớp GCN được sử dụng để tiếp tục học các đặc trưng đó thông qua các lớp tích chập để đưa ra dự đoán cuối cùng.

- Khối CNN để học các biểu diễn cho dòng tế bào bao gồm ba lớp tích chập với hàm kích hoạt là (ReLU) và lớp tổng hợp tối đa (max-pooling)..

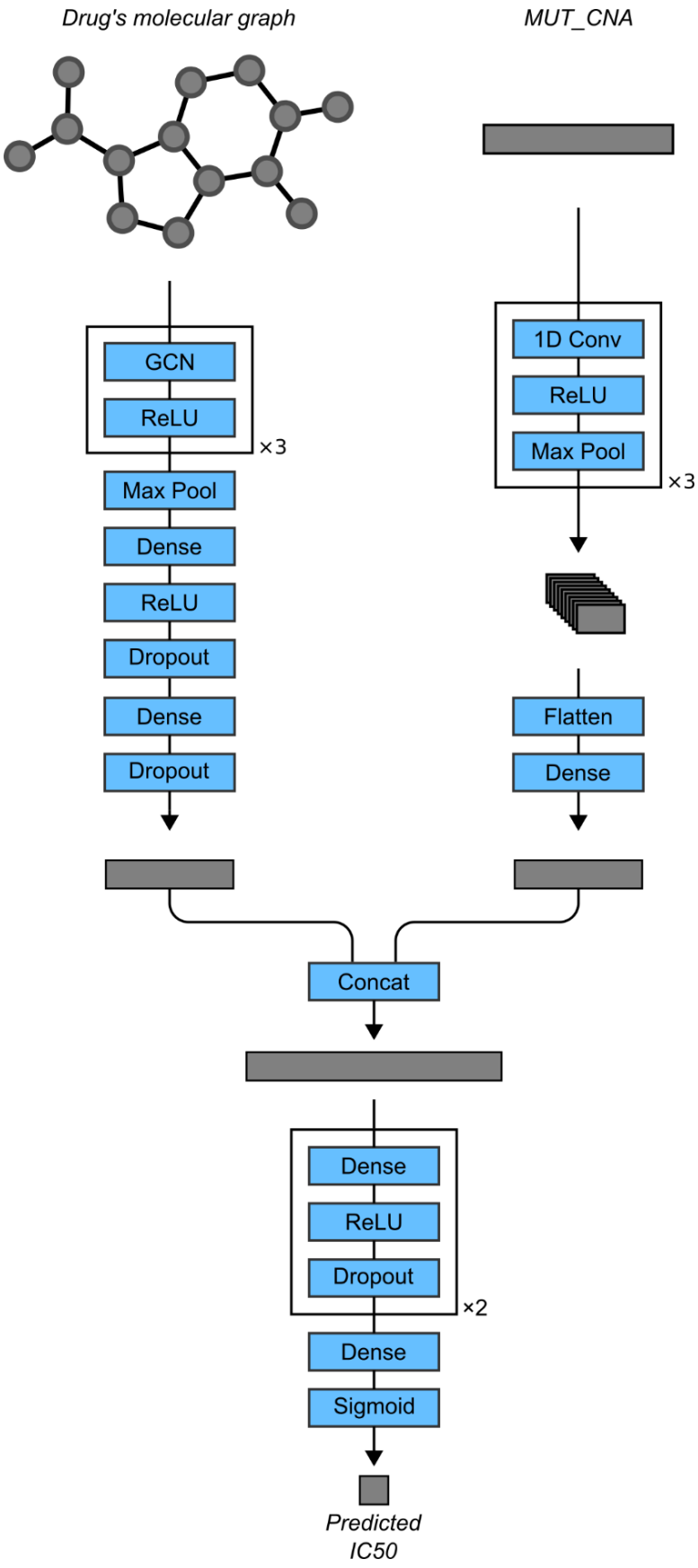
Cài đặt môi trường và siêu tham số huấn luyện

- Learning: 0.001; Batch size: 1024; số epoch: 300 được tinh chỉnh trong quá trình huấn luyện

- Thuật toán học: Tối ưu hóa Adam (Adam Optimization)

- Mô hình được huấn luyện và đánh giá trên server với cấu hình: 32 core CPU, RAM: 32 GB, Disk: 200 GB

Các mô hình, thuật toán được triển khai bằng ngôn ngữ lập trình Python cùng với các thư viện: Pytorch: hỗ trợ triển khai các mô hình deep learning; Rdkit: hỗ trợ biến đổi biểu diễn thuốc từ SMILES sang biểu diễn đồ thị; Pandas: hỗ trợ xử lý dữ liệu dạng .csv; Matplotlib: hỗ trợ trực quan hóa các thông tin



Hình 2.6. Mô hình triển khai GCN trong GraphDRP

2.3.4. Kết quả và đánh giá

Đánh giá hiệu năng dự đoán của mô hình đề xuất GraphDRP và tCNNs với kịch bản Mixed cho thấy mô hình đề xuất hoạt động tốt hơn tCNNs trong tất cả các giải pháp sử dụng mạng nơ-ron đồ thị để học các biểu diễn dữ liệu đồ thị tích chập đồ thị thay vì biểu diễn dữ liệu dạng chuỗi như tCNNs. Bảng 2.2 cho thấy tCNNs chỉ đạt RMSE là 0,0284 và CCp là 0,9160 trong khi GraphDRP(GCN) đạt hiệu năng thấp nhất trong số 4 biến thể mạng nơ-ron đồ thị áp dụng cho mô hình đề xuất, cụ thể chỉ số RMSE và CCp lần lượt là 0,0259 và 0,9216. Hơn nữa GraphDRP(GCN-GAT) và GraphDRP(GIN) đạt hiệu năng tốt nhất trên chỉ số RMSE (0,0243) và CCp (0,9310) tương ứng.

Bảng 2.2. So sánh hiệu năng các phương pháp trên đánh giá CCp và RMSE trong thử nghiệm Mixed

Methods		RMSE	CCp
tCNNs		0.0284	0.9160
GraphDRP	GCN	0.0259	0.9216
	GIN	0.0244	**0.9310**
	GAT	0.0250	0.9270
	GCN-GAT	**0.0243**	0.9308

Có thể thấy rằng khi đánh giá hiệu năng trên chỉ số RMSE thì GraphDRP(GIN) thu được kết quả tốt thứ hai (0,0244), chỉ nhỏ hơn một chút so với kết quả tốt nhất (0,0243). Vì vậy, giải pháp coi GIN là mô hình tốt nhất trong thực nghiệm này.

Bảng 2.3. So sánh hiệu năng các phương pháp trên chỉ số RMSE và CCp trong thử nghiệm Blind-Drug

Methods		RMSE	CCp
tCNNs		0.0680	0.0617
GraphDRP	GCN	**0.0542**	**0.3241**
	GIN	0.0602	0.0481
	GAT	0.0616	0.2751
	GCN-GAT	0.0610	0.1683

Trong thực nghiệm dự đoán đáp ứng cho thuốc mới, Bảng 2.3 cho thấy GraphDRP(GCN) là mô hình vượt trội nhất trên cả chỉ số đánh giá RMSE và CCp. Đặc biệt, xét về chỉ số CCp, GCN đã tăng gấp năm lần (0,3241) so với tCNNs (0,0617). Xét về chỉ số đánh giá RMSE, GIN đã đạt được RMSE (0.0602) thứ hai và

hầu hết các biến thể của mô hình mạng nơ-ron đồ thị áp dụng đạt chỉ số lỗi thấp hơn so với mô hình tCNNs. Trong khi đó, đối với CCp, ngoại trừ GIN, ba phương pháp dựa trên đồ thị khác đạt được hiệu năng tốt hơn khi so sánh với tCNNs. Điều này có thể thấy rằng với Blind-Drug, thuốc mới không xuất hiện trong giai đoạn huấn luyện làm giảm khả năng học các đặc trưng thuốc dẫn đến hiệu quả dự đoán thấp. Tuy nhiên, dữ liệu đồ thị cũng đã biểu thị nhiều thông tin và mô hình học được nhiều đặc trưng ẩn hơn là mô hình với dữ liệu dạng chuỗi (tCNNs) do đó hiệu năng dự đoán các mạng nơ-ron đồ thị cải thiện rõ rệt hơn ở thử nghiệm này. Ngoài ra với 10% thuốc không xuất hiện trong quá trình huấn luyện, cơ chế tích chập tập trung hóa thông tin từ các nút xung quanh có thể cải thiện khả năng biểu diễn mỗi nút, mang lại dự đoán tốt hơn cơ chế chú ý (GAT) và cơ chế tính đẳng cấu đồ thị (GIN).

Với thực nghiệm Blind-Cellline, Bảng 2.4 một lần nữa cho thấy sự vượt trội của GraphDRP so với tCNNs, tương đồng với kết quả hai kịch bản Mixed và Blind-Drug trên cả hai chỉ số đánh giá RMSE và CCp. Riêng phương pháp GIN đạt CCp tốt nhất là 0,8460 và RMSE tốt nhất là 0,0358. Có thể thấy rằng GIN đạt hiệu năng ổn định với thử nghiệm Mixed và Blind-Cellline, khi bộ dữ liệu thuốc mang đầy đủ thông tin hơn. Số lượng thuốc biểu diễn nhiều hơn cũng là một yếu tố ảnh hưởng đến hiệu năng mô hình dự đoán.

Bảng 2.4. So sánh hiệu năng các phương pháp trên chỉ số RMSE và CCp trong thử nghiệm Blind-Cellline

Methods		RMSE	CCp
tCNNs		0.0576	0.3490
GraphDRP	GCN	0.0363	0.8399
	GIN	**0.0358**	**0.8460**
	GAT	0.0380	0.8312
	GCN-GAT	0.0362	0.8402

Đối với cả thử nghiệm Blind-Drug và Blind-Cellline thì, hiệu năng dự đoán không tốt bằng trong thử nghiệm Mixed cho tất cả các mô hình. Điều này chỉ ra rằng khó dự đoán đáp ứng của dòng tế bào thuốc đối với các loại thuốc hay dòng tế bào chưa được học các đặc trưng trước đó. Quan sát thấy rằng hiệu năng dự đoán đáp ứng thuốc đối với dòng tế bào chưa biết đánh giá trên chỉ số CCp (nghĩa là giá trị trung bình là 0,8393 (± 0,0061) và 0,0366 (± 0,001) đối với RMSE tốt hơn so với hiệu năng

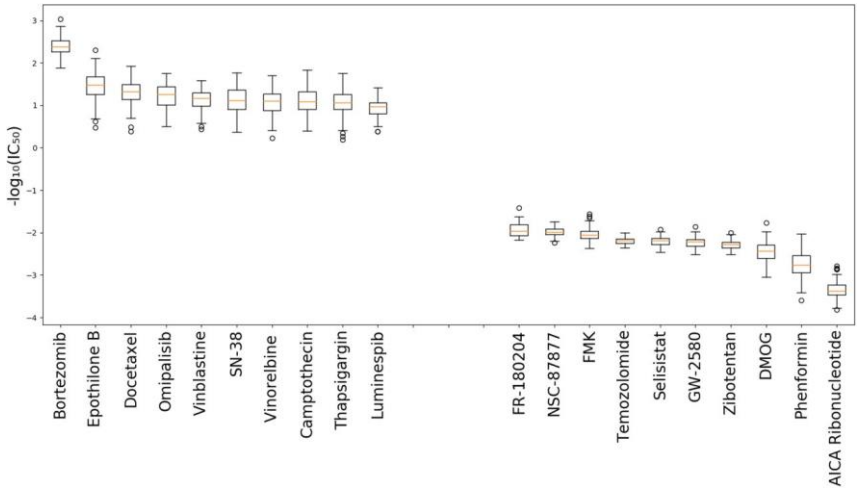
dự đoán đối với thuốc mới (tức là giá trị trung bình là 0,2039 (±0,1226) và 0,059 (±0,0034) đối với CCp và RMSE tương ứng).

Có thể thấy rằng, trong tất cả các kịch bản, mô hình mạng nơ-ron đồ thị vượt trội hơn tCNNs một cách toàn diện. Điều này là do trong tCNNs, định dạng chuỗi SMILES được sử dụng cho biểu diễn dữ liệu thuốc không phải là biểu diễn tự nhiên. Trong khi trong mô hình nghiên cứu, các mạng nơ-ron đồ thị được sử dụng để trích xuất thông tin từ biểu diễn đồ thị của thuốc, vì vậy hiệu năng tốt hơn.

Dự đoán giá trị đáp ứng cho các cặp thuốc – dòng tế bào chưa biết

Trong thực nghiệm này, mô hình tốt nhất được huấn luyện về thử nghiệm Mixed (áp dụng GIN) đã được sử dụng để dự đoán đáp ứng cho 39.290 (18.6%) cặp chưa biết. Hình 2.7 cho thấy mười loại thuốc có IC_{50} dự đoán cao nhất và thấp nhất. Điều đáng chú ý là ba loại thuốc đầu tiên có giá trị IC_{50} cao nhất và thấp nhất đều có kết quả tương tự như trong dự đoán của mô hình tCNNs. Thực nghiệm này cho thấy Bortezomib đạt IC_{50} thấp nhất, nghĩa là đây là loại thuốc nhạy nhất cho điều trị chống ung thư. Điều này được chứng minh trong báo cáo [103] chỉ rằng Bortezomib có tác dụng phân tử và tế bào khác biệt trong các tế bào ung thư vú ở người. Ngoài ra, nó có nhiều ứng dụng trong hoạt động giảm sự phát triển của khối u [104]. Loại thuốc chống ung thư hiệu quả thứ hai trong thực nghiệm này là Epothilone B, hoạt động bằng cách ngăn chặn sự phân chia tế bào thông qua tương tác với tubulin [105]. Ngược lại, AICA Ribonucleotide và Phenformin có IC_{50} cao nhất, có nghĩa là bệnh ung thư ít nhạy hơn với các loại thuốc này. Thật vậy, trong khi AICA Ribonucleotide đã được sử dụng lâm sàng để điều trị và bảo vệ chống lại tổn thương do thiếu máu cục bộ ở tim [106], thì Phenformin là thuốc trị đái tháo đường thuộc nhóm biguanide [21]. Do đó chúng không được sử dụng để chữa bệnh ung thư.

Nhìn chung, nghiên cứu này cho thấy hiệu quả của việc mô hình hóa dữ liệu biểu diễn đồ thị phân tử thuốc từ đó trích xuất các đặc trưng của thuốc thông qua các mạng nơ-ron đồ thị như GCN, GAT, GIN và biến thể GCN-GAT đã cho kết quả dự đoán tốt hơn so với cách biểu diễn dữ liệu phân tử thuốc dạng chuỗi (tCNNs). Kết quả này được thể hiện đồng nhất trong tất cả các kịch bản thử nghiệm như: thử nghiệm hỗn hợp (Mixed) thử nghiệm dự đoán đáp ứng trên thuốc mới (Blind-Drug), thử nghiệm dự đoán đáp ứng thuốc cho dòng tế bào mới (Blind-Cellline).



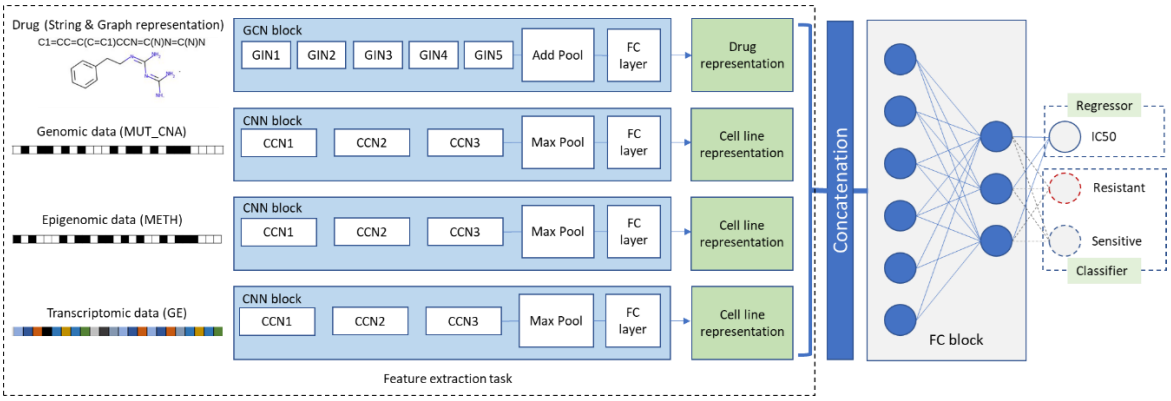
Hình 2.7. Biểu đồ 10 thuốc có giá trị IC_{50} được dự đoán tốt nhất và thấp nhất cho các cặp thuốc – dòng tế bào chưa biết

2.4. ĐỀ XUẤT GIẢI PHÁP TÍCH HỢP ĐA DỮ LIỆU -OMICS VÀ DỮ LIỆU BIỂU DIỄN ĐỒ THỊ PHÂN TỬ THUỐC - GraOmicDRP

2.4.1. Phương pháp GraOmicDRP

Việc tích hợp dữ liệu là một giải pháp tổng hợp các thông tin dữ liệu đầu vào nhằm học được tối đa các đặc trưng và các mối quan hệ giữa chúng của dữ liệu. Đối với một số phương pháp dự đoán đáp ứng thuốc trước đây như tCNNs [21] hay GraphDRP, dữ liệu biểu diễn cho các dòng tế bào mới chỉ dựa trên dữ liệu -omics của hệ gen. Trong khi đó, dữ liệu biểu diễn đặc trưng cho người bệnh hay dòng tế bào bệnh không chỉ có dữ liệu -omics của hệ gen mà còn có các dữ liệu -omics khác như dữ liệu biểu hiện gen, dữ liệu gen di truyền biểu sinh. Với dữ liệu biểu hiện gen cho biết lượng RNA được phiên mã từ DNA, biểu thị mức độ biểu hiện của gen hay mức độ hoạt động của gen ở một trạng thái nhất định (ví dụ: bệnh tật hoặc bình thường) trong một tế bào, thì dữ liệu –omics của hệ di truyền biểu sinh cho biết sự thay đổi methyl hóa DNA (dạng sửa đổi di truyền học biểu sinh) của hệ gen hay không có nghĩa là những thay đổi có thể được quan sát thấy trong kiểu hình, nhưng không có trong kiểu gen (trình tự DNA). Bên cạnh đó, một số nghiên cứu trước đây đã chỉ ra rằng các dữ liệu -omics này mang nhiều thông tin ý nghĩa trong cơ chế sinh học và nâng cao hiệu năng dự đoán trong nhiều hướng nghiên cứu như [55], [107]. Do vậy, để có thể tổng hợp các thông tin trong quá trình phát triển của tế bào thành dạng biểu diễn đặc trưng cho dòng tế bào, nghiên cứu này thực hiện một phương pháp học sâu

tích hợp các dữ liệu -omics khác nhau để biểu diễn đặc trưng dòng tế bào để dự đoán đáp ứng thuốc cho các dòng tế bào.



Hình 2.8. Mô hình đề xuất dự đoán đáp ứng đơn thuốc - GraOmicDRP

Kế thừa nghiên cứu trước đây về mô hình hóa phân tử thuốc biểu diễn dưới dạng đồ thị và áp dụng mô hình mạng nơ-ron đồ thị để học các biểu diễn ẩn đồ thị phân tử thuốc, đề xuất GraOmicDRP triển khai giải pháp tích hợp ba dữ liệu -omics khác nhau với dữ liệu biểu diễn dạng đồ thị của thuốc để giải quyết bài toán này. Trong GraOmicDRP, thuốc được biểu diễn dưới dạng đồ thị liên kết giữa các nguyên tử tương tự như phương pháp GraphDRP, trong khi đó, các dòng tế bào được mô tả không chỉ bằng bộ gen mà còn bằng dữ liệu biểu hiện gen và biểu sinh. Dựa trên khảo sát một số biến thể GNN thử nghiệm cho GraphDRP, GIN được coi là giải pháp cho hiệu năng tốt nhất trong dự đoán đáp ứng thuốc. Do đó, giải pháp đề xuất này áp dụng GIN làm thành phần chính trong khối mạng nơ-ron đồ thị để học các đặc trưng của thuốc.

Trong giải pháp này, năm lớp GIN được triển khai trong khối GCN. Trong đó, MLP bao gồm hai lớp tuyến tính thay vì một lớp tuyến tính như gợi ý của [49] để mô hình phân biệt đồ thị thuốc hiệu quả hơn. Sau mỗi lớp GIN là hàm kích hoạt ReLU và BatchNorm để lớp chuẩn hóa dữ liệu đầu ra đồ thị. Sau đó, một lớp tổng hợp global add pooling được thêm vào để kết hợp một vec-tơ biểu diễn đồ thị và cuối cùng, một lớp được kết nối đầy đủ (FC) làm phẳng các kết quả thành 128 chiều (Hình 2.8).

Đối với việc học các đặc trưng của dòng tế bào, thay vì chỉ sử dụng các đặc trưng bộ gen của các dòng tế bào (MUT và CNA), GraOmicDRP có thể tích hợp nhiều loại dữ liệu -omics (biểu hiện gen và methyl hóa) để trích xuất thêm các đặc

trưng biểu diễn cho dòng tế bào cho mô hình dự đoán. Các đặc trưng -omics của mỗi dòng tế bào được biểu diễn dưới dạng các vec-tơ đặc trưng 1D, nên để học đặc trưng ẩn của mỗi dòng tế bào bệnh của từng omic, các vec-tơ này làm đầu vào cho mạng nơ-ron tích chập 1D. Các lớp tích chập 1D và pooling được sử dụng để trích xuất các dữ liệu ẩn trong dữ liệu. Do đó, mỗi loại dữ liệu -omics, giải pháp đã sử dụng một khối CNN riêng lẻ để học các đặc trưng của dòng tế bào. Cụ thể, kiến trúc một khối CNN bao gồm ba lớp tích chập với lớp tổng hợp tối đa (max-pooling) với hàm kích hoạt là (ReLU). Sau đó, đầu ra được làm phẳng tạo thành một vec-tơ 128 chiều của biểu diễn đặc trưng -omics của dòng tế bào. Mô hình như vậy tương đối linh hoạt để sử dụng dữ liệu đơn -omics cũng như mở rộng tích hợp dữ liệu đa dạng các dữ liệu -omics. Sau đó, vec-tơ kết hợp biểu diễn thuốc và biểu diễn dòng tế bào được coi như vec-tơ biểu diễn cặp tương tác thuốc – dòng tế bào được đưa vào khối dự đoán gồm hai lớp (FC) được kết nối đầy đủ để dự đoán các giá trị đáp ứng thuốc (giá trị IC_{50}). Đối với trường hợp này, GraOmicDRP đã được sử dụng như một mô hình hồi quy. Số nút của lớp FC đầu tiên là 1024 và số nút của FC thứ hai là 128, đầu vào cho khối FC này là một vec-tơ có kích thước 256 hoặc 384 hoặc 512 tùy thuộc vào cài đặt cho single–omics, pair of–omics, multi–omics, tương ứng.

Bên cạnh các mô hình hồi quy dự đoán đáp ứng thuốc dưới dạng giá trị liên tục (IC_{50}), việc xác định một dòng tế bào có khả năng đáp ứng (S: Sensitivity) hoặc kháng thuốc (R: Resitance) đối với thuốc như thế nào có ý nghĩa quan trọng trong y học chính xác. Do đó, mô hình hồi quy đã đề xuất có thể biến đổi thành mô hình phân loại để dự đoán đáp ứng dạng nhị phân. Vì vậy, dữ liệu mẫu ban đầu được nhị phân hóa thành hai lớp tương ứng với đáp ứng (S) và kháng thuốc (R) [20], đồng thời đầu ra mô hình dự đoán được thay thế bằng hàm phân loại softmax.

Mô hình được thử nghiệm với ba kịch bản: Mixed, Blind-Drug, Blind-Cellline như mô tả trong đề xuất GraphDRP.

Bộ dữ liệu:

Trong đề xuất này giải pháp đề xuất tiếp tục sử dụng các bộ dữ liệu GDSC[1] như đã được thu thập và chuẩn hóa trong đề xuất 1 của luận án cho các thử nghiệm

[1] https://www.cancerrxgene.org/downloads/anova

và đánh giá mô hình. Cụ thể bộ dữ liệu gồm 223 loại thuốc, 990 dòng tế bào và các giá trị đáp ứng thuốc theo IC_{50} được chuẩn hóa trong phạm vi (0,1). Ngoài ra, với giải pháp tích hợp dữ liệu đa -omics này, bộ dữ liệu tổng hợp thêm hai loại dữ liệu –omics gồm biểu hiện gen (GE), và dữ liệu methyl hóa (METH) của các dòng tế bào. Hai dữ liệu -omics này kết hợp với dữ liệu đột biến gen (MUT) và biến thể số lượng sao chép (CNA) được tích hợp với nhau tạo thành dữ liệu biểu diễn cho dòng tế bào. Dữ liệu MUT_CNA ở dạng nhị phân biểu diễn liệu một gen có chứa gen đột biến hay không. Tương tự như vậy, dữ liệu methyl hóa cũng được nhị phân hóa để biểu diễn liệu một gen có bị siêu methyl hóa hay giảm methyl hóa hay không, dữ liệu ở dạng [0,1]. Trong khi đó, GE cũng cho biết mức độ biểu hiện gen đo bằng giá trị liên tục, giá trị này được chuẩn hóa trong khoảng (0,1).

Bảng 2.5. Tổng hợp các bộ dữ liệu cho mô hình GraOmicDRP

Datasets	# Cell lines	# Features
Cell-GE	1,018	17,773
Cell-MUT_CNA	990	735
Cell-METH	790	378

Bộ dữ liệu thực nghiệm bao gồm:

- 990 dòng tế bào ung thư, mỗi dòng tế bào có 735 đặc trưng biểu diễn mức độ biến đổi gen (gồm MUT và CNA)

- 1018 dòng tế bào ung thư, mỗi dòng tế bào có 17.773 đặc trưng biểu hiện gen (GE)

- 790 dòng tế bào ung thư, mỗi dòng tế bào có 378 đặc trưng biểu diễn methyl hóa (METH)

- 223 thuốc, mỗi thuốc biểu diễn bằng một chuỗi ký tự hóa học dạng SMILES và tương tác thuốc với dòng tế bào.

- Bộ dữ liệu được chuẩn hóa bằng cách tổ hợp các bộ dữ liệu dựa trên dòng tế bào và bộ dữ liệu tương tác thuốc đối với dòng tế bào. Đối với bộ dữ liệu MUT_CNA, khác với đề xuất GraphDRP, với đề xuất tích hợp đa dữ liệu -omics này, để sử dụng tối đa dữ liệu biểu diễn dòng tế bào, nghiên cứu đã sử dụng tất cả 990 dòng tế bào,

42 dòng tế bào không đủ 735 đặc trưng đột biến gen, được chuẩn hóa chuyển các đặc trưng bị khuyết (chưa biết) thành giá trị 0.

Tập dữ liệu tổng hợp cuối cùng thu được các bộ dữ liệu đơn -omics (single-omics) và đa dữ liệu -omics (multi–omics). Bảng 2.6 tổng hợp các bộ dữ liệu đã được chuẩn hóa gồm số lượng dòng tế bào tương ứng cho mỗi kiểu dữ liệu -omics và số mẫu tương tác của dòng tế bào và thuốc.

Bảng 2.6. Bộ dữ liệu chuẩn hóa cho GraOmicDRP

Datasets		# Cell lines	# Samples
Single -omics	METH	676	150,761
	GE	857	191,034
	MUT_CNA	857	191,049
Multi -omics	GE & METH	663	147,891
	GE & MUT_CNA	838	186,864
	METH & MUT_CNA	676	150,761
	ALL	663	147,891

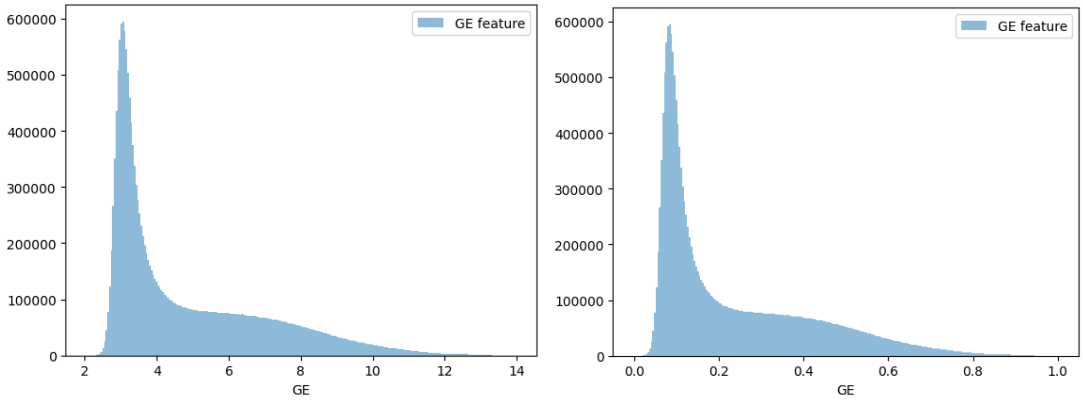
Tiền xử lý dữ liệu:

- Tương tự như Đề xuất 1 – mô hình GraphDRP, các giá trị đáp ứng về IC_{50} được chuẩn hóa về khoảng (0, 1)

- Dữ liệu biểu hiện gen (GE) có các khoảng biểu diễn chênh lệch nhau khá lớn, độ phân phối không đồng đều (GE_{min}= 2.06; GE_{max}= 13.96). GE được chuẩn hóa theo phương pháp min-max scaler với công thức:

$$x_{scale} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (2.1)$$

Trong đó x là giá trị biểu hiện gen, x_{min} là giá trị biểu hiện gen nhỏ nhất, x_{max} là giá trị biểu hiện gen lớn nhất. Hình 2.9 cho thấy phân bố dữ liệu biểu hiện gen trước và sau chuẩn hóa:



Hình 2.9. Biểu đồ phân bố dữ liệu gene expression

2.4.2. Kịch bản thử nghiệm

- Phương pháp chia bộ dữ liệu được thực hiện tương tự phép chia bộ dữ liệu được mô tả trong mục 2.3 của luận án. Cụ thể bộ dữ liệu là đủ lớn nên dữ liệu huấn luyện, kiểm tra và đánh giá được chia theo tỷ lệ 80%, 10%, 10% tương ứng, đảm bảo tương đồng về phân phối dữ liệu.

- Kịch bản thử nghiệm được thực hiện theo ba loại: Mixed, Blind-Drug, Blind-Cellline để kiểm chứng hiệu năng của mô hình cho việc dự đoán đáp ứng thuốc cho các thuốc đã biết, cho các thuốc mới và trên các dòng tế bào mới.

- Hiệu năng mô hình được đánh giá trên các độ đo RMSE và CCp.

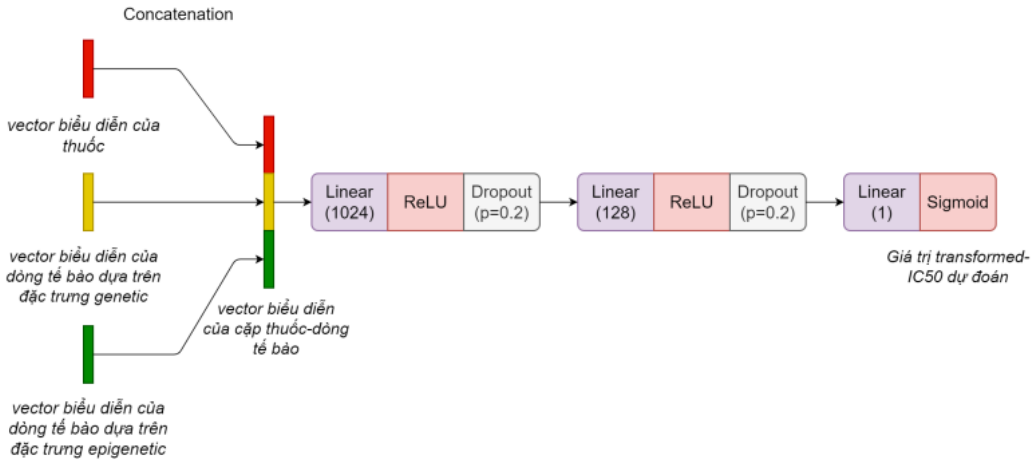
- Các thử nghiệm được tiến hành nhiều lần, các tham số được tinh chỉnh sau mỗi lần thử nghiệm.

2.4.3. Cài đặt mô hình

- Trong mô hình khối GIN gồm năm lớp GIN liên tiếp được sử dụng để học các biểu diễn thuốc, các lớp khác được thực hiện như các mô hình trên hình.

- Khối CNN1D được cài đặt để học các đặc trưng ẩn của các dữ liệu -omics khác nhau. Mỗi kịch bản tích hợp (single-omics, hoặc multi-omics) sẽ triển khai tích hợp các khối CNN tương ứng cho mô hình (Hình 2.11).

- Khối dự đoán gồm hai lớp FC (1024,128) theo sau mỗi lớp là hàm kích hoạt ReLU và dropout (0.2). vec-tơ đầu ra được biến đổi tuyến tính trước khi tính kết quả IC_{50} cuối cùng (ví dụ: Hình 2.10).



Hình 2.10. Khối dự đoán mô hình tích hợp multi-omic

Cài đặt môi trường và siêu tham số huấn luyện

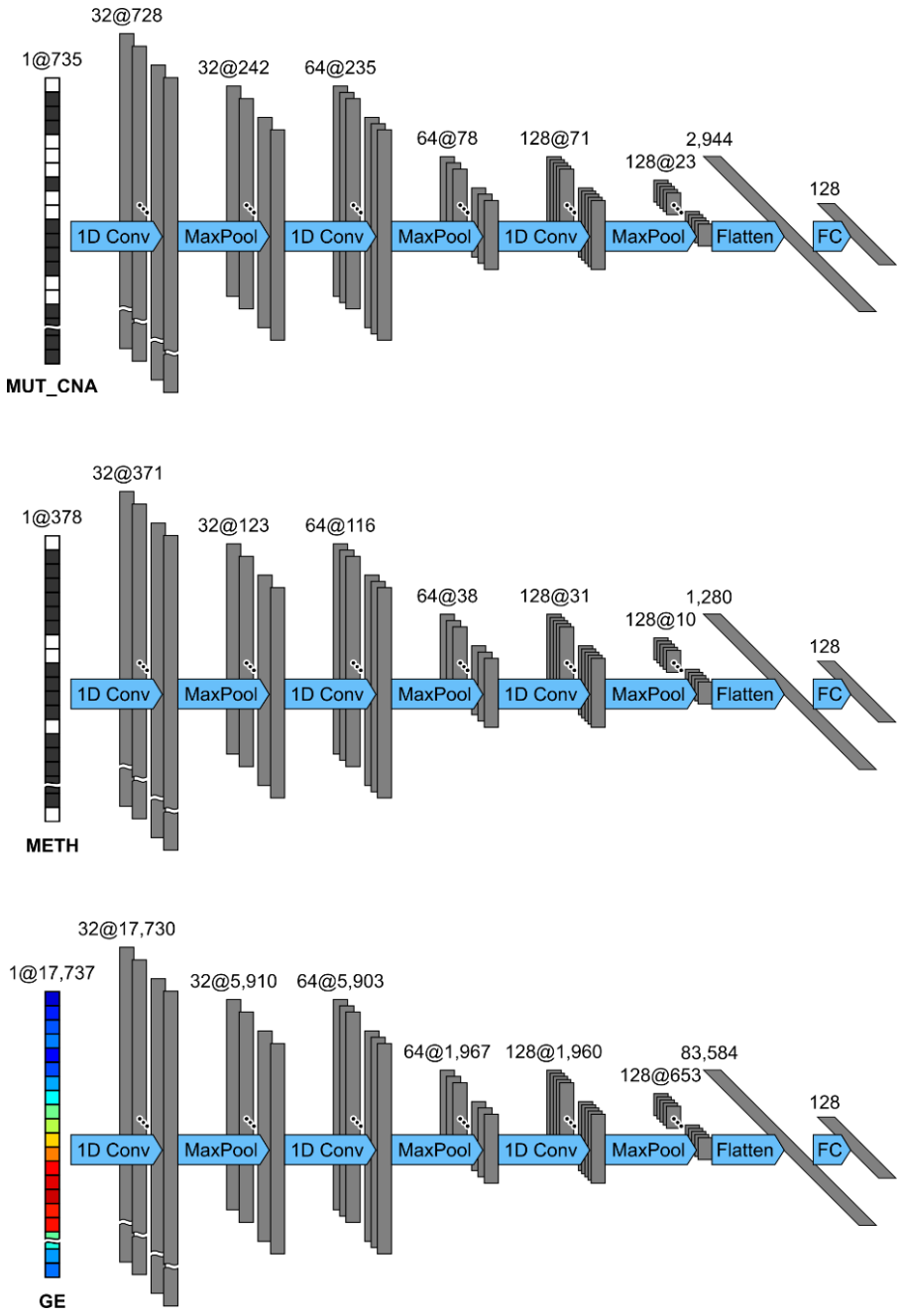
- Learning: 0.001
- Batch size: Tinh chỉnh trong quá trình huấn luyện (1024)
- Số epoch: Tinh chỉnh trong quá trình huấn luyện (300)
- Thuật toán học: Tối ưu hóa Adam (Adam Optimization)

Mô hình được huấn luyện và đánh giá trên server với cấu hình:

- 32 core CPU/ RAM: 32 GB /Disk: 200 GB
- 12 GB RAM GPU/Disk: 200 GB
- Colab Pro

Các mô hình, thuật toán được triển khai bằng ngôn ngữ lập trình Python cùng với các thư viện:

- Pytorch: hỗ trợ triển khai các mô hình deep learning
- Rdkit: hỗ trợ các thao tác phép biến đổi với thuốc
- Pandas: hỗ trợ xử lý dữ liệu dạng .csv
- Matplotlib: hỗ trợ trực quan hóa các thông tin



Hình 2.11. Mô hình học biểu diễn dữ liệu multi-omics của dòng tế bào

2.4.4. Kết quả và đánh giá

Đối với kịch bản Mixed

Đối với thử nghiệm Mixed, trước tiên giải pháp đề xuất đánh giá hiệu năng dự đoán trên dữ liệu single–omic (tức là trên từng tập dữ liệu METH, GE và MUT_CNA), và các kết hợp từng cặp -omics hoặc cả 3 -omics khác nhau. Kết quả cụ thể được minh họa như trong Bảng 2.7. Kết quả này cho thấy mô hình tích hợp đa dữ liệu -omics vượt trội hơn so với mô hình tích hợp đơn -omics ở hầu hết các thử nghiệm.

Đối với với mô hình tích hợp đơn -omics, mô hình kết hợp với GE đạt hiệu năng tốt nhất về cả RMSE (0,0259) và CCp (0,9195), trong khi đó, hiệu năng cho GraphDRP (tức là MUT_CNA) lần lượt là 0,0263 và 0,9120. Kết quả này chỉ ra rằng biểu hiện gen (GE) là dữ liệu giàu thông tin nhất so với các dữ liệu -omics khác khi chúng được sử dụng để dự đoán đáp ứng thuốc. Điều này có thể lí giải là do biểu hiện gen là một loại kiểu hình phân tử gần giống với cách các dòng tế bào đáp ứng với thuốc.

Đối với mô hình tích hợp đa dữ liệu -omics, giải pháp đề xuất đã tích hợp các cặp dữ liệu –omics (tức là GE & METH, GE & MUT_CNA và METH & MUT_CNA) và cả ba -omics (ALL). Bảng 2.7 cũng cho thấy rằng sự kết hợp giữa dữ liệu biểu hiện gen và methyl hóa(GE & METH) là tốt nhất về RMSE (0,0239) và CCP (0,9310). Điều này cũng tốt hơn bất kỳ trường hợp –single-omics nào. Thật thú vị là sự kết hợp không có GE (METH & MUT_CNA) lại đạt được hiệu năng kém hơn. Điều này một lần nữa chỉ ra rằng dữ liệu biểu hiện gen là thông tin phong phú hơn trong việc thể hiện các đặc điểm sinh học của các dòng tế bào trong dự đoán đáp ứng lại thuốc.

Methods		**RMSE**	**CCp**
Single -omics	METH	0.0279	0.9104
	GE	**0.0259**	**0.9165**
	MUT_CNA	0.0263	0.9120
Multi -omics	GE & METH	**0.0239**	**0.931**
	GE & MUT_CNA	0.0246	0.9236
	METH & MUT_CNA	0.0252	0.9277
	ALL	0.0244	0.9295

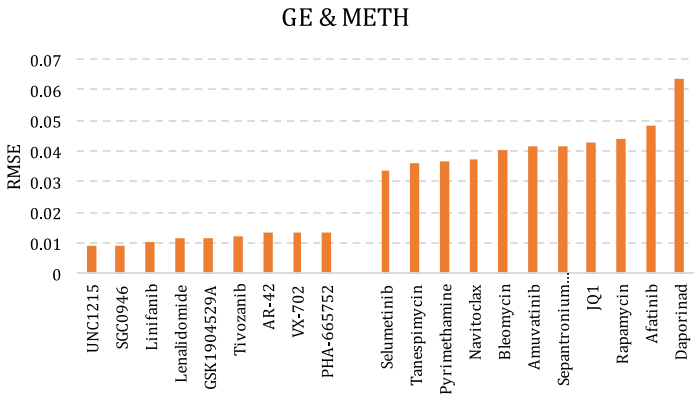
Bảng 2.7. So sánh hiệu năng các phương pháp trên kịch bản thử nghiệm Mixed

Hiệu năng dự đoán tổng thể của GraOmicDRP (Bảng 2.7) phù hợp với hiệu năng dự đoán trên từng thuốc thử nghiệm (Bảng 2.8). Thật vậy, GE là tốt nhất trong số ba dữ liệu đơn -omics thử nghiệm; trong khi đó đối với việc kết hợp đa dữ liệu -omics thì sự kết hợp GE & METH đã đạt được hiệu năng cao nhất, tiếp theo là bộ kết hợp GE & MUT_CNA và cả ba dữ liệu -omics (ALL) trên các chỉ số RMSE và CCp.

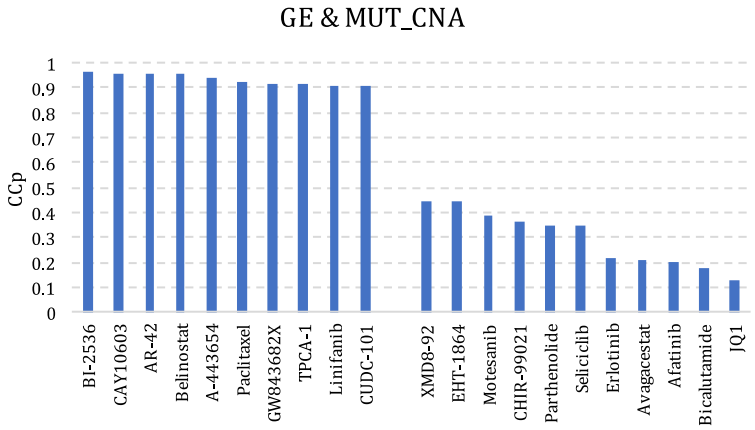
Bảng 2.8. So sánh hiệu năng các phương pháp cho từng thuốc trên kịch bản thử nghiệm Mixed

Methods		RMSE	CCp
Single -omics	METH	0.0278 (± 0.0087)	0.577 (± 0.1483)
	GE	**0.0254 (± 0.0073)**	**0.6603 (±0.1531)**
	MUT_CNA	0.0244 (± 0.0088)	0.6512 (± 0.1595)
Multi -omics	GE & METH	**0.0225 (± 0.0070)**	**0.7084 (± 0.1584)**
	GE & MUT_CNA	0.0229 (± 0.0072)	0.69 (± 0.1534)
	METH & MUT_CNA	0.0236 (± 0.0077)	0.6858 (± 0.1542)
	ALL	0.0234 (± 0.0077)	0.6835 (± 0.1605)

Mô hình dự đoán mười loại thuốc đạt hiệu quả cao nhất và thấp nhất khi kết hợp GE & METH (Hình 2.12) và GE & MUT_CNA (Hình 2.13) tương ứng về RMSE và CCp.



Hình 2.12. Mười thuốc có hiệu năng dự đoán cao nhất trên chỉ số RMSE trong kịch bản tích hợp GE & METH



Hình 2.13 Mười thuốc có hiệu năng dự đoán cao nhất trên chỉ số CCp trong kịch bản tích hợp GE & MUT_CNA

Kịch bản Blind-Cellline

Mục tiêu của thử nghiệm Blind-Cellline là dự đoán giá trị phản hồi lại cho các dòng tế bào không nhìn thấy (nghĩa là các dòng tế bào không ở trong giai đoạn huấn luyện). Để so sánh với kết quả tốt nhất của GraphDRP (tức là GraphDRP-GIN, mô hình GIN được sử dụng để tìm hiểu các đặc trưng của thuốc), giải pháp đề xuất đã chọn sự kết hợp tốt nhất của dữ liệu -omics về RMSE (tức là GE & METH) cho GraOmicDRP. So sánh hiệu năng về cả CCp và RMSE đối với thử nghiệm Blind-Cellline cho thấy giải pháp đề xuất đạt kết quả tốt hơn GraphDRP. Bảng 2.9 cho thấy GraOmicDRP có RMSE nhỏ hơn (0,0327) và CCP cao hơn (0,8766) so với GraphDRP.

Bảng 2.9. So sánh hiệu năng dự đoán đáp ứng thuốc cho dòng tế bào mới

Methods	RMSE	CCp
GraphDRP-GIN	0.0363	0.846
GraOmicDRP (GE & METH)	**0.0327**	**0.8766**

Kịch bản Blind-Drug

Mục tiêu của thử nghiệm Blind-Drug là dự đoán các giá trị đáp ứng đối với các loại thuốc mới chưa được thử nghiệm (tức là các loại thuốc không có trong giai đoạn huấn luyện). Tương tự như thử nghiệm Blind-Cellline, nghiên cứu đã sử dụng tổ hợp tích hợp dữ liệu –omics có hiệu năng dự đoán tốt nhất cho GraOmicDRP (GE & METH) để so sánh với kịch bản thử nghiệm tốt nhất của GraphDRP (tức là GraphDRP-GCN, mô hình GCN được sử dụng để học các biểu diễn thuốc). Bảng 2.10. cho thấy rằng phương pháp nghiên cứu đạt được hiệu năng tương đương trên chỉ số RMSE (nghĩa là 0,0542 và 0,0590 đối với GraphDRP-GCN và GraOmicDRP, tương ứng), nhưng GraOmicDRP tốt hơn trên chỉ số CCp (tức là 0,3241 và 0,4309 đối với GraphDRP-GCN và GraOmicDRP, tương ứng).

Bảng 2.10. So sánh hiệu năng dự đoán đáp ứng cho thuốc mới

Methods	RMSE	CCp
GraphDRP-GCN	**0.0542**	0.3241
GraOmicDRP (GE & METH)	0.059	**0.4309**

Ngoài ra, để điều tra tính hiệu quả của việc tích hợp biểu diễn đồ thị của thuốc với nhiều dữ liệu -omics của các dòng tế bào trong việc dự đoán đáp ứng của thuốc,

nghiên cứu đã so sánh hiệu năng dự đoán giữa GraOmicDRP và hai phương pháp tiên tiến khác là DeepDR và MOLI. Đây là hai phương pháp áp dụng mô hình tích hợp muộn đa dữ liệu -omics nhưng không tích hợp dữ liệu biểu diễn thuốc trong mô hình dự đoán. Khá tương đồng với nghiên cứu GraOmicDRP, DeepDR cũng cho thấy dữ liệu biểu hiện gen (GE) có liên quan trực tiếp đến chức năng sinh học bệnh hơn là dữ liệu đột biến (MUT). Hiệu năng dự đoán của DeepDR trên tập dữ liệu GE hơn hiệu năng trên tập dữ liệu MUT và tương đương với hiệu năng dự đoán của GE và MUT. Do đó, dựa theo kiến trúc và quy trình huấn luyện của DeepDR, nghiên cứu đã xây dựng lại mô hình này với pre-train trên bộ GE của TCGA và đánh giá trên tập dữ liệu GE của GDSC trên kịch bản Blind-Cellline. Kết quả thực nghiệm cho thấy DeepDR đạt 0,0330 RMSE và 0,8722 CCp; trong khi đó, giá trị của GraOmicDRP lần lượt là 0,0327 và 0,8791 (Hình 2.11). Kết quả chỉ này ra rằng mô hình đề xuất vượt trội hơn DeepDR trên cả hai chỉ số RMSE và CCp.

Bảng 2.11. So sánh hiệu năng của GraOmicDRP và DeepDR

Methods	RMSE	CCp
DeepDR (GE)	0.033	0.8722
GraOmicDRP (GE)	**0,0327**	**0,8791**

MOLI là mô hình phân loại dự đoán đáp ứng từng thuốc cụ thể, mô hình này tích hợp dữ liệu đa -omics (MUT, CNA và GE) của các dòng tế bào mà không có thông tin cấu trúc hóa học của thuốc. Cụ thể hơn, MOLI đã xây dựng các mô hình dự đoán riêng biệt cho sáu loại thuốc (tức là Docetaxel, Cisplatin, Gemcitabine, Paclitaxel, Erlotinib và Cetuximab), sau đó thử nghiệm trên bộ dữ liệu trên người và mô ghép PDX/TCGA cho các loại thuốc đó.

Bảng 2.12. So sánh hiệu năng của GraOmicDRP và MOLI

		Docetaxel	Erlotinib	Gemcitabine	Paclitaxel
MOLI (GE & MUT_CNA)		0.6438	0.7295	0.571	0.65
GraOmicDRP	GE_METH	0.7	**0.8304**	**0.8643**	0.8704
	GE & MUT_CNA	0.7304	0.8194	0.8159	**0.9444**
	MUT & METH	0.7281	0.6637	0.6773	0.7407
	ALL	**0.7414**	0.7708	0.8362	0.9259

Để so sánh hiệu năng dự đoán với MOLI, GraOmicDRP được hoạt động như một bộ phân loại, đồng thời mô hình MOLI được triển khai lại và thử nghiệm trên cùng bộ dữ liệu GDSC. Trong số sáu loại thuốc thử nghiệm trong mô hình MOLI, chỉ có bốn loại thuốc (là Docetaxel, Erlotinib, Gemcitabine và Paclitaxel) có sẵn trong cơ sở dữ liệu GDSC. Vì vậy, nghiên cứu đã so sánh hiệu năng dự đoán của mô hình đề xuất với MOLI dựa trên tính toán dự đoán của bốn loại thuốc này. Bảng 2.12. cho thấy rằng GraOmicDRP vượt trội hơn MOLI đối với tất cả các loại thuốc trên chỉ số đánh giá AUC.

2.5. KẾT LUẬN CHƯƠNG

Trong chương này, luận án đã trình bày hai giải pháp nghiên cứu cho dự đoán đáp ứng đơn thuốc là GraphDRP và GraOmicDRP. Trong đó giải pháp GraphDRP áp dụng cách học biểu diễn dữ liệu phân tử thuốc dưới dạng đồ thị - dạng biểu diễn cải tiến hơn so với các nghiên cứu trước đây – thông qua các biến thể mạng nơ-ron đồ thị khác nhau, kết hợp với dữ liệu thông tin di truyền của dòng tế bào để dự đoán đáp ứng thuốc. Giải pháp đề xuất tích hợp đa dữ liệu -omics GraOmicDRP là giải pháp cải tiến tiếp cho GraphDRP. Các giải pháp trình bày trong chương này này nằm trong công trình nghiên cứu số 1 và số 2 của tác giả và các cộng sự.

Trong chương này, thông qua việc nghiên cứu các mô hình tính toán dựa trên mạng nơ-ron đồ thị với dữ liệu đầu vào là dữ liệu biểu diễn đồ thị của phân tử thuốc, các nghiên cứu đề xuất đã cho thấy hiệu quả của mô hình dự đoán được nâng cao trong việc thay đổi cấu trúc dữ liệu biểu diễn phân tử thuốc từ biểu diễn dạng chuỗi sang biểu diễn đồ thị kết hợp dữ liệu biểu diễn dòng tế bào qua mạng nơ ron tích chập 1D. Các thử nghiệm GraphDRP trên các mô hình mạng nơ-ron đồ thị khác nhau cho thấy mô hình GIN có thể khả năng học các cấu trúc tương đồng của đồ thị phân tử thuốc tốt hơn các biến thể mạng nơ-ron đồ thị khác trong dự đoán đáp ứng thuốc.

Việc tích hợp và khai thác nhiều góc nhìn về dữ liệu khác nhau cũng là một thách thức cho bài toán dự đoán đáp ứng thuốc. Hiệu năng dự đoán đáp ứng thuốc cho các dòng tế bào này còn được cải thiện rõ rệt với giải pháp tích hợp đa dữ liệu -omics của các dòng tế bào. Tổng hợp các phân tích và kết quả thử nghiệm cho thấy GraOmicDRP tốt hơn GraphDRP cho tất cả các kịch bản thử nghiệm. Điều này cho

thấy tầm quan trọng của việc tích hợp nhiều dữ liệu –omic, đặc biệt là dữ liệu biểu hiện gen (GE), đối với các vấn đề dự đoán đáp ứng thuốc. Đối với việc tích hợp đa dữ liệu -omics, sự kết hợp của dữ liệu biểu hiện gen với các -omics khác cũng cho hiệu năng dự đoán tốt hơn so với các đơn -omics đó.

Dựa trên các kết quả của hai đề xuất nghiên cứu này, lợi thế của mạng nơ-ron đồ thị và tích hợp đa dữ liệu -omics được tiếp tục phát triển cho các nghiên cứu tiếp sau cũng như hai đề xuất cho dự đoán kết hợp thuốc ở chương 3. Trong cả hai đề xuất, các thử nghiệm Blind-Cellline và Blind-Drug cần được cải thiện hơn.

CHƯƠNG 3 – GIẢI PHÁP TÍCH HỢP DỮ LIỆU TRONG DỰ ĐOÁN ĐÁP ỨNG ĐA THUỐC

3.1. GIỚI THIỆU CHUNG

Để giảm các tác dụng phụ và độc tính cũng như đạt được các đáp ứng lâm sàng hiệu quả, các cách kết hợp đa thuốc là giải pháp quan trọng trong điều trị, đặc biệt với các bệnh phức tạp. So với điều trị đơn thuốc (monotherapy), điều trị đa thuốc (combination therapy) trong ung thư có thể ức chế tế bào ung thư và ngăn chặn sự xuất hiện kháng thuốc hiệu quả hơn, từ đó có thể làm tăng hiệu quả điều trị [108], [109]. Một số các phương pháp học máy và các phương pháp học sâu gần đây đã được đề xuất cho bài toán này để dự đoán giá trị đáp ứng đa thuốc cho các dòng tế bào hoặc phân loại kết hợp thuốc (tương hợp/tương kháng thuốc) [68], [110], [72], [111]. Tuy nhiên, các phương pháp này chưa giải quyết được vấn đề tích hợp nhiều dữ liệu -omics, dữ liệu biểu diễn thuốc hoặc chưa biểu diễn một cách tự nhiên dạng đồ thị hoặc chưa tổng hợp thông tin các cặp thuốc một cách đầy đủ.

Do đó, trong chương này, luận án trình bày hai đề xuất tích hợp dữ liệu cho bài toán dự đoán kết hợp thuốc: (1) đề xuất mô hình GraOmicSynergy, áp dụng mạng nơ-ron đồ thị (GIN – với nhiều ưu điểm được chứng minh trong đề xuất của chương 2) để học các biểu diễn thuốc và tăng cường cơ chế chú ý để tổng hợp biểu diễn độ kết hợp thuốc đối với các dòng tế bào, đồng thời tổng hợp đa dữ liệu -omics của các dòng tế bào để dự đoán khả năng kết hợp nhiều thuốc trong điều trị bệnh; (2) đề xuất AE-XGBSynergy – tích hợp nhiều dữ liệu -omics với thông tin cấu trúc mạng PPI để cải thiện dự đoán phân loại kết hợp thuốc. Trong GraOmicSynergy, mỗi thuốc được biểu diễn dưới dạng đồ thị phân tử của các liên kết giữa các nguyên tử, vec-tơ biểu diễn cặp thuốc tương tác với dòng tế bào được tổng hợp thông qua cơ chế chú ý, các biểu diễn của các dòng tế bào là là dữ liệu được học các biểu diễn ẩn thông qua mạng CNN1D không chỉ bởi một loại dữ liệu -omics (như gennomics) mà còn là tổng hợp bởi dữ liệu biểu hiện gen (transcriptomics) và dữ liệu methyl hóa (epigenomics). Với AE-XGBSynergy dự đoán đáp ứng đa thuốc bằng cách sử dụng thông tin cấu trúc mạng PPI kết hợp việc tăng cường tích hợp đa dữ liệu -omics cụ thể là dữ liệu methyl

hóa (epigenomics) và dữ liệu di truyền (genomics) được trích xuất biểu diễn thông qua bộ mã hóa encoder (trong AE).

Kết quả thử nghiệm của cả hai đề xuất trên được tiến hành và cho thấy hiệu quả của vượt trội của phương pháp đề xuất so với các phương pháp tiên tiến hiện nay trong việc dự đoán kết hợp thuốc. Bên cạnh đó, kết quả thử nghiệm cũng cho thấy các phương pháp tích hợp đa dữ liệu -omics mang lại hiệu năng tốt hơn so với đơn dữ liệu -omics trong hầu hết các cách kết hợp và các kịch bản thử nghiệm dự đoán kết hợp thuốc cho dòng tế bào.

3.2. CÁC NGHIÊN CỨU LIÊN QUAN

Gần đây, các nghiên cứu áp dụng phương pháp học sâu được áp càng phổ biến trong các nghiên cứu khám phá thuốc. Đặc biệt, trong vấn đề nghiên cứu về kết hợp thuốc, một số nghiên cứu gần đây như [68], [110], [72], [111]. DeepSynergy [68] là nghiên cứu đầu tiên đề xuất việc sử dụng DL để dự đoán tác dụng phối hợp thuốc. Đây là một mô hình mạng nơ-ron học sâu (DNN) sử dụng dữ liệu đáp ứng với thuốc, đặc điểm hóa học và dữ liệu biểu hiện gen (GE) để dự đoán đáp ứng thuốc. Mô hình đã được huấn luyện bằng cách sử dụng dữ liệu thuốc từ bộ dữ liệu O'Neil và dữ liệu -omics từ GDSC. DNN đạt hiệu năng tương đối cao hơn so với các phương pháp học máy truyền thống RFs, SVM. Tuy nhiên trong phương pháp này, dữ liệu thuốc mới biểu diễn dữ liệu fingerprint, chưa biểu diễn dạng đồ thị và chưa tích hợp dữ liệu trong dự đoán. Dựa trên thành công của một số nghiên cứu áp dụng "graph" trong dự đoán đáp ứng đơn thuốc, một vài các đề xuất dự đoán đáp ứng đa thuốc[72], [73] đã áp dụng graph trong việc học các dữ liệu đồ thị phân tử thuốc cho thấy hiệu quả tiềm năng của dự đoán. DeepDDS [72] là phương pháp tiên tiến gần đây sử dụng mô hình mạng nơ-ron đồ thị phân tử GAT và GCN để học các biểu diễn của cặp phân tử thuốc kết hợp với dữ liệu biểu diễn dữ liệu biểu hiện gen của dòng tế bào được học qua mô hình MLP dự đoán kết hợp thuốc.

Mạng tương tác protein (PPI) rất quan trọng cho hầu hết mọi quá trình trong tế bào cũng như xác định thuốc nhắm mục tiêu, do đó, việc khám phá thông tin mạng PPI đang được triển khai trong một số nghiên cứu như [111], [112], [113]. Transynergy [111] đã sử dụng thông tin mạng PPI và biểu hiện gen của các dòng tế bào để xây dựng các đặc điểm của dòng tế bào - thuốc trong đó các biểu hiện của

thuốc được trích xuất thông qua bước đi ngẫu nhiên với thuật toán khởi động lại (RWR) chưa được xem xét đến cấu trúc liên kết của mạng. GraphSynergy [112] đã đề xuất một phương pháp dựa trên mạng tích chập đồ thị tập trung vào cấu trúc kết nối toàn cầu và cục bộ của mạng PPI để xác định sự phối hợp giữa các loại thuốc chống lại các dòng tế bào. NEXGB [113] là phương pháp chỉ sử dụng thông tin cấu trúc mạng PPI để trích xuất đặc trưng biểu diễn thuốc và dòng tế bào cho dự đoán phân loại đáp ứng đa thuốc. Tuy nhiên, cả thông tin cấu trúc mạng PPI tiện ích GraphSynergy và NEXGB đều chưa được tích hợp với các dữ liệu sinh học khác của các dòng tế bào.

Do đó trong chương này, luận án trình bày hai đề xuất GraOmicSynergy và AE-XGBSynergy tích hợp đa dữ liệu -omics để cải tiến hiệu năng dự đoán đáp ứng đa thuốc. Các giải pháp đề xuất được đánh giá hiệu năng và so sánh với phương pháp học sâu tiên tiến trên như DeepSynergy, DeepDDS, NEXGB cho thấy hiệu năng được cải tiến rõ rệt.

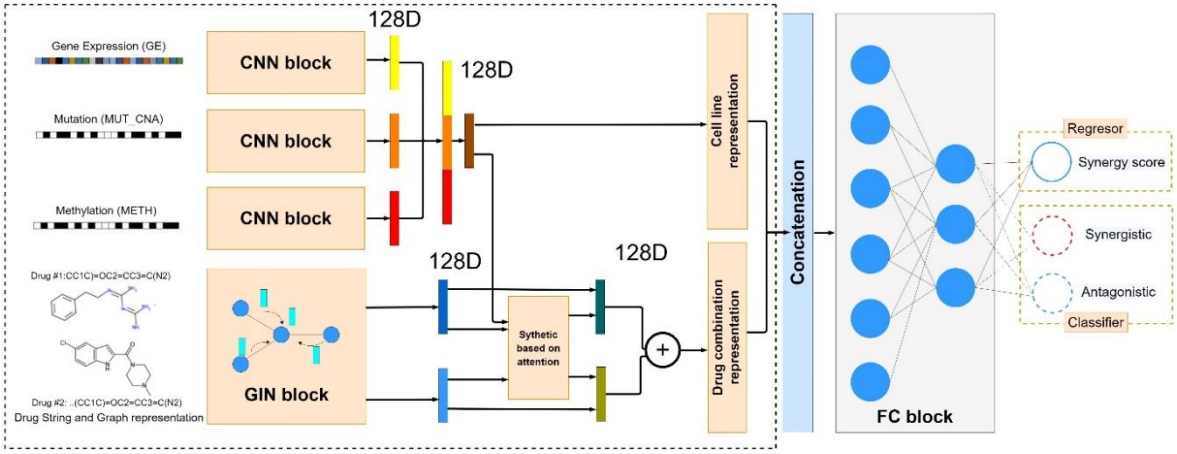
3.3. ĐỀ XUẤT GIẢI PHÁP HỌC BIỂU DIỄN ĐỒ THỊ CỦA ĐA PHÂN TỬ THUỐC VÀ TÍCH HỢP ĐA DỮ LIỆU -OMICS - GraOmicSynergy

3.3.1. Phương pháp

Kế thừa phương pháp tích hợp multi-omics và dữ liệu biểu diễn thuốc dưới dạng đồ thị từ nghiên cứu trước GraOmicDRP cho dự đoán liệu trình đơn thuốc, nghiên cứu này tiếp tục đề xuất phương pháp tích hợp này cho bài toán dự đoán kết hợp thuốc cho dòng tế bào với tên là GraOmicSynergy thực hiện việc kết hợp dữ liệu phân tử của một cặp thuốc và một hoặc nhiều dữ liệu –omic của các dòng tế bào để dự đoán điểm hiệu quả tổng hợp của thuốc. Hình 3.1 minh họa mô hình đề xuất.

Để học các đặc trưng của từng loại thuốc, nghiên cứu tiếp tục sử dụng phương thức mã hóa và biểu diễn dữ liệu từ Đề xuất 1 và Đề xuất 2 với mỗi cặp thuốc biểu diễn dưới dạng đồ thị phân tử từ định dạng chuỗi SMILES [39] sau đó đưa qua mô hình GIN để học các biểu diễn của các phân tử thuốc, kết hợp với module “Synthetic based-on attention” trình bày phần tiếp theo để tổng hợp thành vec-tơ biểu diễn đặc trưng cho cặp thuốc kết hợp và đưa vào mô hình dự đoán. Với khối GIN block, các lớp GIN được triển khai để học các biểu diễn đồ thị phân tử thuốc. Tại mỗi lớp, các

thuộc tính của nút được cập nhật bởi một mạng MLP gồm hai lớp tuyến tính, hàm kích hoạt ReLU và BatchNorm được sử dụng, theo sau là global add-pooling để tổng hợp các biểu diễn của một đồ thị phân tử thuốc. Cuối cùng, một lớp được kết nối đầy đủ làm phẳng kết quả thành vec-tơ ẩn 128 chiều biểu diễn cho mỗi thuốc.



Hình 3.1. Mô hình dự đoán đáp ứng đa thuốc - GraOmicSynergy

Khác với các nghiên cứu trước đây, thường chỉ xem các thuốc có vai trò ngang nhau trong điều trị bệnh thông qua việc biểu diễn cặp thuốc tác động lên dòng tế bào bệnh thì nghiên cứu này xem xét việc kết hợp thuốc dựa trên trên những giá trị đóng góp khác nhau của mỗi thuốc. Mô đun chức năng "Synthetic based on attention" được xây dựng lấy cảm hứng từ hiệu quả của phương pháp tính toán hệ số đóng góp chính của các từ trong một câu trong xử lý ngôn ngữ tự nhiên (NLP) [114], nghiên cứu xét ngữ cảnh kết hợp, coi các thuốc là các "từ" trong một "câu" – các cặp thuốc tác động lên một dòng tế bào. Bên cạnh đó, mô đun này cũng thực hiện việc tổng hợp đáp ứng đa thuốc dựa trên cơ chế kiểm thử kết hợp thuốc thực tế một cách tự nhiên. Cụ thể, theo thứ tự thuốc d_i ức chế dòng tế bào c_n sau một thời gian điều trị, bổ sung thuốc d_j và xác định giá trị đáp ứng của cặp thuốc. Do đó, cơ chế attention được đưa vào để tính toán các đóng góp khác nhau của mỗi thuốc trong cặp thuốc tác động cho dòng tế bào. Ngoài ra, do mạng nơ-ron phân biệt tổ hợp (A,B) theo cách biểu diễn thức tự khác nhau là khác nhau (ví dụ A-B khác B-A), do đó, nghiên cứu đã tính toán các giá trị attention của từng thuốc trên các cách biểu diễn kết hợp khác nhau của các cặp thuốc ($d_{i,}$, d_j) trên dòng tế bào $c_{n,}$. Các cặp thuốc ($d_{i,}$, d_j) tác động trên dòng tế bào c_n được tổng hợp qua phép tính concat (ví dụ: $(concat(d_i, c_n, d_{j)})$ sau đó biến

đổi tuyến tính thành vec-tơ 128 chiều, giá trị attention của mỗi bộ $(d_{i,n}, c_n, d_{j,n)})$ và $d_{j,n}, c_n, d_{i,n)})$ được tính theo công thức sau:

$$a_{i,n,j} = \exp\left(Leaky\_ReLU(Linear(concat(d_i, c_n, d_j))\right) \quad (3.1)$$

$$a_{j,n,i} = \exp\left(Leaky\_ReLU(Linear((concat(d_j, c_n, d_i))\right) \quad (3.2)$$

Giá trị $a_{i,n,j}$ được tính toán như là giá trị attention của thuốc $d_{i,}$ trong cặp thuốc $(d_{i,}, d_j)$ tác động trên dòng tế bào $c_{n,}$

Tương tự như vậy, giá trị $a_{j,n,i}$ được tính toán như là giá trị attention của thuốc $d_{j,}$ trong cặp thuốc $(d_{i,}, d_j)$ tác động trên dòng tế bào $c_{n,}$

Tiếp theo vec-tơ tổng hợp của cặp Vec-tơ biểu diễn cặp thuốc $(d_{i,}, d_j)$ tác động trên dòng tế bào $c_{n,}$ được tính dựa trên các giá trị attention của từng thuốc như sau:

$$\hat{y}\mathrm{i, j, n} = \frac{a_{i,n,j}}{a_{i,n,j} + a_{j,n,i}} * d_i + \frac{a_{j,n,i}}{a_{i,n,j} + a_{j,n,i}} * d_j \quad (3.3)$$

Mô hình học và biểu diễn dữ liệu dòng tế bào

Để học các đặc trưng của dòng tế bào, GraOmicSynergy tiếp tục kế thừa phương pháp tích hợp nhiều loại -omics khác nhau để trích xuất nhiều dữ liệu tiềm ẩn của dòng tế bào cho dự đoán thông qua các mạng nơ-ron tích chập 1 chiều. Tuy nhiên, các dữ liệu biểu diễn cho mỗi loại -omics của dòng tế bào này thay vì được tổng hợp thành một vec-tơ biểu diễn bằng phép ghép nối vec-tơ thông thường thì với GraOmicSynergy, vec-tơ tổng hợp này tiếp tục được biến đổi tuyến tính trở về thành một vec-tơ biểu diễn duy nhất là 128 chiều. Cụ thể hơn, một khối CNN sẽ bao gồm ba lớp tích chập với lớp tổng hợp tối đa (global-max-pooling) và hàm kích hoạt (ReLU). Sau đó, đầu ra được làm phẳng thành một vec-tơ 128 chiều của biểu diễn dòng tế bào. Các vec-tơ này được kết nối qua phép toán concat và chuyển đổi tuyến tính thành vec-tơ 128 chiều. Cuối cùng vec-tơ biểu diễn dữ liệu dòng tế bào này tiếp tục kết hợp với vec-tơ biểu diễn cặp thuốc điều trị cho dòng tế bào được kết nối tạo thành một vec-tơ duy nhất đưa vào khối FC block để dự đoán giá trị kết hợp thuốc cho dòng tế bào. Khối dự đoán FC block là một mạng MLP gồm hai lớp, lớp đầu vào là 1024 chiều, lớp thứ hai là 128 chiều dự đoán giá trị kết hợp thuốc cho mô hình.

Mô hình đánh giá

Mô hình được đánh giá hiệu năng của giải pháp tích hợp đa dữ liệu -omics và tích hợp đơn dữ liệu -omics. Ngoài ra, để so sánh với các mô hình dự đoán kết hợp thuốc tiên tiến hiện nay, nghiên cứu tiến hành so sánh với hai phương pháp gồm (1) DeepSynergy tích hợp dữ liệu GE và dữ liệu biểu diễn thuốc dưới dạng vec-tơ đặc trưng dạng fingerprint, (2) DeepDDS với cách biểu diễn đặc trưng thuốc dưới dạng đồ thị dữ liệu biểu hiện gen của dòng tế bào, trong đó nghiên cứu so sánh với mô hình dự đoán tốt nhất là DeepDDS(GAT).

GraOmicSynergy dự đoán đáp ứng đa thuốc (regressor)

Giải pháp đề xuất hoạt động như một mô hình hồi quy nhằm dư đoán giá trị kết hợp thuốc với đầu vào là các thuộc tính của dòng tế bào, các đặc trưng của thuốc và các giá trị kết hợp đã được quan sát, đầu ra là giá trị dự đoán cho khả năng kết hợp thuốc. Mô hình được đánh giá dựa trên các chỉ số sai số trung bình bình phương (RMSE) và hệ số tương quan Pearson (CCp).

GraOmicSynergy dự đoán phân loại kết hợp thuốc (classifier)

Mô hình GraOmicSynergy được thiết kế để có thể hoạt động như một mô hình phân lớp nhị phân, để dự đoán kết hợp hai thuốc có khả năng tương hợp (synergistic) hay tương kháng (antagonistic). Để làm việc này, mô hình nghiên cứu sử dụng toàn bộ các khối và các lớp trong quá trình huấn luyện, trong khối dự đoán (FC block) lớp phân loại đầu ra là một hàm biến đổi tuyến tính trung bình mũ (softmax).

Khi hoạt động như một mô hình phân lớp, GraOmicSynergy không thiết lập ngưỡng giá trị synergy như trong mô hình DeepDDS và DeepSynergy. Cụ thể, DeepDDS chỉ coi các giá trị nhãn synergy lớn hơn 10 được gắn nhãn là dương (1 – tương hợp) và điểm nhỏ hơn 0 được gắn nhãn là không (0 – tương kháng). Trong khi đó DeepSynergy sử dụng các synergy lớn hơn 30 làm ngưỡng (chiếm 10%). Điều này có thể làm mất đi một lượng lớn các dữ liệu trong quá trình huấn luyện và dự đoán ở cả DeepSynergy và DeepDDS. Do vậy để tăng cường học các đặc trưng của tất cả các mẫu, GraOmicSynergy lấy toàn bộ dữ liệu với ngưỡng giá trị đáp ứng đa thuốc là 0 như trong các quy ước phân loại đáp ứng đa thuốc của [115]. Để đánh giá khả năng dự đoán của mô hình mô hình thực hiện đánh giá độ chính xác của mô hình

bằng cách sử dụng các chỉ số đánh giá như: độ chính xác (ACC), Precision (PREC), Recall và F1-score (F1). Các chỉ số này được thực hiện trên tất cả các kịch bản thử nghiệm.

3.3.2. Cài đặt và thử nghiệm mô hình

Tổng hợp bộ dữ liệu

Bộ dữ liệu cho cho nghiên cứu này bao gồm các tập dữ liệu về kết hợp thuốc cho dòng tế bào, tập dữ liệu về multi-omics cho dòng tế bào; tập dữ liệu biểu diễn thuốc. Các bộ dữ liệu này được thu thập từ nguồn công khai và kế thừa từ các nghiên cứu trước đó. Cụ thể:

- Dữ liệu quan sát tương tác kết hợp thuốc và dòng tế bào: Tương tự như DeepSynergy, nghiên cứu này lựa chọn áp dụng bộ dữ liệu sàng lọc kết hợp thuốc-thông lượng cao cho ung thư do O'Neil [75] cung cấp. Bộ dữ liệu này bao gồm 583 cách kết hợp thuốc theo cặp của 38 loại thuốc riêng biệt, mỗi loại được thử nghiệm trên 39 dòng tế bào bao gồm 7 loại ung thư, có 23.062 mẫu tương tác cặp thuốc và dòng tế bào. Mỗi mẫu quan sát được đo để ước tính hiệu lực kết hợp sau 48 giờ xử lý thuốc theo 4 liều cho mỗi loại thuốc. Kế thừa các giá trị kết hợp thuốc được tính theo chỉ số Loewe như trong nghiên cứu của DeepSynergy[2].

- Bộ dữ liệu biểu diễn hồ sơ sinh học multi-omics cho dòng tế bào bệnh là bộ dữ liệu GDSC [45]. Bộ dữ liệu này được kế thừa và chuẩn hóa theo Đề xuất 2 – GraOmicDRP bao gồm:

  - Bộ dữ liệu dữ liệu đột biến gen và số lượng biến đổi bản sao (MUT_CNA) gồm 990 dòng tế bào, mỗi dòng tế bào gồm 735 đặc trưng biểu hiện đột biến, dữ liệu nhị phân [0,1], cho biết gen có bị đột biến hay không.
  - Bộ dữ liệu biểu hiện gen (GE) của quá trình phiên mã (transcriptomic) gồm: 1018 dòng tế bào, mỗi dòng tế bào gồm 17.737 đặc trưng, dữ liệu biểu hiện gen biểu diễn dạng số thực

[2] https://www.bioinf.jku.at/software/DeepSynergy/

- Bộ dữ liệu biểu hiện methyl hóa (METH) của di truyền biểu gen (epigennomic), dữ liệu dạng nhị phân dạng [0,1] cho biến gen có bị methyl hóa hay không.

- Dữ liệu biểu diễn thuốc: Mô hình hóa dữ liệu biểu diễn thuốc dưới dạng đồ thị cũng được kế thừa theo phương pháp của mô hình GraOmicDRP và GraphDRP.

Tiền xử lý dữ liệu

- Tương tự như Đề xuất 2 – mô hình GraOmicDRP, dữ liệu biểu hiện gen (GE) được chuẩn hóa về khoảng (0,1) theo phương pháp min-max normalization

- Thực hiện tổng hợp các bộ dữ liệu nghiên cứu cho dòng tế bào (GDSC), thuốc (Drugbank) và dữ liệu kết hợp thuốc (O'Neil), bộ dữ liệu cho các thực nghiệm trong nghiên cứu này gồm:

- 38 loại thuốc trong đó 37 thuốc đại diện cho mỗi thuốc trong cặp kết hợp thuốc.
- 25 dòng tế bào thuộc năm loại mô bệnh (tissues) gồm: ung thư vú (breast cancer), ung thư phổi (lung cancer), ung thư da (skin cancer), ung thư hệ niệu đạo (urogenital_system) và ung thư hệ tiêu hóa (digestive_system).
- 14.722 mẫu quan sát được cho các cặp kết hợp thuốc điều trị cho dòng tế bào.

Chia bộ dữ liệu thử nghiệm

Nghiên cứu đã đánh giá tổng thể hiệu năng của mô hình dự đoán của GraOmicSynergy và so sánh với các phương pháp tiên tiến gần đây theo 03 kịch bản thử nghiệm bao gồm: Mixed, Blind-Drugpair, Blind-Cellline. Hiệu năng dự đoán của GraOmicSynergy cũng được đánh giá cho cả dữ liệu –omics đơn và kết hợp đa -omics. Với Mixed, thuốc và dòng tế bào trong giai đoạn huấn luyện có thể xuất hiện trong giai đoạn thử nghiệm; thử nghiệm Blind-Cellline được xây dựng để đưa ra dự đoán cho một dòng tế bào mới không có ở giai đoạn huấn luyện; tương tự, Blind-DrugPair được tạo ra để dự đoán đáp ứng của cặp thuốc chưa biết.

Ngoài ra, để so sánh công bằng trong đánh giá giữa các kịch bản cũng như so sánh hiệu năng với các phương pháp hiện có khác, nghiên cứu đã sử dụng một bộ huấn luyện duy nhất cho tất cả các kịch bản để học các đặc trưng ẩn của các dòng tế

bào và thuốc, trong khi các bộ dữ liệu cho việc đánh giá, điều chỉnh tham số mô hình (tập đánh giá và tập kiểm tra) của mỗi kịch bản riêng được chia độc lập, không trùng lặp nhau. Cụ thể, các bộ dữ liệu này được thực hiện bằng cách xáo trộn dữ liệu một cách ngẫu nhiên, sau đó chia dữ liệu vào các bộ dữ liệu cho huấn luyện (training set), đánh giá (validation set) và kiểm tra (testing set) dựa trên sự tương đồng về giá trị trung bình (mean) và độ lệch chuẩn (standard deviation) của các giá trị kết hợp thuốc (synergy scores) của bộ dữ liệu ban đầu. Việc chia này cũng đảm bảo tính tổng quát của mô hình đánh giá và tránh overfitting. Các tiêu chí lựa chọn này được áp dụng cho tất cả các kịch bản thử nghiệm.

Để thực hiện việc này, trước tiên, nghiên cứu thực hiện việc chọn danh sách các dòng tế bào và danh sách các danh sách thuốc chưa được thử nghiệm cho kịch bản thử nghiệm Blind-Cellline và Blind-DrugPair.

Đối với Blind-Cellline, bằng cách sử dụng hàm train-split-test trong thư viện scikit-learn của Python nghiên cứu phân tách các tập dữ liệu tập dữ liệu trên thành các tập huấn luyện, kiểm tra và đánh giá với tỷ lệ xấp xỉ 80:10:10, các tập dữ liệu này cũng đươc chia đảm bảo có sự tương đồng về phân phối giữa các tập dữ liệu thử nghiệm thông qua kiểm định thống kê Kolmogorov-Smirnov [41]. Các dòng tế bào được lựa chọn cho dòng thử nghiệm Blind-Cell được chia không chỉ dựa trên sự phân bố tương đồng với các bộ khác mà còn dựa trên số lượng dòng tế bào của mỗi loại mô bệnh (theo năm loại tissue). Theo đó, nghiên cứu đã chọn 8/25 dòng tế bào cho thử nghiệm Blind-Cellline. Nghiên cứu thực hiện tương tự với việc lựa chọn các thuốc cho kịch bản Blind-DrugPair, kết quả nghiên cứu chọn được 6/38 loại thuốc cho Blind-DrugPair. Tập dữ liệu còn lại gồm 17 dòng tế bào của năm loại mô và 32 loại thuốc lúc này được tiến hành chia thành 01 tập dữ liệu huấn luyện chung duy nhất, với tỉ lệ 80%; 10% dành cho tập dữ liệu đánh giá và kiểm tra của kịch bản Mixed. Phần còn lại tiếp tục chia cho kịch bản Blind-Cellline và Blind-DrugPair. Chi tiết hơn, bộ dữ liệu đánh giá và thử nghiệm cho kịch bản Blind-Cellline được chia 50:50 từ các mẫu có thuốc thuộc danh sách tập huấn luyện và dòng tế bào thuộc danh sách Blind-Cellline. Tương tự với kịch bản Blind-DrugPair, bộ dữ liệu kiểm tra và đánh giá được chia theo tỉ lệ 50:50 cho các mẫu có cả hai thuốc thuộc danh sách Blind-

Drug và tất cả dòng tế bào. Chi tiết các bộ dữ liệu theo các kịch bản khác nhau được tóm tắt trong Bảng 3.1.

Bảng 3.1. Phân chia bộ dữ liệu thử nghiệm cho các kịch bản đánh giá

	# Drug 1	# Drug 2	# Cell lines	# Train	# Val	# Test
Mixed	31	31	17	5756	720	720
Blind-Cellline	31	31	8	5756	1679	1680
Blind-DrugPair	6	6	17	5756	101	101

Để đảm bảo rằng thứ tự của tổ hợp thuốc AB (dù được biểu diễn dưới dạng A-B hay B-A) không ảnh hưởng đến dự đoán của mô hình, nghiên cứu đã kết hợp cả hai cách trình bày của từng mẫu trong tất cả các dữ liệu cho huấn luyện, đánh giá và kiểm tra. Để làm việc đó, nghiên cứu thực hiện lật các thuốc trên từng cặp thuốc tương tác điều trị cho dòng tế bào sau đó tính trung bình các kết quả để dự đoán để đo hiệu năng chung của mô hình.

Cài đặt mô hình

Tương tự như trong Đề xuất 2, nghiên cứu thực hiện các khối cho mô hình huấn luyện biểu diễn thuốc và dòng tế bào như sau:

- Với thuốc, khối GIN gồm năm lớp GIN liên tiếp được sử dụng để học các biểu diễn thuốc tiếp theo sau lớp tổng hợp thông tin đồ thị thuốc (global-add pooling), một lớp FC để biến đổi tuyến tính biểu diễn thuốc về vec-tơ 128 chiều.

- Với biểu diễn dòng tế bào, nghiên cứu triển khai các khối CNN1D khác nhau tương ứng mỗi dữ liệu dòng tế bào.

    - Mỗi khối bao gồm ba lớp tích chập, tiếp theo sau là lớp tổng hợp các đặc trưng dòng tế bào (max pooling) và lớp FC để biến đổi tuyến tính mỗi biểu diễn dòng tế bào đó thành vec-tơ 128 chiều, được cài đặt để học các đặc trưng ẩn của các dữ liệu -omics khác nhau.
    - Mỗi kịch bản tích hợp (single-omics, multi-omics: 2-omics hoặc 3-omics) đầu ra mỗi khối sẽ được kết hợp thành vec-tơ 128/256/384 chiều tương ứng; vec-tơ này sẽ được biến đổi tuyến tính thành vec-tơ 128 chiều biểu diễn thông tin tổng hợp của dòng tế bào.

- Khối FC block thực hiện nhiệm vụ dự kết quá trình huấn luyện gồm hai lớp đầy đủ (256/128 nút) đầu ra là một nút hoặc hai nút tương ứng với mô hình hoạt động

như tác vụ là bộ dự đoán hồi quy (regressor) hoặc khi mô hình hoạt động như bộ dự đoán phân lớp (classifier).

Cài đặt môi trường và siêu tham số huấn luyện

- Learning: 0.001
- Batch size: Tinh chỉnh trong quá trình huấn luyện (1024)
- Số epoch: Tinh chỉnh trong quá trình huấn luyện (300)
- Thuật toán học: Tối ưu hóa Adam (Adam Optimization)

Mô hình được huấn luyện và đánh giá trên server với cấu hình:

- Colab Pro

Các mô hình, thuật toán được triển khai bằng ngôn ngữ lập trình Python cùng với các thư viện:

- Pytorch: hỗ trợ triển khai các mô hình deep learning
- Rdkit: hỗ trợ các thao tác phép biến đổi với thuốc
- Pandas: hỗ trợ xử lý dữ liệu dạng .csv
- Matplotlib: hỗ trợ trực quan hóa các thông tin

3.3.3. Kết quả và đánh giá

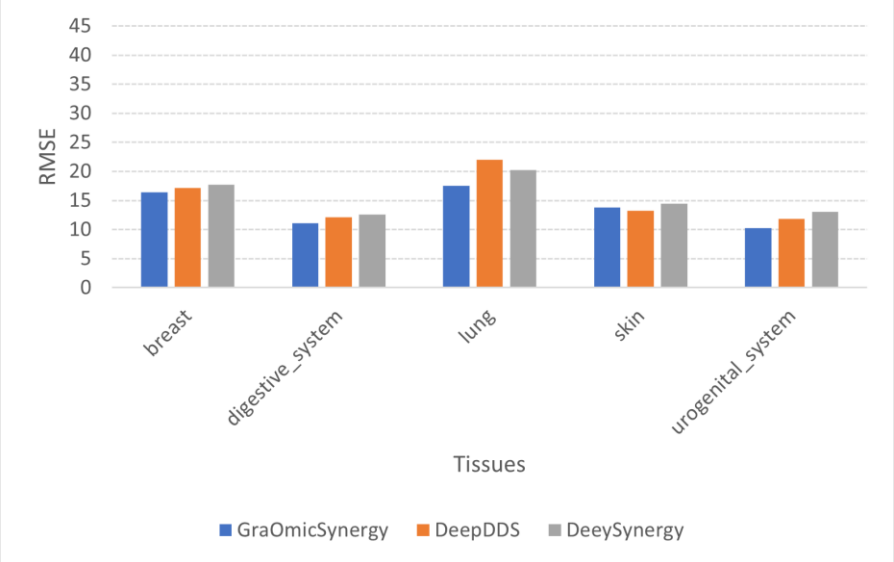
Kịch bản Mixed

Đánh giá hiệu năng của GraOmicSynergy trên kịch bản Mixed cho việc tích hợp cả single-omics và multi-omics cho thấy hầu hết tích hợp đa dữ liệu -omics cho hiệu năng tốt hơn các kịch bản đơn -omics. Trong đó, tích hợp cả ba -omics tổ hợp (ALL: GE & MUT_CNA & METH) đạt hiệu quả tốt nhất về cả RMSE (14,025) và CCp (0,794) (Bảng 3.2) so với tất cả các kịch bản tích hợp cũng như vượt trội hơn so với DeepDDS và DeepSynergy, giá trị RMSE và CCp của DeepDDS lần lượt là 15,871 và 0,744 và giá trị của DeepSynergy là 15,909 và 0,711. Ngoài ra, hầu hết các single-omics và sự kết hợp của phương pháp của nghiên cứu đều vượt trội so với DeepSynergy và DeepDDS ngoại trừ sự kết hợp của MUT & METH. Bên cạnh đó, dữ liệu biểu hiện gen khi kết hợp với các dữ liệu -omics khác cho hiệu năng dự đoán tốt hơn khi so với cặp kết hợp còn lại (MUT_CNA & METH), tương tự như giải pháp tích hợp dữ liệu đa -omics trong dự đoán đáp ứng đơn thuốc (đề xuất GraOmicDRP).

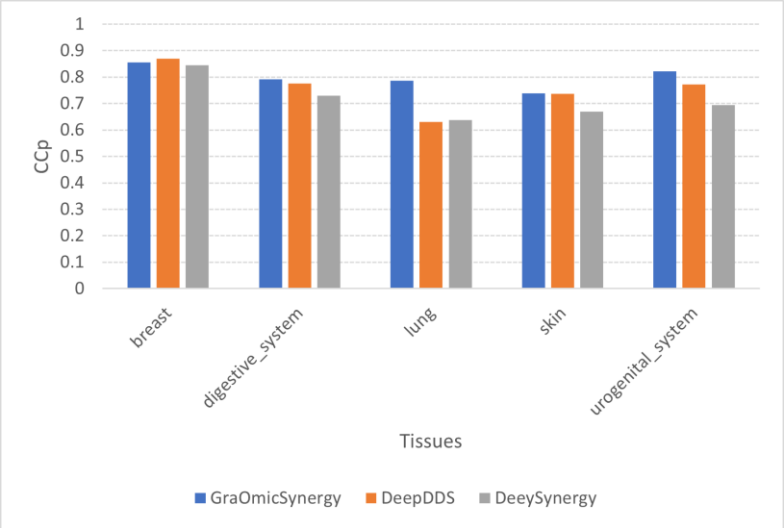
Bảng 3.2. So sánh hiệu năng các phương pháp theo kịch bản Mixed

Methods		RMSE	CCp
DeepDDS(GAT)		15.871	0.744
DeepSynergy		15.909	0.711
GraOmicSynergy (Single-omics)	GE	15.099	0.755
	MUT_CNA	14.971	0.760
	METH	14.546	0.775
GraOmicSynergy (Multi-omics)	GE & MUT_CNA	14.245	0.785
	GE & METH	14.387	0.782
	METH & MUT_CNA	16.401	0.706
	ALL	**14.025**	**0.794**

Tiến hành so sánh hiệu năng dự đoán các phương pháp trên các mô bệnh tổng hợp được (năm mô bệnh) theo chỉ số RMSE và CCp trên kịch bản Mixed. Các bộ dữ liệu thử nghiệm được chia đồng đều theo các mô bệnh nên việc so sánh này cũng đảm bảo công bằng giữa các phương pháp thử nghiệm, kết quả cho thấy GraOmicSynergy có thể dự đoán tốt hơn trong tất cả các loại bệnh. Cụ thể, so với DeepSynergy và DeepDDS, GraOmicSynergy (ALL) đã đạt được RMSE tốt nhất cho tất cả các mô (Hình 3.2), trong khi đó, đạt được CCp tốt nhất cho 4/5 mô bệnh (bao gồm hệ thống tiêu hóa, phổi, da và hệ thống niệu sinh dục (Hình 3.3).



Hình 3.2. So sánh hiệu năng các phương pháp dự đoán các mô bệnh trên đánh giá RMSE theo kịch bản Mixed



Hình 3.3. So sánh hiệu năng các phương pháp dự đoán các mô bệnh trên đánh giá CCp theo kịch bản Mixed

Kịch bản Blind-Cellline

So sánh hiệu năng về cả CCp và RMSE đối với thử nghiệm Blind-Cellline, thấy rằng phương pháp đề xuất tốt hơn Deepsynergy và DeepDD trong hầu hết các single-omics và nhiều omic. Đặc biệt, với bộ dữ liệu -omics – GE – đạt hiệu năng tốt nhất trong cả chỉ số của CCp và RMSE. Bảng 3.3 cho thấy GraOmicDRP RMSE nhỏ hơn (20,73) và CCP cao hơn (0,484) so với DeepDDS và Deepsynergy.

Chi tiết hơn cho việc so sánh hiệu năng các mô hình, nghiên cứu tiến hành đánh giá hiệu quả dự đoán trên từng dòng tế bào mới (Blind-Cellline). Trong kịch bản này, GraOmicSynergy(ALL) đã đạt RMSE tốt nhất cho 5/8 dòng tế bào được dự đoán (gồm MDAMB436, NCIH1650, RKO, OVCAR3, SKMES1) và CCp tốt nhất cho 4/8 dòng tế bào được dự đoán (MDAMB436, NCIH1650, OVCAR3, SKMES1). Tuy nhiên, nghiên cứu nhận thấy rằng hiệu năng dự đoán kết thuốc cho cả 3 phương pháp đối với các dòng tế bào LNCAP, SKMEL30 và OVCAR3 trên đánh giá CCp còn thấp hơn so với năm dòng tế bào còn lại, đặc biệt là đối với LNCAP. Điều này có thể lý giải bởi sự biến thiên khá lớn các dữ liệu quan sát được của các dòng tế bào này. Ví dụ: các giá trị đáp ứng (synergy) quan sát đối với LNCAP (std = 39.227) và SKMEL30(std = 25.176), OVCAR3 (std = 27.576) nằm trong khoảng biến thiên lớn (lần lượt là -169 ÷ 106); (-172.89 ÷ 121.05) và (-46.67 ÷106.63). Do đó độ tương quan Pearson giữa các giá trị dự đoán và quan sát không cao.

Bảng 3.3. So sánh hiệu năng các phương pháp cho dự đoán dòng tế bào mới

Methods		RMSE	CCp
DeepDDS(GAT)		21.893	0.415
DeepSynergy		20.930	0.463
GraOmicSynergy (Single-omics)	GE	20.730	0.484
	MUT_CNA	21.468	0.442
	METH	21.685	0.460
GraOmicSynergy (Multi-omics)	GE & MUT_CNA	21.095	0.469
	GE & METH	**20.458**	**0.512**
	METH & MUT_CNA	20.942	0.468
	ALL	20.742	0.498

Kịch bản Blind-DrugPair

Bảng 3.4 cho thấy phương pháp đề xuất vượt trội về RMSE (19,968) và CCp (0,37) cho DeepSynergy (20,884, 0,131) cũng như DeepDDS (22,745, 0,105) tương ứng. Các mô hình -omics tích hợp khác cũng có thể so sánh với các mô hình của DeepDDS. Kết hợp lại với nhau, kết quả thử nghiệm cho thấy GraOmicSynergy tốt hơn DeepDDS và DeepSynergy cho tất cả các kịch bản thử nghiệm. Điều này cũng cho thấy tầm quan trọng của việc tích hợp nhiều dữ liệu -omics, đặc biệt là dữ liệu biểu hiện gen (GE), đối với các bài toán dự đoán đáp ứng thuốc.

Bảng 3.4. So sánh hiệu năng các phương pháp cho dự đoán cặp thuốc mới

Methods		RMSE	CCp
DeepDDS(GAT)		22.745	0.105
DeepSynergy		20.884	0.131
GraOmicSynergy (Single-omics)	GE	20.620	0.316
	MUT_CNA	21.537	0.158
	METH	20.177	0.378
GraOmicSynergy (Multi-omics)	GE & MUT_CNA	20.927	0.263
	GE & METH	22.036	0.099
	METH & MUT_CNA	20.164	0.384
	ALL	**19.968**	**0.379**

Có thể thấy rằng các kết quả thử nghiệm khi dự đoán giá trị kết hợp các cặp thuốc cho dòng tế bào, cho thấy GraOmicSynergy vượt trội so với DeepDDS và DeepSynergy trong tất cả các kịch bản thử nghiệm, đồng thời cho thấy việc tích hợp nhiều -omics mang lại hiệu quả dự đoán trong các nhiệm vụ dự đoán kết hợp thuốc.

Kết quả thử nghiệm khi mô hình GraOmicSynergy hoạt động như một bộ phân loại

Giống như cách thức đánh giá việc phân loại khả năng kết hợp thuốc của DeepSynergy và DeepDDS, mô hình GraOmicSynergy chuyển đổi trực tiếp kết quả dự đoán giá trị đáp ứng đa thuốc (dạng số thực) thành giá trị phân loại (1,0). Ma trận đánh giá cho mô hình thực hiện trên việc phân loại của bộ dữ liệu thử nghiệm cho toàn bộ các kịch bản. Bảng 3.5 thể hiện kết quả của việc phân loại kết quả dự đoán của kịch bản Mixed đối với bộ dữ liệu kết hợp tất cả các -omics (GE, MUT&CNA và METH) của GraOmicSynergy với DeepDDS và DeepSynergy trên các phép đo Accuracy, Precision, Recall, F1-score. Kết quả cho thấy mô hình đề xuất đạt độ chính xác cao hơn hai mô hình so sánh.

Bảng 3.5. So sánh hiệu năng các phương pháp khi hoạt động như mô hình phân loại trên các kịch bản thử nghiệm

	Methods	ACC	PREC	Recall	F1-score
Mixed	DeepDDS (GAT)	0.758	0.806	0.799	0.802
	DeepSynergy	0.759	**0.816**	0.784	0.8
	GraOmicSynergy (ALL)	**0.783**	0.803	**0.857**	**0.829**
Blind-cellline	DeepDDS	0.712	**0.78**	0.728	0.753
	DeepSynergy	0.691	0.731	0.773	0.751
	GraOmicSynergy (ALL)	**0.717**	0.741	**0.816**	**0.777**
Blind-DrugPair	DeepDDS (GAT)	0.603	0.651	0.7	0.675
	DeepSynergy	0.611	0.647	**0.75**	0.695
	GraOmicSynergy (ALL)	**0.627**	**0.677**	0.7	**0.689**

Các kết quả trên cho thấy mô hình đề xuất trong nghiên cứu thể hiện vượt trội hơn các mô hình so sánh khi hoạt động như mô hình hồi quy cũng như mô hình phân loại khả năng kết hợp thuốc cho dòng tế bào.

Nghiên cứu này cũng đã khảo sát các minh chứng khoa học và liệt kê các công bố trước đây (số PMID/doi) cho thấy khả năng kết hợp thuốc trong điều trị bệnh ung thư tương ứng Bảng 3.6 cho thấy10 cặp kết hợp thuốc được dự đoán tốt nhất cho kịch bản Mixed với điểm lỗi thấp nhất. Các nghiên cứu trước đây cũng cho thấy khả năng kết hợp các cặp thuốc này trong điều trị ung thư. Ví dụ như MK-4827 (Niraparib) và BEZ-235 (Dactolisib) là hai loại thuốc đã được nghiên cứu về tiềm năng sử dụng trong điều trị ung thư. Niraparib là chất ức chế PARP và Dactolisib là chất ức chế kép con đường sinh học PI3K và mTOR, có liên quan đến sự phát triển, sống sót và chuyển hóa của tế bào. Sự ức chế kết hợp PI3K và PARP đã được chứng minh là có

hiệu quả trong điều trị các mô hình ung thư vú tiền lâm sàng [116]. MK-8776 là chất ức chế chọn lọc kinase 1 (Chk1), có thể kết hợp với chất ức chế PARP và có thể đạt được chiến lược điều trị hiệu quả hơn trong ung thư dạ dày [117]. Tương tự, chất ức chế MAP p38 có trong L778123 có thể kết hợp với chất ức chế PI3K và mTOR (BEZ-235) để điều trị tế bào ung thư [118].

Bảng 3.6. Mười kết quả dự đoán tốt nhất và bằng chứng sinh học

Drug1	Drug2	Cell line	Tissue	abs_error	Publications
LAPATINIB	BORTEZOMIB	RKO	digestive_system	0.000115156	PMID: 20701607
SUNITINIB	5-FU	SKMEL30	skin	0.002768755	PMC: 3392575
DASATINIB	VINORELBINE	SKMES1	lung	0.004963875	PMC: 5784669
GELDANAMYC	BORTEZOMIB	RPMI7951	skin	0.006840706	PMID: 15141013
MK-2206	DOXORUBICIN	SKMEL30	skin	0.007393837	PMID: 27499633
MK-5108	5-FU	HT144	skin	0.011643052	PMID: 27499633
MK-2206	SUNITINIB	A2780	urogenital_system	0.022981644	PMID: 32927828
MK-5108	CARBOPLATIN	VCAP	urogenital_system	0.02829051	https://doi.org/10.1186/s12943-020-01305-3
SORAFENIB	5-FU	OCUBM	breast	0.031475782	PMID: 22033636
DASATINIB	PACLITAXEL	HCT116	digestive_system	0.07894993	PMID: 30866697

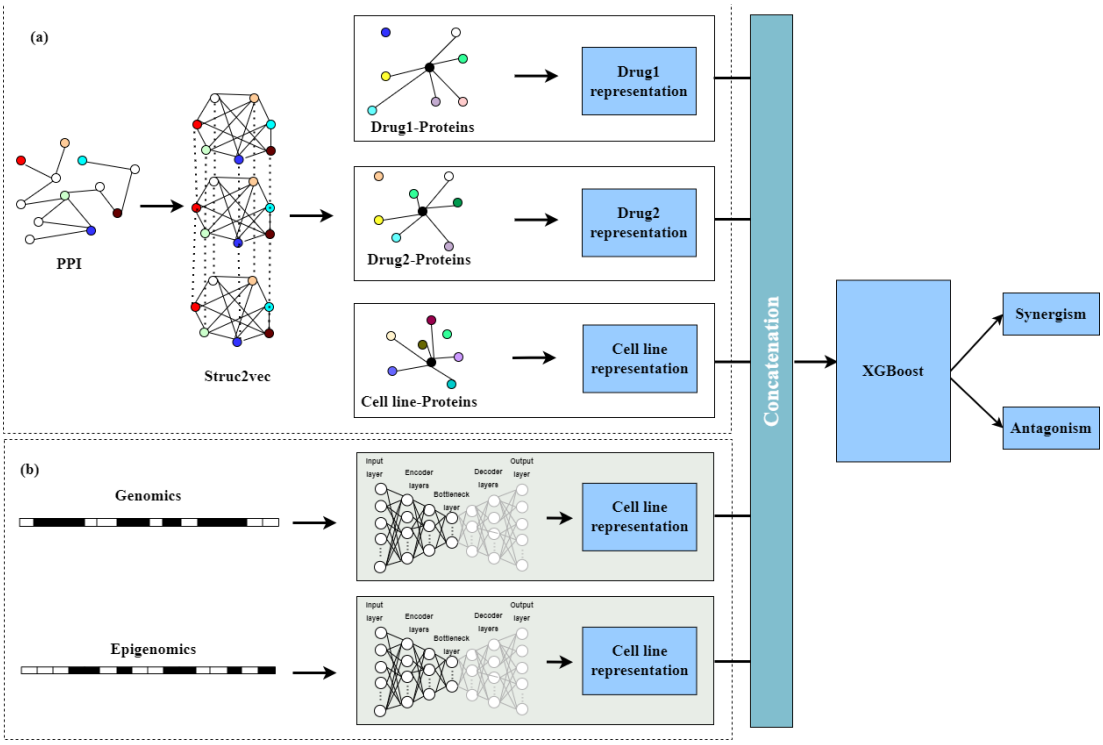
Các thực nghiệm trên cho thấy việc tích hợp đa dữ liệu -omics có hiệu quả vượt trội so với tích hợp đơn dữ liệu -omics đồng thời mô hình GraOmicSynergy hoạt động hiệu quả hơn với các mô hình tiên tiến khác như DeepDDS và DeepSynergy trong hầu hết các kịch bản dự đoán độ kết hợp thuốc. Mô hình học đặc trưng biểu diễn đồ thị phân tử thuốc qua GIN và tổng hợp cặp thuốc tương tác theo cơ chế chú ý mang lại hiệu năng vượt trội hơn so với mô hình tổng hợp thuốc thông qua GAT trong DeepDDS cũng như mô hình không áp dụng biểu diễn đồ thị phân tử thuốc DeepSynergy. Một số bằng chứng y học về kết hợp thuốc cũng tương đồng với độ chính xác cao nhất trong dự đoán của mô hình đề xuất.

3.4. ĐỀ XUẤT GIẢI PHÁP TÍCH HỢP ĐA DỮ LIỆU -OMICS VÀ THÔNG TIN MẠNG SINH HỌC - AE-XGBSynergy

3.4.1. Phương pháp

Các liệu trình kết hợp thuốc đã nổi lên như một phương pháp đầy hứa hẹn trong y học chính xác nhờ khả năng nâng cao hiệu quả điều trị và chống lại tình trạng kháng thuốc. Tuy nhiên, các phương pháp dự đoán kết hợp thuốc hiện tại thiếu cơ chế trích xuất hiệu quả thông tin có ý nghĩa từ đa dữ liệu -omics của các dòng tế bào

và tích hợp nó với thông tin mạng tương tác protein (PPI). Trong nghiên cứu này, luận án trình bày một đề xuất mới gọi là AE-XGBSynergy, nhằm giải quyết hạn chế này của các phương pháp hiện tại bằng cách tích hợp đa dữ liệu -omics từ các dòng tế bào với dữ liệu biểu diễn đặc trưng của dòng tế bào và thuốc được trích xuất từ mạng PPI từ đó dự đoán sự kết hợp thuốc đối với các dòng tế bào. Mô hình AE-XGBSynergy được biểu diễn như trong Hình 3.4.



Hình 3.4. Mô hình đề xuất dự đoán đáp ứng đa thuốc - AE-XGBSynergy

Mô hình AE-XGBSynergy gồm hai phần, trong đó phần đầu tiên (a) để trích xuất đặc trưng của thuốc và dòng tế bào từ mạng tương tác PPI. Cụ thể, việc nhúng mạng PPI được cấu trúc bằng cách sử dụng thuật toán struc2vec. Trong đó, mỗi protein phản ánh dưới dạng một nút và sau đó được nhúng vào không gian nhúng mà vẫn giữ nguyên cấu trúc đồ thị. Hơn nữa, thuật toán struc2vec sử dụng mô hình đồ thị đa tầng để trích xuất đặc điểm tương đồng về cấu trúc và tạo ra bối cảnh cấu trúc cho các nút. Bằng cách đó, nó đảm bảo rằng các cặp gen ở xa nhau nhưng có cấu trúc tương tự nhau được thể hiện chặt chẽ trong mạng lưới tương tác protein-protein. Các biểu diễn đặc trưng dòng tế bào được trích xuất thông qua một mô hình mã hóa pre-trained đã được học các đặc trưng ẩn của các dòng tế bào. Các biểu diễn của các

cặp thuốc và dòng tế bào này được tạo thành vec-tơ đầu vào cho cho mô hình dự đoán sự kết hợp thuốc cho các dòng tế bào ung thư.

Trích xuất đặc trưng của thuốc và dòng tế bào trong mạng PPI

- Mạng PPI là một đồ thị vô hướng đóng vai trò là một đầu vào của mô hình với các protein là các nút và các cạnh là tương tác giữa hai nút. Để tìm hiểu đặc điểm cấu trúc của mạng PPI, nghiên cứu đã sử dụng phương pháp struc2vec. Khác với các phương pháp khác struc2vec mã hóa các điểm tương đồng về cấu trúc bằng cách xây dựng các đồ thị nhiều lớp và tạo ngữ cảnh cấu trúc cho các nút. Các cặp protein cách xa nhau nhưng có cấu trúc tương tự nhau vẫn được suy xét trong thuật toán struc2vec.

- Trong mạng PPI, mỗi đỉnh sẽ được xem xét với đồ thị riêng với các nút hàng xóm quanh nó. Xét quá trình duyệt đường đi từ đỉnh đó đến các nút hàng xóm thì thu được các danh sách chứa tất cả các hành trình được tạo ra của đỉnh đó với các thông tin như: danh sách các nút trong một hành trình và các thông tin liên quan đến hành trình đó.

- Quá trình trích xuất đặc trưng của mạng PPI (P) gồm các bước:

(1) *Tính sự tương đồng về cấu trúc giữa cặp nút đối với các vùng lân cận* k.

- Xét một cặp nút (protein) "u" và "v", $R_k(u)$ xác định là tập các nút ở bán kính k từ "u", $s(R_k(u))$ là tập các bậc của các nút trong $R_k(u)$; G(s(Rk(u)), s(R_k (v))) đo khoảng cách của các chuỗi có thứ tự, trong nghiên cứu này là tập bậc các đỉnh của các nút ở bán kính k của nút "u" và "v". Từ đó các cấu trúc tương đồng giữa u và v xét trong trong vùng lân cận k-hop được xác định bằng khoảng cách cấu trúc và được tính theo công thức:

$$f_k(u,v) = f_{k-1}(u,v) + g\left(s(R_k(u)), s(R_k(v))\right) \quad (3.4)$$

(2) *Xây dựng đồ thị có trọng số nhiều lớp.*

- Sau khi có khoảng cách cấu trúc của mỗi cặp protein ở từng lớp k trong mạng PPI. Thuật toán struc2vec sẽ tính toán trọng số cho mỗi cạnh liên kết cặp đỉnh đó trên nhiều tầng.

- Để xây dựng được đồ thị đa tầng (đồ thị ở mỗi lớp gồm các cạnh vô hướng có trọng số được tạo ra từ các cặp protein theo bán kính k) thì trọng số cạnh giữa mỗi cặp protein (u,v) được xác định bằng công thức:

$$W_{\mathrm{k}}(u,v) = e^{-f_{\mathrm{k}}(u,v)} \quad (3.5)$$

- Mỗi protein u trong lớp k được kết nối tương ứng với protein v trong lớp k+1 và k-1, trọng số cạnh giữa các lớp được tính theo công thức:

$$W_{\mathrm{k}}(u_{\mathrm{k}}, u_{\mathrm{k}}+1) = \log(\Gamma_{\mathrm{k}}(u) + \mathrm{e}) \quad (3.6)$$

- Trong đó $\Gamma_{\mathrm{k}}(u)$ là số cạnh của lớp k mà có trọng số lớn hơn trọng số trung bình của các cạnh mà protein u tương tác với protein khác trong lớp k của đồ thị này.
- Nếu giá trị $\Gamma_{\mathrm{k}}(u)$ tại lớp k càng lớn thì cho thấy ở lớp k có nhiều nút tương tự với protein "u" thì nó sẽ có xu hướng để đi lên lớp cao hơn để có được ngữ cảnh tinh tế hơn.

(3) *Tạo bối cảnh cho các nút thông qua random walk*

- Mỗi protein sẽ thực hiện một bước đi ngẫu nhiên (random walk) trong đồ thị đa lớp P để tạo các chuỗi các nút nhằm xác định ngữ cảnh của protein đó, mỗi protein thường bắt đầu ở lớp 0. Với xác suất mà nút protein tiến hành random walk ở lớp hiện tại là p, $Z_{\mathrm{k}}(u)$ là hệ số chuẩn hóa đỉnh u trong lớp k thì xác suất để chọn đến một nút "v" bất kỳ ở lớp này là:

$$p_{\mathrm{k}}(u,v) = \frac{e^{-f_{\mathrm{k}}(u,v)}}{Z_{\mathrm{k}}(u)} \quad (3.7)$$

- Xác suất để nút protein đó di chuyển sang các tầng khác của đồ thị đa lớp là (1-q) nhưng xác suất để đi lên lớp trên và đi xuống lớp dưới là khác nhau:

$$p_{\mathrm{k}}(u_{\mathrm{k}}, u_{\mathrm{k}+1}) = \frac{W(u_{\mathrm{k}}, u_{\mathrm{k}+1})}{W(u_{\mathrm{k}}, u_{\mathrm{k}+1}) + W(u_{\mathrm{k}}, u_{\mathrm{k}-1})} \quad (3.8)$$

$$p_{\mathrm{k}}\ (u_{\mathrm{k}}, u_{\mathrm{k}-1}) = 1 - p_{\mathrm{k}}\ (u_{\mathrm{k}}, u_{\mathrm{k}+1}) \quad (3.9)$$

- Xem xét sự tương đồng về cấu trúc liên kết của các nút, vec-tơ đặc trưng ẩn 64 chiều của protein sau đó đã được tạo ra. Ngoài ra, dữ liệu protein mục tiêu của các dòng tế bào và thuốc tương tự như dữ liệu mạng PPI, do đó, các đặc trưng của thuốc được trích xuất bằng cách tính giá trị trung bình của vec-tơ protein nhắm mục tiêu thuốc bằng cách tính giá trị trung bình của protein mục tiêu bị ảnh hưởng bởi thuốc P_{Di}. Tương tự, các đặc trưng dòng tế bào được trích xuất bằng cách tính giá trị trung bình của vec-tơ protein nhắm mục tiêu bằng cách tính giá trị trung bình của protein tương tác với dòng tế bào P_{Cj}. Trong đó Di và Cj được ký hiệu là vec-tơ biểu diễn của thuốc i và dòng tế bào j tương tác với n protein.

$$Di = \frac{P_{D1} + P_{D2} + \cdots + P_{Dn}}{n} \quad (3.10)$$

$$Cj = \frac{P_{C1} + P_{C2} + \cdots + P_{Cn}}{n} \quad (3.11)$$

Trích xuất biểu diễn của các dòng tế bào từ dữ liệu -omics

Để trích xuất đặc trưng của các dòng tế bào từ dữ liệu biểu diễn hệ gen (genomics) và dữ liệu methyl hóa (epigenomics), nghiên cứu áp dụng mô hình autoencoders (AE) như là một mô hình pre-trained để học các biểu diễn của mỗi loại dữ liệu -omic. Cụ thể, với mỗi loại dữ liệu MUT và METH, nghiên cứu xây dựng bộ mã hóa MUTenc (với cấu hình 4 lớp, mỗi lớp: 735, 1024, 256 và 64); METHenc (378, 1024.256 và 64) tương ứng để xây dựng trích xuất biểu diễn dòng tế bào. Mỗi dòng tế bào được biểu diễn bởi các vec-tơ 64 chiều. Các vec-tơ này được ghép nối thành vec-tơ biểu diễn sự tương tác giữa cặp thuốc với dòng tế bào. Phương pháp AE-XGBSynergy có khả năng tích hợp dữ liệu đơn -omics (như MUT_CNA hoặc METH) cũng như tích hợp đa -omics (như kết hợp cả MUT_CNA và METH) để tăng cường các đặc trưng của dòng tế bào cho quá trình dự đoán.

Dự đoán đáp ứng đa thuốc cho dòng tế bào

Giải pháp đề xuất đã sử dụng XGBoost làm công cụ phân loại để dự đoán kết hợp thuốc của cặp thuốc cho dòng tế bào dựa trên các biểu diễn được trích xuất trên của thuốc và dòng tế bào. Bộ dự đoán này được áp dụng như một nhiệm vụ phân loại nhị phân với ngưỡng giá trị kết hợp được đặt lớn hơn 0 [23]. Quy trình tổng thể của khung AE-XGBSynergy được hiển thị trong Hình 3.4.

Phương pháp đánh giá mô hình

Hiệu năng của AE-XGBSynergy được đánh giá trên cả kịch bản tích hợp đơn dữ liệu -omics và đa -omics trên sáu chỉ số đánh giá bao gồm độ chính xác (Accuracy), Recall, AUC-ROC, AUC-PR, độ chính xác (Precision) và F1-score. Ngoài ra, các chỉ số này được so sánh hiệu năng của mô hình với hiệu năng của mô hình hiện đại trước đó, NEXGB là mô hình chỉ sử dụng mạng PPI để trích xuất thông tin đặc trưng của các cặp thuốc và dòng tế bào mà không tích hợp thêm dữ liệu -omics cho quá trình dự đoán.

Để so sánh công bằng giữa các kịch bản tích hợp của phương pháp đề xuất và NEXGB, nghiên cứu đã sử dụng một bộ kiểm tra duy nhất. Cụ thể, các tập dữ liệu này được thực hiện bằng cách xáo trộn dữ liệu một cách ngẫu nhiên, sau đó chia toàn bộ tập dữ liệu theo tỷ lệ 80:20 và tỷ lệ tương đồng về số mẫu dương tính và âm tính lần lượt thành tập huấn luyện và tập kiểm tra. Sau đó, nghiên cứu thực hiện xác thực chéo năm lần trên tập huấn luyện. Kết quả từ năm đánh giá này sau đó được tính trung bình để tạo ra thước đo đánh giá cuối cùng cho mô hình.

3.4.2. Cài đặt và thực nghiệm mô hình

Bộ dữ liệu

Kế thừa tập dữ liệu từ các đề xuất ở chương 2 và chương 3, nghiên cứu sử dụng hai loại dữ liệu đột biến gen và methyl hóa làm dữ liệu đầu vào biểu diễn dòng tế bào cho mô hình đề xuất. Các dữ liệu -omics này biểu diễn ở dạng nhị phân (1,0) cho thấy dòng tế bào có hay không biểu hiện đột biến hoặc hoặc có bị methyl hóa hay không.

Để trích xuất thông tin cấu trúc mạng PPI, nghiên cứu sử dụng cùng bộ dữ liệu tương tác thuốc-protein (Drug-Protein association) và Cell line-protein từ nghiên cứu NEXGB[3], cụ thể:

- Drug-Protein association: 15,051 tương tác bởi 4428 thuốc và 2256 proteins.
- Cell line-Protein association: 749,551 tương tác bởi 1035 dòng tế bào ung thư, và 18,022 proteins [44].
- Bộ dữ liệu mạng PPI gồm 15,970 protein, 217,160 tương tác, các protein được biểu diễn bằng số gen với mã hóa được lập bởi bởi GeneCards [67].

Bộ dữ liệu kết hợp thuốc:

- O'Neil dataset [75] gồm 583 cặp thuốc của 38 loại thuốc khác nhau với 39 dòng tế bào, tổng số có 23,062 tương tác. Độ kết hợp thuốc tính toán theo chỉ số Loewe.
- DrugCombDB[4] gồm 764 thuốc khác nhau và 76 dòng tế bào, tổng số 69,436 tương tác kết hợp. Độ kết hợp thuốc tính toán theo chỉ số ZIP.

Kết hợp các bộ dữ liệu trên, nghiên cứu có bộ dữ liệu tổng hợp như Bảng 3.7

Bảng 3.7. Tập dữ liệu thử nghiệm cho AE-XGBSynergy

Datasets	#Drugs	#Cell lines	#Proteins	#Samples
Oncology	21	21	256	3.023
DrugCombDB	69	53	9.923	58.322

3.4.3. Kết quả và đánh giá

Nghiên cứu đã thực hiện đánh giá phương pháp đề xuất trên hai bộ dữ liệu công khai là O'Neil và DrugCombDB cũng như so sánh với một phương pháp không tích hợp dữ liệu -omics là NEXGB. Các kết quả cho thấy AE-XGBSynergy vượt trội hơn so với NEXGB, trong các trường hợp tích hợp dữ liệu đơn dữ liệu -omics và đa dữ liệu -omics.

Cụ thể, hiệu năng của AE-XGBSynergy trên cả hai tập dữ liệu cho thấy MUT_CNA là dữ liệu đóng góp nhiều thông tin quan trọng hơn METH trong việc

[3] https://github.com/lysmfj/NEXGB
[4] http://drugcombdb.denglab.org/main

tích hợp các single - omics trong để dự đoán. Hơn nữa, sự kết hợp giữa MUT_CNA & METH đã đạt được hiệu năng tốt nhất trong bộ dữ liệu O'Neil xét về tất cả sáu chỉ số hiệu năng: ACC (0,790083), Thu hồi (0,810559), AUC-ROC (0,788672), AUC-PR (0,747786), Độ chính xác (0,798165) và F1-score (0,804314). Tương tự, trên tập dữ liệu DrugcombDB, MUT_CNA cũng được xác định là dữ liệu có nhiều thông tin nhất trong dự đoán. Sự kết hợp với MUT_CNA mang lại hiệu năng vượt trội hơn so với NEXGB và so với khi kết hợp với METH

Bảng 3.8. So sánh hiệu năng dự đoán trên bộ dữ liệu O'Neil

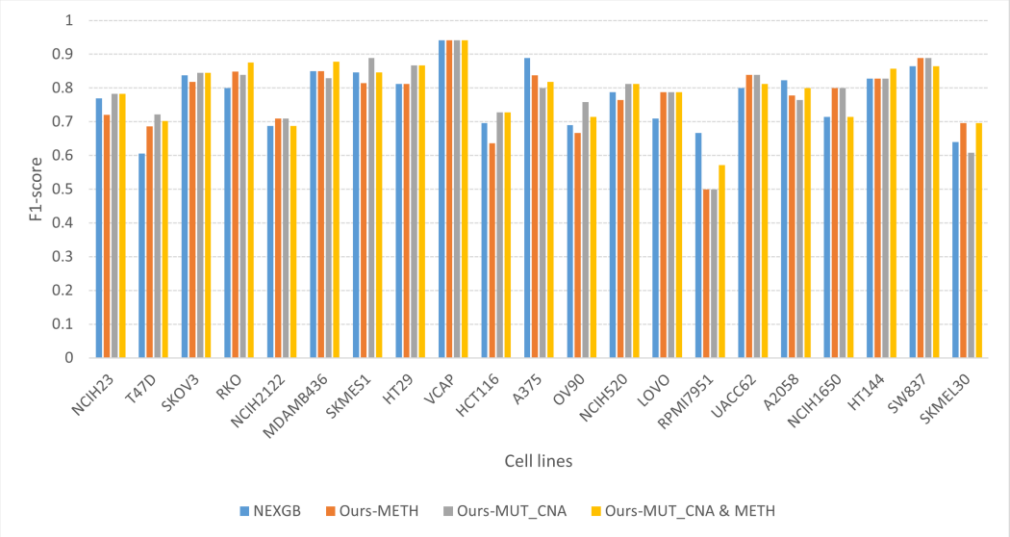
Methods	-Omics	Accuracy	Recall	AUC-ROC	AUC-PR	Precision	F1-score
NEXGB		0.7702	0.7919	0.7688	0.7283	0.7798	0.7858
AE-XGBSynergy	**METH**	0.7736	0.8012	0.7716	0.7303	0.7795	0.7902
	MUT_CNA	0.7884	**0.8106**	0.7869	0.7458	0.7957	0.8031
	MUT_CNA & METH	**0.7901**	**0.8106**	**0.7887**	**0.7478**	**0.7982**	**0.8043**

Bảng 3.9. So sánh hiệu năng dự đoán trên trên bộ dữ liệu DrugCombDB

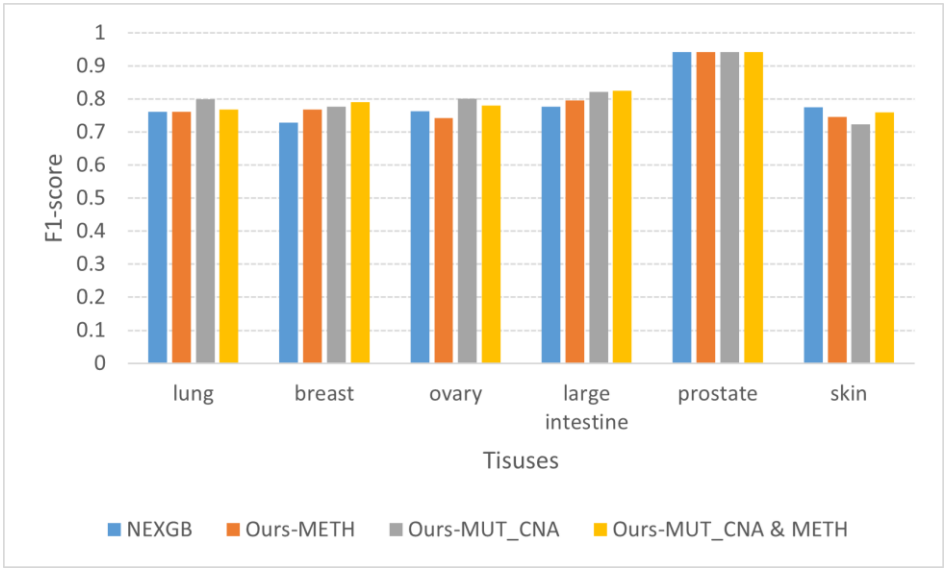
Methods	-Omics	Accuracy	Recall	AUC-ROC	AUC-PR	Precision	F1-score
NEXGB		0.7591	0.6930	0.7524	0.6529	0.7457	0.7184
AE-XGBSynergy	**METH**	0.7602	0.6976	0.7538	0.6539	0.7451	0.7206
	MUT_CNA	**0.7653**	0.7016	**0.7588**	**0.6598**	**0.7520**	**0.7259**
	MUT_CNA & METH	0.7613	**0.7026**	0.7549	0.6550	0.7466	0.7217

Thêm vào đó khi khảo sát hiệu năng mô hình đề xuất trên từng dòng tế bào và trên từng kiểu mô, mô hình đề xuất cũng cho kết quả tốt hơn so với NEXGB. Trong Hình 3.5, AE-XGBSynergy có hiệu năng vượt trội so với NEXGB ở hầu hết các dòng tế bào (19/22 dòng tế bào) khi xem xét việc tích hợp các kịch bản tích hợp dữ liệu -omics trong bộ dữ liệu O'Neil. AE-XGBSynergy thể hiện khả năng dự đoán phân loại kết hợp thuốc tốt, khi khảo sát với hiệu năng của mô hình với cách tích hợp -omics khác nhau trên METH, MUT_CNA, MUT_CNA & METH thì số lượng dòng tế bào dự đoán phân loại kết hợp thuốc trên dòng tế bào đó đạt F1-score lớn hơn 0,7 tương ứng lần lượt là 73%, 86% và 82% trong khi, số dòng tế bào của NEXGB chỉ đạt được các tỷ lệ là 68%. Đáng chú ý, trong số 21 dòng tế bào được phân tích, dòng tế bào RPMI7951 thuộc mô da có hiệu năng thấp nhất. Điều này có thể là do dòng tế bào RPMI7951 có số lượng tương tác protein tương đối thấp hơn so với các dòng tế bào khác (142 protein mục tiêu). Trong khi đó, AE-XGBSynergy và NEXGB đều đạt

được giá trị F1-score cao nhất (0,941) với dòng tế bào VCAP, đáng chú ý là có số lượng protein mục tiêu liên quan cao nhất trong số tất cả các dòng tế bào (830 protein mục tiêu). Quan sát này cho thấy rằng hiệu năng của các phương pháp dự đoán trong các dòng tế bào cụ thể không chỉ bị ảnh hưởng bởi các đặc điểm -omics của các dòng tế bào mà còn bởi số lượng protein mục tiêu liên quan của chúng



Hình 3.5. So sánh hiệu năng dự đoán cho dòng tế bào trên bộ dữ liệu O'Neil



Hình 3.6. So sánh hiệu năng dự đoán cho từng mô bệnh trên bộ dữ liệu O'Neil

Hình 3.6 cho thấy hiệu năng của AE-XGBSynergy được đánh giá và so sánh với các phương pháp khác trên sáu mô ung thư khác nhau như: ung thư vú, ruột già, phổi, buồng trứng, tuyến tiền liệt và da. Trong hầu hết các mô (4 / 5), phương pháp đề xuất của vượt trội hơn NEXGB. Cụ thể, đối với ung thư phổi và ung thư buồng trứng, sự kết hợp MUT_CNA đạt hiệu quả tốt nhất, với giá trị F1-score lần lượt là

0,7987 và 0,7757. Đối với ung thư vú và ung thư ruột già, sự kết hợp giữa MUT_CNA & METH mang lại giá trị F1-score lần lượt là 0,7987 và 0,8244, cao nhất trong số các cách tích hợp được thử nghiệm. Tương tự như hiệu năng so sánh của các dòng tế bào, VCAP là dòng tế bào duy nhất của ung thư tuyến tiền và có nhiều nhất protein (830 protein-targeted) trong tất cả các dòng tế bào, nên có thể hiệu năng ở dự đoán ung thư tuyến tiền liệt (VCAP) đạt được cùng giá trị F1-scorre với NEXGB.

Từ đó có thể thấy AE-XGBSynergy đưa ra một cách tiếp cận tiềm năng để xác định giải pháp tích hợp thông tin đồ thị sinh học trong mạng PPI ở mức cấu trúc và dữ liệu -omics của các dòng tế bào để dự đoán kết hợp thuốc cho các dòng tế bào. Phương pháp này có khả năng tích hợp thêm nhiều dữ liệu -omics có ý nghĩa nhằm tăng hiệu năng và độ chính xác của dự đoán.

3.5. KẾT LUẬN CHƯƠNG

Trong chương này, luận án đã trình bày hai giải pháp nghiên cứu cho dự đoán đáp ứng đa thuốc là GraOmicSynergy và AE-XGBSynergy. Các giải pháp này thực hiện các phương pháp tích hợp dữ liệu khác nhau nhằm nâng cao hiệu quả của dự đoán. Nội dung trình bày cho các giải pháp này nằm trong công trình nghiên cứu số 5 và số 4 của tác giả và các cộng sự.

GraOmicSynergy là một mô hình tích hợp dữ liệu tổng hợp biểu diễn đồ thị phân tử của cặp thuốc theo cơ chế chú ý và đa dữ liệu -omics của dòng tế bào GE, MUT_CNA và METH để dự đoán giá trị kết hợp của thuốc cho các dòng tế bào. Dữ liệu thuốc được biểu diễn dưới dạng đồ thị mang các đặc trưng tự nhiên của thuốc được huấn luyện bởi một biến thể của mạng nơ-ron đồ thị tích chập có khả năng phân biệt đồng phân hiệu quả, đặc biệt cho các dữ liệu phân tử hóa học là GIN. Mỗi thuốc có đáp ứng khác nhau cho dòng tế bào, nghiên cứu giả định mỗi thuốc có một hệ số đóng góp khác nhau trong cặp kết hợp các thuốc điều trị bệnh. Giải pháp đề xuất thực hiện tính toán các hệ số chú ý của mỗi thuốc trong tổ hợp tương tác cặp thuốc cho dòng tế bào thông qua một cơ chế chú ý; vec-tơ tổng hợp cặp thuốc này được tổng hợp bằng một phép tính cộng, toán tử có tính chất giao hoán không phụ thuộc vào thứ tự kết hợp các thuốc với nhau như các phương pháp trước đây. Mô hình tổng hợp tối đa đặc trưng ẩn của các dòng tế bào các thông qua việc kế hợp nhiều dữ liệu -omics của dòng tế bào, thông qua khối các khối mạng nơ-ron tích chập. Các biểu diễn cặp

thuốc và dòng tế bào được kết hợp tạo thành vec-tơ biểu diễn khả năng kết hợp thuốc cho dòng tế bào, được đưa vào khối dự đoán để dự đoán chỉ số kết hợp thuốc. Mô hình cũng được chuyển đổi để dự đoán phân loại khả năng kết hợp hoặc không kết hợp của cặp thuốc cho dòng tế bào.

Mô hình thử nghiệm được tiến hành trên các kịch bản khác nhau cho thấy hiệu quả của tích hợp đa dữ liệu -omics và cơ chế chú ý tổng hợp cặp thuốc đáp ứng dòng tế bào khi so sánh với các phương pháp dự đoán tiên tiến hiện nay là DeepSynergy và DeepDDS. Trong hầu hết các thử nghiệm như Mixed và thử nghiệm Blind-DrugPair, sự kết hợp của dữ liệu biểu hiện gen, methyl hóavà di truyền vượt trội so với những thử nghiệm khác.

Bài toán tích hợp dữ liệu sinh học tiếp tục được mở rộng với mô hình đề xuất AE-XGBSynergy. Trong mô hình này, thông tin cấu trúc mạng PPI khai thác mối quan hệ phức tạp của gen-bệnh-thuốc để trích xuất dữ liệu biểu diễn thuốc và dòng tế bào từ đó tích hợp đa dữ liệu -omics (MUT_CNA và METH) được trích xuất thông qua bộ autoencoder để biểu diễn của các dòng tế bào để dự đoán khả năng kết hợp thuốc qua bộ phân loại XGBoost. Hướng tiếp cận này không chỉ khai thác được các đặc trưng dòng tế bào, mà còn khai thác các mối quan hệ, tương tác qua lại giữa các cặp thuốc-prtotein, cell line-protein nhằm làm tăng độ chính xác của thuật toán cũng như bao quát sâu, rộng các yếu tố ảnh hưởng đến kết hợp thuốc trong điều trị. Các thực nghiệm trên các bộ dữ liệu khác nhau cho thấy hiệu quả của mô hình tích hợp đa dữ liệu -omics so với đơn -omics. Ngoài ra các giải pháp tích hợp dữ liệu -omics cũng cho thấy hiệu năng vượt trội hơn so với mô hình chỉ sử dụng cấu trúc thông tin mạng PPI như NEXGB.

PHẦN KẾT LUẬN

Các kết quả đã đạt được

Những nghiên cứu dự đoán đáp ứng thuốc trong điều trị bệnh hiện đang góp phần nâng cao hiệu quả trong nghiên cứu điều trị tiền lâm sàng và lâm sàng. Các phương pháp dự đoán tiềm năng cũng là cấu phần quan trọng trong việc xây dựng các mô đun tính toán dự đoán trong các hệ thống dự đoán điều trị trong y học chính xác hiện nay. Sự phát triển mạnh mẽ về công nghệ thông lượng cao và các nghiên cứu chuyên sâu về y sinh học đã sinh ra lượng lớn dữ liệu cần khai phá thông tin. Do vậy, luận án tập trung nghiên cứu tổng quan về y sinh học, các phương pháp dự đoán đáp ứng thuốc. Thông qua đó, luận án thấy được ý nghĩa của dự đoán đáp ứng thuốc trong y học chính xác và đề xuất các giải pháp cho hai bài toán dự đoán đáp ứng thuốc là dự đoán đáp ứng đơn thuốc (monotherapy) và dự đoán kết hợp thuốc (combination therapy) trong điều trị. Cụ thể:

Đề xuất các giải pháp để dự đoán đáp ứng đơn thuốc

Dự đoán đáp ứng đơn thuốc nhằm mục đích dự đoán giá trị đáp ứng của từng thuốc cho một một dòng tế bào hoặc người bệnh. Thách thức của bài toán là dữ liệu lớn của gần một nghìn dòng tế bào, mỗi dòng tế bào có vài trăm đến hàng chục nghìn đặc trưng. Trong khi thuốc được biểu diễn bằng các chuỗi phân tử hóa học. Các bài toán dự đoán trước kia thường tập trung vào các bộ dữ liệu nhỏ, sử dụng hoặc không sử dụng dữ liệu về thuốc nên hiệu năng dự đoán không cao. Luận án đã trình bày hai đề xuất cho bài toán này nâng cao hiệu quả dự đoán: (1) GraphDRP: đề xuất áp dụng mô hình biểu diễn dạng đồ thị phân tử thuốc, sử dụng các biến thể của mô hình học sâu dựa trên mạng nơ-ron đồ thị tích hợp với dữ liệu biểu diễn hệ gen di truyền (genomics) được học qua mạng nơ-ron tích chập để dự đoán giá trị đáp ứng thuốc cho từng dòng tế bào; (2) GraOmicDRP: là đề xuất áp dụng phương pháp tích hợp muộn để dự đoán đáp ứng thuốc cho các dòng tế bào. Phương pháp này phát triển dựa trên lợi thế của mô hình học sâu mạng nơ-ron đồ thị tích chập để học các biểu diễn của các phân tử thuốc, trong khi các dòng tế bào được học từ các nhánh kết hợp đa dữ liệu -omics khác nhau không chỉ là dữ liệu gen di truyền (genomics) mà còn là dữ liệu methyl hóa (epigenomics), dữ liệu biểu hiện gen (transciptomics). Hai đề xuất này đã chỉ ra rằng:

- Việc cải tiến cách biểu diễn thuốc một cách tự nhiên hơn dưới dạng đồ thị phân tử (GraphDRP) đã cho thấy hiệu quả vượt trội trên tất cả các kịch bản thử nghiệm như Mixed, Blind-drug, Blind-Cellline so với việc học các biểu diễn thuốc tiên tiến khác như biểu diễn thuốc dưới dạng chuỗi mã hóa one-hot (tCNNs).

- Trong các biến thể của mạng nơ-ron đồ thị như GCN, GAT, GIN, GCN-GAT, thì việc học các đặc trưng ẩn và đặc biệt là các dữ liệu đồng hình dựa trên mô hình GIN cho kết quả tiềm năng hơn các mô hình mạng còn lại trong dự đoán đáp ứng thuốc.

- Áp dụng mô hình học các biểu diễn đặc trưng phân tử thuốc dựa trên mạng nơ-ron đồ thị đồng hình (GIN) và tích hợp đa dữ liệu -omics, GraOmicDRP đã cho thấy khả năng vượt trội của việc tích hợp đa dữ liệu -omics hơn khi tích hợp đơn dữ liệu -omics. So sánh với các nghiên cứu tích hợp muộn tiên tiến khác mà không sử dụng biểu diễn phân tử thuốc dưới dạng đồ thị (no-graph) như MOLI, và DeepDR, nghiên cứu một lần nữa cho thấy việc áp dụng các biểu diễn phân tử thuốc dưới dạng đồ thị cũng cho các kết quả tốt hơn.

- Việc tích hợp đa dữ liệu -omics trong dự đoán đáp ứng thuốc không chỉ tăng độ chính xác của dự đoán mà còn giúp xác định được dữ liệu có ý nghĩa, có đóng góp nhiều vào trong quá trình dự đoán đáp ứng thuốc. Cụ thể với GraOmicsDRP, dữ liệu biểu hiện gen (GE) là dữ liệu đóng góp quan trọng trong việc dự đoán đáp ứng thuốc.

Đề xuất các giải pháp để dự đoán đáp ứng đa thuốc

Kết hợp thuốc là liệu trình điều trị khắc phục tình trạng kháng thuốc sau thời gian dài điều trị theo liệu trình đơn thuốc. Tuy dữ liệu về kết hợp thuốc còn chưa nhiều như dữ liệu đáp ứng đơn thuốc, nhưng các phương pháp dự đoán đáp ứng đa thuốc cũng đang nhận được sự quan tâm của nhiều nhà nghiên cứu. Theo xu hướng tiếp cận các nghiên cứu nâng cao hiệu quả trong điều trị này, luận án đã trình bày hai đề xuất cho bài toán dự đoán đáp ứng đa thuốc GraOmicSynergy và AE-XGBSynergy. Trong đó: (1) GraOmicSynergy: Đề xuất phương pháp tích hợp đa dữ liệu -omics của các dòng tế bào với dữ liệu biểu diễn tổng hợp thuốc biểu diễn dưới dạng đồ thị thông qua mạng nơ-ron đồ thị tích chập (GIN) và cơ chế chú ý đối với cặp thuốc tương tác từ đó dự đoán độ kết hợp thuốc đối với dòng tế bào; (2) AE-XGBSynergy: đề xuất

tích hợp dữ liệu biểu diễn đặc trưng đa -omics và các đặc trưng của dòng tế bào và áp dụng thuật toán struc2vec để trích xuất các đặc trưng của dòng tế bào và thuốc dựa trên các đặc trưng cấu trúc liên kết của từng protein trong mạng PPI để dự đoán phân loại kết hợp thuốc. Hai nghiên cứu này chỉ ra rằng:

- Việc tích hợp đa dữ liệu -omics mang lại hiệu quả dự đoán tốt hơn không chỉ với bài toán dự đoán đáp ứng đơn thuốc mà còn cho cả bài toán dự đoán đáp ứng kết hợp thuốc. Trong GraOmicSynergy, các kịch bản thử nghiệm Mixed, Blind-drugpair, Blind-Cellline cho thấy hiệu năng vượt trội của việc tích hợp đa dữ liệu -omics tốt hơn so với tích hợp đơn dữ liệu -omics. Các kết quả này cũng được thể hiện tương đối đồng nhất khi so sánh với các phương pháp tiên tiến hiện nay là DeepDDS và DeepSynergy.

- Việc cải thiện phương pháp dự đoán trong GraOmicSynergy bằng cách áp dụng mô hình GIN để học các biểu diễn thuốc dạng đồ thị, tổng hợp thông tin biểu diễn cặp thuốc tương tác với dòng tế bào theo cơ chế chú ý đồng thời tổng hợp thông tin biểu diễn đa dữ liệu -omics của dòng tế bào đã cho hiệu năng vượt trội so với phương pháp (DeepDDS) chỉ tích hợp một loại dữ liệu -omics là dữ liệu biểu hiện gen với dữ liệu biểu diễn thuốc dưới dạng đồ thị được học thông qua mạng GAT. Giải pháp tích hợp đề xuất cũng cho thấy việc học các biểu diễn thuốc dưới dạng đồ thị tiếp tục cho thấy hiệu quả hơn học khi mô hình học các biểu diễn thuốc không phải là dạng đồ thị (như fingerprint trong DeepSynergy).

- Việc tích hợp đa dữ liệu -omics không chỉ mang lại hiệu quả dự đoán đối với các mô hình học sâu áp dụng biểu diễn dữ liệu đồ thị mà còn mang lại hiệu quả đối với mô hình dự đoán kết hợp thuốc dựa trên việc trích xuất thông tin cấu trúc mạng đồ thị tương tác protein – protein PPI. AE-XGBSynergy: khai thác các mối quan hệ thuốc – đích (drug-protein), dòng tế bào – protein, áp dụng thuật toán struc2vec, đồ thị đa tầng để trích xuất đặc điểm tương đồng về cấu trúc và tạo ra bối cảnh cấu trúc cho các nút và trích xuất đặc trưng của dòng tế bào và thuốc để tích hợp với các biểu diễn dòng tế bào như dữ liệu methyl hóa, dữ liệu hệ gen di truyền để dự đoán kết hợp thuốc. Các kết quả thử nghiệm đã cho thấy AE-XGBSynergy là mô hình tiềm năng trong việc tích hợp đa dữ liệu -omics với thông tin mạng tương tác protein để dự đoán sự kết hợp thuốc. Khi so sánh với mô hình tiên tiến NEXGB, là mô hình chỉ kết hợp

các đặc trưng của thuốc và dòng tế bào bằng mạng PPI, mô hình AE-XGBSynergy vượt trội hơn về tất cả các độ đo được so sánh.

Hướng phát triển của đề tài luận án

Các đề xuất liên quan đến bài toán dự đoán đáp ứng đơn thuốc và dự đoán kết hợp thuốc được trình bày trong luận án đã giải quyết được một số vấn đề đặt ra cũng như có khả năng so sánh với các phương pháp dự đoán tiên tiến hiện nay. Tuy nhiên, các đề xuất này còn có khả năng được cải thiện hơn nữa về cách mô hình hóa dữ liệu và cải tiến phương pháp dự đoán.

- Về mặt dữ liệu:
    - Thuốc: Đồ thị phân tử thuốc đã cho thấy tiềm năng trong việc dự đoán, tuy nhiên chưa tích hợp các đặc trưng cạnh cũng như nhiều đặc trưng lý hóa khác chưa được khai thác.
    - Dòng tế bào: Các dữ liệu -omics đã được khai thác và chỉ ra dữ liệu quan trọng cho dự đoán, tuy nhiên mới có ba loại dữ liệu -omics và một phần thông tin cấu trúc mạng PPI được triển khai, trong khi còn nhiều các thông tin sinh học khác của dòng tế bào cần khai thác như proteomics, metabomics, .. hay các dữ liệu tương tác pathway, drug-drug interaction, gene-gene interaction.
- Về mặt phương pháp:
    - Các phương pháp mới có thể được triển khai như Molecular-pretrain model, transformer, graphformer.. đã được đề xuất và mang lại kết quả tiềm năng cho các bài toán dự đoán khai phá thuốc nhưng chưa được áp dụng cho bài toán dự đoán đáp ứng thuốc.
- Do đó hướng phát triển tiếp theo có thể triển khai là tiếp tục cải tiến áp dụng các phương pháp mới theo hướng tiếp cận, tăng độ chính xác và hiệu năng lên cao hơn nữa bằng cách cải thiện phương pháp tính toán và trong việc cải thiện phương pháp mô hình hóa dữ liệu thuốc và dòng tế bào:
    - Tích hợp thêm các đặc trưng cạnh, các thông tin lý hóa khác thể hiện rõ mối quan hệ giữa các đỉnh, cạnh của đồ thị cho việc biểu diễn đồ thị phân tử.

- Chuyển mô hình biểu diễn dữ liệu 2D sang dữ liệu 3D, tích hợp đặc trưng góc, cấu trúc không gian 3 chiều đầy đủ của đồ thị phân tử thuốc để học nhiều hơn các đặc trưng ẩn của phân tử thuốc.
- Mô hình hóa dòng tế bào, thuốc bằng cách tích hợp nhiều dạng dữ liệu khác nhau.
- Áp dụng các mô hình tính toán tiên tiến khác như transformer, graphformer,.. và các mô hình Pretrain cho bài toán dự đoán.

DANH MỤC CÁC CÔNG TRÌNH CÔNG BỐ

TẠP CHÍ KHOA HỌC

[1] T. Nguyen, **G. T. T. Nguyen**, T. Nguyen and D. - H. Le, "Graph Convolutional Networks for Drug Response Prediction," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 19, no. 1, pp. 146-154, 1 Jan.-Feb. 2022, doi: 10.1109/TCBB.2021.3060430.

[2] **G. T. T. Nguyen,** H. D. Vu and D. -H. Le, "Integrating Molecular Graph Data of Drugs and Multiple -Omic Data of Cell Lines for Drug Response Prediction," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 19, no. 2, pp. 710-717, 1 March-April 2022, doi: 10.1109/TCBB.2021.3096960

[5] **G. T. T. Nguyen**, D. -H. Vu, Trong-Khanh Nguyen, Tan Phan-Xuan, and D. -H. Le, "Integrating multiple -omic data of cell lines for drug synergy prediction," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2024 [Đã nộp, chờ phản biện]

HỘI NGHỊ KHOA HỌC

[3] **G. T. T. Nguyen**, L. Due Hoang, Q. D. Nguyen, T. T. Nguyen, H. T. T. Dang and D. -H. Le, "An investigation of cancer cell line-based drug response prediction methods on patient data," 2020 12th International Conference on Knowledge and Systems Engineering (KSE), Can Tho, Vietnam, 2020, pp. 306-311

[4] **G. T. T. Nguyen**, K. T. Phuong, K. Nguyen-Trong and D. -H. Le, "A Hybrid Model Integrating Multi-Omic and Topological Information of PPI Network for Drug Synergism Prediction," 2023 RIVF International Conference on Computing and Communication Technologies (RIVF), Hanoi, Vietnam, 2023, pp. 83-88.

TÀI LIỆU THAM KHẢO

[1] S. J. Aronson and H. L. Rehm, “Building the foundation for genomics in precision medicine,” *Nature*, vol. 526, no. 7573, pp. 336–342, Oct. 2015, doi: 10.1038/nature15816.

[2] J. S. Boehm and T. R. Golub, “An ecosystem of cancer cell line factories to support a cancer dependency map,” *Nat. Rev. Genet.*, vol. 16, no. 7, Art. no. 7, Jul. 2015, doi: 10.1038/nrg3967.

[3] G. Caponigro and W. R. Sellers, “Advances in the preclinical testing of cancer therapeutic hypotheses,” *Nat. Rev. Drug Discov.*, vol. 10, no. 3, Art. no. 3, Mar. 2011, doi: 10.1038/nrd3385.

[4] A. Kalamara, L. Tobalina, and J. Saez-Rodriguez, “How to find the right drug for each patient? Advances and challenges in pharmacogenomics,” *Curr. Opin. Syst. Biol.*, vol. 10, pp. 53–62, Aug. 2018, doi: 10.1016/j.coisb.2018.07.001.

[5] B. Mansoori, A. Mohammadi, S. Davudian, S. Shirjang, and B. Baradaran, “The Different Mechanisms of Cancer Drug Resistance: A Brief Review,” *Adv. Pharm. Bull.*, vol. 7, no. 3, pp. 339–348, Sep. 2017, doi: 10.15171/apb.2017.041.

[6] J. S. Boehm and T. R. Golub, “An ecosystem of cancer cell line factories to support a cancer dependency map,” *Nat. Rev. Genet.*, vol. 16, p. 373, Jun. 2015.

[7] G. Caponigro and W. R. Sellers, “Advances in the preclinical testing of cancer therapeutic hypotheses,” *Nat. Rev. Drug Discov.*, vol. 10, p. 179, Mar. 2011.

[8] F. Firoozbakht, B. Yousefi, and B. Schwikowski, “An overview of machine learning methods for monotherapy drug response prediction,” *Brief. Bioinform.*, vol. 23, no. 1, p. bbab408, Oct. 2021, doi: 10.1093/bib/bbab408.

[9] C. De Niz, R. Rahman, X. Zhao, and R. Pal, “Algorithms for Drug Sensitivity Prediction,” *Algorithms*, vol. 9, no. 4, Art. no. 4, Dec. 2016, doi: 10.3390/a9040077.

[10] E. C. Neto, I. S. Jang, S. H. Friend, and A. A. Margolin, “The Stream algorithm: computationally efficient ridge-regression via Bayesian model averaging, and applications to pharmacogenomic prediction of cancer cell line sensitivity,” *Pac. Symp. Biocomput. Pac. Symp. Biocomput.*, pp. 27–38, 2014.

[11] M. P. Menden *et al.*, “Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen,” *Nat. Commun.*, vol. 10, no. 1, Art. no. 1, Jun. 2019, doi: 10.1038/s41467-019-09799-2.

[12] G. T. T. Nguyen and D.-H. Le, “A matrix completion method for drug response prediction in personalized medicine,” in *Proceedings of the Ninth International Symposium on Information and Communication Technology*, in SoICT 2018. New York, NY, USA: Association for Computing Machinery, Dec. 2018, pp. 410–415. doi: 10.1145/3287921.3287974.

[13] I. Tavassoly, J. Goldfarb, and R. Iyengar, “Systems biology primer: the basic methods and approaches,” *Essays Biochem.*, vol. 62, no. 4, pp. 487–500, Oct. 2018, doi: 10.1042/EBC20180003.

[14] L. de Oliveira, D. Hudebine, D. Guillaume, and J. Verstraete, “A Review of Kinetic Modeling Methodologies for Complex Processes,” *Oil Gas Sci. Technol.*, vol. 71, p. 45, May 2016, doi: 10.2516/ogst/2016011.

[15] M. Bansal *et al.*, “A community computational challenge to predict the activity of pairs of compounds,” *Nat. Biotechnol.*, vol. 32, no. 12, Art. no. 12, Dec. 2014, doi: 10.1038/nbt.3052.

[16] J. Wildenhain *et al.*, “Prediction of Synergism from Chemical-Genetic Interactions by Machine Learning,” *Cell Syst.*, vol. 1, no. 6, pp. 383–395, Dec. 2015, doi: 10.1016/j.cels.2015.12.003.

[17] M. Ali and T. Aittokallio, “Machine learning and feature selection for drug response prediction in precision oncology applications,” *Biophys. Rev.*, vol. 11, no. 1, pp. 31–39, Feb. 2019, doi: 10.1007/s12551-018-0446-z.

[18] R. Kurilov, B. Haibe-Kains, and B. Brors, “Assessment of modelling strategies for drug response prediction in cell lines and xenografts,” *Sci. Rep.*, vol. 10, no. 1, Art. no. 1, Feb. 2020, doi: 10.1038/s41598-020-59656-2.

[19] L. Zhang, J. Tan, D. Han, and H. Zhu, “From machine learning to deep learning: progress in machine intelligence for rational drug discovery,” *Drug Discov. Today*, vol. 22, no. 11, pp. 1680–1685, Nov. 2017, doi: 10.1016/j.drudis.2017.08.010.

[20] H. Sharifi-Noghabi, O. Zolotareva, C. C. Collins, and M. Ester, “MOLI: multi-omics late integration with deep neural networks for drug response prediction,” *Bioinformatics*, vol. 35, no. 14, pp. i501–i509, Jul. 2019, doi: 10.1093/bioinformatics/btz318.

[21] P. Liu, H. Li, S. Li, and K.-S. Leung, “Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network,” *BMC Bioinformatics*, vol. 20, no. 1, p. 408, Jul. 2019, doi: 10.1186/s12859-019-2910-6.

[22] L. Rampášek, D. Hidru, P. Smirnov, B. Haibe-Kains, and A. Goldenberg, “Dr.VAE: Improving drug response prediction via modeling of drug perturbation effects,” *Bioinforma. Oxf. Engl.*, vol. 35, Mar. 2019, doi: 10.1093/bioinformatics/btz158.

[23] I. Cortés-Ciriano and A. Bender, “KekuleScope: prediction of cancer cell line sensitivity and compound potency using convolutional neural networks trained on compound images,” *J. Cheminformatics*, vol. 11, no. 1, p. 41, Jun. 2019, doi: 10.1186/s13321-019-0364-5.

[24] D. Baptista, P. G. Ferreira, and M. Rocha, “Deep learning for drug response prediction in cancer,” *Brief. Bioinform.*, vol. 22, no. 1, pp. 360–379, Jan. 2021, doi: 10.1093/bib/bbz171.

[25] W. Peng, T. Chen, H. Liu, W. Dai, N. Yu, and W. Lan, “Improving drug response prediction based on two-space graph convolution,” *Comput. Biol. Med.*, vol. 158, p. 106859, May 2023, doi: 10.1016/j.compbiomed.2023.106859.

[26] H. Liu *et al.*, “Improving anti-cancer drug response prediction using multi-task learning on graph convolutional networks,” *Methods*, vol. 222, pp. 41–50, Feb. 2024, doi: 10.1016/j.ymeth.2023.11.018.

[27] J.-L. Vincent, T. van der Poll, and J. C. Marshall, “The End of ‘One Size Fits All’ Sepsis Therapies: Toward an Individualized Approach,” *Biomedicines*, vol. 10, no. 9, Art. no. 9, Sep. 2022, doi: 10.3390/biomedicines10092260.

[28] J. L. Perez-Gracia *et al.*, “Strategies to design clinical studies to identify predictive biomarkers in cancer research,” *Cancer Treat. Rev.*, vol. 53, pp. 79–97, Feb. 2017, doi: 10.1016/j.ctrv.2016.12.005.

[29] H. Sung *et al.*, “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” *CA. Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[30] R. Q. Wu, X. F. Zhao, Z. Y. Wang, M. Zhou, and Q. M. Chen, “Novel molecular events in oral carcinogenesis via integrative approaches,” *J. Dent. Res.*, vol. 90, no. 5, pp. 561–572, May 2011, doi: 10.1177/0022034510383691.

[31] X. Dai and L. Shen, “Advances and Trends in Omics Technology Development,” *Front. Med.*, vol. 9, p. 911861, Jul. 2022, doi: 10.3389/fmed.2022.911861.

[32] A. Goodspeed, L. M. Heiser, J. W. Gray, and J. C. Costello, “Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics,” *Mol. Cancer Res. MCR*, vol. 14, no. 1, pp. 3–13, Jan. 2016, doi: 10.1158/1541-7786.MCR-15-0189.

[33] T. A. Manolio *et al.*, “Finding the missing heritability of complex diseases,” *Nature*, vol. 461, no. 7265, Art. no. 7265, Oct. 2009, doi: 10.1038/nature08494.

[34] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, “KEGG for integration and interpretation of large-scale molecular data sets,” *Nucleic Acids Res.*, vol. 40, no. D1, pp. D109–D114, Jan. 2012, doi: 10.1093/nar/gkr988.

[35] N. Safari-Alighiarloo, M. Taghizadeh, M. Rezaei-Tavirani, B. Goliaei, and A. A. Peyvandi, “Protein-protein interaction networks (PPI) and complex diseases,” *Gastroenterol. Hepatol. Bed Bench*, vol. 7, no. 1, pp. 17–31, 2014.

[36] M. Bansal *et al.*, “A community computational challenge to predict the activity of pairs of compounds,” *Nat. Biotechnol.*, vol. 32, no. 12, pp. 1213–1222, Dec. 2014, doi: 10.1038/nbt.3052.

[37] S. Lederer, T. M. H. Dijkstra, and T. Heskes, “Additive Dose Response Models: Explicit Formulation and the Loewe Additivity Consistency Condition,” *Front. Pharmacol.*, vol. 9, 2018.

[38] A. Ianevski, L. He, T. Aittokallio, and J. Tang, “SynergyFinder: a web application for analyzing drug combination dose–response matrix data,” *Bioinformatics*, vol. 36, no. 8, p. 2645, Apr. 2020, doi: 10.1093/bioinformatics/btaa102.

[39] N. M. O’Boyle, “Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI,” *J. Cheminformatics*, vol. 4, no. 1, p. 22, Sep. 2012, doi: 10.1186/1758-2946-4-22.

[40] Y. Chang *et al.*, “Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature,” *Sci. Rep.*, vol. 8, no. 1, Art. no. 1, Jun. 2018, doi: 10.1038/s41598-018-27214-6.

[41] A. Gaulton *et al.*, “ChEMBL: a large-scale bioactivity database for drug discovery,” *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D1100-1107, Jan. 2012, doi: 10.1093/nar/gkr777.

[42] A. Gaulton *et al.*, “ChEMBL: a large-scale bioactivity database for drug discovery,” *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D1100-1107, Jan. 2012, doi: 10.1093/nar/gkr777.

[43] T. Sterling and J. J. Irwin, “ZINC 15 – Ligand Discovery for Everyone,” *J. Chem. Inf. Model.*, vol. 55, no. 11, pp. 2324–2337, Nov. 2015, doi: 10.1021/acs.jcim.5b00559.

[44] M. Ghandi *et al.*, “Next-generation characterization of the Cancer Cell Line Encyclopedia,” *Nature*, vol. 569, no. 7757, pp. 503–508, May 2019, doi: 10.1038/s41586-019-1186-3.

[45] W. Yang *et al.*, “Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells,” *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D955-961, Jan. 2013, doi: 10.1093/nar/gks1111.

[46] R. H. Shoemaker, “The NCI60 human tumour cell line anticancer drug screen,” *Nat. Rev. Cancer*, vol. 6, no. 10, pp. 813–823, Oct. 2006, doi: 10.1038/nrc1951.

[47] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, “Convolutional neural networks: an overview and application in radiology,” *Insights Imaging*, vol. 9, no. 4, Art. no. 4, Aug. 2018, doi: 10.1007/s13244-018-0639-9.

[48] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, “Neural Message Passing for Quantum Chemistry,” Jun. 12, 2017, *arXiv*: arXiv:1704.01212. doi: 10.48550/arXiv.1704.01212.

[49] T. N. Kipf and M. Welling, “Semi-Supervised Classification with Graph Convolutional Networks,” Feb. 22, 2017, *arXiv*: arXiv:1609.02907. doi: 10.48550/arXiv.1609.02907.

[50] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph Attention Networks,” Feb. 04, 2018, *arXiv*: arXiv:1710.10903. doi: 10.48550/arXiv.1710.10903.

[51] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, “How Powerful are Graph Neural Networks?,” *ArXiv181000826 Cs Stat*, Feb. 2019.

[52] P. Chen *et al.*, “Identification of Prognostic Groups in High-Grade Serous Ovarian Cancer Treated with Platinum-Taxane Chemotherapy,” *Cancer Res.*, vol. 75, no. 15, pp. 2987–2998, Aug. 2015, doi: 10.1158/0008-5472.CAN-14-3242.

[53] I. Cortés-Ciriano *et al.*, “Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel,” *Bioinformatics*, vol. 32, no. 1, pp. 85–95, Jan. 2016, doi: 10.1093/bioinformatics/btv529.

[54] I. Bayer, P. Groth, and S. Schneckener, “Prediction Errors in Learning Drug Response from Gene Expression Data – Influence of Labeling, Sample Size, and Machine Learning Algorithm,” *PLOS ONE*, vol. 8, no. 7, p. e70294, Jul. 2013, doi: 10.1371/journal.pone.0070294.

[55] I. S. Jang, E. C. Neto, J. Guinney, S. H. Friend, and A. A. Margolin, “Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data,” *Pac. Symp. Biocomput. Pac. Symp. Biocomput.*, pp. 63–74, 2014.

[56] A. Cichonska *et al.*, “Learning with multiple pairwise kernels for drug bioactivity prediction,” *Bioinformatics*, vol. 34, no. 13, pp. i509–i518, Jul. 2018, doi: 10.1093/bioinformatics/bty277.

[57] N. Zhang, H. Wang, Y. Fang, J. Wang, X. Zheng, and X. S. Liu, “Predicting Anticancer Drug Responses Using a Dual-Layer Integrated Cell Line-Drug

Network Model," *PLoS Comput. Biol.*, vol. 11, no. 9, p. e1004498, 2015, doi: 10.1371/journal.pcbi.1004498.

[58] A. Aliper, S. Plis, A. Artemov, A. Ulloa, P. Mamoshina, and A. Zhavoronkov, "Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data," *Mol. Pharm.*, vol. 13, no. 7, pp. 2524–2530, Jul. 2016, doi: 10.1021/acs.molpharmaceut.6b00248.

[59] L. Rampášek, D. Hidru, P. Smirnov, B. Haibe-Kains, and A. Goldenberg, "Dr.VAE: Improving drug response prediction via modeling of drug perturbation effects," *Bioinforma. Oxf. Engl.*, vol. 35, Mar. 2019, doi: 10.1093/bioinformatics/btz158.

[60] L. Kun and H. Wenbin, "TransEDRP: Dual Transformer model with Edge Emdedded for Drug Respond Prediction," Oct. 23, 2022, *arXiv*: arXiv:2210.17401. doi: 10.48550/arXiv.2210.17401.

[61] M. Li *et al.*, "DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, May 2019, doi: 10.1109/TCBB.2019.2919581.

[62] T. Chu, T. T. Nguyen, B. D. Hai, Q. H. Nguyen, and T. Nguyen, "Graph Transformer for Drug Response Prediction," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 20, no. 2, pp. 1065–1072, 2023, doi: 10.1109/TCBB.2022.3206888.

[63] Y. Yang and P. Li, "GPDRP: a multimodal framework for drug response prediction with graph transformer," *BMC Bioinformatics*, vol. 24, no. 1, p. 484, Dec. 2023, doi: 10.1186/s12859-023-05618-0.

[64] C. Holohan, S. Van Schaeybroeck, D. B. Longley, and P. G. Johnston, "Cancer drug resistance: an evolving paradigm," *Nat. Rev. Cancer*, vol. 13, no. 10, pp. 714–726, Oct. 2013, doi: 10.1038/nrc3599.

[65] K. Pang *et al.*, "Combinatorial therapy discovery using mixed integer linear programming," *Bioinforma. Oxf. Engl.*, vol. 30, no. 10, pp. 1456–1463, May 2014, doi: 10.1093/bioinformatics/btu046.

[66] M. M. Gottesman, O. Lavi, M. D. Hall, and J.-P. Gillet, "Toward a Better Understanding of the Complexity of Cancer Drug Resistance," *Annu. Rev.*

Pharmacol. Toxicol., vol. 56, pp. 85–102, 2016, doi: 10.1146/annurev-pharmtox-010715-103111.

[67] F. Cheng, I. A. Kovács, and A.-L. Barabási, "Network-based prediction of drug combinations," *Nat. Commun.*, vol. 10, no. 1, Art. no. 1, Mar. 2019, doi: 10.1038/s41467-019-09186-x.

[68] K. Preuer, R. P. I. Lewis, S. Hochreiter, A. Bender, K. C. Bulusu, and G. Klambauer, "DeepSynergy: predicting anti-cancer drug synergy with Deep Learning," *Bioinformatics*, vol. 34, no. 9, pp. 1538–1546, May 2018, doi: 10.1093/bioinformatics/btx806.

[69] J. Yang *et al.*, "DIGRE: Drug-Induced Genomic Residual Effect Model for Successful Prediction of Multidrug Effects," *CPT Pharmacomet. Syst. Pharmacol.*, vol. 4, no. 2, Feb. 2015, doi: 10.1002/psp4.1.

[70] X. Chen, B. Ren, M. Chen, Q. Wang, L. Zhang, and G. Yan, "NLLSS: Predicting Synergistic Drug Combinations Based on Semi-supervised Learning," *PLOS Comput. Biol.*, vol. 12, no. 7, p. e1004975, Jul. 2016, doi: 10.1371/journal.pcbi.1004975.

[71] T. N. Jarada, J. G. Rokne, and R. Alhajj, "SNF-NN: computational method to predict drug-disease interactions using similarity network fusion and neural networks," *BMC Bioinformatics*, vol. 22, no. 1, p. 28, Jan. 2021, doi: 10.1186/s12859-020-03950-3.

[72] J. Wang, X. Liu, S. Shen, L. Deng, and H. Liu*, "DeepDDS: deep graph neural network with attention mechanism to predict synergistic drug combinations," *ArXiv210702467 Cs Q-Bio*, Jul. 2021.

[73] H. Li *et al.*, "Predicting Drug Synergy and Discovering New Drug Combinations Based on a Graph Autoencoder and Convolutional Neural Network," *Interdiscip. Sci. Comput. Life Sci.*, vol. 15, no. 2, pp. 316–330, 2023, doi: 10.1007/s12539-023-00558-y.

[74] G. Adam, L. Rampášek, Z. Safikhani, P. Smirnov, B. Haibe-Kains, and A. Goldenberg, "Machine learning approaches to drug response prediction: challenges and recent progress," *Npj Precis. Oncol.*, vol. 4, no. 1, Art. no. 1, Jun. 2020, doi: 10.1038/s41698-020-0122-1.

[75] J. O'Neil *et al.*, "An Unbiased Oncology Compound Screen to Identify Novel Combination Strategies," *Mol. Cancer Ther.*, vol. 15, no. 6, pp. 1155–1162, Jun. 2016, doi: 10.1158/1535-7163.MCT-15-0843.

[76] H. Liu, W. Zhang, B. Zou, J. Wang, Y. Deng, and L. Deng, "DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D871–D881, Jan. 2020, doi: 10.1093/nar/gkz1007.

[77] Y. Hu, A. Shmygelska, D. Tran, N. Eriksson, J. Y. Tung, and D. A. Hinds, "GWAS of 89,283 individuals identifies genetic variants associated with self-reporting of being a morning person," *Nat. Commun.*, vol. 7, p. 10448, Feb. 2016, doi: 10.1038/ncomms10448.

[78] M. Hofree, J. P. Shen, H. Carter, A. Gross, and T. Ideker, "Network-based stratification of tumor mutations," *Nat. Methods*, vol. 10, no. 11, pp. 1108–1115, 2013, doi: 10.1038/nmeth.2651.

[79] B. Linghu, E. S. Snitkin, Z. Hu, Y. Xia, and C. Delisi, "Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network," *Genome Biol.*, vol. 10, no. 9, p. R91, 2009, doi: 10.1186/gb-2009-10-9-r91.

[80] J. Menche *et al.*, "Uncovering disease-disease relationships through the incomplete human interactome," *Science*, vol. 347, no. 6224, p. 1257601, Feb. 2015, doi: 10.1126/science.1257601.

[81] N. Zong, H. Kim, V. Ngo, and O. Harismendy, "Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations," *Bioinformatics*, vol. 33, no. 15, pp. 2337–2344, Aug. 2017, doi: 10.1093/bioinformatics/btx160.

[82] R. A. Hodos, B. A. Kidd, S. Khader, B. P. Readhead, and J. T. Dudley, "Computational Approaches to Drug Repurposing and Pharmacology," *Wiley Interdiscip. Rev. Syst. Biol. Med.*, vol. 8, no. 3, pp. 186–210, May 2016, doi: 10.1002/wsbm.1337.

[83] M. Zitnik and J. Leskovec, “Predicting multicellular function through multi-layer tissue networks,” *Bioinformatics*, vol. 33, no. 14, pp. i190–i198, Jul. 2017, doi: 10.1093/bioinformatics/btx252.

[84] J. Bicker, A. Fortuna, G. Alves, P. Soares-da-Silva, and A. Falcão, “Elucidation of the Impact of P-glycoprotein and Breast Cancer Resistance Protein on the Brain Distribution of Catechol-O-Methyltransferase Inhibitors,” *Drug Metab. Dispos. Biol. Fate Chem.*, vol. 45, no. 12, pp. 1282–1291, Dec. 2017, doi: 10.1124/dmd.117.077883.

[85] T. M. Santiago-Rodriguez and E. B. Hollister, “Multi ‘omic data integration: A review of concepts, considerations, and approaches,” *Semin. Perinatol.*, vol. 45, no. 6, p. 151456, Oct. 2021, doi: 10.1016/j.semperi.2021.151456.

[86] I. Subramanian, S. Verma, S. Kumar, A. Jere, and K. Anamika, “Multi-omics Data Integration, Interpretation, and Its Application,” *Bioinforma. Biol. Insights*, vol. 14, p. 1177932219899051, Jan. 2020, doi: 10.1177/1177932219899051.

[87] I. S. Jang, E. C. Neto, J. Guinney, S. H. Friend, and A. A. Margolin, “Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data,” *Pac. Symp. Biocomput. Pac. Symp. Biocomput.*, pp. 63–74, 2014.

[88] J. C. Costello *et al.*, “A community effort to assess and improve drug sensitivity prediction algorithms,” *Nat. Biotechnol.*, vol. 32, no. 12, pp. 1202–1212, Dec. 2014, doi: 10.1038/nbt.2877.

[89] S. B. Amin *et al.*, “Gene expression profile alone is inadequate in predicting complete response in multiple myeloma,” *Leukemia*, vol. 28, p. 2229, Apr. 2014.

[90] X. He, L. Folkman, and K. Borgwardt, “Kernelized rank learning for personalized drug recommendation,” *Bioinforma. Oxf. Engl.*, vol. 34, no. 16, pp. 2808–2816, Aug. 2018, doi: 10.1093/bioinformatics/bty132.

[91] D.-H. Le and V.-H. Pham, “Drug Response Prediction by Globally Capturing Drug and Cell Line Information in a Heterogeneous Network,” *J. Mol. Biol.*,

vol. 430, no. 18, Part A, pp. 2993–3004, Sep. 2018, doi: 10.1016/j.jmb.2018.06.041.

[92] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2012.

[93] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 10, 2015, *arXiv*: arXiv:1512.03385. doi: 10.48550/arXiv.1512.03385.

[94] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, Jan. 1989, doi: 10.1016/0893-6080(89)90020-8.

[95] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2014.

[96] T. Nguyen, H. Le, T. P. Quinn, T. Nguyen, T. D. Le, and S. Venkatesh, "GraphDTA: Predicting drug-target binding affinity with graph neural networks," *Bioinforma. Oxf. Engl.*, Oct. 2020, doi: 10.1093/bioinformatics/btaa921.

[97] Y.-C. Chiu *et al.*, "Predicting drug response of tumors from integrated genomic profiles by deep neural networks," *BMC Med. Genomics*, vol. 12, no. 1, p. 18, Jan. 2019, doi: 10.1186/s12920-018-0460-9.

[98] Greg Landrum., "RDKit: Open-source cheminformatics." Accessed: Mar. 20, 2021. [Online]. Available: https://www.rdkit.org/

[99] Bharath Ramsundar and Peter Eastman and Patrick Walters and Vijay Pande and Karl Leswing and Zhenqin Wu, *Deep Learning for the Life Sciences*. O'Reilly Media, 2019.

[100] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *Proc. Int. Conf. Learn. Represent. ICLR*, 2017.

[101] P. Velicković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *Proc. Int. Conf. Learn. Represent. ICLR*.

[102] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, “How Powerful are Graph Neural Networks?”,” *Proc. Int. Conf. Learn. Represent. ICLR*, 2019.

[103] J. Codony-Servat *et al.*, “Differential cellular and molecular effects of bortezomib, a proteasome inhibitor, in human breast cancer cells,” *Mol. Cancer Ther.*, vol. 5, no. 3, pp. 665–675, Mar. 2006, doi: 10.1158/1535-7163.MCT-05-0147.

[104] A. A. Friedman *et al.*, “Landscape of Targeted Anti-Cancer Drug Synergies in Melanoma Identifies a Novel BRAF-VEGFR/PDGFR Combination Treatment,” *PloS One*, vol. 10, no. 10, p. e0140310, 2015, doi: 10.1371/journal.pone.0140310.

[105] Vincent T. DeVita, Theodore S. Lawrence, Steven A. Rosenberg and Lippincott Williams & Wilkins, *Cancer, Principles and Practice of Oncology*. 2008.

[106] J. M. Corton, J. G. Gillespie, S. A. Hawley, and D. G. Hardie, “5-Aminoimidazole-4-Carboxamide Ribonucleoside,” *Eur. J. Biochem.*, vol. 229, no. 2, pp. 558–565, 1995, doi: 10.1111/j.1432-1033.1995.0558k.x.

[107] N. Aben, D. J. Vis, M. Michaut, and L. F. A. Wessels, “TANDEM: a two-stage approach to maximize interpretability of drug response models based on multiple molecular data types,” *Bioinforma. Oxf. Engl.*, vol. 32, no. 17, pp. i413–i420, Sep. 2016, doi: 10.1093/bioinformatics/btw449.

[108] B. Zagidullin *et al.*, “DrugComb: an integrative cancer drug combination data portal,” *Nucleic Acids Res.*, vol. 47, no. W1, pp. W43–W51, Jul. 2019, doi: 10.1093/nar/gkz337.

[109] M. Lukas *et al.*, “Survey of ex vivo drug combination effects in chronic lymphocytic leukemia reveals synergistic drug effects and genetic dependencies,” *Leukemia*, vol. 34, no. 11, Art. no. 11, Nov. 2020, doi: 10.1038/s41375-020-0846-5.

[110] T. Zhang, L. Zhang, P. R. O. Payne, and F. Li, “Synergistic Drug Combination Prediction by Integrating Multiomics Data in Deep Learning Models,” *Methods Mol. Biol. Clifton NJ*, vol. 2194, pp. 223–238, 2021, doi: 10.1007/978-1-0716-0849-4_12.

[111] Q. Liu and L. Xie, “TranSynergy: Mechanism-driven interpretable deep neural network for the synergistic prediction and pathway deconvolution of drug combinations,” *PLoS Comput. Biol.*, vol. 17, no. 2, p. e1008653, Feb. 2021, doi: 10.1371/journal.pcbi.1008653.

[112] J. Yang, Z. Xu, W. K. K. Wu, Q. Chu, and Q. Zhang, “Erratum to: GraphSynergy: a network-inspired deep learning model for anticancer drug combination prediction,” *J. Am. Med. Inform. Assoc.*, vol. 29, no. 1, p. 220, Jan. 2022, doi: 10.1093/jamia/ocab214.

[113] F. Meng, F. Li, J.-X. Liu, J. Shang, X. Liu, and Y. Li, “NEXGB: A Network Embedding Framework for Anticancer Drug Combination Prediction,” *Int. J. Mol. Sci.*, vol. 23, no. 17, Art. no. 17, Jan. 2022, doi: 10.3390/ijms23179838.

[114] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” May 19, 2016, *arXiv*: arXiv:1409.0473. doi: 10.48550/arXiv.1409.0473.

[115] T.-C. Chou and P. Talalay, “Quantitative analysis of dose-effect relationships: the combined effects of multiple drugs or enzyme inhibitors,” *Adv. Enzyme Regul.*, vol. 22, pp. 27–55, Jan. 1984, doi: 10.1016/0065-2571(84)90007-4.

[116] D. Wang *et al.*, “Combined inhibition of PI3K and PARP is effective in the treatment of ovarian cancer cells with wild-type PIK3CA genes,” *Gynecol. Oncol.*, vol. 142, no. 3, pp. 548–556, Sep. 2016, doi: 10.1016/j.ygyno.2016.07.092.

[117] Y. Yin *et al.*, “Chk1 inhibition potentiates the therapeutic efficacy of PARP inhibitor BMN673 in gastric cancer,” *Am. J. Cancer Res.*, vol. 7, no. 3, pp. 473–483, 2017.

[118] A. Chauhan, D. K. Semwal, S. P. Mishra, S. Goyal, R. Marathe, and R. B. Semwal, “Combination of mTOR and MAPK Inhibitors—A Potential Way to Treat Renal Cell Carcinoma,” *Med. Sci.*, vol. 4, no. 4, Art. no. 4, Dec. 2016, doi: 10.3390/medsci4040016.