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

TÓM TẮT LUẬN ÁN TIẾN SĨ KỸ THUẬT

HÀ NỘI - 2024

Công trình được hoàn thành tại: Học viện Công nghệ Bưu chính Viễn thông

Người hướng dẫn khoa học:

Phản biện 1:

Phản biện 2:

Phản biện 3:

Luận án được bảo vệ trước Hội đồng chấm luận án cấp học viện tại: Học viện Công nghệ Bưu chính Viễn thông

Vào hồi giờ 00', ngày tháng năm 2024

Có thể tìm hiểu luận án tại:

- **Thư viện Quốc gia Việt Nam**
- **Thư viện Học viện Công nghệ Bưu chính Viễn thông**

MỞ ĐẦU

1. GIỚI THIỆU

Mục tiêu quan trọng của y học chính xác là xác định được phương thức điều trị chính xác cho từng bệnh nhân dựa trên hồ sơ sinh học của họ. Trong những năm gần đây, dự đoán đáp ứng thuốc ngày càng thu hút nhiều nhà khoa học, phân tích dữ liệu y sinh học tham gia nghiên cứu và đề xuất các phương pháp mới nhằm cải tiến hiệu năng dự đoán và tìm ra các bằng chứng khoa học, góp phần vào sàng lọc và định hướng điều trị nhanh chóng hơn. Một loạt các nghiên cứu đã được đề xuất cho bài toán dự đoán đáp ứng thuốc. Tuy nhiên, với sự gia tăng ngày càng lớn các dữ liệu y sinh học, các phương pháp này vẫn còn một số hạn chế như: (1) các phương pháp này chưa tích hợp dữ liệu biểu diễn thuốc hoặc mới chỉ biểu diễn thuốc dưới dạng chuỗi hoặc ảnh, chưa biểu diễn dưới dạng tự nhiên hơn (như dạng dữ liệu đồ thị); (2) chưa tích hợp đa dạng các dạng dữ liệu đặc trưng sinh học (-omics) khác nhau; đồng thời chưa áp dụng các các phương pháp tính toán tiên tiến, phù hợp để cải thiện hiệu năng mô hình dự đoán. Luận án này đưa ra các giải pháp liên quan đến tích hợp dữ liệu biểu diễn thuốc theo dạng đồ thị phân tử và tích hợp đa dữ liệu -omics của dòng tế bào nhằm cải thiện hiệu năng dự đoán đáp ứng thuốc trong điều trị bệnh. Các giải pháp trong luận án tập trung vào ứng dụng các mô hình học sâu và khai phá dữ liệu y sinh học cho hai bài toán dự đoán đáp ứng thuốc là dự đoán đáp ứng đơn thuốc và dự đoán đáp ứng đa thuốc cho các dòng tế bào.

2. MỤC TIÊU CỦA LUẬN ÁN

Mục tiêu của luận án được đưa ra dựa trên các vấn đề chưa được giải quyết của bài toán dự đoán đáp ứng thuốc từ đó đề xuất một số giải pháp tính toán để cải thiện hiệu năng dự đoán đáp ứng thuốc đơn thuốc và đáp ứng đa thuốc. Cụ thể

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

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

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

- 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 để dự đoán đáp ứng đa thuốc cho các dòng tế bào
- Đề xuất giải pháp tích hợp đa dữ liệu -omics với thông tin cấu trúc mạng tương tác protein PPI để dự đoán đáp ứng đa thuốc cho các dòng tế bào

3. PHƯƠNG PHÁP NGHIÊN CỨU

Luận án vận dụng các phương pháp nghiên cứu cơ sở lý thuyết nền tảng, khảo sát các nghiên cứu liên quan, đưa ra các vấn đề còn tồn tại từ đó đề xuất giải pháp, xây dựng mô hình thực nghiệm và so sánh đánh giá kết quả. Trước tiên, luận án tổng hợp các lý thuyết nền tảng về dữ liệu sinh học (-omics), dữ liệu về đáp ứng thuốc và các phương pháp đã được đề xuất cho bài toán dự đoán đáp ứng đơn thuốc và đáp ứng đa thuốc đã được công bố. Từ đó đưa ra các vấn đề còn tồn tại và định hướng các giải pháp tính toán áp dụng nhằm nâng cao hiệu năng dự đoán đáp ứng thuốc trong điều trị bệnh. Các kịch bản thử nghiệm được triển khai với mỗi giải pháp đề xuất. Kết quả thực nghiệm được tiến hành và so sánh đánh giá với các nghiên cứu trước đây, đồng thời tìm các dấu ấn sinh học trong nghiên cứu lâm sàng

4. CÁC ĐÓNG GÓP CỦA LUẬN ÁN

Với việc nghiên cứu các phương pháp dự đoán đáp ứng thuốc trong điều trị bệnh, luận án đóng góp 4 giải pháp cho hai bài toán dự đoán đáp ứng đơn thuốc và dự đoán đáp ứng đa thuốc để nâng cao hiệu năng dự đoán.

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 này 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 chỉ ra được loại dữ liệu -omics có ý nghĩa cho mô hình dự đoán.

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.

5. BỐ CỤC CỦA LUẬN ÁN

Chương 1 – Tổng quan về đáp ứng thuốc và dự đoán đáp ứng thuốc

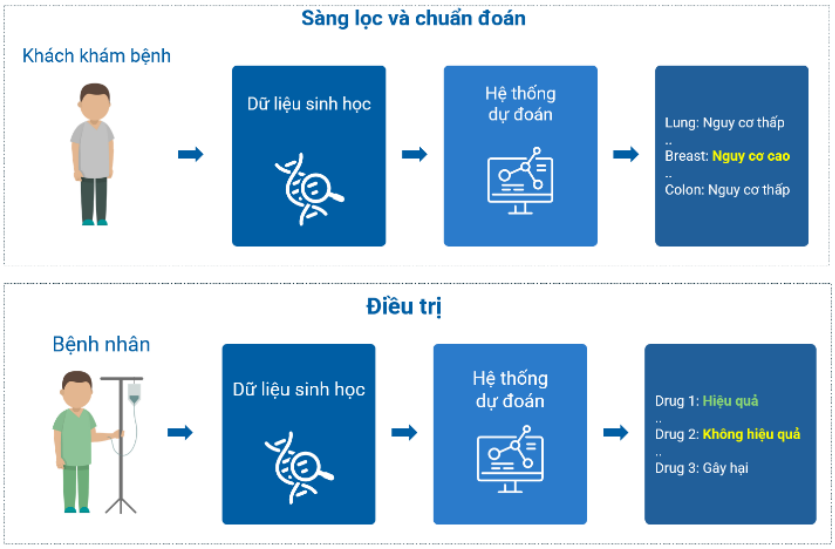
Chương 2 – Giải pháp tích hợp dữ liệu trong dự đoán đáp ứng đơn thuốc

Chương 3 – Giải pháp tích hợp dữ liệu trong dự đoán đáp ứng đa thuốc

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

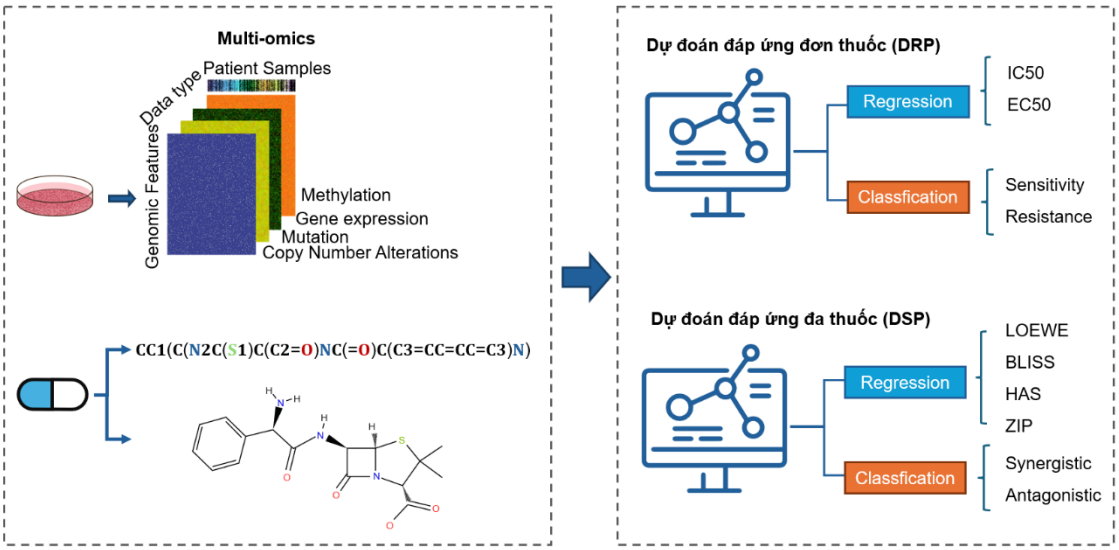
1.1. GIỚI THIỆU CHUNG

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. 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 chuẩn 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, chuẩn đoán và điều trị bệnh.



Hình 1. 1. Hệ thống tổng quan cho dự đoán đáp ứng thuốc

Hiện nay, một số mô hình nghiên cứu phổ biến thường liên quan đến bài toán dự đoán đáp ứng đơn thuốc (monotherapy) và dự đoán đáp ứng đa thuốc (combination therapy).



Hình 1. 2. Các mô hình dự đoán đáp ứng thuốc hiện nay

Trong đó dữ liệu đầu vào là dữ liệu -omics biểu diễn các loại dữ liệu khác nhau của dòng 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

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

1.2.1. Dữ liệu -omics

Các công nghệ -omics như genomics, transciptomics, epigenomics ra đời, cung cấp dữ liệu, tri thức mới về dữ liệu sinh học cho phép khám phá bộ gen, các hiện tượng sinh học trong cơ thể người đồng thời 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.

1.2.2. Dòng tế bào

Các dòng tế bào (cell line) là các khối tế bào bệnh sống được nuôi cấy trong các đĩa nuôi cấy mô trong phòng thí nghiệm, mang đầy đủ thông tin di truyền người bệnh, cung cấp nguồn dữ liệu quan trọng trong các nghiên cứu y sinh học.

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

Dữ liệu di truyền của một cá nhân là một trong những yếu tố quyết định ảnh hưởng đến tình trạng sức khỏe và bệnh tật con người. Có hai dữ liệu hệ gen quan trọng là đột biến gen (MUT) và biến thể số lượng bản sao (CNA).

1.2.4. Biểu hiện gen

Biểu hiện gen (GE) là quá trình chuyển đổi thông tin di truyền trong một gen được truyền vào cấu trúc đang có trong tế bào sống từ đó tính trạng tương ứng được tạo thành ở kiểu hình có thể quan sát được. 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 liệu những thay đổi trong hồ sơ phiên mã có ảnh hưởng đến sức khỏe và bệnh tật như thế nào.

1.2.5. Methyl hóa DNA

Dữ liệu methyl hóa (METH) là dữ liệu cho thấy sự thay đổi chức năng bộ gen dưới tác động ngoại sinh và thường xảy ra ở các đảo CpG (CpG island) trong DNA.

1.2.6. 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. Nghiên cứu dữ liệu trúc mạng tương tác PPI (interactomics) có thể giúp khám phá các đặc tính sinh học của protein hoặc phát triển thuốc nhắm mục tiêu.

1.2.7. Thuốc

1.2.7.1. Đáp ứng thuốc

Thuốc là hợp chất gây ra sự thay đổi trong sinh lý hoặc tâm lý của sinh vật khi được tiêu thụ. 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.

1.2.7.2. Phép đo đáp ứng thuốc

Độ đo phổ biến nhất là IC50: nồng độ thuốc làm chết một nửa số tế bào, tức làm giảm tỷ lệ sống của tế bào 50%. Ngoài IC50, một số độ đo khác cũng được sử dụng để đo độ đáp ứng thuốc như: AUC, EC50. Có thể phân loại đáp ứng thuốc thành đáp ứng (Sensitivity) và kháng thuốc (Resitance)

1.2.7.3. Kết hợp thuốc

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 (đa thuốc) 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. Có 4 độ đo kết hợp thuốc là LOEWE, BLISS, HAS, ZIP trong đó LOEWE được sử dụng rộng rãi hơn cả trong các phương pháp dự đoán đáp ứng đa thuốc. Có thể phân loại kết hợp thuốc thành hai loại là tương hợp (Synergistics) và tương kháng (Antagonistics)

1.2.7.4. Dữ liệu biểu diễn thuốc

SMILES (Simplified Molecular Input Line Entry System) 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 biểu diễn cấu trúc hóa học theo các quy tắc cú pháp cơ bản. 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).

Fingerprints (FP) là kỹ thuật biểu diễn dạng one-hot vector. Kiểu dữ liệu biểu diễn này nhược điểm chính cần dựa trên các quy tắc được định nghĩa trước, chúng thường có số chiều lớn (ví dụ: 881, 1024).

1.2.8. Nguồn dữ liệu y sinh học

Nguồn dữ liệu y sinh học cho dòng tế bào: CCLE, GDSC là hai nguồn dữ liệu quan trọng, 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 và hàng trăm thuốc khác nhau. Ngoài ra còn có các nguồn dữ liệu về thuốc như: CheMBL [42], ZINC [43], KEGG [34].

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

Trong những năm gần đây, các thuật toán học máy (ML), học sâu (DL) được áp dụng trong mọi lĩnh vực nói chung cũng như đối với lĩnh vực y sinh học nói riêng thì ngày càng có nhiều ý nghĩa trong việc phân loại, dự đoán bệnh, dự đoán đáp ứng thuốc trong điều trị bệnh một cách chính xác.

1.3.1. Mô hình học sâu

1.3.1.1. Mạng nơ-ron nhân tạo

Mạng nơ-ron nhân tạo (Artificial Neural Networks) là mạng mô phỏng lại mạng nơ-ron sinh học. Kiến trúc của mạng nơ-ron: lớp kết nối đầy đủ hay Fully Connected (FC) là kiến trúc hay được sử dụng nhất. Các hàm kích hoạt phổ biến: ReLU, LeakyReLU

1.3.1.2. Mạng nơ-ron tích chập

Mạng nơ-ron tích chập (Convolutional Neural Network - CNN) là một trong những mô hình học sâu ứng dụng trong các bài toán thị giác máy tính và nhiều lĩnh vực học máy khác nhau. Mỗi khối tích chập 1-chiều (1D Convolution) bao gồm nhiều bộ lọc. 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 kích hoạt (ví dụ, ReLU).

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). Để tổng hợp thông tin nút, mạng nơ-ron đồ thị thực hiện phương thức truyền thông điệp gồm 2 bước: tạo lập, kết tập thông điệp và cập nhật đỉnh đồ thị.

$$h_u^{(l+1)} = UPDATE^l\left(h_u^{(l)}, AGGREGATE^{(l)}\left(\left\{h_v^{(l)}, \forall\, v \in N(u)\right\}\right)\right)$$
$$= UPDATE^l\left(h_u^{(l)}, m_{N(u)}^{(l)}\right) \quad (1.1)$$

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 $h_u^{(l)}$.

1.3.1.4. Mạng nơ-ron tích chập đồ thị

Mạng nơ-ron tích chập đồ thị (Graph convolutional network – GCN) [49] 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 kết hợp thông tin đặc trưng của các đỉnh và cấu trúc đồ thị để thực hiện phân loại hoặc dự đoán trên đồ thị.

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. , các lớp tích chập đồ thị sẽ cập nhật theo công thức sau:

$$\tilde{D}^{\frac{1}{2}} \tilde{A} \tilde{D}^{\frac{1}{2}} XW \quad (1.11)$$

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. GAT (Graph attention network) 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. $\alpha_{i,j}$ là hệ số attention được định nghĩa:

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

Với một cơ chế multi-head GAT layer, lớp GAT cuối cùng có thể được viết là có thể được biểu diễn như sau:

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

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

1.3.1.6. Mạng nơ-ron đồ thị đẳng cấu

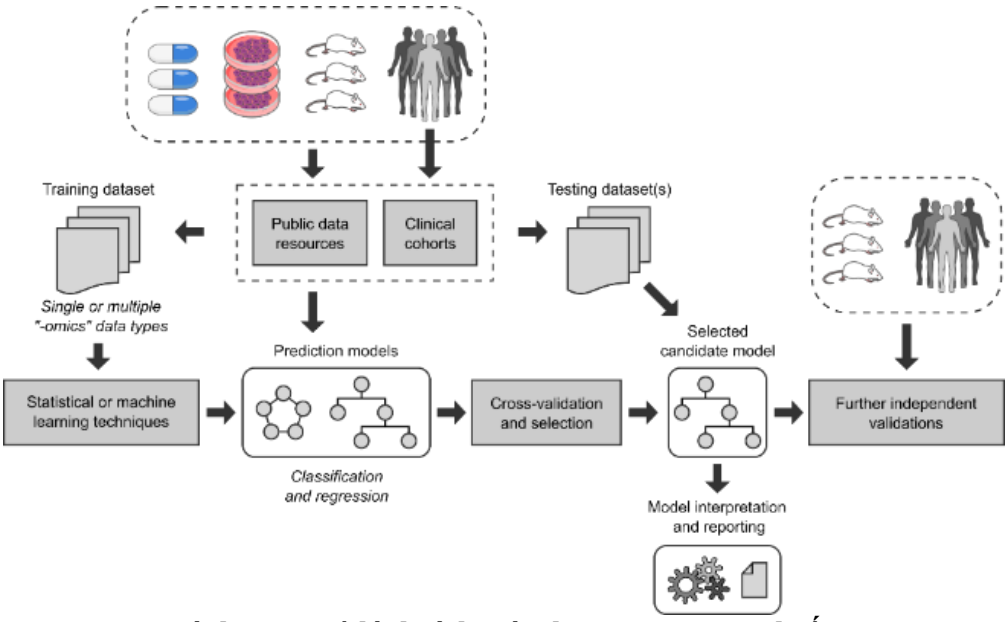
Mạng nơ-ron đồ thị đẳng cấu (GIN) [51] cho phép phân biệt các đồ thị không đẳng cấu với nhau, hay có thể phân biệt cấu trúc đồ thị khác nhau. Sau khi mô hình đã được huấn luyện, nó có thể được sử dụng để tính toán tính đẳng cấu (tương đồng) giữa các đồ thị. Tính tương đồng có thể được đo lường bằng cách so sánh đặc trưng của các đỉnh tương ứng trong các đồ thị. GIN sử dụng một cơ chế cập nhật đồ thị để tính toán vector 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. 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 các giá ϵ là một giá trị được định nghĩa sẵn, N(i) là các lân cận của nút i, $x_i^{(l)}$ 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.

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

Mô hình nghiên cứu phổ biến thường liên quan đến bài toán dự đoán đáp ứng thuốc cho từng thuốc đơn (monotherapy) và dự đoán cho kết hợp thuốc (combination therapy). Các mô hình tính toán đều dựa trên mô hình học có giám sát (Hình 1.22).



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 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 và phân loại. Một loạt các kỹ thuật dựa trên các phương pháp học máy đã được đề xuất [9], [10], [54], [55]. Tuy nhiên không có cách tiếp cận nào có thể vượt trội so với các phương pháp khác trên các tập dữ liệu khác nhau và trên các loại thuốc khác nhau; việc lựa chọn bộ dữ liệu mẫu và kích thước bộ dữ liệu đóng vai trò quan trọng trong mô hình dự đoán. Các hướng nghiên cứu này 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] tuy nhiên khó có thể dự đoán cho các thuốc hoặc bệnh mới. 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 đã được đề xuất [21], [22], [40], [59], [60]. Tuy nhiên các hướng này mới áp dụng đặc trưng thuốc dưới dạng chuỗi hoặc ảnh, có thể coi là các hướng "no-graph". Một số phương pháp gần đây đã cải tiến cách biểu diên dữ liệu thuốc dạng "graph" hoặc bổ sung thêm lớp transformer trong mô hình tính toán biểu diễn đặc trưng dữ liệu được đề xuất [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 kết hợp thuốc

Đã có một số nghiên cứu đề xuất dựa trên mô hình học máy cơ bản để dự đoán đáp ứng thuốc phối hợp nhằm dự đoán đáp ứng đa thuốc (cặp thuốc) 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]. 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] có thể coi là nghiên cứu đầu tiên đề xuất việc sử dụng DL để dự đoán tác dụng phối 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[73], [75] đã áp dụng graph trong việc học các dữ liệu đồ thị phân tử thuốc cho thấy hiệu quả tiềm năng của dự đoán.

1.3.2.3. Phương pháp tích hợp dữ liệu

Các thách thức cho chiến lược tích hợp dữ liệu multi-omics là tích hợp các dữ liệu khác nhau đó như thế nào. Hiện nay có các hướng chính để tích hợp dữ liệu gồm: tích hợp sớm (early integration), tích hợp muộn (late integration).

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. Cách tiếp cận này, về mặt lý thuyết có thể tổng hợp tốt các đặc trưng dữ liệu bởi, tuy nhiên hướng này không xem xét đến các dữ liệu phân bố khác nhau trong các dữ liệu -omics khác nhau, làm tăng chiều dữ liệu.

1.3.2.3.2. Mô hình tích hợp muộn

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, hoạt động với một phân phối duy nhất của mỗi dữ liệu omics.

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 thể đánh giá không chỉ mang tính tổng quát hóa mà còn được đánh giá trên các trường hợp dự đoán cho thuốc mới và dòng tế bào mới. Phân chia bộ dữ liệu thử nghiệm có thể phân chia một tỷ lệ nhất định như (80:10:10). Các chỉ số đánh giá mô hình: RMSE, Pearson (CCp) cho các mô hình hồi quy. Trong khi các mô hình phân loại thường dùng các chỉ số như accuracy, precision, recall, F1-score, ROC, AUC.

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

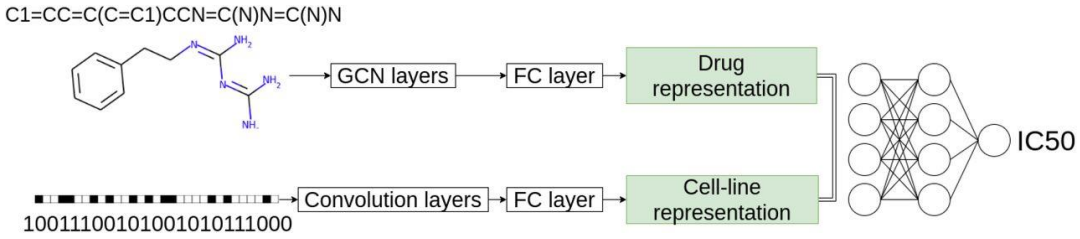
2.1. GIỚI THIỆU CHUNG

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 như tCNNs [20], DeepDR [99], MOLI [20]. Trong đó tCNNs [20] 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), tuy nhiên 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. DeepDR và MOLI, là hai phương pháp tiên tiến tích hợp 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, các đặc trưng -omics mới chỉ áp dụng là dữ liệu đột biến gen và biểu hiện gen chưa tích hợp đa dạng các dữ liệu -omics khác như methyl hóa. Để cải tiến hai vấn đề trên, luận án đề xuất hai giải pháp pháp: (1) GraphDRP - dự đoán đáp ứng thuốc dựa trên một số mô hình mạng nơ-ron đồ thị tích chập; (2) GraOmicDRP – dự đoán đáp ứng thuốc dựa trên mô hình tích hợp đa dữ liệu –omics và dữ liệu đồ thị phân tử thuốc.

2.2. ĐỀ XUẤT GIẢI PHÁP HỌC DỮ LIBIỂU DIỄN ĐỒ THỊ CỦA PHÂN TỬ THUỐC - GraphDRP

2.2.1. 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 (CNV).



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 chuyển đổi thành dữ liệu dạng đồ thị 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 (Hình 2.4). Kết quả là, một đồ thị với các nút được phân bổ đã được xây dựng cho mỗi chuỗi SMILES đầu vào và biến đổi thành 128 chiều biểu diễn dữ liệu thuốc.

Bộ dữ liệu

Dữ liệu dạng nhị phân của 948 dòng tế bào ung thư từ 13 mô, biểu diễn đột biến gen (MUT) và biến thể số lượng bản sao (CNV) được tổng hợp từ GDSC phiên bản 6.0. Mỗi dòng tế bào có 735. - Bộ dữ liệu gồm 223 thuốc, mỗi thuốc biểu diễn dưới dạng một chuỗi ký tự hóa học theo chuẩn CanonicalSMILES. Dữ liệu đáp ứng thuốc được chuẩn hóa về khoảng (0.1), mỗi phân tử biểu diễn bởi vector one-hot 78 chiều.

2.2.2. Kịch bản thử nghiệm

Mixed: Thử nghiệm này đã đánh giá tổng quát hiệu năng dự đoán của các mô hình trên các thuốc - dòng tế bào đã biết. Các cặp thuốc – dòng tế bào đã biết được chia ngẫu nhiên theo tỉ lệ 80:10:10 tương ứng cho các tập huấn luyện, tập đánh giá và tập kiểm tra đảm bảo phân phối tương đồng trên các tập dữ liệu này.

Blind-Drug: là các thử nghiệm dự đoán đáp ứng cho các thuốc mới (Blind-Drug), Các thuốc mới chỉ có 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 IC50 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.

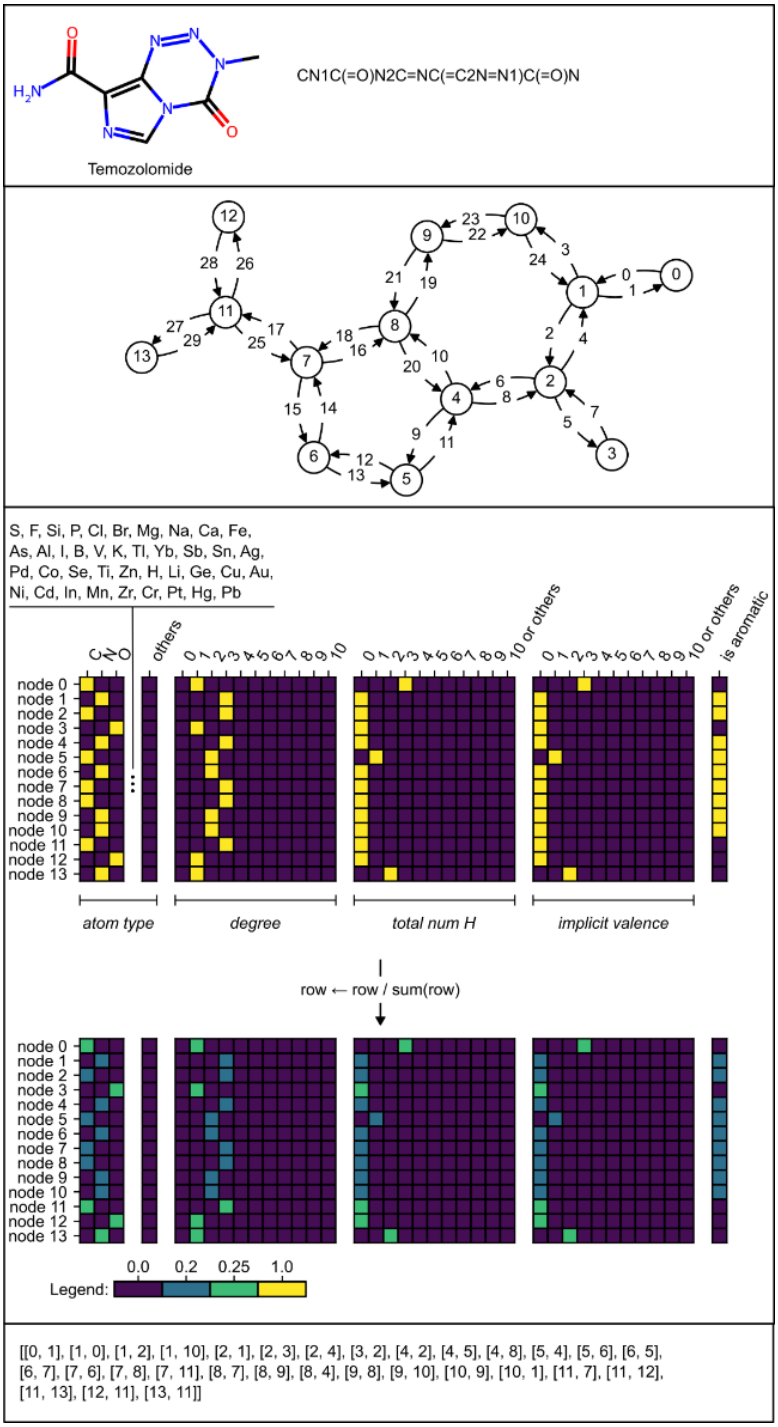
Blind-Cellline: tương tự như Blind-Drug, 10% dòng tế bào mới không có trong bộ dữ liệu huấn luyện và đánh giá, được đưa vào tập dữ liệu thử nghiệm.

Phép đo hiệu năng mô hình: Mô hình sử dụng hai độ đo RMSE và CCp.

2.2.3. Cài đặt mô hình

Mô hình áp dụng một số thực nghiệm trên một số mô hình mạng nơ-ron đồ thị tiên tiến như: GCN (3 lớp), GAT (2 lớp), GIN (5 lớp), GCN-GAT. Bên cạnh đó, mạng nơ-ron tích chập một chiều (CNN1D) được dùng để học các đặc trưng ẩn từ các đặc trưng ban đầu của bộ gen. Cuối cùng, các vecto vec-tơ này được kết nối và đưa vào khối dự đoán ((FC), để dự đoán đáp ứng thuốc cho dòng tế bào.

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



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

2.2.4. Kết quả và đánh giá

Kết quả thử nghiệm cho thấy mô hình đề xuất có hiệu năng vượt trội hơn so với mô hình cơ sở tCNNs trên tất cả các thử nghiệm với cả hai độ đo là CCp và RMSE.

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

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

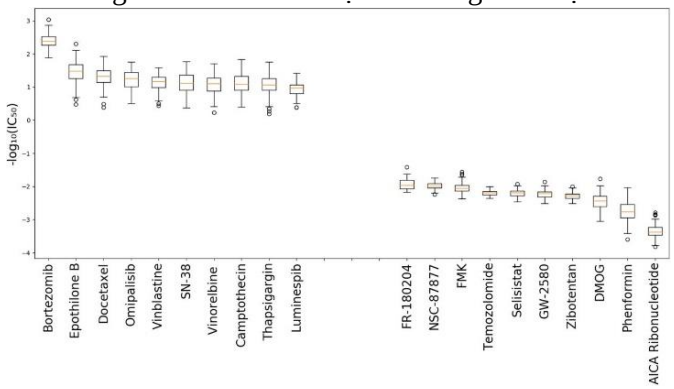
Methods		RMSE	CCp
tCNNs		0.0680	0.0617
GraphDRP	GCN	**0.0542**	**0.3241**
	GIN	0.0602	0.0481
	GAT	0.0616	0.2751
	GCN-GAT	0.0610	0.1683

Trong thí 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). Với thí nghiệm Blind-Cellline, Bảng 2.4 một lần nữa cho thấy sự vượt trội của GraphDRP so với tCNNs, tương đồng với 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 coi GIN là mô hình tốt nhất.

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

Dự đoán giá trị đáp ứng cho các cặp thuốc – dòng tế bào chưa biết: Trong thử nghiệm này, mô hình tốt nhất được huấn luyện về thử nghiệm Mixed-test (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 IC50 dự đoán cao nhất và thấp nhất. Điều đáng chú ý là ba loại thuốc đầu tiên có giá trị IC50 cao nhất và thấp nhất đều có kết quả tương tự như trong dự đoán của mô hình tCNNs. Thí nghiệm này cho thấy Bortezomib đạt IC50 thấp nhất. Ngược lại, AICA Ribonucleotide và Phenformin có IC50 cao nhất, có nghĩa là bệnh ung thư ít nhạy hơn với các loại thuốc này. Các dự đoán này phù hợp với các nghiên cứu trên dữ liệu lâm sàng tìm được.



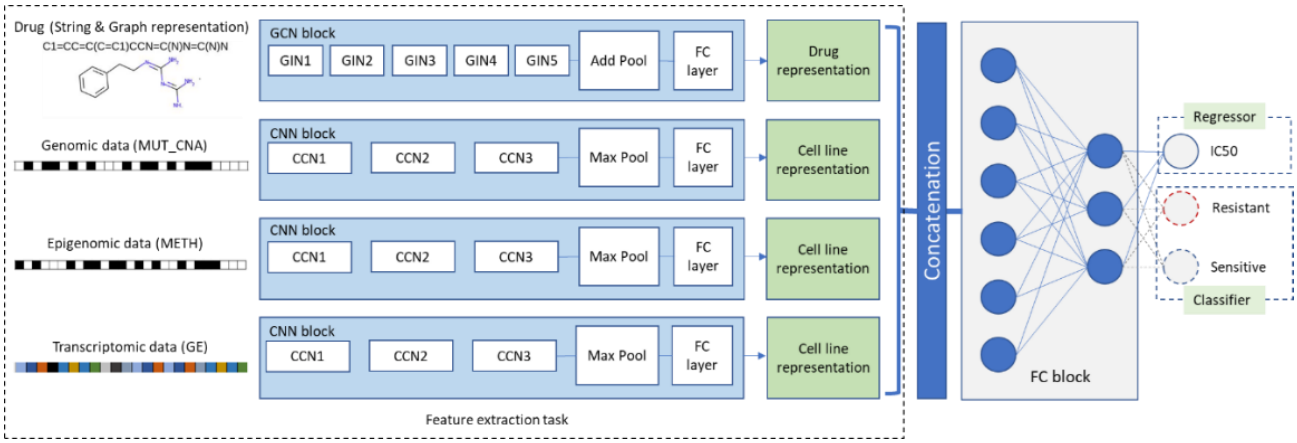
Hình 2.7. Biểu đồ 10 thuốc có giá trị IC50 được dự đoán tốt nhất và thấp nhất cho các cặp thuốc – dòng tế bào chưa biết

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 tốt hơn so với cách biểu diễn dữ liệu phân tử thuốc dạng chuỗi (tCNNs) trên tất cả các kịch bản thử nghiệm.

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

2.3.1. Phương pháp GraOmicDRP

Đề xuất này là một cải tiến của GraphDRP nhằm tích hợp không chỉ dữ liệu biểu diễn thuốc dạng đồ thị mà còn tích hợp ba dữ liệu -omics khác nhau gồm: dữ liệu đột biến gen (MUT) và biến thể số lượng sao chép (CNA); dữ liệu biểu hiện gen (GE) và dữ liệu methyl hóa (METH) của dòng tế bào. GIN được triển khai để học các biểu diễn thuốc và biến đổi thành vector128 chiều. Khối CNN1D được áp dụng để học biểu diễn các dòng tế bào tương ứng (Hình 2.8). Mô hình này có thể được chuyển đổi thành mô hình dự đoán phân loại đáp ứng /kháng thuốc.



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

Bộ dữ liệu:

Bộ dữ liệu MUT_CNA, và đáp ứng thuốc sử dụng tương tự như trong giải pháp GraphDRP, ngoài ra giải pháp tổng hợp thêm hai dữ liệu –omics, bao gồm biểu hiện gen (GE), và dữ liệu methyl hóa (METH) của các 1018 và 790 dòng tế bào với dữ liệu methyl hóa ở dạng [0,1], GE là các 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 chuẩn hóa cho các tập đơn -omics và đa -omics

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

Kịch bản thử nghiệm: Kịch bản thử nghiệm được thực hiện theo ba loại: Mixed test, 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.

2.3.2. Cài đặt mô hình thử nghiệm

Mô hình thực hiện năm lớp GIN liên tiếp được sử dụng để học các biểu diễn thuốc, mỗi nhánh CNN gồm 3 lớp CNN1D theo sau là lớp max pooling và lớp kết nối đầy đủ để tổng hợp biểu diễn dòng tế bào. Khối dự đoán gồm hai lớp FC (1024,128) theo sau mỗi lớp là hàm kích hoạt ReLU và dropout (0.2). Các tham số: Learning rate: 0.001; Batch size: 1024; epoch: 300 được tinh chỉnh trong quá trình huấn luyện.

2.3.3. Kết quả và đánh giá

Kịch bản Mixed

Kết quả cho thấy hầu hết mô hình tích hợp đa dữ liệu -omics cho hiệu năng tốt hơn tích hợp đơn -omics (Bảng 2.7). Trong đó mô hình tích hợp đơn -omics với GE đạt hiệu năng tốt nhất về cả RMSE (0,0259) và CCp (0,9195). Với mô hình tích hợp đa dữ liệu -omics thì sự kết hợp giữa dữ liệu biểu hiện gen và methyl hóa (GE & METH) cho hiệu năng tốt nhất với RMSE (0,0239) và CCP (0,9310). Ngoài ra sự kết hợp không có GE (METH & MUT_CNA) lại đạt được hiệu năng kém hơn so với các cặp kết hợp có GE khác. Điều này một lần nữa chỉ ra rằng dữ liệu biểu hiện gen là thông tin mang nhiều ý nghĩa 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 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). Kết quả cho thấy sự kết hợp GE & METH cũng đạt hiệu năng cao nhất, tiếp theo là bộ kết hợp 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 thử nghiệm Mixed

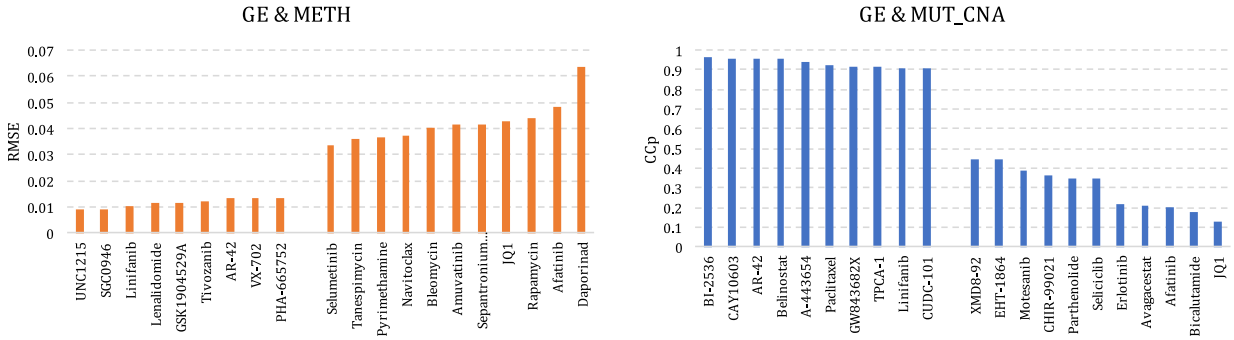
Methods		RMSE	CCp
Single -omics	METH	0.0278 (± 0.0087)	0.577 (± 0.1483)
	GE	**0.0254 (± 0.0073)**	**0.6603 (±0.1531)**
	MUT_CNA	0.0244 (± 0.0088)	0.6512 (± 0.1595)
Multi -omics	GE & METH	**0.0225 (± 0.0070)**	**0.7084 (± 0.1584)**
	GE & MUT_CNA	0.0229 (± 0.0072)	0.69 (± 0.1534)
	METH & MUT_CNA	0.0236 (± 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 trái) và GE & MUT_CNA (Hình 2.13 phải) tương ứng về RMSE và CCp.

Kịch bản Blind-Cellline: So sánh với GraphDRP, Bảng 2.9 cho thấy Gra-OmicDRP có hiệu năng đánh giá trên chỉ số RMSE nhỏ hơn (0,0327) và CCP cao hơn (0,8766) hiệu năng tương ứng của 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**



Hình 2.12 (trái), Hình 2.13. (phải) Mười thuốc có hiệu năng dự đoán cao nhất trên chỉ số RMSE và CCp

Kịch bảnBlind-Drug: Bảng 2.10. cho thấy rằng phương pháp nghiên cứu đạt được hiệu năng tương đương về mặt RMSE (nghĩa là 0,0542 và 0,0590 đối với GraphDRP-GCN và GraOmicDRP, tương ứng), nhưng tốt hơn về mặt 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**

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 hiệu quả 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.

So sánh hiệu năng dự đoán giữa GraOmicDRP và hai phương pháp tiên tiến khác 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 là DeepDR và MOLI, GraOmicDRP cũng cho kết quả vượt trội hơn (Bảng 2.11 và bảng 2.12). Trong đó GraOmicDRP khá tương đồng với nghiên cứu DeepDR khi cho thấy dữ liệu biểu hiện gen (GE) có hiệu năng tốt hơn so với tập dữ liệu đột biến (MUT) trong dự đoán đáp ứng thuốc. Bảng 2.12. cho thấy rằng GraOmicDRP vượt trội hơn MOLI đối với tất cả các loại thuốc trên chỉ số đánh giá AUC.

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**

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

2.4. KẾT LUẬN CHƯƠNG

Chương này trình bày nội dung hai giải pháp để dự đoán đáp ứng đơn thuốc: GraphDRP và GraOmicDRP. Trong đó GraphDRP là giải pháp học dữ liệu biểu diễn đồ thị phân tử thuốc thông qua một số biến thể của mạng nơ-ron đồ thị kết hợp dữ liệu biểu diễn dòng tế bào qua mạng nơ ron tích chập một chiều.

Kết quả thử nghiệm cũng xác định được mô hình mạng nơ-ron đồ thị có hiệu quả nhất trong mô hình 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 với đề xuất GraOmicDRP. Giải pháp này cho thấy hiệu năng dự đoán của mô hình tích hợp đa dữ liệu -omics vượt trội hơn với mô hình tích hợp đơn -omics, đồng thời xác định được loại dữ liệu -omics (GE) có ý nghĩa trong mô hình dự đoán. Nội dung hai giải pháp này là kết quả của hai công trình công bố số 1 và số 2.

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

3.1. GIỚI THIỆU CHUNG

Kết hợp thuốc để giảm các tác dụng phụ và độc tính cũng như khả năng kháng thuốc của quá trình điều trị đơn thuốc [110], [111]. Một số phương pháp học máy, học sâu đã đã được đề xuất cho bài toán này để dự đoán đáp ứng đa thuốc cho các dòng tế bào [68], [72], [73], [74]. Tuy nhiên, các phương pháp này chưa giải quyết được vấn đề tích hợp nhiều dữ liệu -omics hay biểu diễn cặp thuốc kết hợp dưới dạng tự nhiên để dự đoán đa thuốc cũng như chưa tổng hợp thông tin cặp tương tác thuốc một cách tư nhiên hơn.

Do đó trong chương này luận án trình bày hai đề xuất cho dự đoán kết hợp thuốc là GraOmicSynergy và AE-XGBSynergy tích hợp đa dạng các biểu diễn -omics khác nhau. Do đó, trong chương này, luận án trình bày hai đề xuất tích hợp đa dữ liệu -omics cho bài toán dự đoán đáp ứng đa thuốc: (1) đề xuất mô hình GraOmicSynergy học các biểu diễn thuốc dưới dạng đồ thị và tăng cường cơ chế chú ý để tổng hợp biểu diễn kết hợp thuốc đối với các dòng tế bào, từ đó tích hợp nhiều dữ liệu -omics của các dòng tế bào để dự đoán đáp ứng đa thuốc; (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 đáp ứng đa thuốc.

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

3.2.1. Phương pháp

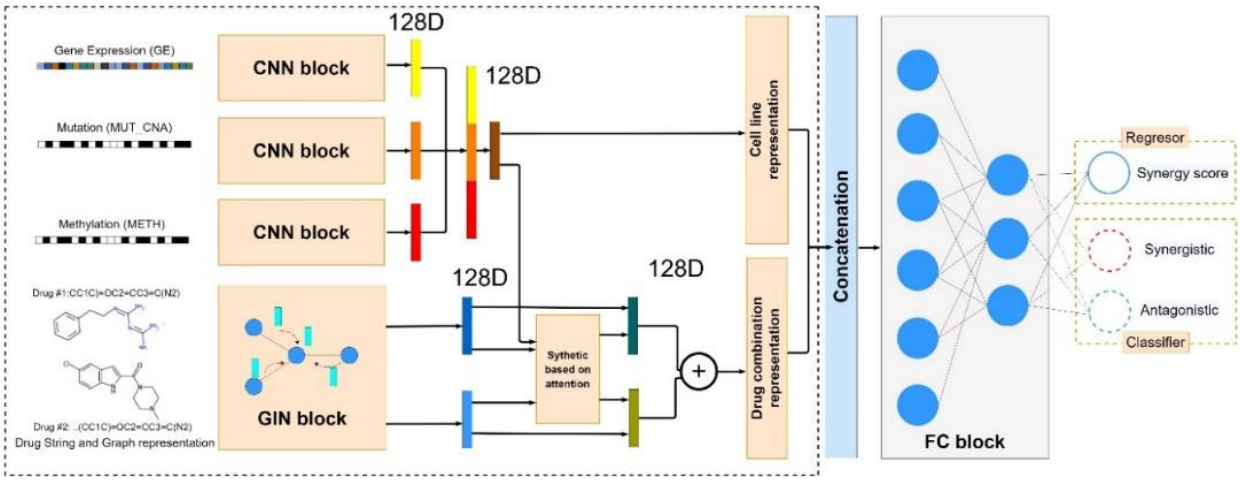
Kế thừa phương pháp tích hợp multi-omics và dữ liệu biểu diễn thuốc dưới dạng đồ thị từ nghiên cứu trước GraOmicDRP cho dự đoán liệu trình đơn thuốc, nghiên cứu này tiếp tục đề xuất phương pháp tích hợp này cho bài toán dự đoán kết hợp thuốc cho dòng tế bào với tên là GraOmicSynergy thực hiện việc kết 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.

Mô đun chức năng "Synthetic based on attention" được xây dựng để tính các đóng góp khác nhau của mỗi thuốc trong tổ hợp tương tác thuốc – dòng tế bào – thuốc. 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 $c_{n,}$



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

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 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 vector 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): 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): chỉ số đánh giá như: độ chính xác (ACC), Precision (PREC), Recall và F1-score (F1).

3.2.2. Cài đặt mô hình thử nghiệm

Tổng hợp bộ dữ liệu

Bộ dữ liệu đáp ứng đa thuốc o O'Neil [77] 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, tổng hợp có 23.062 tương tác cặp thuốc – dòng tế bào. Bộ dữ liệu biểu diễn hồ sơ sinh học multi-omics cho dòng tế bào bệnh là bộ dữ liệu GDSC [45] bao gồm 3 dữ liệu -omics (GE, MUT_CNA, METH) được kế thừa và chuẩn hóa theo Đề xuất 2 – GraOmicDRP . Tổng hợp bộ dữ liệu thực nghiệm thu được gồm: 38 loại thuốc trong đó 37 thuốc đại diện cho mỗi thuốc trong cặp kết hợp thuốc; 25 dòng tế bào thuộc 5 loại mô bệnh (tissues) gồm với 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, đồng thời hiệu năng dự đoán của GraOmicSynergy cũng được đánh giá cho cả dữ liệu –omics đơn và kết hợp đa -omics cho cả dữ liệu –omics đơn và kết hợp đa omic. 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 cho việc kiểm định điều chỉnh tham số mô hình (validation set) và kiểm tra (testing set) của mỗi kịch bản riêng được chia độc lập, không trùng lặp nhau. 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 mạng, nghiên cứu đã kết hợp cả hai cách trình bày của từng mẫu trong tất cả các dữ liệu cho training, testing và validation. 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: năm lớp GIN 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 vector 128 chiều. 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: Các tham số: Learning rate: 0.001; Batch size: 1024; epoch: 300 được tinh chỉnh trong quá trình huấn luyện.

3.2.3. Kết quả và đánh giá

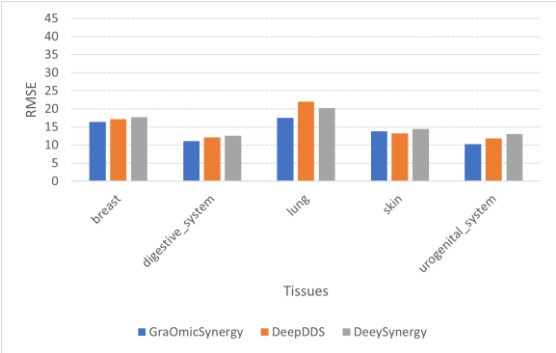
Kịch bản Mixed.

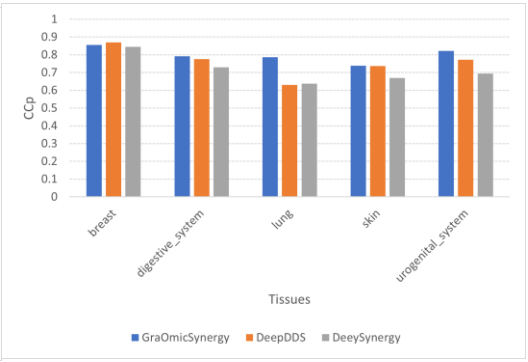
GraOmicSynergy được thực hiện đánh giá trên cả single-omics và multi-omics. Kết quả thực nghiệm đánh giá trên toàn bộ tập dữ liệu thử nghiệm cho thấy tổ hợp ALL gồm 3 -omics (GE & MUT_CNA & METH) đạt hiệu quả tốt nhất về cả RMSE (14,025) và CCp (0,794) (chi tiết ở bảng 3.2); trong khi đó, giá trị của 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. Kết quả này chỉ ra rằng sự tích hợp đa -omics là có thể nâng cao hiệu năng dự đoán. Dữ liệu biểu hiện gen (GE) tiếp tục chứng minh đây là dữ liệu có đóng góp quan trọng trong mô hình tích hợp -omics cho bài toán dự đoán kết hợp thuốc tương tự như bài toán tích hợp dữ liệu trong dự đoán đáp ứng đơn thuốc trên.

Tiến hành so sánh hiệu năng dự đoán các phương pháp trên các loại bệnh, Hình 3.2 và Hình 3.3 cho thấy so với DeepSynergy và DeepDDS, GraOmicSynergy (ALL) đã đạt được RMSE tốt nhất cho tất cả các mô, 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), với mô còn lại (breast cancer).

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**





Hình 3.2. (trái), Hình 3.3. (phải) Hiệu năng dự đoán các mô bệnh trên đánh giá RMSE và CCp theo kịch bản Mixed

Kịch bản Blind-Cellline

So sánh hiệu năng trên chỉ số về cả CCp và RMSE đối với thử nghiệm Blind-Cellline cho thấy giải pháp đề xuất đạt hiệu năng tốt hơn DeepSynergy và DeepDDS trong hầu hết các thử nghiệm. Đặc biệt, với bộ dữ liệu đơn -omics, GE – đạt hiệu năng tốt nhất trong cả hai chỉ sốcủa CCp và RMSE. Bảng 3.3 cho thấy GraOmicDRP lưu trữ RMSE nhỏ hơn (20,73) và CCP cao hơn (0,484) so với DeepDDS và Deepsynergy.

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

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

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 GraOmicDCSP tốt hơn DeepDDS và DeepSynergy cho tất cả các tình huống thử nghiệm. Điều này 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.

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**

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

Bảng 3.5 so sánh 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 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**

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 với một số bằng chứng từ các nghiên cứu trước đây về khả năng kết hợp các cặp thuốc này trong điều trị ung thư. Ví dụ như 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 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].

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. Mô hình đề xuất GraOmicSynergy hoạt động hiệu quả hơn với các mô hình tiên tiến như DeepDDS và DeepSynergy trong hầu hết các kịch bản dự đoán độ kết hợp thuốc.

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

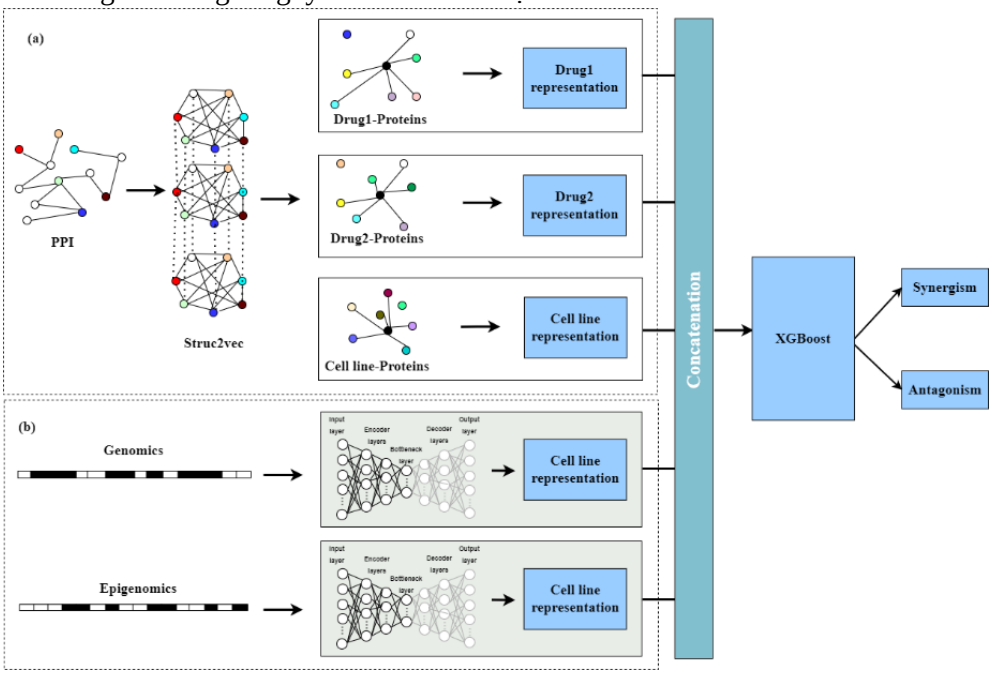
3.3. ĐỀ XUẤT GIẢI PHÁP TÍCH HỢP ĐA DỮ LIỆU -OMICS VÀ MẠNG SINH HỌC - AE-XGBSynergy

3.3.1. Phương pháp

AE-XGBSynergy 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 (Hình 3.3).

AE-XGBSynergy gồm hai phần được mô tả trong hình 3.3. 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.



Hình 3.3. Mô hình đề xuất dự đoán đáp ứng đa thuốc - AE-XGBSynergy

.

Quá trình trích xuất đặc trưng của mạng PPI gồm các bước:

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

(2) *Xây dựng đồ thị có trọng số nhiều lớp:* tính trọng số cạnh giữa mỗi cặp protein (u,v) và trọng số cạnh giữa các lớp:

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

(4) *Trích xuất biểu diễn của các dòng tế bào từ dữ liệu -omics*

Với Di và Cj được ký hiệu là vec-tơ biểu diễn của thuốc i và dòng tế bào j tác động lên protein, P_{Di}, P_{Cj} là vec-tơ biểu diễn protein tương tác với thuốc Di và dòng tế bào Cj tương ứng

$$Di = \frac{P_{D1} + P_{D2} + \cdots + P_{Dn}}{n} \quad (3.4)$$

$$Cj = \frac{P_{C1} + P_{C2} + \cdots + P_{Cn}}{n} \quad (3.5)$$

Để trích xuất đặc trưng của các dòng tế bào từ dữ liệu biểu diễn hệ gen (genomics) và dữ liệu methyl hóa (epigenomics). 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 vector 64 chiều. Các vector này được ghép nối thành vector 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 tính năng của dòng tế bào cho quá trình dự đoán.

Phương pháp đánh giá mô hình

Hiệu năng của AE-XGBSynergy thực hiện xác thực chéo năm lần trên tập huấn luyện. Thực nghiệm được đánh giá trên cả kịch bản dữ liệu đơn dữ liệu -omics và đa -omics tích hợp theo sáu chỉ số đánh giá bao

dự đoán phân loại bao gồm độ chính xác (ACC), Recall, AUC-ROC, AUC-PR, độ chính xác (PRE) và F1-score.

3.3.2. Cài đặt và thực nghiệm mô hình

Bộ dữ liệu

Kế thừa tập dữ liệu multi-omics từ các đề xuất ở chương 2, giải pháp đề xuất 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 dự đoán. Bộ dữ liệu tương tác thuốc-protein (Drug-Protein association) và Cell line-protein áp dụng tương tự trong nghiên cứu NEXGB. Cụ thể gồm: 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; bộ dữ liệu mạng tương tác 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; Bộ dữ liệu kết hợp thuốc: O'Neil dataset kết hợp thuốc tính toán theo chỉ số Loewe và DrugCombDB, 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

3.3.3. 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 và đa -omic.

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

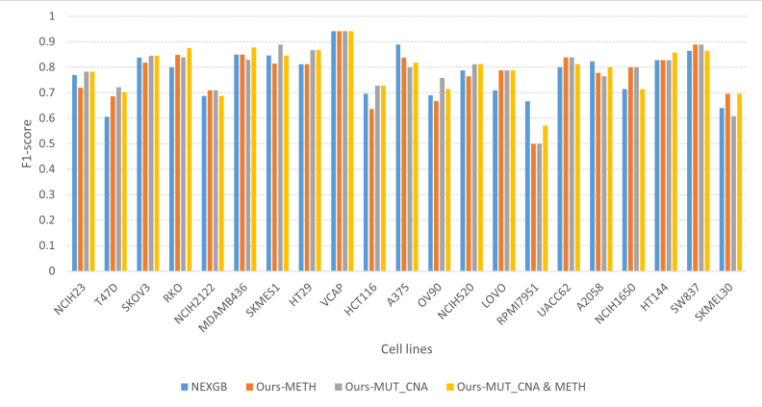
Methods	-Omics	Accuracy	Recall	AUC-ROC	AUC-PR	Precision	F1-score
NEXGB		0.7702	0.7919	0.7688	0.7283	0.7798	0.7858
AE-XGBSynergy	**METH**	0.7736	0.8012	0.7716	0.7303	0.7795	0.7902
	MUT_CNA	0.7884	**0.8106**	0.7869	0.7458	0.7957	0.8031
	MUT_CNA & METH	**0.7901**	**0.8106**	**0.7887**	**0.7478**	**0.7982**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

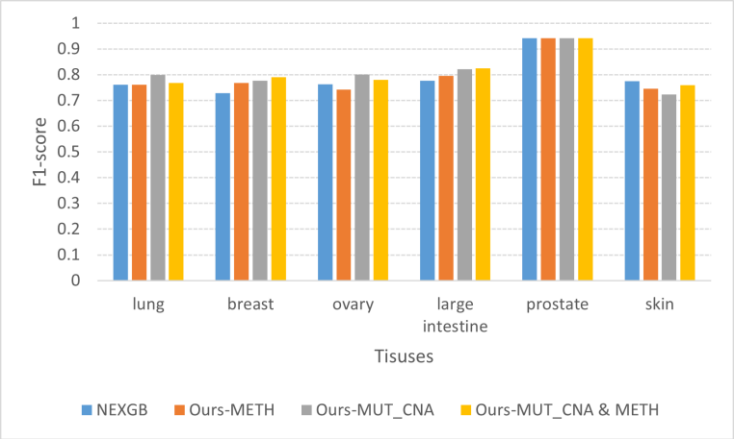
Cụ thể, hiệu năng của AE-XGBSynergy trên cả hai tập dữ liệu cho thấy MUT_CNA là dữ liệu đóng góp nhiều thông tin quan trọng hơn METH trong việc tích hợp các single - omics trong để dự đoán. Hơn nữa, sự kết hợp giữa MUT_CNA & METH đã đạt được hiệu năng tốt nhất trong bộ dữ liệu O'Neil xét về tất cả sáu chỉ số hiệu năng. 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.

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 -omics trong bộ dữ liệu O'Neil. Đáng chú ý, trong số 21 dòng tế bào được phân tích, RPMI7951 thuộc mô da có hiệu năng thấp nhất với142 protein mục tiêu, VCAP cao nhất với 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 3.4. So sánh hiệu năng dự đoán cho dòng tế bào trên bộ dữ liệu O'Neil

So sánh hiệu năng dự đoán trên từng mô bệnh, Hình 3.5, trong hầu hết các mô (4 / 5), phương pháp đề xuất của vượt trội hơn NEXGB ở các giải pháp tích hợp khác nhau. Có thể thấy AE-XGBSynergy đưa ra một cách tiếp cận đầy hứa hẹn 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.



Hình 3.5. So sánh hiệu năng dự đoán cho từng mô bệnh trong bộ dữ liệu O'Neil

3.4. KẾT LUẬN CHƯƠNG

Tiếp tục giải pháp tích hợp dữ liệu, trong chương 3 này, luận án trình bày hai đề xuất mới áp dụng các phương pháp tích hợp dữ liệu sinh học khác nhau để dự đoán khả năng đáp ứng đa thuốc với dòng tế bào với hai đề xuất GraOmicSynergy và AE-XGBSynergy. Nội dung hai giải pháp này là kết quả của hai công trình công bố số 5 và số 4.

Trong đó, GraOmicSynergy tích hợp dữ liệu biểu diễn tổng hợp đa dữ liệu -omics khác nhau như GE, MUT_CNA và METH và dữ liệu biểu diễn các tổng hợp cặp thuốc với cơ chế chú ý để dự đoán giá trị kết hợp của thuốc cho các dòng tế bào. AE-XGBSynergy sử dụng thông tin cấu trúc mạng PPI được khai thác mối quan hệ phức tạp của gen-bệnh-thuốc để tích hợp đa dữ liệu -omics thông qua bộ Autoencoder để trích xuất biểu diễn của các dòng tế bào từ đó dự đoán khả năng kết hợp thuốc qua bộ phân loại XGBoost. Các 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 khác giữa các cặp thuốc 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 thử nghiệm đã chứng minh giải pháp đề xuất mang lại hiệu năng tốt hơn các phương pháp hiện thời.

PHẦN KẾT LUẬN

Các kết quả đã đạt được

Mục tiêu của y học chính xác là xác định được phương thức điều trị chính xác cho từng bệnh nhân dựa trên hồ sơ sinh học của họ. Có nhiều các phương pháp đã được đề xuất cho việc dự đoán đáp ứng thuốc, tuy nhiên các đề xuất này thường hoặc chưa áp dụng dữ liệu biểu diễn thuốc hoặc mới chỉ áp dụng biểu diễn dưới dạng chuỗi chưa biểu diễn dưới dạng tự nhiên hơn như dạng đồ thị nên chưa tận dụng hết thông tin dữ liệu thuốc. Bên cạnh đó, các nghiên cứu này cũng chưa tích hợp đa dạng dữ liệu -omics của dòng tế bào trong dự đoán đáp ưng thuốc. Do đó, luận án đã trình bày 04 giải pháp liên quan đến việc thực hiện tích hợp dữ liệu biểu diễn thuốc dưới dạng đồ thị phân tử và tích hợp đa dữ liệu -omics cho dự đoán đáp ứng thuốc như sau:

Đề xuất các giải pháp dự để 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 từng người bệnh. Luận án đã trình bày hai đề xuất tích hợp dữ liệu biểu diễn thuốc dưới dạng đồ thị phân tử và giải pháp tích hợp đa dữ liệu -omics cho dự đoán đáp ứng thuốc như sau:

(1) Đề xuất giải pháp học dữ liệu biểu diễn đồ thị của phân tử thuốc – GraphDRP: là giải pháp áp dụng cách biểu diễn thuốc dưới dạng đồ thị phân tử một cách tự nhiên hơn so với các cách biểu diễn của các nghiên cứu trước đó (chuỗi, ảnh), 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 một số biến thể của mạng nơ-ron đồ thị (GCN, GAT, GIN, GCN-GAT). Trong khi các dòng tế bào được biểu diễn dưới dạng các vec-tơ nhị phân mô tả thông tin đột biến gen di truyền, các đặc trưng biểu diễn được học thông qua các lớp tích chập một chiều (CNN1D). Các đặc trưng biểu diễn cho thuốc và dòng tế bào sau đó được kết hợp thành các biểu diễn đặc trưng cho từng cặp tương tác thuốc - dòng tế bào để dự đoán đáp ứng thuốc.

(2) Đề xuất giải pháp tích hợp đa dữ liệu -omics và dữ liệu đồ thị phân tử thuốc - GraOmicDRP: là giải pháp tích hợp đa dạng các dữ liệu biểu diễn các dòng tế bào không chỉ là dữ liệu gen di truyền mà còn là dữ liệu methyl hóa, dữ liệu biểu hiện gen kết hợp với dữ liệu biểu diễn đồ thị phân tử thuốc để dự đoán đáp ứng thuốc. Trong đó, các biểu diễn đặc trưng phân tử thuốc được học dựa trên mạng nơ-ron đồ thị đẳng cấu (GIN – mô hình mạng nơ-ron đồ thị hiệu quả nhất trong đề xuất GraphDRP) và dữ liệu biểu diễn dòng tế bào là các cách kết hợp các dữ liệu -omics khác nhau được trích xuất thông qua mạng CNN1D. GraOmicDRP đã cho thấy khả năng vượt trội của việc tích hợp đa dữ liệu -omics hơn khi tích hợp đơn dữ liệu -omics trên các kịch bản thử nghiệm. Việc tích hợp đa dữ liệu -omics trong dự đoán đáp ứng thuốc không chỉ hiệu năng 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 GraOmicsDRP cho thấy dữ liệu biểu hiện gen (GE) là dữ liệu đóng góp quan trọng trong việc dự đoán đáp ứng thuốc. So sánh với các nghiên cứu tích hợp đa dữ liệu -omics tiên tiến khác mà không sử dụng dữ liệu biểu diễn thuốc như MOLI, và DeepDR, giải pháp đề xuất cũng cho thấy hiệu năng dự đoán tốt hơn.

Đề xuất các giải pháp tích hợp dữ liệu để dự đoán đáp ứng đa thuốc

Kết hợp thuốc là kết hợp hai hoặc nhiều thuốc trong quá trình điều trị khắc phục tình trạng kháng thuốc sau thời gian dài điều trị theo liệu trình đơn thuốc. Các giải pháp cho bài toán dự đoán đáp ứng đa thuốc (kết hợp thuốc) cho các dòng tế bào trong luận án tiếp tục hướng tích hợp dữ liệu với hai đề xuất GraOmicSynergy và AE-XGBSynergy như sau:

(3) Đề xuất giải pháp học các biểu diễn đồ thị của đa phân tử thuốc và tích hợp đa dữ liệu -omics – GraOmicSynergy: Đây là đề xuất trong đó mỗi cặp thuốc được biểu diễn dưới dạng đồ thị phân tử và được học thông qua mạng nơ-ron đồ thị đẳng cấu (GIN), sau đó thông tin biểu diễn cặp thuốc kết hợp tương tác với dòng tế bào được tổng hợp thông qua cơ chế chú ý (attention). Dữ liệu biểu diễn dòng tế bào được tổng hợp bằng cách tích hợp đa dữ liệu -omics khác nhau, được học thông qua các mạng nơ-ron tích chập 1 chiều, tạo thành một véc-tơ biểu diễn duy nhất. Các kết quả thử nghiệm cho thấy hiệu năng của việc tích hợp đa dữ liệu -omics tốt hơn so với tích hợp đơn dữ liệu -omics trong mô hình dự đoán đáp ứng đa thuốc đối với dòng tế bào. So sánh với các giải pháp tiên tiến không tích hợp đa dữ liệu -omics khác như DeepDDS (có tích hợp tích hợp dữ liệu biểu diễn đồ thị phân tử thuốc) và DeepSynergy (không tích hợp dữ liệu biểu diễn đồ thị phân tử thuốc) cũng cho thấy hiệu năng vượt trội của GraOmicSynergy.

(4) Đề xuất giải pháp tích hợp đa dữ liệu -omics và thông tin mạng sinh học - AE-XGBSynergy: là đề xuất tích hợp đa dữ liệu -omics của dòng tế bào được trích xuất thông qua bộ mã hóa pre-trained encoder 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. AE-XGBSynergy khai thác các mối quan hệ thuốc – đích (drug-protein), dòng tế bào – protein, xây dựng mạng 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 đáp ứng đa 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. So sánh với mô hình tiên tiến NEXGB (mô hình chỉ sử dụng thông tin cấu trúcmạng PPI), AE-XGBSynergy cho thấy khả năng vượt trội hơn về tất cả các độ đo và trên hai bộ dữ liệu thử nghiệm khác nhau (O'Neil và DrugCombDB).

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 các giải pháp pháp dự đoán.

Về mặt dữ liệu:

- Thuốc: Đồ thị phân tử thuốc đã cho thấy tiềm năng trong việc dự đoán đáp ứng thuốc. Tuy nhiên còn các đặc trưng lý hóa, cấu trúc 3D, đặc trưng góc.. chưa được khai thác. Do đó tích hợp đa dạng thông tin biểu diễn thuốc để học nhiều hơn các đặc trưng ẩn của phân tử thuốc có thể triển khai trong mô hình học biểu diễn thuốc.
- 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 nghiên cứu đề xuất. Do đó mô hình đề xuất có thể nghiên tích hợp thêm dữ liệu sinh học khác của dòng tế bào cần khai thác như proteomics, metabomics, .. hay các dữ liệu, tương tác thuốc-thuốc, tương tác gen.

Về mặt phương pháp: Cấu trúc phân tử hóa học thuốc không chỉ có một số lượng nhỏ các thuốc chống ung thư hiện nay mà có hàng ngàn loại thuốc khác nhau. Do đó có thể nghiên cứu triển khai áp dụng các phương pháp tính tiên tiến khác như Molecular-pretrain model để học tăng cường các biểu diễn phân tử thuốc. Các mô hình tính toán tiên tiến khác như transformer, graphformer... đã mang lại kết quả tiềm năng cho việc khai phá thuốc cũng có thể áp dụng cho bài toán dự đoán đáp ứng thuốc.

DANH MỤC CÁC CÔNG TRÌNH CÔNG BỐ

TẠP CHÍ KHOA HỌC

[1] "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] "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] "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] "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] "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