**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

---

**NGUYỄN ANH TUẤN**

# NGHIÊN CỨU XÂY DỰNG MÔ HÌNH HỌC BIỂU DIỄN SỬ DỤNG DỮ LIỆU ĐA OMICS CHO PHÂN NHÓM UNG THƯ

**ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KĨ THUẬT**

**(Theo định hướng ứng dụng)**

HÀ NỘI - 2025

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

---

NGUYỄN ANH TUẤN

# NGHIÊN CỨU XÂY DỰNG MÔ HÌNH HỌC BIỂU DIỄN SỬ DỤNG DỮ LIỆU ĐA OMICS CHO PHÂN NHÓM UNG THƯ

CHUYÊN NGÀNH: HỆ THỐNG THÔNG TIN
MÃ SỐ: 8.48.01.04

**ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT**

**(Theo định hướng ứng dụng)**

**NGƯỜI HƯỚNG DẪN KHOA HỌC**

**PGS.TSKH. HOÀNG ĐĂNG HẢI**

HÀ NỘI - 2025

# LỜI CAM ĐOAN

Tôi cam đoan rằng công trình nghiên cứu này là thành quả của công sức cá nhân tôi và không sao chép từ bất kỳ nguồn nào khác. Tất cả thông tin được trình bày trong đề án này đều là sản phẩm của công việc cá nhân hoặc được tổng hợp từ nhiều nguồn tài liệu khác nhau. Mọi tài liệu tham khảo đều được trích dẫn một cách hợp pháp và có nguồn gốc rõ ràng. Các dữ liệu và kết quả được trình bày trong đề án đều là trung thực và chưa từng được công bố trong các công trình nào khác.

**Tác giả đề án**

**Nguyễn Anh Tuấn**

# LỜI CẢM ƠN

Để hoàn thành được đề án này, em xin gửi lời cảm ơn chân thành tới những người đã giúp đỡ tôi trong quá trình học tập, nghiên cứu khoa học.

Đầu tiên em xin cảm ơn thầy PGS. TSKH. Hoàng Đăng Hải đã tận tình hướng dẫn truyền đạt những kinh nghiệm quý báu và giúp đỡ tác giả từ những ngày bắt đầu hướng dẫn đến ngày bảo vệ.

Tiếp theo, em cảm ơn các thầy cô trong Trường Học viện Công nghệ Bưu chính Viễn thông đã tận tình dạy dỗ, truyền đạt kiến thức quý báu.

Em cũng xin trân trọng cảm ơn đơn vị nơi tôi công tác và làm việc đã tạo mọi điều kiện thuận lợi cho tôi trong suốt quá trình học cao học.

# MỤC LỤC

*MỤC LỤC* ..... *i*
*DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT* ..... *iv*
*DANH MỤC CÁC BẢNG* ..... *v*
*DANH MỤC CÁC HÌNH ẢNH* ..... *vi*
*MỞ ĐẦU* ..... *1*
1. Lý do chọn đề tài ..... 1
2. Tổng quan về vấn đề cần nghiên cứu ..... 2
3. Mục đích nghiên cứu ..... 3
4. Đối tượng và phạm vi nghiên cứu ..... 3
5. Phương pháp nghiên cứu ..... 4
6. Cấu trúc của luận văn ..... 4
*Chương 1: CƠ SỞ LÝ THUYẾT* ..... *5*
1.1 Khái niệm về omics và đa omics ..... 5
*1.1.1 Genomics* ..... 5
*1.1.2 Transcriptomics* ..... 6
*1.1.3 Proteomics* ..... 6
*1.1.4 Metabolomics* ..... 7
*1.1.5 Epigenomics* ..... 7
1.2 Các phương pháp học biển diễn trên dữ liệu sinh học ..... 8
*1.2.1 Học biểu diễn có giám sát* ..... 9
*1.2.2 Học biểu diễn không có giám sát* ..... 10
*1.2.3 Học biểu diễn dựa trên đồ thị* ..... 11
1.3 Graph Attention Network (GAT) và ứng dụng trong phân tích dữ liệu omics ..... 13
*1.3.1 Khái quát về Graph Attention Network (GAT)* ..... 13
*1.3.2 Khả năng áp dụng GAT cho phân tích dữ liệu omics* ..... 15
1.4 Mô hình GAT không đồng nhất ..... 16
1.5 Vấn đề nghiên cứu đặt ra trong đề án ..... 19
1.6. Kết chương ..... 20
*Chương 2: XÂY DỰNG MÔ HÌNH HỌC BIỂU DIỄN VỚI MULTI OMICS* ..... *21*
2.1 Tổng quan về mô hình omicsGAT ..... 21

2.2 Phân tích các yêu cầu đối với mô hình học biểu diễn Multi-omics ..... 22
*2.2.1 Khả năng tích hợp đồng thời nhiều dạng dữ liệu omics* ..... 22
*2.2.2 Khả năng khai thác cấu trúc đồ thị* ..... 22
*2.2.3 Khả năng học biểu diễn đa nhiệm* ..... 23
*2.2.4 Khả năng xử lý dữ liệu đa chiều* ..... 23
*2.2.5 Khả năng biểu diễn trực quan và hỗ trợ phân tích* ..... 23
2.3 Kiến trúc mô hình học biểu diễn ..... 24
*2.3.1 Đầu vào của mô hình* ..... 25
*2.3.2 Lớp xử lý chính* ..... 26
*2.3.3 Đầu ra của mô hình* ..... 29
2.4 Phương pháp phân cụm, phân nhóm ung thư ..... 31
*2.4.1 Phân nhóm ung thư bằng omicsGAT* ..... 32
*2.4.2 Phân cụm bằng omicsGAT Clustering* ..... 34
2.5 Phương pháp huấn luyện và đánh giá mô hình ..... 37
*2.5.1 Phương pháp huấn luyện* ..... 37
*2.5.2 Đánh giá mô hình* ..... 38
2.6 Xây dựng ứng dụng hiển thị trực quan ..... 39
2.7 Kết chương ..... 40
*Chương 3: TRIỂN KHAI THỬ NGHIỆM* ..... *41*
3.1 Thu thập dữ liệu từ TCGA ..... 41
*3.1.1 Các field (features) chính trong dữ liệu* ..... 42
3.2 Tiền xử lý dữ liệu ..... 42
*3.2.1 Loại bỏ các gene không hợp lệ hoặc ít thông tin* ..... 42
*3.2.2 Chuẩn hoá dữ liệu* ..... 43
*3.2.3 Đồng bộ giữa nhiều loại omics và xuất file* ..... 44
3.3 Chạy thử mô hình omicsGAT và ứng dụng hiển thị ..... 44
*3.3.1 Cấu trúc và workflow của ứng dụng* ..... 44
*3.3.2 Upload dữ liệu và cấu hình tham số* ..... 45
*3.3.3 Huấn luyện mô hình* ..... 46
*3.3.4 Trực quan hoá kết quả* ..... 49
3.4 Phân tích, đánh giá kết quả thử nghiệm ..... 51
*3.4.1 Thiết lập thí nghiệm* ..... 51
*3.4.2 Kết quả* ..... 52
*3.4.3 Phân tích kết quả* ..... 52
3.5 Kết chương ..... 53
*KẾT QUẢ VÀ HƯỚNG PHÁT TRIỂN TIẾP* ..... *54*
1. Kết luận ..... 54
2. Hướng phát triển tiếp ..... 55

*DANH MỤC TÀI LIỆU THAM KHẢO*.....*55*

# DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT

| Viết tắt | Tiếng Anh | Tiếng Việt |
|---|---|---|
| GNN | Graph Neural Networks | Mạng nơ-ron đồ thị |
| GAT | Graph Attention Network | Mạng đồ thị |
| GCN | Graph Convolutional Networks | Mạng tích chập đồ thị |
| BRCA | Breast Invasive Carcinoma | Ung thư vú |
| LUAD | Lung Adenocarcinoma | Ung thư biểu mô tuyến phổi |
| KIRC | Kidney Renal Clear Cell Carcinoma | Ung thư biểu mô tế bào thận |
| GBM | Glioblastoma Multiforme | U nguyên bào thần kinh đệm |
| NGS | Next-Generation Sequencing | Giải trình tự thế hệ mới |
| VAE | Variational Autoencoder | Bộ mã hoá tự động biến thiên |
| NMI | Normalized Mutual Information | Thông tin trộn được chuẩn hoá |
| ARI | Adjusted Rand Index | Chỉ số Rand điều chỉnh |
| MSE | Mean Squared Error | Trung bình bình phương sai số |
| LSTM | Long-Short Term Memory | Bộ nhớ dài ngắn hạn |

# DANH MỤC CÁC BẢNG

Bảng 3.1: Cấu trúc của file RNA-seq 41
Bảng 3.2: Cấu trúc của bảng gán nhãn 49
Bảng 3.3: Cấu hình của thí nghiệm cho các mô hình 51
Bảng 3.4: Bản so sánh kết quả giữa các mô hình 51

# DANH MỤC CÁC HÌNH ẢNH

Hình 1.1: Các loại dữ liệu omics 5
Hình 1.2: Công dụng nổi bật của các loại omics 8
Hình 1.3: Kiến trúc mô hình học có giám sát 9
Hình 1.4: Kiến trúc mô hình học không giám sát 10
Hình 1.5: Mô hình học biểu diễn trên đồ thị 12
Hình 1.6: Kiến trúc của các mô hình đồ thị hiện đại 13
Hình 1.7: GAT gán các trọng số khác nhau cho các nút lân cận khác nhau 13

Hình 2.1: Tổng quan của một mô hình omicsGAT 24
Hình 2.2: Quá trình xử lý dữ liệu đầu vào 25
Hình 2.3: Quan hệ giữa các nút dựa trên trọng số 28
Hình 2.4: Dữ liệu đầu ra được sử dụng để phân loại 30
Hình 2.5: Sơ đồ phân cụm / phân nhóm 32
Hình 2.6: Sơ đồ phân nhóm ung thư bằng mô hình omicsGAT 33
Hình 2.7: Phân cụm omicsGAT bằng Clustering 35
Hình 2.8: Kết quả phân cụm dựa vào Embedding đầu ra từ Encoder 36
Hình 2.9: Ví dụ về một giao diện hiển thị trực quan kết quả 39

Hình 3.3: Code loại bỏ các gene không hợp lệ 41
Hình 3.4: Code chuẩn hoá dữ liệu cho RNA-seq 42
Hình 3.5: Code chuẩn hoá dữ liệu cho Methylation 42
Hình 3.6: Code chuẩn hoá dữ liệu cho CNV 42
Hình 3.7: Code xuất file input 43
Hình 3.8: Workflow của ứng dụng mô phỏng 43
Hình 3.9: Phần upload dữ liệu của người dùng 44
Hình 3. 10: Phần điều chỉnh tham số huấn luyện 45
Hình 3.11: Đồ thị K-NN giữa các mẫu 46
Hình 3.12: Thuật toán phân cụm K-Means 47
Hình 3.13: Biểu đồ PCA 48
Hình 3.14: Biểu đồ UMAP 49

# MỞ ĐẦU

## 1. Lý do chọn đề tài

Ung thư là một trong những nguyên nhân gây tử vong hàng đầu trên thế giới, với các dạng bệnh đa dạng và phức tạp. Trong nhiều trường hợp, các bệnh nhân mắc cùng một loại ung thư lại có phản ứng điều trị khác nhau, do sự khác biệt ở mức độ phân tử. Ung thư có liên quan đến nhiều thông tin sinh học genomics, transcriptomics, proteomics,… Do đó, việc phân nhóm ung thư (cancer subtyping) dựa trên thông tin sinh học sâu hơn có ý nghĩa quan trọng trong chẩn đoán và điều trị cá nhân hóa [1].

Dữ liệu omics là tập hợp của nhiều loại dữ liệu sinh học. Khái niệm Omics xuất phát từ cụm từ cuối trong các dữ liệu sinh học như: Genomics (dữ liệu Gen DNA), Transriptomics (dữ liệu phiên mã RNA), Proteomics (dữ liệu protein),… Phân tích nhiều loại dữ liệu omics (multi-omics) cung cấp cái nhìn toàn diện về cơ chế sinh học của ung thư [2, 3]. Hiện nay, nguồn dữ liệu đa omics đang ngày càng nhiều và được cung cấp công khai cho cộng đồng, điển hình là các tập dữ liệu bản đồ bộ gen ung thư TCGA (The Cancer Genome Atlas) [4].

Phân nhóm ung thư là một bài toán nhỏ trong hỗ trợ chân đoán, tiên lượng và điều trị cá thể hóa [5, 6, 7]. Tuy nhiên, việc phân nhóm theo các phương pháp học máy truyền thống tới nay vẫn chưa tận dụng hết những lợi thế của cấu trúc dữ liệu đa omics [8].

Các mô hình mạng nơ ron đồ thị, đặc biệt là mô hình học biểu diễn đã thể hiện được khả năng trích xuất đặc trưng hiệu quả từ dữ liệu phức tạp. Áp dụng học biểu diễn để phân cụm dữ liệu đa omics là hướng tiếp cận mới mẻ, có thể giúp phân nhóm ung thư tiềm ẩn mà các phương pháp truyền thống không nhận ra.

Xuất phát từ nhu cầu thực tiễn và định hướng công việc nơi em đang công tác, em quyết định chọn đề tài "Nghiên cứu xây dựng mô hình học biểu diễn sử dụng dữ liệu đa omics trong phân nhóm ung thư". Đề án tốt nghiệp hướng tới mục tiêu nghiên cứu cách thứ xây dựng mô hình học biểu diễn trên dữ liệu multi-omics

để phân nhóm ung thư, áp dụng mô hình vào xây dựng một ứng dụng đơn giản nhằm hỗ trợ các bác sỹ cách sử dụng dữ liệu omics trong công tác chân đoán, hỗ trợ điều trị bệnh nhân.

## 2. Tổng quan về vấn đề cần nghiên cứu

Trong những năm gần đây, nhu cầu cá nhân hóa trong điều trị ung thư đã thúc đẩy sự phát triển của các phương pháp phân nhóm bệnh nhân dựa trên dữ liệu phân tử (subtyping). Các chiến lược phân nhóm này đóng vai trò quan trọng trong việc dự đoán tiên lượng, lựa chọn liệu pháp điều trị phù hợp và hiểu rõ hơn về cơ chế sinh học đặc trưng của từng nhóm ung thư. Sự ra đời của công nghệ giải trình tự thế hệ mới đã cung cấp lượng dữ liệu đa omics khổng lồ, bao gồm mRNA, CNV, methyl hóa DNA và miRNA [1, 2, 3]. Tuy nhiên, việc tích hợp hiệu quả các loại dữ liệu này vẫn là một thách thức lớn do sự khác biệt về định dạng, chiều dữ liệu, độ nhiễu và tính phức tạp sinh học giữa các omics [9-16].

Các phương pháp truyền thống như phân tích đơn omics hoặc ghép nối đơn giản (early fusion) thường không đủ khả năng nắm bắt được mối liên hệ phức tạp giữa các lớp dữ liệu [9]. Một số phương pháp hiện đại hơn sử dụng autoencoder, mạng học sâu hoặc kỹ thuật học biểu diễn (representation learning) đã cải thiện phần nào hiệu suất nhưng vẫn thiếu sự tương tác sinh học có ý nghĩa giữa các bệnh nhân [13-16]. Gần đây, các mô hình học sâu dựa trên mạng đồ thị (Graph Neural Networks - GNNs), đặc biệt là các mô hình mạng đồ thị dị thể (Heterogeneous Graph Neural Networks - HeteroGNNs) [14], đã chứng minh tiềm năng mạnh mẽ trong việc biểu diễn và học các mối quan hệ phức tạp giữa các loại dữ liệu sinh học khác nhau. Tuy vậy, phần lớn các nghiên cứu chỉ tập trung vào mối quan hệ nội tại trong từng bệnh nhân, bỏ qua những kết nối sinh học liên bệnh nhân vốn rất quan trọng trong cơ chế bệnh sinh và phân nhóm lâm sàng.

Các tập dữ liệu bản đồ bộ gen ung thư TCGA (The Cancer Genome Atlas) [4] đang mở ra một cơ hội mới cho việc tích hợp và phân tích đồng thời nhiều loại dữ liệu omics nhằm phát hiện các đặc trưng tiềm ẩn, từ đó giúp hỗ trợ cho các bác

sỹ trong công tác chẩn đoán, phân nhóm ung thư, dự báo tiên lượng và điều trị cho bệnh nhân.

Các mô hình học biểu diễn trên đồ thị đã nổi lên như một hướng tiếp cận hiệu quả, cho phép biểu diễn dữ liệu bệnh nhân dưới dạng đồ thị, tận dụng cơ chế attention để học ra đặc trưng quan trọng giữa các mẫu và giữa các tầng dữ liệu. Mô hình không chỉ cải thiện độ chính xác trong phân nhóm ung thư mà còn tạo ra khả năng hỗ trợ các bác sỹ hiểu rõ hơn những loại dữ liệu omics, quan hệ sinh học của các dữ liệu một cách rõ ràng hơn so với các phương pháp học máy truyền thống.

Tuy nhiên, các mô hình biểu diễn dữ liệu omics còn khá mới mẻ, chưa được triển khai nhiều trong thực tiễn. Đề án này hướng tới việc nghiên cứu, tìm hiểu cách thức xây dựng một mô hình học biểu diễn trên dữ liệu multi-omics với các tập dữ liệu TCGA để phân nhóm ung thư, áp dụng mô hình vào xây dựng một ứng dụng đơn giản nhằm hỗ trợ các bác sỹ cách sử dụng dữ liệu omics trong công tác chân đoán, hỗ trợ điều trị bệnh nhân.

## 3. Mục đích nghiên cứu

Mục đích nghiên cứu chính của đề án tốt nghiệp này là:

- Nghiên cứu, xây dựng mô hình học biểu diễn trên dữ liệu multi-omics.
- Triển khai thử nghiệm mô hình với tập dữ liệu TCGA để phân tích, đánh giá khả năng phân nhóm ung thư.
- Áp dụng mô hình vào xây dựng một ứng dụng đơn giản nhằm hỗ trợ các bác sỹ trong công tác chân đoán, hỗ trợ điều trị bệnh nhân.

## 4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu:

- Mô hình học biểu diễn sử dụng dữ liệu đa omics.
- Phương pháp phân nhóm ung thư
- Tập dữ liệu TCGA

Phạm vi nghiên cứu:

Nghiên cứu sẽ được thực hiện trên 5 loại ung thư từ TCGA bao gồm:

- BRCA (Breast Invasive Carcinoma)

- LUAD (Lung Adenocarcinoma)
- GBM (Glioblastoma Multiforme)
- KIRC (Kidney Renal Clear Cell Carcinoma)
- UCEC (Uterine Corpus Endometrial Carcinoma)

Đề tài tập trung vào việc tích hợp bốn loại omics chính: biểu hiện gene mRNA, biến thể số bản sao (CNV), methyl hóa DNA, và biểu hiện miRNA.

## 5. Phương pháp nghiên cứu

Phương pháp nghiên cứu trong đề tài bao gồm hai hướng chính: nghiên cứu lý thuyết và nghiên cứu thực nghiệm.

Về phương pháp nghiên cứu lý thuyết, Về mặt lý thuyết, nghiên cứu tiến hành khảo sát và tổng hợp các công trình liên quan đến phương pháp tích hợp dữ liệu đa omics, đặc biệt là các kỹ thuật học sâu và mạng nơ-ron đồ thị (Graph Neural Networks).

Về mặt thực nghiệm, nghiên cứu sử dụng dữ liệu từ năm loại ung thư phổ biến thuộc cơ sở dữ liệu TCGA (BRCA, LUAD, GBM, KIRC, UCEC), với đầy đủ bốn loại omics cho mỗi bệnh nhân. Dữ liệu được xử lý thông qua các bước tiền xử lý nghiêm ngặt như loại bỏ dữ liệu thiếu, chuẩn hóa theo từng omics, và tổ chức lại các đặc trưng theo mức độ tương quan. Mô hình omicsGAT được huấn luyện và đánh giá dựa trên các chỉ số như C-index, log-rank test, silhouette score, và phân tích Kaplan-Meier nhằm kiểm tra hiệu quả phân nhóm và khả năng dự đoán thời gian sống sót.

## 6. Cấu trúc của luận văn

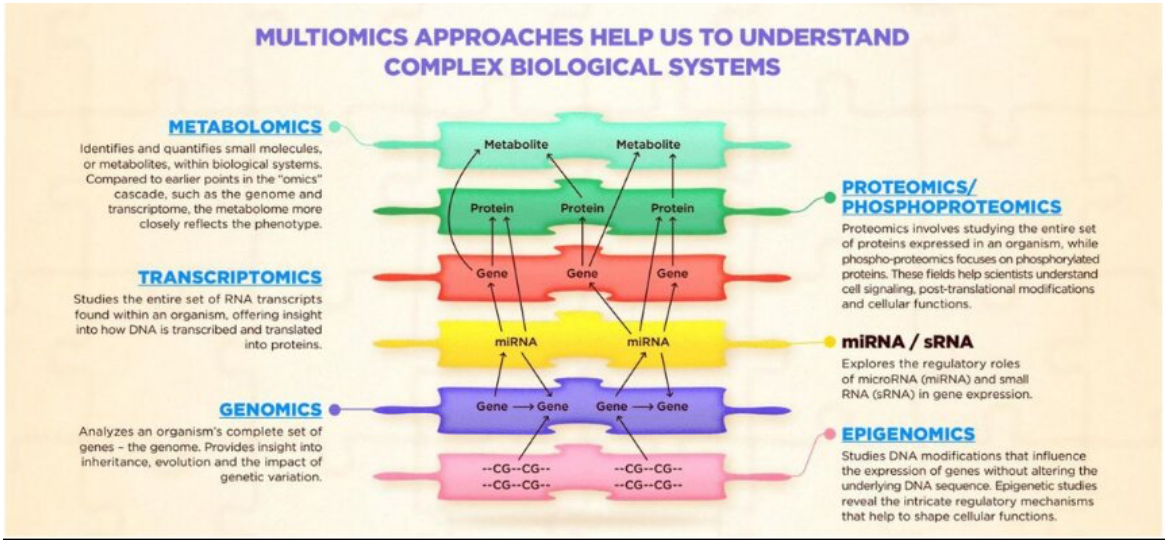
Ngoài phần mở đầu, kết luận, danh mục tài liệu tham khảo và phụ lục, luận văn gồm 03 chương như sau:

- Chương 1 trình bày cơ sở lý thuyết về đa omics và các phương pháp học biểu diễn trên dữ liệu sinh học
- Chương 2 trình bày về xây dựng một mô hình học biểu diễn với đa omics
- Chương 3 trình bày các bước triển khai thử nghiệm đa omics trên mô hình omicsGAT

# Chương 1: CƠ SỞ LÝ THUYẾT

## 1.1 Khái niệm về omics và đa omics

Dữ liệu omics đề cập đến khối dữ liệu cực lớn, được tạo ra từ các việc phân tích hết các tập hợp của một loại phân tử sinh học cụ thể, chẳng hạn như toàn bộ DNA (genomics), RNA (transcriptomics) hoặc protein (proteomics). Những dữ liệu này được dùng với mong muốn giúp hiểu biết toàn diện và tổng thể về một hệ thống sinh học.




**Hình 1.1: Các loại dữ liệu omics**

### *1.1.1 Genomics*

Genomics là một nhánh quan trọng trong lĩnh vực Dữ liệu Omics, tập trung vào việc nghiên cứu toàn bộ bộ gen (genome) của một sinh vật. Thay vì chỉ xem xét từng gen riêng lẻ, genomics phân tích cấu trúc, chức năng, tiến hóa và tương tác của tất cả các gen trong cùng một hệ thống sinh học. Công nghệ giải trình tự gen thế hệ mới đã mở ra khả năng thu thập dữ liệu genomics ở quy mô lớn, giúp hiểu sâu hơn về cơ chế di truyền của bệnh tật, đa dạng sinh học và tiến hóa. Dữ liệu genomics thường bao gồm trình tự DNA, biến thể di truyền (như SNPs, CNVs) và vùng điều hòa gen. Thông qua các phân tích tin sinh học, genomics hỗ trợ phát hiện gen liên quan đến bệnh, phát triển y học cá thể hóa, và tối ưu hóa liệu pháp điều trị. Đây là nền tảng cơ bản giúp kết nối và tích hợp với các loại dữ liệu Omics khác như

transcriptomics hay proteomics để có cái nhìn toàn diện về hệ thống sinh học.

### *1.1.2 Transcriptomics*

Trong dữ liệu Omics, Transcriptomics chuyên nghiên cứu toàn bộ tập hợp RNA được phiên mã từ DNA trong một tế bào hay mô tại một thời gian cụ thể, được gọi là transcriptome. Không giống genomics – vốn tập trung vào thông tin di truyền tĩnh, transcriptomics phản ánh hoạt động động học của gen, cho biết gen nào đang được biểu hiện và ở mức độ nào. Bằng cách sử dụng RNA-Seq hay microarray, transcriptomics cho phép định lượng và so sánh biểu hiện gen giữa các điều kiện sinh học khác nhau, ví dụ như mô bệnh và mô bình thường. Dữ liệu transcriptomics giúp xác định các gen điều hòa, con đường tín hiệu và cơ chế phân tử liên quan đến sự phát triển, bệnh lý hoặc đáp ứng môi trường. Phân tích transcriptome cũng góp phần khám phá các RNA không mã hóa (như miRNA, lncRNA) – những yếu tố quan trọng trong điều hòa biểu hiện gen. Khi được tích hợp với dữ liệu genomics và proteomics, transcriptomics giúp ta có cái nhìn khái quá hơn về hoạt động của tế bào, từ thông tin di truyền đến biểu hiện chức năng, hỗ trợ mạnh mẽ cho phát triển y học chính xác.

### *1.1.3 Proteomics*

Proteomics nghiên cứu toàn bộ tập hợp protein – hay còn gọi là proteome – được tạo ra bởi một tế bào, mô hoặc sinh vật tại một thời điểm cụ thể. Nếu genomics cho biết "khả năng" di truyền của sinh vật, thì proteomics phản ánh trực tiếp "thực tế hoạt động" của hệ thống sinh học, vì protein là những phân tử thực hiện gần như toàn bộ các chức năng trong tế bào. Thông qua các kỹ thuật hiện đại như khối phổ (mass spectrometry – MS) và điện di hai chiều (2D-PAGE), proteomics cho phép xác định, định lượng và phân tích các biến đổi hậu dịch mã của protein. Dữ liệu proteomics giúp hiểu rõ hơn về mạng lưới tương tác protein, con đường tín hiệu và cơ chế phân tử trong các quá trình sinh học hay bệnh lý. Ứng dụng của proteomics rất rộng, từ phát hiện biomarker chẩn đoán, thiết kế thuốc mới, đến cá thể hóa điều trị. Khi được tích hợp với genomics và transcriptomics, proteomics góp phần hoàn thiện về hoạt động của tế bào, giúp kết nối thông tin từ

gen đến chức năng sinh học cuối cùng.

### *1.1.4 Metabolomics*

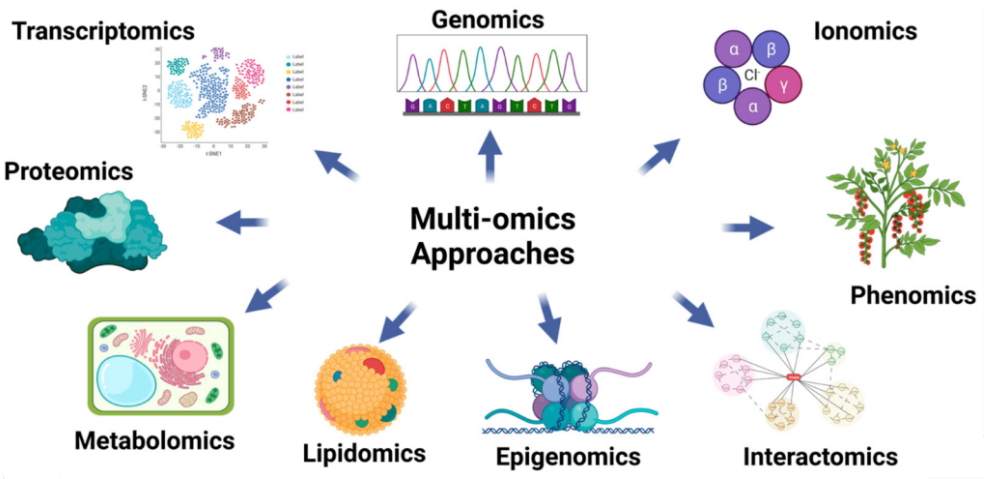
Một loại dữ liệu khác trong dữ liệu Omics là Metabolomics. Loại dữ liệu này kết quả của việc phân tích các chất chuyển hoá nhỏ (metabolites) trong tế bào, mô hoặc dịch cơ thể. Đây là tầng thông tin phản ánh rõ ràng nhất trạng thái sinh lý và sinh hóa của một hệ thống sinh học tại thời điểm đo, vì các chất chuyển hóa chính là sản phẩm cuối cùng của các quá trình gen – protein – enzyme. Bằng các kỹ thuật tiên tiến như sắc ký lỏng ghép khối phổ (LC-MS), sắc ký khí (GC-MS) hay cộng hưởng từ hạt nhân (NMR), metabolomics giúp định danh, định lượng và so sánh nồng độ rất nhiều các tử nhỏ như amino acid, lipid, đường hay acid hữu cơ. Dữ liệu metabolomics cung cấp cái nhìn rõ hơn về con đường trao đổi chất, phản ứng sinh hóa và sự thay đổi của cơ thể trước các yếu tố bệnh lý hoặc môi trường. Trong nghiên cứu y sinh, metabolomics đóng vai trò then chốt trong phát hiện biomarker, theo dõi quá trình bệnh. Khi được tích hợp với genomics, transcriptomics và proteomics, metabolomics giúp xây dựng bản đồ toàn diện về hoạt động sinh học – từ gen đến biểu hiện phân tử và chức năng trao đổi chất.

### *1.1.5 Epigenomics*

Epigenomics là lĩnh vực trong Dữ liệu Omics nghiên cứu toàn bộ các biến đổi biểu sinh (epigenetic modifications) diễn ra trên bộ gen mà không làm thay đổi trình tự DNA, nhưng lại ảnh hưởng đến cách gen được biểu hiện. Những biến đổi này bao gồm chủ yếu methyl hóa DNA, biến đổi histone, và cấu trúc nhiễm sắc thể (chromatin remodeling). Epigenomics giúp giải thích tại sao các tế bào có cùng bộ gen lại có thể biểu hiện khác nhau và thực hiện những chức năng riêng biệt. Với các công nghệ như bisulfite sequencing, ChIP-seq, hay ATAC-seq, lĩnh vực này cho phép lập bản đồ toàn diện về trạng thái biểu sinh trong các loại tế bào và điều kiện sinh học khác nhau. Dữ liệu epigenomics vô cùng giá trị trong quá trình nghiên cứu phát triển phôi, lão hóa, và bệnh lý phức tạp như ung thư, nơi sự điều hòa gen bị rối loạn ở mức biểu sinh. Khi được tích hợp với các tầng dữ liệu khác như genomics và transcriptomics, epigenomics cho thấy rõ mối liên hệ giữa di truyền, môi trường và

biểu hiện gen, góp phần thúc đẩy y học cá thể hóa và liệu pháp điều hòa biểu sinh trong điều trị bệnh.

Dữ liệu đa omics là tập hợp các loại dữ liệu sinh học được thu thập từ nhiều tầng phân tử khác nhau chẳng hạn như genomics, transcriptomics, proteomics, metabolomics hay epigenomics. Việc kết hợp nhiều lớp dữ liệu omics cho phép các nhà khoa học có cái nhìn toàn diện hơn về cơ chế hoạt động của sinh vật, từ đó hiểu rõ mối quan hệ giữa gen, biểu hiện và kiểu hình. Dữ liệu đa omics hiện được ứng dụng mạnh mẽ trong nghiên cứu y sinh, đặc biệt là trong y học chính xác, giúp xác định biomarker, dự đoán bệnh, và lựa chọn liệu pháp điều trị cá thể hóa. Tuy nhiên, việc xử lý và tích hợp dữ liệu đa omics không đơn giản do sự đa dạng về định dạng, kích thước và độ phức tạp của từng loại dữ liệu. Các phương pháp phân tích thống kê, học máy và trí tuệ nhân tạo đang được phát triển để giải quyết thách thức này. Nhờ đó, dữ liệu đa omics đang dần trở nên cực quan trọng trong kỷ nguyên sinh học hệ thống và y học dữ liệu lớn.




**Hình 1.2: Công dụng nổi bật của các loại omics**

Dữ liệu đa omics giúp làm rõ bức tranh toàn diện về hệ sinh học bằng cách liến kết thông tin từ nhiều tầng dữ liệu, qua đó tăng sự chính xác trong dự đoán, hỗ trợ tìm ra cơ chế bệnh phức tạp và thúc đẩy các ứng dụng của y học chính xác.

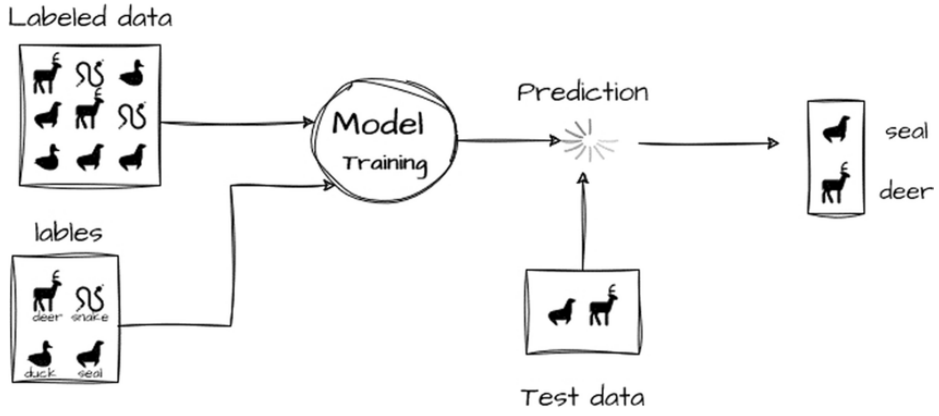
## 1.2 Các phương pháp học biển diễn trên dữ liệu sinh học

Học biểu diễn (representation learning) trên dữ liệu sinh học là một hướng

rất quan trọng trong tin sinh học (bioinformatics) vì dữ liệu sinh học thường có kích thước lớn, nhiều chiều và phức tạp (chuỗi DNA, RNA, protein, biểu hiện gen, hình ảnh mô, v.v.).

### *1.2.1 Học biểu diễn có giám sát*

Học biểu diễn có giám sát (Supervised Representation Learning) là phương pháp mô hình được huấn luyện bằng dữ liệu có nhãn, với mục tiêu học ra các đặc trưng tiềm ẩn phản ánh thông tin quan trọng cho nhiệm vụ dự đoán. Trong lĩnh vực sinh học tính toán, các nhãn này có thể là loại bệnh, loại mô, chức năng protein, hoặc khả năng tương tác giữa các phân tử. Quá trình huấn luyện giúp mô hình không chỉ thực hiện tốt nhiệm vụ cụ thể mà còn học được các biểu diễn tiềm ẩn giàu thông tin sinh học.




**Hình 1.3: Kiến trúc mô hình học có giám sát**

Một trong những hướng phổ biến trong học biểu diễn có giám sát là sử dụng mạng nơ-ron hồi tiếp và mạng Transformer, đặc biệt phù hợp với các loại dữ liệu dạng chuỗi như DNA, RNA, hoặc chuỗi amino acid của protein.

RNNs (Recurrent Neural Networks) được thiết kế để xử lý dữ liệu có tính tuần tự, giúp mô hình ghi nhớ ngữ cảnh và mối quan hệ giữa các phần tử trong chuỗi sinh học. Các biến thể như LSTM hoặc GRU thường được sử dụng để học biểu diễn từ các chuỗi DNA hoặc protein, hỗ trợ các bài toán như:

- Dự đoán vị trí promoter hoặc enhancer trong DNA.
- Phân loại đột biến gen gây bệnh (pathogenic vs benign).

- Dự đoán cấu trúc bậc hai của protein.

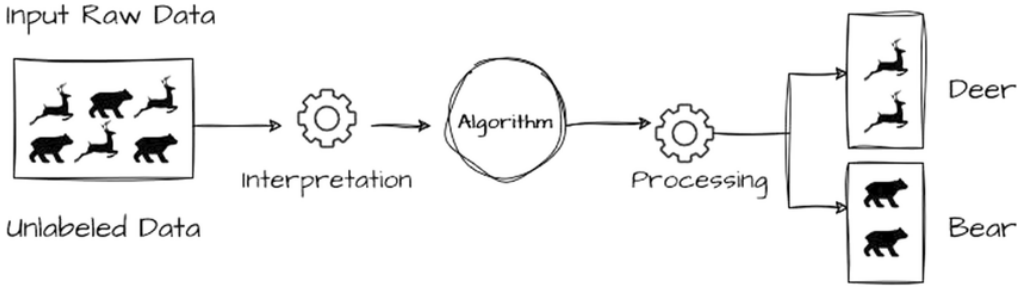
Transformers, với cơ chế self-attention, cho phép mô hình học các phụ thuộc dài hạn trong chuỗi một cách hiệu quả hơn RNNs. Các mô hình dựa trên Transformer như DNABERT, ProtBERT, hoặc ESM (Evolutionary Scale Modeling) đã thể hiện khả năng trích xuất biểu diễn sinh học có ý nghĩa một cách vượt trội.Biểu diễn học được từ các mô hình này có thể được dùng cho đa tác vụ như:

- Dự đoán chức năng protein (protein function prediction).
- Xác định vùng điều hòa gen (regulatory region identification).
- Dự đoán tác động của biến thể di truyền (variant effect prediction).

Nhờ khả năng học biểu diễn mạnh mẽ, các mô hình Transformer hiện đang dần trở thành tiêu chuẩn trong học biểu diễn chuỗi sinh học, tương tự như cách chúng đã cách mạng hóa lĩnh vực xử lý ngôn ngữ tự nhiên (NLP).

### *1.2.2 Học biểu diễn không có giám sát*

Học biểu diễn không giám sát (Unsupervised Representation Learning) là phương pháp trích xuất đặc trưng tiềm ẩn từ dữ liệu mà không cần đến nhãn. Thay vì dựa vào thông tin gán nhãn (như loại bệnh hay chức năng protein), mô hình được huấn luyện để khám phá các cấu trúc, quy luật và mối quan hệ trong dữ liệu. Cách tiếp cận này phù hợp với dữ liệu sinh học, nơi việc gán nhãn chính xác thường khó khăn, tốn kém và đôi khi không khả thi do tính phức tạp của hệ thống sinh học.




**Hình 1.4: Kiến trúc mô hình học không giám sát**

Các phương pháp học không giám sát phổ biến gồm Autoencoder (AE), Variational Autoencoder (VAE), PCA, và đặc biệt là Self-Supervised Learning (học tự giám sát) — một hướng tiếp cận hiện đại đã cho thấy hiệu quả rõ rệt trong việc

học biểu diễn sinh học có ý nghĩa.

Self-Supervised Learning (SSL) là phương pháp trung gian giữa học không giám sát và có giám sát. Thay vì cần nhãn thật, SSL tự tạo ra nhiệm vụ phụ (pretext task) từ chính dữ liệu đầu vào, buộc mô hình phải học hiểu cấu trúc của dữ liệu để giải quyết nhiệm vụ đó. Khi mô hình học xong, biểu diễn thu được có thể được dùng lại (fine-tune) cho nhiều bài toán khác nhau trong sinh học.

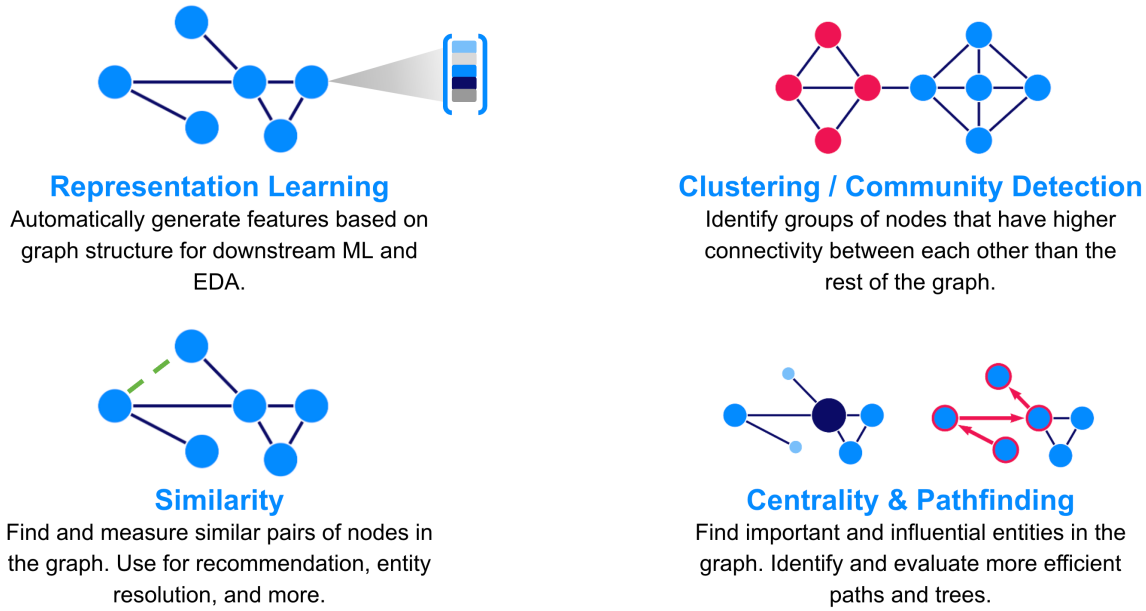
Mô hình DNABERT hoặc GenomeGPT sử dụng cơ chế tương tự như ngôn ngữ tự nhiên (NLP). Mỗi nucleotide hoặc đoạn k-mer được coi như một "từ", và mô hình được huấn luyện với nhiệm vụ dự đoán nucleotide bị che giấu. Cách này giúp mô hình học được ngữ cảnh sinh học và các motif quan trọng trong chuỗi.

Bên cạnh đó còn có các mô hình như ProtBERT, TAPE, hay ESM (Evolutionary Scale Modeling) học biểu diễn chuỗi amino acid bằng cách dự đoán các phần bị che giấu hoặc mối quan hệ tiến hóa giữa các protein. Biểu diễn thu được từ các mô hình này thể hiện thông tin cấu trúc và chức năng của protein, có thể ứng dụng trong:

- Dự đoán cấu trúc bậc ba của protein.
- Phân loại chức năng protein.
- Phân loại chức năng protein.

### *1.2.3 Học biểu diễn dựa trên đồ thị*

Học biểu diễn dựa trên đồ thị (Graph-based Representation Learning) là hướng tiếp cận tận dụng cấu trúc mạng phức tạp vốn có trong nhiều loại dữ liệu sinh học. Thay vì xem dữ liệu như chuỗi, hình ảnh hay ma trận, các phương pháp này biểu diễn dữ liệu dưới dạng đồ thị (graph), trong đó các đỉnh (node) có thể là gene, protein, tế bào, phân tử hoặc bệnh, còn cạnh (edge) mô tả mối quan hệ giữa chúng như tương tác protein–protein, mạng điều hòa gen, hoặc liên kết hóa học trong phân tử.



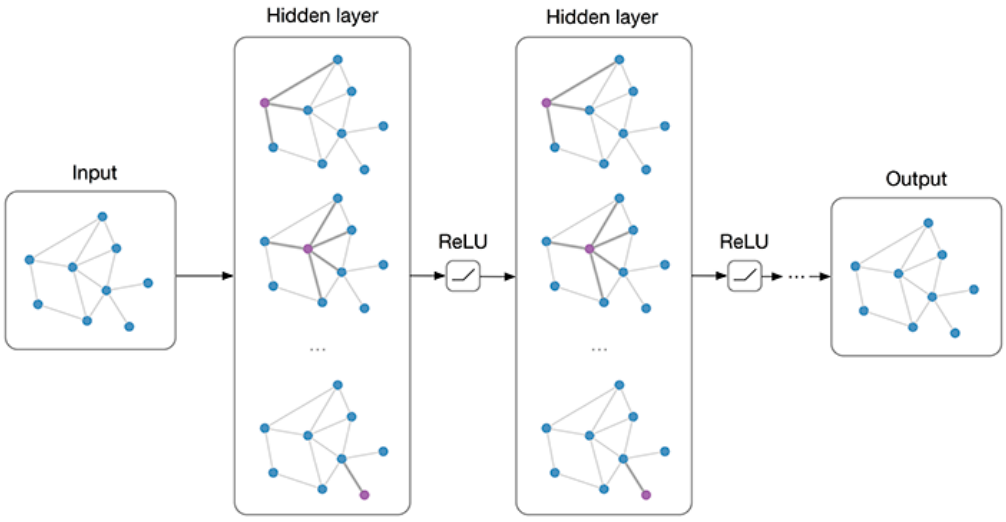

**Hình 1.5: Mô hình học biểu diễn trên đồ thị**

Trong số các mô hình học đồ thị, Graph Convolutional Networks (GCNs) là một trong những mô hình dần trở nên quen thuộc và đem lại hiệu quả nhất. Graph Convolutional Networks (GCN) mở rộng khái niệm tích chập (convolution) từ ảnh sang cấu trúc đồ thị. Thay vì quét trên lưới pixel, GCN "tổng hợp" thông tin từ các node lân cận để cập nhật biểu diễn của một node. Điều này giúp mô hình học được thông tin cục bộ và bán cục bộ trong mạng sinh học.

Một số ứng dụng của GCN trong dữ liệu sinh học:

- Mạng tương tác protein–protein (Protein–Protein Interaction, PPI)
- Dự đoán tương tác thuốc–protein (Drug–Target Interaction prediction)
- Mạng đồng biểu hiện gen (Gene Co-expression Network)

Nhờ khả năng tận dụng trực tiếp cấu trúc đồ thị vốn có trong hệ sinh học, GCN và các mô hình đồ thị hiện đại mạnh mẽ để phân tích dữ liệu đa dạng, từ phân tử đến mạng tương tác cấp hệ thống.

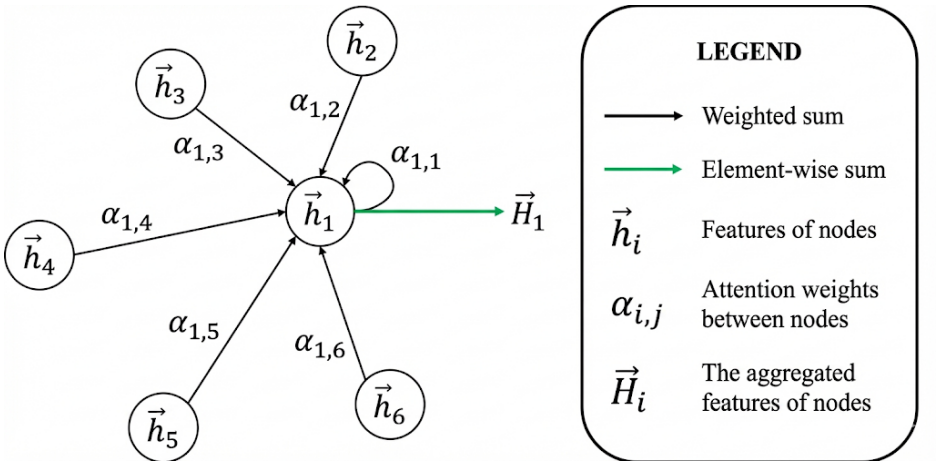



**Hình 1.6: Kiến trúc của các mô hình đồ thị hiện đại**

## 1.3 Graph Attention Network (GAT) và ứng dụng trong phân tích dữ liệu omics

### *1.3.1 Khái quát về Graph Attention Network (GAT)*

GAT là một loại mạng nơ-ron đồ thị mà tại đó mỗi nút (node) trong đồ thị sẽ tập trung (attend) tới các nút lân cận (neighbours) với các trọng số khác nhau, thay vì xử lý tất cả các neighbour như nhau.




**Hình 1.7: GAT gán các trọng số khác nhau cho các nút lân cận khác nhau**

Graph Attention Network (GAT) là một kiến trúc học sâu trên đồ thị sử dụng cơ chế attention để học biểu diễn nút một cách linh hoạt và thích nghi theo cấu trúc lân cận [18].

Khác với các mô hình dựa trên tích chập đồ thị truyền thống (GCN) vốn gán

trọng số cố định cho các nút láng giềng (thường là chuẩn hóa theo bậc), GAT cho phép mô hình tự học mức độ quan trọng của từng láng giềng thông qua các hệ số attention được tính theo nội dung đặc trưng của các nút. Điều này giúp GAT xử lý tốt các đồ thị không đồng nhất, giảm phụ thuộc vào cấu trúc định sẵn và tăng khả năng tổng quát hóa. Kiến trúc GAT tận dụng attention đa chiều (multi-head attention) để ổn định quá trình học và tăng tính biểu diễn. Nhờ đó, GAT trở thành một trong những mô hình nổi bật trong xử lý dữ liệu dạng đồ thị, được ứng dụng rộng rãi trong các bài toán phân loại, dự đoán liên kết và nhiều bài toán khai phá đồ thị phức tạp khác.

Ý tưởng chủ yếu của GAT là thay vì sử dụng giá trị trọng số trung bình hoặc chuẩn hóa đặt cố định cho các nút láng giềng, GAT tự học trọng số phù hợp cho từng mối quan hệ giữa các nút. Điều này được thực hiện thông qua một attention coefficient giữa nút *i* và nút láng giềng *j*, phản ánh mức độ đóng góp của *j* đối với biểu diễn mới của *i*. Trọng số attention được tính dựa trên nội dung đặc trưng (feature) của các nút, nhờ đó mô hình có thể:

- Nhận biết nút nào quan trọng, ngay cả khi bậc nút cao hoặc thấp.
- Xử lý đồ thị không đồng nhất, nơi các cạnh có ý nghĩa khác nhau.
- Giảm phụ thuộc vào cấu trúc tĩnh của đồ thị.

Cơ chế attention trong GAT như sau [18]: Giả sử mỗi nút có vecrtor đặc trưng $h_i$, GAT sẽ thực hiện:

1) Biến đổi tuyến tính đặc trưng thành: $h'_i = \mathrm{W}\ \mathrm{h_i}$
2) Tính Attention giữa cặp nút liền kế: $e_{ij} = a\ (h'_i\ ,\ h'_j\ )$
   Với a là hàm attention học được.
3) Thực hiện chuẩn hóa với hàm Softmax: $\alpha_{ij} = softmax_i\ (\ e_{ij}\ )$
4) Tiếp đó, GAT tổng hợp thông tin với:

$$h_i'' = \sigma\left(\sum_{j \in N(i)} \alpha_{ij}\, h_j'\right) \tag{1.1}$$

Việc tổng hợp thông tin tạo khả năng học được từ nhiều vector đặc trưng, quá trình huấn luyện ổn định hơn, GAT tăng được khả năng học biểu diễn. Hệ số attention giúp cho xác định được nút nào quan trọng hơn. GAT sẽ không phụ thuộc vào ma trận kề chuẩn hóa trước như GCN, có khả năng xử lý đồ thị không đều, không đồng nhất, khả năng diễn giải tốt hơn và có tính linh hoạt, dễ mở rộng.

### *1.3.2 Khả năng áp dụng GAT cho phân tích dữ liệu omics*

Dữ liệu omics (như transcriptomics, genomics, proteomics, metabolomics, methylomics, hay multi-omics) có đặc điểm phức tạp, phi tuyến, và chứa nhiều mối quan hệ ẩn giữa các thực thể sinh học (gene, protein, mẫu bệnh nhân, v.v.). Mô hình Graph Attention Network (GAT) tỏ ra đặc biệt phù hợp vì nó có khả năng học được các mối quan hệ phi tuyến, có trọng số động, giữa các thực thể trong mạng lưới sinh học.

#### *a. GAT mô hình hóa được cấu trúc mạng lưới sinh học*

Thông thường dữ liệu omics sẽ có các mạng tương tác sẵn có. Ví dụ như mạng gene-gene (Gene Regulatory Network) thể hiện quan hệ điều hoà hoặc tương tác giữa các gene/protein hoặc mạng mẫu-mẫu (Sample Similarity Network) biểu diễn độ tương đồng về biểu hiện gene hoặc đặc trưng sinh học giữa các bệnh nhân. Trong GAT, mỗi nút (node) có thể là một gene, protein hoặc là một mẫu bệnh nhân, và các cạnh (edge) sẽ thể hiển mối liên kết sinh học (tương quan, tương tác vật lý hoặc cùng một pathway). Và cũng nhờ cơ chế attention, GAT có thể học được mức độ quan trọng của từng mối liên hệ này, thay vì giả định rằng mọi liên kết đều có ảnh hưởng như nhau giống GCN.

#### *b. GAT học được tầm quan trọng động (dynamic influence) giữa các nút*

Trong sinh học, không phải các gene lân cận đều đóng vai trò giống nhau trong việc quyết định trạng thái của một gene hoặc mẫu. Cơ chế attention coeficient của GAT cho phép mô hình tự động học trọng số cho từng nút lân cận, phản ánh mức độc chú ý mà một gene hoặc bệnh nhân nên dành cho các liên kết của nó. Ví dụ trong mạng gene, một gene điều hoà trung tâm như TP53 sẽ nhận được trọng số cao hơn các gene ít liên quan khác. Điều này hỗ trợ GAT tập trung vào các mối

quan hệ sinh học có ý nghĩa thay vì nhiễu.

### *c. GAT hỗ trợ tích hợp và học biểu diễn đa tầng (multi-omics integration)*

Một thách thức lớn của phân tích omics hiện nay là gồm nhiều loại dữ liệu (gene expression, miRNA, methylation, proteomics,…). Các loại omics này có thể được biểu diễn thành đồ thị khác nhau (mỗi omics = một lớp graph), hoặc đồ thị dị thể (heterogeneous graph) gồm nhiều loại nút và cạnh. GAT dễ dàng mở rộng sang mô hình Heterogeneous Graph Attention Network (HetGAT) hoặc Multi-Head GAT để xử lý đa nguồn dữ liệu cùng lúc, học được cách chú ý khác nhau cho từng loại quan hệ cho phép mô hình hợp nhất thông tin sinh học đa chiều mà không cần giả định tuyến tính hay đồng nhất.

### *d. GAT giúp khai thác ý nghĩa sinh học (biological interpretability)*

Một lợi thế đặc biệt của GAT trong omics là khả năng giải thích kết quả thông qua các trọng số attention. Các hệ số attention ngoài việc xác định gene hoặc protein quan trọng trong mạng điều hòa còn có thể phát hiện mẫu bệnh nhân tương tự về mặt phân tử hay xác định đường dẫn sinh học (pathway) có vai trò quyết định trong phân nhóm bệnh.

### *e. GAT xử lý tốt dữ liệu chiều cao, số mẫu nhỏ và nhiễu lớn*

Dữ liệu omics thường có số lượng đặc trưng (gene, protein, CpG site) gấp hàng nghìn lần số lượng mẫu — điều này khiến nhiều mô hình học máy cổ điển (SVM, RF, MLP) bị quá khớp (overfit). Còn trong GAT mỗi nút chỉ tổng hợp thông tin từ láng giềng trực tiếp, giúp giảm nguy cơ lan truyền nhiễu. Attention cũng cho phép mô hình tự động giảm trọng số cho các neighbour không quan trọng. Và khi được kết hợp với dropout, regularization và multi-head attention, GAT thường ổn định hơn và tổng quát tốt hơn trong môi trường dữ liệu nhiều đặc trưng, ít mẫu.

## 1.4 Mô hình GAT không đồng nhất

Trong thời gian trở lại đây, bài toán phân nhóm ung thư dựa trên dữ liệu đa omics đã được thực hiện nhiều, đặc biệt khi các mô hình mạng nơ-ron đồ thị chứng minh hiệu quả rõ rệt trong việc khai thác mối quan hệ phức tạp giữa các loại dữ liệu

sinh học.

Trong số các hướng tiếp cận hiện đại, mô hình GAT không đồng nhất (Heterogeneous Graph Attention Network) [15] được coi là sáng kiến đột phá, cho phép tích hợp đa omics theo cách phản ánh đúng bản chất sinh học của dữ liệu.

Tabakhi và cộng sự (2024) [15] đề xuất mô hình Heterogeneous Graph Attention Network (HeteroGAT) nhằm giải quyết những nhược điểm còn tồn động của các phương pháp cổ điển trong phân nhóm ung thư. Không giống như các mô hình chỉ đơn thuần ghép nối đặc trưng (feature concatenation) hoặc dùng autoencoder cho từng omics, HeteroGAT mô hình hóa dữ liệu dưới dạng đồ thị dị thể – nơi mỗi loại dữ liệu omics được xem là một loại node khác nhau và được kết nối bằng các mối quan hệ sinh học có ý nghĩa.

### *a. Cấu trúc đồ thị dị thể*

HeteroGAT xây dựng đồ thị gồm nhiều loại nút (heterogeneous nodes), mỗi loại đại diện cho một tầng thông tin sinh học:

- Gene/mRNA node: đặc trưng từ dữ liệu biểu hiện gene.
- miRNA node: đại diện cho các phân tử điều hòa biểu hiện gene.
- CpG/methylation node: phản ánh trạng thái epigenetic.
- CNV node: mô tả biến thể số bản sao ảnh hưởng đến hoạt động gene.

Giữa các node tồn tại nhiều loại cạnh (heterogeneous edges), ví dụ:

- Quan hệ miRNA–mRNA (điều hòa phiên mã).
- Quan hệ CpG–gene promoter (methyl hóa ức chế hoặc kích hoạt phiên mã)
- Quan hệ CNV–gene expression (tăng/giảm bản sao ảnh hưởng mức độ biểu hiện)
- Quan hệ tương đồng giữa bệnh nhân dựa trên profile omics.

Việc đưa các mối quan hệ này vào đồ thị giúp mô hình phản ánh chính xác mạng lưới tương tác sinh học thay vì chỉ xem dữ liệu dạng bảng (tabular).

### *b. Cơ chế Attention trong HeteroGAT*

Điểm nổi bật nhất của HeteroGAT là cơ chế heterogeneous attention, trong đó:

- Mỗi loại cạnh (edge-type) có một bộ tham số attention riêng.
- Mỗi loại node (node-type) có quá trình chuyển đổi đặc trưng độc lập.
- Mô hình học được trọng số quan trọng của từng mối quan hệ sinh học, thay vì coi mọi tương tác có tầm quan trọng như nhau.

Điều này góp phần đảm bảo mô hình chỉ chú ý vào các quan hệ mang ý nghĩa sinh học mạnh, chẳng hạn như miRNA điều hòa trực tiếp gene liên quan ung thư, hoặc những CpG site có mức methyl hóa ảnh hưởng rõ rệt đến hoạt động gene.

***c. Quy trình tích hợp multi-omics***

Quy trình tích hợp của HeteroGAT diễn ra theo các bước:

- Biểu diễn từng loại omics dưới dạng node đặc trưng.
- Kết nối các node dựa trên kiến thức sinh học (biological prior knowledge).
- Truyền thông tin qua mạng GAT đa loại cạnh, tạo ra embedding hợp nhất cho mỗi node và mỗi bệnh nhân.
- Phân nhóm ung thư dựa trên embedding cuối cùng.

***d. Hiệu quả mô hình***

Trong thử nghiệm của bài báo này trên các bộ dữ liệu ung thư từ TCGA đạt được một số thành quả mang tính đột phá:

- Phân nhóm ổn định hơn so với Autoencoder, VAE và các mô hình GCN tiêu chuẩn.
- Tăng rõ rệt các chỉ số như silhouette score, NMI và ARI.
- Các nhóm ung thư thu được tương quan tốt với biểu đồ Kaplan–Meier, cho thấy ý nghĩa lâm sàng trong dự đoán thời gian sống sót.

***e. Hạn chế***

Mặc dù đạt kết quả ấn tượng, mô hình vẫn còn những điểm cần cải thiện:

- Cần tài nguyên tính toán lớn với số lượng node/cạnh cao.
- Khó triển khai trong môi trường lâm sàng vì mô hình phức tạp, khó giải thích hơn các mô hình truyền thống.

Nghiên cứu HeteroGAT cho thấy hướng tiếp cận dựa trên đồ thị dị thể kết hợp attention rất phù hợp với bài toán multi-omics. Việc mô hình hóa trực tiếp quan hệ sinh học ở nhiều tầng giúp tăng chất lượng biểu diễn dữ liệu, nắm bắt tốt các tương tác omics, đồng thời tạo ra các kết quả phân nhóm có giá trị sinh học và lâm sàng. Tuy nhiên, để triển khai rộng rãi, cần tối ưu hóa mô hình và làm giảm sự phức tạp trong tính toán.

## 1.5 Vấn đề nghiên cứu đặt ra trong đề án

Mặc dù các phương pháp phân tích dữ liệu omics và multi-omics đã mang lại nhiều trong nghiên cứu ung thư nhưng vẫn tồn tại hạn chế khiến việc phân nhóm ung thư chưa có kết quả cao. Trước hết, dữ liệu omics có số chiều rất lớn, nhiều nhiễu và phân bố không đồng nhất giữa các loại dữ liệu (RNA-seq, miRNA, methylation, CNV…). Thông thường các kiểu học máy truyền thống khó khai thác đầy đủ cấu trúc phi tuyến và mối quan hệ phức tạp giữa các tầng thông tin sinh học. Bên cạnh đó, nhiều mô hình hiện tại chỉ xử lý độc lập từng loại omics, hoặc đơn thuần ghép nối chúng theo kiểu concatenation mà không xem xét các quan hệ sinh học tiềm ẩn giữa các phân tử và giữa các bệnh nhân.

Một thách thức khác nằm ở việc xây dựng biểu diễn dữ liệu (embeddings) có khả năng khái quát tốt, vừa giữ được thông tin sinh học nội tại của từng mẫu, vừa phản ánh được mối tương quan giữa các mẫu trong không gian biểu hiện gen. Các mô hình mạng nơ-ron đồ thị (GNNs) đã cho thấy tiềm năng mạnh mẽ trong việc học từ cấu trúc mạng, song phần lớn nghiên cứu chỉ sử dụng các mô hình dựa trên tích chập (GCN) và chưa khai thác hiệu quả cơ chế attention – vốn giúp chỉ rõ mức độ quan trọng khác nhau giữa các mối quan hệ sinh học.

Từ những vấn đề trên, đề án đặt ra các câu hỏi nghiên cứu chính sau:

- Làm sao để tạo ra mô hình học biểu diễn tận dụng tốt cấu trúc đồ thị của dữ liệu sinh học, thay vì chỉ dựa vào ma trận biểu hiện gen thuần tuý?
- Có thể xây dựng một pipeline tích hợp từ tiền xử lý dữ liệu, xây dựng đồ thị, học biểu diễn đến phân nhóm nhằm đánh giá hiệu quả của mô hình trên dữ liệu TCGA thực tế?

Từ các câu hỏi nghiên cứu này, đề án hướng tới việc phát triển mô hình omicsGAT – một kiến trúc học biểu diễn dựa trên Graph Attention Network – nhằm khai thác một cách hiệu quả mối quan hệ giữa các mẫu bệnh nhân trong dữ liệu omics. Mục tiêu là đánh giá khả năng của mô hình trong việc phát hiện các nhóm ung thư tiềm ẩn trên dữ liệu TCGA, so sánh với các phương pháp cổ điển, kế hợp xây dựng một công cụ trực quan hỗ trợ ứng dụng.

## 1.6. Kết chương

Chương này đã nêu rõ các lý thuyết cơ sở cho toàn bộ đề án, bao gồm các khái niệm quan trọng về dữ liệu omics và multi-omics, đặc điểm phức tạp của dữ liệu sinh học phân tử cũng như nhu cầu cấp thiết trong việc tích hợp và khai thác hiệu quả các tầng dữ liệu này. Những nội dung về genomics, transcriptomics, proteomics, metabolomics và epigenomics giúp hình thành cái nhìn toàn diện về cấu trúc và chức năng sinh học từ mức DNA đến mức biểu hiện phân tử.

Bên cạnh đó, chương này cũng tổng quan các phương pháp học biểu diễn phổ biến trong sinh học tính toán, từ học có giám sát, học không giám sát đến các mô hình hiện đại dựa trên đồ thị. Đặc biệt, Graph Attention Network (GAT) được phân tích như một hướng tiếp cận tiềm năng và phù hợp cho dữ liệu omics nhờ khả năng học trọng số động, mô hình hóa mối quan hệ phi tuyến giữa các mẫu sinh học và giảm tác động của nhiễu. Các nghiên cứu gần đây cũng chỉ ra rằng các mô hình đồ thị dị thể và cơ chế attention mang lại tiến bộ đáng kể trong phân tích dữ liệu ung thư.

# Chương 2: XÂY DỰNG MÔ HÌNH HỌC BIỂU DIỄN VỚI MULTI OMICS

Chương 2 trình bày quá trình xây dựng mô hình học biểu diễn multi-omics dựa trên kiến trúc omicsGAT, trong đó dữ liệu đầu vào cơ bản là biểu hiện gene và ma trận tương quan giữa các mẫu bệnh mô tả quan hệ tương tác giữa các gene. Mô hình sử dụng omicsGAT encoder với cơ chế multi-head attention, các GAT layer và hàm kích hoạt ELU để trích xuất embedding $Z$ giàu thông tin sinh học. Kiến trúc được thiết kế theo dạng đa nhiệm với hai nhánh ra: nhánh có giám sát thực hiện phân lớp ung thư, và nhánh không giám sát với Decoder phục vụ tái tạo đặc trưng và học biểu diễn ổn định.

Trên cơ sở embedding thu được, nội dung chương cũng trình bày phương pháp phân nhóm và phân cụm ung thư bằng omicsGAT và omicsGAT clustering. Tiếp đó, đề án mô tả quy trình huấn luyện, đánh giá mô hình và cách thức hiển thị trực quan hóa kết quả phân nhóm ung thư.

## 2.1 Tổng quan về mô hình omicsGAT

Trong bài toán phân nhóm ung thư dựa trên dữ liệu đa omics, việc học biểu diễn dữ liệu có ý nghĩa sinh học đóng vai trò quan trọng, do dữ liệu có số chiều lớn và chứa nhiều nhiễu. Các phương pháp học máy truyền thống thường biểu diễn mỗi bệnh nhân một cách độc lập dưới dạng vector đặc trưng, từ đó chưa khai thác được mối quan hệ tiềm ẩn giữa các bệnh nhân trong không gian đặc trưng phân tử.

Để khắc phục hạn chế này, đề án đề xuất mô hình omicsGAT (omics Graph Attention Network) – một mô hình học biểu diễn dựa trên mạng nơ-ron đồ thị với cơ chế attention. Trong mô hình omicsGAT, mỗi nút của đồ thị tương ứng với một bệnh nhân, trong khi các cạnh biểu diễn mức độ tương đồng giữa các bệnh nhân dựa trên dữ liệu omics đã được tiền xử lý. Cách tiếp cận này cho phép mô hình hóa đồng thời đặc trưng phân tử của từng bệnh nhân và cấu trúc quan hệ giữa các bệnh nhân trong toàn bộ tập dữ liệu.

Khác với các mô hình Graph Convolutional Network truyền thống, vốn giả định mức độ ảnh hưởng như nhau giữa các nút lân cận, omicsGAT sử dụng cơ chế attention để tự động học trọng số ảnh hưởng khác nhau cho từng mối quan hệ. Nhờ đó, mô hình có khả năng làm nổi bật các mối quan hệ sinh học quan trọng giữa các bệnh nhân, đồng thời giảm tác động của nhiễu – một vấn đề phổ biến trong dữ liệu đa omics.

Trong phạm vi đề án, omicsGAT đóng vai trò là mô hình trung tâm để học các vector biểu diễn (embedding) ở mức độ bệnh nhân. Các embedding này được sử dụng cho các tác vụ phân nhóm, phân cụm và trực quan hóa dữ liệu ung thư, tạo nền tảng cho các phân tích và đánh giá kết quả được trình bày trong các phần tiếp theo của luận văn.

## 2.2 Phân tích các yêu cầu đối với mô hình học biểu diễn Multi-omics

Để thiết kế kiến trúc mô hình, việc xác định rõ các yêu cầu đặt ra đối với một hệ thống học biểu diễn multi-omics là cần thiết nhằm đảm bảo mô hình vừa phù hợp với bản chất dữ liệu sinh học phức tạp, vừa đáp ứng được mục tiêu phân nhóm ung thư trong đề án này. Các yêu cầu cụ thể gồm:

### *2.2.1 Khả năng tích hợp đồng thời nhiều dạng dữ liệu omics*

Mô hình phải có khả năng tích hợp đồng thời nhiều dạng dữ liệu omics, đặc biệt là dữ liệu biểu hiện gene (gene expression) và thông tin về mạng tương tác gene (gene interaction network). Hai nguồn dữ liệu này không chỉ khác biệt về cấu trúc mà còn tồn tại sự không đồng nhất về phân phối, nhiễu đo lường và mức độ biến thiên giữa các mẫu; do đó mô hình phải học được biểu diễn thống nhất và chặt chẽ ngay cả khi dữ liệu có độ phân tán lớn.

### *2.2.2 Khả năng khai thác cấu trúc đồ thị*

Mô hình cần có khả năng khai thác được cấu trúc đồ thị tiềm ẩn trong tương tác sinh học. Các gene không hoạt động độc lập mà có sự liên kết phức tạp. Vì vậy

việc tận dụng cấu trúc adjacency matrix để lan truyền thông tin giữa các gene là yêu cầu quan trọng.

Do đó, kiến trúc mô hình cần hỗ trợ các cơ chế học trên đồ thị, đồng thời cho phép học trọng số thích nghi giữa các gene thay vì dùng trọng số cố định, nhằm phản ánh đúng trognj số và mức độ ảnh hưởng của từng tương tác sinh học.

### *2.2.3 Khả năng học biểu diễn đa nhiệm*

Mô hình cần đảm bảo khả năng học biểu diễn đa nhiệm (multi-task), bao gồm cả nhiệm vụ có giám sát (supervised) như dự đoán phân nhóm ung thư, và nhiệm vụ không giám sát (unsupervised) như tái tạo dữ liệu để ổn định embedding. Các nhiệm vụ này phải được thiết kế sao cho tương hỗ, giúp mô hình vừa có embedding phân biệt tốt cho phân lớp, vừa giữ được đặc trưng tổng quát cho phân cụm.

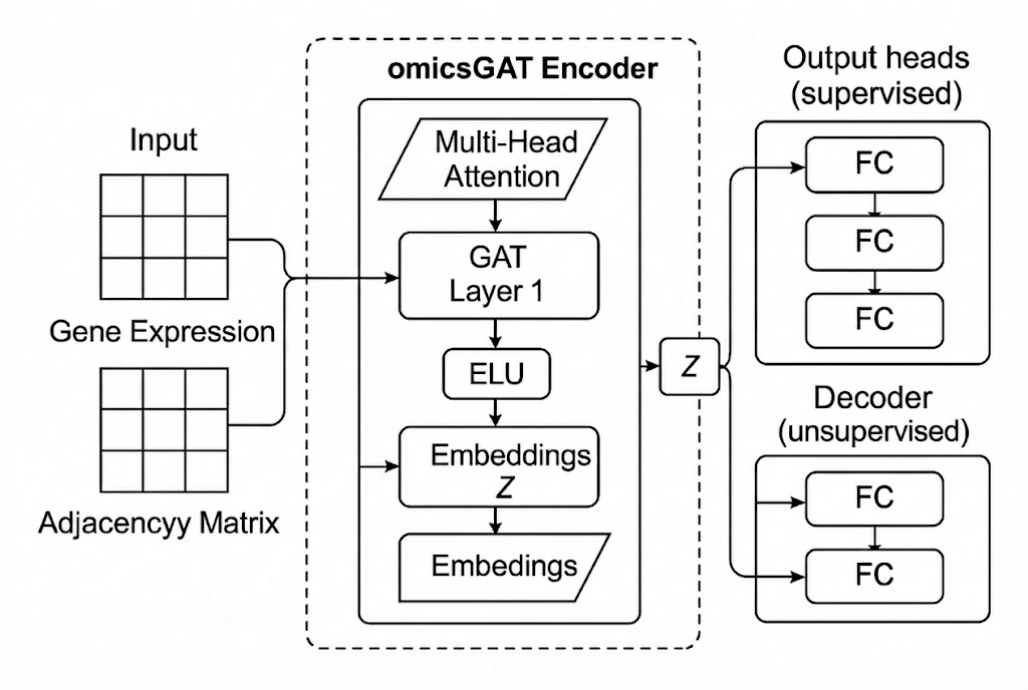
### *2.2.4 Khả năng xử lý dữ liệu đa chiều*

Dữ liệu sinh học thường có số chiều rất lớn nhưng số mẫu lại hạn chế. Điều này đòi hỏi kiến trúc phải tối ưu về số tham số, có cơ chế regularization phù hợp, và đủ linh hoạt để tránh overfitting. Ngoài ra, quá trình huấn luyện cần hỗ trợ khả năng giải thích (interpretability), đảm bảo các embedding và trọng số học được có ý nghĩa sinh học, đặc biệt khi ứng dụng trong bối cảnh phân nhóm ung thư.

### *2.2.5 Khả năng biểu diễn trực quan và hỗ trợ phân tích*

Mô hình cần tạo ra khả năng biểu diễn trực quan và giúp cho việc phân tích chuyên sâu, hỗ trợ chẩn đoán của bác sỹ. Do vậy, đầu ra của mô hình cần cho phép dễ dàng tích hợp vào các ứng dụng hiển thị trực quan kết quả phân nhóm, phân cụm.

## 2.3 Kiến trúc mô hình học biểu diễn




**Hình 2.1: Tổng quan của một mô hình omicsGAT**

Kiến trúc trong mô hình omicsGAT biểu thị ở hình 2.1 được thiết kế nhằm đáp ứng các yêu cầu đã nêu ở phần trên và cho phép khai thác đồng thời thông tin biểu hiện gen và cấu trúc mạng tương tác giữa các gen.

Mô hình sử dụng Graph Attention Network (GAT) để học các biểu diễn ẩn (embeddings) giàu ngữ cảnh sinh học dựa trên cơ chế multi-head attention, cho phép mô hình tập trung vào những mối quan hệ quan trọng trong đồ thị gen. Dữ liệu đầu vào gồm ma trận biểu hiện gen và ma trận kề (adjacency matrix) mô tả kết nối giữa các gen. Những dữ liệu này được đưa vào bộ mã hóa omicsGAT Encoder, nơi một hoặc nhiều lớp GAT được áp dụng và kích hoạt thông qua hàm ELU, tạo ra vector biểu diễn Z cho từng mẫu sinh học.

Từ embedding Z, mô hình chia ra hai hướng xử lý:

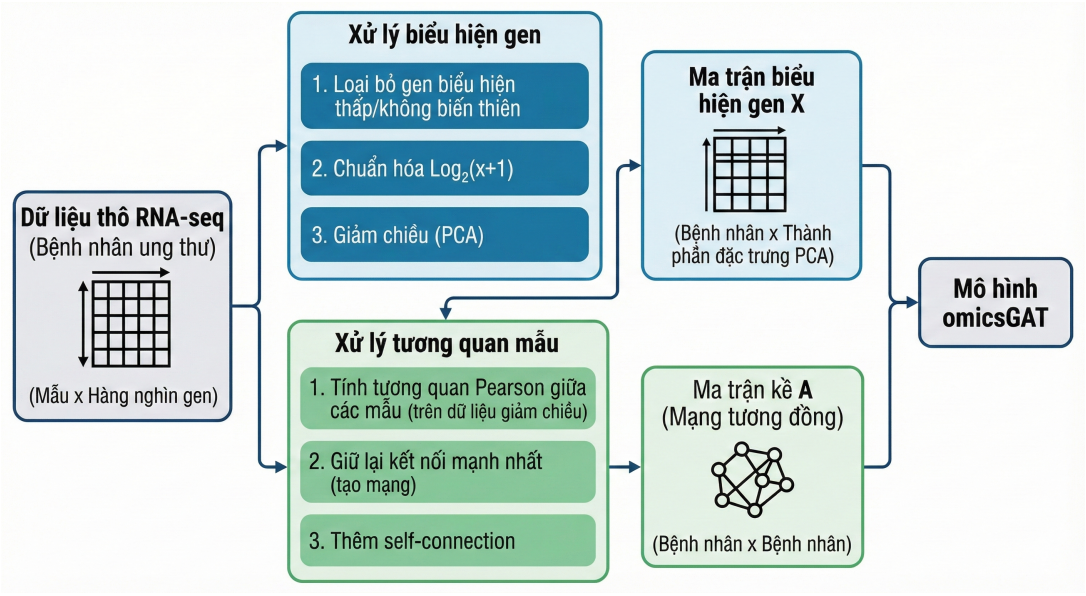
- Nhánh supervised gồm nhiều lớp fully-connected (FC) phục vụ cho các nhiệm vụ dự đoán có giám sát như phân loại hoặc dự đoán đặc tính mẫu.
- Nhánh unsupervised (decoder) tái tạo lại dữ liệu đầu vào nhằm học biểu diễn ổn định và tổng quát hơn.

Cách thiết kế hai nhánh này giúp mô hình vừa học được đặc trưng đặc thù của nhiệm vụ dự đoán, vừa đảm bảo khả năng tổng quát hóa nhờ cơ chế tự mã hóa không giám sát.

### *2.3.1 Đầu vào của mô hình*

Trong các mô hình học biểu diễn, dữ liệu đầu vào đóng vai trò then chốt trong việc xác định phạm vi và chất lượng thông tin mà mô hình có thể học được. Thông thường, dữ liệu đầu vào không chỉ bao gồm các đặc trưng mô tả đối tượng (feature vectors) mà còn có thể bao gồm mối quan hệ giữa các đối tượng, giúp mô hình học được cấu trúc tiềm ẩn của dữ liệu.

Đối với các bài toán trong lĩnh vực sinh học, đặc biệt là phân tích dữ liệu omics (như RNA-seq, scRNA-seq, proteomics,…), dữ liệu đầu vào thường mang tính đa chiều và phức tạp. Do đó, việc biểu diễn dữ liệu ban đầu dưới dạng có cấu trúc – ví dụ như ma trận biểu hiện gen hoặc mạng tương đồng giữa các mẫu – là bước quan trọng để mô hình có thể học được đặc trưng sinh học có ý nghĩa.




**Hình 2.2: Quá trình xử lý dữ liệu đầu vào**

Trong mô hình omicsGAT, dữ liệu đầu vào được xây dựng từ các thông tin biểu hiện gen RNA-seq của bệnh nhân ung thư. Nguồn dữ liệu này ban đầu ở dạng thô, thể hiện mức độ biểu hiện của rất nhiều gen trong mỗi mẫu sinh học. Trước khi

đưa vào mô hình, dữ liệu được xử lý qua nhiều bước để đảm bảo tính ổn định và giảm nhiễu.

Trước hết, các gen có mức biểu hiện thấp hoặc không biến thiên giữa các mẫu được loại bỏ. Tiếp theo, dữ liệu được chuẩn hóa bằng phép $\log_2(x+1)$ nhằm làm trơn phân phối và giảm ảnh hưởng của các giá trị bên ngoài. Để giảm chiều và loại bỏ dư thừa, tác giả áp dụng phân tích thành phần Chính và giữ lại các thành phần chính có giá trị thông tin cao nhất. Kết quả thu được là ma trận biểu hiện gen $X$, trong đó mỗi hàng tương ứng với một bệnh nhân và mỗi cột biểu diễn một thành phần đặc trưng tổng hợp từ các gen.

Song song với đó, mô hình omicsGAT còn khai thác sự tương quan giữa các mẫu bệnh nhân, được biểu diễn thông qua ma trận kề $A$. Ma trận này được xây dựng dựa trên hệ số tương quan Pearson giữa các mẫu trong tập dữ liệu biểu hiện gen đã giảm chiều. Chỉ những kết nối có tương quan mạnh nhất mới được giữ lại, nhằm làm nổi bật các mối quan hệ sinh học đáng tin cậy. Ngoài ra, mô hình thêm self-connection (tự kết nối) để mỗi bệnh nhân có thể duy trì thông tin của chính mình trong quá trình lan truyền trên đồ thị.

### *2.3.2 Lớp xử lý chính*

Trong các mô hình học biểu diễn (Representation Learning Models), lớp xử lý chính đóng vai trò cốt lõi trong việc trích xuất đặc trưng và học biểu diễn tiềm ẩn (latent representation) từ dữ liệu đầu vào. Đây là nơi mô hình thực hiện các phép biến đổi phi tuyến để chuyển đổi dữ liệu gốc có kích thước lớn, nhiễu và phức tạp thành không gian đặc trưng có ý nghĩa hơn.

Tùy vào đặc điểm của dữ liệu, các mô hình học biểu diễn có thể sử dụng những kiến trúc khác nhau cho lớp xử lý chính. Với dữ liệu ảnh, phổ biến là các mạng tích chập (Convolutional Neural Networks – CNNs) để học đặc trưng không gian. Với dữ liệu chuỗi hoặc văn bản, thường dùng mạng tuần tự (Recurrent Neural Networks – RNNs) hoặc Transformers để học phụ thuộc theo thời gian hoặc ngữ cảnh. Còn với dữ liệu có cấu trúc mạng, chẳng hạn như dữ liệu sinh học, xã hội

hoặc giao thông, các mô hình hiện đại thường sử dụng mạng nơ-ron đồ thị để mô hình hóa mối quan hệ giữa các đối tượng.

Trong các mô hình dựa trên GNN, lớp xử lý chính thường bao gồm hai bước chính: Lan truyền thông tin và Tổng hợp đặc trưng. Tuy nhiên, điểm hạn chế của hầu hết các mô hình GNN truyền thống (như GCN – Graph Convolutional Network) là chúng coi tất cả các nút lân cận có tầm quan trọng như nhau, trong khi trên thực tế, mức độ ảnh hưởng giữa các mẫu sinh học không đồng đều.

Trong mô hình omicsGAT, lớp xử lý chính là **Graph Attention Encoder**, được thiết kế để học trọng số chú ý (attention weights) giữa các mẫu trong đồ thị, từ đó cho phép mô hình tự động xác định mức độ ảnh hưởng của từng mẫu lân cận đối với mẫu đang được xem xét.

Cụ thể, mỗi bệnh nhân được biểu diễn bằng một vector đặc trưng từ ma trận đầu vào X, và các mối liên kết giữa họ được thể hiện trong ma trận kề A. Mỗi nút i (đại diện cho một bệnh nhân) học cách tổng hợp thông tin từ các nút lân cận j ∈ Ni thông qua cơ chế attention như trong công thức (2.1):

$$c_{ij} = a^T\left[Wx_i \parallel Wx_j\right] \tag{2.1}$$

trong đó, $W$ là ma trận trọng số dùng để biến đổi không gian đặc trưng, $a$ là vector trọng số attention, và || là phép nối vector.

Sau khi tính toán giá trị tương tác $c_{ij}$, mô hình sử dụng hàm Softmax để chuẩn hóa, tạo ra trọng số chú ý $\alpha_{ij}$ thể hiện mức độ quan trọng của mẫu $j$ đối với mẫu $i$ theo công thức (2.2)

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(c_{ij}\right)\right)}{\sum_{r\in N_i} \exp\left(\text{LeakyReLU}(c_{ir})\right)} \tag{2.2}$$

Nhờ vậy, mô hình có thể "tập trung" vào những láng giềng mang thông tin hữu ích hơn, bỏ qua các kết nối ít liên quan hoặc nhiễu.

Tiếp theo, biểu diễn mới của mẫu $i$ được tính bằng cách tổng hợp có trọng số các đặc trưng từ các láng giềng như công thức (2.3)

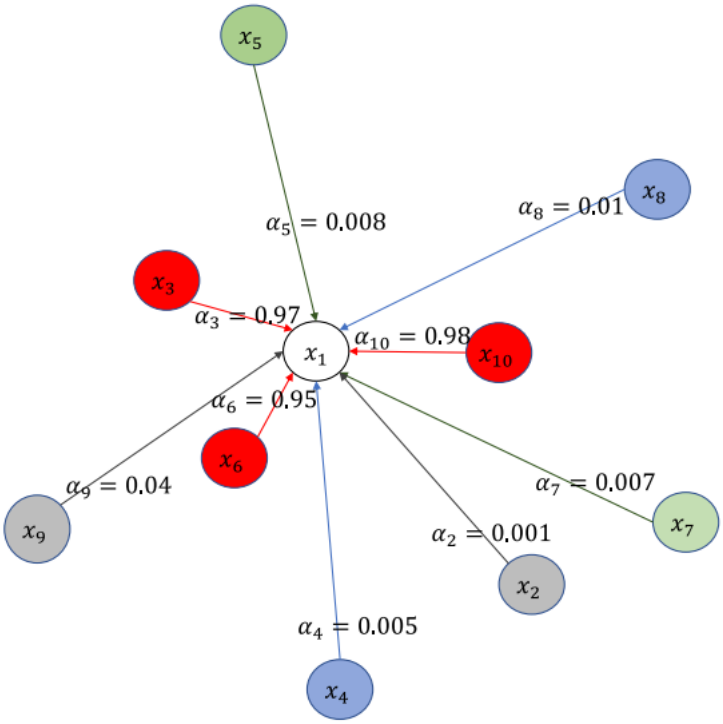
$$x_i' = \sigma\left(\sum_{j \in N_i} \alpha_{ij} W\, x_j\right) \tag{2.3}$$

trong đó $\sigma$ là hàm kích hoạt phi tuyến (thường là ReLU).

Để tăng tính ổn định và khả năng biểu diễn đa chiều, omicsGAT áp dụng multi-head attention, tức là có nhiều "đầu" attention song song cùng học trọng số giữa các mẫu theo các hướng nhìn khác nhau. Kết quả từ các head này được nối lại để tạo thành vector embedding cuối cùng cho mỗi mẫu theo công thức (2.4)

$$z_i = \|_{k=1}^{h} \sigma\left(\sum_{j \in N_i} \alpha_{ij}^{(k)} W^{(k)} x_j\right) \tag{2.4}$$

Trong đó, $h$ là số lượng head, còn $W^{(k)}$ và $\alpha^{(k)}$là tham số riêng của từng head.




**Hình 2.3: Quan hệ giữa các nút dựa trên trọng số**

Nhờ cơ chế này, mô hình omicsGAT có thể học được các mối quan hệ khác nhau giữa các mẫu — chẳng hạn như nhóm mẫu cùng kiểu phân tử, cùng phản ứng thuốc, hoặc có biểu hiện gen tương tự — thay vì chỉ dựa vào mức tương quan tuyến tính đơn giản.

Lớp xử lý chính của omicsGAT không chỉ đóng vai trò trích xuất đặc trưng biểu sinh (molecular features) mà còn mô hình hóa mạng tương tác giữa các bệnh nhân, giúp mô hình hiểu được cấu trúc dữ liệu ở cả cấp độ cục bộ (local) và toàn cục (global). Nhờ việc sử dụng attention đa đầu, mô hình đạt được khả năng học biểu diễn ổn định, linh hoạt và có tính khái quát trội hơn khi đặt cạnh các phương pháp dựa trên GCN truyền thống.

Các embedding được sinh ra từ lớp này chính là nền tảng cho các tác vụ downstream như phân loại phân nhóm ung thư (omicsGAT Classifier) hoặc phân cụm bệnh nhân ung thư (omicsGAT Clustering). Điều này chứng minh rằng, việc đưa cơ chế attention vào quá trình lan truyền thông tin trên đồ thị đã giúp mô hình học được mức độ ảnh hưởng sinh học khác nhau giữa các bệnh nhân, từ đó cải thiện rõ rệt khả năng phân biệt và nhóm mẫu trong không gian đặc trưng.

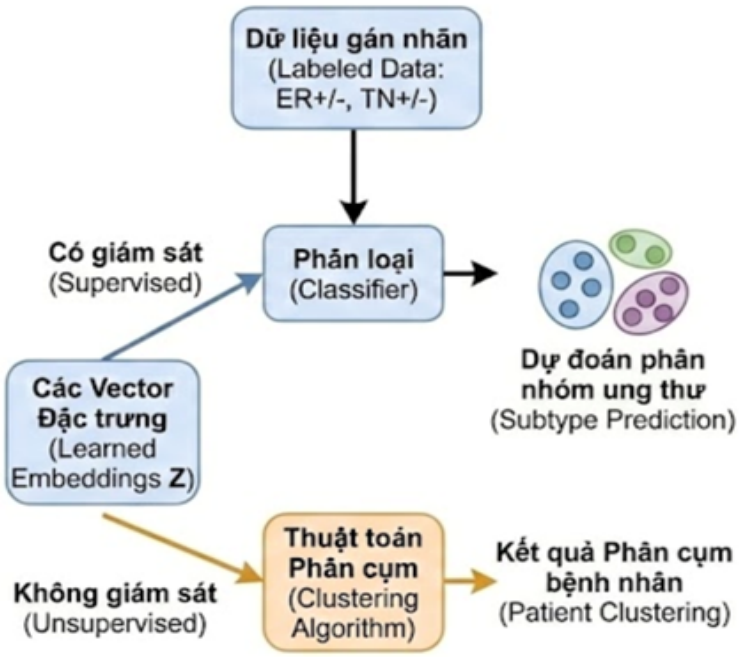
### *2.3.3 Đầu ra của mô hình*

Trong các mô hình học biểu diễn (Representation Learning Models), đầu ra là phần thể hiện kết quả cuối cùng của quá trình trích xuất đặc trưng từ dữ liệu đầu vào. Nếu như lớp xử lý chính đảm nhiệm nhiệm vụ biến đổi dữ liệu thông qua các phép học phi tuyến, thì đầu ra chính là không gian đặc trưng mới (latent space) mà mô hình đã học được.

Đầu ra này đóng vai trò trung gian giữa dữ liệu gốc và các tác vụ downstream như phân loại, phân cụm hay dự đoán. Thay vì làm việc trực tiếp với dữ liệu ban đầu có kích thước lớn, nhiều nhiễu và khó xử lý, các mô hình sẽ dựa vào embedding - các vector đặc trưng gọn nhẹ nhưng mang nhiều thông tin - nhằm tăng độ chính xác và hiệu quả của các thuật toán phía sau.

Trong hầu hết các mô hình học biểu diễn, embedding được thiết kế để thoả mãn ba tiêu chí quan trọng:

- Tính cô đọng: giảm kích thước so với dữ liệu gốc mà không làm xót các thông tin quan trọng.
- Tính phân tách: loại mẫu giống nhau nên có embedding gần nhau, các mẫu khác biệt nên cách xa nhau.
- Tính khái quát: có thể sử dụng được cho nhiều nhiệm vụ khác nhau mà không cần huấn luyện lại từ đầu.




**Hình 2.4: Dữ liệu đầu ra được sử dụng để phân loại**

Trong lĩnh vực sinh học, đặc biệt là phân tích dữ liệu omics, embedding chất lượng cao giúp mô hình nhận diện các nhóm bệnh nhân tiềm ẩn, mối liên hệ về cơ chế phân tử, và thậm chí hỗ trợ dự đoán tiên lượng lâm sàng – điều mà dữ liệu biểu hiện gen thô khó có thể đạt được.

Đối với mô hình omicsGAT, đầu ra của lớp học biểu diễn là ma trận embedding $Z$, trong đó mỗi hàng là một vector đặc trưng học được đại diện cho một bệnh nhân hoặc một mẫu sinh học. Các embedding này được tạo ra thông qua quá trình tổng hợp thông tin từ các láng giềng trên đồ thị với trọng số chú ý (attention

weights), nên chúng không chỉ chứa đặc trưng nội tại của từng mẫu mà còn phản ánh mối tương quan sinh học giữa các mẫu trong toàn bộ tập dữ liệu.

Nếu $p$ là kích thước embedding mỗi head và $h$ là số lượng head attention được sử dụng trong omicsGAT, thì embedding cuối cùng của mỗi mẫu có chiều $z_i \in R^{p \times h}$. Các embedding được nối lại từ tất cả các head để tạo thành biểu diễn giàu thông tin, đảm bảo mô hình học được nhiều kiểu quan hệ sinh học khác nhau giữa các bệnh nhân.

Các đặc trưng sinh ra từ quá trình này thường mang tính trừu tượng hóa, tức là chúng không còn trực tiếp là mức biểu hiện của từng gen, mà là sự mã hóa của các quan hệ phức tạp giữa các mẫu – thứ mà dữ liệu thô không thể hiện một cách rõ ràng. Chính vì vậy, embedding từ omicsGAT được đánh giá là giàu giá trị hơn trong các tác vụ phân tích ung thư.

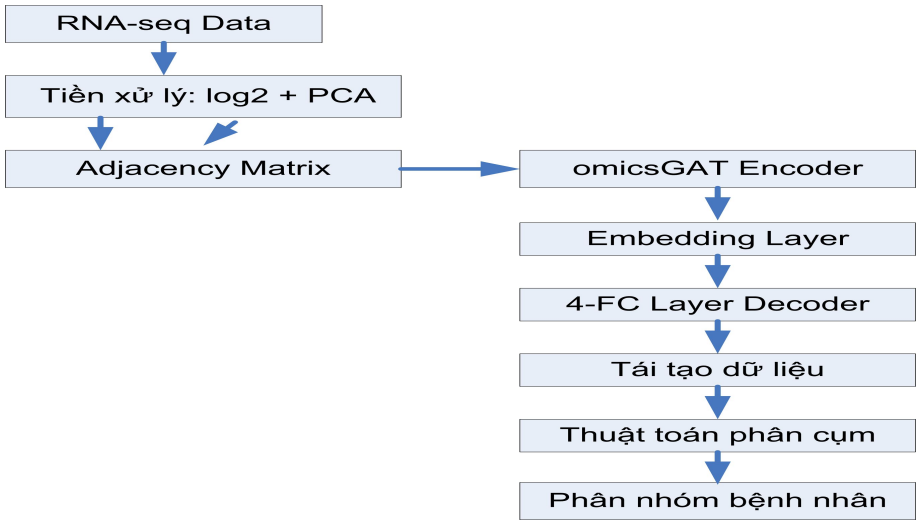
## 2.4 Phương pháp phân cụm, phân nhóm ung thư

Trong nghiên cứu ung thư hiện đại, phân nhóm ung thư (cancer substype classification) và phân cụm bệnh nhân bằng ung thư (cancer patient clustering / stratification) là hai nhiệm vụ quan trọng nhằm hiểu rõ tính dị hợp (heterogeneity) của bệnh nhân

Phân nhóm ung thư là bài toán có giám sát (supervised), trong đó mô hình học từ dữ liệu đã gán nhãn để dự đoán một loại ung thư hoặc trạng thái phân tử (ví dụ: ER+/ER–, TN+/TN–). Mục tiêu là xác định nhóm ung thư đã được định nghĩa trước, thường gắn với cơ chế sinh học hoặc đặc điểm lâm sàng. Ngược lại, phân cụm bệnh nhân ung thư là bài toán không giám sát (unsupervised). Nhiệm vụ là tìm các nhóm bệnh nhân tiềm ẩn, dựa trên sự tương đồng trong biểu hiện gen mà không cần nhãn trước đó. Kết quả phân cụm giúp phát hiện:

- Các phân nhóm mới chưa được biết đến
- Những nhóm bệnh nhân có tiên lượng hoặc đáp ứng điều trị khác nhau
- Những mô hình phân tử đặc trưng của các dạng ung thư dị hợp

Trong cả hai nhiệm vụ, học biểu diễn đóng vai trò trung tâm: các vector đặc trưng tốt sẽ giúp mô hình phân loại chính xác hơn và phân cụm rạch ròi hơn. Đây là lý do các phương pháp hiện đại dựa trên mạng nơ-ron đồ thị, đặc biệt trong dữ liệu sinh học.




**Hình 2.5: Sơ đồ phân cụm / phân nhóm**

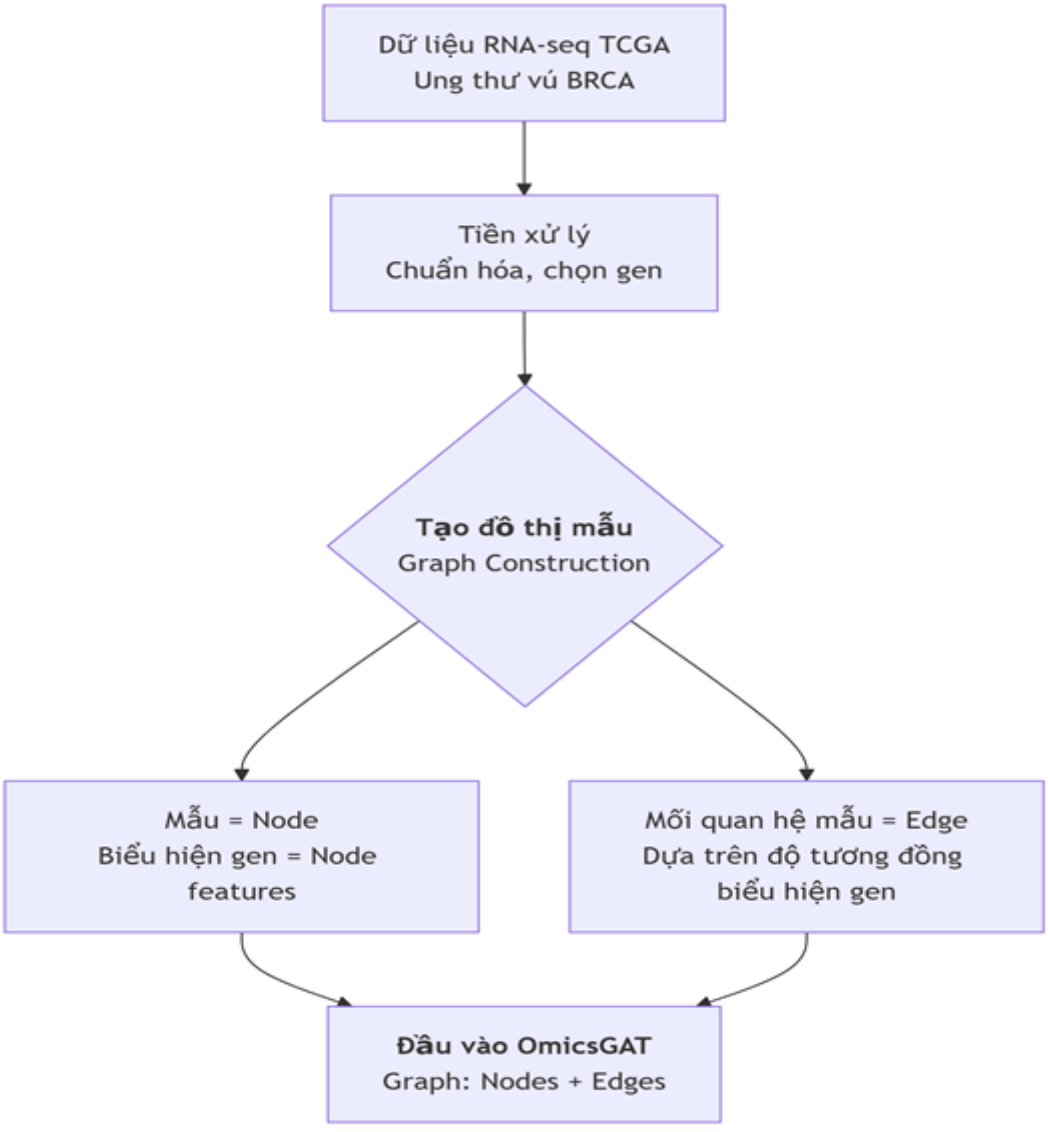
### *2.4.1 Phân nhóm ung thư bằng omicsGAT*

Phương pháp omicsGAT Classifier được thiết kế nhằm khai thác cấu trúc quan hệ giữa các mẫu trong dữ liệu RNA-seq để cải thiện chất lượng phân nhóm bệnh nhân hoặc tế bào. Đây là mô hình học có giám sát (supervised learning), được xây dựng trên nền tảng Graph Attention Network (GAT) – một dạng mạng nơ-ron đồ thị có khả năng học trọng số chú ý (attention weights) giữa các mẫu sinh học.

Mô hình omicsGAT giả định rằng các bệnh nhân (hoặc mẫu tế bào) có biểu hiện gen tương tự nhau sẽ có kết quả lâm sàng hoặc phân nhóm ung thư tương tự. Thay vì coi tất cả các mẫu "hàng xóm" (neighbor) có mức ảnh hưởng như nhau, omicsGAT sử dụng cơ chế multi-head attention để học ra mức độ quan trọng khác nhau của từng mẫu hàng xóm khi xây dựng biểu diễn (embedding) cho một bệnh nhân và kết hợp thông tin từ nhiều "hướng nhìn" khác nhau qua các head độc lập, giúp biểu diễn ổn định và giàu thông tin hơn. Mỗi lớp GAT tính toán embedding cho từng mẫu bằng cách tổng hợp biểu diễn của các láng giềng, có trọng số chú ý

riêng biệt (attention coefficient). Các embedding từ nhiều "head" được nối lại để tạo vector biểu diễn cuối cùng cho mỗi mẫu.

Embedding sẽ được đưa qua ba lớp fully connected. Hai lớp đầu sẽ sử dụng hàm kích hoạt ReLU để học đặc trưng phi tuyến trong khi lớp cuối cùng sử dụng Softmax để dự đoán xác suất mỗi mẫu thuộc về từng nhóm ung thư.




**Hình 2.6: Sơ đồ phân nhóm ung thư bằng mô hình omicsGAT**

Mô hình omicsGAT Classifier được áp dụng lên dữ liệu RNA-seq của ung thư vú (BRCA) trong bộ dữ liệu TCGA, với hai bài toán dự đoán trạng thái thụ thể Estrogen (ER+/ER–) và dự đoán trạng thái Triple Negative (TN+/TN–). So sánh với

các mô hình cơ sở gồm SVM, Random Forest, Deep Neural Network (DNN) và Graph Convolutional Network (GCN), mô hình omicsGAT đạt AUC cao nhất. Ngoài ra, embedding do omicsGAT học được còn thể hiện khả năng dự đoán tốt hơn thời gian sống sót và thời gian tái phát của bệnh nhân khi sử dụng trong mô hình hồi quy Cox, minh chứng rằng embedding này mang thông tin sinh học có giá trị cao hơn so với dữ liệu biểu hiện gen thô.

Khác với phân cụm, bài toán phân loại (phân nhóm) ung thư là bài toán học có giám sát, trong đó mô hình được huấn luyện trên dữ liệu có nhãn nhằm dự đoán loại ung thư hoặc nhóm bệnh nhân tương ứng. Trong đề án, mô hình omicsGAT không trực tiếp đảm nhiệm vai trò phân loại, mà được sử dụng để học các vector biểu diễn ở mức độ bệnh nhân từ dữ liệu đa omics. Các embedding này đóng vai trò là đầu vào cho các mô hình phân loại downstream, giúp cải thiện khả năng phân biệt giữa các nhóm ung thư khác nhau. Cách tiếp cận này cho phép tách biệt rõ ràng giai đoạn học biểu diễn và giai đoạn phân loại, đồng thời cho thấy tiềm năng ứng dụng của omicsGAT trong các hệ thống hỗ trợ chẩn đoán ung thư dựa trên dữ liệu phân tử.

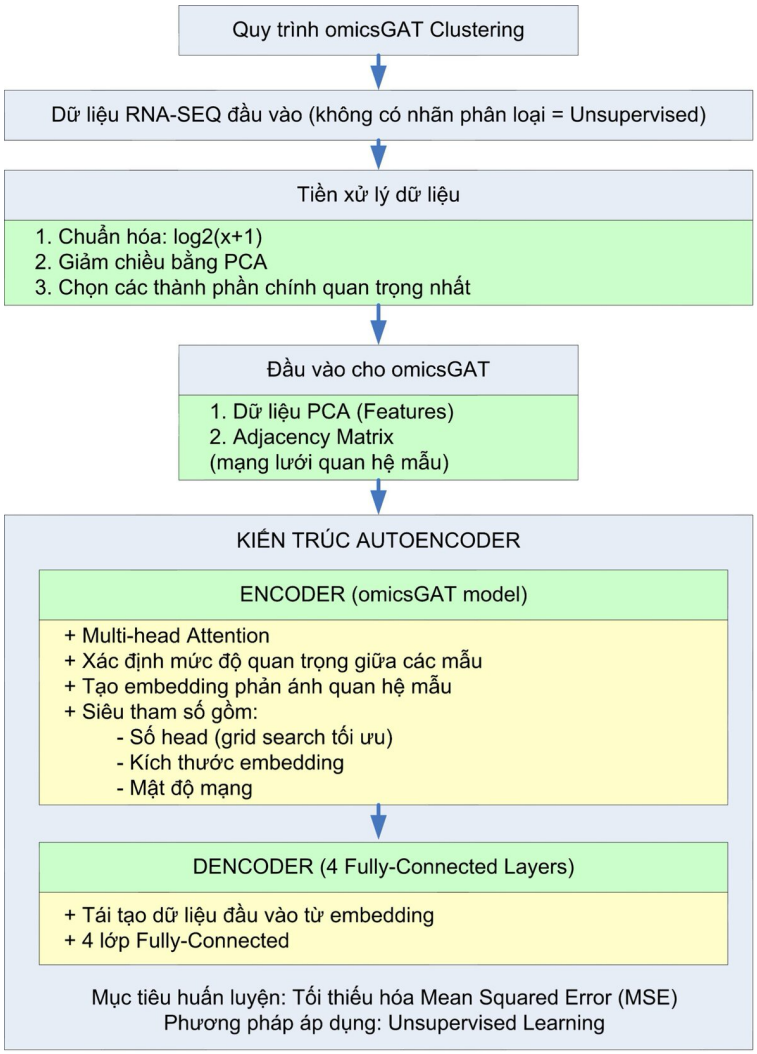
### *2.4.2 Phân cụm bằng omicsGAT Clustering*

OmicsGAT Clustering, một phiên bản học không giám sát (unsupervised) của omicsGAT, để tự động phát hiện các nhóm bệnh nhân tiềm ẩn từ dữ liệu RNA-seq mà không cần nhãn phân loại sẵn có.

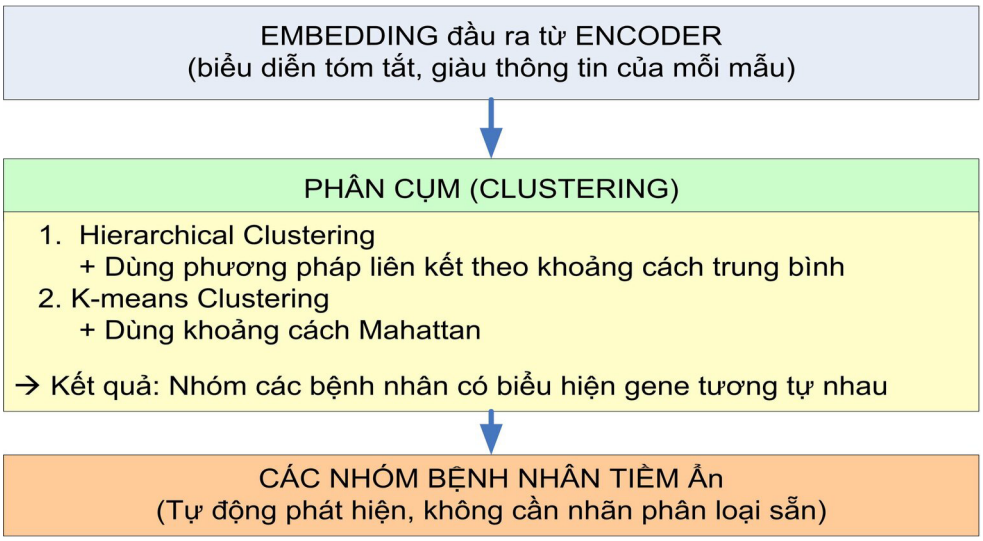
Trong giai đoạn đầu, omicsGAT đóng vai trò là encoder của kiến trúc autoencoder. Mục tiêu của encoder là học một biểu diễn gọn nhưng giàu thông tin của mỗi mẫu, dựa trên cả dữ liệu biểu hiện gene và quan hệ giữa các mẫu. Trước khi đưa vào mô hình, dữ liệu được chuẩn hóa $\log_2(x+1)$ và giảm chiều bằng PCA, chọn ra cácthành phần chính quan trọng nhất làm đầu vào cho omicsGAT. Các siêu tham số như số head, kích thước embedding và mật độ mạng được lựa chọn bằng grid search để đạt hiệu năng tối ưu. Sau đó Encoder sử dụng mô hình omicsGAT để học embedding từ dữ liệu PCA và mạng lưới quan hệ giữa mẫu (adjacency matrix).

OmicsGAT sử dụng cơ chế multi-head attention để xác định múc độ quan trọng của các mẫu và tạo embedding  phản ánh quan hệ giữa các mẫu.

Sau khi encoder tạo embedding, embedding này được đưa vào decoder gồm 4 lớp fully-connected nhằm tái tạo lại dữ liệu đầu vào. Autoencoder được huấn luyện theo phương pháp không giám sát bằng cách tối thiểu hóa sai số bình phương trung bình (MSE). Sau khi huấn luyện, embedding từ encoder được dùng cho bước phân cụm. Sau khi quá trình huấn luyện autoencoder hoàn tất, embedding được xuất ra và đưa vào thuật toán phân cụm phân cấp (Hierarchical Clustering), K-means sử dụng khoảng cách Manhattan và phương pháp liên kết Average linkage để nhóm các bệnh nhân có biểu hiện tương tự nhau.




**Hình 2.7: Phân cụm omicsGAT bằng Clustering**




**Hình 2.8: Kết quả phân cụm dựa vào Embedding đầu ra từ Encoder**

Trong đề án này, phân cụm ung thư được xem là một bài toán học không giám sát, nhằm khám phá các nhóm bệnh nhân có đặc trưng phân tử tương đồng mà không dựa trên nhãn có sẵn. Ý nghĩa chính của bài toán phân cụm không nằm ở việc dự đoán nhãn ung thư cụ thể, mà ở khả năng phát hiện các cấu trúc tiềm ẩn trong dữ liệu đa omics, từ đó gợi ý sự tồn tại của các phân nhóm ung thư mới hoặc các kiểu hình phân tử khác biệt trong cùng một loại ung thư. Việc áp dụng các phương pháp phân cụm trên không gian embedding học được từ mô hình omicsGAT giúp các cụm bệnh nhân thể hiện rõ ràng hơn so với khi sử dụng dữ liệu gốc, qua đó cho thấy embedding thu được đã phản ánh được các đặc trưng sinh học có ý nghĩa.

## 2.5 Phương pháp huấn luyện và đánh giá mô hình

### *2.5.1 Phương pháp huấn luyện*

Trong các mô hình học biểu diễn nói riêng và mô hình học sâu nói chung, phương pháp huấn luyện đóng vai trò kiến quyết trong việc đảm bảo chất lượng của biểu diễn được học và hiệu quả của các tác vụ downstream như phân loại hoặc phân cụm. Việc huấn luyện một mô hình thường gồm các thành phần:

- Hàm mất mát (Loss Function): xác định mục tiêu mà mô hình cần tối ưu

- Thuật toán tối ưu (Optimization Algorithm): điều chỉnh trọng số để hàm giảm mất mát
- Chiến lược huấn luyện (Training Strategy): cách chuẩn hoá dữ liệu, chọn batch, early stoping và các tham số như learning rate hoặc epoch.

Và cũng tuỳ thuộc vào bài toán mà mô hình có thể được huấn luyện theo hai cách phổ biến:

- Học có giám sát (Supervised Learning): mô hình học từ dữ liệu gắn nhãn, phù hợp vơi các bài toán phân loại
- Học không giám sát (Unsupervised Learning): Mô hình học cấu trúc nội tại của dữ liệu mà không cần nhãn, phù hợp cho các bài toán phân cụm hoặc giảm chiều.

Mô hình omicsGAT được huấn luyện theo hai chế độ tùy thuộc vào nhiệm vụ:

- omicsGAT Classifier: là chế độ học có giám sát, cho việc phân loại phân nhóm ung thư
- omicsGAT Clustering: là chế độ học không có giám sát, cho phân cụm bệnh nhân.

Mặc dù hai nhiệm vụ có mục tiêu khác nhau, cả hai đều sử dụng Graph Attention Encoder làm lõi học biểu diễn, nên có nhiều điểm chung trong cách huấn luyện.

Việc kết hợp hai chế độ huấn luyện – supervised và unsupervised – cho phép omicsGAT trở thành mô hình đa năng, áp dụng hiệu quả cho nhiều bài toán sinh học phân tử. Cơ chế attention giúp mô hình đánh trọng số khác nhau cho từng mối quan hệ giữa các mẫu, từ đó cải thiện đáng kể khả năng học biểu diễn so với các phương pháp GCN truyền thống.

### *2.5.2 Đánh giá mô hình*

Đánh giá mô hình (Model Evaluation) là bước thiết yếu để xác định mức độ hiệu quả của mô hình trong việc học đặc trưng từ dữ liệu và thực hiện nhiệm vụ cụ thể như phân loại, phân cụm hay dự đoán. Việc đánh giá thường được thực hiện dựa

trên tập kiểm tra độc lập hoặc thông qua các phương pháp kiểm định chéo (cross-validation), nhằm đảm bảo tính khách quan và khả năng khái quát của mô hình.

Các phương pháp đánh giá có thể khác nhau tùy vào bản chất của nhiệm vụ:

- Với bài toán phân loại (supervised learning) các chỉ số phổ biến gồm: AUC, Accuracy, Recall, Precision, F1-Score,…
- Với bài toán phân cụm (unsupervised learning) do không có nhãn đầu ra tương ứng, các chỉ số đánh giá thường dựa trên mức độ tương đồng giữa các kết quả phân cụm và cấu trúc thực tế, chẳng nhạn như: Normalize Mutual Information (NMI) hay Adjust Rand Index (ARI).
- Trong các bài toán y sinh, ngoài hiệu quả phân loại, phân cụm, mô hình cũng có thể được dánh giá dựa trên khả năng dự đoán thời gian sống sót hoặc tái phát của bệnh nhân, thông qua mô hình Cox Proportional Hazards và các thước đó như C-index.

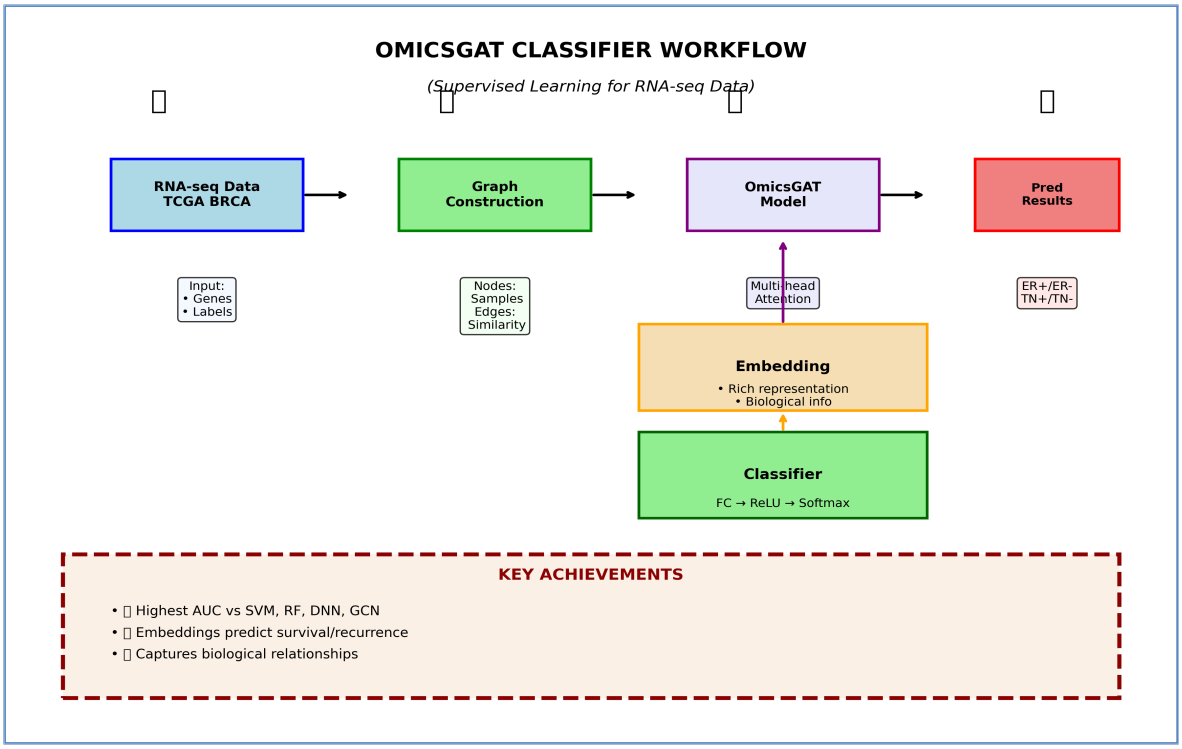
Việc lựa chọn thước đo đánh giá phù hợp là cần thiết để bảo đảm mô hình không chỉ hoạt động tốt về mặt toán học, mà còn phản ánh đúng ý nghĩa sinh học và lâm sàng của dữ liệu.

Mô hình omicsGAT được đánh giá trên nhiều khía cạnh, tương ứng với các nhiệm vụ mà mô hình đảm nhận: phân loại phân nhóm ung thư, phân cụm bệnh nhân, và dự đoán tiên lượng sống còn. Mỗi nhiệm vụ sử dụng bộ chỉ số phù hợp nhằm phản ánh tốt nhất hiệu quả của biểu diễn học được.

## 2.6 Xây dựng ứng dụng hiển thị trực quan

Mục tiêu của ứng dụng hiển thị trực quan là tạo một giao diện cho người dùng để có thể tương tác (nhà nghiên cứu, bác sĩ, hoặc nhà phân tích dữ liệu) để:

- Khám phá và giải thích các embedding do mô hình omicsGAT sinh ra
- So sánh kết quả phân loại và phân cụm
- Truy vết các mối quan hệ mẫu




**Hình 2.9: Ví dụ về một giao diện hiển thị trực quan kết quả**

## 2.7 Kết chương

Sau khi tìm hiểu các kiến thức căn bản về đa omics và các phương pháp học biểu diễn, trong chương này tìm hiểu kĩ quá trình thiết kế mô hình học biểu diễn omicsGAT, từ việc xây dựng kiến trúc tổng thể đến mô tả các thành phần chính của mô hình. Các khối xử lý gồm đầu vào (input), lớp mã hóa dựa trên Graph Attention Network, và đầu ra (embedding) đã được phân tích rõ ràng về vai trò, cơ chế hoạt động cũng như lý do mô hình này phù hợp với dữ liệu omics.

Trong phần trình bày trên còn nói tới hai hướng ứng dụng của mô hình gồm omicsGAT Classifier (phân nhóm ung thư có giám sát) và omicsGAT Clustering (phân cụm bệnh nhân không giám sát). Mỗi hướng được thiết kế nhằm đáp ứng với yêu cầu từng bài toán, nhưng đều khai thác chung nền tảng embedding do mô hình GAT cung cấp, giúp tận dụng hiệu quả cấu trúc mạng giữa các mẫu bệnh nhân. Mô tả phương pháp huấn luyện, lựa chọn tham số, cơ chế tối ưu và tiêu chí đánh giá cũng được làm rõ nhằm đảm bảo mô hình vận hành ổn định và có khả năng tổng quát tốt.

# Chương 3: TRIỂN KHAI THỬ NGHIỆM

Chương này trình bày toàn bộ quy trình triển khai thử nghiệm mô hình học biểu diễn sử dụng dữ liệu đa omics cho bài toán phân nhóm ung thư. Các nội dung bao gồm: thu thập và chuẩn bị dữ liệu từ TCGA, tiền xử lý từng loại dữ liệu omics, xây dựng đồ thị bệnh nhân, huấn luyện mô hình omicsGAT, và cuối cùng là phân tích – đánh giá kết quả thử nghiệm dựa trên các chỉ số phân nhóm và phân tích sống còn.

## 3.1 Thu thập dữ liệu từ TCGA

Dữ liệu được lấy từ cơ sở dữ liệu TCGA (The Cancer Genome Atlas), được đánh giá là nguồn dữ liệu ung thư lớn trên thế giới với hệ thống dữ liệu được chuẩn hóa ở mức độ cao. Đối với đề tài, năm loại ung thư phổ biến được lựa chọn:

- BRCA (Breast Invasive Carcinoma)
- LUAD (Lung Adenocarcinoma)
- GBM (Glioblastoma Multiforme)
- KIRC (Kidney Renal Clear Cell Carcinoma)
- UCEC (Uterine Corpus Endometrial Carcinoma)

Và cho mỗi bệnh nhân sẽ có bốn loại dữ liệu omics được thu thập, bao gồm:

- mRNA expression (RNA-seq)
- DNA methylation
- Copy Number Variation (CNV)
- miRNA expression

Mỗi loại dữ liệu omics sau khi tiền xử lý đều được chuẩn hóa và chuyển về chung một định dạng đầu vào dạng ma trận số (numeric matrix) với cấu trúc:

- Hàng (rows): các mẫu bệnh nhân (Sample ID)
- Cột (columns): các đặc trưng (features) như gene expression, CpG methylation sites, CNV gene-level score hoặc miRNA expression

- Giá trị (cells): giá trị số liên tục (float) biểu diễn mức độ hoạt động sinh học tương ứng

### *3.1.1 Các field (features) chính trong dữ liệu*

Dữ liệu RNA-seq được sử dụng để thử nghiệm ban đầu mang cấu trúc như bảng dưới đây.

**Bảng 3.1: Cấu trúc của file RNA-seq**

| Tên trường | Kiểu dữ liệu | Mô tả |
|---|---|---|
| Sample | String | Mã mẫu bệnh nhân, ví dụ TCGA-AR-A5QQ-01 |
| Gene1 | Float | Biểu hiện gene 1 |
| Gene2 | Float | Biểu hiện gene 2 |
| … | Float | … |
| Gene | Float | Biểu hiện gene N |

## 3.2 Tiền xử lý dữ liệu

### *3.2.1 Loại bỏ các gene không hợp lệ hoặc ít thông tin*

Với đầu vào là một lượng lớn các dữ liệu, ta cần loại bỏ các gene không có biểu hiện hoặc không mang giá trị. Các gene có tổng đếm (count) rất thấp thường là nhiễu hoặc không được biểu hiện trong mô ung thư. Hoặc nếu một loại gene có phương sai thấp, có biểu hiện gần như giống nhau ở mọi bệnh nhân sẽ không giúp phân biệt nhóm cũng sẽ bị loại bỏ

```
# Remove invalid gene
df = df[df.sum(axis=1) > 10]
df = df[df.std(axis=1) > 0.5]
df = df.fillna(df.median(axis=1))
```

**Hình 3.1: Code loại bỏ các gene không hợp lệ**

### *3.2.2 Chuẩn hoá dữ liệu*

#### *a. RNA-seq / miRNA*

RNA-seq có phân phối rất lệch (skewed), với một số gene có biểu hiện cực lớn nên sẽ dùng log2 nhằm giúp làm trơn dữ liệu, giảm ảnh hưởng gene quá lớn và làm phân phối chuẩn

```
df = df.apply(lambda x: np.log2(x + 1))
df = df.apply(lambda x: (x - x.mean()) / x.std(), axis=1)
```

**Hình 3.2: Code chuẩn hoá dữ liệu cho RNA-seq**

#### *b. Methylation Data*

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()

df_scaled = pd.DataFrame(
    scaler.fit_transform(df.T),
    index=df.columns,
    columns=df.index
).T
```

**Hình 3.3: Code chuẩn hoá dữ liệu cho Methylation**

#### *c. CNV (Copy Number Variation)*

CNV từ TCGA thường là log2 copy-number ratio, có nhiễu mạnh và thường không đồng đều. Vậy nên cần phải cắt ngưỡng về [-3,3] để giảm giá trị outliers.

```
df = df.clip(lower=-3, upper=3)
```

**Hình 3.4: Code chuẩn hoá dữ liệu cho CNV**

### *3.2.3 Đồng bộ giữa nhiều loại omics và xuất file*

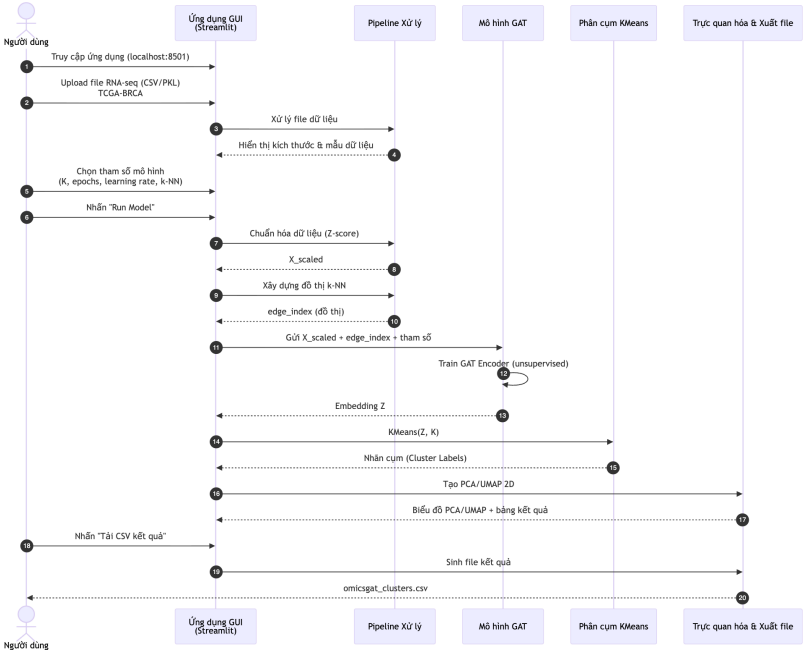
```
common_samples = list(
    set(df_mrna.columns)
    & set(df_methyl.columns)
    & set(df_cnv.columns)
    & set(df_mirna.columns)
)

df_mrna = df_mrna[common_samples]
df_methyl = df_methyl[common_samples]
df_cnv = df_cnv[common_samples]
df_mirna = df_mirna[common_samples]
```

**Hình 3.5: Code xuất file input**

## 3.3 Chạy thử mô hình omicsGAT và ứng dụng hiển thị

### *3.3.1 Cấu trúc và workflow của ứng dụng*




***Hình 3.6: Workflow của ứng dụng mô phỏng***

Trước tiên, người dùng truy cập giao diện ứng dụng thông qua trình duyệt web và tải lên tệp dữ liệu RNA-seq (ví dụ bộ dữ liệu TCGA-BRCA). Sau khi dữ liệu được nhập vào, hệ thống hiển thị thông tin cơ bản để xác nhận người dùng tải đúng định dạng. Tiếp đó, người dùng lựa chọn các tham số mô hình như số cụm K, số epoch huấn luyện, learning rate hoặc số láng giềng k trong đồ thị k-NN.

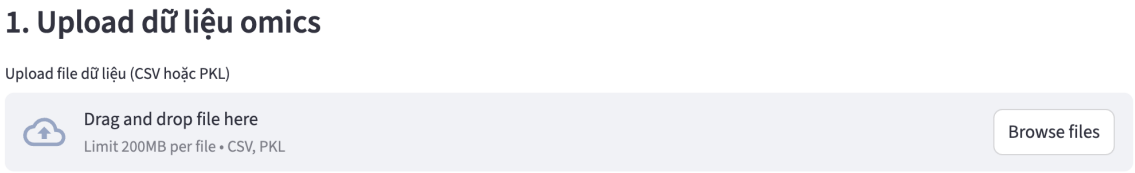
Khi người dùng nhấn nút Run Model, toàn bộ pipeline xử lý được kích hoạt. Ứng dụng thực hiện chuẩn hóa dữ liệu, xây dựng đồ thị k-NN, huấn luyện mô hình GAT để học biểu diễn cho từng mẫu dữ liệu, sau đó áp dụng KMeans để gán nhãn cụm. Cuối pipeline, hệ thống thực hiện giảm chiều PCA/UMAP để trực quan hóa và hiển thị kết quả phân cụm trên giao diện.

Người dùng có thể xem bảng kết quả tương ứng với từng bệnh nhân, đồng thời tải xuống file CSV chứa nhãn phân nhóm để phục vụ cho các phân tích tiếp theo. Sơ đồ này minh hoạ rõ ràng chuỗi hành động từ phía người dùng và phản hồi của hệ thống, đảm bảo cái nhìn tổng quan về cách ứng dụng hoạt động từ đầu đến cuối.

### *3.3.2 Upload dữ liệu và cấu hình tham số*

Ứng dụng được phát trển bằng Streamlit, cho phép chạy trên môi trường local với giao diện web thân thiện. Giao diện gồm hai phần chính:

***a. Khu vực upload***




**Hình 3.7: Phần upload dữ liệu của người dùng**

Người dùng lựa chọn tệp dữ liệu RNA-seq ở định dạng CSV.

Sau khi tải lên, ứng dụng tự động:

- Đọc dữ liệu
- Hiển thị kích thước (số mẫu, gene)

- Xem trước một vài dòng đầu để xác định định dạng

### *b. Cấu hình tham số mô hình*




**Hình 3. 8: Phần điều chỉnh tham số huấn luyện**

Tại sidebar, người dùng có thể điều chỉnh các tham số:

- Số cụm K: số nhóm cần phân
- Số epoch huấn luyện: số vòng lặp tối ưu GAT
- Learning rate: tốc độ học của optimizer
- k của k-NN graph: số hàng xóm khi xây dựng đồ thị mẫu-mẫu

Các tham số này có thể thay đổi linh hoạt tuỳ mục tiêu phân tích.

## *3.3.3 Huấn luyện mô hình*

Sau khi cấu hình, người dùng bấm nút “Run Model” để bắt đầu quá trình huấn luyện. Ứng dụng sẽ được thực hiện theo từng bước như sau.

### *a. Chuẩn hóa dữ liệu bằng Z-score.*

Sau khi người dùng tải tệp CSV/PKL vào hệ thống, dữ liệu được thực hiện qua từng bước:

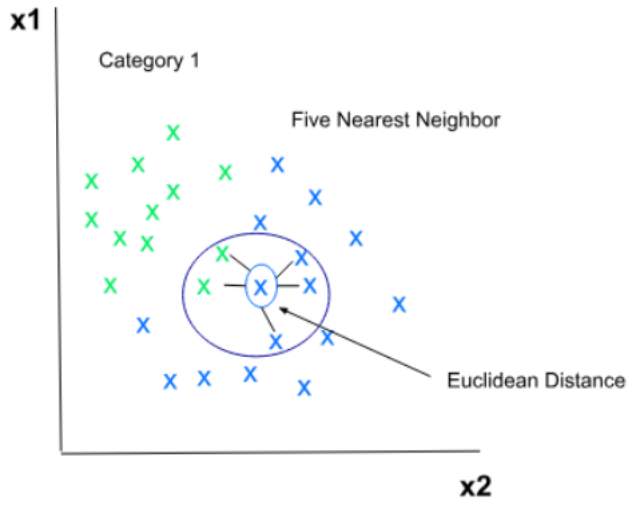
- Chuẩn hoá cấu trúc dữ liệu: Sẽ xác định cột đầu tiên là SampleID, các cột còn lại phải là dạng numeric float và loại bỏ các cột không hợp.

- Lựa chọn đặc trưng: Với các bộ dữ liệu có số lượng đặc trưng lớn, hệ thống sẽ chon lọc top-k đặc trưng có phương sai cao nhất. Việc lựa chọn sẽ giúp giảm nhiễu, giảm chi phí tính toán và cải thiện chất lượng embedding.
- Chuẩn hoá giá trị: Áp dụng Z-score normalization cho từng đặc trưng để đưa các đặc trưng về cùng thang đo, tránh trường hợp những gene có giá trị lớn chi phối mô hình.

***b. Xây dựng đồ thị k-NN giữa các mẫu dựa trên khoảng cách Euclid.***

Dữ liệu omics sau khi chuẩn hóa được xem như một tập các vector đặc trưng có kích thước lớn. Để mô hình GAT có thể khai thác mối quan hệ giữa các mẫu bệnh nhân, hệ thống xây dựng đồ thị k-Nearest Neighbors (k-NN) theo các bước:

- Xác định số láng giềng gần nhất k: Giá trị k được cấu hình bởi người dùng.
- Tính khoảng cách giữa các mẫu: Sử dụng khoảng cách Euclidean để đo mức độ tương đồng.
- Xây dựng danh sách cạnh: với mỗi mẫu i, chọn k mẫu gần nhất để tạo cạnh (i,j). Tập các cạnh được ghép thành ma trận edge_index theo chuẩn PyTorch Geometric. Kết quả sẽ là một đồ thị vô hướng G = (V, E)




**Hình 3.9: Đồ thị K-NN giữa các mẫu**

***c. Huấn luyện mô hình Graph Attention Network để học biểu diễn*** (embeddings) độ chiều thấp.

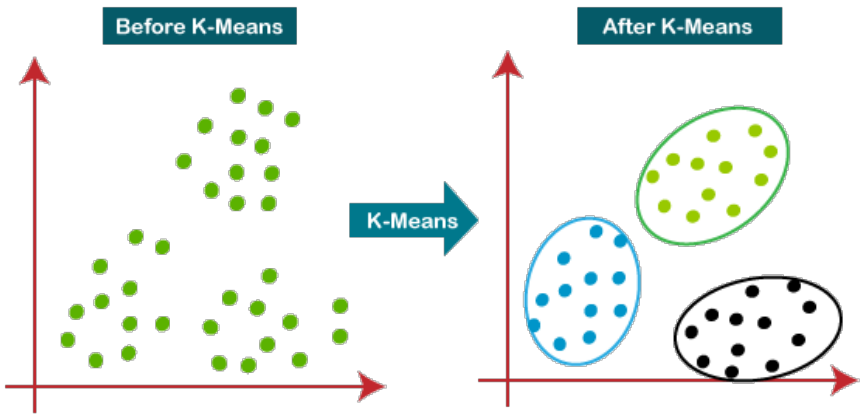
Mô hình huấn luyện sẽ gồm hai lớp chính:

- Lớp GATConv thứ nhất (multi-head attention): Nhận đầu vào là ma trận đặc trưng $X \in R^{N \times D \cdot}$ Học representation phi tuyến bằng cơ chế Attention rồi Sử dụng nhiều head để tăng khả năng biểu diễn.
- Lớp GATConv thứ hai: Nén không gian biểu diễn xuống chiều thấp hơn ($d$ = 16 - 32). Cho ra embedding cuối cùng Z=GAT(X,G)
- Hàm mất mát không giám sát: Mô hình không có nhãn (unsupervised), nên sử dụng smoothness loss, nhằm làm cho embedding của mỗi node "gần" embedding trung bình của hàng xóm.
- Tối ưu: Sử dụng các chỉ số Learning rate, Epochs do người dùng chọn. Trong mỗi epochs truyền xuôi qua hai lớp GATConv. Tính embedding, tính loss theo smoothness và cập nhật trọng số.

***d. Thực hiện phân cụm bằng thuật toán KMeans trên embedding thu được.***

Sau khi có embedding $Z$:

- Áp dụng thuật toán Kmeans: số cụm k do người dùng chọn. Sau đó Kmeans chạy trên embedding.




**Hình 3.10: Thuật toán phân cụm K-Means**

- Nhận nhãn cụm: Mỗi mẫu sẽ được gán 1 cluster ID. Embedding Z giúp dễ tách nhóm hơn vì giảm chiều, khai thác được cấu trúc giữa các bệnh nhân và loại bỏ được nhiễu từ đặc trưng ban đầu.

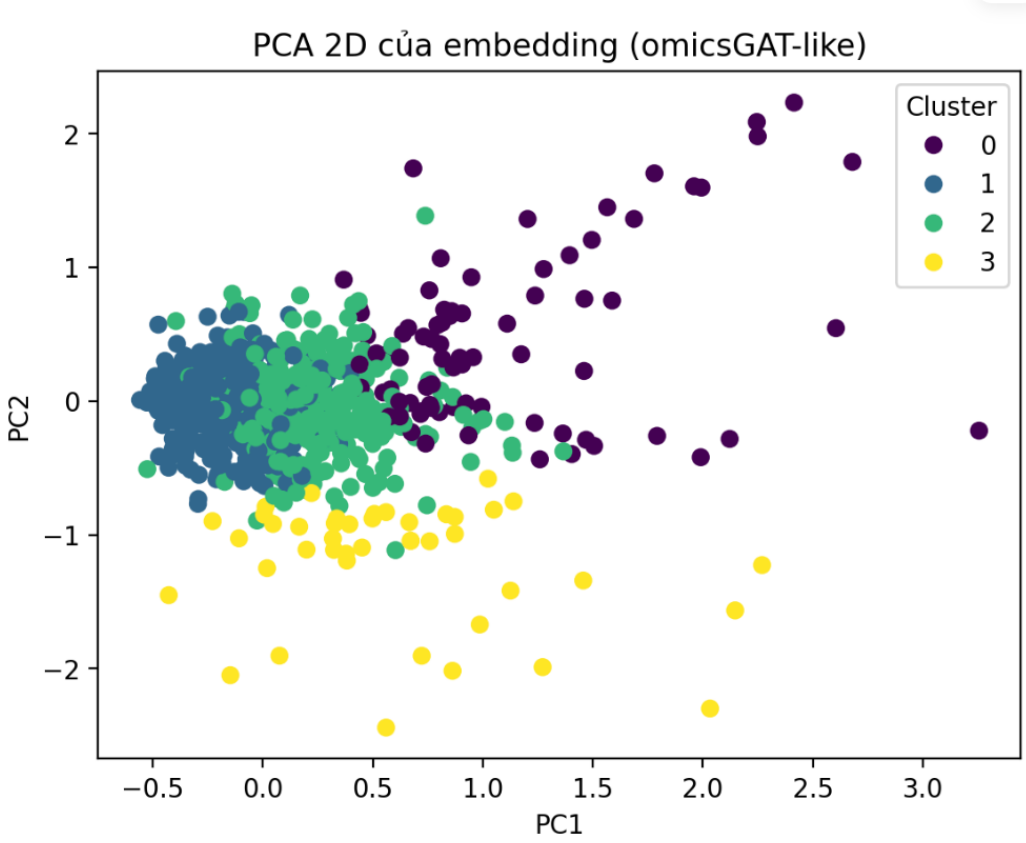
### *3.3.4 Trực quan hoá kết quả*

Sau khi mô hình hoàn thành, ứng dụng tự động hiển thị các kết quả sau:

#### *a. Biểu đồ PCA 2D*

Embedding học được từ mô hình được giảm chiều bằng PCA, sau đó vẽ thành scatter plot:

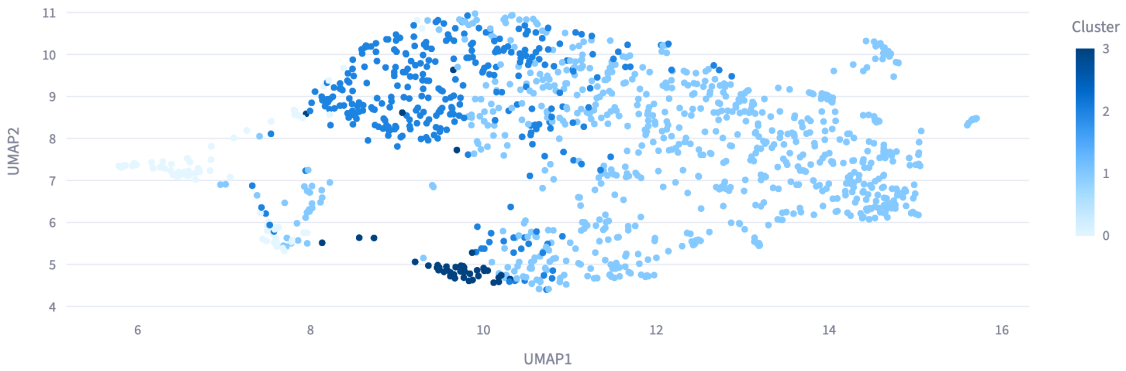
- Mỗi điểm là một bệnh nhân.
- Màu sắc biểu thị cụm mà mô hình gán.
- Biểu đồ cho phép quan sát mức độ tách biệt giữa các nhóm.




**Hình 3.11: Biểu đồ PCA**

#### *b. Biểu đồ UMAP*

Nếu người dùng bật tính năng UMAP, ứng dụng sẽ tính embedding phi tuyến 2D và hiển thị bản đồ UMAP – trực quan hơn với dữ liệu có cấu trúc phức tạp.




**Hình 3.12: Biểu đồ UMAP**

### *c. Bảng gán nhãn*

Một bảng dữ liệu liệt kê:

- ID mỗi bệnh nhân,
- Nhãn cụm mà mô hình dự đoán.

Người dùng có thể đối chiếu hoặc so sánh với các nhóm phân nhóm lâm sàng có sẵn (ví dụ: PAM50).

**Bảng 3.2: Cấu trúc của bảng gán nhãn**

| | Sample | Cluster |
|---|---|---|
| 0 | TCGA-AR-A5QQ-01 | 1 |
| 1 | TCGA-D8-A1JA-01 | 1 |
| 2 | TCGA-BH-A0BQ-01 | 2 |
| 3 | TCGA-BH-A0BT-01 | 2 |
| 4 | TCGA-A8-A06X-01 | 1 |
| 5 | TCGA-A8-A096-01 | 2 |
| 6 | TCGA-BH-A0C7-01 | 1 |
| 7 | TCGA-AC-A5XU-01 | 1 |
| 8 | TCGA-PE-A5DE-01 | 1 |
| 9 | TCGA-PE-A5DC-01 | 1 |

| | | |
|---|---|---|
| … | … | … |

Ứng dụng cho phép tải về kết quả phân cụm dưới dạng tệp CSV, bao gồm: Cột Sample và cột Cluster.

## 3.4 Phân tích, đánh giá kết quả thử nghiệm

Để đánh giá hiệu quả của mô hình omicsGAT trong việc phân cụm các phân nhóm ung thư vú, chúng tôi tiến hành so sánh với ba phương pháp cơ sở (baselines) phổ biến trong lĩnh vực phân tích dữ liệu omics:

- K-Means: Phương pháp phân cụm truyền thống, áp dụng trực tiếp trên dữ liệu biểu hiện gene đã chuẩn hóa, không qua bước học biểu diễn (representation learning).
- VAE (Variational Autoencoder): Mô hình học sâu sinh (generative deep learning) sử dụng kiến trúc encoder-decoder để học biểu diễn tiềm ẩn (latent representation) của dữ liệu.
- GCN (Graph Convolutional Network): Mạng nơ-ron đồ thị truyền thống sử dụng cơ chế tích chập đồ thị để học biểu diễn node, nhưng không có cơ chế attention như GAT.

### *3.4.1 Thiết lập thí nghiệm*

#### *a. Dữ liệu*

- Nguồn dữ liệu: TCGA-BRCA (The Cancer Genome Atlas - Breast Cancer)
- Số mẫu: 956 mẫu khối u vú (sau khi ghép với nhãn PAM50)
- Số đặc trưng: 5,000 genes có phương sai cao nhất
- Ground truth: Phân loại PAM50 gồm 5 phân nhóm phân tử

#### *b. Metrics đánh giá*

- ARI (Adjusted Rand Index): Đo độ tương đồng giữa kết quả clustering và ground truth, có điều chỉnh cho yếu tố ngẫu nhiên. Thang đo: -1 đến 1 (1 = hoàn hảo, 0 = ngẫu nhiên).

- NMI (Normalized Mutual Information): Đo lượng thông tin chung giữa hai phân cụm. Thang đo: 0 đến 1 (1 = hoàn hảo).

### *c. Cấu hình thí nghiệm*

**Bảng 3.3: Cấu hình của thí nghiệm cho các mô hình**

| Tham số | Giá trị |
|---|---|
| Số cụm (n_clusters) | 5 |
| Số epochs | 300 |
| Learning rate | 0.001 |
| k-NN neighbors | 10 |
| Embedding dimension | 32 |

## *3.4.2 Kết quả*

**Bảng 3.4: Bản so sánh kết quả giữa các mô hình**

| Phương pháp | ARI | NMI |
|---|---|---|
| K-Means | 0.4245 | 0.5649 |
| VAE | 0.4172 | 0.5468 |
| GCN | 0.2753 | 0.4507 |
| omicsGAT | 0.4204 | 0.5475 |

## *3.4.3 Phân tích kết quả*

- omicsGAT đạt ARI = 0.4204 và NMI = 0.5475, kết quả tương đương với K-Means (ARI = 0.4245, NMI = 0.5649) và VAE (ARI = 0.4172, NMI = 0.5468).
- omicsGAT nén dữ liệu từ 5000 chiều xuống 32 chiều trong khi vẫn duy trì hiệu suất clustering tương đương K-Means trên dữ liệu gốc.
- Cơ chế attention cho phép mô hình tự động xác định mức độ quan trọng của các mẫu lân cận, vượt trội hơn so với GCN sử dụng trọng số cố định.
- Embeddings học được có thể tái sử dụng cho các tác vụ downstream khác như phân loại, dự đoán tiên lượng.

Quy trình thử nghiệm với dữ liệu TCGA-BRCA cho thấy:

- Mô hình omicsGAT chạy ổn định trên dữ liệu RNA-seq kích thước lớn.
- PCA và UMAP cho thấy các cụm được tách biệt tương đối rõ ràng.

- Hệ thống có khả năng mở rộng cho các bộ multi-omics khác (DNA methylation, miRNA, proteomics,…).

Ứng dụng chứng minh được tính khả thi của việc tích hợp mô hình deep learning đồ thị vào một hệ thống GUI local, phục vụ phân tích dữ liệu ung thư theo hướng trực quan và thực nghiệm.

## 3.5 Kết chương

Chương 3 đã trình bày toàn bộ quy trình triển khai thử nghiệm mô hình omicsGAT trên dữ liệu thực từ TCGA, gồm thu thập, tiền xử lý dữ liệu, xây dựng đồ thị mẫu–mẫu, huấn luyện mô hình và trực quan hóa kết quả. Quy trình thu thập năm loại ung thư phổ biến cùng bốn loại dữ liệu omics đã cho thấy khả năng mở rộng và khả năng tích hợp đa nguồn dữ liệu của hệ thống. Các bước tiền xử lý – như loại bỏ gene nhiễu, chuẩn hóa biểu hiện, cắt ngưỡng CNV và đồng bộ nhiều bộ dữ liệu – đảm bảo đầu vào ổn định, giảm nhiễu và phù hợp với kiến trúc mô hình.

Việc xây dựng đồ thị k-NN giữa các mẫu bệnh nhân là bước quan trọng giúp tạo cơ sở cho GAT học được cấu trúc quan hệ sinh học. Mô hình omicsGAT sau đó được huấn luyện để sinh ra embedding có chất lượng cao, thể hiện rõ sự phân tách giữa các mẫu khi trực quan hóa bằng PCA hoặc UMAP. Kết quả phân cụm từ embedding cho thấy các nhóm bệnh nhân được mô hình xác định có sự khác biệt tương đối rõ trong không gian giảm chiều, chứng minh tính hiệu quả của cơ chế attention trong việc học những mối quan hệ quan trọng giữa các mẫu.

Chương này cũng đưa ra việc triển khai một ứng dụng trực quan dựa trên Streamlit, giúp người dùng dễ dàng tải dữ liệu, điều chỉnh tham số mô hình, theo dõi quá trình huấn luyện và quan sát kết quả phân nhóm. Đây là bước tiến lớn nhằm kéo mô hình từ môi trường nghiên cứu sang công cụ hỗ trợ phân tích thực tế.

# KẾT QUẢ VÀ HƯỚNG PHÁT TRIỂN TIẾP

## 1. Kết luận

Trong đề án này, nghiên cứu đã tập trung xây dựng và đánh giá mô hình học biểu diễn dựa trên mạng nơ-ron đồ thị – omicsGAT – nhằm khai thác hiệu quả dữ liệu đa omics cho bài toán phân nhóm và phân tích ung thư. Kết quả chỉ ra rõ việc mô hình hóa dữ liệu ở dạng đồ thị, kết hợp cơ chế Attention, giúp cải thiện rõ rệt khả năng trích xuất đặc trưng và biểu diễn mối quan hệ sinh học phức tạp giữa các bệnh nhân.

Nghiên cứu đã hệ thống hóa những lý thuyết về dữ liệu omics, multi-omics, cùng các phương pháp học biểu diễn hiện đại, từ đó chỉ ra những hạn chế của các phương pháp truyền thống khi xử lý dữ liệu phân tử có chiều cao và cấu trúc phi tuyến. Cách tiếp cận dựa trên GNN, đặc biệt GAT, được chứng minh là phù hợp hơn nhờ khả năng học trọng số động giữa các mẫu và giảm thiểu nhiễu trong dữ liệu sinh học.

Mô hình omicsGAT được triển khai thử nghiệm trên dữ liệu TCGA với nhiều bước tiền xử lý chặt chẽ, bao gồm lọc gene, chuẩn hóa đa omics, xây dựng đồ thị mẫu–mẫu và huấn luyện encoder dựa trên multi-head attention. Các embedding thu được cho thấy tính phân tách tốt hơn so với PCA truyền thống, thể hiện qua các biểu đồ PCA/UMAP và khả năng phân cụm rõ ràng giữa các nhóm bệnh nhân. Điều này khẳng định rằng omicsGAT không chỉ học được đặc trưng cục bộ từ từng bệnh nhân mà còn nắm bắt được quan hệ sinh học có ý nghĩa ở cấp quần thể.

Nghiên cứu đã xây dựng một ứng dụng trực quan dựa trên Streamlit, cho phép người dùng tải dữ liệu, thiết lập tham số mô hình, quan sát trực quan embedding và kết quả phân cụm. Ứng dụng đã phần nào cho thấy tính khả thi của việc tích hợp mô hình deep learning đồ thị vào các công cụ hỗ trợ bác sĩ, nhà nghiên cứu và phòng xét nghiệm trong quá trình phân tích dữ liệu ung thư.

Nhìn chung, kết quả cho thấy mô hình omicsGAT có tiềm năng lớn trong việc phân nhóm ung thư dựa trên dữ liệu đa omics, giúp cải thiện chất lượng phân tích, hỗ trợ khám phá các phân nhóm tiềm ẩn và mang lại những gợi ý quan trọng cho nghiên cứu y học chính xác. Tuy nhiên, mô hình còn những điều cần cải thiện khi triển khai trên các bộ dữ liệu lớn đa tầng, và yêu cầu tài nguyên tính toán đáng kể. Do đó, hướng nghiên cứu sắp tới dồn vào tối ưu hóa kiến trúc mô hình, mở rộng sang đồ thị dị thể multi-omics đầy đủ, đồng thời kiểm chứng trên các dữ liệu lâm sàng thực tế nhằm nâng cao giá trị ứng dụng.

## 2. Hướng phát triển tiếp

Hiện tại mô hình chủ yếu sử dụng RNA-seq; các hướng mở rộng gồm:

- Tích hợp thêm DNA methylation, copy number variation (CNV), proteomics hoặc metabolomics.
- Thiết kế multi-branch omicsGAT, trong đó mỗi loại omics có một encoder riêng (GAT, GNN, MLP…), sau đó hợp nhất bằng attention hoặc contrastive learning.

# DANH MỤC TÀI LIỆU THAM KHẢO


[1] F. Cunningham, etal., *Ensembl 2022.* Journal Nucleic Acids Research, 2022.

[2] Little, R.J.A. & Rubin, D.B., *Statistical Analysis with Missing Data, 3rd ed. , John Wiley & Sons, Hoboken,* 2019.

[3] M. Love, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2., Journal Genome Biology, 15(12),* 2014.

[4] https://www.genome.gov/Funded-Programs-Projects/Cancer-Genome-Atlas

[5] B. Bolstad, *A comparison of normalization methods for high density oligonucleotide array data., Journal Bioinformatics 19(2),* 2003, pp.185-193.

[6] M. Bibikova, *High density DNA methylation array with single CpG site resolution., Journal Genomics, Vol. 98, Issue 4, October* 2011.

[7] M. Stone, *Cross-validatory choice and assessment of statistical predictions.,* 1974. Journal of the Royal Statistical Society: Series B (Methodological)Volume 36, Issue 2 pp. 111-133

[8] Kaplan, E.L. & Meier, P., *Nonparametric estimation from incomplete observations. ,* 1958. Journal of Science and Education, 53, pp.471-481

[9] A. Levine, *800 million years of evolution and 40 years of discovery.,* 2020. Journal Nature Reviews Cancer 20, 13 May 2020, pp. 471-480

[10] K. Kuchenbaecker, *Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers.,* 2017. JAMA The Journal of the American Medical Association 317(23):2402

[11] T. Lynch, *Activating mutations in the epidermal growth factor receptor underlying responsive- ness of non-small-cell lung cancer to gefitinib.,* 2004. Journal of Medicince, 20 May 2004, Vol 350, No. 21.

[12] F. Harrell, *Evaluating the yield of medical tests,* 1982. Journal of the American… 14 May 1982

[13] N. Mantel, *Evaluation of survival data and two new rank order statistics arising in its consideration.,* 1966. Cancer chemotherapy reports 27 February 1966.

[14] Guo, M., Ye, X., Huang, D, *SubtypeCtAE: Contractive autoencoder for cancer subtyping based on multi-omics data.,* 2025.

[15] Tabakhi,S.,Vandermeulen,C.,Sudbery,I.,Lu,H, *Heterogeneous graph attention network improves cancer multiomics integration..* arXiv preprint arXiv:2408.02845. 2024 Aug 5.

[16] T. Wang, *MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. ,* 2021. Nature Communications, Vol 12 (1):3445, June 2021.

[17] Wang, B., Mezlini, A.M., Demir, F., *Similarity network fusion for aggregating data types on a genomic scale.,* 2014. Nat Methods 11, 333–337 (2014). https://doi.org/10.1038/nmeth.2810.

[18] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y., *Graph Attention Networks*, International Conference on Learning Representations (ICLR). 2018

# 14% Tính tương đồng nói chung

Tổng cộng của tất cả các kết quả trùng khớp, bao gồm cả các nguồn trùng lặp, cho mỗi c...

## Đã lọc khỏi Báo cáo

- Mục lục tham khảo
- Văn bản được trích dẫn
- Văn bản được trích dẫn
- Kết quả trùng khớp nhỏ (ít hơn 8 từ)

## Nguồn hàng đầu

10% Nguồn Internet
6% Ấn bản
6% Bài tập được nộp (bài của học sinh)

PGS.TSKH. Hoàng Đăng Hải

Nguyễn Anh Tuấn

BỘ KHOA HỌC VÀ CÔNG NGHỆ
**HỌC VIỆN CÔNG NGHỆ**
**BƯU CHÍNH VIỄN THÔNG**

**CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM**
**Độc lập - Tự do - Hạnh phúc**

# BÁO CÁO GIẢI TRÌNH
# SỬA CHỮA, HOÀN THIỆN ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Họ và tên học viên: Nguyễn Anh Tuấn

Chuyên ngành: Hệ thống thông tin

Khóa: 2024 đợt 1

Tên đề tài: Nghiên cứu xây dựng mô hình học biểu diễn sử dụng dữ liệu đa omics cho phân nhóm ung thư

Người hướng dẫn khoa học: PGS.TSKH. Hoàng Đăng Hải

Ngày bảo vệ: 24/01/2026

Các nội dung học viên đã sửa chữa, bổ sung trong đề án tốt nghiệp theo ý kiến đóng góp của Hội đồng chấm đề án tốt nghiệp:

| TT | Ý kiến hội đồng | Sửa chữa của học viên |
|---|---|---|
| 1 | Bảng so sánh các giá trị NMI, ARI của mô hình omicsGAT với các phương pháp cơ sở như K-means trên dữ liệu thô, VAE hoặc GCN truyền thống | Đã chạy thử nghiệm với phương pháp khác và có số liệu so sánh ở mục 3.4 |
| 2 | Giải thích rõ mô hình omicsGAT | Đã bổ sung mục 2.1 để giải thích rõ về mô hình omicsGAT |
| 3 | Làm rõ 2 bài toán khác biệt là phân cụm và phân nhóm ung thư (phân lớp) và cách giải quyết sử dụng mô hình omicsGAT. | Đã giải thích thêm trong mục 2.4.1 và 2.4.2 |

*Hà Nội, ngày 9 tháng 2 năm 2026*

**Ký xác nhận của**

| CHỦ TỊCH HỘI ĐỒNG CHẤM ĐỀ ÁN | THƯ KÝ HỘI ĐỒNG | NGƯỜI HƯỚNG DẪN KHOA HỌC | HỌC VIÊN |
|---|---|---|---|
| TS. Nguyễn Duy Phương | TS. Đỗ Thị Liên | PGS.TSKH. Hoàng Đăng Hải | Nguyễn Anh Tuấn |

BỘ KHOA HỌC VÀ CÔNG NGHỆ
**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM**
**Độc lập - Tự do - Hạnh phúc**

# BIÊN BẢN
# HỌP HỘI ĐỒNG CHẤM ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Căn cứ quyết định số Quyết định số 3299/QĐ-HV ngày 26 tháng 12 năm 2025 của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ. Hội đồng đã họp vào hồi ..11.... giờ ..15... phút, ngày 24 tháng 01 năm 2025 tại Học viện Công nghệ Bưu chính Viễn thông để chấm đề án tốt nghiệp thạc sĩ cho:

Học viên: **Nguyễn Anh Tuấn**

Tên đề án tốt nghiệp: **Nghiên cứu xây dựng mô hình học biểu diễn sử dụng dữ liệu đa omics cho phân nhóm ung thư**

Ngành: **Hệ thống thông tin** Mã số: **8.48.01.04**

Các thành viên của Hội đồng chấm đề án tốt nghiệp có mặt: 5...../ 05

| TT | HỌ VÀ TÊN | TRÁCH NHIỆM TRONG HĐ | GHI CHÚ |
|---|---|---|---|
| 1 | **TS. Nguyễn Duy Phương** | Chủ tịch | |
| 2 | **TS. Đỗ Thị Liên** | Thư ký | |
| 3 | **PGS.TS. Nguyễn Long Giang** | Phản biện 1 | |
| 4 | **PGS.TS. Nguyễn Hữu Quỳnh** | Phản biện 2 | |
| 5 | **PGS.TS. Đỗ Trung Tuấn** | Uỷ viên | |

**Các nội dung thực hiện:**

1. Chủ tịch Hội đồng điều khiển buổi họp. Công bố quyết định của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ.
2. Người hướng dẫn khoa học hoặc thư ký đọc lý lịch khoa học và các điều kiện bảo vệ đề án tốt nghiệp của học viên (Có bản lý lịch khoa học và kết quả các môn học cao học của học viên kèm theo).
3. Học viên trình bày tóm tắt đề án tốt nghiệp.
4. Phản biện 1 đọc nhận xét (có văn bản kèm theo)
5. Phản biện 2 đọc nhận xét (có văn bản kèm theo)
6. Các câu hỏi của thành viên Hội đồng:

làm rõ lý do lựa chọn bộ dữ liệu liên quan
Mô hình học biểu diễn và học máy nói chung khác nhau gì?
làm rõ biểu diễn trên đồ thị: nút là gì, cạnh là gì?

7. Trả lời của học viên:

Học viên tiếp thu và chỉnh sửa

8. Thư ký đọc nhận xét về quá trình thực hiện đề án tốt nghiệp của học viên *(có văn bản kèm theo).*

9. Hội đồng họp riêng:

- Bầu Ban kiểm phiếu:
  1. Trưởng Ban kiểm phiếu: Đỗ Trung Tuấn
  2. Uỷ viên Ban kiểm phiếu: Đỗ Thị Liên
  3. Uỷ viên Ban kiểm phiếu: Nguyễn Long Giang
- Hội đồng chấm đề án tốt nghiệp bằng bỏ phiếu kín.
- Ban kiểm phiếu làm việc:
- Trưởng Ban kiểm phiếu báo cáo kết quả kiểm phiếu *(có Biên bản họp Ban kiểm phiếu kèm theo)*
- **Điểm trung bình của đề án tốt nghiệp:** 7.5

**Kết luận:**

1. Các nội dung cần chỉnh sửa, hoàn thiện sau bảo vệ đề án tốt nghiệp:

- làm rõ đóng góp của đề án, giải thích rõ các Graph
- bổ sung và phân tích rõ hết quả

2. Đề nghị Học viện công nhận (hoặc không) và cấp bằng (hoặc không) thạc sĩ cho học viên: Nguyễn Anh Tuấn

3. Đề án tốt nghiệp có thể phát triển thành đề tài nghiên cứu cho NCS: Không

Buổi làm việc kết thúc vào 12h cùng ngày.

| **Chủ tịch** | **Thư ký** |
|---|---|
| **TS. Nguyễn Duy Phương** | **TS. Đỗ Thị Liên** |

**CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**

**Độc lập – Tự do – Hạnh phúc**

# BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

*(Dùng cho người phản biện)*

Tên đề tài đề án tốt nghiệp: NGHIÊN CỨU XÂY DỰNG MÔ HÌNH HỌC BIỂU DIỄN SỬ DỤNG DỮ LIỆU ĐA OMICS CHO PHÂN NHÓM UNG THƯ

Chuyên ngành: HỆ THỐNG THÔNG TIN

Mã chuyên ngành: 8.48.01.04

Họ và tên học viên: Nguyễn Anh Tuấn

Họ và tên người nhận xét: Nguyễn Hữu Quỳnh

Học hàm, học vị: PGS.TS

Chuyên ngành: KHMT

Cơ quan công tác: Trường Đại học CMC

Số điện thoại: 0912174793.....E-mail: nhquynhml@gmail.com.....

## NỘI DUNG NHẬN XÉT

I/ Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài:

Việc khai thác các nguồn dữ liệu công khai như TCGA, kết hợp các kỹ thuật học sâu hiện đại, đặc biệt là mạng nơ-ron đồ thị và cơ chế attention, phù hợp với xu hướng nghiên cứu hiện nay trong tin sinh học và y học chính xác.

II/ Nội dung của đề án tốt nghiệp, các kết quả đã đạt được:

-Học viên đã hệ thống hóa được các khái niệm về dữ liệu đa omics và các phương pháp học biểu diễn hiện đại (có giám sát, không giám sát và dựa trên đồ thị).

- Đã xây dựng được mô hình omicsGAT với cấu trúc đa nhiệm, sử dụng Encoder để trích xuất Embedding và các nhánh Decoder/Classifier để thực hiện phân nhóm.

- Đề án đã triển khai thử nghiệm trên dữ liệu TCGA và xây dựng được ứng dụng hỗ trợ trực quan hóa kết quả thông qua các biểu đồ PCA và UMAP (Bước đầu cho thấy mô hình có khả năng phân tách nhóm bệnh nhân ở mức độ trực quan).

III/ Những vấn đề cần giải thích thêm:

- Trong đề án, học viên trình bày học biểu diễn như một khối tách biệt. Cần làm rõ hơn rằng thực chất đây chính là giai đoạn trích xuất đặc trưng tự động (feature learning) – thành phần cốt lõi của kiến trúc học sâu để tạo ra các vector embedding trung gian. Việc nhấn mạnh vai trò của học biểu diễn như một bước xử lý thông tin giàu ý nghĩa sinh học hơn dữ liệu thô sẽ làm tăng giá trị khoa học của đề tài.

- Mặc dù trong Chương 2 có liệt kê các độ đo như NMI, ARI, silhouette score và phân tích sống sót (C-index), nhưng trong Chương 3 (Triển khai thử nghiệm), các bảng số liệu kết quả định lượng cụ thể vẫn chưa thấy xuất hiện. Danh mục bảng chỉ có cấu trúc dữ liệu và bảng gán nhãn mẫu. Cần bổ sung:

+Bảng so sánh các giá trị NMI, ARI của mô hình omicsGAT với các phương pháp cơ sở (baselines) như K-means trên dữ liệu thô, VAE hoặc GCN truyền thống.

+Kết quả phân tích sống sót (Kaplan-Meier, p-value) để chứng minh ý nghĩa lâm sàng của các nhóm đã phân tách.

IV/ Kết luận:

Đề án có cấu trúc hợp lý, trình bày rõ ràng và có tính ứng dụng cao. Tuy nhiên, để đảm bảo tính thuyết phục về mặt khoa học, học viên cần làm rõ mối quan hệ giữa học biểu diễn với học sâu và đặc biệt là phải bổ sung các bảng dữ liệu thực nghiệm định lượng theo các độ đo đã nêu.

Đồng ý cho phép học viên bảo vệ đề án tốt nghiệp.

Câu hỏi 1:
Trong đề án, học viên nhấn mạnh vào việc xây dựng 'Mô hình học biểu diễn' (Representation Learning). Dưới góc độ kiến trúc của các mạng nơ-ron sâu (Deep Learning), em hãy giải thích bản chất của khâu 'học biểu diễn' này thực chất là gì? Tại sao đối với dữ liệu đa Omics có số chiều rất lớn và độ nhiễu cao, việc sử dụng học biểu diễn thông qua mạng đồ thị (GAT) lại ưu việt hơn so với các phương pháp trích chọn đặc trưng (Feature Extraction) hoặc giảm chiều dữ liệu truyền thống như PCA hay LDA?
Câu hỏi 2: Dựa trên số liệu nào để khẳng định rằng việc kết hợp cơ chế Attention trên đồ thị giúp cải thiện độ chính xác của việc phân nhóm so với các mô hình mạng nơ-ron thông thường?

Ngày 19 tháng 1 năm 2025

NGƯỜI NHẬN XÉT



**CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**
**Độc lập - Tự do - Hạnh phúc**
==========

# BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

*(Dùng cho người phản biện)*

Tên đề án: Nghiên cứu xây dựng mô hình học biểu diễn sử dụng dữ liệu đa OMICS cho phân nhóm ung thư

Chuyên ngành: Hệ thống thông tin

Mã số: 8.48.01.01

Họ và tên học viên: Nguyễn Anh Tuấn

Họ và tên người nhận xét: PGS.TS. Nguyễn Long Giang

Cơ quan công tác: Viện Công nghệ thông tin, Viện Hàn Lâm KH&CN Việt Nam

Số điện thoại: 0904739189 Email: nlgiang@ioit.ac.vn

## NỘI DUNG NHẬN XÉT

*1. Cơ sở khoa học và thực tiễn, sự cần thiết lựa chọn đề tài*

Đề án giải quyết bài toán phân lớp/phân cụm các loại ung thư sử dụng dữ liệu đa omics bằng công cụ omicsGAT và thử nghiệm trên dữ liệu TCGA. Chủ đề nghiên cứu của đề án có ý nghĩa khoa học và thực tiễn.

*2. Nội dung của đề án, các kết quả đã đạt được*

Đề án đã đạt được kết quả chính như sau:

1) Sử dụng mô hình học biểu diễn omicsGAT trong phân lớp/phân cụm ung thư từ dữ liệu omics
2) Thử nghiệm, đánh giá mô hình omicsGAT với bài toán phân lớp/phân cụm dữ liệu từ TCGA, bao gồn tiền xử lý, chạy thử nghiệm và đánh giá kết quả.

Về cơ bản, các kết quả đạt được của đề án đáp ứng yêu cầu đối với đề án tốt nghiệp thạc sĩ ngành Hệ thống thông tin. Tuy nhiên, đề án cần chỉnh sửa, bổ sung để làm rõ các đóng góp của học viên trong việc xây dựng mô hình và ứng dụng mô hình omicsGAT.

*3. Những vấn đề học viên cần chỉnh sửa, bổ sung*

- Rà soát, chỉnh sửa đề án về trình bày, cần bổ sung trích dẫn tài liệu tham khảo cho các nội dung trình bày.
- Cần trình bày, giải thích nhiều khái niệm trong tin sinh học sử dụng trong đề án.

- Làm rõ 2 bài toán khác biệt là phân cụm và phân nhóm ung thư (phân lớp) và cách giải quyết sử dụng mô hình omicsGAT. Mô tả rõ mô hình omicsGAT từ tài liệu nào?

- Phần thử nghiệm (chương 3), đặc tả rõ ràng dữ liệu và quy trình tiền xử lý (dữ liệu sau xử lý như thế nào? Làm rõ Phương pháp huấn luyện và các tham số mô hình, kết quả đạt được về các độ đo như độ chính xác, F1…

*4. Kết luận*

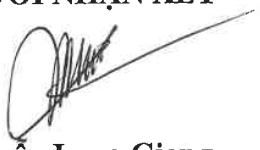
- Đồng ý cho phép học viên bảo vệ đề án tốt nghiệp.
- Đề nghị học viên chỉnh sửa/bổ sung luận văn theo góp ý.

*Câu hỏi*: 1) Làm rõ bộ dữ liệu và tại sao lại sử dụng mô hình omicsGAT cho nghiên cứu này ?

*Hà Nội, ngày 20 tháng 01 năm 2026*

**XÁC NHẬN CỦA CƠ QUAN**

**NGƯỜI NHẬN XÉT**



**Nguyễn Long Giang**