TRƯỜNG ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

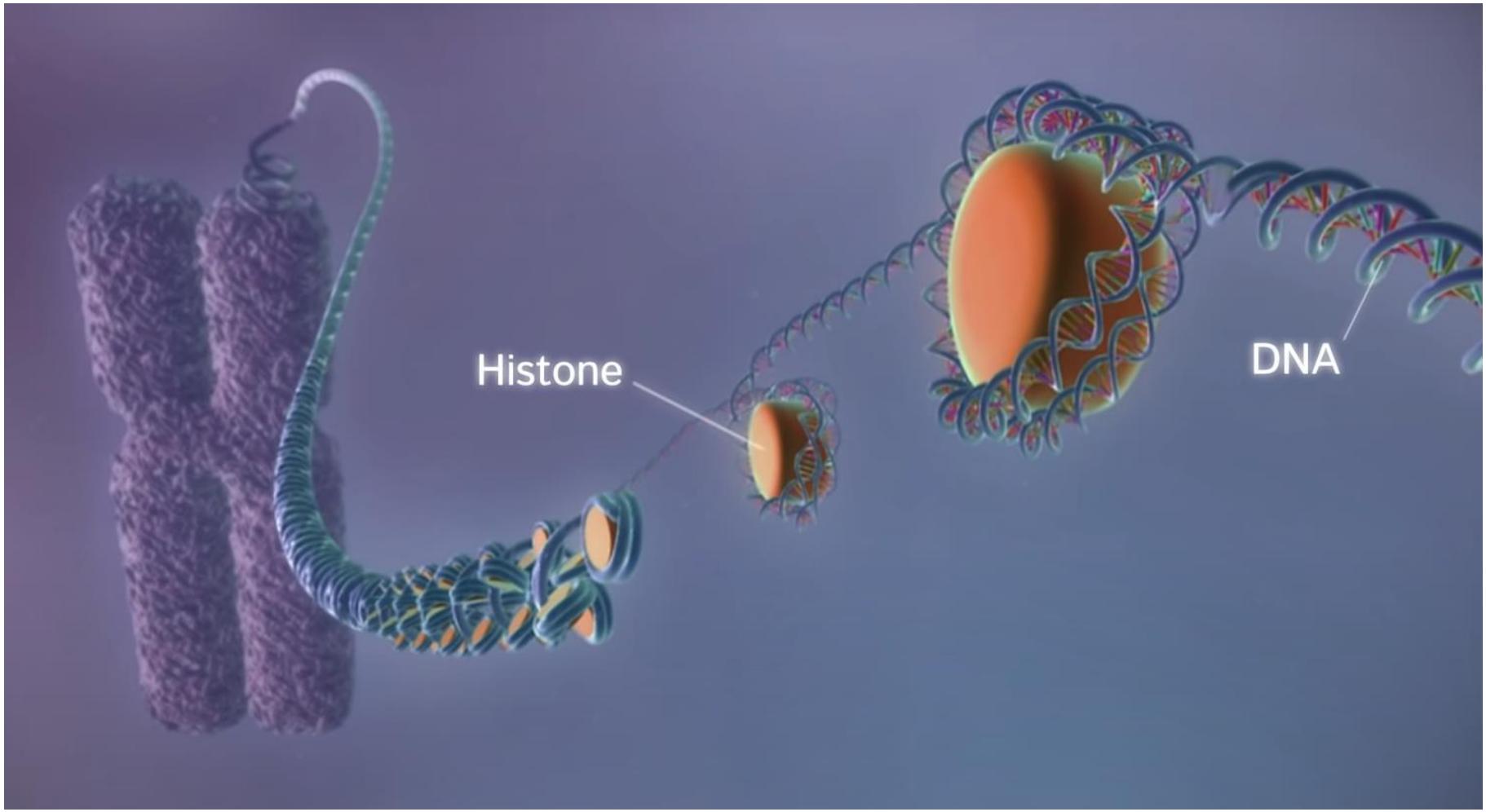
Multimodal Single-Cell Data Integration

Học viên trình bày: Trần Văn Hoàng

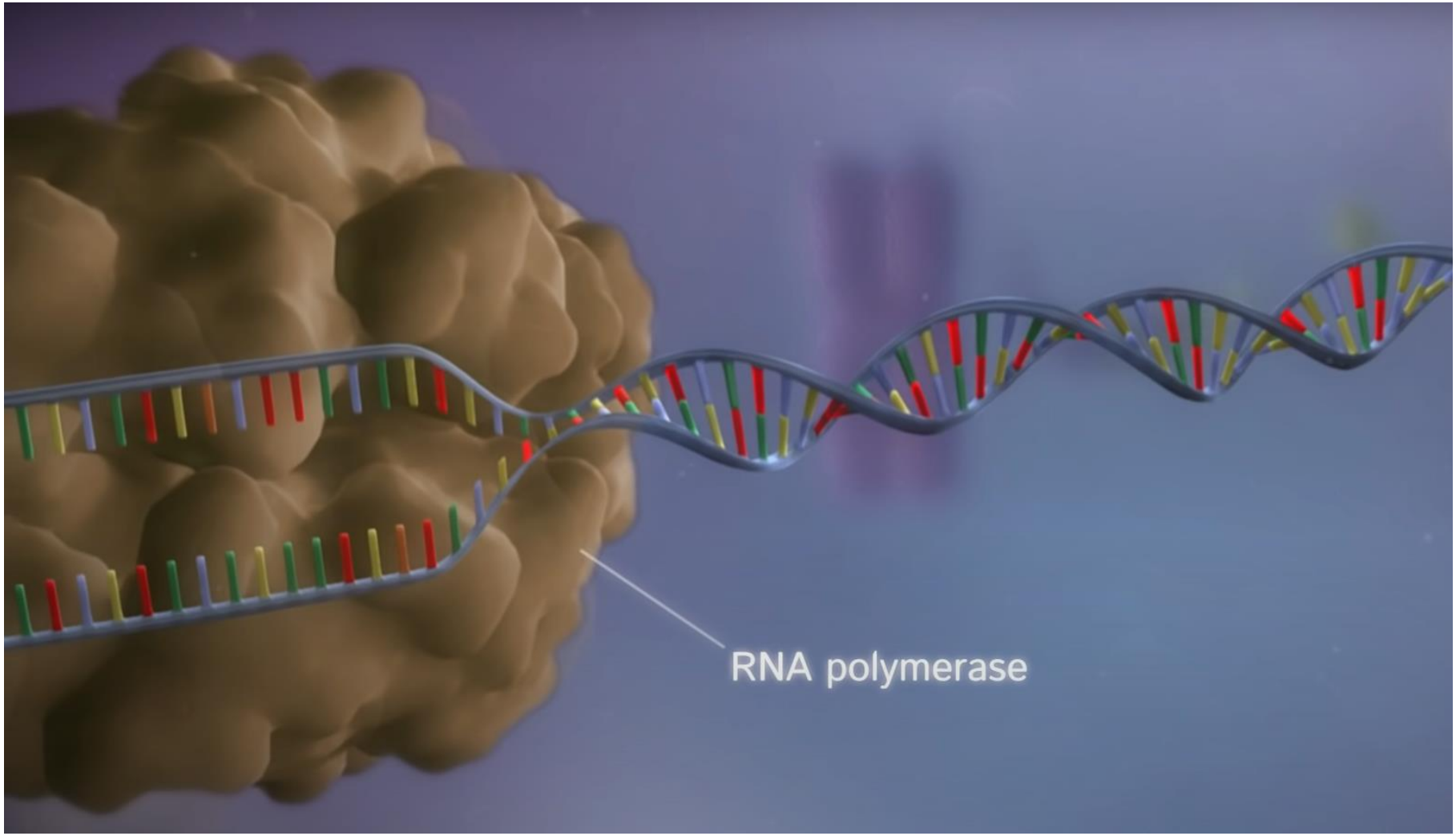
ONE LOVE. ONE FUTURE.

Table of content

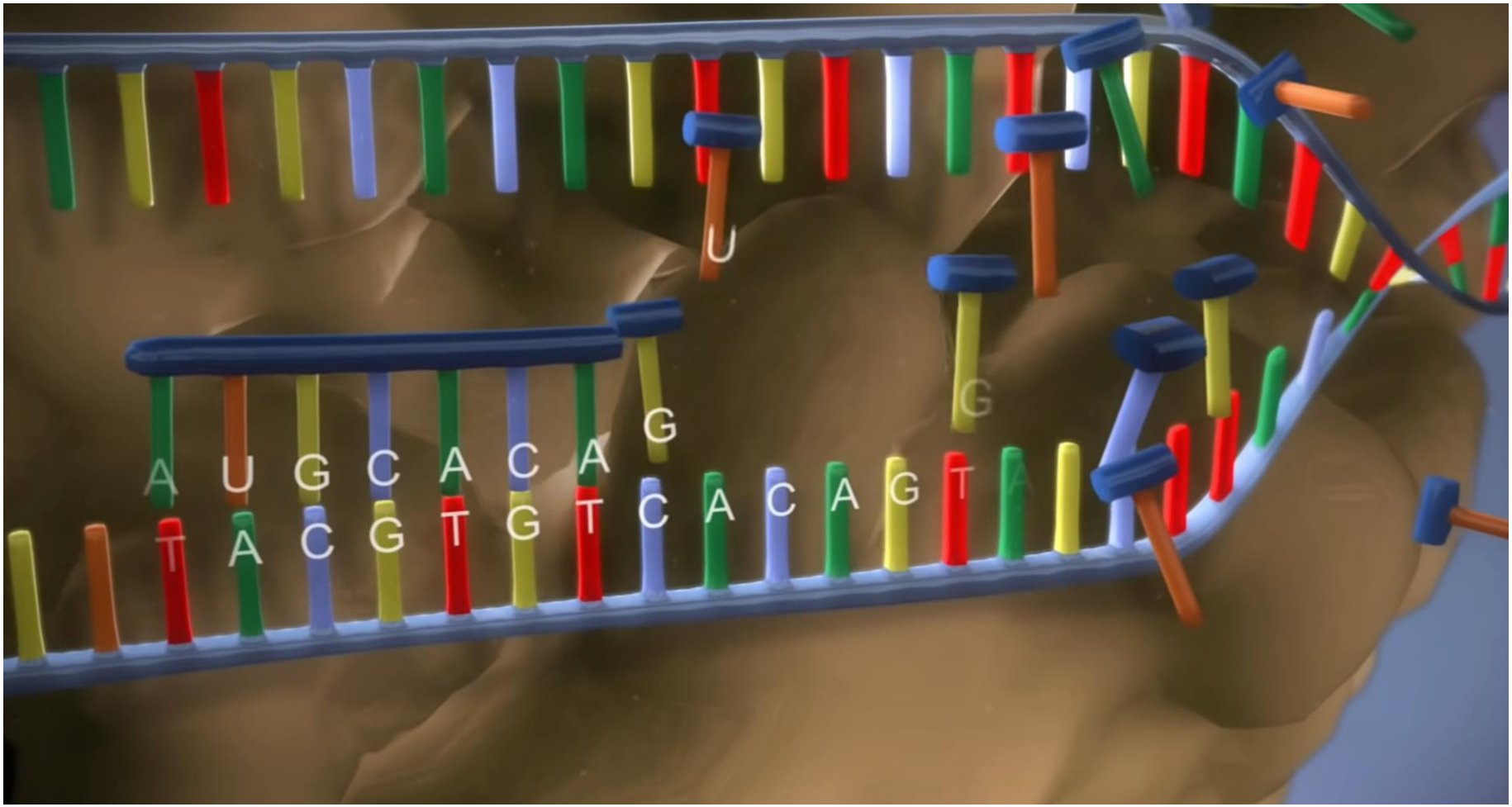
Từ nhiễm sắc thể đến protein



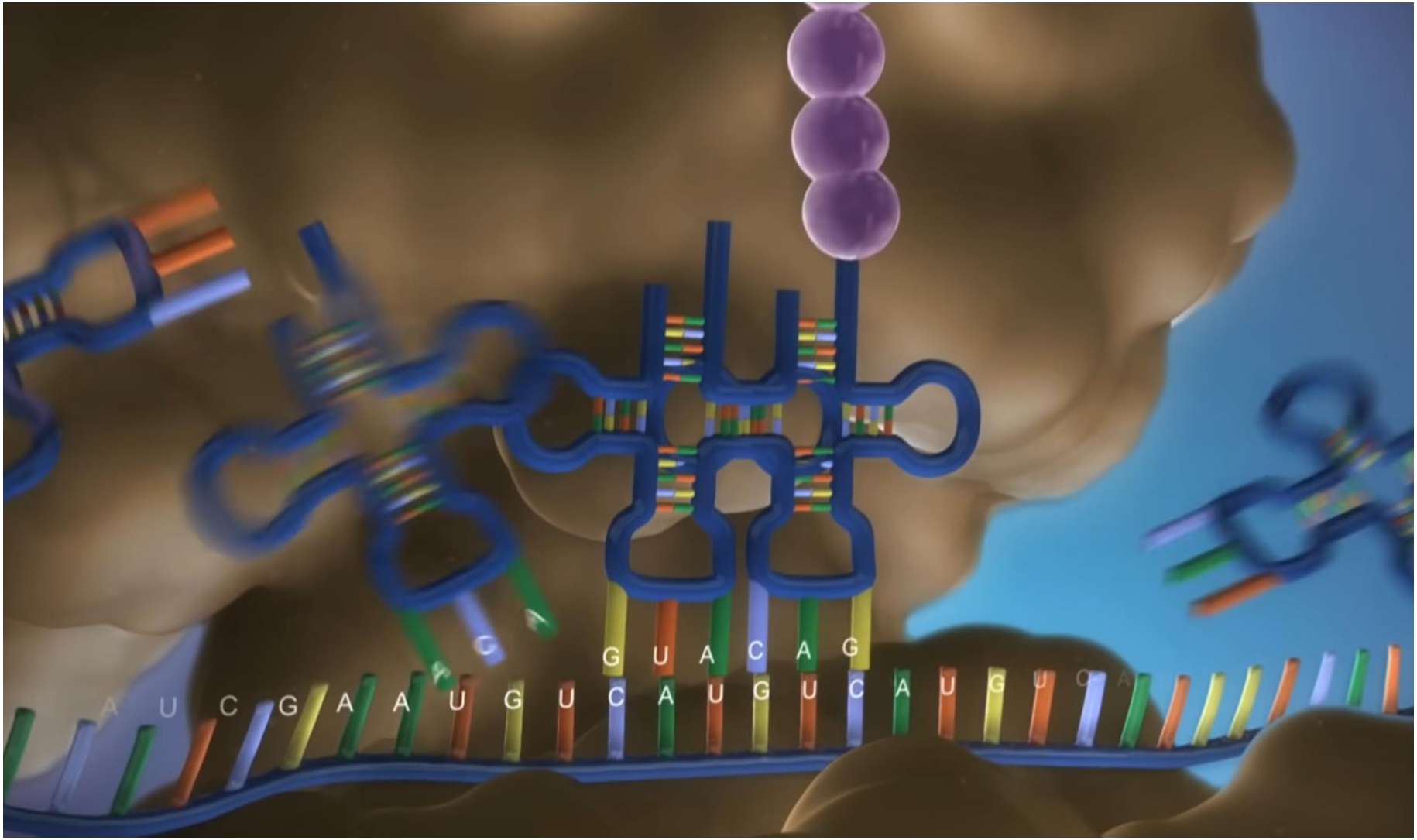
Từ nhiễm sắc thể đến protein



Từ nhiễm sắc thể đến protein



Từ nhiễm sắc thể đến protein

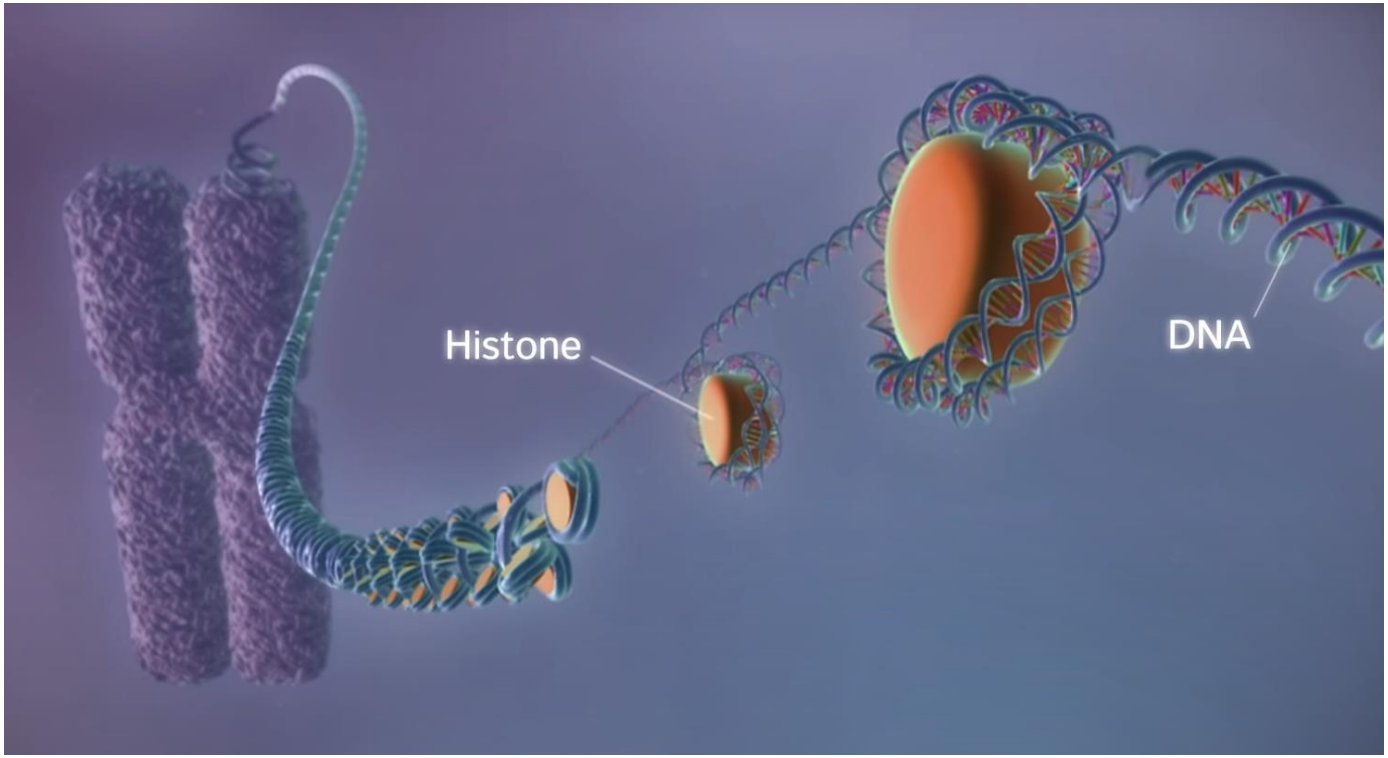


Dữ liệu từ công nghệ Single-cell

- Hiện nay các công nghệ single-cell đã có thể đo lường các loại dữ liệu khác nhau bao gồm:
  - ATAC: dữ liệu về các khu vực có thể tiếp cận của nhiễm sắc thể, các khu vực nào có thể truy cập và mức độ truy cập trên từng tế bào
  - GEX: dữ liệu biểu hiện gene, các gene nào được biểu hiện và biểu hiện bao nhiêu lần trên từng tế bào
  - ADT: dữ liệu biển hiện protein trên bề mặt của tế bào, các protein nào được biểu hiện và biểu hiện bao nhiêu lần trên từng tế bào

Dữ liệu từ công nghệ Single-cell

- Dữ liệu ATAC của một vùng thể hiện khả năng truy cập vào vùng đó.
- Khả năng truy cập này dựa vào việc vùng đó có chứa các đoạn enhancer, promoter hay insulator



Dữ liệu từ công nghệ Single-cell

- Dữ liệu ATAC được lưu dưới dạng ma trận với các hàng thể hiên các tế bào và các cột thể hiện các region. Giá trị của một ô là khả năng truy cập của region tương ứng và tế bào tương ứng

	Region 1	Region 2	Region 3
Cell 1	0	2.8	0
Cell 2	5.2	2.8	0
Cell 3	0	5.2	2.8

Dữ liệu từ công nghệ Single-cell

- Dữ liệu GEX của một gene thể hiện gene đó được biểu hiện bao nhiêu lần trong một tế bào.
- Dữ liệu GEX được lưu dưới dạng ma trận với các hàng thể hiên các tế bào và các cột thể hiện các gene. Giá trị của một ô là số lần được biểu hiện của gene tương ứng và tế bào tương ứng.

	Gene 1	Gene 2	Gene 3
Cell 1	0	100	1
Cell 2	100	200	0
Cell 3	2	300	100

Dữ liệu từ công nghệ Single-cell

- Dữ liệu ADT của một protein thể hiện protein đó được biểu hiện bao nhiêu lần trong một tế bào
- Dữ liệu ADT được lưu dưới dạng ma trận với các hàng thể hiên các tế bào và các cột thể hiện các protein. Giá trị của một ô là số lần được biểu hiện của protein tương ứng và tế bào tương ứng

	Protein 1	Protein 2	Protein 3
Cell 1	0	6	1
Cell 2	5	14	0
Cell 3	2	3	10

Table of content

- Từ nhiễm sắc thể đến protein
- Bài toán
- Các State-of-the-art hiện tại
- Mô hình đề xuất

Đặt vấn đề

- Nhờ sự tiến bộ của công nghệ mà giờ các phương pháp thí nghiệm đã có thể đo cùng lúc 2 modality như ATAC và GEX hay Gex và ADT.
- Tuy nhiên các bộ dữ liệu hiện tại vấn phổ biến là dữ liệu về các modality đơn lẻ
- Thách thức đặt ra khi chúng ta muốn tích hợp nhiều bộ dữ liệu với các modality khác nhau lại với nhau để thực hiện các nghiên cứu hoặc phân tích
- từ đó vấn đề đặt ra cần tìm ra biểu diễn chung của các dữ liệu hay mỗi quan hệ giữa các dữ liệu đó để có thể tích hợp

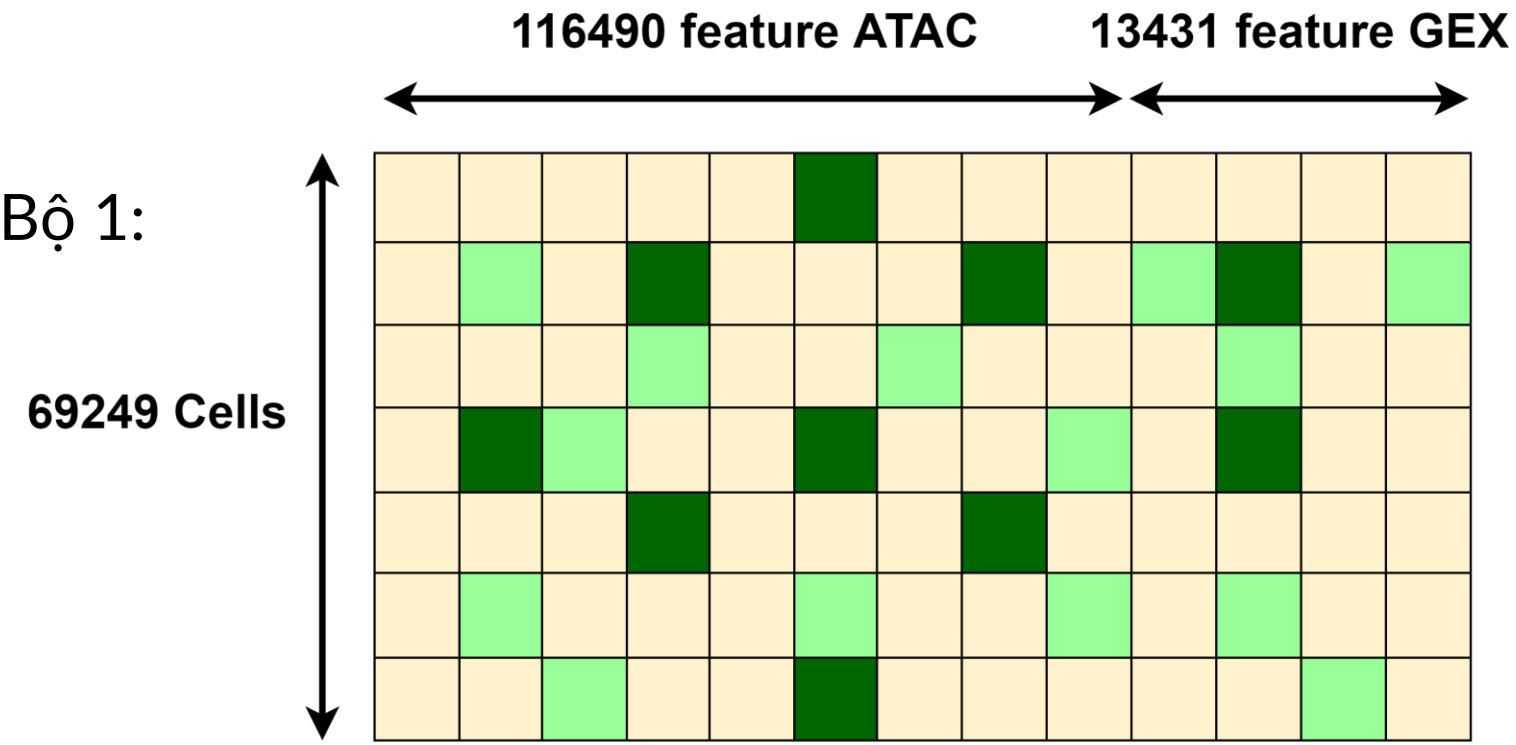
Bài toán

- Luecken và các cộng sự đã đặt ra 3 bài toán dưới dạng các challenge để giải quyết vấn đề trên
- Bài toán 1: Modality Prediction
- Bài toán 2: Modality Matching
- Bài toán 3: Joint Embedding

Bộ dữ liệu

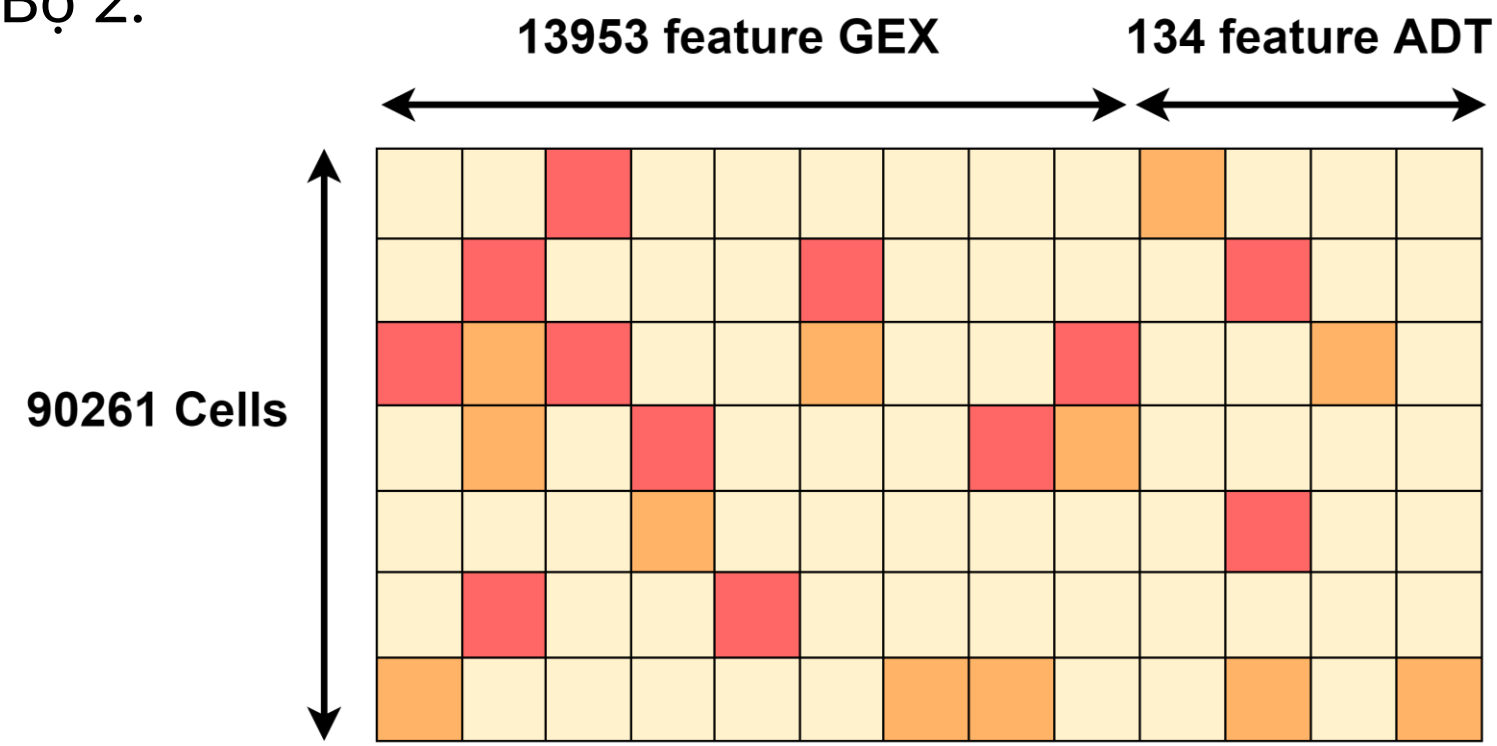
- Luecken và các cộng sự cung cấp 2 bộ dữ liệu mỗi bộ bao gồm một cặp modality của từng cell

- Bộ 1:



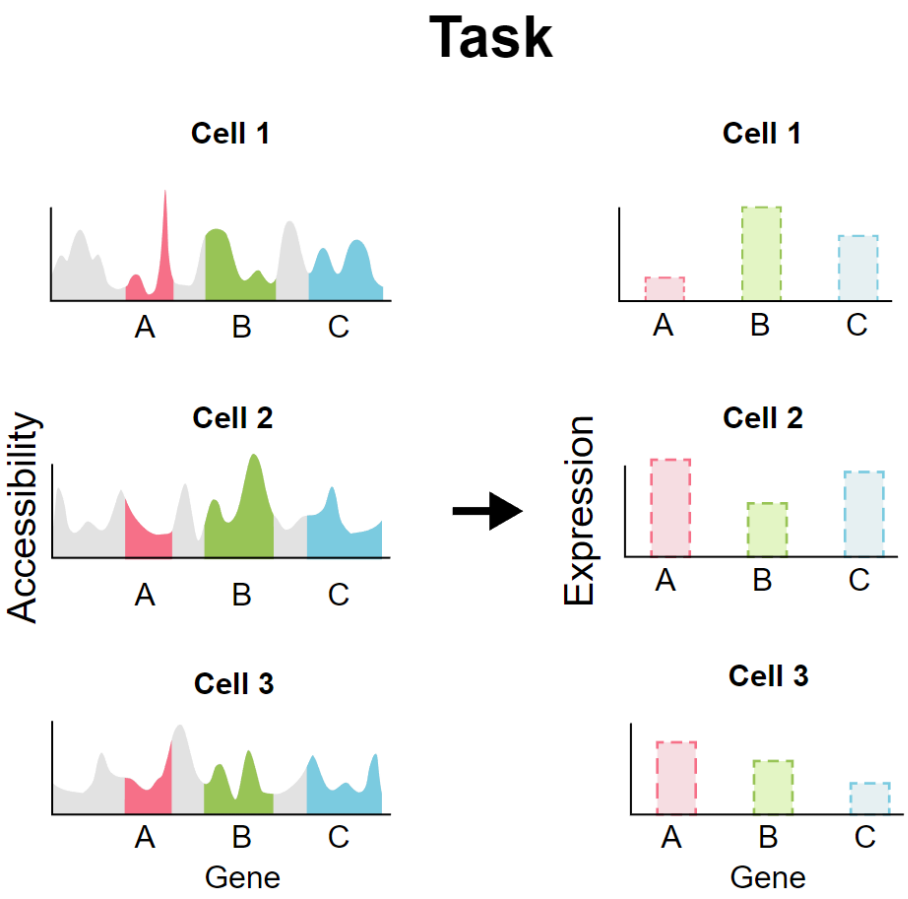
Bộ dữ liệu

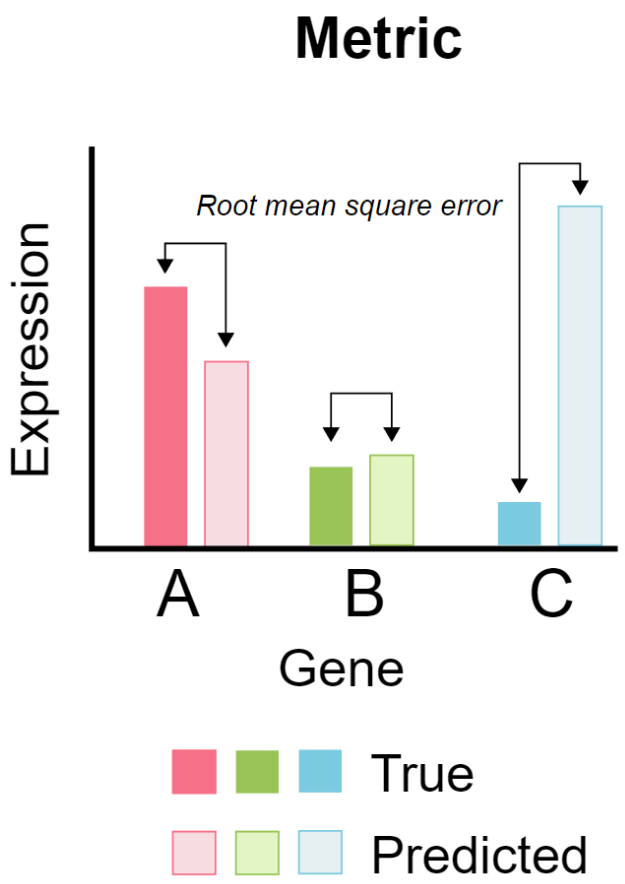
- Bộ 2:



Modality Prediction

- Trong task này chúng ta cần phải dự đoán một modality (ví dụ ATAC) từ đầu vào thành một modality khác (ví dụ GEX)



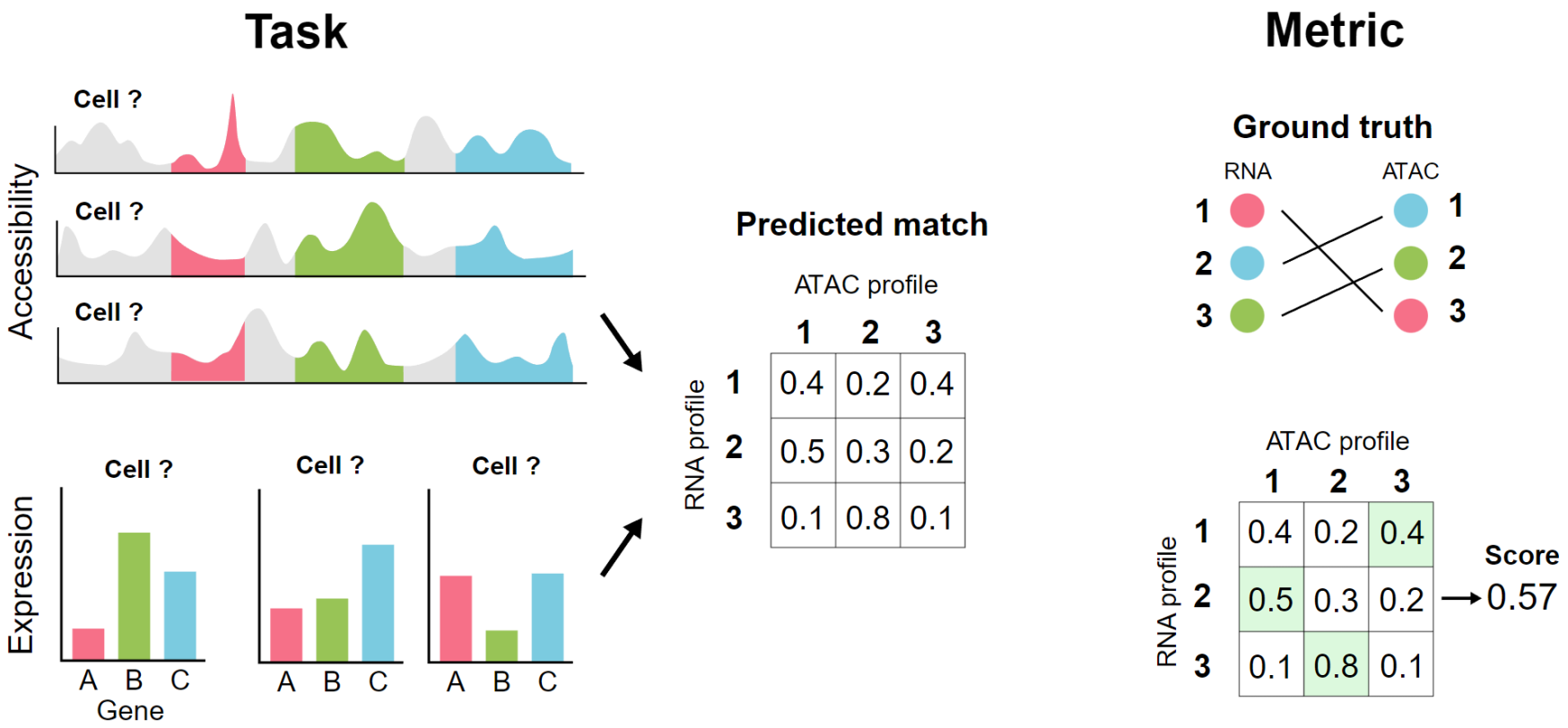


Modality Prediction

- Các cặp modality ta cần phải dự đoán bao gồm:
  - GEX thành ADT
  - ADT thành GEX
  - GEX thành ATAC
  - ATAC thành GEX

Modality Matching

- Mục tiêu của task này là xác định sự tương ứng giữa hai single cell profiles (thuộc 2 modality) và dự đoán xác suất của sự tương ứng này



Modality Matching

- Các cặp modality ta cần phải matching bao gồm:
  - GEX và ADT
  - GEX và ATAC

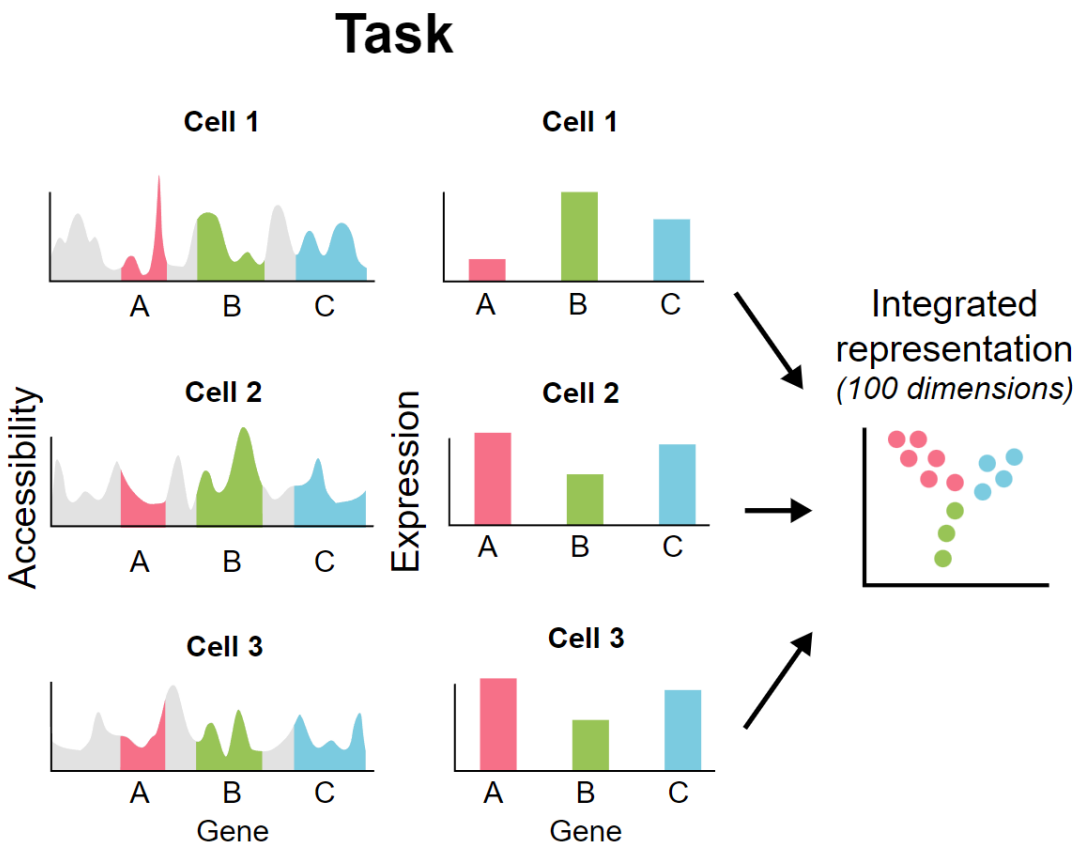
- Metric được đo bằng trung bình xác suất của các cặp dự đoán đúng, cụ thể như sau:

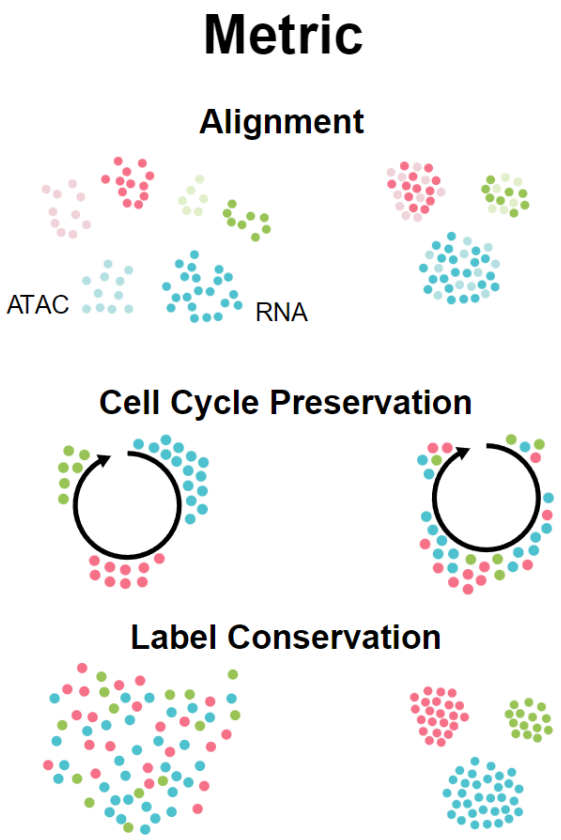
$$Score = \frac{1}{N}\sum_{i}\sum_{j} X_{i,j} \qquad if\ i = j$$

- Trong đó $X_{i,j}$ là xác suất vector *i* của modality 1 match với vector *j* của modality 2, N là số cell

Joint Embedding

- Mục tiêu của bài toán là cần xây dựng mô hình để học embedded representation cho 2 modalities





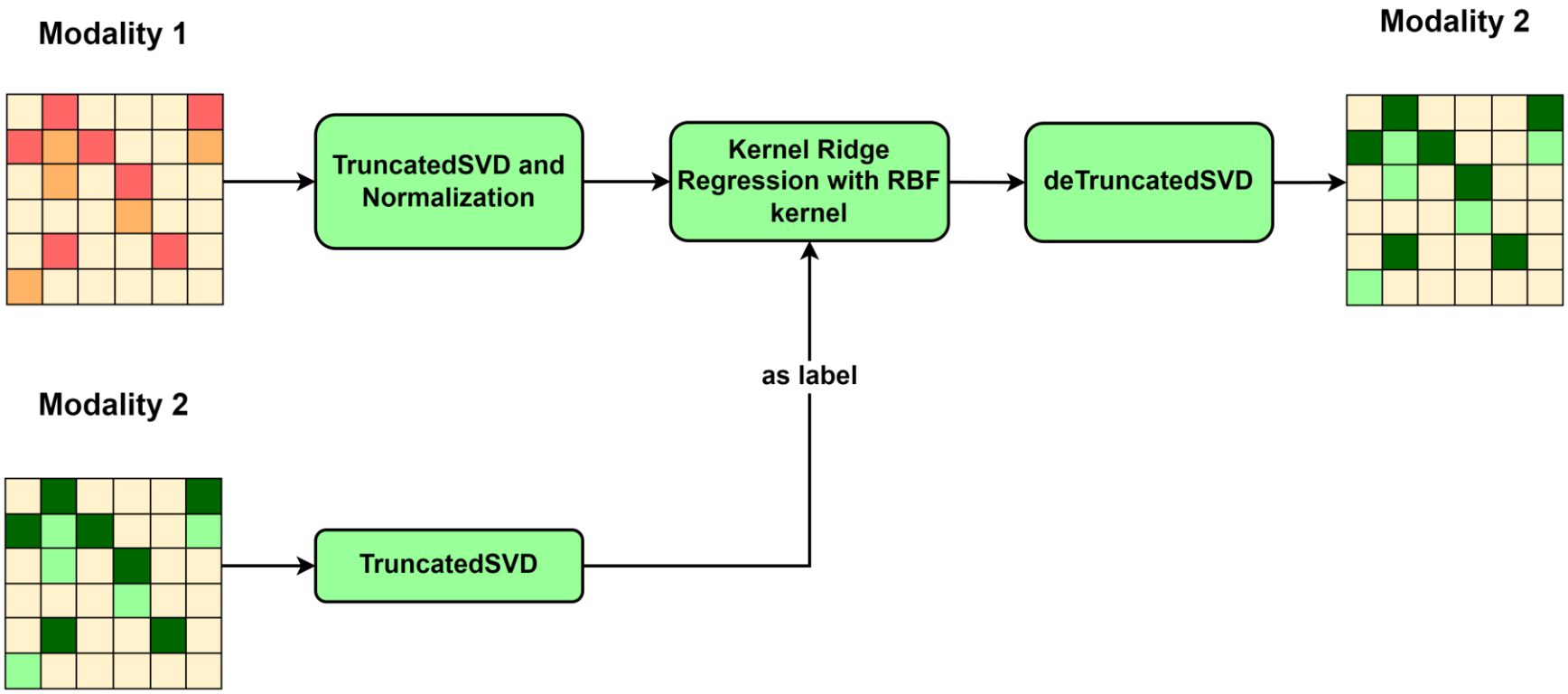
Joint Embedding

- Các cặp modality ta cần phải joint bao gồm:
  - GEX và ADT
  - GEX và ATAC
- Metric sử dụng để đánh giá bao gồm: “NMI cluster/label”, “Cell type ASW”, “Cell cycle conservation”, “Trajectory conservation”, “Batch ASW” and “Graph connectivity”.

Table of content

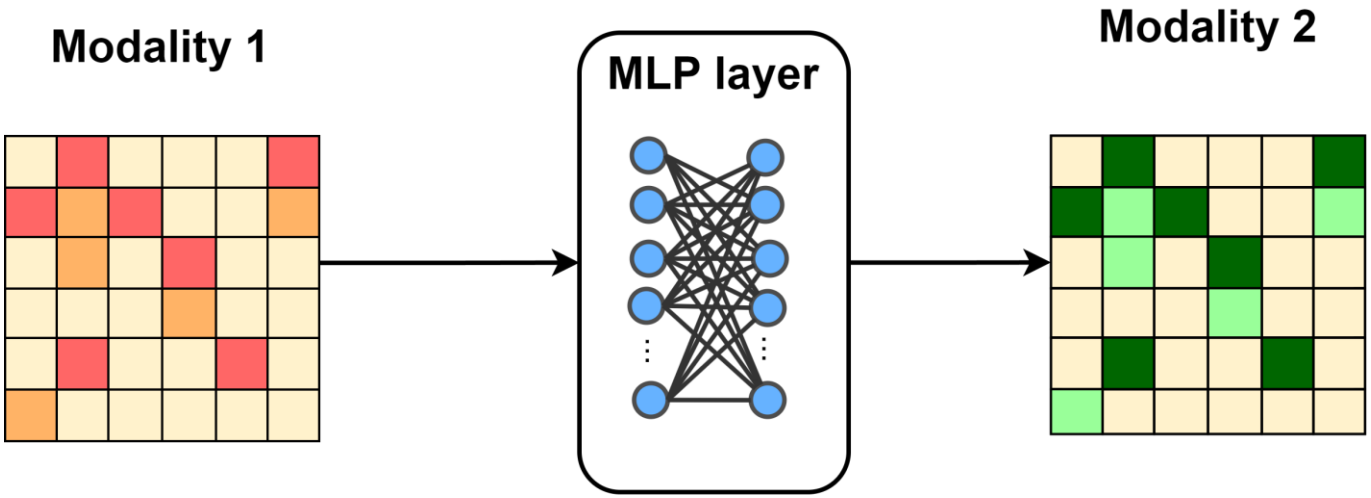
State-of-the-art in Modality Prediction

Model đề xuất bởi Dengkw có kết quả tốt nhất trong predict từ GEX thành ADT



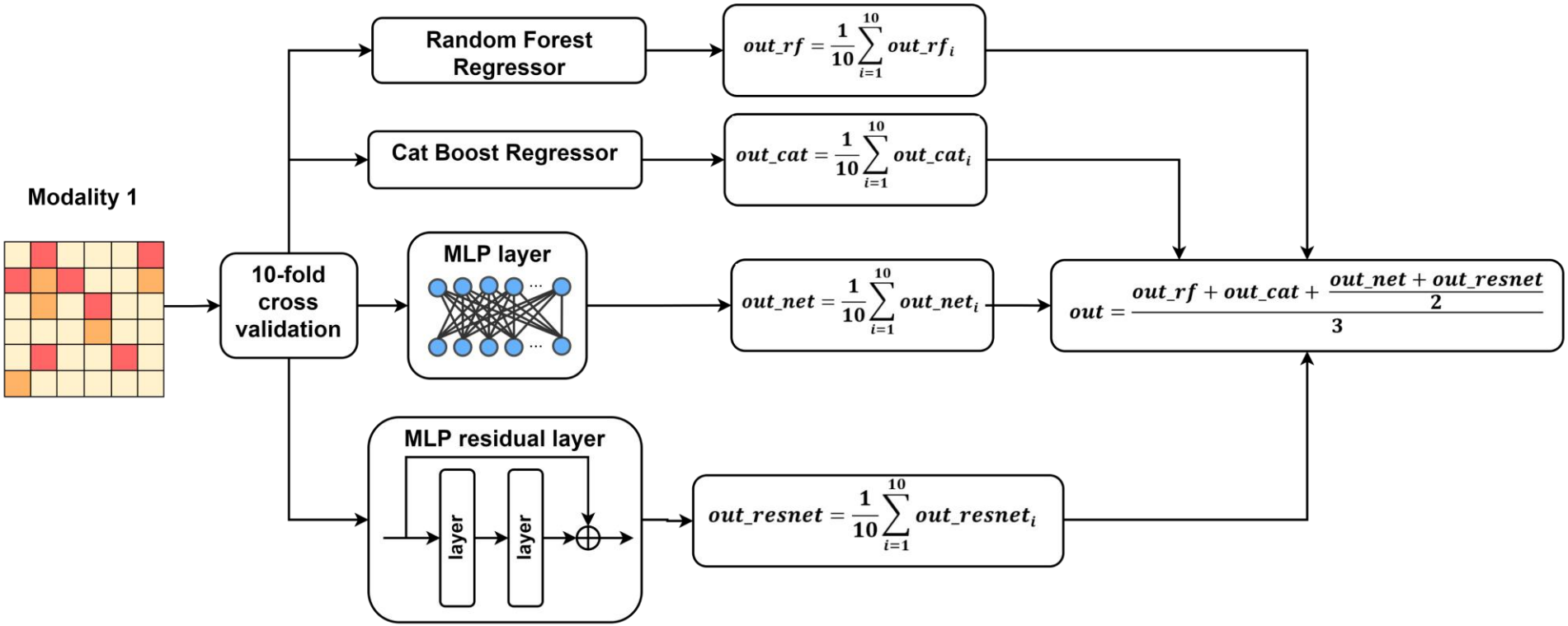
State-of-the-art in Modality Prediction

Model đề xuất bởi nhóm tác giả NOVEL có kết quả tốt nhất trong predict từ ADT thành GEX



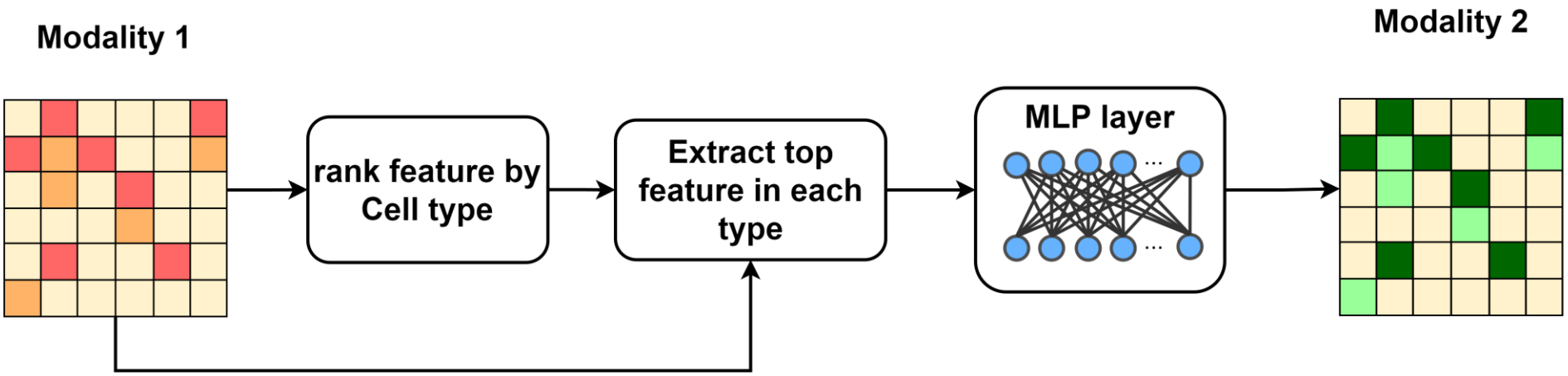
State-of-the-art in Modality Prediction

Model đề xuất bởi nhóm tác giả LSlab có kết quả tốt nhất trong predict từ GEX thành ATAC



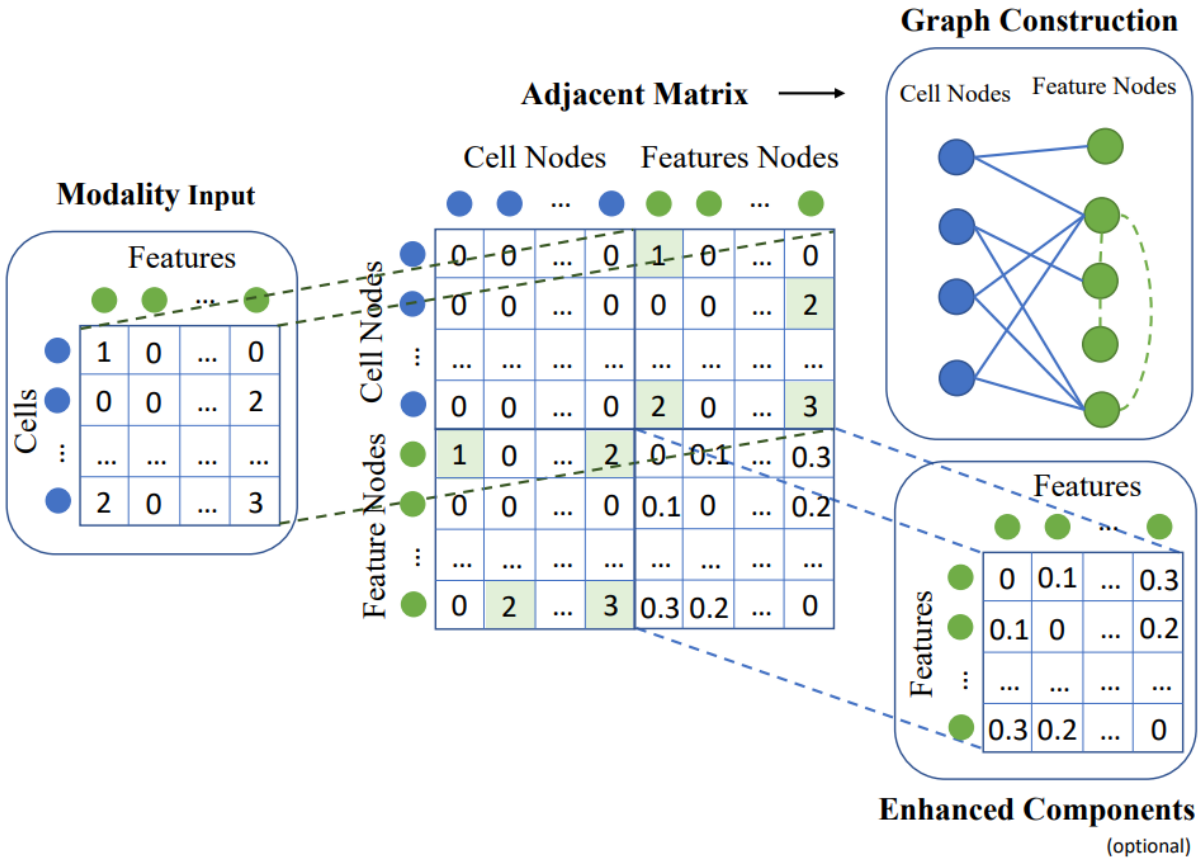
State-of-the-art in Modality Prediction

Model đề xuất bởi nhóm tác giả CAJAL có kết quả tốt nhất trong predict từ ATAC thành GEX

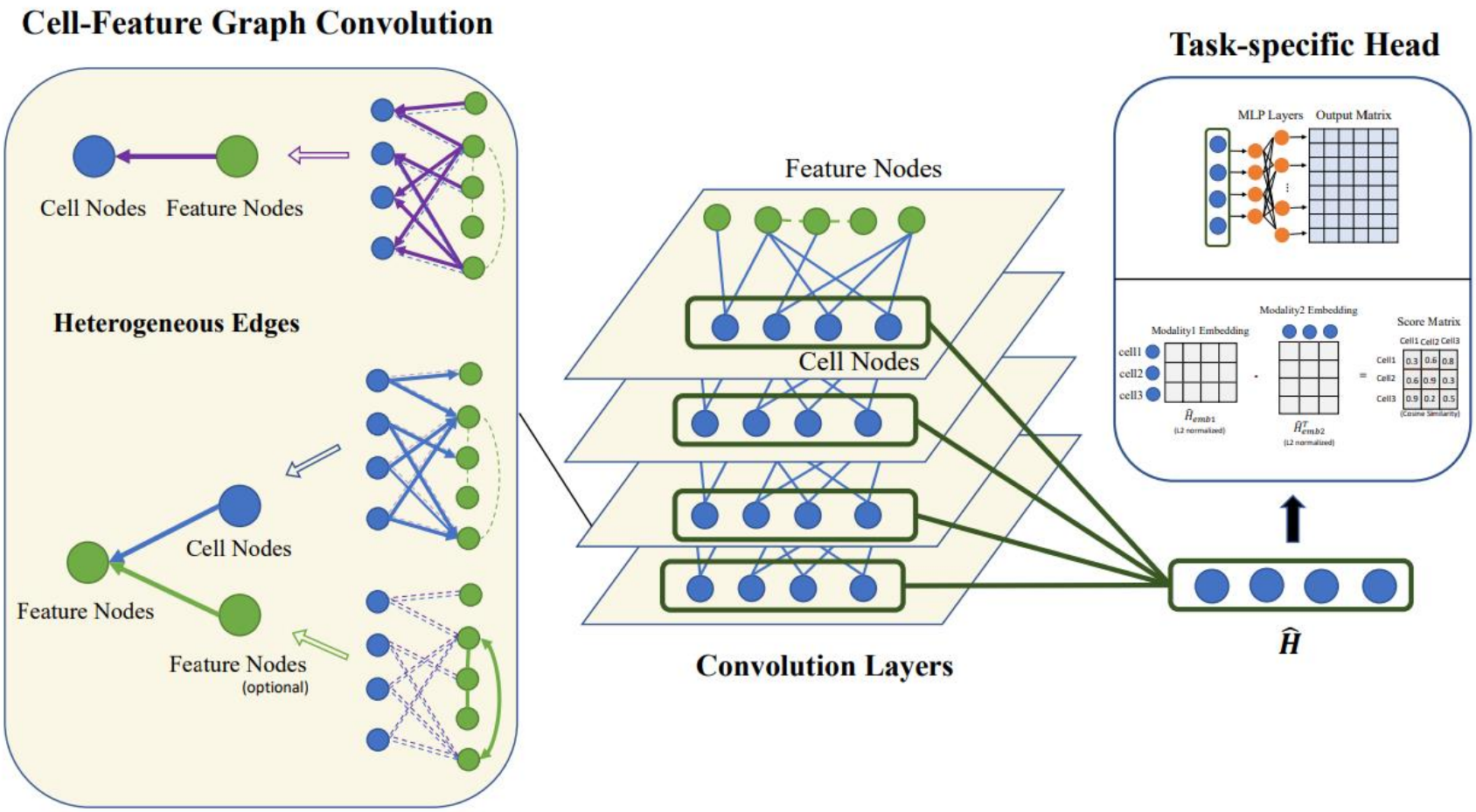


State-of-the-art in Modality Prediction

Wen và các cộng sự đề xuất mô hình sử dụng Graph Convolutional Network và có kết quả trung bình dự đoán tất cả các cặp là tốt nhất

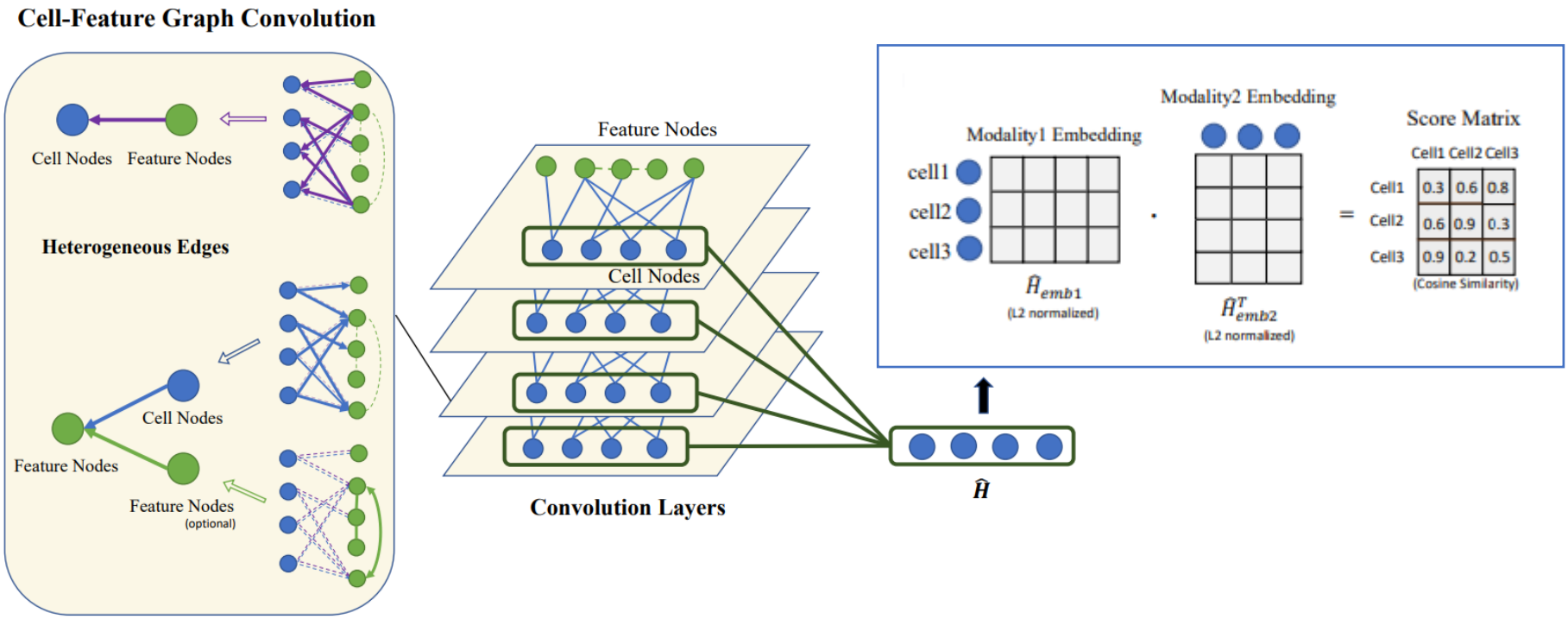


State-of-the-art in Modality Prediction



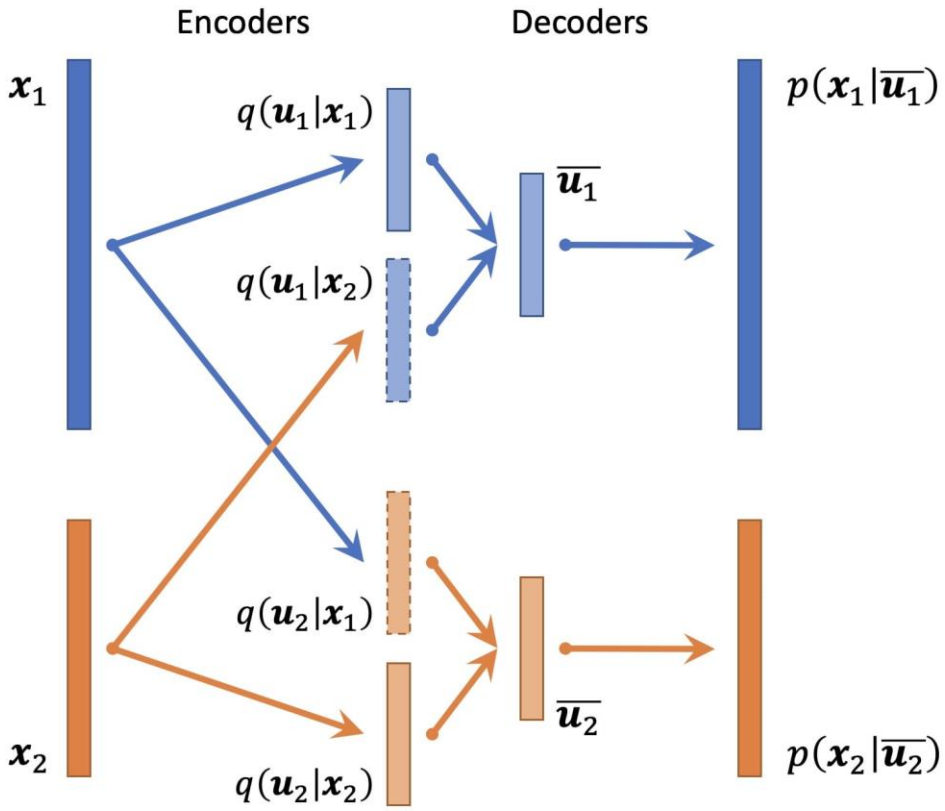
State-of-the-art in Modality Matching

Ở Task này mô hình GCN đề xuất bởi Wen và các cộng sự đạt kết quả matching cao nhất ở mọi cặp modality



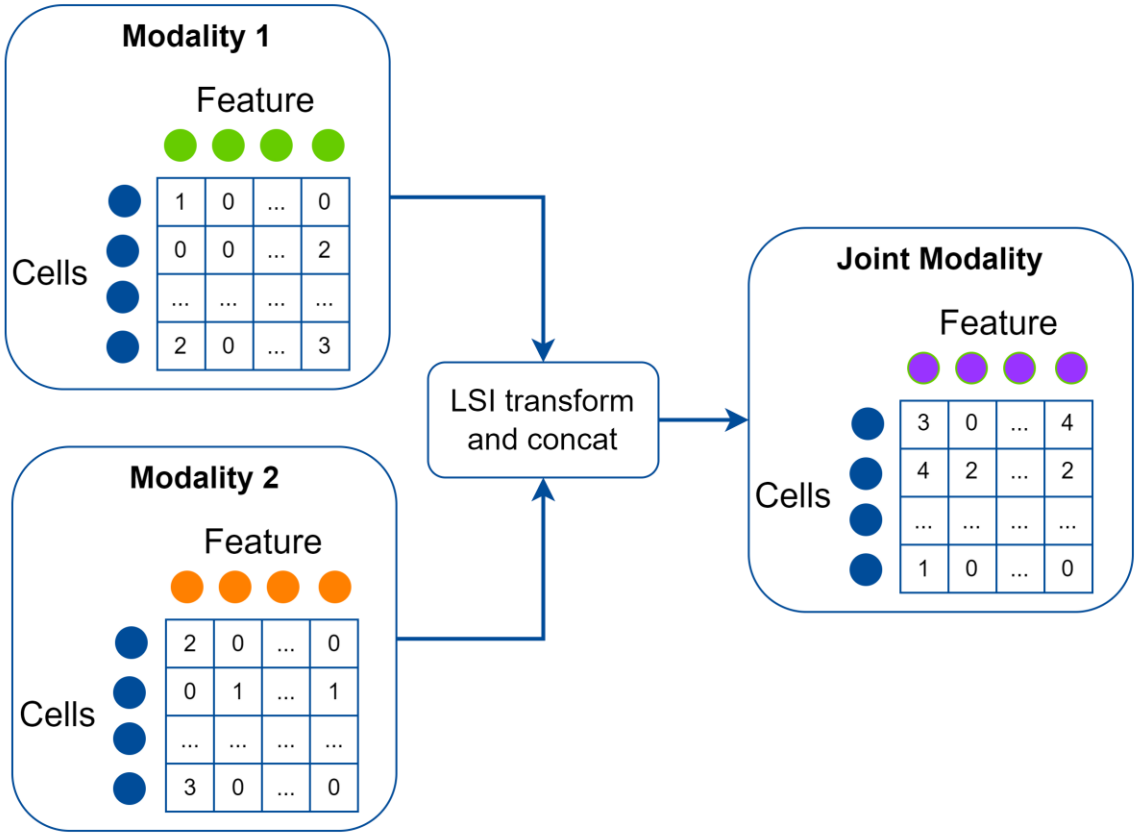
State-of-the-art in Modality Matching

Ngoài ra còn có mô hình autoencoder kép đề xuất bởi nhóm tác giả GLUE đạt kết quả tương đối tốt

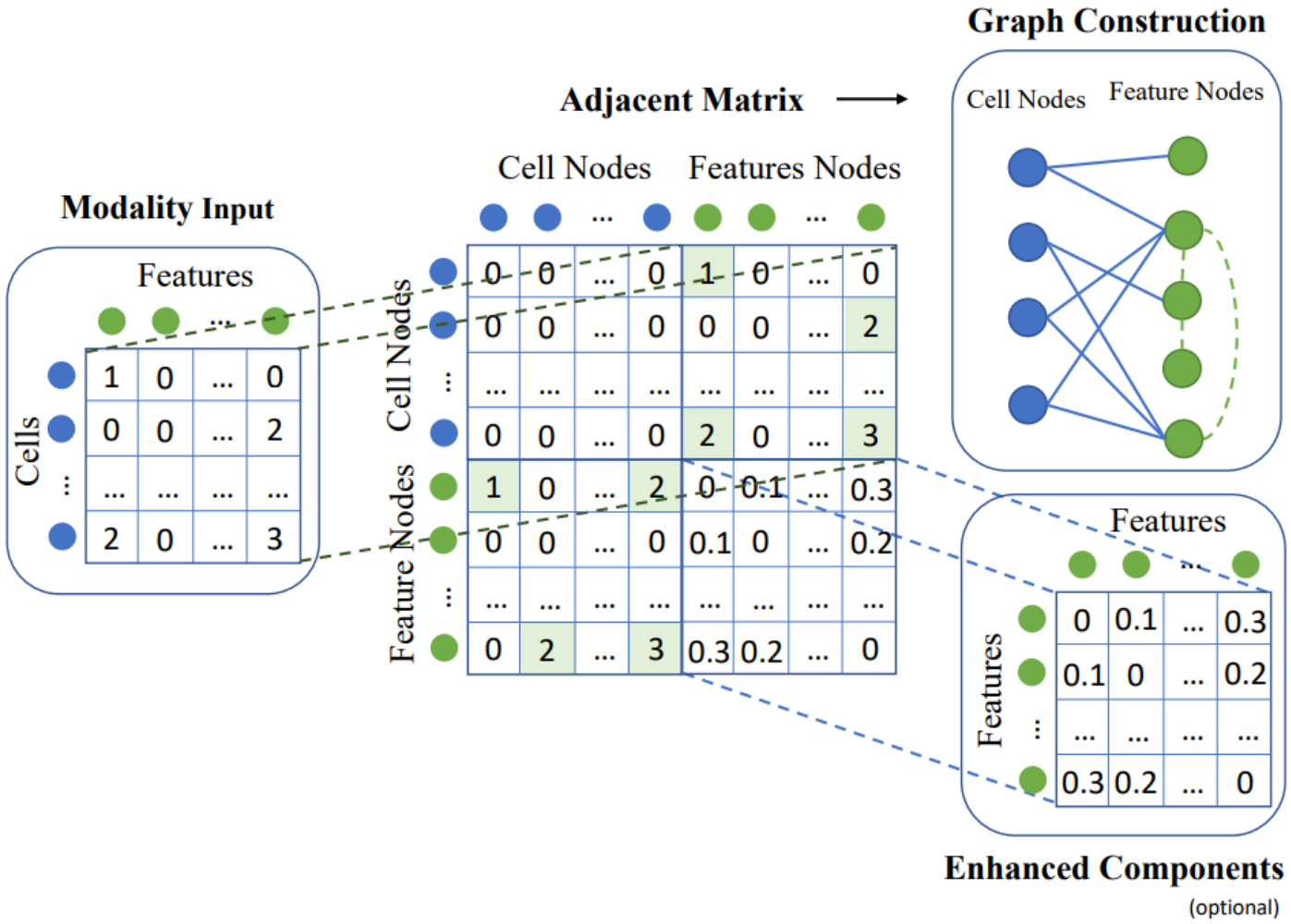


State-of-the-art in Joint Modality

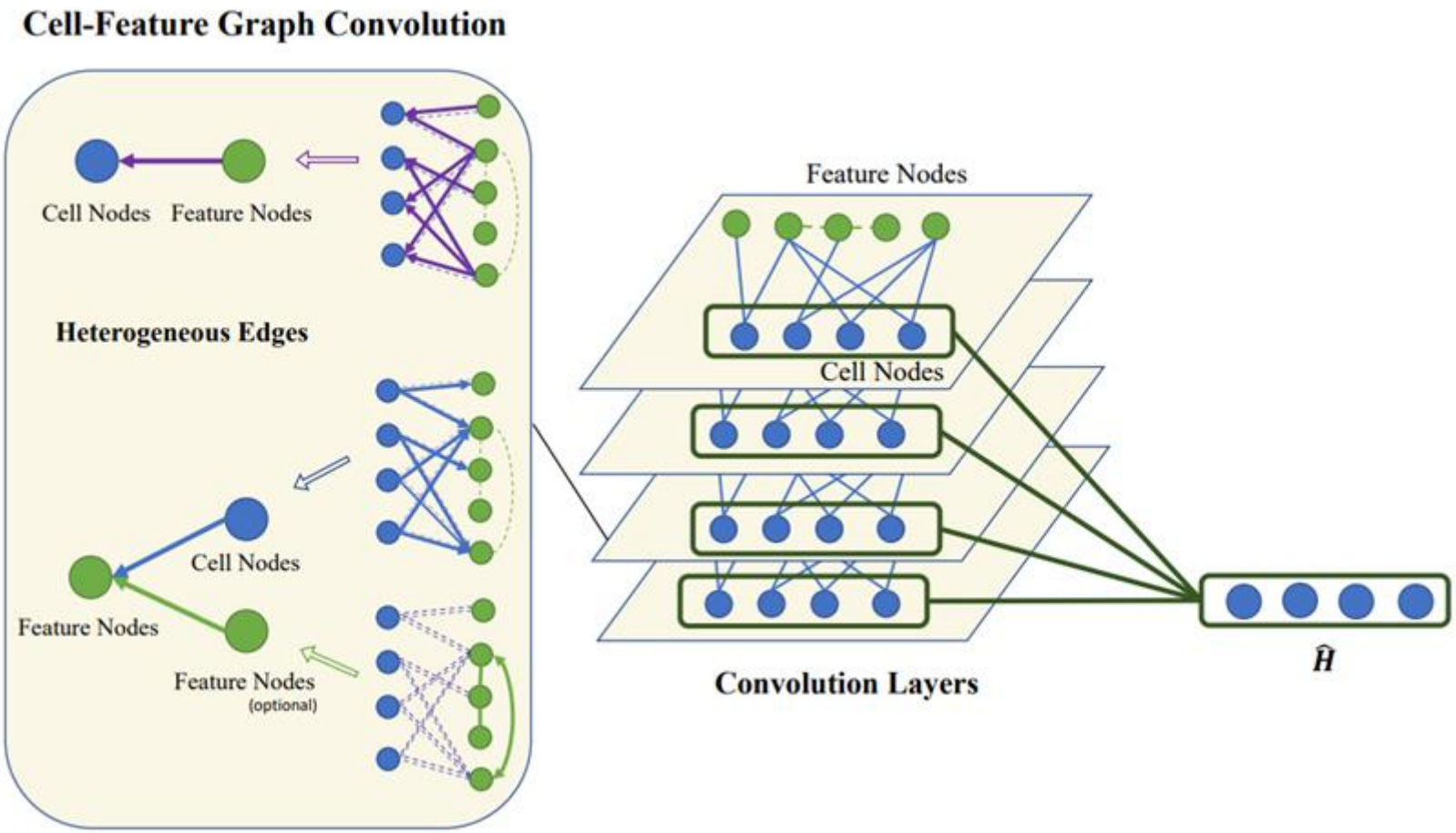
Trong task này mô hình GCN đề xuất bởi Wen và các cộng sự tiếp tục đạt kết quả cao nhất tại các metric NMI cluster/label, Cell type ASW, Batch ASW, Graph connectivity và Average Metric.



State-of-the-art in Modality Matching



State-of-the-art in Modality Matching



State-of-the-art in Joint Modality

Mô hình AutoEncoder đề xuất bởi nhóm tác giả JAE đạt kết quả cao nhất tại các metric Cell cycle conservation, Trajectory conservation.

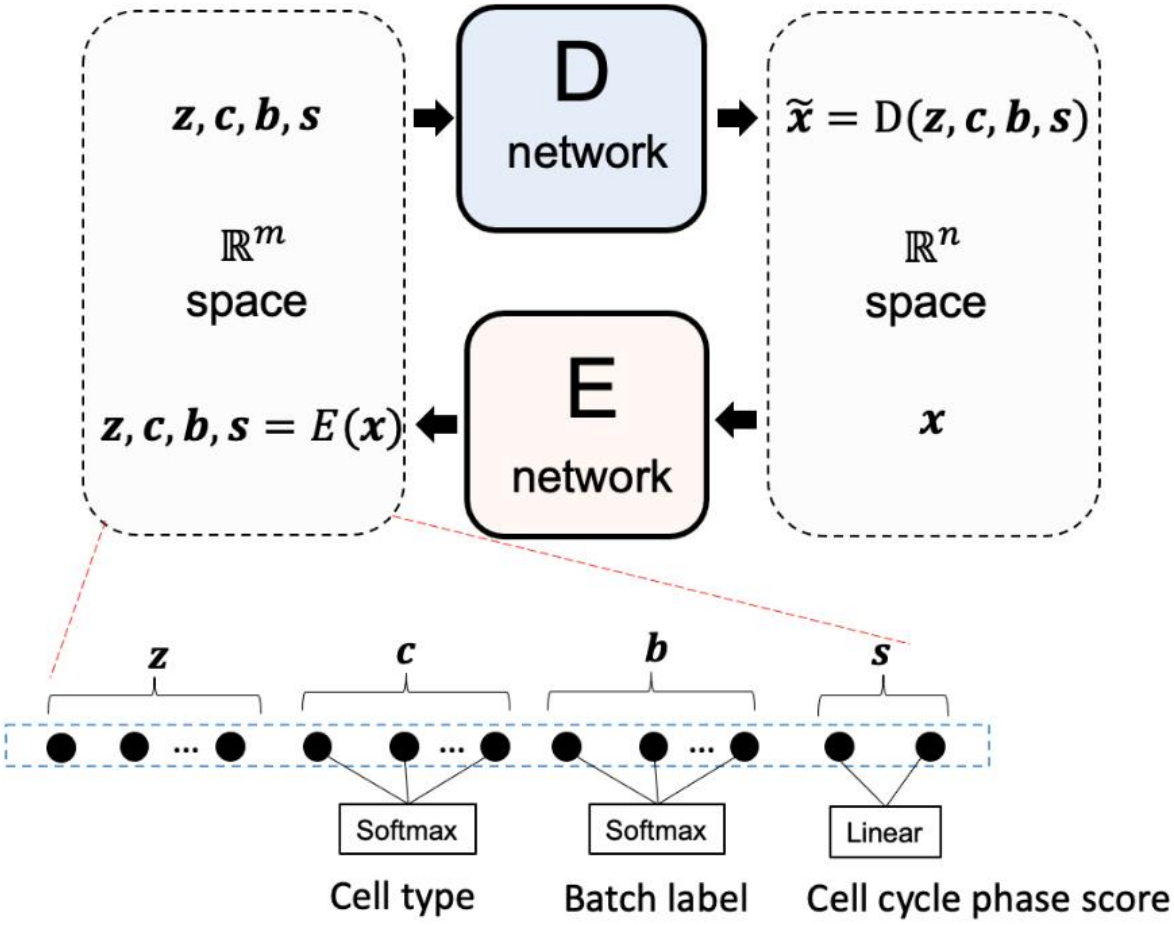
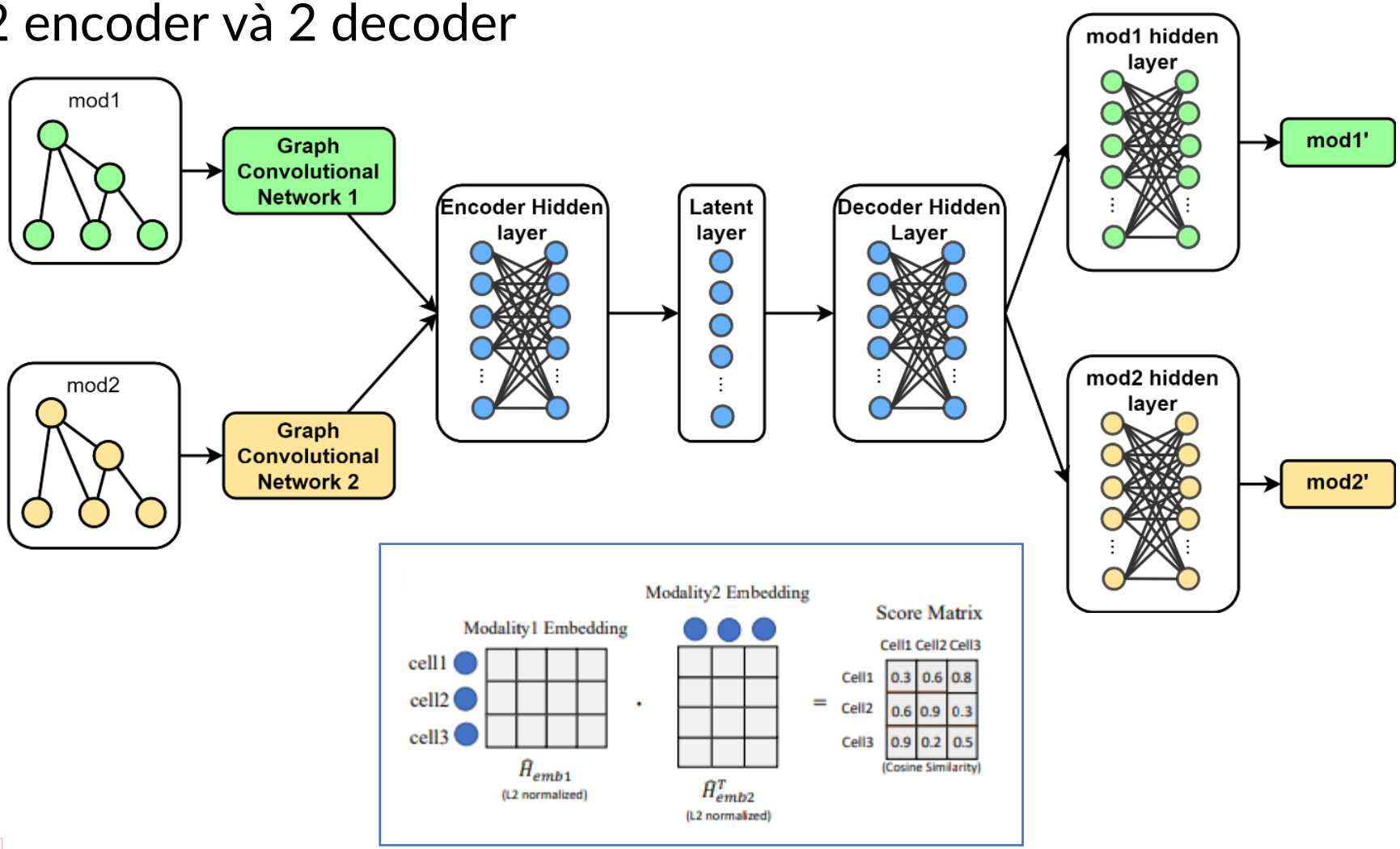


Table of content

Mô hình đề xuất

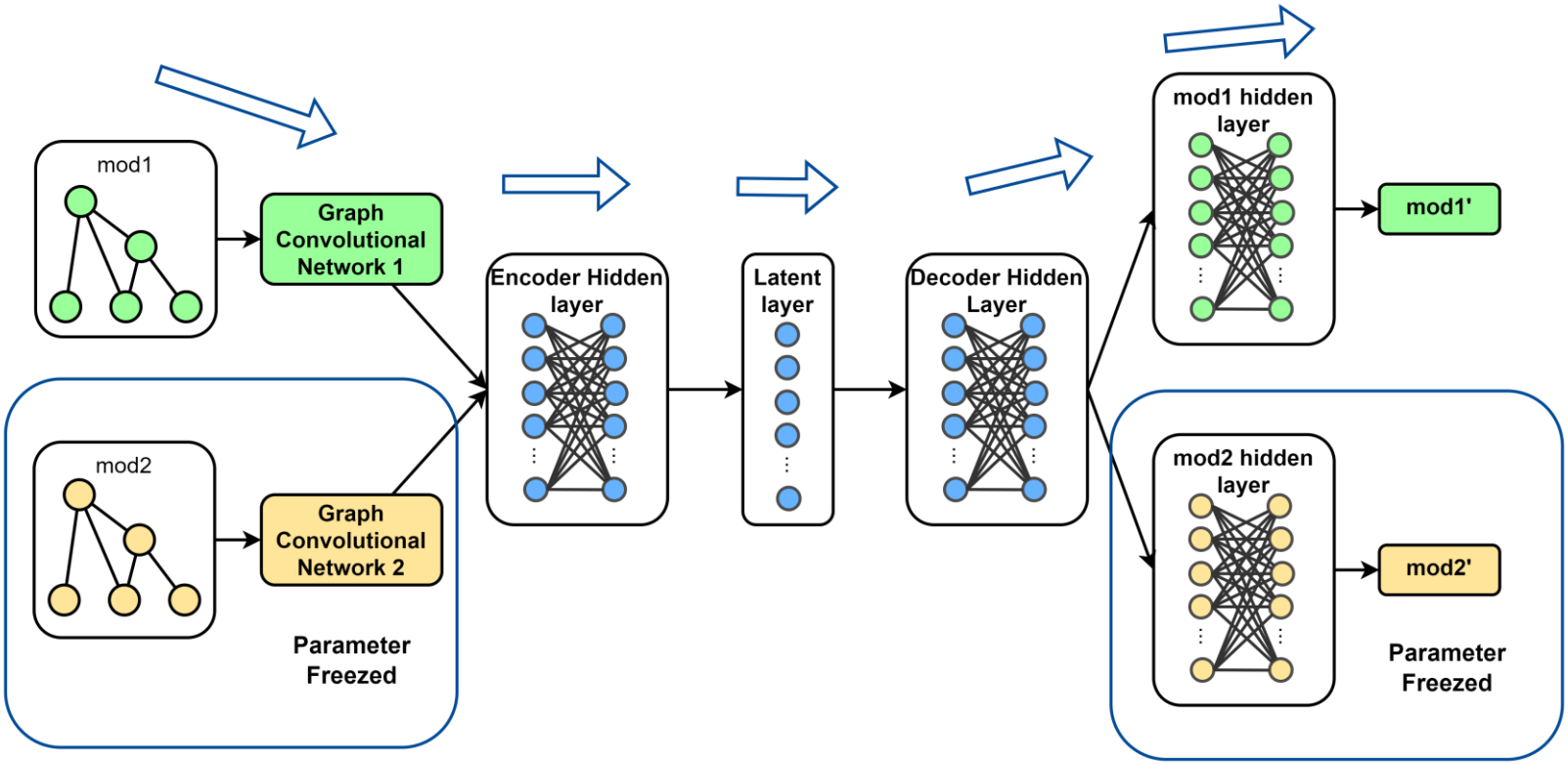
Mô hình đề xuất của em là một mô hình joint auto encoder. Với 2 encoder và 2 decoder



Mô hình đề xuất

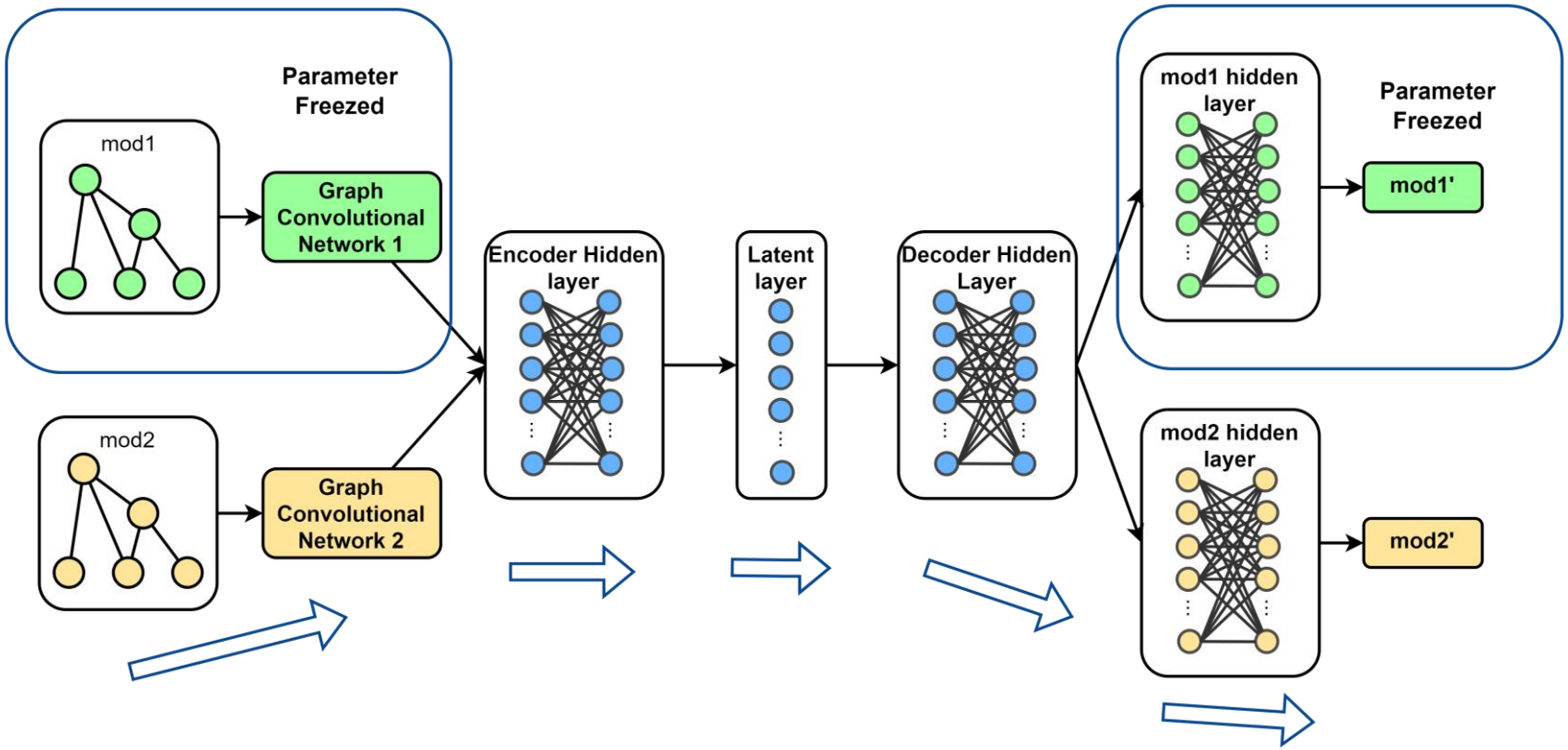
Với mỗi epoch, mô hình được huấn luyện và cập nhật các tham số với bốn bước như sau:

Bước 1:



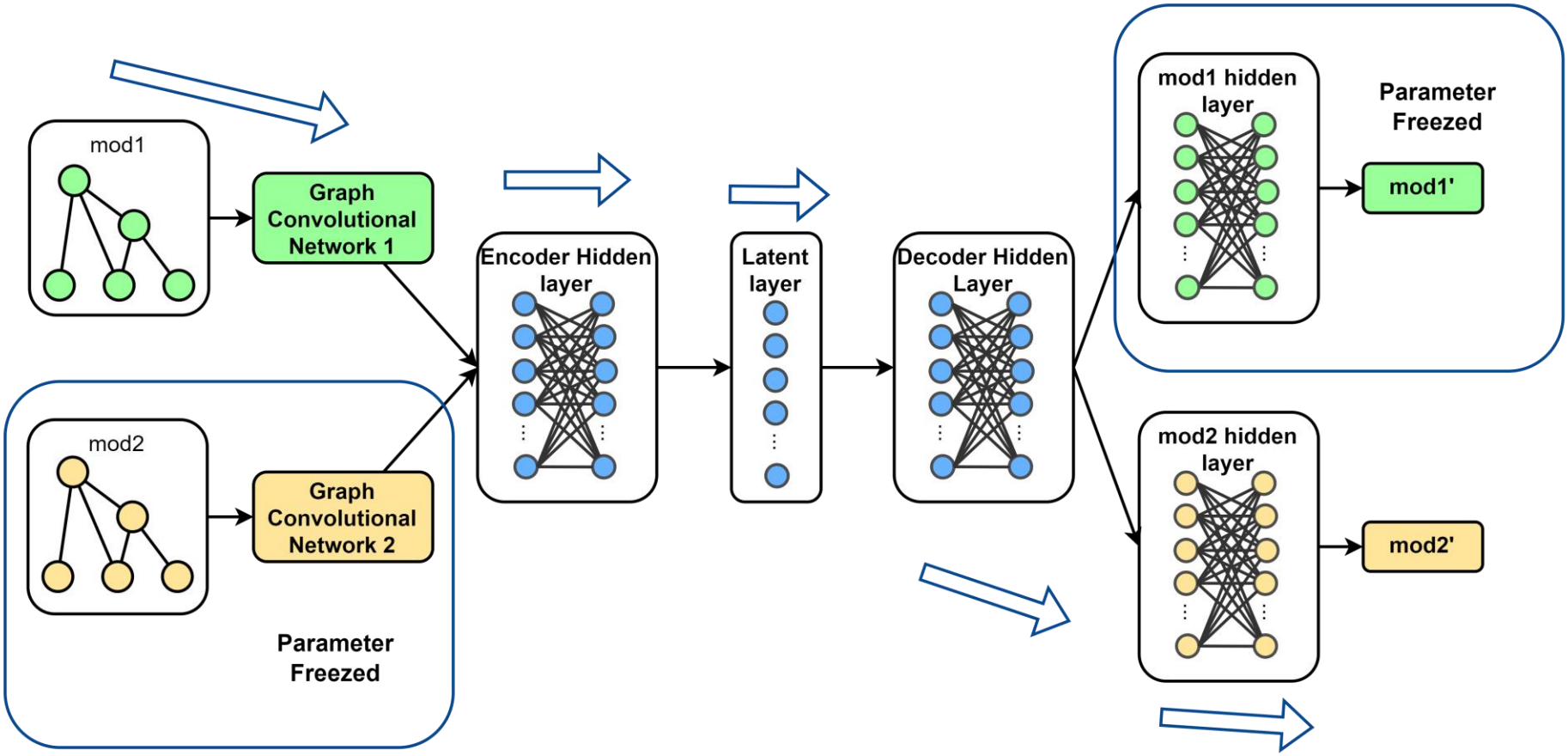
Mô hình đề xuất

Bước 2:



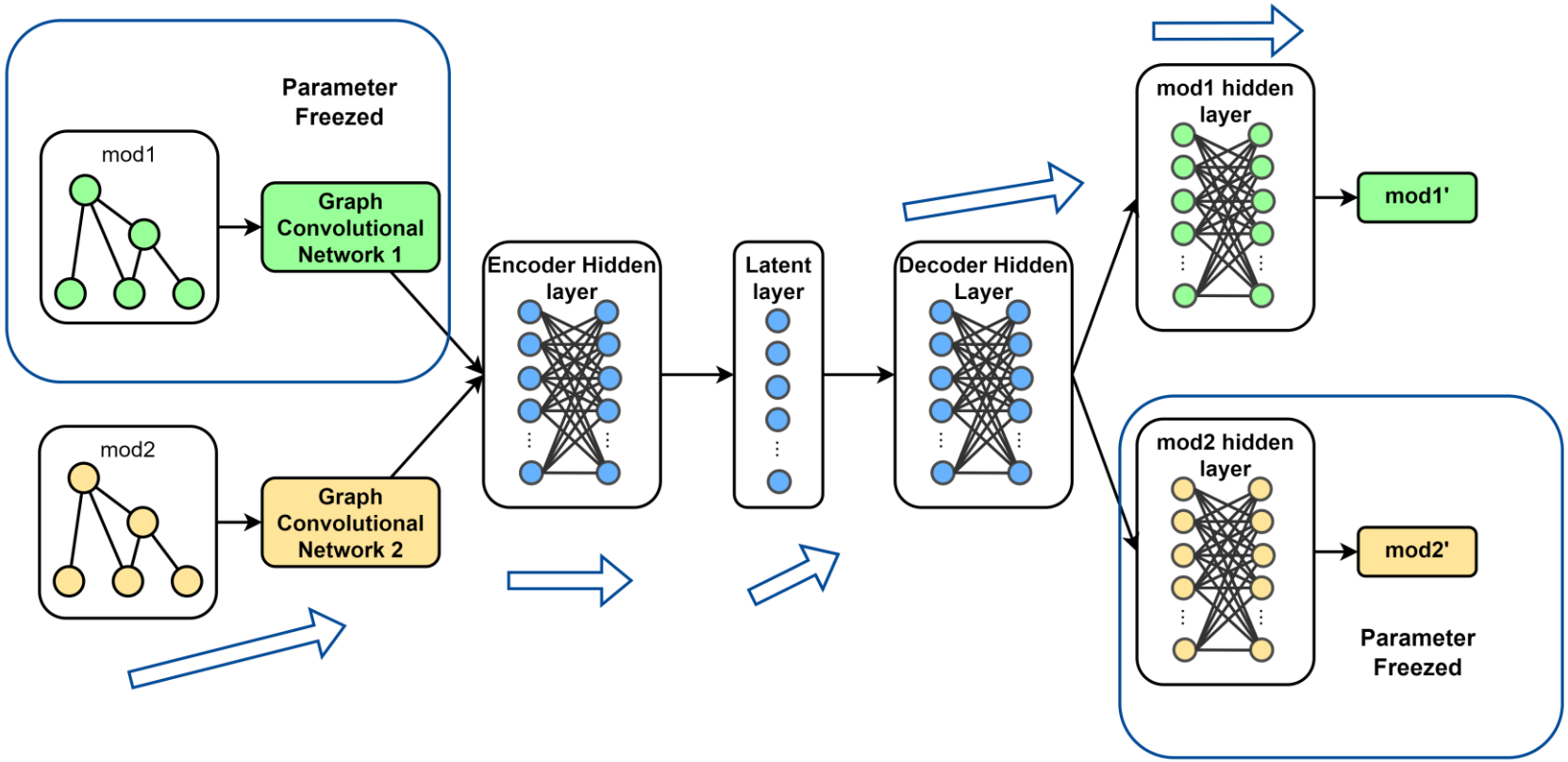
Mô hình đề xuất

Bước 3:



Mô hình đề xuất

Bước 4:



Các vấn đề gặp phải

- Chưa có phương pháp xây dựng đồ thị hiệu quả. Vẫn đang sử dụng đồ thị đơn giản nhất là k nearest neighbor graph
- Bộ dữ liệu quá lớn tốn rất nhiều ram để xây dựng graph dẫn đến k đủ bộ nhớ để training
- Dữ liệu có số chiều rất lớn tuy nhiên hầu hết có giá trị 0. Cần phải giảm chiều dữ liệu tuy nhiên chưa có nhiều kinh nghiệm về các phương pháp giảm chiều dữ

Các hướng triển khai tiếp

- Tìm kiếm và thử nghiệm thêm với các kiểu graph khác nhau
- Tìm hiểu và thử nghiệm thêm với các phương pháp giảm chiều dữ liệu khác
- Thử nghiệm mô hình với bộ nhỏ dữ liệu sau đó thuê các server để giải quyết vấn đề thiếu ram khi training

HUST

hust.edu.vn fb.com/dhbkhn

THANK YOU !