

CAFA6 2025: Project Report

1st Nguyen Duc Dung
Student ID: 23021495

2nd Nguyen Duc Hung
Student ID: 23021583

Tóm tắt nội dung—Dự đoán chức năng protein là một bài toán quan trọng trong sinh học tính toán, đặc biệt đối với các protein chưa được chú thích thực nghiệm. Trong khuôn khổ CAFA 6, nhiệm vụ đặt ra là dự đoán các Gene Ontology (GO) terms cho protein chưa có annotation thuộc ba ontology: Cellular Component (CC), Molecular Function (MF) và Biological Process (BP).

Trong bài báo cáo này, nhóm đề xuất một pipeline dự đoán chức năng protein dựa trên embedding ESM2 kết hợp với mô hình học sâu phân loại đa nhãn, được huấn luyện riêng cho từng ontology. Các bước hậu xử lý dựa trên Gene Ontology hierarchy và GOA negative constraints được áp dụng nhằm cải thiện tính nhất quán sinh học. Ngoài ra, một chiến lược ensemble ở mức submission được sử dụng để kết hợp các nguồn dự đoán bổ sung. Kết quả thực nghiệm cho thấy phương pháp đề xuất đạt điểm Fmax = 0.315 trên bộ đánh giá CAFA 6.

Index Terms—Dự đoán chức năng protein, Gene Ontology, CAFA 6, Deep Learning, Ensemble

I. INTRODUCTION

Dự đoán chức năng protein đóng vai trò quan trọng trong việc hiểu các cơ chế sinh học và hỗ trợ nghiên cứu y sinh. Mặc dù công nghệ giải trình tự protein phát triển nhanh chóng, số lượng protein được chú thích chức năng thực nghiệm vẫn còn hạn chế do chi phí và thời gian thí nghiệm lớn.

CAFA (Critical Assessment of Function Annotation) là một khuôn khổ đánh giá chuẩn nhằm so sánh khách quan các phương pháp dự đoán chức năng protein. Trong CAFA 6, các phương pháp phải dự đoán Gene Ontology (GO) terms cho protein chưa có annotation tại thời điểm đóng băng dữ liệu, và kết quả được đánh giá dựa trên các annotation được phát hiện trong tương lai.

Trong nghiên cứu này, nhóm xây dựng một hệ thống dự đoán chức năng protein kết hợp học sâu và tri thức sinh học. Protein được biểu diễn bằng embedding ESM2, các mô hình phân loại đa nhãn được huấn luyện riêng cho từng ontology, và các bước hậu xử lý dựa trên Gene Ontology được áp dụng nhằm cải thiện độ chính xác và tính nhất quán sinh học.

II. PROBLEM DEFINITION AND DATA

A. Định nghĩa bài toán

Trong khuôn khổ CAFA 6, bài toán dự đoán chức năng protein được mô hình hóa như một bài toán phân loại đa nhãn, trong đó mỗi protein có thể đồng thời gắn với nhiều Gene Ontology (GO) terms.

Cụ thể, với mỗi protein đầu vào, hệ thống cần dự đoán một tập con các GO terms thuộc ba ontology độc lập: Cellular Component (CCO), Molecular Function (MFO) và Biological Process (BPO). Mỗi ontology có tập nhãn, độ phức tạp và

mức độ mất cân bằng khác nhau, dẫn đến sự khác biệt đáng kể trong phân bố nhãn và số lượng annotation trên mỗi protein.

Đầu vào của hệ thống là chuỗi amino acid của protein, được biểu diễn dưới dạng vector embedding có chiều cố định. Đầu ra là một tập các cặp {GO term, xác suất}, biểu diễn mức độ tin cậy của hệ thống đối với từng chức năng được dự đoán. Do đặc tính đa nhãn và phân cấp của Gene Ontology, bài toán này vượt xa phân loại phẳng thông thường và đòi hỏi các ràng buộc sinh học bổ sung trong quá trình suy luận.

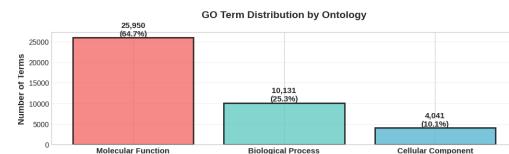
B. Thiết lập đánh giá trong CAFA 6

CAFA 6 sử dụng cơ chế đánh giá đặc thù dựa trên “future annotations”, trong đó các dự đoán được so sánh với các GO annotations được xác nhận sau thời điểm đóng băng dữ liệu. Do đó, các annotation có sẵn tại thời điểm huấn luyện không được xem là ground truth hoàn chỉnh cho tập kiểm tra.

Chỉ số đánh giá chính trong CAFA là Fmax, được xác định là giá trị F1-score lớn nhất đạt được khi quét qua các ngưỡng xác suất khác nhau. Chỉ số này phản ánh sự cân bằng giữa precision và recall, và đặc biệt nhạy cảm với các false positives trong không gian nhãn lớn như Gene Ontology.

Đặc điểm này khiến việc lựa chọn ngưỡng dự đoán, số lượng GO tối thiểu trên mỗi protein, cũng như các bước hậu xử lý nhằm giảm false positives trở nên đặc biệt quan trọng.

C. Bộ dữ liệu huấn luyện và kiểm tra



Hình 1. Phân bố số lượng GO terms theo từng ontology trong Gene Ontology.

Nhóm sử dụng bộ dữ liệu chính thức do ban tổ chức CAFA 6 cung cấp. Tập huấn luyện bao gồm các protein đã có GO annotation tại thời điểm đóng băng dữ liệu, trong khi tập kiểm tra chứa các protein chưa có annotation vào thời điểm này.

Mỗi protein trong tập huấn luyện có thể được gán nhiều GO terms thuộc một hoặc nhiều ontology. Để giảm nhiễu và đảm bảo tính nhất quán dữ liệu, chỉ những protein đồng thời có embedding hợp lệ và ít nhất một annotation thuộc ontology tương ứng mới được sử dụng trong quá trình huấn luyện.

Bài toán dự đoán chức năng protein trong khuôn khổ CAFA6 được mô hình hóa như một bài toán phân loại đa nhãn (multi-label classification) với tập nhãn lớn và có cấu trúc phân cấp theo Gene Ontology (GO).

Cho một protein i với biểu diễn đặc trưng $\mathbf{x}_i \in \mathbb{R}^d$, mục tiêu là học một hàm dự đoán:

$$f : \mathbb{R}^d \rightarrow [0, 1]^{|\mathcal{T}|}, \quad (1)$$

trong đó \mathcal{T} là tập các GO term và mỗi phần tử $\hat{y}_{i,t} = f_t(\mathbf{x}_i)$ biểu diễn xác suất protein i được gán với GO term t .

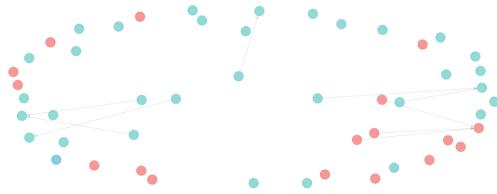
Do cấu trúc phân cấp của Gene Ontology, các nhãn không độc lập mà chịu ràng buộc theo quan hệ tổ tiên-hậu duệ, điều này yêu cầu các bước xử lý hậu kỳ để đảm bảo tính nhất quán sinh học của kết quả dự đoán.

Do sự khác biệt lớn về số lượng nhãn giữa các ontology, dữ liệu được tách và xử lý độc lập cho từng ontology, cho phép mô hình học được các đặc trưng chuyên biệt và giảm ảnh hưởng của mêt cân bằng nhãn.

D. Nguồn tri thức sinh học bổ sung

Ngoài dữ liệu annotation, nhóm khai thác hai nguồn tri thức sinh học quan trọng: Gene Ontology hierarchy và GOA UniProt annotations.

Gene Ontology hierarchy được cung cấp dưới dạng file OBO, mô tả mối quan hệ phân cấp giữa các GO terms thông qua các quan hệ như *is_a* và *part_of*. Cấu trúc này cho phép áp dụng nguyên tắc “true path rule”, theo đó nếu một protein có chức năng cụ thể thì nó cũng sở hữu các chức năng tổng quát hơn tương ứng với các ancestor trong ontology.



Hình 2. Minh họa cấu trúc phân cấp của Gene Ontology và sự phân tách ba ontology chính.

Bên cạnh đó, dữ liệu GOA UniProt cung cấp các annotation phủ định thông qua qualifier “NOT”, biểu thị các chức năng đã được thực nghiệm chứng minh là không đúng đối với một protein nhất định. Các annotation phủ định này được sử dụng như các ràng buộc âm, và được lan truyền xuống các GO terms con nhằm loại bỏ các dự đoán sai lệch về mặt sinh học.

E. Tóm tắt bài toán

Tổng hợp lại, bài toán được xem là một bài toán phân loại đa nhãn, đa ontology, có cấu trúc phân cấp và chịu ràng buộc sinh học. Do đó, một phương pháp hiệu quả cần đồng thời kết hợp mô hình học sâu mạnh mẽ và các bước hậu xử lý dựa trên tri thức sinh học nhằm đảm bảo cả hiệu năng dự đoán và tính hợp lệ sinh học của kết quả.

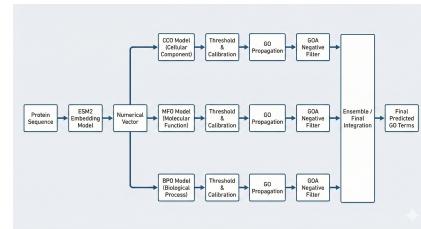
III. PROPOSED METHOD

A. Tổng quan pipeline

Hệ thống đề xuất bao gồm bốn thành phần chính:

- Biểu diễn protein bằng embedding học sâu.

- Huấn luyện các mô hình phân loại đa nhãn riêng biệt cho từng ontology.
- Suy luận có điều chỉnh theo từng ontology.
- Các bước hậu xử lý dựa trên tri thức sinh học và ensemble ở mức submission.



Hình 3. Minh họa Pipeline thiết kế tổng thể.

Pipeline tổng thể được thiết kế nhằm tận dụng đồng thời sức mạnh của mô hình học sâu và các ràng buộc sinh học có cấu trúc, giúp cải thiện hiệu năng dự đoán trong bối cảnh không gian nhãn lớn và phân cấp như Gene Ontology.

B. Biểu diễn protein bằng embedding

Mỗi protein được biểu diễn bằng vector embedding cố định chiều $d = 1280$, được trích xuất sẵn từ mô hình ESM2. Các embedding này được đưa vào một mạng nơ-ron nhiều tầng (MLP) để sinh ra điểm số cho từng GO term.

Cụ thể, với embedding \mathbf{x}_i , mạng dự đoán được mô tả như sau:

$$\mathbf{z}_i = W_2\phi(W_1\mathbf{x}_i + \mathbf{b}_1) + \mathbf{b}_2, \quad (2)$$

trong đó W_1, W_2 và $\mathbf{b}_1, \mathbf{b}_2$ là các tham số học được, $\phi(\cdot)$ là hàm kích hoạt ReLU.

Xác suất dự đoán cuối cùng cho mỗi GO term được tính bằng hàm sigmoid:

$$\hat{y}_i = \sigma(\mathbf{z}_i), \quad (3)$$

với $\hat{y}_{i,t} \in [0, 1]$ biểu diễn mức độ tin cậy của việc gán GO term t cho protein i .

Trong nghiên cứu này, các embedding ESM2 được sử dụng dưới dạng cố định, không fine-tune trong quá trình huấn luyện. Cách tiếp cận này giúp giảm chi phí tính toán và tránh overfitting, đồng thời vẫn đạt hiệu năng cạnh tranh trong bài toán dự đoán chức năng protein.

C. Huấn luyện mô hình theo từng ontology

Do sự khác biệt đáng kể về số lượng nhãn, mức độ mêt cân bằng và độ phức tạp giữa ba ontology, nhóm huấn luyện ba mô hình phân loại đa nhãn độc lập, tương ứng với Cellular Component, Molecular Function và Biological Process.

Mỗi mô hình được huấn luyện trên tập protein có annotation thuộc ontology tương ứng. Nhãn được mã hóa bằng phương pháp multi-hot encoding, và bài toán được mô hình hóa dưới dạng One-vs-Rest classification với hàm mất mát BCEWithLogitsLoss.

Cách tiếp cận theo từng ontology giúp mô hình học được các đặc trưng chuyên biệt, đồng thời cho phép áp dụng các chiến lược suy luận và hiệu chỉnh điểm dự đoán khác nhau cho mỗi ontology trong giai đoạn suy luận.

D. Kiến trúc mạng học sâu

Mỗi mô hình phân loại sử dụng một mạng Multi-Layer Perceptron (MLP) gồm ba tầng fully-connected. Tầng đầu tiên ánh xạ embedding đầu vào sang không gian đặc trưng trung gian, tiếp theo là các tầng ẩn kết hợp Batch Normalization và Dropout nhằm cải thiện khả năng hội tụ và giảm overfitting.

Tầng đầu ra có số neuron tương ứng với số lượng GO terms trong ontology, và sử dụng sigmoid activation để sinh ra xác suất độc lập cho từng nhãn. Thiết kế này phù hợp với bản chất phân loại đa nhãn của bài toán.

E. Chiến lược suy luận theo ontology

Trong giai đoạn suy luận, các chiến lược khác nhau được áp dụng cho từng ontology nhằm phản ánh sự khác biệt về độ phức tạp và phân bố nhãn. Cụ thể, mỗi ontology sử dụng một ngưỡng xác suất tối thiểu và số lượng GO term tối thiểu khác nhau trên mỗi protein.

Ngoài ra, một bước hiệu chỉnh điểm dự đoán được áp dụng thông qua phép biến đổi lũy thừa nhẹ trên xác suất đầu ra. Phép biến đổi này giúp làm sắc nét phân phôi điểm, tăng precision đối với các GO có độ tin cậy cao mà không thay đổi thứ tự tương đối giữa các dự đoán.

Bảng I
CẤU HÌNH SUY LUẬN THEO TỪNG ONTOLOGY

Ontology	Threshold	Min-K	Power
Cellular Component (CCO)	0.02	20	1.1
Molecular Function (MFO)	0.02	15	1.1
Biological Process (BPO)	0.04	10	1.2

Các tham số trong Bảng I được lựa chọn dựa trên đặc điểm phân bố nhãn của từng ontology. Biological Process có số lượng nhãn lớn và cấu trúc phân cấp sâu hơn, do đó sử dụng ngưỡng xác suất cao hơn và hệ số hiệu chỉnh lớn hơn nhằm giảm false positives. Ngược lại, Cellular Component và Molecular Function có phân bố nhãn gọn hơn, cho phép sử dụng ngưỡng thấp hơn và giữ lại nhiều dự đoán hơn trên mỗi protein.

F. Hậu xử lý dựa trên Gene Ontology

Để đảm bảo tính nhất quán sinh học của kết quả dự đoán, các GO terms được lan truyền lên tất cả các ancestor tương ứng trong Gene Ontology hierarchy theo nguyên tắc true path rule. Bước này đảm bảo rằng các chức năng tổng quát hơn cũng được gán cho protein nếu các chức năng cụ thể hơn được dự đoán.

Song song với đó, các annotation phủ định từ GOA UniProt được sử dụng để loại bỏ các dự đoán không hợp lệ. Các GO terms bị phủ định được lan truyền xuống các descendant, giúp giảm false positives trong không gian nhãn lớn.

G. Ensemble ở mức submission

Cuối cùng, một chiến lược ensemble ở mức submission được áp dụng bằng cách kết hợp kết quả dự đoán của hệ thống để xuất với một submission khác đã được xử lý bằng GOA-based filtering. Hai submission được gộp bằng cách lấy giá trị xác suất lớn nhất cho mỗi cặp protein-GO.

Chiến lược này cho phép tận dụng các dự đoán có độ tin cậy cao từ nhiều nguồn khác nhau mà không cần huấn luyện thêm mô hình, đồng thời tránh các vấn đề về không tương thích không gian nhãn giữa các mô hình.

IV. EXPERIMENTAL SETUP AND TRAINING

A. Experimental Setup

Các thí nghiệm trong nghiên cứu này được thực hiện trên bộ dữ liệu chính thức của cuộc thi CAFA 6. Chỉ các protein có vector embedding hợp lệ được sử dụng trong cả quá trình huấn luyện và suy luận. Ba ontology của Gene Ontology bao gồm Molecular Function (MFO), Biological Process (BPO) và Cellular Component (CCO) được xử lý độc lập, tương ứng với ba mô hình phân loại đa nhãn riêng biệt.

Cách tiếp cận huấn luyện theo ontology giúp giảm độ phức tạp của không gian nhãn và cho phép áp dụng các chiến lược huấn luyện và suy luận phù hợp với đặc điểm của từng ontology.

B. Protein Representation

Mỗi protein được biểu diễn bằng một vector embedding chiều 1280, được trích xuất sẵn từ mô hình ngôn ngữ protein ESM-2 đã được tiền huấn luyện. Các embedding này được giữ cố định trong suốt quá trình huấn luyện, nhằm tập trung đánh giá hiệu quả của mô hình phân loại và chiến lược hậu xử lý.

C. Training Protocol

Mô hình được huấn luyện theo cách giám sát với hàm mất mát Binary Cross-Entropy áp dụng độc lập cho từng GO term.

Với nhãn thật $y_{i,t} \in \{0, 1\}$ và logit $z_{i,t}$, hàm mất mát cho một protein được định nghĩa như sau:

$$\mathcal{L}_i = - \sum_{t=1}^{|\mathcal{T}|} [y_{i,t} \log \sigma(z_{i,t}) + (1 - y_{i,t}) \log(1 - \sigma(z_{i,t}))]. \quad (4)$$

Tổng mất mát trên tập huấn luyện được tối ưu hóa bằng thuật toán AdamW, với cơ chế điều chỉnh tốc độ học dựa trên giá trị mất mát của tập validation.

D. Model Architecture

Mỗi mô hình sử dụng kiến trúc Multi-Layer Perceptron (MLP) gồm ba tầng fully-connected. Tầng đầu vào có kích thước 1280, tương ứng với embedding ESM-2. Hai tầng ẩn lần lượt có kích thước 1024 và 512, sử dụng hàm kích hoạt ReLU. Batch Normalization và Dropout với tỉ lệ 0.3 được áp dụng nhằm cải thiện khả năng tổng quát hóa.

Tầng đầu ra có số node tương ứng với số lượng GO terms của ontology đang xét và không sử dụng hàm kích hoạt để phù hợp với hàm mất mát BCEWithLogitsLoss.

E. Training Strategy per Ontology

Ba mô hình độc lập được huấn luyện cho ba ontology MFO, BPO và CCO. Chiến lược này cho phép mô hình học được các đặc trưng riêng của từng ontology và tạo điều kiện thuận lợi cho việc áp dụng các chiến lược suy luận và hiệu chỉnh xác suất khác nhau ở giai đoạn inference, như đã trình bày trong Section III.

V. INFERENCE AND POST-PROCESSING

A. Aspect-aware Inference

Sau khi huấn luyện ba mô hình riêng biệt cho ba ontology, quá trình suy luận được thực hiện độc lập cho từng ontology. Đối với mỗi protein trong tập kiểm tra, mô hình tương ứng sinh ra xác suất dự đoán cho toàn bộ các GO terms thuộc ontology đó.

Do đặc điểm phân bố nhau giữa các ontology, nhóm áp dụng chiến lược suy luận theo ontology (aspect-aware inference), trong đó các tham số suy luận như ngưỡng xác suất, số lượng nhãn tối thiểu được giữ lại và hệ số hiệu chỉnh xác suất được điều chỉnh riêng cho từng ontology. Chiến lược này giúp cân bằng giữa độ bao phủ (recall) và độ chính xác (precision) cho từng nhóm chức năng sinh học.

B. Score Calibration

Để điều chỉnh sự khác biệt về phân bố xác suất giữa các ontology, xác suất dự đoán được hiệu chỉnh bằng phép nâng lũy thừa:

$$\tilde{p}_{i,t} = \hat{p}_{i,t}^{\alpha_a}, \quad (5)$$

trong đó α_a là hệ số calibration phụ thuộc vào ontology $a \in \{C, F, P\}$.

C. Thresholding and Top-k Selection

Sau khi hiệu chỉnh xác suất, các GO terms có xác suất lớn hơn một ngưỡng xác định trước được giữ lại. Trong trường hợp số lượng GO terms thỏa mãn ngưỡng nhỏ hơn một giá trị tối thiểu k , các GO terms có xác suất cao nhất sẽ được chọn sao cho đảm bảo mỗi protein có ít nhất k dự đoán.

Chiến lược kết hợp giữa thresholding và top-k selection giúp tránh hiện tượng thiếu dự đoán ở các protein khó, đồng thời hạn chế việc sinh ra quá nhiều nhãn có độ tin cậy thấp.

Tập GO term dự đoán cho protein i được xác định như sau:

$$\mathcal{T}_i = \begin{cases} \{t \mid \tilde{p}_{i,t} \geq \tau_a\}, & \text{nếu } |\cdot| \geq K_a \\ \text{Top-}K_a(\tilde{p}_{i,*}), & \text{ngược lại} \end{cases} \quad (6)$$

trong đó τ_a và K_a là ngưỡng và số lượng tối thiểu tương ứng với ontology a .

D. GO Hierarchy Propagation

Để đảm bảo tính nhất quán với cấu trúc Gene Ontology, điểm số của mỗi GO term tổ tiên được cập nhật theo:

$$\tilde{p}_{i,u} = \max \left(\tilde{p}_{i,u}, \max_{v \in \text{Desc}(u)} \tilde{p}_{i,v} \right), \quad (7)$$

với $\text{Desc}(u)$ là tập các hậu duệ của term u .

E. Negative Propagation from GOA

Bên cạnh lan truyền dương, nhóm sử dụng thông tin annotation âm (negative annotations) từ cơ sở dữ liệu GOA. Các GO terms được gán với qualifier *NOT* cho một protein, cùng với toàn bộ các GO terms con của chúng trong cấu trúc phân cấp, được coi là các cặp protein-GO không hợp lệ.

Các dự đoán trùng với các cặp protein-GO này sẽ bị loại bỏ trong quá trình hậu xử lý. Cách tiếp cận này giúp giảm đáng

kể số lượng false positives và cải thiện độ chính xác tổng thể của hệ thống.

Với tập các cặp protein-GO bị phủ định \mathcal{N} được trích xuất từ GOA, các dự đoán tương ứng được loại bỏ:

$$\tilde{p}_{i,t} = 0 \quad \text{nếu } (i, t) \in \mathcal{N}. \quad (8)$$

F. Ensemble and GOA Blending

Cuối cùng, kết quả dự đoán từ mô hình được kết hợp với submission dựa trên GOA:

$$p_{i,t}^{final} = \max \left(\tilde{p}_{i,t}^{model}, \lambda \cdot \tilde{p}_{i,t}^{GOA} \right), \quad (9)$$

trong đó λ là hệ số trọng số của GOA.

G. Summary

Toàn bộ quy trình suy luận và hậu xử lý được thiết kế nhằm khai thác đồng thời sức mạnh của mô hình học máy, cấu trúc phân cấp của Gene Ontology và kiến thức sinh học có sẵn từ GOA. Sự kết hợp này đóng vai trò quan trọng trong việc cải thiện chất lượng submission cuối cùng cho bài toán dự đoán chức năng protein trong CAFA 6.

VI. RESULTS AND DISCUSSION

Bảng II
SO SÁNH HIỆU NĂNG GIỮA MÔ HÌNH BASELINE

Phiên bản	Mô tả hệ thống	F_{max}
Baseline	Mô hình đơn giản không sử dụng embedding	0.116
V1	MLP với embedding ESM-2	0.232
V2	V1 + GO hierarchy propagation	0.253
V3	V2 + Negative propagation từ GOA	0.275
V4	V3 + Aspect-aware inference + calibration	0.307
V5	V4 + GOA submission blending	0.315

A. Evaluation Metric

Hiệu năng của hệ thống được đánh giá bằng chỉ số F_{max} , được sử dụng chính thức trong thử thách CAFA.

Với một ngưỡng τ , precision và recall được định nghĩa như sau:

$$P(\tau) = \frac{TP(\tau)}{TP(\tau) + FP(\tau)}, \quad R(\tau) = \frac{TP(\tau)}{TP(\tau) + FN(\tau)}. \quad (10)$$

Giá trị F_{max} được tính bằng:

$$F_{max} = \max_{\tau} \frac{2 \cdot P(\tau) \cdot R(\tau)}{P(\tau) + R(\tau)}. \quad (11)$$

B. Baseline Performance

Mô hình baseline (V0) được xây dựng nhằm phản ánh hiệu năng tối thiểu của hệ thống khi không sử dụng embedding tiền huấn luyện và không khai thác cấu trúc Gene Ontology. Trong cấu hình này, mô hình chỉ sử dụng các đặc trưng đơn giản và không áp dụng bất kỳ chiến lược hậu xử lý nào.

Kết quả cho thấy mô hình baseline chỉ đạt điểm F_{max} ở mức thấp, cho thấy việc thiếu thông tin biểu diễn chuỗi protein và kiến thức ontology làm hạn chế khả năng dự đoán chức năng.

C. Impact of Protein Embedding

Việc bổ sung embedding ESM-2 trong phiên bản V1 mang lại cải thiện lớn về hiệu năng so với baseline. Embedding tiền huấn luyện giúp mô hình khai thác được thông tin ngữ cảnh và tiến hóa từ chuỗi protein, từ đó nâng cao khả năng phân biệt giữa các chức năng sinh học khác nhau.

Kết quả này cho thấy embedding đóng vai trò nền tảng trong toàn bộ hệ thống và là điều kiện cần để các bước cải tiến tiếp theo phát huy hiệu quả.

D. Effect of Ontology-aware Post-processing

Các bước hậu xử lý dựa trên Gene Ontology, bao gồm lan truyền theo cấu trúc phân cấp và loại bỏ các annotation không hợp lệ dựa trên dữ liệu GOA, tiếp tục cải thiện hiệu năng của hệ thống. Lan truyền theo ontology giúp đảm bảo tính nhất quán sinh học của các dự đoán, trong khi negative propagation từ GOA giúp giảm số lượng false positives.

Hai bước này mang lại cải thiện ổn định cho điểm F_{max} mà không làm tăng đáng kể số lượng dự đoán dư thừa.

E. Contribution of Aspect-aware Inference and Calibration

Chiến lược suy luận theo ontology, kết hợp với hiệu chỉnh xác suất, đóng vai trò quan trọng trong việc cân bằng giữa precision và recall. Việc sử dụng các tham số suy luận khác nhau cho từng ontology giúp hệ thống thích ứng tốt hơn với đặc điểm phân bố nhăn của mỗi nhóm chức năng sinh học.

So với việc áp dụng một chiến lược suy luận chung, approach này giúp cải thiện độ ổn định và hiệu năng tổng thể của hệ thống.

F. Effect of GOA Submission Blending

Phiên bản cuối cùng (V5) kết hợp thêm submission dựa trên dữ liệu GOA nhằm bổ sung các annotation có độ tin cậy cao đã được kiểm chứng sinh học. Việc blending được thực hiện với trọng số nhỏ để đảm bảo mô hình học máy vẫn đóng vai trò chính.

Kết quả cho thấy GOA blending mang lại cải thiện nhỏ nhưng nhất quán về điểm số, đóng vai trò hỗ trợ hiệu quả cho hệ thống tổng thể.

G. Discussion

Kết quả thực nghiệm cho thấy rằng hiệu năng của hệ thống không chỉ phụ thuộc vào kiến trúc mô hình mà còn chịu ảnh hưởng lớn từ chiến lược suy luận và hậu xử lý. Việc phát triển hệ thống theo hướng từng bước, bắt đầu từ baseline đơn giản và dần tích hợp embedding, kiến thức ontology và dữ liệu sinh học bên ngoài, cho phép đánh giá rõ ràng đóng góp của từng thành phần.

Mặc dù hệ thống đạt được hiệu năng cạnh tranh, vẫn còn những hạn chế như việc sử dụng embedding cố định và mô hình phân loại tương đối đơn giản. Trong tương lai, việc fine-tune embedding hoặc tích hợp thêm các nguồn thông tin sinh học khác có thể tiếp tục cải thiện kết quả.

VII. CONCLUSION

Trong bài báo này, nhóm đã trình bày một hệ thống dự đoán chức năng protein cho bài toán CAFA 6, được xây dựng theo hướng phát triển từng bước từ một mô hình baseline đơn giản đến một hệ thống hoàn chỉnh kết hợp học máy và kiến thức sinh học. Bắt đầu từ mô hình không sử dụng embedding, hệ thống dần được cải thiện thông qua việc tích hợp embedding ESM-2, khai thác cấu trúc phân cấp của Gene Ontology và sử dụng dữ liệu chú giải từ GOA.

Kết quả thực nghiệm cho thấy embedding tiền huấn luyện đóng vai trò nền tảng trong việc cải thiện hiệu năng, trong khi các chiến lược suy luận và hậu xử lý như lan truyền theo ontology, negative propagation và suy luận theo ontology đóng góp đáng kể vào việc cân bằng giữa precision và recall. Việc kết hợp submission dựa trên GOA mang lại cải thiện nhỏ nhưng ổn định, cho thấy giá trị bổ trợ của kiến thức sinh học đã được kiểm chứng.

Cách tiếp cận theo từng bước giúp làm rõ đóng góp của từng thành phần trong hệ thống và nhấn mạnh rằng hiệu năng cao trong CAFA không chỉ đến từ kiến trúc mô hình mà còn phụ thuộc lớn vào chiến lược suy luận và hậu xử lý. Trong tương lai, hệ thống có thể được mở rộng bằng cách fine-tune embedding tiền huấn luyện hoặc tích hợp thêm các nguồn thông tin sinh học khác nhằm tiếp tục cải thiện hiệu quả dự đoán.

TÀI LIỆU

- [1] M. Ashburner *et al.*, “Gene Ontology: Tool for the unification of biology,” *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.
- [2] I. Radivojac *et al.*, “A large-scale evaluation of computational protein function prediction,” *Nature Methods*, vol. 10, no. 3, pp. 221–227, 2013.
- [3] CAFA Consortium, “Critical Assessment of Functional Annotation,” <https://www.biofunctionprediction.org/>, accessed 2024.
- [4] The UniProt Consortium, “UniProt-GOA: Gene Ontology Annotation,” *Nucleic Acids Research*, vol. 45, D1, pp. D331–D338, 2017.
- [5] A. Rives *et al.*, “Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences,” *Proc. Natl. Acad. Sci. USA*, vol. 118, no. 15, 2021.
- [6] Z. Lin, H. Akin, R. Rao, *et al.*, “Evolutionary-scale prediction of atomic-level protein structure with a language model,” *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023.
- [7] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” *Proc. ICML*, 2017.
- [8] T. G. Dietterich, “Ensemble methods in machine learning,” *Proc. MCS*, 2000.