CAFA 6 Protein Function Prediction K66UET_INT3405E2

Nguyễn Lâm Thái
MSSV: 21020533
Khoa Cơ học Kỹ thuật và Tự động hóa
Đại học Công Nghệ – ĐHQGHN

Nguyễn Công Tuấn Phương
MSSV: 21020660
Khoa Công nghệ Thông tin
Đại học Công Nghệ – ĐHQGHN

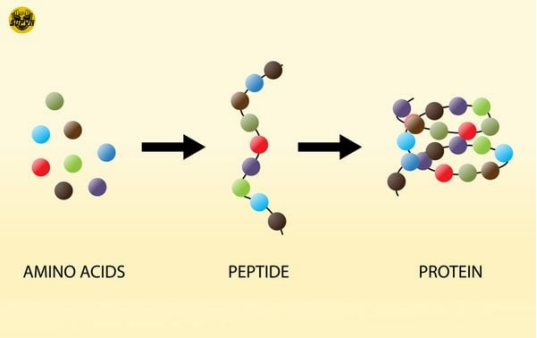
***Tóm tắt nội dung*—Cuộc thi CAFA6 đặt ra bài toán dự đoán chức năng protein trực tiếp từ chuỗi axit amin bằng cách gán các mã Gene Ontology (GO) thuộc ba nhóm MF, BP và CC. Trong báo cáo này, chúng tôi trình bày cách tiếp cận của nhóm K66UET_INT3405E2 dựa trên hai hướng chính: (i) cải thiện biểu diễn dữ liệu từ chuỗi axit amin thông qua nhiều bộ đặc trưng khác nhau, và (ii) xây dựng các mô hình học máy và học sâu bao gồm hồi quy logistic và Multi-Layer Perceptron. Phương pháp của chúng tôi đạt điểm số 0.262 theo thang đo maximum F_1 và xếp hạng 357/1005 vào Ngày 9 tháng 12 năm 2025trên bảng xếp hạng Kaggle. Mã nguồn và các tệp dự đoán được công bố tại https://github.com/lam-thai-nguyen/INT3405E2-CAFA-6-Protein-Function-Prediction.**

I. Giới thiệu

Ngày nay, công nghệ giải trình tự gen tạo ra một lượng lớn chuỗi axit amin, nhưng việc xác định chức năng protein bằng thực nghiệm thủ công lại quá chậm và tốn kém. Các phương pháp truyền thống không dựa trên mô hình, như suy luận dựa trên tương đồng trình tự (Basic Local Alignment Search Tool – BLAST), khai thác văn bản khoa học, hay tổng hợp dữ liệu từ nhiều nguồn sinh học, dù phổ biến, vẫn gặp nhiều hạn chế: chúng phụ thuộc nhiều vào mức độ tương đồng, thất bại khi protein không có họ hàng rõ ràng hoặc khi các protein tương tự thực hiện chức năng khác nhau, và khó nắm bắt bản chất đa chức năng, phi tuyến của hệ sinh học. Trước khối lượng dữ liệu ngày càng tăng, những hạn chế này khiến việc chú thích thủ công trở thành nút thắt cổ chai lớn. Do đó, các phương pháp dựa trên mô hình, đặc biệt là mô hình học máy và học sâu, trở nên cấp thiết để đẩy nhanh quá trình suy luận chức năng, cải thiện độ chính xác và tối ưu hóa ưu tiên thí nghiệm. Chính nhu cầu này là động lực để cuộc thi CAFA Protein Function Prediction ra đời, nhằm tìm ra các mô hình dự đoán chức năng protein trực tiếp từ chuỗi axit amin [2].

Cuộc thi CAFA lần thứ 6 (CAFA6) [1], được tổ chức trên Kaggle[1], thúc đẩy sự phát triển các mô hình học máy và học sâu nhằm dự đoán chức năng sinh học của protein. Như được minh hoạ trong Hình 1, axit amin, peptit và protein có mối quan hệ cấu trúc chặt chẽ: axit amin là đơn vị cơ bản, được ký hiệu bằng mã một hoặc ba chữ cái để biểu diễn trình tự một cách ngắn gọn; các axit amin liên kết với nhau tạo thành peptit; và các chuỗi peptit dài gấp cuộn không gian sẽ tạo nên protein thực hiện nhiều chức năng quan trọng của cơ thể.

[1]https://www.kaggle.com/competitions/cafa-6-protein-function-prediction/



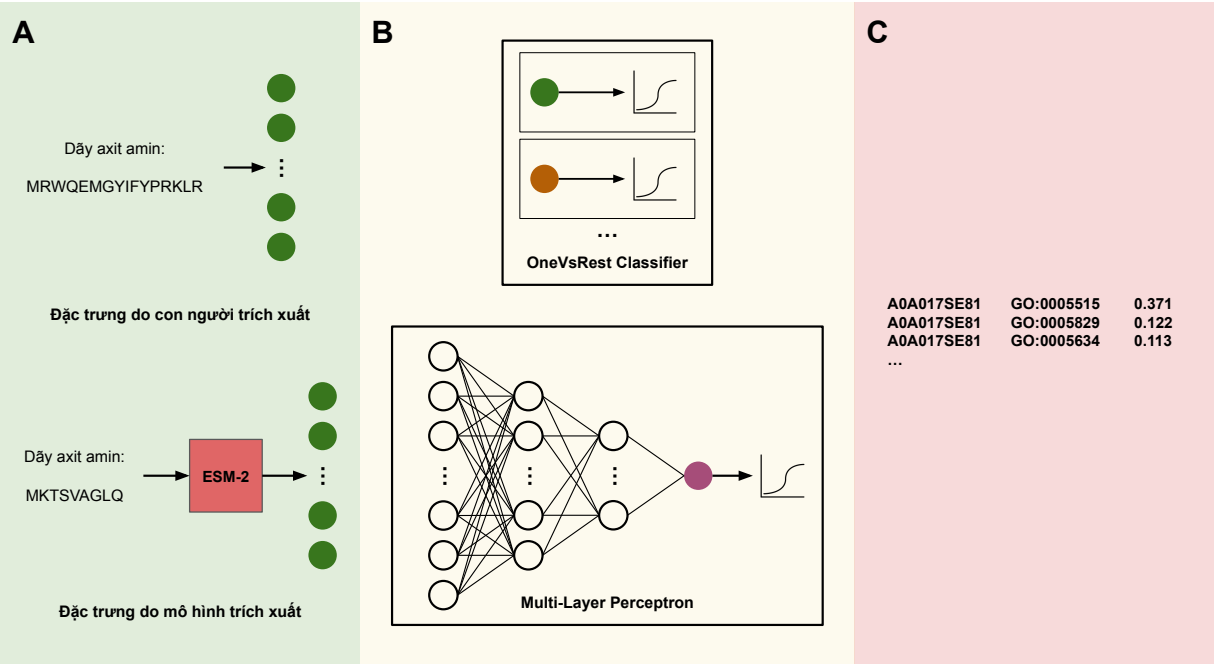
Hình 1. Axit amin, peptit, và protein. Nguồn: https://supvn.net/blogs/kien-thuc/amino-acid-la-gi.

Trong CAFA6, chức năng của protein được mô tả thông qua hệ thống Gene Ontology (GO), trong đó mỗi chức năng được biểu diễn bằng một mã GO chuẩn hoá. Các mã này thuộc ba nhóm chính: Molecular Function (MF) mô tả hoạt động ở mức phân tử, Biological Process (BP) mô tả quá trình sinh học mà protein tham gia, và Cellular Component (CC) mô tả vị trí của protein trong tế bào. Đáng chú ý, một protein có thể đồng thời mang nhiều chức năng thuộc nhiều nhóm khác nhau, phản ánh bản chất đa nhiệm của hệ sinh học. Nhờ hệ thống phân loại này, chức năng protein được biểu diễn một cách có cấu trúc và tạo điều kiện thuận lợi cho việc dự đoán bằng các mô hình tính toán hiện đại.

Trong báo cáo này, chúng tôi trình bày phương pháp dự đoán vai trò sinh học của protein trong cuộc thi CAFA6 của nhóm K66UET_INT3405E2. Trước tiên, chúng tôi phân tích dữ liệu để hiểu bản chất vấn đề trong CAFA6. Sau đó, chúng tôi đưa ra mô hình cơ sở và độ chính xác cơ sở. Cuối cùng, như được minh họa ở Hình 2, chúng tôi đề xuất các phương pháp cải thiện độ chính xác dựa trên (i) dữ liệu, và (ii) mô hình. Nhóm K66UET_INT3405E2 đã đạt được điểm 0.262 và xếp hạng 357/1005 trong cuộc thi CAFA6 vào Ngày 9 tháng 12 năm 2025. Các đoạn mã và file dữ đoán được công bố tại https://github.com/lam-thai-nguyen/INT3405E2-CAFA-6-Protein-Function-Prediction.

II. Phương pháp luận

a) Phương pháp đề xuất: Như được minh họa trên Hình 2, chúng tôi nghiên cứu và cải thiện độ chính xác của



Hình 2. Tổng quan phương pháp đề xuất. (A) Dữ liệu, (B) mô hình, (C) dự đoán.

phương pháp dự đoán chức năng của protein dựa trên (i) dữ liệu, và (ii) mô hình. Cụ thể, chúng tôi sử dụng hai cách biểu diễn chuỗi axit amin đầu vào dựa trên (i) đặc trưng do con người trích xuất, và (ii) đặc trưng do mô hình trích xuất. Véc-tơ biểu diễn với độ dài từ 20 đến 320 giá trị được thí nghiệm và đánh giá về hiệu quả. Đối với mô hình, chúng tôi sử dụng mô hình học máy hồi quy logistic (Logistic Regression) và mạng nơ-ron thần kinh nhân tạo (Multi-Layer Perceptron – MLP). Các thay đổi về mặt cấu hình mô hình được thử nghiệm và đánh giá về hiệu quả.

b) Dữ liệu: Như được mô tả chi tiết ở [1], dữ liệu huấn luyện bao gồm protein dưới dạng chuỗi axit amin, và nhãn gán là mã GO tương ứng. Cụ thể, các protein trong dữ liệu huấn luyện đến từ sinh vật nhân chuẩn (eukaryotes) và một vài loài không phải nhân chuẩn. Với bài toán phân loại đa lớp (multi-class classification) của CAFA6, đối với mỗi một protein, chúng tôi cần dự đoán các mã GO tương ứng. Phân tích cụ thể dữ liệu của CAFA6 được bàn luận ở Mục III.

c) Thông số đánh giá: Thông số maximum F_1 dựa trên Precision và Recall có trọng số được sử dụng để đánh giá độ chính xác của phương pháp. Thông số này được đánh giá cho riêng từng nhóm trong ba nhóm (MF, BP, CC) và giá trị trung bình cộng được sử dụng làm độ chính xác cuối cùng. Công thức của maximum F_1 được đề cập trong [3].

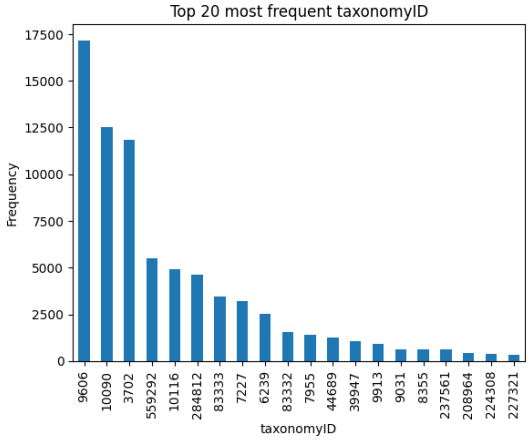
III. Phân tích dữ liệu

Trước tiên, chúng tôi tóm tắt các files được CAFA6 công bố như sau.

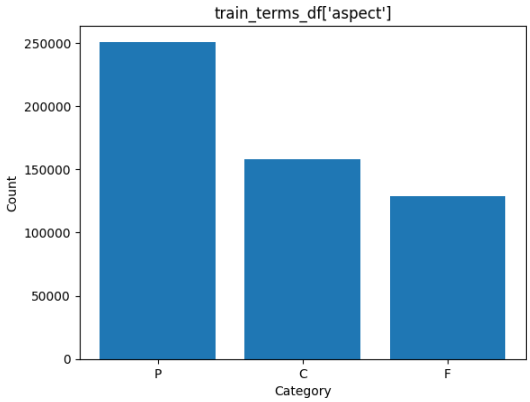
- **train_sequences.fasta.** Trong CAFA6, mỗi một protein được biểu diễn dưới dạng một dãy các axit amin, và đi kèm là một tên gọi (đó là EntryID). File này chứa thông tin về EntryID của các protein huấn luyện và dãy axit amin tương ứng.
- **testsuperset.fasta.** Tương tự như train_sequences.fasta, file này chứa thông tin về EntryID của các protein kiểm thử và dãy axit amin tương ứng.
- **train_terms.tsv.** File này chứa thông tin về EntryID, mã GO, và nhóm (MF, BP, CC) tương ứng.
- **train_taxonomy.tsv.** File này chứa thông tin về EntryID và giống (taxonomyID) tương ứng.

Chúng tôi tiến hành gộp thông tin từ các files trên thành một bảng thống nhất, và trích xuất được các thông tin về dữ liệu như sau: (i) tập dữ liệu huấn luyện gồm 82,404 ví dụ (protein hay chuỗi axit amin); (ii) tập dữ liệu kiểm thử gồm 224,309 ví dụ; (iii) 537,027 mã GO được gán nhãn cho dữ liệu huấn luyện, trong đó có 26,125 mã GO độc nhất (mỗi mã GO là một lớp trong bài toán multi-class classification); và (iv) trong số 26,125 mã GO, MF, BP, CC chiếm lần lượt 6,616, 16,858, và 2,651 mã, lần lượt.

Bên cạnh các thông tin có tính tổng quan trên, chúng tôi còn phân tích được các đặc điểm đặc thù trong tập dữ liệu huấn luyện CAFA6. Đầu tiên, như được minh họa trong Hình 3, có một số taxonomyID như 9606 chiếm tỷ lệ xuất hiện lớn. Cụ thể, 20 trên 1,381 taxonomyIDs xuất hiện nhiều nhất chiếm tới 91.05% toàn bộ tập dữ liệu huấn luyện. Điều này cho thấy tập dữ liệu tập trung xoay quanh protein của một số loài sinh vật nhất định. Thứ hai, 700 trên 2,906 độ dài độc nhất chuỗi axit amin chiếm tới 81.25% tập dữ liệu, và rơi vào khoảng 10 đến 926 ký tự. Điều này cho thấy độ dài chuỗi axit amin tập trung trong khoảng này. Thứ ba, như được minh họa trên Hình 4, trong 527,027 mã GO được gán nhãn, BP là nhóm GO được



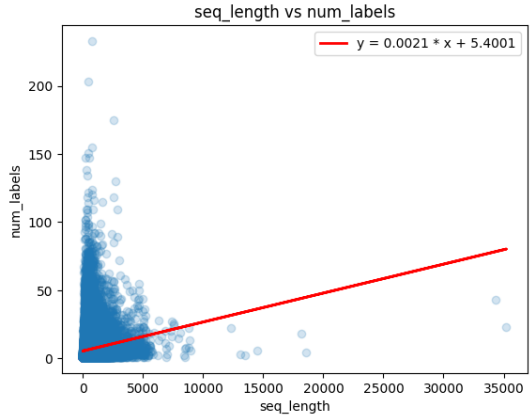
Hình 3. Tần suất của protein theo taxonomyID.



Hình 4. Tần suất của các nhóm GO.

gán nhãn nhiều nhất (250,805), sau đó là CC (157,770) và MF (128,452). Mặc dù MF có tới 16,858 mã GO độc nhất, nó lại được gán nhãn ít nhất, cho thấy các mã GO thuộc nhóm này có xu hướng thưa thớt về mật độ, có tiềm năng gây khó khăn trong việc huấn luyện mô hình. Thứ tư, như được minh họa trong Hình 5, có mối tương quan dương yếu giữa chiều dài chuỗi axit amin và số lượng mã GO. Điều này có nghĩa các protein với độ dài chuỗi axit amin càng dài sẽ có khả năng có nhiều mã GO. Cuối cùng, như được minh họa trên Hình 6, một vài taxonomyIDs như 2592315 có số lượng mã GO trung bình nhiều hơn các loài khác.

Tuy nhiên, các phân tích trên chưa khai thác hết thông tin về bản chất protein và mối liên hệ vị trí giữa các axit amin trong một chuỗi. Cụ thể, các ký tự có thể đại diện cho một axit amin (mã một chữ cái) là: A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V. Vị trí tương đối của các axit amin trong một chuỗi peptit có vai trò quan trọng vì nó quyết định cách chuỗi gấp lại thành cấu trúc ba chiều, và cấu trúc này lại ảnh hưởng trực tiếp đến chức năng của protein. Cùng một tập hợp axit amin nhưng sắp xếp khác nhau có thể dẫn đến các vùng hoạt động, bề mặt liên kết hoặc tính ổn định hoàn toàn khác. Do đó, không chỉ loại axit amin mà còn thứ



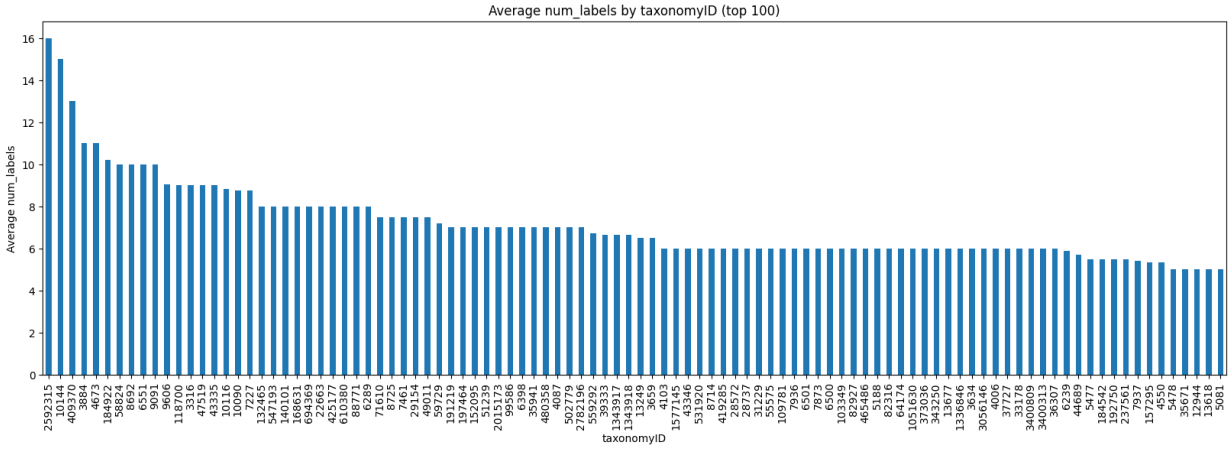
Hình 5. Sự tương quan giữa chiều dài chuỗi axit amin và số lượng mã GO.

Bảng I
CÁC PHƯƠNG PHÁP TRÍCH XUẤT ĐẶC TRƯNG TỪ DÃY AXIT AMIN.

Phương pháp	Ý nghĩa	Độ dài
AAC	Tính tần suất tương đối của 20 axit amin chuẩn, phản ánh thành phần hóa học tổng thể của protein	20
Physicochemical Properties	Đo các tính chất hóa lý toàn cục có liên quan đến cấu trúc và chức năng, gồm: tỷ lệ kị nước, tỷ lệ mang điện, khối lượng phân tử, và độ dài chuỗi	4
CTD-style Group Composition	Nhóm các axit amin theo tính chất chung và đo tỷ lệ của từng nhóm: kị nước, phân cực, tích điện dương/âm, thơm, aliphatic, hoặc kích thước nhỏ	7
k-mer Frequencies	Đếm tần suất các mẫu con ngắn (2-mer, 3-mer); nắm bắt các motif cục bộ mà biểu diễn theo thành phần không thể hiện được	40
Secondary Structure Fractions	Ước lượng tỉ lệ các loại cấu trúc thứ cấp: alpha-helix, beta-sheet, coil; cung cấp ngữ cảnh cấu trúc có ảnh hưởng đến chức năng protein	3
Global Physicochemical Descriptors	Bao gồm các chỉ số hóa lý ở cấp độ toàn chuỗi như pI, GRAVY, hay chỉ số bất ổn; giúp mô tả đặc tính sinh hóa ảnh hưởng đến độ ổn định và tương tác	3

tự và vị trí của chúng trong chuỗi mới quyết định protein thực sự làm được gì trong tế bào.

Trong thực tế, thông tin vị trí tương đối của axit amin đóng vai trò quan trọng, và được khai thác triệt để để mang lại các đặc trưng có ý nghĩa. Cụ thể, các phương pháp như (i) kết cấu axit amin (Amino Acid Composition – AAC), (ii) tính chất hóa lý (Physicochemical Properties), (iii) kết cấu theo dạng CTD (CTD-style Group Composition), (iv) tần suất k-mer (k-mer Frequencies), (v) tỷ lệ cấu trúc bậc hai (Secondary Structure Fractions), và (vi) mô tả hóa lý toàn chuỗi (Global Physicochemical Descriptors). Các phương pháp trích xuất đặc trưng này được tóm tắt ở Bảng I.



Hình 6. Số lượng mã GO trung bình của các taxonomyIDs.

IV. Kết quả

A. Độ chính xác cơ sở

Chúng tôi tiến hành thiết lập độ chính xác cơ sở dựa trên phương pháp trích xuất đặc trưng và mô hình đơn giản như sau.

- **Trích xuất đặc trưng.** Phương pháp AAC được sử dụng để trích xuất một véc-tơ biểu diễn gồm 20 giá trị từ chuỗi axit amin. Phương pháp này tạo ra một ma trận dữ liệu đầu vào với kích cỡ $(82404, 20)$.
- **Mô hình.** Chúng tôi lựa chọn mô hình Logistic Regression và huấn luyện theo phương pháp một-với-tất cả (One Vs. Rest). Điều này có nghĩa là với mỗi một lớp (hay một mã GO), chúng tôi sẽ huấn luyện một mô hình Logistic Regression cho bài toán phân loại nhị phân (binary classification), dẫn tới tổng cộng 26,125 mô hình được huấn luyện cho tất cả mã GO. Từ đây, chúng tôi sẽ gọi mô hình này là OneVsRest Logistic.
- **Mã hóa nhãn (label encoding).** Chúng tôi chia bài toán multi-class classification với 26,125 lớp thành ba bài toán con tương ứng với ba nhóm MF, BP, và CC. Cụ thể, đối với từng nhóm, mỗi nhãn huấn luyện sẽ được mã hóa sử dụng phương pháp multi-hot encoding. Do đó, ma trận nhãn của MF, BP, và CC lần lượt có kích cỡ $(82404, 6616)$, $(82404, 16858)$, và $(82404, 2651)$.

Với cách tiếp cận trên, chúng tôi huấn luyện ba mô hình OneVsRest Logistic cho ba tập mã GO. Kết quả dự đoán cuối cùng là danh sách dự đoán của cả ba mô hình này. Trong thí nghiệm này, mỗi mô hình được huấn luyện trong 100 chu trình, hệ số điều tiết $C = 0.5$, và bộ tối ưu BFGS trong thư viện scikit-learn[2]. Sau khi giới hạn ngưỡng xác suất ở 0.02 cho các dự đoán, cũng như giới hạn 1,500 dự đoán đối với mỗi protein, như thể hiện trong Bảng II, chúng tôi thu được 3,902,841 dự đoán, và đạt số điểm 0.104.

[2]https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

B. Giai đoạn 1: Dữ liệu

Ở giai đoạn đầu tiên, chúng tôi tập trung cải thiện độ chính xác của mô hình dựa trên dữ liệu. Như minh họa ở Hình 2, đặc trưng do con người trích xuất và đặc trưng do mô hình trích xuất sẽ được sử dụng để làm đầu vào cho mô hình.

a) Đặc trưng do con người trích xuất: So với thí nghiệm cơ sở chỉ sử dụng 20 giá trị để biểu diễn chuỗi axit amin, chúng tôi tiến hành sử dụng thêm các phương pháp trích xuất đặc trưng được trình bày trong Bảng I. Cụ thể, ở thí nghiệm đầu tiên, bên cạnh AAC, chúng tôi sử dụng thêm các phương pháp Physicochemical Properties, CTD-style Group Composition, k-mer Frequencies, để trích xuất ra véc-tơ biểu diễn với độ dài 71. Việc tăng thêm số lượng đặc trưng yêu cầu chúng tôi huấn luyện trong 500 chu trình, so với 100 chu trình ở thí nghiệm cơ sở. Các lựa chọn mô hình, thông số không được đề cập được giữ nguyên so với thí nghiệm cơ sở. Như thể hiện trong Bảng II, chúng tôi thu được 4,192,667 dự đoán, và đạt số điểm 0.127, cải thiện hơn độ chính xác cơ sở 0.023 điểm.

Tương tự, ở thí nghiệm thứ hai, chúng tôi sử dụng thêm hai phương pháp Secondary Structure Fractions và Global Physicochemical Descriptors, trích xuất được 77 giá trị để biểu diễn chuỗi axit amin. 600 chu trình được sử dụng để huấn luyện mô hình do số lượng đặc trưng tăng lên. Như thể hiện trong Bảng II, chúng tôi thu được 4,758,132 dự đoán, và đạt số điểm 0.132, cải thiện hơn độ chính xác cơ sở 0.028 điểm.

Hai thí nghiệm trên cho thấy sự hiệu quả của các phương pháp trích xuất đặc trưng trong Bảng I, khi số lượng dự đoán lẫn độ chính xác đều được cải thiện. Tuy nhiên, các phương pháp này đòi hỏi kiến thức chuyên sâu về sinh học, và số lượng đặc trưng sẽ bị giới hạn theo đó. Do đó, chúng tôi đề xuất sử dụng các mô hình ngôn ngữ protein (Protein Language Models – PLM) như ESM-2 [4] để trích xuất ra được nhiều đặc trưng hơn, cũng như không yêu cầu kiến thức chuyên sâu về sinh học.

b) Đặc trưng do mô hình trích xuất: Ở thí nghiệm thứ ba, chúng tôi sử dụng mô hình ESM-2 để trích xuất ra 320 đặc trưng từ chuỗi axit amin. Tuy nhiên, việc sử dụng 320

Bảng II
ĐỘ CHÍNH XÁC CỦA CÁC PHƯƠNG PHÁP.

Thứ tự	Mô hình	Độ phức tạp tương đối	Đặc trưng	Dự đoán	Điểm	Ghi chú
0	OneVsRest Logistic	★	20	3,902,841	0.104	Cơ sở
1	OneVsRest Logistic	★	71	4,192,667	0.127 (↑ 0.023)	Cải thiện trích xuất đặc trưng
2	OneVsRest Logistic	★	77	4,758,132	0.132 (↑ 0.028)	Cải thiện trích xuất đặc trưng
3	OneVsRest Logistic	★	100	5,214,598	0.230 (↑ 0.126)	Cải thiện trích xuất đặc trưng
4	MLP	★	20	3,738,561	0.096 (↓ 0.008)	Cải thiện mô hình
5	MLP	★★	20	4,033,311	0.091 (↓ 0.013)	Cải thiện mô hình
6	MLP	★	320	4,164,029	0.112 (↑ 0.008)	Cải thiện mô hình
7	MLP	★★★	320	5,784,652	**0.262** (↑ 0.158)	Cải thiện mô hình
8	OneVsRest Logistic + MLP	★★★	320	**10,196,836**	0.243 (↑ 0.139)	Quần thể mô hình
9	MLP + CNN1D	★★★	320	7,796,023	0.214 (↑ 0.110)	Quần thể mô hình

đặc trưng để huấn luyện mô hình OneVsRest Logistic sẽ tăng thời gian hơn đáng kể so với các phương pháp trước. Do đó, chúng tôi sử dụng phương pháp phân tích thành phần chính (Principal Component Analysis – PCA) để giảm xuống 100 đặc trưng. Kết quả cho thấy chúng tôi đạt được 5,214,598 dự đoán, với số điểm 0.230, cải thiện 0.126 so với số điểm cơ sở. Đây là cải thiện có tính đột phá trong giai đoạn 1, cho thấy các đặc trưng trích xuất bởi PLM có tính toàn diện và hiệu quả hơn các phương pháp trước đó.

Ở giai đoạn 1, chúng tôi quan sát thấy việc tăng số lượng đặc trưng liên tục mang lại cải thiện về số lượng dự đoán và độ chính xác. Tuy nhiên, đi kèm với sự cải thiện này là thời gian huấn luyện mô hình tăng lên đáng kể. Do đó, chúng tôi đề xuất cải thiện mô hình để có thể tận dụng nhiều đặc trưng hơn.

C. Giai đoạn 2: Mô hình

Ở giai đoạn thứ hai, như minh họa ở Hình 2, chúng tôi tập trung thí nghiệm với mô hình học sâu MLP để cải thiện độ chính xác dự đoán. Lý do chúng tôi lựa chọn MLP là vì mô hình này có thể giải quyết được nút thắt cổ chai của OneVsRest Logistic, đó là việc phải huấn luyện 26,125 mô hình cho mỗi một mã GO, trong đó mỗi mô hình được hồi quy trên 82,404 protein. Việc sử dụng MLP sẽ giúp việc huấn luyện nhanh hơn vì trong một lần dự đoán (forward pass), MLP có thể đưa ra dự đoán cho nhiều mã GO cùng lúc.

Như thể hiện trên Bảng II, ở thí nghiệm thứ 4, chúng tôi sử dụng MLP trên 20 đặc trưng của thí nghiệm cơ sở. Cấu hình MLP được sử dụng trong thí nghiệm này được tóm tắt ở Bảng III, trong đó d_{out} lần lượt bằng 16,858, 2,651, và 6,616 cho nhóm BP, CC, và MF. Mô hình có ba lớp, xen kẽ trong đó là các lớp BatchNorm [5] và Dropout [6] giúp ổn định quá trình huấn luyện và tăng khả năng tổng quát hóa (generalization). Chúng tôi sử dụng hàm mất mát BCEWithLogitsLoss[3] và bộ tối ưu Adam[4] trong Pytorch, và huấn luyện ba mô hình cho ba nhóm BP, CC, MF trong 8 chu trình với tốc độ học (learning rate) là 10^{-3}. Kết quả cho thấy chúng tôi đạt được 3,738,561 dự đoán, với số điểm 0.096. Cả hai thông số này đều giảm nhẹ so với thí nghiệm cơ sở.

Bảng III
CẤU HÌNH MLP Ở THÍ NGHIỆM 4.

Layer	Output size
Input	20
Linear + ReLU + BatchNorm + Dropout(0.3)	64
Linear + ReLU + Dropout(0.2)	32
Linear (output)	d_{out}

Bảng IV
CẤU HÌNH MLP Ở THÍ NGHIỆM 5.

Layer	Output size
Input	20
Linear + ReLU + BatchNorm + Dropout(0.35)	256
Linear + ReLU + BatchNorm + Dropout(0.3)	128
Linear + ReLU + Dropout(0.25)	64
Linear + ReLU + Dropout(0.2)	32
Linear (output)	d_{out}

Ở thí nghiệm thứ 5, chúng tôi giữ nguyên các thiết kế thực nghiệm ở thí nghiệm thứ 4, và tăng độ phức tạp của mô hình MLP, như được thể hiện ở Bảng IV. Kết quả cho thấy mô hình đạt được 4,033,311 dự đoán và 0.091 điểm, kém hơn điểm cơ sở là 0.013.

Hai thí nghiệm trên cho thấy khó khăn khi sử dụng mô hình học sâu, đó là MLP cần thông tin (hay đặc trưng) đủ nhiều và có thể khai thác được để vượt qua mô hình học máy như OneVsRest Logistic. 20 đặc trưng AAC được sử dụng chỉ chứa thông tin về tần suất xuất hiện của các ký tự axit amin, do đó

[3] docs.pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html

[4] docs.pytorch.org/docs/stable/generated/torch.optim.Adam.html

Bảng V
CẤU HÌNH MLP Ở THÍ NGHIỆM 6.

Layer	Output size
Input	320
Linear + ReLU + BatchNorm + Dropout(0.3)	64
Linear + ReLU + Dropout(0.2)	32
Linear (output)	d_{out}

Bảng VI
CẤU HÌNH MLP Ở THÍ NGHIỆM 7.

Layer	Output size
Input	320
Linear + ReLU + BatchNorm + Dropout(0.3)	2048
Linear + ReLU + Dropout(0.25)	1024
Linear (output)	d_{out}

làm mất đi các thông tin có thể khai thác triệt để như vị trí tương đối của các axit amin. Bên cạnh đó, hai mô hình MLP ở Bảng III và IV có độ phức tạp cao, sử dụng nhiều tham số, thêm vào nhiều tính phi tuyến cho một đầu vào đơn giản. Do đó, mô hình MLP đã không thể tận dụng được lợi thế để vượt qua OneVsRest Logistic. Phát hiện này thúc đẩy chúng tôi sử dụng đầu vào có nhiều thông tin hơn, cũng như một mô hình có độ phức tạp hợp lý.

Ở các thí nghiệm thứ 6, 7 trong Bảng II, 320 đặc trưng trích xuất bởi mô hình ESM-2 đã được sử dụng làm đầu vào cho MLP. Cấu hình của MLP trong hai thí nghiệm này được mô tả ở Bảng V và VI. Với lượng thông tin nhiều hơn, mô hình thí nghiệm 6 lập tức cải thiện được số dự đoán (4,164,029) và độ chính xác (0.112) so với thí nghiệm cơ sở. Việc sử dụng mô hình phức tạp hơn ở thí nghiệm 7 cho thấy sức mạnh của MLP đã được tận dụng triệt để, mang lại cải thiện có tính đột phá. Cụ thể, mô hình MLP đạt được 5,784,652 dự đoán và số điểm 0.262.

Qua các thí nghiệm trên, có thể thấy sự cải thiện về mô hình mang lại hiệu quả tốt hơn so với cải thiện về dữ liệu. Tuy nhiên, chúng tôi nhận thấy việc lựa chọn kiến trúc mô hình phù hợp cũng lại một trở ngại lớn, so với giai đoạn 1. Chúng tôi nhận thấy véc-tơ biểu diễn ESM-2 đã chứa rất nhiều thông tin, nên không cần một mô hình MLP quá sâu, dễ dẫn đến quá khớp (overfitting). Một mô hình rộng (nhiều tham số ở mỗi lớp) sẽ là lựa chọn tốt hơn, vì chúng ta cần nhiều sự kết hợp giữa các đặc trưng vốn có rất nhiều thông tin của ESM-2.

D. Các trường hợp thất bại

Các trường hợp thất bại trong báo cáo này là một phần trong giai đoạn 3, đó là cải thiện độ chính xác dự đoán dựa trên quần thể mô hình. Như minh họa trên Bảng II, hai thí nghiệm thứ 8, 9 được mô tả như sau.

Thí nghiệm 8 kết hợp kết quả của hai mô hình tốt nhất ở hai giai đoạn, đó là mô hình trong thí nghiệm 3, và mô hình trong thí nghiệm 7. Trọng số quần thể được sử dụng là 0.53 cho mô hình MLP và 0.47 cho mô hình OneVsRest Logistic. Trọng số này được tính dựa trên số điểm của hai mô hình (đó là 0.262 và 0.230). Sau khi kết hợp kết quả của hai mô hình, phương pháp này đạt được 10,196,836 dự đoán, với số điểm 0.243. Số điểm này thấp hơn số điểm chúng tôi kỳ vọng (đó là 0.262 điểm của thí nghiệm 7) có lẽ vì hai mô hình có độ hiệu chỉnh xác suất khác nhau, và việc trung bình hóa có trọng số đã làm suy yếu các dự đoán đúng mà mô hình MLP vốn tự tin, đồng thời khiến nhiều điểm số rơi xuống dưới ngưỡng dự đoán chính xác.

Thí nghiệm 9 kết hợp kết quả của mô hình MLP trong thí nghiệm 7 và mô hình mới, đó là CNN1D. CNN1D là một mạng nơ-ron thần kinh tích chấp với bộ lọc một chiều. Ý tưởng của thí nghiệm này là kết hợp khả năng phân tích dữ liệu toàn cục của MLP và khả năng phân tích dữ liệu cục bộ của CNN1D. Tuy nhiên, điểm số ban đầu chưa đạt kỳ vọng của chúng tôi (đó là 0.262 điểm của thí nghiệm 7), do đó cần thêm các phân tích chi tiết trong tương lai.

V. KẾT LUẬN

Trong báo cáo này, chúng tôi đã trình bày quy trình xây dựng hệ thống dự đoán chức năng protein cho cuộc thi CAFA6, bao gồm phân tích dữ liệu, xây dựng mô hình cơ sở và đề xuất các hướng cải thiện dựa trên dữ liệu và mô hình. Kết quả phân tích cho thấy tập dữ liệu có cấu trúc phức tạp với số lượng lớn lớp GO, sự mất cân bằng mạnh giữa các nhóm chức năng, và sự lệch phân bố theo taxonomyID. Những đặc điểm này đòi hỏi các phương pháp biểu diễn chuỗi axit amin giàu thông tin hơn cũng như các mô hình đủ khả năng học quan hệ phi tuyến giữa trình tự và chức năng sinh học.

Dựa trên những quan sát đó, chúng tôi khai thác cả đặc trưng do con người trích xuất và đặc trưng do mô hình tạo ra. Các đặc trưng như AAC, tính chất hóa lý, CTD, k-mer và cấu trúc bậc hai giúp bổ sung thông tin tổng quát về thành phần và ngữ cảnh cục bộ của chuỗi, trong khi các véc-tơ biểu diễn do mô hình học sâu tạo ra giúp mô tả mối quan hệ phức tạp hơn. Chúng tôi đồng thời đánh giá hai hướng mô hình là hồi quy logistic theo chiến lược một-với-tất cả và mạng nơ-ron MLP. Những cải thiện theo từng giai đoạn đã nâng hiệu suất hệ thống vượt mức mô hình cơ sở ban đầu.

Cuối cùng, hệ thống của nhóm đạt điểm số 0.262 và đứng ở hạng 357/1005 trên bảng xếp hạng CAFA6 vào Ngày 9 tháng 12 năm 2025. Kết quả này cho thấy ngay cả với các mô hình tương đối đơn giản, việc khai thác hợp lý đặc trưng chuỗi và lựa chọn cấu hình mô hình phù hợp vẫn mang lại hiệu quả đáng kể. Trong tương lai, các hướng mở bao gồm khai thác đầy đủ véc-tơ biểu diễn từ các mô hình ngôn ngữ sinh học hiện đại, và cải thiện xử lý mất cân bằng nhãn.

TUYÊN BỐ ĐÓNG GÓP CỦA TÁC GIẢ

Nguyễn Lâm Thái: Lên ý tưởng, Phương pháp luận, Phân tích dữ liệu, Viết báo cáo. **Nguyễn Công Tuấn Phương:** Huấn luyện mô hình, Viết báo cáo.

TÀI LIỆU

[1] I. Friedberg, P. Radivojac, P. D. Thomas, A. Phan, M. C. D. P. Kaluza, D. Piovesan, P. Joshi, C. Mungall, M. Plomecka, W. Reade, and M. Cruz, "CAFA 6 Protein Function Prediction," Kaggle Competition, 2025. [Online]. Available: https://kaggle.com/competitions/cafa-6-protein-function-prediction

[2] I. Friedberg, P. Radivojac, C. De Paolis, D. Piovesan, P. Joshi, W. Reade, and A. Howard, "CAFA 5 Protein Function Prediction," Kaggle Competition, 2023. [Online]. Available: https://kaggle.com/competitions/cafa-5-protein-function-prediction

[3] Y. Jiang, T. Oron, W. Clark, *et al.*, "An expanded evaluation of protein function prediction methods shows an improvement in accuracy," *Genome Biology*, vol. 17, p. 184, 2016. [Online]. Available: https://doi.org/10.1186/s13059-016-1037-6

[4] 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. [Online]. Available: https://doi.org/10.1126/science.ade2574

[5] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015. [Online]. Available: https://arxiv.org/abs/1502.03167

[6] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014. [Online]. Available: https://jmlr.org/papers/v15/s