Tạp chí Khoa học Đại học Huế: Kỹ thuật và Công nghệ; pISSN 2588-1175 | eISSN 2615-9732

Tập 134, Số 2A, 2025, Tr. 27–46; DOI: 10.26459/hueunijtt.v134i2A.8152

TỐI ƯU HÓA DỰ ĐOÁN TƯƠNG TÁC PROTEIN-PROTEIN TỪ BIỂU DIỄN NGÔN NGỮ THÔNG QUA CƠ CHẾ CHỌN LỌC ĐẶC TRƯNG ĐA GIAI ĐOẠN VÀ HỌC MÁY XẾP CHỒNG

Mai Xuân Văn[1], Trương Khánh Duy[2], Trương Thị Hạnh[3], Trần Tiến Đạt[2], Nguyễn Ngọc Nhớ[2], Nguyễn Tương Tri [1,2]*

[1] Trường Đại học Sư Phạm, Đại học Huế, Huế, Việt Nam
[2] Viện Đào tạo mở và CNTT - Đại học Huế, Huế, Việt Nam
[3] Trường THPT chuyên Quốc Học Huế, Huế, Việt Nam

Tóm tắt. Tương tác protein–protein (PPI) là nền tảng của nhiều hoạt động sinh học bên trong tế bào, và việc dự đoán PPI trực tiếp từ chuỗi axit amin vẫn đang là một hướng nghiên cứu cốt lõi trong sinh học tính toán. Sự ra đời của các mô hình ngôn ngữ protein thế hệ mới như ESM-2 cho phép tạo ra các biểu diễn chuỗi giàu thông tin tiến hóa và tín hiệu cấu trúc tiềm ẩn. Tuy nhiên, các biểu diễn này thường sở hữu số chiều rất lớn với độ nhiễu và tính tương quan nội tại cao. Điều này gây trở ngại cho các mô hình học máy truyền thống trong việc khai thác đặc trưng và dễ rơi vào tình trạng quá khớp. Thách thức này đòi hỏi một phương pháp tiếp cận có khả năng sàng lọc tri thức, loại bỏ sự dư thừa dữ liệu trong khi vẫn bảo toàn các tín hiệu sinh học cốt lõi. Trong công trình này, chúng tôi đề xuất E–StackPPI (Embedding-Stacking Protein-Protein Interaction prediction framework), một khung dự đoán PPI sử dụng hoàn toàn biểu diễn nhúng, trong đó trọng tâm là cơ chế chọn lọc đặc trưng theo tầng, gồm ba bước được áp dụng trực tiếp lên biểu diễn nhúng được tổng hợp từ lớp ẩn cuối cùng của mô hình ESM-2 650M. Cụ thể, (1) quy trình lần lượt loại bỏ các chiều có phương sai thấp; (2) giữ lại các chiều có khả năng phân biệt cao dựa trên độ quan trọng đặc trưng theo LightGBM; (3) loại trừ các chiều có tương quan Pearson lớn nhằm giảm trùng lặp thông tin. Phần đặc trưng đã qua sàng lọc được đưa vào kiến trúc xếp tầng, trong đó hai nhánh LightGBM chạy song song và cuối cùng được hợp nhất ở tầng quyết định bằng hồi quy logistic (Logistic Regression – LR). Thử nghiệm trên hai bộ dữ liệu chuẩn của cơ sở dữ liệu DIP gồm DIP–Yeast và DIP–Human cho thấy E–StackPPI đạt hiệu năng ấn tượng và ổn định trên các chỉ số quan trọng bao gồm độ chính xác, hệ số MCC, cũng như các chỉ số ROC-AUC và PR-AUC. Khi đối chiếu với 12 phương pháp tiên tiến được tổng hợp trong nghiên cứu của Li và cộng sự, mô hình của chúng tôi thể hiện hiệu năng cạnh tranh trên cả hai bộ dữ liệu. Những kết quả này nhấn mạnh vai trò thiết yếu của cơ chế chọn lọc đặc trưng theo tầng trong việc giảm nhiễu và khai thác hiệu quả các biểu diễn nhúng PLM có số chiều rất lớn, qua đó mở ra một hướng tiếp cận khả thi và tiềm năng cho bài toán dự đoán PPI chỉ dựa trên thông tin chuỗi mà không cần bổ sung dữ liệu cấu trúc.

Từ khóa: Tương tác protein protein, Chọn lọc đặc trưng xếp tầng, Mô hình Ngôn ngữ Protein, Mô hình xếp tầng

** Liên hệ: ntuongtri@hueuni.edu.vn*

Ngày nhận bài: 12–11–2025; Ngày nhận đăng: 29–12–2025

Optimizing Protein-Protein Interaction Prediction from Language Representations via Multi-stage Feature Selection and Stacking Ensemble Learning

Mai Xuan Van[1], Truong Khanh Duy[2], Truong Thi Hanh[3], Tran Tien Dat[2], Nguyen Ngoc Nho[2], Nguyen Tuong Tri [1,2]*

[1] University of Education - Hue University, Hue, Vietnam
[2] Institute of Open Education and Information Technology - Hue University, Hue, Vietnam
[3] Quốc Học – Huế High School for the Gifted, Hue, Vietnam

Abstract. Protein–protein interactions (PPIs) form the foundation of many intracellular biological processes, and predicting PPIs directly from amino acid sequences remains a core direction in computational biology. The advent of next-generation Protein Language Models (PLMs), such as ESM-2, enables the generation of sequence representations rich in evolutionary information and latent structural signals. However, these representations often possess extremely high dimensionality, contain significant noise, and exhibit high internal correlation, making it difficult for traditional machine learning models to exploit them effectively and increasing the risk of overfitting. This challenge demands an approach capable of distilling knowledge and eliminating data redundancy while preserving core biological signals. In this work, we propose E–StackPPI (Embedding-Stacking Protein-Protein Interaction prediction framework), a fully embedding-based PPI prediction framework centered on a three-stage layer-wise feature selection mechanism applied directly to embeddings aggregated from the last hidden layers of the ESM-2 650M model. Specifically, the process sequentially: (1) removes dimensions with low variance; (2) retains highly discriminative features based on LightGBM feature importance; and (3) eliminates dimensions with high Pearson correlation to reduce information redundancy. The refined feature set is fed into a stacking architecture, where two parallel LightGBM branches are integrated at the decision layer via Logistic Regression (LR). Experiments on two benchmark datasets from the Database of Interacting Proteins (DIP), including DIP–Yeast and DIP–Human, show that E–StackPPI achieves favorable and stable results across key metrics, including accuracy, MCC, as well as ROC-AUC and PR-AUC indices. When benchmarked against twelve advanced methods summarized in the study by Li et al., our model demonstrates competitive performance on both datasets. These findings highlight the essential role of layer-wise feature selection in mitigating noise and effectively leveraging high-dimensional PLM embeddings, thereby opening a feasible and promising sequence–only approach to PPI prediction without the need for supplementary structural data.

Keywords: Protein-protein interaction, Multi-stage feature selection, Protein Language Models, Stacking model

1 Giới thiệu

Tương tác protein–protein (PPI) là nền tảng của phần lớn các hoạt động sinh học bên trong tế bào, nơi các quá trình như truyền tín hiệu, điều hòa biểu hiện gen và hình thành phức hợp chức năng đều phụ thuộc vào mạng lưới liên kết phân tử này. Khi một protein thay đổi trạng thái hoạt động hoặc kết nối với một đối tác mới, toàn bộ mạng lưới có thể bị tác động dây chuyền, kéo theo các biến đổi sinh lý hoặc rối loạn bệnh lý ở cấp độ hệ thống. Vì vậy, việc nhận diện chính xác các cặp protein tương tác giữ vai trò thiết yếu trong việc giải mã cơ chế phân tử và hỗ trợ thiết kế thuốc, đặc biệt là các phân tử nhỏ có khả năng tác động lên bề mặt tiếp xúc giữa hai protein [1, 2].

Các phương pháp thực nghiệm như lai hai loại nấm men (yeast two-hybrid, Y2H) hoặc đồng miễn dịch kết tủa (co-immunoprecipitation, Co-IP) thường đem lại độ tin cậy cao; tuy nhiên, chúng lại đòi hỏi nhiều thời gian, chi phí và nhân lực phòng thí nghiệm. Do hạn chế này, các hướng tiếp cận dựa trên mô phỏng tính toán (in silico) đã được chú trọng nhằm sàng lọc và ưu tiên những tương tác tiềm năng trước khi đưa vào kiểm chứng thực nghiệm.

Những nỗ lực ban đầu trong lĩnh vực dự đoán PPI dựa vào việc xây dựng đặc trưng thủ công từ hiểu biết về sinh học phân tử, bao gồm thành phần axit amin, các đại lượng hóa lý đơn giản hoặc các dạng thống kê trượt dọc theo chuỗi. Các đặc trưng này được đưa vào những bộ phân loại truyền thống như máy véc tơ hỗ trợ (Support Vector Machine – SVM) hoặc rừng ngẫu nhiên (Random Forest – RF) [3-6]. Cách tiếp cận này có ưu điểm dễ diễn giải và phù hợp trong bối cảnh dữ liệu hạn chế; tuy nhiên, lại khó nắm bắt được tín hiệu tiến hóa sâu hơn và các đặc điểm cấu trúc ẩn trong chuỗi, từ đó hạn chế khả năng khái quát trên các loài hoặc bộ dữ liệu khác nhau. Nhằm mô hình hóa tốt hơn các mối quan hệ phi tuyến, các thuật toán tăng cường dạng cây như XGBoost (Extreme Gradient Boosting) và LightGBM (Light Gradient Boosting Machine) đã được triển khai và chứng minh hiệu quả trong nhiều công trình dự đoán PPI [7-10].

Trong giai đoạn mở rộng của học sâu, nhiều kiến trúc hiện đại lần lượt xuất hiện như mạng tích chập (Convolutional Neural Network – CNN), mạng hồi quy (Recurrent Neural Network – RNN), cơ chế chú ý (attention mechanism) và các cấu trúc mạng nhiều tầng. Nhờ khả năng tự học biểu diễn từ chuỗi mà không phụ thuộc hoàn toàn vào đặc trưng thủ công, các mô hình này đã đạt hiệu năng nổi bật trên nhiều bộ dữ liệu chuẩn [11-13]. Một đóng góp sớm là mô hình mã hóa tự động xếp chồng (stacked autoencoder – SAE) cho dự đoán PPI dựa trên chuỗi của Sun và cộng sự [14].

Ngoài hai xu hướng chính nói trên, một hướng tiếp cận mang tính trung gian cũng đã được phát triển, kết hợp cả tri thức sinh học lẫn các phép biến đổi toán học hoặc hình học nhằm gia tăng mức độ giàu thông tin của đặc trưng. Một ví dụ tiêu biểu là công trình của Khanh Duy Truong và cộng sự [15], nhóm tác giả đã tích hợp phép biến đổi Hilbert vào ma trận điểm tiến

hóa PSSM (Position-Specific Scoring Matrix) để tạo ra biểu diễn tần số, từ đó thu được thông tin dựa trên biên độ và pha tức thời. Kết quả cho thấy việc tổ chức lại đặc trưng theo một cấu trúc phù hợp có thể cải thiện đáng kể hiệu năng của mô hình, thậm chí trong nhiều trường hợp còn mang lại lợi ích đáng kể so với việc gia tăng độ sâu của mạng.

Sự xuất hiện của các mô hình ngôn ngữ protein (Protein Language Model – PLM), tiêu biểu là ESM-2 [16] và ProtT5 [17], đã đánh dấu một bước ngoặt quan trọng. Các mô hình này được huấn luyện tự giám sát trên tập dữ liệu chuỗi lớn và học cách mã hóa mỗi protein thành một dạng biểu diễn nhúng đa chiều, trong đó thông tin tiến hóa, cấu trúc và chức năng được tích hợp dưới dạng đặc trưng tiềm ẩn. Những biểu diễn này đã chứng minh hiệu quả trong nhiều hướng ứng dụng kế tiếp, đặc biệt là các bài toán dự đoán dựa trên chuỗi, bao gồm cả dự đoán PPI. Công trình xCAPT5 của Thanh Hai Dang và Tien Anh Vu [18] cũng cho thấy biểu diễn nhúng T5-XL-UniRef50 có thể vượt trội so với toàn bộ phương pháp dựa trên đặc trưng thủ công.

Tuy vậy, việc đưa trực tiếp biểu diễn nhúng có số chiều lớn vào mô hình truyền thống đối mặt với hai thách thức lớn. Thách thức thứ nhất liên quan đến sự bùng nổ số chiều (dimensionality explosion) khi mà một chuỗi protein từ ESM-2 có thể được mã hóa thành véc tơ dài hơn một nghìn chiều; khi ghép cặp để biểu diễn tương tác, số chiều có thể tăng gấp đôi. Trong số đó, chỉ một phần nhỏ thực sự mang tín hiệu phân loại, phần còn lại chủ yếu gây nhiễu hoặc trùng lặp, tạo ra nguy cơ quá khớp và tăng chi phí tính toán. Điều này đặc biệt nghiêm trọng khi làm việc với các bộ dữ liệu sinh học có kích thước mẫu hạn chế so với số lượng đặc trưng, hệ quả là sự bùng nổ dữ liệu dẫn đến hiện tượng "lời nguyền số chiều" (curse of dimensionality), gây trở ngại lớn cho quá trình tối ưu hóa mô hình. Thách thức thứ hai chính là khả năng lý giải rõ ràng và minh bạch, vì việc đưa toàn bộ véc tơ nhúng vào một bộ phân loại phi tuyến mạnh sẽ làm cho quy trình khó kiểm soát và khó đánh giá vai trò của từng thành phần.

Trong bối cảnh này, bước chọn lọc đặc trưng (feature selection) đóng vai trò như một tầng điều tiết cần thiết giữa PLM và bộ phân loại. Ding và cộng sự [19] đã đề xuất chiến lược chọn lọc hai giai đoạn dựa trên rừng ngẫu nhiên và thuật toán di truyền (genetic algorithm – GA), điều này cho thấy việc kết hợp nhiều chiến lược chọn lọc có thể tạo ra tập đặc trưng gọn nhưng vẫn giàu thông tin. Các tổng quan gần đây [20, 21] cùng kết quả từ bộ chuẩn PEER [22] đều khẳng định rằng việc chuẩn hóa và chọn lọc đặc trưng là bước không thể thiếu khi xử lý biểu diễn nhúng từ PLM.

Trong nghiên cứu này, chúng tôi kế thừa ưu thế biểu diễn của PLM nhưng đồng thời bổ sung một quy trình chọn lọc đặc trưng theo tầng nhằm nén bớt chiều, giảm nhiễu và tái cấu trúc không gian đặc trưng theo hướng phù hợp hơn cho nhiệm vụ phân loại. Trên cơ sở này, chúng tôi xây dựng E–StackPPI, một khung dự đoán PPI hoàn toàn dựa trên biểu diễn nhúng, trong đó thông tin từ ESM-2 được đưa qua ba bước chọn lọc liên tiếp trước khi đi vào hai nhánh mô hình LightGBM chạy song song và cuối cùng được kết hợp tại tầng quyết định bằng hồi quy logistic

(Logistic Regression – LR). Phương pháp tiếp cận E–StackPPI đặt trọng tâm vào tổ chức trình tự xử lý đặc trưng, bao gồm chuẩn hóa thông tin, lựa chọn các chiều quan trọng và loại bỏ các thành phần dư thừa, thay vì gia tăng độ sâu của mô hình.

2 Phương pháp đề xuất

2.1 Trích xuất biểu diễn nhúng protein

Quy trình xử lý bắt đầu bằng việc chuyển đổi trình tự axit amin thành các biểu diễn số học thông qua mô hình ngôn ngữ protein ESM-2 (phiên bản *esm2_t33_650M_UR50D*) với 650 triệu tham số. Mô hình này cung cấp không gian chiều ẩn gốc là $d_{model} = 1280$. Để tối ưu hóa tài nguyên tính toán trong khi vẫn đảm bảo thu nhận đầy đủ các đặc trưng tiến hóa và tín hiệu cấu trúc tiềm ẩn, chúng tôi trích xuất biểu diễn từ lớp ẩn cuối cùng của mô hình.

Đối với mỗi protein, quá trình này tạo ra một ma trận biểu diễn $\mathbf{E} \in \mathbb{R}^{L\times 1280}$, với L là độ dài chuỗi. Để tạo đầu vào nhất quán cho các bước xử lý tiếp theo, chúng tôi rút gọn ma trận này thành một véc tơ toàn cục bằng cách lấy trung bình theo chiều dài chuỗi (Global Average Pooling). Nếu ký hiệu $\mathbf{e}_i$ là biểu diễn nhúng của axit amin tại vị trí thứ i, véc tơ trung bình $\overline{\mathbf{e}} \in \mathbb{R}^{1280}$ được xác định bởi công thức 1:

$$\overline{\mathbf{e}} = \frac{1}{L}\sum_{i=1}^{L} \mathbf{e}_i \quad (1)$$

Thao tác này cô đọng toàn bộ thông tin ngữ nghĩa và ngữ cảnh của protein thành một vector đại diện duy nhất. Khi chuyển sang cặp protein (p, q), chúng tôi thực hiện ghép nối song song hai véc tơ trung bình (công thức 2), thu được véc tơ $\Phi(p, q) \in \mathbb{R}^{2560}$:

$$\Phi(p, q) = [\overline{\mathbf{e}}_p \parallel \overline{\mathbf{e}}_q] \quad (2)$$

Tại giai đoạn này, mỗi cặp protein được biểu diễn bởi 2560 chiều đặc trưng. Đây chính là lớp dữ liệu đầu vào cho quy trình chọn lọc đặc trưng đa giai đoạn.

2.2 Cơ chế chọn lọc đặc trưng đa giai đoạn (Tripartite Feature Selection)

Để giải quyết lời nguyền số chiều và loại bỏ các thành phần nhiễu trong không gian nhúng cao chiều, E–StackPPI thực hiện quy trình chọn lọc qua ba giai đoạn nghiêm ngặt:

Giai đoạn 1: Lọc phương sai (Variance Thresholding). Chúng tôi loại bỏ các chiều đặc trưng có biến thiên cực thấp trên tập huấn luyện với ngưỡng $\tau_v = 0.002$. Bước này giúp loại bỏ các đặc trưng gần như hằng số, vốn không mang giá trị phân biệt trong việc xác định tương tác. Tuy nhiên, thao tác này được thực hiện trước bất kỳ quá trình chuẩn hóa nào, đặc biệt là đối với các đặc trưng nhúng, nhằm loại bỏ triệt để các đặc trưng gần như hằng số, vốn không mang giá

trị phân biệt trong việc xác định tương tác và có thể gây mất ổn định số học cho các bước xử lý kế tiếp.

Giai đoạn 2: Chọn lọc dựa trên tầm quan trọng (Importance–based Selection). Chúng tôi sử dụng mô hình LightGBM như một bộ phân loại thăm dò để trích xuất số *Feature Importance* dựa trên tổng mức tăng thông tin. Thay vì chọn một số lượng đặc trưng cố định, chúng tôi áp dụng ngưỡng tích lũy $q = 0.90$. Cơ chế này đảm bảo giữ lại tập hợp tối thiểu các đặc trưng hội tụ được 90% năng lượng thông tin cần thiết cho việc phân loại.

Giai đoạn 3: Loại bỏ thông tin dư thừa dựa trên tính tương quan (Correlation Filtering). Để tối đa hóa tính trực giao giữa các chiều dữ liệu, chúng tôi tính hệ số tương quan Pearson giữa mọi cặp đặc trưng còn lại. Nếu $|\rho| > 0.90$, chúng tôi chỉ giữ lại chiều có mức độ quan trọng cao hơn và loại bỏ chiều kia. Chiến lược này triệt tiêu hiện tượng đa cộng tuyến, giúp bộ học máy xếp chồng hoạt động ổn định và tránh quá khớp.

2.3 Kiến trúc học máy xếp chồng (Ensemble Stacking Architecture)

Sau khi hoàn tất lọc đặc trưng, không gian đặc trưng súc tích Φ^{final} được đưa vào hệ thống phân loại đa tầng bao gồm:

Tầng cơ sở (Base Layer): Chúng tôi triển khai hai nhánh LightGBM chạy song song với các thiết lập *subsample* và *colsample_bytree* khác biệt nhằm tạo ra sự đa dạng trong việc khai thác không gian đặc trưng. Với mỗi cặp protein, hai mô hình này cung cấp các xác suất dự đoán f_1 và f_2.

Tầng tổng hợp (Meta Layer): Cặp xác suất $[f_1, f_2]$ được chuyển tiếp đến một bộ học Meta dựa trên hồi quy Logistic (Logistic Regression). Mô hình này học tổ hợp tuyến tính tối ưu để đưa ra dự đoán cuối cùng $\hat{y}$ (công thức 3):

$$\hat{y} = \sigma(\beta_0 + \beta_1 f_1 + \beta_2 f_2) \quad (3)$$

Việc sử dụng mô hình tuyến tính ở tầng tổng hợp giúp ổn định ranh giới quyết định, tận dụng tối ưu các góc nhìn bổ sung từ tầng cơ sở và giảm thiểu rủi ro quá khớp so với các mô hình phi tuyến phức tạp.

3 Thử nghiệm và đánh giá kết quả nghiên cứu

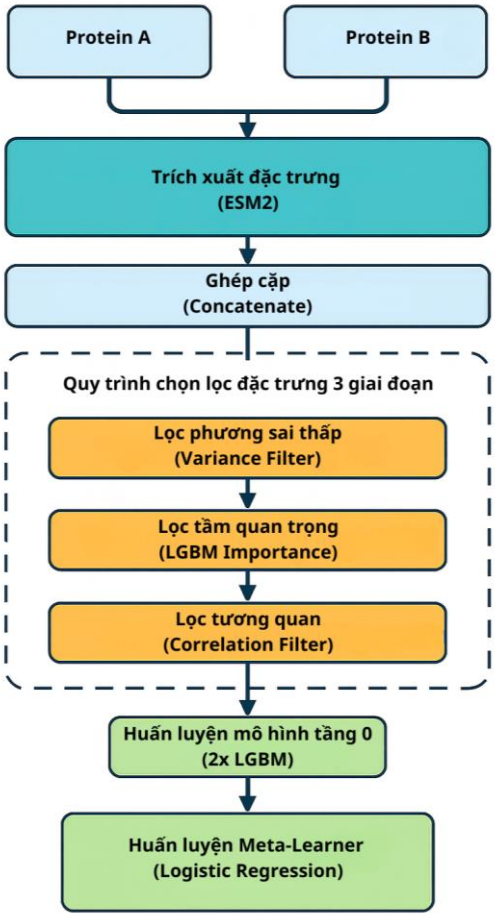
3.1 Chuẩn bị dữ liệu thử nghiệm

Hai bộ dữ liệu được sử dụng trong nghiên cứu này đều có nguồn gốc từ Database of Interacting Proteins (DIP) [23], vốn là một trong những hệ thống dữ liệu PPI có độ tin cậy thử nghiệm cao nhất hiện nay. Đối với mỗi bộ dữ liệu, chúng tôi chỉ giữ lại những cặp tương tác đã

được xác nhận sinh học, đồng thời loại bỏ hoàn toàn các trường hợp tự tương tác và những bản trùng lặp theo mọi phép hoán vị. Bước chuẩn hóa này nhằm thiết lập một tập dữ liệu dương tính tinh gọn, ngăn chặn việc thổi phồng hiệu năng mô hình do sự hiện diện của các cặp trùng lặp.

Việc xây dựng tập âm tính là một bước đặc biệt quan trọng. Theo thông lệ trong cộng đồng, chúng tôi không chọn âm tính một cách ngẫu nhiên, mà ưu tiên những cặp protein không xuất hiện trong bất kỳ báo cáo tương tác nào và có đặc điểm sinh học tách biệt. Quy tắc này giúp giảm nguy cơ đưa nhầm các cặp thực sự có tương tác vào tập âm, từ đó làm tăng độ tin cậy của phép đánh giá. Tỷ lệ dương tính và âm tính được giữ cân bằng 1: 1 trong toàn bộ thử nghiệm.

Tất cả chuỗi protein được ánh xạ sang không gian biểu diễn nhúng của mô hình ESM-2 650M, sau đó đi qua ba tầng chọn lọc đặc trưng xếp tầng, như minh họa ở Hình 1. Các tầng này lần lượt loại bỏ chiều ít biến thiên, chọn lọc theo độ quan trọng, rồi giảm tương quan; từ đó tái cấu trúc lại không gian nhúng theo hướng cô đọng và ít nhiễu hơn.



Hình 1. Sơ đồ hệ thống E–StackPPI: Quy trình từ trích xuất đặc trưng nhúng ESM-2 (650M), lọc đặc trưng 3 giai đoạn đến dự đoán bằng kiến trúc Stacking.

3.2 Thiết lập thử nghiệm

Quá trình kiểm định được tiến hành bằng phương pháp kiểm định chéo 5 lần (5-fold cross-validation). Ở mỗi vòng lặp, một phần dữ liệu (fold) được giữ lại để kiểm thử độc lập, bốn phần còn lại được sử dụng cho huấn luyện và điều chỉnh siêu tham số. Mỗi lần huấn luyện đều bao gồm đầy đủ ba giai đoạn: chọn lọc đặc trưng, hai nhánh LightGBM chạy song song và tầng hợp nhất hồi quy logistic. Toàn bộ quy trình được tái huấn luyện hoàn chỉnh ở mỗi phần nhằm đảm bảo phép đánh giá phản ánh đúng sự ổn định của mô hình.

Hệ thống chỉ số đánh giá bao gồm độ chính xác tổng quan (Accuracy), độ chính xác (Precision), độ bao phủ (Recall), điểm F1 (F1-score), độ đặc hiệu (Specificity), hệ số tương quan Matthews (Matthews Correlation Coefficient – MCC), diện tích dưới đường cong ROC (ROC-AUC) và diện tích dưới đường cong Precision–Recall (PR-AUC).

3.3 Thử nghiệm trên bộ dữ liệu DIP–Yeast

Thay đổi kích thước biểu diễn nhúng theo từng lần kiểm định chéo

Ba tầng chọn lọc đặc trưng có tác dụng thu gọn mạnh không gian biểu diễn ban đầu, vốn có kích thước 2560 chiều do ghép hai véc tơ nhúng 1280 chiều của hai protein trong mỗi cặp. Sau khi áp dụng đầy đủ quy trình chọn lọc, số chiều còn lại giảm đáng kể và hình thành một không gian súc tích hơn nhưng vẫn bảo toàn phần thông tin quan trọng nhất. Kết quả qua năm lần kiểm định chéo của bộ dữ liệu DIP–Yeast được trình bày trong bảng dưới đây.

Bảng 1. Kích thước véc tơ nhúng trước và sau khi chọn lọc trên bộ dữ liệu DIP–Yeast

Lần	Kích thước ban đầu	Sau chọn lọc
Lần 1	2560	561
Lần 2	2560	555
Lần 3	2560	560
Lần 4	2560	556
Lần 5	2560	553
Trung bình	**2560**	**557**

Kết quả cho thấy kích thước đặc trưng sau chọn lọc rất ổn định giữa các lần kiểm định chéo, dao động nhẹ quanh mức ~ 557 chiều. Điều này cho thấy quy trình chọn lọc đặc trưng vận hành nhất quán trên nhiều phân hoạch dữ liệu khác nhau và có khả năng giữ lại những chiều thông tin cốt lõi một cách ổn định trên bộ dữ liệu DIP–Yeast. Sự ổn định này gợi ý rằng tồn tại một tập hợp các đặc trưng lõi mang tính bảo tồn cao, đại diện cho các tính chất hóa lý và tiến hóa nền tảng quy định khả năng tương tác ở nấm men.

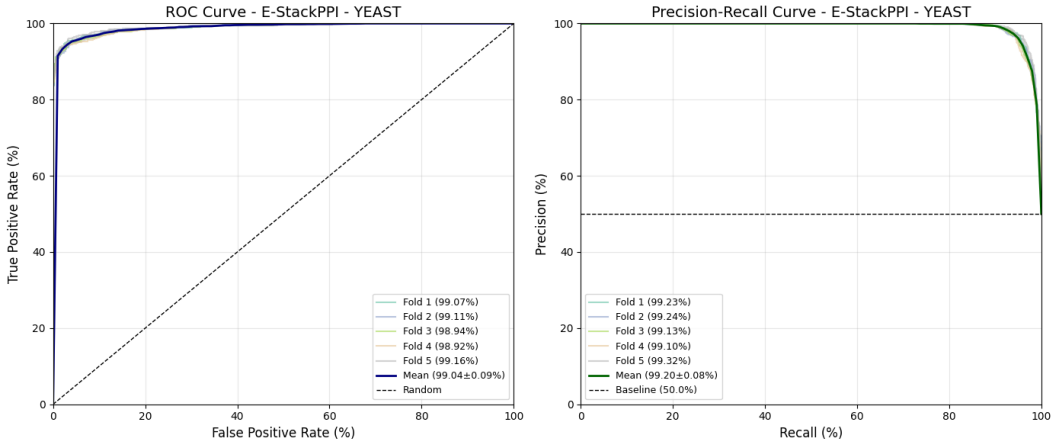
Kết quả thử nghiệm

Trên bộ dữ liệu DIP–Yeast, mô hình đạt được hiệu năng ổn định trên cả năm lần kiểm định chéo. Độ chính xác tổng thể dao động quanh mức 95.67% với sai số nhỏ, trong khi MCC trung bình đạt khoảng 91.36%, cho thấy ranh giới của hai lớp trở nên rõ ràng nhờ vào quá trình chọn lọc đặc trưng. Các giá trị ROC-AUC và PR-AUC đều vượt mức 99%, phản ánh việc mô hình học được một ranh giới quyết định trơn tru và ít phụ thuộc vào cấu trúc của từng phần. Kết quả chi tiết được trình bày trong Bảng 2.

Để có cái nhìn trực quan hơn về năng lực phân loại của mô hình tại các ngưỡng quyết định khác nhau, chúng tôi trình bày đường cong ROC và đường cong Precision–Recall (PR) trong Hình 2. Đồ thị cho thấy các đường cong đều tiệm cận sát góc trên bên trái với diện tích dưới đường cong ROC–AUC trung bình đạt 99.04% và PR–AUC trung bình đạt 99.20%.

Bảng 2. Kết quả kiểm định chéo 5 phần (5-fold cross-validation) trên bộ dữ liệu DIP–Yeast

Lần	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Specificity (%)	MCC (%)	ROC-AUC (%)	PR-AUC (%)
Lần 1	95.98	96.99	94.91	95.93	97.05	91.98	99.07	99.23
Lần 2	95.31	96.60	93.92	95.24	96.69	90.65	99.11	99.24
Lần 3	95.53	96.79	94.19	95.47	96.87	91.10	98.94	99.13
Lần 4	95.53	96.96	94.01	95.46	97.05	91.10	98.92	99.10
Lần 5	95.98	96.90	95.00	95.94	96.96	91.97	99.16	99.32
Trung bình	**95.67 ± 0.30**	**96.85 ± 0.16**	**94.40 ± 0.51**	**95.61 ± 0.31**	**96.93 ± 0.15**	**91.36 ± 0.59**	**99.04 ± 0.09**	**99.20 ± 0.08**



Hình 2. Hiệu năng phân loại trên bộ dữ liệu DIP–Yeast được thể hiện qua đường cong ROC (trái) và đường cong Precision–Recall (phải)

Sự đồng nhất giữa hai chỉ số AUC này khẳng định sự ổn định của khung làm việc E–StackPPI qua các phiên kiểm định chéo. Đặc biệt, giá trị PR–AUC ở mức cao (99.20%) minh chứng cho khả năng duy trì độ chính xác (Precision) ưu việt ngay cả khi mô hình nỗ lực tối đa hóa khả năng nhận diện các tương tác thực (Recall), một đặc tính thiết yếu giúp giảm thiểu tỷ lệ dương tính giả trong thực nghiệm sinh học.

3.4 Kết quả trên bộ dữ liệu DIP–Human

Thay đổi kích thước biểu diễn nhúng theo từng lần kiểm định chéo

Trên bộ dữ liệu DIP–Human, mỗi cặp protein được biểu diễn ban đầu bằng một véc tơ nhúng kích thước 2560 chiều, tương tự như trên DIP–Yeast. Sau khi áp dụng quy trình chọn lọc ba bước, số chiều được giữ lại có xu hướng cao hơn, phản ánh đặc điểm chức năng đa dạng và mức độ phức tạp lớn hơn của protein ở người.

Bảng 3. Kích thước véc tơ nhúng trước và sau khi chọn lọc trên bộ dữ liệu DIP–Human.

Lần	Kích thước ban đầu	Sau chọn lọc
Lần 1	2560	1060
Lần 2	2560	1068
Lần 3	2560	1038
Lần 4	2560	1086
Lần 5	2560	1036
Trung bình	**2560**	**1057.6**

Kết quả cho thấy số chiều sau chọn lọc, số chiều dao động quanh mức ~ 1057, cao hơn so với DIP–Yeast. Mức tăng này cho thấy quy trình chọn lọc thích ứng linh hoạt với mức độ đa dạng của dữ liệu, đồng thời giữ được sự ổn định cần thiết để mô tả hiệu quả các cặp protein trong đặc điểm sinh học phức tạp hơn ở người. Điều này chứng tỏ rằng cơ chế chọn lọc không áp đặt một kích thước cố định, mà "co giãn" tự nhiên theo độ phức tạp nội tại của hệ sinh học đang xét.

Kết quả thử nghiệm

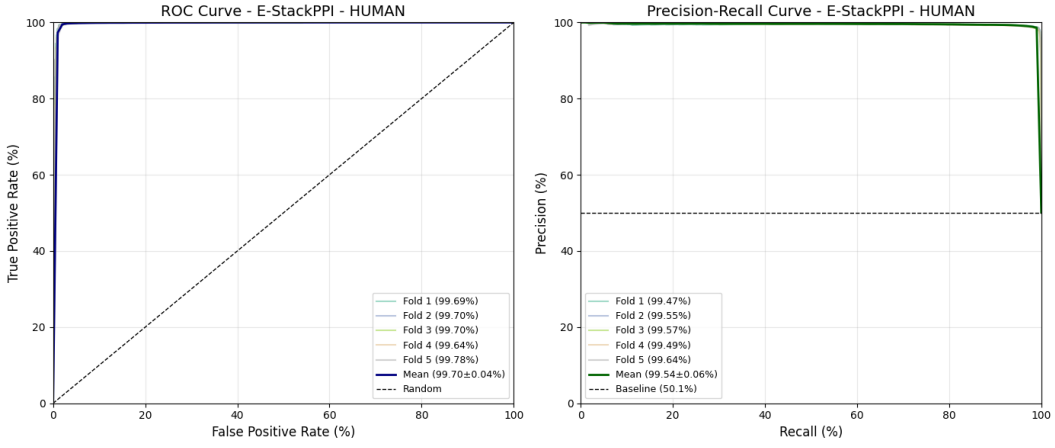
Trên bộ dữ liệu DIP–Human, vốn được xem là một bài toán thách thức hơn do mức độ đa dạng về chức năng và cấu trúc của protein ở người, E–StackPPI vẫn duy trì được hiệu năng rất cao. Độ chính xác tổng quan trung bình đạt 98.60%, trong khi độ chính xác và độ đặc hiệu đều xấp xỉ 98.77%. Đồng thời, diện tích dưới đường cong ROC đạt 99.70% và diện tích dưới đường cong chính xác–bao phủ đạt 99.54%. Việc toàn bộ các chỉ số đánh giá cùng tiến sát giới hạn hiệu năng cho thấy không gian biểu diễn nhúng sau khi chọn lọc đã được mô hình khai thác một cách hiệu quả, ổn định và nhất quán. Kết quả đầy đủ được trình bày trong Bảng 4.

Bảng 4. Kết quả kiểm định chéo 5 phần (5-fold cross-validation) trên bộ dữ liệu DIP–Human.

Lần	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Specificity (%)	MCC (%)	ROC-AUC (%)	PR-AUC (%)
Lần 1	98.67	98.90	98.44	98.67	98.90	97.35	99.69	99.47
Lần 2	98.71	98.82	98.59	98.71	98.82	97.41	99.70	99.55
Lần 3	98.52	98.64	98.39	98.52	98.64	97.03	99.70	99.57
Lần 4	98.44	98.62	98.26	98.44	98.62	96.88	99.64	99.49
Lần 5	98.66	98.87	98.44	98.66	98.88	97.32	99.78	99.64
Trung bình	**98.60 ± 0.12**	**98.77 ± 0.13**	**98.43 ± 0.12**	**98.60 ± 0.11**	**98.77 ± 0.13**	**97.20 ± 0.23**	**99.70 ± 0.04**	**99.54 ± 0.06**

Tương tự như thử nghiệm trên nấm men, hiệu năng phân loại trên bộ dữ liệu người được minh họa chi tiết thông qua đường cong ROC và đường cong Precision–Recall (PR) ở Hình 3. Với giá trị ROC–AUC đạt 99.70% và PR–AUC đạt 99.54%, mô hình thể hiện khả năng phân lớp tiệm cận mức tuyệt đối.

Sự hội tụ của các đường cong qua các lần kiểm định chéo phản ánh rằng quy trình sàng lọc ba giai đoạn đã tinh lọc được một không gian đặc trưng tối ưu, giúp thiết lập ranh giới quyết định minh bạch và sắc sảo. Kết quả này không chỉ khẳng định sức mạnh của kiến trúc Stacking trong việc xử lý dữ liệu phức tạp ở người mà còn chứng minh khả năng kiểm soát nhiễu và phòng chống hiện tượng quá khớp hiệu quả, ngay cả khi đối mặt với sự bùng nổ số chiều từ các mô hình ngôn ngữ protein lớn.



Hình 3. Đường cong ROC (trái) và Precision–Recall (phải) trên bộ dữ liệu DIP–Human, minh họa khả năng phân tách vượt trội của mô hình.

So sánh với các phương pháp hiện hành

Để đánh giá khách quan vị trí của E–StackPPI trong bối cảnh nghiên cứu dự đoán PPI, chúng tôi tiến hành đối chiếu mô hình với nhóm phương pháp tiên tiến (SOTA) đã được công bố và sử dụng rộng rãi trên cùng hai bộ dữ liệu DIP–Yeast và DIP–Human. Các số liệu tham chiếu được trích từ công trình của Li và cộng sự [24], trong đó nhóm tác giả đã tổng hợp hiệu năng của mười hai mô hình tiêu biểu, bao gồm các hướng tiếp cận từ đặc trưng thủ công, mô hình tiến hoá, mạng nơ-ron tích chập đến các kỹ thuật kết hợp nhiều mô hình.

So sánh trên bộ dữ liệu DIP–Yeast

Bảng 5 trình bày hiệu năng của các phương pháp tiên tiến trên bộ dữ liệu DIP–Yeast. Độ chính xác của các mô hình trong nhóm tham chiếu trải dài từ khoảng 86% đến 94.43%, trong đó các phương pháp mạnh nhất như của DL của Du và cộng sự [25], PR-LPQ của Wong và cộng sự. [26], và Bio2Vec CNN của Wang và cộng sự [27] đạt hệ số tương quan Matthews ở mức từ 87.49% đến 88.97%. Nhìn chung, rất ít mô hình vượt được ngưỡng 90% theo chỉ số MCC trên bộ dữ liệu này.

Bảng 5. So sánh hiệu năng với 12 phương pháp khác trên tập dữ liệu DIP–Yeast.

Phương pháp	Accuracy (%)	Sensitivity (%)	Precision (%)	MCC (%)
Du và cộng sự (DL) [25]	94.43 ± 0.30	92.06 ± 0.36	96.55 ± 0.59	88.97 ± 0.62
Wong và cộng sự (PR-LPQ + RoF) [26]	93.92 ± 0.36	91.10 ± 0.31	96.45 ± 0.45	88.56 ± 0.63
Wang và cộng sự (Bio2Vec + CNN) [27]	93.30	92.70	93.55	87.49
You và cộng sự (MCD + SVM) [28]	91.36 ± 0.36	90.67 ± 0.69	91.94 ± 0.62	84.21 ± 0.59
An và cộng sự (PSSMMF + SVM) [29]	90.48 ± 0.76	90.26 ± 0.87	90.58 ± 0.98	82.84 ± 1.27
Wang và cộng sự (3-mers + CNN) [27]	90.26	88.14	91.65	82.38
Li và cộng sự (OLPP + RoF) [24]	90.07 ± 0.60	89.83 ± 1.41	90.24 ± 0.56	82.10 ± 0.97
Guo và cộng sự (ACC + SVM) [3]	89.33 ± 2.67	89.93 ± 3.68	88.87 ± 6.16	**N/A**
Zhou và cộng sự (LD + SVM) [30]	88.56 ± 0.33	87.37 ± 0.22	89.50 ± 0.60	77.15 ± 0.68
Guo và cộng sự (AC + SVM) [3]	87.36 ± 1.38	87.30 ± 4.68	87.82 ± 4.33	**N/A**
You và cộng sự (Multiple + PCA-EELM) [5]	87.00 ± 0.29	86.15 ± 0.43	87.59 ± 0.32	77.36 ± 0.44
Yang và cộng sự (LD + KNN) [31]	86.15 ±1.17	81.03 ± 1.74	90.24 ± 1.34	**N/A**
E-StackPPI (đề xuất)	**95.67 ± 0.30**	**94.40 ± 0.51**	**96.85 ± 0.16**	**91.36 ± 0.59**

Ghi chú: Giá trị **N/A** biểu thị trường hợp không có số liệu tương ứng trong công trình gốc. Các giá trị in đậm thể hiện mô hình đạt hiệu năng cao nhất trong từng cột so sánh.

Khi đối chiếu với Bảng 2, có thể thấy E–StackPPI đạt độ chính xác tổng quan trung bình 95.67%, độ chính xác 96.85% và hệ số tương quan Matthews 91.36%, vượt trội so với toàn bộ các phương pháp tham chiếu. Đáng chú ý hơn, ROC-AUC và PR-AUC đều vượt ngưỡng 99%, cao hơn khoảng 2-3 phần trăm so với mức tốt nhất từng được báo cáo trên bộ dữ liệu này.

3.5 So sánh trên bộ dữ liệu DIP–Human

Hiệu năng của nhóm phương pháp tiên tiến trên bộ dữ liệu DIP–Human được trình bày trong Bảng 6. Đây là bộ dữ liệu có mặt bằng hiệu năng cao hơn so với DIP–Yeast, khi nhiều mô hình đạt độ chính xác trên 96% và hệ số tương quan Matthews xấp xỉ 95%. Mặc dù vậy, không có phương pháp nào trong nhóm đạt được ROC-AUC hoặc PR-AUC vượt ngưỡng 99%.

Bảng 6. So sánh hiệu năng với 12 phương pháp khác trên tập dữ liệu DIP–Human

Phương pháp	Accuracy (%)	Sensitivity (%)	Precision (%)	MCC (%)
Du và cộng sự (DL) [25]	98.14	96.95	**99.13**	96.29
Ding và cộng sự (MMI + NMBAC + RF) [32]	97.56	96.57	98.30	95.13
Pan và cộng sự (LDA + RF) [6]	96.40	94.20	**N/A**	92.80
Huang và cộng sự (DTC + SMR + WSRC) [8]	96.30	92.63	99.59	92.82
OLPP + RoF (Li và cộng sự, 2021) [2]	96.09	95.20	96.56	92.47
Ding và cộng sự (MMI + RF) [32]	96.08	95.05	96.97	92.17
Pan và cộng sự (LDA + RoF) [4]	95.70	97.60	**N/A**	91.80
Ding và cộng sự (NMBAC + RF) [32]	95.59	94.06	96.94	91.21
Pan và cộng sự (AC + RF) [4]	95.50	94.00	**N/A**	91.40
Pan và cộng sự (AC + RoF) [4]	95.10	93.30	**N/A**	91.00
Pan và cộng sự (LDA + SVM) [4]	90.70	89.70	**N/A**	81.30
Pan và cộng sự (AC + SVM) [4]	89.30	94.00	**N/A**	79.20
E-StackPPI (đề xuất)	**98.60**	**98.43**	98.77	**97.20**

Ghi chú: Giá trị **N/A** biểu thị trường hợp không có số liệu tương ứng trong công trình gốc. Các giá trị in đậm thể hiện mô hình đạt hiệu năng cao nhất trong từng cột so sánh.

Kết quả từ bảng 6 cho thấy E–StackPPI với độ chính xác tổng quan trung bình là 98.60% và hệ số tương quan Matthews là 97.20% đã vượt qua toàn bộ các phương pháp tiên tiến đã được báo cáo, bao gồm cả mô hình học sâu của Du và cộng sự [25] (MCC xấp xỉ 96.29%). Bên cạnh đó, hai chỉ số là diện tích dưới đường cong ROC 99.70% và chính xác–bao phủ 99.54% đều cao hơn đáng kể so với mức tốt nhất trong nhóm tham chiếu, cho thấy khả năng phân tách lớp của mô hình vượt trội và nhất quán trên bộ dữ liệu DIP–Human.

3.6 Thử nghiệm loại trừ

Để đánh giá một cách định lượng vai trò của từng tầng trong cơ chế chọn lọc đặc trưng, đồng thời kiểm chứng mức đóng góp của kiến trúc xếp tầng hai nhánh trong mô hình đề xuất, chúng tôi tiến hành một chuỗi thử nghiệm loại trừ trên hai bộ dữ liệu DIP–Yeast và DIP–Human. Mọi thử nghiệm đều sử dụng chung biểu diễn nhúng từ ESM–2 và áp dụng cùng quy trình kiểm định chéo 5 lần; sự khác biệt duy nhất nằm ở số tầng chọn lọc được kích hoạt và số nhánh của bộ phân loại. Bốn cấu hình được khảo sát bao gồm: **Baseline (ESM-2 Raw)** véc tơ nhúng ESM–2 với kích thước đầy đủ 2560 chiều được đưa trực tiếp vào một mô hình LightGBM duy nhất. Đây chính là cấu hình đối chứng nhằm trả lời câu hỏi về hiệu quả thực sự của quy trình đề xuất so với việc sử dụng mô hình ngôn ngữ gốc (chưa qua chọn lọc) trên cùng một bộ dữ liệu; (2) **Var–only**: chỉ loại bỏ các chiều có phương sai thấp, giúp đánh giá vai trò của tầng lọc biến thiên; (3) **Var + LGBM–Imp**: bổ sung thêm bước chọn các chiều có tầm quan trọng tích luỹ do LightGBM xác định nhưng chưa loại trừ các chiều có tương quan mạnh; (4) **E–StackPPI**: cấu hình đầy đủ của mô hình đề xuất, áp dụng tuần tự ba tầng chọn lọc đặc trưng, sau đó huấn luyện hai nhánh LightGBM chạy song song và kết hợp tại tầng quyết định bằng hồi quy logistic. Việc bổ sung hai nhánh LightGBM giúp mô hình quan sát không gian đặc trưng theo hai góc độ bổ sung, trong khi tầng quyết định học cách kết hợp tối ưu hai tín hiệu này trên từng phần dữ liệu. Nhờ cách tổ chức này, sự khác biệt giữa các cấu hình phản ánh trực tiếp vai trò của từng tầng chọn lọc và từng nhánh trong việc ổn định hoá tín hiệu, giảm nhiễu và tăng cường khả năng phân tách lớp của mô hình. Các cấu hình được so sánh dựa trên hai chỉ số quan trọng nhất là độ chính xác tổng quan và hệ số tương quan Matthews.

Kết quả loại trừ trên DIP–Yeast

Bảng 7 cho thấy việc bổ sung từng tầng chọn lọc đặc trưng mang lại cải thiện đáng kể so với baseline. Tầng loại bỏ phương sai thấp giúp giảm nhiễu nền; tầng chọn theo tầm quan trọng của LightGBM nâng cao khả năng phân tách hai lớp; và tầng loại bỏ tương quan Pearson tiếp tục củng cố thêm mức ổn định của mô hình. Mặc dù cấu hình E–StackPPI sử dụng hai nhánh song song thay vì một nhánh duy nhất, kết quả trên DIP–Yeast cho thấy hiệu năng của var–only và var + LGBM–Imp vẫn rất gần mức tối ưu, phản ánh đặc thù của bộ dữ liệu yeast vốn có ranh giới phân tách khá rõ ràng khi sử dụng biểu diễn nhúng từ ESM–2.

Bảng 7. Kết quả ablation study trên bộ dữ liệu DIP–Yeast

Cấu hình	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Specificity (%)	MCC (%)	ROC-AUC (%)	PR-AUC (%)
Baseline	92.72 ± 0.56	93.71 ± 0.49	91.60 ± 0.90	92.64 ± 0.59	93.85 ± 0.49	85.47 ± 1.10	97.64 ± 0.33	98.04 ± 0.27
Var-only	**95.73 ± 0.43**	**97.04 ± 0.19**	**94.33 ± 0.72**	**95.67 ± 0.45**	**97.12 ± 0.17**	**91.49 ± 0.84**	**99.09 ± 0.11**	**99.24 ± 0.09**
Var + LGBM-Imp	95.67 ± 0.30	96.85 ± 0.16	94.40 ± 0.51	95.61 ± 0.31	96.93 ± 0.15	91.36 ± 0.59	99.04 ± 0.11	99.20 ± 0.09
E–StackPPI	95.67 ± 0.30	96.85 ± 0.16	94.40 ± 0.51	95.61 ± 0.31	96.93 ± 0.15	91.36 ± 0.59	99.04 ± 0.11	99.20 ± 0.09

Kết quả loại trừ trên DIP–Human

Kết quả tương tự cũng được ghi nhận trên bộ dữ liệu DIP–Human, như trình bày trong Bảng 8. Mặc dù hiệu năng cơ bản trên bộ dữ liệu người vốn đã cao hơn đáng kể so với yeast, mỗi tầng chọn lọc vẫn giúp hiệu năng được cải thiện một cách ổn định, đặc biệt đối với hệ số MCC. Cấu hình đầy đủ của E–StackPPI tiếp tục cho kết quả tốt nhất, qua đó khẳng định tầm quan trọng của việc loại bỏ sự tương quan nội tại khi làm việc trong một không gian đặc trưng giàu tính đa dạng về chức năng.

Bảng 8. Kết quả ablation study trên bộ dữ liệu DIP–Human

Cấu hình	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Specificity (%)	MCC (%)	ROC-AUC (%)	PR-AUC (%)
Baseline	97.16 ± 0.09	98.53 ± 0.13	95.75 ± 0.12	97.12 ± 0.09	98.56 ± 0.13	94.35 ± 0.17	99.01 ± 0.07	99.03 ± 0.09
Var–only	98.62 ± 0.10	98.80 ± 0.15	98.43 ± 0.10	98.62 ± 0.10	98.80 ± 0.15	97.23 ± 0.20	99.71 ± 0.05	99.55 ± 0.08
Var + Imp	98.60 ± 0.12	98.77 ± 0.13	98.43 ± 0.12	98.60 ± 0.11	98.77 ± 0.13	97.20 ± 0.23	99.70 ± 0.05	99.54 ± 0.07
E–StackPPI	98.60 ± 0.12	98.77 ± 0.13	98.43 ± 0.12	98.60 ± 0.11	98.77 ± 0.13	97.20 ± 0.23	99.70 ± 0.05	99.54 ± 0.07

4 Thảo luận

Kết quả thực nghiệm trên hai bộ dữ liệu DIP–Yeast và DIP–Human là minh chứng rõ ràng cho hiệu quả của chiến lược chọn lọc đặc trưng ba tầng trong E–StackPPI. Các phân tích loại trừ tại Mục 3.6 khẳng định rằng sự tích hợp tuần tự của từng tầng chọn lọc đóng góp tích cực vào việc thiết lập hiệu năng tổng thể so với cấu hình cơ sở (baseline).

Đáng chú ý, kết quả thực nghiệm tại Bảng 7 và 8 cho thấy cấu hình Var–only đã đạt được ngưỡng hiệu năng tối ưu. Điều này gợi mở rằng các thông tin quan trọng nhất trong biểu diễn nhúng của ESM–2 thường tập trung ở những chiều có biến thiên rõ rệt. Tuy nhiên, việc duy trì hai giai đoạn lọc tiếp theo (LGBM-Imp và Correlation) là cần thiết để đảm bảo tính tối giản của vector đặc trưng cuối cùng, giúp loại bỏ triệt để các chiều dữ liệu đa cộng tuyến, từ đó gia tăng độ bền vững của mô hình khi vận hành trên các bộ dữ liệu có độ nhiễu cao hơn.

Những quan sát này củng cố nhận định rằng hướng tiếp cận tập trung vào việc tổ chức lại và chọn lọc không gian biểu diễn có thể là một lựa chọn khả thi, thay thế cho việc gia tăng độ sâu của các kiến trúc học sâu. Điểm mạnh của chiến lược này nằm ở triết lý thiết kế "lọc – chưng cất – tái cấu trúc" thay vì "gia tăng độ sâu", giúp giảm chi phí tính toán nhưng vẫn đạt hiệu quả phân tách cao. Chiến lược này đặc biệt phù hợp trong những bối cảnh dữ liệu không quá lớn nhưng yêu cầu độ chính xác cao và cần bảo đảm tính ổn định giữa các lần huấn luyện.

Về khía cạnh hiệu quả tính toán, mặc dù việc tính toán ma trận tương quan Pearson trong giai đoạn huấn luyện đòi hỏi chi phí nhất định, nhưng mô hình suy diễn cuối cùng (chỉ bao gồm các cây quyết định LightGBM và hồi quy tuyến tính) có độ phức tạp thấp hơn nhiều so với các mạng nơ-ron tích chập hay mạng đồ thị. Điều này cho phép E–StackPPI triển khai dễ dàng trên

các hệ thống không có GPU chuyên dụng. Ngoài ra, việc sử dụng Global Average Pooling, mặc dù có rủi ro làm loãng các tín hiệu motif cục bộ so với cơ chế Attention, lại giúp giảm đáng kể chiều dữ liệu đầu vào, phù hợp với mục tiêu xây dựng một khung dự đoán nhẹ nhưng vẫn tận dụng được sức mạnh ngữ nghĩa từ ESM-2.

Trong tương lai, việc mở rộng quy trình chọn lọc sang các mô hình ngôn ngữ protein thế hệ mới hơn, cũng như thử nghiệm trên các dạng tương tác sinh học khác, có thể mang lại thêm những hiểu biết quan trọng về vai trò của biểu diễn nhúng trong học máy cho sinh học phân tử.

Bên cạnh những ưu điểm nêu trên, E–StackPPI vẫn tồn tại một số hạn chế. Thứ nhất, mô hình phụ thuộc hoàn toàn vào chất lượng biểu diễn của ESM–2; do đó, những dạng tương tác phụ thuộc mạnh vào cấu trúc bậc ba hoặc bậc bốn có thể chưa được mô tả toàn diện khi chỉ dựa trên biểu diễn nhúng theo chuỗi. Điều này cho thấy rằng không gian nhúng tuyến tính hóa theo chiều dài chuỗi vẫn chưa phản ánh trọn vẹn các tín hiệu không gian ba chiều của bề mặt tương tác. Thứ hai, ba tầng chọn lọc vẫn sử dụng các tham số ngưỡng cố định (ngưỡng phương sai và ngưỡng tương quan), chưa khai thác các cơ chế tối ưu hoá thích nghi cho từng bộ dữ liệu. Những ngưỡng tĩnh này có thể chưa phản ánh tốt các phân bố dữ liệu có độ dị biến cao như ở các hệ protein người. Thứ ba, cách ghép cặp bằng cách nối trực tiếp hai vector nhúng là tuyến tính và chưa xem xét các dạng kết hợp phi tuyến giàu ngữ nghĩa hơn giữa hai protein. Những hạn chế này chính là tiền đề mở ra nhiều hướng cải tiến tự nhiên cho các nghiên cứu tiếp theo.

5 Kết luận

Nghiên cứu này đã thiết lập E–StackPPI, một khung tiếp cận tinh gọn nhưng có cấu trúc chặt chẽ cho bài toán dự đoán PPI. Thay vì gia tăng độ sâu của các mô hình học sâu hoặc phụ thuộc hoàn toàn vào những bộ đặc trưng thủ công truyền thống, E–StackPPI lựa chọn một cách tiếp cận cân bằng xuất phát từ biểu diễn nhúng của mô hình ngôn ngữ protein hiện đại, sau đó tái cấu trúc không gian biểu diễn thông qua ba tầng chọn lọc đặc trưng liên tiếp. Ba tầng này lần lượt loại bỏ những chiều ít biến thiên, lựa chọn các chiều mang nhiều thông tin và giảm tương quan nội tại, từ đó tạo ra một không gian biểu diễn cô đặc, ít nhiễu và phù hợp hơn cho bài toán phân tách nhị phân.

Kết quả thực nghiệm trên hai bộ dữ liệu DIP–Yeast và DIP–Human là minh chứng rõ ràng cho hiệu quả của mô hình, với độ chính xác trung bình lần lượt 95.67% và 98.60%, hệ số MCC đạt 91.36% và 97.20%, đồng thời hai chỉ số ROC-AUC và PR-AUC đều vượt ngưỡng 99%. Khi đối chiếu với các phương pháp hiện hành được tổng hợp trong công trình của Li và cộng sự, mô hình đề xuất vượt qua toàn bộ 12 phương pháp tiên tiến trên cả hai bộ dữ liệu, đặc biệt ở các chỉ số đánh giá quan trọng như MCC, ROC–AUC và PR–AUC. Điều này khẳng định ưu thế của E–StackPPI không chỉ đến từ năng lực biểu diễn mạnh của mô hình ngôn ngữ protein, mà còn từ

quá trình chọn lọc có chủ đích giúp ổn định ranh giới quyết định giữa các lần kiểm định chéo. Thử nghiệm loại trừ các tầng chọn lọc đặc trưng cũng củng cố nhận định này, khi cấu hình đầy đủ của E–StackPPI ổn định vượt trội so với các biến thể chỉ sử dụng một hoặc hai tầng chọn lọc.

Ý nghĩa của kết quả không chỉ nằm ở việc cải thiện các chỉ số đánh giá, mà còn ở việc củng cố một triết lý thiết kế mô hình khác cho các bài toán sinh học có lượng dữ liệu hạn chế: thay vì gia tăng độ phức tạp của kiến trúc, hiệu năng cao có thể đạt được thông qua việc tổ chức lại và tinh lọc không gian biểu diễn. Đây là hướng đi phù hợp với nhiều bài toán trong khoa học sự sống hiện nay, nơi dữ liệu lớn không phải lúc nào cũng sẵn có và chi phí huấn luyện mô hình sâu là một rào cản thực tế.

Trong tương lai, có ba hướng mở rộng tự nhiên. Thứ nhất, mở rộng quy trình chọn lọc đặc trưng sang các mô hình ngôn ngữ protein thế hệ mới như ESM–3, ProtT5 hoặc các mô hình tích hợp thông tin không gian–thời gian. Thứ hai, áp dụng E–StackPPI cho các bài toán tương tác sinh học khác như protein–DNA, protein–RNA hoặc tương tác thuốc–protein, nơi vấn đề dư thừa đặc trưng cũng là một thách thức nổi bật. Thứ ba, phân tích sâu hơn cấu trúc của không gian biểu diễn trước và sau khi chọn lọc nhằm nhận diện các nhóm chiều mang ý nghĩa sinh học, từ đó mở ra tiềm năng diễn giải mô hình ở mức phân tử.

Mặc dù đạt hiệu năng cao, E–StackPPI vẫn đối mặt với một số giới hạn. Thứ nhất, việc tuyến tính hóa không gian nhúng theo chiều dài chuỗi có thể làm thất thoát các tín hiệu không gian ba chiều quan trọng tại bề mặt tiếp xúc. Thứ hai, các ngưỡng chọn lọc hiện tại vẫn là tham số tĩnh, chưa thích nghi linh hoạt với các phân bố dữ liệu có độ dị biến cao. Thứ ba, chiến lược ghép cặp thông qua nối véc tơ trực tiếp chưa khai thác được các mối quan hệ phi tuyến phức tạp hơn giữa hai protein. Đây là những tiền đề quan trọng để tiếp tục cải thiện trong các nghiên cứu sau.

Tóm lại, E–StackPPI cho thấy rằng việc kết hợp biểu diễn nhúng từ mô hình ngôn ngữ protein với một quy trình chọn lọc đặc trưng được thiết kế chặt chẽ có thể đạt hiệu năng ngang bằng hoặc vượt qua nhiều mô hình học sâu phức tạp, đồng thời bảo đảm tính ổn định và khả năng khái quát tốt. Đây là một nền tảng hứa hẹn cho các ứng dụng dự đoán tương tác sinh học trong tương lai.

Bộ dữ liệu và mã nguồn

Toàn bộ bộ dữ liệu và mã nguồn được công khai tại đây: https://github.com/mxuanvan02/EStack-PPI.

Tài trợ

Kết quả nghiên cứu này được tài trợ bởi đề tài cấp Đại học Huế, "Nghiên cứu phương pháp nâng cao hiệu suất trích xuất và chọn lọc đặc trưng từ trình tự protein trên các bộ dữ liệu lớn", mã số: DHH2024-19-04.

Tài liệu tham khảo

1. I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, "Dip: the database of interacting proteins," *Nucleic acids research*, vol. 28, no. 1, pp. 289–291, 2000.
2. Y. Li, Z. Wang, L.-P. Li, Z.-H. You, W.-Z. Huang, X.-K. Zhan, and Y.-B. Wang, "Robust and accurate prediction of protein–protein interactions by exploiting evolutionary information," *Scientific Reports*, vol. 11, no. 1, p. 16910, 2021.
3. A. E. Modell, S. L. Blosser, and P. S. Arora, "Systematic targeting of protein–protein interactions," *Trends in pharmacological sciences*, vol. 37, no. 8, pp. 702–713, 2016.
4. J. De Las Rivas and C. Fontanillo, "Protein–protein interactions essentials: key concepts to building and analyzing interactome networks," *PLoS computational biology*, vol. 6, no. 6, p. e1000807, 2010.
5. Y. Guo, L. Yu, Z. Wen, and M. Li, "Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences," *Nucleic acids research*, vol. 36, no. 9, pp. 3025–3030, 2008. [6] X.-Y. Pan, Y.-N. Zhang, and H.-B. Shen, "Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features," *Journal of proteome research*, vol. 9, no. 10, pp. 4992–5001, 2010.
6. Z.-H. You, Y.-K. Lei, L. Zhu, J. Xia, and B. Wang, "Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis," *BMC bioinformatics*, vol. 14, no. Suppl 8, p. S10, 2013.
7. Y.-A. Huang, Z.-H. You, X. Chen, K. Chan, and X. Luo, "Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding," *BMC bioinformatics*, vol. 17, no. 1, p. 184, 2016.
8. C. Chen, Q. Zhang, Q. Ma, and B. Yu, "Lightgbm-ppi: Predicting protein-protein interactions through lightgbm with multi-information fusion," *Chemometrics and Intelligent Laboratory Systems*, vol. 191, pp. 54–64, 2019.
9. A. Sharma and B. Singh, "Ae-lgbm: Sequence-based novel approach to detect interacting protein pairs via ensemble of autoencoder and lightgbm," *Computers in Biology and Medicine*, vol. 125, p. 103964, 2020.
10. B. Yu, C. Chen, H. Zhou, B. Liu, and Q. Ma, "Gtb-ppi: predict protein–protein interactions based on l1-regularized logistic regression and gradient tree boosting," *Genomics, proteomics & bioinformatics*, vol. 18, no. 5, pp. 582–592, 2020.
11. Y. Tian, J. Zhou, and J. Wang, "An ensemble lightgbm framework for robust prediction of protein–protein interactions," *Frontiers in Genetics*, vol. 14, p. 1163943, 2023.
12. S. Hashemifar, B. Neyshabur, A. A. Khan, and J. Xu, "Predicting protein–protein interactions through sequence-based deep learning," *Bioinformatics*, vol. 34, no. 17, pp. i802–i810, 2018.

13. M. Chen, C. J.-T. Ju, G. Zhou, X. Chen, T. Zhang, K.-W. Chang, C. Zaniolo, and W. Wang, "Multifaceted protein–protein interaction prediction based on siamese residual rcnn," *Bioinformatics*, vol. 35, no. 14, pp. i305–i314, 2019.

14. S. Sledzieski, R. Singh, L. Cowen, and B. Berger, "Sequence-based prediction of protein- protein interactions: a structure-aware interpretable deep learning model," *BioRxiv*, pp. 2021– 01, 2021.

15. T. Sun, B. Zhou, L. Lai, and J. Pei, "Sequence-based prediction of protein protein interaction using a deep-learning algorithm," *BMC bioinformatics*, vol. 18, no. 1, p. 277, 2017.

16. K. D. Truong, X. V. Mai, and T. Tri Nguyen, "Predicting protein-protein interactions: A case study using hilbert transform with combining ensemble learning model," in *International Conference on Computational Intelligence in Engineering Science*, pp. 53–64, Springer, 2025.

17. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, *et al.*, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023.

18. A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, *et al.*, "Prottrans: Toward understanding the language of life through self-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 10, pp. 7112–7127, 2021.

19. T. H. Dang and T. A. Vu, "xcapt5: protein–protein interaction prediction using deep and wide multi-kernel pooling convolutional neural networks with protein language model," *BMC bioinformatics*, vol. 25, no. 1, p. 106, 2024.

20. J. Ding, J. Du, H. Wang, and S. Xiao, "A novel two-stage feature selection method based on random forest and improved genetic algorithm for enhancing classification in machine learning," *Scientific Reports*, vol. 15, no. 1, p. 16828, 2025.

21. Y. Murakami and K. Mizuguchi, "Recent developments of sequence-based prediction of protein–protein interactions," *Biophysical Reviews*, vol. 14, no. 6, pp. 1393–1411, 2022.

22. L. Xian and Y. Wang, "Advances in computational methods for protein–protein interaction prediction," *Electronics*, vol. 13, no. 6, p. 1059, 2024.

23. M. Xu, Z. Zhang, J. Lu, Z. Zhu, Y. Zhang, M. Chang, R. Liu, and J. Tang, "Peer: a comprehensive and multi-task benchmark for protein sequence understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 35156–35173, 2022.

24. X. Du, S. Sun, C. Hu, Y. Yao, Y. Yan, and Y. Zhang, "Deepppi: boosting prediction of protein– protein interactions with deep neural networks," *Journal of chemical information and modeling*, vol. 57, no. 6, pp. 1499–1510, 2017.

25. L. Wong, Z.-H. You, S. Li, Y.-A. Huang, and G. Liu, "Detection of protein-protein interactions from amino acid sequences using a rotation forest model with a novel pr-lpq descriptor," in *International conference on intelligent computing*, pp. 713–720, Springer, 2015.

26. Y. Wang, Z.-H. You, S. Yang, X. Li, T.-H. Jiang, and X. Zhou, "A high efficient biological language model for predicting protein–protein interactions," *Cells*, vol. 8, no. 2, p. 122, 2019.

27. Z.-H. You, L. Zhu, C.-H. Zheng, H.-J. Yu, S.-P. Deng, and Z. Ji, "Prediction of protein- protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set," *BMC bioinformatics*, vol. 15, no. Suppl 15, p. S9, 2014.

28. J.-Y. An, Y. Zhou, Y.-J. Zhao, and Z.-J. Yan, "An efficient feature extraction technique based on local coding pssm and multifeatures fusion for predicting protein-protein interactions," *Evolutionary Bioinformatics*, vol. 15, p. 1176934319879920, 2019.

29. Y. Z. Zhou, Y. Gao, and Y. Y. Zheng, "Prediction of protein-protein interactions using local description of amino acid sequence," in *Advances in Computer Science and Education Applications: International Conference, CSE 2011, Qingdao, China, July 9-10, 2011. Proceedings, Part II*, pp. 254–262, Springer, 2011.

30. L. Yang, J.-F. Xia, and J. Gui, "Prediction of protein-protein interactions from protein sequence using local descriptors," *Protein and peptide letters*, vol. 17, no. 9, pp. 1085–1090, 2010.

31. Y. Ding, J. Tang, and F. Guo, "Predicting protein-protein interactions via multivariate mutual information of protein sequences," *BMC bioinformatics*, vol. 17, no. 1, p. 398, 2016.