ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

TRẦN THỊ XUÂN

NÂNG CAO HIỆU QUẢ PHÂN TÍCH PROTEIN SỬA ĐỔI SAU DỊCH MÃ TRÊN CƠ SỞ KẾT HỢP MÔ HÌNH HỌC MÁY VÀ XỬ LÝ NGÔN NGỮ TỰ NHIÊN

TÓM TẮT

LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYÊN - NĂM 2025

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

TRẦN THỊ XUÂN

NÂNG CAO HIỆU QUẢ PHÂN TÍCH PROTEIN SỬA ĐỔI SAU DỊCH MÃ TRÊN CƠ SỞ KẾT HỢP MÔ HÌNH HỌC MÁY VÀ XỬ LÝ NGÔN NGỮ TỰ NHIÊN

Ngành: Khoa học máy tính
Mã số: 9.48.01.01

TÓM TẮT

LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

TẬP THỂ HƯỚNG DẪN KHOA HỌC:

1. PGS.TS. LÊ NGUYỄN QUỐC KHÁNH
2. TS. NGUYỄN VĂN NÚI

THÁI NGUYÊN - NĂM 2025

MỞ ĐẦU

1. Tính cấp thiết của đề tài

Bối cảnh khoa học và thực tiễn:

Sửa đổi sau dịch mã (Post-Translational Modification – PTM) là những biến đổi hóa học diễn ra sau khi quá trình tổng hợp protein hoàn tất. Các dạng PTM phổ biến như glycosyl hóa, phosphoryl hóa, ubiquitin hóa, acetyl hóa, lipid hóa, hay phân giải protein. Có vai trò đặc biệt quan trọng trong việc điều chỉnh cấu trúc, chức năng và hoạt động sinh học của protein.

PTM tác động sâu rộng đến nhiều quá trình sinh học then chốt, chẳng hạn như truyền tín hiệu tế bào, điều hòa miễn dịch, và biểu hiện gen. Sự sai lệch trong quá trình PTM liên quan trực tiếp đến nhiều bệnh lý nguy hiểm như ung thư, rối loạn thần kinh và bệnh truyền nhiễm. Do đó, việc xác định chính xác các vị trí PTM trong chuỗi protein là một nhiệm vụ có ý nghĩa quan trọng trong nghiên cứu y sinh, hỗ trợ làm sáng tỏ cơ chế phân tử, phát triển thuốc và liệu pháp điều trị mới.

Khối phổ (Mass Spectrometry – MS) được coi là phương pháp tiêu chuẩn vàng để phát hiện PTM. Tuy nhiên, kỹ thuật này thường yêu cầu quy trình thí nghiệm phức tạp, tốn kém và mất nhiều thời gian, đồng thời khó mở rộng quy mô. Do đó, sự phát triển của các phương pháp tính toán có khả năng dự đoán vị trí PTM một cách nhanh chóng, chi phí thấp và hiệu quả là hết sức cần thiết nhằm hỗ trợ cho các nghiên cứu trong lĩnh vực y sinh.

Sự phát triển của các phương pháp tính toán:

Trong hơn hai thập kỷ qua, các phương pháp tính toán đã góp phần quan trọng trong dự đoán vị trí PTM, đặc biệt với ba hướng tiếp cận nổi bật: học máy truyền thống, học sâu, và xử lý ngôn ngữ tự nhiên (NLP) và mô hình ngôn ngữ protein (PLMs).

(i) Học máy truyền thống (Machine Learning): Các mô hình học máyđược sử dụng xấy dựng các mô hình dự đoán PTM như SVM, Random Forest, XGBoost hay kNN, tuy nhiên các mô hình này thường dựa trên tập đặc trưng thủ công được thiết kế từ kiến thức sinh học (ví dụ: PseAAC, CKSAAP, BE, PsePSSM). Hướng nghiên cứu này có ưu điểm nổi bật là dễ huấn luyện, triển khai nhanh, và có khả năng diễn giải tốt, đặc biệt phù hợp khi làm việc với dữ liệu nhỏ. Tuy nhiên, nhược điểm lớn là phụ thuộc nhiều vào đặc trưng thủ công vốn mang tính chủ quan và dễ bỏ sót các tín hiệu ngữ cảnh quan trọng, khiến khả năng tổng quát hóa bị hạn chế.

(ii) Học sâu (Deep Learning): Mô hình dự đoán PTM được phát triển dựa trên các

kiến trúc mạng học sâu như CNN, LSTM, Bi-LSTM hoặc các mô hình học sâu lai. Học sâu cho phép tự động trích xuất đặc trưng từ dữ liệu thô và mô hình hóa mối quan hệ phi tuyến phức tạp trong chuỗi protein. Các nghiên cứu gần đây cho thấy mô hình học sâu thường vượt trội hơn so với học máy truyền thống về hiệu quả dự đoán. Tuy nhiên, chúng thường đòi hỏi tập dữ liệu huấn luyện quy mô lớn và tiêu tốn nhiều tài nguyên tính toán. Trong điều kiện dữ liệu sinh học thường hạn chế và mất cân bằng, mô hình học sâu dễ gặp phải vấn đề quá khớp, làm giảm khả năng ứng dụng thực tiễn.

(iii) Xử lý ngôn ngữ tự nhiên (NLP) và mô hình ngôn ngữ protein (PLMs):

Trong hướng tiếp cận này, chuỗi protein được xem như một "ngôn ngữ sinh học", trong đó mỗi axit amin tương ứng với một token, và ngữ cảnh xung quanh token quyết định chức năng sinh học của nó. Quan niệm này mở ra khả năng ứng dụng các kỹ thuật NLP vào dự đoán PTM. Các mô hình ngôn ngữ lớn như BERT và T5 được sử dụng để trích xuất các embedding ngữ cảnh, sau đó các embedding này được đưa vào làm đặc trưng cho các mô hình học máy hoặc học sâu, xây dựng nên các mô hình dự đoán PTM hiệu quả.

Ngoài ra, một số mô hình PTM còn khai thác các mô hình tiền huấn luyện dựa trên BERT chuyên biệt cho protein, chẳng hạn như ProteinBERT, điển hình là DeepPTM. Tuy nhiên, một hạn chế quan trọng là chi phí tính toán rất cao, gây khó khăn trong triển khai thực tế, đặc biệt khi dữ liệu hạn chế hoặc tài nguyên tính toán bị giới hạn.

Các thách thức và khoảng trống nghiên cứu:

Mặc dù đã đạt được nhiều tiến bộ, các nghiên cứu dự đoán vị trí PTM hiện nay vẫn tồn tại một số thách thức sau:

- Phụ thuộc đặc trưng thủ công: Phần lớn các phương pháp học máy truyền thống vẫn dựa nhiều vào đặc trưng do con người thiết kế, mang tính chủ quan và thiếu khả năng khái quát khi áp dụng cho loài mới hoặc dạng PTM khác.

- Nguy cơ quá khớp do dữ liệu hạn chế: Trong bối cảnh dữ liệu PTM thường nhỏ và mất cân bằng, các mô hình học sâu dễ bị quá khớp, làm giảm tính tổng quát trong thực tiễn.

- Chi phí dữ liệu và tài nguyên lớn: Các mô hình học sâu và PLMs/LLMs yêu cầu tập dữ liệu khổng lồ và hạ tầng mạnh, khó áp dụng trong môi trường nghiên cứu hạn chế về tính toán.

- Chưa khai thác kỹ thuật chắt lọc tri thức (Knowledge Distillation-KD). KD đã chứng minh hiệu quả trong thị giác máy tính (Computer Vision) và xử lý ngôn ngữ tự nhiên (NLP), cho phép xây dựng mô hình gọn nhẹ nhưng vẫn duy trì hiệu suất cao. Tuy nhiên, đến nay chưa có công trình nào áp dụng kỹ thuật học chắt lọc tri thức vào dự đoán

PTM, trong khi đây là một hướng hứa hẹn phù hợp với dữ liệu hạn chế và môi trường tài nguyên giới hạn.

Xuất phát từ tầm quan trọng của việc xác định chính xác vị trí PTM trong nghiên cứu y sinh, cùng với nhu cầu phát triển các phương pháp tính toán tiên tiến và những khoảng trống nghiên cứu đã phân tích, NCS lựa chọn đề tài "Nâng cao hiệu quả phân tích protein sửa đổi sau dịch mã trên cơ sở kết hợp mô hình học máy và xử lý ngôn ngữ tự nhiên" làm luận án tiến sĩ ngành Khoa học máy tính.

2. Đối tượng và phạm vi nghiên cứu

(1) Đối tượng thứ nhất là các protein sửa đổi sau dịch mã:

Hiện tại, có hơn 600 loại PTM khác nhau đã được phát hiện và định danh. Mong muốn của NCS là có thể thực hiện nghiên cứu được với nhiều loại PTM khác nhau nhằm bổ sung, góp phần làm giàu tri thức, sự hiểu biết của con người đối với tất cả các loại PTM hiện có. Tuy nhiên, trong phạm vi luận án này, nghiên cứu tập trung vào ba loại phổ biến và có dữ liệu tương đối đầy đủ và còn khoảng trống nghiên cứu: SUMOylation, Succinylation và Ubiquitination.

Ngoài ra, qua khảo sát, cấu trúc protein bậc cao (cấu trúc bậc 2,3,4- thường được lưu trữ dưới dạng ảnh 3D) trong các ngân hàng Protein(UniProt, NCBI, Ensembl...) còn thiếu, chưa đầy đủ và rất tốn kém bộ nhớ để lưu trữ; hơn nữa hầu hết dữ liệu protein hiện nay được lưu trữ dưới dạng chuỗi FASTA (Protein bậc 1). Dạng biểu diễn này không chỉ phổ biến mà còn tiết kiệm tài nguyên và phù hợp với các kỹ thuật học máy, học sâu hiện đại và NLP. Vì vậy, luận án lựa chọn cấu trúc protein bậc 1 làm đầu vào để phát triển mô hình dự đoán vị trí PTM với hiệu năng cao cho ba loại nêu trên.

(2) Đối tượng thứ hai là mô hình dự đoán vị trí PTM dựa trên mô hình học máy kết hợp với xử lý ngôn ngữ tự nhiên:

Kỹ thuật phổ biến để dự đoán vị trí PTM, có độ chính xác cao hiện nay chính là kỹ thuật khối phổ và giải trình tự. Tuy nhiên, kỹ thuật MS này có chi phí rất lớn, thời gian thực hiện lâu, và đặc biệt là khó áp dụng với nhiều protein cùng lúc. Chính vì vậy, việc nghiên cứu các mô hình dự đoán PTM dựa trên mô hình học máy, kết hợp với NLP là một cách tiếp cận phù hợp bởi nó khai thác được những tiến bộ của công nghệ thông tin, các mô hình học máy và kỹ thuật NLP nhằm giúp ngắn thời gian hỗ trợ cho các nhà sinh/y học đưa ra những kết luận nhanh và chính xác, phù hợp nhu cầu và xu hướng phát triển hiện nay.

3. Phương pháp nghiên cứu

Để đạt được mục tiêu nghiên cứu, luận án triển khai song song hai hướng chính, đó là: (1) Nghiên cứu cơ sở lý thuyết để đề xuất mô hình mới và (2) Nghiên cứu thực nghiệm nhằm kiểm chứng hiệu quả các mô hình này.

(1) Về lý thuyết, luận án kế thừa và phát triển các phương pháp hiện đại của khoa học dữ liệu và trí tuệ nhân tạo, bao gồm: Học máy tổ hợp (Ensemble Learning) nhằm khai thác ưu thế của nhiều bộ phân loại để nâng cao độ tin cậy dự đoán. Xử lý ngôn ngữ tự nhiên để biểu diễn "ngôn ngữ protein" và trích xuất ngữ nghĩa, ngữ cảnh từ chuỗi axit amin; Các kiến trúc học sâu lai (Hybrid Deep Learning) kết hợp CNN và LSTM/Bi-LSTM nhằm tận dụng đồng thời khả năng phát hiện đặc trưng cục bộ và quan hệ tuần tự dài hạn; Học chắt lọc tri thức để xây dựng các mô hình gọn nhẹ, thích ứng với dữ liệu hạn chế;

(2) Về thực nghiệm, các mô hình được huấn luyện và kiểm định trên dữ liệu PTM thực tế, sau đó so sánh với các phương pháp tiên tiến hiện có. Kết quả đánh giá giúp khẳng định tính khả thi, hiệu quả và ý nghĩa ứng dụng của các mô hình đề xuất trong việc dự đoán vị trí sửa đổi sau dịch mã trên protein.

4. Các đóng góp của luận án

Luận án tập trung nghiên cứu và đề xuất phương pháp dự đoán ba loại PTM phổ biến, bao gồm: SUMOylation, Succinylation và Ubiquitination. Trên cơ sở kết hợp các phương pháp truyền thống, học máy, học sâu và kỹ thuật xử lý ngôn ngữ tự nhiên, luận án đã giải quyết được các mục tiêu đặt ra của đề tài, đề xuất được những mô hình cải tiến với hiệu suất cao. Trong quá trình thực hiện luận án NCS đã công bố 08 bài báo khoa học trên các tạp chí và hội thảo chuyên ngành trong nước và quốc tế. Trong đó, 06 công bố có nội dung gắn trực tiếp với tên đề tài, phản ánh rõ ràng các kết quả nghiên cứu chính của luận án, được trình bày đầy đủ và chi tiết trong các chương nội dung. Bên cạnh đó, 02 công bố khác có nội dung liên quan và hỗ trợ, góp phần làm phong phú thêm cho luận án, mở rộng phạm vi nghiên cứu, đồng thời khẳng định khả năng ứng dụng, tính tổng quát và hướng phát triển của kết quả nghiên cứu, các công bố bổ trợ này không đi sâu vào nội dung chính của luận án nhưng có ý nghĩa bổ trợ về phương pháp, kỹ thuật và lĩnh vực ứng dụng.

Danh mục chi tiết các công bố đã được liệt kê trong phần Danh mục công trình khoa học kèm theo luận án.

Luận án có ba đóng góp chính sau:

(1) Cơ sở lý luận và tổng quan hệ thống: Luận án đã hệ thống hóa, phân tích và so sánh các phương pháp từ truyền thống, học máy tổ hợp, học sâu lai, kỹ thuật NLP

trong bài toán dự đoán PTM, qua đó xây dựng nền tảng khoa học vững chắc cho các nghiên cứu tiếp theo.

(2) Khai thác NLP cho dữ liệu protein: Luận án đã chứng minh khả năng ứng dụng và hiệu quả của các kỹ thuật NLP trong việc biểu diễn ngữ cảnh của chuỗi protein, giúp vượt qua hạn chế của đặc trưng thủ công và nâng cao độ chính xác trong dự đoán.

(3) Đề xuất và phát triển mô hình mới: Luận án đã đề xuất bốn mô hình PTM với hiệu suất cao, trong đó có các mô hình lai kết hợp học sâu với NLP và đặc biệt là mô hình áp dụng học chắt lọc tri thức cho Ubiquitination, phù hợp với bối cảnh dữ liệu hạn chế và môi trường tính toán hạn chế. Cụ thể, bốn đề xuất chính gồm:

- Đề xuất mô hình dự đoán vị trí PTM (SUMOylation) dựa trên học máy tổ hợp và các đặc trưng lai ghép.

- Đề xuất hai mô hình dự đoán vị trí PTM (SUMOylation và Succinylation) dựa trên kỹ thuật học sâu lai ghép và kỹ thuật xử lý ngôn ngữ tự nhiên.

- Đề xuất mô hình dự đoán PTM (Ubiquitination) dựa trên học chắt lọc tri thức và kỹ thuật xử lý ngôn ngữ tự nhiên.

5. Bố cục của luận án

Luận án bao gồm các phần: Mở đầu, 4 chương nội dung chính, kết luận và hướng phát triển, danh mục các công trình khoa học đã công bố và danh mục tài liệu tham khảo. Nội dung chính của 4 chương được tóm tắt như sau:

Chương 1. Tổng quan dự đoán vị trí sửa đổi sau dịch mã trong chuỗi protein và các kiến thức nền tảng

Trong chương này, NCS trình bày các kiến thức nền tảng về protein và protein sửa đổi sau dịch mã (PTM), vai trò của việc xây dựng mô hình dự đoán vị trí PTM hiệu suất cao. Phần tiếp theo của chương trình bày về bài toán dự đoán vị trí PTM, các bước xây dựng mô hình dự đoán vị trí PTM, các phương pháp mã hoá đặc trưng protein hiện nay, tổng quan tình hình nghiên cứu dự đoán vị trí PTM, một số hạn chế và đề xuất hướng nghiên cứu.

Chương 2. Mô hình học máy tổ hợp dự đoán vị trí sửa đổi sau dịch mã trong chuỗi protein

Trong chương này, NCS trình bày về phương pháp dự đoán vị trí sửa đổi sau dịch mã trên protein dựa trên kỹ thuật học máy tổ hợp. Phương pháp đề xuất sử dụng ba mô hình thành phần gồm Random Forest (RF), Extreme Gradient Boosting (XGBoost) và Support Véc tơ Machine (SVM), được huấn luyện độc lập và kết hợp kết quả dự đoán bằng chiến lược trung bình có trọng số (Weighted Average Voting). Cách tiếp cận này

giúp khai thác thế mạnh riêng biệt của từng mô hình học máy, từ đó cải thiện độ chính xác và tính ổn định trong dự đoán. Các kết quả thực nghiệm cho thấy phương pháp tổ hợp mang lại hiệu suất vượt trội so với các mô hình đơn lẻ.

Chương 3. Mô hình học sâu lai kết hợp xử lý ngôn ngữ tự nhiên dự đoán vị trí sửa đổi sau dịch mã trong chuỗi protein

Trong chương này, NCS giới thiệu mô hình học sâu lai tích hợp với kỹ thuật NLP nhằm nâng cao khả năng dự đoán vị trí PTM. Mô hình được xây dựng dựa trên sự kết hợp giữa mạng CNN1D để khai thác đặc trưng cục bộ và mạng LSTM/Bi-LSTM để học ngữ cảnh tuần tự trong chuỗi protein. Kỹ thuật NLP được áp dụng để biểu diễn chuỗi axit amin dưới dạng véc tơ, giúp mô hình học được các thông tin ẩn trong chuỗi sinh học một cách hiệu quả hơn. Thực nghiệm trên các tập dữ liệu PTM cho thấy mô hình lai cho kết quả chính xác và ổn định hơn so với các mô hình đơn cấu trúc, đồng thời khẳng định được khả năng tổng quát hóa của phương pháp.

Chương 4. Mô hình học chắt lọc tri thức kết hợp xử lý ngôn ngữ tự nhiên dự đoán vị trí sửa đổi sau dịch mã trong chuỗi protein

Trong chương này, NCS trình bày về đề xuất một phương pháp dự đoán vị trí PTM dựa trên kiến trúc học chắt lọc tri thức kết hợp NLP. Phương pháp sử dụng mô hình Giáo viên-Học viên, trong đó mô hình Giáo viên có cấu trúc lớn hơn và được huấn luyện trên tập dữ liệu đa loài để rút trích tri thức, sau đó truyền lại cho mô hình Học viên nhỏ gọn hơn, được huấn luyện trên tập dữ liệu chuyên biệt cho một loài. Việc kết hợp kỹ thuật NLP giúp mã hóa chuỗi protein thành không gian véc tơ, hỗ trợ quá trình học hiệu quả từ dữ liệu thô. Thực nghiệm chứng minh rằng mô hình Học viên không những giảm đáng kể số lượng tham số mà vẫn duy trì được hiệu suất cao, thậm chí vượt trội so với các mô hình không sử dụng học chắt lọc tri thức trong cùng điều kiện huấn luyện.

Phần cuối cùng của luận án NCS trình bày các kết luận chính và đề xuất hướng nghiên cứu tiếp theo.

CHƯƠNG 1. TỔNG QUAN DỰ ĐOÁN VỊ TRÍ SỬA ĐỔI SAU DỊCH MÃ TRONG CHUỖI PROTEIN VÀ CÁC KIẾN THỨC NỀN TẢNG

Trong chương 1, NCS trình bày một số kiến thức nền tảng về protein và protein sửa đổi sau dịch mã (Post-translational modification - PTM), vai trò của việc xây dựng mô hình dự đoán vị trí PTM hiệu suất cao. Phần tiếp theo của chương trình bày về bài toán dự đoán vị trí PTM, các bước xây dựng mô hình dự đoán vị trí PTM, tổng quan tình hình nghiên cứu trong bối cảnh AI (SOTA), các khoảng trống nghiên cứu, thách thức của mô

hình dự đoán vị trí PTM, hướng nghiên cứu trong luận án. Phần cuối của chương trình bày về môi trường sử dụng để thực nghiệm, phương pháp đánh giá mô hình đề xuất. Một phần kết quả nghiên cứu được đăng trong bài báo tổng quan các mô hình học máy trên tạp chí Expert Opinion on Drug Metabolism & Toxicology, SCIE Q1, IF=3.9 (CT1) và hội thảo iFUZZY, Kagawa, Japan (CT2).

1.1 Giới thiệu chung

1.1.1 Protein

1.1.2 Protein sửa đổi sau dịch mã

Sau quá trình dịch mã, các protein mới tổng hợp không hoàn toàn ở trạng thái hoạt động mà thường trải qua các biến đổi hóa học bổ sung gọi là sửa đổi sau dịch mã (Post-Translational Modifications – PTMs). Đây là quá trình trong đó các nhóm chức hóa học (như phosphate, ubiquitin, acetyl, methyl, succinyl...) được gắn hoặc loại bỏ khỏi chuỗi protein, thông qua hệ thống enzyme chuyên biệt.

Mỗi loại PTM chỉ xảy ra tại một hoặc một số axit amin nhất định, điều này dẫn đến hai bài toán phân lớp: Phân lớp nhị phân và phân lớp đa phân.

1.1.3 Vai trò của bài toán dự đoán vị trí PTM và các phương pháp xác dự đoán vị trí PTM hiện nay

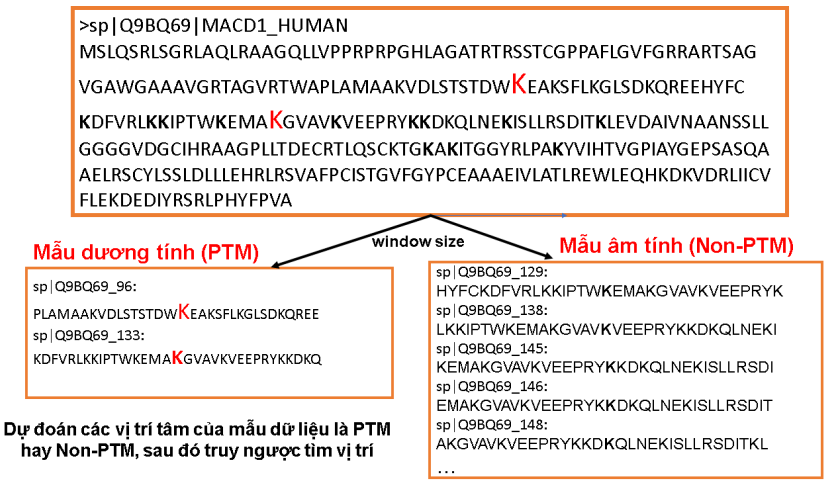
1.2 Bài toán dự đoán vị trí PTM dựa trên học máy

Đầu vào: Chuỗi protein bậc 1, kèm theo các vị trí nghi ngờ có khả năng bị sửa đổi sau dịch mã.

Đầu ra: Nhãn dự đoán cho từng vị trí nghi ngờ, xác định xem đó là vị trí PTM hay Non_PTM (có thể kèm theo xác suất dự đoán).

Mục tiêu: Đề xuất một mô hình học máy hiệu suất cao, có khả năng dự đoán chính xác các vị trí PTM trên chuỗi protein.

Quá trình biến đổi sau dịch mã là một cơ chế sinh học quan trọng, xảy ra tại một hoặc một số axit amin cụ thể trong chuỗi protein. Trong thực tế, chỉ một phần nhỏ các axit amin trên chuỗi protein chịu ảnh hưởng của việc sửa đổi sau dịch mã dưới tác động của enzim như phosphoryl hóa, ubiquitin hóa, succinyl hóa,... Mặt khác, độ dài của các chuỗi protein là khác nhau có chuỗi dài vài nghìn axit amin, các vị trí sửa đổi trên chuỗi protein thường không được biết trước, do đó cần áp dụng kỹ thuật cửa sổ trượt (sliding window) để tạo ra các phân đoạn peptide có độ dài cố định, trong đó tâm cửa sổ tương ứng với vị trí nghi ngờ bị sửa đổi.



Hình 1.1 Chuyển từ bài toán tìm vị trí nghi ngờ sửa đổi sau dịch mã, vị trí nghi ngờ đó nằm ở thứ tự bao nhiêu trong chuỗi về bài toán phân loại nhị phân

Mỗi phân đoạn được xem như một mẫu đầu vào cho mô hình học máy. Nếu axit amin tại tâm cửa sổ là vị trí PTM đã biết, mẫu đó được gán nhãn dương tính (1); ngược lại, nếu không bị sửa đổi, mẫu được gán nhãn âm tính (0). Bài toán vì vậy được quy về một bài toán phân loại nhị phân, với đầu vào là các phân đoạn đã gán nhãn, và đầu ra là xác suất của axit amin ở trung tâm của cửa sổ là PTM hay Non-PTM. Dựa trên kết quả đầu ra của mô hình, có thể ánh xạ ngược lại các tâm cửa sổ về vị trí tương ứng trên chuỗi protein gốc, từ đó xác định các vị trí có khả năng bị sửa đổi sau dịch mã. Cách tiếp cận này cho phép xây dựng các công cụ dự đoán vị trí PTM với độ chính xác cao, góp phần hỗ trợ các nghiên cứu y sinh và thiết kế thuốc.

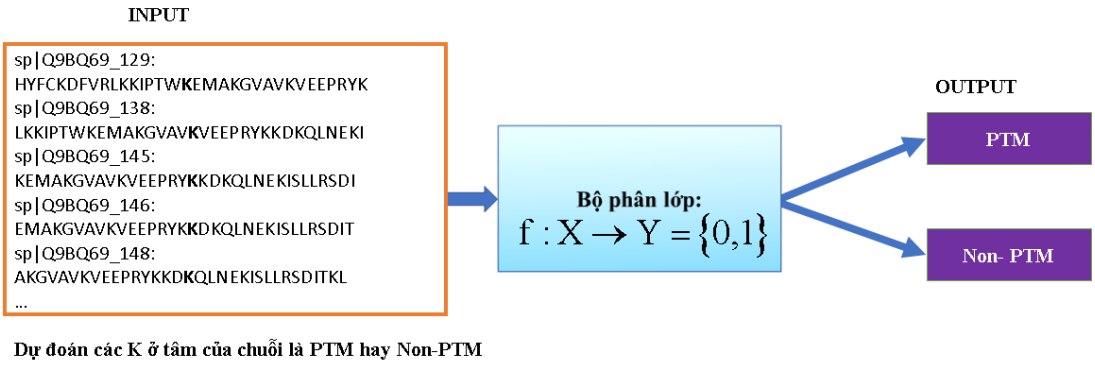
Từ đó, bài toán dự đoán vị trí PTM được chuyển thành một bài toán phân lớp nhị phân, trong đó:

Đầu vào: Là một tập các đoạn peptide $X = \{x_1, x_2, \ldots, x_n\}$, trong đó mỗi x_i là một đoạn peptide cố định có độ dài là w (w là kích thước cửa sổ trượt), được trích xuất từ các chuỗi protein ban đầu bằng phương pháp cửa sổ trượt. Vị trí trung tâm của mỗi đoạn peptide được giả định là vị trí nghi ngờ có thể xảy ra sự kiện sửa đổi sau dịch mã (PTM).

Đầu ra: Vị trí trung tâm của mỗi đoạn peptide là PTM hay Non-PTM, có thể phát biểu đầu ra như một tập các nhãn tương ứng $Y = \{y_1, y_2, \ldots, y_n\}$, trong đó:

Mục tiêu: Xây dựng mô hình phân lớp nhị phân:

$$f : \mathscr{X} \to \mathscr{Y} \tag{1.1}$$



Hình 1.2 Mô tả bài toán dự đoán vị trí PTM

1.3 Xây dựng mô hình dự đoán vị trí PTM

1.3.1 Thu thập và tiền xử lý dữ liệu

1.3.2 Phương pháp mã hoá và trích chọn đặc trưng

1.3.2.1 Phương pháp trích chọn đặc trưng dựa trên chuỗi

1.3.2.2 Phương pháp mã hoá dựa trên kỹ thuật xử lý ngôn ngữ tự nhiên

1.3.3 Xây dựng mô hình

Luận án được thực hiện theo hướng thực nghiệm, tập trung xây dựng, thử nghiệm và đánh giá các mô hình học máy, học sâu để xác định kiến trúc tối ưu cho bài toán dự đoán vị trí PTM. Các mô hình sau được lựa chọn dựa trên khả năng xử lý dữ liệu chuỗi và hiệu quả dự đoán:

- Mô hình học máy truyền thống: SVM, XGBoost, RF.

- Mô hình học sâu: CNN1D, LSTM, Bi-LSTM và các biến thể kết hợp theo hướng mô hình học sâu lai, học chắt lọc tri thức.

1.3.4 Lựa chọn các tham số trong quá trình huấn luyện mô hình dự đoán

1.3.5 Đánh giá mô hình

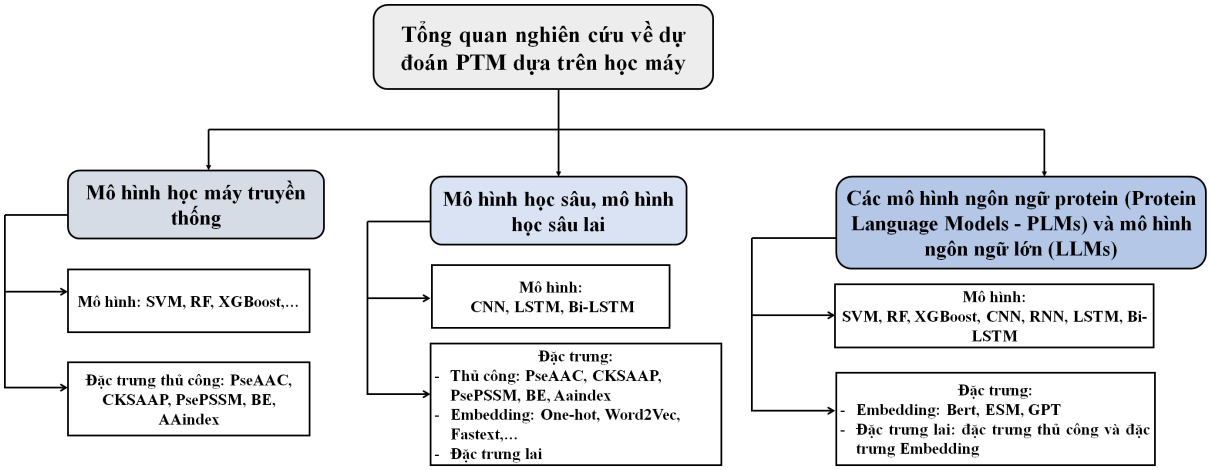
1.3.6 Lựa chọn mô hình

1.3.7 Các yêu cầu hệ thống và môi trường cài đặt

1.4 Thách thức của các mô hình dự đoán vị trí PTM

1.5 Tổng quan nghiên cứu về dự đoán PTM và các phương pháp tiên tiến hiện nay

Hình 1.3 là sơ đồ bức tranh tổng thể về ba hướng nghiên cứu chính trong dự đoán PTM hiện nay.



Hình 1.3 Sơ đồ tổng quan các hướng tiếp cận trong dự đoán vị trí PTM

Khoảng trống nghiên cứu:

Mặc dù các phương pháp học máy, học sâu và mô hình ngôn ngữ protein đã đạt được nhiều tiến bộ quan trọng trong dự đoán vị trí PTM, nhưng hiện tại vẫn tồn tại một số hạn chế cần tiếp tục giải quyết. Các khoảng trống nghiên cứu có thể tóm lược như sau:

- Phụ thuộc đặc trưng thủ công: hạn chế tính tổng quát khi áp dụng trên PTM hoặc loài mới.Các đặc trưng này bao gồm PseAAC, BE, CKSAAP, AAindex hay PSSM – được rút trích dựa trên kiến thức sinh học và thống kê chuỗi – tuy có hiệu quả nhất định nhưng lại mang tính cục bộ, thiếu khả năng biểu diễn các mối quan hệ ngữ cảnh sâu và có thể hạn chế khả năng tổng quát của mô hình. Việc phụ thuộc vào đặc trưng thủ công cũng gây ra thách thức khi áp dụng mô hình cho các loại PTM khác nhau hoặc các loài chưa được nghiên cứu kỹ, do cần thiết kế lại đặc trưng tương ứng. Điều này cho thấy nhu cầu phát triển các mô hình có khả năng tự động học đặc trưng từ dữ liệu thô (end-to-end),

hướng đến giảm thiểu tối đa sự can thiệp của con người trong quá trình tiền xử lý dữ liệu.

- Yêu cầu dữ liệu và tài nguyên lớn: học sâu và PLMs cần tập dữ liệu khổng lồ và hạ tầng mạnh, độ phức tạp cao và nguy cơ quá khớp khi làm việc với các tập dữ liệu PTM có kích thước nhỏ. Nguy cơ quá khớp: dữ liệu PTM thường nhỏ, mất cân bằng, dễ dẫn đến mô hình kém tổng quát.

- Chưa khai thác chắt lọc tri thức (Knowledge Distillation): mặc dù kỹ thuật này đã được ứng dụng thành công trong các lĩnh vực thị giác máy tính và xử lý ngôn ngữ tự nhiên, nhưng đến nay chưa có công trình nào áp dụng cho dự đoán PTM. Đây là một hướng nghiên cứu đầy tiềm năng, đặc biệt phù hợp trong bối cảnh dữ liệu hạn chế và môi trường tài nguyên giới hạn.

1.6 Hướng nghiên cứu trong luận án

Xuất phát từ những khoảng trống trong nghiên cứu dự đoán PTM, nhu cầu phát triển các mô hình có khả năng tự động học đặc trưng từ dữ liệu thô (end-to-end), hướng đến giảm thiểu tối đa sự can thiệp của con người trong quá trình tiền xử lý dữ liệu đặc biệt là sự phụ thuộc vào các đặc trưng trích xuất thủ công và việc chưa khai thác tiềm năng của kỹ thuật học sâu lai, kỹ thuật chắt lọc tri thức, nghiên cứu này được triển khai theo ba giai đoạn chính nhằm giải quyết tuần tự các thách thức hiện tại.

Thứ nhất, trong bối cảnh dữ liệu huấn luyện hạn chế, để nâng cao khả năng khai thác thông tin từ cả đặc trưng sinh học thủ công và biểu diễn theo ngữ cảnh từ NLP, nghiên cứu đã đề xuất một mô hình học máy tổ hợp, sử dụng đặc trưng lai ghép giữa hai nhóm: (i) đặc trưng thủ công (AAindex, BLOSUM62, CKSAAP) và (ii) đặc trưng học được từ mô hình NLP (Word2Vec). Sự kết hợp này giúp tận dụng đồng thời kiến thức sinh học chuyên biệt và khả năng học mẫu ngữ cảnh từ chuỗi protein. Ngoài ra, thiết kế lai ghép còn đóng vai trò như một bước nền quan trọng để đánh giá định lượng mức độ đóng góp của từng nhóm đặc trưng trong nhiệm vụ dự đoán vị trí PTM.

Thứ hai, nhằm giảm thiểu sự phụ thuộc vào đặc trưng thủ công, nghiên cứu phát triển một mô hình học sâu lai kết hợp giữa CNN1D và LSTM/Bi-LSTM, tích hợp kỹ thuật NLP để tự động học biểu diễn đặc trưng từ mẫu dữ liệu thô. Mô hình được huấn luyện theo cơ chế đầu-cuối (end-to-end), trong đó toàn bộ quy trình từ biểu diễn, trích chọn đặc trưng đến phân loại đều được tối ưu hóa đồng thời trong một mạng duy nhất. Cách tiếp cận này giúp mô hình học trực tiếp từ dữ liệu thô, nâng cao tính tổng quát và khả năng tự thích ứng.

Thứ ba, nhận thấy mô hình học sâu lai và tổ hợp tuy hiệu quả nhưng tiêu tốn tài nguyên tính toán đáng kể, nghiên cứu đề xuất ứng dụng kỹ thuật chắt lọc tri thức kết hợp với kỹ thuật NLP để xây dựng mô hình dự đoán PTM hiệu quả hơn về mặt chi phí.

Cụ thể, một mô hình mạnh (giáo viên) sẽ truyền tri thức cho một mô hình nhẹ hơn (học viên), qua đó duy trì hiệu năng dự đoán trong khi giảm đáng kể độ phức tạp và thời gian huấn luyện. Đây là hướng tiếp cận mới, chưa được áp dụng trong lĩnh vực dự đoán PTM, có tiềm năng lớn cả về giá trị học thuật lẫn tính ứng dụng thực tiễn.

1.7 Kết luận chương 1

Trong chương NCS đã trình bày các kiến thức nền tảng về protein, đặc biệt protein sửa đổi sau dịch mã, vai trò việc xác định vị trí PTM trên chuỗi protein. Phát biểu bài toán dự đoán vị trí PTM. Quy trình xây dựng mô hình dự đoán, phương pháp mã hoá đặc trưng hiện nay, phương pháp đánh giá mô hình dự đoán, yêu cầu hệ thống thư viện và môi trường cài đặt. Tổng quan tình hình nghiên cứu trong bối cảnh sự phát triển của AI và SOTA, khoảng trống nghiên cứu, lựa chọn hướng nghiên cứu của luận án.

CHƯƠNG 2. MÔ HÌNH HỌC MÁY TỔ HỢP DỰ ĐOÁN VỊ TRÍ SỬA ĐỔI SAU DỊCH MÃ TRONG CHUỖI PROTEIN

Trong giai đoạn đầu của nghiên cứu, để giải quyết vấn đề dữ liệu hạn chế và để hiểu một số loại đặc trưng trong dự đoán PTM, NCS nghiên cứu các mô hình học máy và các đặc trưng (bao gồm cả thủ công và đặc trưng từ mô hình NLP Word2vec), tiếp theo NCS đề xuất mô hình học tập tổ hợp dựa trên đặc trưng lai để dự đoán PTM. Mặc dù việc kết hợp nhiều đặc trưng làm tăng chiều và độ phức tạp, tuy nhiên bước này có vai trò như một bước nền để đánh giá khả năng đóng góp của loại đặc trưng trong mô hình dự đoán vị trí PTM. Kết quả nghiên cứu được công bố tại Hội thảo quốc tế CITA2023 (Indexed: Scopus Q4) - (CT3).

2.1 Đặt vấn đề

2.2 Kỹ thuật học tập tổ hợp

2.3 Mô hình dự đoán vị trí PTM dựa trên kỹ thuật học tập tổ hợp đề xuất

2.3.1 Tên viết tắt

2.3.2 Dữ liệu thực nghiệm

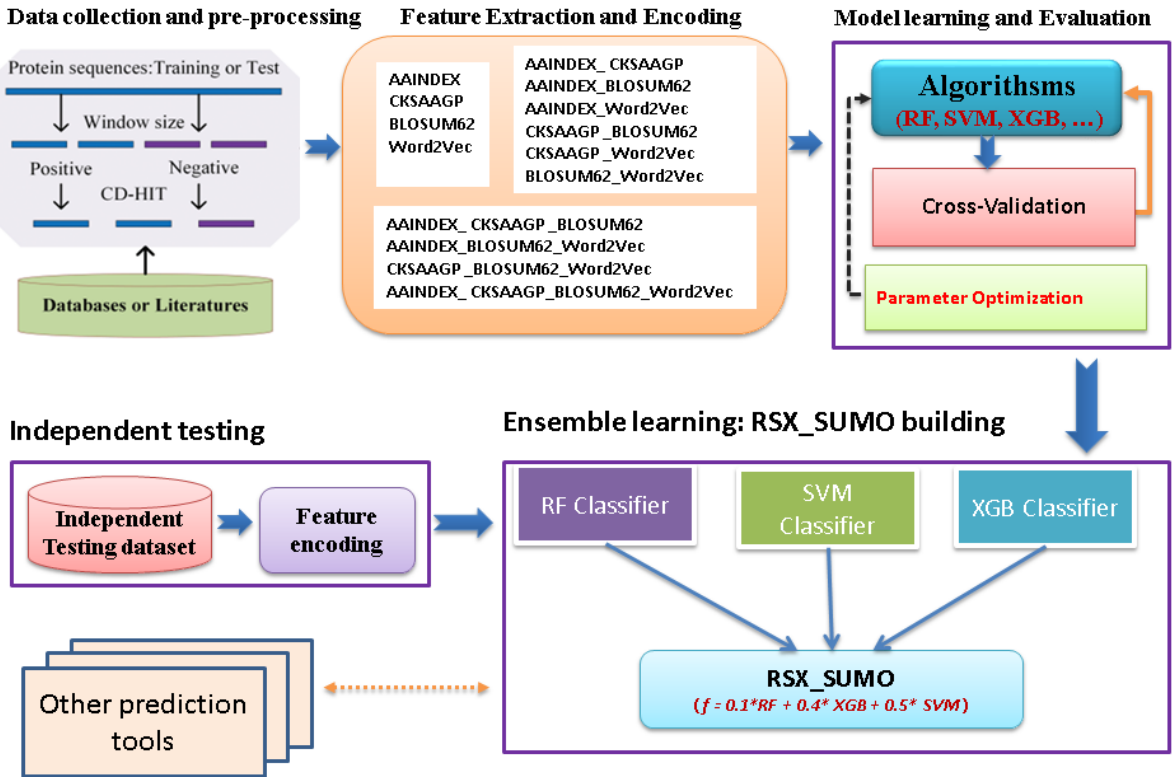
Bảng 2.1 Bộ dữ liệu SUMOylation sử dụng trong nghiên cứu

	SL mẫu dương tính	**SL mẫu âm tính**
Tập dữ liệu huấn luyện	745	1490
Tập dữ liệu kiểm tra	117	234

2.3.3 *Phương pháp trích chọn đặc trưng và mã hoá*

Để xây dựng các mô hình dự đoán cho việc xác định các vị trí SUMOylation, NCS trích chọn các đặc trưng dựa trên chuỗi: AAIndex, CKSAAP, BLOSUM62, các đặc trưng này được trích chọn bởi tool iFeature; đặc trưng dựa trên kỹ thuật NLP được trích chọn bởi mô hình Word2Vec (Skip-gram).

2.3.4 *Kiến trúc mô hình dự đoán PTM đề xuất dựa trên kỹ thuật học tập tổ hợp*



Hình 2.1 Kiến trúc mô hình học tập tổ hợp dự đoán PTM đề xuất

2.3.5 Chiến lược và tham số huấn luyện mô hình

2.3.6 Kết quả và thảo luận

2.4 So sánh mô hình đề xuất với các công cụ dự đoán khác

Bảng 2.2 So sánh hiệu suất mô hình đề xuất và các công cụ dự đoán

Công cụ	Ngưỡng	ACC	SEN	SPE
GPS-SUMO2.0	Low	0.877	0.884	0.875
	Medium	0.794	0.694	0.838
	High	0.877	0.884	0.875
seeSUMO2.0	Low	0.855	0.828	0.865
	Medium	0.769	0.644	0.829
	High	0.836	0.790	0.853
RXS_SUMO (Đề xuất)		**0.886**	**0.835**	**0.911**

2.5 Kết luận chương 2

Trong chương này, NCS đã đề xuất mô hình RSX_SUMO dự đoán vị trí SUMOyaltion. Mô hình RSX_SUMO xây dựng dựa trên kỹ thuật học tập tổ hợp bằng cách kết hợp ba thuật toán học máy: RF, XGBoost và SVM. Việc kết hợp này giúp khai thác ưu điểm của từng thuật toán, nâng cao hiệu suất phân loại và đảm bảo tính ổn định của mô hình. Tuy nhiên mô hình học tập tổ hợp với nhiều mô hình cơ sở và các đặc trưng lai tuy có hiệu suất cao hơn một chút so với mô hình cơ sở nhưng tốn rất nhiều tài nguyên tính toán. Đây cũng là tiền đề để NCS nghiên cứu đề xuất các phương pháp dự đoán ở các chương tiếp theo. Nội dung của chương NCS được công bố trên hội thảo và sau:

CHƯƠNG 3. MÔ HÌNH HỌC SÂU LAI KẾT HỢP XỬ LÝ NGÔN NGỮ TỰ NHIÊN DỰ ĐOÁN VỊ TRÍ SỬA ĐỔI SAU DỊCH MÃ TRONG CHUỖI PROTEIN

Tiếp tục nghiên cứu, trong chương này NCS tập trung vào việc khắc phục các hạn chế của mô hình học máy trong bài toán dự đoán vị trí SUMOylation. Ở thời điểm nghiên cứu trong Chương 2, nguồn dữ liệu SUMOylation còn hạn chế, ảnh hưởng không nhỏ đến hiệu quả huấn luyện và khả năng tổng quát hóa của các mô hình. Nhằm cải thiện điều này, trong Chương 3, NCS đã tiến hành cập nhật và mở rộng tập dữ liệu SUMOylation, thu thập bổ sung dữ liệu, giúp cải thiện chất lượng dữ liệu đầu vào cho

mô hình học sâu.

Song song với việc mở rộng dữ liệu, NCS đề xuất một hướng tiếp cận mới dựa trên mô hình học sâu lai, kết hợp giữa CNN1D và LSTM/Bi-LSTM, đồng thời tích hợp kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) để biểu diễn trình tự protein một cách hiệu quả. CNN1D cho phép trích xuất đặc trưng cục bộ liên quan đến tín hiệu PTM, trong khi LSTM/Bi-LSTM có khả năng học các phụ thuộc dài trong chuỗi, vốn rất cần thiết cho các đặc trưng ngữ cảnh sinh học. Cách tiếp cận theo hướng NLP đề xuất còn giúp giảm sự phụ thuộc vào kỹ thuật trích chọn thủ công, thay vào đó cho phép mô hình học đặc trưng tự động từ chuỗi đầu vào.

Không dừng lại ở SUMOylation, chương này cũng mở rộng phạm vi nghiên cứu sang một loại PTM khác là Succinylation. Việc phát triển mô hình cho cả hai loại PTM này nhằm vừa kiểm chứng khả năng tổng quát hóa của mô hình đề xuất, vừa góp phần làm giàu hiểu biết của cộng đồng khoa học về các dạng sửa đổi sau dịch mã trong hệ gen và proteome sinh vật.

Kết quả thực nghiệm cho thấy các mô hình học sâu đề xuất không chỉ cải thiện độ chính xác và khả năng tổng quát so với phương pháp học máy ở chương trước, mà còn tối ưu hơn về chi phí tính toán và lưu trữ. Một phần kết quả nghiên cứu đã được công bố trên các tạp chí khoa học uy tín như Tạp chí Tin học Điều khiển (CT4) và Computer and Medicine (SCIE Q1, IF 7.0) (CT5), Hội thảo CITA2024 (CT6) và hội thảo ICTA2024 (CT7).

3.1 Mô hình học sâu lai

3.1.1 Mô hình mạng neural tích chập một chiều (CNN1D)

3.1.2 Mô hình LSTM, Bi-LSTM

3.2 Mô hình dự đoán SUMOylation dựa trên kiến trúc học sâu lai (CNN1D_LSTM) và kỹ thuật xử lý ngôn ngữ tự nhiên dự đề xuất

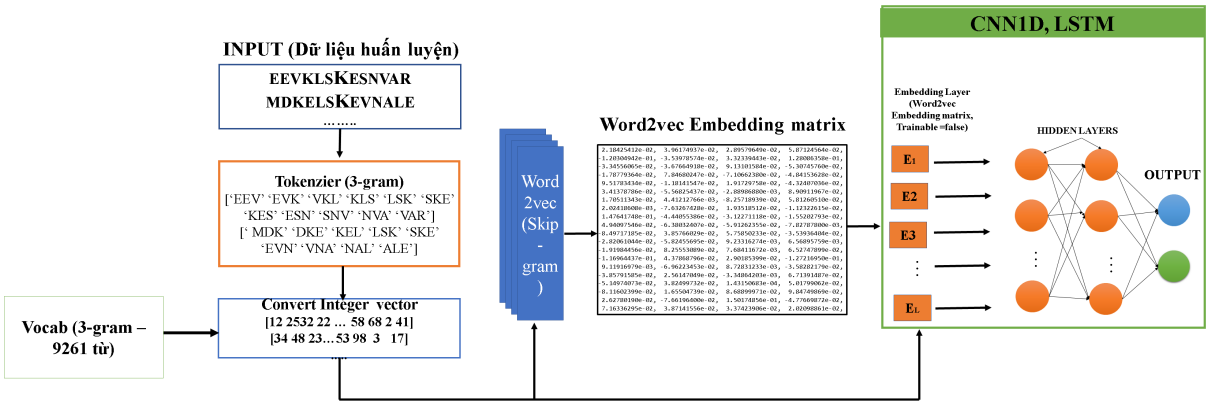
3.2.1 Tên viết tắt

3.2.2 Dữ liệu thực nghiệm

Bảng 3.1 Dữ liệu sử dụng trong nghiên cứu

	SL mẫu dương tính	**SL mẫu âm tính**
Dữ liệu huấn luyện	4985	9967
Dữ liệu kiểm tra	1245	2870

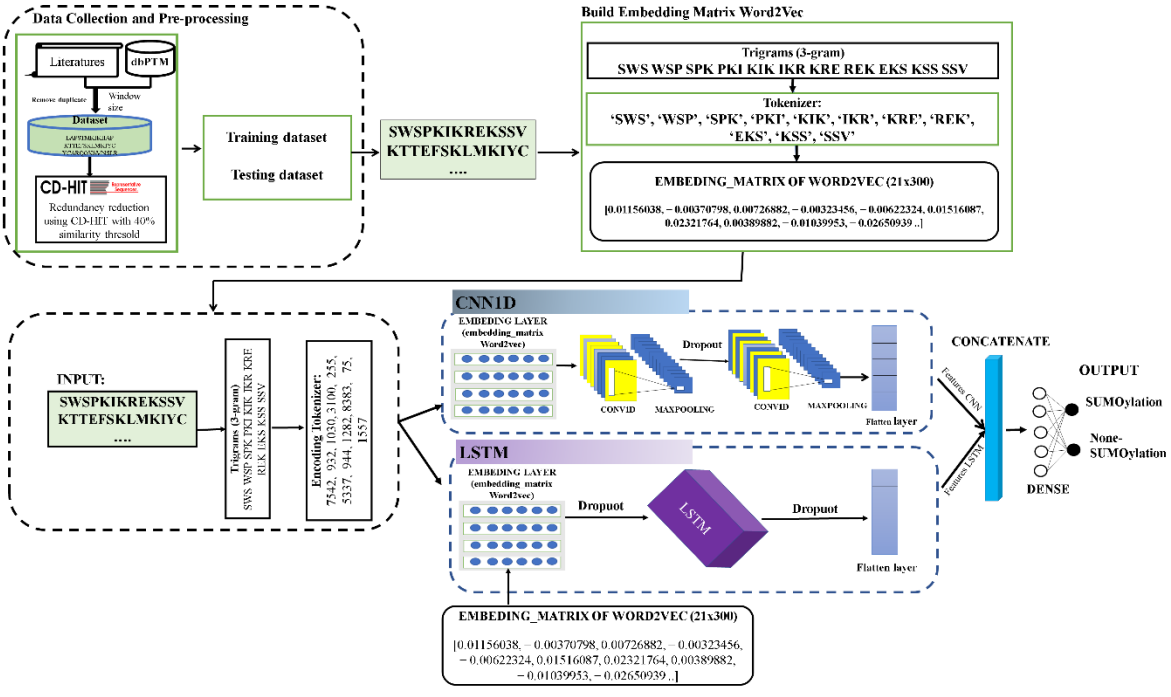
3.2.3 *Phương pháp mã hoá và trích chọn đặc trưng*



Hình 3.1 Quy trình mã hóa dữ liệu đề xuất, bao gồm các bước: (1) tokenization chuỗi protein bằng n-gram, (2) chuyển đổi token thành chỉ số số, và (3) sử dụng ma trận nhúng Word2Vec làm đầu vào cho lớp embedding trong CNN1D và LSTM.

3.2.4 *Kiến trúc mô hình học sâu lai (CNN_LSTM) dự đoán vị trí SUMOylation*

Mô hình học sâu lai dự đoán vị trí SUMOylation xây dựng dựa trên sự kết hợp của hai mô hình học sâu là CNN và LSTM (Hình 3.2):



Hình 3.2 Mô hình học sâu lai dự đoán SUMOylation (CLW_SUMO) đề xuất

3.2.5 Chiến lược và tham số huấn luyện mô hình

3.2.6 Kết quả và thảo luận

3.2.7 So sánh mô hình đề xuất với các công cụ khác

	ACC	SEN	SPE	MCC
GPS-SUMO2.0	0.877	0.884	0.875	0.715
seeSUMO	0.835	0.79	0.853	0.615
JASSA	0.87	0.453	0.864	0.423
ResSUMO	0.727	0.809	0.649	0.456
RXS_SUMO	0.886	0.835	0.91	0.743
CLW_SUMO	0.900	0.880	0.909	0.773

Hình 3.3 So sánh hiệu suất mô hình CLW_SUMO với các công cụ dự đoán SUMOylaiton khác

3.3 Mô hình dự đoán Succinylation dựa trên kiến trúc học sâu lai (CNN1D_Bi-LSTM) và kỹ thuật xử lý ngôn ngữ tự nhiên đề xuất

3.3.1 Tên viết tắt

3.3.2 Dữ liệu thực nghiệm

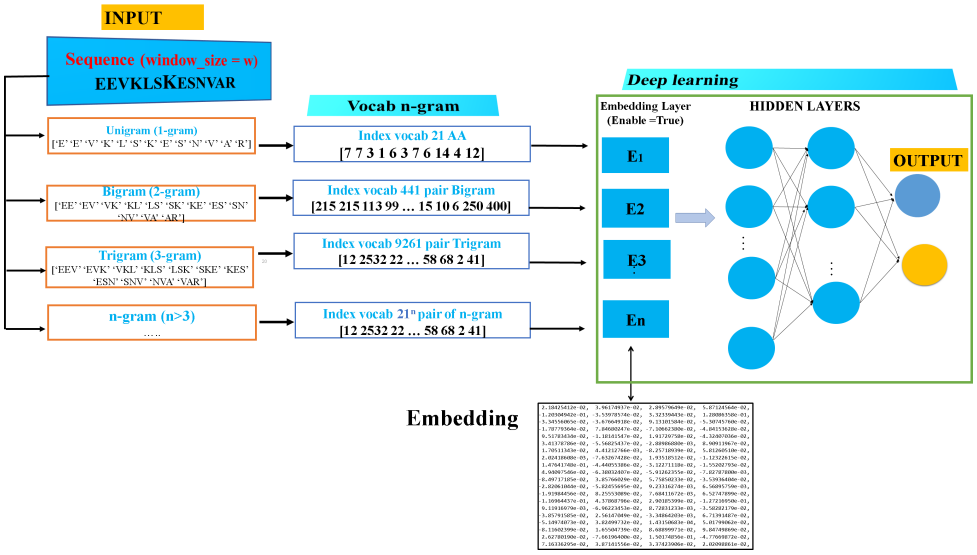
Bảng 3.2 Tập dữ liệu huấn luyện và kiểm tra sử dụng trong nghiên cứu

Tập dữ liệu	SL protein	SL mẫu dương tính	SL mẫu âm tính
Dữ liệu huấn luyện	2192	4750	4750
Dữ liệu kiểm thử 1	124	254	254
Dữ liệu kiểm thử 2	124	254	2977

3.3.3 Phương pháp mã hóa và trích chọn đặc trưng (Embedding động)

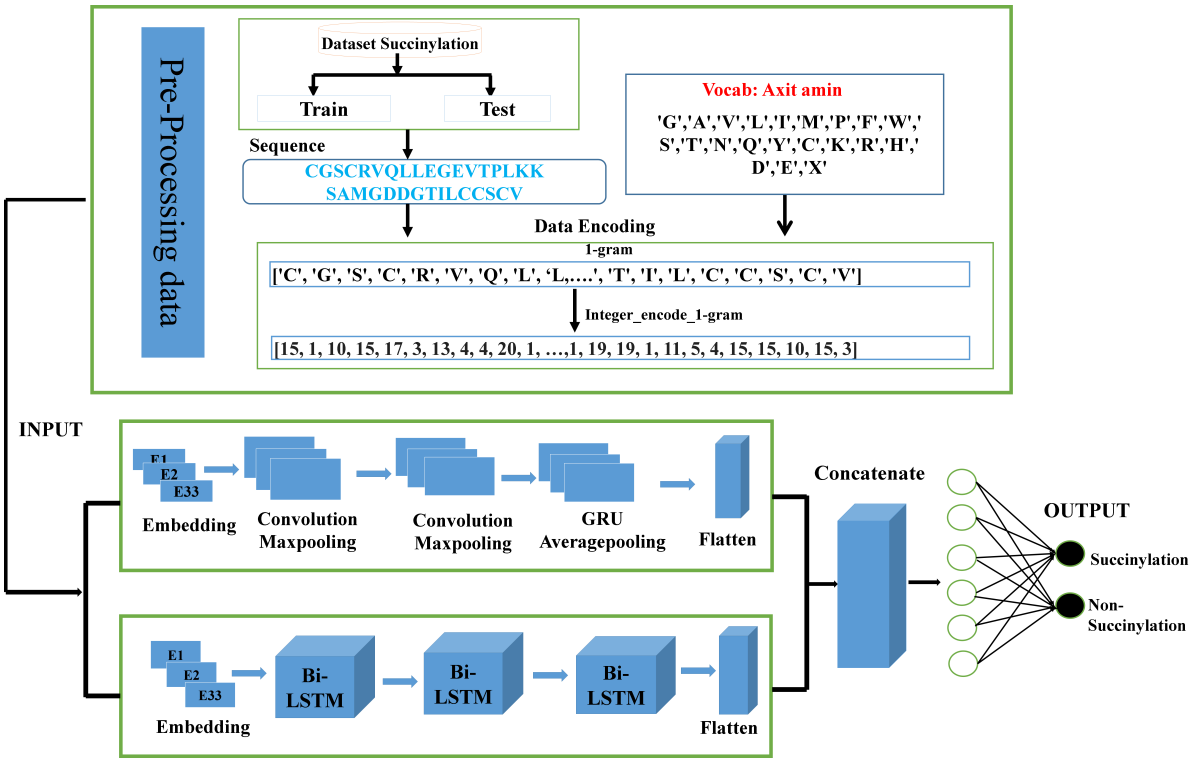
Trong nghiên cưu này, NCS đề xuất phương pháp Embedding động (Hình 3.4) tích hợp trong mô hình học sâu. Khi mô hình học qua nhiều epochs, thông qua quá trình lan truyền xuôi và lan truyền ngược của mạng CNN1D và Bi-LSTM, các giá trị số này được

cập nhật, giúp mô hình học được biểu diễn tối ưu của các axit amin.



Hình 3.4 Quy trình mã hoá dữ liệu protein với phương pháp embedding động

3.3.4 Kiến trúc mô hình học sâu lai (CNN1D_Bi-LSTM) dự đoán vị trí Succinylation



Hình 3.5 Kiến trúc mô hình dự đoán vị trí PTM đề xuất (CBILSuccSite)

3.3.5 Chiến lược và tham số huấn luyện mô hình

3.3.6 Kết quả và thảo luận

3.3.7 So sánh mô hình đề xuất với các công cụ dự đoán Succinylation

Bảng 3.3 So sánh mô hình đề xuất với các công cụ dự đoán succinyl hóa khác

Bộ dữ liệu	Tools	ACC	SEN	SPE	MCC
Dữ liệu kiểm thử 1	GPSuc	0.670	0.660	0.680	0.350
	DeepSuccinylSite	0.700	0.790	0.690	0.480
	LMSuccSite	0.740	0.760	0.730	0.510
	pSuc-EDBAM	0.699	0.748	0.650	0.400
	MDCAN-Lys	0.707	0.768	0.646	0.420
	CBiLSuccSite (Đề xuất)	**0.763**	**0.803**	**0.724**	**0.530**
Dữ liệu kiểm thử 2	GPSuc	0.850	0.880	0.490	0.300
	DeepSuccinylSite	0.700	0.790	0.690	0.270
	LMSuccSite	0.790	0.790	0.790	0.360
	pSuc-EDBAM	0.738	0.760	0.736	0.290
	MDCAN-Lys	0.732	0.705	0.734	0.260
	CBiLSuccSite (Đề xuất)	**0.733**	**0.941**	**0.715**	**0.370**

3.4 Kết luận chương 3

Trong chương này, NCS đã tập trung giải quyết bài toán dự đoán Post-Translational Modification (PTM) bằng cách đề xuất và phát triển hai mô hình học sâu lai tiên tiến: **CLW_SUMO** và CBiLSuccSite. Mô hình **CLW_SUMO** được thiết kế cho dự đoán vị trí SUMOylation, sử dụng kiến trúc kết hợp CNN1D và LSTM với embedding tĩnh dựa trên ma trận Word2Vec nhằm khai thác các đặc trưng ngữ nghĩa từ chuỗi protein. Trong khi đó, CBiLSuccSite hướng tới dự đoán vị trí Succinylation, áp dụng kiến trúc CNN1D kết hợp Bi-LSTM cùng với embedding động, giúp nâng cao khả năng học biểu diễn và tối ưu hóa hiệu suất.

Kết quả thực nghiệm trên nhiều tập dữ liệu độc lập đã khẳng định rằng cả hai mô hình đều vượt trội hơn các mô hình cơ sở và nhiều phương pháp state-of-the-art hiện có. Hiệu quả này bắt nguồn từ sự cộng hưởng của kiến trúc hybrid: CNN1D giúp trích xuất đặc trưng cục bộ, LSTM/Bi-LSTM nắm bắt được ngữ cảnh dài hạn, và Word2Vec embedding hoặc embedding động cung cấp biểu diễn ngữ nghĩa phong phú hơn so với các đặc trưng thủ công truyền thống. Những kết quả đạt được không chỉ chứng minh tính hiệu quả và tiềm năng của việc tích hợp kỹ thuật embedding từ NLP với mô hình học sâu lai trong dự đoán PTM.

CHƯƠNG 4. MÔ HÌNH HỌC CHẮT LỌC TRI THỨC KẾT HỢP XỬ LÝ NGÔN NGỮ TỰ NHIÊN DỰ ĐOÁN VỊ TRÍ SỬA ĐỔI SAU DỊCH MÃ TRONG CHUỖI PROTEIN

Trong chương 3, NCS đã đề xuất một số mô hình học sâu lai kết hợp kỹ thuật NLP nhằm cải thiện hiệu suất dự đoán vị trí Succinylation và SUMOylation. Các kết quả đạt được đã được khẳng định thông qua các công bố trên các tạp chí uy tín trong và ngoài nước, cho thấy tiềm năng của hướng tiếp cận này trong bài toán dự đoán vị trí PTM. Tuy nhiên, với mong muốn tiếp tục mở rộng và làm giàu tri về các loại PTM khác. Trong nghiên cứu này, NCS chọn PTM Ubiquitination. Ubiquitination đóng vai trò thiết yếu trong điều hòa sự ổn định và phân hủy protein, và là mục tiêu nghiên cứu rộng rãi trong lĩnh vực sinh học phân tử và y học.

Bên cạnh đó, từ thực tiễn triển khai mô hình ở các chương trước, một thách thức đáng kể được đặt ra là chi phí tính toán do sử dụng các mô hình sâu lai (CNN, Bi-LSTM). Để giải quyết bài toán này, chương 4 đề xuất một mô hình mới dự đoán vị trí Ubiquitination dựa trên học chắt lọc tri thức (Knowledge Distillation) – một kỹ thuật cho phép huấn luyện mô hình gọn nhẹ (mô hình Học viên) nhưng vẫn duy trì hiệu quả dự đoán tương đương với mô hình lớn (mô hình Giáo viên). Đặc biệt, mô hình được thiết kế kế thừa các ưu điểm của kỹ thuật mã hóa NLP đã chứng minh hiệu quả ở chương 3, giúp biểu diễn tốt hơn thông tin sinh học từ trình tự axit amin và tăng cường khả năng học của mô hình.

Hướng tiếp cận này không chỉ giúp mở rộng nghiên cứu làm giàu tri thức về một loại PTM khác, mà còn góp phần tối ưu hoá chi phí tính toán, giảm độ phức tạp mô hình, và nâng cao tính khả thi triển khai trong thực tiễn đáp ứng yêu cầu thiết yếu của các bài toán sinh tin hiện đại trong bối cảnh dữ liệu ngày càng lớn và đa dạng. Một phần kết quả nghiên cứu được đăng trên tạp chí Methods (SCIE Q1) (CT8).

4.1 Học chắt lọc tri thức

Tuy nhiên, theo khảo sát và tổng quan tài liệu hiện nay, chưa có nghiên cứu nào ứng dụng học chắt lọc tri thức trong bài toán dự đoán vị trí PTM, đặc biệt đối với ubiquitination trong ở thực vật. Khoảng trống này chính là cơ sở để NCS lựa chọn áp dụng phương pháp học chắt lọc tri thức trong bài toán dự đoán vị trí ubiquitination cho Arabidopsis thaliana. Việc kết hợp giữa kiến trúc học sâu và kỹ thuật chắt lọc tri thức được kỳ vọng sẽ góp phần nâng cao hiệu quả dự đoán, đồng thời mở ra hướng tiếp cận mới trong lĩnh vực tin sinh học, nơi các mô hình dự đoán hiện nay còn gặp nhiều hạn chế về khả năng tổng quát hoá và chi phí tính toán. Đây cũng là một đóng góp mới về

mặt phương pháp luận của nghiên cứu này đối với cộng đồng nghiên cứu PTM.

4.2 Mô hình dự đoán Ubiquitination dựa trên học chắt lọc tri thức và kỹ thuật xử lý ngôn ngữ tự nhiên đề xuất

4.2.1 Tên viết tắt

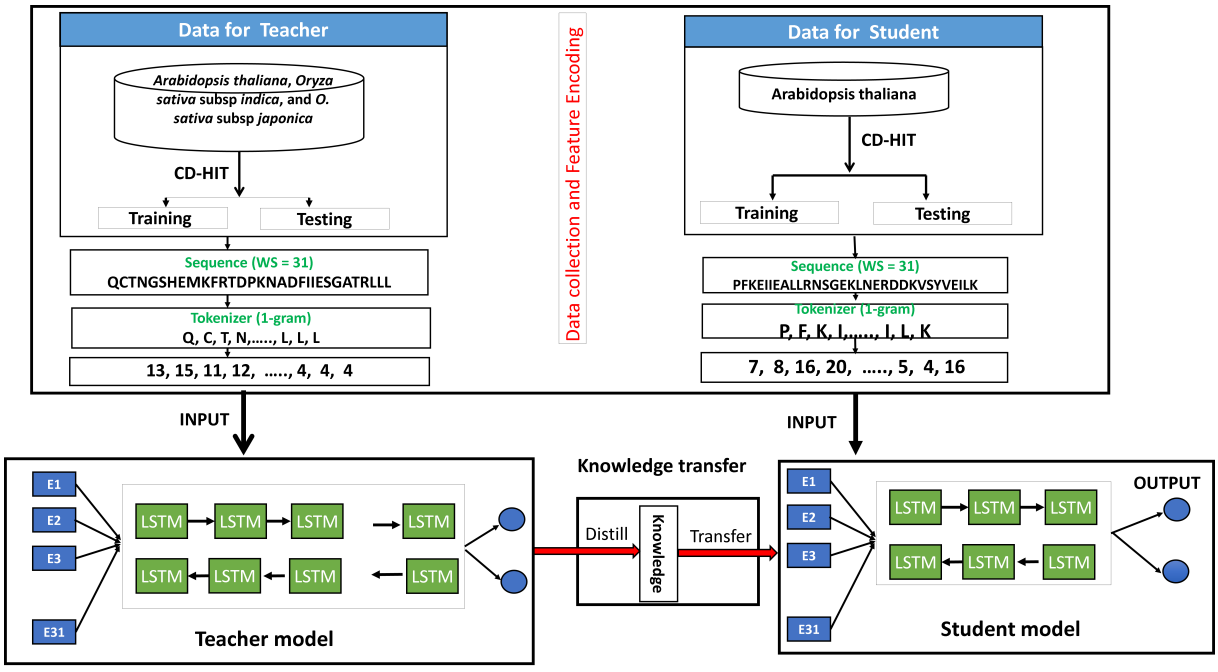
4.2.2 Dữ liệu sử dụng trong nghiên cứu

Bảng 4.1 Bộ dữ liệu huấn luyện và kiểm tra sử dụng trong nghiên cứu

Mô hình	Bộ dữ liệu	SL Protein	SL mẫu dương tính	SL mẫu âm tính
Mô hình Giáo viên	Dữ liệu huấn luyện	25,103	3,373	3,373
	Dữ liệu kiểm tra	–	750	750
Mô hình Học viên	Dữ liệu huấn luyện	1,607	1,532	1,532
	Dữ liệu kiểm tra	–	511	511

4.2.3 Phương pháp mã hoá và trích chọn đặc trưng

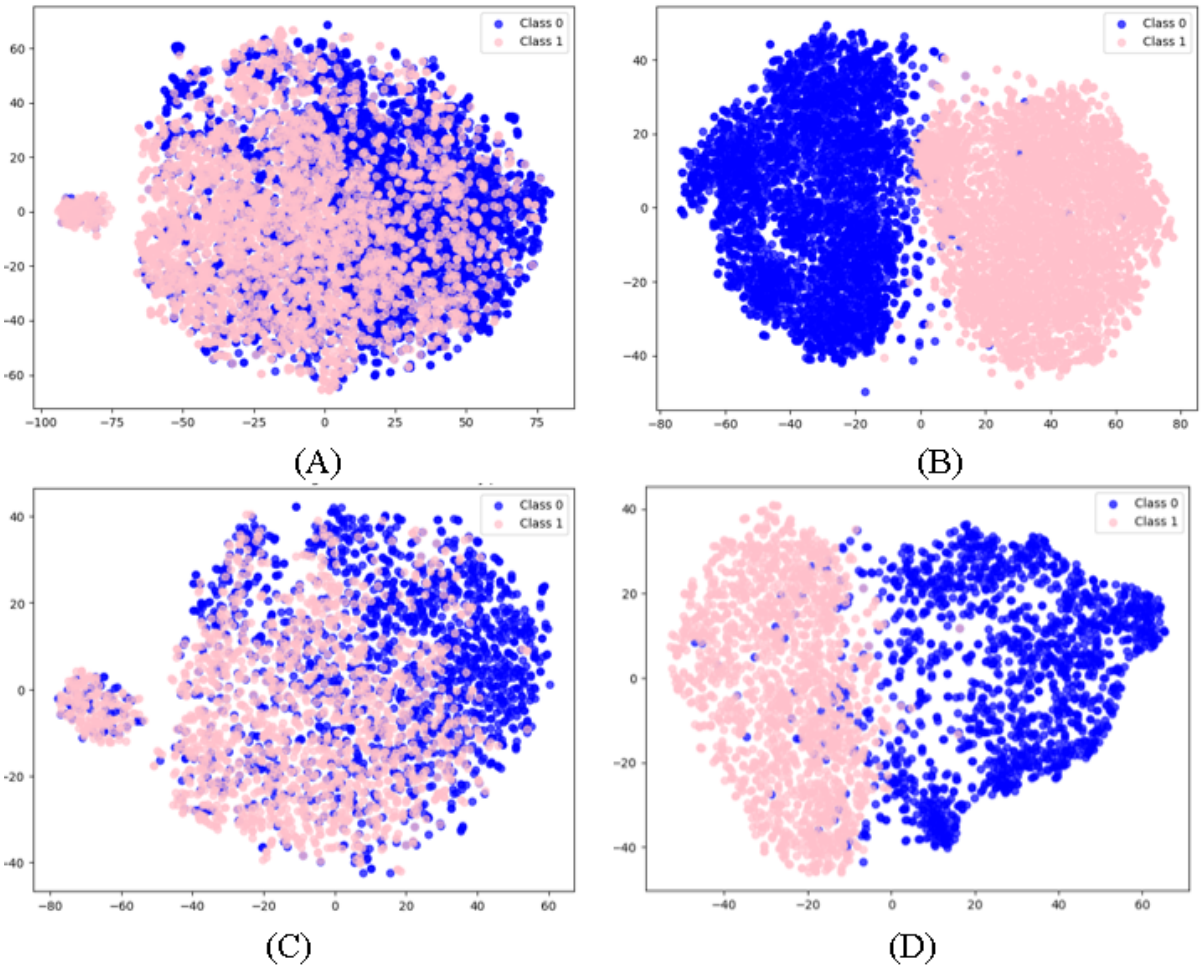
4.2.4 Kiến trúc học chắt lọc tri thức dự đoán vị trí PTM



Hình 4.1 Kiến trúc mô hình học chắt lọc tri thức KD_ArapUbi đề xuất

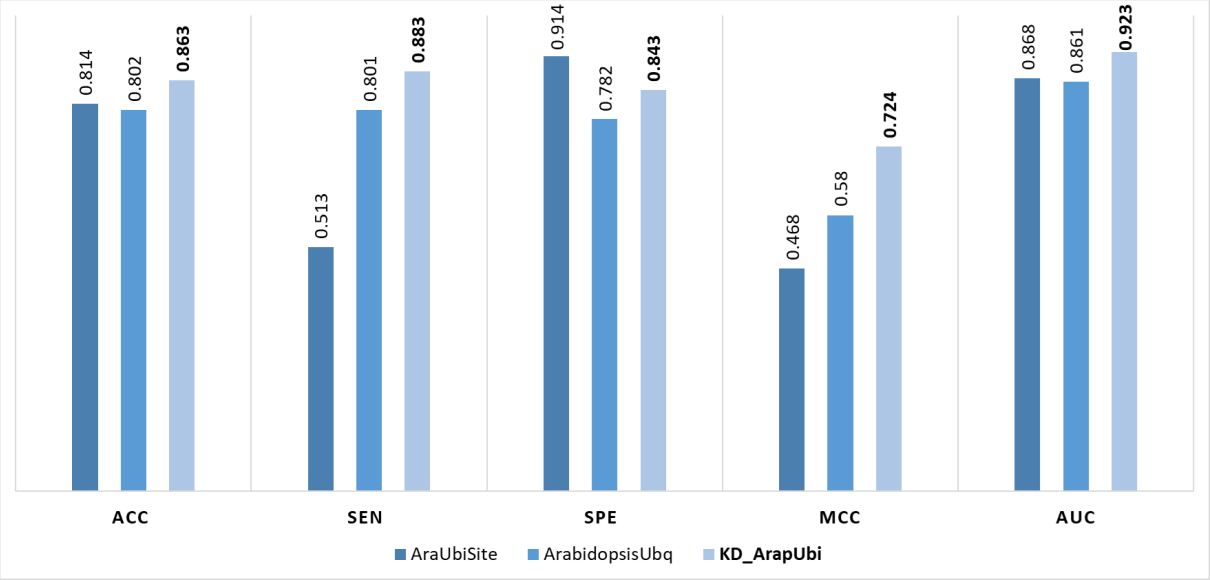
4.2.5 Chiến lược và tham số huấn luyện mô hình

4.2.6 Kết quả và thảo luận



Hình 4.2 Trực quan hoá T-SNE (A) Dữ liệu của mô hình Giáo viên (các loài thực vật) trước khi huấn luyện, (B) Dữ liệu sau khi được huấn luyện bởi mô hình Giáo viên, (C) Dữ liệu loài Arabidopsis thaliana trước khi huấn luyện, (D) Dữ liệu sau khi huấn luyện bởi mô hình học chắt lọc tri thức KD_ArapUbi đề xuất (mô hình Học viên được hướng dẫn bởi mô hình Giáo viên đã được huấn luyện trên bộ dữ liệu đa loài). (Class 0: Mẫu âm tính, Class 1: Mẫu dương tính)

4.3 So sánh mô hình đề xuất với các công cụ hiện có về dự đoán *Arabidopsis thaliana*



Hình 4.3 So sánh mô hình đề xuất và các công cụ dự đoán *Arabidopsis thaliana*

4.4 Phân tích so sánh tổng thể bốn mô hình đề xuất trong luận án

Bảng 4.2 Đánh giá so sánh bốn mô hình được đề xuất trong luận án

Mô hình	RSX_SUMO	CLW_SUMO	CBiLSuccsite	KD_ArapUbi
Kỹ thuật xây dựng mô hình	Học máy tổ hợp (XGBoost, SVM, RF)	Học sâu lai (CNN–LSTM)	Học sâu lai (CNN–BiLSTM)	Học chắt lọc tri thức
Phương pháp mã hoá dữ liệu và trích chọn đặc trưng	Vector đặc trưng thủ công (AAIndex, CKSAAP, BLOSUM62, Word2Vec)	Embedding tĩnh (Word2Vec tiền huấn luyện)	Embedding động (lớp embedding được huấn luyện)	Embedding động (lớp embedding được huấn luyện)
Tổng số tham số	4,387,000 (40.3MB)	5,859,281 (22.35 MB)	451,025 (1.72 MB)	174,538 (681.79 KB)
Tham số huấn luyện được	4,387,000 (40.3MB)	302,081 (1.15 MB)	451,025 (1.72 MB)	174,538 (681.79 KB)
Thời gian huấn luyện	Cao	Cao	Trung bình	Thấp
Phù hợp với dữ liệu	Hạn chế	Vừa,lớn	Vừa	Hạn chế,vừa

4.5 Kết luận chương 4

Trong chương này, NCS đã đề xuất và phát triển mô hình **KD_ArapUbi**, một kiến trúc học chắt lọc tri thức (Knowledge Distillation) ứng dụng cho bài toán dự đoán vị trí ubiquitination trên loài *Arabidopsis thaliana*. Bằng cách kết hợp giữa "mô hình Giáo viên" Bi-LSTM mạnh mẽ, embedding động và "mô hình Học viên" gọn nhẹ hơn, **KD_ArapUbi** đã đạt được hiệu suất vượt trội so với nhiều phương pháp truyền thống và công cụ dự đoán hiện có, đồng thời giảm đáng kể số lượng tham số cần huấn luyện.

Bên cạnh đó, việc so sánh và đánh giá bốn mô hình khác nhau (**RSX_SUMO**, **CLW_SUMO**, CBiLSuccsite và **KD_ArapUbi**) cho thấy sự đa dạng trong cách tiếp cận bài toán. Các mô hình học máy truyền thống đem lại độ ổn định và tốc độ xử lý nhanh, thích hợp cho dữ liệu nhỏ. Các mô hình học sâu lai khai thác tốt cả đặc trưng cục bộ và toàn cục, song đòi hỏi tài nguyên tính toán lớn hơn. Trong khi đó, hướng tiếp cận hiện đại với học chắt lọc tri thức đã chứng minh khả năng tối ưu hoá, vừa giảm chi phí tính toán, vừa duy trì hiệu suất cao, phù hợp cho các ứng dụng thực tiễn.

Từ đó, luận án khẳng định giá trị khoa học không chỉ ở việc đề xuất một mô hình mới có hiệu năng vượt trội, mà còn ở việc cung cấp góc nhìn toàn diện về ưu, nhược điểm và bối cảnh sử dụng của các mô hình dự đoán PTM. Đây là cơ sở quan trọng để định hướng phát triển các công cụ tính toán hiệu quả, phục vụ cho nghiên cứu và ứng dụng trong sinh học tính toán.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong luận án này, NCS đã tập trung nghiên cứu và phát triển các mô hình cải tiến nhằm nâng cao hiệu suất dự đoán các vị trí sửa đổi sau dịch mã (PTM) trên protein. Cụ thể với việc đề xuất kiến trúc Học tập tổ hợp với đặc trưng lai ghép, một số kiến trúc Mô hình học sâu lai, học chắt lọc tri thức kết hợp kỹ thuật NLP mới giúp cải thiện hiệu suất của 3 PTM (SUMOylation, Succinylation và Ubiquitination).

A. Các kết quả đạt được của luận án

Luận án có ba đóng góp chính sau:

(1) Cơ sở lý luận và tổng quan hệ thống: Luận án đã hệ thống hóa, phân tích và so sánh các phương pháp từ truyền thống, học máy tổ hợp, học sâu lai cho đến kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) trong bài toán dự đoán PTM, qua đó xây dựng nền tảng khoa học vững chắc cho các nghiên cứu tiếp theo.

(2) Khai thác NLP cho dữ liệu protein: Luận án đã chứng minh khả năng ứng dụng và hiệu quả của các kỹ thuật NLP trong việc biểu diễn ngữ cảnh của chuỗi protein,

giúp vượt qua hạn chế của đặc trưng thủ công và nâng cao độ chính xác trong dự đoán.

(3) Đề xuất và phát triển mô hình mới: Luận án đã đề xuất bốn mô hình PTM với hiệu suất cao, trong đó có các mô hình lai kết hợp học sâu với NLP và đặc biệt là mô hình áp dụng học chắt lọc tri thức cho Ubiquitination, phù hợp với bối cảnh dữ liệu hạn chế và môi trường tính toán hạn chế. Cụ thể, bốn đề xuất chính gồm:

- Đề xuất mô hình dự đoán vị trí PTM (SUMOylation) dựa trên học máy tổ hợp và các đặc trưng lai ghép.

- Đề xuất hai mô hình dự đón vị trí PTM (SUMOylation và Succinylation) dựa trên kỹ thuật học sâu lai ghép và kỹ thuật xử lý ngôn ngữ tự nhiên.

- Đề xuất mô hình dự đoán PTM (Ubiquitination) dựa trên học chắt lọc tri thức và kỹ thuật xử lý ngôn ngữ tự nhiên.

B. Những điểm mới và ý nghĩa của các kết quả nghiên cứu

Nghiên cứu đã đề xuất phương pháp mã hóa dữ liệu sử dụng kỹ thuật NLP, phù hợp với dữ liệu protein cấu trúc bậc 1, đồng thời tận dụng được sức mạnh tự động học đặc trưng của mô hình học sâu. Điều này giúp cải thiện đáng kể khả năng biểu diễn dữ liệu, tối ưu hóa quá trình huấn luyện mô hình và nâng cao hiệu suất dự đoán.

Bên cạnh đó, luận án đã đề xuất một số mô hình tiên tiến để dự đoán vị trí PTM với hiệu suất cao, trong đó có RSX_SUMO, CLW_SUMO, CBiLSuccSite và KD_ArapUbi. Các mô hình này đều ứng dụng kỹ thuật mã hóa chuỗi protein bằng NLP, giúp khai thác tốt đặc trưng trình tự protein. Hơn nữa, kiến trúc của các mô hình đề xuất đã phát huy sức mạnh của nhiều mô hình học máy và học sâu, từ đó cải thiện hiệu suất tổng thể.

Một số mô hình học sâu lai đặc biệt mô hình học chắt lọc tri thức không chỉ giúp nâng cao độ chính xác mà còn có khả năng tự động học đặc trưng từ dữ liệu thô và thực hiện quá trình học end-to-end. Điều này giúp giảm thiểu sự phụ thuộc vào các phương pháp trích xuất đặc trưng thủ công, đồng thời đảm bảo tính tổng quát và hiệu quả của mô hình trong bài toán dự đoán vị trí PTM.Các mô hình đề xuất không chỉ mang tính học thuật mà còn có ý nghĩa thực tiễn, hỗ trợ các nhà nghiên cứu về sinh học phân tử, dược sĩ, bác sĩ rút ngắn thời gian trong việc phát hiện, phân tích các vị trí sửa đổi trên protein. Bên cạnh việc công bố kết quả nghiên cứu, NCS cũng chia sẻ dữ liệu và toàn bộ codes chương trình thực nghiệm lên nền tảng Github để đóng góp và hỗ trợ tích cực cho các nhà khoa học trong quá trình nghiên cứu, thực nghiệm có liên quan của họ.

Những kết quả nghiên cứu của luận án không chỉ đóng góp vào lĩnh vực dự đoán vị trí PTM mà còn khẳng định tính khả thi và hiệu quả của việc ứng dụng các mô hình Học tập tổ hợp, Mô hình học sâu lai, Học chắt lọc tri thức và kỹ thuật NLP với dữ liệu protein cấu trúc bậc 1 trong dự đoán vị trí PTM.

C. Hướng phát triển của luận án

Thứ nhất: Nâng cao độ chính xác của mô hình

Mặc dù các mô hình trong luận án đã đạt được kết quả đáng khích lệ trong việc dự đoán các vị trí sửa đổi sau dịch mã (PTM), vẫn còn những tiềm năng để cải thiện độ chính xác. Trong các nghiên cứu tiếp theo, cần xem xét kết hợp thêm các kỹ thuật học sâu tiên tiến hơn, tối ưu hóa kiến trúc mô hình hoặc kết hợp thêm thông tin đặc trưng sinh học để cải thiện chất lượng dự đoán.

Thứ hai: Xử lý vấn đề dữ liệu mất cân bằng

Dữ liệu PTM, đặc biệt khi mở rộng sang các loại PTM khác hoặc các loài khác, thường gặp phải tình trạng mất cân bằng nghiêm trọng giữa số lượng mẫu dương tính và âm tính. Trong nghiên cứu này, NCS sử dụng các bộ dữ liệu đã được cân bằng theo các nghiên cứu trước. Tuy nhiên, các hướng nghiên cứu tiếp theo cần tập trung khai thác và so sánh hiệu quả của các phương pháp xử lý dữ liệu mất cân bằng như oversampling, undersampling, áp dụng trọng số cho hàm mất mát, hoặc sử dụng các kỹ thuật như focal loss. Điều này không chỉ giúp cải thiện hiệu quả dự đoán mà còn tăng tính ứng dụng khi triển khai trên các bộ dữ liệu thực tế.

Thứ ba: Mở rộng mô hình cho dự đoán các PTM khác

Luận án đã tập trung vào một số PTM tiêu biểu như SUMOylation, Succinylation và Ubiquitination. Các nghiên cứu tiếp theo có thể mở rộng phạm vi nghiên cứu sang các loại PTM khác như Methylation, Acetylation, Phosphorylation,... để xây dựng một hệ thống dự đoán PTM toàn diện hơn.

Thứ tư: Phát triển phần mềm và công cụ hỗ trợ nghiên cứu

Việc triển khai các mô hình dự đoán vị trí PTM dưới dạng phần mềm hoặc công cụ dễ sử dụng cho các nhà sinh học và nghiên cứu viên sẽ giúp ứng dụng rộng rãi các phương pháp trong thực tiễn, góp phần hỗ trợ các nghiên cứu trong lĩnh vực sinh học phân tử và phát triển dược phẩm.

DANH MỤC CÔNG TRÌNH KHOA HỌC ĐÃ CÔNG BỐ

[CT1] Le N.Q.K, Tran T.X, Nguyen P.A. Nguyen V.N, et al. (2023), Recent progress in machine learning approaches for predicting carcinogenicity in drug development. Expert Opinion on Drug Metabolism & Toxicology. p 621-628, DOI: https://doi.org/10.1080/17425255.2024.2356162.**(SCIE Q1, IF: 3.9)- Bổ trợ- Chương 1**

[CT2] Le N.Q.K, Nguyen V.N, Nguyen T.T, Tran T.X, at al. (2024), Enhancing Protein Sequence Classification with a Fuzzy Neural Network: A Study in Anticancer Peptide Identification, International Conference on Fuzzy Theory and Its Applications (iFUZZY), Kagawa, Japan. pp. 1-6, DOI: https://doi.org/10.1109/iFUZZY63051.2024.10662887. **- Bổ trợ - Chương 1**

[CT3] Tran T.X, Nguyen V.N, and Le N.Q.K. (2023) Incorporating Natural Language-Based and Sequence-Based Features to Predict Protein SUMOylation Sites. Conference on Information Technology and its Applications. DOI: https://doi.org/10.1007/978-3-031-36886-8_7. **(Indexed: Scopus Q4)- Liên quan trực tiếp - Chương 2**

[CT4] Tran T.X., Le N.Q.K., and Nguyen V.N. (2024), CLW-SUMO: A hybrid deep learning model for predicting protein SUMOylation sites. Journal of Computer Science and Cybernetics. DOI: https://doi.org/10.15625/1813-9663/19626. **(Tạp chí Tin học điều khiển 1.25đ) - Liên quan trực tiếp - Chương 3**

[CT5] Tran T.X, Le N.Q.K, and Nguyen V.N. (2025), Integrating CNN and Bi-LSTM for protein succinylation sites prediction based on Natural Language Processing technique. Computers in Biology and Medicine. 186: p. 109664. DOI: https://doi.org/10.1016/j.compbiomed.2025.109664. **(SCIE Q1, IF: 7.0)- Liên quan trực tiếp - Chương 3**

[CT6] Tran T.X., T.T. Nguyen, N.Q.K. Le, et al. (2024), A Novel Deep Learning Approach for the Prediction of Arabidopsis Thaliana Ubiquitination Sites. Proceedings of the 13th International Conference on Information Technology and Its Applications .CITA 2024. p. 48-57. DOI: https://elib.vku.udn.vn/handle/123456789/4010. **(Scopus Index Q4) - Liên quan trực tiếp - Chương 3**

[CT7] Tran T.X, Nguyen T.T, Le N.Q.K, and Nguyen V.N. (2025), A hybrid deep learning and Natural Language Processing Model for Plant Ubiquitination Site Prediction, The 3rd International Conference on Advances in Information and Communication Technology. ICTA 2024. DOI: https://doi.org/10.1007/978-3-031-80943-9_49. **(Indexed: Scopus Q4)- Liên quan trực tiếp - Chương 3**

[CT8] Nguyen V. N., Tran T. X,. Nguyen T. T, N.Q.K. Le. (2024), Enhancing Arabidopsis thaliana ubiquitination site prediction through knowledge distillation and natural language processing. Methods. 232: p. 65-71. DOI: https://doi.org/10.1016/j.ymeth.2024.10.006. **(SCIE Q1 IF: 4.2)- Liên quan trực tiếp - Chương 4**