**THÔNG TIN NGHIỆM THU CẤP ĐẠI HỌC ĐỀ TÀI KH&CN**

**CẤP ĐẠI HỌC NĂM 2023**

**1. Tên đề tài:** Nghiên cứu, kết hợp mô hình học máy ứng dụng phân tích dự đoán Protein sửa đổi sau dịch mã

**2. Mã số:** ĐH2023-TN08-05

**3. Đơn vị chủ trì:** Trường Đại học Kinh tế và QTKD

**4. Chủ nhiệm đề tài:** ThS. Trần Thị Xuân

**5. Quyết định thành lập Hội đồng:** Số 4020/QĐ-ĐHTN ngày 01 tháng 8 năm 2025 của Giám đốc Đại học Thái Nguyên

**6. Thời gian nghiệm thu:** 8h00, thứ sáu, ngày 22 tháng 8 năm 2025

**7. Địa điểm nghiệm thu:** Phòng họp B – Trường Đại học Kinh tế và QTKD

*Trân trọng kính mời các giảng viên, nhà khoa học, nhà quản lý, nghiên cứu sinh, học viên cao học, sinh viên và người quan tâm đến dự.*

**THÔNG TIN KẾT QUẢ NGHIÊN CỨU**

**1. Thông tin chung:**

**-** Tên đề tài: Nghiên cứu, kết hợp mô hình học máy ứng dụng phân tích dự đoán protein sửa đổi sau dịch mã

- Mã số: ĐH2023-TN08-05

- Chủ nhiệm đề tài: ThS. Trần Thị Xuân

- Tổ chức chủ trì: Trường Đại học Kinh tế và Quản trị kinh doanh

- Thời gian thực hiện:24 tháng, từ tháng 07/2023 đến tháng 07/2025

**2. Mục tiêu:**

Nghiên cứu và nắm rõ được các các kiến thức lý thuyết cơ sở liên quan đến cơ chế của biến đổi hậu dịch mã.

Đề xuất được mô hình học máy kết hợp dự đoán protein sửa đổi sau dịch mã với hiệu năng cao.

**3. Tính mới và sáng tạo:**

Đề tài nghiên cứu đã:

- Đề xuất một số mô hình học máy kết hợp trong dự đoán protein sửa đổi sau dịch mã hiệu suất cao.

- Đề xuất mô hình học chắt lọc tri thức trong dự đoán protein sửa đổi sau dịch mã hiệu suất cao.

**4. Kết quả nghiên cứu:**

Tổng hợp được các kiến thức cơ sở liên quan đến cơ chế của biến đổi hậu dịch mã (PTM).

Nghiên cứu đề xuất một số mô hình học máy kết hợp trong dự đoán một số protein sửa đổi sau dịch mã (SUMOylation, Ubiquitination, Succinylation) hiệu suất cao, cụ thể:

* Đề xuất mô hình học tập tổ hợp (kết hợp 3 mô hình học máy SVM, RF, XGBoost có trọng số), cải thiện hiệu suất dự đoán SUMOylation.
* Đề xuất mô hình học sâu lai dự đoán Sumoylation: nghiên cứu này phát triển mô hình học sâu lai kết hợp CNN và LSTM nhằm nâng cao độ chính xác trong dự đoán SUMOylation, mô hình đề xuất này có khả năng tự động học đặc trưng từ dữ liệu thô ban đầu.
* Đề xuất một mô hình học sâu lai kết hợp mạng CNN và mạng LSTM để dự đoán PTM Ubiquitination. Trong nghiên cứu này, đặc trưng sinh học và đặc trưng tuần tự của trình tự protein được khai thác đưa vào các mô hình học sâu giúp cải thiện hiệu suất dự đoán.
* Đề xuất mô hình học chắt lọc tri thức dự đoán Succinylation hiệu suất cao. Mô hình đề xuất tận dụng ưu điểm của cả kỹ thuật học chắt lọc tri thức và embedding động để giảm độ phức tạp tính toán trong khi vẫn duy trì độ chính xác cao trong việc dự đoán các vị trí succinyl hóa protein. Kết quả thử nghiệm chứng minh rằng bộ dự đoán được đề xuất vượt trội hơn các bộ dự đoán hiện có, cung cấp một đóng góp có giá trị cho nghiên cứu PTM và các ứng dụng y sinh.

**5. Sản phẩm:**

**Sản phẩm khoa học:**

+ 02 bài báo đăng trên Tạp chí quốc tế ISI Q1.

+ 01 bài báo đăng tại Kỷ yếu Hội thảo Quốc tế (indexed by Springer Scopus Q4)

+ 01 bài báo đăng tại Kỷ yếu Hội thảo Quốc tế (1 điểm)

+ 01 bài báo đăng trên tạp chí uy tín trong nước (Tạp chí Tin học và điều khiển 1.25 điểm)

1. Nguyen Quoc Khanh Le🖂, Thi Xuan Tran, Phung Anh Nguyen, Thi Trang Ho, Van Nui Nguyen🖂, et al. "Recent progress in machine learning approaches for predicting carcinogenicity in drug development." Expert Opinion on Drug Metabolism & Toxicology just-accepted (2024). Doi: https://doi.org/10.1080/17425255.2024.2356162. ISI Q1(ISI Q1).
2. Thi Xuan Tran; Van Nui Nguyen🖂; LE, Nguyen Quoc Khanh Le. “Incorporating Natural Language-Based and Sequence-Based Features to Predict Protein SUMOylation Sites”. *In: Conference on Information Technology and its Applications. Cham: Springer Nature Switzerland, 2023*. p. 74-88. (Scopus Q4)
3. Thi Xuan Tran, Thi Tuyen Nguyen, Nguyen Quoc Khanh Le, Hong Hai Nguyen, and Van Nui Nguyen🖂. "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)*; pp: 48-57. Doi: <https://elib.vku.udn.vn/handle/123456789/4010> (1 đ)
4. Thi Xuan Tran, Nguyen Quoc Khanh Le, and Van Nui Nguyen🖂. "CLW\_SUMO: A hybrid deep learning model for predicting protein SUMOylation sites." *Journal of Computer Science and Cybernetics 40.4* (2024): 315-325, Doi: <https://doi.org/10.15625/1813-9663/19626> (Tạp chí tin học điều khiển 1.25đ)
5. Tran, T. X., Nguyen, T. T., Le, N. Q. K., V.N Nguyen🖂, KD\_MultiSucc: Incorporating Multi-Teacher Knowledge Distillation and Word Embeddings for Cross-Species Prediction of Protein Succinylation Sites. Biology Methods and Protocols (2025), Doi: <https://doi.org/10.1093/biomethods/bpaf041>. (ISI Q1)

**6. Phương thức chuyển giao, địa chỉ ứng dụng, tác động và lợi ích mang lại của kết quả nghiên cứu:**

Địa chỉ ứng dụng chính của kết quả nghiên cứu là tại Trường Đại học Kinh tế và Quản trị Kinh doanh – Đại học Thái Nguyên, đặc biệt trong các hoạt động giảng dạy, nghiên cứu thuộc các lĩnh vực: học máy, khoa học dữ liệu, khai phá dữ liệu, và tin sinh học. Ngoài ra, kết quả nghiên cứu còn có thể làm nền tảng tham khảo cho các nhóm nghiên cứu, giảng viên và học viên cao học, nghiên cứu sinh trong lĩnh vực khoa học dữ liệu và ứng dụng công nghệ thông tin trong phân tích dữ liệu lớn.

Tác động và lợi ích mang lại:

Về khoa học: Góp phần làm phong phú thêm nguồn tài liệu tham khảo cho các nghiên cứu cơ bản và chuyên sâu về ứng dụng công nghệ thông tin, đặc biệt là trong xử lý và phân tích dữ liệu lớn.

Về đào tạo: Là nguồn tài liệu tham khảo hữu ích phục vụ công tác giảng dạy, xây dựng học liệu và hướng dẫn nghiên cứu khoa học cho sinh viên, học viên cao học, nghiên cứu sinh.

Về ứng dụng thực tiễn: Cung cấp cơ sở lý thuyết và kinh nghiệm thực tiễn để phát triển các ứng dụng tin học hỗ trợ phân tích dữ liệu trong các lĩnh vực kinh tế, quản trị, sinh học và các ngành liên quan.

Về định hướng phát triển: Đặt nền tảng cho việc xây dựng và triển khai các đề tài nghiên cứu tiếp theo có chiều sâu và gắn với thực tiễn ứng dụng tại trường và các cơ sở nghiên cứu khác.

**INFORMATION ON RESEARCH RESULTS**

**1. General information:**

Project title: Research and Integration of Machine Learning Models for Post-Translational Protein Modification Prediction

Code number: ĐH2023-TN08-05

Coordinator: MasterTran Thi Xuan

Implementing institution: Thai Nguyen university of Economics and Business administration

Duration: 24 months, from July 2023 to July 2025

**2. Objective(s):**

Investigate and understand the basic theoretical knowledge related to the mechanism of post-translational transformation (PTM), protein SUMOylation

Propose a model to predict protein SUMOylation site with high performance.

**3. Creativeness and innovativeness:**

The research project has:

Proposed several integrated machine learning models for high-performance prediction of post-translationally modified proteins.

Proposed a knowledge distillation model for high-performance prediction of post-translationally modified proteins.

**4. Research results:**

Summary of basic knowledge related to the mechanism of post-translational transformation (PTM).

It proposes several integrated machine learning models for the high-performance prediction of post-translationally modified proteins, specifically targeting SUMOylation, Ubiquitination, and Succinylation, as follows:

* **A weighted ensemble learning model** is proposed by integrating three machine learning algorithms (SVM, Random Forest, and XGBoost), which significantly improves the prediction performance of SUMOylation.
* **A hybrid deep learning model for SUMOylation prediction** is proposed, which combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. This model is capable of automatically learning hierarchical features from raw protein sequence data, leading to enhanced prediction accuracy.
* **A hybrid CNN-LSTM model for Ubiquitination prediction** is introduced. In this study, both biological and sequential features of protein sequences are extracted and utilized in deep learning frameworks, which substantially boosts the predictive performance.
* **A knowledge distillation-based deep learning model for Succinylation prediction** is proposed. This model leverages the advantages of knowledge distillation techniques and dynamic embeddings to reduce computational complexity while maintaining high accuracy in identifying succinylation sites. Experimental results demonstrate that the proposed predictor outperforms existing tools, providing a valuable contribution to PTM research and biomedical applications.

**5. Products:**

**Scientific products:**

+ 02 article published in the International Journal ISI Q1

+ 01 article published in the International Conference Proceedings (indexed by Springer Scopus Q4).

+ 01 article published in the International Conference Proceedings (1 point)

+ 01 Journal of Computer Science and Cybernetics  (1.25 points)

1. Nguyen Quoc Khanh Le🖂, Thi Xuan Tran, Phung Anh Nguyen, Thi Trang Ho, Van Nui Nguyen🖂, et al. "Recent progress in machine learning approaches for predicting carcinogenicity in drug development." *Expert Opinion on Drug Metabolism & Toxicology just-accepted (2024).* Doi: https://doi.org/10.1080/17425255.2024.2356162. ISI Q1(ISI Q1).
2. Thi Xuan Tran; Van Nui Nguyen; LE, Nguyen Quoc Khanh Le. “Incorporating Natural Language-Based and Sequence-Based Features to Predict Protein SUMOylation Sites”. *In:* *Conference on Information Technology and its Applications. Cham: Springer Nature Switzerland, 2023*. p. 74-88. (Scopus Q4)
3. Thi Xuan Tran, Thi Tuyen Nguyen, Nguyen Quoc Khanh Le, Hong Hai Nguyen, and Van Nui Nguyen🖂. "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);* pp: 48-57. Doi: <https://elib.vku.udn.vn/handle/123456789/4010> (1 đ)
4. Thi Xuan Tran, Nguyen Quoc Khanh Le, and Van Nui Nguyen🖂. "CLW\_SUMO: A hybrid deep learning model for predicting protein SUMOylation sites". *Journal of Computer Science and Cybernetics 40.4 (2024)*: 315-325, Doi: <https://doi.org/10.15625/1813-9663/19626> (Tạp chí tin học điều khiển 1.25đ)
5. Tran, T. X., Nguyen, T. T., Le, N. Q. K., V.N Nguyen🖂, KD\_MultiSucc: Incorporating Multi-Teacher Knowledge Distillation and Word Embeddings for Cross-Species Prediction of Protein Succinylation Sites. Biology Methods and Protocols (2025), Doi: <https://doi.org/10.1093/biomethods/bpaf041>. (ISI Q1)

**6. Transfer alternatives, application institutions, impacts and benefits of research results:**

The primary application of the research results is at the University of Economics and Business Administration – Thai Nguyen University, particularly in teaching and research activities in the fields of machine learning, data science, data mining, and bioinformatics. In addition, the research outcomes can serve as a valuable reference foundation for research groups, lecturers, graduate students, and doctoral candidates working in data science and the application of information technology in big data analysis.

**Impact and Benefits:**

Scientific Impact: Contributes to enriching the body of reference materials for both fundamental and advanced studies on the application of information technology, especially in processing and analyzing large-scale data.

Educational Impact: Serves as a useful reference source for teaching activities, curriculum development, and scientific research guidance for undergraduate students, graduate students, and PhD candidates.s

Practical Application: Provides theoretical foundations and practical experience to support the development of information technology applications for data analysis in the fields of economics, management, biology, and related disciplines.

Development Orientation: Lays the groundwork for developing and implementing further research projects with greater depth and stronger connections to practical applications at the university and other research institutions.