KHÁM PHÁ HỢP CHẤT FLAVONOID VÀ THEAFLAVIN TỪ CHI *CAMELLIA* Ở VIỆT NAM THEO HƯỚNG ỨC CHẾ KEAP1-NRF2 BẰNG PHƯƠNG PHÁP SÀNG LỌC *IN SILICO*

UNVEILING THE FLAVONOID AND THEAFLAVIN COMPOUNDS FROM THE GENUS *CAMELLIA* IN VIETNAM TOWARDS INHIBITING KEAP1-NRF2 BY *IN SILICO* SCREENING METHOD

Nguyễn Minh Quân[1], Đoàn Nguyễn Việt Hà[1], Nguyễn Bùi Quốc Huy[2], Giang Thị Kim Liên[2], Lê Nguyễn Thiên Hân[3], Nguyễn Minh Hiền[3]*

[1]Trường THPT chuyên Trần Đại Nghĩa, Việt Nam
[2]Viện Nghiên cứu và Đào tạo Việt-Anh - Đại học Đà Nẵng, Việt Nam
[3]Trường Đại học Khoa học Sức khoẻ, Đại học Quốc gia Thành phố Hồ Chí Minh, Việt Nam

*Tác giả liên hệ / Corresponding author: nmhien@uhsvnu.edu.vn

(Nhận bài / Received: 24/9/2024; Sửa bài / Revised: 24/11/2024; Chấp nhận đăng / Accepted: 25/11/2024)

Tóm tắt - Các cây thuộc chi Chè *Camellia* từ lâu đã được chứng minh có khả năng chống stress oxy hóa thông qua sự dập tắt các gốc tự do. Nghiên cứu này sử dụng phương pháp sàng lọc ảo *in silico* tích hợp thuật toán học máy để dự đoán khả năng chống oxy hóa thông qua việc tăng cường biểu hiện Nrf2 của 5 loài thuộc chi chè *Camellia* bao gồm *Camellia sinensis*, *Camellia quephongensis*, *Camellia oleifera*, *Camellia amplexicaulis* và *Camellia japonica*. Mô hình học máy phân loại hợp chất được xây dựng dựa trên bốn thuật toán học máy bao gồm Support vector machines (SVM), Random forests (RF), Extreme gradient boosting (XGBoost) và Multilayer Perceptron (MLP). Từ mô hình phân loại có hiệu suất tối ưu, nghiên cứu đã xác định được 33 hợp chất tiềm năng. Các hợp chất được phân loại có khả năng kích hoạt Nrf2 được docking phân tử trên với thụ thể Keap1-Nrf2 (PDB ID: 2FLU). Kết quả cho thấy, có bốn hợp chất với số điểm docking tốt nhất là camellianoside (-10,4 kcal/mol), theaflavin-3-gallate (-9,9 kcal/mol), theaflavin-3'-gallate (-9,8 kcal/mol) và camelliaside B (-9,7 kcal/mol).

Từ khóa - Mô hình học máy; ức chế phức hợp Keap1-Nrf2; chi *Camellia*; docking phân tử; *in silico*.

Abstract - Species from the *Camellia* genus have long been shown antioxidant properties through the quenching of free radicals. This study employed an *in silico* virtual screening approach, integrating machine learning algorithms to predict the antioxidant potential of five species from the *Camellia* genus, namely *Camellia sinensis*, *Camellia quephongensis*, *Camellia oleifera*, *Camellia amplexicaulis*, and *Camellia japonica*, by evaluating the capacity to inhibit Keap1-Nrf2 complex, indirectly enhance Nrf2 expression. Four machine learning algorithms including Support Vector Machines (SVM), Random Forests (RF), Extreme Gradient Boosting (XGBoost), and Multilayer Perceptron (MLP)-were utilized to build a classification model for predicting compound activity. Based on the top-performing model, 33 promising compounds were identified. These Nrf2-activating compounds were further analyzed through molecular docking with the Keap1-Nrf2 complex (PDB ID: 2FLU). The docking results highlighted four compounds with the most favorable binding affinities: camellianoside (-10.4 kcal/mol), theaflavin-3-gallate (-9.9 kcal/mol), theaflavin-3'-gallate (-9.8 kcal/mol), and camelliaside B (-9.7 kcal/mol).

Key words - Machine learning; Keap1-Nrf2 inhibitor; *Camellia* genus; molecular docking; *in silico*.

1. Đặt vấn đề

Stress oxy hóa là tình trạng mất cân bằng giữa các gốc tự do và chất kiểm soát biểu hiện chống oxy hóa do nguyên nhân nội sinh như căng thẳng thần kinh, viêm nhiễm và nguyên nhân ngoại sinh như chế độ ăn uống, vận động quá mức, tia bức xạ, ô nhiễm môi trường [1]. Trong điều kiện bình thường, các chất oxy hóa là sản phẩm phụ của các hoạt động sống của cơ thể, có một số vai trò sinh lý như xúc tác enzyme, kích hoạt con đường truyền tín hiệu tế bào, vận chuyển điện tử nên được duy trì ở một mức độ ổn định trong cơ thể [2]. Song, tình trạng stress oxy hóa mạn tính có thể gây tổn thương đến các tế bào và mô xung quanh, kích hoạt phản ứng viêm làm thúc đẩy tiến trình lão hóa và nhiều bệnh lý nguy hiểm như tim mạch, thoái hóa thần kinh, rối loạn chuyển hóa, ung thư [1].

Nuclear factor erythroid-2 p45-related factor 2 (Nrf2) là yếu tố phiên mã quan trọng điều hòa hoạt động các gen bảo vệ tế bào chống lại stress oxy hóa. Ở điều kiện bình thường, trong cơ thể con người, mật độ Nrf2 tự do thường ở mức thấp vì chúng bị giữ bởi protein Keap1 (Kelch-like ECH-associated protein 1) và bị ubiquitin hóa bởi phức hợp E3 ubiquitin ligase (Cul3 - Rbx1). Keap1 là một protein nhạy cảm với stress oxy hóa [3], gồm 624 acid amin và bốn miền chức năng riêng biệt: miền BTB (Broad complex–Tramtrack–Bric-a-brac), vùng can thiệp (IVR - Intervening Region), miền lặp glycine kép (DGR - Double Glycine Repeats), và miền C-terminal [4]. Miền BTB đảm nhiệm vai trò liên kết với Cullin3/E3 ligase, thúc đẩy quá trình phân hủy Nrf2 trong điều kiện không có stress oxy hóa. Miền DGR có cấu trúc β-propeller sáu cánh, cho phép Keap1 liên

[1] Tran Dai Nghia High School for the Gifted, Vietnam (Nguyen Minh Quan, Doan Nguyen Viet Ha)
[2] The University of Danang - VN-UK Institute for Research and Executive Education, Vietnam (Nguyen Bui Quoc Huy, Giang Thi Kim Lien)
[3] University of Health Sciences, Vietnam National University, Ho Chi Minh City – VNU-HCM, Vietnam (Le Nguyen Thien Han, Nguyen Minh Hien)

kết với miền Neh2 của Nrf2 để điều hòa hoạt động của nó. Vùng IVR, nằm giữa BTB và DGR, giúp Keap1 di chuyển khỏi nhân và hoạt động chủ yếu trong tế bào chất. Đặc biệt, miền IVR và BTB chứa các gốc cysteine nhạy cảm, như Cys151, Cys273, và Cys288, có thể phản ứng với các tác nhân oxy hóa gây stress, dẫn đến thay đổi cấu trúc và chức năng của Keap1, từ đó làm giảm liên kết với Nrf2, giải phóng Nrf2 vào nhân và kích hoạt các gene bảo vệ tế bào [5, 6]. Khi có nhiều gốc chứa oxy có hoạt tính (Reactive oxygen species – ROS), amino acid cystein của Keap1 sẽ bị biến đổi dẫn tới sự phân tách của phức hợp Keap1-Nrf2. Nrf2 sau đó đi vào nhân tế bào, dị hợp với các protein sMAF và tạo ra phức hợp phiên mã có khả năng liên kết với ARE (Antioxidant response element), làm biểu hiện các gen chống oxy hóa do ARE kiểm soát giúp chống lại ROS tấn công tế bào [7].

Hiện nay, các nghiên cứu đã xác định được hai cơ chế chính giúp tăng biểu hiện Nrf2 bao gồm (1) phản ứng ái điện tử với cystein ở Keap1 để bất hoạt Keap1 [8, 9] và (2) ngăn chặn sự hình thành phức hợp giữa Keap1-Nrf2 bằng cách bắt chước cấu trúc miền liên kết Neh2 với Kelch [10]. Chất kích hoạt Nrf2 thành công nhất cho đến nay là este acid fumaric, dimethyl fumarate (DMF) của BG-12 được FDA công nhận vào năm 2013 trong điều trị bệnh đa xơ cứng tái phát-thuyên giảm (Relapsing-Remitting Multiple Sclerosis) [11]. Bên cạnh đó, sulforaphane (SFN), một isothiocyanate tìm thấy dồi dào trong họ Thập tự (Cruciferae), được chứng minh có tiềm năng trong việc tăng biểu hiện của Nrf2 được sử dụng trong điều trị bệnh đái tháo đường tuýp 2 [12]. Cơ chế chống oxy hóa của DMF và SFN đều dựa trên việc làm biến đổi nhiều vùng liên kết khác nhau của Keap1, cụ thể là phản ứng ái điện tử của DMF và SFN với Cys151 ở vùng BTB của Keap1 làm tăng biểu hiện Nrf2 [13, 14]. Khi Cys151 bị biến đổi, khả năng tạo thành phức hợp Cul3/Rbx1 với Keap1 sẽ giảm xuống, dẫn tới việc Keap1 vẫn liên kết với Nrf2 và Nrf2 mới tổng hợp sẽ tự do trong môi trường nội bào do không còn bị ubiquitin hóa và thực hiện hoạt động phiên mã [15]. Tuy nhiên, các chất ức chế tương tác protein-protein giữa Keap1-Nrf2 có tính chọn lọc cao hơn các chất ái điện tử vì dựa trên việc bắt chước mô típ ETGE của Nrf2 có cấu trúc phiến gập β đặc trưng gắn vào miền Kelch của Keap1 thông qua các tương tác kỵ nước và tĩnh điện [12]. Các chất ức chế tương tác Keap1-Nrf2 hiện nay đang được nghiên cứu là LH601A, RA839, tetrahydroisoquinolin, thiopyrimidin, naphthalen, carbazon [16, 17].

Chi Chè (*Camellia*) là một thức uống dân dã, phổ biến ở Việt Nam và đã được chứng minh có nhiều tác dụng về dược lý như chống oxy hóa, kháng viêm, ngăn ngừa ung thư; trong số đó, tác dụng dược lý được quan tâm hơn cả khả năng chống oxy hóa [18, 19]. Nhiều nghiên cứu trước đây đã cho thấy, khả năng dập tắt gốc tự do H_2O_2 và DPPH (2,2-diphenyl-1-picrylhydrazyl) đáng kể của nhiều cây thuộc loài *Camellia sinensis*. Đồng thời, các nghiên cứu cũng chỉ ra các loài *Camellia sinensis*, *Camellia oleifera* và *Camellia japonica* chứa các polyphenol có khả năng dập tắt gốc tự do như epicatechin (EC), epigallocatechin (EGC), epicatechin-3-gallate (ECG) và epigallocatechin-3-gallate (EGCG) [20].

Tuy nhiên, vẫn có rất ít các nghiên cứu đi sâu vào tìm hiểu khả năng kháng oxy hóa thông qua sự ức chế phức hợp Keap1-Nrf2 của các chất có trong chi Chè. Phương pháp sàng lọc ảo *in silico* với ưu thế tiện lợi, tiết kiệm và khả năng xử lý và dự đoán số lượng hợp chất vô cùng lớn đang càng trở nên phổ biến nhưng ở Việt Nam phương pháp *in silico* vẫn ít được sử dụng. Vì vậy, nghiên cứu tập trung sử dụng phương pháp *in silico* để khảo sát khả năng kháng stress oxy hóa thông qua sự ức chế phức hợp Keap1-Nrf2 của 5 loài thuộc chi chè *Camellia* bao gồm *C. sinensis* (Trà xanh), *C. oleifera* (Hoa Sở), *C. quephongensis* (Trà hoa vàng Quế Phong), *C. amplexicaulis* (Hải đường Việt Nam) và *C. japonica* (Trà mi).

2. Nguyên liệu

2.1. Các cơ sở dữ liệu: PubChem, ChEMBL, PDB

Trong nghiên cứu này, 3 cơ sở dữ liệu miễn phí được sử dụng để thu thập các dữ liệu cần thiết bao gồm PubChem, ChEMBL và Protein Data Bank (PDB).

PubChem (https://pubchem.ncbi.nlm.nih.gov) là cơ sở dữ liệu hóa học mở được xây dựng bởi NLM (National Library of Medicine), một viện nghiên cứu trực thuộc NIH (U.S. National Institutes of Health). Nghiên cứu sử dụng cơ sở dữ liệu PubChem để trích xuất định dạng canonical SMILES của các hợp chất trong chi chè *Camellia*.

ChEMBL (https://www.ebi.ac.uk/chembl/) là cơ sở dữ liệu về các phân tử có hoạt tính sinh học được duy trì bởi EBI (European Bioinformatics Institute). Nghiên cứu sử dụng cơ sở dữ liệu ChEMBL để thu thập các hợp chất đã được chứng minh có khả năng ức chế phức hợp Keap1-Nrf2.

Trong nghiên cứu này, cơ sở dữ liệu Protein Data Bank (https://www.rcsb.org/) được sử dụng nhằm thu thập cấu trúc 3D của phức hợp protein Keap1-Nrf2 (PDB ID: 2FLU) để sử dụng trong mô hình docking phân tử.

2.2. Google colaboratory

Google colaboratory là phần mềm cho phép thực thi Python trên nền tảng đám mây. Nghiên cứu đã sử dụng phần mềm để xây dựng các mô hình học máy bằng ngôn ngữ Python và trích xuất các kết quả dữ liệu.

2.3. AutoDock Vina

Để xây dựng mô hình docking phân tử, nhóm nghiên cứu sử dụng phần mềm AutoDock Vina được phát triển bởi viện nghiên cứu Scripps (Mỹ). Phần mềm AutoDock Vina có ưu thế là sử dụng phương pháp tối ưu hóa Gradient trong tính toán và dự đoán vị trí liên kết của phối tử với thụ thể, giúp tối ưu hóa tốc độ tính toán và tăng độ chính xác của các dự đoán [21].

2.4. BIOVIA

Nghiên cứu sử dụng phần mềm BIOVIA Discovery Studio Visualizer (phiên bản 21.1.0) để mô phỏng 2D và 3D kết quả docking phân tử và các liên kết của phối tử với thụ thể trong mô hình docking.

2.5. Swiss ADME

Nghiên cứu đã sử dụng phần mềm SwissADME (http://www.swissadme.ch/index.php) để đánh giá tiềm năng dược lý của các hợp chất thuộc chi chè *Camellia* [22].

3. Phương pháp nghiên cứu

3.1. Chuẩn bị dữ liệu để sàng lọc ảo

Phức hợp Keap1-Nrf2 của người (ID: CHEMBL3038498) được sử dụng là thụ thể gắn kết nhằm

khảo sát các hợp chất có tiềm năng ức chế phức hợp Keap1-Nrf2. Nghiên cứu này sử dụng thư viện hợp chất ChEMBL để thu thập dữ liệu về các hợp chất có khả năng ức chế phức hợp Keap1-Nrf2 bằng cách sử dụng từ khóa "inhibit Keap1-Nrf2" hoặc "inhibit Keap1". Đặc tính và cấu trúc hóa học của các hợp chất sẽ được dùng làm dữ liệu đầu vào; giá trị IC_{50} của hợp chất được sử dụng để phân loại hợp chất thành "có hoạt tính" đối với hợp chất có giá trị IC_{50} nhỏ hơn giá trị ngưỡng và "không có hoạt tính" đối với hợp chất có giá trị IC_{50} lớn hơn hoặc bằng giá trị ngưỡng. Nghiên cứu sử dụng bốn phương pháp để phân loại hợp chất tương ứng dựa trên hai giá trị IC_{50} ngưỡng là 10 µM và 5 µM kết hợp với phương pháp tạo hợp chất không có hoạt tính "decoy" bằng công cụ DUDEZ.

3.1.1. Cơ sở dữ liệu các hợp chất trong chi chè Camellia

Dựa trên các nghiên cứu trước đây về các hợp chất có trong 5 loài thuộc chi chè *Camellia*, nhóm tiến hành tổng hợp dữ liệu từ các bài báo khoa học, cơ sở dữ liệu với các từ khóa như "Chemical constitute", "Chemical composition" và đã xây dựng cơ sở dữ liệu gồm 399 hợp chất để sàng lọc ảo với các mục: loài cây, nhóm hợp chất, tên hợp chất và định dạng canonical SMILES được truy xuất từ cơ sở dữ liệu PubChem.

3.1.2. Phương pháp mã hóa cấu trúc hóa học

Nghiên cứu sử dụng 4 phương pháp mã hóa cấu trúc hóa học khác nhau bao gồm: dấu vân tay Morgan 2, Morgan 3 [23], MACCS [24] và RDKit [25].

3.1.3. Tập dữ liệu huấn luyện và kiểm tra

Đối với từng tập dữ liệu huấn luyện, 80% số hợp chất sẽ được sử dụng để huấn luyện mô hình và 20% số hợp chất sẽ được sử dụng để đánh giá mô hình học máy. Các công thức hóa học của hợp chất được biểu diễn bằng định dạng SMILES, và được chuyển thành vectơ nhị phân bằng các công cụ mã hóa gồm MACCS, Morgan 2, Morgan 3 và Rdk5.

3.1.4. Phương pháp tạo hợp chất không có hoạt tính "decoy"

Nghiên cứu tiến hành sử dụng công cụ DUDEZ (https://tldr.docking.org/) để tạo hợp chất không có hoạt tính "decoy" cho cả 2 tập dữ liệu huấn luyện để tạo ra thêm 2 tập dữ liệu ứng với 2 giá trị ngưỡng IC_{50} tương ứng là 10 µM và 5 µM.

3.2. Xây dựng mô hình học máy dự đoán các chất ức chế phức hợp Keap1-Nrf2

3.2.1. Mô hình thuật toán sử dụng trong sàng lọc ảo

Để xây dựng mô hình học máy, bốn thuật toán phân loại khác nhau bao gồm Support Vector Machines (SVM), Random Forests (RF), Extreme Gradient Boosting (XGBoost) và Multilayer Perceptron (MLP) đã được sử dụng. Các siêu tham số của thuật toán sử dụng cho 4 lần chạy với 4 tập dữ liệu huấn luyện là tương tự nhau:

- Mô hình SVM: Hàm kernel được cài đặt là linear hoặc rbf và tham số gamma được cài đặt là scale. Tham số C được cài đặt là 2 tới 10 và tham số degree là 2 tới 5.

- Mô hình RF: Giá trị max_depth được cài đặt là 10, 20, 30; min_samples_leaf là 1, 2 và min_samples_split là 2, 3. Giá trị n_estimators được cài đặt là 50, 100, 200, 400 và tham số max_features là sqrt hoặc log2.

- Mô hình XGBoost: Giá trị max_depth được cài đặt là 2, 3, 5, 8, 10, 15. Tham số booster là gbtree hoặc gblinear. Giá trị learning_rate là 0,05; 0,1; 0,15; 0,20. Giá trị min_child_weight là 1, 2, 3, 4. Giá trị n_estimators là 100, 200, 300, 500, 900, 1100.

- Mô hình MLP: Tham số hidden_layer_sizes bao gồm các cặp số (16; 8), (32; 16), (64; 32), (128; 64). Tham số activation được cài đặt là relu hoặc tanh và solver là adam hoặc sgd. Tham số learning_rate được cài đặt là constant hoặc adaptive. Giá trị batch_size được cài đặt là 32, 64, 128 và giá trị alpha là 0,001; 0,01; 0,1.

3.2.2. Các thông số đánh giá hiệu suất mô hình

Hiệu suất của mô hình được đánh giá bằng các thông số: Độ chính xác (accuracy), độ nhạy (sensitivity), độ đặc hiệu (specificity), diện tích dưới đường cong biểu diễn đặc điểm hoạt động của thuật toán (Area Under the Curve - Receiver Operating Characteristic - AUC - ROC) [26, 27]. Các thông số được tính dựa trên bốn giá trị là số hợp chất có hoạt tính thật (true positive - TP), số hợp chất có hoạt tính giả (false positive - FP), hợp chất không có hoạt tính thật (true negative - TN), số hợp chất không có hoạt tính giả (false negative - FN). Các thông số được tính toán như sau:

- Độ chính xác = $\frac{TP+TN}{TP+FP+FN+TN}$

- Độ nhạy (tỷ lệ hợp chất có hoạt tính thật (true positive rate – TPR)) = $\frac{TP}{TP+FN}$

- Độ đặc hiệu = $\frac{TN}{TN+FP}$

- Tỷ lệ hợp chất không có hoạt tính thật (false positive rate – FPR) = $\frac{FP}{TN+FP}$

- AUC= $\int_1^0 TPR\,(x).FPR\,(x) = 2\int_1^0 (1-x^2).(1-x)\partial x$

3.2.3. Tối ưu hóa thuật toán

Phương pháp tìm kiếm lưới (gridsearch) và xác thực chéo năm lần (5-fold cross validation) được sử dụng để điều chỉnh siêu tham số cho các mô hình học máy. Siêu tham số tối ưu nhất cho từng thuật toán sẽ được xác định thông qua phương pháp grid search, thuật toán tối ưu nhất của một phương pháp chuẩn bị dữ liệu sẽ được xác định dựa trên giá trị độ chính xác (accuracy), diện tích dưới đường cong (AUC), độ đặc hiệu (specificity). Bốn thuật toán tối ưu nhất ứng với bốn phương pháp chuẩn bị dữ liệu sẽ được so sánh và chọn ra mô hình dự đoán tốt nhất.

3.3. Phân tích tương tác và khả năng gắn kết của các chất được dự đoán có tiềm năng kích hoạt biểu hiện Nrf2 trên protein mục tiêu Keap1 sử dụng mô hình docking phân tử

Nhóm tiến hành docking phân tử với 33 hợp chất được sàng lọc với thụ thể là phức hợp Keap1-Nrf2 (PDB ID: 2FLU). Nghiên cứu sử dụng phần mềm Autodock Vina để mô phỏng và tính toán khả năng liên kết của phối tử và thụ thể. Vị trí docking được chọn là vùng tương tác giữa Nrf2 và Keap1 tại tọa độ: x = 8,11; y = 16,25; z = 9,34 với kích thước: x = 26,65; y = 36,52; z = 43,72. Điểm số docking được sử dụng để đánh giá khả năng liên kết.

3.4. Đánh giá điểm số ADME

ADME là các chỉ số dùng để đánh giá tiềm năng dược lý của một hợp chất, bao gồm Absorption - độ hấp thụ,

Distribution - sự phân bố trong cơ thể, Metabolism - sự chuyển hóa và Excretion - sự thải trừ. Nghiên cứu đã tiến hành đánh giá điểm số ADME của 4 hợp chất có khả năng liên kết với phức hợp Keap1-Nrf2 tốt nhất bằng công cụ SwissADME (http://www.swissadme.ch/index.php).

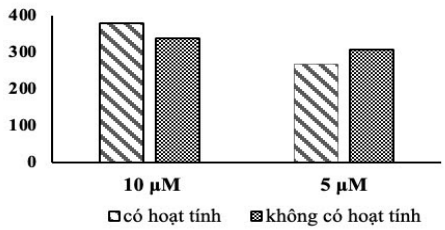
4. Kết quả nghiên cứu

4.1. Kết quả sàng lọc hợp chất từ cơ sở dữ liệu ChEMBL và đánh giá hiệu suất mô hình học máy

4.1.1. Kết quả sàng lọc hợp chất từ cơ sở dữ liệu ChEMBL

Đối với ngưỡng giá trị IC_{50} = 10 µM, có 379 hợp chất được phân loại là "có hoạt tính" (active) và 268 hợp chất được phân loại là "không có hoạt tính" (inactive). Đối với ngưỡng giá trị IC_{50} = 5 µM, có 338 hợp chất được phân loại là "Có hoạt tính" và 309 hợp chất được phân loại là "Không có hoạt tính" (Hình 1). Trong cả 2 trường hợp, số hợp chất "Không có hoạt tính" đều là nhóm ít hơn và đều chiếm >40% tổng số hợp chất, cho thấy cả 2 tập dữ liệu ít bị mất cân bằng và phù hợp để huấn luyện thuật toán.

Khi sử dụng công cụ DUDEZ để tạo hợp chất không có hoạt tính "decoy", kết quả đối với IC_{50} ngưỡng là 10 µM, 379 chất có hoạt tính - 379 chất không có hoạt tính, trong khi đối với IC_{50} ngưỡng = 5 µM, sử dụng decoy: 338 chất có hoạt tính - 338 chất không có hoạt tính.



Hình 2. *Phân loại các hoạt chất có hoạt tính và không có hoạt tính với ngưỡng IC_{50} = 10 µM và IC_{50} = 5 µM*

4.1.2. Hiệu suất mô hình học máy

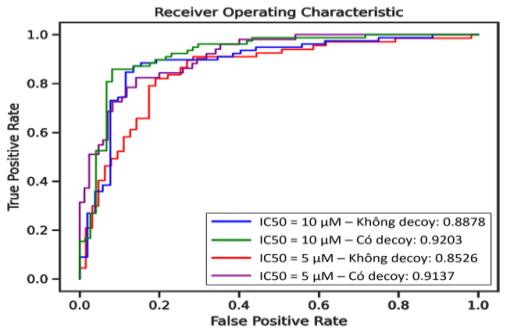
Mỗi tập dữ liệu huấn luyện ứng với từng phương pháp chuẩn bị dữ liệu sẽ được dùng để xây dựng một mô hình phân loại hợp chất. Để xây dựng mô hình, bốn thuật toán phân loại khác nhau đã được sử dụng bao gồm SVM, RF, XGBoost và MLP. Giá trị độ chính xác, AUC, và độ đặc hiệu được chọn làm các giá trị để xác định hiệu suất của mô hình học máy. Siêu tham số tối ưu của các thuật toán ứng với 4 phương pháp mã hóa, sử dụng 4 tập dữ liệu huấn luyện được thể hiện ở Phụ lục. Thuật toán, siêu tham số của thuật toán và phương pháp mã hóa cấu trúc tối ưu của các mô hình ứng với từng phương pháp chuẩn bị bộ dữ liệu huấn luyện và các giá trị đánh giá hiệu suất của mô hình được thể hiện ở Bảng 1.

Nghiên cứu xác định mô hình học máy SVM, phương pháp mã hóa cấu trúc Morgan 3 và phương pháp chuẩn bị bộ dữ liệu huấn luyện với giá trị IC_{50} = 10 µM kết hợp ecoy là mô hình học máy tối ưu nhất với độ chính xác cao nhất là 0,8487 và AUC cao nhất là 0,9203 (Bảng 1).

Nghiên cứu đồng thời tính toán chỉ số tương đồng Tanimoto giữa 647 hợp chất được sử dụng trong xây dựng mô hình học máy và 399 hợp chất trong bộ dữ liệu thử nghiệm được mã hóa bằng phương pháp Morgan 3. Kết quả chỉ số Tanimoto giữa tập dữ liệu huấn luyện và tập dữ liệu thử nghiệm đều nhỏ hơn 0,3 [28]. Điều này cho thấy, tập dữ liệu huấn luyện đủ đa dạng và có thể sử dụng để huấn luyện thuật toán. Từ các kết quả trên, nghiên cứu sử dụng mô hình SVM, kết hợp phương pháp mã hóa cấu trúc Morgan 3 và phương pháp chuẩn bị bộ dữ liệu huấn luyện với giá trị ngưỡng IC_{50} = 10 µM làm mô hình phân loại các hợp chất có khả năng ức chế phức hợp Keap1-Nrf2.

Bảng 1. *Mô hình học máy và hiệu suất mô hình học máy*

	IC_{50} = 10 µM		**IC_{50} = 5 µM**	
	Không decoy	**Có decoy**	**Không decoy**	**Có decoy**
Thuật toán	SVM	SVM	SVM	RF
Phương pháp mã hóa	Morgan3	Morgan3	Morgan3	Morgan2
Siêu tham số	C: 3 degree: 2 gamma: scale kernel: rbf	C: 2 degree: 2 gamma: scale kernel: rbf	C: 3 degree: 2 gamma: scale kernel: rbf	max_depth: 30 max_features: sqrt min_samples_leaf: 2 min_samples_split: 2 n_estimators: 50
Độ chính xác	0,8385	**0,8487**	0,8077	0,8309
AUC	0,8878	**0,9203**	0,8526	0,9137
Độ nhạy	0,8974	**0,8718**	0,8657	0,8235
Độ đặc hiệu	0,7500	**0,8243**	0,7460	0,8353



Hình 3. *Đồ thị biểu diễn đường cong ROC (Receiver Operating Characteristic) và giá trị diện tích dưới đường cong (Area under curve)*

4.2. Kết quả dự đoán hợp chất tiềm năng sử dụng mô hình học máy tối ưu

Sử dụng tập dữ liệu huấn luyện, thuật toán và phương pháp mã hóa tối ưu nhất (IC_{50} = 10 µM, có decoy-SVM-Morgan 3), nghiên cứu tiến hành phân loại các hợp chất tự nhiên trong 5 cây thuộc chi Chè *Camellia*. Công thức phân tử của 399 hợp chất thuộc 5 loài *Camellia* được viết dưới định dạng SMILES và chuyển thành chuỗi vectơ nhị phân bằng phương pháp Morgan 3 để thuật toán phân loại. Nghiên cứu xác định có 33 trong 399 hợp chất được phân loại là có tiềm năng ức chế phức hợp Keap1-Nrf2. Trong các loài, 2 loài được xác định là có nhiều hợp chất tiềm năng nhất là *C. japonica* với 13 hợp chất và *C. sinesis* với 11 hợp chất (Bảng 2).

Nghiên cứu sau đó tiến hành docking phân tử đối với 33 hợp chất tiềm năng trên với protein đích là phức hợp Keap1-Nrf2 (PDB ID: 2FLU) sử dụng phần mềm Autodock Vina. Kết quả cho thấy trong 33 hợp chất, 4 hợp chất có ái lực liên kết với thụ thể Keap1 cao nhất dựa vào điểm số docking lần lượt là camellianoside (-10,4 kcal/mol), theaflavin-3-gallate (-9,9 kcal/mol), theaflavin-3'-gallate (-9,8 kcal/mol), camelliaside B (-9,7 kcal/mol).

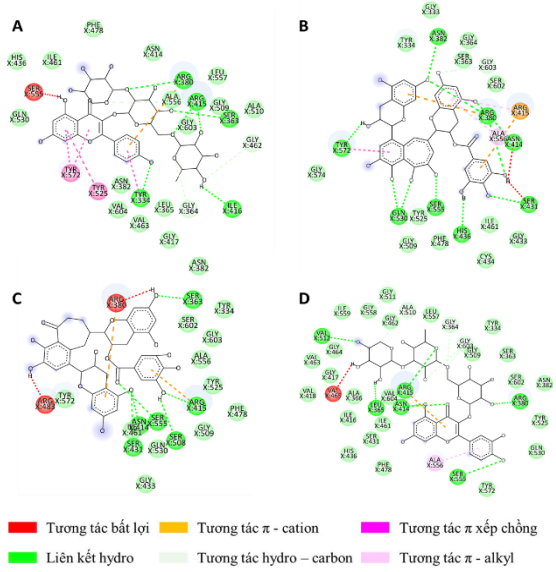
Bảng 2. *Điểm số docking của hợp chất tiềm năng từ kết quả sàng lọc sử dụng mô hình học máy SVM-Morgan 3*

Hợp chất	Điểm số docking	Cây
Rutin	-9,3	*C. quephongensis* *C. oleifera* *C. sinensis*
Camelliaside B	**-9,7**	*C. oleifera*
Camelliaside A	-8,3	*C. oleifera*
Camellianoside	**-10,4**	*C. japonica*
Camelliatannin A	-9,3	*C. japonica*
Camelliatannin C	-9,2	*C. japonica*
Camelliatannin H	NC*	*C. japonica*
Camelliatannin D	NC*	*C. japonica*
Camelliatannin F	-9,6	*C. japonica*
Camelliatannin G	-9	*C. japonica*
Chakasaponins I	-7,3	*C. sinensis*
Chakasaponins III	-7,7	*C. sinensis*
Floratheasaponins A	-7,6	*C. sinensis*
Floratheasaponins B	-7,7	*C. sinensis*
Floratheasaponins C	-7	*C. sinensis*
Floratheasaponins D	-7,7	*C. sinensis*
Floratheasaponins E	-6,6	*C. sinensis*
Floratheasaponins F	-7,3	*C. sinensis*
Camelliasaponins A2	-7,4	*C. japonica*
Camelliasaponins C1	-7,6	*C. japonica*
Theasaponin E1	-7,9	*C. japonica*
Yuchasaponin A	-7,8	*C. oleifera*
Yuchasaponin B	-7,7	*C. oleifera*
Sasanquasaponin	-8,3	*C. oleifera*
Camelliasaponin B2	-7,7	*C. oleifera*
Theaflavin-3'-gallate	**-9,8**	*C. sinensis*
Theaflavin-3-gallate	**-9,9**	*C. sinensis*
Camelliins A	NC*	*C. japonica*
Camelliins B	NC*	*C. japonica*
Camellioside D	-8	*C. japonica*
Oleanolic acid-3-O-β-D-glucopyranoside	-8,5	*C. oleifera*
Camelliquercetiside C	-8,8	*C. amplexicaulis*
Camellioside A	-8,7	*C. amplexicaulis*

*NC: Không tính toán được do phân tử có nhiều hơn 100 nguyên tử khác hydrogen

4.3. Điểm số docking và đánh giá khả năng gắn kết của các chất được dự đoán

Mô phỏng 2D các tương tác ở Hình 3 cho thấy tất cả tương tác của bốn hợp chất với các amino acid của Keap1 là liên kết hydro, tương tác π-cation, π xếp chồng và π-alkyl. Các nhóm hydroxyl của các phối tử đóng vai trò là chất cho hydro cho nhóm amin trên Val465, Leu365, Tyr572, His436, Asn414, Ser431, Arg380, Arg483, Ser508, Ser555 và Ile416. Bốn hợp chất trên có điểm chung là đều có tương tác với Arg380, Arg415, Ser555 với vai trò chủ yếu là chất cho electron. Một số vị trí tương tác mới xuất hiện lần đầu là với các amino acid Val512, Ala 510, Val465, Gly603, His436, Arg483, Gly643, Gly462, chủ yếu là liên kết hydro. Các vị trí gắn kết Tyr334, Ser363, Leu365, Asn382, Tyr525, Ala556, Asn414, Ile416, Gln530, Ser508, Ser555, Arg415, Tyr572 tương đồng với các nghiên cứu khác khi thực hiện docking hợp chất trên thụ thể là phức hợp Keap1-Nrf2 [29,30]. Hơn nữa, một số nghiên cứu trước đã chỉ ra rằng Tyr334, Ser363, Arg380, Asn382, Arg415, Arg483 và Ser508 là các amino acid của Keap1 liên kết với miền Neh2 của Nrf2, đóng vai trò quan trọng trong việc hình thành phức hợp Keap1-Nrf2 [29,30]. Việc bắt chước càng giống tư thế của Nrf2 khi hình thành phức hợp Keap1-Nrf2 được cho là gia tăng khả năng liên kết của các hợp chất với thụ thể Keap1.



Hình 4. *Mô phỏng 2D tương tác các acid amin của protein Keap1 và camellianoside (A), theaflavin-3-gallate (B), theaflavin-3'-gallate (C) và camelliaside B (D)*

Bảng 3. *Loại tương tác và vị trí amino acid của Keap1 tương tác với bốn hợp chất có điểm số docking cao nhất*

Hợp chất	Loại tương tác	Vị trí amino acid tương tác
Camellianoside	Tương tác bất lợi*	Val465
	Liên kết hydro	Val512, Ala510, Gly603, Ala556, Arg380, Ser555, Arg415, Asn414, Leu365
	π-cation	Arg415
	π-alkyl	Ala556
Theaflavin-3-gallate	Tương tác bất lợi	Asn414, Ser431
	Liên kết hydro	Asn382, Arg380, Asn414, Ser431, His436, Ser555, Gln530, Tyr572
	π-cation	Arg415, Arg380
	π xếp chồng	Tyr572
	π-alkyl	Arg415, Ala556
Theaflavin-3'-gallate	Tương tác bất lợi	Arg483, Arg380
	Liên kết hydro	Ser363, Arg415, Ser508, Ser555, Asn414, Ser431
	π-cation	Arg380, Arg415
Camelliaside B	Tương tác bất lợi	Ser555
	Liên kết hydro	Tyr334, Gly364, Ile416, Gly462, Arg415, Ser363, Arg380
	π-cation	Arg380
	π xếp chồng	Tyr572, Tyr525, Tyr334

**Tương tác bất lợi biểu thị các tương tác đẩy/hút bất lợi giữa thụ thể và phối tử, có thể gây ảnh hưởng đến độ bền giữa phức hợp của phối tử và thụ thể trong các nghiên cứu docking [31].*

4.4. Điểm số ADME - dự đoán khả năng hấp thụ các hợp chất của cơ thể con người

Nghiên cứu đã tiến hành đánh giá điểm số ADME của bốn hợp chất có khả năng liên kết với Keap1 tốt nhất bằng công cụ SwissADME (http://www.swissadme.ch/index.php).

Kết quả SwissADME cho thấy cả bốn hợp chất đều thỏa mãn được giá trị Log P ≤ 5, cho thấy khả năng hòa tan trong dịch bào và xuyên qua màng tế bào. Ngoài ra, giá trị độ tan trong nước của hợp chất (Log S) được tính toán theo 2 phương pháp là ESOL và Ali. Theo SwissADME, kết quả Log S sẽ được phân loại như sau: Không tan (Log S < -10), tan kém (-10 < Log S < -6), tan ở mức vừa phải (-6 < Log S < -4), tan trong nước (-4 < Log S < -2), tan tốt trong nước (-2 < Log S < 0) và tan rất tốt trong nước (0 < Log S). Trong cả hai phương pháp tính, camellianoside và camelliaside B đều được xác định là ở mức "tan trong nước". Hai hợp chất theaflavin-3-gallate và theaflavin-3'-gallate đều cho thấy kết quả hầu hết ở mức "tan kém" (Bảng 4).

Bảng 4. *Một số tính chất lý hoá của bốn hợp chất tiềm năng dựa trên điểm số ADME*

Hợp chất	Phân tử khối (amu)	Số lượng nhóm nhận hydro	Số lượng nhóm cho hydro	Tính ưa mỡ	Độ tan trong nước	
				Log P	Log S (ESOL)	Log P
Camellianoside	742,63	20	12	-2,48	-2,97	-4,52
Theaflavin-3-gallate	716,6	16	11	1,35	-5,49	-7,38
Theaflavin-3'-gallate	704,63	15	10	2,03	-6,7	-9,18
Camelliaside B	726,63	19	11	-2,34	-2,5	-3,47

Bảng 5. *Một số tính chất về độ giống thuốc của bốn hợp chất tiềm năng dựa trên điểm số ADME*

	GI absorption	BBB permeant	P-gp substrate	CYP1A2 inhibitor	CYP2C19 inhibitor	CYP2C9 inbitor	CYP2D6 inhibitor	CYP3A4 inhibitor	Log Kp (cm/s)
Camelliaside B	Low	No	No	No	No	No	No	No	-12,48
Theaflavin-3'-gallate (TF2B)	Low	No	No	No	No	Yes	No	No	-7,81
Theaflavin-3-gallate (TF2A)	Low	No	No	No	No	Yes	No	No	-9,40
Camelliano side	Low	No	Yes	No	No	No	No	No	-12,15

5. Thảo luận

Nghiên cứu đã thành công xây dựng mô hình học máy có khả năng sàng lọc và dự đoán các hợp chất trong chi Chè (*Camellia*) có tiềm năng kích hoạt biểu hiện Nrf2 thông qua ức chế phức hợp Keap1-Nrf2. Kết quả thu được 33 chất được xác định là "có hoạt tính", với mô hình tốt nhất là SVM và phương pháp mã hóa Morgan 3, sử dụng decoy và ngưỡng phân loại IC_{50} = 10 μM. Nghiên cứu đồng thời đã mô phỏng tương tác và khả năng liên kết bằng phương pháp docking phân tử để chỉ ra bốn hợp chất có tiềm năng hơn cả với năng lượng liên kết thấp nhất lần lượt là camellianoside (-10,4 kcal/mol) có trong cây *C. japonica*, theaflavin-3-gallate (-9,9 kcal/mol) và theaflavin-3'-gallate (-9,8 kcal/mol) trong cây *C. sinensis*, và camellianoside (-9,7 kcal/mol) trong cây *C. oleifera*. Kết quả docking phân tử của bốn hợp chất tiềm năng nhất cho thấy nhiều vị trí tương tác giữa amino acid của protein Keap1 và phối tử đã được ghi nhận trong các nghiên cứu trước như Tyr334, Ser363, Arg380, Arg415, Ser555 là các amino acid được xác định là đóng vai trò quan trọng trong việc hình thành phức hợp Keap1 – Nrf2. Đồng thời, kết quả cũng ghi nhận một vài vị trí amino acid mới như Val512, Ala 510, Val465.

Bốn hợp chất tiềm năng có điểm chung là đều hình thành liên kết với Arg380, Arg415 và Ser555. Các vị trí này cũng đã được báo cáo trong các nghiên cứu trước đó là vị trí liên kết của Keap1 tạo phức hợp Keap1-Nrf2 (PDB ID: 2FLU) [29, 30]. Cụ thể, các báo cáo trước đây đã chỉ ra rằng miền Neh2 của Nrf2 liên kết với Keap1 ở các vị trí amino acid Tyr334, Ser363, Arg380, Asn382, Arg415, Arg483, Ser508, Ser535 và Ser602 [30]. Thông thường, 2 phân tử Keap1 sẽ dime hóa với nhau thông qua Cul3 và mỗi Keap1 sẽ liên kết với 2 mô-típ DLG và ETGE của Nrf2. DLG có dạng latch (tạm dịch: bản lề) nên sẽ dễ bị đứt liên kết với miền Kelch khi có tác nhân ái điện tử (K_a = 0,1 × 10^7 M^{-1}) [32]. Đối với ETGE có cấu trúc phiến β đặc trưng dạng hinge (tạm dịch: chốt) nên chúng sẽ có ái lực liên kết với Keap1 cao hơn (K_a = 20 × 10^7 M^{-1}) [33]. Vì thế theo lý thuyết các hợp chất bắt chước vùng ETGE của Nrf2 sẽ có hiệu quả kích hoạt cơ chế chống oxy hóa nội sinh tốt hơn [32, 34]. Hai hợp chất thuộc họ theaflavin trong nghiên cứu này cho thấy sự tạo thành liên kết với 6 trên 8 vị trí amino acid đặc trưng của miền Neh2, điều này là phù hợp với nghiên cứu của Xiaodan Han và cộng sự [35] đã chỉ ra rằng, theaflavin có khả năng kích hoạt con đường tín hiệu Nrf2. Tuy nhiên, những vị trí liên kết với Keap1 của 2 hợp chất camellianoside và camelliaside B trong nghiên cứu này không có nhiều sự tương đồng với với các vị trí amino acid liên kết của miền Neh2. Tuy nhiên, cần lưu ý rằng, một số nghiên cứu trước đây ghi nhận hợp chất có hoạt tính trên *in vitro* nhưng không cho thấy sự bắt chước vị trí liên kết của Nrf2 trên Keap1 trong mô hình *in silico* quả ghi nhận các các hợp chất tuy không bắt chước các vị trí amino acid liên kết của Nrf2 trên Keap1 nhưng vẫn có khả năng điều hòa chức năng phiên mã Nrf2 [36]. Mô hình khác của Tianzhu Guan và cộng sự đã chỉ ra rằng các chất có điểm số docking thấp nhất đều không bắt chước vị trí liên kết của Nrf2, nhưng kết quả ghi nhận trên thí nghiệm với tế bào HEK293T đã chuyển plasmid của pARE-Luc và pRL-SV40 cho thấy các hợp chất đó lại có khả năng kích hoạt biểu hiện Nrf2 khá tốt (kích hoạt được 50-60% ở các nồng độ 10, 20, và 40 μM) [37].

Bốn hợp chất nghiên cứu sàng lọc được thuộc 2 nhóm chất là flavonoid và theaflavin với camellianoside và camelliaside B thuộc nhóm flavonoid, theaflavin-3-gallate và theaflavin-3'-gallate thuộc nhóm theaflavin. Các nghiên cứu trước đây cho thấy các chất có hoạt tính chống oxy qua con đường tín hiệu của Nrf2 đa phần thuộc nhóm terpenoid, flavonoid, saponin và chalcon [38-41]. Trên chi Chè (*Camellia*), một số nghiên cứu về hợp chất tiềm năng ức chế phức hợp Keap1–Nrf2 ghi nhận tiềm năng của nhóm hợp chất Catechin trong loài *Camellia sinensis* [42], nhóm Triterpenoids trong loài *Camellia oleifera* [41], và nhóm triterpenoid saponins trong loài *Camellia japonica* [43]. Tiếp nối các nghiên cứu đã thực hiện, nhóm nghiên cứu đã ghi nhận tiềm năng chống oxy hóa của loài *Camellia sinensis* nhưng với nhóm hợp chất mới là theaflavin; đồng thời cũng ghi nhận 2 hợp chất flavonoid mới thuộc loài *Camellia oleifera* (camelliaside B) và *Camellia japonica* (camellianoside). Ngoài ra, cả 4 hợp chất tiềm năng cũng được ghi nhận có cấu trúc carbon carbonyl α,β bất bão hòa

tương đồng với các hợp chất được ghi nhận là có khả năng ức chế phức hợp Keap1 – Nrf2 như dimethyl fumarate (thành phần thuốc Tecfidera điều trị đa xơ cứng) [44], và omaveloxolon (thành phần thuốc Skyclarys điều trị Friedreich's ataxia) [45]. Cấu trúc của bốn hợp chất cũng phù hợp với kết quả nghiên cứu trước đây của Melford Chuka Egbujor và cộng sự ghi nhận về khả năng phản ứng tốt của các hợp chất có cấu trúc carbon carbonyl α,β bất bão hòa với amino acid cystein của Keap1, giúp ức chế phúc hợp Keap1 – Nrf2 [46]. Bên cạnh đó, Kết quả đánh giá điểm số ADME của bốn hợp chất đều thỏa mãn được giá trị Log P ≤ 5, cho thấy khả năng hòa tan trong dịch bào và xuyên qua màng tế bào. Ngoài ra, giá trị độ tan trong nước của hợp chất (Log S) được tính toán theo 2 phương pháp là ESOL và Ali. Theo SwissADME, kết quả Log S sẽ được phân loại như sau: Không tan (Log S < -10), tan kém (-10 < Log S < -6), tan ở mức vừa phải (-6 < Log S < -4), tan trong nước (-4 < Log S < -2), tan tốt trong nước (-2 < Log S < 0) và tan rất tốt trong nước (0 < Log S). Trong cả hai phương pháp tính, camellianoside và camelliaside B đều được xác định là ở mức "tan trong nước". Hai hợp chất theaflavin-3-gallate và theaflavin-3'-gallate đều cho thấy kết quả hầu hết ở mức "tan kém". Các chất này đều có khả năng hấp thu qua đường tiêu hóa thấp, không có khả năng đi qua hàng rào máu não và tính thấm qua da không quá cao. Hầu hết các chất đều không là chất nền P-gp ngoại trừ camellianoside, nghĩa là nó có khả năng bị vận chuyển ra ngoài tế bào bởi P-glycoprotein, làm giảm hiệu quả của thuốc. Theaflavin-3'-gallate (TF2B) có giá trị K_p cao hơn hẳn các chất khác, cho thấy khả năng thẩm thấu qua da tốt nhưng lại có thể bị chuyển hóa bởi CYP2C19 làm giảm tác dụng đối với cơ thể; còn theaflavin-3-gallate (TF2A) có nguy cơ bị chuyển hóa bởi CYP2C9. Các chất này có thể được xem xét cho các ứng dụng ngoài da hoặc khi không cần hấp thụ qua đường tiêu hóa hay tác động lên hệ thần kinh trung ương.

Con đường tín hiệu Nrf2 là một cơ chế tiềm năng giúp tăng khả năng bảo vệ tế bào khỏi stress oxy hóa và các tác nhân ái điện tử của cơ thể, nhưng chỉ khi Nrf2 được kích hoạt tạm thời [8]. Một số nghiên cứu trước đây đã ghi nhận hàm lượng Nrf2 cao ở tế bào ung thư khi Nrf2 bị hoạt hóa quá mức bình thường. Cơ chế này xảy ra khi Keap1 hoặc Nrf2 xảy ra đột biến thể soma, làm cho Keap1 không thể bắt giữ Nrf2 và Nrf2 có khả năng kháng ubiquitin hóa tốt hơn [47]. Một nghiên cứu vào năm 2017 đã chỉ ra rằng những hợp chất có khả năng ức chế phức hợp Keap1-Nrf2, đặc biệt là nhắm vào các miền liên kết với Nrf2 của Keap1 có khả năng ít gây đột biến hơn vì cơ chế gây biến đổi cấu trúc cysteine của Keap1 vẫn còn nhiều lo ngại là có tiềm năng gây tác dụng phụ suy giảm glutathione hoặc oxy hóa-khử các thành phần khác trong tế bào [48].

Nhóm nghiên cứu đã sử dụng học máy để thực hiện sàng lọc một lượng lớn các chất tự nhiên thay cho quy trình trích xuất và phân lập chất; docking phân tử trên phức hợp Keap1-Nrf2 để dự đoán khả năng liên kết và điểm số ADME để dự đoán các tính chất vật lý thay cho quy trình nuôi cấy tế bào và thí nghiệm khảo sát biểu hiện của Nrf2 trên tế bào.

6. Kết luận

Nghiên cứu đã chứng minh Chi chè (*Camellia*) là nguồn dược liệu tiềm năng trong việc tìm kiếm các hợp chất ức chế phức hợp Keap1-Nrf2 dựa trên sàng lọc *in silico* với camellianoside có trong cây *C. japonica*, theaflavin-3-gallate và theaflavin-3'-gallate trong cây *C. sinensis* và camellianoside (-9,7 kcal/mol) trong cây *C. oleifera*. Trong thời gian tới, nghiên cứu tiến hành phân lập bốn hợp chất tiềm năng trên, tiến hành đánh giá khả năng tăng biểu hiện Nrf2 trên *in vitro*. Ngoài ra, việc mở rộng nghiên cứu thêm các tác dụng dược lý của chi Chè *Camellia* giúp nâng tầm giá trị của cây chè Việt Nam, góp phần ngăn ngừa một số vấn đề sức khỏe trong cộng đồng do stress oxy hoá gây ra như lão hóa, tim mạch, ung thư.

TÀI LIỆU THAM KHẢO

[1] N. C. Khanh and N. H. Nam, "Oxidative Stress and Disease", *Journal of Pediatrics*, vol. 15, no. 1, Jun. 2023, doi: https://doi.org/10.52724/tcnk.v15i1.176.

[2] B. Huchzermeyer, E. Menghani, P. Khardia, and A. Shilu, "Metabolic Pathway of Natural Antioxidants, Antioxidant Enzymes and ROS Providence", *Antioxidants*, vol. 11, no. 4, p. 761, Apr. 2022, doi: https://doi.org/10.3390/antiox11040761.

[3] A. T. Dinkova-Kostova *et al.*, "Direct evidence that sulfhydryl groups of Keap1 are the sensors regulating induction of phase 2 enzymes that protect against carcinogens and oxidants", *Proceedings of the National Academy of Sciences*, vol. 99, no. 18, pp. 11908–11913, Aug. 2002, doi: 10.1073/pnas.172398899.

[4] T. Ogura *et al.*, "Keap1 is a forked-stem dimer structure with two large spheres enclosing the intervening, double glycine repeat, and C-terminal domains", *Proceedings of the National Academy of Sciences*, vol. 107, no. 7, pp. 2842–2847, Jan. 2010, doi: 10.1073/pnas.0914036107.

[5] T. Yamamoto *et al.*, "Physiological significance of reactive cysteine residues of KEAP1 in determining NRF2 activity", *Molecular and Cellular Biology*, vol. 28, no. 8, pp. 2758–2770, Feb. 2008, doi: 10.1128/mcb.01704-07.

[6] V. O. Tkachev, E. B. Menshchikova, and N. K. Zenkov, "Mechanism of the Nrf2/Keap1/ARE signaling system", *Biochemistry (Moscow)*, vol. 76, no. 4, pp. 407–422, Apr. 2011, doi: 10.1134/s0006297911040031.

[7] S. M. U. Ahmed, L. Luo, A. Namani, X. J. Wang, and X. Tang, "Nrf2 signaling pathway: Pivotal roles in inflammation", Biochimica et Biophysica Acta (BBA) - *Molecular Basis of Disease*, vol. 1863, no. 2, pp. 585–597, Feb. 2017, doi: https://doi.org/10.1016/j.bbadis.2016.11.005.

[8] F. Chen, M. Xiao, S. Hu, and M. Wang, "Keap1-Nrf2 pathway: a key mechanism in the occurrence and development of cancer", *Frontiers in Oncology*, vol. 14, Apr. 2024, doi: 10.3389/fonc.2024.1381467.

[9] A. Pant, D. Dasgupta, A. Tripathi, and K. Pyaram, "Beyond antioxidation: KEAP1–NRF2 in the development and effector functions of adaptive immune cells", *ImmunoHorizons*, vol. 7, no. 4, pp. 288–298, Apr. 2023, doi: 10.4049/immunohorizons.2200061.

[10] M. Sova and L. Saso, "Design and development of Nrf2 modulators for cancer chemoprevention and therapy: a review", *Drug Design Development and Therapy*, vol. Volume 12, pp. 3181–3197, Sep. 2018, doi: 10.2147/dddt.s172612.

[11] P. Silva, "Tecfidera (dimethyl fumarate) for MS | Uses, side effects, and more", *Multiple Sclerosis News Today*, Dec. 13, 2023. https://multiplesclerosisnewstoday.com/tecfidera-dimethyl-fumarate-multiple-sclerosis [Accessed Sep. 9, 2024].

[12] N. Robledinos-Antón, R. Fernández-Ginés, G. Manda, and A. Cuadrado, "Activators and Inhibitors of NRF2: A Review of Their Potential for Clinical Development", *Oxidative Medicine and Cellular Longevity*, vol. 2019, pp. 1–20, Jul. 2019, doi: https://doi.org/10.1155/2019/9372182.

[13] T. W. Kensler *et al.*, "Keap1-Nrf2 Signaling: A Target for Cancer Prevention by Sulforaphane", *Topics in current chemistry*, vol. 329, pp. 163–177, 2013, doi: https://doi.org/10.1007/128_2012_339.

[14] S. Unni, P. Deshmukh, G. Krishnappa, P. Kommu, and B. Padmanabhan, "Structural insights into the multiple binding modes

of Dimethyl Fumarate (DMF) and its analogs to the Kelch domain of Keap1", *The FEBS journal*, vol. 288, no. 5, pp. 1599–1613, Mar. 2021, doi: https://doi.org/10.1111/febs.15485.

[15] D. D. Zhang and M. Hannink, "Distinct Cysteine Residues in Keap1 Are Required for Keap1-Dependent Ubiquitination of Nrf2 and for Stabilization of Nrf2 by Chemopreventive Agents and Oxidative Stress", *Molecular and Cellular Biology*, vol. 23, no. 22, pp. 8137–8151, Oct. 2003, doi: https://doi.org/10.1128/mcb.23.22.8137-8151.2003.

[16] D. A. Abed, S. Lee, and L. Hu, "Discovery of disubstituted xylylene derivatives as small molecule direct inhibitors of Keap1-Nrf2 protein-protein interaction", *Bioorganic & Medicinal Chemistry*, vol. 28, no. 6, p. 115343, Mar. 2020, doi: https://doi.org/10.1016/j.bmc.2020.115343.

[17] X. Wen, G. Thorne, L. Hu, M. S. Joy, and L. M. Aleksunes, "Activation of NRF2 signaling in HEK293 cells by a First-in-Class direct KEAP1-NRF2 inhibitor", *Journal of Biochemical and Molecular Toxicology*, vol. 29, no. 6, pp. 261–266, Feb. 2015, doi: 10.1002/jbt.21693.

[18] T. Zhao, C. Li, S. Wang, and X. Song, "Green tea (Camellia sinensis): A review of its phytochemistry, pharmacology, and toxicology", *Molecules*, vol. 27, no. 12, p.3909, 2022, doi: 10.3390/molecules27123909

[19] C. Musial, A. Kuban-Jankowska, and M. Gorska-Ponikowska, "Beneficial properties of green tea catechins", *International Journal of Molecular Sciences*, vol. 21, no. 5, p. 1744, Mar. 2020, doi: 10.3390/ijms21051744.

[20] A. M. Teixeira and C. Sousa, "A review on the biological activity of camellia species", *Molecules*, vol. 26, no. 8, p. 2178, Apr. 2021, doi: 10.3390/molecules26082178.

[21] O. Trott and A. J. Olson, "AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading", *Journal of Computational Chemistry*, vol. 31, no. 2, pp. 455–461, Jun. 2009, doi: 10.1002/jcc.21334.

[22] A. Daina, O. Michielin, and V. Zoete, "SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules", *Scientific Reports*, vol. 7, no. 1, Mar. 2017, doi: 10.1038/srep42717.

[23] A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, and G. Pujadas, "Molecular fingerprint similarity search in virtual screening", *Methods*, vol. 71, pp. 58–63, Aug. 2014, doi: 10.1016/j.ymeth.2014.08.005.

[24] H. Kuwahara and X. Gao, "Analysis of the effects of related fingerprints on molecular similarity using an eigenvalue entropy approach", *Journal of Cheminformatics*, vol. 13, no. 1, Mar. 2021, doi: 10.1186/s13321-021-00506-2.

[25] G. Landrum, The RDKit Documentation The RDKit 2019.09.1 documentation. (n.d.), RDKit. September 1, 2019. [Online]. Available: https://www.rdkit.org/docs/index.html [Accessed August 27, 2024].

[26] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning", *Scientific Reports*, vol. 14, no. 1, Mar. 2024, doi: 10.1038/s41598-024-56706-x.

[27] Y. Chushak and R. A. Clewell, "An integrated approach to predict activators of NRF2 - the transcription factor for oxidative stress response", *Artificial Intelligence in the Life Sciences*, vol. 5, p. 100097, Apr. 2024, doi: 10.1016/j.ailsci.2024.100097.

[28] D. Bajusz, A. Rácz, and K. Héberger, "Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?", *Journal of Cheminformatics*, vol. 7, no. 1, May 2015, doi: 10.1186/s13321-015-0069-3.

[29] M. Li *et al.*, "Discovery of Keap1−Nrf2 small−molecule inhibitors from phytochemicals based on molecular docking", *Food and Chemical Toxicology*, vol. 133, p. 110758, Nov. 2019, doi: 10.1016/j.fct.2019.110758.

[30] A. A. Alzain *et al.*, "Modulation of NRF2/KEAP1-Mediated Oxidative Stress for cancer treatment by natural products using Pharmacophore-Based screening, molecular docking, and molecular Dynamics studies", *Molecules*, vol. 28, no. 16, p. 6003, Aug. 2023, doi: 10.3390/molecules28166003.

[31] Hasanain Abdulhameed Odhar, "Molecular docking analysis and dynamics simulation of salbutamol with the monoamine oxidase B (MAO-B) enzyme", *Bioinformation*, vol. 18, no. 3, pp. 304–309, Mar. 2022, doi: https://doi.org/10.6026/97320630018304.

[32] S. Magesh, Y. Chen, and L. Hu, "Small molecule modulators of KEAP1-NRF2-ARE pathway as potential preventive and therapeutic agents", *Medicinal Research Reviews*, vol. 32, no. 4, pp. 687–726, May 2012, doi: 10.1002/med.21257.

[33] Y.-S. Keum and B. Choi, "Molecular and chemical regulation of the KEAP1-NRF2 signaling pathway", *Molecules*, vol. 19, no. 7, pp. 10074–10089, Jul. 2014, doi: 10.3390/molecules190710074.

[34] F. Begnini *et al.*, "Importance of binding site hydration and flexibility revealed when optimizing a macrocyclic inhibitor of the KEAP1–NRF2 Protein–Protein interaction", *Journal of Medicinal Chemistry*, vol. 65, no. 4, pp. 3473–3517, Feb. 2022, doi: 10.1021/acs.jmedchem.1c01975.

[35] X. Han *et al.*, "Theaflavin ameliorates ionizing radiation-induced hematopoietic injury via the NRF2 pathway", *Free Radical Biology and Medicine*, vol. 113, pp. 59–70, Dec. 2017, doi: 10.1016/j.freeradbiomed.2017.09.014.

[36] R. Sahu *et al.*, "LC-MS characterized methanolic extract of zanthoxylum armatum possess anti-breast cancer activity through Nrf2-Keap1 pathway: An in-silico, in-vitro and in-vivo evaluation", *Journal of Ethnopharmacology*, vol. 269, p. 113758, Dec. 2020, doi: 10.1016/j.jep.2020.113758.

[37] T. Guan, C. Bian, and Z. Ma, "In vitro and in silico perspectives on the activation of antioxidant responsive element by citrus-derived flavonoids", *Frontiers in Nutrition*, vol. 10, Aug. 2023, doi: 10.3389/fnut.2023.1257172.

[38] Y.-R. Li *et al.*, "Discovery of natural flavonoids as activators of Nrf2-mediated defense system: Structure-activity relationship and inhibition of intracellular oxidative insults", *Bioorganic & Medicinal Chemistry*, vol. 26, no. 18, pp. 5140–5150, Oct. 2018, doi: 10.1016/j.bmc.2018.09.010.

[39] H. Kumar, I.-S. Kim, S. V. More, B.-W. Kim, and D.-K. Choi, "Natural product-derived pharmacological modulators of Nrf2/ARE pathway for chronic diseases", *Natural Product Reports*, vol. 31, no. 1, pp. 109–139, Nov. 2013, doi: 10.1039/c3np70065h.

[40] H. Zhou, Y. Wang, Q. You, and Z. Jiang, "Recent progress in the development of small molecule Nrf2 activators: a patent review (2017-present)", *Expert Opinion on Therapeutic Patents*, vol. 30, no. 3, pp. 209–225, Feb. 2020, doi: 10.1080/13543776.2020.1715365.

[41] Z. Wu *et al.*, "Discovery of New Triterpenoids Extracted from Camellia oleifera Seed Cake and the Molecular Mechanism Underlying Their Antitumor Activity", *Antioxidants*, vol. 12, no. 1, p. 7, Dec. 2022, doi: 10.3390/antiox12010007.

[42] M. Talebi, M. Talebi, T. Farkhondeh, G. Mishra, S. İlgün, and S. Samarghandian, "New insights into the role of the Nrf2 signaling pathway in green tea catechin applications", *Phytotherapy Research*, vol. 35, no. 6, pp. 3078–3112, Feb. 2021, doi: 10.1002/ptr.7033.

[43] K. Ko, L. D. Wahyudi, Y.-S. Kwon, J.-H. Kim, and H. Yang, "Nuclear Factor Erythroid 2-Related Factor 2 Activating Triterpenoid Saponins from Camellia japonica Roots", *Journal of Natural Products*, vol. 81, no. 11, pp. 2399–2409, Nov. 2018, doi: 10.1021/acs.jnatprod.8b00374.

[44] E. A. Mills, M. A. Ogrodnik, A. Plave, and Y. Mao-Draayer, "Emerging understanding of the mechanism of action for dimethyl fumarate in the treatment of multiple sclerosis", *Frontiers in Neurology*, vol. 9, Jan. 2018, doi: 10.3389/fneur.2018.00005.

[45] D. R. Lynch *et al.*, "Safety, pharmacodynamics, and potential benefit of omaveloxolone in Friedreich ataxia", *Annals of Clinical and Translational Neurology*, vol. 6, no. 1, pp. 15–26, Nov. 2018, doi: 10.1002/acn3.660.

[46] M. C. Egbujor, B. Buttari, E. Profumo, P. Telkoparan-Akillilar, and L. Saso, "An Overview of NRF2-Activating Compounds Bearing α,β-Unsaturated Moiety and Their Antioxidant Effects", *International Journal of Molecular Sciences*, vol. 23, no. 15, p. 8466, Jul. 2022, doi: 10.3390/ijms23158466.

[47] Sandhu, Ivraj Singh, et al. "Sustained NRF2 Activation in Hereditary Leiomyomatosis and Renal Cell Cancer (HLRCC) and in Hereditary Tyrosinemia Type 1 (HT1)", *Biochemical Society Transactions*, vol. 43, no. 4, Aug. 2015, pp. 650–56. https://doi.org/10.1042/bst20150041.

[48] Taguchi, Keiko, and Masayuki Yamamoto. "The KEAP1–NRF2 System in Cancer", *Frontiers in Oncology*, vol. 7, May 2017, https://doi.org/10.3389/fonc.2017.00085.

PHỤ LỤC

Siêu tham số tối ưu của các thuật toán ứng với 4 phương pháp mã hóa, sử dụng 4 tập dữ liệu huấn luyện

PL-1. Kết quả siêu tham số tối ưu của các thuật toán ứng với 4 phương pháp mã hóa, sử dụng tập dữ liệu huấn luyện có giá trị IC_{50} ngưỡng = 10μM và không sử dụng decoy

Phương pháp mã hóa	Thuật toán	Hyperparameter	Giá trị
MACCS fingerprints	Multilayer Perceptron (MLP)	activation	relu
		alpha	0.01
		batch_size	32
		hidden_layer_sizes	64, 32
		learning_rate	constant
		solver	adam
	Random Forest (RF)	max_depth	10
		max_features	sqrt
		min_samples_leaf	1
		min_samples_split	3
		n_estimators	50
	Support Vector Machine (SVM)	C	7
		degree	2
		gamma	scale
		kernel	rbf
	Extreme Gradient Boosting (XGBoost)	booster	gbtree
		learning_rate	0.15
		max_depth	5
		min_child_weight	2
		n_estimators	300
Morgan 2 fingerprints	Multilayer Perceptron (MLP)	activation	tanh
		alpha	0.001
		batch_size	128
		hidden_layer_sizes	16, 8
		learning_rate	constant
		solver	sgd
	Random Forest (RF)	max_depth	30
		max_features	log2
		min_samples_leaf	2
		min_samples_split	2
		n_estimators	50
	Support Vector Machine (SVM)	C	5
		degree	2
		gamma	scale
		kernel	rbf
	Extreme Gradient Boosting (XGBoost)	booster	gbtree
		learning_rate	0.1
		max_depth	10
		min_child_weight	3
		n_estimators	100
Morgan 3 fingerprints	Multilayer Perceptron (MLP)	activation	tanh
		alpha	0.001
		batch_size	128
		hidden_layer_sizes	16, 8
		learning_rate	constant
		solver	sgd
	Random Forest (RF)	max_depth	20
		max_features	log2
		min_samples_leaf	1
		min_samples_split	2
		n_estimators	400
	Support Vector Machine (SVM)	C	3
		degree	2
		gamma	scale
		kernel	rbf
	Extreme Gradient Boosting (XGBoost)	booster	gbtree
		learning_rate	0.15
		max_depth	5
		min_child_weight	3
		n_estimators	100
RDK5 fingerprints	Multilayer Perceptron (MLP)	activation	relu
		alpha	0.001
		batch_size	128
		hidden_layer_sizes	64, 32
		learning_rate	constant
		solver	sgd
	Random Forest (RF)	max_depth	30
		max_features	sqrt
		min_samples_leaf	1
		min_samples_split	3
		n_estimators	50
	Support Vector Machine (SVM)	C	3
		degree	2
		gamma	scale
		kernel	rbf
	Extreme Gradient Boosting (XGBoost)	booster	gbtree
		learning_rate	0.15
		max_depth	2
		min_child_weight	1
		n_estimators	100

PL-2. Kết quả siêu tham số tối ưu của các thuật toán ứng với 4 phương pháp mã hóa, sử dụng tập dữ liệu huấn luyện có giá trị IC_{50} ngưỡng = 5μM và không sử dụng decoy

Phương pháp mã hóa	Thuật toán	Hyperparameter	Giá trị
MACCS fingerprints	Multilayer Perceptron (MLP)	activation	tanh
		alpha	0.1
		batch_size	32
		hidden_layer_sizes	16, 8
		learning_rate	constant
		solver	adam
	Random Forest (RF)	max_depth	30
		max_features	sqrt
		min_samples_leaf	1
		min_samples_split	3
		n_estimators	400
	Support Vector Machine (SVM)	C	10
		degree	2
		gamma	scale
		kernel	rbf
	Extreme Gradient Boosting (XGBoost)	booster	gbtree
		learning_rate	0.1
		max_depth	15
		min_child_weight	1
		n_estimators	100
Morgan 2 fingerprints	Multilayer Perceptron (MLP)	activation	tanh
		alpha	0.001
		batch_size	64
		hidden_layer_sizes	16, 8
		learning_rate	constant
		solver	sgd
	Random Forest (RF)	max_depth	20
		max_features	sqrt
		min_samples_leaf	1
		min_samples_split	3
		n_estimators	100
	Support Vector Machine (SVM)	C	3
		degree	2
		gamma	scale
		kernel	rbf
	Extreme Gradient Boosting (XGBoost)	booster	gbtree
		learning_rate	0.2
		max_depth	5
		min_child_weight	3
		n_estimators	100
Morgan 3 fingerprints	Multilayer Perceptron (MLP)	activation	tanh
		alpha	0.001
		batch_size	128
		hidden_layer_sizes	16, 8
		learning_rate	constant
		solver	sgd
	Random Forest (RF)	max_depth	20
		max_features	log2
		min_samples_leaf	1
		min_samples_split	3
		n_estimators	100
	Support Vector Machine (SVM)	C	3
		degree	2
		gamma	scale
		kernel	rbf
	Extreme Gradient Boosting (XGBoost)	booster	gbtree
		learning_rate	0.2
		max_depth	5
		min_child_weight	3
		n_estimators	100
RDK5 fingerprints	Multilayer Perceptron (MLP)	activation	relu
		alpha	0.1
		batch_size	128
		hidden_layer_sizes	128, 64
		learning_rate	constant
		solver	sgd
	Random Forest (RF)	max_depth	20
		max_features	sqrt
		min_samples_leaf	1
		min_samples_split	3
		n_estimators	50
	Support Vector Machine (SVM)	C	2
		degree	2
		gamma	scale
		kernel	rbf
	Extreme Gradient Boosting (XGBoost)	booster	gbtree
		learning_rate	0.2
		max_depth	5
		min_child_weight	3
		n_estimators	200

PL-3. Kết quả siêu tham số tối ưu của các thuật toán ứng với 4 phương pháp mã hóa, sử dụng tập dữ liệu huấn luyện có giá trị IC_{50} ngưỡng = 10μM và có sử dụng decoy

Phương pháp mã hóa	Thuật toán	Hyperparameter	Giá trị
MACCS fingerprints	Multilayer Perceptron (MLP)	activation	relu
		alpha	0.1
		batch_size	32
		hidden_layer_sizes	16, 8
		learning_rate	constant
		solver	adam
	Random Forest (RF)	max_depth	20
		max_features	sqrt
		min_samples_leaf	2
		min_samples_split	2
		n_estimators	400
	Support Vector Machine (SVM)	C	5
		degree	2
		gamma	scale
		kernel	rbf
	Extreme Gradient Boosting (XGBoost)	booster	gbtree
		learning_rate	0.1
		max_depth	3
		min_child_weight	2
		n_estimators	500
Morgan 2 fingerprints	Multilayer Perceptron (MLP)	activation	relu
		alpha	0.001
		batch_size	128
		hidden_layer_sizes	64, 32
		learning_rate	constant
		solver	sgd
	Random Forest (RF)	max_depth	20
		max_features	log2
		min_samples_leaf	1
		min_samples_split	3
		n_estimators	400
	Support Vector Machine (SVM)	C	2
		degree	2
		gamma	scale
		kernel	rbf
	Extreme Gradient Boosting (XGBoost)	booster	gbtree
		learning_rate	0.1
		max_depth	3
		min_child_weight	1
		n_estimators	100
Morgan 3 fingerprints	Multilayer Perceptron (MLP)	activation	relu
		alpha	0.001
		batch_size	128
		hidden_layer_sizes	64, 32
		learning_rate	constant
		solver	sgd
	Random Forest (RF)	max_depth	30
		max_features	log2
		min_samples_leaf	2
		min_samples_split	2
		n_estimators	400
	Support Vector Machine (SVM)	C	2
		degree	2
		gamma	scale
		kernel	rbf
	Extreme Gradient Boosting (XGBoost)	booster	gbtree
		learning_rate	0.1
		max_depth	15
		min_child_weight	2
		n_estimators	100
RDK5 fingerprints 1024 bits	Multilayer Perceptron (MLP)	activation	relu
		alpha	0.1
		batch_size	128
		hidden_layer_sizes	16,8
		learning_rate	constant
		solver	sgd
	Random Forest (RF)	max_depth	10
		max_features	log2
		min_samples_leaf	1
		min_samples_split	3
		n_estimators	400
	Support Vector Machine (SVM)	C	4
		degree	2
		gamma	scale
		kernel	rbf
	Extreme Gradient Boosting (XGBoost)	booster	gbtree
		learning_rate	0.1
		max_depth	3
		min_child_weight	3
		n_estimators	100

PL-4. Kết quả siêu tham số tối ưu của các thuật toán ứng với 4 phương pháp mã hóa, sử dụng tập dữ liệu huấn luyện có giá trị IC_{50} ngưỡng = 5μM và có sử dụng decoy

Phương pháp mã hóa	Thuật toán	Hyperparameter	Giá trị
MACCS fingerprints	Multilayer Perceptron (MLP)	activation	relu
		alpha	0.001
		batch_size	32
		hidden_layer_sizes	128, 64
		learning_rate	constant
		solver	sgd
	Random Forest (RF)	max_depth	20
		max_features	sqrt
		min_samples_leaf	2
		min_samples_split	2
		n_estimators	100
	Support Vector Machine (SVM)	C	7
		degree	2
		gamma	scale
		kernel	rbf
	Extreme Gradient Boosting (XGBoost)	booster	gbtree
		learning_rate	0.05
		max_depth	5
		min_child_weight	4
		n_estimators	300
Morgan 2 fingerprints	Multilayer Perceptron (MLP)	activation	relu
		alpha	0.01
		batch_size	64
		hidden_layer_sizes	16, 8
		learning_rate	constant
		solver	sgd
	Random Forest (RF)	max_depth	30
		max_features	sqrt
		min_samples_leaf	2
		min_samples_split	2
		n_estimators	50
	Support Vector Machine (SVM)	C	2
		degree	2
		gamma	scale
		kernel	rbf
	Extreme Gradient Boosting (XGBoost)	booster	gbtree
		learning_rate	0.05
		max_depth	5
		min_child_weight	4
		n_estimators	200
Morgan 3 fingerprints	Multilayer Perceptron (MLP)	activation	relu
		alpha	0.1
		batch_size	64
		hidden_layer_sizes	64, 32
		learning_rate	constant
		solver	sgd
	Random Forest (RF)	max_depth	30
		max_features	sqrt
		min_samples_leaf	2
		min_samples_split	2
		n_estimators	100
	Support Vector Machine (SVM)	C	2
		degree	2
		gamma	scale
		kernel	rbf
	Extreme Gradient Boosting (XGBoost)	booster	gbtree
		learning_rate	0.05
		max_depth	5
		min_child_weight	4
		n_estimators	200
RDK5 fingerprints 1024 bits	Multilayer Perceptron (MLP)	activation	relu
		alpha	0.001
		batch_size	128
		hidden_layer_sizes	16, 8
		learning_rate	constant
		solver	sgd
	Random Forest (RF)	max_depth	10
		max_features	log2
		min_samples_leaf	1
		min_samples_split	3
		n_estimators	400
	Support Vector Machine (SVM)	C	2
		degree	2
		gamma	scale
		kernel	rbf
	Extreme Gradient Boosting (XGBoost)	booster	gbtree
		learning_rate	0.1
		max_depth	3
		min_child_weight	4
		n_estimators	100