

INTEGRATING MACHINE LEARNING AND DOCKING SIMULATIONS FOR THE SCREENING AND DESIGN OF NOVEL HDAC2 INHIBITORS AS ANTICANCER AGENTS

Nguyen Ngoc An¹, Dao Quang Tung², Thai Chinh Tam³, Nguyen Thanh Cong¹,
Pham Thi Kim Anh³, Phan Thi Phuong Dung^{3*}, Do Thi Mai Dung³

¹VNU University of Engineering and Technology, ²Stockholm University,

³Hanoi University of Pharmacy

ARTICLE INFO	ABSTRACT
Received: 02/01/2025	Histone deacetylase 2 (HDAC2) is a promising molecular target for anticancer drug design. This study collected and analyzed 2,809 compounds to train machine-learning models for identifying potential HDAC2 inhibitors. Virtual screening of 140 million structures identified 88 compounds with predicted HDAC2 inhibition superior to SAHA, confirmed by the top four models with over 85% accuracy. Docking studies further evaluated these, revealing 60 compounds with ΔG values between -12.3 to -16.5 kcal/mol, surpassing SAHA. Structural analysis led to the design of 16 new compounds with enhanced binding affinity. Two compounds, I and II, were predicted to outperform the original lead, forming additional interactions with the active site while maintaining zinc ion coordination. This research developed a machine learning model and proposed a screening procedure combined with docking to identify new compound structures with good inhibitory potential. This approach is not only suitable for HDAC2 but can be also applied to many biological targets.
Revised: 27/3/2025	
Published: 28/3/2025	
KEYWORDS	
Machine learning	
QSAR	
Drug design	
Docking simulation	
HDAC2 inhibitors	

ỨNG DỤNG HỌC MÁY VÀ MÔ PHỎNG DOCKING TRONG SÀNG LỌC, THIẾT KẾ CÁC CHẤT ỨC CHẾ HDAC2 MỚI HƯỚNG ĐIỀU TRỊ UNG THƯ

Nguyễn Ngọc An¹, Đào Quang Tùng², Thái Chính Tâm³, Nguyễn Thành Công¹,
Phạm Thị Kim Anh³, Phan Thị Phương Dung^{3*}, Đỗ Thị Mai Dung³

¹Trường Đại học Công nghệ - ĐHQG Hà Nội, ²Trường Đại học Stockholm,

³Trường Đại học Dược Hà Nội

THÔNG TIN BÀI BÁO	TÓM TẮT
Ngày nhận bài: 02/01/2025	Histone deacetylase 2 (HDAC2) là một mục tiêu phân tử hấp dẫn trong thiết kế các thuốc chống ung thư. Trong nghiên cứu này, cấu trúc của 2.809 chất cùng kết quả thực nghiệm được dùng để xây dựng các mô hình học máy với mục đích nhận diện các hợp chất có tiềm năng ức chế HDAC2. Kết quả sàng lọc trên 140 triệu cấu trúc xác định được 88 chất có tiềm năng ức chế HDAC2 tốt hơn SAHA (độ chính xác của dự đoán trên 85% với cả 4 mô hình). Phân tích mô phỏng docking, 60/88 chất có tiềm năng tốt hơn chất đối chiếu, ΔG khoảng -12,3 đến -16,5 kcal/mol. Từ cấu trúc của 88 chất, 16 cấu trúc được thiết kế mới để gia tăng khả năng liên kết với trung tâm hoạt động của HDAC2. Áp dụng quy trình sàng lọc ảo gồm 4 mô hình học máy kết hợp docking, kết quả thiết kế được 2 cấu trúc (chất I, II). Phân tích tương tác chỉ ra chất I và II có thể tạo thêm các liên kết mới với trung tâm hoạt động, đồng thời duy trì khả năng tạo phức với ion kẽm. Nghiên cứu đã xây dựng được mô hình học máy và đưa ra quy trình sàng lọc kết hợp docking tìm ra các cấu trúc hợp chất mới có tiềm năng ức chế tốt. Quy trình này không chỉ phù hợp với HDAC2 mà có thể áp dụng cho nhiều mục tiêu sinh học khác.
Ngày hoàn thiện: 27/3/2025	
Ngày đăng: 28/3/2025	
TỪ KHÓA	
Học máy	
QSAR	
Thiết kế thuốc	
Mô phỏng docking	
Chất ức chế HDAC2	

DOI: <https://doi.org/10.34238/tnu-jst.11797>

* Corresponding author. Email: dungptp@hup.edu.vn

1. Giới thiệu

Sự phát triển vượt bậc của sinh học phân tử và sự tiến bộ về sức mạnh tính toán giúp thực hiện các mô phỏng phức tạp trong thời gian tương đối ngắn đã khiến các phương pháp thiết kế thuốc có sự hỗ trợ của máy tính đóng vai trò ngày càng quan trọng hơn trong việc đẩy nhanh quá trình phát triển thuốc mới [1]. Trong đó, việc ứng dụng rộng rãi các thuật toán học máy (Machine Learning - ML) giúp phân loại và sàng lọc dữ liệu tốc độ cao. Phương pháp docking phân tử và mô phỏng động lực học phân tử (Molecular Dynamic Simulations – MD) là những công cụ hữu ích, hỗ trợ quá trình tối ưu hóa dẫn xuất, khám phá liên kết của phối tử với protein. Các phương pháp này góp phần nghiên cứu sâu hơn về mối tương quan cấu trúc – tác dụng của các hợp chất (Structure – Activity Relationship - SAR) và hỗ trợ đắc lực trong việc thiết kế các cấu trúc thuốc mới hiệu quả [2]-[10].

Ung thư luôn là một nhóm bệnh nghiêm trọng với tỷ lệ tử vong cao, thu hút sự quan tâm lớn của các công ty dược phẩm cũng như các nhóm nghiên cứu phát triển thuốc mới. Trong đó, các tế bào ung thư phổi, vú, đại trực tràng, tuyến tiền liệt và dạ dày cho thấy sự biểu hiện mạnh mẽ của nhóm các enzym histon deacetylase (HDAC), đặc biệt là HDAC2 [5]-[7]. Do đó, các chất ức chế chọn lọc HDAC2 được kỳ vọng sẽ tăng hiệu quả điều trị và giảm tác dụng không mong muốn của các chất ức chế HDAC không chọn lọc hiện đang được dùng trong lâm sàng như vorinostat, panobinostat, belinostat, các thuốc này cản trở một số quá trình sinh lý bình thường trong cơ thể được xúc tác bởi HDAC.

Trong nghiên cứu này, chúng tôi kết hợp xây dựng, ứng dụng các mô hình học máy để sàng lọc, dự đoán số lượng lớn các hợp chất chưa có kết quả thực nghiệm về khả năng ức chế enzym histon deacetylase 2. Những chất tiềm năng sẽ là những chất được dự đoán có hoạt tính ức chế với xác suất cao dưới sự đồng thuận của bốn mô hình học máy khác nhau. Từ kết quả sàng lọc ảo, chúng tôi kết hợp với các phương pháp docking phân tử, mô phỏng động lực học phân tử để thiết kế, kiểm tra một số cấu trúc mới có tiềm năng ức chế enzym HDAC2. Nghiên cứu đã xây dựng được mô hình học máy và đưa ra quy trình sàng lọc kết hợp docking tìm ra các cấu trúc hợp chất mới có tiềm năng ức chế tốt. Phương pháp kết hợp này không chỉ phù hợp với HDAC2 mà có thể áp dụng cho nhiều mục tiêu sinh học khác.

2. Phương pháp nghiên cứu

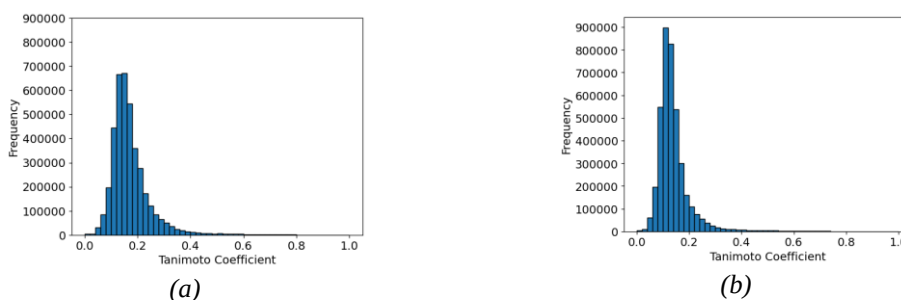
2.1. Thu thập dữ liệu và tiền xử lý dữ liệu

2.1.1. Thu thập dữ liệu

Thu thập tất cả các dữ liệu thử nghiệm IC_{50} trên mục tiêu HDAC2 người từ cơ sở dữ liệu ChEMBL và BindingDB. Tổng cộng 3.493 thử nghiệm sinh học của 2.855 hợp chất liên quan đến HDAC2 được lấy từ cơ sở dữ liệu ChEMBL 33 và 3.795 thử nghiệm sinh học trên HDAC2 của 2.926 hợp chất được tải xuống từ cơ sở dữ liệu Binding DB, thu được bộ dữ liệu thô ban đầu (truy cập vào tháng 8 năm 2023) [11].

2.1.2. Tiền xử lý dữ liệu

Dữ liệu của tất cả các hợp chất sau khi tải về sẽ trải qua quá trình tiền xử lý dữ liệu như sau: (1) Tìm kiếm tất cả những hợp chất có liên quan đến các thử nghiệm sinh học liên quan đến HDAC2 người trên các nền tảng dữ liệu ChEMBL và BindingDB; (2) Loại bỏ các hợp chất bị thiếu giá trị đánh giá khả năng ức chế HDAC2 là IC_{50} và/hoặc giá trị IC_{50} không xác định; (3) Xác định giá trị IC_{50} trung bình của chất đối chiếu dương tính được chọn là vorinostat (SAHA). Giá trị này được sử dụng làm ngưỡng phân loại, các chất có IC_{50} nhỏ hơn hoặc bằng ngưỡng phân loại được dán nhãn có hoạt tính (897 chất), các chất còn lại dán nhãn không có hoạt tính (1.912 chất); (4) Loại những hợp chất không thể mã hoá được bằng thư viện RDKit [8] để chuyển thành fingerprint ECFP4, ECFP6 do những chất không thể mã hoá sẽ không thể sử dụng để xây dựng mô hình. Quá trình tiền xử lý dữ liệu thu được tập dữ liệu sau khi tiền xử lý gồm 2.809 hợp chất.



Hình 1. Giá trị Tc của tập dữ liệu dùng để xây dựng mô hình:
(a) Tc tính theo ECFP4 (1024 bit) và (b) Tc tính theo ECFP6 (1024 bit)

Kết quả đánh giá tính đa dạng về cấu trúc của tập dữ liệu gồm 2.809 cấu trúc thông qua việc tính toán hệ số Tanimoto (Tc) theo cặp được trình bày trong Hình 1a và Hình 1b. Theo đó, phần lớn giá trị Tc của các cặp chất khi được mã hoá bởi cả hai thuật toán ECFP4 và ECFP6 đều có giá trị nhỏ hơn 0,4. Điều này thể hiện các cấu trúc trong tập dữ liệu thu thập được là tương đối đa dạng nhau có thể ứng dụng để dự đoán trong nhiều trường hợp và hỗ trợ thiết kế cấu trúc mới.

2.1.3. Chia tập dữ liệu huấn luyện, đánh giá và kiểm tra

Dữ liệu sau khi tiền xử lý được chia ngẫu nhiên thành tập huấn luyện, tập đánh giá và tập kiểm tra sử dụng thuật toán lấy mẫu phân tầng (stratified sampling) và hàm train_test_split từ thư viện sklearn. Ba tập dữ liệu huấn luyện, đánh giá và kiểm tra có tỷ lệ phần trăm so với toàn bộ dữ liệu lần lượt là 70:15:15, trong đó đảm bảo các tập có tỷ lệ nhãn có hoạt tính (CHT)-không có hoạt tính (KHT) tương tự nhau như trong Bảng 1.

Bảng 1. Thông tin các tập dữ liệu huấn luyện, đánh giá và kiểm tra

Tập dữ liệu	Tập huấn luyện		Tập đánh giá		Tập kiểm tra	
	CHT	KHT	CHT	KHT	CHT	KHT
Số lượng hợp chất	634	1332	126	296	137	284
Tỷ lệ phần trăm (%)	~32	~68	~30	~70	~32	~68

2.2. Huấn luyện và đánh giá mô hình

Bốn thuật toán được sử dụng để xây dựng mô hình phân loại là K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF) và Extreme Gradient Boosting (XGBoost) [9]; các cấu trúc được tham số hóa bằng thư viện rdkit; các mô hình được xây dựng sử dụng thư viện scikit-learn [10] và thư viện xgboost. Các mô hình sẽ được huấn luyện nhiều lần và tối ưu siêu tham số bằng phương pháp grid search.

Nghiên cứu xây dựng tổng cộng 8 mô hình học máy, dựa trên 4 loại thuật toán và 2 loại vận tải cấu trúc gồm ECFP4 và ECFP6. Các mô hình sau khi xây dựng được đánh giá bằng các chỉ số gồm: độ chính xác (accuracy), điểm F_1 (Macro F_1 -Score), MCC (hệ số tương quan Matthews), AUC (diện tích dưới đường cong ROC) và độ chính xác khi xác thực chéo 10 lần (được tính trên tập kết hợp giữa tập huấn luyện và tập đánh giá). Sau đó, 4 mô hình có kết quả đánh giá tốt nhất được lựa chọn để ứng dụng cho quá trình sàng lọc ảo. Đồng thời với quá trình xây dựng mô hình, nghiên cứu cũng tiến hành tính toán miền ứng dụng của mô hình (application domain).

2.3. Sàng lọc ảo

2.3.1. Tính toán miền ứng dụng của mô hình

Miền ứng dụng của mô hình được tính toán dựa trên cấu trúc các chất dùng xây dựng mô hình học máy, nhằm đảm bảo sự tin tưởng của kết quả dự đoán. Vì vậy, trước khi 1 cấu trúc được sàng lọc bởi 4 mô hình học máy, sẽ được xác định trước về sự phù hợp với miền ứng dụng của mô hình thông qua 2 tiêu chí như sau: (1) Tính toán trong lượng phân tử, khối lượng phân tử cần nằm

trong khoảng 200 - 700 dalton; (2) xác định các thông số dấu vân tay ECFP4 – 1024 bits, ECFP6 – 1024 bits; tính giá trị Tc trung bình của hợp chất với 5 chất gần nhất trong tập huấn luyện, giá trị Tc trung bình cần $\geq 0,5$.

Dữ liệu sàng lọc thô ban đầu thu thập bao gồm 140 triệu hợp chất từ cơ sở dữ liệu PubChem [12], sau khi tính toán sự phù hợp với miền ứng dụng của mô hình thu được tập dữ liệu sàng lọc ảo bao gồm 61.901 cấu trúc.

2.3.2. Tiến hành sàng lọc

Nghiên cứu sử dụng đồng thời 4 mô hình tốt nhất để tìm kiếm các hợp chất tiềm năng ức chế HDAC2 từ tập dữ liệu sàng lọc ảo bằng cách dự đoán hoạt tính của từng cấu trúc. Các chất được dự đoán có hoạt tính với xác suất $\geq 85\%$ bởi cả 4 mô hình sẽ được đưa vào để kiểm chứng tiếp bằng docking phân tử.

2.4. Docking phân tử

2.4.1. Đánh giá quy trình docking phân tử

Trước hết, quy trình docking được đánh giá bằng cách docking lại (redocking) phối tử tinh thể có trong các cấu trúc phức hợp protein HDAC2. Độ sai lệch giữa cấu hình của phối tử sau docking lại và phối tử tinh thể được tính toán bằng chỉ số RMSD. Chỉ số tương quan Pearson's R được sử dụng để tính toán mối tương quan giữa năng lượng liên kết tự do thực nghiệm và docking. Năng lượng liên kết thực nghiệm được tính theo công thức (5) [13]:

$$\Delta G = R * T * \ln(IC50) \left(\frac{kcal}{mol} \right) \quad (5)$$

(trong đó $R = 0.001987 \frac{kcal}{mol.K}$, $T = 310K$)

Kết quả cho thấy 10 phức hợp khi tiến hành docking đều có thể tái lập lại trạng thái liên kết như được xác định trong thực nghiệm, RMSD trung bình là 1,05 Å. Bên cạnh đó, hệ số tương quan giữa năng lượng liên kết tính toán từ thực nghiệm được dự đoán là $R \approx 0,84$, cho thấy mối tương quan chặt chẽ. Như vậy, phương pháp docking được sử dụng với mục tiêu phân tử HDAC2 là đáng tin cậy, có thể ứng dụng để tiến hành sàng lọc các hợp chất tiềm năng.

2.4.2. Mô phỏng docking

Cấu trúc tinh thể protein HDAC2 (PDB ID: 4LY1 [14]) được sử dụng để docking phân tử. Các bước tiến hành docking bao gồm: (1) Xử lý protein (xóa các phân tử nước, phân tử acid 2-[N-cyclohexylamino]ethan sulfonic, phân tử tetraethylen glycol và các ion calci, ion natri; thêm hydro; thêm điện tích theo phương pháp Kollman) sử dụng Discovery Studio và Autodock Tools; (2) Chuẩn bị phối tử (xây dựng cấu trúc 3D; thêm hydro; thêm điện tích theo phương pháp Gasteiger) từ file cấu trúc và tối ưu hoá bằng phần mềm Avogadro và xử lý với Autodock Tools; (3) Docking phân tử sử dụng phần mềm mvina [15] với trung tâm (19,402; 18,237; -2,716), kích thước hộp 28 x 28 x 28 Å; (4) Đánh giá kết quả docking bằng PyMOL [16]. Những hợp chất có năng lượng liên kết nhỏ hơn chất đối chiếu SAHA (-12,2 kcal/mol) sẽ được đánh giá là có hoạt tính sinh học.

2.4.3. Thiết kế cấu trúc mới

Từ kết quả phân tích mô phỏng docking, đánh giá và lựa chọn khung cấu trúc có kết quả dự đoán tốt nhất. Tiến hành phân tích tương tác giữa chất ức chế và HDAC2. Từ đó, thiết kế thay đổi cấu trúc hợp lý để tạo thêm các tương tác tốt với trung tâm hoạt động của HDAC2.

Tiến hành dự đoán khả năng ức chế HDAC2 của các hợp chất mới bằng các mô hình học máy đã lựa chọn. Thực hiện mô phỏng docking các hợp chất mới và so sánh với chất đối chiếu (SAHA) để xác định tiềm năng của các cấu trúc mới.

3. Kết quả và bàn luận

3.1. Huấn luyện và đánh giá mô hình

Sau huấn luyện và tối ưu tham số, kết quả đánh giá các mô hình trên tập kiểm tra được trình bày trong Bảng 2.

Bảng 2. Kết quả đánh giá các mô hình với tập dữ liệu kiểm tra

Mô hình	Vân tay cấu trúc	Độ chính xác (%)	MCC (%)	Điểm F1 (%)	AUC (%)
KNN	ECFP4	84,09	63,31	74,91	88,60
RF	ECFP4	81,47	56,72	62,50	83,82
SVM	ECFP4	84,56	63,68	72,57	88,09
XgBoost	ECFP4	84,32	63,22	71,30	87,54
KNN	ECFP6	83,37	61,39	73,28	87,21
RF	ECFP6	81,47	56,72	62,50	86,29
SVM	ECFP6	83,37	60,71	70,34	88,24
XgBoost	ECFP6	83,61	61,42	69,87	87,44

Bảng 2 cho thấy tất cả các mô hình đều có diện tích dưới đường ROC lớn hơn 83%. Điều này thể hiện các mô hình có khả năng phân loại khá tốt. Trong đó, 04 mô hình KNN-ECFP4, SVM-ECFP4, XGBoost-ECFP4, KNN-ECFP6 có độ chính xác, F1-Score, MCC, AUC tốt hơn các mô hình còn lại và được lựa chọn để thực hiện sàng lọc ảo. Các bộ mô hình được tối ưu để không bị quá khớp trên tập huấn luyện và hoạt động ổn định trên tập đánh giá và tập kiểm tra.

Mỗi thuật toán khi sử dụng đều có nhược điểm nhất định, để giảm thiểu tối đa các trường hợp sai số dương tính giả, chúng tôi lựa chọn 04 mô hình có chỉ số đánh giá tốt hơn trong 8 mô hình đã xây dựng, gồm KNN-ECFP4, SVM-ECFP4, XgBoost-ECFP4, KNN-ECFP6

3.2. Sàng lọc ảo

3.2.1. Tiến hành sàng lọc

Bảng 3. Số lượng hợp chất có hoạt tính được dự đoán bởi từng mô hình

Mô hình	Số hợp chất kiểm thử tìm thấy	Số hợp chất có xác suất dương tính $\geq 85\%$	Số hợp chất được chọn bởi tối đa n mô hình			
			n = 1	n = 2	n = 3	n = 4
KNN-ECFP4	20/20	2124				
SVM-ECFP4	20/20	835				
XgBoost-ECFP4	20/20	235	1176	1513	150	88
KNN-ECFP6	20/20	1810				

Tập dữ liệu sàng lọc gồm 61.901 cấu trúc nằm trong miền ứng dụng được dùng để tìm kiếm các hợp chất tiềm năng ức chế HDAC2 bằng cách vận dụng 4 mô hình tốt nhất được lựa chọn ở trên, tính toán khả năng có hoạt tính của từng cấu trúc.

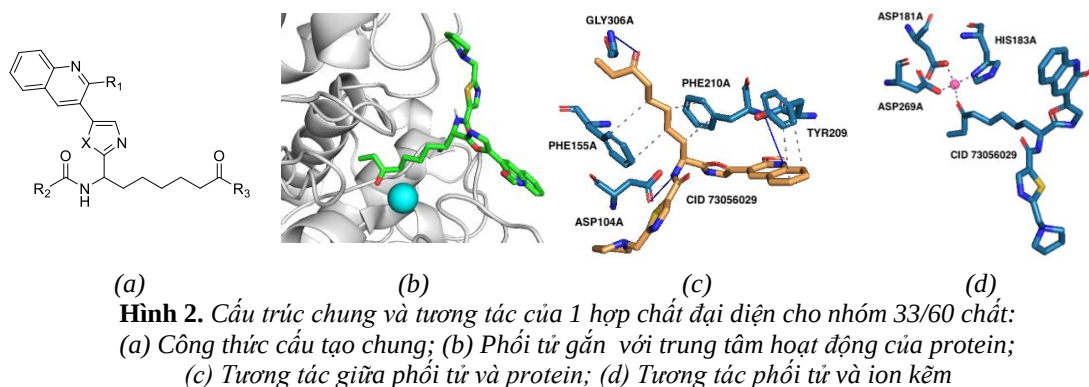
Trong quá trình sàng lọc, lấy ngẫu nhiên từ tập dữ liệu huấn luyện 20 hợp chất đều được gán nhãn là hoạt động với ít nhất 1 mô hình để làm các hợp chất kiểm thử. Kết quả sàng lọc được tổng hợp trong Bảng 3. Kết quả sàng lọc cho thấy 4 mô hình đều phát hiện được tất cả các hợp chất kiểm thử, điều đó chứng tỏ các mô hình có khả năng tìm kiếm các chất hoạt động. Với mong muốn lựa chọn các cấu trúc có khả năng có hoạt tính thực sự, nghiên cứu sẽ ưu tiên chọn nhóm các cấu trúc thỏa mãn được nhiều mô hình nhất. Nhóm các cấu trúc được nghiên cứu tiếp là 88 cấu trúc được lựa chọn bởi cả 4 mô hình, tạo thành tập dữ liệu docking.

3.2.2. Docking các cấu trúc trong tập dữ liệu

Cấu trúc tinh thể protein đích được lựa chọn để docking là 4LY1. 4LX1 là cấu trúc được chụp bằng phương pháp nhiễu xạ tia X với độ phân giải 1.57 Å, có chất đồng kết tinh là 4-(acetylamino)-N-[2-amino-5-(thiophen-2-yl)phenyl]benzamid là một chất ức chế HDAC2. Trong

một nghiên cứu trước đó, Silva Urias, B. cùng cộng sự cũng đã sử dụng 4LY1 để đánh giá sự liên kết của resveratrol với HDAC2, khẳng định tính phù hợp của cấu trúc này cho phân tích docking [17]. Tiến hành docking phân tử với 88 cấu trúc trong tập dữ liệu docking nhằm kiểm tra khả năng liên kết với protein đích (HDAC2).

Kết quả docking thu được 60/88 hợp chất, năng lượng liên kết tự do được dự đoán trong khoảng -12,3 đến -16,5 kcal/mol, khả năng liên kết với HDAC2 tốt hơn SAHA ($\Delta G = -12,2$ kcal/mol). Trong 60 hợp chất, có 33 cấu trúc (55%) đều có chung khung cấu trúc được biểu diễn ở Hình 2. Phân tích tương tác cụ thể của 33 hợp chất này cho thấy cả 33 cấu trúc đều tạo được những liên kết quan trọng với các axit amin Phe155, Phe210, Gly306, là những axit amin đã được chứng minh nằm trong trung tâm hoạt động của HDAC2 [18] và tạo được liên kết với ion kẽm. Hình 2 minh họa cho tương tác cụ thể của một chất đại diện (CID 73056029).



Với các kết quả phân tích trên, có thể thấy những hợp chất có khung cấu trúc chung này có tiềm năng ức chế HDAC2 tốt. Vì vậy, nghiên cứu thực hiện thiết kế các hợp chất mới dựa trên khung cấu trúc này.

3.2.3. Thiết kế cấu trúc mới

Cấu trúc trung tâm hoạt động của HDAC2 có một túi chân nhỏ, bên dưới ion kẽm, khác biệt với hầu hết các HDAC khác. Vì vậy, HDAC2 có thể tương tác tốt với nhiều loại nhóm gắn kẽm khác nhau. Hầu hết nhóm gắn kẽm của 33/60 chất tiềm năng cao đều là các cấu trúc nhỏ. Vì vậy, nhằm tăng khả năng liên kết của phối tử với túi chân, từ đó giúp các phối tử giữ ổn định hơn tại trung tâm hoạt động của protein, đồng thời tăng khả năng gắn chọn lọc của phối tử vào HDAC2, nghiên cứu tiến hành thay đổi cấu trúc phần gắn kẽm. Vùng ZBG của các dẫn chất mới được gắn thêm một nhóm thế chứa dị vòng thơm imidazol với mục tiêu không chỉ duy trì khả năng liên kết với ion kẽm mà còn tăng liên kết với các axit amin ở túi chân [18], [19].

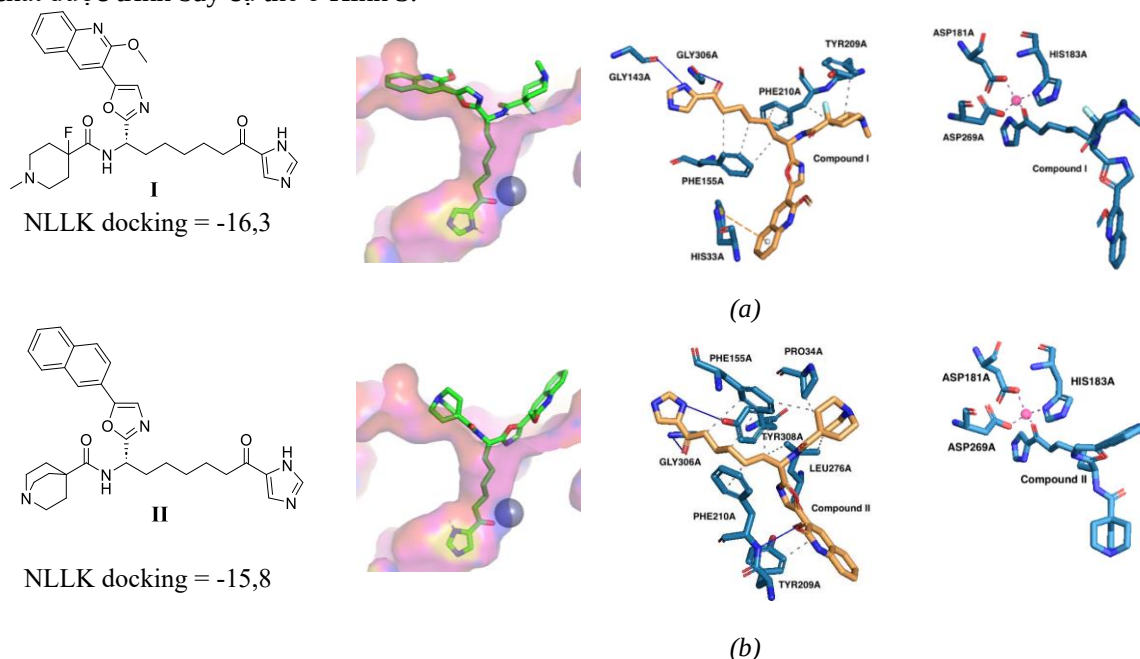
Với mỗi khung cấu trúc, lần lượt thay thế R₃ bằng một nhóm imidazol, sau đó loại bỏ các cấu trúc trùng lặp hoặc các cấu trúc đã từng được công bố, thu được 16 cấu trúc mới chưa từng được thiết kế. 16 cấu trúc mới được sàng lọc bằng 4 mô hình học máy đã lựa chọn trước đó. Các cấu trúc có kết quả tính toán khả năng hoạt động $\geq 85\%$ sẽ được coi là “CHT”.

Kết quả dự đoán bằng học máy cho thấy 16 hợp chất với R₃ là imidazol đều có khả năng ức chế HDAC2 với tiềm năng $\geq 85\%$ với ít nhất 1 mô hình. Trong đó, 2 chất được dự đoán có hoạt tính bởi cả bốn mô hình, 7 chất được dự đoán bởi hai mô hình, và 8 chất được dự đoán bởi một mô hình và không có chất nào được dự đoán là không có hoạt tính.

Tiến hành docking phân tử với 16 cấu trúc thiết kế mới, cũng đưa đến xu hướng đồng thuận với kết quả dự đoán bởi học máy, các cấu trúc có R₃ là imidazol đều có ái lực liên kết tốt, tạo được liên kết với trung tâm hoạt động của HDAC2. Trung tâm hoạt động của HDAC2 bao gồm một kênh enzym chứa ion kẽm (Zn^{2+}), ion này được bao quanh bởi các axit amin quan trọng như His141, His143 và Asp181, giúp duy trì sự ổn định của ion kẽm và hỗ trợ cơ chế xúc tác. Ngoài ra, các axit amin Phe155, Phe210, Leu276, và Gly306 cũng tạo nên vùng tương tác chính giúp

phối tử liên kết bền vững với enzym [18]. Các hợp chất mang nhóm imidazol không chỉ có tương tác với các axit amin quan trọng trong trung tâm hoạt động của HDAC2 mà còn có tương tác với ion kẽm, cho thấy sự phù hợp về khung cấu trúc tổng thể của nhóm các chất thiết kế với HDAC2. Khi quan sát tương tác giữa các tư thể phối tử với protein, có thể thấy vòng imidazol nằm sâu trong phần túi chân, phù hợp với mục tiêu ban đầu.

Kết hợp kết quả dự đoán bởi các mô hình học máy và mô phỏng docking, công thức thiết kế có tiềm năng ức chế HDAC2 tốt nhất được xác định là cấu trúc **I** và **II**, công thức cấu tạo của 2 chất được trình bày cụ thể ở Hình 3.



Hình 3. Công thức cấu tạo và tương tác của cấu trúc I và II với HDAC2: (a) Cấu trúc I; (b) Cấu trúc II.
 Chú thích: tương tác màu xám nét đứt là tương tác kỵ nước, màu xanh dương nét liền là liên kết hydro; màu xanh nhạt nét đứt là liên kết π - π ; điểm màu hồng đại diện cho ion Zn

So sánh các tương tác của cấu trúc mới được thiết kế là I và II với các chất được sàng lọc ra từ 140 triệu chất, khung imidazol là đặc điểm mới nổi bật. Khung này nằm phù hợp với túi chân của HDAC2, tạo thêm tương tác với khu vực túi chân, giúp các chất tạo ra tính chọn lọc với HDAC2 cao hơn. Cụ thể, imidazol trong cấu trúc của chất I thể hiện khả năng tạo liên kết hydro với Gly143; imidazol trong cấu trúc của chất II thể hiện khả năng tạo liên kết hydro với Tyr308 ở phần túi chân. Những liên kết này sẽ đóng góp một phần trong việc tăng ái lực của chất với HDAC2. Bên cạnh đó, cấu trúc trung tâm hoạt động có phần túi chân là đặc điểm khác biệt rất lớn của HDAC2 so với các HDAC khác, thường không có phần mở rộng này. Do đó, cấu trúc I, II có khả năng ức chế chọn lọc HDAC2, ít khả năng tương tác với các HDAC khác. HDAC2 hoạt động quá mức đã được chứng minh là một trong những nguyên nhân dẫn đến phát sinh và phát triển tế bào ung thư, do đó khả năng ức chế HDAC2 sẽ mang đến cho 2 chất I, và II tác dụng kháng tế bào ung thư tốt.

4. Kết luận

Trong nghiên cứu này, chúng tôi đã xây dựng quy trình sàng lọc, thiết kế các chất ức chế HDAC2 mới kết hợp mô hình học máy và mô phỏng docking. Trong đó, nghiên cứu đã xây dựng và lựa chọn 4 mô hình tối ưu để sàng lọc ảo bao gồm RF - ECFP4, RF - ECFP6, XGBoost - ECFP4 và XGBoost - ECFP6 với kết quả đánh giá trên tập dữ liệu kiểm tra là: độ chính xác > 0,80; điểm F1 > 0,63 và MCC > 0,56. Các mô hình được ứng dụng để sàng lọc 61.901 hợp chất

thỏa mãn miền ứng dụng của mô hình, thu được 2.927 chất được dự đoán là hoạt động với tiềm năng $\geq 85\%$ bởi ít nhất 1 mô hình. Cấu trúc của 88 chất được đề xuất bởi cả 4 mô hình có tiềm năng ức chế HDAC2 đã được mô phỏng docking. Nghiên cứu đã thiết kế mới được 16 hợp chất mới; trong đó 2 cấu trúc I, II được xác định có tiềm năng tốt nhất thông qua sàng lọc ảo và phân tích docking. Hai cấu trúc này hoàn toàn mới, cấu trúc có khả năng tổng hợp bằng các phản ứng hữu cơ cơ bản, có tiềm năng để nghiên cứu đưa vào tổng hợp hóa học và phát triển thành chất ức chế tế bào ung thư dựa trên cơ chế ức chế HDAC2. Bên cạnh đó, quy trình sàng lọc kết hợp học máy và mô phỏng docking có khả năng được tích hợp để phát triển phần mềm sàng lọc nhanh hoạt tính ức chế HDAC của các chất, giúp tiết kiệm thời gian nghiên cứu.

Lời cảm ơn

Nghiên cứu được thực hiện với sự hỗ trợ kinh phí từ đề tài nghiên cứu thường niên mã số ĐTCT.23.08 do Trường Đại học Dược Hà Nội tài trợ.

TÀI LIỆU THAM KHẢO/ REFERENCES

- [1] W. Xu, "Current Status of Computational Approaches for Small Molecule Drug Discovery," *Journal of Medicinal Chemistry*, vol. 67, no. 21, pp. 18633-18636, 2024.
- [2] S. Saikia and M. Bordoloi, "Molecular docking: challenges, advances and its use in drug discovery perspective," *Current Drug Targets*, vol. 20, no. 5, pp. 501-521, 2019.
- [3] X. Liu, D. Shi, S. Zhou, *et al.*, "Molecular dynamics simulations and novel drug discovery," *Expert Opinion on Drug Discovery*, vol. 13 no. 1, pp. 23-37, 2017.
- [4] M. D. Vivo, M. Masetti, G. Bottegoni, *et al.*, "Role of molecular dynamics and related methods in drug discovery," *Journal of Medicinal Chemistry*, vol. 59 no. 9, pp. 4035-4061, 2016.
- [5] J. E. Audia and R. M. Campbell, "Histone Modifications and Cancer," *Cold Spring Harbor Perspectives in Biology*, vol. 8, no. 4, 2016, Art. no. a019521.
- [6] N. Basant, X. Lin, T. E. Reid, *et al.*, "Discovery of a Novel HDAC2 Inhibitor by a Scaffold-Merging Hybrid Query," *Combinatorial Chemistry & High Throughput Screening*, vol. 18, no. 7, pp. 693-700, 2015.
- [7] B. M. Müller, L. Jana, A. Kasajima, *et al.*, "Differential expression of histone deacetylases HDAC1, 2 and 3 in human breast cancer-overexpression of HDAC2 and HDAC3 is associated with clinicopathological indicators of disease progression," *BMC Cancer*, vol. 13, pp. 1-8, 2013.
- [8] Greg Landrum, Paolo Tosco, Brain Kelley, *et al.*, "RDKit: Open-source cheminformatics," September 13, 2006 [Online]. <https://www.rdkit.org/>. [Accessed August 20, 2024].
- [9] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 15, no. 1, p. 73, 2016.
- [10] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825-2830, 2011, doi: 10.5555/1953048.2078195.
- [11] K. G. Micheal, L. Tijing, M. Baitaluk, *et al.*, "BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology," *Nucleic Acids Research*, vol. 44, no. D1, pp. D1045-D1053, 2016, doi: 10.1093/nar/gkv1072.
- [12] S. Kim, J. Chen, T. Cheng, *et al.*, "PubChem 2023 update," *Nucleic Acids Research*, vol. 51, no. D1, pp. D1373-D1380, 2023, doi: 10.1093/nar/gkac956.
- [13] J. Aqvist, C. Medina, and J. E. Samuelsson, "A new method for predicting binding affinity in computer-aided drug design," *Protein Engineering*, vol. 7, no. 3, pp. 385-391, 1994, doi: 10.1093/protein/7.3.385.
- [14] H. M. Berman, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235-242, 2000, doi: 10.1093/nar/28.1.235.
- [15] T. N. H. Pham, T. H. Nguyen, N. M. Tam, *et al.*, "Improving ligand-ranking of AutoDock Vina by changing the empirical parameters," *The Journal of Computational Chemistry*, vol. 43, no. 3, pp. 160-169, 2022, doi: 10.1002/jcc.26779.

-
- [16] M. F. Adasme, K. L. Linnemann, S. N. Bolz, *et al.*, "PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA," *Nucleic Acids Research*, vol. 49, no. W1, pp. W530–W534, 2021, doi: 10.1093/nar/gkab294.
- [17] B. S. Urias, A. R. Pavan, G. R. Albuquerque, I. M. Prokopczyk, T. M. F. Alves, T. R. F. de Melo, G. R. Sartori, J. H. M. da Silva, C. M. Chin, and J. L. D. Santos, "Optimization of resveratrol used as a scaffold to design histone deacetylase (HDAC-1 and HDAC-2) inhibitors," *Pharmaceuticals*, vol. 15, 2022, Art. no. 1260, doi: 10.3390/ph15101260.
- [18] S. Tateing and N. Suree, "Decoding molecular recognition of inhibitors targeting HDAC2 via molecular dynamics simulations and configurational entropy estimation," *PLoS One*, vol. 17, no. 8, 2022, Art. no. e0273265, doi: 10.1371/journal.pone.0273265.
- [19] S. Kandasamy, P. Subramani, K. Srinivasan, *et al.*, "Design and synthesis of imidazole-based zinc binding groups as novel small molecule inhibitors targeting histone deacetylase enzymes in lung cancer," *Journal of Molecular Structure*, vol. 1214, 2020, Art. no. 128177, doi: 10.1016/j.molstruc.2020.128177.