STUDY ON THE EFFECT OF MOLECULAR FINGERPRINTS ON THE PERFORMANCE OF MACHINE LEARNING MODEL PREDICTING HIV INTEGRASE BIOACTIVITY

SUMMARY

HIV-1 has been causing severe pandemics since the 1980s by attacking the host immune system. If HIV-1 is not treated, it can eventually lead to AIDS (acquired immunodeficiency syndrome), where death is inevitable due to opportunistic infections. Therefore, discovering new HIV-1 antiviral drugs is urgent. One of the most prominent approaches in Computer-Aided Drug Discovery is constructing Quantitative structure–activity relationship (QSAR) models, especially with the rising of Artificial Intelligence, including machine learning and deep learning. In applied machine learning to create QSAR models, the quality of dataset is a primary criterion to achieve a well-performed binary predictor. The combination of molecular fingerprints can optimize machine learning models, improve prediction accuracy, save time and computational resources. Regarding this research, 3110 HIV integrase inhibitors were extracted from the ChEMBL library and calculated to seven types of molecular fingerprint datasets. After that, a data mining workflow was performed to create a training set and an external validation set. Several machine learning models were initiated to determine the most impactful molecular fingerprint on the model's performance. After finding the best dataset and suitable algorithm, the following stages optimize and evaluate the model's generalized capability on the external validation set. Eventually, RDKit fingerprints had a strong influence on the model's performance. After tuning hyperparameters, the XGBoost algorithm achieved an external validation precision of 0.91, an F1 score of 0.83, and a recall of 0.814. Other metrics, such as an area under the ROC curve of 0.96 and an accuracy of 0.93, were also promising. This result was highly generalizable and reliable for virtual screening potential HIV-1 integrase inhibitors.

Keywords: QSAR, molecular fingerprint, machine learning, HIV integrase, data-centric, average precision.

ĐẶT VẤN ĐỀ

HIV được biết đến là căn bệnh thế kỷ, số người tử vong được ước tính hơn 36 triệu người tính từ khi bắt đầu xuất hiện vào những năm 1980. Các thuốc ức chế HIV hiện nay như dolutegravir, bictegravir, cabotegravir,… với khả năng ức chế enzym integrase, một loại enzym có khả năng tích hợp thông tin di truyền của virus HIV vào tế bào vật chủ. Chính vì vậy enzym này trở thành một trong các hướng tiếp cận tiềm năng để ức chế sự phát triển của HIV[1].

Trí tuệ nhân tạo đã và đang không ngừng phát triển trong nhiều lĩnh vực suốt thập kỷ qua. Sự xuất hiện mang tính đột phá của ChatGPT đã gióng lên hồi chuông thông báo cho cuộc cách mạng trí tuệ nhân tạo sẽ càng bức phá trong những năm sắp tới. Chính vì vậy, việc ứng dụng trí tuệ nhân tạo vào thiết kế thuốc cũng đang dần trở thành xu thế, để có thể, rút ngắn được thời gian, công sức và tài nguyên nghiên cứu sàng lọc ảo được các hoạt chất tiềm năng. Mô hình mô tả mối quan hệ định lượng giữa cấu trúc và tác dụng (QSAR) là một cách tiếp cận phổ biến trong nghiên cứu *in silico* nhằm xây dựng được mối tương quan giữa đặc điểm của phân tử hoá học (thông số mô tả, dấu vân tay phân tử) với hoạt tính sinh học. Việc ứng dụng các thuật toán mạnh mẽ của học máy vào mô hình QSAR đã đạt được nhiều thành công, nâng cao hiệu năng của mô hình như xây dựng được các mô hình dự đoán kháng ung thư, khả năng đáp ứng của thuốc, độc tính [2],… Trước đây, việc

ứng dụng học máy thường chỉ chú trọng đến việc tối ưu thuật toán nhằm nâng cao hiệu năng của mô hình, đây là hướng tiếp cận "model-centric". Tuy nhiên, với quan điểm mới hiện nay, việc có một bộ dữ liệu chuẩn, mô tả đầy đủ các tính năng giúp nâng cao hiệu năng của mô hình vượt trội hơn so với hướng tiếp cận "model-centric", được gọi là "data-centric". Với dạng dữ liệu dấu vân tay, các cấu trúc sẽ được mã hóa dưới dạng bit, mang đến một bộ dữ liệu có chiều rất lớn, do đó cách tiếp cận data-centric sẽ phù hợp với loại dữ liệu này. Một loạt các dấu vân tay phân tử hai và ba chiều đã được phát minh ra để mã hoá các đặc tính cấu trúc liên quan của phân tử nhỏ thành các chuỗi bit, cho phép dễ dàng xây dựng các mô hình QSAR từ các bộ dấu vân tay này [3]. Trên cơ sở đó, các bộ dấu vân tay phân tử khác nhau đã được ứng dụng để xây dựng mô hình máy học dự đoán khả năng ức chế hoạt tính kháng HIV integrase.

ĐỐI TƯỢNG-PHƯƠNG PHÁP NGHIÊN CỨU

Đối tượng nghiên cứu

Tài nguyên tính toán

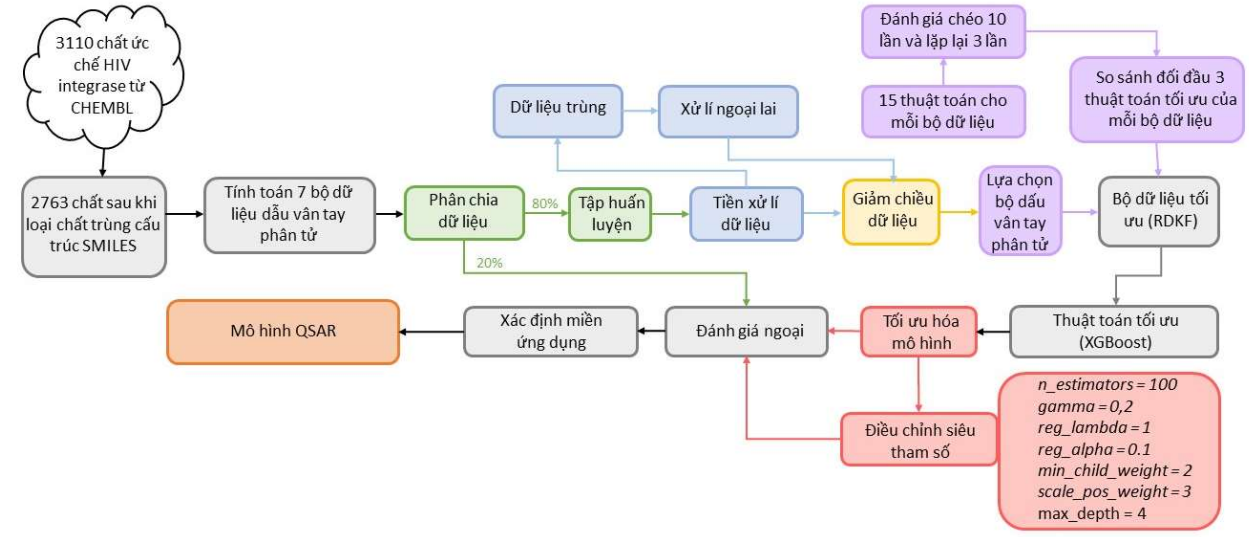
Nghiên cứu *in silico* trong đề tài được tiến hành trên máy tính Macbook Pro 2017, CPU 2.3 GHz Dual-Core Intel Core i5, Ram 8 GB 2133 MHz LPDDR3, GPU Intel Iris Plus Graphics 640 1536 MB, hệ điều hành MacOS 12.3.1 với ngôn ngữ lập trình được sử dụng là Python 3.9.7 cùng thư viện hoá tin RDKit [4] và học máy Scikit-learn 1.1.1 [5].

Dữ liệu xây dựng mô hình

Bộ dữ liệu gồm 3110 chất ức chế HIV integrase được trích xuất từ thư viện ChEMBL 31 [6], được chọn lựa các cấu trúc có cùng phương pháp đo hoặc phép đo tương đồng, với hoạt tính đã được chuẩn hoá thành giá trị "pChEMBL value", giá trị được nhiều nghiên cứu trên các tạp chí uy tín chấp thuận sử dụng [7]. Hoạt tính của các cấu trúc này dao động trong khoảng 0,46 nM đến 10000 nM và đã được chuẩn hoá thành giá trị "pChEMBL value" nằm trong khoảng 4 – 9.

Phương pháp nghiên cứu

Toàn bộ quy trình của nghiên cứu đều được neo giá trị ngẫu nhiên (random_state, seed,…) ở mốc 42, để đảm bảo tính lặp lại của các thử nghiệm. Quy trình tổng quát được thể hiện trong hình 1.



Hình 1. Quy trình xây dựng và so sánh mô hình QSAR

Xây dựng cơ dữ liệu

3110 cấu trúc SMILES (**data.csv*) đã được chuẩn hoá giá trị hoạt tính ức chế HIV integrase được đưa vào ngôn ngữ lập trình Python để xử lý: lọc các cấu trúc SMILES bị trùng, sau đó tối thiểu hoá năng lượng và tính toán các dấu vân tay phân tử bằng thư viện RDKit. Nghiên cứu sử dụng ba loại dấu vân tay khác nhau bao gồm:

- MACCS thuộc nhóm dấu vân tay khoá cấu trúc (structural key fingperprints), mô tả sự hiện diện hoặc vắng mặt của các nhóm chức năng đã được xác định trước [8].
- ECFP4 (*Extended-connectivity fingerprints*): thuộc nhóm dấu vân tay tròn (circular fingperprints), ghi lại các liên kết môi trường xuyên tâm của mỗi nguyên tử với sự gia tăng các lớp vỏ của liên kết nguyên tử [9].
- Dấu vân tay RDKit (RDKF): thuộc nhóm dấu vân tay topo (Topological fingerprints), nắm bắt các đường dẫn của các đặc điểm phân tử, chiếu lên một số liên kết nhất định [10].

Nghiên cứu xây dựng 7 bộ dữ liệu *(*data_fp.csv*) là tổ hợp của 3 loại dấu vân tay phân tử trên bao gồm: MACCS; ECFP4; RDKF; MACCS + ECFP4; MACCS + RDKF; ECFP4 + RDKF; MACCS + ECFP4 + RDKF.

Tiền xử lý dữ liệu

- Mô hình học máy hướng đến việc phân loại nhị phân, nên cần chuyển giá trị "pChEMBL Value" sang các giá trị "0" (không có hoạt tính) và "1" (có hoạt tính) dựa trên ngưỡng phân loại. Nghiên cứu lựa chọn ngưỡng phân loại là 7 (tương đương với giá trị IC50 = 100 nM).
- *Phân chia dữ liệu*: tập dữ liệu phải được phân chia bằng thư viện Scikit-learn thành tập huấn luyện (training set) và tập đánh giá ngoại (external validation set) theo tỷ lệ 4:1, theo nguyên lý "phân tầng".
- *Loại bỏ hàng trùng lặp và cột trùng lặp.*
- *Loại bỏ các dấu vân tay có phương sai thấp*: tính toán phương sai cho từng dấu vân tay, các dấu vân tay có phương sai ≤0,05 được loại bỏ.
- *Xử lý ngoại lai*: các dấu vân tay phân tử chỉ có giá trị 0 hoặc 1 nên sẽ không có dữ liệu ngoại lai đơn biến theo bách phân vị. Nghiên cứu sử dụng phương pháp xử lý ngoại lai đa biến dựa trên thuật toán yếu tố ngoại lai cục bộ (Local Outlier Factor).

Lựa chọn dấu vân tay phân tử

Các dấu vân tay phân tử thường có số chiều rất lớn, như RDKF có đến 2048 bit tương ứng với tạo thêm 2048 cột, việc sử dụng tất cả các dấu vân tay này để xây dựng mô hình dễ gây ra hiện tượng "lời nguyền chiều không gian" [11], dẫn đến quá khớp (overfiting). Chính vì vậy, việc lựa chọn một lượng dấu vân tay ít hơn những vẫn mô tả được khái quát cấu trúc là điều cần thiết. Nghiên cứu sử dụng phương pháp chọn lọc đặc trưng nội tại (intrinsic) dựa trên nền tảng của thuật toán rừng ngẫu nhiên (Random Forest), và được kiểm tra hiệu năng thông qua đánh giá chéo gập 10 lần, lặp lại 3 lần (10x3 RepeatedStratifiedKFold) dựa trên đại lượng độ chính xác trung bình (Average Precision - AP).

Xây dựng mô hình học máy

Các yếu tố để xây dựng được mô hình học máy bao gồm: thuật toán sử dụng (hồi quy logistic, rừng ngẫu nhiên, máy vecto hỗ trợ,…), phương pháp tối ưu mô hình (stochastic gradient descent, adelta, adam, rmsprop,…) và cách đánh giá mô hình (đánh giá nội và đánh giá ngoại). Nắm được các yếu tố này mới có thể xây dựng được mô hình học máy hoàn chỉnh. Nghiên cứu sử dụng 15 thuật toán học máy khác nhau cho từng bộ dữ liệu bao

gồm các thuật toán như hồi quy logistic (Logic), láng giềng gần (KNN), máy vector hỗ trợ (SVM), Gaussian Naive bayes (GNB), Bernoulli Naive bayes (BNB), phân tích phân biệt tuyến tính (LDA), phân tích phân biệt bậc 2 (QDA), thuật toán cây quyết định (DTree), rừng ngẫu nhiên (RF), Extra Tree (ExT), Adaboost (ADA), Gradient Boosting (Grad), XGboost (XGB), Catboost (Catbst) và mạng nơron truyền thẳng nhiều lớp (MLP). Tiến hành đánh giá nội bằng 10x3 RepeatedStratifiedKFold, kết quả đánh giá này được sử dụng để lựa chọn ra bộ dấu vân tay phù hợp nhất.

Lựa chọn bộ dữ liệu tối ưu

Sau khi xây dựng được các biểu đồ hộp và râu, chọn 3 thuật toán tối ưu nhất cho từng loại dữ liệu, tiến hành so sánh đối đầu theo thứ hạng của 3 thuật toán. Nghĩa là các thuật toán tốt nhất (nhì hoặc ba) của các bộ dấu vân tay sẽ so sánh đối đầu với nhau. Bộ dấu vân tay phân tử nào cho kết quả so sánh tốt hơn sẽ được lựa chọn để tiếp tục xây dựng mô hình.

Tối ưu hoá mô hình

Khi chọn được bộ dữ liệu tối ưu, nghiên cứu tiến hành chọn ra thuật toán để thực hiện tối ưu hoá. Các thuật toán được so sánh với nhau với 2 tiêu chí:

- Tiêu chí chính: đánh giá chéo cho trung bình (hoặc trung vị) điểm AP cao hơn các thuật toán còn lại.
- Tiêu chí phụ: đánh giá chéo cho trung bình (hoặc trung vị) F1 score hoặc độ nhạy cao hơn các thuật toán còn lại.

Tối ưu hoá siêu tham số cần chú trọng vào hai yếu tố:

- Kiểm soát vấn đề quá khớp (overfiting): cần cân nhắc giữa độ phức tạp và khả năng tổng quát hoá của mô hình (bias variance tradeoff).
- Giải quyết dữ liệu mất cân bằng: nghiên cứu đưa thêm một giá trị siêu tham số vào trong việc tối ưu mô hình nhằm giải quyết dữ liệu mất cân bằng (cost sensitive).

Đánh giá khả năng tổng quát hoá của mô hình

Đánh giá ngoại (external validation) liên quan đến việc sử dụng các bộ dữ liệu có nguồn gốc độc lập bên ngoài để đánh giá hiệu năng của mô hình dựa trên dữ liệu đầu vào và thường được coi là bằng chứng khách quan cho khả năng tổng quát hoá của mô hình.

Xác định miền ứng dụng của mô hình

Một trong những khía cạnh chính của mô hình QSAR là xác định miền ứng dụng (Applicability Domain - AD) của mô hình. AD được định nghĩa là một không gian hóa học được xây dựng bởi các bộ mô tả và đáp ứng sinh học được sử dụng trong quá trình xây dựng mô hình QSAR trên tập huấn luyện, và hiệu năng của mô hình QSAR sẽ bị giới hạn trong AD của mô hình (khả năng nội suy). Trong khi đó, khả năng ngoại suy của mô hình QSAR cho chất nằm trong không gian sinh học tiềm năng nhưng ngoài AD sẽ bị hạn chế. Tuy nhiên, cũng không thể khẳng định việc dự đoán này là sai, nhưng độ tin cậy của dự đoán sẽ không cao bằng việc nội suy [12]. Có nhiều cách tiếp cận miền ứng dụng như phương pháp dựa trên khoản xác định, phân tích thành phần chính, dựa trên khoảng cách (euclid, độ tương đồng,…), hình học (bao lồi),… Trong phạm vi thực hiện, nghiên cứu sử dụng kết hợp phương pháp phân tích thành phần chính (Principle component analysis – PCA) và bao lồi (convex hull) để xác định miền ứng dụng của mô hình.

KẾT QUẢ

Cơ sở dữ liệu

3110 dữ liệu các chất có giá trị "pChEMBL Value" trên đích tác động HIV integrase, loại bỏ 347 cấu trúc SMILES bị trùng, dữ liệu thô còn 2763 chất được tính toán dấu vân tay phân tử MACCS (167 bit), ECFP4 (4096 bit) và dấu vân tay RDKit (2048 bit) bằng thư viện RDKit, đồng thời tổ hợp 3 bộ dữ liệu trên thành các bộ hai và bộ ba, tổng cộng có 7 bộ dữ liệu dấu vân tay phân tử. Các bộ dữ liệu được xử lý độc lập và so sánh kết quả ở từng giai đoạn với nhau.

Tiền xử lý dữ liệu

- Dữ liệu được chia thành 2210 chất thuộc tập huấn luyện, 553 chất thuộc tập đánh giá ngoại, mức độ mất cân bằng dữ liệu ghi nhận là xấp xỉ 1:4, tức là cứ 5 chất trong cơ sở dữ liệu thì sẽ có 1 chất hoạt tính và 4 chất không hoạt tính.
- Phân tích ngưỡng phương sai, loại bỏ những dấu vân tay có phương sai thấp hơn 0,05, kết quả được mô tả trong bảng 1.
- Xử lý ngoại lai đa biến bằng nhân tố ngoại lai cục bộ (LOF) với thông số *n_neighbors = 20*. Kết quả xử lý ngoại lai được tóm tắt trong bảng 1.

Bảng 1. Kết quả tiền xử lý của các bộ dữ liệu huấn luyện

	MACCS	ECFP4	RDKF	MACCS + ECFP4	MACCS + RDKF	ECFP4+ RDKF	MACCS + ECFP4 + RDKF
Số dấu vân tay	167	4096	2048	4263	2215	6144	6311
Tỷ lệ mất cân bằng	27,90%						
Số dấu vân tay sau phân tích ngưỡng phương sai	115	403	1808	517	1855	2433	2547
Chất ngoại lai	0	95	42	63	41	45	37

Lựa chọn dấu vân tay phân tử

Bảy bộ dữ liệu sau khi được tiền xử lý vẫn chưa thể đưa vào huấn luyện mô hình vì số chiều còn rất lớn. Nghiên cứu đã sử dụng 9 phương pháp khác nhau để giảm số chiều dấu vân tay (chọn lọc đặc trưng), sau đó sử dụng thuật toán rừng ngẫu nhiên để tiến hành đánh giá chéo nội gập 10 lần và lặp lại 3 lần (10x3 RepeatedStratifiedKFold) với đại lượng đánh giá là AP. Kết quả đối với hai phương pháp thống kê mô tả là chi bình phương và thông tin tương hỗ, cho giá trị AP ở 7 bộ dữ liệu đề dưới 0,85. Trong khi đó, các phương pháp chọn lọc đặc trưng nội tại ứng dụng thuật toán máy vector hỗ trợ, hồi quy logistic kết hợp với lasso, rừng ngẫu nhiên và boosting cho kết quả tốt và ổn định hơn và có thể đạt hơn 0,9.

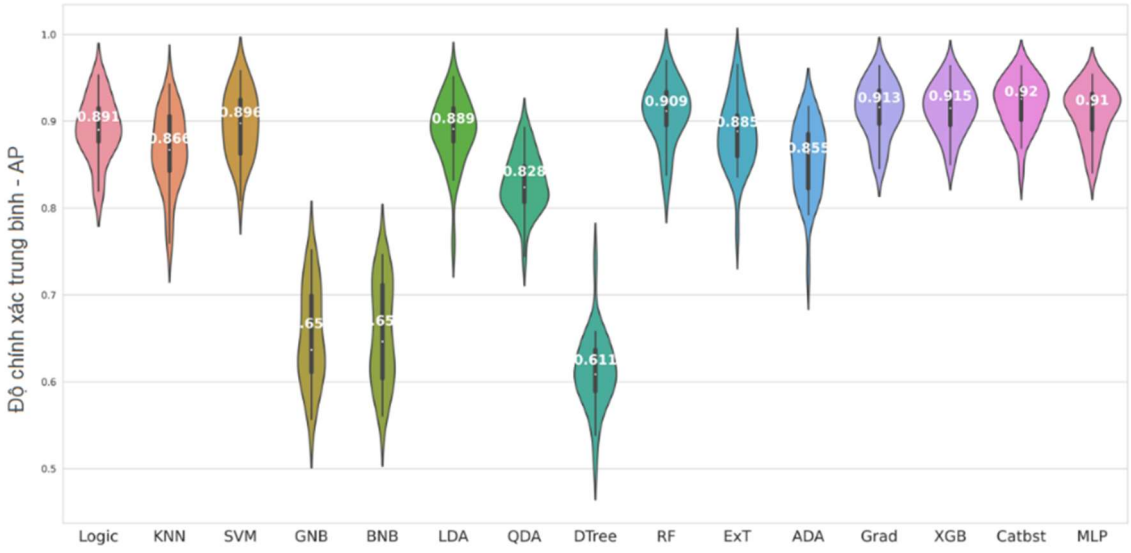
Nghiên cứu đã chọn ra một thuật toán tối ưu để giảm chiều dữ liệu cho từng bộ dấu vân tay phân tử và được tóm tắt trong bảng 2.

Bảng 2. Kết quả chọn lọc thông số mô tả ở các bộ dữ liệu.
Trong đó, đại lượng AP được biểu diễn bằng giá trị trung bình ± sai số chuẩn (trung vị)

	MACCS	ECFP4	RDKF	MACCS + ECFP4	MACCS + RDKF	ECFP4+ RDKF	MACCS + ECFP4 + RDKF
Tối ưu	SVM	SVM	ExT	XGB	ExT	Logic	XGB
AP	0,847±0,036 (0,848)	0,897±0,033 (0,903)	0,896±0,038 (0,899)	0,897±0,025 (0,900)	0,900±0,030 (0,899)	0,905±0,033 (0,903)	0,903±0,028 (0,906)
Số dấu vân tay	105	305	346	196	399	576	400

Đánh giá ảnh hưởng của dấu vân tay phân tử

Để đánh giá được ảnh hưởng của các loại dấu vân tay lên hiệu năng của mô hình học máy, nghiên cứu đã sử dụng 15 thuật toán học máy khác nhau cho từng bộ dữ liệu và tiến hành đánh giá nội bằng 10x3 RepeatedStratifiedKFold, kết quả đánh giá này được sử dụng để lựa chọn ra bộ dấu vân tay phù hợp nhất.



Hình 2. Biểu đồ violin thể hiện phân bố kết quả đánh giá nội cho bộ dữ liệu RDKit

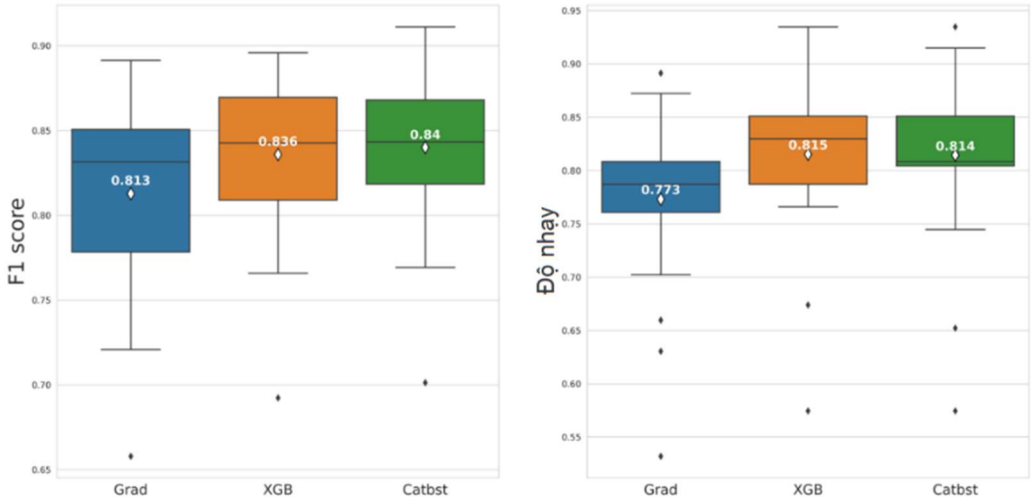
Nghiên cứu tiến hành so sánh đối đầu hiệu năng đánh giá của 3 thuật toán tốt nhất cho từng bộ dữ liệu (Bảng 3). Kết quả, dấu vân tay RDKF cho kết quả tối ưu ở cả 3 thuật toán đều cao hơn các thuật toán còn lại khi so sánh theo thứ hạng.

Bảng 3. So sánh kết quả xây dựng mô hình của các bộ dữ liệu

	MACCS	**ECFP4**	**RDKF**	**MACCS + ECFP4**	**MACCS + RDKF**	**ECFP4+ RDKF**	**MACCS + ECFP4 + RDKF**
FP	105	305	346	196	399	576	400
1	Catbst	XGB	Catbst	Catbst	Catbst	Catbst	XGB
	0,877±0,037 (0,882)	0,910±0,024 (0,917)	**0,920±0,029 (0,926)**	0,911±0,022 (0,909)	0,916±0,024 (0,916)	0,918±0,028 (0,919)	0,916±0,025 (0,919)
2	XGB	MLP	XGB	XGB	XGB	XGB	MLP
	0,860±0,039 (0,860)	0,904±0,028 (0,908)	**0,915±0,028 (0,915)**	0,908±0,021 (0,908)	0,911±0,027 (0,911)	0,914±0,029 (0,916)	0,914±0,024 (0,914)
3	Grad	Catbst	Grad	MLP	RF	MLP	Catbst
	0,855±0,041 (0,852)	0,903±0,029 (0,908)	**0,913±0,031 (0,916)**	0,904±0,022 (0,907)	0,907±0,025 (0,907)	0,906±0,030 (0,903)	0,913±0,021 (0,915)

Tối ưu hoá mô hình

Từ kết quả của bảng 3 cho bộ dữ liệu RDKit có thể thấy giá trị AP đánh giá nội giữa 3 thuật toán Gradient Boosting, XGboost và Catboost chênh lệch nhau không lớn (dưới 0,1 đơn vị), thậm chí có thể thấy ở biểu đồ violin (hình 2), mật độ Kernel của 3 thuật toán này tương đồng với nhau. Nếu chỉ dựa vào 1 đại lượng AP, rất khó để chọn ra thuật toán tốt để tối ưu, vì vậy nghiên cứu đưa thêm 2 tiêu chí phụ để đánh giá khác là F1 score và độ nhạy.



Hình 3. So sánh hiệu năng của 3 thuật toán với đại lượng F1 score (trái) và độ nhạy (phải)

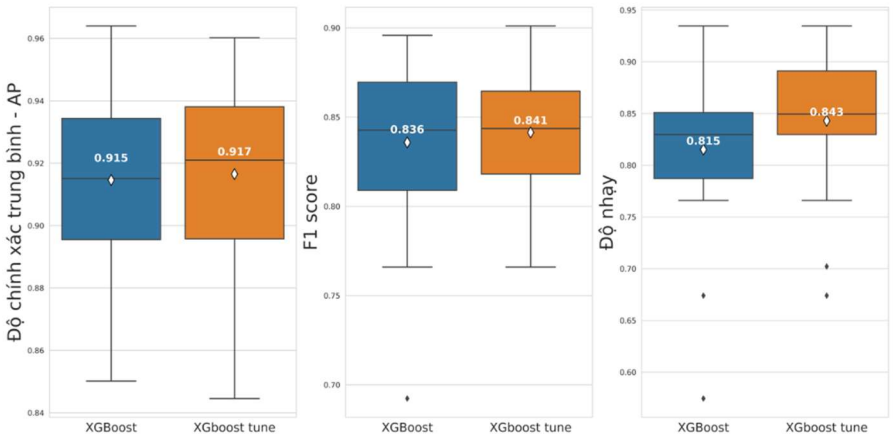
Từ kết quả của hình 3, giá trị F1 score của thuật toán Gradient Boosting thấp hơn so với hai thuật toán còn lại. Điều này càng được khẳng định khi đánh giá thêm đại lượng độ nhạy vào so sánh, thuật toán Gradient Boosting cho điểm số đánh giá nội trung bình dưới 0,8 thấp hơn so với hai thuật toán còn lại. Chính vì vậy, nghiên cứu chọn thuật toán XGBoost để tối ưu cho mô hình học máy vì tốc độ huấn luyện mô hình nhanh hơn so với Catboost.

XGBoost là từ viết tắt của eXtreme Gradient Boosting, đề cập đến kỹ thuật đẩy giới hạn tài nguyên tính toán cho các thuật toán cây tăng cường (boosting). Đây là một thuật toán tổng hợp (ensemble) trong đó các mô hình mới được thêm vào để sửa các lỗi cho các mô hình hiện tại. Các mô hình được thêm vào tuần tự cho đến khi không thể cải thiện hiệu năng được thêm nữa. Nghiên cứu sử dụng các siêu tham số để kiểm soát việc học của XGBoost bao gồm:

Bảng 4. Giá trị siêu tham số được khảo sát và giá trị tối ưu

Siêu tham số	Ý nghĩa	Khoảng khảo sát	Giá trị tối ưu
n_estimators	Số lượng cây quyết định	[50;100;150;200;250;300]	100
max_depth	Độ sâu tối đa của cây	[2;3;4;5;6;7]	4
gamma	Sự giảm hàm mất mát tối thiểu để có thể tạo ra một node lá	[0,2;0,4;0,6;0,8;1,0]	0,2
min_child_weight	Số lượng mẫu tối thiểu trong một node con	[1;2;3;4;5]	2
reg_lambda	Chính quy hoá L2	[0,01; 0,1; 1; 10]	1
reg_alpha	Chính quy hoá L1	[0,01; 0,1; 1; 10]	0,1
scale_pos_weight	Trọng số cân bằng cho lớp hoạt tính và không hoạt tính	[1;2;3;4;5]	3

Kết quả đánh giá mô hình trước và sau khi tối ưu được biểu diễn trong hình 4.



Hình 4. So sánh hiệu năng mô hình XGBoost trước và sau khi tối ưu

Đánh giá khả năng tổng quát hoá của mô hình

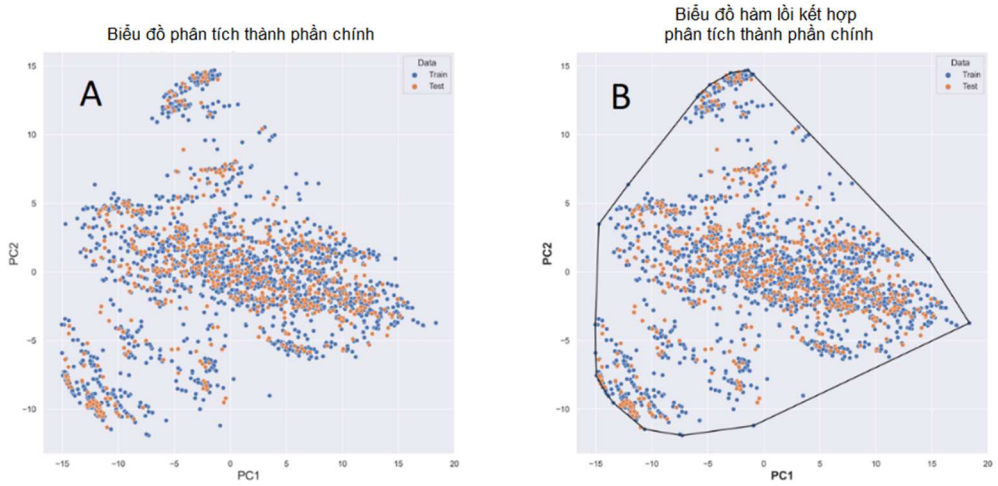
Tập dữ liệu đánh giá ngoại (20%) được chia từ lúc đầu được dùng để đánh giá khả năng tổng quát hoá của mô hình. Kết quả được mô tả chi tiết trong bảng 5.

Bảng 5. Kết quả đánh giá nội và đánh giá ngoại của mô hình XGBoost

	Đánh giá nội			**Đánh giá ngoại**		
	AP	F1	Độ nhạy	AP	F1	Độ nhạy
XGB	0,914	0,836	0,815	0,906	0,829	0,805
XGBoost_tune	0,917	0,841	0,843	0,909	0,828	0,814

Xác định miền ứng dụng của mô hình

Trong phân tích đơn biến, dấu vân tay chỉ có giá trị 0 hoặc 1 nên hầu như không có giới hạn các biến như trong thông số mô tả. Tuy nhiên, khi xem xét một cách tổng quát các dấu vân tay trong không gian có thể có tương quan với nhau nên không thể chỉ phân tích đơn biến. Nghiên cứu sử dụng phương pháp phân tích thành phần chính (PCA) để giảm số chiều của bộ dấu vân tay RDKF xuống còn 2 chiều và xem phân bố của các chất trong không gian (Hình 5A). Miền ứng dụng được xác định là vùng không gian bao phủ tối đa bởi tập huấn luyện. Các chất trong tập đánh giá ngoại nằm ngoài khoảng không gian hai chiều này sẽ được xác định là giá trị ngoài miền ứng dụng và sẽ có độ tin cậy khi dự đoán thấp hơn. Như vậy, phương pháp xây dựng miền giá trị trên tập huấn luyện sẽ quyết định độ lớn của khoảng không gian đó. Nghiên cứu sử dụng phương pháp bao lồi để xác định được khối đa giác (polygon) bao phủ được tập huấn luyện. Dựa vào hình 5B, có thể nhận thấy có một giá trị của tập đánh giá ngoại nằm ngoài miền giới hạn được xác định bởi đường màu đen. Tuy nhiên, có đến 6 điểm nằm trên đường giới hạn. Những chất nằm quá một nửa ra ngoài đường giới hạn được xác định là chất nằm ngoài miền ứng dụng (hoặc giá trị ngoại lai). Từ đó, 3 chất có mã định danh lần lượt là CHEMBL492667, CHEMBL560185, CHEMBL1770429 của tập đánh giá ngoại được xác định là những cấu trúc nằm ngoài miền ứng dụng và có độ tin cậy thấp khi sử dụng mô hình QSAR này để dự đoán.



Hình 5. (A) Biểu đồ phân tán sử dụng phân tích thành phần chính mô tả không gian hóa học của bộ dữ liệu RDKF. (B) Miền ứng dụng dựa trên kỹ thuật phân tích thành phần chính và bao lồi

BÀN LUẬN

Cách tiếp cận học máy truyền thống thường chú trọng vào giải thuật, tuy nhiên, việc tối ưu mô hình thường mất nhiều thời gian và cải thiện hiệu năng không đáng kể. Trong khi xu thế phát triển mô hình học máy dựa trên nền tảng dữ liệu (data-centric) đang là xu hướng mới, với hiệu năng được cải thiện tốt hơn [13]

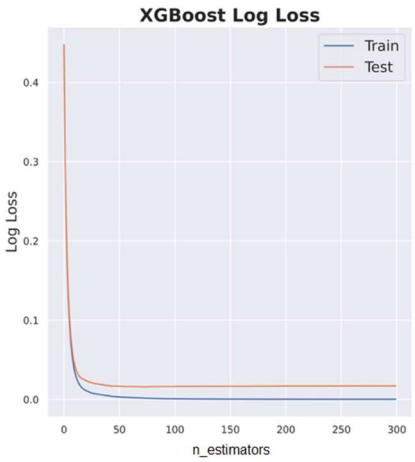
Khi so sánh một cách đơn lẻ ba loại dấu vân tay phân tử với nhau, bộ vân tay RDKF (346) cho hiệu năng vượt trội (đều cao hơn 0,1 đơn vị) so với MACCS (105) và ECFP4 (305). Trong đó, bộ dấu vân tay MACCS cho kết quả kém nhất, khi hầu hết các đánh giá điểm số AP đều dưới 0,9. Tuy nhiên, khi tổ hợp các bộ dấu vân tay, hiệu năng có tăng nhưng vẫn thấp hơn khi so sánh với RDKF:

- Tổ hợp MACCS + ECFP4 (196 dấu vân tay): tổ hợp này cho kết quả tối ưu hơn so với đơn lẻ từng bộ dấu vân tay.
- Tổ hợp MACCS + RDKF (399 dấu vân tay): tổ hợp này cho kết quả tối ưu hơn so với MACCS đơn lẻ, nhưng cho kết quả thấp hơn (0,04 đơn vị) khi so sánh với RDKF, mặc dù số lượng dấu vân tay sử dụng mô hình là nhiều hơn.
- Tổ hợp ECFP4 + RDKF (576 dấu vân tay): với số lượng dấu vân tay sau khi trích lọc đặc trưng là 576, được nhận định là tiềm năng nhất khi có thể mô tả được nhiều đặc tính phân tử nhất, tổ hợp này đã cho kết quả đánh giá cao hơn hầu hết các bộ dữ liệu còn lại, nhưng vẫn thấp hơn so với RDKF.
- Tổ hợp MACCS + ECFP4 + RDKF (400 dấu vân tay): kết quả đánh giá của tổ hợp này để tốt hơn so với MACCS hay ECFP4 đơn lẻ, nhưng vẫn thấp hơn so với RDKF.

Từ các phân tích trên, có thể thấy số lượng dấu vân tay có ảnh hưởng đến hiệu năng của mô hình. Số lượng dấu vân tay nên ở mức 300-400 để có thể đạt được hiệu năng mô hình từ trung bình đến tốt.

Một phân tích biểu đồ lỗi (cross-entropy trong trường hợp phân loại) được thực hiện để chứng minh mô hình không bị quá khớp. Biểu đồ bao gồm trục tung là cross-entropy (log loss) của tập huấn luyện và kiểm tra, trục hoành sẽ biểu diễn số cây quyết định (n_estimators) được sử dụng (n_estimators). Từ đồ thị theo dõi lỗi (Hình 6) có thể thấy mô hình chưa tới điểm bị quá khớp. Tuy nhiên, từ cây quyết định thứ 50, cross-entropy của tập huấn luyện và kiểm tra đã bắt đầu có xu hướng giảm chậm và đi ngang ở epochs thứ 100. Điều này cũng trùng khớp với kết quả tối ưu của mô hình, chọn ra *n_estimators* =

100. Như vậy, có thể thấy với siêu tham *n_estimators = 100*, thì mô hình XGBoost sẽ không bị quá khớp.



Hình 6. Biểu đồ theo dõi lỗi huấn luyện

KẾT LUẬN

Nhóm nghiên cứu đã đánh giá được ảnh hưởng của dấu vân tay phân tử lên hiệu năng của mô hình học máy, kết luận dấu vân tay phân tử RDKF có ảnh hưởng lớn đến hiệu năng dự đoán của mô hình. Thuật toán XGBoost cho thấy hiệu năng tốt và ổn định ở hầu hết các bộ dữ liệu, đặc biệt là dấu vân tay phân tử RDKF. Mô hình XGBoost sau khi tối ưu cho kết quả tốt hơn so với ban đầu với kết quả đánh giá ngoại đều cao như trung bình độ chính xác (0,91), F1 score đạt 0,83, độ đúng là 0,92, độ nhạy là 0,814 và độ chính xác là 0,84. Nghiên cứu sẽ tiến hành triển khai giao diện người dùng cho việc dự đoán khả năng ức chế HIV integrase, kết hợp cùng ứng dụng "học tích cực" trong quá trình kiểm thử, tổng hợp và thử hoạt tính ức chế HIV Integrase của các chất từ thư viện nội bộ, để nâng cao hiệu năng dự đoán của mô hình. Từ đó, sàng lọc quy mô lớn để "tái định vị thuốc" hoặc tìm ra được cấu trúc tiềm năng điều trị HIV.

TÀI LIỆU THAM KHẢO

1. Pau AK, George JM. Antiretroviral therapy: current drugs. *Infectious disease clinics of North America*. 2014;28(3):371-402. doi:10.1016/j.idc.2014.06.001
2. Baptista D, Ferreira PG, Rocha M. Deep learning for drug response prediction in cancer. *Briefings in bioinformatics*. 2021;22(1):360-379. doi:10.1093/bib/bbz171
3. Muegge I, Mukherjee P. An overview of molecular fingerprint similarity search in virtual screening. *Expert opinion on drug discovery*. 2016;11(2):137-48. doi:10.1517/17460441.2016.1117070
4. Landrum G. *Rdkit documentation. Release 1, 4*. 2013.
5. Kramer O, Kramer OJMlfes. Scikit-learn. 2016:45-53.
6. Mendez D, Gaulton A, Bento AP, et al. ChEMBL: towards direct deposition of bioassay data. 2019;47(D1):D930-D940. https://doi.org/10.1093/nar/gky1075
7. Burggraaff L, van Vlijmen HWT, AP IJ, van Westen GJP. Quantitative prediction of selectivity between the A(1) and A(2A) adenosine receptors. *Journal of cheminformatics*. 2020;12(1):33. doi:10.1186/s13321-020-00438-3

8. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. *Journal of chemical information and computer sciences*. 2002;42(6):1273-80. doi:10.1021/ci010132r
9. Rogers D, Hahn M. Extended-connectivity fingerprints. *Journal of chemical information and modeling*. 2010;50(5):742-54. doi:10.1021/ci100050t
10. Carhart RE, Smith DH, Venkataraghavan R. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of chemical information and computer sciences*. 1985;25(2):64-73. doi:10.1021/ci00046a002
11. Köppen, M. The curse of dimensionality. *In 5th online world conference on soft computing in industrial applications*. 2000;1:4-8.
12. Roy K, Kar S, Das RN. Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. *Academic press*; 2015:231-289. doi:10.1016/b978-0-12-801505-6.00007-7
13. Montáns FJ, Chinesta F, Gómez-Bombarelli R, Kutz JN. Data-driven modeling and learning in science and engineering. *Comptes Rendus Mécanique*. 2019;347(11):845-855. doi:https://doi.org/10.1016/j.crme.2019.11.009