BỘ GIÁO DỤC VÀ ĐÀO TẠO BỘ Y TẾ

ĐẠI HỌC Y DƯỢC THÀNH PHỐ HỒ CHÍ MINH

NGÔ TRIỀU DỦ

XÂY DỰNG MÔ HÌNH PHÂN LOẠI VÀ DỰ ĐOÁN CÁC CHẤT ỨC CHẾ BƠM NGƯỢC P-GLYCOPROTEIN, NORA VÀ ỨNG DỤNG TRONG VIỆC SÀNG LỌC CÁC CHALCON CÓ KHẢ NĂNG ỨC CHẾ BƠM NORA CỦA *STAPHYLOCOCCUS AUREUS* ĐA ĐỀ KHÁNG THUỐC

LUẬN ÁN TIẾN SĨ DƯỢC HỌC

TP. HỒ CHÍ MINH, NĂM 2019

BỘ GIÁO DỤC VÀ ĐÀO TẠO BỘ Y TẾ

ĐẠI HỌC Y DƯỢC THÀNH PHỐ HỒ CHÍ MINH

NGÔ TRIỀU DỦ

XÂY DỰNG MÔ HÌNH PHÂN LOẠI VÀ DỰ ĐOÁN CÁC CHẤT ỨC CHẾ BƠM NGƯỢC P-GLYCOPROTEIN, NORA VÀ ỨNG DỤNG TRONG VIỆC SÀNG LỌC CÁC CHALCON CÓ KHẢ NĂNG ỨC CHẾ BƠM NORA CỦA *STAPHYLOCOCCUS AUREUS* ĐA ĐỀ KHÁNG THUỐC

NGÀNH: HÓA DƯỢC
MÃ SỐ: 62720403

LUẬN ÁN TIẾN SĨ DƯỢC HỌC

NGƯỜI HƯỚNG DẪN KHOA HỌC:

1. PGS. TS. THÁI KHẮC MINH

2. PGS. TS. TRẦN THÀNH ĐẠO

TP. HỒ CHÍ MINH, NĂM 2019

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi, các kết quả nghiên cứu được trình bày trong luận án là trung thực, khách quan và chưa từng được công bố ở bất kỳ nơi nào.

Tác giả luận án

Ngô Triều Dủ

MỤC LỤC

DANH MỤC CÁC CHỮ VIẾT TẮT, THUẬT NGỮ

Chữ viết tắt	Chữ viết đầy đủ	Nghĩa tiếng Việt/Định nghĩa
ABC	ATP Binding Cassette	Họ các protein chuyên chở phụ thuộc ATP
ADMET	Absorption, Distribution, Metabolism, Excretion and Toxicity	Hấp thu, phân bố, chuyển hóa, thải trừ và độc tính
Ci	Ciprofloxacin	
EPI	Efflux Pump Inhibitor	Chất ức chế bơm ngược
	Efflux Pump Inhibition	Sự ức chế bơm ngược
GA-PLS	Genetic Algorithm-Partial Least Square	Bình phương tối thiểu-thuật toán di truyền
in silico		Thực hiện trên máy tính hoặc thông qua mô phỏng máy tính
MDR	Multidrug Resistance	Đề kháng đa thuốc
MFS	Major Facilitator Superfamily	Liên họ trợ giúp chính
MIC	Minimum Inhibitory Concentration	Nồng độ ức chế tối thiểu
MRSA	Methicillin-Resistant *Staphylococcus aureus*	*Staphylococcus aureus* đề kháng methicillin
NorA		Bơm ngược MFS của *Staphylococcus aureus*
PaβN	Phenyl-arginin-beta-naphthylamid	
P-gp	P-glycoprotein	
QSAR	Quantitative Structure-Activity Relationship	Mối quan hệ định lượng cấu trúc - tác dụng
QZ59-RRR	Cyclic-tris-(R)-valineselenazol	
RF	Reversal Fold	Số lần đảo ngược
RMSD	Root Mean Square Deviation	Căn bậc hai của độ lệch bình phương trung bình
SAR	Structure-Activity Relationship	Mối quan hệ cấu trúc - tác dụng
SMI	Small Molecule Inhibitor	Chất ức chế phân tử nhỏ
S. aureus	*Staphylococcus aureus*	
2D	Two-Dimension	Hai chiều
3D	Three-Dimension	Ba chiều

DANH MỤC CÁC BẢNG

Trang

DANH MỤC CÁC HÌNH, ĐỒ THỊ

Trang

MỞ ĐẦU

Đề kháng đa thuốc (multidrug resistance - MDR) được nhìn nhận là một trong những vấn đề chính thách thức việc điều trị thành công bệnh ung thư cũng như bệnh nhiễm trùng ở người trong nhiều thập kỷ qua. Các tế bào khối u và các chủng vi khuẩn tự bảo vệ mình khỏi sự tấn công của các thuốc hóa trị bằng nhiều cơ chế khác nhau, trong đó sự đề kháng qua trung gian bơm ngược đóng một vai trò rất quan trọng [121], [122], [141]. Bằng cách bài xuất nhiều loại hợp chất đa dạng về cấu trúc ra khỏi tế bào, các protein màng làm cho sự tích lũy nội bào của thuốc giảm xuống thấp dưới nồng độ có tác dụng và vì vậy giúp cho các tác nhân gây bệnh giảm sự nhạy cảm với thuốc [8], [107]. Trong số các protein thuộc hệ thống bơm ngược của cả tế bào có nhân điển hình và tế bào chưa có nhân điển hình, P-glycoprotein ở động vật có vú và NorA ở vi khuẩn là hai protein được nghiên cứu nhiều nhất, liên quan đến vai trò của chúng trong việc chuyên chở thuốc ra ngoài tế bào [107].

P-glycoprotein của người (P-gp/ABCB1/MDR1) và NorA của *Staphylococcus aureus* tiếp tục là hai mục tiêu thuốc được chọn của đề tài nghiên cứu này bởi vì tầm quan trọng to lớn của chúng về mặt lâm sàng. Với P-gp, bơm ngược này vừa là protein không mục tiêu (antitarget/nontarget) ảnh hưởng đến dược động học và độc tính (ADMET) của nhiều thuốc khác nhau [2], vừa là protein mục tiêu bởi vì sự biểu lộ quá mức của nó đóng góp cho sự đề kháng của ung thư với hóa trị [186]. Trong khi đó, NorA được biết là đóng vai trò chính trong sự phát triển đề kháng của vi khuẩn với các kháng sinh fluoroquinolon [33]. Mặc dù có cấu trúc khác nhau, các bơm ngược của động vật có vú và vi khuẩn lại có sự tương đồng chất nền đủ lớn, với nhiều nghiên cứu đã báo cáo các chất ức chế cả P-gp và NorA như verapamil [120], reserpin [162], piperin [82], capsaicin [79], osthol, curcumin [77], …

Qua nhiều thập kỷ nghiên cứu, ba thế hệ các chất ức chế phân tử nhỏ (small molecule inhibitor - SMI) của P-gp được khám phá và phát triển [136], nhưng vẫn chưa có thuốc nào sẵn có cho mục đích chẹn P-gp trên lâm sàng. Những lý do giải thích hợp lý được đưa ra, bao gồm tính tan kém, tính đặc hiệu kém, tác dụng phụ, độc tính và tương tác dược động [19], [160], [176]. Mặt khác, cũng chưa có chất ức chế bơm NorA nào được đưa vào thử nghiệm trên người [67]. Trong số các phương pháp hợp lý được đề nghị để ức chế P-gp, các thành phần từ tự nhiên nhận được nhiều sự quan tâm bởi vì tính an toàn, không gây độc [173]. Cho ví dụ, CBT-1 là một alkaloid thực vật loại bisbenzylisoquinolin được công ty CBA Pharma Inc. phát triển như một chất ức chế P-gp dùng đường uống và các kết quả lâm sàng ban đầu đầy hứa hẹn của chất này khi phối hợp với doxorubicin [126] và paclitaxel [81], [125] đã khuyến khích các nỗ lực nghiên cứu tiếp theo để tìm kiếm các chất ức chế bơm ngược mới, an toàn và hiệu quả. Cùng với alkaloid, khung flavonoid cũng được xem xét cho hoạt tính ức chế P-gp ở khối u của người [11], [52], [136], [173] và NorA ở vi khuẩn *S. aureus* [67], [215]. Các nghiên cứu trên nhóm cấu trúc này đã thu được các dẫn xuất chalcon có tiềm lực ức chế hai loại bơm ngược [69], [139] và góp phần định hướng cho đề tài thực hiện sàng lọc, thử nghiệm hoạt tính sinh học trên tập dữ liệu gần 100 chalcon nội bộ đã được thiết kế và tổng hợp trước đó với sự đa dạng về các nhóm thế.

Các phương pháp thiết kế thuốc với sự trợ giúp của máy tính (computer-aided drug design - CADD) được xem là một lựa chọn khả thi với chi phí thấp, bao gồm thiết kế dựa vào cấu trúc (structure-based) và dựa vào phối tử (ligand-based), giúp dự đoán và làm sáng tỏ các tương tác phối tử - protein trong giai đoạn sớm của quá trình khám phá thuốc [112], [137]. Trọng tâm của đề tài là xây dựng các mô hình phân loại và dự đoán máy tính giúp giải quyết các vấn đề được nhìn nhận từ các nghiên cứu *in silico* đã được công bố trước đó (tham khảo **Mục 4.1.1** và **Mục 4.1.2**), bao gồm những nghi vấn về khả năng ngoại suy (do được phát triển từ các tập dữ liệu tương đối nhỏ, không đảm bảo tính đa dạng, đồng nhất) và những hạn chế của các mô hình học máy đơn lẻ được báo cáo trong các nghiên cứu này. Ngoài các điều kiện đánh giá thông kê chặt chẽ, khả năng ứng dụng của các công cụ máy tính thu được còn được kiểm chứng bằng các thử nghiệm *in vitro* trên chủng vi khuẩn chuẩn biểu lộ quá mức bơm ngược cũng như trên các chủng đề kháng phân lập từ lâm sàng. Đồng thời qua đó, các ứng viên ức chế bơm ngược tiềm năng được khám phá.

Vì những lý do trên, nghiên cứu này được thực hiện với mục tiêu xây dựng mô hình phân loại và dự đoán các chất ức chế bơm ngược P-glycoprotein, NorA và ứng dụng trong việc sàng lọc các chalcon có khả năng ức chế bơm NorA của *S. aureus* đa đề kháng thuốc. Để đạt được mục tiêu này, cần tiến hành bốn nội dung sau đây:

1. Xây dựng các mô hình máy tính dựa trên phối tử, bao gồm:
   - Các mô hình học máy đơn lẻ và kết hợp giúp phân loại tốt chất ức chế, chất không ức chế P-gp; và dự đoán tốt hoạt tính ức chế bơm ngược này (IC_{50}).
   - Các bản đồ nhận thức về sự chồng phủ phối tử giữa P-gp và NorA, qua đó xác định các tính chất lý hóa, dấu vân tay cần thiết để ức chế ít nhất một trong hai bơm ngược.
   - Mô hình pharmacophore cho các chất ức chế P-gp mạnh trong điều trị ung thư và mô hình pharmacophore cho các chất ức chế NorA nhưng không ức chế P-gp trong điều trị nhiễm trùng.
2. Xây dựng các mô hình máy tính dựa trên cấu trúc (mô hình tương đồng của P-gp) và thực hiện docking phân tử nhằm xác định các tương tác gắn kết, cũng như ái lực gắn kết của phức hợp phối tử-protein.
3. Sàng lọc các chất "hit" là những ứng viên ức chế P-gp, NorA mới và hiệu quả từ hai thư viện nội bộ và Ngân hàng Thuốc bằng các công cụ máy tính thu được.
4. Đánh giá *in vitro* khả năng ức chế bơm ngược NorA của các chalcon "hit" nội bộ, qua đó làm giảm sự đề kháng với ciprofloxacin khi phối hợp trên chủng *S. aureus* SA-1199B (biểu lộ quá mức NorA) và một số chủng SA lâm sàng.

CHƯƠNG 1. TỔNG QUAN

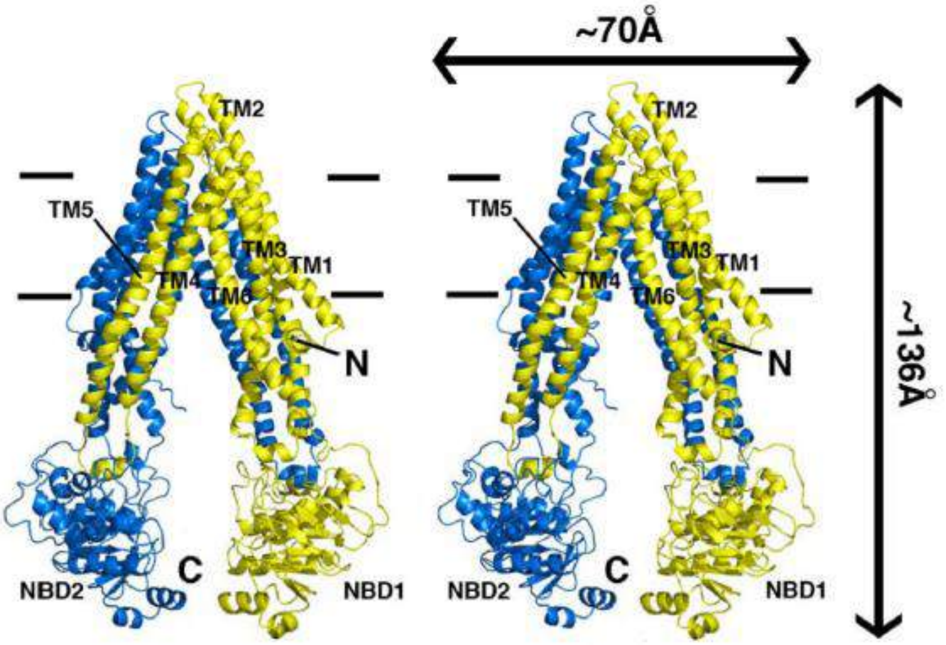
1.1. Tổng quan về các bơm ngược nghiên cứu

1.1.1. Tổng quan về P-glycoprotein

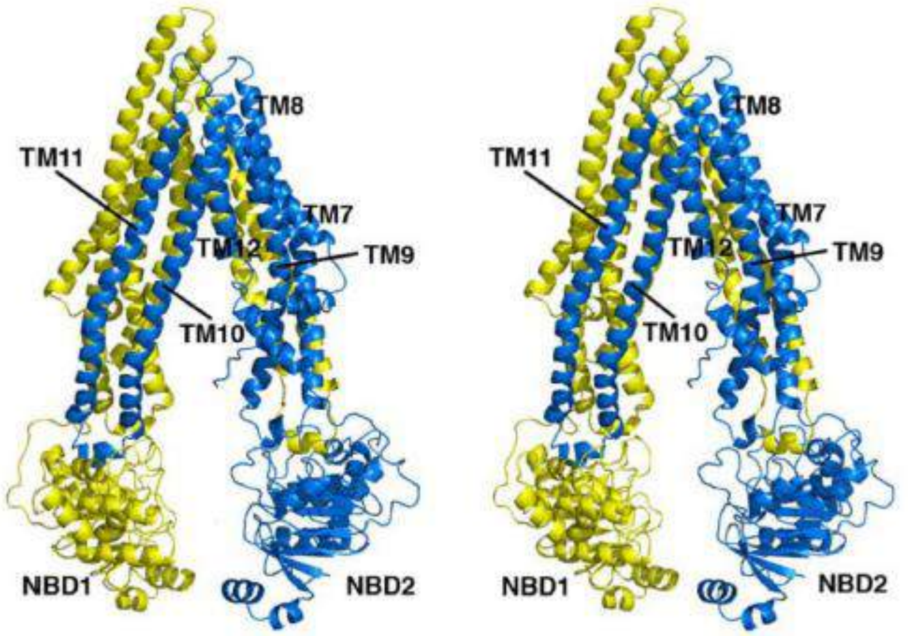
P-glycoprotein của người (P-gp) được mã hóa bởi gen ABCB1/MDR1 và được Dano mô tả lần đầu tiên vào năm 1973 [35]. Đây là một trong những thành viên quan trọng nhất và được nghiên cứu nhiều nhất của liên họ các protein chuyên chở phụ thuộc ATP (ATP Binding Cassette - ABC) [177], [220]. Hoạt tính bơm ngược sử dụng năng lượng, tính đặc hiệu chất nền rộng (các hợp chất tự nhiên, các tác nhân kháng ung thư, các peptid, các steroid, các lipid, các cytokin, thuốc nhuộm và các ion), cùng với sự phân bố ở cả các mô bình thường (ruột, não, tinh hoàn, nhau thai, gan và thận) và khối u, là nền tảng cho các vai trò của protein này trong sinh lý của cơ thể và trong hóa trị liệu [166]. Với sự tham gia vào cơ chế phòng vệ tự nhiên chống lại các chất ngoại sinh như độc tố và thuốc, P-gp được xem là một protein không mục tiêu (antitarget/nontarget) trong quá trình khám phá và phát triển thuốc, cùng với kênh kali hERG (human ether-a-go-go related gene), hệ thống các enzym cytochrom P450s và thụ thể trong nhân PXR (pregnane X-receptor) [62], [182]. Việc chẹn P-gp bằng các chất ức chế (ví dụ ketoconazol) có thể làm thay đổi nồng độ trong máu của các thuốc sử dụng chung (ví dụ terfenadin) hoặc của chất chuyển hóa, dẫn đến các tương tác thuốc - thuốc và các tác dụng dược lý không mong muốn (ví dụ kéo dài khoảng QT/xoắn đỉnh) [190]. Ngoài vai trò bảo vệ cơ thể, P-gp còn đóng vai trò quan trọng trong hiện tượng đề kháng đa thuốc (multidrug resistance - MDR) của các tế bào ung thư dựa trên khả năng chuyên chở chủ động các thuốc gây độc tế bào ra ngoài và cũng được xem là một mục tiêu lâm sàng trong hóa trị liệu [93]. Các nhóm thuốc kháng ung thư là chất nền của P-gp bao gồm anthracyclin (doxorubicin, daunorubicin, epirubicin, idarubicin), alkaloid dừa cạn (vincristin, vinblastin, vinoreblin, vindesin), taxan (paclitaxel, docetaxel), epipodophyllotoxin (etoposid, teniposid), camptothecin (topotecan, irinotecan) và nhóm các thuốc khác (mitoxantron, trimetrexat, actinomycin D, methotrexat, colchicin, tamoxifen, imatinib, mitomycin C, amasacrin)

[11]. Sự ức chế P-gp được nhắm đến để đối phó với kiểu hình MDR ở các bệnh nhân ung thư thông qua việc làm tăng sự tích lũy nội bào của các thuốc chất nền và vì vậy làm tăng độc tính tế bào của những thuốc này [15]. Bên cạnh ung thư, P-gp còn được quan tâm trong một số bệnh lý khác như Alzheimer, động kinh, mất trí liên quan HIV, viêm khớp dạng thấp, ban xuất huyết giảm tiểu cầu miễn dịch và lupus ban đỏ hệ thống [136].

Về mặt cấu trúc, P-gp là một protein xuyên màng với khối lượng 170 kDa được tạo thành bởi hai nửa đối xứng là N tận (N-terminal) và C tận (C-terminal) [150]. Mỗi nửa phân tử chứa sáu vùng xuyên màng (transmembrane domain - TMD), theo sau là một vùng gắn kết nucleotid (nucleotide-binding domain - NBD). Các vùng xuyên màng TMD 4, 5, 10 và 11 tạo thành một khoang gắn kết thuốc có thể tích lớn khoảng 6000 $Å^3$ ở bên trong, mở hướng về cả bào tương và nửa trong của lớp lipid kép giúp cho sự đi vào của thuốc và có thể chứa ít nhất hai chất cùng một lúc (**Hình 1.1**) [4]. Ngoài ra, tính linh hoạt về hình thể cũng là một yếu tố quan trọng cho khả năng gắn kết và chuyên chở nhiều chất nền đa dạng [198]. Cho đến nay, ABCB10 là protein chuyên chở ABC duy nhất của người được phân giải để sử dụng cho các phương pháp dựa vào cấu trúc, bên cạnh các cấu trúc tia X của một vài protein chuyên chở ABC khác có nguồn gốc từ các sinh vật chưa có nhân điển hình như vi khuẩn và có nhân điển hình như chuột [112], [169]. Để khắc phục những khó khăn do sự không sẵn có các cấu trúc tinh thể ba chiều (3D) ở độ phân giải cao của P-gp, các mô hình tương đồng của protein này đã được tạo ra sử dụng các cấu trúc liên quan đã được phân giải làm đĩa mẫu. Cho ví dụ, công trình nghiên cứu gần đây của Ambudkar và cộng sự đã tiết lộ nhiều vị trí gắn kết hoạt tính cho các chất nền và chất điều hòa, bao gồm một vị trí chính yếu nằm trong một túi lớn linh hoạt trong các vùng xuyên màng và các vị trí thứ cấp khác, từ sự kết hợp các phương pháp mô hình hóa tương đồng (homology modeling), docking phân tử (molecular docking), đột biến điểm định hướng (site-directed mutagenesis) với các thử nghiệm dựa trên tế bào và màng tế bào [28].



A



B

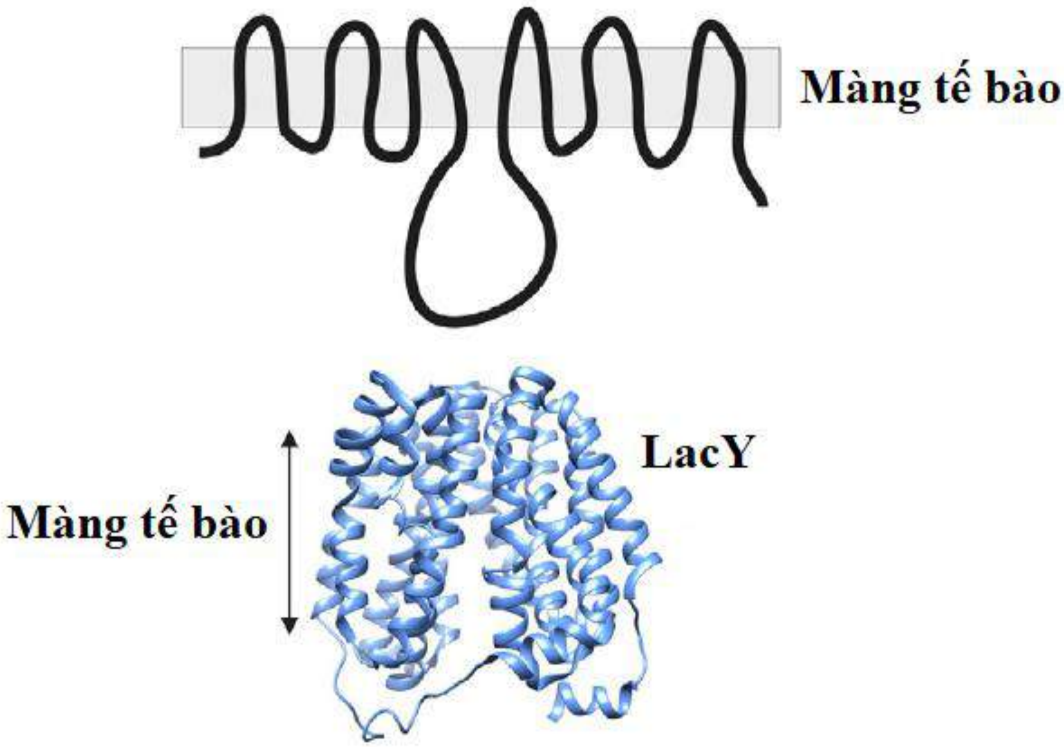
Hình 1.1. Cấu trúc của P-gp chuột: (A) mặt trước và (B) mặt sau. Các domain xuyên màng và domain gắn kết nucleotid lần lượt được đánh dấu từ TM 1-12 và NBD 1-2. Nửa N tận và nửa C tận lần lượt được tô màu vàng và xanh. Các TM 4-5 và TM 10-11 tạo thành các giao diện xoắn vào nhau giúp ổn định hình thể hướng vào trong. Các thanh ngang đại diện cho vị trí xấp xỉ của lớp lipid kép *"Nguồn: Aller S. G., Yu J., Ward A., et al., 2009"* [4].

1.1.2. Tổng quan về NorA

Các bơm ngược được tìm thấy ở hầu hết tất cả các loại vi khuẩn và được chia thành năm họ dựa vào thành phần, số lượng các vùng kéo dài xuyên màn, nguồn năng lượng và chất nền của chúng. Ngoài Họ đề kháng-sự hình thành nốt-phân chia (Resistance-Nodulation-Division (RND) family) chỉ được tìm thấy ở vi khuẩn gram âm; bốn họ còn lại là Liên họ trợ giúp chính (Major Facilitator Superfamily - MFS); Liên họ sử dụng ATP (ATP Binding Cassette (ABC) superfamily); Họ đề kháng đa thuốc nhỏ (Small Multidrug Resistance (SMR) family) và Họ bài xuất chất độc và đa thuốc (Multidrug and Toxic Compound Extrusion (MATE) family) phân bố rộng khắp ở cả vi khuẩn gram dương và gram âm. Trong đó, các thành viên thuộc họ MFS sử dụng nguồn năng lượng từ gradient proton để bài xuất các chất nền của chúng và là những bơm ngược quan trọng về mặt lâm sàng do gây ra hiện tượng MDR trên vi khuẩn gram dương, bao gồm NorA của *S. aureus* [85], [175].

Với gen biểu lộ quá mức ở 43 % số chủng, NorA là mục tiêu được nghiên cứu nhiều nhất trong số các bơm ngược giúp bài xuất các tác nhân kháng khuẩn của *S. aureus* [140]. Gen NorA được nhân bản từ nhiễm sắc thể của một chủng lâm sàng kháng fluoroquinolon và trình tự nucleotid của nó được dự đoán mã hóa cho một protein với 12 mảnh xuyên màng [211]. Do tính đặc hiệu chất nền rộng, protein này có khả năng bơm ngược nhiều chất nền khác nhau về cấu trúc, chẳng hạn như các fluoroquinolon thân nước (ví dụ ciprofloxacin), các kháng sinh, chất diệt khuẩn khác và thuốc nhuộm (ví dụ ethidium bromid) [98].

Cấu trúc tinh thể của một protein chuyên chở thuộc họ MFS là EmrD của *Escherichia coli* được công bố năm 2006 [210] và sự sắp xếp cấu trúc của nó có thể phản ánh cấu trúc chung của các bơm MFS, bao gồm 12 chuỗi xoắn xuyên màn tạo thành một cấu trúc gắn kết với 04 chuỗi H3, H6, H9 và H12 đối diện phía trong và các chuỗi còn lại tạo thành khoang nội tại chứa chủ yếu các thành phần kỵ nước (lõi kỵ nước) để chuyên chở các chất thân dầu. Các đặc tính cấu trúc tương tự của họ MFS cũng được mô tả gần đây (**Hình 1.2**) dựa trên một thành viên khác của họ là LacY, lactose permease của *E. coli* [161]. Việc thiếu thông tin cấu trúc của NorA và sự tương tác phân tử của protein này với các chất ức chế và chất nền đã gây nhiều trở ngại cho những nỗ lực nghiên cứu dựa trên cấu trúc. Tuy nhiên, NorA lại có sự tương đồng với các bơm protein khác của vi khuẩn, chẳng hạn như một số bơm ngược đặc hiệu tetracyclin của vi khuẩn gram âm (20-25 %) hay Bmr của *Bacillus subtilis* (44 %), … [119]. Do đó, các mô hình tương đồng của nó có thể được tạo ra để hỗ trợ cho thiết kế thuốc dựa vào mục tiêu tác động, tương tự như P-gp.



Hình 1.2. Giản đồ cấu trúc của họ các protein bơm ngược đề kháng đa thuốc MFS được tạo ra bằng phần mềm UCSF Chimera 1.10 từ lactose permease của *E. coli* (LacY) *"Nguồn: Schindler B. D., Kaatz G. W., 2016"* [161].

1.2. Các chất ức chế bơm ngược đề kháng đa thuốc

Sử dụng các chất ức chế bơm ngược (efflux pump inhibitor - EPI) như P-gp và NorA là một chiến lược được chấp nhận rộng rãi để khôi phục sự nhạy cảm hóa học trong điều trị kháng ung thư và kháng khuẩn [3], [78], [130], [178], [215].

1.2.1. Các chất ức chế P-gp

Các chất ức chế thế hệ I, II và III lần lượt được phát triển qua ba thập kỷ dựa trên việc sàng lọc các hợp chất có sẵn, tối ưu hóa phân tử mẹ và tổng hợp hóa học, kết hợp với các nghiên cứu tiền lâm sàng và lâm sàng [136].

Các chất ức chế thế hệ I

Nhóm này bao gồm các tác nhân dược lý ban đầu được phát triển cho các chỉ định khác nhưng sau đó được ghi nhận là chất nền kiêm chất ức chế P-gp. Thuốc chẹn kênh calci chống cao huyết áp verapamil, chất đối vận calmodulin trifluoperazin, chất ức chế miễn dịch cyclosporin, các thuốc tim mạch khác như quinidin, reserpin và yohimbin, kháng estrogen tamoxifen và toremifen, và kháng khối u vincristin đều thuộc phân loại này. Hầu hết các chất này cũng là chất nền của P-pg, chúng tương tác với protein, cạnh tranh với các chất nền khác và đóng vai trò như các chất ức chế cạnh tranh. Bởi vì tất cả các chất ức chế thế hệ I được xác định bằng cách này nên rõ ràng chúng không chọn lọc và có tiềm lực kém. Nồng độ ức chế P-gp của chúng đạt đến mức độ gây độc tính cao, do đó nhiều chất ức chế này đã thất bại trong các thử nghiệm lâm sàng [36].

Các chất ức chế thế hệ II

Các chất ức chế thế hệ I được biến đổi cấu trúc, cụ thể là tính quang hoạt (chirality) của chúng được thay đổi để làm vô hiệu tác dụng dược lý ban đầu và qua đó giảm độc tính của các chất mẹ. Dexverapamil là đồng phân R của verapamil không có hoạt tính trên tim, PSC 833 (valspodar) là một đồng đẳng của cyclosporin A không có tính chất ức chế miễn dịch, MS-209 và vài dẫn xuất hoặc đồng đẳng của các thuốc thế hệ I khác được xếp vào phân loại này. Các chất điều hòa này vẫn còn giữ tính chất chất nền của P-gp và cho thấy ái lực thấp với protein. Vì vậy, liều ức chế P-gp của chúng nằm ngoài giới hạn liều có thể dung nạp. Do sự tối ưu hóa tính quang hoạt, các chất ức chế thế hệ II này không còn là các chất nền quen thuộc của CYP450 3A4 trong chuyển hóa, giúp chúng cạnh tranh với các thuốc kháng ung thư là chất nền của P-gp được sử dụng đồng thời và có sự chuyển hóa cũng bị ảnh hưởng bởi cùng hệ thống. Kết quả gây ra các biến đổi dược động học đáng kể, ảnh hưởng đến các cơ chế chuyển hóa và thanh thải của các thuốc chất nền một cách không thể dự đoán, gây khó khăn trong việc điều chỉnh liều hóa trị cho bệnh nhân. Tất cả những vấn đề này làm cho các chất ức chế thuộc phân loại này không được sử dụng [37], [183].

Các chất ức chế thế hệ III

Ứng dụng QSAR cho các kỹ thuật sàng lọc hiệu năng cao (high-throughput screening - HTS) và các phương pháp hóa học kết hợp giúp tạo ra các chất có tác dụng mạnh hơn 10 lần so với thế hệ I và II. Các chất ức chế thế hệ III có tính đặc hiệu cao, không tương tác với hệ thống CYP450 3A4 và không đòi hỏi chỉnh liều hóa trị liệu. Trong nhóm này, một dẫn xuất của anthranilamid XR 9576 (tariquidar) là chất ức chế P-gp không được chuyên chở, được cho là ức chế ATPase thông qua tương tác với một vị trí gắn kết điều hòa phân biệt trên protein. Mặc dù có nhiều triển vọng nhưng việc sử dụng chất này vẫn còn bị trì hoãn bởi vì các báo cáo độc tính bất lợi trong các thử nghiệm pha III ở các trường hợp ung thư biểu mô phổi. Các chất khác được khám phá bởi chiến lược này bao gồm VX-710 (biricodar, một chất điều hòa loại cyclopropyldibenzosuberan được phát triển bởi Eli Lilly Inc.), GF 120918 (elacridar, một dẫn xuất acridonecarboxamid được phát triển bởi GlaxoSmithKline), OC 144-093, mitotan (NSC-38721), annamycin, R101933, ONT-093 và LY335979 (zosuquidar) [130].

Như vậy, hầu hết các chất ức chế P-gp thuộc ba thế hệ đầu tiên này đã không đạt được mục tiêu đặt ra ban đầu bởi vì một số tính chất bất lợi về tính tan, tính đặc hiệu, độ an toàn cũng như tương tác thuốc đã hạn chế việc sử dụng chúng trên lâm sàng [19], [160], [176]. Các chiến lược nghiên cứu mới mở ra triển vọng cho thế hệ thứ tư, như các sản phẩm tự nhiên (cam, bưởi, dâu, …) và bắt chước tự nhiên (flavonoid, alkaloid, coumarin, cannabinoid, taccalonolid, terpenoid, ginsenosid, polyen, lignan); peptidomimetic (các chất ức chế loại peptid, giống valspodar của thế hệ thứ hai); các chất hoạt động bề mặt (giống tween, cremophor EL, …) và lipid (liposom); và các phối tử kép (vừa kháng khối u vừa điều hòa MDR) [136].

1.2.2. Các chất ức chế NorA

Cho đến nay, một số chất ức chế bơm ngược NorA có khả năng khôi phục tính nhạy cảm của thuốc ở các chủng vi khuẩn đề kháng đã được xác định. Trong số đó có thể kể đến reserpin, verapamil, omeprazol, paroxetin, chlorpromazin, … là những thuốc không phải kháng sinh nhưng nồng độ cần thiết cho hoạt tính EPI của chúng là quá cao, dẫn đến cửa sổ trị liệu quá hẹp [7]. Trong quá trình tìm kiếm hơn một thập kỷ qua, các chất ức chế NorA cả tự nhiên (flavon, isoflavon, acylat glycosid, porphyrin phaeophorbid A, kaempferol rhamnosid, …) và tổng hợp (các dẫn xuất chalcon, N-cinnamoylphenalkylamid, indol, piperin, pyridin, fluoroquinolon, phenothiazin, thioxanthen, benzothiophen, macrolid, pyrrolo[1,2-a]quinoxalin, …) đều được báo cáo và hầu hết chúng có sở hữu nối đôi liên hợp [67], [215]. Tuy nhiên hiện tại vẫn chưa có chất ức chế bơm ngược vi khuẩn nào được cấp phép sử dụng để điều trị nhiễm trùng ở người và việc nghiên cứu vẫn tiếp tục [67].

1.2.3. Flavonoid và các thư viện hóa học sẵn có

Trong bối cảnh nghiên cứu hiện tại, một nhóm cấu trúc phổ biến trong tự nhiên là flavonoid, nổi lên như là các tác nhân có tiềm năng đảo ngược MDR qua trung gian bơm ngược với những ưu điểm như cho tác dụng kép (điều hòa P-gp và hoạt tính kháng khối u), an toàn và có thể được xếp vào phân loại không phải dược phẩm thuộc thế hệ thứ ba hoặc vào thế hệ thứ tư của các chất ức chế P-gp [11], [52], [136], [159], [173], [216]. Ngoài ra, các flavonoid cũng được báo cáo là có hoạt tính EPI chống lại NorA của *S. aureus* [67], [215]. Cấu trúc của các chất ức chế loại flavonoid từ tự nhiên đóng vai trò là điểm khởi đầu để thực hiện sự biến đổi hóa học, nghiên cứu các mối quan hệ cấu trúc - tác dụng (structure-activity relationship - SAR) hoặc mối quan hệ định lượng cấu trúc - tác dụng (quantitative structure-activity relationship - QSAR). Cho ví dụ, Holler và cộng sự đã công bố 5 chalcon có hoạt tính ức chế NorA trong tổng số 117 chất được thử nghiệm vào năm 2012, trong đó 2 dẫn xuất N,N-dimethylaminoethoxyphenyl có tiềm lực tương đương reserpin, một chất ức chế bơm đã biết được sử dụng để tham khảo trong các nghiên cứu [69]. Hay trong năm 2014, Ecker và cộng sự đã công bố một loạt các chalcon tổng hợp mới cùng với các kết quả đánh giá hoạt tính sinh học khá tốt của những chất này và chỉ ra tầm quan trọng của các nhóm cấu trúc cụ thể cần thiết cho hoạt tính ức chế P-gp dựa trên các phân tích mối quan hệ định lượng cấu trúc - tác dụng hai chiều và ba chiều (2D- và 3D-QSAR) [139].

Ngoài ra, sàng lọc các cơ sở dữ liệu có uy tín như Zinc, PubChem, ChemSpider, ChEMBL, NuBBE DB, ChemBank, eMolecules, DrugBank, Binding DB [53] cũng là một chiến lược hợp lý được sử dụng để tìm kiếm các chất đã biết có hoạt tính sinh học mong muốn. Trong một nghiên cứu gần đây, Barreca, Sabatini và cộng sự đã công bố ba thuốc không phải kháng sinh đã được phê duyệt là dasatinib, gefitinib và nicardipin có khả năng khôi phục hoạt tính kháng khuẩn của ciprofloxacin trên các chủng *S. aureus* biểu lộ quá mức bơm ngược đề kháng đa thuốc NorA, sử dụng kết hợp sàng lọc ảo dựa vào pharmacophore trên hai thư viện thuốc đã được phê duyệt từ Selleck và Prestwick, và đánh giá sinh học [7]. Hay xa hơn vào năm 2015, nhóm nghiên cứu chúng tôi đã thực hiện sàng lọc ảo trên cơ sở dữ liệu thuốc cổ truyền Trung Quốc (traditional Chinese medicine) để tìm kiếm các chất ức chế NorA mới [181].

1.3. Đề kháng kháng sinh

Việc sử dụng đại trà kháng sinh để kiểm soát các bệnh nhiễm khuẩn đã tạo điều kiện cho sự phát triển đề kháng với các trị liệu này. Theo thời gian, các chủng vi khuẩn kháng thuốc được chọn lọc qua bốn cơ chế chính là biến đổi mục tiêu tác động [6], bất hoạt thuốc bằng enzym [73], giảm hấp thu hoặc tăng cường bơm ngược [188] và thành lập màng sinh học [68]. Hệ quả là hiệu quả điều trị của kháng sinh bị giảm, việc điều trị bệnh trở nên khó khăn, tốn kém hoặc thậm chí thất bại.

Tụ cầu vàng *S. aureus* là nguyên nhân gây ra nhiều loại nhiễm trùng khác nhau và đóng vai trò quan trọng trong sự đề kháng kháng sinh. Sự phát triển đề kháng nhanh chóng với các kháng sinh thông thường đã dẫn đến hậu quả là 11 ngàn ca tử vong được ghi nhận do *S. aureus* đề kháng methicillin (MRSA) gây ra vào năm 2013 tại Mỹ, thậm chí vượt xa số người tử vong do HIV/AIDS là khoảng 8 ngàn ca [7]. Mặt khác, một kháng sinh tiêu chuẩn dùng để điều trị những trường hợp nhiễm MRSA phức tạp là vancomycin đang mất dần hiệu quả cũng bởi vì sự đề kháng lan rộng [7]. Trong danh sách vi khuẩn đề kháng cần ưu tiên cho nghiên cứu phát triển của WHO năm 2017, *S. aureus* kháng methicillin, trung gian và kháng vancomycin cũng được xác định thuộc nhóm 2 với mức độ ưu tiên cao do đã tăng đề kháng với các kháng sinh hiện có trong các bệnh phổ biến [204].

Trong bối cảnh đó, việc tìm kiếm các phương tiện hiệu quả để khắc phục các cơ chế đề kháng kháng sinh hiện tại được đặt ra một cách cấp thiết. Trong đó, đồng sử dụng một kháng sinh với một thuốc có thể khôi phục hoạt tính kháng khuẩn đầy đủ là một chiến lược đáng được xem xét. Với tầm quan trọng đã biết của các loại bơm ngược ở vi khuẩn trong các quá trình khử độc tế bào, chúng có thể là một mục tiêu nghiên cứu đáng giá nhằm tìm kiếm các chất mới giúp giải quyết hiện tượng đề kháng kháng sinh. Bằng cách bài xuất kháng sinh ra ngoài, bơm ngược làm giảm nồng độ thuốc nội bào, gây ra sự đề kháng kháng sinh của vi khuẩn và khi hoạt tính bơm ngược tăng thì giá trị nồng độ ức chế tối thiểu (MIC) của kháng sinh cũng tăng tương ứng. Việc sử dụng các chất ức chế bơm (EPI) có thể làm giảm MIC của kháng sinh chất nền bằng nhiều cách như khôi phục sự nhạy cảm với thuốc, giảm khả năng đề kháng thụ nhận bổ sung (ví dụ đột biến mục tiêu) và ức chế sự thành lập màng sinh học [7].

Có nhiều loại chất ức chế bơm ngược của vi khuẩn, bao gồm cả các chất ức chế bơm eukaryot như P-gp và MRP-1 dùng trong điều trị ung thư. Lợi thế của các chất này là hồ sơ dược động và độc tính của chúng đã được thiết lập [14]. Trong số đó có thể kể đến các chất ức chế P-gp thế hệ thứ ba với hiệu quả ức chế chống lại bơm ngược vi khuẩn là NorA của *S. aureus* (tariquidar và elacridar) và SmeDEF của *Stenotrophomonas maltophilia* (elacridar) [94].

Vì những lý do trên, đề tài đã lựa chọn hướng đi tìm kiếm các chất ức chế bơm ngược có hiệu quả trên vi khuẩn kháng thuốc *S. aureus* từ nguồn cấu trúc có sẵn là các dẫn xuất chalcon, thông qua việc nghiên cứu hoạt tính ức chế trên cả hai loại bơm của người và vi khuẩn là P-gp và NorA.

1.4. Các nghiên cứu trước có liên quan

Các nghiên cứu trước công bố cả mô hình phân loại chất ức chế và chất không ức chế P-gp (biến nhị phân) và mô hình dự đoán hoạt tính ức chế P-gp (biến liên tục). Kết quả thu được từ các công trình trước đó được trình bày tóm tắt trong **Bảng 4.1** và **Bảng 4.2**, cũng như được bàn luận và so sánh với nghiên cứu này trong **Mục 4.1.1** và **Mục 4.1.2**.

1.5. Các thuật toán học máy trong Clementine 12.0

Kỹ thuật học máy (machine learning) là nền tảng để khai phá dữ liệu cho nhiều mục đích khác nhau. Trong quá trình khám phá và phát triển thuốc, các công cụ học máy được ứng dụng ngày càng nhiều để dự đoán các tính chất dược lực học (chất ức chế, chất nền, chất đối vận, chất chủ vận, chất chẹn, chất hoạt hóa, độc tính) và dược động học (ADME) của các chất hóa học [44]. Trong Clementine, hai hạch Binary Classifier và Numeric Predictor lần lượt được sử dụng cho các mục đích phân loại (biến nhị phân) và dự đoán (biến liên tục) [29]. Ngoài ra, hạch Ensemble được đưa vào để kết hợp các dự đoán từ những mô hình đúng nhất, giúp tránh được những hạn chế của các mô hình đơn lẻ và đạt được một giá trị độ đúng tổng thể lớn hơn [29].

Hạch Binary Classifier cho phép ước tính tối đa mười mô hình học máy đơn lẻ là mạng nơron (Neural Network); C5.0; cây phân loại và hồi quy (Classification and Regression Tree - C&R Tree); cây thống kê hiệu quả, không thiên vị, nhanh (Quick, Unbiased, Efficient Statistical Tree - QUEST); máy dò tương tác tự động chi bình phương (Chi-square Automatic Interaction Detector - CHAID); hồi quy logistic (Logistic Regression); mặt nghiêng quyết định (Decision List); mạng Bayesian (Bayesian Network); phân tích phân biệt (Discriminant Analysis) và máy vector hỗ trợ (Support Vector Machine - SVM) [29]. Trong khi đó, hạch Numeric Predictor cho phép ước tính tối đa sáu mô hình học máy đơn lẻ là mạng nơron (Neural Network); cây phân loại và hồi quy (Classification and Regression Tree - C&R Tree); máy dò tương tác tự động chi bình phương (Chi-square Automatic Interaction Detector - CHAID); hồi quy (Regression); tuyến tính suy rộng (Generalized Linear) và máy vector hỗ trợ (Support Vector Machine - SVM) [29]. Nguyên tắc và ví dụ ứng dụng của các phương pháp này được mô tả chi tiết trong nhiều tài liệu [16], [29], [41], [71], [80], [104], [108], [109], [116], [117], [138], [202] và được tóm tắt một phần trong công trình số 4 đã được công bố vào năm 2016 của luận án (tham khảo **Danh mục các công trình đã công bố có liên quan**).

1.6. Các công cụ máy tính khác

1.6.1. Bản đồ nhận thức

Bản đồ nhận thức có thể được xây dựng bằng các phương pháp đo lường đa hướng (multidimensional scaling - MDS) và phân tích tương hợp (correspondence analysis - CA). MDS giúp tìm kiếm cấu trúc hoặc mô hình trong một tập hợp các đo lường khoảng cách giữa các đối tượng hoặc các trường hợp bằng cách chỉ định các quan sát vào những vị trí cụ thể trong một không gian nhận thức để làm cho các khoảng cách giữa các điểm trong không gian phù hợp với những khác biệt được cho trước càng chặt chẽ càng tốt. Trong kỹ thuật này, cần lưu ý hai đại lượng thống kê là: (i) stress: là thông số độ tốt của hit mà MDS cố gắng tối thiểu hóa, bao gồm căn bậc hai của các sai lệch bình phương chuẩn hóa giữa các khoảng cách liên điểm trong đồ thị MDS và các khoảng cách phẳng được dự đoán từ những khác biệt. Stress thay đổi giữa 0 và 1, với các giá trị gần 0 cho thấy một sự phù hợp tốt hơn; (ii) vòng lặp: mỗi vòng lặp là một sự di chuyển của tất cả các điểm trong đồ thị đến một giải pháp tốt hơn. CA giả định các biến định danh có thể mô tả các mối quan hệ giữa các phân loại của mỗi biến cũng như mối quan hệ giữa các biến trong một không gian ít chiều. Trong kỹ thuật này, thông số phương sai (inertia/variance) là phần trăm phương sai được giải thích bởi mỗi hướng. Thông số này thay đổi giữa 0 và 1, với các giá trị gần 1 cho thấy sự tương quan mạnh hơn giữa các trường hợp (các nhóm hoạt tính) và các biến (các dấu vân tay). Với cách thức rõ ràng và trực tiếp hơn so với phân tích thành phần chính (principal component analysis - PCA) [51], các phương pháp này được chọn để kiểm tra trực quan sự hỗn tạp phối tử giữa hai bơm ngược được quan tâm là P-gp và NorA.

1.6.2. Pharmacophore

Theo IUPAC, thuật ngữ "pharmacophore" được định nghĩa là một tập hợp các yếu tố không gian và điện tử cần thiết để đảm bảo cho các tương tác siêu phân tử tối ưu với một mục tiêu sinh học cụ thể và gây ra (hoặc ngăn chặn) đáp ứng sinh học của mục tiêu sinh học đó [199]. Nói cách khác, một pharmacophore bao gồm một tập hợp các yếu tố chung (các nhóm cho/nhận liên kết hydro, các vùng phân cực/kỵ nước) được tìm thấy ở một nhóm các hợp chất có thể tương tác bổ trợ với một tập hợp các yếu tố tương ứng ở vị trí gắn kết của mục tiêu sinh học [88]. Pharmacophore cũng cung cấp thông tin về mô hình sắp xếp trong không gian của các nhóm hóa học hoặc các acid amin chịu trách nhiệm cho việc gắn kết trong một phức hợp phối tử - protein [105]. Mô hình hóa pharmacophore được ứng dụng rộng rãi cho nhiều mục đích nghiên cứu khác nhau, chẳng hạn như sàng lọc ảo để làm giảm số lượng các ứng viên trong giai đoạn sớm của quá trình khám phá thuốc, tối ưu hóa các chất khởi nguồn cho một mục tiêu thuốc cụ thể, thiết kế thư viện các phân tử mới, dự đoán các tương tác tiềm năng dẫn đến các tác dụng không mong muốn [182]. Mặc dù là công cụ máy tính được sử dụng phổ biến trong thiết kế thuốc với nhiều triển vọng, một trong những hạn chế chính cần được xem xét trước khi sử dụng kỹ thuật này chính là tính đơn giản của các giả thuyết pharmacophore khiến cho chúng không thể giải thích hết được tất cả những vấn đề của các tương tác gắn kết giữa protein mục tiêu và phối tử [182].

Các mô hình pharmacophore có thể được xây dựng bằng các phương pháp dựa trên phối tử (đầu vào là một số phối tử) và các phương pháp dựa trên cấu trúc (đầu vào là cấu trúc của protein) [182]. Trong nghiên cứu này, các chất ức chế P-gp mạnh và các chất ức chế NorA nhưng không ức chế P-gp được chọn lọc để xây dựng các mô hình pharmacophore tương ứng.

1.6.3. Mô hình hóa tương đồng

Do không có sẵn các cấu trúc tinh thể ở độ phân giải cao của các protein màng như P-gp ở người, kỹ thuật mô hình hóa tương đồng được xem là một giải pháp khả thi để thu được thông tin cấu trúc của các protein này [148]. Trong nghiên cứu này, server I-TASSER (Iterative Threading ASSEmbly Refinement) là một hệ thống trực tuyến sẵn có việc dự đoán tự động cấu trúc ba chiều (3D) của protein và không thu phí [151], [217], được sử dụng để tạo ra các mô hình tương đồng (mô hình so sánh) của P-gp cho mục đích nghiên cứu docking. Phương pháp của I-TASSER dựa trên các thuật toán hiện đại [206], [218] được mô tả tóm tắt gồm ba giai đoạn: (i) xác định các protein đĩa có cấu trúc hoặc mô hình cấu trúc tương tự với trình tự truy vấn từ các cơ sở dữ liệu cấu trúc đã được phân giải; (ii) lắp ráp cấu trúc; (iii) lựa chọn mô hình và tinh chỉnh [151].

1.6.4. Docking

Docking phân tử là một trong số các phương pháp được sử dụng nhiều nhất trong thiết kế thuốc dựa vào cấu trúc bởi vì nó có khả năng dự đoán hình thể của các phối tử phân tử nhỏ trong vị trí gắn kết mục tiêu phù hợp với một độ đúng đáng kể. Sau khi các thuật toán đầu tiên được phát triển trong những năm 1980, docking phân tử đã trở thành một công cụ quan trọng trong quá trình khám phá thuốc, giúp nghiên cứu các mô hình gắn kết phối tử và các tương tác liên phân tử tương ứng có vai trò ổn định phức hợp phối tử-thụ thể, cũng như ước tính năng lượng tự do gắn kết và xếp hạng các chất dựa trên ái lực gắn kết của phức hợp phối tử-thụ thể [53]. Qua đó, docking có thể được sử dụng để thực hiện sàng lọc ảo các thư viện chất lớn, xếp hạng các kết quả và đề nghị các giả thuyết cấu trúc về cách thức các phối tử ức chế mục tiêu [113].

Sự thiết lập các cấu trúc đầu vào của docking cũng quan trọng như chính bản thân docking [113]. Sau đó, việc xác định các hình thể gắn kết khả thi nhất được thực hiện thông qua hai bước: (i) thử các hình thể của phối tử trong vị trí hoạt động của protein để khám phá một không gian hình thể lớn đại diện cho các kiểu gắn kết tiềm năng khác nhau; (ii) dự đoán năng lượng tương tác tương ứng với mỗi một hình thể gắn kết, sau đó xếp hạng các hình thể này nhờ một hàm tính điểm. Một cách lý tưởng, các thuật toán thử cần có khả năng mô phỏng mô hình gắn kết thực nghiệm và hàm tính điểm cần xếp hạng nó cao nhất trong số tất cả các hình thể được tạo ra. Các chương trình docking phân tử thực hiện các nhiệm vụ này thông qua một quá trình tuần hoàn cho đến khi hội tụ thành một giải pháp có năng lượng tối thiểu [53], [110]. Các phương pháp docking bao gồm [110]:

- Docking phối tử cứng nhắc và thụ thể cứng nhắc (các phần mềm DOCK phiên bản cũ, FLOG, FTDOCK).
- Docking phối tử linh động và thụ thể cứng nhắc (các phần mềm AutoDock 3, FlexX).
- Docking phối tử linh động và thụ thể linh động (các phần mềm GOLD, AutoDock 4, ICM, DOCK, FlexE).

Docking phân tử thụ thể linh động, đặc biệt là sự linh động xương sống (mạch chính) của các thụ thể là một thách thức cho các kỹ thuật docking có sẵn. Phương pháp dựa vào Local Move Monte Carlo gần đây được đưa vào như một giải pháp tiềm năng cho các vấn đề docking thụ thể linh động [110].

1.7. Thử nghiệm tác dụng ức chế bơm ngược trên các chủng vi khuẩn đề kháng

Ngoài dự đoán bằng máy tính, sự ức chế P-gp có thể được nghiên cứu trên *in vitro* bao gồm thử nghiệm độc tính tế bào (cytotoxicity assay), thử nghiệm tích lũy/bơm ngược (accumulation/efflux assay), thử nghiệm chuyên chở (transport assay), thử nghiệm ATPase (ATPase assay), đánh dấu ái lực quang học (P-gp photoaffinity labeling) và trên *in vivo* sử dụng chuột chuyển gen (transgenic/knock-out) hoặc đột biến [11], [189]. Các thử nghiệm này nhìn chung đều có hiệu năng hạn chế, quy trình thực hiện dài dòng, phức tạp (nuôi cấy tế bào, yêu cầu về phân tích, …), dễ bị ảnh hưởng bởi nhiều yếu tố (dòng tế bào sử dụng, mẻ nuôi cấy, tính chất chất nền/chất ức chế, sự hiện diện của vài con đường chuyên chở khác ngoài P-gp hay tính biến thiên sinh học liên quan đến thử nghiệm trên động vật, …) [11], [189], và không phù hợp với điều kiện hiện có tại Việt Nam. Như được đề cập trong phần mở đầu của luận án, sự tồn tại các chất ức chế chung của P-gp và NorA đã gợi ý cho đề tài thực hiện đánh giá tác dụng ức chế bơm ngược trên các chủng vi khuẩn đề kháng để thay thế.

Nguyên tắc của các thử nghiệm được sử dụng trong nghiên cứu này là nếu một chất có khả năng ức chế bơm ngược, nó có thể làm giảm sự đề kháng của chủng vi khuẩn MDR do hệ thống bơm ngược biểu lộ quá mức với các kháng sinh đã bị đề kháng hoặc làm cho nó trở nên nhạy cảm với kháng sinh như chủng tự nhiên [61]. Do đó, hiệu quả ức chế bơm ngược của chất thử nghiệm có thể được xác định thông qua thử nghiệm đánh giá khả năng làm giảm giá trị nồng độ ức chế tối thiểu (MIC) của kháng sinh chất nền trên các chủng vi khuẩn đề kháng bằng cách tăng biểu lộ bơm ngược khi có sự hiện diện của chất thử nghiệm đó ở một nồng độ cụ thể nhỏ hơn MIC của chính nó (kiểm tra bằng mẫu chứng không có kháng sinh) [67].

Với mục đích kiểm tra tác dụng ức chế bơm ngược NorA trên *S. aureus* của một số chalcon nội bộ sàng lọc được, nghiên cứu này tiến hành xác định và so sánh MIC của một kháng sinh bị đề kháng bởi protein chuyên chở này là ciprofloxacin [33], trên các chủng SA đề kháng bằng cơ chế bơm (chủng đột biến SA-1199B có biểu lộ quá mức NorA và các chủng SA phân lập lâm sàng). Các chất tự nhiên bao gồm chalcon cho thấy hoạt tính EPI trên SA ở hàm lượng ≤ 300 µg/mL khi phối hợp với các kháng sinh [85]. Trên cơ sở đó, các chalcon với lượng mẫu giới hạn được chọn thử nghiệm ở các hàm lượng 100, 50 và 20 µg/mL. Ngoài ra, chất ức chế nhiều bơm đã biết là phenyl-arginin-beta-naphthylamid (PaβN) [61], [85], [107] cũng được sử dụng ở hàm lượng 20 µg/mL để sàng lọc các chủng vi khuẩn đề kháng phân lập từ lâm sàng có biểu lộ hệ thống bơm ngược.

Thử nghiệm xác định MIC của kháng sinh được thực hiện bằng các phương pháp pha loãng (dilution methods) [106], [203]. Kết quả được thể hiện bằng nồng độ tối thiểu của chất thử (µg/mL hoặc mg/L) có khả năng ức chế sự mọc của vi khuẩn. Trong đó, chất thử được pha loãng thành một dãy nồng độ từ thấp tới cao theo cấp số nhân trong môi trường nuôi cấy. Mỗi nồng độ chất thử được cấy một lượng vi khuẩn nhất định và được nuôi ủ trong vòng 18 - 24 giờ. Nồng độ chất thử thấp nhất mà ức chế được sự phát triển của vi khuẩn (môi trường không đục hoặc vi khuẩn không mọc trên mặt thạch) được ghi nhận là giá trị MIC. Ngoài vi khuẩn, các phương pháp pha loãng cũng được sử dụng để thử tính nhạy cảm của tác nhân kháng vi sinh vật với nấm men và nấm sợi, dựa trên nhiều hướng dẫn và tiêu chuẩn khác nhau được chấp nhận như CLSI, EUCAST, … Một ứng dụng khác ngoài đánh giá MIC là ước tính hoạt tính diệt khuẩn hoặc diệt nấm thông qua việc xác định nồng độ diệt khuẩn tối thiểu (MBC) hoặc nồng độ diệt nấm tối thiểu (MFC). So với khuếch tán đĩa (disk diffusion), các phương pháp pha loãng linh hoạt hơn do môi trường chuẩn được sử dụng để thử nghiệm các sinh vật thường gặp (ví dụ staphylococci, enterococci, các vi khuẩn họ *Enterobacteriaceae* và *Pseudomonas aeruginosa*) có thể được bổ sung hoặc thay thế bằng môi trường khác để có phép thử chính xác cho các chủng khó hơn. Cách thực hiện các kỹ thuật này được mô tả chi tiết trong tài liệu [10], [76], với những ưu nhược điểm và ứng dụng riêng được tóm tắt như sau:

Phương pháp trong thạch

Chất thử được pha loãng trong thạch. Pha loãng trong thạch thường được khuyến cáo là phương pháp chuẩn cho các sinh vật khó như vi khuẩn kỵ khí và *Helicobacter*. Nó cho thấy mối tương quan tốt với Etest (gradient kháng vi sinh vật) hầu như cho thử nghiệm kháng khuẩn trên cả vi khuẩn Gram dương và Gram âm. Phương pháp này cũng được sử dụng cho các phối hợp các thuốc - tác nhân kháng nấm trên *Candida* sp., *Aspergillus*, *Fusarium* và nấm da.

Ưu điểm: Thích hợp với những chất khó tan trong nước; có thể tiến hành thử nghiệm đồng thời trên cùng một dãy nồng độ chất thử với nhiều chủng vi sinh vật (32 - 60 chủng có thể được cấy lên mỗi đĩa thạch nhờ sự hỗ trợ của thiết bị); dễ phát hiện tình trạng nhiễm vi sinh và không đồng nhất hơn so với phương pháp trong môi trường lỏng.

Nhược điểm: Để có kết quả chính xác đòi hỏi lượng vi khuẩn chấm lên mỗi bản thạch phải như nhau; tốn thời gian, công sức và không có lợi ích kinh tế khi thực hiện thử nghiệm trên nhiều loại vi sinh vật với nhiều loại chất thử; không phải luôn được xem là phương pháp thử tính nhạy cảm cho các tác nhân kháng vi sinh vật mới hơn như ceftarolin, daptomycin và doripenem.

Phương pháp trong môi trường lỏng

Pha loãng trong môi trường lỏng

Chất thử được pha loãng trong các ống nghiệm chứa môi trường lỏng có thể tích $\geq$ 1 mL (thường là 2 mL).

Ưu điểm: Đây là phương pháp hữu ích khi dùng nghiên cứu, thử nghiệm một chất thử với một loại vi sinh vật và dễ thực hiện.

Nhược điểm: Chỉ thích hợp với những chất dễ tan trong nước; không có lợi ích kinh tế khi thực hiện thử nghiệm trên nhiều loại vi sinh vật với một hoặc nhiều loại chất thử.

Vi pha loãng trong môi trường lỏng

Vi pha loãng trong môi trường lỏng hiện tại được xem là phương pháp tham chiếu quốc tế để xác định MIC. Trong phương pháp này, chất thử được pha loãng trong các giếng vi lượng chứa thể tích thường là 0,1 mL.

Ưu điểm: Đây là phương pháp tốt nhất khi thực hiện thử nghiệm trên nhiều loại vi sinh vật với nhiều loại chất thử, đơn giản, dễ thực hiện và chỉ cần lượng nhỏ chất thử.

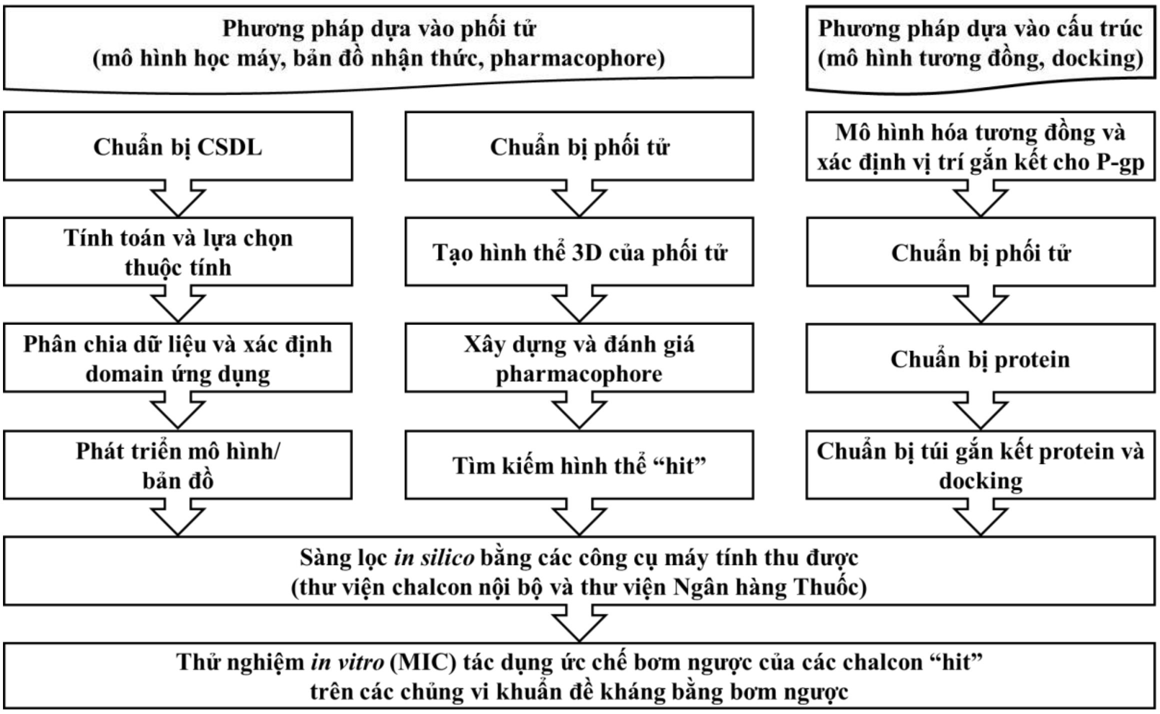
Nhược điểm: Khó đọc kết quả khi chất thử có màu; có thể cần thêm thiết bị, hóa chất như thuốc nhuộm (đo màu) trợ giúp việc đọc thử nghiệm và ghi kết quả để phân biệt sự tăng trưởng trong các giếng.

Với những ưu và nhược điểm nói trên, phương pháp vi pha loãng trong môi trường lỏng sử dụng đĩa 96 giếng được lựa chọn để thực hiện nghiên cứu do các chalcon thử nghiệm chỉ có sẵn với lượng nhỏ.

Trong nghiên cứu này, hai chiến lược khác nhau lần lượt được sử dụng nhằm tìm kiếm các phân tử có tác dụng ức chế bơm ngược: (i) kết hợp các phương pháp sàng lọc *in silico* bằng các công cụ máy tính được nêu trong **Mục 1.5** và **Mục 1.6** để tìm kiếm trong số các thuốc đã biết từ nguồn thư viện hóa học sẵn có; (ii) kết hợp sàng lọc *in silico* và thử nghiệm *in vitro* được nêu trong **Mục 1.7** để tìm kiếm trong số các chalcon từ nguồn nội bộ. Các chiến lược này sẽ tiếp tục được làm rõ trong các chương tiếp theo của luận án.

CHƯƠNG 2. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU

Nghiên cứu đã sử dụng kết hợp cả hai phương pháp dựa vào phối tử và dựa vào cấu trúc để xây dựng các mô hình *in silico* khác nhau và ứng dụng để sàng lọc tác dụng ức chế các bơm ngược P-gp và NorA cho hai cơ sở dữ liệu (chalcon nội bộ và Ngân hàng Thuốc). Sau đó, thử nghiệm *in vitro* trên vi khuẩn được thực hiện để chứng minh các kết quả dự đoán bằng máy tính, đồng thời khám phá ra các chalcon "lead" cho mục đích ức chế bơm ngược. Quy trình thực hiện được trình bày tóm tắt trong **Hình 2.1** và được làm rõ ở các nội dung tiếp theo trong chương này.



Hình 2.1. Quy trình nghiên cứu của đề tài

2.1. Đối tượng nghiên cứu

2.1.1. Các tập dữ liệu dùng cho việc dự đoán chất ức chế và chất không ức chế P-gp

Hai tập dữ liệu lớn và đa dạng về cấu trúc bao gồm 1935 và 1273 chất ức chế/không ức chế P-gp lần lượt được thu thập từ các nghiên cứu của Ecker và cộng sự [142] và Hou và cộng sự [23]. Trong đó, hoạt tính của các chất được thể hiện bằng một biến nhị phân (1 là chất ức chế; 0 là chất không ức chế). Sau đó, hai cơ sở dữ liệu này được hợp nhất lại và loại bỏ 1048 chất trùng bằng phần mềm MOE 2008.10 [111], dựa trên các tiêu chí chính là cấu trúc phân tử, tên và hoạt tính sinh học. Trong số 2160 chất còn lại, 22 cặp chất có tên giống nhau được phát hiện. Sau khi kiểm tra các cấu trúc hai chiều (2D) bằng cơ sở dữ liệu hóa học ChemSpider và so sánh các giá trị hoạt tính được thể hiện trong hai tập dữ liệu gốc, 26 chất đã được loại bỏ (5 chất của 5 cặp không phát hiện bất cứ khác biệt nào: chlorprothixen, lovastatin lacton, saquinavir, cortison, testosteron; 13 chất của 13 cặp có cấu trúc không đúng: laudanosin, NSC 633043, clonidin, hydramethylnon, paraquat, 1-phenylpiperazin, adiphenin, coralyn, memotin, becyclan, tamoxifen, mevinphos, eletriptan; 6 chất của 3 cặp có các giá trị hoạt tính mâu thuẫn: syrosingopin, lobelin, diethylstilbestrol; 2 chất của 1 cặp có cấu trúc không thể kiểm tra bằng ChemSpider hoặc bất cứ đâu: LY296097). Kết quả thu được một tập dữ liệu 2134 chất, bao gồm 1328 chất ức chế và 806 chất không ức chế cho mục đích xây dựng các mô hình phân loại.

Ngoài ra, 22 dẫn xuất chalcon cùng với các giá trị IC_{50} của chúng cũng được thu thập từ một nghiên cứu khác của Ecker và cộng sự [139] (**tập tin TLBS.xlsx, Sheet2** và **bản in Tài liệu bổ sung, TLBS2**) làm tập đánh giá ngoại để kiểm tra khả năng ngoại suy của các mô hình phân loại mà sẽ được sử dụng để sàng lọc các chalcon nội bộ có hoạt tính ức chế P-gp trong quá trình sàng lọc ảo sau đó. Dựa trên các giá trị ngưỡng được Polli và cộng sự đề xuất [147] và cũng đã được Cruciani và cộng sự sử dụng [18] để phân biệt chất ức chế và chất không ức chế P-gp, hai chất 18 và 19 với các giá trị IC_{50} của chúng > 100 μM được chỉ định là các chất không ức chế. Với IC_{50} = 20,4 μM > 15 μM, chất 24 là một chất ức chế P-gp yếu và vì vậy cũng được chỉ định là chất không ức chế. Tất cả 19 chất còn lại đều có IC_{50} < 15 μM và được xếp vào nhóm các chất ức chế.

2.1.2. Các tập dữ liệu dùng cho việc dự đoán hoạt tính ức chế P-gp

Theo Ecker và cộng sự [213], việc tạo ra một tập dữ liệu lớn và đa dạng về mặt hóa học của các chất ức chế P-gp mà các giá trị hoạt tính sinh học của chúng thu được từ các thử nghiệm khác nhau là có thể chấp nhận được. Trên cơ sở đó, bốn tập dữ liệu dựa trên cùng loại thử nghiệm ức chế bơm ngược daunorubicin/daunomycin ở các tế bào MDR CCRF vcr1000 được thu thập và kết hợp lại sử dụng phần mềm MOE 2008.10 [111] để mở rộng không gian hóa học của các mô hình QSAR, bao gồm: (i) 198 chất từ công trình của Ecker và cộng sự [213]; (ii) 159 chất được thử nghiệm từ PubChem BioAssay với mã số thử nghiệm (Assay Identification - AID) là 281137, bao gồm 141 chất có hoạt tính và 18 chất không được xác định; (iii) 303 chất được thử nghiệm từ PubChem BioAssay với AID là 781331, bao gồm 274 chất có hoạt tính và 29 chất không được xác định. Tất cả các chất này đều có trong cơ sở dữ liệu nguồn mở ChEMBL [13]; (iv) 22 dẫn xuất chalcon từ một nghiên cứu khác của Ecker và cộng sự [139]. Sau khi kiểm tra và loại bỏ hai cặp chất trùng (CID 13504844 và CID 9976755 trong cơ sở dữ liệu AID 281137 có cấu trúc lần lượt giống với CID 73357260 và CID 73346637 trong cơ sở dữ liệu AID 781331), tập dữ liệu sau cùng bao gồm cấu trúc hóa học và hoạt tính sinh học (IC_{50}) của 499 chất được chọn ra để xây dựng các mô hình dự đoán (**tập tin TLBS.xlsx, Sheet3** và **bản in Tài liệu bổ sung, TLBS3**). Thuật toán logarit âm của IC_{50} là pIC_{50} được sử dụng làm biến phụ thuộc. Trong số đó, một tập con 99 chất (20 %) được chọn ngẫu nhiên làm tập đánh giá ngoại để kiểm tra khả năng ngoại suy của các mô hình được xây dựng.

2.1.3. Các tập dữ liệu dùng cho việc xây dựng bản đồ nhận thức về sự hỗn tạp phối tử giữa P-gp và NorA

Khả năng ức chế P-gp và/hoặc NorA của 32 chất có cấu trúc đa dạng từ nghiên cứu của Carosati và cộng sự [17] và 13 chất có nguồn gốc tự nhiên từ nghiên cứu của Vishwakarma và cộng sự [77] được xác nhận bằng các thử nghiệm sinh học khác nhau. Ngoài ra, 9 ứng viên thuốc khác có nguồn gốc tự nhiên (reserpin [162], piperin [82], capsaicin [79]) và tổng hợp (verapamil [120], tariquidar [94], elacridar [60], biricodar, timcodar [114], SK-20 [86]) cũng được báo cáo là chất ức chế kép của hai loại bơm ngược trong một số nghiên cứu khác gần đây. Tất cả 54 chất này được sử dụng để xây dựng bản đồ nhận thức cho sự chồng phủ phối tử giữa P-gp và NorA (**tập tin TLBS.xlsx, Sheet4** và **bản in Tài liệu bổ sung, TLBS4**), bao gồm 19 chất ức chế P-gp nhưng không ức chế NorA (P); 7 chất ức chế NorA nhưng không ức chế

P-gp (A); 19 chất ức chế cả P-gp và NorA (D); và 9 chất không ức chế cả P-gp lẫn NorA (N).

2.1.4. Các tập dữ liệu dùng cho việc mô hình hóa pharmacophore của các chất ức chế P-gp/NorA

Trong nghiên cứu của Carosati và cộng sự [17], các chất 7 (aripiprazol), 8 (ebastin) biểu lộ khả năng ức chế 99 % sự bơm ngược rhodamin 123 qua trung gian P-gp ở các tế bào T lymphoma L5178 MDR1 của chuột; và các chất 22, 23 và 32 biểu lộ khả năng ức chế trên 80 % sự bơm ngược ethidium bromid qua trung gian NorA ở các tế bào SA-1199B. Hai chất 7, 8 cùng với hai chất ức chế P-gp mạnh khác là tariquidar và elacridar [42] được sử dụng để mô hình hóa pharmacophore của chất ức chế P-gp mạnh, trong khi ba chất 22, 23 và 32 được sử dụng để mô hình hóa pharmacophore của chất ức chế NorA nhưng không ức chế P-gp.

2.1.5. Các tập dữ liệu dùng trong sàng lọc ảo

Trong nghiên cứu này, một tập dữ liệu bao gồm 95 chalcon nội bộ (**Phụ lục 1**) và một tập dữ liệu khác bao gồm 6874 chất thuộc sáu nhóm khác nhau là "đã được phê duyệt/approved", "đang thử nghiệm/experimental", "đang nghiên cứu/investigational", "dinh dưỡng/nutraceutical", "đã bị thu hồi/withdrawn" và "bị cấm/illicit" tải về từ Ngân hàng Thuốc (Drug Bank) [84], [91], [200], [201] (**tập tin TLBS.xlsx, Sheet5** và **bản in Tài liệu bổ sung, TLBS5**) được sử dụng cho sàng lọc *in silico* nhằm tìm kiếm các chất ức chế bơm ngược mới cũng như khai thác các thuốc có sẵn, giúp khắc phục hiện tượng MDR ở các khối u và vi khuẩn.

2.1.6. Tập dữ liệu dùng để thử nghiệm khả năng làm giảm hiện tượng đề kháng kháng sinh ciprofloxacin qua trung gian NorA trên *S. aureus*

Một tập con các chalcon nội bộ thu được sau quá trình sàng lọc ảo (thuộc tập 95 chalcon được nêu trong **Mục 2.1.5**) với mẫu thử nghiệm sẵn có tại Bộ môn Hóa Dược, Khoa Dược, Đại học Y Dược Thành phố Hồ Chí Minh được sử dụng để đánh giá khả năng ức chế bơm ngược NorA, qua đó làm giảm giá trị nồng độ ức chế tối thiểu (MIC) của ciprofloxacin khi phối hợp trên chủng *S. aureus* SA-1199B đề kháng ciprofloxacin (biểu lộ quá mức NorA) và một số chủng SA phân lập từ lâm sàng.

2.2. Phương pháp nghiên cứu *in silico*

Đề tài sử dụng các phần mềm máy tính bao gồm: ChemBioDrawUltra 12.0, MOE 2008.10, PaDEL-Descriptor 2.21, RapidMiner 5.3.008, Weka 3.7.9, Applicability domain using standardization approach, Clementine 12.0, XternalValidationPlus, SPSS 20.0, server tự động I-TASSER và FlexX/LeadIT 2.0.2. Các phần mềm được liệt kê không có yêu cầu đặc biệt về cấu hình máy tính. Nghiên cứu *in silico* trong đề tài này được tiến hành trên laptop Dell Inspiron 3421, CPU Intel Core i3 1.90 GHz, Ram 4 GB, card màn hình VGA, hệ điều hành Windows 7 Ultimate 32-bit; với cách thức thực hiện như sau:

2.2.1. Tính toán và lựa chọn thông số mô tả

Các cấu trúc hai chiều (2D) được chuẩn bị trong phần mềm ChemBioDrawUltra 12.0 [21] nếu không có sẵn và sau đó được tối thiểu hóa năng lượng trong MOE [111] trước khi tính toán thông số mô tả. Các thông số mô tả giúp biến đổi các chất hóa học thành các vector mô tả trên máy tính và có tầm quan trọng đặc biệt trong việc dự đoán các tương tác phối tử - protein [214]. 184 thông số hai chiều (2D) mô tả các tính chất hóa lý, các diện tích bề mặt được chia nhỏ, số lượng các nguyên tử và liên kết, các chỉ số liên kết Kier&Hall và hình dạng Kappa, các ma trận gần kề và cách xa, các yếu tố pharmacophore và điện tích từng phần; và 1444 thông số mô tả một chiều (1D), hai chiều (2D) đại diện cho 63 loại tính chất phân tử khác nhau được tính toán cho các tập dữ liệu dùng để xây dựng các mô hình phân loại, dự đoán và bản đồ nhận thức, sử dụng lần lượt các phần mềm MOE [111] và PaDEL-Descriptor 2.21 [209]. Ngoài ra, cho mục đích phân loại và lập bản đồ nhận thức, 166 dấu vân tay MACCS (Molecular ACCess System), 881 dấu vân tay Pubchem và 307 dấu vân tay dưới cấu trúc cũng được tính toán, sử dụng phần mềm PaDEL-Descriptor 2.21 [209].

Việc lựa chọn thông số nhằm mục đích loại bỏ các thông số dư thừa hoặc không có liên quan, giúp cải thiện chất lượng mô hình và giảm bớt tài nguyên máy tính [41]. Đầu tiên, các chất với đầy đủ thông số mô tả được lọc ra bằng toán tử "Filter Examples", sử dụng phần mềm RapidMiner 5.3.008 [145]. Tiếp theo, việc loại bỏ các thông số vô nghĩa và/hoặc có tương quan mạnh với nhau (> 0,95) bằng các toán tử

"Remove Useless Attributes", "Remove Correlated Attributes" và tối ưu hóa sự lựa chọn dựa trên thuật toán di truyền bằng toán tử "Optimize Selection (Evolutionary)" cũng được thực hiện trong phần mềm RapidMiner [145]. Sau cùng, phương pháp BestFirst trong phần mềm Weka 3.7.9 [66] được sử dụng để lựa chọn các thông số mô tả, kết hợp đánh giá chéo 10 lần. Phương pháp này thực hiện tìm kiếm không gian của các tập con thuộc tính dựa trên thuật toán "greedy hill-climbing augmented" với khả năng học trở lại (backtracking). Tất cả các thông số trong quá trình lựa chọn biến đều được thiết lập như mặc định.

2.2.2. Phân chia dữ liệu thành các tập huấn luyện và tập đánh giá nội

Cơ sở dữ liệu dùng cho việc phân loại chất ức chế và chất không ức chế P-gp được phân chia theo tỷ lệ 4:1, trong khi cơ sở dữ liệu dùng cho việc dự đoán hoạt tính ức chế P-gp (400 chất còn lại) được phân chia theo tỷ lệ 3:1, thành tập huấn luyện và tập đánh giá nội cho mục đích đánh giá nội, sử dụng hai công cụ là "Rand" và "Diverse Subset" trong MOE [111]. Hàm Rand được sử dụng để phân chia dữ liệu một cách ngẫu nhiên, trong đó mỗi chất được gán một số ngẫu nhiên bất kỳ từ 0 đến 1. Ngược lại, ứng dụng Diverse Subset xếp hạng các chất trong một tập dữ liệu dựa trên khoảng cách giữa chúng với nhau, tạo ra một tập con bao gồm các chất ở xa các chất khác nhất. Trong trường hợp này, khoảng cách giữa hai chất được tính toán sử dụng các thông số mô tả/dấu vân tay có liên quan nhất, từ đó xác định được các chất ở xa nhất.

2.2.3. Xác định phạm vi khả năng ứng dụng

Việc xác định phạm vi khả năng ứng dụng (applicability domain - AD) là cần thiết để dự đoán đúng một chất bằng một mô hình QSAR. Thuật ngữ này được diễn giải theo nhiều cách khác nhau [74], [118], [127] nhưng có thể được hiểu một cách đơn giản là "đáp ứng và không gian cấu trúc hóa học trong đó mô hình QSAR thực hiện dự đoán với một độ tin cậy nhất định" [154]. Một số phương pháp có sẵn để xác định phạm vi khả năng ứng dụng, chẳng hạn như các phạm vi trong không gian thông số mô tả; các phương pháp hình học; các phương pháp dựa trên khoảng cách; phân bố mật độ xác suất và phạm vi của biến phụ thuộc và nhiễu [74], [118], [153]. Gần đây, Roy và cộng sự đã đề xuất một phương pháp mới và đơn giản để xác định các chất lạ (X-outlier) trong trường hợp tập huấn luyện và các chất nằm ngoài phạm vi khả năng ứng dụng (outside AD) trong trường hợp tập đánh giá [154]. Nguyên tắc của phương pháp này dựa trên lý thuyết của phương pháp chuẩn hóa, xem số trung bình (mean) ± 3*độ lệch chuẩn (standard deviation - SD) là vùng của hầu hết các chất trong tập huấn luyện (99,7 %) và phần còn lại là vùng của các chất khác biệt. Theo phương pháp này, một "X-outlier" trong tập huấn luyện hoặc một "outside AD" trong tập đánh giá có thể được xác định bằng cách tính toán thông số chuẩn hóa $S_{i(k)}$, giá trị $S_{i(k)}$ tối đa ($[S_i]_{max(k)}$), giá trị $S_{i(k)}$ tối thiểu ($[S_i]_{min(k)}$) (nếu cần), giá trị $S_{new(k)}$ (nếu cần) và so sánh chúng với giá trị ngưỡng bằng 3. Phần mềm "Applicability domain using standardization approach" do Roy và cộng sự phát triển [154] được sử dụng để tiến hành xác định phạm vi khả năng ứng dụng trong nghiên cứu này.

2.2.4. Các phương pháp học máy

Nghiên cứu này sử dụng phần mềm Clementine 12.0 [29], với hai hạch Binary Classifier và Ensemble được dùng cho mục đích phân loại dựa trên phối tử chất ức chế và không ức chế P-gp, và hai hạch Numeric Predictor và Ensemble được dùng cho mục đích dự đoán hoạt tính ức chế bơm ngược này. Trong đó, các mô hình được tạo ra từ hạch Binary Classifier hoặc hạch Numeric Predictor sẽ được gộp lại thành một mô hình kết hợp bằng hạch Ensemble. Các điều kiện mặc định được chọn của từng thuật toán trong hạch được mô tả tóm tắt như sau:

Mạng nơron

Với phương pháp Quick mặc định, một mạng nơron đơn giản được huấn luyện sử dụng thuật toán lan truyền về phía sau (back-propagation), dựa trên quy tắc delta tổng quát [158]. Mạng này bao gồm một lớp đầu vào với n_i = 24 nơron (tương ứng với 24 thuộc tính phân tử có liên quan nhất) trong hạch Binary Classifier và n_i = 34 nơron (tương ứng với 34 thuộc tính phân tử có liên quan nhất) trong hạch Numeric Predictor; một lớp đầu ra với n_o = 1 nơron (tương ứng với hoạt tính sinh học) và một lớp ẩn với n_h = max (3; (n_i + n_o)/20) = 3 nơron trong cả hai hạch. Thông số mặc định "persistence" được sử dụng để xác định thời điểm ngừng huấn luyện mạng, trong đó mạng sẽ tiếp tục được huấn luyện cho đến hết số chu kỳ được quy định là 250 chu kỳ, mặc dù có thể không có sự cải thiện nào giữa các chu kỳ.

C5.0

Hệ thống xây dựng bộ phân loại dưới dạng cây quyết định, sử dụng mô hình đơn giản thiên về độ đúng. Cây được tạo ra có độ sâu bằng 11, nghĩa là chỉ 11 trong số 24 thuộc tính phân tử được sử dụng cho việc phân loại.

Cây phân loại và hồi quy (C&R Tree)

Các quy tắc ngừng xây dựng cây được sử dụng bao gồm: Số mức dưới gốc (độ sâu tối đa của cây) = 5; số lượng đại diện tối đa = 5; số lượng chất tối thiểu trong nhánh mẹ = 2 % và số lượng chất tối thiểu trong nhánh con = 1 %. Sự phân chia hạch tốt nhất tạo ra sự giảm độ không tinh khiết nhỏ hơn sự thay đổi độ không tinh khiết tối thiểu (0,0001) cũng được sử dụng để ngăn một hạch phân chia. Các điều kiện này là giống nhau giữa hai hạch Binary Classifier và Numeric Predictor.

Cây thống kê hiệu quả, không thiên vị, nhanh (QUEST)

Các quy tắc ngừng xây dựng cây được sử dụng bao gồm: Số mức dưới gốc (độ sâu tối đa của cây) = 5; số lượng đại diện tối đa = 5; số lượng chất tối thiểu trong nhánh mẹ = 2 % và số lượng chất tối thiểu trong nhánh con = 1 %. Sự phân chia hạch tốt nhất tạo ra một giá trị p lớn hơn giá trị phân chia α (0,05) cũng được sử dụng để ngăn một hạch phân chia.

Máy dò tương tác tự động chi bình phương (CHAID)

Các quy tắc ngừng xây dựng cây được sử dụng bao gồm: Số mức dưới gốc (độ sâu tối đa của cây) = 5; số lượng chất tối thiểu trong nhánh mẹ = 2 %; số lượng chất tối thiểu trong nhánh con = 1 %; hệ số ε cho sự hội tụ = 0,001 và số vòng lặp tối đa = 100. Sự phân chia hạch tốt nhất tạo ra một giá trị p lớn hơn giá trị phân chia α (0,05) cũng được sử dụng để ngăn một hạch phân chia. Các điều kiện này là giống nhau giữa hai hạch Binary Classifier và Numeric Predictor.

Hồi quy logistic

Các thiết lập bao gồm quy trình đa thức, phương pháp Enter và hệ số chịu đựng dị biệt = $1{,}0*10^{-8}$ được sử dụng để tiến hành mô hình hóa.

Mặt nghiêng quyết định

Việc tìm kiếm được thực hiện theo chiều đi lên; số lượng phân đoạn tối đa, kích thước phân đoạn tối thiểu (theo phần trăm) và kích thước phân đoạn tối thiểu (theo giá trị tuyệt đối) được thiết lập lần lượt là 5, 5 % và 50. Các quy tắc phân đoạn bao gồm số lượng thuộc tính tối đa = 5 và khoảng tin cậy cho mỗi điều kiện mới = 95 % và cho phép sử dụng lại các thuộc tính này.

Mạng Bayesian

Một mô hình mạng Bayesian đơn giản được tạo ra cho mục đích phân loại, sử dụng phương pháp Tree Augmented Naive (TAN) Bayes giúp hoàn thiện phương pháp Naive Bayes bằng cách cho phép mỗi biến độc lập phụ thuộc vào một biến độc lập khác ngoài biến mục tiêu. Ngoài ra, khả năng tối đa được chọn làm phương pháp học thông số.

Phân tích phân biệt

Phương pháp Enter, các xác suất bằng nhau cho tất cả các nhóm và ma trận hiệp phương sai trong nội bộ các nhóm được sử dụng để phát triển mô hình.

Máy vector hỗ trợ (SVM)

Loại hàm kernel RBF được sử dụng để ánh xạ dữ liệu với điều kiện ngừng lại = $1{,}0*10^{-3}$. Các thiết lập mặc định khác bao gồm thông số quy tắc C = 10; độ chính xác hồi quy $\varepsilon = 0{,}1$ và gamma RBF (γ_{RBF}) = 0,1. Các điều kiện này là giống nhau giữa hai hạch Binary Classifier và Numeric Predictor.

Hồi quy

Phương pháp Enter được chọn mặc định để đưa vào, loại bỏ biến và bao gồm hằng số trong phương trình.

Tuyến tính suy rộng (Generalized Linear)

Loại mô hình chỉ những ảnh hưởng chính được áp dụng và bao gồm hệ số chặn (intercept) trong mô hình.

2.2.5. Đánh giá mô hình học máy

2.2.5.1. Đánh giá mô hình phân loại

Các mô hình phân loại được đánh giá bằng các thông số được tính toán dựa trên bốn đại lượng của ma trận nhầm lẫn là số chất dương tính thật (true positives - TP), số chất dương tính giả (false positives - FP), số chất âm tính thật (true negatives - TN) và số chất âm tính giả (false negatives - FN). Các thông số này được mô tả ngắn gọn như sau [41]:

Độ đúng tổng thể: Là tỷ lệ của các kết quả dự đoán đúng (cả dương tính thật và âm tính thật) trong dân số tất cả các chất.

$$\text{Độ đúng tổng thể} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Độ nhạy (độ đúng cho dự đoán dương tính): Là tỷ lệ các kết quả dự đoán dương tính thật trong tất cả các chất có hoạt tính.

$$\text{Độ nhạy} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Độ đặc hiệu (độ đúng cho dự đoán âm tính): Là tỷ lệ các kết quả dự đoán âm tính thật trong tất cả các chất không có hoạt tính.

$$\text{Độ đặc hiệu} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Độ chính xác (giá trị dự đoán dương): Là tỷ lệ các kết quả dự đoán dương tính thật trong tất cả các chất dương tính do bộ phân loại xác định.

$$\text{Độ chính xác} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Giá trị dự đoán âm: Là tỷ lệ các kết quả dự đoán âm tính thật trong tất cả các chất âm tính do bộ phân loại xác định.

$$\text{Giá trị dự đoán âm} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

Hệ số tương quan Matthews (Matthews Correlation Coefficient - MCC)

MCC là một dạng của hệ số tương quan Pearson và có thể được sử dụng để đánh giá sự dự đoán cân bằng của các mô hình phân loại. Thông số này xem xét bốn đại lượng của ma trận nhầm lẫn được đề cập ở trên và có giá trị thay đổi từ +1 đến -1 (+1:

Sự tương quan cùng chiều hoàn toàn hay dự đoán hoàn hảo; 0: Không có tương quan hoặc dự đoán ngẫu nhiên; -1: Sự tương quan ngược chiều hoàn toàn hay dự đoán tệ nhất có thể).

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN} - \mathrm{FP}\times\mathrm{FN}}{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}$$

G-mean

Thông số này xem xét cả độ nhạy và độ đặc hiệu và cũng được sử dụng để đo lường sự dự đoán cân bằng như MCC.

$$\text{G-mean} = \sqrt{\text{độ nhạy}\times\text{độ đặc hiệu}}$$

Chỉ số Youden's

Thông số này do Youden [212] đề xuất và góp phần xác định các mô hình tốt nhất.

$$\text{Chỉ số Youden's} = \text{độ nhạy} + \text{độ đặc hiệu} - 1$$

Điểm số độ tốt của hit (GH)

Thông số này xem xét cả kết quả (tỷ lệ các cấu trúc hoạt tính được dự đoán trúng) và phần trăm các chất có hoạt tính thu được từ cơ sở dữ liệu [65]. Mô hình được xem là tốt nhất nếu có các điểm số GH của tất cả các lớp gần với 1 [179].

$$\text{Điểm số GH cho các chất có hoạt tính} = \frac{\mathrm{TP}\times((\mathrm{TP}+\mathrm{FN})+(\mathrm{TP}+\mathrm{FP}))}{2\times(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TP}+\mathrm{FP})}$$

$$\text{Điểm số GH cho các chất không có hoạt tính} = \frac{\mathrm{TN}\times((\mathrm{FP}+\mathrm{TN})+(\mathrm{TN}+\mathrm{FN}))}{2\times(\mathrm{FP}+\mathrm{TN})\times(\mathrm{TN}+\mathrm{FN})}$$

Trong nghiên cứu này, phương pháp đánh giá chéo k lần với k = 10 và phương pháp ngẫu nhiên hóa biến phụ thuộc của tập huấn luyện y-randomization cũng được áp dụng cho mục đích đánh giá nội các mô hình [64]. Trong quá trình đánh giá chéo 10 lần [129], tập huấn luyện gốc được phân chia ngẫu nhiên thành 10 tập con và mỗi tập con lần lượt đóng vai trò là tập đánh giá cho mô hình được huấn luyện bởi 9 tập con còn lại. Trong quá trình ngẫu nhiên hóa biến phụ thuộc [156], các giá trị của biến phụ thuộc bị xáo trộn ngẫu nhiên 10 lần và các mô hình phân loại được tạo ra từ mỗi tập huấn luyện được ngẫu nhiên hóa biến y. Các quy trình đánh giá này nhằm mục đích ước tính sai số phân loại và phát hiện hiện tượng học quá (overfitting).

2.2.5.2. Đánh giá mô hình dự đoán biến liên tục

Các thông số thống kê và điều kiện đánh giá được sử dụng trong nghiên cứu này để đo lường khả năng dự đoán của các mô hình QSAR bao gồm:

Đánh giá nội

Chất lượng nội tại của các mô hình được đánh giá bằng hệ số tương quan bình phương R^2 và hệ số tương quan bình phương đánh giá chéo Q^2. Đánh giá chéo bỏ một (leave-one-out - LOO) là một trường hợp đặc biệt của đánh giá chéo k lần với k bằng số lượng các chất trong tập dữ liệu [12]. Trong trường hợp đánh giá chéo LOO, thông số Q^2 được tính theo công thức sau [156]:

$$Q^2_{LOO} = 1 - \frac{\sum_{i=1}^{n_{TR}}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{TR}}(y_i - \bar{y})^2} \quad (1)$$

Trong phương trình (1), y_i và $\hat{y}_i$ lần lượt là các giá trị hoạt tính được quan sát và được dự đoán, $\bar{y}$ là giá trị trung bình của y_i trong tập huấn luyện. $R^2 \geq 0{,}7$; $Q^2_{LOO} \geq 0{,}6$ và $|R^2 - Q^2_{LOO}| \leq 0{,}1$ là các điều kiện cần thiết để mô hình được chấp nhận [26], [27], [63], [64], [184], [185].

Ngoài ra, phương pháp ngẫu nhiên hóa biến phụ thuộc y-randomization cũng được áp dụng trên tập huấn luyện cho mục đích đánh giá nội. Trong quá trình này, các giá trị của biến phụ thuộc bị xáo trộn ngẫu nhiên 10 lần và các mô hình mới được tạo ra từ mỗi tập huấn luyện được ngẫu nhiên hóa biến y. Thông số R_p^2 được Roy và cộng sự [156] đề xuất để đảm bảo các mô hình không được phát triển một cách may rủi và được tính toán theo công thức sau:

$$R_p^2 = R^2\sqrt{R^2 - R_r^2}) \quad (2)$$

Trong phương trình (2), R^2 và R_r^2 lần lượt là hệ số tương quan bình phương của mô hình không được ngẫu nhiên hóa và hệ số tương quan trung bình bình phương của các mô hình được ngẫu nhiên hóa. Đối với một mô hình QSAR dự đoán, giá trị R_p^2 nên lớn hơn 0,5 [156].

Đánh giá ngoại

Các thông số phổ biến là Q^2_{F1} [168]; Q^2_{F2} [163]; Q^2_{F3} [30], [31]; r_m^2; $\overline{r_m^2}$; Δr_m^2 [123], [124], [157]; và hệ số tương quan phù hợp (concordance correlation coefficient

- CCC) [100] được áp dụng để đánh giá ngoại các mô hình khi thực hiện dự đoán cho các chất không liên quan đến sự phát triển mô hình. Ngoài ra, sai số tuyệt đối trung bình (mean absolute error - MAE) cũng được sử dụng để đánh giá khả năng dự đoán trên tập đánh giá ngoại [31]. Các thông số này được tính toán bằng các công thức sau:

$Q_{F1}^2 = 1 - \frac{\sum_{i=1}^{n_{EXT}}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{EXT}}(y_i - \bar{y}_{TR})^2}$ (3)

$Q_{F2}^2 = 1 - \frac{\sum_{i=1}^{n_{EXT}}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{EXT}}(y_i - \bar{y}_{EXT})^2}$ (4)

$Q_{F3}^2 = 1 - \frac{\left[\sum_{i=1}^{n_{EXT}}(y_i - \hat{y}_i)^2\right]/n_{EXT}}{\left[\sum_{i=1}^{n_{TR}}(y_i - \bar{y}_{TR})^2\right]/n_{TR}}$ (5)

$r_m^2 = r^2(1 - \sqrt{r^2 - r_0^2})$ (6)

$r'^2_m = r'^2(1 - \sqrt{r'^2 - r'^2_0})$ (7)

$\overline{r_m^2} = \frac{r_m^2 + r'^2_m}{2}$ (8)

$\Delta r_m^2 = |r_m^2 - r'^2_m|$ (9)

$CCC = \frac{2\sum_{i=1}^{n_{EXT}}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sum_{i=1}^{n_{EXT}}(y_i - \bar{y})^2 + \sum_{i=1}^{n_{EXT}}(\hat{y}_i - \bar{\hat{y}})^2 + n_{EXT}(\bar{y} - \bar{\hat{y}})^2}$ (10)

$MAE = \frac{1}{n} \times \sum_{i=1}^{n_{EXT}}|y_i - \hat{y}_i|$ (11)

Trong các phương trình (3), (4), (5), (10) và (11), y_i và $\hat{y}_i$ lần lượt là các giá trị hoạt tính được quan sát và được dự đoán, trong khi $\bar{y}$ và $\bar{\hat{y}}$ lần lượt là các giá trị trung bình của y_i và $\hat{y}_i$. Trong các phương trình (6) và (7), r^2 và r_0^2 lần lượt là các hệ số xác định trong hàm hồi quy khi có và không có hệ số tự do trong trường hợp giá trị thực nghiệm được biểu diễn trên trục y và giá trị dự đoán được biểu diễn trên trục x, trong khi r'^2 và r'^2_0 lần lượt là các hệ số xác định tương tự trong trường hợp ngược lại. Các điều kiện đánh giá chặt chẽ bao gồm $Q_{Fn}^2 \geq 0{,}7$; $\overline{r_m^2} \geq 0{,}65$; CCC ≥ 0,85 [26], [27]; $r_m^2 \geq 0{,}5$ [156]; và $\Delta r_m^2 \leq 0{,}2$ [155] được chọn để xác nhận khả năng dự đoán ngoại của các mô hình tốt. Còn với các điều kiện dựa vào MAE được Roy và cộng sự đề xuất gần đây [152], các dự đoán được xem là tốt khi MAE ≤ 0,1*độ rộng của các giá trị hoạt tính được quan sát trên tập huấn luyện ($(y_{i(\max)} - y_{i(min)})_{train}$) và MAE + 3σ ≤ 0,2*$(y_{i(\max)} - y_{i(min)})_{train}$; và là xấu khi MAE > 0,15*$(y_{i(\max)} - y_{i(min)})_{train}$

hoặc MAE + 3σ > 0,25*$(y_{i(\max)} - y_{i(min)})_{train}$. Phần mềm "XternalValidationPlus" do Roy và cộng sự phát triển [152] được sử dụng để hỗ trợ quá trình đánh giá ngoại trong nghiên cứu này.

2.2.6. Các công cụ máy tính khác

2.2.6.1. Bản đồ nhận thức

Các bản đồ tương tự biểu diễn đồng thời các biến (thông số mô tả/dấu vân tay) và các trường hợp (bốn nhóm hoạt tính: P, A, D, N) được tạo ra sử dụng hai phương pháp là đo lường đa hướng (multidimensional scaling - MDS) và phân tích tương hợp (correspondence analysis - CA) trong phần mềm SPSS 20.0 [172]. Gói MDS ALSCAL đo lường các ma trận sự giống nhau và sự khác nhau, sử dụng hàm chi phí Young với phương pháp đo lường Young's S-STRESS.

2.2.6.2. Mô hình hóa pharmacophore

Sử dụng phương pháp dựa vào phối tử, các giả thuyết pharmacophore trong nghiên cứu này được phát triển từ một vài tập hợp các chất khác nhau bằng công cụ Pharmacophore Elucidation trong phần mềm MOE [111]. Các thông số được sử dụng để phân tích các truy vấn pharmacophore bao gồm: (i) Độ bao phủ (cover): Là số lượng các chất có hoạt tính phù hợp với truy vấn; (ii) Độ chồng phủ (overlap): Là điểm số gióng hàng, thay đổi giữa 0 và số lượng các phân tử có hoạt tính. Điểm số càng cao cho thấy sự gióng hàng càng tốt; (iii) Độ đúng (accuracy) của truy vấn trong việc phân biệt giữa các chất có hoạt tính và các chất không có hoạt tính (các hình thể). Giá trị độ đúng bằng 1 nghĩa là truy vấn phù hợp với tất cả các chất có hoạt tính và không phù hợp với bất kỳ chất không có hoạt tính nào [111].

2.2.6.3. Mô hình hóa tương đồng

Trước khi vận hành server I-TASSER, trình tự P-gp bao gồm 1280 acid amin được tải lên dưới định dạng FASTA [115], không chỉ định các ràng buộc và protein đĩa kèm theo.

Các thông số đầu ra của server bao gồm điểm số tin cậy (confidence score - C-score), điểm số mô hình hóa đĩa (template modeling score - TM-score), căn bậc hai của độ lệch bình phương trung bình (root mean square deviation - RMSD), số lượng các mồi (decoy) và mật độ đám (cluster density) được cung cấp cho việc đánh giá định lượng các mô hình. C-score là độ tin cậy ước tính của cấu trúc dự đoán. Nằm trong giới hạn điển hình [-5; 2], giá trị C-score càng lớn thì chất lượng mô hình càng tốt và C-score > -1,5 chỉ ra dạng hình học topo đúng của mô hình dự đoán. RMSD là khoảng cách trung bình của tất cả các cặp acid amin trong cấu trúc dự đoán và cấu trúc đĩa, nằm trong khoảng 1 - 2 Å đối với các mô hình có độ phân giải cao và trong khoảng 2 - 5 Å đối với các mô hình có độ phân giải trung bình. Bởi vì RMSD có thể bị ảnh hưởng bởi một sai số cục bộ, một thông số khác cũng đo lường sự tương tự giữa hai cấu trúc là TM-score được đề nghị để giải quyết vấn đề này. TM-score < 0,17 là dự đoán ngẫu nhiên và TM-score > 0,5 chỉ ra dạng hình học topo đúng cho tất cả các kích thước của protein [208]. Mật độ đám là số lượng các bản sao cấu trúc nhiệt độ thấp tại một đơn vị không gian trong cụm SPICKER, với giá trị càng cao phản ánh chất lượng mô hình càng tốt.

Bên cạnh các thông số kể trên, chất lượng hóa lập thể của mô hình sau cùng được kiểm tra sử dụng chương trình PROCHECK [90]. Cấu trúc protein ở định dạng PDB được tải lên PDBsum để xây dựng đồ thị Ramachandran của các góc xoắn phi-psi cho tất cả các acid amin trong cấu trúc, ngoại trừ những acid amin ở các điểm cuối chuỗi. Bởi vì acid amin glycin không bị giới hạn cho bất kỳ vùng đặc biệt nào của đồ thị, chúng được xác định riêng bởi các tam giác. Căn cứ trên sự phân tích 118 cấu trúc có độ phân giải ít nhất 2,0 Å và yếu tố R (R-factor) không lớn hơn 20,0, một mô hình chất lượng tốt được kỳ vọng là có trên 90 % acid amin nằm trong các vùng được ưa thích nhất [A,B,L] (còn được gọi là các vùng lõi).

2.2.6.4. Docking phân tử

Bao gồm giai đoạn chuẩn bị và giai đoạn docking.

Chuẩn bị phối tử và protein

Cả phối tử và protein đều phải được chuẩn bị trước khi tiến hành docking. Các cấu trúc hai chiều (2D) của các phối tử được xây dựng trong phần mềm ChemBioDrawUltra 12.0 [21] nếu không có sẵn và sau đó được tối thiểu hóa năng lượng trong MOE [111]. Mô hình tương đồng tốt nhất của P-gp trong phức hợp với phối tử được thêm hydro, giới hạn lại và tối thiểu hóa bằng công cụ LigX trong MOE [111] và sau đó loại bỏ phối tử đi kèm. Vị trí gắn kết của protein mục tiêu cũng được server I-TASSER dự đoán dựa trên vị trí gắn kết tương tự của protein đĩa.

Docking

Các chất ức chế P-gp tiềm năng sau khi được xác định trước từ dự đoán của các mô hình học máy trên hai cơ sở dữ liệu sàng lọc ảo là tập nội bộ và tập Ngân hàng Thuốc sẽ được dock vào túi gắn kết phối tử của mô hình tương đồng của P-gp đã được chuẩn bị bằng gói FlexX trong phần mềm LeadIT 2.0.2 [92] để tìm hiểu về phương thức nhận diện phân tử thông qua các tương tác phối tử - protein. Trong quá trình này, thuật toán hợp tam giác (triangle matching) được sử dụng với số lượng giải pháp tối đa cho mỗi vòng lặp là 1000 và cho mỗi sự phân mảnh là 200. Các chất được phân loại là chất ức chế, có giá trị pIC_{50} dự đoán và điểm số docking tốt sẽ có nhiều khả năng là các chất ức chế P-gp hiệu quả.

2.3. Phương pháp nghiên cứu *in vitro*

Nguyên tắc của thử nghiệm *in vitro* đánh giá tác dụng ức chế bơm ngược trên các chủng vi khuẩn *S. aureus* đề kháng trong nghiên cứu này đã được trình bày trong **Mục 1.7**, bao gồm ba thử nghiệm khác nhau với thông tin về chất thử, chủng vi khuẩn, môi trường, hóa chất, trang thiết bị và cách thức bố trí, tiến hành được mô tả chi tiết như sau [1]:

2.3.1. Chất thử nghiệm

- Các chalcon nội bộ do PGS. TS. Trần Thành Đạo - Bộ môn Hóa Dược, Khoa Dược, Đại học Y Dược Thành phố Hồ Chí Minh cung cấp.
- Kháng sinh ciprofloxacin mua từ Công ty Nam Khoa Biotek.

2.3.2. Vi khuẩn thử nghiệm

- Các chủng *S. aureus* SA-1199 (tự nhiên, phân lập từ máu của bệnh nhân nhiễm trùng huyết) và SA-1199B đề kháng ciprofloxacin (đột biến, phân lập từ thỏ với thử nghiệm viêm màng trong tim, có biểu lộ quá mức NorA) do GS. TS. Michael Joseph Rybak - Đại học Wayne State, Mỹ cung cấp [3], [67].

- 156 chủng *S. aureus* phân lập lâm sàng từ các mẫu bệnh phẩm (mủ, đàm, máu, nước tiểu, đầu catheter, ống thông tiểu, dịch, ...) tại Việt Nam do GS. TS. Phạm Hùng Vân - Công ty Nam Khoa Biotek cung cấp.

2.3.3. Môi trường

- Môi trường tăng sinh: TSA (Tryptic Soy Agar), TSB (Tryptic Soy Broth) của Oxoid
- Môi trường thử nghiệm kháng sinh: Mueller-Hinton (MH) broth của Oxoid

Các môi trường được tiệt trùng ở 121 ºC, 1 atm trong 20 phút. Môi trường đã hấp tiệt trùng nhưng chưa sử dụng được bảo quản trong tủ lạnh ở 2 - 8 ºC.

2.3.4. Hoá chất, dụng cụ, thiết bị

- Dimethyl sulfoxid (DMSO) của Merck
- Phenyl-arginin-beta-naphthylamid (PaβN) của Sigma Aldrich
- Đĩa 96 giếng, micropipet, ống tuýp 5 mL, que cấy, đèn cồn, bếp, tủ sấy, tủ ấm, máy vortex, máy đo quang, …

Các nguyên liệu, hóa chất sử dụng đáp ứng tiêu chuẩn Dược dụng.

2.3.5. Thử nghiệm tác dụng ức chế bơm ngược NorA trên các chủng vi khuẩn *S. aureus* SA-1199 và SA-1199B của một số chalcon nội bộ

a. Chuẩn bị

Pha mẫu

- Cân 0,01 g mỗi chất thử nghiệm (trong trường hợp này là chalcon) cho vào ống tuýp 5 mL vô trùng. Sau đó, cho tiếp 1 mL DMSO vào và lắc cho tan hoàn toàn để thu được dung dịch mẹ có hàm lượng chất thử là 10 mg/mL.
- Pha môi trường MHB chứa chất thử X với hàm lượng 50 µg/mL (MHB XA) hoặc 100 µg/mL (MHB XB) bằng cách lấy 25 µL hoặc 50 µL dung dịch mẹ của chất thử X đã pha ở trên (hàm lượng 10 mg/mL) cho vào ống tuýp chứa 5 mL MHB.
- Pha dung dịch kháng sinh ciprofloxacin (Ci) với hàm lượng 256 µg/mL trong MHB bằng cách lấy 128 µL dung dịch ciprofloxacin trong DMSO với hàm lượng 2 mg/mL cho vào 772 µL MHB. Lưu ý là sử dụng môi trường MHB đã pha (A hoặc B) chứa chất thử (hàm lượng 50 hoặc 100 µg/mL) và môi trường MHB không chứa chất thử (đối chứng) để pha dung dịch kháng sinh.

Chuẩn bị vi khuẩn

- Cấy ria vi khuẩn thử nghiệm trên môi trường thạch TSA, ủ ở 37 ^oC trong 24 giờ.
- Lấy 3 - 5 khuẩn lạc riêng rẽ cấy vào môi trường TSB.
- Ủ từ 2 - 6 h ở 37 ^oC để hoạt hóa vi khuẩn.
- Chỉnh độ đục vi khuẩn bằng nước muối sinh lý, sao cho mật độ thu được tương đương với độ đục chuẩn McFarland 0,5 (khoảng $1,5*10^8$ CFU/mL).
- Tiếp tục pha loãng huyền dịch vi khuẩn 100 lần bằng cách lấy 30 µL huyền dịch vi khuẩn đã chỉnh độ đục ở trên cho vào 3 mL nước muối sinh lý vô trùng. Vi khuẩn đã chuẩn bị cần được sử dụng trong vòng 15 phút.

b. Tiến hành

Mỗi chất sẽ được thử nghiệm ở 02 nồng độ (50 và 100 µg/mL), trên 02 chủng vi khuẩn vi khuẩn (SA-1199 và SA-1199B), tương ứng với 4 hàng trên đĩa 96 giếng (**Hình 2.2**). Cách thực hiện ở mỗi hàng như sau:

- Dùng micropipet cho 90 µL MHB vào mỗi giếng trên đĩa nhựa, tổng cộng 12 giếng mỗi hàng.
- Dùng micropipet lấy 90 µL dung dịch kháng sinh ciprofloxacin (256 µg/mL) cho vào giếng số 1, trộn đều bằng cách hút lên xuống 3 - 4 lần.
- Thực hiện pha loãng ½ từ giếng số 1 sang giếng số 2, bằng cách lấy 90 µL dung dịch ở giếng số 1 cho sang giếng số 2 và trộn đều bằng cách hút lên xuống 3 - 4 lần.
- Tiếp tục thực hiện pha loãng ½ cho đến giếng số 11 thì ngưng. Sau khi trộn đều thì hút bỏ 90 µL ở giếng số 11.
- Dùng micropipet lấy 10 µL huyền dịch vi khuẩn (đã pha loãng 1/100) cho vào mỗi giếng từ số giếng số 1 cho đến giếng số 12, trừ giếng số 11 không cho huyền dịch vi khuẩn.
- Ủ ở 35 - 37^oC trong 24 giờ.
- Lưu ý là với kháng sinh ciprofloxacin không có sự hiện diện của chất thử phải sử dụng MHB đối chứng, còn với cặp kháng sinh ciprofloxacin - chất thử XA/XB phải sử dụng MHB XA/XB.

c. Đọc kết quả

- Thực hiện đọc kết quả bằng mắt thường với đĩa thử nghiệm được đặt trên một bề mặt sẫm màu, không phản xạ ánh sáng. Trong đó MIC là nồng độ kháng sinh tại giếng có độ pha loãng thấp nhất và vi khuẩn bắt đầu không mọc. Kết quả chỉ có giá trị khi vi khuẩn trong các giếng chứng số 12 mọc bình thường. Nếu có vi khuẩn mọc ở nồng độ cao hơn và bị ức chế ở nồng độ thấp, quá trình thử nghiệm có thể đã bị nhiễm và phải được thực hiện lại.

- Hiệu quả ức chế bơm ngược NorA của một chất thử nghiệm được xác định bằng cách so sánh các giá trị MIC của ciprofloxacin trên chủng đột biến SA-1199B (biểu lộ quá mức bơm ngược) khi không có và khi có mặt chất thử. Nếu MIC của ciprofloxacin giảm khi có mặt chất thử so với khi không có mặt chất thử, chất thử được xác định là có khả năng ức chế NorA; ngược lại là chất thử không có hiệu quả.

2.3.6. Thử nghiệm sàng lọc các chủng vi khuẩn *S. aureus* phân lập từ lâm sàng đề kháng ciprofloxacin qua trung gian bơm ngược bằng PaβN

a. Chuẩn bị

Pha mẫu

- Cân 0,01 g PaβN cho vào ống tuýp 5 mL vô trùng. Sau đó, cho tiếp 1 mL nước cất vô trùng vào và lắc cho tan hoàn toàn để thu được dung dịch mẹ có hàm lượng PaβN là 10 mg/mL. Pha lặp lại nếu cần.

- Pha môi trường MHB chứa PaβN với hàm lượng 20 µg/mL (MHB P) bằng cách lấy 10 µL dung dịch mẹ của PaβN đã pha ở trên (hàm lượng 10 mg/mL) cho vào ống tuýp chứa 5 mL MHB, pha lặp lại nếu cần.

- Pha dung dịch kháng sinh ciprofloxacin (Ci) với hàm lượng 256 µg/mL trong MHB bằng cách lấy 128 µL dung dịch ciprofloxacin với hàm lượng 2 mg/mL cho vào 772 µL MHB, pha lặp lại nếu cần. Lưu ý là sử dụng môi trường MHB P đã pha (chứa PaβN với hàm lượng 20 µg/mL) và môi trường MHB đối chứng (không chứa chất thử) để pha dung dịch kháng sinh.

Chuẩn bị vi khuẩn

Tương tự như **Mục 2.3.5**.

b. Tiến hành

Thử nghiệm được thực hiện theo bố trí như mô tả trong **Hình 2.3**, trong đó cách pha loãng ở mỗi hàng tương tự như **Mục 2.3.5**.

c. Đọc kết quả

Việc đọc kết quả MIC tương tự như **Mục 2.3.5**. Một chủng SA lâm sàng được xác định là có đề kháng ciprofloxacin qua trung gian bơm ngược nếu MIC của ciprofloxacin giảm khi có mặt PaβN so với khi không có mặt chất ức chế bơm chuẩn này và ngược lại.

2.3.7. Thử nghiệm tác dụng ức chế bơm ngược trên các chủng vi khuẩn *S. aureus* lâm sàng có đề kháng ciprofloxacin qua trung gian bơm ngược của một số chalcon nội bộ

a. Chuẩn bị

Pha mẫu

- Cân 0,01 g mỗi chất thử nghiệm (trong trường hợp này là chalcon) cho vào ống tuýp 5 mL vô trùng. Sau đó, cho tiếp 1 mL DMSO vào và lắc cho tan hoàn toàn để thu được dung dịch mẹ có hàm lượng chất thử là 10 mg/mL. Pha lặp lại nếu cần.
- Pha môi trường MHB chứa chất thử X_i với hàm lượng 20 µg/mL (MHB X_i) bằng cách lấy 10 µL dung dịch mẹ của chất thử X_i đã pha ở trên (hàm lượng 10 mg/mL) cho vào ống tuýp chứa 5 mL MHB, pha lặp lại nếu cần.
- Pha dung dịch kháng sinh ciprofloxacin (Ci) với hàm lượng 256 µg/mL trong MHB bằng cách lấy 128 µL dung dịch ciprofloxacin trong DMSO với hàm lượng 2 mg/mL cho vào 772 µL MHB, pha lặp lại nếu cần. Lưu ý là sử dụng môi trường MHB đã pha (MHB X_i) chứa chất thử (hàm lượng 20 µg/mL) và môi trường MHB không chứa chất thử (đối chứng) để pha dung dịch kháng sinh.

Chuẩn bị vi khuẩn

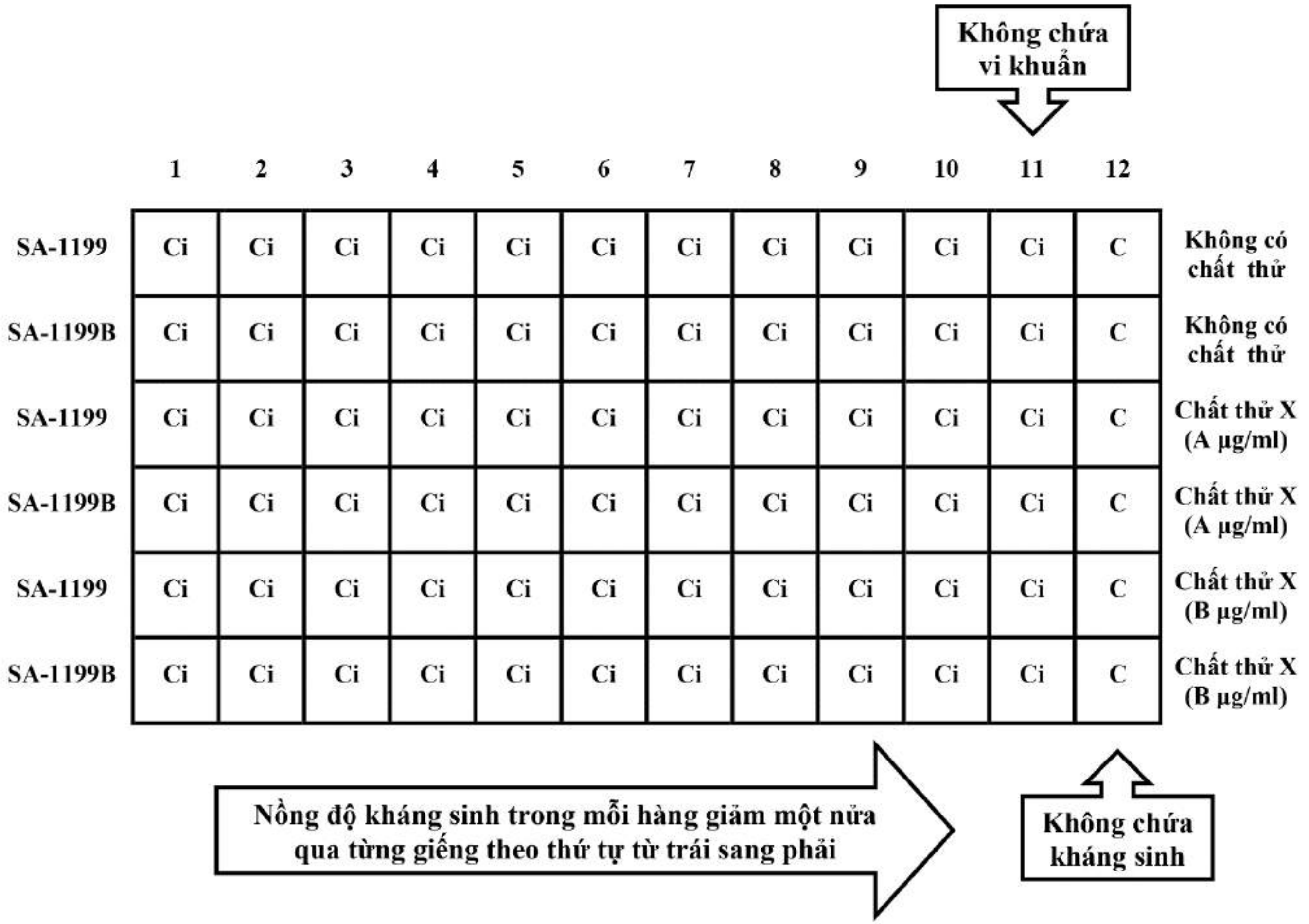
Các chủng vi khuẩn thu được từ **Mục 2.3.6** được chuẩn bị tương tự như **Mục 2.3.5**.

b. Tiến hành

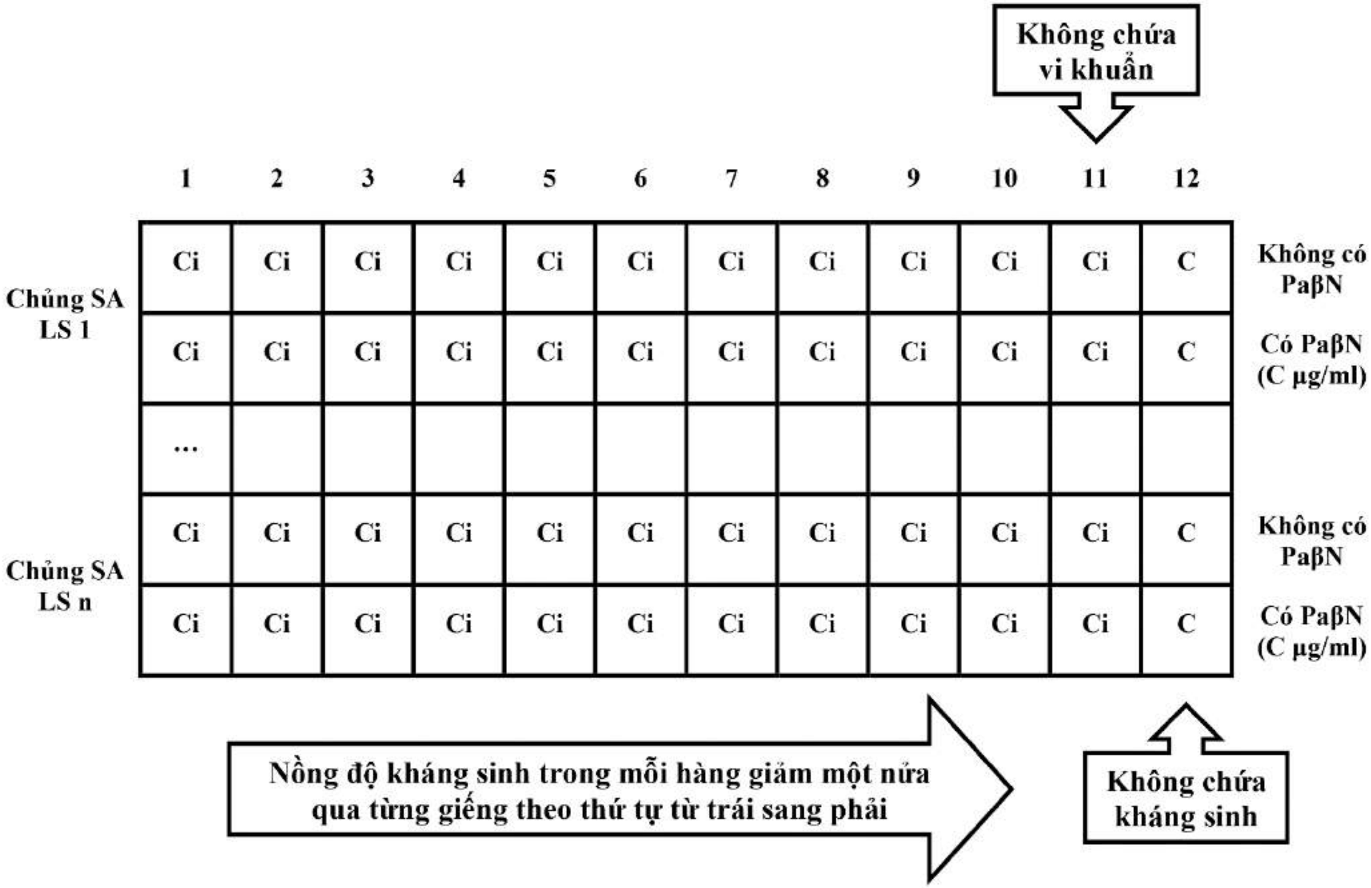
Mỗi chất sẽ được thử nghiệm ở nồng độ 20 µg/mL, trên các chủng vi khuẩn *S. aureus* lâm sàng đề kháng ciprofloxacin bằng bơm ngược được sàng lọc. Quá trình này được thực hiện theo bố trí như mô tả trong **Hình 2.4**, trong đó cách pha loãng ở mỗi hàng tương tự như **Mục 2.3.5**.

c. Đọc kết quả

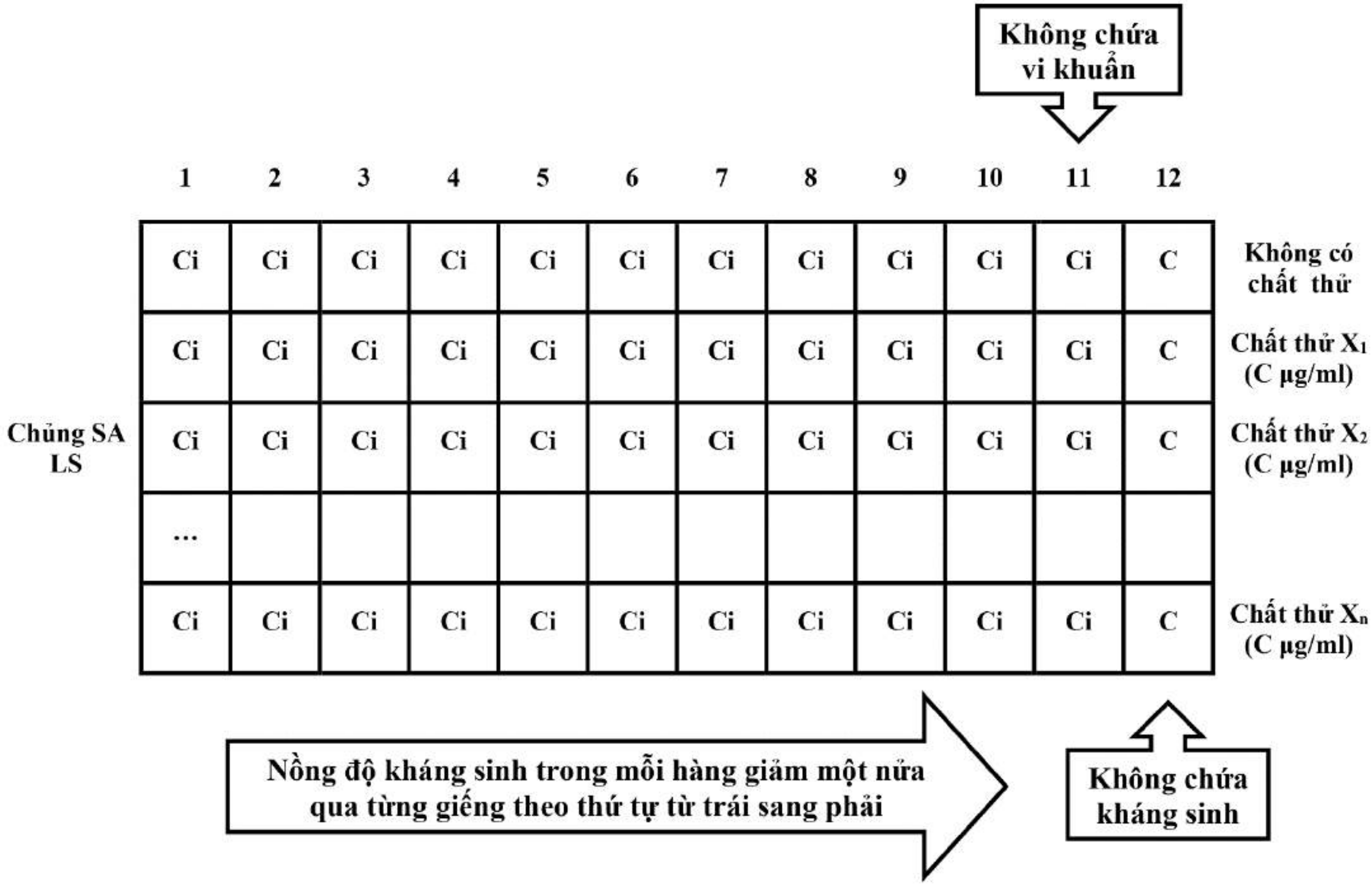
Việc đọc kết quả MIC tương tự như **Mục 2.3.5**. Một chất thử nghiệm được xác định là có khả năng ức chế bơm ngược của một chủng SA lâm sàng nếu MIC của ciprofloxacin giảm khi có mặt chất thử đó so với khi không có mặt nó và ngược lại.



Hình 2.2. Bố trí thử nghiệm *in vitro* xác định MIC của ciprofloxacin (Ci) trên các chủng *S. aureus* SA-1199 và SA-1199B khi vắng mặt và khi có mặt chất thử nghiệm X ở các nồng độ khác nhau (A, B µg/mL), qua đó đánh giá khả năng ức chế bơm ngược NorA của SA của chất thử nghiệm. Trong mỗi hàng ngang của đĩa, tất cả các giếng chứa kháng sinh (trừ giếng số 11) được cho cùng lượng và loại vi khuẩn như giếng kiểm soát C (chứa vi khuẩn nhưng không có kháng sinh).



Hình 2.3. Bố trí thử nghiệm *in vitro* xác định MIC của ciprofloxacin (Ci) trên các chủng *S. aureus* phân lập từ lâm sàng khi vắng mặt và khi có mặt chất ức chế bơm đã biết là PaβN ở nồng độ C = 20 µg/mL, qua đó chọn lọc ra các chủng SA lâm sàng có biểu lộ quá mức bơm ngược. Trong mỗi hàng ngang của đĩa, tất cả các giếng chứa kháng sinh (trừ giếng số 11) được cho cùng lượng và loại vi khuẩn như giếng kiểm soát C (chứa vi khuẩn nhưng không có kháng sinh).



Hình 2.4. Bố trí thử nghiệm *in vitro* xác định MIC của ciprofloxacin (Ci) trên các chủng *S. aureus* phân lập từ lâm sàng có biểu lộ quá mức bơm ngược, khi vắng mặt và khi có mặt các chất thử nghiệm X_1, X_2, …, X_n ở nồng độ C = 20 µg/mL, qua đó đánh giá khả năng ức chế bơm ngược của các SA lâm sàng của từng chất thử nghiệm. Trong mỗi hàng ngang của đĩa, tất cả các giếng chứa kháng sinh (trừ giếng số 11) được cho cùng lượng và loại vi khuẩn như giếng kiểm soát C (chứa vi khuẩn nhưng không có kháng sinh).

CHƯƠNG 3. KẾT QUẢ

3.1. Các mô hình máy tính dựa trên phối tử

3.1.1. Các mô hình phân loại chất ức chế và chất không ức chế P-gp

Tổng cộng 2982 thông số mô tả và dấu vân tay MOE và PaDEL đã được tính toán cho toàn bộ tập dữ liệu 2134 chất. Trong đó, 5 thông số mô tả (Kier1, Kier2, Kier3, apol và bpol) được xác định là trùng lặp và bị loại bỏ. 25 chất bị thiếu các giá trị thuộc tính cũng được loại bỏ trước khi tiến hành lựa chọn biến cho mục đích phân loại. Với đầu vào là 2109 chất còn lại (**tập tin TLBS.xlsx, Sheet1** và **bản in Tài liệu bổ sung, TLBS1**), quá trình giảm biến trong RapidMiner bao gồm lọc thô và lựa chọn tối ưu thu được 759 thuộc tính. Số lượng thuộc tính tiếp tục được giảm xuống còn 71 thuộc tính với số lần đánh giá chéo ≥ 1 (10 %) trong Weka. Tuy nhiên, chỉ 24 thuộc tính có liên quan nhất với số lần đánh giá chéo ≥ 8 (80 %), bao gồm 13 thông số mô tả hai chiều (2D) và 11 dấu vân tay (**Phụ lục 2**) được lựa chọn để phát triển các mô hình học máy. Trong đó, PEOE_VSA_FPPOS là thông số duy nhất được tính bằng MOE; thông số này phụ thuộc vào điện tích từng phần của mỗi nguyên tử trong một cấu trúc hóa học. Như vậy, toàn bộ tập dữ liệu các chất ức chế và không ức chế P-gp được đại diện chủ yếu bởi các thuộc tính PaDEL. 12 thông số mô tả PaDEL hai chiều (2D) lần lượt mô tả các tính chất tự tương quan (2 thông số), các giá trị eigen được biến đổi Burden (4 thông số), nguyên tử hóa học topo mở rộng (1 thông số), số lượng vòng (2 thông số), điện tích topo (2 thông số) và ma trận khoảng cách hình học topo (1 thông số). Các dấu vân tay được chọn bao gồm 1 chuỗi MACCS, 5 chuỗi Pubchem và 5 chuỗi dưới cấu trúc lần lượt được mã hóa bởi các ký hiệu, SMILES và SMARTS; và đại diện cho các nhóm cấu trúc và chức năng như được mô tả trong **Phụ lục 2**.

Với 1690 chất trong các tập huấn luyện thu được từ sự phân chia đa dạng và ngẫu nhiên, hoạt tính ức chế P-gp là một biến nhị phân (chất ức chế hoặc chất không ức chế) được mô hình hóa tự động, sử dụng các thuật toán học máy tích hợp trong hạch Binary Classifier trong Clementine. Trong cả hai kiểu phân chia dữ liệu (**Bảng**

3.1 và **Bảng 3.2**), 9/10 mô hình ứng viên là mạng nơron, C5.0, C&R Tree, QUEST, CHAID, hồi quy logistic, mặt nghiêng quyết định, mạng Bayesian và SVM (trừ phân tích phân biệt) đều được tạo ra theo các thông số mặc định để dự đoán biến kết quả, dựa vào 24 thuộc tính đã chọn. Trong đó, các mô hình mặt nghiêng quyết định chỉ dự đoán các chất có hoạt tính, nghĩa là với những mô hình này thì một chất bất kỳ hoặc được phân loại là chất ức chế hoặc không được phân loại. Mặc dù không thể phân biệt giữa chất ức chế và chất không ức chế P-gp như kỳ vọng nhưng với độ chính xác cao nhất (các giá trị dự đoán dương PPV $\geq$ 0,9), các mô hình mặt nghiêng quyết định này đã cho thấy chúng có khả năng tốt nhất trong việc xác định các chất ức chế trong các tập dữ liệu lớn và phức tạp.

Các mô hình tạo ra được so sánh với nhau bằng thông số độ đúng tổng thể. Trong sự phân chia đa dạng, các giá trị độ đúng tổng thể trên tập huấn luyện nhỏ hơn nhiều so với các giá trị độ đúng tổng thể trên tập đánh giá nội trong tất cả các mô hình. Ngược lại, các giá trị độ đúng tổng thể trên tập huấn luyện lại lớn hơn hoặc nhỏ hơn không đáng kể so với các giá trị độ đúng tổng thể trên tập đánh giá nội trong sự phân chia ngẫu nhiên. Ngoài ra, các giá trị độ đúng tổng thể của tất cả các mô hình khi phân chia đa dạng nhỏ hơn khi phân chia ngẫu nhiên trong trường hợp tập huấn luyện và ngược lại trong trường hợp tập đánh giá nội. Các kết quả thu được cho thấy tập huấn luyện đa dạng thích hợp hơn tập huấn luyện ngẫu nhiên để sử dụng cho mục đích phát triển các mô hình học máy.

Bảng 3.1. Kết quả dự đoán trên tập huấn luyện và tập đánh giá nội với sự phân chia đa dạng.

	C5.0		Mạng nơron		SVM		Hồi quy logistic		CHAID		C&R Tree		Mạng Bayesian		QUEST		Mặt nghiêng quyết định		Ensemble	
	HL	**ĐGN**	**HL**	**ĐGN**	**HL**	**ĐGN**	**HL**	**ĐGN**	**HL**	**ĐGN**	**HL**	**ĐGN**	**HL**	**ĐGN**	**HL**	**ĐGN**	**HL**	**ĐGN**	**HL**	**ĐGN**
Dương tính thật	878	347	827	347	856	350	857	351	820	344	856	347	828	333	805	337	326	190	860	350
Dương tính giả	182	26	155	20	196	22	201	25	163	22	215	30	210	26	202	24	36	4	184	24
Âm tính thật	556	35	583	41	542	39	537	36	575	39	523	31	528	35	536	37	0	0	554	37
Âm tính giả	74	11	125	11	96	8	95	7	132	14	96	11	124	25	147	21	0	0	92	8
Độ đúng tổng thể	0,85	0,91	0,83	0,93	0,83	0,93	0,82	0,92	0,83	0,91	0,82	0,90	0,80	0,88	0,79	0,89	0,90	0,98	0,84	0,92
Độ nhạy	0,92	0,97	0,87	0,97	0,90	0,98	0,90	0,98	0,86	0,96	0,90	0,97	0,87	0,93	0,85	0,94	1,00	1,00	0,90	0,98
Độ đặc hiệu	0,75	0,57	0,79	0,67	0,73	0,64	0,73	0,59	0,78	0,64	0,71	0,51	0,72	0,57	0,73	0,61	0,00	0,00	0,75	0,61
Độ chính xác	0,83	0,93	0,84	0,95	0,81	0,94	0,81	0,93	0,83	0,94	0,80	0,92	0,80	0,93	0,80	0,93	0,90	0,98	0,82	0,94
Giá trị dự đoán âm	0,88	0,76	0,82	0,79	0,85	0,83	0,85	0,84	0,81	0,74	0,84	0,74	0,81	0,58	0,78	0,64	-	-	0,86	0,82
MCC	0,69	0,61	0,66	0,69	0,65	0,69	0,64	0,66	0,64	0,64	0,63	0,56	0,60	0,51	0,58	0,56	-	-	0,67	0,67
G-mean	0,83	0,75	0,83	0,81	0,81	0,79	0,81	0,76	0,82	0,78	0,80	0,70	0,79	0,73	0,78	0,76	0,00	0,00	0,82	0,77
Chỉ số Youden's	0,68	0,54	0,66	0,64	0,63	0,62	0,63	0,57	0,64	0,60	0,61	0,48	0,59	0,50	0,57	0,55	0,00	0,00	0,65	0,58
Điểm số GH cho chất có hoạt tính	0,88	0,95	0,86	0,96	0,86	0,96	0,86	0,96	0,85	0,95	0,85	0,94	0,83	0,93	0,82	0,94	0,95	0,99	0,86	0,96
Điểm số GH cho chất không có hoạt tính	0,82	0,67	0,81	0,73	0,79	0,73	0,79	0,71	0,80	0,69	0,78	0,62	0,76	0,58	0,76	0,62	-	-	0,80	0,71

*HL: Tập huấn luyện; ĐGN: Tập đánh giá nội.

Bảng 3.2. Kết quả dự đoán trên tập huấn luyện và tập đánh giá nội với sự phân chia ngẫu nhiên.

	C5.0		**Mạng nơron**		**SVM**		**Hồi quy logistic**		**CHAID**		**C&R Tree**		**Mạng Bayesian**		**QUEST**		**Mặt nghiêng quyết định**		**Ensemble**	
	HL	**ĐGN**	**HL**	**ĐGN**	**HL**	**ĐGN**	**HL**	**ĐGN**	**HL**	**ĐGN**	**HL**	**ĐGN**	**HL**	**ĐGN**	**HL**	**ĐGN**	**HL**	**ĐGN**	**HL**	**ĐGN**
Dương tính thật	971	224	960	242	967	243	973	244	921	223	958	233	943	236	928	230	467	111	975	243
Dương tính giả	101	45	183	49	183	48	186	50	146	48	166	45	187	51	188	45	18	13	172	43
Âm tính thật	538	115	456	111	456	112	453	110	493	112	473	115	452	104	451	115	0	0	467	117
Âm tính giả	80	35	91	17	84	16	78	15	130	36	93	26	108	22	123	29	0	0	76	16
Độ đúng tổng thể	0,89	0,81	0,84	0,84	0,84	0,85	0,84	0,84	0,84	0,80	0,85	0,83	0,83	0,82	0,82	0,82	0,96	0,90	0,85	0,86
Độ nhạy	0,92	0,86	0,91	0,93	0,92	0,94	0,93	0,94	0,88	0,86	0,91	0,90	0,90	0,91	0,88	0,89	1,00	1,00	0,93	0,94
Độ đặc hiệu	0,84	0,72	0,71	0,69	0,71	0,70	0,71	0,69	0,77	0,70	0,74	0,72	0,71	0,67	0,71	0,72	0,00	0,00	0,73	0,73
Độ chính xác	0,91	0,83	0,84	0,83	0,84	0,84	0,84	0,83	0,86	0,82	0,85	0,84	0,83	0,82	0,83	0,84	0,96	0,90	0,85	0,85
Giá trị dự đoán âm	0,87	0,77	0,83	0,87	0,84	0,88	0,85	0,88	0,79	0,76	0,84	0,82	0,81	0,83	0,79	0,80	-	-	0,86	0,88
MCC	0,77	0,59	0,65	0,66	0,66	0,67	0,66	0,67	0,65	0,57	0,67	0,64	0,62	0,62	0,60	0,62	-	-	0,68	0,70
G-mean	0,88	0,79	0,81	0,81	0,81	0,81	0,81	0,80	0,82	0,78	0,82	0,80	0,80	0,78	0,79	0,80	0,00	0,00	0,82	0,83
Chỉ số Youden's	0,77	0,58	0,63	0,63	0,63	0,64	0,63	0,63	0,65	0,56	0,65	0,62	0,60	0,59	0,59	0,61	0,00	0,00	0,66	0,67
Điểm số GH cho chất có hoạt tính	0,91	0,85	0,88	0,88	0,88	0,89	0,88	0,89	0,87	0,84	0,88	0,87	0,87	0,87	0,86	0,86	0,98	0,95	0,89	0,89
Điểm số GH cho chất không có hoạt tính	0,86	0,74	0,77	0,78	0,78	0,79	0,78	0,78	0,78	0,73	0,79	0,77	0,76	0,75	0,75	0,76	-	-	0,80	0,81

*HL: Tập huấn luyện; ĐGN: Tập đánh giá nội.

Trong nghiên cứu này, quá trình xác định phạm vi khả năng ứng dụng phát hiện ra 59/1690 chất của tập huấn luyện đa dạng là các chất chất lạ; 1/419 chất của tập đánh giá nội và không có chất nào của tập đánh giá ngoại nằm ngoài phạm vi khả năng ứng dụng (**tập tin TLBS.xlsx, Sheet6**). Việc xây dựng và đánh giá các mô hình học máy vẫn được tiến hành với sự hiện diện của các chất này. Dựa trên độ đúng tổng thể, các mô hình C5.0 cho kết quả phân loại tốt nhất trên cả tập huấn luyện đa dạng và tập huấn luyện ngẫu nhiên với độ đúng lần lượt là 85 % và 89 %, nhưng các mô hình khác cũng gần đúng như C5.0. Tuy nhiên trên các tập đánh giá nội tương ứng, ba mô hình tốt nhất lại là SVM, mạng nơron và hồi quy logistic khi chúng lần lượt dự đoán đúng trên 92 % và 84 % tổng số chất. Ngoài ra, có sáu chất trong tập đánh giá nội ngẫu nhiên không được phân loại bởi mô hình mạng Bayesian là endosulfan, acid ferulic, N-acetylaspartat, acetaminophen, mesna và busulfan. Để tránh những hạn chế như vậy của các mô hình đơn lẻ, giải pháp kết hợp các dự đoán từ nhiều mô hình được đặt ra trong nghiên cứu này.

Ngoại trừ mặt nghiêng quyết định, tám mô hình còn lại (**Phụ lục 3**) được gộp chung thành một mô hình kết hợp bằng hạch Ensemble. Để so sánh từng mô hình đơn lẻ với mô hình kết hợp, tùy chọn “Filter out fields generated by ensembled models” không được lựa chọn. Phương pháp kết hợp là bầu chọn dựa trên trọng số độ tin cậy (confidence-weighted voting), trong đó các bầu chọn được đo lường căn cứ vào giá trị độ tin cậy cho mỗi dự đoán. Một lợi ích thường gặp của phương pháp này là khả năng tạo ra những dự đoán đúng hơn so với bất kỳ mô hình đơn lẻ nào [29]. Thật vậy, mô hình kết hợp đã cho thấy khả năng phân loại tốt với độ đúng tổng thể thu được lần lượt là 84 % và 85 % trên các tập huấn luyện đa dạng và ngẫu nhiên, và 92 % và 86 % trên các tập đánh giá nội tương ứng. Sự kết hợp nhiều mô hình đã thực hiện dự đoán đúng nhất trong trường hợp tập đánh giá nội ngẫu nhiên. Mặc dù không hoàn toàn tốt như một vài mô hình đơn lẻ trong các trường hợp khác, nhưng những khác biệt ghi nhận được là không đáng kể. Các mô hình kết hợp đã được chứng minh là mô hình mạnh, có thể thực hiện việc dự đoán chính xác cho các tập dữ liệu khác nhau trong điều kiện chung mà không cần đi sâu lựa chọn và tối ưu các thông số của một mô hình đơn lẻ bất kỳ.

Các quá trình đánh giá chéo 10 lần và ngẫu nhiên hóa biến phụ thuộc tiếp tục được thực hiện trên tập huấn luyện đa dạng cho mục đích đánh giá nội (**Bảng 3.3**). Tất cả các mô hình thu được đều có độ đúng tổng thể từ 77 % trở lên khi đánh giá chéo, trong đó hai mô hình đơn lẻ SVM và mạng nơron cùng với mô hình kết hợp là các mô hình phân loại tốt nhất với độ đúng bằng 82 %. Khi ngẫu nhiên hóa biến phụ thuộc, giá trị độ đúng tổng thể bị giảm ít nhất là 19 % với mô hình mạng Bayesian và nhiều nhất là 29 % với mô hình C5.0. Trong quá trình đánh giá ngoại, sáu mô hình đơn lẻ là C5.0, mạng nơron, SVM, hồi quy logistic, CHAID, mạng Bayesian và mô hình kết hợp đã dự đoán đúng 100 % số chất của tập đánh giá ngoại; hai mô hình đơn lẻ còn lại là C&R Tree và QUEST có khả năng dự đoán kém hơn một chút khi lần lượt phân loại đúng 21/22 và 20/22 chất của tập đánh giá ngoại (**Bảng 3.4**).

Bảng 3.3. Kết quả đánh giá chéo 10 lần và y ngẫu nhiên trên tập huấn luyện đa dạng.

	C5.0		**Mạng nơron**		**SVM**		**Hồi quy logistic**		**CHAID**		**C&R Tree**		**Mạng Bayesian**		**QUEST**		**Ensemble**	
	ĐGC	**YNN**	**ĐGC**	**YNN**	**ĐGC**	**YNN**	**ĐGC**	**YNN**	**ĐGC**	**YNN**	**ĐGC**	**YNN**	**ĐGC**	**YNN**	**ĐGC**	**YNN**	**ĐGC**	**YNN**
Dương tính thật	829	-	823	-	852	-	500	-	548	-	829	-	475	-	800	-	843	-
Dương tính giả	210	-	179	-	207	-	131	-	165	-	219	-	144	-	217	-	198	-
Âm tính thật	528	-	559	-	531	-	326	-	366	-	519	-	297	-	521	-	540	-
Âm tính giả	123	-	129	-	100	-	57	-	104	-	123	-	81	-	152	-	109	-
Độ đúng tổng thể	0,80	0,56	0,82	0,57	0,82	0,58	0,81	0,57	0,77	0,58	0,80	0,59	0,77	0,61	0,78	0,57	0,82	0,57
Độ nhạy	0,87	1,00	0,86	0,99	0,90	0,97	0,90	0,93	0,84	0,94	0,87	0,91	0,85	0,83	0,84	0,97	0,89	0,98
Độ đặc hiệu	0,72	0,00	0,76	0,02	0,72	0,07	0,71	0,10	0,69	0,11	0,70	0,18	0,67	0,34	0,71	0,05	0,73	0,04
Độ chính xác	0,80	0,56	0,82	0,57	0,80	0,57	0,79	0,57	0,77	0,58	0,79	0,59	0,77	0,62	0,79	0,57	0,81	0,57
Giá trị dự đoán âm	0,81	0,73	0,81	0,59	0,84	0,63	0,85	0,48	0,78	0,65	0,81	0,61	0,79	0,60	0,77	0,56	0,83	0,68
MCC	0,60	0,08	0,63	0,06	0,63	0,08	0,63	0,04	0,54	0,11	0,59	0,14	0,54	0,19	0,55	0,06	0,63	0,09
G-mean	0,79	0,02	0,81	0,09	0,80	0,25	0,80	0,28	0,76	0,22	0,78	0,37	0,76	0,53	0,77	0,16	0,80	0,17
Chỉ số Youden's	0,59	0,00	0,62	0,01	0,61	0,04	0,61	0,02	0,53	0,05	0,57	0,09	0,53	0,16	0,55	0,02	0,62	0,03
Điểm số GH cho chất có hoạt tính	0,83	0,78	0,84	0,78	0,85	0,77	0,85	0,75	0,80	0,76	0,83	0,75	0,81	0,72	0,81	0,77	0,85	0,78
Điểm số GH cho chất không có hoạt tính	0,76	0,38	0,79	0,32	0,78	0,35	0,78	0,29	0,73	0,40	0,76	0,41	0,73	0,47	0,74	0,32	0,78	0,37

*ĐGC: Đánh giá chéo; YNN: Y ngẫu nhiên.

Bảng 3.4. Kết quả dự đoán trên tập đánh giá ngoại của các mô hình được tạo ra từ tập huấn luyện đa dạng.

	C5.0	Mạng nơron	SVM	Hồi quy logistic	CHAID	C&R Tree	Mạng Bayesian	QUEST	Ensemble
Dương tính thật	19	19	19	19	19	19	19	18	19
Dương tính giả	0	0	0	0	0	1	0	1	0
Âm tính thật	3	3	3	3	3	2	3	2	3
Âm tính giả	0	0	0	0	0	0	0	1	0
Độ đúng tổng thể	1,00	1,00	1,00	1,00	1,00	0,95	1,00	0,91	1,00
Độ nhạy	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,95	1,00
Độ đặc hiệu	1,00	1,00	1,00	1,00	1,00	0,67	1,00	0,67	1,00
Độ chính xác	1,00	1,00	1,00	1,00	1,00	0,95	1,00	0,95	1,00
Giá trị dự đoán âm	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,67	1,00
MCC	1,00	1,00	1,00	1,00	1,00	0,80	1,00	0,61	1,00
G-mean	1,00	1,00	1,00	1,00	1,00	0,82	1,00	0,79	1,00
Chỉ số Youden’s	1,00	1,00	1,00	1,00	1,00	0,67	1,00	0,61	1,00
Điểm số GH cho chất có hoạt tính	1,00	1,00	1,00	1,00	1,00	0,98	1,00	0,95	1,00
Điểm số GH cho chất không có hoạt tính	1,00	1,00	1,00	1,00	1,00	0,83	1,00	0,67	1,00

3.1.2. Các mô hình dự đoán hoạt tính ức chế P-gp

Tổng cộng 1628 thông số mô tả MOE và PaDEL đã được tính toán cho toàn bộ tập dữ liệu 499 chất. Trong đó, 5 thông số mô tả (Kier1, Kier2, Kier3, apol và bpol) được xác định là trùng lặp và bị loại bỏ. Tất cả 400 chất của tập huấn luyện và tập đánh giá nội không bị thiếu thông số mô tả nào và đều được sử dụng cho việc lựa chọn biến. Quá trình giảm biến trong RapidMiner bao gồm lọc thô và lựa chọn tối ưu thu được 383 thông số. Số lượng thông số tiếp tục được giảm xuống còn 89 thông số với số lần đánh giá chéo ≥ 1 (10 %) trong Weka. Tuy nhiên, chỉ 34 thông số có liên quan nhất với số lần đánh giá chéo ≥ 8 (80 %) (**Phụ lục 4**) được lựa chọn để phát triển các mô hình học máy. Trong đó, các thông số PaDEL chiếm đa số so với các thông số MOE (27 so với 7) để đại diện cho toàn bộ tập dữ liệu. Các thông số MOE mô tả các tính chất ma trận gần kề và cách xa (2 thông số), số lượng các nguyên tử và liên kết (1 thông số) và điện tích từng phần (4 thông số); trong khi các thông số PaDEL mô tả các tính chất tự tương quan (4 thông số), ma trận Barysz (4 thông số), các giá trị eigen được biến đổi Burden (9 thông số), trạng thái điện tử topo loại nguyên tử (4 thông số), nội dung thông tin (1 thông số), chuỗi béo dài nhất (1 thông số), quan hệ năng lượng tự do tuyến tính phân tử (1 thông số), số lượng vòng (1 thông số), điện tích topo (1 thông số) và ma trận khoảng cách hình học topo (1 thông số).

Với 300 chất trong các tập huấn luyện thu được từ sự phân chia đa dạng và ngẫu nhiên, mối quan hệ giữa cấu trúc hóa học (các thông số mô tả) và hoạt tính ức chế P-gp (biến liên tục) được mô hình hóa tự động, sử dụng các thuật toán học máy tích hợp trong hạch Numeric Predictor trong Clementine. Trong cả hai kiểu phân chia dữ liệu, tất cả sáu mô hình ứng viên là mạng nơron, C&R Tree, CHAID, hồi quy, tuyến tính suy rộng và SVM đều được tạo ra theo các thông số mặc định để dự đoán biến kết quả, dựa vào 34 thông số mô tả đã chọn (**Bảng 3.5**). Do hai mô hình hồi quy và tuyến tính suy rộng có kết quả dự đoán giống nhau cho tất cả các chất trong tập huấn luyện trong cả hai trường hợp phân chia, nên chỉ có mô hình hồi quy được chọn để tiếp tục xây dựng mô hình kết hợp.

Bảng 3.5. Sáu mô hình đơn lẻ được tạo ra cùng với các giá trị R^2 của chúng trên tập huấn luyện và tập đánh giá nội, theo hai kiểu phân chia dữ liệu đa dạng và ngẫu nhiên.

Kiểu phân chia	Mô hình tạo ra	$R^2_{\text{tập huấn luyện}}$	$R^2_{\text{tập đánh giá nội}}$
Đa dạng	CHAID	0,86	0,79
	C&R Tree	0,83	0,81
	Hồi quy	0,75	0,77
	Tuyến tính suy rộng	0,75	0,77
	SVM	0,74	0,81
	Mạng nơron	0,73	0,75
Ngẫu nhiên	CHAID	0,86	0,62
	C&R Tree	0,84	0,74
	Hồi quy	0,76	0,73
	Tuyến tính suy rộng	0,76	0,73
	SVM	0,74	0,71
	Mạng nơron	0,72	0,74

Các mô hình tạo ra được so sánh với nhau bằng hệ số tương quan bình phương R^2 (càng gần 1 thì mối quan hệ càng mạnh). Trong sự phân chia đa dạng, các giá trị R^2 trong tập huấn luyện nhỏ hơn các giá trị R^2 trong tập đánh giá nội trong trường hợp các mô hình hồi quy, SVM và mạng nơron; và ngược lại trong trường hợp các mô hình CHAID và C&R Tree. Trong sự phân chia ngẫu nhiên, giá trị R^2 trong tập huấn luyện nhỏ hơn giá trị R^2 trong tập đánh giá nội trong trường hợp mô hình mạng nơron; và ngược lại trong trường hợp các mô hình còn lại (CHAID, C&R Tree, hồi quy và SVM). Ngoài ra, mặc dù các giá trị R^2 của tất cả các mô hình khi phân chia đa dạng xấp xỉ các giá trị R^2 của tất cả các mô hình khi phân chia ngẫu nhiên trong trường hợp tập huấn luyện, hầu hết các giá trị này khi phân chia đa dạng cao hơn đáng kể khi phân chia ngẫu nhiên trong trường hợp tập đánh giá nội. Các kết quả thu được cho thấy tập huấn luyện đa dạng thích hợp hơn tập huấn luyện ngẫu nhiên để sử dụng cho mục đích phát triển các mô hình học máy.

Dựa trên các giá trị R^2, các mô hình CHAID và C&R Tree được xếp hạng tốt nhất trong cả các tập huấn luyện đa dạng (86 % và 83 %) và ngẫu nhiên (86 % và 84 %); trong khi đó các mô hình thực hiện dự đoán tốt nhất là C&R Tree và SVM trong trường hợp tập đánh giá nội đa dạng (cùng 81 %), và là C&R Tree và mạng nơron trong trường hợp tập đánh giá nội ngẫu nhiên (cùng 74 %). Sự khác biệt lớn nhất về giá trị R^2 giữa các tập huấn luyện và tập đánh giá nội là ở mô hình CHAID khi phân chia ngẫu nhiên. Ngược lại, không có sự khác biệt đáng kể về thông số này giữa các tập huấn luyện đa dạng và ngẫu nhiên, cũng như giữa các tập đánh giá nội đa dạng và ngẫu nhiên trong trường hợp mạng nơron. Để tránh những hạn chế như vậy của các mô hình đơn lẻ, giải pháp kết hợp các dự đoán từ nhiều mô hình được đặt ra trong nghiên cứu này.

Ngoại trừ tuyến tính suy rộng, năm mô hình còn lại (**Phụ lục 5**) được gộp chung thành một mô hình kết hợp bằng hạch Ensemble. Để so sánh từng mô hình đơn lẻ với mô hình kết hợp, tùy chọn "Filter out fields generated by ensembled models" cũng không được lựa chọn. Các dự đoán kết hợp cho biến mục tiêu được tạo ra bằng cách lấy trung bình giá trị dự đoán của các mô hình đơn lẻ. Phương pháp này cũng thường mang lại lợi ích là khả năng tạo ra những dự đoán đúng hơn so với bất kỳ mô hình đơn lẻ nào [29]. Thật vậy, trong khi các mô hình đơn lẻ không đạt ít nhất một trong số các điều kiện được sử dụng để đánh giá nội trên tập huấn luyện đa dạng, cụ thể là Q^2_{LOO} (< 0,6) và $|R^2 - Q^2_{LOO}|$ (> 0,1) trong các mô hình CHAID, C&R Tree; $\overline{r^2_m}$ trong các mô hình hồi quy, SVM, mạng nơron (< 0,65) và Δr^2_m trong mô hình hồi quy (> 0,2); mô hình kết hợp đã cho thấy khả năng dự đoán tốt với R^2 = 0,84; Q^2_{LOO} = 0,70; $|R^2 - Q^2_{LOO}|$ = 0,14; r^2_m = 0,80; r'^2_m = 0,64; $\overline{r^2_m}$ = 0,72 và Δr^2_m = 0,16 (**Bảng 3.6**). Giá trị $|R^2 - Q^2_{LOO}|$ của mô hình kết hợp mặc dù vượt 0,1, có thể được giải thích là do ảnh hưởng của hai mô hình CHAID và C&R Tree, tuy nhiên khác biệt này còn nhỏ hơn 0,3 nên vẫn đảm bảo được khả năng dự đoán [127], [192]. Tất cả sáu mô hình đều có $|R^2 - R^2_{Yi}| \geq 0{,}2$ và đáp ứng điều kiện $R^2_p \geq 0{,}5$ khi ngẫu nhiên hóa biến phụ thuộc (y-randomization) trong tập huấn luyện (**Bảng 3.6**). Mặc dù không hoàn toàn tốt như các mô hình đơn lẻ trong một vài điều kiện cụ thể, nhưng xét trên phương diện tổng thể thì sự kết hợp nhiều mô hình đã cho thấy kết quả dự đoán tốt nhất. Mô hình kết hợp đã được chứng minh là một mô hình mạnh, có thể thực hiện việc dự đoán chính xác cho các tập dữ liệu khác nhau trong điều kiện chung mà không cần đi sâu lựa chọn và tối ưu các thông số của một mô hình đơn lẻ bất kỳ.

Các mô hình QSAR tiếp tục được kiểm tra bằng các tập đánh giá nội và ngoại. Trên tập đánh giá nội, mô hình kết hợp tiếp tục thực hiện tốt nhất với Q^2_{F1} = 0,83; Q^2_{F2} = 0,83; Q^2_{F3} = 0,81; r^2_m = 0,80; r'^2_m = 0,64; $\overline{r^2_m}$ = 0,72; Δr^2_m = 0,16 và CCC = 0,90; trong khi mô hình hồi quy lại không đạt hai điều kiện về $\overline{r^2_m}$ (< 0,65) và Δr^2_m (> 0,2) (**Bảng 3.7**). Còn trên tập đánh giá ngoại, mô hình tốt nhất lại là SVM (**Bảng 3.8**), mặc dù hệ số tương quan bình phương của mô hình này (R^2 = 0,74) chỉ cao hơn mô hình mạng nơron (R^2 = 0,73) khi xét trên tập huấn luyện đa dạng; còn mô hình kém nhất là CHAID khi chỉ đạt các điều kiện về r^2_m, r'^2_m (cùng > 0,5) và Δr^2_m (< 0,2) (**Bảng 3.6**). Trong trường hợp này, mô hình kết hợp cho thấy các kết quả dự đoán có thể so sánh với mô hình tốt nhất là SVM với Q^2_{F1} = 0,83; Q^2_{F2} = 0,82; Q^2_{F3} = 0,83; r^2_m = 0,82; r'^2_m = 0,67; $\overline{r^2_m}$ = 0,74; Δr^2_m = 0,16 và CCC = 0,90 (**Bảng 3.8**). Các kết quả này một lần nữa giúp chứng minh lợi ích của việc kết hợp nhiều mô hình trong việc giải quyết các vấn đề đa dạng phát sinh khi thực hiện mô hình hóa dữ liệu sinh học.

Bảng 3.6. Kết quả đánh giá nội các mô hình dự đoán được tạo ra từ tập huấn luyện đa dạng.

Thông số	CHAID	C&R Tree	Hồi quy	SVM	Mạng nơron	Ensemble
R^2	0,86	0,83	0,75	0,74	0,73	0,84
Q^2_{LOO}	0,50	0,52	0,67	0,72	0,68	0,70
$\lvert R^2 - Q^2_{LOO}\rvert$	0,36	0,30	0,08	0,02	0,04	0,14
r^2_m	0,86	0,83	0,75	0,73	0,69	0,80
r'^2_m	0,73	0,67	0,54	0,55	0,54	0,64
$\overline{r^2_m}$	0,79	0,75	0,65	0,64	0,62	0,72
Δr^2_m	0,13	0,16	0,21	0,18	0,15	0,16
R^2_r	0,08	0,28	0,11	0,14	0,01	0,29
R^2_p	0,76	0,61	0,61	0,58	0,61	0,62

Bảng 3.7. Kết quả đánh giá các mô hình dự đoán trên tập đánh giá nội.

Thông số	CHAID	C&R Tree	Hồi quy	SVM	Mạng nơron	Ensemble
Q^2_{F1}	0,79	0,81	0,77	0,81	0,75	0,83
Q^2_{F2}	0,79	0,81	0,77	0,81	0,75	0,83
Q^2_{F3}	0,77	0,80	0,75	0,80	0,73	0,81
r^2_m	0,74	0,80	0,74	0,80	0,71	0,80
r'^2_m	0,68	0,63	0,54	0,63	0,60	0,64
$\overline{r^2_m}$	0,71	0,72	0,64	0,71	0,66	0,72
Δr^2_m	0,06	0,17	0,21	0,17	0,11	0,16
CCC	0,89	0,89	0,86	0,89	0,86	0,90

Bảng 3.8. Kết quả đánh giá truyền thống các mô hình dự đoán trên tập đánh giá ngoại.

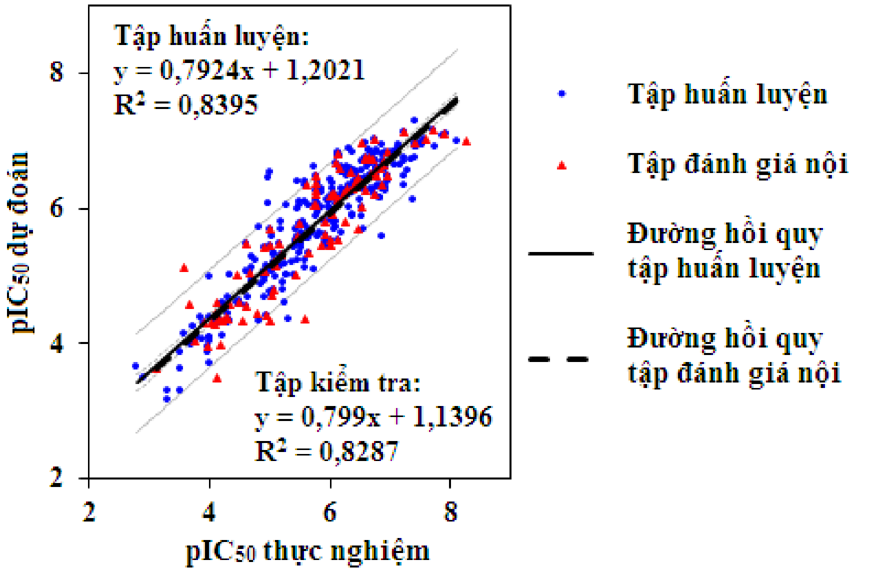
Thông số	CHAID	C&R Tree	Hồi quy	SVM	Mạng nơron	Ensemble
Q^2_{F1}	0,68	0,75	0,81	0,85	0,78	0,83
Q^2_{F2}	0,68	0,75	0,81	0,85	0,78	0,82
Q^2_{F3}	0,69	0,76	0,81	0,85	0,79	0,83
r^2_m	0,61	0,67	0,77	0,83	0,75	0,82
r'^2_m	0,58	0,68	0,59	0,68	0,62	0,67
$\overline{r^2_m}$	0,59	0,67	0,68	0,75	0,69	0,74
Δr^2_m	0,02	0,01	0,18	0,14	0,13	0,16
CCC	0,84	0,88	0,89	0,91	0,88	0,90

Trong nghiên cứu này, quá trình xác định phạm vi khả năng ứng dụng phát hiện ra 7/300 chất của tập huấn luyện đa dạng (89, 288, 381, 411, 422, 433 và 472) là các chất chất lạ; 1/100 chất của tập đánh giá nội (358) và 2/99 chất của tập đánh giá ngoại (40, 407) là các chất nằm ngoài phạm vi khả năng ứng dụng (**tập tin TLBS.xlsx, Sheet3** và **bản in Tài liệu bổ sung, TLBS3**). Với sự hiện diện của các chất này, hai mô hình đơn lẻ (hồi quy, SVM) và mô hình kết hợp được phân loại là "tốt" và ba mô hình đơn lẻ khác (CHAID, C&R Tree và mạng nơron) được phân loại là "trung bình" theo các thông số dựa vào MAE được ước tính sau khi loại bỏ 5 % chất trong tập đánh giá ngoại có các giá trị sai số cao (**Bảng 3.9**; **tập tin TLBS.xlsx, Sheet7**). Các vấn đề liên quan đến các thông số dựa vào R^2 cổ điển và CCC (các giá trị của chúng bị ảnh hưởng bởi phạm vi và sự phân bố của các giá trị biến phụ thuộc của các chất trong tập đánh giá xung quanh giá trị trung bình của tập huấn luyện/tập đánh giá) có thể dẫn đến kết luận sai về khả năng chấp nhận mô hình [152]. Các điều kiện dựa vào MAE do Roy và cộng sự đề xuất [152] giúp kiểm tra các giá trị sai số dự đoán bằng cách xác định giới hạn sai số cho phép, sử dụng phạm vi biến phụ thuộc của tập huấn luyện. Kết quả xác định phạm vi khả năng ứng dụng và kiểm tra các điều kiện dựa vào MAE đã cho thấy các chất của tập đánh giá nội và ngoại trong nghiên cứu này đều nằm trong phạm vi hóa học và đáp ứng của tập huấn luyện. Nói cách khác, khả năng dự đoán "tốt" của các mô hình QSAR xây dựng được xác nhận với độ tin cậy cao hơn khi kết hợp các phương pháp đánh giá truyền thống và đánh giá mới.

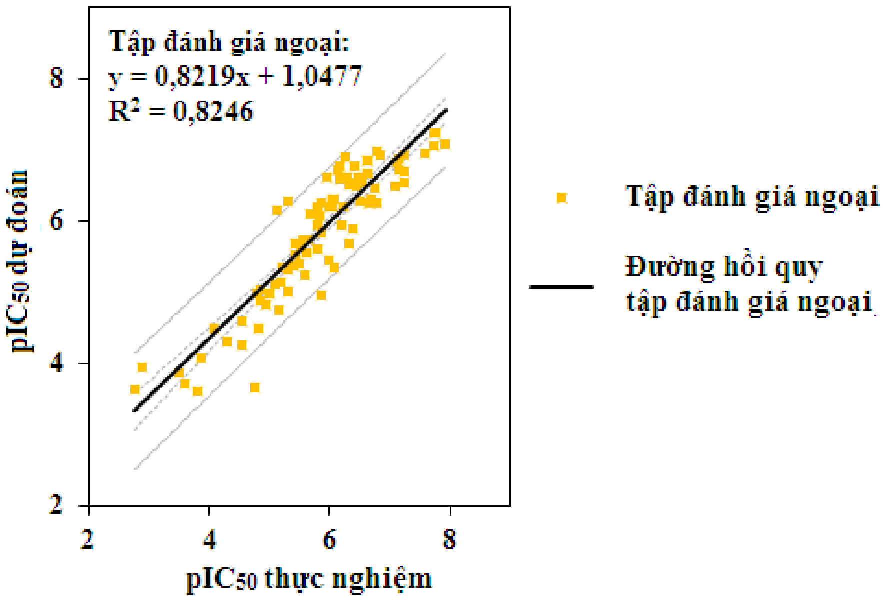
Bảng 3.9. Kết quả đánh giá các mô hình dự đoán trên tập đánh giá ngoại, sử dụng các điều kiện dựa trên MAE áp dụng cho 95 % dữ liệu.

Thông số	**CHAID**	**C&R Tree**	**Hồi quy**	**SVM**	**Mạng nơron**	**Ensemble**
MAE	0,37	0,38	0,33	0,29	0,36	0,31
MAE+3*SD	1,19	1,10	0,98	0,90	1,11	0,96
Chất lượng dự đoán	Trung bình	Trung bình	Tốt	Tốt	Trung bình	Tốt

Bằng phân tích định lượng, Gramatica và Chirico [27] đã nhấn mạnh tầm quan trọng của việc đánh giá đồ thị phân tán của dữ liệu thực nghiệm và dữ liệu dự đoán, giúp phát hiện ra những mô hình QSAR không thể chấp nhận được chỉ với các giá trị thống kê tốt. Các đồ thị hồi quy giá trị pIC_{50} dự đoán bằng giá trị pIC_{50} quan sát được trên bơm ngược P-gp được biểu diễn cho mô hình kết hợp trong trường hợp tập huấn luyện và tập đánh giá nội (**Hình 3.1A**) và tập đánh giá ngoại (**Hình 3.1B**). Các điểm dữ liệu nằm dọc theo đường hồi quy cho thấy sự tương quan tốt giữa các giá trị thực nghiệm và dự đoán trên tất cả các tập dữ liệu. Nói cách khác, mô hình kết hợp được tạo ra là phù hợp tốt với các dữ liệu này.



A



B

Hình 3.1. Đồ thị phân tán của mô hình kết hợp trên các tập dữ liệu: (A) Trên tập huấn luyện và tập đánh giá nội; (B) Trên tập đánh giá ngoại.

3.1.3. Bản đồ nhận thức về sự ức chế bơm ngược qua trung gian P-gp và NorA

Trước đó, Carosati và cộng sự [17] và Vishwakarma và cộng sự [77] đã xem xét sự hỗn tạp phối tử giữa các bơm ngược P-gp ở người và NorA ở *S. aureus* và tiết lộ có một sự chồng phủ giữa các chất ức chế hai loại bơm ngược này. Do số lượng hạn chế của các chất được thu thập từ tài liệu mà sẵn có thông tin hoạt tính trên cả hai protein [17], [60], [77], [79], [82], [86], [94], [114], [120], [162], các bản đồ nhận thức được phát triển thay cho các mô hình phân loại đa lớp đã được xây dựng trước đó [180] để làm rõ các tính chất cần thiết cho sự ức chế ít nhất một trong số hai protein chuyên chở này.

Quá trình tính toán và lựa chọn thuộc tính mô tả ở trên được áp dụng cho tập dữ liệu nhỏ 54 chất. Sự lựa chọn biến không bao gồm kỹ thuật tối ưu hóa bằng thuật toán di truyền đã thu được 18 thông số mô tả lý hóa và 3 dấu vân tay từ 2977 thông số mô tả và dấu vân tay được tính toán ban đầu. Trong số đó, 5 thông số mô tả MOE biểu thị các tính chất ma trận gần kề và cách xa (4 thông số) và điện tích từng phần (1 thông số), trong khi 13 thông số mô tả PaDEL biểu thị các tính chất tự tương quan (6 thông số), ma trận Barysz (1 thông số), con đường chi (2 thông số), trạng thái điện tử topo loại nguyên tử (2 thông số), nguyên tử hóa học topo mở rộng (1 thông số) và đường biên khoảng cách phân tử (1 thông số) của các chất trong tập dữ liệu. Cả hai loại thông số mô tả được chia tỷ lệ trong khoảng [0, 1] trước khi xây dựng các bản đồ. Trong 3 dấu vân tay được chọn, có 2 chuỗi MACCS, 1 chuỗi Pubchem và không có chuỗi dưới cấu trúc nào. Tất cả các thuộc tính và cách viết tắt của chúng được liệt kê và trình bày trong **Phụ lục 6**.

Tiếp theo, kỹ thuật đo lường đa hướng MDS ALSCAL được sử dụng trong trường hợp các thông số mô tả (thang đo khoảng) và kỹ thuật phân tích tương hợp CA được sử dụng trong trường hợp các dấu vân tay nhị phân (thang đo danh nghĩa) để thực hiện lập bản đồ nhận thức. Với các điều kiện độ hội tụ S-stress bằng 0,001; giá trị Stress tối thiểu bằng 0,005 và số vòng lặp tối đa bằng 30, quá trình MDS đã dừng lại ở vòng lặp thứ 4 trong trường hợp các thông số mô tả và ở vòng lặp thứ 3 trong trường hợp các lớp hoạt tính, bởi vì mức độ cải thiện S-stress nhỏ hơn 0,001. Các giá trị Stress nhỏ hơn 0,1 và các giá trị hệ số tương quan bình phương (squared correlation - RSQ) lớn hơn 0,9 (**Bảng 3.10**) cho thấy mối tương quan mạnh giữa các khác biệt và các khoảng cách. Tất cả các giá trị phương sai trong quá trình giảm hướng CA được liệt kê trong **Bảng 3.11**. Phương sai tổng cộng bằng 0,125 phản ánh

12,5 % khác biệt được giải thích bởi hai hướng hoặc ba dấu vân tay hoặc bốn lớp hoạt tính.

Bảng 3.10. Các giá trị thống kê trong quá trình chia tỷ lệ, sử dụng kỹ thuật đo lường đa hướng (MDS ALSCAL).

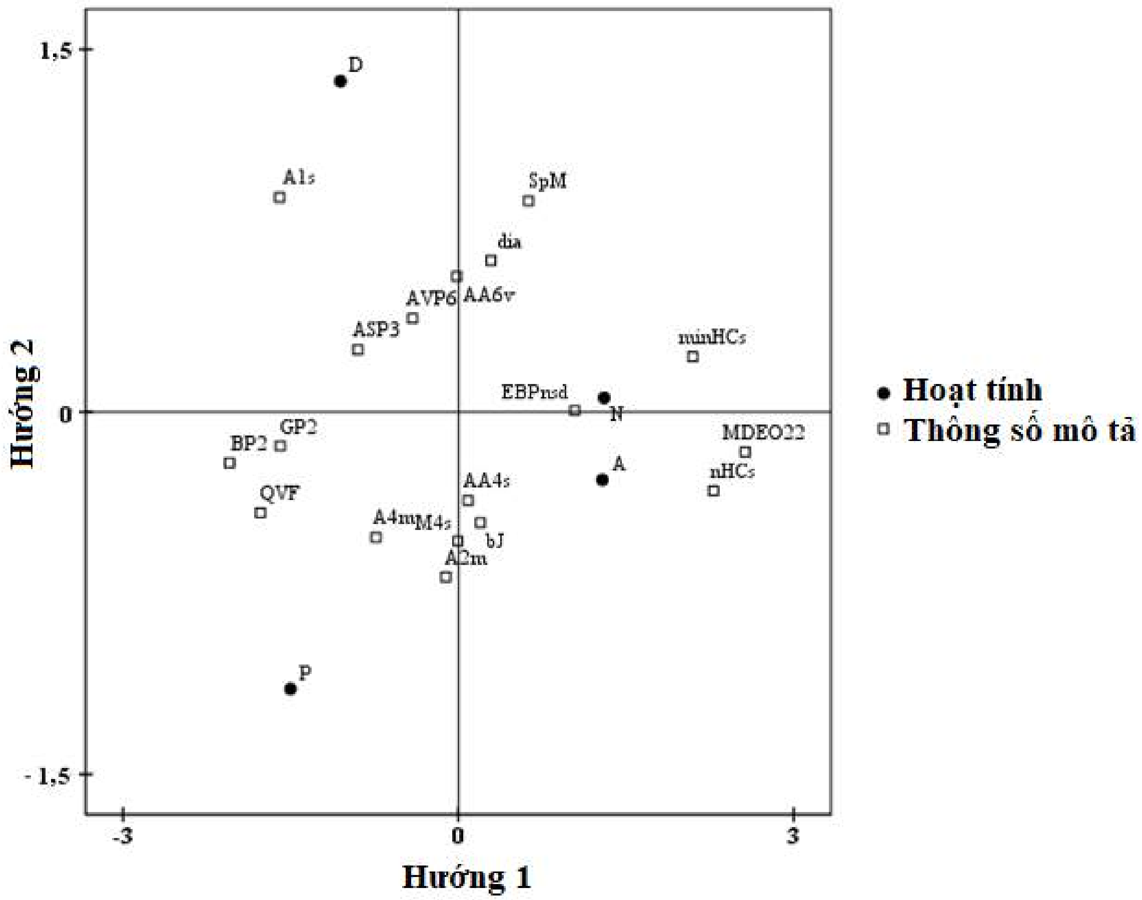
	Lịch sử vòng lặp			**Stress và tương quan bình phương (RSQ) trong các khoảng cách cho ma trận**	
	Vòng lặp	**S-stress[a]**	**Sự cải thiện**	**Stress[b]**	**RSQ[c]**
Giữa các biến (các thông số mô tả)	1	0,09731		0,06562	0,98219
	2	0,07076	0,02656		
	3	0,06776	0,00300		
	4	0,06747	0,00029		
Giữa các trường hợp (các lớp hoạt tính)	1	0,05764		0,09656	0,94377
	2	0,05516	0,00248		
	3	0,05513	0,00003		

[a]Công thức S-stress 1 của Young được sử dụng.
[b]Các giá trị Stress là công thức stress 1 của Kruskal.
[C]Các giá trị RSQ (squared correlation) là tỷ lệ của phương sai của dữ liệu được chia tỷ lệ (các khác biệt) trong sự phân chia (hàng, ma trận, hoặc toàn bộ dữ liệu), được giải thích bởi các khoảng cách tương ứng của chúng.

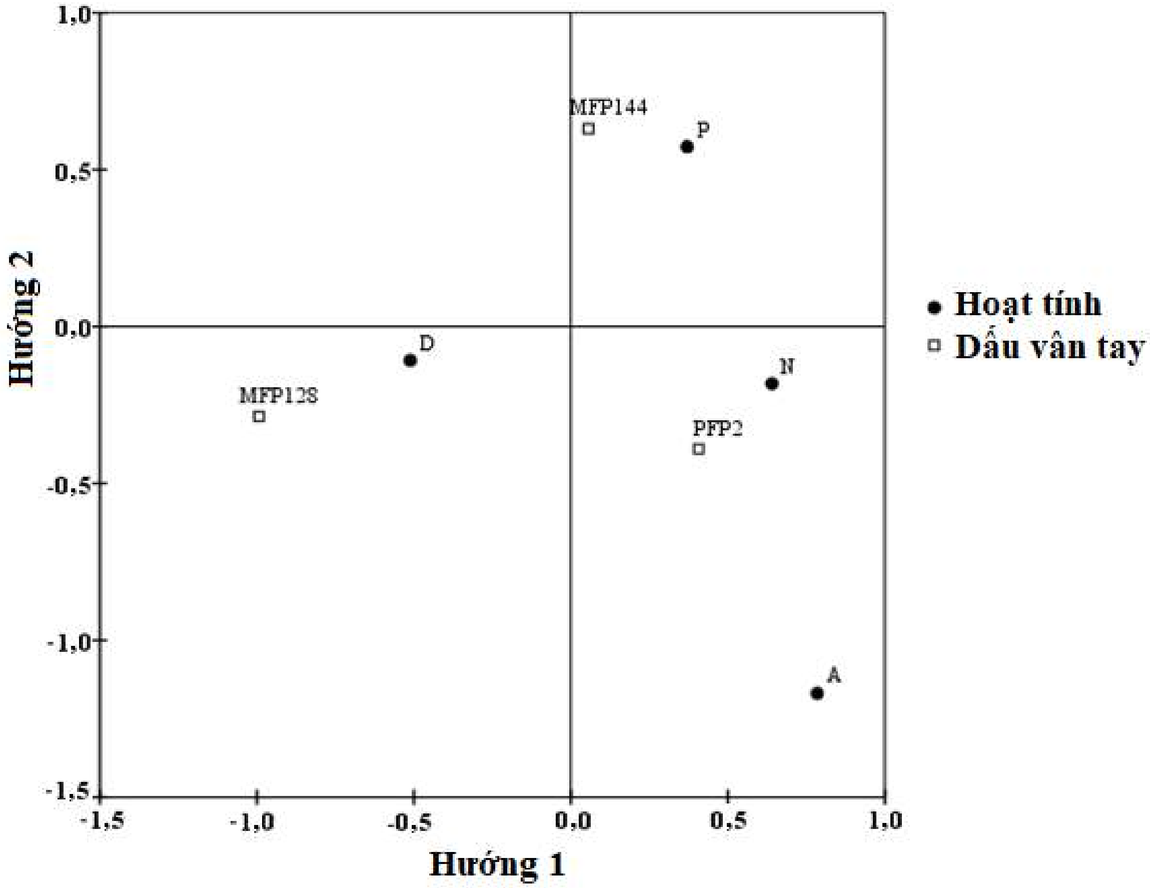
Bảng 3.11. Giá trị phương sai của các hướng, các dấu vân tay và các lớp hoạt tính trong quá trình giảm hướng, sử dụng kỹ thuật phân tích tương hợp (CA).

Hướng	**Phương sai**	**Dấu vân tay**	**Phương sai**	**Hoạt tính**	**Phương sai**
1	0,073	MACCSFP128	0,057	A	0,041
2	0,051	MACCSFP144	0,033	D	0,036
Tổng	0,125	PubchemFP2	0,035	N	0,014
		Tổng	0,125	P	0,034
				Tổng	0,125

Các bản đồ nhận thức MDS và CA của bốn trường hợp đại diện cho các lớp hoạt tính (P, A, D, N) và 21 biến được chọn đại diện cho các thông số mô tả và các dấu vân tay lần lượt được thể hiện trong **Hình 3.2** và **Hình 3.3**. Trong bản đồ MDS, các điểm đại diện cho các trường hợp được định vị ở bốn góc phần tư khác nhau và các điểm đại diện cho các biến phân tán xung quanh điểm giao nhau của hai trục. Trong bản đồ CA, các điểm của các trường hợp và các biến cũng được phân bố theo cách tương tự nhưng để trống góc phần tư phía trên bên trái. Hai điểm A và N nằm gần nhau trong cả hai bản đồ ngụ ý là các chất ức chế NorA nhưng không ức chế P-gp và các chất không ức chế cả hai protein là khá giống nhau ở các tính chất được xem xét.



Hình 3.2. Bản đồ nhận thức đo lường đa hướng (MDS) của các lớp hoạt tính và các thông số mô tả. P: Chất ức chế chỉ P-gp; A: Chất ức chế chỉ NorA; D: Chất ức chế cả P-gp và NorA; N: Chất không ức chế cả P-gp và NorA; dia: diameter; BP2: BCUT_PEOE_2; GP2: GCUT_PEOE_2; bJ: balabanJ; QVF: Q_VSA_FNEG; A2m: ATSC2m; A4m: ATSC4m; A1s: ATSC1s; AA6v: AATSC6v; AA4s: AATSC4s; M4s: MATS4s; SpM: SpMAD_DzZ; ASP3: ASP-3; AVP6: AVP-6; nHCs: nHCsatu; minHCs: minHCsatu; EBPnsd: ETA_BetaP_ns_d; MDEO22: MDEO-22.



Hình 3.3. Bản đồ nhận thức phân tích tương hợp (CA) của các lớp hoạt tính và các dấu vân tay. P: Chất ức chế chỉ P-gp; A: Chất ức chế chỉ NorA; D: Chất ức chế cả P-gp và NorA; N: Chất không ức chế cả P-gp và NorA; MFP128: MACCSFP128; MFP144: MACCSFP144; PFP2: PubchemFP2.

3.1.4. Các mô hình pharmacophore

Trước khi chạy ứng dụng Pharmacophore Elucidation trong MOE, các hình thể ba chiều (3D) của các chất nghiên cứu được tạo ra sử dụng một ứng dụng khác cũng của MOE là Conformation Import [111]. Các thiết lập mặc định được giữ nguyên khi xây dựng các mô hình pharmacophore, cụ thể là các hình thể đầu vào được phân cụm để loại bỏ hình thể trùng; kế hoạch chú thích "thống nhất/unified" được sử dụng để xác định các yếu tố pharmacophore của mỗi phân tử; không thiết lập ràng buộc về tần số xuất hiện của biểu lộ trong các truy vấn pharmacophore được tạo ra; bán kính của các biểu lộ yếu tố truy vấn được thiết lập bằng 1,4 Å; số lượng các phân tử mà một truy vấn tạo ra phải phù hợp được thiết lập bằng 0,9; số lượng các yếu tố được giới hạn bằng 5 và không cho phép bất kỳ lựa chọn gióng hàng nào. Từ các pharmacophore xây dựng được, các phân tử "hit" dễ dàng được xác định thông qua các hình thể "hit" trong quá trình tìm kiếm pharmacophore.

3.1.4.1. Các mô hình pharmacophore cho hoạt tính ức chế P-gp mạnh

Từ cơ sở dữ liệu đầu vào chứa bốn phân tử hoạt tính là aripiprazol, ebastin, tariquidar và elacridar, tổng cộng 39 truy vấn pharmacophore được tạo ra sử dụng công cụ Pharmacophore Elucidation. Ba truy vấn tốt nhất cùng với các thông số đánh giá của chúng được trình bày trong **Bảng 3.12** theo thứ tự điểm số chồng phủ giảm dần. Các truy vấn này có cùng số lượng và loại yếu tố, bao gồm ba nhóm kỵ nước (H) và một nhóm nhận liên kết hydro (a). Santos và cộng sự [54] cũng đã công bố một mô hình pharmacophore bốn điểm giống với nghiên cứu này, dựa trên các diterpen vòng lớn. Tầm quan trọng của tính kỵ nước và các nhóm nhận điện tử như là các yếu tố cần thiết cho sự ức chế P-gp hiệu quả cũng được nhấn mạnh trong nhiều nghiên cứu trước [17], [20], [32], [48], [49], [54], [56], [89], [95], [131], [132], [133], [134], [135], [149], [219].

Bảng 3.12. Ba giả thuyết pharmacophore tốt nhất cho các chất ức chế P-gp mạnh và các chất ức chế chọn lọc NorA cùng với các giá trị thống kê của chúng.

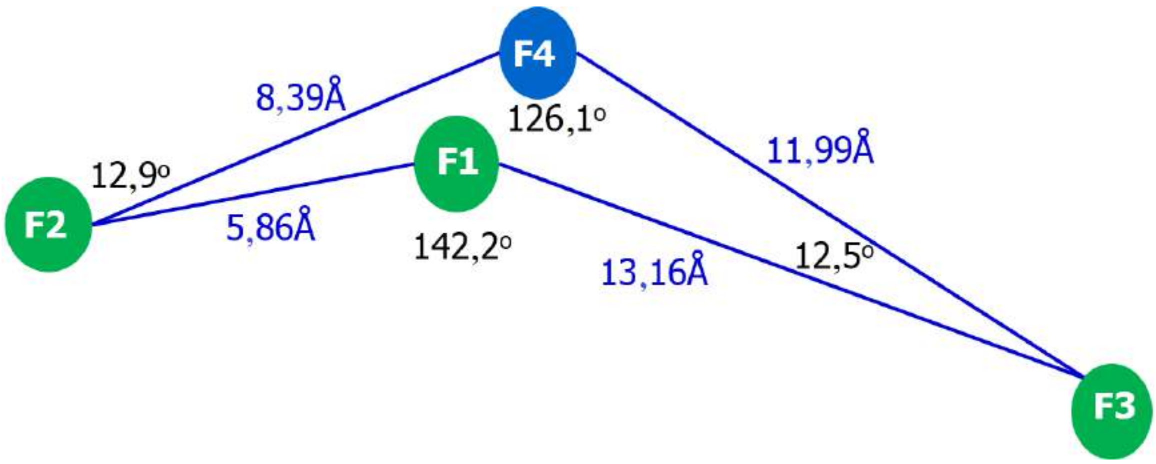
	STT	Ph4	Độ bao phủ	Độ chồng phủ	Độ đúng	Tìm kiếm Ph4			
						P	A	D	N
Chất ức chế P-gp mạnh	1	HHHa	4	2,60	1	5/19	4/7	13/19	3/9
	2	HHHa	4	2,59	1	0/19	0/7	10/19	0/9
		HHHaV	4	2,59	1	0/19	0/7	10/19	0/9
	3	HHHa	4	2,58	1	0/19	0/7	9/19	0/9
Chất ức chế chọn lọc NorA	1	RRHa	3	2,27	1	1/19	5/7	6/19	4/9
	2	RHHa	3	2,21	1	7/19	6/7	12/19	6/9
	3	RRHd	3	2,21	1	2/19	6/7	1/19	3/9
		RRHdV	3	2,21	1	2/19	6/7	0/19	3/9

*Ph4: Pharmacophore; H: Yếu tố kỵ nước; a: Nhóm nhận liên kết hydro; R: Yếu tố vòng thơm/vòng Pi; d: Nhóm cho liên kết hydro; V: Thể tích bên ngoài.

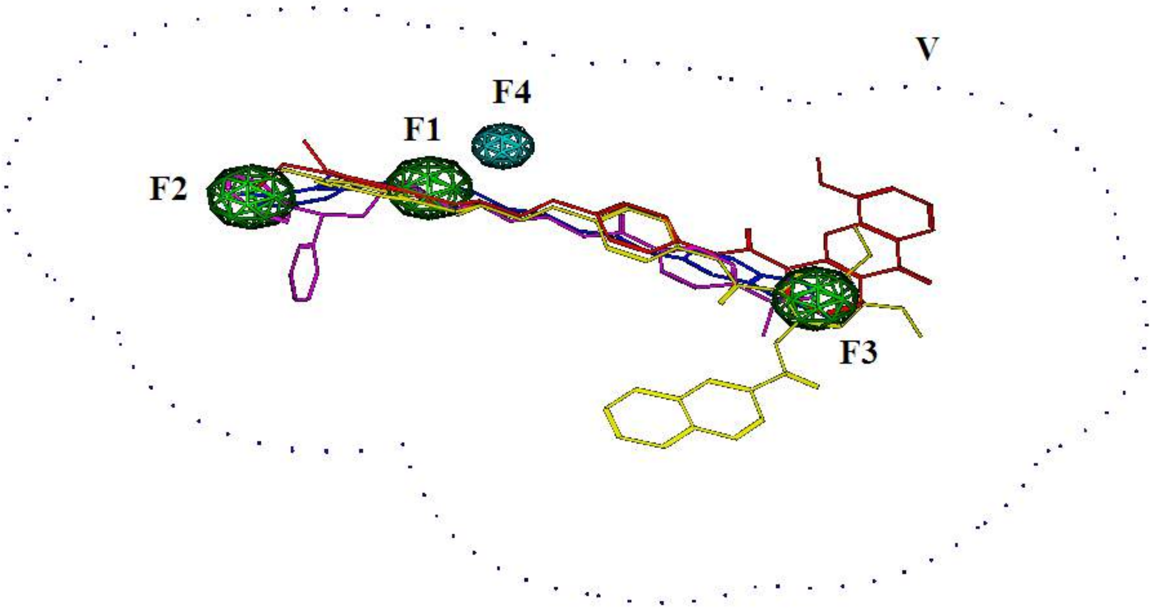
Khi áp dụng để tìm kiếm cho tập dữ liệu 54 chất đã được sử dụng để xây dựng bản đồ nhận thức ở trên, truy vấn số 2 đã cho thấy kết quả tìm kiếm tốt nhất so với hai truy vấn còn lại khi nhận dạng được 10/38 chất có hoạt tính (các chất D), trong đó có 4 chất được sử dụng để mô hình hóa và không nhận dạng các chất không có hoạt tính (**Bảng 3.12**). Việc không có bất kỳ chất P nào (hầu hết không phải là chất ức chế P-gp mạnh) thỏa truy vấn số 2 đã chứng minh đây là mô hình gắn kết phù hợp giữa các chất có hoạt tính mạnh và thụ thể. Mặc dù nhận dạng được nhiều chất có hoạt tính hơn (18/38 chất P và D), truy vấn số 1 cũng đã nhận dạng không chọn lọc 7 chất không có hoạt tính (các chất A và N). Trong khi không nhận dạng các chất không có hoạt tính như truy vấn số 2, truy vấn số 3 lại chỉ nhận dạng được ít chất có hoạt tính hơn (9/38 chất D, trừ astemizol). Các khoảng cách và các góc trong không gian được tạo thành từ các điểm của truy vấn số 2 được biểu diễn trong **Hình 3.4A**.

Trên truy vấn số 2, ràng buộc về thể tích (V) bao gồm một số khối cầu có tâm đặt tại các nguyên tử nặng được thêm vào. Loại thể tích được thiết lập là "bên ngoài/exterior" và bán kính của nó được điều chỉnh bằng 6,5 Å. Truy vấn số 2 mới chặt chẽ hơn (HHHaV) vẫn cho kết quả nhận dạng tương tự là 10/38 chất D khi được sử dụng để lặp lại việc tìm kiếm trên tập 54 chất nói trên. Hình ảnh của truy vấn và các hình thể tốt nhất (có RMSD thấp nhất) của bốn phân tử sử dụng để mô hình hóa pharmacophore được biểu diễn trong **Hình 3.4B**.

Trong quá trình đánh giá tiếp theo, mô hình pharmacophore mới xây dựng (HHHaV) được áp dụng để tìm kiếm trên hai tập dữ liệu hình thể thu được từ 2134 chất và 22 chalcon mà đã được sử dụng để phát triển các mô hình phân loại chất ức chế và chất không ức chế P-gp ở trên. Trong số 1028 chất tính được hình thể (thuộc tập 2134 chất) bao gồm 485 chất ức chế và 543 chất không ức chế, có 38 chất thỏa mãn mô hình pharmacophore bao gồm 34 chất ức chế và 4 chất không ức chế (**tập tin TLBS.xlsx, Sheet8**). Trong số 22 chalcon, mô hình pharmacophore đã nhận dạng thành công 6 chất có hoạt tính mạnh là các chất 12, 13, 20, 21, 22 và 23. Ba chalcon 18, 19 và 24 không ức chế P-gp (IC_{50} > 15 µM) đã không thỏa truy vấn. Các kết quả thu được giúp chứng minh mô hình pharmacophore cho hoạt tính ức chế P-gp mạnh có thể được sử dụng để sàng lọc ảo nhằm tìm kiếm các chất ức chế P-gp mới và hiệu quả.



A



B

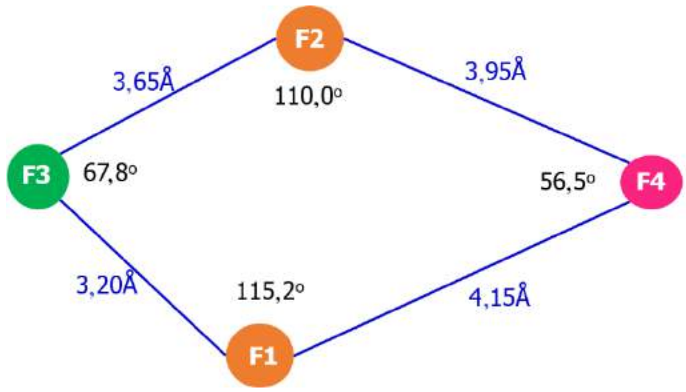
Hình 3.4. Mô hình pharmacophore chất ức chế P-gp mạnh (F1, F2, F3: Nhóm kỵ nước; F4: Nhóm nhận liên kết hydro; V: Giới hạn thể tích): (A) Các khoảng cách và góc; (B) Với sự hiện diện của các chất có hoạt tính (aripiprazol, ebastin, tariquidar và elacridar).

3.1.4.2. Các mô hình pharmacophore cho hoạt tính ức chế NorA nhưng không ức chế P-gp (sự ức chế chọn lọc NorA)

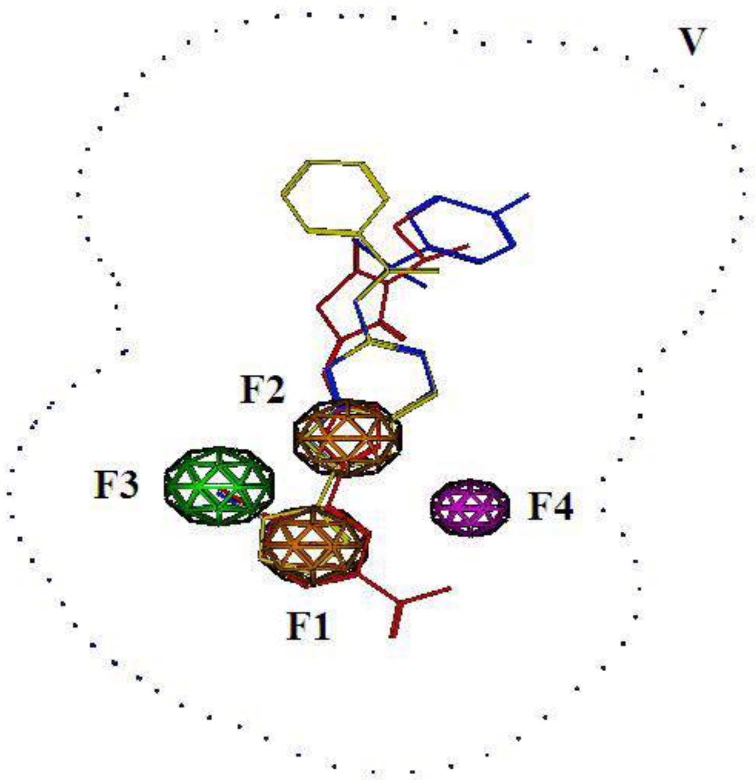
Từ cơ sở dữ liệu đầu vào chứa ba phân tử hoạt tính là các chất 22, 23 và 32 [17], tổng cộng 18 truy vấn pharmacophore được tạo ra sử dụng công cụ Pharmacophore Elucidation. Ba truy vấn tốt nhất cùng với các thông số đánh giá của chúng được trình bày trong **Bảng 3.12** theo thứ tự điểm số chồng phủ giảm dần. Các truy vấn này có cùng số lượng yếu tố nhưng khác nhau về loại yếu tố: Truy vấn số 1 chứa hai vòng thơm, một nhóm kỵ nước và một nhóm nhận liên kết hydro (RRHa); truy vấn số 2 chứa một vòng thơm, hai nhóm kỵ nước và một nhóm nhận liên kết hydro (RHHa) và truy vấn số 3 chứa hai vòng thơm, một nhóm kỵ nước và một nhóm cho liên kết hydro (RRHd). Sự hiện diện của các yếu tố vòng thơm, kỵ nước và nhóm nhận liên kết hydro trong các giả thuyết pharmacophore này là phù hợp với các nghiên cứu của Dupree [46] và Carosati và cộng sự [17].

Khi áp dụng để tìm kiếm cho tập dữ liệu 54 chất như các pharmacophore cho hoạt tính ức chế P-gp mạnh, truy vấn số 3 đã cho thấy kết quả tìm kiếm tốt nhất so với hai truy vấn còn lại khi nhận dạng được 6/7 chất có hoạt tính (các chất A), trong đó có 3 chất được sử dụng để mô hình hóa và cả 6 chất không có hoạt tính (các chất P, D và N) (**Bảng 3.12**). Các truy vấn số 1 và 2 không chỉ nhận dạng được số chất có hoạt tính ít hơn hoặc bằng truy vấn số 3 mà còn nhận dạng không chọn lọc nhiều chất không có hoạt tính hơn (lần lượt là 11 và 25 chất không có hoạt tính thuộc các lớp P, D và N). Các khoảng cách và các góc trong không gian được tạo thành từ các điểm của truy vấn số 3 được biểu diễn trong **Hình 3.5A**.

Trên truy vấn số 3, ràng buộc về thể tích (V) bao gồm một số khối cầu có tâm đặt tại các nguyên tử nặng cũng được thêm vào. Loại thể tích được thiết lập là "bên ngoài/exterior" và bán kính của nó được điều chỉnh bằng 5 Å. Truy vấn số 3 mới chặt chẽ hơn (RRHdV) cho kết quả nhận dạng tốt hơn với số chất có hoạt tính thỏa mãn truy vấn được giữ nguyên trong khi số chất không có hoạt tính phù hợp với truy vấn giảm xuống còn năm chất (trừ astemizol) khi được sử dụng để lặp lại việc tìm kiếm trên tập 54 chất nói trên. Hình ảnh của truy vấn và các hình thể tốt nhất (có RMSD thấp nhất) của ba phân tử sử dụng để mô hình hóa pharmacophore được biểu diễn trong **Hình 3.5B**.



A



B

Hình 3.5. Mô hình pharmacophore chất ức chế NorA nhưng không ức chế P-gp (F1, F2: Yếu tố vòng thơm/vòng Pi; F3: Nhóm kỵ nước; F4: Nhóm cho liên kết hydro; V: Giới hạn thể tích): (A) Các khoảng cách và góc; (B) Với sự hiện diện của các chất có hoạt tính (20, 21, 30).

Sau đó, mô hình pharmacophore mới xây dựng (RRHdV) được áp dụng để lặp lại việc tìm kiếm đã được thực hiện trước đó trên hai tập dữ liệu hình thể của 1028/2134 chất và 22 chalcon. Trong trường hợp của cơ sở dữ liệu lớn hơn, 81/485 chất ức chế và 75/543 chất không ức chế P-gp thỏa mãn mô hình pharmacophore (**tập tin TLBS.xlsx, Sheet8**). Trong trường hợp của cơ sở dữ liệu nhỏ hơn, tất cả chalcon bao gồm 19 chất ức chế P-gp mạnh ($IC_{50} \leq 15$ μM) đã không thỏa truy vấn. Các kết quả thu được giúp chứng minh mô hình pharmacophore cho hoạt tính ức chế NorA nhưng không ức chế P-gp có thể được sử dụng để sàng lọc ảo nhằm tìm kiếm các chất ức chế NorA mới và an toàn.

3.2. Các mô hình máy tính dựa trên cấu trúc (mô hình tương đồng của P-gp)

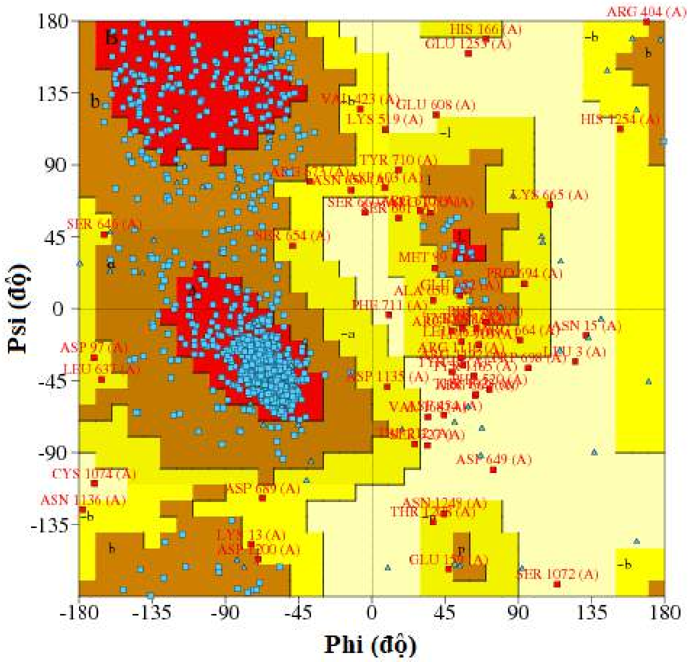
Kết quả mô hình hóa phân tử của server tự động I-TASSER thu được bốn mô hình tương đồng của P-gp cho mục đích docking phân tử trong bước tiếp theo; với chuỗi A của bốn protein có mã số định danh trong ngân hàng dữ liệu protein là 3g61 [4], 4m1m [97], 4f4c [75] và 3g5u [4] được chọn làm các đĩa cấu trúc theo thứ tự. Các thông số ước tính chất lượng mô hình được trình bày trong **Bảng 3.13**, trong đó chỉ mô hình đầu tiên (mô hình 1) có các giá trị TM-score và RMSD được dự đoán. Bởi vì sự tương quan của C-score và chất lượng của các mô hình được xếp hạng thấp hơn là yếu hơn nhiều so với mô hình đầu tiên, chất lượng tuyệt đối (TM-score và RMSD) của các mô hình này không thể được ước tính một cách có ý nghĩa. Tuy nhiên, chất lượng tương đối của các mô hình xếp hạng thấp hơn có thể được dự đoán dựa trên thứ hạng tương đối và thông tin C-score của chúng [151].

Bảng 3.13. Bốn mô hình tương đồng tốt nhất của P-gp được dự đoán bởi I-TASSER với thông tin đĩa và các thông số ước tính.

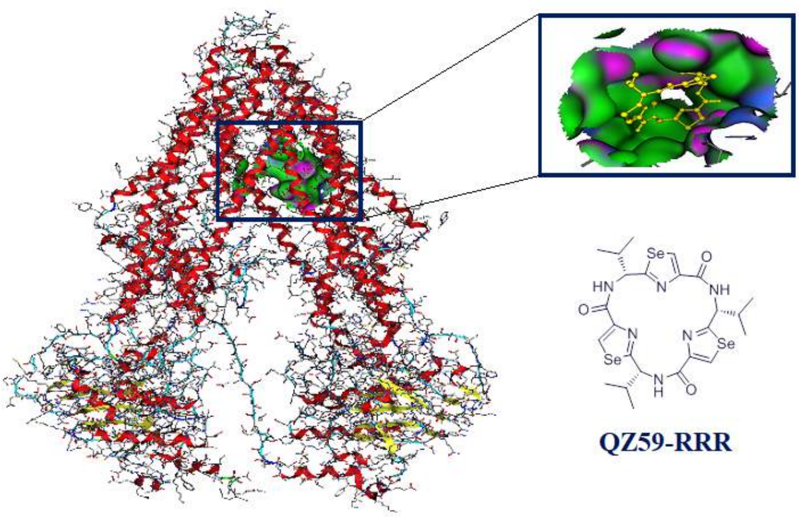
Mô hình	Đĩa sử dụng	C-score	TM-score	RMSD	Số lượng decoy	Mật độ đám
1	3g61A	0,58	0,79 ± 0,09	8,1 ± 4,4 Å	1850	0,17
2	4m1mA	0,96	-	-	1200	0,24
3	4f4cA	0,38	-	-	1212	0,14
4	3g5uA	-2,11	-	-	84	0,01

Với chất lượng được đánh giá là tốt nhất theo các chỉ số của I-TASSER (C-score = 0,58; TM-score = 0,79; RMSD = 8,1; số lượng decoy = 1850 và mật độ đám

= 0,17), mô hình 1 dự đoán từ chuỗi A của P-gp ở chuột với mã là 3g61 [4] được tiếp tục sử dụng để tạo ra đồ thị Ramachandran. Trong đồ thị này (**Hình 3.6**), các vùng được ưa thích nhất có màu đỏ và các vùng ít được ưa thích hơn có màu sáng dần. Theo các thông số thống kê, các vùng được ưu thích nhất [A,B,L]; các vùng cho phép bổ sung [a,b,l,p]; các vùng cho phép hào phóng [~a,~b,~l,~p] và các vùng không cho phép [XX] lần lượt chứa 905 acid amin (78,7 %); 189 acid amin (16,4 %); 35 acid amin (3,0 %) và 21 acid amin (1,8 %) trong tổng số 1150 acid amin không phải là glycin và prolin. Dưới 90 % acid amin được phát hiện trong các vùng lõi có thể được giải thích là do độ phân giải chưa cao của cấu trúc đĩa dùng để xây dựng mô hình tương đồng (4,35 Å). Túi gắn kết của mô hình 1 cũng được dự đoán bởi server cho mục đích docking trong quá trình sàng lọc ảo, dựa trên phức hợp của chuỗi B của protein đĩa được sử dụng với phối tử QZ59-RRR (cyclic-tris-(R)-valineselenazol) (**Hình 3.7**). Kết quả của quá trình mô hình hóa tương đồng là hoàn toàn phù hợp với thông tin cấu trúc được Chang và cộng sự [4] tiết lộ về khoang nội phân tử có kích thước lớn của P-gp chịu trách nhiệm cho khả năng gắn kết thuốc đa dạng.



Hình 3.6. Đồ thị Ramachandran của mô hình tương đồng P-gp tốt nhất, trong đó các vùng được ưa thích nhất (the most favoured regions), các vùng được cho phép thêm (the additional allowed regions), các vùng được cho phép rộng rãi (the generously allowed regions) và các vùng không được cho phép (the disallowed regions) được ký hiệu lần lượt là [A,B,L]; [a,b,l,p]; [~a,~b,~l,~p] và [XX]. Khu vực màu đậm hơn tượng trưng cho kết hợp phi-psi được ưa thích hơn.



Hình 3.7. Mô hình tương đồng tốt nhất của P-gp với vị trí gắn kết phối tử QZ59-RRR (cyclic-tris-(R)-valineselenazol) được dự đoán bởi I-TASSER.

3.3. Sàng lọc *in silico* trên P-gp

Các mô hình học máy (phân loại, dự đoán) và pharmacophore sau khi xây dựng được ứng dụng cho hai tập dữ liệu bao gồm 95 chalcon nội bộ và 6874 hợp chất Ngân hàng Thuốc để sàng lọc các chất ức chế mạnh và an toàn của các bơm ngược MDR là P-gp và NorA, nhằm khôi phục hoạt tính kháng khối u và/hoặc kháng khuẩn của các trị liệu có sẵn hiện tại. Ngoài ra, mô hình tương đồng của P-gp tạo ra ở trên cũng được sử dụng cho nghiên cứu docking các chalcon nội bộ và các hợp chất Ngân hàng Thuốc được chọn.

3.3.1. Phân loại chất ức chế và chất không ức chế P-gp

Dựa trên 24 thông số mô tả và dấu vân tay được trình bày trong **Phụ lục 2**, các chất có đầy đủ giá trị thuộc tính cùng với 2109 chất của toàn bộ tập dữ liệu và 22 chalcon của tập đánh giá ngoại dùng cho mục đích phân loại được kiểm tra nhanh bằng hạch "Anomaly Detection" trong Clementine để phát hiện các chất bất thường. Tỷ lệ của các chất bất thường được thiết lập ở mức 1 %. Ngoài ra, phân tích thành phần chính (PCA) của tất cả các chất này cũng được thực hiện với số lượng tối đa là năm thành phần.

Sau quá trình sàng lọc sơ bộ, 79 chất bất thường được phát hiện từ 6755 hợp chất Ngân hàng Thuốc không bị thiếu thông số mô tả nào (**tập tin TLBS.xlsx, Sheet9**). Sự phân bố của các chất thuộc các cơ sở dữ liệu khác nhau theo hai thành phần chính đầu tiên được thể hiện trong **Hình 3.8**. Đồ thị phân tán đầu tiên (**Hình 3.8A**) cho thấy tính hợp lý của việc loại bỏ các chất bất thường là những chất có phần tách biệt khỏi các chất khác trong không gian hai chiều được chọn. Trong khi đó, hai đồ thị khác (**Hình 3.8B** và **Hình 3.8C**) lại cho thấy sự phân tán của 95 chalcon nội bộ và 6676 hợp chất Ngân hàng Thuốc còn lại nằm trong không gian hóa học của cơ sở dữ liệu 2109 chất được sử dụng để phân loại chất ức chế và chất không ức chế P-gp. Nói cách khác, các mô hình phân loại có nhiều khả năng dự đoán đúng hoạt tính sinh học của các chất được quan tâm.