XÂY DỰNG MÔ HÌNH ML, ANN, GNN VÀ KẾT HỢP DOCKING PHÂN TỬ

Sàng lọc các chất ức chế thụ thể ALK trong điều trị bệnh ung thư phổi không tế bào nhỏ

Sinh viên thực hiện: TRỊNH THẾ CHƯƠNG
GVHD: PSG. TS. TRƯƠNG NGỌC TUYỀN

NỘI DUNG

1 **TỔNG QUAN**
Tổng quan NSCLC, NSCLC dạng ALK dương tính, ML, ANN, GNN, Docking phân tử

2 **ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP**
Dữ liệu xây dựng, quy trình xây dựng các mô hình, và sàng lọc ảo

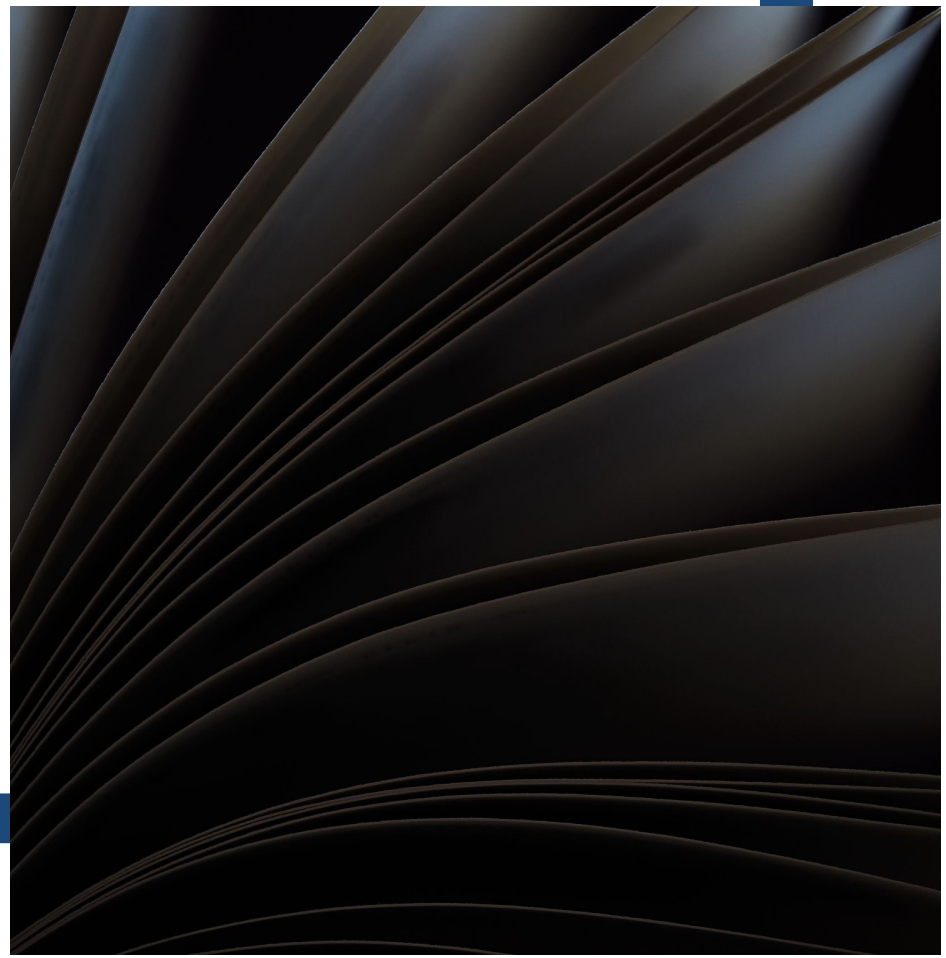
3 **KẾT QUẢ VÀ BÀN LUẬN**
Kết quả đánh giá các mô hình và các ứng viên ức chế ALK tiềm năng.

4 **KẾT LUẬN VÀ ĐỀ NGHỊ**

1. TỔNG QUAN

Tổng quan về ung thư phổi và ung thư phổi không tế bào nhỏ (NSCLC), NSLC dạng ALK dương tính, các thuốc ức chế ALK.

Tổng quan các nghiên cứu *in silico* ứng dụng trí tuệ nhân tạo trong khám phá và thiết kế thuốc nói chung và nhóm ức chế ALK nói riêng.

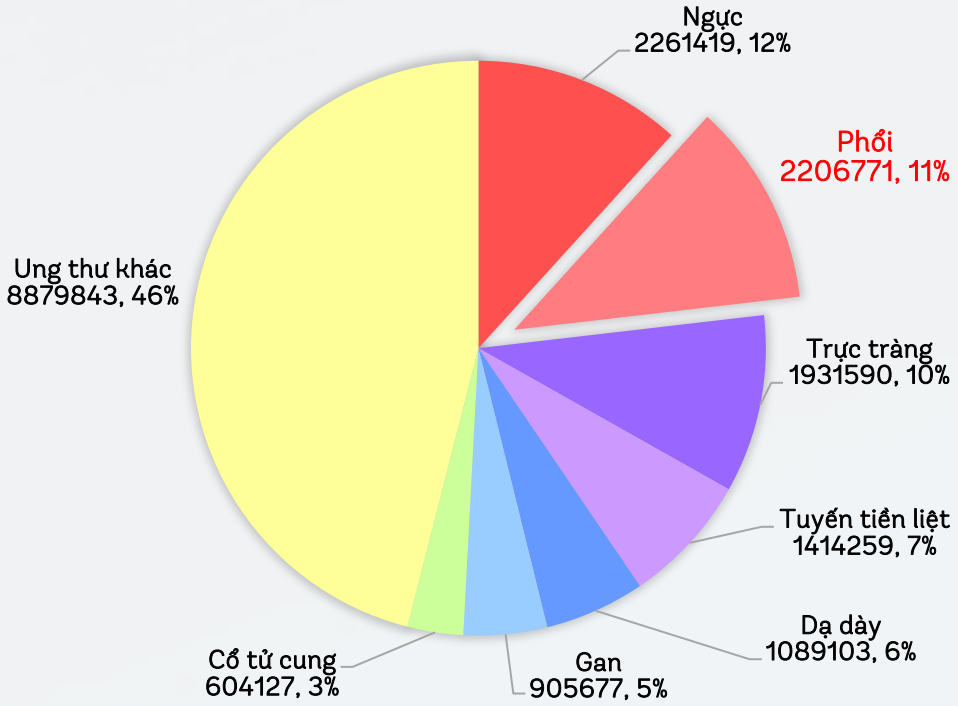


TÌNH HÌNH UNG THƯ PHỔI

1

SỐ LIỆU THỐNG KÊ

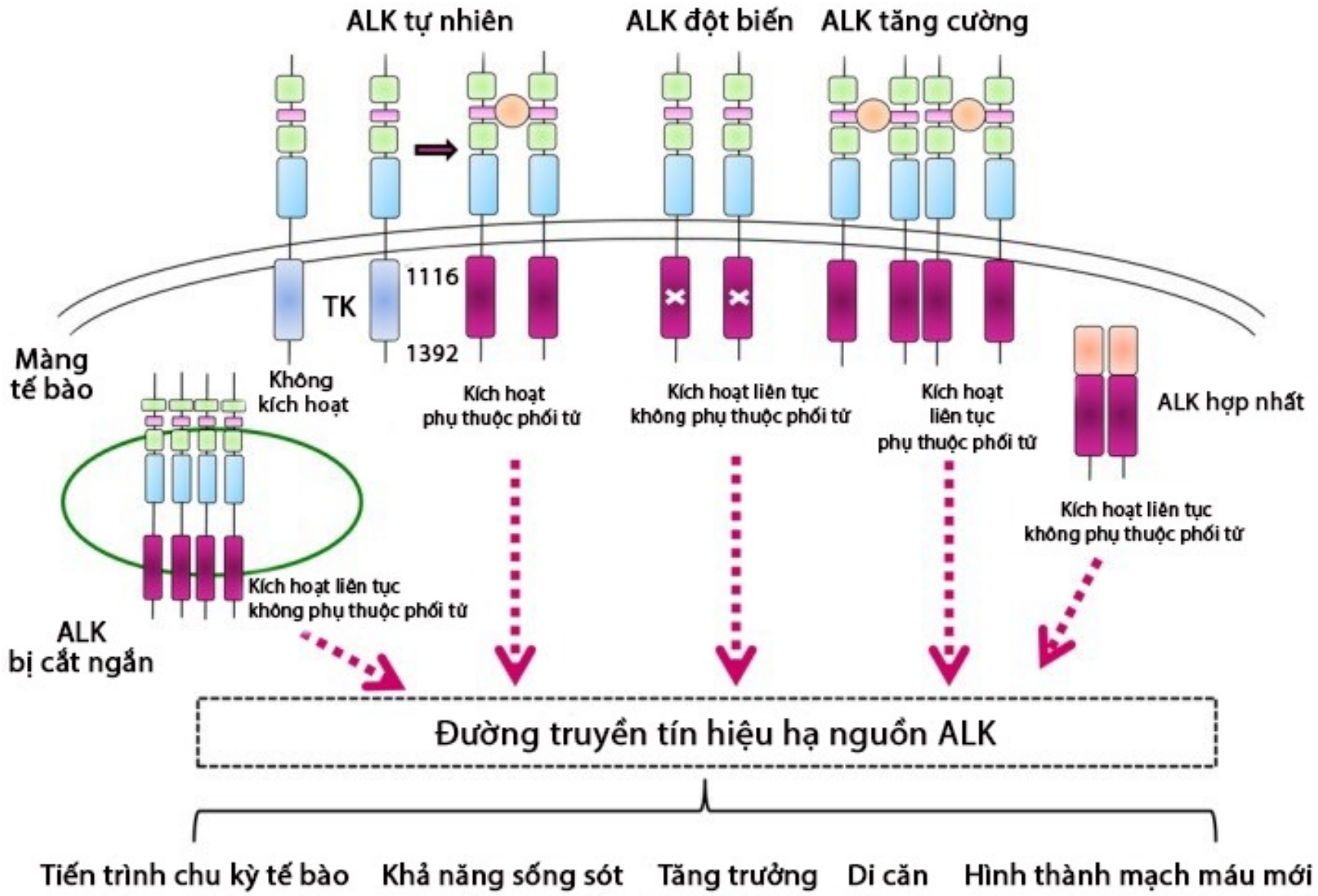
- Tỉ lệ tử vong chiếm 18%
- Tại Việt Nam, bệnh ung thư phổ biến thứ 2.
- NSCLC chiếm 85%.
- NSCLC dạng ALK dương tính chiếm 5%



Ước tính số ca ung thư mới mắc 2020

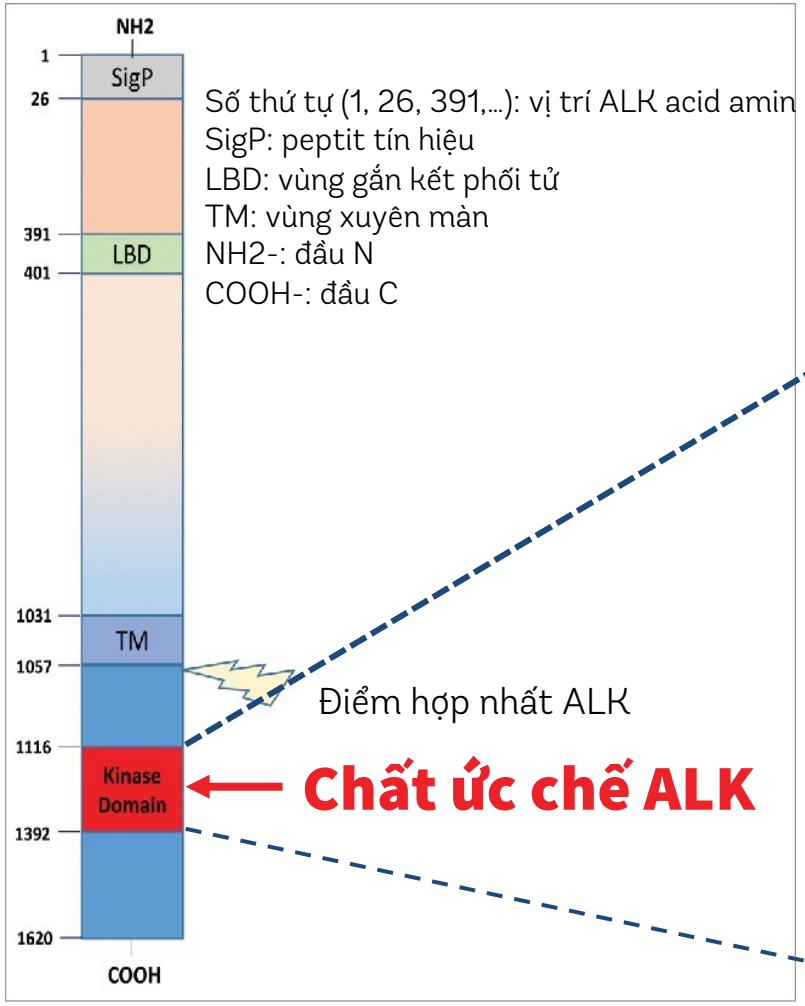
https://gco.iarc.fr/today/data/factsheets/populations/900-world-fact-sheets.pdf

CƠ CHẾ BỆNH SINH

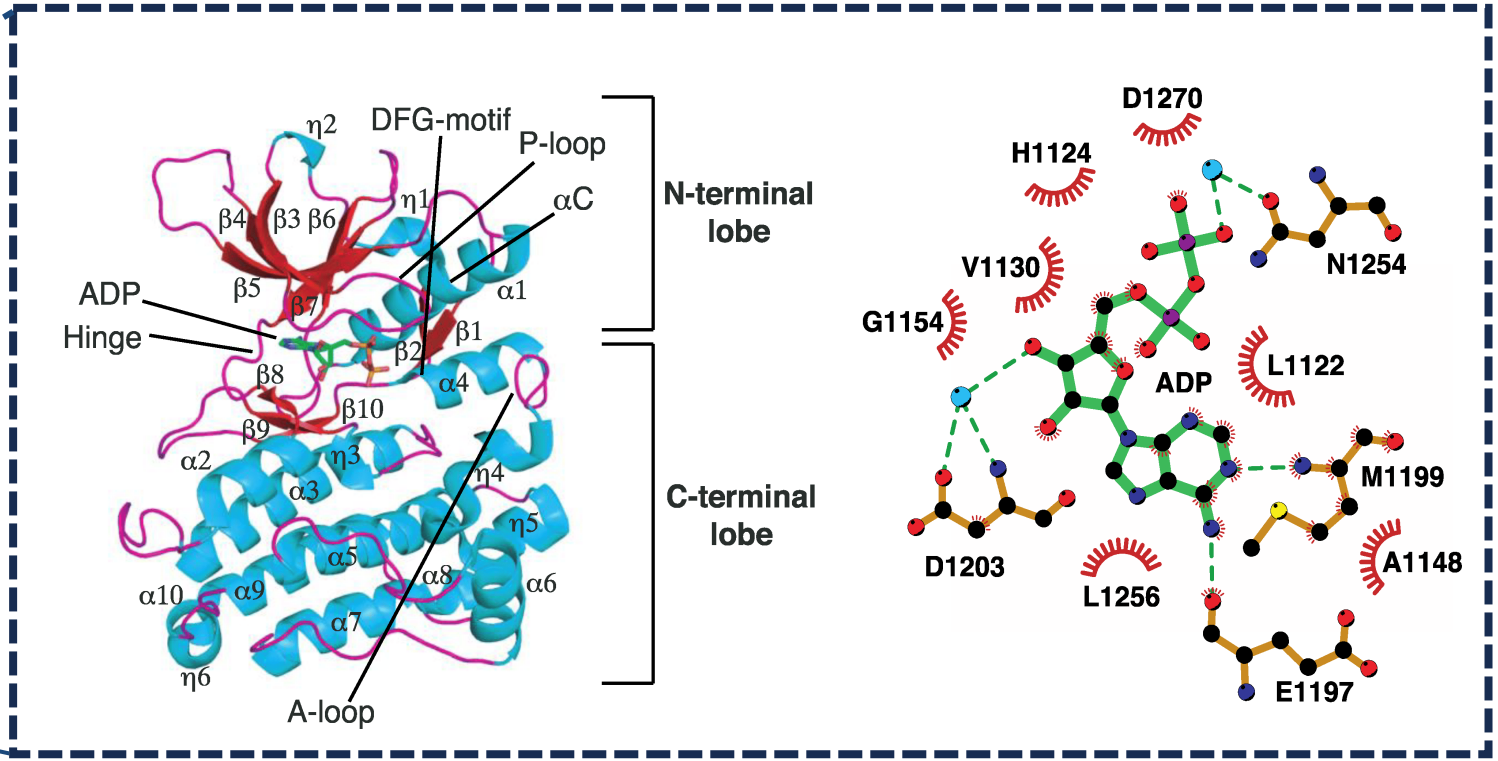


Takita J. The role of anaplastic lymphoma kinase in pediatric cancers. Cancer science. 2017

CẤU TRÚC THỤ THỂ ALK



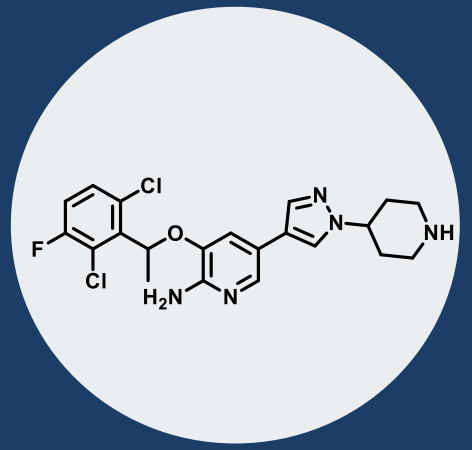
ATP gắn vào vùng kinase của thụ thể để kích hoạt quá trình phosphoryl hóa. **Vùng bản lề** là nơi ATP gắn vào. **Các chất ức chế ALK bắt chước cách gắn của ATP** ở vùng bản lề để ngăn chặn quá trình phosphoryl hóa và sự phát triển của ung thư.



Zhao Z. Anaplastic lymphoma kinase: role in cancer and therapy perspective. *Cancer Biology & Therapy*

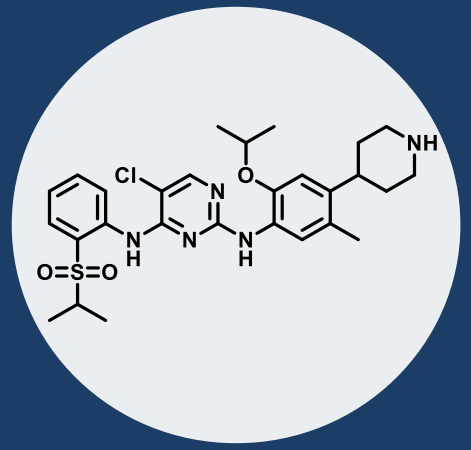
CÁC THUỐC ỨC CHẾ ALK

1



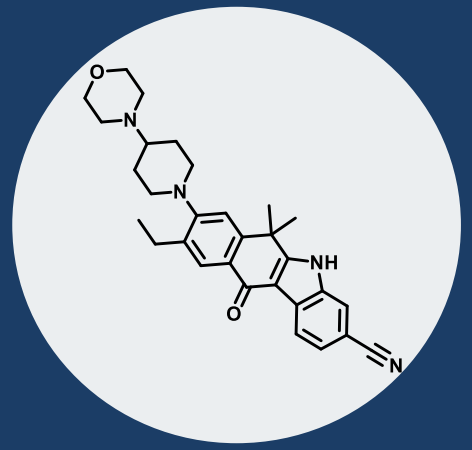
Crizotinib

Tên thương mại là Xalkori®, được FDA chấp thuận năm 2011



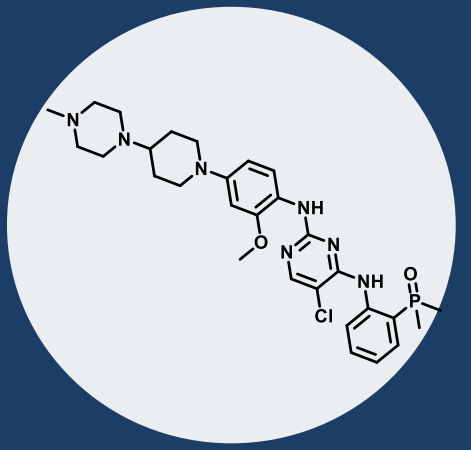
Ceritinib

Tên thương mại là Zykadia®, được FDA chấp thuận năm 2014



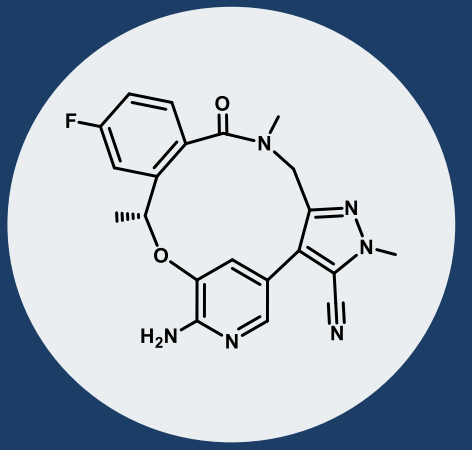
Alectinib

Tên thương mại là Alecensa®, được FDA chấp thuận năm 2015



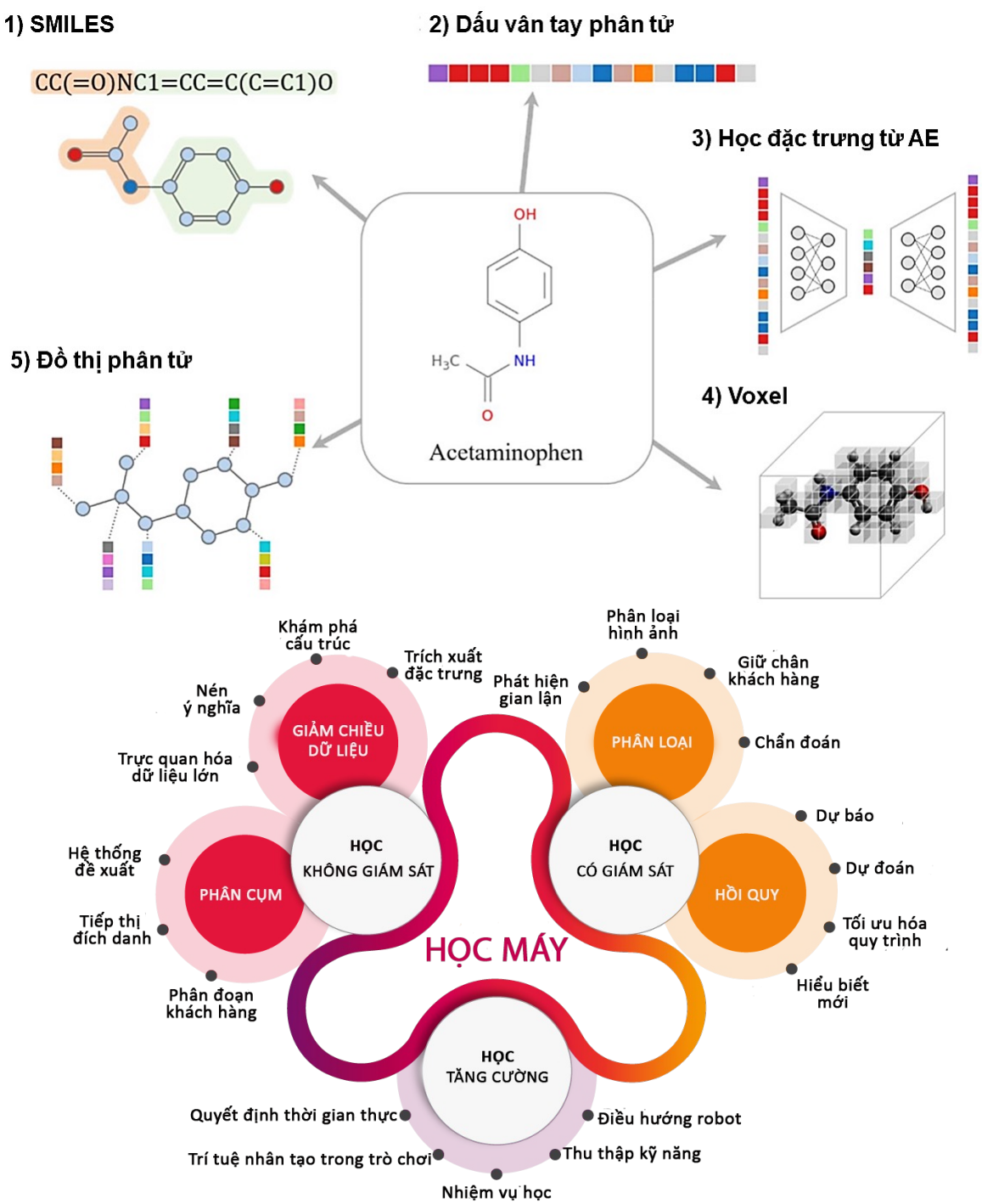
Brigatinib

Tên thương mại là Alunbrig®, được FDA chấp thuận năm 2016



Lorlatinib

Tên thương mại là Lorbrena®, được FDA chấp thuận năm 2015



VAI TRÒ CỦA AI TRONG THIẾT KẾ THUỐC

1

DỰ ĐOÁN CHẤT TIỀM NĂNG

Halicin: kháng sinh do AI tìm ra

DỰ ĐOÁN ÁI LỰC GẮN KẾT

GNINA: docking phân tử dùng AI

DỰ ĐOÁN CẤU TRÚC PROTEIN

Alphafold: AI dự đoán cấu trúc 3D protein

BIỂU DIỄN CẤU TRÚC PHÂN TỬ

SMILES, dấu vân tay ,đồ thị phân tử

CÁC THUẬT TOÁN HỌC MÁY PHỔ BIẾN

1

MẠNG THẦN KINH NHÂN TẠO (ANN)

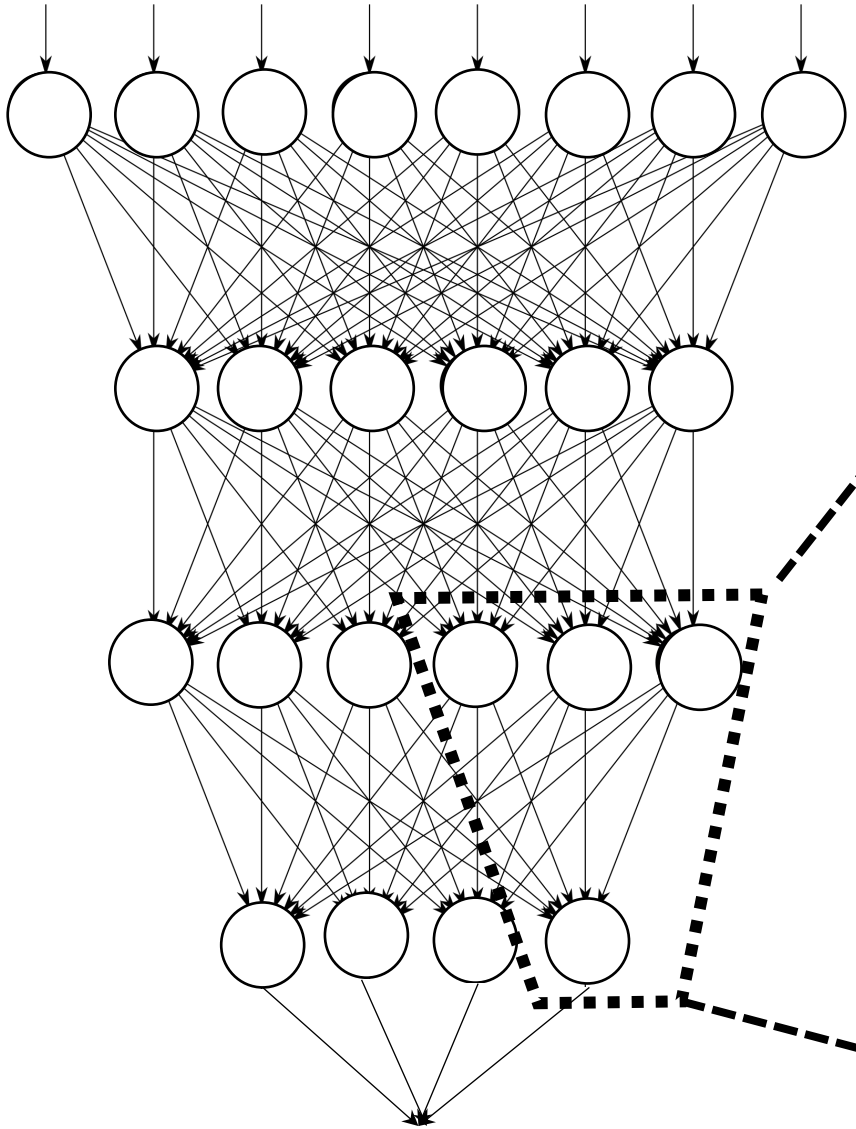
Lớp đầu vào

Lớp ẩn 1

Lớp ẩn 2

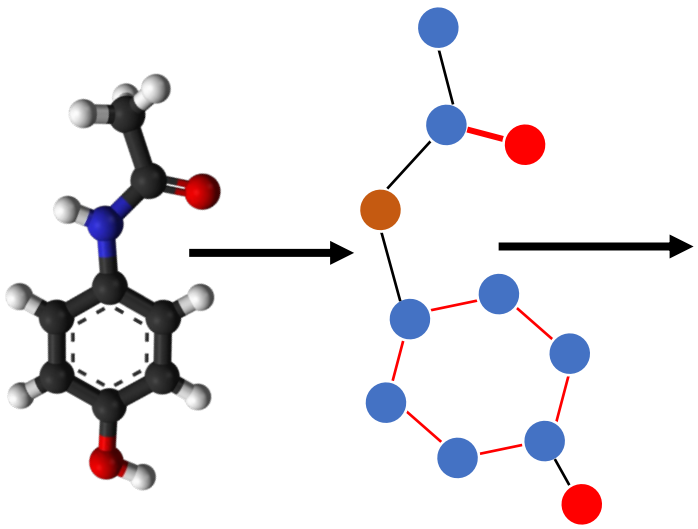
Lớp ẩn 3

Đầu ra

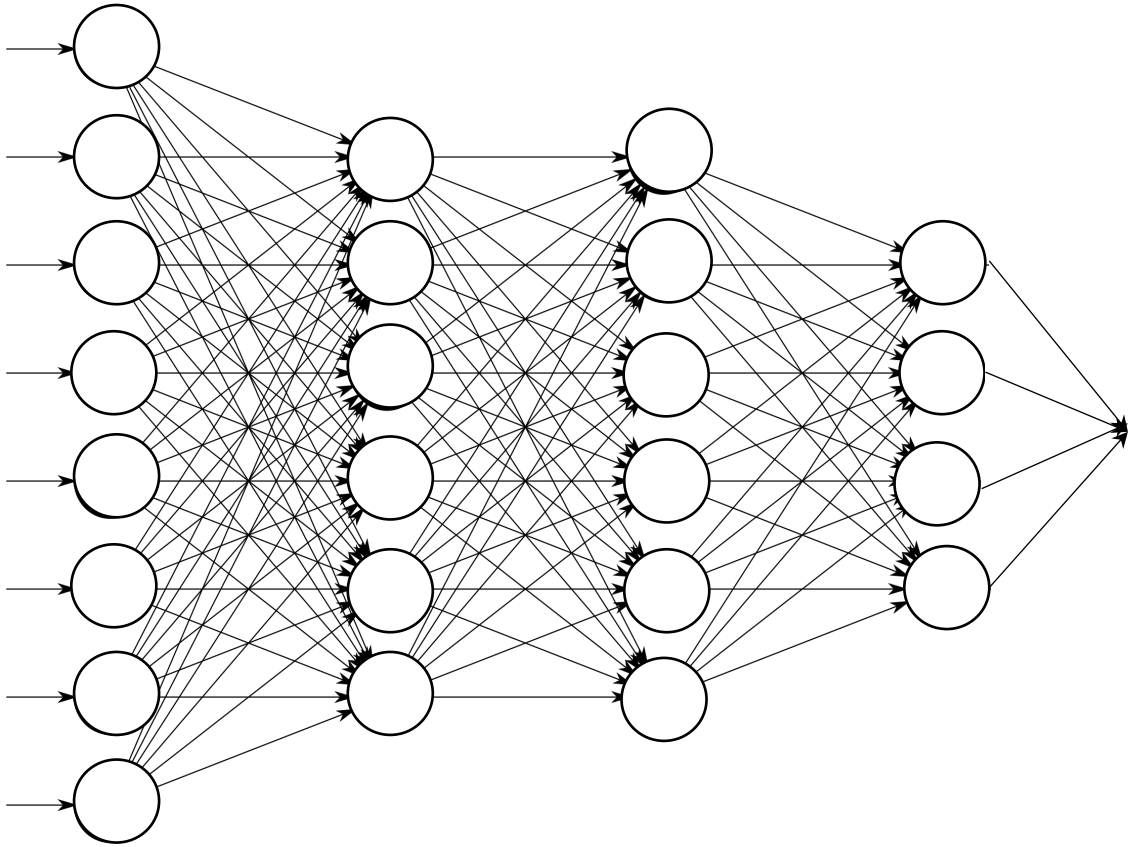


ACTIVE
INACTIVE

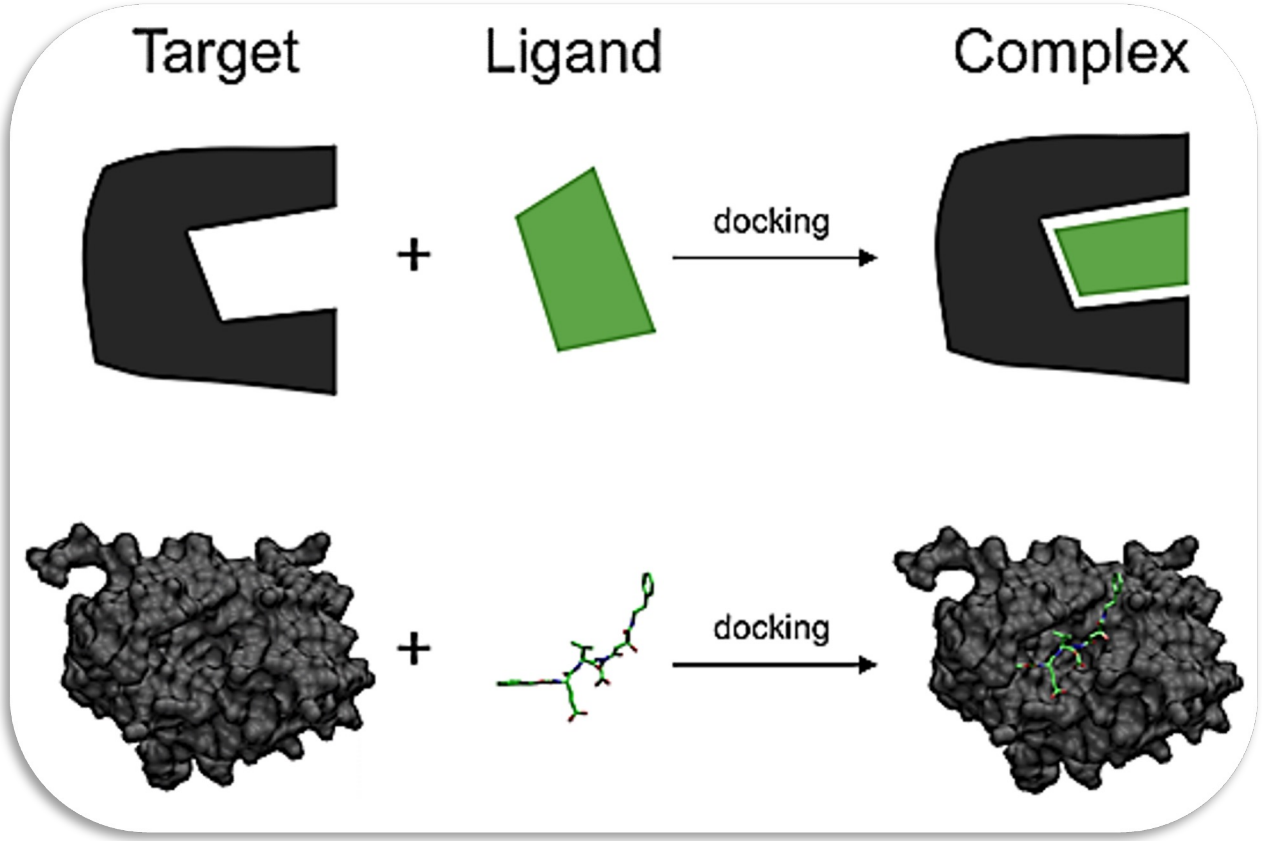
MẠNG THẦN KINH ĐỒ THỊ (GNN)



LỚP XỬ LÝ ĐỒ THỊ



MÔ HÌNH DOCKING PHÂN TỬ



PROTEIN

PHỐI TỬ

CHUẨN BỊ

DOCKING

TÍNH ĐIỂM

PHÂN TÍCH

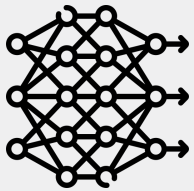
2. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP

Xây dựng dữ liệu huấn luyện, dữ liệu sàng lọc và protein đích.

Xây dựng các mô hình học máy, ANN, GNN, Docking. Phương pháp đánh giá và cải thiện mô hình.

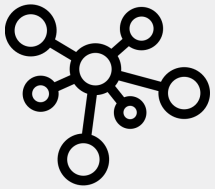
Mô hình học máy

Quy trình xây dựng và đánh giá mô hình học máy.



ANN

Thiết kế cấu trúc mạng ANN đánh giá mô hình ANN.



GNN

Thiết kế cấu trúc mạng GNN đánh giá mô hình GNN.

Docking

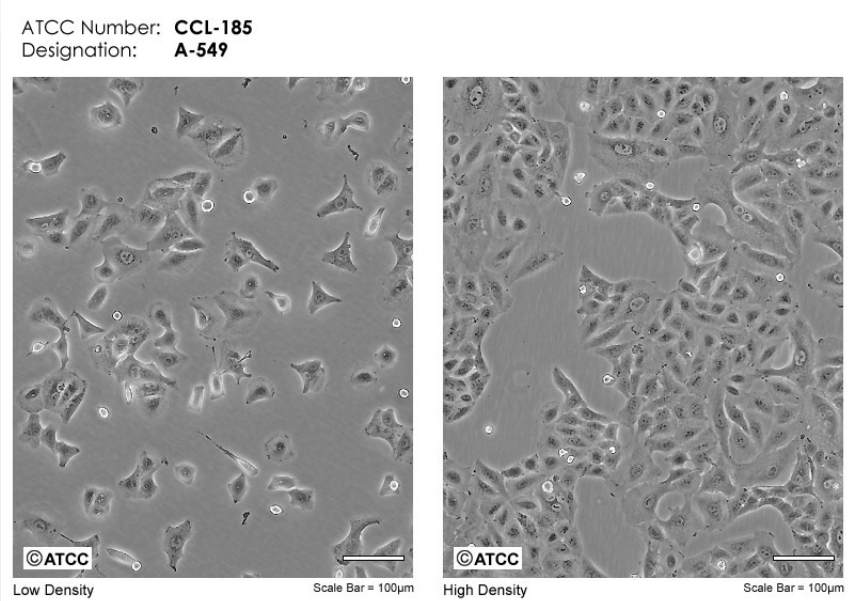
Docking phân tử để đánh giá hiệu năng mô hình AI .

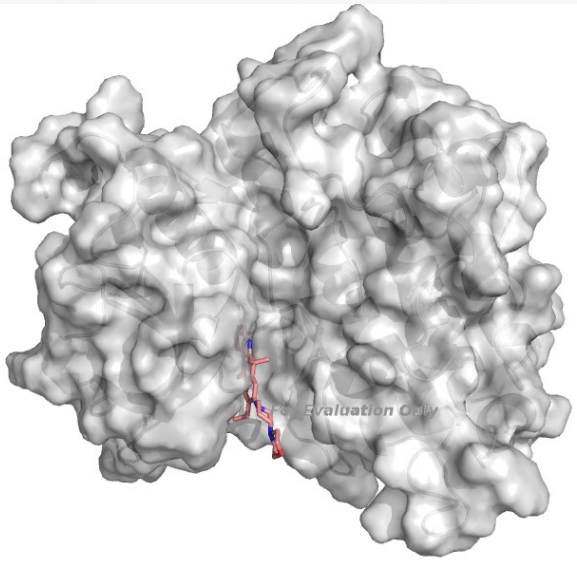
CƠ SỞ DỮ LIỆU

2

Reaxys

ELSEVIER





DỮ LIỆU XÂY DỰNG

26.168 cấu trúc được nghiên cứu trên ALK từ thư viện Reaxys.

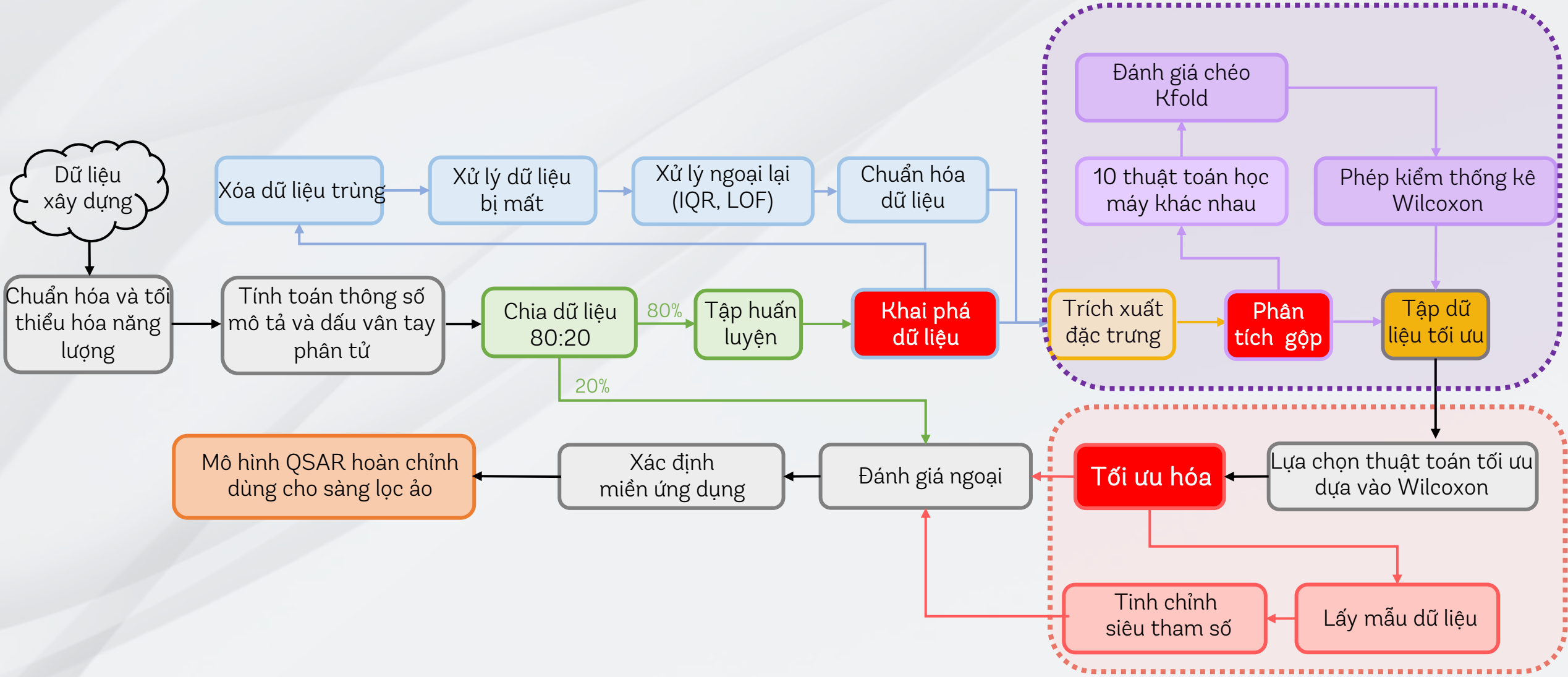
DỮ LIỆU SÀNG LỌC

Cấu trúc được nghiên cứu độc tính trên dòng tế bào A549 của ChEMBL mã định danh 392.

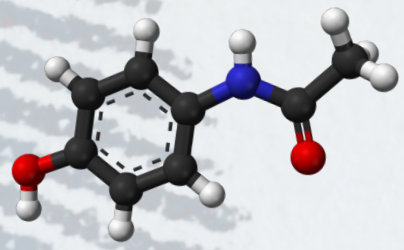
PROTEIN

Cấu trúc ALK liên kết với phối tử alectinib với mã protein là 3AOX.

MÔ HÌNH HỌC MÁY



Quy trình Phân tích gộp



FP1 FP2 FP3 FP4 FP4 FP5 FP6 FP7 FP8 FP8

FP9 FP10 FP11 FP12 FP13 FP14 FP15 FP16 FP17 FP18 FP19

KHAI PHÁ DỮ LIỆU

Model 10 Model 9 Model 8 Model 7 Model 6 Model 5 Model 4 Model 3 Model 2 Model 1

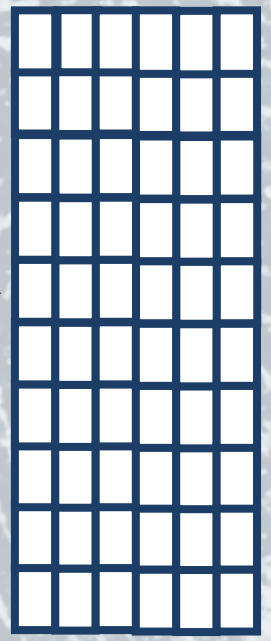
Meta-model

Quy trình Phân tích gộp

Meta-model

- Model 1
- Model 2
- Model 3
- Model 4
- Model 5
- Model 6
- Model 7
- Model 8
- Model 9
- Model 10

Ma trận kết quả

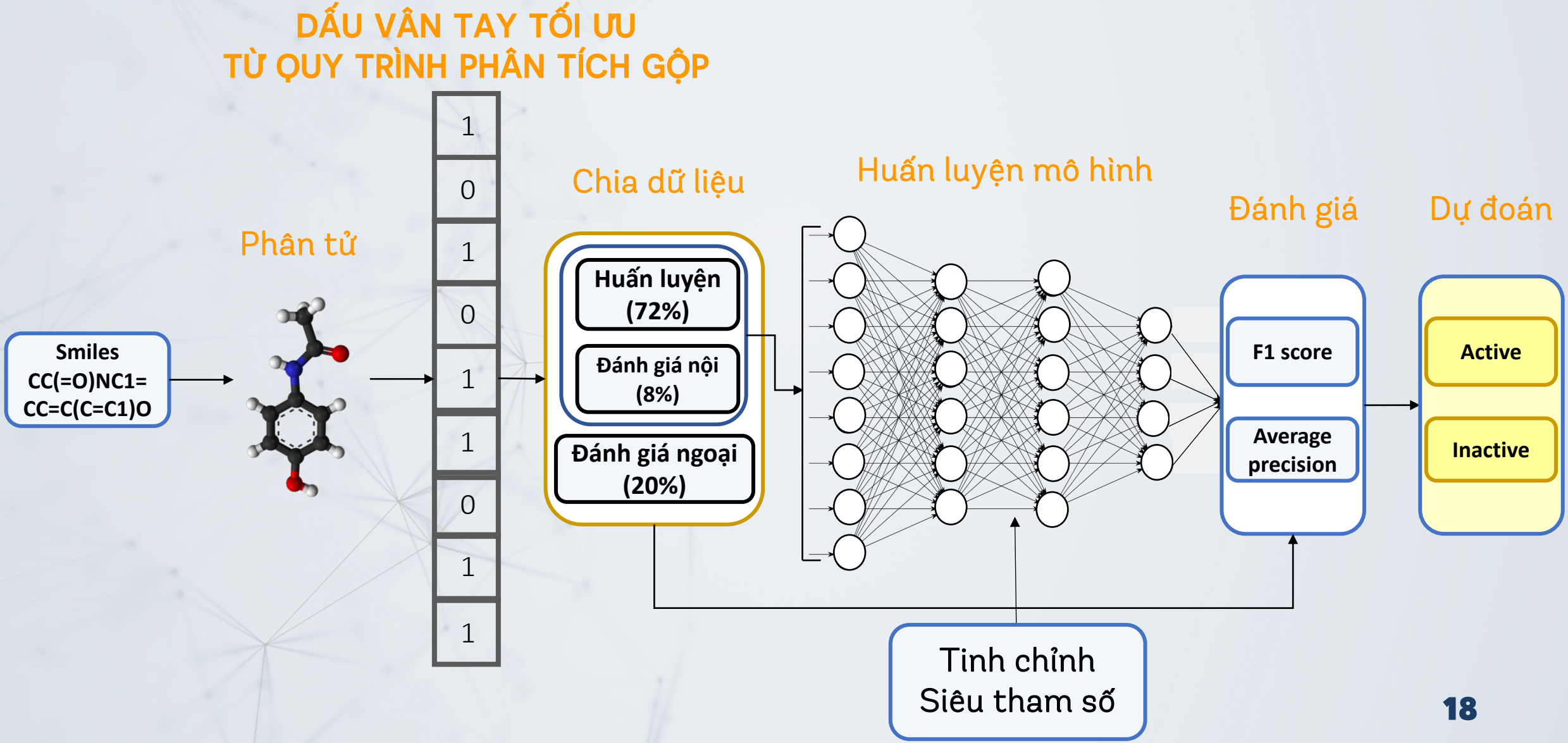


x19

Phép kiểm thống kê **Wilcoxon**

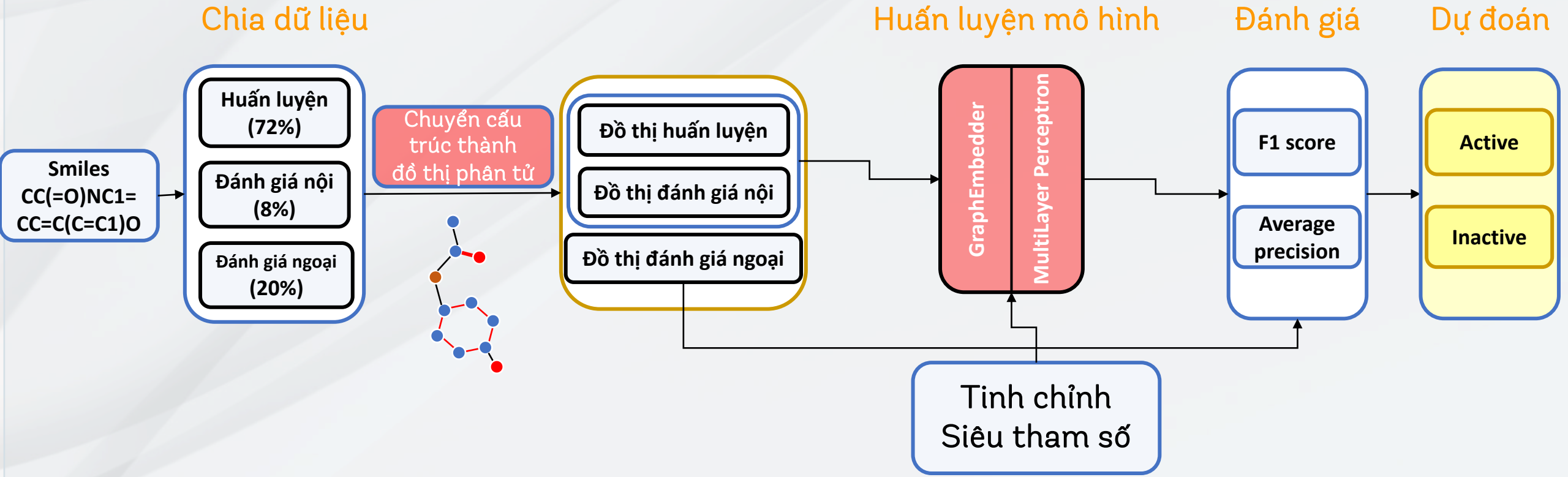
Fingerprint tốt nhất

MẠNG THẦN KINH NHÂN TẠO (ANN)

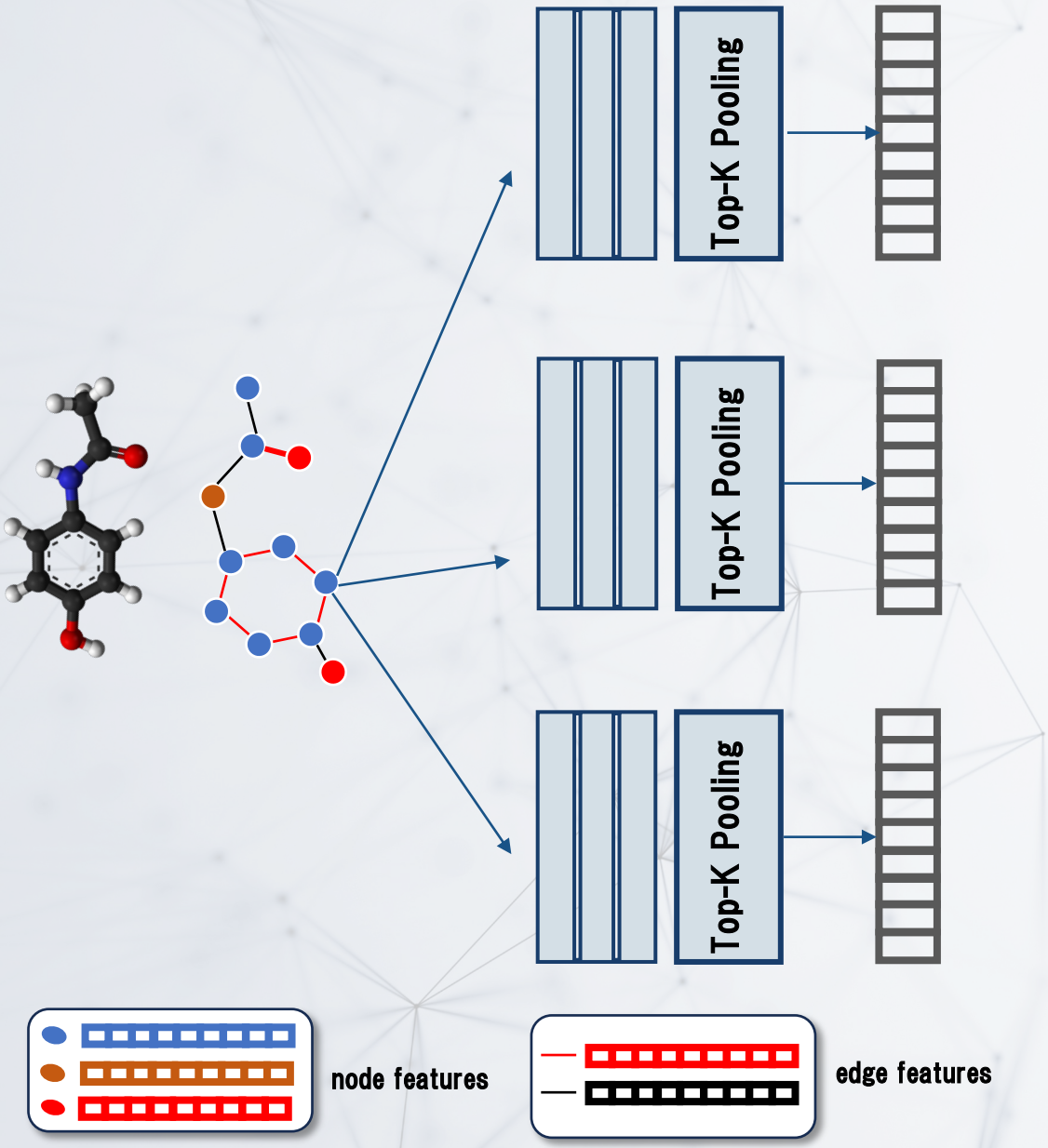


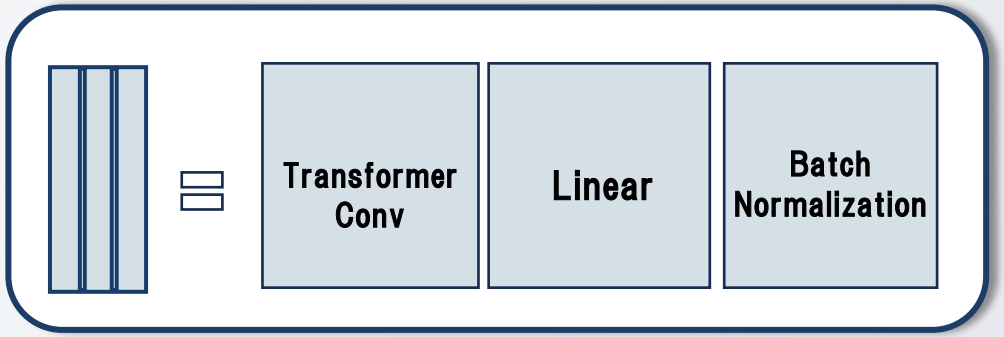
MẠNG THẦN KINH ĐỒ THỊ (GNN)

2

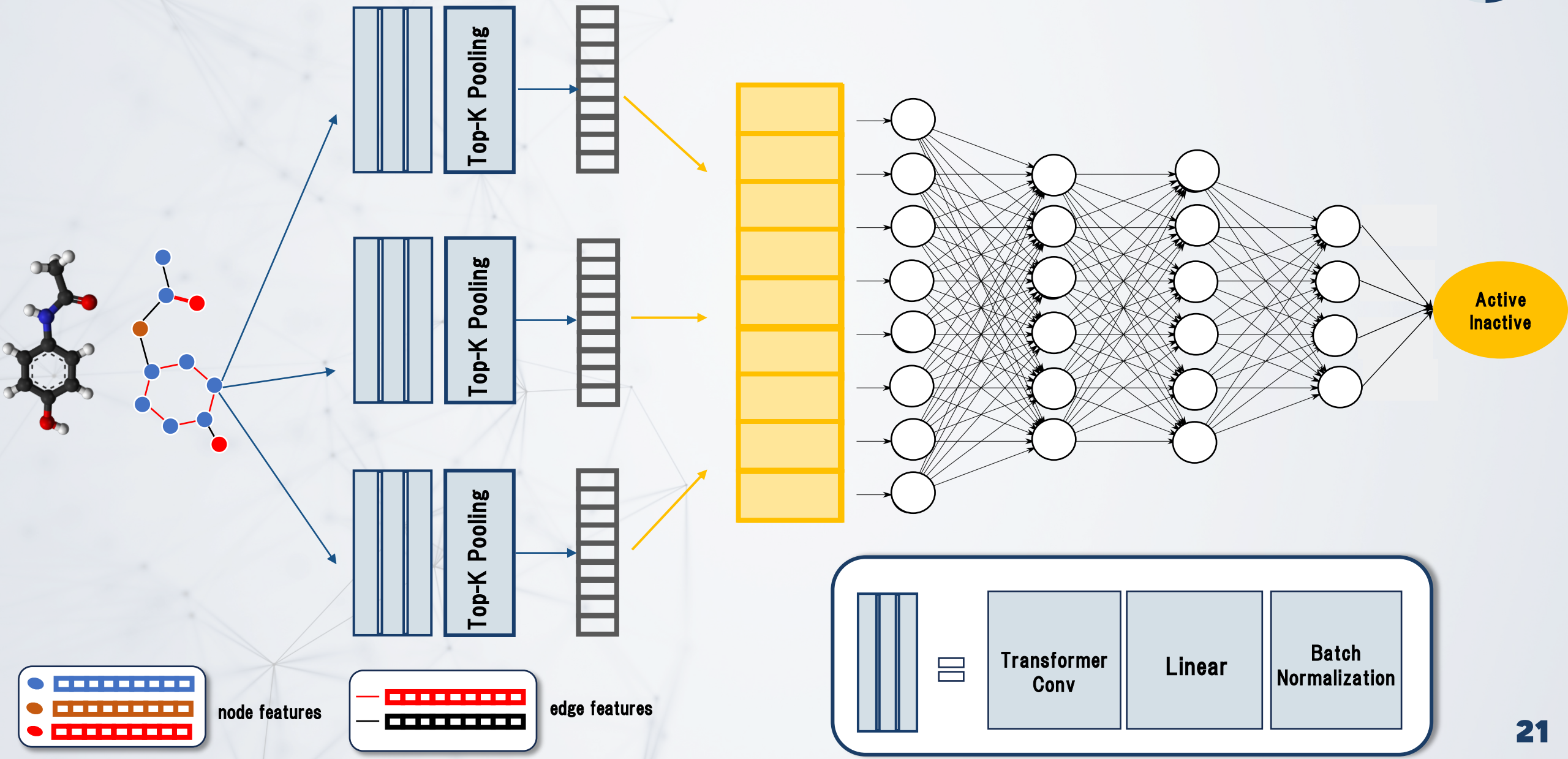


MẠNG THẦN KINH ĐỒ THỊ (GNN)





MẠNG THẦN KINH ĐỒ THỊ (GNN)



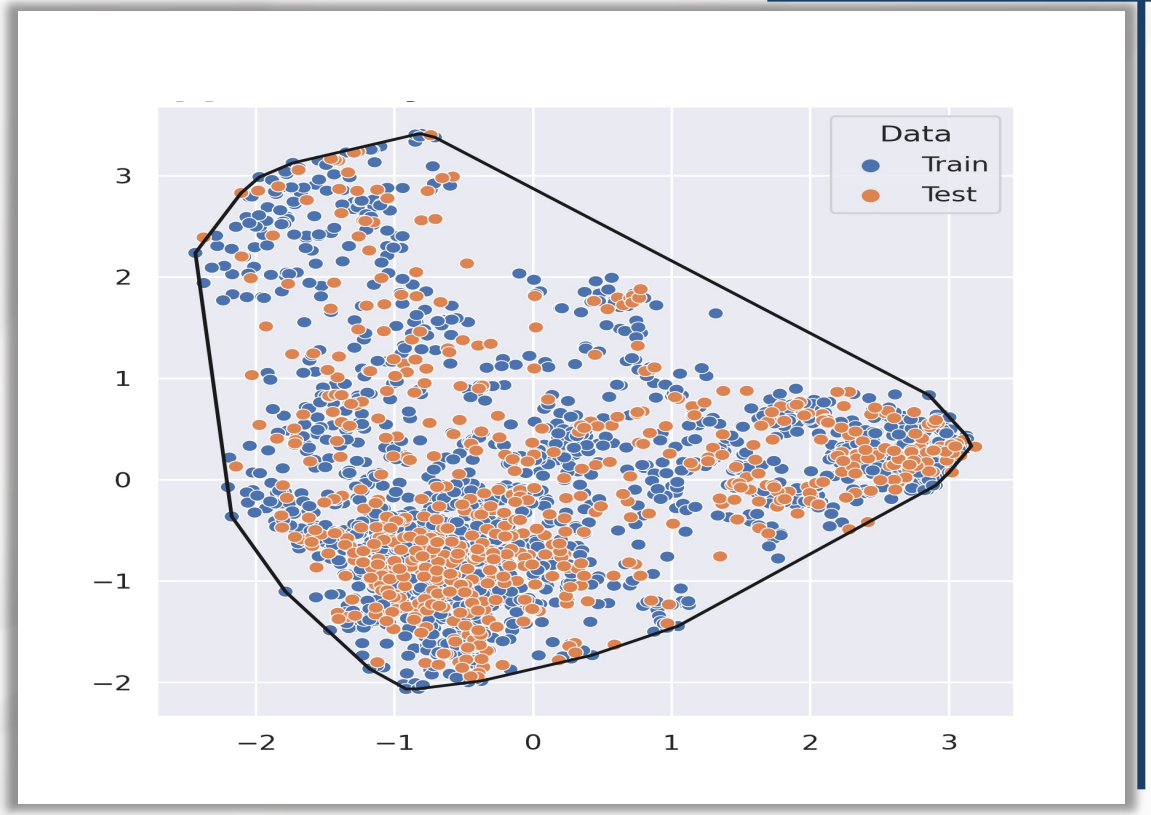
MIỀN ỨNG DỤNG (APPLICABILITY DOMAIN)

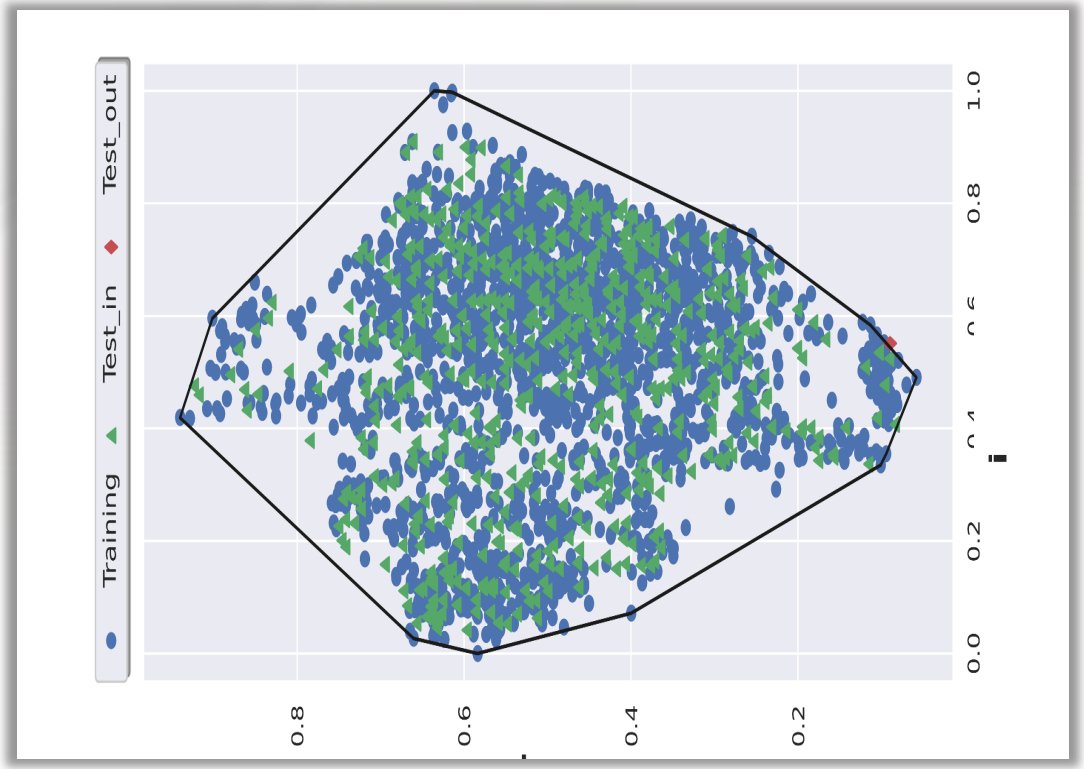
2

DẤU VÂN TAY + PCA + HÀM BAO LỒI

Ưu điểm: đơn giản, tính toán nhanh.

Nhược điểm: Có nhiều khoảng trống trong AD





MA TRẬN TƯƠNG ĐỒNG TANIMOTO + MDS + HÀM BAO LỒI

Ưu điểm: thể hiện độ tương đồng của các chất với nhau, khắc phục nhược điểm vùng trống.

Nhược điểm: tính toán tốn nhiều thời gian.

TỐI ƯU HÓA MÔ HÌNH

2

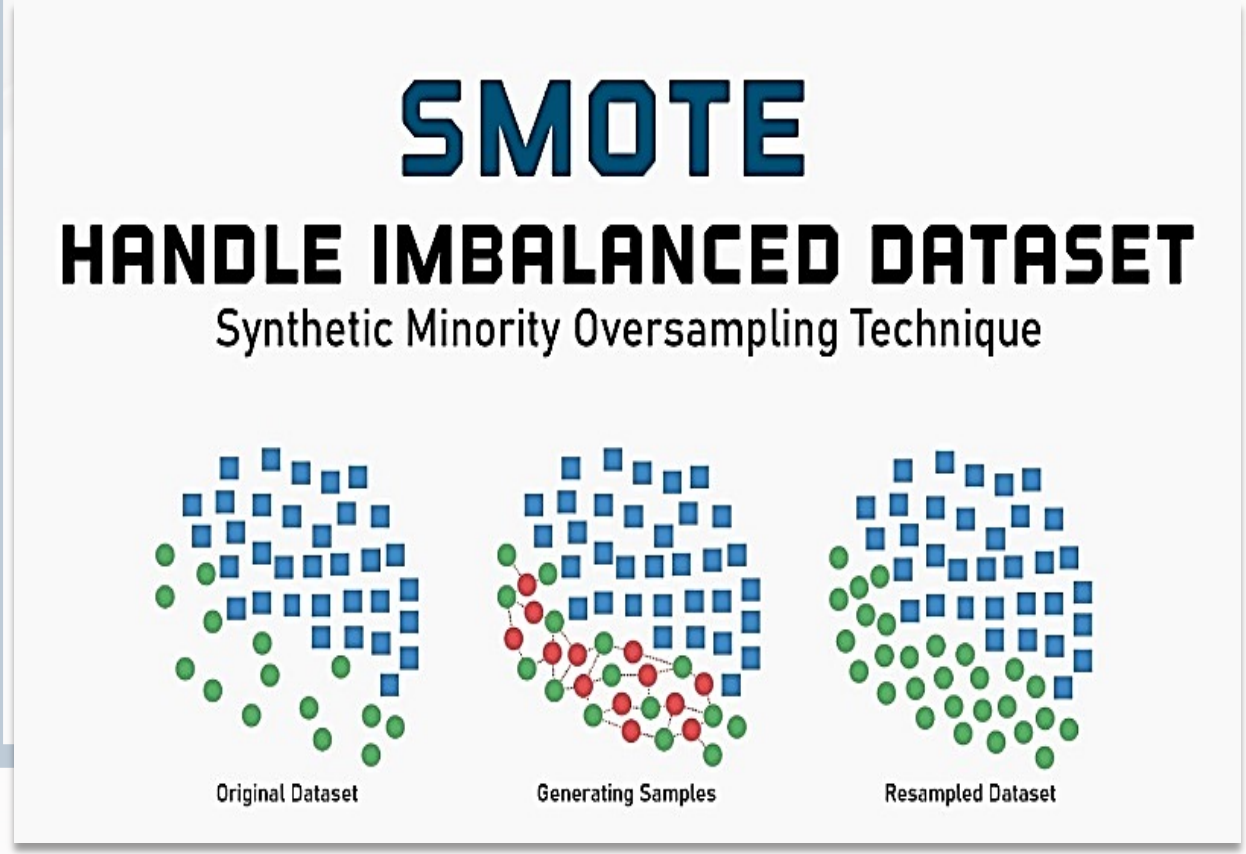
LẤY MẪU DỮ LIỆU

Phù hợp tập dữ liệu bị mất cân bằng.

SMOTE (Synthetic Minority Over-sampling Technique) là mô-đun của thư viện imbalanced-learn phiên bản 0.11.0.

Hoạt động bằng cách **tạo ra các mẫu giả trong lớp thiểu số** giúp cân bằng tỷ lệ giữa các lớp.

Đề tài áp dụng phương pháp này cho **mô hình học máy**.



TỐI ƯU HÓA MÔ HÌNH

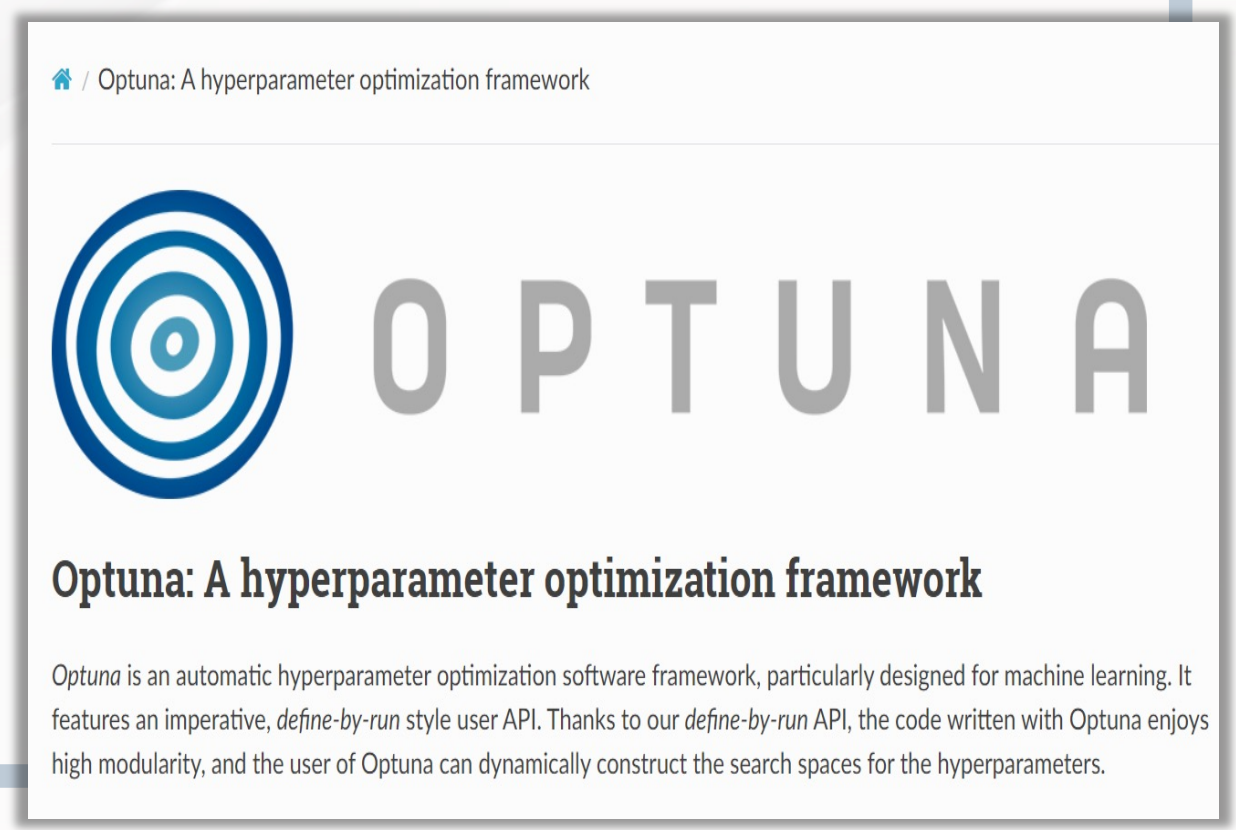
TINH CHỈNH SIÊU THAM SỐ

Siêu tham số là các thông số cấu hình được đặt sẵn trước khi huấn luyện mô hình.

Optuna là một thư viện Python được thiết kế cho việc tối ưu hóa siêu tham số.

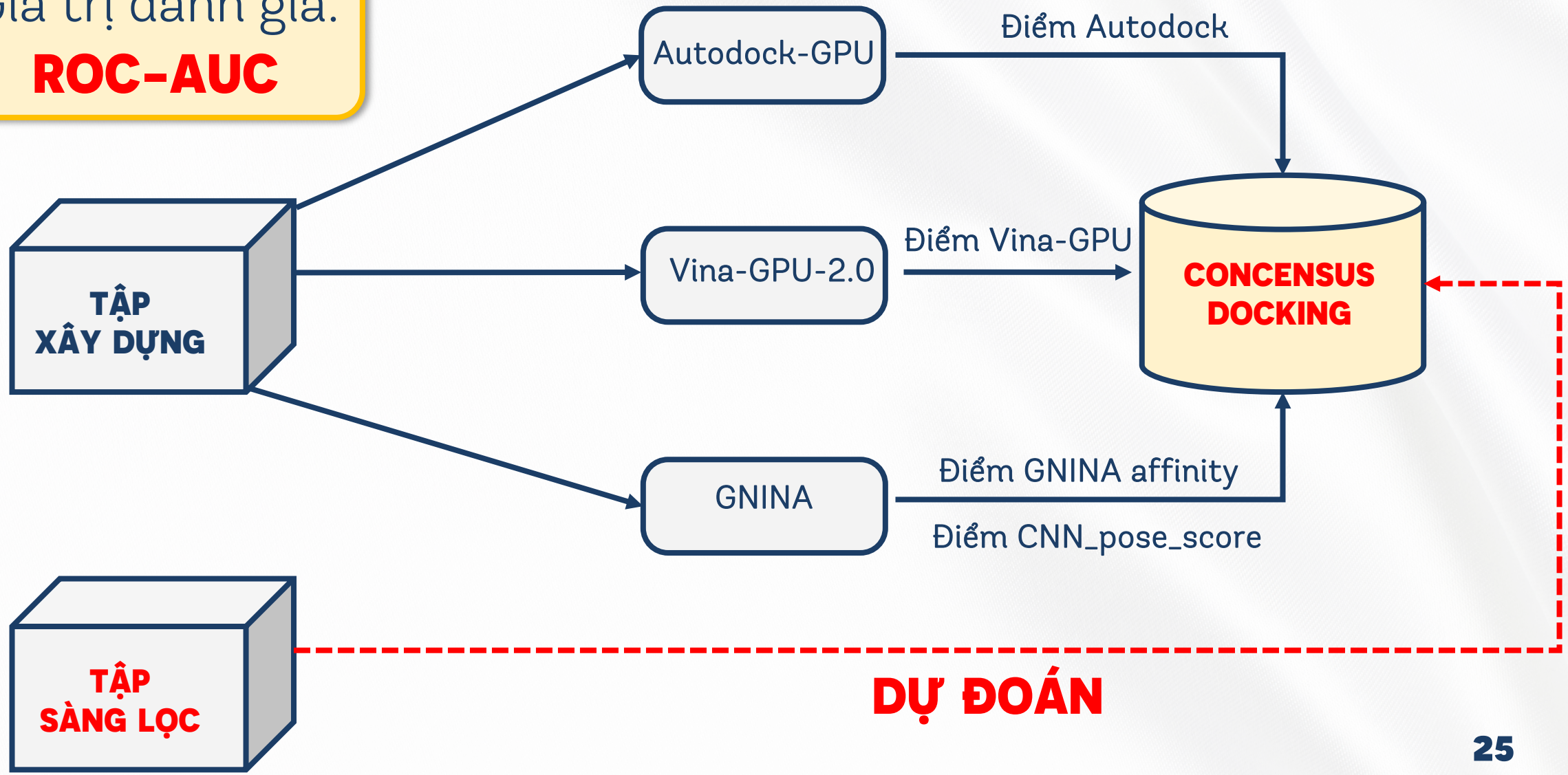
Optuna sử dụng thuật toán Tree-structured Parzen Estimator (TPE)để thực hiện tối ưu hóa Bayesian.

Đề tài lần lượt thử nghiệm 300 lần (trials) cho mô hình học máy, 100 lần cho mô hình ANN và 50 lần cho mô hình GNN.

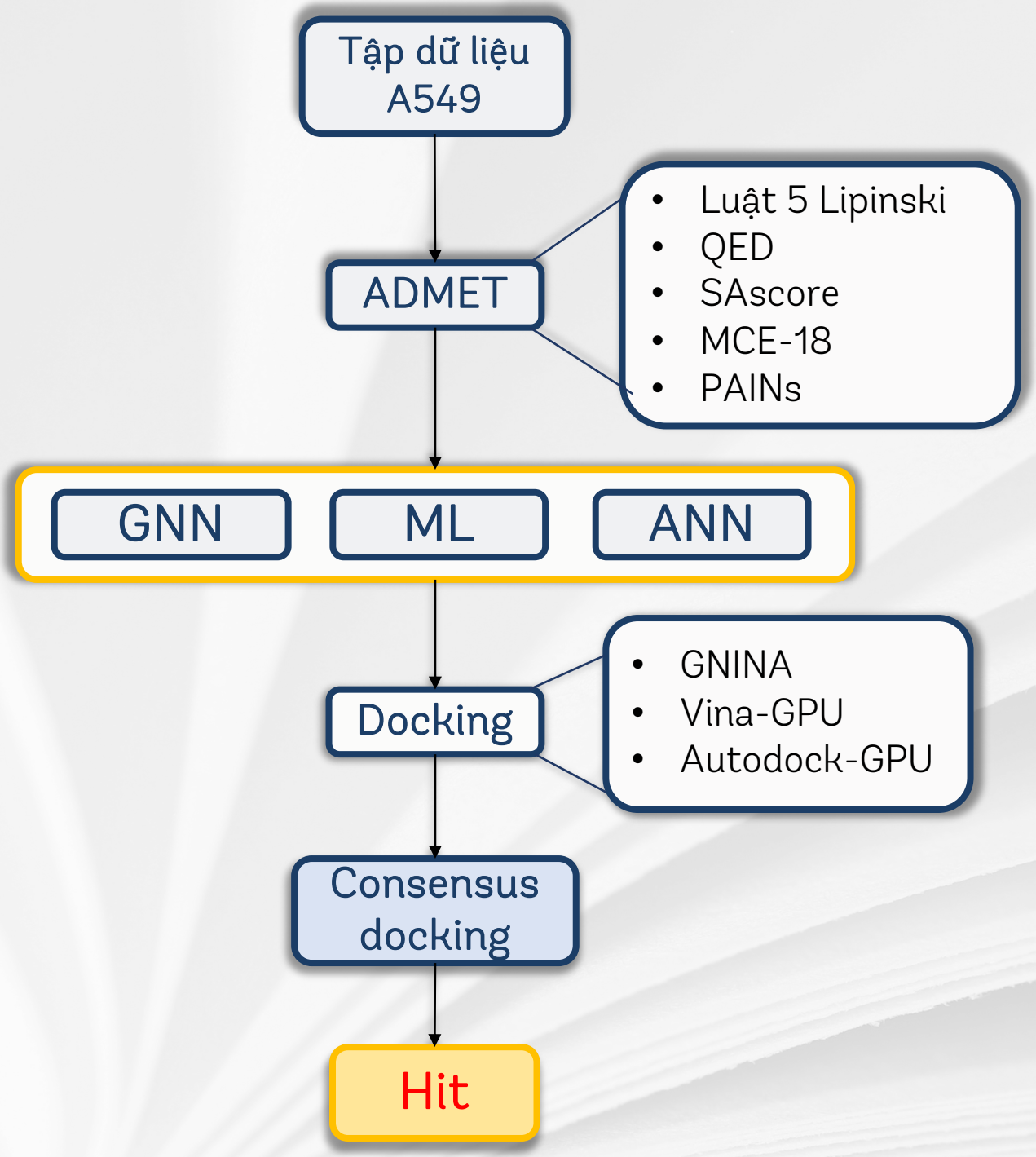


DOCKING PHÂN TỬ

Giá trị đánh giá:
ROC-AUC



SÀNG LỌC ẢO


Tập dữ liệu A549
ADMET
Luật 5 Lipinski
QED
SAscore
MCE-18
PAINs
GNN
ML
ANN
Docking
GNINA
Vina-GPU
Autodock-GPU
Consensus docking
Hit

3. KẾT QUẢ

Kết quả của quy trình sàng lọc dữ liệu xây dựng và dữ liệu sàng lọc

Kết quả đánh giá mô hình học máy, mạng thần kinh nhân tạo và mạng thần kinh đồ thị

Triển khai các mô hình cho việc sàng lọc ảo để tìm ra chất ức chế ALK tiềm năng.

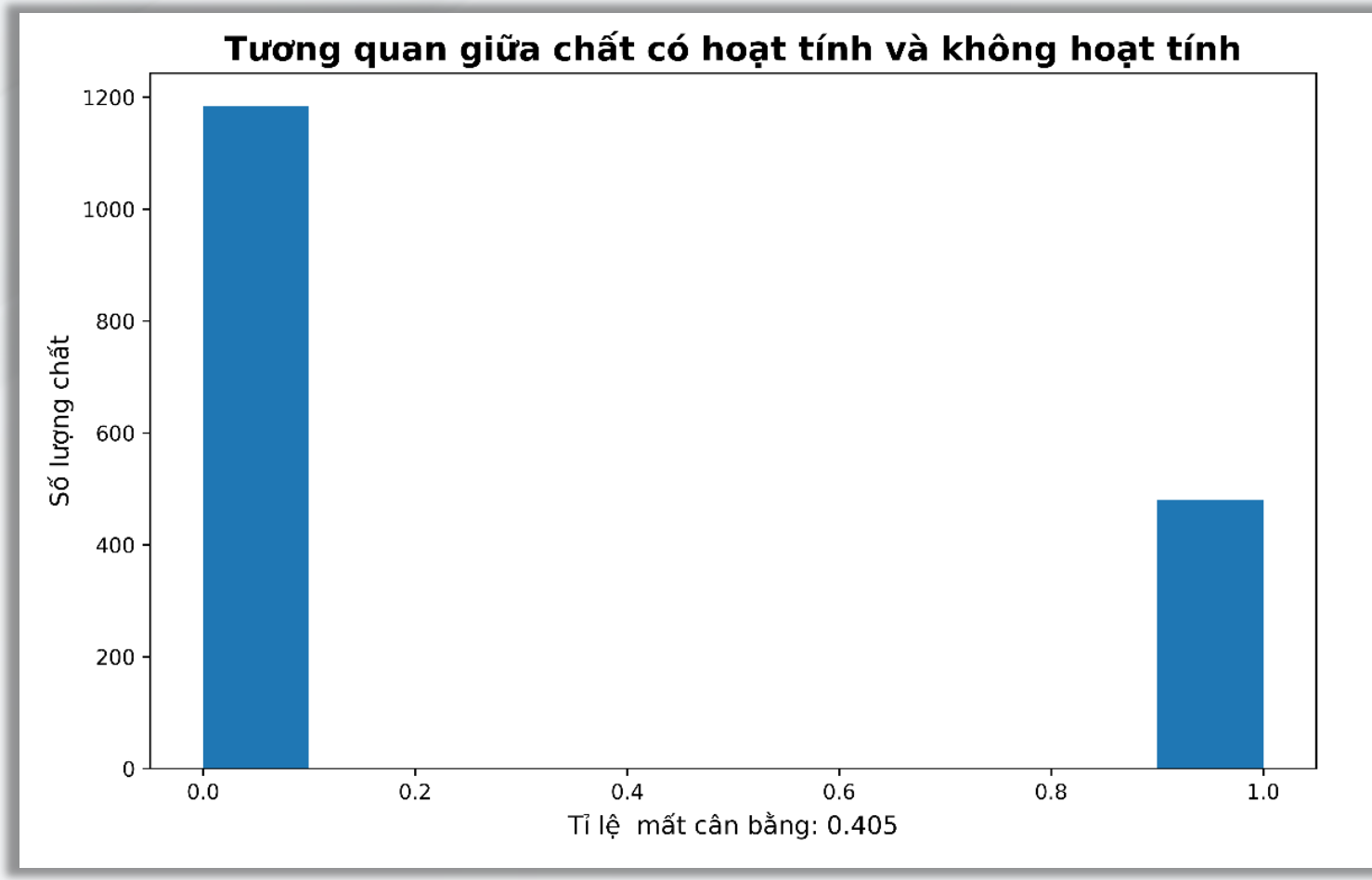
QUY TRÌNH THU THẬP VÀ XỬ LÝ DỮ LIỆU

3

Cơ sở dữ liệu Reaxys: 26.168 chất.

Số chất giữ lại sau quy trình lọc: 1.664

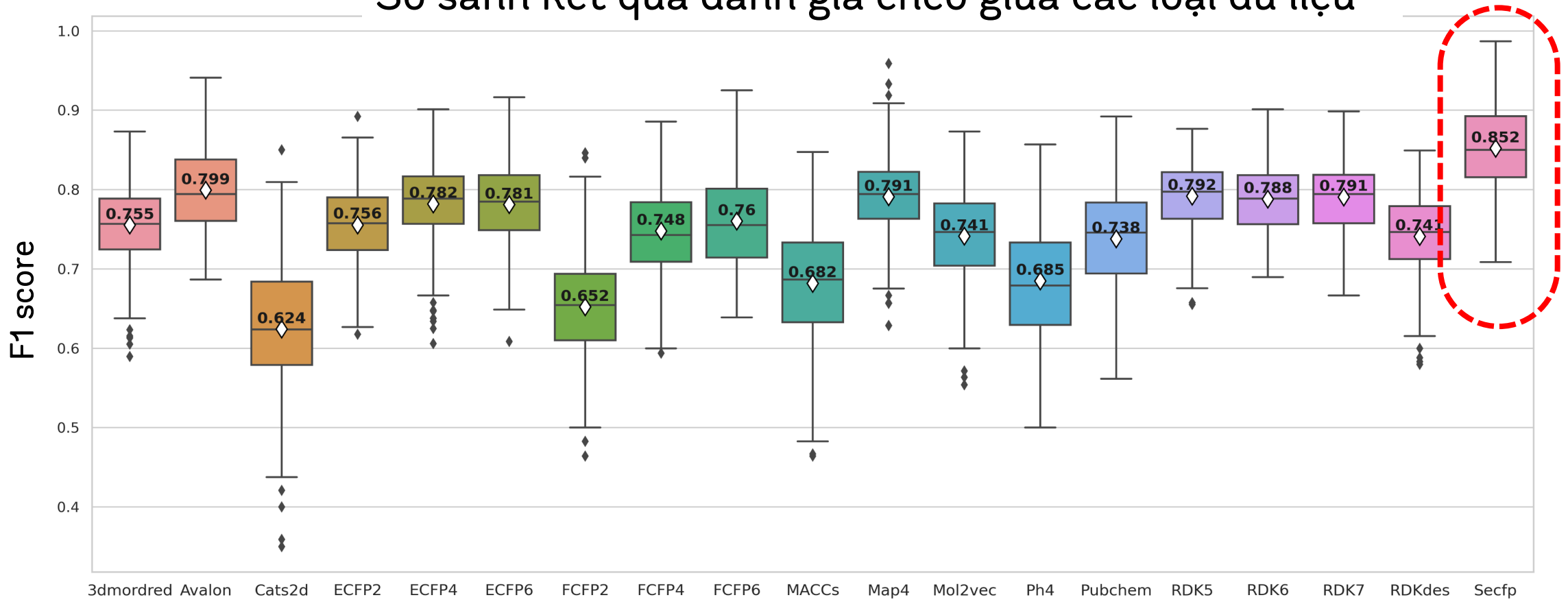
Tỉ lệ mất cân bằng: 40,5%.



MÔ HÌNH HỌC MÁY

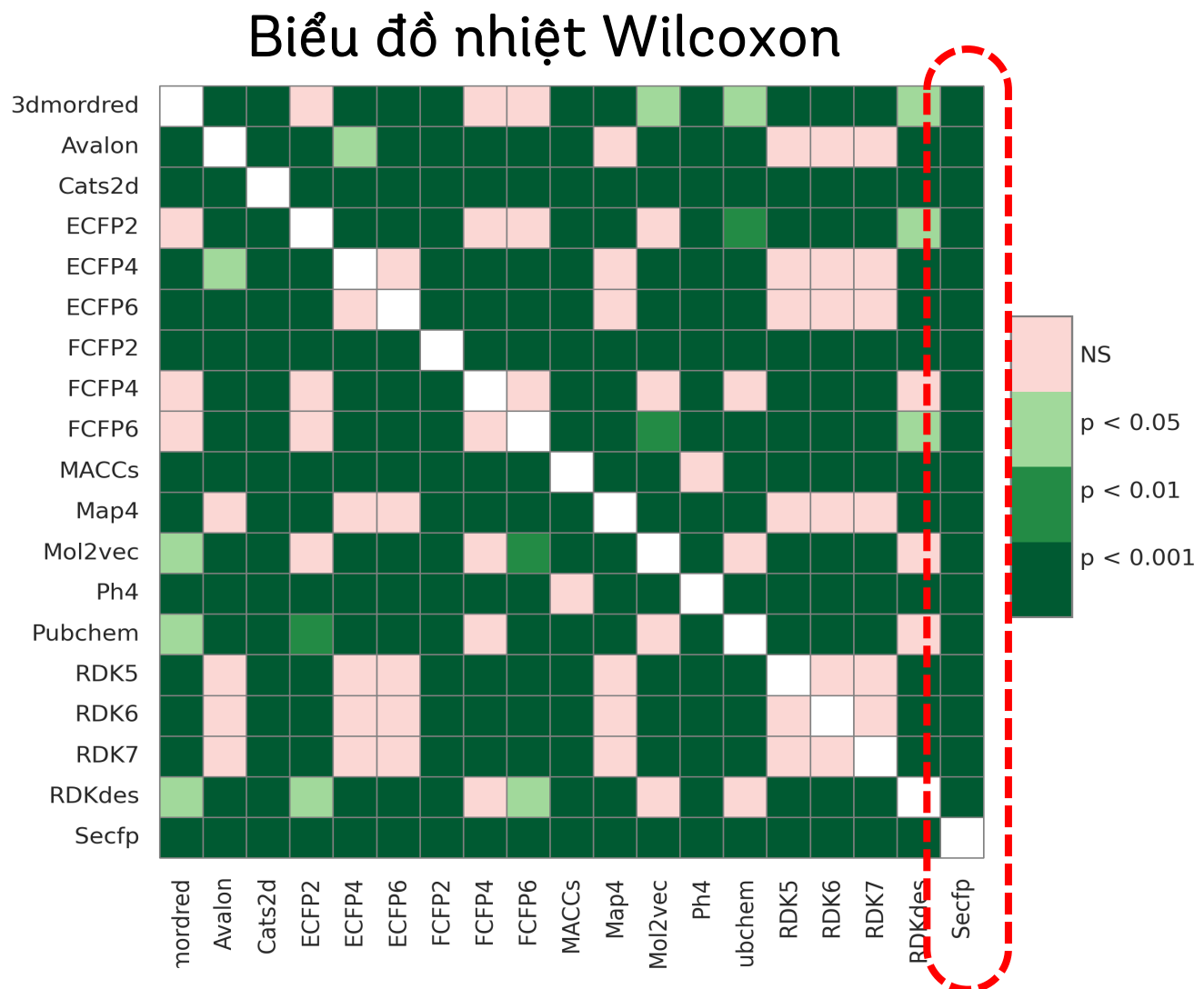
Phân tích gộp

So sánh kết quả đánh giá chéo giữa các loại dữ liệu



MÔ HÌNH HỌC MÁY

Phân tích gộp



SECFP là bộ dấu vân tay tối ưu

MÔ HÌNH HỌC MÁY

3

Kết quả quy trình khai thác dữ liệu áp dụng trên Secfp.

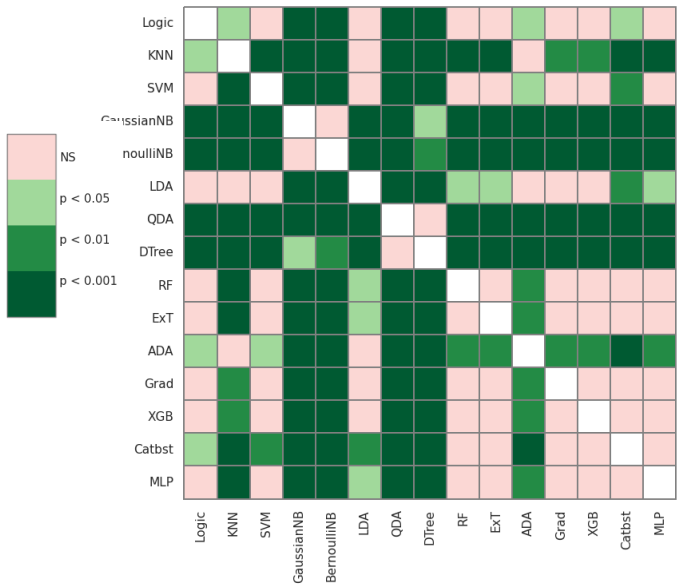
BƯỚC THỰC HIỆN	KẾT QUẢ
Phân chia tập dữ liệu	Tập huấn luyện có 1.331 chất, tỉ lệ mất cân bằng 40,5%. Tập đánh giá ngoại có 333 chất, tỉ lệ mất cân bằng 40,5%.
Làm sạch dữ liệu	Tập huấn luyện còn 1.330 chất và 320 đặc trưng. Tập đánh giá ngoại vẫn giữ 333 chất và 320 đặc trưng.
Xử lý dữ liệu bị mất	Không tìm thấy dữ liệu bị mất
Lựa chọn đặc trưng	Gradient Boosting được chọn là thuật toán lựa chọn đặc trưng. Số lượng đặc trưng mà Gradient Boosting giữ lại còn 54.

MÔ HÌNH HỌC MÁY

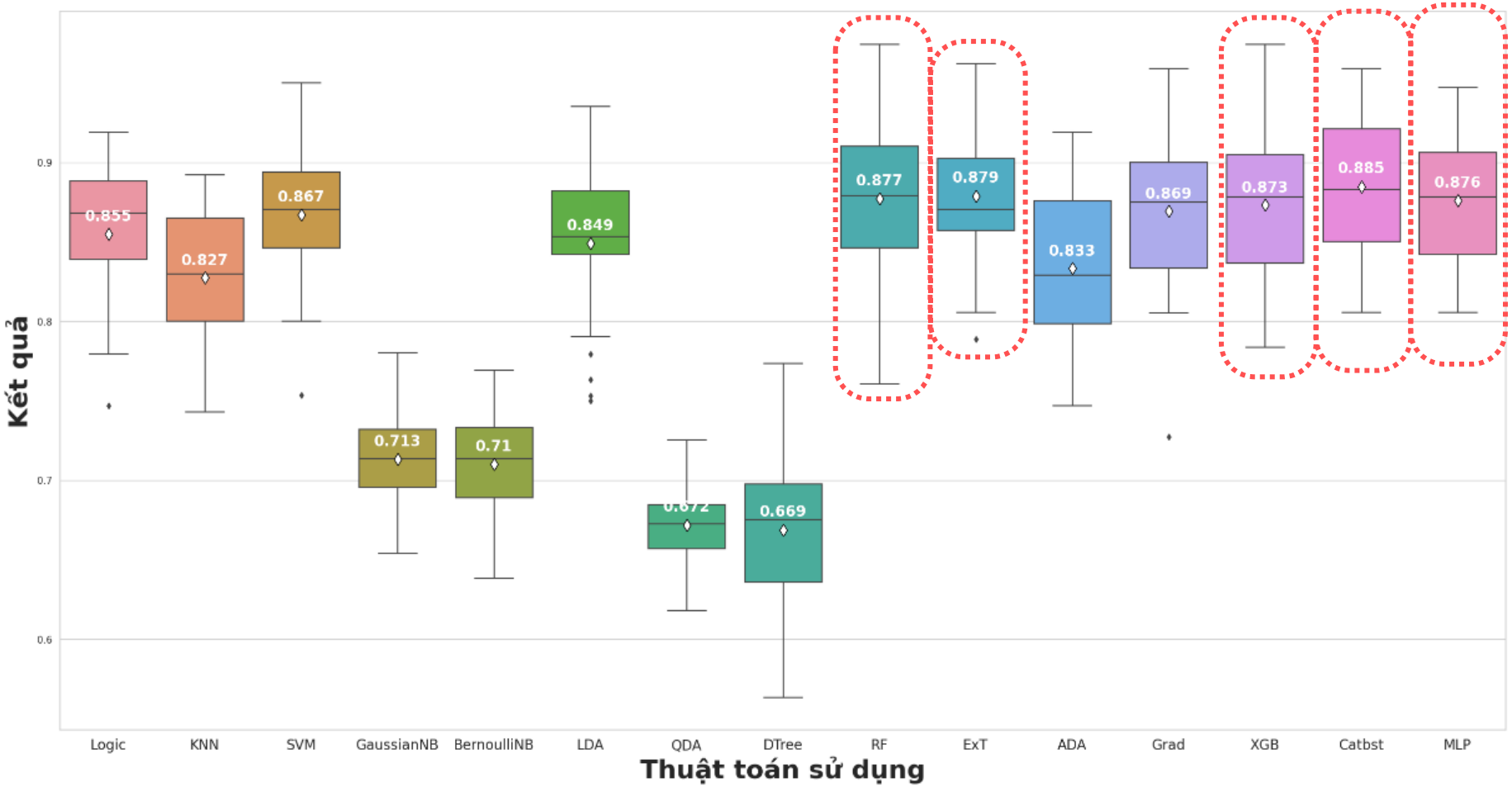
3

So sánh hiệu năng thuật toán

Biểu đồ nhiệt Wilcoxon



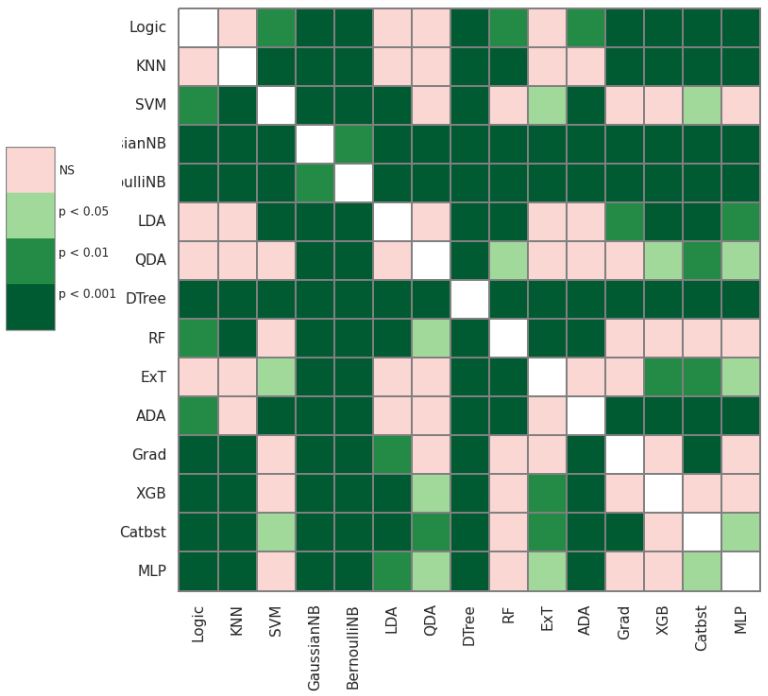
So sánh kết quả CV-f1 score giữa các thuật toán



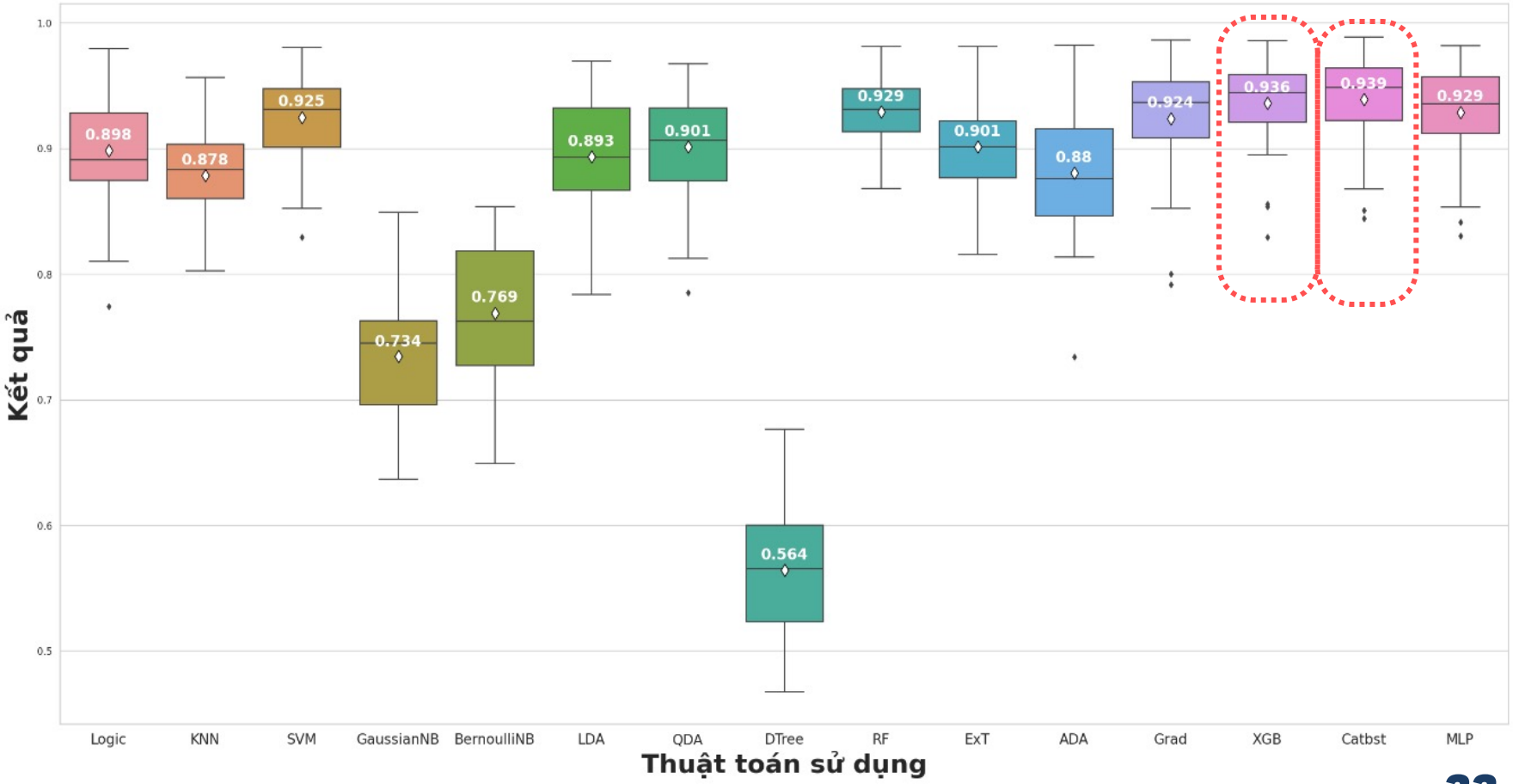
MÔ HÌNH HỌC MÁY

So sánh hiệu năng thuật toán

Biểu đồ nhiệt Wilcoxon



So sánh kết quả CV-AP giữa các thuật toán



MÔ HÌNH HỌC MÁY

So sánh hiệu năng thuật toán

Mô hình cuối cùng

MÔ HÌNH HỌC MÁY

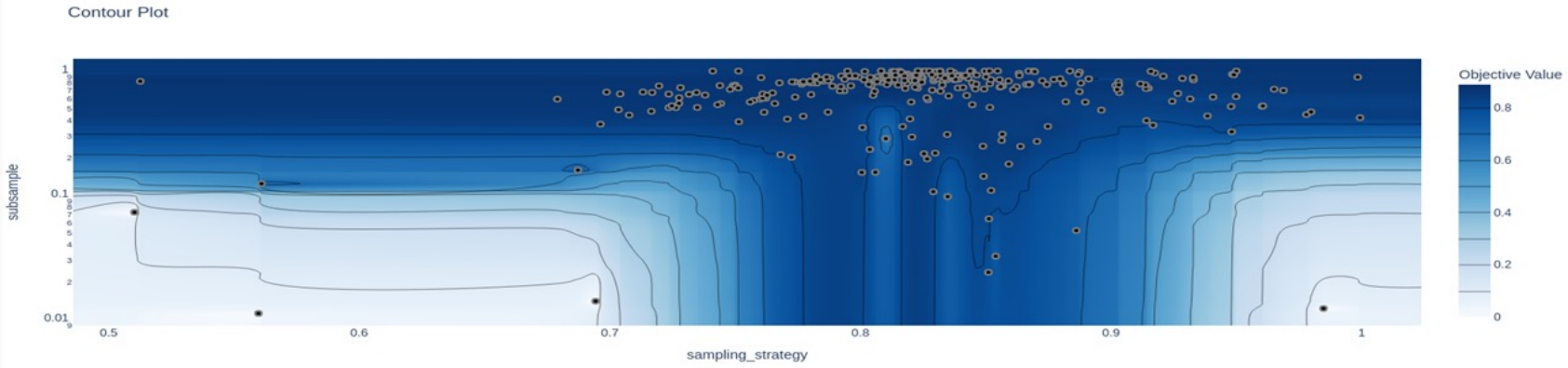
Tối ưu hóa siêu tham số trong 300 thử nghiệm.

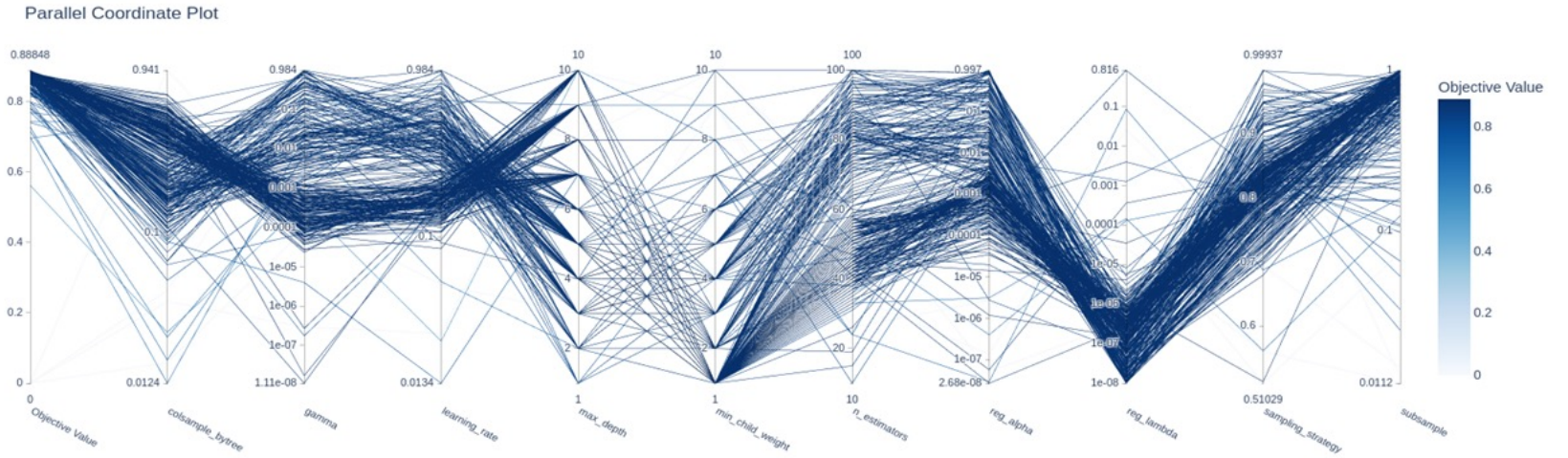
Giá trị tối ưu: CV-f1 score

Thời gian: 15 phút

Các siêu tham số bao gồm:
sampling_strategy, max_depth, min_child_weight, learning_rate, n_estimators, gamma, reg_alpha, reg_lambda, subsample, colsample_bytree

Tối ưu hóa mô hình





MẠNG THẦN KINH NHÂN TẠO (ANN)

Tối ưu hóa mô hình

Tối ưu hóa trong 100 thử nghiệm.

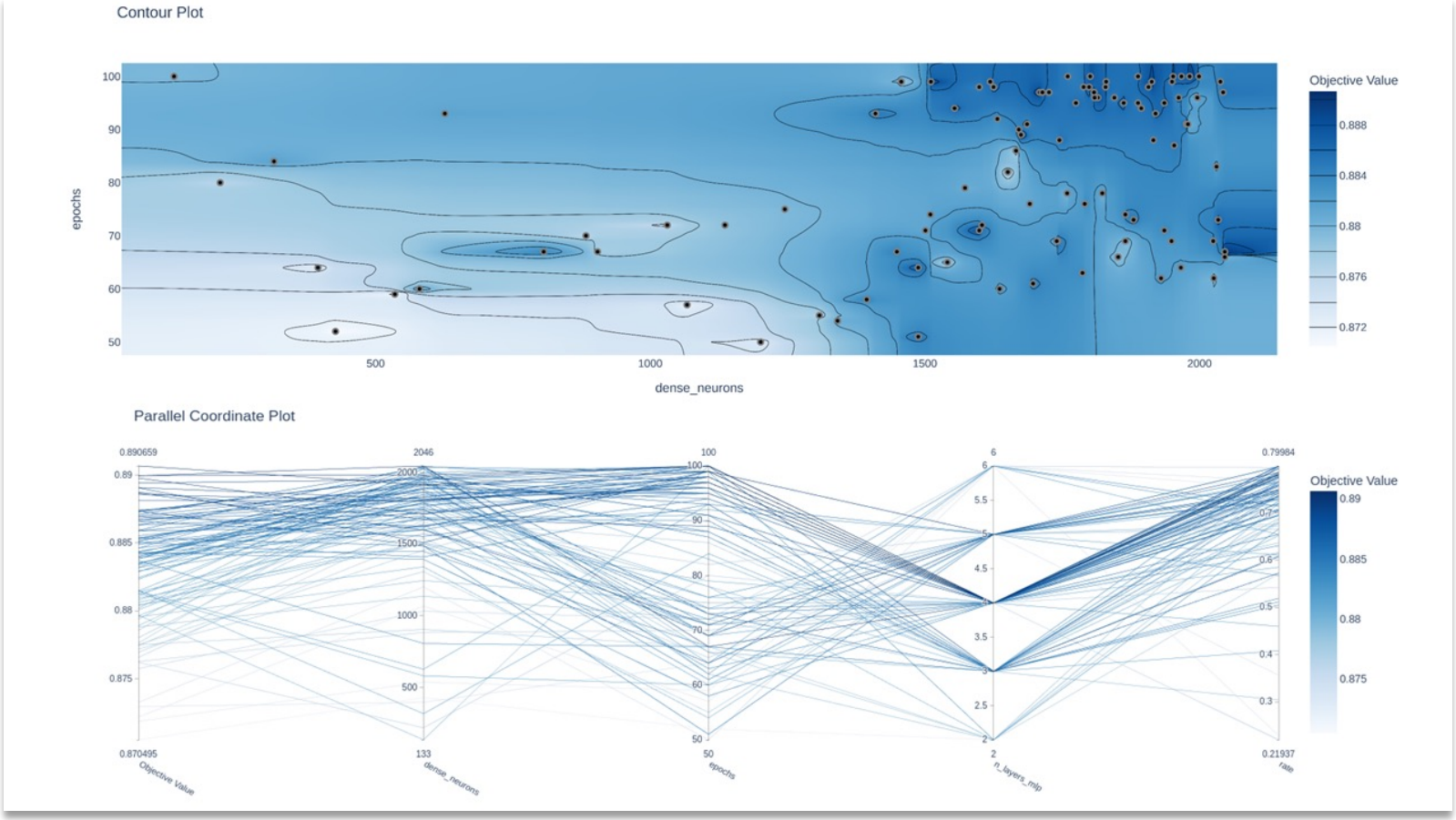
Giá trị tối ưu: CV-f1 score

Thời gian: 8 giờ.

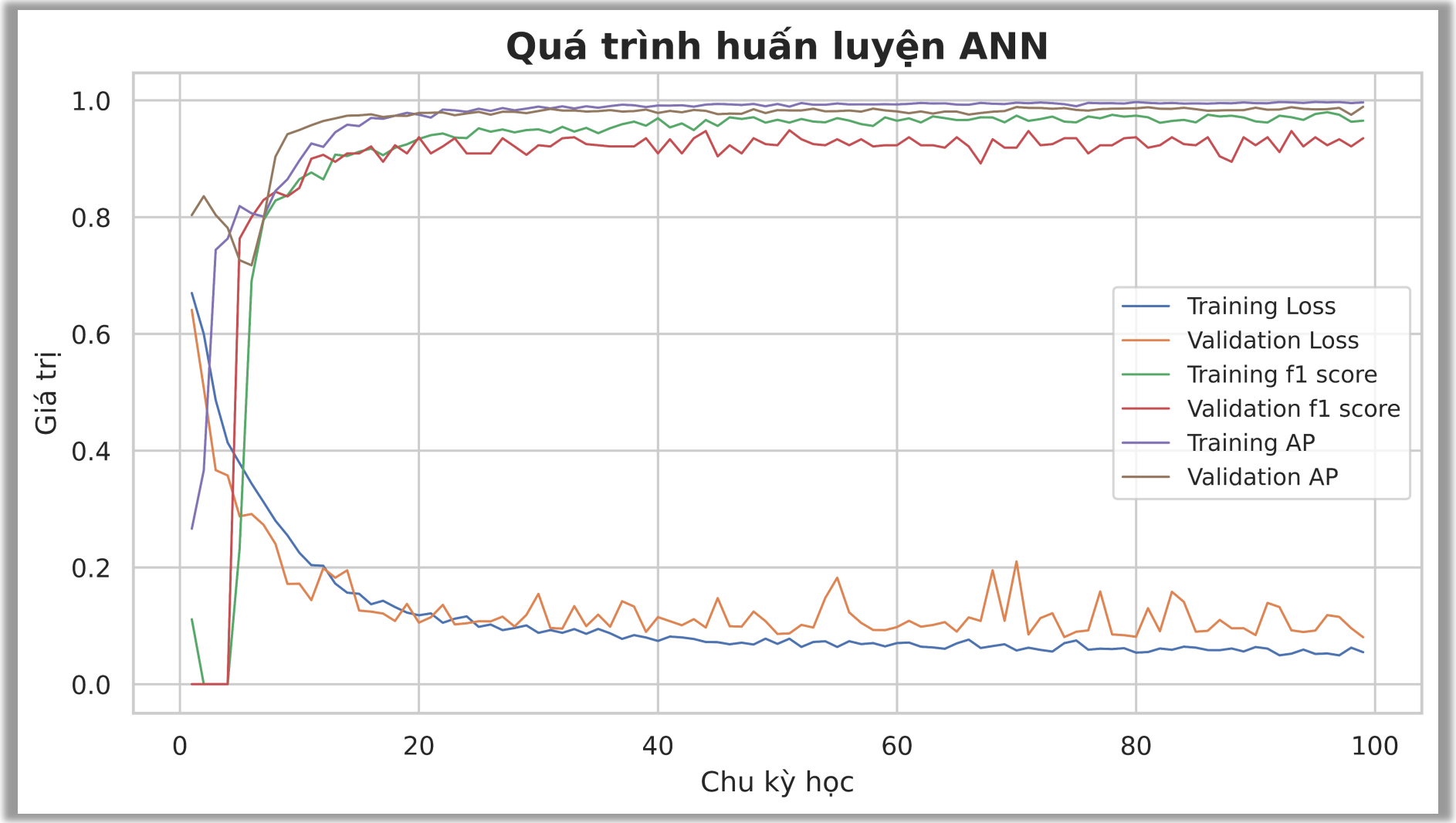
Các siêu tham số bao gồm: *số lớp ẩn, số nơ-ron lớp ẩn, tỉ lệ dropout và chu kỳ học.*

Hàm mất mát: **Binary Cross Entroy**

Hàm tối ưu: **Adam**, tốc độ học = 0,0001, weight_decay = 0,01.

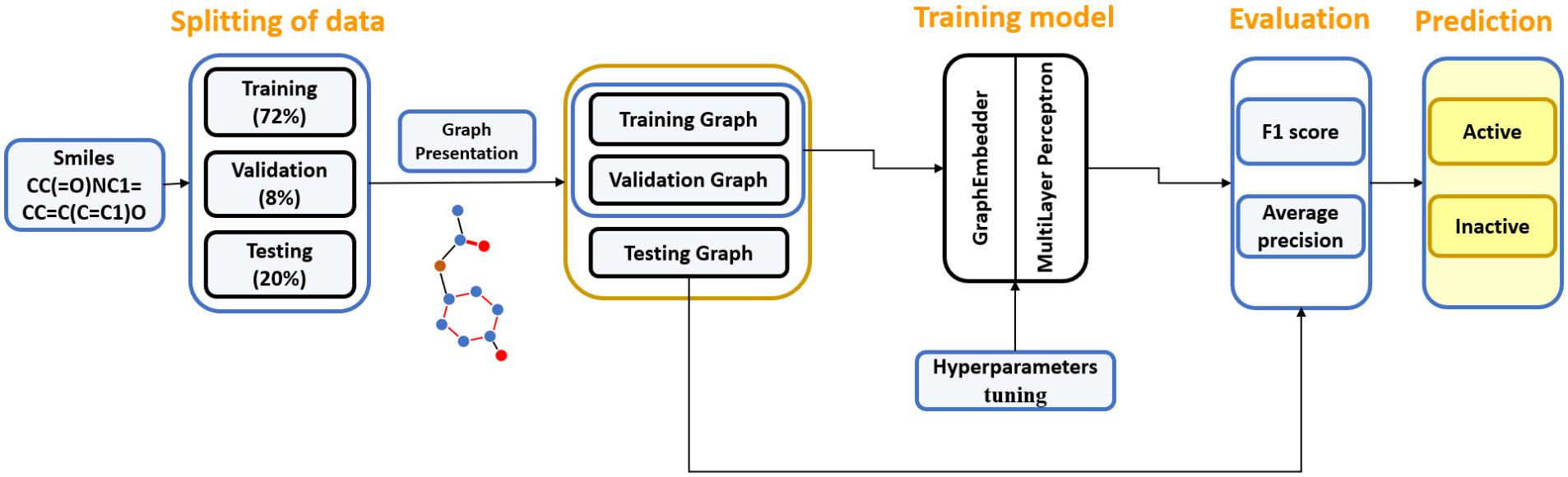


Thông số huấn luyện mô hình



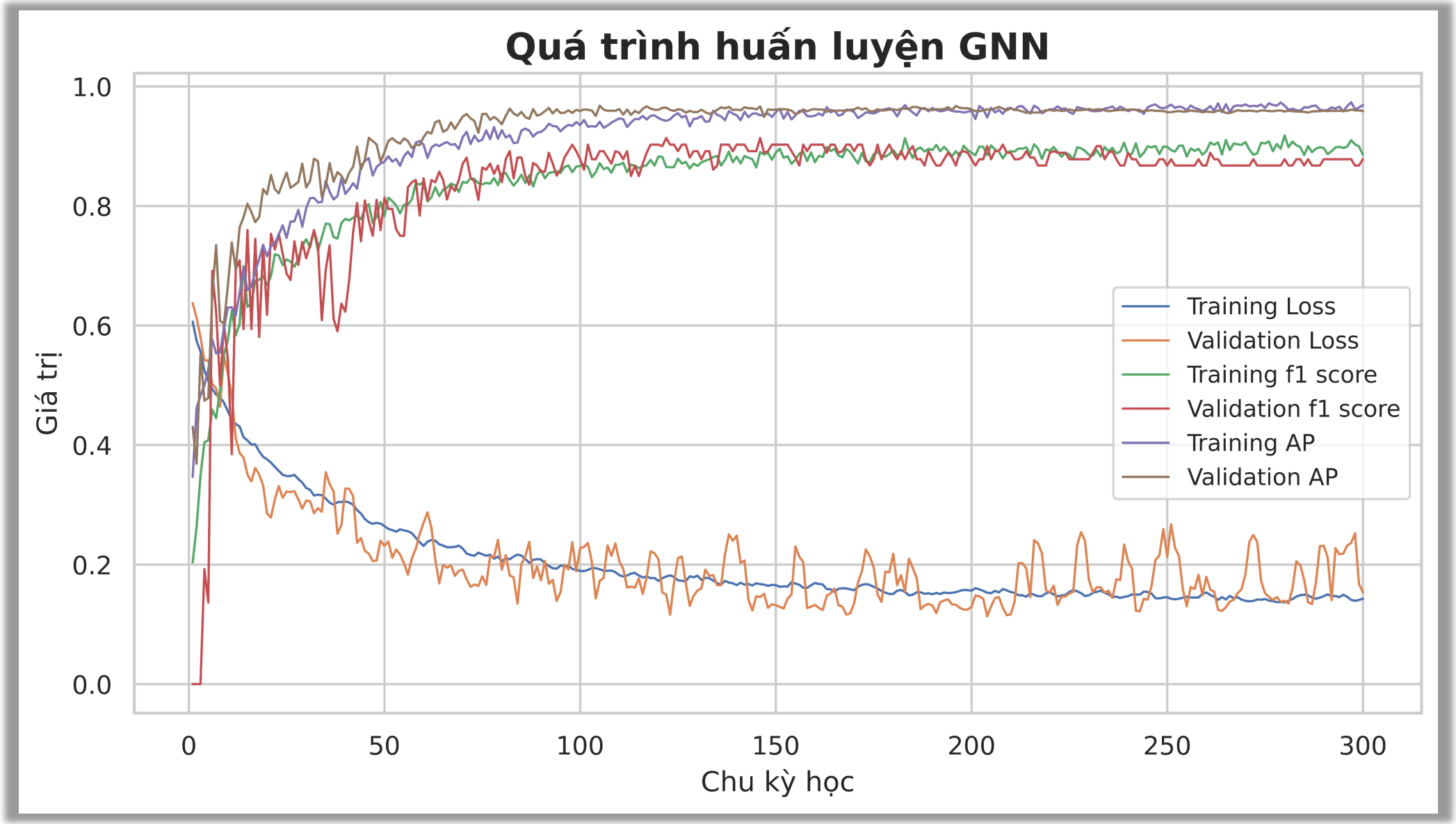
MẠNG THẦN KINH ĐỒ THỊ (GNN)

VẤN ĐỀ



- Tốn rất nhiều tài nguyên tính toán.
- Sai số của tập đánh giá nội (validation loss) dao động lớn
- Pytorch Geometric không hỗ trợ khả năng tái tạo mô hình(reproducibility) khi huấn luyện trên GPU.

MẠNG THẦN KINH ĐỒ THỊ (GNN)



MẠNG THẦN KINH ĐỒ THỊ (GNN)

Tối ưu hóa trong 50 thử nghiệm.

Giá trị tối ưu: validation loss

Thời gian: 16 giờ

Các siêu tham số bao gồm:
batch_size, n_blocks, embedding_size, drop_out_rate, top_k_ratio, number_hidden_node, ann_dropout_rate, sgd_momentum, scheduler_gamma

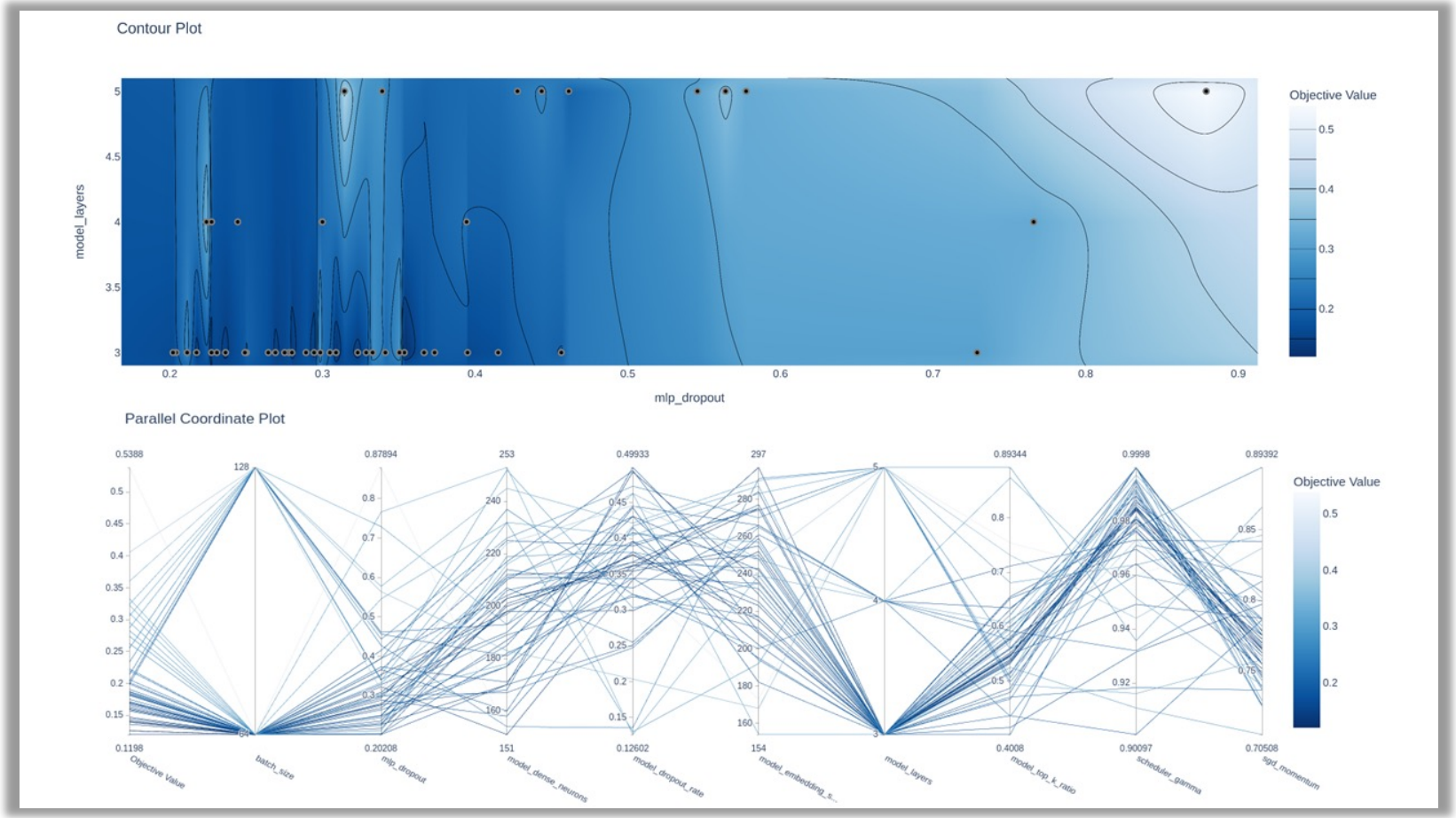
Hàm mất mát: Binary cross entropy

Hàm tối ưu: SGD, tốc độ học 0,01 và weight_decay 0,0001, momentum

ExponentialLR một cơ chế điều chỉnh tốc độ học theo lịch trình,

Số chu kỳ học: 300

Tối ưu hóa mô hình

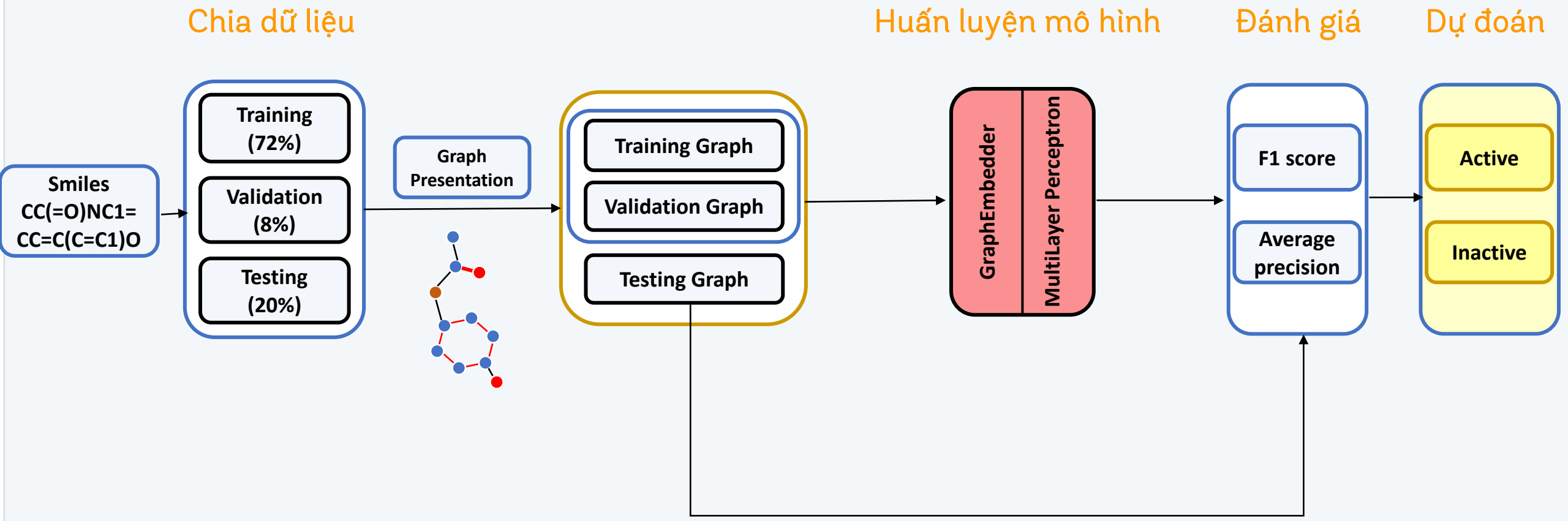


MẠNG THẦN KINH ĐỒ THỊ (GNN)

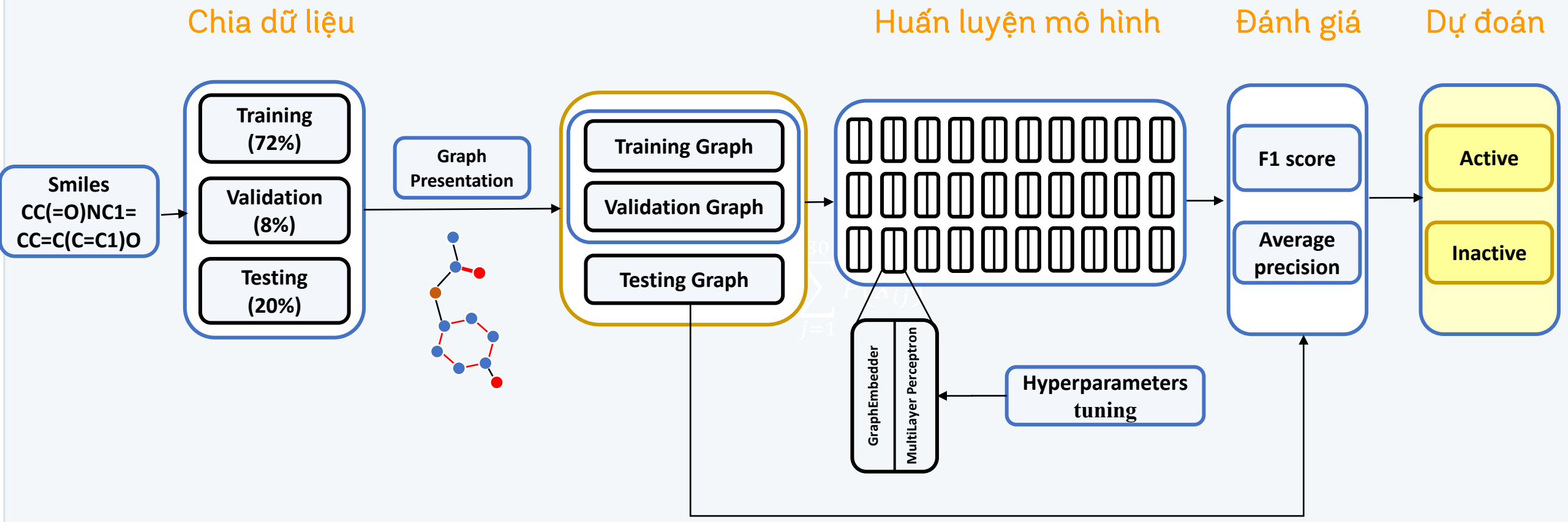
Phân tích STD của 100 chu kỳ

ID TRIALS	MEAN	STD	MIN	MAX	LASTED LOSS
17	0,226	0,047	0,159	0,437	0,171
29	0,195	0,050	0,136	0,398	0,140
15	0,182	0,060	0,121	0,376	0,125
30	0,238	0,061	0,167	0,437	0,168
19	0,174	0,071	0,109	0,423	0,119
18	0,214	0,076	0,139	0,492	0,139
21	0,192	0,077	0,117	0,488	0,141

MẠNG THẦN KINH ĐỒ THỊ (GNN)



MẠNG THẦN KINH ĐỒ THỊ (GNN)



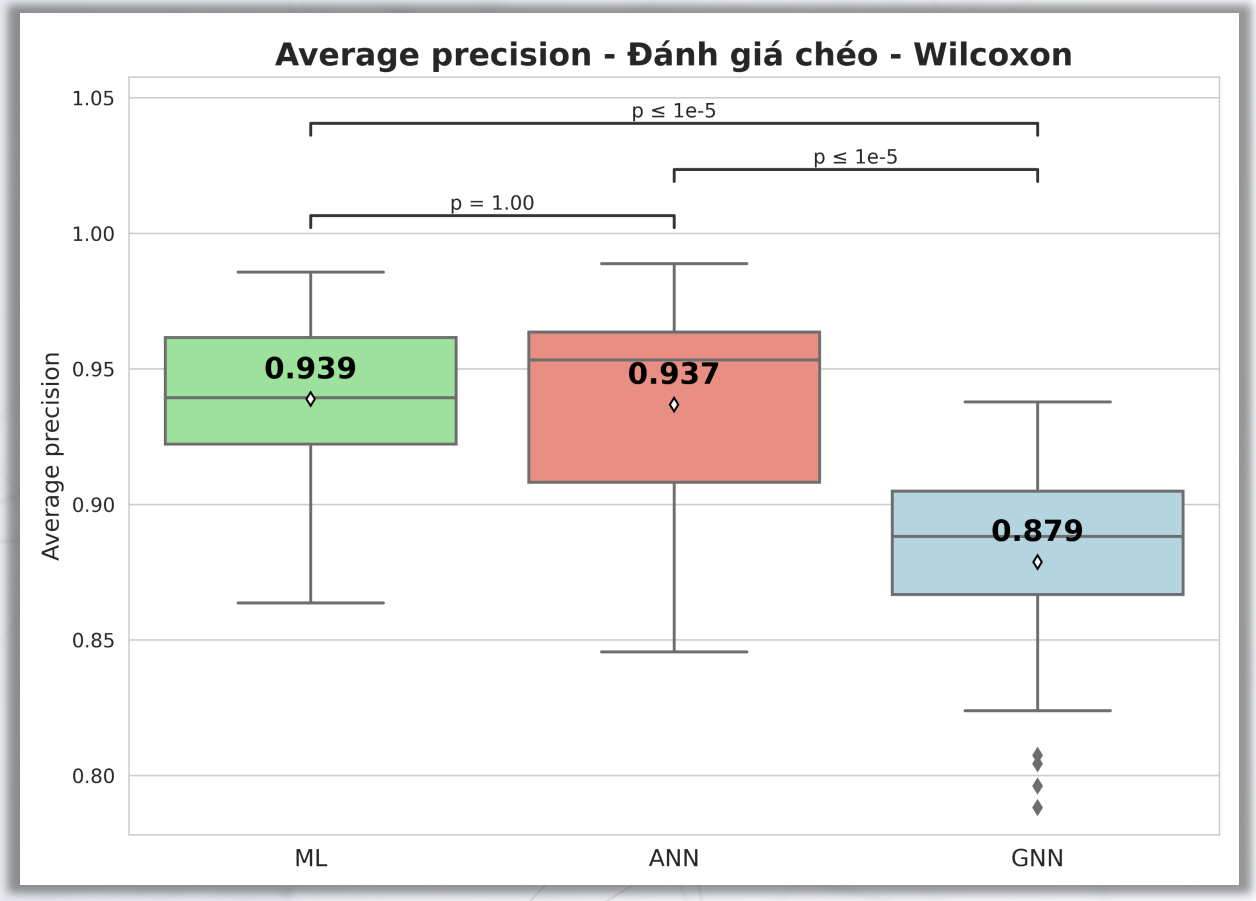
EV-f1 score = 0,808±0,009

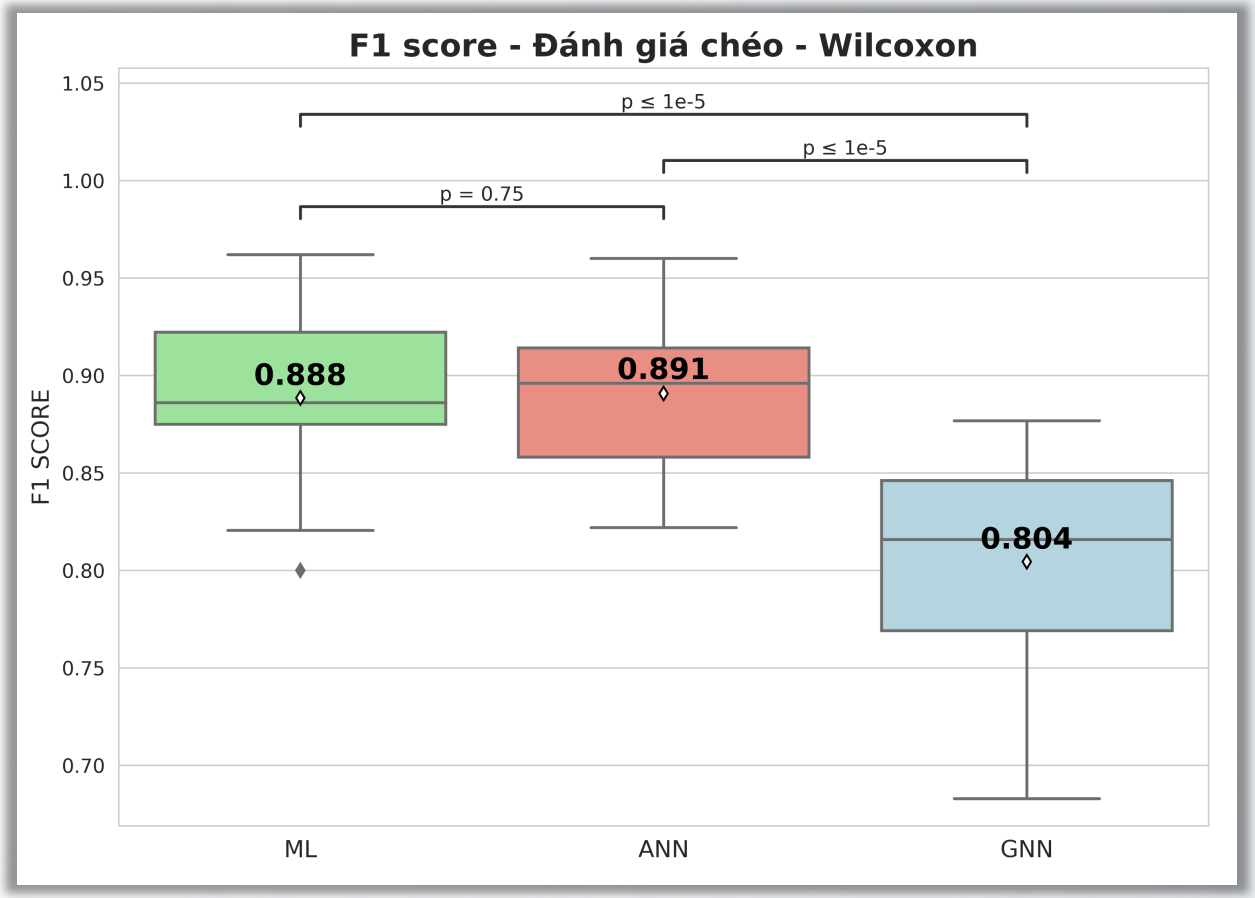
EV-AP = 0,827±0,013

KẾT QUẢ 3 MÔ HÌNH

	HỌC MÁY	ANN	E-GNN
CV-f1 score	0,888±0,039	0,891±0,037	0,804±0,049
EV-f1 score	0,921	0,930	0,808±0,009
CV-AP	0,939±0,032	0,934±0,040	0,879±0,041
EV-AP	0,961	0,955	0,827±0,013
Thời gian huấn luyện	< 2 phút	< 3 phút	3 giờ
Thời gian tối ưu	15 phút	8 giờ	16 giờ
Số thử nghiệm	300	100	50
Thiết bị	CPU	GPU	GPU
Triển khai ứng dụng	Phức tạp	Đơn giản	Đơn giản

KẾT QUẢ 3 MÔ HÌNH





DOCKING PHÂN TỬ

Re-docking

Kết quả đánh giá khả năng tái tạo dữ liệu của phần mềm được thể hiện qua giá trị RMSD của quá trình re-docking của phối tử alectinib so với phối tử đồng kết tinh của protein.

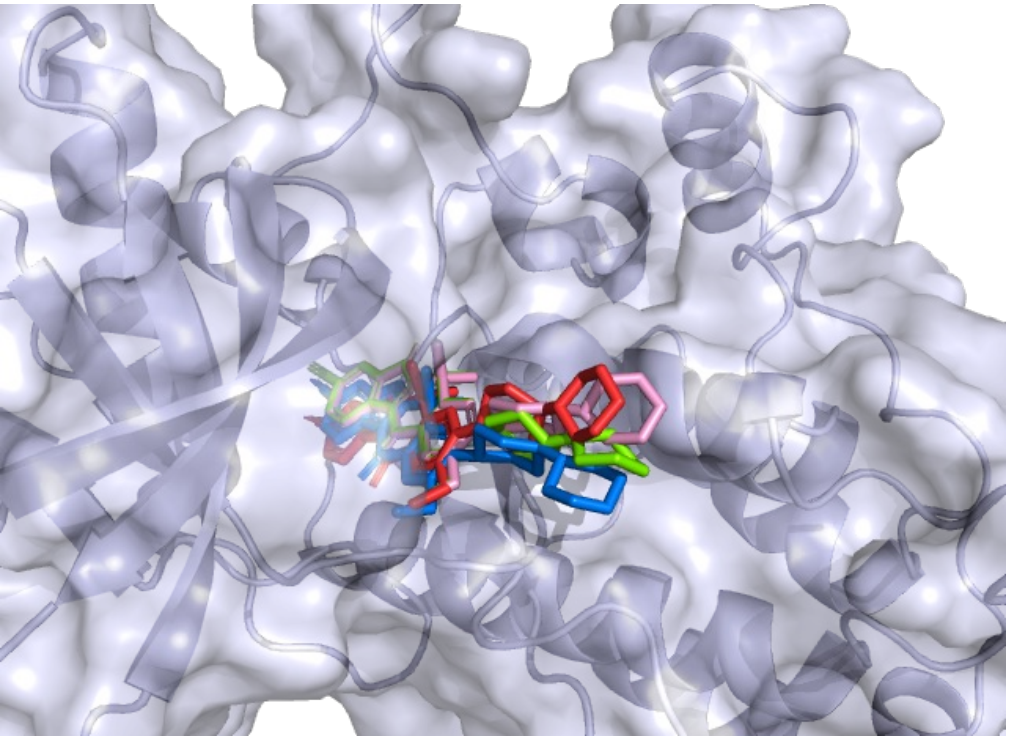
Cả 3 phần mềm **Vina-GPU-2.0, GNINA, Autodock-GPU** đều cho giá trị RMSD dưới 2Å.

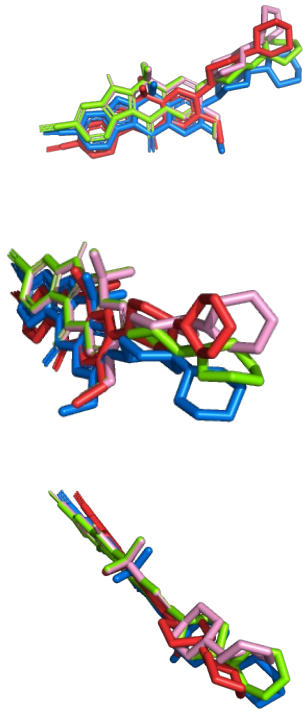
PHẦN MỀM DOCKING	RMSD (Å)	ĐIỂM SỐ DOCKING CẤU DẠNG TỐT NHẤT
Vina-GPU-2.0	1,605	-9,9 (kcal/mol)
GNINA	1,584	GNINA affinity: 7,487 CNN_pose_score: 0,86
Autodock-GPU	1,64	-8,9 (kcal/mol)

DOCKING PHÂN TỬ

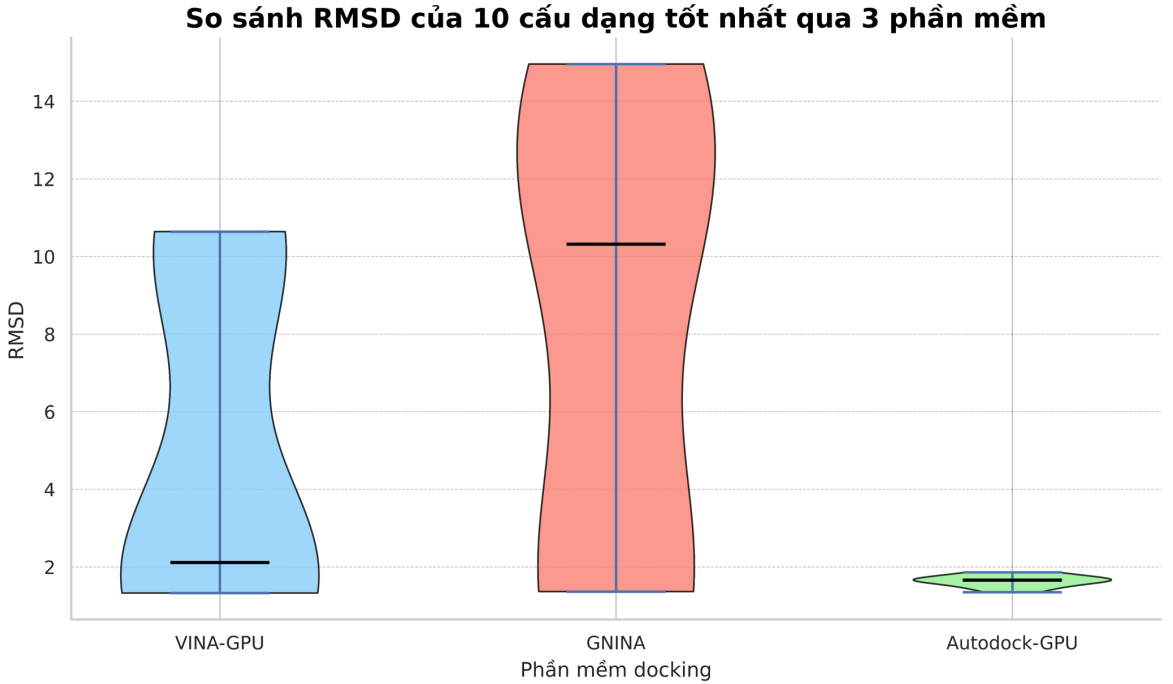
Re-docking

Cấu dạng của alectinib của 3 phần mềm





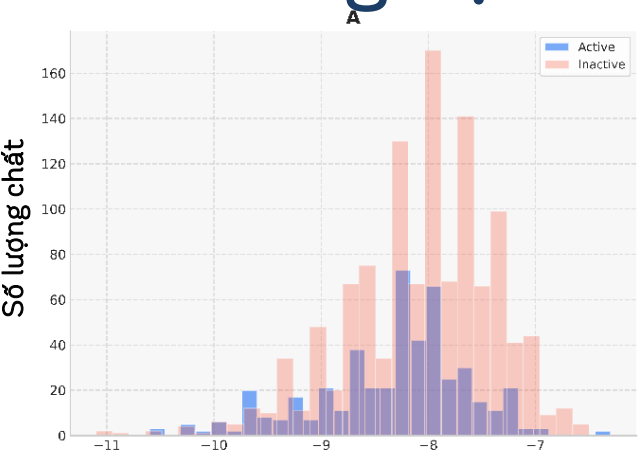
Sự tương đồng về cấu dạng alectinib qua 3 phần mềm. Trong đó màu đỏ là cấu dạng của phối tử đồng kết tinh, màu xanh lá là phần mềm GNINA,màu xanh dương là autodock-GPU và màu hồng là Vina-GPU.

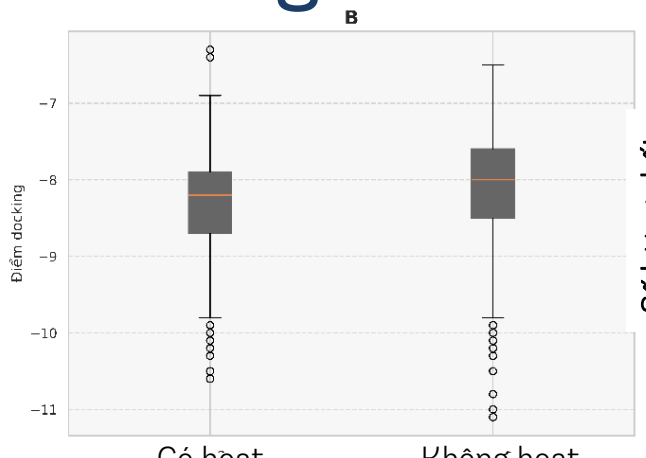


Biểu đồ Violin thể hiện phân bố giá trị RMSD của 10 cấu dạng tốt nhất được các phần mềm tìm ra.

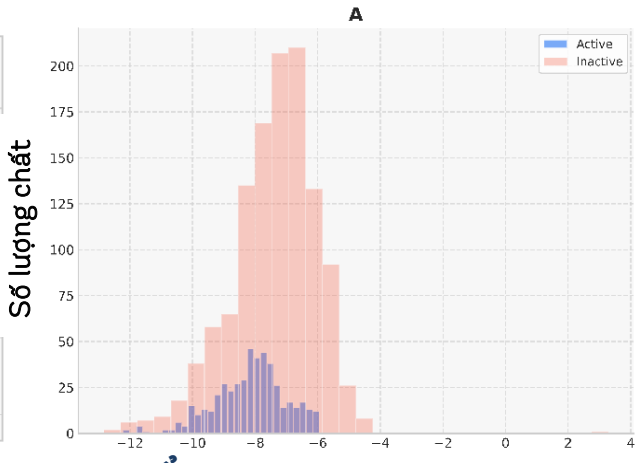
DOCKING PHÂN TỬ

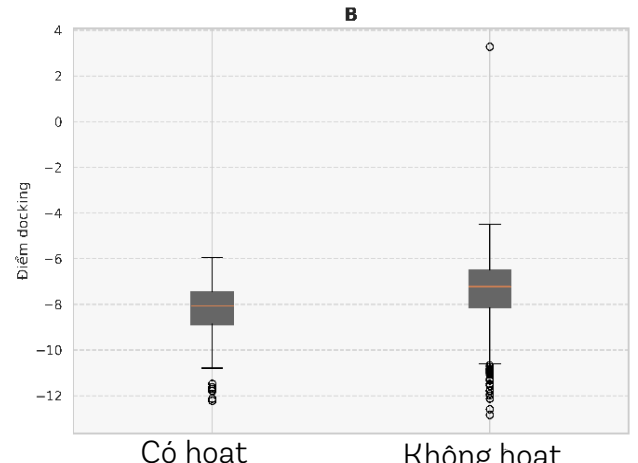
Thử nghiệm hồi chứng



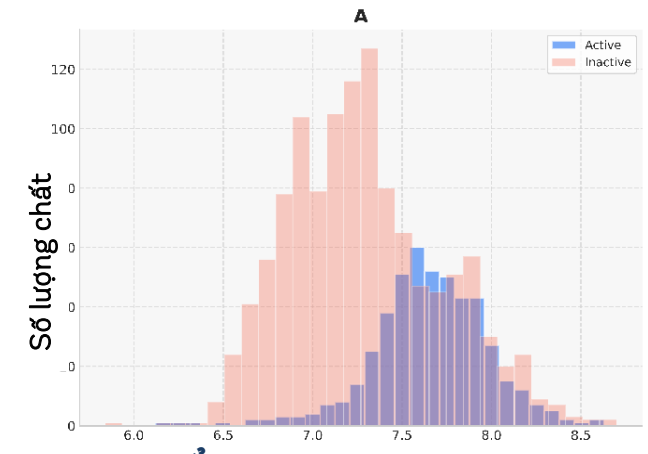


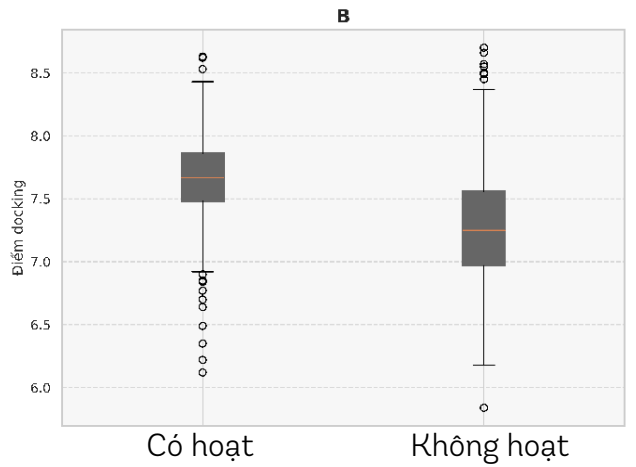
VINA-GPU



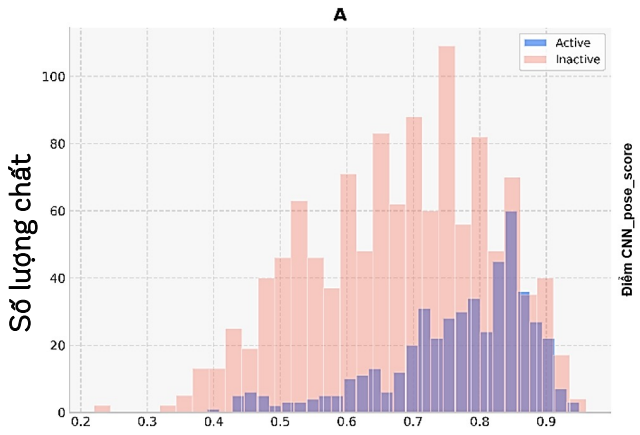


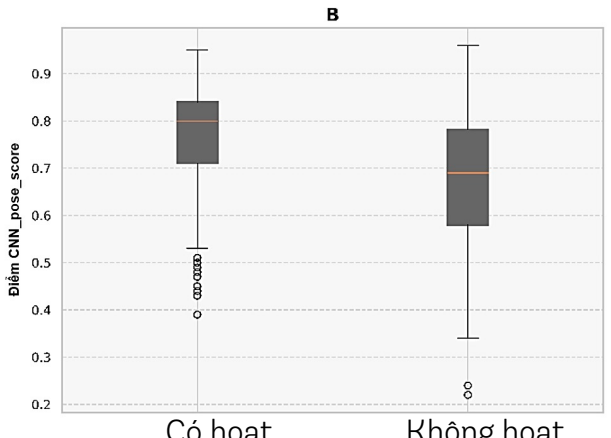
ĐIỂM AUTODOCK-GPU





ĐIỂM GNINA AFFINITY

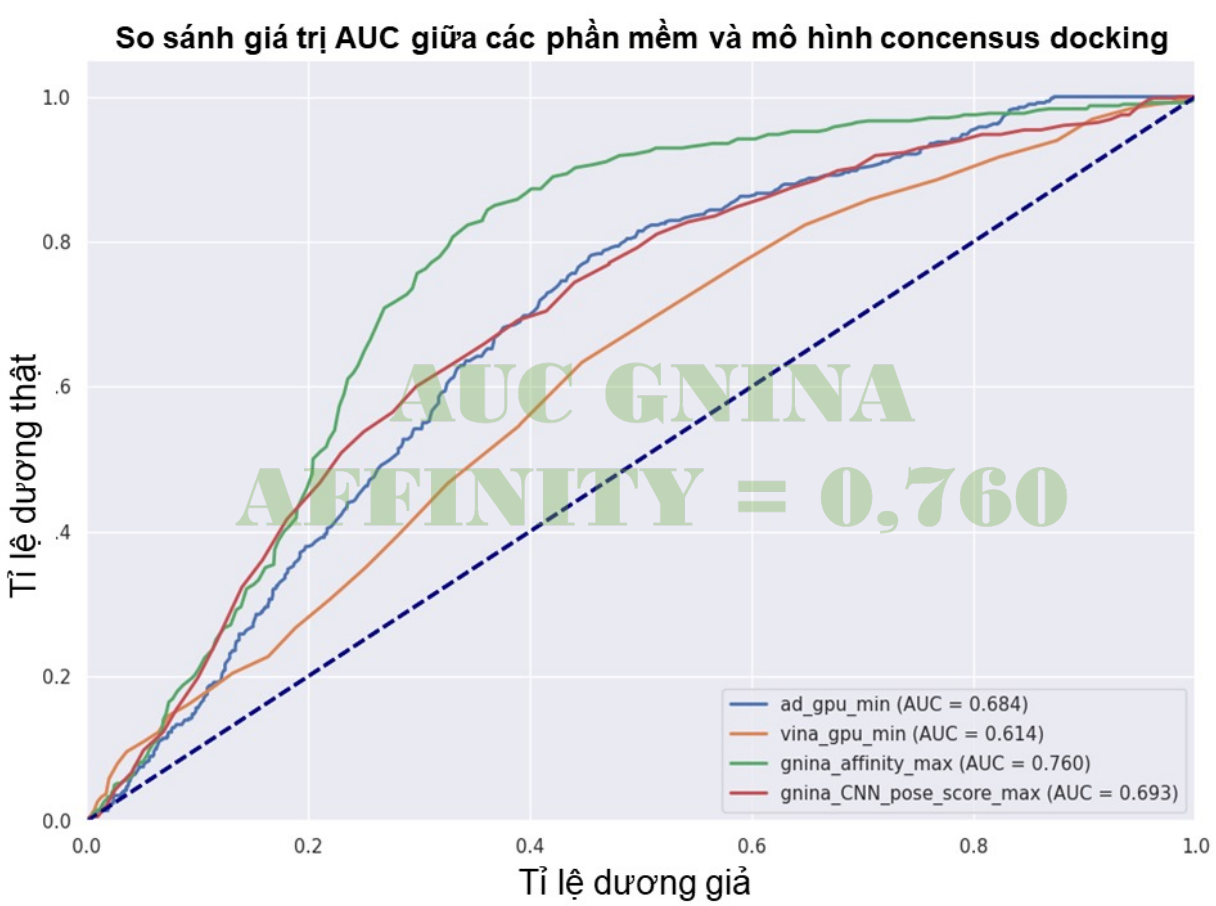


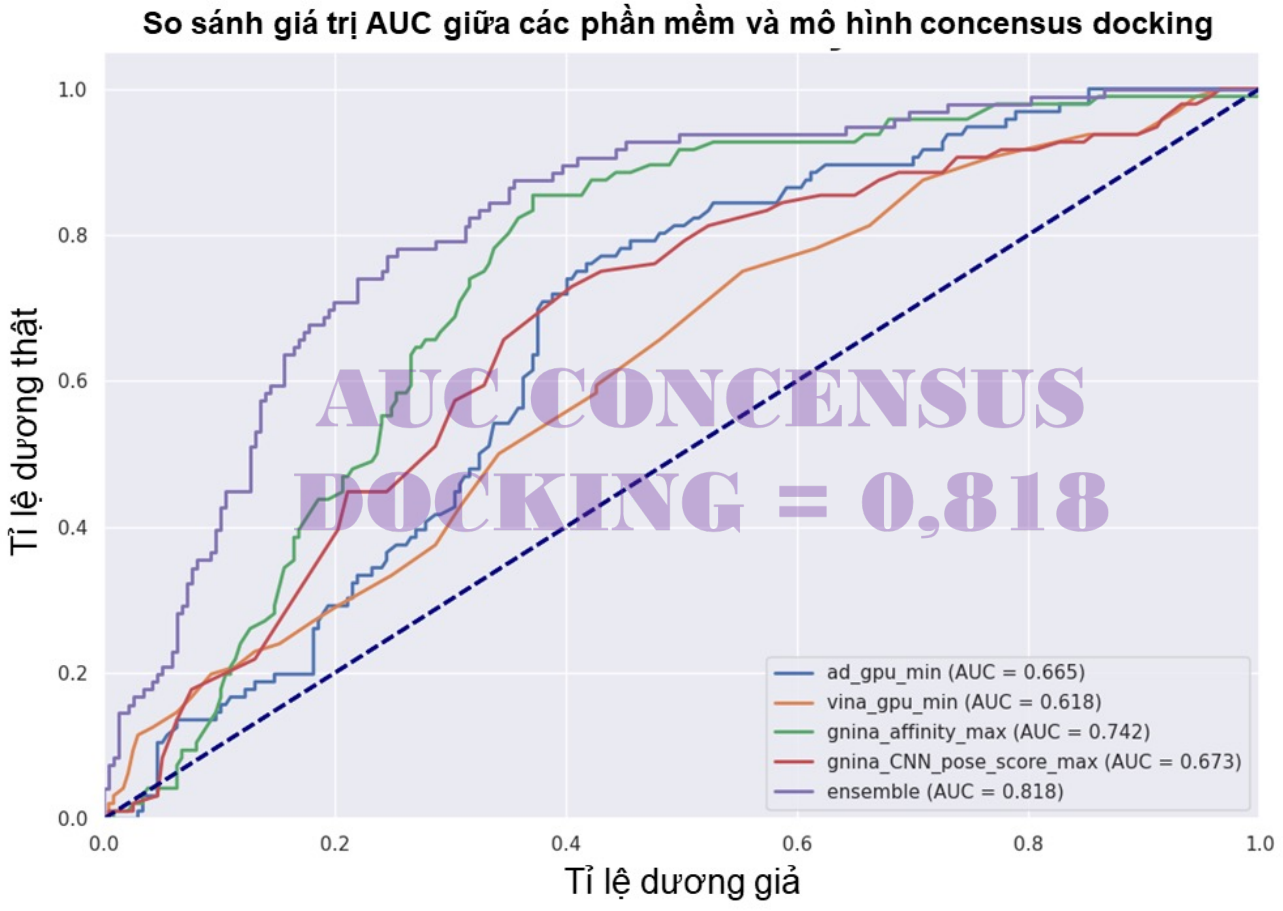


ĐIỂM GNINA_POSE_SCORE

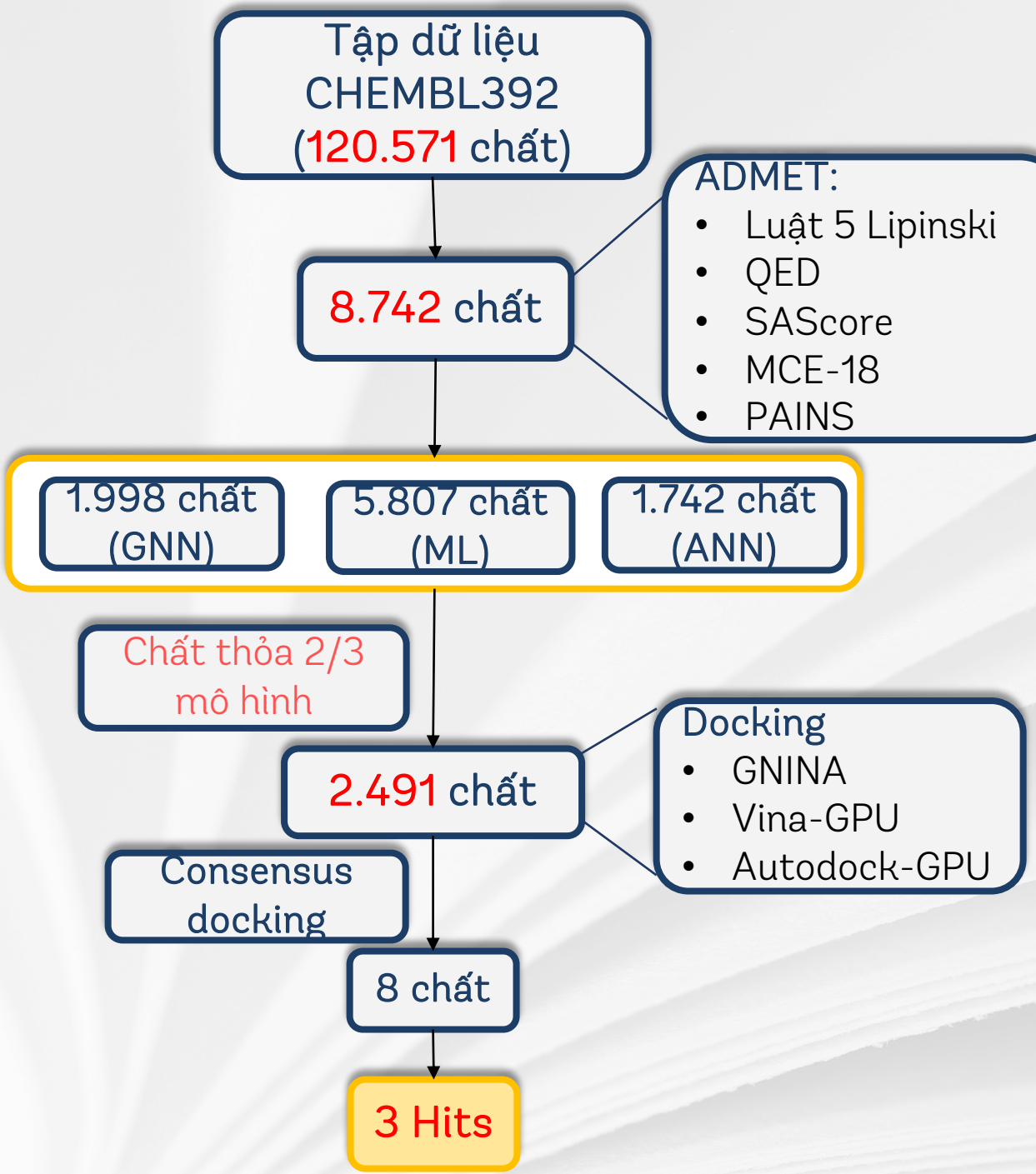
DOCKING PHÂN TỬ

Thử nghiệm hồi chứng





SÀNG LỌC ẢO

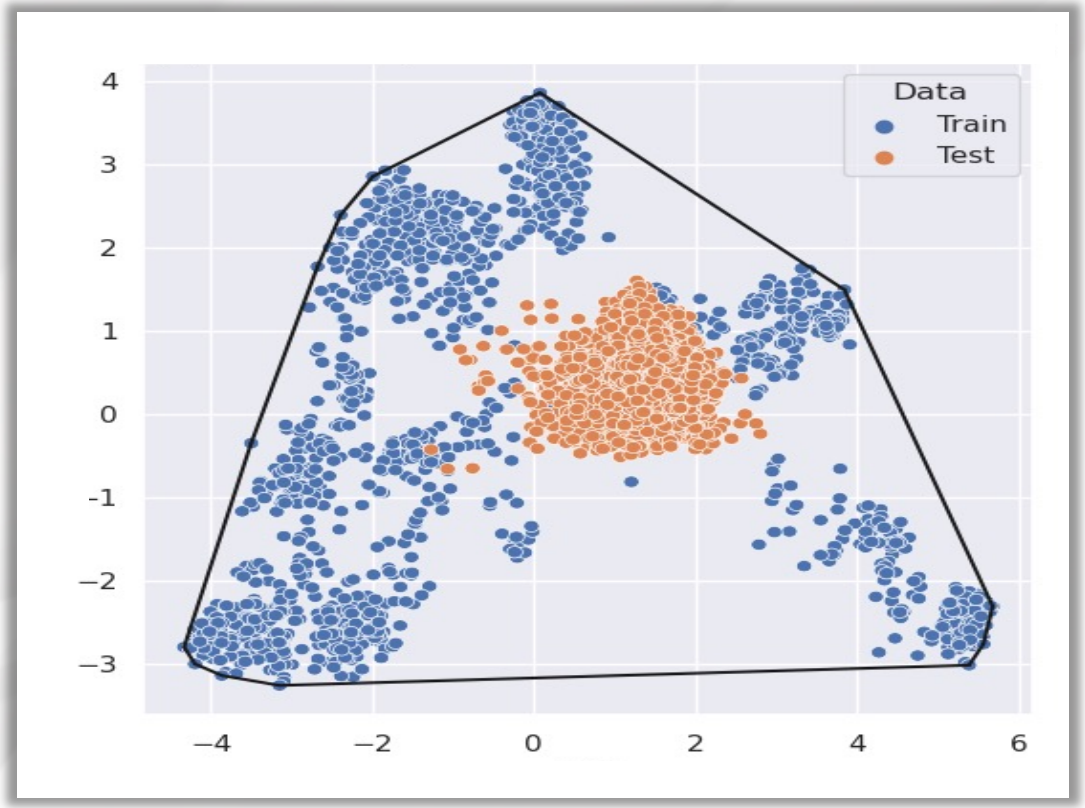

Tập dữ liệu CHEMBL392 (120.571 chất)
8.742 chất
ADMET:
Luật 5 Lipinski
QED
SAScore
MCE-18
PAINS
1.998 chất (GNN)
5.807 chất (ML)
1.742 chất (ANN)
Chất thỏa 2/3 mô hình
2.491 chất
Docking
GNINA
Vina-GPU
Autodock-GPU
Consensus docking
8 chất
3 Hits

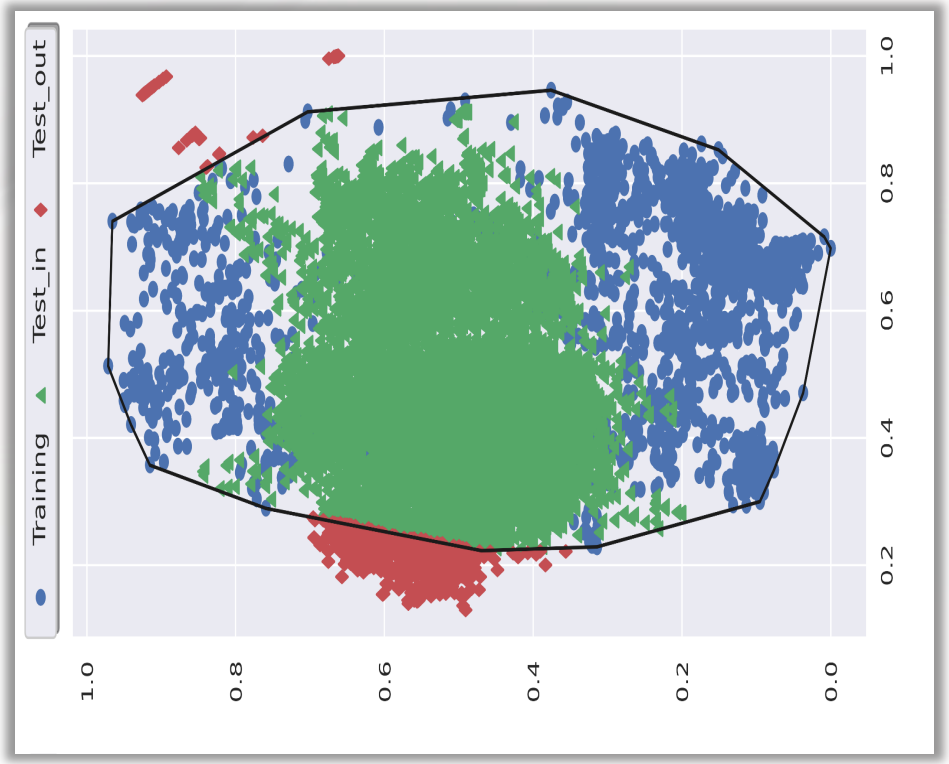
SÀNG LỌC ẢO

Miền ứng dụng

DẤU VÂN TAY + PCA + HÀM BAO LỒI

Toàn bộ đều thuộc miền ứng dụng.
Tuy nhiên, có nhiều vùng trống trong hàm bao lồi mà các chất dự đoán xuất hiện ở đó

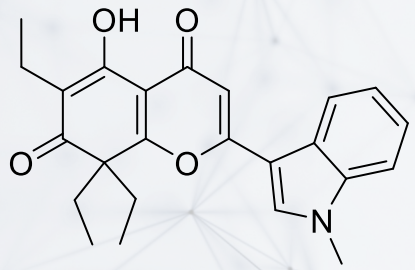
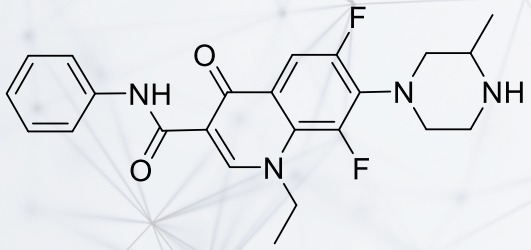
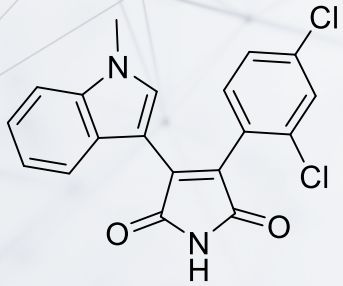




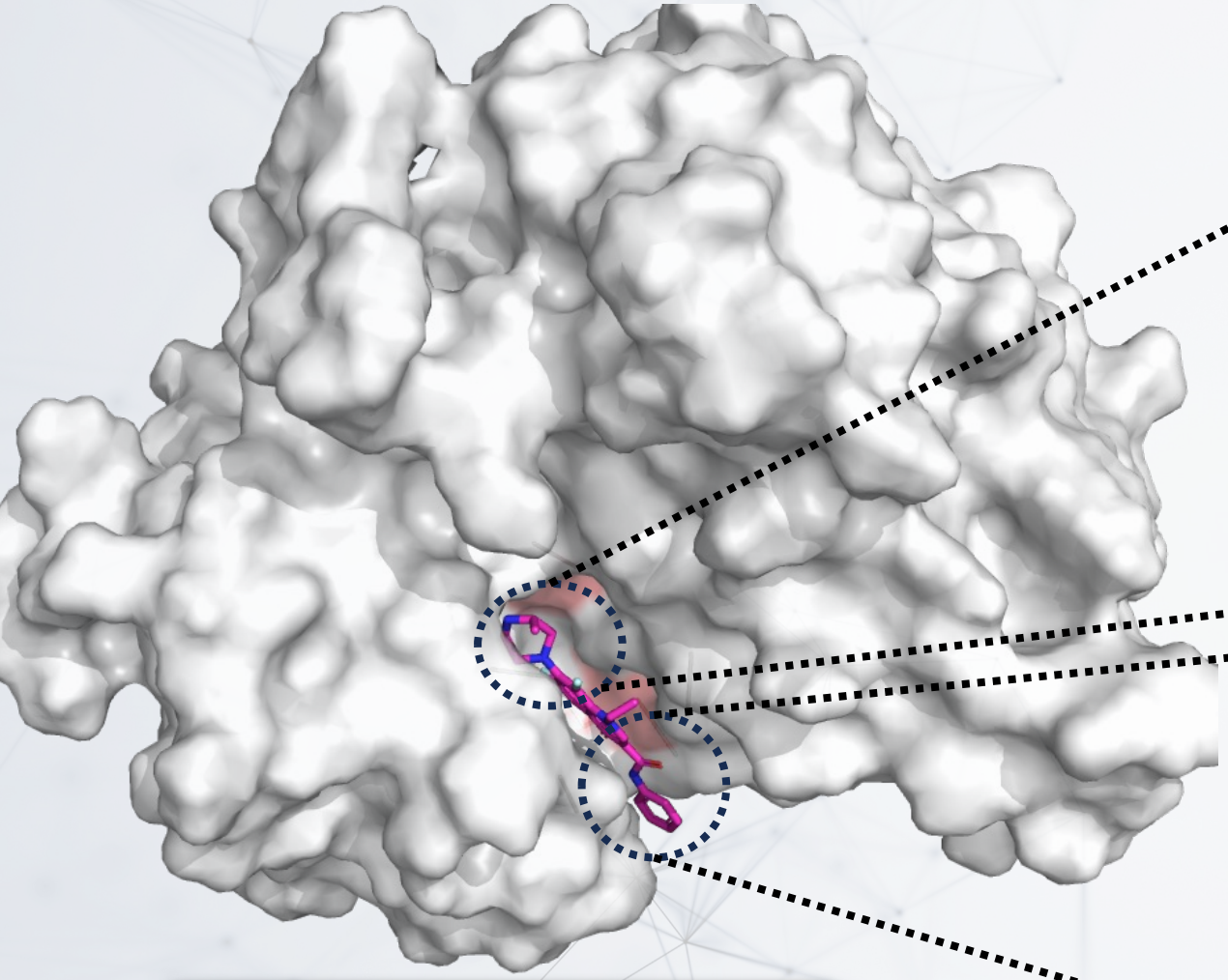
MA TRẬN TƯƠNG ĐỒNG TANIMOTO + MDS + HÀM BAO LỒI

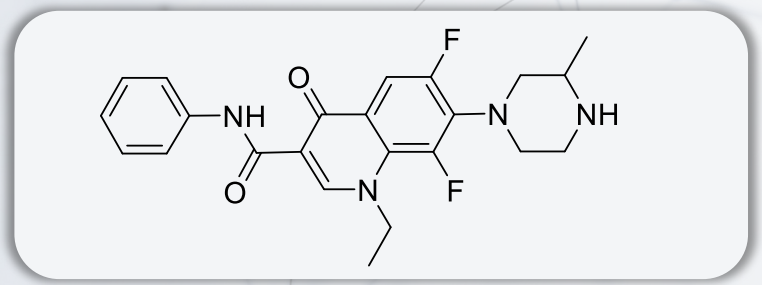
862 cấu trúc được phát hiện nằm ngoài miền ứng dụng.
Thời gian tính toán: 3 giờ

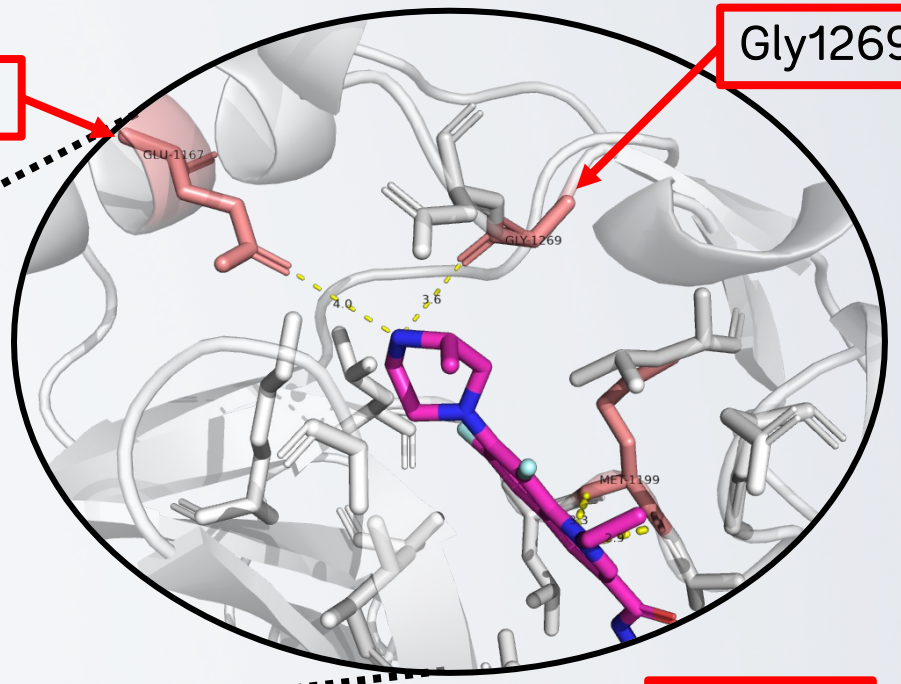
SÀNG LỌC ẢO

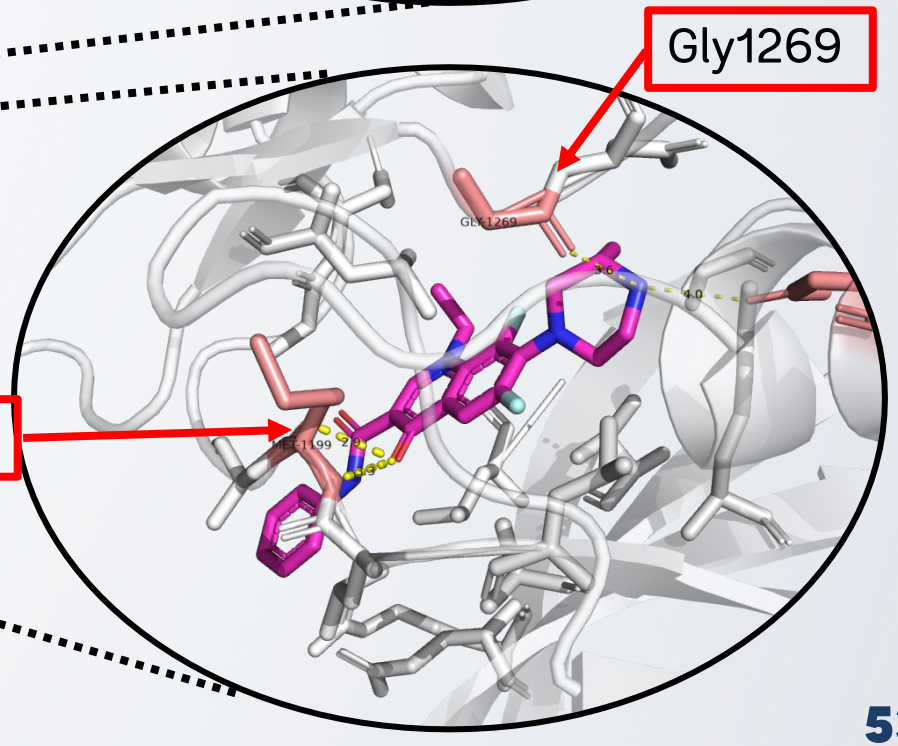
CHEMBL ID	CẤU TRÚC	XÁC SUẤT DỰ ĐOÁN CONCENSUS DOCKING	ĐIỂM SỐ DOCKING	XÁC SUẤT DỰ ĐOÁN MÔ HÌNH PHÂN LOẠI
CHEMBL1689515		74,9%	AD-GPU: -8,44 kcal/mol Vina-GPU: -9,10 kcal/mol GNINA affinity: 7,36 CNN-score: 0,89	Học máy: 96,81% ANN: 97,04% E-GNN: 45,63% Trung bình: 79,83%
CHEMBL2380351		66,8%	AD-GPU: -8,61 kcal/mol Vina-GPU: -8,80 kcal/mol GNINA affinity: 7,83 CNN-score: 0,90	Học máy: 89,81% ANN: 77,88% E-GNN: 90,24% Trung bình: 85,98%
CHEMBL102714		65,9%	AD-GPU: -7,0 kcal/mol Vina-GPU : -9,30 kcal/mol GNINA affinity: 7,32 CNN-score: 0,85	Học máy: 93,03% ANN: 85,66% E-GNN: 51,86% Trung bình: 76,85%

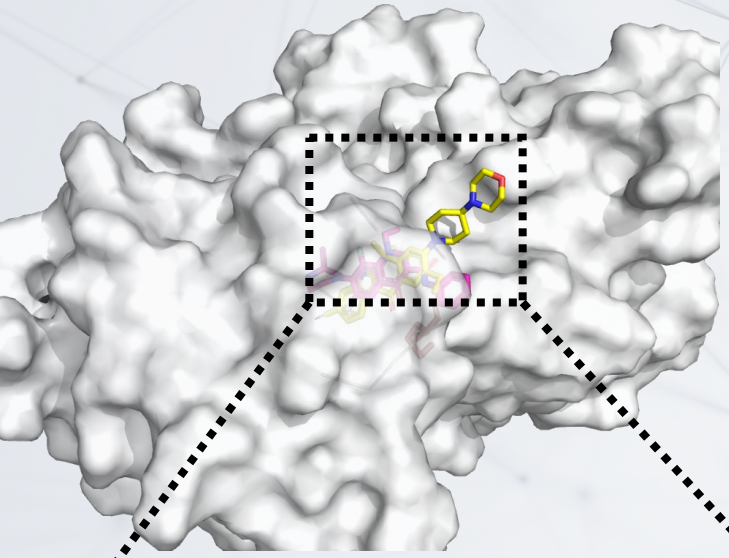
CHEMBL2380351

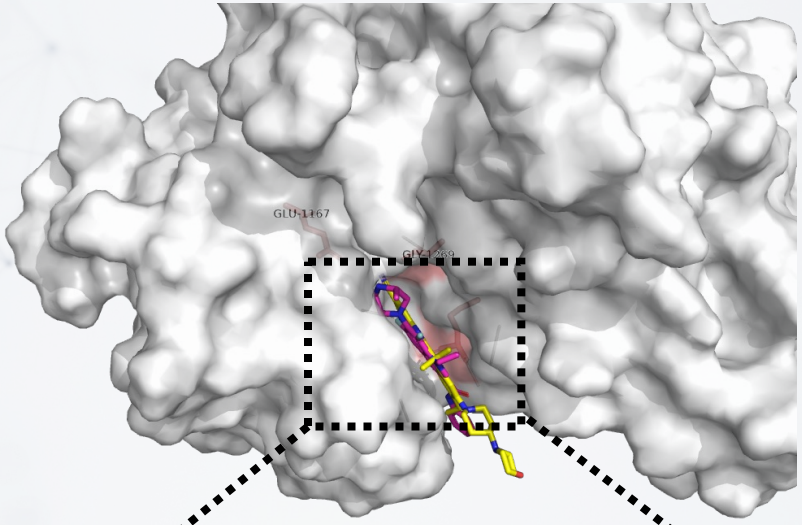


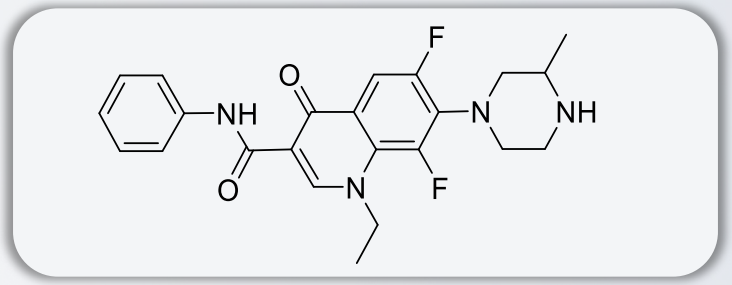


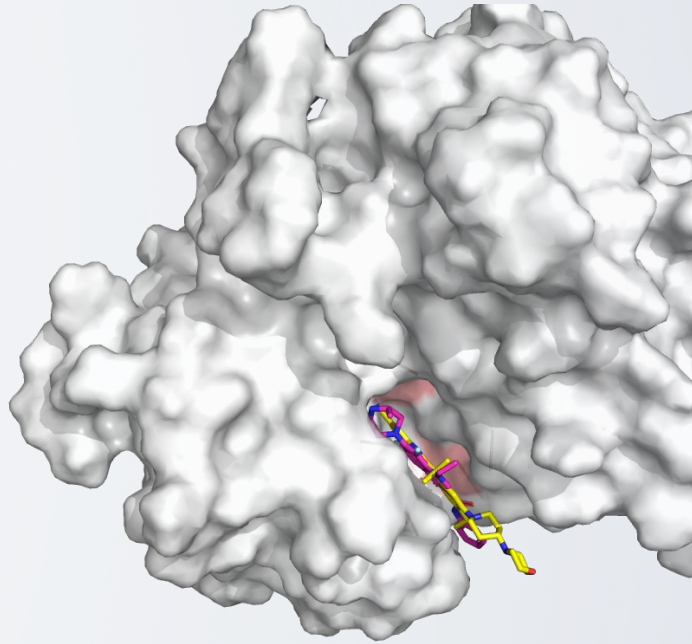


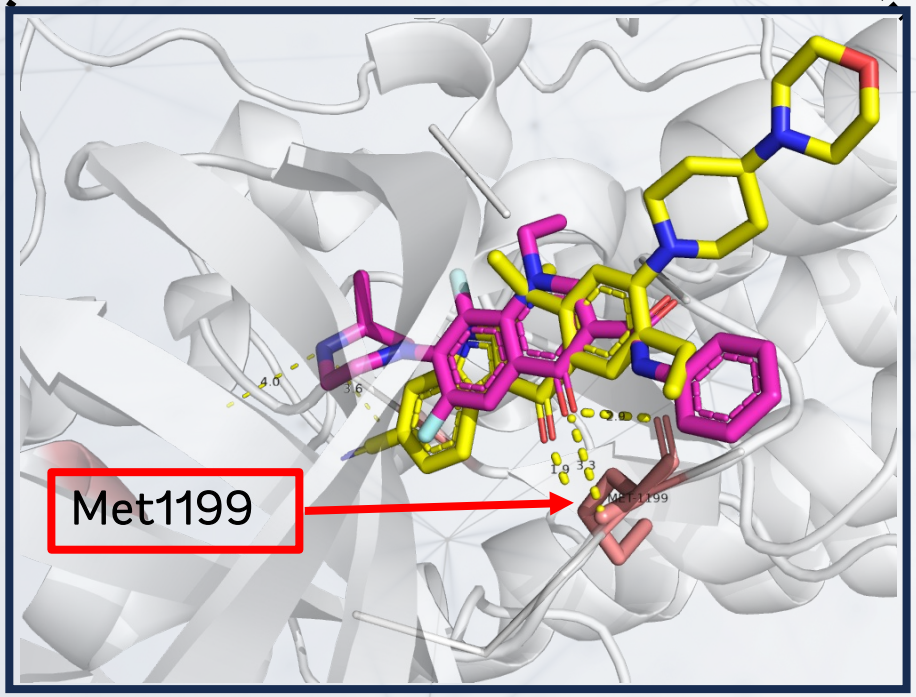


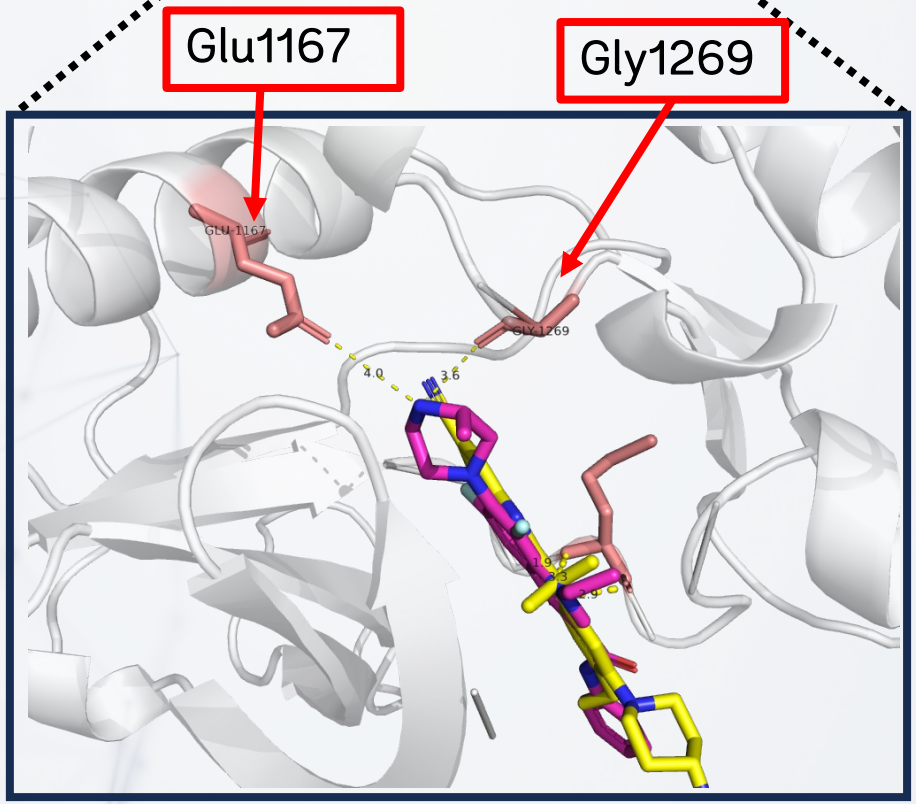


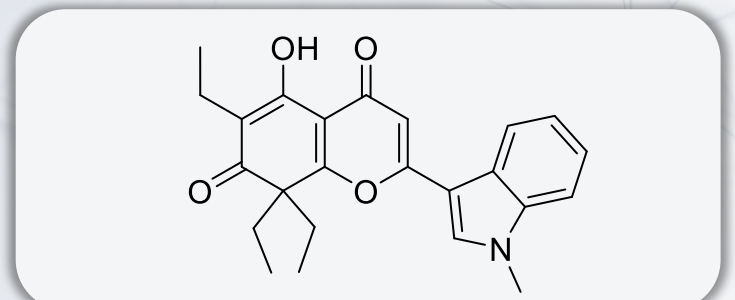

GLU-1167
GLY-1269

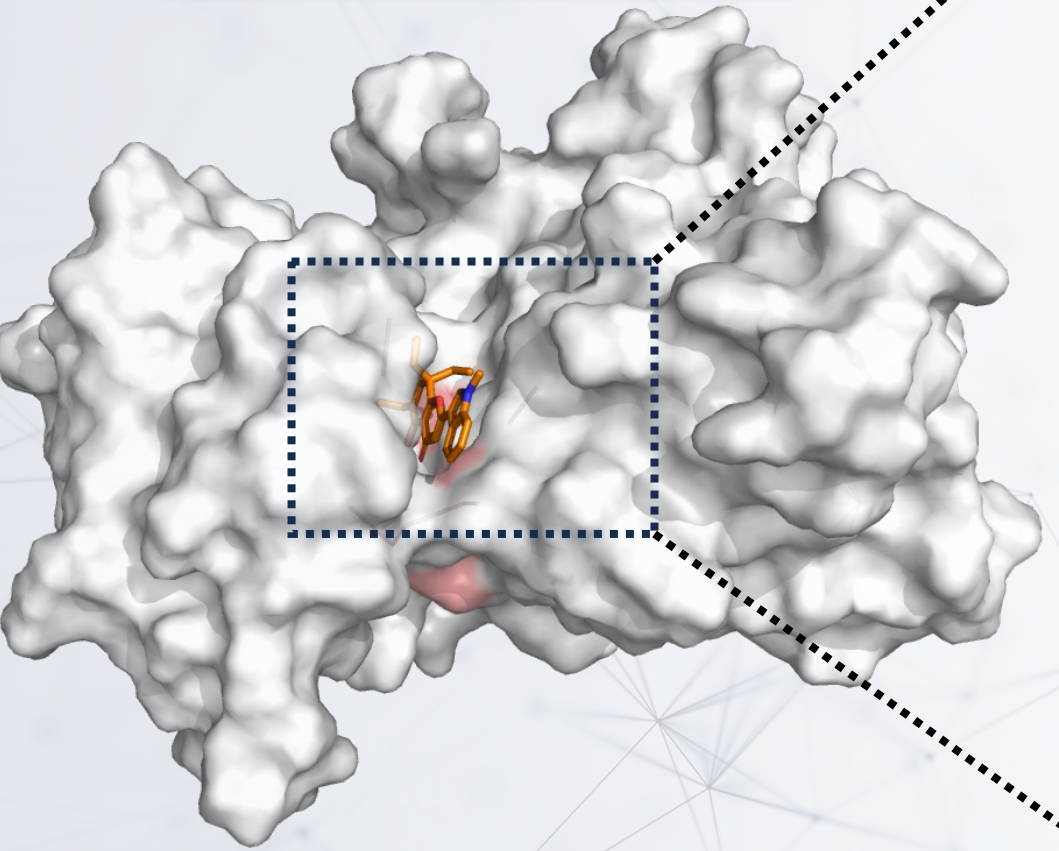

F
O
NH
N
NH
O
N
F

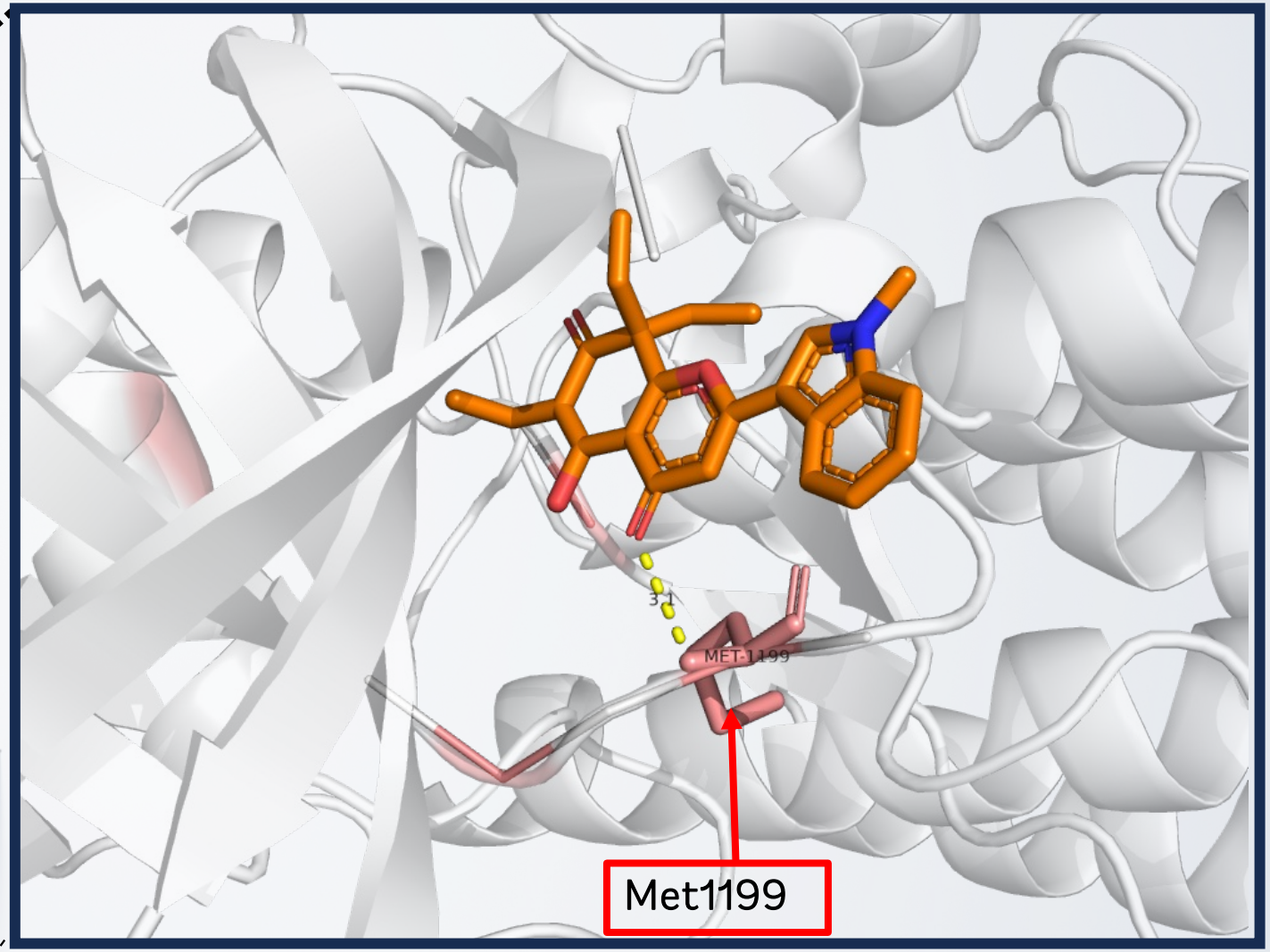



Met1199
4.0
3.6
1.9
3.3
MET-1199


Glu1167
Gly1269
GLU-1167
GLY-1269
4.0
3.6

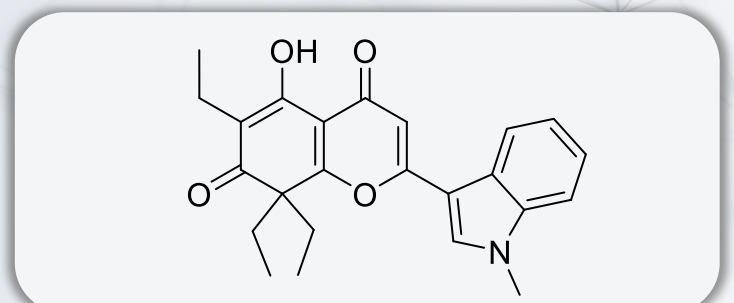


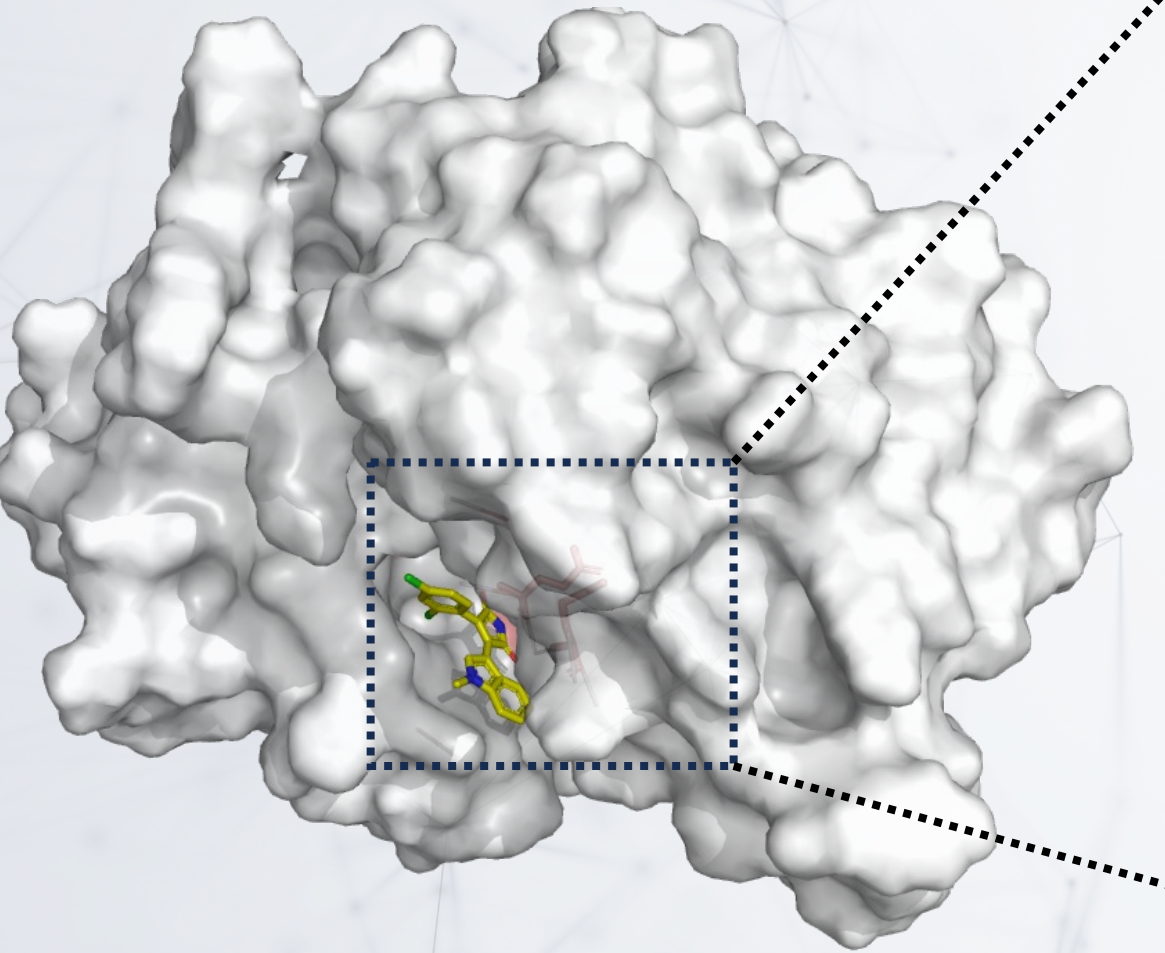


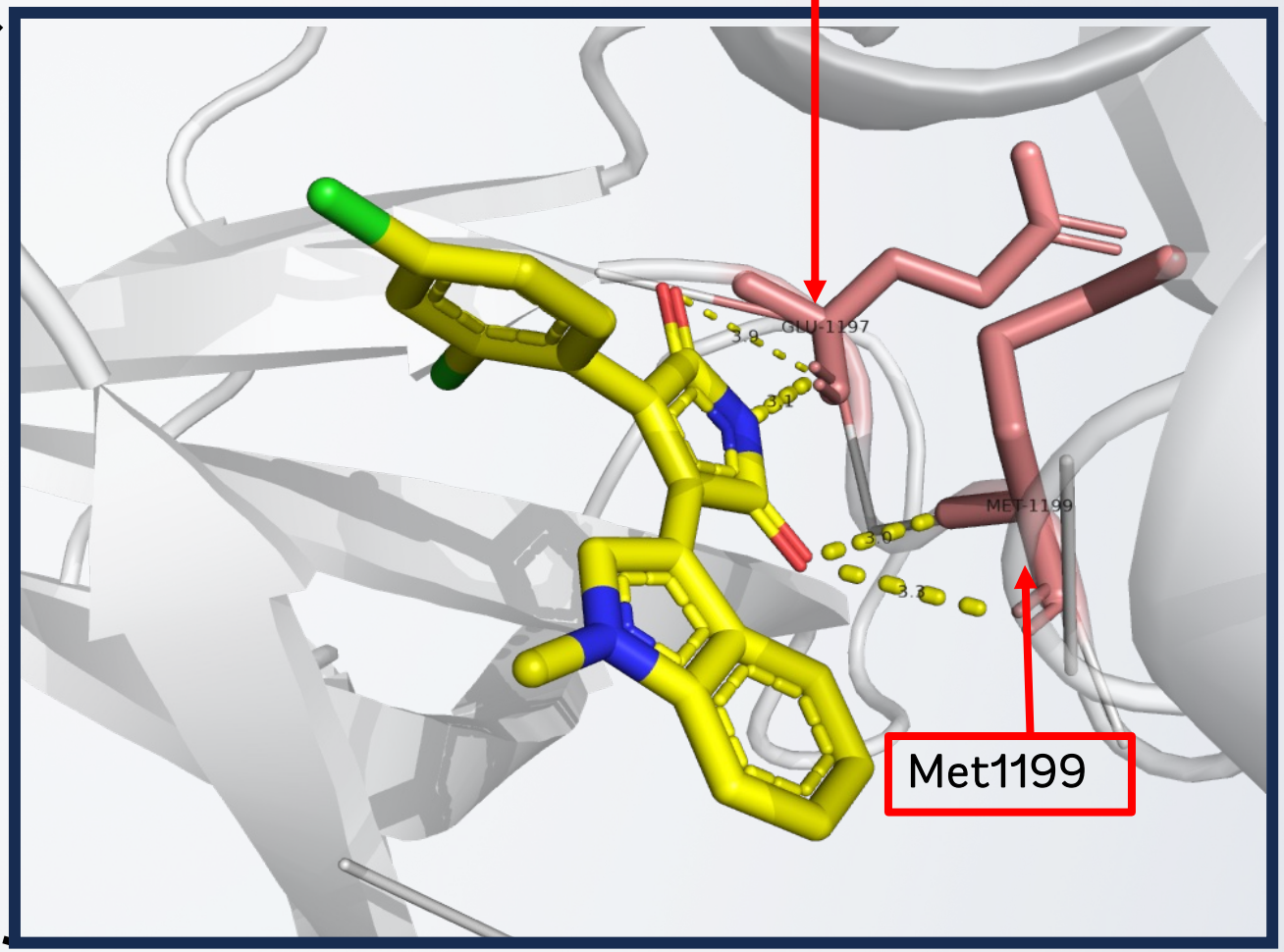


CHEMBL1689515

CHEMBL102714







SÀNG LỌC ẢO

Ứng dụng ALK Classification

3

Streamlit

HUGGING FACE

Mô hình

Sơ lược về ALK

Dự đoán

app

ALK

Project

ALK INHIBITORS CLASSIFICATION PROJECT

Main Menu

Dự đoán nhiều chất

Predict a batch

Predict one Molecule

About

Thông tin tác giả

Dự đoán một chất

1. Upload csv file

Please, upload your csv file

Drag and drop file here

Limit 200MB per file

Browse files

approved.csv 349.0B

Nút tải tệp *csv

Bảng ID và SMILE của tệp tải lên

Your file has 5 molecules

	ID	SMILE
0	alec	CCc1cc2c(cc1N1CCC(N3CCOCC3)CC1)C(C)(C)c1[nH]c3cc(C#N)ccc3c1C2=O
1	crizo	C[C@@H](Oc1cc(-c2cnn(C3CCNCC3)c2)cnc1N)c1c(Cl)ccc(F)c1Cl
2	ceri	Cc1cc(Nc2ncc(Cl)c(Nc3ccccc3S(=O)(=O)C(C)C)n2)c(OC(C)C)cc1C1CCNCC1
3	lorla	C[C@H]1Oc2cc(cnc2N)-c2c(nn(C)c2C#N)CN(C)C(=O)c2ccc(F)cc21
4	briga	COc1cc(N2CCC(N3CCN(C)CC3)CC2)ccc1Nc1ncc(Cl)c(Nc2ccccc2P(C)(C)=O)n1

app

ALK

Project

MedAI

1	crizo	C[C@@H](Oc1cc(-c2cnn(C3CCNCC3)c2)cnc1N)c1c(Cl)ccc(F)c1Cl
2	ceri	Cc1cc(Nc2ncc(Cl)c(Nc3ccccc3S(=O)(=O)C(C)C)n2)c(OC(C)C)cc1C1CCNCC1
3	lorla	C[C@H]1Oc2cc(cnc2N)-c2c(nn(C)c2C#N)CN(C)C(=O)c2ccc(F)cc21
4	briga	COc1cc(N2CCC(N3CCN(C)CC3)CC2)ccc1Nc1ncc(Cl)c(Nc2ccccc2P(C)(C)=O)n1

2. Preprocessing

Choose the number of criterions in Rule of 5 Lipinski

Number of chosen RO5 rules

○ 1 ○ 2 ◉ 3 ○ 4

Phễu Lipinski

Nút đặc trưng hóa

Processing

This is your featurized data frame

Bảng ID và SMILE

và dấu vân tay phân tử

	ID	Canomicalsmiles	0	1	2	3
0	alec	CCc1cc2c(cc1N1CCC(N3CCOCC3)CC1)C(C)(C)c1[nH]c3cc(C#N)ccc3c1C2=O	0	0	0	
1	crizo	C[C@@H](Oc1cc(-c2cnn(C3CCNCC3)c2)cnc1N)c1c(Cl)ccc(F)c1Cl	0	0	0	
2	ceri	Cc1cc(Nc2ncc(Cl)c(Nc3ccccc3S(=O)(=O)C(C)C)n2)c(OC(C)C)cc1C1CCNCC1	0	0	0	
3	lorla	C[C@H]1Oc2cc(cnc2N)-c2c(nn(C)c2C#N)CN(C)C(=O)c2ccc(F)cc21	0	0	0	
4	briga	COc1cc(N2CCC(N3CCN(C)CC3)CC2)ccc1Nc1ncc(Cl)c(Nc2ccccc2P(C)(C)=O)n1	0	0	0	

Your data have 2 molecules violated Rule of 5 Lipinski

Preprocessing done!

app

ALK

Project

MedAI

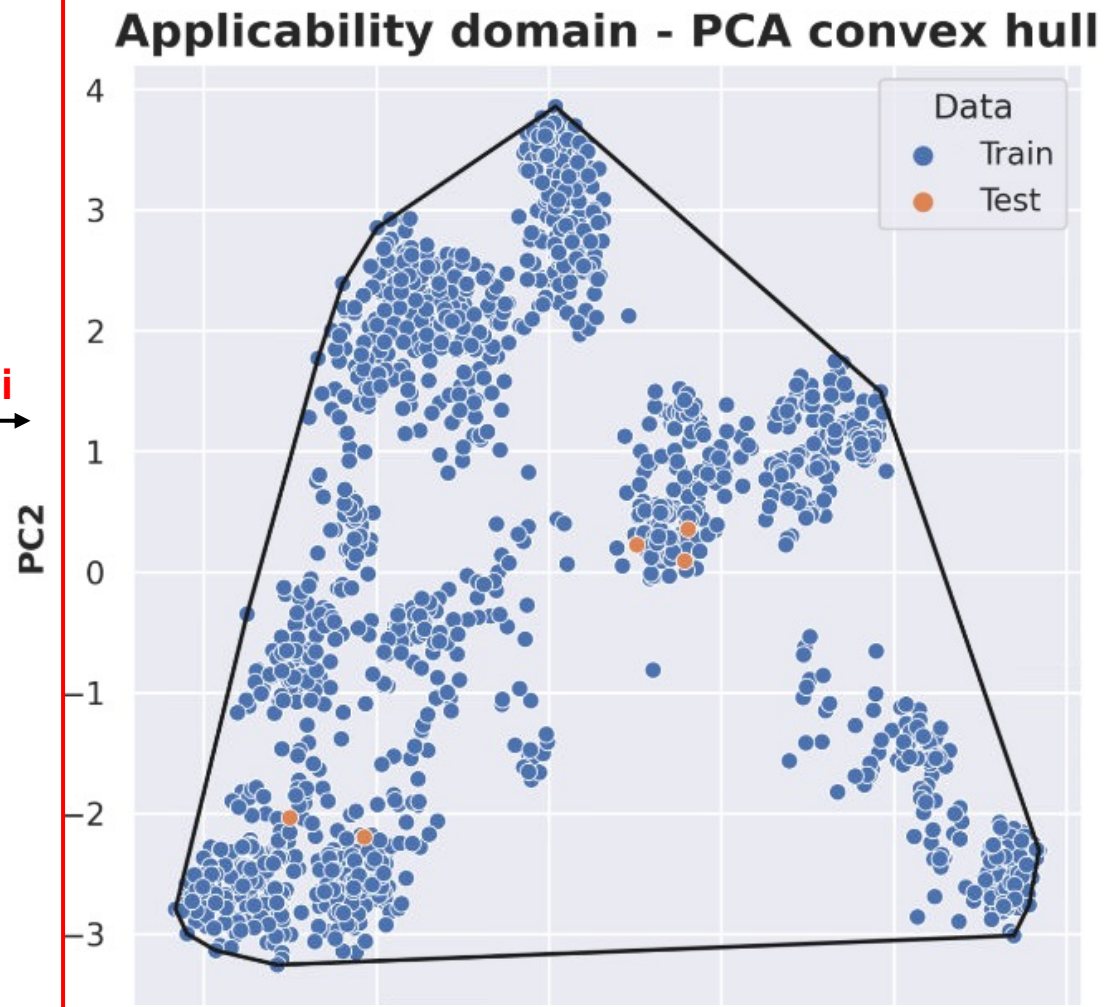
3. Application domain

Check application domain

Nút kiểm tra miền ứng dụng

All predicted molecules are in model's application domain

Miền ứng dụng PCA + bao lồi



ALK CLASSIFICATION

app

ALK

Project

MedAI

Processing

3. Application domain

Check application domain

4.Predict

Predict

Nút dự đoán

This is your results

	ID	SMILES	Predict_ANN	Proba_ANN
0	alec	CCc1cc2c(cc1N1CCC(N3CCOCC3)CC1)C(C)(C)c1[nH]c3cc(C#N)ccc3c1C2=O	1	75.81
1	crizo	C[C@@H](Oc1cc(-c2cnn(C3CCNCC3)c2)cnc1N)c1c(Cl)ccc(F)c1Cl	1	98.9
2	ceri	Cc1cc(Nc2ncc(Cl)c(Nc3ccccc3S(=O)(=O)C(C)C)n2)c(OC(C)C)cc1C1CCNCC1	0	24.64
3	lorla	C[C@H]1Oc2cc(cnc2N)-c2c(nn(C)c2C#N)CN(C)C(=O)c2ccc(F)cc21	0	49.43
4	briga	COc1cc(N2CCC(N3CCN(C)CC3)CC2)ccc1Nc1ncc(Cl)c(Nc2ccccc2P(C)(C)=O)n1	1	92.91

Bảng kết quả của 3 mô hình

Nút tải bảng kết quả

Download your result data frame as CSV

app

ALK

Project

MedAI

ALK INHIBITORS CLASSIFICATION PROJECT

Main Menu

Predict a batch | **Predict one Molecule** | About

1. Input smile

Please, input your smile

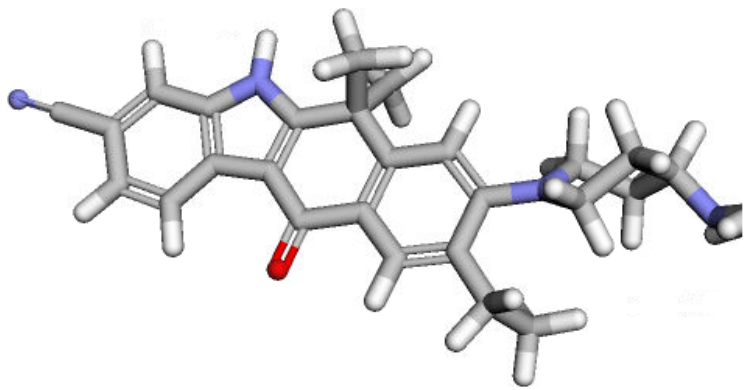
CCC1=CC2=C(C=C1N3CCC(CC3)N4CCOCC4)C(C5=C(C2=O)C6=C(N5)C=C(C=C6)C#N)(C)C

This is your 3D molecule

Dự đoán một chất

Nhập SMILE

Cấu trúc 3D của chất dự đoán

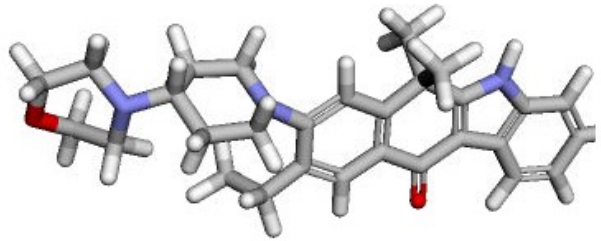


app

ALK

Project

MedAI



Choose the number of criterions in Rule of 5 Lipinski

Number of chosen RO5 rules

○ 1 ○ 2 ◉ 3 ○ 4

Phễu Lipinski

2. Predict

Predict

Nút dự đoán

This is your results

	ID	SMILES	Predict_ANN	Proba_
0	1	CCC1=CC2=C(C=C1N3CCC(CC3)N4CCOCC4)C(C5=C(C2=O)C6=C(N5)C=C(C=C6)C#N)(C)(	1	

Bảng kết quả

Download your result data frame as CSV

Nút tải bảng kết quả

https://huggingface.co/spaces/thechuongtrinh/ALK_inhibitors_classification

4. KẾT LUẬN VÀ ĐỀ NGHỊ

KẾT LUẬN

4

Đề tài đã xây dựng thành công **3 mô hình AI** để phân loại các chất ức chế thụ thể ALK, với cơ sở dữ liệu từ thư viện Reaxys.

Xây dựng thành công mô hình **concensus docking** với sự kết hợp từ 3 kết quả docking của 3 phần mềm Autodock-GPU, Vina-GPU-2.0 và GNINA.

Xác định **miền ứng dụng** của mô hình.

Triển khai mô hình thành ứng dụng sàng lọc ảo **ALK Classification** trên trang Huggingface.com

Quá trình sàng lọc ảo từ 120.571 chất có hoạt tính trên tế bào A549, đề tài thu được 3 ứng viên tiềm năng đó là: **CHEMBL1689515, CHEMBL2380351, CHEMBL102714**.

ĐỀ NGHỊ

4

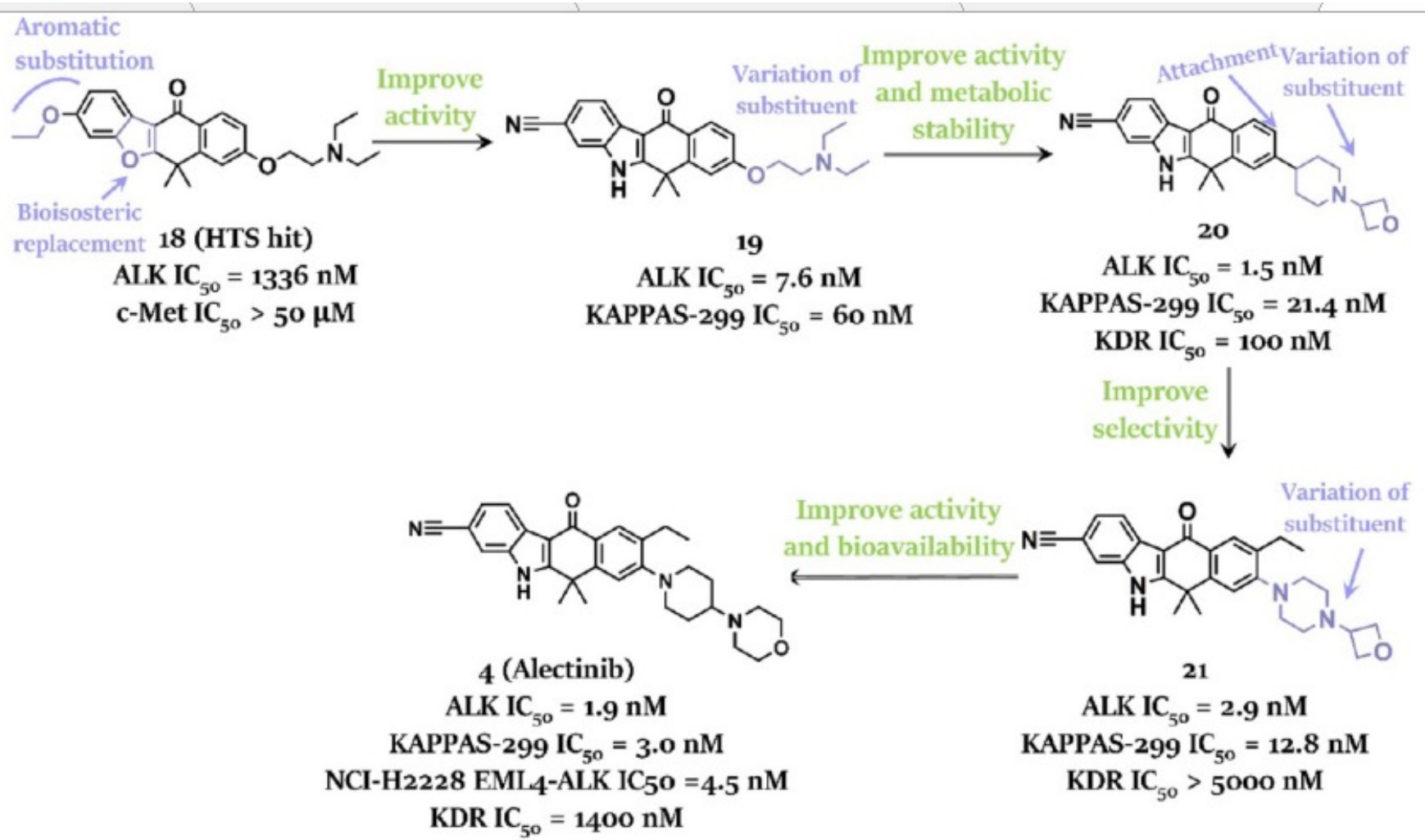
Thực hiện chạy động học phân tử (Molecular dynamics - MD) với 3 chất tiềm năng

Tổng hợp và thử nghiệm *in vitro* để xác định hoạt tính sinh học.

Cải thiện hiệu suất mô hình GNN bằng cách tìm thêm dữ liệu từ các nguồn khác, thiết kế cấu trúc mạng mới, áp dụng các mô-đun khác trong thư viện Pytorch Geometric. Sử dụng đồ thị protein kết hợp với đồ thị phân tử .Tiếp cận các mô hình AI khác như CNN, RNN để tối đa hóa tiềm năng của AI trong thiết kế thuốc

CẢM ƠN QUÝ
THẦY CÔ ĐÃ LẮNG NGHE

Email: ttchuong.d18@ump.edu.vn


Aromatic substitution
Bioisosteric replacement
18 (HTS hit)
ALK IC50 = 1336 nM
c-Met IC50 > 50 μM
Improve activity
Variation of substituent
19
ALK IC50 = 7.6 nM
KAPPAS-299 IC50 = 60 nM
Improve activity and metabolic stability
Attachment
Variation of substituent
20
ALK IC50 = 1.5 nM
KAPPAS-299 IC50 = 21.4 nM
KDR IC50 = 100 nM
Improve selectivity
Variation of substituent
21
ALK IC50 = 2.9 nM
KAPPAS-299 IC50 = 12.8 nM
KDR IC50 > 5000 nM
Improve activity and bioavailability
4 (Alectinib)
ALK IC50 = 1.9 nM
KAPPAS-299 IC50 = 3.0 nM
NCI-H2228 EML4-ALK IC50 =4.5 nM
KDR IC50 = 1400 nM

CÁC THUỐC ỨC CHẾ ALK

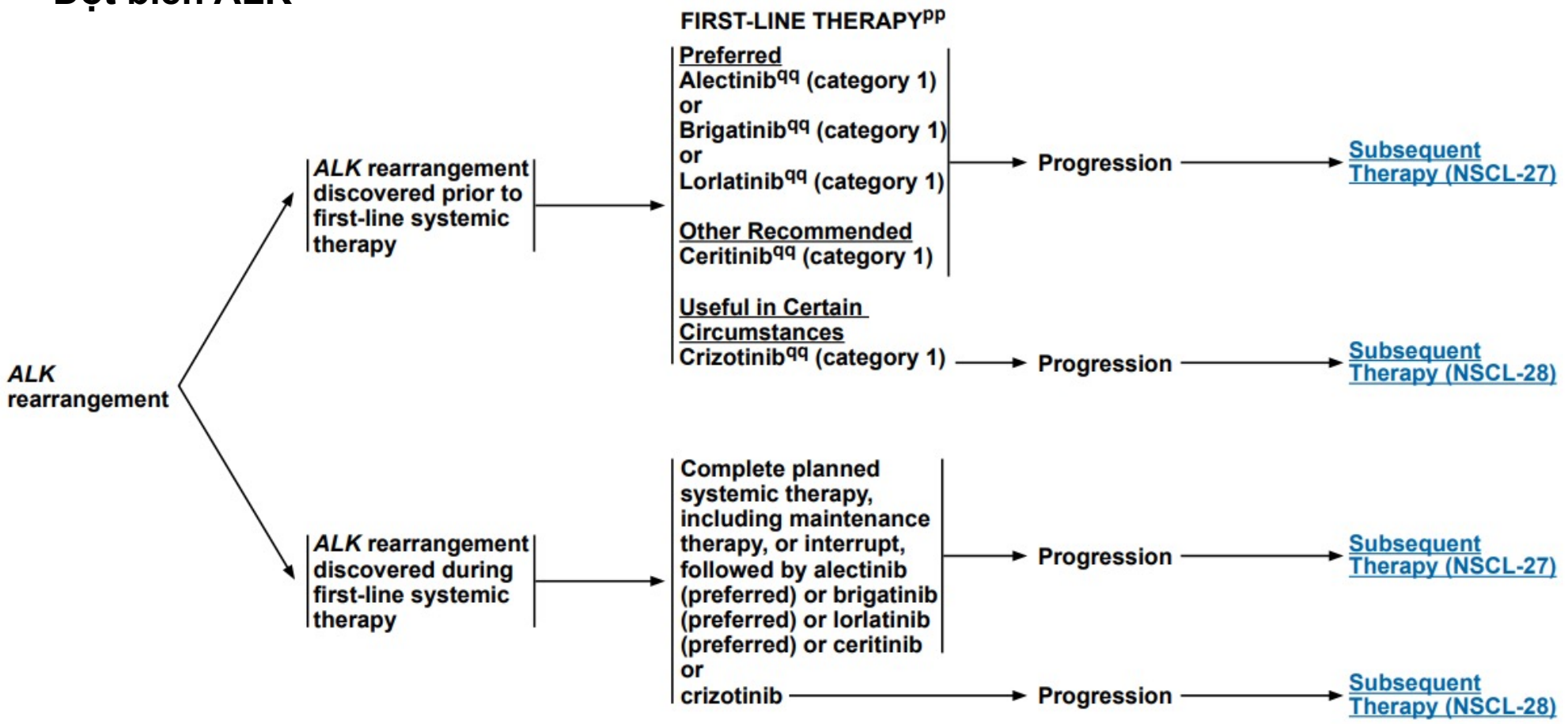
1

National Comprehensive Cancer Network®

HƯỚNG DẪN ĐIỀU TRỊ NCCN NĂM 2023
UNG THƯ PHỔI KHÔNG TẾ BÀO NHỎ

NCCN Guidelines Index
Table of Contents
Discussion

Đột biến ALK



CÁC THUỐC ỨC CHẾ ALK

	Alectinib	Brigatinib	Ceritinib	Crizotinib	Lorlatinib
C1156Y	**Likely beneficial** Relapsing: 1/82 (1%) Sensitivity: 5/5 (100%)	Possibly beneficial Relapsing: 0/32 (0%) Sensitivity: 1/1 (100%)	Possibly beneficial Relapsing: 3/53 (6%) Sensitivity: 2/3 (67%)	No benefit expected Relapsing: 22/220 (10%) No patients treated	Insufficient evidence Relapsing: 1/34 (3%) Sensitivity: 1/2 (50%)
I1171N	No benefit expected Relapsing: 17/82 (21%) No cases treated	No benefit expected Relapsing: 1/32 (3%) Sensitivity: 0/3 (0%)	**Likely beneficial** Relapsing: 0/53 (0%) Sensitivity: 6/6 (100%)	Insufficient evidence Relapsing: 0/220 (0%) No patients treated	Possibly beneficial Relapsing: 1/34 (3%) Sensitivity: 1/1 (100%)
I1171S	No benefit expected Relapsing: 6/82 (7%) No cases treated	Possibly beneficial Relapsing: 0/32 (0%) Sensitivity: 1/1 (100%)	Insufficient evidence Relapsing: 0/53 (0%) No patients treated	Insufficient evidence Relapsing: 1/220 (0.5%) No patients treated	Insufficient evidence Relapsing: 0/34 (0%) No patients treated
I1171T	Possibly beneficial Relapsing: 4/82 (5%) Sensitivity: 2/2 (100%)	No benefit expected Relapsing: 0/32 (0%) Sensitivity: 0/1 (0%)	**Likely beneficial** Relapsing: 0/53 (0%) Sensitivity: 3/3 (100%)	Insufficient evidence Relapsing: 10/220 (4%) No patients treated	Insufficient evidence Relapsing: 0/34 (0%) No patients treated
F1174C	Possibly beneficial Relapsing: 0/82 (0%) Sensitivity: 1/1 (100%)	Insufficient evidence Relapsing: 0/32 (0%) No patients treated	Possibly beneficial Relapsing: 3/53 (6%) Sensitivity: 1/1 (100%)	No benefit expected Relapsing: 2/220 (0.9%) Sensitivity: 0/1 (0%)	Possibly beneficial Relapsing: 0/34 (0%) Sensitivity: 1/1 (100%)
F1174L	Possibly beneficial Relapsing: 0/82 (0%) Sensitivity: 1/1 (100%)	Possibly beneficial Relapsing: 1/32 (3%) Sensitivity: 1/1 (100%)	No benefit expected Relapsing: 3/53 (6%) Sensitivity: 0/1 (0%)	Insufficient evidence Relapsing: 9/220 (4%) No patients treated	Insufficient evidence Relapsing: 0/34 (0%) No patients treated
F1174V	Possibly beneficial Relapsing: 0/82 (0%) Sensitivity: 1/1 (100%)	Insufficient evidence Relapsing: 1/32 (3%) No patients treated	Insufficient evidence Relapsing: 2/53 (4%) No patients treated	Insufficient evidence Relapsing: 4/220 (2%) No patients treated	Possibly beneficial Relapsing: 0/34 (0%) Sensitivity: 2/2 (100%)
L1196M	Conflicting evidence Relapsing: 9/82 (11%) Sensitivity: 9/11 (82%)	No benefit expected Relapsing: 2/32 (6%) Sensitivity: 1/2 (50%)	Possibly beneficial Relapsing: 3/53 (6%) Sensitivity: 10/10 (100%)	No benefit expected Relapsing: 55/220 (25%) No cases treated	**Likely beneficial** Relapsing: 1/34 (3%) Sensitivity: 7/10 (70%)
G1202R	Conflicting evidence Relapsing: 26/82 (32%) Sensitivity: 4/6 (67%)	No benefit expected Relapsing: 10/32 (31%) Sensitivity: 2/4 (50%)	No benefit expected Relapsing: 19/53 (36%) Sensitivity: 1/2 (50%)	No benefit expected Relapsing: 16/220 (7%) No cases treated	**Likely beneficial** Relapsing: 1/34 (3%) Sensitivity: 17/22 (77%)
G1269A	**Likely beneficial** Relapsing: 0/82 (0%) Sensitivity: 2/3 (67%)	Insufficient evidence Relapsing: 0/32 (0%) Sensitivity: 1/2 (50%)	**Likely beneficial** Relapsing: 0/53 (0%) Sensitivity: 2/3 (67%)	No benefit expected Relapsing: 34/220 (16%) No cases treated	**Likely beneficial** Relapsing: 1/34 (3%) Sensitivity: 5/5 (100%)
L1196M G1202R	Insufficient evidence Relapsing: 2/82 (2%) No patients treated	Insufficient evidence Relapsing: 1/32 (3%) No patients treated	Insufficient evidence Relapsing: 0/53 (0%) No patients treated	Insufficient evidence Relapsing: 0/220 (0%) No patients treated	No benefit expected Relapsing: 4/34 (12%) Sensitivity: 0/1 (0%)

Clinical benefit

% of patients with only this mutation achieving clinical benefit or sensitivity on sequential treatment with this inhibitor

Relapsing

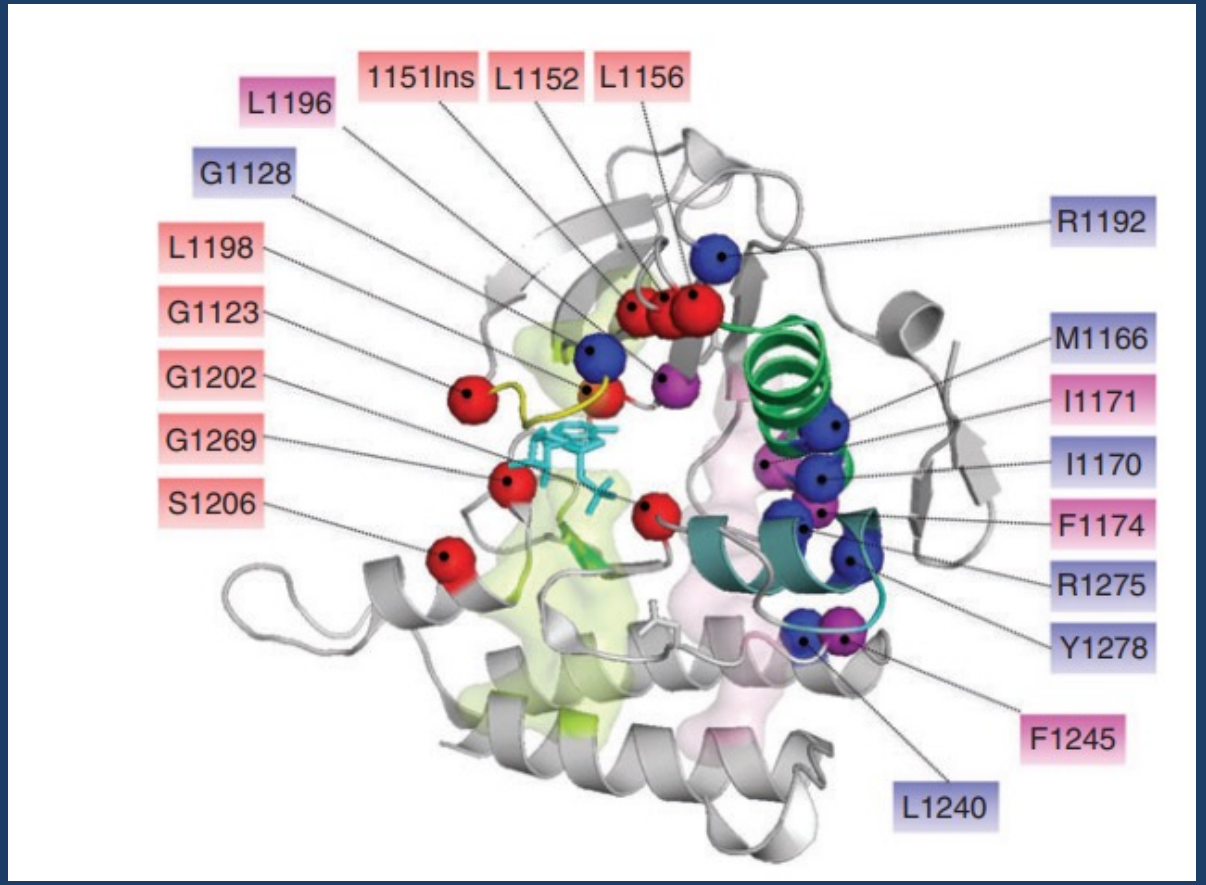
% of patients relapsing on this inhibitor harboring only this mutation

Three or more patients with treatment results

	>66%	33-66%	<33%
<5%	Likely beneficial	Possibly beneficial	No benefit expected
5-9%	Possibly beneficial	Conflicting evidence	No benefit expected
≥10%	Conflicting evidence	No benefit expected	No benefit expected

Two or less patients with treatment results

	100%	50%	0%	No patients treated
<5%	Possibly benefidial	Insufficient evidence	No benefit expected	Insufficient evidence
5-9%	Possibly beneficial	No benefit expected	No benefit expected	No benefit expected
≥10%	No benefit expected	No benefit expected	No benefit expected	No benefit expected



QUY TRÌNH THU THẬP VÀ XỬ LÝ DỮ LIỆU

2

BƯỚC 1

Thu thập dữ liệu

Cơ sở dữ liệu Reaxys

BƯỚC 2

Xử lý dữ liệu

Xóa dữ liệu không hoàn thiện.

Chuẩn hóa đích tác động, đơn vị đo.

Chuẩn hóa pIC50.

BƯỚC 3

Chuẩn hóa cấu trúc

Loại bỏ SMILES lỗi.

Chuẩn hóa dạng muối

BƯỚC 4

Luật 5 Lipinski

Loại bỏ các cấu trúc vi phạm hơn 1 luật

QUY TRÌNH KHAI PHÁ DỮ LIỆU

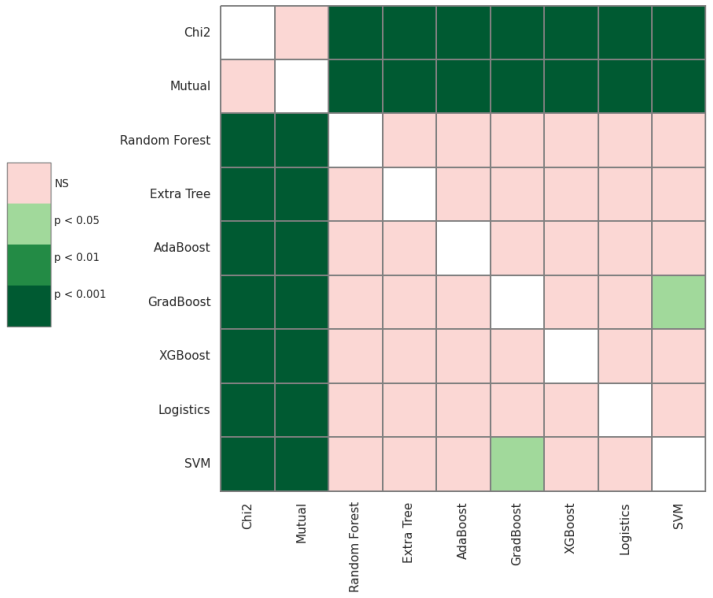
2

Chia dữ liệu	Tập huấn luyện chiếm 80% và tập đánh giá ngoại chiếm 20%
Làm sạch dữ liệu	Xóa dữ liệu bị trùng, xóa các cột có phương sai thấp hơn 0,05
Xử lý dữ liệu bị mất	Sử dụng thuật toán kNN imputer để điền vào các ô dữ liệu bị mất
Chuẩn hóa	Min_max scaler để chuẩn hóa các giá trị trong cùng một cột
Lựa chọn đặc trưng	Sử dụng các thuật toán học máy để chọn ra các đặc trưng quan trọng
Lựa chọn thuật toán	Sau khi dữ liệu đã được xử lý, lần lượt đưa dữ liệu vào các thuật toán học máy và so sánh hiệu quả của từng thuật toán

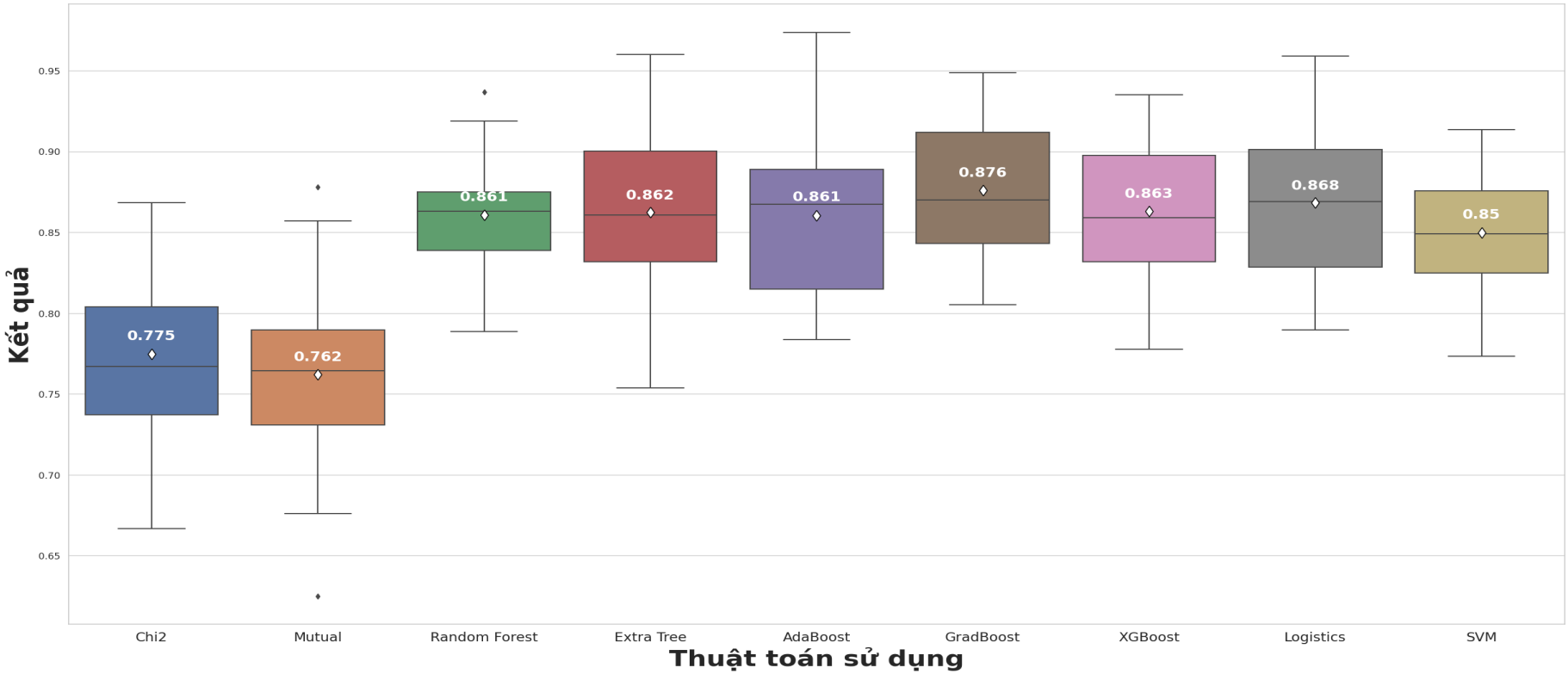
MÔ HÌNH HỌC MÁY

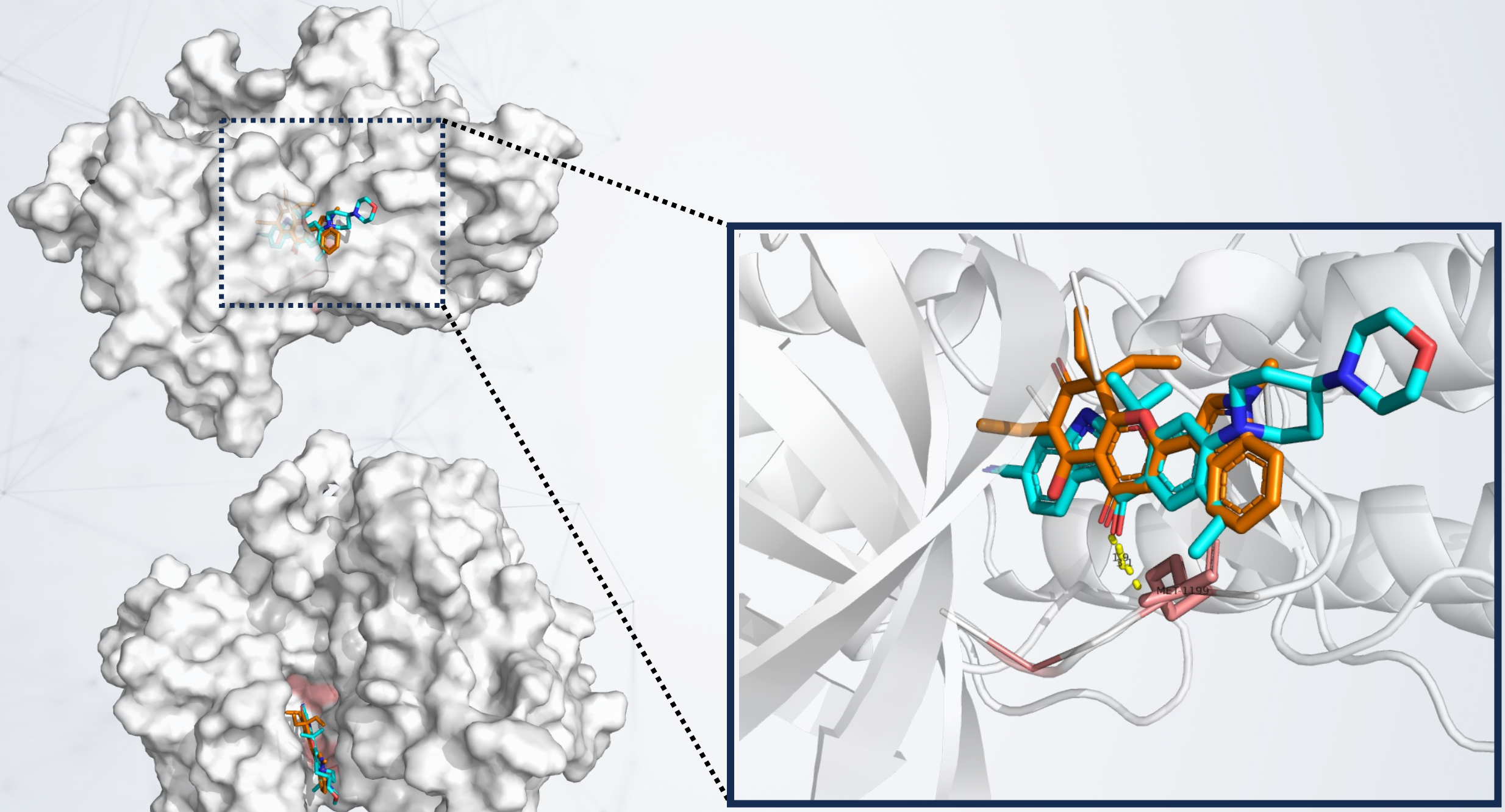
So sánh thuật toán chọn đặc trưng

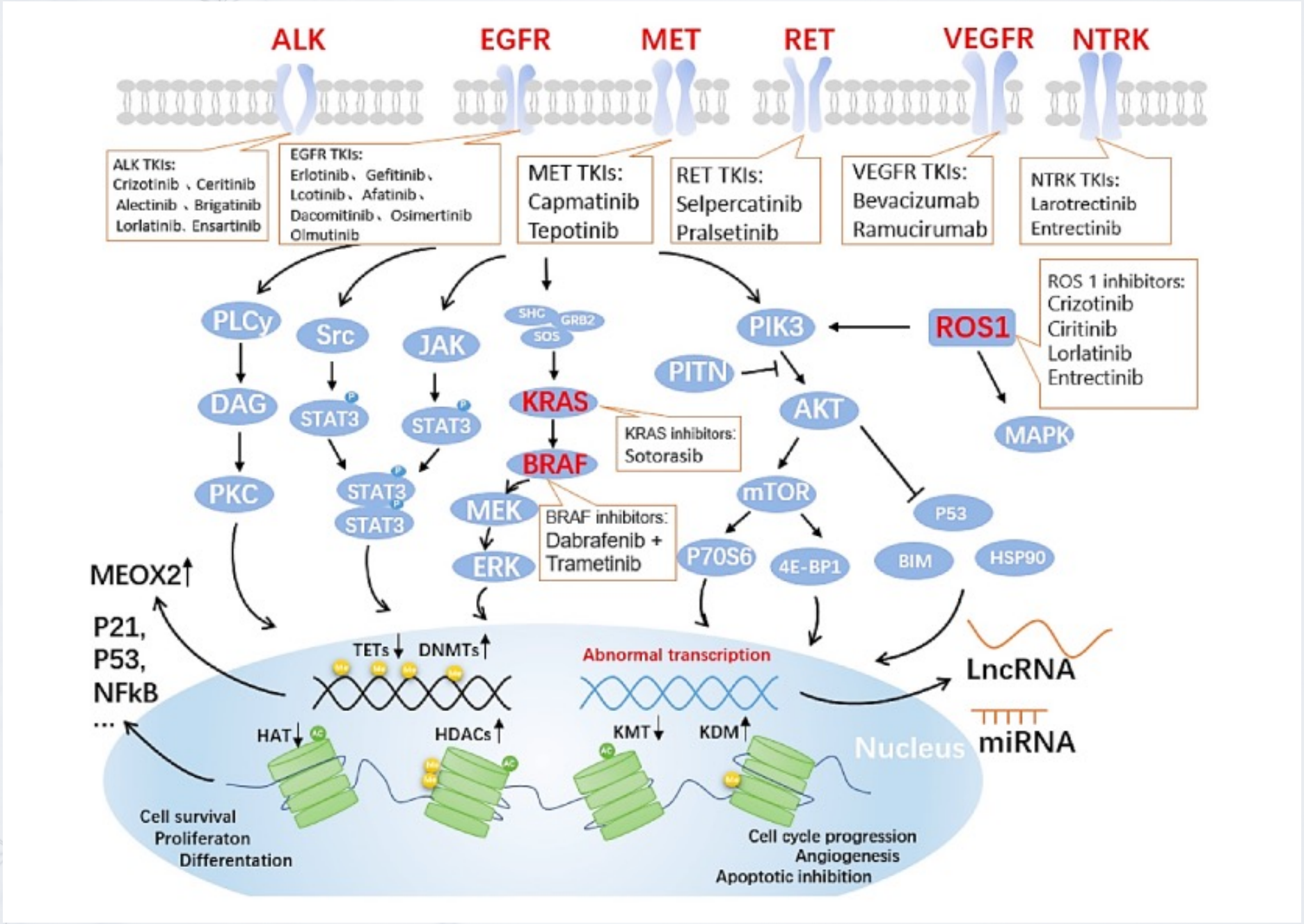
Biểu đồ nhiệt Wilcoxon

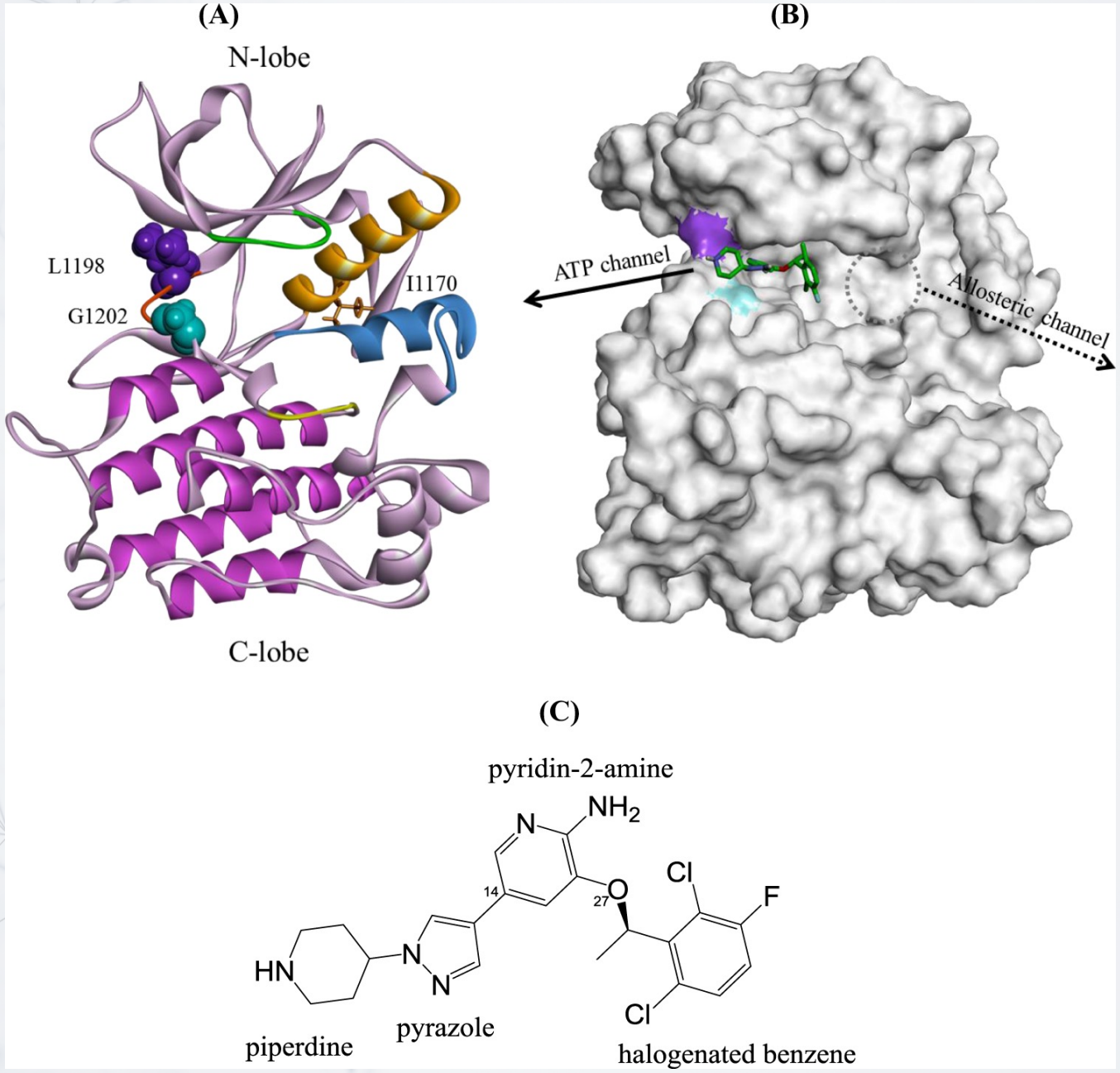


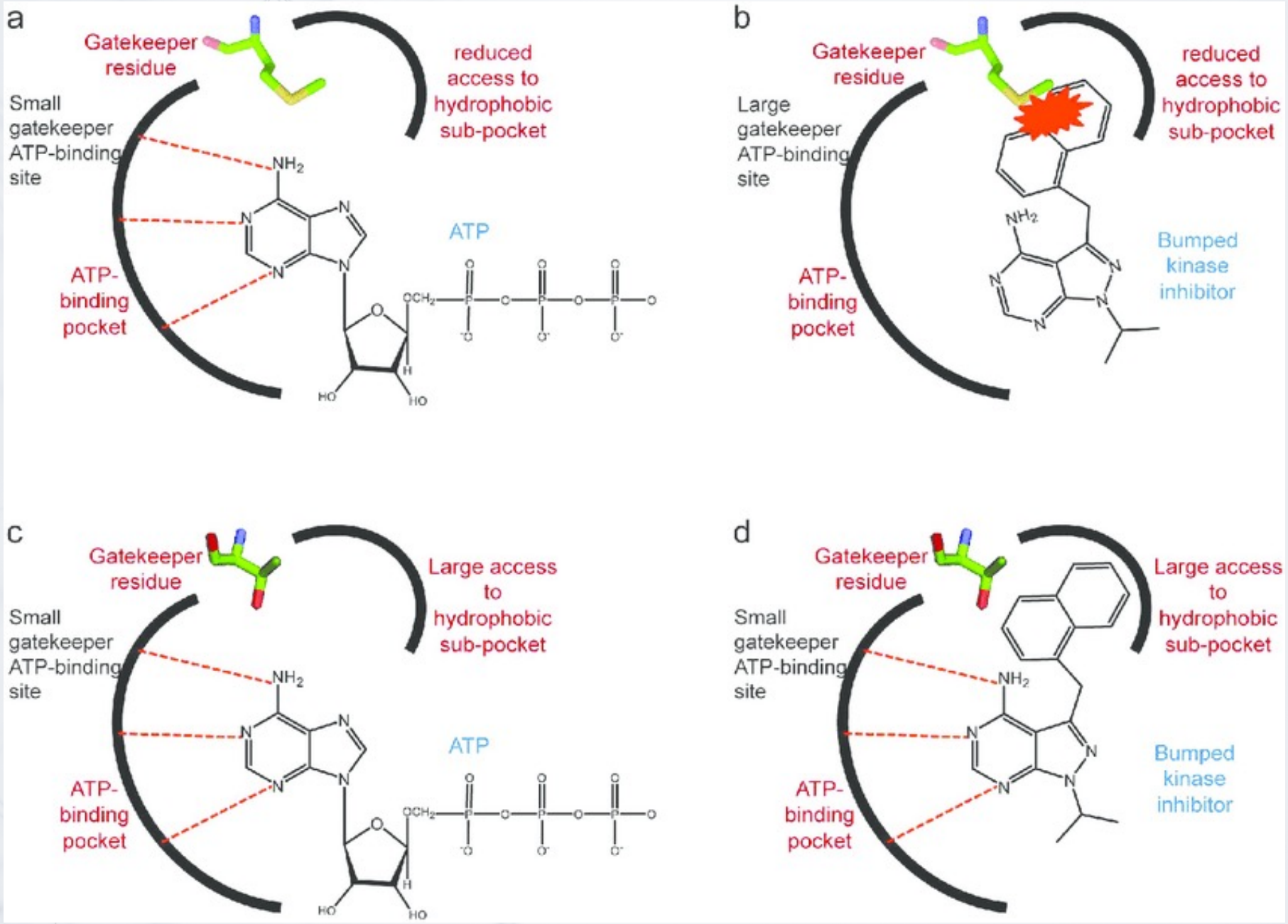
So sánh kết quả CV-f1 score giữa các thuật toán chọn đặc trưng

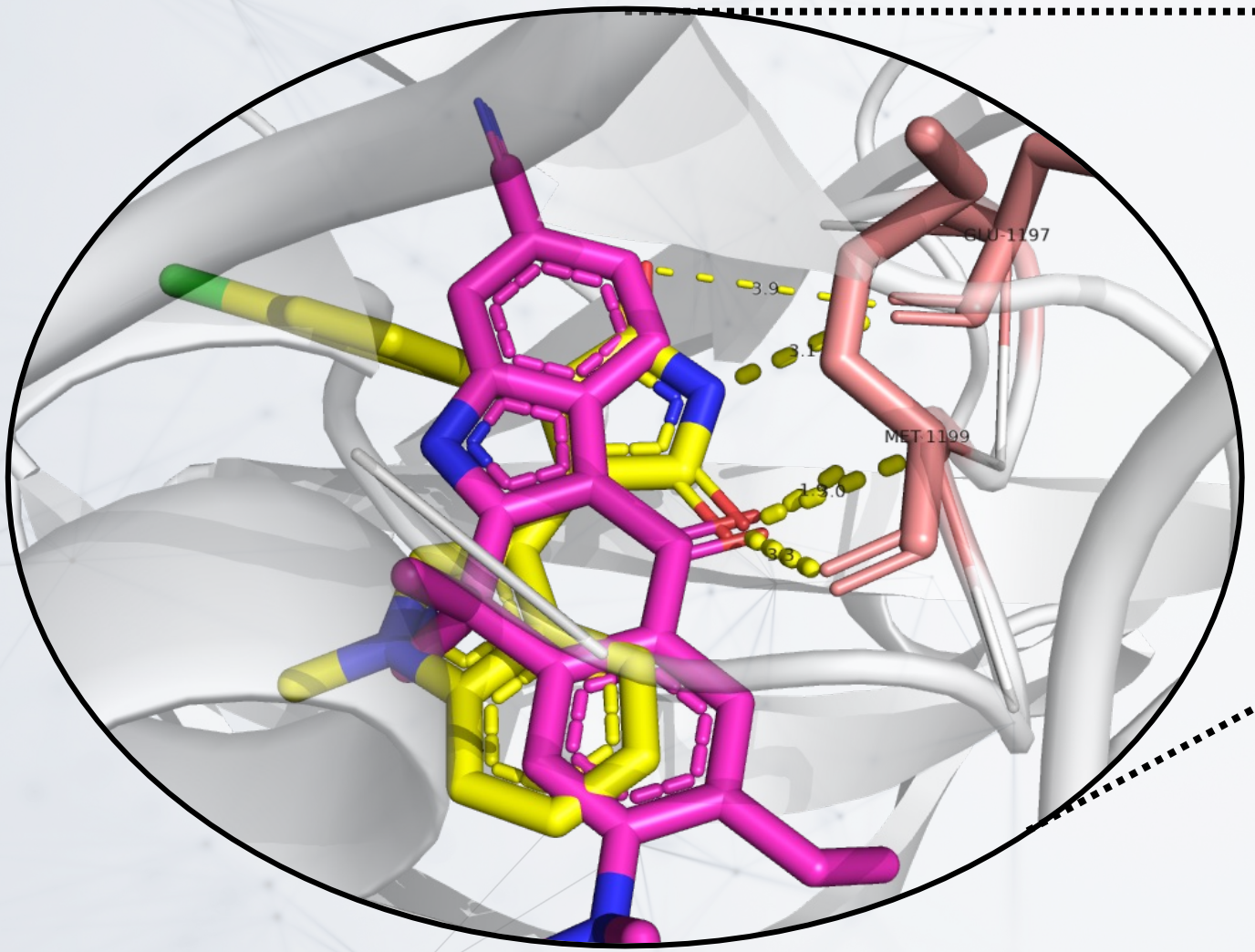


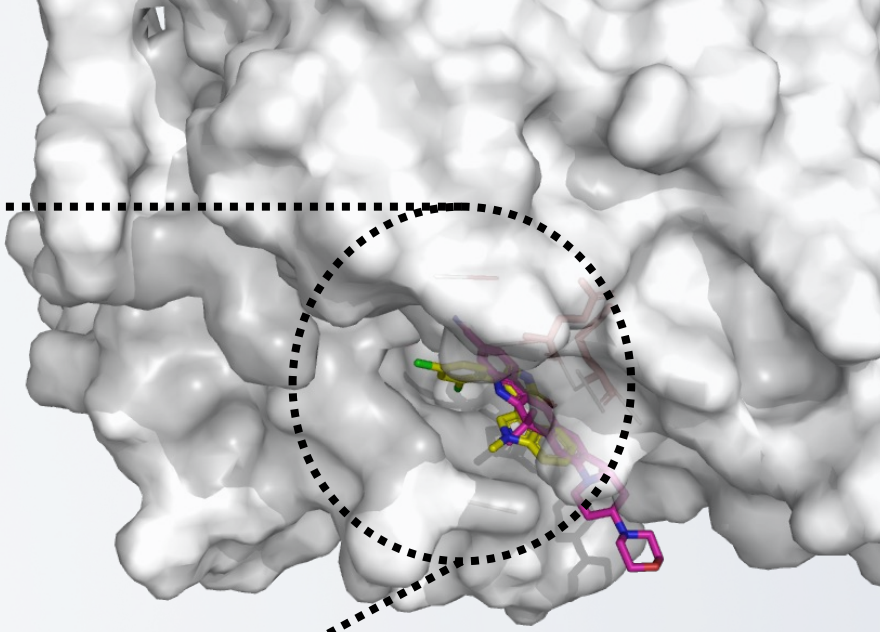

MET 1199

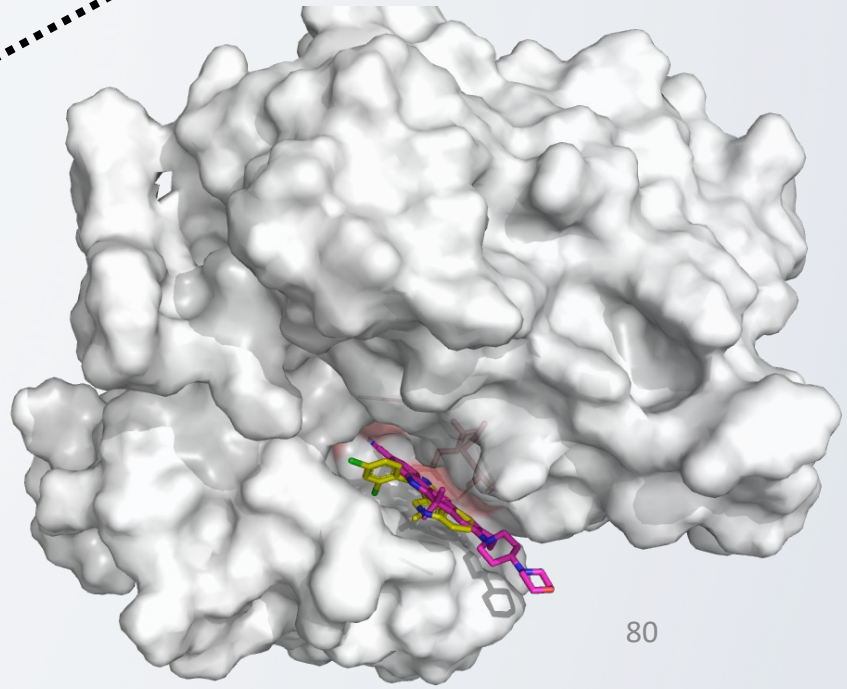

ALK
EGFR
MET
RET
VEGFR
NTRK
ALK TKIs:
Crizotinib 、Ceritinib
Alectinib 、Brigatinib
Lorlatinib、Ensartinib
EGFR TKIs:
Erlotinib、Gefitinib、
Lcotinib、Afatinib、
Dacomitinib、Osimertinib
Olmutinib
MET TKIs:
Capmatinib
Tepotinib
RET TKIs:
Selpercatinib
Pralsetinib
VEGFR TKIs:
Bevacizumab
Ramucirumab
NTRK TKIs:
Larotrectinib
Entrectinib
PLCy
Src
JAK
SHC
GRB2
SOS
PIK3
ROS1
ROS 1 inhibitors:
Crizotinib
Ciritinib
Lorlatinib
Entrectinib
PITN
DAG
STAT3
STAT3
KRAS
AKT
MAPK
KRAS inhibitors:
Sotorasib
BRAF
PKC
STAT3
STAT3
MEK
mTOR
P53
BRAF inhibitors:
Dabrafenib +
Trametinib
ERK
P70S6
4E-BP1
BIM
HSP90
MEOX2↑
P21,
P53,
NFkB
…
TETs↓ DNMTs↑
Abnormal transcription
LncRNA
miRNA
HAT↓
HDACs↑
KMT↓
KDM↑
Nucleus
Cell survival
Proliferaton
Differentation
Cell cycle progression
Angiogenesis
Apoptotic inhibition


(A)
N-lobe
L1198
G1202
I1170
C-lobe
(B)
ATP channel
Allosteric channel
(C)
pyridin-2-amine
NH2
N
14
O
27
Cl
F
Cl
HN
N
N
piperdine
pyrazole
halogenated benzene


a
Gatekeeper residue
reduced access to hydrophobic sub-pocket
Small gatekeeper ATP-binding site
ATP
ATP-binding pocket
b
Gatekeeper residue
reduced access to hydrophobic sub-pocket
Large gatekeeper ATP-binding site
Bumped kinase inhibitor
ATP-binding pocket
c
Gatekeeper residue
Large access to hydrophobic sub-pocket
Small gatekeeper ATP-binding site
ATP
ATP-binding pocket
d
Gatekeeper residue
Large access to hydrophobic sub-pocket
Small gatekeeper ATP-binding site
Bumped kinase inhibitor
ATP-binding pocket


GLU-1197
3.9
MET-1199

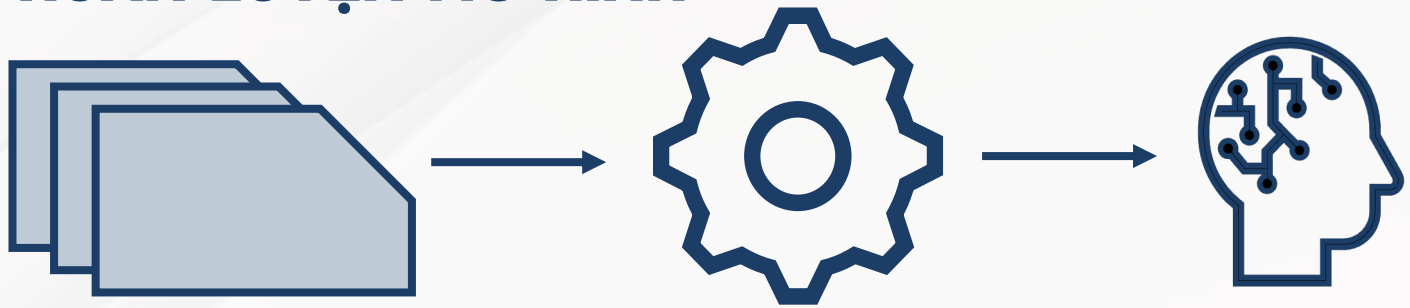




MÔ HÌNH HỌC MÁY

1

HUẤN LUYỆN MÔ HÌNH

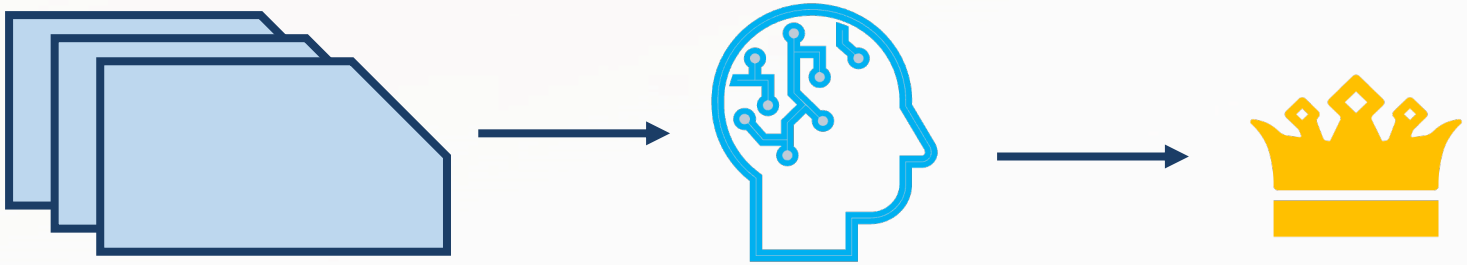


Dữ liệu huấn luyện

Thuật toán học máy

Mô hình dự đoán

ĐÁNH GIÁ MÔ HÌNH

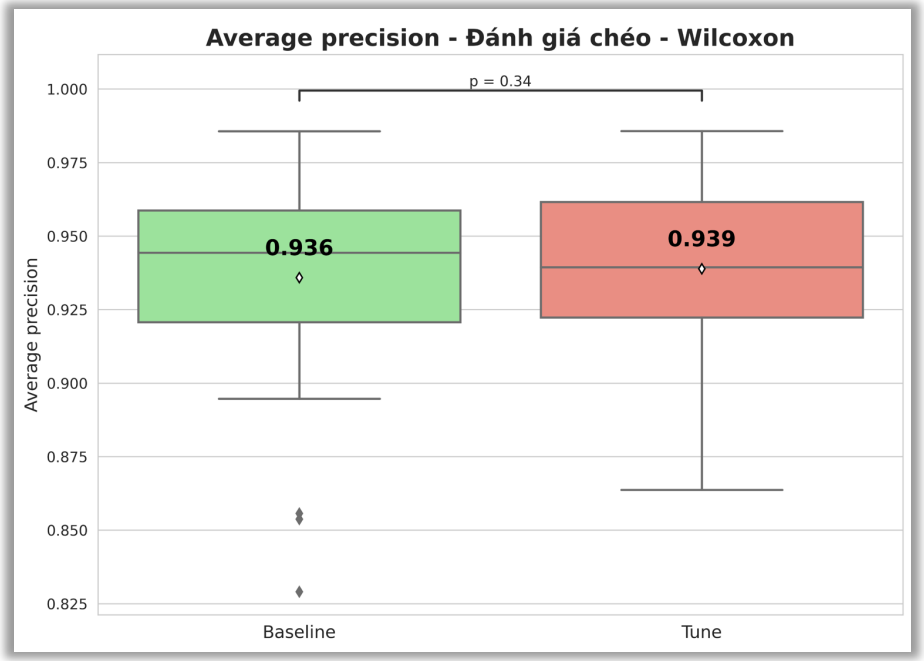


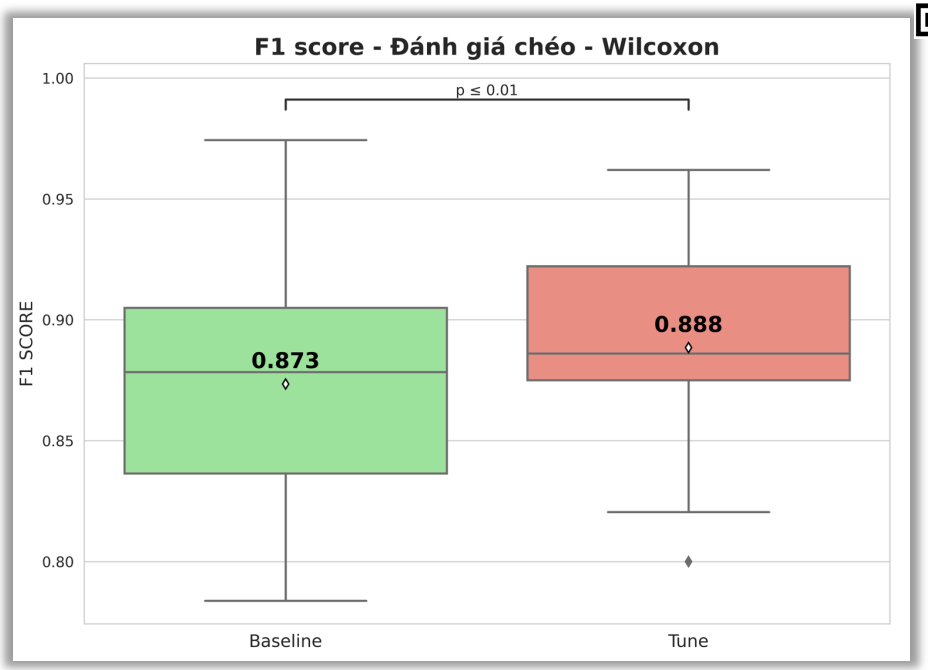
Dữ liệu dự đoán

Mô hình dự đoán

Kết quả

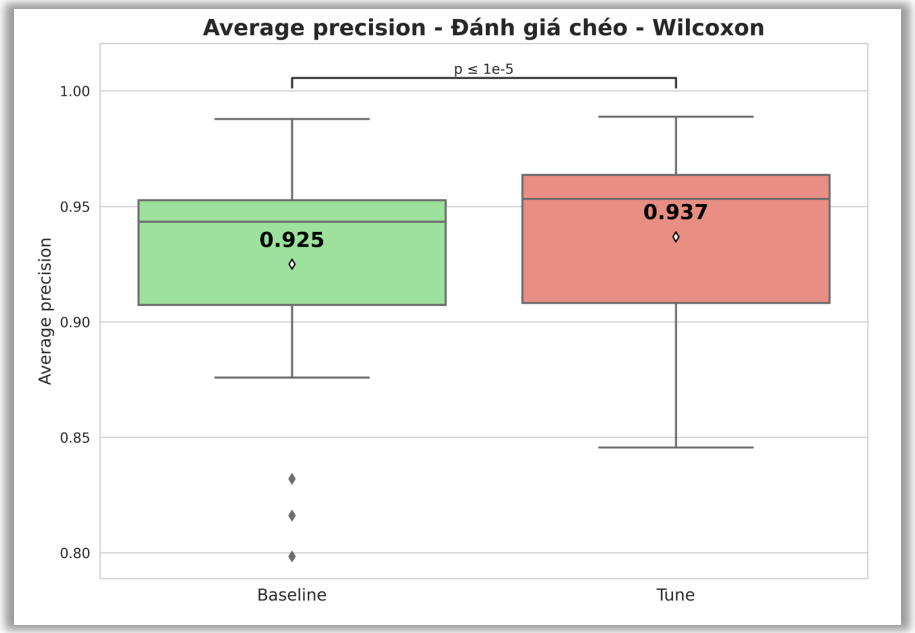
MÔ HÌNH HỌC MÁY

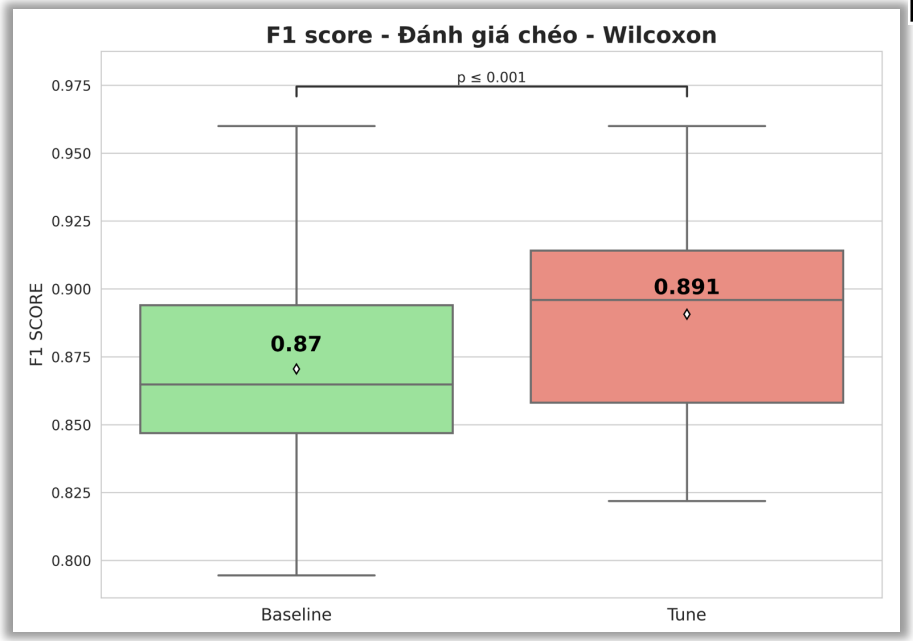




ĐIỂM	ĐÁNH GIÁ CHÉO (CV)			ĐÁNH GIÁ NGOẠI (EV)	
	Trước tối ưu	Sau tối ưu	p-value	Trước tối ưu	Sau tối ưu
f1 score	0,873±0,045	0,888±0,039	0,0062	0,894	0,921
AP	0,936±0,037	0,939±0,032	0,339	0,921	0,961

MẠNG THẦN KINH NHÂN TẠO (ANN)





ĐIỂM	ĐÁNH GIÁ CHÉO (CV)			ĐÁNH GIÁ NGOẠI (EV)	
	Trước tối ưu	Sau tối ưu	p-value	Trước tối ưu	Sau tối ưu
f1 score	0,870±0,038	0,891±0,037	$1,37.10^{-4}$	0,940	0,930
AP	0,925±0,046	0,937±0,040	$4,42.10^{-6}$	0,962	0,955

Thông số huấn luyện mô hình

SIÊU THAM SỐ	KHOẢNG TÌM KIẾM	SIÊU THAM SỐ TỐI ƯU	SIÊU THAM SỐ MÔ HÌNH CƠ SỞ
n-layers	int(2;6)	4	2
dense-neurons	int(128;2048)	1913	427
dropout-rate	float(0,1; 0,8)	0,782035574792785	0,7063233020424546
epochs	int(50;100)	99	52

Tham số huấn luyện:

- Hàm mất mát: Binary Cross Entroy
- Hàm tối ưu: Adam, tốc độ học = 0,0001, weight_decay = 0,01.

MẠNG THẦN KINH ĐỒ THỊ (GNN)

Thông số huấn luyện mô hình

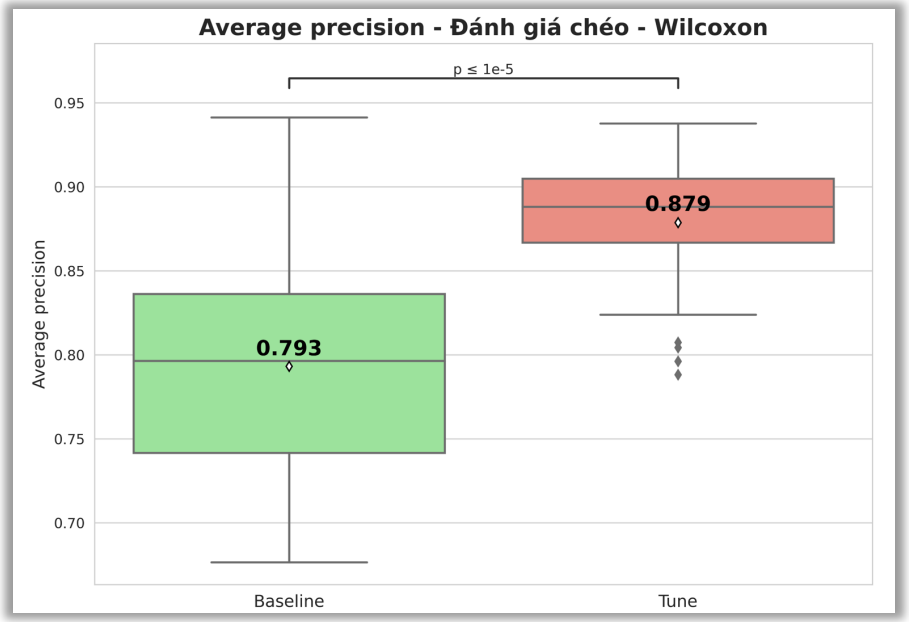
Siêu tham số	Khoảng tìm kiếm	Siêu tham số tối ưu	Siêu tham số của mô hình cơ sở
batch-size	categoriacal(128; 64)	64	128
n-blocks	categorical(3; 4; 5)	3	3
embedding-size	int(150; 300)	215	195
drop-out-rate	uniform(0,1; 0,5)	0,3767639976718497	0,3
top-k-ratio	uniform(0,4; 0,9)	0,5805249184752477	0,5
number-hidden-node	int(150; 256)	198	181
ann-dropout-rate	uniform(0,2; 0,9)	0,35364435414672357	0,5
sdg-momentum	uniform(0,7; 0,9)	0,765050961383454 1	0,8
scheduler-gamma	uniform(0,9; 1,0)	0,9844021689216244	0,99

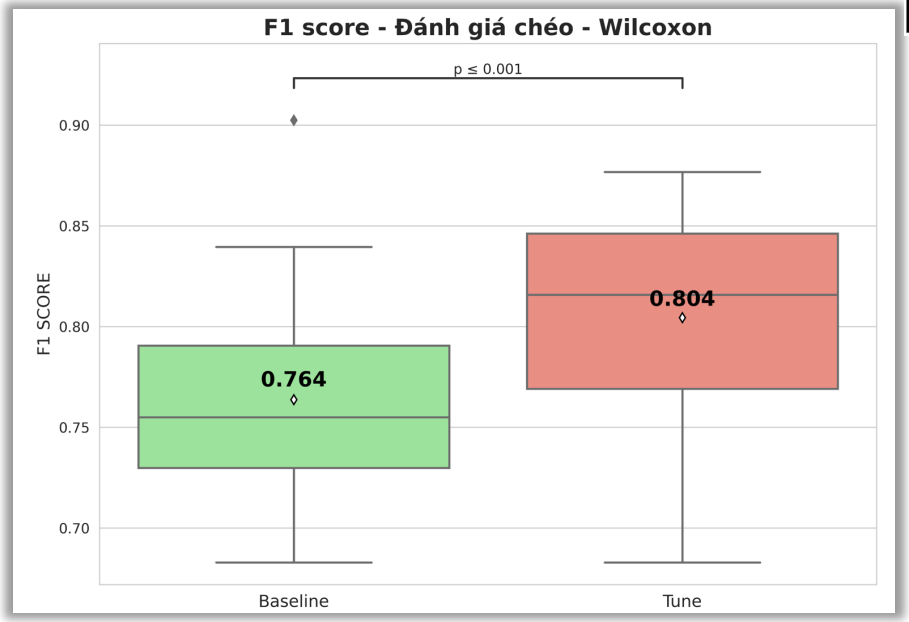
Hàm mất mát: Binary cross entropy,

Hàm tối ưu: SGD, tốc độ học 0,01 và weight_decay 0,0001, momentum

ExponentialLR một cơ chế điều chỉnh tốc độ học theo lịch trình, Số chu kỳ học: 300

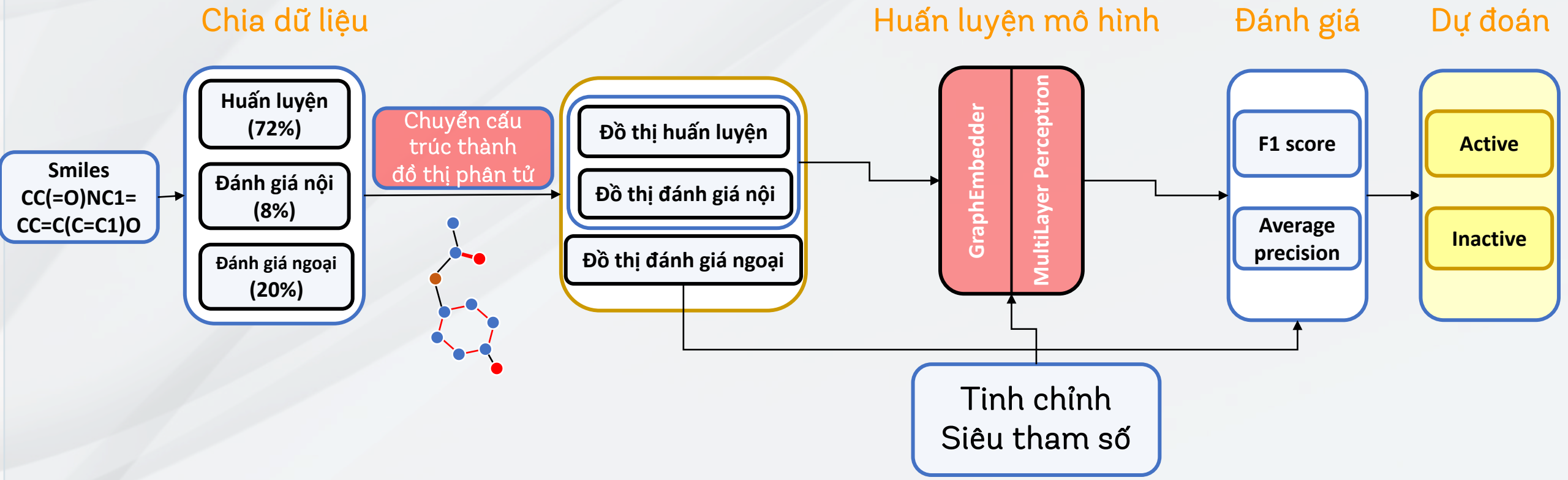
MẠNG THẦN KINH ĐỒ THỊ (GNN)

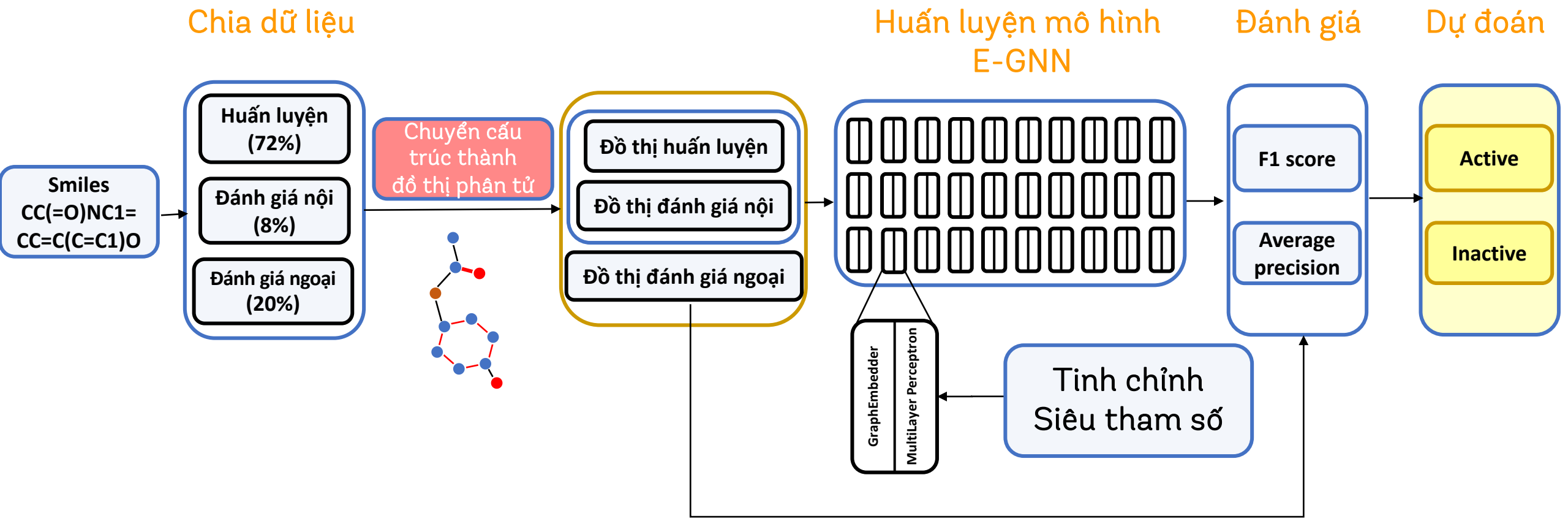




ĐIỂM	ĐÁNH GIÁ CHÉO (CV)			ĐÁNH GIÁ NGOẠI (EV)	
	Trước tối ưu	Sau tối ưu	p-value	Trước tối ưu	Sau tối ưu
f1 score	0,729±0,045	0,804±0,049	$5,75.10^{-6}$	0,794	0,863
AP	0,769±0,069	0,879±0,041	$1,92.10^{-6}$	0,823	0,938

MẠNG THẦN KINH ĐỒ THỊ (GNN)




Chia dữ liệu
Huấn luyện mô hình E-GNN
Đánh giá
Dự đoán
Smiles
CC(=O)NC1=CC=C(C=C1)O
Huấn luyện (72%)
Đánh giá nội (8%)
Đánh giá ngoại (20%)
Chuyển cấu trúc thành đồ thị phân tử
Đồ thị huấn luyện
Đồ thị đánh giá nội
Đồ thị đánh giá ngoại
GraphEmbedder
MultiLayer Perceptron
Tinh chỉnh Siêu tham số
F1 score
Average precision
Active
Inactive