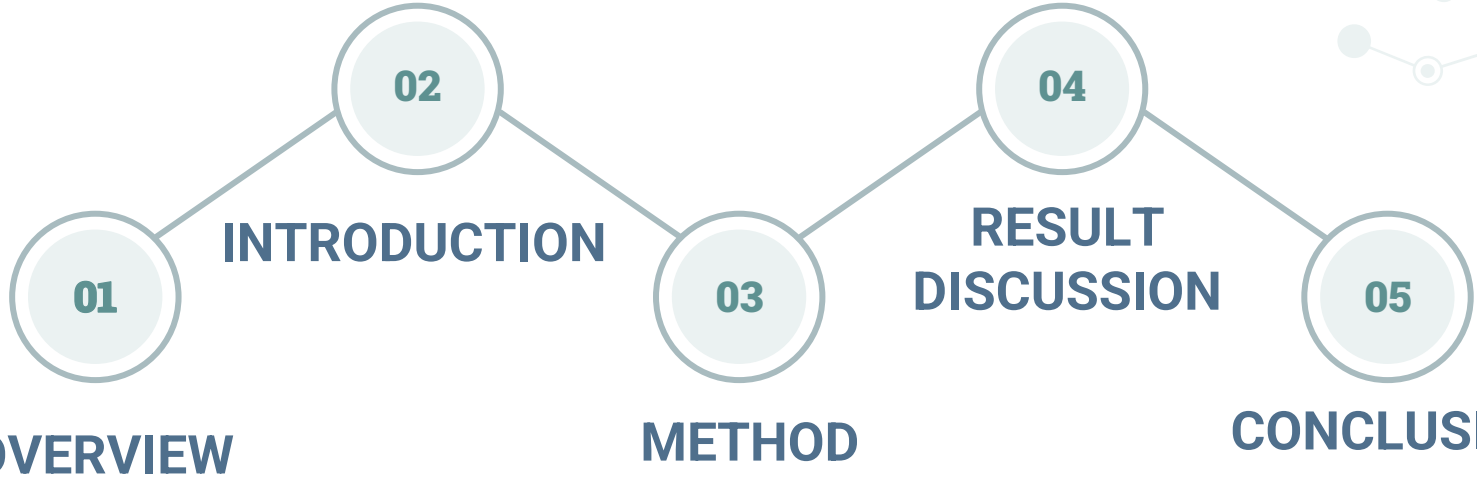
GRADUATION THESIS

Department of Organic Chemistry

IN SILICO MODELING FOR PREDICTION OF POTENTIAL HIV-1 INTEGRASE INHIBITORS

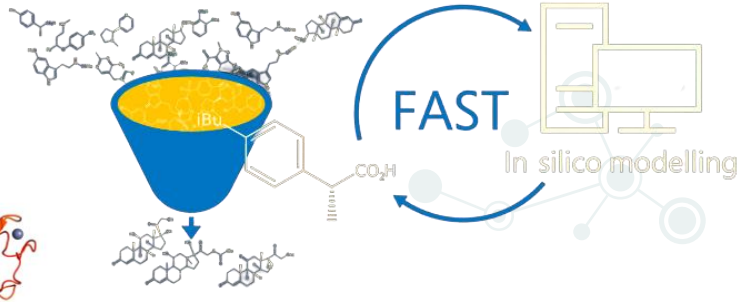
Presenter: Phan Tieu Long
Supervisor: Assoc.Prof.Truong Ngoc Tuyen

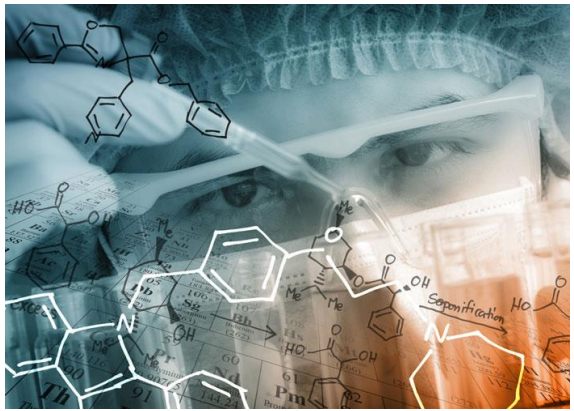
CONTENT


01
OVERVIEW
02
INTRODUCTION
03
METHOD
04
RESULT
DISCUSSION
05
CONCLUSION

OVERVIEW

OVERVIEW





OVERVIEW

MOLECULAR DOCKING

QSAR

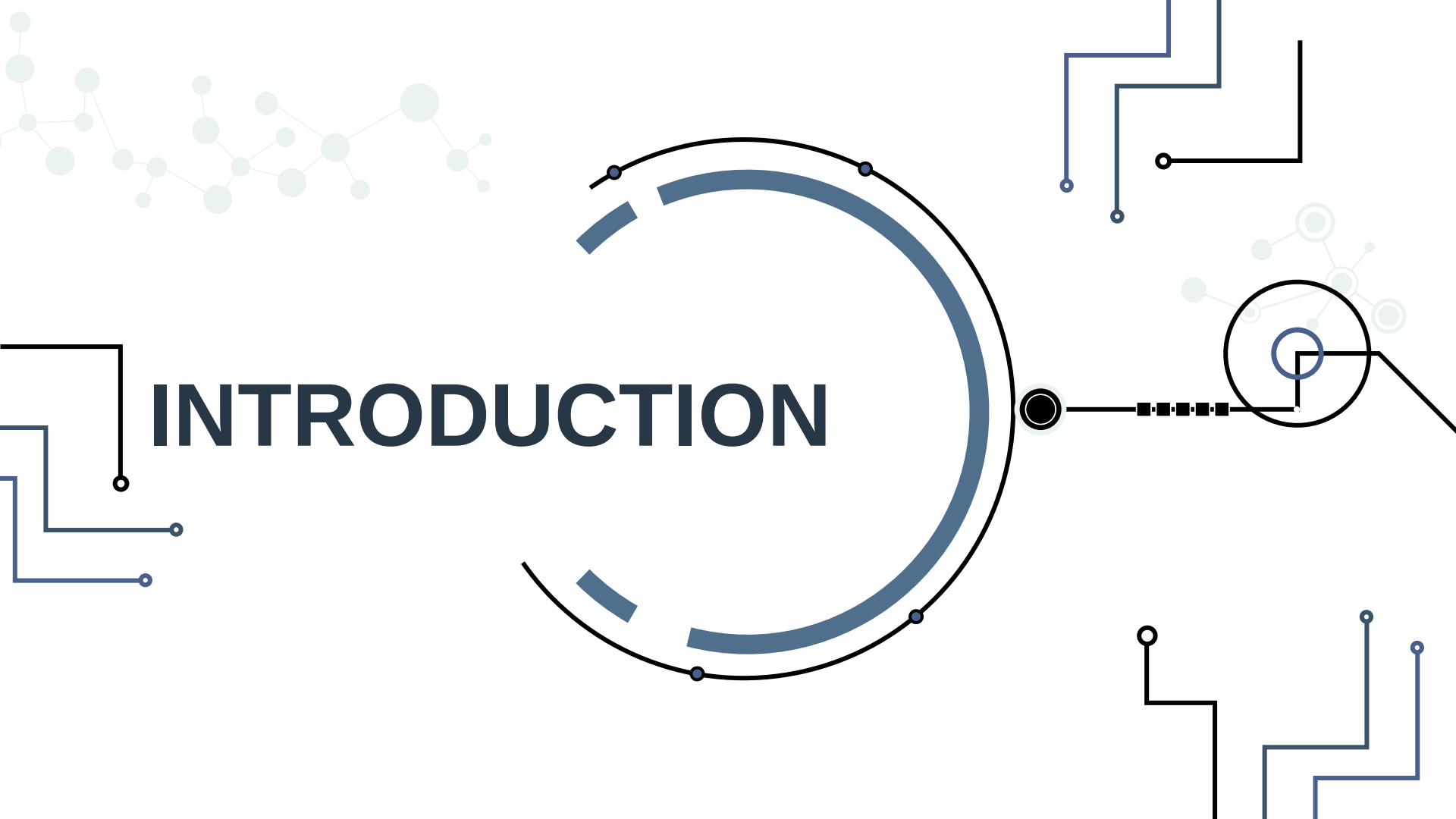
PHAMACOPHORE

VIRTUAL SCREENING

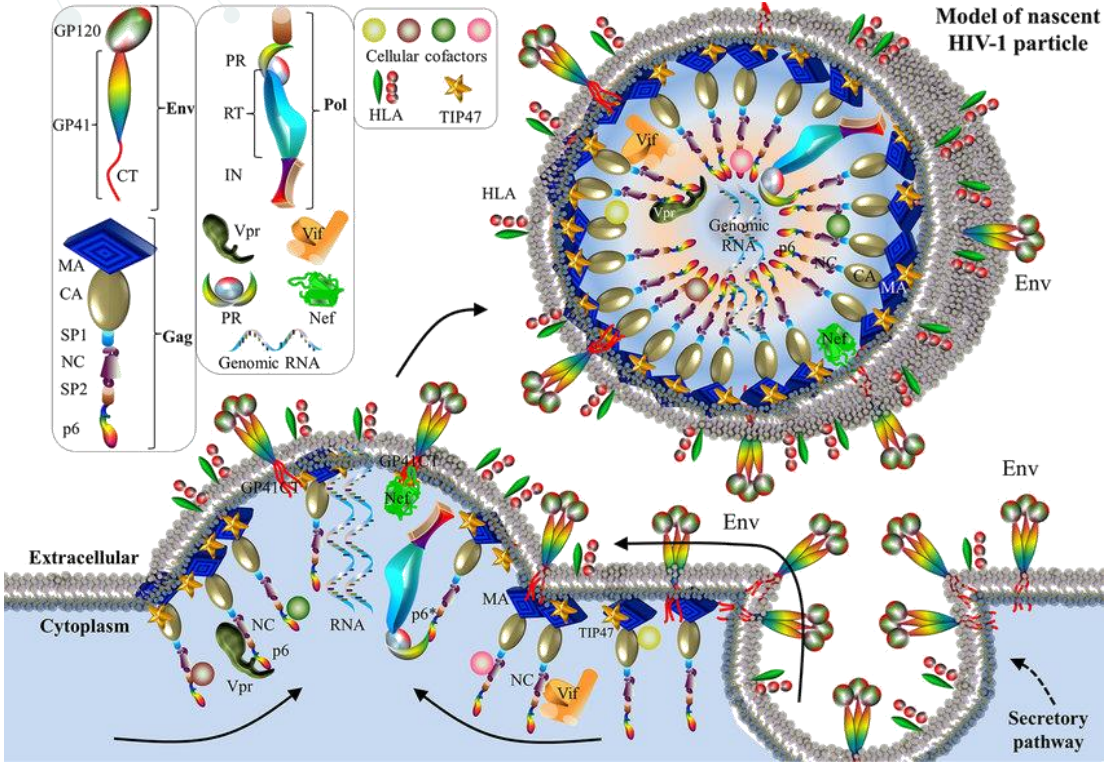
OVERVIEW


01
Pharmacophore
02
QSAR classification
03
QSAR regression
04
Docking
05
Screening

INTRODUCTION

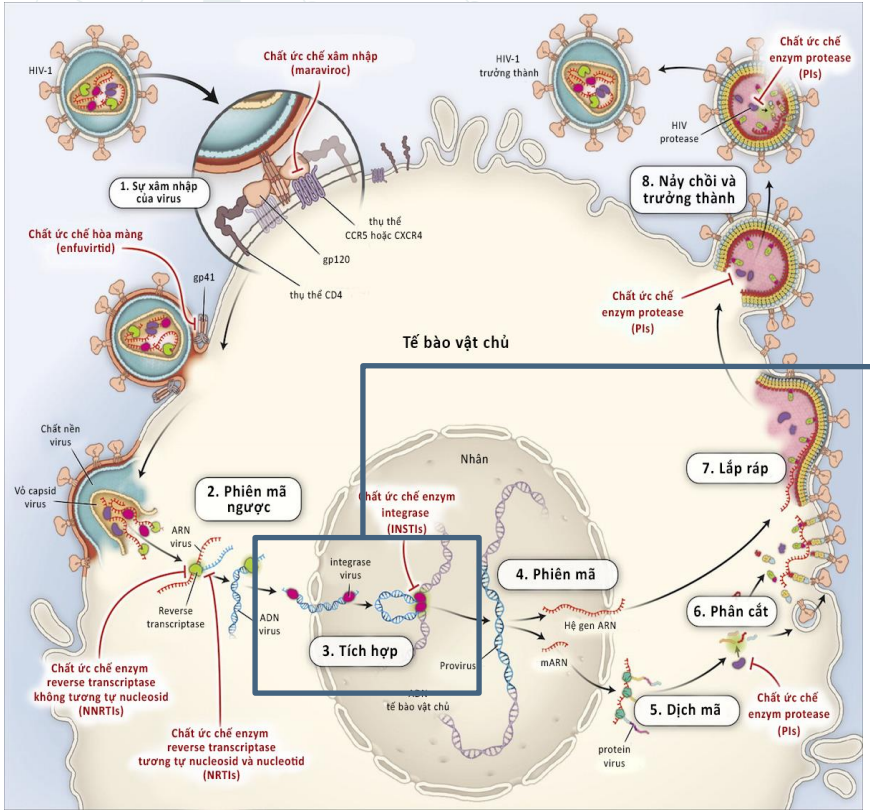


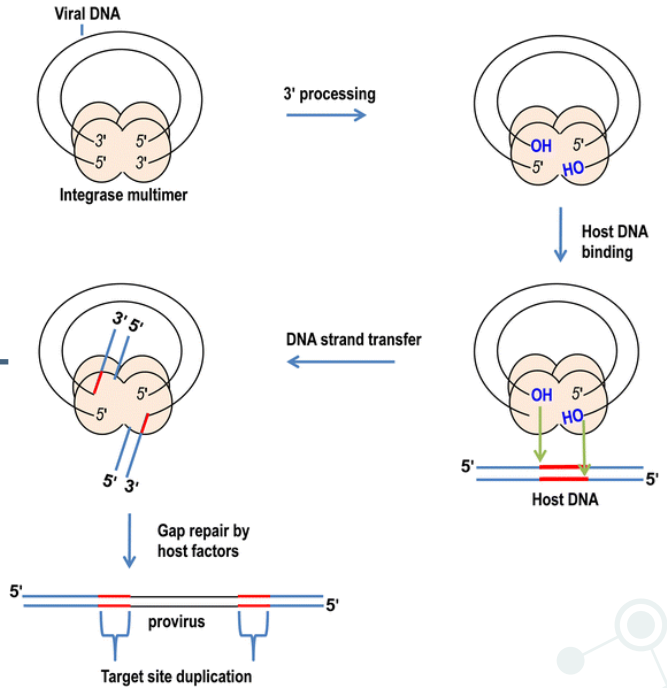
INTRODUCTION



Li G, De Clercq E. HIV Genome-Wide Protein Associations: a Review of 30 Years of Research. Microbiology and molecular biology reviews : MMBR. 2016;80(3):679-731. doi:10.1128/mmbr.00065-15

OVERVIEW

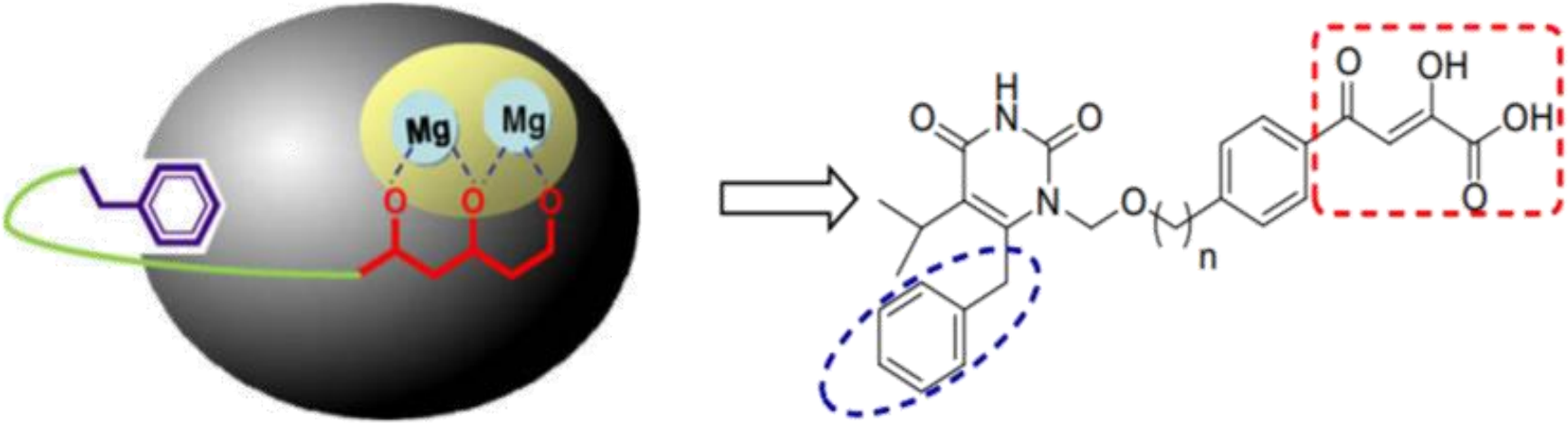




INTEGRATION MECHANISM

Gandhi M. , Gandhi R. T. J. N. E. J. o. M. (2014), "Single-pill combination regimens for treatment of HIV-1 infection".

SAR OF INSTIS



Wang Z., Tang J., Salomon C. E. et al. (2010), "Pharmacophore and structure-activity relationships of integrase inhibition within a dual inhibitor scaffold of HIV reverse transcriptase and integrase", Bioorganic & Medicinal Chemistry. 18 (12), pp. 4202-4211

FDA APPROVED DRUGS

Cabotegravir
2021

Bictegravir
2018

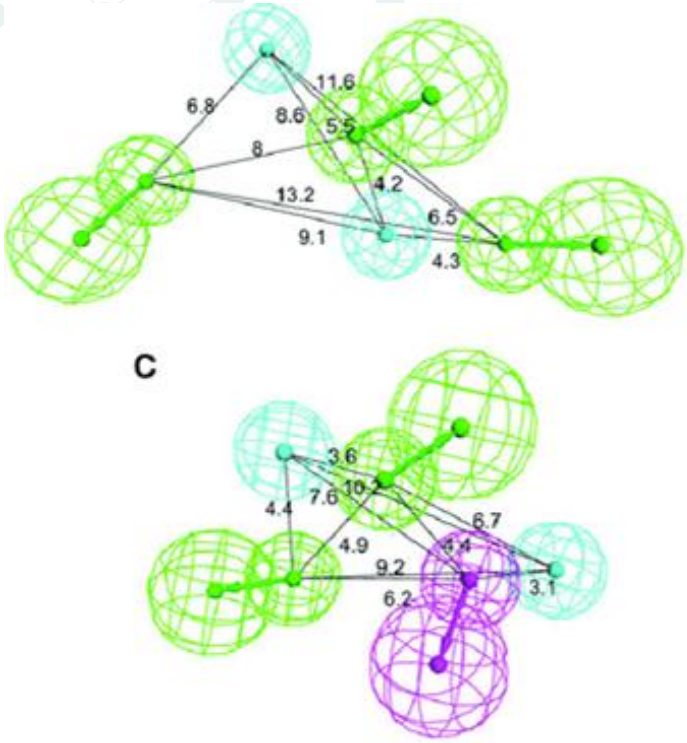
Dolutegravir
2013

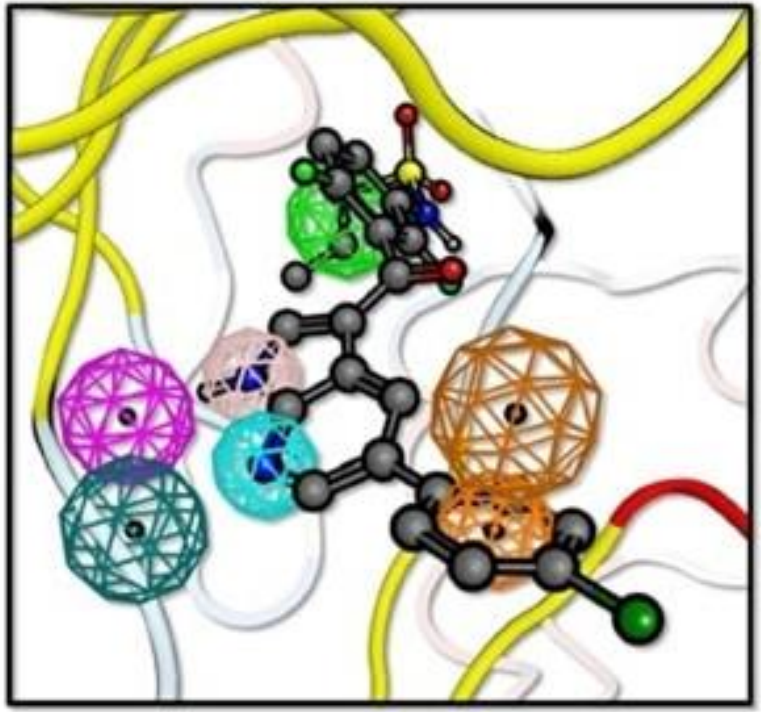
Elvitegravir
2012

Raltegravir
2007

OVERVIEW

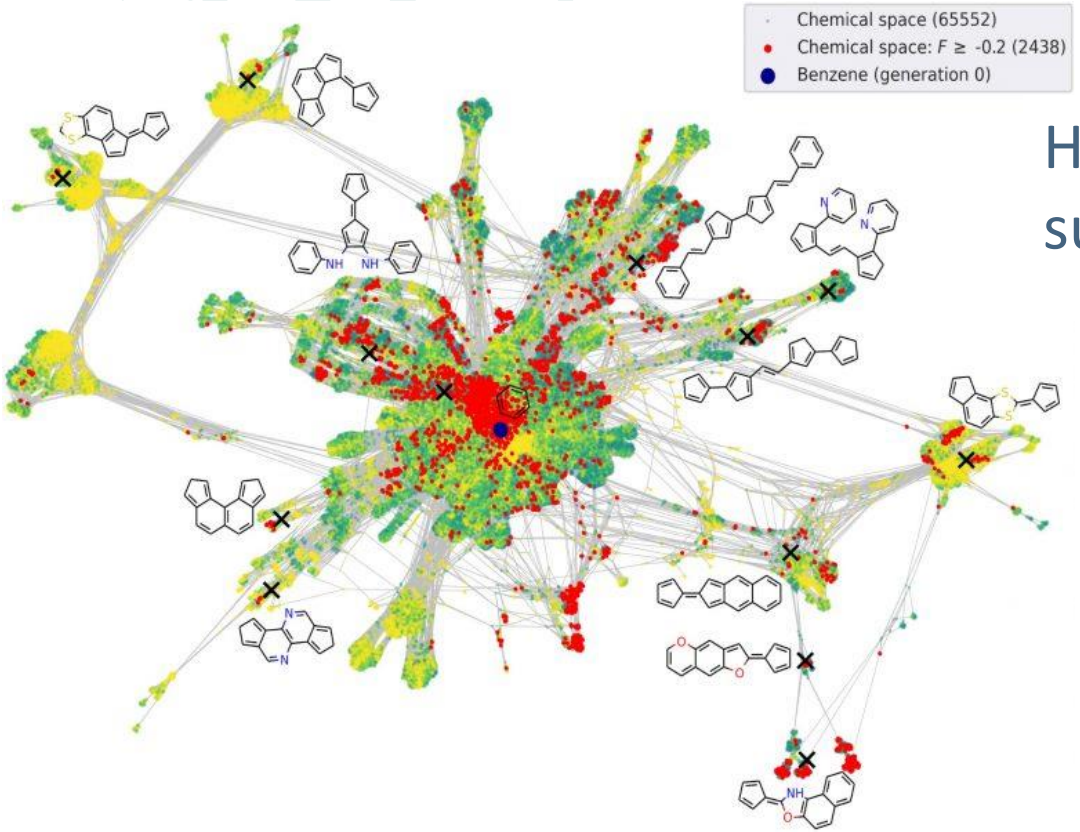
PHARMACOPHORE





Ligand-based

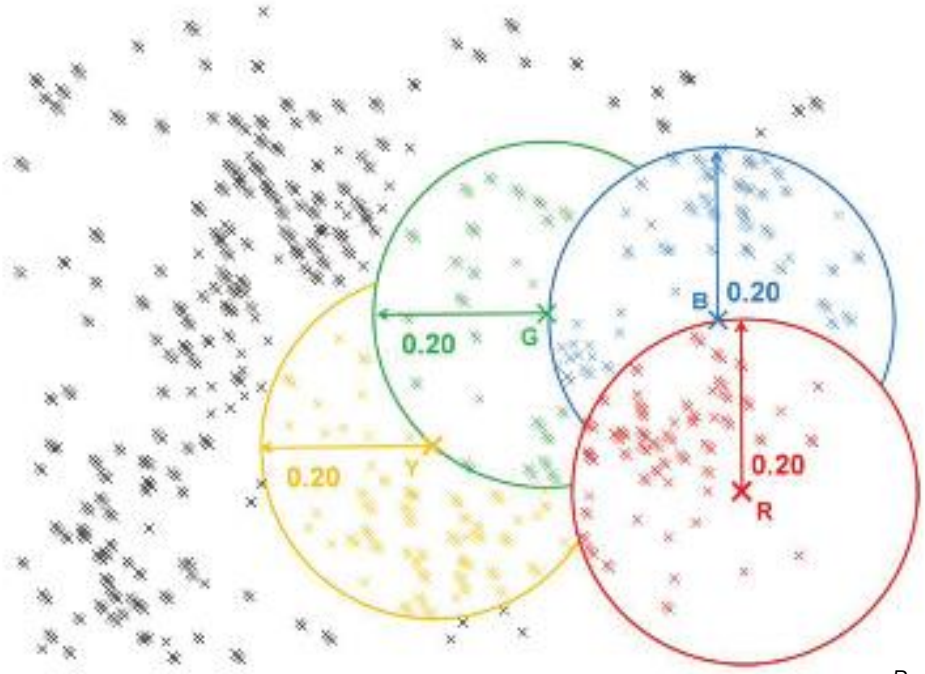
Protein-based



How can we select diverse subset for training model?

PHARMACOPHORE

Butina algorithm



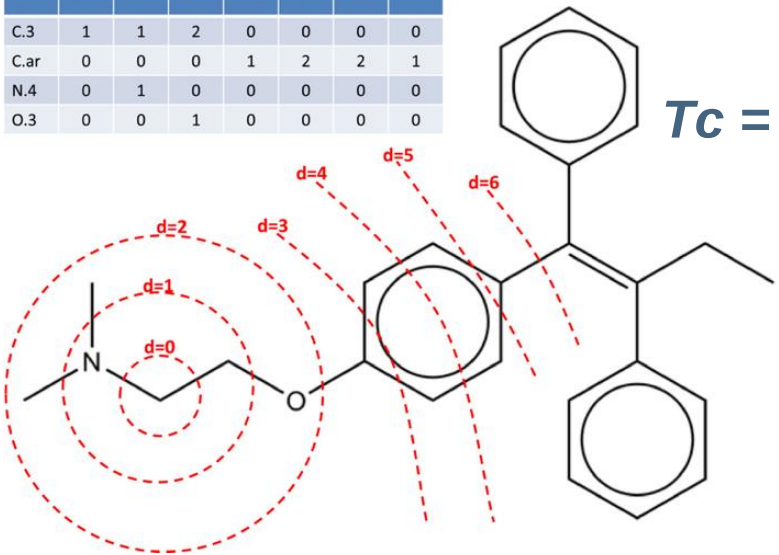
- Calculate similarity matrix
- Select centroid
- ”Single Pass” technique

Downs GM, Barnard JM. Clustering methods and their uses in computational chemistry. Reviews in computational chemistry. 2002;18:1-40.

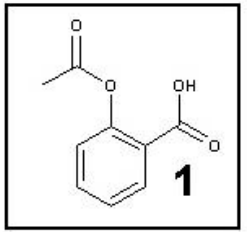
PHARMACOPHORE

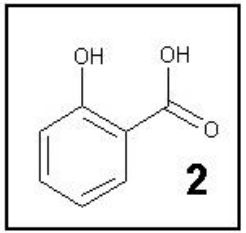
Molecular fingerprint

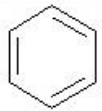
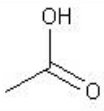
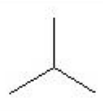
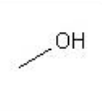
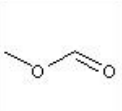
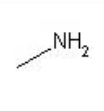
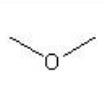
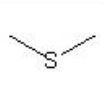
	d=0	d=1	d=2	d=3	d=4	d-=5	d=6
C.3	1	1	2	0	0	0	0
C.ar	0	0	0	1	2	2	1
N.4	0	1	0	0	0	0	0
O.3	0	0	1	0	0	0	0



$$Tc = \frac{c}{a+b-c} = 0{,}6$$





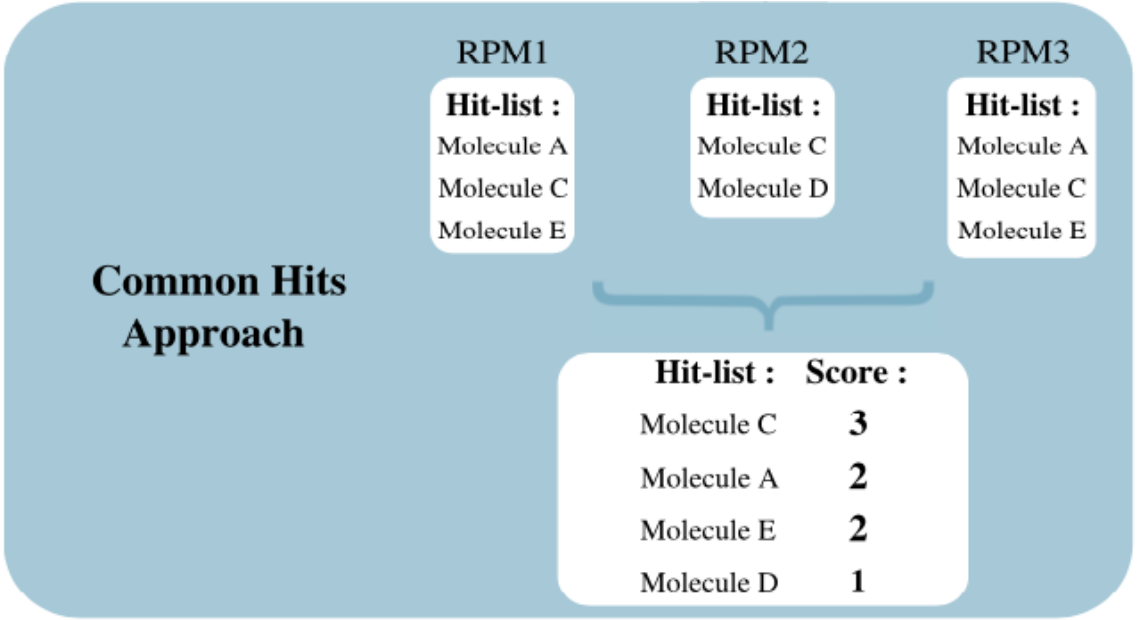
1	1	1	0	1	1	0	1	0
2	1	1	0	1	0	0	0	0
								

Rogers D, Hahn M. Extended-connectivity fingerprints. Journal of chemical information and modeling. 2010;50(5):742-54. doi:10.1021/ci100050t

Bender A, Glen RC. Molecular similarity: a key technique in molecular informatics. Organic & biomolecular chemistry. 2004;2(22):3204-18. doi:10.1039/b409813g

PHARMACOPHORE

Model optimization

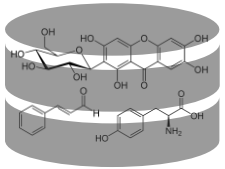


- Change query threshold
- Change overlap ratio
- Add exclusion volume
- Change ligand's shape
- Common Hits Approach

Wieder M, Garon A, Perricone U, et al. Common Hits Approach: Combining Pharmacophore Modeling and Molecular Dynamics Simulations. Journal of chemical information and modeling. 2017;57(2):365-385. doi:10.1021/acs.jcim.6b00674

OVERVIEW

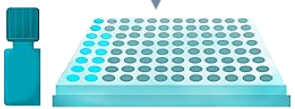
Database



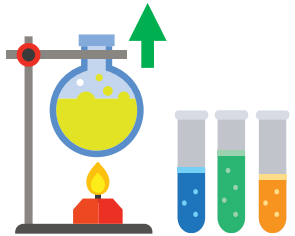
QSAR MODEL

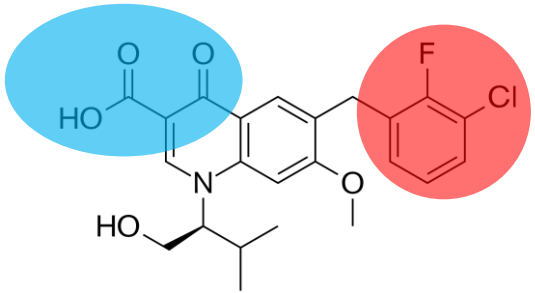
Lead compound

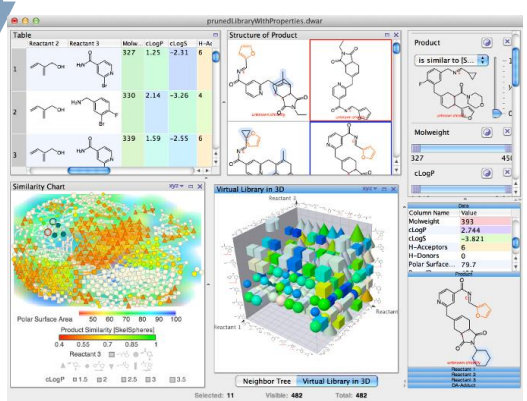
in vitro



Synthesis

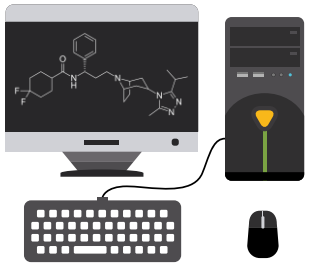






QSAR

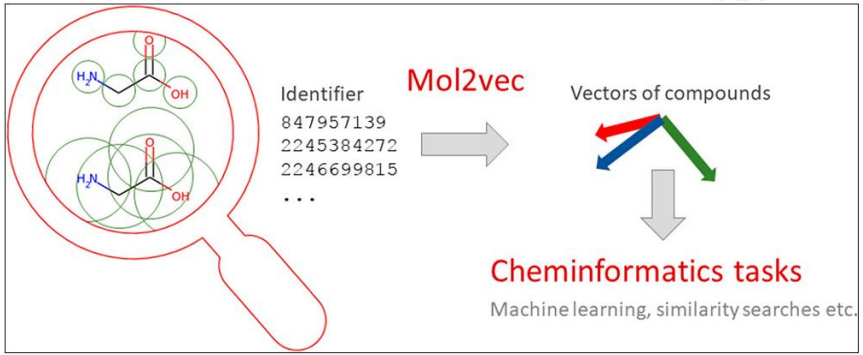
Design
in silico



OVERVIEW

QSAR MODEL

Molecular representation



mordred-descriptor/
documentation

http://mordred-descriptor.github.io/documentation/master

1 Contributor | 0 Issues | 0 Stars | 0 Forks

and Machine Learning

phi-grib/**PaDEL-descriptor-ws**

PaDEL ws descriptors engine

1 Contributor | 1 Issue | 11 Stars | 6 Forks

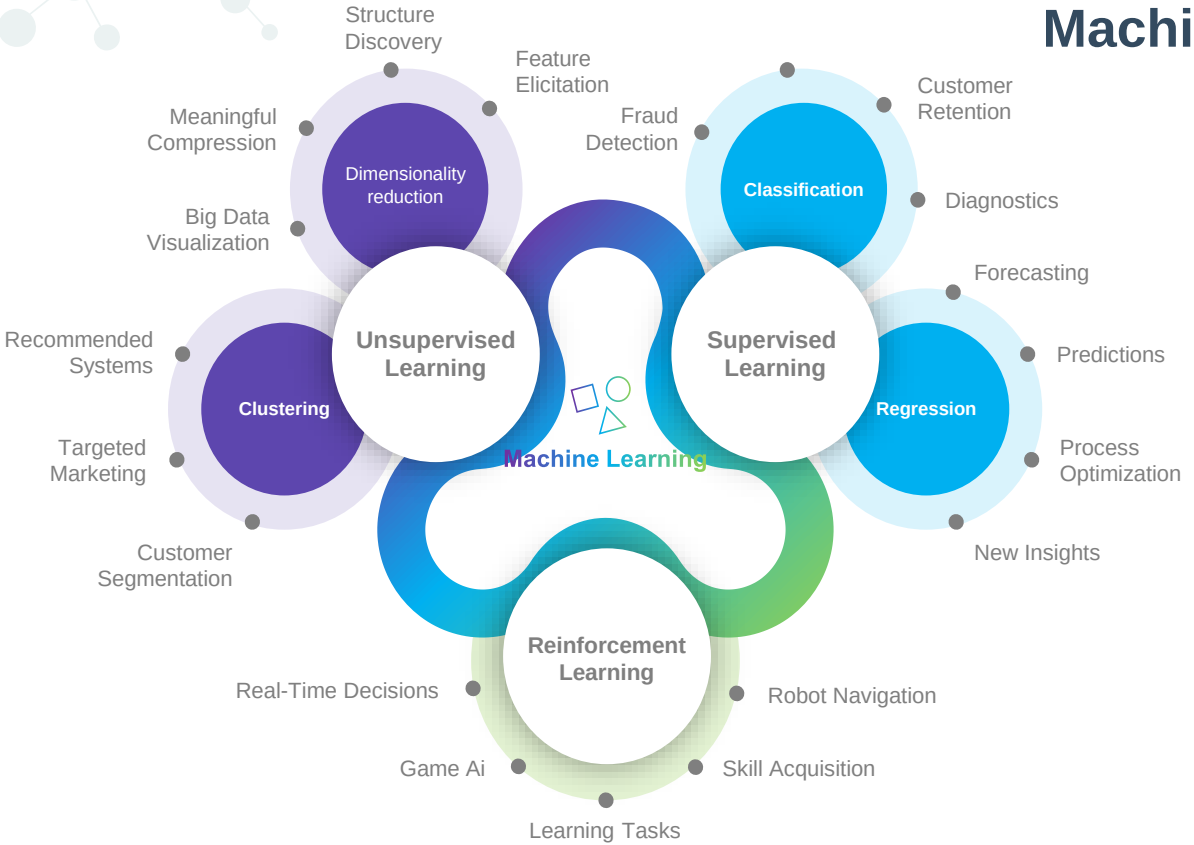
RDKit

Open-Source Cheminformatics and Machine Learning

OVERVIEW

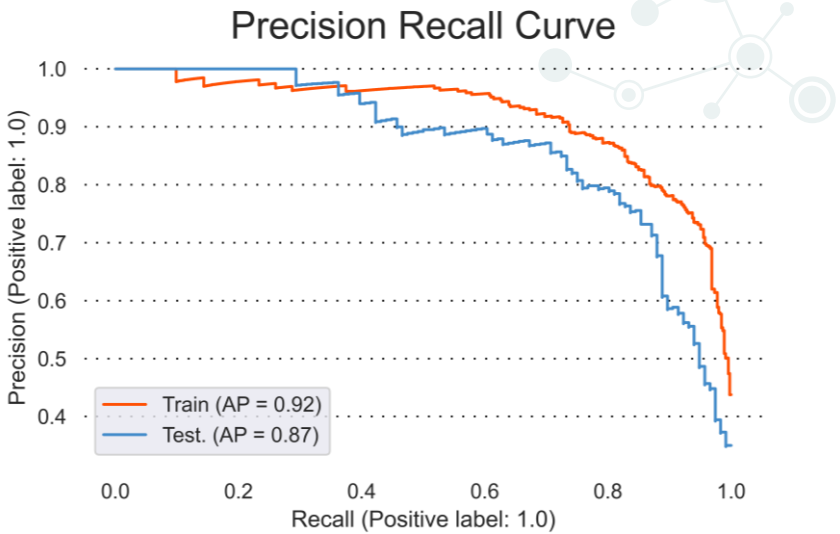
QSAR MODEL

Machine Learning



Evaluation metric - Classification

		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$



$$F_1 = \frac{2(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

$$AP = \sum_n (R_n - R_{n-1}) P_n$$

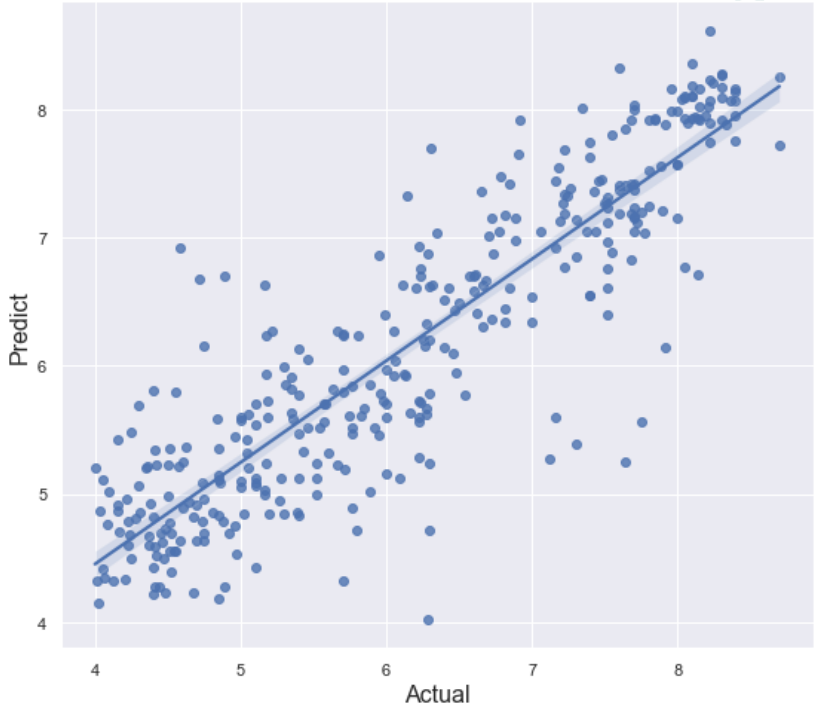
QSAR MODEL

Evaluation metric - Regression

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}$$

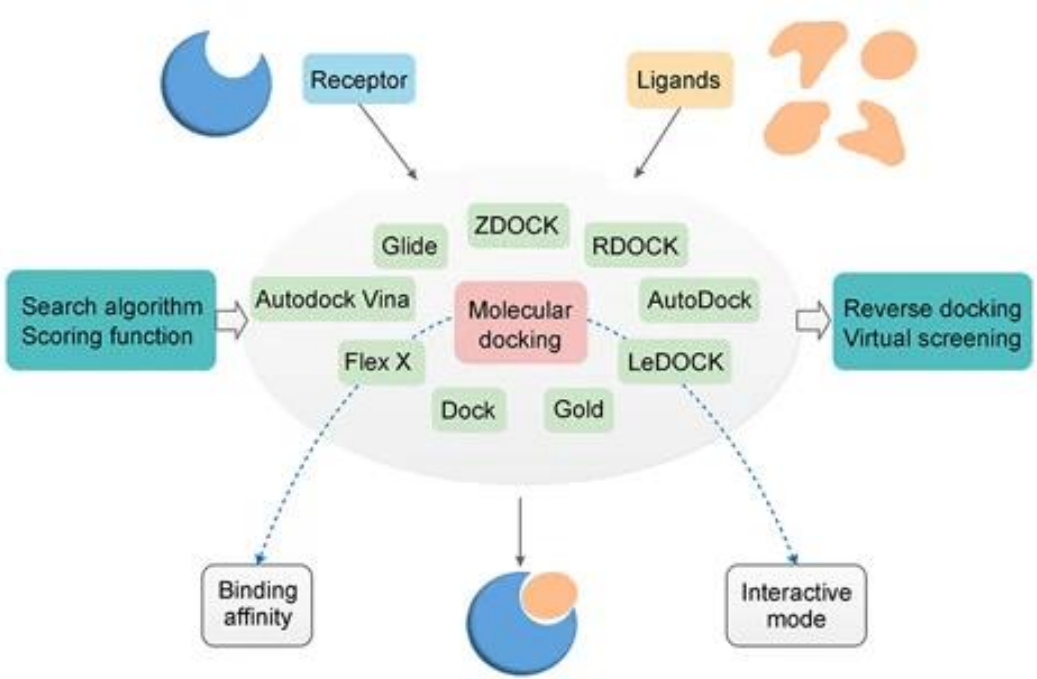
$$RMSE = \frac{1}{n} \times \sqrt{\|y - \hat{y}\|_2^2}$$

$$MAE = \frac{1}{n} \times \|y - \hat{y}\|_1$$



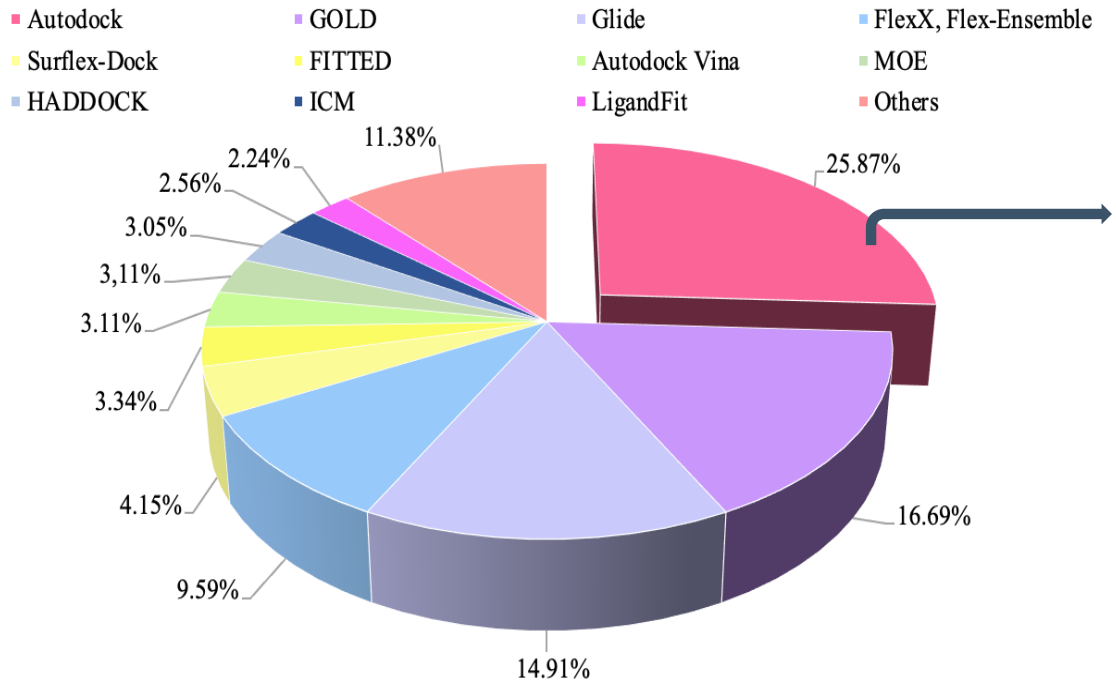
OVERVIEW

MOLECULAR DOCKING



1. Conformation
2. Scoring function
3. Searching

MOLECULAR DOCKING



Autodock had highest citation in ISI Web of Science (2005)

MOLECULAR DOCKING

01 **1989**

Autodock

Physics-based scoring functions

02 **2010**

Autodock vina 1.1.2

Empirical Scoring Functions
Multi-core processing

03 **2013**

Smina

Can modify scoring function

04 **2017**

Qvina2

Increase docking speed by BFGS

05 **2021**

Autodock vina 1.2.3

Python integrated
Batch-dock
AD4 scoring function

06 **2021**

Autodock-GPU

GPU
Batch-dock

07 **2022**

Vina-GPU

GPU

OVERVIEW

MOLECULAR DOCKING

Retrospective control

Active compounds

Active reference for docking model

Decoy

Inactive reference for docking model

Metrics

AUC-ROC; G-mean, TPR

Meaning

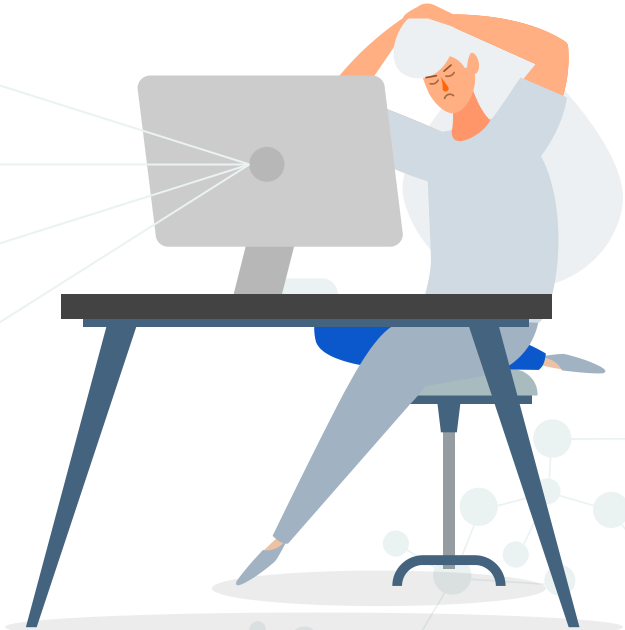
Performance of docking softwares

01

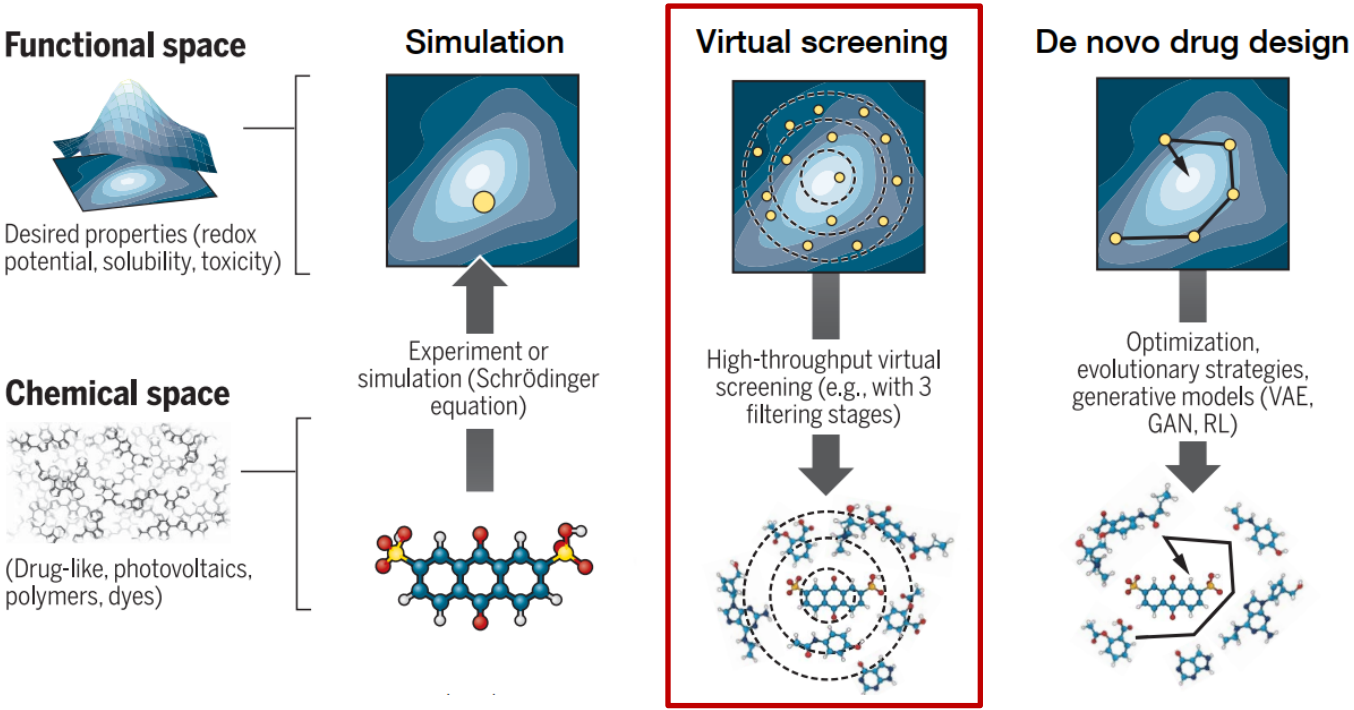
02

03

04



Computational drug discovery: three schemes

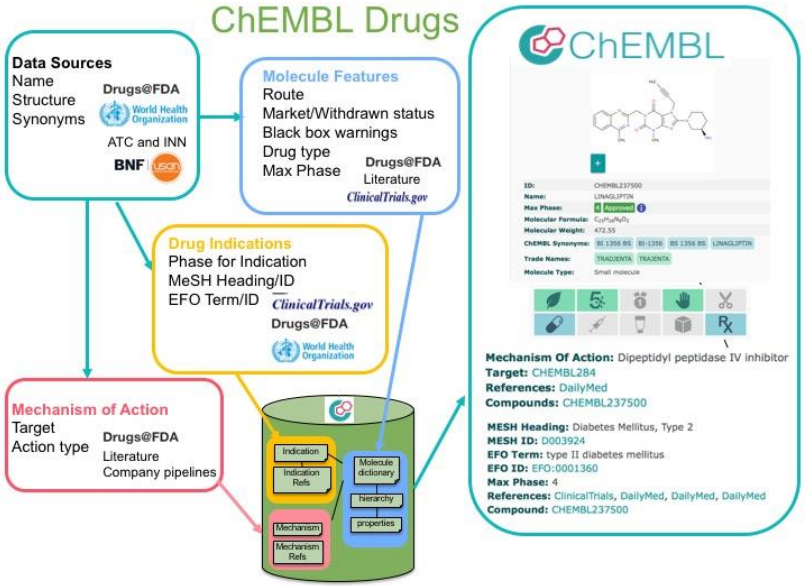


Sanchez-Lengeling *et al.*, Science 361, 360-365 (2018)


DATA

DATA

DATA FOR DEVELOPING MODELS



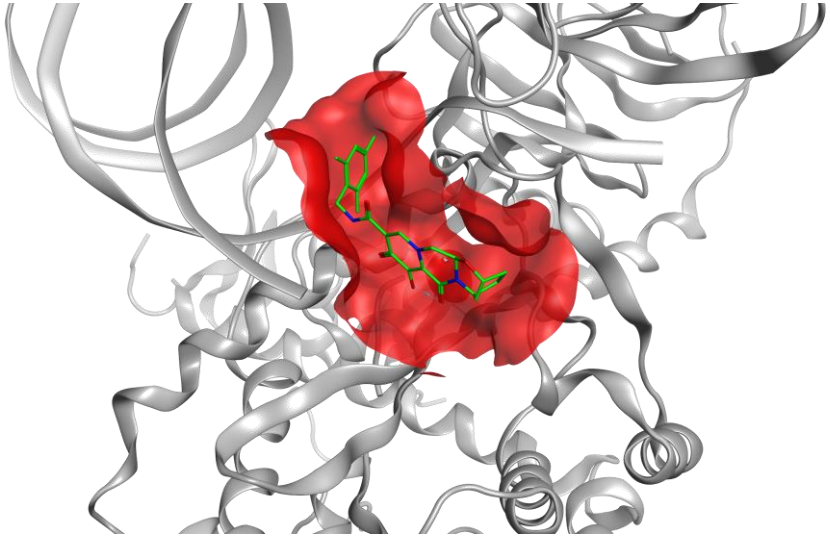
DATA FOR SCREENING

Internal database

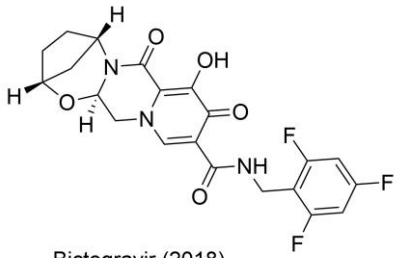
2016 compouds

DATA

PROTEIN

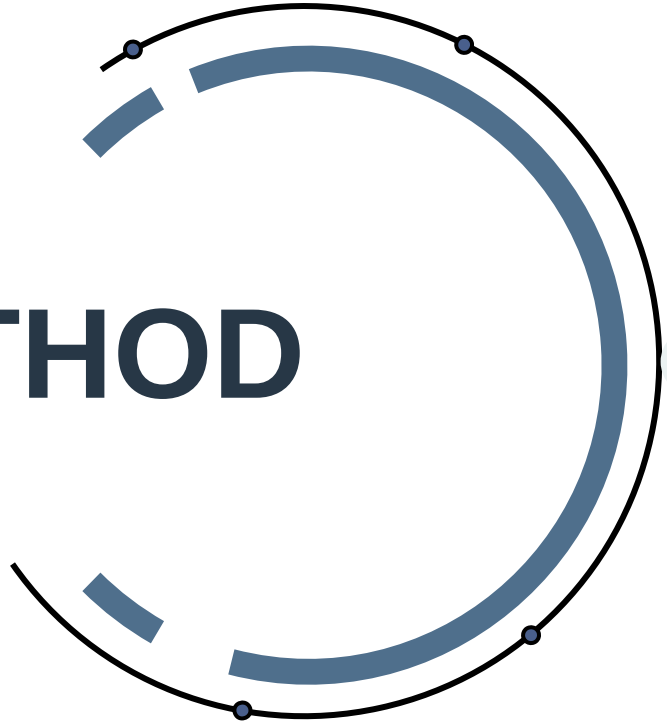


- Cryo-em
- 2,9 Å
- PDB ID: 6PUW



Bictegravir (2018)

METHOD



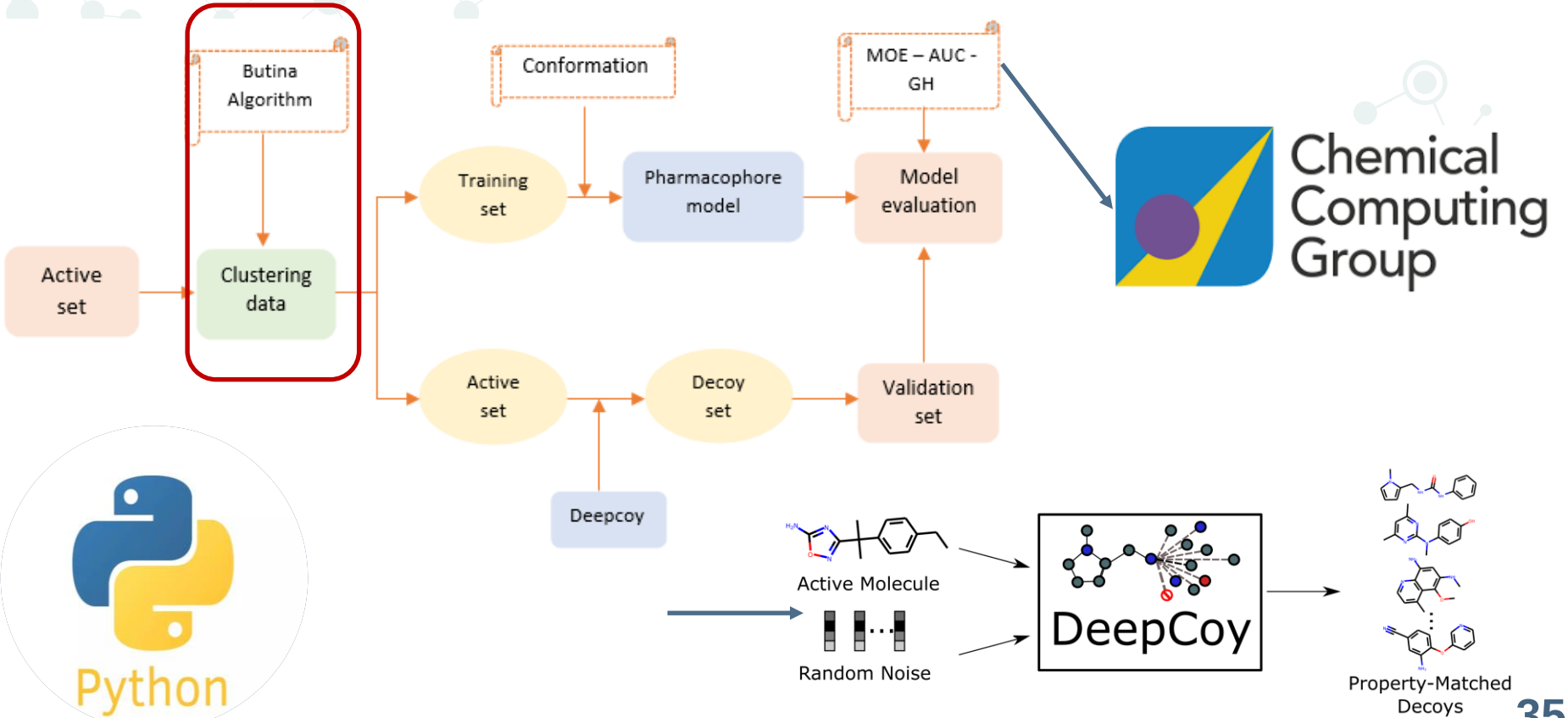
Method

01

PHARMACOPHORE

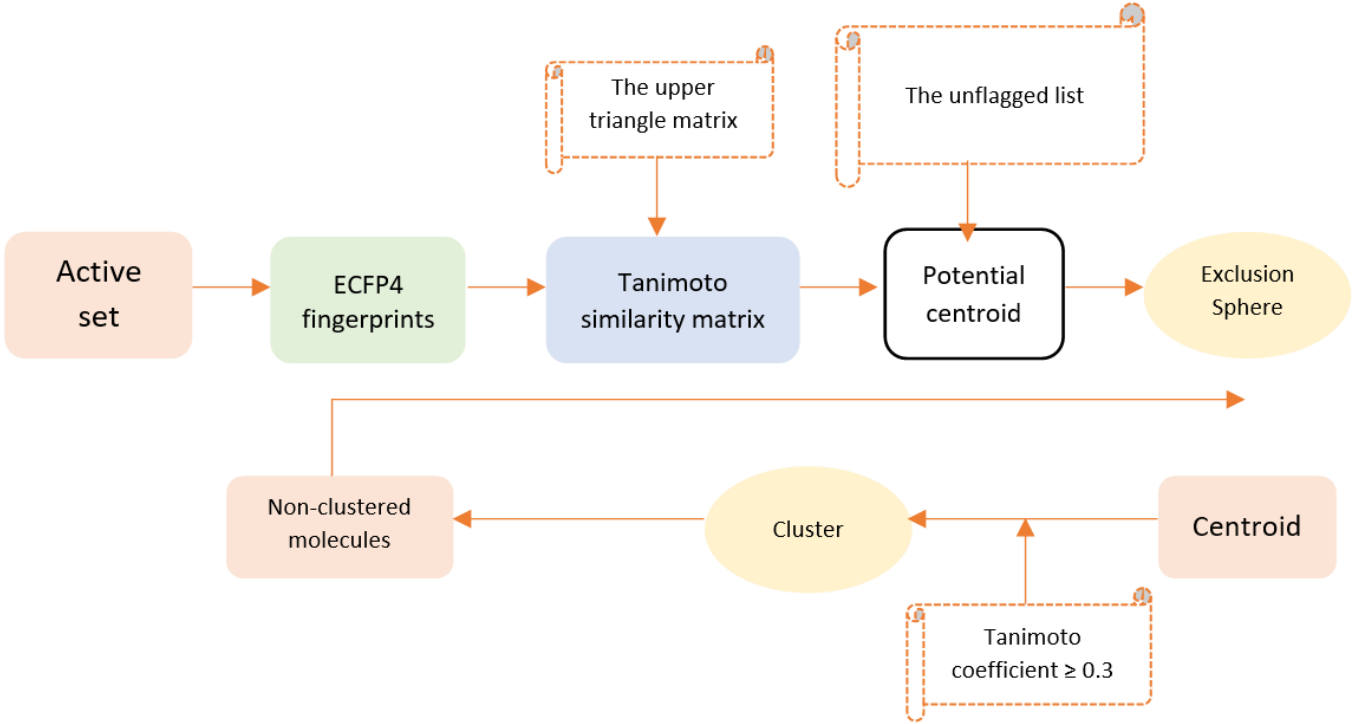
METHOD

PHARMACOPHORE



PHARMACOPHORE

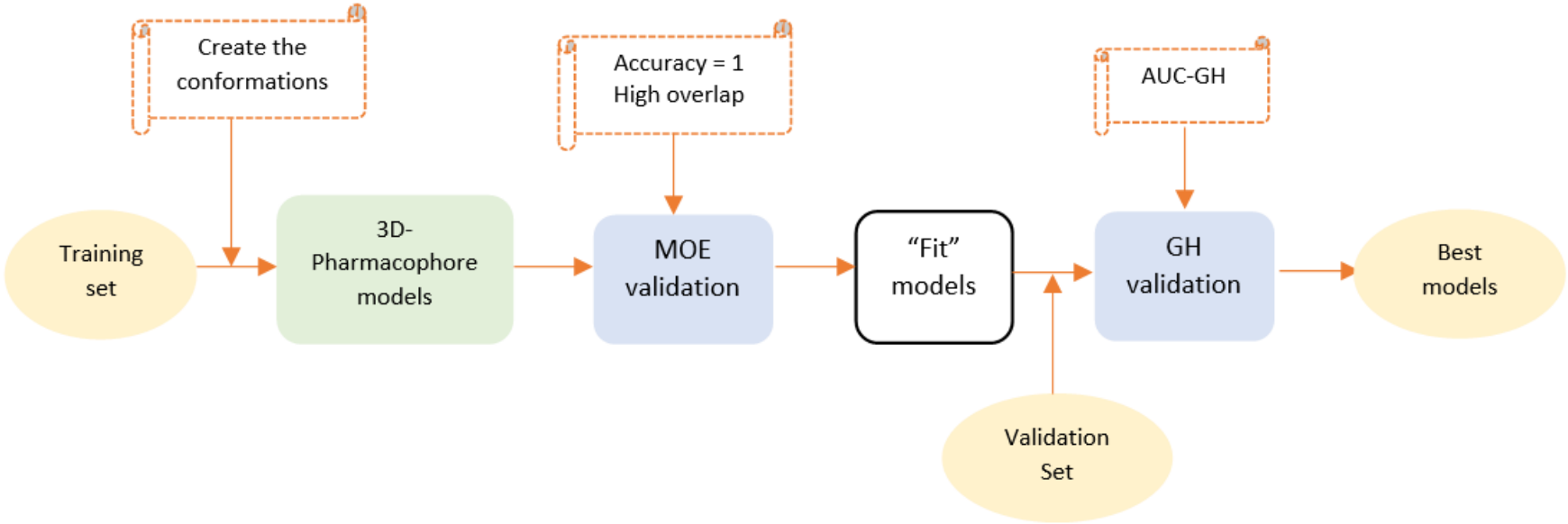
Molecules clustering



METHOD

PHARMACOPHORE

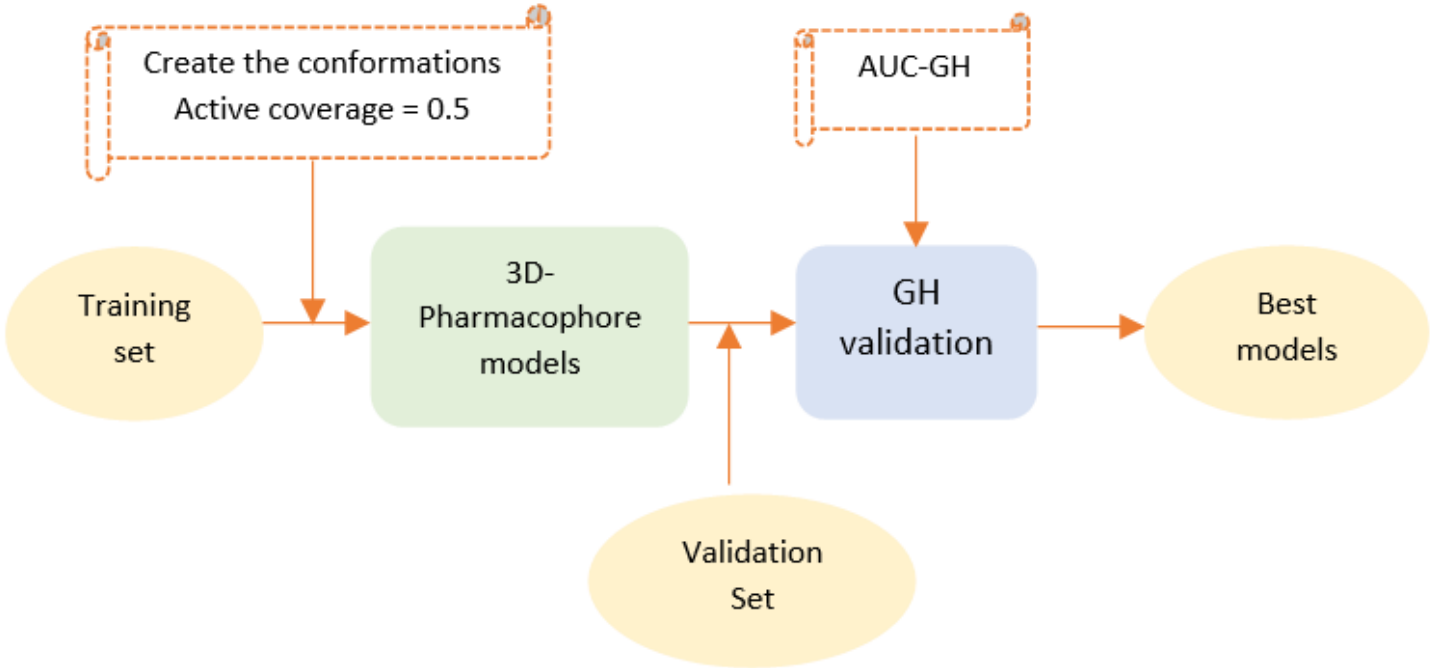
Local Search



METHOD

PHARMACOPHORE

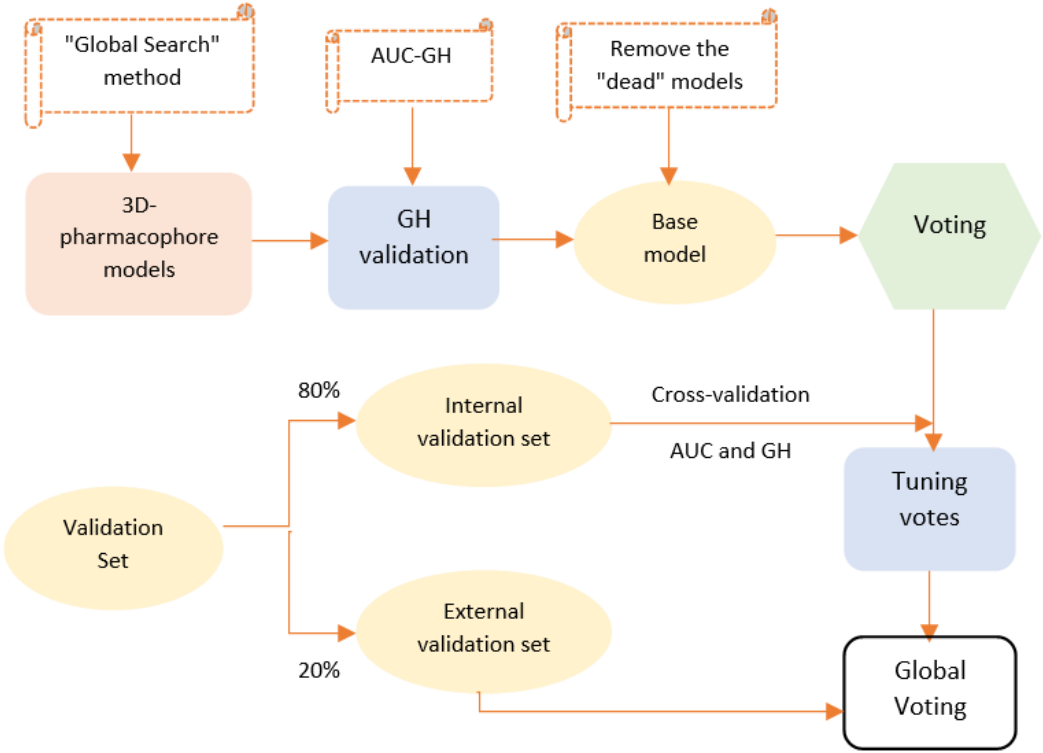
Global Search



METHOD

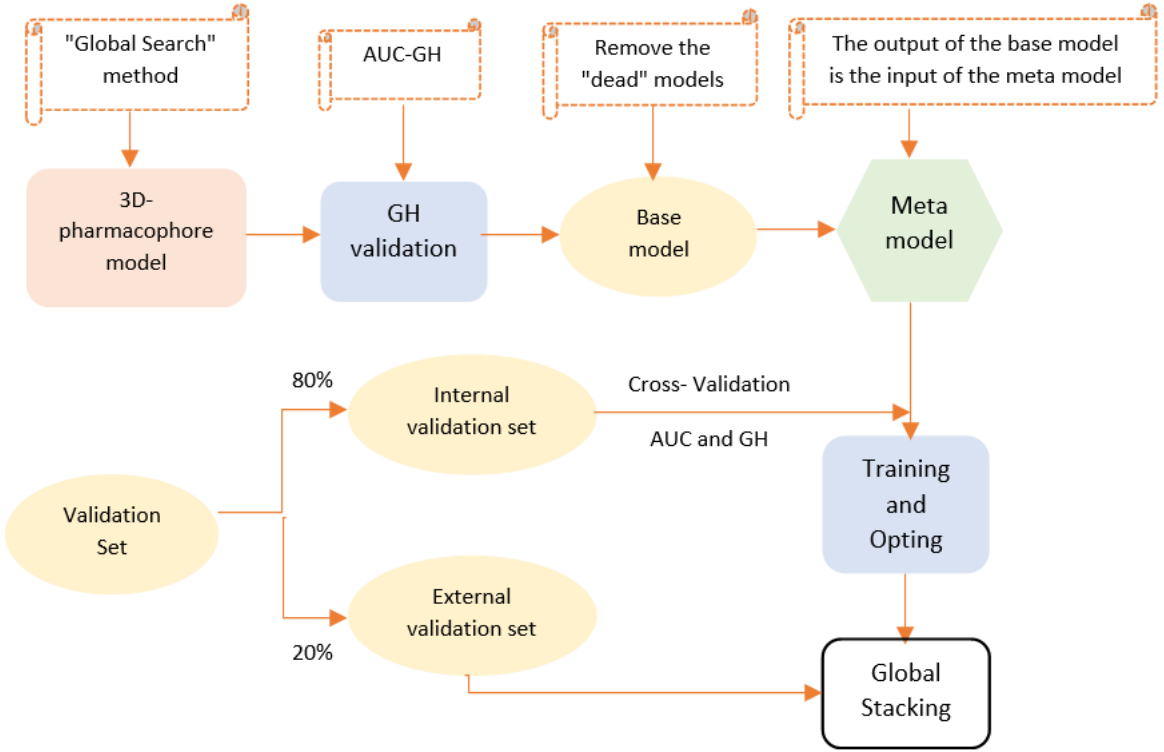
PHARMACOPHORE

Voting



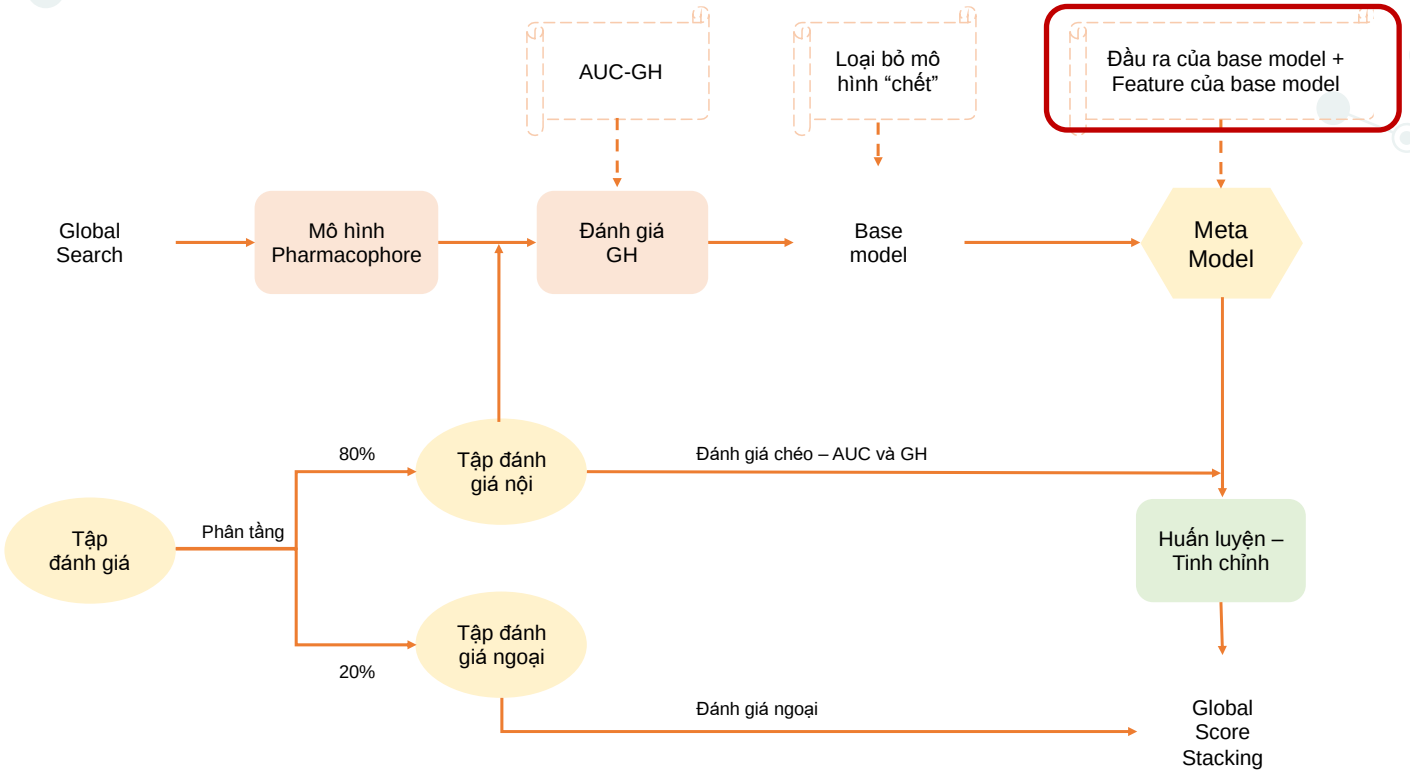
MÔ HÌNH PHARMACOPHORE

Stacking



PHARMACOPHORE

Score Stacking



PHARMACOPHORE EVALUATION

Model

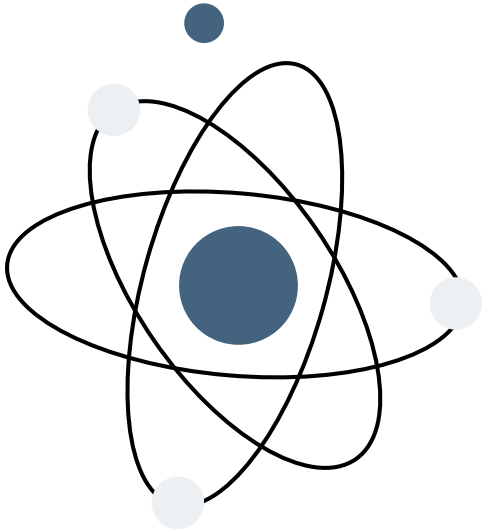
1. Local Search
2. Global Search
3. Voting
4. Stacking
5. Score Stacking
6. **Feature Score Stacking**
7. **Optimized Feature Score Stacking**

Evaluation method

- **Cross validation**: post hoc Wilcoxon
- **External validation**: generalization
- **Metric**:

AUC > GH > F1

METHOD

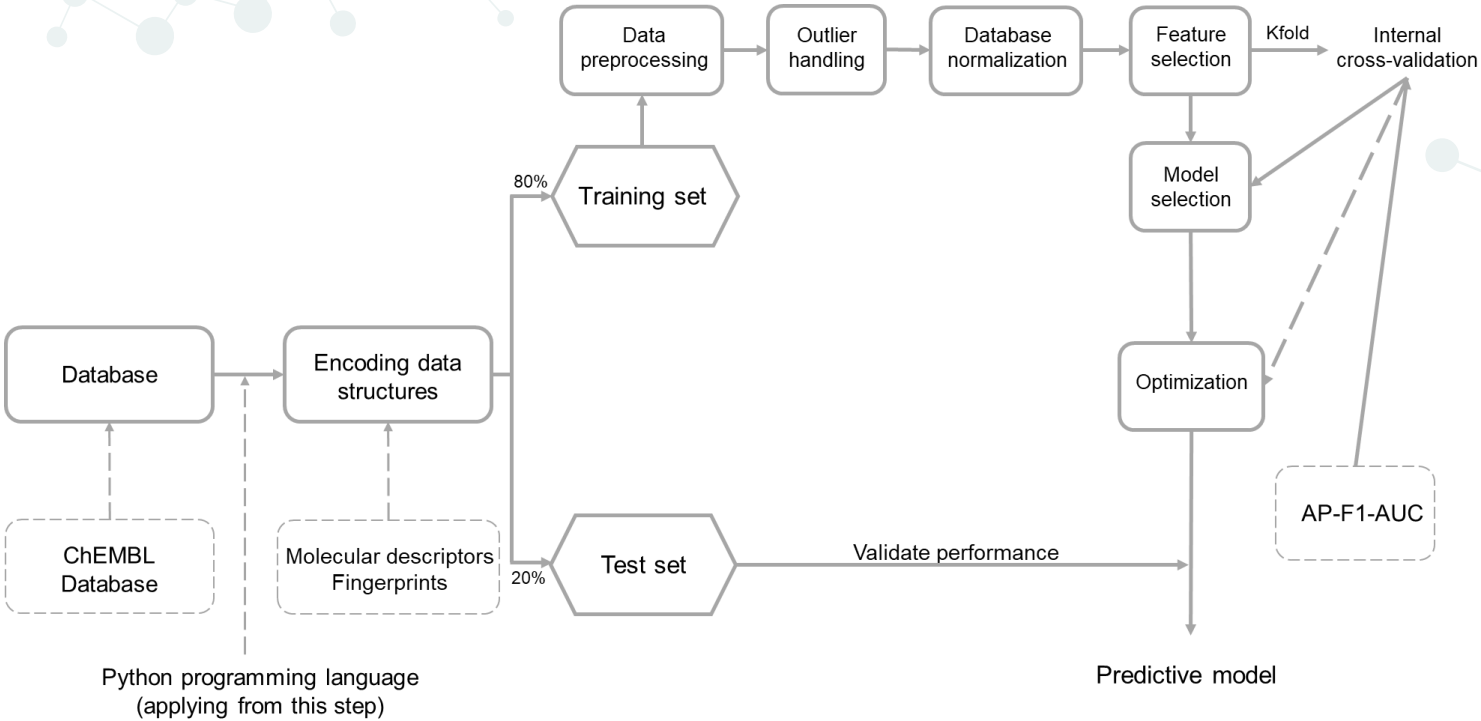


02

QSAR Model

METHOD

MÔ HÌNH QSAR



Source code

MÔ HÌNH QSAR

Biểu diễn phân tử

01

Thông số mô tả

2D-Mordred
RDKit
Mol2vec

02

Dấu vân đường dẫn

RDK5
RDK6
RDK7

03

Dấu vân tay từ điển

MACCS
PubChem
Avalon

04

Dấu vân tay tròn

ECFP2
ECFP4
ECFP6

05

Dấu vân tay PH4

Cats2D
Pharmacophore Gobbi

06

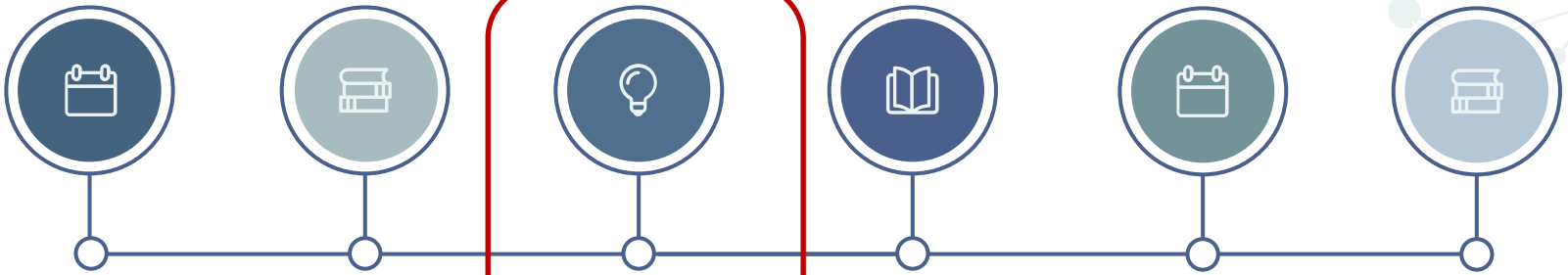
Dấu vân tay khác

map4
secfp

METHOD

QSAR

Data-centric



01

Xây dựng cơ sở dữ liệu

02

Khai phá dữ liệu

03

Lựa chọn bộ dữ liệu tối ưu

04

Lựa chọn đặc trưng

05

Lựa chọn mô hình

06

Tinh chỉnh mô hình

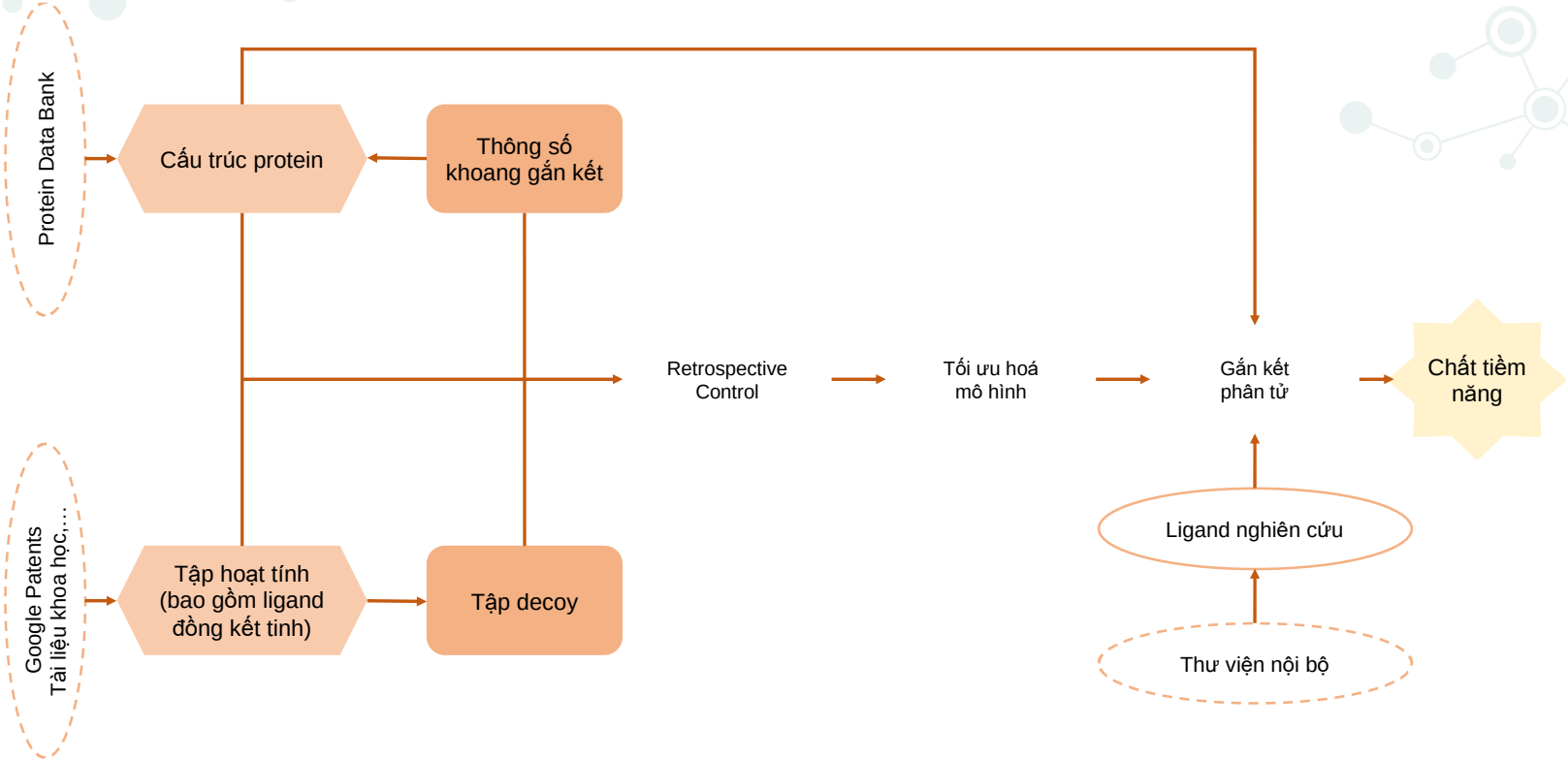
PHƯƠNG PHÁP

03

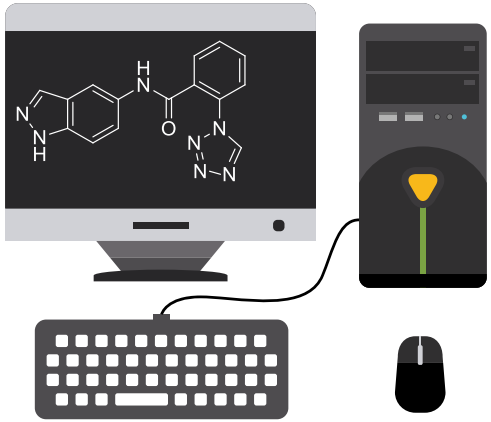
MOLECULAR DOCKING

METHOD

MOLECULAR DOCKING



METHOD

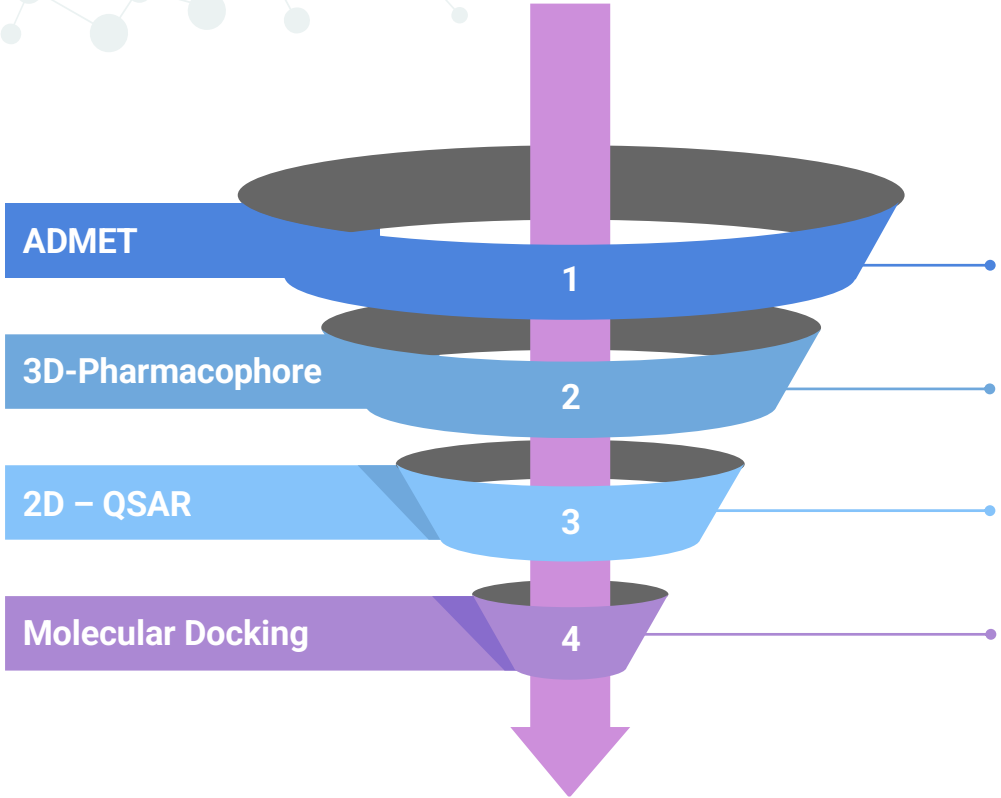


04

VIRTUAL SCREENING

METHOD

VIRTUAL SCREENING



Đánh giá sự hấp thu, phân bố, chuyển hoá, thải trừ và độc tính của các hợp chất

Dự đoán thoả hay không thoả mô hình 3D-Pharmacophore

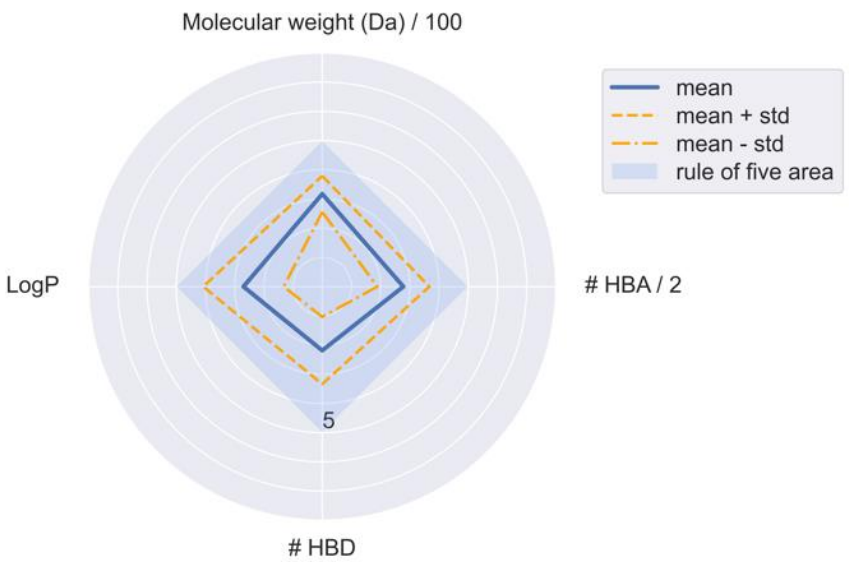
Dự đoán có hay không có hoạt tính và tính giá trị pIC50

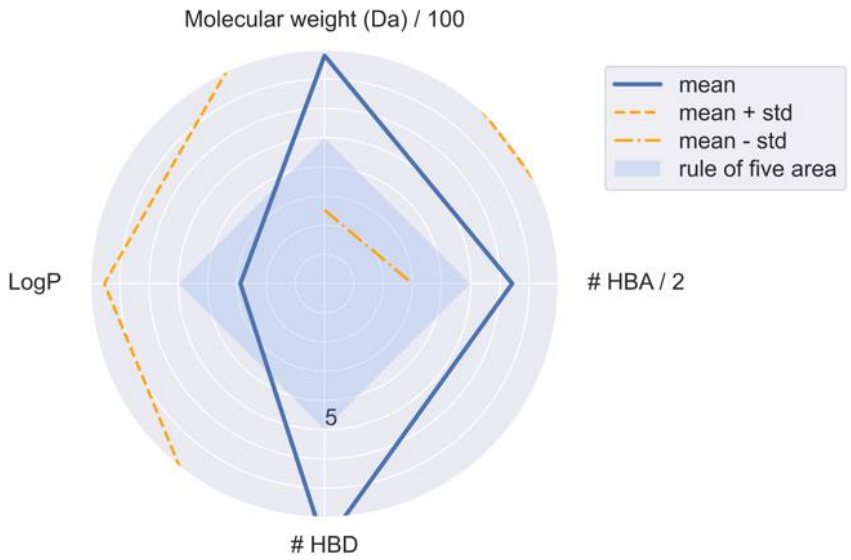
Đánh giá và dự đoán khả năng gắn kết của phân tử nhỏ (ligand) vào cấu trúc mục tiêu (receptor)

RESULT- DISCUSSION

RESULT

2445 compounds





RESULT

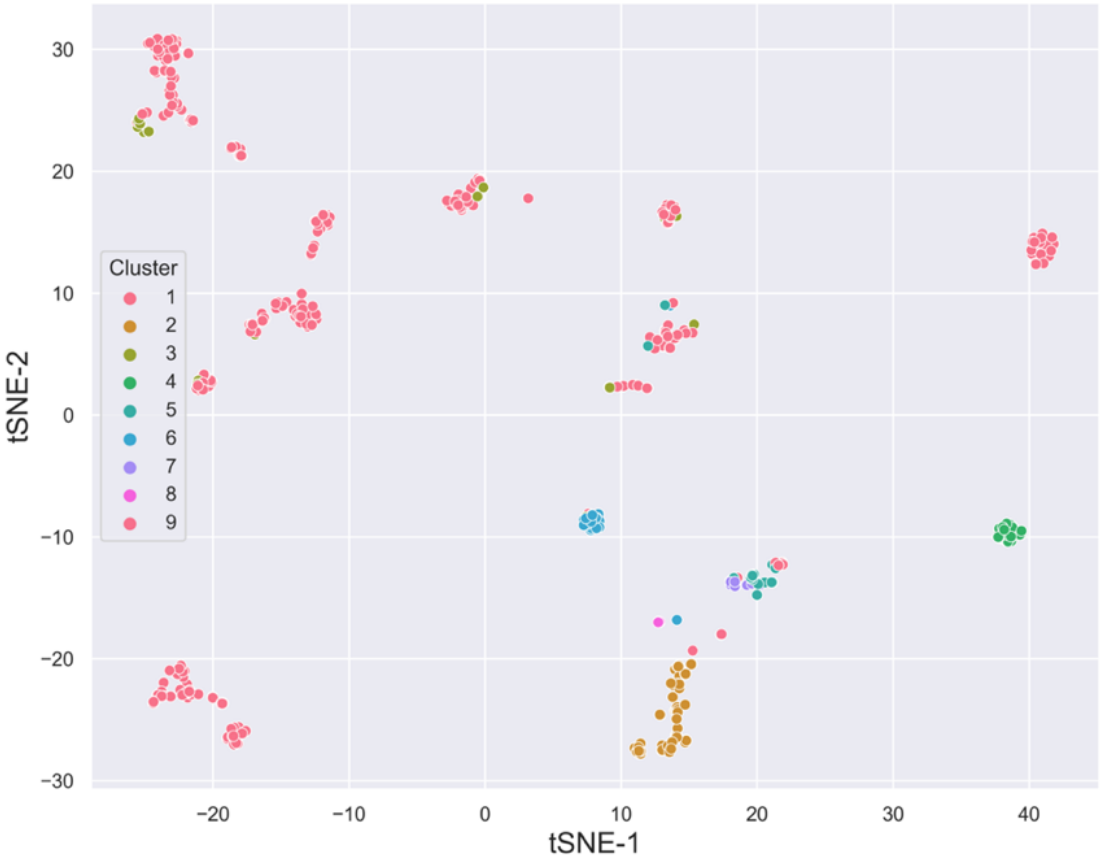
1

Pharmacophore

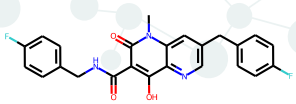
Ligand-based pharmacophore

RESULT

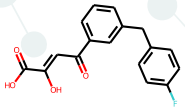
Pharmacophore Clustering



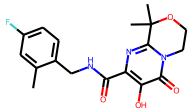
RESULT



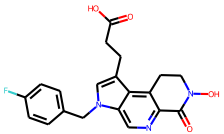
CHEMBL1914556



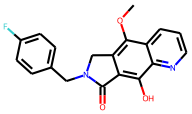
CHEMBL584360



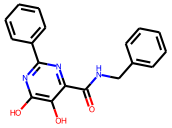
CHEMBL4126686



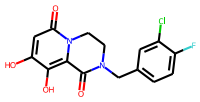
CHEMBL1773405



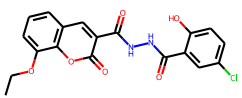
CHEMBL209440



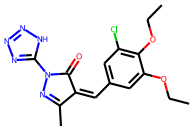
CHEMBL385951



CHEMBL237727

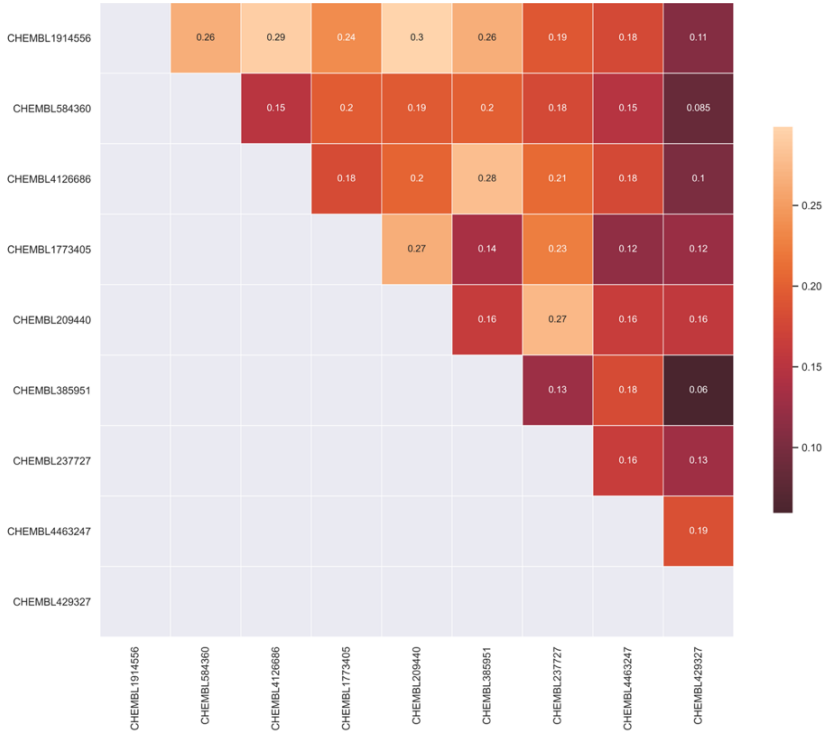


CHEMBL4463247



CHEMBL429327

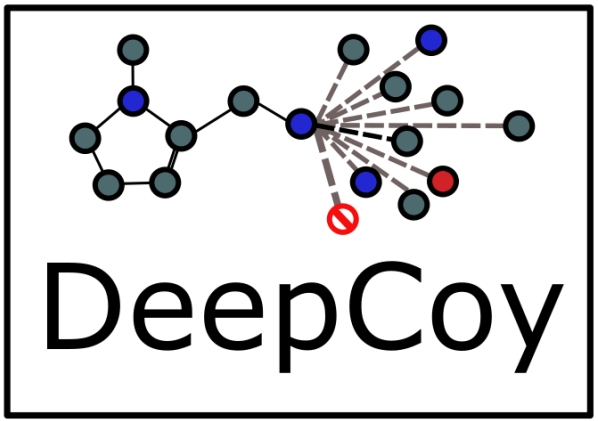
Pharmacophore



RESULT

Pharmacophore

Decoy



Metric	Values
AUC-ROC 1NN	0,533
AUC-ROC RF	0,702
DOE score	0,065
Doppelganger score mean	0,271
Doppelganger score max	0,421

RESULT

Pharmacophore

Local search

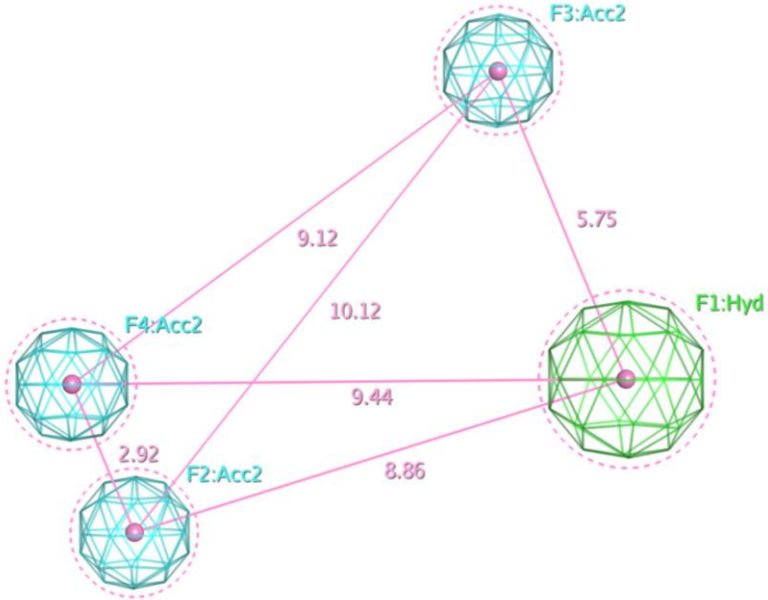
Model	S_e	S_p	P_r	AUC	GH
Haaa_7	0,814	0,573	0,037	0,794	0,133
Haaa_3	0,765	0,482	0,029	0,632	0,103
Haaa_4	0,8	0,6	0,039	0,748	0,138
Haaa_6	0,403	0,631	0,021	0,489	0,074
….					

RESULT

Pharmacophore

Local search

Haaa_7 model



RESULT

Pharmacophore

Global Search

12 models have AUC > 0,7

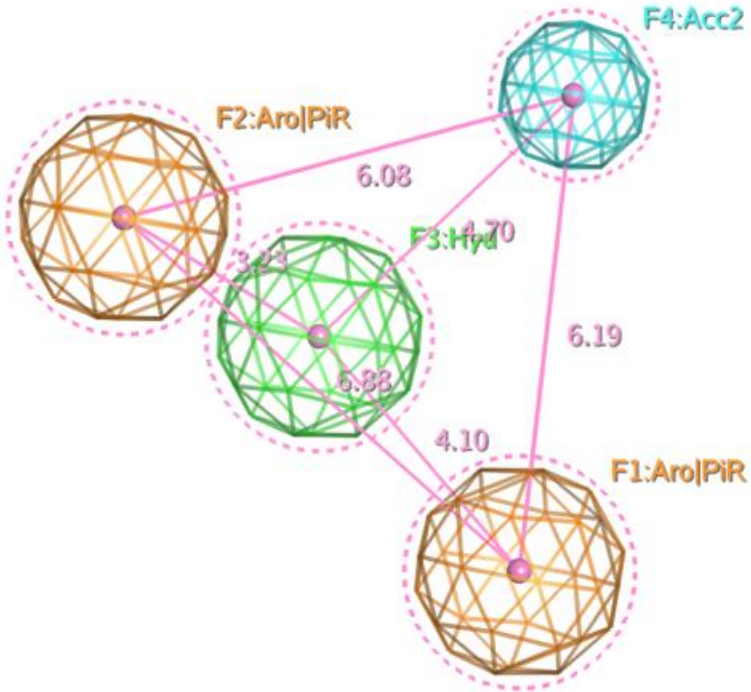
Model	S_e	S_p	P_r	AUC	GH
RRHa_4	0,834	0,894	0,137	0,87	0,278
Haaa_7	0,814	0,573	0,037	0,794	0,133
RRHa_1	0,761	0,812	0,075	0,791	0,200
RHaa_7	0,846	0,546	0,036	0,775	0,130
….					

RESULT

Pharmacophore

Global search

RRHa_4 model



- **Cover 7**
- F1 and F2 – Aromatic (Aro: r= 1,4Å)
- F3 is hydrophobic (Hyd:r=1,4Å)
- F4 hydro acceptor (Acc2: r= 1,0Å)

RESULT

Pharmacophore Voting

Local Voting

17 model with cover 9 vote

12 model (AUC > 0,7) vote

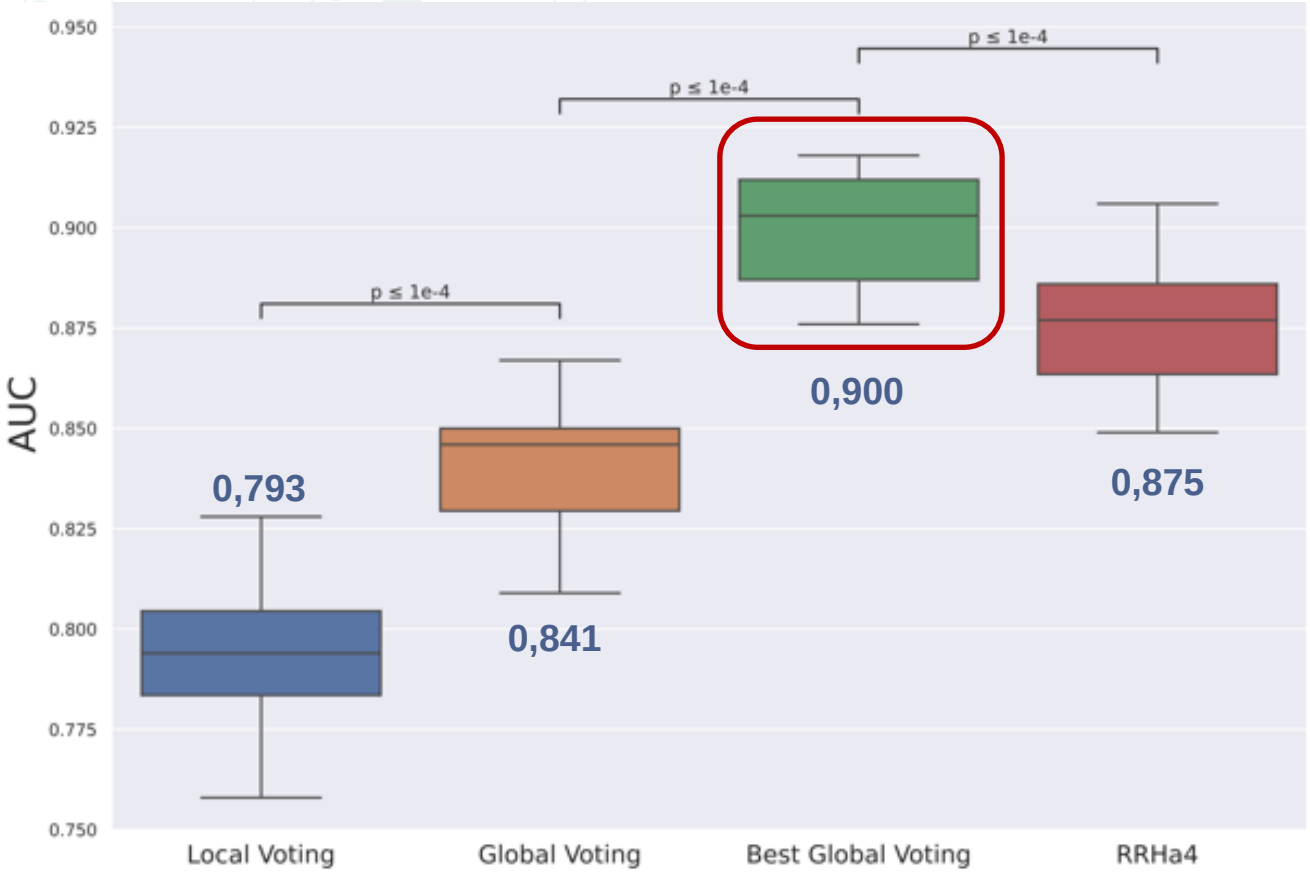
Best Global Voting

Global Voting

80 model vote

RESULT

Pharmacophore Voting



Best global voting:

- AUC = 0,900 ± 0,015
- GH = 0,334 ± 0,049

RESULT

Pharmacophore

Stacking

01

Base model

BGV
Best global voting

02

Local Stacking

17 model cover 9

Stacking

Pharmacophore output is input of meta-model

03

Best Global Stacking

12 model
AUC > 0,7

04

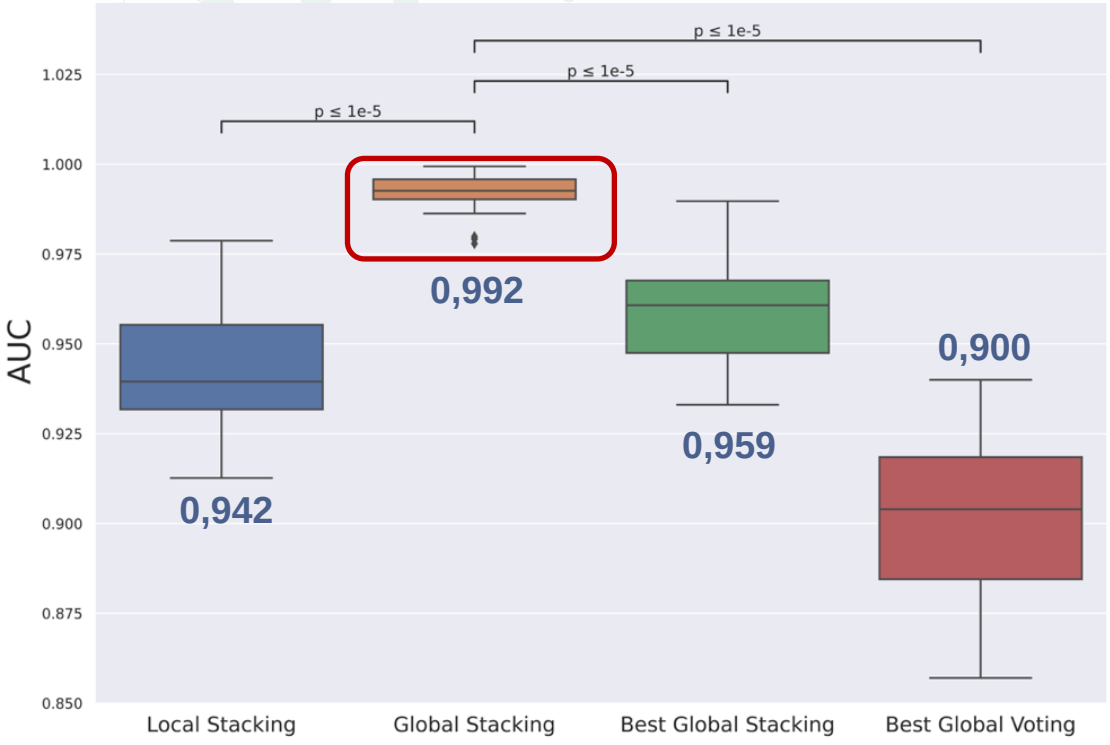
Global Stacking

80 model

RESULT

Pharmacophore

Stacking

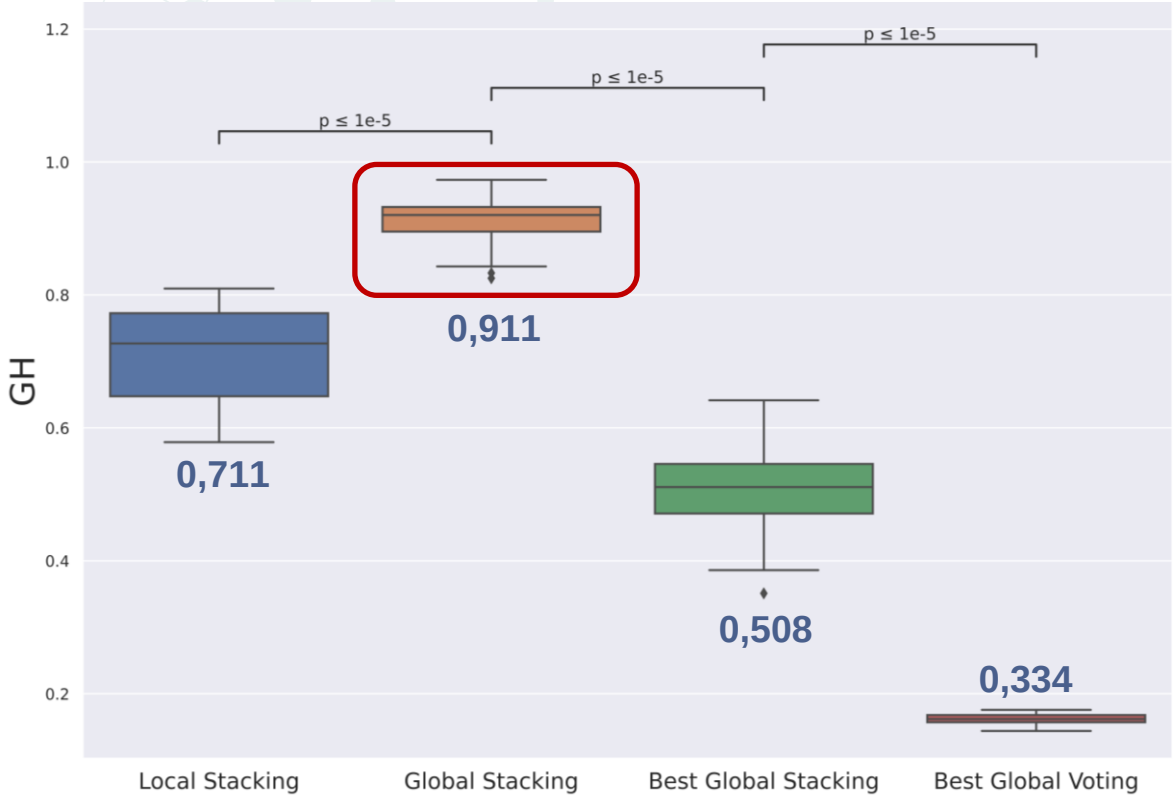


Global stacking:

- AUC = 0,992 ± 0,017
- GH = 0,911 ± 0,035

RESULT

Pharmacophore Stacking



Global stacking:

- AUC = 0,992 ± 0,017
- GH = 0,911 ± 0,035

RESULT

Pharmacophore

Score Stacking

01

Ehance data

GSS model
Using rescore from pharmacophore model

02

Feature selection

F-GSS Model
GSS feature selection
XGBoost: n_estimators = 300 max_depth = 4

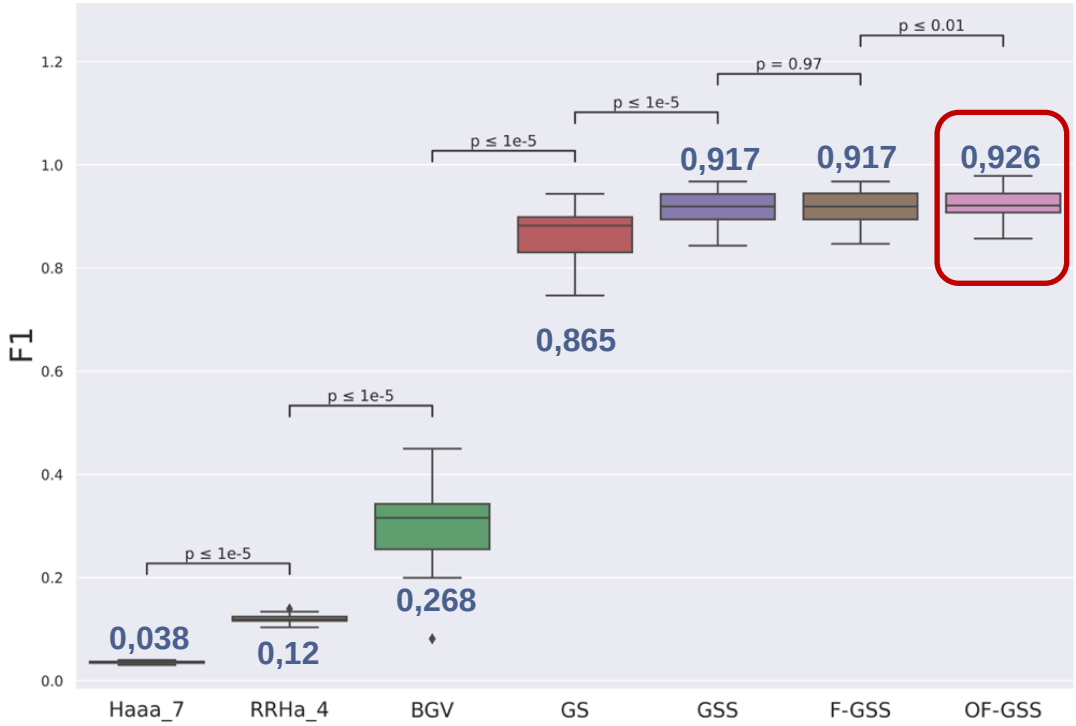
03

Model optimization

OF-GSS Model
Optimize F-GSS

Pharmacophore

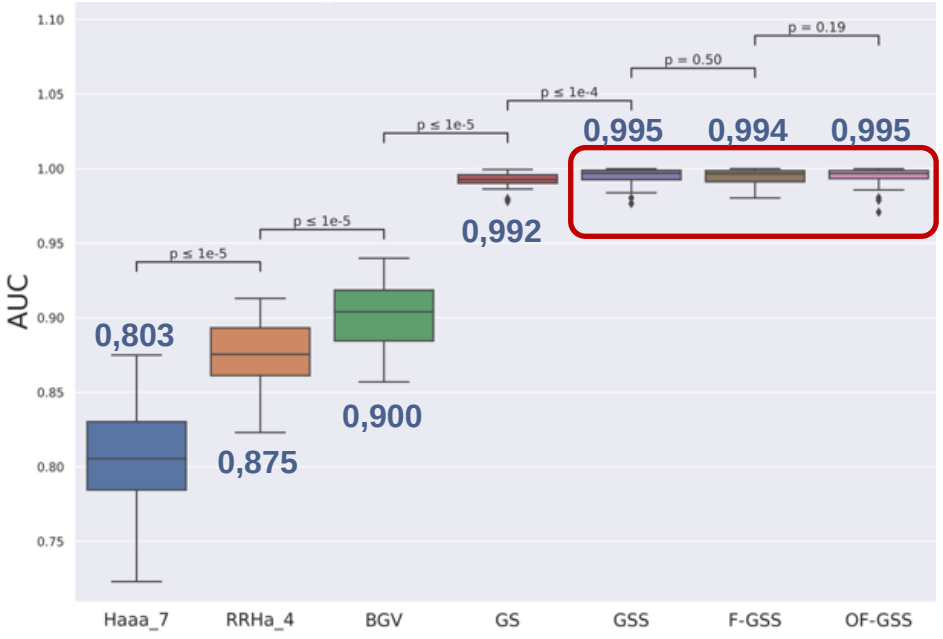
Performance comparison

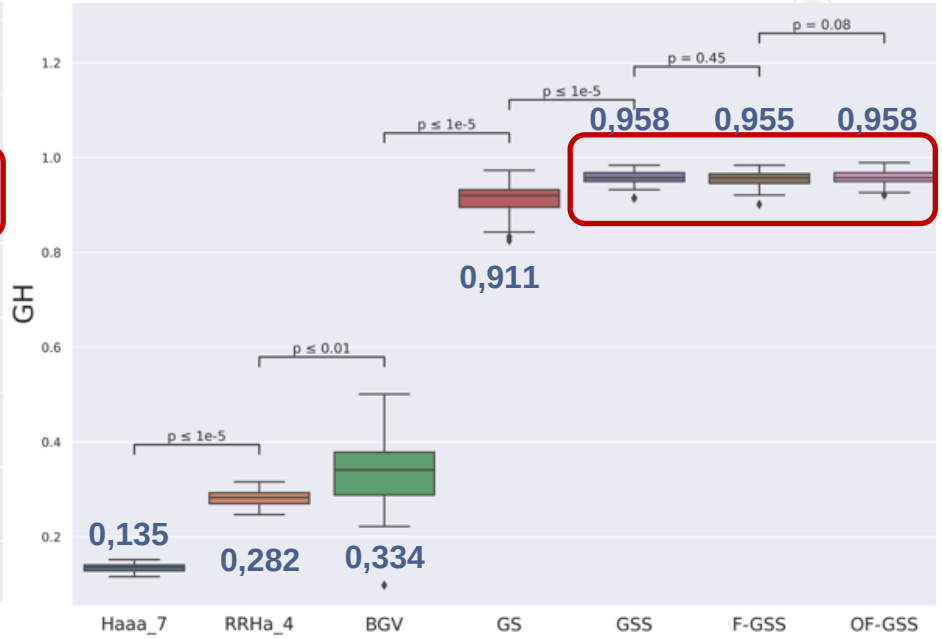


1. Local search: **Haaa_7**
2. Global search: **Rhaa**
3. **BGV**: 12 model with AUC > 0,7
4. **GS**: stacking 80 model
5. **GSS**: using rescore
6. **F-GSS:** feature selection of GSS
7. **OF-GSS**: optimization of F-GSS

Pharmacophore

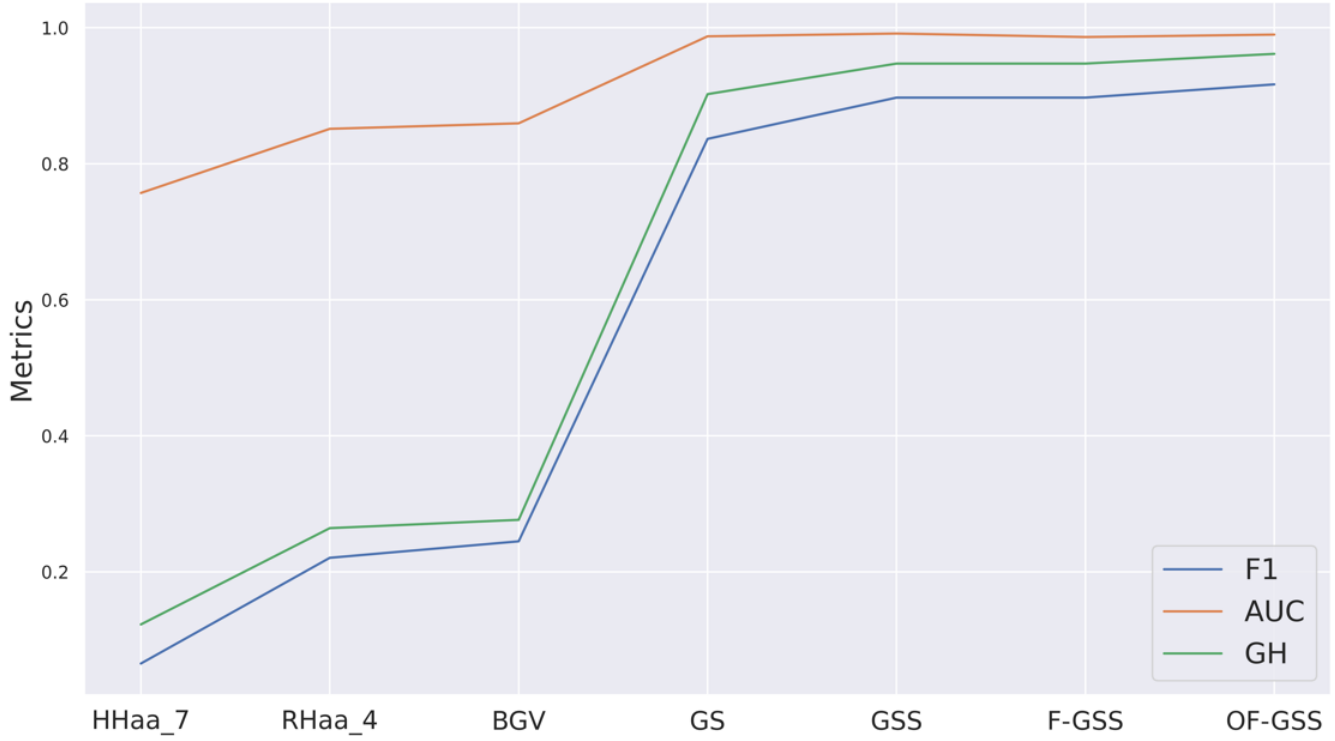
Performance comparison





Pharmacophore

External Validation



OF-GSS

- F1 = 0,917
- AUC = 0,990
- GH = 0,962

RESULT

2

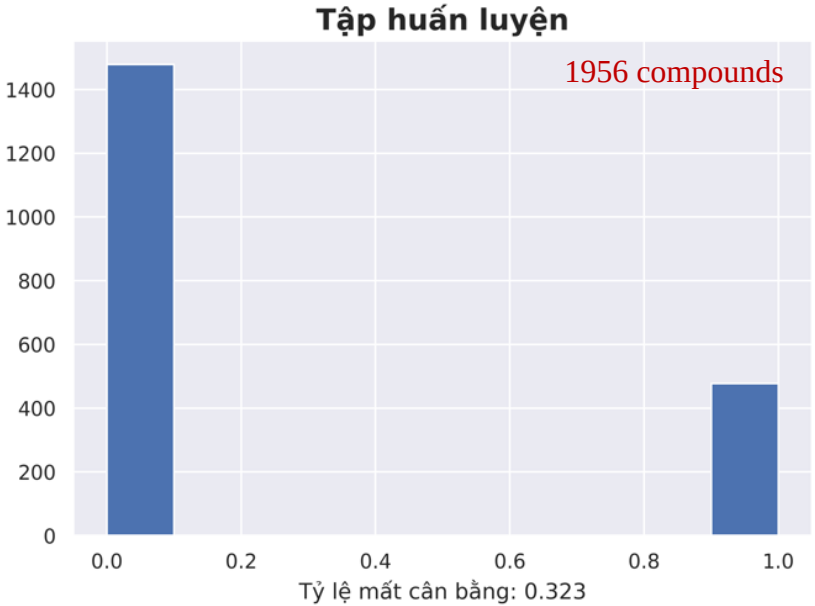
QSAR

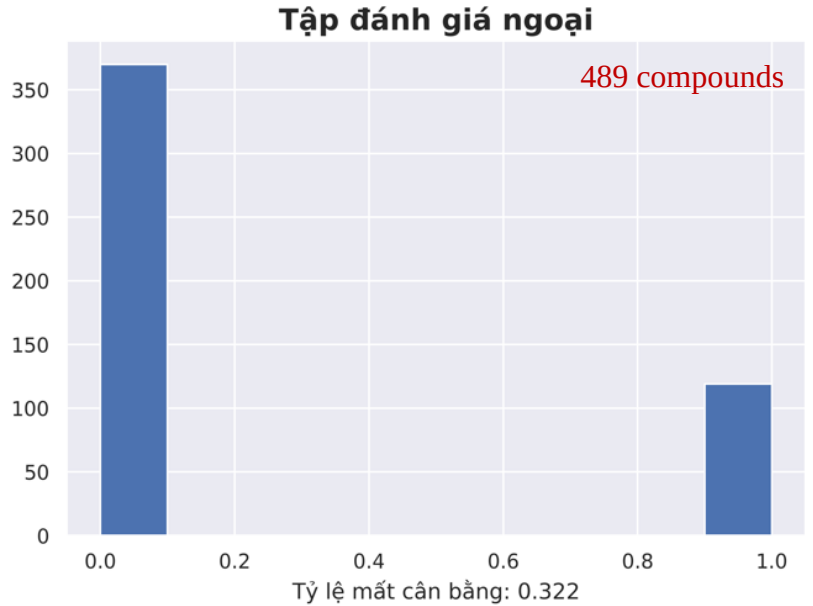
2.1. Mô hình học máy phân loại

RESULT

QSAR – Classification

Data preparation

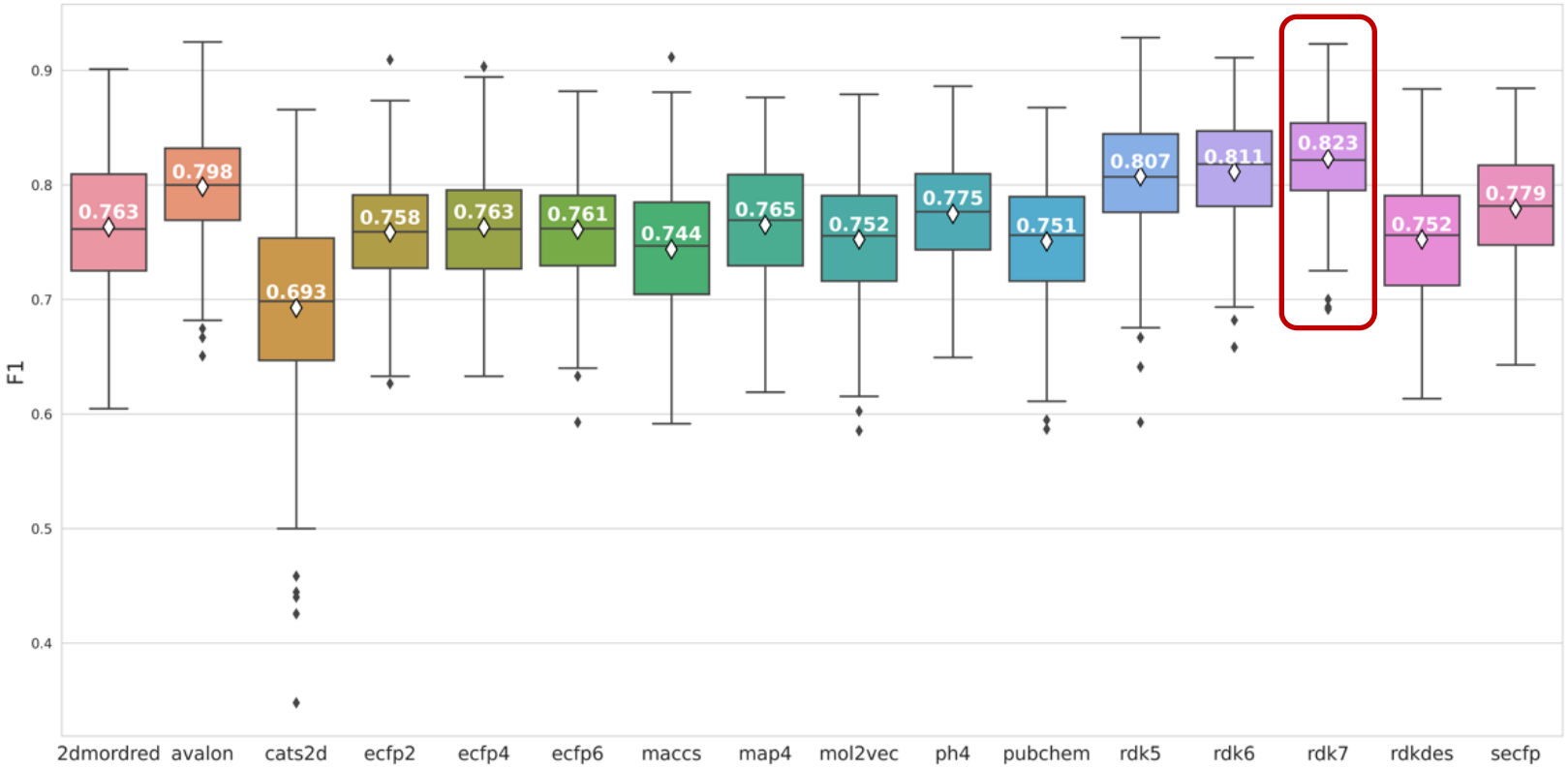




RESULT

QSAR – Classification

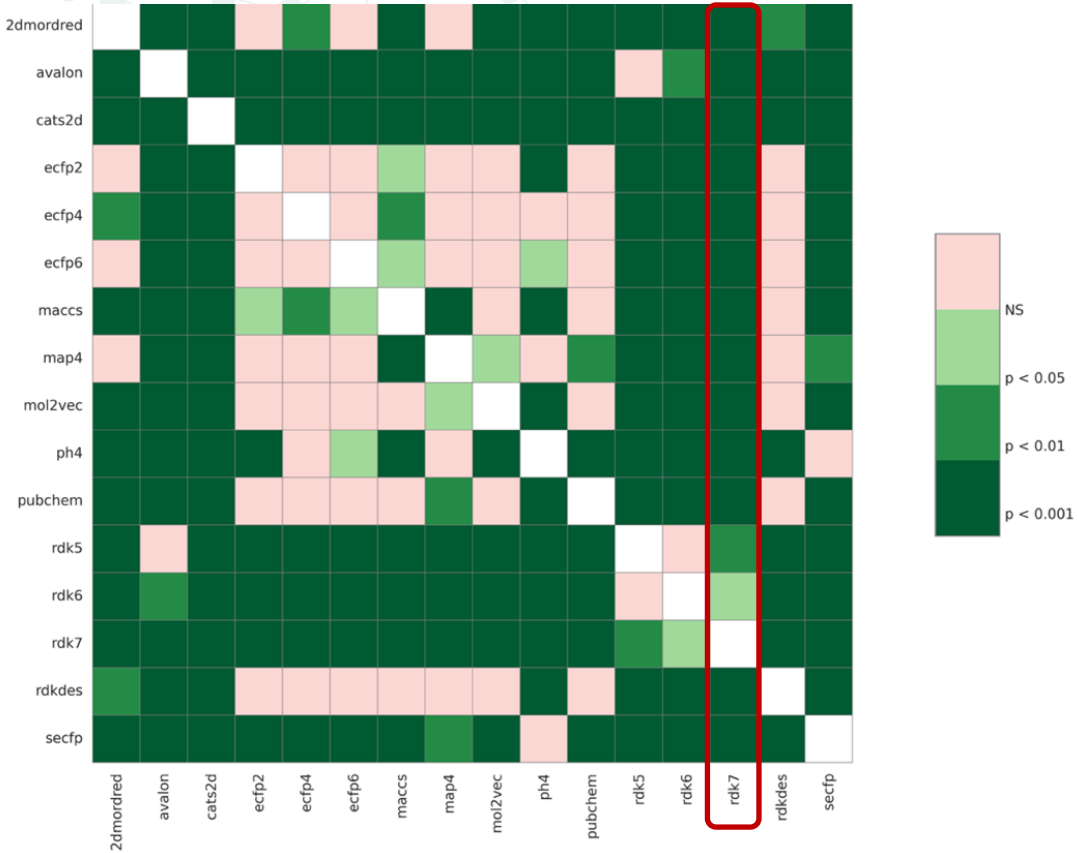
Feature selection



RESULT

QSAR – Classification

Data selection



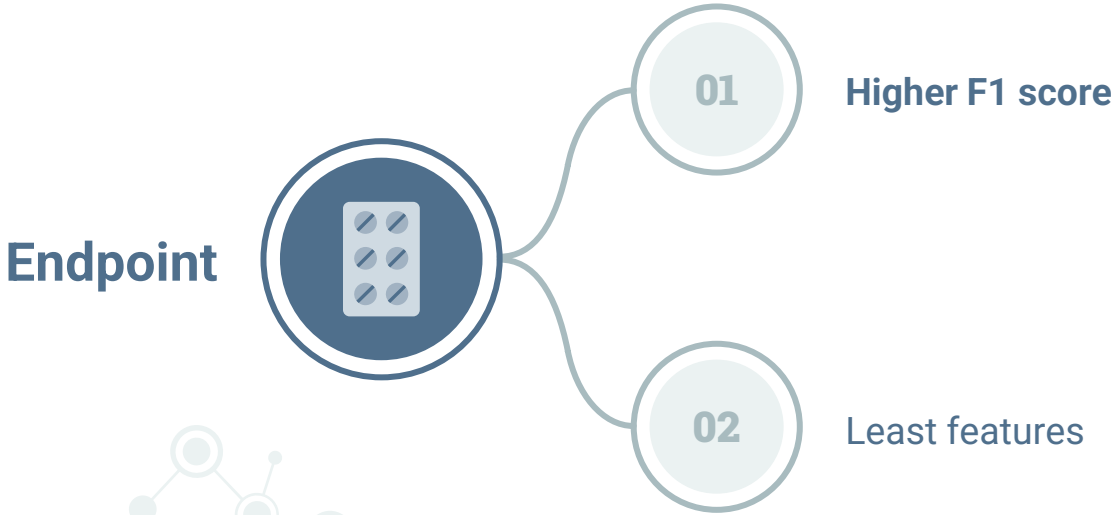
Primary endpoint

Highest F1 score

RESULT

QSAR – Classification

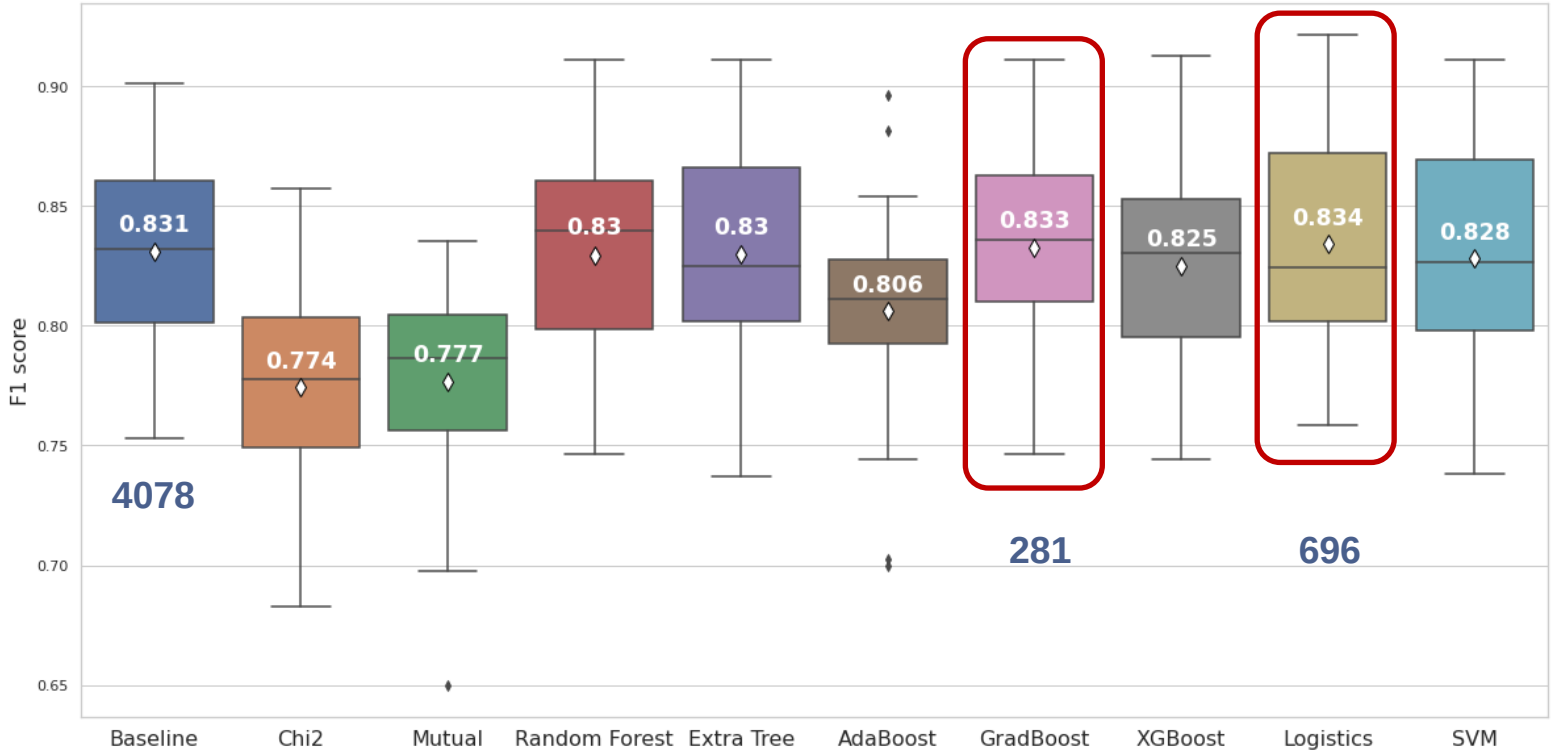
Feature selection



RESULT

QSAR – Classification

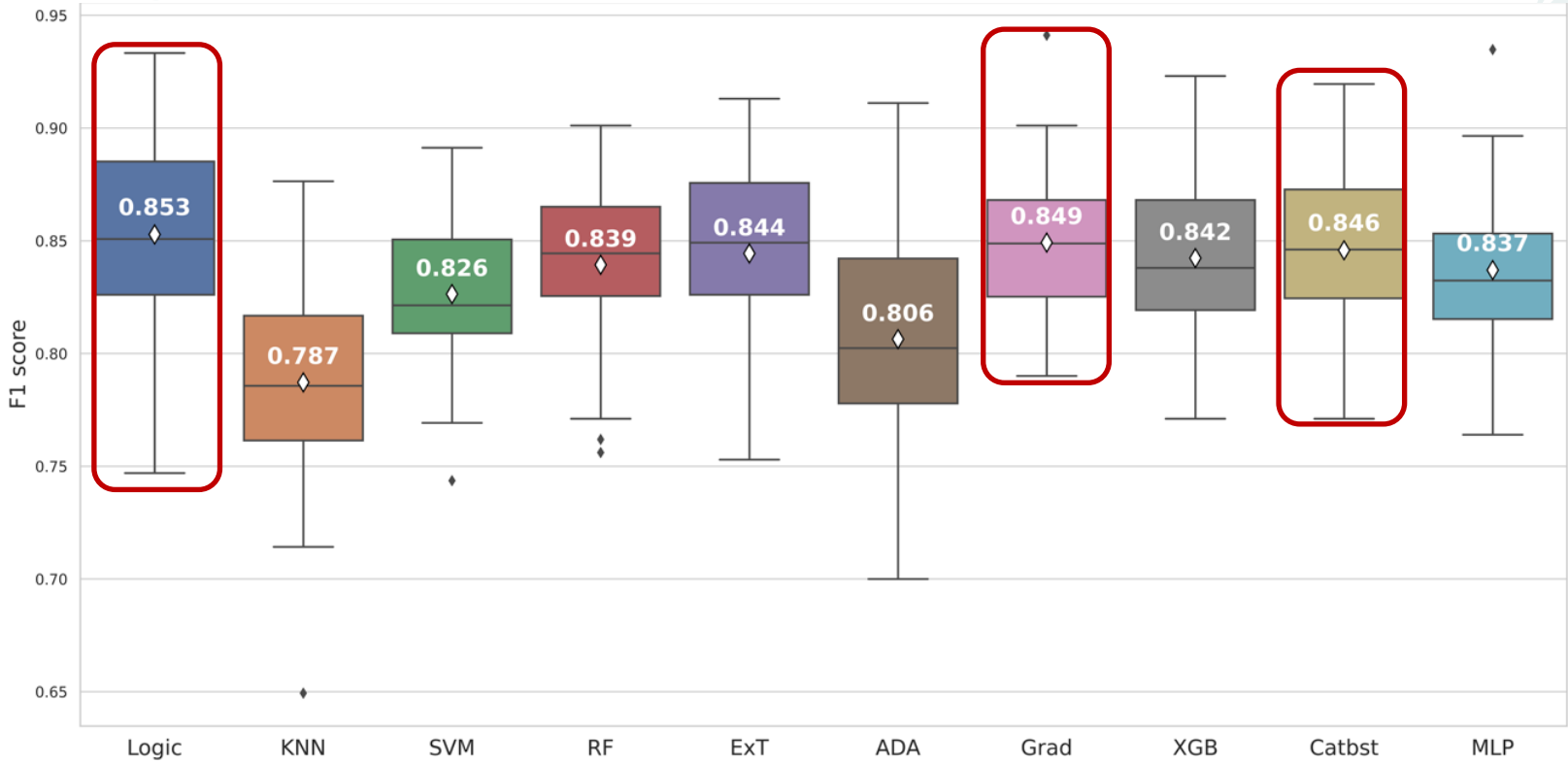
Feature selection



RESULT

QSAR – Classification

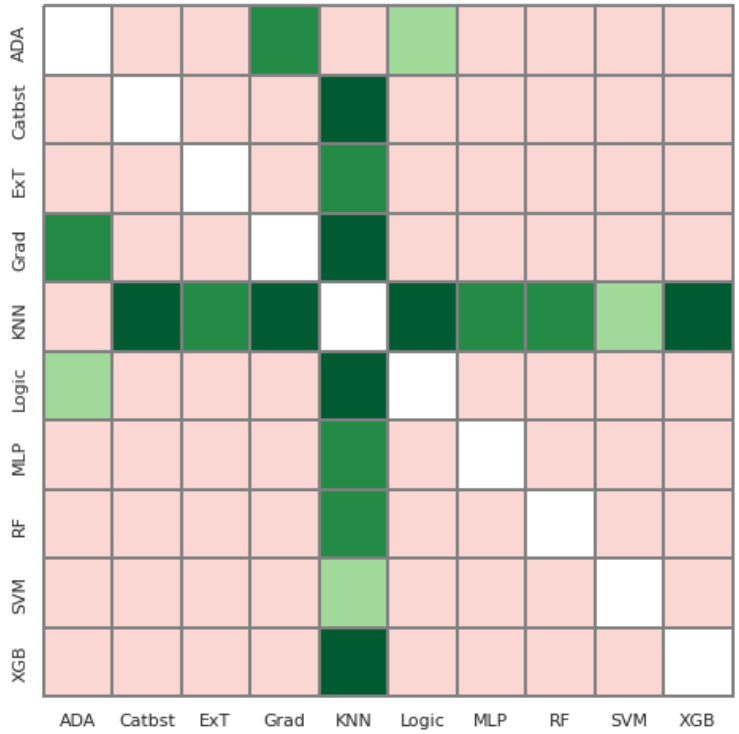
Model selection

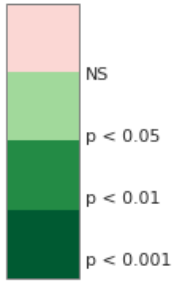


RESULT

QSAR – Classification

Model selection





0

Optimized Model

- Logistic regression
- **Gradient Boosting**
- CatBoost

RESULT

QSAR – Classification

Model Optimization

Data sampling

Over-sampling, under-sampling

Hyperparameter Tuning

Grid search,

Probability

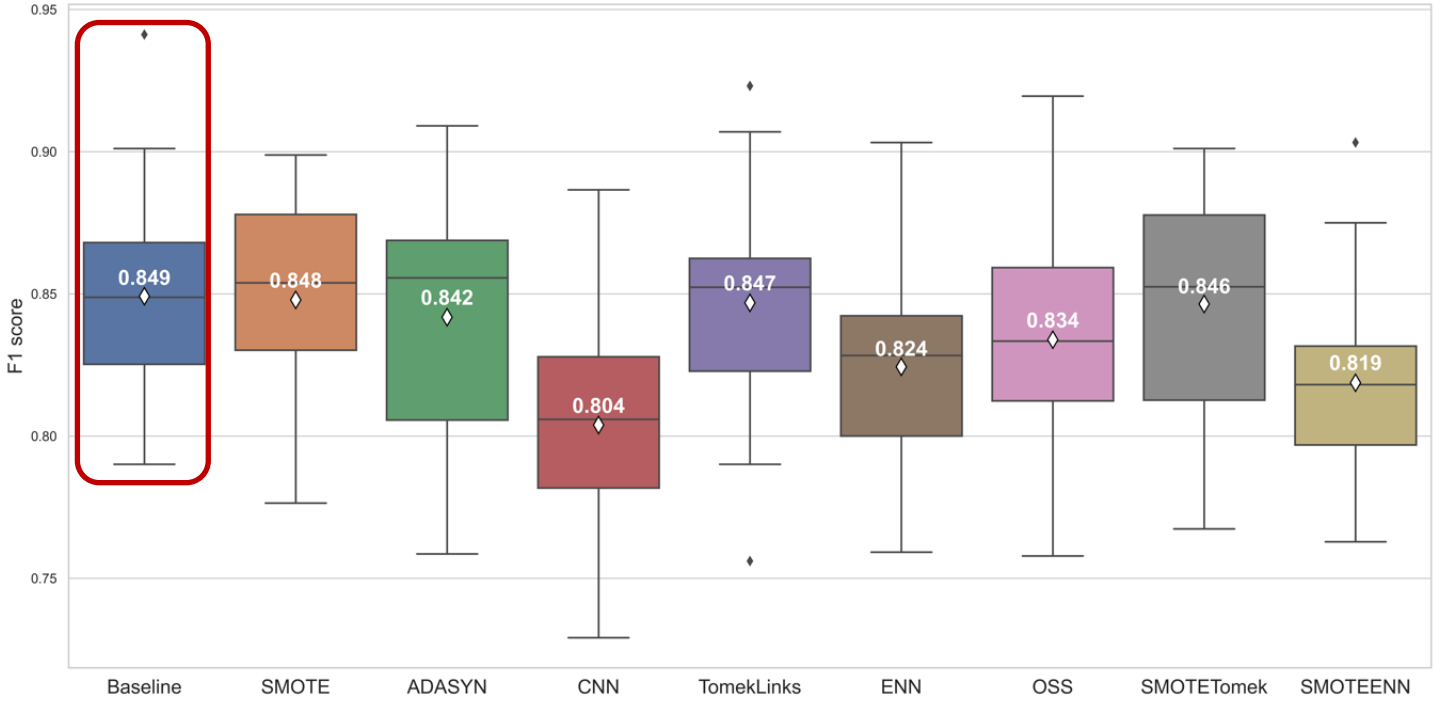
Sigmoid (Platt Scalling) or isotonic (Isotonic Regression).

RESULT

QSAR – Classification

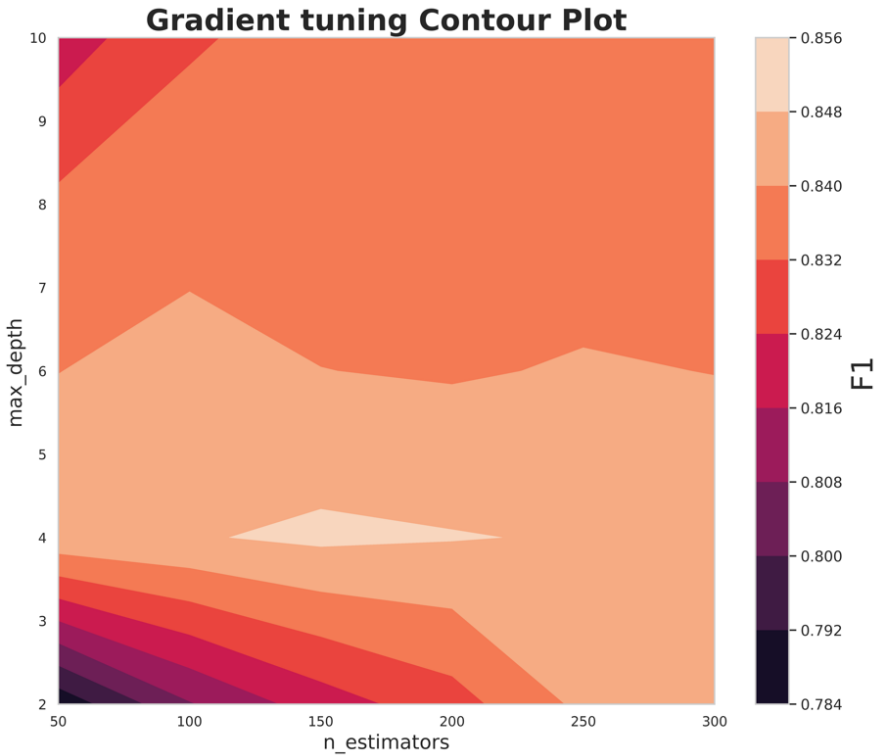
Model Optimization

Data sampling



RESULT

Tuning



QSAR – Classification

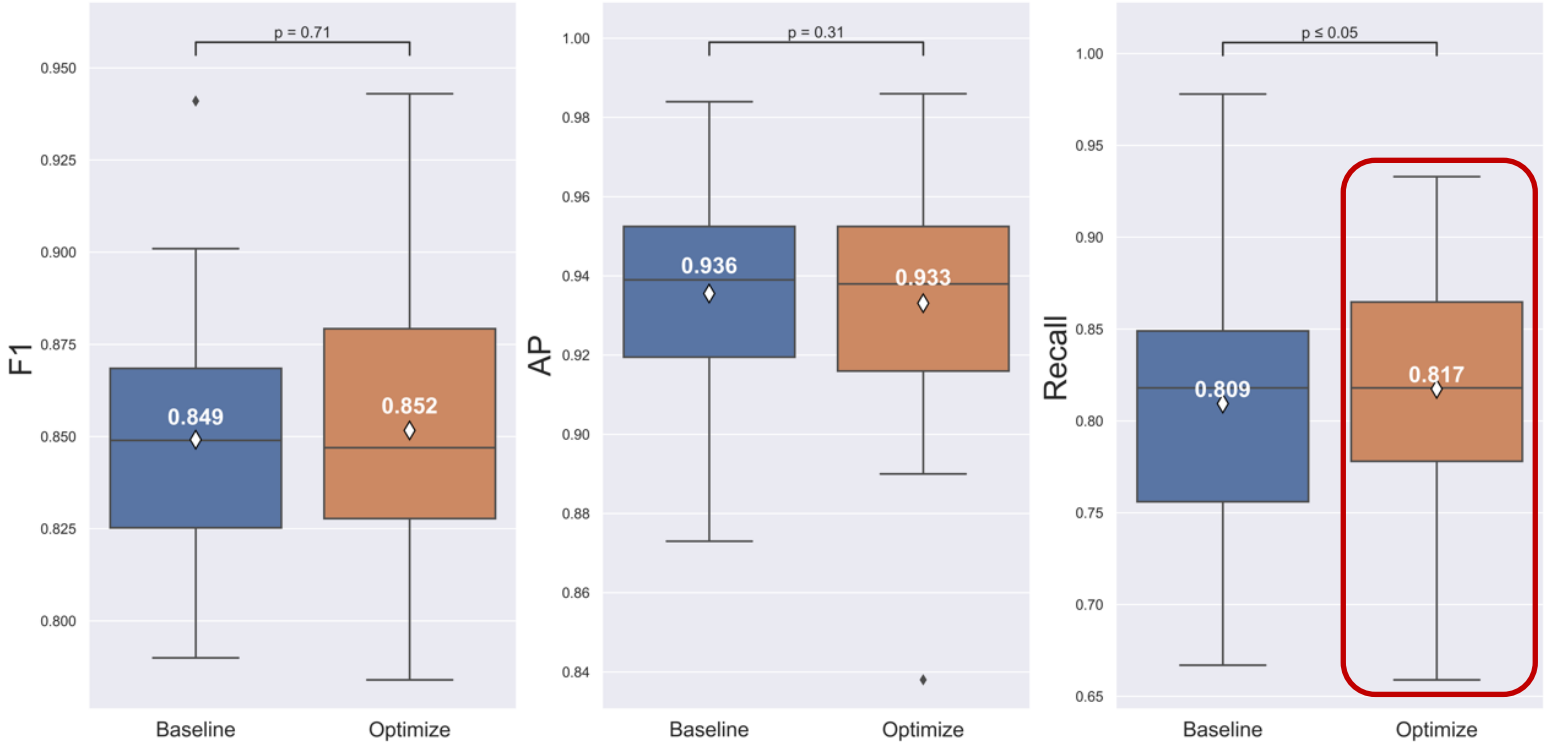
Model Optimization

n_estimators	150
max_depth	4
l2_leaf_reg	0,1
random_state	42

RESULT

QSAR – Classification

Model Optimization



RESULT

QSAR – Classification

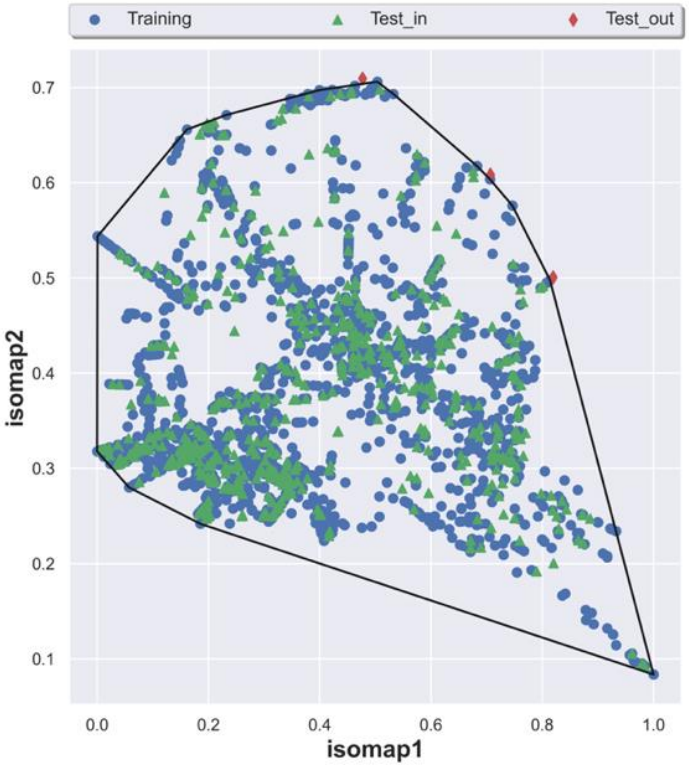
External Validation

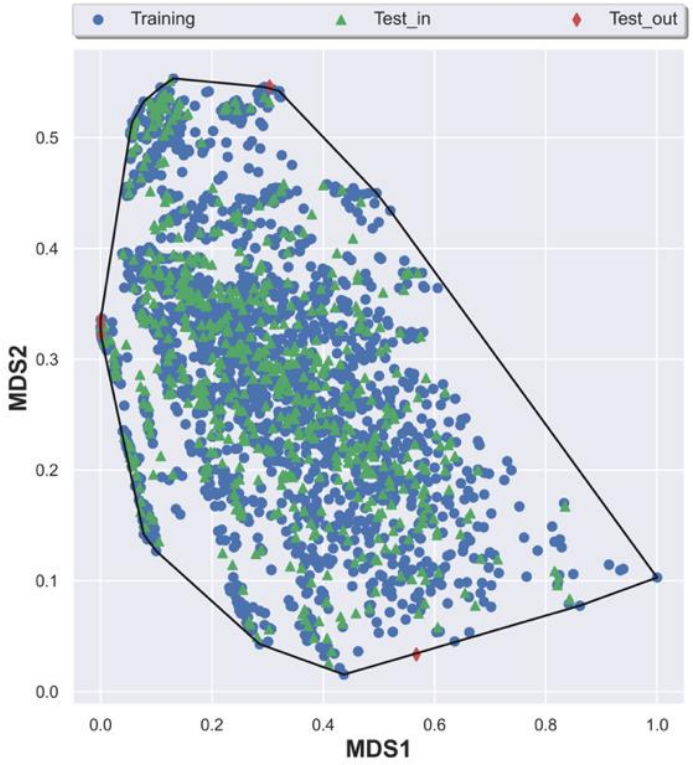
	Internal Validation			External Validation		
	AP	F1	Độ nhạy	AP	F1	Độ nhạy
Baseline	0,936	0,849	0,809	0,928	0,873	0,864
Optimize	0,933	0,852	**0,817**	**0,938**	**0,874**	**0,865**

QSAR – Classification

Applicability domain

Convex hull





RESULT

QSAR – Classification

Applicability domain

Isomap	MDS
	CHEMBL3310415
CHEMBL394650	CHEMBL2236598
CHEMBL252831	CHEMBL3310409
CHEMBL19043	CHEMBL1914564
	CHEMBL252757

RESULT

3

Docking

Re-docking

Retrospective control

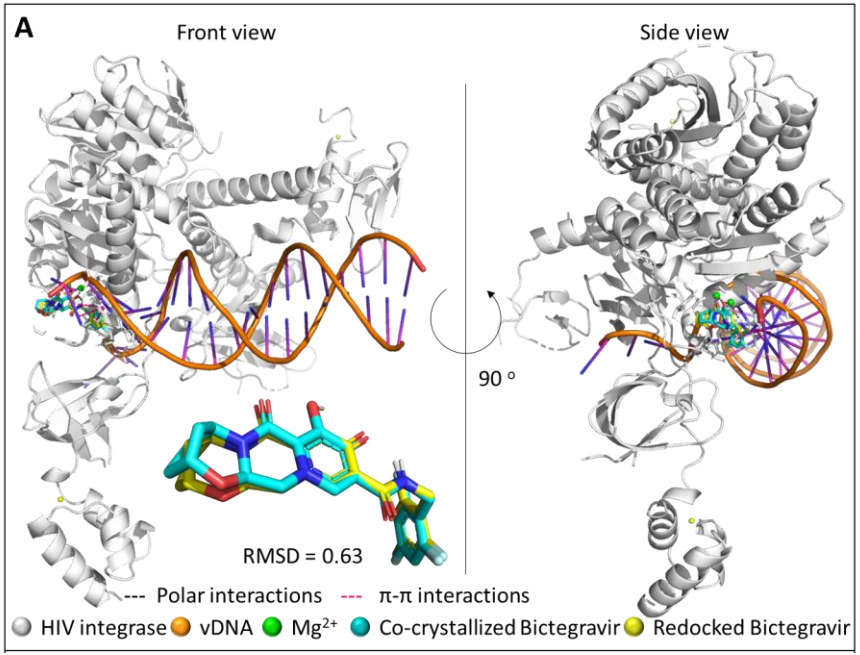
Software selection

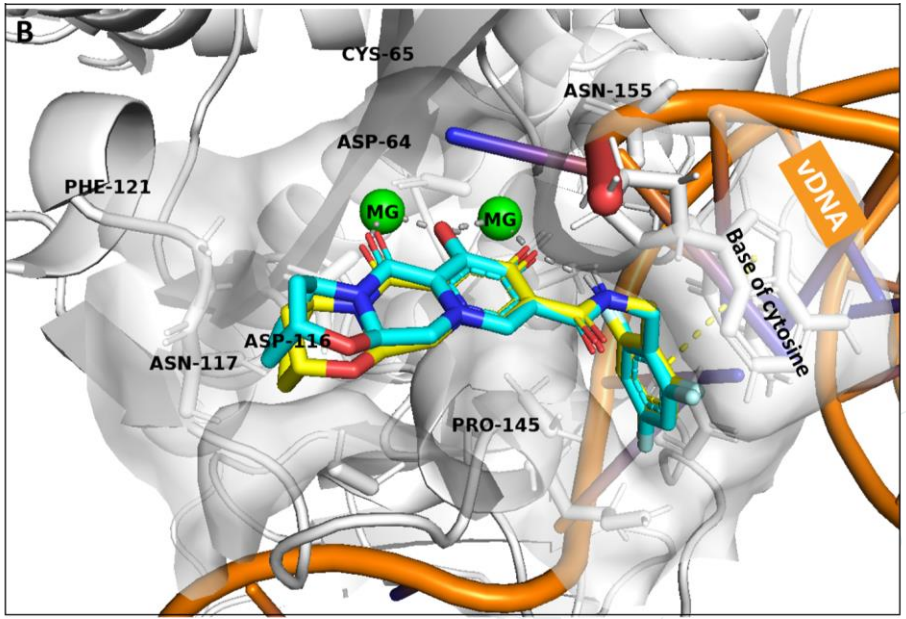
RESULT

MOLECULAR DOCKING

Re-docking

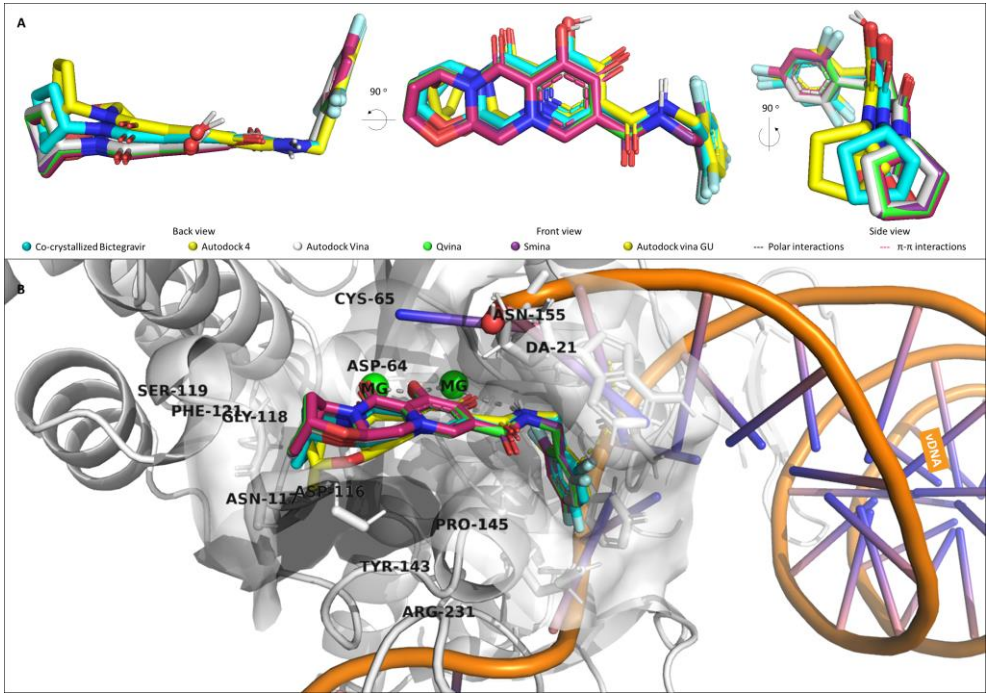
Autodock-GPU





MOLECULAR DOCKING

Re-docking



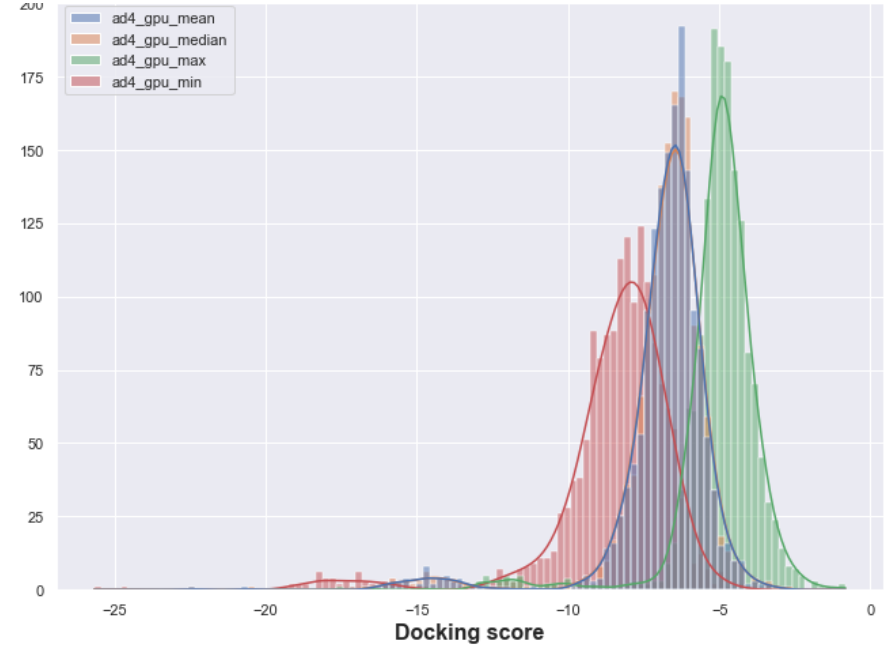
Softwares	RMSD (Å)	Best docking score(kcal/mol)
Smina	0,677	-10,10
Qvina2	0,716	-9,30
Autodock Vina 1.2.3	0,636	-9,86
Vina-GPU	0,792	-9,80
Autodock-GPU	0,630	-11,11

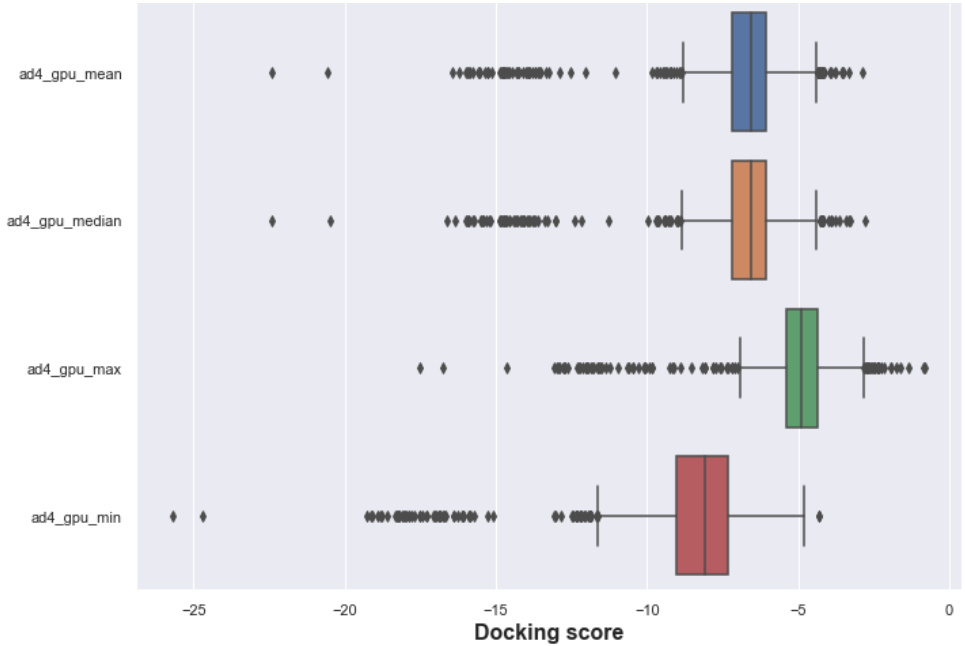
RESULT

MOLECULAR DOCKING

Autodock-GPU

Retrospective control



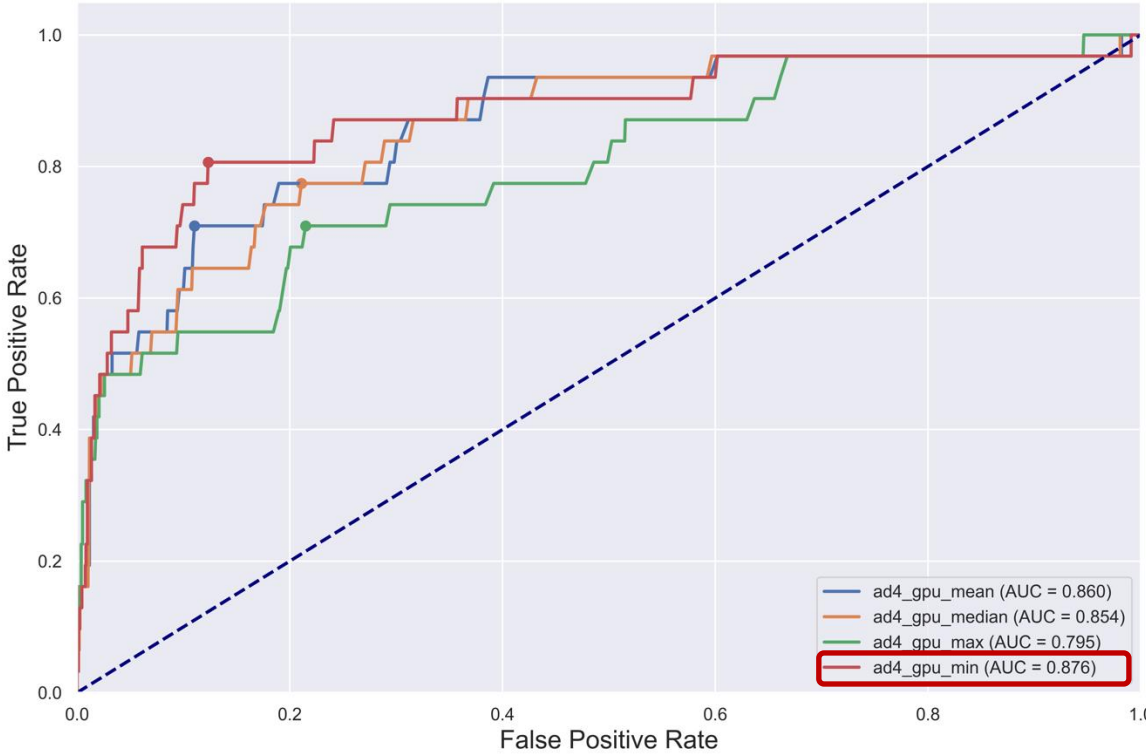


RESULT

MOLECULAR DOCKING

Autodock-GPU

Retrospective control



Autodock-GPU_min:

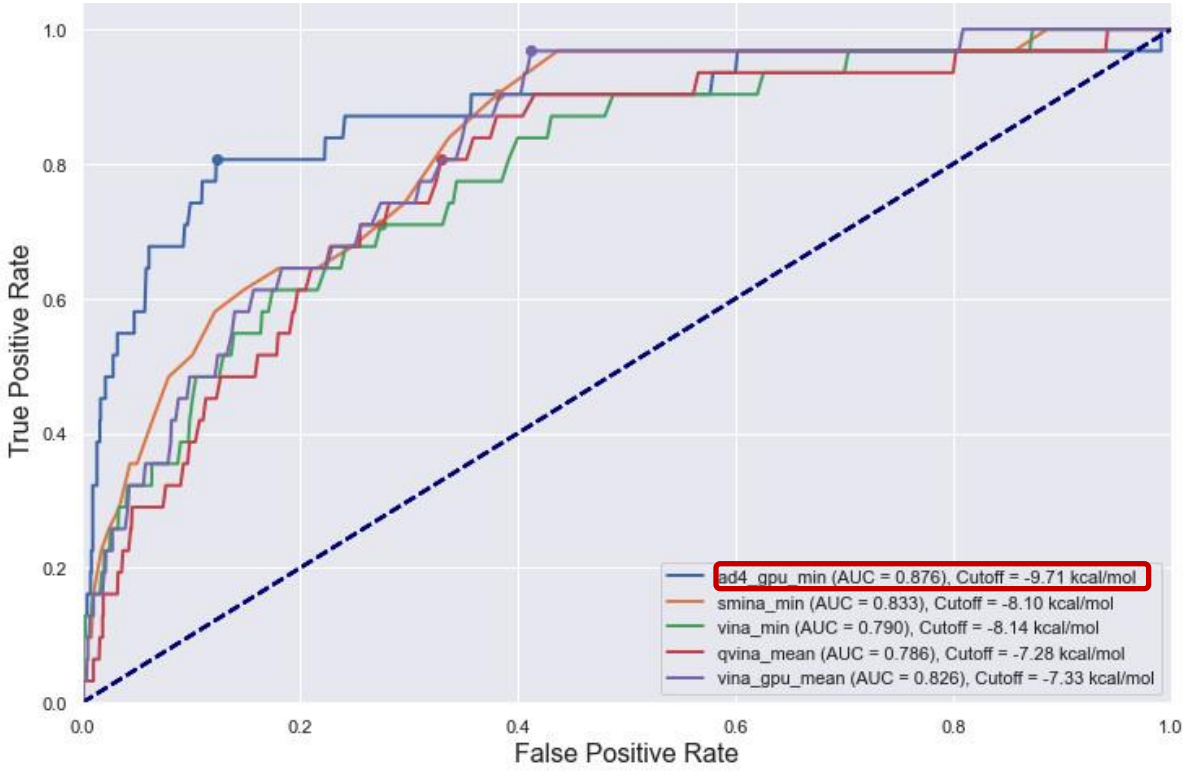
- AUC = 0,876
- G-mean max = 0,841
- TPR = 0,806
- FPR = 0,123
- cutoff = -9,71 kcal/mol

RESULT

MOLECULAR DOCKING

Autodock-GPU

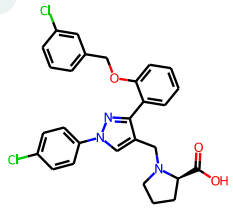
Retrospective control



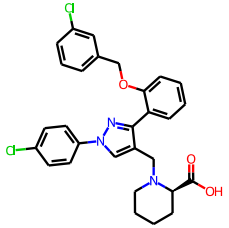
4 Screening

RESULT

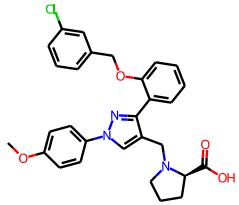
VIRTUAL SCREENING



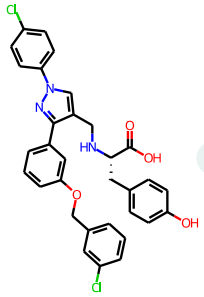
4.3.Pro
pChEMBL = 7,00



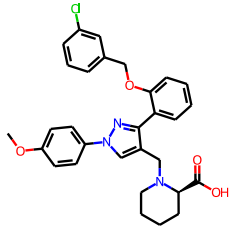
4.3.Pipe
pChEMBL = 6,92



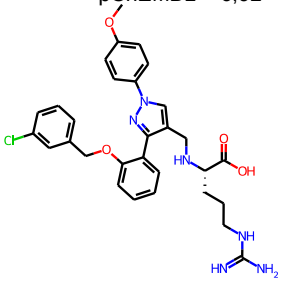
6.3.Pro
pChEMBL = 7,13



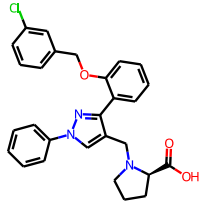
4.3.Tyr
pChEMBL = 7,06



6.3.Pipe
pChEMBL = 7,08



6.3.Arg
pChEMBL = 7,40

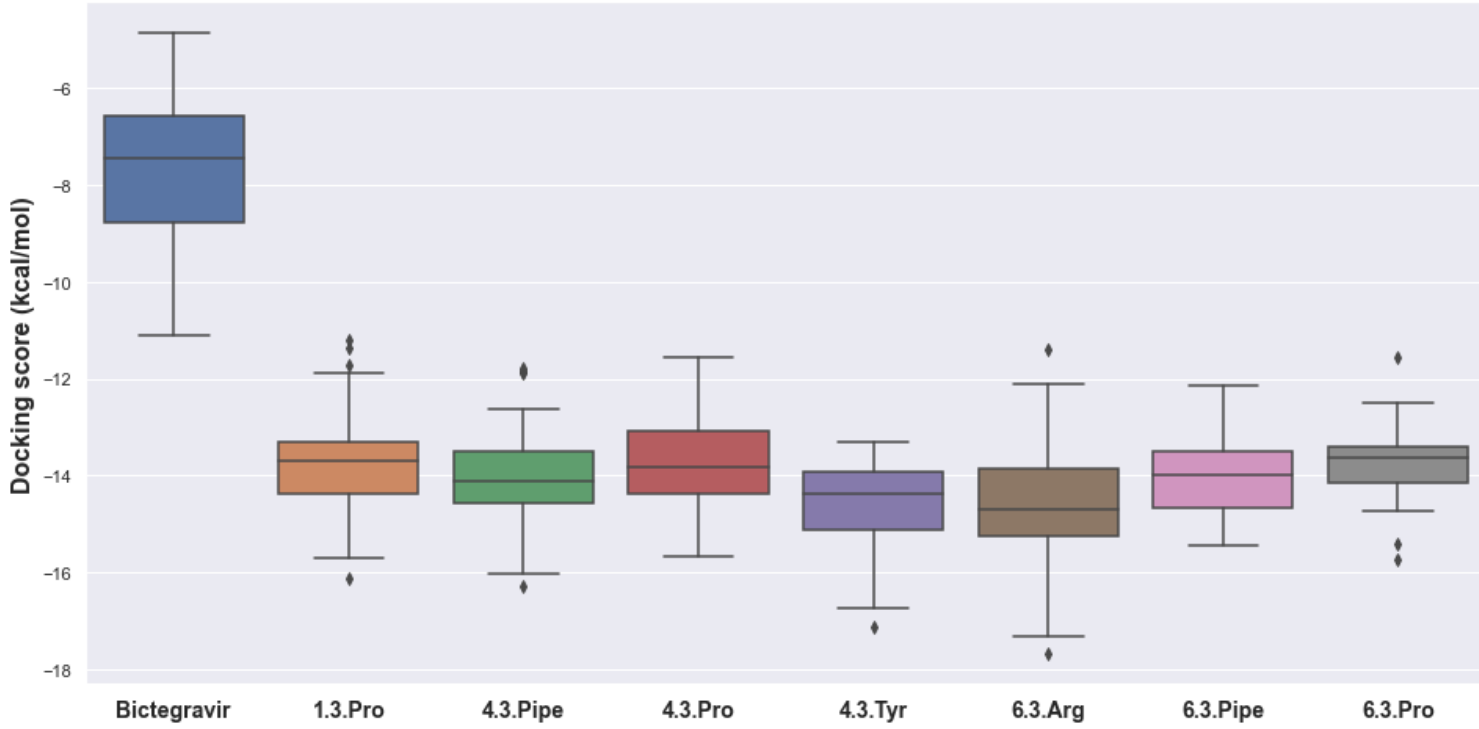


1.3.Pro
pChEMBL = 7,16

RESULT

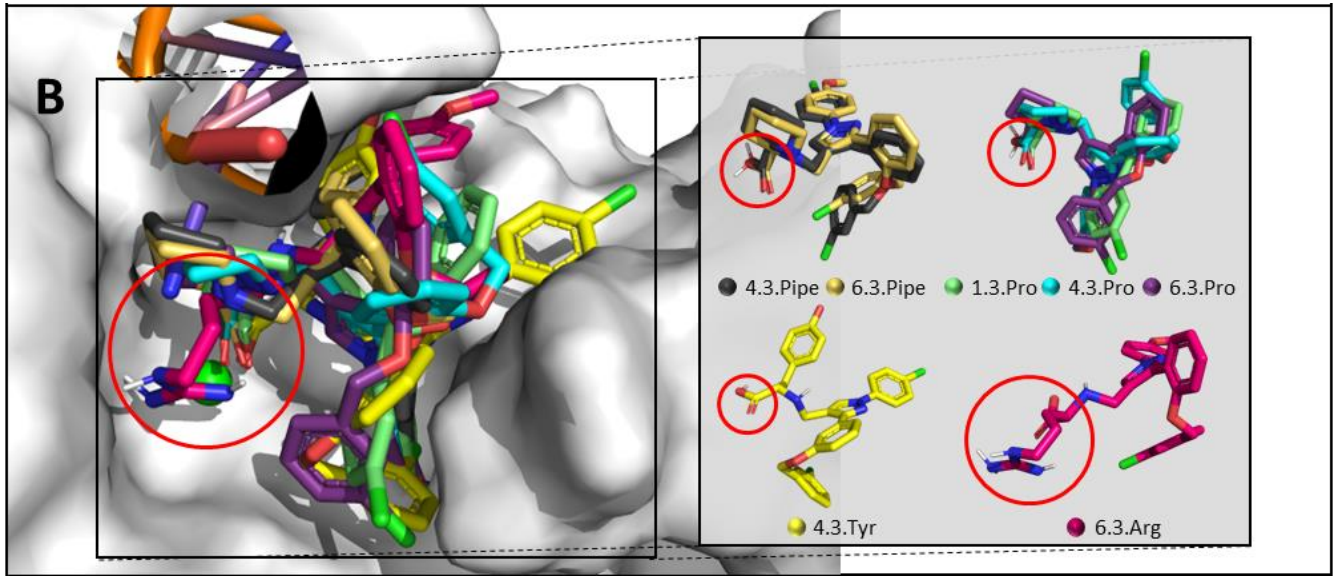
VIRTUAL SCREENING

Molecular Docking



VIRTUAL SCREENING

Molecular Docking



Group 1: 4.3.Pipe, 6.3.Pipe

Group 2: 1.3.Pro, 4.3.Pro, 6.3.Pro

Group 3: 4.3.Tyr

Group 4: 6.3.Arg

VIRTUAL SCREENING

Group 4

Molecular Docking