Sviluppo di modelli QSAR predittivi mediante tecniche di Machine Learning: applicazione sugli inibitori di HDAC4

Tesi Sperimentale in Chimica Farmaceutica

FACOLTÀ DI FARMACIA E MEDICINA

DIPARTIMENTO DI CHIMICA E TECNOLOGIA DEL FARMACO

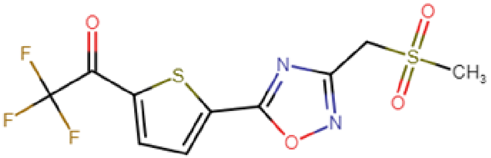
CORSO DI LAUREA MAGISTRALE IN
CHIMICA E TECNOLOGIA FARMACEUTICHE

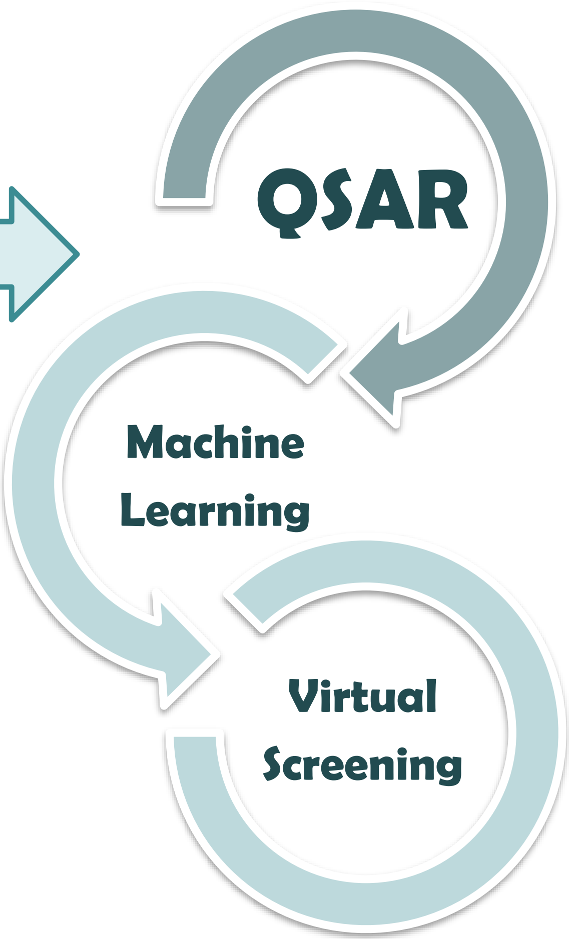
LAUREANDA:
Fantera Giulia

RELATORE:
Prof. Rino Ragno

Obiettivo della ricerca e strumenti impiegati

Apprendimento relazione struttura-attività

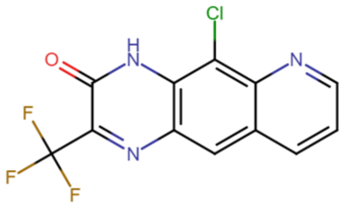




QSAR

Machine Learning

Virtual Screening

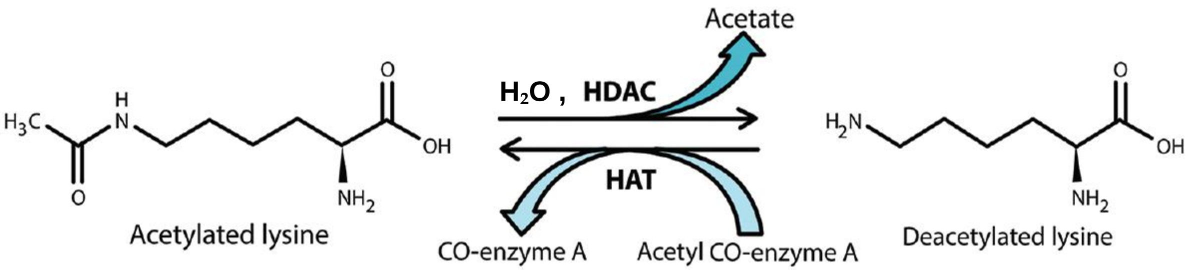


Identificazione potenziali *hit compounds*

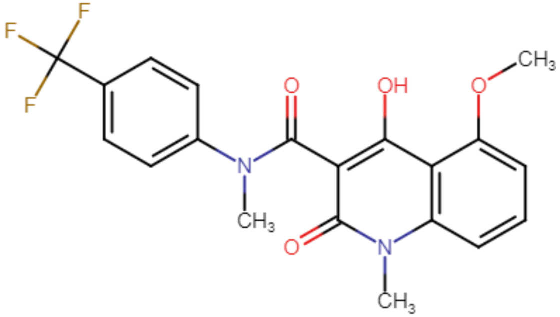
Rome Center for Molecular Design

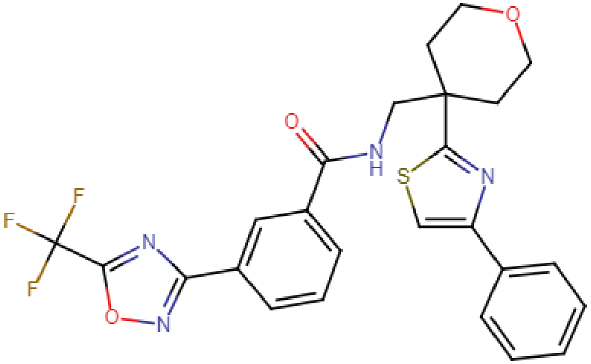
Target biologico di interesse: HDAC4

CLASSE	PROTEINA
I	HDAC1 HDAC2 HDAC3 HDAC8
IIa	HDAC4 HDAC5 HDAC7 HDAC9
IIb	HDAC6 HDAC10
III	SIRTUINE
IV	HDAC11


Acetate
H₂O , HDAC
HAT
H₃C
H N
O
OH
NH₂
Acetylated lysine
CO-enzyme A
Acetyl CO-enzyme A
H₂N
Deacetylated lysine

Inibitori di HDAC4





Tasquinimod

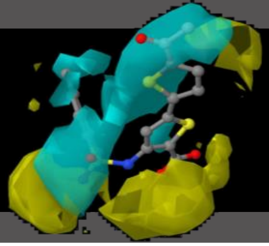
- fase I e III
- antitumorale

TMP-269

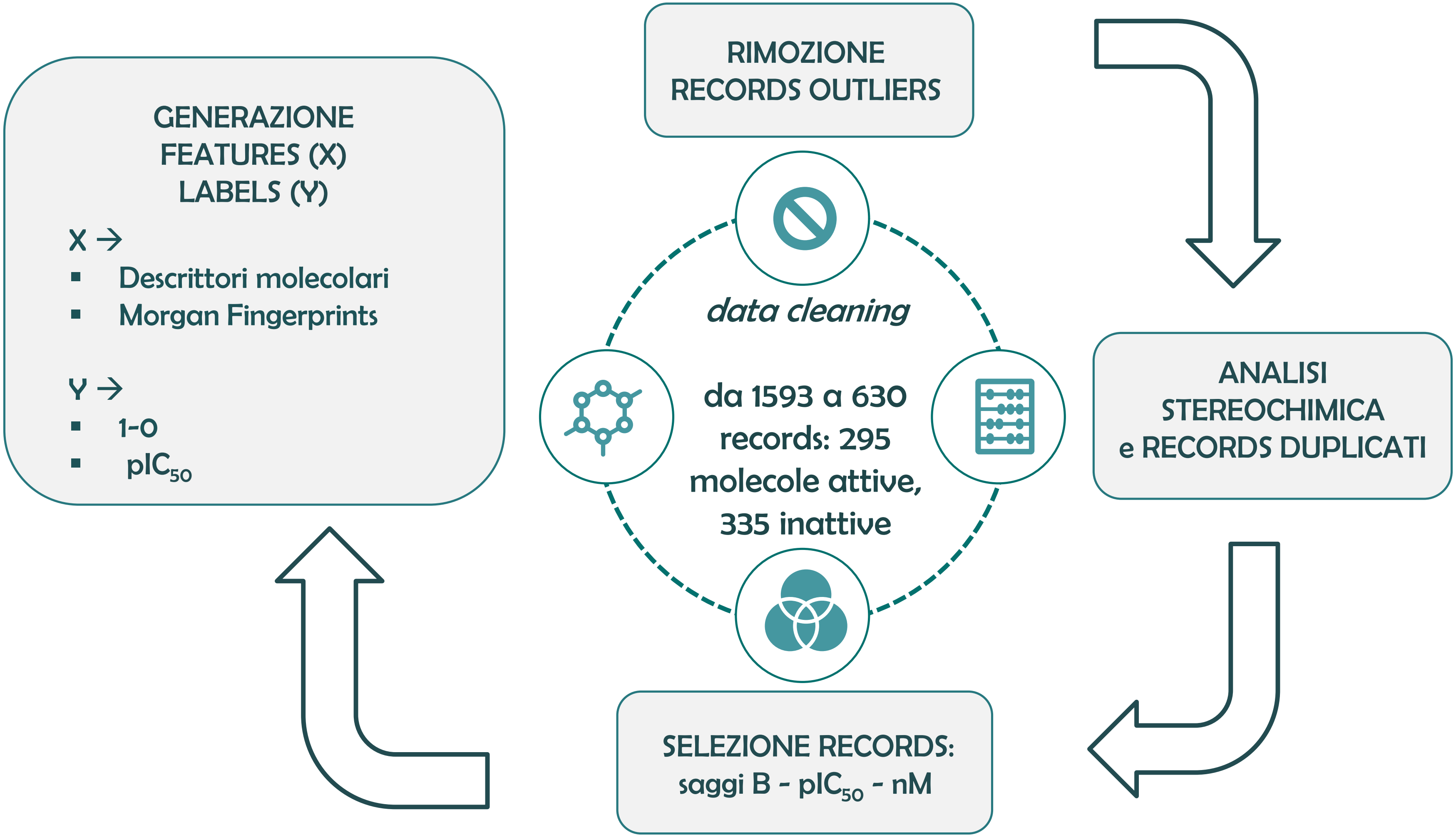
- fase preclinica
- neuroprotezione

Metodologia Sperimentale

- Generazione *training set*
- Costruzione dei modelli di classificazione e regressione
- Validazione esterna con *test set*
- Analisi similarità molecolare *sets*
- Analisi selettività: costruzione dei relativi modelli
- Virtual Screening



Generazione training set



RIMOZIONE RECORDS OUTLIERS

ANALISI STEREOCHIMICA e RECORDS DUPLICATI

SELEZIONE RECORDS: saggi B - pIC_{50} - nM

GENERAZIONE FEATURES (X) LABELS (Y)

X →
- Descrittori molecolari
- Morgan Fingerprints

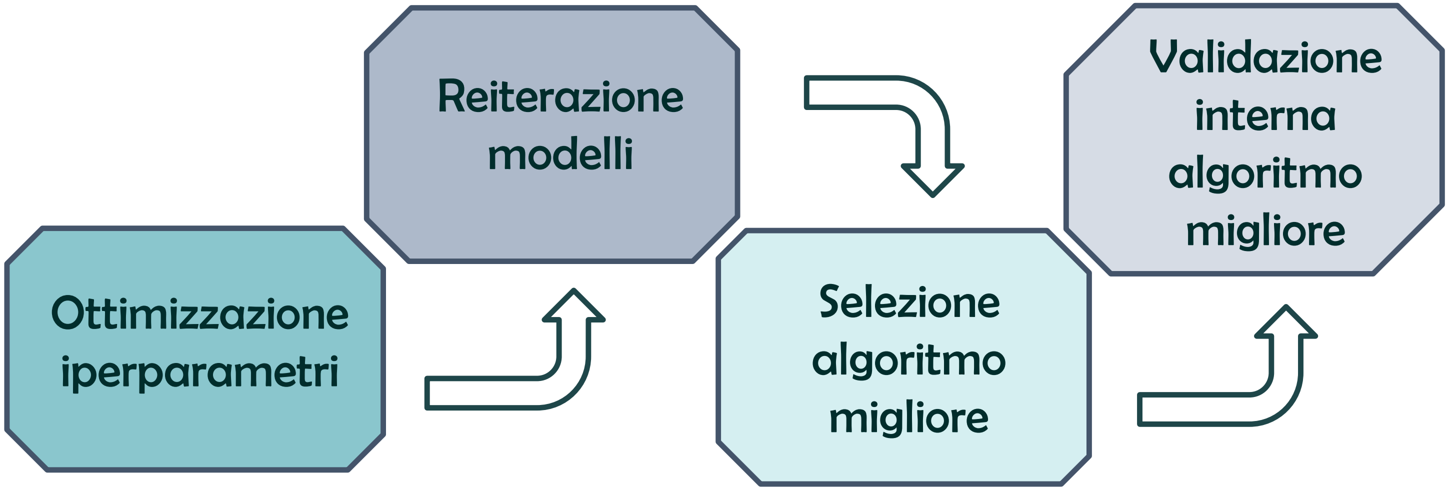
Y →
- 1-0
- pIC_{50}

data cleaning

da 1593 a 630 records: 295 molecole attive, 335 inattive

Costruzione dei modelli per classificazione e regressione

Rome Center for Molecular Design



Ottimizzazione iperparametri

Reiterazione modelli

Selezione algoritmo migliore

Validazione interna algoritmo migliore

```
est.fit(x_train, y_train)
y_pred = est.predict(x_test)
y_fit = est.predict(x_train)
y_cvpred = make_cv(x_train,
y_train, 5, est)
```

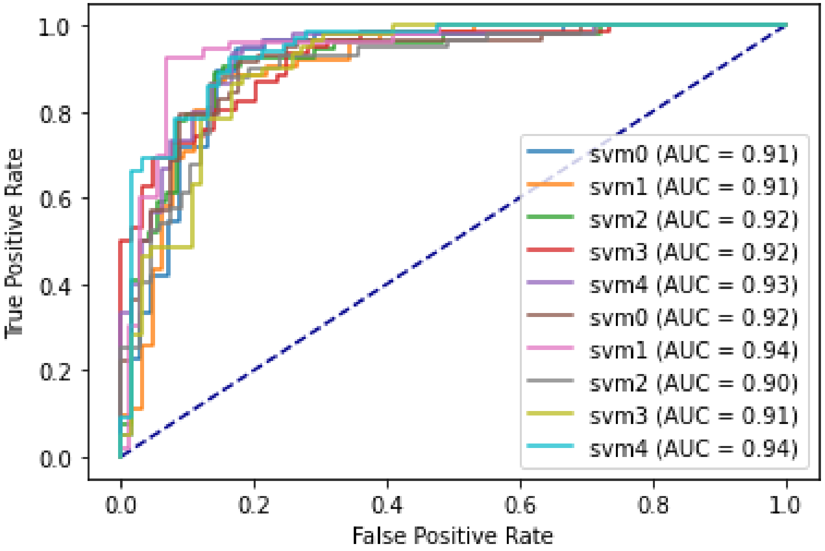
Risultati Training Classificazione

M.C.C. Fingerprints	y_pred	y_fit	y_cvpred
SupportVectorMachine	0,726	1	0,645
RandomForest	0,545	0,887	0,645
GradientBoosting	0,656	1	0,649

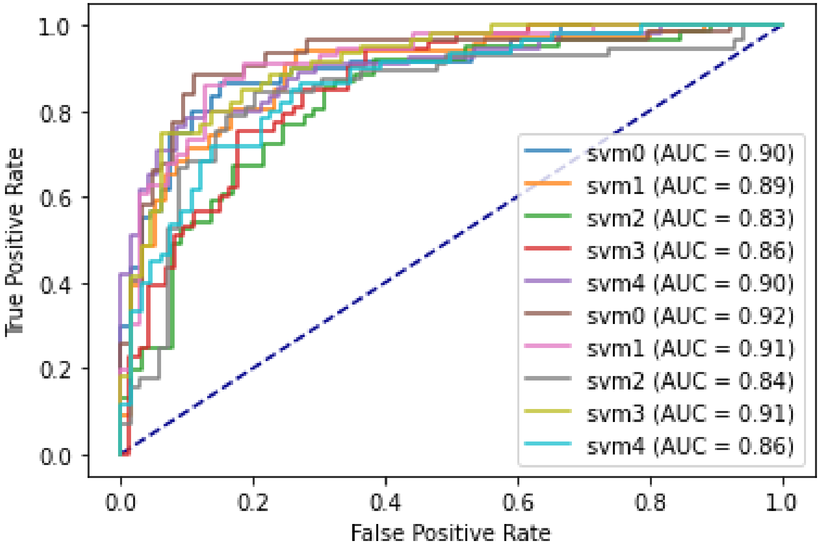
M.C.C. Descrittori Molecolari	y_pred	y_fit	y_cvpred
SupportVectorMachine	0,635	0,934	0,619
RandomForest	0,585	0,989	0,599
GradientBoosting	0,618	1	0,607

Risultati Training Classificazione

Rome Center for Molecular Design



FINGERPRINTS



DESCRITTORI MOLECOLARI

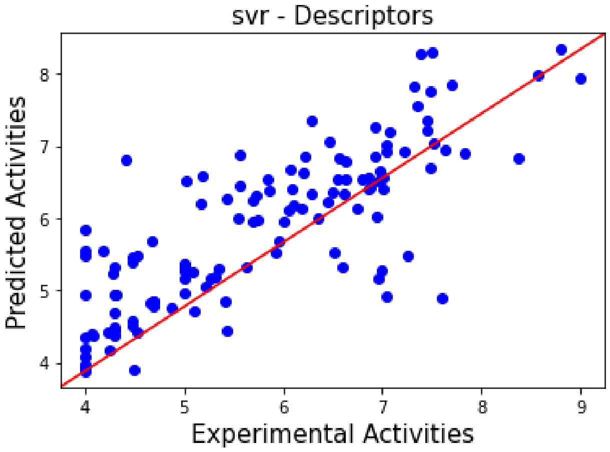
Risultati Training Regressione

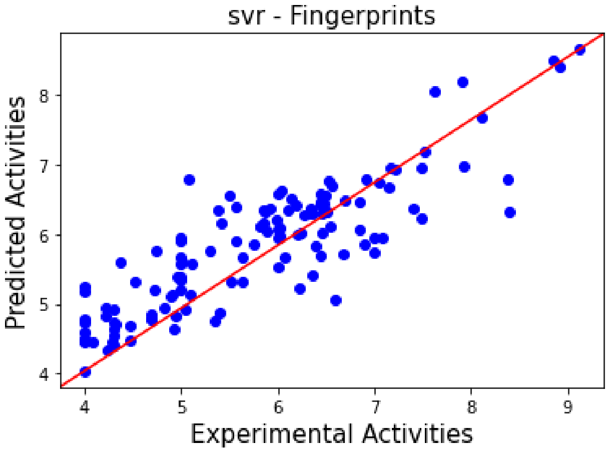
Fingerprints	y_pred R² pred	y_fit R²	y_cvpred Q²
SupportVectorMachine	0,707	0,993	0,695
RandomForest	0,666	0,921	0,654
Ridge	0,695	0,957	0,681

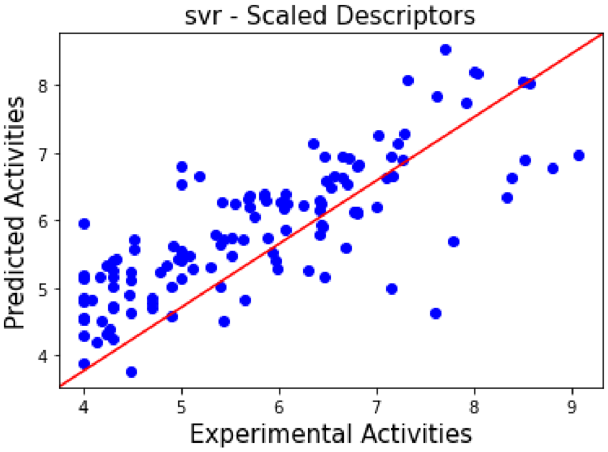
Descrittori Molecolari	y_pred R² pred	y_fit R²	y_cvpred Q²
SupportVectorMachine	0,585	0,886	0,609
RandomForest	0,539	0,937	0,557
Ridge	0,468	0,487	0,441

Risultati Training Regressione

Desc. Molec. Scalati	y_pred R² pred	y_fit R²	y_cvpred Q²
SupportVectorMachine	**0,630**	**0,875**	**0,589**
RandomForest	**0,628**	**0,941**	**0,555**
Ridge	**0,415**	**0,599**	**0,339**



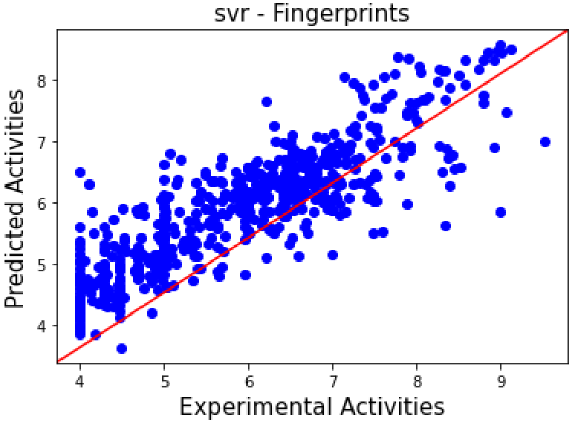


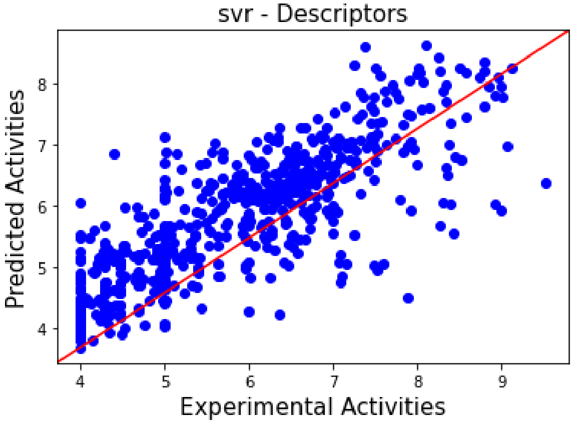


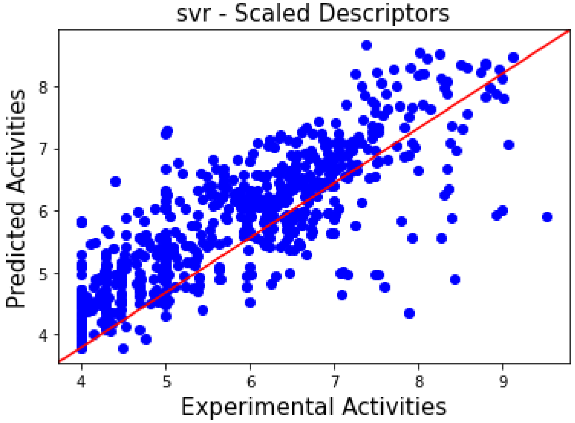
Risultati Validazione interna Classificazione e Regressione

SVC	Fingerprints	Descrittori Molecolari
Accuracy	**0,844**	**0,825**
M.C.C.	**0,687**	**0,652**

SVR	Fingerprints	Descrittori Molecolari	Desc. Molec. Scalati
Q²	**0,729**	**0,655**	**0,651**
RMSE	**0,677**	**0,764**	**0,769**







Validazione esterna ed analisi similarità training-test sets

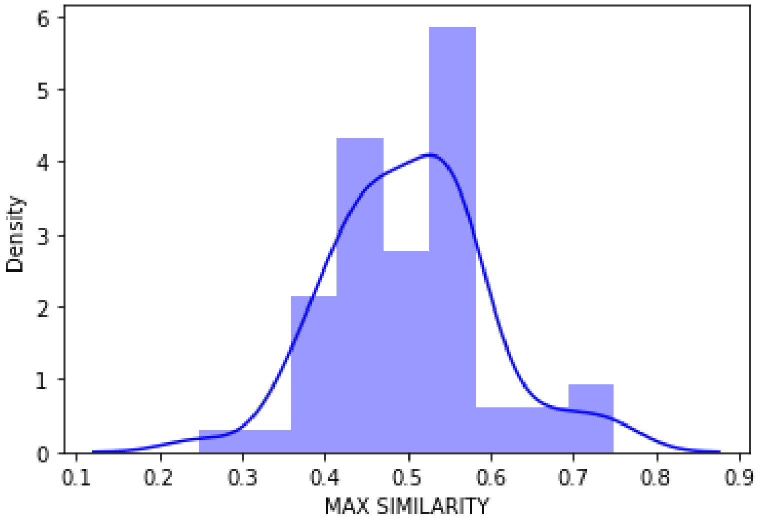
Ricerca *test set*

Confronto con *training set*

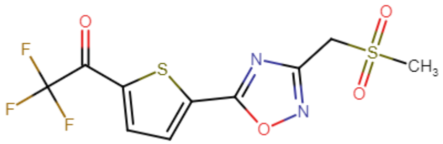
Generazione *features* e *labels*

58 molecole:

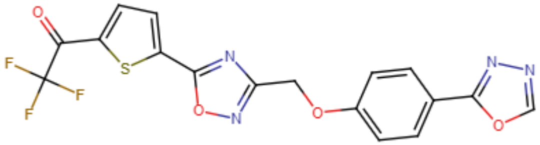
37 attive, 21 inattive



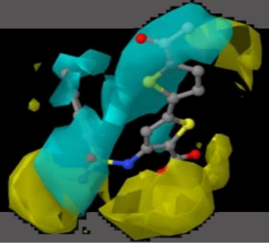
Training



Test



Max similarity: 0,48



Risultati validazione esterna

CLASSIFICAZIONE	DESCRITTORI MOLECOLARI	FINGERPRINTS
Accuracy	0,689	0,551
F1 score	0,769	0,566
Confusion Matrix	10-11 7-30	15-6 20-17
M.C.C.	0,303	0,171

REGRESSIONE	DESC. MOLEC.	DESC. MOLEC. SCALATI	FINGERPRINTS
M.A.E.	0,642	0,789	0,669

Analisi Selettività: costruzione modelli

Confronto datasets HDAC4-HDAC1

Selezione molecole con doppia attività

Definizione del rapporto tra i due pIC_{50}

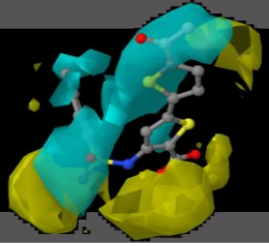
```
selectivity = []
for i in df_hdac4_1['RATIO_HDAC4_1']:
    if i>=0.5:selectivity.append(1)
    if i<=-0.5:selectivity.append(0)
    if i>-0.5 and i<0.5:selectivity.append(-1)
```

545 molecole:
185 HDAC4 selettive,
144 HDAC1 selettive,
216 non selettive

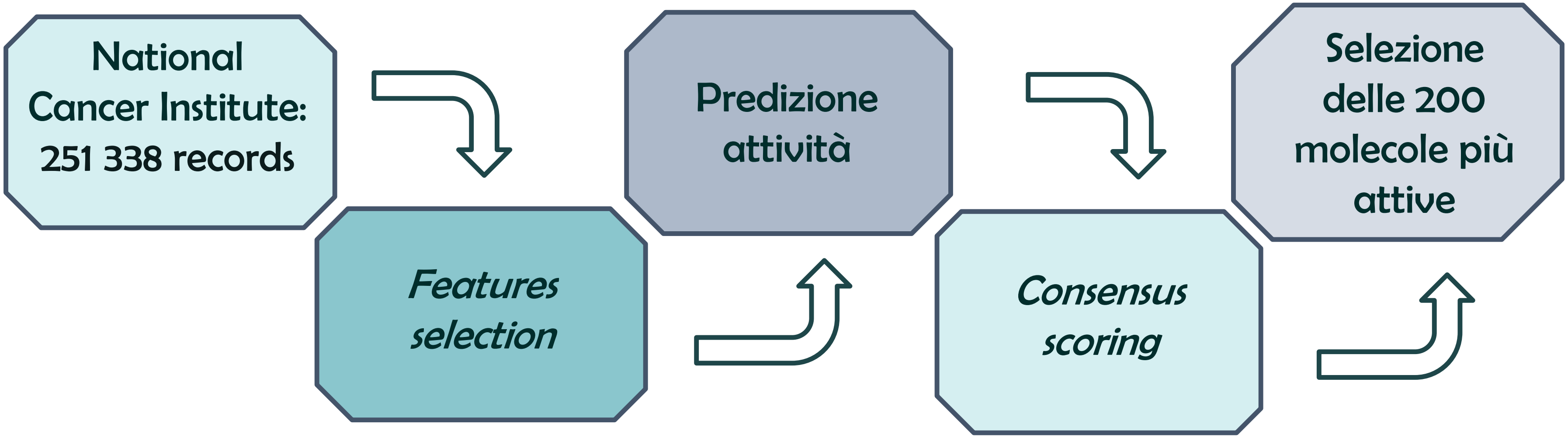
Risultati Training Classificazione e Regressione Selettività

M.C.C.	FINGERPRINTS	DESC. MOLEC.
SupportVectorMachine	0,900	0,883
RandomForest	0,838	0,875
GradientBoosting	0,905	0,878

R^2 pred	FINGERPRINTS	DESC. MOLEC.	DESC. MOLEC. SCAL.
SupportVectorMachine	0,854	0,814	0,810
RandomForest	0,808	0,818	0,820
Ridge	0,827	0,636	0,833



Virtual Screening

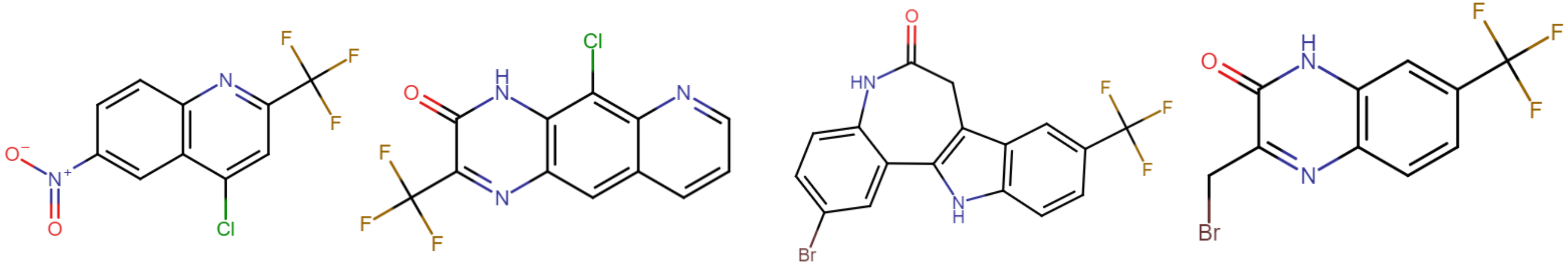


National Cancer Institute: 251 338 records

Features selection

Predizione attività

Consensus scoring

Selezione delle 200 molecole più attive