

Applicazione di tecniche Machine Learning per la Costruzione di Modelli QSAR sugli inibitori delle Sirtuine



SAPIENZA
UNIVERSITÀ DI ROMA

**Facoltà di Farmacia e Medicina
Corso di Laurea in Farmacia
Tesi Sperimentale in Chimica Farmaceutica
a.a. 2015/2016**

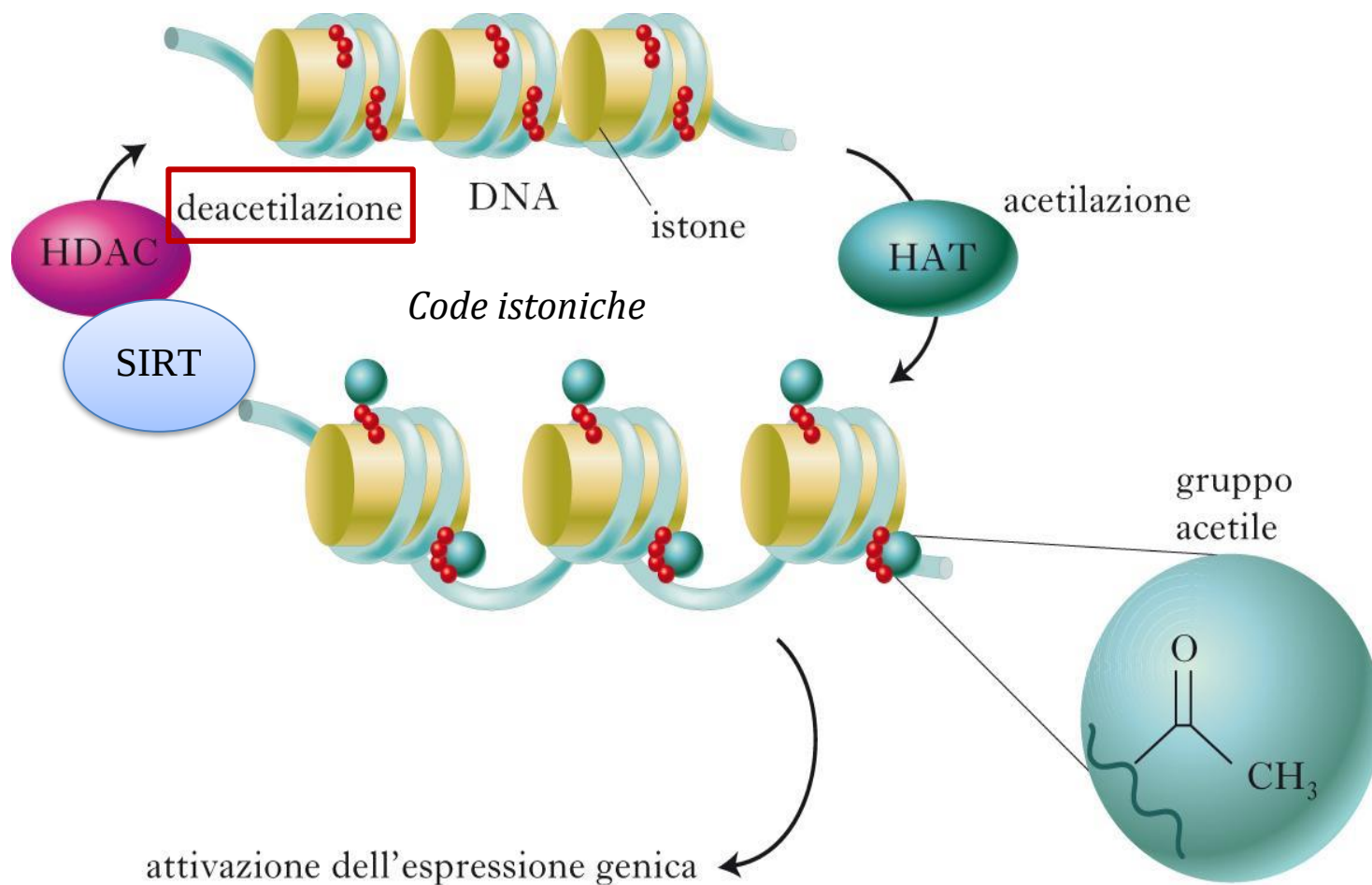
**Laureando: Nicola Adriano Pallottino
Matricola: 1472949**

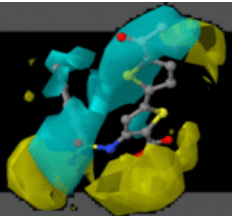
Relatore: Prof. Rino Ragno



Epigenetica

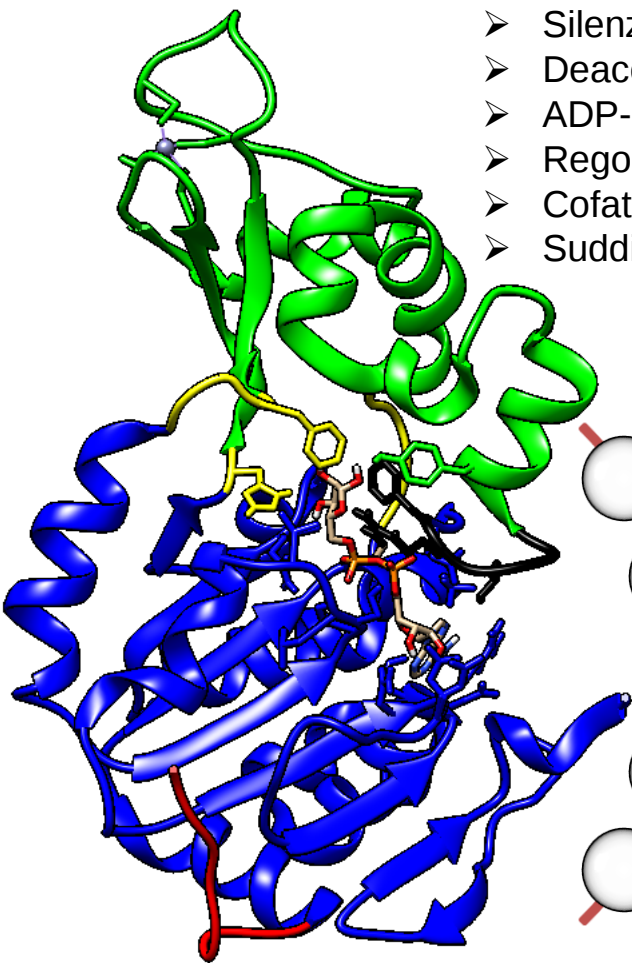
by www.RCMD.it





Famiglia Sir2

by **www.RCMD.it**



- Silenziamento Genico
- Deacetilazione
- ADP-ribosilazione
- Regolazione dei fattori di trascrizione
- Cofattore NAD⁺
- Suddivise in V classi:



Dominio Catalitico

Dominio Zn

Complesso Cofattore Ligando

Tasca di Rossmann

Dominio C-terminale

Classe I

SIRT1 SIRT2: Nucleo e Citoplasma
SIRT3: Nucleo e mitocondrio

Classe II

SIRT4: Mitocondrio

Classe III

SIRT5: Mitocondrio

Classe IV

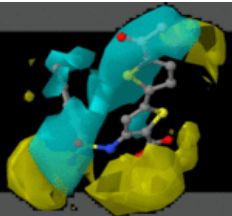
SIRT6: Nucleo
SIRT7: Nucleo e citoplasma

Classe V

Batteri Gram positivi

SIRT1 PDB 4IF6

Frye, R. A. Phylogenetic Classification of Prokaryotic and Eukaryotic Sir2-like Proteins. Biochem. Biophys. Res. Commun. 2000, 273, 793–798.



SIRT1

by www.RCMD.it

Enzima

SIRT1

Substrati epigenetici

H1K26
H3K9
H3K14
H4K16
DNMT1
BMAL1
RelA/p65
STAT3
P53
BCL6
PARP-1
HSF-1
PGC-1α
TLE1
YAP2
TIP
IRS-1
MCM10
eNOS
Recettore Androgeno
hMOF
PCAF
STAB1..

Funzioni

Deacetilando numerosi substrati determina:
Soppressione apoptotica
Formazione di eterocromatina
Implicazioni metaboliche
Regolazione del ritmo circadiano
Infiammazione
Degradazione del substrato e del gene
Stimola la sua trascrizione per quanto riguarda la gluconeogenesi, l'ossidazione degli acidi grassi, la biogenesi mitocondriale
Promuove l'autofagia
Incrementa l'accumulo dei lipidi epatici

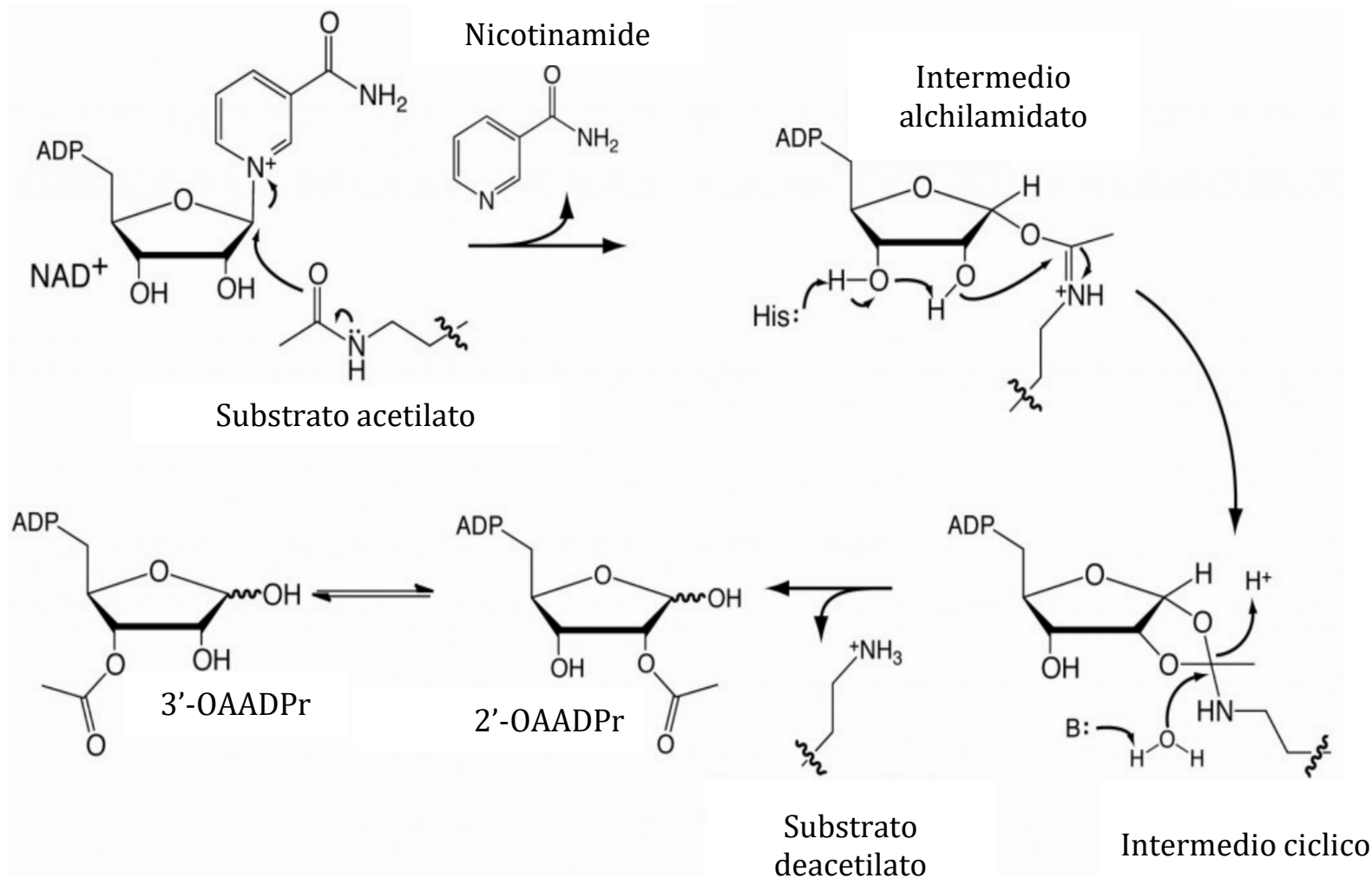
Patologie

Cancerogenesi
Infiammazione
Alterazione metabolica
Alterazione del ritmo circadiano
Alzheimer
Diabete
Neurodegenerazione
Favorisce la migrazione delle cellule tumorali



Meccanismo d'Azione

by www.rcmd.it

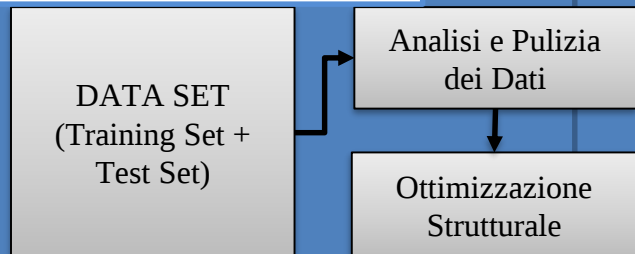




Costruzione Modelli QSAR

by www.RCMD.it

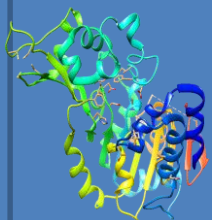
STEP 1. SELEZIONE DATASET



FINGERPRINTS

Hash Fingerprints

Bits



DESCRITTORI

Costituzionali

Topologici

Geometrici

Sterici

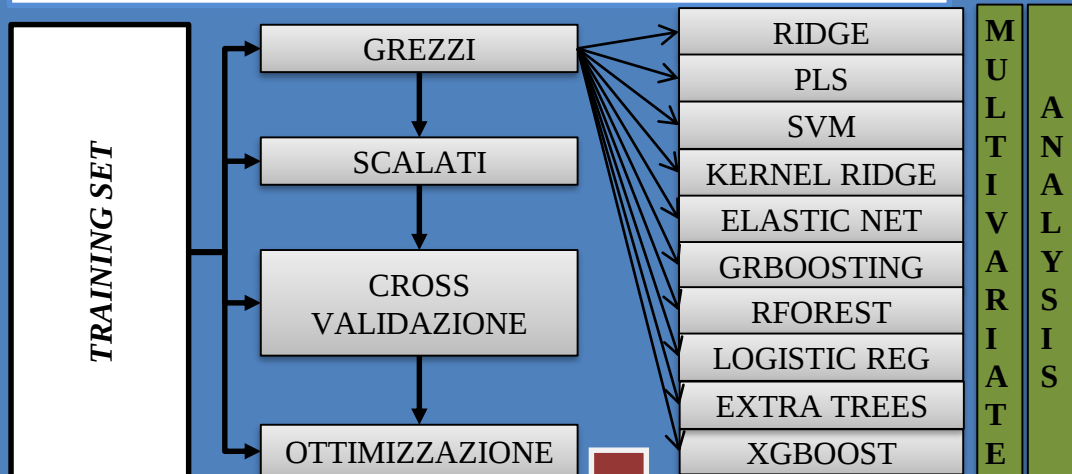
Lipofilici

Idrofobici

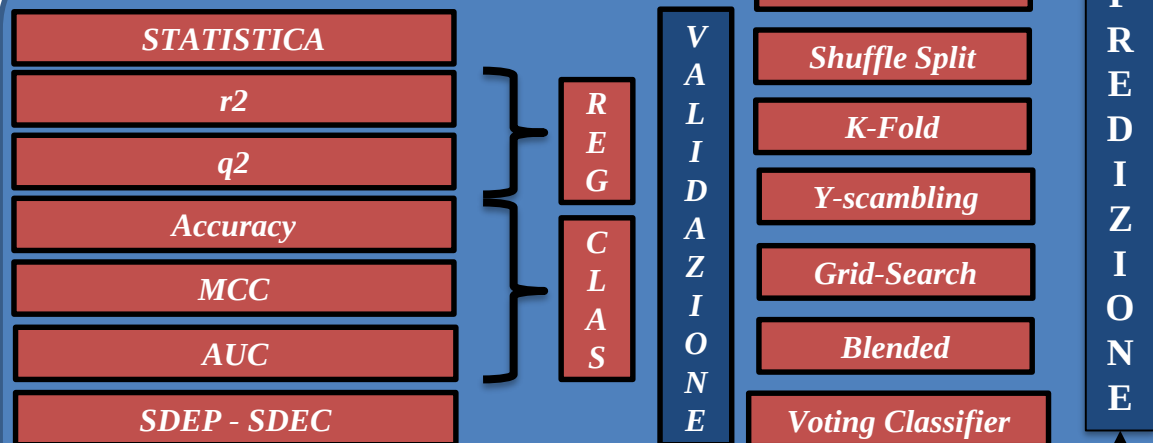
Elettronici

TEST SETS

STEP 2. PROCESSAMENTO DATI & ANALISI MULTIVARIATA



STEP 3. PREDIZIONE





Analisi dei dati

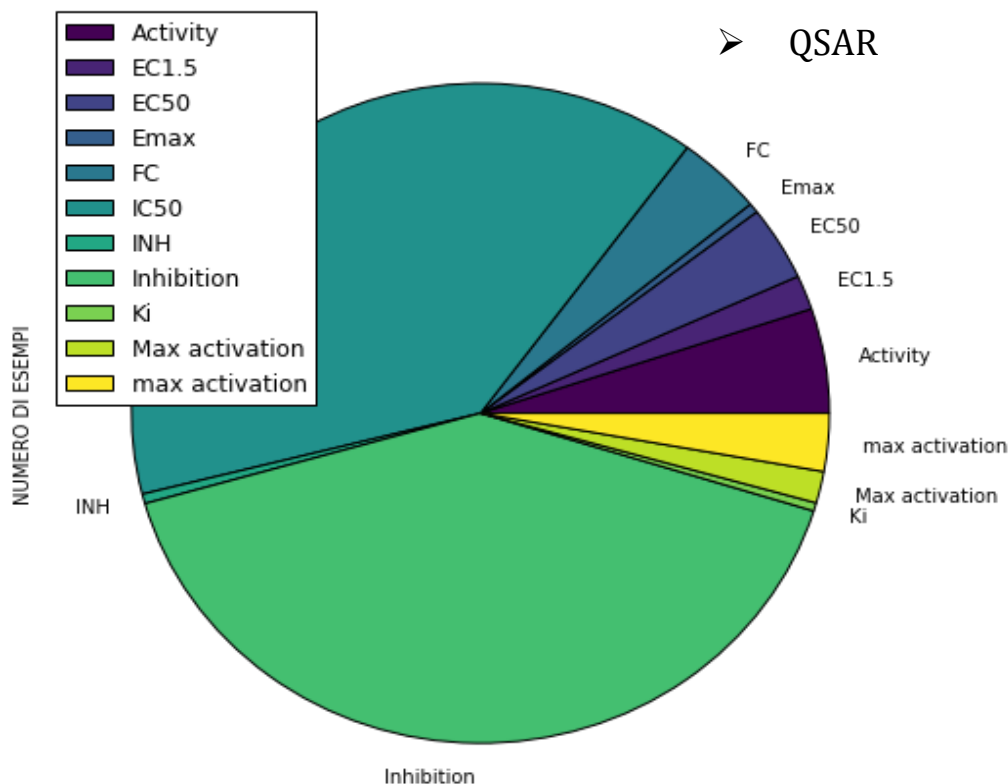
by **www.RCMD.it**



Acquisizione dei dati da ChEMBL (2133 composti saggiati)

Analisi dei dati:

- Esplorare ed avere familiarità con i dati
- Colonne e righe
- Intervallo
- Media & DevStd



$IC_{50} = 34.06\%$
Inibizione = 35.59%
 $K_i = 0.35\%$
 $EC_{50} = 3.18\%$
Attività = 4.47%

Pulizia del Dataset:

- Costruire un percorso
- Creare una Matrice X e vettore y
- QSAR

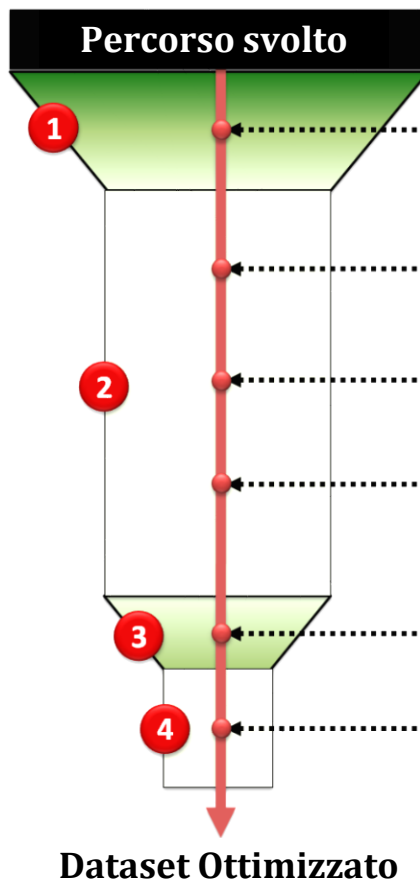
STEP 1. SELEZIONE DATASET



Pulizia dei dati

by **www.RCMD.it**

Pulizia dei dati



Rimozione di composti inorganici ed organometallici

CC([Se])NCCCC[C@H](NC(=O)OCc1ccccc1)C(=O)Nc2ccccc2

Conversione strutturale

Rimozione/pulizia dei sali

Cl.CN(C)CCCCC(=O)Nc1ccc(NC(=O)NC(=O)c2ccc(cc2)C(C)(C)C)cc1

Scelta dei Descrittori

Normalizzazione di specifici chemiotipi

Trattamento della chiralità

Trattamento delle forme tautomeriche

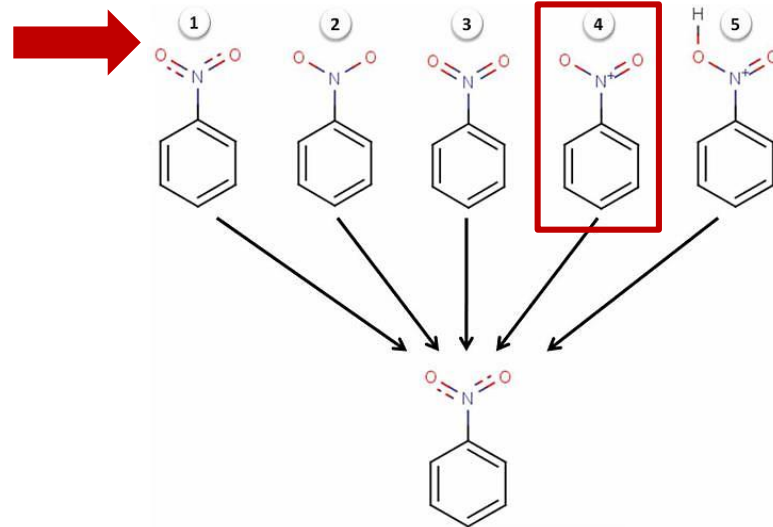
Escludere gli outlier

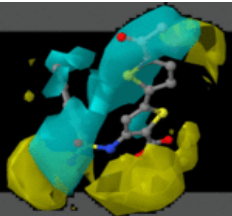
Distinguere i valori finiti

Analisi dei duplicati

Calcolare la media e la DevStd

Ispezione manuale





Descrittori & Fingerprints

by www.rcmd.it



0D: numero di legami, peso molecolare, numero degli atomi

1D: numero dei frammenti, leg-H acc/don, Crippen, PSA, SMARTS

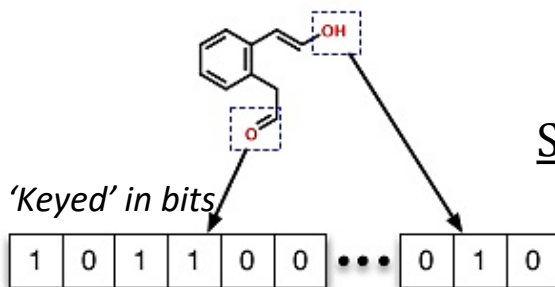
2D: descrittori topologici (Balaban, Randic, Wiener, BCUT, kappa, chi)

3D: descrittori geometrici, proprietà di superfici, CoMFA

4D: coordinate 3D + conformazioni

Distanza euclidea

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$



Similarità di Tanimoto

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

- Hash fingerprints & Dictionary Fragment
- 'Dictionary Keyed' Indica la presenza o l'assenza di una substruttura nel bit
- 'Hash Keyed' Indica la presenza o l'assenza di un descrittore nel bit
- Lunghezza dei bits varia da 166 a 4096 o maggiore
- La distanza fra i fingerprints è calcolata con la Similarità di Tanimoto



Risultati

by www.RCMD.it

Performance dei Modelli Machine Learning di Regressione QSAR: predizione esterna con il blended dei Fingerprints (test set 20 + 4)

	Desc	Scalato	Metodo ML	q^2 CV	r^2 pred	Y-s	SDEP
I _R	Fp	No	SVMl	0.63	0.61	-0.29	0.76
II _R	Fp	No	GBRh	0.66	0.7	-0.31	0.60
III _R	Fp	No	GBRls	0.61	0.58	-0.73	0.77
IV _R	Fp	No	RR	0.68	0.64	-0.27	0.73
V _R	Fp	No	PLS	0.645	0.65	-0.28	0.67
VI _R	Fp	No	KR	0.68	0.65	-0.26	0.73
VII _R	Fp	No	SVMRbf	0.68	0.61	-0.27	0.82
VIII _R	Fp	No	RFR	0.68	0.52	-0.28	0.75
IX _R	Fp	No	ExtraR	0.68	0.5	-0.46	0.85
X _R	Fp	No	EN	0.685	0.66	-0.29	0.70
XI _R	Fp	No	EN2	0.64	0.65	-0.26	0.74
XII _R	-	-	Blended	0.66	0.615		0.8



Risultati

by www.RCMD.it

**Performance dei Modelli Machine Learning di Classificazione QSAR SIRT1:
predizione esterna con il Voting Classifier dei Descrittori
(110 composti in predizione esterna)**

	Scalato	Metodo	Peso	Accuracy CV	Accuracy ext	MCC CV	MCC ext
I _C	Si	SVC	1.8	0.8	1	0.62	1
II _C	Si	ExtraC	0.1	0.84	0.99	0.66	0.70
III _C	Si	RFR	0.1	0.84	0.99	0.65	0.70
IV _C	Si	GBRex	0.2	0.84	0.99	0.645	0.70
V _C	Si	GBRd	0.2	0.97	0.99	0.66	0.70
VI _C	Si	LG	0.8	0.8	0.99	0.645	0.82
VII _C	Si	XBG	1	1.0	1.0	0.66	1.0
			TOT		=		=
VIII _C		VC	0.53		1.0		1.0



Risultati

by **www.RCMD.it**

0 - Not Active, 1 - Active

TEST3

(1, 2048)

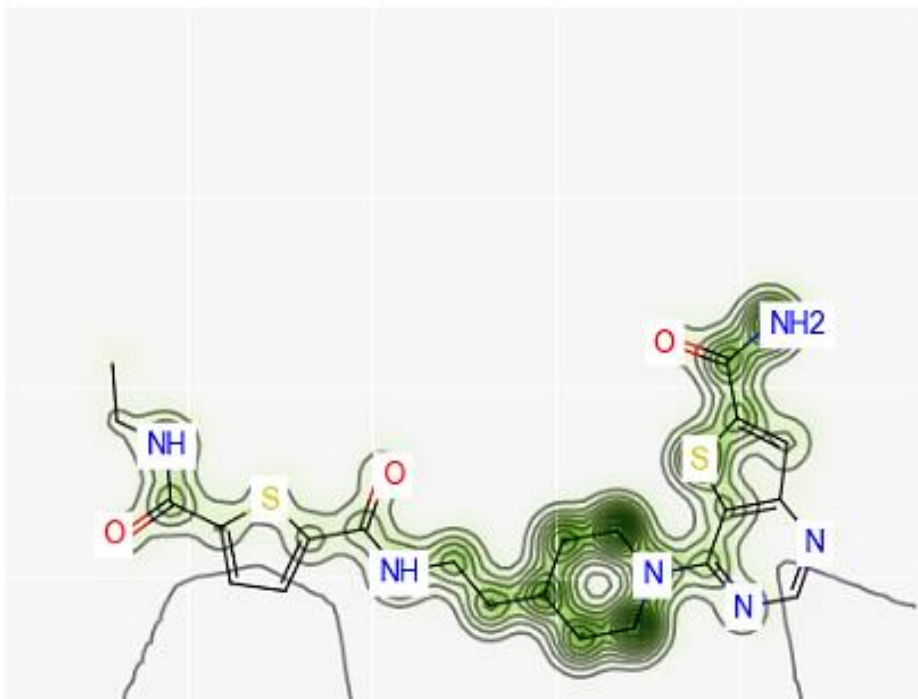
Experimental Outcome: 4 True

Name: IS_ACTIVE, dtype: bool

Predicted Outcome [1]

Probability of being Not Active: 0.2277161572596027

Probability of being Active: 0.7722838427403973



Voting Classifier dei Descrittori (110 pred)

Y true	Y pred	Y proba _{pred}	Molid
0	0	0.40	MC3417
0	0	0.40	MC3412
1	1	0.78	TEST3
1	1	0.55	TEST6

Svm Classifier dei Descrittori (110 pred)

Y true	Y pred	Y proba _{pred}	Molid
0	0	0.42	MC3417
0	0	0.40	MC3412
1	1	0.77	TEST3
1	1	0.52	TEST6



SIRT1 vs SIRT2

by www.rcmd.it

Modelli Machine Learning di Classificazione per valutare la selettività SIRT1/2 :
Validazione interna con CV - Shuffle split

	Metodo	Accuracy	MCC fit	M Accuracy CV	Roc AUC	SDEP AUC	MCC CV
I _{S1/2}	LG	1	1	0.99	1	0	0.98
II _{S1/2}	GBC	1	1	0.98	0.97	0	0.98
III _{S1/2}	RFC	1	1	0.99	1	0	0.98
IV _{S1/2}	EXC	1	1	0.99	1	0	0.98
V _{S1/2}	RFC	1	1	0.99	1	0	0.98



Mapping SIRT1 vs SIRT2

by **www.RCMD.it**

CHEMBL492308

(1, 2048)

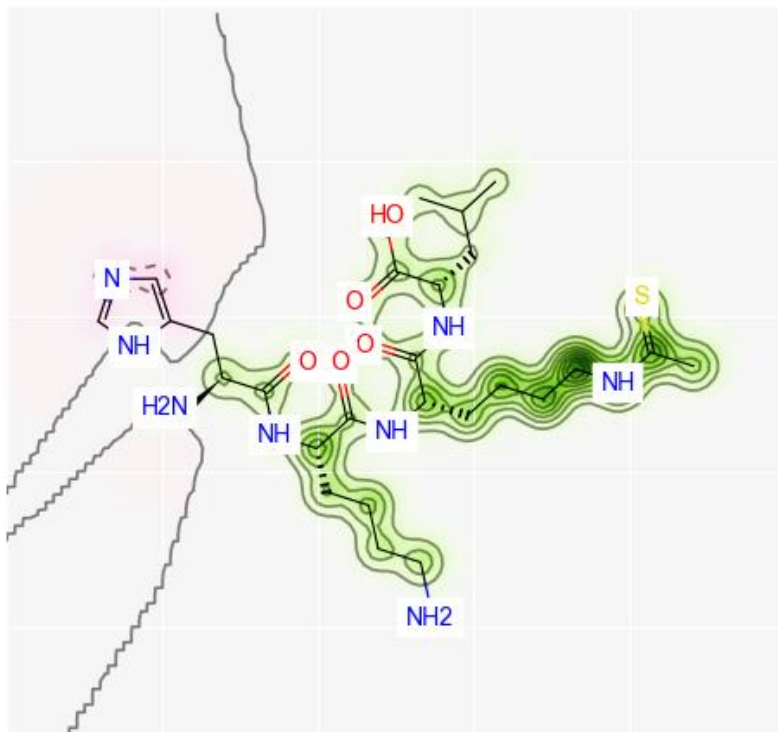
Experimental SIRT1 Outcome: 6.318758762624412

Experimental SIRT2 Outcome: 4.7011469235902945

Predicted CLASS [True]

Probability of being SIRT2 selective: 0.12968195141132188

Probability of being SIRT1 selective: 0.8703180485886781



CHEMBL3343652

(1, 2048)

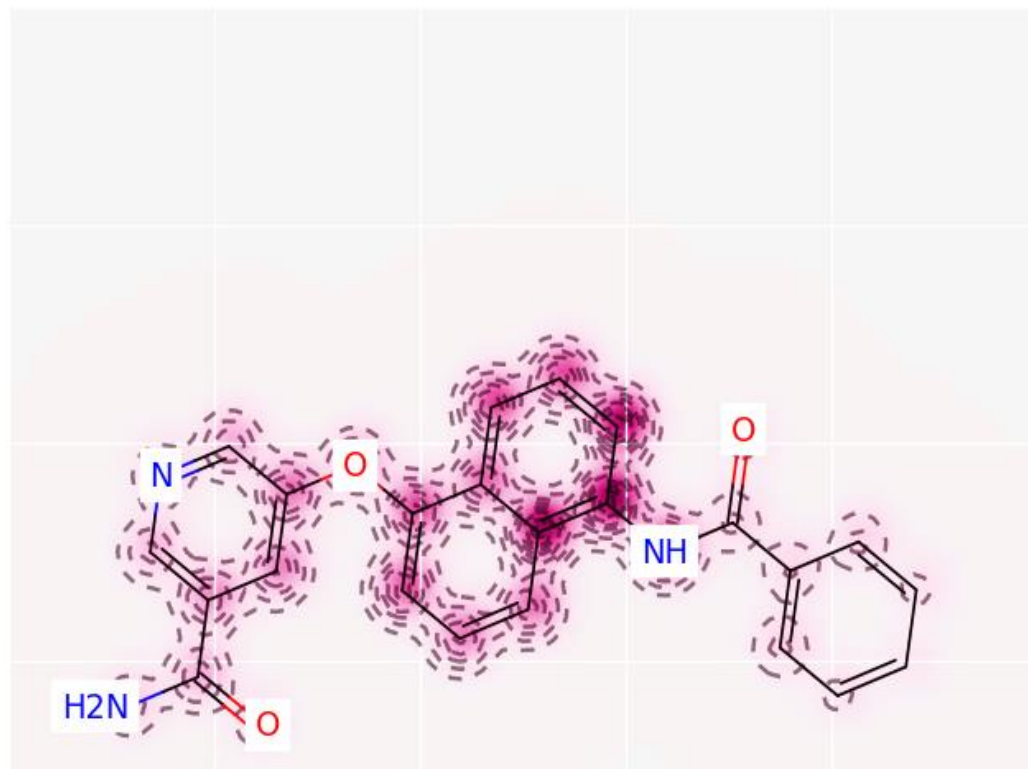
Experimental SIRT1 Outcome: 4.59950406936991

Experimental SIRT2 Outcome: 7.3160528692484865

Predicted CLASS [False]

Probability of being SIRT2 selective: 0.9861897516963828

Probability of being SIRT1 selective: 0.013810248303617005



STEP 3. PREDIZIONE



Mapping Classificazione

by www.rcmd.it

- Porzione Tiofenica
- Piperazina
- **Piperidina**

CHEMBL2332040

(1, 2048)

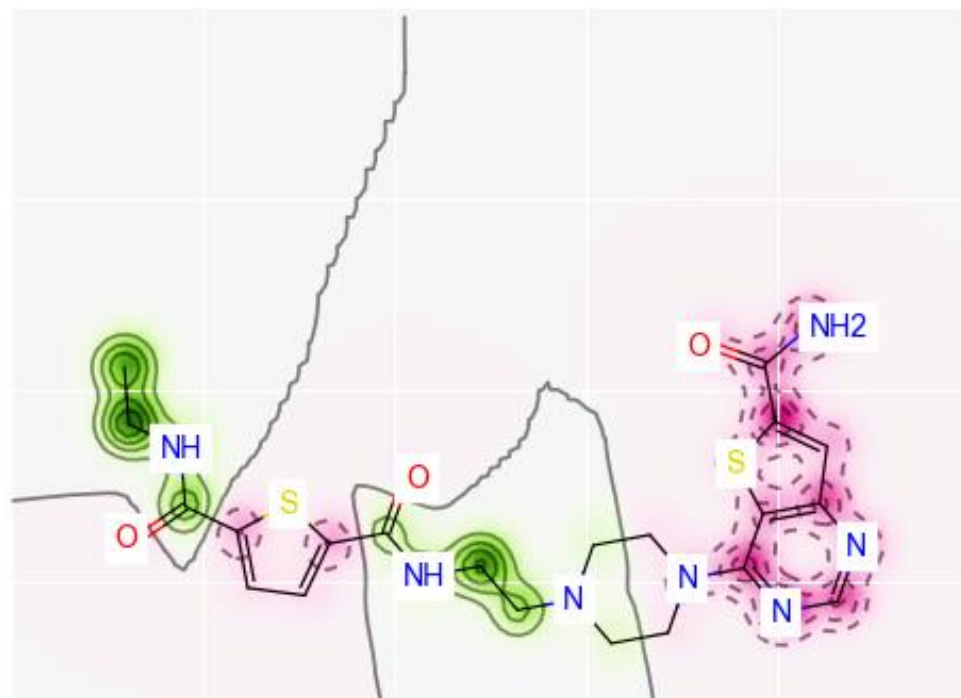
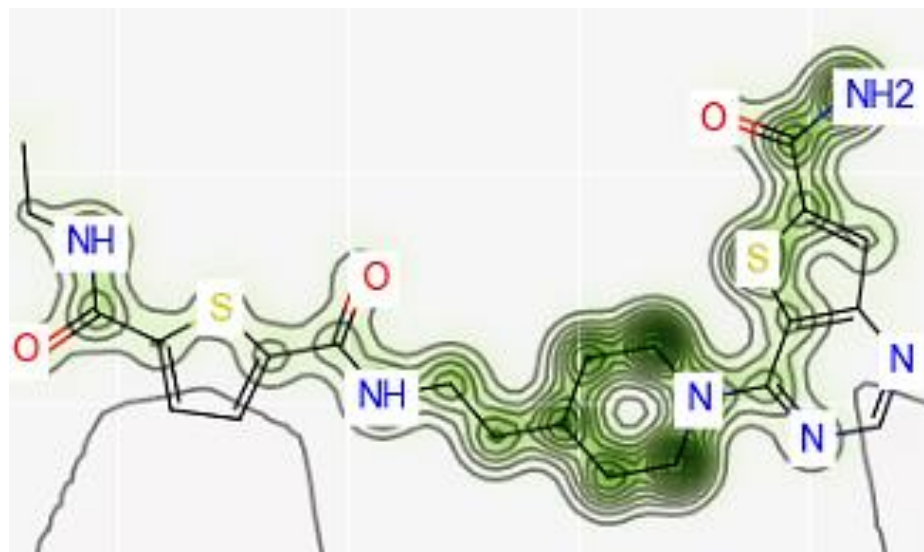
Experimental SIRT1 Outcome: 7.508638306165731

Experimental SIRT2 Outcome: 8.318758762624412

Predicted CLASS [True]

Probability of being SIRT2 selective: 0.49461999432186066

Probability of being SIRT1 selective: 0.5053800056781393



STEP 3. PREDIZIONE

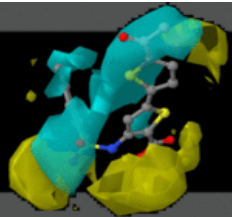


Prospettive future

by www.RCMD.it

- 3-D QSAR (modelli preliminari)
- Virtual Screening
- Progettazione di nuovi composti

					Fitting		Pret_FS		Y-scrambling		
	Cluster	Allineamento	Probe	PC	r^2	SDEC	q^2	SDEP	r^2	q^2	SDEP
I _{RTr}	Cl7	piuattiva	d	3	0.66	0.59	0.45	0.75	0.33	-0.53	1.26
II _{RTr}	Cl4	menoattiva	d	3	0.744	0.28	0.461	0.404	0.21	-0.40	0.647
III _{RTr}	Cl7	menoattiva	d	3	0.577	0.66	0.413	0.77	0.29	-0.7	1.31



Ringraziamenti

by www.RCMD.it

Dott. Alexandros Patsilidakos

Dott.ssa Manuela Sabatino

Gli studenti dell'RCMD

Grazie per l'attenzione