Sviluppo di modelli QSAR predittivi mediante tecniche di Machine Learning. Applicazione ad inibitori di PKC-θ.

SAPIENZA
UNIVERSITÀ DI ROMA

Facoltà di Farmacia e Medicina
Corso di Laurea in Chimica e Tecnologia Farmaceutiche
Tesi Sperimentale in Chimica Farmaceutica
a.a. 2019/2020

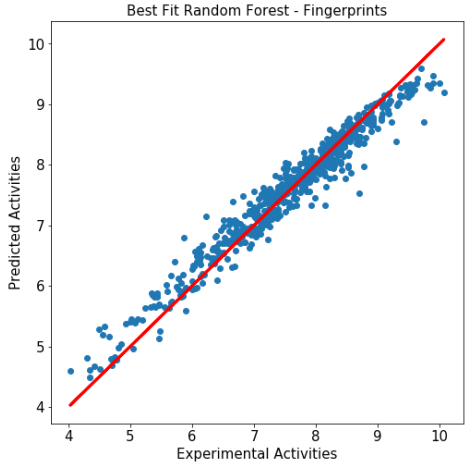
Laureanda: Chiara Consolini
Matricola: 1601470

Relatore: Prof. Rino Ragno

Un'overview del lavoro

PKC-θ

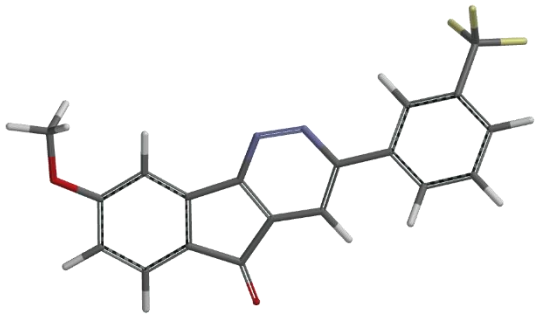
Metodologie QSAR e procedura sperimentale



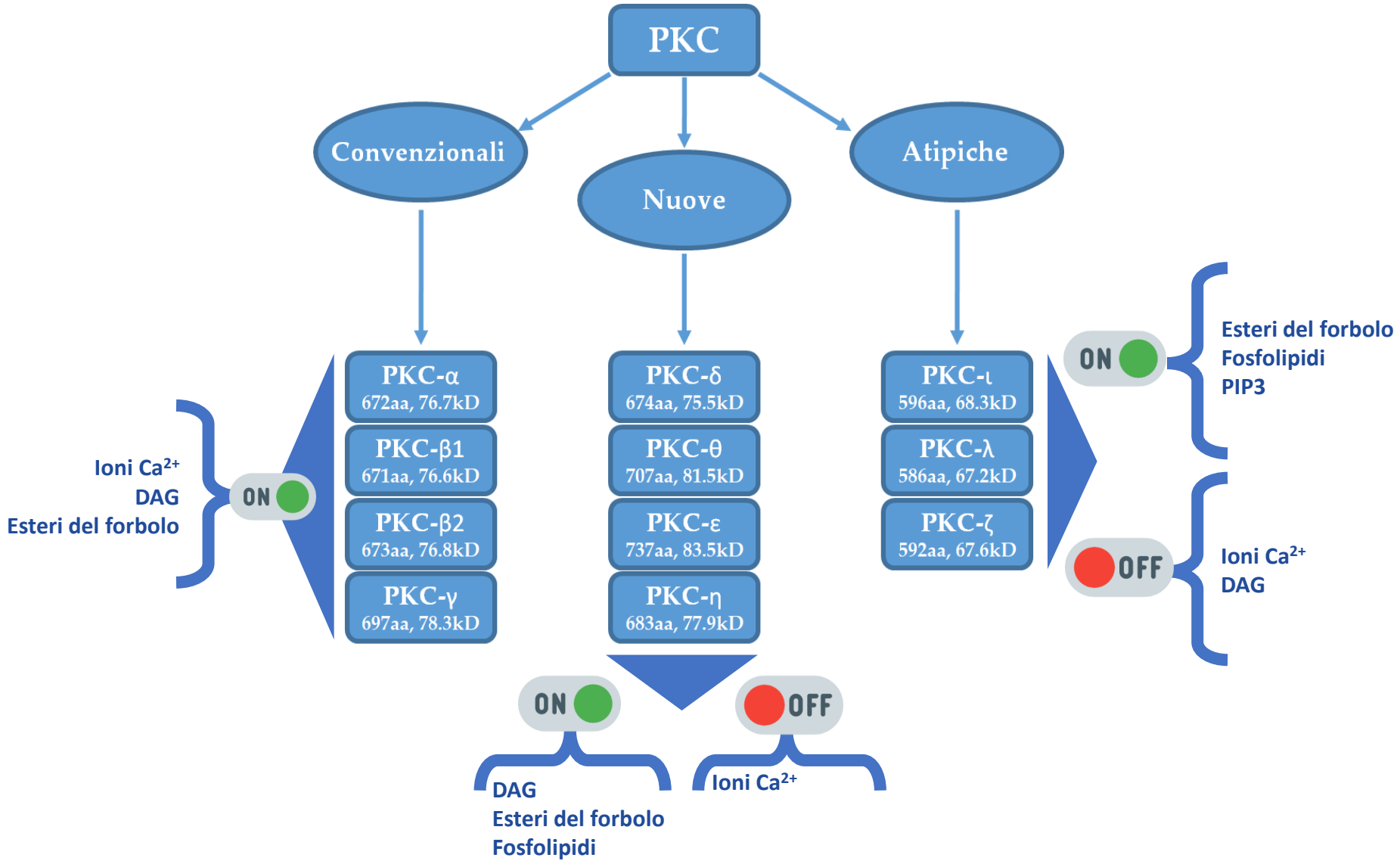
Implicazione della PKC-θ nella biologia delle cellule T

Nuovi Inibitori

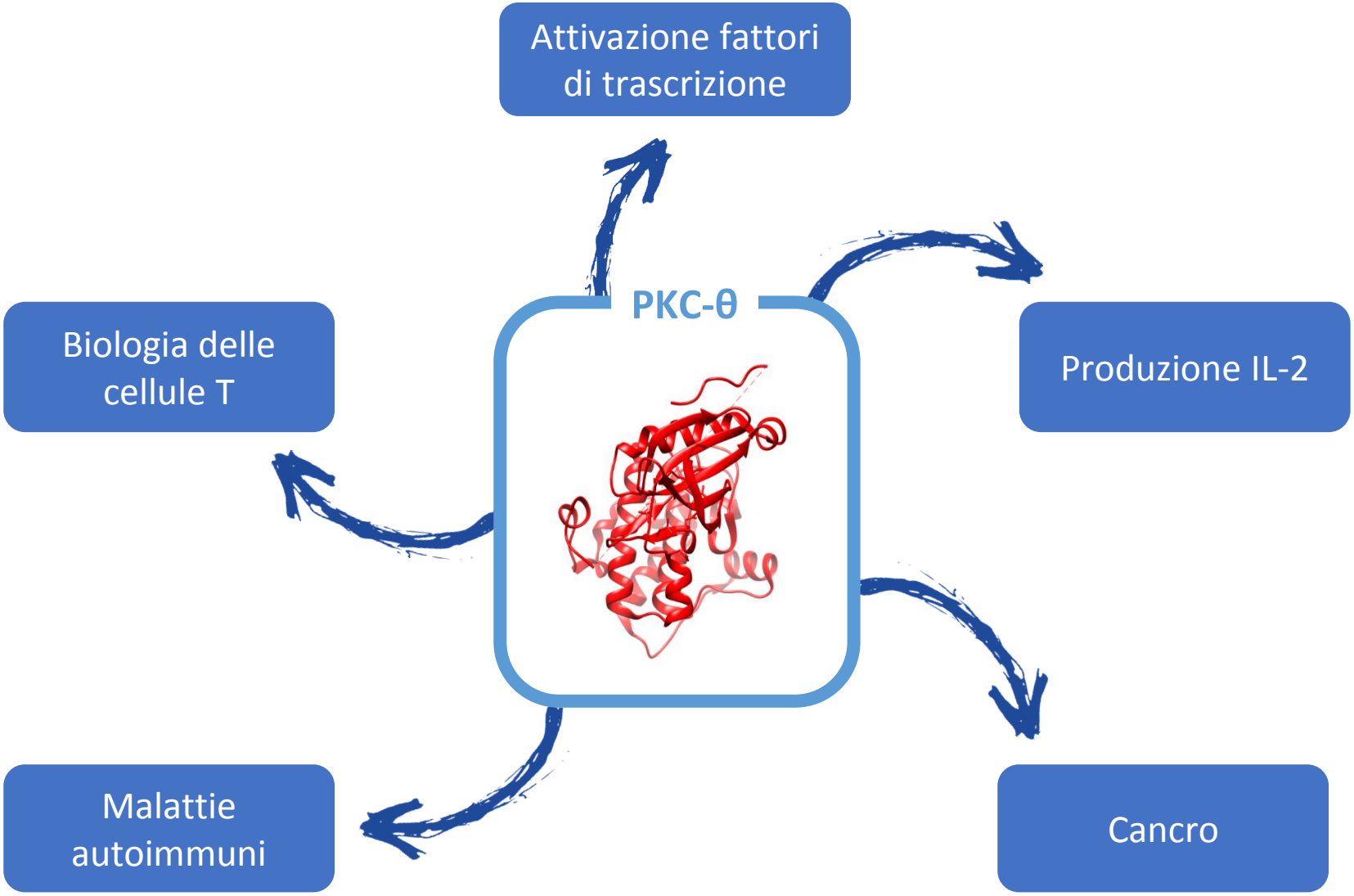
Risultati



La famiglia delle protein chinasi

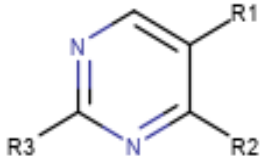


Target biologico di interesse: la PKC-θ

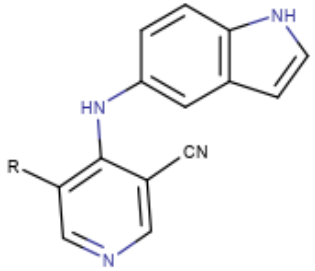


Categorie principali di inibitori di PKC-θ

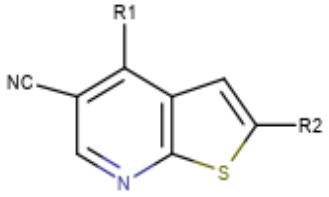
Derivati amminopirimidinici



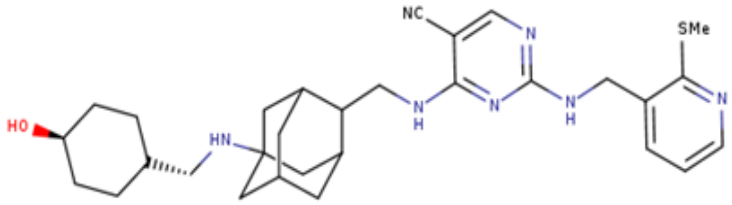
Derivati cianopiridinici



Derivati 5-cianotieno[2,3b]piridinici

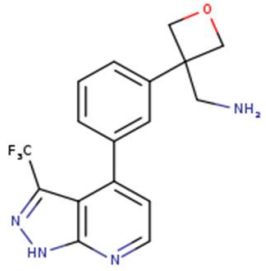


Derivati 2,4-diammino-5-cianopirimidinici



IC_{50} = 0,44 nM

Derivati pirazolopiridinici

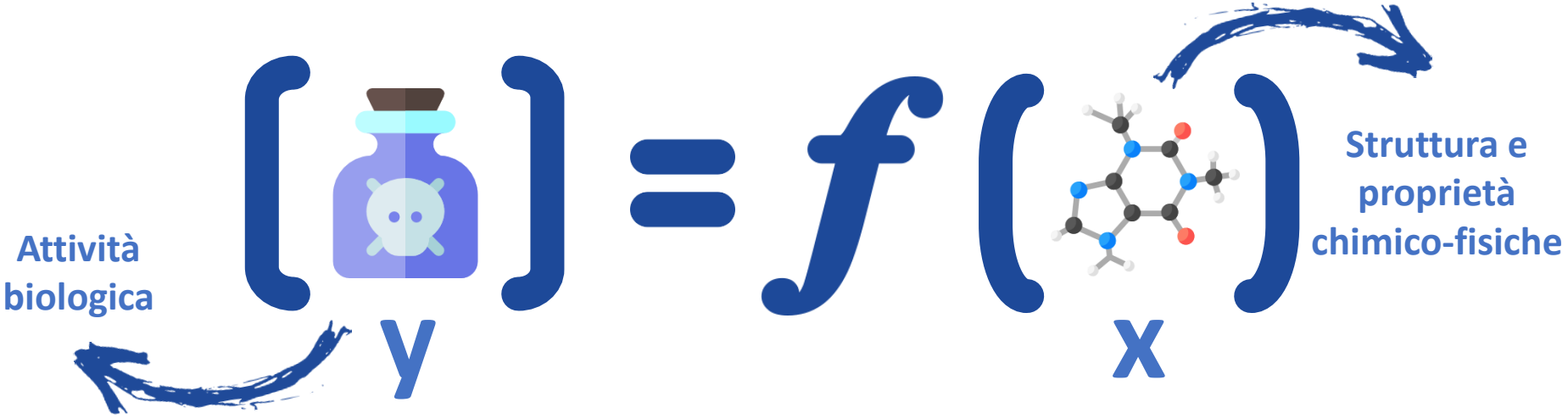


Ki = 6 nM

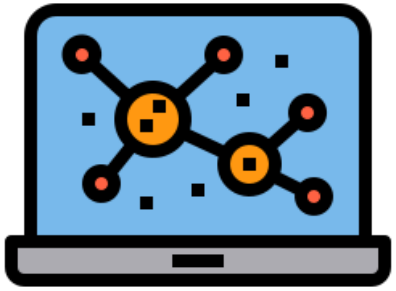
(Q)SAR

=

(Quantitative) Structure – Activity Relationship



Scopo della tesi



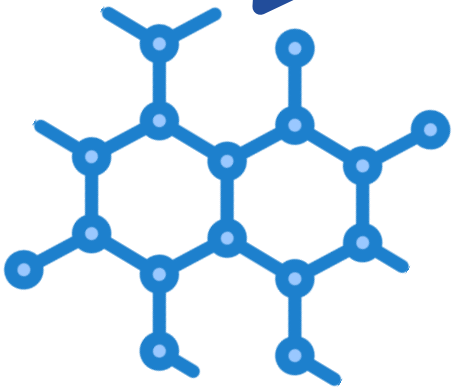
Linux

PYTHON

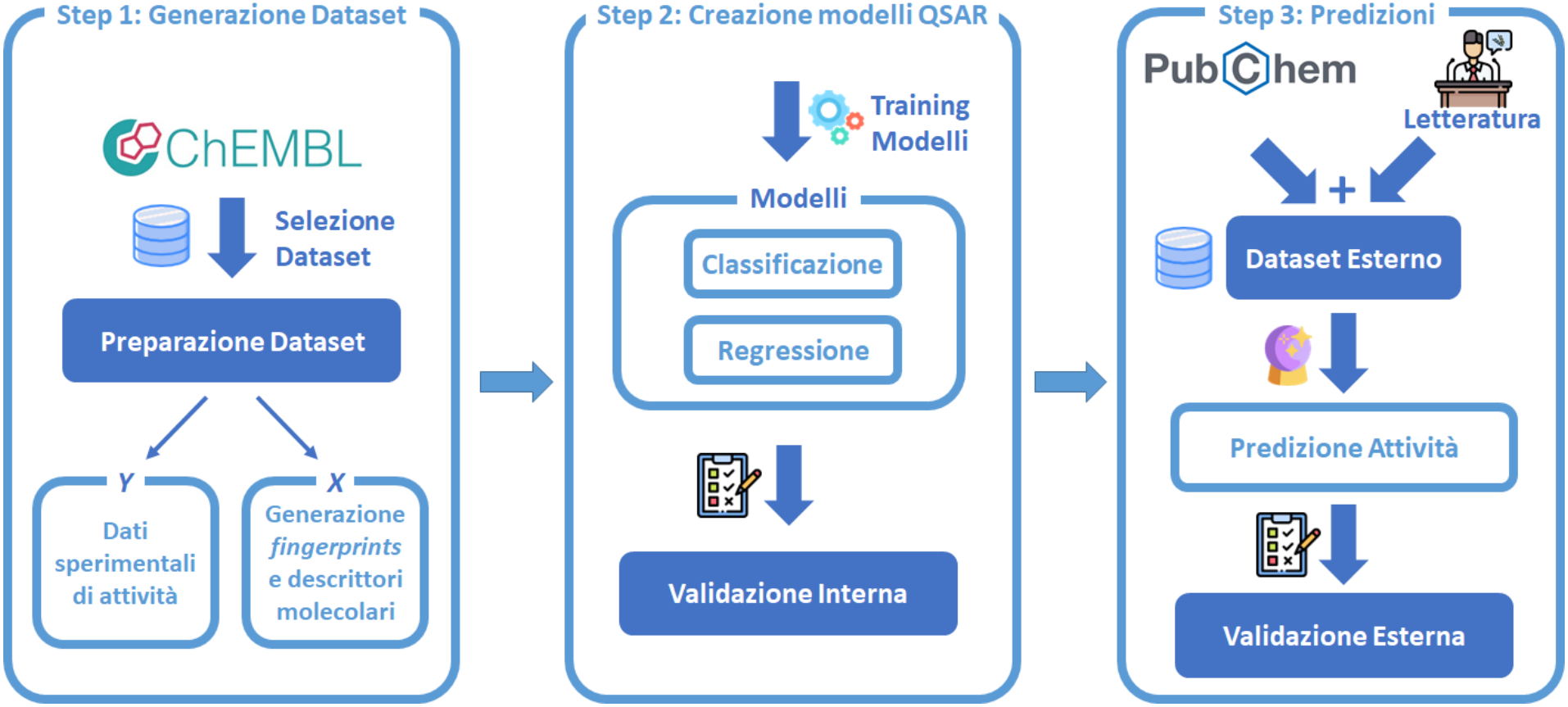
QSAR

Machine Learning

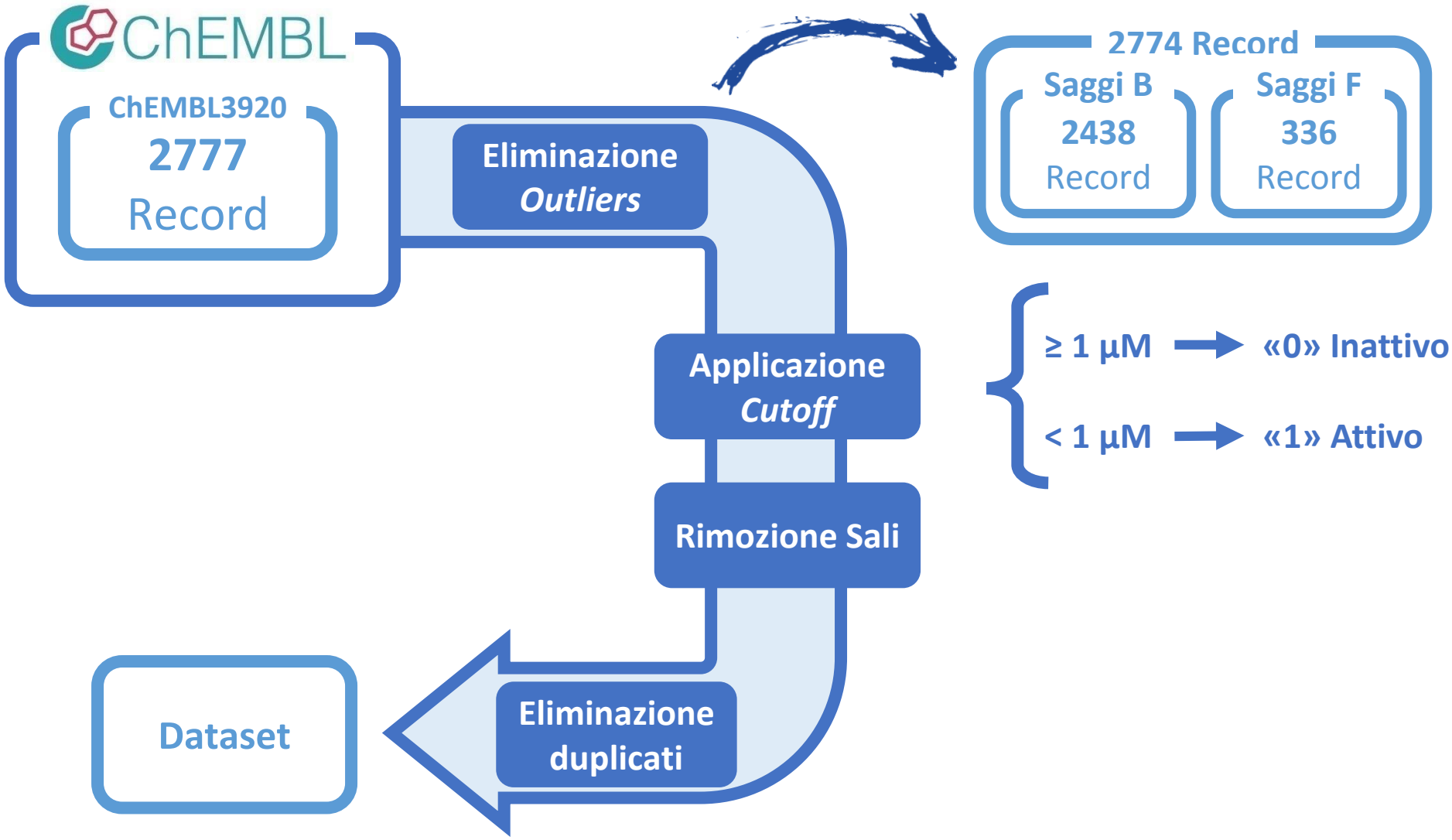
Potenziali nuovi inibitori di PKC-θ



Procedura sperimentale



Step 1 – Compilazione Dataset



Step 2 – Classificazione: calcolo delle *features*

SMILES

NCCc1c[nH]c2ccc(O)cc12

CARATTERISTICHE STRUTTURALI

CARATTERISTICHE CHIMICO-FISICHE

FINGERPRINTS

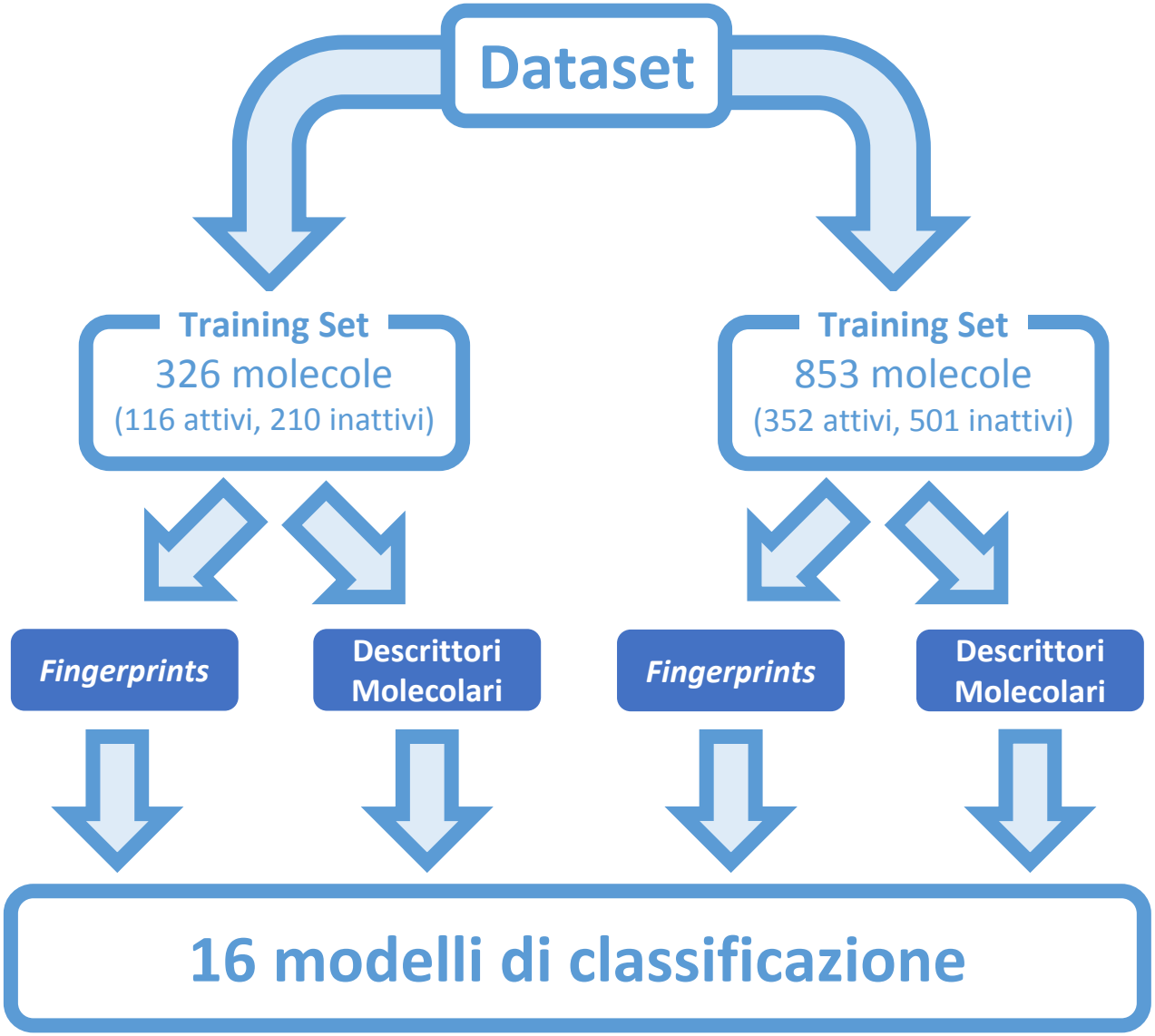
(101000100…001100010001)

DESCRITTORI

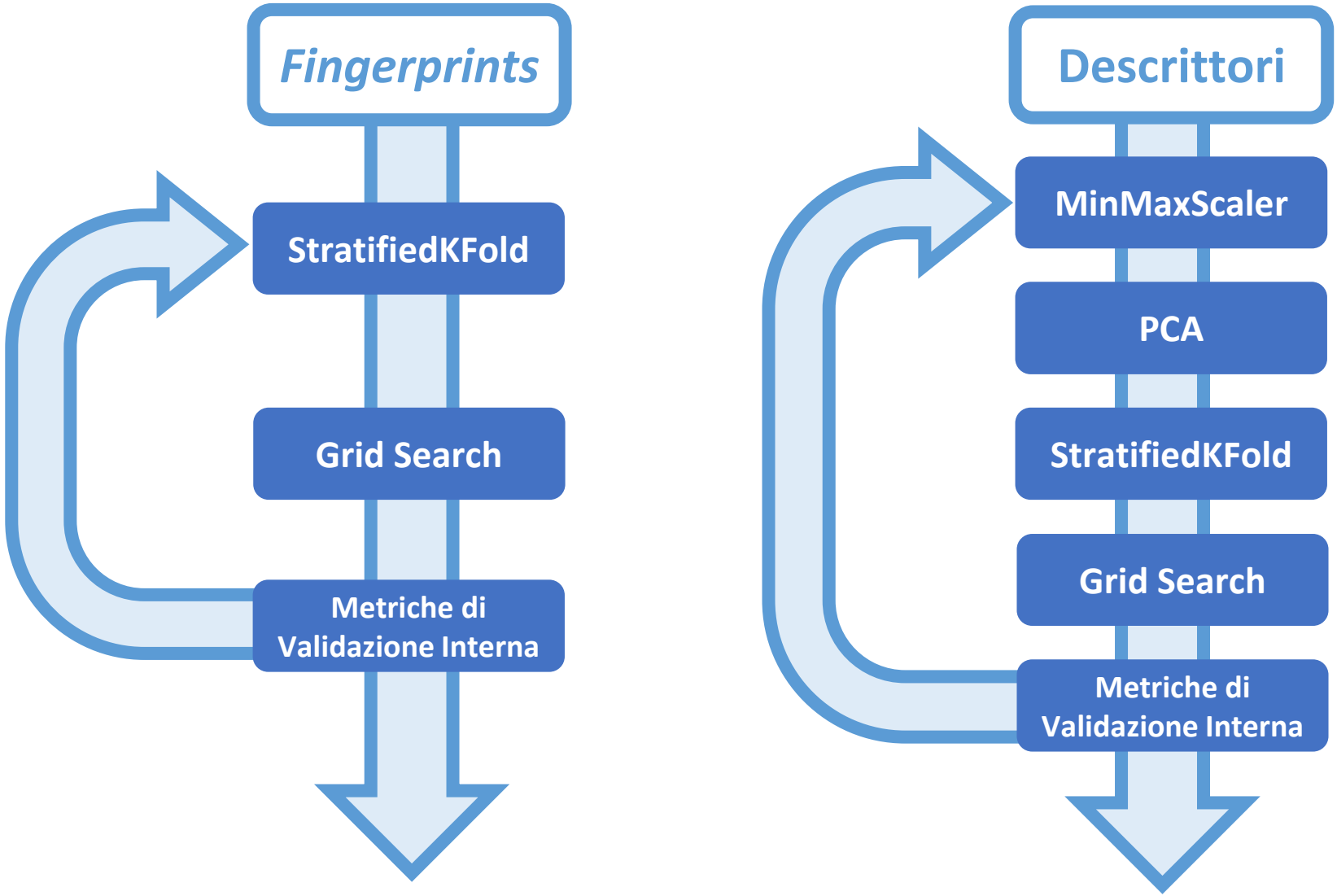
(37.3, 1.384, …, -0.0474, 0.05)

RDKit
Open-Source Cheminformatics and Machine Learning

Step 2 – Classificazione: creazione modelli



Step 2 – Classificazione: Training e validazione



Accuracy

$$\frac{TP + TN}{TP + TN + FP + FN}$$

F_1-score

	Predetto 0	Predetto 1
Reale 0	TN	FP
Reale 1	FN	TP

MCC

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

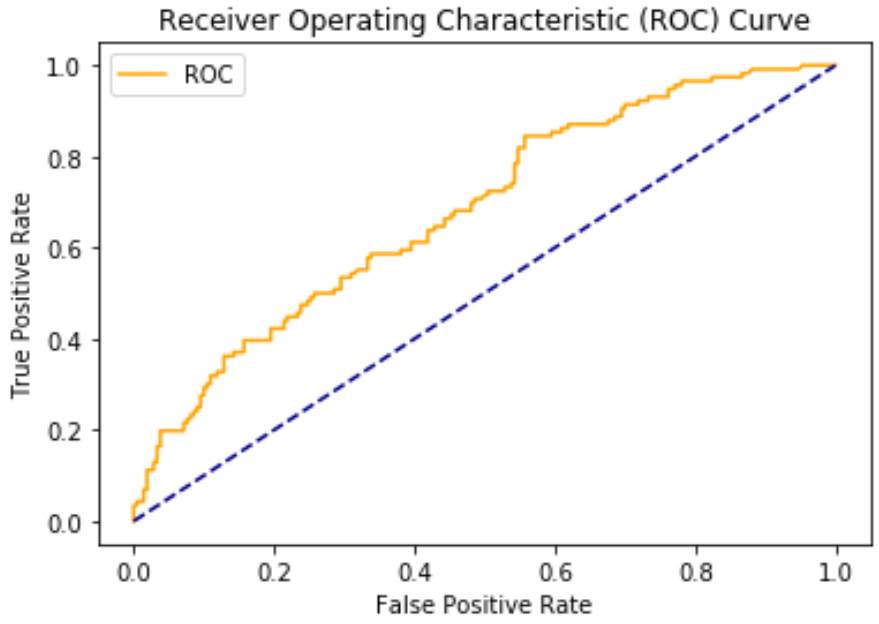
ROC AUC

Training set generato a partire dai saggi cellulari (F) : 326 molecole

	Algoritmo	*Accuracy*	MCC	ROC AUC	F1 *score*
Fingerprints	Random Forest	0,70	0,32	0,63	0,45
	KNN	0,70	0,40	0,63	0,50
	GBM	0,63	0,33	0,57	0,40
	Extra Trees	0,63	0,33	0,52	0,21
Descrittori	Random Forest	0,63	0,33	0,57	0,29
	KNN	0,70	0,35	0,62	0,40
	GBM	0,52	0,28	0,45	0,20
	Extra Trees	0,72	0,32	0,60	0,38

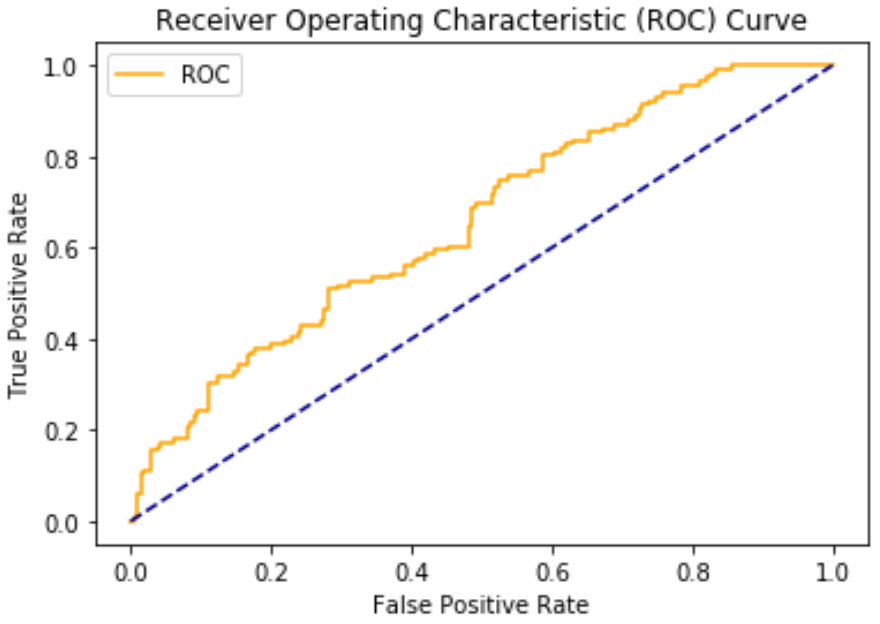
Training set generato a partire dai saggi cellulari (F) : 326 molecole

Fingerprints



Miglior classificatore:
KNN (ROC AUC = 0,63)

Descrittori



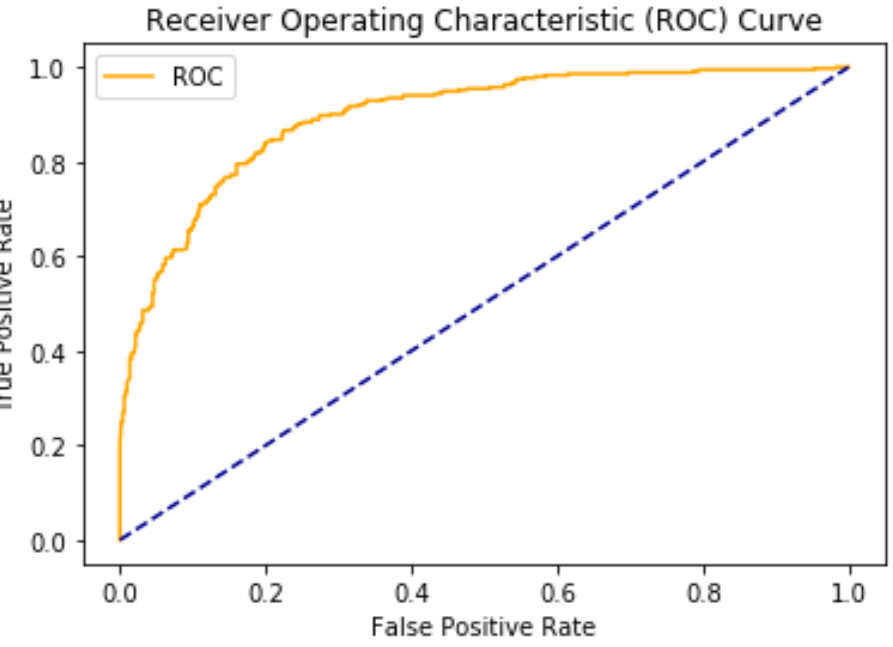
Miglior classificatore:
KNN (ROC AUC = 0,62)

Training set generato a partire da saggi biochimici (B) : 853 molecole

	Algoritmo	*Accuracy*	MCC	ROC AUC	F1 *score*
Fingerprints	Random Forest	0,81	0,65	0,80	0,75
	KNN	0,80	0,65	0,80	0,76
	GBM	0,80	0,64	0,79	0,75
	Extra Trees	0,80	0,64	0,80	0,76
Descrittori	Random Forest	0,80	0,64	0,79	0,76
	KNN	0,80	0,61	0,80	0,77
	GBM	0,79	0,61	0,77	0,73
	Extra Trees	0,83	0,64	0,82	0,78

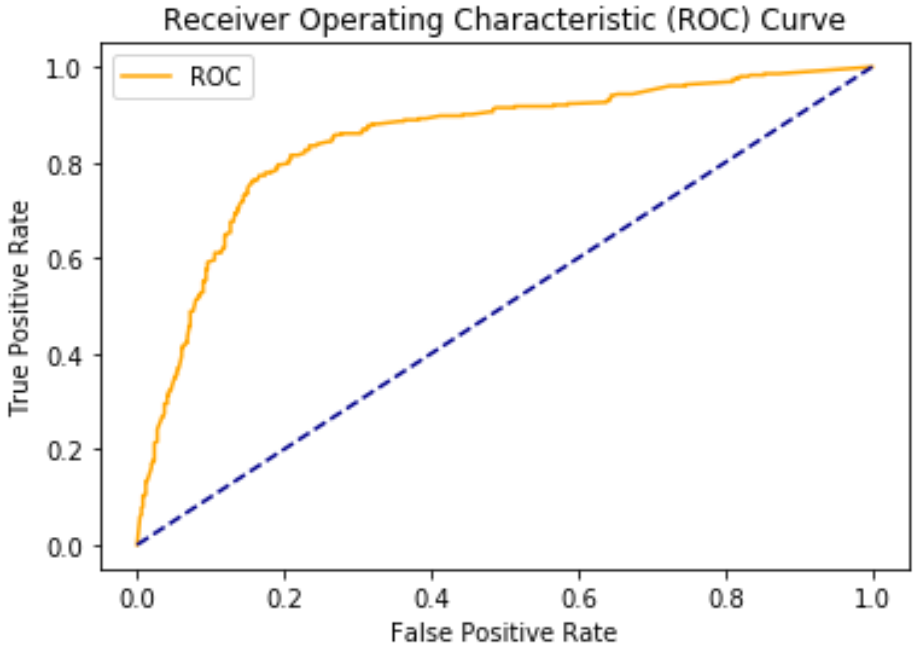
Training set generato a partire dai saggi biochimici (B) : 853 molecole

Fingerprints



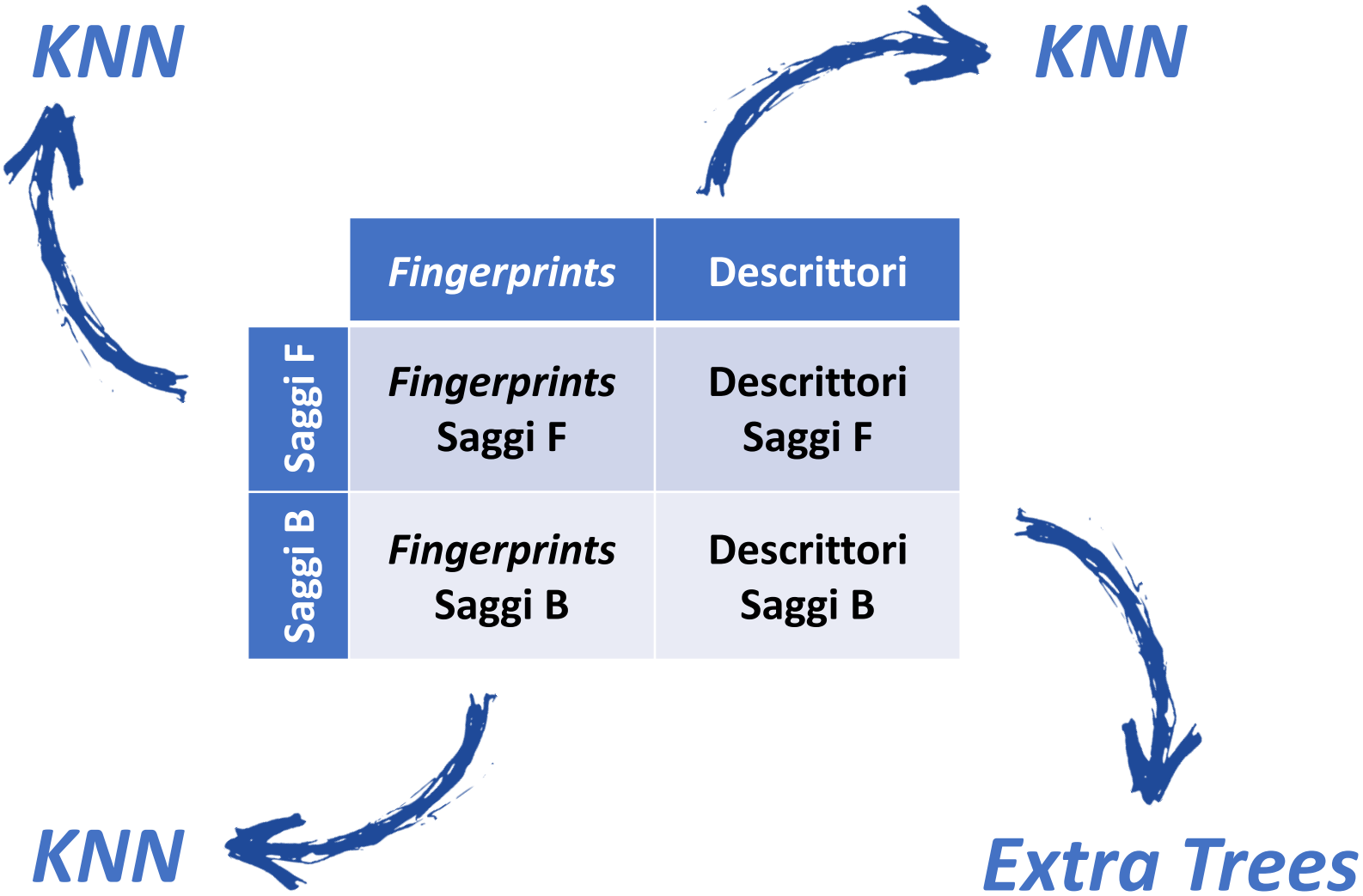
Miglior classificatore:
KNN (ROC AUC = 0,80)

Descrittori



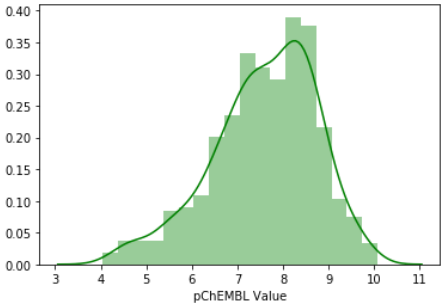
Miglior classificatore:
Extra Trees (ROC AUC = 0,80)

Step 2 – Classificazione: Risultati finali



	Fingerprints	Descrittori
Saggi F	***Fingerprints*** **Saggi F**	**Descrittori Saggi F**
Saggi B	***Fingerprints*** **Saggi B**	**Descrittori Saggi B**

Step 2 – Regressione: creazione modelli

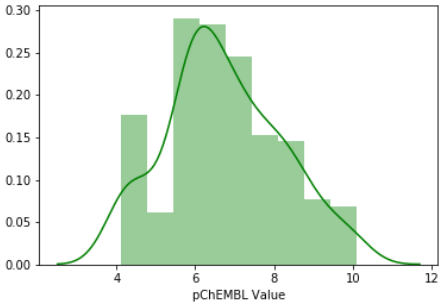


Dataset

Risposte biologiche espresse in:

pIC_{50}
624 molecole

pKi
196 molecole



Fingerprints

Descrittori Molecolari

Fingerprints

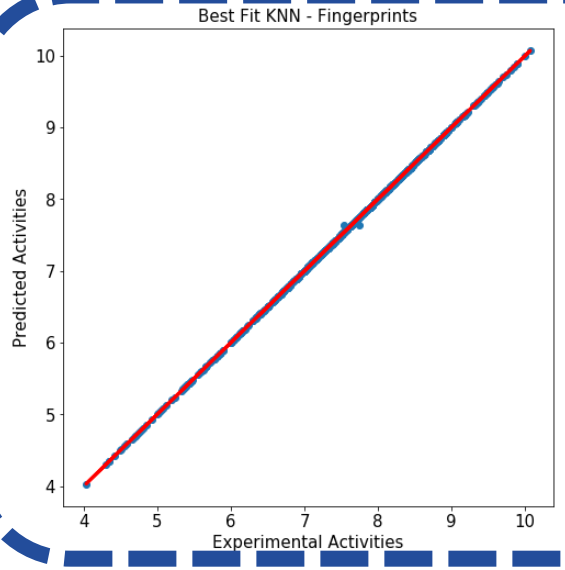
Descrittori Molecolari

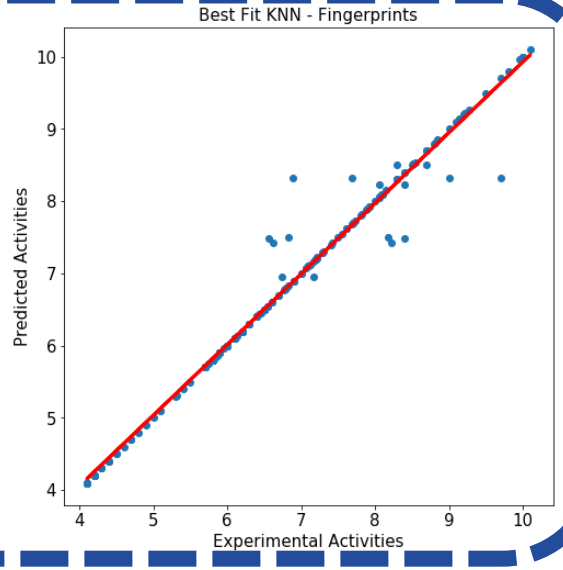
12 modelli di regressione

Training set generato utilizzando come *features* i *fingerprints*

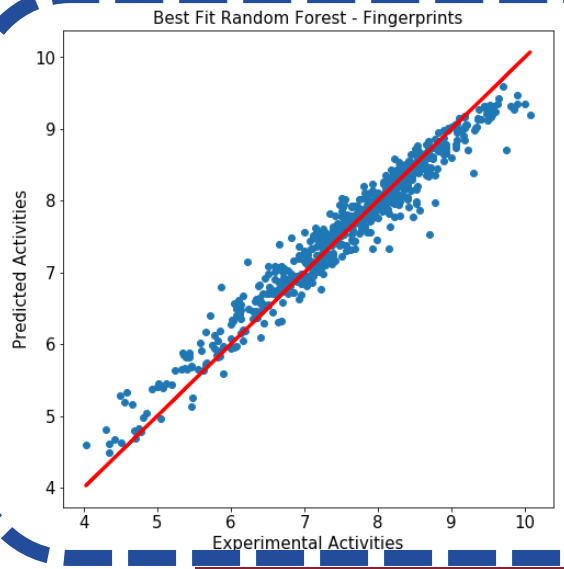
	Algoritmo	N	r^2	q^2	SDEC	SDEP
pIC_{50}	Random Forest	624	0,95	0,63	0,26	0,88
	KNN	624	0,99	0,64	0,005	1,17
	Ridge	624	0,93	0,64	0,31	0,92
pKi	Random Forest	196	0,95	0,52	0,34	0,92
	KNN	196	1	0,58	0,0	2,53
	Ridge	196	0,97	0,60	0,25	0,92

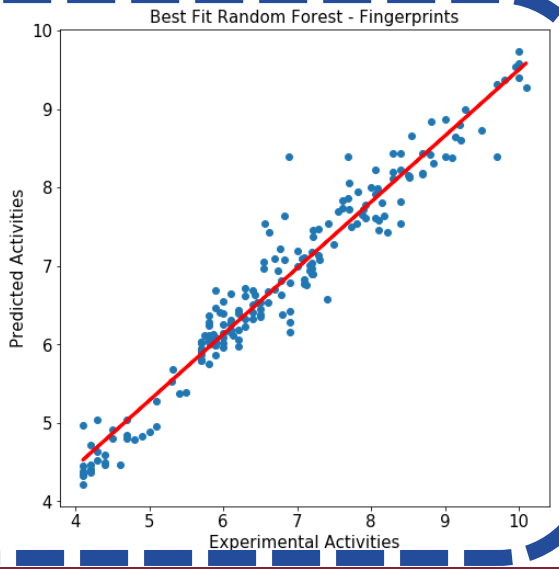
Step 2 – Regressione: Rette di *Best Fit (fingerprints)*





Overfitting!



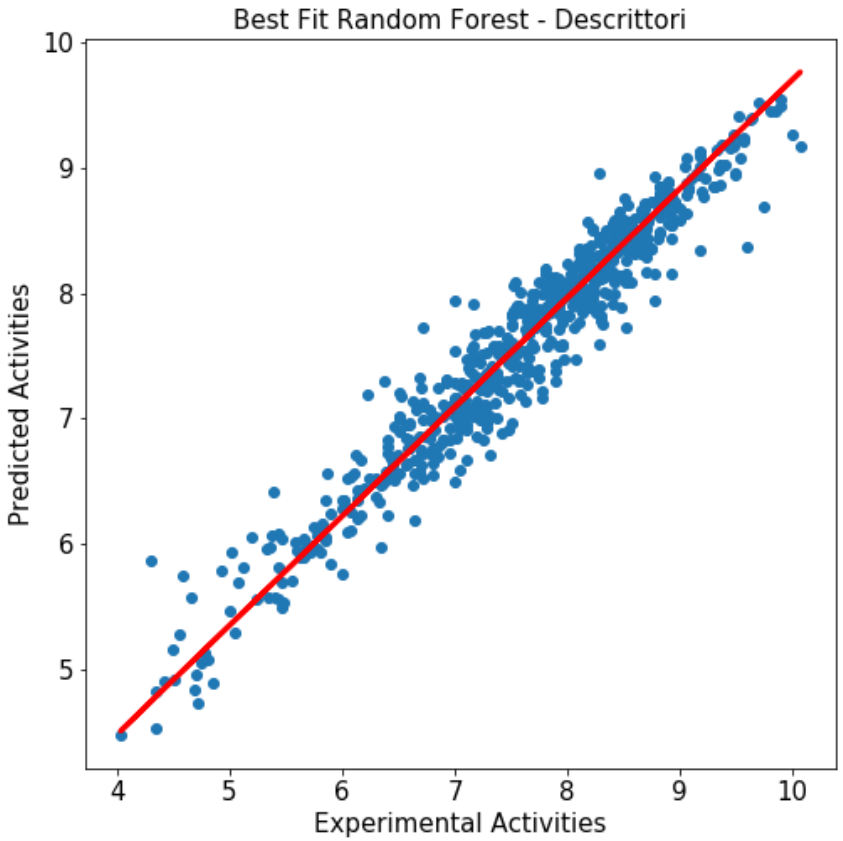


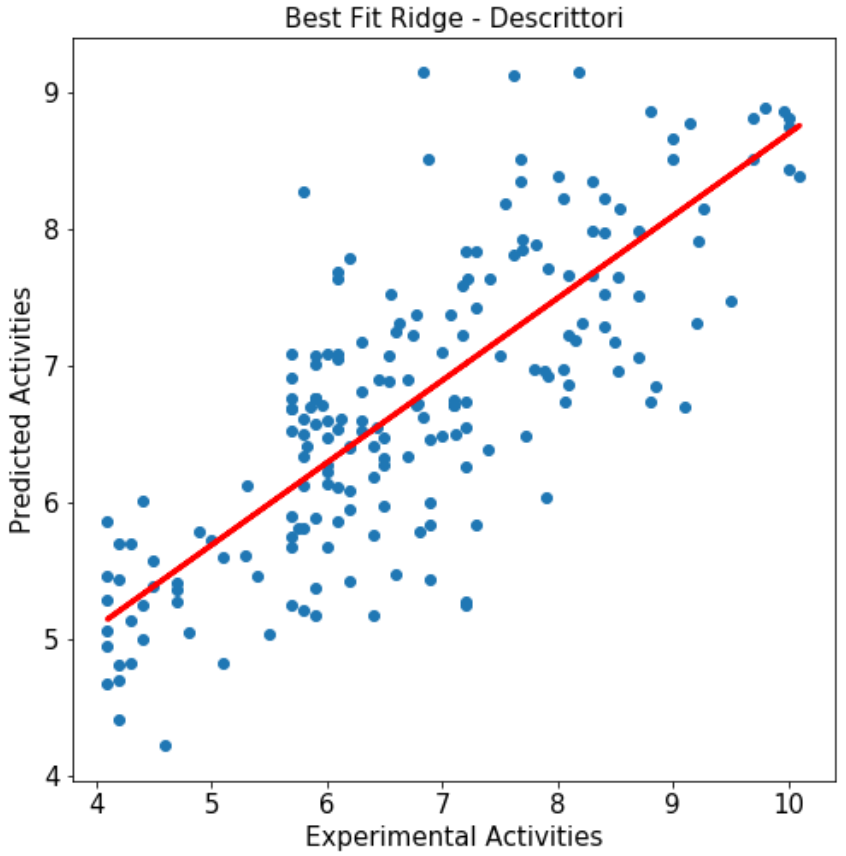
Buon modello!

Training set generato utilizzando come *features* i descrittori

	Algoritmo	N	r^2	q^2	SDEC	SDEP
pIC_{50}	Random Forest	624	0,95	0,50	0,25	0,88
	KNN	624	0,99	0,64	0,005	0,90
	Ridge	624	0,92	0,42	0,30	0,92
pKi	Random Forest	196	0,94	0,58	0,35	0,89
	KNN	196	1	0,61	0,21	0,88
	Ridge	196	0,94	0,55	0,33	0,86

Step 2 – Regressione: Rette di *Best Fit* (descrittori)





Step 2 – Regressione: Risultati finali

Random Forest

Random Forest

	Fingerprints	Descrittori
pIC_{50}	***Fingerprints*** **pIC50**	**Descrittori pIC50**
pKi	***Fingerprints*** **pKi**	**Descrittori pKi**

Random Forest

Ridge

Step 3 – Predizioni : Creazione del test set esterno

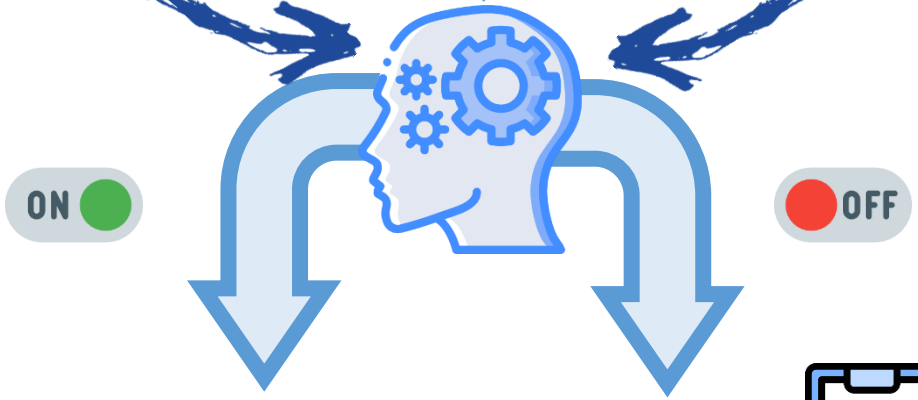
PubChem

Letteratura

Modelli di classificazione

Test set esterno (50 molecole)

Modelli di regressione



Validazione esterna

Training set generato a partire dai saggi cellulari (F) : 326 molecole

	Algoritmo	*Accuracy*	MCC	ROC AUC	F1 *score*
Fingerprints	Random Forest	0,41	0,23	0,67	0,52
	KNN	0,68	0,40	0,82	0,78
	GBM	0,27	0,15	0,50	0,31
	Extra Trees	0,63	0,33	0,52	0,21
Descrittori	Random Forest	0,56	0,31	0,75	0,68
	KNN	0,48	0,26	0,71	0,60
	GBM	0,52	0,28	0,73	0,63
	Extra Trees	0,56	0,32	0,75	0,68

Grazie per l'attenzione!