

3-D QSAR.

Validazione di una nuova metodologia per la costruzione automatica di modelli predittivi.

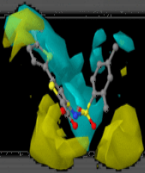


SAPIENZA
UNIVERSITÀ DI ROMA

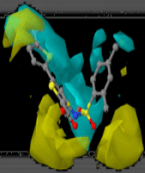
Facoltà di Farmacia e Medicina
Corso di Laurea in Chimica e Tecnologia Farmaceutiche
Tesi Sperimentale in Chimica Farmaceutica
a.a. 2010/2011

Laureanda : Roberta Fazi
Matricola: 1060903

Relatore: prof. Rino Ragno



- Validazione di un metodo freeware per la costruzione di modelli 3-D QSAR
- Ridefinizione regole allineamento
- Creazione di un database online

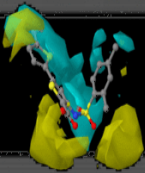


- Quantitative **S**tructure-**A**ctivity **R**elationship (QSAR)

$$\log P (1/C) = \log D + K (\Delta G_h + \Delta G_e + \Delta G_s + etc.) + \text{costante}$$

MIF
+
PLS

3-D QSAR



Comparative Molecular Field Analysis

J. Am. Chem. Soc. **1988**, *110*, 5959–5967

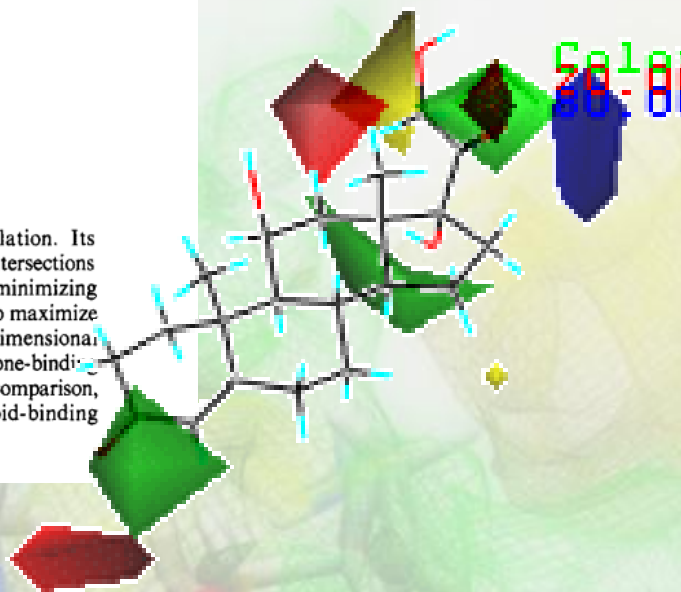
5959

Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins

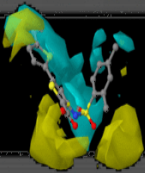
Richard D. Cramer, III,* David E. Patterson, and Jeffrey D. Bunce

Contribution from Tripos Associates, 1699 South Hanley Road, St. Louis, Missouri 63144. Received January 5, 1988

Abstract: Comparative molecular field analysis (CoMFA) is a promising new approach to structure/activity correlation. Its characteristic features are (1) representation of ligand molecules by their steric and electrostatic fields, sampled at the intersections of a three-dimensional lattice, (2) a new "field fit" technique, allowing optimal mutual alignment within a series, by minimizing the RMS field differences between molecules, (3) data analysis by partial least squares (PLS), using cross-validation to maximize the likelihood that the results have predictive validity, and (4) graphic representation of results, as contoured three-dimensional coefficient plots. CoMFA is exemplified by analyses of the affinities of 21 varied steroids to corticosteroid- and testosterone-binding globulins. Also described are the sensitivities of results to the nature of the field and the definition of the lattice and, for comparison, analyses of the same data using various combinations of other parameters. From these results, a set of ten steroid-binding affinity values unknown to us during the CoMFA analysis were well predicted.

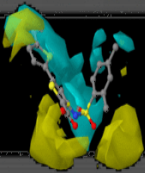


- ✓ Spaziatura nodi della griglia 2 Å
- ✓ Contributi sterici e elettrostatici
- ✓ Visualizzazione grafica mediante CoMFA Steric/Electrostatic contour maps



Obiettivi 3-D QSAR

- Spiegazione dei dati Biologici su base tridimensionale
- Ottimizzazione molecole esistenti
- Predizione composti non ancora saggiati



Costruzione di un modello

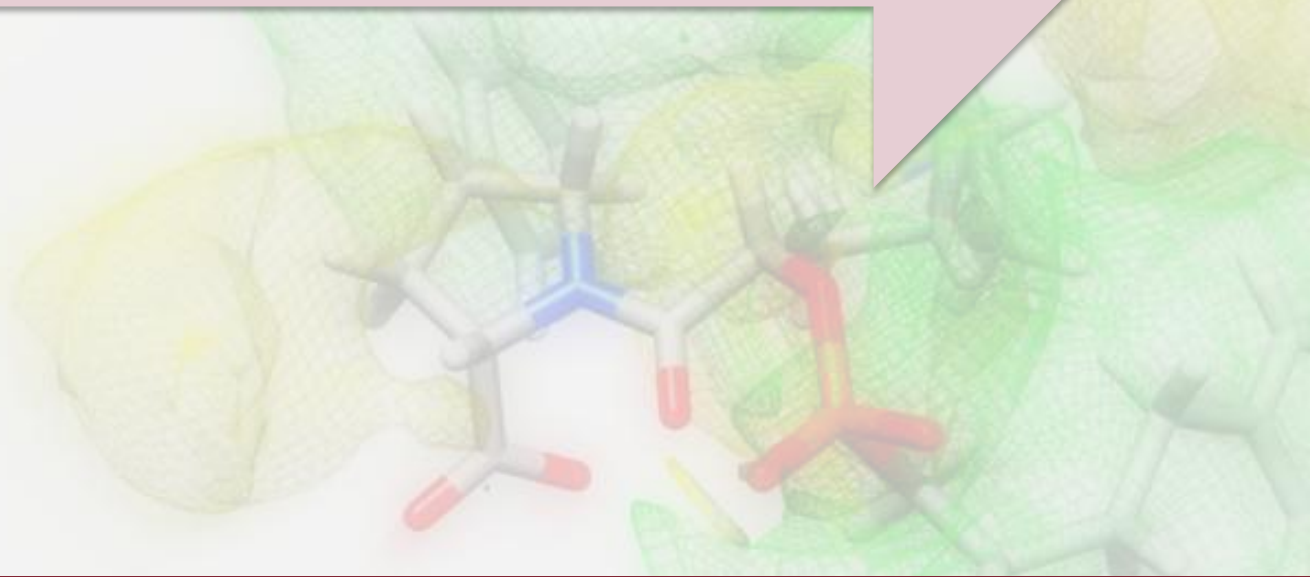
Scelta del
training set

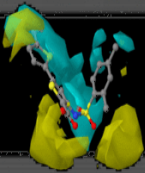
Allineamento
delle
molecole

Calcolo del
campi di
interazione
molecolare
(MIF)

Generazione
del modello
statistico

Validazione
interna,
esterna e
y-scrambling





Costruzione di un modello

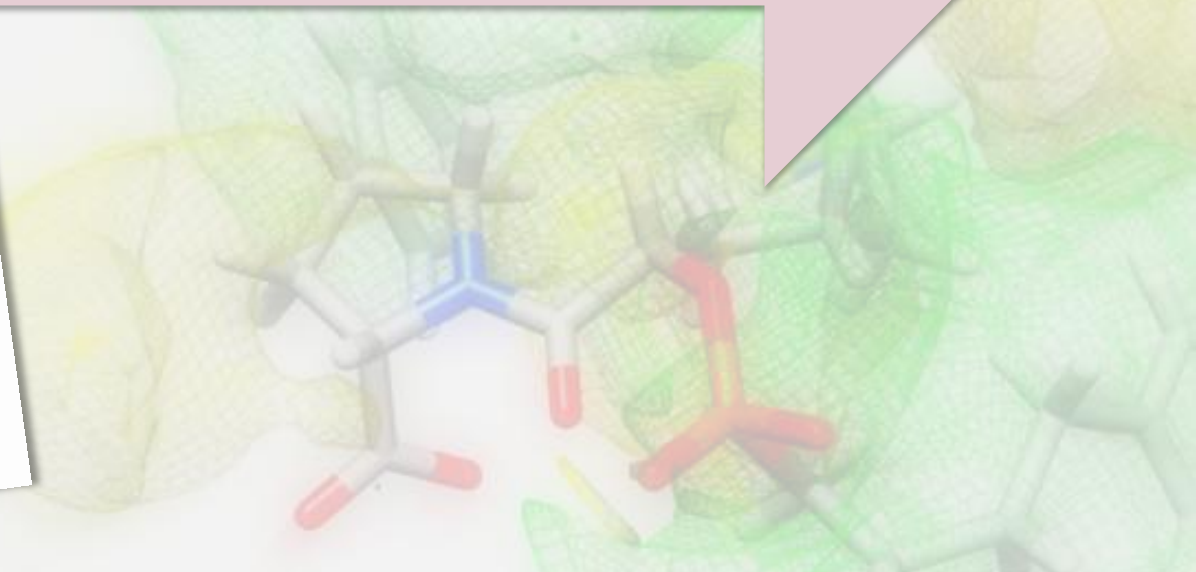
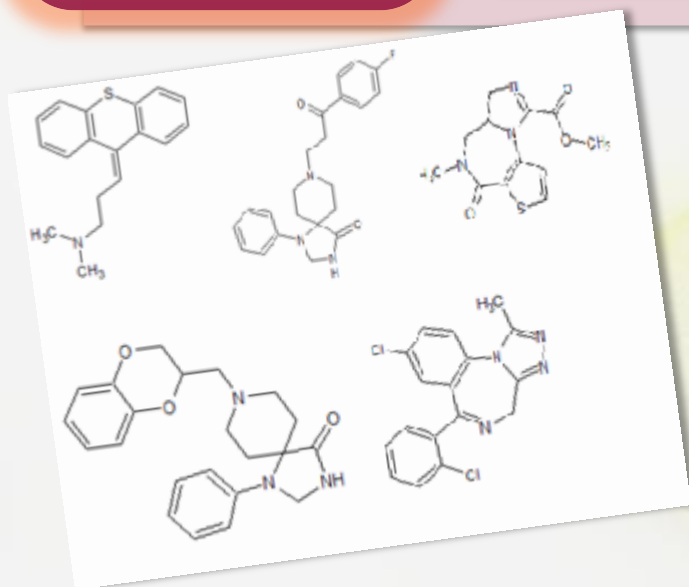
**Scelta del
training set**

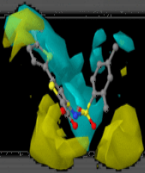
Allineamento
delle
molecole

Calcolo del
campi di
interazione
molecolare
(MIF)

Generazione
del modello
statistico

Validazione
interna,
esterna e
y-scrambling





Costruzione di un modello

Scelta del
training set

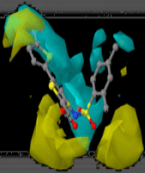
**Allineamento
delle molecole**

Calcolo del
campi di
interazione
molecolare
(MIF)

Generazione
del modello
statistico

Validazione
interna,
esterna e
y-scrambling





Costruzione di un modello

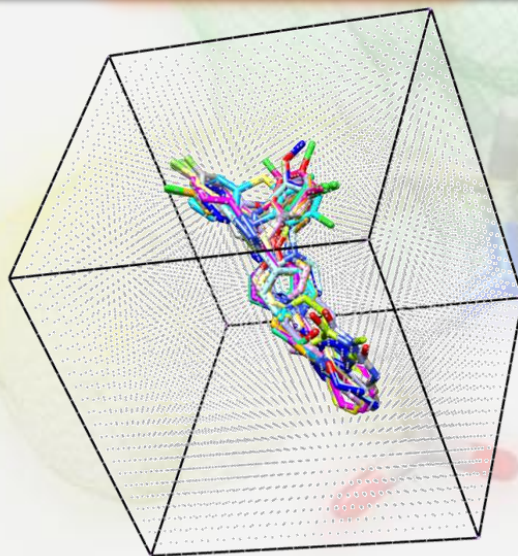
Scelta del
training set

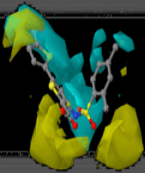
Allineamento
delle
molecole

**Calcolo dei
campi di
interazione
molecolare
(MIF)**

Generazione
del modello
statistico

Validazione
interna,
esterna e
y-scrambling





Costruzione di un modello

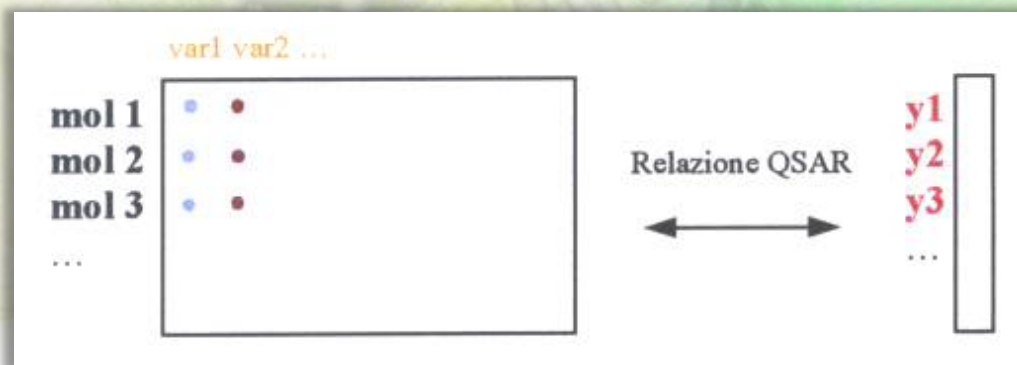
Scelta del
training set

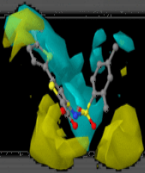
Allineamento
delle
molecole

Calcolo del
campi di
interazione
molecolare
(MIF)

**Generazione del
modello statistico**

Validazione
interna,
esterna e
y-scrambling





Costruzione di un modello

Scelta del
training set

Allineamento
delle
molecole

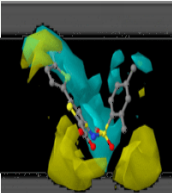
Calcolo del
campi di
interazione
molecolare
(MIF)

Generazione
del modello
statistico

**Validazione
interna, esterna e
y-scrambling**

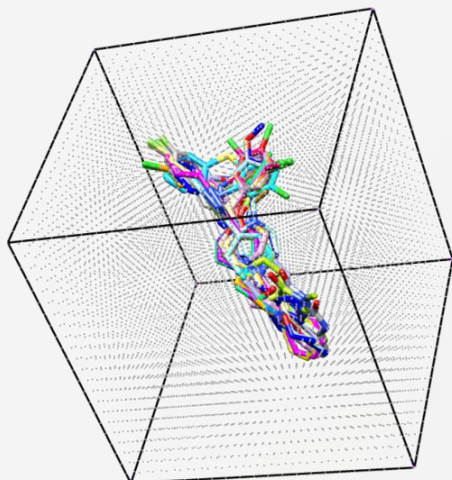
$$SDEP = \sqrt{\sum (y_{\text{exp}} - y_{\text{pred}})^2 / n - 1}$$

$$Q^2 = 1 - \sum (y_{\text{exp}} - y_{\text{pred}})^2 / \sum (y_{\text{exp}} - \bar{y})^2$$



AutoGrid

Calcolo dei Campi di Interazione Molecolare (MIF)



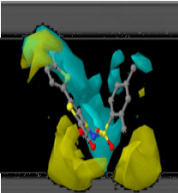
Otto differenti probe considerati
(A, C, HD, N, NA, OA, e, d)

Distanza dei nodi della griglia
definibile dall'utente
(default 1Å)

Analisi di regressione PLS (Partial Least Square)

- ✓ Modello Pretrattato
- ✓ Modello con selezione delle variabili (GA)
- ✓ Cross-Validazione (LOO, K-F5)
- ✓ Generazione contour maps CoMFA-like



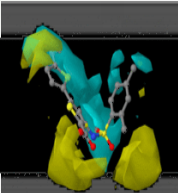


Dataset

Nome modello	Training Set	Test Set
ACE	76	38
AchE	74	37
BZR	98	49
GPB	44	22
COX2	188	94
DHFR	237	124
THERM	51	25
THR	59	29
ATA	72	22
ARB	28	//
CCR5	63	12
YOPH	35	4
KOA	31	8
MX	29	//
DAT	36	6

Nome modello	Training Set	Test Set
TP2A	24	//
CBRA	32	//
EDC	121	10
AI	78	//
HIVPR	93	20
GSK3B	34	8
STEROIDS	21	//
RYP	18	//
GHSM	31	//
D2R	32	6
D4R	32	6
PTC	40	//
DIAZEPAM	42	//
TAPAP2	72	16
AC	30	5

30 dataset diversi
2341 molecole suddivise in Training Set e Test Set

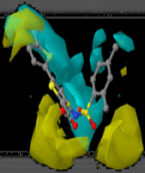


Risultati

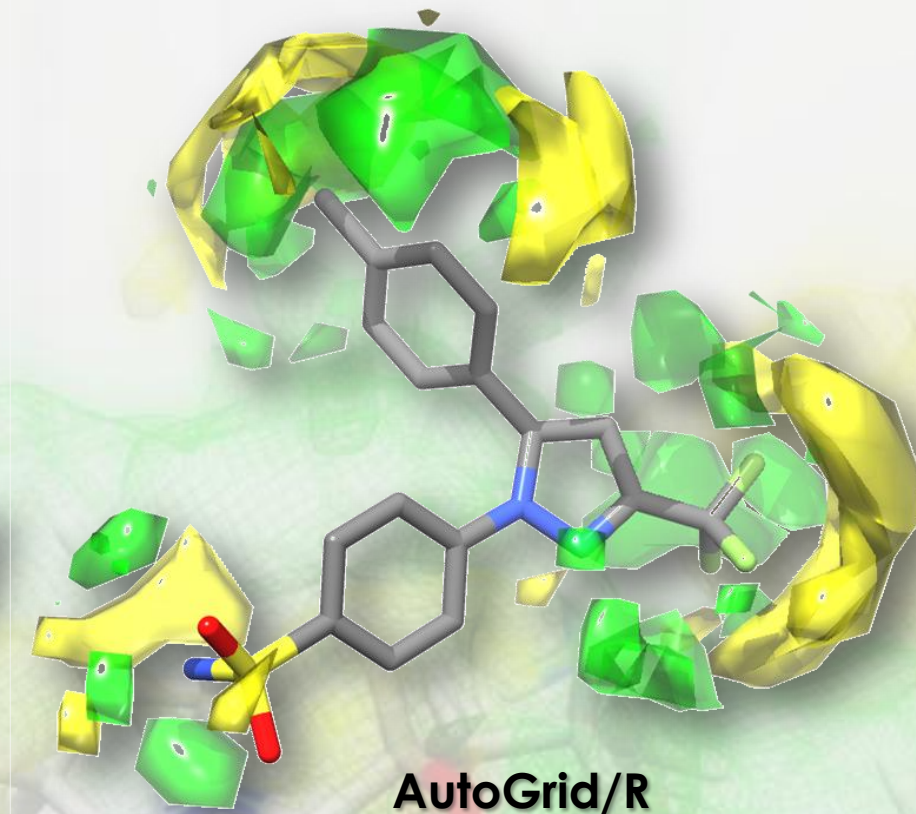
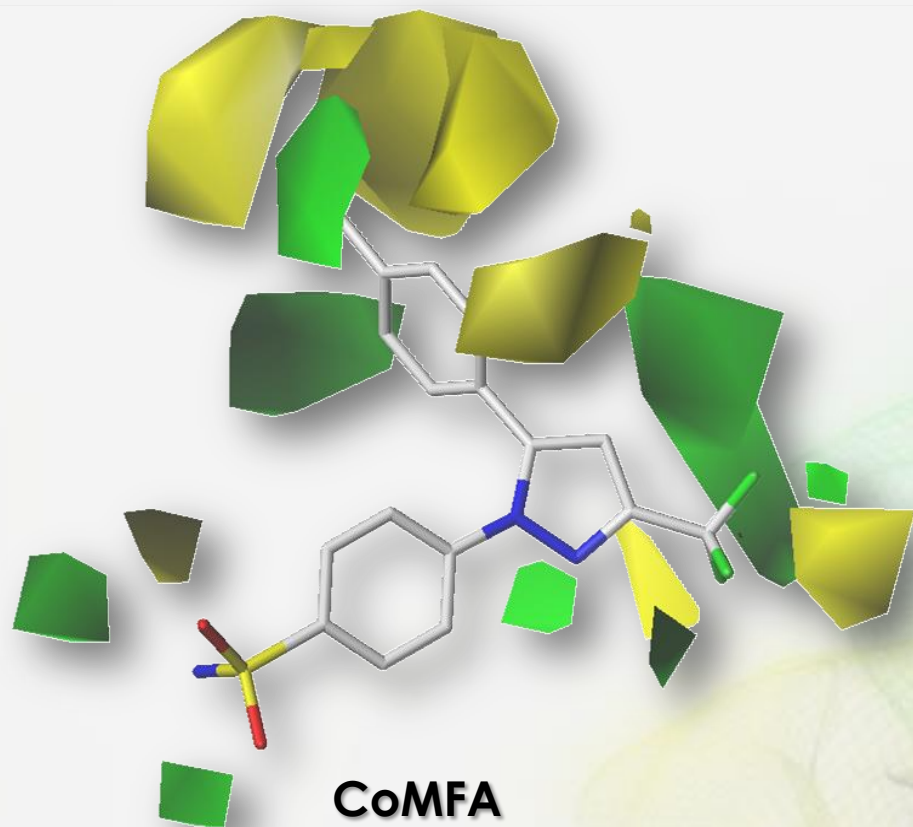
Nome modello	CoMFA	Autogrid/R pretrattato		Autogrid/R GA	
		q2	q2	PC	q2
ACE	0,644	0,64	2	0,708	3
AchE	0,544	0,40	3	0,532	3
BZR	0,375	0,34	3	0,441	3
GPB	0,141	0,34	4	0,661	5
COX2	0,358	0,41	5	0,479	3
DHFR	0,682	0,59	5	0,622	3
THERM	0,458	0,47	3	0,549	3
THR	0,595	0,48	3	0,606	3
ATA	0,115	0,16	3	0,403	3
ARB	0,406	0,35	3	0,616	3
CCR5	0,780	0,76	2	0,831	2
YOPH	0,771	0,76	1	0,839	2
KOA	0,695	0,62	4	0,702	3
MX	0,723	0,71	2	0,867	2
DAT	0,773	0,07	3	0,452	3
TP2A	0,644	0,64	2	0,708	3
CBRA	0,544	0,40	3	0,532	3

Nome modello	CoMFA	Autogrid/R pretrattato		Autogrid/R GA	
		q2	q2	PC	q2
EDC	0,413	0,35	1	0,506	3
AI	0,610	0,63	3	0,759	3
HIVPR	0,597	0,45	1	0,549	2
GSK3B	0,698	0,69	3	0,811	3
STEROIDS	0,76	0,49	4	0,663	3
RYR	0,466	neg		0,502	2
GHSN	0,306	0,32	1	0,634	4
D2R	0,757	0,62	4	0,732	3
D4R	0,491	0,46	3	0,546	3
PTC	0,768	0,67	3	0,813	4
DIAZEPAM DS/DI	0.79	0,73	3	0,817	3
DIAZEPAM DI	0.70	0.30	3	0,691	5
DIAZEPAM DS	0.73	0,41	3	0,596	4
TAPAP2 thrombin	0.613	0,58	3	0,686	3
TAPAP2 trypsin	0.636	0,56	2	0,692	3
TAPAP2 Xa	0.429	0,41	4	0,555	3
AC		0.35	5	0.412	4

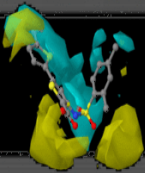
272 modelli generati



Confronto contour maps

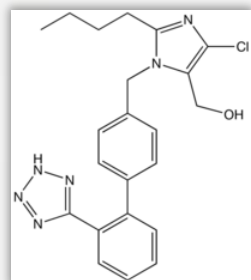
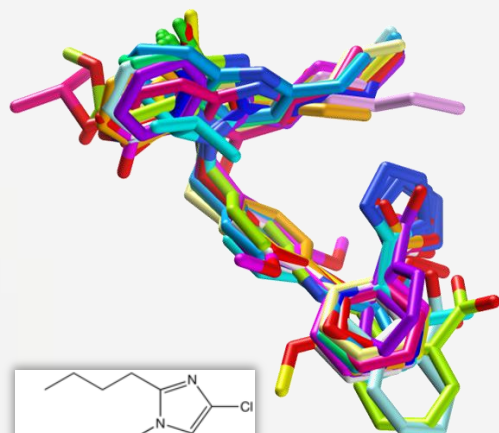


Le mappe realizzate con la procedura Autogrid/R risultano simili e comparabili a quelle ottenute con CoMFA.



Regole allineamento

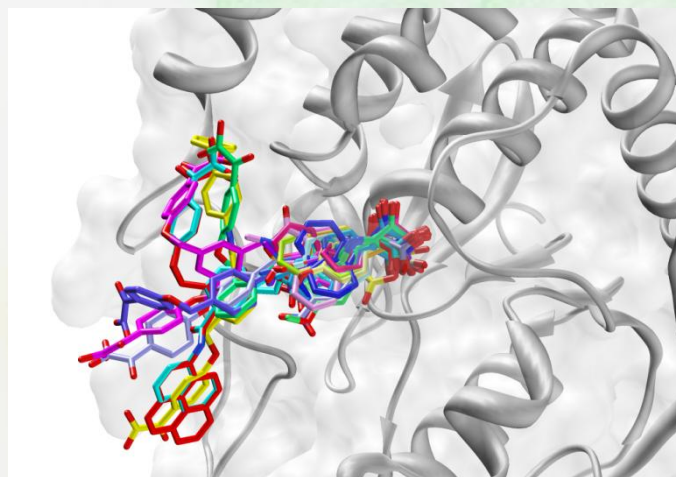
Allineamento Ligand-Based (atom by atom)



Losartan

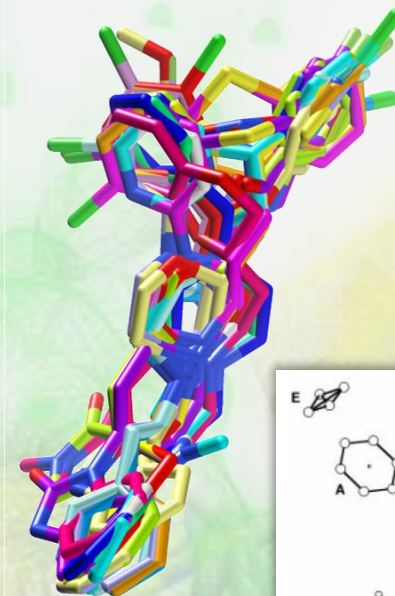
Modello
ARB

Allineamento Structure-Based

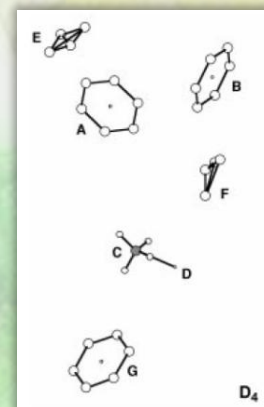


Modello **YOPH**

Allineamento Ligand-Based su farmacoforo

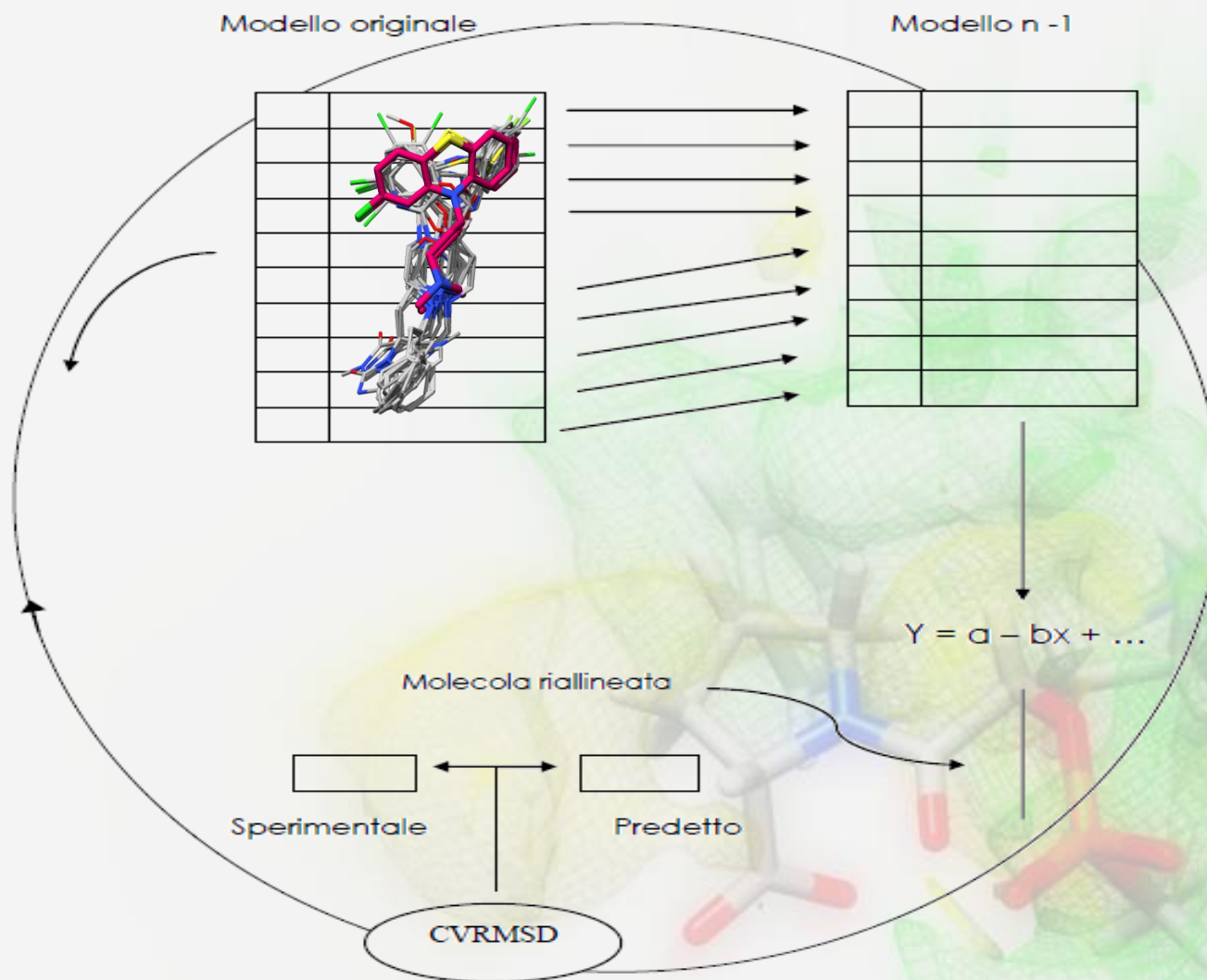


Modello **D2R**

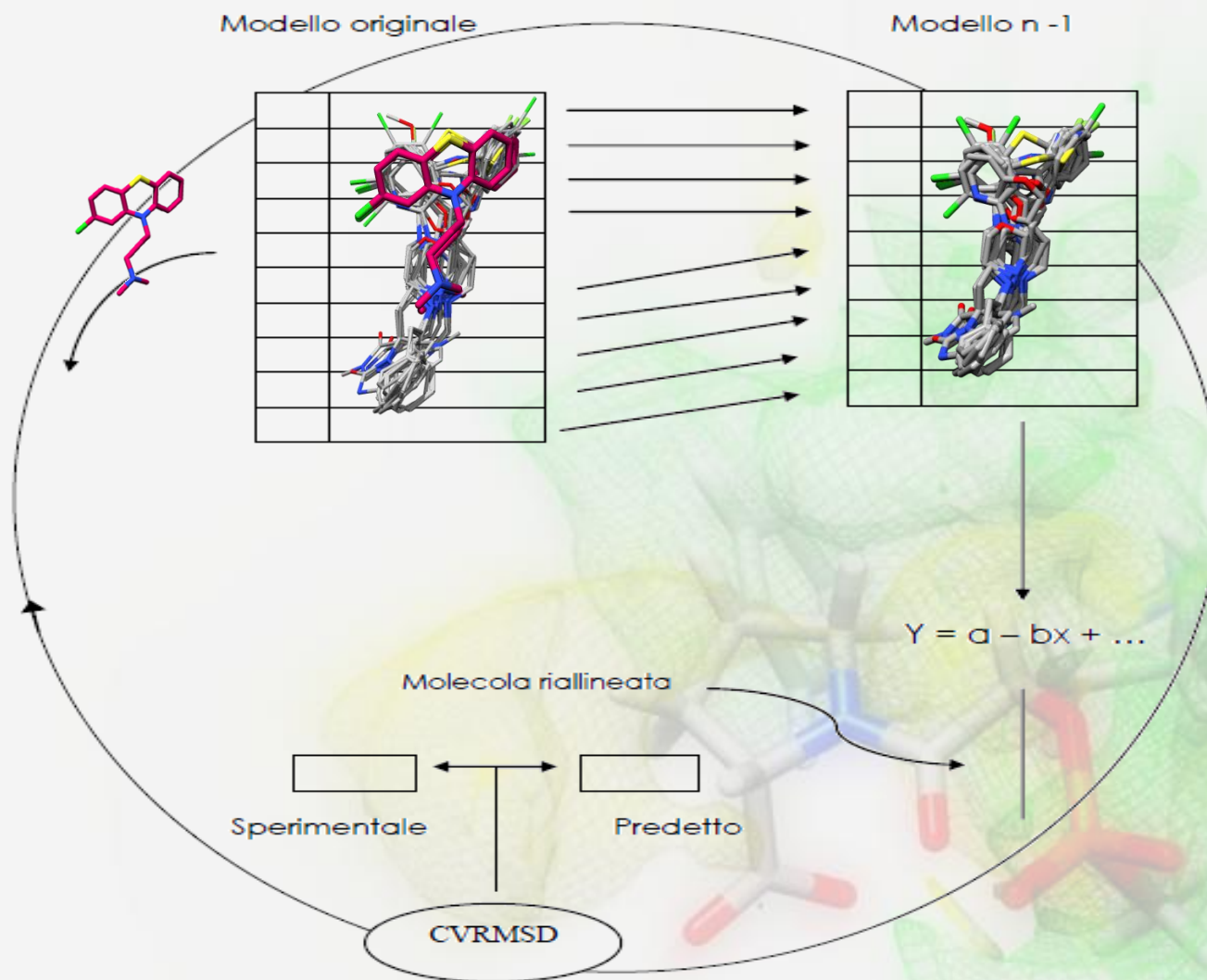




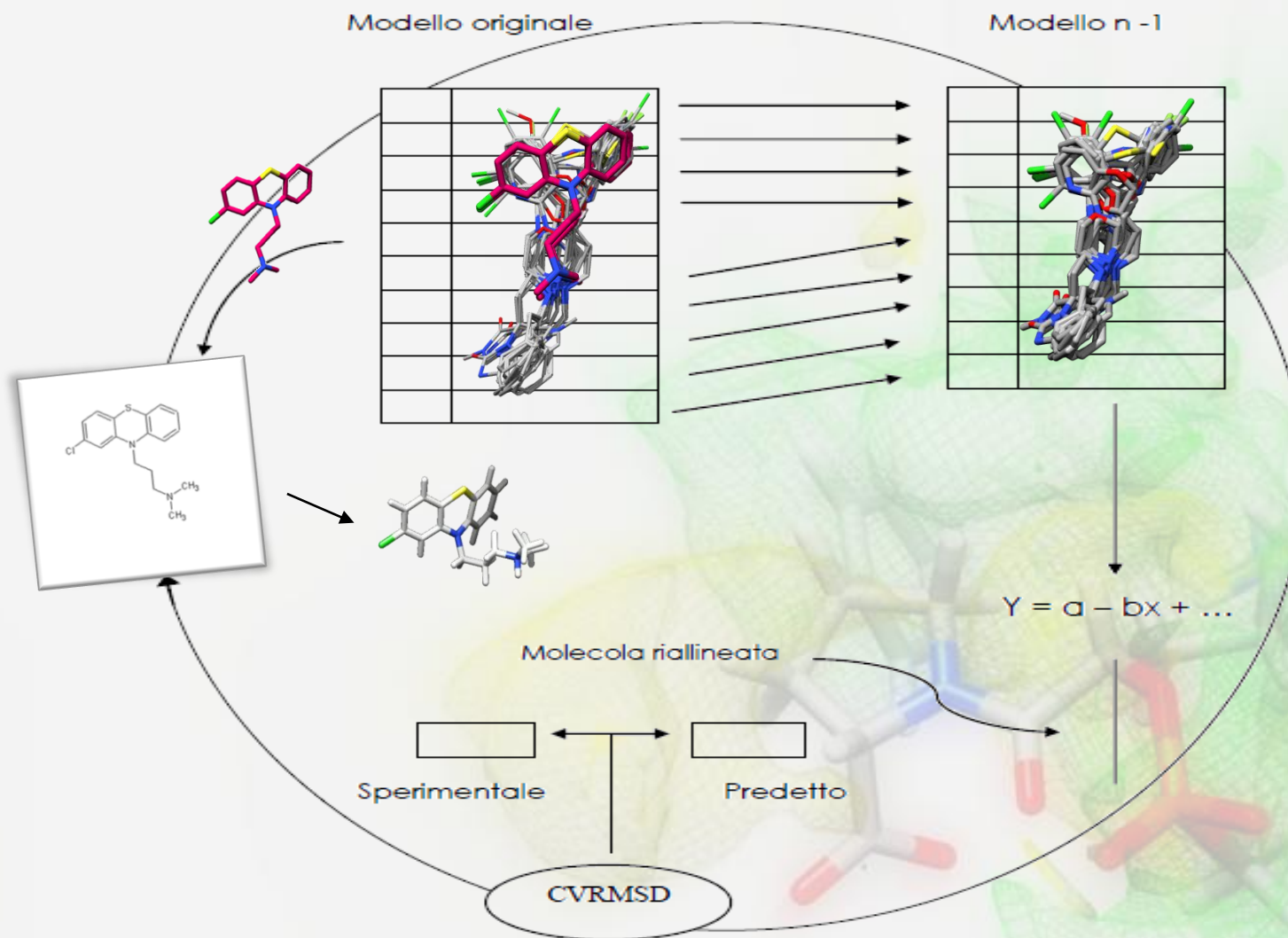
Ridefinizione regole allineamento



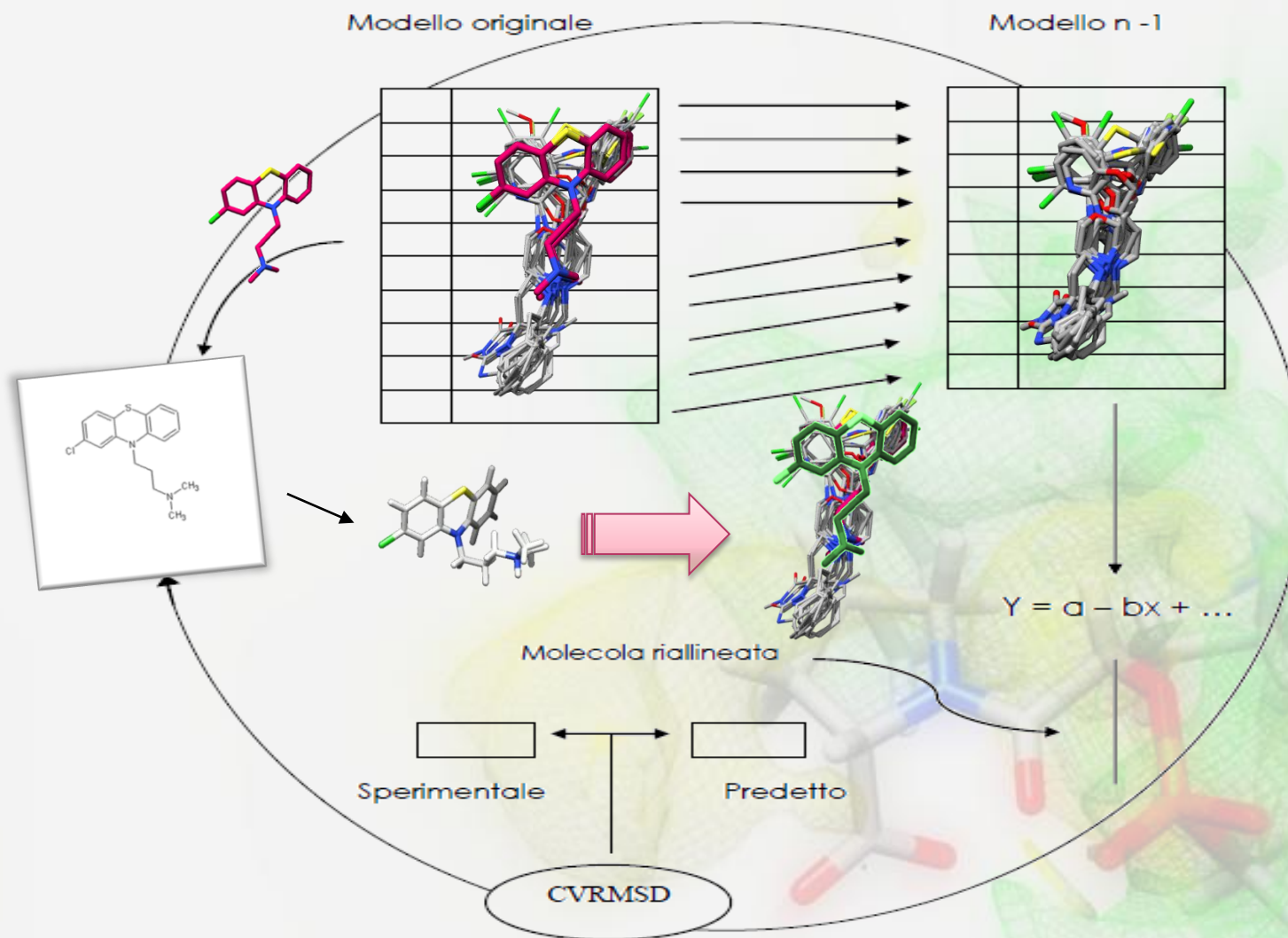
Ridefinizione regole allineamento

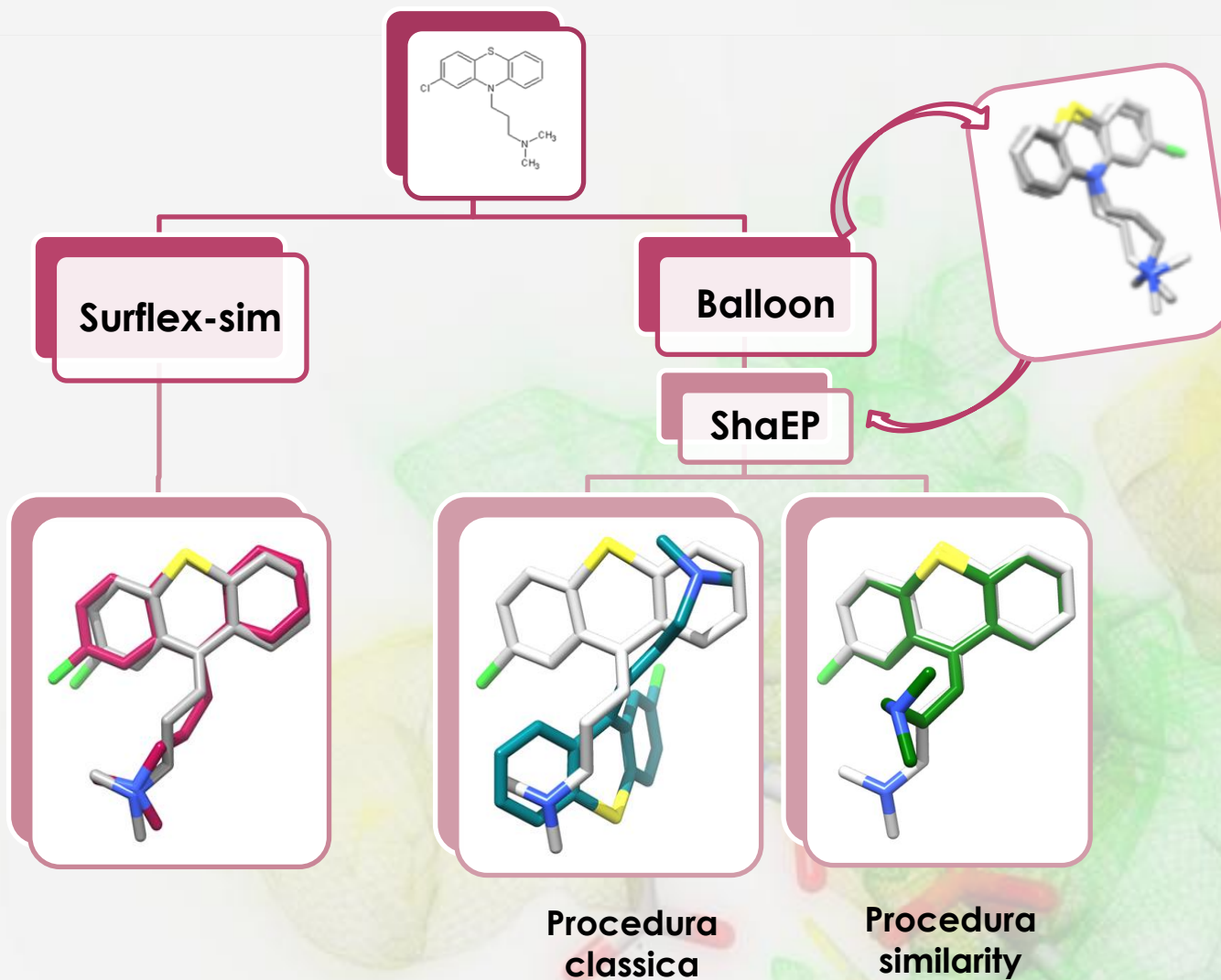


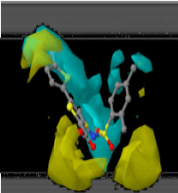
Ridefinizione regole allineamento



Ridefinizione regole allineamento



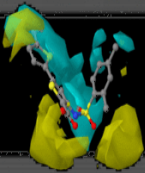




Risultati riallineamento

Modello	LOO		CVRMSD		Tipo di allineamento
	q2	PC	q2	PC	
ACE	0,67	3	0,61	2	Surflex-sim
AchE	0,40	3	0,35	2	Surflex-sim
BZR	0,34	3	0,15	2	Surflex-sim
GPB	0,34	4	0,22	2	Surflex-sim
COX2	0,37	4	0,19	2	Surflex-sim
ATA	0,16	3	0,22	4	Surflex-sim
ARB	0,42	3	0,27	1	Surflex-sim
CCR5	0,77	3	0,51	2	Surflex-sim
YOPH	0,79	2	0,69	2	Surflex-sim
KOA	0,62	4	0,59	2	Surflex-sim
MX	0,76	4	0,83	3	ShaEP-sim
TP2A	0,37	2	0,05	4	Surflex-sim

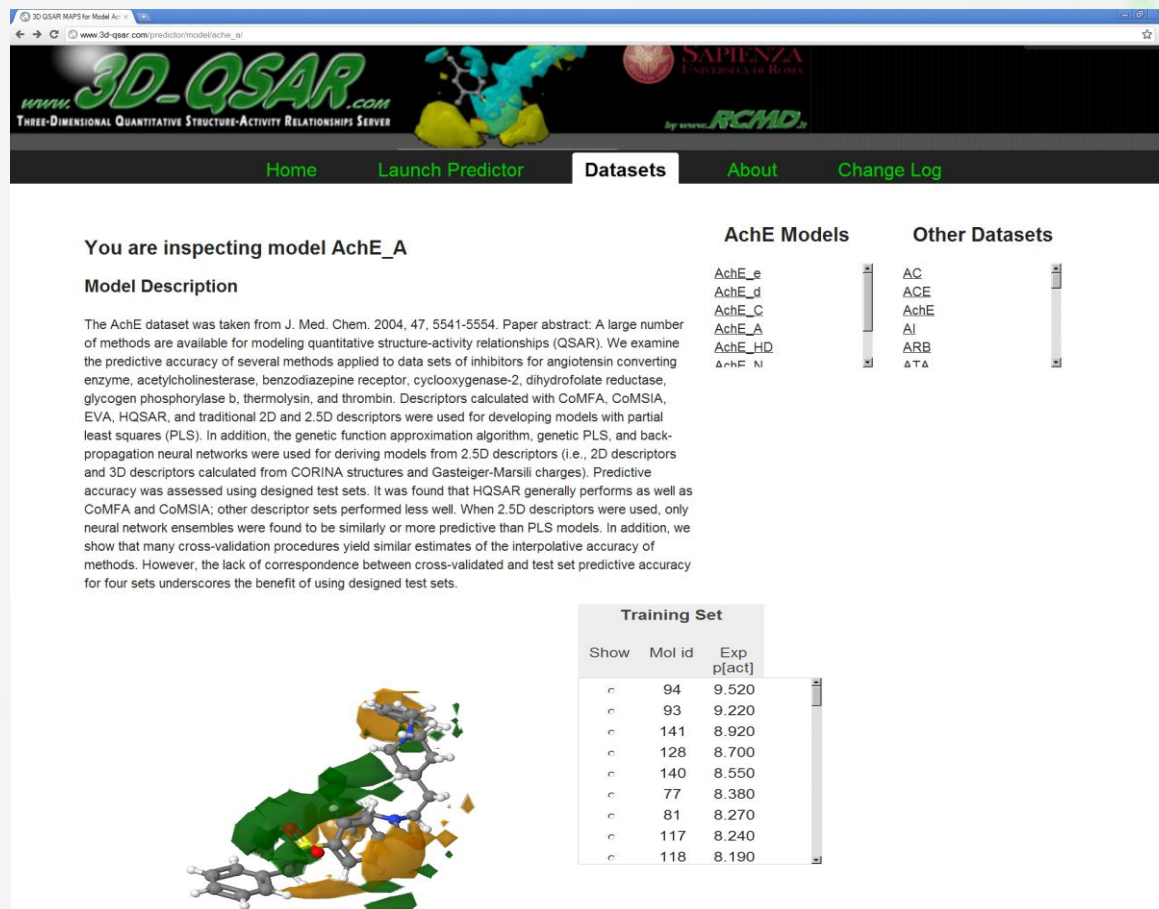
Modello	LOO		CVRMSD		Tipo di allineamento
	q2	PC	q2	PC	
CBRA	0,61	2	0,56	2	Surflex-sim
AI	0,64	3	0,57	3	Surflex-sim
HIVPR	0,47	2	0,37	1	Surflex-sim
GSK3B	0,73	3	0,70	3	Surflex-sim
D2R	0,36	4	0,48	2	Surflex-sim
D4R	0,68	4	0,16	2	Surflex-sim
DIAZEPAM DS/DI	0,51	3	0,58	3	Surflex-sim
DIAZEPAM DI	0,81	3	0,11	1	ShaEP-sim
DIAZEPAM DS	0,42	3	0,39	2	Surflex-sim
TAPAP2_thrombin	0,49	2	//		none
TAPAP2_trypsin	0,58	4	0,23	1	Surflex-sim
TAPAP2_Xa	0,43	3	0,22	2	Surflex-sim



- Efficacia e robustezza del metodo Autogrid/R
 - Validato su 272 modelli
- Allineamento di molecole esterne
 - Metodo CVRMSD
- Creazione di un database online di modelli 3-D QSAR
 - In grado di predire l'attività biologica di molecole incognite

Conclusioni

www.3d-qsar.com



The screenshot displays the 3D-QSAR website interface. At the top, there is a navigation bar with links: Home, Launch Predictor, Datasets, About, and Change Log. The main content area is titled "You are inspecting model AchE_A". Below this, the "Model Description" section provides a detailed paragraph about the AchE dataset, its source (J. Med. Chem. 2004, 47, 5541-5554), and the various methods and descriptors used for modeling. To the right of the description, there are two lists: "AchE Models" and "Other Datasets". The "AchE Models" list includes AchE_e, AchE_d, AchE_C, AchE_A, AchE_HD, and AchE_N. The "Other Datasets" list includes AC, ACE, AchE, AI, ARB, and ATΔ. Below the description, there is a "Training Set" table with columns for "Show", "Mol id", and "Exp p[act]". The table lists 10 molecules with their IDs and predicted activity values. At the bottom left of the page, there is a 3D molecular model of a ligand bound to a protein, with a green surface representing the protein's binding pocket.

You are inspecting model AchE_A

Model Description

The AchE dataset was taken from J. Med. Chem. 2004, 47, 5541-5554. Paper abstract: A large number of methods are available for modeling quantitative structure-activity relationships (QSAR). We examine the predictive accuracy of several methods applied to data sets of inhibitors for angiotensin converting enzyme, acetylcholinesterase, benzodiazepine receptor, cyclooxygenase-2, dihydrofolate reductase, glycogen phosphorylase b, thermolysin, and thrombin. Descriptors calculated with CoMFA, CoMSIA, EVA, HQSAR, and traditional 2D and 2.5D descriptors were used for developing models with partial least squares (PLS). In addition, the genetic function approximation algorithm, genetic PLS, and back-propagation neural networks were used for deriving models from 2.5D descriptors (i.e., 2D descriptors and 3D descriptors calculated from CORINA structures and Gasteiger-Marsili charges). Predictive accuracy was assessed using designed test sets. It was found that HQSAR generally performs as well as CoMFA and CoMSIA; other descriptor sets performed less well. When 2.5D descriptors were used, only neural network ensembles were found to be similarly or more predictive than PLS models. In addition, we show that many cross-validation procedures yield similar estimates of the interpolative accuracy of methods. However, the lack of correspondence between cross-validated and test set predictive accuracy for four sets underscores the benefit of using designed test sets.

AchE Models

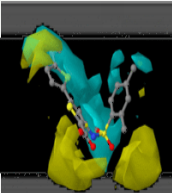
- AchE_e
- AchE_d
- AchE_C
- AchE_A
- AchE_HD
- AchE_N

Other Datasets

- AC
- ACE
- AchE
- AI
- ARB
- ATΔ

Training Set

Show	Mol id	Exp p[act]
☐	94	9.520
☐	93	9.220
☐	141	8.920
☐	128	8.700
☐	140	8.550
☐	77	8.380
☐	81	8.270
☐	117	8.240
☐	118	8.190



Grazie per l'attenzione
Ringrazio il Prof. Ragno e l'RCMD.

