SOCIETÀ ITALIANA DI TOSSICOLOGIA

SITOX

1972

21° Congresso Nazionale

Società Italiana di Tossicologia

Pericolo, rischio e rapporto rischio-beneficio

BOLOGNA

20-22 Febbraio 2023

www.sitox.org

Stato dell'arte sugli strumenti *in silico* per la predizione di tossicità di proteine

Luca Palazzolo e Ivano Eberini
Laboratorio di Biochimica e Biofisica Computazionale
Dipartimento di Scienze Farmacologiche e Biomolecolari
Università degli Studi di Milano

Domanda:
Quali sono gli strumenti e le metodologie attualmente disponibili per prevedere la potenziale attività delle proteine? Come vengono applicate per classificare le proteine come tossine o non tossine, con particolare attenzione nel contesto degli alimenti e dei mangimi?

A causa della natura descrittiva della domanda, gli elementi chiave Popolazione (P) e Risultato (O) sono definiti come segue :

Popolazione	Insieme di letteratura identificata su uno specifico strumento/ metodologia
Risultato	Lo strumento/metodologia è utile per prevedere se una proteina può essere classificata come positiva (tossina) o negativa (non tossica) (Sì/Nessun risultato)

Attività	Stringa di ricerca	Numero di risultati
Ricerca di letteratura per i *tool* di predizione di tossine	• english[Language] AND "last 10 years"[dp] AND ("mechanism of action" OR moa OR "mode of action" OR toxic) AND (protein OR peptide) AND (prediction OR "in silico" OR in-silico OR computational OR predictive) AND (tool OR software OR application OR program OR server)	4,245

Attività	Stringa di ricerca	Numero di risultati
Ricerca di letteratura per le metodologie *in silico*	• english[Language] AND "last 10 years"[dp] AND ("protein modeling" OR "protein modelling") AND protein AND ("mode of action" OR moa OR toxin) AND prediction	8
	• english[Language] AND "last 10 years"[dp] AND ("support vector machine" OR svm) AND protein AND ("mode of action" OR moa OR toxin) AND prediction	25
	• english[Language] AND "last 10 years"[dp] AND ("machine learning" OR ml) AND protein AND ("mode of action" OR moa OR toxin) AND prediction	65
	• english[Language] AND "last 10 years"[dp] AND ("hidden markov model" OR hmm) AND protein AND ("mode of action" OR moa OR toxin) AND prediction	6
	• english[Language] AND "last 10 years"[dp] AND "artificial intelligence" AND protein AND ("mode of action" OR moa OR toxin) AND prediction	571

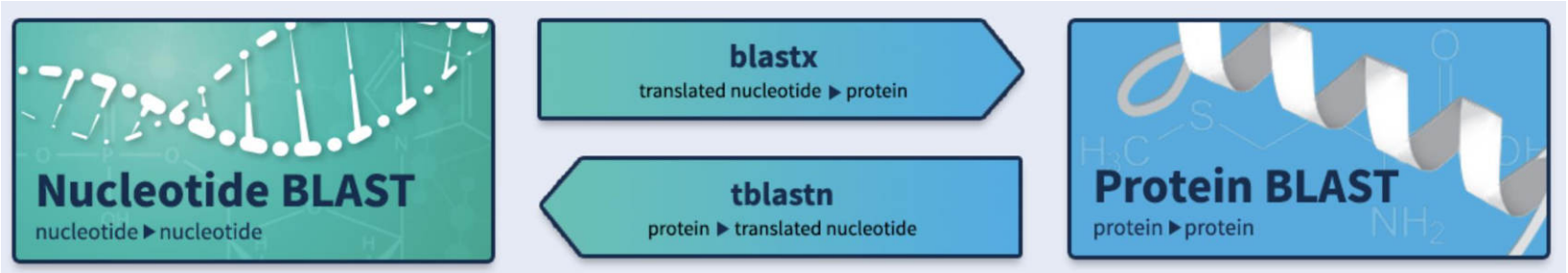
Tool	Campo di applicazione	Training/test datasets
BTXPred	Tossine batteriche	TP: 150 tossine batteriche TN: 500 proteine non tossiche
NTXPred	Neurotossine	TP: 582 neurotossine TN: 582 proteine non tossiche
PredCSF	Conotossine	TP: 261 conotossine TN: 60 proteine non tossiche ricche di cisteine
ToxinPred	Tossine	TP (main): 1805 peptidi tossici TN (main): 3593 proteine non tossiche
ToxClassifier	Tossine contenute nel veleno animale	TP: 8093 tossine contenute nel veleno animale TN (*easy*): 47144 proteine non tossiche TN (*moderate*): 8034 proteine non tossiche Tn (*difficult*): 7403 proteine non tossiche

Tool	Campo di applicazione	Training/test datasets
NNTox	Tossine	TP: 488 tossine TN: 6497 proteine non tossiche
TOXIFY	Tossine contenute nel veleno animale	TP: 4808 tossine contenute nel veleno animale TN: 32391 proteine non contenute nel veleno animale
ToxDL	Tossine	TP: 6164 tossine TN_1: 6164 proteine non tossiche TN_2: 903 proteine contenute nel veleno animale
ToxIBTL	Tossine	TP (1): 4472 tossine animali TN (1): 6341 proteine animali non tossiche TP (2): 3864 peptidi tossici TN (2): 3864 peptide non tossici
ToxinPred2	Tossine	TP (Main): 8233 tossine TN (Main): 8233 proteine non tossiche TP (Alternate): 1924 tossine TN (Alternate): 1924 proteine non tossiche

Tools		BTXPred	NTXPred	PredCSF	ToxinPred	ToxClassifier	NNTox	TOXIFY	ToxDL	ToxIBTL	ToxinPred2
Methods	PSI-BLAST	X	X	X		X	X				X
	PSIPRED	X		X							X
	HMMER	X				X					X
	CLUSTAL-W	X									X
	SVM	X	X	X	X	X					X
	NN		X			X	X	X	X	X	X
Domains	InterPro								X		
Motifs	MEME		X		X				X		
	LOGO				X						
	MAST		X								
	MERCI										X
	TOMTOM								X		
Year		2007	2007	2011	2013	2016	2019	2019	2021	2022	2022

BLAST

Il Basic Local Alignment Search Tool (BLAST) trova regioni di **somiglianza locale tra sequenze**. Il programma confronta le sequenze della proteina ai database di sequenza e ne calcola la **significatività statistica**. BLAST può essere usato per inferire le relazioni funzionali ed evolutive fra le sequenze come pure per contribuire ad identificare i membri delle famiglie del gene.



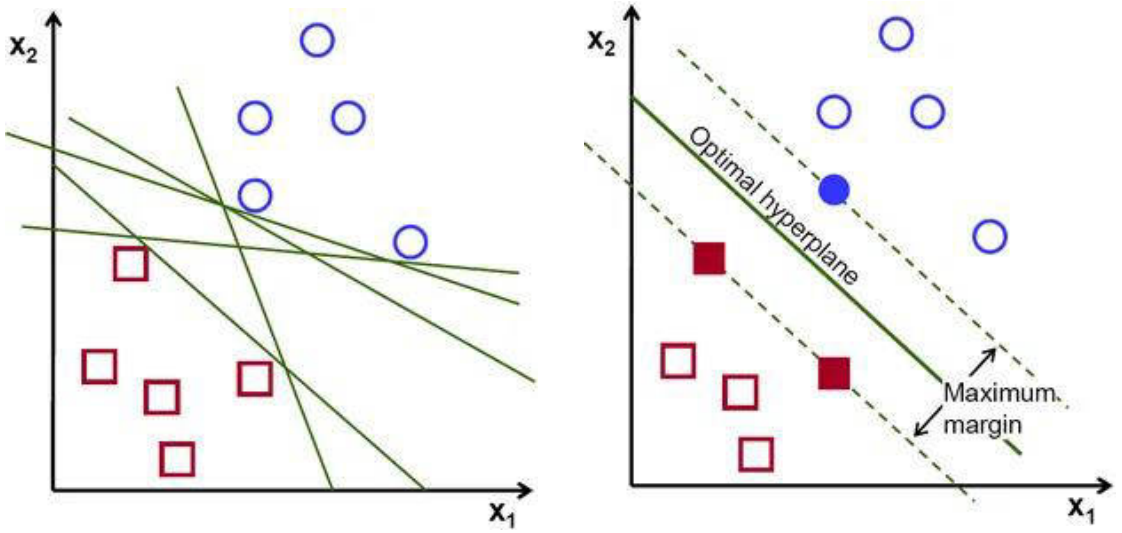
Hidden Markov Model

Questi modelli statistici hanno avuto molto successo con l'applicazione alla bioinformatica per il **riconoscimento di sequenze simili**. Nel caso dell'analisi delle proteine, si sa che due proteine con la stessa funzione possono essere differenti nella sequenza di aminoacidi perché ci sono state delle inserzioni o delle cancellazioni o delle sostituzioni di aminoacidi simili; **il problema è stabilire se due sequenze simili possono essere la stessa proteina**. Tra gli stati nascosti si definiscono gli stati di inserzione, cancellazione, sostituzione di un aminoacido e le probabilità di transizione verso questi stati che permette di stabilire quanto sono **simili due proteine**.

Il successo nella bioinformatica è dato dalla possibilità di gestire sequenze di lunghezza variabile, dalla possibilità di eseguire **allineamenti, analisi strutturali e rilevazione di pattern**, dal fatto che il modello statistico è molto robusto ed efficiente.

SVM

L'obiettivo dell'algoritmo della macchina vettoriale di supporto (Support Vector Machine) è quello di trovare un iperpiano in uno spazio N-dimensionale (N - il numero di caratteristiche) che classifica distintamente i punti dati.



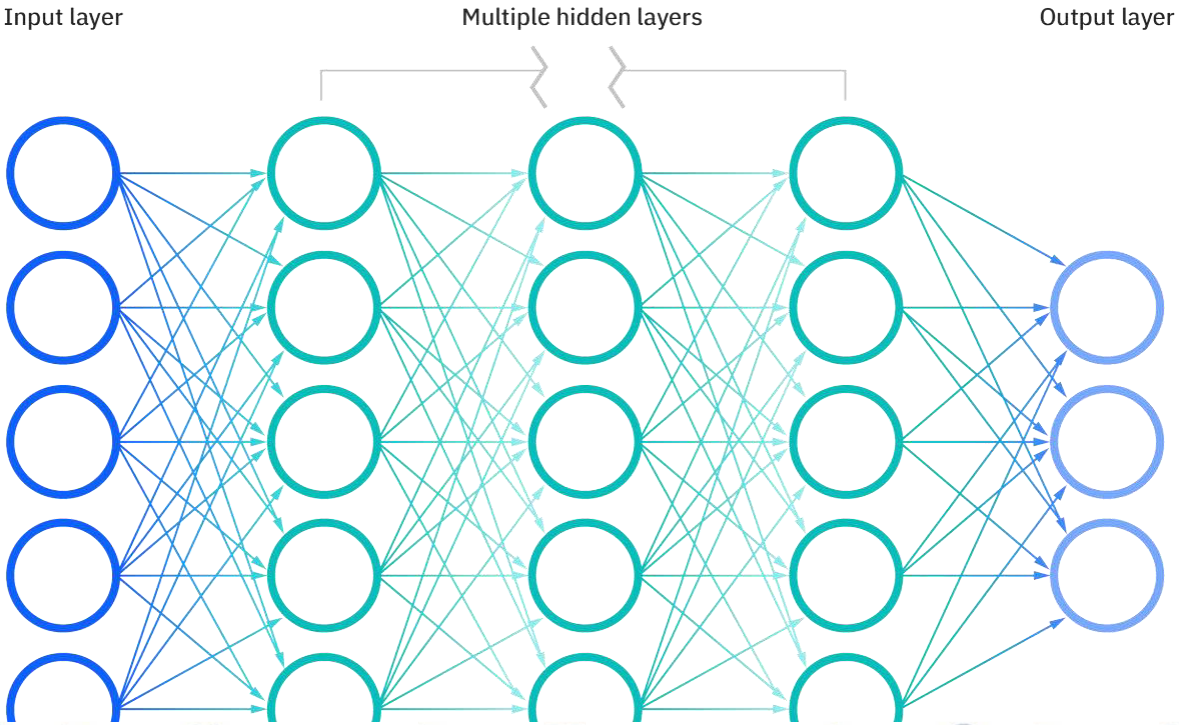
Per separare le due classi di punti dati, ci sono molti iperpiani possibili che potrebbero essere scelti. **L'obiettivo è quello di trovare un piano che abbia il margine massimo, cioè la distanza massima tra i punti dati di entrambe le classi.** La massimizzazione della distanza di margine fornisce un certo rinforzo in modo che i futuri punti dati possano essere classificati con maggiore sicurezza.

Gli SVM sono punti dati più vicini all'iperpiano e influenzano la posizione e l'orientamento dell'iperpiano. Usando questi vettori di supporto, massimizziamo il margine del classificatore. L'eliminazione degli SVM cambierà la posizione dell'iperpiano. Questi sono i punti che ci aiutano a costruire il nostro SVM.

Neural Network

Le reti neurali, anche conosciute come reti neurali artificiali (Anns) o reti neurali simulate (SNNs), sono un sottoinsieme di apprendimento automatico e sono al centro degli algoritmi di apprendimento profondo. Il loro nome e la struttura sono ispirati dal cervello umano, imitando il modo in cui i neuroni biologici si segnalano l'un l'altro.

Le reti neurali artificiali (Anns) sono costituite da un livello di nodo, contenente **un livello di input, uno o più livelli nascosti e un livello di output**. Ogni nodo, o neurone artificiale, si collega ad un altro e ha un peso e una soglia associati. Se l'output di un singolo nodo è superiore al valore di soglia specificato, tale nodo viene attivato, inviando i dati al livello successivo della rete. Altrimenti, nessun dato viene passato al livello successivo.

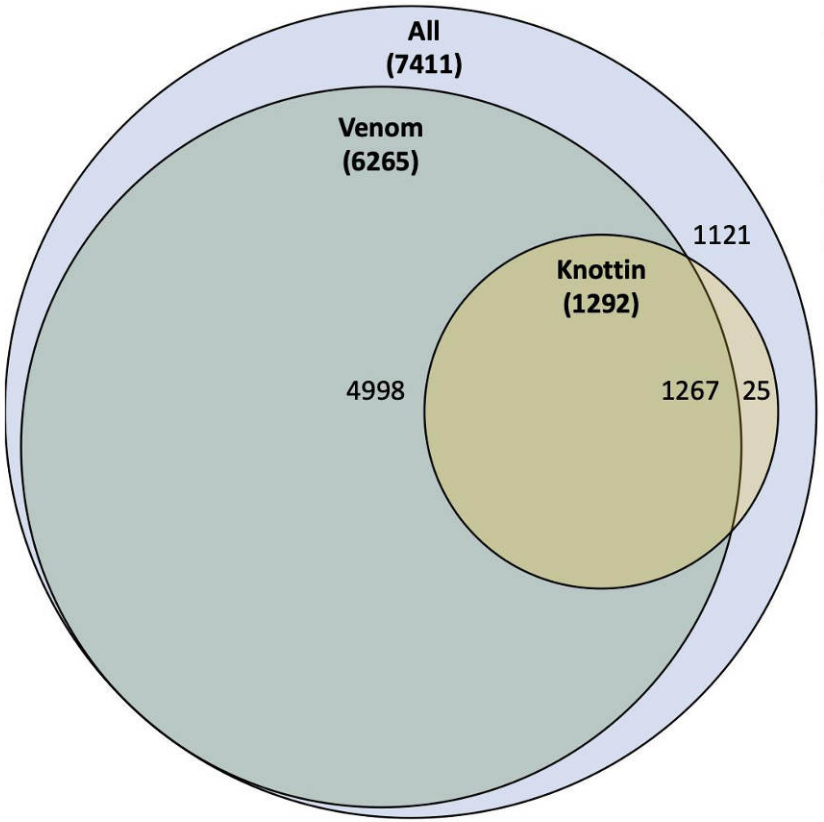


Tool	Campo di applicazione	Limitazioni tecniche
ToxClassifier	Tossine animali (veleno)	N.A.
NNTox	Tossine	N.A.
TOXIFY	Tossine animali (veleno)	Sequenza ≤500 amminoacidi
ToxDL	Tossine	N.A.
ToxIBTL	Tossine	N.A.
ToxinPred2	Tossine	N.A.
KNOTTIN	Knottine	Sequenza ≤ 200 amminoacidi

Ogni voce del dataset contiene i seguenti campi: *Identificatore UniProt, GO - funzione molecolare, Organismo, Pfam, Motivi, Sequenza, Lunghezza della sequenza.*

Questi campi sono stati scelti per fornirci dati che possono essere utili per classificare proteine e tossine sulla base di informazioni rilevanti, come la funzione molecolare, la famiglia proteica di Pfam e la presenza di specifici motivi funzionali

Dataset	query	#Entries
alltox	((go:0090729*) AND (reviewed:true))	7411
venom	((go:0090729) AND (reviewed:true)) AND venom	6265
knottin	((go:0090729) AND (reviewed:true)) AND (keyword:KW-0960**)	1292
allnontox	(NOT (go:0090729)) AND (reviewed:true)	560591



A: 7411 total toxins
V: 6265 venom toxins
K: 1292 knottin toxins

A \ ***V***: 1146 toxins that are *not* venoms
A \ ***K***: 6119 toxins that are *not* knottins
A \ (***V*** ∪ ***K***): 1121 toxins that are *neither* venoms *nor* knottins

V ∩ ***K***: 1267 venoms that are knottins
V \ ***K***: 4998 venoms that are *not* knottins
K \ ***V***: 25 knottins that are *not* venoms

Il webserver di ToxClassifier è offline e non è possibile testarlo da GitHub poiché mancano alcuni codici.

Il webserver di KNOTTIN ha alcuni problemi e non possiamo procedere con i test.

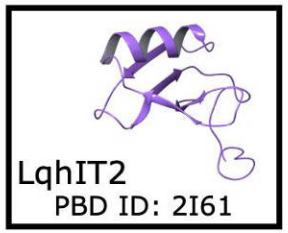
2 lines in sequence INPUT:

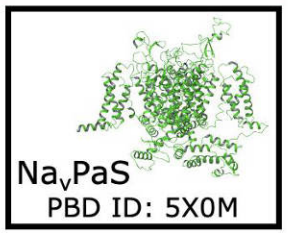
>mytest

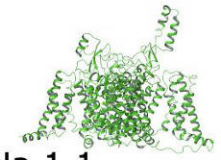
GCPRILMRCKRDSDCLAECVCWPNGFCG

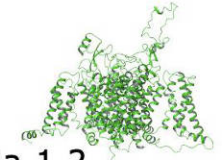
The sequence is probably not a knottin: No action taken.

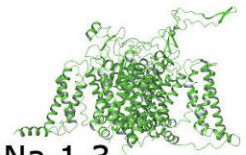
Modelli e Strutture

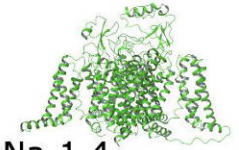

LqhIT2
PBD ID: 2I61

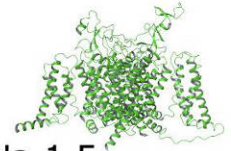

Na_vPaS
PBD ID: 5X0M

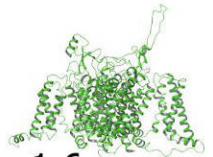

$hNa_v1.1$
PBD ID: 7DTD


$hNa_v1.2$
PBD ID: 6J8E

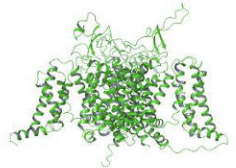

$hNa_v1.3$
PBD ID: 7W77


$hNa_v1.4$
PBD ID: 6AGF


$hNa_v1.5$
PBD ID: 6LQA


$hNa_v1.6$
HM (template $hNa_v1.2$)


$hNa_v1.7$
PBD ID: 6J8I

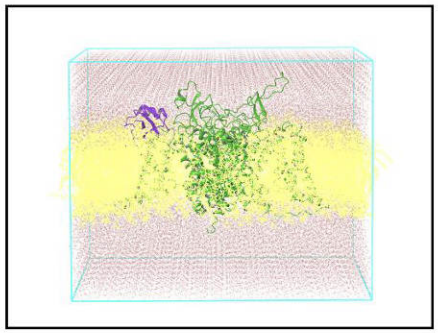

$hNa_v1.8$
HM (template $hNa_v1.5$)


$hNa_v1.9$
HM (template $hNa_v1.4$)

Docking proteina::proteina

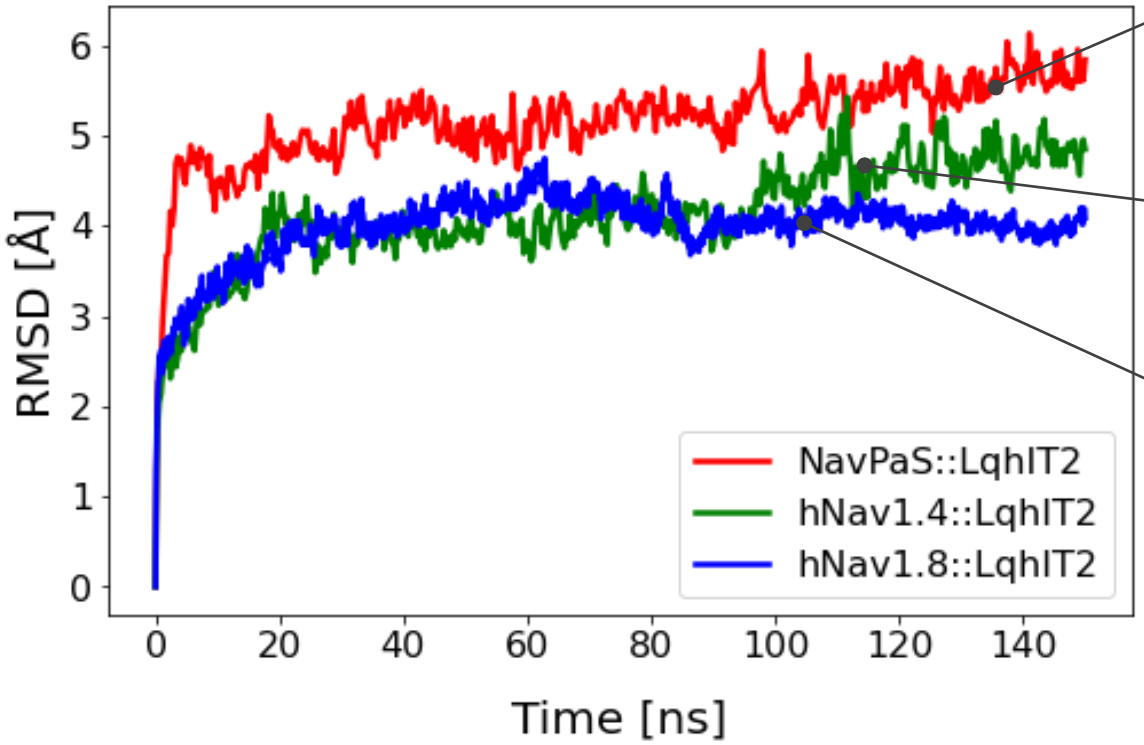
- **Na_vPaS::LqhIT2**
- $hNa_v1.1$::LqhIT2
- $hNa_v1.2$::LqhIT2
- $hNa_v1.3$::LqhIT2
- **$hNa_v1.4$::LqhIT2**
- $hNa_v1.5$::LqhIT2
- $hNa_v1.6$::LqhIT2
- $hNa_v1.7$::LqhIT2
- **$hNa_v1.8$::LqhIT2**
- $hNa_v1.9$::LqhIT2

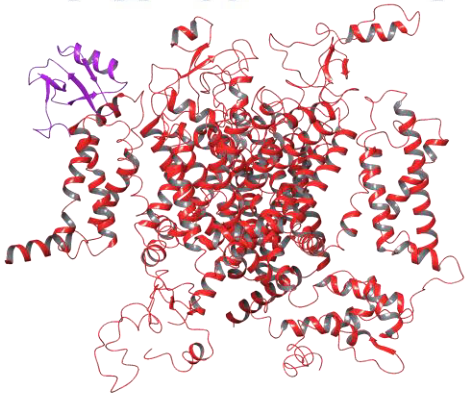
Dinamica molecolare

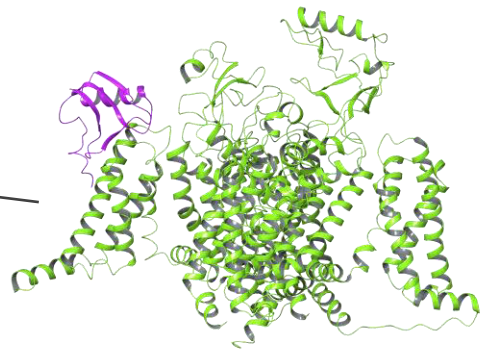


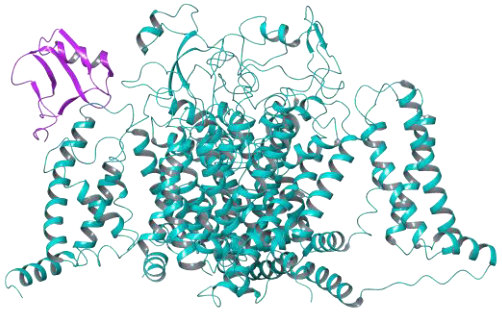
Descrittori termodinamici

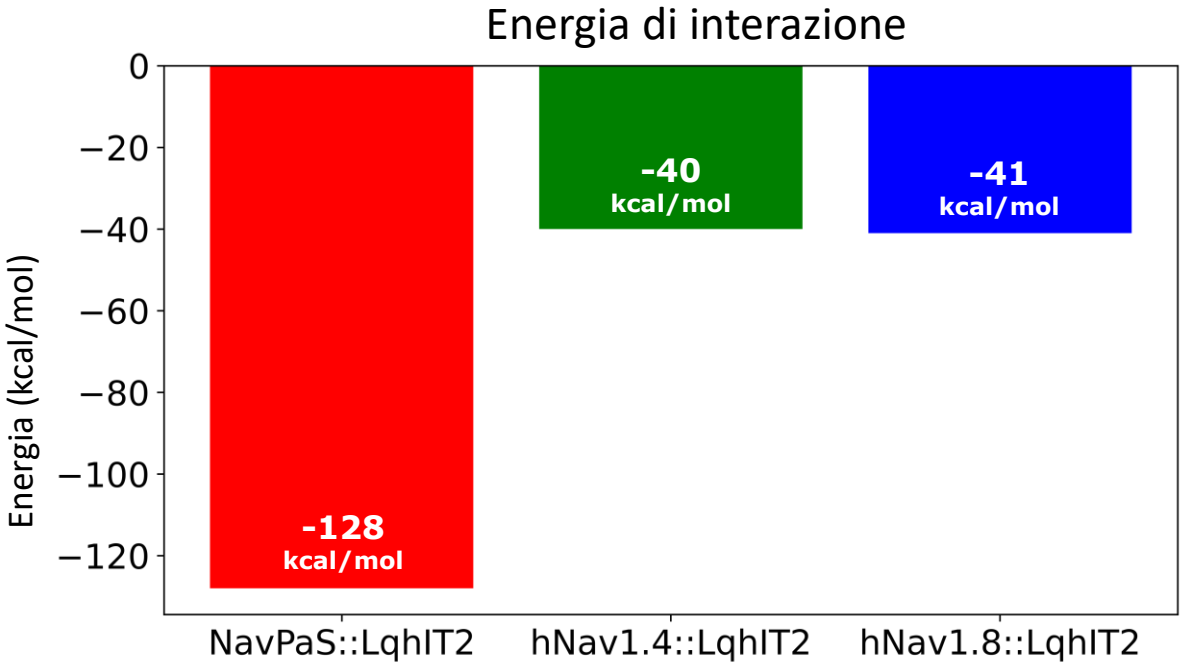
- Energia di interazione
- Energia libera di legame
- Interazioni

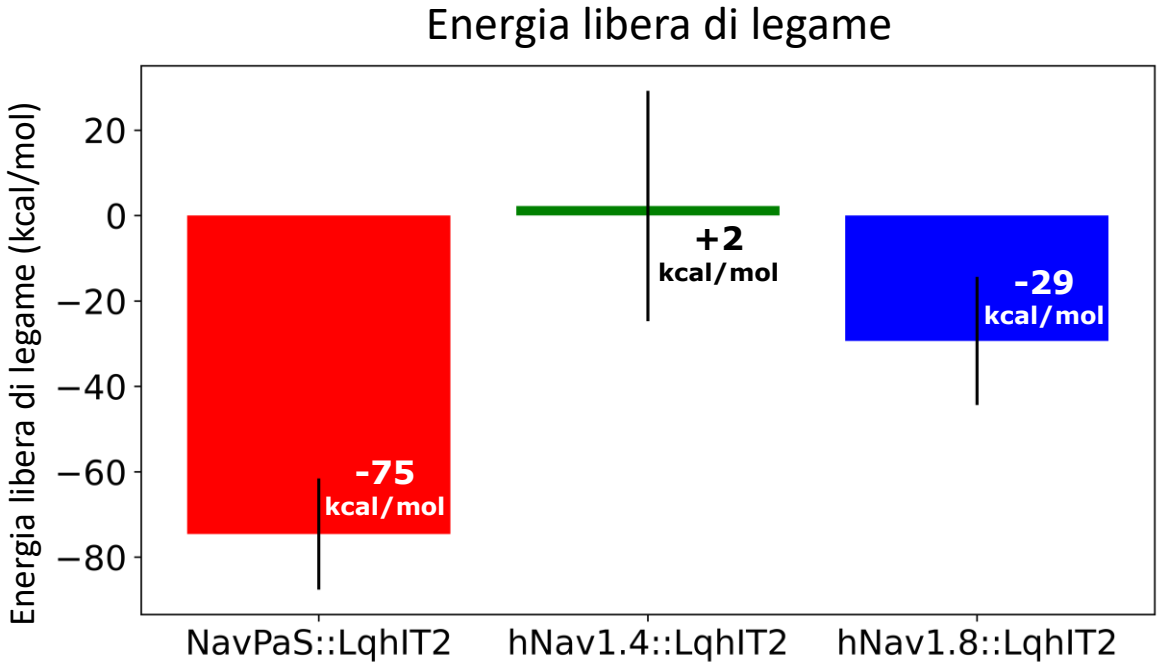

RMSD [Å]
Time [ns]
0
20
40
60
80
100
120
140
1
2
3
4
5
6
NavPaS::LqhIT2
hNav1.4::LqhIT2
hNav1.8::LqhIT2







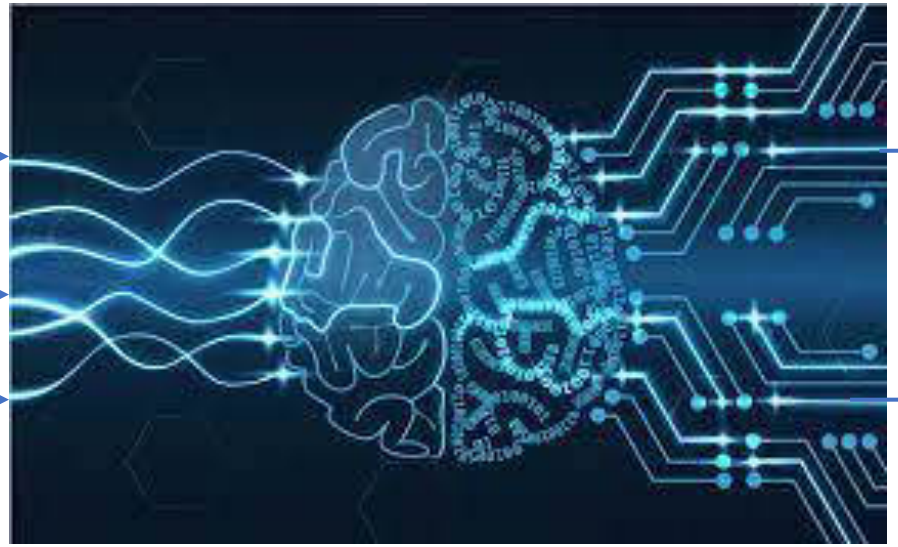

Energia di interazione
Energia (kcal/mol)
0
−20
−40
−60
−80
−100
−120
-128
kcal/mol
-40
kcal/mol
-41
kcal/mol
NavPaS::LqhIT2
hNav1.4::LqhIT2
hNav1.8::LqhIT2


Energia libera di legame
Energia libera di legame (kcal/mol)
20
0
−20
−40
−60
−80
-75
kcal/mol
+2
kcal/mol
-29
kcal/mol
NavPaS::LqhIT2
hNav1.4::LqhIT2
hNav1.8::LqhIT2

Consenso dei *tools*

Consenso delle metodologie

Biochimica e biofisica strutturale



TOSSINA

NON TOSSINA