Un Sistema Adattivo di Reservoir Computing per Grafi Applicato in Tossicologia

Andrea Zanelli

Laurea Magistrale in Informatica, Università di Pisa

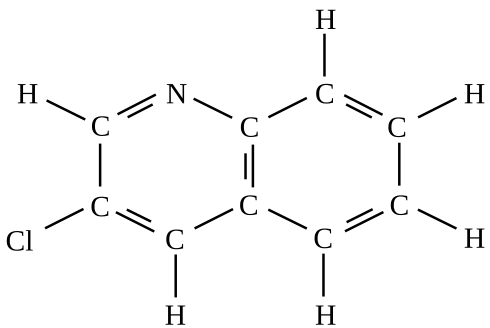
12 aprile 2013

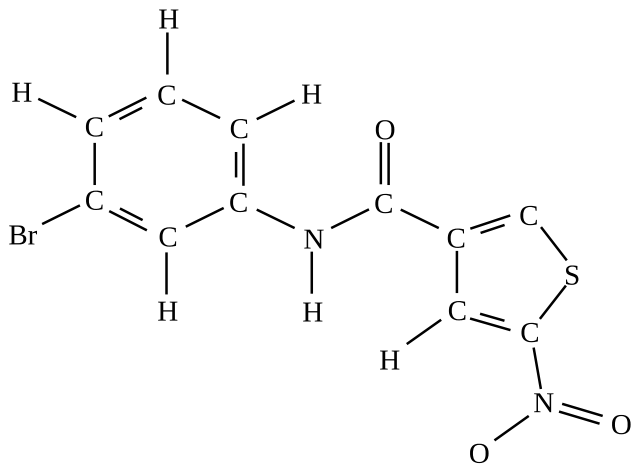
Introduzione: Dati Strutturati

I modelli di **apprendimento automatico** maggiormente diffusi riguardano il trattamento di dati vettoriali a **dimensione fissata**.

In molti ambiti però i dati sono rappresentati attraverso **strutture complesse** (e.g. liste, alberi o grafi).

- Esempio: in ambito chimico una **molecola** è tipicamente rappresentata come un **grafo** (i vertici sono gli atomi, gli archi i legami chimici tra gli atomi).





Introduzione: Dati Strutturati

Una tipica soluzione per gestire questo tipo di dati, chiamati **dati strutturati**, è utilizzare un **vettore di descrittori** per rappresentare il dato a dimensione fissata.

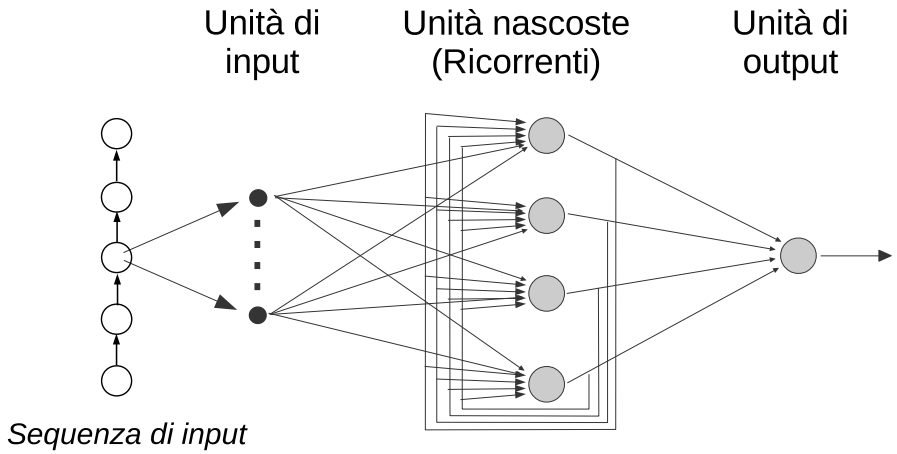
Svantaggi dell'approccio a descrittori:

- Mancanza di **generalità**: per ogni problema può essere necessario scegliere un insieme di descrittori differenti.
- In genere è rischiesto il **supporto di un esperto** del dominio applicativo per la scelta di descrittori più appropriati.
- Si possono perdere importanti **informazioni strutturali**.

Sono stati studiati e sono oggetto di recenti avanzamenti metodi di apprendimento automatico che permettono di elaborare **direttamente dati strutturati**.

Introduzione: Reservoir Computing

Il lavoro svolto in questa tesi si concentra su modelli di **Reti Neurali**.



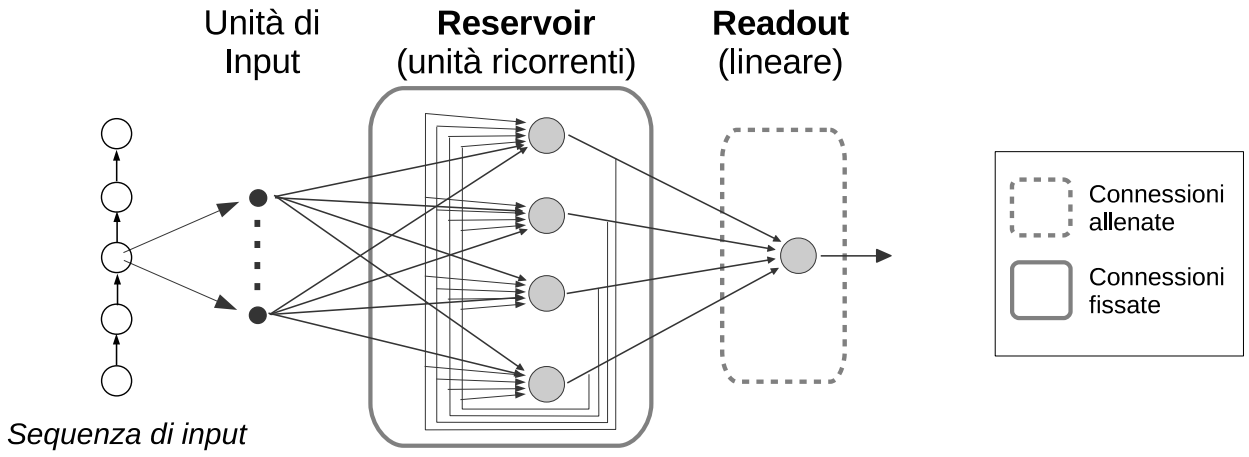
Le **Reti Neurali Ricorrenti** (RNNs) sono una classe di reti neurali che permettono di **apprendere** e **calcolare** trasduzioni strutturali su **sequenze**.

Il **Reservoir Computing** (RC) è un paradigma per RNNs che ne permette un training **efficiente** grazie ad una separazione concettuale in un **reservoir** (unità ricorrenti) e in un **readout** (lineare).

- Le **Echo State Networks** (ESNs): uno dei principali modelli di RC.

Introduzione: Reservoir Computing

Il lavoro svolto in questa tesi si concentra su modelli di **Reti Neurali**.



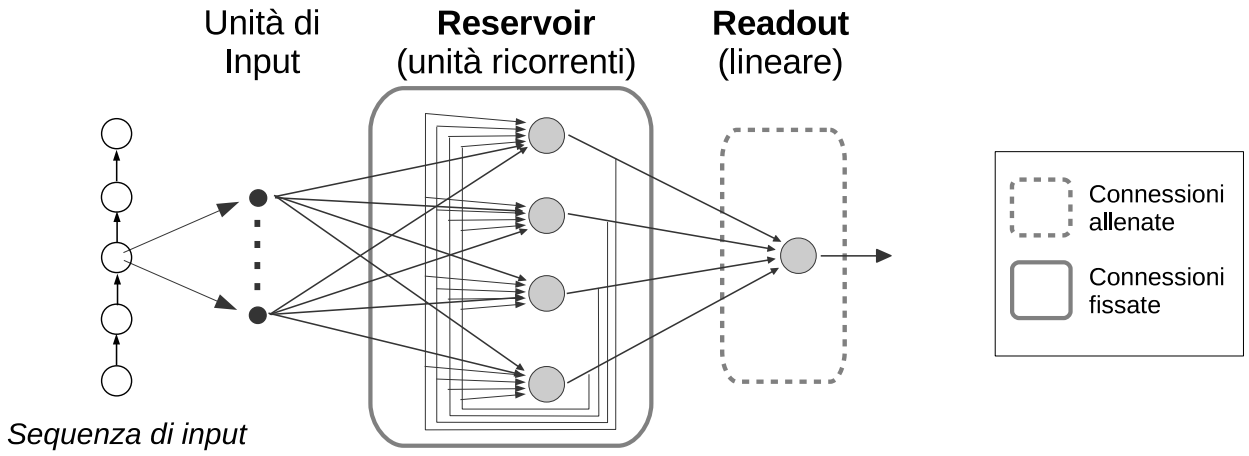
Le **Reti Neurali Ricorrenti** (RNNs) sono una classe di reti neurali che permettono di **apprendere** e **calcolare** trasduzioni strutturali su **sequenze**.

Il **Reservoir Computing** (RC) è un paradigma per RNNs che ne permette un training **efficiente** grazie ad una separazione concettuale in un **reservoir** (unità ricorrenti) e in un **readout** (lineare).

- Le **Echo State Networks** (ESNs): uno dei principali modelli di RC.

Introduzione: Reservoir Computing

Il lavoro svolto in questa tesi si concentra su modelli di **Reti Neurali**.



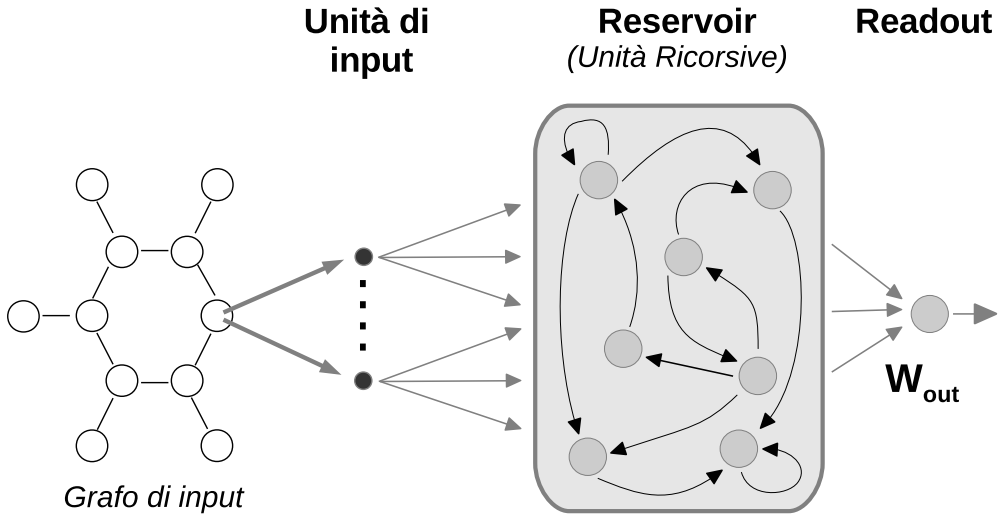
Le **Reti Neurali Ricorrenti** (RNNs) sono una classe di reti neurali che permettono di **apprendere** e **calcolare** trasduzioni strutturali su **sequenze**.

Il **Reservoir Computing** (RC) è un paradigma per RNNs che ne permette un training **efficiente** grazie ad una separazione concettuale in un **reservoir** (unità ricorrenti) e in un **readout** (lineare).

- Le **Echo State Networks** (ESNs): uno dei principali modelli di RC.

Introduzione: Graph Echo State Network

Di recente sono state introdotte le **Graph Echo State Networks**.

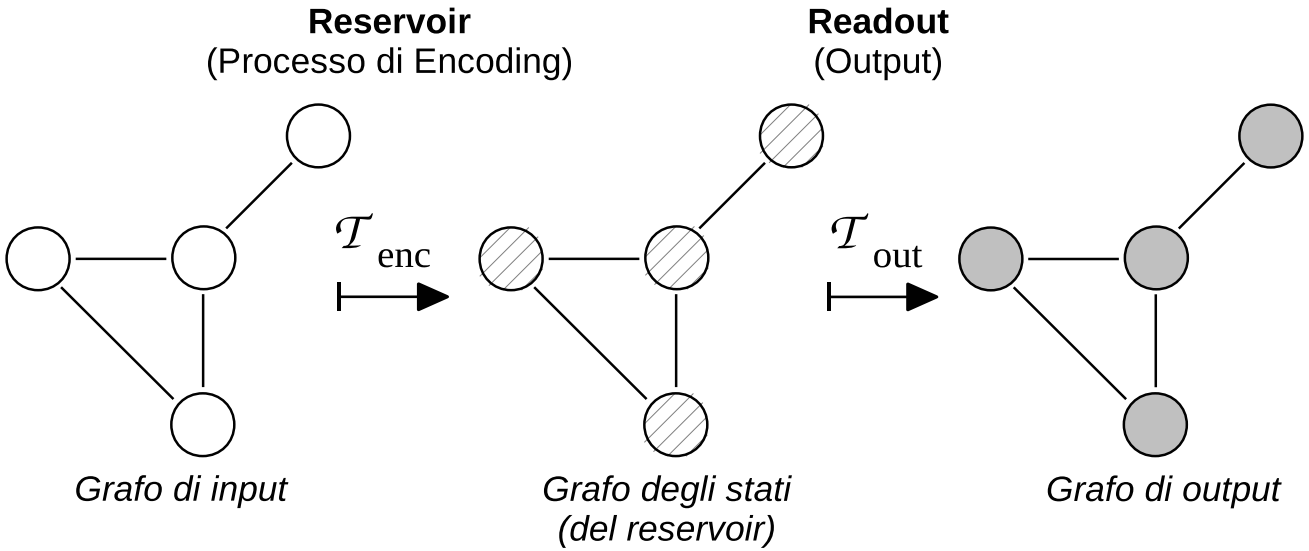


Estendono le Echo State Networks all'elaborazione di **grafi** conservando i punti di forza del **Reservoir Computing** (semplicità ed efficienza del training).

- Reservoir di **unità ricorsive**.
- Una **inizializzazione** delle unità del reservoir rispettando opportuni **vincoli** permette di trattare strutture sia **cicliche** che **acicliche**, sia **grafi diretti e indiretti**.

Introduzione: State Mapping Function

Le Graph Echo State Networks (GraphESNs) producono (in modo naturale) un output in corrispondenza di ogni vertice di un grafo di input.

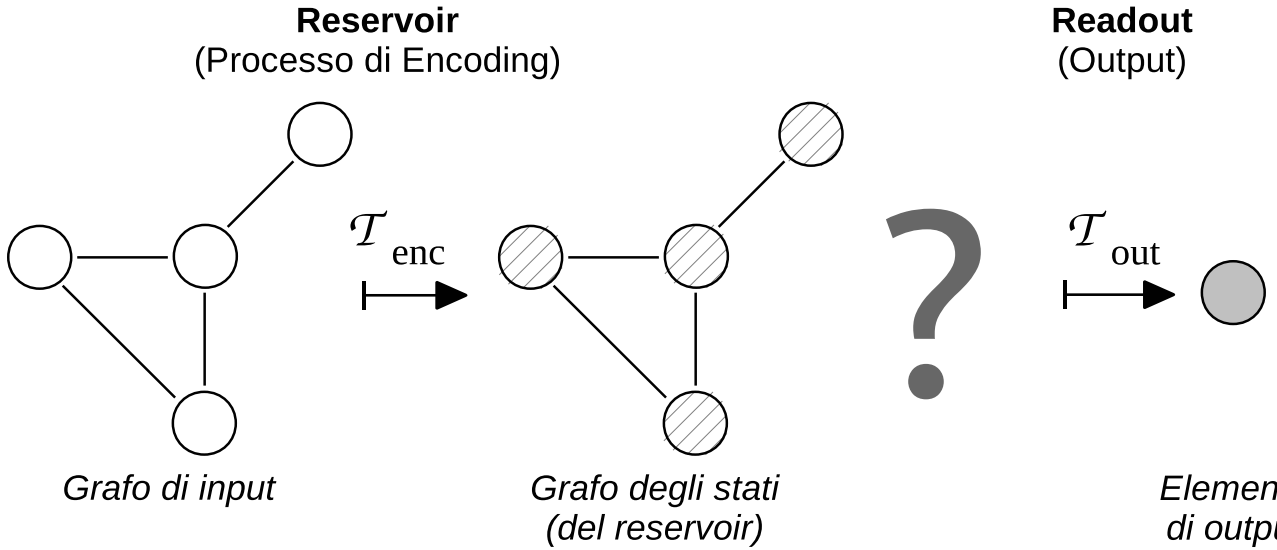


Classificazione: un grafo dev'essere mappato in un singolo elemento di output (e.g. in **tossicologia**).

Si utilizza una **State Mapping Function** (SMF) $\mathcal{X}$ che restituisce uno **stato unico** per l'intero grafo (e.g. media). Implementazione di SMF **avanzate** oggetto di recentissimi studi.

Introduzione: State Mapping Function

Le Graph Echo State Networks (GraphESNs) producono (in modo naturale) un output in corrispondenza di ogni vertice di un grafo di input.

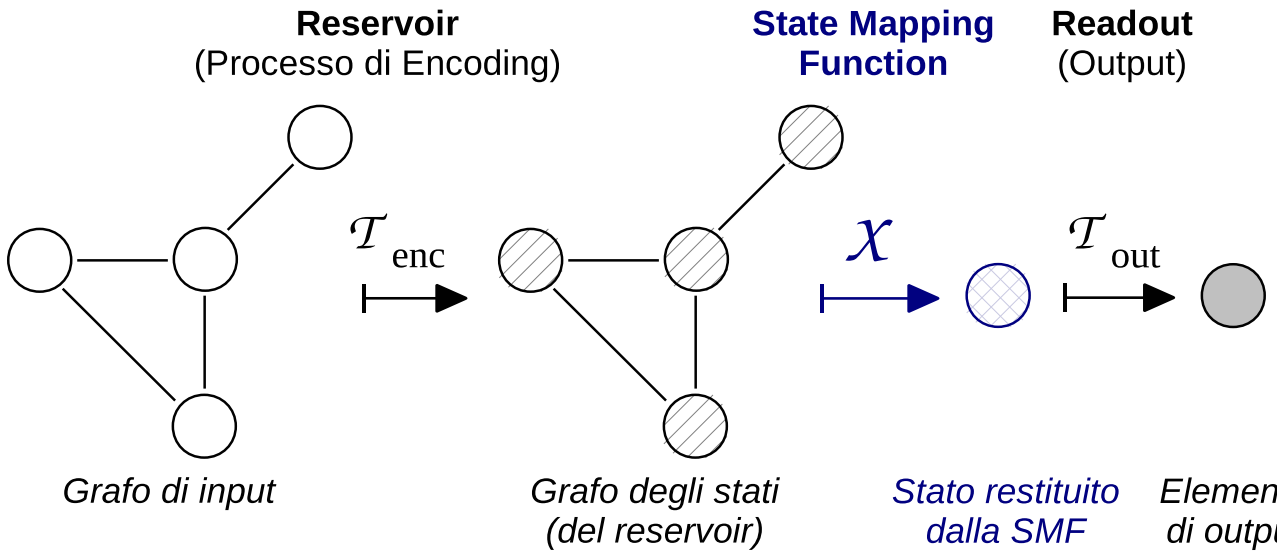


Classificazione: un grafo dev'essere mappato in un singolo elemento di output (e.g. in **tossicologia**).

Si utilizza una **State Mapping Function** (SMF) $\mathcal{X}$ che restituisce uno **stato unico** per l'intero grafo (e.g. media). Implementazione di SMF **avanzate** oggetto di recentissimi studi.

Introduzione: State Mapping Function

Le Graph Echo State Networks (GraphESNs) producono (in modo naturale) un output in corrispondenza di ogni vertice di un grafo di input.



Classificazione: un grafo dev'essere mappato in un singolo elemento di output (e.g. in **tossicologia**).

Si utilizza una **State Mapping Function** (SMF) $\mathcal{X}$ che restituisce uno **stato unico** per l'intero grafo (e.g. media). Implementazione di SMF **avanzate** oggetto di recentissimi studi.

Introduzione: Obiettivi

Obiettivo: realizzare un sistema di Reservoir Computing per grafi in grado di fornire elementi per un'**analisi qualitativa della risposta** (caratteristica fin'ora assente da modelli di questo tipo):

- Estendendo il modello **GraphESN**.
- Nuova **SMF**, basata su Self Organizing Map (SOM), in grado di fornire elementi utili all'analisi della risposta.
- **Pruning** dei pesi del readout per ridurre il numero di elementi che determinano la risposta del sistema.

Importante applicazione in **tossicologia computazionale**:

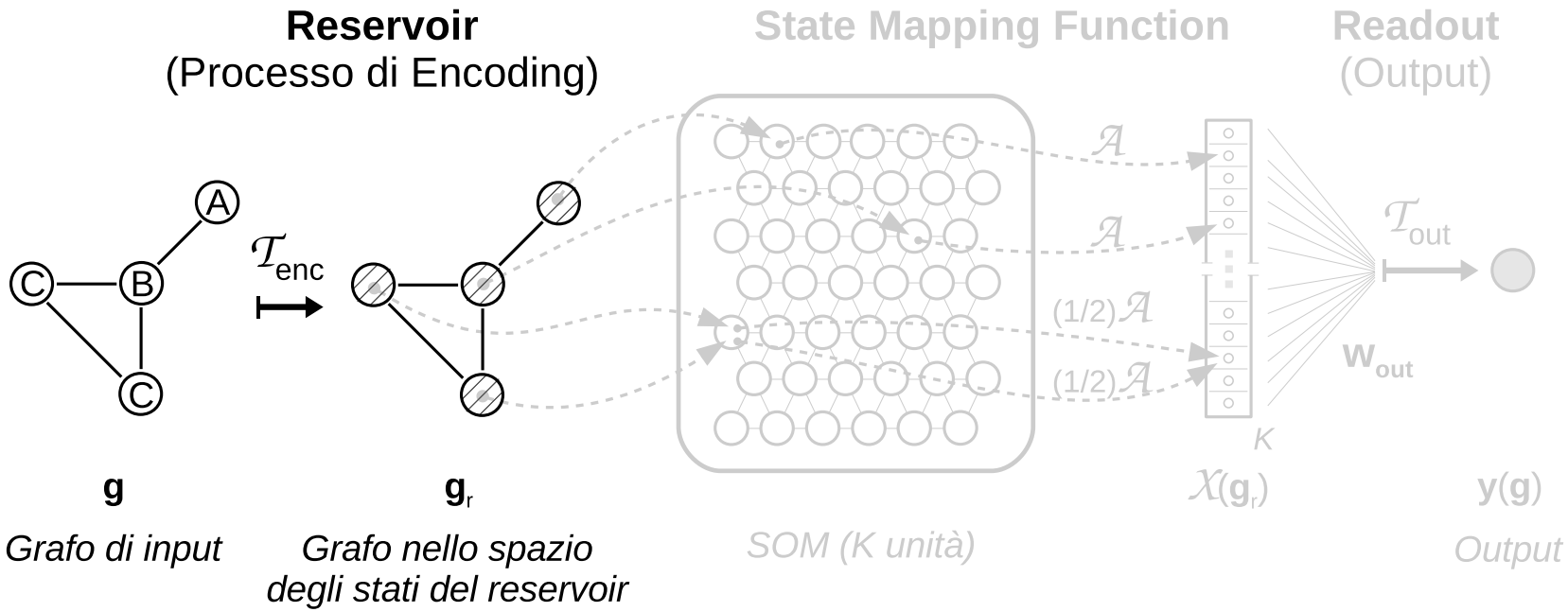
- Per **valutare la pericolosità** di un composto chimico.
- In grado di **trattare grafi** con cui si rappresentano i composti.
- **Supporto all'analisi** di utenti esperti del campo: devono fornire informazioni qualitative.

Introduzione: Obiettivi

Recenti leggi europee (**REACH** in particolare) **incentivano** e **regolamentano** l'utilizzo di sistemi predittivi per la valutazione di composti chimici (metodi alternativi alla sperimentazione su animali):

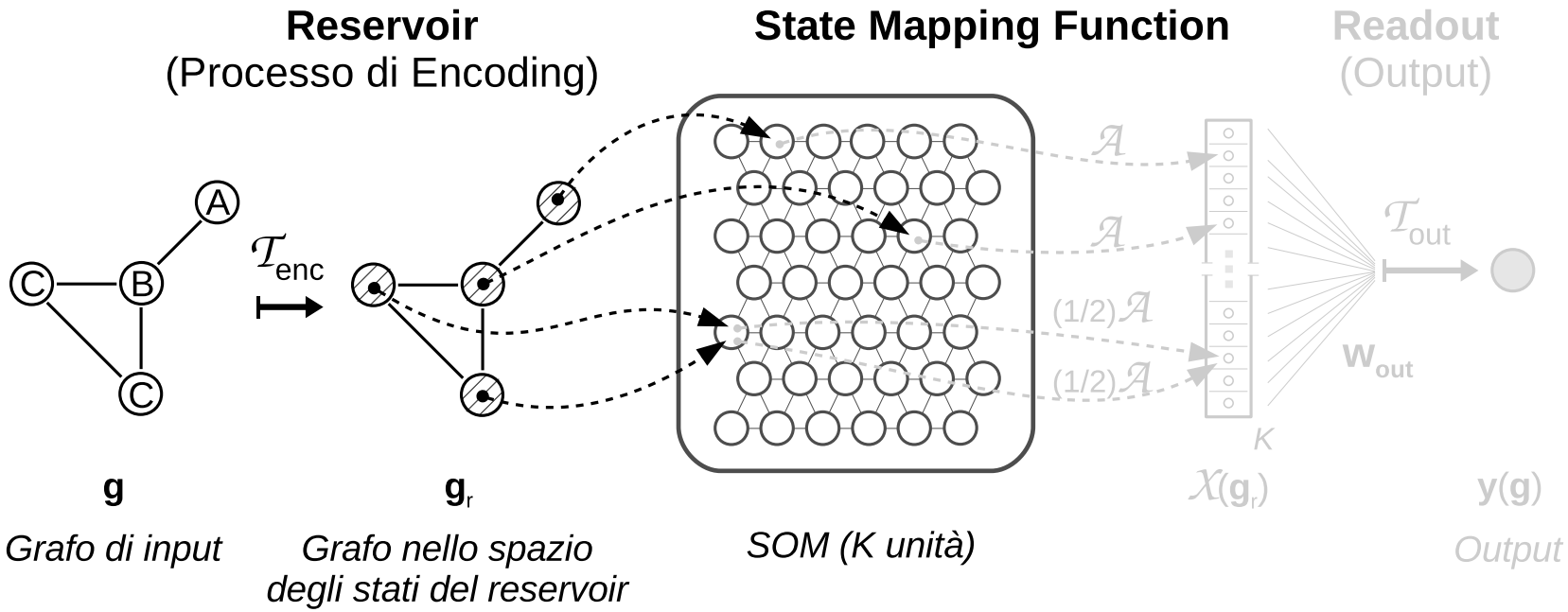
- Il lavoro svolto in questa tesi è stato quindi orientato anche alla **realizzazione di modelli validi** sotto queste regolamentazioni.

Descrizione del Sistema Proposto



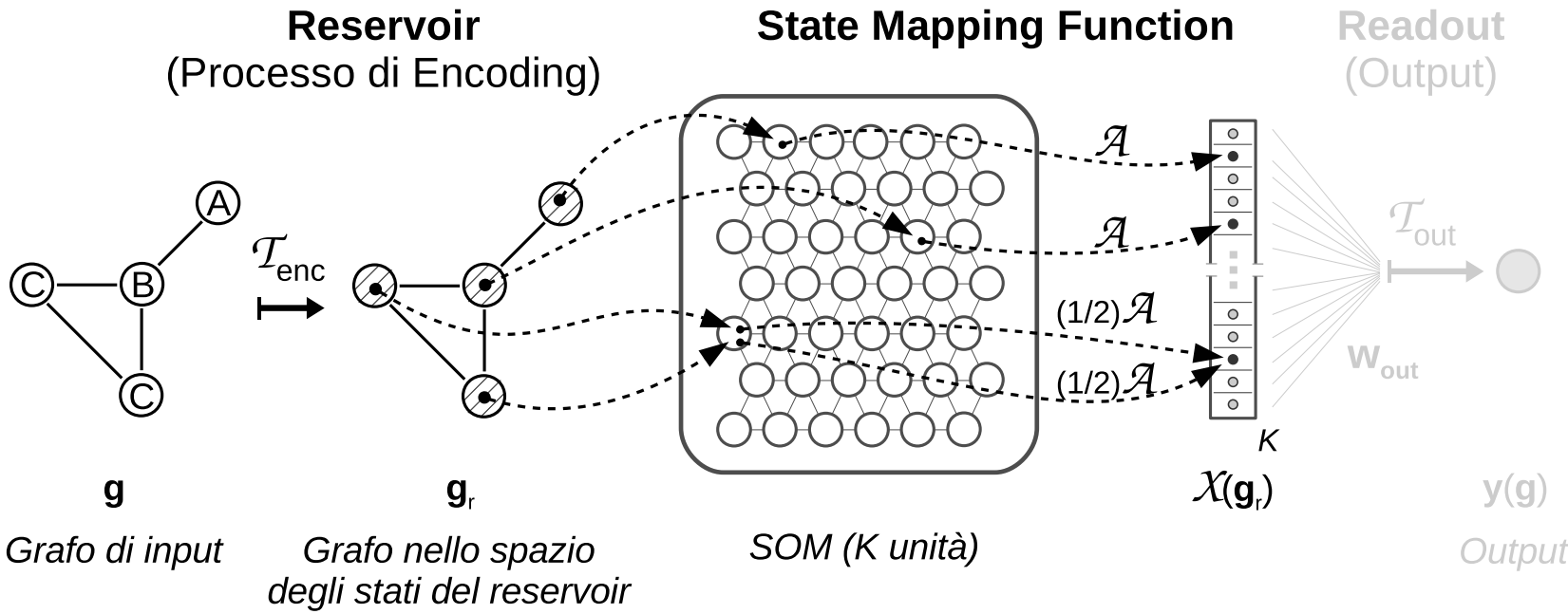
Funzionamento di **GraphESN-SOM** (problemi di **classificazione**): dato un **grafo di input g** viene calcolato, dal processo di *encoding* (i.e. dal reservoir), un **grafo nello spazio degli stati** $\mathbf{g}_r$.

Descrizione del Sistema Proposto



Ogni **vertice** del grafo nello spazio degli stati è **mappato in una SOM** di K unità, allenata in modo **supervisionato** sull'insieme degli stati corrispondenti a vertici di grafi di un **training set**.

Descrizione del Sistema Proposto

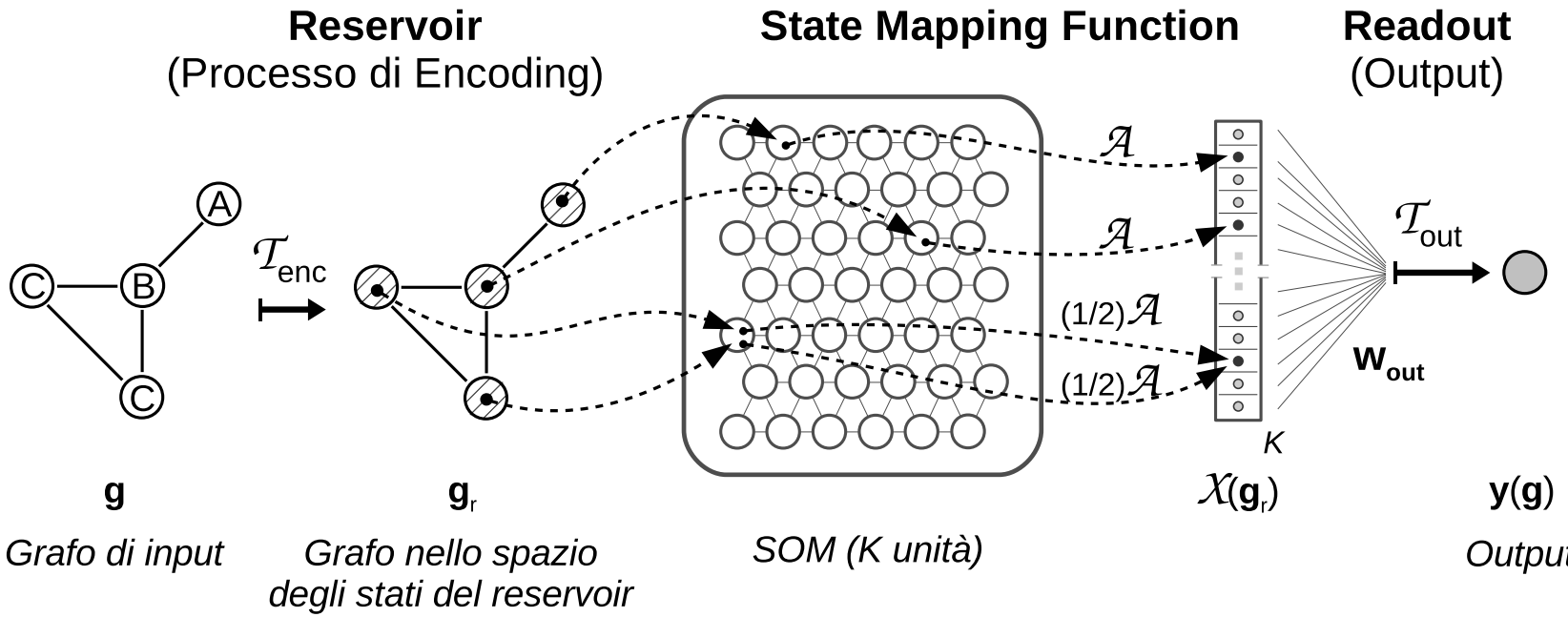


Per ogni **unità della SOM** si crea un **cluster**, eventualmente vuoto, composto dagli **stati catturati** da tale unità.

Per ogni **cluster** (i.e. per ogni unità della SOM) si ottiene un **singolo valore** applicando una **funzione di aggregazione** $\mathcal{A}$ ad ogni stato, e prendendo la **media locale** ad ogni cluster dei valori restituiti dalla funzione $\mathcal{A}$.

La **concatenazione** di tali valori crea il vettore finale $\mathcal{X}(\mathbf{g}_r)$ **restituito dalla SMF**, che è una rappresentazione a dimensione fissata di $\mathbf{g}$.

Descrizione del Sistema Proposto

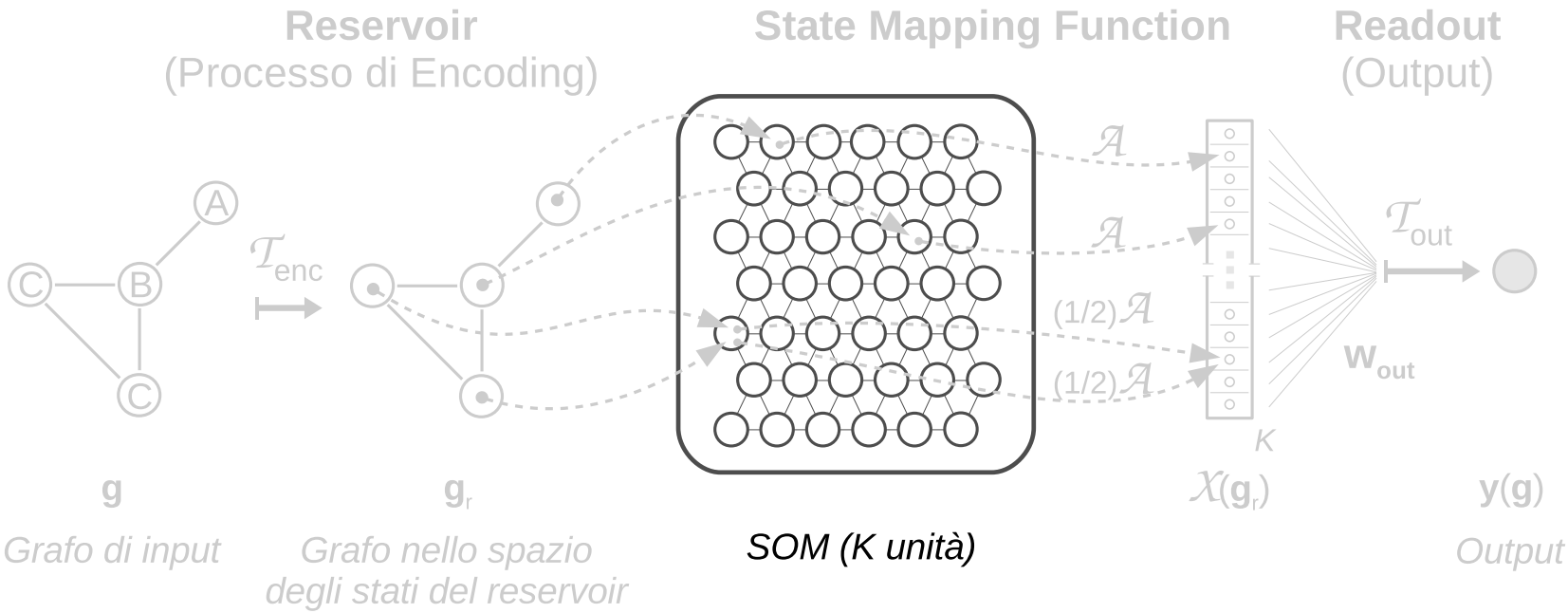


Il **readout**, allenato con **Elastic Net** sui dati di training ottenendo il vettore dei pesi $\mathbf{w}_{out}$, calcola l'**output finale** $\mathbf{y}(\mathbf{g})$ con l'equazione **lineare**:

- $\mathbf{y}(\mathbf{g}) = \mathbf{w}_{out}\mathcal{X}(\mathbf{g}_r)$

La riduzione dello stato del reservoir ad un **solo valore** permette di assegnare, nel readout, **un unico peso** ad ogni unità della SOM: Elastic Net effettua un pruning dei pesi che permette di **annullare** alcune unità SOM.

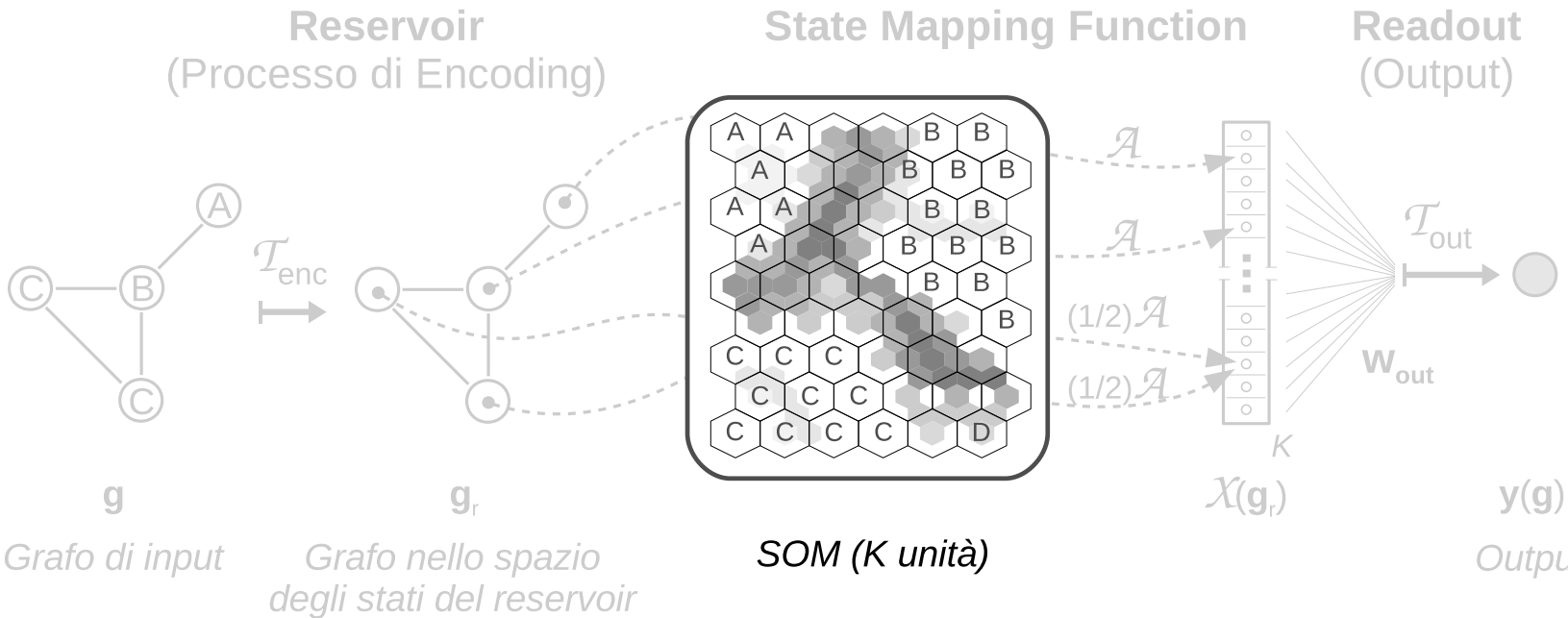
Descrizione del Sistema Proposto



La SOM ha la notevole caratteristica di poter essere **facilmente visualizzata** in una mappa bi-dimensionale, ad esempio tramite una **U-Matrix**.

Nel readout viene **associato un singolo peso** ad ogni unità della SOM che può essere visualizzato direttamente sulla mappa: rappresentazione grafica del funzionamento del modello.

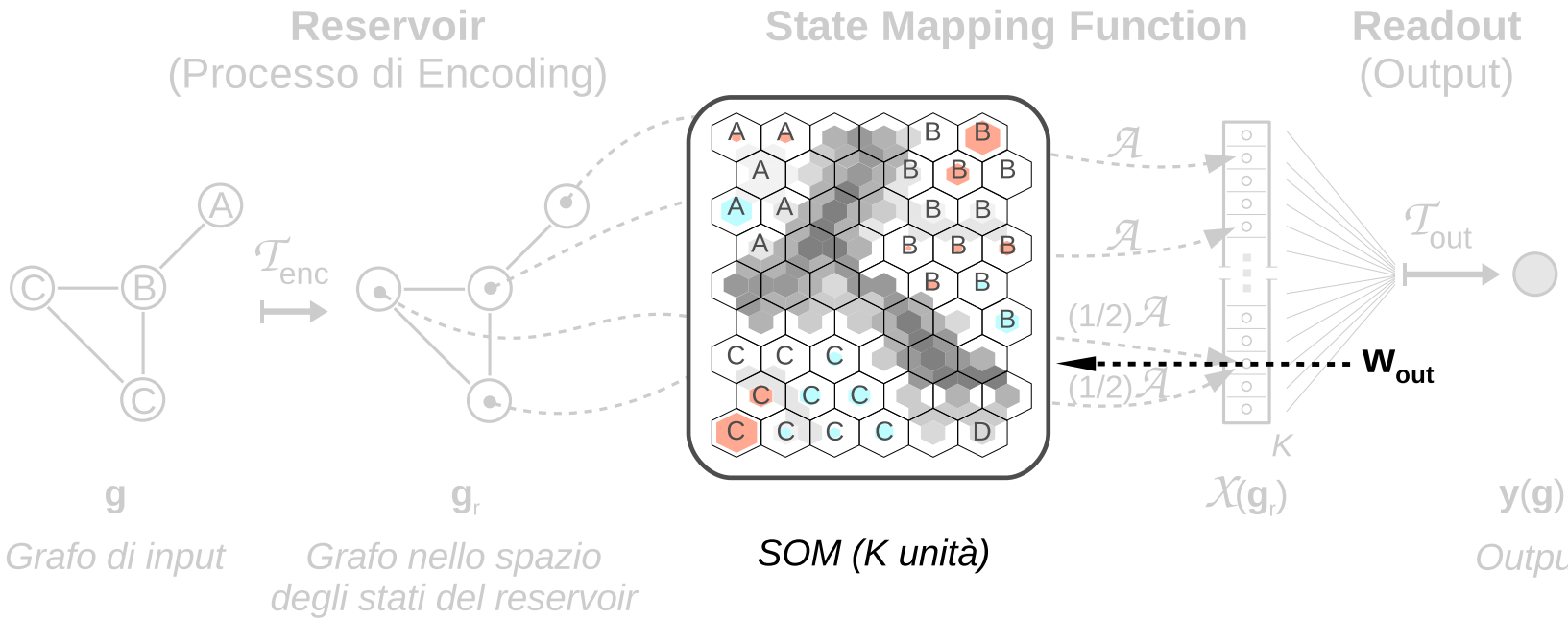
Descrizione del Sistema Proposto



La SOM ha la notevole caratteristica di poter essere **facilmente visualizzata** in una mappa bi-dimensionale, ad esempio tramite una **U-Matrix**.

Nel readout viene **associato un singolo peso** ad ogni unità della SOM che può essere visualizzato direttamente sulla mappa: rappresentazione grafica del funzionamento del modello.

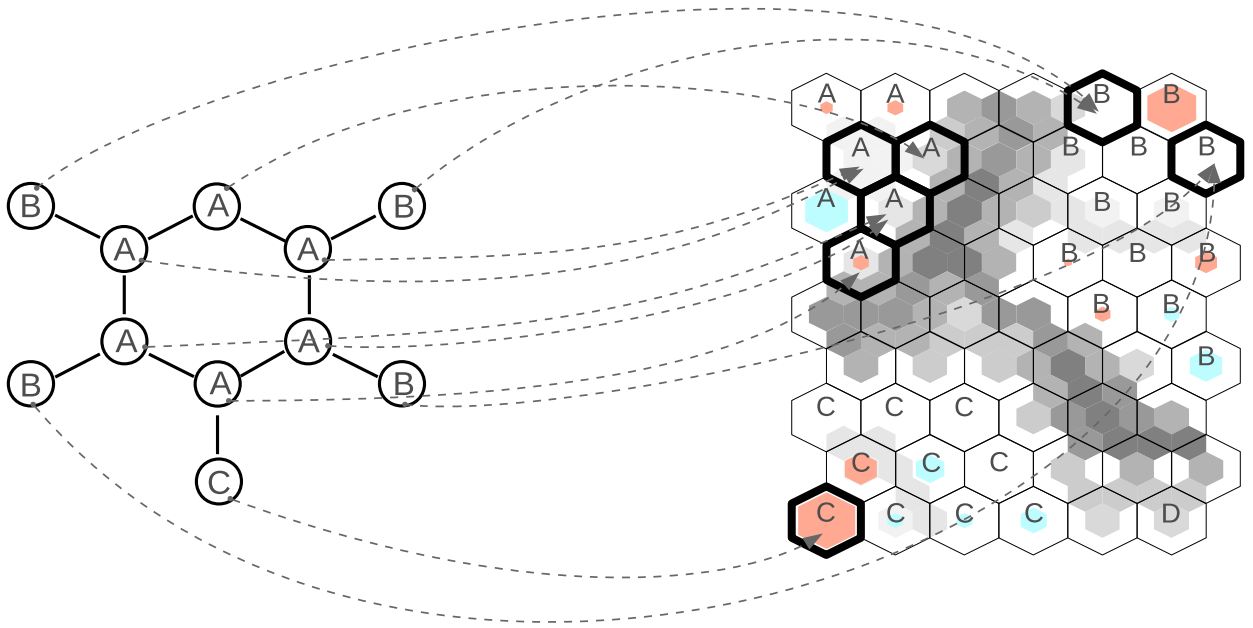
Descrizione del Sistema Proposto



La SOM ha la notevole caratteristica di poter essere **facilmente visualizzata** in una mappa bi-dimensionale, ad esempio tramite una **U-Matrix**.

Nel readout viene **associato un singolo peso** ad ogni unità della SOM che può essere visualizzato direttamente sulla mappa: rappresentazione grafica del funzionamento del modello.

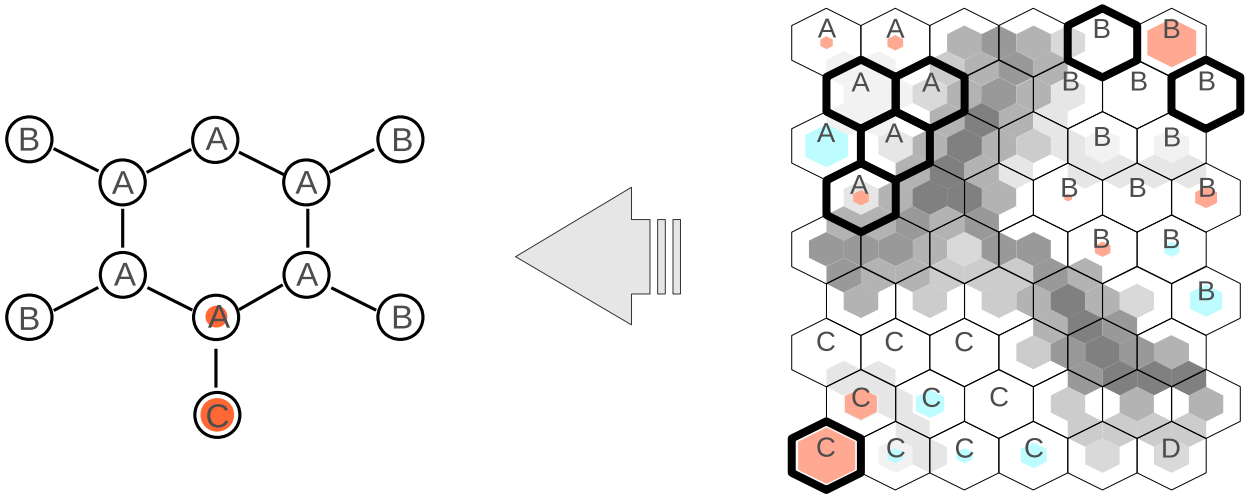
Descrizione del Sistema Proposto



Mappatura di un **grafo di input**: ogni vertice è mappato in una unità della SOM in base al valore dello **stato del reservoir** associato ad esso, il quale dipende dall'**etichetta** del vertice e dal **suo contesto strutturale**.

Il **peso** associato alle celle attivate in questo modo **guida la classificazione** del grafo (esempio in figura: positivo). Dalla visualizzazione della mappa si può leggere il **funzionamento** del modello e capire i motivi della risposta.

Descrizione del Sistema Proposto



Riportando il peso su ogni vertice del grafo, si mettono in luce le **componenti strutturali** che hanno contribuito alla classificazione.

Il **pruning** (nel readout) permette, infatti, di ridurre il numero di pesi lasciando solamente quelli associati alle **celle ritenute più importanti** per il problema affrontato.

Descrizione del Sistema Proposto

La **State Mapping Function** realizzata:

- Permette di avere una **rappresentazione grafica** di un modello allenato che ne descrive il funzionamento.
- Permette di associare un **peso** ai vertici di un grafo rilevanti per la classificazione, mettendo in luce le componenti strutturali che di più contribuiscono alla risposta finale.
- Il training **supervisionato** della SOM crea una disposizione, delle unità della SOM, che cattura **aspetti significativi** per il problema in esame, in modo **adattivo**.
  - Differenziandosi efficacemente dai **metodi a descrittori** nei quali un insieme di caratteristiche devono essere scelte a priori e a seconda del problema.
- Tutto questo porta ad una SMF **innovativa** sia per gli aspetti supervisionati e sia nel contesto delle SMFs.

Descrizione del Sistema Proposto

Il training del readout via **Elastic Net**:

- $\mathbf{w}_{\text{out}} = \text{argmin}_{\mathbf{w}}\{\frac{1}{2}\|\mathbf{wX} - \mathbf{y}_{\text{target}}\|_2^2 + \lambda_1\|\mathbf{w}\|_1 + \frac{1}{2}\lambda_2\|\mathbf{w}\|_2^2\}$
- **Regolarizzazione** e **pruning** contemporaneamente.
- La **regolarizzazione** è fondamentale (in particolare in modelli di Reservoir Computing) per ottenere buone prestazioni di **generalizzazione** (evitando il problema dell'*overfitting*).
- Il **pruning** è utile a semplificare il modello e concentrare su un numero limitato di features la predizione, a favore di una migliore **interpretazione** della risposta del sistema.

Descrizione del Sistema Proposto

Costi computazionali della fase di training di GraphESN-SOM, dato un training set di T di grafi, per un totale di T_V vertici:

- Il processo di *encoding*: $\mathcal{O}(T_V N_R)$
- State Mapping Function (training SOM): $\mathcal{O}(T_V K N_R)$
- Training del readout via Elastic Net: $\mathcal{O}(K^3 + TK^2)$

Complessivamente il costo di training è dello **stesso ordine** del training con **GraphESN** classica:

- Tipicamente più **efficiente** di altri metodi (e.g. metodi a **kernel** o reti neurali allenate con algoritmi basati su **discesa del gradiente**).
- GraphESN-SOM conserva i vantaggi del **Reservoir Computing**.

Risultati Sperimentali

Prove in 6 problemi di **tossicologia**: classificare un composto chimico come **tossico** oppure **non tossico**.

Test sulle singole componenti di GraphESN-SOM:

- SOM supervisionata.
- Valori binari o valori continui restituiti dalla SMF.
- Effetti del pruning.

Confronto con **metodi simili** di Reservoir Computing per grafi e risultati allo **stato dell'arte** nei 6 problemi.

Risultati ottenuti con Double Cross Validation.

Risultati: SOM supervisionata

Effetti del training **supervisionato** della SOM confrontato con un training **non supervisionato**, nel funzionamento della GraphESN-SOM:

Task	SOM non superv. (test acc.)	SOM superv. (test acc.)
Bursi	82.03%	**82.82%**
ISSCAN (SAL)	73.07%	**73.64%**
PTC (FM)	61.41%	**61.65%**
PTC (FR)	67.76%	**68.49%**
PTC (MM)	65.29%	**66.84%**
PTC (MR)	56.18%	**57.91%**

Il training supervisionato della SOM porta ad un **miglioramento sistematico** in ognuno dei 6 problemi, confermando i vantaggi (in termini predittivi) di una SMF supervisionata.

Risultati: Valori binari

Confronto tra l'utilizzo di valori binari per la riduzione dello stato del reservoir ad un singolo valore ed un altro metodo che restituisce la **somma delle componenti** dello stato (un valore continuo):

Task	Valori binari (test acc.)	Somma delle comp. (test acc.)
Bursi	**82.82%**	81.41%
ISSCAN (SAL)	**73.64%**	72.72%
PTC (FM)	**61.65%**	60.32%
PTC (FR)	**68.49%**	64.45%
PTC (MM)	**66.84%**	64.51%
PTC (MR)	**57.91%**	56.62%

L'uso di **valori binari** nella SMF porta ad migliori risultati in termini di **generalizzazione** rispetto all'utilizzo di valori continui.

- Ulteriormente motivato l'utilizzo di valori binari, che portano a vantaggi in termini di **interpretazione** della risposta.

Risultati: Effetti del pruning

Confronto del training del readout via **Elastic Net** (che porta al **pruning** dei pesi nel readout) con un training via **Ridge Regression** (tipica regolarizzazione senza pruning):

Task	Elastic Net		Ridge Regression	
	Test acc.	Unità annull.	Test acc.	Unità annull.
Bursi	**82.82%**	68.41%	82.60%	36.87%
ISSCAN (SAL)	**73.64%**	81.49%	73.38%	50.33%
PTC (FM)	61.65%	92.61%	**62.09%**	57.91%
PTC (FR)	**68.49%**	94.58%	65.87%	54.47%
PTC (MM)	66.84%	89.14%	**66.95%**	58.15%
PTC (MR)	**57.91%**	87.17%	57.37%	45.85%

Il training via EN porta ad **annullare** un numero molto maggiore di unità della SOM rispetto al training con RR, mantenendo anche **ottime prestazioni** di generalizzazione.

Risultati: Confronto con altri metodi (1)

Confronto con **approcci simili** di Reservoir Computing per grafi (test accuracy):

Task	GraphESN-SOM	GraphESN	GraphESN-NG
Bursi	**82.8**%	75.8%	79.2%
PTC (FM)	61.7%	60.4%	**62.5**%
PTC (FR)	**68.5**%	67.1%	66.7%
PTC (MM)	**66.8**%	65.0%	64.8%
PTC (MR)	57.9%	57.4%	**65.7**%

GraphESN-SOM ottiene **migliori** performance predittive rispetto a GraphESN su ognuno dei problemi e sulla maggior parte dei problemi rispetto a GraphESN-NG.

Risultati: Confronto con altri metodi (2)

Confronto con risultati allo **stato dell'arte** (test accuracy):

Task	GraphESN-SOM	Altri metodi (best results)
Bursi	82.8%	83.2% (Descrittori + SVM)
ISSCAN (SAL)	73.6%	78.0% (SAs Benigni/Bossa)
PTC (FM)	61.7%	64.5% (Graph Kernels)
PTC (FR)	68.5%	66.9% (Graph Kernels)
PTC (MM)	66.8%	66.4% (Graph Kernels)
PTC (MR)	57.9%	65.7% (Graph Kernels)

Performance **comparabili** a risultati allo stato dell'arte, ottenuti da metodi più **costosi** e che non hanno componenti per un'**analisi qualitativa**.

L'**obiettivo** del sistema realizzato **non** è ottenere le migliori performance.

Significativo il risultato su **Bursi** (dataset caratterizzato da un'ottima quantità e qualità dei dati).

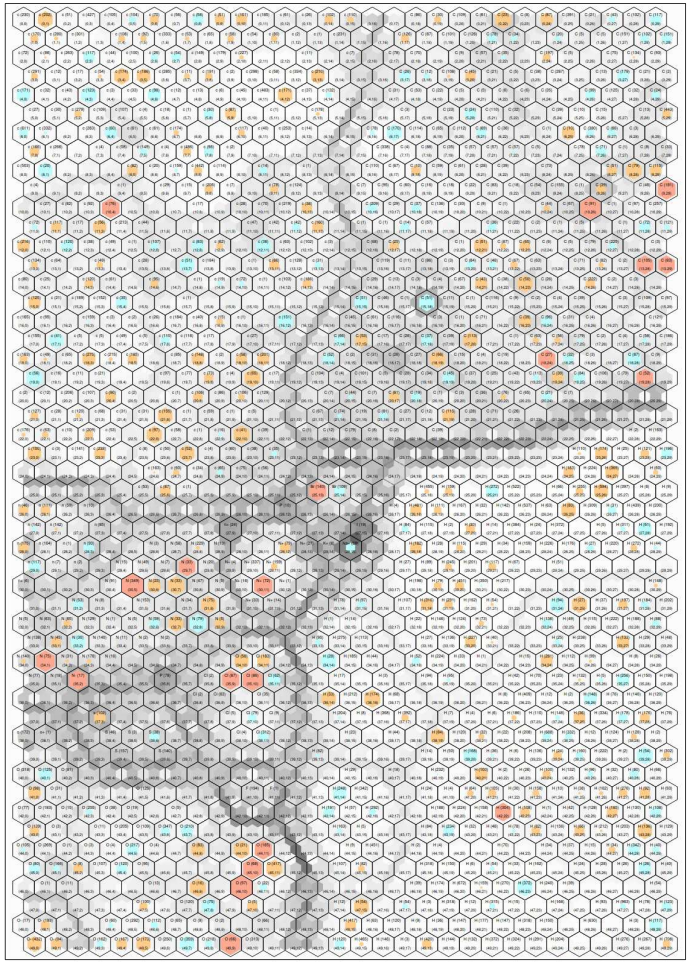
Esempio

Esempio di funzionamento del sistema GraphESN-SOM nella classificazione di un **composto chimico**:

- Un **modello** allenato sul training set Bursi.
- **Problema**: classificare un composto chimico come **tossico** oppure **non tossico**.
  - Tossico: **segno positivo** (valore target $+1$)
  - Non tossico: **segno negativo** (valore target -1)

Esempio: Mappa complessiva su tutto il training set Bursi

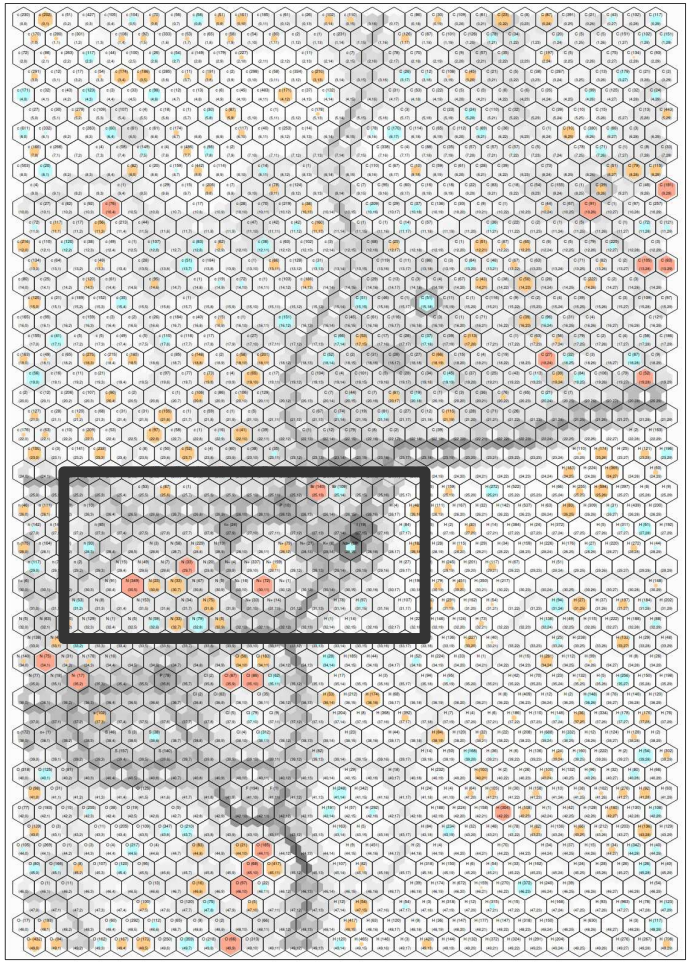
SOM 50x30 unità (K = 1500)



Mappa del **modello** allenato sul training set **Bursi** (U-Matrix di sfondo).

Esempio: Mappa complessiva su tutto il training set Bursi

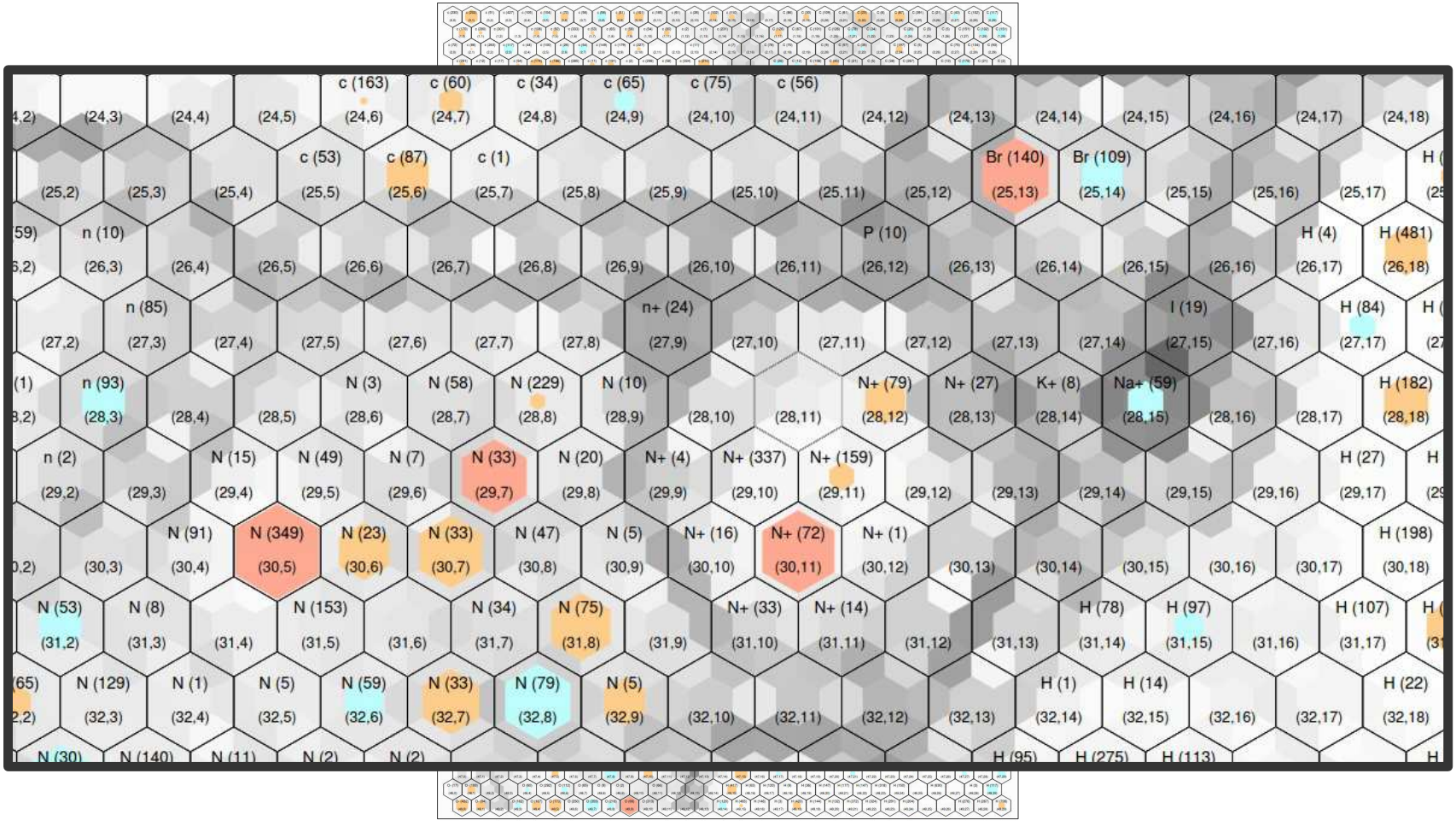
SOM 50x30 unità (K = 1500)



Mappa del **modello** allenato sul training set **Bursi** (U-Matrix di sfondo).

Esempio: Mappa complessiva su tutto il training set Bursi

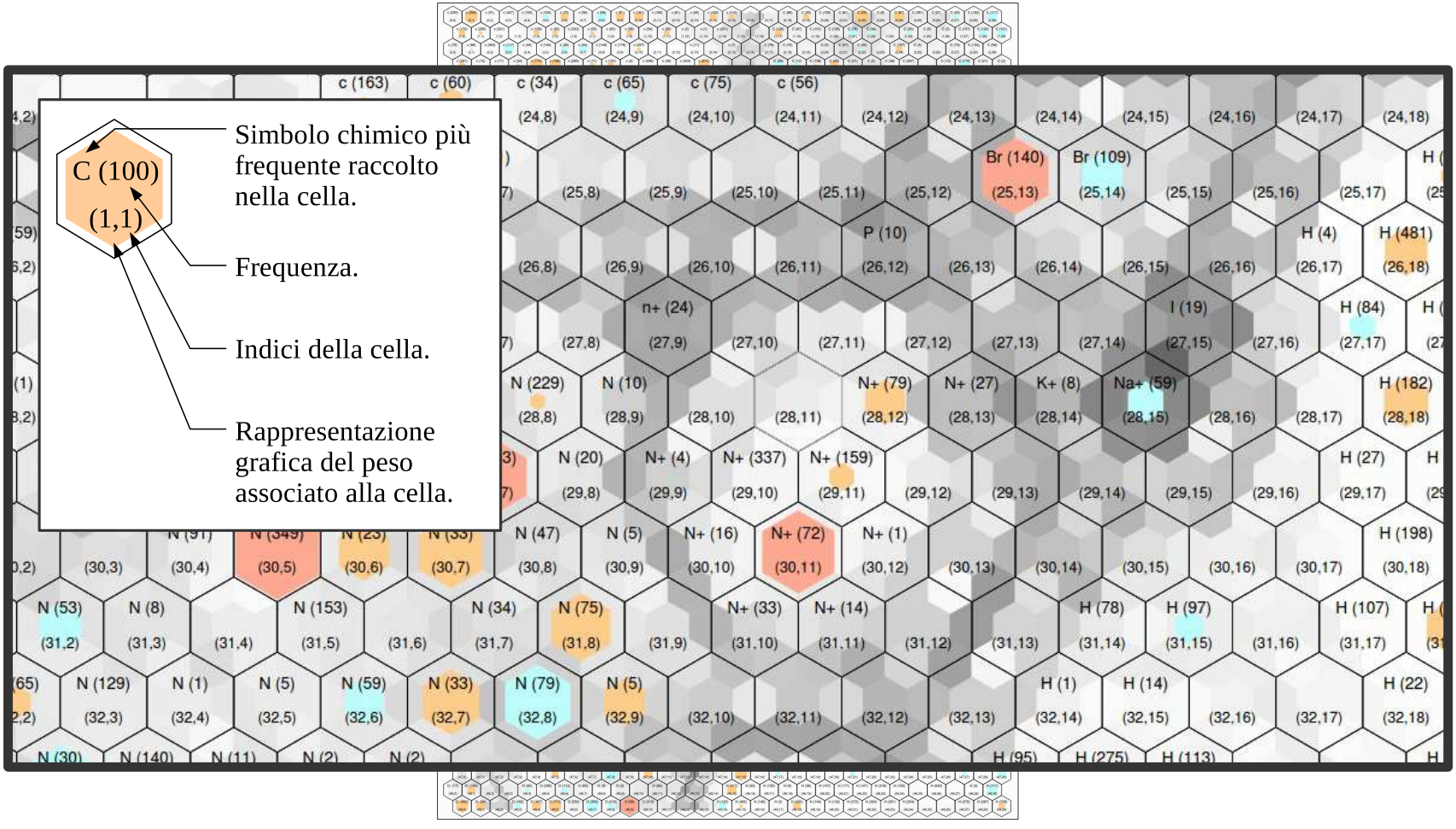
SOM 50x30 unità (K = 1500)



Si distinguono **diverse aree** che raccolgono vertici simili.

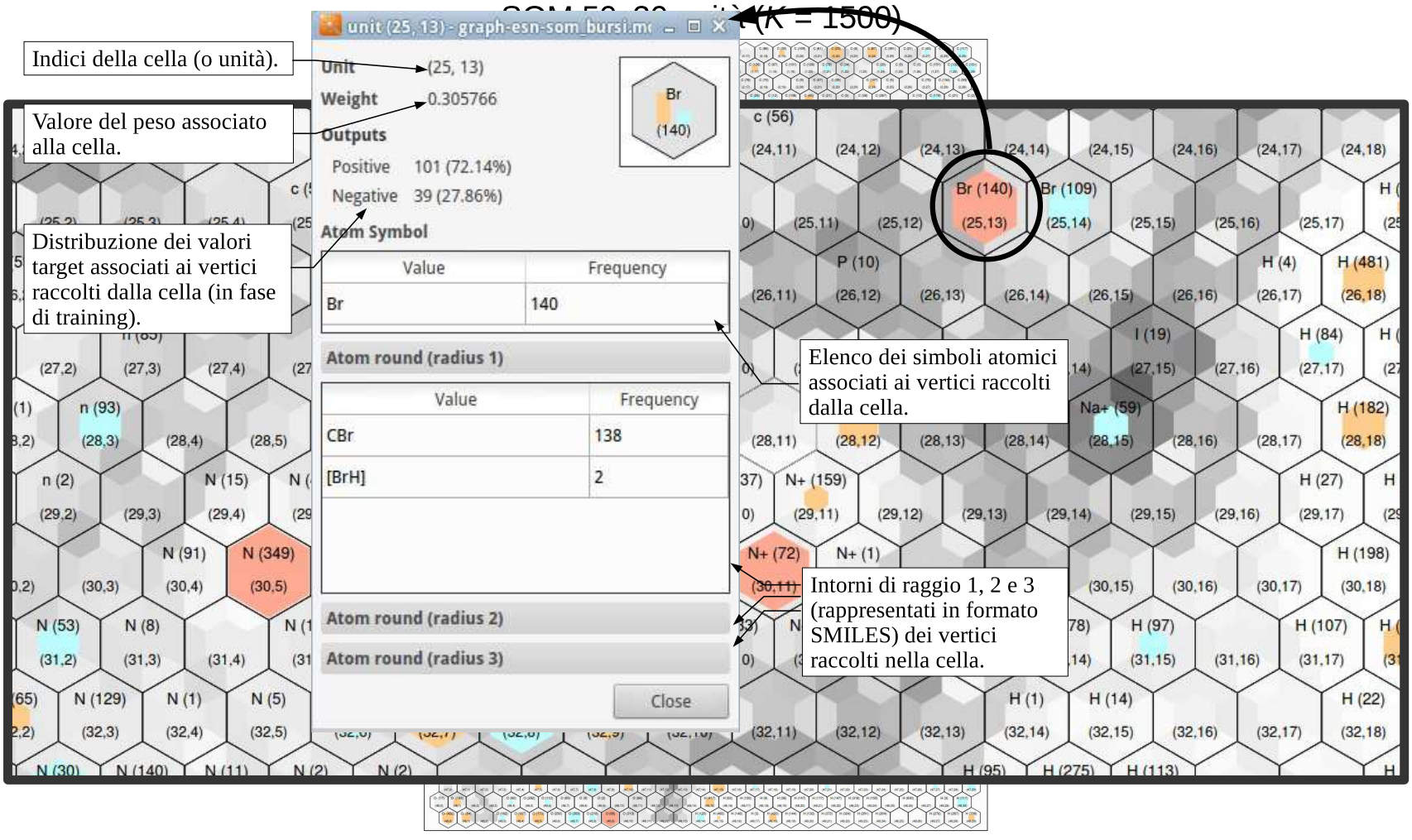
Esempio: Mappa complessiva su tutto il training set Bursi

SOM 50x30 unità (K = 1500)



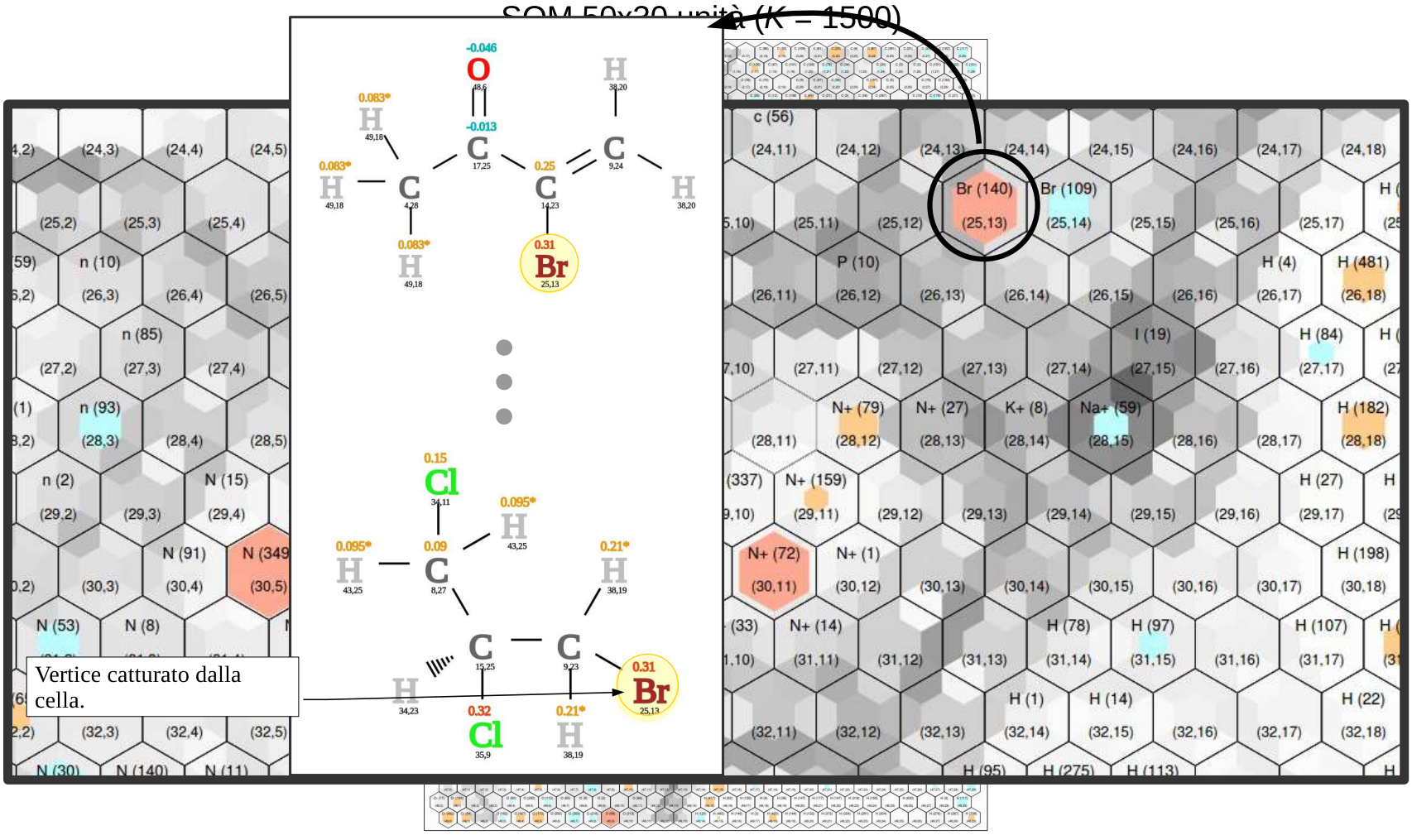
Una molecola è **classificata** in base ai pesi associati alle celle attivate.

Esempio: Info contenute nelle celle



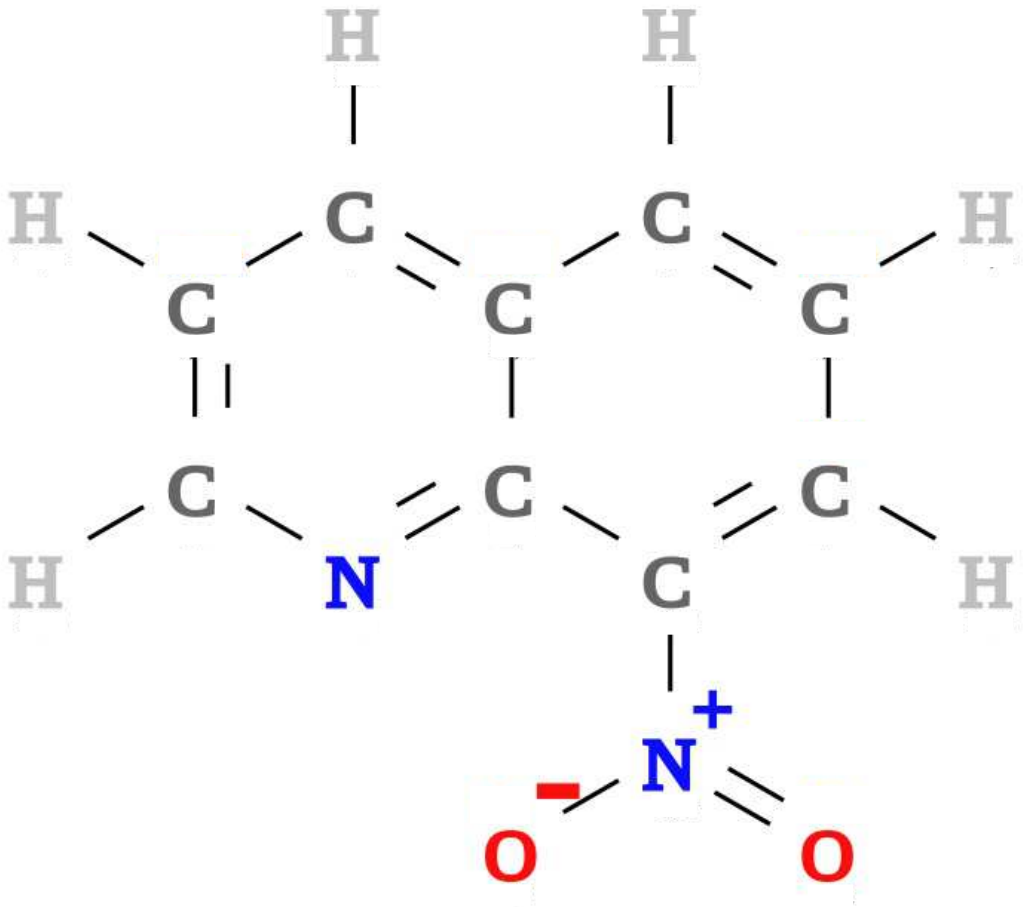
Informazioni utili a comprendere il tipo di vertici raccolti dalla cella.

Esempio: Info contenute nelle celle



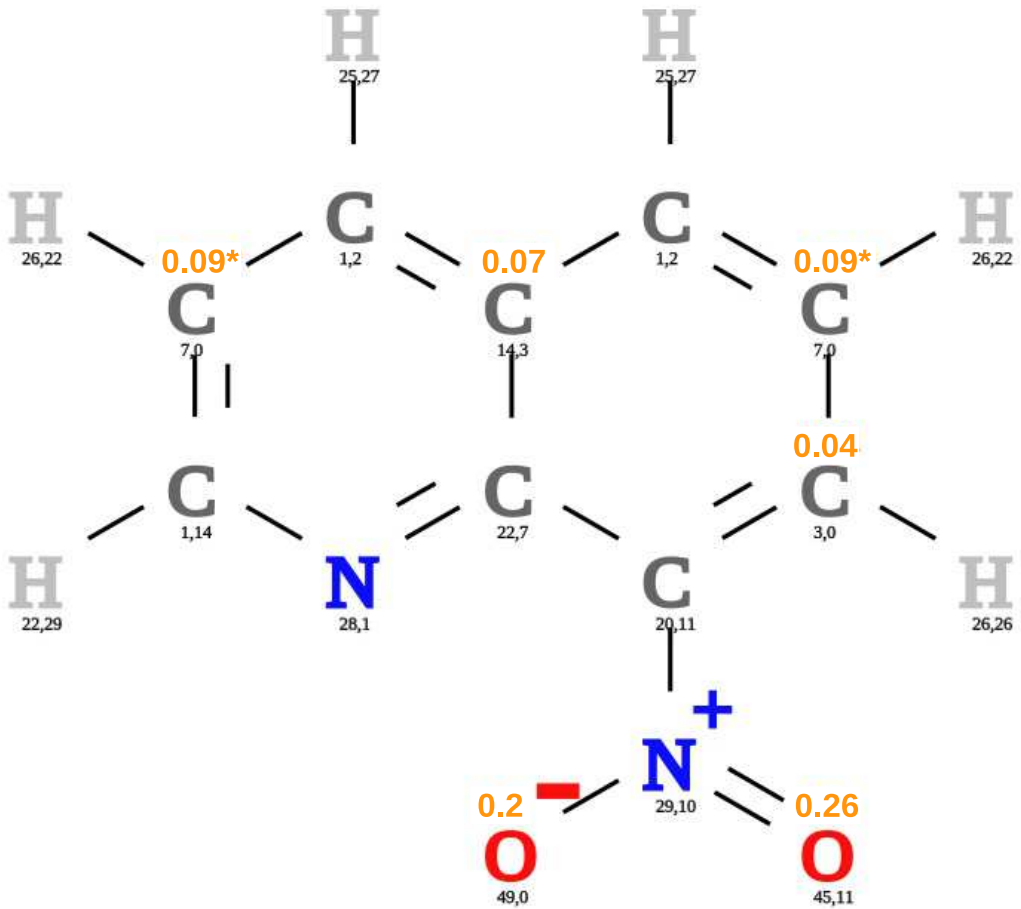
Ad esempio: un elenco delle molecole di training con un vertice nella cella.

Esempio: Classificazione di un composto



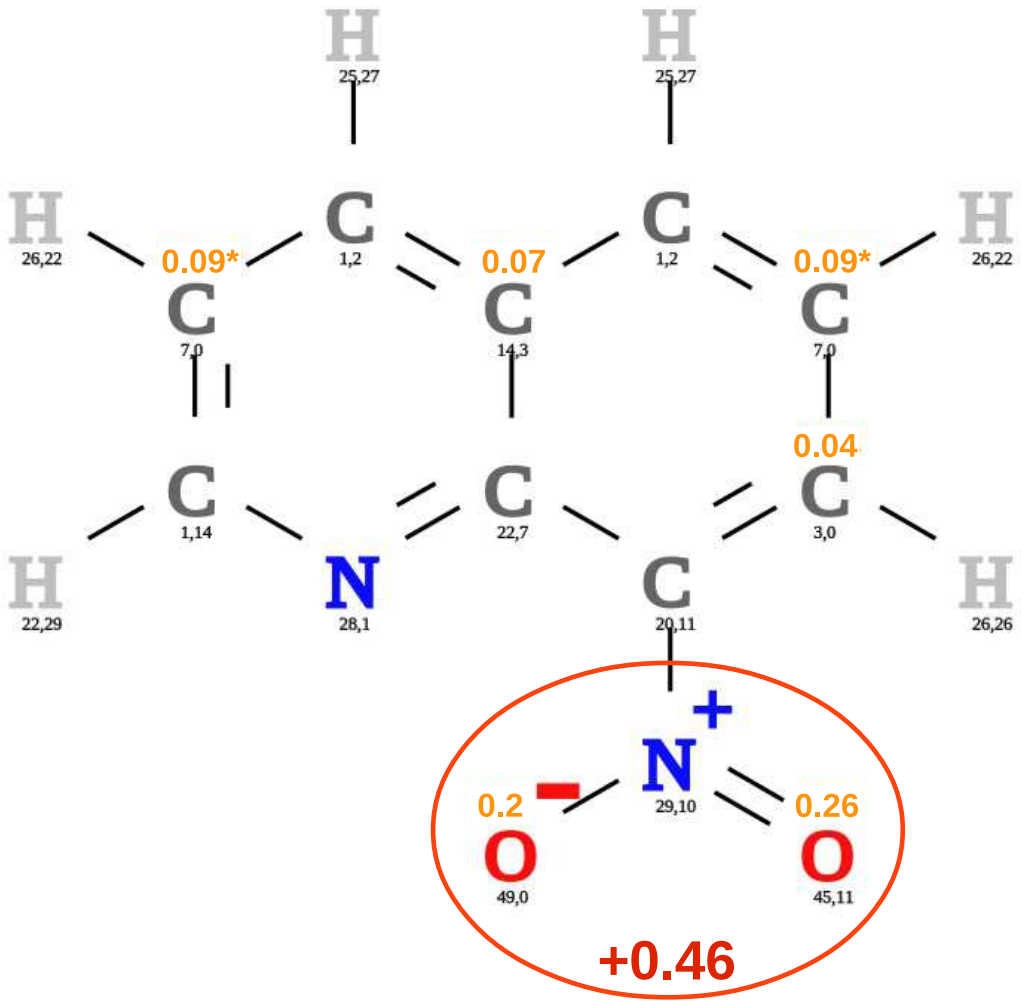
Un composto chimico da classificare.

Esempio: Classificazione di un composto



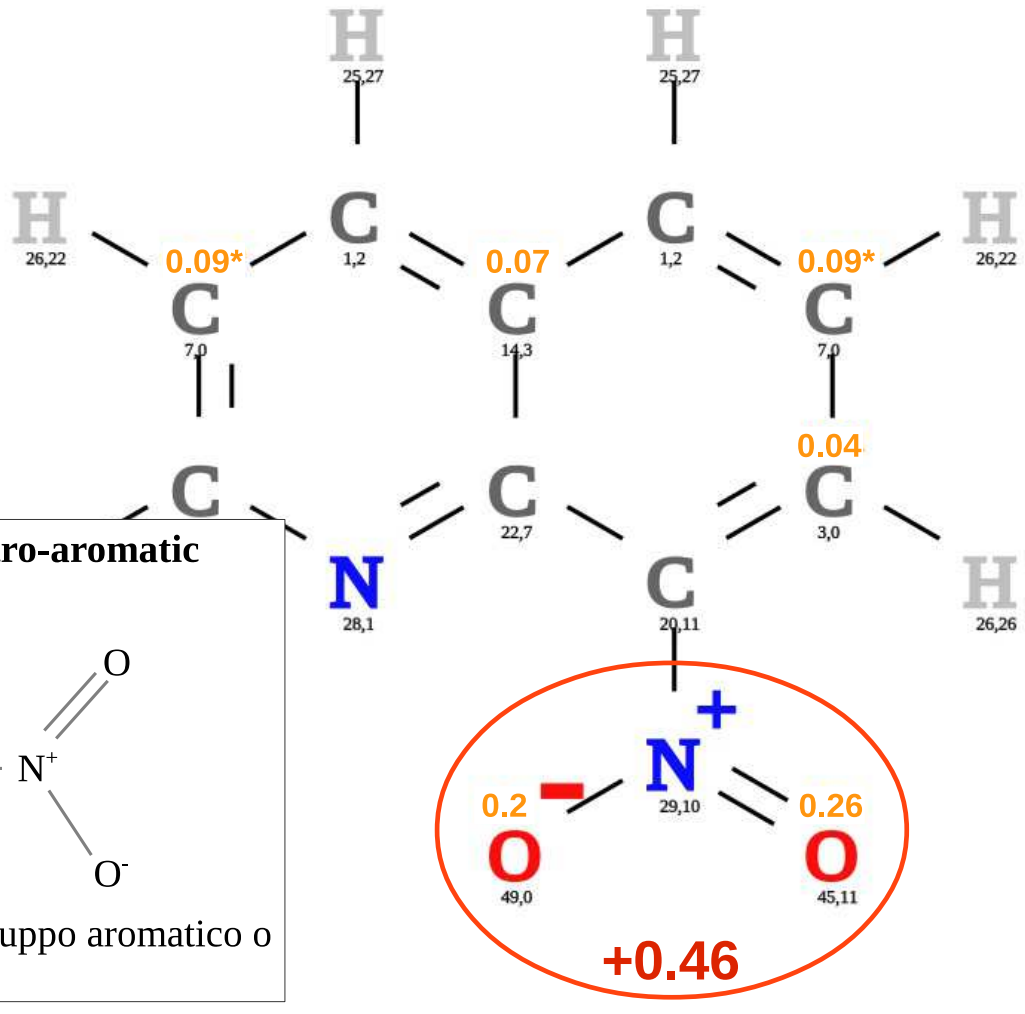
Bias $= -0.28$ | **Valore predetto** $= \text{sign}(+0.38) = +1$ (**Tossica**)

Esempio: Classificazione di un composto

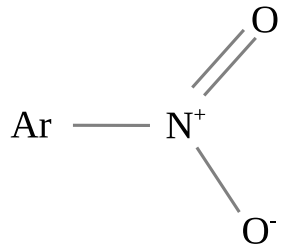


Bias $= -0.28$ | **Valore predetto** $= \text{sign}(+0.38) = +1$ (**Tossica**)

Esempio: Classificazione di un composto



SA_27: Nitro-aromatic



Ar = qualsiasi gruppo aromatico o eteroaromatico

Bias $= -0.28$ | **Valore predetto** $= \text{sign}(+0.38) = +1$ (**Tossica**)

Conclusioni

Si è introdotto un nuovo modello di **Reservoir Computing per grafi**, chiamato GraphESN-SOM, che combina diversi aspetti:

- Una nuova e innovativa **State Mapping Function**: fornisce una **rappresentazione grafica** del modello a supporto dell'analisi qualitativa.
- Training del readout via **Elastic Net**: **pruning** per evidenziare gli elementi determinanti nella risposta restituita dal sistema.
- È possibile visualizzare **direttamente sulla struttura** da elaborare il modo in cui un modello ottiene una predizione.
- Vantaggi tipici del Reservoir Computing: **efficienza** e **buone performance** sperimentali.
- Differente da metodi basati su **descrittori**: GraphESN-SOM si **adatta in modo automatico** al problema da affrontare.