

# CONTRASTIVE 3D PROTEIN PREDICTION

Strumenti Formali per la Bioinformatica – A.A. 23/24

# INTRODUZIONE



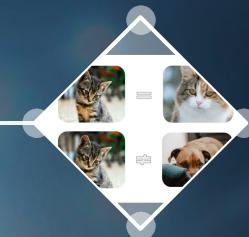
## PROTEOMICA

La predizione delle strutture proteiche è fondamentale per la comprensione dei processi biologici, nonché per lo sviluppo di terapie mediche mirate e farmaci innovativi.



## BIOCOMPUTAZIONE

Le **GNNs** si rivelano strumenti potenti per analizzare e predire le strutture proteiche, mentre l'elaborazione delle sequenze di DNA adotta modelli avanzati come DNABERT-2.

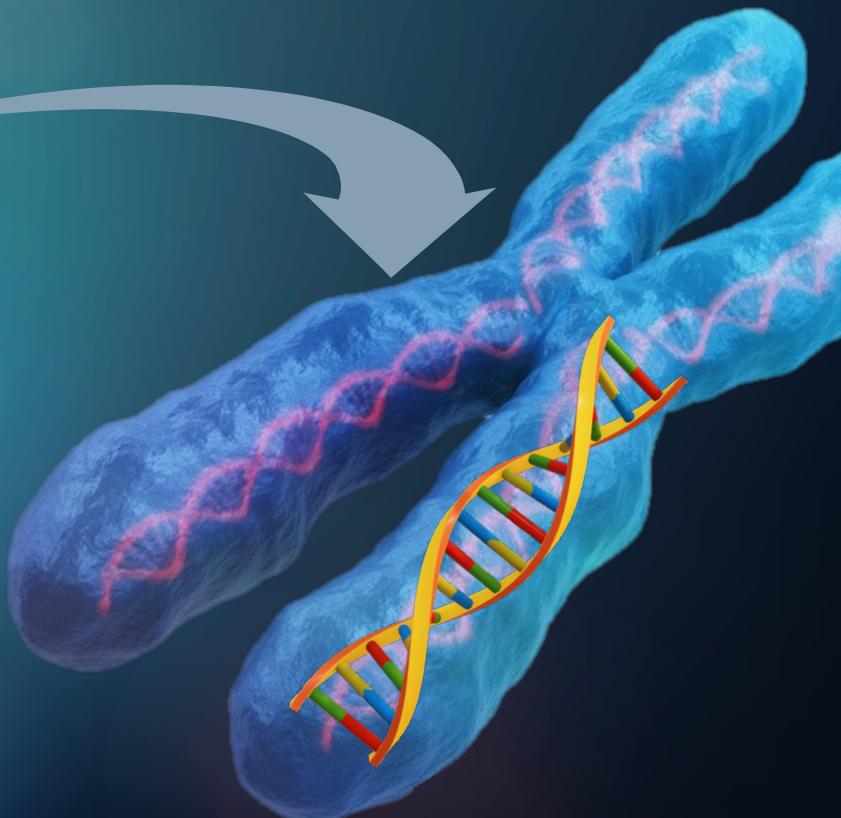
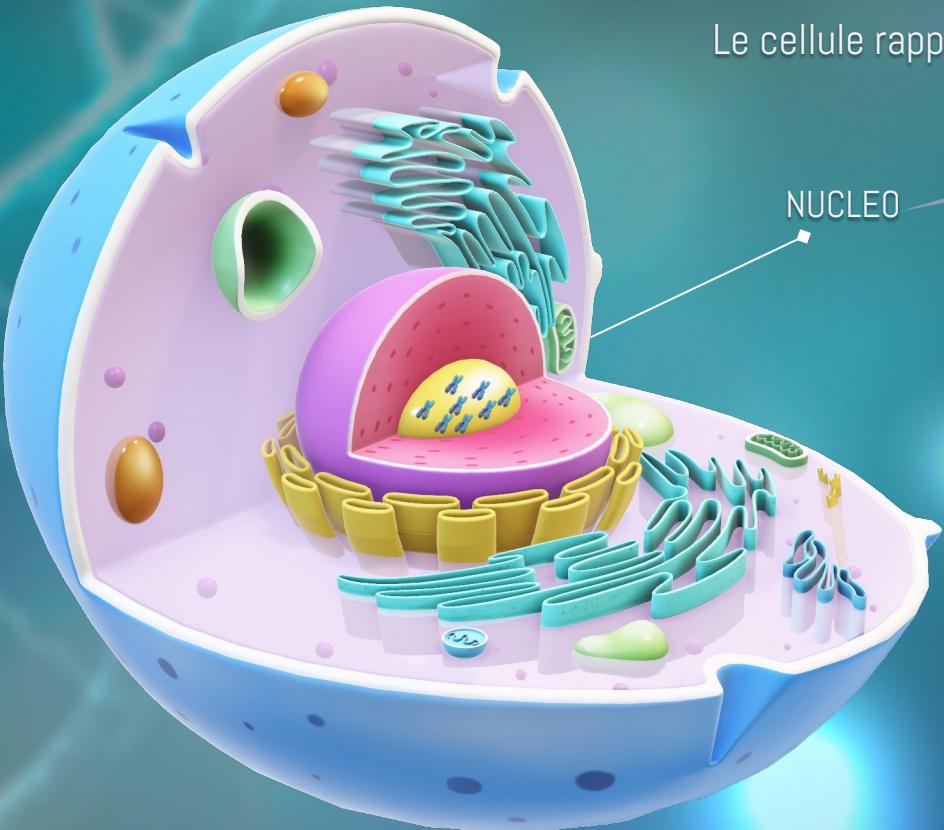


## APPRENDIMENTO CONTRASTIVO

L'approccio contrastivo ha dimostrato efficacia nell'apprendimento di rappresentazioni informative in svariati ambiti come la computer vision.

# PROTEINE

Le cellule rappresentano le unità di base di tutto il tessuto vivente.





# PROTEINE

DNA

(Trascrizione + splicing)

RNA

(Traduzione)

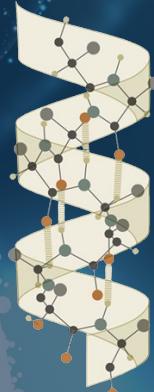
PROTEINA

GTGCATCTGACTCCTGAGGAGAAG  
CACGTAGACTGAGGACTCCTCTTC

GUG CAU CUG ACU CCU GAG GAG AAG  
V H L T P E E K

# PROTEINE

## 02 STRUTTURA SECONDARIA



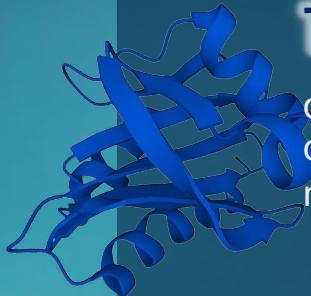
si riferisce ai modelli di piegamento localizzato o ripetitivo della catena polipeptidica ( $\alpha$ -elica, foglietto- $\beta$ ).

## 01 STRUTTURA PRIMARIA



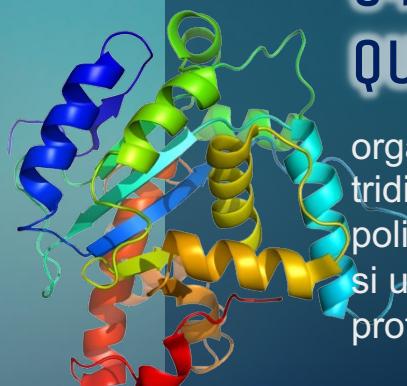
sequenza lineare degli aminoacidi nella catena polipeptidica.

## 03 STRUTTURA TERZIARIA



disposizione tridimensionale della catena polipeptidica nel suo complesso

## 04 STRUTTURA QUATERNARIA



organizzazione tridimensionale di più catene polipeptidiche (subunità) che si uniscono per formare una proteina complessa.

# PROTEIN FOLDING

**Dogma di Anfinsen:** la sequenza di amminoacidi di una proteina racchiude tutte le informazioni essenziali per determinare la sua corretta struttura tridimensionale.



HOMOLOGY MODEL

Proteine con sequenze simili presenteranno strutture altrettanto simili.



FOLD RECOGNITION

Confronta la sequenza di una struttura sconosciuta con tutte le strutture note in un DB.



AB INITIO

Mira a costruire modelli tridimensionali partendo da zero.

# INTRODUZIONE AL MACHINE LEARNING



CRISTALLOGRAFIA A  
RAGGI X



MICROSCOPIA  
ELETTRONICA



SPETTROSCOPIA DI  
RISONANZA MAGNETICA  
NUCLEA



DEEP  
LEARNING



Costi ridotti Flessibilità

Velocità

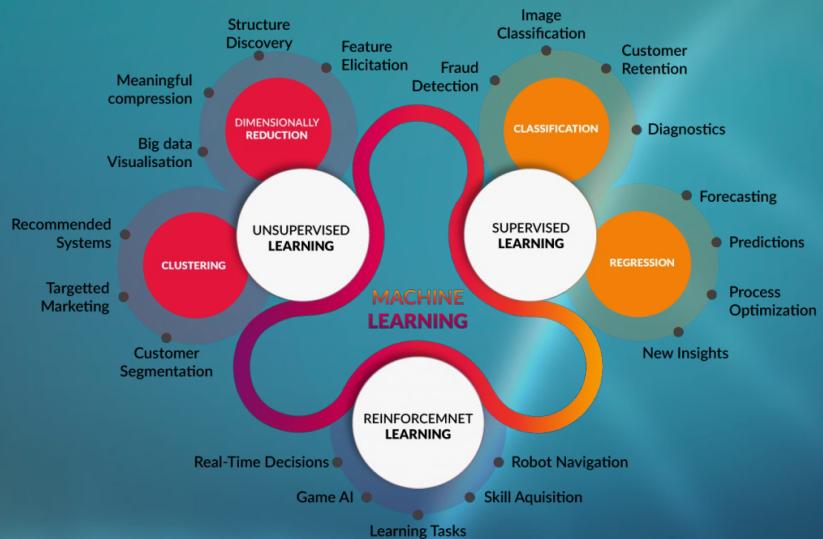
Scalabilità

Risorse disponibili

VANTAGGI

# MACHINE LEARNING

**Machine learning:** branca dell'intelligenza artificiale che si occupa dello sviluppo di algoritmi capaci di imparare dai dati e di generalizzare sulla base di essi in modo da applicare le conoscenze acquisite per fare previsioni o prendere decisioni su nuovi dati.

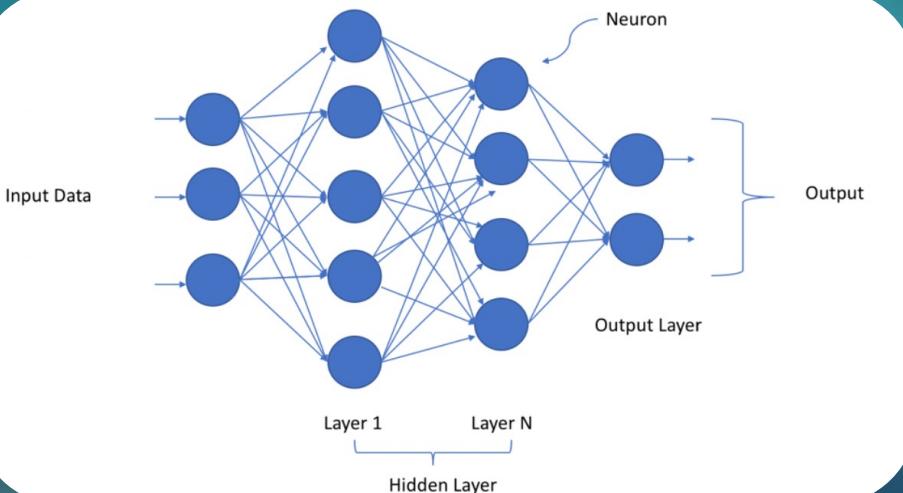


## TIPOLOGIE DI ML

- Supervisionato
- Non supervisionato
- Semi-supervisionato
- Apprendimento per rinforzo

# DEEP LEARNING

**Deep learning:** specializzazione del machine learning che si concentra sullo studio, sviluppo e addestramento di reti neurali profonde su grandi quantità di dati per compiere tasks complesse.



- **Input Layer:** i dati vengono elaborati attraverso una funzione di attivazione iniziale che può introdurre non linearità.
- **Hidden Layer:** combinano linearmente l'output del livello di input e applicano una funzione di attivazione.
- **Output Layer:** fornisce il risultato finale della rete.

# GRAPH NEURAL NETWORKS

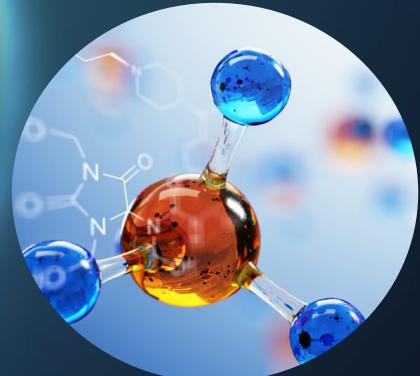
Classe di modelli di deep learning progettati per interpretare le relazioni non-lineari nei grafi attraverso operazioni di aggregazione e propagazione dell'informazione attraverso i nodi e gli archi del grafo. Negli ultimi anni hanno acquisito notorietà per la loro efficacia nell'affrontare diverse sfide.



RILEVAMENTO DI OGGETTI



SCOPERTA DI FARMACI



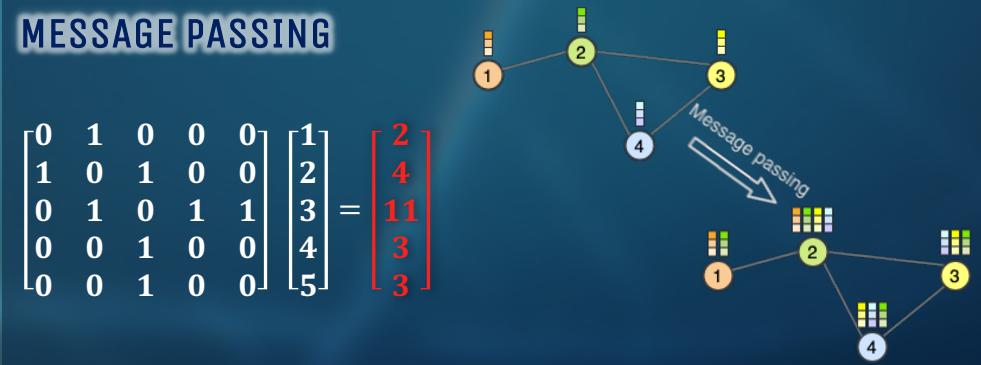
CLASSIFICAZIONE DI MOLECOLE

# MESSAGE PASSING



**MATRICE DI ADIACENZA**

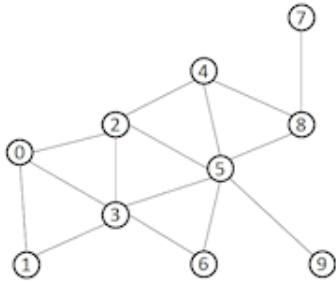
1	2	3	4	5
1	0	1	0	0
2	1	0	1	0
3	0	1	0	1
4	0	0	1	0
5	0	0	1	0



- Passaggio di informazioni tra i nodi e gli archi di un grafo durante il processo di apprendimento.
- Ad ogni passaggio, ciascun nodo aggrega le informazioni provenienti dai nodi vicini e aggiorna il proprio stato interno di conseguenza.
- Consente di catturare le relazioni strutturali e contestuali all'interno dei grafi.

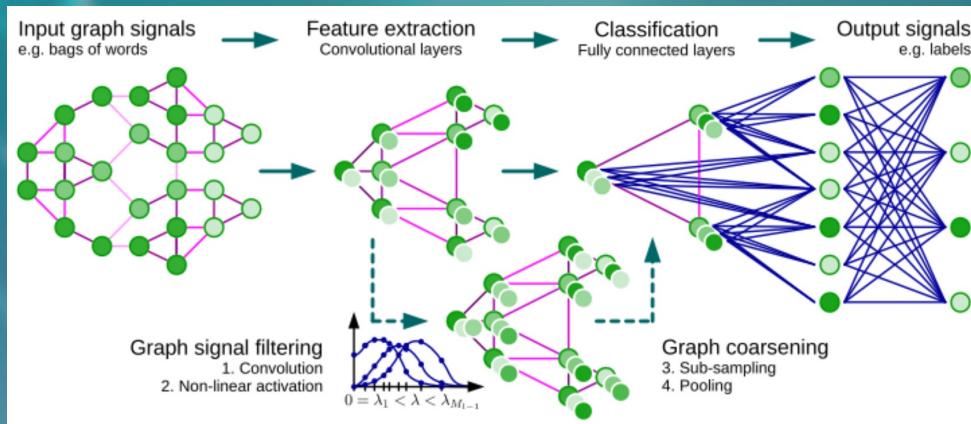
# MESSAGE PASSING

# GRAPH NEURAL NETWORKS



# GRAPH CONVOLUTIONAL NETWORKS

Classe di reti neurali progettate per l'elaborazione e l'apprendimento di dati strutturati rappresentati come grafi.



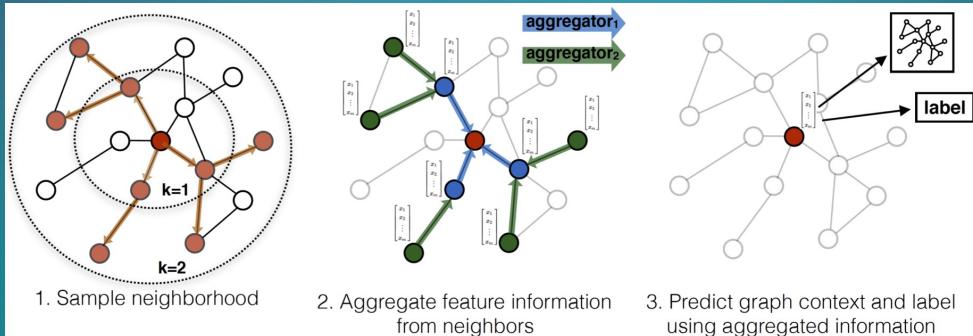
**FEATURE SMOOTHING:** processo di aggregazione e diffusione delle caratteristiche dei nodi lungo le relazioni nel grafo durante il **message passing**, consentendo una rappresentazione più ricca e complessa dei nodi stessi.

## LAYER PRINCIPALI

- **Layer di Input:** riceve in input la rappresentazione del grafo;
- **Layer di Convolution:** vengono applicate operazioni di convoluzione ai nodi del grafo;
- **Layer di Pooling (Opzionale):** utilizzato per ridurre la dimensione del grafo o per estrarre caratteristiche salienti;
- **Layer di output:** a seconda del problema, può essere progettato per compiti come la classificazione, la regressione o altre operazioni specifiche.

# GRAPHSAGE

Modello di rete neurale per l'apprendimento delle rappresentazioni dei nodi in grafi, che utilizza il campionamento e l'aggregazione delle informazioni dai vicini di ciascun nodo per generare embedding.



**OBIETTIVI:** superare le limitazioni connesse alla **superficialità** e alla **trasduttività**. Questo obiettivo viene raggiunto mediante l'introduzione di un insieme di funzioni aggregatrici  $K$ , le quali, una volta addestrate, consentono l'applicazione diretta a nuovi nodi senza la necessità di riaddestrare l'intero modello.

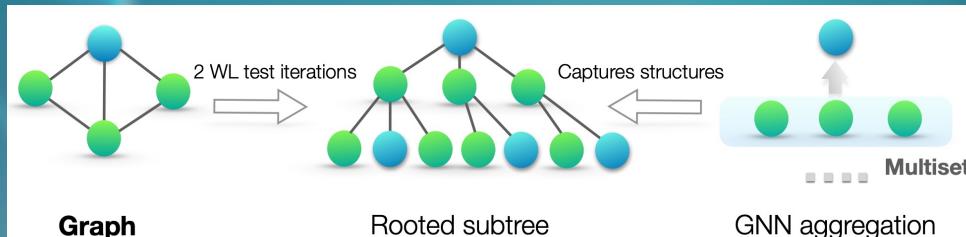
## FASI DI GRAPHSAGE

- **Campionamento dei vicini:** per ciascun nodo, vengono campionati un certo numero di vicini nel suo intorno;
- **Propagazione dei messaggi:** le informazioni dai vicini vengono aggregate per generare una rappresentazione per il nodo centrale;
- **Predizione del contesto e dell'etichetta del grafo:** utilizzo delle rappresentazioni dei nodi per prevedere il contesto o le etichette del grafo.

# GRAPH ISOMORPHISM NETWORKS

Classe di modelli di rete neurale progettati per apprendere rappresentazioni per grafi isomorfi, cioè grafi che hanno la stessa struttura, ma con etichette di nodi e archi diverse.

**ISOMORFISMO TRA GRAFI:** due grafi sono isomorfi se esiste una corrispondenza biunivoca tra i loro nodi che preserva la connettività tra di essi.



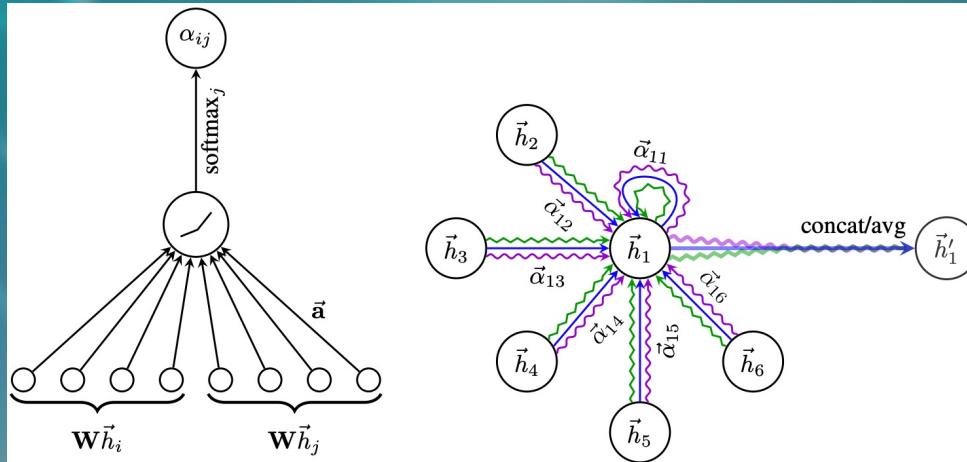
**TEST DI WEISFEILER-LEHMAN:** strumento per determinare l'isomorfismo tra grafi, che coinvolge l'aggregazione e l'hashing delle etichette aggregate dei nodi e dei loro vicini per generare nuove etichette univoche.

## FUNZIONAMENTO PRINCIPALE

- **Rappresentazione dei grafi:** i grafi di input vengono convertiti in rappresentazioni vettoriali o matriciali;
- **Propagazione dei messaggi:** i dati vengono passati attraverso la rete e successivamente le informazioni vengono aggregate da nodi e archi;
- **Apprendimento:** la rete apprende le rappresentazioni che catturano le caratteristiche dei grafi di input;
- **Output:** la rete neurale può essere utilizzata per classificare o confrontare grafi.

# GRAPH ATTENTION NETWORKS

Classe di modelli di rete neurale che operano su grafi, estendendo il concetto di reti neurali convoluzionali (CNN) ai dati basati su grafi

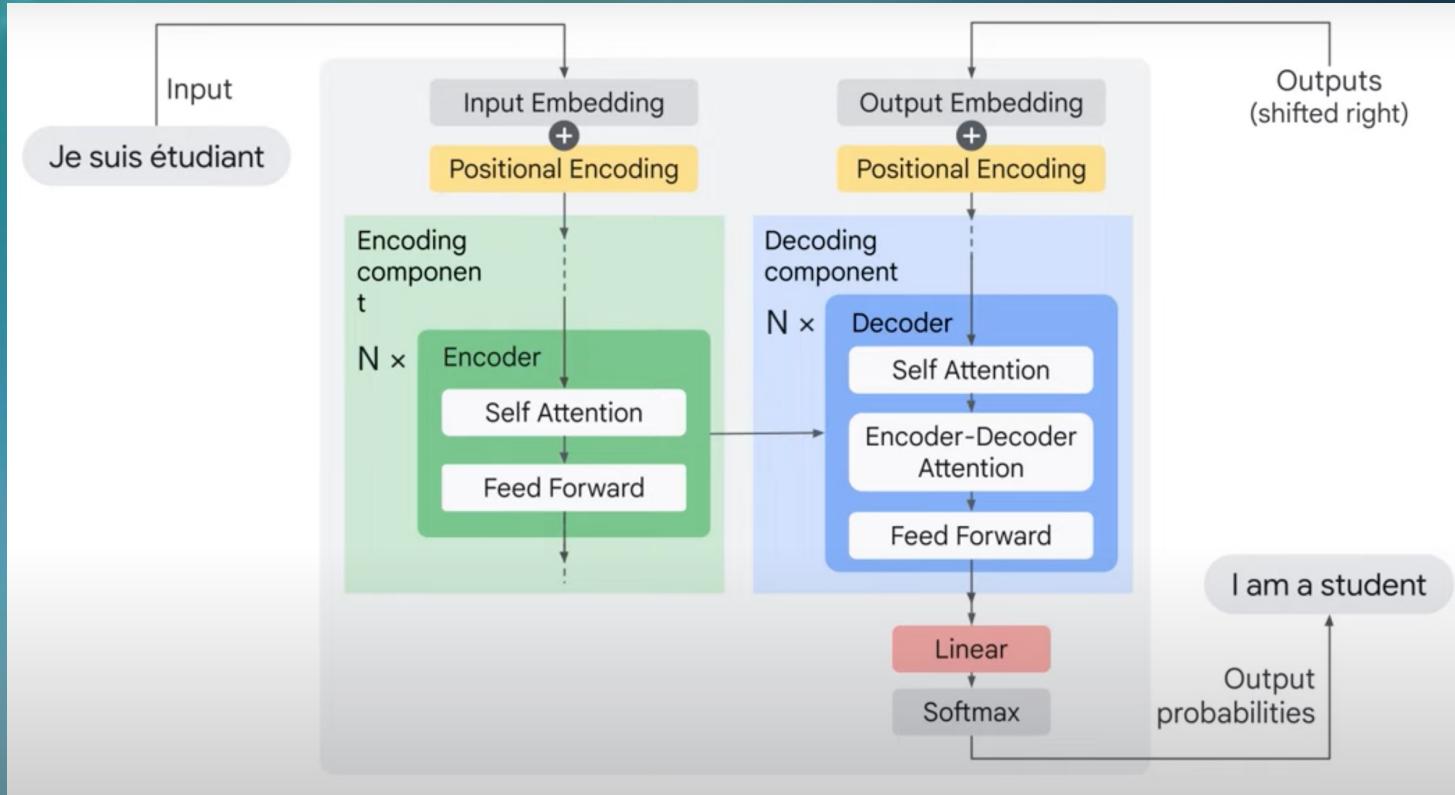


**MECCANISMO DI ATTENZIONE:** consente ai nodi di pesare diversamente le informazioni dei loro vicini in base alla rilevanza delle caratteristiche, migliorando così l'aggregazione delle informazioni nel grafo.

## FUNZIONAMENTO PRINCIPALE

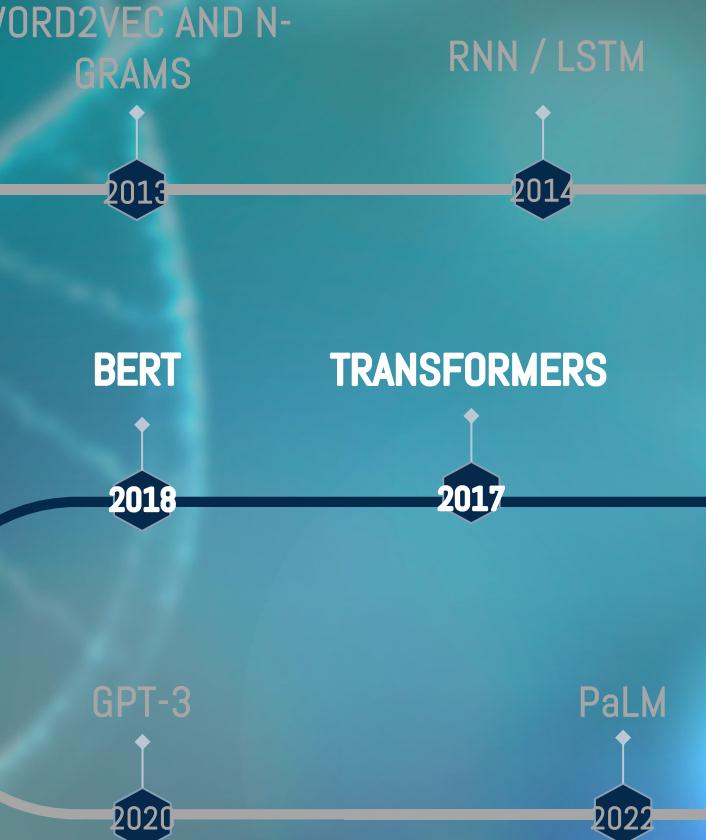
- GAT utilizza meccanismi di **attenzione** per calcolare pesi differenti per i vicini di un nodo durante l'aggregazione delle informazioni;
- Ogni nodo calcola i **pesi di attenzione** per i suoi vicini basandosi sulle caratteristiche dei nodi stessi e dei vicini.
- Le caratteristiche dei vicini sono pesate con i pesi di attenzione e aggregate per ottenere la **rappresentazione del nodo**.

# TRANSFORMERS



# BERT

(Bidirectional Encoder Representations from Transformers)



## MASKED LANGUAGE MODELING

The man went to the [MASK] to buy a [MASK] of milk

STORE

GALLON

## NEXT SENTENCE PREDICTION

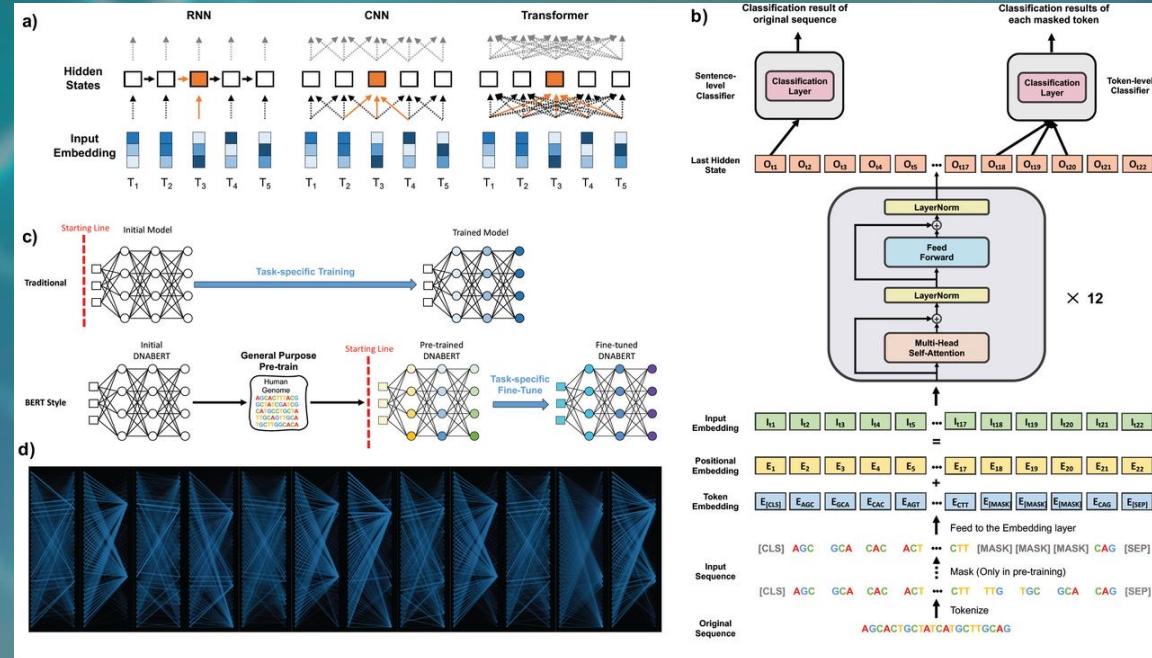
SENTENCE A

The man went to the store.

SENTENCE B

He bought a gallon of milk.

# DNABERT

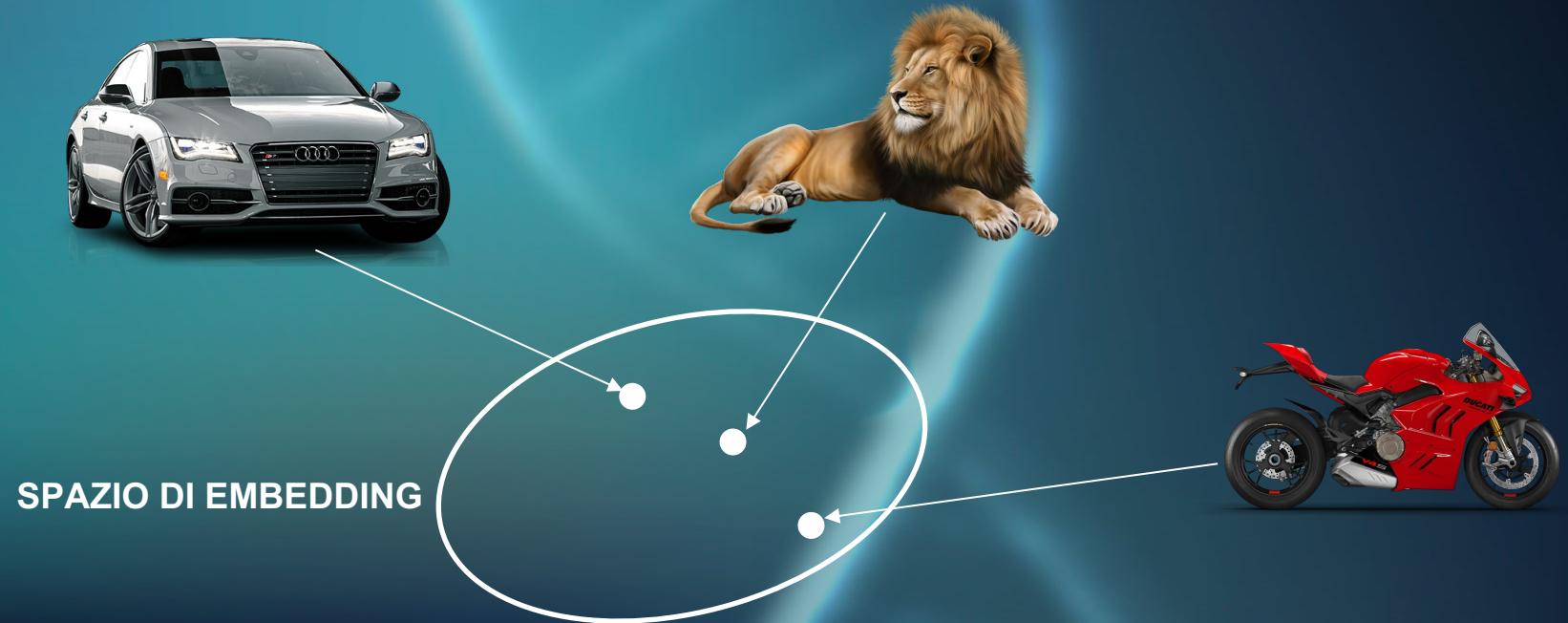


- **INPUT:** insieme di sequenze rappresentate come token  $k$ -mer.
- **RAPPRESENTAZIONE:** ogni sequenza è rappresentata come una matrice incorporando ciascun token in un vettore numerico e le informazioni contestuali vengono catturate applicando la multi-head self-attention.
- **PRE-TRAINING:** simile a BERT, tralasciando la fase di NPS.
- **PREDICTION:** la lunghezza della sequenza viene modificata e il modello è costretto a prevedere  $k$  token contigui.

**DNABERT-2:** variante di DNABERT che adotta la metodologia **Byte Pair Encoding (BPE)**, un metodo statistico per la costruzione dei token, invece della tokenizzazione  $k$ -mer, raggiungendo prestazioni paragonabili ai modelli all'avanguardia pur essendo più piccolo.

# CONTRASTIVE LEARNING

Tecnica di apprendimento automatico basata sul principio secondo il quale le coppie della stessa classe mostrano una stretta vicinanza tra loro nello spazio di embedding, mentre quelle appartenenti a classi differenti siano posizionate più distanti tra loro.



# CONTRASTIVE LEARNING

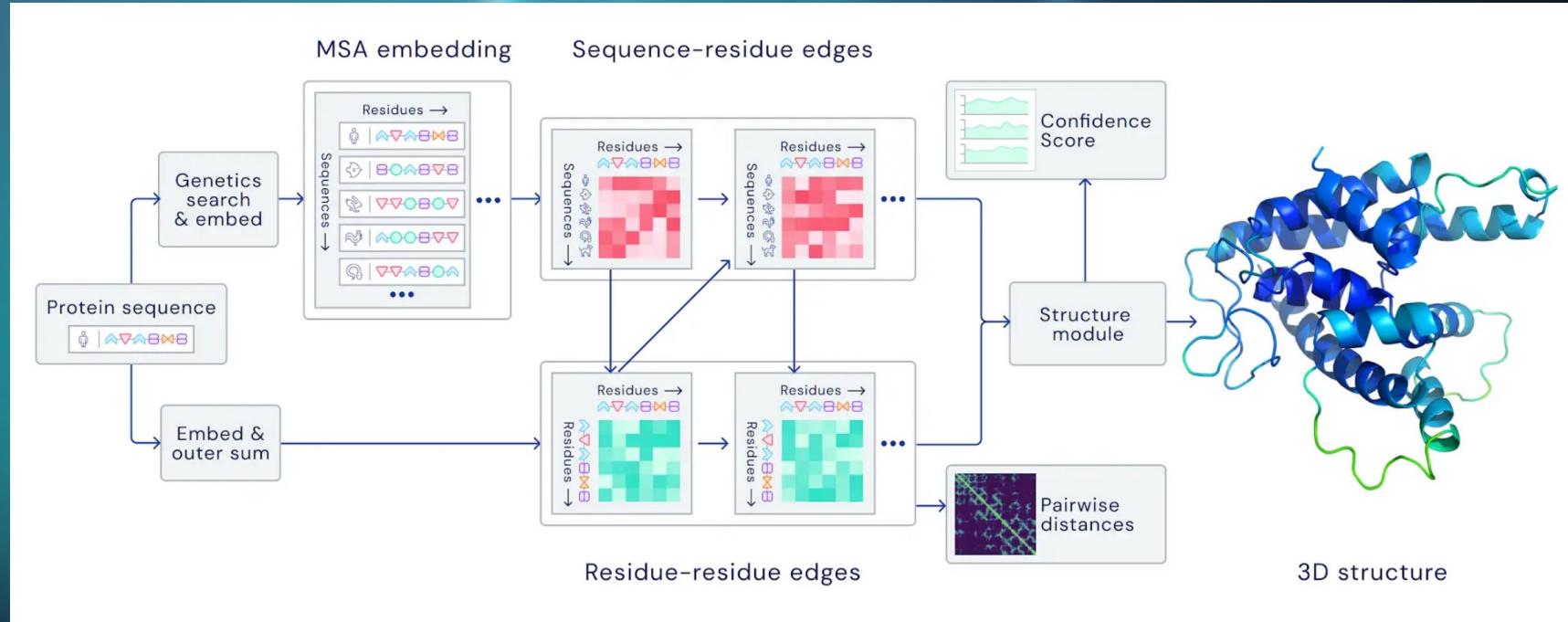
Tecnica di apprendimento automatico basata sul principio secondo il quale le coppie della stessa classe mostrano una stretta vicinanza tra loro nello spazio di embedding, mentre quelle appartenenti a classi differenti siano posizionate più distanti tra loro.



# ALPHAFOLD

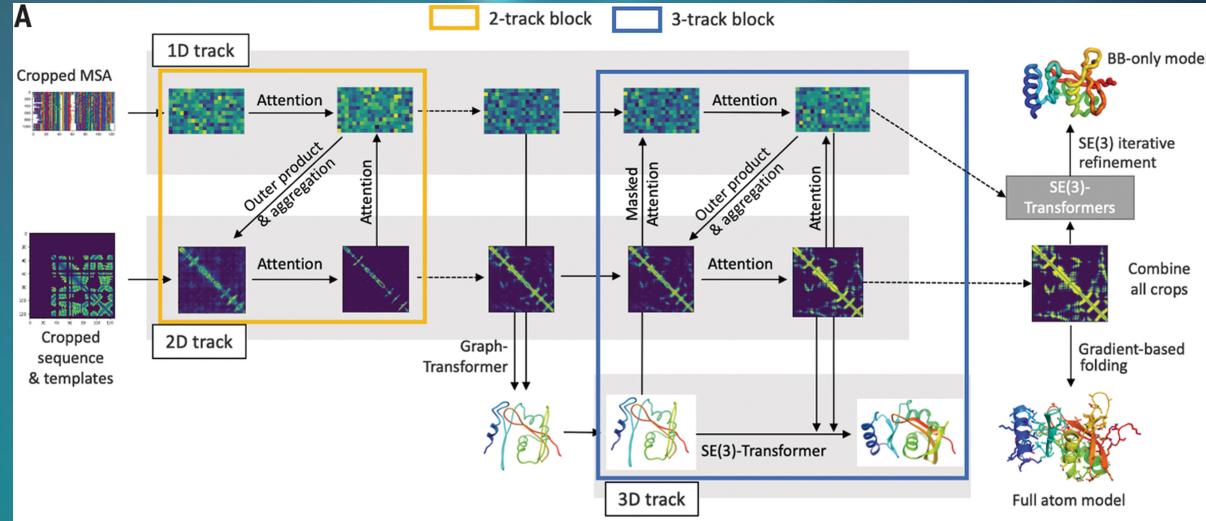
# LAVORI CORRELATI

AlphaFold: tecnologia sviluppata da DeepMind per la predizione accurata della struttura tridimensionale delle proteine.



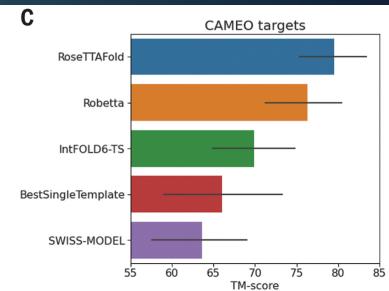
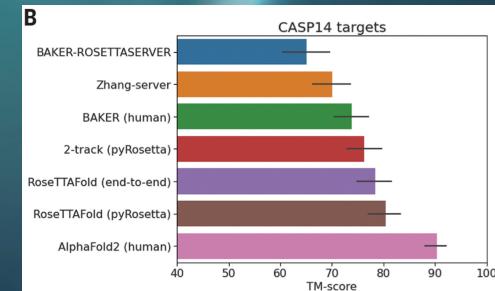
# ROSETTA FOLD

# LAVORI CORRELATI



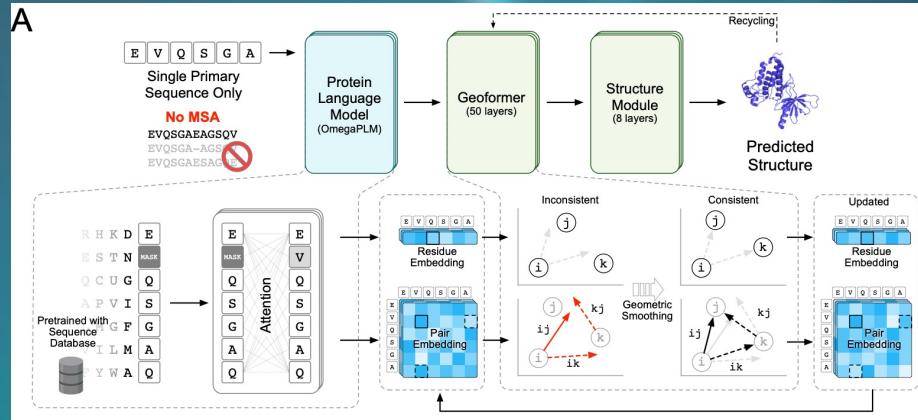
Con l'obiettivo di migliorare le previsioni delle strutture proteiche e promuovere il design delle proteine, i ricercatori hanno adottato una "rete a due tracce" e ampliato l'architettura con una terza traccia in 3D per stabilire connessioni più strette tra sequenza, distanze, orientamenti residuo-residuo e coordinate atomiche.

Risultati promettenti specialmente nei test di benchmark su CAMEO, mostrando una maggiore precisione rispetto ad altri server di previsione della struttura.



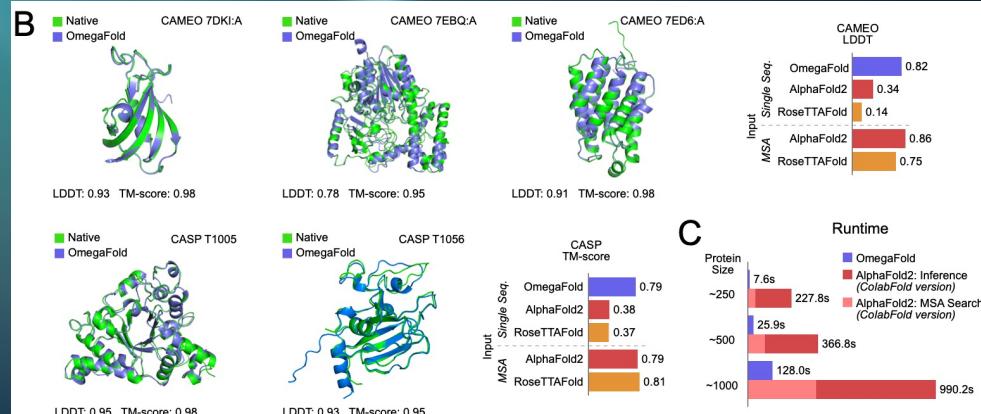
# OMEGAFOOLD

# LAVORI CORRELATI



Il blocco base dell'architettura è **OmegaPLM**, un modello di linguaggio proteico, capace di catturare informazioni strutturali e funzionali codificate nell'amminoacido sequenze attraverso gli incorporamenti. Le sequenze vengono poi immesse in **Geoformer** per distillare ulteriormente le relazioni strutturali e fisiche a coppie tra gli amminoacidi. Infine, un **modulo strutturale** prevede le coordinate 3D di tutti gli atomi pesanti.

Notevole capacità di predire con precisione le strutture proteiche su dataset di riferimento come **CASP** e **CAMEO**. I risultati, ottenuti senza l'uso di MSA e basati su una singola sequenza di input, hanno mostrato un livello di accuratezza comparabile o superiore ai metodi avanzati basati su MSA, come AlphaFold e RoseTTAFold.

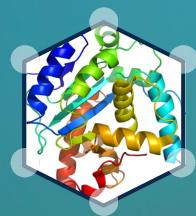


# METODI CONTRASTIVI

# LAVORI CORRELATI

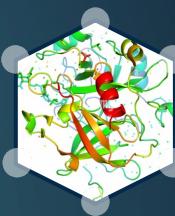
## SIMGRACE

Framework per il learning di rappresentazioni grafiche che utilizza l'apprendimento contrastivo senza data augmentation. Include **perturbazione dell'encoder**, **projection head** e **loss contrastiva** per migliorare la coerenza delle rappresentazioni apprese.



## HERMOSILLA & ROPINSKI

Il modello viene addestrato per massimizzare la similarità tra rappresentazioni derivanti da sottostrutture all'interno della stessa proteina, mentre minimizza la similarità tra sottostrutture provenienti da proteine diverse.



## GEARNET

Encoder structure-based che integra informazioni spaziali attraverso uno sparse edge message mechanism. Utilizza un framework di CL per allineare le rappresentazioni delle sottostrutture proteiche nello spazio latente, facilitando il passaggio di messaggi a livello degli archi e massimizzando le informazioni reciproche tra le visualizzazioni biologicamente correlate.

# DATASET

Il dataset utilizzato costituisce un sottoinsieme **AlphaFold Protein Structure Database**, riguardante specificamente i proteomi umani e comprendendo previsioni di circa **24.000 proteine umane**.

01

## PRE-PROCESSING

Utilizzo di file **PDB** e della libreria **Graphein** per la costruzione dei grafi corrispondenti.

02

## GRANULARITÀ

Grafo a **livello di atomo**: ogni atomo della proteina è un nodo del grafo, consentendo una rappresentazione più dettagliata delle interazioni atomiche presenti nella struttura proteica.

03

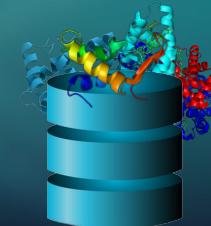
## COSTRUZIONE DEI NODI

- $\text{coords}_a$  = vettore che descrive le coordinate spaziali dell'atomo  $a$ ;
- $\text{meiler}_a$  = vettore di embeddings di Meiler;
- $\text{expasy}_a$  = proprietà aminoacidiche della scala proteica EXPASY;
- $\text{amino}_a$  = encoding one-hot dei tipi di amminoacidi associati ad  $a$ .

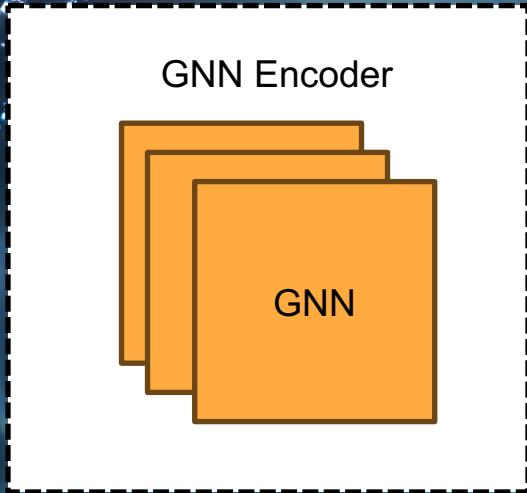
04

## COSTRUZIONE DEGLI ARCHI

- Criterio dei KNN
- Informazioni chimiche e strutturali
  - Legami ad idrogeno
  - Legami peptidici
  - Legami aromatici



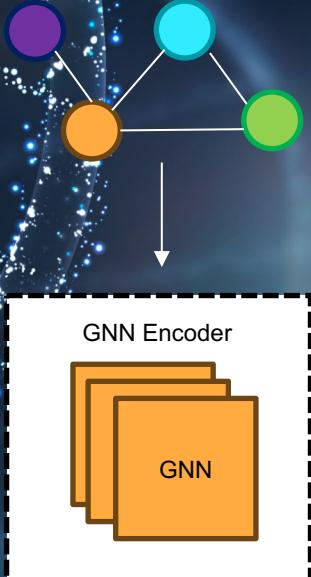
# C3DPNet



Graph-encoder basato su GNN per estrarre informazioni strutturali dai grafi.



Encoder testuale basato su DNABERT-2 per acquisire informazioni semantiche e relazioni contestuali dalla sequenza di DNA.



G

Dot product

L

Loss

AGACGTAGCA  
TCGATC

DNABERT-2

D

# FUNZIONE DI LOSS

La funzione InfoNCE valuta l'allineamento della rappresentazione del contesto con le future caratteristiche rispetto ai campioni negativi. Incentiva il modello a massimizzare la somiglianza tra il contesto e le rappresentazioni future, mentre riduce al minimo la similarità con i campioni negativi.

$$L_{InfoNCE} = -\mathbb{E}_s \left[ \log \frac{f_{\theta}(s_{t+k}, c_t)}{\sum_{s_j \in S} f_{\theta}(s_j, c_t)} \right]$$

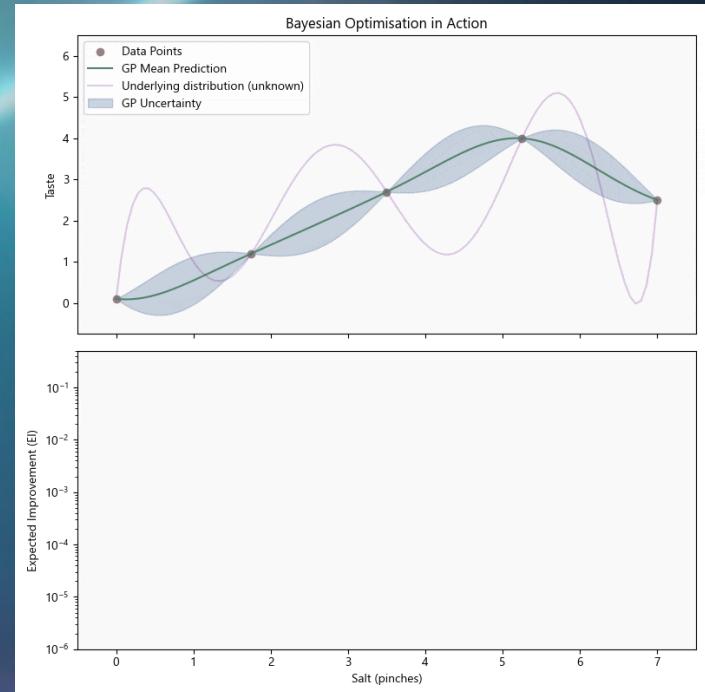
- $f_{\theta}$  = funzione che mappa la rappresentazione del contesto  $c_t$  in uno spazio di caratteristiche.
- $S$  = insieme di campioni, che include un campione positivo e campioni negativi  $N - 1$ .

# ESPERIMENTI

L'ottimizzazione bayesiana è una tecnica di ottimizzazione che utilizza il **teorema di Bayes** per guidare la ricerca di minimi di una funzione obiettivo, considerando anche l'incertezza associata ai risultati. L'obiettivo consiste in trovare il miglior set di parametri per minimizzare una funzione obiettivo, utilizzando il minor numero possibile di valutazioni.

## PROCEDIMENTO

1. Si costruisce un modello probabilistico della funzione obiettivo.
2. Si sceglie quale insieme di parametri testare nel prossimo passaggio, considerando sia l'incertezza che la conoscenza attuale.
3. Si valuta la funzione obiettivo nei punti selezionati.
4. Si aggiorna il modello probabilistico incorporando le nuove valutazioni.
5. Si ripetono i passaggi precedenti fino a raggiungere una convergenza o un limite di tempo prefissato.



# ESPERIMENTI

- Per il learning rate scheduler, è stata adottata una strategia lineare.
- Per l'ottimizzatore sono state esplorate le opzioni tra Adam, AdamW e SGD.
- La batch size è stata fissata a 8 per tutti gli esperimenti.

Learning Rate

MIN	MAX
1e-8	1e-5
0.01	0.1
3	6

Weight Decay

Layers

	GSAGE	GCN	GIN	GAT
Learning Rate	6.25e-6	3.34e-6	9.28e-6	5.12e-6
Weight Decay	0.038	0.098	0.083	0.062
Layers	6	6	6	3

Migliori combinazioni di iper-parametri trovate per ogni tipologia di GNN considerata.

# METRICHE

ACCURACY

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

PRECISION

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

F-1 SCORE

$$\frac{TP}{TP + 0.5(FP + FN)}$$

RECALL

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

# RISULTATI

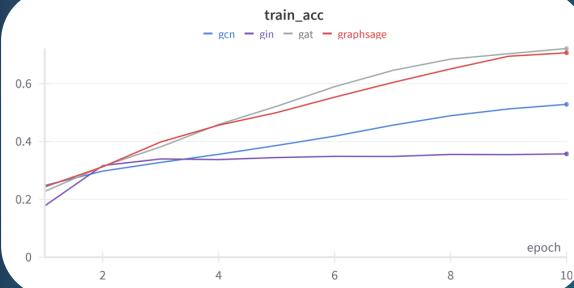
Dai risultati della valutazione delle varianti di GNNs considerate sul dataset impiegato per l'addestramento, emerge che il modello migliore si ottiene utilizzando un GNN encoder basato su GraphSAGE.

	Accuracy	Precision	F1-Score	Recall
<b>GSAGE</b>	<b>0.778</b>	<b>0.685</b>	<b>0.715</b>	<b>0.778</b>
<b>GIN</b>	0.373	0.211	0.253	0.373
<b>GAT</b>	0.731	0.617	0.653	0.731
<b>GCN</b>	0.580	0.450	0.489	0.580

*Risultati delle valutazione dei modelli sul dataset impiegato durante il training. I valori riportati sono i valori intra-batch.*

# RISULTATI

## TRAINING ACCURACY



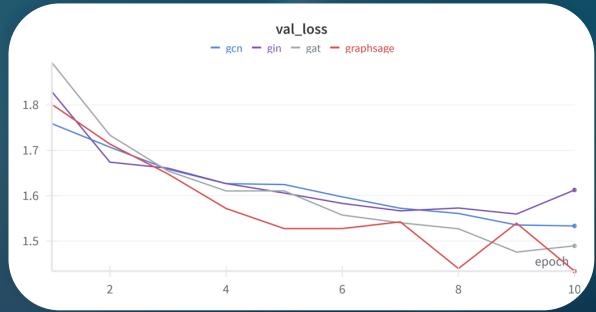
## VALIDATION ACCURACY



## TRAINING LOSS

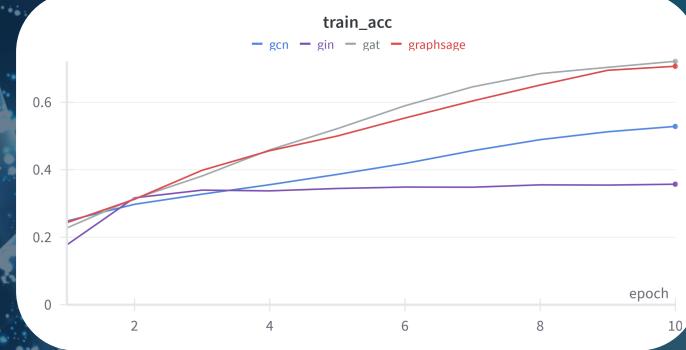


## VALIDATION LOSS

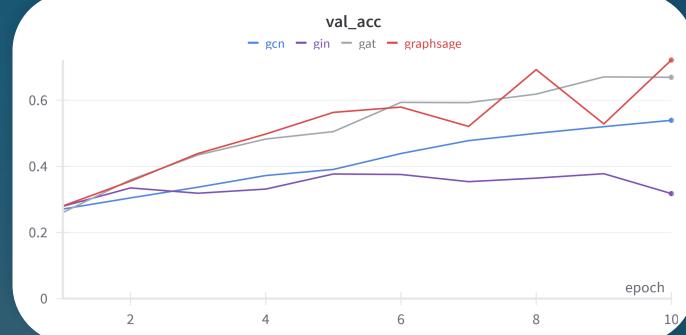


# RISULTATI

## TRAINING ACCURACY



## VALIDATION ACCURACY

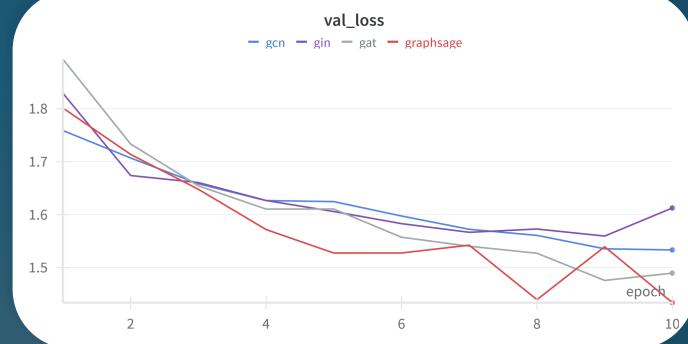


# RISULTATI

## TRAINING LOSS



## VALIDATION LOSS



# RISULTATI

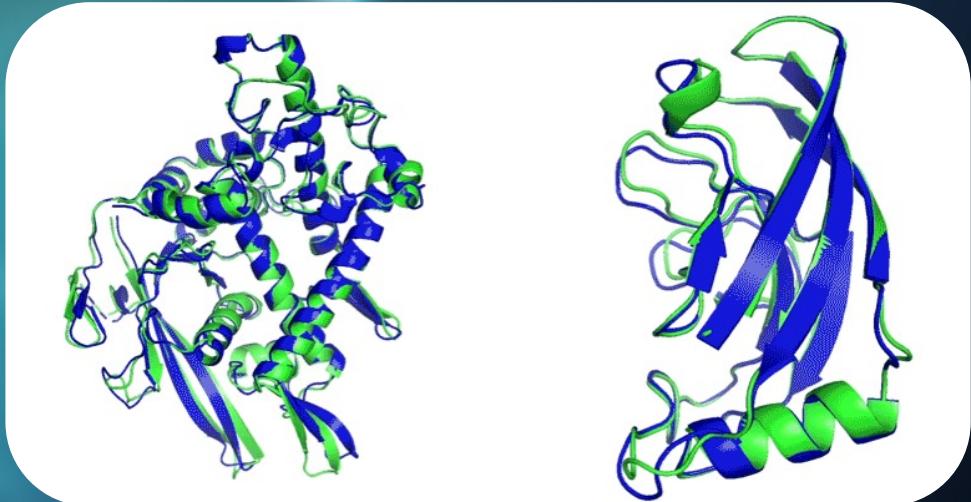


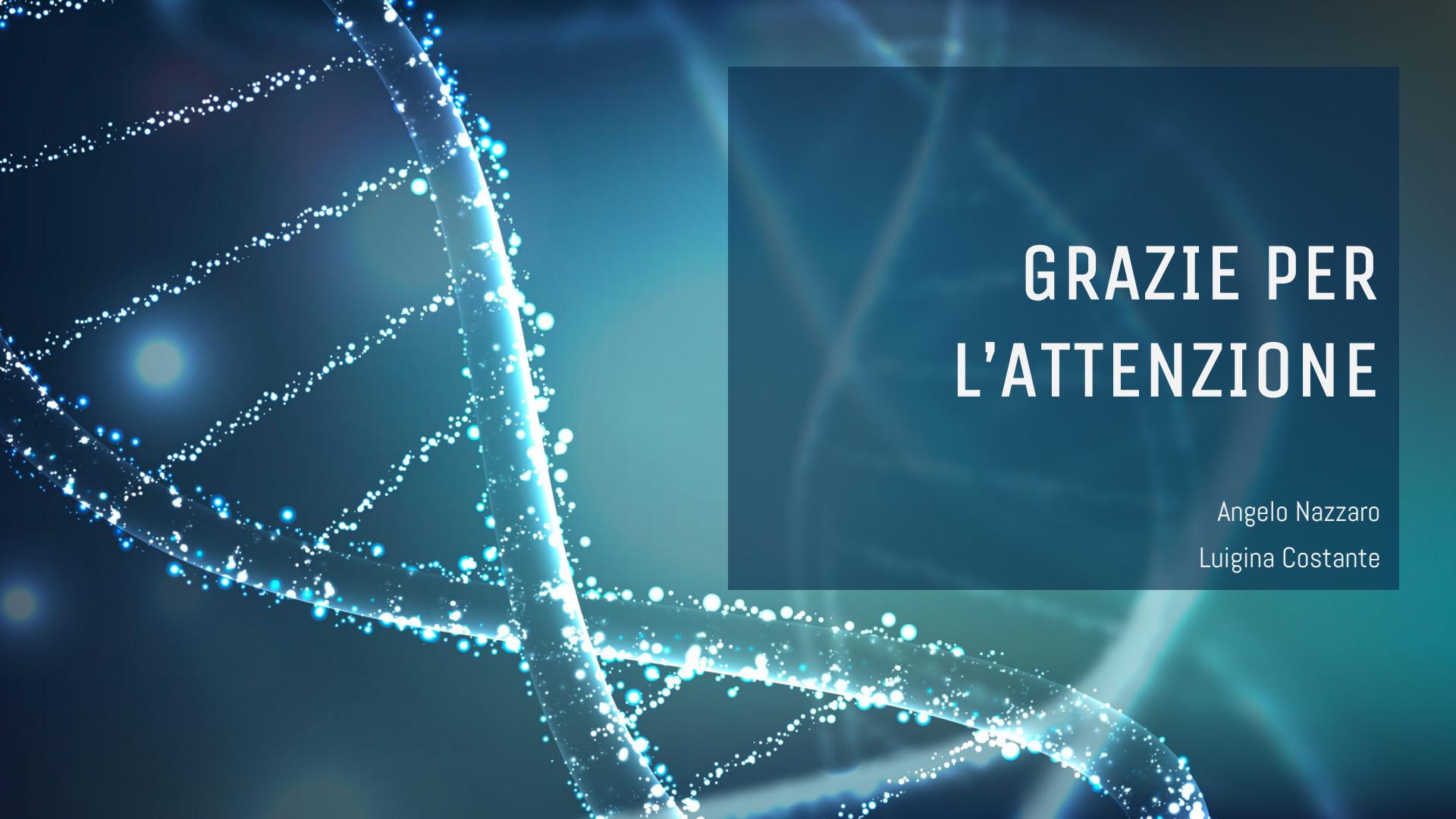
	Accuracy	Precision	F1-Score	Recall
C3DPNet	0.542	0.613	0.572	0.667
GSAGE	0.508	0.376	0.402	0.499

# CONCLUSIONI

Nonostante i risultati promettenti, riconosciamo la necessità di ulteriori miglioramenti:

- esplorare altre architetture di GNN, come **DiffPool** e **Graph-UNets**;
- valutare l'utilizzo di altre loss contrastive, come la **Sigmoid loss**;
- considerare l'utilizzo di modelli pre-addestrati specifici per le proteine, come **ProteinBERT**;
- impiegare tecniche di **data augmentation** per incrementare il numero di coppie positive;
- estendere la valutazione del modello su una vasta gamma di **dataset proteici**.





**GRAZIE PER  
L'ATTENZIONE**

Angelo Nazzaro  
Luigina Costante