
Università degli Studi di Padova – Dipartimento di Ingegneria dell'Informazione
Corso di Laurea in Ingegneria Informatica

Relazione per la prova finale
***«Analisi comparativa di modelli di link prediction
per la drug discovery»***

Relatore: Prof. Stefano Marchesin

Laureanda: Sara Gandini
Matricola n.: 2101768

Padova, 21/07/2026

Limiti della ricerca tradizionale

- processo complesso e dispendioso
- tasso di fallimento altissimo



Tecnologie ad alto rendimento

- moli di dati biomedici e farmacologici senza precedenti
- Network Medicine

Biologia dei sistemi

- sistemi cellulari e fisiologici sono rappresentati come reti



LINK PREDICTION

La modellazione a grafo permette di riformulare la ricerca farmacologica come un problema di Link Prediction



POTENZIALITA'

Riduce drasticamente lo spazio di ricerca abbattendo i costi di screening

Individua nuove applicazioni terapeutiche per molecole già approvate e sicure

Prevede interazioni e reazioni avverse studiando gli effetti combinati su bersagli multipli.

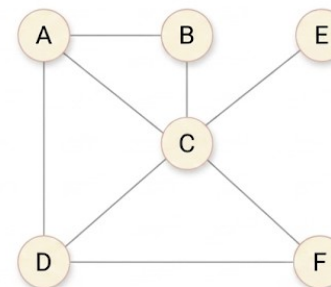
La rete biologica è definita come: $G = (V, E)$

Distinzione tra **tipologie** di grafi rilevanti

- rete **monopartita**: connette nodi della medesima natura
- rete **bipartita**: i nodi V sono divisi in due sottoinsiemi disgiunti.

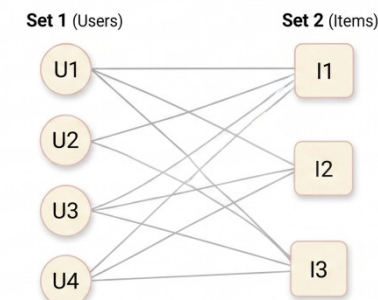
I legami avvengono solo tra set diversi

Grafo Monopartito
Unipartite Graph



Un solo tipo di nodo
Tutti i nodi appartengono
alla stessa natura.

Grafo Bipartito
Bipartite Graph



Collegamenti solo tra insiemi diversi
Le connessioni esistono esclusivamente
tra i nodi di set differenti.

TOPOLOGIA LOCALE

Calcolano la similarità tra due nodi

- **Common Neighbors (CN):** $CN(x,y)=|\Gamma(x)\cap\Gamma(y)|$
- **Adamic-Adar (AA)**
- **Resource Allocation (RA)**

BASATI SUI PERCORSI

Esplorano distanze maggiori di 2 applicando fattori di attenuazione per percorsi lunghi

- **Indice di Katz (Globale):** $S_{ij} = \sum_{l=1}^{\infty} \beta^l |paths_{ij}^l|$
- **Local Path Index (LPI)**
- **Leicth-Holme-Newman**

SPETTRALI E RANDOM WALK

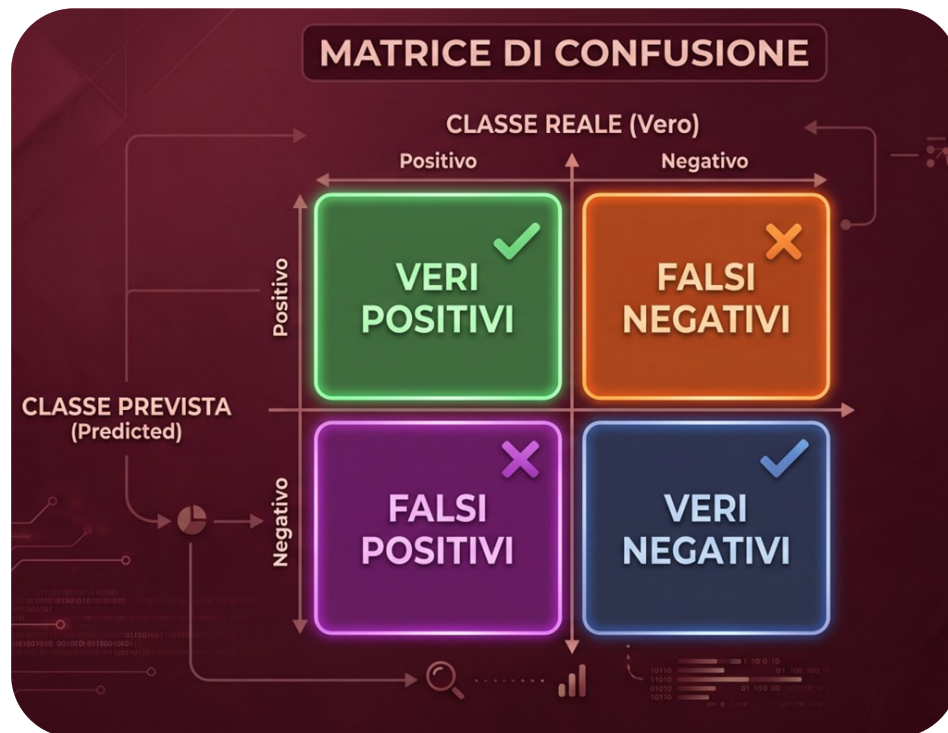
Simulano dinamiche diffusive tramite camminate aleatorie (Random Walk) per catturare la struttura globale della rete

- **Random Walk with Restart (RWR)**
- **Local Random Walk (LRW)**
- **Average Commute Time (ACT)**

NETWORK EMBEDDING

Traslano la rete in spazi vettoriali a bassa dimensionalità tramite tecniche derivate dal Deep Learning

- **DeepWalk**
- **Node2vec**
- **NetMF**
- **Prone**



FIXED-THRESHOLD METRICS

Basate sulla matrice di confusione

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN)$$

$$\text{Precision} = TP/(TP+FP)$$

$$\text{Recall} = TP/(TP+FN)$$

$$\text{F1-score} = (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

$$\text{FPR} = FP / (TN + FP)$$

TRESHOLD CURVES

Curva **ROC** (Receiver Operating Characteristic)

Curva **PR** (Precision-Recall)

L'analisi comparativa dei modelli è stata condotta su cinque reti estratte da repository pubblici, caratterizzate da elevata eterogeneità e sparsità

Dataset	Topologia	Nodi	Archi	Applicazione
MATADOR	Bipartita	3702	15 843	Interazione Farmaco-Target
DTI	Bipartita	3932	18 690	Interazione Farmaco-Target
DDI	Monopartita	1514	48 514	Reazioni Farmaco-Farmaco
DGA	Bipartita	7813	21 357	Associazione Malattia-Gene
DDA	Bipartita	7197	466 656	Drug Repurposing

Da una prima analisi risultano le seguenti evidenze

DGA, DDA

- modello ACT miglior punteggio AUROC
- modello LRW5 con 5 passi ha ottenuto il miglior punteggio AUPR
- modello LHN2 con parametro 0,95 ha ottenuto il miglior punteggio F1

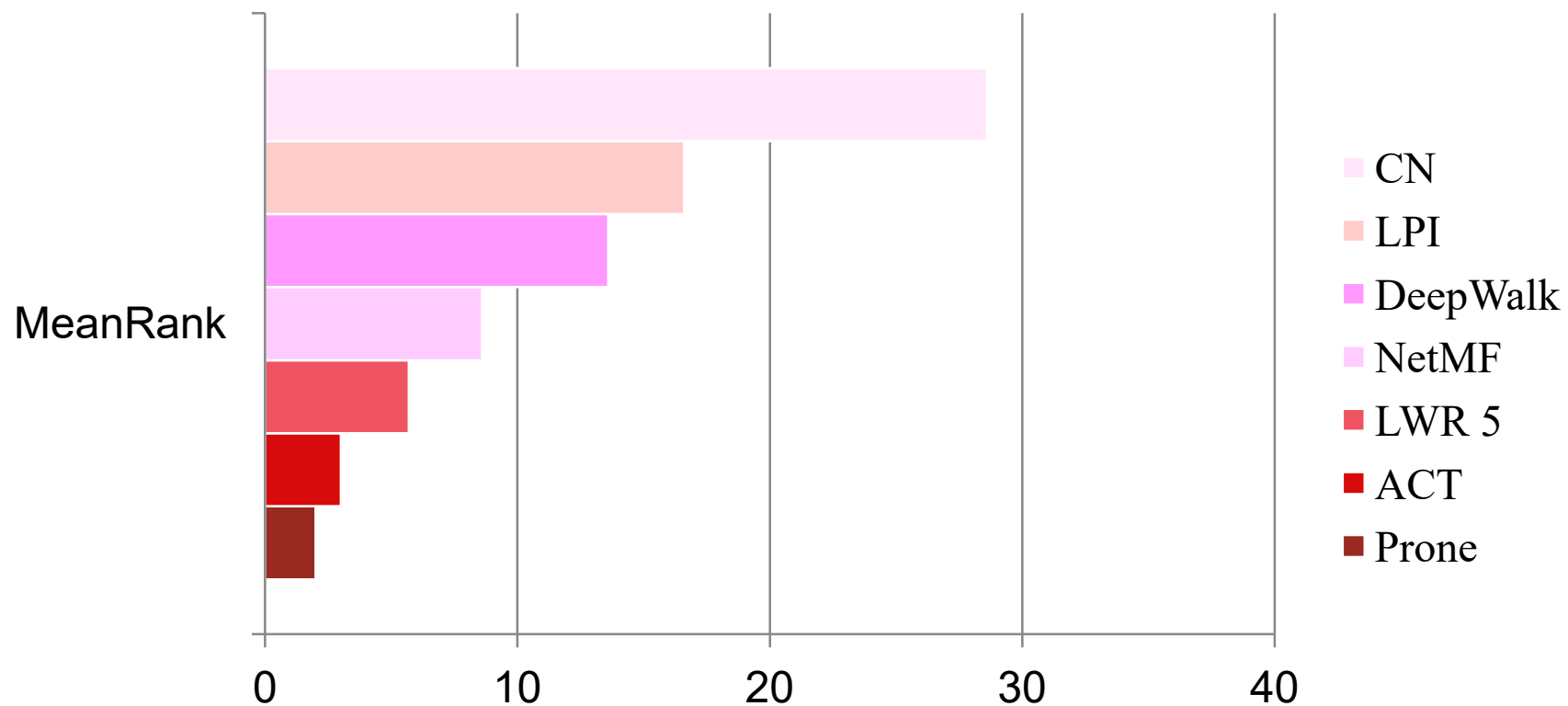
DTI, MATADOR

NetMF migliori prestazioni in tutte e tre le metriche

DDI

In quanto unica rete monopartita, il modello Prone ha ottenuto i risultati migliori in tutte e tre le metriche

Converte le performance AUROC in un Mean Rank globale: valori inferiori indicano prestazioni migliori



I risultati sperimentali dimostrano che **non esiste un modello universale**; l'efficacia dipende da due assi principali:

1. Topologia della rete

- Reti monopartite: il modello di embedding Prone domina in modo assoluto su tutte le metriche (AUROC, AUPR, F1-score)
- Reti Bipartite: eccellono i modelli NetMF e ACT, in grado di catturare le relazioni tra set disgiunti

2. Grado di Sparsità

- Nelle reti molto sparse (con alto tasso di informazioni mancanti), i metodi puramente locali degradano rapidamente.
- In questi scenari, le estensioni basate sui percorsi o sui cammini aleatori (Random Walk) mostrano una resilienza nettamente superiore

- La riformulazione della Drug Discovery tramite link prediction **accelera la ricerca** in modo inestimabile rispetto al laboratorio tradizionale.
- **Non esiste però un modello universale**: va scelto in base alla tipologia e alla sparsità della rete.
- Un forte **limite metodologico** dei modelli puramente topologici analizzati è che non possono prevedere i legami di una molecola o proteina totalmente nuova che non ha ancora nessun arco assegnato nel database di addestramento.

Grazie per l'attenzione