Utilizzo di CNN e GNN per la Predizione del Docking Score a Partire da Strutture Molecolari

Giorgia Roselli - Kevin Attarantato

Università degli Studi di Urbino Carlo Bo
Dipartimento di Scienze Pure e Applicate

Corso di Laurea Magistrale in Informatica Applicata

16 Gennaio 2025

1506 UNIVERSITÀ DEGLI STUDI DI URBINO CARLO BO

- Dettagli Architetturali della GNN

6. Addestramento dei Modelli
   - Metriche di Valutazione
   - Funzione di Perdita e Strategie di Addestramento
   - Andamento della Funzione di Loss nelle Reti CNN e GNN
   - Commento sui Grafici della Loss
   - Visualizzazione Grafica delle Metriche
   - Commento sull'Andamento delle Metriche per CNN e GNN
   - Analisi delle Metriche di Valutazione sui Dati di Test

7. Conclusioni
   - Conclusioni Generali
   - Limiti e Prospettive Future

8. EXTRA
   - Architettura della Rete GNN2
   - Dettagli Architetturali della GNN2

Introduzione allo Studio

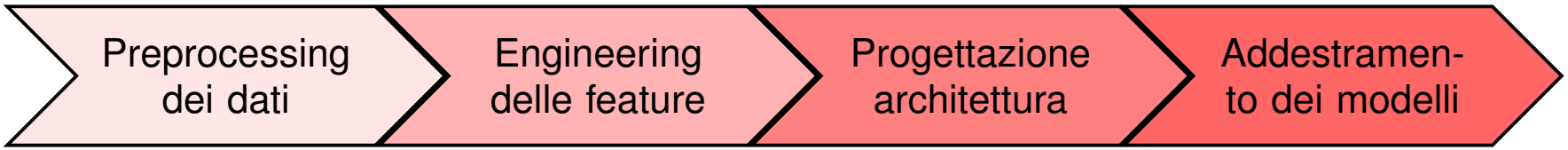
Questo progetto, svolto nell'ambito dell'esame di Deep Learning presso l'Università degli Studi di Urbino Carlo Bo, si propone di predire i *docking score* partendo dalle molecole rappresentate in formato *SMILES* (Simplified Molecular Input Line Entry System).
A tal fine, sono state implementate e confrontate due diverse architetture di reti neurali profonde: le **Convolutional Neural Networks** (CNN) e le **Graph Neural Networks** (GNN).

Il lavoro si inserisce nel contesto del *drug discovery*, uno dei processi più complessi del settore farmaceutico.

- **Docking Score:** Rappresentano la forza dell'interazione tra una molecola candidata e il target biologico. È un valore quantitativo che tiene conto di fattori come interazioni elettrostatiche, legami idrogeno e interazioni idrofobiche.
- **Smiles:** Stringa lineare che rappresenta la struttura chimica della molecola.
  Per esempio:

  CN1C(=O)NC2=CC=CC(NC(=O)C(F)C3CC(N)C3)=C21

Pipeline del Progetto



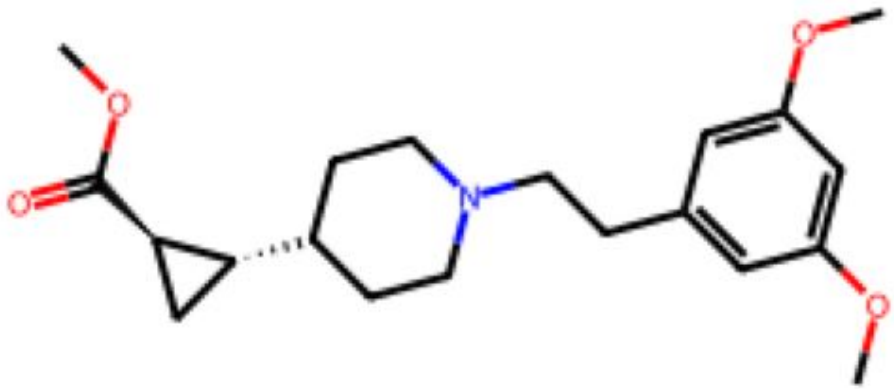
Molecular Fingerprints

Per sfruttare appieno il potenziale delle reti neurali, queste stringhe SMILES vengono trasformate in rappresentazioni piu' strutturate utilizzando la libreria **RDKit**:

- **Molecular Fingerprints:** Sequenze binarie che codificano la presenza di specifiche caratteristiche strutturali nella molecola, come sottostrutture e gruppi funzionali. Questi fingerprint permettono di sintetizzare informazioni complesse in un formato compatto, elaborabile dalle *Convolutional Neural Networks* (CNN).

Molecular Graphs

- **Molecular Graphs:** Ogni molecola è stata rappresentata come un grafo, dove i nodi corrispondono agli atomi e gli archi ai legami chimici. Questa rappresentazione è particolarmente utile per reti neurali progettate per dati strutturati, come le *Graph Neural Networks* (GNN).



Queste rappresentazioni sono state elaborate attraverso due diverse architetture di deep learning:

- **Convolutional Neural Networks (CNN)** per i fingerprints come vettori monodimensionali.
- **Graph Neural Networks (GNN)** per i grafi molecolari, sfruttando la loro struttura intrinseca per apprendere rappresentazioni che combinano informazioni locali e globali della molecola.

Questo approccio ha permesso di valutare quale rappresentazione e architettura fossero più efficaci per la predizione del docking score, fornendo indicazioni preziose per futuri sviluppi nel campo del *virtual screening*.

Validazione degli SMILES

Con l'ausilio della libreria RDKit, tutte le stringhe SMILES sono state preliminarmente validate per accertarne la corretta conversione in strutture molecolari, sia nel caso delle reti neurali CNN sia per le GNN.

Nel nostro caso nessuna entry è stata rimossa per errori di parsing o invalidità, garantendo così l'utilizzo dell'intero dataset in entrambi gli approcci.

Eliminazione dei Duplicati

Il processo di eliminazione dei duplicati è stato implementato in modo diverso per ciascun modello, tenendo conto della rappresentazione dei dati molecolari utilizzata:

CNN:

- Le fingerprint molecolari sono state convertite in stringhe binarie, consentendo un confronto diretto e semplice.
- L'analisi ha rilevato che il *19.64%* delle entry erano duplicate.

GNN:

- I grafi sono stati rappresentati come tuple immutabili contenenti nodi e archi, permettendo un confronto efficace.
- L'analisi ha rilevato che il *10.84%* delle entry erano duplicate.

In entrambi i casi, è stata mantenuta solo la prima occorrenza per ciascuna fingerprint/grafo, preservando l'unicità del dataset.

Normalizzazione e Dati

Inizialmente, i dati sono stati mescolati casualmente per evitare che il modello apprendesse schemi legati all'ordine dei campioni anziché alle loro caratteristiche.
Successivamente, è stata applicata la normalizzazione utilizzando la funzione `StandardScaler`, che trasforma i dati in modo che abbiano media pari a 0 e deviazione standard pari a 1. Questa operazione è stata valutata separatamente per le due reti:

- **CNN**: La normalizzazione non ha migliorato le prestazioni del modello, pertanto si è deciso di utilizzare i dati nella loro forma originaria.
- **GNN**: La normalizzazione ha portato a un miglioramento delle prestazioni ed è stata quindi mantenuta.

Suddivisione e Gestione del Dataset

Il dataset è stato suddiviso per garantire una corretta fase di addestramento e validazione:

- **80%** allocato al *training set*.
- Il restante **20%** suddiviso in *validation set* e *test set*.

La gestione dei dati è stata effettuata tramite `DataLoader`, che ha organizzato i campioni in batch da 64 elementi:

- Per il **training set**, i dati sono stati mescolati casualmente ad ogni epoca, per evitare che il modello si adattasse a sequenze specifiche.
- Per il **validation set** e il **test set**, l'ordine dei campioni è stato mantenuto invariato, garantendo una valutazione consistente e riproducibile.

Architettura della Rete CNN

La rete CNN utilizza strati convoluzionali e completamente connessi per estrarre rappresentazioni di alto livello dai dati molecolari.
Composta da:

- **Tre blocchi convoluzionali:** Ogni blocco comprende convoluzioni 1D, Batch Normalization, ReLU e Max-Pooling per ridurre la dimensionalità e stabilizzare l'apprendimento.
- **Strati fully connected:** Due livelli con unità dense e dropout al 50% per migliorare la generalizzazione.
- **Livello di output:** Un'unica unità lineare per la previsione del docking score.

Tecniche di regolarizzazione, come dropout e batch normalization, sono state integrate per ridurre il rischio di overfitting.

Dettagli Architetturali della CNN

Livelli Convoluzionali:

- **Primo livello:** 64 filtri, kernel di dimensione 3, seguito da Batch Normalization, ReLU e Max-Pooling.
- **Secondo livello:** 128 filtri, kernel di dimensione 3, seguito da Batch Normalization, ReLU e Max-Pooling.
- **Terzo livello:** 128 filtri, kernel di dimensione 3, seguito da Batch Normalization, ReLU e Max-Pooling.

Livelli Fully Connected:

- 256 unità con ReLU e dropout (50%).
- 64 unità con ReLU e dropout (50%).
- Livello lineare di output per la previsione.

1506
UNIVERSITÀ
DEGLI STUDI
DI URBINO
CARLO BO

Architettura della Rete GNN

La rete GNN combina strati convoluzionali specifici per grafi e livelli completamente connessi per rappresentare efficacemente le strutture molecolari.

Caratteristiche principali:

- **Integrazione di nodi e archi:** Utilizza 5 caratteristiche per nodo e 2 attributi per legame.
- **Strati convoluzionali sui grafi:** Due livelli con 256 canali di output, Batch Normalization, ReLU e Dropout (30%).
- **Pooling globale:** Aggregazione con Global Mean Pooling.
- **Regolarizzazione:** Dropout e Batch Normalization per ridurre il rischio di overfitting.

Dettagli Architetturali della GNN

Livelli convoluzionali:

- Due strati `GCNConv` con 256 canali.
- Batch Normalization, ReLU e Dropout (30%) applicati a ciascun livello.

Pooling globale:

- Global Mean Pooling per aggregare le informazioni dei nodi.

Livelli fully connected:

- Primo livello con 256 unità, ReLU, Batch Normalization e Dropout (30%).
- Livello di output con una singola unità per la previsione del docking score.

Metriche di Valutazione

Metriche utilizzate:

- *Mean Squared Error (MSE):* Media degli errori al quadrato, penalizzando maggiormente gli errori grandi.
- *Mean Absolute Error (MAE):* Media degli errori assoluti, fornendo una misura intuitiva della deviazione media.
- R^2 *(Coefficiente di Determinazione):* Proporzione di variabilità spiegata dal modello, con valori più alti indicativi di una migliore accuratezza.

Funzione di Perdita e Strategie di Addestramento

Funzione di perdita:

- La **MSE** è stata scelta per il training, grazie alla sua sensibilità agli errori grandi, utile per migliorare la precisione del modello.

Ottimizzatore:

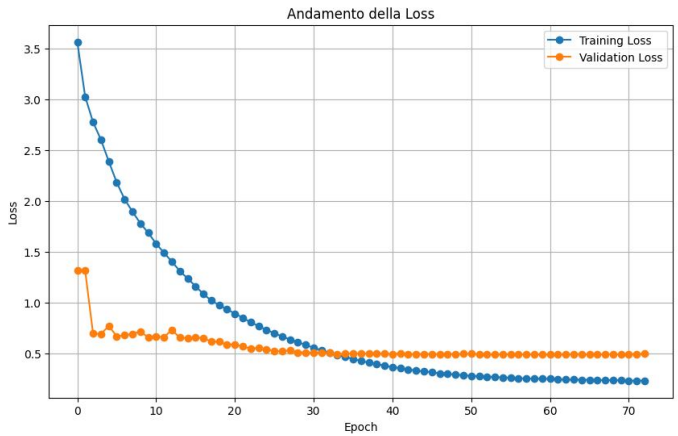
- È stato utilizzato **Adam**, un ottimizzatore robusto ed efficiente, per garantire stabilità nell'aggiornamento dei pesi.

Early Stopping:

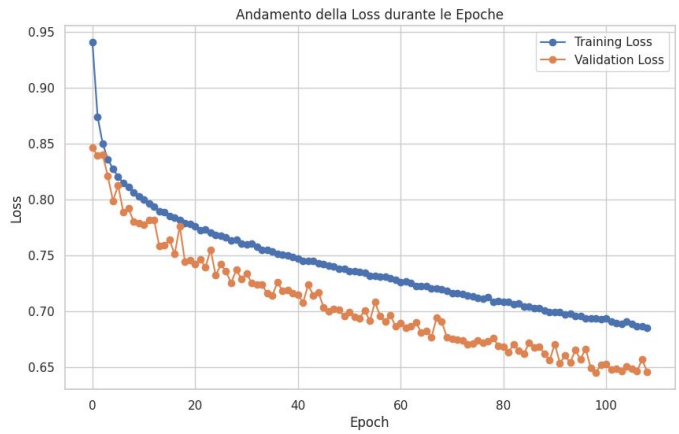
- È stata applicata una strategia di **early stopping**, interrompendo l'addestramento se la perdita sul validation set non migliorava per 10 epoche consecutive.
- Questa tecnica ha permesso di ridurre il rischio di overfitting e i tempi di training.

Confronto della Loss: CNN vs GNN

L' andamento della funzione di loss durante l'addestramento e la validazione fornisce indicazioni fondamentali sull'efficacia e sulla capacità di generalizzazione del modello.
Di seguito sono riportati i grafici che mostrano come varia la loss per le reti CNN e GNN:



Loss per la CNN.



Loss per la GNN.

Commento sui Grafici della Loss

CNN:

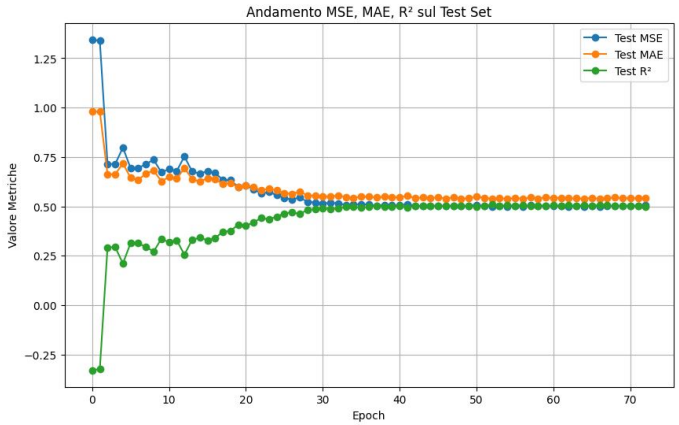
- Rapida discesa della *training loss*, segnale di un apprendimento veloce dai fingerprint molecolari.
- Stabilizzazione precoce della *validation loss*, indicando buona generalizzazione.
- La differenza tra *training loss* e *validation loss* riflette un effetto positivo di regolarizzazione.

GNN:

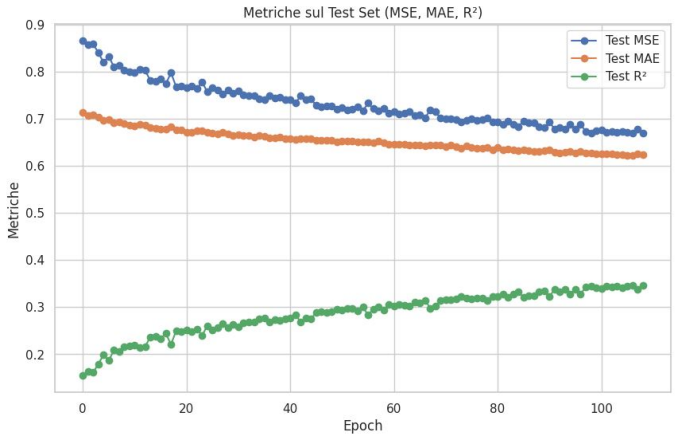
- Diminuzione più graduale della *training loss*, dovuta alla maggiore complessità dei dati grafici.
- *Validation loss* inizialmente inferiore alla *training loss*, seguita da un calo progressivo.
- Oscillazioni più marcate nella *validation loss*, legate alla minore regolarizzazione (dropout al 30%).

Visualizzazione delle Metriche per CNN e GNN

I grafici seguenti illustrano l'andamento delle metriche *MSE*, *MAE* e R^2 durante l'addestramento e la validazione, evidenziando le differenze tra le due architetture:



Metriche per la CNN.



Metriche per la GNN.

Commento sulle Metriche di CNN e GNN

CNN:

- Riduzione graduale e stabile di *MSE* e *MAE*, segnale di apprendimento affidabile.
- R^2 in continuo aumento, indicativo di una buona capacità di spiegare la varianza dei dati.

GNN:

- Oscillazioni iniziali di *MSE* e *MAE*, con stabilizzazione successiva.
- Crescita più lenta di R^2, dovuta alla maggiore complessità dei dati grafici.

La CNN garantisce una convergenza più rapida e accurata, mentre la GNN evidenzia potenziale in applicazioni che sfruttano relazioni topologiche complesse.

Analisi delle Metriche di Valutazione sui Dati di Test

I valori finali delle metriche di valutazione sui dati di test per entrambe le reti sono riportati nella seguente tabella:

Metrica	**CNN**	**GNN**
Validation Loss	0.4872	0.6269
Test Loss	0.5003	0.6503
Test MSE	0.5003	0.6503
Test MAE	0.5406	0.6169
R^2	0.5053	0.3649

Confronto delle metriche tra CNN e GNN.

Conclusioni Generali

Lo studio ha esplorato l'uso di tecniche di deep learning per predire i docking score di molecole rappresentate tramite SMILES, sviluppando due modelli distinti:

- **CNN:** Elabora i fingerprint molecolari e si distingue per la precisione predittiva.
- **GNN:** Modella relazioni topologiche complesse sfruttando grafi molecolari, offrendo potenziale per scenari in cui la struttura delle molecole è centrale.

Risultati principali:

- La CNN ha mostrato migliori performance complessive.
- La GNN ha evidenziato un potenziale maggiore nell'analisi di relazioni strutturali.

Lo studio sottolinea l'importanza di adattare preprocessing e architettura al tipo di dati per massimizzare la capacità predittiva.

1506 UNIVERSITÀ DEGLI STUDI DI URBINO CARLO BO

Limiti e Prospettive Future

Limiti dello studio:

- Risorse computazionali limitate (es. Google Colab) hanno influito sull'ottimizzazione dei modelli.
- Dataset non sufficientemente ampio e diversificato, soprattutto per la GNN.

Sviluppi futuri:

- Integrazione di approcci ibridi tra CNN e GNN per combinare pattern locali e relazioni strutturali.
- Uso di dataset arricchiti con descrittori chimico-fisici e misurazioni sperimentali per migliorare la generalizzazione.
- Miglioramento dell'interpretabilità dei modelli per comprendere meglio le caratteristiche molecolari che influenzano i docking score.

La scelta tra CNN e GNN dipende dai requisiti applicativi: precisione o analisi strutturale.

Architettura della Rete GNN2

La rete GNN utilizza convoluzioni specifiche per grafi e trasformazioni avanzate per elaborare informazioni sia dai nodi che dai legami in modo efficace.

- **Trasformazione degli attributi dei legami:** Una rete `MLP` elabora gli attributi dei legami per generare rappresentazioni significative.
- **Strati convoluzionali sui grafi:** Due strati `NNConv` con 256 canali di output, ciascuno seguito da Batch Normalization, ReLU e Dropout (30%).
- **Pooling globale:** Aggregazione con Global Mean Pooling per combinare le informazioni dei nodi in un'unica rappresentazione.
- **Regolarizzazione:** Uso di Dropout e Batch Normalization per prevenire overfitting e migliorare la generalizzazione.

Dettagli Architetturali della GNN2

Trasformazione degli attributi dei legami:

- `MLP` iniziale: Trasforma gli attributi dei legami con due strati lineari e una funzione di attivazione ReLU.
- Output della `MLP`: Dimensione compatibile con il prodotto tra le feature dei nodi e i canali nascosti.

Livelli convoluzionali:

- Primo livello `NNConv`: 256 canali, BatchNorm, ReLU e Dropout.
- Secondo livello `NNConv`: 256 canali, BatchNorm, ReLU e Dropout.

Pooling globale:

- Global Mean Pooling per aggregare le informazioni a livello di grafo.

Livelli fully connected:

- Primo livello con 256 unità, ReLU, Batch Normalization e Dropout.
- Livello finale con una singola unità per la previsione del docking score.

Risultati GNN2

I risultati presentati per la nuova rete GNN2 sono stati ottenuti al termine dell'epoca 83. Purtroppo, a causa delle limitazioni di Colab, l'addestramento non ha potuto essere completato. Tuttavia, è ragionevole supporre che con risorse computazionali più avanzate, la rete avrebbe continuato a migliorare.

Risultati ottenuti:

Metrica	**GNN2**	**GNN**
Validation Loss	0.5338	0.6269
Test Loss	0.5559	0.6503
Test MSE	0.5559	0.6503
Test MAE	0.5641	0.6169
R^2	0.4570	0.3649