Analisi degli Embedding dei Codici ATC

Sommario

Questo report presenta un'analisi degli embedding dei codici ATC (Anatomical Therapeutic Chemical), una rappresentazione vettoriale di 88 dimensioni per 6440 codici farmacologici. L'obiettivo è valutare la qualità di questi embedding attraverso diverse metriche quantitative, verificando se la rappresentazione vettoriale cattura efficacemente la struttura gerarchica del sistema di classificazione ATC. Viene inoltre presentata un'analisi comparativa degli embedding ottenuti con diverse dimensioni (44, 66, 88, 100, 120 dimensioni) per identificare la configurazione ottimale.

Indice

1 Introduzione

1.1 Contesto

Il sistema di classificazione ATC (Anatomical Therapeutic Chemical) è uno standard internazionale per classificare i farmaci secondo criteri anatomici, terapeutici e chimici. Ogni codice ATC è organizzato in modo gerarchico su più livelli:

- **ATC1** (1 carattere): macro-categoria anatomica (es. "A" = apparato gastrointestinale)
- **ATC2** (3 caratteri): sottogruppo terapeutico
- **ATC3** (4 caratteri): sottogruppo farmacologico
- **ATC4** (5 caratteri): sottogruppo chimico
- **ATC5** (7 caratteri): sostanza chimica specifica

Ad esempio, il codice `A10BA02` viene scomposto come:

- **A**: Apparato gastrointestinale e metabolismo
- **A10**: Farmaci usati nel diabete
- **A10B**: Farmaci per la riduzione della glicemia, escluse le insuline
- **A10BA**: Biguanidi
- **A10BA02**: Metformina

1.2 Obiettivo dell'analisi

Gli embedding sono rappresentazioni vettoriali dense che mappano ciascun codice ATC in uno spazio di 88 dimensioni. L'obiettivo principale del lavoro è apprendere una rappresentazione numerica continua che rispetti la struttura gerarchica della classificazione farmaceutica, utilizzando:

1. Un **embedding gerarchico** che tratta esplicitamente i diversi livelli del codice ATC
2. Un **classificatore multilivello** per predire ATC1, ATC2 e ATC3
3. Una **triplet loss gerarchica** per imporre vincoli geometrici sullo spazio vettoriale

L'obiettivo di questa analisi è verificare se questi vettori riflettono correttamente la similarità semantica e la struttura gerarchica dei codici farmacologici attraverso diverse metriche quantitative.

2 Metodologia

2.1 Pre-processing dei codici ATC

Il dataset viene caricato da un file CSV contenente i codici ATC in forma testuale. Il processo di pre-processing include:

1. **Normalizzazione**: applicazione di `strip()` e `upper()` per uniformare i codici;
2. **Estrazione di codici unici**: identificazione di 6440 codici ATC distinti;
3. **Costruzione di vocabolari**: creazione di vocabolari separati per ogni livello gerarchico.

La funzione `atc_to_level_tokens` scompone ogni codice in cinque prefissi corrispondenti ai livelli canonici con lunghezze $[1, 3, 4, 5, 7]$ caratteri. Ad esempio:

$$\text{“A10BA02”} \longrightarrow [\text{A}, \text{A10}, \text{A10B}, \text{A10BA}, \text{A10BA02}]$$

Se un codice è troppo corto per un certo livello, viene utilizzato il token speciale `<PAD>`.

La funzione `build_level_vocabs` costruisce un vocabolario (token $\rightarrow$ indice intero) per ogni livello, includendo i token speciali `<PAD>` e `<UNK>`. Le dimensioni dei vocabolari risultanti sono:

Tabella 1: Dimensioni dei vocabolari per livello gerarchico

Livello	Lunghezza prefisso	Numero token
ATC1	1 carattere	16
ATC2	3 caratteri	96
ATC3	4 caratteri	271
ATC4	5 caratteri	911
ATC5	7 caratteri	5156

Infine, la funzione `encode_atc_to_level_ids` converte ogni codice ATC in una lista di 5 interi (uno per livello), che costituirà l'input del modello di embedding.

2.2 Modello di embedding gerarchico

La classe `HierarchicalATCEmbedding` implementa un layer di embedding che rispetta la struttura gerarchica del codice ATC. I parametri principali del modello sono:

- **level_vocabs**: lista dei vocabolari costruiti per ciascun livello gerarchico
- **level_dims**: dimensioni degli embedding per ogni livello, definite come $[8, 16, 16, 16, 32]$
- **combine_mode**: modalità di combinazione impostata a `"concat"` (concatenazione)

Per ciascun livello gerarchico viene creato un layer `nn.Embedding` indipendente, permettendo al modello di apprendere rappresentazioni specifiche per ogni livello della tassonomia ATC.

2.2.1 Processo di forward propagation

Dato un tensore di input contenente gli ID numerici di shape $(\text{batch}, 5)$, dove ogni riga rappresenta un codice ATC scomposto nei suoi 5 livelli il modello esegue le seguenti operazioni:

1. **Estrazione degli embedding**: per ogni livello, il corrispondente ID viene trasformato nel suo vettore embedding attraverso la matrice di embedding specifica del livello
2. **Applicazione dei pesi**: vengono applicati eventuali pesi di normalizzazione o importanza ai diversi livelli
3. **Concatenazione**: tutti i vettori vengono concatenati in sequenza, producendo un unico vettore di dimensione:
$$8 + 16 + 16 + 16 + 32 = 88$$

Il risultato finale è una rappresentazione continua in $\mathbb{R}^{88}$ per ciascun codice ATC. Questa struttura consente alle diverse componenti del vettore di catturare informazioni specifiche dei vari livelli gerarchici: i primi 8 valori codificano il livello ATC1, i successivi 16 il livello ATC2, e così via, preservando esplicitamente la struttura tassonomica del sistema di classificazione.

2.3 Costruzione delle etichette gerarchiche

Per l'addestramento del classificatore multilivello, vengono estratti tre livelli di etichette da ogni codice:

- **ATC1**: primo carattere
- **ATC2**: primi 3 caratteri
- **ATC3**: primi 4 caratteri

Le stringhe vuote (quando il livello non esiste) vengono mappate a -1 e ignorate tramite l'opzione `ignore_index=-1` nella funzione di loss.

I vocabolari di etichette `(build_label_vocab)` costruiti hanno le seguenti cardinalità:

$$\text{Classi ATC1} = 14$$
$$\text{Classi ATC2} = 94$$
$$\text{Classi ATC3} = 269$$

2.4 Caricamento dei dati

Il processo inizia con il caricamento dei codici ATC da un file CSV. I codici vengono normalizzati e scomposti nei 5 livelli gerarchici (da ATC1 a ATC5), ottenendo 6440 sequenze di ID numerici.

Durante il training, il modello `HierarchicalATCEmbedding` apprende gli embedding di 88 dimensioni concatenando le rappresentazioni dei 5 livelli.

2.5 Rete neurale per la classificazione

La classe `ATCHierarchicalClassifier` implementa il modello completo per la classificazione multilivello dei codici ATC. Questo modello integra l'embedding gerarchico con una rete neurale che elabora le rappresentazioni attraverso strati successivi, producendo predizioni simultanee a tre livelli della gerarchia farmaceutica.

2.5.1 Architettura del modello

Il classificatore è strutturato in tre componenti principali che operano in sequenza:

1. Layer di embedding condiviso Utilizza `HierarchicalATCEmbedding` per convertire ogni codice ATC in un vettore denso di 88 dimensioni. Questa rappresentazione cattura informazioni da tutti e cinque i livelli della gerarchia ATC, con componenti dedicate a ciascun livello (8, 16, 16, 16 e 32 dimensioni rispettivamente).

2. Rete neurale di elaborazione (MLP condiviso) Una rete Multi-Layer Perceptron elabora il vettore di embedding per estrarre caratteristiche di livello superiore:

- **Layer lineare**: trasforma il vettore da 88 a 128 dimensioni, combinando le informazioni dei diversi livelli gerarchici
- **Funzione di attivazione ReLU**: introduce non-linearità nel modello, permettendo di apprendere relazioni complesse
- **Dropout (20%)**: tecnica di regolarizzazione che disattiva casualmente il 20% dei neuroni durante il training per prevenire overfitting

Questa componente è *condivisa* tra tutti i livelli di classificazione, garantendo che le rappresentazioni utilizzate per le diverse predizioni siano coerenti tra loro.

3. Tre teste di classificazione indipendenti Ciascuna testa è specializzata nella predizione di un livello specifico della gerarchia:

- **head_atc1**: 128 $\rightarrow$ 14 classi, predice la macro-categoria anatomica (Es. "A" = apparato gastrointestinale)
- **head_atc2**: 128 $\rightarrow$ 94 classi, predice il sottogruppo terapeutico (Es. "A10" = farmaci per diabete)
- **head_atc3**: 128 $\rightarrow$ 269 classi, predice il sottogruppo farmacologico (Es. "A10B" = ipoglicemizzanti)

2.5.2 Flusso di elaborazione

Il processo di forward propagation può essere rappresentato come una sequenza di trasformazioni:

$$\text{ID numerici} \xrightarrow{\text{embedding}} z \in \mathbb{R}^{88} \xrightarrow{\text{MLP}} h \in \mathbb{R}^{128} \xrightarrow{\text{teste}} \begin{cases} \text{logits}_{\text{ATC1}} \in \mathbb{R}^{14} \\ \text{logits}_{\text{ATC2}} \in \mathbb{R}^{94} \\ \text{logits}_{\text{ATC3}} \in \mathbb{R}^{269} \end{cases}$$

dove:

- z è la rappresentazione embedding del codice ATC
- h è la rappresentazione elaborata dalla rete neurale
- i *logits* sono i punteggi non normalizzati per ciascuna classe.

2.5.3 Funzione di loss per la classificazione

La funzione di loss complessiva per la classificazione utilizza la Cross-Entropy su tre livelli con pesi differenziati:

$$\mathscr{L}_{\text{class}} = 0.3\,\mathscr{L}_{\text{ATC1}} + 0.3\,\mathscr{L}_{\text{ATC2}} + 0.4\,\mathscr{L}_{\text{ATC3}}$$

dove $\mathscr{L}_{\text{ATC1}}$, $\mathscr{L}_{\text{ATC2}}$ e $\mathscr{L}_{\text{ATC3}}$ sono le Cross-Entropy sui rispettivi livelli. La maggiore enfasi su ATC3 (peso 0.4) riflette l'importanza di predire correttamente il livello più dettagliato della gerarchia.

2.5.4 Triplet loss gerarchica

Oltre alla loss di classificazione, il modello utilizza una **triplet loss** per imporre vincoli geometrici sullo spazio degli embedding. L'obiettivo è garantire che codici farmacologicamente simili abbiano rappresentazioni vettoriali vicine, mentre codici diversi siano ben separati.

Definizione di similarità Due codici ATC sono considerati *simili* se condividono lo stesso prefisso ATC3 (primi 4 caratteri). Ad esempio, A10BA02 (Metformina) e A10BA03 (Buformina) sono simili perché entrambi appartengono al sottogruppo farmacologico A10B (Biguanidi).

Costruzione delle triplette Per ogni codice del dataset, chiamato *anchor*, vengono selezionati due codici di confronto:

- **Positivo**: un codice con lo stesso prefisso ATC3 dell'anchor (farmacologicamente simile)
- **Negativo**: un codice con prefisso ATC3 diverso (farmacologicamente diverso)

L'implementazione utilizza una strategia semplificata che seleziona il primo positivo e il primo negativo trovati durante l'iterazione sul dataset.

Ad esempio, per l'anchor A10BA02:

- Positivo: A10BA03 (condivide A10B)
- Negativo: N02BE01 (prefisso N02B diverso)

Calcolo della loss Per ogni tripletta (a, p, n) vengono calcolate le distanze euclidee nello spazio degli embedding:

$$d_{ap} = \|e_a - e_p\|_2 \quad \text{(distanza anchor-positivo)}$$
$$d_{an} = \|e_a - e_n\|_2 \quad \text{(distanza anchor-negativo)}$$

dove e_a, e_p ed e_n sono i vettori di embedding di anchor, positivo e negativo rispettivamente.

La triplet loss è definita come:

$$\mathscr{L}_{\text{triplet}} = \max\big(0, d_{ap} - d_{an} + m\big)$$

dove $m = 2.5$ è il *margine*.

Questa funzione di loss penalizza le configurazioni indesiderate in cui l'embedding positivo (simile) non è sufficientemente più vicino all'anchor rispetto all'embedding negativo (diverso).

In dettaglio:

- Se $d_{ap} < d_{an} - m$: il positivo è già sufficientemente più vicino del negativo (oltre il margine) implica loss = 0 (nessuna penalità)
- Se $d_{ap} \geq d_{an} - m$: il positivo non è abbastanza più vicino implica loss > 0 (penalità proporzionale alla violazione)

Il margine $m = 2.5$ impone che la distanza tra anchor e positivo sia inferiore di almeno 2.5 unità rispetto alla distanza tra anchor e negativo, creando una "zona di sicurezza" che separa chiaramente codici simili da codici diversi nello spazio vettoriale.

Effetto sul training Durante l'ottimizzazione, questa loss guida il modello a:

1. Avvicinare gli embedding di codici con lo stesso prefisso ATC3
2. Allontanare gli embedding di codici con prefissi ATC3 diversi
3. Creare uno spazio in cui la distanza euclidea riflette la similarità farmacologica

Questa componente si integra con la loss di classificazione, con ruoli complementari:

- **Cross-Entropy**: garantisce che il modello assegni le etichette corrette (accuratezza predittiva)
- **Triplet loss**: organizza lo spazio degli embedding in modo che farmaci simili siano rappresentati da vettori vicini e farmaci diversi da vettori lontani (coerenza geometrica)

Questa componente complementa la loss di classificazione: mentre la Cross-Entropy ottimizza l'accuratezza predittiva, la triplet loss ottimizza la geometria dello spazio degli embedding, garantendo che la struttura gerarchica della classificazione ATC si rifletta nelle distanze vettoriali.

2.5.5 Loss totale

La funzione di loss complessiva combina i due obiettivi:

$$\mathscr{L}_{\text{totale}} = \mathscr{L}_{\text{class}} + \lambda\, \mathscr{L}_{\text{triplet}}, \qquad \lambda = 6.0$$

Il termine λ controlla il peso relativo della triplet loss rispetto alla loss di classificazione. Nel nostro esperimento è stato fissato a $\lambda = 6.0$, in modo da rendere la componente metrico-geometrica sufficientemente influente sull'organizzazione dello spazio degli embedding, senza compromettere la stabilità dell'addestramento supervisionato.

Questo approccio consente di:

- Ottimizzare la capacità predittiva del classificatore
- Imporre vincoli geometrici che riflettano la similarità semantica dei codici

2.6 Procedura di training

Il modello viene addestrato sull'intero dataset di codici ATC utilizzando le seguenti configurazioni:

Tabella 2: Configurazione del training

Parametro	**Valore**
Dataset	6440 codici ATC unici
Numero di epoche	40
Ottimizzatore	Adam
Learning rate	0.001
Batch size	Full-batch (intero dataset)

2.6.1 Evoluzione della loss

Durante il training, vengono monitorate due componenti della loss per valutare l'apprendimento del modello:

Tabella 3: Evoluzione delle loss durante il training

Epoca	**Loss totale**	**Loss triplet**
1	2.37	0.0302
10	1.87	0.0263
20	1.38	0.0225
30	0.98	0.0193
40	0.70	0.0166

2.6.2 Analisi dell'apprendimento

L'evoluzione delle loss durante il training rivela diversi aspetti importanti del processo di apprendimento:

Convergenza regolare La loss totale diminuisce in modo costante e monotono da 2.37 a 0.70 nel corso delle 40 epoche, con una riduzione complessiva del 70%. Questa progressione regolare indica un apprendimento stabile senza oscillazioni o instabilità, suggerendo che il learning rate scelto (0.001) è appropriato e che non si verifica overfitting sul dataset.

Composizione della loss La loss totale è composta da due componenti con ruoli distinti ma complementari:

- **Loss di classificazione**: rappresenta la componente dominante dell'ottimizzazione e guida il modello verso un'elevata accuratezza predittiva nella classificazione dei codici ATC ai diversi livelli gerarchici
- **Loss triplet**: contribuisce in misura quantitativamente inferiore ma non trascurabile alla loss totale. Grazie al fattore di peso $\lambda = 6.0$, questa componente esercita un'influenza concreta sull'organizzazione geometrica dello spazio degli embedding, favorendo la separazione tra codici farmacologicamente diversi e la vicinanza tra codici gerarchicamente simili.

Questo rapporto indica che l'ottimizzazione è principalmente guidata dall'obiettivo di classificazione, con la triplet loss che fornisce un vincolo secondario sulla struttura dello spazio vettoriale.

Riduzione coordinata Entrambe le componenti diminuiscono simultaneamente durante il training:

- **Loss totale**: $2.37 \rightarrow 0.7$ (riduzione del 70%)
- **Loss triplet**: $0.0302 \rightarrow 0.0166$ (riduzione del 45%)

Ruolo della triplet loss Nonostante il contributo numerico ridotto, la triplet loss svolge un ruolo importante nell'apprendimento:

- Impone vincoli geometrici che preservano le relazioni gerarchiche tra i codici
- Non domina l'ottimizzazione, evitando di compromettere l'accuratezza predittiva
- La sua riduzione progressiva indica che lo spazio diventa sempre più coerente semanticamente

L'efficacia di questa componente, pur numericamente piccola, sarà confermata dai risultati successivi che mostrano eccellente coerenza K-NN (98.4%) e un gradiente di distanze perfettamente ordinato secondo la gerarchia ATC.

2.6.3 Output del training

Al termine delle 40 epoche, il modello produce una matrice di embedding di dimensioni $(6440, 88)$, dove ogni riga rappresenta un codice ATC come vettore in $\mathbb{R}^{88}$. Questi embedding integrano informazioni dai cinque livelli gerarchici e sono ottimizzati simultaneamente per:

1. **Accuratezza predittiva**: attraverso la loss di classificazione multilivello
2. **Coerenza geometrica**: attraverso la triplet loss gerarchica

Questi vettori costituiscono la base per tutte le analisi successive di qualità e struttura semantica dello spazio degli embedding.

2.7 Metriche di valutazione

Per valutare la qualità degli embedding appresi, viene utilizzato un insieme di metriche che esaminano diversi aspetti della rappresentazione: capacità predittiva, struttura a cluster, coerenza semantica locale e preservazione della gerarchia. Ogni metrica fornisce informazioni complementari sulla qualità complessiva del modello.

2.7.1 Metriche di classificazione

Accuracy L'accuracy misura la percentuale di codici ATC per cui il modello predice correttamente la categoria di appartenenza. Vengono valutati tre livelli della gerarchia:

- **ATC1** (1 carattere): predizione della macro-categoria anatomica

  Esempio: dato il codice `A10BA02`, il modello deve predire `A` (apparato gastrointestinale e metabolismo) tra 14 possibili classi.

- **ATC2** (3 caratteri): predizione del sottogruppo terapeutico

  Esempio: il modello deve predire `A10` (farmaci usati nel diabete) tra 94 possibili classi.

- **ATC3** (4 caratteri): predizione del sottogruppo farmacologico

  Esempio: il modello deve predire `A10B` (farmaci ipoglicemizzanti, escluse le insuline) tra 269 possibili classi.

L'accuracy viene calcolata come:

$$\text{Accuracy} = \frac{\text{Numero di predizioni corrette}}{\text{Numero totale di codici}}$$

Accuratezze elevate indicano che gli embedding contengono informazioni discriminative sufficienti per ricostruire i livelli della gerarchia ATC.

Balanced Accuracy La balanced accuracy è la media delle recall per ciascuna classe, fornendo una misura più robusta rispetto alla semplice accuracy quando le classi sono sbilanciate:

$$\text{Balanced Accuracy} = \frac{1}{N_{\text{classi}}} \sum_{i=1}^{N_{\text{classi}}} \text{Recall}_i$$

Questa metrica è importante per il dataset ATC, dove alcune categorie possono essere molto più popolate di altre.

F1-Score L'F1-score è la media armonica tra precision e recall, fornendo un bilancio tra questi due aspetti:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Vengono calcolate due varianti:

- **F1 Macro**: media non pesata degli F1-score di ogni classe (tratta tutte le classi con uguale importanza)
- **F1 Weighted**: media degli F1-score pesata per la numerosità di ciascuna classe

Precision e Recall

- **Precision**: frazione di predizioni corrette tra tutte quelle assegnate a una classe

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- **Recall**: frazione di esempi di una classe correttamente identificati

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Come per l'F1-score, vengono calcolate le versioni macro (non pesata) e weighted (pesata).

2.7.2 Metriche di clustering

Queste metriche confrontano la struttura a cluster osservata nello spazio degli embedding con la classificazione ufficiale ATC1 dei farmaci. Le etichette ATC1 note rappresentano la "verità di riferimento" (ground truth) rispetto alla quale viene valutata la qualità del clustering.

Silhouette Score Il Silhouette Score misura quanto ciascun punto è ben assegnato al proprio cluster rispetto agli altri cluster. Per ogni embedding i, viene calcolato:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

dove:

- $a(i)$: distanza media tra i e tutti gli altri punti dello stesso cluster (coesione intra-cluster)
- $b(i)$: distanza media tra i e tutti i punti del cluster più vicino diverso dal proprio (separazione inter-cluster)

In base al risultato ottenuto, si ha:

- $s(i) \approx 1$: il punto è ben assegnato al proprio cluster (molto vicino ai membri del cluster, lontano dagli altri)

- $s(i) \approx 0$: il punto è sul confine tra due cluster
- $s(i) < 0$: il punto potrebbe essere assegnato al cluster sbagliato

Il valore finale è la media su tutti i punti. Valori vicini a 1 indicano cluster ben definiti e separati.

Davies-Bouldin Index Questo indice misura il rapporto tra la dispersione interna dei cluster (quanto i punti sono sparsi al loro interno) e la separazione tra cluster diversi (quanto i centri dei cluster sono distanti).

- Valore basso (< 1): cluster compatti e ben separati
- Valore alto (> 3): cluster dispersi o sovrapposti

Calinski-Harabasz Index Questo indice valuta la qualità del clustering confrontando due tipi di variabilità:

- **Separazione tra cluster** (varianza between): quanto i diversi cluster sono distanti tra loro nello spazio. Si misura calcolando quanto i centri dei cluster sono lontani dal centro globale di tutti i dati.
- **Compattezza interna** (varianza within): quanto i punti all'interno di ciascun cluster sono vicini al proprio centro.

Un clustering di qualità presenta cluster ben separati (alta varianza between) e internamente compatti (bassa varianza within), risultando in un indice alto.

- Valori alti (> 100): clustering buono (cluster distinti e compatti)
- Valori bassi (< 30): clustering scarso (cluster sovrapposti o dispersi)

2.7.3 K-NN Semantic Consistency

Questa metrica valuta la **coerenza semantica locale** dello spazio degli embedding, verificando se codici vicini nello spazio vettoriale appartengono effettivamente alla stessa categoria farmacologica.

Per ogni codice ATC:

1. Si identificano i $k = 10$ vicini più prossimi nello spazio degli embedding (utilizzando la distanza euclidea)
2. Si calcola la frazione di questi vicini che appartengono alla stessa macro-categoria ATC1 dell'anchor
3. Si media questa frazione su tutti i codici del dataset

Esempio Consideriamo il codice A10BA02 (Metformina, classe A):

- I suoi 10 vicini più prossimi sono: A10BA03, A10BA04, A10BB01, ..., A10BG02
- Se 9 su 10 appartengono alla classe A $\rightarrow$ consistency locale = 90%
- Questo processo viene ripetuto per tutti i 6440 codici
- La K-NN consistency finale è la media di tutte le consistency locali

In base al valore ottenuto si può dire che:

- Valore vicino a 100%: eccellente coerenza semantica locale (codici farmacologicamente simili sono rappresentati da vettori vicini)
- Valore basso ($< 50\%$): lo spazio è disorganizzato (la distanza vettoriale non riflette la similarità farmacologica)

Questa metrica è particolarmente significativa perché valuta la struttura locale dello spazio, complementando le metriche globali di clustering.

2.7.4 Analisi delle distanze intra/inter categoria

Questa analisi confronta la distribuzione delle distanze euclidee tra embedding appartenenti alla stessa macro-categoria rispetto a quelli di categorie diverse.

1. Si campionano casualmente fino a 2000 codici per efficienza computazionale
2. Per ogni coppia di codici (i, j), si calcola la distanza euclidea $d_{ij} = \|e_i - e_j\|_2$
3. Le coppie vengono divise in due gruppi:
   - **Intra-classe**: coppie con la stessa macro-categoria ATC1
   - **Inter-classe**: coppie con macro-categorie ATC1 diverse
4. Si calcola la distanza media per ciascun gruppo

Un embedding di qualità dovrebbe soddisfare:

$$\text{Distanza media intra-classe} < \text{Distanza media inter-classe}$$

Questa proprietà indica che lo spazio rispetta la struttura categoriale: farmaci della stessa categoria sono rappresentati da vettori più vicini rispetto a farmaci di categorie diverse.

2.7.5 Test gerarchico delle distanze

Questo test verifica se lo spazio degli embedding preserva l'intera struttura gerarchica della classificazione ATC, non solo il primo livello.

Si definiscono quattro gruppi di coppie di codici in base al livello gerarchico condiviso:

1. **Gruppo 1 – Stessa ATC3** (4 caratteri identici)

   Esempio: A10BA02 e A10BA03 (entrambi biguanidi)

   Massima similarità: appartengono allo stesso sottogruppo farmacologico

2. **Gruppo 2 – Stessa ATC2, ATC3 diversa** (3 caratteri identici, il 4° diverso)

   Esempio: A10BA02 (biguanidi) e A10BB01 (sulfaniluree)

   Similarità intermedia: stesso gruppo terapeutico (antidiabetici) ma meccanismi diversi

3. **Gruppo 3 – Stessa ATC1, ATC2 diversa** (1 carattere identico, i successivi diversi)

   Esempio: A10BA02 (antidiabetici) e A02BC01 (inibitori di pompa protonica)

   Bassa similarità: stessa macro-categoria anatomica ma funzioni terapeutiche diverse

4. **Gruppo 4 – ATC1 diversa** (primo carattere diverso)

   Esempio: A10BA02 (apparato gastrointestinale) e N02BE01 (sistema nervoso)

   Minima similarità: categorie anatomiche completamente diverse

Per ogni gruppo, si calcola la distanza euclidea media tra tutti i membri.

Un embedding che preserva perfettamente la gerarchia dovrebbe mostrare un **gradiente monotono crescente** delle distanze:

$$d_{\text{ATC3}} < d_{\text{ATC2}} < d_{\text{ATC1}} < d_{\text{diversi}}$$

Questo gradiente ordinato dimostra che:

- La distanza vettoriale riflette la similarità semantica a tutti i livelli della gerarchia
- Codici che condividono più livelli gerarchici sono sistematicamente più vicini
- Lo spazio degli embedding ha appreso la struttura tassonomica completa, non solo le distinzioni di primo livello

Questo test è particolarmente rigoroso perché richiede che l'ordinamento delle distanze sia coerente con l'intera tassonomia farmaceutica.

3 Visualizzazione dello spazio degli embedding

3.1 Proiezione t-SNE

Per ispezionare visivamente la struttura dello spazio degli embedding viene applicata la tecnica t-SNE (t-distributed Stochastic Neighbor Embedding) per ridurre le 88 dimensioni a 2 dimensioni visualizzabili. Ogni punto nel grafico rappresenta un codice ATC, colorato in base alla sua macro-categoria ATC1.

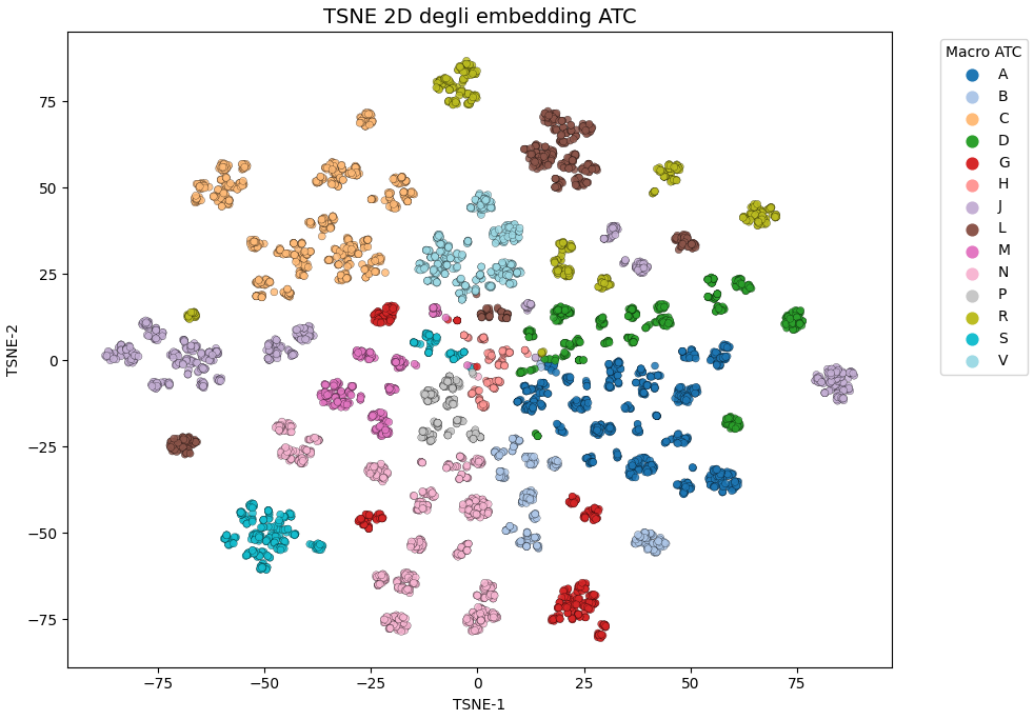

Figura 1: Proiezione t-SNE 2D degli embedding ATC.

Dal grafico t-SNE emergono diverse caratteristiche rilevanti riguardo alla struttura degli embedding appresi:

- **Formazione di cluster**: le macro–categorie ATC1 tendono a organizzarsi in gruppi visivamente distinguibili. Ogni colore rappresenta una classe ATC1 e molti punti appartenenti alla stessa categoria risultano concentrati in regioni coerenti dello spazio bidimensionale
- **Separazione globale debole fra categorie**: pur essendo presenti cluster localmente riconoscibili, questi non risultano completamente isolati a livello globale. Alcune macro–categorie mostrano aree di contatto o parziale sovrapposizione, indicando che gli embedding non impongono una separazione netta fra tutte le classi, ma riflettono relazioni farmacologiche più sfumate
- **Compattezza variabile**: alcune categorie producono cluster molto compatti e ben definiti, segnalando un'elevata similarità interna dei codici corrispondenti. Altre appaiono invece più disperse, a testimonianza di una maggiore eterogeneità semantica all'interno della stessa macro–categoria
- **Struttura semantica emergente**: la tendenza dei punti a raggrupparsi in base alla macro–categoria suggerisce che gli embedding catturino informazione semantica significativa sul prefisso ATC. La disposizione nello spazio ridotto riflette infatti, almeno qualitativamente, la gerarchia del sistema ATC.

In sintesi, la visualizzazione t-SNE fornisce un primo riscontro qualitativo sulla qualità degli embedding: pur non mostrando una separazione perfetta, evidenzia una struttura coerente con le macro–categorie ATC.

4 Risultati

4.1 Metriche di classificazione

Il modello neurale addestrato sugli embedding di 88 dimensioni ha prodotto i seguenti risultati:

Tabella 4: Metriche di classificazione complete per livello gerarchico ATC (88 dim)

Metrica	ATC1	ATC2	ATC3
Accuracy	0.755	0.634	0.583
Balanced Accuracy	0.634	0.380	0.287
F1 Macro	0.653	0.401	0.285
F1 Weighted	0.733	0.591	0.512
Precision Macro	0.733	0.597	0.395
Recall Macro	0.634	0.380	0.287

- **Livello ATC1 (macro-categorie)**: Con un'accuracy del 75.5% e un F1 weighted del 73.3%, il modello mostra una capacità moderata di identificare le macro-categorie anatomiche. La balanced accuracy del 63.4% è significativamente inferiore, rivelando prestazioni ridotte sulle classi minoritarie.
- **Livello ATC2 (sottogruppi terapeutici)**: L'accuracy del 63.4% su 94 classi indica prestazioni limitate su questo livello gerarchico. Il gap sostanziale tra F1 weighted (59.1%) e F1 macro (40.1%) evidenzia un forte sbilanciamento nel dataset, con molte categorie che il modello fatica a classificare correttamente.
- **Livello ATC3 (sottogruppi farmacologici)**: Con 269 classi, l'accuracy del 58.3% indica una capacità predittiva non trascurabile, ma le metriche robuste allo sbilanciamento evidenziano limiti marcati: la balanced accuracy (28.7%) e l'F1 macro (28.5%) mostrano che il *recall medio per classe* è basso e che le prestazioni risultano concentrate sulle classi più rappresentate. Il divario tra F1 weighted (51.2%) e F1 macro conferma che molte classi meno frequenti vengono riconosciute con difficoltà. Nel complesso, gli embedding a 88 dimensioni catturano solo parzialmente l'informazione discriminativa necessaria a separare in modo affidabile le classi ATC3.

Le metriche di accuratezza sono calcolate sugli stessi codici utilizzati nel training (full-batch). Di conseguenza, questi valori misurano principalmente la capacità del modello di *adattarsi* alla struttura gerarchica presente nei dati più che una prestazione di generalizzazione su codici non visti.

4.2 Metriche di clustering

Tabella 5: Metriche di clustering basate sulle macro-categorie ATC1

Metrica	Valore
Silhouette Score	0.050
Davies-Bouldin Index	4.028
Calinski-Harabasz Index	88.2

- **Silhouette Score (0.050)**: valore vicino allo zero, che indica una separazione globale debole tra i cluster. Non si osserva una netta separazione geometrica, tuttavia le macro-categorie non risultano completamente sovrapposte nello spazio degli embedding
- **Davies-Bouldin Index (4.028)**: valore relativamente alto, che suggerisce la presenza di cluster con confini non nettamente separati. Questo comportamento è coerente con la natura del dominio farmacologico, in cui alcune macro–categorie ATC possono presentare sovrapposizioni funzionali o terapeutiche, pur appartenendo a rami distinti della classificazione.
- **Calinski-Harabasz Index (88.2)**: indica la presenza di una struttura di clustering non banale nello spazio degli embedding. Sebbene il valore non indichi una separazione netta tra i cluster, esso mostra che le categorie ATC presentano una struttura globale non casuale nello spazio degli embedding, con una varianza tra categorie superiore a quella interna.

4.3 K-NN Semantic Consistency

Tabella 6: Coerenza semantica locale tramite K-NN

Metrica	**Valore**
K-NN Consistency (k=10, macro ATC)	98.2%

Un valore del 98.2% è eccellente e indica che quasi tutti i vicini più prossimi di un codice ATC appartengono alla stessa macro–categoria. Questo risultato evidenzia una forte coerenza semantica locale: codici farmacologicamente simili nello spazio vettoriale tendono ad appartenere alla stessa categoria ATC1. La metrica conferma quindi che lo spazio degli embedding preserva in modo coerente la struttura locale della classificazione ATC.

4.4 Analisi delle distanze intra/inter categoria

Tabella 7: Confronto distanze euclidee medie

Tipo di confronto	**Distanza media**
Stessa macro-categoria (intra-classe)	11.919
Macro-categorie diverse (inter-classe)	12.966
Differenza	1.047

- La distanza intra-classe è inferiore alla distanza inter-classe, come desiderato.
- Tuttavia, la differenza è relativamente piccola, suggerendo che lo spazio degli embedding non presenta cluster completamente separati.
- Questo è coerente con il Silhouette Score basso: le categorie sono distinguibili ma con confini graduali piuttosto che netti.

4.5 Test gerarchico delle distanze

Tabella 8: Distanze medie in funzione della condivisione gerarchica

Livello di condivisione	Distanza media
Stessa ATC3 (4 caratteri)	8.493
Stessa ATC2, ATC3 diversa	10.922
Stessa ATC1, ATC2 diversa	12.331
ATC1 diversa	12.944

Si osserva un **gradiente di distanze perfettamente ordinato** che rispecchia la gerarchia ATC:

- **8.493**: farmaci nella stessa sottocategoria farmacologica (ATC3) sono i più vicini nello spazio vettoriale.
- **10.922**: farmaci nello stesso sottogruppo terapeutico (ATC2) ma con differenze farmacologiche (ATC3 diversa) hanno una distanza maggiore.
- **12.331**: farmaci nella stessa macro-categoria (ATC1) ma con scopi terapeutici diversi (ATC2 diversa) sono ancora più distanti.
- **12.944**: farmaci in macro-categorie anatomiche completamente diverse hanno la massima distanza.

Questo gradiente ordinato dimostra che gli embedding non solo catturano l'appartenenza categoriale, ma **preservano la struttura gerarchica del sistema ATC**. La similarità vettoriale riflette accuratamente la similarità farmacologica a tutti i livelli della classificazione.

5 Analisi Comparativa: Impatto della Dimensionalità

Per comprendere l'effetto della dimensione degli embedding sulla qualità delle rappresentazioni apprese, sono stati condotti esperimenti sistematici con cinque diverse configurazioni dimensionali. Tutti i modelli utilizzano la stessa architettura e procedura di training, variando solo le dimensioni degli embedding per livello.

5.1 Configurazioni testate

Tabella 9: Configurazioni dimensionali testate

Nome	Lvl 1	Lvl 2	Lvl 3	Lvl 4	Lvl 5	Totale
44 dim	4	8	8	8	16	44
66 dim	6	12	12	12	24	66
88 dim	8	16	16	16	32	88
100 dim	10	20	20	20	30	100
120 dim	12	24	24	24	36	120

5.2 Confronto delle metriche di classificazione

5.2.1 Livello ATC1 (macro-categorie)

Tabella 10: Metriche di classificazione ATC1 al variare della dimensionalità

Metrica	44 dim	66 dim	88 dim	100 dim	120 dim
Accuracy	0.611	0.671	0.755	0.778	0.862
Balanced Acc.	0.486	0.524	0.634	0.640	0.746
F1 Macro	0.494	0.529	0.653	0.631	0.751
F1 Weighted	0.576	0.623	0.733	0.733	0.837
Precision Macro	0.661	0.769	0.733	0.734	0.837
Recall Macro	0.486	0.524	0.634	0.640	0.746

5.2.2 Livello ATC2 (sottogruppi terapeutici)

Tabella 11: Metriche di classificazione ATC2 al variare della dimensionalità

Metrica	44 dim	66 dim	88 dim	100 dim	120 dim
Accuracy	0.458	0.479	0.634	0.654	0.692
Balanced Acc.	0.237	0.248	0.380	0.382	0.431
F1 Macro	0.244	0.258	0.401	0.392	0.442
F1 Weighted	0.398	0.417	0.591	0.592	0.633
Precision Macro	0.376	0.450	0.597	0.568	0.616
Recall Macro	0.237	0.248	0.380	0.382	0.431

5.2.3 Livello ATC3 (sottogruppi farmacologici)

Tabella 12: Metriche di classificazione ATC3 al variare della dimensionalità

Metrica	44 dim	66 dim	88 dim	100 dim	120 dim
Accuracy	0.462	0.542	0.583	0.636	0.677
Balanced Acc.	0.205	0.257	0.287	0.332	0.367
F1 Macro	0.196	0.251	0.285	0.322	0.366
F1 Weighted	0.388	0.468	0.512	0.557	0.611
Precision Macro	0.302	0.350	0.395	0.413	0.471
Recall Macro	0.205	0.257	0.287	0.332	0.367

5.2.4 Analisi comparativa delle metriche di classificazione

Trend generali tra i tre livelli:

- **Trend complessivo**: all'aumentare della dimensionalità, le prestazioni migliorano in modo coerente su tutti e tre i livelli (ATC1/ATC2/ATC3). L'effetto risulta più evidente sulle metriche robuste allo sbilanciamento (balanced accuracy e F1 macro), che misurano il recupero medio per classe.

- **ATC1**: le prestazioni crescono in modo regolare passando da 0.611 (44 dim) a 0.862 (120 dim). Anche la balanced accuracy aumenta (0.486 $\rightarrow$ 0.746), indicando un miglioramento del recall medio per classe e non soltanto delle classi più frequenti.
- **ATC2**: si osserva un incremento netto passando da 66 a 88 dimensioni (accuracy 0.479 $\rightarrow$ 0.634), con un miglioramento parallelo di balanced accuracy (0.248 $\rightarrow$ 0.380) e F1 macro (0.258 $\rightarrow$ 0.401). Il divario persistente tra F1 weighted e F1 macro in tutte le configurazioni evidenzia l'impatto dello sbilanciamento tra classi.
- **ATC3**: il livello più complesso (269 classi) beneficia dell'aumento dimensionale ma mantiene metriche macro relativamente basse (balanced accuracy $\leq$ 0.367, F1 macro $\leq$ 0.366), indicando difficoltà nel riconoscimento delle classi meno rappresentate. Il passaggio da 88 a 100 dimensioni mostra un miglioramento apprezzabile (accuracy 0.583 $\rightarrow$ 0.636), suggerendo una soglia di capacità del modello in quell'intervallo.
- **Incrementi marginali**: oltre 100 dimensioni i miglioramenti continuano, ma tendono a essere più contenuti rispetto ai salti osservati nelle configurazioni più compatte; questo è particolarmente visibile nelle metriche macro su ATC2 e ATC3.

5.3 Confronto delle metriche di clustering

Tabella 13: Metriche di clustering al variare della dimensionalità

Metrica	44 dim	66 dim	88 dim	100 dim	120 dim
Silhouette	0.042	0.049	0.050	0.063	0.060
Davies-Bouldin	4.275	3.894	4.028	3.700	3.638
Calinski-Harabasz	109.8	99.5	88.2	106.5	99.5

Analisi clustering:

- Il Silhouette Score rimane costantemente basso (0.042-0.063) per tutte le configurazioni, confermando che la separazione globale tra macro-categorie è debole indipendentemente dalla dimensionalità.
- Il Davies-Bouldin Index mostra un leggero miglioramento (diminuzione) per dimensioni maggiori: da 4.275 (44 dim) a 3.638 (120 dim).
- Il Calinski-Harabasz Index fluttua senza un trend chiaro, suggerendo che la struttura globale dei cluster è relativamente stabile rispetto alla dimensionalità.
- Nel complesso, le metriche di clustering globale sono meno sensibili alla dimensionalità rispetto alle metriche di classificazione

5.4 K-NN Semantic Consistency

Tabella 14: K-NN Consistency al variare della dimensionalità

Dimensione	44 dim	66 dim	88 dim	100 dim	120 dim
K-NN Consistency	94.4%	96.6%	98.2%	98.7%	98.7%

Osservazioni K-NN: La coerenza semantica locale aumenta con la dimensionalità da 94.4% a 98.7%) e raggiunge un plateau oltre 100 dimensioni.

5.5 Analisi delle distanze gerarchiche

Tabella 15: Distanze gerarchiche al variare della dimensionalità

Livello	44 dim	66 dim	88 dim	100 dim	120 dim
Stessa ATC3	6.007	7.443	8.439	8.455	9.397
Stessa ATC2, ATC3 $\neq$	7.854	9.593	10.922	11.613	12.902
Stessa ATC1, ATC2 $\neq$	8.843	10.732	12.331	13.298	14.559
ATC1 diversa	9.470	11.447	12.944	14.212	15.442
Ratio max/min	1.577	1.538	1.534	1.681	1.643

Analisi distanze gerarchiche:

- **Tutte le configurazioni preservano perfettamente il gradiente gerarchico**: le distanze crescono monotonamente secondo l'ordine ATC3 $<$ ATC2 $<$ ATC1 $<$ diversi
- Le distanze assolute scalano proporzionalmente con la dimensionalità: configurazioni più grandi producono spazi "più ampi" con distanze maggiori
- Il **ratio max/min** (distanza ATC1 diversa / distanza stessa ATC3) rimane notevolmente stabile tra 1.53 e 1.68, indicando che la *proporzione relativa* delle distanze gerarchiche è preservata indipendentemente dalla dimensionalità
- Le configurazioni 100 e 120 dim mostrano la maggiore separazione relativa (ratio $\approx$ 1.64-1.68), suggerendo una migliore discriminazione gerarchica

5.6 Distanze intra/inter categoria

Tabella 16: Distanze intra/inter categoria al variare della dimensionalità

Tipo	44 dim	66 dim	88 dim	100 dim	120 dim
Intra-classe	8.533	10.350	11.919	12.779	14.021
Inter-classe	9.487	11.446	12.966	14.202	15.434
Differenza	0.954	1.096	1.047	1.423	1.413
Diff. %	11.2%	10.6%	8.8%	11.1%	10.1%

Osservazioni distanze intra/inter:

- Tutte le configurazioni rispettano correttamente la proprietà fondamentale: distanza intra-classe $<$ distanza inter-classe
- La differenza assoluta aumenta con la dimensionalità (da 0.954 a 1.413), ma la **differenza percentuale rimane stabile** intorno al 10%, oscillando tra 8.8% e 11.2%
- La configurazione 88 dim presenta la separazione percentuale più bassa (8.8%), mentre 100 e 120 dim raggiungono valori intorno al 10-11%, indicando una separazione relativa leggermente migliore.