

# GENERAZIONE DI MOLECOLE ORGANICHE UTILIZZANDO LATENT DIFFUSION MODELS

Duccio Meconcelli

Relatore: Franco Scarselli

Correlatori: Pietro Bongini, Niccolò Pancino

A.A. 2023-2024

UNIVERSITÀ  
DI SIENA

1240



# OVERVIEW

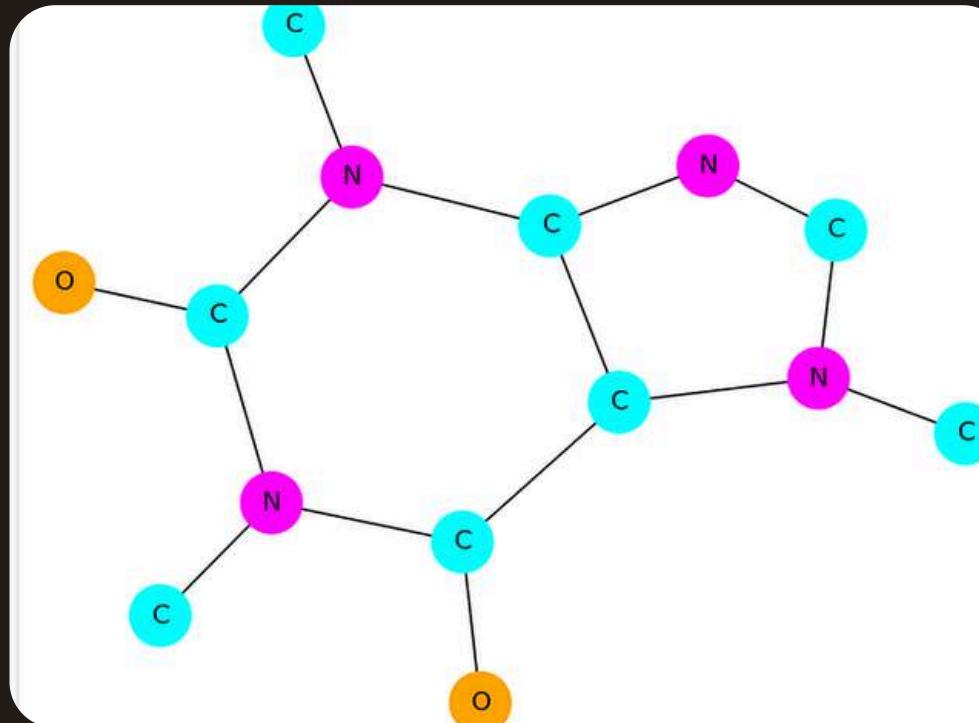
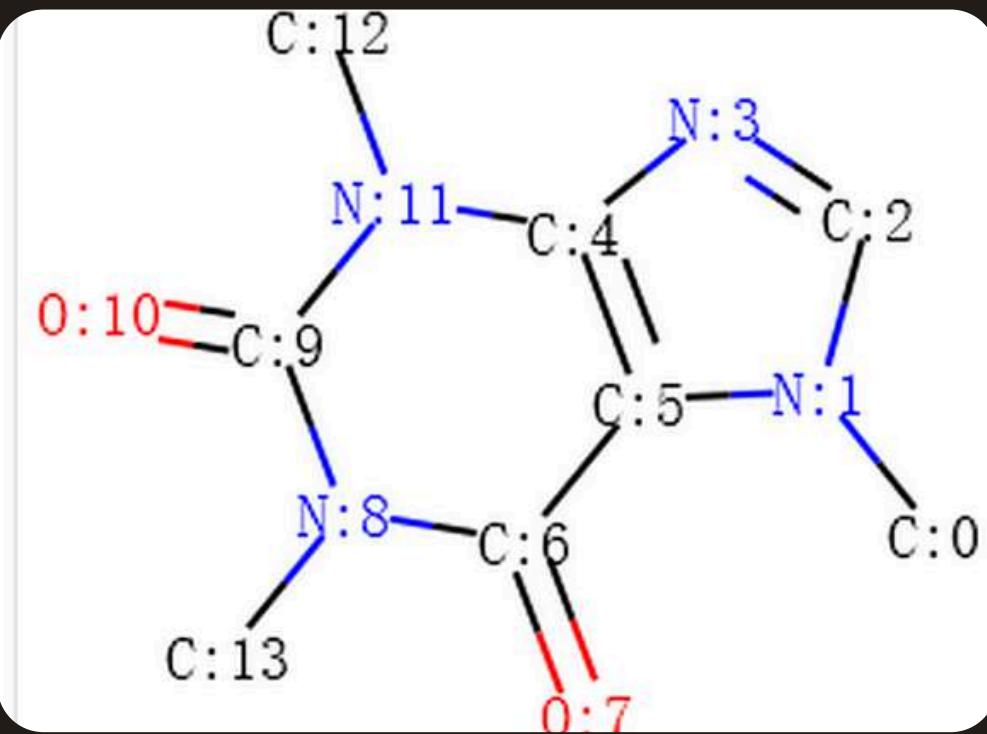
- 1 Introduzione
- 2 Background Teorico
- 3 Il Modello
- 4 Setup Sperimentale
- 5 Risultati e Analisi
- 6 Conclusioni
- 7 Sviluppi Futuri



# INTRODUZIONE



CHE  
MAS  
PE  
TIA  
IN  
FORNO



Problema:  
Generazione di Molecole

La generazione di grafi molecolari è cruciale per la scoperta di farmaci e la scienza dei materiali

Sfide:

- Generare molecole chimicamente valide
- Mantenere diversità e novità

Molecole come grafi:

- atomi (nodi) e legami (archi)

LA  
VIA

CIANI ASCOLTA CHIO UNA BUSTA DI INSALATA INSALATA MISTA COSI EH CHE DICE

LA  
VOLTE  
IO

SE  
LE  
VOLTE

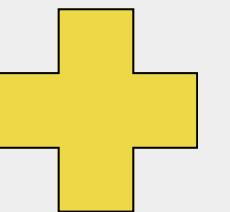
MANGIARE  
VOGLIE

# INTRODUZIONE

Tradizionalmente la generazione di grafi si basava su metodi statistici come i modelli Erdős-Rényi o Barabási-Albert.

- Il nostro approccio:
  - **GraphVAE**
  - **Integrazione con Latent Diffusion Model**

GRAPHVAE



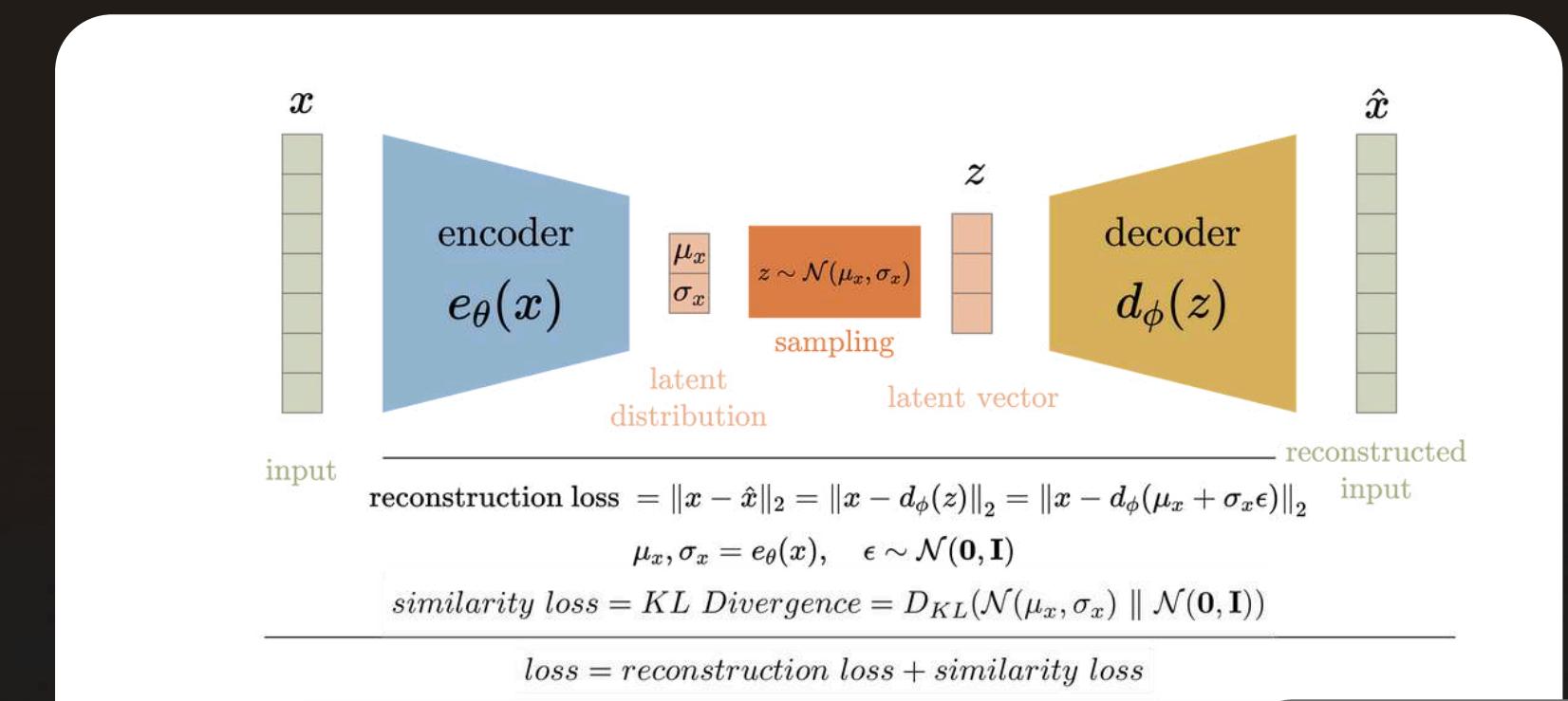
LATENT  
DIFFUSION  
MODEL



# VARIATIONAL AUTOENCODERS

## Variational Autoencoder (VAE):

- Encoder:
  - input: dati (es. immagini) → distribuzione latente
- Decoder:
  - input: campione spazio latente → ricostruzione
- Obiettivo:
  - minimizzare errore ricostruzione e KL divergence

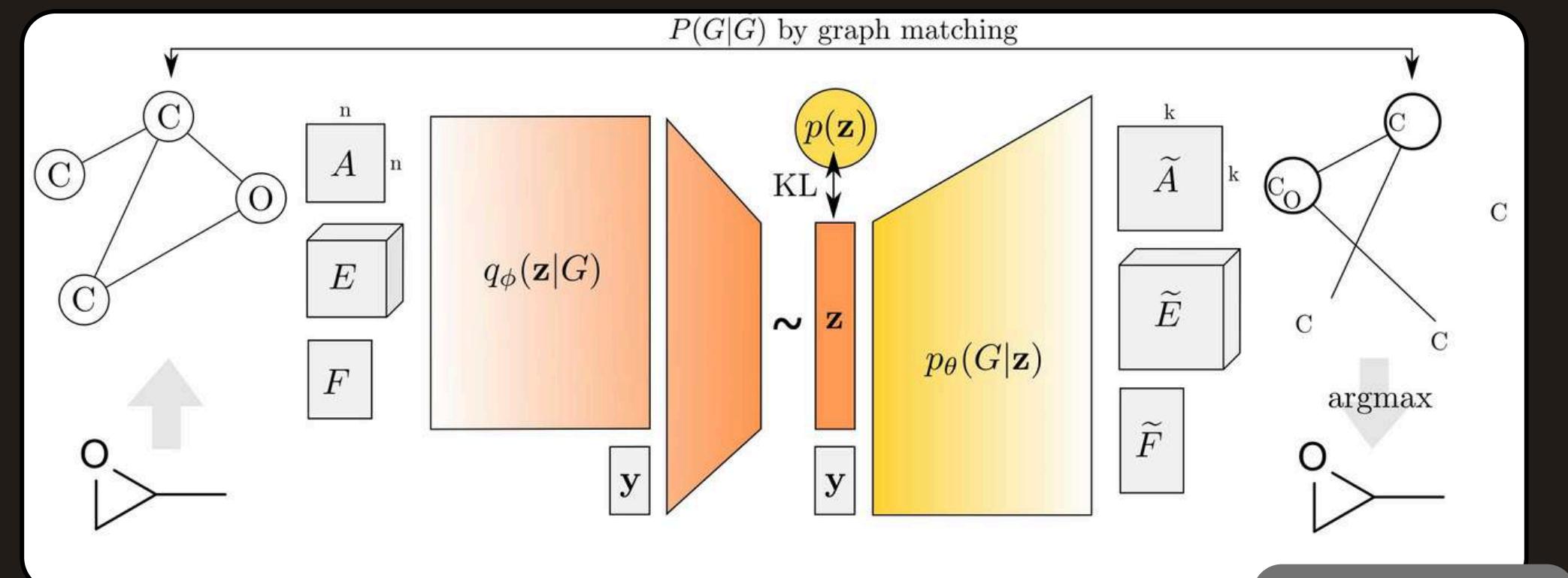


Aqeel Anwar

2 - Background

# GRAPH VARIATIONAL AUTOENCODERS

Il **GraphVAE** estende i VAE al dominio dei grafi, utilizzando layer convoluzioni su grafi nell'encoder e tecniche specifiche nel decoder per ricostruire la struttura del grafo.



## GraphVAE:

- Adatta VAE per strutture a grafo
- Encoder/Decoder specializzati per grafi
- Loss di ricostruzione che tiene conto anche di mantenere la struttura dei grafi



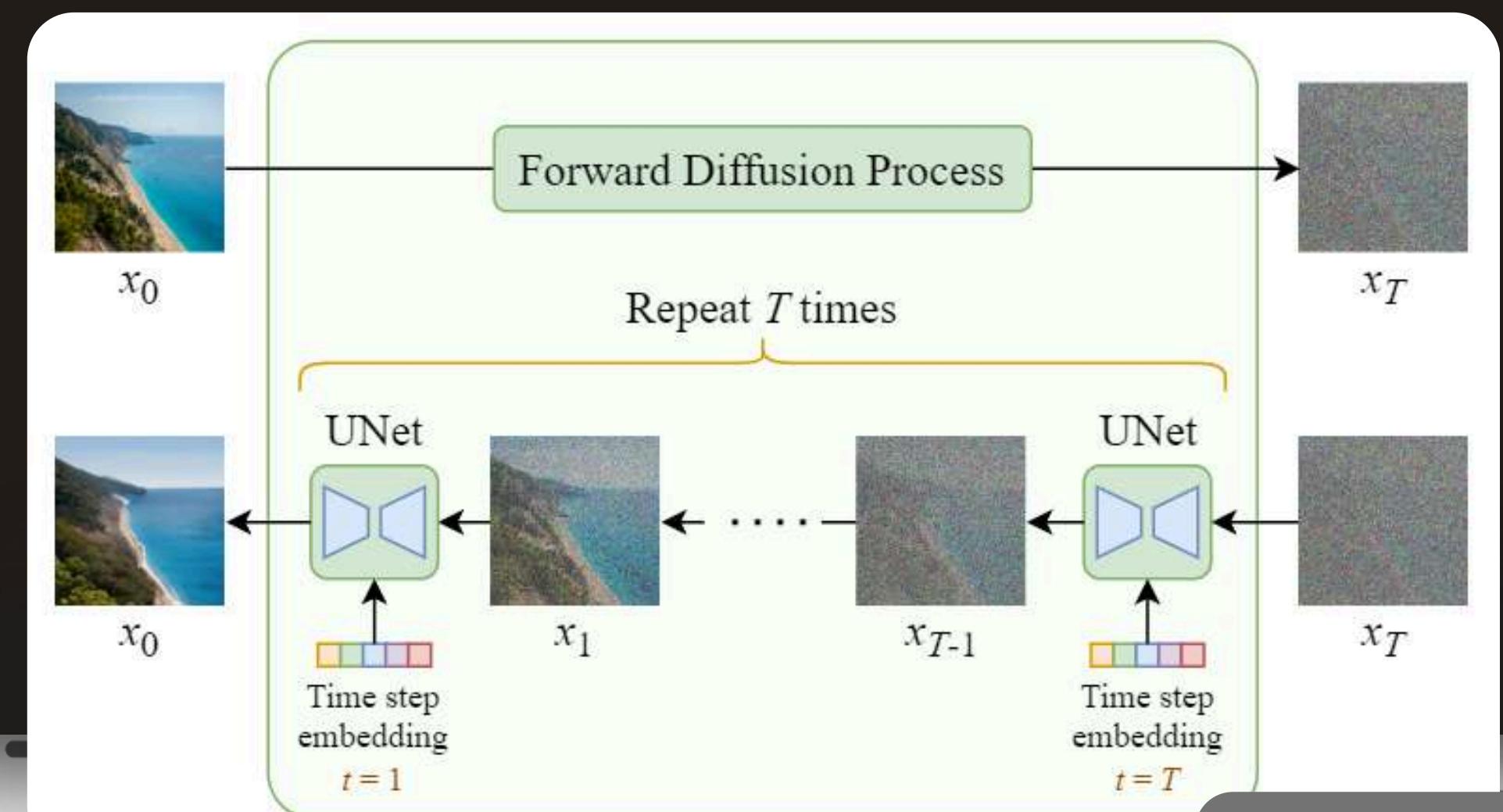
# DIFFUSION MODEL

Diffusion Model:

- Forward Process: aggiunge gradualmente rumore
- Processo inverso (denoising): rimuove rumore per generare dati

**Latent Diffusion Model (LDM):**

- Opera nello spazio latente definito da un VAE invece che sui dati
- Riduce complessità computazionale
- Cattura più informazioni "astratte"



Steins - Medium

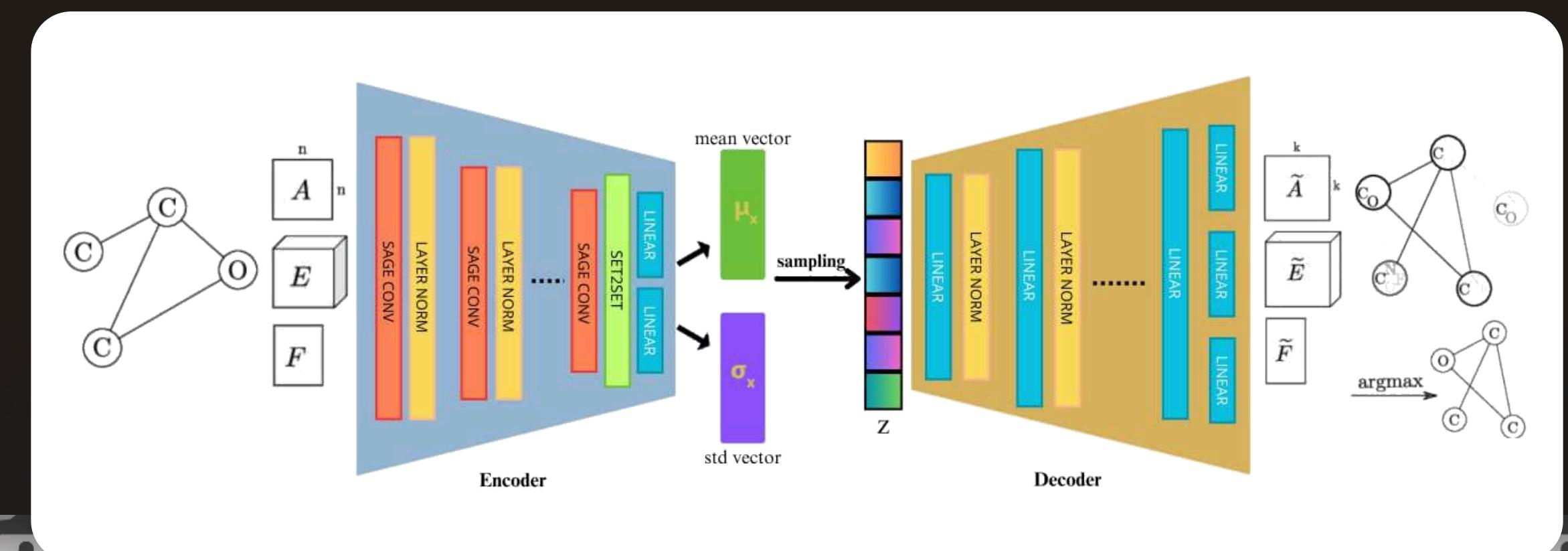
2 - Background

# IL NOSTRO MODELLO: GRAPHVAE

Differenze dal GraphVAE originale:

- Architettura: **SAGEConv layers** (informazioni aggiuntive date dall'edge index)
- **Fingerprint Loss**:
  - Cattura informazioni strutturali complesse
  - Calcolata usando RDKit o Morgan fingerprints
  - Guida il modello a rispettare la struttura del grafo

Fingerprint loss per preserva la struttura del grafo durante la ricostruzione.  
 Utilizzando una Neural Network addestrata, confronta le fingerprint molecolari delle molecole originali e ricostruite.



3 - Il Modello

# IL NOSTRO MODELLO: LATENT DIFFUSION MODEL

Architettura:

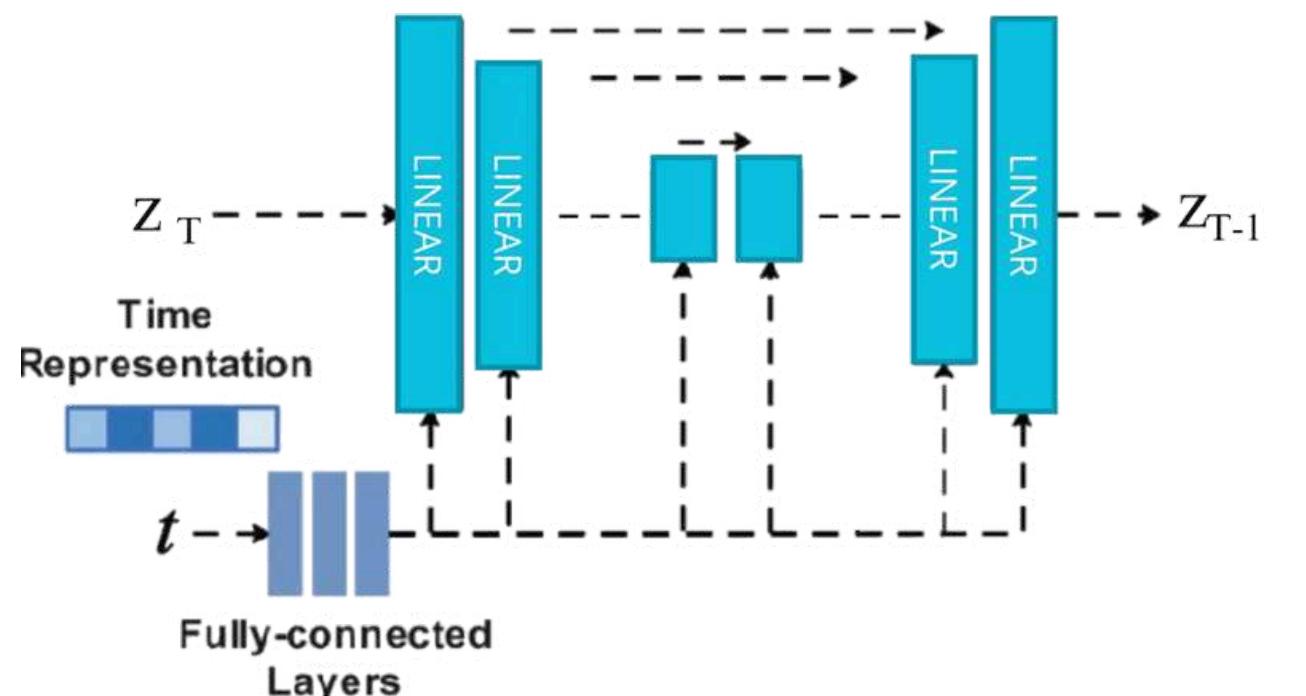
- U-Net modificata per vettori latenti
- Layer Lineari

Opera nello spazio latente del GraphVAE

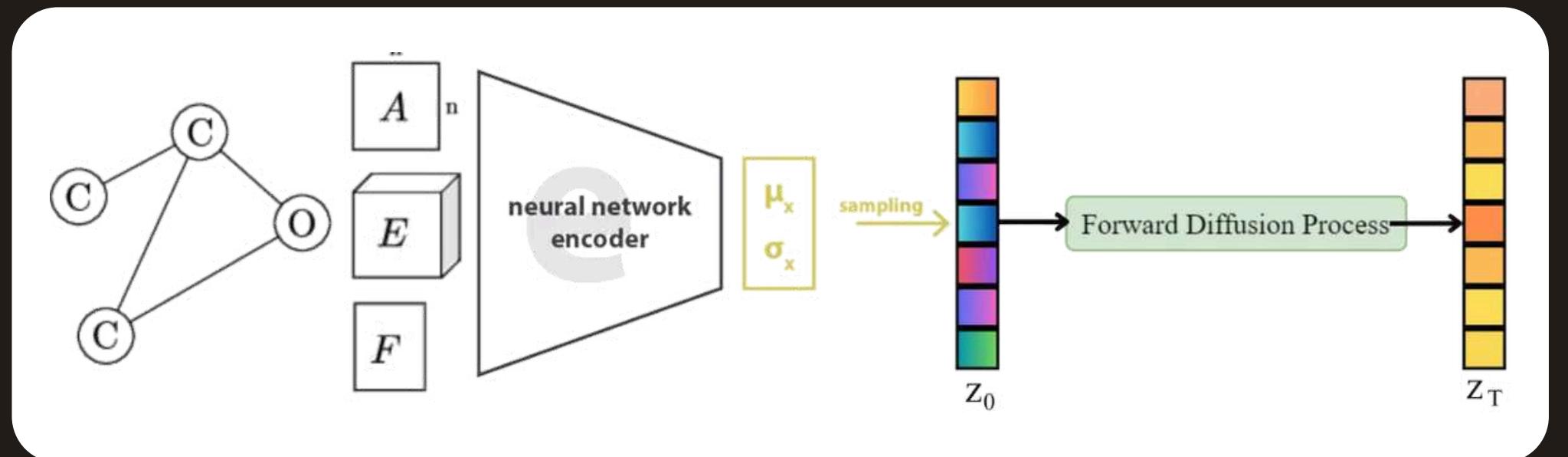
Vantaggi:

- Efficienza computazionale
- Migliore esplorazione dello spazio latente

Incorpora embedding temporali per condizionare il modello sul livello di rumore in ogni step del processo di diffusione.

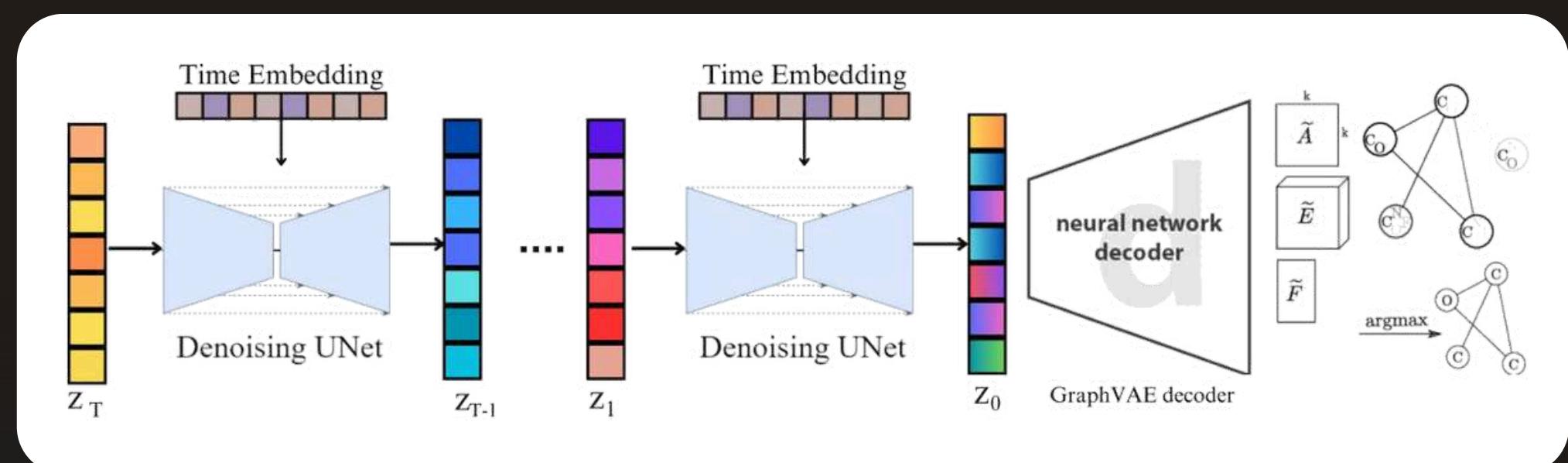


# IL NOSTRO MODELLO: ARCHITETTURA FINALE



- GraphVAE:  
codifica molecole  
in spazio latente

- LDM: genera nuovi punti nello spazio latente
- Decoder GraphVAE:  
converte punti latenti in molecole

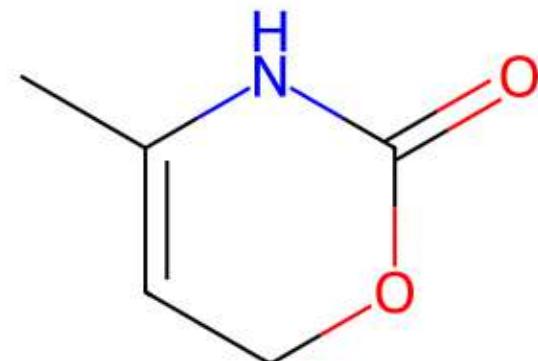


# SETUP SPERIMENTALE

Dataset: QM9 (130,828 molecole organiche)

Metriche calcolate su test set di 10k molecole:

- Validity, Uniqueness, Novelty
- Distribuzioni proprietà delle molecole (QED, LogP, peso molecolare, numero anelli)



Prove iniziali per scegliere un architettura di base,  
successivamente vari esperimenti:

- Uso di Graph Matching Loss Approssimata
- Variazione dimensione spazio latente (20, 80, 120)
- Confronto tipi di fingerprint (Morgan, RDKit Base)
- Confronto Diverse dimensioni di fingerprint (128, 2048)
- Allenamento senza fingerprint loss
- Variazione dimensioni dataset (50k vs 100k molecole)



QM9: contenente molecole organiche con al massimo 9 atomi  
pesanti (C, N, O, F).

4 - Exp. Setup

# RISULTATI - CONFRONTO BASE

Model	Validity (%)	Uniqueness (%)	Novelty (%)
GraphVAE (original)	57.1	52.0	71.9
Graphusion	97.35	99.41	98.49
Our GraphVAE (50k)	78.52	17.65	80.81
Our LDM (50k)	77.57	18.81	80.05
GraphVAE (100k)	86.52	6.97	83.91
LDM (100k)	82.89	8.51	84.54



# RISULTATI - CONFRONTO BASE

- LDM: maggiore diversità a scapito di lieve calo in validità
  - distribuzioni più simili al dataset originale
    - Miglior Fréchet distance per LogP e QED
- Entrambi i modelli superano significativamente il GraphVAE originale in validità e novità
- Ma molto inferiori al Graphusion (Stato dell'arte)
  - Architettura più avanzata
  - Meccanismo di self-guidance latente per modellare l'intera distribuzione usando pseudo etichette



# RISULTATI - BEST MODEL

Model	Validity (%)	Uniqueness (%)	Novelty (%)
Graphusion	97.35	99.41	98.49
Baseline GraphVAE (50k)	78.52	17.65	80.81
Baseline LDM (50k)	77.57	18.81	80.05
RDKit GraphVAE (50k)	71.77	23.14	90.85
RDKit LDM (50k)	68.06	27.12	91.22
Approx. Loss GraphVAE	6.62	69.03	95.84
Latent Space 20 GraphVAE	66.42	24.63	92.23
Latent Space 20 LDM	84.28	10.89	86.27

## RISULTATI - GLI ALTRI ESPERIMENTI

Riassunto dei principali risultati degli esperimenti:

- Impatto positivo della fingerprint loss sulla validità e novità
- Miglior performance con RDKit fingerprint rispetto a Morgan fingerprint
- Dimensione ottimale dello spazio latente intorno a 80

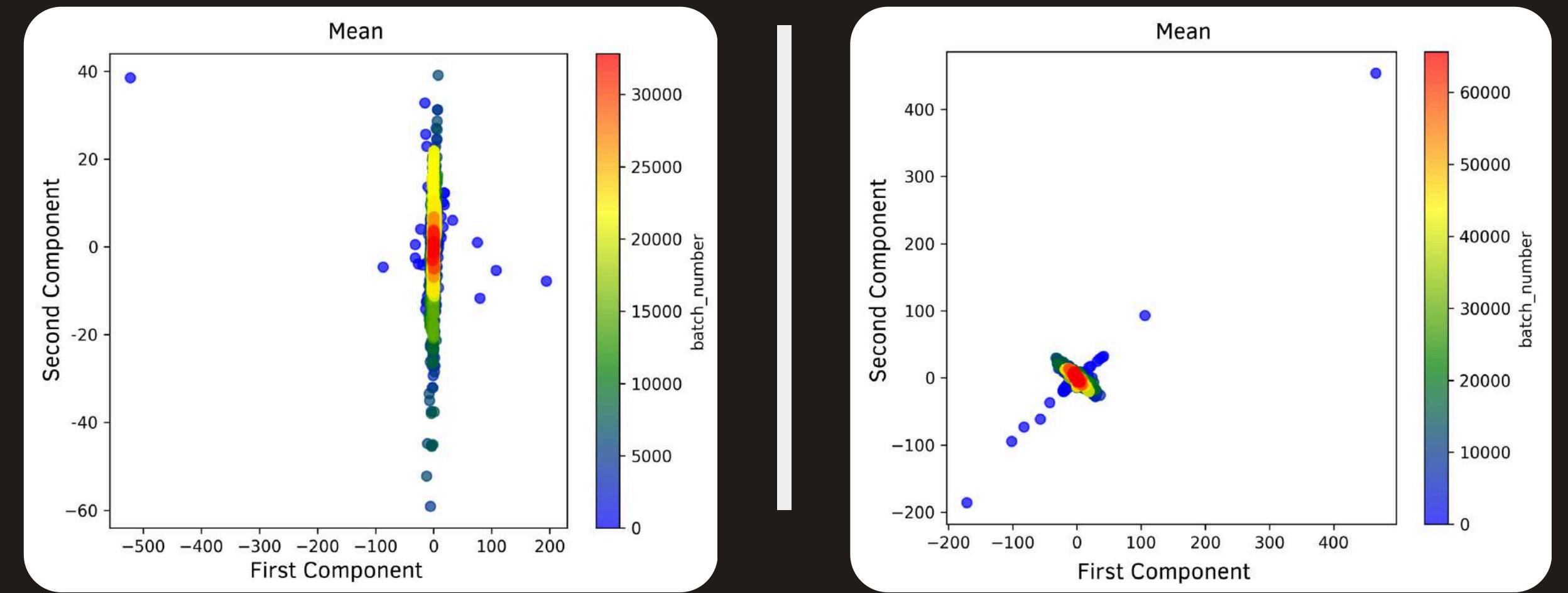
Dataset 100,000 molecole:

- Aumento validità
- Diminuzione unicità
- Possibile overfitting con dataset più grande
- Distribuzione dei campioni del GraphVAE più concentrata nello spazio latente
- Risultato -> miglior bilanciamento con 50,000 molecole



ORA VOGLIO PROPRIO VEDERE CHE TANTO IO VOGLIO A CORRETTO  
I GIORNI E HO VOGLIO PROPRIO

## RISULTATI - 50K VS 100K IN DETTAGLIO



Best GraphVAE 50k

Best GraphVAE 100k

Distribuzione dei campioni del GraphVAE 100k più concentrata  
nello spazio latente



NON CHI VAI A CORRE GUARDA, NON ASPETTO ALTRO

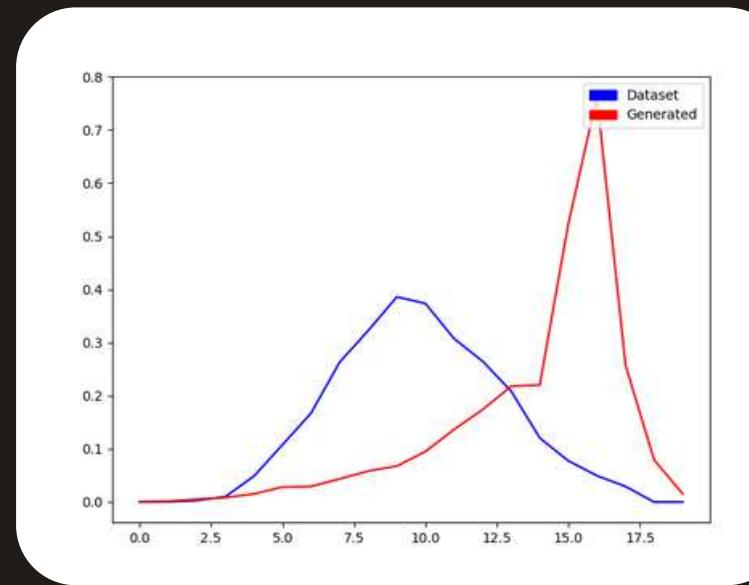
5 - Risultati

# RISULTATI - GRAPHVAE VS LDM

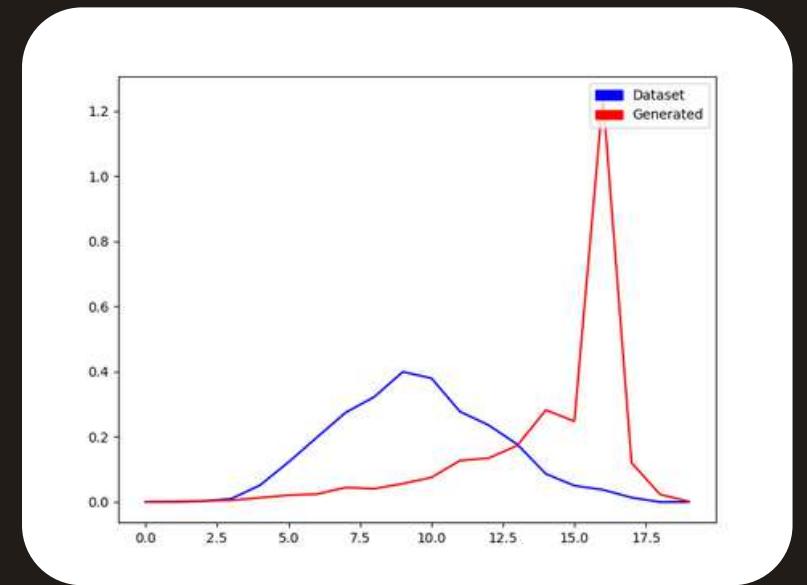
Differenze principali tra GraphVAE e LDM:

- GraphVAE: maggiore validità, minore unicità e novità
- LDM: maggiore unicità e novità, minore validità
- LDM cattura meglio la distribuzione delle proprietà molecolari

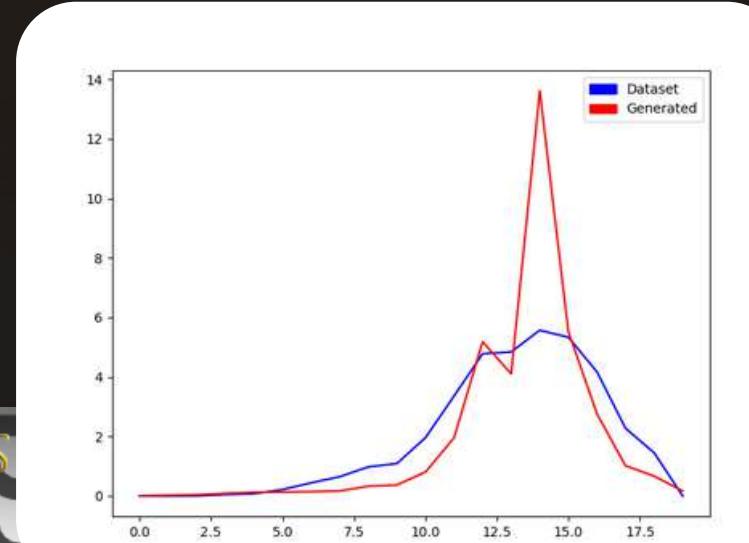
LogP Best LDM  
(50k)



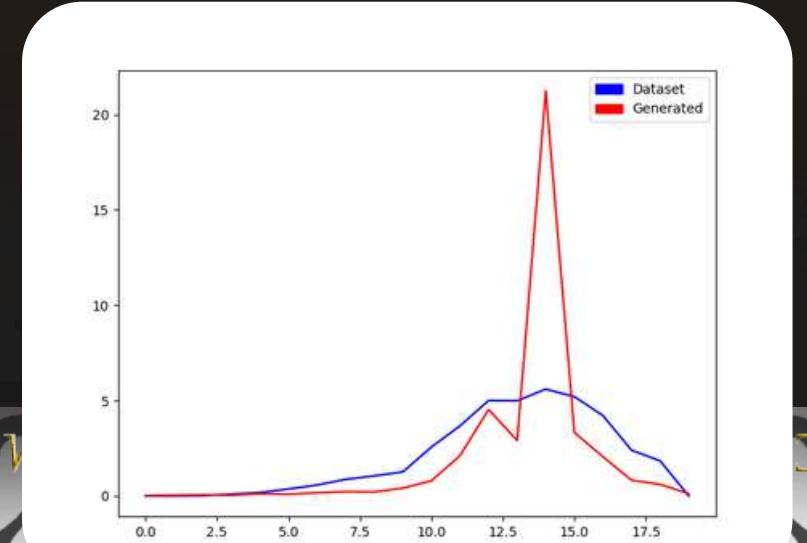
LogP Best  
GraphVAE (50k)



QED Best LDM  
(50k)



QED Best  
GraphVAE (50k)



# CONCLUSIONE

- Miglioramento significativo rispetto al GraphVAE originale
- LDM promettente per generazione diversificata
- Trade-off persistente tra validità e diversità
- Importanza del bilanciamento dataset e iperparametri
- Distanti dallo stato dell'arte (Graphusion)

Il nostro approccio ha dimostrato miglioramenti significativi rispetto al GraphVAE originale, soprattutto in termini di validità e novità. L'introduzione del LDM ha portato benefici nella diversità delle molecole generate. Tuttavia, c'è ancora un gap considerevole con lo stato dell'arte, indicando ampio spazio per miglioramenti futuri.

# SVILUPPI FUTURI

Possibili Miglioramenti:

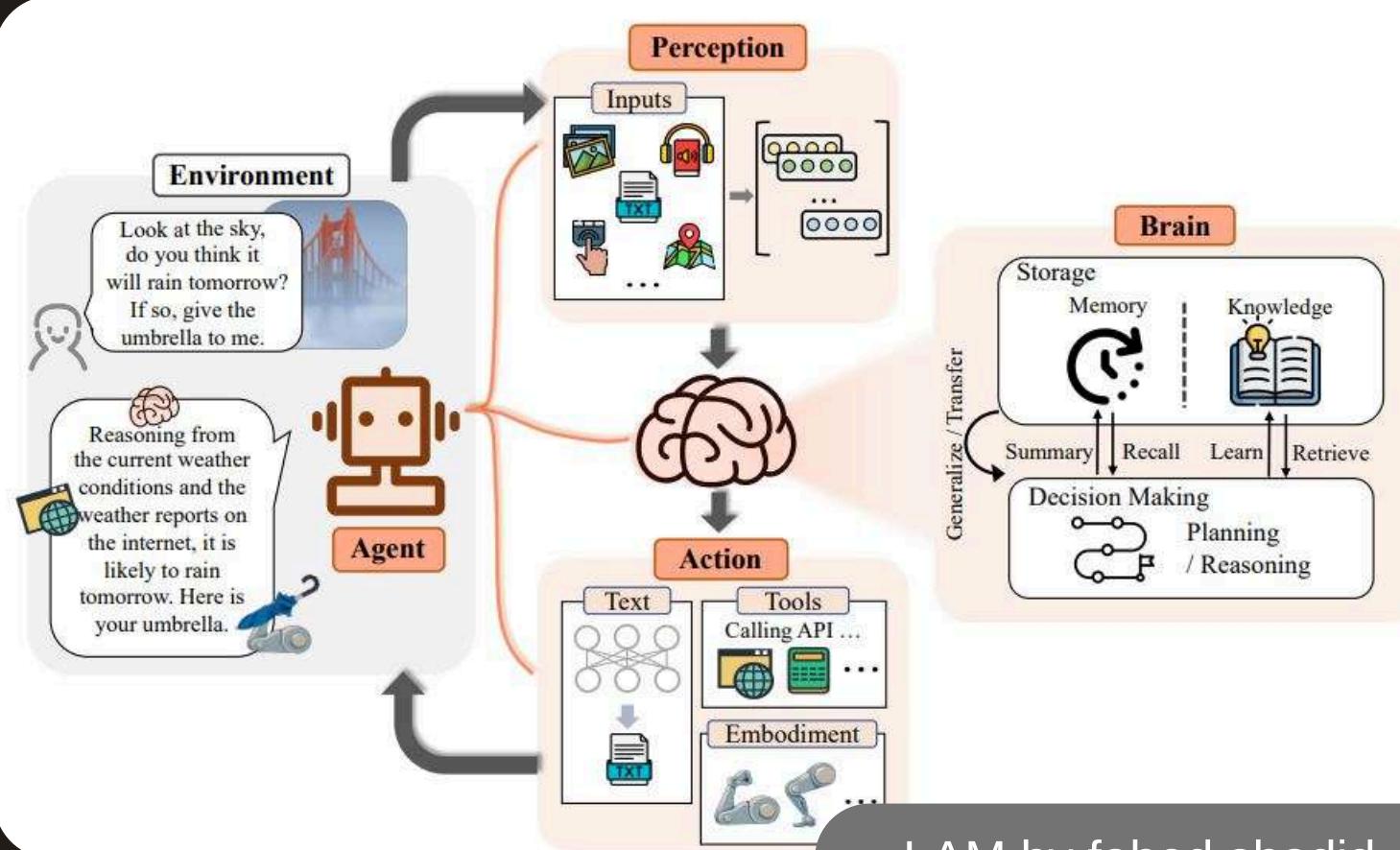
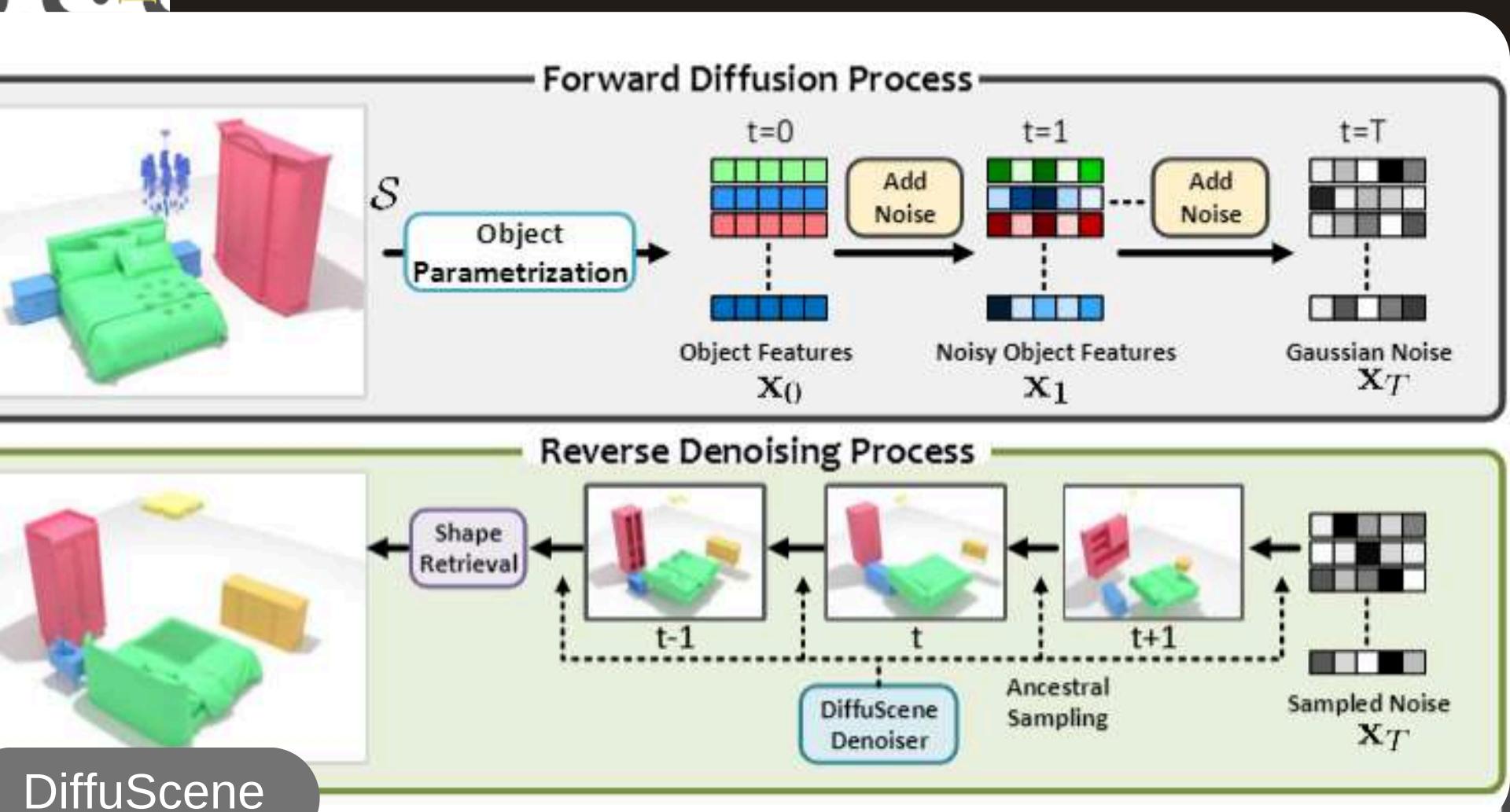
- Esplorazione di dataset più grandi e diversificati (ZINC250K)
- Condizionamento
- Ottimizzazione architettura e loss di ricostruzione
- Tecniche avanzate di regolarizzazione
- Sperimentare architetture più complesse per encoder e decoder

PORCO \*\*\* SONO UNA STUFA A LEGNA !!! MERDEEEEEE

# SVILUPPI FUTURI

Ambiti promettenti:

- Bioinformatica:
  - generazione di proteine, ricostruzione molecole
- **Large Action Model (LAM):**
  - generazione di azioni per navigazione UI in sistemi complessi



LAM by fahed shadid

## • Virtual Environment Generation:

- creazione di ambienti 3D per simulazioni (Diffuscene)

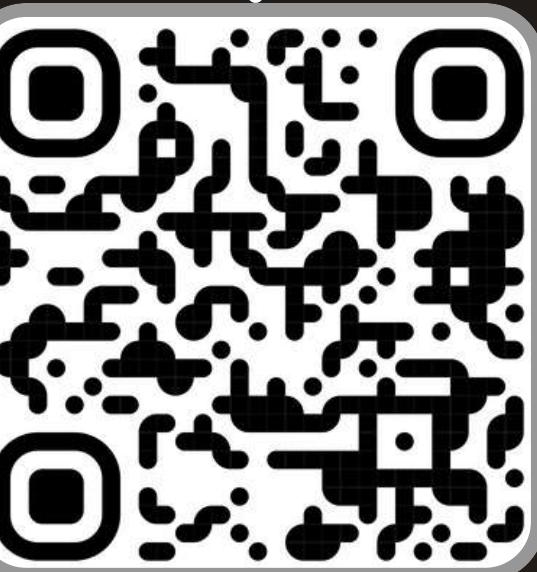
6 - Sviluppi Futuri

CIANI ASCOLTA C'HO UNA BUSTA DI INSALATA INSALATA MISTA COSÌ EH CHE DICE

**GRAZIE!  
DOMANDE?!**

[www.duccio.me](http://www.duccio.me)

Codice e Risultati  
disponibili qui



Fine

DAMIANO SE SEI DIGIUNO OSAT OIL

Codice e Risultati  
disponibili qui



"If there were more goodness in the world  
And each one saw a brother in another,  
There'd be fewer worries, fewer sorrows,  
And the world would shine much brighter."

WWW.DUCCIO.ME



- Fine?

MANDI QUELLO DI COSO LI... DI ZUCCHINO PORCO DI \*\*\* NON CAPISCI PROPRIO NIENNE GUAD