

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI FISICA E ASTRONOMIA “GALILEO GALILEI”

CORSO DI LAUREA IN FISICA

**COARSE-GRAINING AUTO ENCODERS PER LA
DINAMICA MOLECOLARE**

Relatore:

DOTT. EMANUELE LOCATELLI

Laureando:

GIACOMO DI PRIMA

Correlatore:

DOTT. FRANCESCO MAMBRETTI

Anno Accademico 2021/2022

Indice

1 Coarse-Graining	3
1.1 Modelli CG Consistenti	4
1.2 Modelli Top-Down	6
1.3 Modelli Bottom-Up	6
2 Autoencoders	7
2.1 Autoencoder Undercomplete	8
2.2 Autoencoder Regolarizzati	8
2.2.1 Autoencoders Sparsi e Penalizzazione delle Derivate	9
2.3 Profondità della Rete	9
3 Coarse-graining Auto-encoders for Molecular Dynamics	11
3.1 Architettura Utilizzata	11
3.1.1 Encoder	12
3.1.2 Struttura del Decoder	13
3.1.3 Loss Function	14
3.1.4 Potenziale Coarse-Grained	14
3.2 Training	14
3.3 Risultati	16
3.4 Discussione	18
4 Applicazione del Modello AE ai Polimeri ad Anello	19
4.1 Dati	20
4.2 Training	20
4.3 Risultati	21
4.4 Discussione	23
Bibliografia	25

Introduzione

Le simulazioni di dinamica molecolare (MD) costituiscono uno strumento importante per la comprensione della struttura microscopica di materiali e di sistemi biologici, e concorrono alla formulazione di previsioni e nuove teorie su di essi. Simulazioni a livello atomistico risultano essere piuttosto efficaci per investigare le dinamiche di strutture molecolari su scale temporali del nanosecondo e spaziali del nanometro, con risoluzioni rispettivamente del femtosecondo e dell' Ångstrom. Tuttavia, modelli che sfruttano un minor numero di dettagli, come quelli coarse-grained (CG), offrono la possibilità di studiare gli stessi sistemi su scale temporali e spaziali maggiori [1]. Tali rappresentazioni consentono di simulare il comportamento di sistemi atomistici trovando una rappresentazione semplificata di quest'ultimi, associando gruppi di atomi a singole particelle CG. Il coarse-graining comporta due problemi di apprendimento associati: definire la mappatura da una rappresentazione atomistica ad una ridotta e parametrizzare un'hamiltoniana su coordinate CG.

Nell'articolo del 2019 di Wujie Wang e Rafael Gómez-Bombarelli dal titolo “Coarse-graining auto-encoders for molecular dynamics” [2] entrambi i problemi sono stati affrontati usando il framework degli Autoencoders (AE), un particolare modello di rete neurale (NN) in grado di estrarre una rappresentazione dell'input che sfrutta un numero di gradi di libertà inferiore alla sua dimensione, e partendo da essa di restituire in output una ricostruzione dei dati originali. Nell'articolo la rete è stata addestrata ad ottenere rappresentazioni CG utilizzando posizioni e forze di tutti gli atomi provenienti da simulazioni atomistiche di alcune molecole. Una volta ottenute le rappresentazioni CG a diverse risoluzioni (numero di atomi CG usati) delle molecole in esame, sono state studiate le configurazioni ricostruite da esse per verificare che la rete avesse catturato le caratteristiche salienti della molecola. Inoltre sono state effettuate simulazioni CG per verificare l'applicabilità del framework a sistemi estesi. Questa tesi riporta un riassunto esteso dell'articolo e la riproduzione di alcuni risultati in esso riportati relativi all'ottenimento di una rappresentazione CG della molecola di alanina dipeptide, utilizzando gli esempi di codice che gli autori hanno messo a disposizione su GitHub¹. Inoltre ho studiato il comportamento della rete applicata a polimeri ad anello, della quale ho modificato i parametri legati all'apprendimento ed alla dimensione delle coordinate CG, ottenendo delle rappresentazioni del sistema con diverse risoluzioni.

¹<https://github.com/learningmatter-mit/Coarse-Graining-Auto-encoders>

Capitolo 1

Coarse-Graining

Simulazioni atomistiche di dinamica molecolare forniscono informazioni sulla struttura, la dinamica ed il funzionamento di molti e significativi sistemi soft matter, utilizzando modelli con dettagli dell'ordine di grandezza dell'Ångstrom e risoluzioni del femtosecondo [1]. Grazie allo sviluppo di hardware e software sempre più potenti, quotidianamente vengono simulati sistemi di biomolecole all'equilibrio per decine di nanosecondi e per distanze di svariati nanometri. Nonostante ciò, esistono diversi processi, come il ripiegamento delle proteine [3, 4], che avvengono su scale temporali dell'ordine del microsecondo o maggiori ed i loro meccanismi non possono essere studiati utilizzando convenzionali metodologie di MD [5]. Tra le tecniche sviluppate per poter investigare sistemi estesi, che non è possibile analizzare adeguatamente con dettagli atomistici, si trovano i modelli CG. Questi consistono nella rappresentazione del sistema molecolare attraverso siti di interazione (atomi/siti CG) che corrispondono a gruppi di atomi. Poiché delle buone rappresentazioni CG consentono di ridurre il numero di gradi di libertà del sistema originale, sono di gran lunga più efficienti in termini computazionali.

Il requisito fondamentale nella modellizzazione CG è che i risultati osservati nelle simulazioni a basse risoluzioni devono essere consistenti con quelli che si osserverebbero utilizzando modelli atomistici più dettagliati. Sebbene le rappresentazioni CG possano permettere lo studio esaustivo di configurazioni molecolari per grandi scale spaziali e temporali, i risultati così ottenuti potrebbero essere fuorvianti, a meno che l'ensemble di strutture a bassa risoluzione costituito da atomi CG non sia una rappresentazione a bassa risoluzione dell'ensemble che si osserverebbe utilizzando un modello atomistico [6]. Di conseguenza, lo sviluppo di una teoria formale di meccanica statistica per ottenere modelli a bassa risoluzione che siano consistenti con quelli atomistici risulta fondamentale.

1.1 Modelli CG Consistenti

La rappresentazione CG di un sistema è consistente con un particolare modello atomistico del medesimo sistema se [6]:

1. ogni coordinata e momento CG sono definiti come una combinazione lineare delle coordinate e momenti di un gruppo di atomi del modello atomistico;
2. la distribuzione all'equilibrio delle coordinate e dei momenti del modello CG è uguale a quella del modello atomistico.

Consideriamo un modello atomistico ed uno CG, siano T la temperatura del sistema in esame e k_b la costante di Boltzmann.

Lo stato dinamico istantaneo del sistema atomistico costituito da n atomi è descritto dalle coordinate cartesiane $r^n = \{r_1, \dots, r_n\}$ ed i momenti $p^n = \{p_1, \dots, p_n\}$. L'hamiltoniana atomistica sarà data da:

$$h(r^n, p^n) = \sum_{i=1}^n \frac{1}{2m_i} p_i^2 + u(r^n) \quad (1.1)$$

Nell'ensemble canonico la densità di probabilità degli stati dinamici all'equilibrio è:

$$p_{rp}(r^n, p^n) = p_r(r^n)p_p(p^n) \quad (1.2)$$

con

$$p_r(r^n) \propto \exp\left(-\frac{u(r^n)}{k_b T}\right), \quad p_p(p^n) \propto \exp\left(-\sum_{i=1}^n \frac{p_i^2}{2m_i k_b T}\right)$$

In modo analogo scriveremo le coordinate cartesiane del sistema CG con N siti di interazione come $R^N = \{R_1, \dots, R_N\}$ ed i momenti $P^N = \{P_1, \dots, P_N\}$. L'hamiltoniana sarà data da:

$$H(R^N, P^N) = \sum_{I=1}^N \frac{1}{2M_I} P_I^2 + U(R^N) \quad (1.3)$$

Nell'ensemble canonico la densità di probabilità degli stati dinamici all'equilibrio è:

$$P_{RP}(R^N, P^N) = P_R(R^N)P_P(P^N) \quad (1.4)$$

con

$$P_R(R^N) \propto \exp\left(-\frac{U(R^N)}{k_b T}\right), \quad P_P(P^N) \propto \exp\left(-\sum_{I=1}^N \frac{P_I^2}{2M_I k_b T}\right)$$

Il modello CG descrive i siti di interazione come punti dotati di massa e privi di struttura. Ogni coordinata CG viene costruita in termini delle coordinate del modello atomistico. Per esempio, uno specifico atomo CG potrebbe corrispondere al centro di massa di uno specifico

set di atomi o di una molecola. La definizione delle posizioni dei siti CG è data da un operatore lineare $M_R^N(r^n) = \{M_{R_1}(r^n), \dots, M_{R_N}(r^n)\}$ dalla forma:

$$M_{R_I}(r^n) = \sum_{i=1}^n c_{Ii} r_i \quad \text{per } I = 1, \dots, N \quad (1.5)$$

mentre per i momenti:

$$M_{P_I}(p^n) = M_I \sum_{i=1}^n c_{Ii} \frac{p_i}{m_i} \quad \text{per } I = 1, \dots, N. \quad (1.6)$$

Seguendo la definizione data in precedenza, un modello CG si dice consistente con quello atomistico se la probabilità congiunta delle coordinate CG e dei momenti (Eq. 1.4) è uguale a quella prodotta dalla densità di probabilità atomistica (Eq. 1.2) insieme agli operatori di mappatura (Eq. 1.5 e Eq. 1.6). Inoltre ci deve essere una relazione tra il potenziale CG $U(R^N)$ e quello atomistico $u(r^n)$: si dimostra [6] che il potenziale atomistico e l'operatore $M_R^N(r^n)$ determinano univocamente $U(R^N)$ a meno di una costante. Si rende necessario anche ridefinire le masse nel modello CG, che devono dipendere dalle masse di quello atomistico.

Nel caso di sistemi che non presentino vincoli intra-molecolari rigidi [6] è possibile fornire un set di condizioni sufficienti che implicano la consistenza tra il modello CG e quello atomistico di uno stesso sistema:

1. la posizione di ciascun sito CG è definita da un'espressione nella forma dell' Eq. 1.5;
2. ogni sito CG ha almeno un atomo che è assegnato a quel sito specifico;
3. le forze CG devono soddisfare la seguente equazione: $F_I(R^N) = < \mathcal{F}_I(r^n) >_{R^N}$, con $\mathcal{F}_I = \sum_j f_j(r^n) d_{Ij}/c_{Ij}$, dove f sono le forze atomistiche e la sommatoria viene calcolata sugli atomi che sono assegnati unicamente al sito I -esimo. L'espressione mette in relazione le forze medie all'equilibrio nell'ensemble canonico atomistico con quelle del modello CG.

Queste sono condizioni sufficienti affinché i due modelli siano consistenti nello spazio delle configurazioni. Perchè lo siano anche nello spazio delle fasi vi sono due condizioni sufficienti aggiuntive:

5. nessun atomo è coinvolto nella definizione di più di un sito CG;
6. le masse CG soddisfano l'equazione: $M_I = \left(\sum_i \frac{c_{Ii}^2}{m_i} \right)^{-1}$, dove la sommatoria viene effettuata sugli atomi assegnati al sito I -esimo.

1.2 Modelli Top-Down

Nei modelli CG “top-dow” le interazioni tra i siti sono parametrizzate senza considerare esplicitamente il modello atomistico. In particolare, le interazioni CG non vengono definite come un’approssimazione del potenziale a molti corpi della forza media di un sistema specifico. Invece esse vengono determinate sulla base di intuizioni chimico-fisiche oppure per riprodurre alcune proprietà fisiche generali o determinate proprietà strutturali o termodinamiche che vengono misurate.

Questi modelli possono essere utilizzati per descrivere un particolare sistema con una sufficiente specificità chimica [7]. In tal caso, spesso i potenziali di interazione sono modellizzati da semplici funzioni che vengono parametrizzate in modo da riprodurne le proprietà termodinamiche. Frequentemente ai siti CG vengono associati 3 o 4 atomi pesanti o molecole. Questo tipo di approccio è stato utilizzato, ad esempio, per simulare la densità e la tensione superficiale di un particolare sistema osservate sperimentalmente [8, 9].

1.3 Modelli Bottom-Up

Invece approcci di tipo “bottom-up” sviluppano modelli CG in cui le interazioni tra i siti vengono calcolate a partire da quelle dei modelli atomistici dei sistemi in esame. Il potenziale di forza media a molti corpi (PMF) è la grandezza centrale per questo tipo di approccio al coarse-graining. Esso viene definito dal modello atomistico e dalla mappa di assegnazione (Eq. 1.5) [7]. E’ importante notare che il PMF non è una convenzionale funzione di energia potenziale, ma deve essere considerata una funzione di energia libera che tiene conto dei contributi energetici ed entropici. Le forze CG determinate come il gradiente del PMF sono uguali alle medie delle forze atomistiche. Dunque il PMF è un potenziale che genera forze medie e quantifica il lavoro reversibile associato al cambiamento della configurazione CG di un modello atomistico per un sistema molecolare.

La sfida per questo tipo di approccio consiste nel determinare delle approssimazioni del PMF che siano trattabili dal punto di vista del calcolo, computazionalmente efficienti ed abbastanza accurate per descrivere un dato fenomeno o nel migliore dei casi, per modellizzare sistemi molecolari diversi da quelli per i quali erano state parametrizzate.

Può essere difficile ottenere informazioni quantitative e capacità predittive utilizzando metodi di coarse-graining top-down che, approssimando e semplificando le interazioni atomistiche, si focalizzano sulle proprietà che il sistema studiato manifesta piuttosto che sulle loro origini, come invece fanno i modelli bottom-up. Il problema dell’identificazione di variabili CG può anche essere affrontato sfruttando delle NN con particolari architetture. Nel capitolo successivo introdurrò le caratteristiche fondamentali di quella implementata nell’articolo di Wujie Wang e Rafael Gomez-Bombarelli del 2019.

Capitolo 2

Autoencoders

Un autoencoder (AE) è una rete neurale che viene addestrata per produrre un output simile all'input che riceve. La sua architettura può variare in termini di numero e dimensione dei layer interni, delle funzioni di attivazione utilizzate, ma in generale può essere vista come la composizione di una funzione parametrizzata encoder $h = f(x; \theta)$, con x input della rete, e una funzione parametrizzata decoder $r = g(h; \phi)$, con θ e ϕ vettori.

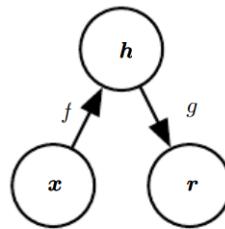


Figura 2.1: Struttura generale autoencoder [10]

Il processo di apprendimento avviene attraverso la minimizzazione di una funzione $L(x, g(f(x)))$, detta “loss function”, che penalizza $(g \circ f)(x)$ per essere dissimile dall’input. Un esempio può essere lo scarto quadratico medio (MSE) calcolato tra l’input e l’output, che è tanto più grande quanto maggiore è la differenza tra i due. Tale processo, iterazione dopo iterazione, permette di aggiornare i parametri θ e ϕ di h e r fino alla convergenza del valore della loss function. Nel caso ideale, tramite il processo di apprendimento, viene raggiunto il minimo globale della funzione, mentre nella pratica è possibile che la minimizzazione si fermi ad un minimo locale. Questo tipo di rete può essere utilizzata per imparare le caratteristiche salienti dei dati, i quali vengono compressi e proiettati in uno spazio h detto “spazio latente”, e a partire da esse viene prodotto un nuovo set di dati simili a quelli di partenza. Se l’AE apprendesse semplicemente una funzione identità, ovvero a copiare l’input nell’output, la rete non sarebbe in grado di imparare nulla sui dati che le vengono presentati. Per impedire ciò, possono essere utilizzate diverse tecniche. Esamineremo di seguito quelle che sono

state implementate nella costruzione del modello per la rappresentazione CG nell'articolo studiato.

2.1 Autoencoder Undercomplete

Tipicamente la struttura dell'AE presenta una dimensione di h inferiore a quella di x . In questa configurazione la rete si dice “undercomplete”. L'effetto a collo di bottiglia (“bottleneck”) che si viene a creare costringe la rete a proiettare l'input in uno spazio di dimensione inferiore, nel quale dovrebbero essere codificate le informazioni essenziali del sistema da cui ricostruire i dati.

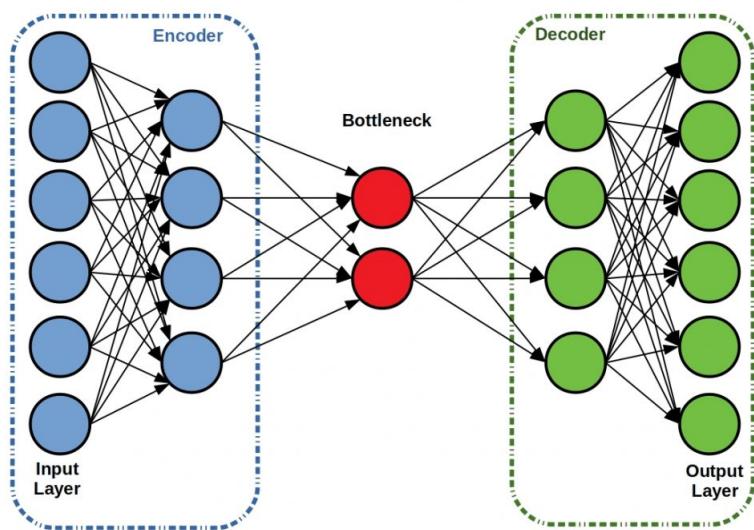


Figura 2.2: Architettura AE.(<https://starship-knowledge.com/autoencoder>)

La dimensione del boottleneck è critica per il corretto apprendimento delle principali caratteristiche dei dati. Se la dimensione è troppo piccola rispetto a quella dell'input, la rete non riesce a descrivere tutti gli aspetti salienti del sistema. Al contrario, se la dimensione dello spazio latente è troppo grande, la rete può imparare ad effettuare una semplice copia dell'input senza però acquisire una capacità di generalizzazione nel comprimere nuovi dati a cui la rete non è stata esposta durante la fase di training.

2.2 Autoencoder Regolarizzati

Un'altra tecnica utilizzata per ottenere una compressione dei dati significativa dal punto di vista delle informazioni catturate prevede l'aggiunta di un termine di regolarizzazione alla loss function tale che la rappresentazione cercata presenti delle particolari proprietà come la robustezza al rumore ed agli input mancanti, la sparsità (in questo contesto il termine indica che il valore ottimale di alcuni parametri della rete è 0) e valori piccoli della derivata.

2.2.1 Autoencoders Sparsi e Penalizzazione delle Derivate

Gli AE sparsi presentano una loss function nella forma:

$$L(x, g(f(x))) + \Omega(h)$$

con $\Omega(h)$ temine di penalizzazione di sparsità. Un AE che sia stato regolarizzato per essere sparso deve tenere conto di caratteristiche statistiche uniche dei dati anziché agire come una funzione identità.

Un’altro modo di regolarizzare la loss function consiste nella penalizzazione del valore del gradiente delle variabili dello spazio latente. In questo caso la loss function assume la forma:

$$L(x, g(f(x))) + \Omega(h, x)$$

con Ω definito come:

$$\Omega(h, x) = \lambda \sum_i ||\nabla_x h_i||^2$$

dove h_i sono le variabili dello spazio latente. Ciò costringe il modello ad imparare una funzione, il cui valore non cambia molto in seguito ad una piccola variazione dell’input.

2.3 Profondità della Rete

Nelle reti “feedforward”, le connessioni tra i nodi delle quali non formano cicli, come gli AE, ci possono essere diversi vantaggi nell’impiego di architetture che utilizzano più layer (“Deep neural networks”).

In base alla classe di funzioni che si vogliono rappresentare, l’addestramento di deep AE consente:

- un’approssimazione arbitrariamente accurata della mappa tra l’input e le variabili dello spazio latente;
- una possibile riduzione dei costi computazionali e un decremento esponenziale dei dati necessari al training della rete;
- una migliore compressione rispetto ad AE con singolo layer e che rappresentano funzioni lineari.

Capitolo 3

Coarse-graining Auto-encoders for Molecular Dynamics

Nell'articolo da me studiato il problema di trovare una rappresentazione CG di molecole viene riformulato come l'apprendimento, da parte di una NN, delle variabili latenti di configurazioni atomistiche. Il modello sviluppato si basa sul framework degli AE e presenta le seguenti caratteristiche:

1. impara delle rappresentazioni CG ed è in grado di decodificarle in coordinate atomiche;
2. riesce a trovare il potenziale CG che corrisponde alla forza media istantanea che agisce sui singoli atomi.

Ciò consente di effettuare simulazioni che comportano minori costi computazionali e dalle quali possono essere ricavate le informazioni relative alla struttura delle molecole in esame.

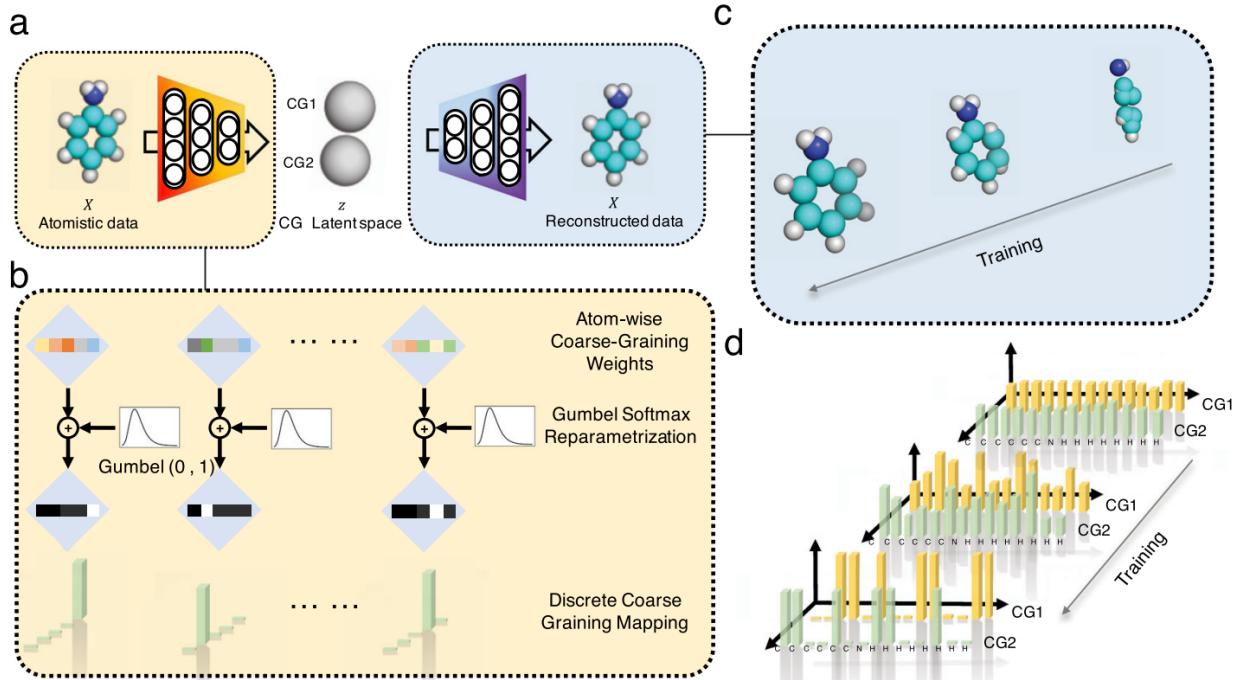
3.1 Architettura Utilizzata

La rete è formata da un encoder ed un decoder lineari, costituiti da un singolo layer. Definendo N_{atoms} il numero di atomi nella molecola e stabilita la dimensione dello spazio latente, che in questo caso si traduce nel numero di siti CG (N_{CG}) che vogliamo rappresentino l'intera molecola, entrambe le parti dell'AE saranno costituite da matrici con dimensioni riportate nella tabella 3.1 .

Tabella 3.1: Dimensioni Matrici Encoder e Decoder

Encoder	$N_{CG} \times N_{atoms}$
Decoder	$N_{atoms} \times N_{CG}$

Figura 3.1: Framework Coarse-Graining AE. **a** Il modello consiste in un encoder ed un decoder, ed è addestrato a ricostruire i dati originali proiettando le traiettorie atomistiche in uno spazio di dimensione inferiore. **b** Grafico della parametrizzazione della mappatura nello spazio CG utilizzando la gumbel-softmax. **c** Processo di ricostruzione delle molecole condizionato dalla rappresentazione CG ottenuta. **d** Evoluzione dei parametri della rete durante il training :l'asse x rappresenta i singoli atomi, mentre l'asse y rappresenta i due siti CG usati per la rappresentazione. Durante l'addestramento della rete i parametri di assegnazione, inizializzati casualmente, vengono aggiornati fino a raggiungere una codifica one-hot encoded.



3.1.1 Encoder

Siano x le coordinate atomistiche e z le coordinate CG, $n \equiv N_{atoms}$ e $N \equiv N_{CG}$ la funzione di encoding è una funzione proiettiva tale che $E(x) : \mathbb{R}^{3n} \rightarrow \mathbb{R}^{3N}$. L'encoder deve soddisfare le seguenti proprietà:

1. $z_{ik} = E(x) = \sum_{j=1}^n E_{ij} x_{jk} \in \mathbb{R}^3$, con $i = 1, \dots, N$ e $j = 1, \dots, n$. i è l'indice della variabile CG, j è l'indice atomico, k rappresenta l'indice della coordinata cartesiana (x, y, z). La matrice z è definita come il prodotto matriciale tra la matrice dell'encoder e quella dei dati.
2. $\sum_j E_{ij} = 1$, con $E_{ij} \geq 0$. Le variabili CG vengono definite come la media geometrica delle coordinate cartesiane degli atomi.
3. Ogni atomo contribuisce al massimo ad una variabile CG.

La matrice x ha dimensioni $n \times 3$ mentre quelle della matrice z sono $N \times 3$. Per essere consistenti nello spazio dei momenti dopo la mappatura nelle variabili CG, vengono ridefinite

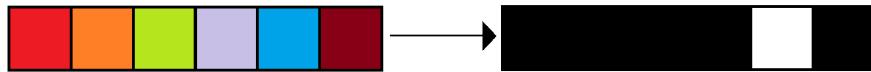
le masse:

$$M_i = \left(\sum_j \frac{E_{ij}^2}{m_j} \right)^{-1}$$

con M_i massa dell'atomo CG i-esimo, m_j massa dell'atomo j-esimo.

I parametri della funzione encoder vengono inizializzati casualmente come vettori $\vec{\phi}$ le cui entrate rappresentano la probabilità degli atomi di essere assegnati ad un particolare sito CG. Essi formano una matrice di elementi ϕ_{ij} che costituiscono le probabilità di assegnazione dell'atomo j alla variabile coarse-grained i . Questi vengono aggiornati durante il training della rete in modo da ottenere dei vettori che possiedono un'unica entrata diversa da zero, responsabile dell'assegnazione degli atomi alle rispettive variabili CG. Questa codifica si chiama “one-hot encoded” e può essere formalizzata come $\text{one-hot}(\vec{C}_j) = \arg \max_i \phi_{ij}$. A questa definizione, che è discontinua, si preferisce costruire i parametri della rete utilizzando una parametrizzazione differenziabile.

Figura 3.2: One-Hot Encoding. I colori rappresentano i valori relativi alle entrate dei vettori. Gli elementi del vettore, in generale diversi tra loro, diventano tutti nulli tranne uno.



Quindi:

$$E_{ij} = \frac{C_{ij}}{\sum_j^n C_{ij}}$$

I coefficienti C_{ij} vengono parametrizzati (figura 3.1 b) utilizzando la funzione gumbel-softmax [11], una distribuzione continua che può essere ricondotta ad una categorica con continuità:

$$C_{ij} = \frac{e^{\frac{g_{ij} + \log(\phi_{ij})}{\tau}}}{\sum_j e^{\frac{g_{ij} + \log(\phi_{ij})}{\tau}}}$$

dove g_{ij} è un numero estratto casualmente dalla gumbel-softmax e τ un parametro chiamato “temperatura”. Per $\tau \rightarrow 0$ (figura 3.1 d) la parametrizzazione di C_{ij} tende alla funzione $\arg \max$, rendendo possibile il one-hot encoding. Questa parametrizzazione ha il pregio di essere differenziabile rispetto ai parametri della rete E_{ij} e permette di utilizzare l'algoritmo di back-propagation per l'ottimizzazione dell'AE. Quest'ultimo sfrutta il gradiente della loss function rispetto ai parametri del modello, dopo averlo calcolato a partire dalle derivate parziali di ogni operazione effettuata sui dati.

3.1.2 Struttura del Decoder

L'approccio adottato è identico a quello utilizzato per l'encoder: viene utilizzata una funzione di proiezione che mappa le variabili CG nello spazio atomistico. Questo avviene tramite una

matrice D , di dimensioni $N_{atoms} \times N_{CG}$, tale che:

$$x_{recon} = D(z) = \sum_{i=1}^N D_{ji} z_{ik}$$

3.1.3 Loss Function

La loss function che viene ottimizzata durante l'apprendimento della rete è la seguente:

$$L_{ae} = \frac{1}{N} \mathbb{E}_x [(D(E(x)) - x)^2 + \rho F_{inst}(E(x))^2]$$

con \mathbb{E}_x la media aritmetica rispetto ai valori di x . Essa consiste in due addendi:

1. il primo termine computa l'errore di ricostruzione come lo MSE, calcolando la differenza tra le coordinate atomistiche originali e quelle ricostruite;
2. il secondo consiste in un termine di regolarizzazione che permette l'apprendimento di un profilo di energia libera regolare. Questo viene calcolato come la media delle forze istantanee agenti sugli atomi estratte dalle simulazioni di MD, moltiplicato per un iper-parametro ρ , scelto dallo sperimentatore, che misura il peso relativo del termine di regolarizzazione rispetto a quello di ricostruzione.

3.1.4 Potenziale Coarse-Grained

Per l'apprendimento del potenziale Coarse-Grained viene impiegata una rete diversa, che utilizza lo schema della corrispondenza delle forze istantanee [12, 13]. La loss function di questa nuova rete è definita come:

$$L_{inst} = \mathbb{E}[(F(z) + \nabla_z V_{CG}(z))^2]$$

3.2 Training

Al modello sono state fornite in input traiettorie atomistiche in condizioni di equilibrio composte da qualche migliaio di frame e le forze, estratte da simulazioni di MD, associate a ciascun atomo per ciascun frame.

Prima l'AE è stato addestrato al fine di ottenere la rappresentazione CG delle molecole in esame. Per l'OTP (orto-terfenile) e l'anilina sono stati impiegati 3000 frame, mentre per lalanina dipeptide 5000, ottenuti da mdshare¹ utilizzando il pacchetto PYEMMA². In quest'ultimo caso il termine di regolarizzazione non è stato calcolato poiché nei dati ottenuti

¹<https://markovmodel.github.io/mdshare/>

²<http://www.emma-project.org/latest/>

non erano presenti le forze. Per l'etano, il propano e il $C_{24}H_{50}$ l'addestramento è stato effettuato utilizzando 10000 frames senza il termine di regolarizzazione poiché la topologia delle molecole è relativamente semplice.

Iterazione dopo iterazione i parametri della rete vengono ottimizzati in modo da diminuire l'errore di ricostruzione, migliorando via via la rappresentazione CG della molecola in esame. Il termine di regolarizzazione viene introdotto a partire dalla 400-esima.

L'algoritmo implementato è riportato di seguito.

Algorithm 1 Addestramento AE

 $\phi_{ij}, D_{ji}, \tau, \Delta\tau \leftarrow$ inizializzazione parametri

repeat
 $x \leftarrow$ posizioni degli atomi di un frame della traiettoria molecolare

 $g_{ij} \leftarrow$ gumbel-softmax(0,1)

$$C_{ij} \leftarrow \frac{e^{\frac{g_{ij} + \log(\phi_{ij})}{\tau}}}{\sum_j e^{\frac{g_{ij} + \log(\phi_{ij})}{\tau}}}$$

$$E_{ij} \leftarrow \frac{C_{ij}}{\sum_j^n C_{ij}}$$

$$g \leftarrow \nabla_{\phi_{ij}, D_{ji}} L_{AE}(\phi_{ij}, D_{ji}; \tau, g_{ij})$$

 $\phi_{ij}, D_{ji} \leftarrow$ aggiorna i parametri utilizzando g

$$\tau \leftarrow \tau - \Delta\tau$$

until L_{AE} converge

Per fare in modo che la mappatura dalle coordinate atomistiche a quelle CG non ignori gradi di libertà disponibili al sistema, nel caso di grandi forze istantanee, è importante scegliere valori di ρ bassi, che sono presentati nella tabella 3.3.

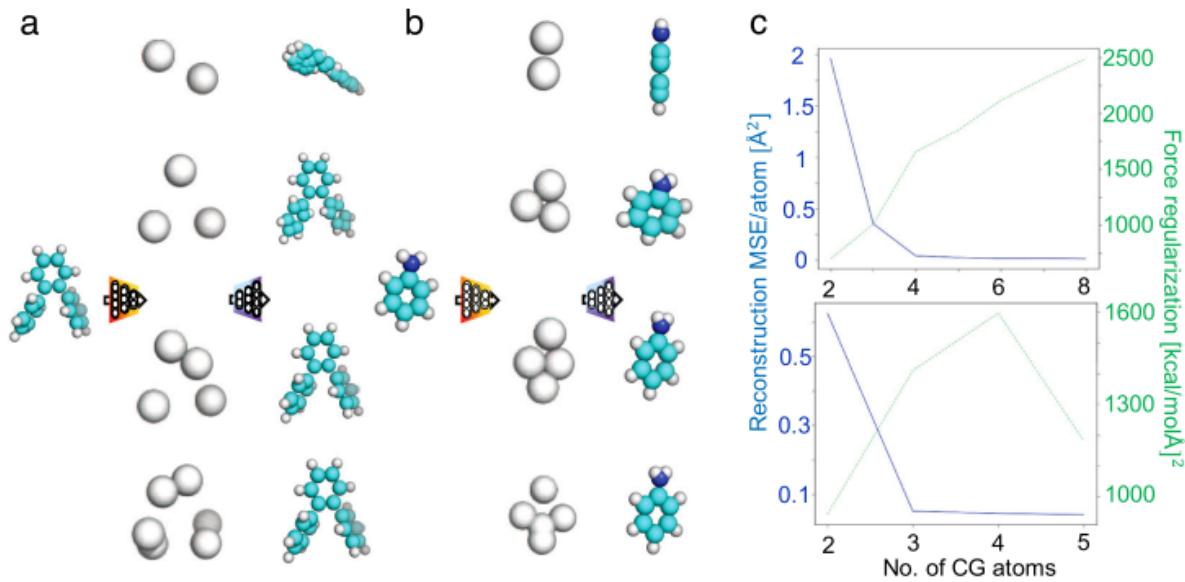
Tabella 3.2: Valori di ρ in funzione delle risoluzioni CG impiegate

Risoluzione OTP	ρ	Risoluzione Anilina	ρ
2	2e-2	2	1e-2
3	1e-2	3	2e-3
4	1e-3	4	5e-4
5	5e-5	5	1e-4
6	2e-5		
7	1e-5		
8	5e-6		

3.3 Risultati

Il processo di auto-encoding di variabili CG è stato applicato alle molecole di orto-terfenile (OTP), di anilina ed di alanina dipeptide.

Figura 3.3: Loss di ricostruzione di OTP e anilina. a – b Visualizzazione di OTP e anilina, della loro rappresentazione CG e della ricostruzione della configurazione per diverse risoluzioni. c Grafici che mettono in relazione i contributi di ricostruzione (in blu) e regolarizzazione (in verde) della loss function per l' OTP e l'anilina.

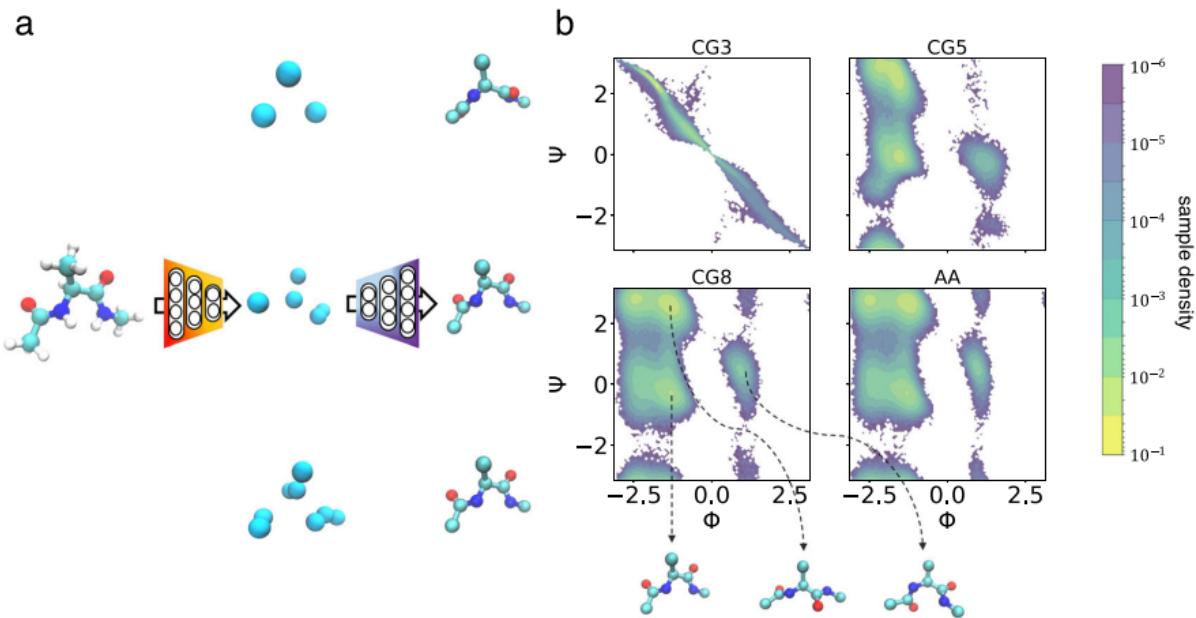


Esaminando figura 3.3 c notiamo come all'aumentare del numero di atomi CG l'errore di ricostruzione della molecola diminuisca. Ad esempio, prendendo in esame la molecola di OTP, figura 3.3 a, costituita da 3 anelli di carbonio, si nota come 2 o 3 variabili CG non riescano a catturare le posizioni relative degli anelli e le loro orientazioni, fatto messo in evidenza dall'elevato valore della loss function. A partire da 4 variabili CG, il numero maggiore di siti consente di catturare più informazioni sulla configurazione e la rappresentazione CG viene decodificata con maggiore accuratezza (valore del termine di ricostruzione più basso). Il contributo di regolarizzazione tende a crescere poiché, all'aumentare del numero di variabili CG, il profilo dell'energia libera diventa sempre più irregolare.

Applicando il modello alla molecola di alanina dipeptide, è stata studiata la capacità da parte dell'AE di catturare caratteristiche critiche delle configurazioni. Ciò è stato verificato ricorrendo alle mappe di Ramachandran ottenute dalla ricostruzione delle variabili CG, figura 3.4 b. Si tratta di grafici che consentono di visualizzare le possibili distribuzioni degli angoli, energeticamente permesse, tra gli amminoacidi residui di una struttura proteica. I maggiori contributi alla costruzione delle mappe sono forniti dagli atomi più pesanti presenti nella molecola (ciò avviene per ragioni legate al momenti angolari e di inerzia molecolari). Nonostante il ridotto numero di gradi di libertà le proprietà fondamentali della molecola

non vengono distorte dalla ricostruzione del sistema ed all'aumentare del numero di siti CG queste vengono riprodotte più fedelmente.

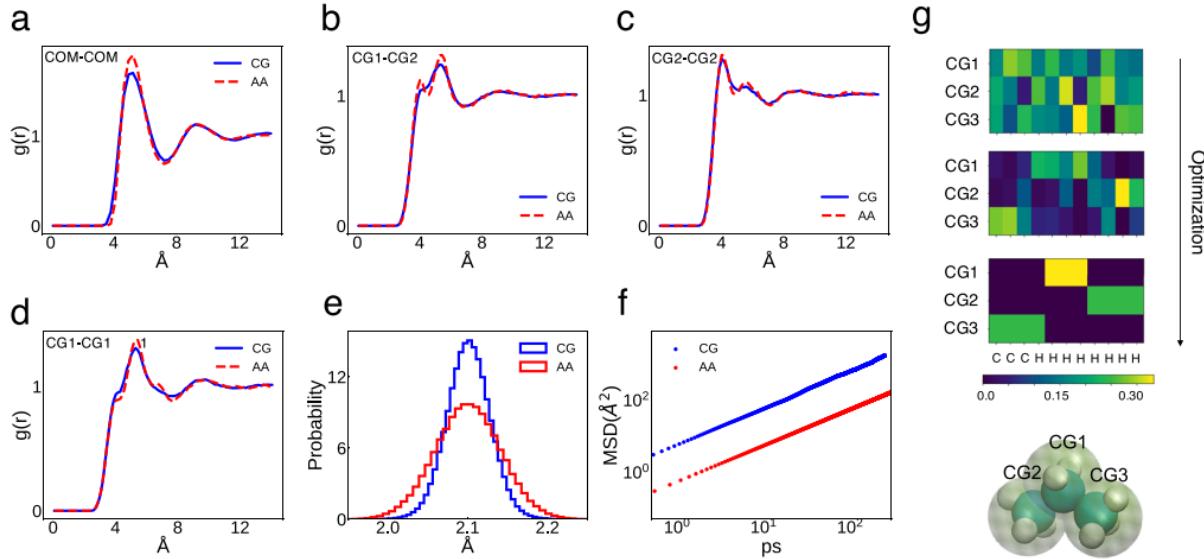
Figura 3.4: Coarse-graining encoding e decoding per l'alanina dipeptide. **a** Visualizzazione dell'alanina dipeptide, della sua rappresentazione CG e della ricostruzione della configurazione per diverse risoluzioni (3, 5 e 8 atomi CG). **b** Comparazione tra le mappe di Ramachandran ottenute dai dati atomistici (AA) e dalle distribuzioni atomiche ricostruite. La rete non è in grado di catturare l'informazione riguardante gli atomi di idrogeno, ma all'aumentare del numero di siti CG riesce a ricostruire fedelmente la posizione degli atomi pesanti e le loro configurazioni.



Infine il framework proposto viene applicato per la simulazione di alcani liquidi a catena corta (C_2H_6 , C_3H_8) e lunga ($C_{24}H_{50}$). Sono state utilizzate rispettivamente 2 e 3 variabili CG per l'etano ed il propano, mentre sono state impiegate due risoluzioni (8 e 12 siti CG) per il $C_{24}H_{50}$. A questo punto le simulazioni CG delle tre molecole in esame sono state eseguite alla stessa temperatura e densità, ed infine sono state estratte le statistiche strutturali relative al liquido in esame.

Le pair-correlation functions $g(r)$ estratte dalle simulazioni atomistiche sono compatibili con quelle calcolate dalle simulazioni CG (figura 3.5 **a – d**) che sfruttano un numero di gradi di libertà inferiore. Queste funzioni descrivono come cambia la densità di un sistema in funzione della distanza da una particella di riferimento. Simili sono le distribuzioni delle lunghezze di legame (figura 3.5 **e**): mentre i dati atomistici risentono della variabilità del sistema, le molecole ricostruite sono ottenute a partire da rappresentazioni CG che vengono calcolate mediando le configurazioni istantanee atomistiche diminuendo la varianza della distribuzione. Calcolando il mean-squared displacement (MSD), ovvero la misura della deviazione della posizione di una particella nel tempo rispetto ad una posizione di riferimento, si nota come esso sia maggiore per le simulazioni CG. Nello spazio CG la perdita di attrito causato dagli atomi genera una dinamica più veloce, provocando un aumento del MSD.

Figura 3.5: Comparazione tra le statistiche della simulazione atomistica e la simulazione CG per il propano liquido con una risoluzione CG di 3 siti per molecola. a – e Grafici delle pair-correlation functions e della distribuzione delle lunghezze di legame delle simulazioni atomistiche e CG. f Comparazione tra il mean-squared displacement (MSD) delle simulazioni atomistiche e CG. g Apprendimento di una mappatura discreta durante l'addestramento: la matrice colorata è una rappresentazione di E_{ij} , dove i colori rappresentano i valori relativi degli elementi della matrice.



3.4 Discussion

Il framework consente di ottenere rappresentazioni CG che catturano abbastanza accuratamente le caratteristiche principali del sistema in esame, sebbene presenti alcune limitazioni. Il modello proposto nell'articolo è deterministico: date in input le coordinate atomistiche, l'output è univocamente determinato dalla matrice dell'encoder e da quella del decoder, e ciò comporta una perdita di informazioni irreversibile. Questo si riflette nella ricostruzione di strutture atomistiche medie anziché di configurazioni istantanee. Un modello probabilistico può imparare una distribuzione di probabilità di ricostruzione che riflette le proprietà termodinamiche del sistema, le quali vengono mediate dal processo di coarse-graining. Inoltre metodi basati sul force-matching per trovare il potenziale CG non garantiscono la possibilità di catturare proprietà di trasporto al non-equilibrio.

Nel 2022 è stato pubblicato un follow-up dell'articolo studiato, in cui viene proposto un modello di AE non deterministico “Variational Auto-Encoders” [14], che permette di imparare le distribuzioni di probabilità alle quali i dati appartengono e potrebbe consentire di catturare i principali aspetti di sistemi che non si trovano in condizioni di equilibrio.

Capitolo 4

Applicazione del Modello AE ai Polimeri ad Anello

Seguendo il lavoro pubblicato su GitHub¹ dagli autori, ho utilizzato l'architettura precedentemente discussa per studiare il comportamento della rete applicata ad un polimero ad anello costituito da 70 monomeri (Figura 4.1). L'AE è stato sviluppato utilizzando il linguaggio di programmazione Python, nello specifico la rete è stata implementata sfruttando il pacchetto Pytorch².

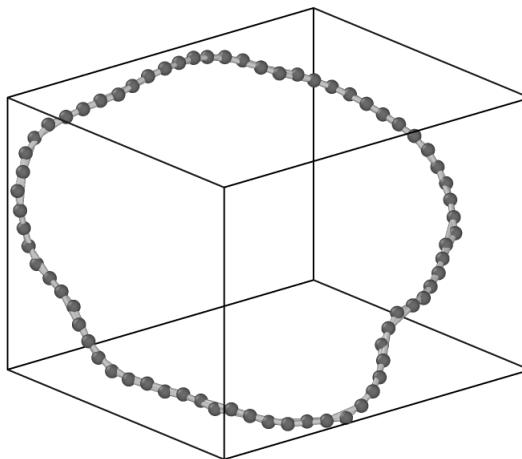


Figura 4.1: Polimero ad Anello

Ho studiato la rappresentazione CG della molecola in esame con 4 risoluzioni diverse (6, 12, 18, 24 siti CG) ed esaminato l'efficacia del processo di apprendimento comparando i valori degli errori di ricostruzione per ciascuna risoluzione.

¹<https://github.com/learningmatter-mit/Coarse-Graining-Auto-encoders>

²<https://pytorch.org/>

4.1 Dati

Le configurazioni dei polimeri ad anello sono state ottenute con il codice di simulazione open-source LAMMPS³, utilizzando il modello coarse-grained di Kremer e Grest [15]. In questo tipo di simulazioni si usano unità ridotte: come unità di energia si è scelto $k_B T = 1$, come unità di lunghezza e di massa si sono scelti il diametro e la massa di un singolo monomero, $\sigma = 1$ e $m = 1$, rispettivamente. Le simulazioni a temperatura costante ($k_B T = 1$), dalle quali sono stati estratti i dati, sono state realizzate usando due diversi termostati, ovvero due algoritmi che permettono di simulare le fluttuazioni statistiche di un sistema a temperatura costante: il termostato di Langevin e il termostato di Nosé-Hoover [16]. Il termostato di Langevin implementa l'equazione di Langevin, imponendo su ogni monomero la forza

$$\vec{F}_L = -\gamma m \vec{v} + \vec{f}(t)$$

dove γ è il coefficiente di frizione e $\vec{f}(t)$ è una forza casuale, che soddisfa il teorema di fluttuazione-dissipazione. Nelle simulazioni effettuate $\gamma = 1$. Il termostato di Nosé-Hoover, al contrario, non aggiunge alcuna forza nel sistema ma utilizza una variabile dinamica ausiliaria, ζ , che regolarizza le fluttuazioni delle velocità nel sistema. E' possibile, partendo fuori dall'equilibrio, che il termostato di Nosé-Hoover raggiunga l'equilibrio solo dopo moltissimo tempo: è stato questo il caso in una delle simulazioni effettuate. Abbiamo comunque testato l'AE per osservare come si comportava con delle configurazioni non all'equilibrio.

Sono stati quindi preparati 3 diversi set di dati da 10100 frame ciascuno, 10000 dei quali sono stati utilizzati per il training della rete mentre 100 per il testing. Il primo set è costituito da configurazioni all'equilibrio generate utilizzando il termostato di Langevin (L_{eq}), il secondo da configurazioni all'equilibrio generate con il termostato di Nosé-Hoover (NH_{eq}) mentre il terzo da configurazioni al non-equilibrio generate con il termostato di Nosé-Hoover (NH_{neq}).

4.2 Training

Per tutti i tre i set di dati l'addestramento della rete è stato effettuato con 600 iterazioni (epoch). Il termine di regolarizzazione è stato introdotto solo a partire dall'epoca numero 300 ed i valori dell'iper-parametro ρ per le diverse risoluzioni sono stati scelti in modo che la rete assegnasse almeno un atomo a ciascun sito CG. I valori di ρ utilizzati sono riportati nella tabella 4.1.

³<https://www.lammps.org>

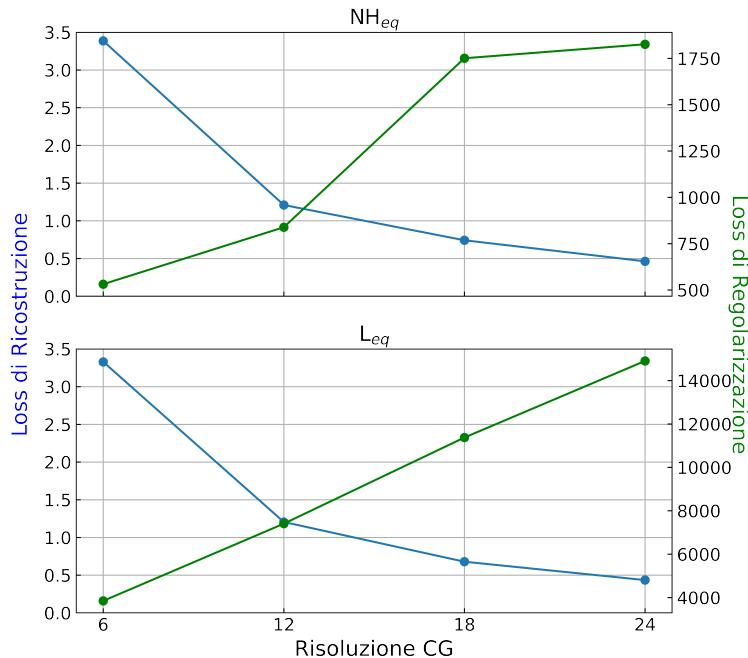
Tabella 4.1: Valori di ρ per i diversi set di dati a diverse risoluzioni CG

Risoluzione CG	L_{eq}	NH_{eq}	NH_{neq}
6	1e-5	1e-5	-
12	7e-6	1e-6	-
18	1e-6	5e-7	-
24	7e-7	1e-7	-

4.3 Risultati

All'aumentare del numero di siti CG utilizzati per ottenere la rappresentazione della molecola si nota un decremento dell'errore di ricostruzione ed un aumento del valore del termine di regolarizzazione. Ciò è coerente con i risultati dell'articolo esposti nel capitolo precedente.

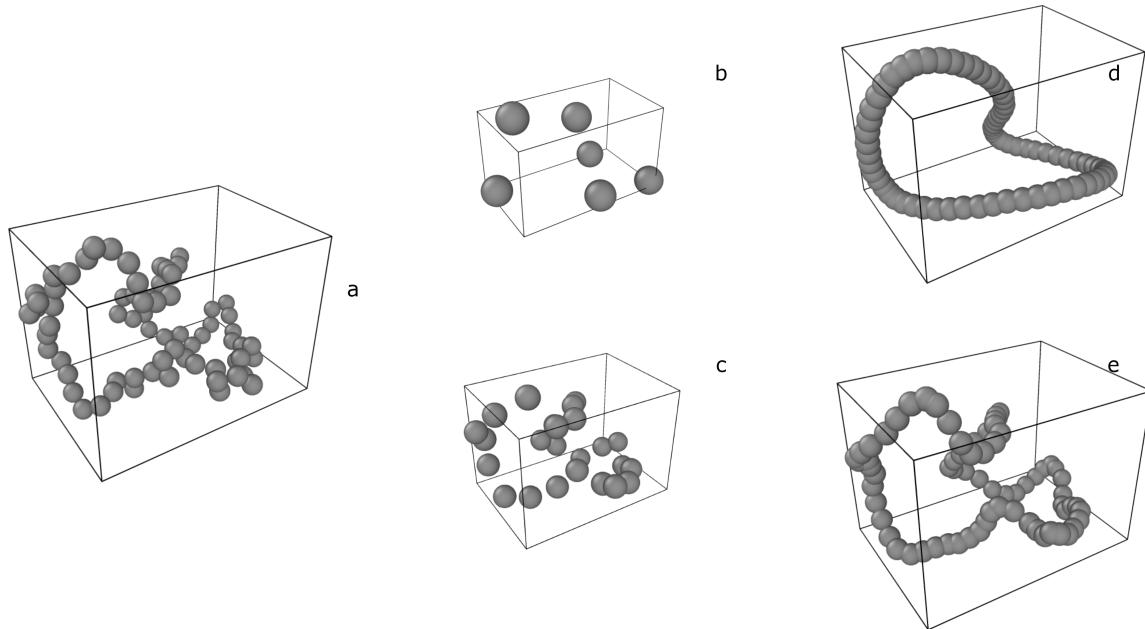
Figura 4.2: Loss di ricostruzione e di regolarizzazione. Il grafico superiore fa riferimento al set di dati NH_{eq} , quello inferiore a quello L_{eq} . In blu si vede l'andamento dell'errore di ricostruzione che diminuisce all'aumentare dei siti CG utilizzati, mentre in verde si nota l'andamento crescente della forza CG.



Il comportamento della rete applicata ai dati L_{eq} e NH_{eq} è il medesimo: utilizzando degli adeguati valori di ρ l'AE sfrutta tutti i gradi di libertà che gli vengono concessi e restituisce valori molto simili dell'errore di ricostruzione, come si può notare dalla figura 4.2. In figura 4.3 si vede come il maggior numero di siti CG consente una ricostruzione più dettagliata della molecola originale. I risultati dell'addestramento sono coerenti con quanto ci si potrebbe

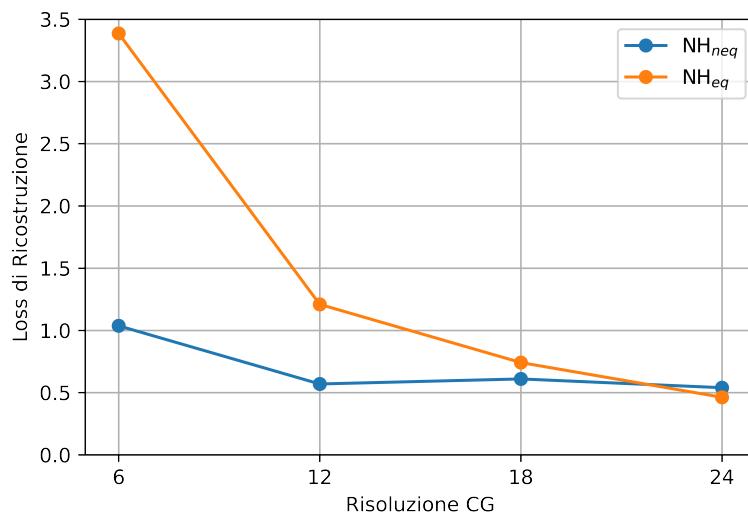
aspettare: sebbene i due set di dati siano stati ottenuti con due diversi termostati allo equilibrio, la Fisica è la stessa e perciò i risultati del training dovrebbero essere simili, come si osserva.

Figura 4.3: Visualizzazione polimero ad anello. **a** Configurazione atomistica data in input alla rete. **b,d** Rappresentazione a 6 siti CG e ricostruzione della molecola in coordinate atomistiche. **c,e** Rappresentazione a 24 siti CG e ricostruzione della molecola in coordinate atomistiche.



Confrontando i valori dell'errore di ricostruzione ottenuti in seguito al training dell'AE con il set di dati NH_{neq} e quelli con NH_{eq} , si nota come i primi siano sensibilmente inferiori per basse risoluzioni CG, figura 4.4. Inoltre, indipendentemente dal valore di ρ , la rete addestrata a partire dai dati NH_{neq} non sfrutta tutti i siti CG che ha a disposizione, bensì ne utilizza al massimo circa 12.

Figura 4.4: Confronto tra le loss di ricostruzione del modello applicato a NH_{eq} e NH_{neq}



4.4 Discussione

Applicato a configurazioni all'equilibrio, il modello riesce ad imparare una rappresentazione del polimero ad anello sfruttando tutti i gradi di libertà che gli vengono concessi. In accordo con i risultati ottenuti dall'articolo precedentemente discusso, l'AE riesce a catturare tante più caratteristiche della molecola quanto maggiore è la risoluzione della rappresentazione, comportando un minore errore di ricostruzione.

La rete non può però essere applicata con successo a configurazioni molecolari che sono fuori dall'equilibrio, poiché, nonostante il basso valore della loss di ricostruzione, non riesce a sfruttare tutti i gradi di libertà che le si forniscono.

In futuro potrebbe essere interessante applicare questo framework a configurazioni molecolari fuori dall'equilibrio e studiare le caratteristiche che la rete riesce a catturare.

Bibliografia

1. Ayton, G. S., Noid, W. G. & Voth, G. A. Multiscale modeling of biomolecular systems: in serial and in parallel. *Current opinion in structural biology* **17**, 192–198 (2007).
2. Wang, W. & Gómez-Bombarelli, R. Coarse-graining auto-encoders for molecular dynamics. *npj Computational Materials* **5**, 1–9 (2019).
3. Snow, C. D., Nguyen, H., Pande, V. S. & Gruebele, M. Absolute comparison of simulated and experimental protein-folding dynamics. *nature* **420**, 102–106 (2002).
4. Onuchic, J. N. & Wolynes, P. G. Theory of protein folding. *Current opinion in structural biology* **14**, 70–75 (2004).
5. Tuckerman, M. E. & Martyna, G. J. *Understanding modern molecular dynamics: Techniques and applications* 2000.
6. Noid, W. G. *et al.* The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *The Journal of chemical physics* **128**, 244114 (2008).
7. Noid, W. G. Perspective: Coarse-grained models for biomolecular systems. *The Journal of chemical physics* **139**, 09B201_1 (2013).
8. Shelley, J. C. *et al.* Simulations of phospholipids using a coarse grain model. *The Journal of Physical Chemistry B* **105**, 9785–9792 (2001).
9. Shinoda, W., DeVane, R. & Klein, M. L. Multi-property fitting and parameterization of a coarse grained model for aqueous surfactants. *Molecular Simulation* **33**, 27–36 (2007).
10. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* <http://www.deeplearningbook.org>, 500 (MIT Press, 2016).
11. Jang, E., Gu, S. & Poole, B. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).
12. Izvekov, S. & Voth, G. A. A multiscale coarse-graining method for biomolecular systems. *The Journal of Physical Chemistry B* **109**, 2469–2473 (2005).
13. Izvekov, S. & Voth, G. A. Multiscale coarse-graining of mixed phospholipid/cholesterol bilayers. *Journal of Chemical Theory and Computation* **2**, 637–648 (2006).

14. Wang, W. *et al.* Generative coarse-graining of molecular conformations. *arXiv preprint arXiv:2201.12176* (2022).
15. Kremer, K. & Grest, G. S. Dynamics of entangled linear polymer melts: A molecular-dynamics simulation. *The Journal of Chemical Physics* **92**, 5057–5086 (1990).
16. Thijssen, J. M. *Computational Physics* (2nd ed.) 226–231. ISBN: 978-0-521-83346-2 (Cambridge University Press, 2007).