Il mio lavoro di tesi riguarda lo studio e l'applicabilità del framework degli autoencoders per ottenere rappresentazioni CG di sistemi atomistici.

Il coarse-graining consiste nel rappresentare gruppi di atomi o intere molecole come siti di interazione CG, ovvero punti dotati di massa e privi di struttura. Questa tecnica consente di ridurre il numero di gradi di libertà necessari per descrivere un sistema atomistico ed un conseguente risparmio computazionale nel momento in cui lo si simula al computer.

E' possibile impiegare tecniche di ML per trovare rappresentazioni CG di questi sistemi. Gli AE sono un modello di NN lineare che impara a copiare in output l'input che riceve. La prima parte della rete è detta encoder, ed impara a proiettare i dati in uno spazio di dimensione inferiore, la quale dipende dalla dimensione del bottleneck. Questo spazio è detto spazio latente. A partire da quest'ultimo il decoder impara a ricostruire i dati partendo dalla rappresentazione ridotta precedentemente ottenuta. I parametri della rete vengono ottimizzati minimizzando una funzione, detta loss funtion, che penalizza l'output per essere dissimile dall'input, un esempio può essere lo MSE.

Nell'articolo del 2019 'CG autoencoders per la dinamica molecolare' gli autori si pongono due obbiettivi il primo, sul quale mi sono concentrato, è di sviluppare un AE in grado di trovare una rappresentazione CG di sistemi atomistici, mentre il secondo è di sviluppare un modello in grado di parametrizzare un potenziale CG a partire dal potenziale atomistico.

Nel modello da loro sviluppato i parametri dell'encoder corrispondono alle probabilità che un particolare atomo venga assegnato ad un sito CG piuttosto che ad un altro. Inizialmente i parametri vengono inizializzati casualmente. Durante il training, per un dato atomo le probabilità di essere assegnato ad i siti CG si azzereranno a meno di una che sarà massima e ne comporterà l'assegnazione al sito corrispondente. La loss function della rete è costituita da due termini, il primo è detto di ricostruzione ed è equivalente all'MSE, il secondo è un termine di regolarizzazione che dipende dalle forze atomistiche e serve a regolarizzare il profilo dell'energia libera. I dati che vengono utilizzati sono frames di traiettorie atomistiche di molecole all'equilibrio che contengono le posizioni di tutti gli atomi.

Gli autori dell’articolo verificano come utilizzando una rappresentazione CG con un maggior numero di siti l’errore di ricostruzione diminuisca. Nel caso dell’ OTP e dell’ anilina lo si può vedere nei grafici a destra: presentano il valore della loss function di ricostruzione, in blu, in funzione del numero di siti usati per la rappresentazione. Il valore del termine di regolarizzazione, in verde, aumenta al crescere della risoluzione CG utilizzata per via dell'aumento delle irregolarità del potenziale.

Similmente sono state studiate le mappe di ramachandran delle configurazioni atomistiche ricostruite a partire da diverse risoluzioni CG della molecola di analina dipeptine. Quello che si nota è come all'aumentare del numero dei siti le mappe diventino sempre più simili a quella generata dal modello atomistico originale.

Ho applicato l'architettura proposta nell'articolo precedentemente esposto a configurazioni atomistiche di polimeri ad anello costituiti da 70 monomeri. I dati sono stati generati a partire da simulazioni di dinamica molecolare mantenute a temperatura costante. Per fare ciò sono state impiegate due diverse tecniche: la prima utilizza il termostato di Langevin, per il quale ad ogni monomero viene imposta una forza casuale, con cui sono state generate 10000 configurazioni all'equilibrio, mentre la seconda utilizza il termostato di nose-hoover, che sfrutta una variabile dinamica aggiuntiva per regolarizzare le fluttuazioni del sistema, con la quale sono state generate 10000 configurazioni all'equilibrio ed altre 10000 fuori dall'equilibrio.

Le gif a destra presentano la visualizzazione del processo di training. Sull’asse x si trovano gli atomi della molecola mentre sull’asse y i siti CG ai quali possono essere assegnati. Le matrici colorate rappresentano i valori relativi dei parametri dell’encoder della rete. In alto vediamo il training utilizzando NH all'equilibrio e in basso quello al non equilibrio.

In accordo con i risultati dell'articolo precedentemente esposto possiamo vedere dalla visualizzazione della molecola come un maggior numero di siti CG riesce a catturare maggiori caratteristiche del polimero, in particolare nell'immagine è riportato il caso di una rappresentazione che sfrutta 6 e 24 siti CG.

Nei grafici a destra è messo a confronto l'andamento delle loss function di ricostruzione dei set di dati all’equilibrio, in blu, che decresce all'aumentare della risoluzione impiegata. Il training che utilizza questi set riporta andamenti praticamente equivalenti. Confrontando i valori delle loss function ottenuti a partire dai set all'equilibrio e non, si nota come la loss di ricostruzione sia sensibilmente inferiore per il set al non equilibrio per basse risoluzioni, inoltre la rete addestrata a partire da questo set non riesce a sfruttare tutti i gradi di libertà che le vengono concessi, non assegnando atomi ad alcuni siti CG.

Concludendo, il framework può essere applicato ai polimeri ad anello all'equilibrio con successo, ottenendo delle rappresentazioni CG della molecola, Purtroppo se si utilizzano configurazioni fuori dall'equilibrio l'AE non riesce a sfruttare tutti i siti CG che gli vengono concessi, indipendentemente dai valori del termine di regolarizzazione.