UNIVERSITÀ DEGLI STUDI DI MILANO – BICOCCA

Scuola di Scienze

Corso di Laurea in Scienze e Tecnologie Chimiche

ANALISI DEI RISULTATI DI SIMULAZIONI MOLECOLARI DEL PROCESSO DI ADSORBIMENTO DI ACQUA SU MODELLI DI PARTICOLATO ATMOSFERICO

Relatore:

Prof. Claudio GRECO

Correlatore:

Prof. Ugo Renato COSENTINO

Tesi di Laurea di:

Giorgio CARBONE

Matricola n. 811974

Anno accademico 2019/2020

Indice

Introduzione

Scopo del tirocinio è stato quello di studiare il processo di adsorbimento di molecole di acqua su superfici modello di particolato atmosferico, attraverso l'analisi di risultati di simulazioni computazionali effettuate in precedenti lavori di Tesi.

In particolare, è stato considerato un sistema modello di una superficie di cloruro di sodio, componente principale dei *sea-salt aerosol (SSA)*, di origine marina.

Dal punto di vista ambientale il processo studiato assume grande rilevanza: il grado di idratazione dell'aerosol atmosferico risulta infatti strettamente correlato alle proprietà catalitiche, ottiche e igroscopiche dello stesso, e quindi all'effetto diretto provocato dal particolato sul clima e sulla chimica della troposfera.

La tecnica utilizzata nelle simulazioni considerate è il metodo Monte Carlo nell'insieme termodinamico Gran Canonico (μVT), che prevede potenziale chimico, volume e temperatura constanti. Le simulazioni sono state condotte alla temperatura di 298K, in un intervallo di pressione di H_2O tra 7.500 e 9.000 matm, campionato con un passo di 0.125 milliatmosfere. Il potenziale utilizzato è classico (*force field* di tipo AMBER per il cloruro di sodio; e modello SPC/E per rappresentare le molecole di acqua).

Nel corso del tirocinio è stato sviluppato uno *script*, scritto nel linguaggio di programmazione Python, in grado di effettuare una *data analysis* automatizzata, approfondita ed efficiente delle configurazioni generate durante ogni simulazione, condotta ad uno specifico valore di pressione di H_2O.

Principalmente è stata eseguita una *cluster analylis* dei risultati, con lo scopo di studiare i fenomeni di tipo aggregativo che coinvolgono le molecole d'acqua adsorbite sulla superficie, le quali, interagendo tramite legame ad idrogeno, formano spontaneamente "isole" oppure "strati" sovrapposti, in funzione del grado di copertura della superficie del sale, come emerso dai risultati sperimentali degli studi

di Foster e Ewing[1] e da quelli delle simulazioni computazionali effettuate da Engkvist e Stone[2].

L'algoritmo implementato per il *clustering* è DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*), proposto da Martin Ester *et al.*[3], il quale permette l'individuazione di aggregati tramite la separazione di aree dello spazio tridimensionale ad alta densità di molecole d'acqua, rispetto a zone a bassa densità.

I *cluster* individuati sono stati poi classificati in "isole" o "strati" e sono ne sono state studiate le diverse proprietà.

Il primo capitolo di questo elaborato contiene un'introduzione sulla tematica degli aerosol atmosferici, facendo riferimento ai risultati di precedenti studi, sperimentali e computazionali, descritti in letteratura, nonché una descrizione del metodo e del sistema modello alla base delle simulazioni di cui sono stati analizzati i risultati. Il secondo capitolo si concentra sulla descrizione della *data analysis* eseguita sui risultati delle simulazioni. Nella prima parte vengono riportati i metodi di analisi dati utilizzati durante il tirocinio, vengono esposti i principali algoritmi di *clustering* e in particolare viene descritto il funzionamento, e i parametri caratteristici, dell'algoritmo DBSCAN. Nella seconda sezione viene delineata l'implementazione dell'algoritmo di clusterizzazione nello *script* sviluppato e la sua applicazione al sistema modello. Sono inoltre descritti gli ulteriori algoritmi di analisi sviluppati, e implementati, al fine di elaborare i risultati della *cluster analysis* e ricavarne dati di interesse chimico.

Nel terzo capitolo vengono riportati e discussi i risultati della *data analysis.* Infine, le discussioni riassumono i principali risultati ottenuti e le considerazioni riguardanti il ruolo dell'analisi dati nello studio della grande mole di dati ricavata dalle simulazioni computazionali.

[1] (Foster & Ewing, 2000)
[2] (Engkvist & Stone, Adsorption of water on the NaCl (001) surface. III. Monte Carlo simulations at ambient temperatures, 2000)
[3] (Ester, Kriegel, Sander, & Xu, 1996)

Capitolo 1

1. Processi di adsorbimento di acqua su aerosol atmosferici

1.1 Aerosol atmosferici

Gli aerosol atmosferici sono definiti come sistemi colloidali, nei quali la fase disperdente è gassosa, mentre la fase dispersa, sospesa nell'aria, è costituita da particelle liquide o solide con un diametro compreso nell'intervallo 0.001-10 μm.
Le particelle che compongono gli aerosol possono variare molto in termini di dimensione, forma, distribuzione nello spazio e nel tempo e in base al tempo di vita medio nell'atmosfera.
Le caratteristiche degli aerosol risultano fortemente connesse alla fonte da cui si originano, la quale può essere antropogenica, come nel caso degli aerosol carboniosi, derivanti dalla combustione delle biomasse, o naturale, come per i *sea-salt aerosol (SSA),* di origine marina e prevalentemente costituiti da cloruro di sodio.
Ogni fonte sarà correlata ad uno specifico processo di formazione, in funzione delle fasi in cui il processo avviene. Possiamo distinguere: aerosol atmosferici primari, immessi direttamente nell'atmosfera, e secondari, generati in situ, tramite reazioni di nucleazione (conversione gas-particella). Una volta formati gli aerosol possono accrescere ulteriormente le proprie dimensioni per condensazione dei gas o attraverso lo scontro con altre particelle solide (coagulazione).[4]
Le fonti di aerosol atmosferico più comuni sono state individuate, nel 2007, dall' Intergovernmental Panel for Climate Change[5]:

- polvere proveniente dal suolo: componente principale dell'aerosol soprattutto nelle regioni tropicali e subtropicali. Le *soil dust* provengono principalmente da deserti, letti di laghi prosciugati e zone

[4] (Sokolik, 2002)
[5] (Forster, et al., 2007)

secche, dove la vegetazione è stata rimossa o il terreno è stato manipolato dall'uomo;

- sale marino: i *sea-salt aerosol (SSA),* di origine marina, sono generati da diversi processi fisici, in particolare la rottura delle creste delle onde. La composizione ionica dei *SSA* è dominata da Cl^-(55.04% w/w) e Na+ (30.61% w/w)[6];
- polveri industriali: costituiscono la fonte antropogenica primaria di particolato atmosferico, sono associate ad attività industriali e tecniche come l'incenerimento dei rifiuti, trasporti, metallurgia, manifattura del cemento e così via;
- aerosol carboniosi: compongono una larga, ma altamente variabile, porzione degli aerosol atmosferici. Sono ottenuti principalmente dalla combustione di biomasse e di combustibili fossili;
- aerosol biogenici primari: consistono in detriti di piante, come frammenti di foglie, e particelle di origine microbica (batteri, funghi, virus, alghe, spore, polline, etc.). I bioaerosol fungono da *cloud condensation nuclei* in certe regioni del mondo[7];
- aerosol solfati: aerosol secondari, originano da reazioni nell'atmosfera a partire da precursori gassosi;
- aerosol nitrati: la presenza di questi aerosol risulta strettamente correlata all'abbondanza relativa di solfati e sali di ammonio;
- aerosol di origine vulcanica: le polveri primarie e i solfuri gassosi sono le componenti delle emissioni vulcaniche maggiormente correlate agli aerosol.

[6] (Einfeld & Pandis, 1998)
[7] (Christner, Morris, Foreman, Cai, & Sands, 2008)

Gli aerosol atmosferici assumono grande rilevanza, dal punto di vista ambientale, a causa degli effetti diretti provocati sulle proprietà chimiche dell'atmosfera e sul clima.

Essi presentano proprietà ottiche di *scattering* e assorbimento della radiazione solare, con conseguenze sulla visibilità e sulla temperatura dell'atmosfera. A causa della loro complessità e varietà risulta difficile studiare l'effetto complessivo esercitato dagli aerosol sulla luce solare in arrivo sulla terra, ma in termini generali si può dire che particelle colorate in modo acceso o traslucide fungano da centri di diffusione, tendendo quindi a rifrangere la radiazione incidente in tutte le direzioni e verso lo spazio (gli aerosol mandano nello spazio circa un quarto della radiazione che arriva alla terra dal sole), mentre aerosol più scuri tendono ad assorbire la luce che li colpisce. Il fenomeno di assorbimento della luce da parte degli aerosol più scuri (formati, ad esempio, da particelle di *black carbon*[8]) porta ad un surriscaldamento locale dello strato dell'atmosfera in cui sono contenute le particelle, provocando però un ombreggiamento e un raffreddamento netto della superficie sottostante. Inoltre, gli aerosol che si depositano sulla superficie della terra ne incrementano l'albedo, aumentando la frazione di radiazione solare riflessa nello spazio.

Gli aerosol atmosferici presentano inoltre proprietà igroscopiche, sono infatti costituiti da un nucleo organico o inorganico, il quale funge da nucleo di condensazione per le molecole organiche e inorganiche presenti nell'atmosfera, e in particolare per l'acqua, coadiuvando la formazione di nebbie e nubi.

Nuvole che contengono molti aerosol riflettivi appaiono bianche e sono caratterizzate dalla presenza di un numero maggiore di particelle di acqua di piccole dimensioni, che portano a un aumento della diffusione e della riflessione della luce.

L'acqua adsorbita influenza le caratteristiche del particolato in vari modi, determinando: la natura dell'aerosol (solido, liquido, o vetroso/amorfo), le sue dimensioni e la crescita igroscopica, la capacità di adsorbimento di gas e ossidanti e influenzando le reazioni multi-fase che si verificano nelle goccioline di aerosol.

[8] (Ramanathan, et al., 2007)

Infatti, come accennato precedentemente, la presenza degli aerosol determina modifiche nella chimica dell'atmosfera, questo effetto si manifesta soprattutto come interazione tra i gas atmosferici (in particolare vapore d'acqua) e gli aerosol stessi. La fase condensata, sotto forma di *bulk* water o di sottili film di acqua (spessore compreso fra 1nm e i 50μm): assicura un'elevata area superficiale, permettendo l'assorbimento o l'adsorbimento di specie chimiche presenti nell'atmosfera, come composti organici a bassa solubilità e bassa pressione di vapore[9]; fornisce siti attivi ad alta attività catalitica implicati in numerosi processi di catalisi eterogenea che coinvolgono gas (inorganici e organici) e ossidanti atmosferici (ozono, radicale idrossile, ossigeno singoletto, radicale nitrato, cloro, …); funge da mediatore per la deposizione di inquinanti sulla terra e sui corpi d'acqua e da solvente per reazioni di conversione di gas in aerosol secondari. Nel caso in cui il contenuto di acqua fosse inferiore a quello richiesto per ottenere la formazione di un *monolayer*, quindi di uno strato monomolecolare di acqua sulla superficie, questa agisce invece da adsorbato competitivo per i siti di adsorbimento presenti sulla superficie.

Un esempio dell'azione catalitica dell'acqua adsorbita sulla superficie è la reazione di ossidazione in fase acquosa di SO_2 a dare ione solfato, responsabile della dell'acidificazione delle piogge. L'anidride solforosa presente nell'atmosfera deriva da emissioni prevalentemente antropogeniche, parte di essa viene rimossa dall'atmosfera per deposizione secca o deposizione umida, mentre una certa percentuale (~40%) viene ossidata attraverso reazioni in fase gassosa o acquosa. È generalmente accettato che l'ossidazione in fase acquosa, ad opera principalmente di O_3, H_2O_2 (reazione riportata come esempio) e O_2 + TMI (*Transition Metal Ion*), domini sulla reazione in fase gassosa, ad opera del radicale idrossile (·OH)[10]:

$$SO_{2(ads)} + H_2O_{(l)} \rightleftharpoons H^+_{(aq)} + HSO^-_{3(aq)}$$

$$HSO^-_{3(aq)} + H_2O_{2(aq)} + H^+_{(aq)} \rightleftharpoons {SO_4}^{2-}{}_{(aq)} + H_2O_{(l)} + 2H^+_{(aq)}$$

[9] (Valsaraj, Ehrenhauser, Heath, & Vaitilingom, 2015)
[10] (Yang, et al., 2018)

Molte proprietà (sia chimico-fisiche sia ottiche) e molti processi (reazioni multifase, corrosione, ecc.) sono influenzati dalla fase (solida o liquida) in cui le particelle esistono nell'atmosfera, e quindi dal grado di idratazione dell'aerosol.

Il livello di idratazione dell'aerosol dipende dall'umidità relativa (RH, Relative Humidity: definita come il rapporto fra la pressione del vapore acqueo presente nell'atmosfera rispetto alla pressione di saturazione dell'acqua alla temperatura considerata) alla quale la particella di aerosol è esposta nell'atmosfera.

L'esposizione di aerosol al processo di umidificazione porta inizialmente all'adsorbimento di acqua sulla superficie del particolato. Una diminuzione dell'umidità relativa risulta nell'evaporazione delle molecole di acqua adsorbite, raggiunto il valore minimo di soglia, definito come ERH (*Efflorescence Relative Humidity*) o CRH (*Crystallization Relative Humidity)*, si ha efflorescenza, quindi transizione di fase con passaggio allo stato solido. Nel processo opposto, invece, si osserva il fenomeno della deliquescenza, ovvero la dissoluzione del solido con formazione di una soluzione satura. La transizione di fase avviene quando l'RH supera il caratteristico valore di soglia, definito come DRH (*Deliquescence Relative Humidity)*, se l'RH incrementa ulteriormente si ha crescita igroscopica del particolato per condensazione dell'acqua.

Dunque, è comunemente noto che, per un particolare valore di umidità relativa, una particella possa esistere contemporaneamente come solido e come gocciolina di soluzione[11].

Analizzando le dimensioni dell'aerosol, risulta che ad alti livelli di umidità relativa le particelle acquistano dimensioni significative a causa del loro contenuto di acqua; ciò implica che diffondono la luce di lunghezza d'onda visibile molto più efficacemente, quindi la crescita igroscopica delle particelle atmosferiche è responsabile anche della ridotta visibilità associata allo smog.

[11] (Martin, 2000)

1.2 Studi sperimentali e computazionali dell'adsorbimento di acqua su NaCl

Il processo di adsorbimento di acqua su una superficie di cloruro di sodio è stato largamente studiato, tuttavia sono stati pochi gli studi intrapresi considerando temperature rilevanti dal punto di vista ambientale e che quindi permettessero di rivelare, in tali condizioni, la struttura molecolare del sottile film adsorbito. Infatti, considerando temperature comprese tra i 20 e i 30°C e pressioni del vapore acqueo comprese tra i 5 e i 20 mbar, il tempo di vita di una molecola di acqua adsorbita è dell'ordine dei microsecondi e la struttura mobile dello strato di adsorbimento si presenta come un insieme di molte configurazioni irregolari[12].

Ad esempio, uno strato monomolecolare di H_2O (*monolayer)* su NaCl (001) è stato studiato, intorno alla temperatura criogenica di 100K, tramite dispersione dell'atomo di elio (*Helium Atom Scattering*, HAS) e diffrazione di elettroni a bassa energia (*Low-Energy Electron Diffraction, LEED)*[13]*;* a queste temperature le molecole di acqua sono fissate in configurazioni di minima energia per intervalli di tempo nell'ordine delle ore, rendendo possibile l'identificazione di diverse strutture monostrato possibili per i film d'acqua.

Gli studi sperimentali a temperatura ambiente sono stati compiuti sfruttando la spettroscopia infrarossa a trasformata di Fourier (FTIR)[14] e gli spettri ottenuti hanno permesso di ricavare e illustrare le possibili strutture degli strati d'acqua prima della dissoluzione al punto di deliquescenza. Lo studio del processo di adsorbimento mediante spettroscopia FTIR ha permesso di ottenere risultati molto utili a causa della sensibilità nei confronti dei legami a idrogeno da parte della banda vibrazionale corrispondente allo stretching di O-H.

Il fisiadsorbimento delle molecole d'acqua è possibile grazie ai campi elettrici, con tipico raggio molecolare, che caratterizzano la struttura quadrata degli ioni Na^+ e Cl^-

[12] (Engkvist & Stone, 2000)
[13] (Foelsch & Henzler, 1995)
[14] (Foster & Ewing, 2000)

presenti sulla superficie. Le molecole d'acqua adsorbite si aggregano formando *cluster* sulla superficie (isole) o *monolayer*, instaurando legami a idrogeno laterali, oppure, nel caso siano presenti più strati (*multilayer)*, questi risultano legati tra di loro tramite legami ad idrogeno, avendo quindi un passaggio da legami ad idrogeno laterali ad isotropici, simili a quelli presenti nell'acqua liquida. La formazione di aggregati, anche in condizioni di bassa copertura della superficie, risulta favorita perché permette una minimizzazione dell'energia, correlata ad una stabilizzazione dello strato adsorbito, dovuta alla presenza dei legami a idrogeno, migliore rispetto a quella ottenuta dalla sola interazione con gli ioni Na^+ e Cl^- della superficie in una molecola di H_2O isolata[15].

Osservando lo spettro di assorbimento dell'acqua gassosa (Figura 1) è possibile individuare tre bande caratteristiche: una a 3756 cm^{-1}, che corrisponde allo stretching asimmetrico; una a 3652 cm^{-1}, che corrisponde allo stretching simmetrico; una a 1545 cm^{-1}, associata al movimento di bending della molecola.

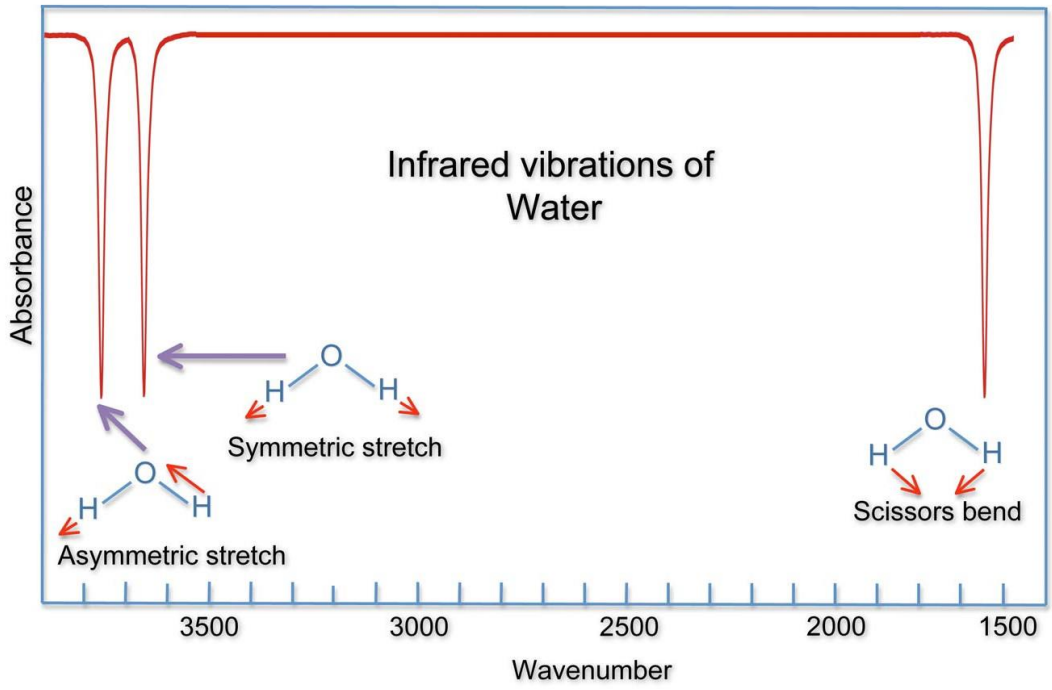

Figura 1: Spettro di assorbimento IR dell'acqua in fase gas.

[15] (Engkvist & Stone, 1999)

Se l'acqua gassosa condensa, venendo adsorbita sulla superficie di NaCl, lo spettro di assorbimento si riduce ad un'unica banda vibrazionale di stretching (Figura 2).

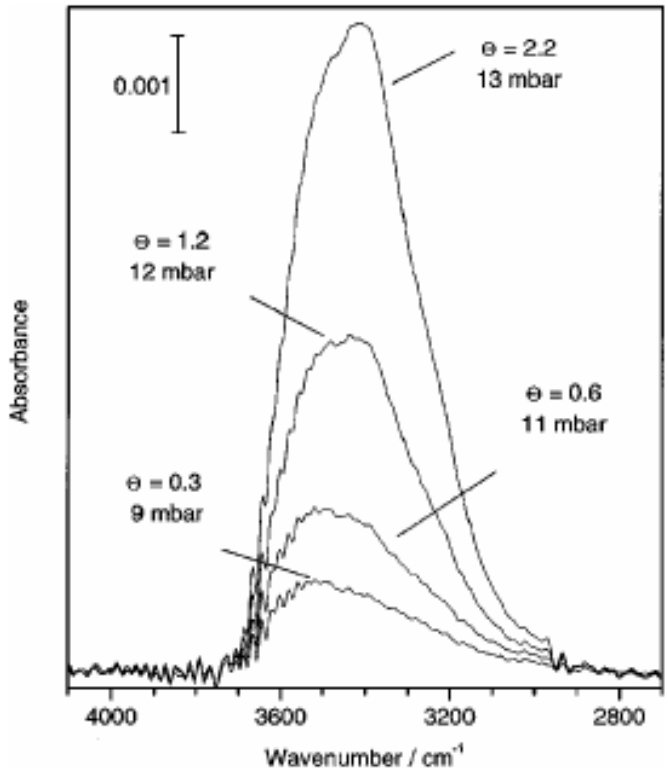

Figura 2: Spettro di assorbimento IR dell'acqua adsorbita su NaCl(001) a 24 °C per diversi valori di pressione e coverage.[16]

Alla pressione di 9 mbar, valore minore campionato, la banda di assorbimento è asimmetrica e il massimo è posizionato a 3515 cm^{-1}; tende a diventare più simmetrica al crescere della pressione e si sposta verso numeri d'onda minori, stabilizzandosi a 3415 cm^{-1} ad una pressione di circa 13 mbar (Figura 2). L'assorbanza non cresce in modo lineare con la pressione, un incremento di pressione da 9 a 11 mbar risulta in un valore di assorbanza due volte maggiore rispetto a quello iniziale, invece passando da 11 a 13 mbar il valore viene quadruplicato.

Il valore integrato dell'assorbanza ($\tilde{A}$), risulta correlabile quantitativamente al grado di copertura dello strato di adsorbimento (*coverage* θ), calcolabile come rapporto tra

[16] (Foster & Ewing, 2000)

la densità di superficie del cristallo di NaCl e la densità di superficie dell'acqua adsorbita (Equazione 1).

$$\Theta = \frac{S_{H_2O}}{S_{NaCl}}$$

Equazione 1

La densità di superficie dell'acqua adsorbita è ricavabile da una versione modificata della legge di Lambert-Beer (Equazione 2).

$$\tilde{A} = \frac{N\bar{\sigma}S_{H_2O}}{S_{NaCl}}$$

Equazione 2[17]: *N* è il numero di superfici di NaCl(001) e $\tilde{A}$ è l'assorbanza integrata per un range di frequenze uguale a quello usato per ottenere $\bar{\sigma}$.

La possibilità di stimare il valore di *coverage* a partire da misure spettroscopiche di assorbanza hanno permesso a Foster e Ewing di costruire la curva isoterma ad una temperatura di 24° C, che mostra la dipendenza del valore di *coverage* dalla pressione di acqua in mbar (Figura 4)[18].

.

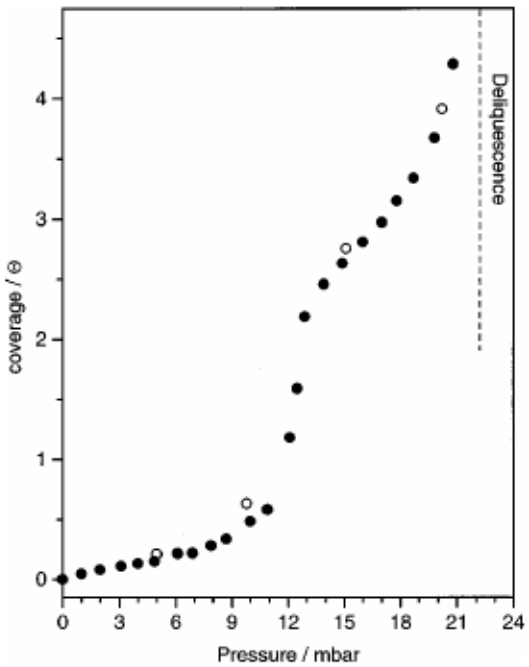

Figura 3: Isoterma di adsorbimento di acqua su NaCl a 24 °C. I dati relativi a pressioni crescenti e decrescenti sono rispettivamente rappresentati da cerchi pieni e vuoti.

[17] (Richardson, Chang, Noda, & Ewing, 1989)
[18] (Foster & Ewing, 2000)

Nella curva sono presenti quattro regioni distinte:

1) La regione a basso *coverage* ($\theta \leq 0.5$) è caratterizzata da una crescita lineare leggera del grado di copertura e da una concavità verso la fine. Per valori di *coverage* estremamente bassi ($\theta << 0.1$), viene ipotizzata la presenza di una banda di assorbimento centrata a 3700 cm^{-1}, in corrispondenza delle due frequenze di stretching del vapore acqueo, associata alla presenza di molecole d'acqua adsorbite su NaCl e isolate, quindi non coinvolte in legami a idrogeno. Da $\theta = 0.1$ a $\theta = 0.5$ diventano rilevanti le interazioni intermolecolari, si ha assorbimento IR a 3500 cm^{-1}, una frequenza maggiore rispetto all'acqua liquida, manifestazione di uno stato intermedio tra l'acqua liquida pura e il vapore acqueo.

   Lo stato del sistema viene detto di *submonolayer*, le molecole di acqua adsorbite non formano uno strato completo ma si aggregano a formare isole sulla superficie, caratterizzate da legami a idrogeno laterali. La stabilizzazione di queste strutture risulta ancora principalmente dovuta all'interazione dell'acqua con il substrato solido.

2) Per valori di *coverage* compresi tra 0.5 e 2.5 si ha la regione di transizione, θ cresce molto (in modo quasi verticale) in funzione di piccoli incrementi di pressione ed è presente un punto di flesso in corrispondenza di $\theta = 1.5$.

   L'assorbimento IR avviene ad un valore di frequenza comunque maggiore rispetto all'acqua liquida, la regione rappresenta infatti una zona di passaggio da legami a idrogeno di tipo laterale, tipici delle isole, a legami isotropici, paragonabili a quelli dell'acqua liquida e riscontrabili nella terza regione della curva.

   Per le strutture della zona di transizione e, di conseguenza, anche per quelle della regione superiore, i legami a idrogeno tra molecole d'acqua vicine rappresentano il principale contributo alla stabilizzazione, mentre le

interazioni con il substrato sottostante divengono secondarie (situazione opposta rispetto allo stato di basso coverage).

Per la regione di transizione viene proposta la possibile coesistenza di isole e strutture multistrato (*multilayer)* di molecole d'acqua adsorbite sulle superficie del sale.

3) La terza regione, con $2.5 \leq \theta \leq 3.5$ è detta regione ad alta copertura ed è caratterizzata da un aumento meno ripido del *coverage* in funzione della pressione. In questa regione le molecole di acqua si dispongono in sistemi multistrato, instaurando legami a idrogeno di natura isotropica, simili a quelli dell'acqua liquida. L'ipotesi di una struttura *multilayer* è stata proposta dagli autori perché la frequenza di vibrazione, a tali pressioni di vapore acqueo, coincide con quella dell'acqua liquida. In corrispondenza di alti valori di coverage, le molecole d'acqua del primo strato di adsorbimento sono legate sia al substrato sia a quelle soprastanti. Data l'alta concentrazione d'acqua, ogni molecola del primo strato è notevolmente più influenzata dalle molecole d'acqua superiori rispetto a quelle ai suoi lati.
4) A $\theta = 3.5$ c'è un ulteriore punto di flesso e la regione che corrisponde a $\theta > 3.5$ viene chiamata regione di presoluzione, in quanto il sistema tende alla deliquescenza. Un significativo numero di ioni viene incorporato nel film di acqua, passando in soluzione, prima che abbia luogo la dissoluzione completa del sale.

Risulta chiaro, analizzando le isoterme e le strutture identificate alle varie pressioni, quanto il processo di adsorbimento dell'acqua su NaCl sia influenzato dai legami a idrogeno, i quali assumono maggiore rilevanza al crescere del grado di copertura. La curva isoterma, infatti, non ha nessuna somiglianza con quella del modello di Langmuir, correlata ad un adsorbimento randomico sui siti disponibili sulla superficie (Figura 4).

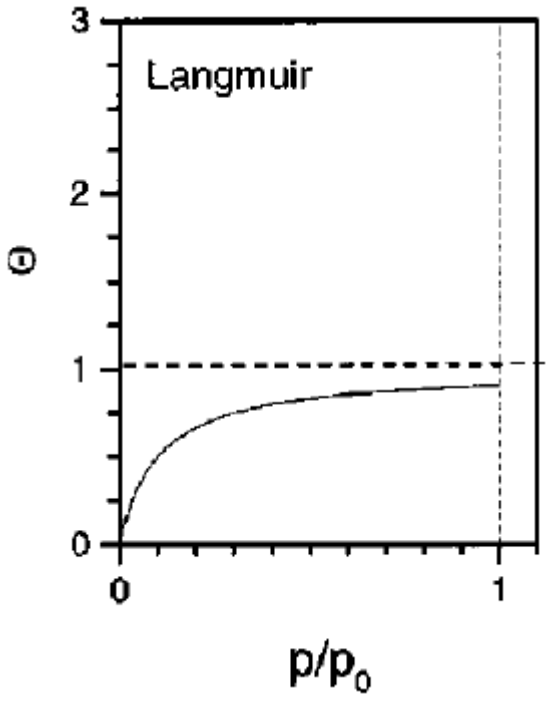

Figura 4:Isoterma di adsorbimento dal modello di Langmuir[19], *p/p₀* è la pressione ridotta, dove *p₀* è la pressione di equilibrio della fase *bulk* dell'adsorbente.

Secondo il modello, quando tutti i siti di adsorbimento vengono occupati, la superficie risulta saturata da un *monolayer* e non avviene ulteriore adsorbimento. Le interazioni laterali vengono ignorate e viene esclusa la formazione di un *multilayer.* Ad esempio, il fisiadsorbimento di CO su NaCl(001) a 55 K risulta accuratamente descritto dall'isoterma di Langmuir, la quale prevede una curva convessa, in netto contrasto con la concavità presente nella zona finale della regione a basso *coverage* dell'isoterma di adsorbimento di acqua su NaCl(001) a temperatura ambiente. A 55 K, si ottiene θ = 0.5 a $\frac{p}{po} = 2 \times 10^{-8}$, il *monolayer* si forma a $\frac{p}{po} = 10^{-7}$. Per incrementi ulteriori di *p/p₀* di CO, anche di diversi ordini di grandezza, non si ha alterazione della struttura adsorbita. Anche quando si forma una struttura *multilayer* cristallina di CO, a $\frac{p}{po} = 1$, la struttura monostrato originaria non subisce alterazioni apprezzabili.

[19] (Brunauer, Deming, Deming, & Teller, 1949)

Ola Engkvist e Anthony J. Stone[20] hanno condotto delle simulazioni Monte Carlo a temperatura ambiente per comprovare gli studi delle analisi sperimentali di Foster e Ewing, permettendo di comprendere più approfonditamente la struttura dei film molecolari dell'acqua di adsorbimento e, contemporaneamente, i risultati ottenuti sono stati utilizzati per interpretarne gli spettri IR.

Calcoli quantistici, condotti alla temperatura dello zero assoluto[21], hanno portato alla conclusione che, per la minimizzazione dell'energia, è favorita la formazione di cluster, i quali coinvolgono legami idrogeno che favoriscono la stabilizzazione del sistema.

Lo studio computazionale è stato effettuato mediante un approccio basato sullo sviluppo di potenziali intermolecolari (acqua-acqua e acqua substrato) molto accurati, derivati da calcoli quantistici. Per l'interazione acqua-acqua gli autori hanno sviluppato una versione migliorata del potenziale ASP-W4, che prevede un modello multipolare dell'acqua, molto accurato ma dispendioso dal punto di vista delle risorse di calcolo utilizzate per le simulazioni.

Il modello di acqua ASP-W4 è stato utilizzato nel contesto di simulazioni Monte Carlo nell'insieme NVT (cioè con numero di particelle N, volume V e temperatura T costanti), essendo l'approccio Gran Canonico (ensemble μVT) troppo oneroso da implementare in questo caso. La cella considerata è periodica lungo due dimensioni e la superficie di NaCl è rigida, dal momento che i valori dei gradi di copertura scelti precedono il punto di deliquescenza.

A causa dell'onerosità dei calcoli gli autori hanno scelto due soli specifici valori di coverage (θ = 0.5 e θ = 3.0) a cui studiare il sistema, rappresentativi rispettivamente delle regioni a basso e alto *coverage*, in modo da evidenziare le due diverse fasi suggerite dall'isoterma sperimentale.

È stata rilevata una differenza nei fenomeni aggregativi che interessano le molecole d'acqua adsorbite sulle superficie ai due valori di coverage studiati:

[20] (Engkvist & Stone, 2000)
[21] (Engkvist & Stone, 1999)

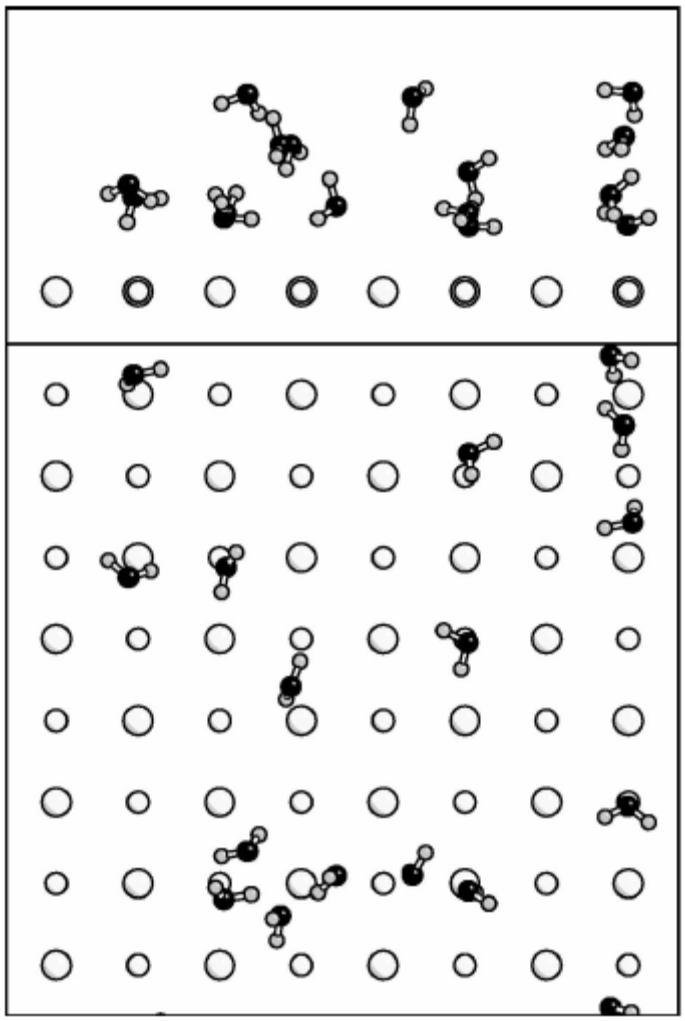

Figura 5: visione laterale (in alto) e ortogonale (in basso) di una tipica configurazione di molecole di acqua adsorbite su una superficie di NaCl a *coverage* θ = 0.5.

Lo *snapshot* della simulazione in Figura 5 rappresenta quella che è la tipica configurazione delle molecole d'acqua adsorbite sulla superficie di NaCl nella regione a basso *coverage*. La caratteristica principale che emerge dalle simulazioni a bassa copertura è la tendenza spontanea delle molecole d'acqua ad aggregarsi in isole, il fenomeno accade in ognuna delle simulazioni eseguite, partendo da una distribuzione uniforme delle molecole sulla superficie. Le molecole d'acqua sono libere di muoversi sulla superficie, le isole quindi hanno natura dinamica, le dimensioni delle stesse diminuiscono e aumentano nel corso della simulazione.

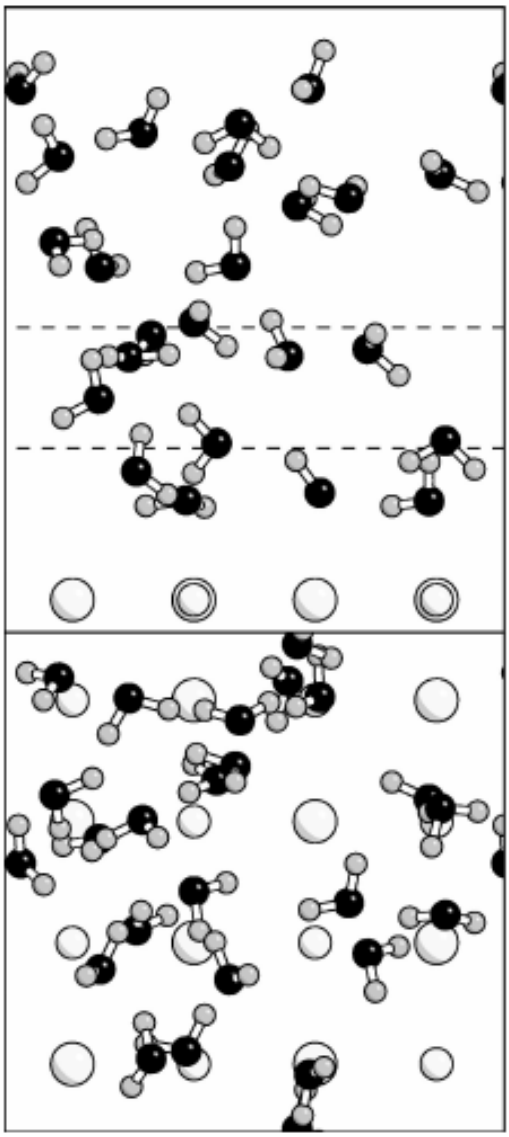

Figura 6: Visione laterale (in alto) e ortogonale (in basso) di una tipica configurazione di molecole di acqua adsorbite su una superficie di NaCl a coverage θ = 3.0.

Come è possibile osservare dallo *snapshot* (Figura 6), ad alti valori di *coverage* le molecole di acqua perdono la tendenza ad aggregarsi in isole separate, mentre formano invece strutture tridimensionali multistrato.

Le deduzioni ottenibili dall'osservazione qualitativa degli *snapshot* delle simulazioni sono avvalorate dalla valutazione della densità in funzione della distanza dalla superficie del sale, per il caso ad alto *coverage* sono distinguibili tre picchi correlati alla presenza di tre strati sovrapposti.

Al fine di investigare come la superficie di NaCl influenzi la struttura delle molecole d'acqua adsorbite, Engkvist e Stone hanno inoltre calcolato tre diverse funzioni di distribuzione radiale.

$$S_1(z) = \langle cos\beta \rangle_z$$

Equazione 3

Una di queste funzioni (Equazione 3) prevede lo studio del valore del coseno di β, ovvero l'angolo formato dalla normale alla superficie e l'asse C_2 di una molecola di acqua, in funzione della distanza (z) dalla superficie.

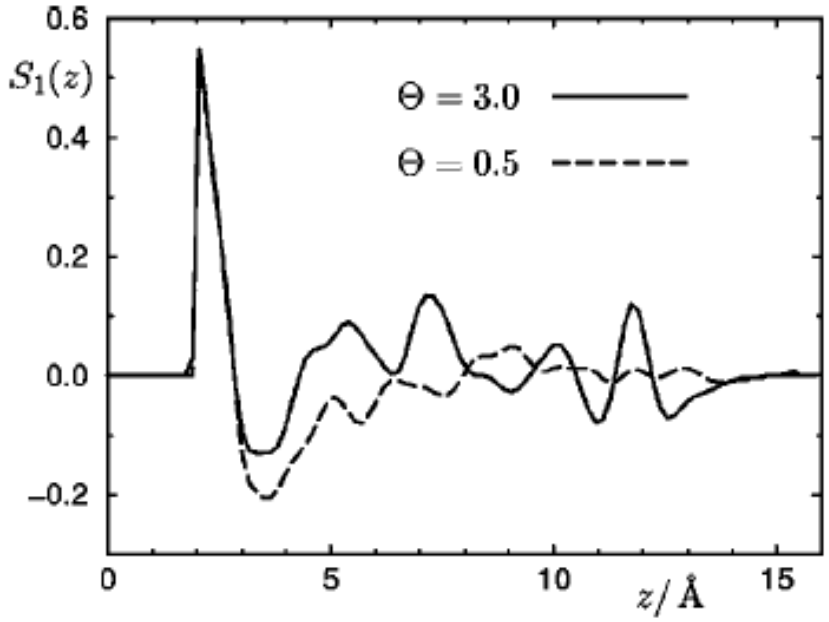

Figura 7: funzione della distribuzione di $S_1(z)$ per θ = 0.5 e θ = 3.0.

Interpretando $S_1(z)$ (Figura 7) è possibile dedurre come le molecole di acqua molto vicine alla superficie tendano a posizionarsi sopra Na^+ con gli atomi di idrogeno che puntano lontano dalla superficie ($\beta \approx 60°$). Le molecole leggermente più distanti presentano soprattutto legami a idrogeno con Cl^-, la bisettrice di HOH risulta quindi orientata prevalentemente verso la superficie. Per distanze maggiori dalla superficie invece $S_1(z)$ tende a 0, l'indebolimento progressivo dell'interazione con la superficie si manifesta come assenza di orientazione preferenziale. Queste tendenze sono rispettate sia nel caso a basso che ad alto *coverage.*

1.3 Simulazioni molecolari svolte nei precedenti lavori di Tesi

Lo scopo del tirocinio è stato quello di analizzare i risultati di simulazioni molecolari del processo di adsorbimento di acqua su una superficie modello di NaCl, le quali sono state effettuate in precedenti lavori di Tesi.

1.3.1 Metodi di simulazione molecolare: il metodo Monte Carlo Gran Canonico[22]

Nello studio del processo in esame, la tecnica utilizzata nelle simulazioni considerate è il metodo Monte Carlo nell'insieme termodinamico Gran Canonico, che prevede potenziale chimico, volume e temperatura costanti.
Nella configurazione sperimentale, infatti, la fase adsorbita risulta essere in equilibrio con la fase gas, dalla definizione di equilibrio termodinamico possiamo dedurre che il potenziale chimico e la temperatura della sostanza nelle due fasi siano uguali: un sistema di questo tipo individua un insieme termodinamico Gran Canonico a μ, T e V costanti. Dal punto di vista computazionale il sistema viene simulato considerando il sale adsorbente come in contatto con un serbatoio alla medesima temperatura T, nel quale è contenuto il vapore d'acqua. La pressione non è definita all'interno del sistema, ma risulta sempre ricavabile da un'equazione di stato, in modo da ottenere dati confrontabili con i risultati sperimentali determinati a pressione costante.
Al fine di ottenere risultati confrontabili con quelli sperimentali, è necessario tenere conto del fatto che in ambito sperimentale alcune condizioni (che definiscono l'*ensemble* termodinamico dell'esperimento) sono tenute costanti, le proprietà molecolari sono misurate come medie rispetto all'insieme termodinamico in cui si svolge l'esperimento e bisogna quindi sfruttare approcci computazionali in grado di riprodurre tali condizioni nel contesto della simulazione teorica.

[22] (Cramer, 2004)

Inoltre, nel caso in cui vengano studiati processi nei quali non avvengano importanti riorganizzazioni della distribuzione elettronica, come nel processo di adsorbimento di acqua su NaCl, il quale coinvolge solo interazioni di non-legame e di tipo ionico, il sistema può essere descritto in termini di meccanica classica. Lo stato del sistema classico può essere descritto dalla definizione delle tre coordinate spaziali (x_1, y_1, z_1) e le tre coordinate del momento $(p_{x,1}, p_{y,1}, p_{z,1})$ per ognuna delle N particelle presenti nel sistema, per un totale di 6N coordinate. Le 6N coordinate definiscono uno spazio multidimensionale che prende il nome di spazio delle fasi. In ogni istante il sistema occupa un punto dello spazio delle fasi che viene indicato con $X = (q, p)$, dove i termini q e p contengono l'insieme delle coordinate spaziali e del momento di ogni molecola presente nel sistema. Poiché lo spazio delle fasi comprende tutti i possibili stati di un sistema, il valore medio ⟨A⟩ di una proprietà, per un sistema all'equilibrio, dipenderà dalla probabilità che il sistema si trovi in un determinato punto dello spazio delle fasi e quindi dalla probabilità che la proprietà considerata assuma uno specifico valore A. La probabilità $P(q, p)$ che il sistema si trovi in un determinato punto dello spazio delle fasi risulta, a sua volta, dipendente dall'energia totale (come somma di energia cinetica ed energia potenziale) associata allo stesso. Al fine di determinare gli stati più probabili, la simulazione dovrebbe quindi calcolare il valore di energia per tutti i punti dello spazio delle fasi, l'operazione risulterebbe però molto complessa. Per questo motivo vengono utilizzati approcci, di cui un esempio è proprio il metodo Monte Carlo, che permettono di campionare solamente gli stati a più bassa energia, ovvero i più probabili.

Il metodo Monte Carlo si basa sulla perturbazione del sistema mediante spostamenti casuali delle particelle, sul calcolo della probabilità di esistenza dello stato specifico e una sua eventuale accettazione nel caso questa superi un valore di probabilità di soglia. Il metodo Monte Carlo Gran Canonico è una variazione del metodo Monte Carlo, opera in condizioni di potenziale chimico, temperatura e volume costanti, mentre il numero di particelle (*N*) può variare. Vengono scelte configurazioni perturbate mediante tre diversi movimenti: lo spostamento (traslazione o rotazione)

di una molecola, l'inserimento di una molecola in una posizione casuale della cella e la rimozione di una molecola. Viene calcolata la probabilità p associata alla configurazione perturbata e viene accettata se maggiore di un valore casuale z.

1.3.2 Studio dell'adsorbimento di acqua su una superficie modello di cloruro di sodio

Lo studio del processo di adsorbimento di acqua su particolato atmosferico è stato effettuato considerando una superficie modello di cloruro di sodio. La scelta è correlata al fatto che il cloruro di sodio risulta essere la componente maggioritaria dei *sea-salt aerosol (SSA),* di origine marina. Inoltre, per questo sistema sono state determinate le isoterme di adsorbimento dell'acqua per via sperimentale, tramite spettroscopia FTIR. Segue la descrizione dei parametri iniziali della simulazione.

Il sistema modello prevede la presenza di un cristallo di cloruro di sodio. La modellizzazione della superficie di NaCl (001) è stata effettuata sulla base dei parametri della cella cristallografica della salgemma, ovvero del cloruro di sodio cristallino, ricavati sperimentalmente tramite difrattometria a raggi X:

- a = b = c = 5.658 Å
- $\alpha = \beta = \gamma = 90°$

La struttura cristallina della salgemma è caratterizzata da impaccamento cubico compatto (ccp) e viene ottenuta per ripetizione nello spazio di una cella elementare cubica a facce centrate (Figura 8).

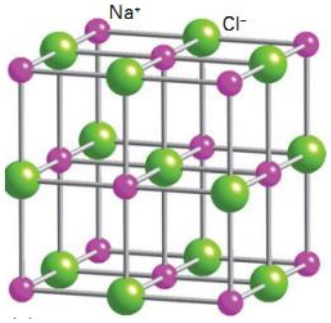

Figura 8: cella elementare della salgemma (cfc)[23].

[23] (6.11A: Structure - Rock Salt (NaCl), 2020)

La coordinazione è 6:6, entrambi gli ioni risultano coordinati ottaedricamente, la struttura cristallina può essere ottenuta per riempimento dello spazio con poliedri (tassellazione), in particolare ottaedri $NaCl_6$ o $(ClNa_6)$ con spigoli in comune (Figura 9). Le proprietà di adsorbimento della superficie di NaCl sono correlate alla bassa coordinazione che caratterizza gli ioni superficiali.

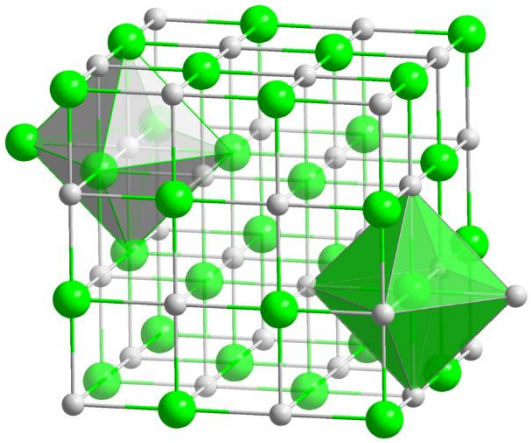

Figura 9: struttura cristallina di NaCl in cui è evidenziata la coordinazione ottaedrica[24].

Il modello del cristallo di NaCl è composto da cinque *layer* (strati bidimensionali) sovrapposti, ogni *layer* presenta 98 cationi Na^+ e 98 anioni Cl^- (196 ioni per strato) e complessivamente il sistema è costituito da 980 ioni. Il cristallo possiede un certo grado di mobilità, in particolare le simulazioni sono state effettuate con il primo, secondo, quarto e quinto *layer* mobili (solo il *layer* centrale possiede ioni vincolati).
La cella di simulazione è costituita da un parallelepipedo delle dimensioni di 39.606 Å x 39.606 Å x 50.000 Å, al centro del quale sono stati posti i 5 layer di NaCl (la distanza tra i nuclei dei due *layer* superficiali lungo l'asse maggiore della cella è pari a 11.316 Å).
La temperatura impostata per le simulazioni è 297.15K e il volume è mantenuto costante.
Al sistema considerato sono state applicate opportune condizioni periodiche al contorno nelle tre dimensioni spaziali, al fine di garantire che le molecole d'acqua ai

[24] (NaCl polyhedra.png, 2008)

limiti del sistema non soffrano di “effetti parete”. Essendo il sistema molto piccolo gli effetti sono importanti, imponendo una periodicità le molecole al limite esterno del parallelepipedo non interagiranno con una “parete” ma con le molecole contenute nella cella adiacente; di fatto il sistema da analizzare può essere visto come la cella elementare di un cristallo formato dalla traslazione periodica della stessa nelle tre dimensioni spaziali.

Tutte le simulazioni sono state avviate con un sistema iniziale che prevede la presenza di una sola molecola d’acqua (Figura 10).

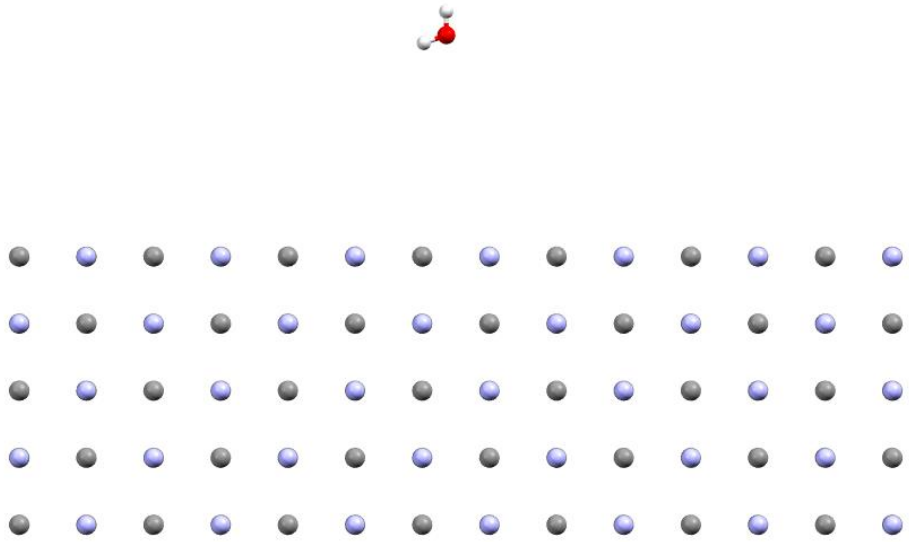

Figura 10: cella di simulazione (sistema di avvio simulazione).

Sono state considerate simulazioni nell’intervallo di pressione di H_2O tra 7.500 e 9.000 matm, il quale è stato campionato con un passo di 0.125 matm.

Per ogni valore di pressione di H_2O considerato è stata condotta una simulazione produttiva di 600 milioni di *step*, a valle di una fase di equilibrazione di 2.2×10^9 *step*, necessaria al fine di ottenere una condizione per cui il numero di molecole d’acqua presenti nel sistema oscilli intorno ad un valore medio pressoché costante, potendo

quindi considerare il sistema come all'equilibrio termodinamico. Per ogni simulazione sono state inoltre condotte tre repliche, le quali differiscono per la posizione dell'unica molecola d'acqua presente inizialmente nel sistema. Il campionamento delle energie e coordinate degli atomi presenti è stato effettuato ogni 100000 *step* (per ogni simulazione produttiva sono stati salvati 6000 *frame*, divisi in due "tronconi" da 3000 *frame*).

Per tutte le simulazioni eseguite sono state impostate frequenze costanti per le varie tipologie di perturbazione (traslazione Na^+ superficiali: 10%; traslazione Cl^- superficiali: 10%; traslazione molecola di H_2O: 20%; rotazione di un angolo massimo di 0.005° di una molecola di H_2O: 20%; inserimento/rimozione di una molecola di H_2O: 40%).

Nel metodo di simulazione molecolare adottato, al fine di determinare la configurazione perturbata più probabile, risulta necessario quindi valutare il valore dell'energia potenziale del sistema in funzione della sua posizione nello spazio delle fasi. Dato l'insieme di atomi presenti nella cella di simulazione è definibile una superficie di energia potenziale (*PES; Potential Energy Surface)*, che descrive l'energia potenziale dell'insieme di atomi in funzione della loro posizione relativa. Il calcolo dell'energia potenziale in funzione delle coordinate atomiche è stato effettuato considerando un'espressione empirica, chiamata campo di forze, basata interamente sulle leggi della meccanica classica. Infatti, sulla base dell'ipotesi che nel processo di fisiadsorbimento dell'acqua non avvengano rilevanti riorganizzazioni della distribuzione elettronica, vengono considerati i contributi derivanti dalle interazioni di non-legame (van der Waals), per le quali è stato imposto un *cutoff* di 12 Å, e ioniche (elettrostatiche), calcolate con il metodo della sommatoria di Ewald *(Particle Mesh Ewald, PME)*. Nel caso del processo di adsorbimento sono quindi studiante le interazioni a lungo raggio tra le molecole d'acqua (dipolo-dipolo), tra H_2O e gli ioni della superficie (ione-dipolo) e tra gli ioni che compongono il cristallo modellizzato (ione-ione).

Il campo di forza contiene i parametri necessari per il calcolo delle energie in gioco per quanto riguarda gli ioni Na^+ e Cl^-, poiché i parametri per l'acqua dipendono dal modello di acqua preso in considerazione. Il campo di forze utilizzato durante le simulazioni è di tipo AMBER (*Assisted Model Building with Energy Refinement*), modificato secondo i parametri del campo di forze AMBER99, secondo i risultati del lavoro di Joung e Cheatham[25] e sulla base di risultati di simulazioni precedenti.

Per quanto riguarda il modello di acqua (Figura 11), è stato utilizzato il SPC/E (*Extendend Simple Point Charge)*, basato sul modello rigido a tre centri SPC, nel quale sono stati cambiati i valori di carica atomica associati agli atomi di idrogeno e ossigeno (Tabella 1).

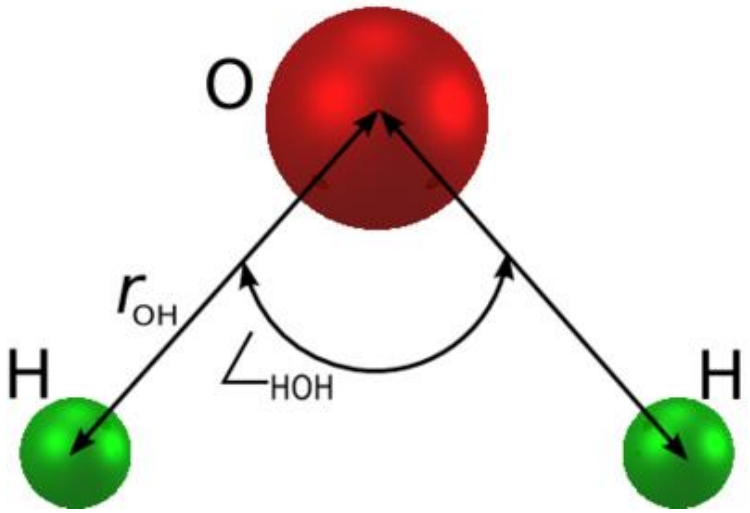

Figura 11: Modello SPC/E.

Parametro	***Valore***
σ	3.166 Å
ε	0.650 kJ/mol
r_{OH}	1.000 Å
$\angle_{HOH}$	109.47°
q_O	-0.8476
q_H	+0.4238

Tabella 1: Parametri caratteristici per il modello SPC/E.

[25] (Joung & Cheatham, 2008)

Lo studio effettuato in precedenti lavori di Tesi ha permesso di ottenere l'isoterma teorica a 24 °C (Figura 12) relativa all'adsorbimento di acqua sulla superficie di NaCl(001), confrontata con l'isoterma sperimentale ottenuta dai dati campionati da Foster e Ewing (Figura 13).

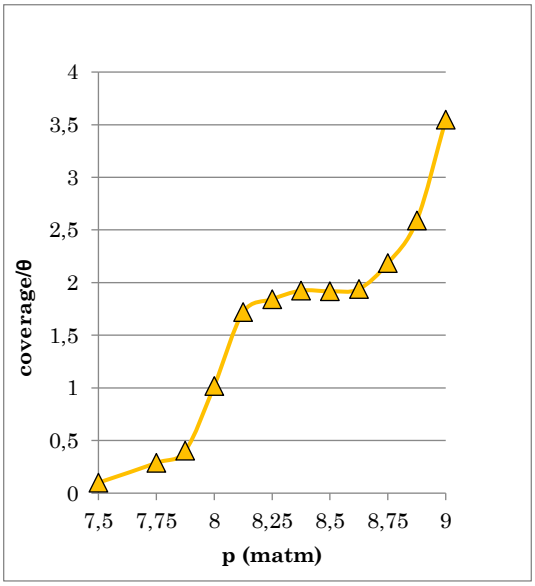

Figura 13: Isoterma di adsorbimento teorica di acqua su NaCl a 24 °C, ricavata dai risultati delle simulazioni dei precedenti lavori di Tesi.

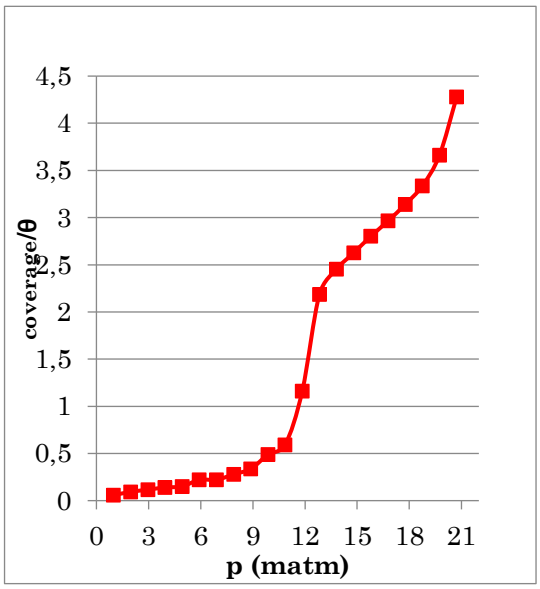

Figura 12: Isoterma sperimentale di adsorbimento di acqua su NaCl a 24 °C relativa allo studio sperimentale di Foster e Ewing.

L'isoterma teorica risulta avere forma simile a quella sperimentale, presenta però una compressione lungo l'asse delle ascisse, legata a una sovrastima delle interazioni acqua-acqua rispetto alle interazioni acqua-superfici.

È importante osservare, al fine di giustificare i risultati delle analisi descritte nei capitoli successivi, una discrepanza tra i *range* che definiscono le regioni di *coverage* nelle isoterme teoriche e sperimentali: la regione a basso *coverage* ($\theta \leq 0.5$) è compresa tra 7.500 e 7.875 matm; la regione di transizione ($0.5 < \theta \leq 1.8$) tra 8.000 e 8.250 matm; e la regione ad alto *coverage* ($1.8 < \theta \leq 3.5$) per valori di pressione di H_2O compresi tra 8.375 e 9.000 milliatmosfere. La regione ad alto *coverage* è caratterizzata dalla presenza di una zona di *plateau* ($1.8 < \theta < 2.0$) compresa tra 8.375 e 8.625 milliatmosfere.

Capitolo 2

2. Metodologie e algoritmi per l'analisi dei dati di simulazioni molecolari nello studio del processo di adsorbimento di acqua su NaCl

Lo scopo del tirocinio è stato studiare il processo di adsorbimento di acqua su particolato atmosferico, attraverso una *data analysis* dei risultati delle simulazioni computazionali descritte nel Capitolo 1.4, eseguite considerando un sistema modello di una superficie di cloruro di sodio. L'attività svolta si è concentrata sullo studio dei fenomeni di tipo aggregativo che coinvolgono le molecole d'acqua adsorbite sulla superficie del sale e interagenti tramite legame a idrogeno. L'obiettivo è stato quello di verificare la presenza, e le proprietà, delle strutture ipotetiche descritte negli studi sperimentali di Foster ed Ewing, sulla base della posizione della banda di *stretching* dell'acqua, negli spettri di assorbimento IR ottenuti a diversi valori di *coverage*. Le caratteristiche degli aggregati sono state studiate a livello computazionale da Engkvist e Stone, ma lo studio si è limitato a due valori di *coverage*, rispettivamente nelle regioni a bassa e alta copertura, non considerando quindi la regione di transizione. L'analisi dei risultati ottenuti nelle simulazioni molecolari effettuate nei precedenti lavori di tirocinio ha permesso invece di studiare i fenomeni di aggregazione, da un punto di vista microscopico, in un intervallo di pressioni (e di θ) tale da consentire l'esplorazione di tutte le regioni dell'isoterma di adsorbimento.

Il problema principale connesso all'analisi del sistema oggetto del nostro studio è l'elevata quantità dati da considerare. Infatti, per ogni simulazione (in tre repliche), ad una fissata pressione di H_2O, sono disponibili le posizioni, espresse in coordinate cartesiane, di tutti gli atomi nel sistema, per ognuno dei 6000 *step* salvati. Inoltre, il numero medio delle molecole d'acqua presenti nel sistema varia in funzione della pressione, passando da circa 20 molecole a 7.500 matm fino a circa 700 molecole a 9.000 matm (Figura 14). Non essendo nativamente presente nel software

DL_MONTE[26], utilizzato per le simulazioni, un *tool* di analisi dei dati utile allo studio dei fenomeni aggregativi, il lavoro svolto ha visto lo sviluppo di uno *script*, scritto nel linguaggio di programmazione Python (tra i più utilizzati per la computazione scientifica), che permettesse una *data analysis* automatizzata e ottimizzata, in termini di performance e tempistiche di calcolo, delle configurazioni generate dalle simulazioni molecolari del processo di adsorbimento di acqua su NaCl. Al fine di studiare le proprietà dei *cluster* formati sulla superficie del sale, è stato in primo luogo necessario individuare un metodo che permettesse, per ognuno dei *frame* generati (e salvati) durante la simulazione, di verificare la presenza di aggregati, e quindi di assegnare ad ogni molecola di acqua presente nel sistema una *label*, che ne definisse l'appartenenza a uno specifico aggregato (tutte le molecole dello stesso aggregato sono indicate dalla stessa etichetta) o eventualmente la natura di particella isolata.
Conoscendo il *cluster* di appartenenza di ogni molecola, e quindi il numero di molecole nello stesso, è possibile successivamente classificare il *cluster* come "isola" o come "strato" e studiare separatamente le proprietà delle molecole appartenenti alle due tipologie di aggregato. I risultati ottenuti per ogni *frame* vengono poi mediati sull'intera simulazione e, dai valori ottenuti da tutte le simulazioni, si ottiene uno studio della dipendenza di determinate proprietà in funzione della classe di *cluster* e della pressione di H_2O nel sistema.

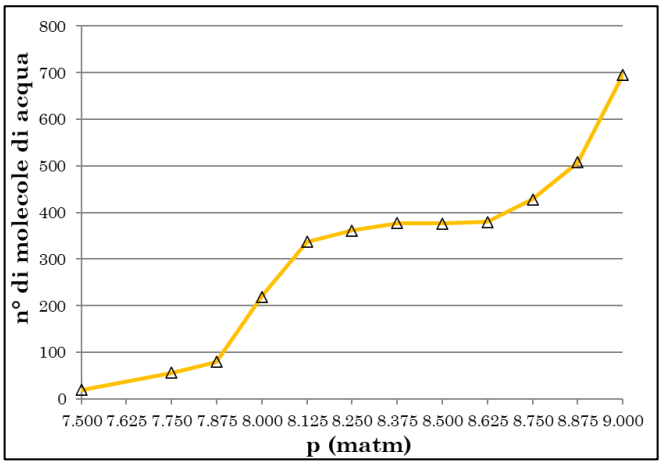

Figura 14: n° medio di molecole d'acqua nel sistema (valore mediato sulle 3 repliche) in funzione della pressione di H_2O.

[26] (Crabtree, Parker, & Purton, 2013)

L’approccio adottato ha quindi visto una fase iniziale di selezione di un algoritmo di clusterizzazione adatto al tipo di dati da analizzare (coordinate atomiche) ed efficiente in termini di risorse di computazione utilizzate.

2.1 Metodi di Analisi

Il *clustering* (o analisi dei gruppi) è una tecnica di *data analysis* che, quando applicata ad un set di oggetti eterogenei, identifica sottogruppi omogenei, sulla base di un modello fornito o di una misura di similarità (o di distanza) tra gli elementi[27].

La clusterizzazione è una tecnica comune nell’analisi statistica dei dati (*statistical data analysis)*, dove viene sfruttata prevalentemente per scopi di caratterizzazione e descrizione dei fenomeni, e nel *machine learning*, con un obiettivo di natura maggiormente predittiva; in entrambi i casi il traguardo comune resta quello di estrarre e derivare conoscenza dai dati.

La *cluster analysis trova* utilizzo in molti campi, come la bioinformatica, l’automazione industriale, l’astronomia, l’ingegneria, la medicina e la chemioinformatica. Un esempio di utilizzo della clusterizzazione su *set* di dati di interesse chimico è l’estrazione di sottogruppi rappresentativi in dati ottenuti da processi di *high throughput screening* (screening ad alta capacità) e dalla chimica combinatoriale, applicata alla scienza dei materiali o al processo di *drug discovery*, correlata alla sintesi rapida e all’analisi automatizzata di un grande numero di molecole con caratteristiche strutturali simili. Gli algoritmi di *machine learning* (*ML*), come gli algoritmi di *clustering***,** permettono di estrarre informazioni di interesse chimico da grandi database di composti, rendendo possibile la progettazione di farmaci con importanti proprietà biologiche, con un costo minore e un’efficienza maggiore rispetto ai metodi tradizionali[28].

Se da una parte la statistica classica prevede prevalentemente uno studio dei dati attraverso il *fitting* di una distribuzione di punti, quindi la determinazione dei

[27] (Downs & Barnard, 2002)
[28] (Lo, Rensi, Torng, & Altman, 2018)

parametri di una distribuzione prevista (un modello matematico noto, solitamente una distribuzione normale) per la quale questa presenti la migliore corrispondenza con la distribuzione iniziale, i metodi di *machine learning* sono caratterizzati da una forte capacità di modellizzazione non-lineare e utilizzano i dati per costruire modelli che si basano su strutture di dati complesse. È proprio grazie a questa proprietà che i metodi del *machine learning*, tra cui il *clustering*, hanno trovato grande applicazione in diversi campi delle scienze ambientali[29], come nelle previsioni metereologiche e della qualità dell'aria, nel monitoraggio dei ghiacci e delle foreste, ma anche nella classificazione degli aerosol atmosferici in funzione della composizione chimica e della loro dimensione[30].

Nel dettaglio, il *clustering* è una tecnica di apprendimento non supervisionato (*unsupervised learning),* permette l'identificazione autonoma di strutture e *pattern* nascosti nei dati in *input* senza che siano presenti *label* preesistenti associate ai dati e con una supervisione umana minima. La clusterizzazione si distingue quindi dalle tecniche di *supervised learning*, le quali hanno lo scopo di stabilire relazioni tra dati di *input* e *output* già esistenti, così da permette di predire successivamente *output* a partire da nuovi *input*. Il complementare supervisionato del *clustering*, da questo punto di vista, è la *classificazione*, che prevede la presenza di un *set* di classi predefinite nelle quali gli oggetti vengono inseriti in funzione di alcune proprietà che li caratterizzano. Nello *script* sviluppato sono esempi di classificazione: la distinzione degli aggregati in "isole" e "strati" in funzione del numero di particelle contenute negli stessi e la classificazione delle molecole d'acqua in funzione della distanza dalla superficie.

Nonostante le tecniche di clusterizzazione siano nate con lo scopo di gestire grandi set di dati ad alta dimensionalità, la maggior parte di esse funzionano bene anche su piccoli *dataset* a bassa dimensionalità, e sono di conseguenza adatte allo studio dei singoli *frame* contenenti le coordinate atomiche tridimensionali delle molecole

[29] (Hiesh, 2009)
[30] (Christopoulos, Garimella, Zawadowicz, Möhler, & Cziczo, 2018)

d’acqua nella cella di simulazione. Nonostante questo, però, non tutti gli algoritmi di *clustering* risultano efficacemente applicabili al caso studiato. Il metodo scelto infatti deve presentare le seguenti caratteristiche:

1) la conoscenza del dominio richiesta per determinare i parametri di *input* deve essere minima e coerente con quanto già si conosce del sistema, i valori appropriati non sono infatti di facile determinazione;
2) il metodo deve riuscire ad individuare aggregati di forma arbitraria, anche non-globulare;
3) deve essere efficiente dal punto di vista delle risorse e dei tempi di calcolo.

2.1.1 Classificazione dei metodi di *clustering*

Una possibile classificazione dei metodi di clusterizzazione può essere ottenuta in funzione del tipo di algoritmo utilizzato per dividere lo spazio. Sulla base di questo criterio sono individuabili tre tipologie principali di algoritmi: algoritmi di *clustering* partizionale, algoritmi di *clustering* gerarchico e algoritmi di *clustering* basati sulla densità.

Gli algoritmi di *clustering* partizionale prevedono la partizione di un database *D*, di *n* oggetti, in un *set* di *k cluster* ($k \leq n$). Il numero di aggregati *k* è un parametro di ingresso di questo tipo di algoritmi, ciò significa che è necessaria una buona conoscenza del dominio, che non è sempre assicurata in tutte le applicazioni (ad esempio, nel caso studiato, non è possibile sapere in principio il numero di aggregati di molecole d’acqua che saranno presenti in una certa configurazione del sistema).

Dato che lo studio di ogni possibile partizione è infattibile a livello computazionale, questa classe di metodi comporta tipicamente l’individuazione di una singola partizione iniziale di *D* e un successivo utilizzo di una strategia di ottimizzazione iterativa, che prevede la riassegnazione dei punti nei *k cluster* individuati, attraverso

degli schemi di ricollocazione che variano in funzione dell'algoritmo considerato[31]. Il fine del processo iterativo è quello di ottimizzare una funzione di partizionamento oggettivo, o funzione di similarità, associata ad esempio alla distanza tra i punti, in modo tale da massimizzare la similarità *intra-cluster* e minimizzare quella *inter-cluster*. Ogni *cluster* è rappresentato dalla media dei suoi punti o *centroide* del *cluster* (algoritmo *k-means*), oppure da un oggetto del *cluster* localizzato vicino alla sua media (algoritmi *k-medoid).* La funzione di partizionamento oggettivo nell'algoritmo *k-means*, attualmente uno dei più utilizzati algoritmi di clusterizzazione per le applicazioni scientifiche e industriali, è tipicamente la somma degli errori quadratici tra un set di punti e il loro centroide, la quale esprime la varianza *intra-cluster* totale (Equazione 4).

$$E(C) = \sum_{j=1:k} \sum_{x_i \epsilon C_j} ||x_i - c_j||^2$$

Equazione 4: E(C) è l'errore quadratico totale, c_j è la media del cluster j-esimo, x_i è la posizione dell'oggetto *i*-esimo nel cluster j-esimo, *k* è il numero totale di cluster.

L'algoritmo *k-means* funziona in step, vengono inizialmente individuati in modo casuale *k* centri dei *cluster*, assegnati a *k* degli *n* oggetti del sistema, successivamente ogni oggetto rimanente viene assegnato al *cluster* che presenta il centroide più "vicino" all'oggetto stesso, al termine dell'operazione viene ricalcolata la media del *cluster*. Il processo viene poi iterato fino alla convergenza della funzione di partizionamento oggettivo. Il metodo è particolarmente efficiente nel trattare grandi insiemi di dati, ma risulta sensibile agli *outlier*, i punti periferici possono infatti provocare una forte distorsione della distribuzione dei dati.

L'algoritmo *k-medoid* presenta una minore sensibilità agli *outlier*, il punto rappresentativo del cluster è infatti l'oggetto più vicino alla media del *cluster* (medoide), la funzione minimizzata durante il partizionamento esprimerà la somma

[31] (Berkhin, 2006)

delle dissimilarità tra ogni oggetto e l'oggetto rappresentativo del *cluster* corrispondente, ovvero l'errore assoluto totale (Equazione 5).

$$E(C) = \sum_{j=1:k} \sum_{x_i \epsilon C_j} ||x_i - a_j||$$

Equazione 5: Equazione 6: E(C) è l'errore assoluto totale, a_j è la posizione del medoide del cluster j-esimo, x_iè la posizione dell'oggetto *i*-esimo del cluster j-esimo, *k* è il numero totale di cluster.

Tutti i metodi di *clustering* partizionale presentano dei limiti comuni, per i quali sono risultati inadatti allo studio degli aggregati di molecole d'acqua adsorbite sulla superficie di NaCl:

- sono contemplati solo *cluster* di forma convessa, la quale non è necessariamente una caratteristica degli aggregati di molecole d'acqua;
- il risultato dipende fortemente dai centroidi (o medoidi) scelti inizialmente;
- è necessario conoscere preventivamente il numero di aggregati presenti nel sistema, mentre nel caso specifico è un dato ignoto che deve essere ricavato dalla clusterizzazione stessa;
- non contemplano l'esistenza di *noise point*, quindi di molecole d'acqua isolate e non interagenti tramite legame ad idrogeno.

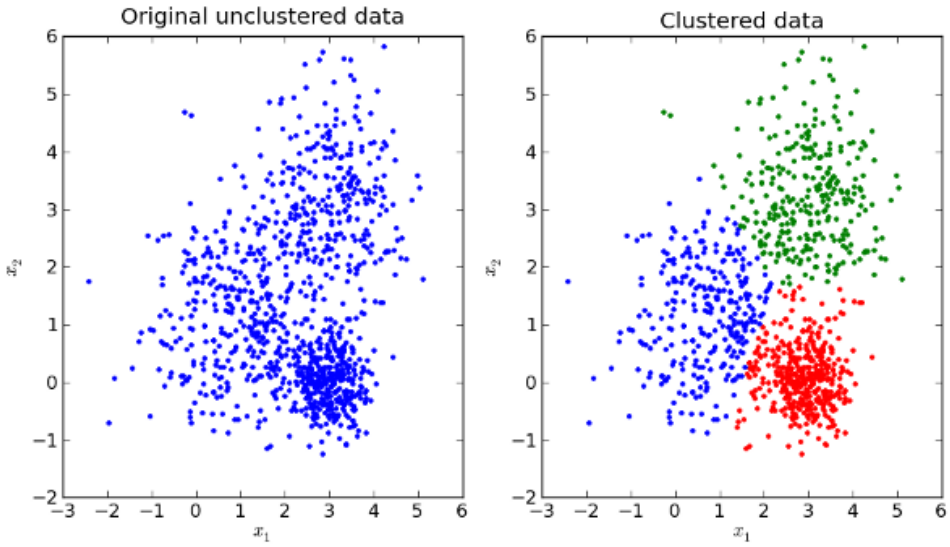

Figura 15: Esempio di clusterizzazione su dati bidimensionali tramite *k-means* con *k*=3. A sinistra sono rappresentati i dati originali, a destra i risultati del *clustering*.

Una seconda classe di metodi di clustering molto diffusa è quella degli algoritmi di clustering gerarchico (*hierarchical algorithms*), il quale funzionamento prevede il raggruppamento degli oggetti in *alberi* di *cluster*.

Il *set* di dati è analizzato in modo iterativo, in ogni *step* una coppia di *cluster* viene unita o un singolo *cluster* viene diviso, il risultato è gerarchico, con una relazione *genitore-figlio* stabilita tra i *cluster* ad ogni livello successivo dell'iterazione.

Un albero raffigurante questa gerarchia, chiamato *dendrogramma* (Figura 16), rappresenta la divisione di *D* iterativamente in *subset* più piccoli fino al raggiungimento di uno stato per cui ogni *subset* è composto da un singolo oggetto.

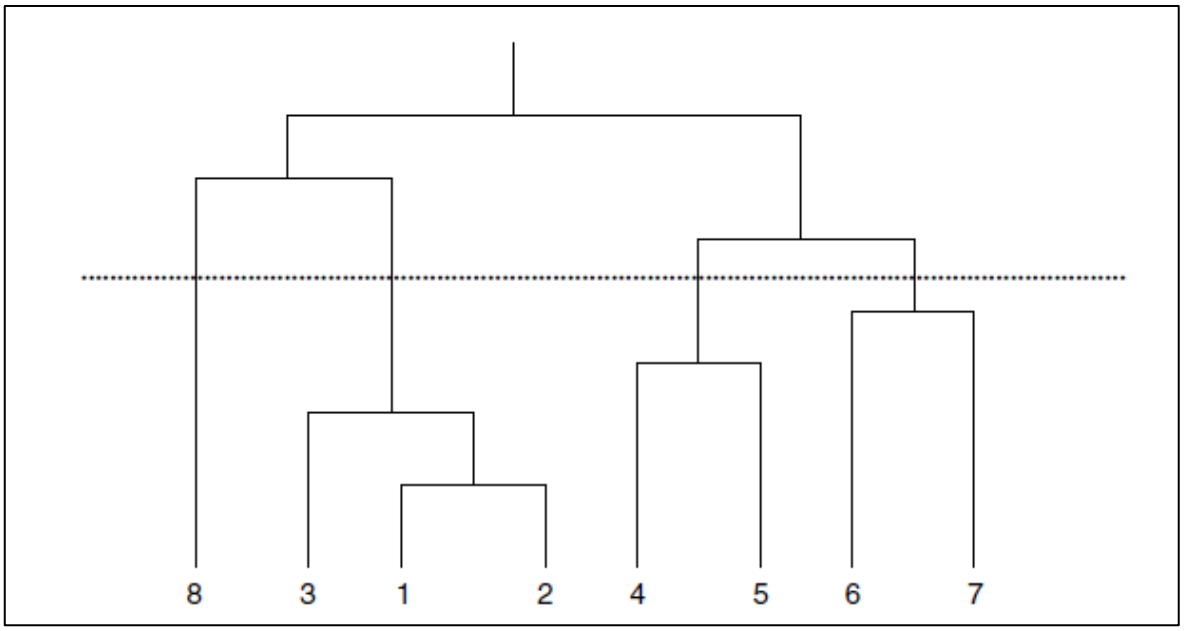

Figura 16: Un esempio di gerarchia (dendrogramma) generato dalla clusterizzazione di otto oggetti (etichettati dai numeri presenti in basso). In alto un *cluster* contenente tutti gli otto elementi. La linea tratteggiata orizzontale rappresenta una singola partizione contenente quattro cluster [8], [3,1,2], [4,5] e [6,7].

In un dendrogramma gli oggetti individuali sono le *foglie* dell'albero, e sono posti in basso, ogni *nodo* interno dello stesso rappresenta un cluster di *D*, in alto (*radice* dell'albero) è rappresentato un singolo *cluster* che contiene tutti gli elementi. La posizione verticale delle linee orizzontali che uniscono coppie di oggetti o *cluster* ne indica il grado di similarità.

I metodi di clusterizzazione gerarchica sono classificati in *agglomerativi* (*bottom-up*) e *divisivi* (*top-down*).

I metodi gerarchici agglomerativi prevedono una costruzione del dendrogramma partendo dalle foglie e arrivando alla radice. Inizialmente tutti gli oggetti sono inseriti in un *cluster* individuale, negli *step* successivi avviene la fusione ricorsiva di due o più *cluster* dalle proprietà simili, fino a ottenere un unico *cluster* oppure fino alla soddisfazione di specificati criteri di terminazione. I metodi appartenenti a questa classe si differenziano in funzione della misura di similarità *inter-cluster*.

I metodi gerarchici divisivi seguono un approccio inverso rispetto ai metodi agglomerativi. Inizialmente si ha un singolo *cluster*, contenente tutti gli oggetti del sistema, il quale viene ricorsivamente diviso in *subset* dalle caratteristiche opportune, fino alla soddisfazione del criterio di terminazione.

I vantaggi del *clustering* gerarchico includono:

- flessibilità riguardo il livello di *granularità* nella clusterizzazione;
- facilità di gestione di tutte le forme di similarità o distanza.

Gli svantaggi principali del *clustering* gerarchico sono:

- la maggior parte degli algoritmi gerarchici non modifica i *cluster* una volta che sono stati costruiti;
- efficienza minore rispetto ad altri algoritmi nella gestione di grandi *database*;
- difficoltà nella scelta del giusto criterio di terminazione.

I metodi di *clustering* gerarchico, contrariamente ai metodi di *clustering* partizionale, non necessitano che sia specificato il numero *k* di *cluster* da individuare, è richiesta però la definizione di parametri di terminazione che indichino all'algoritmo quando la fusione o la divisione dei *cluster* debba essere fermata. Un esempio di criterio di terminazione è la distanza critica D_{min} tra tutti i *cluster* del sistema, la quale deve essere abbastanza piccola da permettere la separazione di tutti i *cluster* "naturalmente" presenti nel sistema, ma al contempo abbastanza grande da non provocare la divisione degli stessi.

L’aspetto più problematico associato ai metodi di *clustering* gerarchico è proprio la difficoltà nella definizione dei parametri di terminazione appropriati e per questo spesso il criterio di terminazione viene fatto coincidere con il numero *k* di *cluster* da individuare.

2.1.2 DBSCAN: Un algoritmo di clusterizzazione basato sulla densità

Osservando i tre insiemi di punti bidimensionali rappresentati nella Figura 17, risulta intuitiva e chiara la definizione di gruppi di punti e di punti isolati (*noise point*) non appartenenti a nessuno dei *cluster* individuati.

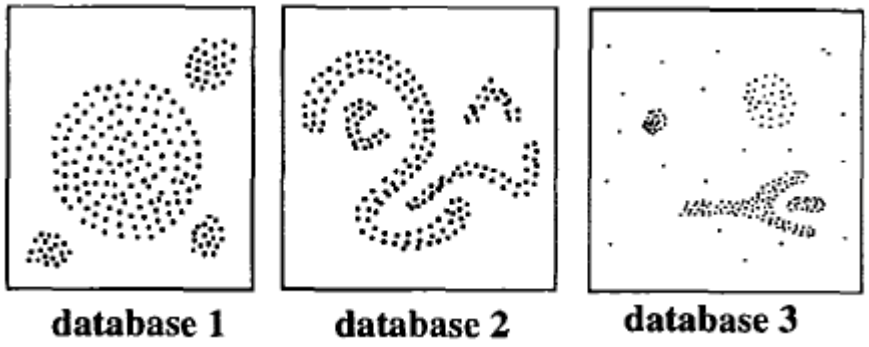

Figura 17: Set di *database* di esempio.

La ragione principale per cui siamo in grado di riconoscere intuitivamente i *cluster* è correlata al fatto che all’interno dei *cluster* la densità dei punti è considerevolmente maggiore rispetto che all’esterno degli stessi. Inoltre, la densità caratteristica delle zone di rumore è minore della densità in ognuno dei *cluster* individuati.
Un approccio alla clusterizzazione di questo tipo è caratteristico dei metodi di *clustering* basati sulla densità, che effettuano una divisione dello spazio dei dati euclideo, bidimensionale o tridimensionale (anche se in alcuni casi risultano applicabili anche in spazi ad alta dimensionalità), in regioni ad alta densità di oggetti, che individuano i cluster, e regioni a bassa densità, che ospitano il rumore.

L'esempio più noto, e più citato in letteratura, di algoritmo di *clustering* basato sulla densità, è DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*), proposto nel 1996 da Martin Ester *et al.*[32], metodo selezionato per eseguire l'analisi dei gruppi durante il lavoro di tirocinio. Il metodo definisce ogni *cluster* connettendo regioni dello spazio con densità sufficientemente elevata, i *cluster* quindi "crescono" in tutte le direzioni in cui vengono condotti dalla densità. Di conseguenza i metodi basati sulla densità presentano una protezione naturale verso gli *outlier* e permettono di identificare *cluster* con forma arbitraria, anche non globulare. In Figura 18 sono illustrate forme di *cluster* problematiche per gli algoritmi di *clustering* partizionale (come *k-means*) ma gestite propriamente dagli algoritmi *density-based*.

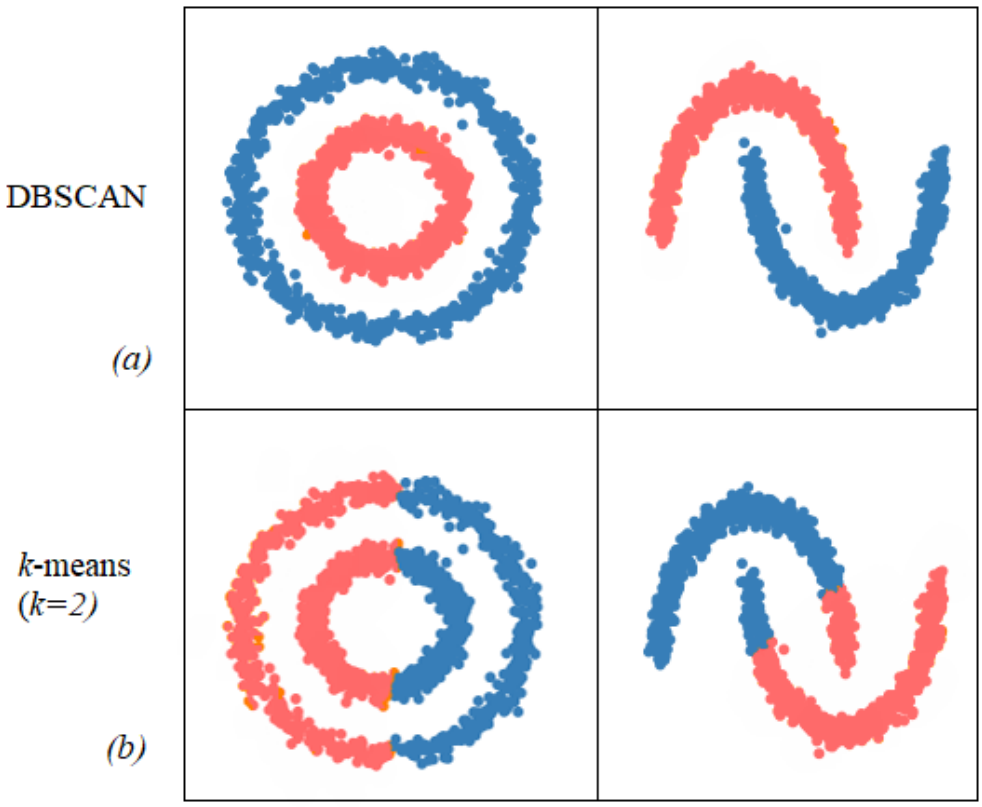

Figura 18: Clustering su due insiemi di punti bidimensionali ad opera di (a) DBSCAN e (b) *k*-means, con *k=2*.

[32] (Ester, Kriegel, Sander, & Xu, 1996)

DBSCAN, come tutti gli algoritmi di clustering basati sulla densità, risulta particolarmente efficiente nello studio di *database* a bassa dimensionalità (il numero di variabili che definisce ogni oggetto è piccolo), è quindi ottimale per studiare le configurazioni generate durante la simulazione del processo di adsorbimento, in cui ogni atomo nella cella di simulazione, quindi ogni oggetto di *D*, è definito solo dalle coordinate atomiche tridimensionali.

Un ulteriore vantaggio di DBSCAN è il fatto che non necessiti la conoscenza del numero di *cluster* (parametro *k*) in cui dividere gli *n* oggetti di *D*, a differenza gli algoritmi di *clustering* partizionale, e non preveda l'inserimento delle complesse condizioni di terminazione tipiche dei metodi di *clustering* gerarchico.

Il funzionamento di DBSCAN si basa su un'idea fondamentale: il vicinato (*neighborhood*) di ognuno degli oggetti di un *cluster*, definita come una regione intorno all'oggetto associata ad un determinato raggio, deve contenente un certo numero di oggetti, ovvero deve essere caratterizzata da una densità superiore ad un valore di soglia. La forma del *neighborhood* dipende dalla funzione utilizzata per definire la distanza tra due oggetti nei punti *p* e *q* nello spazio, indicata con *dist(p,q)*. L'approccio di DBSCAN permette di utilizzare una qualsiasi formula per definire la distanza tra due punti, in modo appropriato al tipo di applicazione, nel caso specifico è stata utilizza la distanza euclidea (Equazione 7).

$$d(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

Equazione 7: Distanza euclidea tra due punti P_1 e P_2 nello spazio tridimensionale.

Per una migliore comprensione del ruolo dei due parametri di *input*, ovvero *Eps* (o ε) e *MinPts*, e del funzionamento di DBSCAN è necessario introdurre una serie di definizioni:

1) Il vicinato di raggio ε di un oggetto in un punto *p*, indicato con $N_{Eps}(p)$, è chiamato *Eps-neighborhood* (o *Eps-vicinato*) ed è definito come $N_{Eps}(p) = \{q \in D \mid dist(p,q) \leq Eps\}$. Un'ipotesi di approccio al *clustering* basato su

questa semplice definizione potrebbe richiedere che ogni oggetto in un *cluster* possieda un numero minimo (*MinPts*) di oggetti nel suo *Eps*-vicinato. Tuttavia, questo approccio fallisce perché esistono due tipologie di oggetti in un *cluster* (Figura 19a), gli oggetti interni (*core point*) e quelli posizionati sul confine (*border point*). I *border point* presentano un numero inferiore di oggetti nel loro *Eps*-vicinato rispetto ai *core point.*

2) Un *core point* è un oggetto, in un punto *p* di un *cluster*, che presenta un numero minimo (*MinPts*) di punti nel suo *Eps*-vicinato, vale quindi la relazione $|N_{\epsilon}(p)| \geq MinPts$.
3) Un oggetto in un punto *p* è *direttamente density-reachable* da un oggetto in un punto *q* se sono soddisfatte entrambe le seguenti condizioni:
   a. $p \in N_{Eps}(q)$.
   b. $\left|N_{Eps}(q)\right| \geq MinPts$ (*core point condition*, indica che *q è un* core *point*).

   La relazione è simmetrica per due *core point*, ma non è simmetrica se un'oggetto è un *core point* e l'altro è un *border point*, poiché quest'ultimo non rispetta la *core point condition* (Figura 19b).

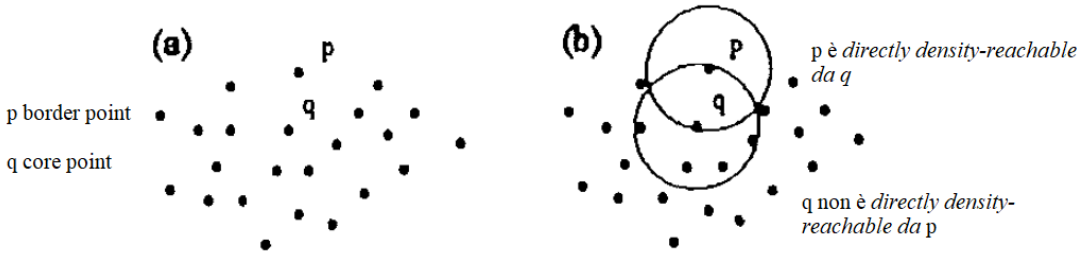

Figura 19: ***(a) Core point* e *border point* di un *cluster*. *(b)* Asimetria della *direttamente density-reachability*.**

4) Un punto *p* è *density-reachable* da un punto *q* se esiste una catena di oggetti $p_1, \ldots, p_n$ ($p_1 = q$, $p_n = p$) per cui p_{i+1} è *direttamente density-reachable* da p_i. Tra *p* e *q* è quindi presente una sequenza di *core point* per cui ogni punto p_{i+1} si trova nel *Eps*-vicinato del suo predecessore p_i.

   Il concetto di *density-reachable* è l'estensione del concetto di *direttamente density-reachable*, la relazione è transitiva ma non è sempre simmetrica (è

sempre simmetrica se p e q sono entrambi *core point*). In Figura 20a è rappresentato il concetto di *density-reachable* ed in particolare il caso non simmetrico.

Due *border point* non sono tra loro *density-reachable*, perché non è rispettata la *core point condition* all'interno della catena, il primo punto della stessa deve essere un *core point*. Nonostante questo, in un c*luster* deve esistere un *core point* dal quale entrambi i *border point* siano *density-reachable*, viene quindi introdotto il concetto di *density-connected.*

5) Un punto *p* ed un punto *q* sono *density-connected* se esiste un punto *o* dal quale sia *p* sia *q* siano *density-reachable*. La relazione in questo caso è sia riflessiva che simmetrica (Figura 19b).

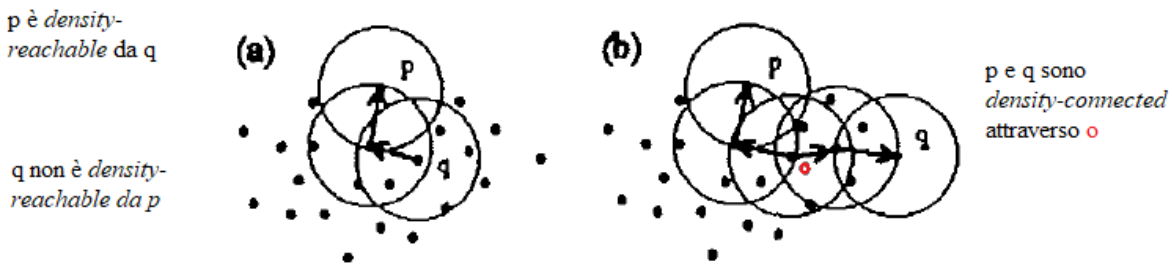

Figura 20: Cluster di punti in cui è rappresentata (a) l'asimmetria nella relazione di *density-reachability* e (b) la simmetria della relazione di *density-connectivity*.

6) Dato un insieme di *k cluster* $C_1, \dots, C_k$ in un database *D* è definito rumore un punto *p* che non appartiene a nessuno dei *cluster*, vale quindi la relazione: $p \in D | \forall\, i: p \notin C_i$. Un *noise point* non è un *core point* e non è *direttamente density-reachable* o *density-reachable* da nessun *core point.*
7) Dato un database di punti *D*, viene definito *cluster C*, considerando opportuni valori dei parametri *MinPts* e *Eps*, un sottoinsieme non vuoto di *D* per il quale valgono le seguenti condizioni:
    a. $\forall\, p, q$: se $p \in C$ e q è *density-reachable* da p allora $q \in C$.
    b. $\forall\, p, q \in C: p$ è *density-connected* a q.

Sulla base del contenuto dei punti 7a e 7b è possibile asserire che metodi di *clustering* basati sulla densità, come DBSCAN, formano *cluster* composti da oggetti *density-connected* per il quale si ha massimizzazione della *density-reachability.*
Dall'insieme delle definizioni si possono inoltre trarre alcune considerazioni generali sui parametri di *input* dell'algoritmo:

- *Eps*: definisce la distanza massima tra due punti *p* e *q* per cui questi siano considerabili vicini.
  La determinazione di un valore di *Eps* appropriato è fondamentale per la riuscita del processo di clusterizzazione e richiede una buona conoscenza del dominio, relativa alla distanza che caratterizza gli oggetti presenti nei *cluster*. Nel caso specifico un valore consono è definibile sulla base della natura del legame a idrogeno instaurato tra le molecole d'acqua presenti negli aggregati.
- *MinPts:* determina il numero minimo di oggetti che un *core point* deve avere nel suo *Eps-vicinato* affinché sia considerabile tale, compreso esso stesso.
  Il valore del parametro coincide con il numero minimo di oggetti di cui un *cluster* deve essere composto. Questo perché: ogni *cluster C* non vuoto deve essere formato da almeno un punto *p*, *p* deve essere *density-connected* a se stesso (la *density-connectivity* è riflessiva), deve quindi esistere un *core point o* (che può coincidere con *p*) dal quale *p* è *density-reachable* e, per la *core point condition*, *o* deve contenere almeno *MinPts* punti nel suo vicinato.

A causa della definizione di parametri globali, validi per tutti i *cluster* del sistema, DBSCAN non opera adeguatamente con *cluster* caratterizzati da variabilità della densità *intra-cluster*. L'algoritmo, per quanto sia ottimale nella definizione di regioni ad alta e bassa densità di elementi nello spazio, distinguendo punti appartenenti a *cluster* da punti di rumore, non distingue *cluster* adiacenti (o sovrapposti) che presentino densità diversa, fondendoli in un singolo aggregato. Nel caso studiato, il legame a idrogeno impone una certa rigidità alle strutture di molecole d'acqua

adsorbite, si ipotizza quindi l'assenza di strutture aggregate che siano caratterizzate da grande variabilità della densità interna.
Per quanto riguarda il funzionamento dell'algoritmo, DBSCAN effettua la clusterizzazione definendo inizialmente l'*Eps*-vicinato di ogni oggetto del *database*. Vengono identificati successivamente tutti i *core point* nel sistema, ovvero tutti i punti che presentano un numero di oggetti maggiore di *MinPts* nel loro intorno di raggio ε, e per ogni *core point* viene definito un singolo *cluster*. Sono poi individuati tutti gli oggetti che risultano *density-reachable* dai vari *core point* e uniti ai rispettivi *cluster*. Se due *core point* sono reciprocamente *density-reachable* allora i due *cluster* vengono fusi. Il calcolo ha fine quando non risulta più possibile aggiungere oggetti ai *cluster*, quindi ogni punto è inserito in un *cluster* o etichettato come rumore.

2.2 Sviluppo degli algoritmi di analisi

Il lavoro di tirocinio, in seguito alla fase di selezione dell'appropriato algoritmo di *clusterizzazione*, ha visto lo sviluppo di uno *script*, che implementasse DBSCAN e altri algoritmi appositamente ideati, con lo scopo di estrarre informazioni rilevanti dal punto di vista chimico dalle configurazioni generate durante ogni simulazione del processo di adsorbimento di acqua sulla superficie di NaCl.
Lo *script* permette:

- l'importazione automatizzata ed efficiente dei file *.pdb* (*Protein Data Bank*) di *output* delle simulazioni, contenenti le configurazioni generate e salvate;
- la correzione degli errori presenti nelle configurazioni, relativi alla presenza di molecole posizionate oltre limiti della cella di simulazione;
- la clusterizzazione dei *frame* di ogni simulazione, grazie alla scelta degli opportuni parametri di DBSCAN; inoltre, lo *script* considera, nella clusterizzazione, la periodicità del sistema, non contemplata nativamente da DBSCAN;

- la classificazione dei *cluster* come “isole” o come “strati”, permettendo uno studio separato delle proprietà delle molecole d’acqua appartenenti alle due tipologie di aggregato o al rumore;
- l’ottenimento di *output* come file in formato *.csv* (*comma-separated values*), importabili in fogli di calcolo;
- la visualizzazione dei risultati dell’analisi attraverso la rappresentazione grafica opportuna, senza necessità di utilizzare *software* esterni.

Il linguaggio di programmazione scelto per lo sviluppo dello *script* è Python perché: è caratterizzato da una sintassi semplice e chiara, è un *linguaggio interpretato* (il codice può essere direttamente eseguito senza effettuare la compilazione) e presenta un grande numero di librerie utili alla computazione scientifica.

2.2.1 Importazione, manipolazione e visualizzazione dei dati

Durante ogni simulazione, le configurazioni associate ad ogni *frame* sono salvate progressivamente in un unico *file .pdb* (*Protein Data Bank*), un formato standard per i *file* contenenti coordinate atomiche tridimensionali. I *file* considerati contengono le coordinate atomiche degli atomi di ossigeno e idrogeno delle molecole d’acqua presenti nella cella di simulazione. Ogni riga del *file* di tipo ATOM è associata ad un singolo atomo, e contiene, separate in colonne, le coordinate atomiche ortogonali X,Y,Z in Å, il nome dell’elemento e ulteriori informazioni poco rilevanti per l’analisi compiuta. Alcune specifiche righe del *file* indicano l’inizio e la fine di un *frame* della simulazione.
Per importare i dati contenuti nei *file .pdb* viene utilizzata la funzione *read_csv()* di pandas[33], libreria che fornisce un insieme di utili funzioni di manipolazione dei dati. L’intero *file .pdb* viene importato e salvato in un unico *dataframe*, ovvero una matrice

[33] (McKinney, 2010)

bidimensionale in grado di contenere dati di natura eterogenea. Lo *script* successivamente ricerca e seleziona, all'interno del *dataframe*, le coordinate relative agli atomi di ossigeno e idrogeno nel primo *frame* della simulazione, salvandole in un *dataset*¸ questa volta di natura numerica e omogenea, sul quale poi vengono effettuate le operazioni di analisi. Terminati i calcoli sul singolo *frame*, viene caricato il *frame* successivo e quindi progressivamente tutti i *frame* della simulazione. I valori associati alle analisi effettuate sul singolo *frame* (ad esempio il numero di isole individuate nel sistema) vengono progressivamente aggiunti a dei vettori (*1D array*), ossia matrici monodimensionali di dati omogenei, sui quali poi vengono eseguite operazioni di media, che permettono di individuare i valori medi di determinate proprietà sull'intera simulazione eseguita a una specifica pressione di H_2O. Per la creazione e manipolazione dei *dataset* e dei vettori, nonché per la maggior parte delle operazioni di computazione numerica eseguite sugli stessi, sono state utilizzate le funzioni della libreria *open source* NumPy[34], su cui è basata la libreria pandas.

Al termine dell'analisi dei dati su una singola simulazione, lo *script* è in grado di cercare, all'interno della *directory* opportuna, e importare autonomamente il *file .pdb* della simulazione successiva, proseguendo nell'analisi. I risultati mediati ottenuti da ogni simulazione vengono progressivamente salvati in *file* in formato *.csv*.

La natura dell'algoritmo permette quindi di effettuare l'analisi dei risultati, di tutte le simulazioni, in modo completamente automatizzato, senza che sia necessario selezionare manualmente i *file* da analizzare.

Inoltre, l'utilizzo di pandas per importare l'intero *.pdb*, garantisce tempi di calcolo nettamente inferiori rispetto al caso in cui si eseguano continue operazioni di lettura e chiusura del *file* per estrarre i singoli *frame*.

Per la visualizzazione dei dati, in particolare per la produzione di grafici (2D e 3D) e istogrammi, è stata implementata la libreria Matplotlib[35].

[34] (Harris, et al., 2020)
[35] (Hunter, 2007)

2.2.2 Correzione degli errori nei dati

Le configurazioni generate da DL_MONTE durante le simulazioni presentano degli "errori", nella maggioranza dei *frame* sono presenti alcune molecole d'acqua posizionate all'esterno dei limiti della cella di simulazione.

Al fine di risolvere il problema, è stata appositamente sviluppata una funzione di correzione che permettesse di modificare le coordinate atomiche delle molecole d'acqua esterne alla scatola e di riposizionarle in un punto interno alla cella di simulazione, considerando la periodicità imposta al contorno.

Ipotizzando di rappresentare i confini della cella di simulazione come superfici di un parallelepipedo rettangolo di dimensioni 39.606 Å x 39.606 Å x 50.000 Å, l'origine del sistema di riferimento cartesiano tridimensionale è posizionata nel punto di intersezione delle diagonali del parallelepipedo.

Una molecola d'acqua i-esima nel sistema, con ossigeno di coordinate x_i y_i z_i, è definita interna alla cella di simulazione se:

- $-19.803\ Å \leq x_i \leq 19.803\ Å$.
- $-19.803\ Å \leq y_i \leq 19.803\ Å$.
- $-25.000\ Å \leq z_i \leq 25.000\ Å$.

La scelta di considerare solo la posizione dell'atomo di ossigeno è arbitraria, ipotizzando infatti di avere una molecola parzialmente esterna alla scatola, correggendone la posizione, secondo la periodicità della cella, si ottiene nuovamente una molecola parzialmente esterna alla scatola, si sceglie quindi in questi casi di mantenere sempre interni alla scatola gli atomi di ossigeno. Nel caso in cui l'atomo di ossigeno di una molecola d'acqua risulti esterno alla cella, la correzione viene ovviamente applicata anche agli atomi di idrogeno.

Di seguito è riportata una porzione di pseudocodice della funzione, considerando una molecola d'acqua *i*-esima con ossigeno di coordinate x_i y_i z_i.

se x_i > 19.803 Å:

    se [(x_i//19.803)%2]=1: $x_{icorretto}$=-19.803+(x_i%19.803)

    se [(x_i//19.803)%2]=0: $x_{icorretto}$= x_i%19.803

se x_i < -19.803 Å:

    se [(x_i//-19.803)%2]=1: $x_{icorretto}$=+19.803+(x_i%-19.803)

    se [(x_i//-19.803)%2]=0: $x_{icorretto}$= x_i%-19.803

Il controllo e l'eventuale correzione vengono poi effettuati anche sulle coordinate y_i z_i. L'operatore // è chiamato *divisione intera* ed approssima per difetto il risultato decimale di una divisione tra due valori. % è l'operatore *modulo*, il quale fornisce il resto della divisione tra due valori.

Nella Figura 21 è rappresentato un *frame* esemplificativo, contenente una molecola d'acqua prima e dopo la correzione.

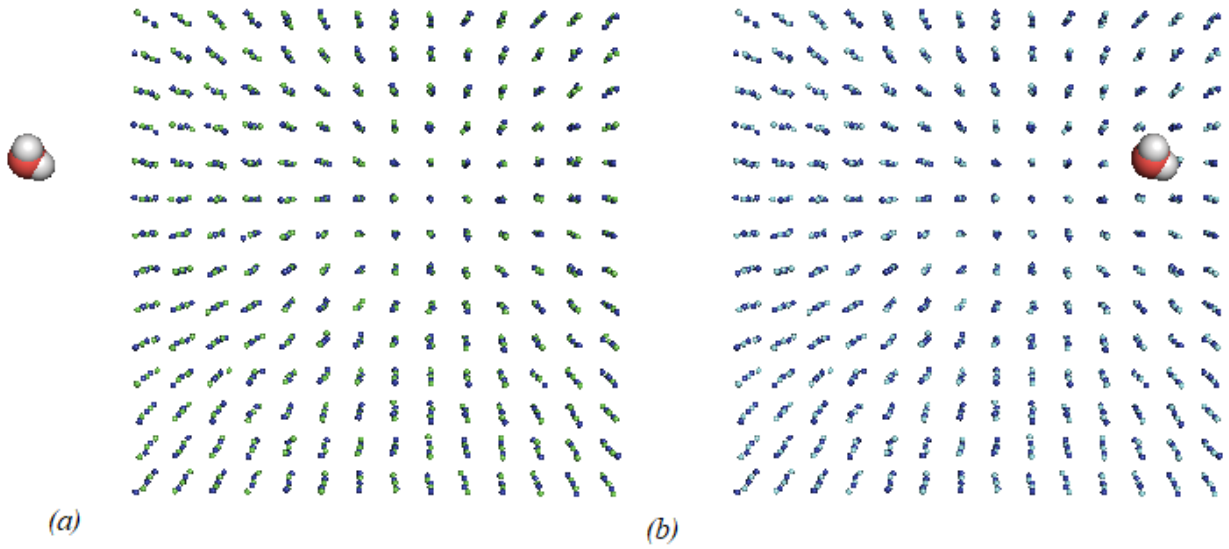

Figura 21: Vista ortogonale al piano xy di una molecola d'acqua (a) esterna alla cella di simulazione, prima della correzione e (b) interna, dopo la correzione. (O: rosso, H: bianco, Na: blu; Cl: ciano).

2.2.3 Parametri di DBSCAN e clusterizzazione in un sistema periodico

DBSCAN è stato importato tramite l'apposito pacchetto presente nella libreria *open source* per il *machine learning* scikit-learn[36].

L'algoritmo di clusterizzazione DBSCAN è particolarmente sensibile ai parametri di ingresso, la scelta dei valori appropriati risulta quindi cruciale per permettere un *clustering* efficace. I parametri scelti sono:

- *Eps* = 4;
- *minPts* = 3;
- *dist(p,q)*: distanza euclidea modificata per il sistema periodico.

Per comprendere il valore assegnato a *Eps* è necessario specificare che la clusterizzazione è stata effettuata considerando solo le coordinate atomiche degli atomi di ossigeno delle molecole d'acqua presenti nel sistema. *Eps* indica la distanza massima entro cui due oggetti siano considerabili *Eps*-vicini, nel caso specifico è legato alla distanza massima O···O per cui due molecole d'acqua adsorbite sulla superficie siano considerabili interagenti. La distanza di legame intramolecolare O-H nell'acqua liquida è circa 0.97 Å, mentre il tipico valore della distanza O···H, tra due molecole interagenti tramite legame a idrogeno nell'acqua liquida vale circa 1.75 Å. Inoltre, nello studio computazionale di Engkvist e Stone[37] è stata prodotta la funzione di distribuzione radiale a coppie $g_{OO}(r)$, considerando solo molecole d'acqua a distanza inferiore a 4.5 Å dalla superficie, nella quale si può osservare, sia ad alto che basso *coverage*, la presenza di un massimo in corrispondenza di distanza O···O uguale a 3 Å, in modo del tutto analogo all'acqua liquida. Il picco è però più largo rispetto all'acqua liquida, riflettendo un conflitto tra la formazione del legame a idrogeno e una tendenza ad attaccare gli ioni Na^+, con una conseguente distanza O···O che tende verso i 4 Å. Sulla base dei dati teorici e considerando una certa

[36] (Pedregosa, et al., 2011)
[37] (Engkvist & Stone, 2000)

flessibilità del sistema, è stato deciso arbitrariamente di considerare interagenti due molecole con distanza O···O inferiore a 4 Å.

Nel caso studiato, *MinPts* coincide con il numero di molecole d'acqua (o meglio di atomi di ossigeno) che devono essere presenti in un intorno di raggio *Eps* di un atomo di ossigeno di una molecola d'acqua, affinché questa sia considerabile *core point* di un *cluster*. Il valore del parametro rappresenta inoltre il numero minimo di molecole d'acqua che può contenere un *cluster*. La regola comune vuole che, dato un *dataset* di *D* dimensioni, si abbia $MinPts \geq D + 1$. Sulla base di un approccio *trial and error* e considerando la natura bidimensionale delle isole di molecole d'acqua adsorbite, è risultato però ottimale scegliere $MinPts = 3$.

La funzione della distanza *dist(p,q),* scelta per calcolare la distanze tra tutte le possibili coppie di atomi di ossigeno delle molecole d'acqua presenti nel sistema, è la distanza euclidea (Equazione 7). L'algoritmo DBSCAN implementato da scikit-learn calcola la distanza tra le coppie di punti presenti nel sistema sulla base del tipo di metrica scelta dall'utente, definita da un parametro *metric*, che di *default* è impostato come '*euclidean'*. DBSCAN non contempla però la presenza di condizioni periodiche al contorno, applicate al sistema che ospita l'insieme dei punti su cui viene utilizzato, il che si traduce in un calcolo scorretto delle distanze. La funzione di clusterizzazione è però in grado, impostando il parametro *metric='precomputed',* di accettare come argomento una matrice contenente le distanze tra i punti, scomposte nelle componenti *x,y* e *z*, calcolate separatamente. L'algoritmo sviluppato per calcolare la distanza corretta prevede l'ottenimento delle componenti *x*, *y* e *z* delle distanze, tra tutte le possibili coppie di atomi di ossigeno presenti nel *dataset* relativo ad un determinato *step* della simulazione, utilizzando la funzione *pdist* della libreria SciPy[38], impostando la metrica euclidea. Successivamente viene effettuato un controllo e una eventuale correzione di tutti i valori delle componenti *x,y* delle distanze. É stato arbitrariamente scelto di non considerare la periodicità della cella lungo l'asse *z*, il confine di cella si trova infatti ad una distanza di circa 19 Å dalla

[38] (Virtanen, et al., 2020)

superficie modello di NaCl, per la quale è da escludere la presenza di molecole d'acqua adsorbite interagenti tramite legame a idrogeno attraverso il confine.

È riportata una porzione di pseudocodice dell'algoritmo di controllo e correzione delle componenti delle distanze d(x) e d(y) tra una coppia di atomi di ossigeno O_i e O_j:

```
se dij(x) > ½ * 39.606 Å:
      dij(x)corretto = | dij(x) - 39.606 |
se dij(y) > ½ * 39.606 Å:
      dij(x)corretto = | dij(x) - 39.606 |
```

L'algoritmo si basa sul principio per cui, data una cella di simulazione che si ripete periodicamente nello spazio tridimensionale, se la componente x o y della distanza calcolata tra due punti *i* e *j* è maggiore di metà del valore della dimensione della cella lungo x o y, allora la distanza minore, e corretta, sarà quella tra il punto *i* e il punto *j'*, immagine di *j* nella cella adiacente (Figura 22).

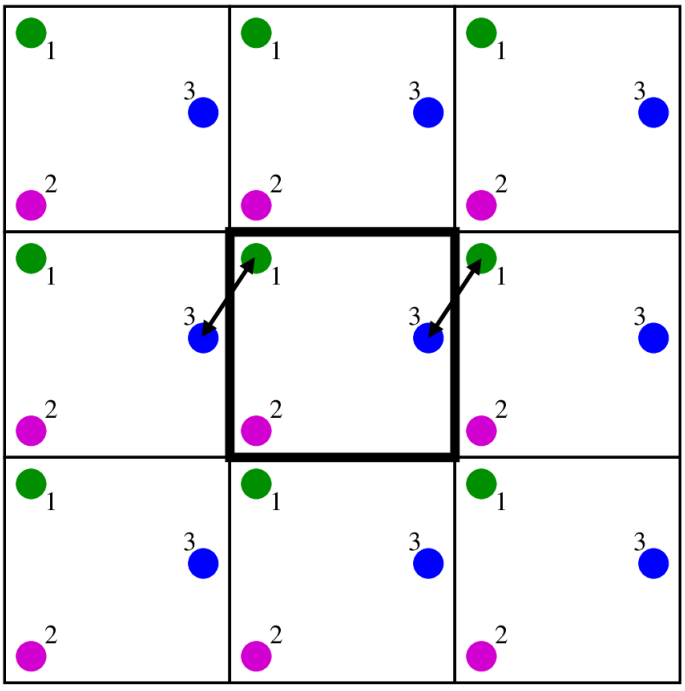

Figura 22: Condizioni periodiche al contorno. Una cella di simulazione centrale bidimensionale è circondata dalle copie di se stessa. Le frecce indicano la distanza minore tra le particelle 1 e 3.[39]

[39] (Introduction to Molecular Simulation and Statistical Thermodynamics, 2020)

L’utilizzo degli opportuni parametri di DBSCAN e della corretta funzione della distanza permettono di ottenere una clusterizzazione efficace dei *dataset* (Figura 23).

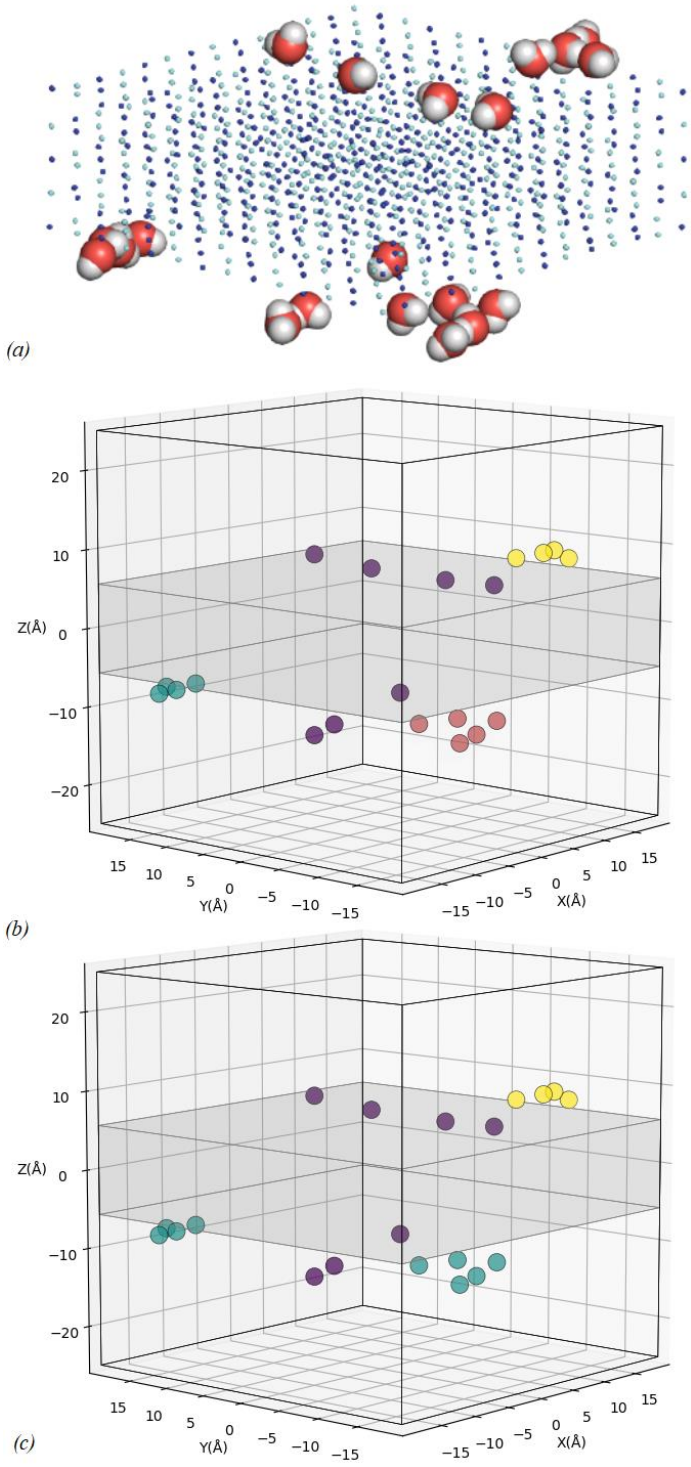

Figura 23: *(a)* un frame della simulazione (O: rosso, H: bianco, Na: blu; Cl: ciano), *(b)* relativo risultato della clusterizzazione con metrica euclidea nativa in DBSCAN (sono individuati tre *cluster*) e (*c)* risultato della clusterizzazione con funzione della distanza corretta (sono individuati due *cluster*).

2.2.4 Classificazione dei *cluster*

Partendo da un *dataset* delle coordinate atomiche tridimensionali degli atomi di ossigeno (Figura 24a) delle molecole d'acqua presenti nel sistema, DBSCAN effettua il *clustering*, sulla base della matrice delle distanze generata dall'algoritmo adibito, e restituisce come risultato un vettore contenente delle etichette, relative ad ogni atomo di ossigeno del *set* di dati (Figura 24b).

O_i	x_i	y_i	z_i
O_1	x_1	y_1	z_1
O_2	x_2	y_2	z_2
O_3	x_3	y_3	z_3
O_4	x_4	y_4	z_4
O_5	x_5	y_5	z_5
O_6	x_6	y_6	z_6
O_7	x_7	y_7	z_7
O_8	x_8	y_8	z_8
O_9	x_9	y_9	z_9
O_{10}	x_{10}	y_{10}	z_{10}

(a)

$$\begin{matrix} O_1 & O_2 & O_3 & O_4 & O_5 & O_6 & O_7 & O_8 & O_9 & O_{10} \\ labels = [-1\,, 0\,, 1\,, 0\,, 0\,, 1\,, 1\,, -1, 1\,, -1\,] \end{matrix}$$

(b)

Figura 24: Esempio di (a) *dataset* in *input* e relativo (b) *output* della clusterizzazione via DBSCAN.

Il vettore generato in *output* dall'algoritmo di clusterizzazione contiene un numero di variabili, in particolare *label* (etichette), uguale al numero di atomi di ossigeno presenti nel *dataset* in *input.* L'etichetta con indice *i* all'interno del vettore caratterizza l'atomo di ossigeno nella riga *i*-esima nella matrice delle coordinate atomiche inziale. Un valore di etichetta uguale a -1 identifica atomi di ossigeno di molecole d'acqua appartenenti a regioni di bassa densità, ovvero il rumore. Un valore di $label \geq 0$ indica invece atomi di ossigeno di molecole d'acqua appartenenti a

cluster, due molecole d'acqua appartenenti allo stesso *cluster* saranno indicate da un congruo valore di etichetta. Un valore crescente di etichetta non fornisce informazioni sulla dimensione del *cluster*, un valore uguale a 0 indica semplicemente che il primo *core point* del *cluster* specifico è stato individuato prima del primo *core point* individuato nei *cluster* con etichetta maggiore.

L'algoritmo successivamente applica la funzione *np.bincount()*, della libreria NumPy, la quale conta il numero di volte in cui una stessa etichetta è presente nel vettore *labels*. La funzione produce un vettore di *output* contenente i risultati dei conteggi posti in ordine crescente in base al valore crescente dell'etichetta. Di fatto si ottiene quindi un vettore i quali valori indicano il numero di molecole d'acqua identificate come rumore ed il numero di molecole d'acqua appartenenti ad ognuno dei *cluster* individuati (Figura 25).

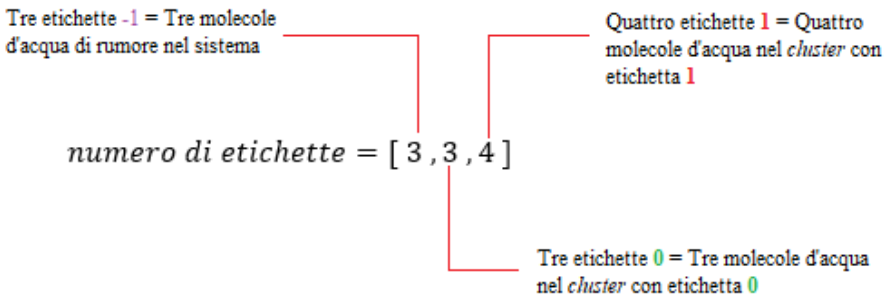

Figura 25: Vettore di *output* ottenuto dall'applicazione della funzione *np.bincount* al vettore delle etichette in Figura 24*b*.

A questo punto l'algoritmo effettua la classificazione dei *cluster* come "isole" o come "strati" in funzione del numero di molecole d'acqua presenti negli stessi. Viene classificata come "isola" un *cluster* che contiene un numero di molecole d'acqua inferiore o uguale a 70, è invece classificato come "strato" un *cluster* che presenta un numero di molecole d'acqua maggiore di 70. La scelta di 70 come discriminante è basata sul fatto che, sapendo che lo strato superficiale del modello di NaCl presenta 98 ioni Na^+, e considerando una struttura di minima energia caratterizzata da molecole d'acqua posizionate sopra gli ioni Na^+, si ha uno strato bidimensionale

monomolecolare (*monolayer*) quando in un aggregato sono presenti 98 molecole d'acqua. Scegliendo 71 come il numero minimo di molecole d'acqua necessario a definire un aggregato come "strato", si includono in questa definizione "isole" bidimensionali molto grandi, strutture *monolayer* e *multilayer*.

In Figura 26*a* si ha un *frame* di una simulazione condotta in condizioni di basso *coverage*. Il risultato della clusterizzazione (Figura 26*b*) mostra la presenza di tre *cluster* classificabili come "isole" ed un certo numero di molecole d'acqua appartenenti al rumore.

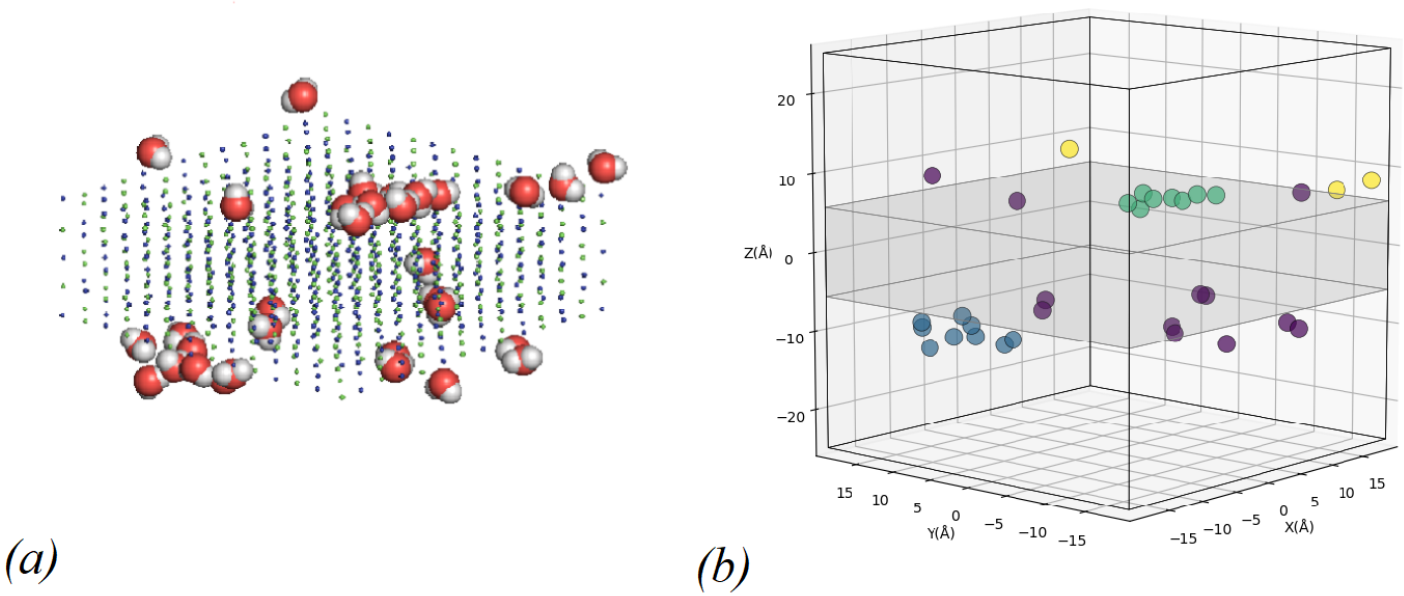

Figura 26: (a) un frame di una simulazione condotta a un valore di pressione di H_2O uguale a 7.500 matm (O: rosso, H: bianco, Na: blu; Cl: ciano). (b) relativo risultato della clusterizzazione.

La Figura 27*a* è invece un *frame* di una simulazione condotta in condizioni di alto *coverage*, nel risultato del *clustering* (Figura 27*b*) sono distinguibili due *cluster*. Entrambe le strutture adsorbite sui due *layer* superficiali del modello di NaCl sono classificabili come "strati".

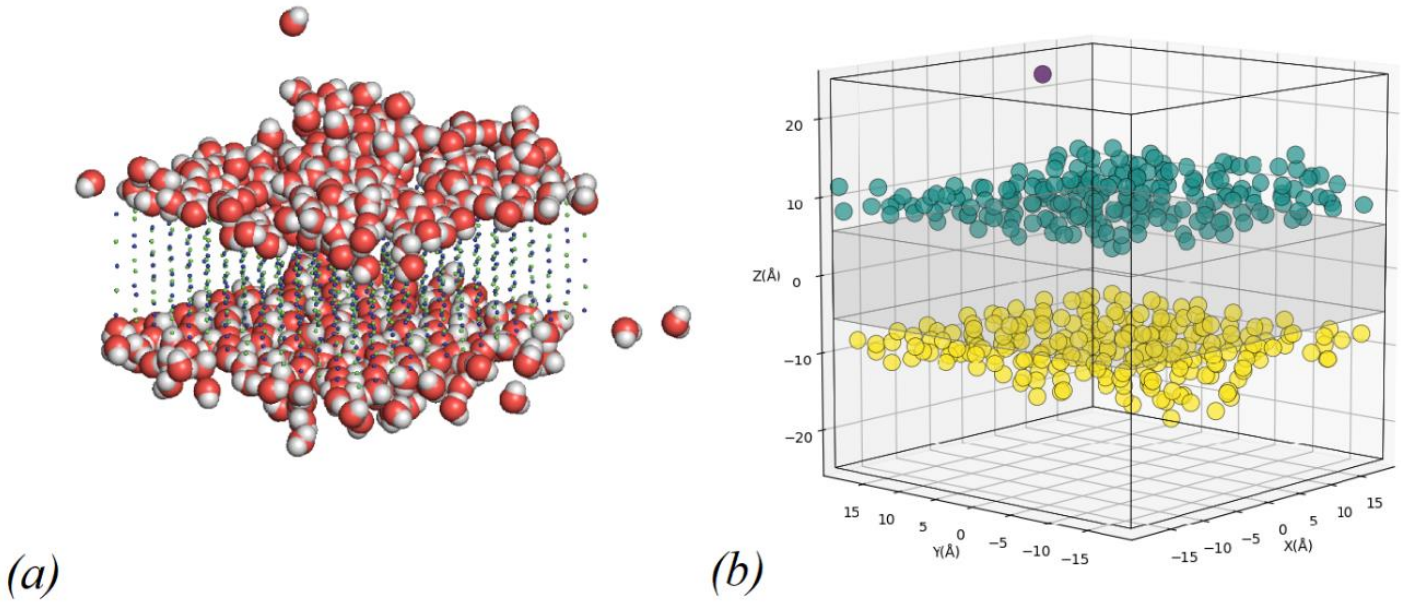

Figura 27: (a) un frame di una simulazione condotta a un valore pressione di H_2O uguale a 8.750 matm (O: rosso, H: bianco, Na: blu; Cl: ciano). (b) relativo risultato della clusterizzazione.

Dall'operazione di clusterizzazione e classificazione effettuata su un *dataset* relativo ad un singolo *frame* della simulazione si ottiene: il numero di "isole" e "strati" individuati nel sistema, il numero di molecole d'acqua appartenenti al rumore e il numero di molecole d'acqua appartenenti a ogni "isola" o struttura stratificata individuata. I valori ottenuti vengono salvati in appositi vettori, progressivamente riempiti durante l'analisi di tutti gli *step*, per poi essere mediati sull'intera simulazione.

Inoltre, sapendo che l'etichetta -1 è legata al rumore, e avendo associato a ogni altra etichetta, attraverso la classificazione, una tipologia di *cluster* specifica, è possibile risalire alle coordinate delle singole molecole d'acqua nel *dataset* iniziale e generare tre ulteriori *dataset* separati, contenenti le coordinate degli atomi di idrogeno e ossigeno delle molecole d'acqua appartenenti al rumore, alle "isole" e agli "strati", al fine di studiarne separatamente l'orientazione.

2.2.5 Studio dell'orientazione delle molecole d'acqua adsorbite in funzione della distanza dalla superficie

Al fine di investigare come la superficie influenzi le molecole d'acqua nello strato adsorbito è stata studiata la distribuzione orientazionale di H_2O in funzione della distanza della superficie. In particolare, l'orientazione di una molecola d'acqua è stata definita come il coseno dell'angolo θ, formato dal vettore momento dipolare di una molecola d'acqua ed il vettore normale alla superficie di NaCl.

In funzione del fatto che la molecola sia adsorbita sulla superficie superiore o inferiore del modello di NaCl sono state considerati due vettori normale unitari, con uguale direzione ma verso opposto, rispettivamente $\overrightarrow{v_n} = (0,0,1)$ e $\overrightarrow{v_n} = (0,0,-1)$. Un valore di $cos\,(\theta)$ tendente a 1 (Figura 28*a*) sarà correlato a molecole d'acqua posizionate sopra i cationi Na^+, con gli atomi di idrogeno direzionati lontano dalla superficie. Un valore di $cos\,(\theta)$ tendente a -1 (Figura 28*b*) indicherà molecole d'acqua con vettore momento dipolare orientato verso la superficie, sono molecole prevalentemente legate tramite legame a idrogeno agli anioni Cl^-.

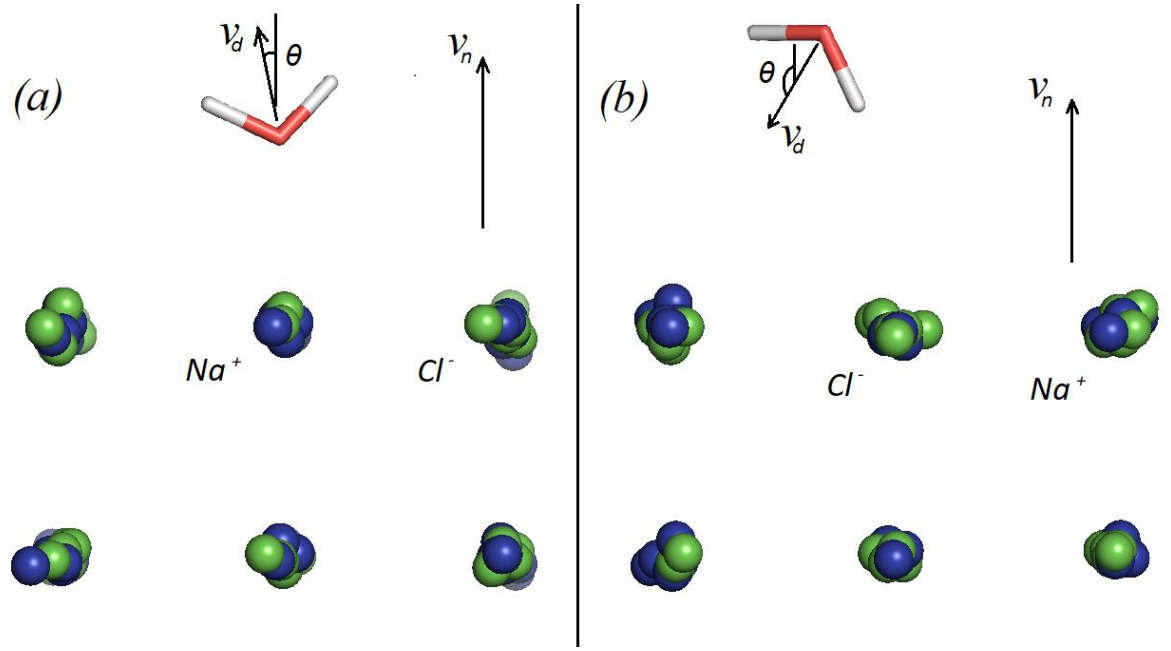

Figura 28: Molecola d'acqua adsorbita sulla superficie modello di NaCl con (*a*) valore di *cos(θ)* tendente a 1 e (*b*) valore di *cos(θ)* tendente a -1. (O: rosso, H: bianco, Na: blu; Cl: ciano; v_d : vettore momento dipolare di H_2O; v_n : vettore normale alla superficie).

Al fine di poter studiare come $cos\ (\theta)$ varia allontanandosi dalla superficie, lo *script* classifica le molecole d'acqua in funzione della distanza dalla stessa. I tre *dataset* precedentemente ottenuti, in cui sono salvate le coordinate atomiche delle molecole d'acqua, vengono divisi in ulteriori *subset*. Ogni *subset* contiene le coordinate degli atomi di idrogeno e ossigeno delle molecole d'acqua la cui distanza dalla superficie è compresa in un certo intervallo di valori di *z*, i cui estremi differisco per un valore di 0.5 Å. Ad esempio, il primo *subset* sarà associato alle molecole d'acqua caratterizzate da una distanza dalla superficie compresa tra 0 e 0.5 Å, il secondo tra 0.5 e 1 Å e così di seguito. Al termine dell'analisi orientazionale effettuata sul singolo *frame*, i valori di $cos(\theta)$ calcolati per un determinato *subset* verranno depositati in un vettore, il quale viene progressivamente riempito con i risultati ottenuti dai *frame* successivi. Al termine del processo, saranno disponibili vettori contenenti i valori di $cos(\theta)$ di tutte le molecole d'acqua, in certo intervallo di distanza dalla superficie, salvati nell'intera simulazione, su cui poi viene effettuato uno studio statistico.

Capitolo 3

3. Risultati

In questo capitolo verranno presentati i risultati ottenuti dalle analisi, svolte durante il periodo di tirocinio, delle simulazioni. Verranno prima riportati i dati relativi ottenuti dal *clustering* e dalla classificazione dei *cluster*, seguiti dai risultati dello studio orientazionale sulle molecole d'acqua in funzione della distanza dalla superficie di NaCl.

3.1 *Clustering* e classificazione degli aggregati

Partendo dai risultati di ogni simulazione, condotta a una specifica pressione di H_2O, sono stati calcolati, per ogni valore di pressione:

- il numero medio ($\bar{n}°$) di isole individuate per *frame*;
- il numero medio ($\bar{n}°$) di molecole d'acqua di cui è composta un'isola;
- il numero medio ($\bar{n}°$) di superfici interamente ricoperte individuate per *frame*. In particolare, si fa riferimento al numero medio di "strati" individuati, quindi di strutture che comprendono grandi isole (tendenti a *monolayer*), strutture *monolayer* e strutture *multilayer*. Ad esempio, un valore uguale a 1 indicherà che mediamente solo una delle due superfici di NaCl è completamente ricoperta;
- il numero medio ($\bar{n}°$) di molecole d'acqua di cui è composto uno "strato", ovvero il numero medio di molecole per superficie interamente ricoperta;
- l'errore standard della media ($S_{\bar{X}}$) per ognuna delle medie calcolate, ottenuto come rapporto tra la deviazione standard associata al *set* di valori e la radice quadrata del numero di valori mediati: $S_{\bar{X}} = \frac{s}{\sqrt{n}}$.

I dati della simulazione alla pressione di 8.000 matm sono disponibili solo per la terza replica. Per ciascuna replica sono stati analizzati separatamente i primi 300 milioni di step e i successivi 300 milioni di step.

Pressione di H_2O (matm)	$\bar{n}°$ di isole per *frame*	$S_{\bar{X}}$	$\bar{n}°$ di H_2O per isola	$S_{\bar{X}}$	$\bar{n}°$ di superfici interamente coperte	$S_{\bar{X}}$	$\bar{n}°$ di H_2O per superficie interamente coperta	$S_{\bar{X}}$
7.500	1.74	0.02	6.27	0.06	0.00	0.00	0.00	0.00
7.750	0.77	0.01	7.67	0.13	1.00	0.00	167.01	0.58
7.875	2.06	0.02	7.95	0.09	0.00	0.00	0.00	0.00
8.125	0.00	0.00	4.62	0.73	2.00	0.00	176.70	0.32
8.250	0.00	0.00	3.50	0.35	2.00	0.00	180.60	0.25
8.375	0.00	0.00	3.00	0.00	2.00	0.00	186.12	0.26
8.500	0.00	0.00	0.00	0.00	2.00	0.00	178.56	0.20
8.625	0.00	0.00	0.00	0.00	2.00	0.00	190.00	0.21
8.750	0.00	0.00	0.00	0.00	2.00	0.00	226.99	0.47
8.875	0.00	0.00	0.00	0.00	2.00	0.00	285.69	0.69
9.000	0.00	0.00	0.00	0.00	2.00	0.00	375.38	0.64

Tabella 2: Prima replica, primi 300 milioni di step. Risultati del clustering e della classificazione.

Pressione di H_2O (matm)	$\bar{n}°$ di isole per *frame*	$S_{\bar{X}}$	$\bar{n}°$ di H_2O per isola	$S_{\bar{X}}$	$\bar{n}°$ di superfici interamente coperte	$S_{\bar{X}}$	$\bar{n}°$ di H_2O per superficie interamente coperta	$S_{\bar{X}}$
7.500	1.60	0.02	6.78	0.11	0.00	0.00	0.00	0.00
7.750	1.60	0.02	9.66	0.17	0.31	0.01	129.18	1.14
7.875	1.89	0.02	10.98	0.16	0.29	0.01	147.04	1.12
8.125	0.00	0.00	3.00	0.00	2.00	0.00	178.48	0.20
8.250	0.00	0.00	3.00	0.00	2.00	0.00	180.12	0.22
8.375	0.00	0.00	0.00	0.00	2.00	0.00	193.98	0.26
8.500	0.00	0.00	0.00	0.00	2.00	0.00	194.38	0.22
8.625	0.00	0.00	0.00	0.00	2.00	0.00	186.76	0.18
8.750	0.00	0.00	0.00	0.00	2.00	0.00	203.95	0.22
8.875	0.00	0.00	0.00	0.00	2.00	0.00	379.85	1.07
9.000	0.00	0.00	0.00	0.00	2.00	0.00	399.38	0.54

Tabella 3: Prima replica, ultimi 300 milioni di step. Risultati del clustering e della classificazione.

Pressione di H_2O (matm)	$\bar{n}°$ di isole per *frame*	$S_{\bar{X}}$	$\bar{n}°$ di H_2O per isola	$S_{\bar{X}}$	$\bar{n}°$ di superfici interamente coperte	$S_{\bar{X}}$	$\bar{n}°$ di H_2O per superficie interamente coperta	$S_{\bar{X}}$
7.500	1.53	0.02	6.55	0.09	0.00	0.00	0.00	0.00
7.750	1.96	0.02	8.27	0.10	0.00	0.00	0.00	0.00
7.875	1.81	0.02	7.05	0.09	0.00	0.00	0.00	0.00
8.125	0.59	0.02	24.05	0.47	1.58	0.01	175.32	0.38
8.250	0.00	0.00	3.00	0.00	2.00	0.00	172.44	0.25
8.375	0.00	0.00	3.17	0.15	2.00	0.00	186.09	0.25
8.500	0.00	0.00	3.25	0.22	2.00	0.00	176.51	0.22
8.625	0.00	0.00	0.00	0.00	2.00	0.00	186.73	0.18
8.750	0.00	0.00	0.00	0.00	2.00	0.00	219.91	0.38
8.875	0.00	0.00	3.00	0.00	2.00	0.00	199.02	0.21
9.000	0.00	0.00	0.00	0.00	2.00	0.00	388.81	0.70

Tabella 4: Seconda replica, primi 300 milioni di step. Risultati del clustering e della classificazione.

Pressione di H_2O (matm)	$\bar{n}°$ di isole per *frame*	$S_{\bar{X}}$	$\bar{n}°$ di H_2O per isola	$S_{\bar{X}}$	$\bar{n}°$ di superfici interamente coperte	$S_{\bar{X}}$	$\bar{n}°$ di H_2O per superficie interamente coperta	$S_{\bar{X}}$
7.500	1.60	0.02	7.30	0.09	0.00	0.00	0.00	0.00
7.750	2.01	0.02	6.78	0.07	0.00	0.00	0.00	0.00
7.875	2.05	0.02	8.50	0.10	0.00	0.00	0.00	0.00
8.125	0.00	0.00	4.00	0.00	2.00	0.00	182.94	0.23
8.250	0.00	0.00	3.25	0.22	2.00	0.00	175.56	0.17
8.375	0.00	0.00	3.22	0.21	2.00	0.00	183.87	0.23
8.500	0.00	0.00	0.00	0.00	2.00	0.00	185.72	0.18
8.625	0.00	0.00	3.50	0.31	2.00	0.00	189.13	0.24
8.750	0.00	0.00	0.00	0.00	2.00	0.00	212.15	0.27
8.875	0.00	0.00	0.00	0.00	2.00	0.00	226.34	0.28
9.000	0.00	0.00	0.00	0.00	2.00	0.00	384.67	0.36

Tabella 5: Seconda replica, ultimi 300 milioni di step. Risultati del clustering e della classificazione.

Pressione di H_2O (matm)	$\bar{n}$° di isole per *frame*	$S_{\bar{X}}$	$\bar{n}$° di H_2O per isola	$S_{\bar{X}}$	$\bar{n}$° di superfici interamente coperte	$S_{\bar{X}}$	$\bar{n}$° di H_2O per superficie interamente coperta	$S_{\bar{X}}$
7.500	1.48	0.02	5.34	0.05	0.00	0.00	0.00	0.00
7.750	2.19	0.02	7.88	0.08	0.00	0.00	0.00	0.00
7.875	0.95	0.02	18.98	0.41	0.90	0.01	137.50	0.82
8.000	1.12	0.02	24.47	0.37	1.07	0.00	166.56	0.52
8.125	0.37	0.01	43.92	0.72	1.72	0.01	157.60	0.38
8.250	0.00	0.00	3.00	0.00	2.00	0.00	197.36	0.40
8.375	0.00	0.00	3.00	0.00	2.00	0.00	198.06	0.31
8.500	0.00	0.00	0.00	0.00	2.00	0.00	203.02	0.35
8.625	0.00	0.00	0.00	0.00	2.00	0.00	193.83	0.22
8.750	0.00	0.00	0.00	0.00	2.00	0.00	204.74	0.36
8.875	0.00	0.00	3.00	0.00	2.00	0.00	227.02	0.43
9.000	0.00	0.00	0.00	0.00	2.00	0.00	259.20	0.45

Tabella 6: Terza replica, primi 300 milioni di step. Risultati del clustering e della classificazione. Sono presenti i dati della simulazione a 8.000 matm, non disponibili per le altre repliche.

Pressione di H_2O (matm)	$\bar{n}$° di isole per *frame*	$S_{\bar{X}}$	$\bar{n}$° di H_2O per isola	$S_{\bar{X}}$	$\bar{n}$° di superfici interamente coperte	$S_{\bar{X}}$	$\bar{n}$° di H_2O per superficie interamente coperta	$S_{\bar{X}}$
7.500	1.57	0.02	6.21	0.07	0.00	0.00	0.00	0.00
7.750	1.73	0.02	6.95	0.08	0.00	0.00	0.00	0.00
7.875	0.78	0.02	5.88	0.09	1.00	0.00	173.75	0.24
8.000	1.05	0.02	7.24	0.12	1.00	0.00	177.96	0.32
8.125	0.00	0.00	4.50	0.35	2.00	0.00	181.53	0.19
8.250	0.00	0.00	3.25	0.22	2.00	0.00	176.49	0.24
8.375	0.00	0.00	3.00	0.00	2.00	0.00	182.98	0.22
8.500	0.00	0.00	0.00	0.00	2.00	0.00	189.64	0.20
8.625	0.00	0.00	3.00	0.00	2.00	0.00	193.07	0.33
8.750	0.00	0.00	3.00	0.00	2.00	0.00	216.87	0.44
8.875	0.00	0.00	0.00	0.00	2.00	0.00	204.99	0.29
9.000	0.00	0.00	0.00	0.00	2.00	0.00	278.07	0.72

Tabella 7: Terza replica, ultimi 300 milioni di step. Risultati del clustering e della classificazione. Sono presenti i dati della simulazione a 8.000 matm, non disponibili per le altre repliche.

ISOLE

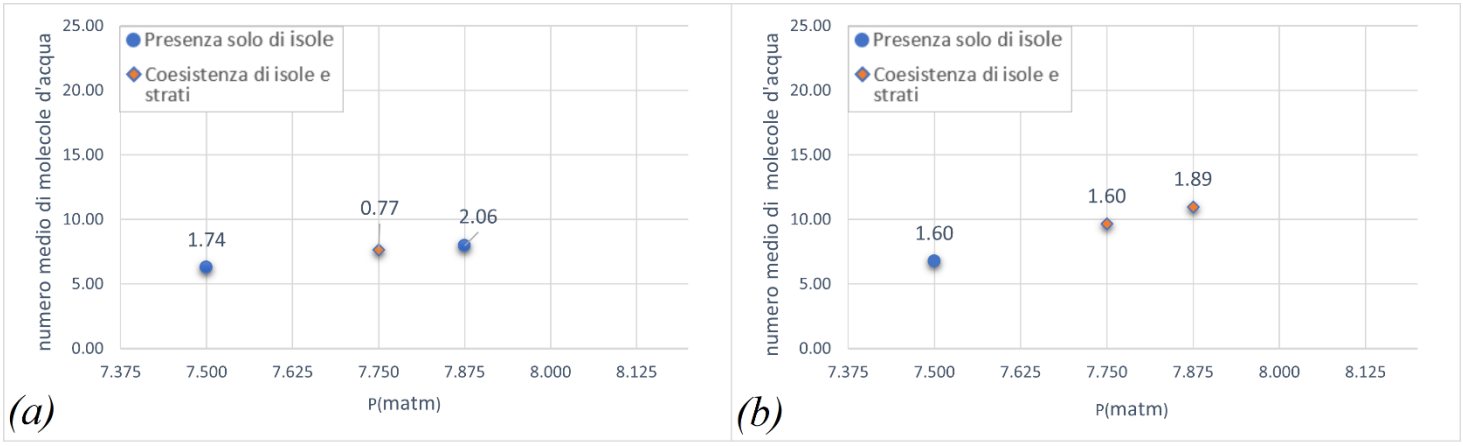

Figura 29: Replica 1. Grafico del numero medio di molecole d'acqua in un'isola in funzione della pressione di H_2O. Le etichette indicano il numero medio di isole per *frame*. (a) Primi 300 milioni di *step* e (b) ultimi 300 milioni di *step*.

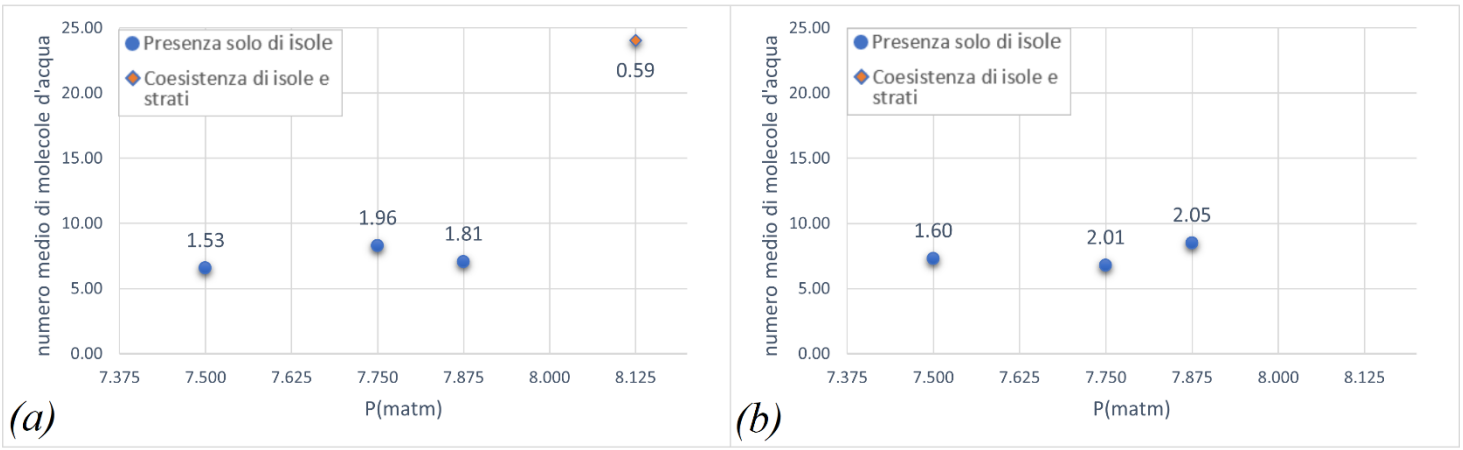

Figura 30: Replica 2. Grafico del numero medio di molecole d'acqua in un'isola in funzione della pressione di H_2O. Le etichette indicano il numero medio di isole per frame. (a) Primi 300 milioni di step e (b) ultimi 300 milioni di step.

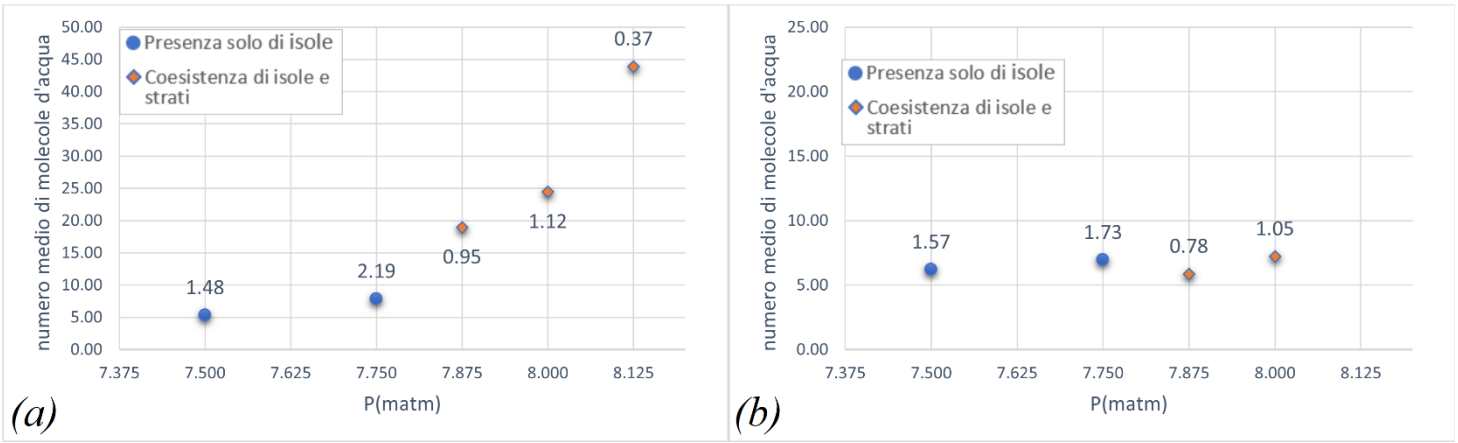

Figura 31: Replica 3. Grafico del numero medio di molecole d'acqua in un'isola in funzione della pressione di H_2O. Le etichette indicano il numero medio di isole per frame. (a) Primi 300 milioni di step e (b) ultimi 300 milioni di step.

STRATI

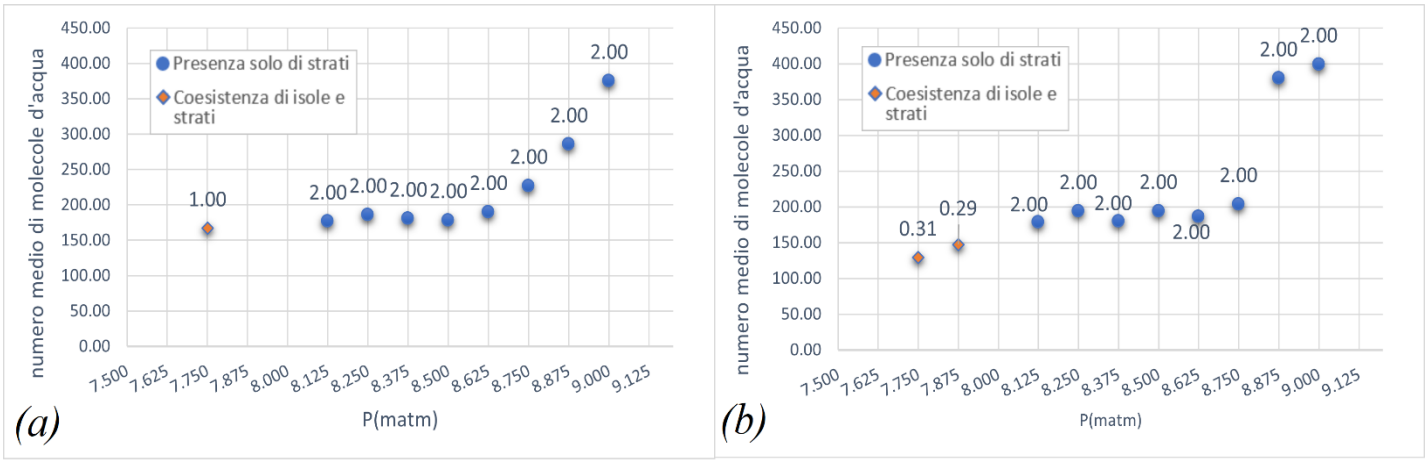

Figura 32: Replica 1. Grafico del numero medio di molecole d'acqua in uno strato in funzione della pressione di H_2O. Le etichette indicano il numero medio di isole per frame. (a) Primi 300 milioni di step e (b) ultimi 300 milioni di step.

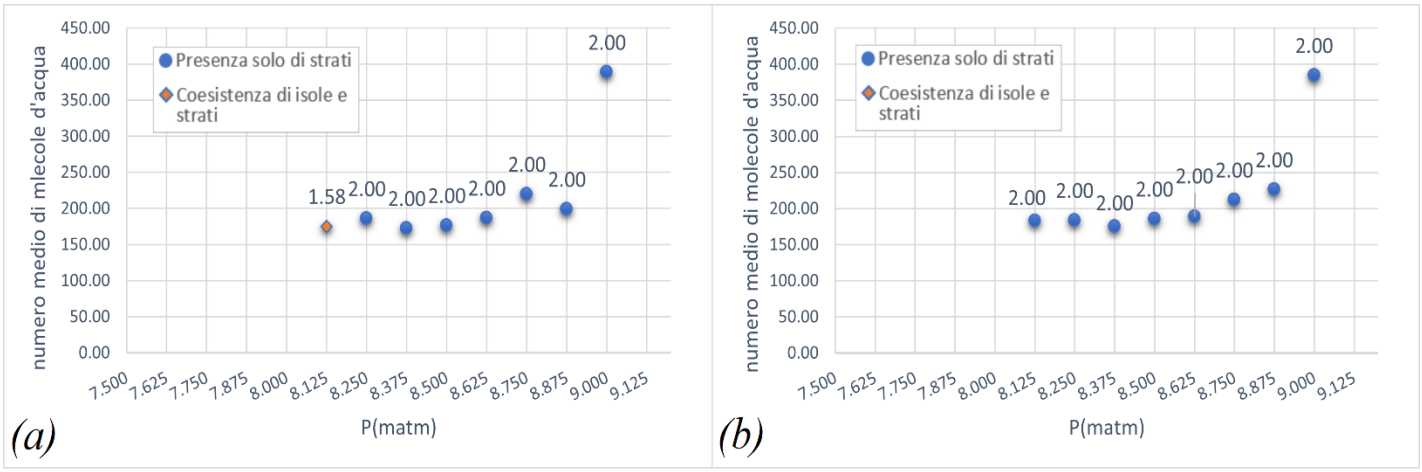

Figura 33: Replica 2. Grafico del numero medio di molecole d'acqua in uno strato in funzione della pressione di H_2O. Le etichette indicano il numero medio di isole per *frame*. (a) Primi 300 milioni di *step* e (b) ultimi 300 milioni di *step*.

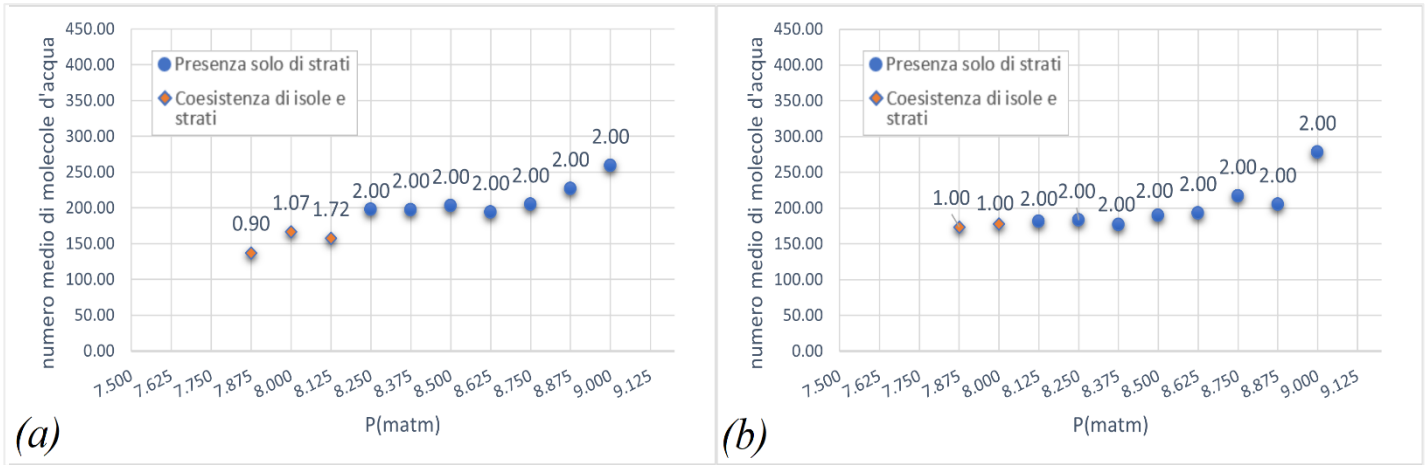

Figura 34: Replica 3. Grafico del numero medio di molecole d'acqua in uno strato in funzione della pressione di H_2O. Le etichette indicano il numero medio di isole per frame. (a) Primi 300 milioni di step e (b) ultimi 300 milioni di step.

Discussione dei risultati:

I risultati ottenuti dalla *cluster analysis*, riportati nelle Tabelle 2-7 e rappresentati nei grafici (Figure 29-34), rispecchiano perlopiù quanto atteso sulla base delle ipotesi fatte nello studio sperimentale di Foster e Ewing.

I dati relativi alle simulazioni condotte in condizioni di basso *coverage* ($\theta \leq 0.5$), quindi per pressioni di H_2O comprese tra 7.500 e 7.875 matm, verificano la presenza costante di isole nel sistema, in particolare un numero medio che oscilla tra circa 1.5 e 2 isole per *frame*, di piccole dimensioni (6-8 molecole d'acqua per isola). Il dato è coerente con la frequenza di assorbimento IR misurata nello studio sperimentale, legata alla presenza di strutture aggregate bidimensionali caratterizzate da legami a idrogeno laterali, intermedia tra quella dell'acqua in fase gas e liquida. Un risultato inatteso è la presenza di coesistenza tra strutture stratificate e isole in alcune delle simulazioni a basso *coverage*. In particolare, nei primi 300 milioni di *step* della prima replica della simulazione a 7.750 matm (Figura 32a) e per tutti i 600 milioni di *step* della terza replica della simulazione a 7.875 matm (Figura 34a e 34b), si osserva la formazione di uno strato permanente adsorbito su una delle due superfici di NaCl. Il numero di molecole presenti mediamente nello strato (circa 150) manifesta la presenza di una struttura intermedia tra *monolayer* e *bilayer*. Alla luce di questo risultato è opportuno includere la pressione di 7.875 matm nella regione di transizione, per la quale erano stati inizialmente considerati valori di pressione compresi tra 8.000 e 8.250 matm. Nella regione di transizione (facendo riferimento al *range* originale di 8.000-8.250 matm) i risultati sono diversificati in funzione della replica considerata. Per tutti i 600 milioni di *step* delle prime repliche delle simulazioni a 8.125 e 8.250 matm (Figura 32) si ha assenza di isole e completa copertura delle due superfici di NaCl, con strutture stratificate composte da circa 180 molecole, coerenti con una struttura tendente al *bilayer*.

Per gli stessi valori di pressione, la seconda replica esprime una congruenza con quanto detto per le prime repliche. L'unica differenza è osservabile nei primi 300 milioni di step della seconda replica alla pressione di 8.125 matm: si hanno

mediamente 1.58 (Figura 33a) superfici completamente ricoperte e la presenza media di 0.59 isole (Figura 30a) di medie dimensioni (25 molecole d'acqua). I valori sembrano indicare uno stato per cui una delle due superfici di NaCl risulta costantemente ricoperta, mentre l'altra presenta inizialmente uno stato di *submonolayer* (Figura 35) e solo successivamente viene completamente coperta. L'ipotesi è coerente con il fatto che negli ultimi 300 milioni di *step* (Figura 33b) si hanno entrambe le superfici costantemente ricoperte.

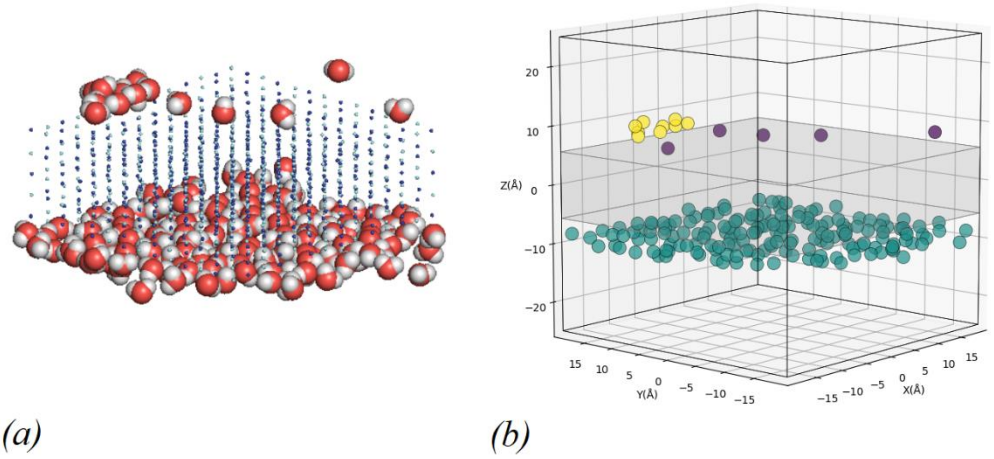

Figura 35: Coesistenza di un'isola e di uno strato in (a) un frame di una simulazione condotta a un valore pressione di H_2O uguale a 7.750 matm (O: rosso, H: bianco, Na: blu; Cl: ciano). (b) relativo risultato della clusterizzazione.

Per la terza replica (Figure 31 e 34) sono disponibili i risultati anche per la pressione di 8.000 matm, coerenti a quanto detto per la simulazione condotta a 7.875 matm, si ha la formazione di uno strato permanente su una delle due superfici e di un'isola di dimensioni variabili sull'altra.

I valori ottenuti dai primi 300 milioni di *step* alla pressione di 8.125 matm evidenziano la presenza di una grande isola (composta mediamente da 45 molecole d'acqua) su una superfice e di uno strato sulla seconda superficie e, dato il piccolo valore medio di isole individuate per *frame* (0.37), probabilmente questo stato interessa meno della metà degli *step* campionati. Nella restante frazione della prima parte di *step* e in tutti gli ultimi 300 milioni di *step* si manifesta una copertura completa di entrambe le superfici di NaCl.

I risultati relativi a simulazioni condotte nella regione ad alto *coverage*, quindi per pressioni di H_2O comprese tra 8.375 e 9.000 matm, sono coerenti nelle tre repliche. Le due superfici di NaCl sono costantemente completamente ricoperte. La dimensione degli strati non varia per valori di pressione compresi tra 8.375 e 8.625 matm, come atteso sulla base dell'isoterma di adsorbimento teorica in Figura 13, che in questo intervallo di pressioni è caratterizzata da una regione di *plateau* della curva. In particolare, ogni struttura stratificata è di *bilayer*, accoglie mediamente circa 200 molecole d'acqua. Dalla pressione di 8.750 matm la dimensione degli strati inizia a crescere. Nella prima e nella seconda replica (Figure 32 e 33) della simulazione condotta a 9.000 matm, ovvero l'ultima pressione campionata prima della deliquescenza, ognuno dei due strati adsorbiti conta circa 400 molecole d'acqua, si ipotizza una struttura formata da quattro *layer* sovrapposti. Nella terza replica si raggiunge un numero di molecole d'acqua compreso all'incirca tra 250 e 300 per strato.

Riassumendo, i risultati confermano le strutture dello strato adsorbito ipotizzate nello studio sperimentale di Foster e Ewing. Però, contrariamente a quanto ipotizzato nello studio, la coesistenza di isole bidimensionali e strutture stratificate non interessa solamente la regione di transizione. Il fenomeno risulta osservabile per pressioni di H_2O vicine al limite superiore della regione a basso *coverage* e nella prima parte dell'intervallo di pressioni caratteristico della regione di transizione. Il risultato ha portato a una nuova definizione delle regioni dell'isoterma di adsorbimento teorica: la regione a basso *coverage* è quindi considerata per $\theta < 0.4$ e per le pressioni di 7.500 e 7.750 matm, mentre la regione di transizione per $0.4 \leq \theta \leq 1.8$ e pressione compresa tra 7.875 e 8.250 matm.

Nonostante alcune differenze, l'andamento del numero di isole e strati individuati, e della dimensione dei *cluster*, in funzione della pressione di H_2O, risulta piuttosto coerente nelle tre repliche effettuate per ogni simulazione. Il fatto di aver ottenuto risultati molto simili dalle tre repliche è indice di una buona riproducibilità delle simulazioni.

3.2 Orientazione delle molecole d'acqua in funzione della distanza dalla superficie di NaCl

Al fine di investigare come la superficie influenzi l'orientazione delle molecole d'acqua nello strato adsorbito è stata studiata la dipendenza del coseno dell'angolo θ, formato dal vettore momento dipolare di una molecola d'acqua ed il vettore normale alla superficie, in funzione della distanza dalla superficie di NaCl.

Per ogni simulazione è stato ottenuto:

- Un grafico del $cos\,(\theta)$ medio in funzione della classe di *cluster* di appartenenza e della distanza dalla superficie. Nelle figure mostrate nella presente sezione, l'intensità del colore associato a ogni un punto del grafico è proporzionale al numero di valori mediati. Ogni punto del grafico, quindi ogni valore medio di $cos\,(\theta)$, sarà ottenuto come media delle orientazioni di tutte le molecole d'acqua che nel corso dell'analisi dei *frame* sono state classificate come appartenenti ad un determinato "settore di distanza" dalla superficie di NaCl. Ogni "settore di distanza" include molecole d'acqua che presentano un valore di distanza dalla superficie compreso tra due estremi separati da uno *step* di 0.5 Å. Ogni valore medio sarà quindi associato ad un valore di distanza dalla superficie, sull'asse delle ascisse del grafico, che indicherà il limite superiore del "settore di distanza". Ad esempio, un valore di $cos\,(\theta)$, associato ad una distanza di 0.5 Å dalla superficie, è media delle orientazioni di tutte le molecole d'acqua caratterizzate da una distanza dalla superficie compresa tra 0 e 0.5 Å.
- Per ogni "settore di distanza" associato ad una certa classe di *cluster*, viene generato un istogramma, il quale rappresenta la distribuzione delle orientazioni nell'intera simulazione. L'istogramma viene ottenuto andando a dividere il *range* di valori di $cos\,(\theta)$, che vanno da -1 a +1, in intervalli di 0.2 unità. Gli intervalli (sull'asse delle x) definiscono la base di rettangoli adiacenti, la cui altezza è la frequenza percentuale associata al numero

orientazioni individuate nell'intervallo specifico. Le frequenze mostrate sono quindi frequenze relative, ottenute dal rapporto tra il numero di molecole d'acqua per cui è stata determinata un'orientazione coerente con un certo intervallo, ed il numero totale di molecole d'acqua appartenenti al "settore di distanza".

L'istogramma è quindi normalizzato, dalla somma delle altezze di tutti i rettangoli si ha una percentuale del 100%.

Per un fine esemplificativo, quanto di confronto con gli istogrammi relativi alle simulazioni, è riportato in Figura 36 l'istogramma relativo a un sistema contenente solo acqua pura *bulk*, caratterizzato da una distribuzione discreta uniforme.

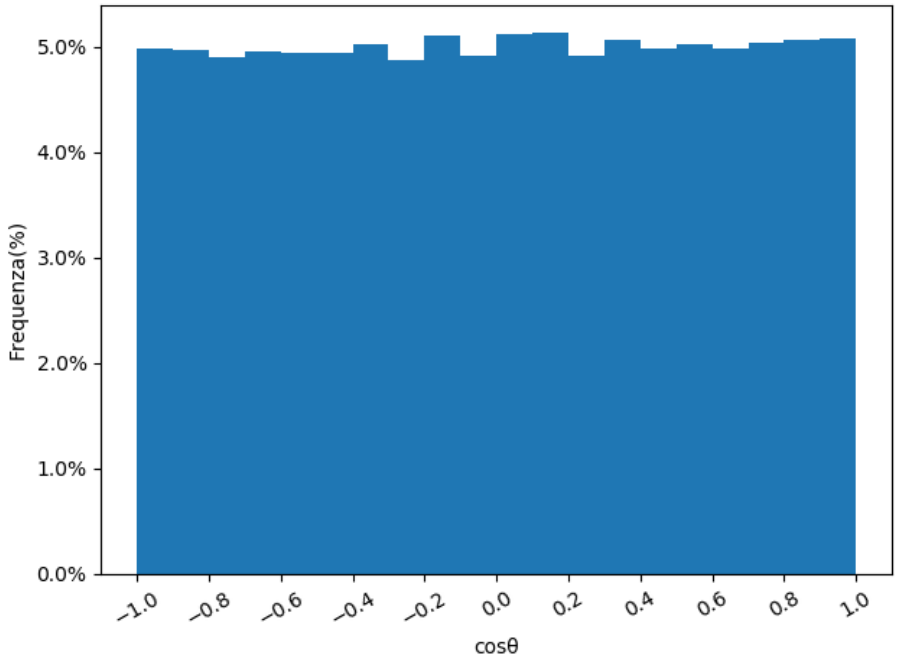

Figura 36: Istogramma relativo alla distribuzione uniforme delle orientazioni delle molecole d'acqua in un sistema contenente solo acqua *bulk*.

Dato l'elevato numero di grafici ottenuti, ne vengono di seguito proposti alcuni relativi a simulazioni rappresentative delle diverse regioni di *coverage*, a cui sono associati gli istogrammi delle "sezioni di distanza" più rilevanti.

- Simulazione condotta in condizioni di basso *coverage*: primi 300 milioni di *step* della seconda replica alla pressione di H_2O di 7.500 matm.

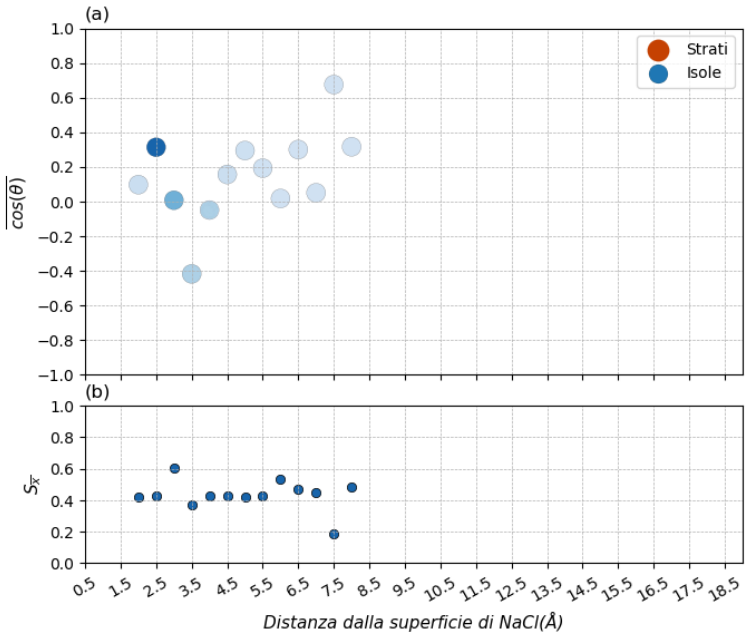

Figura 37: Grafico del (a) valore medio di cos(θ), e del (b) relativo errore standard della media, in funzione della distanza *d* dalla superficie. L'intensità del colore è proporzionale al numero di valori mediati. Primi 300 milioni di *step* della seconda replica alla pressione di H_2O di 7.500 matm.

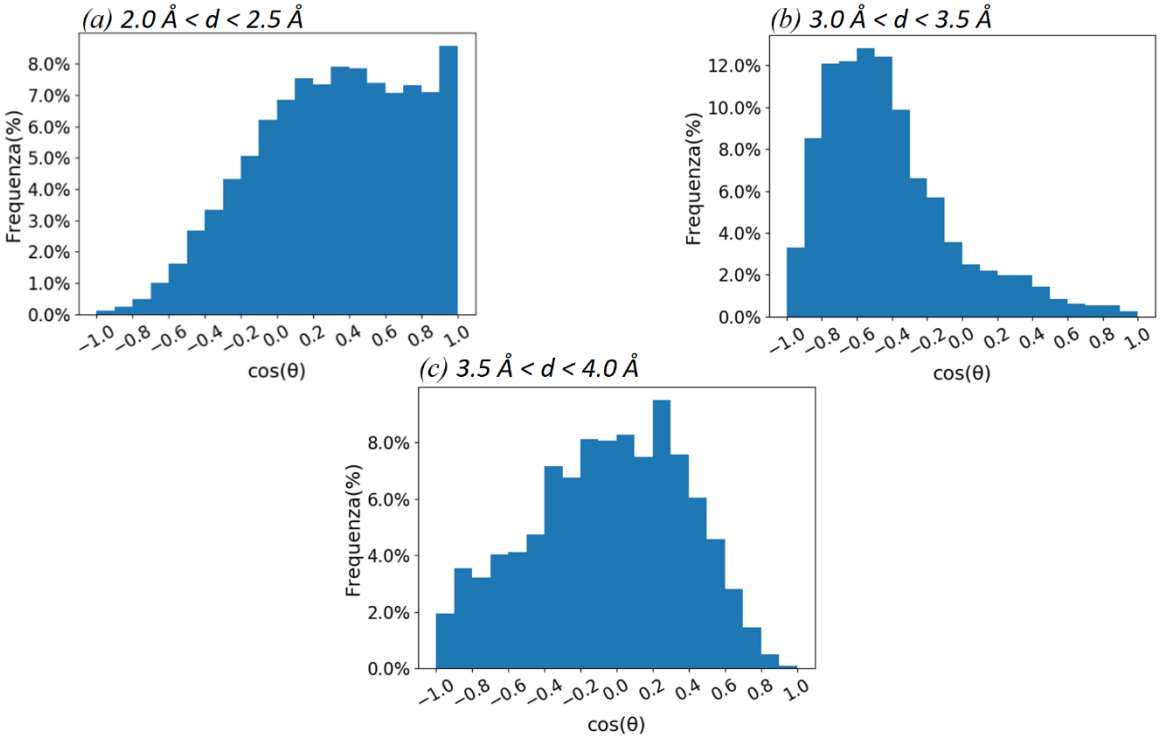

Figura 38: ISOLE. Istogrammi relativi alla distribuzione delle orientazioni nei "settori di distanza" più rilevanti: *(a) 2.0 Å < d < 2.5 Å; (b) 3.0 Å < d < 3.5 ; (c) 3.5 Å < d < 4.0 Å;*

- Simulazione condotta nella regione di transizione: primi 300 milioni di *step* della seconda replica alla pressione di H_2O di 8.125 matm.

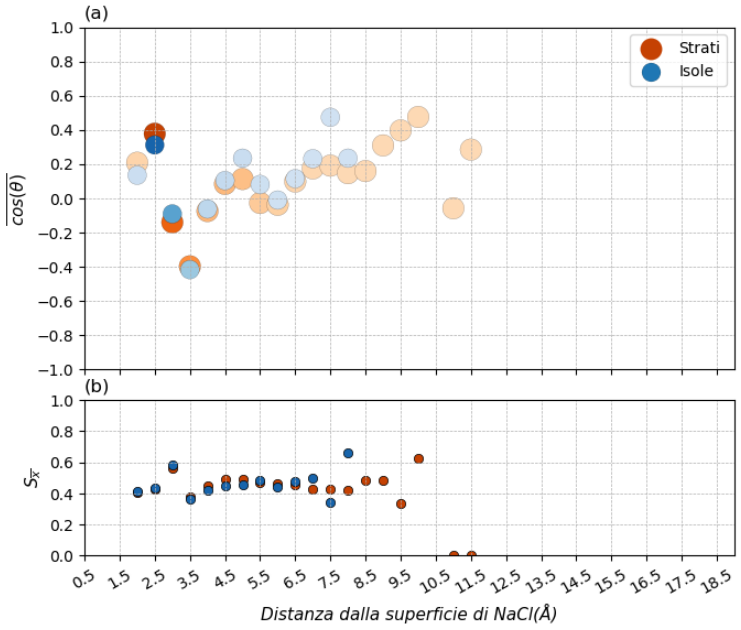

Figura 39: Grafico del (a) valore medio di cos(θ), e del (b) relativo errore standard della media, in funzione della distanza *d* dalla superficie. L'intensità del colore è proporzionale al numero di valori mediati. Primi 300 milioni di *step* della seconda replica alla pressione di H_2O di 8.125 matm.

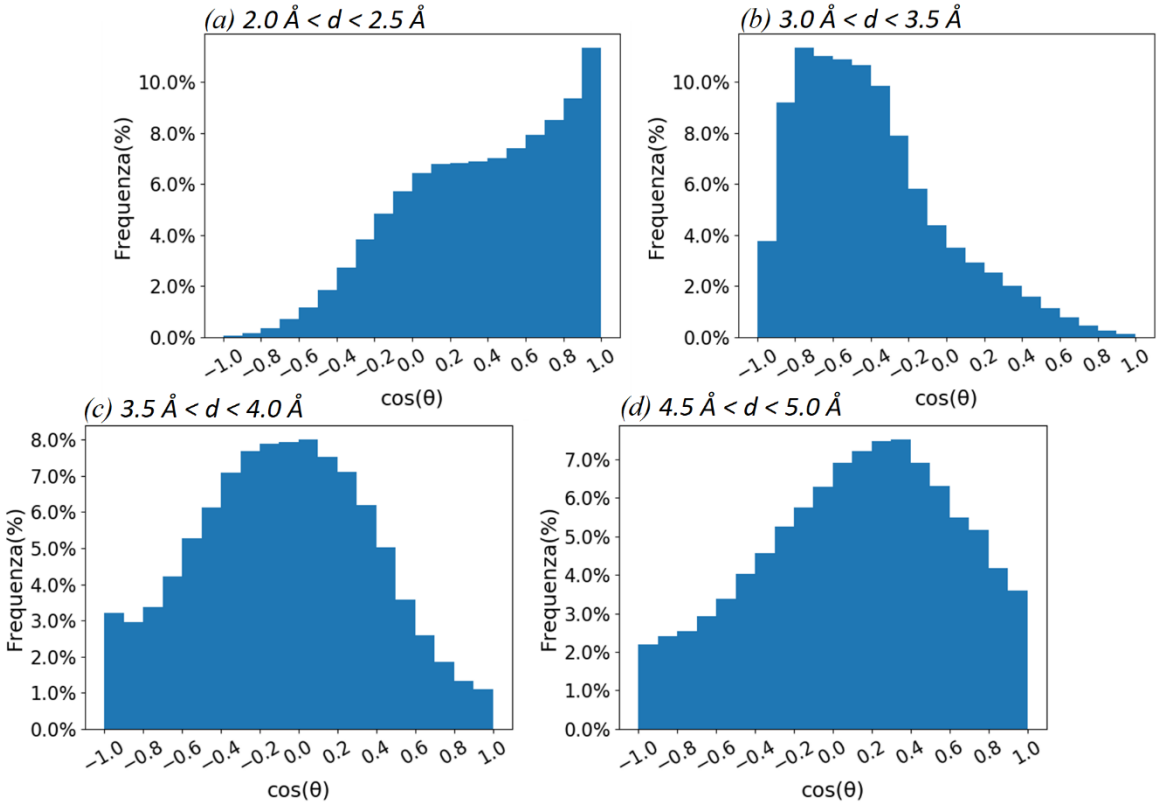

Figura 40: STRATI. Istogrammi relativi alla distribuzione delle orientazioni nei "settori di distanza" più rilevanti: *(a) 2.0 Å < d < 2.5 Å; (b) 3.0 Å < d < 3.5 ; (c) 3.5 Å < d < 4.0 Å; (d) 4.5 Å < d < 5.0 Å.*

- Simulazione condotta in condizioni di alto *coverage*: primi 300 milioni di *step* della seconda replica alla pressione di H_2O di 9.000 matm.

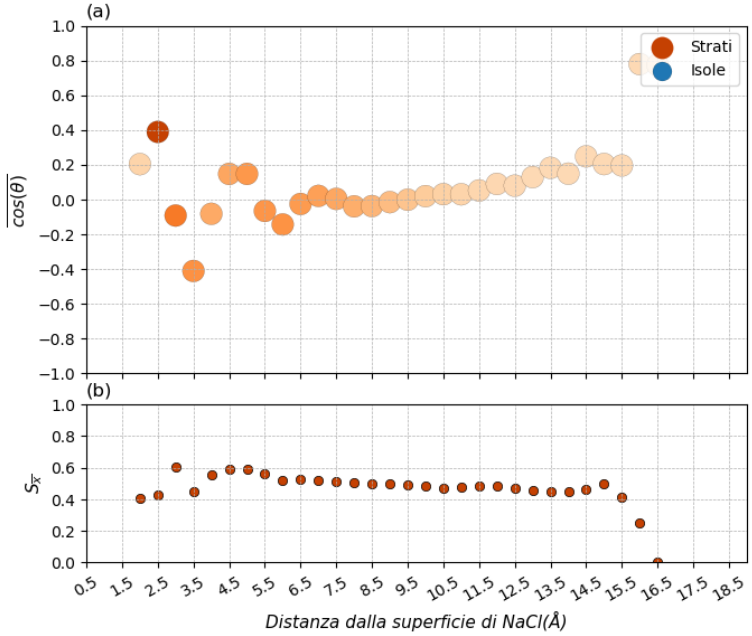

Figura 41: Grafico del (a) valore medio di cos(θ), e del (b) relativo errore standard della media, in funzione della distanza *d* dalla superficie. L'intensità del colore è proporzionale al numero di valori mediati. Primi 300 milioni di *step* della seconda replica alla pressione di H_2O di 9.000 matm.

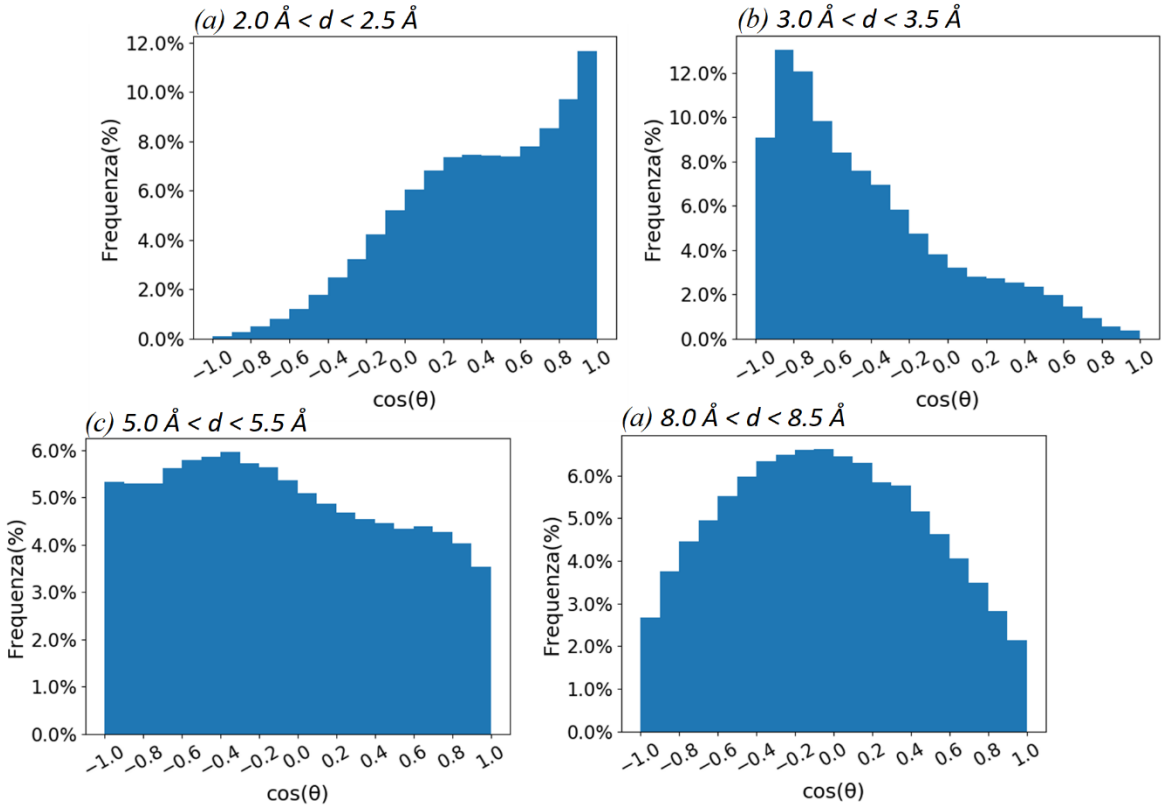

Figura 42: STRATI. Istogrammi relativi alla distribuzione delle orientazioni nei "settori di distanza" più rilevanti: *(a) 2.0 Å < d < 2.5 Å; (b) 3.0 Å < d < 3.5 ; (c) 5.0 Å < d < 5.5 Å; (d) 8.0 Å < d < 8.5 Å.*

Discussione dei risultati:

Osservando il grafico relativo alla simulazione condotta in condizioni di basso *coverage* (Figura 37) è possibile distinguere tre valori di $\overline{cos\,(\theta)}$ relativi a tre "settori di distanza" particolarmente rilevanti nella descrizione delle isole che caratterizzano lo strato adsorbito. Nel grafico è presente un massimo, in corrispondenza di $\overline{cos\,(\theta)} \cong 0.4 \pm 0.4$. Il valore e l'errore standard sono interpretabili sulla base dell'istogramma (Figura 38a) relativo allo specifico "settore di distanza" *(2.0 Å < d < 2.5 Å),* nel quale è osservabile una frequenza relativa molto maggiore per valori di $cos(\theta) > 0$, e massima per $\overline{cos(\theta)} \cong +1$. A una piccola distanza dalla superficie, le molecole d'acqua nelle isole saranno ipoteticamente posizionate sopra i cationi Na^+, con gli atomi di idrogeno direzionati lontano dalla superficie. Allontanandosi leggermente dalla superficie si ha un minimo per $\overline{cos\,(\theta)} \cong -0.4 \pm 0.4$, l'istogramma corrispondente (Figura 38b, *3.0 Å < d < 3.5 Å)* mostra una frequenza relativa massima per $cos(\theta) \cong -1$, decrescente per valori crescenti di $cos(\theta)$. Le molecole d'acqua in questo "settore di distanza" presentano il vettore momento dipolare prevalentemente orientato verso la superficie e saranno quindi legate tramite legame a idrogeno agli anioni Cl^-. Per una distanza ancora maggiore si ha $\overline{cos\,(\theta)} \cong 0.0 \pm 0.4$, c'è perdita di un'orientazione preferenziale definita dall'interazione delle molecole d'acqua con Na^+ o Cl^-.

Nonostante questo, osservando gli istogrammi relativi a distanze maggiori (Figura 38b, *3.5 Å < d < 4.0 Å),* è verificabile come non si ottenga una distribuzione discreta uniforme. Le frequenze relative sono massime per $-0.4 < cos(\theta) < +0.4$ e decrescono per valori di $cos(\theta)$ che tendono a +1 o -1. Le molecole d'acqua presenteranno quindi il vettore momento dipolare parallelo al piano della superficie o parzialmente orientato verso, o lontano, dalla stessa. Si può dedurre che, a una certa distanza dalla superficie, i legami a idrogeno laterali tra le molecole d'acqua stabilizzino maggiormente la struttura rispetto alle interazioni ione-dipolo. Per distanze superiori il numero di molecole d'acqua individuate è esiguo, i dati sono quindi poco rilevanti.

La seconda simulazione, di cui sono stati riportati i risultati (Figura 39), è stata scelta perchè condotta nella regione di transizione e perché caratterizzata da coesistenza di isole e strutture stratificate (intermedie tra *monolayer* e *bilayer*). Osservando il grafico, risulta evidente una forte corrispondenza tra i valori di $\overline{cos(\theta)}$ che caratterizzano le isole e gli strati, il dato è indice di quanto la superficie influenzi la struttura delle due tipologie di *cluster* in modo pressoché identico. Inoltre, per gli istogrammi degli strati (Figure 40a, 40b e 40c), relativi ai "settori di distanza" discussi nel caso a basso *coverage* per le isole, valgono le medesime considerazioni strutturali, l'aumento del numero di molecole d'acqua adsorbite non ha quindi modificato sostanzialmente l'effetto della superficie sulla struttura dello strato adsorbito. Per distante superiori a quelle considerate nel caso a basso *coverage* ($d > 4.5$ Å), $\overline{cos(\theta)}$ oscilla intorno a ad un valore uguale a 0. Il fatto che $\overline{cos(\theta)}$ oscilli, prima di andare a 0, potrebbe essere dovuto a un effetto orientante della superficie sulle molecole d'acqua, via via sempre meno rilevante allontanandosi dalla stessa.

Ad esempio, l'istogramma in Figura 40d (che mostra la distribuzione delle orientazioni per 4.5 Å $< d < 5.0$ Å), nonostante sia molto simile a quello osservato per 3.5 Å $< d < 4.0$ Å (Figura 40c), presenta un leggero *shift* del massimo della distribuzione verso valori positivi di $cos(\theta)$, il che potrebbe essere dovuto all'interazione con gli ioni Na^+, anche se molto indebolita rispetto all'interazione ione-dipolo che caratterizza le molecole d'acqua molto vicine alla superficie.

Le simulazioni ad alto *coverage* hanno tutte restituito risultati orientazionali molto simili ai casi a basso *coverage* e nella regione di transizione, il valore di $\overline{cos(\theta)}$ è: massimo per molecole molto vicine alla superficie, minimo allontanandosi leggermente dalla stessa e oscilla, fino a stabilizzarsi, intorno a 0 per valori di distanza maggiori. È stato scelto di riportare il risultato dello studio orientazionale effettuato sulla simulazione condotta a 9.000 perché, nonostante l'andamento di $\overline{cos(\theta)}$ sia coerente con quanto detto precedentemente, presenta una distribuzione delle orientazioni, ad una certa distanza dalla superficie, particolarmente significativa.

Osservando l'istogramma in Figura 42c, correlato ad una distanza dalla superficie compresa tra *5.0 Å* e *5.5 Å*, si osserva una distribuzione assimilabile, per quanto non perfettamente, ad una distribuzione discreta uniforme. Per un valore di pressione di H_2O di 9.000 matm, lo strato adsorbito è infatti di *multilayer*, è quindi probabile che le molecole d'acqua, presenti nello specifico "settore di distanza", appartengano a un *layer* intermedio, e quindi formino legami a idrogeno isotropici.

Il fatto che la distribuzione non sia perfettamente omogenea è probabilmente dovuto all'asimmetria del sistema e quindi all'effetto orientante della superficie: le molecole d'acqua nel *layer* intermedio interagiscono con la superficie di NaCl presente sotto di esse, anche se debolmente. Si può quindi concludere che, anche per pressioni di H_2O molto elevate, lo strato adsorbito *multilayer* risulta strutturalmente diverso dall'acqua liquida *bulk*. Le molecole d'acqua adsorbite che presentano comportamento più simile alle molecole dell'acqua liquida sono quelle dei *layer* interni.

Conclusioni

Durante il periodo di tirocinio è stato studiato il processo di adsorbimento, a temperatura ambiente, di molecole d'acqua su superfici modello di cloruro di sodio, componente primaria del particolato atmosferico di origine marina, attraverso l'analisi dei risultati di simulazioni computazionali effettuate in precedenti lavori di Tesi. La mia attività di tirocinio si è concentrata sullo sviluppo di uno *script*, che permettesse un'analisi dati, automatizzata e approfondita, delle configurazioni generate nel corso delle simulazioni e permettesse di ricavare risultati rilevanti nello studio dei fenomeni aggregativi che coinvolgono le molecole di acqua adsorbite sulla superficie del sale. In primo luogo, lo *script* ha permesso (implementando l'algoritmo DBSCAN) di effettuare il *clustering* su ogni *frame* delle simulazioni, quindi di rilevare e quantificare la presenza di strutture aggregate nel sistema, e successivamente di classificare i *cluster* rilevati, in funzione della loro dimensione. Sono state definite due tipologie di *cluster*: isole, ovvero aggregati bidimensionali di piccole dimensioni, che coprono solo parzialmente la superficie del sale, e strati, ossia strutture stratificate, la cui dimensionalità e le cui proprietà dipendono dal numero di *layer* di molecole d'acqua che li compongono. I risultati ricavati dallo studio dei singoli *frame*, mediati poi sull'intera simulazione, hanno permesso di comprendere come varia la struttura dello strato adsorbito in funzione della pressione di H_2O nel sistema. I risultati sono stati confrontati con le ipotesi strutturali avanzate, sulla base degli spettri di assorbimento IR a diversi valori di *coverage*, nello studio sperimentale di Foster e Ewing[40], le quali sono state perlopiù confermate. Nella regione a basso *coverage* ($\theta < 0.4$) della curva di adsorbimento, per i valori di pressione di H_2O di 7.500 e 7.750 matm, è sempre verificata la presenza di isole di piccole dimensioni nel sistema. La regione di transizione ($0.4 \leq \theta \leq 1.8$ e pressione compresa tra 7.875 e 8.250 matm) presenta caratteristiche strutturali diversificate, fenomeno correlabile all'elevata pendenza della curva di adsorbimento. Per piccoli valori di pressione si

[40] (Foster & Ewing, 2000)

osserva coesistenza tra strutture stratificate, intermedie tra *monolayer* e *bilayer*, e isole, mentre al crescere della pressione, e quindi del *coverage*, si ha completa copertura della superficie del sale. Nella regione ad alto *coverage*, ($1.8 < \theta \leq 3.5$) per valori di pressione di H_2O compresi tra 8.375 e 9.000 milliatmosfere, è confermata l'ipotesi di strutture *multilayer* dello strato adsorbito. In particolare, si osservano strutture di *bilayer* nella zona di *plateau* della curva, mentre, per valori di pressione superiori, aumenta la dimensione e quindi il numero di *layer* della struttura multistrato.

Inoltre, lo *script* sviluppato ha permesso di studiare come varia l'orientazione delle molecole d'acqua adsorbite in funzione della distanza dalla superficie e come tale dipendenza sia influenzata dal tipo di *cluster* di appartenenza e dalla pressione di H_2O nel sistema. Le molecole d'acqua molto vicine alla superficie *($d < 2.5$ Å)* sono posizionate sopra i cationi Na^+, con gli atomi di idrogeno direzionati lontano dalla superficie. A distanze leggermente maggiori (*3.0 Å $< d < 3.5$ Å)* il vettore momento dipolare delle molecole d'acqua risulta prevalentemente orientato verso la superficie e sono quindi legate tramite legame a idrogeno agli anioni Cl^-. A una distanza maggiore è osservabile la perdita di un'orientazione preferenziale definita dall'interazione delle molecole d'acqua con la superficie. Le considerazioni esposte sono risultate valide nella descrizione delle isole quanto delle strutture stratificate, indipendentemente dalla pressione di H_2O nel sistema.

Per le strutture multistrato, individuate nelle simulazioni condotte a 9.000 matm, ultima pressione di H_2O campionata prima della deliquescenza, è stata verificata la presenza di legami a idrogeno isotropici per le molecole d'acqua dei *layer* interni alla struttura.

In conclusione, il lavoro di tirocinio ha permesso di dimostrare il grande potenziale offerto dai metodi del *machine learning*, di cui le tecniche di *clustering* sono un esempio, nello studio dei risultati di simulazioni computazionali. La dimensione dei *database* generati nelle simulazioni è destinata a crescere con l'aumento della potenza di calcolo degli elaboratori, i metodi del *machine learning* sono in grado di gestire

una grande mole di dati in modo migliore rispetto ai metodi di analisi statistica classica. Inoltre, la capacità di modellizzazione non-lineare caratteristica di questi metodi permette la costruzione di modelli che si basano su strutture di dati complesse e di identificare *pattern* nascosti nei dati, con scopi descrittivi quanto predittivi. Nel caso specifico, la possibilità di analizzare singolarmente tutte configurazioni generate dalle simulazioni, ha consentito di ottenere una visione di dettaglio, a livello microscopico, dei fenomeni che coinvolgono l'adsorbimento di acqua su superfici di particolato atmosferico.

Bibliografia

Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques. In K. J., N. C., & T. M., *Grouping Multidimensional Data.* (p. p. 25-71). Springer, Berlin, Heidelberg.

Brunauer, S., Deming, L. S., Deming, W. E., & Teller, E. (1949). *J. Am. Chem. Soc. 62*, 1723.

Christner, B., Morris, C., Foreman, C., Cai, R., & Sands, D. (2008). *Science, 319*, p. 1214.

Christopoulos, C. D., Garimella, S., Zawadowicz, M. A., Möhler, O., & Cziczo, D. J. (2018). A machine learning approach to aerosol classification for single-particle mass spectrometry. *Atmos. Meas. Tech. 11*, p. 5687–5699.

Crabtree, J., Parker, S., & Purton, J. (2013). DL MONTE: a general purpose program for parallel Monte Carlo simulation. *Molecular Simulation 39*, pp. 1240.

Cramer, C. J. (2004). Essential of Computational Chemistry, Theories and Models.

Downs, G. M., & Barnard, J. M. (2002). Chapter 1: Clustering Methods and Their Uses in Computational Chemistry. In K. B. Lipkowitz, & D. B. Boyd, *Reviews in Computational Chemistry 18* (p. p. 1-40). John Wiley & Sons, Inc.

Einfeld, J. H., & Pandis, S. N. (1998). *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change.* New York: Wiley.

Engkvist, O., & Stone, A. J. (1999). Adsorption of water on NaCl(001). I. Intermolecular potentials and low temperature structures. *J. Chem. Phys. 110*, 12089.

Engkvist, O., & Stone, A. J. (2000). Adsorption of water on the NaCl (001) surface. III. Monte Carlo simulations at ambient temperatures. *The Journal of Chemical Physics, 112*, p. 6817.

Engkvist, O., & Stone, A. J. (2000). Adsorption of water on the NaCl(001) surface. III. Monte Carlo simulations at ambient temperatures. *The Journal of Chemical Physics, 112*, p. 6827.

Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *In Proc. KDD,* p. 226–231.

Foelsch, S., & Henzler, M. (1995). *Surf. Sci. 264*, 65.

Forster, P., Ramaswamy, V., Artaxo, P., Berntsen, T., Betts, R., Fahey, D., . . . Prinn, R. (2007). "Changes in Atmospheric Constituents and in Radiative Forcing". In *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge, UK, New York, NY .: Cambridge University Press.

Foster, M. C., & Ewing, G. E. (2000). Adsorption of water on the NaCl (001) surface. II. An infrared study at ambient temperatures. *The Journal of Chemical Physics, 112*, 6817.

Harris, C. R., Millman, K. J., Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., . . . A. (2020). Array programming with NumPy. *Nature 585*, p. 357–362.

Hiesh, W. W. (2009). *Machine Learning Methods in the Environmental Sciences.*

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering 9*, p. 90-95.

Joung, S., & Cheatham, T. E. (2008). Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular. *J. Phys. Chem. B 112*, 9020.

Lo, Y.-C., Rensi, S. E., Torng, W., & Altman, R. B. (2018). Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today 23*, p. 1538-1546.

Martin, S. T. (2000). Phase Transitions of Aqueous Atmospheric Particles. *Chemical Reviews, 100*, 3403–3454.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, (p. p. 51-56).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Perrot, M. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, p. 2825-2830.

Ramanathan, V., Ramana, M., Roberts, G., Kim, D., Corrigan, C., Cheng, C., & Winkler, D. (2007). *Nature, 448*, 575-578.

Richardson, H. H., Chang, H., Noda, C., & Ewing, G. E. (1989). *Surf. Sci., 216*, 43.

Sokolik, D. I. (2002). *Regional radiative effects due to anthropogenic aerosols.* . Tratto da irina.eas.gatech.edu: http://irina.eas.gatech.edu/ATOC3500_Fall1998/Lecture25.pdf

Valsaraj, K. T., Ehrenhauser, F. S., Heath, A. A., & Vaitilingom, M. (2015). Mass Transport and Chemistry at the Air-Water Interface of Atmospheric Dispersoids Food. In *Energy, and Water: the Chemistry Connection.* (p. p. 93-112).

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., . . . Mayorov, N. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods, 17(3)*, p. 261-272.

Yang, D., G.Bardoux, N.Assayag, C.Laskar, D.Widory, & P.Cartigny. (2018). Atmospheric SO2 oxidation by NO2 plays no role in the mass independent sulfur isotope fractionation of urban aerosols. *Atmospheric Enviroment, 193*, p. 109-117.

Introduction to Molecular Simulation and Statistical Thermodynamics. (2020, 10 15). Tratto da Scientific Figure on ResearchGate: https://www.researchgate.net/figure/Periodic-boundary-conditions-A-central-box-is-surrounded-by-copies-of-itself-The_fig6_238778314

NaCl polyhedra.png. (2008, 06 04). Tratto da https://commons.wikimedia.org/: https://commons.wikimedia.org/wiki/File:NaCl_polyhedra.png

(s.d.). Tratto da http://openchemistryhelp.blogspot.com/2012/08/rock-salt-structure.html

6.11A: Structure - Rock Salt (NaCl). (2020, 08 31). Tratto da chem.libretexts.org: https://chem.libretexts.org/Bookshelves/Inorganic_Chemistry/Map%3A_Inorganic_Chemistry_(Housecroft)/06%3A_Structures_and_energetics_of_metallic_and_ionic_solids/6.11%3A_Ionic_Lattices/6.11A%3A_Structure_-_Rock_Salt_(NaCl)

Ringraziamenti

Il primo ringraziamento va al mio relatore, il Prof. Greco, e il mio correlatore, il Prof. Cosentino, per la gentilezza, la pazienza e la disponibilità con cui mi hanno guidato e consigliato sia durante il Tirocinio che nella redazione della Tesi.

Un grazie speciale va alla mia famiglia: ai miei genitori, per aver sempre creduto in me e per avermi supportato durante tutto il percorso universitario; a mio fratello Lorenzo, che è stato sempre al mio fianco, nei momenti migliori e in quelli più difficili; a mia nonna Maria per l'incondizionato affetto e a tutti i nonni che non ci sono più, per aver contribuito a rendermi la persona che sono oggi.

Ringrazio Laura per la stima, l'amore e tutti i bei momenti passati insieme, per aver sempre visto oltre la superficie, esortandomi ogni giorno ad essere la migliore versione di me stesso.

Infine, ma non per questo meno importanti, ringrazio i miei più cari amici: Teo, Gara, Marci, Tino e Lampre, su cui posso sempre contare e con cui condivido ricordi indelebili.