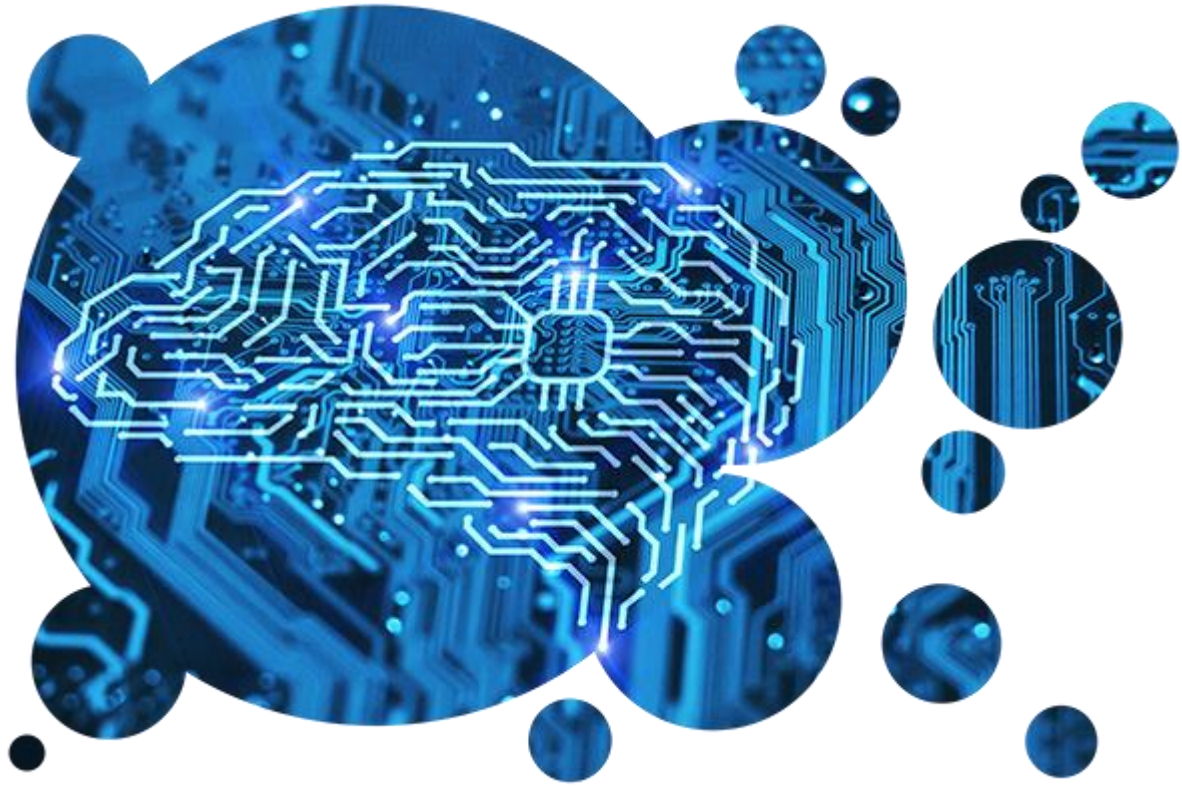
CASE STUDY ATTRACTOR

Obiettivo tematico I - Ricerca, sviluppo tecnologico e innovazione Azione I.1.b.1.2 “Sostegno alle attività collaborative di R&S per lo sviluppo di nuove tecnologie sostenibili, di nuovi prodotti e servizi - **Bando PASS**

Studio di fattibilità: ATTRACTOR



strumenti di intelligenza ArTificiale per oTtimizzare il RilAscio ConTrollatO di faRmaci

Pagina intenzionalmente vuota

Contenuti

1 Introduzione

1.1 Acronimi e definizioni

Ogni acronimo che compare nel documento deve essere inserito nella tabella seguente ordinata in ordine alfabetico

Termine	**Descrizione**
CIF	Crystallographic Information **File**
MOF	Metal-Organic Framework
POC	Proof of Content
SdF	Studio di Fattibilità
ML	Machine Learning
OdR	Organismo di Ricerca

2 ATTRACTOR

Lo studio di fattibilità svolto con il partner Aethia(https://www.aethia.com/) e in collaborazione con l'Università di Torino (dip. Chimica - https://www.chimica.unito.it/do/home.pl) in ambito medicina di precisione si è concentrato sull'analisi e sull'identificazione di differenti materiali (carrier) adatti al rilascio di farmaci nell'organismo. La simulazione del comportamento dei diversi carrier tramite metodi quanto-meccanici è molto dispendiosa, mentre il machine learning (ML) costituisce una valida alternativa.

Questo PoC tipo potrebbe essere un notevole aiuto per un'azienda farmaceutica o centro di ricerca, la quale potrebbe sfruttarlo per studiare e preparare carrier adatti senza eseguire tutte le simulazioni. Infatti, con tempi e costi contenuti, si potrebbe restringere il campo di ricerca a un numero minore di candidati su cui concentrare in un secondo momento ricerche più approfondite. L'ipotesi di business legata al successo di questo studio di fattibilità si basa inizialmente sulla possibilità di offrire un servizio di supporto all'indirizzamento di analisi per centri dei ricerca, per poi successivamente diventare un portale di servizi per aziende farmaceutiche. Lo studio ha consentito a entrambe le aziende partecipanti di maturare l'esperienza necessaria in tecniche A.I.

Questo studio di fattibilità è stato realizzato grazie al contributo ricevuto tramite il Bando **PASS**:

Progetto in sviluppo con fondi di investimento europei 2014/2020

Obiettivo tematico I - Ricerca, sviluppo tecnologico e innovazione
Azione I.1.b.1.2 "Sostegno alle attività collaborative di R&S per lo sviluppo di nuove tecnologie sostenibili, di nuovi prodotti e servizi - Bando PASS

Lo studio di fattibilità ha caratterizzato una piattaforma adeguata a esigenze e dimensioni di una realtà professionale media e alla valutazione del software applicativo necessario in termini di risorse hardware, software e network.
Nella realizzazione dello studio di fattibilità si sono sviluppati sia pur in modo prototipale strumenti di ML per il loro utilizzo nello studio delle relazioni struttura-proprietà di materiali microporosi metallo-organici (metal-organic frameworks, MOF).

I MOF, infatti, hanno la caratteristica di essere chimicamente molto versatili perché costituiti da unità strutturali modulabili e assemblabili con topologie diverse, come in un "lego molecolare". Inoltre, i reticoli tridimensionali porosi sono caratterizzati da gabbie e canali di forma e

dimensione diverse che li rendono selettivi alla cattura e al rilascio di molecole. I MOF rappresentano quindi una fonte di dati molto grande e un banco di prova ideale per sviluppare e mettere a punto strumenti di ML.

Nel nostro caso specifico è stata necessaria inizialmente la raccolta dei dati e la creazione di uno o più database; in particolare sono stati effettuati dei calcoli quanto-meccanici ab-initio usando il programma CRYSTAL a partire da alcune decine di strutture. In seguito sono state aggiunte altre strutture prelevandole dal database Opensource COREMOF.

Si è proceduto alla definizione di opportuni descrittori da usare come input per l'algoritmo di apprendimento e la messa a punto degli algoritmi stessi per l'analisi dei dati (modelli della statistica-multivariata). Infine, si è predisposto un sito internet pubblico in cui gli strumenti sviluppati permetteranno agli utenti di ottenere le proprietà d'interesse per nuove strutture attraverso il motore di apprendimento automatico. Il sito pubblico consentirà di diffondere più facilmente informazioni riguardo alla tecnologia che si svilupperà, di conseguenza la sua visibilità favorirà la creazione di opportunità di business.

Più in dettaglio, lo studio ha previsto una fase preliminare di predizione delle proprietà elettriche dei MOF sulla base di dati già a disposizione che hanno messo in evidenza una relazione diretta tra la porosità del materiale e la carica elettrica.

In una seconda fase, si è proceduto alla messa a punto di uno strumento di apprendimento automatico per la predizione delle cariche atomiche dei MOF. Quest'ultimo passaggio risulta essere essenziale per la predisposizione di modelli classici, tipo campi di forza, da usare nella simulazione dell'adsorbimento di molecole all'interno della struttura microporosa. Infatti, a differenza di altri approcci, si utilizzeranno cariche di tipo-Hirshfeld che derivano direttamente dalla densità di carica del materiale. Queste sono state calcolate con metodi ab-initio usando il programma CRYSTAL su un ampio insieme di strutture, sia sintetizzate sia ipotetiche.

In parallelo a queste fasi, è stata svolta l'attività fondamentale di definizione dei descrittori, creazione del modello di rappresentazione dei dati e messa a punto degli algoritmi di analisi che costituiranno il cuore del motore di apprendimento automatico.

Le fasi fondamentali per la realizzazione del modello di machine learning efficace sono 4:

- raccolta dei dati, che devono essere esenti da errori per evitare che gli algoritmi del ML vengano fuorviati
- rappresentazione dei dati, che devono essere convertiti in un formato comprensibile per l'algoritmo

- scelta del tipo di apprendimento, che può essere supervisionato, semi-supervisionato e non supervisionato a seconda della richiesta che viene avanzata alla macchina
- ottimizzazione del modello, che prevede l'eliminazione o la diminuzione dei tre errori fondamentali ovvero bias (o tendenza del modello), varianza del modello ed errori irriducibili

In prospettiva, le cariche atomiche ottenute dai modelli potranno eventualmente essere usate all'interno di campi di forza sviluppati per i MOF (es. MOF-FF) e validate nella simulazione dell'adsorbimento di piccole molecole (es. H2O, CO2, …) e di farmaci (es. 5-fluoro-uracile). Quest'obiettivo di massima potrebbe essere perseguito in un progetto successivo.

Per la realizzazione dello studio è stato necessario avvalersi della consulenza di un OdR. L'OdR coinvolto è UniTO, in particolare il gruppo di chimica teorica del Dipartimento di Chimica dell'università di Torino nella persona del Prof. Bartolomeo Civalleri.

L'Sdf è nato da un'idea progettuale che unisce competenze, risultati scientifici e tecnologie che i partner, unitamente all'OdR, mettono a fattor comune. Pertanto il TRL iniziale si può considerare come TRL2. L'obiettivo dell'SdF della realizzazione di un prototipo funzionante di software che dimostri la validità dell'idea di partenza, e che quindi può essere considerato un TRL4 è stato raggiunto. Il prototipo sarà presentato e valutato in modo da poter poi avviare un passaggio a livelli successivi.

La collaborazione tra i partner è stata ottimale e non ci sono state criticità e necessita di attivare procedure di gestione del rischio.

Le attività si sono svolte e organizzate in base ai work package previsti.

- WP1 Project Management

il Project Manager si è fatto carico di gestire la pianificazione delle diverse attività, l'organizzazione delle riunioni di progetto volte al coordinamento delle varie attività e all'aggiornamento sullo stato progettuale, il monitoraggio dell'avanzamento del progetto, il controllo dei risultati intermedi e finali. Il Project Manager ha, inoltre, sovrainteso alla realizzazione e verificare la correttezza, la consistenza e la qualità dei diversi Milestones e Deliverables. Si è occupato della redazione della reportistica legata alle due Milestones.

- WP2 Audit tecnologico e Analisi Funzionale

Durante il WP2 si è eseguita l'analisi di dataset disponibile e degli algoritmi e l'installazione dei serve virtuali necessari assieme alle librerie scelte (ambiente linux, librerie open source, linguaggi di programmazione pyhon, php, SQL) Redazione delle specifiche funzionali

- WP3 **Sviluppo POC**

Durante il WP3 si è costruito il POC con la produzione dei dataset, lo sviluppo di due modelli di ML e lo sviluppo dell'interfaccia Web tramite WebApp (http://attractor.netsurf.it).

3 Disponibilità del PoC

Le attività si sono concluse cogliendo tutti gli obiettivi previsti dallo studio di fattibilità.

Il POC è disponibile online e accessibile alla seguente url: http:/attractor.netsurf.it

Sono stati realizzati due modelli di Machine learning sviluppati con tecnologie diverse e hanno portato risultati analoghi, sia pure con performance diverse e precisioni simili.

In particolare gli obiettivi raggiunti:

- sviluppato modello con tecnologia open source XGBurst su ispirazione dell'articolo [1]
- Nuovo modello creato con TensorFlow e rete neurali, più dinamico e con ottimi risultati
- In ogni caso con entrambi i modelli l'errore tra dati calcolati rispetto ai predetti è assolutamente accettabile, con grande soddisfazione per entrambi i modelli realizzati
- DI conseguenza si è fatto in modo che l'utente che richieda una predizione possa scegliere quale modello richiamare
- **Uno degli indici utilizzati nel modello ha creato qualche problema causando lo scarto di alcune configurazioni. Si è trattato dell'Indice di Voronoi**: discussione lato Unito per capire se escluderlo o no. Si è concordato di agire nel modo seguente:
  - Caricamento in MatMiner di tutti gli atomi escludendo l'indice di Voronoi e verificato che il dataset si è ampliato (erano prima scartati il 27% dei dati sottomessi)

  Lo si è per tanto escluso in quanto già considerato in un altro parametro e con poco impatto nel modello (la precisione rimaneva simile nella predizione con o senza indice di voronoi nei dati sottomessi)

  Il training set (UNITO) fornito da UNITo si è arricchito durante tutto il periodo del SdF Attractor.

- **Disponibilità file CIF da database pubblico core CoREMOF**: sono state prelevate molecole MOF da questo DB, normalizzate e utilizzare dal modello ML risultato.
- Realizzata la webapp utilizzabile su qualunque device (mobile incluso) che permette di richiedere la predizione e di conservare una storia dei cif file e delle predizioni.

- **Possibilità di visualizzare le molecole a partire dal file cif dalla webapp**: visualizzazione di una molecola a partire dal file cif caricato tramite la libreria js jmol:

http://jmol.sourceforge.net

- **Bontà della predizione e risultato per l'utente**: si è valutata la bontà della predizione tramite la tecnica di cross-validation per addestrare il modello, tecnica che può essere usata trasversalmente indipendentemente dal modello utilizzato.
- Proposta un'architettura senza implementarla nel SdF che consenta un'integrazione con l'applicazione Crystal per popolare in modo automatico il training set ed eventualmente l'integrazione con altre applicazioni disponibili simili a Crystal.

Il successo dei risultati ottenuti con lo SDF dà motivo di creare nuove opportunità per la realizzazione di un servizio professionale basato su i modelli realizzati, che sia proponibile sul mercato per i centri di ricerca che intendono selezionare molecole MOF adatte a completare analisi computazionale con l'obiettivo di creare 'carrier' di farmaci per la "medicina di precisione".

4 Conclusioni

Lo studio di fattibilità ha raggiunto gli obiettivi prefissati con accuratezza maggiore di quanto previsto.

I MOF sono materiali ibridi organici-inorganici che negli ultimi anni hanno stimolato l'interesse della comunità scientifica per le loro caratteristiche peculiari. Sono costituiti dall'assemblaggio di nodi inorganici (metalli o cluster) e di leganti organici che formano delle strutture stabili, ordinate e porose. Inoltre, hanno la caratteristica di essere chimicamente molto versatili perchè costituiti da unità strutturali modulabili e assemblabili con topologie diverse, come in un "lego molecolare". Questo li rende particolarmente interessanti per screening predittivi come quello del progetto ATTRACTOR.

Si sono, infatti, verificati i risultati ottenuti con molecole MOF per cui si è calcolato con esattezza la distribuzione delle cariche elettriche e i risultati ottenuti con il modello predittivo e l'accuratezza è impressionante.

NetSurf S.r.l. con l'Università di Torino ha lavorato per l'utilizzo della libreria MATMIner [1] per la costruzione di un modello predittivo per ottenere cariche atomiche di Metal-Organic Frameworks (MOF).

Il lavoro svolto all'interno del progetto si è articolato su due direzioni:

(i) Messa a punto di un protocollo di lettura di dati cristallografici di strutture di MOF, provenienti da banche dati o ottenuti da calcoli quanto-meccanici, utilizzando il programma CRYSTAL, attraverso la libreria MATMiner [1].

(ii) Utilizzo di MATMiner [1] per generare le informazioni descrittive della struttura necessarie per la predisposizione di un algoritmo di machine learning (ML) che è stato impiegato per la predizione di cariche atomiche di MOF

La strategia per la definizione dei descrittori strutturali da utilizzare all'interno dell'algoritmo di ML si è basata sull'assunzione di località, per la quale la predizione delle cariche può essere determinata in funzione della natura chimica dell'atomo in esame e del suo intorno chimico. La procedura seguita è analoga a un recente lavoro sullo stesso soggetto [2].

Lo schema del processo per generare le informazioni da fornire all'algoritmo di ML è rappresentato nella Figura 1. All'interno del progetto è stato messo a punto uno script python che gestisce le diverse fasi:

1) Estrazione della struttura del MOF da file cristallografico (formato cif) proveniente da database esperimentali (es. CSD) o dataset di strutture di MOF adattate per la simulazione (es. CoReMOF-2019 [3] e Quantum MOF [4].)
2) Simmetrizzazione della struttura e creazione automatica di un meta-input per CRYSTAL

3) Ottimizzazione della struttura del MOF e calcolo delle cariche con CRYSTAL (metodo di calcolo: PBEsol-3c [5])
4) Analisi della struttura per determinare i descrittori strutturali e creazione della tabella dati da passare all'algoritmo di ML.
5) Predisposizione di un file cif con la struttura ottimizzata e le cariche atomiche calcolate.

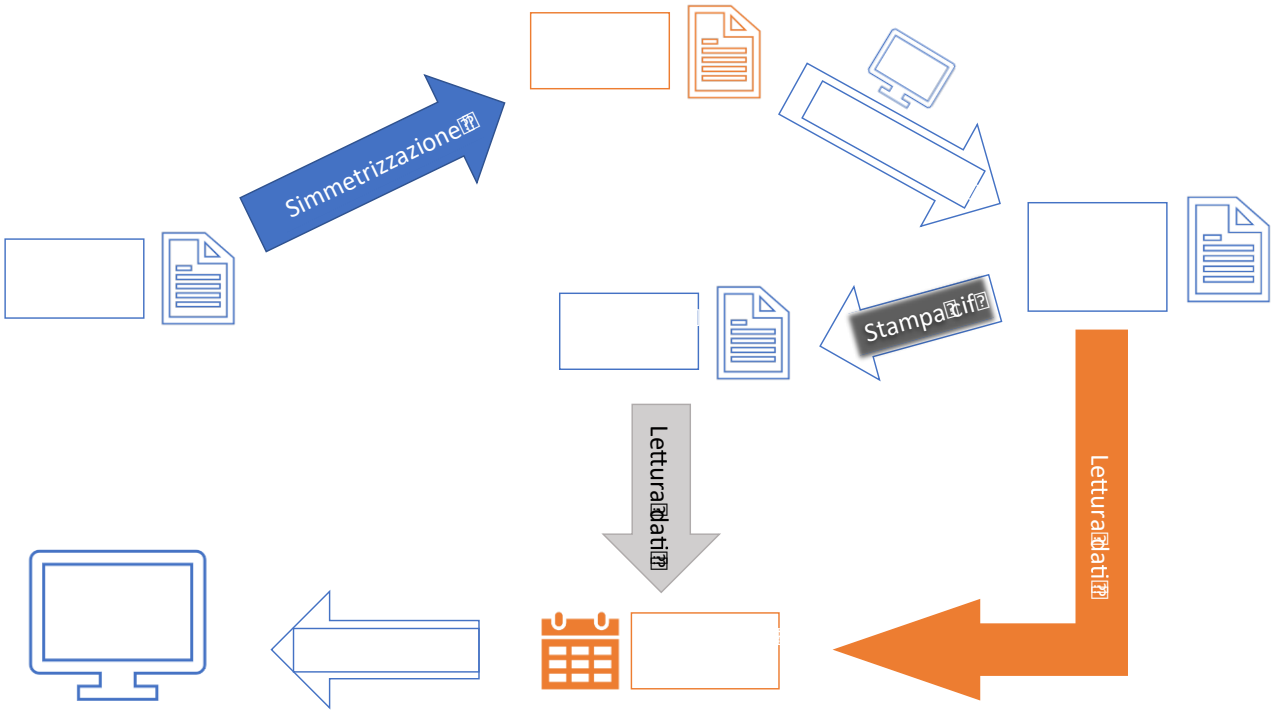


Figura 1. Schema del diagramma di flusso dello script messo a punto per gestire i diversi passaggi richiesti per ottenere i dati da fornire all'algoritmo di ML.

Come descrittori sono state utilizzate informazioni strutturali relative all'intorno chimico degli atomi nella struttura del MOF (estratte attraverso MATMiner [1]) e alle loro proprietà atomiche (estratte dalla libreria Mendeleev [6])

Nel dettaglio, i descrittori strutturali individuati attraverso MATMiner [1] sono:

- Numero di coordinazione, cioè il numero di primi vicini rispetto ad un dato sito atomico (CoordinationNumber())
- Indice di Voronoi e altri parametri collegati (VoronoiFingerprint())
- Caratteristiche cristallografiche del sito atomico (CrystalNNFingerprint ())
- Parametri d'ordine locali relativi all'intorno chimico del sito atomico (OPSiteFingerprint())
- Funzione di distribuzione radiale e proprietà collegate (AGNIFingerprints())
- Funzioni gaussiane di simmetria radiale e angolare (GaussianSymmFunc())

Per le proprietà atomiche sono state invece considerate: il gruppo e il periodo della tavola periodica cui appartiene l'atomo, l'elettronegatività di Pauling e il raggio covalente.

Infine, è stata predisposta all'interno della procedura la possibilità che lo stesso script python possa essere utilizzato per la lettura dei dati esterni relativi a un ipotetico MOF, compatibilmente con l'attuale set di training dell'algoritmo di ML, e dare inizio alla procedura di predizione delle cariche attraverso il portale web sviluppato all'interno del progetto.

Aethia S.r.l. in collaborazione con l'Università di Torino (UNITO) ha contribuito alla creazione di un database di strutture di Metal-Organic Framework (MOF) per le quali sono stati effettuati i calcoli quanto-meccanici ab-initio usando il programma CRYSTAL per generare le cariche atomiche di Hirshfeld.

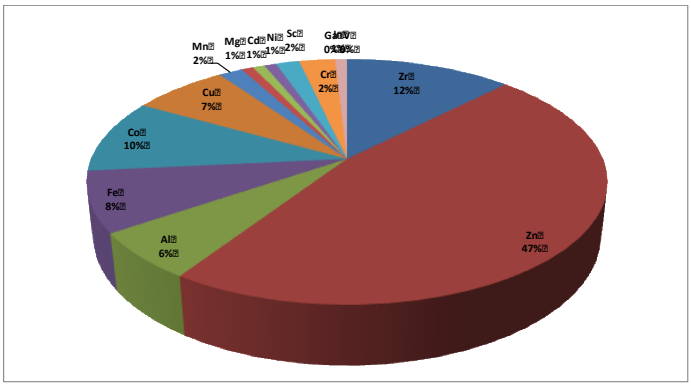
La selezione dei MOF che costituiscono il database è stata tale da cercare di garantire un ampio screening dello spazio chimico dei materiali non solo in termini quantitativi, ma anche qualitativi. Si è cercato di estendere il più possibile la tipologia di elementi chimici che entrano nella struttura del nodo inorganico del MOF sia in termini di specie chimica e, a parità di metallo, di tipo di coordinazione sia includendo strutture con la stessa composizione chimica ma con diversa topologia reticolare. Per questo tipo di screening computazionali sui MOF, è importante sottolineare l'importanza di avere delle strutture cristalline ripulite (es. senza le molecole di solvente) non disordinate e prive di difetti, in modo che le coordinate atomiche possano essere usate in modo affidabile per i successivi calcoli.

Per la costruzione del database è stato fatto riferimento a due set:

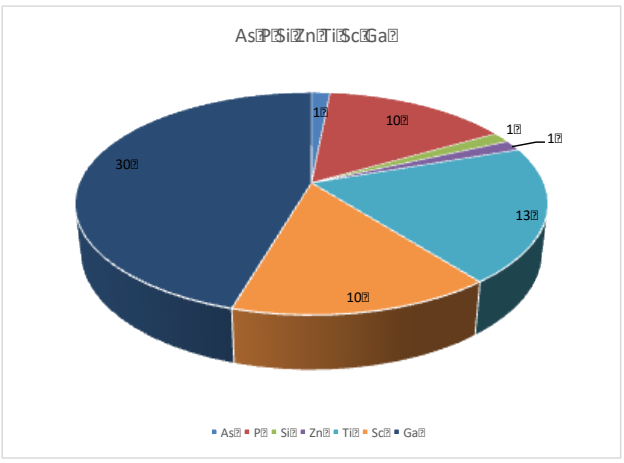
1) Un insieme di strutture raccolte negli anni dal gruppo di ricerca di UNITO sulle quali erano già stati condotti calcoli quanto meccanici di vario tipo. Questo costituisce un primo set di circa 110 MOF.

2) Il secondo set è stato estratto da due database sviluppati da altri gruppi di ricerca e disponibili pubblicamente: (i) il database CoReMOF-2019 (Computation-Ready MOF) [1] e il database QMOF (Quantum MOF) [2]. Entrambi sono derivati dal database cristallografico CSD in cui si trovano raccolte le strutture dei MOF sintetizzati sperimentalmente. Dai due database sono state estratte 53 strutture circa per estendere il tipo di specie chimiche presenti nel dataset.

Nella Figura 2 è riportata la distribuzione percentuale degli elementi chimici presenti nei nodi inorganici per i due set di dati. Come si può vedere, il primo set contiene strutture di MOF in cui è presente maggiormente lo Zn, seguito da altri metalli di transizione come Co, Fe e Zr. Il secondo set, invece, è stato arricchito con strutture contenenti Ti, Ga, P e Si.



(a)



(b)

Figura 2. Distribuzione percentuale degli elementi chimici presenti nei nodi inorganici nel primo dataset (a) di strutture raccolte da UNITO e nel secondo dataset (b) di strutture proveniente da database esterni.

L'insieme dei MOF selezionati è stato quindi processato attraverso lo script, messo a punto in collaborazione con NetSurf S.r.l., per generare automaticamente gli input CRYSTAL delle strutture, opportunamente simmetrizzate, e procedere con il calcolo delle cariche atomiche di Hirshfeld.

Per i calcoli quanto-meccanici è stato usato un metodo composito ibrido HF/DFT denominato PBEsol0-3c, opportunamente sviluppato per la simulazione di solidi [5], che impiega il funzionale ibrido PBEsol0, in combinazione con una base di qualità doppio-zeta, e che include le interazioni dispersive attraverso lo schema D3 e una correzione (gCP) per eliminare il BSSE intrinseco nel set base utilizzato. Il metodo di calcolo è particolarmente indicato per lavori di screening, come per il progetto ATTRACTOR, in quanto è stato dimostrato il suo ridotto costo computazionale [7] e la sua accuratezza nella simulazione dei MOF [8].

In totale sono stati eseguiti i calcoli su 165 MOF. La struttura del MOF in termini di parametri reticolari e coordinate atomiche è stata prima ottimizzata e sulla geometria di equilibrio sono state calcolate le cariche di Hirshfeld. Alla fine, sono state ottenute oltre 20000 cariche atomiche.

I dati così generati sono poi stati utilizzati da Aethia per completare la procedura di allenamento e validazione dell'algoritmo di machine learning.

5 Referenze

[1] Logan Warda, Alexander Dunnc, Alireza Faghaninia, Nils E.R. Zimmermann, Saurabh Bajaj, Qi Wangc Joseph Montoya, Jiming Chen, Kyle Bystrom, Maxwell Dylla, Kyle Chard, Mark Asta, Kristin A. Persson, G. Jeffrey Snyder, Ian Foster, Anubhav Jain

Matminer: An open source toolkit for materials data mining

Computational Materials Science 152 (2018) 60–69

[2] Vadim V. Korolev, Artem Mitrofanov, Ekaterina I. Marchenko, Nickolay N. Eremin, Valery Tkachenko, and Stepan N. Kalmykov

Transferable and Extensible Machine Learning-Derived Atomic Charges for Modeling Hybrid Nanoporous Materials

Chemistry of Materials 32 (2020) 7822−7831

[3] Andrew S. Rosen, Shaelyn M. Iyer, Debmalya Ray, Zhenpeng Yao, Alán Aspuru-Guzik, Laura Gagliardi, Justin M. Notestein, Randall Q. Snurr

Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery

Matter 4 (2021) 1578-1597

[4] Yongchul G. Chung, Emmanuel Haldoupis, Benjamin J. Bucior, Maciej Haranczyk, Seulchan Lee, Konstantinos D. Vogiatzis, Sanliang Ling, Marija Milisavljevic, Hongda Zhang, Jeff S. Camp, Ben Slater, J. Ilja Siepmann, David S. Sholl, & Randall Q. Snurr.

Computation-Ready Experimental Metal-Organic Framework (CoRE MOF) 2019 Dataset (1.1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3370144

[5] Lorenzo Donà, J. Gerit Brabdenburg, Bartolomeo Clvalleri,

Extending and assessing composite electronic structure methods to the solid state

Journal of Chemical Physics 151 (2019) 121101

[6] https://pypi.org/project/mendeleev/0.3.1/

[7]Lorenzo Donà, J. Gerit Brabdenburg, Ian J. Bush, Bartolomeo CIvalleri,

Cost-effective composite methods for large-scale solid-state calculations

Faraday Discussions 224 (2020) 292–308

[8] Lorenzo Donà, J. Gerit Brabdenburg, Bartolomeo CIvalleri,

Metal–organic frameworks properties from hybrid density functional approximations

Journal of Chemical Physics 156 (2022) in press; doi: 10.1063/5.0080359