Studio dei siti di binding di complessi proteici e analisi dell'affinità di legame attraverso modelli di machine learning

Francesco D'Amico

9 Luglio 2022

1 Abstract

Recenti progressi nel campo del machine learning [Jum+21] hanno mostrato la possibilità di compiere eccezionali predizioni della struttura tridimensionale degli atomi di una proteina, partendo dalla sequenza dei suoi aminoacidi. Quindi, parzialmente, il problema del folding è stato risolto, perché se si conoscono con buona approssimazione le posizioni atomiche, si conoscerà bene di conseguenza la superficie molecolare e la forma complessiva del fold. Invece ad oggi un metodo affidabile e rapido per la predizione delle interazioni tra due o più differenti proteine non è ancora disponibile. Il presente lavoro esplora possibili metodi di studio e predizione delle caratteristiche dei siti di binding tra due proteine, avvalendosi sia di analisi statistiche, sia dell'uso di modelli di machine learning.
Come grandezza rappresentativa dell'intensità del legame tra due proteine si sceglie la costante di associazione K_a del complesso, e si cerca di comprendere da quali caratteristiche del sito di legame questa grandezza è influenzata. Data la complessità dell'intricato network tridimensionale di interazioni che si stabiliscono nel sito di legame tra due proteine, si semplifica la trattazione considerando solamente poche caratteristiche degli aminoacidi presenti e proiettandole su un piano bidimensionale. Tramite questa metodologia semplificata, si dimostra che è possibile comunque ricostruire delle informazioni sulla fisica del problema, e addestrare un modello di rete neurale artificale in grado di predire in modo soddisfacente la costante di associazione K_a tra due proteine.

Indice

2 Introduzione

2.1 Lo scopo del lavoro

Lo scopo di questo lavoro è esplorare nuovi possibili metodi per studiare la natura del binding tra biomolecole, e nel particolar caso tra proteine. Comprendere il meccanismo con il quale le proteine si assembrano tra loro e predire la formazione dei complessi proteici partendo dalle singole proteine nella loro forma unbound, è ancora una sfida aperta nel campo della biologia computazionale. In particolare, si investiga l'organizzazione delle interazioni intermolecolari attraverso un metodo di descrizione geometrica e biochimica del binding che sia in grado di ridurre la dimensione e la complessità del sistema di input. L'intenzione è di sviluppare un modello del sito di legame di bassa complessità che riesca comunque a mantenere al suo interno le informazioni utili a ricostruirne la stabilità. L'organizzazione della complessa rete delle interazioni intermolecolari, così come la composizione aminoacidica dei siti di legame, sono i due fattori chiave analizzati separatamente nell'analisi qui presentata. Nel dettaglio, il metodo di modellizzazione scelto consiste in un algoritmo di proiezione che diminuisca la dimensionalità del problema da 3 a 2 dimensioni, e che tenga conto solamente di un insieme minimale di informazioni geometriche e chimiche. Come grandezza rappresentativa della stabilità del legame si sceglie il valore sperimentale della costante di associazione di legame K_a.

2.2 La fisica del legame

Le proteine sono polimeri di aminoacidi che nell'ambiente biologico in cui svolgono le loro funzioni sono ripiegate in una struttura tridimensionale ben precisa. Quando due differenti proteine si trovano ravvicinate a formare un complesso, prende forma un complicatissimo network di interazioni, che dipende dalle precise posizioni e tipologie di aminoacidi presenti e dalla posizione e orientazione dei loro residui. Tra di questi si sviluppano legami intermolecolari, principalmente dovuti alle caratteristiche della distribuzione delle cariche, e alla presenza di dipoli permanenti e dipoli indotti.
Fintanto che le due proteine sono a distanza sufficientemente grande da non interagire, queste si trovano in una forma unbound. Nel transiente in cui si avvicianano e quando poi si trovano legate, le interazioni in gioco causano spostamenti dei nuclei e deformazioni degli orbitali molecolari: noi facciamo l'ipotesi che per determinare la stabilità del legame sia più importante la struttura statica dell'interfaccia tra le due proteine rispetto a questi fenomeni dinamici. L'approccio seguito è specifico per questo caso, mentre per studiare la dinamica del legame servirebbe l'elaborazione di un metodo più sofisticato. Non è esclusa però la possibilità che si possa partire da metodi statici come questo per ottenere degli approcci che invece tengano conto della dinamica del processo.
Di tutte le possibili caratteristiche fisiche degli aminoacidi presenti nel sito di legame, si sceglie di tenere conto solamente di informazioni geometriche (è sicuramente necessario in quanto le principali interazioni in gioco sono locali) e dell'idropatia, ovvero dell'idrofobicità e dell'idrofilicità. L'idropatia è importante poiché è legata alla capacità degli aminoacidi di formare determinati legami intermolecolari, in particolare i legami a idro-

geno. L'idrofilicità e l'idrofobicità hanno la caratteristica di poter essere descritte tramite un singolo numero in una certa scala. Questo è ottimo per ridurre la complessità del problema, ma di sicuro è una semplificazione della fisica realmente in atto. Ad esempio non conta solo la posizione dei residui polari e apolari, ma anche la loro orientazione e la presenza di altri aminoacidi nei dintorni.
Per quanto riguarda la misura della stabilità del legame, si sceglie come grandezza osservabile il valore della costante di associazione K_a. Questa si ottiene sperimentalmente misurando in un ambiente opportuno (ad esempio a pH e temperatura tipiche di una cellula) le concentrazioni molari [A] e [B] delle due proteine separate, e la concentrazione del composto [AB]:

$$K_a = \frac{[AB]}{[A][B]}$$

Questa grandezza ha come scala naturale quella logaritmica, per cui d'ora in poi si considererà sempre come K_a il suo logaritmo in base 10: composti con un alto valore di K_a sono quindi caratterizzati da un legame stabile.

2.3 Il dataset

Si dispone dei file in formato PDB delle posizioni di tutti gli atomi di 274 composti proteici. Nella scala naturale logaritmica, i valori delle K_a dei composti presi in considerazione seguono una distribuzione circa gaussiana. Le posizioni atomiche sono state ottenute tramite X-ray cristallography, eseguita dopo che i complessi di proteine sono stati fatti cristallizare. Successivamente sono stati eliminati dai dati eventuali residui di molecole d'acqua e imperfezioni artefatte.
Invece, il valore delle K_a è stato ottenuto misurando in ambiente fisiologico le concentrazioni molari dei composti e delle proteine dissociate. E' evidente come le due misurazioni siano effettuate in ambienti molto diversi, e quindi è importante l'assunzione che la struttura del sito di legame non sia modificata in modo eccessivo tra i due ambienti.

3 Il metodo

3.1 L'algoritmo di proiezione

Il sito di binding tra due proteine consiste in un complicato network di interazione tridimensionale. In questo lavoro, il primo passo consiste nella semplificazione del problema. Si comincia dai 274 file PDB, che contengono le informazioni sulla posizione tridimensionale e la tipologia degli atomi della catena proteica. Si sceglie di tenere conto solamente dei residui degli amminoacidi e non degli atomi di backbone, in quanto si ipotizza che i primi siano più rilevanti ai fini della comprensione della capacità di legame.
Si sceglie una threshold dell'ordine dei ($\approx 10\mathring{A}$), e si definisce "sito di interazione" l'insieme dei residui poizionati ad una distanza minore della threshold da elementi dell'altra proteina del complesso. Questo perché entro i $10\mathring{A}$ siamo sicuri di considerare tutte le interazioni atomo-atomo di Van der Waals, poiché dopo questa distanza l'energia tende a zero.

Tenendo conto solamente dei residui nel sito di interazione, si procede a costruire il piano su cui sarà proiettato il sistema, che chiamiamo "piano di interazione". Questo viene costruito a partire dai seguenti tre punti:

1. Il centroide c_A della prima proteina, cioè il baricentro geometrico degli atomi dei suoi residui (esclusi gli idrogeni) presenti nel sito di interazione.

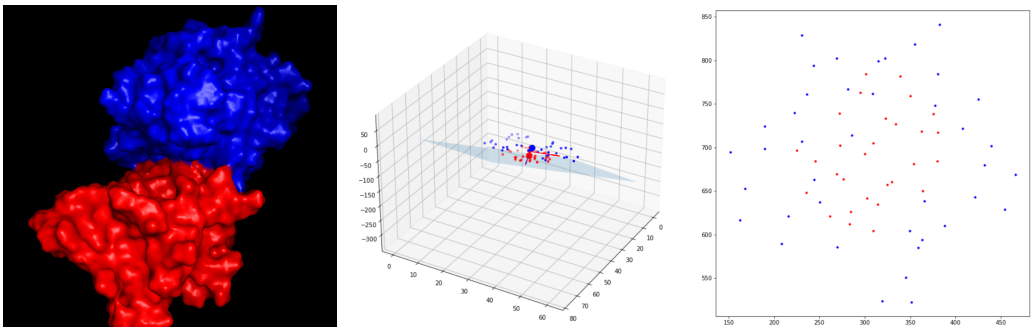
2. L'analogo della seconda proteina, c_B.

3. Il centroide di tutti i residui del sito di interazione di entrambe le proteine, c_{tot}.

Si traccia il segmento $\overline{c_A c_B}$ che collega i centroidi delle due proteine, e si definisce "piano di interazione" l'unico piano perpendicolare a $\overline{c_A c_B}$ e passante per c_{tot}, come è possibile osservare in Figura 1 (b).
Allo scopo di ridurre di dimensionalità il problema, si vuole proiettare su tale piano qualcosa che codifichi l'interazione tra le due proteine. Si sono seguite due differenti scelte: prima si considerano solamente informazioni geometriche, e poi si aggiungono anche informazioni chimiche. Nelle sezioni 4 e 5 sarà spiegato con maggior dettaglio che tipo di informazione è stato proiettato sul piano di interazione in ciascuno dei due casi. In sostanza, nei due casi l'algoritmo di proiezione è lo stesso, la differenza è che si proiettano sul piano di interazione differenti informazioni fisiche.
In conclusione, in entrambi i casi, si dispone di un piano su cui sono stati proiettati dei punti. A ciascun punto sono associate una o più grandezze fisiche. Nel caso dell'analisi geometrica, a ogni punto è associata una sola grandezza fisica (che chiamiamo canale); nell'altra analisi, a ciascun punto sono associati 5 canali. A partire da questi piani si mettono in atto due diverse metodologie di studio: una analisi radiale per studiare le caratteristiche del sito di interazione, e un modello di machine learning che impari a predire i valori della costante di associazione K_a.

3.2 L'analisi radiale

Come prima cosa, si analizza la distribuzione radiale delle features proiettate. Questo sarebbe il metodo ideale nel caso in cui il sito di binding abbia una simmetria circolare perfetta; in una certa misura i siti di interazione del dataset in esame tendono ad avere una forma circolare (come in Fig. 2 (b)), ma ci sono casi molto asimmetrici. Nonostante questa osservazione, si procede comunque a fare una analisi di questo tipo data la sua semplicità e interpretabiltà e il fatto che si studiano risultati statistici, e quindi anche se sono presenti alcune eccezioni rimane la possibilità che si riescano ad osservare comportamenti generali. Ipotizziamo quindi che il modo in cui si distribuiscono le interazioni tra il centro e l'esterno del sito di interazione sia un fattore importante per determinare la costante di associazione K_a del legame.
L'analisi si svolge nel seguente modo: partendo dal baricentro della distribuzione dei punti sul piano di interazione, si calcola la distanza di ciascun punto dal baricentro, e si ragruppano i punti in bin equispaziati. Si può immaginare questa operazione come un binnaggio effettuato tramite corone circolari. In ciascun bin si calcola la media di



(a) (b) (c)

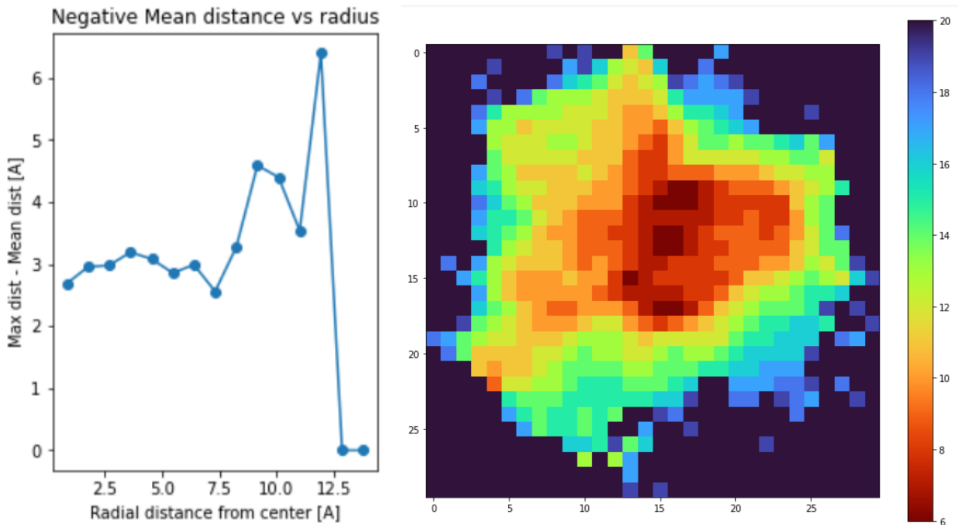
Figura 1: Operazioni eseguite dall'algoritmo di proiezione. (a) Rappresentazione del complesso proteico originale, ottenuto tramite X-ray cristallography. (b) Considerati solo i residui vicini al sito di interazione, si costruisce il piano di interazione. Per come è definito, taglia a metà in modo perpendicolare il segmento che collega i centroidi delle due proteine, che nella figura sono i due punti più grandi. (c) Esempio di proiezione di informazioni tridimensionali sul piano di interazione. In questo caso, i punti proiettati sono semplicemente le posizioni dei centroidi dei residui delle due proteine. Nell'immagine i punti sono colorati a seconda della proteina di appartenenza; in base alle grandezze fisiche che si tengono in considerazione, a ciascun punto sono associati uno o più valori.

ciascun canale: data la feature f, a seguito di questa operazione si ottiene quindi una funzione $f(r)$, come è posibile osservare in Fig. 2 (a).
Ripetendo questa operazione per ciascun complesso del dataset, e analizzando separatamente ciascun canale, si è interessati a trovare correlazioni tra gli andamenti di queste funzioni radiali e il coefficiente K_a.

3.3 Studio con Machine Learning

3.3.1 La rete CNN

Nell'analisi radiale, si era interessati a comprendere se gli andamenti delle $f(r)$ sono in qualche modo legati alla capacità delle due proteine di formare un legame stabile. Ora, invece, si cerca un metodo per predire la K_a di un complesso.
Una volta costruito il piano di interazione e proiettati i punti, si costruisce una griglia bidimensionale e si raggruppano i punti in ciascun elemento della griglia. Separatamente per ciascun canale si fa la media in ogni bin: interpretando ogni cella come un pixel, il risultato è una immagine a un colore per ogni canale, come l'esempio in Fig. 2 (b). Si vuole addestrare un modello di machine learning a predire la K_a di un complesso partendo da questa immagine.
Si sceglie come modello la Convolutional Neural Network (CNN), date le sue ottime capacità di lavorare con griglie di pixel [OSh+15]. E' una artificial neural network, cioè una rete di neuroni artificiali collegati in modo feed-forward (ovvero senza formare cicli).



(a) (b)

Figura 2: Comparazione dei differenti oggetti da cui si parte per effettuare l'analisi radiale e quella tramite machine learning. Entrambe le immagini si riferiscono a una sola grandezza fisica: nel caso si considerino più canali, si attua la stessa procedura per ciascun canale separatamente. (a) Nello studio radiale, si divide il piano di interazione in corone circolari, si raggruppano tutti i punti che cadono nello stesso bin, e si calcola la media per ciascun valore del raggio. Si ottiene dunque una funzione $f(r)$. (b) Per lo studio tramite machine learning, i punti vengono inseriti in una griglia, e si fa la media della grandezza fisica in ciascuna cella. L'immagine ottenuta è l'input del modello di machine learning.

E' divisa in layer: i primi layer sono la CNN vera e propria, i successivi layer sono fully-connected (ogni neurone è collegato con tutti quelli del layer precedente e successivo). Ogni neurone riceve un input dal layer precedente, e invia un output ai neuroni del layer successivo. Contiene al suo interno dei pesi che hanno il compito di modulare i segnali di input per creare il segnale di output: addestrare il modello significa modificare tale set di $O(10^5)$ pesi al fine di ottenere la migliore accuracy di previsione delle K_a. In realtà non si cerca di minimizzare l'accuracy, ma una funzione di loss differenziabile (in questo caso semplicemente la Mean Squared Error, che è uno stimatore dell'accuracy).
Si può interpretare un modello costruito in questo modo come costituito da una prima parte in cui l'immagine viene elaborata e trasformata in un encoding di alto livello, e una seconda parte in cui a partire dal segnale elaborato si calcola il valore della K_a. Decisa la struttura della rete, si lascia poi al modello stesso apprendere i pesi che devono essere associati ad ogni connessione tra neuroni al fine di effettuare la predizione con il minore errore possibile.
Riassumendo, il primo layer riceve in input la mesh creata dalla griglia sul piano di interazione (come l'esempio in Fig. 2 (b)), e l'ultimo layer restituisce in output il valore

della K_a predetto per quel particolare complesso di proteine.

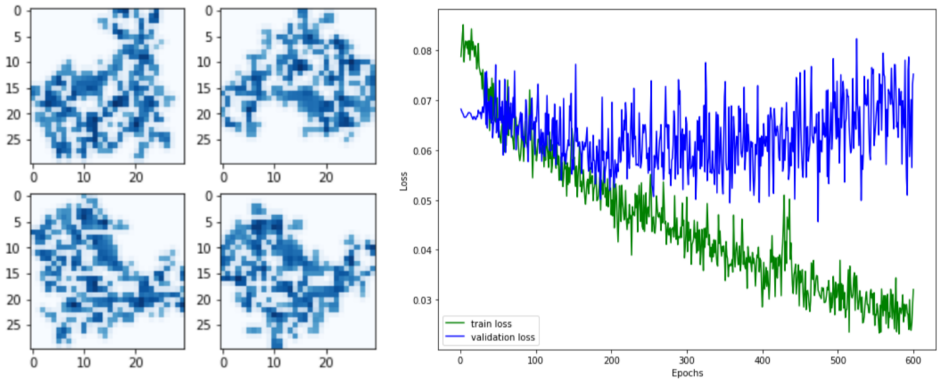
3.3.2 L'addestramento

Si divide il dataset in due parti: training set (circa l'80% dei complessi, scelti casualmente), e i rimanenti si inseriscono nel validation set. La limitatezza del dataset composto da 274 elementi non permette di distinguere un ulteriore gruppo, il test set. Durante l'addestramento, gli elementi del training set vengono presentati al modello, che viene lasciato libero di "apprendere", ovvero modificare i propri pesi per predire con minore errore la K_a di ciascun complesso. In realtà l'immagine che si presenta alla rete è leggermente differente ogni volta, poiché si applicano trasformazioni di data augmentation, come descritto in Fig. 3 (a). Dopo che tutti gli elementi del training set sono stati analizzati per una volta, si presentano al modello quelli del validation set. In questo caso si impedisce al modello di addestrarsi, e si valutano solamente le prestazioni a cui è giunto. L'insieme di queste due operazioni si chiama "epoca", e si itera per qualche centinaio di epoche. Alla fine, si sceglie il modello che ha avuto le migliori prestazioni sul validation set, poiché su di esso la rete non ha mai avuto la possibilità di addestrarsi. Si fa questa scelta per minimizzare l'overfitting, cioè il rischio che la rete abbia imparato a memoria i valori delle K_a dei complessi, e che non abbia appreso la fisica dietro al problema. L'idea è che siamo interessati a stimare le prestazioni che si avrebbero su dei complessi nuovi, su cui potenzialmente ancora non è disponibile il valore sperimentale della K_a.
Per definizione, il validation set è l'insieme dei modelli su cui la rete addestrata ha le migliori prestazioni. Questo induce un bias sulle sue reali prestazioni. Sarebbe importante disporre di un test set per stimare la reale performance della rete, ma la limitatezza del dataset non permette questa possibilità. Pertanto, sicuramente l'accuracy del modello è leggermente sovrastimata. Il punto di questo lavoro non è ottenere uno strumento in grado di prevedere il valore delle K_a nel modo migliore possibile, ma comprendere le potenzialità e le possibilità di utilizzo di metodi di machine learning nello studio dell'interazione tra proteine.

4 Risultati dello studio utilizzando solamente informazioni geometriche

4.1 Descrizione

Come primo tentativo, si è provato ad affrontare il problema provando a mantenere come unica informazione di cui si tiene conto la disposizione geometrica dei centroidi dei residui delle catene proteiche. Lavorando come descritto nell'algoritmo di proiezione, si crea il piano di interazione. Per ogni residuo nel sito di binding, se questo si trova entro una certa threshold dal centroide di un residuo dell'altra proteina, si calcola il punto medio delle due posizioni e si proietta tale punto sul piano di interazione. Ripetendo per ogni residuo, si ottengono quindi tutti i punti medi proiettati sul piano, come è possibile



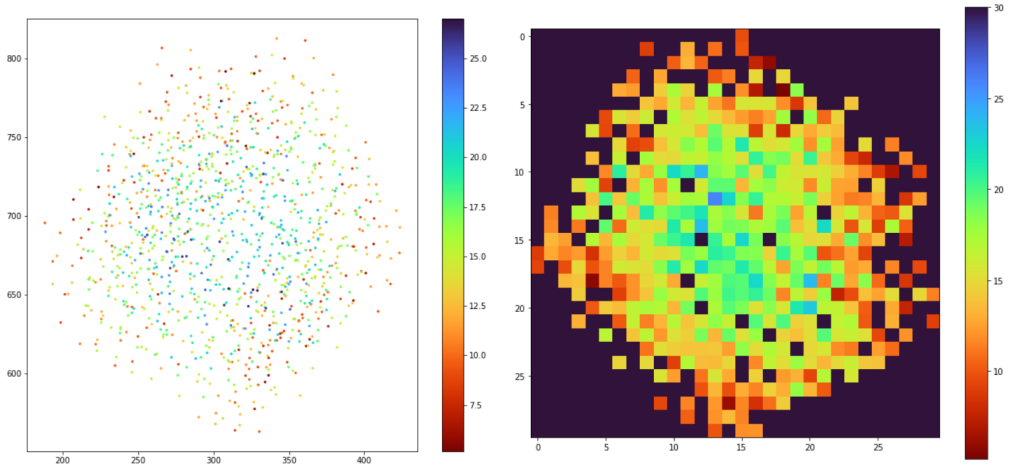
(a) (b)

Figura 3: (a) Esempio di data augmentation utilizzata per l'addestramento della rete: la stessa immagine viene presentata alla rete leggermente diversa a ogni iterazione, in particolare ruotata di un angolo casuale, oppure con i valori dei bin affetti da un piccolo rumore gaussiano. E' importante che le trasformazioni applicate siano fisiche, ovvero trasformino l'immagine secondo reali simmetrie del sistema. In questo caso, i due assi coordinati del piano di interazione sono scelti casualmente dall'algoritmo di proiezione, e quindi la figura ottenuta ruotandoli è fisicamente equivalente. Inoltre un leggero rumore gaussiano ricostruisce quella variabilità dovuta all'incertezza dei dati ottenuti sperimentalmente. (b) Esempio di addestramento: all'aumentare delle epoche, la rete ottiene prestazioni sempre migliori sul training set (curva verde). Per evitare overfitting, si decide quando fermare l'addestramento non appena l'errore sul validation set inizia a riaumentare (curva blu), ovvero circa all'epoca 200.

osservare in Fig. 4 (a). A ciascun punto medio si associa un'unica grandezza fisica, la distanza tra i due centroidi da cui è stato ricavato. Osserviamo quindi che questo primo metodo è ad un solo canale, per cui per ciascun complesso si ottiene un solo grafico $f(r)$ e una sola mesh.

4.2 Risultati dello studio radiale

Come prima analisi, si analizza l'andamento della feature radiale $f(r)$. L'unico canale in questo caso è il valore della distanza tra le coppie di centroidi. Il raggio di tutti i siti di interazione è stato riscalato, poiché siamo interessati a informazioni topologiche e non metriche. Non è una modifica drastica poiché i siti di binding sono tutti di dimensioni comparabili (circa $30\mathring{A}$ di diametro). Si fa quindi l'ipotesi che la forma del sito di binding sia più determinante nella stabilità del legame rispetto alle dimensioni metriche delle zone di contatto.
Come è possibile osservare in Fig. 5 (a), considerando solamente informazioni geometriche e topologiche, l'andamento della funzione radiale $f(r)$ sembra essere leggermente correlato al valore della costante di associazione. Si osserva in particolare che la cur-

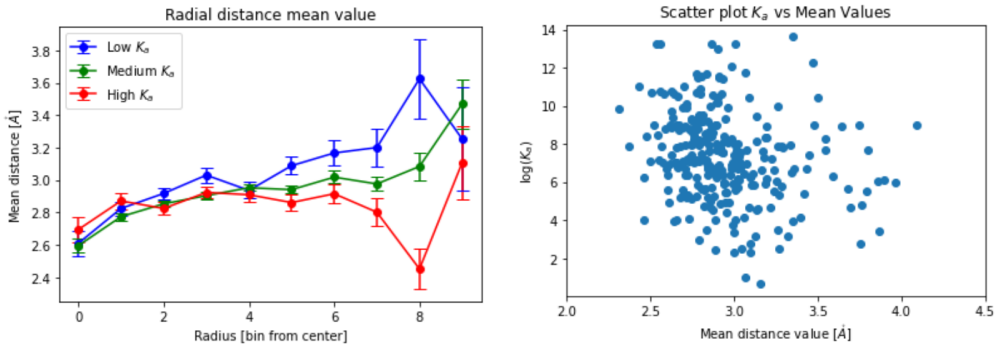


(a) (b)

Figura 4: (a) Risultato della proiezione sul piano di interazione di un complesso di proteine. Sono stati proiettati i punti medi, e ad ogni punto è associata la distanza tra i due centroidi di partenza (il colore, in $\mathring{A}$). (b) Esempio di una mesh ottenuta facendo un binning tramite una griglia. Nelle celle in cui non sono presenti punti, si inserisce un valore pari alla threshold. Sono state esplorate varie threshold (in questo caso, 30 $\mathring{A}$), ma sono stati ottenuti risultati interessanti in tutti i metodi di studio solamente per threshold dell'ordine dei 10 $\mathring{A}$.

va rossa (corrispondente a valori di K_a alti) è correlata a valori mediamente minori di distanze tra i centroidi dei residui. Questo diventa significativo a grandi distanze dal centro, cioè nei pressi dei bordi del sito di legame. Da questa analisi si comprende quindi che se nel bordo del sito di interazione i residui delle due catene sono vicini, mediamente il complesso sarà più stabile. Sembra quindi che ai fini dell'affinità di legame tra due proteine sia importante la presenza di una sorta di "cerniera" lungo il bordo del sito di interazione, piuttosto che le distanze medie nel centro del sito di interazione.
Nel grafico (b), per ciascun complesso è stato calcolato il valore medio di tutte le distanze tra residui, cioè senza tenere conto della posizione radiale. Come si osserva, non è presente correlazione tra stabilità del legame del complesso e distanza media tra residui. Questa informazione avvalora maggiormente l'intuizione che non è la distanza media tra i residui ad essere importante, ma la sua distribuzione radiale. Se al bordo i residui sono più vicini, il complesso tenderà ad essere più stabile.
Se da un lato i legami tra amminoacidi in una catena proteica sono un fenomeno complesso da analizzare e prevedere, poiché dipendono dalla tipologia di residui, dalla loro orientazione e dall'ambiente che li circonda, con una grande semplificazione si può immaginare che se due residui stanno interagendo, tenderanno ad essere più vicini. Con questa ipotesi si potrebbe spiegare l'andamento visto sperimentalmente: tendono ad essere più stabili quei complessi proteici i cui siti di interazione hanno forti legami ai bordi.

Sembrerebbe quindi da questa analisi che la forma geometrica dei siti di interazione sia un fattore importante per la stabilità del legame .



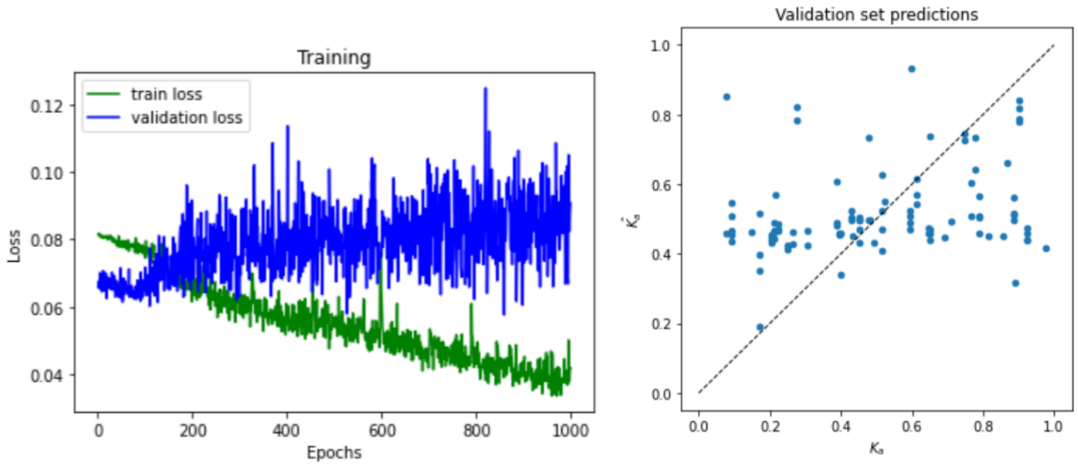
(a) (b)

Figura 5: (a) Andamenti sperimentali delle funzioni radiali della distanza tra i residui delle due catene di ciascun complesso. I complessi sono stati divisi in tre classi (alta, media e bassa K_a), ed è stata fatta la media per ciascuna classe. Ai bordi del sito di interazione (cioè a grandi valori del raggio) emerge un andamento significativo: la presenza di residui ravvicinati indica una maggiore stabilità di legame. (b) Assenza di correlazione tra il valore della K_a e la distanza media tra tutti i residui, senza tenere conto di informazioni radiali. Ciascun punto corrisponde a uno dei 274 complessi. La K_a si riporta in scala logaritmica poiché è la sua scala naturale.

4.3 Analisi con metodi di machine learning

Si addestra un modello CNN a predire la K_a di un complesso data la sua mesh sul piano di interazione. Poiché lo studio è ad un solo canale, ogni input consiste in una sola immagine di 30x30 pixel, come quella in Fig. 4 (b), e l'output è il valore predetto della K_a. In dettaglio, si sceglie casualmente il training set (con l'80% dei complessi), e i rimanenti dati vengono trattati come validation set. Come è possibile osservare in Fig. 6 (a), il modello si addestra con successo sul training set, cioè all'aumentare delle epoche diminuisce la Loss (e quindi migliora l'accuracy delle previsioni). Ma la curva della Loss sul validation set aumenta sin dal principio all'aumentare delle epoche, segnale che il modello sta overfittando. Non sta imparando una regola generale che permette di prevedere la K_a di un complesso, piuttosto sta imparando a memoria come associare correttamente a ogni esempio del training la sua K_a, e quindi più migliora in questa operazione, più statisticamente tende a commettere errori a prevedere i complessi del validation set (che sono a lui sconosciuti).
Il grafico (b) delle previsioni dei valori delle K_a rispetto al loro reale valore mostra esattamente l'incapacità del modello di generalizzare su esempi mai visti. Si può concludere quindi che il modello non riesce ad apprendere una regola che permette di recuperare correttamente i valori delle K_a dalle mesh costruite sul piano di interazione. Questo,

unitamente all'assenza di forti correlazioni tra le K_a e gli andamenti radiali analizzati nella sezione precedente, suggerisce la necessità di inserire nel metodo fin'ora seguito anche informazioni sulla chimica delle interazioni. Sembrerebbe quindi che usare esclusivamente informazioni geometriche riguardo l'interazione tra residui semplifichi troppo la descrizione della fisica del sistema, al punto che non è più possibile recuperare i valori delle K_a dei complessi.
Il successivo passo consisterà quindi nel differenziare i residui in base all'amminoacido a cui appartengono, inserendo informazioni chimiche oltre che geometriche nello studio del problema.



(a) (b)

Figura 6: (a) Esempio di addestramento del modello: mentre il modello migliora le prestazioni sul training set nel corso delle epoche, invece peggiora l'accuracy sul validation set. Questo è un segnale molto evidente della presenza di overfitting. Poiché l'overfitting inizia sin da subito, è possibile affermare che il modello non riesce a trovare l'informazione della K_a nei dati che sta studiando. (b) Incapacità del modello di compiere predizioni sul validation set. Sull'asse delle ascisse il reale valore della K_a di un complesso, su quello delle ordinate la predizione della CNN. I valori delle K_a del dataset considerato, in scala logaritmica, seguono una distribuzione gaussiana, e quindi molti hanno valori delle K_a nella media, mentre sono di meno quelli con valori alti o bassi. E' stata effettuata una trasformazione non lineare che ha portato tale distribuzione gaussiana in una distribuzione uniforme tra 0 e 1, perché i modelli di machine learning lavorano meglio con distribuzioni di questo tipo.

5 Analisi con informazioni geometriche e chimiche

5.1 L'uso dell'idropatia

Per aggiungere maggiori informazioni sulle interazioni in gioco nel sito di binding dei complessi proteici, è necessario scegliere che tipo di informazioni chimico-fisiche utilizzare. In particolare, per continuare a seguire un approccio minimale, è necessario che le informazioni inserite siano di semplice manipolazione e interpretazione. Si sceglie quindi di considerare l'idrofilicità e l'idrofobicità degli amminoacidi, e questo per varie ragioni:

1. A tutti gli atomi di un residuo si assegnano due numeri, corrispondenti all'idrofilicità e all'idrofobicità. Si approssimano quindi queste due proprietà come egualmente diffuse in tutto il residuo. In questo modo, la trattazione del problema continua ad essere di semplice analisi e interpretazione. Ovviamente questa è una semplificazione: la realtà è che l'idropatia dipende non solo dal residuo, ma è in parte indotta dall'ambiente circostante. Per uno studio con informazioni più avanzate, sarebbe quindi importante trovare un modo per tenere conto di questa caratteristica [Rie+21].

2. E' ben nota l'importanza di queste due proprietà nel comportamento delle proteine e quindi è probabile che, nonostante la semplicità della loro implementazione, queste due quantità possano arricchire lo studio del problema con informazioni utili.

In particolare, si utilizzano la scala di misura e i valori ottenuti computazionalmente in [Bon+14].

Type	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile
I_n	1.33	0.97	0.92	1.13	1.26	1.02	1.10	1.35	0.96	1.46
I_y	0.51	1.22	2.95	5.98	1.02	2.37	3.15	0.55	0.62	0.51

Type	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
I_n	1.47	0.99	1.38	1.36	1.38	1.00	1.13	1.25	1.20	1.45
I_y	0.54	1.71	0.65	0.54	0.53	1.95	1.92	1.22	0.66	0.51

Tabella 1: Tabella dei valori utilizzati di idrofobicità (I_n) e idrofilicità (I_y).

5.2 La scelta delle features

Rispetto a quanto fatto nella sezione precedente, oltre a considerare informazioni differenti, cambiano anche i punti che vengono proiettati sul piano di interazione. Ora si procede nel seguente modo: considerati solamente i residui nel sito di interazione, si proiettano sul piano di interazione tutti gli atomi dei residui della proteina che ne ha di più. Vengono esclusi da questa operazione gli atomi di idrogeno. Si compie la scelta di proiettare gli atomi di una proteina, e non più il punto medio del segmento che collega

due residui, perché ora utilizzando informazioni chimiche si è interessati strettamente ai legami che vengono a formarsi. Prima si cercava di studiare la forma generale del sito di legame, e quindi era utile conoscere informazioni geometriche di entrambe le proteine. Ora invece ci interessa la posizione dei legami, e in linea con il modello Lock and Key, nel limite in cui una proteina ha un sito di interazione molto più grande dell'altra, la superficie della prima avvolge la seconda, e quindi considerare solo la prima è un modo semplice per determinare le posizioni dei legami. Il modello Lock and Key assume che le posizioni dei siti di legame rimane invariata prima e dopo il binding. Questa è una assunzione, e sicuramente la scelta di proiettare solamente una proteina non è l'unica possibile, ma la sua semplicità e la capacità di non creare artefatti ne giustificano l'adozione.
A questo punto, per ciascun complesso, si ottiene quindi un insieme di punti su un piano, ciascuno corrispondente ad un atomo di un residuo della proteina con sito di interazione più grande. A ciascuno di questi punti vengono associate 5 grandezze:

- L'idrofobicità del residuo da cui deriva
- L'idrofilicità del residuo di provenienza
- La "cross-hydrophobicity" X_{phob}, definita come il valore medio del prodotto delle idrofobicità tra l'atomo del reiduo di partenza e di tutti quelli dell'altra proteina entro la threshold. Chiamando A la proteina con più residui nel sito di interazione, e considerando il suo i-esimo residuo, chiamiamo n_i l'insieme degli N_i atomi della proteina B che distano meno della threshold. Allora

  $$X^i_{phob} = \frac{1}{N_i} \sum_{j \in n_i} I_{n,i} I_{n,j}$$

  Si sceglie una grandezza di questo tipo perché è un modo molto semplice per combinare le proprietà delle due proteine.
- La "cross-hydrophilicity" X_{phil}, ovvero la stessa grandezza ma ottenuta a partire dalle idrofilicità:

  $$X^i_{phil} = \frac{1}{N_i} \sum_{j \in n_i} I_{y,i} I_{y,j}$$

- Il numero di vicini entro una threshold, cioè il numero N_i definito in precedenza.

Ora, scelti i punti da proiettare e le 5 feature da associare a ciascun punto, è possibile procedere all'analisi dei risultati che si ottengono.

5.3 Risultati dello studio radiale

Si lavora in modo analogo a quanto già fatto nella sezione precedente. La differenza ora è che ci sono 5 canali e non più uno, e ciascuno di questi si analizza separatamente. I primi e gli ultimi bin sono affetti da distorsioni dovute a effetti di bordo. Inoltre, gli stessi

grafici sono stati riprodotti per differenti valori della threshold. Gli andamenti qualitativi osservati in Fig. 7 sono gli stessi per una threshold compresa tra 4 $\AA$ ed 8 $\AA$, mentre nell'intervallo tra i 10-14 $\AA$ gli andamenti continuano ad essere uguali, ma tutti i punti delle curve tendono sempre più a diventare compatibili entro le barre di errore. Questo si potrebbe spiegare considerando che a valori minori di 10 $\AA$ si considerano solamente residui alle distanze tipiche dell'interazione di Van der Waals, e cioè che stanno realmente interagendo. Utilizzando invece threshold maggiori si considerano anche legami artefatti, non plausibili a causa della distanza tra gli atomi, e quindi i segnali si confondono con il rumore causato da questi legami fittizi. Vengono mostrati come rappresentativi i grafici ottenuti con una threshold di 8$\AA$, e di nuovo si dividono i grafici tra classi di valori del coefficiente K_a.

- La prima osservazione è che sia idrofobicità che cross-idrofobicità seguono lo stesso andamento (stessa cosa per idrofilicità e cross-idrofilicità). Poiché il primo grafico descrive solo una proteina, mentre l'altro tiene conto degli andamenti di entrambe, si intuisce che entrambe le proteine distribuiscono l'idrofobicità e l'idrofilicità in modo simile nel sito di interazione.

- Per quanto riguarda idrofobicità e cross-idrofobicità, nella zona di interesse a medi valori del raggio, l'andamento è circa costante, e compatibile tra tutte le classi di affinità di legame. Questo significa che tutti i complessi nella zona di legame distribuiscono allo stesso modo residui apolari. E' un effetto ben noto che inserire nella zona di legame porzioni idrofobiche consente di schermarle dall'acqua in cui le proteine sono immerse, riducendo di conseguenza l'energia libera e rendendo la configurazione più stabile. L'osservazione interessante è che sembra che tutti i complessi seguano questa regola allo stesso modo, indipendentemente dalla classe di K_a di appartenenza. L'unica deviazione da questo andamento circa costante si ha a raggi molto grandi, ovvero nei bordi del sito di interazione. Questa cosa è facilmente spiegabile considerando di nuovo che residui idrofobici stabilizzano la configurazione se sono lontani dall'acqua, cosa che non avviene più ai bordi.

- Guardando l'idrofilicità e la cross-idrofilicità, anche in questo caso, è interessante osservare come radialmente l'idrofilicità viene distribuita circa allo stesso modo indipendentemente dal valore del coefficiente K_a. Per questi due canali, è ben evidente un trend crescente dei valori medi all'aumentare del raggio, differentemente dall'idrofobicità che rimaneva circa a valori costanti. Di nuovo si potrebbe spiegare questo andamento con argomentazioni di energia libera totale e stabilità del legame date dalla presenza di più residui idrofilici nelle zone di bordo, dove possono interagire con le molecole di acqua circostanti.

- Infine, il grafico dell'andamento radiale del numero di vicini entro la threshold mostra un risultato interessante, poichè è l'unico nel quale ci sono diffenze tra le classi di K_a a tutte le threshold nell'intervallo 4-10 $\AA$. Nella zona centrale del sito di binding, cioè a valori del raggio medi e bassi, i residui dei complessi hanno un minor numero di vicini all'aumentare dell'affinità di legame. Questa evidenza è in

contrasto con l'idea che più vicini comportano mediamente più legami e quindi più stabilità. Questa evidenza potrebbe essere messa in relazione con il fatto che le proteine con minore densità tendono ad essere più termostabili [AGD17]. Quindi sembrerebbe essere importante un pò di spazio aggiuntivo tra i residui nel sito di binding. Per comprenderne meglio la motivazione potrebbe essere utile fare uno studio più approfondito.

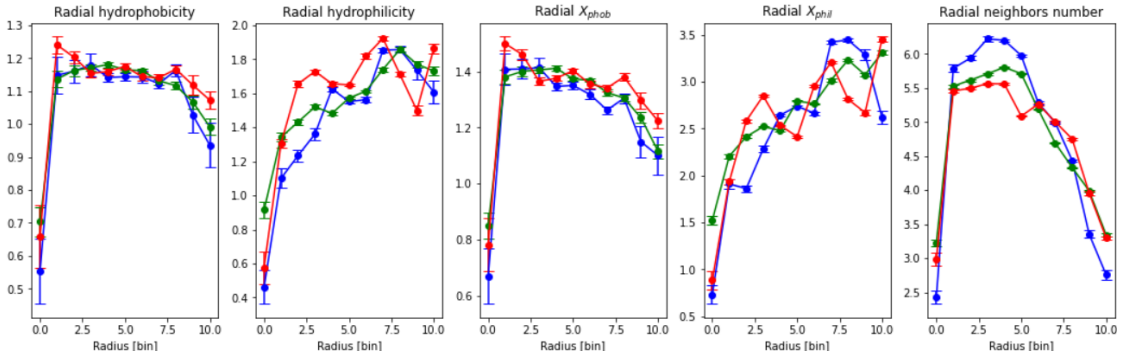


Figura 7: Funzioni radiali dei 5 canali analizzati, analizzate contemporaneamente su tutti i 274 complessi, divisi tra classi di K_a alte (in rosso), medie (in verde), basse (in blu).

5.4 Analisi con metodi di machine learning

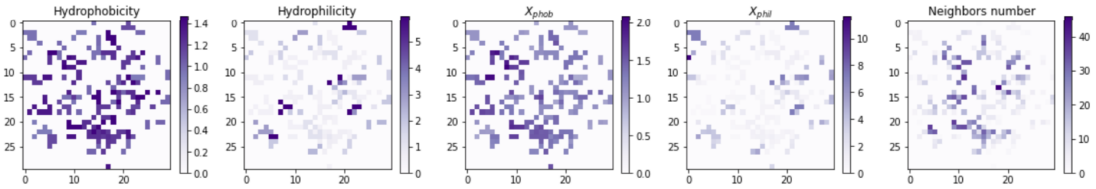
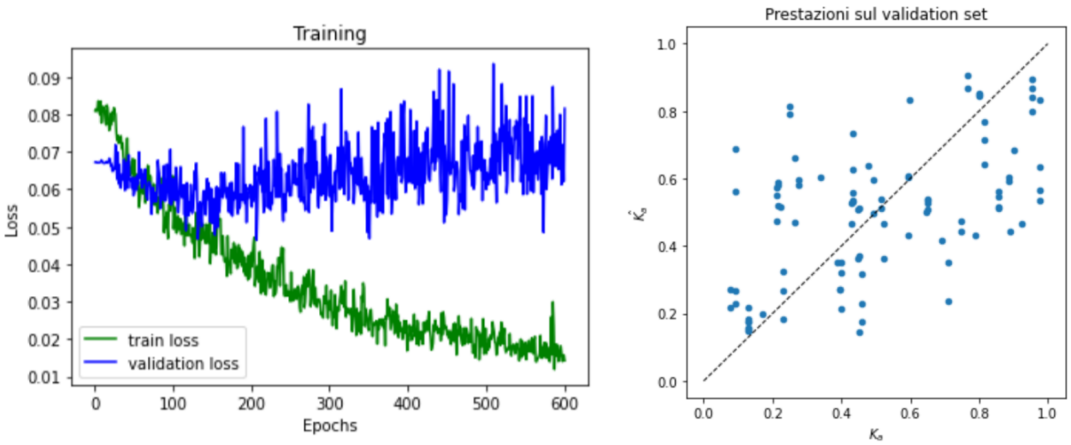


Figura 8: Esempio delle 5 mesh (una per canale) che si ottengono per ciascun complesso proteico, e che servono da input per il modello CNN. Si osserva come la figura è identica per ogni canale poiché costruita a partire dallo stesso insieme di punti proiettati sul piano di interazione; a cambiare sono i valori delle features.

Nuovamente, si addestra una CNN per prendere come input la mesh a 5 canali di ciascun complesso e predire il suo valore della K_a. Si ripetono le stesse divisioni tra training e validation set, e si procede all'addestramento e alla valutazione delle prestazioni ottenute allo stesso modo della sezione precedente.
In questo caso, differentemente al precedente, per le threshold minori o uguali di 10 $\AA$, il

modello riesce ad addestrarsi oltre che sul training set anche sul validation set. Per circa 150 epoche le prestazioni su quest'ultimo continuano a migliorare, che significa che fino a quel momento il modello sta realmente imparando qualcosa sulle regole che legano le mesh e il coefficiente K_a; dopodiché il miglioramento delle prestazioni è solamente overfitting. Si procede quindi con un early stopping, cioé si salva il modello che ha raggiunto le migliori prestazioni sul validation set intorno all'epoca 150, scartando quindi tutto l'addestramento successivo.
Questo modello non raggiunge prestazioni eccezionali o migliori di altri metodi di analisi senza Machine Learning, ma considerando l'estrema limitatezza del Dataset e le conseguenti limitazioni nell'efficacia dell'addestramento, i risultati raggiunti sono sicurmanete degni di nota. Inoltre è da considerare che l'interazione dei complessi originari è stata compressa e semplificata, perché si tiene conto solamente di un problema bidimensionale e con un insieme di informazioni minimo.
Date queste considerazioni, il fatto che il modello raggiunge un certo potere predittivo suggerisce che l'uso del machine learning nello studio dei complessi proteici - con dataset più grandi, l'utilizzo di informazioni più sofisticate e di conseguenza modelli più avanzati - potrebbe portare a risultati molto interessanti.
Ultimo aspetto da considerare è la velocità e la leggerezza computazionale della pipeline che porta a realizzare una predizione. La trasformazione di un complesso in formato PDB in una mesh a 5 canali richiede un tempo di circa 2 secondi (con un processore di basso livello). E per la CNN, ciascun addestramento ha una durata totale minore di 5 minuti; ma una volta addestrato, non è più necessario impiegare questo tempo e la GPU necessaria. Dopo la fase di training, per predire una K_a dall'immagine di input, la CNN impiega un tempo dell'ordine dei decimi di secondo. Quindi, utilizzando processori e dispositivi di basso livello, in un'ora è possibile trasformare in mesh e predire la K_a di un numero di siti di interazione dell'ordine dei 10^3. La notevole velocità computazionale potrebbe aprire la strada a utilizzi differenti di quelli descritti in questo lavoro. Un esempio potrebbe essere quello di determinare dove si trova il sito di interazione di un complesso, se questo non è noto a priori, tra molte possibili scelte fornite precedentemente da un algoritmo di docking molecolare. Un'altra applicazione potrebbe riguardare uno studio che tenga conto della dinamica del legame tra due proteine: dividendo in numerosi frame, e generando una immagine per ciascuno, si potrebbe applicare ripetutamente il modello, provando a gestire la dinamicità del processo studiando il variare delle informazioni statiche nel corso del tempo.



(a) (b)

Figura 9: (a) Un esempio di addestramento del modello con threshold ad 8 $\AA$. Durante le prime circa 150 epoche il modello realmente impara qualcosa della fisica del problema, poiché migliora le prestazioni sul validation set, composto di complessi a lui sconosciuti. La successiva fase di training viene scartata poiché è evidente la presenza di overfitting. (b) Esempio di predizioni sui complessi del validation set del modello addestrato. Il coefficiente di Pearson del grafico è di 0.52. Una volta effettuata una predizione, questa sarà compresa tra 0 e 1; per ottenere il valore della K_a predetta, è poi necessario effettuare la traformazione inversa di quella che era stata operata sui coefficienti di associazione.

6 Conclusione

Il lavoro compiuto è di tipo esplorativo: si è decisa una procedura per semplificare e trattare il sistema, e si sono analizzati i risultati lavorando in due modi leggermente differenti.
Nel primo caso, utilizzando informazioni esclusivamente geometriche e topologiche, dall'analisi radiale è emerso il comportamento interessante che un complesso di proteine chimicamente più stabile tende ad avere ai bordi i residui posizionati più vicini. Non conta invece la distanza media di tutti i residui del sito di legame, che non è correlata al valore della costante di associazione.
Nel secondo caso, inserendo nello studio anche l'idropatia, l'analisi radiale mostra che i residui idrofobici si distribuiscono in modo uniforme nel sito di interazione, mentre quelli idrofilici tendono a essere posizionati nei pressi del bordo. E' notevole il fatto che da questa analisi risulta che questi andamenti sono compatibilmente identici per tutte le classi di K_a. E' interessante e per certi versi inaspettato il risultato che l'unico trend significativo che distingue complessi con classi di K_a differenti è il numero medio di vicini dell'altra proteina di ciascun residuo. Sembrerebbe che siti di interazione con residui con maggiore spazio a disposizione siano più stabili. Sarebbe interessante investigare più in dettaglio questa proprietà.
Il modello CNN ha mostrato capacità di addestrarsi correttamente, e di avere potere predittivo dei valori della K_a. Le prestazioni raggiunte sono fortemente limitate dalla grande semplificazione del problema e dalla limitatezza dei dati, ma comunque sono significative. E' da sottolineare che l'assenza di un reale test set rende l'analisi delle prestazioni raggiunte affetta da problemi metodologici. In ogni caso, il punto importante di questo lavoro non è l'efficacia del metodo, ma mostrare che è possibile seguire questo approccio ed ottenere risultati interessanti.
Lo studio del comportamento microscopico delle proteine è un argomento di elevata complessità, dato il grande numero di interazioni in gioco. Allo stesso tempo, è di primaria importanza dato il loro ruolo come elementi di base della vita organica. Gli approcci sperimentali sono stati utilissimi negli ultimi 30 anni per la determinazione della forma tridimensionale delle catene proteiche, come la X-ray cristallography o la NMR spectroscopy. Nel frattempo, gli approcci computazionali sono stati un valido strumento per la comprensione della fisica e della chimica sottostante. Negli ultimi anni, l'emergenza di metodi di machine learning si è rivelata molto promettente, e sono stati ottenuti anche risultati rivoluzionari, come in [Jum+21]. Questi approcci sono indicati per lo studio delle proteine, poiché particolarmente adatti all'applicazione in sistemi molto complessi e con un grande numero di dati a disposizione. La principale limitazione di approcci del genere è la loro tendenza ad essere una "black box", cioè molto efficaci ma anche incapaci di spiegare cosa è stato appreso sulla fisica in gioco. E' vero che esistono modelli e metodi più "aperti", in grado di far comprendere allo scienziato quanto hanno appreso, ma nella maggioranza delle applicazioni questo è difficile o impossibile. Ma in ogni caso, disporre di strumenti in grado di prevedere in modo rapido e preciso caratteristiche fisiche e chimiche delle proteine e dei complessi che queste formano, potrebbe essere comunque utile:

non solo in loro applicazioni pratiche, ma anche come importanti strumenti di supporto in studi di ricerca di biofisica di base.

Riferimenti bibliografici

[Bon+14] S. Bonella et al. "Mapping the Hydropathy of Amino Acids Based on Their Local Solvation Structure". In: *J. Phys. Chem. B 2014, 118, 24, 6604–6613* (2014).

[OSh+15] O'Shea et al. "An introduction to convolutional neural networks". In: *arXiv preprint arXiv:1511.08458* (2015).

[AGD17] A. Amadei, S. Del Galdo e M. D'Abramo. "Density discriminates between thermophilic and mesophilic proteins". In: *Journal of Biomolecular Structure and Dynamics* (2017).

[Jum+21] J. Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature https://doi.org/10.1038/s41586-021-03819-2* (2021).

[Rie+21] L. Di Rienzo et al. "Characterizing Hydropathy of Amino Acid Side Chain in a Protein Environment by Investigating the Structural Changes of Water Molecules Network". In: *Front. Mol. Biosci. 8:626837* (2021).