UNIVERSITÀ DEGLI STUDI DEL SANNIO Benevento
DST
DIPARTIMENTO DI SCIENZE E TECNOLOGIE

Corso di Laurea Triennale in Biotecnologie

Tesi di Laurea in

BIOINFORMATICA

Virtual Screening basato su Machine Learning per l'identificazione di inibitori di CDK9

RELATORE

Prof. Francesco Napolitano

CANDIDATA

Diana Parente

Matr. 171001810

Anno Accademico 2025/2026

Indice

ABSTRACT

La scoperta di nuovi inibitori farmacologici rappresenta una fase cruciale nello sviluppo di terapie antitumorali mirate. Tra i target di interesse, la Cyclin-Dependent Kinase 9 (CDK9) svolge un ruolo chiave nella regolazione della trascrizione genica e della sopravvivenza cellulare, risultando frequentemente deregolata in diversi tipi di tumore. L'identificazione di nuovi composti inibitori attraverso approcci sperimentali tradizionali richiede tempi e costi elevati; pertanto, metodologie computazionali come il virtual screening basato su Machine Learning rappresentano una strategia promettente per ottimizzare le fasi iniziali della drug discovery.

L'obiettivo di questo lavoro è stato sviluppare un modello predittivo in grado di stimare l'attività biologica di composti chimici nei confronti di CDK9 a partire dalla loro struttura molecolare. È stato costruito un dataset di 426 molecole selezionate dal database pubblico ChEMBL. Le strutture chimiche sono state standardizzate e trasformate in rappresentazioni numeriche mediante Morgan fingerprints. Successivamente, sono stati implementati modelli supervisionati di regressione e classificazione utilizzando l'algoritmo Random Forest, con validazione mediante K-Fold Cross-Validation.

I risultati mostrano una capacità predittiva adeguata per la regressione (R^2 = 0,45; RMSE = 0,73) e un'elevata performance per la classificazione (ROC-AUC = 0,94), evidenziando una buona capacità del modello di distinguere composti attivi da inattivi. Nel complesso, il modello rappresenta quindi uno strumento promettente anche per l'identificazione di nuovi inibitori, fornendo un valido strumento per lo screening virtuale di grandi librerie di composti e contribuendo alla selezione razionale di molecole promettenti per successive validazioni sperimentali.

1. INTRODUZIONE

Il presente lavoro si inserisce nel contesto della ricerca di nuovi inibitori farmacologici diretti verso CDK9, una chinasi frequentemente associata a processi tumorali.
L'obiettivo principale della tesi è lo sviluppo di un modello predittivo basato su tecniche di Machine Learning.
In particolare, il lavoro si propone di costruire un classificatore capace di distinguere molecole attive da inattive, utilizzando fingerprint molecolari come descrittori numerici.
Parallelamente è stato sviluppato un modello di regressione finalizzato alla predizione del valore di pKi.
L'approccio adottato mira a dimostrare come l'integrazione tra chemoinformatica e Machine Learning possa costruire uno strumento efficace per supportare le fasi preliminari della drug discovery.

1.1 Stato dell'arte

Sono stati condotti numerosi lavori di virtual screening basato su machine learning per l'identificazione di inibitori di altre chinasi ciclina-dipendenti oltre a CDK9. In particolare, recenti studi hanno applicato modelli ML e deep learning per la selezione di composti attivi e selettivi contro CDK1, CDK2, CDK4, e CDK5, oltre a CDK9, utilizzando dataset ampi e descrittori molecolari diversificati. Questi approcci hanno permesso di derivare mappe di attività e selettività, evidenziando regioni chimiche favorevoli per l'inibizione specifica di ciascun CDK, con accuratezza predittiva elevata e validazione su test esterni [1].

L'inclusione di questi studi nel contesto della ricerca sugli inibitori di CDK9 è altamente rilevante, poiché la selettività tra le diverse chinasi CDK rappresenta una delle principali sfide nella scoperta di nuovi farmaci. I modelli multivariati e le strategie di screening virtuale sviluppate per altre CDK forniscono insight utili per

migliorare la selettività e ridurre gli effetti collaterali, e possono essere adattati o integrati nei workflow per CDK9. Inoltre, la comparazione tra le strutture attive e i profili di selettività dei diversi CDK consente di identificare scaffolds chimici e proprietà molecolari che favoriscono l'inibizione mirata [2][3].

In sintesi, la ricerca sugli inibitori di CDK9 beneficia significativamente dall'esperienza maturata con altre CDK, sia in termini di metodologie ML/DL sia di strategie di selettività, e l'integrazione di questi dati è fondamentale per lo sviluppo di nuovi agenti antitumorali.

Alla luce di queste evidenze, appare evidente come lo sviluppo di inibitori selettivi richieda non solo approcci computazionali avanzati, ma anche una solida comprensione delle caratteristiche strutturali e funzionali delle cicline chinasi-dipendenti, che verranno approfondite nel paragrafo seguente.

1.2 Chinasi ciclina – dipendenti

Le **chinasi ciclina-dipendenti (CDK)** sono enzimi serina/treonina fondamentali per la regolazione del ciclo cellulare e della trascrizione. Il loro ruolo principale consiste nel controllare le transizioni tra le diverse fasi del ciclo cellulare tramite l'interazione con le cicline, garantendo una proliferazione cellulare ordinata. La **deregulation delle CDK** è una caratteristica chiave della cancerogenesi: l'iperattivazione o la perdita di controllo di queste chinasi porta a proliferazione incontrollata, instabilità genomica e progressione tumorale. Le chinasi ciclina-dipendenti (Cdk), sono state identificate da Leland Hartwell, Paul Nurse e Timothy Hunt negli anni '70 e '80. L'attivazione cronologica delle rispettive Cdk in base alla rispettiva fase del ciclo cellulare G1, S, G2 o M è mediata dall'associazione con una subunità ciclina regolatrice, dalla fosforilazione delle Cdk e dal legame di attivatori e inibitori endogeni, nonché dalla localizzazione subcellulare [4].

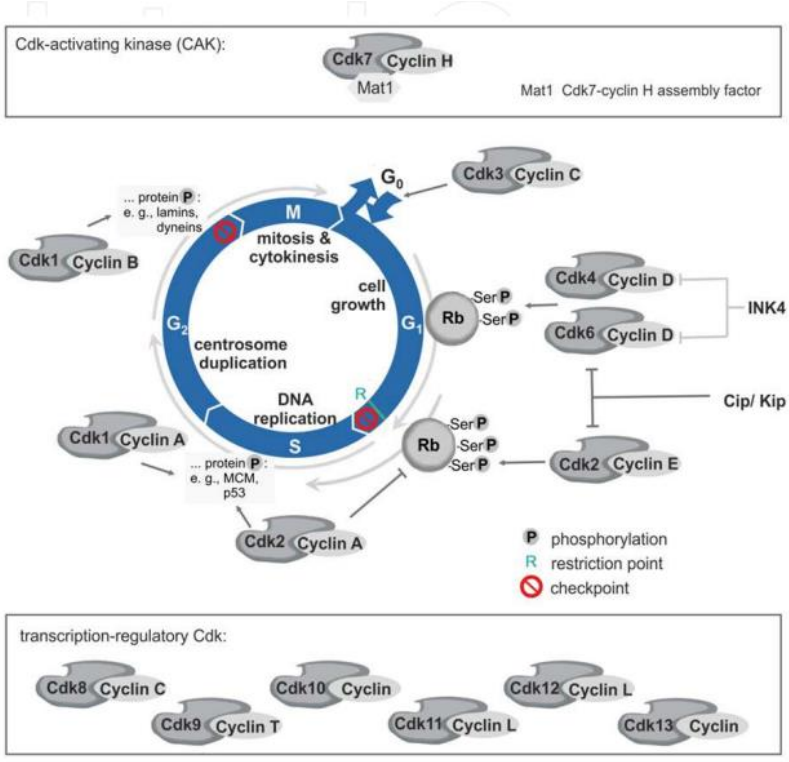

Figura 1. Panoramica dell'attivazione del ciclo cellulare umano e della regolazione trascrizionale attraverso i complessi CdkCyclin. Adattata da [4].

Il ciclo cellulare, che consiste in quattro fasi diverse, è un insieme di processi necessari per replicare una cellula eucariotica in due cellule figlie identiche. Durante l'interfase, che a sua volta comprende tre fasi, ovvero la fase di intervallo 1 (G1), la sintesi del DNA (S) e la fase di intervallo 2 (G2), la cellula si prepara ad entrare nella fase di mitosi (M) in cui il contenuto di DNA viene copiato e quindi avviene la divisione cellulare.

Nella fase G1, come prima fase di crescita, la cellula aumenta di dimensioni, gli organelli si dividono e vengono create le strutture molecolari necessarie per la crescita successiva. Come fase successiva, nella fase S, la cellula non solo sintetizza una copia completa del DNA, ma riproduce anche una struttura organizzativa dei microtubuli chiamata centrosoma. Nella fase G2, come ultimo stadio dell'interfase e secondo stadio della crescita, la cellula cresce sempre di più, produce più proteine e organelli e inizia a riorganizzare il suo contenuto per la mitosi. Se le cellule non riescono a mantenere la loro proliferazione a causa della mancanza di stimoli mitogenici sufficienti o della presenza di segnali antimitogenici, possono uscire dalla fase G1 ed entrare nella fase G0 (riposo), una fase in cui la cellula subiscono una fase di quiescenza. In particolare, la fase G0 è permanente per alcune cellule, mentre altre possono rientrare nel ciclo cellulare se ricevono segnali appropriati e riprendono la divisione. La transizione delle cellule nelle diverse fasi è sottoposta a un'intensa sorveglianza da parte dei meccanismi sensoriali chiamati checkpoint, in cui le CDK svolgono un ruolo fondamentale. I checkpoint sono infatti le penne correttrici che impediscono l'incidenza delle mutazioni acquisite durante la replicazione del DNA [5].

Le CDKs hanno strutture bilobate simili ad altre proteine chinasi, per cui la sezione principale della struttura monomerica contiene un'elica α nel terminale C, mentre il terminale N è costituito da fogli β. La fessura dell'adenosina trifosfato (ATP) all'interno del sito attivo si trova tra questi due siti. Da notare che il ciclo di attivazione o ciclo T nel lobo C-terminale occupa un'area ampia e flessibile per creare una barriera che impedisce l'ingresso del substrato proteico nel sito attivo. Vale la pena ricordare che il legame della ciclina al CDK determina solo un'attivazione parziale, e che per ottenere la piena attivazione è necessaria la fosforilazione del residuo di treonina vicino al sito attivo della chinasi da parte della chinasi attivante il CDK (CAK) [6][7].

A seconda delle cicline a cui si legano e delle loro subunità regolatorie, le CDK possono regolare la transizione delle cellule da qualsiasi fase del ciclo cellulare. Tuttavia, va notato che le funzioni biologiche di queste chinasi non sono limitate solo al ciclo cellulare, poiché questi gruppi di proteine regolatorie possono anche modulare il processo di trascrizione, l'elaborazione dell'RNA messaggero (mRNA) e la differenziazione cellulare.

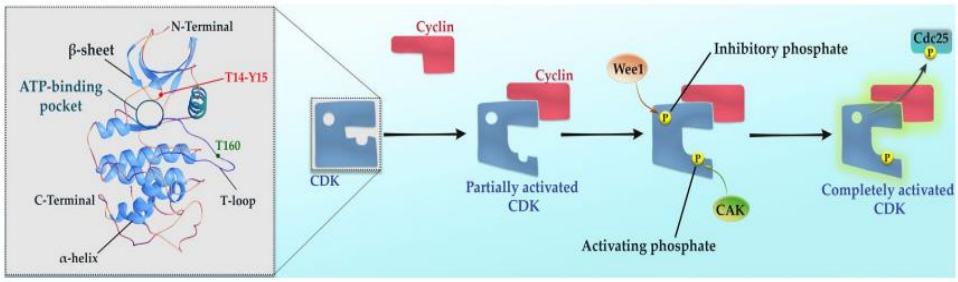

Figura 2. Rappresentazione schematica della struttura e del processo di attivazione delle CDKs. Adattata da [5].

Sulla base del sequenziamento, le CDK sono classificate in "CDK correlate al ciclo cellulare", tra cui CDK1,−2,−4,−6, e "CDK trascrizionali", costituite da CDK7,-8,-9,-11,-12,-13,-19 e -20 [8][9].

Una delle caratteristiche principali del cancro è la crescita incontrollata delle cellule, spesso causata da proteine note come chinasi ciclina-dipendenti, che controllano quando e come le cellule si dividono o producono molecole vitali come l'RNA. La crescita tumorale può derivare dal funzionamento improprio di queste proteine. I pazienti con alcuni tumori, come il cancro al seno, traggono beneficio da farmaci che già prendono di mira questi particolari tipi di proteine. È più difficile colpire in modo sicuro ed efficace altri tipi di queste chinasi, in particolare quelle coinvolte nell'espressione genica.

Nella pratica clinica, le CDK sono diventate target cruciali per lo sviluppo di farmaci antitumorali. In particolare, gli **inibitori selettivi di CDK4/6** (come palbociclib, ribociclib e abemaciclib) hanno mostrato efficacia significativa nel trattamento del carcinoma mammario HR-positivo/HER2-negativo, sia in fase avanzata che in setting

adiuvante. Oltre alle CDK4/6, sono in fase di studio inibitori di altre CDK coinvolte nella trascrizione (es. CDK7, CDK8, CDK9), che regolano l'espressione genica e la sopravvivenza delle cellule neoplastiche, ampliando le prospettive terapeutiche.
CDK7, CDK8 e CDK9 sono bersagli terapeutici desiderabili, poiché hanno dimostrato un ruolo oncogenico in diversi tipi di cancro, come quelli del colon-retto, delle ovaie e della mammella [10].

1.3 CDK9 e complesso P-TEFb (trascrizione)

Il fattore di allungamento della trascrizione positivo b (P-TEFb) è composto dalle cicline T1 o T2 e dalla chinasi ciclina-dipendente 9, che regolano la fase di allungamento della trascrizione da parte della RNA polimerasi II.
Nelle cellule umane esistono diversi complessi P-TEFb. Essi contengono CycT1, CycT2a o CycT2b e forme più corte o più lunghe di CDK9 (42 e 55 kDa). Le due isoforme di CDK9 sono prodotte da diversi siti di inizio della trascrizione nel primo esone del gene CDK9 [11].

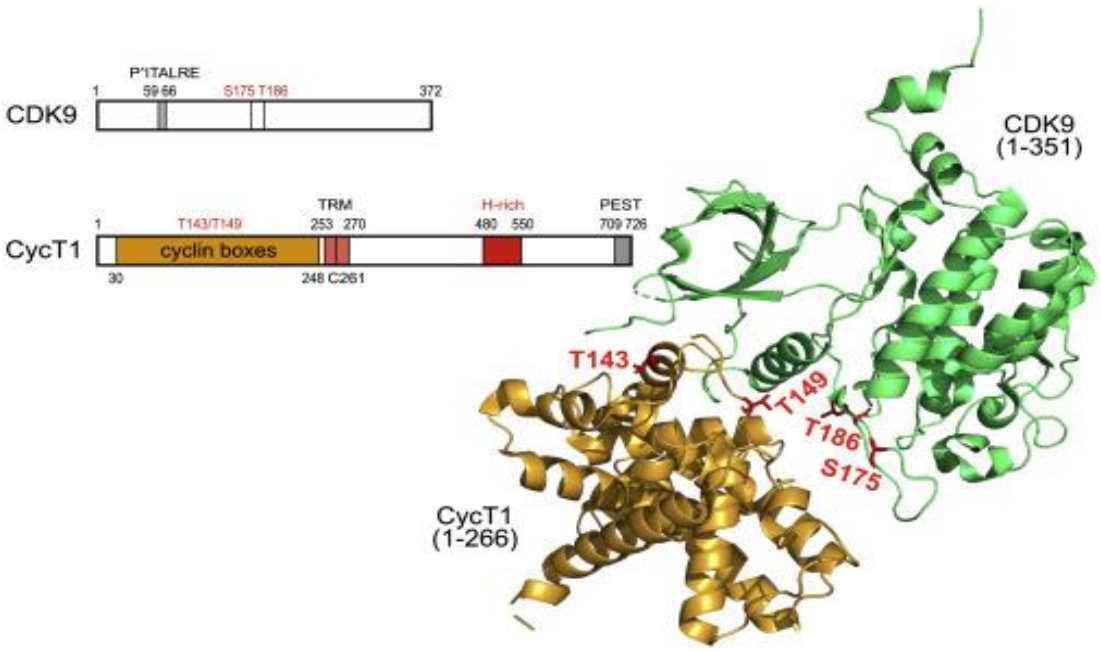

Figura 3. Rappresentazione schematica e struttura 3D delle proteine umane CDK9 e CycT1. Adattata da [11].

La chinasi ciclina-dipendente 9 (CDK9) è un regolatore critico dell'allungamento trascrizionale , operando all'interno del complesso del fattore di allungamento della trascrizione positivo b (P-TEFb) insieme alla ciclina T1 .

P-TEFb facilita il rilascio dell'RNA polimerasi II (RNAPII) dalla pausa prossimale al promotore, consentendo così un allungamento trascrizionale produttivo. L'attività di CDK9 è strettamente controllata dal complesso della piccola ribonucleoproteina nucleare 7SK (7SK snRNP), comprendente 7SK snRNA , LARP7, MEPCE e HEXIM1/2. In condizioni omeostatiche, 7SK snRNP sequestra e inattiva una frazione di P-TEFb, mantenendola in uno stato represso. Tuttavia, in risposta allo stress cellulare o all'aumento della domanda trascrizionale, P-TEFb viene rilasciato da 7SK snRNP, attivando CDK9 per garantire un controllo trascrizionale preciso e dipendente dal contesto. Questo interruttore regolatorio consente un adattamento dinamico a stimoli ambientali e intracellulari. Nuove evidenze implicano la deregolazione di 7SK snRNP nella progressione del cancro. L'interazione tra 7SK snRNP e CDK9, evidenzia come le alterazioni nei singoli componenti di 7SK snRNP determinino squilibri trascrizionali, amplifichino i programmi oncogenici e promuovano un ambiente tumorigenico.

L'inibizione dell'attività di CDK9 ha il potenziale per essere una terapia antitumorale altamente efficace [12].

1.4 CDK9 come target terapeutico (cancro)

CDK9 rappresenta un target farmacologico emergente: inibitori selettivi e degradatori di CDK9 sono in sviluppo clinico, con risultati promettenti in modelli preclinici e studi di fase precoce, soprattutto in neoplasie dipendenti da trascrizione aberrante. L'inibizione di CDK9 induce downregulation rapida di proteine oncosoppressive a breve emivita e può riattivare geni soppressori silenziati, potenziando anche l'efficacia di immunoterapie [13].

CDK9 è stato originariamente scoperto da Graña X. et.al. nel 1994, nel tentativo di identificare potenziali chinasi correlate a CDC2 (CDK1), come una proteina da 42 kDa e 372 aminoacidi (aa), denominata PITALRE a causa della presenza di un motivo

contenente Pro-Ile-Thr-Ala-Lue-Arg-Glu [14]. Questo motivo corrispondeva alla casella PSTAIRE altamente conservata, osservata in più CDK.

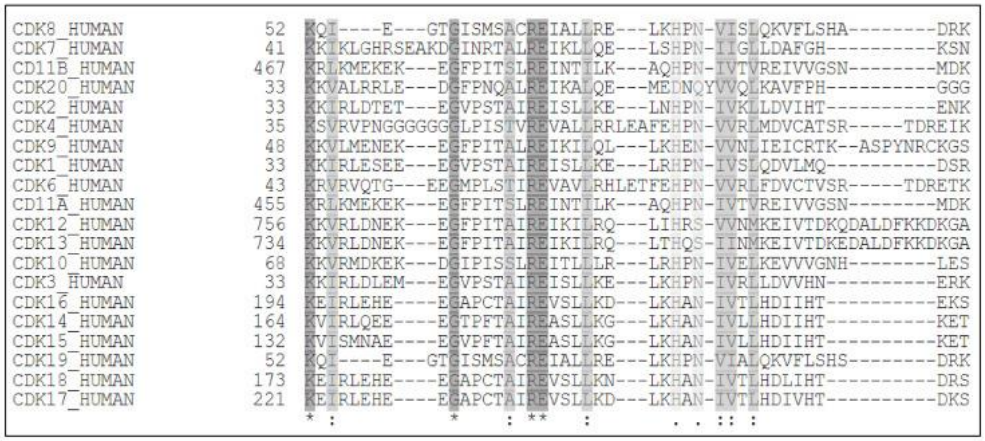

Figura 1. La sequenza PITALRE (Pro-Ile-Thr-Ala-Lue-Arg-Glu) di CDK9 che si allineava con la casella PSTAIRE altamente conservata, osservata in più CDK. Adattata da [14].

CDK9 esiste in due isoforme, quella originariamente identificata e più abbondante di 42 KDa e quella meno abbondante di 55 kDa, quest'ultima con 117 aa aggiuntivi al suo N-terminale. Queste due isoforme sono trascritte dallo stesso gene *CDK9* , ma da due promotori diversi, situati a più di 500 bp di distanza sul gene *CDK9* , con il promotore di 42 kDa che è significativamente più forte di quello di 55 kDa. All'interno del nucleo, CDK9 è localizzato principalmente nel nucleoplasma mentre CDK9 si accumula principalmente nel nucleolo [15].

L'attivazione di CDK9 non è regolata in modo dipendente dal ciclo cellulare, ma dalla sua associazione con i suoi partner ciclinici per formare un complesso eterodimerico Positive-Transcription Elongation Factor-b (P-TEFb). Anche se la ciclina T1 è il principale partner ciclinico di CDK9, partner minori come le cicline T2a e T2b possono anche attivare CDK9 [16].

L'espressione di CDK9 è stata implicata nella prognosi e nella resistenza alle terapie antitumorali in un gran numero di tipi di cancro come quelli del seno, del polmone, della prostata, dell'endometrio, oltre al melanoma, all'osteosarcoma, alla leucemia mieloide, ai sarcomi dei tessuti molli ecc.

Nel corso degli anni sono stati segnalati numerosi inibitori di CDK9, inclusi molti inibitori pan-CDK. Cassandri et al. nel loro articolo di revisione, hanno elencato una serie di composti naturali che prendono di mira le CDK come le indirubine, isolate da diverse piante produttrici di indaco. Sono stati segnalati anche un paio di derivati sintetici delle indirubine come la 6-bromoindirubina e la 6-bromoindirubina-3′-monoossima [17].

La maggior parte di questi inibitori si è mostrata promettente durante le sperimentazioni precliniche, molti hanno provocato gravi effetti collaterali o non sono stati efficaci nel generare SD o nel migliorare la OS e la PFS nei pazienti [13].

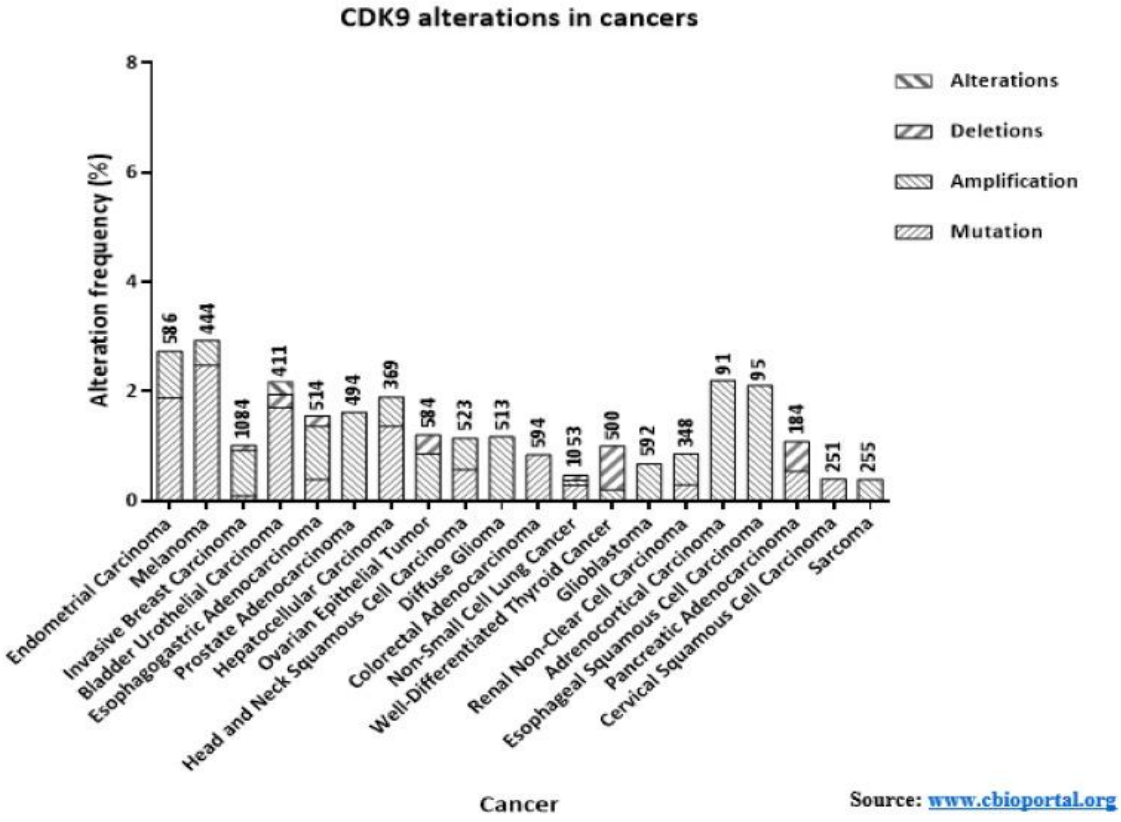

Figura 2. Alterazioni del gene CDK9 in diverse entità tumorali. Adattata da [13].

1.5 Virtual screening

Il virtual screening è una metodologia computazionale fondamentale nella ricerca farmacologica moderna, utilizzata per identificare rapidamente molecole potenzialmente attive da grandi librerie chimiche, riducendo drasticamente tempi e costi rispetto agli screening sperimentali tradizionali. Il suo ruolo principale consiste nell'analizzare e selezionare composti che presentano caratteristiche strutturali o funzionali compatibili con un target biologico di interesse, come una proteina coinvolta in una patologia.

L'importanza del virtual screening risiede nella sua capacità di esplorare uno spazio chimico vastissimo, anche di miliardi di composti, grazie all'integrazione di tecniche di docking molecolare, farmacofori, QSAR e, più recentemente, algoritmi di intelligenza artificiale e machine learning, che aumentano la precisione e la velocità del processo.

L'idea dello screening in silico, chiamato anche virtual screening, consiste nell'utilizzare un approccio computazionale per predire la capacità di una piccola molecola di interagire con un bersaglio biologico. È stato stimato che lo spazio chimico totale delle piccole molecole adatte alla scoperta di farmaci contenga oltre 10^60 composti possibili [18]. Gli screening tradizionali ad alta processività sono spesso limitati a decine di migliaia o centinaia di migliaia di composti, e la fascia più alta di questo intervallo è generalmente realizzabile solo in contesti di grandi aziende farmaceutiche. Gli screening virtuali, invece, sono attualmente in grado di analizzare miliardi di piccole molecole.

Il risultato di una procedura di virtual screening è una predizione delle affinità di legame tra ciascuna piccola molecola e il biomolecola bersaglio, tipicamente una proteina. Il processo di predizione della posa di legame e della relativa affinità di legame di una singola molecola per il bersaglio, noto anche come recettore, viene definito molecular docking. Una procedura di docking può essere suddivisa in due fasi indipendenti. La prima fase consiste nella ricerca dello spazio conformazionale sulla base dei gradi di libertà del recettore e del ligando. Per individuare disposizioni spaziali adatte di un ligando rispetto al bersaglio, lo spazio di ricerca può essere restrittivo, limitandosi a un singolo sito della proteina di interesse, come il sito attivo di un enzima, oppure molto ampio, comprendendo l'intera superficie del recettore, nel

cosiddetto blind docking. Questa prima fase viene spesso definita anche come generazione della posa.

La seconda fase consiste in un metodo di scoring, che assegna a ciascuna conformazione un valore correlato con l'affinità di legame del ligando verso il recettore. Se l'affinità di legame è elevata, allora il ligando rappresenta un buon candidato per ulteriori studi come potenziale composto hit o lead. Negli ultimi decenni sono stati sviluppati numerosi metodi per affrontare sia il campionamento dello spazio conformazionale sia la funzione di scoring dello screening, oltre che per esplorare lo spazio chimico [19]. Ad esempio, nuovi algoritmi provenienti dal campo dell'informatica sono stati adattati per campionare efficacemente lo spazio conformazionale, inclusi algoritmi ispirati a fenomeni naturali come l'ottimizzazione tramite grey wolf e ant colony optimization [20][21].

Le funzioni di scoring possono essere ampiamente suddivise in tre categorie principali: funzioni basate su force-field, funzioni empiriche e funzioni basate sulla conoscenza. Nelle implementazioni attuali delle procedure di docking esiste spesso un compromesso tra rigore computazionale e velocità [22]. Questo, combinato con le inaccuratezze intrinseche dei metodi di docking stessi, può portare a falsi positivi nello screening virtuale. Un modo per affrontare questo problema consiste nel rendere il metodo di docking più accurato. Un approccio promettente per migliorare l'accuratezza del docking è l'utilizzo di metodi basati sulla meccanica quantistica, che promettono risultati più accurati ma risultano computazionalmente molto onerosi.

Un approccio alternativo adottato da altri screening sfrutta metodi basati sul machine learning, che hanno il potenziale di essere computazionalmente più efficienti. Tuttavia, entrambi questi approcci sono utili solo se lo spazio chimico esplorato contiene piccole molecole adatte, capaci di legarsi strettamente al recettore. Studi recenti hanno indicato che gli ultra-large virtual screens, che esplorano uno spazio vastissimo di candidati possibili, non solo identificano composti hit e lead altamente potenti, ma presentano anche un tasso ridotto di falsi positivi grazie a una maggiore tolleranza all'errore tra i migliori hit virtuali [19]. Le principali applicazioni del virtual screening includono l'identificazione di nuovi inibitori enzimatici, modulatori di recettori, e molecole attive su bersagli difficili come le interfacce proteina-proteina, oltre alla riproposizione di farmaci già noti per nuove indicazioni terapeutiche. L'adozione diffusa di queste

tecniche da parte di industrie farmaceutiche e centri accademici ha reso il virtual screening uno strumento centrale per accelerare la scoperta di nuovi farmaci e ottimizzare le risorse disponibili.

1.6 Machine Learning applicato alla drug discovery

Le complessità insite nello sviluppo di farmaci sono molteplici. I bersagli, spesso proteine o enzimi complessi, possono nascondere meccanismi nascosti o siti allosterici, rendendo difficile l'intervento. Il vasto spazio chimico, con un numero astronomico di potenziali molecole, rappresenta una sfida ardua quando si tratta di identificare il giusto candidato farmacologico. Garantire la sicurezza, l'efficacia e l'accessibilità economica dei farmaci richiede anche di superare gli ostacoli normativi. Queste esigenze insoddisfatte hanno alimentato la ricerca di soluzioni innovative e l'intelligenza artificiale si è affermata come un potente strumento per rivoluzionare il processo di scoperta dei farmaci.

Il machine learning (ML) è un campo emergente derivato dall'intelligenza artificiale (AI), un'idea che ebbe inizio con Alan Turing negli anni Quaranta e che ha accompagnato la rivoluzione informatica negli ultimi quattro decenni. La scoperta dei nanomateriali a base di carbonio, inclusi i nanotubi di carbonio e il grafene, nel 2004 ha fornito un impulso allo sviluppo dei primi algoritmi di AI applicati a giochi come la dama, al riconoscimento di immagini facciali e alle auto a guida autonoma. In breve, il machine learning comprende lo studio e l'addestramento di diversi tipi di algoritmi che utilizzano input provenienti da dataset per predire in modo autonomo un output. Attraverso numerosi tentativi, l'utente migliora l'insieme di algoritmi, noto come modello, al fine di ottenere predizioni sempre più accurate.

Le complesse difficoltà legate ai bersagli patologici e l'enorme vastità dello spazio chimico hanno da tempo ostacolato la continua ricerca di farmaci efficaci. Gli approcci tradizionali basati sul target, pur avendo rivoluzionato la scoperta di farmaci, hanno mostrato anche alcuni limiti. Lo screening ad alta processività (HTS) ha consentito una rapida identificazione di composti lead, ma la presenza di falsi positivi e la scarsa considerazione delle proprietà drug-like hanno rallentato i progressi. Approcci basati su frammenti e sulla struttura, così come il virtual screening, hanno fornito maggiore precisione, ma alcune sfide sono rimaste irrisolte. Fortunatamente, l'avvento

dell'intelligenza artificiale attraverso metodi di machine learning e deep learning ha dato inizio a un'era trasformativa, potenziando ciascun approccio con capacità senza precedenti.

Nel contesto del machine learning, il campo può essere ampiamente suddiviso in apprendimento supervisionato e non supervisionato. L'apprendimento supervisionato consiste nell'addestrare algoritmi su dati etichettati per predire nuovi dati non osservati, mentre l'apprendimento non supervisionato identifica pattern e strutture all'interno di dati non etichettati. Esempi comuni di compiti di apprendimento supervisionato includono la classificazione e la regressione, mentre il clustering e la riduzione della dimensionalità rappresentano tipiche attività di apprendimento non supervisionato [23].

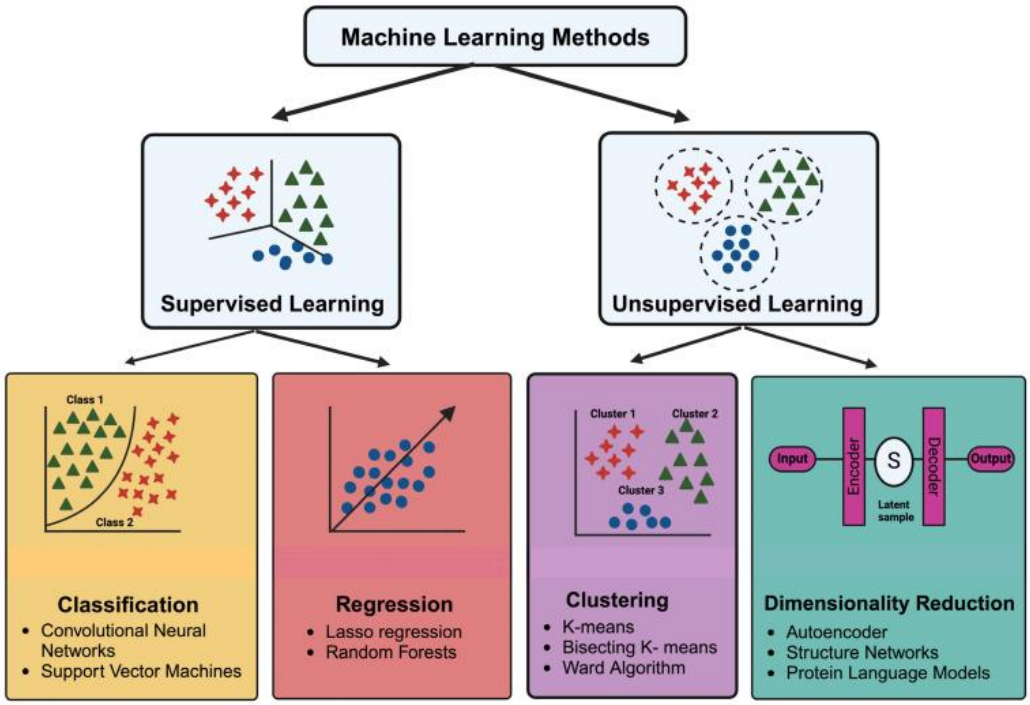

Figura 3. Esempio di algoritmi e classificatori nei modelli ML. Adattato da [23].

I vantaggi del machine learning sono la capacità di gestire dati ad alta dimensionalità, ridurre tempi e costi rispetto agli approcci tradizionali, aumentare la probabilità di successo nella selezione di candidati promettenti e automatizzare compiti complessi come la generazione di nuove strutture chimiche tramite modelli generativi e reti

neurali profonde. Inoltre, ML consente di integrare dati eterogenei (genomici, strutturali, clinici) per decisioni più informate e personalizzate.

Le sfide principali riguardano la qualità e la quantità dei dati disponibili, la scarsa interpretabilità dei modelli complessi (soprattutto deep learning), la riproducibilità dei risultati, la necessità di validazione sperimentale e le questioni etiche e regolatorie emergenti

Attualmente sono disponibili numerose librerie software per il calcolo matematico ad alte prestazioni su diverse piattaforme hardware, tra cui CPU, GPU e TPU, e su sistemi che vanno dai computer desktop ai cluster di server. Tra i framework più utilizzati vi sono TensorFlow, originariamente sviluppato da ricercatori e ingegneri del team Google Brain, oltre a PyTorch, Keras e Scikit-learn [24].

Gli algoritmi di machine learning attualmente più utilizzati e validati nella pratica clinica e preclinica per la drug discovery sono principalmente le reti neurali profonde (deep neural networks, DNNs), in particolare le reti neurali convoluzionali (CNNs) e le reti neurali ricorrenti (RNNs), con le varianti LSTM-RNNs per la gestione di dati sequenziali come SMILES e sequenze peptidiche [25]. Questi modelli sono impiegati per virtual screening, predizione dell'affinità di legame, generazione de novo di molecole e analisi di immagini biomediche.

Sebbene le reti neurali profonde rappresentino oggi uno degli strumenti più avanzati nel panorama del machine learning applicato alla drug discovery, numerosi studi dimostrano come algoritmi ensemble basati su alberi decisionali, quali Random Forest, offrano eccellenti performance predittive soprattutto in ambito QSAR e virtual screening ligand-based. Random Forest combina molteplici alberi decisionali addestrati su sottoinsiemi casuali dei dati e delle feature, migliorando la robustezza del modello, riducendo il rischio di overfitting e garantendo una buona interpretabilità attraverso l'analisi dell'importanza delle variabili. Per tali ragioni, nel presente lavoro è stato scelto di implementare un modello Random Forest per la classificazione e la regressione dell'attività biologica verso CDK9, in quanto rappresenta un compromesso efficace tra accuratezza, stabilità e interpretabilità dei risultati.

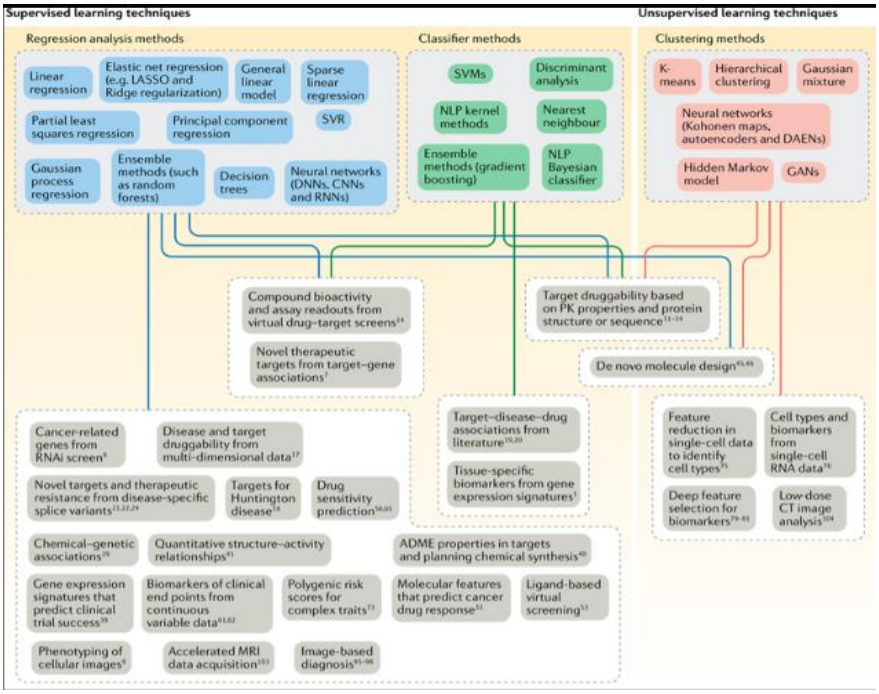

Figura 4. Strumenti di machine learning e loro applicazioni nella scoperta di farmaci. Adattato da [24].

Altri algoritmi ampiamente validati includono le autoencoder (AEs, VAEs, WAEs) per la generazione e ottimizzazione di nuove entità chimiche, i message passing neural networks (MPNNs) per la rappresentazione molecolare e la predizione di proprietà chimico-fisiche, e i modelli multitask deep neural networks (MTDNNs) per la predizione simultanea di molteplici endpoint farmacologici [25][26]. I modelli classici come random forest, support vector machines (SVMs) e naive Bayes sono ancora largamente impiegati per QSAR, classificazione e regressione di proprietà molecolari [27][28].

Questi algoritmi sono validati sia in fase preclinica (screening, ottimizzazione, predizione ADMET) sia in contesti clinici, come l'analisi di dati di trial, biomarcatori prognostici e digital pathology [24].

L'adozione di deep learning, in particolare, ha portato a una significativa accelerazione e miglioramento dell'accuratezza nella scoperta di nuovi farmaci, con validazione in vivo e in modelli animali per molecole generate o selezionate tramite ML.

L'adozione efficace di ML richiede anche competenze multidisciplinari e una stretta collaborazione tra informatici, chimici e biologi.

In sintesi, il machine learning sta rivoluzionando la scoperta di farmaci, ma il suo impatto dipende dalla qualità dei dati, dalla validazione dei modelli e dalla capacità di superare le barriere interpretative e regolatorie.

2. MATERIALI E METODI

Dopo aver delineato il contesto biologico e teorico alla base del presente lavoro, in questa sezione vengono descritti in dettaglio i materiali utilizzati e le procedure adottate per lo sviluppo del modello predittivo. L'obiettivo è illustrare in modo chiaro e riproducibile le fasi che hanno condotto dalla raccolta e preparazione dei dati fino alla costruzione e valutazione del modello di machine learning per l'identificazione di potenziali inibitori di CDK9.
In particolare verranno descritti il dataset impiegato, le strategie di preprocessing, la rappresentazione molecolare mediante fingerprint e la costruzione del modello Random Forest, insieme ai criteri di validazione adottati.

2.1 DATASET

Il dataset utilizzato in questo studio è stato estratto dal database ChEMBL, una risorsa pubblica che raccoglie dati bioattivi relativi a composti chimici e target proteici.
ChEMBL è un database open access, ampiamente utilizzato nella ricerca e scoperta di nuovi farmaci, che raccoglie dati curati manualmente su molecole bioattive con proprietà drug-like, inclusi farmaci approvati e candidati in sviluppo clinico. Le sue principali caratteristiche comprendono la presenza di dati standardizzati su attività biologica, strutture chimiche, bersagli molecolari, proprietà farmacocinetiche (ADMET), indicazioni terapeutiche, meccanismi d'azione, avvertenze e fasi di sviluppo clinico. ChEMBL integra dati provenienti dalla letteratura primaria di chimica farmaceutica, brevetti, screening di malattie trascurate, dati di metabolismo e studi in vivo, offrendo un contesto sperimentale dettagliato e una copertura trasversale di target e composti.
Sin dalla prima pubblicazione di ChEMBL nel 2009, gli scienziati che lavorano in diversi settori, tra cui il mondo accademico, gli istituti no-profit, le organizzazioni benefiche, le aziende biotecnologiche e le grandi organizzazioni globali, hanno potuto

accedere a grandi quantità di dati di alta qualità e curati su molecole bioattive provenienti dalla letteratura di chimica farmaceutica [29].

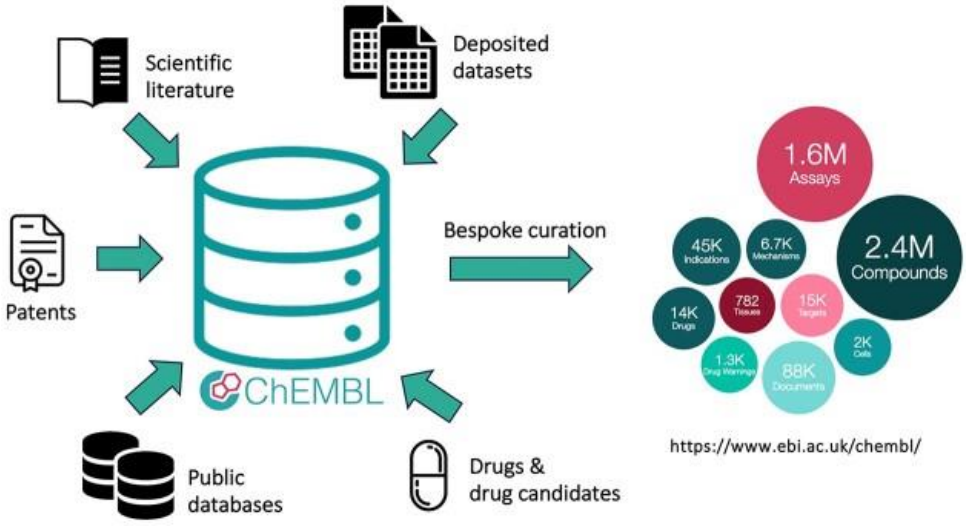

Figura 5. Adattato da [29].

L'inclusione di dati sui farmaci ricchi, di alta qualità e ben curati all'interno della struttura di un database FAIR di molecole bioattive su larga scala e ad accesso libero rende ChEMBL una risorsa di biologia chimica sempre più preziosa e affidabile. Svolge un ruolo centrale nel panorama delle risorse di biologia chimica ad accesso libero, mentre la comunità scientifica si muove verso un più ampio utilizzo dell'intelligenza artificiale nella scoperta di farmaci e la sua intrinseca dipendenza da dati di alta qualità.

I dati curati sui farmaci consentono alla comunità scientifica di rispondere a domande scientifiche importanti e pratiche [30].
Le applicazioni principali di ChEMBL includono il supporto al virtual screening, la modellistica QSAR, la selezione e validazione di target, la riproposizione di farmaci, la generazione di dataset per algoritmi di machine learning e la valutazione di relazioni struttura-attività. Il database è accessibile tramite interfaccia web, download diretto e servizi web RESTful, facilitando l'integrazione in workflow computazionali e applicazioni di chemoinformatica [31].
L'avvento delle tecnologie web, del cloud computing e delle architetture orientate ai microservizi ha reso le API REST un ingrediente essenziale dei moderni modelli di

sviluppo software. L'ampia disponibilità di strumenti che utilizzano risorse RESTful li ha resi utili per molti gruppi di utenti. L'API ChEMBL è una preziosa risorsa di dati sulla bioattività della scoperta di farmaci per chimici professionisti, studenti di chimica, data scientist, sviluppatori scientifici e web.

Il primo passo di ogni flusso di lavoro chemioinformatico efficiente è il recupero dei dati. ChEMBL facilita questo processo offrendo diversi canali per l'accesso e la distribuzione dei dati, tra cui un'interfaccia web, una piattaforma RDF, download FTP e API RESTful. [32] Dopo aver introdotto il database ChEMBL come risorsa pubblica di riferimento per la raccolta di dati bioattivi, è stato effettuato un processo di estrazione mirato allo studio di composti inibitori della **Cyclin-Dependent Kinase 9 (CDK9)**. CDK9 rappresenta un target farmacologico di grande interesse, in quanto coinvolto nella regolazione della trascrizione e associato a numerosi processi patologici, in particolare in ambito oncologico. Per tale motivo, l'identificazione di nuove molecole in grado di modulare l'attività di questa chinasi costituisce un obiettivo rilevante nel contesto della drug discovery.

Il dataset utilizzato in questo studio è stato ottenuto scaricando da ChEMBL un insieme di dati sperimentali relativi all'attività biologica di diversi composti contro CDK9 (Target ChEMBL ID: **CHEMBL3116**). In particolare, è stato generato un file in formato tabulare (activitiescdk9.tsv) contenente informazioni chimiche e farmacologiche relative ai ligandi associati al target selezionato.
Il dataset grezzo iniziale comprendeva **426 record sperimentali**, ciascuno associato a una specifica molecola identificata tramite il proprio **Molecule ChEMBL ID**, insieme alla rappresentazione strutturale in formato **SMILES** e ai valori quantitativi di attività biologica. Tuttavia, il file originale conteneva numerose colonne descrittive aggiuntive, non tutte necessarie ai fini della modellazione predittiva.
Per questa ragione, al fine di semplificare l'analisi e mantenere esclusivamente le informazioni rilevanti, sono state selezionate le seguenti variabili principali: identificativo della molecola (Molecule ChEMBL ID); struttura chimica in formato SMILES; valore sperimentale standardizzato (Standard Value); tipo di misura (Standard Type); identificativo del target (Target ChEMBL ID).
Successivamente, per garantire omogeneità nei dati sperimentali, è stato applicato un filtraggio mantenendo esclusivamente le misurazioni di tipo **Ki**, in quanto tale

parametro rappresenta una misura diretta dell'affinità di legame tra ligando e target ed è frequentemente utilizzato negli studi di inibizione enzimatica.

I valori di Ki sono stati convertiti in formato numerico ed è stata effettuata la rimozione di eventuali record incompleti o privi di misurazione valida. Inoltre, poiché alcune molecole risultavano presenti più volte nel dataset con misurazioni replicate, è stata eseguita una procedura di gestione dei duplicati. In tali casi, i valori multipli di Ki associati alla stessa molecola sono stati aggregati calcolandone la media, al fine di ottenere un singolo valore rappresentativo per ciascun composto. Al termine di queste operazioni di preprocessing, è stato ottenuto un dataset finale costituito da **414 molecole uniche**, ciascuna caratterizzata da una struttura chimica definita e da un valore sperimentale medio di attività biologica. Questo dataset rappresenta la base su cui sono state successivamente applicate tecniche di rappresentazione molecolare mediante fingerprint e algoritmi di Machine Learning supervisionato, con l'obiettivo di sviluppare modelli predittivi per l'identificazione di potenziali inibitori di CDK9.

2.2 Preprocessing dei dati

Prima dell'applicazione di algoritmi di Machine Learning, è fondamentale sottoporre i dati a una fase di **preprocessing**, al fine di garantire la qualità del dataset e migliorare l'affidabilità dei modelli predittivi.

Nel contesto della drug discovery e degli studi QSAR (Quantitative Structure–Activity Relationship), la fase di preprocessing rappresenta un passaggio cruciale, poiché i dataset provenienti da database pubblici come ChEMBL possono contenere valori mancanti, misurazioni eterogenee e molecole duplicate.

Nel presente lavoro, il preprocessing è stato eseguito sul dataset relativo a CDK9 con l'obiettivo di ottenere un insieme di dati coerente, pulito e adatto alla successiva fase di modellazione.

Il dataset iniziale estratto da ChEMBL includeva numerosi record sperimentali associati a differenti tipologie di misure farmacologiche.

Per garantire uniformità, sono stati selezionati esclusivamente i record contenenti valori di attività di tipo **Ki** (costante di inibizione). Tale parametro rappresenta una misura quantitativa dell'affinità di legame tra un ligando e il target proteico ed è

considerato uno degli endpoint più robusti negli studi di inibizione enzimatica. Successivamente, i valori di Ki sono stati convertiti in formato numerico, e sono stati eliminati eventuali record privi di misurazione valida o contenenti valori mancanti. Questa fase ha permesso di ridurre la presenza di rumore nei dati e di ottenere un dataset più affidabile.

Un aspetto comune nei dataset di bioattività è la presenza di molecole ripetute, dovuta al fatto che uno stesso composto può essere stato testato in differenti studi sperimentali o in condizioni diverse.

Nel dataset analizzato, alcune molecole risultavano presenti più volte con misurazioni replicate di Ki. Per evitare che tali duplicati potessero introdurre bias durante l'addestramento dei modelli, è stata implementata una procedura di aggregazione: per ciascuna molecola duplicata, i valori multipli di Ki sono stati combinati calcolandone la media.

Questa strategia consente di ottenere un singolo valore rappresentativo dell'attività biologica per ogni composto, riducendo la ridondanza e migliorando la consistenza del dataset. Al termine di questo processo, è stato ottenuto un insieme finale composto da **414 molecole uniche**, ciascuna associata a un singolo valore medio di attività. I valori di Ki presentano generalmente una distribuzione altamente asimmetrica, poiché possono variare su diversi ordini di grandezza (da valori molto bassi per composti altamente attivi a valori molto elevati per composti inattivi). Per rendere i dati più adatti all'analisi statistica e alla modellazione predittiva, è prassi comune applicare una trasformazione logaritmica. Nel presente studio, i valori di Ki (espressi in nanomolare, nM) sono stati convertiti in **pKi** utilizzando la seguente relazione: $pKi = -\log_{10}(Ki)$

Questa trasformazione consente di comprimere la scala dei valori sperimentali; ridurre la skewness della distribuzione; facilitare l'interpretazione, poiché valori più alti di pKi corrispondono a maggiore attività. Ad esempio, un composto con Ki molto basso (alta affinità) avrà un valore di pKi elevato, mentre un composto con Ki alto sarà associato a un pKi basso. Dopo la trasformazione in pKi, è stato necessario definire una strategia per trasformare il problema in una classificazione binaria supervisionata. A tal fine, i composti sono stati suddivisi in due classi principali: **composti attivi (classe 1); composti inattivi (classe 0).** La classificazione è stata effettuata applicando soglie comunemente utilizzate in letteratura per distinguere molecole con elevata

affinità verso un target: molecole con **pKi ≥ 7** sono state considerate attive; molecole con **pKi ≤ 6** sono state considerate inattive. I composti con valori intermedi sono stati esclusi o trattati separatamente, al fine di ottenere una separazione più netta tra le due categorie. Questa scelta consente di ridurre l'ambiguità nei dati e migliorare la capacità dei modelli di apprendere pattern strutturali associati all'attività biologica. Un elemento critico emerso durante l'analisi è la distribuzione non uniforme delle classi. Il dataset finale risulta infatti caratterizzato da un numero significativamente maggiore di composti inattivi rispetto agli attivi, condizione nota come **class imbalance**.

Questo fenomeno è particolarmente frequente nei dataset di drug discovery, poiché solo una piccola frazione dei composti testati mostra effettiva attività biologica significativa. Lo sbilanciamento può influenzare negativamente l'addestramento dei modelli, portandoli a predire prevalentemente la classe maggioritaria. Per tale motivo, nella valutazione delle prestazioni non è stata considerata esclusivamente l'accuracy, ma sono state utilizzate metriche più informative come precision, recall e F1-score. Per sviluppare modelli predittivi robusti, il dataset è stato suddiviso in: un insieme di addestramento (**training set**); un insieme di valutazione (**test set**). Questa procedura consente di valutare la capacità del modello di generalizzare su dati mai osservati durante l'addestramento. Inoltre, la suddivisione è stata effettuata in modo stratificato, mantenendo proporzioni simili tra composti attivi e inattivi in entrambi gli insiemi, condizione essenziale in presenza di class imbalance. Al termine delle operazioni descritte, è stato ottenuto un dataset pulito e pronto per la fase successiva, costituito da: molecole uniche rappresentate in formato SMILES; valori di attività trasformati in pKi; etichette binarie per la classificazione (Class 0/1). Questo dataset preprocessato è stato successivamente utilizzato per il calcolo dei descrittori molecolari e per l'addestramento dei modelli di Machine Learning.

2.3 Descrittori molecolari

Per applicare algoritmi di Machine Learning a problemi di drug discovery, è necessario convertire le strutture chimiche in una rappresentazione numerica. Questa trasformazione avviene tramite l'utilizzo di descrittori molecolari, ovvero

caratteristiche calcolate a partire dalla struttura che descrivono proprietà fisicochimiche e/o pattern strutturali.

I descrittori molecolari svolgono un ruolo centrale nella **predizione dell'attività biologica** dei composti chimici. Questi parametri numerici codificano le proprietà strutturali, fisico-chimiche e topologiche delle molecole, permettendo di correlare la struttura chimica con l'attività biologica tramite modelli computazionali come QSAR (Quantitative Structure-Activity Relationship).

La selezione accurata dei descrittori è fondamentale: descrittori rilevanti migliorano la capacità predittiva dei modelli, mentre quelli ridondanti o rumorosi possono causare overfitting e ridurre la robustezza della predizione [33].

La cheminformatica è una disciplina che integra metodi computazionali, gestione di dati chimici e algoritmi di machine learning per analizzare, modellare e prevedere le proprietà e l'attività biologica dei composti chimici. Il suo ruolo è fondamentale nella gestione di grandi librerie molecolari, nell'estrazione di descrittori molecolari e nella costruzione di modelli predittivi, accelerando la selezione di candidati promettenti per la sintesi e la sperimentazione biologica [34].

La Quantitative Structure-Activity Relationship (QSAR) rappresenta una delle principali applicazioni della cheminformatica. QSAR si basa sulla correlazione quantitativa tra i descrittori molecolari (che codificano proprietà chimico-fisiche e strutturali) e l'attività biologica osservata. I modelli QSAR, tramite algoritmi statistici e di machine learning, permettono di prevedere l'attività di nuovi composti, riducendo il numero di esperimenti necessari e ottimizzando il processo di drug discovery e valutazione tossicologica.

Vi è una procedura standard per lo sviluppo di modelli di relazione quantitativa struttura-attività (QSAR) implementata in due flussi di lavoro gratuiti e di facile utilizzo. Il primo flusso di lavoro aiuta l'utente a recuperare dati chimici (SMILES) dal web, verificandone la correttezza e curandoli per produrre set di dati coerenti e pronti all'uso per la chemioinformatica. Il secondo flusso di lavoro implementa sei metodi di apprendimento automatico per sviluppare modelli QSAR di classificazione. I modelli possono essere utilizzati anche per prevedere sostanze chimiche esterne. Il flusso di lavoro include anche il calcolo e la selezione dei descrittori chimici, la messa a punto degli iperparametri dei modelli e i metodi per gestire lo sbilanciamento dei dati [34].

L’affidabilità dei modelli QSAR dipende dalla qualità dei dati, dalla selezione dei descrittori e dalla validazione statistica, e la loro applicazione è ormai consolidata sia nell’industria farmaceutica che nella regolamentazione chimica [35].

In sintesi, **cheminformatica e QSAR** sono strumenti chiave per la predizione dell’attività biologica, consentendo screening virtuali ad alto rendimento, identificazione di lead compounds e supporto alle decisioni in ambito farmacologico e tossicologico [34]. I descrittori molecolari sono il ponte computazionale tra la struttura chimica e la funzione biologica, e la loro scelta e ottimizzazione sono cruciali per il successo delle strategie di drug discovery e sviluppo [36].

Recentemente, l’integrazione di descrittori biologici e chimici, anche tramite tecniche di machine learning, ha permesso di espandere la predizione dell’attività a composti poco caratterizzati, migliorando la comprensione delle relazioni struttura-attività e facilitando la scoperta di nuovi farmaci.

La maggior parte dello spazio chimico rimane ancora inesplorata, e individuare le regioni biologicamente rilevanti è fondamentale per la chimica medicinale e la biologia chimica. Per esplorare e catalogare questo enorme spazio, gli scienziati hanno sviluppato diversi descrittori chimici, capaci di codificare le proprietà fisico-chimiche e strutturali delle piccole molecole. Tra questi, le impronte molecolari (molecular fingerprints) rappresentano una delle forme più diffuse: si tratta di vettori binari (1/0) che indicano la presenza o assenza di specifiche sottostrutture molecolari. Queste codifiche sono alla base della chemioinformatica e risultano essenziali per la ricerca di similarità tra composti, il clustering, la drug discovery computazionale, l’ottimizzazione strutturale e la predizione dei bersagli farmacologici. I primi tentativi di sviluppare descrittori biologici per i composti chimici includevano l’affinità di legame ligando-recettore, mentre successivamente fingerprint basati sul profilo dei bersagli delle piccole molecole hanno rivelato numerose associazioni inattese e fisiologicamente rilevanti. Attualmente, le banche dati pubbliche contengono informazioni sperimentali di bioattività per circa un milione di molecole, che rappresentano però solo una piccola parte dei composti disponibili commercialmente e una frazione trascurabile dello spazio chimico teoricamente sintetizzabile [37].

I descrittori rappresentano il ponte tra la chimica e la biologia, consentendo di prevedere l’attività biologica di nuove molecole in modo efficiente e sistematico [38].

Le principali tipologie di **descrittori molecolari** utilizzate nella predizione dell'attività biologica dei composti chimici includono: **Descrittori fisico-chimici**, che rappresentano proprietà come peso molecolare, logP, punto di fusione, polarità e solubilità. Questi parametri sono fondamentali per correlare la struttura con la bioattività e sono ampiamente impiegati nei modelli QSAR [39].

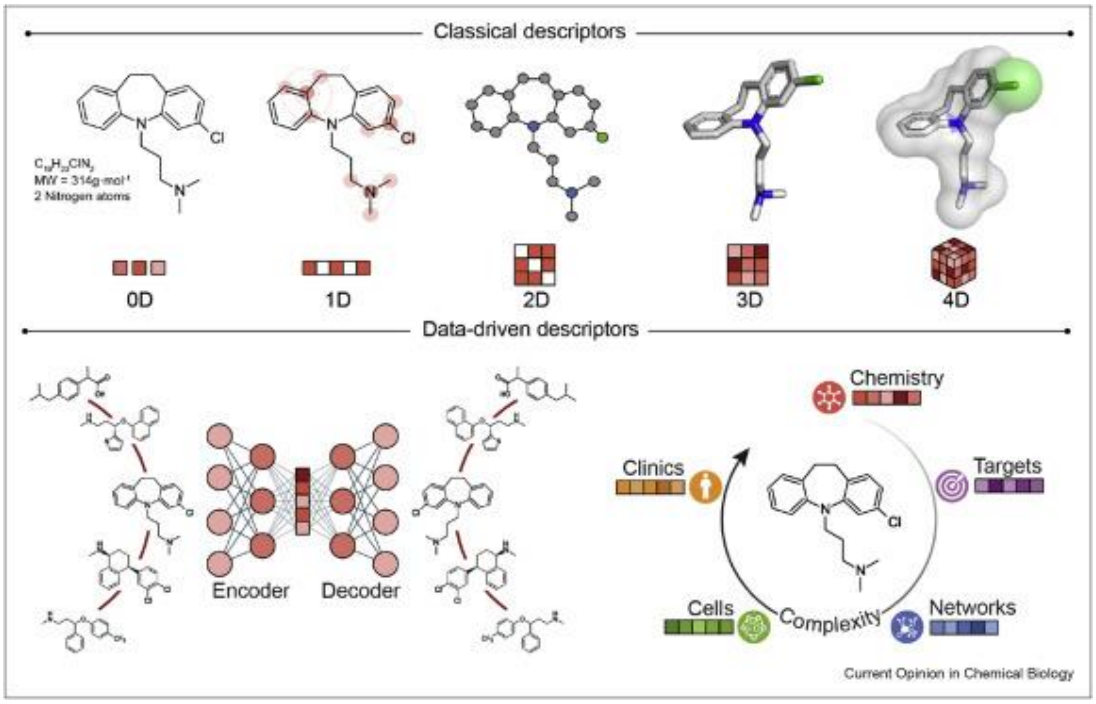

Figura 6. Descrittori fisico-chimici. Adattata da [39].

Descrittori strutturali/topologici, che codificano la connettività atomica, la presenza di specifici gruppi funzionali, la complessità molecolare e la forma. Esempi sono gli indici di Kier-Hall, i descrittori di connettività e i fingerprint molecolari (come ECFP4, MACCS keys) [40].

Descrittori farmacoforici, che identificano le caratteristiche essenziali per l'interazione con il target biologico, come donatori/accettori di idrogeno, centri aromatici e cariche [40].

Descrittori bioattività/data-driven, che derivano da dati sperimentali di attività biologica, profili di risposta cellulare o outcome clinici, e sono generati tramite tecniche di machine learning e deep learning per arricchire la rappresentazione molecolare [41].

La scelta e la combinazione di queste tipologie di descrittori sono cruciali per migliorare la performance predittiva dei modelli e la copertura dello spazio di bioattività.

Nel presente lavoro, la rappresentazione molecolare è stata effettuata tramite l'utilizzo di **fingerprint**, ovvero vettori binari che descrivono la presenza o assenza di particolari sottostrutture chimiche. Questa metodologia è ampiamente utilizzata nel virtual screening e nella classificazione di composti bioattivi. I fingerprint molecolari svolgono un ruolo centrale nella **drug discovery** grazie alla loro capacità di rappresentare in modo computazionale le caratteristiche strutturali e chimiche dei composti, facilitando la predizione dell'attività biologica. Questi algoritmi trasformano le molecole in stringhe binarie o vettori numerici che codificano la presenza di specifici frammenti, sub-strutture o pattern chimici, consentendo confronti

rapidi tra grandi insiemi di composti e la selezione di candidati promettenti tramite virtual screening [42].

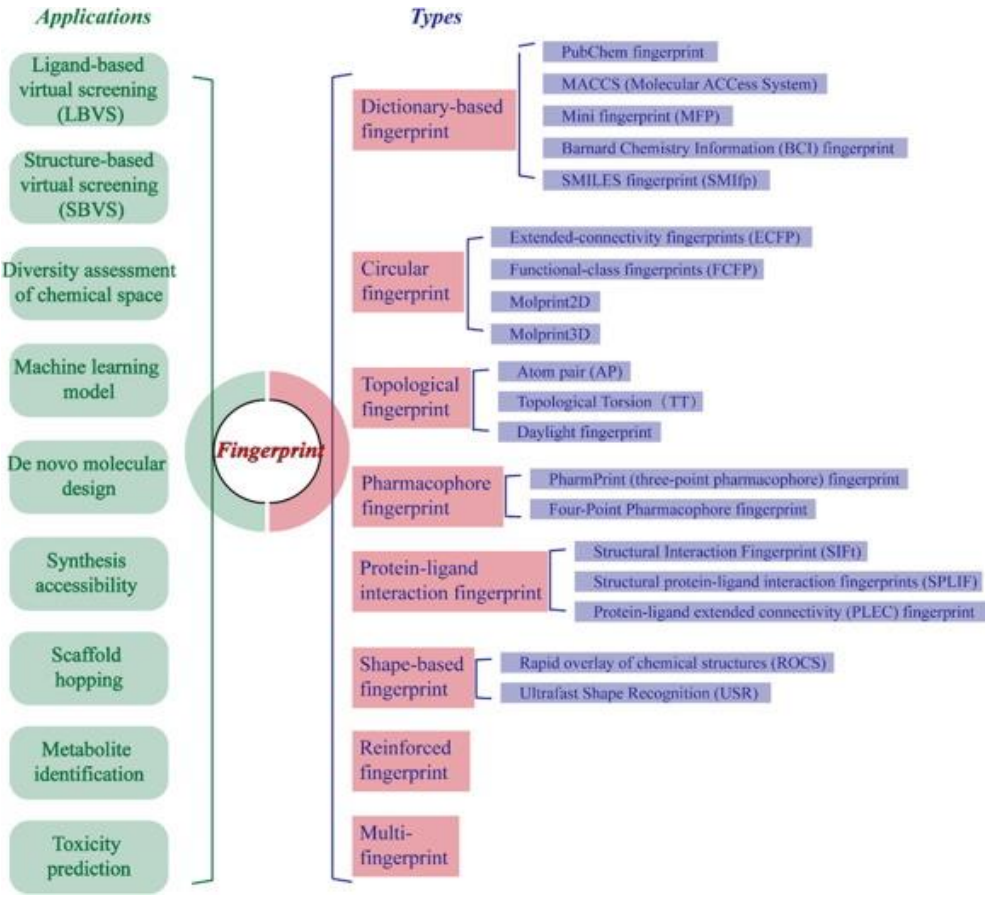

Figura 7. Panoramica dei tipi e delle applicazioni più comuni dei fingerprints. Adattata da [42].

I fingerprint vengono utilizzati per identificare composti con similarità strutturale rispetto a molecole note attive, sotto il principio che molecole simili tendono ad avere attività simili. Sono impiegati sia in modelli QSAR che in algoritmi di machine learning, dove fungono da input per la classificazione o la regressione dell'attività biologica su specifici target.

I Morgan Fingerprints, noti anche come Extended Connectivity Fingerprints (ECFP), rappresentano uno degli strumenti più utilizzati nella drug discovery e

nella predizione dell'attività biologica dei composti chimici. Questi fingerprint codificano la struttura molecolare in vettori binari o numerici, identificando e rappresentando in modo efficiente i pattern di connettività atomica e i sottostrutture chimiche rilevanti per l'attività biologica [43].

L'efficacia degli ECFP nei modelli QSAR deriva dalla loro capacità di catturare dettagli strutturali rilevanti, inclusa la stereochimica, e dalla loro flessibilità nell'adattarsi a molecole di diversa complessità. Gli ECFP sono particolarmente adatti per la classificazione di composti attivi/inattivi, la predizione quantitativa di attività biologica e il virtual screening, mostrando performance predittive competitive rispetto ad altri tipi di fingerprint e descrittori. [40] In particolare, l'utilizzo di ECFP4 e ECFP6 come input per algoritmi di machine learning (come reti neurali e random forest) consente di ottenere modelli robusti e generalizzabili per la scoperta di nuovi lead e la prioritizzazione di composti [44]. Inoltre, i Morgan Fingerprints sono apprezzati per la loro interpretabilità: la presenza di specifici bit può essere ricondotta a sottostrutture chimiche definite, facilitando l'analisi delle relazioni struttura-attività [45].

Per il calcolo dei descrittori molecolari è stata utilizzata la libreria open-source **RDKit**, uno degli strumenti più diffusi per l'analisi computazionale di strutture chimiche in Python.

Le molecole sono state convertite in fingerprint numeriche e successivamente utilizzate come input per i modelli di Machine Learning.

RDKit è una libreria open source ampiamente utilizzata in cheminformatica per il calcolo dei descrittori molecolari e la conversione delle strutture chimiche in fingerprint numeriche. RDKit consente di importare e gestire strutture chimiche (ad esempio da SMILES o file SDF), calcolare una vasta gamma di descrittori 1D, 2D e 3D, e generare fingerprint come i Morgan Fingerprints (ECFP4, ECFP6), che rappresentano la presenza di sottostrutture chimiche in forma vettoriale binaria o numerica.

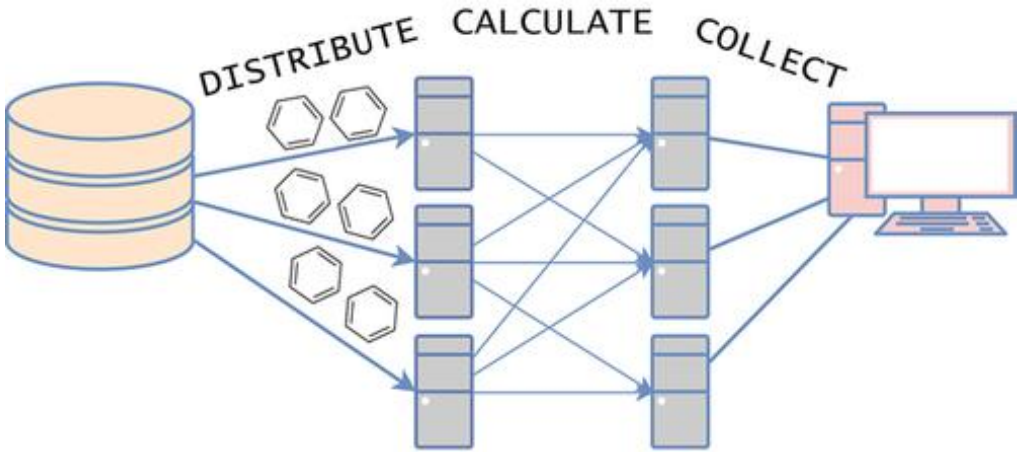

Figura 8. Adattata da [45].

RDKit rappresenta uno standard per la generazione e gestione di descrittori e fingerprint molecolari, abilitando la predizione computazionale dell'attività biologica tramite modelli QSAR e screening virtuale su larga scala.

I modelli di machine learning più efficaci e validati per la predizione dell'attività biologica che utilizzano fingerprint molecolari generati dalla libreria open source RDKit sono principalmente **Random Forest**, **Support Vector Machine (SVM)**, **XGBoost**, **reti neurali artificiali (ANN)** e, in alcuni contesti, **k-Nearest Neighbors (KNN)** e **Naive Bayes**. Questi algoritmi sono frequentemente applicati a fingerprint come Morgan (ECFP4/ECFP6), MACCS e RDKit fingerprints, tutte ottenibili tramite RDKit, per costruire modelli QSAR robusti e generalizzabili [46].

L'input nei modelli di machine learning per la predizione dell'attività biologica è una rappresentazione tabellare in cui ogni riga corrisponde a una molecola e ogni colonna a una feature molecolare (descrittore, fingerprint, embedding) prende il nome di feature matrix. La costruzione di questa matrice permette di integrare diverse modalità di rappresentazione delle molecole, come fingerprint binari (ad esempio ECFP4/ECFP6), descrittori numerici (fisico-chimici, topologici) o rappresentazioni apprese tramite deep learning, ognuna con specifici vantaggi in termini di informazione strutturale e predittività [47]. Un problema centrale è la **dimensionalità elevata**: molte rappresentazioni, come i fingerprint estesi, generano migliaia di feature, spesso con una prevalenza di zeri (vettori sparsi). L'uso di **vettori sparsi** è vantaggioso perché consente di gestire grandi matrici riducendo il carico computazionale e la

memoria necessaria, oltre a mitigare il rischio di overfitting e migliorare la robustezza dei modelli. Tuttavia, la dimensionalità elevata può introdurre rumore e ridondanza, rendendo necessarie tecniche di selezione o riduzione delle feature, come PCA o metodi basati su regolarizzazione, per ottimizzare la performance predittiva [48]. Le modalità di rappresentazione delle feature molecolari includono fingerprint binari, descrittori numerici, rappresentazioni 3D, e feature apprese tramite reti neurali, ciascuna con impatti diversi sulla capacità del modello di cogliere relazioni struttura-attività e sulla scalabilità computazionale.

Nel presente studio, per rappresentare numericamente le molecole del dataset, sono state utilizzate Morgan Fingerprints (ECFP) generate tramite la libreria RDKit. In particolare, ciascuna molecola è stata convertita in un vettore binario di lunghezza 2048 bit (radius = 2), che costituisce l'input per l'addestramento dei modelli di Machine Learning descritti nella sezione successiva.

2.4 Modelli di Machine Learning

Una volta ottenuta la rappresentazione numerica delle molecole attraverso l'impiego di descrittori molecolari, è stato possibile applicare algoritmi di Machine Learning supervisionato per sviluppare modelli predittivi in grado di stimare l'attività biologica dei composti nei confronti del target CDK9.

Nel presente studio, l'obiettivo principale è la costruzione di un modello QSAR basato su fingerprint molecolari, finalizzato alla discriminazione tra molecole attive e inattive. Questo approccio rientra nell'ambito del ligand-based virtual screening, in cui la predizione dell'attività avviene sulla base delle caratteristiche strutturali dei ligandi noti, senza la necessità di utilizzare direttamente la struttura tridimensionale del recettore. In particolare, ciascuna molecola del dataset è stata codificata mediante Morgan Fingerprints (ECFP), generate tramite la libreria RDKit, ottenendo vettori binari di lunghezza 2048 bit. Tali fingerprint costituiscono la matrice delle feature utilizzata come input per l'addestramento dei modelli di classificazione.

L'attività biologica sperimentale, espressa come valore di pKi, è stata utilizzata per assegnare a ciascun composto un'etichetta di classe, distinguendo tra composti attivi e

inattivi secondo soglie definite nella fase di preprocessing. Il problema è stato pertanto formulato come una classificazione binaria supervisionata.

Tra i diversi algoritmi di Machine Learning applicabili in ambito QSAR, nel presente lavoro è stato scelto un approccio basato su Random Forest, un metodo ensemble particolarmente adatto alla gestione di dataset ad alta dimensionalità e caratterizzati da feature sparse, come nel caso delle fingerprint molecolari. Random Forest consente inoltre di ridurre il rischio di overfitting grazie alla combinazione di molteplici alberi decisionali e all'aggregazione delle predizioni tramite voto di maggioranza. Poiché il dataset risulta fortemente sbilanciato, con una netta prevalenza di composti inattivi rispetto agli attivi, la valutazione delle prestazioni del modello è stata condotta utilizzando metriche specifiche per problemi di classificazione, tra cui precision, recall, F1-score e confusion matrix, al fine di ottenere una stima più accurata della capacità del modello di identificare potenziali molecole attive.

Per sviluppare un modello di Machine Learning in grado di generalizzare correttamente su nuovi composti, è fondamentale suddividere il dataset disponibile in insiemi distinti per l'addestramento e la valutazione.

Nel presente lavoro, il dataset preprocessato, costituito da 327 molecole uniche (260 inattive 67 attive) rappresentate tramite fingerprint molecolari, è stato suddiviso in due sottoinsiemi principali: **Training set**, utilizzato per l'addestramento del modello;

Test set, utilizzato esclusivamente per la valutazione finale delle prestazioni predittive. Questa separazione consente di simulare una condizione realistica, in cui il modello deve effettuare predizioni su molecole mai osservate durante la fase di apprendimento. La suddivisione è stata effettuata mediante una procedura di **train-test split**, mantenendo una proporzione tipica dell'80% dei dati destinati all'addestramento e del 20% riservati al test. Inoltre, data la presenza di un forte sbilanciamento tra composti inattivi e attivi, la divisione è stata eseguita in modo **stratificato**, preservando la stessa distribuzione delle classi in entrambi gli insiemi. Questa scelta è particolarmente importante nei problemi di drug discovery, poiché un modello addestrato su dati sbilanciati potrebbe tendere a predire prevalentemente la classe maggioritaria (inattivi), riducendo la capacità di identificare correttamente i composti realmente attivi.

Una volta definita la suddivisione, il training set è stato utilizzato per addestrare un classificatore supervisionato basato su Random Forest. Durante questa fase, il modello apprende associazioni tra i pattern strutturali codificati nei fingerprint e l'attività biologica assegnata alle molecole. L'insieme di test, mantenuto separato e non utilizzato durante l'addestramento, è stato successivamente impiegato per valutare la capacità del modello di generalizzare, attraverso il calcolo di metriche di performance e la costruzione della confusion matrix. Questa procedura rappresenta uno standard consolidato nello sviluppo di modelli QSAR predittivi, garantendo una valutazione affidabile e riproducibile dell'efficacia del metodo proposto.

L'**algoritmo di Random Forest** è una tecnica di machine learning supervisionato che costruisce una "foresta" composta da molti **alberi decisionali** indipendenti. Un **albero decisionale** suddivide i dati in modo gerarchico, creando nodi che rappresentano scelte basate sui valori delle feature molecolari, fino a raggiungere foglie che corrispondono a una predizione (ad esempio, attività biologica). La "foresta" è l'insieme di questi alberi, ciascuno addestrato su un campione casuale dei dati e su un sottoinsieme casuale delle feature, e la predizione finale è ottenuta aggregando (mediante voto o media) le predizioni di tutti gli alberi [49].

I passaggi usuali intrapresi durante lo sviluppo di un modello Random Forest (RF) includono il preprocessamento e il filtraggio dei dati, l'estrazione delle feature, la selezione delle feature, l'inizializzazione del modello, l'addestramento del modello, la validazione e la valutazione.

Random Forest funziona particolarmente bene con fingerprint molecolari perché è in grado di gestire dati ad alta dimensionalità e vettori sparsi, tipici delle rappresentazioni come ECFP4/ECFP6. L'algoritmo è robusto rispetto a feature irrilevanti o rumorose, riduce il rischio di overfitting grazie all'aggregazione di molteplici alberi e può identificare pattern non lineari tra struttura chimica e attività biologica [50].

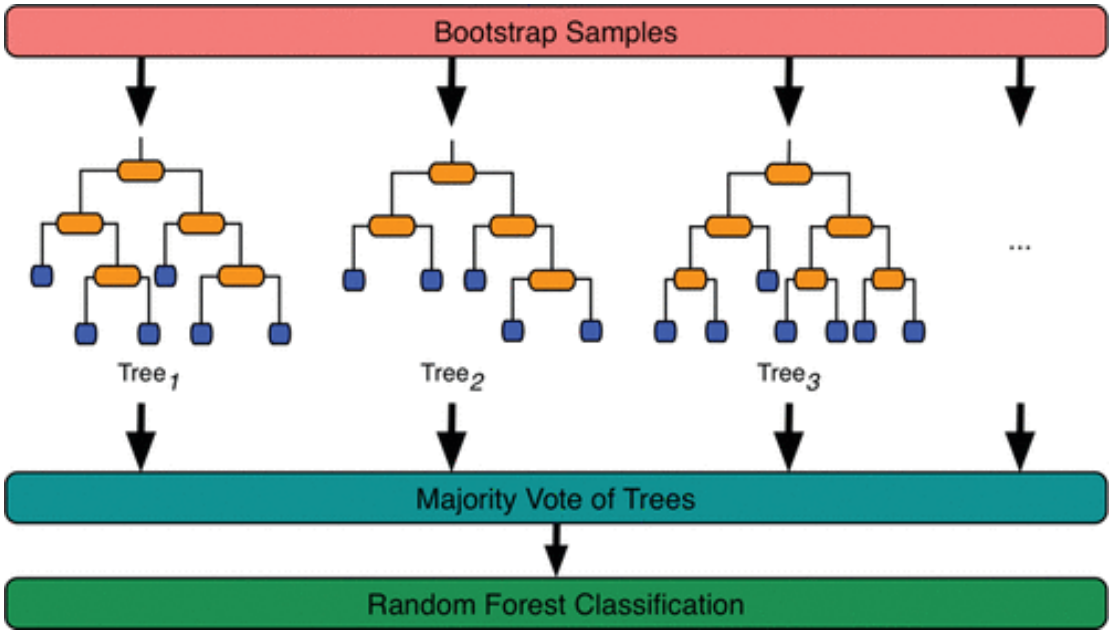

Figura 9. Adattata da [50].

I **principali vantaggi di Random Forest nei modelli QSAR** includono: elevata accuratezza predittiva rispetto ai modelli lineari, robustezza rispetto a dati rumorosi e feature non informative, capacità di gestire grandi insiemi di descrittori/fingerprint, e possibilità di interpretare l'importanza delle variabili per la bioattività. Questi aspetti rendono Random Forest uno degli algoritmi più utilizzati e validati per la predizione dell'attività biologica in drug discovery [51].

I principali iperparametri del modello Random Forest, come il numero di alberi e la profondità massima, sono stati impostati e ottimizzati empiricamente sulla base del training set, al fine di ottenere un compromesso tra prestazioni predittive e robustezza del modello.

Una volta definito l'algoritmo di classificazione e la procedura di addestramento, si è reso necessario stabilire criteri quantitativi per la valutazione delle prestazioni del modello predittivo.

La valutazione delle prestazioni di un modello di classificazione supervisionata rappresenta un passaggio fondamentale per stimarne l'affidabilità e la capacità di generalizzazione. In presenza di dataset sbilanciati, come nel caso in esame, l'utilizzo della sola accuracy può risultare fuorviante, poiché un modello potrebbe ottenere valori elevati semplicemente predicendo la classe maggioritaria.

Per questo motivo, nel presente studio le prestazioni del modello sono state valutate utilizzando metriche più informative, tra cui precision, recall e F1-score. La **precision** misura la proporzione di composti correttamente identificati come attivi rispetto al totale dei composti predetti come attivi, mentre la **recall** indica la capacità del modello di identificare correttamente i composti realmente attivi presenti nel dataset. Il **F1-score**, definito come la media armonica tra precision e recall, fornisce una misura complessiva dell'equilibrio tra le due metriche.

A supporto dell'analisi quantitativa, è stata inoltre utilizzata la **confusion matrix**, che consente di visualizzare il numero di veri positivi, veri negativi, falsi positivi e falsi negativi, offrendo una rappresentazione intuitiva delle prestazioni del classificatore.

L'impiego combinato di queste metriche permette una valutazione più robusta e affidabile del modello, particolarmente rilevante in ambito di drug discovery, dove l'identificazione dei composti attivi riveste un'importanza prioritaria.

3. RISULTATI

Il **dataset finale**, ottenuto dopo le fasi di raccolta e preprocessing descritte nel Capitolo 2, è composto da **414 molecole uniche**, ciascuna rappresentata tramite **fingerprint molecolari** (Morgan Fingerprint, raggio 2, 2048 bit) e associata a un valore sperimentale di **pKi** (–log10(Ki)). La conversione da Ki a pKi ha permesso di standardizzare i valori di affinità, facilitando l'analisi statistica e l'applicazione di algoritmi di machine learning.

A seguito della definizione di **soglie di classificazione** (pKi ≥ 7 per composti attivi, pKi ≤ 6 per composti inattivi), e dall'esclusione delle molecole con valori intermedi (6 < pKi < 7) il dataset usato per la classificazione è stato suddiviso in due classi:

260 composti inattivi (79.5%)

67 composti attivi (20.5%)

L'analisi della distribuzione delle classi (Figura 14) evidenzia un **marcato sbilanciamento**, con una netta prevalenza di composti inattivi. Questo squilibrio ha reso necessario l'utilizzo di **metriche di valutazione robuste** (precision, recall, F1-score, ROC-AUC) e tecniche di stratificazione durante la validazione, al fine di garantire una valutazione affidabile delle prestazioni del modello.

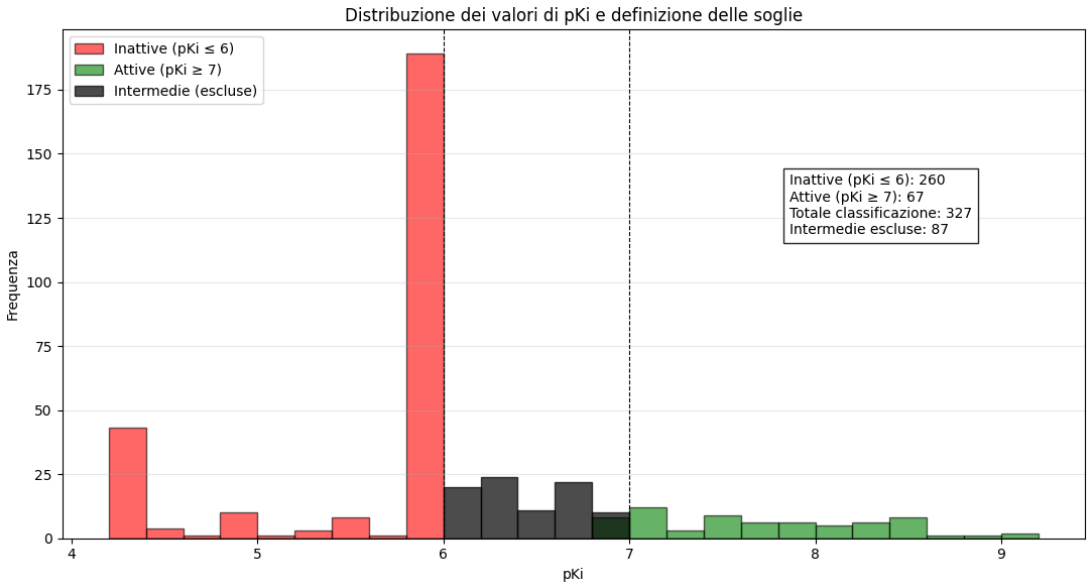

Figura 10. Distribuzione pKi per classe.

Il modello di classificazione basato su **Random Forest** (200 alberi, random_state=42) è stato addestrato utilizzando il **training set** (80% dei dati), mentre il **test set** (20%) è stato impiegato esclusivamente per la valutazione delle prestazioni predittive. Le metriche ottenute sul test set sono riportate di seguito.

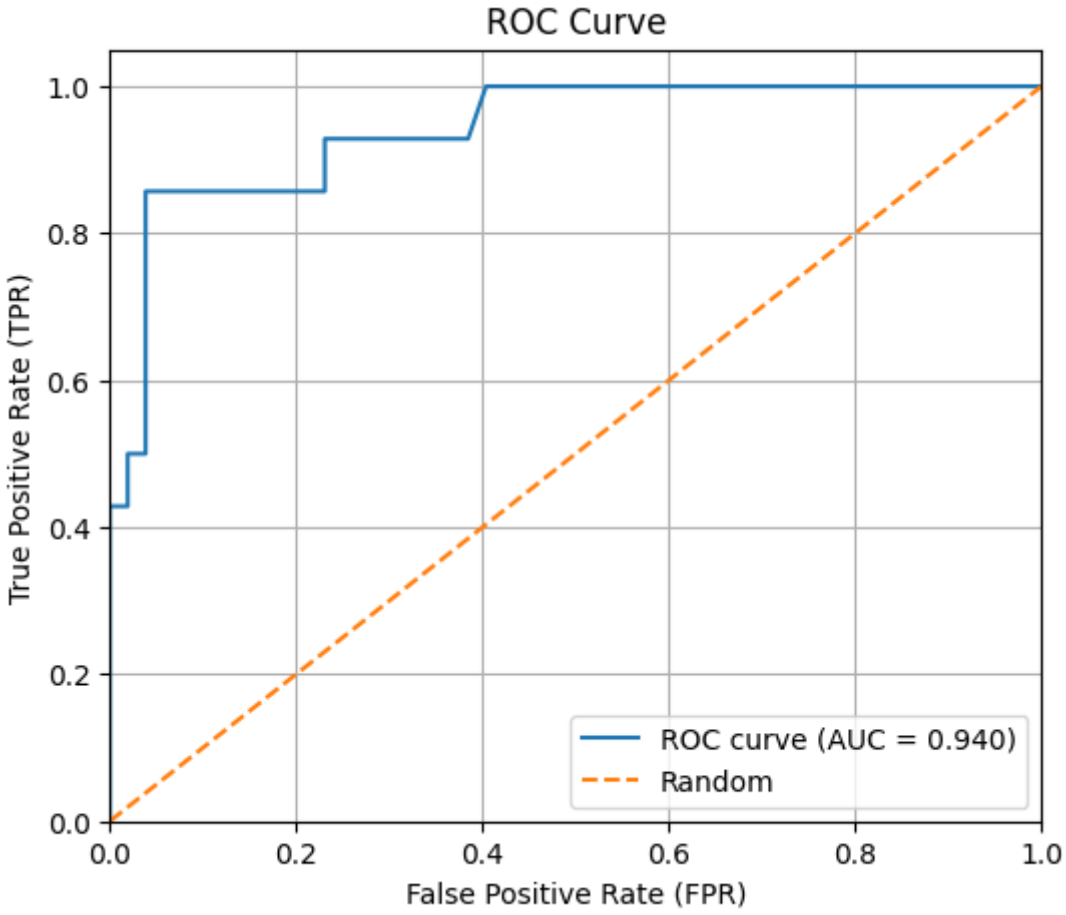

Figura 11. ROC-AUC

Questi risultati indicano che il modello è in grado di **discriminare efficacemente** tra composti attivi e inattivi, con una particolare capacità di identificare correttamente i composti inattivi (specificità del 96%). Tuttavia, il **recall della classe attiva** (79%) suggerisce che una piccola percentuale di composti attivi viene erroneamente classificata come inattiva, evidenziando la presenza di un numero limitato di falsi negativi.

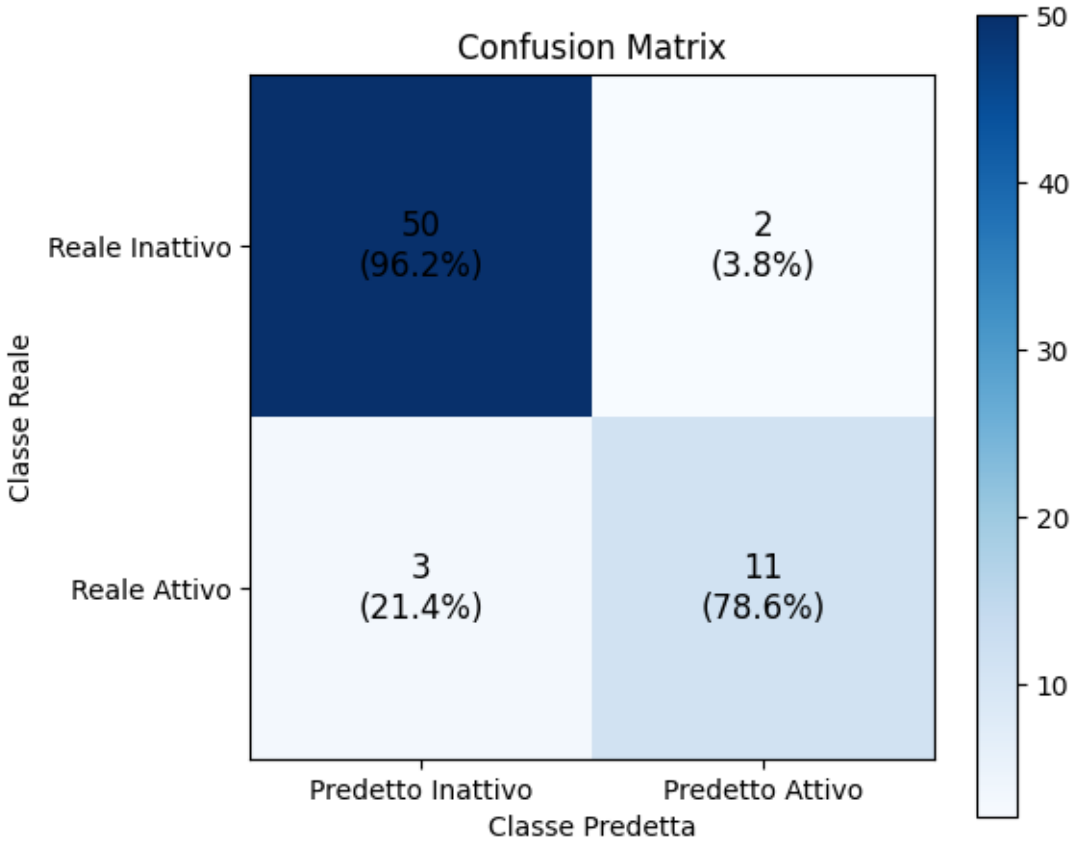

Figura 12. Confusion Matrix

Dopo aver valutato le performance del modello di classificazione, è stata analizzata anche la rilevanza delle caratteristiche molecolari nella fase predittiva.

Attraverso l'analisi dell'importanza delle feature del Random Forest, è stato possibile identificare i bit delle Morgan Fingerprint che hanno contribuito maggiormente alla discriminazione tra composti attivi e inattivi.

Alcune feature risultano significativamente più informative rispetto alle altre, suggerendo che specifici pattern strutturali siano maggiormente associati all'attività verso CDK9.

Sebbene i singoli bit delle fingerprint non siano direttamente interpretabili come specifici gruppi funzionali, tale analisi evidenzia che il modello non opera in maniera casuale, ma basa le proprie predizioni su informazioni strutturali coerenti e riproducibili.

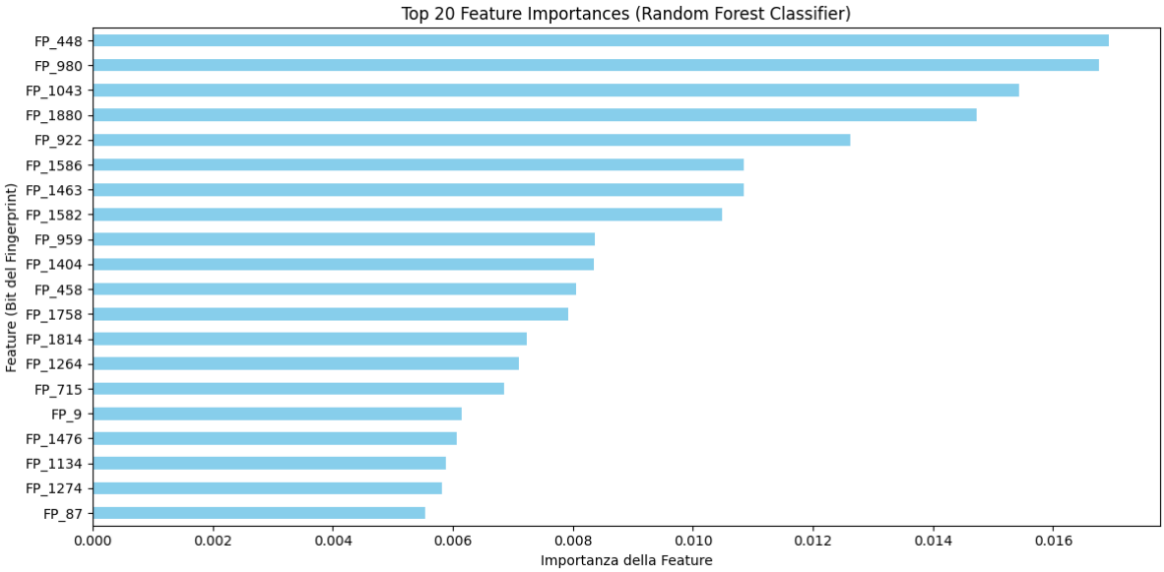

Figura 13. Analisi dell'importanza delle feature del Random Forest

Oltre alla classificazione, è stato sviluppato un modello di **regressione** (Random Forest Regressor, 200 alberi, random_state=42) per prevedere il valore esatto di **pKi**. Le prestazioni del modello sono state valutate tramite **validazione incrociata (5-fold CV)** e test set, ottenendo i seguenti risultati:

R^2 medio (5-fold CV): 0.45

RMSE (test set): 0.73

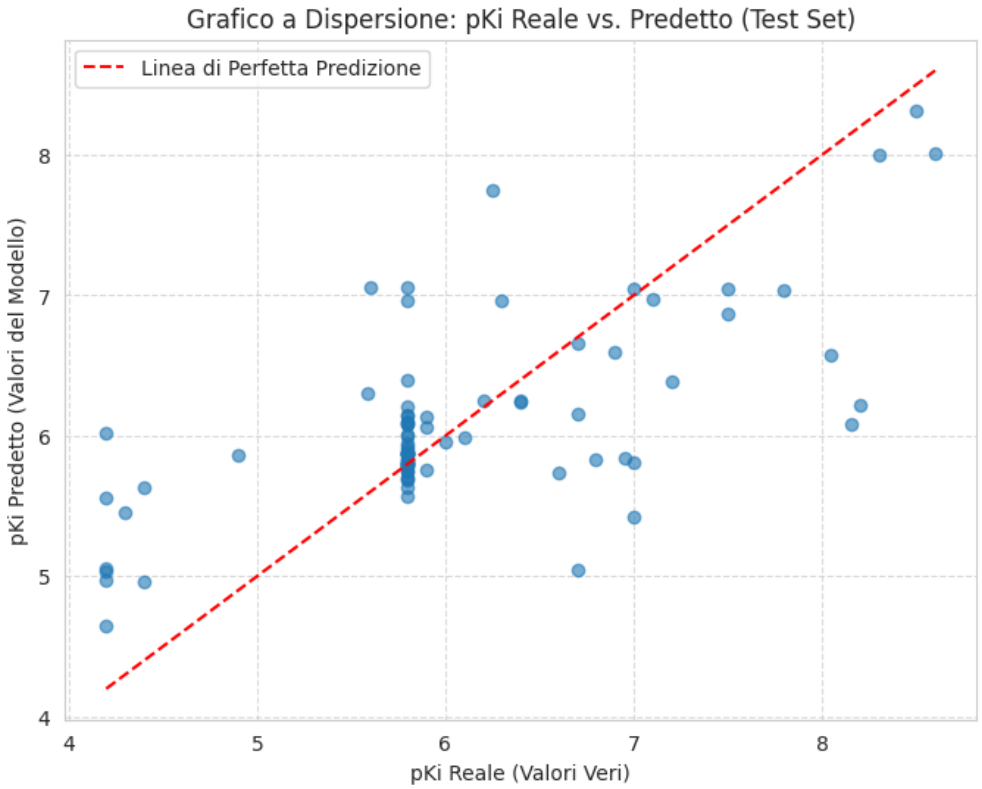

Figura 14. Grafico di dispersione tra i valori reali e predetti di pKi

Il grafico di dispersione tra i valori reali e predetti di pKi ha mostrato una **tendenza lineare**, confermando che il modello cattura una parte significativa della variabilità dei dati.

DISCUSSIONE

I risultati ottenuti nel presente studio dimostrano come l'approccio basato su Machine Learning applicato a fingerprint molecolari sia in grado di fornire un valido supporto nella predizione dell'attività biologica verso CDK9. In particolare, il modello di classificazione ha mostrato performance elevate, con un valore di ROC-AUC pari a 0,94, indicando un'ottima capacità discriminativa tra composti attivi e inattivi.

La migliore performance della classificazione rispetto alla regressione può essere attribuita a diversi fattori. In primo luogo, la natura binaria del problema riduce la complessità della predizione: distinguere tra attivo e inattivo rappresenta un compito meno oneroso rispetto alla stima quantitativa precisa del valore di pKi. La regressione, infatti, richiede di catturare variazioni sottili nell'affinità di legame, che possono dipendere da fattori strutturali complessi e non completamente rappresentati dalle sole Morgan Fingerprint.

Il valore di R^2 pari a circa 0,45 indica che il modello di regressione è in grado di spiegare una parte significativa, ma non esaustiva, della variabilità dei dati. Questo risultato è coerente con la complessità intrinseca delle interazioni proteina-ligando, che non dipendono esclusivamente dalla topologia molecolare bidimensionale, ma anche da aspetti tridimensionali, conformazionali e dinamici non inclusi nel modello.

Un elemento rilevante emerso dall'analisi è lo sbilanciamento delle classi, con una netta prevalenza di composti inattivi. Tale squilibrio può influenzare le performance del modello, favorendo una maggiore accuratezza nella predizione della classe dominante. Tuttavia, l'utilizzo di metriche come precision, recall e ROC-AUC, unitamente alla stratificazione durante la validazione, ha consentito di ottenere una valutazione più equilibrata delle prestazioni, evidenziando comunque una buona capacità di identificare anche i composti attivi.

L'analisi dell'importanza delle feature ha ulteriormente confermato la coerenza del modello, mostrando che specifici bit delle fingerprint contribuiscono in maniera più significativa alla classificazione. Sebbene tali bit non siano direttamente interpretabili come gruppi funzionali specifici, la loro rilevanza suggerisce l'esistenza di pattern strutturali ricorrenti associati all'attività verso CDK9. Questo aspetto rafforza l'ipotesi

che il modello abbia effettivamente appreso informazioni strutturali significative.

Nonostante i risultati incoraggianti, il presente studio presenta alcune limitazioni. Il numero relativamente contenuto di composti attivi e l'assenza di validazione su un dataset esterno possono limitare la generalizzabilità del modello. Inoltre, l'utilizzo esclusivo di descrittori topologici bidimensionali potrebbe non catturare completamente le caratteristiche molecolari determinanti per l'affinità di legame.

In prospettiva futura, l'integrazione di ulteriori descrittori molecolari, quali parametri fisico-chimici o rappresentazioni tridimensionali, potrebbe migliorare la capacità predittiva del modello di regressione. Ulteriori sviluppi potrebbero includere l'applicazione del modello a librerie esterne di composti o l'estensione dell'approccio ad altri target appartenenti alla famiglia delle CDK.

Nel complesso, i risultati ottenuti dimostrano come l'approccio basato su Machine Learning possa rappresentare uno strumento efficace per supportare le fasi iniziali della drug discovery, contribuendo alla selezione razionale di composti promettenti da sottoporre a validazione sperimentale.

CONCLUSIONI

Il presente lavoro ha dimostrato come l'applicazione di tecniche di Machine Learning alla chemoinformatica possa rappresentare un approccio efficace per supportare l'identificazione di nuovi potenziali inibitori di CDK9, un target di rilevante interesse nella ricerca oncologica.

Attraverso la costruzione di un dataset accuratamente preprocessato e la rappresentazione delle molecole mediante Morgan Fingerprint, è stato possibile sviluppare modelli predittivi in grado di correlare la struttura chimica con l'attività biologica. In particolare, il modello di classificazione ha evidenziato un'elevata capacità discriminativa, dimostrando l'efficacia dell'approccio nel distinguere composti attivi da inattivi. Il modello di regressione, pur mostrando una precisione moderata, ha confermato la presenza di una relazione strutturale significativa tra le caratteristiche molecolari e l'affinità di legame verso CDK9.

I risultati ottenuti sottolineano il valore strategico dell'integrazione tra strumenti computazionali e ricerca farmacologica. L'utilizzo di modelli predittivi consente infatti di orientare in modo razionale il processo di screening, riducendo tempi, costi e complessità della fase sperimentale e contribuendo a rendere il percorso di drug discovery più efficiente e sostenibile.

Sebbene siano presenti alcune limitazioni, legate principalmente alla dimensione del dataset e all'impiego esclusivo di descrittori bidimensionali, il lavoro svolto costituisce una base solida per ulteriori sviluppi. L'integrazione di rappresentazioni molecolari più avanzate e l'estensione del modello a nuovi target potrebbero ampliare ulteriormente le potenzialità dell'approccio proposto.

In conclusione, questo studio evidenzia come l'integrazione tra biotecnologie, bioinformatica e Machine Learning rappresenti non soltanto un supporto metodologico, ma uno strumento concreto e innovativo capace di contribuire allo sviluppo di nuove strategie terapeutiche.

BIBLIOGRAFIA

[1] Kaveh S, Mani-Varnosfaderani A, Neiband MS. Deriving general structure-activity/selectivity relationship patterns for different subfamilies of cyclin-dependent kinase inhibitors using machine learning methods. Sci Rep. 2024 Jul 3;14(1):15315. doi: 10.1038/s41598-024-66173-z. PMID: 38961127; PMCID: PMC11222421.

[2] Srivastava N, Saxena AK. Novel Small Molecule Inhibitors of Cyclin-dependent Kinases
as Anticancer Agents. Curr Med Chem. 2025;32(34):7512-7553. doi: 10.2174/0109298673331685241205180307. PMID: 40468937.

[3] Tutone M, Almerico AM. Recent advances on CDK inhibitors: An insight by means of in silico methods. Eur J Med Chem. 2017 Dec 15;142:300-315. doi: 10.1016/j.ejmech.2017.07.067. Epub 2017 Aug 4. PMID: 28802482.

[4] Gupta A, Dagar G, Chauhan R, Sadida HQ, Almarzooqi SK, Hashem S, Uddin S, Macha MA, Akil ASA, Pandita TK, Bhat AA, Singh M. Cyclin-dependent kinases in cancer: Role, regulation, and therapeutic targeting. Adv Protein Chem Struct Biol. 2023;135:21-55. doi: 10.1016/bs.apcsb.2023.02.001. Epub 2023 Mar 16. PMID: 37061333.

[5] Zabihi M, Lotfi R, Yousefi AM, Bashash D. Cyclins and cyclin-dependent kinases: from biology to tumorigenesis and therapeutic opportunities. J Cancer Res Clin Oncol. 2023 Apr;149(4):1585-1606. doi: 10.1007/s00432-022-04135-6. Epub 2022 Jul 4. PMID: 35781526; PMCID: PMC11796601.

[6] Malumbres M (2014) Cyclin-dependent kinases. Genome Biol 15(6):122. https://doi.org/10.1186/gb4184

[7] De Vivo M, Bottegoni G, Berteotti A, Recanatini M, Gervasio FL, Cavalli A (2011) Cyclin-dependent kinases: bridging their structure and function through computations. Future Med Chem 3(12):1551–1559. https://doi.org/10.4155/fmc.11.113

[8] Wood DJ, Endicott JA (2018) Structural insights into the functional diversity of the CDK-cyclin family. Open Biol 8(9):180112. https://doi.org/10.1098/rsob.180112

[9] Malumbres M, Harlow E, Hunt T, Hunter T, Lahti JM, Manning G et al (2009) Cyclin-dependent kinases: a family portrait. Nat Cell Biol 11(11):1275–1276. https://doi.org/10.1038/ncb11 09-1275

[10] Alrouji M, Alshammari MS, Anwar S, Venkatesan K, Shamsi A. Mechanistic Roles of Transcriptional Cyclin-Dependent Kinases in Oncogenesis: Implications for Cancer Therapy. Cancers (Basel). 2025 May 3;17(9):1554. doi: 10.3390/cancers17091554. PMID: 40361480; PMCID: PMC12071579.

[11] Fujinaga K, Huang F, Peterlin BM. P-TEFb: The master regulator of transcription elongation. Mol Cell. 2023 Feb 2;83(3):393-403. doi: 10.1016/j.molcel.2022.12.006. Epub 2023 Jan 3. PMID: 36599353; PMCID: PMC9898187.

[12] Puidebat O, Egloff S. The 7SK snRNP complex: a critical regulator in carcinogenesis. Biochimie. 2025 Nov;238(Pt A):3-8. doi: 10.1016/j.biochi.2025.05.003. Epub 2025 May 12. PMID: 40368082.

[13] Mandal R, Becker S, Strebhardt K. Targeting CDK9 for Anti-Cancer Therapeutics. Cancers (Basel). 2021 May 1;13(9):2181. doi: 10.3390/cancers13092181. PMID: 34062779; PMCID: PMC8124690.

[14] Graña X, De Luca A, Sang N, Fu Y, Claudio PP, Rosenblatt J, Morgan DO, Giordano A. PITALRE, a nuclear CDC2-related protein kinase that phosphorylates the retinoblastoma protein in vitro. Proc Natl Acad Sci U S A. 1994 Apr 26;91(9):3834-8. doi:

10.1073/pnas.91.9.3834. PMID: 8170997; PMCID: PMC43676.

[15] Morales F, Giordano A. Overview of CDK9 as a target in cancer research. Cell Cycle. 2016;15(4):519-27. doi: 10.1080/15384101.2016.1138186. PMID: 26766294; PMCID: PMC5056610.

[16] Ramakrishnan R, Yu W, Rice AP. Limited redundancy in genes regulated by Cyclin T2 and Cyclin T1. BMC Res Notes. 2011 Jul 26;4:260. doi: 10.1186/1756-0500-4-260. PMID: 21791050; PMCID: PMC3160394.

[17] Cassandri M, Fioravanti R, Pomella S, Valente S, Rotili D, Del Baldo G, De Angelis B, Rota R, Mai A. CDK9 as a Valuable Target in Cancer: From Natural Compounds Inhibitors to Current Treatment in Pediatric Soft Tissue Sarcomas. Front Pharmacol. 2020 Aug 13;11:1230. doi: 10.3389/fphar.2020.01230. PMID: 32903585; PMCID: PMC7438590.

[18] Bohacek RS, McMartin C, Guida WC. The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev. 1996 Jan;16(1):3-50. doi: 10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6. PMID: 8788213.

[19] Gorgulla C, Jayaraj A, Fackeldey K, Arthanari H. Emerging frontiers in virtual drug discovery: From quantum mechanical methods to deep learning approaches. Curr Opin Chem Biol. 2022 Aug;69:102156. doi: 10.1016/j.cbpa.2022.102156. Epub 2022 May 13. PMID: 35576813; PMCID: PMC9990419.

[20] Gorgulla C, Fackeldey K, Wagner G, Arthanari H. Accounting of Receptor Flexibility in Ultra-Large Virtual Screens with VirtualFlow Using a Grey Wolf Optimization Method. Supercomput Front Innov. 2020 Nov 7;7(3):4-12. doi: 10.14529/jsfi200301. PMID: 34693068; PMCID: PMC8530406.

[21] Gorgulla C, Çınaroğlu SS, Fischer PD, Fackeldey K, Wagner G, Arthanari H. VirtualFlow Ants-Ultra-Large Virtual Screenings with Artificial Intelligence Driven Docking Algorithm Based on Ant Colony Optimization. Int J Mol Sci. 2021 May 28;22(11):5807. doi: 10.3390/ijms22115807. PMID: 34071676; PMCID: PMC8199267.

[22] Brooijmans N, Kuntz ID. Molecular recognition and docking algorithms. Annu Rev Biophys Biomol Struct. 2003;32:335-73. doi: 10.1146/annurev.biophys.32.110601.142532. Epub 2003 Jan 28. PMID: 12574069.

[23] Patne AY, Dhulipala SM, Lawless W, Prakash S, Mohapatra SS, Mohapatra S. Drug Discovery in the Age of Artificial Intelligence: Transformative Target-Based Approaches. Int J Mol Sci. 2024 Nov 14;25(22):12233. doi: 10.3390/ijms252212233. PMID: 39596300; PMCID: PMC11594879.

[24] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, Madabhushi A, Shah P, Spitzer M, Zhao S. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019 Jun;18(6):463-477. doi: 10.1038/s41573-019-0024-5. PMID: 30976107; PMCID: PMC6552674.

[25] Koutroumpa NM, Papavasileiou KD, Papadiamantis AG, Melagraki G, Afantitis A. A Systematic Review of Deep Learning Methodologies Used in the Drug Discovery Process with Emphasis on In Vivo Validation. Int J Mol Sci. 2023 Mar 31;24(7):6573. doi: 10.3390/ijms24076573. PMID: 37047543; PMCID: PMC10095548.

[26] Sumathi S, Suganya K, Swathi K, Sudha B, Poornima A, Varghese CA, Aswathy R. A Review on Deep Learning-driven Drug Discovery: Strategies, Tools and Applications. Curr Pharm Des. 2023 May 19;29(13):1013-1025. doi: 10.2174/1381612829666230412084137. PMID: 37055908.

[27] Patel L, Shukla T, Huang X, Ussery DW, Wang S. Machine Learning Methods in Drug Discovery. Molecules. 2020 Nov 12;25(22):5277. doi: 10.3390/molecules25225277. PMID: 33198233; PMCID: PMC7696134.

[28] Ekins S, Puhl AC, Zorn KM, Lane TR, Russo DP, Klein JJ, Hickey AJ, Clark AM. Exploiting machine learning for end-to-end drug discovery and development. Nat Mater. 2019 May;18(5):435-441. doi: 10.1038/s41563-019-0338-z. Epub 2019 Apr 18. PMID: 31000803; PMCID: PMC6594828.

[29] Zdrazil B, Felix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, de Veij M, Ioannidis H, Lopez DM, Mosquera JF, Magarinos MP, Bosc N, Arcila R, Kizilören T, Gaulton A, Bento AP, Adasme MF, Monecke P, Landrum GA, Leach AR. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res. 2024 Jan 5;52(D1):D1180-D1192. doi: 10.1093/nar/gkad1004. PMID: 37933841; PMCID: PMC10767899.

[30] Hunter FMI, Ioannidis H, Bento AP, Bosc N, Corbett S, Felix E, Magarinos MP, Manners E, Smit IA, de Veij M, O'Boyle NM, Zdrazil B, Leach AR. Drug and Clinical Candidate Drug Data in ChEMBL. J Med Chem. 2025 Oct 9;68(19):19800-19827. doi: 10.1021/acs.jmedchem.5c00920. Epub 2025 Sep 19. PMID: 40971497; PMCID: PMC12516679.

[31] Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F, Bellis L, Overington JP. ChEMBL web services: streamlining access to drug discovery data and utilities. Nucleic Acids Res. 2015 Jul 1;43(W1):W612-20. doi: 10.1093/nar/gkv352. Epub 2015 Apr 16. PMID: 25883136; PMCID: PMC4489243.

[32] Nowotka MM, Gaulton A, Mendez D, Bento AP, Hersey A, Leach A. Using ChEMBL web services for building applications and data processing workflows relevant to drug discovery. Expert Opin Drug Discov. 2017 Aug;12(8):757-767. doi: 10.1080/17460441.2017.1339032. Epub 2017 Jun 12. PMID: 28602100; PMCID: PMC6321761.

[33] Xia LY, Wang YW, Meng DY, Yao XJ, Chai H, Liang Y. Descriptor Selection via Log-Sum Regularization for the Biological Activities of Chemical Structure. Int J Mol Sci. 2017 Dec 22;19(1):30. doi: 10.3390/ijms19010030. PMID: 29271922; PMCID: PMC5795980.

[34] Gadaleta D. Automated Workflows for Data Curation and Machine Learning to Develop Quantitative Structure-Activity Relationships. Methods Mol Biol. 2025;2834:115-130. doi: 10.1007/978-1-0716-4003-6_5. PMID: 39312162.

[35] Gini GC. QSAR: Using the Past to Study the Present. Methods Mol Biol. 2025;2834:3-39. doi: 10.1007/978-1-0716-4003-6_1. PMID: 39312158.

[36] Danishuddin, Khan AU. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. Drug Discov Today. 2016 Aug;21(8):1291-302. doi: 10.1016/j.drudis.2016.06.013. Epub 2016 Jun 18. PMID: 27326911.

[37] Bertoni M, Duran-Frigola M, Badia-I-Mompel P, Pauls E, Orozco-Ruiz M, Guitart-Pla O, Alcalde V, Diaz VM, Berenguer-Llergo A, Brun-Heath I, Villegas N, de Herreros AG, Aloy P. Bioactivity descriptors for uncharacterized chemical compounds. Nat Commun. 2021 Jun 24;12(1):3932. doi: 10.1038/s41467-021-24150-4. PMID: 34168145; PMCID: PMC8225676.

[38] Fernández-Torras A, Comajuncosa-Creus A, Duran-Frigola M, Aloy P. Connecting chemistry and biology through molecular descriptors. Curr Opin Chem Biol. 2022 Feb;66:102090. doi: 10.1016/j.cbpa.2021.09.001. Epub 2021 Oct 6. PMID: 34626922.

[39] Danishuddin, Khan AU. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. Drug Discov Today. 2016 Aug;21(8):1291-302. doi: 10.1016/j.drudis.2016.06.013. Epub 2016 Jun 18. PMID: 27326911.

[40] Koutsoukas A, Paricharak S, Galloway WR, Spring DR, Ijzerman AP, Glen RC, Marcus D, Bender A. How diverse are diversity assessment methods? A comparative analysis and benchmarking of molecular descriptor space. J Chem Inf Model. 2014 Jan 27;54(1):230-42. doi: 10.1021/ci400469u. Epub 2013 Dec 13. PMID: 24289493.

[41] Bertoni M, Duran-Frigola M, Badia-I-Mompel P, Pauls E, Orozco-Ruiz M, Guitart-Pla O, Alcalde V, Diaz VM, Berenguer-Llergo A, Brun-Heath I, Villegas N, de Herreros AG, Aloy P. Bioactivity descriptors for uncharacterized chemical compounds. Nat Commun. 2021 Jun 24;12(1):3932. doi: 10.1038/s41467-021-24150-4. PMID: 34168145; PMCID: PMC8225676.

[42] Yang J, Cai Y, Zhao K, Xie H, Chen X. Concepts and applications of chemical fingerprint for hit and lead screening. Drug Discov Today. 2022 Nov;27(11):103356. doi: 10.1016/j.drudis.2022.103356. Epub 2022 Sep 13. PMID: 36113834.

[43] Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010 May 24;50(5):742-54. doi: 10.1021/ci100050t. PMID: 20426451.

[44] Myint KZ, Wang L, Tong Q, Xie XQ. Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions. Mol Pharm. 2012 Oct 1;9(10):2912-23. doi: 10.1021/mp300237z. Epub 2012 Aug 31. PMID: 22937990; PMCID: PMC3462244.

[45] Zhong S, Guan X. Count-Based Morgan Fingerprint: A More Efficient and Interpretable Molecular Representation in Developing Machine Learning-Based Predictive Regression Models for Water Contaminants' Activities and Properties. Environ Sci Technol. 2023 Nov 21;57(46):18193-18202. doi: 10.1021/acs.est.3c02198. Epub 2023 Jul 5. PMID: 37406199.

[46] Lane TR, Foil DH, Minerali E, Urbina F, Zorn KM, Ekins S. Bioactivity Comparison across Multiple Machine Learning Algorithms Using over 5000 Datasets for Drug Discovery. Mol Pharm. 2021 Jan 4;18(1):403-415. doi:

10.1021/acs.molpharmaceut.0c01013. Epub 2020 Dec 16. PMID: 33325717; PMCID: PMC8237624.

[47] Chuang KV, Gunsalus LM, Keiser MJ. Learning Molecular Representations for Medicinal Chemistry. J Med Chem. 2020 Aug 27;63(16):8705-8722. doi: 10.1021/acs.jmedchem.0c00385. Epub 2020 May 15. PMID: 32366098.

[48] Li Y, Ngom A. Sparse representation approaches for the classification of high-dimensional biological data. BMC Syst Biol. 2013;7 Suppl 4(Suppl 4):S6. doi: 10.1186/1752-0509-7-S4-S6. Epub 2013 Oct 23. PMID: 24565287; PMCID: PMC3854665.

[49] Pantic IV, Paunovic Pantic J, Valjarevic S, Corridon PR, Topalovic N. Artificial intelligence - based approaches based on random forest algorithm for signal analysis: Potential applications in detection of chemico - biological interactions. Chem Biol Interact. 2025 Sep 5;418:111624. doi: 10.1016/j.cbi.2025.111624. Epub 2025 Jun 27. PMID: 40582691.

[50] Bruce CL, Melville JL, Pickett SD, Hirst JD. Contemporary QSAR classifiers compared. J Chem Inf Model. 2007 Jan-Feb;47(1):219-27. doi: 10.1021/ci600332j. PMID: 17238267.

[51] Marchese Robinson RL, Palczewska A, Palczewski J, Kidley N. Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Data Sets. J Chem Inf Model. 2017 Aug 28;57(8):1773-1792. doi: 10.1021/acs.jcim.6b00753. Epub 2017 Aug 2. PMID: 28715209.