



INTELLIGENZA ARTIFICIALE E MEDICINA DI PRECISIONE

L'intelligenza artificiale rivoluziona la scoperta di farmaci rendendo accessibile uno spazio chimico infinito. Il contributo descrive PyRMD, tool automatizzato di screening virtuale basato sull'algoritmo Random Matrix Discriminant. Si illustrano l'integrazione con il docking (PyRMD2Dock), la suite PyRMD Studio e la scoperta di un modulatore per il recettore GPR183.

La storia del machine learning (ML) ha le sue radici negli anni Cinquanta, quando pionieri come Alan Turing e Arthur Samuel iniziarono a immaginare macchine in grado di apprendere dai dati. Nei decenni successivi, lo sviluppo di algoritmi statistici e delle prime reti neurali, tra cui il *perceptron* di Frank Rosenblatt, pose le basi teoriche di un campo che sarebbe rimasto per lungo tempo limitato, soprattutto nelle sue applicazioni in ambito biologico, dalla scarsità di dati e di potenza computazionale. Ai suoi albori, il ML veniva usato principalmente per migliorare l'accuratezza dei metodi statistici per l'interpretazione di dati con comportamenti non lineari [1]. La svolta arrivò tra il 2000 e il 2010, con l'esplosione dei big data grazie a Internet, il potenziamento della componente hardware e il miglioramento degli algoritmi, guidati da ricercatori come Geoffrey Hinton. Tecniche basate su reti neurali profonde iniziarono a superare i metodi tradizionali in compiti complessi, come il riconoscimento di immagini e la traduzione automatica, ampliando le applicazioni nelle scienze della vita [2]. Il campo della biologia, infatti, presenta forti componenti di non linearità dovute alla complessità dei sistemi biologici ed è quindi uno degli ambiti in cui l'impatto del ML può essere particolarmente utile per identificare simmetrie e pattern nei dati biologici. Una svolta parallela, cruciale per le scienze della vita, arrivò con il Progetto Genoma Umano, completato nei primi anni Duemila,

che rese disponibile la prima sequenza di riferimento del genoma umano [3, 4]. Da quel momento, lo sviluppo delle tecnologie di sequenziamento di nuova generazione (NGS) ha portato a una crescita esponenziale dei dati biologici: esperimenti di genomica, RNA-seq ed epigenomica producono oggi grandi moli di informazioni, una parte significativa delle quali è accessibile attraverso database pubblici [5] (Fig. 1). In questa crescita, un tassello fondamentale è stato l'impiego delle schede grafi-

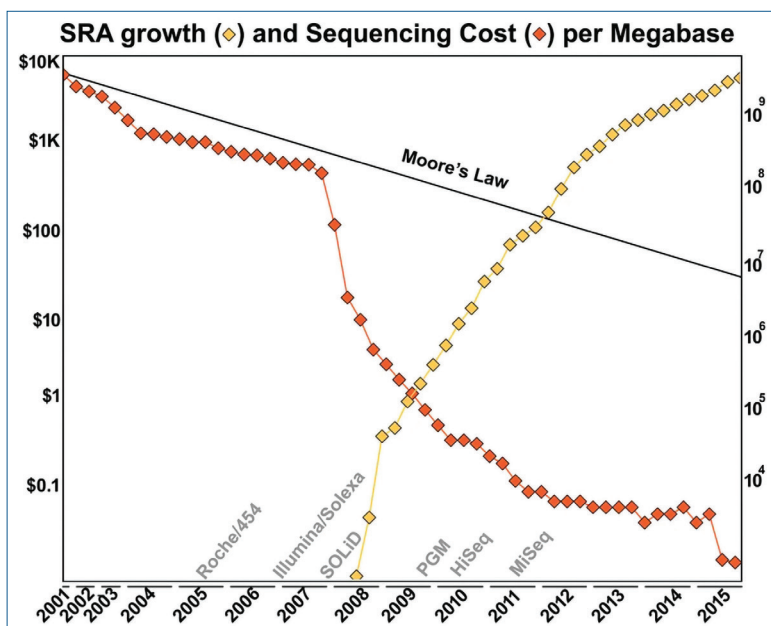


Fig. 1 - Costo (in scala logaritmica) per megabase di sequenziamento del DNA nel corso del tempo mostra che il prezzo del sequenziamento è diminuito drasticamente dopo l'affermazione delle tecnologie NGS, portando a un aumento del numero di megabase sequenziati (in scala logaritmica) nel database Sequence Read Archive (SRA). Nella figura sono mostrate anche le principali tecnologie di sequenziamento.

Fonte: <https://doi.org/10.1093/bioinformatics/btv584>



Fig. 2 - La rilevanza dell'AI in biologia è evidenziata dal risalto e dalla frequenza con cui modelli come AlphaFold, AlphaMissense, Evo e AlphaGenome hanno conquistato le copertine di *Nature* e di *Science* negli ultimi 5 anni

che (GPU), che ha reso possibile processare enormi quantità di dati in parallelo.

Nell'ambito biomedico, il ML è oggi uno strumento fondamentale: viene utilizzato per analizzare immagini diagnostiche, prevedere la risposta ai farmaci e interpretare dati genomici su larga scala. Il successo di questi approcci è stato consacrato dai premi Nobel del 2024. In Fisica, il riconoscimento è andato a John Hopfield e Geoffrey Hinton per le scoperte e le invenzioni che hanno reso possibile il ML basato su reti neurali artificiali. In Chimica, il premio è stato assegnato a David Baker per il design computazionale di proteine tramite Rosetta [6] e a Demis Hassabis e John Jumper per la predizione della struttura proteica mediante AlphaFold [7]. Modelli come Rosetta e AlphaFold2 hanno dimostrato che il deep learning è in grado di affrontare e risolvere problemi biologici storicamente complessi, aprendo la strada a una nuova generazione di strumenti computazionali per le scienze della vita. In particolare, nella medicina di precisione, questi modelli consentono di collegare le varianti genetiche ai fenotipi clinici, migliorando la diagnosi e i trattamenti personalizzati. Oggi, il ML non è più soltanto un supporto alla ricerca, ma un elemento strutturale della biomedicina, capace di trasformare il modo in cui comprendiamo e curiamo le malattie (Fig. 2).

I primi passi del ML nel campo della scoperta di nuovi farmaci (Drug Discovery), risalgono agli anni Sessanta e Settanta, con approcci come il QSAR (Quantitative Structure-Activity Relationship), introdotto da Corwin Hansch [8], che utilizzava mo-

delli di regressione lineare per correlare la struttura chimica di una molecola alla sua attività biologica. Oggi, software come DeepChem, Chemprop, Metasite e Schrödinger Suite integrano ML e chimica computazionale per predire proprietà molecolari e ottimizzare candidati farmaci. Un ruolo più centrale è svolto dai modelli basati su Graph Neural Networks (GNN), in cui le molecole vengono rappresentate come grafi: nodi (atomi) e archi (legami). Questo consente di preservare naturalmente la struttura chimica e di catturare relazioni sia locali sia globali. Modelli come MolGAN [9] o D-MPNN [10] consentono sia la predizione di proprietà chimiche sia la generazione di nuove molecole. In modo intuitivo, i GNN funzionano facendo "aggregare" informazioni tra atomi connessi (*message passing*), aggiornando iterativamente le rappresentazioni dei nodi: questo consente al modello di apprendere pattern chimici complessi, come effetti elettronici o sterici, difficili da catturare con descrittori tradizionali. Accanto a questi, la frontiera più avanzata della ricerca è oggi rappresentata dai modelli di diffusione, come GeoDiff [11] o MiDi [12], che permettono di generare strutture molecolari tridimensionali in modo efficiente, catturando vincoli geometrici e proprietà strutturali con elevata precisione.

Negli ultimissimi anni, la rivoluzione dei dati nelle scienze della vita sta trovando alleati importanti nei programmi di intelligenza artificiale (AI). Tra questi ci sono quelli sviluppati da Google e, in particolare, da Google DeepMind. Modelli come AlphaFold e AlphaMissense [13] hanno dimostrato che il deep learning può risolvere problemi biologici storica-

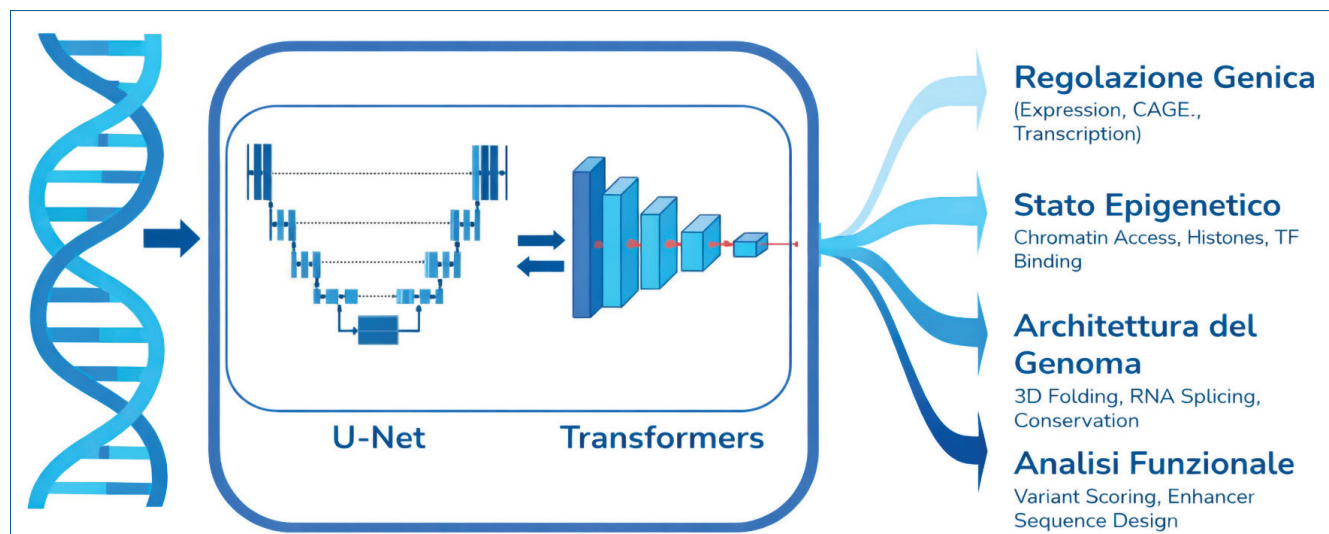


Fig. 3 - AlphaGenome utilizza architetture U-Net e Transformers per elaborare dati biologici reali, convertendo intervalli di sequenze genomiche in mappe dettagliate di informazioni molecolari

mente complessi, aprendo la strada a una nuova generazione di strumenti progettati per operare direttamente sui dati genomici. Accanto alle iniziative di Google DeepMind, l'applicazione del ML alla genomica e alla biologia molecolare è oggi un campo altamente competitivo e diffuso. Aziende e gruppi accademici stanno sviluppando modelli analoghi per la predizione funzionale del genoma: tra questi, Enformer (originariamente sviluppato da DeepMind ed ampiamente utilizzato nella comunità), DeepSEA e Basenji, insieme a piattaforme come Evo2 [14] e a modelli sviluppati da realtà come Insilico Medicine e Recursion Pharmaceuticals. Questi strumenti condividono un obiettivo comune: prevedere l'effetto delle varianti genetiche, modellare l'espressione genica e identificare regioni regolatorie funzionali a partire dalla sola sequenza del DNA.

Nell'ambito della medicina di precisione e, in particolare, del sequenziamento genomico, il ML sta aiutando a risolvere alcuni dei problemi più critici legati alle malattie genetiche: distinguere le varianti genetiche con impatto funzionale o clinico (mutazioni) da quelle neutre. In questo contesto si inserisce AlphaGenome [15] (Fig. 3), un modello concepito per analizzare il DNA su larga scala e collegare le eventuali mutazioni presenti nelle sequenze genomiche al cambiamento della loro funzione biologica. AlphaGenome apprende rappresentazioni globali del genoma, integrando informazioni sulla

regolazione e sull'espressione genica e sul contesto epigenetico. L'obiettivo è prevedere l'effetto funzionale delle varianti genetiche, distinguendo le mutazioni rilevanti dal rumore di fondo e contribuendo a chiarire il legame tra genotipo e fenotipo. In ambito di medicina di precisione, questo significa poter interpretare più rapidamente i dati di sequenziamento, identificare possibili varianti patogeniche, in particolare in regioni non codificanti proteine, e, in futuro, sostenere lo sviluppo di terapie mirate.

In questo panorama, AlphaGenome si distingue per un'impostazione più integrata e orientata all'uso pratico. Più che un singolo modello "chiuso", AlphaGenome è concepito come un'API che consente ai ricercatori di interrogare modelli di deep learning addestrati su grandi dataset genomici. Questo approccio consente di integrare direttamente capacità predittive avanzate all'interno di pipeline bioinformatiche già esistenti.

Dal punto di vista più tecnico, AlphaGenome utilizza architetture neurali profonde (tipicamente basate su transformer o su modelli convoluzionali per sequenze) per apprendere rappresentazioni del DNA su larga scala. Per quanto riguarda la fase di learning, in input riceve sequenze genomiche e restituisce una varietà di predizioni funzionali. Nella fase di predizione, AlphaGenome richiede come input una o più mutazioni e i diversi risultati ottenibili possono essere di vario tipo. A scopo esempli-

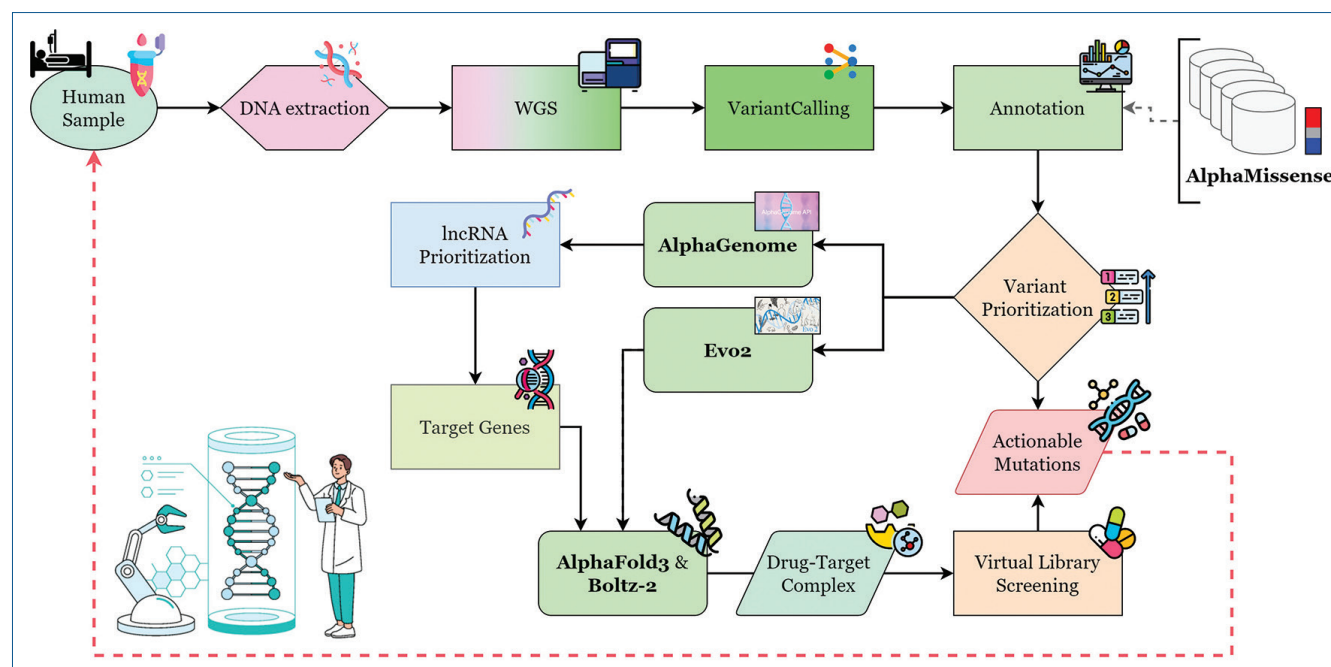


Fig. 4 - Esempio di come l'AI supporta il flusso della ricerca in campo biomedico

ficativo, vengono riportate alcune operazioni che possono essere condotte sia sul genoma umano sia su quello murino:

- Espressione Genica: prevede i livelli di produzione di mRNA e l'attività dei siti di inizio della trascrizione (dati RNA-seq, CAGE e PRO-cap);
- Modelli di Splicing: forniscono predizioni ad alta risoluzione sui siti di splicing, sull'efficienza del loro utilizzo e sulla forza delle giunzioni tra esoni;
- Stato della Cromatina: valuta quanto il DNA sia "accessibile" o compatto, predicendo segnali di DNase-seq, ATAC-seq e le modificazioni istoniche (i "tag" chimici che regolano l'impacchettamento del DNA);
- Legame Regolatorio: identifica i siti di legame dei fattori di trascrizione (TF), le proteine che agiscono come interruttori molecolari per i geni;
- Struttura 3D: predice le mappe di contatto della cromatina (come Micro-C e Hi-C), ovvero come il filamento di DNA si ripiega nello spazio tridimensionale per far interagire elementi distanti tra loro.

Un'altra pietra miliare nel panorama dell'AI applicata alla biologia è stata la pubblicazione su *Nature* del modello di linguaggio genomico (gLM) Evo2, avvenuta nel marzo 2026. Evo2 è un modello fondazionale per la biologia, cioè un modello progettato per apprendere regolarità e schemi generali

presenti nel codice genetico. È stato co-sviluppato da un team che comprende ricercatori dell'Arc Institute, della Stanford University e di NVIDIA, con il contributo dell'Università della California. Allenato su 128 mila genomi provenienti dai diversi domini della vita, per un totale di oltre 9,3 trilioni di nucleotidi, ad oggi rappresenta il più grande modello di AI nell'ambito della biologia con una complessità paragonabile a quella dei moderni Large Language Model (LLM). Questo modello ha dimostrato una sorprendente capacità di apprendere caratteristiche quali confini fra le regioni introniche ed esoniche, siti di legame per fattori di trascrizione ed elementi strutturali delle proteine, senza uno specifico allenamento (*Zero shot performance*). Grazie alle caratteristiche apprese, si è dimostrato in grado di generare genomi sintetici e di predire l'effetto di mutazioni nel genoma umano, sia in regioni codificanti sia in non codificanti. Evo2 rappresenta dunque un efficace strumento che può guidare il ricercatore nella selezione delle varianti di significato non noto (VUS) sulle quali effettuare approfondimenti, come mostrato da Spirito *et al.* [16].

Conclusioni

La rivoluzione dell'intelligenza artificiale [17] sta trasformando anche la pratica clinica e la ricerca



biomedica (Fig. 4). L'utilizzo dell'AI abbraccia l'intero flusso di analisi dei dati omici: dalla decodifica e dall'interpretazione dei segnali grezzi generati dagli strumenti di laboratorio [18, 19], all'identificazione delle varianti [20] e alla predizione dei loro effetti biologici, sia a livello fisiologico sia patologico. A questi passaggi si affiancano applicazioni emergenti e ad alto impatto, come la costruzione di gemelli digitali (*human digital twin*) per simulare l'evoluzione della malattia e la risposta ai trattamenti [21], e l'uso di modelli computazionali per scoprire o riposizionare farmaci, individuando target terapeutici promettenti e ottimizzando i percorsi di sviluppo [22, 23]. Questa rivoluzione poggia su due pilastri fondamentali: da un lato l'accessibilità e la standardizzazione di una crescente mole di dati su cui allenare i modelli; dall'altra la necessità di infrastrutture di calcolo ad alte prestazioni su cui eseguire l'allenamento dei modelli. Entrambi presuppongono la collaborazione e il coordinamento tra attori pubblici e privati in un quadro normativo che favorisca la condivisione responsabile delle informazioni, garantendo al contempo la tutela della privacy e dei diritti degli individui.

Oggi è possibile immaginare un flusso di analisi guidato dai dati omici del paziente (*patient data-driven precision medicine*) in grado di identificare specifiche caratteristiche genetiche e cliniche, così da indirizzare la terapia in modo mirato e, potenzialmente, disegnare un farmaco personalizzato in base alle mutazioni presenti nel singolo individuo. Questo è reso possibile anche grazie al supporto che l'AI offre nel velocizzare l'analisi della grande mole di dati (identificazione di varianti, target terapeutici e screening farmacologico) che i moderni strumenti mettono a disposizione, trasformandoli in informazioni utili per i clinici e i ricercatori.

BIBLIOGRAFIA

- [1] W.G. Baxt, J. Skora, *The Lancet*, 1996, **347**(8993), 12-5.
- [2] E.J. Topol, *Nature Medicine*, 2019, **25**(1), 44.
- [3] E.S. Lander *et al.*, *Nature*, 2001, Feb. 15, **409**(6822), 860.
- [4] J.C. Venter *et al.* *Science*, 2001, **291**(5523), 1304.
- [5] Z.D. Stephens *et al.*, *PLoS biology*, 2015, **13**(7), e1002195.
- [6] R. Das, D. Baker, *Annu. Rev. Biochem.*, 2008, **77**, 363.
- [7] J. Jumper *et al.*, *Nature*, 2021, **596**(7873), 583.
- [8] C. Hansch, T. Fujita, *Journal of the American Chemical Society*, 1964, **86**(8), 1616.
- [9] N. De Cao, T. Kipf, ICML 2018 workshop on Theoretical Foundations and Applications of Deep Generative Models. 2018.
- [10] K. Yang *et al.*, *Journal of Chemical Information and Modeling*, 2019, **59**(8), 3370.
- [11] X. Minkai, 2022, <https://doi.org/10.48550/arXiv.2203.02923>
- [12] C. Vignac, 2023, <https://doi.org/10.48550/arXiv.2302.09048>
- [13] J. Cheng *et al.*, *Science*, 2023, **380**(6651), 1230.
- [14] G. Brixi *et al.*, *Nature*, 2026, **652**, 1349.
- [15] Ž. Avsec *et al.*, *Nature*, 2026, **18**, 1196.
- [16] G. Spirito *et al.*, *npj Genom. Med.*, 2026, **11**, 11.
- [17] S. Makridakis, *Futures*, 2017, **90**, 46.
- [18] R. Wick *et al.*, *Genome biology*, 2019, **20**, 1.
- [19] M.U. Ahsan *et al.*, *Nature Comm.*, 2024, **15**, 1234.
- [20] R. Poplin *et al.*, *Nature biotechnology*, 2018, **36**, 983.
- [21] A. Vallée, *Frontiers in Digital Health*, 2023, **5**, 123456.
- [22] S. Pushpakom *et al.*, *Nature Reviews Drug discovery*, 2019, **18**, 41.
- [23] K. Park, *Translational and clinical pharmacology*, 2019, **27**(2), 59.

Artificial Intelligence and Precision Medicine

Artificial intelligence has a significant impact on precision medicine thanks to its ability to analyze large volumes of genomic and clinical data. Neural networks and deep learning algorithms are now helping to improve diagnosis, prediction, and the development of personalized therapies, opening new perspectives for basic science and translational medicine programs.