Andrea Zaliani[a,b]
http://dx.medra.org/10.17374/CI.2023.105.4.41
[a]Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP), Hamburg (Germany)
[b]Fraunhofer Cluster of Excellence for Immune-Mediated Diseases (CIMD) Frankfurt (Germany)
Andrea.Zaliani@itmp.fraunhofer.de

BANCHE DATI ED AI IN DRUG DISCOVERY

L'uso di database pubblici e di algoritmi di AI/ML nella scoperta di farmaci e nella progettazione di reazioni è promettente per lo sviluppo di nuovi farmaci. I modelli di AI/ML possono automatizzare l'ottimizzazione delle condizioni di reazione, prevedere le proprietà di nuovi composti e la fattibilità delle reazioni, accelerando, in ultima analisi, la scoperta di farmaci, riducendo tempi e costi e portando a farmaci più efficaci e accessibili.

Introduzione

I database pubblici rivoluzionano la scoperta di farmaci e la progettazione di vie sintetiche, fornendo ai ricercatori accesso a informazioni molecolari, target e reazioni chimiche. L'integrazione di intelligenza artificiale (AI) e apprendimento automatico (ML) permette l'analisi di grandi quantità di dati provenienti da diverse fonti. Esploreremo l'impatto dei database pubblici e dell'AI/ML nella scoperta di nuovi farmaci, discutendo benefici, limiti e direzioni future.

Un vantaggio dell'uso di database pubblici è l'identificazione di nuovi target e strutture chimiche. Analizzando proteine, geni e biomolecole, i ricercatori possono individuare bersagli farmacologici e progettare composti interagenti con essi. Ad esempio, gli algoritmi di AI sono stati utilizzati per prevedere l'attività di composti contro diverse proteine, la così detta multi-farmacologia **[1]**, consentendo la scoperta di farmaci multi-target con maggiore efficacia e con ridotti effetti collaterali **[2]**.

Gli algoritmi di intelligenza artificiale sono stati utilizzati anche per ottimizzare le reazioni chimiche per la sintesi di ingredienti farmaceutici attivi, riducendo i tempi e i costi necessari per lo sviluppo dei farmaci **[3]**. Riguardo la reattività chimica, le uniche banche dati gratuitamente accessibili sono banche dati di brevetti come l'USPTO o come Espacenet, dato che la stragrande maggioranza delle banche dati di reazioni è commerciale. Nonostante ciò, analizzando grandi insiemi di dati di reazioni chimiche **[4]**, gli algoritmi di AI possono prevedere i rendimenti delle reazioni e identificare le condizioni ottimali di reazione **[5]**, consentendo ai ricercatori di progettare nuove vie di sintesi con maggiore efficienza e flessibilità.

Esistono anche limitazioni e sfide associate all'uso di database pubblici e AI/ML nella scoperta di farmaci e nella progettazione di nuove reazioni. Ad esempio, la qualità e la completezza dei dati presenti nei database pubblici possono variare notevolmente **[6]** e l'integrazione di più set di dati può essere difficile a causa delle differenze nei formati e negli standard utilizzati **[7]**. Inoltre, l'uso di algoritmi di AI/ML può essere computazionalmente intensivo e richiedere risorse significative, rendendo difficile per i gruppi di ricerca più piccoli utilizzare appieno questi strumenti.

Da non dimenticare anche considerazioni etiche e problemi di privacy associati all'uso di database pubblici e AI/ML nella scoperta di farmaci e nella progettazione di nuove reazioni **[8]**. L'uso di dati sensibili dei pazienti per identificare bersagli farmacologici o per ottimizzare le proprietà dei farmaci solleva preoccupazioni sulla privacy dei pazienti e sulla sicurezza dei dati **[9]**. Inoltre, l'uso di algoritmi di AI/ML per prendere decisioni sullo sviluppo di farmaci e sulla cura dei pazienti solleva questioni relative alla parzialità e alla trasparenza degli algoritmi **[10, 11]**.

Nella presente rassegna cercheremo di coprire i principali aspetti legati all'uso dell'AI/ML nella scoperta di farmaci e alla generazione di nuovi spazi chimici per la scoperta di farmaci.

Banche dati

Tra i principali database accessibili pubblicamente possiamo citare *Protein Data Bank* (PDB) **[12]** con informazioni strutturali tridimensionali su proteine, acidi nucleici e ligandi ad essi complessati; *Chemical Entities of Biological Interest* (ChEBI) **[13]**, focalizzato sulle terminologie per una ontologia chimica necessaria per la costruzione di qualunque banca dati contenente molecole; *PubChem* **[14]**, un database di sostanze chimiche e delle loro attività biologiche; *ChEMBL* **[15]** contenente molecole bioattive e i loro bersagli proteici, con particolare attenzione a dati preclinici; *RCSB Ligand Explorer* **[16]**, un sottoinsieme del database PDB focalizzato su ligandi di piccole dimensioni e le loro interazioni con proteine e acidi nucleici; *OrphaNet* **[17]**, che raccoglie sia composti attivi sia target genetici associati a malattie rare; *Drugbank* **[18]**, un database semicommerciale che raccoglie una grande quantità di annotazioni su farmaci autorizzati; *UniProt* **[19]**, che contiene sequenze, strutture e informazioni funzionali sulle proteine; *Gene Expression Omnibus* (GEO), dove sono riportati dati di espressione genica provenienti da una varietà di organismi e tipi cellulari; *The Human Protein Atlas* **[21]**, un database che raccoglie una serie di dati di espressione delle proteine nei tessuti e nelle cellule umane; *MetaboLights* **[22]**, che raccoglie dati di metabolomica, compresi i profili dei metaboliti e le vie metaboliche note; *KEGG* **[23]**, basato su vie metaboliche note e relativi geni, proteine e piccole molecole; *ChemSpider* **[24]**, un database di strutture chimiche e delle loro proprietà, compresi i dati spettrali e reazioni chimiche associate.

La qualità dei dati contenuti in questi database è un aspetto importante: cosa succede se i dati che possono esserne scaricati non sono documentati, controllati, facilmente formattabili, non aggiornati o semplicemente incompleti? L'intera utilizzazione per le applicazioni elettroniche di questi dati può essere messa a rischio dalla loro mancanza di qualità. Negli ultimi anni ci sono stati molti successi nei tentativi di rendere questi dati *FAIR* (cioè *Findable, Accessible, Interoperable, Reusable*) **[25, 26]**.

E adesso?

Ora che tutti questi dati sono stati resi a portata di mano dell'utente, quali strumenti di AI/ML possiamo utilizzare per la scoperta dei farmaci? Portare un composto sul mercato come farmaco è un processo complesso che integra molte informazioni e risorse. Stime recenti collocano i costi totali associati tra gli oltre 300 milioni e i 2,8 miliardi di dollari **[27]**. Una delle prime e fondamentali domande in ricerca farmaceutica è se una proteina/gene possa essere identificata come fortemente correlata a una malattia e se sia "accessibile ad un farmaco", cioè modulabile nella sua concentrazione o nelle sue funzioni nelle cellule da una molecola. Uno dei primi tentativi di identificazione automatica di un target proteico in questa direzione risale al 2011 **[28]**. In seguito sono stati condotti innumerevoli altri tentativi e questo campo dell'identificazione di target guidata dal computer all'interno di un approccio di biologia sistematica è sempre più attivo e promettente **[29-31]**.

La tecnologia dell'informazione utilizza da tempo grafi per studiare reti di qualsiasi tipo e natura. In linea di principio, ogni database citato più sopra può essere trasformato e utilizzato come una rete (*knowledge graph*). La progressiva interoperabilità dei dati prodotti nella ricerca farmaceutica ha recentemente aperto nuove opportunità di integrazione delle informazioni attraverso questo approccio grafico **[32]**. Ogni nodo di un *knowledge graph* può essere costituito da molecole, test biologici, proteine, geni, documenti brevettuali, risultati di letteratura scientifica ecc. Essi possono essere correlati l'uno all'altro quando un'affermazione o un dato sperimentale li lega. Naturalmente, la qualità delle affermazioni è rilevante, poiché spesso si possono osservare affermazioni contraddittorie, dato che è fondamentale il contesto biologico che genera le affermazioni. Così, per esempio, gli stessi composti possono produrre attività biologiche diverse a seconda dei tipi di cellule utilizzate per il saggio o a seconda del tempo scelto per la lettura dei dati cellulari. Recentemente sono stati pubblicati diversi *knowledge graphs*, in particolare per condensare le conoscenze sulle malattie o su agenti patogeni come virus o batteri **[33-36]**. L'enorme accessibilità dei dati dovuta alla digitalizzazione ha fatto emergere anche la necessità di nuovi algoritmi statistici veloci e stabili. I modelli predittivi basati sui dati sperimentali del mondo reale accelerano la generazione di ipotesi, la valutazione dei

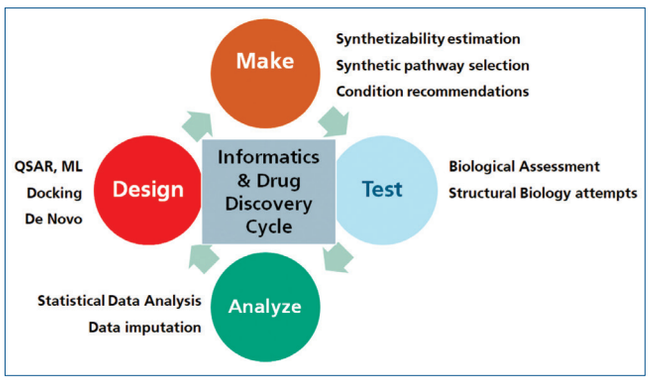

Fig. 1 - Alcune opportunità di intervento informatico nel ciclo classico di attività di drug discovery MTAD (make-test-analyze-design)

rischi e delle tossicità ed i processi decisionali necessari in ricerca. L'uso di diverse strutture di reti neurali (NN) ed il cosiddetto *deep-learning* (DL) **[37]** sono in grado non solo di modellare in modo statisticamente robusto la quantità di dati noti prodotti nel passato (modelli retrospettivi) **[38]** ma anche di prevedere in modo affidabile caratteristiche o risultati per molecole esistenti o virtuali. Questo sviluppo di metodiche sta cambiando definitivamente il modo in cui la ricerca farmaceutica preclinica viene gestita **[39]**. Anche gli enti regolatori che guidano il processo di autorizzazione dei farmaci, come la FDA in US ed EMA in Europa, stanno progressivamente accettando dati solo previsti dai calcoli e non sperimentali (Fig. 1). Le caratteristiche chimico-fisiche dei composti e le previsioni *ADME-Tox (Absorption, Distribution, Metabolism, Excretion and Toxicity)* sono state sempre più accettate dalla FDA, soprattutto nei casi in cui i dati farmacologici o farmacocinetici sono difficili da prodursi sperimentalmente **[40-42]**.

E quindi quale sarà il farmaco del futuro?

I chimici di sintesi e farmaceutici sono sempre più professionisti essenziali nel processo di scoperta dei farmaci. Ci sarà in futuro per la progettazione e la sintesi automatica di farmaci? Sono già stati riportati diversi successi dalla progettazione di farmaci basati su studi strutturali (*SBDD*) e sullo sviluppo sintetico di frammenti (*FBDD*), dove la forte integrazione della biologia strutturale e degli approcci AI/ML hanno permesso una progettazione più rapida (*de-novo*) **[43]** e più diversificata (*scaffold-hopping*) **[44]** di nuovi composti.

De-Novo?

La progettazione *de-novo* di molecole è iniziata molto prima della comparsa dei metodi di AI generativa. Si basavano principalmente sull'approccio retrosintetico dei farmaci attivi, seguita dalla generazione e dalla ricombinazione dei diversi frammenti, a volte vincolati da limitazioni nei descrittori chimico-fisici o dalle previsioni di modelli statistici esterni. Questi progetti *de-novo* ottimizzati **[45, 46]** hanno raggiunto una certa popolarità, ma è attraverso approcci guidati dalla ML o da deep-learning e diretti alla valutazione della sintetizzabilità che questo campo di studio ha raggiunto un nuovo livello di maturità **[47, 48]**. La nascita di start-up focalizzate sulla ricerca di farmaci basata sull'AI, come BenevolentAI **[49]**, Exscientia **[50]** e HealX **[51]**, combina nei loro metodi generativi di nuovi composti, previsioni sulla sintetizzabilità dei composti, in modo da poter raggiungere facilmente un gran numero di molecole da un limitato numero di reazioni **[52-54]**.

Le start-up basate sull'intelligenza artificiale, ma non solo loro, sostengono di ridurre enormemente i tempi di scoperta di nuovi candidati farmaci (Fig. 2).

Da dove vengono i nuovi composti?

La reattività chimica rimane alla base della creazione della diversità chimica. La necessità di reazioni di sintesi nuove, efficienti e pulite è molto elevata. Partendo da uno spazio chimico previsto di 10^{60} molecole, le librerie commerciali chimiche virtuali stanno raggiungendo dimensioni senza precedenti di ca $10^{10\text{-}15}$ molecole sintetizzabili. Il vero problema computazionale in questo campo si è spostato

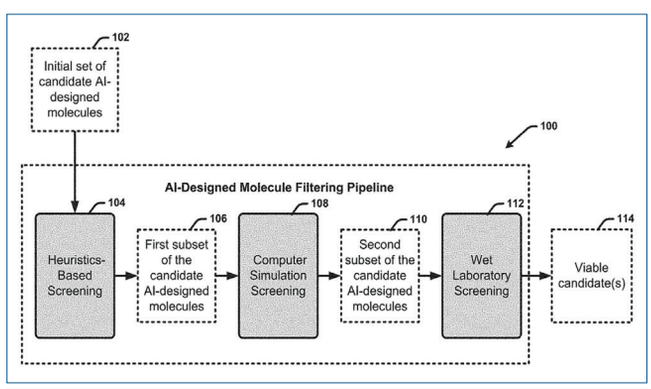

Fig. 2 - Workflow IBM per la generazione di composti ad alta probabilità di successo. Brevetto US20210366580A1 (https://patents.google.com/patent/US20210366580A1/en)

dalla generazione alla capacità di ricerca rapida e precisa su queste enormi collezioni **[55-58]**.

È tutto vero o ci sono dei limiti?

Una delle principali limitazioni di questo approccio è la difficoltà di distinguere tra correlazione e causalità nei dati **[59]**. Sebbene i modelli AI/ML siano in grado di identificare le correlazioni tra diverse variabili, non possono sempre determinare quali variabili siano realmente causative di un particolare risultato. Questa indeterminazione può implicare previsioni e conclusioni imprecise non sempre evidenti. Un'altra sfida è la generalizzazione delle conclusioni **[60]**. I modelli di AI/ML sono addestrati su set di dati specifici e le loro prestazioni possono variare in modo significativo quando vengono applicati a nuovi dati che possono differire dal set di dati di training. Questo può portare a previsioni veramente imprecise o fallaci.

Inoltre, i modelli AI/ML sono spesso percepiti come una scatola nera, il che significa che i loro processi decisionali non sono facilmente interpretabili dall'uomo. Questa mancanza di trasparenza può rendere difficile capire come il modello sia arrivato alle sue conclusioni, rendendo difficile la convalida delle sue previsioni e la garanzia della sua sicurezza **[61]**.

Per far fronte a queste limitazioni e sfide, c'è una crescente necessità di "AI spiegabile" (XAI, *eXplainable AI*) nella ricerca farmaceutica **[62, 63]**. XAI si riferisce all'uso di specifici modelli AI/ML in grado di fornire risultati trasparenti e interpretabili, consentendo ai ricercatori di capire come il modello sia arrivato alle sue previsioni **[64, 65]**. I modelli XAI possono contribuire a migliorare l'accuratezza delle previsioni, garantendo al contempo la sicurezza e l'affidabilità del processo.

Considerazioni etiche

I modelli di AI di per sé sono eticamente agnostici, poiché sono progettati e addestrati in base ai dati che ricevono. Tuttavia, le implicazioni etiche dei modelli di AI risiedono nelle loro applicazioni e nel modo in cui vengono utilizzati specialmente quando applicati in fase clinica. I modelli di AI possono perpetuare pregiudizi, discriminazioni e ingiustizie se non vengono progettati e addestrati tenendo conto di considerazioni etiche **[66]**. Pertanto, è responsabilità di sviluppatori, ricercatori e responsabili politici garantire che i modelli di AI siano progettati e utilizzati in modo etico. Ciò include la promozione della trasparenza, della responsabilità e dell'equità nello sviluppo e nell'impiego dell'AI per prevenire danni agli individui e alla società nel suo complesso **[67]**.

Direzioni future

Nonostante queste sfide, l'uso di database pubblici e algoritmi di AI/ML nella scoperta di farmaci e nella progettazione di nuove reazioni è molto promettente. Modelli AI/ML hanno il potenziale per rivoluzionare il campo della sintesi dei farmaci, accelerando la scoperta e l'ottimizzazione di nuovi composti chimici. Alcune possibili direzioni future per i modelli AI/ML applicati alla sintesi dei farmaci includono l'ottimizzazione automatizzata delle condizioni di reazione prevedendo il solvente, la temperatura e altri parametri ottimali per una determinata reazione. Modelli AI/ML possono aiutare a prevedere la fattibilità delle reazioni chimiche analizzando la struttura chimica dei reagenti e dei prodotti. Ciò può aiutare a identificare potenziali ostacoli e a guidare la sintesi di nuovi composti chimici anche dal punto di vista brevettuale **[68]**.

BIBLIOGRAFIA

[1] **https://pubs.acs.org/doi/full/10.1021/acs.jcim.8b00677**
[2] **https://www.biorxiv.org/content/10.1101/2022.12.16.520738v2.abstract**
[3] **https://doi.org/10.1016/j.coche.2021.100749**
[4] **https://doi.org/10.1016/j.coche.2021.100749**
[5] **https://doi.org/10.3390/pr11020330**
[6] **https://doi.org/10.1517/17460441.2011.579100**
[7] **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7778943/**
[8] **https://www.nature.com/articles/s42256-022-00465-9**
[9] **https://doi.org/10.3390/app12041927**
[10] **https://doi.org/10.1016/j.inffus.2023.03.008**
[11] **https://link.springer.com/article/10.1007/s43681-021-00131-7**

[12] https://www.r csb.org/
[13] https://www.ebi.ac.uk/chebi/
[14] https://pubchem.ncbi.nlm.nih.gov
[15] https://www.ebi.ac.uk/chembl
[16] http://ligand-expo.rcsb.org/
[17] https://www.orpha.net/consor/cgi-bin/index.php
[18] https://go.drugbank.com
[19] https://www.uniprot.org/
[20] https://www.ncbi.nlm.nih.gov/geo
[21] https://www.proteinatlas.org
[22] https://www.ebi.ac.uk/metabolights
[23] https://www.genome.jp/kegg
[24] http://www.chemspider.com
[25] https://faircookbook.elixir-europe.org/content/home.html
[26] https://www.go-fair.org/resources/more-on-fair/
[27] https://pubmed.ncbi.nlm.nih.gov/32125404/
[28] https://www.science.org/doi/10.1126/scisignal.2001950
[29] https://doi.org/10.1016/j.copbio.2016.04.007
[30] https://www.doi.org/10.1186/S13321-015-0055-9
[31] https://www.doi.org/10.1021/ACS.JMEDCHEM.9B01989
[32] https://doi.org/10.1080/17460441.2021.1910673
[33] https://doi.org/10.1016/j.csbj.2020.05.017
[34] https://doi.org/10.1093/bib/bbac543
[35] https://doi.org/10.1016/j.artmed.2020.101817
[36] https://doi.org/10.1093/bioadv/vbad045
[37] https://doi.org/10.1080/17460441.2016.1200126
[38] https://doi.org/10.1080/17460441.2016.1201262
[39] https://doi.org/10.1016/j.drudis.2018.01.039
[40] https://doi.org/10.1093/toxsci/kft189
[41] https://doi.org/10.3390/md21010024
[42] https://doi.org/10.1002/jcph.1478
[43] https://doi.org/10.1073/pnas.2206240119
[44] https://doi.org/10.1080/17425255.2020.1777280
[45] https://doi.org/10.1002/wcms.49
[46] https://jcheminf.biomedcentral.com/articles/10.1186/s13321-022-00582-y
[47] https://doi.org/10.1016/j.compbiomed.2022.105403
[48] https://www.nature.com/articles/s42256-022-00463-x
[49] https://www.benevolent.com/
[50] https://www.exscientia.ai/
[51] https://healx.ai/
[52] https://www.nature.com/articles/nature25978
[53] https://doi.org/10.1021%2Fci300116p
[54] https://doi.org/10.1007%2Fs10822-006-9099-2
[55] https://doi.org/10.1021/acs.jcim.2c00390
[56] https://doi.org/10.1021/acs.jcim.1c00811
[57] https://doi:10.1021/acs.jcim.1c00640
[58] https://doi.org/10.1021/acs.jcim.1c00975
[59] https://link.springer.com/article/10.1007/s10462-022-10381-4
[60] https://doi.org/10.1016/j.drudis.2022.05.009
[61] https://doi.org/10.1158/1538-7445.AM2022-454
[62] https://link.springer.com/article/10.1007/s10462-022-10306-1
[63] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4390887
[64] https://ieeexplore.ieee.org/abstract/document/9916585
[65] https://doi.org/10.1016/j.inffus.2021.07.016
[66] https://link.springer.com/article/10.1007/s43681-023-00258-9
[67] https://link.springer.com/chapter/10.1007/978-3-031-09846-8_12
[68] https://doi.org/10.1016/j.ailsci.2023.100069

Databases and AI in Drug Discovery

The use of public databases and AI/ML algorithms in drug discovery and reaction design offers promising possibilities for the development of more effective drugs. AI/ML models can optimize reaction conditions, predict properties of new compounds and identify potential roadblocks in chemical reactions. These future directions have the potential to accelerate drug discovery, reduce time and cost of development, and ultimately result in the creation of more effective and affordable medicines.