

A pesca di molecole per arricchire l'arsenale farmaceutico

Il drug discovery con il machine learning: una strada percorribile o una scatola nera?

L'idea alla base dell'utilizzo dell'intelligenza artificiale in farmacologia è di semplificare la complessità e far emergere potenziali linee di ricerca che altrimenti "a occhio nudo" i ricercatori difficilmente sarebbero in grado di individuare. Come se l'universo di proteine nascondesse messaggi scritti con inchiostro simpatico, invisibili ai più, e che richiedono degli occhiali nuovi per essere interpretati.

Il primo bias nello sviluppo di un farmaco è proprio l'avvio della ricerca, l'idea di provare a lavorare su una certa proteina e non su un'altra. L'intuizione, seppure evidence based. Una volta selezionata la proteina da cui partire, si innesca quel noto processo a imbuto che, sfortunatamente solo in pochi casi, può tradursi in un medicinale vero e proprio. Un processo però lungo, anche fino a dieci anni, e che lascia dietro di sé numerose "vittime", molecole cioè candidate a diventare un farmaco ma che per diverse ragioni si rivelano non adatte e quindi vengono scartate. Per citare una statistica nota, in media solo il 14 per cento delle molecole che supera la prima fase di test è adatta a diventare un farmaco.

Anche ai non addetti ai lavori risulta evidente che l'intero processo è condizionato dal limite umano nel gestire e setacciare grandi moli di dati. Nella fattispecie, capire fra miliardi di proteine esistenti qual è la candidata più adatta per bersagliare la malattia che vorremmo colpire. Il risultato è che siamo obbligati a utilizzare un processo top-down, quando invece a rigor di logica dovremmo procedere bottom-up.

“*Serve un modo per scegliere dove scavare.*”

— Jean Louis Reymond

Nel 2017 Jean Louis Reymond dell'Università di Berna ha pubblicato una "mappa" delle molecole note: un database di 166 miliardi di composti organici chimicamente realizzabili fino a 17 atomi che ha chiamato Gdb-17. È lapalissiano che la mente umana non riesca a confrontare nei tempi dettati dalla ricerca scientifica una mole così imponente di composti per individuare i migliori candidati sui quali puntare e investire a partire dalle caratteristiche del bersaglio terapeutico.

Ebbene, proprio qui si può inserire secondo molti il contributo dell'intelligenza artificiale. Come? Mettendo a punto algoritmi sempre più "intelligenti" in grado di setacciare a tappeto l'oceano di composti, fra cui le varianti di proteine già studiate e magari scartate, e individuare nuovi candidati promettenti. Si tratta di una branca della ricerca che per alcuni rappresenta oggi una speranza, per altri una strada percorribile, per altri ancora una scatola nera per la medicina del futuro. Se infatti l'idea è semplice, realizzarla pone ancora diverse questioni che taluni reputano ancora troppo aperte per poter andare oltre il parlare di speranza.

Il gruppo di Reymond a Berna ha provato a

catalogare questo enorme database, in modo da rendere più agile l'azione dell'algoritmo e identificare sostanze terapeuticamente promettenti "molto vicine" a farmaci già testati, ricercando le somiglianze tra i composti. L'algoritmo sviluppato dal team in soli tre minuti può scorrere i 166 miliardi di molecole catalogate in Gdb-17. Per esempio il gruppo è partito da una molecola che lega il recettore nicotinico dell'acetilcolina, utile nell'ambito dei disturbi che riguardano il sistema nervoso o la funzione muscolare, selezionando una prima lista di 344 composti correlati. Di questi il gruppo ne ha sintetizzati tre, due dei quali potrebbero rivelarsi – secondo i ricercatori – promettenti per il target esaminato¹.

Il drug discovery attraverso l'apprendimento automatico

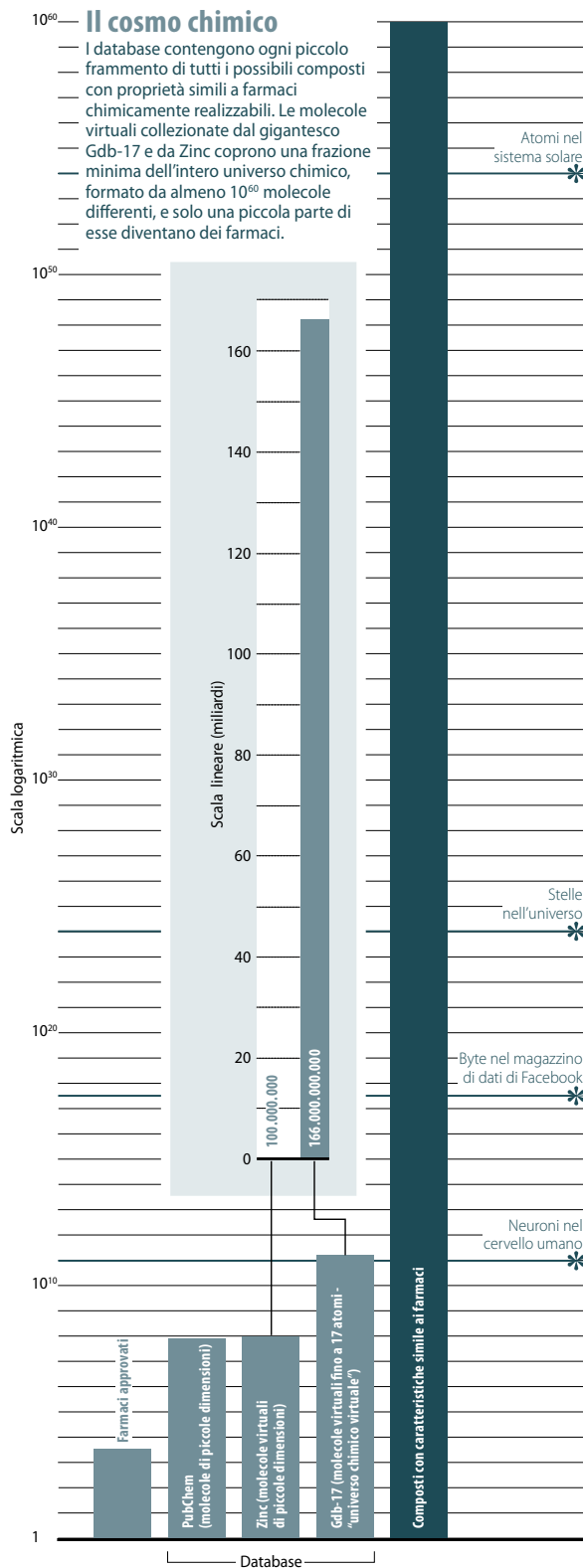
L'esempio citato ha come caratteristica quella di partire da un farmaco già noto per esaminare delle possibili varianti promettenti. Vi sono però algoritmi in fase di studio che si pongono l'obiettivo di fare un passo ulteriore: individuare nuovi candidati senza aver prima definito il target terapeutico. Si tratta di un approccio decisamente più radicale, che si basa sulla ricerca di corrispondenze, cioè di molecole che si legano a una determinata proteina. L'idea di fondo è quella di utilizzare algoritmi detti di "legame molecolare" su un'istantanea di una proteina.

Un altro gruppo, questa volta presso l'Università di San Francisco, ha usato tecniche simili nel 2016 per individuare possibili candidati a diventare una nuova classe di antidolorifici. Su oltre 3 milioni di composti (molti di meno dunque rispetto alla mappa di Reymond), il team ne ha selezionati 23, di cui 7 in provetta hanno mostrato l'attività desiderata, e uno di essi è attualmente in fase di sperimentazione per diventare – forse – un nuovo farmaco.

Si tratta di un approccio diverso rispetto al primo, e che pone diverse difficoltà, legate alla frontiera rappresentata dall'apprendimento automatico. Gli algoritmi, per poter selezionare proteine senza bias e individuare modelli invisibili all'uomo, devono conoscere i composti, le loro caratteristiche e i diversi tipi di legame. Si chiama deep learning, apprendimento profondo, quel processo gerarchico tramite cui si "insegna" agli algoritmi a categorizzare i dati entro diversi livelli di rappresentazione, a partire dai concetti più in basso, cioè più semplici, per poi salire di gerarchia fino a costruire quelli più complessi. Un po' come partire dal concetto di insieme, per poi passare a quello di funzione, fino ad arrivare a quello di integrale.

Il cosmo chimico

I database contengono ogni piccolo frammento di tutti i possibili composti con proprietà simili a farmaci chimicamente realizzabili. Le molecole virtuali collezionate dal gigantesco Gdb-17 e da Zinc coprono una frazione minima dell'intero universo chimico, formato da almeno 10^{60} molecole differenti, e solo una piccola parte di esse diventano dei farmaci.



Il deep learning apre le porte – per lo meno virtualmente – a un terzo percorso percorribile con l'intelligenza artificiale, ancora più radicale: quello della progettazione di nuove molecole in silico, ovvero al computer. Spostare quindi l'intero processo nella dimensione del virtuale, cioè della modellistica in silico, per estrapolare quello che potrebbe essere un buon candidato per un nuovo farmaco, da sintetizzare materialmente in laboratorio in un secondo momento. In questa direzione lavora un team di ricercatori dell'Mit di Boston, che sta provando a "insegnare" all'algoritmo a lavorare su 250 molecole (ancora una volta siamo a diversi ordini di grandezza di distanza rispetto alla mappa di Reymond) per progettare nuove molecole, prevederne il comportamento e suggerire possibili combinazioni.

Investire nelle reti intelligenti

Siamo qui a un livello di simulazione completamente virtuale, che rappresenta – si legge su *Nature* – uno scollamento importante con la reale capacità di realizzare in laboratorio siffatte molecole. In soldoni, lo sforzo di elaborare il candidato perfetto potrebbe vanificarsi se non possedessimo poi le conoscenze ingegneristiche tali da permetterci di fabbricare concretamente queste molecole potenzialmente perfette. O di farlo a costi contenuti. C'è infatti il problema sempiterno dei costi. Un algoritmo per produrre ottimi risultati deve avere a disposizione moltissimi dati su cui lavorare, e i dati costano. È quindi necessario sviluppare algoritmi che siano in grado di lavorare sempre con meno dati.

Una cosa è certa: l'industria farmaceutica sta investendo parecchio in questa direzione. Lo testimoniano le cifre raccontate da un altro articolo apparso su *Nature Biotechnology*²: dal 2015 a oggi aziende farmaceutiche come Merck, Janssen Pharmaceutica, Sanofi, Genentech, Takeda, Santen Pharmaceuticals hanno stipulato collaborazioni con realtà note nell'ambito dell'intelligenza artificiale per lo sviluppo di nuove linee terapeutiche in diversi ambiti, da quello oncologico, a quello delle malattie infettive, malaria in *primis*, fino alle malattie rare.

Il modello di apprendimento più conosciuto per la scoperta di nuovi farmaci è forse il sistema Watson sviluppato dall'Ibm, racconta Eric Smalley su *Nature Biotechnology*². Nel dicembre 2016 l'Ibm ha firmato un accordo con Pfizer per lo sviluppo di farmaci in ambito immuno-oncologico, una delle strade più promettenti e più acclamate per la lotta contro il cancro.

Bicchieri mezzo pieno o mezzo vuoto, dunque? "Si tratta di un campo senza dubbio interessante, ma prima di capire se possiamo essere ottimisti è necessario avere fra le mani concrete storie di successo", conclude Smalley. "Dobbiamo prima vedere la prima auto senza conducente, prima di poter affermare di avere fra le mani la nuova Tesla del Pharma".

Cristina Da Rold, freelance data journalist
e scientific communicator

- Mullard A. The drug-maker's guide to the galaxy. How machine learning and big data are helping chemists search the vast chemical universe for better medicines. *Nature*, 26 settembre 2017
- Smalley E. AI-powered drug discovery captures pharma interest. *Nature Biotechnology* 2017; 35:604-5.

L'assistente chimico virtuale

Esplora le molecole intuitivamente, usando la conoscenza chimica

Quali sono le principali difficoltà che voi ricercatori state affrontando in questo momento? Qual è la sfida principale?

Dal mio punto di vista la sfida principale per chi lavora nel campo dell'intelligenza artificiale applicata alla farmaceutica è quella di avere a disposizione dei buoni dati dai quali "imparare". La questione chiave è avere accesso a set di dati di alta qualità per essere in grado di fare previsioni. A questo proposito alla Harvard university stiamo sviluppando modelli molto sofisticati in ambito chimico che stanno generando interesse da parte di aziende farmaceutiche. L'idea è che i nostri algoritmi possano essere usati anche nell'ambito della drug discovery, anche se non lavoriamo specificamente in quella direzione.

Più nel dettaglio, su che cosa state lavorando in questo momento?

Lavoriamo prevalentemente sulla ricerca di materiali organici funzionali utilizzati per assemblare i materiali elettronici e per lo stoccaggio di energia (batterie organiche, led organici, fotovoltaici organici) che abbiano dimensioni e complessità analoghe a quelle dei farmaci. Abbiamo impiegato con successo la chimica quantistica e i metodi di



Intervista a
Alan Aspuru-Guzik
Responsabile del gruppo
di fisica e chimica teorica
Dipartimento di chimica
e biologia chimica
Harvard university

machine learning per progettare diodi organici emettitori di luce, a dimostrazione della potenza di queste tecniche. Inoltre abbiamo sviluppato modelli generativi, che permettano al computer di selezionare composti candidati per applicazioni particolari. Siamo molto eccitati per i risultati ottenuti dai nostri ecoder e dalle reti antagoniste generative (generative adversarial network).

Ci può spiegare in che cosa consistono?

I generative adversarial network sono una classe di algoritmi per l'intelligenza artificiale utilizzabili nell'apprendi-

mento automatico. Si tratta di tecniche applicabili anche nella ricerca di nuovi farmaci, perché possono individuare nuovi candidati che non sono già inclusi nei database esistenti. Questa è una differenza fondamentale rispetto all'approccio di Reymond e colleghi (vedi p. 24) e di altri che sviluppano invece "librerie" specifiche. I nostri algoritmi possono essere utilizzati per "addestrare" i nostri modelli, affinché diventino capaci di "immaginare" nuove molecole candidate simili a quelle già note e catalogate, ma con caratteristiche migliori rispetto a determinate proprietà molecolari.

Come e perché ha scelto di dedicarsi a questo particolare ambito di ricerca?

Il campo di ricerca del mio team è l'intersezione fra calcolo quantistico e chimica. In questo momento ci occupiamo di interfacciare la chimica con settori in rapida crescita, come il mondo della computazione. Da alcuni anni abbiamo cominciato a lavorare proprio con l'intelligenza artificiale, perché crediamo sia un settore in rapido movimento e per questo molto emozionante. Siamo certi che l'intelligenza artificiale arricchirà le possibilità dei chimici computazionali negli anni a venire.

[A cura di **Cristina Da Rold**]

