

28. marzo

Proteomica e IA

*Ci sono cose che l'intelligenza è capace di cercare,
ma che, da sola, non troverà mai.*

Henri Bergson

Negli ultimi mesi **ChatGPT** e altri **chatbot** di intelligenza artificiale con straordinarie capacità di rispondere ai suggerimenti con una scrittura fluida e simile a quella umana hanno scatenato torrenti di angoscia che fluiscono da diversi ambienti della società; i **chatbot** potrebbero aiutare gli studenti a imbrogliare, invadere posti di lavoro o produrre disinformazione in massa.

Al di fuori dei riflettori puntati sui **chatbot**, anche i ricercatori delle scienze della vita hanno lanciato una tecnologia simile basata sull'intelligenza artificiale, ma con molto meno clamore.

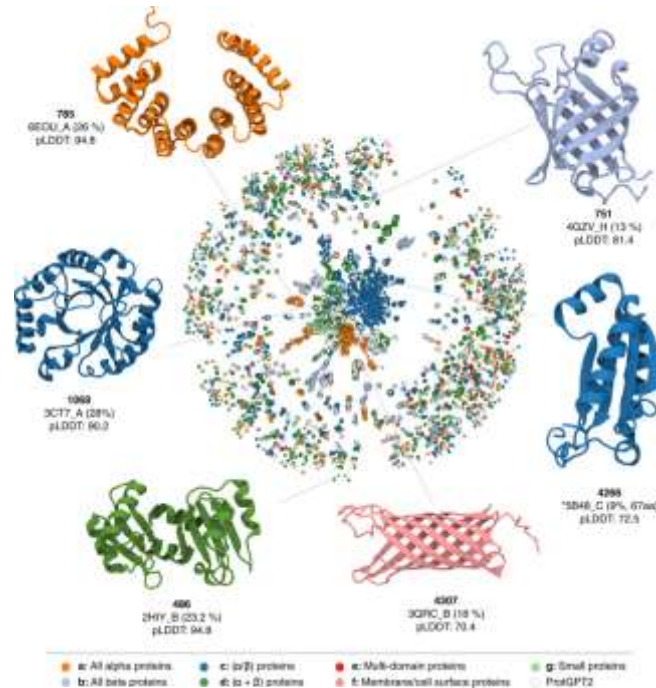
Questo è preoccupante, perché i nuovi algoritmi per la progettazione delle proteine, mentre potenzialmente migliorano la capacità di combattere le malattie, possono anche creare significative opportunità di uso improprio.



Poiché i processi di produzione biotecnologica si stanno evolvendo per rendere più facile per i creatori realizzare il DNA sintetico e altri prodotti che hanno progettato, nuovi modelli di intelligenza artificiale come **ProtGPT2** e **ProGen** consentiranno ai ricercatori di concepire una gamma molto più ampia di molecole e proteine che mai.

La natura ha impiegato milioni di anni per progettare le proteine. L'intelligenza artificiale può generare sequenze proteiche significative in pochi secondi

Ferruz N et al . [ProtGPT2 is a deep unsupervised language model for protein design](#). Nat Commun. 2022 Jul 27;13(1):4348.



Una panoramica dello spazio proteico ed esempi di proteine generate da ProtGPT2.

Ogni nodo rappresenta una sequenza. Due nodi sono collegati quando hanno un allineamento di almeno 20 aminoacidi e una probabilità di ricerca HH del 70%. I colori rappresentano le diverse classi SCOPe e le sequenze ProtGPT2 sono mostrate in bianco. Come esempio, selezioniamo le proteine di ciascuna delle cinque principali classi SCOP: strutture all- β (751), α/β (4266 e 1068), proteina di membrana (4307), $\alpha+\beta$ (486) e all- α (785). Le strutture selezionate sono colorate in base alla classe del loro colpo più simile. Le strutture sono state previste con AlphaFold, e indichiamo il codice della struttura più simile nel PDB trovata da FoldSeek [58](#), ad eccezione della proteina 4266, dove non sono state trovate strutture. (da Ferruz N 2022)

Mentre ci sono buone ragioni per sviluppare la tecnologia AI per la progettazione biologica, ci sono anche rischi per tali sforzi che gli scienziati del settore non sembrano aver valutato. L'intelligenza artificiale potrebbe essere utilizzata per progettare nuove armi biologiche o tossine che non possono essere rilevate.

Man mano che questi sistemi si sviluppano insieme a nuove capacità di produzione più facili, più economiche e più veloci, gli scienziati dovrebbero parlare e imparare da colleghi che si concentrano sui rischi per la biosicurezza.

Qual è il problema con le proteine ingegnerizzate?

Sebbene molte applicazioni dell'ingegneria proteica siano utili, le armi a base di tossine proteiche sono state a lungo una preoccupazione. Questi sono veleni creati da organismi, come piante o funghi. La **ricina**, ad esempio, una tossina ricavata dai semi di ricino, è stata probabilmente utilizzata da agenti bulgari a Londra nel 1978

Nel ben documentato assassinio di **Georgi Markov**, un dissidente bulgaro.



[Murder on Waterloo Bridge: placing the assassination of Georgi Markov in past and present context, 1970 - 2018](#)
[Daniel Salisbury & Karl Dewey](#). Pages 128-156 | Published online: 18 Jan 2023

Un altro esempio è la neurotossina botulinica; è 100.000 volte più tossico dell'agente nervino sarin ed è stato un punto fermo degli arsenali statali di armi biologiche .

Botulinum Toxin (Botulism)

Agent Fact Sheet del February 09, 2023 della J. Hopkidss

Se la tecnologia per lo sviluppo e la produzione di tossine proteiche migliorasse, ad esempio rendendo più semplice la progettazione di nuove tossine o migliorando quelle esistenti, ciò potrebbe rappresentare un rischio elevato.

Gli autori di un primo articolo sui rischi dell'ingegneria proteica, il campo coinvolto nella progettazione e produzione di proteine, ha osservato nel 2006 che la produzione di tossine in quantità maggiori stava diventando più fattibile.

Tucker JB, Hooper C. Protein engineering: security implications. The increasing ability to manipulate protein toxins for hostile purposes has prompted calls for regulation. EMBO Rep. 2006 Jul;7 Spec No(Spec No):S14-7.

Invece di dover estrarre le tossine dalle loro fonti, le nuove tecnologie hanno permesso di produrle nei batteri o in altre cellule. Gli esperti di biosicurezza **Jonathan Tucker e Craig Hooper** erano particolarmente preoccupati per il crescente interesse medico per le cosiddette "tossine da fusione", che potrebbero combinare elementi di una tossina come la **ricina** con anticorpi che potrebbero azzerare le cellule tumorali. Insieme alle cellule tumorali, una **tale** tecnologia potrebbe potenzialmente essere utilizzata per distruggere altre cellule sane.

Mentre l'ingegneria delle proteine stava facendo passi da gigante, gli autori la vedevano come un'area di rischio a duplice uso "spesso trascurata" dalla comunità della biosicurezza.

Una cosa che limitava questi rischi, secondo **Tucker e Hooper**, era la mancanza di potenza di calcolo per progettare nuove proteine. Circa 17 anni dopo, con l'avvento della progettazione proteica basata sull'intelligenza artificiale e molta più potenza di calcolo, questa limitazione sembra essersi erosa.

Perché generare nuove proteine?

La progettazione di proteine assistita dall'intelligenza artificiale è un'area potenzialmente rivoluzionaria nelle scienze della vita. Le proteine sono le macchine molecolari della vita. Gli esseri umani e altre specie producono migliaia di queste molecole complesse, che svolgono un'ampia gamma di funzioni critiche.

Questi includono consentire ai muscoli di muoversi, far entrare e uscire le molecole dalle cellule e abbattere i nutrienti in cibi e bevande. Laddove l'evoluzione ha impiegato centinaia di milioni di anni o più per perfezionare le proteine, gli scienziati possono ora "addestrare" un'intelligenza artificiale, utilizzando centinaia di milioni di sequenze proteiche come esempi, per farlo molto più rapidamente.

Ma perché qualcuno dovrebbe voler produrre nuove proteine?

Immagina un enzima proteico naturale che catalizza una reazione chimica utile per produrre un farmaco. Un'intelligenza artificiale potrebbe evocare un enzima che rende più veloce il prezioso prodotto.

Oppure potrebbe inventare un farmaco a base di proteine leggermente modificato, con qualità migliorate. Forse un giorno l'intelligenza artificiale potrebbe essere utile per sviluppare proteine con reazioni completamente nuove.

I ricercatori potrebbero progettare proteine che si legano specificamente alle proteine umane e le inibiscono per correggere una malattia genetica. O proteine che neutralizzano batteri o virus quando si legano a loro. O proteine che abbattano gli inquinanti. Le applicazioni commerciali della progettazione di proteine generative sono potenzialmente infinite.

*Cao L et al. [Design of protein-binding proteins from the target structure alone](#). *Nature*. 2022 May;605(7910):551-560.*

Quali sono i rischi delle proteine progettate dall'IA?

I modelli di linguaggio AI per la progettazione delle proteine vengono sviluppati rapidamente. ad esempio **ProtGPT2** è stato addestrato su 45 milioni di sequenze proteiche

*Ferruz N et al. [ProtGPT2 is a deep unsupervised language model for protein design](#). *Nat Commun*. 2022 Jul 27;13(1):4348.*

ProGen su 280 milioni di sequenze.

*Madani A et al. [Large language models generate functional protein sequences across diverse families](#). *Nat Biotechnol*. 2023 Jan 26.*

Entrambi sono stati descritti in recenti articoli pubblicati dal 2022. Gli scienziati coinvolti nei modelli hanno utilizzato il loro software per progettare nuove proteine che sono state poi realizzate e testate per verificare che fossero effettivamente funzionalmente competenti. I modelli stanno generando confronti diretti con la natura innovativa di ChatGPT e dei suoi recenti predecessori.

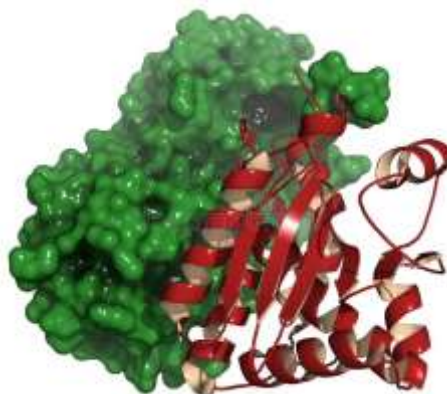
Alla fine di gennaio, [Ali Madani](#), parte del team che ha sviluppato ProGen,



ha twittato: *"ChatGPT per la biologia? Sono entusiasta di condividere oggi il nostro lavoro su [modelli di intelligenza artificiale] per la progettazione di proteine". Ispirati in parte dai generatori di testo AI, i ricercatori dietro ProtGPT2 hanno scritto: "ci siamo chiesti se potessimo addestrare un modello generativo per... imparare efficacemente il linguaggio delle proteine..."*.

Gli scienziati della vita non sono ingenui riguardo alle dimensioni della sicurezza del loro mestiere. Sebbene pochi vorrebbero che la loro scienza fosse usata per causare deliberatamente danni, molti, a quanto pare, non apprezzano appieno come tali rischi potrebbero applicarsi alle proprie aree di ricerca. I progettisti di proteine non sono diversi. Almeno non ci sono prove nei documenti o nelle pre stampa su [ProtGPT2](#) e [ProGen](#) che tale considerazione sia in corso. I progettisti di proteine rendono anche prontamente disponibili gli strumenti software, depositandoli in repository ampiamente accessibili, apparentemente senza considerare potenziali problemi di sicurezza. Questo è un problema. Nuove proteine possono aggirare le biodifese e altri controlli.

Prendiamo ad esempio il **sito attivo** della **ricina**, la tossina proteica vegetale.



La **ricina** può inattivare i ribosomi, le strutture cellulari che assemblano le proteine. In altre parole, è un pericoloso veleno. Un'IA di ingegneria proteica potrebbe, ad esempio, riprogettare la struttura proteica attorno al sito attivo della ricina, rimuovendo potenzialmente qualsiasi somiglianza di sequenza con la ricina. Chiunque lavori su tecnologie di rilevamento delle tossine, ad esempio laboratori di biodifesa o sicurezza alimentare, potrebbe avere difficoltà a identificare nuove tossine invisibili. I sistemi consolidati per rilevare e prevenire l'esportazione illegale di

tossine non le riconoscerebbero. Sebbene, fortunatamente, l'attività della ricina dipenda da una combinazione di fattori, e il cambiamento della struttura proteica circostante non ostacolerà necessariamente di per sé il rilevamento e il controllo, i progressi nella sintesi genica, le tecniche di espressione e purificazione delle proteine stanno aggiungendo velocità e facilità allo sviluppo di proteine progettate al computer. Questi probabilmente aumenteranno il rischio di uso improprio.

Produrre proteine progettate dall'intelligenza artificiale.

Progettare una molecola o una proteina che svolga un determinato compito può sembrare innocuo, ma unendo tale capacità alla tecnologia per realizzare effettivamente il prodotto, entriamo in un regno diverso. Potrebbe davvero accadere? Bene, ci sono già stati progressi nel campo della chimica robotica che creano una *pipeline integrata* che progetta e produce piccole molecole.

Coley CW et al. *A robotic platform for flow synthesis of organic compounds informed by AI planning*. *Science*. 2019 Aug 9;365(6453):eaax1566.

Mentre espandere questo concetto alla progettazione automatizzata e alla sintesi di molecole più grandi come le proteine sarebbe un grande salto tecnologico, i *recenti sviluppi* nella sintesi proteica stanno consentendo la produzione su piccola scala e ad alto rendimento di alcune proteine create dall'intelligenza artificiale.

Hartrampf N Et al. *Synthesis of proteins by automated flow chemistry*. *Science*. 2020 May 29;368(6494):980-987. doi: 10.1126/science.abb2491. PMID: 32467387

E le biofonderie, strutture all'avanguardia per la progettazione e la produzione di prodotti biologici molecolari, stanno consentendo la creazione di altre proteine. Tutto ciò che serve è che i nuovi algoritmi di progettazione delle proteine siano integrati con quei processi di produzione.

Per noi questo è un momento déjà vu. In precedenza, ci siamo imbattuti nel potere dell'utilizzo della IA generativa per sviluppare analoghi *dell'agente nervino VX* una potente arma chimica.

Urbina F et al. *Use of Artificial Intelligence-powered Drug Discovery*. *Nat Mach Intell*. 2022 Mar;4(3):189-191.

Cambia le impostazioni su un'intelligenza artificiale per la scoperta di droghe e, invece di molecole non tossiche, il software fornirà le "droghe" più tossiche. Il campo della progettazione di proteine generative offre parallelismi immediati, ma c'è poco riconoscimento nel campo della necessità di affrontare il potenziale rischio di duplice uso della tecnologia.

Le proteine possono avere una complessità 3D molto maggiore rispetto ai composti chimici che compongono molti farmaci, e potrebbe essere più difficile "vedere" lo scopo nefasto delle proteine/enzimi di nuova concezione. In effetti, potrebbero non essere paragonabili a nulla di noto. Ma la scienza responsabile richiede ancora la considerazione degli usi alternativi a cui la scienza può essere destinata, diversi da quelli previsti.

Un modo per affrontare il potenziale uso improprio è seguire un semplice codice: fermati, guarda e ascolta. Se stai sviluppando una nuova tecnologia o applicazione di IA generativa, fermati e considera il potenziale del duplice uso. Guarda cosa hanno fatto altri scienziati per risolvere questo problema nel loro software. E ascolta quando altri scienziati ti avvertono del potenziale uso improprio della tua tecnologia. La mancata adozione di tali precauzioni potrebbe portare alla continua ripetizione di momenti insegnabili nell'uso improprio dell'intelligenza artificiale o conseguenze molto peggiori.

Urbina F, Lentzos F, Invernizzi C, Ekins S. *A teachable moment for dual-use*. *Nat Mach Intell*. 2022 Jul;4(7):607.

Al di là anche dei rischi sostanziali della generazione di proteine, esiste un potenziale senza precedenti per utilizzare l'IA generativa per sviluppare nuove e invadenti forme sorveglianza alimentate da dati biologici e genomici

Driving Responsible Innovation of AI, Life Sciences and Next Generation Biotech

Nancy D. Connell PhD, IC20-027 Industry Connections (IC) Program Co-Chair Kobi Leins,

e persino armi biologiche

How to protect the world from ultra-targeted biological weapons

By Filippa Lentzos, December 7, 2020

Il potere di trasformazione della combinazione di intelligenza artificiale e scienze della vita potrebbe condurre l'umanità lungo un percorso irreversibile, con possibilità senza rivali di interferenza coercitiva e letale nei processi più intimi della vita. Deve essere aperto lo spazio per le conversazioni sia nella comunità della sicurezza internazionale che nella comunità delle scienze della vita, nonché tra queste comunità, su cosa potrebbe essere la ricerca sulle scienze della vita sicura, protetta e responsabile che coinvolge l'IA.

È facile eludere le nostre responsabilità, ma non possiamo eludere le conseguenze dell'aver eluso le nostre responsabilità. *Non è solo per quello che facciamo che siamo ritenuti responsabili, ma anche per quello che non facciamo. (Molière)*

A chi legge:

questo report è stato costruito a partire dai dati di:

-Sean Ekins

Sean Ekins è fondatore e CEO di Collaborations Pharmaceuticals, Inc., che si concentra sull'utilizzo di approcci di apprendimento automatico per rari e trascurati

-Filippa Lentzos

Filippa Lentzos è Reader (Associate Professor) in Science & International Security presso il King's College di Londra, dove è nominata congiuntamente.

-Max Brackmann

Max Brackmann lavora nel controllo degli armamenti NBC presso il laboratorio Spiez dell'Ufficio federale della protezione civile, dove si occupa di.

-Cédric Invernizzi

Cédric Invernizzi è un esperto senior nel controllo degli armamenti biologici presso il Laboratorio Spiez, l'Istituto federale svizzero per la protezione dal nuclear

Baedeker/Replay del 28marzo 2022

REACT-1 per la sorveglianza efficace di possibili nuove varianti

Mentre i paesi entrano nel terzo anno della pandemia di COVID-19, continuano ad affrontare la minaccia di nuove varianti e la sfida del loro monitoraggio.

In Inghilterra, lo studio Real-time Assessment of Community Transmission (REACT-1) monitora i livelli di infezione da maggio 2020, raccogliendo regolarmente campioni da centinaia di migliaia di individui selezionati casualmente e testando la presenza di sindrome respiratoria acuta grave coronavirus 2 (SARS-CoV-2). REACT-1 ha segnalato un rapido aumento della variante Omicron nei dati nel mese di dicembre 2021, dimostrando che questa variante sia cresciuta dal 10% di tutte le infezioni al 90% in poco più di una settimana. Al contrario, Delta ha impiegato circa un mese per raggiungere tale prevalenza. Il programma REACT-1 ha la capacità di dimostrare il valore dei programmi di sorveglianza della comunità man mano che l'immunità si sviluppa ed emergono nuove varianti. Mentre l'emergere della variante Delta è stato accompagnato da scene di ospedali sovraccaricati in India, l'identificazione iniziale di Omicron in Sud Africa non è stata seguita dallo stesso grave impatto delle onde precedenti. Di fronte a una nuova variante,

pertanto è fondamentale comprenderne la capacità di diffondersi e causare malattie. Le prime analisi dei dati clinici in Sud Africa hanno suggerito che Omicron fosse associato a un minor rischio di ospedalizzazione rispetto ad altre varianti circolanti di SARS-CoV-2, utilizzando un sistema di sorveglianza che collega le cartelle cliniche e ospedaliere con i dati del genoma

Anche così, è sempre più difficile estrapolare conclusioni sulle varianti emergenti a causa delle variazioni dell'immunità della popolazione dovute a infezioni naturali e vaccinazioni nei diversi paesi. Quattro varianti principali sono ad oggi circolate in Sud Africa, causando grandi epidemie, il che significa che esistono diverse combinazioni multiple di infezione e vaccinazione precedenti che gli individui potrebbero aver avuto. Al contrario, una percentuale maggiore di immunità in Europa e negli Stati Uniti è derivata dalla vaccinazione, complicando paradossalmente la misura in cui i ricercatori possono ottenere informazioni generalizzabili. Tuttavia i sistemi di sorveglianza basati sui sintomi, come quelli utilizzati per analizzare le prime tendenze in Sud Africa, potrebbero non riflettere infezioni lievi o asintomatiche nella comunità. Durante tutta la pandemia, sono stati necessari dati aggiuntivi per comprendere la vera relazione tra la malattia osservata e i livelli di infezione di fondo, evidenziando il valore di test comunitari ripetuti sistematici che campionano gli individui indipendentemente dai sintomi. Il loro quadro di indagini prospettiche di struttura simile ha consentito il confronto con le tendenze specifiche per età osservate nelle ondate precedenti e prima della disponibilità della vaccinazione in questi gruppi di età. Si è potuto così dedurre che con l'aumento della copertura vaccinale tra gli adolescenti di età compresa tra 12 e 17 anni, il carico di infezione per Delta si è spostato sui gruppi di età più giovani non vaccinati. Inoltre, la spinta a garantire che i gruppi di età più anziani ricevessero dosi di richiamo è stata seguita anche da una riduzione dei livelli di infezione tra le persone anziane.

Il team di Paul Elliot dell Imperial College di Londra ha invitato più di 800.000 persone a partecipare tra la fine di novembre e l'inizio di dicembre 2021, con circa 97.000 campioni raccolti e analizzati, di cui 1192 positivi. Questo sforzo consente la quantificazione delle dinamiche di trasmissione della comunità, anche tra individui infetti in modo asintomatico, e sono di fondamentale importanza per comprendere correttamente le dinamiche alla base di SARS-CoV-2, in particolare durante i periodi di improvvisa crescita o declino epidemico. Tali studi diventeranno ancora più preziosi man mano che la relazione tra infezione e malattia continua a cambiare: i paesi con livelli crescenti di immunità contro SARS-CoV-2 avranno una proporzione crescente di infezioni che rappresentano una svolta o reinfezione del vaccino, con una frazione decrescente di infezioni avere sintomi o malattie gravi. Il tasso di crescita della popolazione di una nuova variante può essere influenzato da molteplici fattori: la capacità del virus di diffondersi tra gli ospiti, le interazioni sociali dei gruppi all'interno dei quali si sta diffondendo e la misura in cui la vaccinazione o l'infezione precedenti protegge dal nuovo variante. Distinguere e differenziare questi fattori continuerà a essere essenziale nello sviluppo di politiche efficaci per la protezione dalle varianti SARS-CoV-2. L'incertezza sul futuro percorso evolutivo di SARS-CoV-2 significa che sarà fondamentale utilizzare la triangolazione di più linee di evidenza durante l'analisi delle caratteristiche delle varianti. In REACT-1, tuttavia, è stato osservato che due dei tre campioni sequenziati a Londra all'inizio di maggio 2021 erano Delta, a sostegno dell'ipotesi di un'ampia diffusione nella comunità ***S. Riley et al., "REACT-1 round 11 report: Low prevalence of SARS-CoV-2 infection in the community prior to the third step of the English roadmap out of lockdown," Working paper, Imperial College London (2021).***

Sia per Omicron che per le varianti precedenti, tali dati hanno fornito informazioni preliminari sulla distribuzione geografica della diffusione, in alcuni casi con infezioni importate provenienti da paesi che non hanno ancora segnalato la variante a livello nazionale. Oltre a supportare la risposta al COVID-19 e alle future pandemie, studi di comunità attentamente progettati potranno aiutarci a "ripensare" alle infezioni familiari. Ad esempio....

(Per continuare vai all'originale)

Un anno fa... Baedeker/Replay del 28marzo 2021

L'incerto confine tra "lieve" e "grave": i markers di laboratorio negli adulti e nei bambini.

Parte Prima: gli adulti