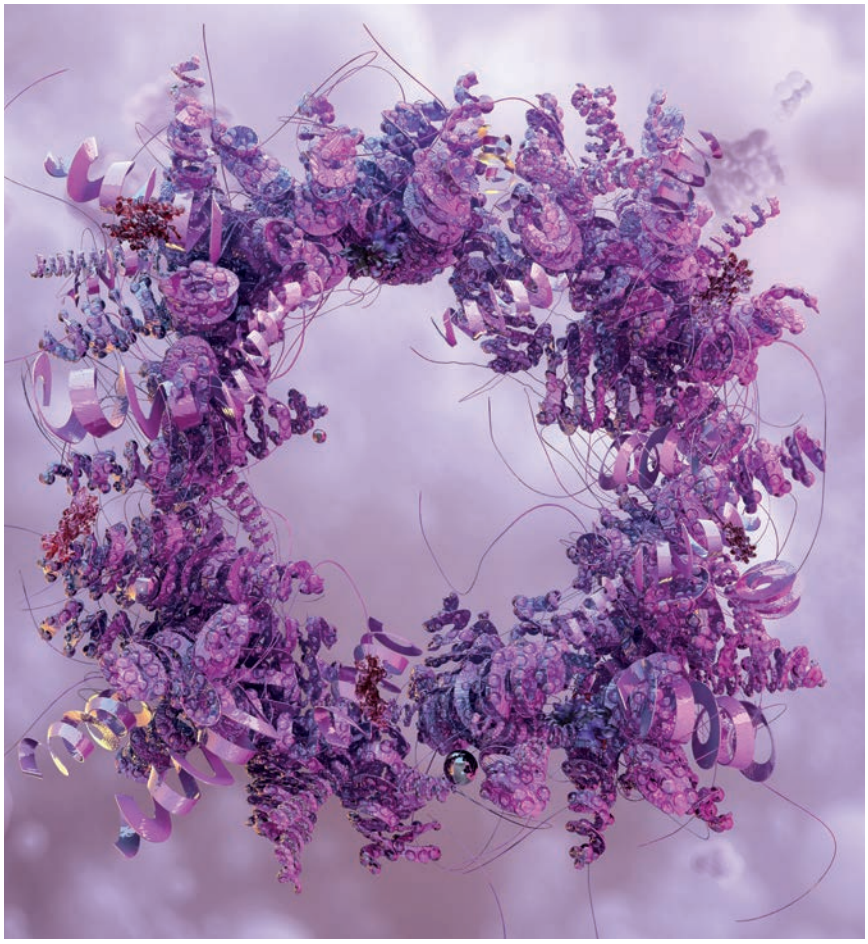
LA REVOLUCIÓN DE LA INTELIGENCIA ARTIFICIAL EN LA BIOLOGÍA ESTRUCTURAL DE PROTEÍNAS

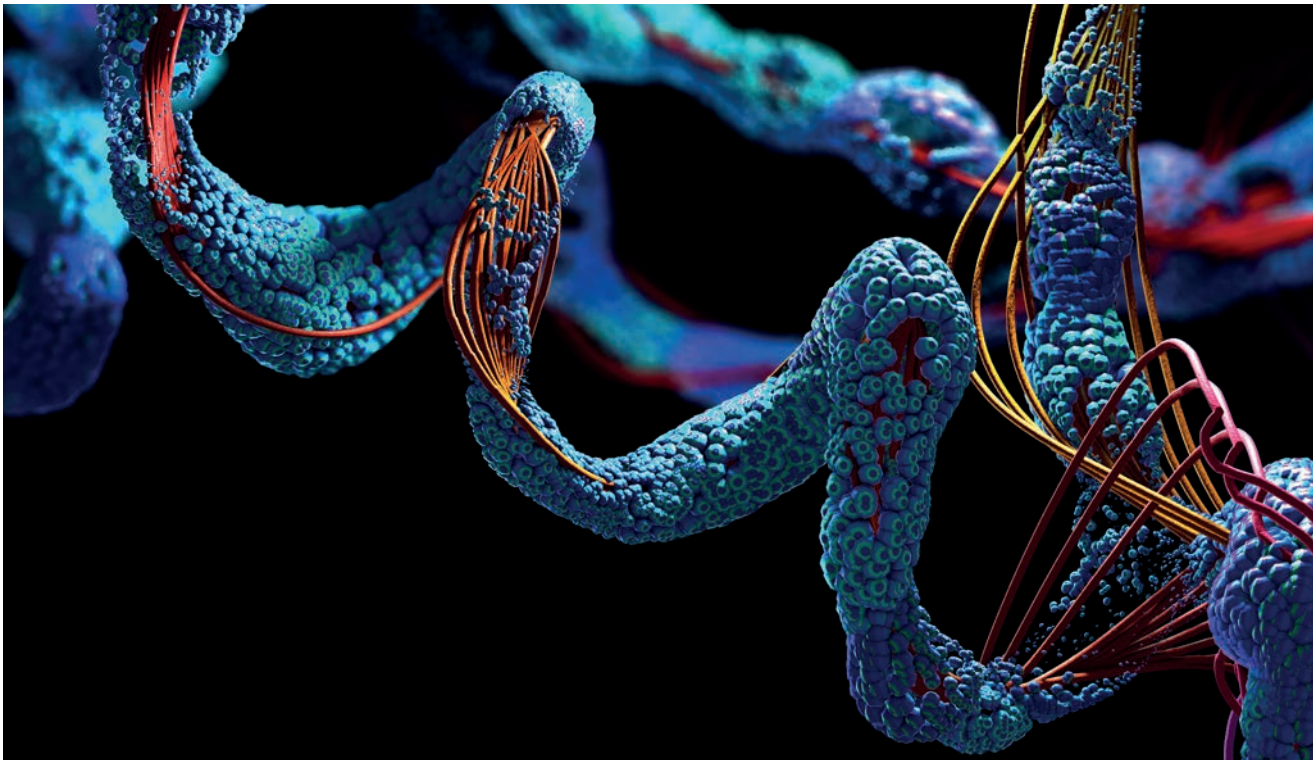
Pablo Chacón Montes
Instituto de Química Física Blas Cabrera (IQF 'Blas Cabrera'), CSIC, Madrid
Dpto. Química Física Biológica
Grupo de Bioinformática Estructural

La Inteligencia artificial (IA) ha revolucionado la biología estructural en tiempo récord. La capacidad de predecir con cierta precisión la disposición espacial de una proteína a partir de su secuencia nos permite explorar el espacio estructural de formas antes inimaginables, acelerar su caracterización experimental, y abrir nuevas ventanas para el desarrollo de herramientas computacionales capaces de diseñar proteínas que van más allá de lo explorado por la evolución natural. Las IA generativas, como los modelos de lenguaje y los procesos de difusión, han demostrado su capacidad para generar nuevas proteínas con propiedades específicas, alcanzando un éxito experimental notable. En este momento disruptivo y dinámico, en el que surgen a diario nuevas herramientas y aplicaciones, este artículo se centra en la estructura de proteínas, y excluye otros campos relacionados como el diseño de fármacos y ácidos nucleicos, donde la IA también es uno de los principales motores de avance.

AlphaFold es un ejemplo más de cómo la IA está transformado nuestro modo de vida, en este caso, revolucionando la biología estructural, un campo clave en el avance de la biología, la biotecnología y la investigación biomédica durante las últimas cinco décadas. En diciembre de 2020, AlphaFold 2 (AF2), desarrollado por Google-DeepMind, resolvió uno de los mayores desafíos de la biología: la predicción de la estructura tridimensional de las proteínas a partir de su secuencia de aminoácidos con una precisión comparable, en muchos casos, a la de las técnicas experimentales. Utilizando redes neuronales profundas, esta herramienta emplea la información de las estructuras conocidas y la co-evolución recogida en sus secuencias homólogas para predecir la estructura 3D de las proteínas. Seis meses después, la herramienta fue publicada en código abierto y permitió a la comunidad científica validar su eficacia (y limitaciones), provocando un impacto general y profundo (p. ej. la publicación original acumula más de 18.000 referencias) que ha culminado con la concesión en tiempo récord del Nobel en química para sus principales autores Demis Hassabis y John M. Jumper, junto a David Baker el gran pionero en modelado y diseño de proteínas que también tiene su propio desarrollo similar a AF2, RoseTTAFold. El uso de estas herramientas ha permitido expandir tres órdenes de magnitud la información estructural que disponemos. La base de datos AlphaFold DB (https://alphafold.ebi.ac.uk) ya almacena más de 214 millones de estructuras predichas por AF2, incluyendo el proteoma humano y de otras 47 especies, en contraposición con las 200.000 estructuras resueltas experimentalmente almacenadas en el Protein Data Bank (PDB). Al tiempo que intentamos racionalizar esta ingente información estructural, su potencial es evidente como también que se nos presentan grandes retos que afrontar. Por ejemplo, para la identificación de proteínas estructuralmente similares en una base de datos de estas dimensiones se ha tenido que desarrollar una herramienta de búsqueda ultra-rápida (*Foldseek*) de nuevo con elementos de IA ante la imposibilidad de aplicar métodos tradicionales. Es importante resaltar que la capacidad de generar buenas predicciones permite acelerar la propia determinación experimental de estructuras. Podemos utilizar las predicciones para generar construcciones más estables eliminado zonas potencialmente desestructuradas, para resolver la estructura más rápidamente por reemplazo molecular en cristalografía de rayos-X, y, de forma similar, en crio-microscopía electrónica su localización dentro de los mapas de densidad facilita su interpretación a resolución atómica. Como pescadilla que se muerde la cola, la capacidad de aprendizaje de las herramientas de IA y su precisión, aumentará al disponer de más estructuras experimentales.



Sin duda, AF2 y RF se han convertido en herramientas muy populares al permitir el acceso a información estructural de las proteínas en minutos a partir de su secuencia. Si no lo ha hecho ya, el lector puede comprobar lo fácil que es obtener una predicción, bien buscándola directamente en AlphaFold DB o bien en la nube donde podrá encontrar implementaciones optimizadas de AF2 como Colabfold, o RoseTTAFold. Además de la predicción 3D de la estructura, por cada amino ácido se obtiene una estimación de la fiabilidad en su predicción, llamada pLDDT, que nos indica qué partes



de la estructura pueden ser fiables y cuáles no. Valores de pLDDT>90 corresponden a regiones predichas con precisión, valores entre 90 y 70 suelen corresponder a una predicción donde la cadena principal está bien modelada pero las cadenas laterales pueden contener algún error, y valores pLDDT<50 caracterizan zonas cuya predicción no es fiable, y suelen corresponder con regiones desordenadas. Cuando visualizamos una de estas predicciones, solemos ver que la mayoría son regiones con alta fiabilidad, más o menos salpicada con regiones de menor precisión muchas veces localizadas en la superficie en regiones flexibles como bucles. También son frecuentes zonas de baja precisión en el N y C terminales en forma de espagueti, que son particularmente extensas en organismos eucariotas más evolucionados. En general, los modelos de AlphaFold o RoseTTAFold son muy buenos, pero no siempre lo son, e incluso las regiones de alta fiabilidad pueden contener errores. En la red se pueden encontrar un gran número de predicciones erróneas, que no invalidan la bondad y el carácter generalista de estas aproximaciones, pero que deben encender una luz amarilla de precaución sobre todo cuando queramos utilizar esas predicciones, por ejemplo en simulaciones atomísticas.

Aunque la entrada principal para la red neuronal de AlphaFold2 es una secuencia, en el primer paso se genera un alineamiento múltiple (MSA) buscando en distintas bases de datos secuencias similares a la de entrada. Precisamente, la calidad del alineamiento es uno de los factores determinantes para obtener una predicción más o menos precisa de la estructura. Un MSA diverso y profundo, con cientos o miles de secuencias alineadas, ayudará a identificar señales coevolutivas y utilizarlas para averiguar la correcta estructura 3D de la proteína. Por el contrario, un MSA poco profundo con pocas secuencias y baja variabilidad entre ellas, aumentará la posibilidad de errores en la predicción por la pobre señal coevolutiva. Otras limitaciones incluyen la dificultad de modelar cambios conformacionales e interacciones proteína-proteína.

Potentes modelos de IA han cambiado el panorama de la comprensión del lenguaje, la maravillosa generación de lenguaje de ChatGPT no deja de asombrarnos. Precisamente, el éxito de los modelos de lenguaje de gran tamaño (LLM por sus siglas en inglés) en procesamiento del lenguaje natural (PLN) y la similitud entre nuestro lenguaje y el «lenguaje de las proteínas» motivaron el desarrollo de los modelos del lenguaje de las proteínas (PLM). Estos modelos, en lugar de aprender de las distribuciones de palabras/frase/textos, aprenden de las distribuciones de aminoácidos/secuencias/funciones. Los modelos PLM tratan las secuencias como datos de entrada de forma similar a como se trata el texto en modelos de PLN como GPT (*Generative Pre-trained Transformer*) o BERT (*Bidirectional Encoder Representations from Transformers*). El modelo de lenguaje ESM-2 desarrollado por

Meta y basado en BERT, entrenado con 250 millones de secuencias de proteínas, fue capaz de aglutinar y codificar a distintos niveles jerárquicos sus distintas propiedades (relaciones coevolutivas, propiedades bioquímicas y biofísicas de los aminoácidos, etc.) captando las complejas reglas que rigen la "gramática" de las proteínas. Tras el entrenamiento, este modelo de lenguaje puede transformar cualquier secuencia de proteínas en un vector que encapsula estas propiedades denominado *embeddings*.

Los *embeddings* son una forma de representar la información de un texto en vectores para que los modelos de aprendizaje automático puedan medir la similitud de distintos textos y documentos en función de su significado, y así poder establecer relaciones que faciliten búsquedas, clasificaciones, etc. En este caso, los *embeddings* de ESM-2 son vectores que contienen información relevante de la secuencia como la conservación evolutiva, relaciones funcionales, motivos estructurales, etc. Utilizando modelos de aprendizaje profundo sobre estos *embeddings* de proteínas, se han logrado resultados espectaculares en la predicción de contactos de largo alcance, efectos mutacionales, de función e incluso también en la predicción de la estructura tridimensional. Esta última aplicación, denominada ESMFold, aunque no llega a los niveles de precisión de AF2, es muchísimo más rápida, lo que permite predecir eficazmente un ingente número de proteínas sin grandes recursos y puede ser una alternativa a AF2 en casos donde el MSA sea limitado. Hace pocos meses se ha presentado una nueva versión, ESM3, que es el primer modelo generativo que razona simultáneamente sobre la secuencia, la estructura y la función de las proteínas. Mejora a su antecesor en muchos aspectos: su capacidad de predicción rivaliza con AF2, permite cierta especialización ya que está entrenado con diversos organismos y biomas, y permite diseñar proteínas con funcionalidades específicas. Como ejemplo ilustrativo de diseño con ESM3, han presentado una nueva proteína verde fluorescente (GFP) a partir de las estructuras y de los residuos críticos que definen su función. De los miles de diseños generados, se identificó uno con una similitud del 58% con la proteína fluorescente conocida más cercana con una fluorescencia similar a la observada en medusas y corales (Figura 1). Este resultado muestra la capacidad de ESM3 para explorar espacios proteicos que la naturaleza necesitaría millones de años de evolución en descubrirlos, y su potencial en la generación de nuevas proteínas funcionales.

En lugar de predecir cómo una secuencia se pliega, el aspecto crucial en el diseño *de novo* es resolver el problema inverso: diseñar una secuencia que, cuando se sintetice y pliegue, adopte la estructura específica con la función que deseamos. De los enfoques iniciales que estaban basados en modelos físicos (p. ej. Rossetta) hemos pasado al desarrollo de herramientas más potentes de aprendizaje profundo basadas en grafos (GNNs por sus siglas inglés). En este tipo de redes neuronales cada átomo o residuo del esqueleto proteico queda definido por un nodo que está conectado con los nodos cercanos a una distancia dada, que forman las aristas del grafo. En cada nodo de la red podemos añadir información como el tipo de aminoácido o las coordenadas tridimensionales, mientras que

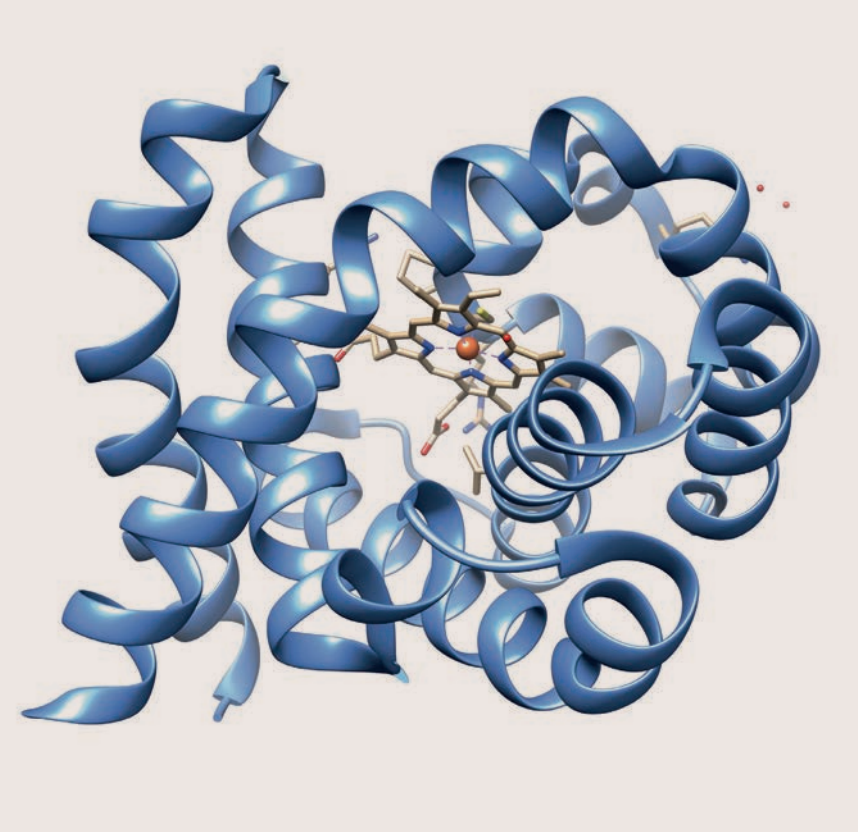


Figura 1

en las aristas se pueden incluir vectores que describen la distancia, dirección y orientación entre nodos conectados. Mediante el paso de mensajes, en el que cada nodo envía y recibe mensajes basados en las características de sus nodos y aristas vecinos, la red va agregando la información local y aprendiendo patrones complejos entre estructura y secuencia. ProteinMPNN es el paradigma de este tipo de redes, y su capacidad para predecir secuencias de aminoácidos compatibles con una determinada estructura ha sido validada experimentalmente en multitud de escenarios como el diseño de proteínas solubles, de nuevas proteínas, rediseño de estructuras y la generación de complejos simétricos. En su nueva versión, LigandMPNN, añade pequeños ligandos y ácidos nucleicos al proceso de diseño, y añade casos de éxito al diseño de proteínas que interaccionan con gran afinidad y especificidad con moléculas no proteicas.

Es importante resaltar que las validaciones experimentales de este tipo de métodos tienen un alto grado de éxito, entre el 10-20% de las secuencias predichas se obtiene la estructura deseada. Por esa razón no es de extrañar que el número y variedad de aplicaciones de diseño de proteínas crezca a pasos agigantados. Sin embargo, para su aplicación necesitamos una estructura de entrada. En un paso previo necesitamos definir un esqueleto o boceto estructural inicial y asegurarnos que sea "diseñable", *i.e.* que al menos una secuencia se pliegue en esa estructura. Aunque siempre podemos partir de una estructura conocida intentando mejorar sus propiedades (p. ej. estabilidad o solubilidad), o reciclarla para otros propósitos, de nuevo, la IA nos ofrece una solución con los modelos de difusión. Durante el entrenamiento de estos modelos generativos de aprendizaje profundo, primero se añade ruido gaussiano a las estructuras de proteínas conocidas para luego entrenar la red para recuperar las estructuras originales mediante un proceso iterativo de eliminación de ruido. Así, el modelo aprende las distribuciones de probabilidad del espacio de estructuras proteicas, lo que le permite, en el momento de la generación, recibir ruido gaussiano y transformarlo de forma iterativa en nuevos esqueletos proteicos cuya secuencia tendremos que diseñar.

RFDiffusion, aprovechando los pesos preentrenados de RoseTTAFold y las capacidades de ProteinMPNN para resolver el problema inverso, obtiene resultados realmente espectaculares. Con esta herramienta se han generado nuevas arquitecturas de barriles alfa-beta visibles mediante CryoEM que superan las variaciones estructurales del clásico pliegue de barril TIM –un diseño de proteínas para cinco dianas proteicas terapéuticas, con una tasa de éxito del 18% con 95 diseños con especificidad picomolar-. La nueva versión RFdiffusionAA, que utiliza RoseTTAFold All-Atom, la versión mejorada de RF, expande el rango de aplicabilidad al diseño de proteínas que unen pequeños ligandos. Por ejemplo, ha diseñado proteínas que se unen a bilina, hemo y digoxigenina con estructuras muy diferentes a las proteínas que unen estos compuestos que se encuentran en el PDB. Además, los modelos de difusión pueden incorporar restricciones en generación que nos permiten incorporar propiedades y funciones deseadas en el diseño.

Chroma, otro modelo de difusión, también puede condicionarse con texto lo que abre la posibilidad de que en algún futuro tengamos una herramienta a la que podamos decirle «Diseña una proteína pequeña, soluble y que se una a la proteína X». A pesar de estos increíbles avances queda mucho camino por recorrer, en particular, poder incorporar en el diseño la exploración del espacio conformacional y de interacciones, lo que nos permitiría crear proteínas que respondan a determinados estímulos o incluso generar nuevos mecanismos enzimáticos.

El impacto y el rango de aplicación de las herramientas basadas en IA (Figura 2) es impresionante. Muy recientemente, Google DeepMind ha publicado AlphaFold3 (AF3), que extiende su capacidad más allá de proteínas individuales con la predicción de sus interacciones con ácidos nucleicos, iones y pequeñas moléculas. A pesar de que reportan resultados muy prometedores que mejoran la capacidad y el rango predictivo de la versión anterior, de momento, AF3 sólo está accesible de forma limitada en un servidor web junto con unas draconianas licencias de uso. De igual forma ESM3, aunque el código sea accesible, prácticamente solo sus desarrolladores disponen de los recursos computacionales necesarios para su entrenamiento, e incluye una oscura licencia no comercial. Aunque estamos acostumbrados a que estas compañías controlen nuestra vida digital sabiendo donde estamos, lo que buscamos, compramos, etc., parece excesivo que vayan a controlar nuestra capacidad de investigación. Esperemos que la comunidad científica logre que recapaciten y vuelvan a la senda del código abierto, donde el Prof. D. Baker ha demostrado su activismo a lo largo de los años al permitir el acceso a sus herramientas de predicción y el diseño de proteínas, que desde aquí agradecemos.

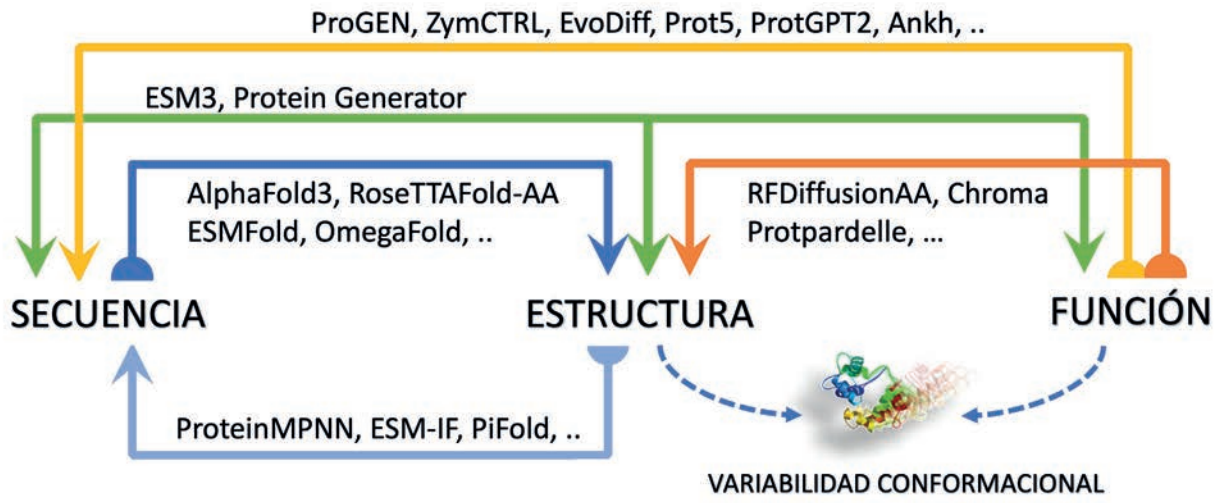


Figura 2

Para leer más

Abramson J, *et al.* "Accurate structure prediction of biomolecular interactions with AlphaFold3". *Nature* 630 (2024) 493–500. https://doi.org/10.1038/s41586-024-07487-w

Krishna R, *et al.* "Generalized biomolecular modeling and design with RoseTTAFold All-Atom". *Science* 384 (2024) 2528. DOI: 10.1126/science.adl2528

Varadi M, *et al.* "AlphaFold Protein Structure Database in 2024: Providing structure coverage for over 214 million protein sequences". *Nucleic Acids Research* 52 (2024) D368–D375. https://doi.org/10.1093/nar/gkad1011

Watson JL, *et al.* "De novo design of protein structure and function with RFdiffusion". *Nature* 620 (2023) 1089–1100. https://doi.org/10.1038/s41586-023-06415-8

Winnifrith A, *et al.* "Generative artificial intelligence for de novo protein design". *Current Opinion in Structural Biology* 86 (2024) 102794. https://doi.org/10.1016/j.sbi.2024.102794