Design de Proteínas e a Previsão da sua Estrutura – O Prémio Nobel da Química 2024, atribuído a David Baker, Demis Hassabis e John Jumper

O Prémio Nobel da Química 2024

Em 2024 o comité Nobel decidiu atribuir dois Prémios Nobel, o da Física e o da Química, a avanços científicos e tecnológicos dependentes da Inteligência Artificial e outros métodos computacionais. O que pode ser considerado, possivelmente, um sinal dos tempos. Enquanto o Prémio Nobel da Física, atribuído a John J. Hopfield (*Princeton University*) e a Geoffrey Hinton (*University of Toronto*), foi mesmo sobre redes neuronais, o Prémio Nobel da Química, atribuído a David Baker (*University of Washington*), Demis Hassabis e John Jumper (ambos da *Google DeepMind*), consagrou, em certa medida, o impacto da Inteligência Artificial e das redes neuronais nos enormes desenvolvimentos no campo no *design* de proteínas e na previsão da sua estrutura tridimensional.

David Baker | Demis Hassabis | John Jumper

Ill. Niklas Elmehed © Nobel Prize Outreach

Como se poderá ver em seguida, os desenvolvimentos que deram origem ao Prémio Nobel da Química 2024 estão muito relacionados. David Baker recebeu metade do Prémio Nobel pelos desenvolvimentos em *design* de proteínas, enquanto Demis Hassabis e John Jumper receberam a outra metade do mesmo Prémio pela previsão da estrutura de proteínas. Por uma razão de perspetiva histórica e de clareza conceptual, falar-se-á primeiro sobre a previsão da estrutura de proteínas e só depois sobre o *design* de proteínas.

Previsão de estrutura de proteínas

As proteínas, as máquinas da vida, têm, na maioria dos casos, uma estrutura predominante definida, que é determinante para a sua função. A busca desta estrutura tridimensional, ou seja, a organização espacial de todos os átomos que as constituem, é uma investigação de longa data. Esta pesquisa tem dois aspetos: (i) compreender como uma proteína sintetizada por um ribossoma adquire a sua forma, ou seja, qual o caminho que segue, desde uma cadeia de resíduos de aminoácidos até à sua forma final (ou mais provável), ou (ii) qual é a forma final da estrutura da proteína, independentemente do caminho para lá chegar. Poder-se ia pensar que os dois aspetos são semelhantes, mas o primeiro, que contém o segundo, é muito mais difícil.

A primeira proteína que teve a sua estrutura resolvida a uma resolução quase atómica (2 Å) foi a mioglobina (Figura 1), uma proteína composta por 153 resíduos de aminoácidos e contendo um grupo heme, que existe no citoplasma das células do músculo dos vertebrados e fixa oxigénio, necessário à respiração celular. Esta estrutura foi determinada por John Kendrew e colaboradores em 1960 [1], usando cristalografia por difração de raios-X e a sua determinação representou um feito espantoso para a época, garantindo a Kendrew o Nobel da Química de 1962, laureado conjuntamente com Max Perutz.

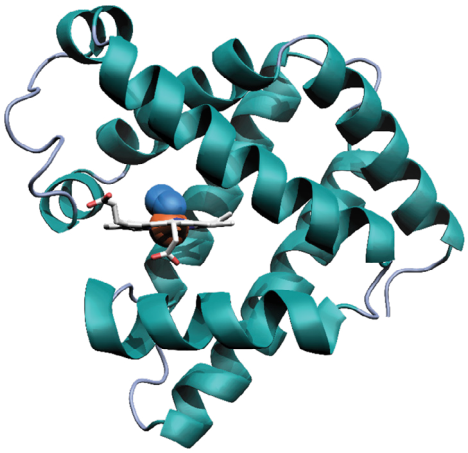


Figura 1 – Estrutura da mioglobina determinada usando cristalografia por difração de raios-X. Nesta figura molecular gerada usando o software PyMOL [2] a partir das coordenadas depositadas no PDB (código PDB: 1MBO [3]), a estrutura secundária da mioglobina está retratada usando uma representação denominada *cartoon* e colorida a verde, os átomos do grupo heme que coordenam o átomo de ferro estão representados usando *sticks* brancos, o átomo de ferro está representado usando uma esfera laranja e a molécula de oxigénio (O_2) está representada usando esferas azuis.

A determinação da estrutura da mioglobina por John Kendrew e colaboradores exigiu muitos anos de dedicação e foi um processo laborioso. Desde então as metodologias evoluíram e é possível determinar a estrutura de proteínas e outras biomoléculas usando cristalografia por difração de raios-X (*X-ray*), ressonância magnética nuclear (NMR) multidimensional e, mais recentemente, Crio-microscopia eletrónica (Cryo-EM). O repositório de estruturas experimentais de biomoléculas

(PDB – *Protein Data Bank*) tem hoje (30 de novembro de 2024) 227933 estruturas diferentes. Apesar de terem existido muitos avanços técnicos neste campo, determinar a estrutura de uma biomolécula continua a ser um esforço laborioso, requer conhecimentos especializados e instrumentação dispendiosa.

Assim, um dos grandes objetivos da Biologia Estrutural, a par com o desejo de desenvolver métodos experimentais cada vez mais eficientes na determinação da estrutura, sempre foi encontrar um meio para prever a estrutura tridimensional de proteínas (e outras biomoléculas) com base no conhecimento da sua sequência primária, ou seja, a sequência de resíduos de aminoácidos que a constituem.

As primeiras tentativas de prever a estrutura de proteínas vieram da aplicação do campo da Mecânica Molecular, usando funções de energia potencial empíricas. Estas metodologias, que eram inicialmente usadas para estudar a conformação de moléculas pequenas, começaram a ser aplicadas a pequenas proteínas, sendo de salientar os trabalhos pioneiros de Gibson & Sheraga 1967 [4], Levitt & Warshel [5], que usaram grandes aproximações para permitir estudar sistemas tão complexos. No entanto, o problema revelou-se muito desafiante, mesmo com o surgimento de computadores cada vez mais rápidos.

Os impressionantes avanços computacionais (quer ao nível da capacidade de cálculo do *hardware*, quer ao nível da eficiência do *software*) permitiram idealizar que seria possível simular as alterações conformacionais das cadeias polipeptídicas a partir de uma conformação não enrolada, com o objetivo de se reconstruir o processo de adquirir estrutura enrolada usando métodos de mecânica molecular, em particular a simulação de dinâmica molecular. Este objetivo esteve na cabeça de inúmeros investigadores, mas os primeiros que conseguiram simular este processo complexo de uma maneira verdadeiramente global foram Duan & Kolmann em 1998 [6] que, usando um supercomputador, simularam a *Villin headpiece subdomain*, um polipéptido de 36 resíduos de aminoácidos, durante 1 μs ($1x10^{-6}$ s). Este tempo de simulação para uma proteína em solvente explicito (aproximadamente 3000 moléculas de água) foi algo inaudito na altura e só por si consistiu num feito tecnológico. No entanto, embora se tenha observado a formação de estrutura secundária, a conformação final da proteína era consideravelmente diferente da estrutura obtida por NMR. Este feito suscitou grande entusiasmo na comunidade científica (incluindo dos autores), mas os recursos computacionais necessários para atingir este resultado (para não falar das longas simulações que seriam necessárias para se observar o *folding* de proteínas um pouco maiores ou mais complexas), estavam e estão apenas ao alcance de poucos. Muitos estudos foram feitos, até que um competidor inusitado entrou nesta corrida: David E. Shaw, um cientista computacional, que se transformou em financeiro computacional (e em bilionário) e mais tarde voltou à ciência, desta vez como Bioquímico Computacional. Em 2010, Shaw e os seus colaboradores (incluindo um dos premiados com o Nobel deste ano, John Jumper) publicaram um artigo em que reportaram o *folding* de diversas proteínas, usando simulações muito longas para a época, alcançando a escala do milissegundo ($1x10^{-3}$ s) [7]. Estes estudos basearam-se num avanço tecnológico notável, a construção de um computador, denominado Anton [8], construído especificamente para fazer simulações de dinâmica molecular (e não sendo de uso geral, como a maioria dos computadores). David E. Shaw usou a sua vasta fortuna para construir o Anton e criou um instituto de investigação para fazer estes estudos, supervisionado e financiado por ele, o D. E. Shaw Research, LLC. Não há muitos *Anton* no planeta e o comum dos mortais não tem acesso a estas máquinas, mas é uma questão de esperar que os computadores se tornem mais rápidos para que esta metodologia, que não só prevê estrutura, mas também o caminho de *folding*, se possa generalizar. De qualquer forma, este sucesso mostra que, contrariamente ao que é veiculado, conhecemos os fatores que determinam o *fold* de proteínas sem alterações pós-traducionais: são a sequência e a física molecular introduzida nas funções de energia potencial. O fator limitante é unicamente computacional.

Dado que prever a estrutura de uma proteína não passa necessariamente por conhecer o caminho através do qual esta estrutura é adquirida, desde cedo surgiram estratégias alternativas para inferir a estrutura de uma proteína a partir da sua sequência. Uma das possibilidades exploradas para "construir" a estrutura de uma proteína com base noutra foi tirar partido da homologia entre sequências de proteínas conhecidas. Tal foi feito pela primeira vez por Browne e colaboradores [9], que previram a estrutura da alfa-lactalbumina a partir da estrutura experimental da lisozima, com base em modelos físicos (construídos com arames, literalmente). Este trabalho abriu o caminho para o campo da modelação por homologia (ou modelação comparativa), que, mais tarde, se tornou computacional. Nesta metodologia, usa-se a estrutura de uma (ou várias) proteínas, para prever a estrutura de outra proteína homóloga, considerando as estruturas experimentais, o alinhamento de sequências e mecânica molecular. Vários investigadores dedicaram-se a desenvolver *software* e métodos para fazer modelação por homologia, mas os mais sucedidos e continuados esforços foram feitos no laboratório de Tom L. Blundell, inicialmente por Mike Sutcliffe [10] (*Composer*) e depois por Andrei Sali [11], que escreveu o mais difundido e sólido *software*: *Modeller*. Durante longos anos, a modelação por homologia foi a metodologia rainha na previsão de estrutura de proteínas, originando grandes avanços no seu conhecimento estrutural, uma vez que

muitas sequências homólogas de proteínas, facilmente obtidas por sequenciação genómica, podiam dar origem a previsões da estrutura tridimensional dessas proteínas, desde que houvesse no PDB uma estrutura com uma suficiente identidade sequencial (30-40%) para servir de "molde". É especulado que o PDB contém já a maioria das formas naturais possíveis das proteínas, pelo que, para a maioria dos casos, a aplicação de modelação por homologia era possível, muito embora muitos modelos pudessem ter erros significativos e, em muitos casos, se observassem desvios significativos das estruturas modeladas quando estas eram comparadas com a estrutura da mesma proteína determinada experimentalmente. Este problema é tanto mais severo, quanto menor a homologia entre a proteína "alvo" e a proteína "molde". Para piorar a situação, há numerosos casos em que nenhum homólogo da proteína cuja estrutura se quer prever se encontra no PDB. O que fazer nestes casos?

Para estes casos restavam as metodologias de previsão denominadas *ab initio*, onde nenhuma informação experimental explicita era utilizada. As metodologias de Mecânica Molecular, já referidas acima, eram muito utilizadas na previsão *ab initio*, mas o campo evoluía lentamente, embora David Baker, um dos destinatários do Prémio Nobel de 2024, fosse uma das estrelas do CASP (*Critical Assessment of Protein Structure Prediction*), a competição de previsão de estrutura de proteínas criada por John Moult e colaboradores em 1994 [12]. O laboratório de David Baker criou o *software* ROSETTA [13] que, entre outras funcionalidades, previa estruturas de proteínas com base em fragmentos.

Devemos fazer aqui um parêntesis para referir um outro fator determinante para as previsões de estruturas modernas: os alinhamentos múltiplos de sequências (MSA – *Multiple Sequence Alignments*). Um alinhamento de múltiplas sequências de proteínas homólogas permite identificar sinais de correlação de mutações, que nos contam a história evolutiva das proteínas. Se a mutação de dois resíduos de aminoácido ocorre de forma correlacionada, é possível (mas não é determinístico) que esses dois resíduos estejam em contacto na estrutura das proteínas da família analisada. Esta correlação foi formalizada inicialmente por Chris Sander e colaboradores [14], e tornou-se um fator importante naquilo que veio muito mais tarde. E o que veio a seguir dependeu também da enorme informação que hoje temos sobre sequências primárias de proteínas, resultante da sequenciação massiva (e pouco dispendiosa) de genomas e meta-genomas.

Agora que já conhecemos a paisagem complexa da previsão da estrutura de proteínas e a forma como esta paisagem evolui ao longo do tempo, é o tempo para fazermos um *fast-forward* para parte do Prémio Nobel de 2024: os desenvolvimentos que levaram a grandes avanços nas metodologias de previsão de estrutura. No CASP13, em 2018, apareceu um grupo da DeepMind, empresa subsidiária da Google, que apresentou um *software*, baseado em redes neuronais que, para além de bater toda a competição, atingiu quase uma *Global Distance Test* de 60 (ver Figura 2) - muito aumentada em relação à mais bem obtida no CASP12, de cerca de 40. Este *software* foi denominado AlphaFold e mostrou as potencialidades da inteligência artificial e do *deep learning* na previsão da estrutura de proteínas [15]. No CASP14, em 2020, uma versão redesenhada da rede neuronal original, designada AlphaFold2 [16], bateu toda a competição, alcançando um valor de cerca de 90 GDT (Figura 2), comparável com métodos experimentais de previsão de estrutura de proteínas. Parte do segredo desta nova versão do AlphaFold foi ter tirado partido dos novos desenvolvimentos no campo da inteligência artificial e de uma nova abordagem conhecida como "mecanismos de atenção" que é muito mais poderosa do que as anteriores a encontrar relações globais e não apenas relações mais locais entre os elementos de um sistema. No caso da previsão de estrutura de proteínas, este fator é muito importante, porque a estrutura final depende de um conjunto enorme de interações entre resíduos, alguns dos quais estão muito distantes em termos de sequência primária. O mundo da biologia estrutural mudou nesse dia e foi isso que foi reconhecido em metade do Prémio Nobel da Química de 2024, consagrando Demis Hassabis e John Jumper.

Os modelos AlphaFold (já estamos neste momento no AlphaFold3 [17]), são baseados em *deep neuronal networks* treinadas nas estruturas existentes no PDB e também usam para a previsão a informação de correlações obtidas no MSA de sequências homólogas

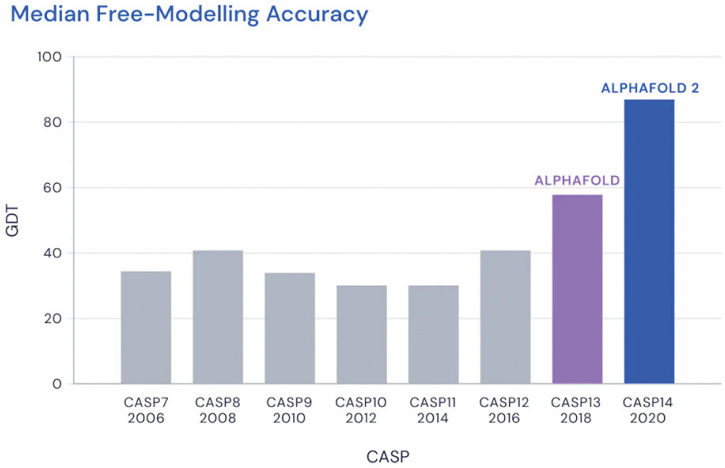


Figura 2 - Evolução da medida *Global Distance Test* (GDT) ao longo de sucessivas edições do concurso CASP. O valor de GDT, que varia de 0 a 100, é uma medida da percentagem de resíduos de aminoácido que estão abaixo de um valor limite de distância em relação à sua posição correta, sendo um valor de 90 indicativo de que a previsão é comparável aos resultados obtidos por métodos experimentais. As barras correspondem ao valor médio obtido pela equipa vencedora de cada edição considerando os cinco casos em que obtiveram as melhores previsões. A barra a roxo evidencia o aumento substancial obtido pela primeira versão do AlphaFold e a barra a azul mostra os resultados impressionantes obtidos pela segunda versão do AlphaFold – AlphaFold2 (extraído de deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology).

da proteína que se quer prever (Figura 3), sendo possível o uso de estruturas experimentais de proteínas homólogas para aumentar a confiança do modelo.

Agora, é possível prever a estrutura de qualquer proteína, seja homóloga ou não com uma proteína cuja estrutura tenha sido determinada anteriormente. Um exemplo encontra-se na Figura 4, onde a proteína Rv2102 de *Mycobacterium tuberculosis*, que não tem homologia sequencial com nenhuma proteína de estrutura experimentalmente conhecida, é prevista pelo AlphaFold3.

Os modelos AlphaFold não só permitem a previsão de estruturas de proteínas, mas também permitem obter uma medida, global e local, de qualidade dessas estruturas. O caso da proteína Rv2102 dá-nos uma imagem clara disso mesmo, já que algumas regiões (coloridas a azul) apresentam uma grande confiança, enquanto outras regiões (coloridas a amarelo e a laranja) apresentam baixa confiança. Esta baixa confiança pode ser consequência de limitações do modelo (não nos esqueçamos que o modelo é treinado em dados conhecidos e que depende parcialmente do MSA) ou corresponder a zonas desordenadas da proteína. Ainda há muito caminho a percorrer para se conhecerem todas as características das proteínas, mas o primeiro passo determinante foi dado.

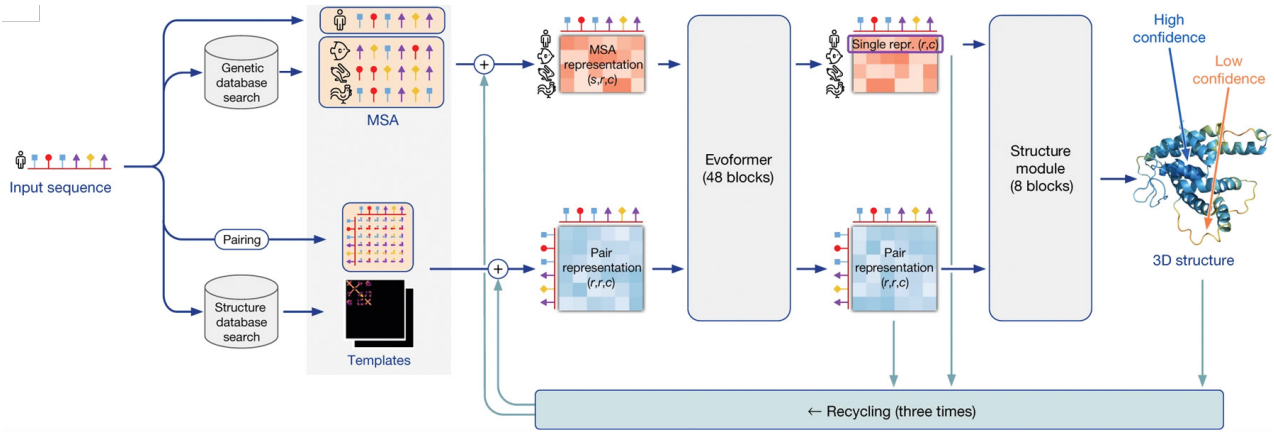


Figura 3 - Arquitetura do *software* AlphaFold 2. Este *software* recebe como *input* a sequência de uma proteína cuja estrutura se pretende determinar e a partir desta gera um alinhamento múltiplo de sequências (MSA). O *software* usa então dois tipos de representação: o MSA e uma "representação de pares", podendo também usar informação contida em estruturas conhecidas de proteínas homólogas. Seguidamente, o módulo principal do *software*, designado "Evoformer", usa uma rede neuronal profunda para refinar informação do MSA e da representação de pares. O módulo final, chamado "Structure Module", mapeia as representações refinadas do MSA e de pares em coordenadas tridimensionais da proteína. O processo é iterado várias vezes, de forma a diminuir o ruído e aumentar a confiança da estrutura prevista. Esta figura foi reproduzida da referência [16].

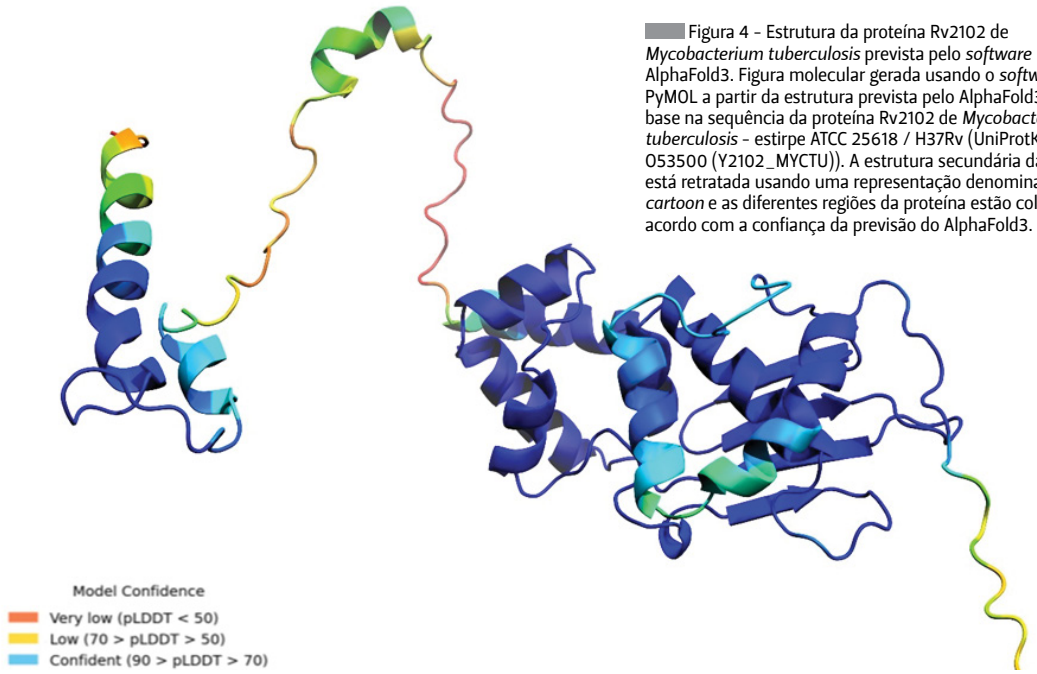


Figura 4 - Estrutura da proteína Rv2102 de *Mycobacterium tuberculosis* prevista pelo *software* AlphaFold3. Figura molecular gerada usando o *software* PyMOL a partir da estrutura prevista pelo AlphaFold3 com base na sequência da proteína Rv2102 de *Mycobacterium tuberculosis* - estirpe ATCC 25618 / H37Rv (UniProtKB - O53500 (Y2102_MYCTU)). A estrutura secundária da proteína está retratada usando uma representação denominada *cartoon* e as diferentes regiões da proteína estão coloridas de acordo com a confiança da previsão do AlphaFold3.

Design de proteínas

A outra face do prémio Nobel da Química de 2024 é também a outra face do problema descrito acima. Se a previsão de estrutura tem como objetivo determinar a estrutura de uma proteína a partir da sequência, a outra face desta moeda, conhecida como desenho de proteínas, tem como objetivo criar uma sequência de resíduos de aminoácido que seja capaz de adotar uma estrutura desejada e realizar uma função pretendida. Criar novas proteínas para tratar doenças como o cancro, defender-nos de vírus e bactérias ou degradar poluentes como o plástico, parece ficção, mas é hoje possível graças a esta área da biologia estrutural.

David Baker recebeu parte do Prémio Nobel da Química pelos avanços em desenho de proteínas, dado que teve o mais sólido e continuado trabalho no campo (assim como referido acima, na previsão de estrutura de proteínas). Mas não foi o pioneiro. O primeiro esforço consistente reportado neste campo foi feito por Regan & DeGrado em 1988 [18], quando desenharam e produziram experimentalmente uma proteína não natural com quatro hélices alfa. Embora um feito para a altura, o desenho desta proteína não foi propriamente computacional. O primeiro esforço computacional foi feito Dahiyat & Mayo em 1997 [19], que desenharam uma proteína com a forma de um *zinc finger*, mas sem os resíduos de aminoácidos para ligar o ião zinco, nem o zinco (Figura 5). De notar que não existem proteínas naturais conhecidas que adotem este *fold* na ausência de um átomo de zinco que estabilize a estrutura, tendo sido, por isso, um feito criar uma proteína que adota esta arquitetura sem necessitar da ligação ao zinco.

Dahiyat & Mayo fixaram a cadeia principal do *zinc finger* e pesquisaram combinações de resíduos e conformações de cadeias laterais para otimizar a rede de interações moleculares, usando modelos físicos para classificar a qualidade de cada uma das proteínas geradas, proibindo a inclusão de resíduos de cisteína e histidina nos locais que coordenam o ião zinco. A proteína desenhada foi sintetizada e a sua estrutura foi determinada por NMR, evidenciando uma forma muito semelhante ao do *zinc finger* original.

Mas foi David Baker, recipiente de parte do Prémio Nobel da Química de 2024, que deu os passos mais arrojados e consistentes no campo do desenho de proteínas. Em 2003, Baker e colaboradores publicaram uma estrutura com uma topologia não vista na Natureza [20], a proteína Top7 (Figura 6), desenhada pelo *software* ROSETTA. Quando a estrutura desta proteína foi determinada experimentalmente, verificou-se que era muito similar ao modelo previsto e otimizado pelos métodos de Mecânica Molecular implementados no ROSETTA e pelo resultado da pesquisa massiva de sequências executada por este *software*.

Mais recentemente, os avanços no campo da inteli-

Figura 5 - Comparação da estruturas experimentais de um domínio *zinc finger* nativo com um domínio desenhado para adquirir o mesmo *fold* na ausência de zinco. A figura mostra o segundo domínio de *zinc finger* da proteína (PDB 1zaa) – em cima, e uma proteína desenhada por Dahiyat e Mayo para adquirir a mesma forma, mas que não contém os resíduos de aminoácidos necessários para ligar o ião zinco (PDB 1fsd) – em baixo. As figuras moleculares foram geradas usando o *software* PyMOL, sendo a estrutura secundária das proteínas retratada usando uma representação denominada *cartoon* colorida a laranja. As cadeias laterais dos resíduos que compõem as proteínas estão representadas usando *sticks*, com os átomos de carbono, oxigénio e nitrogénio coloridos a laranja, vermelho e azul, respetivamente.

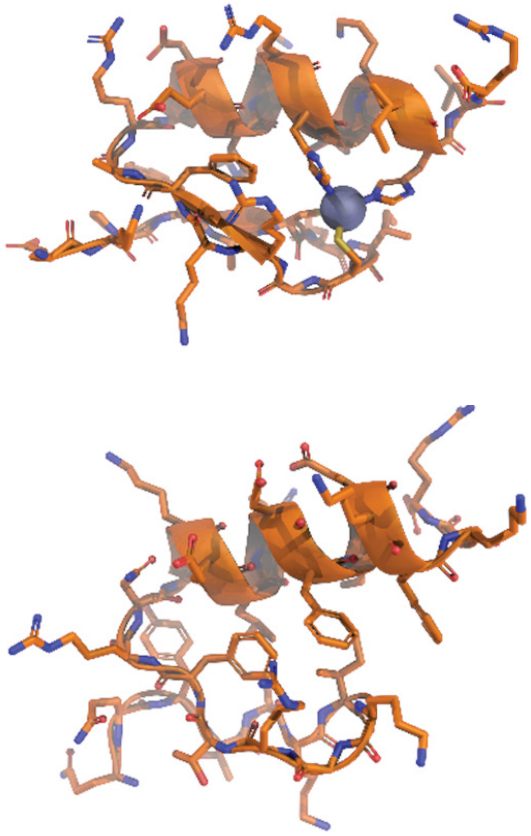
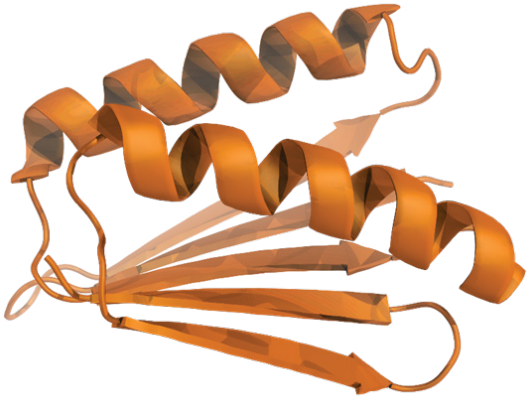


Figura 6 - Estrutura da proteína Top7 desenhada pelo laboratório de David Baker usando o *software* ROSETTA. A figura molecular foi gerada usando o *software* PyMOL a partir da estrutura experimental depositada no PDB (PDB 1QYS), sendo a estrutura secundária retratada usando uma representação denominada *cartoon* colorida a laranja.



gência artificial (IA) também mudaram a forma como se faz desenho de proteínas. Entre os métodos de IA mais usados nesta área, encontram-se duas ferramentas desenvolvidas no laboratório de David Baker: *RF diffusion* [21] e *Protein MPNN* [22]. Estas duas ferramentas são complementares entre si e são usadas em diferentes etapas do processo de *design*. Numa primeira etapa, o *RF diffusion* é usado para gerar um esqueleto da proteína que tenha a estrutura desejada. Se o objetivo for desenhar uma proteína que possa encaixar-se bem numa outra proteína alvo, esta ferramenta irá gerar uma estrutura tridimensional que tenha um bom encaixe geométrico com o alvo. Para isso, usa uma abordagem semelhante à que é usada para criar imagens em servidores como o DALL-E, mas, neste caso, usa-se esta abordagem para gerar o esqueleto da estrutura tridimensional de uma proteína. O *RF diffusion* é um modelo generativo (gera estruturas) baseado numa técnica que é conhecida como difusão. O processo, que quase parece magia, começa por gerar uma forma tridimensional muito mal definida e cheia de ruído, que se vai moldando numa proteína com a forma pretendida à medida que o ruído é progressivamente removido usando uma rede neuronal muito poderosa, que foi treinada com estruturas reais depositadas no PDB. No final deste processo obtém-se uma proteína que tem a forma ideal para o propósito que temos em mente, que pode ser, por exemplo, ligar-se a uma proteína alvo.

Um ponto interessante desta metodologia é que a solução que obtemos não é ainda a solução final, porque o *RF diffusion* gera apenas o esqueleto da nossa proteína, colocando resíduos de glicina em todas as posições da sequência. Podemos fazer uma analogia com a construção de uma casa: temos os alicerces, mas precisamos de construir as paredes e todo o recheio, ou seja, precisamos de preencher todas as posições da nossa sequência. Para isso, precisamos de um método robusto que consiga prever quais são os resíduos mais adequados para cada posição. No final de contas, precisamos de desenhar a sequência certa que dará origem à estrutura tridimensional que acabámos de gerar no passo anterior. Um dos métodos mais usados para este fim é o *Protein MPNN*, que usa uma rede neuronal profunda, dividida em dois módulos principais. O primeiro módulo, chamado "codificador da estrutura do esqueleto da proteína", codifica a estrutura que recebeu como *input* e transforma-a numa espécie de mapa que contém informação sobre como os resíduos da proteína interagem uns com os outros. O segundo módulo, chamado "descodificador da sequência" é responsável por gerar uma sequência compatível com a estrutura do esqueleto da proteína que foi codificada pelo primeiro módulo. Para cada posição, a rede neuronal prevê as probabilidades para cada um dos 20 resíduos de aminoácido possíveis e, com base nas probabilidades previstas, seleciona um resíduo para uma dada posição. Isto é feito de forma iterativa e aleatória (a ordem de previsão ao longo da cadeia não é fixa) e o processo só termina quando toda a sequência está preenchida. Uma vez que se trata de um modelo probabilístico e estocástico, pode gerar soluções diferentes para o mesmo esqueleto de proteína. Muitas vezes, nas campanhas de *design* de proteínas são geradas milhares de sequências com esta ferramenta, sendo seguidamente selecionadas as melhores com base em propriedades, definidas pelos investigadores, que sejam indicativas de uma alta probabilidade de sucesso. Tal como acontecia nas campanhas de *design* baseadas em métodos "físicos" como o ROSETTA, as sequências das proteínas desenhadas são depois transformadas em sequências de DNA que as codificam e que são otimizadas para permitirem uma expressão eficiente no organismo escolhido para as produzir.

Nos projetos internacionais BioPlaTTAR e EvaMobs que coordenamos, usamos estratégias semelhantes à que foi descrita acima para desenhar proteínas com uma forma otimizada para bloquear a superfície dos vírus de modo a impedir a sua entrada nas nossas células, e consequentemente a infeção. Como podemos criar proteínas específicas para diferentes alvos, esta estratégia pode ser aplicada a uma panóplia de doenças e permitir-nos combater futuras pandemias e outros flagelos. Um exemplo pode ser apreciado na Figura 7, onde apresentamos um *design* de uma proteína helicoidal que liga ao RBD (*Receptor Binding Domain*) da proteína-S do vírus SARS-CoV-2.

O que nos espera o futuro?

Os avanços em inteligência artificial estão a mudar a face da ciência e também da sociedade, numa acelerada revolução, com características ímpares. É de esperar o surgimento de novas metodologias que permitam ir ainda mais além na previsão da estrutura de proteínas e outras biomoléculas, assim como no seu desenho. Essas metodologias serão, possivelmente, um misto entre inteligência artificial e modelos físicos, combinando o melhor dos dois mundos. O futuro será um lugar interessante e a Biologia é a Engenharia do século XXI.

Mas existem desafios que precisamos de enfrentar. A inteligência artificial, tradicionalmente um campo académico exótico, tornou-se recentemente uma corrente dominante do desenvolvimento tecnológico, dominada por empresas, dado que a maioria dos académicos (e mesmo as empresas de menores dimensões) não têm recursos computacionais suficientes para treinar modelos de inteligência artificial de última geração. É aqui importante notar que esta situação levanta ainda mais questões pelo facto de iniciativas como o AlphaFold deverem a sua existência ao enorme investimento público que resultou na obtenção dos grandes dados que foram usados para treinar estes métodos. Podemos dizer que estes avanços, embora tenham sido implementados por empresas privadas, foram

alimentados por dinheiro público.

O uso de dados públicos sem garantido retorno público e a desigualdade de recursos, quer na academia quer nas SMEs, em relação aos gigantes tecnológicos, podem ser perniciosos para a inovação. Não é por acaso que o AlphaFold e dois dos laureados de 2024 (Demis Hassabis e John Jumper) pertencem a uma empresa subsidiária de uma das maiores empresas mundiais, com recursos de ordens de grandeza superiores às mais ricas Universidades e Instituições sem fins lucrativos. Naturalmente, as empresas quererão capitalizar os seus investimentos, mas o setor público necessita de encontrar maneiras de potenciar o desenvolvimento tecnológico nestas áreas e ocupar o seu devido lugar no ecossistema de inovação – no futuro das coisas inauditas e cuja aplicabilidade é ainda desconhecida. Cabe também à ciência o papel de não querer apenas ter bolas de cristal que nos permitam fazer previsões e criar proteínas, mas também de gerar novo conhecimento a partir destes métodos e de tornar os seus resultados interpretáveis, passíveis de ser apreendidos pela mente humana e explorados para o desenvolvimento das sociedades humanas.

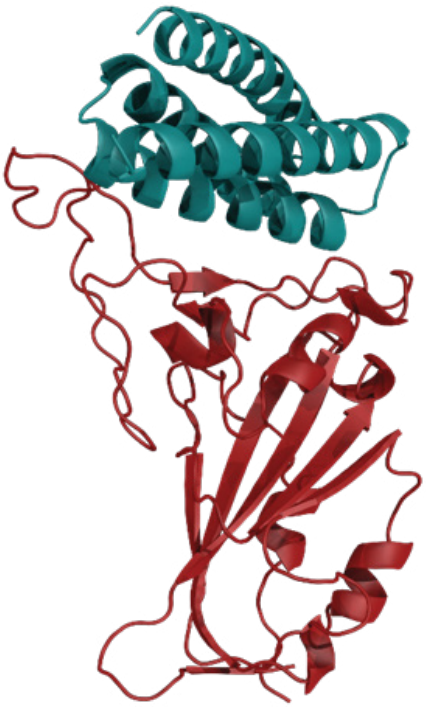


Figura 7 - Estrutura de uma mini-proteína helicoidal (representada em *cartoon* verde) criada usando uma estratégia de *design* computacional baseada nas ferramentas *RF diffusion* e *Protein MPNN* com o objetivo de se ligar ao domínio de ligação ao recetor da proteína *spike* do vírus SARS-CoV-2 (representada em *cartoon* vermelho) e neutralizar este vírus, pelos investigadores do projeto BioPlaTTAR [23].

Referências

[1] J. C. Kendrew, R. E. Dickerson, B. E. Strandberg, R. G. Hart, D. R. Davies, D. C. Phillips, V. C. Shore, *Nature* **1960**,*185*, 422-427. DOI: 10.1038/185422a0.

[2] *The PyMOL Molecular Graphics System, Version 2.5.0 Schrödinger, LLC.*

[3] S. E. Phillips, *J. Mol. Biol.* **1980**, *142*, 531-554. DOI: 10.1016/0022-2836(80)90262-4.

[4] K. D. Gibson, H. A. Scheraga, *Proc. Natl. Acad. Sci.* **1967**, *58*, 420-427. DOI: 10.1073/pnas.58.2.420.

[5] M. Levitt, A. Warshel, *Nature* **1975**, *253*, 694-698. DOI: 10.1038/253694a0.

[6] Y. Duan, P. A. Kollman, *Science* **1998**, *282*, 740-744. DOI: 10.1126/science.282.5389.740.

[7] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, W. Wriggers, *Science* **2010**, *330*, 341-346. DOI: 10.1126/science.1187409.

[8] D. E. Shaw, R. O. Dror, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers, E. Chow, M. P. Eastwood, D. J. Ierardi, J. L. Klepeis, J. S. Kuskin, R. H. Larson, K. Lindorff-Larsen, P. Maragakis, M. A. Moraes, S. Piana, Y. Shan, B. Towles, *Proc. Int. Conf. High Perform. Comput. Netw. Storage Anal. SC 2009*, **2009**, *65*, 1-11. DOI: 10.1145/1654059.1654126.

[9] W. J. Browne, A. C. North, D. C. Phillips, K. Brew, T. C. Vanaman, R. L. Hill, *J. Mol. Biol.* **1969**, *42*, 65-86. DOI: 10.1016/0022-2836(69)90487-2.

[10] a) M. J. Sutcliffe, I. Haneef, D. Carney, T. L. Blundell, *Protein Eng. Des. Sel.* **1987**, *1*, 377-384. DOI: 10.1093/protein/1.5.377; b) M. J. Sutcliffe, F. R. F. Hayes, T. L. Blundell, *Protein Eng. Des. Sel.* **1987**, *1*, 385-392. DOI: 10.1093/protein/1.5.385.

[11] A. Sali, T. L. Blundell, *J. Mol. Biol.* **1993**, *234*, 779-815. DOI: 10.1006/jmbi.1993.1626.

[12] J. Moult, J. T. Pedersen, R. Judson, K. Fidelis, *Proteins: Struct., Funct., Genet.* **1995**, *23*, ii–iv. DOI: 10.1002/prot.340230303.

[13] a) K. F. Han, D. Baker, *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 5814-5818. DOI: 10.1073/pnas.93.12.5814; b) K. T. Simons, C. Kooperberg, E. Huang, D. Baker, *J. Mol. Biol.* **1997**, *268*, 209-225. DOI: 10.1006/jmbi.1997.0959; c) C. Bystroff, D. Baker, *J. Mol. Biol.* **1998**, *281*, 565-577. DOI: 10.1006/jmbi.1998.1943; d) K. T. Simons, R. Bonneau, I. Ruczinski, D. Baker, *Proteins: Struct., Funct., Genet.* **1999**, *37*, 171-176. DOI: 10.1002/(sici)1097-0134(1999)37:3+%3C171::aid-prot21%3E3.3.co;2-q.

[14] a) I. N. Shindyalov, N. A. Kolchanov, C. Sander, *Protein Eng.* **1994**, *7*, 349-358. DOI: 10.1093/protein/7.3.349; b) U. Göbel, C. Sander, R. Schneider, A. Valencia, *Proteins: Struct., Funct., Genet.* **1994**, *18*, 309-317. DOI: 10.1002/prot.340180402.

[15] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, D. Hassabis, *Nature* **2020**, *577*, 706-710. DOI: 10.1038/s41586-019-1923-7.

[16] K. Tunyasuvunakool, J. Adler, Z. Wu, T. Green, M. Zielinski, A. Žídek, A. Bridgland, A. Cowie, C. Meyer, A. Laydon, S. Velankar, G. J. Kleywegt, A. Bateman, R. Evans, A. Pritzel, M. Figurnov, O. Ronneberger, R. Bates, S. A. A. Kohl, A. Potapenko, A. J. Ballard, B. Romera-Paredes, S. Nikolov, R. Jain, E. Clancy, D. Reiman, S. Petersen, A. W. Senior, K. Kavukcuoglu, E. Birney, P. Kohli, J. Jumper, D. Hassabis, *Nature* ***2021***, *596*, 583-589. DOI: 10.1038/s41586-021-03828-1.

[17] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, J. M. Jumper, *Nature* **2024**, *630*, 493-500. DOI: 10.1038/s41586-024-07487-w.

[18] L. Regan, W. DeGrado, *Science* **1998**, *241*, 976-978. DOI: 10.1126/science.3043666.

[19] B. I. Dahiyat, S. L. Mayo, *Science* **1997**, *278*, 82-87. DOI: 10.1126/science.278.5335.82.

[20] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, D. Baker, *Science* **2003**, *302*, 1364-1368. DOI: 10.1126/science.1089427.

[21] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, D. Baker, *Nature* **2023**, *620*, 1089-1100. DOI: 10.1038/s41586-023-06415-8.

[22] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, D. Baker, *Science* **2022**, *378*, 49-56. DOI: 10.1126/science.add2187.

[23] D. H. Silva *et al., manuscrito em preparação.*

Cláudio M. Soares
ITQB NOVA, Instituto de Tecnologia Química e Biológica António Xavier.
claudio@itqb.unl.pt

Diana Lousa
ITQB NOVA, Instituto de Tecnologia Química e Biológica António Xavier.
dlousa@itqb.unl.pt