


O impacto da inteligência artificial nas ciências da vida através da bioinformática

Lucas Moraes dos Santos
Universidade Federal de Minas Gerais (UFMG)

 <https://orcid.org/0000-0003-4214-1576>
moraes.lsanos@gmail.com

Diego Mariano
Universidade Federal de Minas Gerais (UFMG)

 <https://orcid.org/0000-0002-5899-2052>
dcbmariano@gmail.com

Raquel Cardoso de Melo-Minardi
Universidade Federal de Minas Gerais (UFMG)

 <https://orcid.org/0000-0001-5190-100X>
raquelcm@dcc.ufmg.br

RESUMO

Nos últimos anos, as ciências da vida testemunharam uma profunda transformação, impulsionada pelos notáveis avanços na inteligência artificial (IA). Esta transformação tornou-se possível através da convergência de novos métodos e tecnologias que possibilitaram a geração de dados biológicos em larga escala e de alta qualidade. Além disso, a bioinformática tem desempenhado um papel fundamental viabilizando a modelagem e resolução de problemas biológicos complexos, criando assim um terreno fértil à aplicação do aprendizado de máquina e, conseqüentemente, catalisando insights e perspectivas inovadoras na área. Neste trabalho, discutiremos sobre o profundo impacto da IA nas ciências da vida, com particular ênfase naqueles mediados pela bioinformática, na evolução contínua dos algoritmos de IA e nas implicações de longo alcance à pesquisa nas

ciências da vida. Além disso, procuramos elucidar aspectos arquitetônicos inerentes às redes convolucionais e aos modelos generativos, demonstrando o motivo pelo qual cada técnica é aplicada a diferentes problemas biológicos.

Palavras-chave: bioinformática; inteligência artificial; aprendizagem de máquina.

The impact of artificial intelligence in life sciences through bioinformatics

ABSTRACT

En los últimos años, las ciencias de la vida han sido testigos de una profunda transformación, impulsada por avances notables en la inteligencia artificial (IA). Esta transformación fue posible gracias a la convergencia de nuevos métodos y tecnologías que permitieron la generación de datos biológicos a gran escala y de alta calidad. Además, la bioinformática ha jugado un papel fundamental al permitir el modelado y la resolución de problemas biológicos complejos, creando así un terreno fértil para la aplicación del aprendizaje automático y, en consecuencia, catalizando conocimientos y perspectivas innovadoras en el área. En este trabajo, analizamos el profundo impacto de la IA en las ciencias de la vida, con especial énfasis en aquellas mediadas por la bioinformática, la evolución continua de los algoritmos de IA y las implicaciones de gran alcance para la investigación en ciencias de la vida. Además, buscamos dilucidar aspectos arquitectónicos inherentes a las redes convolucionales y modelos generativos, demostrando por qué cada técnica se aplica a diferentes problemas biológicos.

Keywords: bioinformática; inteligencia artificial; aprendizaje automático.

Submissão em: 06/09/2023 | **Aprovação em:** 16/10/2023

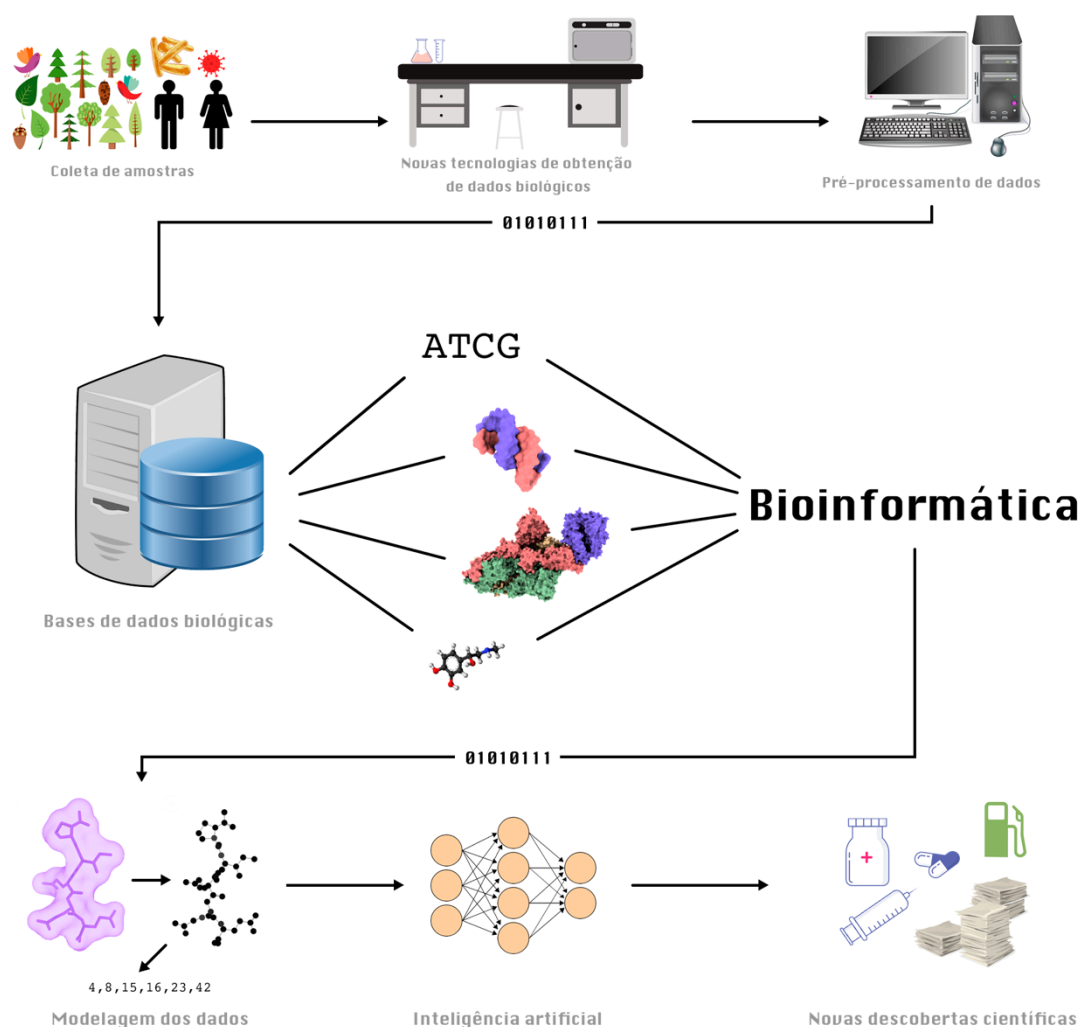
1. INTRODUÇÃO

Nos últimos anos, novas tecnologias, como sequenciadores de próxima geração ou equipamentos para realização de crio-microscopia eletrônica (Crio-ME), têm permitido a produção de uma imensa quantidade de dados biológicos. Além disso, diversas evoluções metodológicas e tecnológicas têm proporcionado um aumento expressivo do volume e da qualidade de bases de dados relacionados às biomoléculas. Nesse contexto, surge a bioinformática, uma área de pesquisa que envolve o uso e desenvolvimento de técnicas computacionais para análises de dados biológicos (Mariano, Ferreira, *et al.*, 2020; Mishra, Das, *et al.*, 2023).

A priori, estudos em bioinformática concentraram-se principalmente na análise de sequências de DNA, RNA e proteínas. No entanto, as pesquisas em bioinformática não se limitam a isso. Por exemplo, o interesse no estudo de estruturas tridimensionais de macromoléculas fomentou o surgimento de uma subárea conhecida como bioinformática estrutural. A bioinformática estrutural se baseia no uso de metodologias da química computacional em combinação com técnicas de modelagem molecular para lidar com questões biológicas de um ponto de vista tridimensional (Verli, 2014).

Recentemente, análises de bioinformática combinadas a algoritmos baseados em inteligência artificial (IA) têm possibilitado a pesquisadores obter novas percepções na abordagem de problemas por anos sem solução. A título de exemplo, o problema do enovelamento de proteínas, um desafio que havia sido estudado por mais de 50 anos (Dill e Maccallum, 2012), teve avanços significativos nos últimos anos com novas soluções envolvendo IA. Em 2021, pesquisadores do Google DeepMind propuseram uma nova arquitetura de redes neurais profundas denominada AlphaFold para predição de estruturas de proteínas a partir da estrutura primária de aminoácidos. Um artigo recente na revista *Nature* (Method of the Year 2021: Protein structure prediction, 2022) o indicou como “o método do ano de 2021” e previu ainda que seus resultados causariam um impacto duradouro e de longo alcance.

Figura 1 - Papel da bioinformática e das técnicas de inteligência artificial para obtenção de novas descobertas a partir de dados biológicos.



Fonte: autoria própria.

A Figura 1 apresenta um pipeline comum na pesquisa em bioinformática baseada em IA. Dado que a coleta de amostras de diversos organismos tem se tornado mais acessível, avanços recentes na tecnologia de sequenciamento de próxima geração (*Next-Generation Sequencing*, NGS) e técnicas para determinação de estruturas tridimensionais de proteínas por Crio-ME, têm possibilitado a obtenção rápida e eficiente de grandes volumes de informações biológicas (Gao, Mahajan, *et al.*, 2020). Em paralelo, o desenvolvimento da tecnologia da informação viabilizou uma capacidade ainda maior

de processamento das bases de dados biológicas em larga escala. Além disso, metodologias de modelagem computacional têm permitido a preparação e representação dos dados (e.g., *one-hot encoding*, mapas de distâncias, grafos) para sua utilização em algoritmos de IA (Defresne, Barbe e Schiex, 2021). Assim, é possível aplicar técnicas de aprendizado de máquina, viabilizando a descoberta de conhecimento.

Nesse sentido, revisões acerca da aprendizagem profunda aplicada à bioinformática têm abordado diferentes pipelines que possibilitam da predição de estrutura até a geração *de novo* de sequências de proteínas. Em dos principais e predecessores trabalhos na área, Min et al. (2017) desenvolveram uma revisão abrangente, na qual descrevem sobre aplicações categorizadas por domínio na bioinformática (ômicas e processamento de imagens ou sinais biomédicos) e arquitetura de aprendizagem profunda (perceptron multicamadas, redes neurais convolucionais e recorrentes). Contudo, com o desenvolvimento da IA generativa, modelos generativos vêm sobrepondo arquiteturas como as redes recorrentes, na predição e modelagem de moléculas baseada em estrutura, sequência ou propriedades físico-químicas, aparecendo majoritariamente em revisões recentes (Huang, Boyken e Baker, 2016; Gao, Mahajan, *et al.*, 2020; Torrisi, Pollastri e Le, 2020; Bai, Liu, *et al.*, 2021; Defresne, Barbe e Schiex, 2021; Abbasi,, Santos, *et al.*, 2022).

Neste trabalho, serão discutidos os impactos da IA na bioinformática e nas ciências da vida. Nas seções seguintes, serão apresentadas técnicas de aprendizagem profunda tradicionais como as redes neurais convolucionais, *autoencoders* variacionais, redes adversárias generativas e as redes neurais recorrentes, além de arquiteturas recentes como os modelos largos de linguagem (também conhecidos como modelos de linguagem baseados em *transformer*) e os modelos de difusão, dentre outros tópicos que serão descritos a partir de aplicações na bioinformática e ciências da vida.

2. APRENDIZAGEM PROFUNDA

Nos últimos anos, o principal desafio à inteligência artificial envolveu a resolução de problemas que são inerentemente simples para humanos resolverem intuitivamente, mas que se tornam complexos quando é necessária uma descrição formal. Isso foi possível por meio de uma abordagem conhecida como *hierarquia de conceitos* que permite ao computador compreender conceitos complexos, construindo-os a partir de componentes mais elementares. Nesse sentido, ao relacionarmos esses conceitos através de um grafo, este será profundo (ou seja, composto por muitas camadas). Por esse motivo essa abordagem é conhecida como aprendizagem profunda em IA (Goodfellow, Bengio e Courville, 2016).

Em outras palavras, a solução dos problemas se dá pela extração de conhecimento (ou padrões não triviais) de bancos de dados através de algoritmos de *aprendizado de máquina*. No campo da *Bioinformática*, a IA tem sido empregada há mais de uma década impulsionada pelo crescente volume de dados biológicos advindos das ômicas (genômica, transcriptômica, proteômica, entre outras) ou ainda, de imagens e outros tipos de sinais biomédicos (Min, Lee e Yoon, 2017).

Contudo, o desempenho desses métodos ainda é dependente da representação dos dados (Goodfellow, Bengio e Courville, 2016). Para contornar esse desafio, tem sido utilizada uma abordagem alternativa - utilizar o aprendizado de máquina na descoberta de representações que podem ser expressas a partir de componentes mais simples (Duda, Hart e Stork, 2001). A necessidade de intervenção humana no processo de preparação dos dados pode ser significativamente diminuída com essa abordagem que é conhecida como *aprendizado de representação* (Bengio, Courville e Vincent, 2013). Dentre as diversas metodologias para aprender representações tem-se destacado a *aprendizagem profunda* (*deep learning*, DL), ou seja, um tipo especializado de aprendizado de máquina que alcança grande poder e flexibilidade aprendendo a representar o mundo a partir de uma hierarquia aninhada de conceitos (Goodfellow, Bengio e Courville, 2016).

Uma característica particular das arquiteturas baseadas em aprendizagem profunda é o uso de múltiplas camadas transformações não lineares, representadas por redes neurais, nas quais a saída de um nó é conectada a entrada de cada nó da camada seguinte (Bai, Liu *et al.*, 2021). Esse arranjo das unidades neuronais forma camadas ocultas e densas, constituindo uma rede totalmente conectada (Gonzalez e Woods, 2008, p. 945). Conseqüentemente, os neurônios ocultos possibilitam à rede extrair progressivamente as características mais significativas dos padrões de entrada, devido à profundidade das camadas (Haykin, 1999). Por esse motivo, essas redes neurais também são denominadas redes neurais profundas (*deep neural networks*, DNN).

Dessa forma, através de grandes volumes de dados sobre biomoléculas (DNA, RNA, proteínas, metabólitos, etc) e das redes neurais profundas, capazes de abstrair conceitos complexos compreendidos nos dados, muitos problemas têm sido resolvidos com maior precisão que se comparado a abordagens baseadas no desenvolvimento de programas clássicos, como fazia a bioinformática tradicionalmente.

Contudo, ainda que realizem predições com precisão igual ou superior ao desempenho humano, a maioria dos modelos baseados em aprendizagem profunda são incapazes de fornecer resultados interpretáveis, frequentemente sendo comparados a uma “caixa preta” (Bai, Liu, *et al.*, 2021). Isso ocorre porque as redes neurais profundas são capazes de mapear a entrada para representações internas de alta dimensionalidade, cujos padrões aprendidos encontram-se distribuídos nas camadas ocultas (Goodfellow, Bengio e Courville, 2016). Assim, torna-se difícil compreender diretamente algumas decisões específicas tomadas pela rede, limitando sua aplicação em sistemas biomoleculares (Li, Liu, *et al.*, 2022). Visando contornar esse problema, um novo campo denominado aprendizado de máquina interpretável (*interpretable machine learning*, IML) tem alcançado interesse e popularidade. A partir do uso de técnicas como visualização de recursos, maximização de ativação e mecanismos de atenção, tornou-se possível obter *insights* sobre o que o modelo aprendeu (Molnar, 2022).

Vamos introduzir a seguir alguns dos principais tipos de redes neurais profundas que têm sido amplamente empregadas em bioinformática com significativa relevância.

2.1 REDES CONVOLUCIONAIS

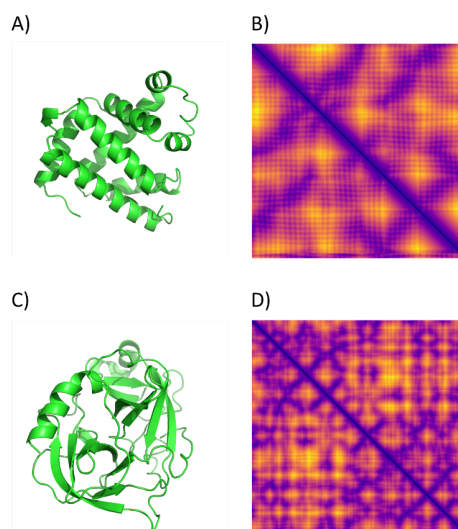
As redes neurais convolucionais (*convolutional neural networks*, CNN) são uma classe de rede neural profunda (LECUN, BOSER, *et al.*, 1989), especializada no processamento de dados que possuem uma topologia em grade/matriz (Goodfellow, Bengio e Courville, 2016, p. 330). Essas arquiteturas são inspiradas na hierarquia do córtex visual humano, na qual os neurônios iniciais respondem a padrões elementares em sub-regiões específicas de estímulos visuais, enquanto neurônios de nível superior sintetizam as informações dessas células mais simples, identificando padrões mais complexos (Min, Lee e Yoon, 2017).

Em alusão ao seu nome, a CNN realiza uma operação linear denominada convolução que pode ser assimilada a um produto escalar de duas matrizes. Uma matriz contendo o conjunto de parâmetros aprendidos, conhecida como *filtro*, desliza pela imagem, envolvendo uma outra matriz que representa uma região local da entrada (Goodfellow, Bengio e Courville, 2016). Nesse caso, as camadas convolucionais não aprendem a entrada globalmente de uma única vez, mas concentram-se no padrão local do campo receptivo (Bian e Xie, 2021). Essa característica possibilita as CNNs aprenderem de padrões elementares (*e.g.*, bordas, texturas) a características e conceitos abstratos (*e.g.*, faces, objetos). Isso se destaca como um exemplo de como princípios neurocientíficos influenciam a aprendizagem profunda (Molnar, 2022).

No contexto da bioinformática estrutural, em que se estuda a estrutura 3D de proteínas e sua interação com outras biomoléculas, as CNNs têm sido amplamente utilizadas na predição de estrutura ou função de proteínas (Gao, Mahajan, *et al.*, 2020; Torrisi, Pollastri e Le, 2020). Isto tem sido possível pois as CNNs se beneficiam das simetrias das representações estruturais em duas dimensões, como por exemplo dos mapas de distâncias que são representações em imagens das distâncias entre todos os pares de átomos de uma proteína (Defresne, Barbe e Schiex, 2021).

O exemplo mais notório é a primeira versão do AlphaFold¹ (Alquraishi, 2019; Senior, Evans, *et al.*, 2020), cujo componente central é uma rede residual convolucional treinada com estruturas do PDB para prever distogramas² de proteínas, com base em sequências de aminoácidos. Além disso, as CNNs demonstraram ser relevantes na identificação dos estados funcionais de proteínas e peptídeos estáticos (Santos e Melo-Minardi, 2022; Santos *et al.*, 2023), bem como daqueles simulados através da dinâmica molecular (Han, Lee, *et al.*, 2022), usando uma representação por mapas de distâncias (Figura 2), ou pixels (Plante, Shore, *et al.*, 2019; Li, Liu, *et al.*, 2022), e na modelagem *ab initio* de proteínas (Anishchenko, Pellock, *et al.*, 2021).

Figura 2 – Mapas de distâncias. A) Estrutura tridimensional de uma mioglobina (PDB ID: 1TES). B) Mapa de distâncias correspondente a mioglobina. C) Estrutura tridimensional de uma tripsina (PDB ID: 1SGT). D) Mapa de distâncias correspondente a tripsina.



Fonte: autoria própria.

¹ AlphaFold. Disponível em: <https://alphafold.ebi.ac.uk/>.

² Representação 2D na qual cada pixel corresponde distribuição de probabilidade do quão próximo encontram-se os pares de resíduos no mapa de distâncias (SENIOR, EVANS, *et al.*, 2020).

3. INTELIGÊNCIA ARTIFICIAL GENERATIVA

Os modelos baseados em IA generativa, também conhecidos como *modelos generativos*, têm o objetivo de modelar uma distribuição a partir dos dados de treinamento e, posteriormente, produzir resultados semelhantes (Sanchez-Lengeling e Aspuru-Guzik, 2018; Zeng, Wang, *et al.*, 2022). Um exemplo bastante simples é treinar a rede com fotos de faces de seres humanos e depois utilizar o modelo generativo para produzir faces nunca vistas. Essa abordagem tem se mostrado promissora ao desenho de novas moléculas (inclusive proteínas), através da exploração não apenas do espaço de sequência, mas também de propriedades estruturais e funcionais (Huang, Boyken e Baker, 2016; Tong, Liu, *et al.*, 2021). Isso se deve ao fato de que cada molécula possui uma representação única em um espaço latente multidimensional, que codifica suas características. Por exemplo, considere uma molécula composta por três átomos A, B e C. Nesse caso, cada átomo poderia ser descrito por características, como seu elemento químico (e.g., carbono, hidrogênio, oxigênio), carga (positiva, neutra e negativa), polaridade (apolar e polar) e suas coordenadas espaciais (x, y, z), por exemplo. Assim, a combinação dessas características resulta em potenciais moléculas em um espaço de possibilidades – referido como *espaço latente* – nesse caso 4D (elemento, carga, polaridade, posição).

Os modelos generativos exploram esse espaço variável, gerando vetores de características (ou latentes) a partir de uma distribuição de probabilidade (Sanchez-Lengeling e Aspuru-Guzik, 2018). Pesquisadores têm constatado que esses modelos treinados com sequências e estruturas de proteínas nativas, têm grande potencial na geração de moléculas *de novo* (do zero) (Callaway, 2023). Essas redes neurais têm transformado o desenho e a otimização de pequenas moléculas e macromoléculas, apresentando novos candidatos parcialmente otimizados, em alguns casos em um tempo inferior ao normalmente executado por abordagens convencionais (Zeng, Wang, *et al.*, 2022; Thomas, Bender e Graaf, 2023). O desempenho desses modelos na química generativa se deve ao grande volume de dados disponíveis em bases de dados (Min, Lee e Yoon, 2017, Bian e Xie, 2021).

3.1 AUTOENCODERS VARIACIONAIS

Um *autoencoder* é uma rede neural treinada para tentar reconstruir a entrada em sua saída, com a maior precisão possível (Goodfellow, Bengio e Courville, 2016, p. 499). O aprendizado nessa arquitetura baseia-se em uma função, denominada codificadora, que transforma os dados em alta dimensionalidade (entrada) para uma representação de baixa dimensionalidade em um espaço latente, e um decodificador que reconstrói a entrada original a partir dessa representação ((Goodfellow, Bengio e Courville, 2016; Tong, Liu, *et al.*, 2021). No entanto, no contexto de aplicações generativas, é mais frequente encontrar uma variante desse modelo conhecida como *autoencoder* variacional (*variational autoencoder*, Vae), sendo a entrada representada como uma distribuição de probabilidade no espaço latente (Kingma e Welling, 2013; Sousa, Correia, *et al.*, 2021).

Essas arquiteturas têm sido amplamente empregadas no desenho de novas moléculas (Chenthamarakshan, Hoffman, *et al.*, 2023), uma vez que possibilitam a exploração de regiões desconhecidas do espaço de design químico a partir de representações latentes de baixa dimensionalidade (Anstine e Isayev, 2023). Nesse caso, o codificador mapeia a molécula em um *embedding* contínuo, ou seja, um vetor latente com dimensionalidade inferior à representação da entrada (Bian e Xie, 2021), enquanto o decodificador tenta recuperar a molécula a partir do *embedding* aprendido. Assim, o objetivo no treinamento de VAEs é minimizar o erro de reconstrução, ou seja, a diferença entre a representação de entrada fornecida ao codificador e a saída do decodificador (Anstine e Isayev, 2023).

3.2 REDES ADVERSÁRIAS GENERATIVAS

As redes adversárias generativas (*generative adversarial networks*, GAN) são arquiteturas de redes neurais profundas compostas por dois módulos concorrentes: gerador e discriminador (Goodfellow, Pouget-Abadie, *et al.*, 2014). O aprendizado em GANs pode ser compreendido como um jogo de soma zero, onde o gerador aprende a distribuição dos dados de treinamento criando instâncias sintéticas para classificação

pelo discriminador. Conforme o treinamento avança, o gerador melhora, produzindo resultados cada vez mais realistas para enganar o discriminador e classificá-los como amostras autênticas. Simultaneamente, o discriminador se torna mais eficiente em distinguir instâncias reais das falsas. Eventualmente, as amostras do gerador tornam-se indistinguíveis dos dados reais, tornando o discriminador desnecessário à medida que o treinamento converge (Goodfellow, Bengio e Courville, 2016, p. 702), condição conhecida como *Equilíbrio de Nash* (Nash, 1950).

Esses modelos podem ser conceitualmente simples de compreender a partir de um exemplo de geração de moléculas. O discriminador visa maximizar a taxa de erro das moléculas sintéticas do gerador, enquanto este, por sua vez, tenta minimizar o erro a partir da criação de moléculas suficientemente realistas, capazes de enganar o discriminador (Bilodeau, Jin, *et al.*, 2022; Zeng, Wang, *et al.*, 2022). Ou seja, mesmo um discriminador bem treinado pode classificar instâncias geradas como reais, mostrando a capacidade do gerador em criar compostos promissores a partir de uma entrada aleatória latente (Bian e XIE, 2021). Assim, esses modelos têm sido adaptados para lidar com diversos desafios na química generativa, abrangendo a criação de moléculas *de novo* (Anand e Huang, 2018; Surana, Arora, *et al.*, 2023; Xie, Valiente e Kim, 2023), predição da estrutura de proteínas (Ding e Gong, 2020) e modelagem de *loop* (Li, Nguyen, *et al.*, 2017).

3.3 REDES NEURAIIS RECORRENTES

As redes recorrentes (*recurrent neural networks*, RNN) são uma família de redes neurais para processamento de dados sequenciais (Rumelhart, Hinton e Williams, 1986; Goodfellow, Bengio e Courville, 2016). Essas arquiteturas apresentam *loops* internos, nos quais as saídas das camadas no estado anterior são transferidas para o estado atual (Mitchell, 1997). Dessa forma, elas conseguem processar a entrada consecutivamente, ao invés de uma única etapa (Bian e Xie, 2021). Contudo, o aumento das etapas na RNN pode ocasionar o desaparecimento de gradientes na retropropagação, dificultando o aprendizado de dependência a longo prazo. A inclusão de unidades LSTM (*Long*

Short-Term Memory) contorna esse problema, introduzindo parâmetros aprendíveis para controlar o fluxo de informações (Sousa, Correia, *et al.*, 2021).

Essas arquiteturas têm sido empregadas em tarefas de modelagem de sistemas que possuem um componente sequencial, como a geração de moléculas a partir da sequência primária (Min, Lee e Yoon, 2017; Zeng, Wang, *et al.*, 2022). Neste contexto, é comum ser utilizada uma notação conhecida como Smiles (*Simplified Molecular Input Line Entry System*), ou seja, codificações textuais simplificadas, como *tokens*, para descrever estruturas moleculares, tornando-o uma escolha predominante para representar pequenas moléculas (Sousa, Correia, *et al.*, 2021). Após treinar a RNN com muitas strings Smiles, seria possível prever novas sequências válidas não presentes no conjunto de dados inicial (Tong, Liu, *et al.*, 2021). Por exemplo, a partir de uma sequência inicial (e.g., "CC"), a RNN atribui probabilidades a caracteres subsequentes. Neste caso, "1" (um) corresponderia a alta probabilidade, podendo ser escolhido como o próximo caractere. "1" seria a entrada de feedback à RNN. Este ciclo continua até o token final, "\n", denotando o fim da sequência (Zeng, Wang, *et al.*, 2022).

3.4 MODELOS LARGOS DE LINGUAGEM

Modelos largos de linguagem (*large language models*, LLM) são uma categoria especializada de modelo de linguagem caracterizada por uma rede neural profunda contendo um extenso número de parâmetros, comumente na ordem de 10^9 . Recentemente, eles tornaram-se o estado da arte no processamento de linguagem natural (Sousa, Correia, *et al.*, 2021) uma vez que utilizam uma arquitetura denominada *transformer* (Vaswani, Shazeer, *et al.*, 2017). Esse modelo consiste em uma estrutura codificador-decodificador, na qual um componente crucial, conhecido como *mecanismo de atenção*, permite a ele focar em partes relevantes da sentença. Além disso, devido ao paralelismo do *transformer*, as sentenças são processadas como um todo, prevenindo o desaparecimento do gradiente durante o treinamento (Zhang, Fan, *et al.*, 2023).

As similaridades intrínsecas às linguagens naturais e sequências biológicas têm impulsionado a aplicação de LLMs (e.g., Bert, GPT-3) na pesquisa em química e bioinformática, abrangendo da análise de sequência a modelagem *de novo* (Vert, 2023; Zhang, FAN, *et al.*, 2023). Isso, pois, dados biológicos heterogêneos e de alta dimensionalidade (e.g., descritores moleculares, estruturas de proteínas etc.) podem ser representados de maneira uniforme, através de *embeddings* (Tong, Liu, *et al.*, 2021; Vert, 2023). Um *embedding* é um vetor contínuo no espaço n -dimensional usado para codificar padrões complexos a partir dos dados de entrada (e.g., texto ou imagens). Eles são estruturados de forma a posicionar conceitos semelhantes próximos uns dos outros no espaço de *embeddings*, capturando relações semânticas. Esses vetores podem ser alimentados diretamente nos LLMs, que extraem informações dependentes do contexto biológico da correlação de segmentos de *embeddings* (Zhang, Fan, *et al.*, 2023).

Nesse sentido, tanto os *transformers* como os LLMs, têm sido o núcleo de métodos computacionais para predição de estruturas tridimensionais de proteínas com alta precisão, representando o estado da arte nesse campo. A partir da utilização de *embeddings* derivados de alinhamentos múltiplos de sequência e um modelo baseado em *transformer*, o AlphaFold 2 (Jumper, Evans, *et al.*, 2021) obteve previsões de estruturas de proteínas em paridade com métodos experimentais, alcançando uma pontuação GDT de 92,4 no CASP14³. Outro exemplo é o ESM-2⁴ (Lin, Akin, *et al.*, 2023), uma nova família de modelos de linguagem de proteínas baseados em *transformers* pré-treinados generativos (*generative pre-trained transformer*, GPT) para inferir uma estrutura diretamente da sequência primária. Nesse caso, o ESM-2 aproveita informações de coevolução de resíduos em sequências para produzir previsões de nível atômico (Lin, Akin, *et al.*, 2023; Zhang, Fan, *et al.*, 2023).

³ Em 2020, o AlphaFold 2 (SENIOR, EVANS, *et al.*, 2020) alcançou uma precisão de 92,4 (numa escala de 0 a 100) no Teste de Distância Global (Global Distance Test, GDT), uma métrica utilizada pela Avaliação Crítica de Técnicas para Predição de Estrutura de Proteínas (*Critical Assessment of protein Structure Prediction*, CASP). O resultado significa que, em mais da metade das previsões realizadas pelo programa, a corretude dos átomos na estrutura prevista encontra-se acima de 92,4%. Esse nível de precisão é comparável a técnicas experimentais, como a cristalografia de raios-X.

⁴ ESM-2. Disponível em: <https://esmatlas.com/>.

3.5 MODELOS DE DIFUSÃO

Os modelos de difusão (*diffusion models*, Diff) (Sohl-Dickstein, Weiss, *et al.*, 2015) têm sido bem-sucedidos na realização de tarefas como geração de imagens a partir de texto, produzindo amostras de alta qualidade (Anstine e Isayev, 2023). Essas redes têm como objetivo modelar uma distribuição de probabilidade condicional dos dados, a partir de etapas em sequência. Nesse caso, a estrutura de entrada (*e.g.*, imagem) é transformada gradualmente através de um procedimento iterativo conhecido como *difusão*. Esse processo envolve a introdução de um ruído gaussiano, que desloca os dados em direção a uma distribuição alvo, onde a estrutura resultante não possui qualquer semelhança com a estrutura inicial. Durante o treinamento, a rede aprende a remover o ruído a partir de uma estimativa da máxima verossimilhança, na qual o modelo visa maximizar a probabilidade de geração da entrada a partir dos dados ruidosos (Sohl-Dickstein, Weiss, *et al.*, 2015).

Recentemente, uma equipe de pesquisadores desenvolveu um programa de IA denominado *RoseTTAFold diffusion*⁵ (RFdiffusion) (Watson, Juergens, *et al.*, 2023), que utiliza difusão guiada na criação de proteínas personalizadas (Callaway, 2023). Nesse caso, os autores ajustam a rede de predição de estrutura RoseTTAFold para o design de proteínas de acordo com restrições específicas de projeto durante a remoção de ruído – processo conhecido como condicionamento (Callaway, 2023; Watson, Juergens, *et al.*, 2023) – alcançando excelente desempenho no design de backbones de proteínas. Além disso, modelos baseados em difusão têm sido aplicados na geração de conformação molecular, onde o modelo aprende a modificar iterativamente os pares de distâncias atômicas para gerar conformações estáveis (Bilodeau, Jin, *et al.*, 2022). Contudo, ainda que representem uma abordagem promissora, os modelos de difusão ainda são pouco explorados (Anstine e Isayev, 2023).

⁵ RFdiffusion. Disponível em: <https://github.com/RosettaCommons/RFdiffusion>.

4. APRENDIZAGEM PROFUNDA APLICADA A PESQUISA EM BIOINFORMÁTICA

A diversidade de arquiteturas de aprendizagem profunda tem oferecido uma gama de alternativas à representação de dados biológicos, o que tem impulsionado o crescente e extenso volume de trabalhos em diferentes campos da bioinformática (Min Lee e Yoon, 2017). Além disso, programas de IA como o trRosetta⁶ (Du *et al.*, 2021) e, mais recentemente, o AlphaFold2 (Jumper, Evans, *et al.*, 2021) têm avançado para além da modelagem por homologia, sendo utilizados no design *de novo* de moléculas, e pavimentado o caminho para novas aplicações. Um exemplo é o AlphaFold Multimer, uma extensão do AlphaFold2 que foi desenvolvida especificamente para prever complexos proteína-proteína. Nesse sentido, a aprendizagem profunda e a IA generativa têm auxiliado no desenvolvimento ágil de moléculas personalizadas, explorando um espaço desconhecido de candidatos terapêuticos (Callaway, 2023). Com isso, torna-se possível a descoberta, ou otimização, de moléculas *in silico* para um determinado alvo, ou função, servindo de base para a descoberta de medicamentos e desenvolvimento de vacinas (Vert, 2023). O apoio computacional por meio da IA torna os processos de desenvolvimento de moléculas, tanto com fins biotecnológicos quanto farmacológicos, mais ágil, barato e promissor.

O Quadro 1 apresenta uma síntese de trabalhos recentes em diferentes campos da bioinformática, onde as técnicas de aprendizagem profunda discutidas anteriormente, foram aplicadas em objetivos específicos com sucesso.

⁶ trRosetta. Disponível em: <https://yanglab.nankai.edu.cn/trRosetta/>.

Quadro 1 - Síntese de aplicações representativas baseadas em aprendizagem profunda e IA generativa nos diferentes campos da pesquisa em bioinformática

Modelagem de proteínas <i>de novo</i>			
Arquitetura	Objetivo	Fonte	Repositório
CNN	Modelagem <i>de novo</i> de proteínas baseada em uma combinação de DNNs e simulações de refinamento estrutural, para gerar sequências alucinadas similares a sequências de proteínas nativas	ANISHCHENKO, PELLOCK, <i>et al.</i> , 2021	https://github.com/gjoni/trDesi-gn
GAN	Modelagem generativa de estruturas de proteínas, codificadas como distâncias pareadas entre os carbonos- α no backbone da proteína	ANAND e HUANG, 2018	https://github.com/collinarnett/protein-gan
GAN	Geração e otimização de moléculas, usando uma representação de SMILES	ABBASI, SANTOS, <i>et al.</i> , 2022	https://github.com/larngroup/GAN-Drug-Generator
GAN	Geração de sequências para novos peptídeos antivirais	SURANA, ARORA, <i>et al.</i> , 2023	https://github.com/thoughtworks/antiviral-peptide-predictions-using-gan/
GAN	Design <i>de novo</i> de α -hélices representadas como vetores de sequência e <i>features</i> estruturais	XIE, VALIENTE e KIM, 2023	https://github.com/xxiexuezhi/helix-gan
Modelos de difusão	Geração de estruturas moleculares 3D condicionadas à forma de uma determinada molécula	CHEN, PENG, <i>et al.</i> , 2023	https://github.com/ehoogeboom/e3-diffusion-for-molecules
Modelos de linguagem baseados em <i>transformer</i>	Um modelo baseado em GPT pré-treinado para geração de sequências semelhantes a proteínas naturais a partir "do zero"	FERRUZ, SCHMIDT e HÖCKER, 2022	https://huggingface.co/nferruz/ProtGPT2
RNN	Descoberta de potenciais inibidores de quinase. As moléculas foram representadas por SMILES	LI, XU, <i>et al.</i> , 2020	https://github.com/Xyqii/RNN-generator
<i>Transformer</i>	Design racional de peptídeos antimicrobiais	MAO, GUAN, <i>et al.</i> , 2023	https://github.com/AspirinCod

			e/AMPTrans-Istm
VAE	Design de inibidores para o domínio ligante ao receptor (RBD ⁷) da proteína spike e a principal protease do SARS-CoV-2 (M ^{pro})	CHENTHAMAR AKSHAN, HOFFMAN, <i>et al.</i> , 2023	https://zenodo.org/records/7863805
Previsão de estrutura e função			
Arquitetura	Objetivo	Fonte	Repositório
CNN	Classificação de estados funcionais específicos a ligantes de GPCRs, usando conformações moleculares codificadas em trajetórias de DM transformadas em representações de pixel	PLANTE, SHORE, <i>et al.</i> , 2019	-
CNN	AlphaFold 1	ALQURAISHI, 2019; SENIOR, EVANS, <i>et al.</i> , 2020	https://github.com/google-deepmind/alphafold
CNN	Modelo DL interpretável baseado em representação por pixels para identificação de estados funcionais em trajetórias de DM de GPCRs, revelando resíduos-chave para compreensão do mecanismo funcional	LI, LIU, <i>et al.</i> , 2022	https://github.com/Jane-Liu97/ICNNMD
CNN	Reconhecimento de mudanças conformacionais relacionadas a padrões espaço-temporais induzidos por ligantes de receptores β 2-AR usando mapas de distâncias	HAN, LEE, <i>et al.</i> , 2022	https://github.com/MinwooHan84/beta2AR
GAN	Modelagem de <i>loop</i> ⁸ usando uma representação de mapas de distâncias dos carbonos- α da proteína	LI, NGUYEN, <i>et al.</i> , 2017	-
GAN	Previsão das distâncias inter-resíduos para proteínas, a partir de informações de coevolução 1D e 2D	DING e GONG, 2020	https://github.com/Wenze-Codebase/DistancePrediction-Protein-GAN.git

⁷ Receptor Binding Domain

⁸ Predição de regiões ausentes em uma estrutura de proteína.

Modelos de linguagem baseados em <i>transformer</i>	ESM-1b	RIVES, MEIER, <i>et al.</i> , 2019; RAO, MEIER, <i>et al.</i> , 2020	https://github.com/facebookresearch/esm
Modelos de linguagem baseados em <i>transformer</i>	Modelo de linguagem profunda auto-supervisionado projetado especificamente para proteínas, o qual captura representações locais e globais de proteínas de maneira natural	BRANDES, OFER, <i>et al.</i> , 2022	https://github.com/nadavbra/protein_bert
Modelos de linguagem baseados em <i>transformer</i>	ESM-2	LIN, AKIN, <i>et al.</i> , 2023	https://github.com/facebookresearch/esm
<i>Transformer</i>	AlphaFold 2	JUMPER, EVANS, <i>et al.</i> , 2021	https://github.com/google-deepmind/alphafold
Expressão gênica			
Arquitetura	Objetivo	Fonte	Repositório
VAE	Geração de novas moléculas a partir de uma entrada constituída por compostos representados como SMILES e assinaturas de expressão gênica	PRAVALPHRUE KUL, PIRIYAJITAKON KIJ, <i>et al.</i> , 2023	https://github.com/ChemEXL/BiCEV
Modelos de linguagem baseados em <i>transformer</i>	Previsão da patogenicidade de variantes <i>missense</i> no genoma humano	BRANDES, GOLDMAN, <i>et al.</i> , 2023	https://github.com/ntranoslab/esm-variants

Fonte: autoria própria.

5. CONSIDERAÇÕES FINAIS

Ao longo deste trabalho, discorreremos sobre as principais técnicas de um subcampo da IA, conhecido como aprendizagem profunda, que têm sido amplamente adotadas na pesquisa em bioinformática. Buscamos também elucidar aspectos arquiteturais específicos às redes convolucionais e aos modelos generativos, demonstrando o motivo pelo qual cada técnica é aplicada em diferentes problemas biológicos.

A inteligência artificial tem um grande potencial para transformar a pesquisa em biologia e áreas afins. Estas áreas de pesquisa têm sido totalmente transformadas pelo uso de modelos computacionais nos últimos anos. Há um volume de dados crescente sobre os sistemas biológicos e que cresce também em qualidade e complexidade, de forma que tiram proveito dos novos modelos e algoritmos desenvolvidos pela bioinformática. A aplicação, ainda incipiente, de técnicas de IA nas ciências da vida por meio da bioinformática já trazem impactos consideráveis em diversas áreas de pesquisa e na computação, com a demanda por novos modelos e algoritmos. O que se vislumbra hoje é apenas o começo de um futuro brilhante desta parceria entre as ciências da vida e a ciência da computação.

REFERÊNCIAS

- ABBASI, Maryam et al. *Designing optimized drug candidates with Generative Adversarial Network*. Journal of Cheminformatics, 14, n. 1, 2022. 40.
- ALQURAIISHI, Mohammed. *AlphaFold at CASP13*. Bioinformatics, 35, n. 22, 2019. 4862-4865.
- ANAND, Namrata; HUANG, Po-Ssu. *Generative modeling for protein structures*. Proceedings of the 32nd International Conference on Neural Information Processing Systems, 31, 2018. 7505-7516.
- ANISHCHENKO, Ivan. et al. *De novo protein design by deep network hallucination*. Nature, 600, n. 7889, 2021. 547-552.
- ANSTINE, Dylan M.; ISAYEV, Olexandr. *Generative Models as an Emerging Paradigm in the Chemical Sciences*. Journal of the American Chemical Society, 145, n. 16, 2023. 8736-8750.
- BAI, Qifeng et al. *Application advances of deep learning methods for de novo drug design and molecular dynamics simulation*. WIREs Computational Molecular Science, 12, n. 3, 2021. e1581.
- BENGIO, Yoshua; COURVILLE, Aaron; VINCENT, Pascal. *Representation Learning: A Review and New Perspectives*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35, n. 8, 2013. 1798-1828.
- BIAN, Yuemin; XIE, Xiang-Qun. *Generative chemistry: drug discovery with deep learning generative models*. Journal of Molecular Modeling, 27, n. 3, 2021. 71-89.

- BILODEAU, Camille et al. *Generative models for molecular discovery: Recent advances and challenges*. WIREs Computational Molecular Science, 12, n. 5, 2022.
- BRANDES, Nadav et al. *ProteinBERT: a universal deep-learning model of protein sequence and function*. Bioinformatics, 38, n. 8, 2022. 2102-2110.
- BRANDES, Nadav et al. *Genome-wide prediction of disease variant effects with a deep protein language model*. Nature Genetics, 2023.
- CALLAWAY, Ewen. *AI tools are designing entirely new proteins that could transform medicine*. Nature, 619, n. 7969, 2023. 236-238.
- CHEN, Ziqi et al. *Shape-conditioned 3D Molecule Generation via Equivariant Diffusion Models*. Preprint, 2023. Disponível em: <<https://arxiv.org/abs/2308.11890>>. Acesso em: 5 Setembro 2023.
- CHENTHAMARAKSHAN, Vijil et al. *Accelerating drug target inhibitor discovery with a deep generative foundation model*. Science Advances, 9, n. 25, 2023.
- DEFRESNE, Marianne; BARBE, Sophie; SCHIEX, Thomas. *Protein Design with Deep Learning*. International Journal of Molecular Sciences, 22, n. 21, 2021. 11741.
- DILL, Ken A.; MACCALLUM, Justin L. *The Protein-Folding Problem, 50 Years On*. Science, 338, n. 6110, 2012. 1042-1046.
- DING, Wenzhe; GONG, Haipeng. *Predicting the Real-Valued Inter-Residue Distances for Proteins*. Advanced Science, 7, n. 19, 2020. 2001314.
- DU, Zongyang et al. *The trRosetta server for fast and accurate protein structure prediction*. Nature Protocols, 16, n. 12, 2021. 5634-5651.
- DUDA, Richard O.; HART, Peter E.; STORK, David G. *Pattern Classification*. 2^a. ed.
- FERRUZ, Noelia; SCHMIDT, Steffen; HÖCKER, Birte. *ProtGPT2 is a deep unsupervised language*. Nature Communications, 13, n. 1, 2022. 4348.
- GAO, Wenhao et al. *Deep Learning in Protein Structural Modeling and Design*. Patterns, 1, n. 9, 2020. 100142.
- GONZALEZ, Rafael C.; WOODS, Richard E. *Digital Image Processing*. 4^a. ed.
- GOODFELLOW, Ian J. et al. *Generative Adversarial Networks*. Advances in Neural Information Processing System, 2014. 2672-2680.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. Deep Learning.

HAN, Minwoo et al. *Recognition of the ligand-induced spatiotemporal residue pair pattern of β 2-adrenergic receptors using 3-D residual networks trained by the time series of protein distance maps*. Computational and Structural Biotechnology Journal, 20, 2022. 6360-6374.

HAYKIN, Simon. Neural Networks: A Comprehensive Foundation. 2^a. ed.

HOLT, Charles A.; ROTH, Alvin E. *The Nash equilibrium: A perspective*. Proceedings of the National Academy of Sciences, 101, n. 12, 2004. 3999-4002.

HOPFIELD, J. J. *Neural networks and physical systems with emergent collective computational abilities*. Proceedings of the National Academy of Sciences, 79, n. 8, 1982. 2554-2558.

HUANG, Po-Ssu; BOYKEN, Scott E.; BAKER, David. *The coming of age of de novo protein design*. Nature, 537, n. 7620, 2016. 320-327.

JUMPER, John et al. *Highly accurate protein structure prediction with AlphaFold*. Nature, 596, n. 7873, 2021. 583-589.

KINGMA, Diederik P.; WELING, Max. *Auto-Encoding Variational Bayes*, 2013.

LECUN, Yan et al. *Backpropagation Applied to Handwritten Zip Code Recognition*. Neural Computation, 1, n. 4, 1989. 541-551.

LI, Chuan et al. *An Interpretable Convolutional Neural Network Framework for Analyzing Molecular Dynamics Trajectories: a Case Study on Functional States for G-Protein-Coupled Receptors*. Journal of Chemical Information and Modeling, 62, n. 6, 2022. 1399-1410.

LI, Xuanyi et al. *Chemical space exploration based on recurrent neural networks: applications in discovering kinase inhibitors*. Journal of Cheminformatics, 12, n. 1, 2020. 42.

LI, Yuesen et al. *DrugGPT: A GPT-based Strategy for Designing Potential Ligands Targeting Specific Proteins*. Preprint, 2023. Disponível em: <<https://doi.org/10.1101/2023.06.29.543848>>. Acesso em: 5 Setembro 2023.

LI, Zhaoyu et al. *Protein Loop Modeling Using Deep Generative Adversarial Network*. IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI), 2017. 1085-1091.

LIN, Zeming et al. *Evolutionary-scale prediction of atomic-level protein structure with a language model*. Science, 379, n. 6637, 2023. 1123-1130.

MAO, Jiashun et al. *Application of a deep generative model produces novel and diverse functional peptides against microbial resistance*. *Computational and Structural Biotechnology Journal*, 21, 2023. 463-471.

MARIANO, Diego et al. *A Brief History of Bioinformatics Told by Data Visualization*. *Advances in Bioinformatics and Computational Biology*. BSB 2020. *Lecture Notes in Computer Science*. Belo Horizonte: Springer, Cham. 2020. p. 235-246.

METHOD of the Year 2021: Protein structure prediction. *Nature Methods*, 19, n. 1, 2022.

MIN, Seonwoo; LEE, Byunghan; YOON, Sungroh. *Deep learning in bioinformatics*. *Briefings in Bioinformatics*, 18, n. 5, 2017. 851-869.

MISHRA, Sarbani et al. *Introduction to the World of Bioinformatics*. *A Guide to Applied Machine Learning for Biologists*, 2023. 105-126.

MITCHELL, Tom M. *Machine learning*.

MOLNAR, Christoph. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2^a. ed.

NASH, John F. *Equilibrium points in n-person games*. *Proceedings of the National Academy of Sciences*, 36, n. 1, 1950. 48-49.

PLANTE, Ambrose et al. *A Machine Learning Approach for the Discovery of*

Ligand-Specific Functional Mechanisms of GPCRs. *Molecules*, 24, n. 11, 2019. 2097.

PRAVALPHRUEKUL, Nutaya et al. *De Novo Design of Molecules with Multi-action Potential from Differential Gene Expression using Variational Autoencoder*. *Journal of Chemical Information and Modeling*, 63, n. 13, 2023. 3999-4011.

RAO, Roshan. et al. *Transformer protein language models are unsupervised structure learners*. *International Conference on Learning Representations*, 2020.

RIVES, Alexander. et al. *Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences*. *PNAS*, 118, n. 15, 2019.

ROSSETTO, Allison; ZHOU, Wenjin. *GANDALF: Peptide Generation for Drug Design Using Sequential and Structural Generative Adversarial Networks*. *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2020.

RUMELHART, David E.; HINTON, Geoffrey E.; WILLIAMS, Ronald J. *Learning representations by back-propagating errors*. *Nature*, 323, n. 6088, 1986. 533-536.

SANCHEZ-LENGELING, Benjamin; ASPURU-GUZIK, Alán. *Inverse molecular design using machine learning: Generative models for matter engineering*. *Science*, 361, n. 6400, 2018. 360-365.

SANTOS, Lucas M. D. et al. *Peptide-Protein Interface Classification Using Convolutional Neural Networks*. *Advances in Bioinformatics and Computational Biology*. BSB 2023. *Lecture Notes in Computer Science*. Curitiba: Springer, Cham. 2023. p. 112-122.

SANTOS, Lucas M. D.; MELO-MINARDI, Raquel C. D. *Identifying Large Scale Conformational Changes in Proteins Through Distance Maps and Convolutional Networks*. *Advances in Bioinformatics and Computational Biology*. BSB 2022. *Lecture Notes in Computer Science*. Búzios: Springer, Cham. 2022. p. 56-67.

SENIOR, Andrew W. et al. *Improved protein structure prediction using potentials from deep learning*. *Nature*, 577, n. 7792, 2020. 706–710.

SOHL-DICKSTEIN, Jascha et al. *Deep Unsupervised Learning using Nonequilibrium Thermodynamics*. *ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning*. Lille: [s.n.]. 2015. p. 2256–2265.

SOUSA, Tiago et al. *Generative Deep Learning for Targeted Compound Design*.

Journal of Chemical Information and Modeling, 61, n. 11, 2021. 5343-5361.

SURANA, Shraddha. et al. *PandoraGAN: Generating antiviral peptides*. *SN COMPUT. SCI.*, 4, n. 607, 2023.

THOMAS, Morgan; BENDER, Andreas; GRAAF, Chris D. *Integrating structure-based approaches in generative molecular design*. *Current Opinion in Structural Biology*, 79, 2023. 102559.

TONG, Xiaochu et al. *Generative Models for De Novo Drug Design*. *Journal of Medicinal Chemistry*, 64, n. 19, 2021. 14011-14027.

TORRISI, Mirko; POLLASTRI, Gianluca; LE, Quan. *Deep learning methods in protein structure prediction*. *Computational and Structural Biotechnology Journal*, 18, 2020. 1301-1310.

VASWANI, Ashish et al. *Attention is All you Need*. *Advances in Neural Information Processing Systems*, 30, 2017. 6000–6010.

VERLI, Hugo. *Bioinformática: da biologia à flexibilidade molecular*.

VERT, Jean-Philippe. How will generative AI disrupt data science in drug discovery?. *Nature Biotechnology*, 41, n. 6, 2023. 750-751.

WATSON, Joseph L. et al. *De novo design of protein structure and function with RFDiffusion*. *Nature*, 620, n. 7976, 2023. 1089-1100.

XIE, Xuezhi; VALIENTE, Pedro A.; KIM, Philip M. *HelixGAN a deep-learning methodology for conditional de novo design of α -helix structures*. *Bioinformatics*, 39, n. 1, 2023.

ZENG, Xiangxiang et al. *Deep generative molecular design reshapes drug discovery*. *Cell Reports Medicine*, 3, n. 12, 2022. 100794.

ZHANG, Shuang et al. *Applications of transformer-based language models in bioinformatics: a survey*. *Bioinformatics Advances*, 3, n. 1, 2023.