Capítulo 6

Aprendizado de Máquina Aplicado à Descoberta de Fármacos

Karina dos Santos Machado, Adriano Velasque Werhli,
Frederico Kremer, Rafael Junqueira Borges

Abstract

Drug discovery is a long, complex, and expensive process, motivating the development of computational strategies capable of accelerating different stages of this pipeline. In this context, Machine Learning (ML) and Artificial Intelligence (AI) have played an important role in applications such as protein structure prediction, co-folding, virtual screening, molecular docking, and scoring function development. This chapter presents the main theoretical foundations of ML methods applied to rational drug discovery, covering concepts of structural bioinformatics, molecular data representation, classical supervised models, and modern deep learning-based approaches. In addition, collaborative open science initiatives such as CACHE and DREAM Challenge are discussed, highlighting their role in the development and evaluation of computational methods under realistic scenarios.

Resumo

A descoberta de fármacos é um processo longo, complexo e de alto custo, impulsionando o desenvolvimento de estratégias computacionais capazes de acelerar diferentes etapas desse pipeline. Nesse contexto, o Aprendizado de Máquina (AM) e a Inteligência Artificial (IA) têm desempenhado um papel cada vez mais relevante em aplicações como predição estrutural de proteínas e complexos, VS, docking molecular, desenvolvimento de FEs. Este capítulo apresenta os principais fundamentos teóricos de métodos de AM aplicados à descoberta racional de fármacos, abordando conceitos de bioinformática estrutural, representação de dados moleculares, modelos supervisionados clássicos e abordagens modernas baseadas em aprendizado profundo. Além disso, são discutidas iniciativas colaborativas de ciência aberta, como CACHE e DREAM Challenge, que têm impulsionado o desenvolvimento e avaliação de métodos computacionais em cenários realistas.

6.1. Introdução

A necessidade de inovação na descoberta de novos fármacos é urgente, especialmente diante de problemas de saúde pública de escala global, como pandemias, resistência antimicrobiana, crescimento anual dos casos de câncer e doenças genéticas ainda sem tratamento conhecido [Murray et al. 2022, Bray et al. 2024, Kernohan and Boycott 2024]. Todos esses cenários destacam a necessidade urgente de desenvolver novas estratégias para a busca de fármacos. Apesar de todos esses problemas de saúde mundial, a proposta de um novo fármaco ainda é um processo longo e caro. Estima-se que sejam necessários 10-15 anos para que uma nova droga seja aprovada e um investimento médio é de aproximadamente 2,3 bilhões de dólares. Como alternativa para acelerar e reduzir custos desse processo, o uso de ferramentas computacionais tem se mostrado promissor.

Recentemente, avanços significativos foram alcançados pela Ciência da Computação - principalmente em **Inteligência Artificial** (IA) - e pela área de Desenho Racional de fármacos (*Rational Drug Design* - RDD). O RDD pode ser categorizado em dois tipos: planejamento de fármacos baseado na estrutura do ligante (*Ligand-Based Drug Design* - LBDD) e planejamento de fármacos baseado na estrutura do receptor (*Structure-Based Drug Design* - SBDD) [Lima et al. 2016]. A abordagem SBDD, o foco desse capítulo, é baseada no conhecimento da estrutura tridimensional do alvo biológico, usualmente uma proteína (receptor) e como ele interage com o(s) candidato(s) a inibidor (ligante).

Em 2024, os avanços de IA no SBDD foram notáveis, destacando-se os prêmios Nobel de Física para G. Hinton e J. Hopfiled pela proposta das Redes Neurais Artificiais e o prêmio Nobel de Química, para D. Baker pelo método computacional para projetar proteínas *de novo*, além de D. Hassabis e J. Jumper pela proposta do algoritmo de predição de estrutura de proteínas Alphafold2 [Xu 2024, Jumper et al. 2021]. Esses avanços têm permitido o desenvolvimento de métodos aplicados na indústria farmacêutica durante os primeiros estágios de descoberta de um novo fármaco [Kuntz 1992, Meng et al. 2011].

Nesse contexto, a **triagem virtual** (*Virtual Screening* - VS) é uma técnica computacional que permite avaliar e priorizar grandes bibliotecas de compostos para identificar potenciais candidatos a fármacos, geralmente utilizando ***docking*** (docagem) **molecular**. O *docking* molecular consiste em uma simulação em nível atômico, onde milhares de orientações e conformações do ligante no sítio de ligação do receptor são avaliadas e ranqueadas de acordo com a estabilidade do complexo estimada por uma **energia livre de ligação** (*Free Energy of Binding* - FEB) ou escore, predita/calculada por uma função de escore (FE) [Lybrand 1995, Crampon et al. 2022] .

Os desafios e avanços na área de VS com uso de **Aprendizado de Máquina** (AM) têm sido destacados em iniciativas como o **Challenge** [Li et al. 2024], uma competição que explora o uso de IA em VS para busca de novas moléculas promissoras para alvos de doenças complexas. Nas duas primeiras edições do **CACHE**, foram abordadas doença como Parkinson (#1) e Covid-19 (#2), os quais os participantes utilizaram estratégias de AM para selecionar compostos candidatos promissores. No Cache Challenge #2, os participantes tinham que encontrar moléculas promissoras para a proteína NSP13 do Sars-Cov-2. A equipe composta pelos proponentes deste minicurso foi a vencedora entre 22 times participantes. Durante o desafio que ocorreu de 2022 a 2024, foram selecionadas 46 moléculas promissoras determinadas após ex-

tensivo conjunto de testes experimentais, sendo 8 dessas indicadas pelo nosso grupo [Herasymenko et al. 2025, CACHE Initiative 2024]. Nosso método combinou o uso de ferramentas *open-source* e métodos de AM, sendo a única equipe do Sul Global e a única liderada por uma mulher.

Há também o **DREAM Challenge**, uma iniciativa dentro do consórcio **MAINFRAME** (*MAchine learning Innovation Network For Research to Advance MEdicinal chemistry*) [Edwards et al. 2025] que promove competições abertas voltadas à aplicação de AM e IA na descoberta de fármacos. Essa plataforma segue os princípios da ciência aberta, oferecendo grandes conjuntos de dados experimentais e computacionais sobre proteínas, ligantes e resultados de ensaios biológicos para que equipes do mundo inteiro possam propor e comparar modelos preditivos. O objetivo é acelerar a inovação na área de descoberta de fármacos baseada em dados, estimulando a reprodutibilidade científica, a colaboração entre grupos acadêmicos e industriais, e a criação de soluções de IA transparentes e acessíveis. Além disso, o DREAM Challenge/MAINFRAME busca integrar dados heterogêneos de diferentes fontes para permitir o desenvolvimento de modelos robustos e generalizáveis, consolidando-se como um ambiente de referência para o uso ético e colaborativo de IA em bioinformática e saúde.

Além das competições, Xu *et al*. destacam uma lista de exemplos de descoberta de fármacos baseados em IA, já em estudo clínico, mostrando o sucesso desse tipo de estratégia [Xu 2024]. No trabalho publicado recentemente, Xu *et al*. relatam o desenvolvimento do rentosertib, um inibidor da quinase TNIK identificado por meio de algoritmos de IA generativa, desde o desenho molecular até a otimização estrutural e seleção de candidatos para ensaio clínico [Xu et al. 2025]. Uma das moléculas selecionadas neste trabalho tem apresentado resultados clínicos promissores em pacientes com fibrose pulmonar idiopática (IPF). Esse é um dos primeiros casos documentados em que um fármaco proposto a partir de diferentes algoritmos de IA avança até a fase 2 de testes clínicos, evidenciando que modelos generativos e *pipelines* baseados em AM podem produzir candidatos farmacologicamente eficazes e seguros.

Nos últimos anos, o avanço de arquiteturas de **aprendizado profundo**, **modelos generativos** e **modelos de linguagem de larga escala** (*Large Language Models* – LLMs) tem ampliado significativamente as possibilidades da descoberta de fármacos assistida por IA. Essas abordagens têm sido aplicadas em diferentes etapas do *pipeline* de desenvolvimento de fármacos, incluindo predição de estrutura de proteínas-alvo, geração de novas moléculas, VS, desenvolvimento de FEs e predição de propriedades farmacocinéticas e toxicológicas [Jumper et al. 2021, Abramson et al. 2024, Crampon et al. 2022, Xu 2024]. Além disso, a disponibilidade crescente de grandes bases de dados biológicos e químicos, associada ao aumento do poder computacional, tem permitido o treinamento de modelos cada vez mais robustos e generalizáveis. Apesar desses avanços, ainda existem desafios importantes relacionados à interpretabilidade dos modelos, viés nos conjuntos de dados, custo computacional e capacidade de generalização para sistemas biológicos complexos. Nesse contexto, compreender os fundamentos dessas abordagens e suas aplicações práticas torna-se essencial para a formação de pesquisadores capazes de desenvolver e aplicar soluções computacionais modernas na descoberta racional de fármacos.

6.1.1. Objetivo

Apresentar, de forma teórica e prática, a aplicação de Aprendizado de Máquina na descoberta de fármacos, com enfoque na preparação de equipes e divulgação no Brasil de competições com iniciativa ciência aberta (*Open Science*) para esse fim.

6.1.2. Organização do Capítulo

Este capítulo está organizado da seguinte forma: inicialmente, a Seção *Referencial Teórico* apresenta conceitos fundamentais relacionados à bioinformática estrutural, dados biológicos, Aprendizado de Máquina, e iniciativas colaborativas como CACHE e DREAM Challenge. Posteriormente, a Seção *Estratégias Computacionais para a descoberta de fármacos* descreve os métodos de *docking* molecular, com destaque para os diferentes tipos de FEs e os principais métodos de Co-folding em uso atualmente. Em seguida, a Seção *Estudo de Caso: WDR91* apresenta de forma resumida um estudo de caso para o alvo WDR91, demonstrando o uso integrado de ferramentas computacionais e métodos de IA em um problema real de descoberta de fármacos. Por fim, a seção *Considerações finais* discute os desafios atuais e as perspectivas futuras para o uso de IA e AM na área.

6.2. Referencial Teórico

Esta subseção apresenta o referencial teórico para fundamentar as abordagens computacionais discutidas ao longo do capítulo, situando a descoberta de fármacos no contexto da bioinformática estrutural, da disponibilidade de dados moleculares e da aplicação de métodos de AM. Inicialmente, são introduzidos os conceitos centrais da bioinformática estrutural, enfatizando o papel da estrutura tridimensional na compreensão da função biológica. Em seguida, são descritos os principais tipos de dados utilizados na área, incluindo bases de dados de receptores, moléculas e interações experimentais, bem como os formatos e representações computacionais empregados para sua manipulação. A seção também aborda, de maneira introdutória, técnicas de AM relevantes para a descoberta de fármacos, desde a seleção de atributos e modelos supervisionados clássicos até abordagens modernas baseadas em aprendizado profundo, além de estratégias adequadas de validação. Por fim, são apresentadas iniciativas de avaliação colaborativa, como as competições CACHE e DREAM, que posiciona o uso dessas metodologias em cenários realistas e desafiadores.

6.2.1. Bioinformática estrutural

A bioinformática estrutural é uma área interdisciplinar que integra conceitos e métodos da Biologia Estrutural, da Química Computacional, da Física e da Computação. Esta área de estudos dedica-se à análise e modelagem computacional de estruturas tridimensionais de macromoléculas biológicas, partindo do princípio de que a função biológica está intimamente ligada à estrutura tridimensional (3D) das moléculas [Branden and Tooze 2012, Petsko and Ringe 2004]. A forma espacial assumida por proteínas e ácidos nucleicos determina propriedades fundamentais como especificidade molecular, afinidade de ligação e atividade catalítica, uma vez que processos biológicos tem origem diretamente da disposição dos átomos e resíduos no espaço. Alterações estruturais, sejam elas induzidas por mutações, interações intermoleculares ou efeitos dinâmicos, podem resultar em mudanças significativas de função, evidenciando a inseparabilidade

entre estrutura tridimensional e função biológica em nível molecular.

Embora grande parte das aplicações em descoberta de fármacos concentre-se em proteínas globulares, a bioinformática estrutural abrange um escopo mais amplo. A área dedica-se também ao estudo estrutural de ácidos nucleicos, como DNA e RNA, incluindo ribozimas, ribossomos e complexos ribonucleoproteicos, bem como de complexos macromoleculares, membranas biológicas e sistemas biomoleculares heterogêneos [McCammon and Harvey 1987]. Diferentes representações estruturais podem ser empregadas, variando desde modelos atômicos completos, nível de átomos, até descrições *coarse-grained*[1], dependendo da escala espacial e temporal do fenômeno de interesse. Do ponto de vista metodológico, a bioinformática estrutural engloba um conjunto diversificado de técnicas computacionais, entre as quais se destacam a modelagem por homologia, a comparação e alinhamento estrutural, o *docking* molecular (veja na subseção 6.3.1), a Dinâmica Molecular clássica e acelerada, simulações de Monte Carlo, análises de flexibilidade e estudos de estabilidade estrutural. Essas abordagens permitem investigar não apenas estruturas estáticas, mas também aspectos dinâmicos e funcionais dos sistemas biológicos, consolidando a bioinformática estrutural como uma área central para a compreensão mecanicista de processos moleculares e para o desenvolvimento racional de fármacos.

A estrutura tridimensional das biomoléculas fornece a base para a identificação e caracterização de sítios funcionais, como bolsões de ligação, cavidades catalíticas e regiões de interação proteína-ligante. A descrição espacial detalhada dessas estruturas também possibilita a determinação dos tipos de interações atômicas que estabilizam os complexos biomoleculares, incluindo ligações de hidrogênio, interações hidrofóbicas, forças eletrostáticas e empilhamentos aromáticos, fatores determinantes tanto da afinidade quanto da seletividade molecular [Petsko and Ringe 2004]. Do ponto de vista computacional, esses modelos estruturais 3D constituem a entrada essencial para métodos amplamente empregados na bioinformática estrutural. Em particular, o *docking* molecular explora a complementaridade geométrica e energética entre ligantes e alvos biológicos, enquanto a Dinâmica Molecular investiga a relação entre conformação, flexibilidade e função ao longo do tempo, permitindo capturar efeitos que não são acessíveis por abordagens puramente estáticas. Além disso, a estrutura tridimensional é central em análises como comparação estrutural, modelagem por homologia, avaliação de mutações e estudos de estabilidade, consolidando-se como um elemento fundamental nos *pipelines* modernos de bioinformática estrutural e descoberta de fármacos.

Como exemplo representativo de organização estrutural, são apresentadas em maior detalhe as proteínas, que figuram entre as macromoléculas mais extensivamente estudadas na bioinformática estrutural, tanto pela diversidade de suas funções biológicas quanto pela ampla disponibilidade de dados estruturais tridimensionais. As proteínas podem ser descritas em diferentes níveis hierárquicos, tradicionalmente classificados como estrutura primária, secundária, terciária e quaternária [Branden and Tooze 2012]. Essa hierarquia estrutural está representada da esquerda para direita na Figura 6.1. A estrutura primária corresponde à sequência linear de aminoácidos, cuja composição e ordem

[1] resolução reduzida para conjuntos de átomos, possibilitando a captura de fenômenos em escalas maiores

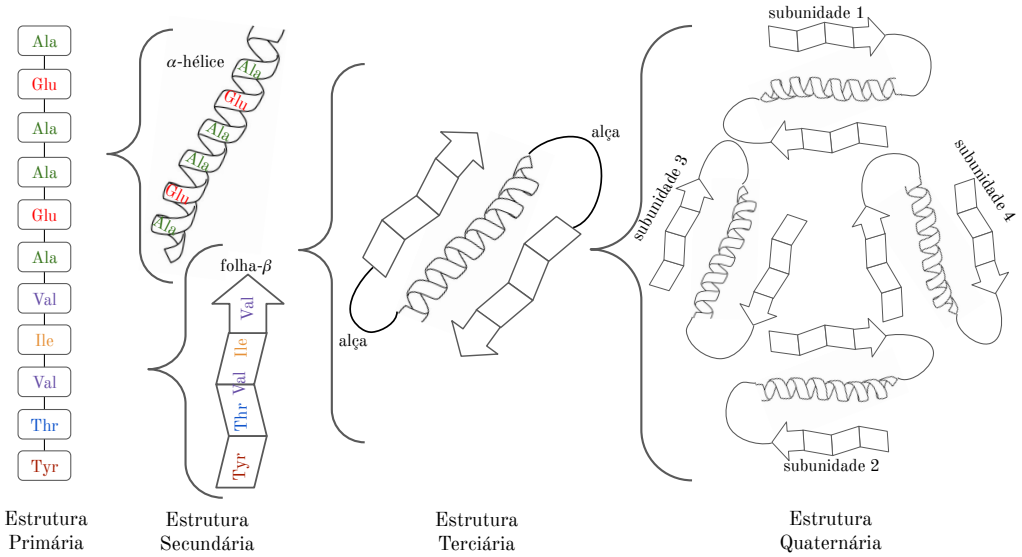

Figura 6.1: Organização estrutural das proteínas. Da esquerda para direita são apresentadas representações das estruturas primária, secundária, tericiária e quaternária. A estrutura primária consiste na sequência específica de aminoácidos. A cadeia peptídica resultante pode se dobrar em uma hélice-α (alfa-hélice) ou em uma folha-β que são tipos de estrutura secundária. Estes segmentos são incorporados à estrutura terciária da cadeia polipeptídica dobrada e conectados por alças. A estrutura quaternária é composta por subunidades de estruturas terciárias fomando complexos proteícos.

determinam as possibilidades conformacionais da molécula. A partir dessa sequência, interações locais entre resíduos dão origem às estruturas secundárias, como hélices-α e folhas-β, estabilizadas principalmente por ligações de hidrogênio. Por sua vez, a estrutura terciária emerge do arranjo tridimensional completo da cadeia polipeptídica, resultante da interação entre elementos secundários e de forças físico-químicas de maior alcance, como interações hidrofóbicas, eletrostáticas e ligações dissulfeto. Em muitas proteínas funcionalmente relevantes, a atividade biológica emerge não apenas da conformação de uma única cadeia, mas da associação específica entre múltiplas subunidades, caracterizando a estrutura quaternária. Esse nível estrutural descreve a organização espacial de complexos proteicos formados por duas ou mais cadeias polipeptídicas, cujas interações são essenciais para processos como regulação alostérica, estabilidade estrutural e reconhecimento molecular.

6.2.2. Dados

Esta subseção discutirá os diferentes tipos de dados utilizados em estudos de descoberta de fármacos assistida por computador. Serão abordados dados estruturais de proteínas (experimentais e modelados), dados químicos de pequenas moléculas, dados experimentais de interação proteína-ligante e dados derivados de triagens químicas de larga escala.

6.2.2.1. Bancos de dados de receptores

Fundado em 1971 como o primeiro arquivo digital de biologia de acesso aberto, o Protein Data Bank (PDB) é o repositório global primário de dados estruturais experi-

mentais de macromoléculas biológicas. Atualmente mantido pelo consórcio internacional **Worldwide Protein Data Bank** (wwPDB) (https://www.wwpdb.org/), o repositório opera através de um sistema colaborativo onde pesquisadores de todo o mundo depositam suas descobertas empíricas, derivadas majoritariamente de cristalografia de raios X, Ressonância Magnética Nuclear (RMN) e criomicroscopia eletrônica. Em termos de funcionamento, cada submissão passa por um rigoroso processo de validação, anotação e padronização biocuratorial antes de ser disponibilizada publicamente, garantindo a integridade e reprodutibilidade científica.

O banco disponibiliza uma vasta gama de dados que vão além das coordenadas atômicas tridimensionais (eixos x, y, z). Os arquivos (disponibilizados primariamente no formato moderno *mmCIF* e no formato legado *PDB*) incluem estruturas primárias, secundárias, terciárias e complexos quaternários de proteínas, ácidos nucleicos (DNA/RNA) e complexos proteína-ligante fundamentais para o desenho de fármacos. Além da geometria da molécula, os dados contêm metadados experimentais detalhados, como fatores de temperatura (*B-factor*), métricas de resolução, informações sobre a taxonomia do organismo de origem e condições do ensaio bioquímico.

Mais recentemente, acompanhando a evolução da IA, o PDB expandiu sua infraestrutura para abrigar modelos preditos computacionalmente através do portal *Computed Structure Models* (CSM). Dentre os principais provedores de CSMs destaca-se o *AlphaFold Protein Structure Database* (AFDB) (https://alphafold.ebi.ac.uk/), fruto da parceria entre a Google DeepMind e o *European Bioinformatics Institute* (EMBL-EBI). O AFDB revolucionou a disponibilidade de dados estruturais ao fornecer modelos preditos com alta confiança para mais de 200 milhões de sequências catalogadas no UniProt usando o algoritmo AlphaFold2, preenchendo lacunas cruciais no RDD para alvos que ainda não possuem resolução empírica.

6.2.2.2. Bancos de dados de moléculas químicas

A exploração do espaço químico, a VS e o treinamento de modelos de AM dependem da disponibilidade de bancos de dados de moléculas. Essas plataformas catalogam desde compostos prontos para síntese até fármacos aprovados e suas respectivas atividades biológicas, fornecendo os dados de entrada (*inputs*) essenciais para campanhas de VS e modelagem preditiva. Abaixo, destacam-se os principais repositórios utilizados na área:

ZINC (https://zinc.docking.org/): Desenvolvido e mantido pela Universidade da Califórnia, São Francisco (UCSF), o ZINC é um banco de dados gratuito focado em coleções de compostos disponíveis comercialmente, otimizados especificamente para VS. Suas versões mais recentes (como ZINC15 e ZINC20) catalogam bilhões de moléculas, oferecendo representações em formatos tridimensionais (como MOL2 e SDF) já preparados para *docking*, além de anotações sobre propriedades físico-químicas e viabilidade biológica.

PubChem (https://pubchem.ncbi.nlm.nih.gov/): Mantido pelo National Center for Biotechnology Information (NCBI), o PubChem atua como um repositório aberto massivo que integra estruturas químicas a dados de ensaios biológicos. A arquitetura de dados da plataforma é estruturada em três bancos primários interconectados, operando

com diferentes níveis de curadoria e aplicação: *Substances*, *Compounds* e *BioAssays*.

O ***PubChem Substances*** armazena registros químicos brutos exatamente como submetidos pelos depositantes originais, abrigando misturas complexas, extratos não caracterizados e registros redundantes sob identificadores SID (*Substance ID*). O ***PubChem Compounds*** extrai e padroniza os dados da base *Substances* para gerar representações de moléculas químicas únicas e estruturalmente validadas, designando a cada uma um identificador CID (*Compound ID*). Esta etapa elimina ambiguidades e centraliza o espaço químico rastreável. O ***PubChem BioAssays*** consolida os resultados de testes biológicos, vinculando CIDs e SIDs a painéis de toxicidade, ensaios enzimáticos e projetos de triagem de alto rendimento (*High-Throughput Screening* - HTS), registrando tanto métricas quantitativas empíricas (IC_{50}, EC_{50}, K_i) quanto classificações categóricas (ativo/inativo/inconclusivo).

A integração sistemática do módulo *BioAssays* com o *Compounds* serve de base para o treinamento de modelos QSAR (*Quantitative Structure-Activity Relationship*). No fluxo de trabalho de AM, os dados do *BioAssays* fornecem os rótulos de classe ou as variáveis contínuas alvo (*ground truth* experimental) essenciais para a aprendizagem supervisionada.

Catálogos de Fornecedores (Enamine, MolPort e outros): Além dos repositórios acadêmicos, os catálogos de empresas químicas fornecedoras desempenham um papel crucial para garantir que as moléculas preditas possam ser fisicamente testadas. A **Enamine** destaca-se por fornecer bibliotecas massivas baseadas em síntese química sob demanda (*make-on-demand*), como o *REAL Space*, que expande o espaço químico explorável para dezenas de bilhões de compostos. O **MolPort**, por sua vez, atua como um agregador global, consolidando catálogos de múltiplos fornecedores em uma única interface, o que facilita a aquisição logística de moléculas promissoras (*hits*) para ensaios *in vitro*.

ChEMBL (https://www.ebi.ac.uk/chembl/): Gerenciado pelo European Bioinformatics Institute (EMBL-EBI), o ChEMBL é um banco de dados de quimioinformática focado em compostos bioativos com propriedades do tipo fármaco (*drug-like*). Seu grande diferencial é a curadoria de dados extraídos manualmente da literatura científica primária, fornecendo medidas quantitativas rigorosas de afinidade de ligação, como constantes de dissociação (K_d), constantes de inibição (K_i) e concentrações inibitórias meias-máximas (IC_{50}). Essa riqueza de dados experimentais torna o ChEMBL uma das bases mais importantes para o treinamento de FEs empíricas e modelos de regressão em AM. Dados do ChEMBL e PubChem são geralmente complementares e interconectados.

DrugBank (https://go.drugbank.com/): Consiste em um banco de dados abrangente que combina quimioinformática e farmacologia, contendo informações detalhadas sobre fármacos aprovados, compostos em fase de investigação clínica e farmacêuticos. O DrugBank integra a estrutura química de pequenas moléculas com anotações biológicas profundas, incluindo mecanismos de ação, vias de metabolismo, interações medicamentosas e o mapeamento preciso de seus respectivos alvos proteicos.

Therapeutic Target Database (TTD) (https://ttd.idrblab.cn/): Diferentemente das bases puramente químicas, o TTD foca na relação bidirecional entre alvos terapêuti-

cos (proteínas, ácidos nucleicos) e as moléculas direcionadas a eles. O banco fornece informações detalhadas sobre alvos associados a doenças específicas, o *status* de aprovação ou fase de ensaio clínico dos compostos correspondentes e as vias bioquímicas envolvidas. O TTD é uma ferramenta particularmente valiosa para estudos de reposicionamento de fármacos (*drug repurposing*) e avaliação de perfis de eficácia estrutural.

6.2.2.3. Bases de Dados experimentais de interação

Para o desenvolvimento de modelos preditivos baseados em estrutura, é imperativo o mapeamento direto entre a conformação tridimensional de um complexo biomolecular e sua respectiva força de interação experimental [Wang et al. 2004]. Nesse contexto, bases de dados especializadas desempenham papel central ao integrar informações estruturais e biofísicas utilizadas no treinamento e validação de modelos de IA aplicados à descoberta de fármacos.

O **PDBbind** (https://www.pdbbind-plus.org.cn/) constitui o principal repositório global curado para integração entre estruturas tridimensionais e dados quantitativos de afinidade molecular [Wang et al. 2004]. A base extrai sistematicamente complexos biomoleculares depositados no PDB, incluindo sistemas proteína-ligante, proteína-proteína e proteína-ácido nucleico, associando essas estruturas a medições experimentais reportadas na literatura primária. Entre os parâmetros documentados destacam-se K_d, K_i e IC_{50}.

A arquitetura do PDBbind é organizada em diferentes níveis de curadoria e rigor experimental. O *general set* contém o conjunto mais amplo de complexos disponíveis; o *refined set* aplica filtros rigorosos relacionados à resolução cristalográfica, consistência estrutural e qualidade molecular; enquanto o *core set*, também denominado CASF *benchmark* (*Comparative Assessment of Scoring Functions*), é amplamente utilizado para *benchmarking* e validação comparativa de FEs e modelos preditivos [Su et al. 2018].

No ecossistema de AM aplicado à descoberta de fármacos, o PDBbind atua como uma das principais fontes de dados rotulados (*ground truth*) [Wang et al. 2004]. A geometria atômica tridimensional do complexo fornece a entrada para extração de descritores espaciais, modelagem baseada em grafos e representação volumétrica tridimensional, enquanto os valores experimentais de afinidade constituem o sinal supervisionado utilizado no treinamento de redes neurais convolucionais 3D, modelos de regressão e FEs híbridas baseadas em IA [Jiménez et al. 2018].

O **BindingDB** (https://www.bindingdb.org) é um banco de dados público especializado em afinidades de ligação experimentalmente determinadas entre proteínas alvo e pequenas moléculas bioativas [Liu et al. 2007]. Diferentemente do PDBbind, a inclusão de entradas no BindingDB não depende da disponibilidade de estruturas cristalográficas tridimensionais resolvidas.

A plataforma agrega dados oriundos de ensaios bioquímicos, testes enzimáticos e medições biofísicas, incluindo valores de K_i, K_d, IC_{50} e constantes cinéticas, extraídos sistematicamente da literatura científica e de documentos de patentes [Gilson et al. 2015]. Essa característica amplia significativamente o volume de dados disponíveis para treinamento de modelos QSAR bidimensionais, redes neurais baseadas em sequências proteicas

e arquiteturas fundamentadas exclusivamente em grafos moleculares.

Além disso, devido ao seu elevado volume de interações anotadas experimentalmente, o BindingDB tornou-se uma importante fonte de dados para aplicações de *deep learning*, aprendizado multitarefa e modelos de predição de afinidade proteína-ligante em larga escala [Gilson et al. 2015].

AIRCHECK (Artificial Intelligence-Ready CHEmiCal Knowledge-base) https://aircheck.ai/, é uma iniciativa aberta desenvolvida para disponibilizar dados químicos estruturados e padronizados voltados ao desenvolvimento e avaliação de modelos de AM aplicados à descoberta de fármacos. A plataforma integra dados experimentais provenientes de Bibliotecas Codificadas por DNA (*DNA-Encoded Libraries* - DEL), Espectrometria de Massas com Seleção por Afinidade (*Affinity Selection Mass Spectrometry* - ASMS) e validação de compostos, incluindo exemplos positivos e negativos de interação molecular, além de fornecer recursos para treinamento, *benchmark* e avaliação de modelos preditivos em cenários mais realistas.

6.2.2.4. Métodos Experimentais para Ensaios de *High-Throughput Screening* (HTS)

6.2.2.4.1. Bibliotecas Codificadas por DNA (DEL)

A tecnologia de DEL revolucionou a escala da triagem física de compostos químicos [Goodnow et al. 2016, Gironda-Martínez et al. 2021]. Diferentemente dos paradigmas clássicos, nos quais moléculas são avaliadas individualmente, as DELs permitem a síntese combinatória em esquema *split-and-pool*, combinando sequencialmente diferentes *building blocks* químicos em múltiplos ciclos reacionais (Figura 6.2).

Teoricamente, considerando a utilização de 10^3 *building blocks* distintos em dois ciclos sucessivos de combinação química, o espaço químico gerado alcançaria aproximadamente 10^6 moléculas únicas. Em bibliotecas mais complexas, contendo múltiplas etapas sintéticas adicionais, esse número pode atingir escalas superiores a bilhões de compostos [Goodnow et al. 2016, Franzini et al. 2014].

Cada molécula sintetizada é fisicamente conjugada a uma sequência curta e única de DNA, que atua como um código de barras molecular identificador, permitindo a rastreabilidade do histórico sintético e a identificação posterior dos compostos enriquecidos durante os ensaios de afinidade.

Durante o ensaio de afinidade, a biblioteca completa é incubada com a proteína alvo imobilizada. Compostos sem afinidade significativa são removidos por lavagem, enquanto os ligantes retidos permanecem associados ao alvo biológico. Após dissociação do complexo, os códigos de DNA correspondentes aos compostos enriquecidos são amplificados via PCR e identificados por Sequenciamento de Nova Geração (*Next-Generation Sequencing* – NGS) [Franzini et al. 2014].

No contexto do AM, ensaios DEL produzem matrizes de dados massivas contendo contagens de enriquecimento associadas à ligação molecular. O enorme volume de dados permite o treinamento de modelos robustos de classificação, VS e aprendi-

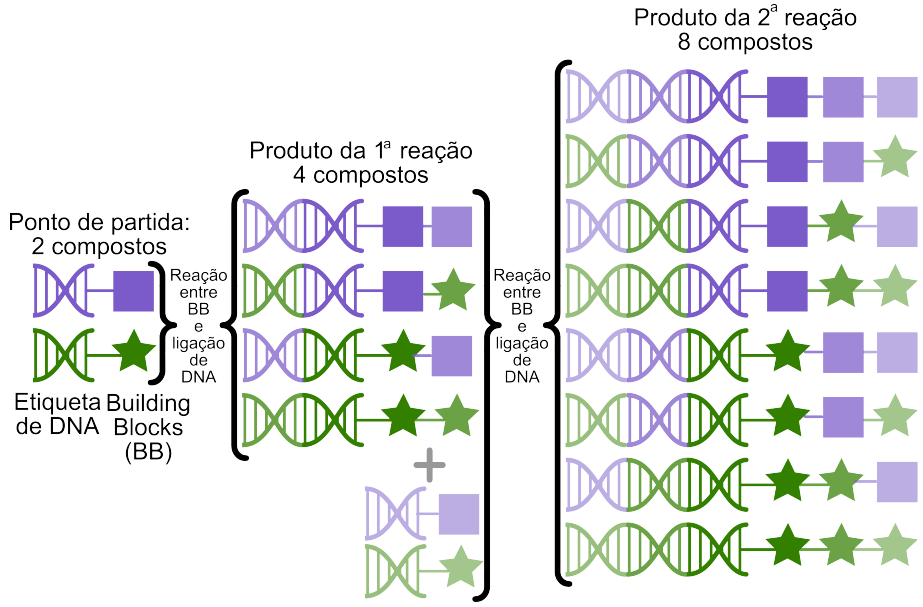

Figura 6.2: Representação esquemática da estratégia de síntese combinatória *split-and-pool* empregada em Bibliotecas Codificadas por DNA (DELs). Inicialmente, diferentes *building blocks* químicos são conjugados a sequências específicas de DNA que atuam como códigos de barras moleculares. Após cada ciclo de reação química, os compostos são reunidos (*pool*), redistribuídos (*split*) e submetidos a novas etapas de síntese com conjuntos adicionais de fragmentos químicos. O processo iterativo permite a geração exponencial de diversidade molecular, possibilitando a construção de bibliotecas contendo milhões a bilhões de compostos distintos em um único experimento.

zado profundo, inclusive em cenários sem estruturas tridimensionais explícitas disponíveis [Goodnow et al. 2016]. Resultados recentes apontam a descoberta de ligantes na ordem de interação de µM, requerendo ciclos de otimização [Ahmad et al. 2023].

Apesar de seu alto poder de escala, esses dados apresentam características que exigem cuidado na análise computacional. O sinal experimental pode ser influenciado por fatores como eficiência de síntese, qualidade da biblioteca, interações inespecíficas, efeitos do DNA conjugado, abundância inicial dos compostos e ruído de sequenciamento. Além disso, a definição de compostos positivos e negativos depende de critérios de enriquecimento e de processamento dos dados. Portanto, embora DEL seja uma fonte extremamente valiosa de dados para AM, esses dados devem ser interpretados como evidências experimentais de priorização, e não como medidas diretas e definitivas de afinidade.

6.2.2.4.2. Espectrometria de Massas com Seleção por Afinidade (ASMS)

Diferentemente de ensaios dependentes de fluoróforos, marcadores radioativos ou códigos de DNA, a ASMS constitui uma abordagem *label-free*, eliminando potenciais interferências físico-químicas decorrentes da marcação molecular [Muchiri and van Breemen 2020]. O método consiste na incubação da proteína alvo com misturas contendo centenas ou milhares de pequenas moléculas simultaneamente.

Posteriormente, técnicas como cromatografia de exclusão por tamanho (*Size Exclusion Chromatography* – SEC) ou ultrafiltração molecular são empregadas para separar

rapidamente os complexos proteína-ligante das moléculas livres em solução. Em seguida, os ligantes associados são liberados sob condições desnaturantes e identificados por espectrometria de massas de alta resolução [Prudent et al. 2023].

Uma variação recente dessa abordagem é a **eASMS** (*enantioselective* ASMS). Frequentemente, pequenas moléculas quirais existem como pares de enantiômeros, isto é, formas moleculares que possuem a mesma composição química e conectividade, mas que se organizam no espaço como imagens especulares não sobreponíveis. Como proteínas também são estruturas tridimensionais quirais (formada por L-aminoácidos - levógiros), seus sítios de ligação podem reconhecer de forma preferencial um enantiômero em relação ao outro. Assim, a observação de ligação enantiosseletiva fornece uma evidência adicional de que a interação detectada é específica e dependente de complementaridade molecular tridimensional.

Bibliotecas ricas em compostos quirais são particularmente úteis porque pares de enantiômeros funcionam como controles internos muito próximos, que compartilham massa, fórmula molecular e muitas propriedades físico-químicas, mas diferem na orientação espacial de seus grupos químicos. Portanto, diferenças consistentes de enriquecimento entre enantiômeros podem ajudar a distinguir interações específicas de artefatos inespecíficos de retenção, agregação ou ruído experimental, elevando a taxas de verdadeiros positivos.

Essa característica torna a eASMS especialmente atraente para a descoberta de ligantes fracos ou iniciais, que são frequentes em campanhas voltadas a alvos pouco explorados. Em um estudo recente, essa abordagem foi aplicada a 31 proteínas humanas utilizando uma biblioteca de mais de 8 mil compostos quirais, permitindo identificar 16 ligantes fracos para 12 alvos desafiadores, gerar evidências de seletividade e oferecer confirmação ortogonal de ligação [Wang et al. 2025]. Os autores destacam que a eASMS é útil para identificar e caracterizar ligantes seletivos para proteínas previamente sem ligantes conhecidos ou consideradas desafiadoras para descoberta química.

Entretanto, assim como em DEL, os dados de eASMS exigem interpretação cuidadosa. Um sinal de enriquecimento não deve ser confundido automaticamente com uma medida quantitativa de afinidade, como um valor de K_d. A intensidade observada pode depender de condições experimentais, composição da biblioteca, concentração da proteína, competição entre compostos, eficiência de separação e sensibilidade da espectrometria de massas. Portanto, *hits* identificados por eASMS geralmente devem ser confirmados por métodos ortogonais, como Ressonância Plasmônica de Superfície (*Surface Plasmon Resonance* - SPR), ITC, Ensaios de Mudança Térmica (*Thermal Shift Assay* - TSA), ensaios funcionais ou estudos estruturais.

Para aplicações em IA, os dados oriundos de ASMS são considerados de alta qualidade experimental devido ao reduzido nível de ruído e à evidência física direta de interação não covalente. Frequentemente, esses conjuntos são utilizados como referência experimental (*gold standard*) para validação de modelos treinados com bases mais volumosas e ruidosas, como DELs [Muchiri and van Breemen 2020]. Dados de DEL e ASMS já vêm sendo disponibilizados para diferentes alvos na plataforma AIRCHECK.

Além de dados estruturais oriundos de complexos previamente cristalizados, como

aqueles disponíveis em bases especializadas a exemplo do PDBBind [Wang et al. 2004], campanhas modernas de HTS produzem volumes massivos de dados experimentais de interação molecular. Esses ensaios desempenham papel central no treinamento de modelos preditivos, pois fornecem o "gabarito" experimental necessário para abordagens supervisionadas, incluindo classificação binária (*ativo* versus *inativo*) e estimativas quantitativas de afinidade. Entre as tecnologias experimentais mais relevantes neste contexto destacam-se metodologias de larga escala e alta precisão, capazes de alimentar arquiteturas modernas de IA aplicada à descoberta de fármacos.

6.2.2.5. Métodos Ortogonais de Validação de Interação

6.2.2.5.1. Ressonância Plasmônica de Superfície

A técnica de SPR constitui um método óptico avançado para monitoramento em tempo real de interações biomoleculares sem necessidade de marcadores fluorescentes [Homola 2008]. Nesse sistema, a proteína alvo é imobilizada sobre um biossensor metálico, geralmente um *chip* de ouro, enquanto soluções contendo os ligantes fluem continuamente por um sistema microfluídico. A interação molecular altera localmente o índice de refração da superfície, produzindo um sinal óptico mensurável instantaneamente.

Diferentemente de ensaios de ponto final, a SPR permite determinar simultaneamente parâmetros termodinâmicos e cinéticos da interação molecular, incluindo a constante de dissociação (K_d), as taxas de associação (k_{on}) e dissociação (k_{off}) [Rich and Myszka 2008]. Em descoberta de fármacos assistida por IA, o chamado "tempo de residência" molecular ($1/k_{off}$) emergiu como um importante parâmetro preditivo de eficácia *in vivo*, motivando o desenvolvimento de modelos focados em dinâmica cinético-estrutural [Copeland et al. 2006].

6.2.3. Formato dos dados

A aplicação de métodos de AM na descoberta de fármacos depende fortemente da forma como as estruturas moleculares e os dados químicos são representados e armazenados. Ao longo das últimas décadas, diferentes formatos de arquivos foram desenvolvidos para atender a necessidades específicas das áreas de bioinformática e quimioinformática. Dentre essas necessidades, destacam-se a descrição tridimensional de macromoléculas, o armazenamento de pequenas moléculas com propriedades associadas e a codificação compacta de estruturas químicas, permitindo o armazenamento eficiente, a interoperabilidade entre ferramentas computacionais e o processamento de grandes volumes de dados em aplicações de modelagem preditiva e VS. Nesta subseção, são apresentados alguns dos principais formatos utilizados em bioinformática estrutural e química computacional; PDB, mmCIF, MOL2, SDF e SMILES.

O formato **PDB** (Protein Data Bank) é o padrão histórico para a representação de estruturas tridimensionais de macromoléculas biológicas, incluindo proteínas, ácidos nucleicos e complexos proteína-ligante, veja um exemplo na FIGURA 6.3. Criado na década de 1970 juntamente com o repositório homônimo, o formato PDB armazena co-

ordenadas atômicas em três dimensões, informações sobre conectividade, elementos químicos, fatores de ocupação e temperatura, além de metadados experimentais, como o método de determinação estrutural (por exemplo, cristalografia por raios X, RMN ou criomicroscopia eletrônica) [Berman et al. 2000, Burley et al. 2021]. Apesar de sua ampla adoção, o formato PDB apresenta limitações conhecidas, como restrições rígidas de largura de campos, ausência de suporte nativo a propriedades químicas detalhadas e descrições incompletas de estados de protonação ou ordens de ligação. Essas limitações motivaram o desenvolvimento do formato mmCIF, mais estruturado e extensível. Ainda assim, arquivos PDB continuam sendo largamente utilizados em *pipelines* de *docking* molecular, modelagem molecular e AM baseado em estruturas tridimensionais de proteínas, principalmente devido à grande quantidade de estruturas disponíveis publicamente [Westbrook and Fitzgerald 2003].

```
ATOM      1  N   VAL A   2     -78.391 -19.186  13.764  1.00 76.19           N
ATOM      2  CA  VAL A   2     -78.926 -20.535  13.907  1.00 80.33           C
ATOM      3  C   VAL A   2     -79.769 -20.649  15.173  1.00 83.98           C
ATOM      4  O   VAL A   2     -79.303 -20.341  16.269  1.00 92.42           O
ATOM      5  CB  VAL A   2     -77.800 -21.580  13.919  1.00 79.68           C
ATOM      6  CG1 VAL A   2     -78.368 -22.972  14.144  1.00 73.58           C
ATOM      7  CG2 VAL A   2     -77.010 -21.523  12.620  1.00 97.52           C
```

Figura 6.3: Trecho ilustrativo de um arquivo no formato PDB referente à helicase NSP13 do SARS-CoV-2 (código PDB: 7NN0), no qual são apresentadas linhas do tipo ATOM contendo a identificação dos átomos, resíduos, cadeias e suas coordenadas tridimensionais.

O formato **mmCIF** (macromolecular Crystallographic Information File) é atualmente o padrão oficial adotado pelo Protein Data Bank para o armazenamento e distribuição de estruturas tridimensionais de macromoléculas biológicas. Diferentemente do formato PDB clássico,o mmCIF utiliza uma estrutura baseada em pares chave-valor organizados em tabelas, permitindo maior flexibilidade, extensibilidade e consistência na descrição dos dados estruturais, veja um exemplo em 6.4. O mmCIF foi projetado para acomodar estruturas macromoleculares de grande porte e elevada complexidade, como complexos multiproteicos, ribossomos, vírus e estruturas determinadas por criomicroscopia eletrônica em alta resolução, que frequentemente excedem os limites do formato PDB. Além das coordenadas tridimensionais dos átomos, arquivos mmCIF permitem a representação mais completa de metadados experimentais, informações cristalográficas, relações de simetria e anotações estruturais, facilitando a interoperabilidade entre bases de dados e ferramentas computacionais.

O formato **MOL2** foi desenvolvido originalmente pela Tripos Associates e é amplamente utilizado para representar pequenas moléculas em contextos de química computacional e modelagem molecular, veja na Figura 6.5. Diferentemente do PDB, o MOL2 inclui informações químicas mais detalhadas, como tipos atômicos específicos do campo de força, ordens de ligação, cargas parciais e identificação explícita de subestruturas, [Tripos 2005]. Essas características tornam o MOL2 particularmente adequado para aplicações como *docking* molecular, parametrização de campos de força e geração de descritores moleculares. No contexto de AM, arquivos MOL2 são frequentemente usados como ponto de partida para extração de grafos moleculares enriquecidos, nos quais átomos e ligações podem ser representados por atributos mais informativos do que aqueles


```
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_alt_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_entity_id
_atom_site.label_seq_id
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.pdbx_PDB_model_num

ATOM 1  N   N   . VAL A 1 2 -78.391 -19.186 13.764 1.00 76.19 1
ATOM 2  C   CA  . VAL A 1 2 -78.926 -20.535 13.907 1.00 80.33 1
ATOM 3  C   C   . VAL A 1 2 -79.769 -20.649 15.173 1.00 83.98 1
ATOM 4  O   O   . VAL A 1 2 -79.303 -20.341 16.269 1.00 92.42 1
ATOM 5  C   CB  . VAL A 1 2 -77.800 -21.580 13.919 1.00 79.68 1
ATOM 6  C   CG1 . VAL A 1 2 -78.368 -22.972 14.144 1.00 73.58 1
ATOM 7  C   CG2 . VAL A 1 2 -77.010 -21.523 12.620 1.00 97.52 1
```

Figura 6.4: Trecho ilustrativo de um arquivo no formato mmCIF (.cif) referente à helicase NSP13 do SARS-CoV-2 (PDB: 7NN0). O bloco `loop_` define a tabela `_atom_site`, que armazena informações equivalentes às linhas `ATOM` do formato PDB, incluindo identificação atômica, resíduo, cadeia e coordenadas tridimensionais.

disponíveis em formatos mais simples [Morris et al. 2009]. Entretanto, o MOL2 não é um formato padronizado por uma entidade internacional, e variações na forma como tipos atômicos e cargas são atribuídos podem impactar a reprodutibilidade entre diferentes ferramentas.

```
@<TRIPOS>ATOM
1 N1   -0.012  1.245  0.000 N.ar  1 CAF  -0.347
2 C2    1.234  0.678  0.000 C.2   1 CAF   0.215
3 N3    1.198 -0.745  0.000 N.ar  1 CAF  -0.289
4 C4   -0.034 -1.312  0.000 C.2   1 CAF   0.181
5 O5   -1.245 -0.845  0.000 O.2   1 CAF  -0.512
@<TRIPOS>BOND
1 1 2 ar
2 2 3 ar
3 3 4 2
4 4 5 2
```

Figura 6.5: Trecho ilustrativo de um arquivo no formato MOL2 representando a molécula de cafeína (PubChem CID: 2519), no qual são explicitadas a conectividade química, os tipos atômicos e as cargas parciais, informações comumente utilizadas em estudos de *docking* molecular e modelagem computacional.

O **SDF** (Structure Data File) é uma extensão do formato MOL desenvolvida pela Molecular Design Limited (MDL) e amplamente adotada para o armazenamento de conjuntos de pequenas moléculas, especialmente em bases de dados químicas e bibliotecas de compostos. Uma de suas principais vantagens é a capacidade de associar, a cada molécula, um conjunto arbitrário de propriedades, como atividades biológicas, valores de afinidade, descritores fisicoquímicos ou informações experimentais [Dalby et al. 1992]. Cada entrada em um arquivo SDF pode conter coordenadas 2D ou 3D, conectividade química e

um bloco de dados estruturados em pares chave–valor. Essa flexibilidade torna o SDF um formato central em tarefas de aprendizado supervisionado, como predição de atividade, toxicidade ou propriedades ADMET, sendo amplamente utilizado em ferramentas como RDKit, Open Babel e *pipelines* de machine learning em química [Landrum 2006]. Como limitação, o SDF tende a ser relativamente verboso, o que pode impactar desempenho e armazenamento em grandes coleções de compostos.

```
2519
  -OEChem-04192618422D

 24 25  0     0  0  0  0  0  0999 V2000
    3.7321    2.0000    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    2.0000   -1.0000    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
6  8  8
7  8  8
7  9  8

$$$$
```

Figura 6.6: Trecho ilustrativo de um arquivo no formato SDF (Structure Data File) representando a molécula de cafeína (PubChem CID: 2519), no qual a estrutura química é acompanhada de propriedades associadas, permitindo o armazenamento conjunto de dados estruturais e informações químicas ou biológicas.

O **SMILES** (Simplified Molecular Input Line Entry System) é uma representação textual linear que codifica a estrutura de uma molécula como uma sequência de caracteres ASCII. Desenvolvido com o objetivo de ser compacto e facilmente legível por máquinas, o SMILES descreve conectividade atômica, ramificações, ordens de ligação e estereoquímica por meio de uma gramática formal bem definida [Weininger 1988]. Devido à sua simplicidade e eficiência, o SMILES tornou-se extremamente popular em aplicações de AM, particularmente em modelos baseados em redes neurais recorrentes, transformers e modelos generativos, onde moléculas podem ser tratadas como sequências ou tokens [Gómez-Bombarelli et al. 2018]. No entanto, uma mesma molécula pode ser representada por múltiplos SMILES equivalentes, o que levou ao desenvolvimento de conceitos como SMILES canônico e randomized SMILES para aumentar a robustez de modelos estatísticos e redes neurais [Bjerrum 2017]. Apesar de não conter explicitamente coordenadas tridimensionais, o SMILES é frequentemente combinado com etapas posteriores de geração de conformações 3D. Como exemplo, a molécula de cafeína é representada em SMILES como: `CN1C=NC2=C1C(=O)N(C(=O)N2C)C`.

6.2.4. *Fingerprints* de moléculas

As *fingerprints* circulares, exemplificadas pelo algoritmo Morgan (*Extended-Connectivity Fingerprints*, ECFP) implementado no RDKit, operam por meio de um processo iterativo que mapeia a vizinhança atômica de maneira concêntrica. O cálculo inicia-se pela atribuição de identificadores numéricos a cada átomo individual com base em propriedades fundamentais, como número atômico, valência, carga formal e hibridização. A cada iteração, ou "raio", o identificador de um átomo é atualizado para incorporar as informações de seus vizinhos imediatos, gerando representações capazes de codificar subestruturas de complexidade crescente [Rogers and Hahn 2010, Landrum 2013]. O principal diferencial dessa abordagem reside em sua elevada capacidade de representar o ambiente químico local

com alta resolução estrutural, sendo amplamente considerada o padrão-ouro para capturar centros reativos e padrões específicos de substituição química [Rogers and Hahn 2010].

Em contraste, as *fingerprints* baseadas em caminhos (*path-based fingerprints*) utilizam uma lógica fundamentada na conectividade linear da molécula. Nesse método, o algoritmo identifica exaustivamente todas as sequências possíveis de átomos e ligações até um comprimento máximo predefinido, geralmente limitado a sete ligações consecutivas [Daylight Chemical Information Systems 2024]. Cada caminho identificado é posteriormente processado por uma função de *hash*, responsável por ativar posições específicas em um vetor binário. Enquanto as representações circulares modelam "conchas" atômicas ao redor de um centro, as *fingerprints* baseadas em caminhos priorizam a topologia do esqueleto molecular e a continuidade das ligações químicas. Consequentemente, essas representações tendem a apresentar desempenho superior na discriminação de famílias de *scaffolds* lineares, sistemas ramificados e arquiteturas moleculares que não possuem um centro radial claramente definido [Cereto-Massagué et al. 2015].

As chaves estruturais de dicionário, exemplificadas pelas *MACCS Keys*, abandonam o paradigma de *hashing* em favor de uma comparação determinística contra um conjunto fixo de fragmentos químicos previamente definidos [Durant et al. 2002]. O cálculo consiste na verificação binária da presença ou ausência de cada um dos 166 padrões estruturais descritos no dicionário químico da ferramenta. O principal diferencial dessa abordagem é sua elevada interpretabilidade, uma vez que cada posição do vetor possui um significado químico explícito e invariável. Entretanto, sua principal limitação em relação às representações circulares ou baseadas em caminhos reside na incapacidade de capturar novidades estruturais, padrões emergentes ou fragmentos químicos que não tenham sido previamente incluídos no dicionário original [Cereto-Massagué et al. 2015].

A lógica das *fingerprints* de pares de átomos (*Atom Pair Fingerprints*) difere substancialmente das abordagens anteriores ao priorizar relações de distância topológica em vez de conectividade local imediata. Para cada par de átomos pesados presentes na molécula, o algoritmo registra simultaneamente o tipo químico dos dois átomos e o número mínimo de ligações que os separa [Carhart et al. 1985]. Dessa forma, a representação passa a capturar aspectos relacionados à forma global e às dimensões topológicas da molécula. Essa estratégia permite que modelos de AM identifiquem similaridades estruturais entre compostos que apresentam esqueletos químicos distintos, mas preservam distribuições espaciais equivalentes de grupos funcionais no espaço bidimensional molecular [Cereto-Massagué et al. 2015].

Por fim, as *fingerprints* de torsão topológica (*Topological Torsion Fingerprints*) concentram-se na descrição da rigidez molecular e de aspectos associados à estereoquímica teórica. O cálculo identifica sistematicamente todas as sequências compostas por quatro átomos consecutivamente ligados, codificando propriedades específicas de cada átomo da sequência, como hibridização, aromaticidade e número de elétrons π [Nilakantan et al. 1987]. Ao focar explicitamente em unidades de quatro átomos, essa abordagem torna-se capaz de distinguir isômeros e padrões conformacionais que poderiam ser considerados equivalentes por *fingerprints* circulares ou baseadas em caminhos. Como consequência, essa representação fornece aos modelos de IA informações relevantes sobre conformação molecular, liberdade rotacional e restrições geométricas internas

da estrutura química [Cereto-Massagué et al. 2015].

6.2.5. Aprendizado de Máquina na descoberta de fármacos

Uma vez estabelecidas as representações moleculares e os conjuntos de dados utilizados na descoberta de fármacos, partimos para a aplicação sistemática de métodos de Aprendizado de Máquina para a construção de modelos preditivos. Nesta subseção, apresentamos uma visão integrada das principais etapas dessa metodologia, abrangendo desde técnicas de seleção de atributos e redução de dimensionalidade, passando por modelos supervisionados clássicos baseados em ensembles, até abordagens modernas de aprendizado profundo, bem como estratégias de validação e avaliação adequadas às particularidades da área de descoberta de fármacos.

6.2.5.1. Seleção de atributos e redução de dimensionalidade

A seleção de atributos constitui uma etapa fundamental em *pipelines* de AM, especialmente no contexto da descoberta de fármacos. As representações computacionais de moléculas, como *fingerprints*, descritores físico-químicos ou vetores derivados de grafos moleculares, frequentemente resultam em espaços de atributos de alta dimensionalidade. Embora tais representações sejam altamente informativas, é comum que contenham atributos redundantes, irrelevantes ou fortemente correlacionados, com potencial de impactar negativamente o desempenho dos modelos preditivos e aumentar o risco de *overfitting*.

Nesse cenário, técnicas de seleção de atributos desempenham um papel central ao identificar subconjuntos de variáveis mais relevantes para a tarefa de predição. A redução adequada do espaço de atributos contribui para melhorar a capacidade de generalização dos modelos, reduzir o custo computacional e facilitar a interpretação dos resultados, aspecto particularmente importante em aplicações de bioinformática estrutural.

De forma geral, as abordagens de seleção de atributos podem ser classificadas em métodos do tipo **filtro**, ***wrapper*** e **embutidos** [Guyon and Elisseeff 2003]. Métodos do tipo **filtro** avaliam a relevância dos atributos de maneira independente do modelo preditivo, utilizando critérios estatísticos como correlação, informação mútua ou testes univariados. Já os métodos ***wrapper*** empregam o desempenho de um modelo específico como critério de seleção, explorando diferentes subconjuntos de atributos. Essa combinação de subconjuntos de atributos inevitavelmente resulta em um maior custo computacional. Por fim, métodos **embutidos** realizam a seleção de atributos durante o próprio processo de treinamento do modelo, sendo comuns em algoritmos baseados em árvores de decisão.

Em descoberta de fármacos, algoritmos como Random Forest e métodos de *gradient boosting* oferecem mecanismos naturais para estimar a importância relativa dos atributos, sendo frequentemente utilizados tanto como modelos preditivos quanto como ferramentas auxiliares para seleção de características relevantes. Alternativamente, técnicas de redução de dimensionalidade, como a Análise de Componentes Principais (PCA), podem ser empregadas quando o objetivo é compactar a representação dos dados preservando a maior parte da variância explicada, ainda que à custa de uma menor interpretabilidade dos atributos originais.

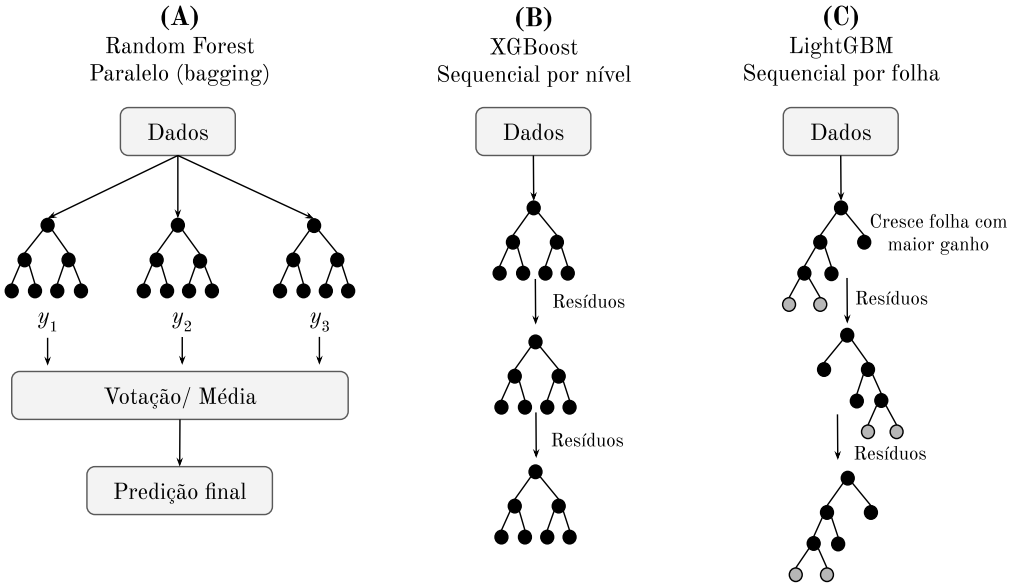

Figura 6.7: Comparação entre métodos de Aprendizado de Máquina baseados em árvores: (A) Random Forest: Ilustração do método de bagging, onde múltiplas árvores são treinadas de forma independente e paralela, gerando uma predição final por meio de votação ou média. (B) XGBoost: Estrutura de boosting sequencial com crescimento por nível (level-wise), focada na redução progressiva de resíduos. (C) LightGBM: Estrutura sequencial com crescimento por folha (leaf-wise), priorizando a expansão do nó que oferece o maior ganho de informação.

6.2.5.2. Modelos supervisionados clássicos baseados em Ensemble

Esses modelos desempenham um papel fundamental na etapa de indução de modelos preditivos, que se inicia após a preparação dos dados. Nesse contexto, técnicas de AM são empregadas para explorar representações previamente definidas e permitir a aprendizagem a partir de exemplos rotulados. Entre essas abordagens, modelos baseados em ensembles (ou conjuntos de modelos) destacam-se por sua ampla adoção em estudos acadêmicos e aplicações industriais. Tal predominância decorre da capacidade desses métodos de combinar múltiplos modelos simples, resultando em predições mais robustas, com menor variância e maior estabilidade diante de dados ruidosos e de alta dimensionalidade, frequentemente encontrados em conjuntos de dados químicos e biológicos [Hastie 2009, Murphy 2012].

Em tarefas típicas da área, como predição de atividade biológica, afinidade proteína-ligante ou propriedades ADMET, os dados disponíveis frequentemente apresentam número limitado de amostras, forte correlação entre atributos e distribuição desequilibrada entre classes. Nesse cenário, algoritmos de ensemble baseados em árvores de decisão têm se mostrado particularmente eficazes, pois são capazes de capturar relações não lineares complexas sem exigir suposições explícitas sobre a distribuição dos dados, além de apresentarem bom desempenho mesmo na presença de atributos irrelevantes.

Dentre estes modelos, o **Random Forest** destaca-se por sua simplicidade conceitual e robustez [Breiman 2001]. Essa características garantem uma grande aplicabilidade em problemas de descoberta de fármacos. O método baseia-se no princípio de bagging

(bootstrap aggregating), cujo objetivo é reduzir a variância de modelos de alta complexidade, como árvores de decisão, por meio da combinação de múltiplos modelos treinados de forma independente. Nesse contexto, o Random Forest constrói um conjunto de B árvores de decisão, cada uma treinada a partir de diferentes subconjuntos aleatórios dos dados de treinamento e dos atributos disponíveis, o que contribui para mitigar o risco de sobreajuste e aumentar a capacidade de generalização do modelo, veja na Figura 6.7(A). A predição final resulta da agregação das predições individuais das árvores que compõem o ensemble, sendo calculada pela média no caso de tarefas de regressão ou por votação majoritária em problemas de classificação. Formalmente, para uma amostra de entrada x, a predição do modelo pode ser expressa como:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} f_b(x) \tag{1}$$

onde $f_b(x)$ representa a predição produzida pela b-ésima árvore. Além de seu desempenho preditivo, o Random Forest fornece, como subproduto do processo de treinamento, estimativas da importância relativa dos atributos, característica particularmente relevante em descoberta de farmácos. Nessa área, compreender quais descritores ou características estruturais exercem maior influência sobre a atividade molecular pode ser tão importante quanto alcançar uma alta acurácia preditiva [Svetnik et al. 2003].

Enquanto o Random Forest explora o princípio de bagging, no qual múltiplos modelos são treinados de forma independente para reduzir a variância, outras abordagens de ensemble, como **XGBoost** e **LightGBM**, adotam uma estratégia distinta baseada em boosting [Friedman 2001]. Nessas abordagens, os modelos são construídos de maneira sequencial, com o objetivo de corrigir progressivamente os erros cometidos nas etapas anteriores. Esses algoritmos oferecem maior flexibilidade e, em muitos casos, melhor desempenho preditivo, sendo amplamente utilizados em competições e *benchmarks* de descoberta de fármacos.

Nesse contexto, o XGBoost (eXtreme Gradient Boosting) implementa o *Gradient Boosting* de forma otimizada, expandindo as árvores nível a nível (*level-wise*) para manter o equilíbrio da estrutura, [Chen and Guestrin 2016]. O algoritmo minimiza de forma iterativa uma função de objetivo regularizada, onde cada nova árvore tenta ajustar os resíduos das predições anteriores, veja Figura 6.7(B). A função de objetivo no passo t é dada por:

$$\mathscr{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

onde $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \|w\|^2$ representa o termo de regularização para controlar a complexidade do modelo. Nessa formulação, x_i e y_i representam, respectivamente, o vetor de atributos e o rótulo verdadeiro da i-ésima amostra do conjunto de treinamento, enquanto $\hat{y}_i^{(t-1)}$ corresponde à predição acumulada do modelo até a iteração anterior. A função de perda $l(\cdot)$ mede o erro entre a predição e o valor real, podendo assumir diferentes formas conforme a tarefa de regressão ou classificação. O termo $f_t(x_i)$ representa a contribuição da árvore adicionada na iteração t, que é ajustada para modelar os resíduos das predições

anteriores. O termo de regularização $\Omega(f_t)$ penaliza a complexidade da árvore, onde γ controla o número de folhas T e λ regula a magnitude dos pesos w, favorecendo modelos com melhor capacidade de generalização.

Distinto do XGBoost, o LightGBM, [Ke et al. 2017], utiliza uma estratégia de crescimento por folha (*leaf-wise*), escolhendo a folha que resulta na maior redução da função de perda (*loss*) para expansão, veja Figura 6.7(C). Essa abordagem é matematicamente eficiente pois foca no erro residual local, embora exija maior controle de profundidade para evitar *overfitting*. O ganho de divisão para uma folha é calculado como:

$$\text{Gain} = \frac{1}{2}\left[\frac{(\sum g_i)_L^2}{\sum h_i + \lambda} + \frac{(\sum g_i)_R^2}{\sum h_i + \lambda} - \frac{(\sum g_i)_{Total}^2}{\sum h_i + \lambda}\right]$$

Nessa expressão, g_i e h_i representam, respectivamente, o gradiente de primeira ordem e o hessiano de segunda ordem da função de perda em relação à predição do modelo para a i-ésima amostra. Os índices L e R denotam os subconjuntos de amostras atribuídos às folhas esquerda (*left*) e direita (*right*) após a divisão candidata, enquanto o índice $Total$ refere-se ao conjunto de amostras presente na folha original antes da divisão. Dessa forma, os termos $(\sum g_i)_L$ e $(\sum g_i)_R$ correspondem às somas dos gradientes das amostras associadas às folhas resultantes, enquanto $(\sum g_i)_{Total}$ representa a soma dos gradientes da folha original. O parâmetro λ é um hiperparâmetro de regularização L_2, introduzido para penalizar valores elevados dos pesos das folhas e controlar a complexidade do modelo, contribuindo para a redução do *overfitting*. O valor do ganho quantifica a redução esperada da função de perda decorrente da divisão da folha, sendo utilizado pelo LightGBM para selecionar, de forma gulosa, a divisão que maximiza o benefício preditivo.

Em conjunto, esses modelos supervisionados clássicos estabelecem uma linha de base sólida para aplicações de AM em descoberta de fármacos, servindo tanto como ferramentas preditivas de alto desempenho quanto como referências para a avaliação de abordagens mais complexas baseadas em aprendizado profundo.

6.2.5.3. Modelos baseados em Transformers

Os recentes avanços em aprendizado profundo têm impulsionado o uso de modelos baseados em **Transformers** em diversas áreas da ciência, incluindo a descoberta de fármacos. Originalmente desenvolvidos no contexto do processamento de linguagem natural [Vaswani et al. 2017], esses modelos foram projetados para lidar com sequências simbólicas (principalmente texto) e capturar dependências complexas presentes na linguagem humana escrita. Esse mesmo arcabouço conceitual tem sido atualmente sido estendido para outras "linguagens" que não são naturalmente interpretáveis por humanos, como a linguagem química, bioquímica e de proteínas, permitindo que modelos computacionais operem diretamente sobre essas representações simbólicas para capturar e modelar os efeitos das interações moleculares, ainda que a interpretação semântica explícita destas "linguagens" não seja conhecida.

Diferentemente de modelos supervisionados clássicos, esses métodos são capazes de aprender representações distribuídas diretamente a partir dos dados de entrada, redu-

zindo a dependência de descritores manualmente projetados. A arquitetura dos Transformers permite modelar dependências de longo alcance de forma eficiente e paralelizável.

No contexto da química e da bioinformática, os Transformers têm sido aplicados com sucesso em diferentes tipos de representações, como sequências SMILES, sequências de aminoácidos e, mais recentemente, representações estruturais e multimodais [Schwaller et al. 2019]. Em tarefas como predição de atividade biológica, geração de novas moléculas, modelagem de propriedades fisicoquímicas e afinidade proteína-ligante, esses modelos demonstram elevada capacidade de capturar padrões complexos presentes em grandes volumes de dados.

Ademais, estratégias de *pré-treinamento* em grandes bases de dados químicos ou biológicos, seguidas de *fine-tuning* em tarefas específicas, têm se mostrado bastante eficazes em cenários com dados rotulados limitados, uma situação comum na descoberta de fármacos [Chithrananda et al. 2020]. Apesar de seu elevado custo computacional, e menor interpretabilidade em comparação a modelos clássicos, os Transformers representam uma classe poderosa de modelos para aplicações em larga escala e continuam a expandir seu papel em *pipelines* modernos de AM na área.

6.2.5.4. Validação e avaliação de modelos

A validação adequada de modelos de AM é um aspecto crítico na descoberta de fármacos, uma vez que desempenhos preditivos inflados podem levar a conclusões incorretas e decisões experimentais custosas. Diferentemente de aplicações tradicionais de AM, conjuntos de dados químicos e biológicos frequentemente apresentam dependências estruturais, desbalanceamento entre classes, *data leakage* e viés de amostragem, exigindo estratégias de validação cuidadosamente projetadas [Wallach and Heifets 2018].

Abordagens clássicas, como validação cruzada aleatória, são amplamente utilizadas, mas podem superestimar o desempenho do modelo quando compostos estruturalmente semelhantes aparecem simultaneamente nos conjuntos de treinamento e teste. Nesse contexto, estratégias mais rigorosas, como *scaffold splitting*, são frequentemente empregadas para avaliar a capacidade de generalização do modelo para novas classes químicas.

A escolha das métricas de avaliação deve ser compatível com a tarefa em questão. Em problemas de **classificação**, são comumente usadas métricas como AUC (*Area Under the Curve*), ROC (*Receiver Operating Characteristic*), que avalia a capacidade do modelo de discriminar entre classes ao longo de diferentes limiares de decisão, precisão e revocação, que quantificam respectivamente a proporção de predições positivas corretas e a cobertura dos exemplos positivos. Em contextos de descoberta de fármacos, métricas de enriquecimento precoce também são frequentemente empregadas, pois avaliam a capacidade do modelo de priorizar compostos ativos nas primeiras posições de uma lista ordenada, um aspecto particularmente relevante em cenários de VS. Em tarefas de **regressão**, métricas como RMSE (*Root Mean Squared Error*) e MAE (*Mean Absolute Error*) medem o erro médio das predições em relação aos valores reais, enquanto o coeficiente de determinação $\left(R^2\right)$ indica a fração da variância dos dados explicada pelo modelo [Hastie 2009]. Uma validação criteriosa, aliada à interpretação adequada dessas

métricas, é fundamental para garantir que os modelos desenvolvidos apresentem desempenho confiável e relevância prática em cenários reais de descoberta de fármacos.

6.2.6. Competições de proposta de novas moléculas: CACHE e DREAM

A avaliação de modelos preditivos em descoberta de fármacos apresenta viés de sobreajuste (*overfit*) quando restrita a validações retrospectivas em bancos de dados estáticos. Desafios científicos estruturados sob os princípios da ciência aberta reduzem essa limitação ao exigir predições cegas prospectivas atreladas à validação experimental. Essas competições padronizam métricas de desempenho, expõem a taxa real de sucesso (*hit rate*) de diferentes arquiteturas de AM e forçam a construção de *pipelines* reprodutíveis sob condições de estrito isolamento de dados de teste, evitando assim o vazamento de dados.

O CACHE Challenge atua como um mecanismo direto de validação empírica para *pipelines* computacionais de VS. Um resumo das edições anunciadas até o presente momento (maio de 2026) do CACHE Challenge está apresentado na Tabela 6.1.

Tabela 6.1: Resumo das edições do CACHE Challenge (*Critical Assessment of Computational Hit-finding Experiments*)

#	Alvo	Categoria	Objetivo
1	Domínio WD40 da LRRK2	Alvo sem ligantes conhecidos	Descoberta de ligantes para o domínio WD40 da proteína LRRK2 associada à Doença de Parkinson
2	Helicase NSP13 do SARS-CoV-2	Sítio de ligação ao RNA	Identificação de compostos ligantes para a helicase viral NSP13 visando antivirais contra coronavírus
3	Macrodomínio Mac1 da NSP3 do SARS-CoV-2	Ligantes competitivos de ADP-ribose	Descoberta de novos ligantes para o macrodomínio viral evitando grupos ácido carboxílico
4	Domínio TKB da CBLB	Interação proteína-proteína	Identificação de ligantes para o domínio TKB da ubiquitina ligase CBLB
5	Tau K18 fibrils	Agregação proteica	Descoberta de moléculas capazes de interagir com fibrilas Tau associadas a doenças neurodegenerativas
6	Receptor Sigma-2/TMEM97	GPCR/alvo de membrana	Identificação de novos ligantes para o receptor Sigma-2/TMEM97
7	MPro variante resistente	Resistência antiviral	Descoberta de ligantes ativos contra variantes resistentes da protease principal do SARS-CoV-2
8	GID4 (CTLH E3 ligase)	Degradação proteica direcionada	Identificação de moléculas capazes de ocupar o bolso de reconhecimento de substrato da GID4

A organização define alvos biológicos de alta prioridade terapêutica e recebe submissões de estruturas moleculares ranqueadas *in silico* pelas equipes participantes. O consórcio executa a aquisição ou síntese química das moléculas selecionadas e conduz os

ensaios biofísicos e estruturais em laboratório. O desempenho dos algoritmos é medido exclusivamente pela identificação *in vitro* de ligantes ativos reais, reduzindo significativamente artefatos estatísticos associados a modelos superajustados e conjuntos de validação artificiais.

O *First DREAM Target 2035 Drug Discovery Challenge*, operado via consórcio MAINFRAME foi estruturado em múltiplas etapas sequenciais, nas quais os participantes desenvolveram modelos capazes de identificar compostos ativos em cenários de avaliação cega. Nas etapas iniciais, os modelos foram treinados utilizando uma biblioteca DEL contendo milhões de compostos representados por *fingerprints* moleculares, juntamente com dezenas de moléculas conhecidas em formato SMILES, sendo posteriormente avaliados em conjuntos ASMS compostos majoritariamente por exemplos negativos e um número reduzido de ligantes verdadeiros (determinados por SPR) que precisam ser identificados participantes. Além da classificação e ranqueamento molecular, o desafio avalia a capacidade dos modelos em identificar compostos estruturalmente diversos e generalizar para espaços químicos previamente não observados. Na etapa final, os melhores modelos são utilizados para selecionar compostos a partir de bibliotecas químicas contendo milhões de moléculas, cujos candidatos priorizados passam por validação experimental posterior. Dessa forma, o DREAM Challenge, associado à iniciativa MAINFRAME, estabelece um ambiente colaborativo para comparação sistemática de métodos de IA em cenários mais próximos das condições reais da descoberta de fármacos.

6.3. Estratégias Computacionais para a descoberta de fármacos

Esta seção apresenta estratégias computacionais para a descoberta de fármacos, incluindo o *docking* molecular (algoritmos e FEs) e a técnica de *co-folding*.

6.3.1. *Docking* Molecular

Avanços tanto na Ciência da Computação quanto na Bioinformática têm permitido o desenvolvimento de novas estratégias computacionais aplicadas na indústria farmacêutica durante os primeiros estágios de desenvolvimento de um novo fármaco, um processo definido como RDD [Kuntz 1992, Meng et al. 2011]. A aplicação em conjunto de métodos computacionais e métodos experimentais tem importante impacto na identificação de novos compostos promissores para alvos terapêuticos, reduzindo o tempo e os custos envolvidos nesse processo [Ferreira et al. 2015].

A interação entre receptores e ligantes pode ser simulada em um nível atômico por algoritmos de *docking* molecular [Lybrand 1995, Lengauer and Rarey 1996]. O *docking* molecular é tradicionalmente formulada como um problema composto por dois componentes principais: a amostragem conformacional (*sampling*), responsável por explorar poses, orientações e conformações do ligante no sítio de ligação do receptor-alvo, e a avaliação dessas soluções por FEs, responsáveis por estimar a afinidade e selecionar as poses mais prováveis [Lengauer and Rarey 1996, Trott and Olson 2010, Pagadala et al. 2017]. Nessas simulações, centenas de milhares de orientações e conformações do ligante no sítio de ligação do receptor são avaliadas e ranqueadas de acordo com a estabilidade do complexo estimada por uma FEB ou escore predita/calculada por uma FE) [Lengauer and Rarey 1996, Halperin et al. 2002,

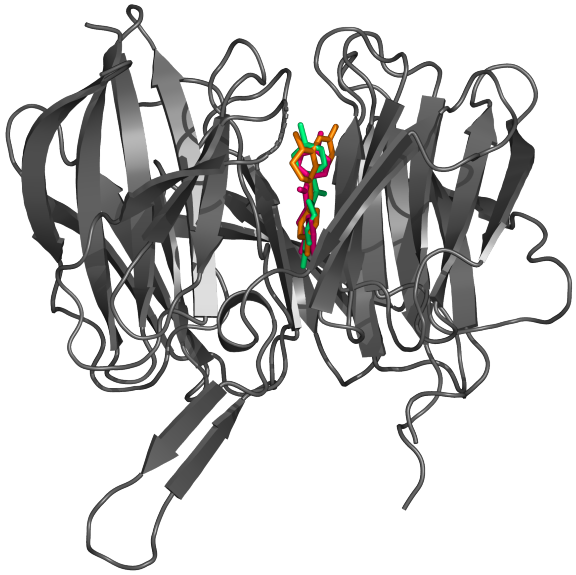

Figura 6.8: Exemplo de resultado de *docking* molecular. O receptor está em cinza, no formato de *cartoon* (Código PDB 9DTA, proteína WDR91). O ligante de exemplo está em 3 poses geradas pelo *docking* molecular com o programa AutoDock Vina 2021, no formato de *sticks* em rosa, laranja e verde. Figura gerada pelo software Pymol com resultado de *docking* gerado pelo grupo.

Meng et al. 2011, Crampon et al. 2022].

A Figura 6.8 mostra um exemplo de resultado de *docking* molecular. O receptor é a proteína WDR91 (Código PDB 9DTA), e três diferentes poses finais do ligante após a *docking* molecular são exibidas em rosa (a melhor pose), verde e laranja.

As simulações de *docking* molecular não são simples pois muitos fatores entrópicos e entálpicos influenciam a interação entre o receptor e o ligante. Sendo assim, muitos desafios estão envolvidos nesse processo e cada vez mais há a necessidade de aproximar métodos *in-silico* de métodos *in-vitro*, tornando as ferramentas computacionais de simulação mais próximas do que acontece na realidade. Entre esses desafios destacam-se a adequada representação da solvatação, efeitos induzidos por ajuste conformacional (*induced fit*), tratamento da flexibilidade do receptor e limitações inerentes das FEs em reproduzir afinidades experimentais [Pagadala et al. 2017, Pinzi and Rastelli 2019]. Além disso, há a necessidade de que novas tecnologias sejam desenvolvidas para aumentar a eficiência e eficácia da descoberta de novos fármacos. Mais recentemente, avanços em AM e aprendizado profundo vêm sendo incorporados nesse processo, ampliando o papel da *docking* molecular em *pipelines* modernos de descoberta de fármacos [Crampon et al. 2022, Scantlebury et al. 2020, Xu 2024].

6.3.1.1. Amostragem conformacional

Um dos principais desafios da *docking* molecular consiste na busca pela pose correta do ligante no sítio de ligação do receptor. Esse processo é conhecido como amostragem conformacional (*conformational sampling*) e envolve a exploração computacional das possí-

veis orientações, posições e conformações do ligante durante a interação proteína-ligante [Lengauer and Rarey 1996].

O problema da amostragem conformacional é complexo devido ao alto número de graus de liberdade envolvidos. Em geral, os algoritmos de *docking* molecular consideram o receptor como rígido e permitem graus de liberdade em algumas ligações dos ligantes. Porém, mesmo pequenas moléculas podem apresentar diversas ligações rotacionáveis, resultando em um grande número de conformações possíveis. Além disso, diferentes orientações e translações do ligante no sítio ativo precisam ser consideradas simultaneamente. Em cenários mais complexos, e para aproximar o resultado *in-silico* do *in-vitro* também pode ser necessário considerar flexibilidade do receptor, aumentando ainda mais o espaço conformacional explorado.

De forma geral, os algoritmos de amostragem utilizados em *docking* molecular podem ser classificados em:

- Métodos determinísticos/sistemáticos: exploram o espaço conformacional de forma sistemática avaliando combinações discretas de ângulos torcionais e orientações possíveis. Podem apresentar alto custo computacional, especialmente para moléculas mais flexíveis;
- Métodos estocásticos: é o método utilizado pela maioria dos programas de *docking* molecular. Nesse método, o espaço conformacional é explorado de maneira probabilística, onde diferentes conformações do ligante são geradas aleatoriamente e refinadas iterativamente, buscando minimizar a energia estimada do complexo [Morris et al. 2009] calculada por uma FE. Exemplos de algoritmos comuns utilizados nesta estratégia são os algoritmos genéticos e métodos de Monte Carlo;
- Métodos baseados em fragmentos: nessa abordagem, o ligante é dividido em partes menores e menos flexíveis para reduzir a complexidade do problema. Posteriormente, os fragmentos são reconectados e refinados para reconstruir a molécula completa. Essa estratégia reduz o custo computacional associado à exploração direta de todas as conformações possíveis do ligante completo;
- Métodos baseados em AM e aprendizado profundo: utilizam modelos computacionais treinados em grandes conjuntos de dados de complexos proteína-ligante (como os disponíveis no PDBbind) para aprender padrões estruturais e energéticos associados ao processo de interação molecular. Diferentemente das abordagens tradicionais, esses métodos podem auxiliar diretamente na priorização de regiões conformacionais mais favoráveis, previsão de poses de ligação e redução do espaço conformacional explorado durante a *docking* [Crampon et al. 2022, Masters et al. 2023]

Um dos mais conhecidos métodos de amostragem é o algoritmo utilizado pelo AutoDock, baseado em Algoritmos Genéticos Lamarckianos (Lamarckian Genetic Algorithm - LGA). Nesse método, populações de poses do ligante evoluem ao longo de múltiplas gerações utilizando operações inspiradas em evolução biológica, como mutação, recombinação e seleção [Morris et al. 2009]. O termo "Lamarckiano" decorre da

incorporação de refinamentos locais diretamente nos indivíduos gerados, permitindo acelerar a convergência para soluções energeticamente favoráveis. O AutoDock Vina, por sua vez, introduziu uma estratégia distinta baseada em busca global iterativa combinada com otimização local [Trott and Olson 2010]. O Vina utiliza uma representação eficiente dos graus de liberdade do ligante e explora o espaço conformacional por meio de heurísticas estocásticas associadas a refinamentos locais rápidos. Essa abordagem permitiu ganhos significativos tanto em velocidade quanto em precisão em comparação com versões anteriores do AutoDock.

Além da flexibilidade do ligante, a flexibilidade do receptor constitui outro grande desafio no *docking* molecular. Um grande número de abordagens alternativas para incorporar a flexibilidade do receptor na *docking* molecular tem sido desenvolvido, conforme revisado em [Ganesan et al. 2017, Amaro et al. 2018]. Entre as principais estratégias destacam-se o *soft-docking*, a flexibilidade das cadeias laterais, métodos de *induced fit docking* e o *ensemble docking*. No *soft-docking*, a flexibilidade do receptor é parcialmente incorporada por meio da suavização dos termos energéticos relacionados às interações de *van der Waals*, permitindo um certo grau de sobreposição entre os átomos do ligante e do receptor [Ferrari et al. 2004]. Já os métodos baseados na flexibilidade das cadeias laterais consideram mudanças conformacionais locais em resíduos do sítio de ligação, geralmente utilizando bibliotecas de rotâmeros previamente determinadas experimentalmente ou por análises estatísticas [Morris et al. 2009]. Entretanto, essas abordagens normalmente são limitadas a movimentos locais e não conseguem representar adequadamente alterações conformacionais globais da proteína [Ganesan et al. 2017].

Entre as diferentes abordagens, destaca-se o método de *ensemble docking*. Nessa estratégia, diferentes conformações do receptor podem ser combinadas em uma única representação estrutural ou utilizadas individualmente em múltiplas execuções independentes de *docking* molecular [Machado et al. 2010]. As diferentes conformações da proteína podem ser obtidas experimentalmente (e buscadas em bases de dados como o *Protein Data Bank*), ou ainda geradas computacionalmente por meio de simulações de dinâmica molecular (DM) [Amaro et al. 2018, Machado et al. 2010]. De acordo com Ganesan et al. [Ganesan et al. 2017], o *ensemble docking* tornou-se uma das abordagens mais aceitas em SBDD para incorporar a flexibilidade do receptor. Entretanto, um dos principais desafios desse método está relacionado ao elevado custo computacional, uma vez que a utilização de múltiplas estruturas do receptor implica na execução de múltiplas simulações de *docking* molecular [Scaini et al. 2019, Machado et al. 2010, Amaro et al. 2018].

Mais recentemente, estratégias baseadas em AM e aprendizado profundo têm sido propostas para auxiliar na incorporação da flexibilidade do receptor em métodos de *docking* molecular. Esses métodos buscam identificar conformações mais relevantes da proteína, reduzir o conjunto de estruturas necessárias em abordagens de *ensemble docking* ou aprender padrões conformacionais diretamente de grandes conjuntos de complexos proteína-ligante [Crampon et al. 2022]. Nesse contexto, modelos como o EDM-Dock [Masters et al. 2023] utilizam redes neurais profundas para prever diretamente matrizes de distância intermoleculares entre proteína e ligante, permitindo reconstruir poses de ligação sem a necessidade de algoritmos de busca iterativos e incorporando a flexibilidade do receptor de forma implícita por meio de representações coarse-grained das cadeias laterais. Essa abordagem reduz significativamente o custo computacional e contorna limi-

tações associadas à amostragem explícita de múltiplas conformações do sítio de ligação. Essas abordagens baseadas em AM representam uma alternativa promissora para reduzir os custos computacionais associados à flexibilidade do receptor, além de possibilitar a incorporação mais eficiente de movimentos conformacionais em FEs e algoritmos de *docking* molecular

Apesar dos avanços, a amostragem conformacional ainda representa um dos principais gargalos da *docking* molecular. A dificuldade em explorar adequadamente o espaço conformacional, especialmente em sistemas altamente flexíveis, continua sendo um desafio importante para a obtenção de poses biologicamente relevantes e afinidades confiáveis.

6.3.1.2. Funções de escore - enfoque nas baseadas em Aprendizado de Máquina

De acordo com Lin et al. [Li et al. 2019], as FEs para *docking* molecular proteina-ligante podem ser organizadas em 4 tipos, conforme resume a Tabela 6.2.

Como este capítulo trata sobre AM aplicado a descoberta de fármacos, esse tipo de FE é detalhado a seguir. As FE baseadas em AM surgiram como uma alternativa às funções clássicas (baseadas em fisica e empiricas), principalmente devido à capacidade desses modelos em aprender relações complexas entre proteínas e ligantes a partir de grandes conjuntos de dados. Diferentemente das funções empíricas, baseadas em força de campo ou conhecimento, as FE baseadas em AM utilizam dados experimentais previamente conhecidos para aprender padrões associados à afinidade de ligação [Ballester and Mitchell 2010, Shen et al. 2020].

As FE baseadas em AM, conforme revisado por [Shen et al. 2020], tem utilizado diferentes algoritmos de AM, como Random Forest, *Support Vector Machines*, XGBoost, entre outros. Em geral, o desenvolvimento dessas FE envolve algumas etapas principais: obtenção do conjunto de dados para treinamento, extração de descritores moleculares, treinamento do modelo e validação [Arrua et al. 2024].

A primeira etapa consiste na obtenção de um conjunto de complexos proteína-ligante com dados experimentais de afinidade de ligação, normalmente provenientes de bases como o PDBbind [Liu et al. 2017]. A partir desses complexos, são calculados descritores que representam as interações proteína-ligante. Esses descritores podem incluir termos energéticos, propriedades físico-químicas, *fingerprints* moleculares, contatos atômicos, informações estruturais [Ballester and Mitchell 2010, Shen et al. 2020]. A qualidade da representação dos dados possui impacto direto no desempenho do modelo [Shen et al. 2020].

A validação do modelo é uma etapa fundamental para avaliar a capacidade de generalização da FE. Diversas métricas podem ser utilizadas, incluindo coeficiente de correlação de Pearson (R), Root Mean Square Error (RMSE), *Mean Absolute Error* (MAE). Além dessas métricas tradicionais de regressão, um dos *benchmarks* mais utilizados para avaliação de FEs é o CASF (Comparative Assessment of Scoring Functions) [Su et al. 2018], que avalia diferentes aspectos das FE, incluindo capacidade de previsão de afinidade de ligação (scoring power), predição de identificaçao de ligantes nativos entre *decoys* (*docking power*), ranqueamento de ligantes (*ranking power*) e enriquecimento

Tabela 6.2: Classificação das funções de escore para *docking* proteína-ligante

Tipo	**Princípio**	**Vantagens**	**Limitações**	**Exemplos**
Baseadas em física (*Physics-based*)	Baseadas em termos físicos derivados da mecânica molecular, modelando interações como van der Waals e eletrostáticas	Interpretação física clara; base teórica consistente	Custo computacional mais alto; tratamento simplificado de solvente	DOCK [Allen et al. 2015]
Empíricas (*Empirical*)	Combinação de termos que representam diferentes interações proteína-ligante, com pesos ajustadosa partir de dados experimentais de afinidade	Maior eficiência computacional; boa correlação com dados experimentais no domínio de treino	Dependência do conjunto de treinamento; menor capacidade de generalização	Vina [Trott and Olson 2010]
Basedas em Conhecimento (*Knowledge-based*)	Potenciais derivados de análises estatísticas de complexos proteína-ligante	É o tipo de FE mais rápida; capturam padrões estruturais recorrentes	Dependência da qualidade e tamanho do banco de dados; interpretação indireta	Convex-PL [Kadukova et al. 2021]
Baseadas em Aprendizado de Máquina	Modelos de AM supervisionados que aprendem relações entre descritores estruturais e afinidade de ligação experimental	Bom desempenho preditivo; capturam relações não lineares complexas	Baixa interpreta-bilidade; forte dependência de dados de treinamento	RFL-Score[Arrua et al. 2024], CRRF-Score [Werhli et al. 2025], MedusaGraph [Jiang et al. 2022], Pafnucy [Stepniewska-Dziubinska et al. 2018]

em VS (*screening power*) [Su et al. 2018].

O ComBi-Lab, na FURG, tem desenvolvido FE baseadas em AM. Um exemplo é a RFL-Score [Arrua et al. 2024] que consiste em uma FE desenvolvida a partir de um conjunto híbrido de dados contendo complexos proteína–ligante experimentais e estruturas *decoys*. A representação dos complexos é construída por meio da extração de um conjunto amplo de 722 descritores provenientes do receptor, do ligante e de suas interações. Para reduzir a dimensionalidade e identificar atributos relevantes, foi aplicada a regressão *Lasso* como método de seleção de atributos. A FE propriamente dita é então construída utilizando o algoritmo *Random Forest*. A validação da FE RFL-Score foi realizada utilizando o *benchmark* CASF-2016 [Su et al. 2018] e os resultados mostraram que o RFL-Score apresenta desempenho competitivo, alcançando valores de correlação próximos ao estado da arte.

Duas áreas ganharam atenção crescente nos últimos anos nesse contexto: o *Deep Learning* - DL (Aprendizado Profundo) e os modelos generativos. Em FE para *docking* molecular, modelos de *Deep Learning* têm sido utilizados para aprender automaticamente representações complexas das interações proteína-ligante diretamente a partir de estruturas tridimensionais dos complexos moleculares [Shen et al. 2020, Meli et al. 2022]. Diferentes arquiteturas têm sido propostas para essa finalidade, incluindo redes neurais convolucionais tridimensionais (3D-CNNs), redes baseadas em grafos moleculares, entre outros. Essas abordagens permitem capturar relações espaciais e padrões não lineares das interações moleculares com maior capacidade de generalização quando comparadas a FEs tradicionais.

Entre as muitas FE baseadas em DL, uma das precursoras foi oa Pafnucy [Stepniewska-Dziubinska et al. 2018] que propôs redes neurais convolucionais 3D aplicadas a representações em grade dos complexos proteína–ligante para predizer afinidade de ligação. Também destaca-se o GNINA, cuja versão 1.3 [McNutt et al. 2025] utiliza redes neurais convolucionais 3D como FE e para avaliar e reclassificar as poses geradas no *docking*, a partir de representações tridimensionais da interação proteína-ligante, melhorando a precisão sem alterar o processo tradicional de amostragem. No contexto de FE baseadas em DL, tem-se as redes do tipo *Graph Neural Network* (GNN), que permitem representar o complexo proteína–ligante como um grafo, no qual os átomos correspondem a nós e as interações entre eles são modeladas como arestas, incorporando diretamente informações estruturais do sistema [Jiang et al. 2022]. Nesse sentido, o método MedusaGraph utiliza duas redes neurais baseadas em grafos: a primeira ajusta a posição dos átomos do ligante a partir de uma pose inicial, refinando a conformação no sítio de ligação, enquanto a segunda avalia a qualidade da pose resultante, classificando-a como próxima ou distante da estrutura experimental [Jiang et al. 2022]. Dessa forma, o modelo reduz a dependência de FE tradicionais e melhora a identificação de poses relevantes.

Mais recentemente, os *Large Language Models* (LLMs) também passaram a despertar interesse em aplicações relacionadas à descoberta de fármacos e bioinformática estrutural. Embora originalmente desenvolvidos para tarefas de processamento de linguagem natural, esses modelos podem ser adaptados para aprender padrões presentes em sequências biológicas, representações moleculares e interações proteína-ligante. Nesse contexto, arquiteturas baseadas em Transformers [Vaswani et al. 2017] têm sido utiliza-

das para modelagem de sequências de proteínas, geração de representações moleculares e aprendizado multimodal integrando informações estruturais e químicas. Revisões recentes indicam que esses LLMs também vêm sendo aplicados à predição de afinidade de ligação proteína-ligante a partir de sequências e representações textuais [Zheng et al. 2025], ainda que, diferentemente das FE tradicionais, não operem diretamente sobre a geometria tridimensional de poses.

Essas abordagens baseadas em DL e LLMs representam uma mudança importante no desenvolvimento de FE para *docking* molecular, permitindo reduzir a dependência de descritores definidos manualmente e possibilitando que os próprios modelos aprendam representações relevantes diretamente dos dados. Dessa forma, esses modelos tem potencial para contribuir tanto na predição de afinidade de ligação quanto na identificação de padrões moleculares complexos associados ao reconhecimento proteína-ligante.

6.3.2. Co-folding

Esta seção apresenta o *co-folding* é tratado como uma classe de métodos computacionais que predizem simultaneamente a geometria de múltiplos componentes moleculares (proteínas, ácidos nucleicos, ligantes, íons ou modificações pós-traducionais) a partir de representações de sequência e estrutura química.

Como discutido nas seções anteriores, o *docking* molecular é uma das estratégias mais utilizadas em estudos de SBDD. De modo geral, essa abordagem parte de uma estrutura tridimensional do alvo biológico, usualmente uma proteína, e busca predizer a pose mais provável de uma pequena molécula em seu sítio de ligação. Apesar de sua ampla aplicação, o *docking* molecular ainda apresenta limitações importantes, especialmente no tratamento da flexibilidade do receptor, na descrição de ajustes conformacionais induzidos pelo ligante e na estimativa quantitativa da afinidade proteína-ligante.

6.3.2.1. Da modelagem do receptor ao aprendizado profundo

Uma alternativa para explorar novas conformações de alvos não disponíveis experimentalmente é realizar a predição tridimensional do receptor. A modelagem por homologia foi, por décadas, uma das principais estratégias utilizadas. Essa abordagem depende da existência de uma estrutura molde relacionada, normalmente com identidade de sequência acima de 30% e cobertura suficientes para permitir a construção de um modelo confiável. Essa qualidade depende também da conservação das regiões funcionais, da qualidade do alinhamento e da resolução da estrutura experimental utilizada como referência. Entretanto, para proteínas sem moldes próximos, regiões flexíveis ou domínios pouco caracterizados, a acurácia desses métodos era frequentemente limitada inviabilizando o *docking* molecular para diversos potenciais alvos.

Nesse contexto, métodos recentes baseados em aprendizado profundo têm surgido como uma alternativa ou complemento às abordagens tradicionais, ao tentar predizer diretamente a estrutura de complexos biomoleculares. Essa estratégia é frequentemente denominada *co-folding*, pois busca modelar simultaneamente a conformação dos componentes moleculares e suas interações no complexo [Senior et al. 2020].

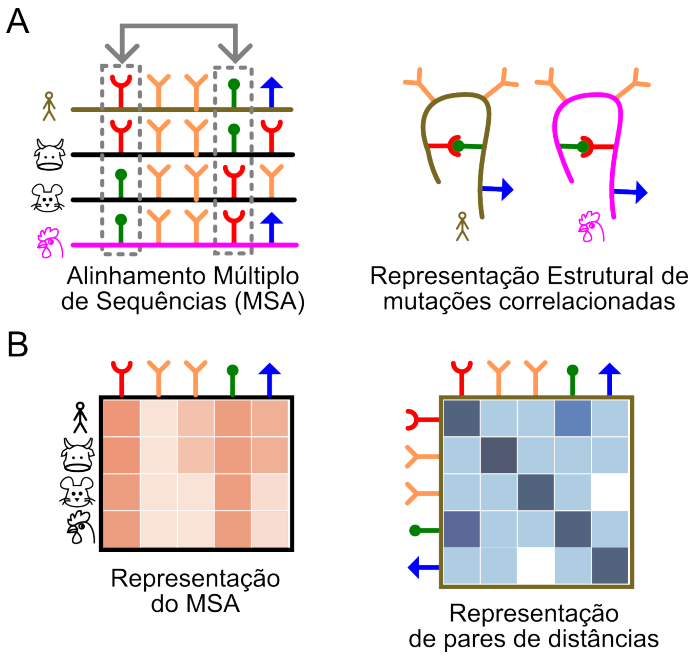

Figura 6.9: Esquema do uso de informação evolutiva na predição estrutural de proteínas. (A) Alinhamentos múltiplos de sequências permitem identificar posições que variam de forma correlacionada ao longo da evolução. Essas correlações podem indicar pares de resíduos que interagem ou que estão espacialmente próximos na estrutura tridimensional, uma vez que mutações em uma posição podem ser compensadas por mutações em outra posição. (B) Em modelos modernos de predição estrutural, como o AlphaFold2, as informações extraídas do MSA são convertidas em representações numéricas e integradas a representações de pares de resíduos, que codificam relações geométricas, como distâncias esperadas entre aminoácidos. Essa integração entre representação evolutiva e representação geométrica auxilia a inferência da estrutura tridimensional da proteína. Figura adaptada de Jumper et al. [Jumper et al. 2021].

Esse cenário foi profundamente alterado com o desenvolvimento de métodos de predição estrutural baseados em aprendizado profundo, em especial a partir das contribuições de Dr. John Jumper, Dr. Demis Hassabis e colaboradores da DeepMind com o AlphaFold (AF). A primeira versão do método, AF1, explorava informações de covariação evolutiva extraídas de alinhamentos múltiplos de sequência, ou MSA, do inglês *multiple sequence alignment* [Altschuh et al. 1987]. A premissa central é que resíduos espacialmente próximos em uma proteína tendem a apresentar padrões correlacionados de mutação ao longo da evolução (Figura 6.9). Assim, quando uma mutação ocorre em uma posição da sequência, outra posição próxima no espaço tridimensional pode sofrer uma mutação compensatória, preservando interações estruturais ou funcionais. Essas informações evolutivas são utilizadas para inferir relações entre pares de resíduos, como contatos ou distribuições de distância e ângulos de torção, que auxiliam na construção do modelo tridimensional [Senior et al. 2020]. Apesar do avanço em relação a métodos anteriores, a acurácia ainda era limitada, especialmente para alvos sem bons moldes.

6.3.2.2. AlphaFold2: predição estrutural, atenção e métricas de confiança

O novo conceito principal do AF2 [Jumper et al. 2021] foi o uso de mecanismos inspirados em Transformers, arquitetura baseada em atenção e também presente em modelos

como chatGPT (*Generative Pre-trained Transformer*). A atenção funciona de forma parecida com a tradução de uma frase: para traduzir uma palavra, o modelo não olha apenas para essa palavra isoladamente, mas também para outras palavras da frase que ajudam a definir seu significado. Por exemplo, em "*bank of the river*", a palavra "*river*" recebe mais peso para traduzir "*bank*" como "margem", e não como "banco". No AF, ocorre algo análogo: para atualizar a informação de um resíduo, o modelo aprende a dar mais peso a outros resíduos, posições do MSA ou pares de resíduos que ajudam a definir sua posição na estrutura 3D. Todos esses elementos são convertidos em representações numéricas, que são o formato compreendido pelo computador (*embeddings*) e que permite estabelecer relações [Jumper et al. 2021].

AF2 propôs o Evoformer (Figura 6.10 e Tabela 6.3), modelo que integra uma representação evolutiva (gerado a partir do MSA) e uma representação geométrica (relacionado aos pares de distâncias). A representação de pares é atualizada por operações triangulares, nas quais a relação entre dois resíduos é refinada usando um terceiro resíduo como contexto, favorecendo consistência geométrica em 3D. Quando uma hipótese geométrica se mostra inconsistente, as atualizações entre MSA e representação de pares permitem redistribuir a atenção e refinar a representação. Após 48 blocos de análises, um módulo estrutural constrói os resíduos através da predição da rotação e translação de cada resíduo a partir do modelo anterior. Após 3 ciclos entre o Evoformer e o módulo estrutural, hipótese(s) de estruturas da proteína (em formato PDB) são geradas [Jumper et al. 2021].

Um aspecto particularmente importante do AF2 foi a associação dos modelos gerados a métricas internas de confiança conforme ilustrado na Figura 6.11 contendo o modelo predito da WDR91 a partir de sua sequência inteira. O pLDDT fornece uma estimativa local da confiabilidade da predição, indicando quais regiões da estrutura tendem a estar bem modeladas e quais devem ser interpretadas com maior cautela [Jumper et al. 2021]. Já a matriz PAE, ou *Predicted Aligned Error*, estima o erro esperado na posição relativa entre pares de resíduos ou regiões da proteína. Essas métricas são fundamentais para o uso prático dos modelos, pois permitem avaliar se uma região de interesse, como um sítio de ligação, domínio funcional ou interface de interação, apresenta suporte suficiente para análises posteriores. Outra interpretação para resíduos em loops com baixa confiança é serem regiões intrinsecamente desordenadas, uma observação empírica feita por alguns estudos. Assim, o AF2 forneceu ferramentas para interpretar a confiabilidade dessas predições.

O sucesso do AF2 foi obtido quando esta ferramenta desbancou todos os antigos métodos baseados em conceitos físicos na edição do 14º CASP (*Critical Assessment of Structure Prediction*) ao predizer o enovelamento de proteínas com um erro semelhante ao observado experimentalmente [Kryshtafovych et al. 2021]. Nesta competição, estruturas com enovelamentos inéditos são determinadas experimentalmente sem o conhecimento dos competidores, que têm a chance de predizê-las a partir da sequência.

Outro importante impacto do AF2 foi a criação do AlphaFold Protein Structure Database, ou AFDB, desenvolvido em parceria entre a Google DeepMind e o EMBL-EBI [Varadi et al. 2022] descrito na seção 6.2.2.1. Para a SBDD, esse avanço foi relevante porque aumentou o número de proteínas potencialmente analisáveis por métodos estru-

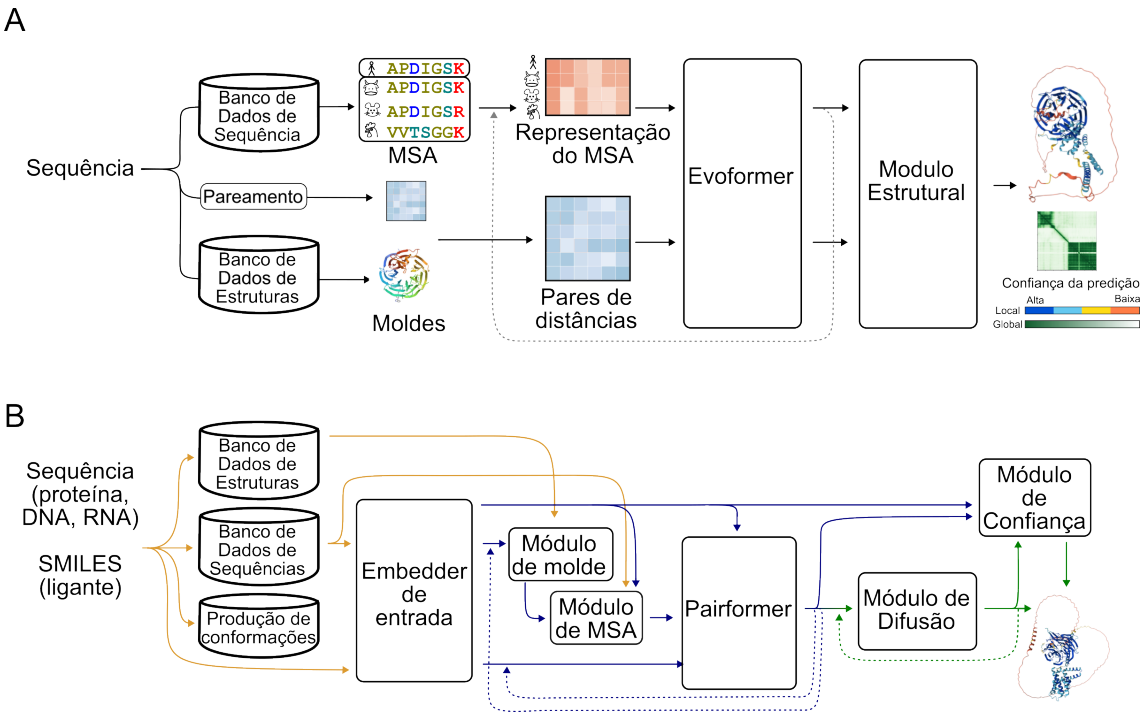

Figura 6.10: Comparação esquemática entre as arquiteturas do AlphaFold2 e do AlphaFold3. (A) O AlphaFold2 utiliza a sequência proteica, alinhamentos múltiplos de sequência (MSA) e moldes estruturais para gerar representações de MSA e de pares de distâncias, que são processadas pelo Evoformer e convertidas em uma estrutura tridimensional pelo módulo estrutural. (B) O AlphaFold3 expande essa lógica para complexos biomoleculares, incorporando proteínas, ácidos nucleicos e ligantes, e substitui o módulo estrutural por um módulo de difusão condicionado por representações processadas pelo Pairformer. Figura adaptada de Jumper *et al.* [Jumper et al. 2021] e Abramson *et al.* [Abramson et al. 2024].

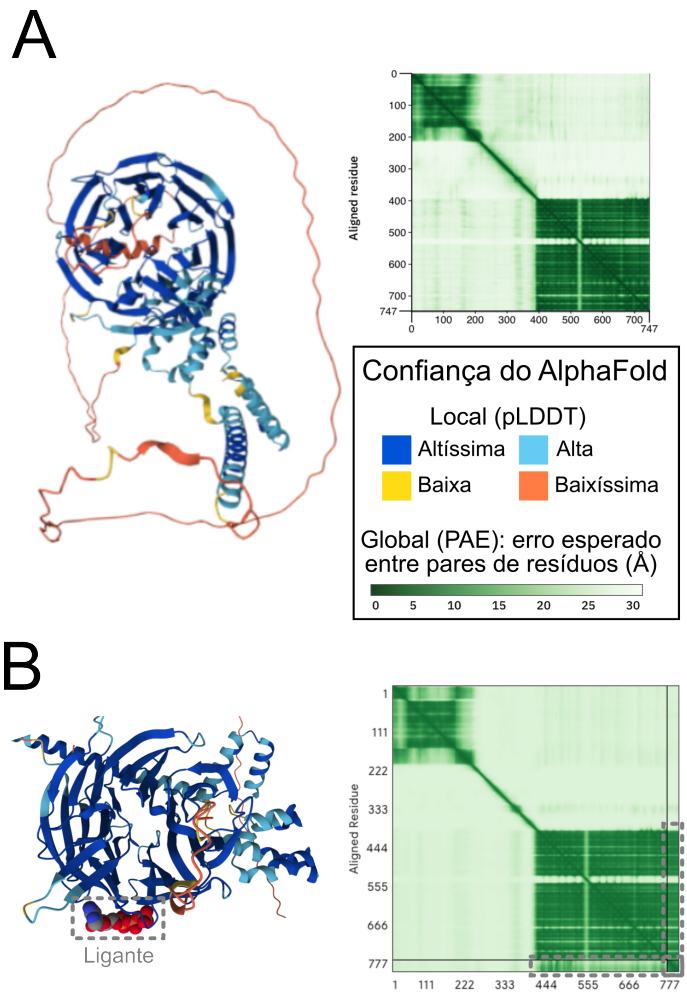

Figura 6.11: Comparação entre modelos da WDR91 preditos por AlphaFold2 (A) e AlphaFold3 (B). Em A, é mostrado o modelo da proteína completa, acompanhado das métricas de confiança local (pLDDT) e global (PAE). Em B, apenas o domínio WDR é representado em *cartoon*, enquanto o ligante predito pela estratégia de *co-folding* é representado por esferas e destacado por um retângulo tracejado.

turais. No entanto, modelos do AFDB devem ser utilizados com cautela em aplicações de *docking* molecular, principalmente quando o sítio de ligação apresenta baixa confiança local, quando há regiões desordenadas ou quando a conformação funcional depende da presença de ligantes, cofatores, membranas ou parceiros biomoleculares.

6.3.2.3. AlphaFold3 e a emergência do co-folding para complexos biomoleculares

O AF3 expandiu esse paradigma ao propor um modelo voltado à predição conjunta de complexos biomoleculares, incluindo proteínas, DNA, RNA, pequenas moléculas e modificações pós-traducionais [Abramson et al. 2024] (Figura 6.10B e Tabela 6.3). Essa abordagem é frequentemente descrita como co-folding, pois busca predizer simultaneamente a conformação dos componentes e suas interações no complexo, diferentemente do *docking* molecular.

Na arquitetura do AF3 [Abramson et al. 2024] (Figura 6.10), diferentes entradas são processadas por diferentes abordagens: i) a sequência da macromolécula é utilizada para realizar uma busca por moldes (templates) e ser processada no módulo de molde; essa sequência também será utilizada para construir o MSA e processá-lo para gerar a matriz de pares de distâncias que alimentará o módulo de MSA, que reduz a matriz e os cálculos reduzidos; iii) moléculas representadas por SMILES têm seus confôrmeros gerados pelo RDKit. Diferentemente do AF2, o AF3 substitui o Evoformer por uma arquitetura denominada Pairformer e utiliza um modelo de difusão no módulo estrutural. Em termos conceituais, modelos de difusão aprendem a reconstruir dados a partir de versões dos dados progressivamente corrompidas por ruído. Seria como transformar uma imagem em um chiado de televisão antiga e aprender como resgatar o início a partir do ruído. No caso do AF3, a estrutura inicial pode ser entendida como um conjunto de átomos distribuídos de forma ruidosa no espaço, que é progressivamente refinado com base nas representações aprendidas pelo modelo.

6.3.3. Novos modelos generativos para co-folding com novas funções

Diante do sucesso do AF3, outros grupos desenvolveram modelos generativos ou foundation models para co-folding. Um dos primeiros foi o RoseTTAFold All-Atom do Dr. David Baker [Krishna et al. 2024] (Tabela 6.3), também utilizando um modelo de difusão, complexos iniciam com todos os constituintes aleatoriamente dispostos no espaço que vão se ordenando com informações evolutivas, sobre moldes e no caso dos ligantes, por sua representação em grafos.

Outras abordagens, como Chai-1 (Tabela 6.3) e Boltz, também avançaram ao permitir a inclusão de restrições espaciais como parte da entrada do modelo [Chai Discovery Team 2024]. Essas restrições podem ter origem em dados experimentais ou em conhecimento prévio sobre o sistema, como contatos esperados, distâncias aproximadas ou regiões envolvidas em interação. Essa capacidade é particularmente relevante para aplicações em biologia estrutural integrativa, nas quais informações incompletas de diferentes fontes podem ser combinadas para orientar a predição de complexos biomoleculares.

Mais recentemente, propostas como Boltz-2 (Tabela 6.3) passaram a sugerir tam-

bém a predição de afinidade proteína-ligante e o desenho de novas moléculas (*de novo design*) com possibilidade de síntese [Passaro et al. 2025]. Essas capacidades, se acuradas, poderiam não apenas encontrar novos *hits* como também otimizá-los a leads. Para alcançar tais objetivos, os modelos foram treinados após a aplicação de curadoria a diversos bancos de dados, como ChEMBL, BindingDB, PubChem, entre outros. No entanto, como algumas dessas abordagens ainda foram divulgadas inicialmente como preprints, suas capacidades devem ser interpretadas com cautela até validação mais ampla pela comunidade.

Tabela 6.3: Comparação entre programas de predição estrutural baseados em aprendizado profundo e métodos recentes de *co-folding* para complexos biomoleculares.

Programa	**Princípio**	**Vantagens**	**Referência**
AF2	Evoformer + módulo estrutural	Marco na predição estrutural de proteínas; não utiliza difusão generativa e tende a sinalizar regiões incertas por métricas de confiança	[Jumper et al. 2021]
AF3	Pairformer + difusão	Acelera a predição estrutural e expande a modelagem para complexos biomoleculares heterogêneos	[Abramson et al. 2024]
RFAA	Difusão + representação atômica	Modelagem em nível atômico com refinamento da geometria final dos complexos	[Krishna et al. 2024]
Chai1	Modelo generativo com difusão	Permite incorporar restrições espaciais ou conhecimento prévio como entrada do modelo	[Chai Discovery Team 2024]
Boltz2	Modelo generativo / foundation model	Propõe integração entre predição estrutural e estimativa de afinidade proteína-ligante	[Passaro et al. 2025]

6.3.4. Limitações e uso crítico em SBDD

Apesar do entusiasmo em torno dessas abordagens, é importante interpretar seus resultados com cautela. Em uma avaliação sistemática, Škrinjar *et al*. demonstraram que abordagens atuais de *co-folding* tendem a memorizar poses de ligantes presentes no treinamento [Škrinjar et al. 2026]. De forma semelhante, o *benchmark* FoldBench mostrou que o sucesso na modelagem de ligantes aumenta com a similaridade química em relação ao conjunto de treinamento, sugerindo que esses modelos são mais robustos para recapitular complexos conhecidos ou quimicamente próximos do que para generalizar para ligantes verdadeiramente novos [Xu et al. 2026]. Essas observações são indicativos que *benchmarkings* mais realistas e desconhecidos precisam ser adotados e que a capacidade destes modelos de generalizar para complexos verdadeiramente inéditos e impõe cautela em aplicações de descoberta de moléculas em abordagem *de novo*.

Alguns estudos reportaram limitações do AF3 na predição de estados conformacionais alternativos já caracterizados experimentalmente. Este é o caso para ubiquitina ligase E3 que um estado aberto e outro fechado foram determinados experimentalmente,

contudo apenas o último é predito [Abramson et al. 2024]. Outros exemplos são de proteínas órfãs, i.e., sem ou com poucas proteínas homólogas nos bancos de dados de sequências. Estratégias de aumentar número de modelos gerados ou aumentar as sementes (seeds) que podem melhor a confiança dos modelos gerados.

A arquitetura generativa do AF3 implica que às vezes produz categorias de erro decorrente de ordem estrutural espúria (alucinações) em regiões desordenadas que não foram observadas no AF2. Neste, regiões alucinadas geralmente são marcadas com baixíssima confiança, com pontuações pLDDT bem abaixo de 50, e em conformação de loop disperso no espaço. No entanto, AF3 pode apresentá-las como hélices [Abramson et al. 2024].

Outra limitação importante é o maior tempo de predição dos métodos de *co-folding*, geralmente na ordem de minutos por complexo, em comparação ao *docking* molecular, que costuma operar na escala de segundos por ligante. Esse custo computacional ainda limita sua aplicação direta em etapas de *virtual screening* em larga escala, nas quais é necessário avaliar milhares ou milhões de moléculas. Assim, para que métodos de *co-folding* sejam incorporados de forma rotineira em *pipelines* de descoberta de fármacos, será necessário reduzir substancialmente seu custo computacional ou utilizá-los de maneira complementar, por exemplo, na reclassificação de conjuntos menores previamente filtrados por *docking*, AM ou triagens experimentais.

A predição correta da geometria de um complexo não implica necessariamente a predição correta de afinidade, seletividade ou atividade funcional. Interações proteína-ligante dependem de fatores físico-químicos complexos, incluindo efeitos entrópicos, solvatação, protonação, tautomeria, flexibilidade conformacional e estados funcionais alternativos da proteína. Assim, o *co-folding* deve ser entendido menos como substituto direto do *docking* molecular e mais como uma nova camada de modelagem estrutural probabilística. Seu maior valor está em gerar hipóteses tridimensionais para complexos biomoleculares, explorar modos de interação plausíveis e integrar informações estruturais, evolutivas e químicas em *pipelines* de descoberta de fármacos. Entretanto, sua aplicação em SBDD ainda exige validação experimental, controle rigoroso de vazamento de dados e interpretação crítica das métricas de confiança, especialmente quando o objetivo envolve afinidade, seletividade ou desenho de moléculas inéditas.

6.4. Estudo de Caso: WDR91

A proteína WDR91 é uma escolha interessante do ponto de vista didático e científico porque representa um alvo desafiador, pouco explorado farmacologicamente e associado a dados experimentais recentes de triagem molecular de DEL. Proteínas desse tipo são exemplos importantes do problema enfrentado por iniciativas como o Target 2035: há muitos alvos biologicamente relevantes para os quais ainda não existem ligantes químicos bem caracterizados ou ferramentas moleculares amplamente disponíveis.

A WDR91 é uma proteína contendo um domínio de repetição Trp-Asp (WDR) com enovelamento característico de β-*propeller*. Essa classe de proteínas são codificadas por aproximadamente 350 genes e intrinsicamente envolvidos em doenças [Ackloo et al. 2025]. A WDR91 regula a conversão precoce do endossomo e desempenha papéis vitais na fusão, reciclagem e transporte do endossomo, sendo importante em

processos do desenvolvimento neuronal e apontada como potencial fator hospedeiro para infecção viral por experimentos de *wide-genome CRISPR* [Ahmad et al. 2023]. Dessa forma, essa classe de proteínas é frequentemente associada à formação de plataformas de interação proteína-proteína, participando de processos celulares regulatórios relevantes para DD. Esse tipo de arquitetura estrutural pode tornar a descoberta de ligantes particularmente desafiadora, pois seus sítios de interação nem sempre correspondem a cavidades profundas e bem definidas, como aquelas observadas em enzimas clássicas. Por isso, a identificação de pequenas moléculas capazes de interagir com proteínas desse tipo é um problema relevante tanto para a biologia química quanto para o desenvolvimento de novas estratégias computacionais.

No contexto do minicurso, o domínio WDR da proteína WDR91 (região compreendida entre os resíduos 392 a 747) será tratada como um alvo-modelo para demonstrar como dados experimentais de triagem podem ser convertidos em um problema de classificação molecular. Cada molécula presente no conjunto de dados é associada à uma representação química computacional e a uma classe experimental: positiva ou negativa. O objetivo do modelo será aprender padrões moleculares associados ao enriquecimento experimental observado na triagem DEL. A partir disso, os alunos poderão treinar modelos capazes de estimar a probabilidade de uma nova molécula pertencer à classe de compostos potencialmente ligantes.

É importante destacar que, em dados de triagem experimental, as classes positiva e negativa não devem ser interpretadas de forma absoluta. Um composto classificado como positivo não é necessariamente um ligante validado com alta afinidade, assim como um composto classificado como negativo não é necessariamente incapaz de se ligar à proteína. Em DEL, por exemplo, a classificação pode refletir enriquecimento relativo sob determinadas condições experimentais. Portanto, é mais adequado interpretar os positivos como compostos enriquecidos ou candidatos a ligantes, e os negativos como compostos não enriquecidos ou sem evidência de ligação nas condições da triagem. Essa distinção é fundamental para que os alunos compreendam as incertezas e limitações inerentes a dados experimentais reais.

6.4.1. Subconjunto utilizado no minicurso

Para fins didáticos e escalonáveis para computadores domésticos, o minicurso utilizará uma versão simplificada de dados DEL associados à WDR91 [Ahmad et al. 2023]. O subconjunto será composto por aproximadamente 30 mil exemplos positivos e 150 mil exemplos negativos descritos por apenas um *fingerprint*, ECFP4, resultando em uma proporção aproximada de um positivo para cada cinco negativos. Essa distribuição mantém uma característica importante de problemas reais de triagem molecular: o desbalanceamento entre classes. Em campanhas experimentais, compostos com evidência de ligação costumam representar apenas uma pequena fração do total avaliado. Assim, mesmo em uma versão reduzida, o conjunto de dados preserva um desafio central para modelos de classificação em descoberta de fármacos.

Cada instância apresenta os ID identificadores dos *building blocks* usados para gerar a molécula final, que é representada pelo *fingerprint* ECFP4. *Fingerprints* moleculares são representações numéricas da estrutura química, utilizadas para converter moléculas

em vetores que podem ser interpretados por algoritmos de AM. O ECFP4, em particular, descreve ambientes atômicos circulares ao redor de cada átomo da molécula, capturando padrões locais de conectividade química. Essa representação é amplamente utilizada em quimioinformática porque permite comparar moléculas, calcular similaridade química e treinar modelos preditivos de atividade biológica.

A escolha de utilizar apenas o ECFP4 tem uma finalidade pedagógica. Em aplicações reais, é comum combinar diferentes tipos de descritores, como *fingerprints* adicionais, propriedades físico-químicas, representações por grafos, modelos tridimensionais, informações estruturais da proteína, *docking* ou *co-folding*. No entanto, para um primeiro contato com AM aplicado à descoberta de fármacos, o uso de uma única representação molecular reduz a complexidade técnica e permite que os alunos compreendam com clareza as etapas fundamentais do processo. Assim, o foco do minicurso será entender como preparar os dados, dividir o conjunto em treino e teste, lidar com desbalanceamento, treinar classificadores, avaliar métricas de desempenho e interpretar os resultados obtidos.

Além do treinamento e da avaliação inicial dos classificadores com os dados DEL, os modelos desenvolvidos no minicurso serão posteriormente pontuados em um conjunto independente de teste composto por moléculas identificadas por ASMS e validadas experimentalmente por SPR. Essa etapa permitirá avaliar se os padrões aprendidos a partir dos dados de enriquecimento em DEL são capazes de generalizar para compostos detectados por uma tecnologia experimental distinta e confirmados por um método ortogonal de ligação. Dessa forma, o conjunto de teste baseado em ASMS/SPR funcionará como uma validação mais rigorosa e biologicamente relevante dos modelos, aproximando a atividade prática de um cenário real de priorização de moléculas em descoberta de fármacos.

Após a descrição do estudo de caso específico da proteína WDR91, é útil abstrair os detalhes particulares e apresentar o fluxo de trabalho geral que orienta a aplicação de AM nesse tipo de problema. A Figura 6.12 apresenta uma estratégia conceitual e genérica para a seleção de moléculas em desafios de identificação de ligantes para um determinado alvo biológico.

No fluxo ilustrado na Figura 6.12, o painel (1) representa os conjuntos de dados de entrada e suas respectivas representações computacionais, destacando o uso de *fingerprints* moleculares para descrever compostos provenientes de bibliotecas codificadas por DNA (DEL), nas quais as estruturas químicas não estão explicitamente disponíveis na forma de SMILES. O painel (2) apresenta diferentes estratégias de seleção de atributos, aplicadas com o objetivo de reduzir a dimensionalidade e ressaltar características moleculares relevantes. Em seguida, o painel (3) descreve as etapas de treinamento e validação dos modelos de AM. Por fim, os painéis (4) e (5) ilustram a aplicação do conjunto de teste, composto por dados do tipo ASMS, ao modelo treinado, culminando no ranqueamento das moléculas candidatas de acordo com seu ranking estimado de atividade.

É importante ressaltar que a Figura 6.12 tem caráter conceitual e genérico, ilustrando um fluxo de trabalho típico para a aplicação de AM na identificação e priorização de moléculas bioativas a partir de dados de triagem molecular. A estratégia apresentada não corresponde exclusivamente ao caso da proteína WDR91, mas representa um arcabouço metodológico geral, aplicável a diferentes alvos biológicos e cenários experimentais, nos quais dados de treinamento e de teste podem ser provenientes de fontes

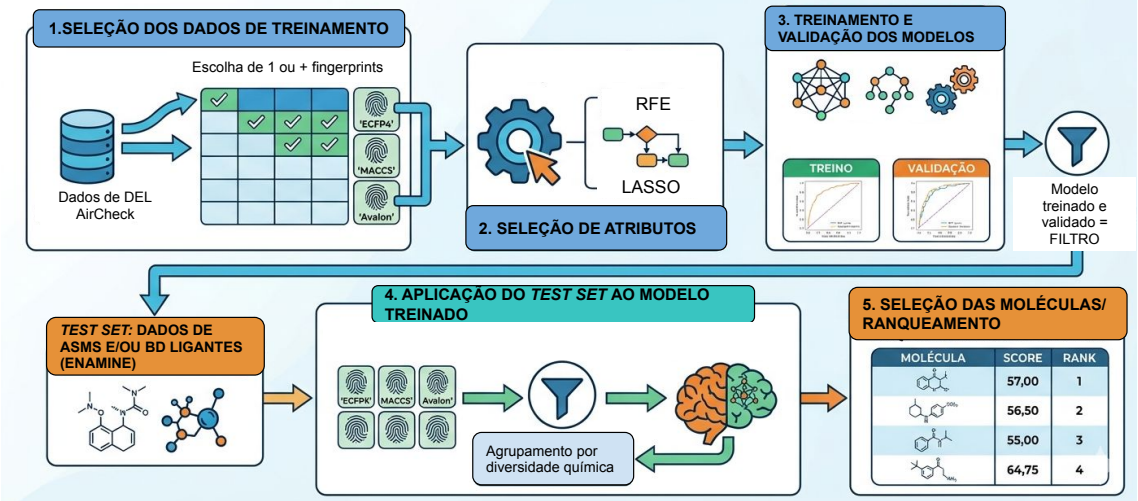

Figura 6.12: Proposta de estratégia de aplicação de Aprendizado de Máquina para a seleção de moléculas ativas em um cenário onde os dados de treinamento são provenientes de bibliotecas codificadas por DNA (DEL) e os dados de teste do tipo ASMS. A figura apresenta os dados e suas representações como *fingerprints* no painel 1. Já no painel 2 são apresentadas algumas opções de seleção de atributos, para logo em seguida no painel 3 mostrar o treino e a validação. Os painéis 4 e 5 mostram respectivamente a aplicação do test set ao modelo e a o ranqueamento das moléculas.

distintas.

6.5. Considerações finais

Esta seção sintetiza os principais conceitos discutidos ao longo do capítulo, destacando o papel crescente do AM na descoberta de fármacos e a importância da integração entre dados experimentais e métodos computacionais.

Ao longo deste capítulo, foram apresentados conceitos fundamentais para compreender como métodos de AM e IA vêm sendo incorporados à descoberta de fármacos. Partindo de fundamentos de bioinformática estrutural, representação molecular, bancos de dados químicos e estruturais, técnicas experimentais de triagem e modelos supervisionados, discutimos como diferentes tipos de dados podem ser transformados em entradas computacionais para a construção de modelos preditivos. Também foram abordadas estratégias clássicas e modernas de SBDD, incluindo *docking* molecular, FEs baseadas em AM e métodos recentes de *co-folding*. Essas técnicas foram discutidas em suas vantagens e limitações, com o objetivo de preparar a comunidade acadêmica para a escolha crítica das metodologias mais adequadas a diferentes desafios em SBDD.

Um ponto central é que o desempenho de modelos computacionais não depende apenas da escolha do algoritmo. A qualidade, a origem, a padronização e a interpretação dos dados são fatores igualmente determinantes. Em descoberta de fármacos, os dados experimentais frequentemente são escassos, heterogêneos, ruidosos ou gerados em condições específicas, o que dificulta a comparação entre métodos e a generalização para novos alvos ou novas classes químicas.

Para enfrentar esse problema, diferentes iniciativas internacionais vêm buscando organizar a geração, padronização e compartilhamento de dados experimentais para

SBDD. Uma delas é o projeto **LIGAND-AI**, que propõe disponibilizar de forma aberta no **AIRCHECK.ai** dados de triagens biofísicas, como DEL e ASMS, para uma fração representativa do proteoma humano. A obtenção de dados tabulados, homogêneos e comparáveis entre diferentes proteínas poderá fornecer uma base experimental robusta para treinar e avaliar métodos preditivos capazes de priorizar ligantes e, progressivamente, apoiar estimativas de potência e seletividade de pequenas moléculas. Em longo prazo, esse tipo de iniciativa se alinha à visão do **Target 2035**, que busca identificar moduladores farmacológicos ou sondas químicas para todas as proteínas humanas até 2035, incluindo aquelas ainda pouco exploradas experimentalmente, e finalmente acelerar a descoberta de fármacos.

O estudo de caso com **WDR91** ilustra essa lógica de integração entre dados experimentais disponíveis no **AIRCHECK.ai** e AM no SBDD. A utilização de um subconjunto simplificado de dados DEL, representado por *fingerprints* ECFP4, permite demonstrar de forma prática como um problema de descoberta de ligantes pode ser formulado como uma tarefa supervisionada de classificação e como moléculas disponíveis comercialmente podem ser priorizadas.

Nesse contexto, desafios comunitários, como o da WDR91 no **DREAM Target 2035 Drug Discovery Challenge** e outros no **CACHE**, têm papel complementar, pois transformam conjuntos de dados experimentais em *benchmarks* abertos para a comunidade, permitindo estabelecer o estado da arte, mapear desafios atuais e identificar potenciais soluções. Essas iniciativas podem desempenhar, para a descoberta de fármacos, um papel análogo ao que o **CASP** exerceu para a predição de estruturas de proteínas, ao criar um ambiente competitivo, padronizado e rigoroso que favoreceu o surgimento de métodos como o AlphaFold. Dessa forma, a comunidade científica brasileira encontra-se em um momento estratégico para participar dessas competições e contribuir para avanços relevantes possibilitados pela IA em SBDD. A equipe brasileira RFL_Bambu ilustra esse cenário ao vencer o CACHE Challenge #2, sendo a única equipe do Sul Global na competição.

Por fim, espera-se que os conceitos apresentados neste capítulo forneçam aos leitores e participantes uma visão crítica e aplicada da área. O uso de AM na descoberta de fármacos exige domínio técnico, mas também compreensão das limitações experimentais, dos vieses dos dados, das métricas de validação e do contexto biológico dos alvos estudados. Ao integrar fundamentos teóricos, exemplos práticos e iniciativas colaborativas de ciência aberta, este capítulo busca preparar os participantes para avaliar, desenvolver e aplicar modelos computacionais de forma responsável em problemas reais de descoberta de fármacos, contribuindo para a formação de uma nova geração de *drug hunters*.

6.6. Referências

Referências

[Abramson et al. 2024] Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. (2024). Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630:493–500.

[Ackloo et al. 2025] Ackloo, S., Li, F., Szewczyk, M., et al. (2025). A target class ligandability evaluation of wd40 repeat-containing proteins. *Journal of Medicinal Chemistry*, 68(2).

[Ahmad et al. 2023] Ahmad, S., Xu, J., Feng, J., Hutchinson, A., Zeng, H., Ghiabi, P., Dong, A., Centrella, P., Clark, M., Guié, M.-A., et al. (2023). Discovery of a first-in-class small-molecule ligand for wdr91 using dna-encoded chemical library selection followed by machine learning. *Journal of Medicinal Chemistry*, 66(23):16051–16061.

[Allen et al. 2015] Allen, W. J., Balius, T. E., Mukherjee, S., Brozell, S. R., Moustakas, D. T., Lang, P. T., Case, D. A., Kuntz, I. D., and Rizzo, R. C. (2015). Dock 6: impact of new features and current docking performance. *Journal of Computational Chemistry*, 36(15):1132–1156.

[Altschuh et al. 1987] Altschuh, D., Lesk, A. M., Bloomer, A. C., and Klug, A. (1987). Correlation of coordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *Journal of Molecular Biology*, 193(4):693–707.

[Amaro et al. 2018] Amaro, R. E., Baudry, J., Chodera, J., Demir, O., McCammon, J. A., Miao, Y., and Smith, J. C. (2018). Ensemble docking in drug discovery. *Biophysical journal*, 114(10):2271–2278.

[Arrua et al. 2024] Arrua, O. E., Aderhold, A., Werhli, A. V., and Machado, K. D. S. (2024). Rfl-score: random forest with lasso scoring function for protein-ligand molecular docking. In *2024 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–8. IEEE.

[Ballester and Mitchell 2010] Ballester, P. J. and Mitchell, J. B. (2010). A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26(9):1169–1175.

[Berman et al. 2000] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, 28(1):235–242.

[Bjerrum 2017] Bjerrum, E. J. (2017). Smiles enumeration as data augmentation for neural network modeling of molecules. *arXiv preprint arXiv:1703.07076*.

[Branden and Tooze 2012] Branden, C. I. and Tooze, J. (2012). *Introduction to protein structure*. Garland Science.

[Bray et al. 2024] Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., and Jemal, A. (2024). Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3):229–263.

[Breiman 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[Burley et al. 2021] Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V., Christie, C. H., Dalenberg, K., Di Costanzo, L., Duarte, J. M., et al. (2021). Rcsb protein data bank: powerful new tools for exploring 3d structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic acids research*, 49(D1):D437–D451.

[CACHE Initiative 2024] CACHE Initiative (2024). Cache challenge #2 results. Accessed March 2026.

[Carhart et al. 1985] Carhart, R. E., Smith, D. H., and Venkataraghavan, R. (1985). Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*, 25(2):64–73.

[Cereto-Massagué et al. 2015] Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., and Pujadas, G. (2015). Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63.

[Chai Discovery Team 2024] Chai Discovery Team (2024). Chai-1: Decoding the molecular interactions of life. *bioRxiv*. Preprint.

[Chen and Guestrin 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *KDD*, pages 785–794.

[Chithrananda et al. 2020] Chithrananda, S., Grand, G., and Ramsundar, B. (2020). Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*.

[Copeland et al. 2006] Copeland, R. A., Pompliano, D. L., and Meek, T. D. (2006). Drug–target residence time and its implications for lead optimization. *Nature Reviews Drug Discovery*, 5(9):730–739.

[Crampon et al. 2022] Crampon, K., Giorkallos, A., Deldossi, N., Baud, S., and Steffenel, L. A. (2022). Machine-learning methods for ligand–protein molecular docking. *Drug Discovery Today*, 27(1):151–164.

[Dalby et al. 1992] Dalby, A., Nourse, J. G., Hounshell, W. D., Gushurst, A. K., Grier, D. L., Leland, B. A., and Laufer, J. (1992). Description of several chemical structure file formats used by computer programs developed at molecular design limited. *Journal of chemical information and computer sciences*, 32(3):244–255.

[Daylight Chemical Information Systems 2024] Daylight Chemical Information Systems (2024). Daylight theory manual: Fingerprints. Accessed: 2026-05-11.

[Durant et al. 2002] Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G. (2002). Reoptimization of mdl keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280.

[Edwards et al. 2025] Edwards, A. M. et al. (2025). Protein–ligand data at scale to support machine learning. *Nature Reviews Chemistry*, 9:634–645.

[Ferrari et al. 2004] Ferrari, A. M., Wei, B. Q., Costantino, L., and Shoichet, B. K. (2004). Soft docking and multiple receptor conformations in virtual screening. *Journal of medicinal chemistry*, 47(21):5076–5084.

[Ferreira et al. 2015] Ferreira, L. G., dos Santos, R. N., Oliva, G., and Andricopulo, A. D. (2015). Molecular docking and structure-based drug design strategies. *Molecules*, 20(7):13384–13421.

[Franzini et al. 2014] Franzini, R. M., Neri, D., and Scheuermann, J. (2014). Dna-encoded chemical libraries: Advancing beyond conventional small-molecule libraries. *Accounts of Chemical Research*, 47(4):1247–1255.

[Friedman 2001] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

[Ganesan et al. 2017] Ganesan, A., Coote, M. L., and Barakat, K. (2017). Molecular dynamics-driven drug discovery: leaping forward with confidence. *Drug discovery today*, 22(2):249–269.

[Gilson et al. 2015] Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., and Chong, J. (2015). Bindingdb in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research*, 44(D1):D1045–D1053.

[Gironda-Martínez et al. 2021] Gironda-Martínez, A., Donckele, E. J., Samain, F., and Neri, D. (2021). Dna-encoded chemical libraries: A comprehensive review with succession stories and future challenges. *ACS Pharmacology & Translational Science*, 4(4):1265–1279.

[Gómez-Bombarelli et al. 2018] Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276.

[Goodnow et al. 2016] Goodnow, R. A., Dumelin, C. E., and Keefe, A. D. (2016). Dna-encoded chemistry: enabling the deeper sampling of chemical space. *Nature Reviews Drug Discovery*, 16(2):131–147.

[Guyon and Elisseeff 2003] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182.

[Halperin et al. 2002] Halperin, I., Ma, B., Wolfson, H., and Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function, and Bioinformatics*, 47(4):409–443.

[Hastie 2009] Hastie, T. (2009). The elements of statistical learning: data mining, inference, and prediction.

[Herasymenko et al. 2025] Herasymenko, O. et al. (2025). Cache challenge #2: Targeting sars-cov-2 nsp13. *Journal of Chemical Information and Modeling*.

[Homola 2008] Homola, J. (2008). Surface plasmon resonance sensors for detection of chemical and biological species. *Chemical Reviews*, 108(2):462–493.

[Jiang et al. 2022] Jiang, H., Wang, J., Cong, W., Huang, Y., Ramezani, M., Sarma, A., Dokholyan, N. V., Mahdavi, M., and Kandemir, M. T. (2022). Predicting protein–ligand docking structure with graph neural network. *Journal of chemical information and modeling*, 62(12):2923–2932.

[Jiménez et al. 2018] Jiménez, J., Škalič, M., Martínez-Rosell, G., and De Fabritiis, G. (2018). k_{deep}: Protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of Chemical Information and Modeling*, 58(2):287–296.

[Jumper et al. 2021] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589.

[Kadukova et al. 2021] Kadukova, M., Machado, K. d. S., Chacón, P., and Grudinin, S. (2021). Korp-pl: a coarse-grained knowledge-based scoring function for protein–ligand interactions. *Bioinformatics*, 37(7):943–950.

[Ke et al. 2017] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

[Kernohan and Boycott 2024] Kernohan, K. D. and Boycott, K. M. (2024). The expanding diagnostic toolbox for rare genetic diseases. *Nature Reviews Genetics*, 25(6):401–415.

[Krishna et al. 2024] Krishna, R., Wang, J., Ahern, W., Sturmfels, P., Venkatesh, P., Kalvet, I., Lee, G. R., Morey-Burrows, F. S., Anishchenko, I., Humphreys, I. R., et al. (2024). Generalized biomolecular modeling and design with rosettafold all-atom. *Science*, 384:eadl2528.

[Kryshtafovych et al. 2021] Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2021). Critical assessment of methods of protein structure prediction (casp)—round xiv. *Proteins: Structure, Function, and Bioinformatics*, 89(12):1607–1617.

[Kuntz 1992] Kuntz, I. D. (1992). Structure-based strategies for drug design. *Science*, 257:1078–1082.

[Landrum 2006] Landrum, G. (2006). Rdkit: Open-source cheminformatics. 2006. *Google Scholar*.

[Landrum 2013] Landrum, G. (2013). Rdkit: Open-source cheminformatics software. Accessed: 2026-05-11.

[Lengauer and Rarey 1996] Lengauer, T. and Rarey, M. (1996). Computational methods for biomolecular docking. *Current Opinion in Structural Biology*, 6(3):402–406.

[Li et al. 2024] Li, F., Ackloo, S., Arrowsmith, C. H., Ban, F., Barden, C. J., Beck, H., Beránek, J., Berenger, F., Bolotokova, A., Bret, G., et al. (2024). Cache challenge# 1: targeting the wdr domain of lrrk2, a parkinson's disease associated protein. *Journal of Chemical Information and Modeling*, 64(22):8521–8536.

[Li et al. 2019] Li, J., Fu, A., and Zhang, L. (2019). An overview of scoring functions used for protein–ligand interactions in molecular docking. *Interdisciplinary Sciences: Computational Life Sciences*, 11(2):320–328.

[Lima et al. 2016] Lima, T. A., Bezerra, I. C., Rocha, B. C. G., Viana, J. R., and Silva, C. R. (2016). Use of docking approaches to predict the affinity and orientation between molecules: a review. *Biophysical Reviews*, 8(2):157–165.

[Liu et al. 2007] Liu, T., Lin, Y., Wen, X., Jorissen, R. N., and Gilson, M. K. (2007). Bindingdb: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Research*, 35(Database):D198–D201.

[Liu et al. 2017] Liu, Z., Su, M., Han, L., Liu, J., Yang, Q., Li, Y., and Wang, R. (2017). Forging the basis for developing protein–ligand interaction scoring functions. *Accounts of chemical research*, 50(2):302–309.

[Lybrand 1995] Lybrand, T. P. (1995). Ligand–protein docking. *Current Opinion in Structural Biology*, 5(2):224–228.

[Machado et al. 2010] Machado, K. S., Winck, A. T., Ruiz, D. D., and de Souza, O. N. (2010). Mining flexible-receptor docking experiments to select promising protein receptor snapshots. *BMC genomics*, 11(5):1–13.

[Masters et al. 2023] Masters, M. R., Mahmoud, A. H., Wei, Y., and Lill, M. A. (2023). Deep learning model for efficient protein–ligand docking with implicit side-chain flexibility. *Journal of Chemical Information and Modeling*, 63(6):1695–1707.

[McCammon and Harvey 1987] McCammon, J. A. and Harvey, S. C. (1987). *Dynamics of Proteins and Nucleic Acids*. Cambridge University Press.

[McNutt et al. 2025] McNutt, A. T., Li, Y., Meli, R., Aggarwal, R., and Koes, D. R. (2025). Gnina 1.3: the next increment in molecular docking with deep learning. *Journal of Cheminformatics*, 17(1):28.

[Meli et al. 2022] Meli, R., Morris, G. M., and Biggin, P. C. (2022). Scoring functions for protein-ligand binding affinity prediction using structure-based deep learning: a review. *Frontiers in bioinformatics*, 2:885983.

[Meng et al. 2011] Meng, X. et al. (2011). Molecular docking: a powerful approach. *Current Computer-Aided Drug Design*, 7(2):146–157.

[Morris et al. 2009] Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., and Olson, A. J. (2009). Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry*, 30(16):2785–2791.

[Muchiri and van Breemen 2020] Muchiri, R. N. and van Breemen, R. B. (2020). Affinity selection–mass spectrometry for the discovery of pharmacologically active compounds from combinatorial libraries and natural products. *Journal of Mass Spectrometry*, 56(5).

[Murphy 2012] Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

[Murray et al. 2022] Murray, C. J. L., Ikuta, K. S., Sharara, F., Swetschinski, L., Robles Aguilar, J., Gray, A., Han, C., Bisignano, C., Rao, P., et al. (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, 399(10325):629–655.

[Nilakantan et al. 1987] Nilakantan, R., Bauman, N., Dixon, J. S., and Venkataraghavan, R. (1987). Topological torsion: a new molecular descriptor for sar applications. comparison with other descriptors. *Journal of Chemical Information and Computer Sciences*, 27(2):82–85.

[Pagadala et al. 2017] Pagadala, N. S., Syed, K., and Tuszynski, J. (2017). Software for molecular docking: a review. *Biophysical Reviews*, 9(2):91–102.

[Passaro et al. 2025] Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V. R., Getz, N., Portnoi, T., Roy, J., Stark, H., et al. (2025). Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*. Preprint.

[Petsko and Ringe 2004] Petsko, G. A. and Ringe, D. (2004). *Protein Structure and Function*. Primers in Biology. New Science Press; distributed by Oxford University Press.

[Pinzi and Rastelli 2019] Pinzi, L. and Rastelli, G. (2019). Molecular docking: shifting paradigms in drug discovery. *International Journal of Molecular Sciences*, 20(18):4331.

[Prudent et al. 2023] Prudent, R., Lemoine, H., Walsh, J., and Roche, D. (2023). Affinity selection mass spectrometry speeding drug discovery. *Drug Discovery Today*, 28(11):103760.

[Rich and Myszka 2008] Rich, R. L. and Myszka, D. G. (2008). Survey of the year 2007 commercial optical biosensor literature. *Journal of Molecular Recognition*, 21(6):355–400.

[Rogers and Hahn 2010] Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754. PMID: 20426451.

[Scaini et al. 2019] Scaini, J. L. R., Camargo, A. D., Seus, V. R., von Groll, A., Werhli, A. V., da Silva, P. E. A., and dos Santos Machado, K. (2019). Molecular modelling and competitive inhibition of a mycobacterium tuberculosis multidrug-resistance efflux pump. *Journal of Molecular Graphics and Modelling*, 87:98–108.

[Scantlebury et al. 2020] Scantlebury, J. et al. (2020). Data set augmentation allows deep learning-based virtual screening to better generalize to unseen target classes. *Journal of Chemical Information and Modeling*, 60(8):3722–3730.

[Schwaller et al. 2019] Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., and Lee, A. A. (2019). Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583.

[Senior et al. 2020] Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577:706–710.

[Shen et al. 2020] Shen, C., Ding, J., Wang, Z., Cao, D., Ding, X., and Hou, T. (2020). From machine learning to deep learning: Advances in scoring functions for protein–ligand docking. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 10(1):e1429.

[Škrinjar et al. 2026] Škrinjar, P., Eberhardt, J., Studer, G., et al. (2026). Evaluating generalization in protein–ligand cofolding methods. *Nature Structural & Molecular Biology*.

[Stepniewska-Dziubinska et al. 2018] Stepniewska-Dziubinska, M. M., Zielenkiewicz, P., and Siedlecki, P. (2018). Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*, 34(21):3666–3674.

[Su et al. 2018] Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y., and Wang, R. (2018). Comparative assessment of scoring functions: the casf-2016 update. *Journal of chemical information and modeling*, 59(2):895–913.

[Svetnik et al. 2003] Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958.

[Tripos 2005] Tripos, S. (2005). Tripos mol2 file format.

[Trott and Olson 2010] Trott, O. and Olson, A. J. (2010). Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461.

[Varadi et al. 2022] Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). Alphafold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444.

[Vaswani et al. 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[Wallach and Heifets 2018] Wallach, I. and Heifets, A. (2018). Most ligand-based classification benchmarks reward memorization rather than generalization. *Journal of chemical information and modeling*, 58(5):916–932.

[Wang et al. 2004] Wang, R., Fang, X., Lu, Y., and Wang, S. (2004). The pdbbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry*, 47(12):2977–2980.

[Wang et al. 2025] Wang, X. et al. (2025). Enantioselective protein affinity selection mass spectrometry (e-asms). *Nature Communications*, 17(651).

[Weininger 1988] Weininger, D. (1988). Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36.

[Werhli et al. 2025] Werhli, A. V., Lopes, P. P., Arrua, O. E., Aderhold, A., and Machado, K. d. S. (2025). Crrf-score-cumulative ranking random forest scoring function for free energy of binding prediction in protein-ligand docking. In *Brazilian Conference on Intelligent Systems*, pages 199–213. Springer.

[Westbrook and Fitzgerald 2003] Westbrook, J. D. and Fitzgerald, P. (2003). The pdb format, mmcif, and other data formats. *Methods Biochem Anal*, 44:161–179.

[Xu et al. 2026] Xu, S., Feng, Q., Qiao, L., Wu, H., Shen, T., Cheng, Y., Zheng, S., and Sun, S. (2026). Benchmarking all-atom biomolecular structure prediction with foldbench. *Nature Communications*, 17:442.

[Xu 2024] Xu, W. (2024). Current status of computational approaches for small molecule drug discovery. *Journal of Medicinal Chemistry*, 67(21):18633–18636.

[Xu et al. 2025] Xu, Z. et al. (2025). A generative ai-discovered tnik inhibitor for idiopathic pulmonary fibrosis. *Nature Medicine*, 31:2602–2610.

[Zheng et al. 2025] Zheng, Y., Koh, H. Y., Ju, J., Yang, M., May, L. T., Webb, G. I., Li, L., Pan, S., and Church, G. (2025). Large language models for drug discovery and development. *Patterns*, 6(10).

