

*Uma Abordagem para o Uso de Aprendizado de Máquina em Bases Abertas para Apoiar a Predição de Distribuição de Fármacos*

**Helder Pestana**

XXXX / 20XX

Projeto de pesquisa de Mestrado em Ciência da

Computação

Uma Abordagem para o Uso de Aprendizado de Máquina em Bases Abertas para Apoiar a Predição de Distribuição de Fármacos

Esse documento corresponde ao Projeto de Pesquisa apresentado à Banca Examinadora para qualificação no curso de Mestrado em Ciência da Computação da Faculdade Campo Limpo Paulista.

Campo Limpo Paulista, XX de XXXX de 20XX.

Helder Pestana

Orientador: Prof. Dr. Rodrigo Bonacin

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001

**Resumo**. Aqui o aluno deve escrever um breve resumo do trabalho que pretende executar, apresentando claramente o problema e sua contribuição.

**Abstract**: Aqui o aluno deve escrever um breve resumo do trabalho em inglês que pretende executar, apresentando claramente o problema e sua contribuição.

Sumário

|  |  |  |
| --- | --- | --- |
| 1. | Introdução | 1 |
| 2. | Detalhes do Documento | 1 |
|  | 2.1. Figuras, Tabelas e Legendas | 1 |
|  | 2.2. Incluindo Referências | 2 |
|  | Referências | 2 |
|  |  |  |

Glossário

Lista de Tabelas

|  |  |  |
| --- | --- | --- |
| 1. | Exemplo de tabela. | 2 |

Lista de Figuras

|  |  |  |
| --- | --- | --- |
| 1 | Uma figura típica | 2 |

# Introdução

A eficácia terapêutica de um medicamento envolve fatores como a biodisponibilidade, ou seja a quantidade de um medicamento que realmente é absorvida pelo corpo humano e se torna disponível para uso, e a absorção que é processo pelo qual um medicamento é transferido do local de administração para a corrente sanguínea, tornando-se disponível para distribuição para os tecidos e órgãos do corpo. Estes fatores dependem da estrutura química do fármaco, suas propriedades físico-químicas, interações com proteínas e tecidos biológicos, entre outros (Shargel,1999).

Pesquisadores do campo da farmacologia desenvolveram bases de dados abertas contendo conjuntos de dados acessíveis publicamente que fornecem informações sobre uma ampla gama de drogas, incluindo aquelas desenvolvidas pela indústria farmacêutica, bem como substâncias experimentais e drogas ilícitas (Wishart,2006). Essas bases de dados geralmente incluem uma variedade de informações, como estrutura química dos fármacos, propriedades físico-químicas, dados de ensaios clínicos e pré-clínicos, interações medicamentosas, etc. A análise dessas bases de dados tem o potencial de identificar padrões e correlações, assim os algoritmos de Aprendizado de Máquina (AM) podem ajudar a prever como os medicamentos são distribuídos nos tecidos e órgãos do corpo humano, permitindo avaliar a eficácia e a segurança dos compostos em estudo, contribuindo para a tomada de decisões e para a redução do tempo e dos custos envolvidos na produção farmacêutica.

O propósito deste projeto de pesquisa é aliar a capacidade dos algoritmos de AM, incluindo Aprendizagem Profunda (AP), com as informações contidas em bases abertas no campo da farmacologia para apoiar a predição de distribuição de fármacos.

## 1.1. Contexto e Motivação

Predizer os efeitos clínicos de um determinado medicamento é essencial para o desenvolvimento de medicamentos eficazes e seguros.

*Drug Delivery* (distribuição de fármacos, como é conhecida no contexto farmacêutico em português) concentra-se na compreensão de como os medicamentos são transportados pelo corpo após sua administração e a concentração do fármaco nos tecidos-alvo (Pattni,2015). O propósito é evitar que na administração destes medicamentos ocorra subdosagens em locais críticos ou toxicidades elevadas em locais não desejados, o que compromete a ação do fármaco e aumenta o risco de efeitos adversos.

Na abordagem mais comumente utilizada no desenvolvimento de formulações depende de processos incertos de tentativa e erro, exigindo assim um grande número de experimentos *in vitro* e *in vivo* que consomem muitos recursos e tempo (Shargel,1999). Entre as dificuldades deste processo está o desenvolvimento de medicamentos bom boa distribuição de fármacos. Assim, uma forma de acelerar o processo de desenvolvimento de medicamentos é otimiza-la, identificando como diferentes compostos interagem com o corpo humano e como suas características físico-químicas influenciam sua distribuição e eficácia terapêutica. Isso pode acelerar o processo ao identificar candidatos promissores de maneira mais rápida e eficiente, e otimizar suas formulações para maximizar sua eficácia clínica.

O aumento da capacidade computacional, a evolução dos algoritmos de AM e o conjunto de dados disponíveis ofertados em bases abertas motivaram o desenvolvimento deste projeto. Espera-se que por meio destes recursos será possível ofertar soluções que otimizem o tempo e reduzam o custo, possibilitando a fabricação de fármacos de forma mais eficiente e mais acessível.

## 1.2. Problemática e Justificativa

Uma abordagem baseada em AM para o apoio à predição de propriedades ligadas a distribuição de fármacos encontra obstáculos como a falta de qualidade dos dados, qualidade do modelo de aprendizagem e uma compreensão profunda da natureza dos modelos de Inteligência Artificial (IA) (Tran TTV et al. 2023). Embora as fontes de dados públicos para Pesquisa e Desenvolvimento (P&D) de medicamentos tenha aumentando significativamente nos últimos anos, os algoritmos de IA necessitam não apenas de uma quantidade considerável de dados, mas também de dados de alta qualidade e integrados para criar modelos precisos.

Para além da qualidade dos dados, é crucial ter um modelo de aprendizagem apropriado para explorar o potencial da IA na previsão de propriedades distributivas. Isso inclui investigar e construir novos modelos baseados em AM e AP.

Outro desafio enfrentado é a dificuldade em compreender a natureza dos modelos de IA. Apesar do desempenho dos modelos de previsão de propriedades distributivas baseados em IA, ainda há uma carência de interpretações mais precisas. Isso torna difícil avaliar e confiar nas hipóteses geradas pela IA devido à sua natureza de "caixa preta", dificultando o aprimoramento do modelo e a otimização de compostos com propriedades de distribuição indesejáveis. Assim como na descoberta de medicamentos, a distribuição de medicamentos é uma disciplina multidisciplinar, para estudar e construir modelos preditivos de características de distribuição é preciso ter conhecimentos relevantes em áreas como biologia, bioinformática, farmacologia, química e informática química. Sendo necessário, portanto, equipes multidisciplinares.

Uma abordagem crucial para enfrentar os desafios na previsão da distribuição de medicamentos com base em IA é otimizar a parametrização dos algoritmos utilizados. Isso requer um foco específico na melhoria tanto da qualidade quanto da quantidade dos dados empregados para treinar e validar os modelos. É fundamental incorporar uma diversidade mais ampla de fontes de dados, como ensaios clínicos, registros eletrônicos de saúde e evidências do mundo real. Além disso, a implementação de técnicas avançadas de limpeza e pré-processamento de dados se mostra essencial para mitigar o ruído e o viés inerentes aos conjuntos de dados, resultando em previsões mais precisas e confiáveis. Além disso, é essencial maximizar a interpretabilidade das previsões geradas pelos modelos de IA. Para isso, torna-se necessário desenvolver métodos que permitam visualizar e interpretar os mecanismos subjacentes às previsões do modelo, aumentando assim a confiança nos resultados e facilitando a tomada de decisões no âmbito do desenvolvimento de medicamentos.

A literatura atual sobre o tema tem abordado principalmente aspectos relacionados ao uso de algoritmos de AM e AP e sua aplicabilidade no desenvolvimento de soluções para distribuição de medicamentos em contextos específicos, no entanto, uma lacuna significativa persiste em relação ao entendimento dos algoritmos utilizados, bem como a forma de parametrização e ajuste destes algoritmos na construção de um modelo confiável. Muitos estudos também não evidenciaram os resultados obtidos a partir do uso desses modelos e o ganho de tempo com o uso desta ferramenta. O presente trabalho visa preencher essas lacunas e investigar detalhadamente a relação entre os algoritmos utilizados e o ganho, utilizando métodos e abordagens inovadoras para explorar essas interações. Isso contribuirá para uma compreensão mais completa do uso de AM e AP em bases abertas como ferramenta de apoio a predição na distribuição de fármacos.

## 1.3. Objetivos, Contribuições e Métodos

O objetivo geral deste trabalho é pesquisar e propor uma solução que possa predizer propriedades ligadas à distribuição de fármacos por meio de modelos de AM e AP treinados utilizando bases internacionais abertas em vistas à otimizar a produção de fármacos.

Portanto, este trabalho se propõe a responder a seguinte pergunta: “Como criar uma abordagem e solução baseada em AM (e AP) que faça uso de bases internacionais abertas para apoiar a predição de propriedades ligadas à distribuição de fármacos?”. A partir da pergunta principal, outras perguntas específicas são abordadas:

* “Quais bases de dados abertas serão úteis ao treinamento de modelos de AM e AP?”
* “Quais técnicas de seleção e pré-processamento deverão ser aplicadas sobre estes dados antes de utiliza-los para treinar os modelos?”
* “Como codificar (*embedding*) as informações (moléculas) das bases abertas de modo a serem utilizadas pelos algoritmos?”
* “Como escolher os algoritmos adequados para o modelo de aprendizagem de máquina segundo o objetivo proposto?”
* “Como ajustar os hiper parâmetros dos algoritmos de AM e AP?”
* “Como mensurar a eficácia do modelo? Quais métricas serão necessárias para este propósito?”
* “De que forma apresentar os resultados e maximizar a interpretabilidade dos modelos?”

A partir deste objetivo principal e das questões de pesquisa, definimos as seguintes metas:

1. Revisar a literatura sobre o uso de algoritmos de ML (e DL - *Deep Learning*) para a predição de propriedades ligadas a distribuição de fármacos.
2. Avaliar bases abertas nacionais e internacionais existentes que podem ter dados de interesse para treinamento, validação e testes dos modelos.
3. Aplicar o processo de ETL (Extração, Transformação e Carga) sobre estes dados no intuito de deixá-los formatados aos algoritmos.
4. Conceber e criar modelos candidatos (de AM e AP) ajustando adequadamente os hiper parâmetros dos algoritmos.
5. Validar, ajustar e efetuar os testes finais nos algoritmos com o melhor desempenho.
6. Analisar os resultados obtidos e bem como a interpretabilidade dos modelos com maior desempenho.

Assim o presente trabalho se propõe a criação e otimização de modelos de AM e AP específicos para analisar dados de fármacos em bases abertas, isso inclui o desenvolvimento de novos modelos e métodos para lidar com os desafios únicos apresentados por esses dados, avaliando de forma abrangente o desempenho de diferentes algoritmos de aprendizado de máquina na tarefa de análise de dados de fármacos em bases abertas, incluindo comparações de precisão, eficiência computacional e robustez em diferentes cenários de aplicação.

Espera-se que com este trabalho seja possível identificar padrões importantes nos dados de fármacos, fornecendo *insights* valiosos para a comunidade científica e com isso ajudar a identificar e produzir medicamentos com melhor distribuição de fármacos. Além disso, este trabalho pode resultar no desenvolvimento de ferramentas e recursos computacionais para facilitar a análise de dados sobre distribuição de fármacos em bases abertas.

Enfim esta pesquisa irá contribuir para a aplicação prática de algoritmos e desenvolvimento de modelos de AM e AP na descoberta de fármacos, envolvendo a identificação de candidatos a medicamentos promissores do ponto de vista de distribuição de fármacos.

## 1.4. Estrutura da Proposta

A estrutura dos capítulos restantes desta dissertação é a seguinte:

* **Capitulo 2 – Referencial Teórico e Metodológico:** Apresenta a fundamentação dos principais conceitos e tecnologias abordados nesta pesquisa, tais com ADMET (Absorção, Distribuição, Metabolismo e Excreção e Toxicidade), AM e AP, bem como Bases de Dados Abertas contendo dados químicos e farmacológicos.
* **Capitulo 3 - Revisão da Literatura e Trabalhos Relacionados:** Apresenta um estudo abrangente sobre a pesquisa em temas relacionados por meio de uma revisão da literatura, enfatizando a metodologia utilizada e os resultados mais significativos alcançados, bem como a discussão dos trabalhos relacionados.
* **Capitulo 4 - O Uso de Aprendizado de Máquina no Apoio à Predição de Distribuição de Fármacos:** apresenta a metodologia de pesquisa, a base de dados e a seleção de atributos, as fases de pré-processamento e codificação, modelos e implementação e a descrição formal do problema, execução e métricas de avaliação.
* **Capitulo 5 - Análise dos Resultados e Discussão:** apresenta os resultados experimentais obtidos, bem como uma discussão sobre as implicações dos resultados, resultados de outros estudos, aplicabilidade do modelo na pratica, limitações e desafios futuros.
* **Capitulo 6 – Conclusão:** apresenta as contribuições, limitações, trabalhos futuros e considerações finais desta dissertação.
* **Capitulo 7 - Cronograma de Atividades**: apresenta as atividades que serão desenvolvidas entre a qualificação e a defesa por meio de um diagrama de gantt.

# Referencial Teórico e Metodológico

Neste capítulo, é apresentada a fundamentação teórica e metodológica, além das tecnologias utilizadas no desenvolvimento desta dissertação. O conteúdo deste capítulo é embasado na revisão de trabalhos relacionados ao tema central desta pesquisa.

A seção 2.1 aborda os conceitos básicos sobre o desenvolvimento de fármacos, as propriedades farmacológicas e a distribuição dos fármacos no organismo humano; na seção 2.2 são explicitados conceitos sobre AM e AP; na seção 2.3 são detalhadas as principais bases de dados abertas sobre dados químicos e farmacológicos (DrugBank, PubChem, etc.) e o SMILES (*Simplified Molecular-Input Line-Entry System*) , padrão de representação de estruturas químicas de compostos orgânicos e inorgânicos por meio de uma linha de texto simples.

## Desenvolvimento de fármacos, propriedades farmacológicas e distribuição de fármacos

O desenvolvimento de fármacos é um processo complexo e de elevado custo financeiro que envolve diversas etapas, desde a descoberta inicial até a disponibilização do medicamento no mercado. Segundo Franke (2020) para compreendermos o processo inicialmente é importante entender quatro conceitos básicos:

* Droga: Qualquer substancia que interaja com o organismo produzindo algum efeito;
* Fármaco: Uma substancia definida com propriedades ativas produzindo efeito terapêutico;
* Medicamento: É quando no fármaco são adicionados todos os componentes (incipientes) para que este seja administrado terapeuticamente;
* Forma Farmacêutica: É a forma final de como um medicamento se apresenta: comprimido, capsulas, injetáveis, etc.

Quanto às etapas no desenvolvimento de fármacos (Franke 2020), podem ser subdivididas nos respectivos períodos:

* Período de descoberta da droga:
  + Descoberta de Fármacos: Nesta fase, os pesquisadores identificam e caracterizam moléculas que têm potencial para se tornarem medicamentos. Isso pode envolver triagem de compostos químicos, estudos de biologia molecular e modelagem computacional.
  + Desenvolvimento Pré-clínico: Os compostos promissores identificados na fase de descoberta passam por testes pré-clínicos em laboratório e em animais para determinar sua segurança, eficácia e toxicidade. Esta fase também inclui a otimização da formulação do medicamento.
* Período de desenvolvimento do fármaco:
  + Ensaios Clínicos: Se um composto passa nos testes pré-clínicos, ele avança para ensaios clínicos em seres humanos. Estes ensaios são divididos em três fases (Fase I, Fase II e Fase III) e visam determinar a segurança, eficácia e dosagem adequada do medicamento.
  + Submissão Regulatória: Após concluir os ensaios clínicos, os dados são submetidos às agências regulatórias, como a Food and Drug Administration (FDA) nos Estados Unidos ou a Agência Nacional de Vigilância Sanitária (ANVISA) no Brasil, para revisão e aprovação.
* Período de fabricação, comercialização e farmacovigilância (Fase IV):
  + Fabricação: Uma vez aprovado, o medicamento é fabricado em larga escala de acordo com padrões rigorosos de qualidade e boas práticas de fabricação (GMP, Good Manufacturing Practices).
  + Comercialização: Após a aprovação regulatória e a fabricação em escala comercial, o medicamento é lançado no mercado, onde é distribuído para farmácias, hospitais e outros pontos de venda.
  + Farmacovigilância: Após o lançamento no mercado, o medicamento continua sendo monitorado para identificar e avaliar quaisquer efeitos colaterais ou problemas de segurança que possam surgir.

Para compreendermos como ocorre a distribuição de fármacos no organismo, precisamos entender o conceito de farmacocinética (Pereira 2007) que consiste no estudo do movimento de uma substancia química, em particular, um fármaco no interior de um organismo vivo. Na farmacocinética existem propriedades farmacológicas importantes como a ADMET.

A absorção é processo pelo qual o fármaco entra na corrente sanguínea a partir do local de administração. A distribuição é como o fármaco se espalha pelos tecidos do corpo, influenciado por fatores como perfusão sanguínea, ligação a proteínas plasmáticas e características da própria molécula do fármaco. O metabolismo é o processo pelo qual o fármaco é modificado no organismo, geralmente ocorrendo no fígado, para facilitar sua excreção. A excreção é a remoção do fármaco e seus metabólitos do organismo, principalmente pelos rins e a toxicidade refere-se à capacidade de uma substância causar danos a um organismo vivo.

Na distribuição de fármacos, uma série de fatores desempenham um papel crucial, incluindo barreiras biológicas, biodisponibilidade e meia-vida. No organismo, os fármacos enfrentam desafios como a superação da barreira hematoencefálica e da barreira placentária, conforme sua aplicação. A biodisponibilidade refere-se à fração do fármaco administrado que alcança a circulação sistêmica, pronta para desencadear seus efeitos farmacológicos. Por sua vez, a meia-vida representa o período necessário para que a concentração do fármaco no sangue seja reduzida pela metade, o que pode impactar diretamente na frequência requerida para sua dosagem.

## Aprendizado de Máquina e Aprendizagem Profunda

O AM é um campo que se destaca por sua abordagem de aprendizagem da IA, focando no desenvolvimento de algoritmos capazes de representar eficientemente os conjuntos de dados (Choi et al. 2020). Ao contrário da programação clássica, onde algoritmos são codificados de forma explícita utilizando características conhecidas, o AM emprega subconjuntos de dados para criar algoritmos que podem explorar combinações novas ou diferentes de características e pesos. Na AM, quatro métodos de aprendizagem são comumente utilizados, cada um adequado para resolver diferentes tipos de tarefas: supervisionada, não supervisionada, semi-supervisionada e aprendizado por reforço.

### Aprendizagem Supervisionada

Rimal (2024) ilustra a aprendizagem supervisionadada em um cenário de uma empresa imobiliaria que deseja prever o preço de uma casa com base em características específicas. Para iniciar, a empresa coletaria um conjunto de dados contendo diversas instâncias, cada uma representando uma observação singular de uma casa e suas características associadas. Essas características incluem propriedades registradas da casa que podem ser relevantes para a predição de preços, como área total, número de andares e presença de quintal. O alvo a ser previsto é o preço do imóvel.

Geralmente, os conjuntos de dados são divididos em conjuntos de treinamento, validação e teste, sendo que os modelos tendem a ter um desempenho melhor nos dados nos quais foram treinados. A aprendizagem supervisionada utiliza padrões presentes no conjunto de treinamento para mapear características para o alvo, permitindo que um algoritmo faça previsões, tais como preços de imóveis, em conjuntos de dados futuros.

Essa abordagem é denominada supervisionada pois o modelo infere um algoritmo a partir de pares de características e alvo, sendo informado pelo alvo se suas previsões foram corretas. Em outras palavras, as características (X) são mapeadas para o alvo (Y), aprendendo a função de mapeamento (F), de modo que os preços futuros dos imóveis possam ser aproximados utilizando o algoritmo desenvolvido.

A avaliação da performance do algoritmo ocorre no conjunto de dados de teste, que consiste em dados não vistos pelo algoritmo durante o treinamento. Os passos fundamentais do aprendizado de máquina supervisionado são os seguintes:

* Aquisição e Divisão dos Dados: O primeiro passo é adquirir um conjunto de dados e dividi-lo em conjuntos separados de treinamento, validação e teste.
* Modelagem e Ajuste: Em seguida, utiliza-se os conjuntos de treinamento e validação para informar o modelo sobre a relação entre as características e o alvo.
* Avaliação da Performance: O modelo é avaliado utilizando o conjunto de dados de teste para determinar sua capacidade de prever com precisão os preços dos imóveis para instâncias não vistas.

Durante cada iteração, o desempenho do algoritmo nos dados de treinamento é comparado com seu desempenho no conjunto de validação. O algoritmo é então ajustado com base na avaliação feita no conjunto de validação. No entanto, é importante ressaltar que a generalização do desempenho do algoritmo pode variar, uma vez que o conjunto de validação pode diferir do conjunto de teste.

As tarefas mais comuns no aprendizado supervisionado são a regressão e a classificação. Na regressão, o objetivo é prever dados numéricos, como notas de testes, valores de laboratório ou preços de itens, semelhante ao exemplo de preços de imóveis. Por outro lado, na classificação, o objetivo é prever a categoria à qual uma instância pertence.

Tomando como exemplo a previsão de faixas de preços para venda de imóveis em um mercado volátil, os cientistas de dados poderiam transformar a variável alvo numérica em uma variável categórica, agrupando os preços dos imóveis em classes separadas, como (0, 125 mil), (125 mil, 250 mil), (250 mil, 375 mil) e (375 mil, ∞). Essas classes seriam ordinais, ou seja, há uma ordem natural associada às categorias. No entanto, se a tarefa fosse determinar se as casas tinham revestimento de madeira, plástico ou metal, as classes seriam nominais, independentes umas das outras e sem uma ordem natural.

### Aprendizagem Não Supervisionada

Em contraste com o aprendizado supervisionado, o aprendizado não supervisionado visa identificar padrões em conjuntos de dados e categorizar instâncias individuais nesses padrões. Esses algoritmos são chamados de não supervisionados porque os padrões presentes ou ausentes nos dados não são informados por um alvo específico, sendo deixados para serem determinados pelo algoritmo.

Algumas das tarefas mais comuns do aprendizado não supervisionado incluem agrupamento, associação e detecção de anomalias. No agrupamento, como o próprio nome sugere, as instâncias de um conjunto de dados são agrupadas em clusters distintos com base em combinações específicas de suas características.

### Aprendizagem Semi-Supervisionada

A aprendizagem semi-supervisionada representa um ponto intermediário entre os métodos supervisionados e não supervisionados, sendo particularmente valiosa em conjuntos de dados que combinam dados rotulados e não rotulados. Este cenário é comum em situações onde a rotulagem de dados, especialmente em imagens, é trabalhosa ou onerosa. Em campos como imagens médicas, por exemplo, um médico pode rotular apenas uma pequena fração das imagens disponíveis, utilizando-as para treinar um modelo. Esse modelo, então, é empregado para classificar o restante das imagens não rotuladas, ampliando assim o conjunto de dados rotulados. Posteriormente, este conjunto expandido é utilizado para treinar um modelo final, com a expectativa de superar abordagens puramente não supervisionadas.

### Aprendizagem por Reforço

A aprendizagem por reforço é uma técnica que visa treinar algoritmos para realizar tarefas específicas em cenários onde não há uma resposta única correta, mas sim um resultado geral desejado. É considerada uma abordagem próxima à forma como os seres humanos aprendem, uma vez que envolve tentativa e erro, em vez de depender apenas de dados. Embora seja uma ferramenta poderosa, e com boas perspectivas, suas aplicações na saúde ainda são limitadas.

Imagine que queremos treinar um algoritmo para um determinado jogo. Não existe uma sequência única de comandos de controle; há várias combinações que levam à vitória e outras que não. Na aprendizagem por reforço, o algoritmo é deixado para "jogar" por conta própria. Ele tenta diversas combinações de comandos de controle e, quando consegue avançar sem sofrer danos, é "recompensado" (ou seja, seu comportamento é reforçado). Ao longo desse processo, o algoritmo começa a discernir quais ações são mais vantajosas por meio de recompensas e ajustes (por exemplo, avançar é preferível a recuar, pular sobre inimigos é melhor do que colidir com eles). Com o tempo, o algoritmo aprende a completar os níveis com sucesso.

### Redes Neurais e Aprendizagem Profunda

Uma Rede Neural Artificial (RNA) é um algoritmo de aprendizado de máquina que se inspira nas complexas redes neurais encontradas no cérebro humano. Em uma RNA, cada nó desempenha um papel semelhante ao de um neurônio biológico, comunicando-se com outros nós por meio de conexões, que podem ser comparadas aos axônios e dendritos encontrados no sistema nervoso (Ferneda 2006).

Assim como no cérebro humano, onde as sinapses entre neurônios são fortalecidas quando os neurônios têm saídas correlacionadas (de acordo com a teoria Hebbiana (Abreu et. al. 2020) que postula "células nervosas que se disparam juntas, se conectam juntas"), as conexões entre nós em uma RNA são ponderadas com base na contribuição que cada uma oferece para alcançar o resultado desejado. Esse processo permite que a rede aprenda e adapte seus pesos ao longo do tempo, otimizando seu desempenho na realização de tarefas específicas.

### Redes Neurais *Feedforward*

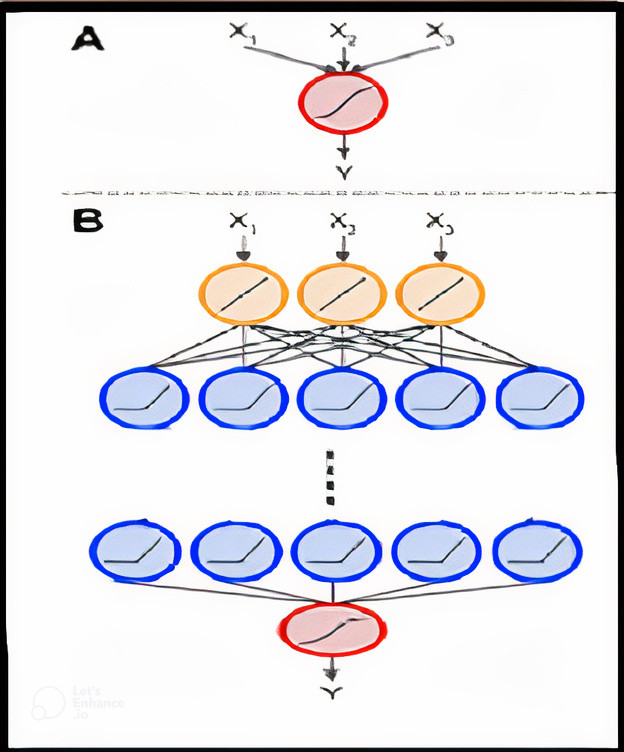
Um *perceptron* é um algoritmo de aprendizado de máquina que recebe um conjunto de características e seus alvos como entrada, buscando identificar uma linha, plano ou hiperplano que separe as classes em um espaço bidimensional, tridimensional ou hiper dimensional, respectivamente. Essas características são então transformadas usando a função sigmoide.

Quando vários *perceptrons* são conectados (Figura 2), formando um modelo conhecido como *perceptron* multicamadas RNA, é comum incluir uma camada de nós de entrada, uma camada de nós de saída e uma ou mais "camadas ocultas" entre elas. Enquanto em RNA simples pode haver uma camada de entrada, zero a três camadas ocultas e uma camada de saída, redes neurais profundas podem conter dezenas ou até centenas de camadas ocultas.

Em redes neurais *feedforward*, que são as mais comuns, as informações fluem em uma única direção, da camada de entrada para a camada de saída. Cada nó em uma camada subsequente recebe informações de todos os nós na camada anterior, realiza transformações e repassa os resultados para a próxima camada. Já em redes neurais recorrentes, as informações podem ser passadas entre nós dentro da mesma camada ou retroceder para camadas anteriores, permitindo a retroalimentação de dados.

Cada camada em uma RNA pode conter qualquer número de nós, sendo que o número de nós na camada de saída geralmente corresponde ao número de classes sendo previstas. Por exemplo, em classificação multiclasse, há um nó de saída para cada classe, enquanto em classificação binária há um único nó com uma função de ativação sigmoide. Para tarefas de regressão, a camada de saída pode usar uma função de ativação linear. Essas funções de ativação transformam a entrada de um nó em uma saída desejada, e cada nó em uma RNA possui uma função de ativação associada, não se limitando apenas à camada de saída.

Essas funções de ativação, embora não sejam sempre lineares, não precisam ser intrinsecamente complexas. Por exemplo, a unidade linear retificada (ReLU - *Rectified Linear Unit*) aplica uma transformação linear às entradas ≥ 0 e define as entradas < 0 como 0. Essa simplicidade é sua força: ao longo das camadas de uma RNA, as entradas são progressivamente ajustadas, de modo que, na camada final, elas podem não se parecer mais com o estado original. No entanto, essa representação final da entrada é teoricamente mais preditiva do resultado desejado.



**Figura 1**. -Perceptron(A) e Rede Neural Profunda(B) (Adaptado de Choi et al. 2020)

### Redes Neurais Convolucionais

Quando utilizamos uma rede neural *feedforward* para reconhecer imagens, cada pixel é tratado como uma entrada separada. No entanto, essa abordagem tem uma limitação: não leva em consideração a relação entre os pixels vizinhos. Por exemplo, em uma imagem, pixels próximos tendem a estar mais correlacionados do que pixels distantes, o que significa que a informação espacial importante pode ser perdida.

É aqui que entra a rede neural convolucional (CNN - *Convolutional Neural Network*). Ao contrário da RNA tradicional, a CNN (Silva 2018) preserva o contexto espacial das características da imagem. Em vez de alimentar cada pixel separadamente, a CNN utiliza "patches" ou pequenas regiões da imagem, permitindo que ela capture relações espaciais entre os pixels. Esses patches são processados por filtros convolucionais, que aprendem a extrair características específicas da imagem.

As convoluções são amplamente utilizadas no campo do processamento de imagens por sua capacidade de desfocar, realçar ou detectar bordas. Por exemplo, ao aplicar um filtro convolucional, podemos realçar as bordas de um objeto na imagem. Uma imagem digital em escala de cinza é representada por uma única matriz, enquanto uma imagem colorida é representada por três matrizes empilhadas para os canais de cor vermelho, verde e azul.

Um filtro convolucional é uma matriz menor que é deslizada sobre a imagem original. Em cada posição, ocorre uma multiplicação elemento a elemento entre o filtro e a região correspondente da imagem. A saída dessa operação é um novo conjunto de valores que representam características específicas da imagem. Em resumo, as CNNs são poderosas para o reconhecimento de imagens, pois preservam o contexto espacial e permitem a extração eficiente de características importantes para a tarefa em questão.

Nas CNNs, os filtros são treinados para identificar características específicas em imagens, como linhas verticais ou objetos com formato de U, e registrar sua localização no mapa de características. Uma CNN profunda usa esses mapas de características como entrada para a camada seguinte, que aplica novos filtros para gerar outro mapa de características. Esse processo continua por várias camadas, e à medida que avança, as características extraídas se tornam mais abstratas, mas também mais úteis para a previsão. No final, os mapas de características são comprimidos e passados para uma rede neural *feedforward*, onde ocorre a classificação da imagem com base nas características e texturas extraídas. Esse processo é conhecido como aprendizado profundo (DL - *Deep Learning*).

Além da classificação de imagens, o DL para dados tabulares aplica redes neurais profundas para analisar e identificar padrões em conjuntos de dados estruturados, como planilhas e bancos de dados relacionais. Ao contrário dos dados não estruturados, como imagens ou texto, os dados tabulares possuem colunas com tipos de dados específicos (numéricos, categóricos, etc.). Modelos de DL, como redes neurais totalmente conectadas (fully connected networks) e arquiteturas avançadas como TabNet, são treinados para capturar relações complexas e interações entre as colunas, frequentemente superando métodos tradicionais de ML em precisão e capacidade de generalização. Esses modelos são utilizados em tarefas como classificação, regressão e previsão, proporcionando uma abordagem poderosa para a análise de grandes volumes de dados tabulares.

## Bases de Dados Abertas de dados químicos e farmacológicos

As bases de dados abertas de química e farmacologia servem como ferramentas essenciais para a pesquisa, descoberta e desenvolvimento de novos medicamentos, bem como para o avanço do conhecimento científico nessas áreas. Elas proporcionam acesso a uma vasta quantidade de informações que são fundamentais para avançar no entendimento e tratamento de doenças.

Existem diversas bases de dados importantes, abaixo algumas das principais:

* PubChem: É uma base de dados mantida pelo National Center for Biotechnology Information (NCBI) que armazena informações sobre compostos químicos, incluindo estruturas, propriedades químicas, atividades biológicas e referências bibliográficas (Sunghwan et al. 2016).
* ChEMBL: É uma base de dados que se concentra em dados de atividade biológica para compostos químicos, oferecendo informações sobre alvos moleculares, bioensaios, atividades biológicas e referências (Anna et al. 2017)
* DrugBank: É uma base de dados que contém informações detalhadas sobre medicamentos, incluindo suas estruturas químicas, mecanismos de ação, interações, propriedades farmacocinéticas, indicações clínicas e referências bibliográficas (David S et al. 2018).
* ChemSpider: É uma base de dados de propriedades químicas mantida pela Royal Society of Chemistry, que contém informações sobre milhões de compostos químicos, incluindo estruturas, propriedades físicas e químicas, identificadores e referências cruzadas para literatura científica (Williams A et al. 2010).
* ZINC: É uma base de dados fundamental para a triagem virtual e descoberta de novos medicamentos (B. K. et al. 2005), oferecendo uma vasta quantidade de informações sobre compostos comercialmente disponíveis. O banco de dados é especialmente útil para biólogos estruturais e químicos medicinais, permitindo triagem rápida e eficiente de hipóteses. As informações dos compostos estão disponíveis em vários formatos, como SMILES, mol2, SDF 3D, entre outros, facilitando a integração com diferentes ferramentas de triagem virtual e docking molecular (técnica computacional utilizada para prever a orientação preferencial de uma molécula, geralmente uma pequena molécula chamada ligante, quando ligada a uma segunda molécula geralmente uma proteína, chamada receptor).

Portanto, é possível identificar características especificas que diferenciam cada uma das bases de dados abertas. Para o propósito deste projeto será utilizada a base de dados DrugBank como fonte principal, e as outras como apoio.

### DrugBank

O DrugBank surgiu no ano de 2006 como uma plataforma digital com um banco de dados habilitado para a web que contém informações moleculares abrangentes sobre drogas, os mecanismos, as interações e respectivos alvos. O DrugBank contém informações detalhadas sobre 1.480 medicamentos aprovados pela FDA dos EUA, abrangendo 28.447 nomes de marcas e sinônimos (Wishart, D. S. 2008). Esta coleção inclui 1.281 medicamentos sintéticos de pequenas moléculas, 128 medicamentos biotecnológicos e 71 suplementos ou medicamentos nutracêuticos (substâncias derivadas de alimentos que têm efeitos positivos na saúde além do seu valor nutricional básico). Além disso, dispõe de dados sobre 1.669 alvos diferentes (moléculas de proteínas, lipídios ou DNA) e enzimas metabolizadoras com as quais esses medicamentos interagem. A base de dados também cobre 187 drogas ilícitas (aquelas legalmente proibidas ou seletivamente proibidas na maioria dos países desenvolvidos) e 64 drogas retiradas do mercado devido a questões de segurança.

As informações químicas, farmacêuticas e biológicas contidas no DrugBank são essenciais não só para a compreensão das reações adversas dos medicamentos, mas também para prever se uma nova entidade medicamentosa pode apresentar semelhanças químicas ou funcionais inesperadas com uma droga perigosa ou altamente viciante. A base de dados é manualmente curada por uma equipe de bioinformáticos, farmacêuticos, bioquímicos e médicos. Atualizações do DrugBank são normalmente lançadas a cada seis meses, incluindo informações sobre medicamentos recentemente aprovados, correções ou atualizações de medicamentos antigos, a adição de novos campos de dados e melhorias na interface ou nos utilitários de pesquisa.

Atualmente, estão em andamento esforços para expandir as capacidades de consulta da base de dados (com pesquisas de estrutura aprimorada), aumentar a cobertura de nutracêuticos ou medicamentos fitoterápicos, incluir plug-ins para facilitar a triagem virtual de drogas e a modelagem farmacológica (absorção, distribuição, excreção, metabolismo e toxicidade).

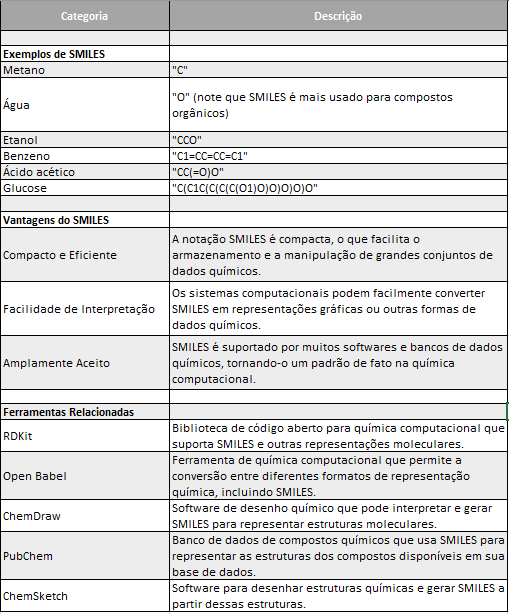
### SMILES (Simplified Molecular-Input Line-Entry System)

Método de notação que permite representar a estrutura química de moléculas utilizando uma linha de texto ASCII (Weininger, D. 1988). Esse sistema é amplamente utilizado em química computacional e bioinformática para a entrada e manipulação de dados químicos.

Aspectos importantes do SMILES:

1. **Representação Linear**: SMILES transforma a representação bidimensional de uma molécula em uma cadeia linear de caracteres que descrevem a composição e a conectividade dos átomos na molécula.
2. **Símbolos de Átomos**: Os átomos são representados pelos seus símbolos químicos, como "C" para carbono, "O" para oxigênio, "N" para nitrogênio, etc.
3. **Ligações**: As ligações entre átomos são representadas implicitamente (ligações simples) ou explicitamente (ligações duplas com "=", triplas com "#", e aromáticas com ":").
4. **Ramificações**: As ramificações na estrutura molecular são representadas por parênteses. Por exemplo, a estrutura isopropil ("-C(CH3)2-") pode ser escrita como "CC(C)C".
5. **Anéis**: Os anéis são representados utilizando números para indicar a posição inicial e final do anel. Por exemplo, o benzeno pode ser escrito como "C1=CC=CC=C1".
6. **Estereoisomeria**: A estereoisomeria (configuração espacial dos átomos) pode ser representada usando símbolos como "@", "@@", "/" e "".

O uso de SMILES facilita a troca de informações químicas e o processamento computacional de dados químicos, sendo uma ferramenta essencial na bioinformática, descoberta de medicamentos e química computacional. A Tabela 1 apresenta exemplos de SMILES, como o metano ("C"), etanol ("CCO") e benzeno ("C1=CC=CC=C1"), além de destacar suas vantagens, como ser uma notação compacta, fácil de interpretar por sistemas computacionais, e amplamente aceita em bancos de dados e softwares de química computacional. Ferramentas como RDKit, Open Babel, ChemDraw, PubChem e ChemSketch suportam SMILES, permitindo a conversão, visualização e análise de moléculas de forma eficiente.

****

**Tabela 1**. - Exemplos, vantagens e ferramentas relacionadas ao SMILES

# Revisão da Literatura e Trabalhos Relacionados

O desenvolvimento de produtos farmacêuticos envolve tempo e um alto custo financeiro, e uma fase crítica nesse processo é a formulação 1. As formulações farmacêuticas envolvem um grande conjunto de fatores, e a abordagem tradicional para desenvolver essas formulações depende de processos incertos baseados em tentativa e erro. Esse processo requer um grande número de experimentos in vitro e in vivo, que consomem muitos recursos e tempo 2. Outro fator chave na indústria farmacêutica é o rápido crescimento dos avanços tecnológicos, incluindo o aumento da quantidade de dados científicos disponíveis, o que resulta em vários desafios computacionais, como aqueles ligados ao armazenamento e análise de dados 3.

Nesse cenário, o AM tem se destacado como uma ferramenta para otimizar e acelerar o desenvolvimento de formulações farmacêuticas. Por exemplo, técnicas de ML podem ser usadas para analisar informações complexas sobre as propriedades físico-químicas de ingredientes ativos e excipientes, bem como dados de estabilidade e eficácia de medicamentos existentes. O ML permitiu várias melhorias no domínio da saúde, por exemplo, no diagnóstico de doenças 4, na previsão de estruturas de proteínas e na identificação de novos medicamentos 2.

A falta de revisões da literatura focadas na relação do uso de ML e suas aplicações, bem como o potencial e rápido desenvolvimento do uso de ML para apoiar a formulação de medicamentos, nos motivou a conduzir esta revisão da literatura. Os principais objetivos e contribuições esperadas da revisão realizada são: (1) identificar as técnicas de ML e sua relação com grupos de aplicações no contexto de formulações farmacêuticas; e (2) destacar os aspectos positivos e limitações dos estudos analisados, apontando lacunas na literatura e desafios de pesquisa atuais e futuros.

## Trabalhos relacionados a revisão da literatura e diferencial desta revisão

Existem várias revisões da literatura sobre a aplicação de ML e Aprendizado Profundo (DL) na saúde, e várias revisões sobre tecnologias e métodos para formulação farmacêutica e descoberta de medicamentos (por exemplo, 5). Esta seção é focada em apresentar e analisar revisões sobre o uso de ML ou DL na formulação de medicamentos e fases relacionadas.

A referência 2 apresenta uma revisão da literatura sobre os métodos e ferramentas aplicados ao desenvolvimento de formulações químicas usando ML. Segundo os autores, a formulação de medicamentos orientada por ML oferece oportunidades para acelerar os esforços de desenvolvimento, descobrir novos materiais e formulações inovadoras, e gerar novos conhecimentos na ciência da formulação de medicamentos. A revisão enfatiza as mais recentes tecnologias de inteligência artificial, incluindo modelos generativos, DL bayesiano, aprendizado por reforço e laboratórios autônomos.

A referência 6 apresenta uma revisão da literatura com o objetivo de analisar o potencial do ML em várias áreas do desenvolvimento biofarmacêutico. O foco deles foi nos desenvolvimentos e aplicações de ML nos desafios relacionados à adoção da Indústria 4.0 na indústria biofarmacêutica.

A referência 7 propõe o uso de algoritmos de IA e ML em conjunto com modelagem molecular, modelagem matemática, simulação de processos e modelagem farmacocinética baseada em fisiologia para o desenvolvimento da técnica de entrega de medicamentos.

Outras revisões focam em diferentes fases e abordagens específicas na formulação de medicamentos. A referência 8, por exemplo, focou no uso de ML generativo para descoberta de medicamentos de novo, ou seja, no design de novas entidades químicas que se encaixam em um conjunto de restrições usando algoritmos computacionais.

A referência 9 revisou o estado da arte sobre o uso de redes neurais na fabricação farmacêutica. Os estudos selecionados foram classificados em análise e melhoria de processos, controle de qualidade e fabricação aditiva.

Nosso estudo difere dos outros ao fornecer uma nova visão focada no uso de técnicas de ML e sua aplicação em três grupos: entrega de medicamentos, desenvolvimento de proteínas e formulações.

## Metodologia e Execução da Revisão

Esta revisão é metodologicamente baseada nas diretrizes apresentadas por Kitchenham e no protocolo PRISMA.

Uma pesquisa exploratória preliminar foi realizada para obter os insumos necessários para esta revisão. Isso resultou na definição dos parâmetros de pesquisa, período de cobertura da pesquisa, bases de dados científicas, palavras-chave a serem usadas e área de busca nos artigos. Utilizamos as seguintes bases de dados na pesquisa exploratória: Google Scholar, IEEE Xplore, Springer Link, ACM DL e PubMed.

A partir dessa pesquisa exploratória, uma Questão Principal (MQ) foi definida: Quais são as abordagens, aplicações, técnicas, limitações e desafios ao usar ML para apoiar a seleção e desenvolvimento de formulações farmacêuticas?

Como focamos em um tópico de pesquisa recente, o período de busca foi definido de 2019 a 2023. As seguintes bases de dados foram escolhidas porque fornecem resultados relevantes e baixa sobreposição na revisão exploratória preliminar: PubMed, Springer Link e IEEE Xplore. A seguinte string de busca foi usada (adaptada à sintaxe de cada base de dados): (“formulação”) AND (“aprendizado de máquina”) AND (“farmacêutica”) AND (“previsão” OR “predição”).

Os critérios de inclusão e exclusão foram definidos por três pesquisadores da área de pesquisa em computação, em um processo iterativo de leitura de artigos e proposição de critérios até alcançar um consenso. Os critérios estão detalhados na Tabela 3.

A pesquisa inicial resultou em um total de 71 artigos, distribuídos da seguinte forma: 2 artigos encontrados no IEEE Xplore, 22 no Springer Link e 47 no PubMed.

A primeira avaliação considerou o título, resumo e palavras-chave. Inicialmente, 65 artigos foram selecionados por estarem relacionados à MQ, e foram posteriormente avaliados completamente contra os critérios. Dois pesquisadores de computação e um pesquisador de biomedicina analisaram os artigos selecionados e criaram a lista final por consenso após discussões.

Durante a análise de elegibilidade (ver Figura 2), 45 estudos foram identificados como não conformes (de acordo com E2) e 3 trabalhos foram identificados como revisões de literatura (de acordo com E6).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Referencia | Aplicação | | | | Principais técnicas de ML |
| M | D | P | F |
| Deng et al. [12] | X | X |  |  | XGB,ANN,DT,PLS DNN,KNN,SVM,RF LGBM,MLR,RN,RR |
| Lou e Hageman [13] | X | X |  |  | AB,RF,DT,ANN  GNB,KNN |
| He et al. [14] | X | X |  |  | LGBM,RF,KNN,PLS DNN,SVM,DT,MLR |
| Gao et al. [15] | X | X |  |  | RF, KNN, AD, NB, SVM, LGBM, XGB |
| Noorain et al. [16] | X | X |  |  | GP |
| Damiati e Damiati [17] | X | X |  |  | ANN |
| Gentiluomo et al. [18] | X |  | X |  | ANN |
| Gentiluomo et al. [19] | X |  | X |  | ANN |
| Lai et al. [20] | X |  | X |  | LR,SVM,KNN,DT |
| Kamerzell e Middaugh [3] | X |  | X |  | KNN,DT,LR,NB |
|  |  |  |  |  | ANN,RF,SVN... |
| Zhao et al. [21] | X |  |  | X | DNN, LGBM, RF |
| Schmitt et al. [22] | X |  |  | X | PLS,R,LRG KR,XGB,ANN |
| Patel et al. [23] | X |  |  | X | ANN |
| Gao et al. [24] | X |  |  | X | LGBM |
| Dong et al. [25] | X |  |  | X | RF, SVM, LGBM, XGB, SVM |
| Glišić et al. [26] | X |  |  | X | AB, NB, RL |
| Yang et al. [1] | X |  |  | X | MLR,PLSR,SVM ANN,KNN,DNN |
| Yoo et al. [27] | X |  |  | X | CNN |

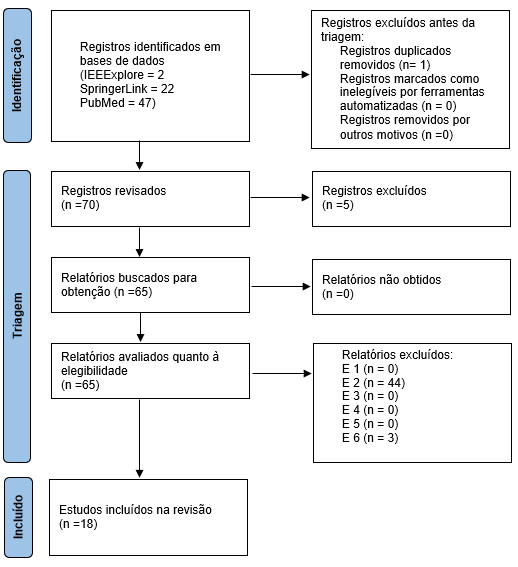
**Tabela 2 -** Resumo das aplicações de ML e técnicas utilizadas nos trabalhos selecionados

**Aplicações:** M: Aprendizado de Máquina; D: Entrega de Medicamentos; P: Desenvolvimento de Proteínas; F:Formulações.

**Técnicas de ML:** AB: AdaBoost; ANN: Rede Neural Artificial; CNN: Rede Neural Convolucional; DNN: Rede Neural Profunda; DT: Árvore de Decisão; GNB: Gaussian Naive Bayes; GP: Processo Gaussiano; KNN: K-Nearest Neighbors; KR: Regressão Kernel Ridge; LGBM: Máquina de Gradiente LightBoost; LRG: Regressão Linear Ridge; LR: Regressão Logística; LRM: Regressão Linear Múltipla; NB: Naive Bayes; PLS: Mínimos Quadrados Parciais; RF: Random Forest; RN: ResNet DNN; RR: Regressão Ridge; SVM: Máquina de Vetores de Suporte; SVR: Regressão de Vetores de Suporte; XGB: eXtreme Gradient Boosting.

|  |  |  |
| --- | --- | --- |
| **Tipo** | **Sigla** | **Critério** |
| **Inclusão** | I1 | Pesquisas sobre Machine Learning , Farmacêutica e Predição. |
| I2 | Estudos que utilizam aprendizado de maquina em formulações farmacêuticas. |
| **Exclusão** | E1 | Artigos escritos em idiomas diferentes do Inglês e do Português. |
| E2 | Artigos que não estejam relacionados com Machine Learning e formulações farmacêuticas. |
| E3 | Artigos que não sejam da área de computação ou multidisciplinar com computação. |
| E4 | Textos que não sejam publicações científicas. |
| E5 | Resumos com menos de 4 páginas e que não tenham profundidade ou resultados relevantes. |
|  | E6 | Revisões sistemáticas e Livros. |

**Tabela 3 –** Critérios de inclusão e exclusão de artigos



**Figura 2**-Diagrama Prisma com resultados quantitativos

## Análise e Discussão dos Resultados

Apresentamos uma análise dos 18 estudos selecionados. A Tabela 2 apresenta uma visão geral da classificação dos artigos de acordo com suas técnicas e aplicações. As siglas referentes às técnicas de ML, usadas no restante deste artigo, também são definidas na Tabela 2. Esta seção está estruturada da seguinte forma: A Seção 3.3.1 apresenta soluções baseadas em ML para otimizar o desenvolvimento de Drug Delivery; A Seção 3.3.2 apresenta soluções baseadas em ML para o desenvolvimento de proteínas; e a Seção 3.3.3 apresenta o restante dos artigos, que usam o ML como ferramenta para otimizar a formulação farmacêutica.

### Soluções baseadas em aprendizado de máquina para otimizar o desenvolvimento de entrega de medicamentos

Após analisar o texto completo dos 18 artigos incluídos nesta revisão, 6 foram selecionados para serem apresentados nesta subseção. A referência [12] apresentou um modelo de previsão para acelerar o desenvolvimento de produtos de microesferas para medicamentos de moléculas pequenas por meio de técnicas de ML. Para tanto, foram utilizadas as seguintes técnicas de ML: XGB, ANN, DT, PLS, DNN, KNN, SVM, RF, LGBM, MLR, RN e RR.

A referência [13] usou as técnicas AB, RF, DT, ANN, GNB e KNN para prever a biodisponibilidade após administração subcutânea de anticorpos monoclonais, mesmo sem entender completamente o mecanismo e a causalidade entre entradas e saídas.

A referência [14] apresentou o uso de técnicas LGBM, RF, KNN, PLS, DNN, SVM, DT e MLR para o desenvolvimento de formulações de nanocristais substituindo processos de tentativa e erro que consomem tempo e recursos.

A referência [15] propôs o uso de uma metodologia computacional integrada com o objetivo de reduzir o trabalho tradicional de design de formulações de medicamentos e trazer novas ideias para futuros projetos de formulação, modelagem molecular e abordagens experimentais para o design racional de formulações SEDDS (Self-Emulsifying Drug Delivery Systems). Para tanto, foram utilizadas as seguintes técnicas de ML: RF, KNN, DT, NB, SVM, LGBM e XGB.

A referência [16] abordou a combinação de nano terapêutica e técnica GP ML para simplificar os sistemas de desenvolvimento de medicamentos antivirais automatizando a análise.

Por fim, a referência [17] utilizou a combinação das técnicas microfluídica e LR, SVM, KNN e DT, juntamente com o uso de biomateriais, a fim de gerar micropartículas poliméricas carregadas com medicamentos.

### Soluções baseadas em aprendizado de máquina para desenvolvimento de proteínas

Após a leitura de todos os artigos na íntegra, 4 trabalhos foram selecionados para serem apresentados nesta subseção. O rápido crescimento dos avanços tecnológicos e a quantidade de dados científicos na última década levaram a vários desafios, incluindo armazenamento e análise de dados. Isso se reflete no estudo e desenvolvimento de modelos precisos de conjuntos de dados complexos de proteínas.

A referência [3] utilizou algoritmos ML populares (por exemplo, KNN, DT, LR, NB, ANN, RF, SVN) focados no desenvolvimento de proteínas farmacêuticas. Os autores afirmam que os modelos ML aplicados podem ser usados para entender a viscosidade não linear dependente da concentração de soluções proteicas, prever taxas de oxidação e desamidação de proteínas, classificar partículas subvisíveis e comparar a estabilidade física.

A referência [18] abordou o desenvolvimento de plataformas para a formulação de proteínas usando uma abordagem baseada em RNA com o objetivo de diminuir os custos financeiros e também facilitar o desenvolvimento de medicamentos à base de proteínas.

A referência [19] aplicou RNAs para prever a estabilidade a longo prazo em condições reais de armazenamento a partir de estudos de estabilidade acelerada e outras propriedades biofísicas de alto rendimento, por exemplo, a primeira temperatura aparente de desdobramento.

Finalmente, a referência [20] usou modelos preditivos baseados em LR, SVM, KNN e DT para prever taxas de agregação de anticorpos terapêuticos e viscosidade em altas concentrações no desenvolvimento de proteínas terapêuticas.

### Soluções baseadas em aprendizado de máquina com uma variedade de aplicações

Entre os 18 estudos avaliados, 8 trabalhos foram selecionados para serem apresentados nesta subseção.

A referência [21] usou técnicas de DNN, LGBM e RF para construir modelos preditivos para simplificar e reduzir o número de experimentos no desenvolvimento de formulações farmacêuticas para sistemas de medicamentos com ciclodextrina.

A referência [22] usou modelos preditivos baseados em PLS, RR, LRG, KR, XGB e ANN para obter estimativas iniciais para valores de parâmetros de processo que melhor atingem um tamanho de partícula alvo para uma determinada formulação.

A referência [23] desenvolveu uma ferramenta preditiva para identificar incompatibilidades de excipientes de medicamentos usando modelos de RNAs. Referência [24] integrou ferramentas, incluindo modelagem molecular e métodos de ML, com base em um modelo LGBM, para desenvolver uma nova formulação ternária de Andrographolide-Ciclodextrinas-Tocoferol Polietileno Glicol Succinato (AG-CD-TPGS), obtendo melhorias na solubilidade aquosa, taxa de dissolução e biodisponibilidade.

A referência [25] apresentou uma plataforma de previsão de formulação de dispersão sólida usando técnicas de RF, SVM, LGBM, XGB e SVM. O processamento de sistemas líquido-sólido (LSS), uma abordagem para melhorar a biodisponibilidade oral de medicamentos pouco solúveis, tem se mostrado desafiador devido à quantidade relativamente alta de fase líquida incorporada a eles. Referência [26] utilizou ferramentas baseadas nas técnicas AB, NB e LR, para melhor compreender os efeitos dos fatores de formulação e dos parâmetros do processo de compressão nas propriedades de fluidez e compressão do LSS. A referência [1] usou modelos MLR, PLSR, SVM, ANN, KNN e DNN para prever formulações farmacêuticas. Os resultados mostram que as precisões das redes de DL foram maiores do que outras. O complexo fármaco-fosfolipídio é uma tecnologia de fórmula promissora para melhorar a baixa biodisponibilidade de ingredientes farmacêuticos ativos. No entanto, identificar o fosfolipídio fosfórico e o candidato a medicamento por meio de testes in vitro pode ser caro e demorado devido às propriedades físico-químicas e ao ambiente experimental. Nesse contexto, a referência [27] usou modelos de CNN para prever a formação do complexo fármaco-fosfolipídio.

### Discussão sobre a Revisão da Literatura

De acordo com a referência[14], o uso de ML na criação e seleção de formulações farmacêuticas tem o potencial de impulsionar a indústria, pois tem a capacidade de acelerar um extenso processo manual de tentativa e erro, reduzindo custos e tempo. Neste artigo, apresentamos uma revisão da literatura sobre o uso de ML na criação e seleção de formulações farmacêuticas. Este estudo identifica e analisa tendências, por exemplo, o uso de redes neurais treinadas com uma grande quantidade de dados. De um total de 71 artigos, 18 trabalhos foram criteriosamente selecionados, classificados e sintetizados a fim de analisar as abordagens do estado da arte, com o objetivo de avaliar o uso de ferramentas computacionais de ML e sua aplicabilidade no campo do desenvolvimento de medicamentos.

Os trabalhos analisados mostram um cenário muito promissor para a aplicação de algoritmos de ML, pois o volume de dados científicos [3] está crescendo e tem possibilitado a criação de modelos preditivos mais confiáveis nos últimos anos. Em consonância com [28], destacamos que os estudos devem se concentrar mais na qualidade das decisões sobre qual composto levar adiante, incluindo a eficácia e segurança dos medicamentos. No que diz respeito às técnicas de ML adotadas, destacamos que a maioria dos trabalhos utilizou técnicas baseadas em redes neurais (profundas ou não), estando presente em 12 dos 18 trabalhos analisados. Esta revisão também destaca que vários trabalhos experimentaram muitas técnicas, arquiteturas e parâmetros diferentes, revelando que não há uma solução dominante ou previamente estabelecida para todos os casos. As pesquisas analisadas também apontam que os modelos preditivos devem ser utilizados de forma criteriosa e complementar à experimentação oratória laboratorial, pois as validações experimentais são essenciais para confirmar as previsões do modelo, bem como para garantir a qualidade das formulações. A interação de ML e cientistas farmacêuticos é essencial para explorar plenamente o potencial dessa abordagem e garantir o desenvolvimento responsável de medicamentos, considerando questões éticas e de segurança relacionadas à privacidade de dados e ao uso responsável da tecnologia. Ao mesmo tempo em que os artigos analisados apontam para o uso de diversas técnicas para diferentes aplicações com bons resultados, destacamos também desafios relacionados à necessidade de criação de bancos de dados integrados. O uso de novas técnicas generativas e de DL também está na agenda de pesquisas futuras.

## Trabalhos Relacionados e Diferencial da Pesquisa

A pesquisa sobre o uso de Machine Learning (ML) na entrega de medicamentos (drug delivery) tem mostrado avanços significativos, com várias revisões da literatura abordando diferentes aspectos dessa aplicação. No entanto, ainda há uma lacuna na literatura existente no que diz respeito a uma visão abrangente que integre as diversas técnicas de ML e suas aplicações específicas na entrega de medicamentos.

## 3.4.1 Trabalhos Relacionados

Revisões Gerais sobre ML na Saúde: Diversos estudos revisaram o uso de ML e Aprendizado Profundo (DL) na saúde, destacando suas aplicações em diagnóstico, previsão de estruturas de proteínas e descoberta de novos medicamentos. Essas revisões fornecem uma base sólida sobre as capacidades gerais de ML, mas não se aprofundam nas especificidades da entrega de medicamentos.

Aplicações de ML na Formulação Química: A revisão apresentada na referência 2 foca nos métodos e ferramentas de ML aplicados ao desenvolvimento de formulações químicas. Os autores destacam como o ML pode acelerar o desenvolvimento de novos materiais e formulações inovadoras, utilizando tecnologias como modelos generativos e aprendizado por reforço.

Desafios da Indústria 4.0 na Biofarmacêutica: A referência 6 analisa o potencial do ML em enfrentar os desafios da Indústria 4.0 na biofarmacêutica, abordando desenvolvimentos e aplicações de ML em várias áreas do desenvolvimento biofarmacêutico.

Modelagem e Simulação na Entrega de Medicamentos: A referência 7 propõe a integração de algoritmos de IA e ML com modelagem molecular, matemática e farmacocinética para aprimorar as técnicas de entrega de medicamentos. Essa abordagem permite otimizar a liberação controlada de fármacos, melhorando a eficácia terapêutica e reduzindo efeitos colaterais.

Descoberta de Medicamentos de Novo: A referência 8 foca no uso de ML generativo para o design de novas entidades químicas, destacando como algoritmos computacionais podem criar novas moléculas que atendem a critérios específicos, o que é crucial para a entrega de medicamentos inovadores.

Redes Neurais na Fabricação Farmacêutica: A revisão na referência 9 explora o uso de redes neurais na fabricação farmacêutica, classificando os estudos em análise e melhoria de processos, o que inclui a otimização de sistemas de entrega de medicamentos.

## 3.4.2 Diferencial da Pesquisa

O diferencial desta pesquisa reside na sua abordagem integrada e abrangente, que visa:

* Mapear as Técnicas de ML: Identificar e categorizar as diversas técnicas de ML utilizadas na entrega de medicamentos, proporcionando uma visão clara das metodologias mais promissoras e suas aplicações específicas.
* Analisar Limitações e Desafios: Destacar as limitações dos estudos existentes e identificar lacunas na literatura, oferecendo uma base para futuras pesquisas.
* Propor Direções Futuras: Sugerir direções futuras para a pesquisa, considerando os avanços tecnológicos e as necessidades emergentes na entrega de medicamentos.

Ao fornecer uma análise detalhada e integrada das técnicas de ML aplicadas à entrega de medicamentos, esta pesquisa contribui para um entendimento mais profundo e abrangente do campo, facilitando o desenvolvimento de soluções inovadoras e eficientes.

# O Uso de Aprendizado de Máquina no Apoio à Predição de Distribuição de Fármacos

O uso de aprendizado de máquina tem se mostrado uma ferramenta valiosa na predição de características farmacológicas, incluindo a distribuição de fármacos no organismo. Este trabalho se concentra em explorar como técnicas de aprendizado de máquina podem ser aplicadas para melhorar a precisão e a eficiência dessas predições. A primeira etapa envolve a construção de um conjunto de dados (*dataset*) e seleção de atributos, onde são escolhidas bases de dados apropriadas contendo informações moleculares, como SMILES, e atributos relevantes para a predição, como absorção e outras propriedades farmacocinéticas. Em seguida, realiza-se o pré-processamento e codificação dos dados, abordando o tratamento de dados ausentes, normalização e a conversão das representações moleculares em vetores numéricos adequados para os algoritmos de aprendizado de máquina.

Na fase de modelos e implementação, foram utilizados modelos de regressão e redes neurais implementados com bibliotecas Python, acompanhados de ajustes de hiper parâmetros e validação dos modelos para garantir seu desempenho. Por fim, o estudo inclui uma descrição formal do problema, caracterizando-o como um problema de regressão, com a aplicação de técnicas de validação cruzada e a utilização de métricas como MSE, RMSE e R-quadrado para avaliar predições das variáveis contínuas e medidas de acuraria, f1-score e AUC para avaliar predições de variáveis discretas, além de métodos de interpretação para análise dos resultados.

## Base de Dados e Seleção de Atributos

O sucesso de uma análise no campo farmacêutico não depende apenas da escolha de bases de dados confiáveis, mas também de como os atributos extraídos dessas bases são utilizados em modelos de ML. A combinação de dados de qualidade e algoritmos avançados permite construir modelos preditivos robustos, capazes de identificar padrões relevantes para a descoberta e análise de moléculas que possuem as melhores características para DrugDelivery.

Neste estudo, a seleção criteriosa da base de dados foi essencial para fornecer subsídios aos algoritmos de ML, como Support Vector Machines (SVM), Random Forest, Gradient Boosting e Redes Neurais Artificiais. A seguir, apresenta-se a base de dados selecionada para este estudo, bem como os critérios que justificam sua escolha.

### Base de Dados

Neste estudo, utilizou-se o DrugBank como base de dados, contendo um total de 11.583 moléculas analisadas. Essa escolha se deu porque comparada a outras bases de dados, como PubChem, ChEMBL, ChemSpider e ZINC, o DrugBank se destaca quando a pesquisa envolve informações detalhadas sobre medicamentos, mecanismos de ação e interações farmacológicas. Podendo essas outras serem utilizadas em etapas futuras desta pesquisa.

O DrugBank se destaca por fornecer informações detalhadas sobre medicamentos, abrangendo não apenas suas estruturas químicas, mas também mecanismos de ação, interações medicamentosas, propriedades farmacocinéticas e indicações clínicas. Isso o torna essencial para pesquisas que envolvem medicamentos aprovados ou com aplicações terapêuticas. Já as bases como PubChem e ChemSpider são focadas em propriedades químicas úteis em um aprofundamento da pesquisa do ponto de vista da química. O DrugBank inclui dados clínicos detalhados, como efeitos colaterais e ensaios clínicos, sendo vital para estudos que buscam compreender a segurança e eficácia dos medicamentos. Além disso oferece uma compreensão profunda sobre as interações entre medicamentos e alvos moleculares. Já bases como o ChEMBL, que se concentram em bioensaios e atividades biológicas. Suas informações sobre interações medicamentosas são fundamentais para estudos de combinações de fármacos, enquanto os dados farmacocinéticos ajudam a entender como o corpo processa os medicamentos. Esse nível de detalhe, aliado às referências bibliográficas robustas, torna o DrugBank uma ferramenta indispensável para pesquisas farmacêuticas focadas em aplicações clínicas.

### Seleção de Atributos

No intuito de prever moléculas que possuem a melhor biodisponibilidade, absorção e propriedades de distribuição foram elencados os seguintes atributos: Bioavailability, LogP, CYP450 2C9, CYP450 2D6. Esta seleção foi realizada com auxílio de uma especialista no domínio, uma biomédica com doutorado em química computacional e experiência em simulações biomoleculares.

A Biodisponibilidade (Bioavailability) e o coeficiente de partição octanol-água (LogP), são parâmetros que auxiliam na compreensão da capacidade do composto de atravessar membranas celulares e ser eficazmente absorvido (Pereira, D. G. 2007). Além disso, as interações com as enzimas do citocromo P450, especialmente as isoformas CYP450 2C9 e CYP450 2D6 (Smith G. et al. 1998), são essenciais, pois essas enzimas estão entre as principais responsáveis pelo metabolismo de drogas, determinando a velocidade de biotransformação dos compostos e, consequentemente, sua permanência e distribuição no organismo. Portanto, ao eleger esses atributos para análise, busca-se um conjunto de parâmetros que possa fornecer previsões consistentes e confiáveis sobre o desempenho das moléculas no organismo humano, contribuindo para a seleção de compostos promissores no desenvolvimento de novos fármacos. Cada atributo selecionado atua como uma referência central para o treinamento e teste de modelos de ML e DL, contribuindo com informações fundamentais para a previsão de propriedades moleculares. A origem dos dados está no atributo SMILES, uma notação padrão e amplamente utilizada para representar a estrutura de moléculas em química computacional, essencial para transformar as características moleculares em uma forma legível e interpretável pelos modelos.

## Pré-processamento e Codificação

Para o desenvolvimento de modelos preditivos confiáveis, o pré-processamento dos dados é uma etapa crucial, especialmente em aplicações como a análise de propriedades moleculares. Diversas técnicas de pré-processamento foram utilizadas para extrair atributos, tratar dados faltantes, valores ambíguos, *outliers* e padronização dos dados, contribuindo para uma maior robustez do modelo e para resultados mais consistentes.

### Extração de Atributos

O formato XML (eXtensible Markup Language) foi utilizado para a extração inicial de dados da base empregada neste projeto. No entanto, para acessar atributos específicos e seus respectivos valores, foi necessária a conversão do arquivo para o formato JSON (JavaScript Object Notation), permitindo maior flexibilidade na manipulação dos dados. A partir dessa estrutura, uma planilha no formato CSV (Comma-Separated Values) foi gerada, através de algoritmos na linguagem Python, facilitando o processamento e a análise dos dados em etapas subsequentes. Essa transformação possibilitou o uso de ferramentas analíticas adicionais e a integração com modelos preditivos de forma eficiente[[1]](#footnote-1).

### Análise Exploratória e Tratamento de Outliers

A análise inicial dos dados, como a visualização das primeiras linhas e das estatísticas descritivas, fornece uma visão geral das variáveis e possibilita a identificação de valores inconsistentes ou extremos. Para reduzir o impacto de outliers, foi aplicada a técnica do **Intervalo Interquartil (IQR)** às variáveis alvo (Freitas, 2019). Os valores que estavam fora dos limites definidos pelo IQR (valores menores que: Q1 - 1.5\*IQR e maiores que: Q3 + 1.5\*IQR) foram removidos, reduzindo o viés que esses outliers poderiam introduzir no modelo de regressão.

### Padronização dos Dados

Para tratar potenciais diferenças de escala entre as variáveis, foi aplicado o **RobustScaler** aos dados de entrada, uma técnica robusta contra outliers. Esta técnica transforma os dados de maneira a torná-los mais homogêneos, calculando os parâmetros de escalamento baseados na mediana e no intervalo interquartil em vez da média e desvio padrão. Essa abordagem ajuda o modelo a convergir mais rapidamente e evita que variáveis com escalas maiores dominem a análise.

### Codificação de moléculas para análise quantitativa

A técnica de codificação de moléculas para análise quantitativa transforma sequências SMILES (Simplified Molecular Input Line Entry System) em vetores numéricos, permitindo a representação das propriedades químicas para modelos de aprendizado de máquina. Através deste procedimento fingerprints (FPs) moleculares capturam padrões estruturais através de representações vetoriais binárias.

Os passos fundamentais para codificação de moléculas são os seguintes:

* Transformação SMILES para objetos químicos: SMILES são convertidos em objetos Mol, um formato que facilita a manipulação de informações químicas computacionais.
* Geração de Morgan Fingerprints: Utilizando uma função especifica do modulo RDKit da linguagem Python (AllChem.GetMorganGenerator) o código gera vetores que representam subestruturas ao redor de cada átomo. Cada bit do vetor binário representa a presença ou ausência de subestruturas específicas, crucial para modelagem química.
* Formato de Saída: Esses vetores binários são unidos a atributos alvo (Bioavailability, LogP, CYP450 2C9 ou CYP450 2D6), criando uma matriz de atributos que relaciona estrutura química e propriedades, permitindo análises exploratórias e preditivas com métodos de ML, como redes neurais (MLP) e regressão linear.

Essa técnica facilita a interpretação química e a generalização dos dados em modelos preditivos, como mencionado em métodos como redes neurais e validação cruzada, resultando em uma abordagem robusta para prever propriedades moleculares com base em representações numéricas confiáveis e consistentes.

## Modelos e Implementação

- Aqui entra a parte mais técnica como bibliotecas Python utilizadas, modelos reaproveitados ou construídos, etc

- Entra também como foi realizada a validação e ajustes de hiperparâmetros

- As camadas das redes neuras, por exemplo, devem ser descritas aqui

## Descrição formal do problema, execução e métricas de avaliação

- Descrever formalmente o problema como um problema de aprendizado de máquina (de regressão neste caso específico)

- Aqui entra as técnicas de validação cruzada, split e testes

- Descrever as métricas utilizadas (ex: MSE, RMSE, MAE, MAPE, R-squared ...), e justificar o uso delas

- Descrever métodos de intepretação de modelos (caso use)

# Análise dos Resultados e Discussão

- Colocar preâmbulo

## Resultados

- Descrever resultados (tabelas com métricas comparativas entre os resultados)

- Colocar interpretação de modelos (caso use)

## Discussão

- Discutir implicações dos resultados

- Discutir aplicabilidade na prática

- Comentar perante resultados de outros estudos/artigos prévios

- Falar sobre possíveis vieses e “ameaças a validade”

- Falar sobre limitações

- Falar sobre desafios futuros (não necessariamente sobre a sua pesquisa, mas da área como um todo)

# Conclusão

- Revisitar a questão de pesquisa e destacar como ela foi respondida

## Contribuições

- Destacar as contribuições da sua pesquisa (prática e teórica)

- Destacar publicações realizadas e a ser submetidas

## Trabalhos Futuros

- Listar os próximos passos da sua pesquisa

## Considerações Finais

- Breve fechamento e reafirmação

# Cronograma de Atividades

- Listar cada atividade que vai desenvolver (ente a qualificação e a defesa), e descrever

- Colocar em um cronograma com datas de execução destas atividades até a defesa da dissertação (diagrama de gantt vem aqui ...), ou seja é o que você vai fazer entre a qualificação e a defesa

##### Referências

Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. **DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006** Jan 1;34(Database issue):D668-72. doi: 10.1093/nar/gkj067. PMID: 16381955; PMCID: PMC1347430.

Shargel, Leon; ANDREW, B. C.; WU-PONG, Susanna. **Applied biopharmaceutics & pharmacokinetics.** Stamford: Appleton & Lange, 1999.

Pattni, B.S., Torchilin, V.P. (2015). **Targeted Drug Delivery Systems: Strategies and Challenges**. In: Devarajan, P., Jain, S. (eds) **Targeted Drug Delivery : Concepts and Design. Advances in Delivery Science and Technology. Springer**, Cham. doi: 10.1007/978-3-319-11355-5\_1

Tran TTV, Tayara H, Chong KT. **Recent Studies of Artificial Intelligence on In Silico Drug Distribution Prediction.** Int J Mol Sci. 2023 Jan 17;24(3):1815. doi: 10.3390/ijms24031815. PMID: 36768139; PMCID: PMC9915725.

Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to Machine Learning, Neural Networks, and Deep Learning. Transl Vis Sci Technol. 2020 Feb 27;9(2):14. doi: 10.1167/tvst.9.2.14. PMID: 32704420; PMCID: PMC7347027.

GUIDO, R. V. C.; ANDRICOPULO, A. D.; OLIVA, G. Planejamento de

fármacos, biotecnologia e química medicinal: aplicações em doenças infecciosas.

Revista Estudos Avançados. São Paulo, v.24, n.70, p.81-98, 2010.

Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH. PubChem Substance and Compound databases. Nucleic Acids Res. 2016 Jan 4;44(D1):D1202-13. doi: 10.1093/nar/gkv951. Epub 2015 Sep 22. PMID: 26400175; PMCID: PMC4702940.

Gaulton, Anna et al. “The ChEMBL database in 2017.” Nucleic acids research vol. 45,D1 (2017): D945-D954. doi:10.1093/nar/gkw1074

Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018 Jan 4;46(D1):D1074-D1082. doi: 10.1093/nar/gkx1037. PMID: 29126136; PMCID: PMC5753335.

Pence, H. E., & Williams, A. (2010). ChemSpider: An online chemical information resource. Journal of Chemical Education, 87(11), 1123-1124.

Irwin, J. J., & Shoichet, B. K. (2005). ZINC--a free database of commercially available compounds for virtual screening. Journal of chemical information and modeling, 45(1), 177–182. <https://doi.org/10.1021/ci049714>+

Wishart, D. S. (2008). DrugBank and its Relevance to Pharmacogenomics. Pharmacogenomics, 9(8),1155–1162. <https://doi.org/10.2217/14622416.9.8.1155>

WEININGER, David. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences, v. 28, n. 1, p. 31-36, 1988. DOI: 10.1021/ci00057a005.

Franke, F. S. (2020). Como são desenvolvidos os medicamentos? farmacoLÓGICA. Universidade Federal do Rio Grande do Sul. Disponível em: <https://www.ufrgs.br/farmacologica/2020/06/30/como-sao-desenvolvidos-os-medicamentos/>

Pereira, D. G.. (2007). Importância do metabolismo no planejamento de fármacos. Química Nova, 30(1), 171–177. https://doi.org/10.1590/S0100-40422007000100029

Rimal, R., Rimal, B., Bhandari, H.N. et al. Real Estate Market Prediction Using Deep Learning Models. Ann. Data. Sci. (2024). <https://doi.org/10.1007/s40745-024-00543-2>

Ferneda E. Redes neurais e sua aplicação em sistemas de recuperação de informação. Ci Inf [Internet]. 2006 Jan;35(1):25–30. Available from: <https://doi.org/10.1590/S0100-19652006000100003>

Silva, Rodrigo Emerson Valentim da. Um estudo comparativo entre redes neurais convolucionais para a classificação de imagens / Rodrigo Emerson Valentim da Silva. – 2018.

Pereira, D. G. (2007). Importância do metabolismo no planejamento de fármacos. Química Nova, 30(1), 171–177. <https://doi.org/10.1590/S0100-40422007000100029>

Smith G, Stubbins MJ, Harries LW, Wolf CR. Molecular genetics of the human cytochrome P450 monooxygenase superfamily. Xenobiotica. 1998 Dec;28(12):1129-65. doi: 10.1080/004982598238868. PMID: 9890157.

Freitas, Igor Wescley Silva de. Um estudo comparativo de técnicas de detecção de outliers no contexto de classificação de dados / Igor Wescley Silva de Freitas. - 2019.

1. O repositório com os arquivos pode ser acessado em: <https://github.com/helderpestanagit/DrugDeliveryPredictor> (Acesso em: 20 nov. 2024). [↑](#footnote-ref-1)