UNIVERSIDADE ESTADUAL DE SANTA CRUZ
PRO-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL EM CIÊNCIA E TECNOLOGIA

LÉO RODRIGUES BISCASSI

METODOLOGIA PARA DESENVOLVIMENTO DE MODELOS REPRODUTÍVEIS PARA PREDIÇÃO DE PROPRIEDADES ADMET UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

PPGMC – UESC

ILHÉUS-BA
2017

LÉO RODRIGUES BISCASSI

METODOLOGIA PARA DESENVOLVIMENTO DE MODELOS REPRODUTÍVEIS PARA PREDIÇÃO DE PROPRIEDADES ADMET UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

PPGMC – UESC

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional em Ciência e Tecnologia da Universidade Estadual de Santa Cruz, como parte das exigências para obtenção do título de Mestre em Modelagem Computacional em Ciência e Tecnologia.

Orientador: Prof. Dr. Paulo Eduardo Ambrósio

Coorientador: Prof. Dr. Rodrigo Antônio Faccioli

ILHÉUS-BA
2017

B621 Biscassi, Léo Rodrigues.
Metodologia para o desenvolvimento de modelos reprodutíveis para predicação de propriedades ADMET utilizando algoritmos de aprendizado de máquinas / Léo Rodrigues Biscassi – Ilhéus, BA : UESC, 2017.
59 f.: il.

Orientador: Paulo Eduardo Ambrósio
Dissertação (Mestrado) – Universidade Estadual de Santa Cruz.Programa de Pós-Graduação em Modelagem Computacional em Ciência e Tecnologia.
Inclui referências.

1. Aprendizado do computador. 2. Big data. 3. Farmacologia. 4. Medicamentos – desenvolvimento. I. Título.

CDD 006.31

LÉO RODRIGUES BISCASSI

METODOLOGIA PARA DESENVOLVIMENTO DE MODELOS REPRODUTÍVEIS PARA PREDIÇÃO DE PROPRIEDADES ADMET UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

PPGMC – UESC

Ilhéus-BA, 20/02/2017

Comissão Examinadora

Prof. Dr. Paulo Eduardo Ambrósio
UESC
(Orientador)

Profa. Dra. Marta Magda Dornelles
UESC

Profa. Dra. Rafaela Salgado Ferreira
UFMG

À minha família, pelo apoio e dedicação.

Agradecimentos

- Agradeço a toda minha família pelo apoio nessa caminhada. Em especial à meus pais, José Carlos e Neide, que desde o princípio me apoiaram e me deram força para continuar. Aos meus avós, João e Maria, e aos meus padrinhos, Odete e José, pelas palavras de apoio durante esta caminhada.
- Aos meus orientadores, prof. Dr. Paulo Eduardo Ambrósio e prof. Dr. Rodrigo Antônio Faccioli. Paulo, sem o seu acolhimento e apoio desde o primeiro dia que cheguei em Ilhéus, nada disso teria sido possível. Obrigado por sua amizade e orientação, mas acima de tudo, pela confiança e pela oportunidade de ter trabalhado contigo nesses últimos dois anos. Rodrigo, sem o seu apoio e incetivo na reta final da graduação eu não teria vivido nenhuma dessas experiências nos últimos dois anos. Obrigado por sua amizade, conselhos e orientação durante essa caminhada.
- Aos amigos Anderson, Gabriel Ganem, Gabriel Mello, Leandro, Lucas Moura e Marlesson, vocês foram a minha família em minha estadia na Bahia. Muito obrigado, pelo apoio e incentivo, muitas vezes regado a cerveja, que vocês me deram nesses últimos dois anos.
- Aos meus amigos de infância Iago, Marco Antônio, Olegário, Ricardo e Tarsio, que estão comigo desde sempre e me apoiaram desde o primeiro momento. Gostaria de agradecer especialmente aos meus grandes amigos(as) Daniele Freitas, Gabriel Pontin, Luan Celso, não só pelo apoio, mas também por seus valiosos ensinamentos e companheirismo, sem vocês eu não teria conseguido superar alguns desafios nessa jornada.
- Enfim, agradeço a todos que contribuíram na elaboração dessa dissertação, mas por um esquecimento não constam seus nomes, apesar disso sua contribuição foi de igual importância, além dos meus agradecimentos, fica registrado minhas sinceras desculpas.

"A persistência é o caminho do êxito" (Charles Chaplin)

Metodologia para desenvolvimento de modelos reprodutíveis para predição de propriedades ADMET utilizando algoritmos de aprendizado de máquina

PPGMC – UESC

Resumo

Nos últimos anos, a utilização de modelos computacionais em projetos de descoberta de fármacos é evidente. Esses modelos têm como principal função a seleção de moléculas promissoras e a investigação de propriedades relacionadas à Absorção, Distribuição, Metabolismo, Excreção e Toxicidade (ADMET) na tentativa de otimizar o processo como um todo, consequentemente reduzindo custos e a quantidade de experimentos em animais. Porém, muitas vezes esses modelos são limitados e não estão aptos a lidar com Big Data, além de possuírem problemas com reprodutibilidade. Este trabalho propõe uma metodologia para o desenvolvimento de modelos reprodutíveis para a predição de propriedades ADMET. O ambiente computacional, com todas as ferramentas e dados necessários para a execução da metodologia, é disponibilizado através de uma imagem de container do Docker. O Jupyter Notebook é utilizado como editor para a publicação das etapas da metodologia, que podem ser executadas de modo interativo, facilitando a reprodução da mesma. A demonstração da metodologia é realizada através de um estudo de caso que visa desenvolver modelos preditivos para as enzimas CYP450, principal família de enzimas envolvida no metabolismo de medicamentos no corpo humano.

Palavras-chave: Reprodutibilidade, Aprendizado de Máquina, Big Data, ADMET

Methodology for the development of reproducible models to predict ADMET properties using machine learning algorithms.

PPGMC - UESC

Abstract

In the last few years, the use of computational modelos in drug discovery projects is evident. The main function of these models is the selection of promising molecules and optimization of Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET) properties in an attempt to reducing costs and quantity of animal experiments. However, often these models are limited and unable to handle with Big Data, besides have reproducibility problems. This work proposes a methodology for the development of reproducible models for the ADMET properties prediction. The computational environment, with all the necessary tools and data, is available through an Docker image. Jupyter Notebook is used as an editor for publication of methodology steps, which can be performed in an interactive way, facilitating the reproduction. The demonstration of methodology is carried out through a case study that aims develop predictive models of CYP450 enzymes.

Keywords: Reproducibility, Machine Learning, Big Data, ADMET

Lista de figuras

Lista de quadros

Lista de abreviaturas e siglas

FDA	*Food and Drug Administration*
NME	*New Molecular Entities*
BLA	*Biologic License Applications*
P&D	Pesquisa & Desenvolvimento
VS	*Virtual Screening*
ADMET	Absorção, Distribuição, Metabolismo, Excreção e Toxicidade
ADME	Absorção, Distribuição, Metabolismo e Excreção
EC	*Effective Concentration*
ED	*Effective Dose*
LD	*Lethal Dose*
TD	*Toxic Dose*
IT	*Índice Terapêutico*
SLC	*Solute Linked Carrier*
OAT	*Organic Anion Transporter*
OCT	*Organic Cation Transporter*
SNC	Sistema Nervoso Central
RDD	*Resilient Distributed Dataset*
ALS	*Alternating Least Square*
ISO	*Internacional Organization for Standardization*
NITE	*National Institute of Technology and Evaluation*
US-FDA	*United States Food and Drug Administration*
TP	*True Positive*
TN	*True Negatives*
FP	*False Positive*

FN	*False Negative*
ROC	*Receiver Operating Characteristic*
TPR	*True Positive Rate*
FPR	*False Positive Rate*
AUC	*Area Under Curve*

Sumário

1 Introdução

O processo para identificar novas moléculas para serem avaliadas como candidatas a novos fármacos é chamado planejamento de fármacos. Ele faz parte do ciclo de desenvolvimento para descoberta de novos medicamentos. De acordo com DiMasi (2002) este ciclo é composto por sete fases antes da comercialização, como pode ser observado na Figura 1.

As fases iniciais, pré-desenvolvimento e desenvolvimento, tem como principal fim a identificação de compostos químicos (ligantes) para o desenvolvimento do medicamento. Na fase pré-clínica se inicia o desenvolvimento do novo medicamento com as moléculas selecionadas nas fases anteriores. Na fase de pesquisas clínicas o objetivo é aperfeiçoar os compostos para que seja possível definir o novo fármaco de forma segura e eficiente para que este possa ser utilizado pela população. A última fase é a de aprovação dos órgãos regulamentadores para o início da comercialização do medicamento.

Figura 1 – Fases no ciclo de descoberta de novos fármacos

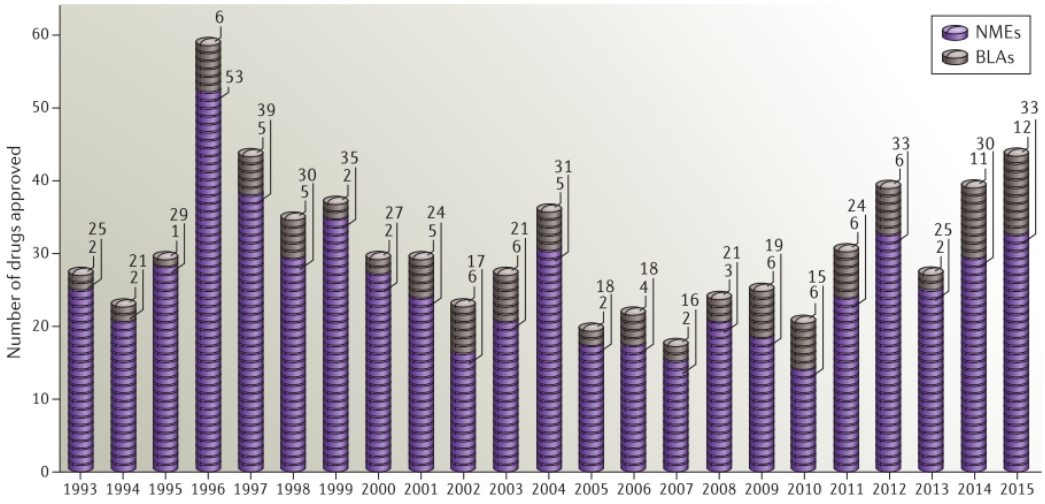
Fonte: Imagem desenvolvida pelo autor.

Estima-se que o custo total de um medicamento, do desenvolvimento à comercialização, gira em torno de 2.6 bilhões de dólares, e 300 milhões de dólares, em média, com estudos pós-clínicos que são realizados após o início das vendas (MULLARD, 2016). Além disso, são necessários, em média, de 10 a 12 anos para um novo medicamento chegar ao mercado (MATHEWS, 2015).

O número de medicamentos aprovados pela (*Food and Drug Administration*) em 2015 aumentou consideravelmente se comparado aos últimos anos, como pode ser observado na Figura 2 (MULLARD, 2016), onde a sigla NMEs (*New Molecular Entities*) representa a quantidade de novas moléculas descobertas e a sigla BLAs (*Biologic License Applications*) representa o número de licenças de aplicação biológicas, ou seja, o número de licença de novos medicamentos.

Apesar da quantidade de fármacos comercializáveis ter aumentado, ela ainda não corresponde ao esperado diante dos altos investimentos realizados em pesquisa e desenvolvimento (P&D) pelas indústrias farmacêuticas (LABBE et al., 2015). Podemos

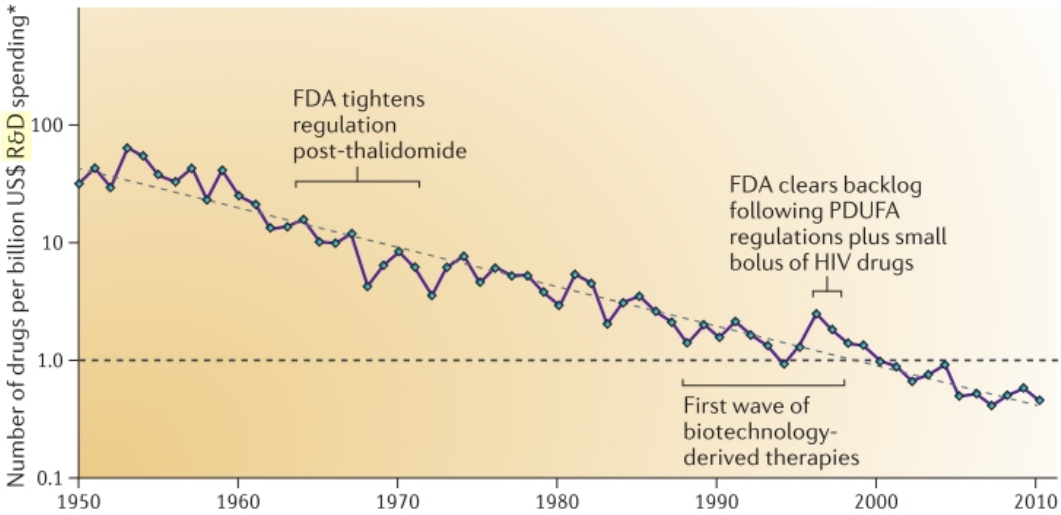
Figura 2 – Número de medicamentos aprovados pela FDA em 2015.



Fonte: Mullard (2016)

observar na Figura 3, a queda de eficiência na descoberta de novos fármacos por bilhão de dólares investidos entre 1950 e 2010.

Figura 3 – Tendência geral de queda na eficiência de pesquisa e desenvolvimento de fármacos por bilhões de dólares investidos, com ajuste de inflação.



Fonte: Scannell et al. (2012)

Na última década, empresas farmacêuticas têm se reestruturado buscando a redução de custos e riscos, realizando cortes principalmente nas fases de pré-desenvolvimento, desenvolvimento e pré-clínica, onde a maioria dos projetos de pesquisa baseiam-se em ideias especulativas e de alto risco. O período onde falta investimento e produtividade no processo de descoberta de medicamentos é chamado de "vale da morte" (OSAKWE; RIZVI, 2016).

Um número crescente de pequenas empresas têm se dedicado ao processo de planejamento de fármacos, tomando os altos riscos das fases iniciais dos projetos de descoberta de medicamentos. A modelagem computacional têm desempenhado um papel importante nesse cenário, visto que grande parte dessas pequenas empresas trabalham com abordagens *in silico* na tentativa de otimizar o processo evitando uma grande quantidade de experimentos que têm chance de falhar. Alguns desses métodos ganharam destaque recentemente com o prêmio Nobel em Química de 2013 (NOBELPRIZE, 2013).

O *docking* molecular é um procedimento computacional que tem como objetivo realizar a simulação da interação entre pequenas moléculas (ligantes), candidatas à se tornarem novos medicamentos, com um alvo biológico (receptor) de interesse, normalmente uma ou mais proteínas relacionadas a doença em questão, predizendo a afinidade de ligação entre elas (TROTT; OLSON, 2010).

O *Virtual Screening* (VS) consiste na execução de simulações de docking em alta escala, ou seja, ao invés de fazer simulações para um ligante, são realizadas simulações com uma biblioteca de ligantes para tentar predizer os que apresentam resultados mais satisfatórios em relação a atividade biológica no receptor de interesse (BIESIADA et al., 2011). Essa técnica leva a aceleração do processo, pois permite focar os esforços e recursos em experimentos com moléculas promissoras, evitando sequências de longos experimentos que poderiam falhar.

Uma molécula com alta afinidade de ligação não é suficiente para a definição de um fármaco eficaz, outras características como a concentração em que o medicamento se encontra no organismo no momento da ligação com o receptor de interesse a fim de se atingir o efeito terapêutico desejado, a menor quantidade possível de efeitos colaterais e ser um fármaco seguro, também são atributos importantes.

As propriedades relacionadas à Absorção, Distribuição, Metabolismo, Excreção e Toxicidade (ADMET) também são de suma importância na definição de um fármaco eficaz. De acordo com Hecht (2011), uma análise das falhas na fase de testes clínicos nas décadas de 1980 e 1990 mostraram que aproximadamente 40% estavam relacionadas a um baixo índice de Absorção, Distribuição, Metabolismo e Excreção. Em razão do alto custo para se realizar estudos relacionados a essas propriedades, que eram predominantemente experimentos *in vivo*, estes eram realizados nos estágios finais dos projetos de descoberta de novos fármacos, contribuindo para essa alta taxa de falhas nas etapas finais dos projetos de descoberta de fármacos (HECHT, 2011).

Com o advento das tecnologias que possibilitaram experimentos *in vitro*, a otimização das propriedades ADMET nos estágios iniciais do planejamento de medicamentos se tornou uma realidade, aumentando o número de informações disponíveis sobre os compostos químicos, que viabilizaram a criação de métodos *in silico* para a investigação entre a relação da estrutura do composto com suas propriedades ADMET (GOLA et al.,

2006).

A integração entre métodos *in vitro* e *in silico* na triagem e modelagem de compostos com propriedades ADME favoráveis e com boa afinidade de ligação contra um alvo de interesse se mostraram bem-sucedidas, visto que a taxa de falhas em testes clínicos devido ao baixo índice de Absorção, Distribuição, Metabolismo e Excreção caiu para 10-14% a partir de 2010 (HECHT, 2011).

Uma das principais dificuldades de se realizar estudos sobre toxicidade no começo do processo de desenvolvimento de um novo medicamento é que as causas e consequências relacionadas a toxicidade dependem de vários fatores, onde a reação tóxica observada pode ser o resultado final de uma série de eventos bioquímicos, que podem ser dependentes da relação entre a dose-tempo na administração do medicamento (GOLA et al., 2006). Com a otimização das propriedades ADME no início do desenvolvimento de medicamentos, uma das principais causas de falhas nos testes clínicos é a toxicidade, que se tornou um objeto de estudo de interesse nas abordagens *in silico*.

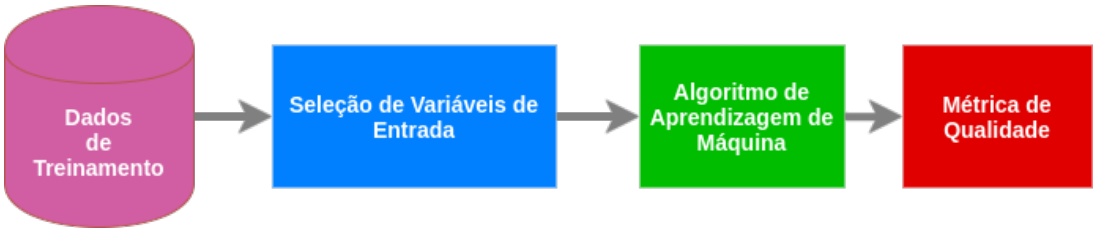
O principal objetivo em toxicologia é compreender os mecanismos bioquímicos envolvidos e associados a atividades toxicológicas em relação a um *endpoint* bem definido. A principal característica dos métodos *in silico* empregados em estudos toxicológicos é a exploração de conhecimento sobre um grupo de compostos químicos conhecidos com o objetivo de predizer o grau de toxicidade de outros compostos (Konstantin V. Balakin, 2010).

Segundo Hecht (2011), a adoção de algoritmos de inteligência computacional na busca de desenvolvimento de melhores modelos de predição de atividades toxicológicas têm sido crescente. A área de inteligência computacional emprega algoritmos que são capazes de aprender uma determinada tarefa a partir dos dados disponíveis, selecionando, analisando e interpretando automaticamente informações relevantes extraídas dos dados disponibilizados. Uma das famílias de algoritmos mais utilizada na predição de toxicidade é a dos algoritmos de aprendizagem de máquina. Na Figura 4, podemos observar um diagrama de fluxo simplificado explicando o funcionamento do processo de treinamento de um algoritmo de aprendizado de máquina.

Os dados de treinamento são os dados disponíveis em que os algoritmos de aprendizado de máquina estão sendo empregados. Em nosso contexto os dados de treinamento correspondem ao subconjunto de moléculas que possuem atividades toxicológica conhecida em relação ao *endpoint* que pretendemos analisar.

A etapa de seleção de variáveis de entrada é onde buscamos definir os dados de entrada mais relevantes para que nosso modelo tenha maiores chances de funcionar de maneira adequada. Após a seleção dos dados de entrada, aplicamos o algoritmo de

Figura 4 – Diagrama de fluxo simplificado explicando o processo de treinamento de um algoritmo de aprendizagem de máquina.



Fonte: Imagem desenvolvida pelo autor.

aprendizado de máquina escolhido em cima destes dados, esse processo é chamado de treinamento. Por último, é aplicada uma métrica de qualidade, que é responsável por nos mostrar a acurácia de predição do nosso modelo em relação aos dados de treinamento. O subconjunto de moléculas que possuem atividades toxicológicas desconhecidas em relação ao *endpoint* para qual o algoritmo de aprendizado de máquina foi treinado, é chamado de conjunto de busca, são essas moléculas as quais queremos realizar a predição da atividade toxicológica.

De acordo com Konstantin V. Balakin (2010), o processo de construção de um modelo para a predição de toxicidade pode ser dividido em cinco etapas: 1) Geração de dados, 2) Seleção de dados, 3) Treinamento do modelo, 4) Validação do modelo e 5) Interpretação do modelo. Os algoritmos de aprendizado de máquina podem ser aplicados na etapa de seleção de dados de entrada, onde são utilizados para remover dados irrelevantes ou correlatos, e também na construção do modelo de predição em si.

Os descritores moleculares são o resultado final de procedimentos matemáticos que transformam informações químicas em números, os quais podem ser utilizados como dados de entrada em modelos que utilizam algoritmos de inteligência computacional (Konstantin V. Balakin, 2010). Uma das principais etapas no desenvolvimento dos modelos de predição é a identificação dos descritores moleculares a serem utilizados, além da redução no número de descritores utilizados quando possível.

O termo *Big Data* foi caracterizado por Laney (2001) como modelo dos 3 V's, que são Volume, Velocidade e Variedade, respectivamente. O Volume refere-se à quantidade de informação em si, a Velocidade refere-se à rapidez em que os dados são gerados e a Variedade refere-se à variedade de tipos de dados disponíveis em determinado domínio.

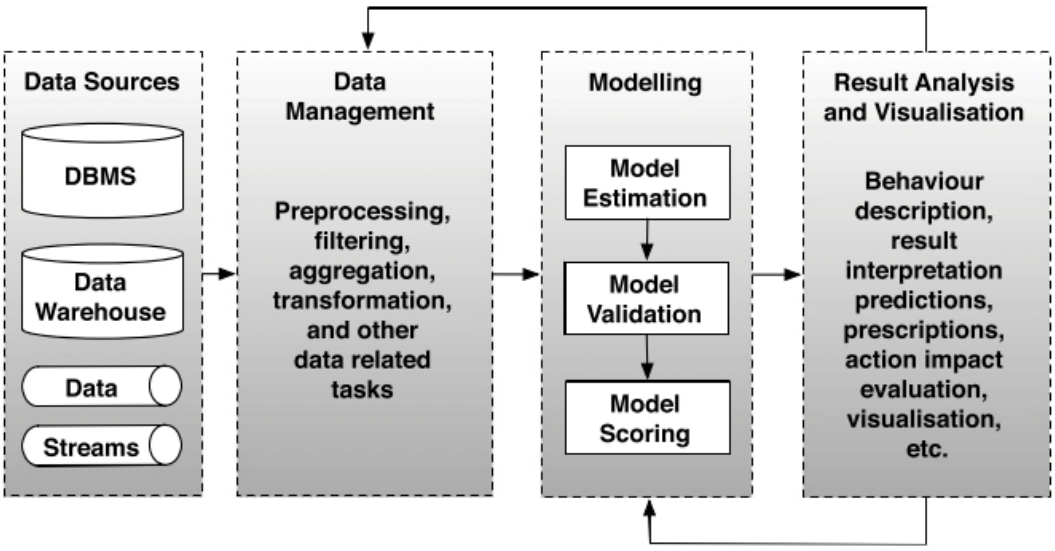
Segundo Assunção et al. (2015), os tipos de dados normalmente são classificados em: 1) Estruturados, 2) Não-Estruturados, 3) Semi-Estruturados e 4) Mistos.Os dados classificados como estruturados possuem um modelo de dados bem definido, ao contrário dos classificados como não-estruturados; os dados classificados como semi-

estruturados possuem um modelo de dados flexível e, por fim, dados que mesclam a característica de dois ou mais tipos apresentados anteriormente são considerados mistos.

Assunção et al. (2015) também define que soluções analiticas podem ser classificadas como descritivas, as quais usam dados históricos para identificar padrões, preditivas, as quais tentam predizer situações desconhecidas através dos dados atuais e históricos, e prescritivas, que são soluções que apoiam analistas na tomada de decisão, por exemplo, avaliando o impacto de determinadas escolhas no processo como um todo.

O termo *Big Data Analytics* tem sido usado para descrever a tarefa de analisar dados que se encaixam no paradigma do *Big Data*, ou seja, dados provenientes de várias fontes, que são gerados muitas vezes em alta velocidade e possuem um grande volume. Realizar essa tarefa exige métodos eficientes para armazenar, filtrar, transformar e recuperar dados (ASSUNÇÃO et al., 2015). Na Figura 5 podemos observar um fluxo de *Big Data Analytics*.

Figura 5 – Fluxo do processo de *Big Data Analytics*.



Fonte: (ASSUNÇÃO et al., 2015)

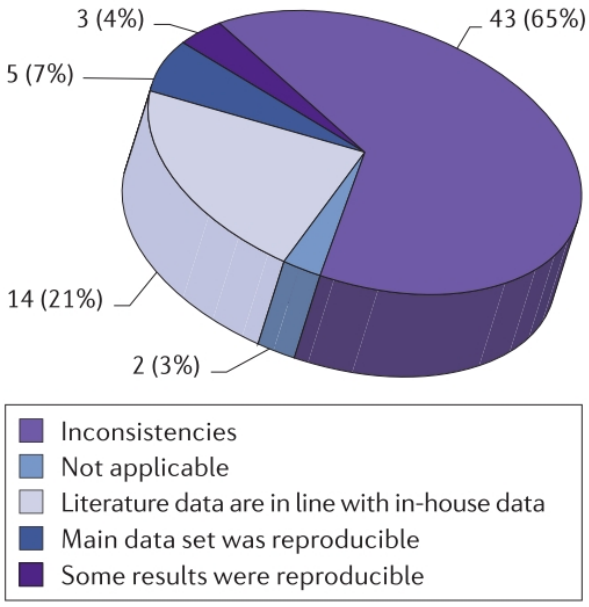
A problemática do *Big Data* têm se difundido de maneira crescente em questões de interesse da indústria farmacêutica, como os projetos para descoberta de novos fármacos, uma vez que moléculas candidatas a se tornarem medicamentos são oriundas de várias fontes de dados, como dados da literatura, campanhas internas para identificação de moléculas, dentre outras; além do volume de informações disponíveis aumentar consideravelmente a cada ano (BANERJEE et al., 2016).

A validação de dados obtidos por meio da literatura e outras fontes de dados

públicas é crucial no processo de tomada de decisão de um laboratório quando o mesmo considera investir recursos no estudo de um grupo de moléculas candidatas a se tornarem fármacos, o que destaca a importância da reprodutibilidade das metodologias e modelos computacionais utilizados em projetos de descoberta de novos medicamentos (BANERJEE et al., 2016).

Prinz et al. (2011) realizou uma análise de 67 artigos relacionados ao tema, coletando dados e tentando reproduzir a metodologia, o resultado pode ser observado na Figura 6.

Figura 6 – Fluxo do processo de *Big Data Analytics*.



Fonte: (PRINZ et al., 2011)

Podemos observar na Figura 6 que somente 7% dos trabalhos foram reproduzidos totalmente e apenas 4% foram reproduzidos parcialmente. A análise realizada por Prinz et al. (2011) também constatou que mesmo publicações em revistas de prestígio não garantem a reprodutibilidade dos estudos publicados.

Apesar da adesão de simulação computacional nas diversas áreas da ciência proporcionar avanços significativos, ela também trás alguns desafios quando abordamos o tema reprodutibilidade. Até a obtenção dos resultados finais, os dados são analisados diversas vezes, com modificações de métodos e parâmetros, ou até mesmo atualização nos dados, porém, em grande parte das ocasiões os detalhes computacionais recebem pouca atenção no texto publicado (MESIROV, 2010).

De acordo com Mesirov (2010), dois elementos são de suma importância na tentativa de realizar e publicar uma pesquisa que faz uso da simulação computacional de maneira reprodutível. O primeiro elemento é um ambiente de pesquisa reprodutível,

o qual fornece as ferramentas e dados necessários para se realizar as computações e análises a fim de se obter os mesmos resultados. O segundo elemento é um sistema de publicação de pesquisas reprodutíveis, como um editor de textos padrão, porém, que forneça uma integração com o ambiente de pesquisa mencionado anteriormente.

Com o evidenciamento da importância da reprodutibilidade e de uma necessidade crescente de modelos computacionais utilizados no contexto de descoberta de novos fármacos estarem aptos a lidarem com *Big Data*, este trabalho tem como principal objetivo promover o desenvolvimento de uma metodologia para a criação de modelos computacionais reprodutíveis e escaláveis, aptos a liderem com a problemática do *Big Data*, para a predição de propriedades de Absorção, Distribuição, Metabolismo, Excreção e Toxicidade (ADMET) de medicamentos a partir de bases de dados que possuam a estrutura 2D de moléculas no formato SMILES, onde as mesmas estejam devidamente classificadas.

2 Revisão de Literatura

Este capítulo tem como principal objetivo elucidar os conceitos relacionados ao tema do trabalho, buscando atender o cunho multidisciplinar do mesmo, explicando o domínio do trabalho, bem como as técnicas utilizadas no desenvolvimento do mesmo.

2.1 Farmacologia

Como podemos observar anteriormente, o desenvolvimento de novos fármacos tem se tornado cada vez mais desafiador devido ao declínio da eficiência no processo de pesquisa e desenvolvimento de medicamentos. Um dos principais fatores na elaboração bem sucedida de uma nova droga é o balanceamento entre eficácia e segurança, que também é um dos principais fatores para a maioria das drogas que chegam na fase de teste clínicos não serem lançadas no mercado (PIRES et al., 2015).

Muitos projetos de descoberta de fármacos concentram-se inicialmente na identificação de moléculas que possuem boa afinidade de ligação com um receptor de interesse, abordando as propriedades farmacocinéticas e de toxicidade em estágios finais do processo. A interação entre propriedades farmacocinéticas, propriedades de toxicidade e potência são essenciais para a definição de um fármaco efetivo, sendo as duas primeiras cruciais para que um medicamento tenha o efeito terapêutico desejado e consequentemente seja lançado no mercado (PIRES et al., 2015).

Para compreender e delimitar termos chaves para o desenvolvimento desse trabalho como, por exemplo, fármacos, farmacocinética, toxicidade, ligantes, receptores, afinidade de ligação, dentre outros, é necessário discutir o que é farmacologia e qual a importância dessa disciplina para o entendimento dos referidos termos.

De acordo com Hollinger (2007), farmacologia é a ciência dos fármacos, a origem da palavra vem do grego *pharmakos*, que significa fármaco ou droga, e *logos*, que significa estudo. Definindo de forma mais ampla, farmacologia é a ciência que estuda o efeito de medicamentos em organismos vivos, tentando descrever as respostas biológicas que são produzidas pela aplicação de medicamentos, além de tentar entender os mecânismos pelos quais essas respostas biológicas são geradas.

Segundo Hollinger (2007), um fármaco ou droga, é uma substância química que pode alterar ou influenciar a capacidade de resposta de um sistema biológico, seja imitando, facilitando ou contrapondo um fenômeno que ocorre naturalmente. A *Food and Drug Administration* (FDA), agência governamental que é responsável pela fiscalização e regulação de medicamentos nos EUA, define que todos fármacos são

compostos químicos, mas nem todos compostos químicos são fármacos, onde compostos químicos são substâncias constituídas de uma combinação de elementos (elétrons, prótons e nêutrons). Ainda de acordo com a FDA um fármaco é um composto químico utilizado para diagnóstico, prevenção, cura ou melhoria de uma condição indesejada de saúde.

As raízes da farmacologia remontam o tempo em que nossos ancestrais viviam nas savanas africanas entre 5 a 10 milhões de anos atrás. A utilização de plantas na alimentação, rituais religiosos, entre outras atividades culturais, permitiram que nossos ancestrais desenvolvessem conhecimento empírico em relação à propriedades terapêuticas de determinadas plantas (HOLLINGER, 2007).

Claudius Galen pode ser considerado um dos primeiros praticantes da farmacologia, visto que, por volta do ano 150 d.C., chamou atenção para a importância do trabalho teórico e da realização de experimentos no uso racional de medicamentos (VALLANCE; SMART, 2006). Segundo a teoria de Galen, existiam quatro tipos de humores corporais, o sangue, fleuma, bílis amarela e bílis preta, quando estes estavam em harmonia, gozávamos de saúde, quando estavam desequilibrados, doenças surgiriam (HOLLINGER, 2007).

Um médico suíço chamado *Phillippus Theophrastus von Hohenheim* (1493-1541), que a partir de 1516 assumiu o nome *Paracelsus*, começou a questionar as doutrinas transmitidas desde a antiguidade (HOLLINGER, 2007). *Paracelsus* impulsionou a farmacologia encorajando a investigação dos princípios ativos e ingredientes presentes nos medicamentos de muitas receitas medievais, descartando a teoria humoral de Galen, que até então era o fundamento da medicina na época (VALLANCE; SMART, 2006).

Conforme Hollinger (2007), a comunidade médica da época não aceitou a teoria de *Paracelsus*, que chegou a ser considerado um envenenador. Um dos motivos dados na época para a comunidade médica dar-lhe essa alcunha era o uso de substâncias inorgânicas na medicina, que eram consideradas tóxicas demais para serem utilizadas como agentes terapêuticos. *Paracelsus* se defendeu com a tese, que posteriormente se tornou um axioma na área de toxicologia, que todos os fármacos podem ser venenosos dependendo da quantidade administrada e reconheceu que uma única dose pode determinar se um composto é ou não terapeuticamente útil ou tóxico (VALLANCE; SMART, 2006).

A farmacologia contemporânea, surgiu em meados do século XIX com *Rudolph Buchheim e Oswald Schmiedeberg*. De acordo com Hollinger (2007), a inclusão da farmacodinâmica como um componente da farmacologia é atribuída aos esforços de *Buchheim*, além de acreditar-se que ele tenha criado o primeiro laboratório de farmacologia do mundo na Universidade de Dorpat na Hungria. *Schmiedeberg* realiza estudos referentes ao efeito da aplicação de muscarina no coração, funda a primeira revista cientifica

de farmacologia do mundo e também funda o Instituto de Farmacologia de Strasbourg, no qual realiza estudos da farmacologia do clorofórmio. Foi nesse período da história da farmacologia que vários conceitos como receptores de fármacos, relação estrutura-atividade, dentre outros foram desenvolvidos (VALLANCE; SMART, 2006).

Dentro da disciplina de farmacologia, os assuntos presentes nos domínios da farmacodinâmica, farmacocinética e toxicologia são importantes para entendermos as motivações para realização de estudos como este trabalho. Alguns tópicos importantes para este trabalho presentes nesses domínios serão apresentados nas próximas seções.

2.2 Farmacodinâmica

Segundo Pires et al. (2015), muitos projetos de descoberta de fármacos focam inicialmente na seleção de moléculas que possuem boa afinidade de ligação com os receptores de interesse, tornando a potência das moléculas contra o receptor um fator determinante para as decisões nas etapas iniciais dos projetos. Para entendermos mais sobre afinidade de ligação entre receptor-ligante, se faz necessário definirmos o que é um receptor, além de entender como funciona a interação entre fármaco-receptor. O estudo dos efeitos dessa interação entre fármaco e receptor no organismo como um todo, gerando métricas quantitativas que permitem estabelecer doses apropriadas dos medicamentos em questão para os pacientes, bem como comparar a potência, eficácia e a segurança de um fármaco, é chamado de farmacodinâmica.

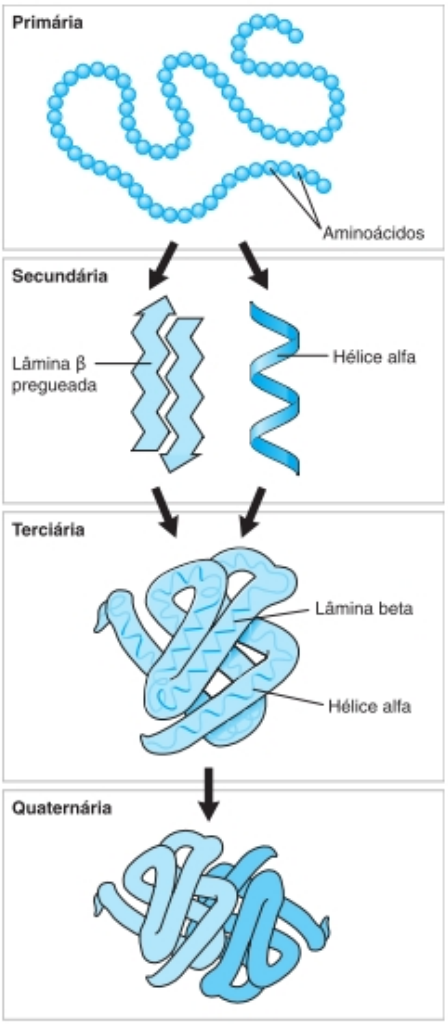
Para produzir o efeito terapêutico desejado, um fármaco, apesar de poder se ligar com praticamente qualquer tipo de alvo tridimensional, interage com moléculas específicas, comumente chamadas de receptores, que desempenham papéis importantes na função fisiológica e fisiopatológica. Muitas vezes essa interação com o receptor também pode ser responsável pelo desencadeamento de efeitos adversos de um fármaco (GOLAN et al., 2009).

Segundo Bacq (2013), os receptores de fármacos são macromoléculas que, através de sua ligação a determinado fármaco, permeiam as alterações bioquímicas e fisiológicas causas por este.

Devido a grande parte dos receptores de fármacos serem proteínas, é conveniente entendermos os níveis estruturais destas. Como podemos observar na Figura 7, as proteínas possuem quatro níveis de estrutura. A sequência de aminoácidos de uma proteína, que é determinada pelas sequências do DNA que codificam a proteína, é chamada de estrutura primária. A estrutura secundária é resultado da interação de átomos de hidrogênio de carga positiva com átomos de oxigênio de carga negativa em carbonos da mesma proteína, que determinam diversos padrões secundários, característicos da conformação da proteína, incluindo hélices alfa e a lâmina β pregueada. A estrutura

terciária refere-se ao arranjo tridimensional resultante da interação de todos os resíduos de aminoácidos da proteína (WALSH, 2014). A estrutura quaternária é estabelecida a partir da interação de ligação entre duas subunidades proteicas independentes.

Figura 7 – Níveis estruturais das proteínas



Fonte: (GOLAN et al., 2009)

A forma da proteína também é influenciada pela afinidade de suas partes pela água. Como a água está presente tanto no meio extracelular quanto no intracelular, os segmentos proteicos hidrofóbicos estão frequentemente retraídos no interior da proteína ou protegidos da água pela sua inserção em membranas de dupla camada lipídica, ao contrário dos segmentos hidrofílicos, que frequentemente se localizam na superfície externa da proteína. É importante ressaltar que, uma vez que a forma da proteína esteja estável, a mesma se torna responsável pela determinação de sua função, sua localização no corpo, sua relação com as membranas celulares e como se dá a interação de ligação com os fármacos e as demais moléculas (WALSH, 2014).

A ligação fármaco-receptor resulta da soma de múltiplas interações químicas

entre as duas moléculas em uma região do receptor chamada sítio de ligação. Cada sítio de ligação possui características químicas singulares, determinadas pelas propriedades específicas dos aminoácidos que o compõe (GOLAN et al., 2009).

O termo afinidade é utilizado para designar a probabilidade de ocorrer interação entre o fármaco e o sítio de ligação no receptor de interesse, fatores como hidrofobicidade, hidrofilicidade e pK_a dos aminoácidos próximos ao sítio de ligação podem afetar a afinidade dessa interação (GOLAN et al., 2009).

De acordo com Golan et al. (2009) as principais forças que promovem a afinidade fármaco-receptor são: 1) forças de van der Waals, 2) ligações de hidrogênio, 3) ligações iônicas, 4) ligações covalentes. Essas forças são descritas no Quadro 1, que também indica a força relativa de cada um dos tipos de ligação.

Quadro 1 – Forças que promovem afinidade fármaco-receptor

Tipo de Ligação	**Mecanismo**	**Força da ligação**
van der Waals	A mudança de densidade de elétrons em áreas de uma molécula ou em uma molécula como um todo resulta na geração de cargas positivas ou negativas transitórias. Essas áreas interagem com áreas transitórias de carga oposta sobre uma outra molécula.	+
Hidrogênio	Os átomos de hidrogênio ligados ao nitrogênio ou oxigênio tornam-se mais positivamente polarizados, permitindo a sua ligação a átomos de polarização mais negativa, como oxigênio, nitrogênio ou enxofre.	++
Iônica	Os átomos com excesso de elétrons (conferindo ao átomo uma carga negativa global) são atraídos por átomos com deficiência de elétrons (conferindo ao átomo uma carga positiva global).	+++
Covalente	Dois átomos em ligação compartilham elétrons.	++++

Fonte: Golan et al. (2009)

Quando um fármaco consegue se ligar ao receptor de interesse, uma reação, decorrente dessa interação de ligação, pode ocorrer. O efeito da reação resultante da ligação fármaco-receptor pode inclusive ser observada a nível de um órgão ou até mesmo do paciente. Diante deste cenário, um modelo o qual consiga descrever a ligação de um fármaco com um receptor para prever o efeito do fármaco em todos os seus

níveis (celular, tecidual, organismo) se torna interessante. O estudo, desenvolvimento e aplicação desse modelo são papéis da farmacodinâmica (GOLAN et al., 2009).

A equação (1) descreve um caso simples de ligação fármaco-receptor em que o receptor se encontra livre ou reversivelmente ligado à um fármaco.

$$L + R \underset{k_{livre}}{\overset{k_{ligado}}{\rightleftarrows}} LR \tag{1}$$

Os coeficientes L, R e LR da equação 1 são respectivamente, o fármaco, o receptor livre e o complexo fármaco-receptor. A fração de receptores em cada um desses estados depende da constante de dissociação, K_d, onde $K_d = k_{livre}/k_{ligado}$. K_d é uma propriedade intrínseca de qualquer par fármaco-receptor, que varia com a temperatura, porém, como a temperatura do corpo humano é relativamente constante, pode-se considerar que K_d é uma constante para cada combinação de fármaco-receptor (PAN et al., 2013).

A associação entre receptor livre e receptor pode ser descrita como a equação (2) apresentada abaixo:

$$K_d = \frac{[L][R]}{[LR]} \text{ reorganizada para } [LR] = \frac{[L][R]}{K_d} \tag{2}$$

Os coeficientes $[L]$, $[R]$ e $[LR]$ são respectivamente, a concentração de ligante livre, a concentração de receptor livre, a concentração do complexo fármaco-receptor. Como K_d é uma constante, algumas propriedades relativas à interação fármaco-receptor podem ser deduzidas a partir da equação (2). Podemos deduzir primeiramente que, medida que aumentamos a concentração de medicamento, a concentração de receptores ligados também aumentará. A segunda propriedade que podemos deduzir, que não é tão evidente, é que à medida que a concentração de receptores livres aumenta, a concentração de receptores ligados também aumenta. Consequentemente, pode acontecer um aumento no efeito de um fármaco em decorrente a um aumento na concentração do ligante ou do receptor (PAN et al., 2013).

Com o intuito de facilitar a explicação da teoria, iremos assumir que a concentração do total de receptores é uma constante, de modo que $[LR] + [R] = [R_0]$, permitindo-nos ordenar a equação (3) da seguinte forma:

$$[R_0] = [R] + [LR] = [R] + \frac{[L][R]}{K_d} = [R]\left(1 + \frac{[L]}{K_d}\right) \tag{3}$$

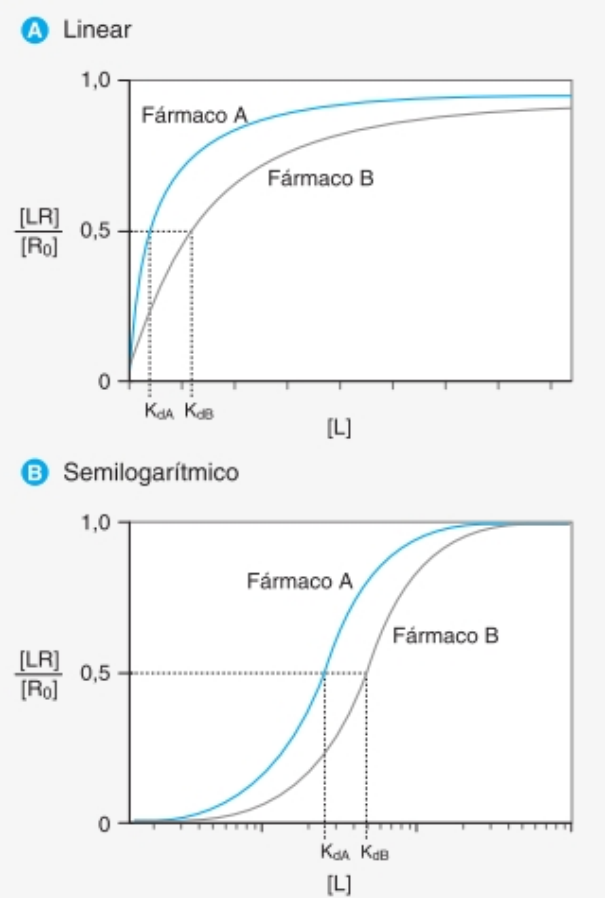
Resolvendo $[R]$ e fazendo as devidas substituições na equação (2) a partir da

equação (3), obtemos:

$$[LR] = \frac{[R_0][L]}{[L] + K_d}, \text{ reorganizada para } \frac{[LR]}{[R_0]} = \frac{[L]}{[L] + K_d} \tag{4}$$

Na Figura 8 apresentamos os gráficos de curva de ligação fármaco-receptor, em escala linear na Figura 8a, e em escala semilogarítmica na Figura 8b. Podemos observar na Figura 8 que a ligação fármaco-receptor máxima ocorre quando $[LR]$ é equivalente a $[R_0]$, ou $[LR]/[R_0] = 1$. Também podemos perceber que quando $[L] = K_d$, então $[LR]/[R_0] = K_d/2K_d = 1/2$, consequentemente a constante K_d pode ser definida como a concentração de ligante em que 50% dos receptores disponíveis estão ocupados (BRUNTON et al., 2006).

Figura 8 – Gráficos de curva de ligação fármaco-receptor



Fonte: (GOLAN et al., 2009)

De acordo com Golan et al. (2009), de maneira intuitiva é esperado que a realização dose-resposta esteja relacionada estreitamente com a relação de ligação fármaco-receptor, o que realmente ocorre para muitas combinações fármaco-receptor, partindo do princípio de que a resposta de um fármaco é proporcional à concentração de re-

ceptores ocupados pelo fármaco. Essa pressuposição pode ser quantificada através da seguinte equação:

$$\frac{\text{resposta}}{\text{resposta máx.}} = \frac{[DR]}{[R_0]} = \frac{[D]}{[D] + K_d} \tag{5}$$

$[D]$ é a concentração do fármaco livre, $[DR]$ a concentração do complexo fármaco-receptor, $[R_0]$ a concentração total de receptores e K_d a constante de dissociação e equilibrio para a interação fármaco-receptor.

Existem dois tipos principais de relações dose-resposta, as relações graduadas, que descrevem o efeito de várias doses de um medicamento sobre o indivíduo, e as relações quantais, que descrevem o efeito de várias doses de um medicamento sobre uma população de indivíduos (HOLLINGER, 2007).

A potência e a eficácia de um medicamento podem ser deduzidos a partir da curva de dose-resposta graduada. A potência (EC_{50}) de um medicamento refere-se à concentração em que o mesmo produz 50% de sua resposta máxima. A eficácia ($EC_{mx.}$) refere-se à resposta máxima produzida pelo medicamento (BRUNTON et al., 2006).

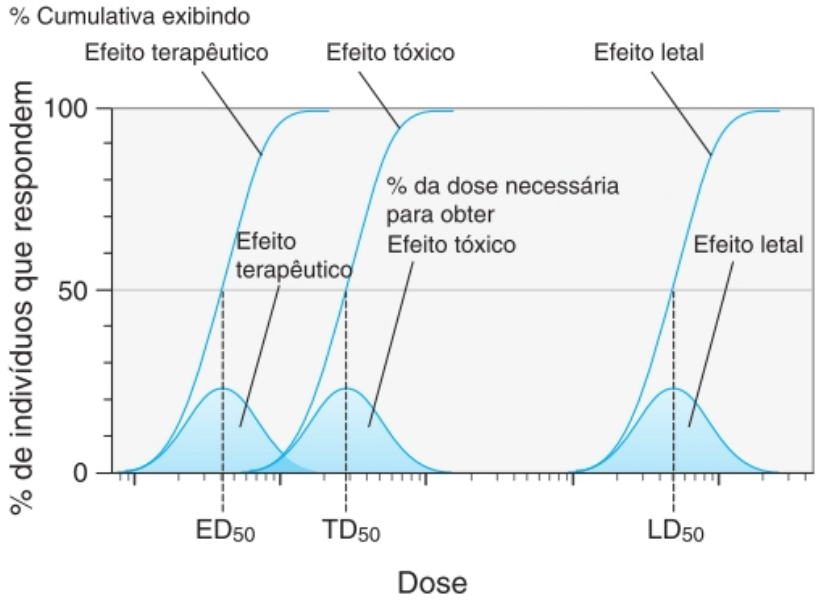
Segundo Golan et al. (2009), a relação dose-resposta quantal representa graficamente a fração da população que responde a determinada dose de um medicamento como função da dose deste medicamento. Como as respostas biológicas podem ser diferentes entre indivíduos, os efeitos de um medicamento são observados ao longo de uma faixa de doses, onde as respostas são definidas em termos de presentes ou ausentes, por exemplo, "com sono / sem sono", o objetivo é generalizar um resultado para uma população. Três métricas principais podem ser obtidas com a relação de dose-resposta quantal, essas métricas são efetividade (efeito terapêutico), toxicidade (efeito adverso) e letalidade (efeito letal).

De acordo com Brunton et al. (2006), as doses que produzem essas métricas em 50% da população analisada são conhecidas como dose efetiva mediana (ED_{50}), dose tóxica mediana (TD_{50}) e dose letal mediana (LD_{50}), respectivamente. Na Figura 9 podemos observar uma representação gráfica dessas métricas relacionadas à dose-resposta.

A grande maioria dos receptores podem ser categorizados dentro de dois estados de conformação, ativo ou inativo. Os fármacos possuem algumas denominações de acordo com a ação que exercem sobre o estado de conformação do receptor. Essas denominações são: 1) agonistas, 2) agonistas parciais, 3) agonistas inversos, 4) antagonistas competitivos e, 5) antagonistas não competitivos (GOLAN et al., 2009).

De acordo com Brunton et al. (2006), um agonista é um composto que se acopla a um receptor e o ativa. A equação (7) fornece um modelo para entender a relação entre

Figura 9 – Representação gráfica das métricas de relação dose-resposta.



Fonte: (GOLAN et al., 2009)

ligação do agonista e a ativação do receptor:

$$\begin{array}{ccc} D + R & \rightleftarrows & D + R^* \\ \uparrow\downarrow & & \uparrow\downarrow \\ DR & \rightleftarrows & DR^* \end{array} \tag{6}$$

Os coeficientes D, R, DR e R^* representam respectivamente a concentração do fármaco, concentração dos receptores livres, concentração do complexo agonista-receptor e a indicação da conformação ativa do receptor. Para a maioria dos pacientes e dos agonistas, R^* e DR são espécies instáveis que têm apenas uma existência breve, sendo quantativamente insignificantes em comparação com R e DR^*, consequentemente, na grande maioria dos casos podemos simplificar a equação (7) para:

$$D + R \rightleftarrows DR^* \tag{7}$$

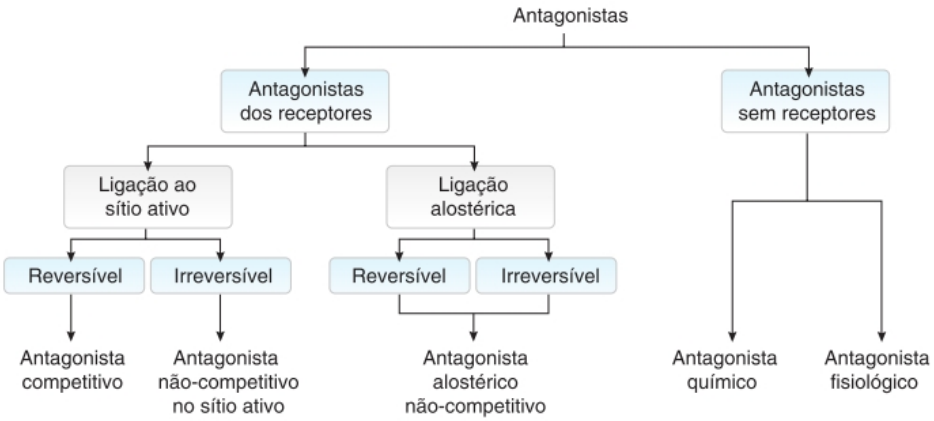
Observe que a equação (7) é idêntica a equação (1), que foi utilizada para análise da ligação fármaco-receptor, sugerindo que para a maioria dos receptores, a ligação do agonista é proporcional à ativação do receptor (GOLAN et al., 2009).

De acordo com Brunton et al. (2006), um agonista parcial é um composto que se acopla a um receptor em seu sítio ativo, mas que só produz uma resposta parcial. Os agonistas inversos normalmente atuam ligando-se à um receptor em sua forma inativa

e realizando sua estabilização, isso tem como efeito a desativação dos receptores que se encontram na forma ativa na ausência do fármaco (GOLAN et al., 2009).

Na Figura 10 podemos observar a classificação dos fármacos antagonistas.

Figura 10 – Classificação dos antagonistas.



Fonte: (GOLAN et al., 2009)

Segundo Brunton et al. (2006), um antagonista é um composto que se liga a um receptor, porém, não o ativa e, consequentemente, não gera nenhuma resposta. Essa intervenção influência na ação dos agonistas, visto que estes não conseguem se ligar ao receptor em questão. Os antagonistas podem desmembrados em antagonistas de receptores e antagonistas sem receptores.

O antagonista de receptor liga-se ao sítio ativo (sítio de ligação do agonista) ou a um sítio alostérico de um receptor. A ligação do antagonista no sítio ativo impede a ligação do agonista ao receptor, enquanto a ligação do antagonista a um sítio alostérico altera K_d, o que impede a mudança de conformação necessária para ativação do receptor. Os antagonistas de receptores também podem ser desmembrados em antagonistas reversíveis e irreversíveis, ou seja, antagonistas que se ligam a seus receptores de forma reversível e antagonistas que se ligam a seus receptores de forma irreversível (GOLAN et al., 2009).

O antagonista competitivo liga-se reversivelmente ao sítio de um receptor, porém, não estabiliza a conformação necessária para a ativação do receptor, bloqueando a ligação do agonista a seu receptor e mantendo o receptor em seu estado inativo (BRUNTON et al., 2006).

O antagonista não-competitivo pode se ligar a um sítio ativo ou a um sitio alostérico de um receptor, sendo que, quando se liga ao sítio ativo do receptor essa ligação é realizada de modo covalente ou com uma afinidade muito alta, tornando-a irreversível. Quando o antagonista se liga ao sítio alostérico o mesmo atua ao impedir a

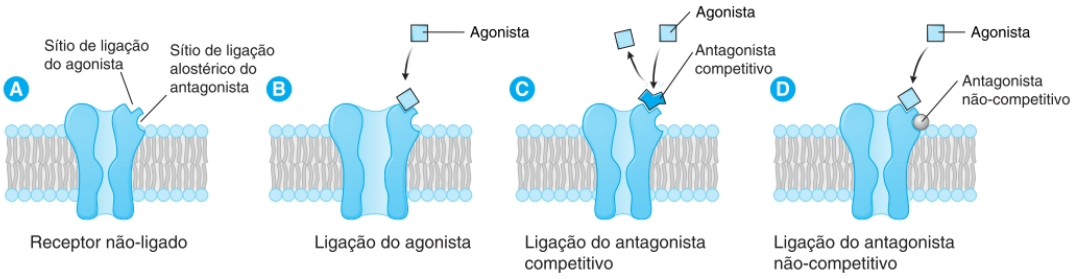
ativação do receptor, mesmo quando o agonista está acoplado ao sítio ativo (BRUNTON et al., 2006).

De acordo com Golan et al. (2009), "uma característica entre antagonistas competitivos e não-competitivos reside no fato de que os antagonistas competitivos reduzem a potência do agonista, enquanto os não-competitivos diminuem a eficácia do agonista". O fato do antagonistas competitivo competir continuamente pela sua ligação ao receptor, diminuindo efetivamente a afinidade do receptor pelo seu agonista, porém, sem limitar o número de receptores disponíveis e de que os antagonistas não competitivos removem os receptores funcionais do sistema, limitando o número de receptores disponíveis, pode explicar essa diferença (GOLAN et al., 2009).

Os antagonistas sem receptores podem ser classificados como químicos ou fisiológicos, no primeiro caso, o antagonista químico inativa o agonista específico ao modificá-lo ou sequestra-lo, de modo que o agonista não é mais capaz de acoplar-se ao receptor e ativá-lo, no segundo caso, o antagonista fisiológico ativa ou bloqueia um receptor que produz uma resposta fisiológica oposta àquela do receptor do agonista (GOLAN et al., 2009).

Na Figura 11, podemos observar como funciona o processo de ligação de antagonistas competitivos e não-competitivos.

Figura 11 – Processo de ligação de antagonistas competitivos e não-competitivos.

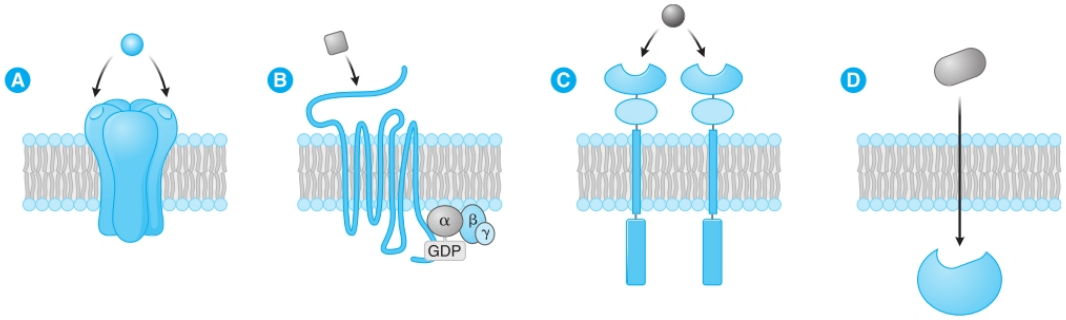


Fonte: (GOLAN et al., 2009)

Apesar da grande diversidade de moléculas de fármacos, a maioria das interações fármaco-receptor conhecidas atualmente podem ser classificadas em seis grupos, 1) canais iônicos transmembrana, 2) receptores transmembrana acoplados a proteínas G intracelulares, 3) receptores transmembrana com domínios citosólicos enzimáticos, 4) receptores intracelulares, incluindo enzimas, reguladores da transcrição e proteínas estruturais, 5) enzimas extracelulares, 6) receptores de adesão de superfície celular (HOLLINGER, 2007). Na Figura 12, podemos observar quatro tipos de interação fármaco-receptores.

Numerosas funções celulares são dependentes da passagem de íons e outras moléculas hidrofílicas através da membrana plasmática, essas funções são reguladas

Figura 12 – Tipos de interação fármaco-receptor. A) Interação através dos canais iônicos transmembrana. B) Interação através de proteínas G. C) Interação através de receptores transmembrana. D) Interação através de receptores intracelulares.



Fonte: (GOLAN et al., 2009)

por canais transmembrana especializados. Os canais iônicos possuem diversas atribuições, incluindo funções fundamentais na neurotransmissão, na condução cardíaca, na contração muscular e na secreção. Consequentemente os medicamentos cuja interação é direcionada para os canais iônicos podem exercer impacto significativos sobre as principais funções orgânicas (HOLLINGER, 2007).

Existem três mecanismos principais na regulação da atividade dos canais iônicos transmembrana, regulação por ligante, regulação por voltagem e regulação por segundo mensageiro, esses mecanismos são apresentados de forma resumida no Quadro 2. O domínio o qual os canais iônicos são regulados por ligantes (que nesse caso está sendo utilizado como um sinônimo de fármaco), pode ser extracelular, localizado dentro do canal, ou intracelular, enquanto o domínio que interage com outros receptores ou moduladores é, com mais frequência, intracelular (HOLLINGER, 2007).

Na Figura 13, apresentamos um exemplo o qual o canal iônico receptor nicotínico de acetilcolina (Ach), que nessa figura está com a resolução de sua estrutura estabelecida em 4,6 Å , é regulado por um ligante.

Esse receptor é formado por cinco subunidades, e cada uma delas atravessa a membrana plasmática. Duas subunidades foram denominadas como α; cada uma contém um único sítio de ligação extracelular para a ACh. No estado livre do receptor, o canal encontra-se fechado por cadeias laterais de aminoácidos e, dessa forma, não permite a passagem de íons. A ligação de duas moléculas de acetilcolina ao receptor causa uma alteração de sua conformação, que abre o canal e permite a passagem de íons (HOLLINGER, 2007).

Segundo Hollinger (2007), os receptores acoplados à proteína G são a classe mais abundante de receptores no corpo humano. Esses receptores, que estão expostos na superfície extracelular da membrana, atravessam a membrana e possuem áreas

Quadro 2 – Três mecanismos principais na regulação da atividade dos canais iônicos.

Tipo de Canal	**Mecanismo de ativação**	**Função**
Regulado por ligante	Ligação do ligante ao canal.	Alteração da condutância iônica
Regulado por voltagem	Alteração no gradiente de voltagem transmembrana.	Alteração da condutância iônica
Regulado por segundo mensageiro	Ligação do ligante ao receptor transmembrana com domínio citosólico acoplado à proteína G, resultando em geração de segundo mensageiro.	O segundo mensageiro regula a condutância iônica do canal

Fonte: Golan et al. (2009)

intracelulares que ativam uma classe singular de moléculas de sinalização, denominadas proteínas G, que são assim denominadas em razão de sua ligação aos nucleotídeos de guanina, GTP e GDP. Esses mecanismos de sinalização estão envolvidos em numerosos processos, incluindo visão, neurotransmissão, entre outros.

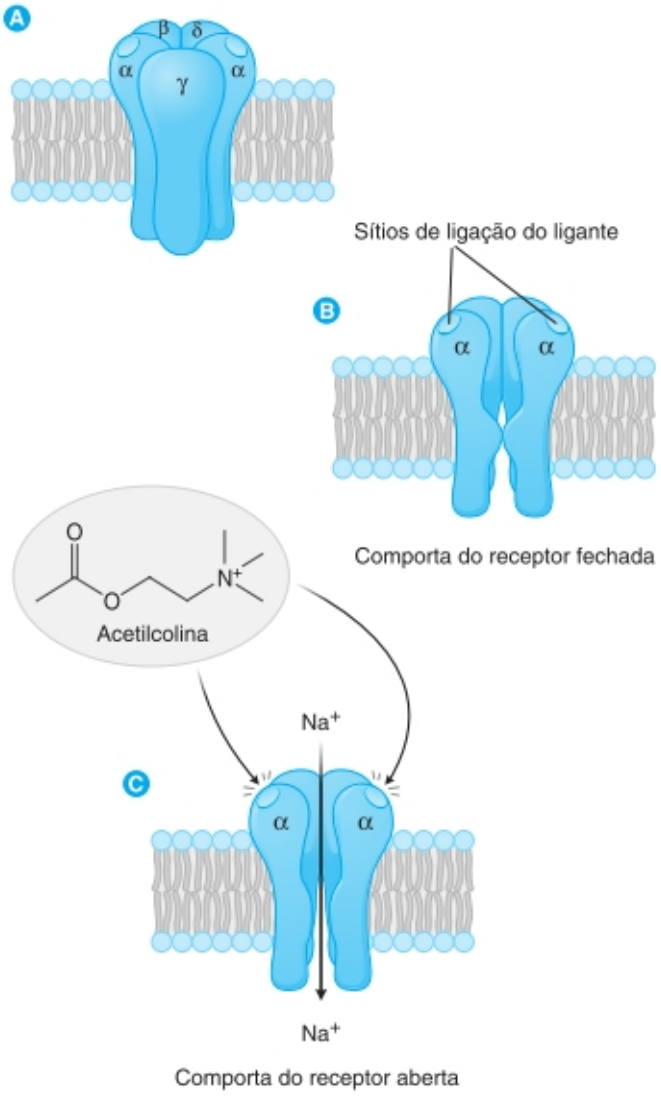
Os receptores acoplados às proteínas G possuem sete regiões transmembrana inclusos dentro de uma única cadeia polipeptídica. Cada região transmembrana corresponde a uma única hélice α, e essas hélices estão atribuídas em um modelo estrutural característico, que se assemelha em todos os membros dessa classe de receptores. O domínio extracelular dessa classe de proteínas possui normalmente a região de ligação do ligante, apesar de alguns receptores acoplados à proteína G se ligarem aos fármacos dentro do domínio transmembrana (HOLLINGER, 2007).

No estado não-estimulado, o domínio citoplasmático do receptor está acoplado de forma não-covalente a uma proteína G, constituída por subunidades α e β_y. Com o processo de ativação, a subunidade α efetua a troca de GDP por GTP. Posteriormente, a subunidade α-GTP se separa da subunidade β_y, e a subunidade α ou β_y difunde-se ao longo do folheto interno da membrana plasmática para interagir com diversos efetores diferentes. Os sinais intermediados pelas proteínas G normalmente são interrompidos pela hidrólise do GTP a GDP, que é catalisada pela atividade inerente de GTPase da subunidade α, como pode ser observado na Figura 14 (HOLLINGER, 2007).

A classe dos receptores transmembrana com domínios citosólicos enzimáticos transformam uma interação de ligação com ligantes extracelulares numa ação intracelular através da ativação de um domínio enzimática ligado (BRUNTON et al., 2006).

Diversas funções são desempenhadas por esses receptores em vários processos fisiológicos. Ao contrário dos receptores acoplados à proteína G, que atravessam sete vezes a membrana, estes receptores o fazem uma única vez. Vários receptores com do-

Figura 13 – Receptor nicotínico de acetilcolina regulado por ligante.



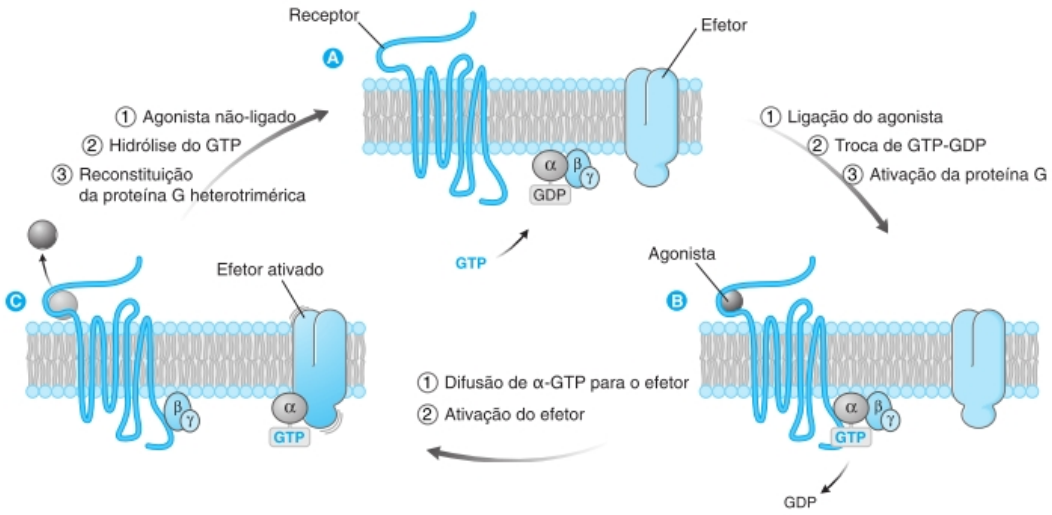
Fonte: (GOLAN et al., 2009)

mínios citosólicos enzimáticas foram dímeros ou complexos de múltiplas subunidades para transdução de seus sinais (BRUNTON et al., 2006).

A quarta classe de alvos importantes para fármacos é a dos receptores intracelulares, especialmente as enzimas, que constituem um alvo citosólico comum, e muitos medicamentos que são direcionados para enzimas intracelulares produzem suas respostas ao alterar a produção enzimática de moléculas sinalizadoras ou metabólicas críticas (BRUNTON et al., 2006).

A quinta classe de alvos importantes para fármacos é a das enzimas extracelulares, cujos sítios ativos estão localizados fora da membrana plasmática. O ambiente fora das células é formado por um meio de proteínas e moléculas de sinalização. Muitas dessas proteínas desempenham papel estrutural, porém, outras são usadas na comunicação

Figura 14 – Processo de interação fármaco-receptor envolvendo proteínas G.



Fonte: (GOLAN et al., 2009)

da informação entre células, consequentemente as enzimas que modificam as moléculas que intermedeiam esses sinais importantes podem influenciar processos fisiológicos, como a vasoconstrição e a neurotransmissão (GOLAN et al., 2009).

A sexta, e última, classe de alvos importantes para fármacos é a dos receptores de adesão da superfície celular. Frequentemente as células interagem entre si para a realização de funções específicas ou a comunicação de informações. A região de contato entre duas células é denominada adesão, e as interações de adesão entre células são intermediadas por pares de receptores de adesão sobre a superfície de cada célula. Muitos receptores de adesão envolvidos na resposta inflamatória são alvos interessantes para inibidores seletivos (GOLAN et al., 2009).

Após termos revisado alguns dos principais tópicos da farmacodinâmica, como quais são os principais tipos de receptores que intermedeiam a interação fármaco-receptor, quais as classificações dos fármacos de acordo com sua ação sobre a conformação desses receptores e as principais métricas utilizadas para analisar essa interação em um organismo como um todo, podemos observar que o estudo da farmacodinâmica é essencial para auxiliar a maneira como os medicamentos são administrados e como é realizado o monitoramento das respostas terapêuticas proporcionadas pelo mesmo.

Segundo Golan et al. (2009), "a janela terapêutica é a faixa de doses (concentrações) de um fármaco que produz uma resposta terapêutica, sem efeitos adversos inaceitáveis (toxicidade), numa população de pacientes". A quantificação da janela terapêutica pode ser realizada pelo índice terapêutico (IT), que costuma ser definido como na equação (8), onde TD50 é a dose do fármaco que produz uma resposta tóxica em 50% da população, e ED50 é a dose do fármaco terapeuticamente efetiva em 50% da

população. O IT fornece um número que quantifica a margem de segurança relativa de um fármaco numa população (GOLAN et al., 2009).

$$\text{Índice Terapêutico (IT)} = \frac{TD_{50}}{ED_{50}} \tag{8}$$

O conhecimento prático da farmacodinâmica é essencial em todos os casos em que se efetua uma comparação entre fármacos com base na sua potência ou eficácia, além de também ser necessário em casos que se necessita estabelecer a dose apropriada de um medicamento para um paciente específico. Na próxima seção revisaremos os principais tópicos referente à farmacocinética, que é a sub-área da farmacologia responsável por estudar as principais propriedades responsáveis pela efetividade clínica dos fármacos: Absorção, Distribuição, Metabolismo e Excreção.

2.3 Farmacocinética

Uma molécula com boa interação fármaco-receptor não é o suficiente para se obter bons resultados nos estudos clínicos realizados antes de um medicamento ser lançado no mercado. O medicamento deve atingir o órgão-alvo em uma concentração suficiente para gerar a resposta terapêutica desejada (GOLA et al., 2006).

Grande parte das características que tornam o corpo humano resistente à danos causados por invasores e substâncias tóxicas também limitam a capacidade dos medicamentos modernos de combater doenças apresentadas pelos pacientes. Um medicamento eficaz deve ser capaz de superar as barreiras fisiológicas existentes no corpo a fim gerar a resposta desejada sobre determinada patologia, para isso é essencial que este apresente boas taxas de absorção, distribuição, metabolismo e excreção. A absorção pode ocorrer por meio de vários mecanismos, que foram desenvolvidos para explorar ou romper essas barreiras fisiológicas (GOLAN et al., 2009).

Após ser absorvido, os sistemas de distribuição do organismo, como vasos sanguíneos e alvos linfáticos, conduzem o medicamento até o seu órgão-alvo, porém, o medicamento também pode ter o seu acesso limitado por processo que ocorrem no paciente, durante a fase de transporte. Esses processos normalmente são divididos em duas categorias: o metabolismo, em que o organismo inativa o fármaco por meio da degradação enzimática, que ocorre primariamente no fígado, e a excreção, em que o medicamento é eliminado do corpo, principalmente pelos rins, fígado e fezes. As propriedades de absorção, distribuição, metabolismo e excreção são comumente abreviadas com a sigla ADME na literatura (GOLAN et al., 2009).

As etapas de absorção e distribuição são as duas primeiras etapas a serem realizadas pelo fármaco após a administração, estas etapas são influenciadas diretamente pelas

barreiras fisiológicas presentes no corpo. A seguir abordaremos algumas das principais barreiras enfrentadas pelos medicamentos.

Todas as células humanas são delimitadas por uma membrana com dupla camada lipídica. A membrana é constituída por uma superfície hidrofílica, que fica em contato com os ambientes extracelular e intracelular aquosos, além de conter um cerne de lipídios hidrofóbicos (BRUNTON et al., 2006).

Moléculas anfifílicas como fosfolipídios, colesterol e outras de espécies de menor importância são os principais componentes lipídicos que constituem as membranas biológicas. Os grupos hidrofílicos que contém fosfato da cabeça dos fosfolipídios e os grupos hidroxila polares do colesterol são expostos às superfícies externa e interna da membrana, ao passo que as caudas hidrofóbicas dos lipídios estão voltadas para o interior da membrana. Além de serem constituídas por componentes lipídicos, as membranas biológicas contêm numerosas proteínas, algumas apenas expostas na superfície extracelular ou intracelular, outras, denominadas proteínas transmembranas, penetram na dupla cama lipídica e ficam expostas a ambas as superfícies da membrana, desempenhando um importante papel na terapia farmacológica. Para que um fármaco consiga agir sobre alvos intracelulares ou permear uma célula, deve ser capaz de atravessar pelo menos uma e, em geral, várias membranas biológicas (BRUNTON et al., 2006).

Algumas proteínas transmembrana relacionadas à superfamília do carregador ligado à solutos (SLC, *Solute Linked Carrier*) humano, que engloba 43 famílias de proteínas, como a família do transportador de ânions orgânicos (OAT, *Organic Anion Transporter*) e a família do transportador de cátions orgânicos (OCT, *Organic Cation Transporter*), permitem a passagem de medicamentos e moléculas polares através da membrana (GOLAN et al., 2009).

Após a ligação do fármaco à superfície extracelular da proteína, a mesma sofre uma alteração em sua conformação, que é denominada difusão facilitada quando não depende de energia, ou transporte ativo, quando existe a demanda de energia, essa alteração na conformação possibilita que o medicamento tenha acesso ao interior da célula. De forma alternativa, alguns medicamentos se ligam à receptores específicos na superfície celular e deflagram um processo denominado endocitose, onde a membrana celular envolve a molécula para formar uma cavidade fechada ou vesícula, a partir da qual o medicamento é liberado (GOLAN et al., 2009).

De acordo com Golan et al. (2009), na ausência de outros fatores, um fármaco irá penetrar em uma célula até que as concentrações intracelular e extracelular sejam equivalentes. A velocidade de difusão depende do gradiente de concentração do medicamento através da membrana e da espessura desta última. De acordo com a lei de Fick,

o fluxo efetivo de um medicamento através da membrana é o seguinte:

$$\text{Fluxo} = \frac{(C1 - C2) \text{ x Área x Permeabilidade}}{\text{Espessura}_{membrana}} \tag{9}$$

os coeficientes $C1$ e $C2$ representam as concentrações intracelular e extracelular do fármaco. Essa definição se aplica a uma situação em que não há fatores complicantes, como gradientes iônicos, de pH e de cargas através da membrana, entretanto, *in vivo*, esses fatores adicionais afetam a tendência de um fármaco a penetrar nas células. Como exemplo, podemos citar a situação em que em que uma maior concentração do fármaco está presente fora da célula, isso normalmente tende a favorecer a entrada efetiva do fármaco na célula, porém, se tanto o interior da célula quanto o fármaco possuírem cargas negativas, é possível que a sua entrada na célula seja impedida. (GOLAN et al., 2009).

A difusão de fármacos, ácidos e básicos, por meio das membranas também pode ser afetada por um fenômeno associado à carga, conhecido como sequestro de pH. O grau de sequestro de um medicamento em um dos lados da membrana é determinado pela constante de dissociação de ácido (pk_a) do fármaco e pelo gradiente de pH através da membrana. Em termos quantitativos, a pK_a de um medicamento representa o valor de pH em que metade do fármaco encontra-se em sua forma iônica (GOLAN et al., 2009).

A equação Henderson-Hasselbalch descreve a relação entre a pK_a de um fármaco A, ácido ou básico, e o pH do meio que contém este medicamento:

$$pK_a = pH + \log \frac{[HA]}{[A^-]} \tag{10}$$

onde o coeficiente HA é a forma protonada do fármaco A.

A barreira hematoencefálica se encontra no sistema nervoso central (SNC) e representa um desafio especial para a terapia farmacológica, ao contrário da maioria das outras regiões anatômicas, visto que o SNC está particularmente bem isolado de substâncias estranhas (HOLLINGER, 2007).

Essa barreira faz uso de junções especializadas para impedir a difusão passiva da maioria dos fármacos da circulação sistêmica para a circulação cerebral. Consequentemente, os medicamentos destinados a atuar no SNC devem ser pequenos e hidrofóbicos para atravessar as membranas biológicas, caso contrário, devem utilizar as proteínas de transporte presentes na barreira hematoencefálica para penetrar nas estruturas centrais. Os fármacos hidrofílicos que não conseguem se acoplar a proteínas de transporte facilitado ou ativo na barreira hematoencefálica, são incapazes de penetrar no sistemas nervoso central. (HOLLINGER, 2007).

Além das membranas celulares e da barreira hematoencefálica, barreiras como a depuração mucociliar na traqueia, a secreção de lisozima dos ductos lacrimais, o ácido no estômago e a base do duodeno são mecanismos de defesa não específicos que precisam ser superados ou a quantidade do medicamento disponível para o órgão alvo, designada como biodisponibilidade do medicamento, nunca será alta o suficiente para que o mesmo seja eficaz (BRUNTON et al., 2006).

Segundo Brunton et al. (2006), a via de administração do medicamento, a sua forma química e alguns fatores específicos do paciente, como transportadores e enzimas gastrintestinais e hepáticas, influenciam na determinação da biodisponibilidade de um fármaco.

Podemos definir a biodisponibilidade de maneira quantitativa da seguinte forma:

$$\text{Biodisponibilidade} = \frac{\text{Quantidade de fármaco que alcança a circulação sistêmica}}{\text{Quantidade de fármaco administrado}} \tag{11}$$

Essa definição de biodisponibilidade se baseia no fato de que a maioria dos medicamentos alcançam os seus locais de ação diretamente a partir da circulação sistêmica. Os fármacos que são administrados de maneira intravenosa são injetados diretamente na circulação sistêmica, nesse caso a quantidade administrada equivale à quantidade que alcança a circulação sistêmica, portanto, sua biodisponibilidade é igual a 1,0 (HOLLINGER, 2007).

Todo medicamento em desenvolvimento é planejado e testado em uma forma posológica que é administrada por uma via específica, isso influencia diretamente na taxa de absorção do mesmo, ou seja, na capacidade de atravessar as barreiras apresentadas pelo corpo. As principais vias de administração de fármacos são: 1) enteral, 2) parenteral, 3) membrana mucosa e 4) transdérmica (GOLAN et al., 2009).

A administração enteral de um medicamento, ou por via oral, representa a mais simples das vias de administração de medicamentos. A via oral explora os pontos fracos nas barreiras de defesa humanas, porém, expõe o fármaco a ambientes rigorosos passíveis a limitar a sua absorção. Essa via oferece muitas vantagens ao paciente, visto que é fácil, conveniente e tem menos tendência a introduzir infecções sistêmicas se comparada a outros métodos (HOLLINGER, 2007).

O medicamento administrado pela via oral deve permanecer estável durante a sua absorção pelo epitélio do trato gastrintestinal, apesar das junções das células epiteliais gastrintestinais dificultarem o transporte paracelular através do mesmo. De maneira geral, os fármacos hidrofóbicos e neutros permeiam membranas celulares mais eficientemente do que fármacos hidrofílicos ou com carga elétrica, a não ser que a membrana contenha uma molécula carreadora que facilite a passagem das substâncias

hidrofílicas (GOLAN et al., 2009).

Após atravessar o epitélio gastrintestinal, o sistema porta transporta os medicamentos até o fígado antes de passar para a circulação sistêmica. O sistema porta pode complicar o processo de passagem dos medicamentos para a circulação sistêmica, visto que ele tem como função proteger o corpo dos efeitos sistêmicos de toxinas ingeridas as entregando para o fígado realizar a desintoxicação, porém, isso pode afetar também os medicamentos administrados (HOLLINGER, 2007).

Todos os fármacos administrados por via oral estão sujeitos ao metabolismo de primeira passagem, onde as enzimas hepáticas, presentes no fígado, podem inativar uma fração do medicamento ingerido. Qualquer fármaco que sofra metabolismo de primeira passagem significativo precisa ser administrado em quantidade suficiente para que haja uma concentração efetiva de fármaco ativo na circulação sistêmica, a partir da qual é possível alcançar o órgão-alvo. As vias não-enterais de administração de medicamentos não estão sujeitas ao metabolismo de primeira passagem (HOLLINGER, 2007).

A administração parenteral consiste na introdução direta de um medicamento através das barreiras de defesa do corpo na circulação sistêmica ou em algum outra espaço tecidual, isso que o medicamento supere imediatamente as barreiras que limitam a eficiência encontradas quando a administração é realizada pela via oral. Nessa forma de administração, os medicamentos podem ser injetados no tecido vascularizado, diretamente no sangue ou no líquido cefalorraquidiano. Quando a administração é tecidual a velocidade de início de ação do medicamento varia de acordo com o fluxo sanguíneo do tecido (BRUNTON et al., 2006).

Essa via de administração possui um maior potencial de risco de infecção, além de ter necessidade de ser administrada por um profissional de saúde. A velocidade de início da ação dos fármacos administrados por essa via é, normalmente, rápida, o que resulta num aumento potencial da toxicidade caso o fármaco seja administrado com muita rapidez ou em doses incorretas (BRUNTON et al., 2006).

As membranas mucosas consistem em uma via de administração que pode proporcionar potencialmente uma rápida absorção, baixa incidência de infecção e conveniência em sua administração, além de evitar o metabolismo de primeira passagem. Os epitélios sublinguais, oculares, pulmonares, nasais, retais e urinários podem ser utilizados para a administração de medicamentos na forma de gotas, comprimidos de rápida dissolução, aerossóis e supositórios, dentre outras formas posológicas. Isso se deve ao fato das mucosas serem muito vasculares, permitindo ao fármaco penetrar rapidamente na circulação sistêmica e alcançar seu órgão-alvo em um tempo menor (GOLAN et al., 2009).

Os fármacos administrados através da via transcutânea são absorvidos pela pele e pelos tecidos subcutâneos, diretamente no sangue. Essa via de administração é adequada para um medicamento que precisa ser administrado lentamente e de forma contínua por um longo período de tempo. Não existem riscos associados de infecção, além de ser uma via simples e conveniente (GOLAN et al., 2009).

Apesar da absorção constituir um pré-requisito para um medicamento atingir níveis plasmáticos adequados, ele também precisa alcançar seu órgão alvo em uma concentração terapêutica a fim de exercer o efeito desejado (BRUNTON et al., 2006).

A distribuição de um medicamento ocorre primariamente por meio do sistema circulatório. Após ser absorvido na circulação sistêmica, o fármaco é capaz de alcançar qualquer órgão-alvo, com uma possível exceção de compartimentos como o cérebro e os testículos, que são os chamados "compartimentos santuários"(BRUNTON et al., 2006).

A concentração de fármaco no plasma é utilizada frequentemente na definição dos níveis terapêuticos e na monitoração deste, visto que é difícil medir a quantidade que é realmente captada pelo órgão-alvo. Porém, em alguns casos essa concentração pode representar uma medida relativamente precária de sua verdadeira concentração tecidual, porém, na grande maioria dos casos o efeito do medicamento no tecido alvo se correlaciona bem com sua concentração plasmática (BRUNTON et al., 2006).

A capacidade dos órgãos e tecidos de captar diferentes fármacos varia acentuadamente, bem como a sua proporção de fluxo sanguíneo sistêmico. Esses fatores cinéticos determinam a quantidade de medicamento que necessita ser administrada para atingir a concentração desejada do fármaco no compartimento vascular. A capacidade dos tecidos não-vasculares e das proteínas plasmáticas de captar e/ou ligar-se ao medicamento contribui para a complexidade dos esquemas de dosagem e também deve ser considerada para se alcançar níveis terapêuticos do fármaco (BRUNTON et al., 2006).

De acordo com Golan et al. (2009), o volume de distribuição de um medicamento (V_d) representa o volume líquido necessário para conter a quantidade total do medicamento absorvido no corpo numa concentração uniforme, equivalente à do plasma no estado de equilíbrio dinâmico:

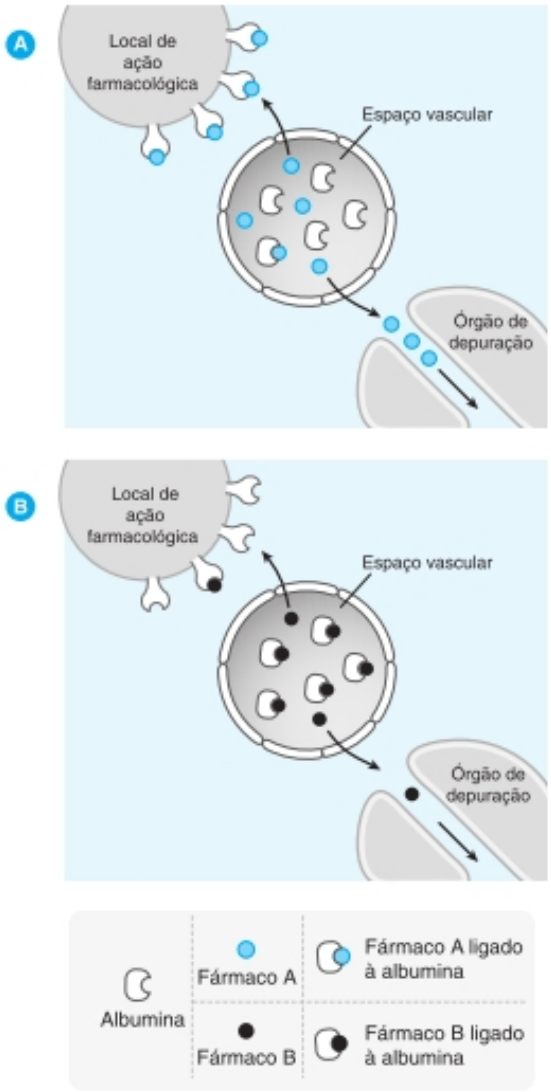
$$V_d = \frac{\text{Dose}}{[\text{Fármaco}]_{\text{plasma}}} \tag{12}$$

Quando o fármaco distribui-se amplamente pelos tecidos corporais a proporção captada pelo corpo é maior, consequentemente o volume de distribuição é relativamente baixo para fármacos que são retidos no compartimento vascular e relativamente alto para os que sofrem ampla distribuição no músculo, tecido adiposo e outros compartimentos não-vasculares (HOLLINGER, 2007).

A capacidade do medicamento ligar-se mais facilmente aos músculos e ao tecido adiposo aumenta a tendência desse fármaco de sofre a difusão do sangue para compartimentos não-vasculares, porém, essa tendência pode ser contrabalanceada, em certo grau, pela ligação do fármaco às proteínas plasmáticas, dentre estas, a albumina constitui a proteínas plasmática mais abundante, sendo uma das principais responsáveis pela ligação dos medicamentos (HOLLINGER, 2007).

Essa ligação tende a reduzir a disponibilidade de um medicamento para difusão ou transporte no órgão alvo, visto que, em geral, apenas a forma não-ligada do medicamento é capaz de difundir-se através das membranas, como podemos observar na Figura 15 (GOLAN et al., 2009).

Figura 15 – Ligação às proteínas plasmáticas e sequestro do fármaco.



Fonte: (GOLAN et al., 2009)

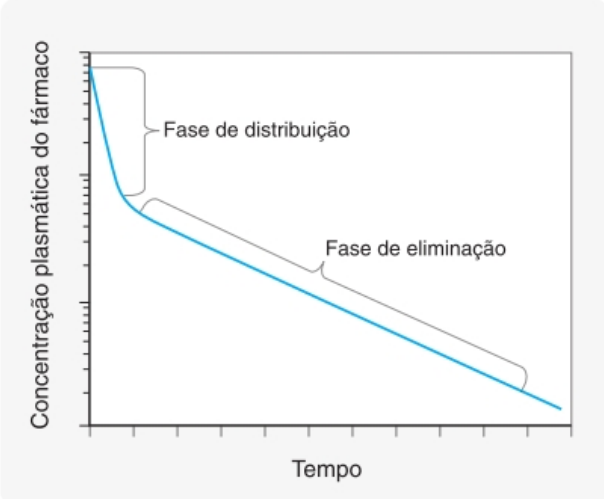
A ligação às proteínas plasmáticas também pode reduzir o transporte dos medicamentos em compartimentos não-vasculares, como o tecido adiposo e o músculo.

Essa ligação também pode ser um importante mecanismo em algumas interações entre medicamentos. A co-administração de dois ou mais medicamentos, em que todos se ligam altamente às proteínas plasmáticas, pode resultar numa concentração plasmática maior do que a esperada de pelo menos um dos fármacos, que pode causar efeitos secundários tóxicos, sendo necessário ajustar o esquema de dosagem dos fármacos (GOLAN et al., 2009).

Grande parte dos medicamentos presentes na circulação sistêmica se distribui rapidamente para outros compartimentos do corpo. No caso de fármacos administrados por meio de injeção intravenosa direta, essa fase de distribuição causa uma diminuição na concentração plasmática do fármaco pouco depois de sua administração e, até mesmo quando este está equilibrado entre seus reservatórios teciduais, a concentração plasmática continua declinando, devido à sua eliminação do corpo (GOLAN et al., 2009).

Durante a fase de eliminação a velocidade de declínio é mais lenta que durante a fase de distribuição, visto que durante a eliminação, o "reservatório"de fármaco nos tecidos pode se difundir novamente para o sangue com intuito de substituir o fármaco eliminado, como podemos observar nas Figuras 16 e 17 (GOLAN et al., 2009).

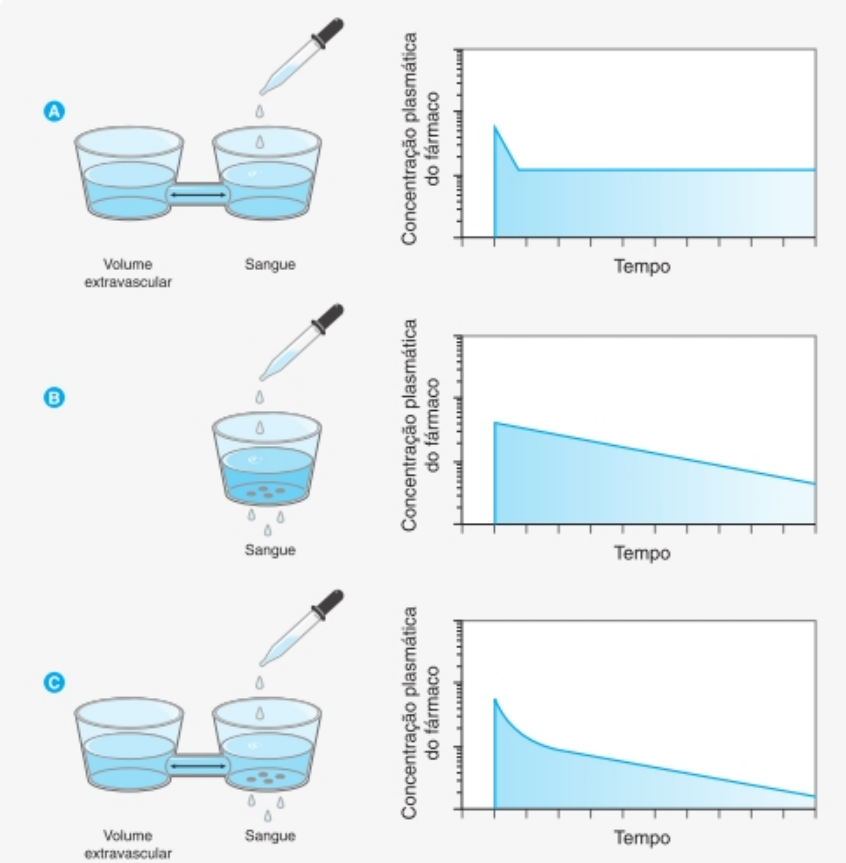
Figura 16 – Distribuição e eliminação dos fármacos após administração intravenosa.



Fonte: (GOLAN et al., 2009)

A tendência de um medicamento ser retido pelo tecido adiposo e/ou tecido muscular durante a fase de distribuição pode gerar um conjunto de equilíbrios dinâmicos entre as concentrações nos diversos compartimentos corporais. O compartimento altamente vascularizado é o primeiro onde a concentração do fármaco aumenta, devido ao elevado fluxo sanguíneo, que favorece a entrada do medicamento. A capacidade de um compartimento em captar um medicamento e a taxa de fluxo sanguíneo nesse compartimento também afetam a taxa de saída do fármaco. Os fármacos tendem a

Figura 17 – Modelo esquemático de distribuição e eliminação de fármacos.



Fonte: (GOLAN et al., 2009)

sair primeiramente do compartimento altamente vascularizado, em seguida, do tecido muscular e, por fim, do tecido adiposo (HOLLINGER, 2007).

Vários órgãos tem a capacidade de metabolizar, em certo grau, os medicamentos por meio de reações enzimáticas. Os rins, o trato gastrintestinal, os pulmões, a pele e outros órgãos contribuem na metabolização sistêmica dos medicamentos, porém, o fígado é o que contém maior diversidade e quantidade de enzimas metabólicas, sendo que a maior parte do metabolismo dos fármacos ocorre nesse órgão (BRUNTON et al., 2006).

A capacidade do fígado de modificar os fármacos varia de acordo com a quantidade de fármaco que penetra nos hepatócitos. Este órgão metaboliza preferencialmente os fármacos hidrofóbicos, porém, devido à numerosa quantidade de transportadores da família do SNC, que também permite a entrada de alguns fármacos hidrofílicos nos hepatócitos, que também podem passar pelo processo de metabolização (BRUNTON et al., 2006).

As enzimas hepáticas têm a propriedade de modificar quimicamente uma variedade de substituintes nas moléculas dos medicamentos, inativando-os ou facilitando

sua eliminação. Essas modificações são chamadas de biotransformação. As reações de biotransformação são classificadas em dois tipos: as reações de oxidação / redução, ou reações de Fase I, e as reações de conjugação / hidrólise, ou reações de Fase II (BRUNTON et al., 2006).

As reações de oxidação / redução modificam a estrutura química de um medicamento através da oxidação ou redução, sendo que as principais enzimas que intermedeiam as reações oxidativas são às pertencentes à família CYP450. Alguns fármacos utilizam a estratégia pró-fármaco, onde são administrados em sua forma inativa, de modo a serem alterados metabolicamente à forma ativa por meio de reações de oxidação / redução no fígado. Essa estratégia normalmente é utilizada na tentativa de facilitar a biodisponibilidade oral, diminuir a toxicidade gastrintestinal e/ou prolongar a meia-vida de eliminação de um fármaco (HOLLINGER, 2007).

As reações de conjugação / hidrólise inativam ou aumentam a solubilidade e excreção de um fármaco na urina ou na bile. Em certos casos, a hidrólise ou a conjugação podem resultar em ativação metabólica de pró-fármacos (HOLLINGER, 2007).

A excreção renal consiste no mecanismo mais comum de excreção de medicamentos e baseia-se em sua natureza hidrofílica ou em seu metabólito. Um número relativamente pequeno de fármacos é excretado primariamente na bile. Vários fármacos administrados por via oral não são absorvidos totalmente pelo trato gastrintestinal superior e são eliminados por excreção fecal. Além disso, os fármacos podem ser excretados em quantidade mínimas através das vias respiratória e dérmica (HOLLINGER, 2007).

O entendimento dos processos farmacocinéticos como absorção, distribuição, metabolismo e excreção é essencial para que possamos entender a maneira como um fármaco chega em seu órgão alvo em uma concentração terapêutica adequada e quais as barreiras que devem ser superadas para isso, sendo que o estudo e a criação de métodos *in silico* que sejam capazes de modelar essas propriedades computacionalmente têm sido essenciais para a diminuição de falhas nos testes clínicos de fármacos nas últimas décadas.

2.4 Toxicologia

Fármacos podem ser tóxicos para certos pacientes, devido à predisposição genética, ação não-seletiva, uso ou administração inapropriados do fármaco. É preciso reconhecer que não existe nenhuma substância totalmente específica. Todos os fármacos possuem efeitos pretendidos primários e efeitos não-pretendidos secundários; os efeitos não-pretendidos são conhecidos como efeitos colaterais ou efeitos adversos (BRUNTON et al., 2006).

A toxicologia farmacológica enfoca os efeitos prejudiciais, que derivam da ativação ou inibição inapropriadas do alvo pretendido da substância (efeitos adversos direcionados para o alvo) ou de alvos não-pretendidos (efeitos adversos não direcionados para o alvo), de fármacos em animais e no corpo humano (BRUNTON et al., 2006).

Um conceito importante na toxicidade de substâncias é que um efeito adverso pode representar um exagero da ação farmacológica desejada, devido a alterações na exposição à substância. Isso pode ocorrer através de um erro deliberado ou acidental de dose, alterações na farmacocinética da substância (por exemplo, devido a doença hepática ou renal ou a interações com outras substâncias) e alterações na farmacodinâmica da interação substância-receptor, alterando a resposta farmacológica (por exemplo, mudanças no número de receptores). Todas essas alterações podem levar a um aumento na concentração efetiva da substância e, portanto, a um aumento da resposta biológica (HOLLINGER, 2007).

Uma importante classe de efeitos adversos sobre o alvo pode ocorrer em consequência da interação do fármaco ou de um de seus metabólitos com o receptor apropriado, porém, no tecido incorreto. Muitos alvos de fármacos são expressos em mais de um tipo celular ou tecido. Algumas vezes, os efeitos colaterais sobre o alvo revelam funções importantes e previamente desconhecidas do alvo biológico (HOLLINGER, 2007).

Os efeitos adversos não relacionados ao alvo ou não-pretendidos ocorrem quando o fármaco interage com alvos não-pretendidos. Um exemplo de efeito não-pretendido é fornecido pelo anti-histamínico terfenadina, que inibe um canal de potássio cardíaco (hERG). Infelizmente, a inibição não pretendida do canal iônico levou a arritmias cardíacas fatais em alguns pacientes, e em consequência, a terfenadina foi retirada do mercado (GOLAN et al., 2009).

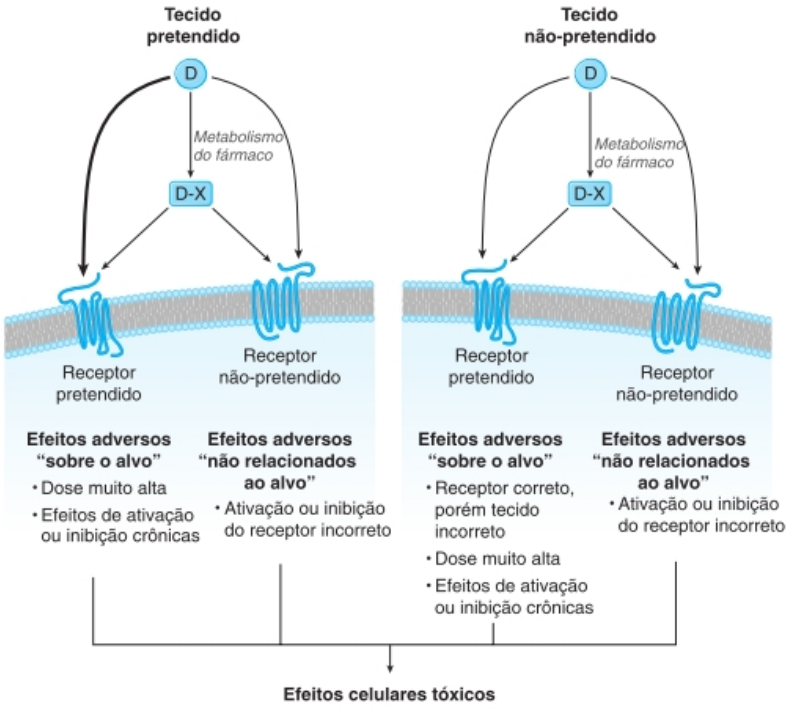
Outro efeito comum não relacionado ao alvo é a ativação não-pretendida de diferentes subtipos de receptores. Por exemplo, o receptor β_1-adrenérgico é expresso no coração, e a sua ativação aumenta a frequência cardíaca e a contratilidade miocárdica. Receptores β_2-adrenérgicos estreitamente relacionados são expressos primariamente nas células musculares lisas das vias respiratórias e na vasculatura, e a ativação desses receptores β_2 leva ao relaxamento do músculo liso e dilatação desses tecidos (GOLAN et al., 2009).

Os usos clínicos dos antagonistas dos receptores β-adrenérgicos (os denominadas β-bloqueadores) são frequentemente direcionados para o receptor β_1, a fim de controlar a frequencia cardíaca e reduzir a demanda de oxigênio do miocárdio em pacientes com angina ou com insuficiência cardíaca. Entretanto, alguns antagonistas dos receptores β_1 não são totalmente seletivos para o receptor β_1 e também podem antagonizar o

receptor β_2. Por conseguinte, os antagonistas dos receptores beta-adrenérgicos com efeitos não-seletivos estão contra-indicados para pacientes com asma, visto que esses fármacos têm a capacidade de causar inadvertidamente obstrução das vias respiratórias através do antagonismo dos receptores β_2 (GOLAN et al., 2009).

Na Figura 18 podemos observar um esquema com os efeitos adversos dos fármacos sobre o alvo e não relacionados ao alvo.

Figura 18 – Efeitos adversos dos fármacos sobre o alvo e não relacionados ao alvo.



Fonte: (GOLAN et al., 2009)

As interações farmacocinéticas entre fármacos surgem quando um fármaco modifica a absorção, a distribuição, o metabolismo ou a excreção de outro fármaco, alterando, assim, a concentração do fármaco ativo no organismo (BRUNTON et al., 2006).

Os fármacos podem inibir ou induzir as enzimas hepáticas do citocromo P450. Quando dois fármacos são metabolizados pela mesma enzima P450, a inibição competitiva ou irreversível dessa enzima P450 por um fármaco pode levar a um aumento na concentração plasmática do segundo fármaco. Por outro lado, a indução de uma enzima P450 específica por um fármaco pode levar a uma redução nas concentrações plasmáticas dos outros fármacos que são metabolizados pela mesma enzima (BRUNTON et al., 2006).

Algumas vezes, uma interação farmacocinética pode ser desejável. Assim, por

exemplo, como a penicilina é depurada através de secreção tubular nos rins, a meia-vida de eliminação desse fármaco pode aumentar se for administrado concomitantemente com probenecid, um inibidor do transporte tubular renal (GOLAN et al., 2009).

Um segundo exemplo é fornecido pela combinação de imipeném, um antibiótico de amplo espectro, com a cilastatina, um inibidor seletivo de uma dipeptidase da borda em escova renal (desidropeptidase I). Como o imipeném é rapidamente inativado pela desidropeptidase I, a co-administração de imipeném com cilastatina é necessária para produzir concentrações plasmáticas terapêuticas do antibiótico (BRUNTON et al., 2006).

Um fármaco que se liga às proteínas plasmáticas, como a albumina, pode deslocar um segundo fármaco da mesma proteína, aumentando a sua concentração plasmática livre e, consequentemente, a sua biodisponibilidade para tecidos-alvo e não-alvo. Esse efeito pode ser intensificado em uma situação em que os níveis circulantes de albumina estão baixos, como na insuficiência hepática ou desnutrição (síntese diminuída de albumina) ou na síndrome nefrótica (excreção aumentada de albumina) (GOLAN et al., 2009).

Surgem interações farmacodinâmicas quando um fármaco modifica a resposta dos tecidos-alvo ou não-alvo a outro fármaco. Podem ocorrer interações farmacodinâmicas tóxicas quando dois fármacos ativam vias complementares, resultando em efeito biológico exagerado (GOLAN et al., 2009).

A segurança e a eficácia de um fármaco também podem ser alteradas pela co-exposição a vários produtos não-farmacêuticos, como alimentos, bebidas, ervas e outros suplementos dietéticos. Muitos produtos herbáceos consistem em misturas complexas de compostos biologicamente ativos, e a sua segurança e eficiência raramente foram testadas em estudos controlados. O largo uso de produtos herbáceos não regulamentados entre o público deve levar o médico a investigar o uso desses produtos pelo paciente (HOLLINGER, 2007).

A literatura contém diversos relatos de falha terapêutica de fármacos utilizados juntamente com produtos herbáceos, bem como alguns relatos de toxicidade. Por exemplo, a preparação ginkgo biloba (da árvore de mesmo nome) inibe a agregação plaquetária. O uso simultâneo de ginkgo e de antiinflamatório não-esteroides, que também inibem a agregação plaquetária, pode aumentar o risco de sangramento (GOLAN et al., 2009).

O entendimento em torno dos conceitos da toxicologia são essenciais, visto que o estudo e a criação de métodos *in silico* que sejam capazes de predizer atividades tóxicas de fármacos é um dos principais desafios em projetos de descoberta de fármacos atualmente, onde a toxicidade é tem uma participação relevante nas falhas na etapa de testes clínicos.

2.5 Big Data

O termo Big Data foi caracterizado por Laney (2001) como modelo dos 3 V's, que são Volume, Velocidade e Variedade, respectivamente. O Volume refere-se à quantidade de informação em si, a Velocidade refere-se à rapidez em que os dados são gerados e a Variedade refere-se à variedade de tipos de dados disponíveis em determinado domínio.

Os tipos de dados podem ser desmembrados em: 1) estruturados: que possuem um modelo de dados formal e bem definido, 2) semi-estruturados: que possuem um modelo de dados, porém, este é mais flexível se comparado com o modelo de dados anterior, 3) não-estruturados: não possui modelo de dados pré-definido e 4) misto: é quando dois ou mais tipos de dados apresentados anteriormente estão presentes na massa de dados analisada (ASSUNÇÃO et al., 2015).

A velocidade em que os dados são gerados normalmente é divida em três categorias: 1) *batch*: os dados são gerados em intervalos de tempo bem definidos, 2) *near-time*: os dados são gerados em pequenos intervalos de tempo, 3) *real-time*: os dados são gerados e processados de forma contínua (ASSUNÇÃO et al., 2015).

A problemática do Big Data têm se difundido de maneira crescente em questões de interesse da indústria farmacêutica, como os projetos para descoberta de novos fármacos, uma vez que moléculas candidatas a se tornarem medicamentos são oriundas de várias fontes de dados, como dados da literatura, campanhas internas para identificação de moléculas, dentre outras; além do volume de informações disponíveis aumentar consideravelmente a cada ano (BANERJEE et al., 2016).

A validação de dados obtidos por meio da literatura e outras fontes de dados públicas é crucial no processo de tomada de decisão de um laboratório quando o mesmo considera investir recursos no estudo de um grupo de moléculas candidatas a se tornarem fármacos, o que destaca a importância da reprodutibilidade das metodologias e modelos computacionais utilizados em projetos de descoberta de novos medicamentos (BANERJEE et al., 2016).

Na química, o termo *Big Data* normalmente está relacionado à bases de dados com informações químicas e/ou bioquímicas, ou então, à mineração de novos dados químicos a partir de um repositório de patentes, literatura, dentre outros. Alguns exemplos de bases de dados que armazenam dados referentes a atividade de compostos químicos são: *PubChem*, *ChEMBL*, *BindingDB* e *AdmetSAR*, todas elas são disponíveis para uso do público em geral. O desafio de como trabalhar em torno desse grande volume de dados químicos se tornou um importante problema para o desenvolvimento futuro da indústria química, incluindo o braço farmacêutico, agroquímico, biotecnológico, dentre outras (TETKO et al., 2016).

2.6 Apache Spark

A demanda por novas plataformas que consigam lidar com a problemática do *Big Data* é crescente desde a década passada. De acordo com Zaharia et al. (2010), o modelo *MapReduce* e seus variantes têm obtido sucesso ao lidar com aplicações que necessitam trabalhar com grandes quantidade de dados através de processamento paralelo em *clusters*, garantindo tolerância à falhas, balanceamento de carga e escalabilidade.

Esse modelo de programação gerencia a entrada de dados de forma acíclica, onde os dados são transmitidos através de um conjunto de operadores. Essa estrutura permite que o modelo consiga lidar com falhas sem a intervenção do usuário (ZAHARIA et al., 2010).

Apesar desse modelo de programação ser útil para uma grande quantidade de aplicações, existem algumas que não podem ser implementadas de forma eficiente usando-o, destacando-se as que trabalham de forma iterativa, ou seja, que necessitam reutilizar dados frequentemente. Cada vez que o algoritmo necessitar reutilizar um conjunto de dados, ele dispara um serviço *MapReduce* que lê os dados do disco, causando lentidão nas iterações (ZAHARIA et al., 2010).

O Apache Spark é um *framework* de computação paralela que suporta aplicações iterativas de forma eficiente e oferece escalabilidade, tolerância a falhas e balanceamento de carga, como o *MapReduce*. Para trabalhar com o *framework*, o programador necessita criar um *driver program* que implementa o controle de fluxo da aplicação, ou seja, é responsável por rodar as operações paralelas (ZAHARIA et al., 2010).

As duas abstrações principais para programação paralela oferecidas pelo Spark são: *Resilient Distributed Datasets*, ou simplesmente RDD's, e *Parallel Operations*, que são operações executadas sobre os RDD, normalmente por meio da passagem de uma função no próprio RDD. Além disso, o Spark suporta dois tipos de variáveis compartilhadas, que podem ser utilizadas em funções executas no *cluster* (ZAHARIA et al., 2010).

Um RDD é uma coleção de dados imutável que é criada somente através de operações em dados armazenados na memória permanente ou a partir de outros RDDs. As operações que criam novos RDD's são chamadas de *transformations* com o objetivo de diferenciar-se de outras operações possíveis (ZAHARIA et al., 2012).

Os RDD's não precisam ser computados todas as vezes que são utilizados, na verdade, o Spark gera um fluxograma (chamado internamente de *lineage*) com o passo-a-passo necessário para a construção do RDD, em certas ocasiões, essa é uma propriedade importante, visto que a partir desse recurso é possível reconstruir um RDD caso ocorra alguma falha. Essa característica do RDD de não necessitar ser computado toda vez em que é utilizado é chamada de *lazy evaluation* (ZAHARIA et al., 2012).

A representação de um RDD é baseada em grafos, esta representação possibilita suportar uma grande quantidade de *transformations* sem a necessidade de adicionar uma lógica especial na estrutura de dados para isso. Essa interface baseada em grafos é dividida em cinco informações, que são apresentadas no quadro 3 (ZAHARIA et al., 2012).

Quadro 3 – Interface usada para representar RDD's no Spark.

Operation	**Meaning**
partitions()	Retorna uma lista de objetos *Partition*
preferredLocations(p)	Lista os nós onde a partição p pode ser acessada com maior velocidade devido a localização dos dados.
dependencies()	Retorna uma lista de dependências.
iterator(p, parentIters)	Computa os elementos da partição p a partir dos elementos da partição do RDD base.
partitioner()	Retorna metadados com a informação se é um RDD particionado em *hash / range*.

Fonte: Zaharia et al. (2012)

As dependências entre RDD's são classificadas em dois tipos, dependências diretas (*narrow dependencies*), onde cada partição do RDD base é utilizada no máximo em uma partição no RDD filho e, dependências amplas (*wide dependencies*), onde existe a possibilidade de várias partições do RDD filho podem depender de uma partição no RDD base. A vantagem das dependências diretas é que em caso de uma falha em um processamento, apenas a partição perdida é recalculada, ao contrário das dependências amplas, que em alguns casos pode exigir que uma reexecução completa seja executada (ZAHARIA et al., 2012).

O programador tem controle sobre outros dois aspectos de um RDD, a persistência e o particionamento. O controle sobre a persistência ou não do RDD é interessante, por exemplo, nos casos mencionados dos algoritmos iterativos, onde o programador pode persistir o RDD que irá ser reutilizado na memória, aumentando a velocidade das iterações. O controle sobre o particionamento dos dados pode ser utilizados para otimizações em termos de localização dos dados, diminuindo a necessidade de tráfego na rede (ZAHARIA et al., 2012).

Existem dois tipos de *parallel operations*, as *transformations* e as *actions*, a primeira tem como principal função transformar um RDD, após a transformação ela retorna um novo RDD com os valores correspondentes às alterações realizadas no RDD base, esse tipo de operação não requer que o RDD seja computado. O segundo tipo de *parallel operation* são as *actions*, que normalmente retornam alguma informação para o *driver*

program, salvam dados provenientes de um RDD, dentre outras atividades, esse tipo de operação computa o RDD em questão através do seu fluxograma, ou *lineage*, como mencionado anteriormente (ZAHARIA et al., 2012).

Os tipos de variáveis compartilhadas suportadas pelo Spark são: 1) variáveis de *broadcast*, que distribui um determinado valor para todos os nós envolvidos nos cálculos da aplicação que está em execução, 2) acumuladores, que são variáveis que os nós em que estão sendo realizados podem adicionar um determinado valor, que somente o *driver program* pode ler (ZAHARIA et al., 2010).

De acordo com Zaharia et al. (2012), as tarefas de execução são alocadas pelo gerenciador de acordo com a localidade dos dados que irão ser processados. Se os dados necessários para a execução de uma rotina estiverem armazenados na memória, o gerenciador de tarefas irá transmitir a execução para este nó, dentre outras otimizações que são feitas em tempo de execução.

A ML Lib é a biblioteca de algoritmos de aprendizagem de máquina distribuídos do *framework* Spark, que tem como principal objetivo atender a demanda em casos que se beneficiam do paralelismo de dados ou modelos. A biblioteca consiste em uma coleção de algoritmos clássicos de aprendizagem de máquina, implementados com alta performance e escalabilidade. Algumas das classes de algoritmos de aprendizagem de máquina providos pela biblioteca são: 1) classificação, 2) regressão, 3) filtro colaborativo, 4) agrupamento e 5) redução de dimensionalidade (MENG et al., 2016).

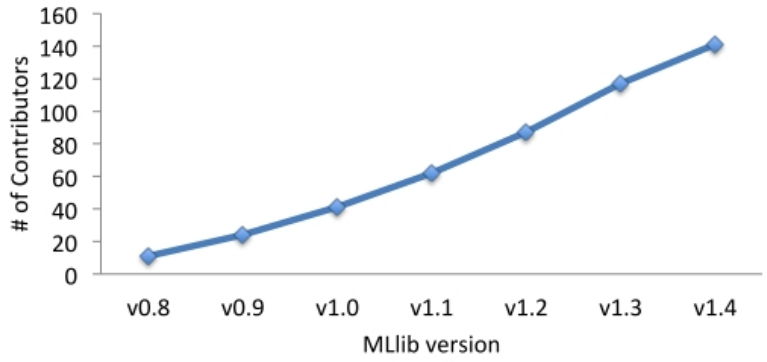
A integração da biblioteca com o Spark trás uma série de benefícios. Uma vez que o *framework* foi projetado visando proporcionar bom desempenho para algoritmos de natureza iterativa, o desenvolvimento de implementações eficientes de algoritmos de aprendizagem de máquina de maneira distribuída é facilitado, além de que melhorias nas estruturas básicas do Spark normalmente se traduz em ganhos de performance nas implementações presentes na ML Lib, sem qualquer mudança direta nos algoritmos da biblioteca (MENG et al., 2016).

Pipelines para o treinamento de modelos que utilizam algoritmos de aprendizagem de máquina como base, nem sempre são fáceis de se implementar devido a complexidade do processo e ao fato de que a maioria dos programas utilizados no mercado não possuem soluções para todos os estágios do processo de construção de um modelo preditivo. Como a biblioteca ML Lib possui uma série de funcionalidades que visam atender a todas as etapas do processo de construção de um modelo, a mesma também provê uma uma API para a construção de *pipelines* que busca atender a essa demanda pela criação de *pipelines* (MENG et al., 2016).

O desenvolvimento da biblioteca começou em meados de 2012, como parte do projeto ML_{BASE}, tendo o código disponibilizado como *opensource* para o público

em setembro de 2013, onde começou a ser empacotada com o *framework* a partir da versão 0.8. O time inicial de desenvolvimento era composto por onze desenvolvedores, menos de dois anos depois, no lançamento da versão 1.4, o time já contava com 140 contribuintes de mais de cinquenta organizações diferentes, indicando a existência de uma grande comunidade por trás do projeto (MENG et al., 2016). Na Figura 19 podemos observar o crescimento na quantidade de contribuintes de acordo com o lançamento das versões.

Figura 19 – Aumento do número de contribuintes do projeto ML Lib.

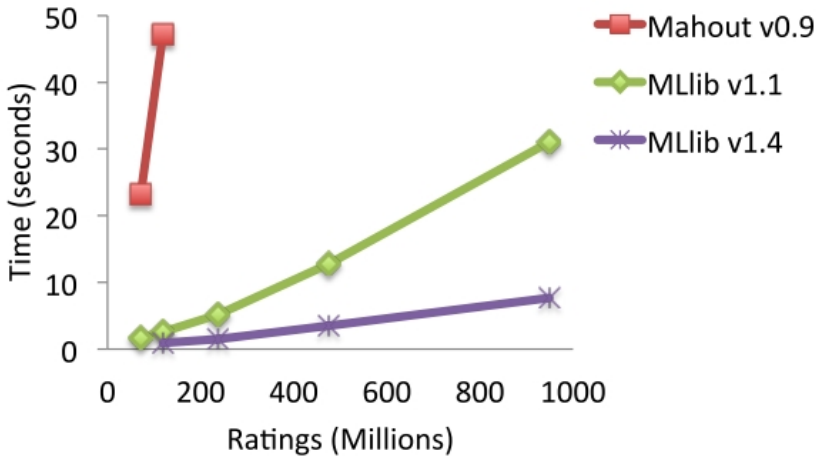


Fonte: (MENG et al., 2016)

A velocidade, escalabilidade e melhoria contínua da biblioteca ML Lib ao longo do tempo, pode ser observada no gráfico da Figura 20, onde apresentamos o resultado um teste de desempenho com o treinamento do algoritmo ALS, que é comumente utilizado em técnicas de filtragem colaborativa. Esse teste foi realizado em um *cluster* EC2 com 16 máquinas m3.2xlarge utilizando a biblioteca ML Lib distribuída com o Apache Spark nas versões 1.1 e 1.4, além da utilização da ferramenta Apache Mahout 0.9, que roda sobre o *Hadoop MapReduce*, para fins comparativos. O treinamento foi realizado utilizando o conjunto de dados *Amazon Reviews* (MCAULEY; LESKOVEC, 2013), que foi duplicado a fim de aumentar a quantidade de dados disponíveis (MENG et al., 2016).

Notamos na Figura 20, que o desempenho da biblioteca ML Lib é superior ao da ferramenta Apache Mahout 0.9. Também notamos que existe melhoria entre as versões do *framework*, nessa ocasião a versão da biblioteca que roda sobre o Spark 1.4 teve uma performance muito melhor que a versão que roda sobre o Spark 1.1, o que nos mostra que o ecossistema em torno do Spark tem tido melhoria contínua com o decorrer do tempo (MENG et al., 2016).

Figura 20 – Desempenho da biblioteca ML Lib.



Fonte: (MENG et al., 2016)

2.7 Árvore de Decisão

Os algoritmos de aprendizagem de máquina voltados para classificação visam identificar classes que possuem algumas características em comum, sendo uteis em uma ampla variedade de aplicações, particularmente nas que envolvem tomada de decisão automatizada (HSSINA et al., 2014).

O algoritmo árvore de decisão representa um dos métodos de classificação mais populares, ele toma como entrada um conjunto de dados, previamente classificados, e emite uma árvore que se assemelha a um diagrama de orientação em que cada nó final é uma classe (HSSINA et al., 2014).

Na área de aprendizado de máquina, grande parte dos estudos são baseados na teoria da informação, para compreendermos como o algoritmo funciona, é necessário definirmos alguns conceitos matemáticos utilizados em teoria da informação.

O primeiro deles, entropia, foi desenvolvido por Claude Shannon, é utilizada como uma métrica de incerteza. De acordo com Hssina et al. (2014), a entropia pode ser definida de acordo com a seguinte equação:

$$Entropia(P) = -\sum_{i=1}^{n} p_i * \log(p_i) \tag{13}$$

O segundo conceito interessante no contexto dos algoritmos de árvore de decisão é o de ganho de informação, que é uma métrica usada para medir a efetividade de um atributo em classificar um conjunto de dados, ou seja, o ganho de informação calcula a redução na incerteza ao selecionar um atributo no conjunto de dados (HSSINA et al.,

2014). Segundo Hssina et al. (2014) o ganho de informação pode ser definido de acordo com a seguinte equação:

$$GI(p,T) = Entropia(p) - \sum_{j=1}^{n}(p_j * Entropia(p_j)) \tag{14}$$

onde os valores (p_j) é o conjunto de todos os possíveis valores para o atributo T. Essa métrica pode ser utilizada para ranquear os atributos e construir uma árvore de decisão onde os atributos com maior ganho de informação ocupam os nós mais próximos do nó raiz ou o próprio nó raiz (HSSINA et al., 2014).

O algoritmo ID3 é um algoritmo de aprendizagem supervisionada que cria uma árvore de decisão a partir de um conjunto de dados pré-definido. A árvore de decisão gerada é então utilizada para classificar futuros exemplares. Esse algoritmo constrói uma árvore de decisão baseado na métrica de ganho de informação dos dados de treinamento, como mencionado anteriormente (HSSINA et al., 2014).

2.8 Ferramentas Computacionais

A seguir apresentaremos as seguintes ferramentas utilizados junto à fonte de dados e ao *framework* Apache Spark em nossa proposta de metodologia reprodutível para predição de propriedades ADMET: 1) Docker, 2) Conda, 3) Jupyter Notebooks, 4) RD Kit, 5) Numpy, 6) Pandas e, 7) scikit-learn.

O *Docker* é uma plataforma que implementa e padroniza a virtualização baseada em *containers*. Essa plataforma é aberta e consiste na *Docker Engine* que é o pacote de ferramentas para criar e rodar *containers* e o *Docker Hub*, um serviço na nuvem que permite compartilhar imagens dos *containers* criados, automatizando o processo de configuração e publicação do ambiente computacional (LIU; ZHAO, 2014).

O *Conda* é um gerenciador de pacotes que auxilia no processo de distribuição de programas. Essa ferramenta permite a criação de ambientes virtuais isolados para a instalação dos programas necessários para a execução da metodologia proposta.

Utilizaremos o *Conda* para instalar os programas *Jupyter Notebooks*, que é um editor de textos que nos permite mesclar texto puro com trechos de código-fonte e disponibilizar o passo-a-passo da nossa metodologia de pesquisa de maneira interativa; RD Kit que é uma biblioteca para se trabalhar com moléculas em seus diversos formatos de entrada, utilizaremos esse programa para calcular os descritores moleculares que iremos utilizar no treinamento dos modelos computacionais; *Numpy* que é uma biblioteca para cálculo numérico na linguagem python; *Pandas* que é uma biblioteca para a manipulação e visualização de estruturas de dados na linguagem python; e o *scikit-learn*

que é uma biblioteca de aprendizado de máquina desenvolvida de forma sequencial, iremos utilizar essa biblioteca para gerar nossos arquivos de dados no formato *libsvm*.

3 Metodologia

Esta seção tem como principal objetivo explanar as etapas necessárias para a obtenção de um modelo computacional reprodutível para a predição de propriedades ADMET a partir de bases de dados que possuam a estrutura 2D das moléculas no formato SMILES, onde estas estejam devidamente classificadas. As seguintes etapas serão abordadas: 1) base de dados, 2) pré-processamento dos dados e 3) treinamento do modelo computacional.

Uma das propriedades mais importantes na definição de um fármaco seguro é a de metabolização de uma molécula candidata a se tornar fármaco, pois essa propriedade influencia na concentração do medicamento no organismo, afetando diretamente a eficácia do fármaco, além de levar à reações tóxicas, caso o medicamento atinja uma concentração maior do que a desejada (MISHRA, 2011).

A família de enzimas CYP450 é considerada uma das mais relevantes no metabolismo do corpo humano. Um dos casos de reações tóxicas decorrentes à metabolização de fármacos ocorre quando duas ou mais drogas co-administradas no tratamento de um paciente competem para serem metabolizadas pela mesma enzima CYP, aumentando a concentração plasmática de um dos fármacos, tornando a predição de inibição de enzimas CYP um problema relevante na definição de um medicamento seguro e nos motivando a trabalhar com esse tema no estudo de caso realizado para a demonstração da metodologia (MISHRA, 2011).

As enzimas CYP1A2, CYP2C9, CYP2C19 e CYP2D6 serão utilizadas na elaboração do estudo de caso, onde aplicaremos nossa proposta de metodologia reprodutível, disponibilizando publicamente os dois elementos apontados por Mesirov (2010) como essenciais para a reprodutibilidade de pesquisas que utilizam simulações computacionais na solução de um problema: 1) ambiente computacional reprodutível e 2) um editor de texto integrado com o ambiente computacional onde é possível executar a metodologia passo-a-passo obtendo o mesmo resultado.

Nesse estudo de caso criaremos doze modelos computacionais, três para cada enzima, onde cada um terá uma distribuição do conjunto de treinamento e conjunto de testes diferente. Esses modelos tem como principal objetivo prever se uma molécula candidata a fármaco inibe ou não a enzima em questão.

Entendemos por modelo, o conjunto de dados utilizado no treinamento e o algoritmo de aprendizagem de máquina utilizado no processo de desenvolvido do modelo. Nesse caso de uso iremos utilizar a base de dados admetSAR e seus respectivos conjunto de dados para cada enzima e o algoritmo árvore de decisão para a predição da

atividade biológica do fármaco sobre a enzima em questão.

3.1 Base de Dados

Nos últimos anos, estudo *in silico* tem recebido grande atenção da comunidade farmacêutica como uma tentativa de aprimorar o processo de descoberta de fármacos. Os modelos são aplicados na seleção de moléculas promissoras e na investigação das propriedades de Absorção, Distribuição, Metabolismo, Excreção e Toxicidade (ADMET), proporcionando mais assertividade na escolha de moléculas para testes clínicos e redução na quantidade de experimentos animais necessários para a definição de medicamentos seguros (BANERJEE et al., 2016).

Órgãos regulamentadores como *Organization for Economic Co-operation and Development*, *Internacional Organization for Standardization (ISO)*, *National Institute of Technology and Evaluation (NITE)*, *United States Food and Drug Administration (US-FDA)*, tem desenvolvido ferramentas computacionais para auxiliar estudos voltados às propriedades ADMET ao longo dos anos (CHENG et al., 2012).

Uma das iniciativas da *US-FDA* é o portal *openFDA*, que disponibiliza dados referentes às moléculas aprovadas pela instituição, como relatórios de efeitos adversos, vias de administração, dosagem adequada, dentre outros (CHENG et al., 2012).

O estudo relacionado a toxicidade de fármacos ganhou destaque a partir do ano de 2010, quando aproximadamente 10-14% das falhas em testes clínicos estavam relacionadas ao tema (HECHT, 2011). Muitas abordagens têm sido empregadas com o intuito de auxiliar pesquisadores na definição de um fármaco seguro; desde metodologias que investigam a relação entre estrutura-atividade de moléculas à algoritmos de aprendizado de máquina, porém, muitas vezes essas abordagens possuem limitações (PRINZ et al., 2011).

Uma das principais limitações no desenvolvimento de modelos computacionais eficientes para a predição de propriedades ADMET é a qualidade dos dados de moléculas com atividade conhecida disponibilizados publicamente, apesar da quantidade de trabalhos relacionados na última decada ser considerável, como por exemplo *FragmentStore*, *SuperToxic*, *SuperTarget*, *FAF-Drugs*, *PK/DB*, dentre outras (CHENG et al., 2012).

O trabalho desenvolvido por Cheng et al. (2012), o banco de dados admetSAR, disponibiliza uma base de dados aberta com a estrutura de aproximadamente 96000 moléculas no formato SMILES, devidamente classificadas em relação aos *endpoints* mais utilizados no estudo de propriedades ADMET, que foram selecionadas e validadas a partir de uma grande coleção de artigos da literatura.

A primeira etapa para a construção da base de dados admetSAR foi a seleção dos artigos que seriam validados. De acordo com Cheng et al. (2012) foi realizada uma pesquisas das publicações no portal *PubMed* (http://www.ncbi.nlm.nih.gov/pubmed) e no *Google Scholar* (http://scholar.google.com/) no período de 2002 à 2011, utilizando palavras relacionadas às propriedades ADMET, além de filtrar trabalhos publicados em jornais de prestígio com intuito com o intuito de restringir a qualidade inicial dos dados que seriam escolhidos papra a validação manual.

Posteriormente, para cada publicação selecionada, foram extraídos os dados e as metodologias validadas manualmente pelos autores, que removeram dados incorretos ou que aumentavam o nível de incerteza das publicações selecionadas. Uma vez que os dados foram validados manualmente pelos autores, a estrutura completa da molécula, em sua representação 2D, foi convertida no formato SMILES canônico, utilizando o programa *OpenBabel v2.3.1* (CHENG et al., 2012).

Após a seleção das publicações, os dados foram baixados das informações de apoio presente em cada artigo, os métodos utilizados em cada artigo foram avaliados e testados por especialistas. De acordo com Cheng et al. (2012), durante esse processo de verificação, os dados que possuíam algum tipo de problema foram removidos. Após a seleção dos dados, a estrutura de cada composto foi baixada no formato SMILES de bases como US-EPA ACToR e DrugBank, depois foram convertidas para SMILES canônicos utilizando o software OpenBabel v2.3.1 (CHENG et al., 2012).

No Quadro 4 podemos observar algumas informações referentes aos dados disponibilizados pelo projeto admetSAR, como *endpoints* comtemplados, quantidade de moléculas disponíveis e se são provenientes de experimentos em alta escala (HTS) (CHENG et al., 2012).

Inúmeras publicações (AMIN et al., 2017; NISHA et al., 2016; SHAIKH; JOSHI, 2016; PARAMASHIVAM et al., 2015; PIRES et al., 2015) têm utilizado o conjunto de dados disponibilizado por Cheng et al. (2012) na construção de modelos computacionais preditivos, que além de possuir alta qualidade, segundo o autor, desde que foi publicado em 2012, o projeto recebe dados adicionais mensalmente. A boa qualidade dos dados disponibilizados e a notável adoção da base de dados admetSAR pela comunidade científica nos motivaram a selecioná-la como fonte de dados para este trabalho.

O processo de obtenção do conjunto de dados admetSAR pode ser realizado acessando o endereço virtual do trabalho (http://lmmd.ecust.edu.cn/admetsar1/). Na seção "Downloads"é necessário preencher um cadastro com informações como nome, e-mail, instituição de ensino a qual se está vinculado, dentre outras. Em seguida, deve-se selecionar os arquivos de interesse para download, que são separados por *endpoints*; Em nosso estudo de caso iremos selecionar os arquivos M_CYP1A2I_I, M_CYP2C9I_I, M_CYP2C19I_I e M_CYP2D6I_I.

Quadro 4 – admetSAR - Informações Adicionais

Endpoints	**Qtd. Moléculas**	**HTS**
aqueous solubility (I)	1708	não
aqueous solubility (II)	46315	sim
human intestinal absorption	578	não
Caco-2 permeability	674	não
blood-brain barrier	1839	não
P-gp substrate	332	não
P-gp inhibitor (I)	1273	não
P-gp inhibitor (II)	1275	sim
CYP1A2 inhibitor	14903	sim
CYP2C9 inhibitor	14709	sim
CYP2C19 inhibitor	14576	sim
CYP2D6 inhibitor	14741	sim
CYP3A4 inhibitor	18561	sim
CYP2C9 substrate	673	não
CYP2D6 substrates	671	não
CYP3A4 substrates	671	não
hERG inhibitor (I)	368	não
hERG inhibitor (II)	806	não
AMES mutagenicity	8445	não
chemical carcinogens	293	não
fathead minnow toxicity	554	não
honey bee toxicity	195	não
tetrahymena pyriformis toxicity	1571	não
rat acute toxicity	10207	não
hepatotoxicity	2154	não
reproductive toxicity	4621	sim
maximum recommended daily dose	1214	não
biodegradation	947	não
bioconcentration factors	916	não

Fonte: Cheng et al. (2012)

Cada arquivo selecionado contém o código de identificação na base de dados *PubChem* na coluna *PubChem ID*; a representação 2D da molécula, no formato SMILES, está na coluna SMILES; e por fim, a coluna *label* se refere à classificação da atividade da molécula, a qual utilizaremos no treinamento dos modelos, quando o valor dessa coluna é 0, o mesmo indica que o composto não inibe a enzima, quando o valor é 1, o mesmo indica que o composto inibe a enzima em questão.

O arquivo M_CYP1A2I_I possuí um total de 14903 moléculas, onde 7415 inibem a enzima e 7488 não inibem. O arquivo M_CYP2C9I_I possuí 14709 moléculas, sendo 4978 inibidoras e 9731 não-inibidoras. O arquivo M_CYP2C19I_I possuí 14756 moléculas, onde 6041 inibidoras e 8535 não-inibidoras. O arquivo M_CYP2D6I_I possuí 14741, onde 3060 são inibidoras e 11681 não-inibidoras. A classificação apresentada na coluna *label*

foi realizada considerando principalmente o valor da métrica AC_{50}, onde moléculas que possuem essa métrica com valor $< 10\mu M$ são classificadas como inibidoras e, moléculas que possuem essa métrica com valor $> 57\mu M$ são classificadas como não-inibidoras (CHENG et al., 2012).

Apesar de termos escolhido o conjunto de dados disponilizados pelo projeto admetSAR para a demonstração da metodologia proposta neste trabalho, a mesma pode ser utilizada em qualquer base de dados que disponibilize a estrutura das moléculas em sua representação 2D no formato SMILES e suas respectivos rótulos, inibe ou não-inibe.

3.2 Pré-Processamento dos Dados

Após a obtenção dos dados, iremos iniciar a etapa de pré-processamento dos dados, checando se a estrutura das moléculas disponibilizadas são válidas e calculando os descritores moleculares que utilizaremos no treinamento do modelo, que serão disponibilizados no formato *libsvm*.

Para reprodução das etapas que serão descritas a seguir, dois passos se fazem necessários: 1) Download do repositório asklepios-dados no *github* (https://goo.gl/Qi7zzU) que contém todos os dados e código-fonte necessários para a obtenção dos modelos. 2) Download da imagem do container com o ambiente computacional necessário para a execução do código-fonte disponibilizado no repositório asklepios-dados.

O programa *git*, necessário para clonar o repositório asklepios-dados, pode ser obtido através deste link (https://git-scm.com/downloads). Após a instalação do programa, basta fazer download do repositório utilizando o comando *git clone https://github.com/leobiscassi/asklepios-dados.git*.

O programa *docker community edition*, necessário para realizar o download da imagem do *container* com o ambiente computacional, pode ser obtido através deste link (https://docs.docker.com/engine/installation/). Após a instalação do programa basta fazer download da imagem através do comando *docker pull leobiscassi/asklepios-dev:0.1* . Essa imagem possui as ferramentas ipython 5.1.0, jupyter 4.2.1, apache spark 2.0.2, pandas 0.19.2, rdkit 2016.09.2, numpy 1.11.2 e scikit learn 0.18, instaladas.

Após o download da imagem, o ambiente computacional pode ser utilizado executando o seguinte comando *docker run -d -p 8880:8888 -v /home/leobiscassi/projects/asklepios-dados/:/home/jovyan/work/ leobiscassi/asklepios-dev:0.1*, onde o caminho */home/leobiscassi/projects/asklepios-dados/* deve ser substituído pelo caminho onde foi realizado o download do repositório asklepios-dados. Depois da execução do comando, basta acessar o endereço http://localhost:8880/tree no navegador e selecionar a pasta notebooks, onde estarão os arquivos com as etapas realizadas para o pré-processamento

dos dados.

A primeira etapa do pré-processamento dos dados, é a leitura das moléculas a partir dos arquivos disponibilizamos pelo website do projeto admetSAR. Após esse processo, realizamos a leitura das estruturas presentes na coluna SMILES dos arquivos com o programa RD Kit e verificamos se a molécula é válida ou não através do resultado da função *isValidaMol()* presente notebook, essa função tenta obter o número de átomos da estrutura em questão, caso não consiga computar, assumimos que a mesma possui algum problema. Após a identificação das estruturas problemáticas, removemos estas do nosso conjunto de dados.

Segundo Todeschini e Consonni (2009), um descritor molecular é o resultado final da aplicação de uma lógica ou procedimento matemático que traduz informações químicas em números. Descritores moleculares normalmente são utilizados como dados de entradas na construção de modelos computacionais que trabalham com predição de atividades de moléculas. Neste trabalho utilizaremos algums descritores moleculares que podem ser calculados através do software RD Kit.

O cálculo dos descritores moleculares é realizado para cada estrutura de molécula válida, caso o valor determinado seja *NaN* podemos concluir que houve um erro no cálculo; caso exista alguma ocorrência de valor *NaN* para um descritor calculado o eliminaremos do nosso conjunto de descritores calculados. Disponibilizaremos os descritores válidos em um arquivo no formato *libsvm*, que será consumido no treinamento do algoritmo de aprendizado de máquina.

3.3 Treinamento do Modelo

De acordo com Mishra (2011), o algoritmo de árvore de decisão é bastante empregado na construção de modelos que trabalham com a predição de inibição de enzimas da família CYP450. Optamos pelo uso deste algoritmo na demonstração da metodologia proposta devido sua popularidade no domínio em questão e facilidade de uso.

Utilizaremos a implementação do algoritmo disponibilizada pela biblioteca de algoritmos de aprendizado de máquina do Apache Spark, a *ML Lib*. A biblioteca *ML Lib* consiste em uma coleção de algoritmos de aprendizado de máquina implementados de forma escalável e aptos a tirar proveito do modelo de programação do *framework* Apache Spark (MENG et al., 2016).

A utilização da biblioteca *ML Lib* proporciona confiabilidade e facilidade na reprodução dos algoritmos, uma vez que esta é desenvolvida e testada por uma grande comunidade de desenvolvedores e testada em diversos domínios de aplicação.

A *ML Lib* consome arquivos de dados no formato *libsvm*, por este motivo geramos um arquivo nesse formato contendo os descritores moleculares calculados anteriormente. Inicialmente, carregamos o arquivo de cada conjunto de dados utilizando o Apache Spark e indexamos os dados para que o *framework* consiga trabalhar com eles.

Posteriormente, dividimos os dados em dois conjuntos, o conjunto de treinamento e o conjunto de teste. Neste caso, utilizamos uma estratégia na divisão dos conjuntos, treinando o algoritmo com os dados divididos em conjuntos com 70% para o treinamento e 30% para teste, 80% para o treinamento e 20% para teste, e por fim, 90% para treinamento e 10% para teste. O que nos permite analisar se a divisão dos conjuntos de treinamento e teste influenciou de forma significativa nos resultados apresentados pelos modelos.

Após a divisão dos dados, realizamos o treinamento dos modelos, validando-os com o conjunto de testes, o que nos possibilitou gerar as métricas para análise do desempenho do classificador, que apresentaremos na seção a seguir.

4 Resultados e Discussão

Um total de doze modelos foram obtidos seguindo a metodologia proposta, um para cada enzima e respectivas divisões de conjunto de treinamento e teste; estratégia utilizada para observar se alguma alteração no desempenho ocorreria.

As métricas escolhidas para a avaliação dos modelos foram: 1) Acurácia e, 2) Área sob a curva ROC. Uma das métricas mais utilizas para avaliar a qualidade de um modelo preditivo é a acurácia, que calcula basicamente a quantidade de predições corretas obtidas pelo modelo, porém, essa métrica não se adequa bem à situações em que um desbalanceamento entre as classes do conjunto de dados é evidente. A acurácia pode ser calculada através da equação 15:

$$\text{acurácia} = \frac{TP + TN}{TP + FP + FN + TN} \tag{15}$$

onde TP (*True Positive*, em português Verdadeiro Positivo), são os valores que são positivos e foram preditos como positivos, TN (*True Negatives*, em português Verdadeiro Negativo), são os valores que são negativos e foram preditos como negativos, FP (*False Positive*, em português Falso Positivo), são os valores negativos que foram preditos como positivos, FN (*False Negative*, em português Falso Negativo), são os valores positivos que foram preditos como negativos.

No quadro 5 podemos observar os valores obtidos através da métrica acurácia para cada enzima e suas respectivas divisões no conjunto de treinamento e teste.

Quadro 5 – Acurácia dos modelos desenvolvidos.

Enzima	**Divisão dos dados**	**Acurácia**
CYP1A2	70% - 30%	0.765443
CYP1A2	80% - 20%	0.766877
CYP1A2	90% - 10%	0.764974
CYP2C9	70% - 30%	0.743800
CYP2C9	80% - 20%	0.762649
CYP2C9	90% - 10%	0.760885
CYP2C19	70% - 30%	0.749319
CYP2C19	80% - 20%	0.746676
CYP2C19	90% - 10%	0.722104
CYP2D6	70% - 30%	0.816355
CYP2D6	80% - 20%	0.817482
CYP2D6	90% - 10%	0.830034

É possível identificar a partir dos dados apresentados no Quadro 5, que os modelos preditivos desenvolvidos para a enzima CYP2D6 apresentaram uma acurácia

maior que os demais, porém, o conjunto de dados referente à enzima CYP2D6 é o que possui o maior desequilíbrio de classes, 3060 exemplares pertencem à classe "inibe" e 11681 exemplares pertencem à classe "não-inibe". Esse desequilíbrio afeta na acurácia do modelo, uma vez que o número de acertos sempre será grande para a classe em evidência.

Em contrapartida, a curva ROC é uma das métricas mais recomendadas para a avaliação de classificadores binários, visto que não é afetada pelo balanceamento de classes, caso ocorra, em um conjunto de dados. A curva ROC basicamente é um gráfico onde plotamos o valor TPR, definida na equação (16), em função da FPR, definida na equação (17).Para realizar a comparação de classificadores é desejável que consigamos reduzir a curva ROC a um número escalar; uma das possíveis formas de se alcançar isso é calculando a área sob a curva ROC.

$$True\ Positive\ Rate\ (TPR) = \frac{True\ Positives}{Total\ Positives} \tag{16}$$

$$False\ Positive\ Rate\ (FPR) = \frac{False\ Positives}{Total\ Negatives} \tag{17}$$

Podemos observar a partir dos dados apresentados no Quadro 5 que a divisão do conjunto de dados não teve uma influência significativa na acurácia dos modelos; o que nos levou a calcular a área sob a curva ROC somente para a divisão 80% - 20% para compararmos com outros trabalhos.

No Quadro 6, podemos observar os valores obtidos através da métrica área sob a curva ROC, para nossos modelos com a divisão de dados 80% - 20% e para os modelos desenvolvidos nos trabalhos admetSAR (CHENG et al., 2012) e pkCMS (PIRES et al., 2015).

Quadro 6 – AUC ROC deste trabalho em comparação com outros trabalhos.

Enzima	**AUC**	**AUC admetSAR**	**AUC pkCMS**
CYP1A2	0.767	0.815	0.876
CYP2C9	0.723	0.802	0.868
CYP2C19	0.729	0.805	0.879
CYP2D6	0.611	0.855	0.843

A interpretação dos valores escalares obtidos através da métrica área sob a curva ROC pode ser feita da seguinte maneira: 1) entre 0.90 e 1.0: modelo excelente, 2) entre 0.80 e 0.90: modelo bom, 3) entre 0.70 e 0.80: razoável, 4) entre 0.60 e 0.70: ruim e 5) entre 0.50 e 0.60: péssimo.

Analisando os resultados obtidos por nossos modelos podemos identificar três modelos com qualidade razoável e um com qualidade ruim. Ao realizar a comparação

com os trabalhos admetSAR (CHENG et al., 2012) e (PIRES et al., 2015), podemos notar a superioridade destes na qualidade de predição.

A utilização de algoritmos mais robustos e complexos do que o algoritmo de árvore de decisão, como *Support Vector Machine* e *LIBSVM Vector Machine Classification* no caso do admetSAR (CHENG et al., 2012); *Random Forest* e *Logistic Regression* no caso do pkCMS (PIRES et al., 2015), além dos descritores moleculares e outras características serem distintas à utilizadas em nossos modelos contribuíram para obtermos um desempenho reduzido.

Apesar disso, de forma geral, obtemos modelos preditivos razoáveis em nosso estudo de caso; o qual tinha como principal objetivo demonstrar nossa proposta de metodologia para o desenvolvimento de modelos computacionais para a predição de propriedades ADMET reprodutíveis e aptos a lidarem com a problemática do *Big Data*.

A principal contribuição deste trabalho para a comunidade acadêmica é a proposição de uma metodologia que faz uso e disponibiliza publicamente os dois elementos apontados como essenciais por Mesirov (2010) para a reprodutibilidade de pesquisas que utilizam simulações computacionais na solução de um problema, uma vez que utilizamos o *Docker* para a automatização do processo de configuração e publicação do ambiente computacional e o editor *Jupyter Notebook* para a publicação de cada uma das etapas realizadas para a obtenção dos modelos de forma integrada ao ambiente computacional disponibilizado.

5 Conclusão

O desenvolvimento de novos medicamentos é uma tarefa complexa, de alto custo e com uma baixa taxa de sucesso. Os principais fatores causadores de falhas nos estágios finais dos projetos de desenvolvimento de novos fármacos estão relacionados às propriedades de Absorção, Distribuição, Metabolismo, Excreção e Toxicidade. Com a popularização de tecnologias que permitem a triagem experimental em alta escala, a quantidade de dados disponíveis para análises têm aumentado com o passar dos anos, permitindo a criação e aperfeiçoamento de técnicas *in silico* e a incorporação destas no processo, aumentando a eficiência na P&D de fármacos.

O benefício da adoção de técnicas *in silico*, principalmente nas etapas iniciais, na P&D de fármacos é inegável. Um exemplo da vantagem da adoção dessas técnicas é a otimização de propriedades ADME nas etapas iniciais do processo, entre a década de 1980 e 1980 aproximadamente 40% das falhas nos testes clínicos eram referentes a taxas baixas de absorção, distribuição, metabolismo e excreção, que caiu para 10-14% a partir do ano de 2010, após a criação e adoção de técnicas *in silico* de baixo custo nas etapas iniciais do processo. A partir de 2010, a toxicidade se tornou um alvo de interesse de estudos *in silico*, visto que esta se tornou uma das principais causas de falhas nos testes clínicos.

Diante da tendência de crescimento no volume e diversidade dos dados disponíveis publicamente, com cada vez mais parcerias entre setor público e privado, propostas de metodologias para a criação de modelos para a predição de propriedades ADMET que estejam aptas a gerar modelos capazes de trabalhar com *Big Data* se tornam interessantes. Este trabalho propõe uma metodologia para a criação de modelos aptos a lidarem com *Big Data*, utilizando o *framework* de computação paralela Apache Spark, que foi projetado especialmente para suportar algoritmos de natureza iterativa, como os algoritmos de aprendizagem de máquina, além de utilizar uma base de dados e o software para cálculo dos descritores moleculares que são disponibilizados publicamente e com código aberto. A adoção de ferramentas de análise exploratória de código aberto, como o projeto *Jupyter Notebook*, e ferramentas de automação da configuração e publicação dos ambientes de trabalho, como o *Docker*, facilitam a reprodutibilidade e a escalabilidade das simulações computacionais realizadas.

5.1 Trabalhos Futuros

Os modelos desenvolvidos no caso de uso apresentaram uma qualidade razoável devido à escolha do algoritmo de aprendizado de máquina e o conjunto de descritores

moleculares utilizados no treinamento. Como trabalho futuro, pretendemos adicionar mais descritores moleculares para o treinamento dos modelos através de programas opensource, além de disponibilizar mais algoritmos de aprendizado de máquina para o desenvolvimento de modelos preditivos e adotar a estratégia de validação cruzada (*cross-validation*) para a divisão de dados e validação do modelo.

Referências

AMIN, S. A.; BHATTACHARYA, P.; BASAK, S.; GAYEN, S.; NANDY, A.; SAHA, A. Pharmacoinformatics study of piperolactam a from piper betle root as new lead for non steroidal anti fertility drug development. **Computational Biology and Chemistry**, Elsevier, 2017.

ASSUNÇÃO, M. D.; CALHEIROS, R. N.; BIANCHI, S.; NETTO, M. A. S.; BUYYA, R. Big Data computing and clouds: Trends and future directions. **J. Parallel Distrib. Comput.**, Elsevier Inc., v. 79-80, p. 3–15, 2015. ISSN 0743-7315. Disponível em: <http://dx.doi.org/10.1016/j.jpdc.2014.08.003>.

BACQ, Z. M. **Fundamentals of biochemical pharmacology**. [S.l.]: Elsevier, 2013.

BANERJEE, P.; SIRAMSHETTY, V. B.; DRWAL, M. N.; PREISSNER, R. Computational methods for prediction of in vitro effects of new chemical structures. **Journal of cheminformatics**, Springer, v. 8, n. 1, p. 51, 2016.

BIESIADA, J.; POROLLO, A.; VELAYUTHAM, P.; KOURIL, M.; MELLER, J. Survey of public domain software for docking simulations and virtual screening. **Human genomics**, v. 5, n. 5, p. 497–505, 2011. ISSN 1479-7364.

BRUNTON, L. L.; LAZO, J.; PARKER, K. **The pharmacological basis of therapeutics. Goodman and Gilmans**. [S.l.]: New York: McGraw-Hill, 2006.

CHENG, F.; LI, W.; ZHOU, Y.; SHEN, J.; WU, Z.; LIU, G.; LEE, P. W.; TANG, Y. AdmetSAR: A comprehensive source and free tool for assessment of chemical ADMET properties. **Journal of Chemical Information and Modeling**, v. 52, n. 11, p. 3099–3105, 2012. ISSN 15499596.

DIMASI, J. A. The Value of Improving the Productivity of the Drug Development. **Original Research Article**, p. 1–10, 2002.

GOLA, J.; OBREZANOVA, O.; CHAMPNESS, E.; SEGALL, M. **ADMET property prediction: The state of the art and current challenges**. 2006. 1172–1180 p.

GOLAN, D.; JUNIOR, T.; ARMEN, H.; ARMSTRONG, E. J.; ARMSTRONG, A. W. et al. **Princípios de farmacologia: a base fisiopatológica da farmacoterapia**. [S.l.]: Guanabara Koogan, 2009.

HECHT, D. Applications of Machine Learning and Computational Intelligence to Drug Discovery and Development. **Drug Development Research**, v. 72, p. 53–65 ST – Applications of Machine Learning and C, 2011. ISSN 0272-4391.

HOLLINGER, M. A. **Introduction to pharmacology**. [S.l.]: CRC Press, 2007.

HSSINA, B.; MERBOUHA, A.; EZZIKOURI, H.; ERRITALI, M. A comparative study of decision tree id3 and c4. 5. **International Journal of Advanced Computer Science and Applications**, Citeseer, v. 4, n. 2, 2014.

Konstantin V. Balakin, S. E. **Pharmaceutical Data Mining: Approaches and Applications for Drug Discovery**. [S.l.: s.n.], 2010. 220 p. ISBN 9780470196083.

LABBE, C. M.; REY, J.; LAGORCE, D.; A, M. V.; BECOT, J.; SPERANDIO, O.; VILLOUTREIX, B. O.; TUFFERY, P.; MITEVA, M. a. MTiOpenScreen: a web server for structure-based virtual screening. **Nucleic Acids Research**, v. 43, n. 18, p. 1–7, 2015. ISSN 0305-1048. Disponível em: <http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv306>.

LANEY, D. 3d data management: Controlling data volume, velocity and variety. **META Group Research Note**, v. 6, p. 70, 2001.

LIU, D.; ZHAO, L. The research and implementation of cloud computing platform based on docker. In: IEEE. **Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2014 11th International Computer Conference on**. [S.l.], 2014. p. 475–478.

MATHEWS, M. **The High Cost Of Inventing New Drugs – And Of Not Inventing Them**. 2015. Disponível em: <http://www.forbes.com/sites/merrillmatthews/2015/04/11/the-high-cost-of-inventing-new-drugs-and-of-not-inventing-them/>. Acesso em: 30 de novembro de 2016.

MCAULEY, J.; LESKOVEC, J. Hidden factors and hidden topics: understanding rating dimensions with review text. p. 165–172, 2013.

MENG, X.; BRADLEY, J.; YAVUZ, B.; SPARKS, E.; VENKATARAMAN, S.; LIU, D.; FREEMAN, J.; TSAI, D.; AMDE, M.; OWEN, S. et al. Mllib: Machine learning in apache spark. **Journal of Machine Learning Research**, v. 17, n. 34, p. 1–7, 2016.

MESIROV, J. P. Accessible reproducible research. **Science**, American Association for the Advancement of Science, v. 327, n. 5964, p. 415–416, 2010.

MISHRA, N. K. Computational modeling of P450s for toxicity prediction. **Expert opinion on drug metabolism & toxicology**, v. 7, n. 10, p. 1211–1231, 2011. ISSN 1744-7607.

MULLARD, A. 2015 FDA drug approvals. **Nature Publishing Group**, Nature Publishing Group, v. 15, n. 2, p. 73–76, 2016. ISSN 1474-1776. Disponível em: <http://dx.doi.org/10.1038/nrd.2016.15>.

NISHA, C. M.; KUMAR, A.; NAIR, P.; GUPTA, N.; SILAKARI, C.; TRIPATHI, T.; KUMAR, A. Molecular docking and in silico admet study reveals acylguanidine 7a as a potential inhibitor of β-secretase. **Advances in bioinformatics**, Hindawi Publishing Corporation, v. 2016, 2016.

NOBELPRIZE. **The Nobel Prize in Chemistry 2013**. 2013. Disponível em: <https://www.nobelprize.org/nobel_prizes/chemistry/laureates/2013/>. Acesso em: 30 de novembro de 2016.

OSAKWE, O.; RIZVI, S. A. **Social Aspects of Drug Discovery, Development and Commercialization**. [s.n.], 2016. 85–108 p. ISBN 9780128022207. Disponível em: <http://www.sciencedirect.com/science/article/pii/B9780128022207000041>.

PAN, A. C.; BORHANI, D. W.; DROR, R. O.; SHAW, D. E. Molecular determinants of drug – receptor binding kinetics. **Drug Discovery Today**, Elsevier Ltd, v. 18, n. 13-14, p. 667–673, 2013. ISSN 1359-6446. Disponível em: <http://dx.doi.org/10.1016/j.drudis.2013.02.007>.

PARAMASHIVAM, S. K.; ELAYAPERUMAL, K.; NATARAJAN, B. bhagavan; RAMAMOORTHY, M. devi; BALASUBRAMANIAN, S.; DHIRAVIAM, K. N. In silico pharmacokinetic and molecular docking studies of small molecules derived from indigofera aspalathoides vahl targeting receptor tyrosine kinases. **Bioinformation**, Biomedical Informatics Publishing Group, v. 11, n. 2, p. 73, 2015.

PIRES, D. E. V.; BLUNDELL, T. L.; ASCHER, D. B. pkCSM: Predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. **Journal of Medicinal Chemistry**, v. 58, n. 9, p. 4066–4072, 2015. ISSN 15204804.

PRINZ, F.; SCHLANGE, T.; ASADULLAH, K. Believe it or not: how much can we rely on published data on potential drug targets? **Nature reviews Drug discovery**, Nature Publishing Group, v. 10, n. 9, p. 712–712, 2011.

SCANNELL, J. W.; BLANCKLEY, A.; BOLDON, H.; WARRINGTON, B. Diagnosing the decline in pharmaceutical R&D efficiency. **Nature reviews. Drug discovery**, v. 11, n. 3, p. 191–200, 2012. ISSN 1474-1784. Disponível em: <http://dx.doi.org/10.1038/nrd3681>.

SHAIKH, U. P. A.; JOSHI, Y. N. Molecular docking studies of e-bola virus protein vp30. 2016.

TETKO, I. V.; ENGKVIST, O.; KOCH, U.; REYMOND, J. L.; CHEN, H. BIGCHEM: Challenges and Opportunities for Big Data Analysis in Chemistry. **Molecular Informatics**, v. 35, n. 11-12, p. 615–621, 2016. ISSN 18681751.

TODESCHINI, R.; CONSONNI, V. **Molecular descriptors for chemoinformatics, volume 41 (2 volume set)**. [S.l.]: John Wiley & Sons, 2009. v. 41.

TROTT, O.; OLSON, A. J. AutoDock Vina. **J. Comput. Chem.**, v. 31, p. 445–461, 2010.

VALLANCE, P.; SMART, T. G. The future of pharmacology. **British journal of pharmacology**, Wiley Online Library, v. 147, n. S1, 2006.

WALSH, G. **Proteins: Biochemistry and Biotechnology**. [S.l.]: John Wiley & Sons, 2014.

ZAHARIA, M.; CHOWDHURY, M.; DAS, T.; DAVE, A. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. **NSDI'12 Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation**, p. 2–2, 2012. ISSN 00221112. Disponível em: <https://www.usenix.org/system/files/conference/nsdi12/nsdi12-final138.pdf>.

ZAHARIA, M.; CHOWDHURY, M.; FRANKLIN, M. J.; SHENKER, S.; STOICA, I. Spark: Cluster Computing with Working Sets. **HotCloud'10 Proceedings of the 2nd USENIX conference on Hot topics in Cloud Computing**, p. 10, 2010. ISSN 03642348.