UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

DEPARTAMENTO DE QUÍMICA

ATTILIO CHIAVEGATTI NETO

MACHINE LEARNING EM QUÍMICA ORGÂNICA

SÃO CARLOS – SP

2020

ATTILIO CHIAVEGATTI NETO

MACHINE LEARNING EM QUÍMICA ORGÂNICA

Trabalho de Conclusão de Curso apresentado ao Departamento de Química da Universidade Federal de São Carlos para obtenção do título de Bacharel em Química

Orientador: Prof. Dr. Marco Antonio Barbosa Ferreira

SÃO CARLOS – SP

2020

À memória de Sandra Lúcia Bezan

AGRADECIMENTOS

Agradeço aos meus familiares, que me apoiaram muito durante todo o curso. Em especial meus pais, Rogério e Daniele, que são meu porto seguro, meus exemplos e minha inspiração.

À todos os professores e técnicos da UFSCar que participaram direta ou indiretamente da minha formação.

Ao professor Dr. Marco Antonio Barbosa Ferreira pela orientação durante a iniciação científica, por todos os seus ensinamentos e pela confiança em meu trabalho.

À Amanda Aline Barboza, por dividir comigo as alegrias e pela companhia nos momentos difíceis.

Aos meus amigos da graduação, em especial os "manjadores" Breno, Bruna, Fábio, Lucas, Maria, Nathalia, Samuel, e Thais pela amizade em todos os momentos dessa longa jornada.

Aos amigos do grupo de pesquisa, Ariel, Arrocha, Guilherme, Henrique, Isac, Ives, Juliana, Luis, Matheus, Meire, Sajjad e João Vitor pelas boas discussões.

Aos meus amigos do LQBO, pelos bons momentos que passei durante a iniciação científica.

À FAPESP pelo auxílio financeiro.

RESUMO

A busca por sistemas computacionais inteligentes capazes de resolver problemas que tradicionalmente são reservados à mente humana é de longa data. Diversas tentativas de desenvolver tais sistemas para resolver questões como identificação estrutural e síntese de moléculas orgânicas foram realizadas a partir dos anos 70, mas a baixa capacidade dos computadores disponíveis e a falta de algoritmos apropriados na época eram limitações severas que inviabilizaram muitos desses projetos. Atualmente, com o crescente aumento na capacidade de processamento e com o enorme volume de informação química acumulada em bancos de dados públicos e comerciais, ressurgiu o interesse em desenvolver tais sistemas. Diversos trabalhos que vêm sendo publicados mostram que utilizando algoritmos de aprendizado de máquina é possível criar programas capazes de gerar automaticamente caminhos sintéticos para moléculas complexas de interesse industrial e acadêmico bem como otimizar reações de maneira eficiente e autônoma.

Palavras-chave: Aprendizado de Máquina. Retrossíntese. Otimização de reação

ABSTRACT

The search for intelligent computational systems capable of solving problems that are traditionally reserved for the human mind are a long-standing crusade. Several attempts to solve problems such as structural identification and synthesis of organic molecules using those toolboxes started in the 1970s, but the low capacity of available computational power and the lack of appropriate algorithms at the time were severe limitations, rendering many of these projects unfeasible. Currently, with a continuous increase in the processing capacity and with an enormous amount of chemical information accumulated in public and commercial databases, the interest in developing these systems has resurged. Several papers that have been published show that using machine learning algorithms it is possible to create programs capable of automatically generate synthetic paths for complex molecules of industrial and academic interest and also optimize reactions in an efficient and autonomous way.

Keywords: Machine Learning. Retrossynthesis. Reaction optimization

LISTA DE ILUSTRAÇÕES

LISTA DE SIGLAS

ANN – *Artificial Neural Network*

CAS – *Chemical Abstract Service*

CASP – *Computer Aided Synthesis Planning*

CIP – *Chemical Information Program*

DCIP – 2,6-diclorofenolindolfenol

DNN – *Deep Neural Network*

DU – Dugundji-Ugi

ESI – *Electrospray ionization*

HTE – *High-Throughput Screening*

InChI – *International Chemical Identifier*

KNN – k-*Nearest Neighbors*

MCS – *Maximum Common Substructure*

MCTS – *Monte Carlo Tree Search*

ML – *Machine Learning*

OCSS – *Organic Chemical Simulation of Synthesis*

SMARTS – *Smiles Arbitrary Target Specification*

SMILES – *Simplified Molecular Input Line Entry Specification*

SUMÁRIO

1 - INTRODUÇÃO

1.1 - Histórico da Química Computacional

A sociedade industrial do século XIX experimentou uma mudança de paradigma iniciada em meados do século XX e mais intensamente a partir dos anos 2000, estabelecendo o que conhecemos hoje como "Era da Informação" – na qual o uso de informações pela sociedade é onipresente, depende de um suporte tecnológico e é intimamente relacionado com aspectos econômicos, socioculturais e políticos[1]. Uma inovação tecnológica iniciada por cientistas da *Moore School of Electrical Engineering* em meados dos anos 40 contribuiu imensamente para essa mudança do cenário global. Motivados pelo esforço durante a Segunda Guerra Mundial eles desenvolveram um equipamento denominado *Electronic Numerical Integrator and Computer* (ENIAC) – o primeiro computador digital eletrônico de grande escala[2]. Esse equipamento foi utilizado brevemente para cálculos de tabelas de trajetória para artilharia e também nas pesquisas do *Los Alamos Scientific Laboratory* – o berço das bombas atômicas deflagradas sobre o Japão em 1945[3]. Após a guerra novos computadores surgiram, impulsionados pelo sucesso de Eckert Mauchly no desenvolvimento do UNIVAC, o primeiro computador disponível comercialmente[2].

O uso de computadores na Química começou na área de Química Quântica. Foi rapidamente percebido que usar uma máquina seria mais eficiente que contratar assistentes munidos de calculadoras para resolver as complicadíssimas equações que emergiam da Teoria Quântica, como já vinha sendo feito desde os anos 30[3]. Um exemplo interessante foi a solução da equação de Schrödinger para a molécula de N_2 que em 1966 levou apenas 2 minutos para ser obtida no melhor computador disponível, enquanto Scherr e dois assistentes levaram 2 anos para resolver o mesmo problema 11 anos antes, em 1955.[4; 5].

O desenvolvimento da teoria de Hückel em 1937 para simplificar o tratamento quântico de sistemas π-conjugados foi o primeiro exemplo na área de

química orgânica teórica aplicada[6]. O Método de Hückel Estendido criado por R. Hoffmann nos anos 60 permitiu incluir também os orbitais σ no tratamento computacional dessas moléculas e foi usado pelo autor em trabalhos conjuntos com R. B. Woodward para elucidar uma série de mecanismos de reações pericíclicas por meio da conservação da simetria de orbitais[7]. Por seu trabalho no estudo do mecanismo de reações, Hoffmann foi laureado com o Nobel em Química de 1981[8]. Outras aplicações dos computadores nos anos 50 e 60 incluem trabalhos de Derek Barton no desenvolvimento da Análise Conformacional[9] – laureado com Nobel em Química em 1969 – e também de Hendrickson e Allinger no desenvolvimento da mecânica molecular[10]. Um passo importante na associação de computadores e ciência foi quando se percebeu que além de resolver equações complexas esses equipamentos poderiam atuar também como uma forma de armazenar e processar informações[3].

Nesse contexto surgiu em meados dos anos 60 o *Chemical Information Program* (CIP), financiado por diversas iniciativas americanas em cooperação com o *Chemical Abstracts Service* (CAS), subsidiária da *American Chemical Society*. O escopo do projeto era modernizar a documentação de insumos químicos existente e desenvolver ferramentas para lidar com essas informações em larga escala. Foi criado um banco de dados no qual toda informação disponível sobre uma dada substância era associada a um único registro numérico. Cada *input* nesse sistema podia ser realizado com a estrutura da substância, seu nome ou uma cifra. Uma tabela de conexão mostrando os átomos da molécula, os átomos a ele conectados e o tipo de ligação entre eles, além do número de massa, valência, número de coordenação e carga era a linguagem utilizada para comunicar ao computador a estrutura da molécula. Além disso o sistema continha a fórmula molecular e também referências bibliográficas relacionadas[11]. O número de registro CAS é muito utilizado atualmente tanto pela academia, iniciativas privadas e instituições governamentais e seu banco de dados contém mais de 100 milhões de registros[12].

A representação de estruturas químicas por meio de tabelas de conexão, sistemas de registro numérico, identificador químico internacional IUPAC (InChI), notação SMILES e SMARTS, entre outros são formas eficientes de armazenar informação química em bancos de dados digitais e permite a implementação de sistemas de busca dentro dessas bases[13]. Representar reações químicas, por outro lado, é uma tarefa muito mais complexa. Isso porque procurar uma reação química em uma base de dados envolve responder questões do tipo "Quais reações convertem o composto A em C?", "Quais reações podem ocorrer entre os compostos A e B?" ou ainda "Como sintetizar o composto C?" e para obter tais respostas a identificação do centro reacional – o conjunto de átomos e ligações envolvidas na transformação – é fundamental[14]. Os métodos de *atom-atom mapping* são empregados para identificar de maneira automática quais átomos dos reagentes estão presentes no produto e muitos desses métodos são algoritmos baseados em *Maximum Common Substructure* (MCS), aos quais são fornecidas duas ou mais moléculas e o objetivo é identificar a maior porção da estrutura (subestrutura) comum entre elas[15]. Apesar da formulação simples, essa é uma questão extremamente complexa de interesse da Química, Biologia, Matemática e Ciência da Computação que pertence à classe dos problemas NP-completos, que não possuem algoritmos eficientes para encontrar uma solução exata[16]. Em particular, o *atom-atom mapping* têm recebido grande atenção na literatura em vista da necessidade de extração automática de regras gerais e padrões de reações químicas para uso em ferramentas de Inteligência Artificial[17]. O desenvolvimento de algoritmos de MCS fomentou a criação de bancos de dados comerciais como o Reaxys da Elsevier (derivado do Beilstein Crossfire) e o CASREACT da CAS que são utilizados até hoje[14].

A partir dos anos 2000 a capacidade de gerar, armazenar e processar dados por meio de sistemas computacionais aumentou rapidamente, em partes devido à popularização da Internet e de sistemas móveis[18]. Esse fenômeno é conhecido por "*Big Data*" – caracterizado por bancos de dados grandes cujo

tamanho, velocidade de aquisição ou representação dos dados limitam a aplicação de meios tradicionais de processamento[19]. O problema de lidar com esses bancos de dados vai além do enorme volume de informação, pois podem também apresentar falta de estrutura coerente e alta dimensionalidade[20]. A ferramenta de escolha para lidar com esses bancos de dados, inclusive os de informações químicas, têm sido os algoritmos de Aprendizagem de Máquina, uma subárea da Inteligência Artificial que busca desenvolver agentes capazes de aprender com a experiência[18].

1.2 - Inteligência Artificial e Aprendizagem de máquina

Os primeiros trabalhos em inteligência artificial foram feitos em 1943 por Warren McCulloch e Walter Pitts e seu estabelecimento como uma disciplina independente começou em meados dos anos 50 quando houve um workshop no *Dartmouth College* para promover o estudo da ciência e da engenharia envolvida na criação de "máquinas inteligentes"[21]. Atualmente a Inteligência Artificial é um campo multidisciplinar, transitando entre ciência da computação, matemática, psicologia, linguística, filosofia, neurociência, ética e outras (**Figura 1**). Sua definição é complexa em vista da subjetividade do conceito de inteligência. Por um lado, sistemas de inteligência artificial podem ser aqueles que buscam emular o ser humano tanto do ponto de vista comportamental – o raciocínio, o aprendizado, a memória – quanto do ponto de vista físico – a visão, a linguagem, o movimento – e por outro, o pensamento ou ação de tais sistemas podem ser medidos em relação a alguma métrica de racionalidade[21].

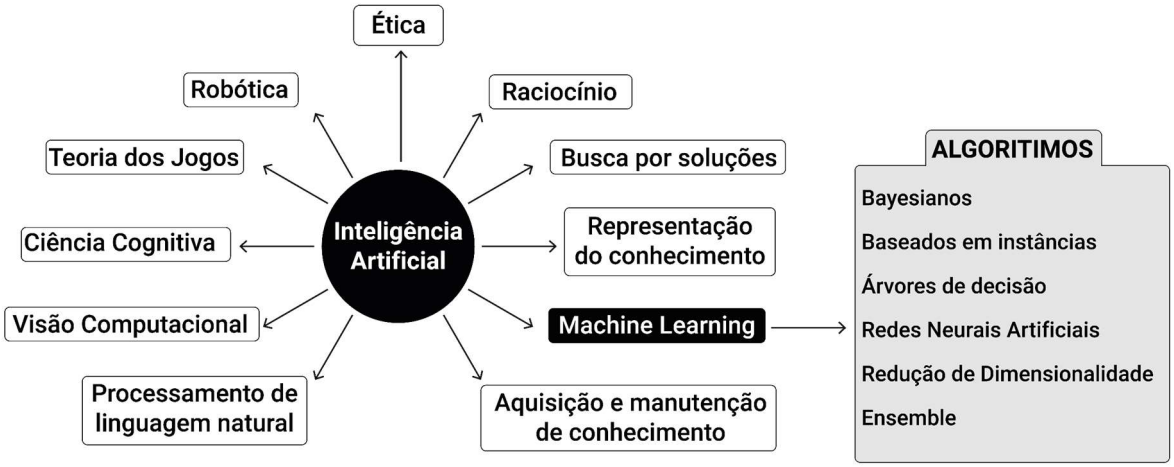


Figura 1 – Visão geral da área de Inteligência Artificial do ponto de vista das aplicações (à esquerda) e dos métodos utilizados (à direita).

No contexto da Inteligência Artificial uma subárea cujo objetivo é explorar de que forma os agentes – em geral computadores – podem aprender a partir de dados, ou ainda uma área que fornece aos computadores meios de aprender sem que sejam explicitamente programados é denominada Aprendizado de Máquina (do inglês *Machine Learning*). Um agente aprende quando sua performance em uma tarefa melhora após realizar observações e esse aprendizado é relevante em situações nas quais não é possível antever todos os cenários em um sistema estático nem antecipar todas as mudanças que nele podem atuar com o tempo – por exemplo um sistema de reconhecimento de texto em imagens[21]. Tradicionalmente os computadores são explicitamente programados por meio de regras para executar tarefas a partir de dados e nesse sentido o aprendizado de máquina representa uma quebra desse paradigma tradicional ao transferir para o computador a tarefa de gerar tais regras e com elas processar dados inéditos[22] (**Figura 2**). Os sistemas de aprendizado usados na prática são diversos, desde modelos lineares e não lineares até os não-paramétricos e outros[21].

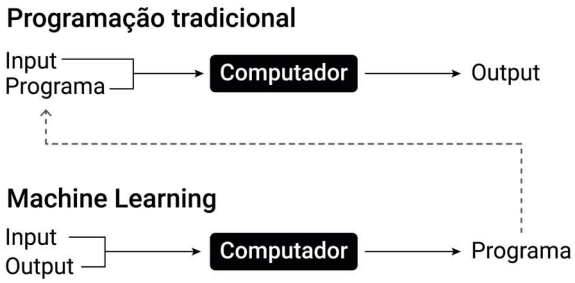


Figura 2 – Os algoritmos de *Machine Learning* permitem gerar um programa a partir dos resultados desejados e esse programa pode ser usado para processar novos dados.

1.2.1 - Formas de aprendizagem

O aprendizado de máquina depende das variáveis presentes nos dados fornecidos ao sistema. As variáveis independentes (*inputs*, *features* ou variáveis)[i] podem ser um conjunto de características discretas ou contínuas tais como: grupos funcionais, energias de orbitais de fronteira, parâmetros estéricos, entre outros. Na ausência de variáveis dependentes (*outputs*, respostas, classes ou *labels*)[i] associadas a cada variável esses dados são ditos **não rotulados** e, caso contrário, são **rotulados**. O aprendizado **não supervisionado** consiste em organizar uma série de variáveis não-rotuladas em grupos. Por outro lado, o aprendizado **supervisionado** parte de um conjunto rotulado em que para cada variável x_i existe uma resposta Y_i associada e o objetivo é encontrar uma função $f(x_1,x_2,...,x_i) = Y_i$ que as relacione. Especificamente, um problema de **classificação** é quando o domínio de tal função é composto por valores discretos ao passo que uma **regressão** envolve valores contínuos. Uma terceira forma de aprendizado é **por reforço**, quando um agente é treinado para atingir um certo objetivo por meio de ações, sendo recompensado por aquelas que o aproximem do objetivo ou punido caso contrário[21].

[i] Por convenção, no presente trabalho utiliza-se o termo variáveis referir às variáveis independentes e o termo respostas para as variáveis dependentes afim de se evitar ambiguidades. Nos textos sobre Machine Learning, entretanto, o comum é utilizar *features* e *labels*, respectivamente.

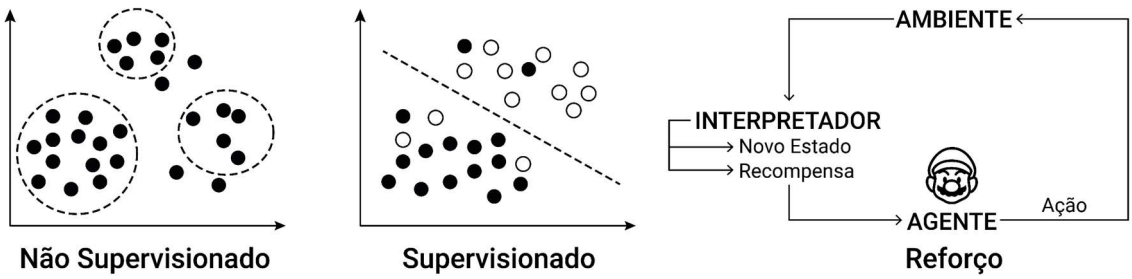


Figura 3 – Formas de aprendizado utilizadas por algoritmos de *Machine Learning*.

1.2.2 - Algoritmos para aprendizado de máquina

O aprendizado ocorre por meio de **algoritmos**. Um algoritmo consiste na descrição de um processo em instruções elementares e o fluxo no qual são executadas[22]. Tipicamente, os dados disponíveis para construção de um modelo com algoritmos de *Machine Learning* são divididos aleatoriamente em dois grupos: o conjunto de treino (*training set*) e o conjunto de teste (*test set*). O conjunto de treino é justamente o grupo de dados em que o modelo é construído de fato e os parâmetros de interesse são ajustados. Já o conjunto de teste é a parte dos dados originais empregada para avaliar a performance do modelo em dados não observados em sua construção – essa prática é conhecida como validação cruzada. A acurácia do modelo em cada conjunto é uma métrica importante pois pode indicar problemas de sobre-ajuste (*overfitting*) – quando o modelo descreve perfeitamente o conjunto de treino, mas não apresenta boa performance para novos dados – ou de sub-ajuste (*underfitting*) – quando o modelo é insuficiente para descrever o conjunto de teste[23] (**Figura 4**).

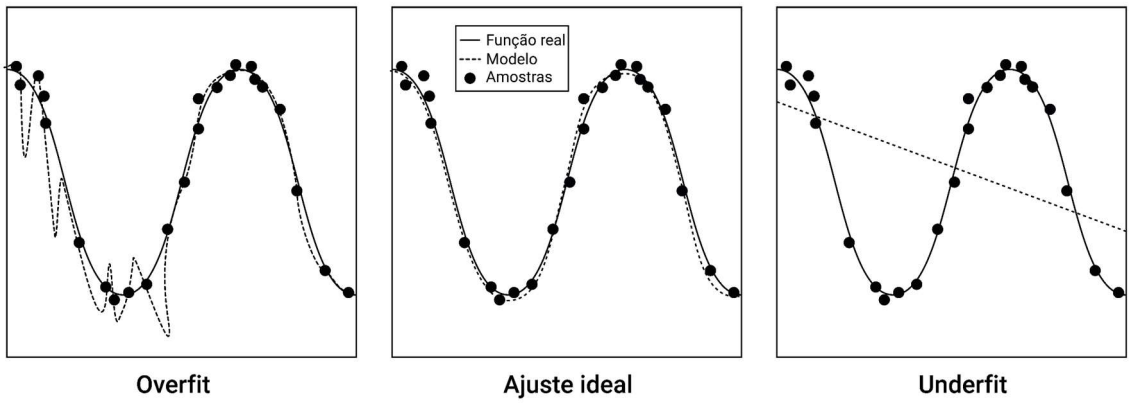


Figura 4 – Possíveis problemas que podem reduzir a performance de modelos de *Machine Learning*.

O algoritmo de k vizinhos mais próximos (kNN, do inglês k *Nearest Neighboors*) é o algoritmo de aprendizado supervisionado mais simples e pode ser utilizado em problemas de classificação e de regressão. O modelo consiste no próprio conjunto de treino e sua função é determinar respostas com base na distância entre as amostras. Seu funcionamento é facilmente visualizado em um problema contendo duas respostas possíveis (**Figura 5**). Nesse exemplo, se k = 1 a resposta atribuída à nova amostra é equivalente à do ponto mais próximo a ela. Mas quando k é maior que 1 a resposta atribuída é aquela presente em maior quantidade dentre os k vizinhos considerados em um problema de classificação e a média desses valores em um problema de regressão. Dois critérios são importantes nesse algoritmo, o número de vizinhos e a métrica de distância utilizada, que em geral é a euclidiana. Esse é um modelo simples que pode ser utilizado como base para medir a performance de outros mais elaborados. Por outro lado, o algoritmo kNN pode ser lento para realizar predições e lida com dificuldade com conjuntos contendo muitas variáveis[24].

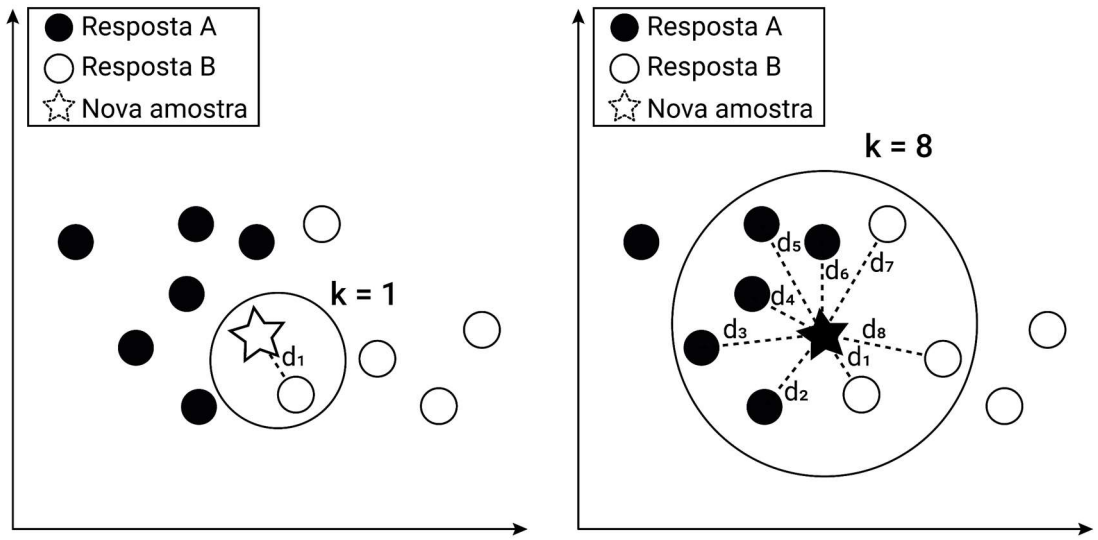


Figura 5 – K vizinhos mais próximos

Os modelos lineares são aqueles que relacionam as variáveis independentes x_i às respostas Y por meio de uma função linear. A forma geral é:

$$Y = a_0 x_0 + a_1 x_1 + ... + a_i x_i + b$$

Onde $\mathbf{a_i}$ é o coeficiente angular em relação a cada variável $\mathbf{x_i}$ e $\mathbf{b}$ é o termo independente. Existem diferentes modelos lineares para regressão que diferem na forma como os parâmetros $\mathbf{a}$ e $\mathbf{b}$ são determinados e de que maneira a complexidade do modelo pode ser controlada. A estratégia mais comum utiliza o método dos mínimos quadrados, em que $\mathbf{a_i}$ e $\mathbf{b}$ são tais que o erro quadrático médio entre as respostas preditas e as respostas originais no conjunto de treino é o menor possível. Outros métodos incluem a regressão "ridge" e a regressão "lasso", as quais permitem incluir restrições na determinação dos coeficientes angulares a de forma que assumam valores sempre próximos a zero (regularização L2) ou exatamente zero (regularização L1), respectivamente. Além disso os modelos lineares também podem ser empregados em problemas de classificação[24].

As árvores de decisão são algoritmos não-paramétricos empregados para classificação e regressão. O modelo é composto por diversos nós organizados de maneira hierárquica tal como um fluxograma. O primeiro desses nós (a "raiz"

da árvore) contém todos os pontos do conjunto de treino e o objetivo do algoritmo é encontrar o melhor critério para dividi-lo. Esse processo segue recursivamente até que todos os nós sejam "puros", isto é, contenham apenas uma classe. Entretanto limitar a profundidade da árvore é importante para evitar o *overfit*. Outras restrições que podem ser impostas ao modelo incluem o número máximo de nós terminais (as "folhas") e a quantidade mínima de pontos contidos em um nó para que ele seja dividido. Diferentes algoritmos de *Machine Learning* podem ser combinados para gerar um modelo – esses métodos compostos são denominados *ensembles*. A combinação de diversas árvores de decisão em um único modelo preditivo é conhecida por *Random Forest*, cuja vantagem é usar a boa capacidade de previsão dos modelos de árvore de decisão ao mesmo tempo que reduz o *overfit* global[23].

O *Deep Learning* é o nome moderno de uma família de algoritmos conhecida originalmente por Redes Neurais Artificiais (ANN, do inglês *Artificial Neural Network*)[24]. Esses modelos diferem entre si principalmente na natureza da grande quantidade de parâmetros utilizados por eles para realizar predições. Assim como nos modelos lineares, as predições realizadas são avaliadas em termos de sua diferença em relação a reposta desejada utilizando uma função custo – tipicamente o desvio quadrático médio ou a entropia cruzada – e por meio de um algoritmo de retropropagação os parâmetros do modelo são ajustados de forma a minimizar essa função custo, em um processo iterativo[25]. O modelo mais simples é o *Deep Neural Network* (DNN) contendo três camadas, a entrada, a oculta e a saída cada uma delas composta por "neurônios" (**Figura 6**). Cada um dos neurônios das camadas ocultas e de saída possuem dois parâmetros ajustáveis, um coeficiente de ativação w e o viés b. A topologia de uma rede neural refere-se à quantidade de camadas utilizadas, de que maneira essas camadas estão conectadas e a função desempenhada por cada neurônio. A transferência de informação entre essas camadas envolve operações matriciais cujo resultado é submetido à uma função de ativação $f(\theta)$ ou $g(\theta)$ – que normalmente é uma função

não linear como a Unidade Linear Retificada ou a Tangente Hiperbólica – o que permite modelar fenômenos não lineares[25].

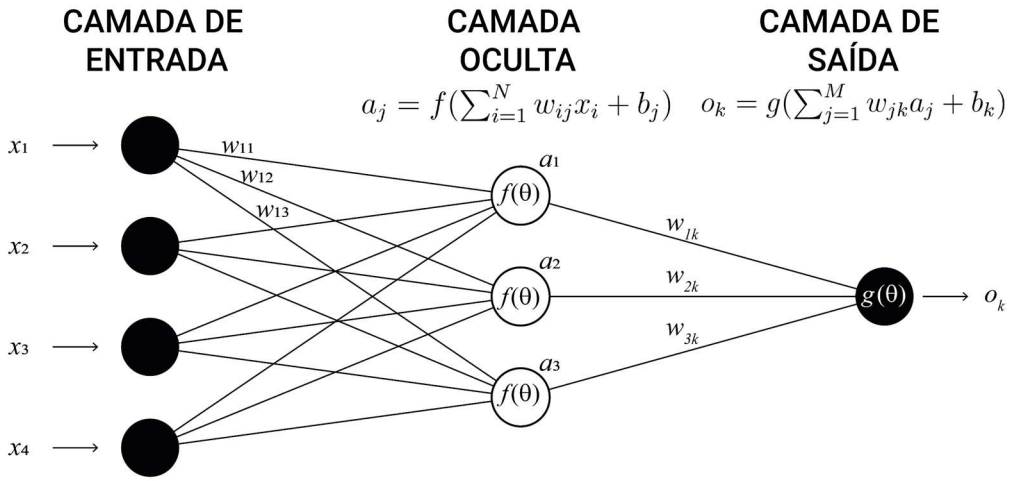
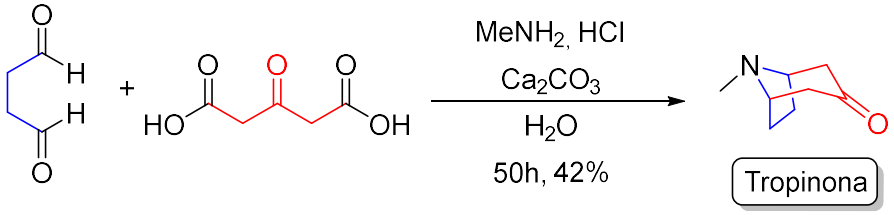


Figura 6 – Exemplo de Rede Neural Artificial contendo uma única camada oculta

Utilizando os bancos de dados construídos com base nas publicações científicas os algoritmos de *Machine Learning* podem ser empregados em Química Orgânica para resolver problemas de retrossíntese (dado um produto encontrar um material de partida conveniente), planejamento de síntese (dado um reagente encontrar os possíveis produtos) e otimização de condições reacionais, entre outros[26].

1.3 - Síntese orgânica e computadores

No ano de 1917 Sir Robert Robinson publicou um trabalho seminal a respeito da síntese da tropinona[27; 28] (**Esquema 1**). Essa molécula é um precursor importante na síntese da atropina, um medicamento capaz de bloquear um tipo específico de receptor de acetilcolina presente nos neurônios permitindo, por exemplo, neutralizar os efeitos de agentes neurotóxicos como o gás sarin – uma arma química amplamente utilizada na Primeira Guerra Mundial, em curso naquela época[29].



Esquema 1 – Síntese da tropinona por Robinson

Nesse trabalho, Robinson selecionou os materiais de partida baseado na simetria do produto, propondo "hidrólises hipotéticas" em ligações específicas. Entretanto, não havia naquela época um método definido para síntese de moléculas orgânicas – hoje conhecido como análise retrossintética. Na realidade, foi somente em 1967 quando Elias J. Corey publicou seu trabalho intitulado "*General Methods For The Construction Of Complex Molecules*" que regras sistemáticas foram propostas[30]. O problema de obter uma molécula alvo a partir de reagentes disponíveis passou a ser visto de maneira reversa, isto é, partindo da molécula desejada a questão se torna identificar quais são as desconexões estratégicas que levam aos reagentes comerciais. Corey ilustrou a análise retrossintética para o sesquiterpeno longifoleno **4** (**Figura 7**).

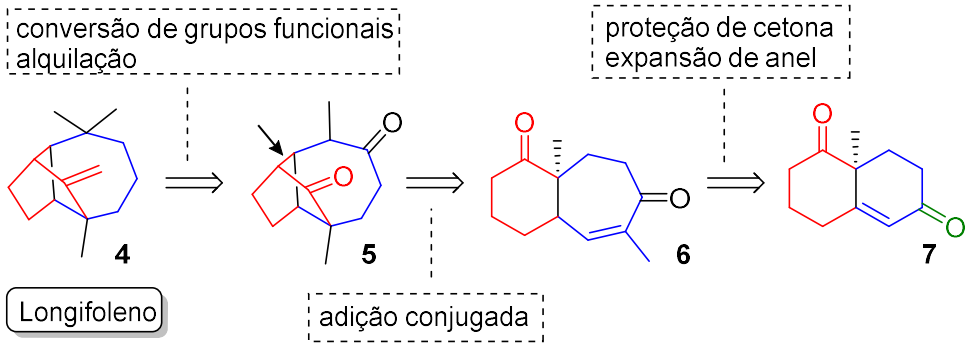


Figura 7 – Retrossíntese do sesquiterpeno longifoleno.

Assim, o planejamento de síntese pode ser definido como o processo de determinar como sintetizar um composto químico partindo de materiais de partida disponíveis por meio de reações químicas viáveis[26]. Desde o princípio da história da retrossíntese foram realizadas tentativas de incorporar essa estratégia

em computadores para obter protocolos de planejamento sintético assistidos por computador (CASP, do inglês *Computer Aided Synthesis Planning*). Dentre as primeiras tentativas de utilizar inteligência artificial em química destacam-se o projeto DENDRAL, criado para elucidar espectros de massas, o projeto OCSS conduzido por Corey e Wipke, considerado a primeira aplicação de inteligência artificial em síntese orgânica, e o projeto IGOR conduzido por Ugi[31].

1.3.1 - O projeto DENDRAL

O primeiro trabalho envolvendo química orgânica e inteligência artificial envolve um conjunto de softwares cujo objetivo era determinar a estrutura de moléculas dado seu espectro de massas, intitulado Projeto DENDRAL[32]. Conduzido por Lederberg e colaboradores, é considerado pioneiro não só na aplicação de inteligência artificial na resolução de um problema químico mas também em um problema envolvendo raciocínio científico[33]. Seu funcionamento consistia na sinergia entre dois módulos distintos, o *meta*-DENDRAL e o *heuristic*-DENDRAL (**Figura 8**). O primeiro era o mecanismo de aprendizado de máquina capaz de gerar e validar as regras necessárias para interpretar um espectro de massas. Ao *meta*-DENDRAL fornecia-se estruturas químicas juntamente com seu respectivo espectro de massas (aprendizado supervisionado). O segundo módulo era o responsável por interpretar um espectro de massas desconhecido e inferir a presença de certos grupos funcionais, baseando-se nas regras geradas pelo *met*a-DENDRAL. Essas informações juntamente com a fórmula molecular eram usadas por um terceiro programa para gerar uma lista de possíveis moléculas associadas ao espectro fornecido[32]. Apesar de sua capacidade em gerar um bom conjunto de estruturas baseado apenas no espectro de massas esse software não alcançou ampla utilização dentre os químicos por diversos fatores tais como alto custo dos computadores na época, curva de aprendizagem muito longa, ausência de demanda, dentre outros[33].

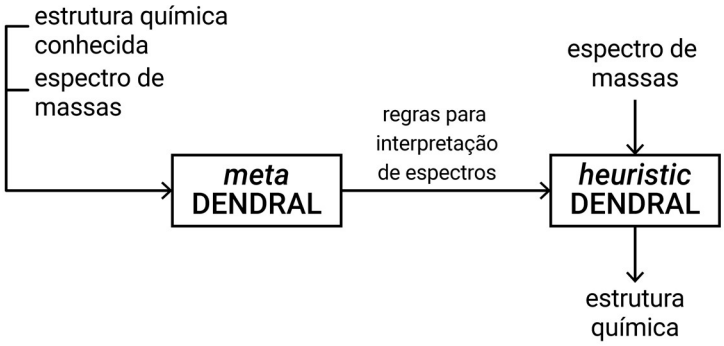


Figura 8 – Diagrama de blocos mostrando o funcionamento do Projeto DENDRAL.

1.3.2 - OCSS - o primeiro programa de retrossíntese

Corey classificou a análise sintética em três estratégias gerais, mostradas na **Figura 9**. Uma delas é a síntese direta, ilustrada pelo alvo sintético **7**, que pode ser divido em diversas subunidades distintas. Essas subunidades são moléculas disponíveis comercialmente e que podem ser acopladas por meio de reações conhecidas. A síntese intermediária envolve converter uma molécula cuja síntese já se conhece em um alvo sintético que apresente estrutura semelhante. Essa estratégia foi aplicada na síntese da cortisona **14** a partir do ácido deoxicólico **13**. Outra estratégia, a síntese lógica, não requer nenhuma hipótese a respeito dos materiais de partida. Partindo do alvo é construída uma "árvore sintética" que segue em direção aos materiais de partida comerciais com redução global na complexidade estrutural[34].

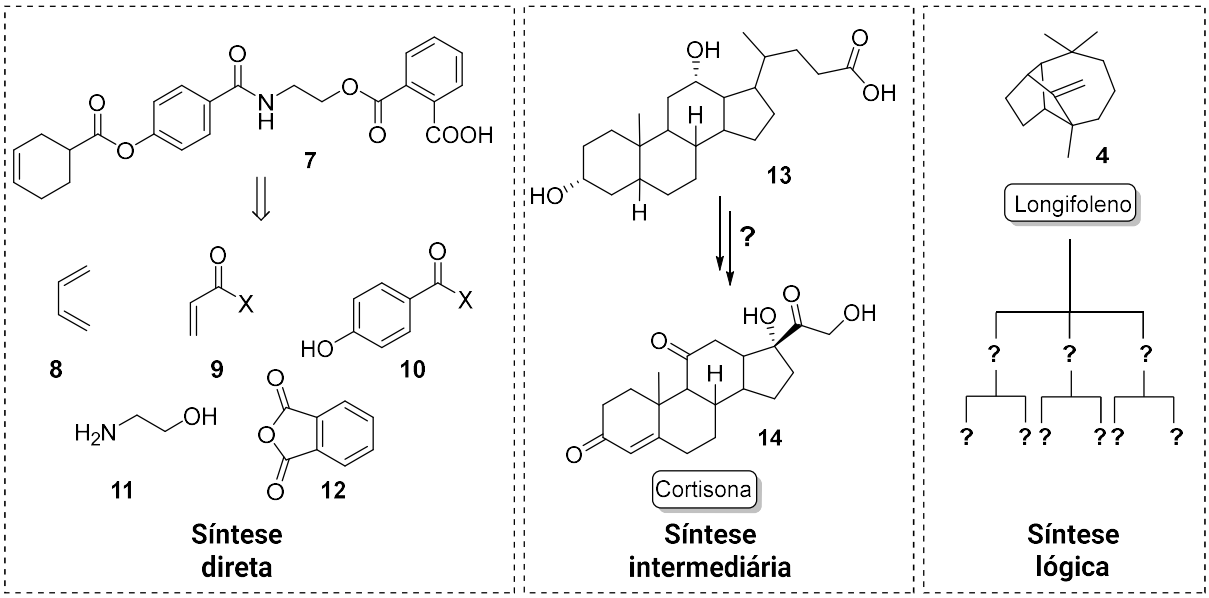


Figura 9 – Estratégias de síntese formalizadas por Corey.

Em vista de sua característica sistemática, a análise lógica é limitada apenas pela fronteira da química – o limite entre quais reações são possíveis e quais não são. Entretanto, Corey reconheceu que até para os melhores químicos esse processo de análise é lento e complexo, uma vez que uma só pessoa dificilmente pode conhecer toda a vasta literatura química[34]. Assim, a possibilidade de "ensinar" um computador as regras elementares da reatividade orgânica e implementar essa sistemática de retrossíntese em um programa para auxiliar na síntese de novos compostos despertou o interesse de alguns grupos de pesquisa[35].

O programa *Organic Chemical Simulation of Synthesis* (OCSS) foi o primeiro de uma série de projetos que visavam atingir tal objetivo. Os desafios dessa empreitada começaram em algo que os químicos atualmente possuem amplo acesso – a representação gráfica de moléculas em um computador[35]. Nesse sentido, todo o controle do programa desde a seleção de opções até o *input* de estruturas moleculares era realizado por meio de um "*Rand Tablet*" e uma caneta e a visualização de dados se dava por meio de três osciloscópios aos quais estava conectada uma impressora para imprimir os resultados finais[34]. Curiosamente,

parte dos resultados gerados no desenvolvimento dessa interface entre representação molecular no papel e o computador foram utilizados no desenvolvimento da suíte de programas ChemDraw, que atualmente é um software amplamente utilizado pelos químicos orgânicos[36].

O OCSS era dividido em 5 módulos, cada um responsável por uma etapa da análise sintética. A primeira ação do OCSS era traduzir a representação gráfica humana em algo que pudesse ser tratado pelo computador. Isso era realizado por meio do módulo de comunicação gráfica, que gerava uma tabela de conexão – tal como citado anteriormente, na construção do banco de dados CAS – na qual eram incluídos todos os átomos da molécula, exceto os hidrogênios ligados aos carbonos, e as suas respectivas ligações químicas.

Essa tabela era interpretada pelo módulo de "percepção" cujo objetivo era identificar as características presentes na molécula tais como grupos funcionais, anéis e seus substituintes, simetria e propriedades eletrônicas simples. Além disso esse módulo podia agrupar grupos funcionais que apresentassem reatividade semelhante e também identificar as relações entre grupos distintos. Para um químico treinado esse tipo de análise é trivial, todavia programar um computador para desempenhar essa tarefa é até hoje é um problema complexo[35].

Os resultados do módulo de "percepção" eram enviados para o módulo de "estratégia e controle", composto por heurísticas da química orgânica, utilizadas para desconectar a molécula. Elas constituíam um conjunto de "regras de ouro", os princípios gerais que compõe o paradigma da síntese orgânica ou ainda o conjunto de conhecimentos empregados, conscientemente ou não, por um químico orgânico na solução de um problema sintético[34]. Por exemplo, o OCSS poderia propor a remoção de grupos reativos nas primeiras etapas de retrossíntese bem como delinear estratégias para contração de anéis de 7 a 10 membros para estruturas mais comuns contendo ciclos de 5 ou 6 membros por meio de rearranjos ou eliminação. Para que o computador tome decisões acerca de qual ligação

desconectar ou de qual grupo funcional trocar essas regras são fornecidas ou explicitamente na forma de ordens ou – na abordagem mais atual – por meio de algoritmos de aprendizagem de máquina. No caso do OCSS as instruções eram embutidas no programa por uma linguagem denominada *Chemistry Translator* (CHMTRN), algo próximo a escrever instruções em língua inglesa para o computador, mas bem distante de uma linguagem de programação tradicional[35].

O módulo de "manipulação" era responsável por executar as estratégias geradas, realizando por exemplo quebras ou formação de ligações, adição ou remoção de átomos e manipulações de carga. A cada atuação desse módulo uma nova camada na "árvore sintética" era adicionada. Uma vez gerados os possíveis caminhos sintéticos o *software* fazia a classificação dessas rotas, identificando por exemplo aqueles que apresentam intermediários comuns.

O OCSS foi um projeto de curta duração. Foi dividido em duas iniciativas, o *Logic and Heuristic Applied to Synthetic Analysis* (LHASA) conduzido no grupo de Corey, sendo aprimorado ao longo dos anos para aumento de sua base de dados que em 1994 já contava com 2100 reações, e o *Simulation and Evaluation of Chemical Synthesis* (SECS) conduzido por Wipke. O SECS possuía nativamente a capacidade de lidar com centros estereogênicos e geometria de duplas ligações, funcionando de maneira muito semelhante ao LHASA. Ele foi implementado na farmacêutica Merck nos anos 80 e avaliado por 50 químicos orgânicos sintéticos dessa companhia. A principal limitação apontada foi o tamanho da base de dados do programa, cuja ampliação mobilizou muitos esforços[37].

1.3.3 - Trabalhos posteriores

Tanto o OCSS quanto o SECS utilizam a implementação de conhecimentos técnicos em um computador visando resolver problemas de forma semelhante ao que seria realizado por um humano. Em ciência da computação tais programas são conhecidos por "sistemas especialistas" e no caso dos programas

de retrossíntese esse tipo de abordagem era utilizado para impor limites na construção da "árvore retrossintética", evitando a explosão combinatória, quando a quantidade de caminhos possíveis se torna grande demais para ser tratada pelo computador[32].

Por outro lado, existem programas construídos em definições abstratas de reações químicas, também denominadas "abordagens formais", que não dependem de um conjunto de regras de reatividade. Nesse caso as transformações químicas são expressas diretamente em termos matemáticos, possibilitando que todo o espaço químico seja explorado. Como consequência esses programas tem sérias dificuldades em lidar com o problema da "explosão combinatorial", de forma que a intervenção do usuário é solicitada quando necessário[35].

Desenvolvido por Ivar Ugi, o programa *Intermediate Generation of Organic Reactions* (IGOR) foi uma implementação da abordagem formal. A representação química era realizada por meio do modelo matemático de Dugundji-Ugi (DU), no qual as moléculas são representadas por matrizes (**Figura 10**). Nesse sistema, uma molécula de n átomos é expressa em uma matriz BE (*bond-electron*) de dimensão n x n cuja diagonal contém o número de elétrons de valência livres para cada átomo e os elementos fora da diagonal fornecem a ordem de ligação. A diferença entre a matriz dos reagentes e dos produtos fornece a matriz R, que representa a própria reação química[38].

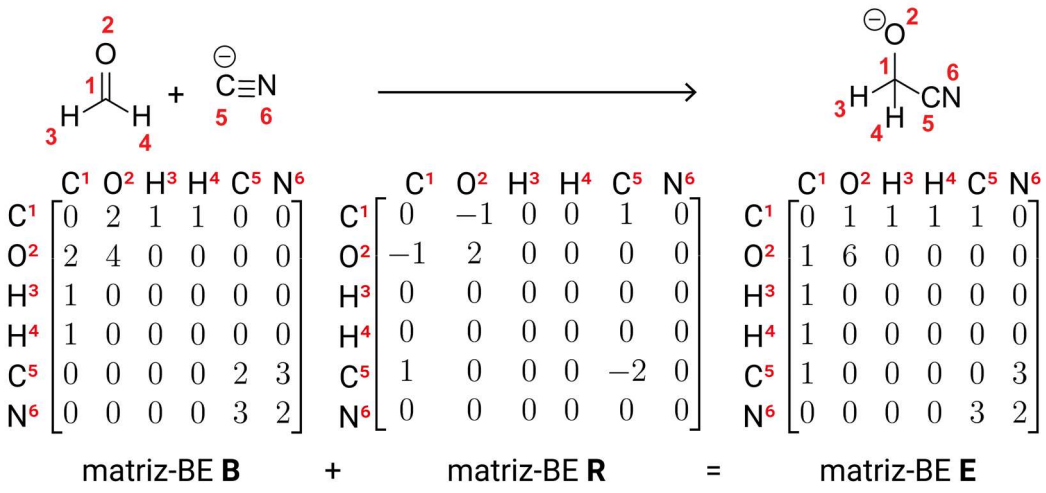


Figura 10 – Modelo de Dugundji-Ugi para uma reação de ataque à carbonila.

O modelo DU permite representar o espaço químico na forma de diagramas nos quais as moléculas são nós conectados por vetores. Essa representação permite medir o progresso sintético em termos de uma "distância química", utilizada como métrica para sugerir os caminhos reacionais mais favorecidos.

Dentre algumas dificuldades relatadas no desenvolvimento desse programa destaca-se a necessidade de incluir a estequiometria completa da reação e também o problema de explosão combinatória, sendo necessária alguma interação do usuário durante o planejamento sintético. Além de pouco prática, essa solução poderia levar ao descarte de propostas contendo reações inéditas, que eram o principal atrativo do programa[35].

1.3.4 - Paradigma atual

O desenvolvimento de tecnologias com potencial de simplificar e automatizar a síntese química é uma tarefa que vem sendo realizada por mais de 50 anos e ainda permanece em desenvolvimento[39]. Apesar dos esforços pioneiros descritos acima e outros, a síntese assistida por computador falhou em atingir ampla utilização de rotina[31]. O entusiasmo dos anos 70 e 80 se dissipou até praticamente desaparecer nos anos 2000, quando o desafio de ensinar o

computador como sintetizar moléculas ganhou a reputação de “missão impossível”[40]. As razões são diversas, entre elas a dificuldade em construir bancos de dados de alta qualidade, baixa capacidade computacional e ausência de algoritmos apropriados. Entretanto, suportados pelo contínuo aumento na capacidade de processamento e armazenamento dos computadores os algoritmos de *Machine Learning* atuais são capazes de lidar com a sempre crescente quantidade de dados disponíveis publicamente – o *Big Data*. Por essa razão, atualmente a síntese orgânica assistida por computador experimenta um interesse renovado entre os químicos[39].

2 - OBJETIVOS

O presente trabalho tem como objetivos apresentar uma breve visão histórica do uso de ferramentas de inteligência artificial no contexto da química orgânica, apresentar os principais conceitos de aprendizado de máquina e mostrar alguns trabalhos recentes que empregam algoritmos para resolver questões do interesse da química orgânica como análise retrossintética de moléculas de interesse comercial, predição de produtos de reação e otimização de condições reacionais.

3 - DISCUSSÃO

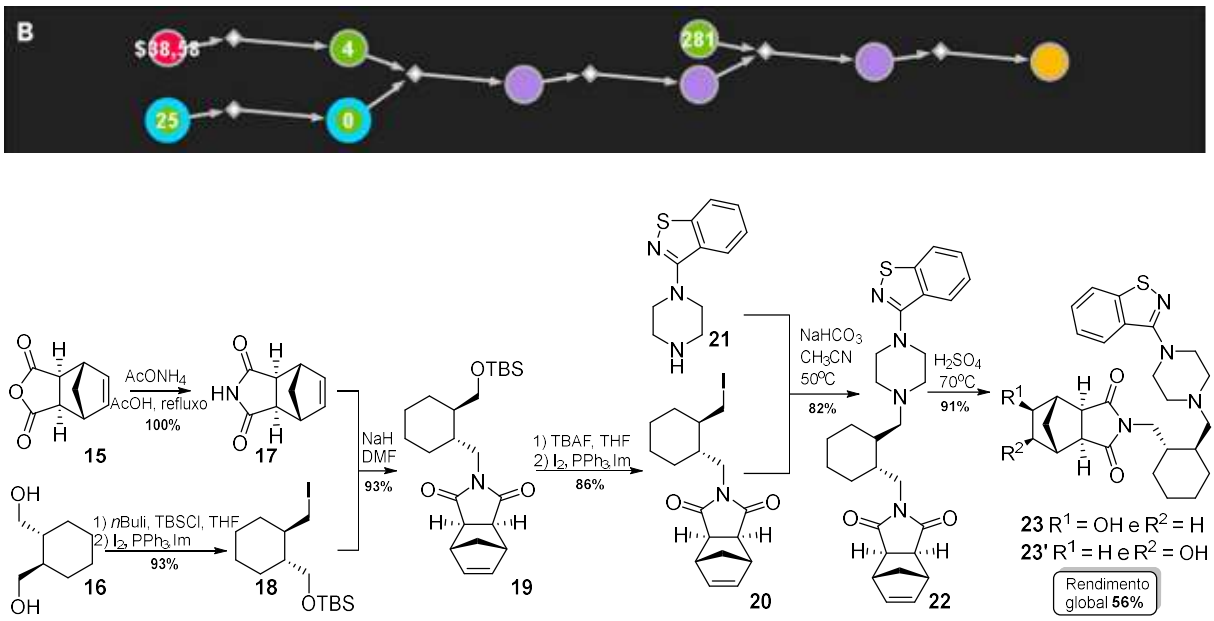
3.1 - *Machine Learning* em planejamento sintético

O desenvolvimento de programas de planejamento sintético assistido por computador (CASP) atualmente envolve tanto a busca por melhores soluções na área de retrossíntese quanto na previsão de produtos de uma reação, uma vez que essas aplicações são complementares.

Existem diferentes formas utilizadas para construir a base de dados de um programa de retrossíntese, dentre as quais a tradicional é a construção de sistemas especialistas. Nesse caso, o conjunto de regras que regem a reatividade das moléculas devem ser explicitamente declaradas ao programa, por exemplo na forma de centros reacionais. Essas bases de dados dificilmente são grandes o suficiente para explorar o espaço químico de maneira satisfatória e consequentemente suas previsões são limitadas a aplicações simples. Entretanto, desde que tempo suficiente seja empenhado para incluir a maior quantidade de regras possível, resultados impressionantes podem ser obtidos.

Essa abordagem foi utilizada por Grzybowski e colaboradores[40] no desenvolvimento do Chematica – produto de mais de 10 anos de trabalho que atualmente conta com aproximadamente 50 mil regras implementadas em sua base de dados. Esse programa representa os caminhos sintéticos na forma de uma rede, utilizando algoritmos altamente sofisticados para explorá-las e gerar rotas sintéticas de acordo com critérios como custo e periculosidade dos reagentes[41]. Atualmente esse programa é disponível comercialmente sob o nome de Synthia® de propriedade do conglomerado MilliporeSigma/Merck KGaA[42]. O Chematica foi o primeiro programa capaz de propor rotas sintéticas completas e que foram validadas em laboratório para moléculas de interesse medicinal e industrial[41]. Seu funcionamento é ilustrado na síntese do composto 5β/6β-hidroxilurasidona **23**, cuja única rota conhecida era patenteada (**Esquema 2**). O caminho proposto pelo programa começa com o anidrido **15**, posteriormente convertido para a imida **17**,

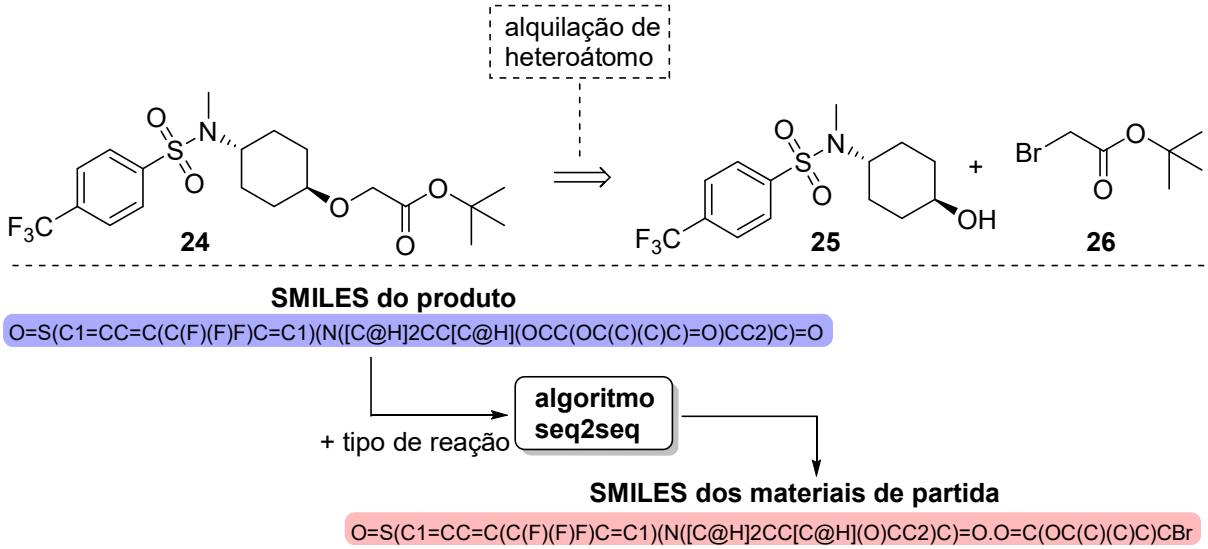
e o diol **16** que é protegido seletivamente com cloreto de *tert*-butildimetilsilil e iodado gerando o composto **18**. A imida **17** é alquilada com **18**, gerando o intermediário **19**. A seguir ocorre a desproteção da hidroxila, que é iodada gerando **20**. Esse intermediário é utilizado na N-alquilação de **21** gerando o intermediário **22**, que leva ao produto de interesse após a hidratação da dupla ligação. Essa rota mostrou-se eficiente em escala de grama e gerou um rendimento global 2,5 vezes maior que o oferecido pela rota patenteada[41].



Esquema 2 – Rota gerada pelo Chematica para síntese da 5β/6β-hidroxilurasidona. Acima é mostrado o diagrama gerado pelo programa.

Uma segunda maneira de construir a base de dados de programas de retrossíntese é a abstenção completa do uso de regras de reatividade. Um dos caminhos possíveis é tratar a questão retrossintética como um problema de tradução, em que o objetivo é transformar uma sequência de texto que representa o material de partida naquela que representa o produto. Essa questão foi abordada por Liu e colaboradores[43] em um programa que toma uma molécula expressa em notação de texto SMILES e um tipo de reação como alquilação, arilação, formação de heterociclo, oxidações entre outras e fornece os materiais de partida mais prováveis (**Esquema 3**). Como modelo foi utilizada uma rede neural especializada chamada *sequence-to-sequence* (seq2seq) utilizada em sistemas de

tradução computacional como o Google Tradutor[44] treinada em uma base de dados de patentes químicas[45]. O protocolo desenvolvido por Liu permitiu realizar uma única etapa de retrossíntese de moléculas simples.



Esquema 3 – O modelo baseado no algoritmo de seq2seq desenvolvido por Liu e colaboradores utiliza fragmentos da notação SMILES do produto desejado para propor possíveis materiais de partida.

Utilizando um algoritmo de seq2seq semelhante, Lin e colaboradores[46] aprimoraram a estratégia proposta por Liu[43] para permitir a geração de caminhos retrossintéticos completos. Foi empregado o algoritmo de MCTS (*Monte Carlo Tree Search*) para expandir uma árvore de intermediários (**Figura 11**). Dado um certo estágio hipotético na retrossíntese de um alvo, a primeira ação desse algoritmo é selecionar um certo nó **A,** que representa um intermediário reacional, partindo da raiz **R** do diagrama. Se o nó **A** não é a solução do problema retrossintético, então o algoritmo realiza a expansão da árvore gerando novos intermediários **B** e **C**. Em seguida o algoritmo seleciona um desses dois nós criados, no caso o **C**, e realiza simulações aleatórias (*rollout*), gerando diversos novos intermediários (não mostrados) até atingir um nó terminal. Esse nó terminal contém um valor associado, que é propagado de volta à raiz **R** para atualizar todos os estados da árvore retrossintética. Esse novo estado permite ao

algoritmo avaliar se a melhor decisão é ampliar a árvore a partir de **C** ou explorar outros nós.

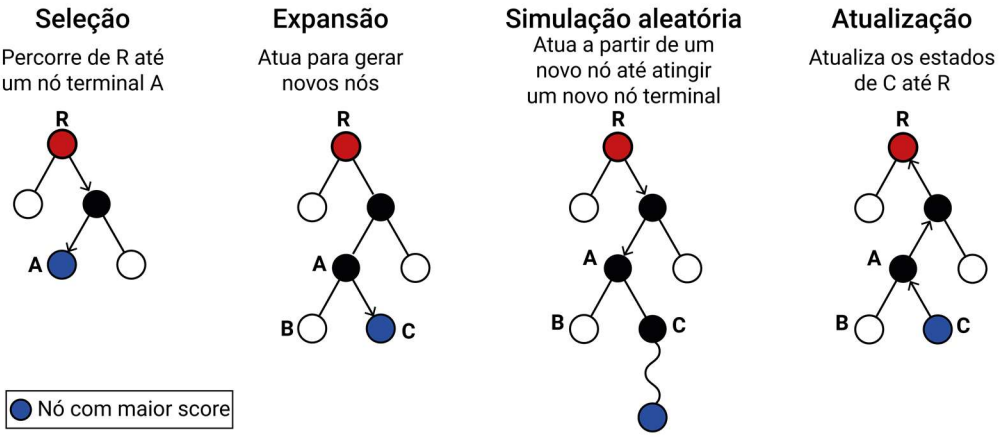
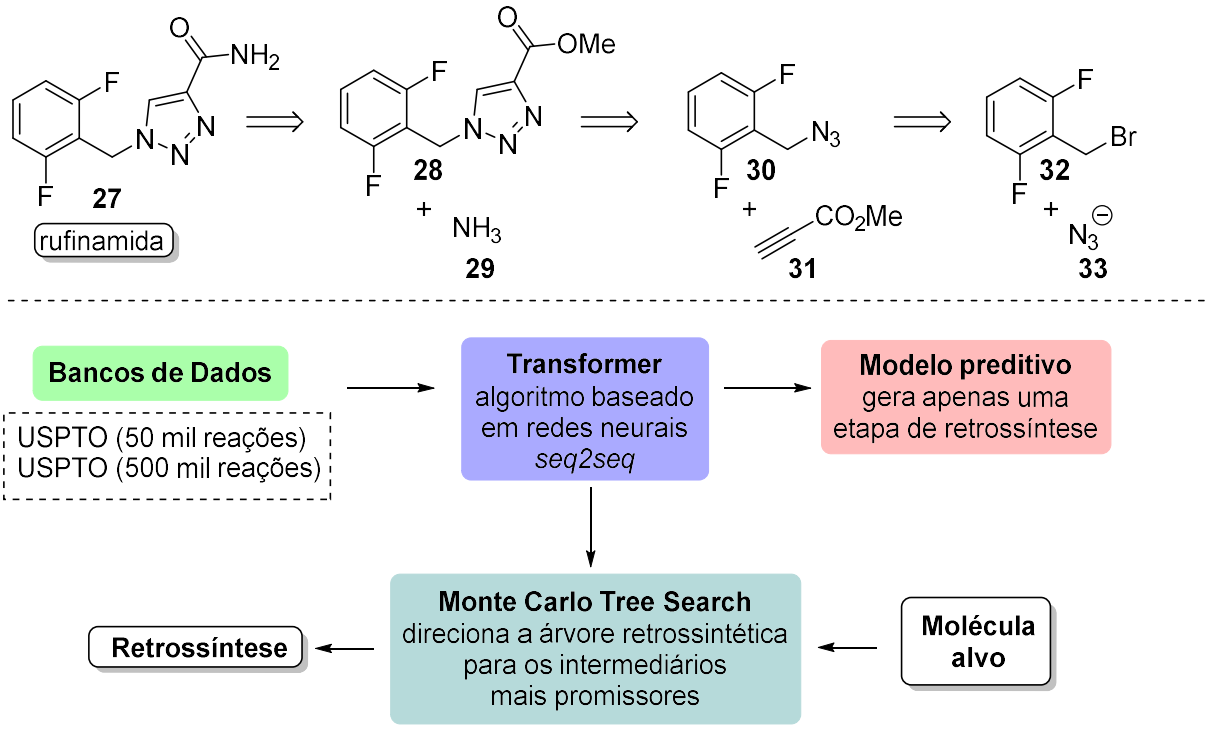


Figura 11 – Representação esquemática do algoritmo de *Monte Carlo Tree Search* (MCTS).

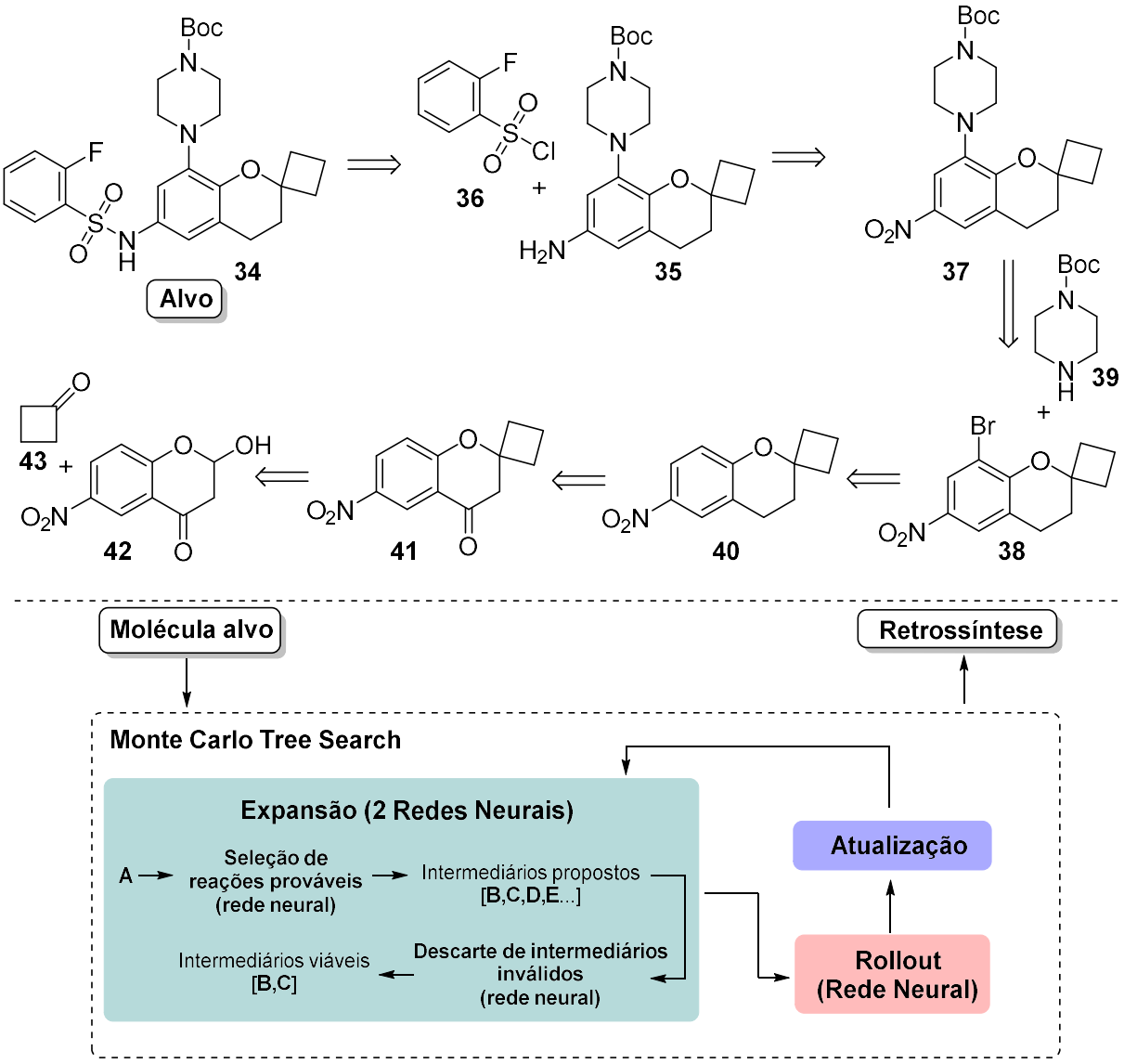
O modelo de Lin e colaboradores[46] utiliza como algoritmo de expansão e de *rollout* para o MCTS um "*Transformer*" baseado em seq2seq. Esse protocolo apresentou performance superior ao proposto por Liu quando considerada uma única etapa. Seu funcionamento é ilustrado na retrossíntese proposta para a rufinamida, um medicamento anticonvulsionante. (**Esquema 4**).



Esquema 4 – O modelo de Lin e colaboradores utiliza um algoritmo *sequence-to-sequence* acoplado com MCTS para gerar rotas retrossintéticas completas.

Outra sistemática utilizada nos programas de retrossíntese envolve selecionar algumas regras específicas, em detrimento de aplicar uma biblioteca completa. Esse método foi utilizado por Segler e colaboradores[47] para desenvolver um sistema baseado em redes neurais capaz de extrair e avaliar a importância de regras de reações químicas, compostas pelo centro reacional e seus átomos vizinhos. Posteriormente esse sistema foi implementado numa solução mais ampla que empregava MCTS para desenvolver rotas de retrossíntese completas em velocidades impressionantes, como no planejamento sintético de 6 etapas proposto para a molécula **34** completado em apenas 5 segundos[48] (**Esquema 5**). O programa é composto por 3 redes neurais distintas, sendo uma responsável por propor transformações possíveis em cada posição, outra para verificar a probabilidade de a reação proposta funcionar e a terceira para realizar simulações de Monte Carlo em cada ponto, expandindo a árvore retrossintética. O autor aponta como deficiências desse protocolo a dificuldade em lidar com moléculas

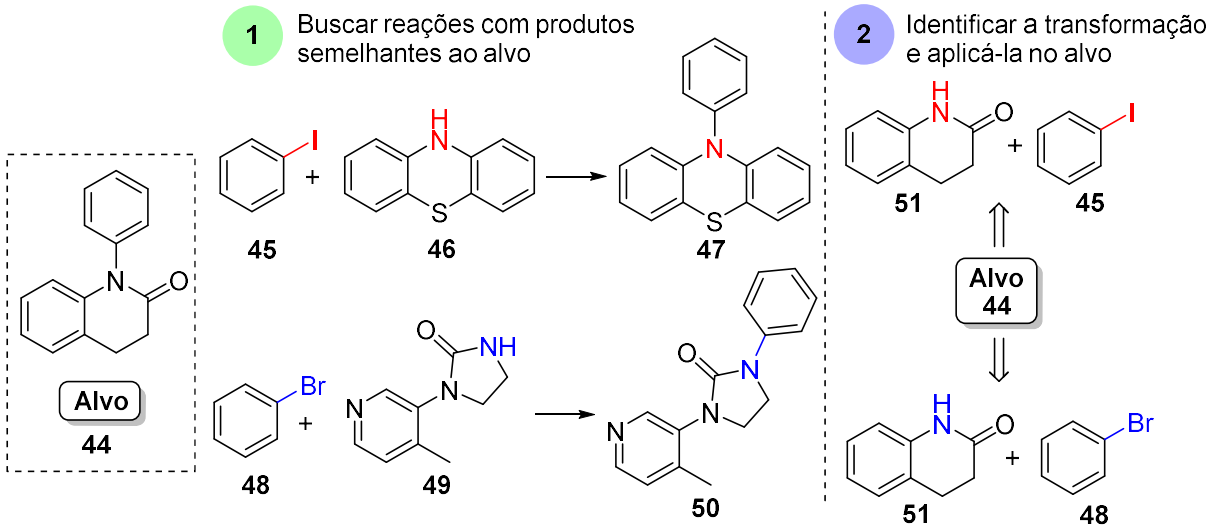
de produtos naturais e com centros estereogênicos e a ausência de condições reacionais nas rotas propostas.



Esquema 5 – O modelo de Segler e colaboradores utiliza 3 redes neurais acopladas com MCTS para realizar propostas de retrossíntese.

Foi relatado por Coley e colaboradores[49] um modelo baseado em similaridade molecular cuja base de dados é um conjunto de 40 mil reações depositadas em um banco de patentes[45]. Inicialmente, o programa procura no banco de dados por reações que apresentem produtos semelhantes ao alvo sintético. Em seguida, o centro reacional de cada reação encontrada é determinado e a mesma transformação é aplicada ao alvo desejado, gerando possíveis

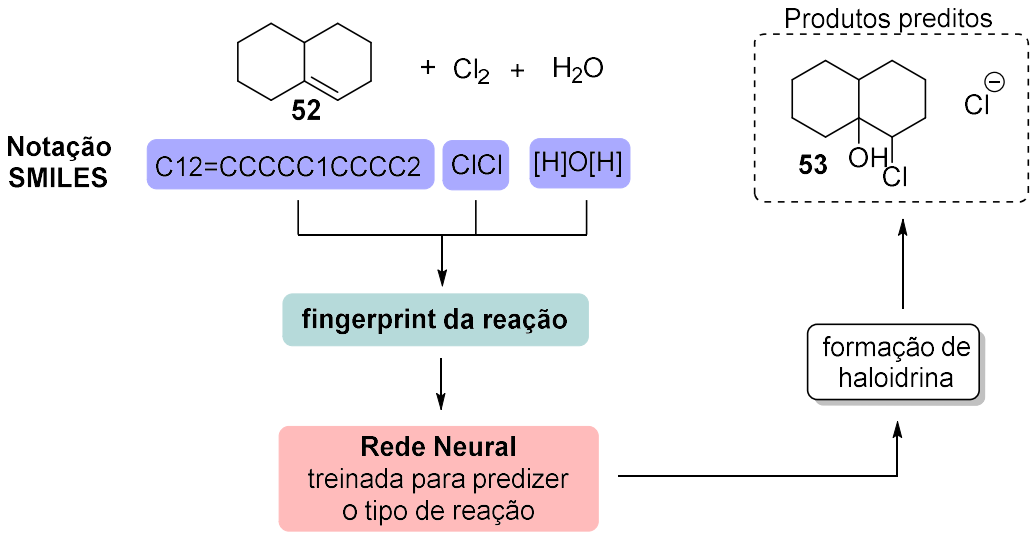
precursores. Por fim, os precursores gerados são avaliados em termos de sua similaridade com os precursores originais (**Esquema 6**). Foi demonstrado que essa abordagem também pode ser estendida para mais de uma etapa de retrossíntese. Por definição, esse programa foi criado para aplicar conhecimento químico existente em novos substratos e não é capaz de gerar desconexões inovadoras.



Esquema 6 – Modelo proposto por Coley e colaboradores para realizar análises retrossintéticas por meio da similaridade molecular.

Determinar qual a distribuição de produtos de uma reação química a partir de variáveis experimentais como material de partida, reagentes, catalisador, solvente, concentração, temperatura, tempo, entre outras é um desafio. Extrair dados tão específicos da literatura é uma tarefa complexa. Não obstante, soluções para esse problema têm sido propostas porque a predição de produtos pode auxiliar na redução de falsos positivos – reações que não podem ser realizadas ou que não levam ao produto de interesse – propostos por algoritmos de retrossíntese. Um dos caminhos é usar uma estratégia mais simples: a determinação do produto majoritário dado um material de partida, reagentes e, quando possível, o solvente, catalisador e a temperatura[26].

No trabalho publicado por Wei e colaboradores[50] foi desenvolvido um modelo que permite classificar materiais de partida e reagentes em termos de tipos de reações químicas (substituição nucleofílica, eliminação, hidrogenação, entre outras) que eles podem participar. Para tanto, os materiais de partida e os reagentes de reações conhecidas foram descritos por *fingerprints* – que são identificadores numéricos atribuídos a cada átomo presente em uma molécula gerados por um algoritmo especializado[51]. Esses *fingerprints* foram submetidos à uma rede neural para criar o sistema de classificação por aprendizado supervisionado. Sabendo a classe de reações possíveis para um dado material de partida, foi possível predizer os produtos majoritários por meio de uma transformação SMARTS – um tipo de notação para expressar na forma de texto variações estruturais em uma molécula – associada a cada reação (**Esquema 7**).



Esquema 7 – Programa desenvolvido por Wei e colaboradores para predizer produtos de reação baseado em *fingerprints* moleculares.

Coley e colaboradores[52] desenvolveram um modelo bastante elaborado para predizer o produto majoritário de reações orgânicas. Foram utilizadas duas bases de dados compostas por reações patenteadas entre 1976 e 2013 que foram compiladas por Lowe[45], uma contendo 1.122.662 reações das quais foi extraído o centro reacional (conjunto de átomos e ligações envolvidos na reação química) e outra contendo 15.000 reações utilizadas para treinar e

validar o modelo de aprendizado de máquina, todas elas com o *atom-atom mapping* definido e codificadas em notação de texto.

Utilizando o banco de dados de 1.122.662 reações foram extraídos em torno de 140 mil centros reacionais distintos dos quais foram selecionados apenas aqueles que se repetem mais de 50 vezes, totalizando 1689 possíveis reações químicas das quais as mais populares são hidrólise de ésteres, redução de grupo nitro à amina e N-desalquilação de aminas terciárias e secundárias. Basicamente, esses centros reacionais são subestruturas que podem ser expressas em códigos como a notação SMARTS para descrever precisamente quais ligações presentes em um reagente se modificam para se tornar ligações no produto e quais são quebradas gerando grupos de saída. Esses códigos representam transformações e podem ser aplicados em um conjunto de reagentes, também expressos em notação SMARTS, para gerar os produtos (**Figura 12**).

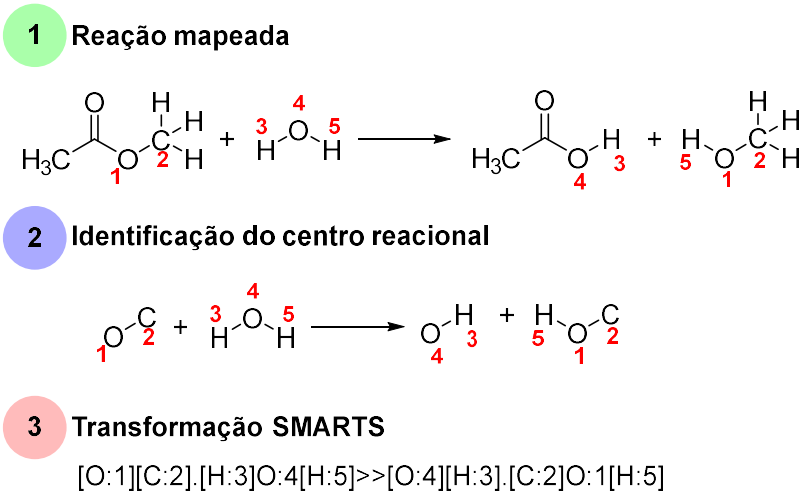
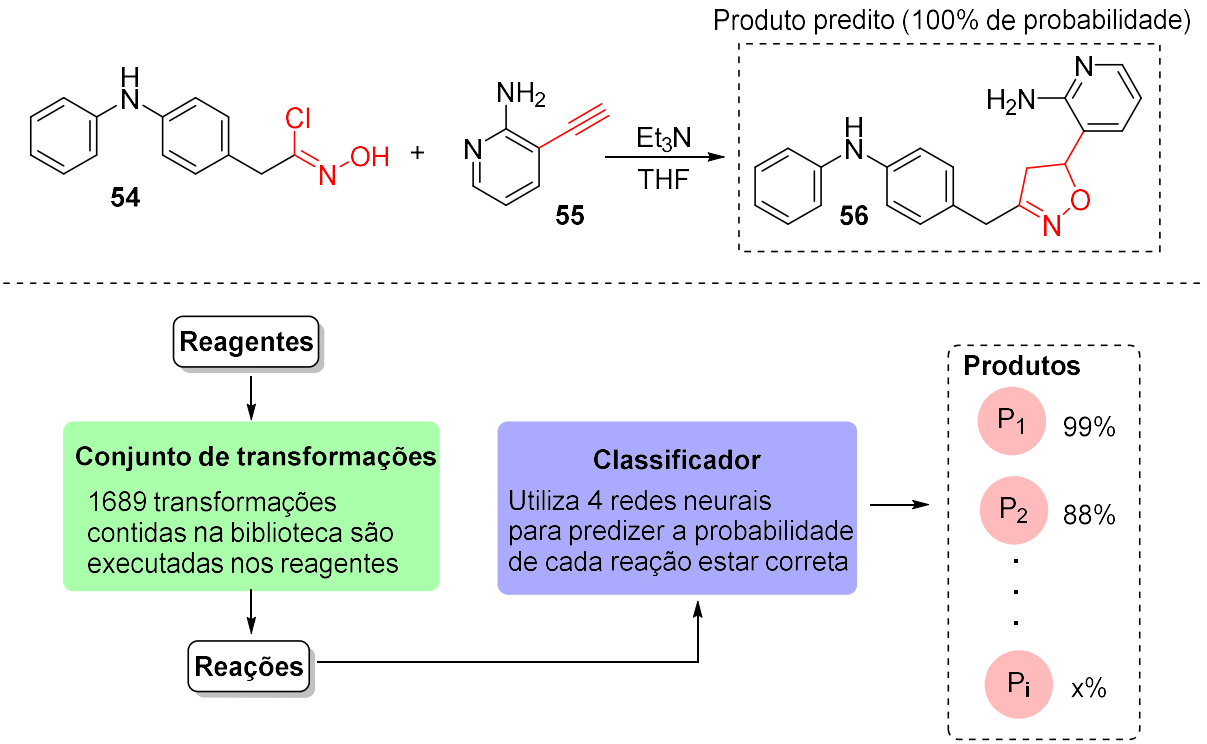


Figura 12 – Visão qualitativa do processo de extração de regras reacionais a partir de bancos de dados de reações mapeadas. A transformação SMARTS mostrada em **3** pode ser aplicada em qualquer éster reagindo com uma molécula de água para obter os produtos de hidrólise.

O modelo é construído a partir das 1689 regras supracitadas. Elas são aplicadas aos reagentes das 15.000 reações do segundo banco de dados, gerando para cada uma delas 353 produtos, em média. Como o produto majoritário verdadeiro é conhecido, esse é um problema de aprendizado supervisionado.

Curiosamente, ao invés de classificar os produtos gerados em termos da probabilidade de ser o verdadeiro produto majoritário, o modelo de Coley e colaboradores visa calcular a probabilidade de cada reação ser a que leva ao produto correto. Essa mudança sutil permite dividir o problema de classificação em 4 redes neurais, em que cada uma avalia uma possível transformação (perda de hidrogênio, ganho de hidrogênio, quebra de ligação química e formação de ligação química), atribuindo uma pontuação final que permite identificar qual reação é a mais provável para os reagentes fornecidos e, consequentemente, identificar os produtos associados (**Esquema 8**).



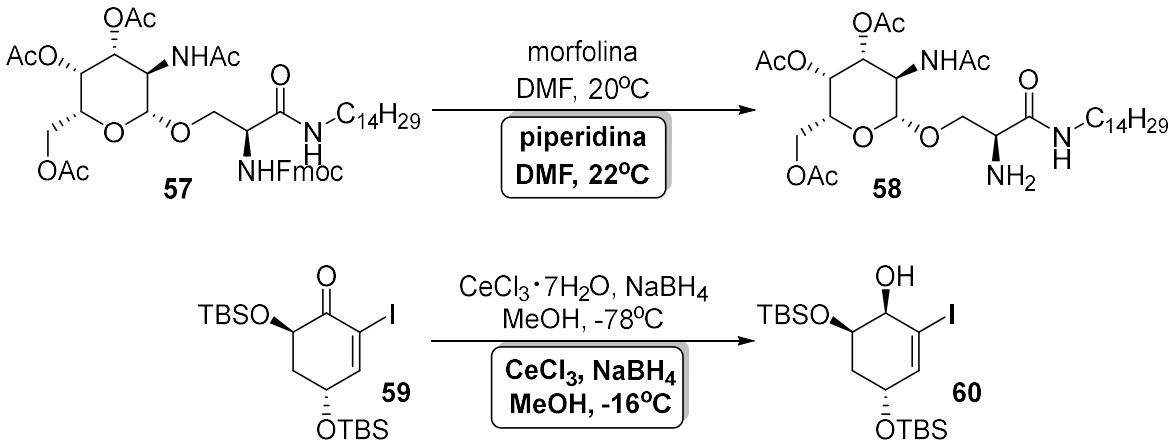
Esquema 8 – Esquema mostrando uma reação predita em cima, e abaixo a explicação do modelo.

3.2 - *Machine Learning* em otimização de reações

A escolha e otimização de condições experimentais para realizar uma reação química são etapas fundamentais de um planejamento sintético. A definição das variáveis experimentais como solventes e catalisadores pode ser realizada com base em reações análogas, experiências prévias e com ferramentas

da físico-química como os modelos cinéticos[25]. Eventualmente as condições empregadas em precedentes da literatura podem não funcionar em novos exemplos e o viés humano pode limitar o horizonte de possibilidades a serem testadas. Nesse sentido, o desenvolvimento de programas capazes de sugerir condições experimentais detalhadas para uma certa transformação seria de grande auxílio para o químico experimental e um complemento importante nos programas de planejamento sintético.

Gao e colaboradores[53] desenvolveram um modelo treinado em 10 milhões de reações disponíveis no banco de dados Reaxys para predizer condições para reações químicas de diversas classes utilizando uma rede neural. A proposta consiste em uma rede neural treinada predizer a temperatura, um catalisador, duas opções de solvente e dois reagentes para uma dada reação química. A avaliação do modelo utilizando 1 milhão de reações no conjunto de testes mostrou que em 70% dos casos foi possível propor o catalisador, pelo menos um solvente e um reagente de maneira correta e temperatura adequada com precisão de ± 20 ºC. Testes de aplicação mostraram que o sistema foi capaz de fornecer condições para reações de hidrólise, esterificação, reação de Wittig, reação de Suzuki-Miyaura, entre outras (**Esquema 9**).



Esquema 9 – Condições reacionais preditas pelo modelo desenvolvido por Gao e colaboradores para uma reação de desproteção de Fmoc (acima) e uma redução de Luche (abaixo). Condições preditas em negrito.

Em muitos exemplos os algoritmos de ML são empregados como "caixas-pretas", pois produzem modelos com alto poder preditivo, mas com poucas possibilidades de interpretação dos parâmetros utilizados. Nesse sentido, Walker e colaboradores[54] desenvolveram um protocolo para predição de solventes em reações clássicas, especificamente a de Friedel-Crafts, adição aldólica, condensação de Claisen, reação de Diels-Alder e reação de Wittig. Dentre os algoritmos de ML testados o kNN apresentou a melhor performance, com acurácia acima de 90% considerando as três melhores sugestões para cada reação. Conforme discutido, esse algoritmo realiza predições para respostas de novas amostras com base na distância delas aos pontos contidos no conjunto de treino. No modelo de Walker a distância entre as reações químicas é determinada pela similaridade entre elas. A similaridade, por sua vez, depende tipo de catalisador e a estrutura molecular do reagente, apenas. Curiosamente, o conjunto de treino classificado de acordo com essas variáveis se distribui naturalmente em função do solvente utilizado, mesmo que essa informação não seja fornecida para o treinamento do modelo. Foi verificado que uma pequena porção dos catalisadores aparecem frequentemente nas reações consideradas e se organizam em grupos grandes (**Figura 13-a**). Além disso, as reações que se distribuem em grupos menores geralmente utilizam o mesmo solvente, ao passo que nos grupos grandes a variedade é maior (**Figura 13-b**). Com isso, o algoritmo é capaz de atribuir o solvente para uma reação com base no catalisador utilizado e na similaridade entre o reagente proposto e os contidos na base de dados, mimetizando o raciocínio de um químico ao procurar precedentes na literatura.

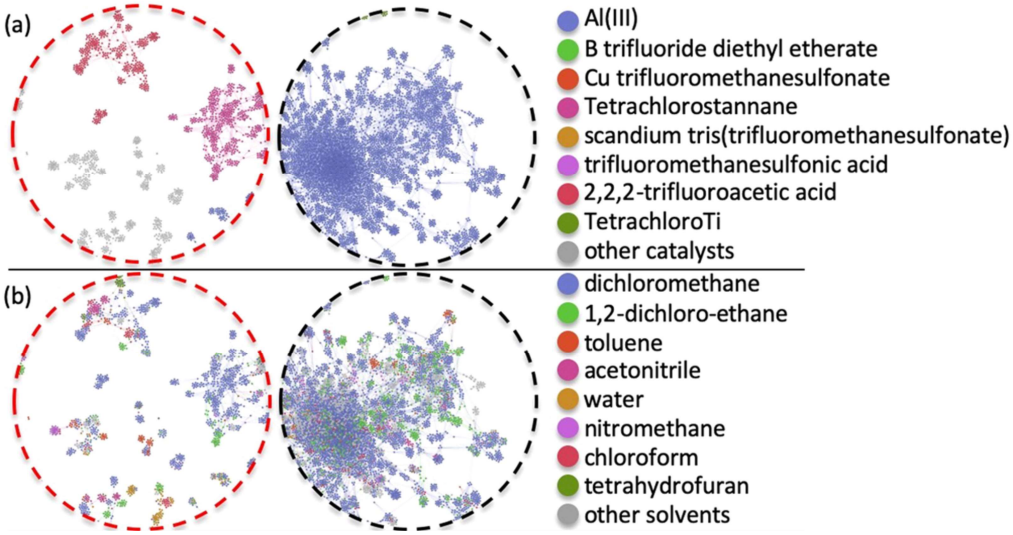
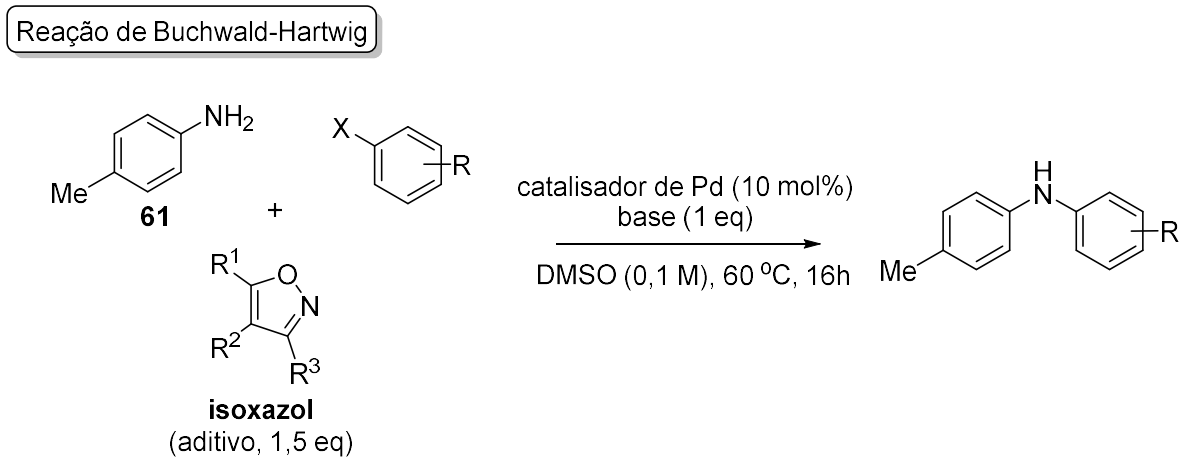


Figura 13 – Dois fragmentos distintos (círculo vermelho e círculo preto) da rede de reações de Friedel-Crafts organizadas de acordo com o modelo de kNN proposto por Walker e colaboradores[54]. A legenda indica em **(a)** o catalisador utilizado e em **(b)** o solvente.

Ahneman e colaboradores[55] desenvolveram um protocolo de ML para predizer o rendimento de reações de Buchwald-Hartwig entre haletos de arila ou alquila e 4-metilanilina, com o objetivo de melhorar a performance da reação para substratos contendo heterociclos que apresentam uma ligação entre heteroátomos como o isoxazol. Para evitar o imenso trabalho de sintetizar e purificar diversas moléculas contendo o núcleo isoxazol foi utilizado o *screening* intermolecular de Glorius[56], em que o grupo funcional de interesse é adicionado a reação na forma de aditivo (**Esquema 10**).

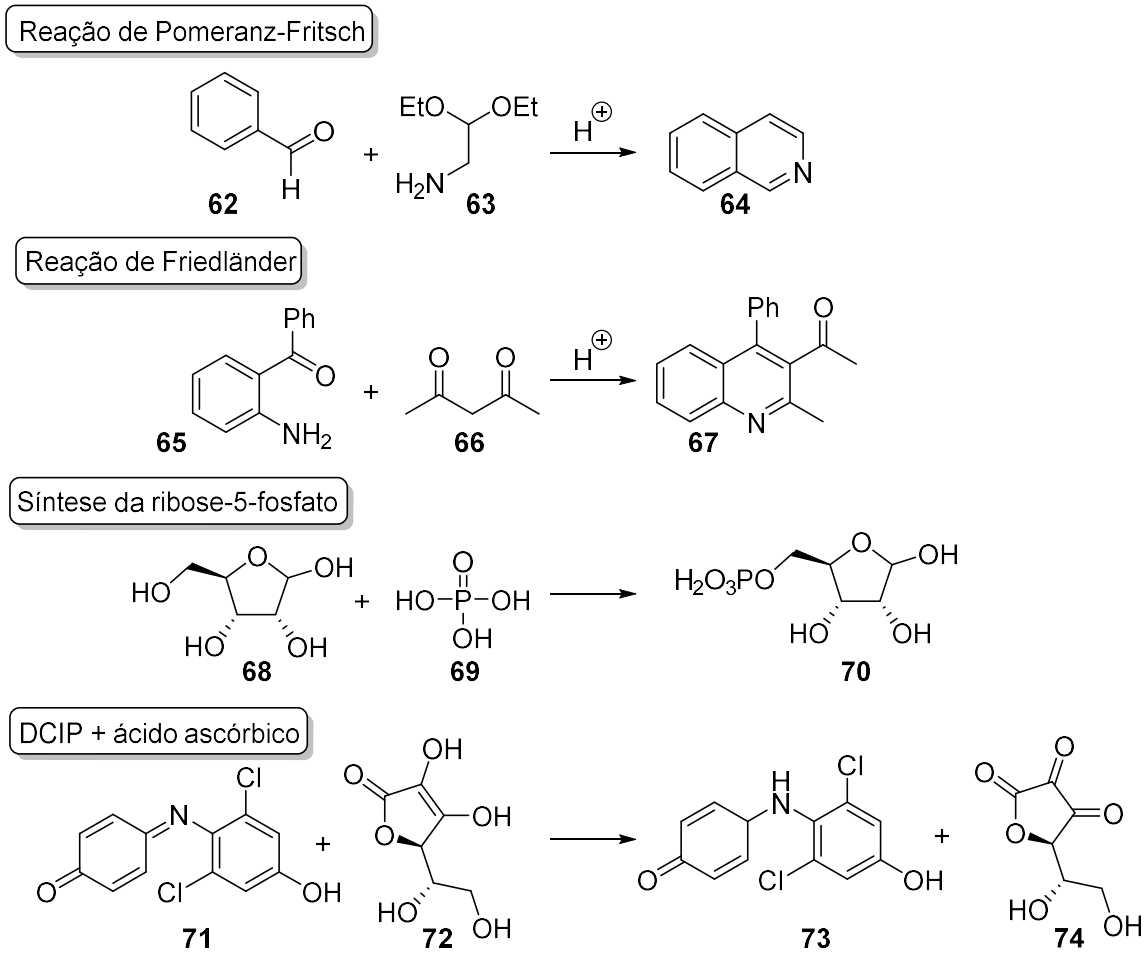


Esquema 10 – Reação de Buchwald-Hartwig utilizada por Ahneman e colaboradores na elaboração de um modelo baseado em aprendizado de máquina para predizer rendimentos.

Como base de dados foram utilizados os rendimentos de 4608 reações realizadas por *high-throughput experimentation* (HTE), utilizando diferentes combinações entre 15 haletos, 4 ligantes de Buchwald, 3 bases e 23 isoxazóis. Foram escolhidas como variáveis para o modelo 120 descritores vibracionais, moleculares e atômicos calculados utilizando DFT (B3LYP/6-31G*) e utilizados para representar os reagentes. Diversos algoritmos de aprendizado supervisionado foram avaliados. Dentre eles o *Random Forest* foi o que apresentou melhor performance quando comparado a regressão linear (R^2 0,92 e 0,67 respectivamente). O modelo apresentou excelente capacidade de modelar os rendimentos e além disso conseguiu identificar a melhor combinação de bases e ligantes para reações contendo isoxazóis e haletos de arila não utilizados no conjunto de treino.

Zhou e colaboradores[57] utilizaram um método de aprendizado por reforço para otimizar reações ocorrendo em microvolumes[58] gerados por ionização em electrospray (ESI). Ao contrário dos outros trabalhos, esse não empregou nenhuma base de dados químicos para o treinamento do modelo. Na realidade, o conjunto de treino foram "reações simuladas", representadas por uma mistura de funções gaussianas para gerar uma superfície de resposta contendo vários mínimos e máximos locais. O procedimento de otimização ocorre de maneira iterativa. Inicialmente, as condições reacionais (taxa de fluxo, voltagem e pressão do ionizador) são definidas aleatoriamente e a reação é realizada gerando o rendimento inicial. Em seguida, o algoritmo de ML, uma rede neural recorrente, utiliza esse e todos os outros rendimentos que forem obtidos posteriormente para definir novas condições reacionais com o objetivo de maximizar o rendimento ao longo do tempo. O modelo foi testado em quatro reações, a síntese de isoquinolina de Pomeranz-Fritsch, a síntese de quinolina de Friedländer, síntese da ribose fostato e a reação entre 2,6-diclorofenolindolfenol (DCIP) e ácido arcórbico (**Esquema 11**). Esse sistema apresentou excelente performance em comparação aos outros algoritmos testados, otimizando as

reações em 30 minutos após 40 iterações. O modelo também apresentou boa performance na otimização da síntese de nanopartículas de prata, mostrando assim que o método pode ser estendido para reações ocorrendo em volumes usuais.



Esquema 11 – Reações que foram otimizadas utilizando *Deep Reinforcement Learning* como proposto por Zhou e colaboradores.

4 - CONSIDERAÇÕES FINAIS

Em suma, os exemplos apresentados permitem demonstrar que os algoritmos de aprendizado de máquina podem ser empregados para gerar modelos capazes de predizer caminhos sintéticos e promover a otimização de condições reacionais, que são problemas de interesse da síntese orgânica. Além disso, tais modelos podem ser empregados juntamente com ferramentas bem estabelecidas como os cálculos DFT e técnicas experimentais para auxiliar no estudo do mecanismo de reações orgânicas. Existem, entretanto, muitas dificuldades a serem superadas para que essa tecnologia possa ser amplamente utilizada. Uma delas é a necessidade de melhorar as técnicas de representação molecular, muitas das quais não lidam corretamente com centros estereogênicos. Além disso, apesar da imensa disponibilidade de reações químicas nos bancos de dados públicos e comerciais, a extração de dados da literatura para abastecer os algoritmos de aprendizado de máquina é um procedimento difícil que demanda muito esforço para normalização e organização dessas informações. Por fim, a tendência em publicar apenas os bons resultados prejudica o desenvolvimento desses modelos, uma vez que a informação das reações que não funcionam é tão valiosa quanto saber quais transformações levam aos maiores rendimentos.

5 - REFERÊNCIAS

1 KOHN, K.; MORAES, C. H. D. **O impacto das novas tecnologias na sociedade: conceitos e características da Sociedade da Informação e da Sociedade Digital**. XXX Congresso Brasileiro de Ciências da Comunicação. Santos**:** 1-15 p. 2007.

2 CAMPBELL-KELLY, M. **Computer: A History of the Information Machine**. Taylor & Francis, 2018. ISBN 9780429975004.

3 BOLCER, J. D.; HERMANN, R. B. The Development of Computational Chemistry in the United States. In: (Ed.). **Reviews in Computational Chemistry**, v.V, 1994. p.1-63.

4 MULLIKEN, R. S. Spectroscopy, Molecular Orbitals, and Chemical Bonding (Nobel Lecture). 1966. Disponível em: < <https://www.nobelprize.org/prizes/chemistry/1966/mulliken/lecture/> >.

5 SCHERR, C. W. An SCF LCAO MO Study of N2. **The Journal of Chemical Physics,** v. 23, n. 3, p. 569-578, 1955.

6 BACHRACH, S. M. Brief History of Applied Theoretical Organic Chemistry. In: (Ed.). **Applied Theoretical Organic Chemistry**, 2018. p.69-95.

7 SEEMAN, J. I. Woodward–Hoffmann's Stereochemistry of Electrocyclic Reactions: From Day 1 to the JACS Receipt Date (May 5, 1964 to November 30, 1964). **The Journal of Organic Chemistry,** v. 80, n. 23, p. 11632-11671, 2015.

8 NOBELPRIZE.ORG. The Nobel Prize in Chemistry 1981. 2020. Disponível em: < https://www.nobelprize.org/prizes/chemistry/1981/summary/ >.

9 BARTON, D. H. R. The conformation of the steroid nucleus. **Experientia,** v. 6, n. 8, p. 316-320, 1950.

10 POLTEV, V. Molecular Mechanics: Principles, History, and Current Status. In: LESZCZYNSKI, J. (Ed.). **Handbook of Computational Chemistry**. Dordrecht: Springer Netherlands, 2016. p.1-48. ISBN 978-94-007-6169-8.

11 **Uses of Electronic Computers in Chemistry**. Washington, DC: The National Academies Press, 1967. 35 Disponível em: < https://www.nap.edu/catalog/21329/uses-of-electronic-computers-in-chemistry >.

12 CAS. A Division of American Chemical Society. Disponível em: < https://www.cas.org/ >.

13 WARR, W. A. Representation of chemical structures. **WIREs Computational Molecular Science,** v. 1, n. 4, p. 557-579, 2011.

14 WARR, W. A. A Short Review of Chemical Reaction Database Systems, Computer-Aided Synthesis Design, Reaction Prediction and Synthetic Feasibility. **Molecular Informatics,** v. 33, n. 6-7, p. 469-476, 2014.

[15] CHEN, W. L.; CHEN, D. Z.; TAYLOR, K. T. Automatic reaction mapping and reaction center detection. **WIREs Computational Molecular Science,** v. 3, n. 6, p. 560-593, 2013.

[16] RAYMOND, J. W.; WILLETT, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. **Journal of Computer-Aided Molecular Design,** v. 16, n. 7, p. 521-533, 2002.

[17] JAWORSKI, W. et al. Automatic mapping of atoms across both simple and complex chemical reactions. **Nature Communications,** v. 10, n. 1, p. 1434, 2019.

[18] JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science,** v. 349, n. 6245, p. 255-260, 2015.

[19] CHEN, M.; MAO, S.; LIU, Y. Big Data: A Survey. **Mobile Networks and Applications,** v. 19, n. 2, p. 171-209, 2014.

[20] COLEY, C. W.; EYKE, N. S.; JENSEN, K. F. Autonomous Discovery in the Chemical Sciences Part II: Outlook. **Angewandte Chemie International Edition**, 2020.

[21] RUSSELL, S. J. et al. **Artificial Intelligence: A Modern Approach**. Prentice Hall, 2010. ISBN 9780136042594.

[22] GUTTAG, J. **Introduction to Computation and Programming Using Python: With Application to Understanding Data**. MIT Press, 2016. ISBN 9780262529624.

[23] HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition**. Springer New York, 2009. ISBN 9780387848587.

[24] MÜLLER, A. C.; GUIDO, S. **Introduction to Machine Learning with Python: A Guide for Data Scientists**. O'Reilly Media, 2016. ISBN 9781449369897.

[25] MATER, A. C.; COOTE, M. L. Deep Learning in Chemistry. **Journal of Chemical Information and Modeling,** v. 59, n. 6, p. 2545-2559, 2019.

[26] COLEY, C. W.; GREEN, W. H.; JENSEN, K. F. Machine Learning in Computer-Aided Synthesis Planning. **Accounts of Chemical Research,** v. 51, n. 5, p. 1281-1289, 2018.

[27] MEDLEY, J. W.; MOVASSAGHI, M. Robinson's landmark synthesis of tropinone. **Chemical Communications,** v. 49, n. 92, p. 10775-10777, 2013.

[28] ROBINSON, R. LXIII.—A synthesis of tropinone. **Journal of the Chemical Society, Transactions,** v. 111, p. 762-768, 1917.

[29] MCCALLUM, J. E. **Military Medicine: From Ancient Times to the 21st Century**. ABC-CLIO, 2008. ISBN 9781851096930.

[30] COREY, E. J. General methods for the construction of complex molecules. **Pure and Applied Chemistry** v. 14, n. 1, p. 19, 1967.

[31] COOK, A. et al. Computer-aided synthesis design: 40 years on. **WIREs Computational Molecular Science,** v. 2, n. 1, p. 79-107, 2012.

[32] DOLATA, D. P. Artificial Intelligence in Chemistry. In: (Ed.). **Encyclopedia of Computational Chemistry**, 1998.

[33] LINDSAY, R. K. et al. DENDRAL: A case study of the first expert system for scientific hypothesis formation. **Artificial Intelligence,** v. 61, n. 2, p. 209-261, 1993.

[34] COREY, E. J.; WIPKE, W. T. Computer-Assisted Design of Complex Organic Syntheses. **Science,** v. 166, n. 3902, p. 178, 1969.

[35] TODD, M. H. Computer-aided organic synthesis. **Chemical Society Reviews,** v. 34, n. 3, p. 247-266, 2005.

[36] EVANS, D. A. History of the Harvard ChemDraw Project. **Angewandte Chemie International Edition,** v. 53, n. 42, p. 11140-11145, 2014.

[37] GUND, P. et al. Computer-Assisted Synthetic Analysis at Merck. **Journal of Chemical Information and Computer Sciences,** v. 20, n. 2, p. 88-93, 1980.

[38] DUGUNDJI, J.; UGI, I. **An algebraic model of constitutional chemistry as a basis for chemical computer programs**. Berlin, Heidelberg: Springer Berlin Heidelberg, 1973. 19-64 p.

[39] DE ALMEIDA, A. F.; MOREIRA, R.; RODRIGUES, T. Synthetic organic chemistry driven by artificial intelligence. **Nature Reviews Chemistry,** v. 3, n. 10, p. 589-604, 2019.

[40] SZYMKUĆ, S. et al. Computer-Assisted Synthetic Planning: The End of the Beginning. **Angewandte Chemie International Edition,** v. 55, n. 20, p. 5904-5937, 2016.

[41] KLUCZNIK, T. et al. Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. **Chem,** v. 4, n. 3, p. 522-532, 2018.

[42] GRZYBOWSKI, B. A. et al. Chematica: A Story of Computer Code That Started to Think like a Chemist. **Chem,** v. 4, n. 3, p. 390-398, 2018.

[43] LIU, B. et al. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. **ACS Central Science,** v. 3, n. 10, p. 1103-1113, 2017.

[44] WU, Y. et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. **CoRR**, 2016.

[45] LOWE, D. **Extraction of chemical structures and reactions from the literature**. 2012. (Tese de Doutorado). Departamento de Química, Universidade de Cambridge

[46] LIN, K. et al. Automatic retrosynthetic route planning using template-free models. **Chemical Science,** v. 11, n. 12, p. 3355-3364, 2020.

[47] SEGLER, M. H. S.; WALLER, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. **Chemistry – A European Journal,** v. 23, n. 25, p. 5966-5971, 2017.

[48] SEGLER, M. H. S.; PREUSS, M.; WALLER, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. **Nature,** v. 555, n. 7698, p. 604-610, 2018.

[49] COLEY, C. W. et al. Computer-Assisted Retrosynthesis Based on Molecular Similarity. **ACS Central Science,** v. 3, n. 12, p. 1237-1245, 2017.

[50] WEI, J. N.; DUVENAUD, D.; ASPURU-GUZIK, A. Neural Networks for the Prediction of Organic Chemistry Reactions. **ACS Central Science,** v. 2, n. 10, p. 725-732, 2016.

[51] ROGERS, D.; HAHN, M. Extended-Connectivity Fingerprints. **Journal of Chemical Information and Modeling,** v. 50, n. 5, p. 742-754, 2010.

[52] COLEY, C. W. et al. Prediction of Organic Reaction Outcomes Using Machine Learning. **ACS Central Science,** v. 3, n. 5, p. 434-443, 2017.

[53] GAO, H. et al. Using Machine Learning To Predict Suitable Conditions for Organic Reactions. **ACS Central Science,** v. 4, n. 11, p. 1465-1476, 2018.

[54] WALKER, E. et al. Learning To Predict Reaction Conditions: Relationships between Solvent, Molecular Structure, and Catalyst. **Journal of Chemical Information and Modeling,** v. 59, n. 9, p. 3645-3654, 2019.

[55] AHNEMAN, D. T. et al. Predicting reaction performance in C–N cross-coupling using machine learning. **Science,** v. 360, n. 6385, p. 186, 2018.

[56] COLLINS, K. D.; GLORIUS, F. A robustness screen for the rapid assessment of chemical reactions. **Nature Chemistry,** v. 5, n. 7, p. 597-601, 2013.

[57] ZHOU, Z.; LI, X.; ZARE, R. N. Optimizing Chemical Reactions with Deep Reinforcement Learning. **ACS Central Science,** v. 3, n. 12, p. 1337-1344, 2017.

[58] YAN, X.; BAIN, R. M.; COOKS, R. G. Organic Reactions in Microdroplets: Reaction Acceleration Revealed by Mass Spectrometry. **Angewandte Chemie International Edition,** v. 55, n. 42, p. 12960-12972, 2016.