

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE GOIÁS
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
CÂMPUS GOIÂNIA
MESTRADO PROFISSIONAL EM TECNOLOGIA, GESTÃO E SUSTENTABILIDADE

DHIULLY PRISCILLA SOUSA SILVA

Utilização de modelos de *machine learning* para prever resultados de reações químicas em aplicações ambientais

Goiânia, 2026.



INSTITUTO FEDERAL
Goiás

MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
SISTEMA INTEGRADO DE BIBLIOTECAS

TERMO DE AUTORIZAÇÃO PARA DISPONIBILIZAÇÃO NO REPOSITÓRIO DIGITAL DO IFG - ReDi IFG

Com base no disposto na Lei Federal nº 9.610/98, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia de Goiás, a disponibilizar gratuitamente o documento no Repositório Digital (ReDi IFG), sem ressarcimento de direitos autorais, conforme permissão assinada abaixo, em formato digital para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IFG.

Identificação da Produção Técnico-Científica

- | | |
|--|---|
| ☐ Tese | ☐ Artigo Científico |
| ☒ Dissertação | ☐ Capítulo de Livro |
| ☐ Monografia – Especialização | ☐ Livro |
| ☐ TCC - Graduação | ☐ Trabalho Apresentado em Evento |
| ☐ Produto Técnico e Educacional - Tipo: _____ | |

Nome Completo do Autor: DHIULLY PRISCILLA SOUSA SILVA

Matrícula: 20232011280310

Título do Trabalho: UTILIZAÇÃO DE MODELOS DE *MACHINE LEARNING* PARA PREVER RESULTADOS DE REAÇÕES QUÍMICAS COM APLICAÇÕES AMBIENTAIS

Autorização - Marque uma das opções

1. ☒ Autorizo disponibilizar meu trabalho no Repositório Digital do IFG (acesso aberto);
2. ☐ Autorizo disponibilizar meu trabalho no Repositório Digital do IFG somente após a data ____/____/____ (Embargo);
3. ☐ Não autorizo disponibilizar meu trabalho no Repositório Digital do IFG (acesso restrito).

Ao indicar a opção **2 ou 3**, marque a justificativa:

- ☐ O documento está sujeito a registro de patente.
☐ O documento pode vir a ser publicado como livro, capítulo de livro ou artigo.
☐ Outra justificativa: _____

DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O/A referido/a autor/a declara que:

- i. I. o documento é seu trabalho original, detém os direitos autorais da produção técnico-científica e não infringe os direitos de qualquer outra pessoa ou entidade;
- ii. II. obteve autorização de quaisquer materiais inclusos no documento do qual não detém os direitos de autor/a, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia de Goiás os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;
- iii. III. cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia de Goiás.



Documento assinado digitalmente

DHIULLY PRISCILLA SOUSA SILVA

Data: 01/07/2026 14:55:09-0300

Verifique em <https://validar.iti.gov.br>

Goiânia, 22/06/2026.

Local

Data

Assinatura do Autor e/ou Detentor dos Direitos Autorais

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE GOIÁS
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
CÂMPUS GOIÂNIA
MESTRADO PROFISSIONAL EM TECNOLOGIA, GESTÃO E SUSTENTABILIDADE

DHIULLY PRISCILLA SOUSA SILVA

Utilização de modelos de *machine learning* para prever resultados de reações químicas em aplicações ambientais

Programa de Pós-Graduação *Stricto Sensu* em Tecnologia, Gestão e Sustentabilidade do Instituto Federal de Educação, Ciência e Tecnologia de Goiás - Dissertação de Mestrado Profissional. Área de Concentração: Tecnologia de Sistemas de Produção Limpa. Linha de Pesquisa: Tecnologia e Gerenciamento de Resíduos.

Orientador: Prof. Dr. Fernando Schimidt

Coorientador: Prof. Dr. Raphael de Aquino Gomes

Goiânia, 2026.

S586u Silva, Dhiully Priscilla Sousa.

Utilização de modelos de machine learning para prever resultados de reações químicas em aplicações ambientais / Dhiully Priscilla Sousa Silva. – Goiânia: Instituto Federal de Educação, Ciência e Tecnologia de Goiás, 2026.
82f.

Orientação: Prof. Dr. Fernando Schmidt.
Coorientação: Prof. Dr. Raphael de Aquino Gomes.

Dissertação (mestrado) – Programa de Pós-graduação *Stricto Sensu* em Tecnologia, Gestão e Sustentabilidade, Instituto Federal de Educação, Ciência e Tecnologia de Goiás.

1. Compostos orgânicos voláteis. 2. Poluentes atmosféricos. 3. Quimioinformática. 4. Simulação computacional. 5. Aprendizado de máquina. I. Schmidt, Fernando (orientação). II. Gomes, Raphael de Aquino (coorientação). III. Instituto Federal de Educação, Ciência e Tecnologia de Goiás. IV. Título.

CDD 540.285

Ficha catalográfica elaborada pela bibliotecária Lana Cristina Dias Oliveira CRB1/ 2.631
Biblioteca Professor Jorge Félix de Souza,
Instituto Federal de Educação, Ciência e Tecnologia de Goiás, Câmpus Goiânia.



INSTITUTO FEDERAL
Goiás

MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE GOIÁS
CÂMPUS GOIÂNIA

PARECER 61/2026 - GYN-CMTGS/GYN-GPPEX/CP-GOIANIA/IFG

**SERVIÇO PÚBLICO FEDERAL
INSTITUTO FEDERAL DE EDUCAÇÃO CIÊNCIA E TECNOLOGIA DE GOIÁS
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
IFG CÂMPUS GOIÂNIA
MESTRADO EM TECNOLOGIA, GESTÃO E SUSTENTABILIDADE**

DHIULLY PRISCILLA SOUSA SILVA

**UTILIZAÇÃO DE MODELOS DE *MACHINE LEARNING* PARA PREVER RESULTADOS DE REAÇÕES
QUÍMICAS COM APLICAÇÕES AMBIENTAIS**

Dissertação apresentada ao Programa de Pós-Graduação em Tecnologia,
Gestão e Sustentabilidade do Instituto Federal de Educação, Ciência e
Tecnologia de Goiás como requisito para a obtenção do título de Mestre.
Área de Concentração: Tecnologia de Sistemas de Produção Limpa.
Linha de Pesquisa: Tecnologia e Gerenciamento de Resíduos.

Aprovação em: 23 de abril de 2026.

Banca examinadora:

Presidente e orientador: Prof. Dr. Fernando Schmidt - PPGTGS IFG

Coorientador: Prof. Dr. Raphael de Aquino Gomes - PPGTGS IFG

Avaliador: Prof. Dr. Paulo Roberto Filgueiras - UFES

Avaliador: Prof. Dr. Thiago Eduardo Pereira Alves - PPGTGS IFG

Documento assinado eletronicamente por:

- Paulo Roberto Filgueiras, Paulo Roberto Filgueiras - Docente - Ifg - Câmpus Goiânia (10870883000225), em 22/06/2026 12:11:24.
- Thiago Eduardo Pereira Alves, PROFESSOR ENS BASICO TECN TECNOLÓGICO, em 04/06/2026 10:04:04.
- Raphael de Aquino Gomes, PROFESSOR ENS BASICO TECN TECNOLÓGICO, em 30/05/2026 14:02:01.
- Fernando Schmidt, PROFESSOR ENS BASICO TECN TECNOLÓGICO, em 26/05/2026 14:49:45.

Este documento foi emitido pelo SUAP em 13/04/2026. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse <https://suap.ifg.edu.br/autenticar-documento/> e forneça os dados abaixo:

Código Verificador: 766878
Código de Autenticação: 534a75969b



Instituto Federal de Educação, Ciência e Tecnologia de Goiás
Rua 75, nº 46, Centro, GOIÂNIA / GO, CEP 74055-110
Sem Telefones cadastrados

Dedicatória

Dedico este trabalho a todas as mulheres que vieram antes de nós, abrindo caminhos mesmo diante das dificuldades, e às que hoje podem sonhar, escolher e ocupar qualquer espaço que desejarem. Que nunca nos falte coragem para continuar.

Agradecimentos

Agradeço, primeiramente, ao ser superior que torna todas as coisas possíveis.

Ao meu orientador, Prof. Dr. Fernando Schmidt, minha profunda admiração, confiança e gratidão. Obrigada pela orientação, pela paciência e por ser uma inspiração acadêmica e profissional ao longo desta caminhada.

Ao meu coorientador, Prof. Dr. Raphael de Aquino Gomes, por ter ampliado meus horizontes, apresentado novos desafios e contribuído de forma tão significativa para o desenvolvimento desta pesquisa. Sua participação foi uma grata e valiosa adição a este trabalho.

À minha família, por toda a paciência, incentivo e apoio incondicional durante essa trajetória. Em especial às minhas irmãs, Daniele e Laura, que nunca me deixaram duvidar da minha capacidade, sempre acreditando que eu era capaz de conquistar tudo aquilo que sonhasse.

Aos meus amigos, por terem sido leveza, acolhimento e alívio nos dias mais difíceis, tornando essa caminhada mais suportável e feliz.

Por fim, agradeço a todos que, de alguma forma, contribuíram para que este momento se tornasse possível.

“Palavras são, na minha nada humilde opinião, nossa inesgotável fonte de magia.”

J.K. Rowling — Harry Potter e as Relíquias da Morte,
(1965-).

Resumo

Título: Utilização de modelos de *machine learning* para prever resultados de reações químicas em aplicações ambientais

Autor: Dhiully Priscilla Sousa Silva

Orientador: Dr. Fernando Schmidt

Coorientador: Prof. Dr. Raphael de Aquino Gomes

Os compostos orgânicos voláteis (COVs), emitidos por fontes naturais e antrópicas, participam de processos fotoquímicos complexos na atmosfera, influenciando a qualidade do ar e o clima por meio da formação de oxidantes e aerossóis secundários. A compreensão de suas transformações químicas e de seus potenciais impactos ambientais permanece um desafio relevante para o desenvolvimento de estratégias de monitoramento e controle. Neste trabalho, foram aplicadas ferramentas de quimi-informática e modelos de *Machine Learning* para simular a formação de produtos de oxidação a partir de moléculas presentes no diesel combustível e avaliar seu potencial toxicológico. Reações atmosféricas específicas envolvendo radical hidroxila (OH), ozônio (O₃) e oxigênio atômico (O) foram modeladas em ambiente Python com auxílio da biblioteca RDKit. A partir das estruturas geradas, descritores moleculares foram calculados e utilizados na construção de modelos de classificação para predição de mutagenicidade, com base em dados do teste de Ames. Entre os algoritmos avaliados, o modelo Extra Tree Classifier apresentou melhor desempenho em comparação à rede neural profunda implementada em Keras, demonstrando maior robustez diante do desbalanceamento da base de dados. Os resultados evidenciam que a integração entre simulação computacional de reações atmosféricas e modelagem preditiva constitui uma estratégia consistente para a investigação e a triagem de compostos derivados de combustíveis fósseis sob a perspectiva da Química Ambiental.

Palavras-chave

reações químicas; COVs, RDKit, poluentes.

Abstract

Title: Using machine learning models to predict the outcomes of chemical reactions in environmental applications

Author: Dhiully Priscilla Sousa Silva

Advisor: Dr. Fernando Schimidt

Coadvisor: Prof. Dr. Raphael de Aquino Gomes

Volatile organic compounds (VOCs), emitted by natural and anthropogenic sources, participate in complex photochemical processes in the atmosphere, influencing air quality and climate by forming oxidants and secondary aerosols. Understanding their chemical transformations and potential environmental impacts remains a relevant challenge for the development of monitoring and control strategies. In this work, chemoinformatics tools and machine learning models were applied to simulate the formation of oxidation products from molecules present in diesel fuel and to evaluate their toxicological potential. Specific atmospheric reactions involving hydroxyl radicals (OH), ozone (O₃), and atomic oxygen (O) were modeled in a Python environment using the RDKit library. From the generated structures, molecular descriptors were calculated and used to build classification models for mutagenicity prediction, based on Ames test data. Among the algorithms evaluated, the Extra Tree Classifier achieved better performance than the deep neural network implemented in Keras, demonstrating greater robustness to database imbalance. The results show that integrating computational simulation of atmospheric reactions with predictive modeling constitutes a consistent strategy for investigating and screening compounds derived from fossil fuels from an Environmental Chemistry perspective.

Keywords

chemical reactions; VOCs, RDKit, pollutants.

Lista de Figuras

Figura 1	– Reação orgânica com formação de benzeno (por eliminação) a partir do hexano, com o produto intermediário cicloexano.	22
Figura 2	– Estrutura molecular dos HPAs benzo[a]pireno, antraceno e fluoranteno	23
Figura 3	– Fluxo geral de funcionamento de um modelo de aprendizado de máquina.	31
Figura 4	– Exemplo simplificado de uma árvore de decisão utilizada em tarefas de classificação.	32
Figura 5	– Estrutura básica de uma rede neural artificial.	33
Figura 6	– Pipeline computacional para simulação de reações atmosféricas e avaliação de toxicidade de moléculas do diesel	41
Figura 7	– Parte do painel principal do software Marvin Sketch, mostrando a reação de ataque (#2) à dupla ligação de alcenos alifáticos, via radical OH.	51
Figura 8	– Distribuição do número de átomos de carbono entre as 261 moléculas geradas pelo molecular generator.	56
Figura 9	– Distribuição do peso molecular das 261 moléculas geradas pelo molecular generator.	57
Figura 10	– Distribuição de aromaticidade e classificação das 261 moléculas geradas pelo molecular generator	57
Figura 11	– Matriz de confusão obtida para o conjunto de teste, no modelo Extra Trees Classifier.	60
Figura 12	– Função de perda em relação ao número de épocas considerando o recurso <i>Early Stopping</i>	63
Figura 13	– Matriz de confusão obtida para o conjunto de teste, no modelo de Aprendizado Profundo.	64
Figura 14	– Ataque nas ligações duplas das cadeias aromáticas	78
Figura 15	– Ataque às ligações duplas das cadeias alifáticas.	79
Figura 16	– Ataque às ligações duplas de cadeias alifáticas.	79
Figura 17	– Ataque nas ligações duplas das cadeias aromáticas.	80
Figura 18	– Reações com cadeias alifáticas, mecanismo A.	80
Figura 19	– Reações com cadeias alifáticas, mecanismo B.	81
Figura 20	– Reações envolvendo cadeias alifáticas, mecanismo C.	81
Figura 21	– Ataque às ligações simples em alcanos.	82
Figura 22	– Ataque às ligações simples em alcanos.	82
Figura 23	– Ataque aos hidrocarbonetos aromáticos.	83

Lista de Tabelas

Tabela 1	– Exemplos de moléculas presentes no diesel e suas representações SMILES.	47
Tabela 2	– Número de produtos únicos produzidos nas reações químicas estudadas considerando as 261 moléculas de reagentes.	54
Tabela 3	– Métricas de avaliação no conjunto de teste do modelo Extra Trees Classifier.	59
Tabela 4	– Métricas de avaliação no conjunto de teste do modelo de Aprendizado Profundo.	62

Lista de Quadros

Quadro 1	– Lista dos HPAs analisados.	42
Quadro 2	– Alguns exemplos de moléculas reagentes e produtos, no formato SMILES, da reação química da Figura 7 (Reação #2)	53

Lista de Códigos de Programas

Código 1	– Exemplo de aplicação de funções do RDKit	38
----------	--	----

Lista de Abreviaturas e Siglas

CAS	<i>Chemical Abstracts Service</i>
COAs	Compostos Orgânicos Aromáticos
COVs	Compostos Orgânicos Voláteis
DNN	<i>Deep Neural Networks</i>
HPAs	Hidrocarbonetos Policíclicos Aromáticos
IA	Inteligência Artificial
InChI	<i>International Chemical Identifier</i>
ISSSTY	<i>In vitro mutagenicity in Salmonella TYphimurium</i>
ML	<i>Machine Learning</i>
OMS	Organização Mundial da Saúde
PAHs	<i>Polycyclic Aromatic Hydrocarbons</i>
PCA	<i>Principal Component Analysis</i>
QSAR	<i>Quantitative Structure-Activity Relationship</i>
SDF	<i>Structure Data File</i>
SMARTS	<i>SMiles ARbitrary Target Specification</i>
SMILES	<i>Simplified Molecular Input Line Entry System</i>
USEPA	<i>United States Environmental Protection Agency</i>
USPTO	<i>United States Patent and Trademark Office</i>

Sumário

1	Introdução	18
2	Revisão Bibliográfica	21
2.1	Poluentes atmosféricos	21
2.1.1	Formação de <i>smog</i> com origem em Ox, NOx, hidrocarbonetos e luz solar	23
2.2	Reações Orgânicas	25
2.2.1	Reações de Adição	27
2.2.2	Reação de Eliminação	28
2.2.3	Reações de Oxidação	28
2.2.4	Reações de Substituição	30
2.3	<i>Machine Learning</i>	31
2.3.1	<i>Pipeline</i> de desenvolvimento de modelos de ML	34
2.3.2	<i>Machine Learning</i> na Química	35
2.3.3	Biblioteca RDKit	36
2.3.4	Codificação SMILES	38
2.3.5	Codificação SMARTS	39
3	Metodologia	40
3.1	Simulação de Reações Atmosféricas	41
3.1.1	Seleção das moléculas representativas do diesel	41
3.1.2	Representação molecular e codificação química	43
3.2	Geração molecular via redes generativas e algoritmo genético	43
3.3	Desenvolvimento dos modelos de classificação de toxicidade	44
3.3.1	Tratamento de dados	44
3.3.2	Arquitetura dos modelos e treinamento	45
4	Resultados e Discussão	46
4.1	Modelagem de reações atmosféricas via SMARTS e RDKit	46
4.1.1	Reações atmosféricas	50
4.2	Gerador Molecular Computacional (<i>Molecular Generator</i>)	55
4.3	Modelos de toxicidade	58
4.3.1	Modelo Extra Trees Classifier	59
4.3.2	Modelo de Aprendizado Profundo	61
4.4	Considerações finais do capítulo	64
5	Conclusão	65
	Referências	67

Apêndice A – Padrões de reação e códigos SMARTS	78
---	----

Introdução

O desenvolvimento industrial e as matrizes energéticas globais, fortemente dependentes da queima de combustíveis fósseis, resultaram em um aumento sem precedentes das emissões de compostos químicos na atmosfera. Essa emissão massiva, proveniente em grande parte de fontes veiculares, é a principal causa da poluição do ar, um dos maiores desafios ambientais e de saúde coletiva do século XXI, impactando diretamente ecossistemas e populações em escala global (Vormittag *et al.*, 2021). Dentre a vasta gama de substâncias liberadas, os compostos orgânicos, em especial os de natureza aromática presentes na composição desses combustíveis, merecem atenção particular devido à sua alta reatividade com agentes oxidantes atmosféricos, e aos seus efeitos nocivos (Venkatraman *et al.*, 2024; Dai *et al.*, 2023).

Neste grupo de substâncias, destacam-se os Hidrocarbonetos Policíclicos Aromáticos (HPAs), também conhecidos como *Polycyclic Aromatic Hydrocarbons* (PAHs), como o antraceno, cuja presença no ambiente representa uma preocupação relevante devido às suas propriedades tóxicas e carcinogênicas. Os HPAs são compostos orgânicos caracterizados pela presença de dois ou mais anéis aromáticos, alta estabilidade estrutural e notável persistência no meio ambiente, o que favorece sua bioacumulação (Venkatraman *et al.*, 2024). Os HPAs são considerados poluentes ambientais e podem ser emitidos tanto por fontes primárias, que os lançam diretamente no ambiente, quanto por fontes secundárias, nas quais reações atmosféricas formam novos compostos poluentes (Wu *et al.*, 2024).

O entendimento e a previsão dessas reações secundárias são importantes, pois ajudam a esclarecer as dinâmicas de formação e dispersão desses poluentes no ar atmosférico urbano (Zhou *et al.*, 2021), em que a concentração de Compostos Orgânicos Aromáticos (COAs) é elevada, o que contribui para a formação de *smog* fotoquímico. Este fenômeno, caracterizado pela geração de oxidantes, como o ozônio troposférico (O₃), e de material particulado fino (MP_{2.5}), está diretamente associado a uma série de agravos à saúde coletiva, incluindo o aumento da incidência de doenças respiratórias e cardiovasculares (Souza *et al.*, 2023a).

Tradicionalmente, a investigação sobre o destino de compostos orgânicos na

atmosfera, como os liberados pela queima de diesel, apoia-se em duas abordagens principais: a experimental e a computacional (Martins, 2007; Gavidia-Calderón *et al.*, 2024). A abordagem experimental, por meio de monitoramento em campo e simulações em laboratório, é fundamental para obter dados concretos sobre os produtos e as taxas de reação (Kim *et al.*, 2024; Chen *et al.*, 2023). Por sua vez, os modelos computacionais pioneiros nesta área eram, em grande parte, determinísticos, baseados em um conjunto limitado de regras de reatividade conhecidas ou em correlações estatísticas mais simples entre a estrutura molecular e seu comportamento (Liu *et al.*, 2023; Meylan; Howard, 2003; Mallet; Sportisse, 2008). Juntas, essas estratégias formaram o alicerce para a compreensão do problema.

Tendo em vista que a melhora da qualidade do ar depende de uma redução das emissões e da formação subsequente de poluentes (Gentner *et al.*, 2017), o principal desafio científico reside no fato de ainda haver falta de compreensão sobre as práticas específicas de formação secundária desses poluentes (Cruz *et al.*, 2019). Essa dificuldade se justifica pelas limitações das abordagens atuais. A previsão de reações enfrenta o obstáculo da grande diversidade de reações orgânicas que podem ocorrer na atmosfera, o que leva à formação de um número vasto e, muitas vezes, imprevisível de compostos (Barber; Green; Kroll, 2021). Tais barreiras, tanto experimentais quanto teóricas, criam uma lacuna para o desenvolvimento de metodologias preditivas que, ao mesmo tempo, sejam rápidas, acessíveis e precisas.

É neste cenário de limitações que a quimioinformática se estabelece como uma área estratégica, oferecendo um conjunto de métodos computacionais para análise e interpretação de dados químicos em larga escala (Niazi; Mariam, 2023). Dentre suas abordagens mais poderosas, destaca-se o Aprendizado de Máquina, ou *Machine Learning* (ML), um ramo da Inteligência Artificial (IA) focado no desenvolvimento de algoritmos que aprendem padrões diretamente dos dados (Zhou *et al.*, 2021). O sucesso de sua aplicação em reações químicas depende de dois fatores-chave: a disponibilidade de bases de dados robustas e o uso de descritores moleculares ou reacionais adequados, que traduzem a informação química em um formato que o modelo computacional possa compreender (Park *et al.*, 2022).

A partir dessa abordagem, o uso de ML para realizar previsões é capaz de aperfeiçoar, de forma eficiente e com menor geração de resíduos, as análises de formação de poluentes (Park *et al.*, 2022). Consequentemente, ela oferece um caminho mais rápido e preciso para a investigação científica, ao mesmo tempo em que contribui para reduzir o impacto ambiental causado pelas pesquisas experimentais tradicionais. Portanto, o desenvolvimento de modelos preditivos de ML para prever reações químicas ambientais representa uma alternativa inovadora e sustentável para a pesquisa em química (Park *et al.*, 2022; Silva; Evans, 2024).

Diante disso, esta dissertação insere-se na interface entre a quimioinformática e a química ambiental, com o objetivo de transformar o modo como se compreendem e se preveem reações complexas envolvendo poluentes. O caráter distintivo deste trabalho reside em investigar o emprego de métodos de aprendizado de máquina para antecipar a formação de compostos aromáticos, como o antraceno, a partir de reações com moléculas presentes no óleo diesel. Essa proposta configura uma alternativa inovadora e mais sustentável, que possibilita não apenas um maior aprofundamento na compreensão dos mecanismos reacionais, mas também uma avaliação mais rápida, precisa e com menor geração de resíduos dos impactos ambientais associados a essas moléculas.

Neste contexto, o objetivo geral desta dissertação é a aplicação de modelos de *Machine Learning* para prever a formação de compostos orgânicos a partir de moléculas (reagentes) presentes no diesel combustível. Para alcançar este propósito, definiram-se os seguintes objetivos específicos:

- Estudar reações químicas de oxidação específicas, utilizando os recursos computacionais da biblioteca RDKit, em linguagem Python;
- Analisar as moléculas selecionadas como reagentes em reações químicas com radical OH, ozônio e oxigênio atômico;
- Compilar e pré-processar bases de dados representativas contendo moléculas presentes no óleo diesel e seus respectivos produtos de reação atmosférica;
- Extrair e selecionar descritores moleculares e reacionais adequados, traduzindo as propriedades físico-químicas e estruturais dos compostos para o treinamento dos algoritmos;
- Construir modelos de classificação para prever a toxicidade dessas moléculas (com base em bases de dados de mutagênese).
- Analisar a importância dos recursos (descritores) nos modelos treinados, visando aprofundar a compreensão sobre os mecanismos reacionais e os fatores que governam a formação de poluentes secundários;
- Comparar a eficiência preditiva e computacional dos modelos desenvolvidos com as abordagens tradicionais (determinísticas e experimentais) descritas na literatura.

Sendo assim, o desenvolvimento da presente pesquisa, ao cumprir os objetivos definidos, apresenta-se como um vínculo entre a ciência de dados e a química ambiental. Com isso, a expectativa é oferecer não apenas um modelo preditivo funcional, mas também contribuir com uma nova metodologia para o estudo de reações complexas em sistemas ambientais (Park *et al.*, 2022; Silva; Evans, 2024). Os detalhes, o desenvolvimento e os resultados desta abordagem serão explorados nos capítulos seguintes.

Revisão Bibliográfica

Neste capítulo, serão abordados aspectos teóricos da literatura relevantes ao desenvolvimento do trabalho. A abordagem inclui a qualidade do ar em áreas urbanas, mecanismos de reações orgânicas, emissões de antraceno, bem como reações que podem originá-lo na atmosfera. Também serão exploradas as técnicas de aplicação de ML empregadas nos estudos químicos e os métodos de programação computacional utilizados.

2.1 Poluentes atmosféricos

A expressão “poluição do ar” normalmente se refere a elementos que, em níveis locais e regionais, impactam diretamente ou indiretamente os seres vivos, incluindo animais, plantas e pessoas. A qualidade do ar é fundamental para o bem-estar humano e ambiental, e há séculos há queixas sobre a qualidade do ar, principalmente nos centros urbanos (Spiro; Stigliani, 2008).

O ar atmosférico, composto majoritariamente por nitrogênio e oxigênio, pode conter traços de poluentes que comprometem sua pureza e prejudicam a saúde e os ecossistemas. Esses poluentes incluem gases tóxicos, partículas sólidas e líquidas, bem como compostos orgânicos e inorgânicos. Estudos da Organização Mundial da Saúde (OMS) indicam que mais de sete milhões de mortes anuais estão associadas à exposição a poluentes atmosféricos, evidenciando a gravidade do problema (OPAS, 2018). Além dos impactos à saúde humana, a poluição do ar também afeta os ecossistemas, causando a degradação de habitats, alterações climáticas e a perda de biodiversidade (Raposo; Assis; Norton, 2021). Por exemplo, a deposição de poluentes em corpos d’água pode provocar eutrofização, enquanto o aumento dos gases de efeito estufa contribui para o aquecimento global.

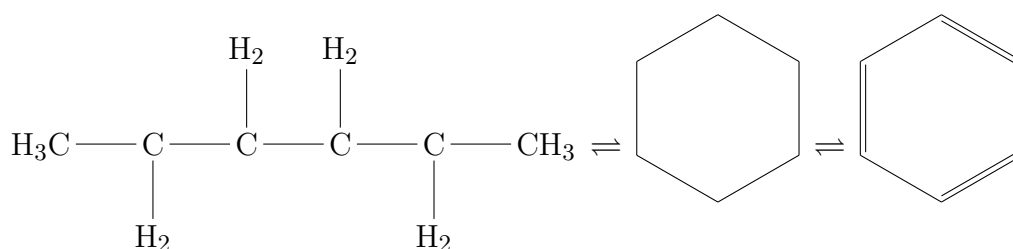
Os efeitos amplamente disseminados das emissões são cada vez mais evidentes, e a necessidade de controlá-las influencia, em maior grau, o desenvolvimento de tecnologias. Os principais poluentes de foco de controle são os óxidos de carbono (COx) e de nitrogênio (NOx), substâncias orgânicas tóxicas e material particulado

(Souza *et al.*, 2022). Entre as principais substâncias presentes nos poluentes orgânicos tóxicos, podem ser citados os aldeídos, os hidrocarbonetos aromáticos e os aromáticos policíclicos, sendo a formação desses últimos o foco desta pesquisa.

Os poluentes atmosféricos podem ser classificados em duas categorias principais: primários e secundários. Os poluentes primários são aqueles emitidos diretamente por fontes naturais ou antrópicas, como o dióxido de enxofre (SO₂), o monóxido de carbono (CO) e os óxidos de nitrogênio (NO_x). Esses poluentes são liberados por processos como a combustão de combustíveis fósseis, as erupções vulcânicas e as queimadas. Os poluentes secundários formam-se na atmosfera por meio de reações químicas entre poluentes primários e outros componentes atmosféricos como o ozônio troposférico (O₃), formado pela interação de óxidos de nitrogênio e Compostos Orgânicos Voláteis (COVs) na presença de luz solar, e os nitratos e sulfatos presentes no material particulado (Lenzi; Favero, 2009).

De acordo com Baird e Cann (2011), a maioria das partículas do ar urbano é secundária, portanto, sua concentração só pode ser controlada por meio da redução da emissão de gases poluentes primários de onde são criados. Os Hidrocarbonetos Policíclicos Aromáticos (HPAs) se formam como subprodutos da queima de combustíveis de carbono (Spiro; Stigliani, 2008); um exemplo é a formação do benzeno a partir do hexano, composto presente na gasolina, por meio de reação orgânica de eliminação, conforme apresentado na Figura 1.

Figura 1 – *Reação orgânica com formação de benzeno (por eliminação) a partir do hexano, com o produto intermediário ciclohexano.*

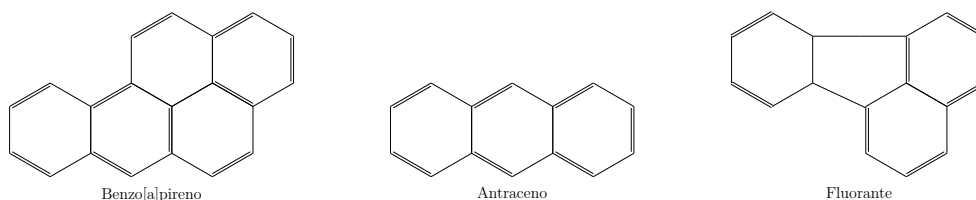


Fonte: Adaptado de (Baird; Cann, 2011).

Os HPAs constituem uma classe de compostos orgânicos altamente tóxicos, formados pela queima incompleta de materiais orgânicos. Esses compostos são frequentemente encontrados em áreas urbanas e industriais, onde a queima de combustíveis fósseis e de biomassa é intensa. Eles possuem estruturas químicas complexas compostas por dois anéis aromáticos condensados. Essa estrutura confere a esses compostos alta estabilidade química e elevada bioacumulação no ambiente. Substâncias como o benzo[a]pireno, o antraceno e o fluoranteno, apresentados na

Figura 2, são reconhecidas por seu potencial carcinogênico e mutagênico e são objeto de regulações ambientais em diversos países (Baird; Cann, 2011).

Figura 2 – Estrutura molecular dos HPAs benzo[a]pireno, antraceno e fluoranteno



Fonte: Elaborado pelos autores, com base em (Baird; Cann, 2011).

Os Compostos Orgânicos Voláteis (COVs) e os óxidos de nitrogênio (NOx) também são compostos primários que não são considerados poluentes diretos do ar por não oferecerem nenhum risco de forma direta a saúde humana. Porém, são os principais responsáveis pela formação do *smog* fotoquímico, a névoa de cor marrom que cobre muitas cidades no mundo. As três maiores fontes de emissão de COVs são os setores de processamento industrial, a utilização de solventes e os veículos rodoviários e não rodoviários (Lenzi; Favero, 2009).

O *smog* fotoquímico pode ser formado sempre que uma grande quantidade de gases de exaustão automotivos e industriais são confinados por uma camada de inversão térmica sobre uma localidade que seja, ao mesmo tempo, exposta ao sol. Ele se caracteriza por um acúmulo de fumaça marrom e nebulosa, que contém ozônio e outros agentes oxidantes e provoca efeitos prejudiciais à saúde humana. Embora a maior parte do dano causado pelo *smog* resulte da ação do ozônio e de outros oxidantes, estes não se acumulam sem a ação combinada entre NOx e COVs, uma vez que o controle da formação de *smog* fotoquímico requer a redução das emissões de NOx e COVs (Spiro; Stigliani, 2008).

2.1.1 Formação de *smog* com origem em Ox, NOx, hidrocarbonetos e luz solar

A formação de *smog* fotoquímico é um fenômeno complexo que decorre de uma série de reações químicas entre compostos presentes na atmosfera. Esse processo é impulsionado pela presença de luz solar, oxigênio molecular (O₂), óxidos de nitrogênio (NOx) e hidrocarbonetos voláteis (Spiro; Stigliani, 2008). Entender as bases químicas e ambientais desse processo é de suma importância para a prática de ações voltadas a mitigar a poluição atmosférica (Lima; Borges, 2021).

Os principais precursores do *smog* fotoquímico são o oxigênio atômico (O), óxidos de nitrogênio (NO e NO₂, frequentemente referidos como NOx) e compostos

orgânicos voláteis (COVs). Esses compostos reagem entre si sob a influência da radiação ultravioleta (UV) presente na luz solar, formando uma mistura de poluentes secundários, como o ozônio troposférico (O_3) e compostos orgânicos oxigenados, que compõem o *smog* (Silva; Arbilla, 2022).

O oxigênio molecular é essencial para a formação do *smog* pois atua como substrato na formação do ozônio troposférico. Em presença de luz UV, o oxigênio molecular é dissociado em átomos de oxigênio (O) (reação 1), que subsequentemente reagem para formar ozônio (O_3) (reação 2) (Spiro; Stigliani, 2008).

Fotólise do Oxigênio Molecular:



Formação do Ozônio:



Os óxidos de nitrogênio, derivados principalmente de processos de combustão, desempenham um papel crucial no ciclo fotoquímico. O NO_2 , em particular, é fotolisado pela radiação UV, formando óxido nítrico (NO) e um átomo de oxigênio reativo (O) (reação 3), que, posteriormente, contribui para a geração de ozônio (reações 4 e 5). Essas reações estabelecem um ciclo dinâmico entre NO, NO_2 e O_3 , fundamental para a manutenção do *smog* fotoquímico (Wark; Warner, 1980).

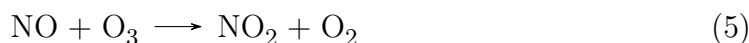
Fotólise do NO_2 :



Formação de Ozônio a partir do Oxigênio Atômico:



Reação entre NO e O_3 :



Os COVs, emitidos por fontes naturais e antropogênicas, como vegetação, combustíveis fósseis e solventes industriais, são oxidados na atmosfera, formando radicais de alquila e água (reação 6). O radical alquila ($R\bullet$) reage com O_2 para formar um radical peroxila ($RO_2\bullet$) (reação 7). Esses produtos promovem a oxidação do NO a NO_2 (reação 8), essencial para a reciclagem do ozônio no *smog* fotoquímico

(Spiro; Stigliani, 2008).

Oxidação de Hidrocarbonetos:



Formação de Peróxidos Radicais:



Oxidação de NO a NO₂:



O *smog* fotoquímico tem impactos significativos na saúde humana e no meio ambiente. A presença de ozônio troposférico e de outros poluentes secundários pode causar problemas respiratórios, irritação ocular e danos à vegetação. Além disso, o *smog* contribui para a formação de partículas finas (PM_{2,5}), associadas a doenças cardiovasculares e à redução da visibilidade atmosférica. A compreensão detalhada das reações envolvidas no processo de formação do *smog* é fundamental para o desenvolvimento de estratégias eficazes de controle da poluição atmosférica (Seinfeld; Pandis, 2016).

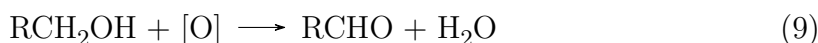
Optar por uma abordagem multidisciplinar pode ser uma atitude inteligente e eficaz para o controle da poluição atmosférica, pois a integração de diferentes abordagens é fundamental para enfrentar esse desafio. Modelos químicos atmosféricos permitem simular e prever a formação e a dispersão do *smog*, ajudando a identificar as principais fontes de poluentes e os mecanismos de reação envolvidos. Políticas públicas bem estruturadas são de grande importância para reduzir as emissões de poluentes (UNEP, 2021). O monitoramento ambiental é responsável por fornecer dados em tempo real sobre a qualidade do ar, permitindo a avaliação da eficácia das políticas implementadas e a detecção de áreas críticas que requerem intervenção. A combinação dessas ações, modelos químicos, monitoramento e políticas pode levar a uma redução significativa dos impactos negativos do *smog*, melhorando a qualidade do ar e protegendo a saúde coletiva e os ecossistemas.

2.2 Reações Orgânicas

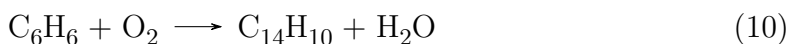
As reações orgânicas desempenham um papel importante na formação de poluentes atmosféricos, por meio de processos diretos ou indiretos, especialmente em ambientes urbanos. Essas reações são responsáveis pela transformação química de

compostos orgânicos voláteis (COVs) em produtos secundários que podem contribuir para a poluição do ar. Esses compostos são emitidos por processos industriais, pela combustão de combustíveis fósseis, como o diesel e a gasolina, e por fontes naturais. Eles passam por transformações químicas na atmosfera, o que pode gerar diversos poluentes. As reações mais comuns neste contexto incluem adição, eliminação, substituição e oxidação, que desempenham papéis fundamentais na formação de poluentes orgânicos (Spiro; Stigliani, 2008).

A oxidação é uma das reações mais comuns na atmosfera, envolvendo a interação entre COVs radicais de hidroxila ($\text{OH}^\bullet$), ozônio (O_3) e oxigênio atômico (O). Esse processo pode levar à formação de compostos oxigenados, como álcoois, aldeídos, cetonas e ácidos carboxílicos. A reação apresentada na Equação 9 mostra a oxidação de um álcool primário formando um aldeído. Tais compostos desempenham um papel significativo na formação de smog fotoquímico e de material particulado (Spiro; Stigliani, 2008).



Além disso, a oxidação dos COVs, como o benzeno, pode contribuir para a formação de hidrocarbonetos aromáticos policíclicos (PAHs), como o antraceno, conforme mostrado na reação representada na Equação 10. O antraceno, assim como outros PAHs, apresenta elevado potencial cancerígeno e é extremamente danoso à saúde humana (Solomons; Fryhle; Snyder, 2024).



A reação de adição é um tipo de reação orgânica simples que consiste na ruptura de uma ligação pi (π) e na adição de um átomo ou grupo de átomos substituinte para suprir a valência livre do carbono. É um tipo de reação em que moléculas simples se combinam para formar produtos ligeiramente mais complexos. Essas reações podem ocorrer na presença de catalisadores atmosféricos, como óxidos de nitrogênio (NO_x) e radicais livres, o que contribui para a formação de compostos orgânicos oxigenados e de peróxidos (Solomons; Fryhle; Snyder, 2024).

A reação de substituição consiste na troca de ligantes por outros. Trata-se de um tipo de reação em que um átomo ou grupo de átomos ligado à cadeia carbônica é substituído por outro átomo ou grupo. Esse tipo de reação é mais comum em compostos saturados, como alcanos e haletos orgânicos, em que uma ligação sigma (σ) é rompida e um novo grupo é incorporado à molécula. Reações de substituição podem ser nucleofílicas ou eletrofílicas, dependendo do mecanismo envolvido nas condições reacionais, como a presença de luz e catalisadores (Solomons;

Fryhle; Snyder, 2024).

Por fim, as reações de eliminação são um tipo de reação orgânica em que uma molécula elimina átomos ou grupos de átomos, resultando na formação de uma ligação pi (π) entre carbonos adjacentes. Esse processo geralmente leva à formação de compostos insaturados, como alcenos e alcinos. As reações de eliminação podem ocorrer por diferentes mecanismos, sendo os principais a eliminação intramolecular e a eliminação bimolecular (Spiro; Stigliani, 2008).

O entendimento detalhado das reações orgânicas na atmosfera é importante para o desenvolvimento de modelos preditivos voltados à mitigação dos impactos da poluição do ar. Nesse contexto, a integração de técnicas de *machine learning* permite simular e prever a formação de poluentes com maior precisão e eficiência (Park *et al.*, 2022).

2.2.1 Reações de Adição

As reações de adição ocorrem quando dois ou mais reagentes se combinam para formar um único produto. Esse tipo de reação é típico de compostos insaturados, como alcenos (hidrocarbonetos com dupla ligação) e alcinos (hidrocarbonetos com tripla ligação), em que a presença de ligações π permite a formação de novas ligações σ . A adição pode ocorrer em diferentes tipos de ligações, como duplas ou triplas, e envolve a quebra dessas ligações para a formação de novos grupos funcionais (Carey; Sundberg, 2000).

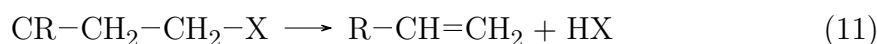
O mecanismo reacional de uma reação orgânica de adição geralmente envolve a formação de um intermediário reativo, como um carbocátion (C^+) ou um íon halogênio (X^-), dependendo do reagente utilizado (Carey; Sundberg, 2000). A reação de adição pode ser de dois tipos: eletrofílica, quando um eletrófilo (espécie química que aceita um par de elétrons) ataca a ligação π , ou radicalar, quando ocorre a formação de radicais livres.

As reações de adição do tipo hidratação são processos químicos em que moléculas de água (H_2O) são adicionadas a compostos insaturados, como alcenos e alcinos. (Solomons; Fryhle; Snyder, 2024). As reações de adição do tipo hidratação são fundamentais na química orgânica, pois permitem a síntese de álcoois, cetonas e aldeídos a partir de compostos insaturados (Solomons; Fryhle; Snyder, 2024). Compreender essas reações é importante para o planejamento de sínteses orgânicas e para a previsão dos produtos formados.

2.2.2 Reação de Eliminação

As reações de eliminação são um tipo de reação orgânica em que uma molécula perde átomos ou grupos de átomos, resultando na formação de uma ligação dupla ou tripla entre carbonos adjacentes. Essas reações são comuns e muito importantes na química orgânica e ocorrem frequentemente com compostos como haletos de álcool ou de alquila, nos quais a perda de átomos ou grupos de átomos leva à formação de uma ligação π entre carbonos. As reações de eliminação são amplamente utilizadas na síntese orgânica para a produção de alcenos e alcinos, que são intermediários importantes na fabricação de polímeros, fármacos e outros compostos orgânicos (March, 1985).

Em uma reação de eliminação, dois grupos substituintes são removidos de átomos de carbono adjacentes, resultando na formação de uma ligação dupla ($C=C$) ou tripla ($C\equiv C$), conforme apresentado na Equação 11 (Carey; Sundberg, 2000). O processo geral pode ser representado pela Equação 11:



onde:

- R é um grupo alquila ou hidrogênio;
- X é um grupo de saída, como um grupo hidroxila (OH) ou até mesmo um grupo hidrogênio (H).

As reações químicas podem ser melhor compreendidas e manipuladas por meio de seus mecanismos reacionais, que descrevem as etapas pelas quais os reagentes são transformados em produtos. No contexto das reações de eliminação, dois mecanismos fundamentais predominam: o E1 (eliminação unimolecular ou intramolecular) e o E2 (eliminação bimolecular). Ambos os processos envolvem a remoção de átomos ou grupos de átomos de uma molécula, resultando na formação de uma ligação insaturada, dupla ($C=C$) ou tripla ($C\equiv C$), entre carbonos adjacentes (Carey; Sundberg, 2000). As reações de eliminação são processos fundamentais na síntese orgânica.

2.2.3 Reações de Oxidação

As reações de oxidação sempre ocorrem acompanhadas de uma reação de redução; por isso, são chamadas de reações de oxirredução. As reações de oxirredução (ou redox) envolvem a transferência de elétrons entre espécies químicas: uma delas cedendo elétrons (agente redutor) e a outra recebendo-os (agente oxidante), resultando em alterações no número de oxidação (NOX) dos átomos envolvidos. Essas

reações desempenham um papel de suma importância na síntese e na modificação de compostos, influenciando diretamente suas propriedades estruturais e funcionais (Spiro; Stigliani, 2008).

A oxidação em compostos orgânicos ocorre pelo aumento no número de ligações de um átomo de carbono com átomos mais eletronegativos, como o oxigênio, ou pela perda de hidrogênios ligados ao carbono, elevando o estado de oxidação do carbono (NOX). Já a redução ocorre na direção oposta, envolvendo a adição de hidrogênios ou a remoção de oxigênios, o que reduz o NOX do carbono (Solomons; Fryhle; Snyder, 2024). Dentre diversas reações de oxidação, três se destacam por sua relevância no âmbito das reações ambientais: as reações com agentes oxidantes, como o ozônio (O_3), também chamadas de ozonólise, com a hidroxila ($-OH$) e com oxigênio atômico (O).

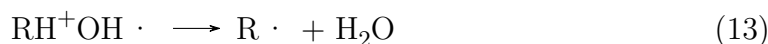
A ozonólise é um tipo específico de reação de oxirredução; trata-se de uma reação de oxidação que ocorre na presença de ozônio (O_3), conforme Equação 12. Esse tipo de reação é específico para compostos orgânicos como os alcenos, hidrocarbonetos com dupla ligação na cadeia. Tais reações promovem a cisão da dupla ligação e, posteriormente, a formação de compostos carbonílicos (March, 1985). A ozonólise pode ser empregada na degradação de poluentes orgânicos persistentes, como compostos aromáticos e hidrocarbonetos, em processos de tratamento de água e de ar.

Exemplo de uma ozonólise:



A oxidação por radicais hidroxila ($\cdot OH$) é um processo fundamental na química atmosférica, caracterizado pela alta reatividade com a maioria dos compostos orgânicos. Esse radical atua como um poderoso agente oxidante, iniciando reações em cadeia que levam à degradação de poluentes orgânicos voláteis (COVs) na atmosfera. O mecanismo típico envolve a abstração de um átomo de hidrogênio do composto orgânico, formando um radical alquila ($R \cdot$) e água (H_2O).

Esse radical livre, subsequentemente, reage com o oxigênio molecular (O_2), formando peróxidos orgânicos ($RO_2 \cdot$) que participam de ciclos fotoquímicos complexos, contribuindo para a formação de oxidantes atmosféricos, como o ozônio (O_3), e de compostos carbonílicos, conforme Equações 13 e 14.





O oxigênio molecular (O_2) atua como um agente oxidante fundamental em processos de combustão e em reações atmosféricas. Sua reatividade varia conforme as condições do meio, podendo participar tanto de oxidações brandas (como em sistemas biológicos e na catálise enzimática) quanto de reações energéticas, como na combustão de combustíveis fósseis (Seinfeld; Pandis, 2016). Na atmosfera, o O_2 reage com radicais livres orgânicos ($R \cdot$), formando radicais peróxila ($RO_2 \cdot$), que desempenham um papel crucial na formação do ozônio troposférico (O_3) e de outros oxidantes secundários (Solomons; Fryhle; Snyder, 2024).

A oxidação por O_2 é essencial nos ciclos de degradação de poluentes, mas também contribui para a formação de **smog** fotoquímico quando associada a NO_x e compostos orgânicos voláteis (COVs). A combustão incompleta, comum em motores a diesel, libera partículas finas ($PM_{2.5}$) e hidrocarbonetos policíclicos aromáticos (HPAs), que são carcinogênicos (Zhou *et al.*, 2021). Esses processos destacam a importância do O_2 tanto em reações naturais quanto em fenômenos de poluição atmosférica, o que influencia diretamente a qualidade do ar e a saúde coletiva.

2.2.4 Reações de Substituição

A reação de substituição é um tipo de reação orgânica em que um átomo ou grupo de átomos ligado a um carbono é substituído por outro átomo ou grupo. Esse processo pode ocorrer tanto em compostos saturados quanto em insaturados, sendo mais comum em alcanos e compostos aromáticos. Existem dois mecanismos principais para esse tipo de reação: substituição nucleofílica e substituição eletrofílica. No primeiro caso, um nucleófilo (espécie rica em elétrons) ataca o carbono parcialmente positivo, deslocando um grupo de saída. Já na substituição eletrofílica, comum em anéis aromáticos, um eletrófilo (espécie pobre em elétrons) substitui um hidrogênio ligado ao anel. Essas reações desempenham papel essencial na síntese de derivados orgânicos, como haletos, álcoois e nitrocompostos, e são influenciadas por fatores como a natureza do solvente, do grupo de saída e da espécie reagente (Solomons; Fryhle; Snyder, 2024).

A compreensão aprofundada das reações orgânicas, especialmente as envolvendo adição, eliminação e oxidação, é fundamental para o avanço das pesquisas acadêmicas e para o desenvolvimento de novas tecnologias na área da química, em suas diversas aplicações. Esse conhecimento não apenas viabiliza a síntese e a análise de compostos orgânicos, mas também contribui significativamente para a produção de energia e para a redução dos impactos ambientais. Assim, o estudo sistemático

das reações orgânicas evidencia o papel central da química orgânica na resolução de desafios científicos e industriais, ressaltando a importância contínua de investimentos em pesquisa e inovação nesse campo.

2.3 Machine Learning

Machine Learning (ML), ou Aprendizado de Máquina, constitui um sub-campo fundamental da Inteligência Artificial (IA) que se distingue pela capacidade de conferir aos sistemas computacionais a aptidão para aprender a partir de dados, em vez de depender estritamente de instruções programadas logicamente (Sharifani; Amini, 2023). Diferentemente da computação convencional, os algoritmos de ML utilizam métodos estatísticos para identificar padrões em grandes volumes de dados, permitindo que o modelo generalize o conhecimento adquirido para tomar decisões ou realizar previsões sobre novos conjuntos de dados (Goodfellow *et al.*, 2016). De forma geral, o processo de treinamento e aplicação de um modelo de aprendizado de máquina pode ser representado conforme ilustrado na Figura 3.

Figura 3 – Fluxo geral de funcionamento de um modelo de aprendizado de máquina.



Fonte: Elaborado pelos autores com base em (Géron, 2019).

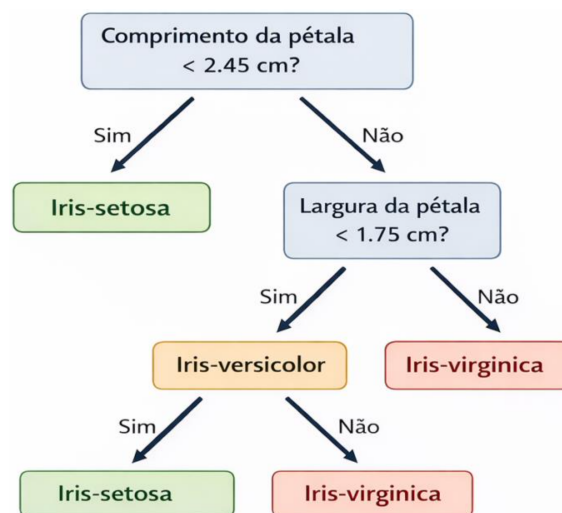
A classificação principal de ML baseia-se na natureza da experiência de aprendizado e na presença ou ausência de supervisão durante o treinamento. De acordo com Mehta *et al.* (2019) e Ramsundar *et al.* (2019b), as abordagens dividem-se em três categorias principais: Aprendizado Supervisionado, Aprendizado Não Supervisionado e Aprendizado por Reforço.

O Aprendizado Supervisionado é a técnica mais difundida, na qual o modelo é treinado com um conjunto de dados previamente rotulados. Ou seja, para cada entrada fornecida ao sistema, existe uma resposta correta (rótulo) conhecida. O objetivo é aprender uma função que mapeie as entradas para as saídas, sendo

amplamente utilizado em tarefas de classificação (ex.: diagnóstico de doenças) e de regressão (ex.: previsão de propriedades químicas).

Entre os modelos mais utilizados no aprendizado supervisionado destacam-se as árvores de decisão e as redes neurais profundas. As árvores de decisão são modelos hierarquicamente estruturados, em que os dados são sucessivamente divididos em subconjuntos com base em critérios estatísticos que visam maximizar a separação entre classes ou minimizar o erro de previsão. Entre os critérios mais comuns estão o índice de Gini (Gastwirth, 1972) e a entropia. Uma das principais vantagens desse modelo é sua interpretabilidade, pois o processo de decisão pode ser representado por regras claras e facilmente visualizadas em forma de árvore, como ilustrado na Figura 4. No entanto, árvores individuais podem apresentar alta variância e tendência ao sobreajuste (*overfitting*). Para reduzir esse problema, utilizam-se métodos de aprendizado em conjunto (*ensemble learning*), como Random Forest e Extra Trees, que combinam múltiplas árvores para aumentar a estabilidade e a capacidade preditiva do modelo.

Figura 4 – Exemplo simplificado de uma árvore de decisão utilizada em tarefas de classificação.

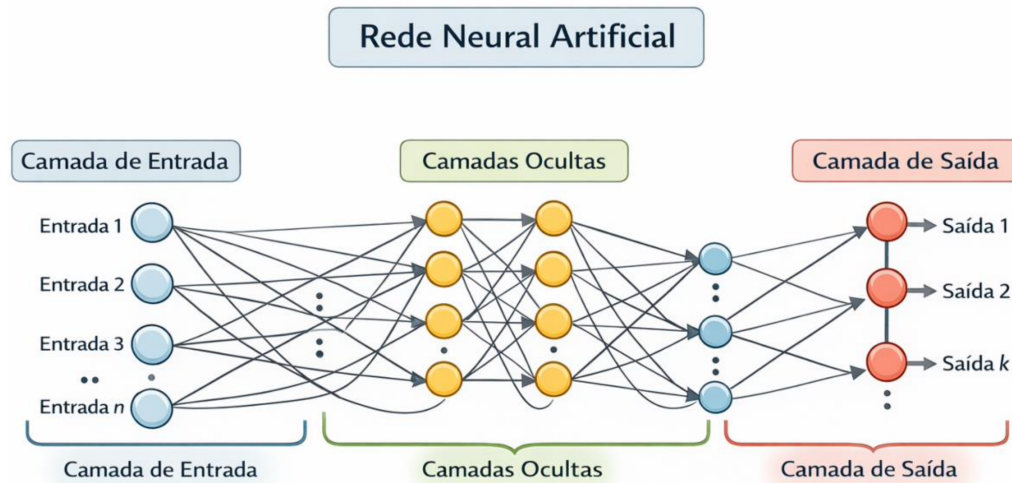


Fonte: Elaborado pelos autores com base em (Géron, 2019).

As redes neurais profundas, ou *Deep Neural Networks* (DNN), por sua vez, são modelos inspirados no funcionamento do sistema nervoso biológico e compostos por múltiplas camadas de neurônios artificiais interconectados, como ilustrado na Figura 5. Cada neurônio realiza uma combinação ponderada das entradas, adiciona um termo de viés e aplica uma função de ativação não linear para gerar uma saída. Durante o treinamento, os pesos das conexões entre os neurônios são ajustados por meio do algoritmo de retropropagação do erro (*backpropagation*), que ajusta iterativamente os pesos com base no gradiente da função de perda. A presença

de múltiplas camadas ocultas permite que a rede aprenda representações mais complexas dos dados e capture relações não lineares, o que torna esses modelos particularmente úteis em problemas que envolvem grande volume de informações ou padrões difíceis de identificar (Géron, 2019; Sahu; Mokhade; Bokde, 2023).

Figura 5 – *Estrutura básica de uma rede neural artificial.*



Fonte: Elaborado pelos autores com base em (Géron, 2019).

Para a implementação prática dessas arquiteturas, bibliotecas computacionais têm sido utilizadas no desenvolvimento de modelos de aprendizado de máquina. Entre elas, destaca-se o Keras (Gulli; Pal, 2017), uma interface de alto nível desenvolvida para facilitar a construção, o treinamento e a avaliação de redes neurais profundas, geralmente operando sobre o framework TensorFlow (Abadi *et al.*, 2016). Sua estrutura baseada em camadas permite definir modelos de forma sequencial, especificando, de forma simplificada, funções de ativação, otimizadores e métricas de desempenho. Dessa forma, ferramentas como esta contribuem para reduzir a complexidade da implementação dos modelos, permitindo que o pesquisador concentre seus esforços na análise e modelagem do problema e na interpretação dos resultados obtidos (Géron, 2019).

Já no Aprendizado Não Supervisionado, o algoritmo analisa dados sem rótulos ou categorias pré-definidas. O sistema busca identificar, por conta própria, estruturas ocultas ou agrupamentos intrínsecos nos dados. É essencial para a descoberta de novos padrões e para a segmentação de dados (*clustering*), permitindo, por exemplo, identificar novos tipos de compostos químicos com características semelhantes (Park *et al.*, 2022).

Por fim, o Aprendizado por Reforço, diferentemente das categorias anteriores, baseia-se na interação de um agente com o ambiente. O aprendizado ocorre

por meio de um sistema de recompensas e punições, no qual o algoritmo busca aperfeiçoar uma estratégia para alcançar um objetivo específico a longo prazo.

A robustez desses modelos permite sua aplicação em domínios científicos complexos. Na medicina, os algoritmos de ML transcendem a análise estatística básica, sendo empregados na personalização de tratamentos e na predição de eficácia farmacológica (Ramsundar *et al.*, 2019b). No âmbito da ciência dos materiais e da química, ML acelera a descoberta de novos compostos ao prever propriedades moleculares e otimizar reações que, por métodos experimentais convencionais, demandam tempo e recursos significativamente superiores (Alves *et al.*, 2018; Park *et al.*, 2022). Assim, a integração de ML na pesquisa científica atua como catalisador da inovação, permitindo a gestão e a interpretação de volumes de dados experimentais em escalas sem precedentes.

2.3.1 Pipeline de desenvolvimento de modelos de ML

O desenvolvimento de modelos de *machine learning* geralmente não ocorre de forma isolada, mas sim seguindo um conjunto de etapas organizadas, conhecido como *pipeline*. O *pipeline* pode ser entendido como um fluxo de trabalho que abrange desde o tratamento dos dados até o treinamento e ajuste do modelo (Géron, 2019; Belcic; Stryker, 2026), sendo essencial para garantir resultados mais confiáveis (Nkemdilim *et al.*, 2024).

A primeira etapa envolve a limpeza e preparação dos dados (*data cleaning*). Nessa fase, são tratados problemas comuns, como dados ausentes, inconsistências e duplicidades, que podem comprometer o desempenho do modelo (Géron, 2019). Em aplicações químicas, essa etapa pode envolver a padronização de estruturas moleculares, a remoção de compostos inválidos e a verificação de inconsistências nos descritores moleculares. Esse processo é importante porque os algoritmos de aprendizado de máquina dependem diretamente da qualidade dos dados utilizados (Landrum, 2006; Cherkasov *et al.*, 2014).

Em seguida, tem-se a etapa de engenharia de atributos (*feature engineering*), que consiste em transformar os dados brutos em informações mais úteis ao modelo. Isso pode incluir a criação de novas variáveis ou a modificação das já existentes, com o objetivo de facilitar a identificação de padrões (Géron, 2019). No contexto da quimioinformática, essa etapa inclui a geração de descritores moleculares, como propriedades físico-químicas e características estruturais, obtidos a partir de representações como SMILES e SMARTS. Essa transformação permite converter informações químicas em variáveis numéricas interpretáveis pelos algoritmos (Elton *et al.*, 2019).

Após a construção dos atributos, realiza-se a seleção de variáveis (*feature selection*), com o objetivo de selecionar apenas as mais relevantes para o problema. Essa etapa ajuda a reduzir a complexidade do modelo e a evitar problemas como sobreajuste (*overfitting*), além de tornar os resultados mais fáceis de interpretar (Géron, 2019).

Após a etapa anterior, é possível avaliar a importância das variáveis (*feature importance*), que permite identificar quais atributos têm maior influência nas previsões do modelo. Essa análise é muito útil em modelos baseados em árvores de decisão, pois contribui para a interpretação dos resultados (Géron, 2019).

A etapa seguinte corresponde ao treinamento do modelo, no qual o algoritmo aprende padrões a partir dos dados disponíveis (Nkemdilim *et al.*, 2024). Normalmente, os dados são divididos em conjuntos de treino e teste, o que permite avaliar o desempenho do modelo em dados que não foram utilizados no treinamento (Géron, 2019; Belcic; Stryker, 2026). Diferentes algoritmos podem ser empregados, como árvores de decisão, métodos de ensemble e redes neurais, dependendo da natureza do problema e da complexidade dos dados.

Por fim, destaca-se a etapa de definição e ajuste de hiperparâmetros, que são configurações do modelo definidas antes do treinamento e que influenciam diretamente o desempenho. A escolha adequada desses parâmetros é fundamental para obter melhores resultados e, muitas vezes, é realizada de forma iterativa (Géron, 2019).

De maneira geral, o pipeline de *Machine Learning* é um processo contínuo e iterativo, no qual diferentes etapas podem ser revisitadas ao longo do desenvolvimento (Nkemdilim *et al.*, 2024). No contexto deste trabalho, essas etapas são fundamentais para garantir a construção de modelos capazes de prever, com maior precisão, a formação e a toxicidade de compostos derivados de reações químicas atmosféricas.

2.3.2 *Machine Learning* na Química

No domínio da Química, a integração de *Machine Learning* tem transformado a modelagem molecular, a previsão de propriedades físico-químicas e a otimização de rotas sintéticas. A capacidade desses algoritmos de processar grandes volumes de dados permite a identificação de padrões complexos em bases de dados estruturais e cinéticas, como as do PubChem, do *United States Patent and Trademark Office* (USPTO) e da *United States Environmental Protection Agency* (USEPA) (Park *et al.*, 2022). Esse suporte computacional é fundamental para acelerar a descoberta de novos compostos e aperfeiçoar processos químicos, tornando-os mais eficientes e menos onerosos (Alves *et al.*, 2018; Duarte *et al.*, 2025).

Uma das frentes mais promissoras é a predição dos resultados de reações químicas. Tradicionalmente, a identificação de produtos e a análise de reatividade dependem de experimentos exaustivos em laboratório, o que demanda tempo e gera resíduos químicos consideráveis. A aplicação de modelos preditivos baseados em dados históricos permite antecipar com alta precisão o desfecho de reações, minimizando a necessidade de testes empíricos e alinhando a pesquisa aos princípios da Química Verde (Park *et al.*, 2022). Para viabilizar essa análise, ferramentas como o RDKit (Landrum, 2006) e o Deep Chem (Ramsundar *et al.*, 2019a) são essenciais, pois permitem a manipulação de estruturas moleculares, o cálculo de descritores químicos e a execução de simulações computacionais robustas (Duarte *et al.*, 2025).

No contexto da Química Ambiental, o aprendizado de máquina torna-se uma ferramenta estratégica para o monitoramento e a mitigação da poluição atmosférica. O ML tem sido empregado com sucesso para prever a formação de poluentes, identificar fontes de emissão e otimizar estratégias de controle (Zhou *et al.*, 2021). Dado que a degradação da qualidade do ar é um desafio global, a compreensão das transformações químicas na atmosfera é vital para a formulação de políticas públicas de saúde e de preservação ambiental (Lima; Borges, 2021).

Particularmente, o ML destaca-se na análise de mecanismos complexos, como a formação de compostos orgânicos voláteis (COVs) e de hidrocarbonetos aromáticos policíclicos (HPAs). Esses compostos resultam de interações fotoquímicas entre poluentes primários, como os óxidos de nitrogênio (NOx), sob influência da radiação solar. Modelar essas reações por métodos convencionais é altamente complexo devido ao número de variáveis envolvidas. Contudo, o uso de algoritmos de ML permite mapear essas correlações de forma ágil, auxiliando no desenvolvimento de estratégias eficazes para reduzir a concentração desses poluentes na troposfera (Zhou *et al.*, 2021).

Um exemplo prático dessas previsões é a estimativa da geração de antraceno decorrente de interações químicas envolvendo compostos presentes no diesel. O antraceno é um hidrocarboneto aromático policíclico (HPA) reconhecido por sua alta toxicidade e potencial cancerígeno. Ele pode ter sua produção atmosférica antecipada por meio de técnicas de aprendizado de máquina. Esses modelos são alimentados com informações sobre reações previamente documentadas e, posteriormente, aplicados para simular a síntese de antraceno a partir de diversas interações entre substâncias químicas (Huang *et al.*, 2008).

2.3.3 Biblioteca RDKit

Python é uma linguagem de programação interpretada, o que significa que o código é executado linha por linha, sem a necessidade de compilação prévia.

Essa característica facilita o desenvolvimento rápido de aplicações e a realização de testes iterativos (Lutz, 2025). Sua sintaxe simples e legível, aliada a uma vasta gama de bibliotecas especializadas, torna o Python uma ferramenta essencial para pesquisadores e desenvolvedores (VanRossum; Drake, 2010).

Uma das principais vantagens do Python é sua extensa biblioteca padrão, que inclui módulos para manipulação de strings, operações matemáticas, acesso a bancos de dados e muito mais. Além disso, a comunidade Python é ativa e contribui com uma grande variedade de pacotes de terceiros, que ampliam as funcionalidades da linguagem para áreas específicas, como ciência de dados, aprendizado de máquina e química computacional (Pedregosa *et al.*, 2011).

A versatilidade do Python como linguagem de programação científica tem encontrado diversas aplicações no campo da química computacional, onde a capacidade de manipular e analisar estruturas moleculares de forma eficiente é essencial. Neste contexto, a biblioteca RDKit (Landrum, 2006) emerge como uma das ferramentas mais poderosas e versáteis disponíveis para pesquisadores. Desenvolvido como um projeto de código aberto, o RDKit oferece um conjunto abrangente de funcionalidades que abrange desde a representação básica de moléculas até análises complexas de propriedades químicas, integrando-se perfeitamente ao ecossistema científico do Python.

O RDKit se destaca por sua capacidade de lidar com diversos formatos de representação molecular, incluindo *Simplified Molecular Input Line Entry System* (SMILES), *International Chemical Identifier* (InChI), e arquivos no formato *Structure Data File* (SDF). Essa flexibilidade permite aos pesquisadores converter facilmente entre diferentes representações, o que facilita a integração com bancos de dados químicos e outras ferramentas computacionais.

Uma das características mais valiosas do RDKit é sua capacidade de visualização molecular integrada. A biblioteca permite a renderização de estruturas químicas em 2D e 3D diretamente em ambientes como notebooks do Jupyter, oferecendo ferramentas para destacar características específicas, como grupos funcionais ou ligações de interesse. Essa funcionalidade não apenas auxilia na análise visual de moléculas, mas também é indispensável para a comunicação de resultados em publicações científicas e em apresentações.

Para ilustrar a aplicação prática do RDKit, considere o exemplo apresentado no Código 1 que demonstra como carregar uma molécula a partir de sua representação SMILES, calcular seus descritores moleculares e gerar sua visualização.

Este exemplo simples demonstra como o RDKit pode ser utilizado para obter rapidamente propriedades moleculares relevantes. Em aplicações mais avançadas, como no desenvolvimento de modelos de *machine learning* para prever resultados de

reações químicas (um dos focos desta pesquisa), o RDKit pode ser usado para gerar representações numéricas de moléculas que capturam suas características estruturais essenciais. Essas representações podem então ser alimentadas em algoritmos de aprendizado de máquina para prever, por exemplo, a formação de produtos em reações orgânicas ou a toxicidade de compostos químicos.

Código 1 Exemplo de aplicação de funções do RDKit

```
1 from rdkit import Chem
2 from rdkit.Chem import Descriptors
3 from rdkit.Chem.Draw import IPythonConsole
4
5 # Carregar molécula a partir de SMILES
6 mol = Chem.MolFromSmiles('C1=CC=CC=C1') # Benzeno
7
8 # Calcular massa molar e logP
9 massa_molar = Descriptors.MolWt(mol)
10 logP = Descriptors.MolLogP(mol)
11
12 print(f"Massa Molar: {massa_molar:.2f} g/mol")
13 print(f"Coeficiente de Partição (logP): {logP:.2f}")
```

Na pesquisa aqui apresentada, que utiliza modelos de *machine learning* para prever resultados de reações químicas com aplicações ambientais, o RDKit desempenha um papel central. Ele é utilizado para processar e representar as moléculas alvo presentes no diesel, gerar descritores moleculares relevantes, e facilitar a integração desses dados com algoritmos de aprendizado de máquina. Essa abordagem permitirá não apenas acelerar o processo de investigação das reações químicas, mas também reduzir a necessidade de experimentos laboratoriais extensivos, alinhando-se com os princípios da química verde e sustentável.

2.3.4 Codificação SMILES

O sistema *Simplified Molecular Input Line Entry System* (SMILES) estabelece-se como um dos padrões mais amplamente adotados na representação computacional de estruturas químicas, oferecendo uma notação linear eficiente para descrever moléculas complexas. Desenvolvido como uma linguagem simplificada, o SMILES permite a codificação precisa de informações estruturais através de uma sequência de caracteres ASCII, tornando-se fundamental para armazenamento, busca e intercâmbio de dados químicos em plataformas computacionais (Weininger, 1988).

O SMILES distingue-se por sua abordagem intuitiva, onde átomos são representados por seus símbolos químicos e ligações por caracteres específicos (- para simples, = para duplas, # para triplas). A notação incorpora regras claras para representar características estruturais essenciais, incluindo cadeias lineares, ramificações (indicadas por parênteses), ciclos (numerados) e estereoquímica (usando os símbolos /, \ e @). O código SMILES serve como base para formatos mais avançados como SMARTS, consolidando seu papel como linguagem fundamental no ecossistema de química computacional (Weininger, 1988).

2.3.5 Codificação SMARTS

A codificação *SMiles Arbitrary Target Specification* (SMARTS) destaca-se como uma ferramenta essencial no campo da quimioinformática, complementando e expandindo as capacidades do padrão SMILES para aplicações avançadas em busca molecular. Desenvolvida como uma extensão, a SMARTS permite a representação precisa de padrões estruturais complexos, tornando-se indispensável para pesquisas que requerem identificação de sub estruturas químicas específicas (Daylight Chemical Information Systems, 2007).

O sistema SMARTS diferencia-se por sua capacidade de especificar padrões de busca sofisticados através de uma sintaxe que incorpora operadores lógicos e símbolos especializados. Essa abordagem oferece flexibilidade incomparável na descrição de características moleculares, incluindo tipos específicos de átomos, ligações químicas e ambientes estruturais, possibilitando buscas precisas em grandes bancos de dados moleculares. A linguagem suporta a construção de consultas complexas que são fundamentais para aplicações como triagem virtual de compostos, estudo isômeros, análise de similaridade molecular e estudos de relação estrutura atividade (Daylight Chemical Information Systems, 2007).

Metodologia

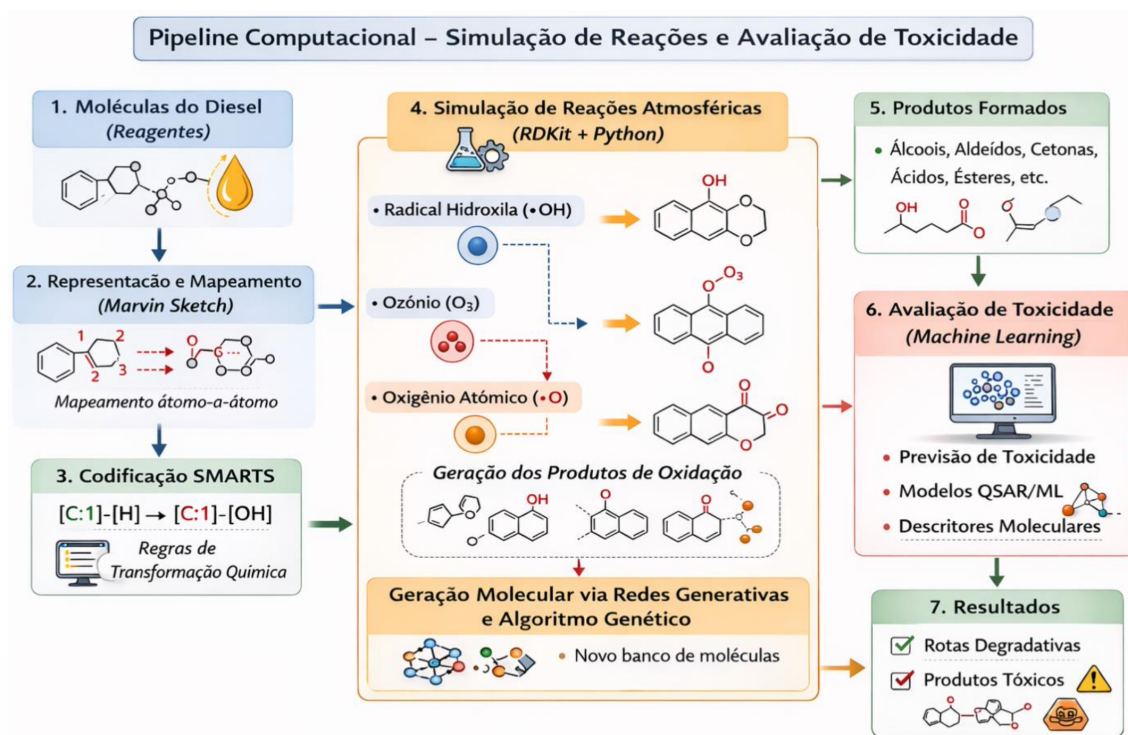
Esta pesquisa fundamenta-se no desenvolvimento de um pipeline computacional para a predição de rotas degradativas de moléculas constituintes do diesel combustível e a subsequente avaliação de toxicidade dos produtos formados, utilizando técnicas de *Machine Learning* e Quimioinformática.

De forma geral, o procedimento metodológico adotado pode ser compreendido como uma sequência estruturada de etapas computacionais que abrange desde a representação molecular dos compostos presentes no diesel até a predição das propriedades toxicológicas dos produtos gerados. Inicialmente, as moléculas de interesse são representadas e preparadas para simulação computacional e, posteriormente, submetidas a regras de transformação química que simulam processos de oxidação atmosférica. A partir dessas transformações, obtêm-se possíveis produtos de degradação, que passam por etapas adicionais de geração molecular e de caracterização estrutural.

A organização geral desse fluxo metodológico é apresentada na Figura 6, que ilustra o pipeline computacional desenvolvido neste estudo. Nessa representação, destacam-se as principais etapas do processo, incluindo a preparação das moléculas iniciais, a simulação de reações atmosféricas com a biblioteca RDKit, a geração de novas estruturas moleculares por meio de redes generativas e algoritmos genéticos e, por fim, a aplicação de modelos de *Machine Learning* para a predição de toxicidade.

A partir dessa estrutura geral, as subseções a seguir descrevem detalhadamente cada etapa do processo metodológico, incluindo os procedimentos de simulação de reações atmosféricas, de geração molecular e de análise toxicológica das estruturas obtidas.

Figura 6 – *Pipeline computacional para simulação de reações atmosféricas e avaliação de toxicidade de moléculas do diesel*



Fonte: Elaborado pelos autores.

3.1 Simulação de Reações Atmosféricas

As reações de oxidação foram simuladas com a biblioteca RDKit no ambiente Python, no Google Colab. O procedimento seguiu as etapas:

- 1. Representação e mapeamento:** Os padrões das partes principais das moléculas dos reagentes foram desenhados no software Marvin Sketch (v.22.13), no qual se realizou o mapeamento átomo-a-átomo.
- 2. Codificação SMARTS:** A representação SMILES foi expandida para o padrão SMARTS, que define regras de transformação química.
- 3. Simulação estrutural:** Foram modeladas reações com três agentes oxidantes atmosféricos principais: radical hidroxila (OH), ozônio (O_3) e oxigênio atômico (O). A execução das reações foi processada por funções, em linguagem Python, específicas da biblioteca RDKit para gerar os produtos principais.

3.1.1 Seleção das moléculas representativas do diesel

Foram selecionadas dez moléculas representativas da composição do diesel combustível, com base em dados reportados na literatura especializada, considerando

sua abundância relativa, diversidade estrutural e relevância como potenciais precursores de Hidrocarbonetos Aromáticos Policíclicos (HAPs). A escolha contemplou diferentes classes químicas (alifáticos, aromáticos e heterocíclicos), visando representar adequadamente a complexidade do combustível e suas possíveis transformações químicas na atmosfera.

Com base nos dados obtidos na revisão teórica, destacam-se os dados do trabalho realizado por Júnior (2013), que visou identificar poluentes atmosféricos na cidade de Goiânia - GO. O autor analisou dezesseis hidrocarbonetos aromáticos polinucleados, apresentados no Quadro 1, e identificou a presença significativa de alguns desses HAPs (tais como benzo[a]pireno, antraceno e naftaleno) no ar atmosférico de bairros da cidade.

Quadro 1 – *Lista dos HPAs analisados.*

nº	Nome	nº	Nome
1	Naftaleno	9	Criseno
2	Acenafteno	10	Benzo(b)fluoranteno
3	Fluoreno	11	Benzo(k)fluoranteno
4	Antraceno	12	Benzo(a)pireno
5	Fenantreno	13	Benzo(e)pireno
6	Fluorantreno	14	Benzo(g,h,i)perileno
7	Pireno	15	Indeno(1,2,3,cd)pireno
8	Benz(a)antraceno	16	Dibenz(a,h)antraceno

Fonte: (Júnior, 2013)

A partir da análise dos resultados obtidos no trabalho de Júnior (2013), a pesquisa seguiu para o próximo passo na busca pela identificação das fontes emissoras, com ênfase na correlação entre os hidrocarbonetos aromáticos policíclicos (HAPs) detectados e possíveis origens antropogênicas, especialmente o tráfego veicular, com ênfase no diesel (Yao *et al.*, 2024).

Diante da presença desses compostos nos dados analisados, tornou-se conveniente investigar a composição molecular do diesel, uma vez que este combustível é composto por uma complexa mistura de hidrocarbonetos orgânicos, muitos dos quais apresentam elevado potencial para a formação de HAPs. Considerando a relevância dessa via de emissão, aprofundou-se a avaliação dos precursores químicos presentes no diesel, buscando compreender as transformações atmosféricas e sua

contribuição para a poluição ambiental, conforme evidenciado em estudos de Júnior (2013) e Venkatraman *et al.* (2024).

3.1.2 Representação molecular e codificação química

Os dados coletados sobre as moléculas-alvo e as reações químicas foram processados e analisados em Python, utilizando a biblioteca RDKit. O Python foi escolhido por ser uma linguagem de alto nível, com sintaxe intuitiva, amplamente adotada na ciência de dados e na química computacional, além de oferecer uma ampla gama de bibliotecas para *machine learning*, manipulação de dados e visualização. Já a biblioteca RDKit é ideal por ser especializada em quimiointormática, capaz de manipular estruturas moleculares, calcular descritores químicos e representar reações orgânicas de forma eficiente.

As moléculas representativas do diesel foram convertidas em códigos SMILES, enquanto as reações químicas foram descritas por meio de códigos SMARTS. Para a realização dessa etapa, foi utilizado o software Marvin Sketch (ChemAxon, 2024), um programa proprietário e pago, para o qual foi concedida uma licença acadêmica gratuita específica para o desenvolvimento desta pesquisa. O Marvin Sketch, desenvolvido pela empresa ChemAxon, foi empregado para o desenho, a edição e a visualização de estruturas químicas, bem como para a conversão entre diferentes formatos de representação molecular. Esse software foi selecionado por permitir a construção e a manipulação detalhada de estruturas químicas de forma intuitiva, sendo amplamente utilizado em atividades acadêmicas e científicas que envolvem modelagem molecular.

Após a geração dos códigos SMILES, as estruturas moleculares foram convertidas em objetos moleculares compatíveis com a biblioteca RDKit por meio da função `rdkit.Chem.MolFromSmiles()`. Em seguida, o padrão de reação química SMART foi lido e aplicado de forma sistemática a todas as moléculas reagentes, com o objetivo de simular as transformações químicas propostas. O procedimento foi configurado para exibir os produtos gerados na tela, com a possibilidade de armazenar os dados em arquivos no formato TXT e SMILES. Os resultados destes procedimentos estão apresentados no Apêndice A.

3.2 Geração molecular via redes generativas e algoritmo genético

Para trabalhar com um amplo conjunto de moléculas orgânicas específicas presentes no diesel, optou-se por empregar um *molecular generator*. Baseado em

uma rede neural generativa, utiliza um algoritmo evolucionário multiobjetivo (Silva; Evans, 2024). Para gerar estruturas semelhantes à composição do diesel, um conjunto de 27 moléculas validadas experimentalmente foi utilizado como alvo de similaridade (Guan *et al.*, 2022). Para otimizar a descoberta de novos compostos, utilizou-se uma arquitetura híbrida:

- **Variational Autoencoder (VAE):** Um modelo generativo pré-treinado com 1,8 milhão de compostos do banco ChEMBL (Gaulton *et al.*, 2012), capaz de comprimir e reconstruir estruturas químicas em um espaço vetorial contínuo.
- **Otimização multiobjetivo:** Aplicou-se o algoritmo NSGA-III (Algoritmo Genético de Ordenação Não Dominada) (Deb; Jain, 2013) para navegar no espaço de soluções. O processo evolutivo utilizou cruzamento de dois pontos ($P=0,9$) e mutação gaussiana ($\sigma = 0,01$; $P=0,1$), operando por 100 gerações com população fixa de 640 indivíduos.

3.3 Desenvolvimento dos modelos de classificação de toxicidade

O banco de dados *In vitro mutagenicity in Salmonella Typhimurium* (ISSSTY) (Benigni *et al.*, 2013), disponibilizado em Istituto Superiore di Sanità (2021), constitui uma base de dados sobre mutagenicidade, desenvolvida com o teste de Ames (Ames; Lee; Durston, 1973), com cepas da bactéria *Salmonella Typhimurium* e possui 7.367 linhas (dados brutos).

A planilha que representa o banco contém os seguintes atributos: nome do composto, número *Chemical Abstracts Service* (CAS), código SMILES, peso molecular, fórmula geral, resultado geral do teste AMES (*overall*) e resultados de testes específicos.

A avaliação do potencial mutagênico foi estruturada como um problema de classificação multiclasse (Rótulos: 1 - Negativo, 2 - Duvidoso, 3 - Positivo), com base no teste de Ames.

3.3.1 Tratamento de dados

O conjunto de dados original ISSSTY (7.367 amostras) passou por um processo de limpeza:

- a) Validação de códigos SMILES via número CAS no banco PubChem;
- b) Remoção de resultados “inconclusivos”, resultando em um conjunto final de 5.827 amostras;

- c) Geração de 123 descritores *Quantitative Structure-Activity Relationship* (QSAR) calculados com o RDKit.
- d) Definição de um total de 5.827 amostras, sendo 1.166 para o conjunto de teste e 4.661 para o conjunto de treinamento.

3.3.2 Arquitetura dos modelos e treinamento

Para a redução de dimensionalidade, aplicaram-se métodos de *Feature Selection* e *Feature Importance* por meio da biblioteca Scikit-Learn, reduzindo a matriz de 123 para 70 descritores fundamentais. Os dados foram divididos em 80% para treinamento e 20% para teste.

Foram avaliados dois algoritmos:

- Extra Tree Classifier: Configurado com 100 estimadores e um mínimo de 2 folhas por nó.
- Rede Neural Artificial, utilizando o framework Keras: Otimizada via framework Optuna. A arquitetura consiste em uma camada de entrada (70 neurônios), três camadas ocultas (128, 64 e 32 neurônios) e uma camada de saída Softmax (3 neurônios). Utilizou-se o otimizador Adam ($\eta = 8,103 \times 10^{-4}$), função de perda categorical crossentropy e técnica de *Early Stopping* para prevenir *overfitting*.

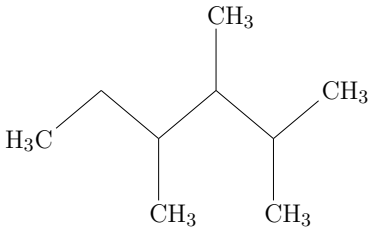
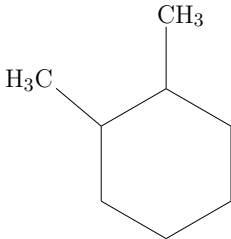
Resultados e Discussão

Os resultados apresentados neste capítulo correspondem à aplicação de modelos de ML para prever a formação de compostos aromáticos a partir de moléculas presentes no diesel, utilizando a biblioteca RDKit

4.1 Modelagem de reações atmosféricas via SMARTS e RDKit

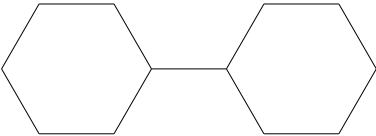
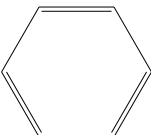
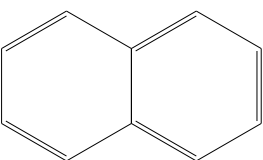
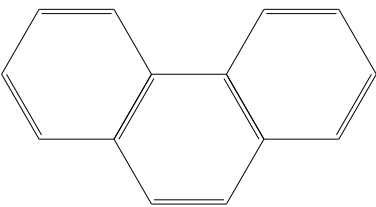
As moléculas selecionadas presentes no diesel, apresentadas na Tabela 1, foram convertidas em representações SMILES pelo software MarvinSketch, obtido com licença acadêmica para uso na realização da pesquisa, permitindo a padronização e a análise computacional.

Tabela 1 – Exemplos de moléculas presentes no diesel e suas representações SMILES.

Nº	Molécula	Fórmula Molecular	Código SMILES	Fórmula estrutural
1	Hexano	C ₆ H ₁₄	CCCCC	
2	2,3,4 Trimetilhexano	C ₉ H ₂₀	CCC(C)C(C)C(C)C	
3	1,2 Dimetilciclohexano	C ₈ H ₁₆	CC1CCCCC1C	

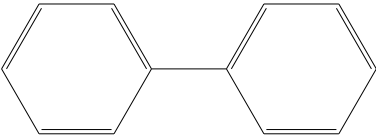
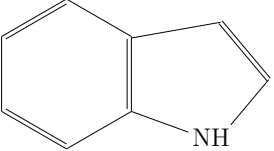
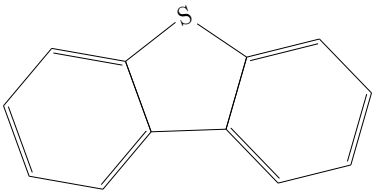
(Continua)

Tabela 1 – Continuação

Nº	Molécula	Fórmula Molecular	Código SMILES	Fórmula estrutural
4	Biciclohexano	C ₁₂ H ₂₂	<chem>C1CCC(CC1)C1CCCCC1</chem>	
5	Benzeno	C ₆ H ₆	<chem>C1=CC=CC=C1</chem>	
6	Naftaleno	C ₁₀ H ₈	<chem>C1=CC2=CC=CC=C2C=C1</chem>	
7	Fenantreno	C ₁₄ H ₁₀	<chem>C1=CC2=C(C=C1)C1=C(C=CC=C1)C=C2</chem>	

(Continua)

Tabela 1 – Continuação

Nº	Molécula	Fórmula Molecular	Código SMILES	Fórmula estrutural
8	Fenilbenzeno	C ₆ H ₅	<chem>C1=CC=C(C=C1)C1=CC=CC=C1</chem>	
9	Indol	C ₈ H ₉ N	<chem>N1C=CC2=C1C=CC=C2</chem>	
10	Tetrahidrodibenzo ti- ofeno	C ₁₂ H ₈ S	<chem>S1C2=CC=CC=C2C2=CC=CC=C12</chem>	

Fonte: Elaborada pelos autores.

4.1.1 Reações atmosféricas

Foram simuladas reações atmosféricas de oxidação envolvendo radicais hidroxila ($\cdot\text{OH}$) oxigênio atômico (O) e ozônio (O_3), utilizando funções da biblioteca RDKit, separando as classes mais comuns de reagentes específicos para teste, mas não todos os representados na Tabela 1. O Apêndice A apresenta os produtos gerados em reações-teste com compostos específicos, como compostos oxigenados e intermediários reativos, que evidenciam a formação de poluentes secundários. Por exemplo, a oxidação do fenantreno resultou em estruturas que são precursores diretos do antraceno, reforçando a hipótese de que moléculas aromáticas presentes no diesel contribuem para a poluição atmosférica urbana, na forma de material particulado (Wang *et al.*, 2021).

As reações envolvendo os oxidantes atmosféricos radicais hidroxila ($\cdot\text{OH}$), ozônio (O_3) e oxigênio atômico (O) foram estruturadas e implementadas por meio de algoritmos desenvolvidos em Python, os quais estão disponíveis publicamente no repositório https://github.com/Schimidt99/atmospheric_reactions, juntamente com as demais etapas metodológicas desenvolvidas no trabalho. A formalização computacional dessas reações permitiu a padronização dos mecanismos considerados e a aplicação automatizada posteriormente às moléculas selecionadas neste trabalho.

A seleção dos mecanismos reacionais foi fundamentada em referências clássicas e consolidadas da química atmosférica (Manahan, 2022; Baird; Cann, 2011; Akimoto, 2016) e também em estudos específicos (Atkinson; Arey, 2003; Burlakova *et al.*, 2005; Kruk, 1998; Miller, 2017; Rajkiewicz; Tyszkiewicz; Wertejuk, 2014; Jewell; Holbrook; Oldershaw, 1981; Jewell; Holbrook; Oldershaw, 1982; Hassner; Galle, 1970), que descrevem os principais processos de degradação de compostos orgânicos na troposfera. Esses processos estão diretamente associados à formação do *smog* fotoquímico, fenômeno característico de regiões urbanas e industriais, resultante principalmente das emissões de motores a combustão interna e de determinadas atividades industriais. Nessas condições, a presença de radiação solar favorece a geração de espécies reativas capazes de oxidar compostos orgânicos voláteis (COVs), desencadeando cadeias reacionais complexas.

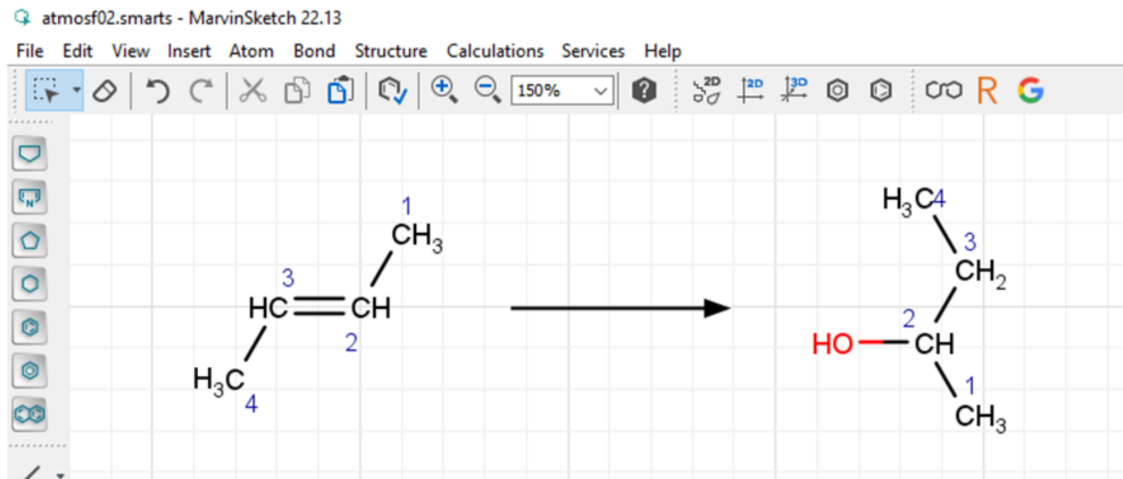
Embora a literatura descreva múltiplos caminhos mecanísticos capazes de conduzir a produtos finais semelhantes, optou-se, neste trabalho, pela seleção de nove reações específicas (descritas detalhadamente no Apêndice A). Essa escolha foi orientada pela objetividade mecanística e pela clareza na formação dos produtos, buscando garantir a coerência com as referências adotadas e a viabilidade computacional na modelagem. A limitação deliberada do conjunto de reações permitiu maior controle na análise dos resultados e reduziu as ambiguidades associadas a mecanismos concorrentes.

Todas as reações foram inicialmente modeladas no software Marvin Sketch, no qual foram construídas representações estruturais dos reagentes e dos principais produtos. O processo de construção envolveu a definição de fragmentos moleculares representativos das regiões reativas, que, posteriormente, foram organizados em um padrão de reação. Em seguida, realizou-se o mapeamento átomo a átomo (*atom mapping*), etapa fundamental para explicitar quais átomos participam diretamente do processo reacional.

Esse mapeamento permitiu identificar, de forma inequívoca, quais ligações químicas são rompidas e quais são formadas ao longo da transformação, garantindo a consistência estrutural entre os reagentes e os produtos. A partir dessa etapa, o padrão reacional foi exportado no formato SMARTS, possibilitando sua integração ao ambiente Python e sua aplicação sistemática por meio da biblioteca RDKit.

A Figura 7 apresenta um exemplo simplificado de reação modelada, ilustrando o procedimento de construção do padrão reacional e a definição explícita das transformações estruturais consideradas neste exemplo.

Figura 7 – Parte do painel principal do software Marvin Sketch, mostrando a reação de ataque (#2) à dupla ligação de alcenos alifáticos, via radical OH.



Fonte: Elaborada pelos autores..

A reação química representada na Figura 7 (Reação #2) foi codificada em linguagem SMARTS segundo o seguinte padrão estrutural:

$$[\#6:1] \backslash [\#6:2] = [\#6:3] \backslash [\#6:4] >> [\#6:4] - [\#6:3] - [\#6:2] (-[\#6:1]) - [\#8]$$

Essa codificação descreve explicitamente como ocorre a transformação estrutural do reagente em produto, indicando, por meio da numeração dos átomos,

quais centros atômicos estão envolvidos na quebra e na formação de ligações. A indexação atômica presente na notação SMARTS é essencial para garantir que o mapeamento entre o reagente e o produto seja realizado corretamente, mantendo a correspondência estrutural durante a aplicação e o processamento da reação.

A modelagem computacional dessa reação foi realizada por meio da execução sequencial de comandos em Python, utilizando funções da biblioteca RDKit. O código foi executado na plataforma Google Colab. Antes da aplicação da reação, o programa principal carrega um arquivo de texto no formato TXT, denominado “diesel-smiles.txt”. Esse arquivo contém 261 moléculas reagentes, e cada linha corresponde a uma estrutura química representada na notação SMILES.

Após a leitura do padrão SMARTS pela função `AllChem.ReactionFromSmarts`, a transformação é aplicada automaticamente a todas as moléculas do arquivo de entrada que obedecem ao padrão de reação SMARTS. A partir desse momento, os produtos formados seguem rigorosamente a regra de indexação definida no código SMARTS, assegurando que a modificação estrutural ocorra exatamente nos átomos especificados no padrão reacional.

Um ponto importante é que a codificação SMARTS utilizada distingue explicitamente estruturas alifáticas de aromáticas. Essa diferenciação já estabelece, previamente, se determinada molécula reagente poderá ou não gerar produto ao ser submetida à reação. Assim, apenas as estruturas compatíveis com o padrão definido sofrem transformação química.

Todos os produtos gerados são convertidos para notação SMILES e gravados automaticamente em um arquivo de saída no formato ASCII, denominado “saida.txt”. Durante esse processo, estruturas repetidas são removidas, evitando duplicidade nos resultados. No Quadro 2 estão apresentados exemplos de moléculas reagentes extraídas do arquivo “diesel-smiles.txt”, juntamente com os respectivos produtos obtidos após o processamento da reação ilustrada na Figura 7.

O padrão reacional apresentado na Figura 7 corresponde a um processo de oxidação, simulando o ataque do radical hidroxila ($\cdot\text{OH}$) a ligações duplas presentes exclusivamente em cadeias alifáticas (Reação #2). Essa representação busca reproduzir, de forma simplificada, um dos principais mecanismos de degradação de compostos orgânicos na atmosfera.

Os códigos em Python desenvolvidos para cada reação específica encontram-se disponíveis no repositório https://github.com/Schmidt99/atmospheric_reactions, permitindo a reprodução do procedimento descrito.

O padrão de reação mostra todos os possíveis produtos formados. Cada molécula gerada como produto da reação passa por uma etapa de validação denominada “sanitização”. Esse procedimento consiste na verificação estrutural completa da mo-

Quadro 2 – *Alguns exemplos de moléculas reagentes e produtos, no formato SMILES, da reação química da Figura 7 (Reação #2)*

	Reagentes	Produtos
1	<chem>CCCCCCC(CC(C)C)n1ccnc1</chem>	—
2	<chem>CCCC(C)C1(C(C)C)C=CC(C)C(=O)N1</chem>	<chem>CCCC(C)C1(C(C)C)NC(=O)C(C)CC1O</chem>
3	<chem>CCC1CCCC1</chem>	—
4	<chem>CC1=C(C(C)C)C(C(C)C)NCC1</chem>	<chem>CC(C)C1NCCC(C)C1(O)C(C)C</chem>
5	<chem>CCN1CC=C(CC(C)=O)CC1</chem>	<chem>CC(C)C1NCCC(C)(O)C1C(C)C</chem>

Fonte: Elaborado pelos autores.

lécua formada, por meio de uma função específica que avalia individualmente cada átomo da estrutura. Durante essa análise, são examinados três critérios principais: o estado de valência do átomo, a coerência das ligações químicas estabelecidas e a configuração da sua vizinhança imediata, ou seja, os átomos diretamente conectados a ele (RDKit Development Team, 2025).

Observando o Quadro 2, nota-se que não há produtos formados a partir dos reagentes localizados nas primeiras e terceiras linhas. Esse resultado está diretamente relacionado às regras estabelecidas no padrão de reação definido pela representação SMARTS utilizada. No caso do primeiro reagente, a molécula apresenta caráter aromático, o que a torna incompatível com o padrão reacional adotado. Já para o terceiro reagente, não há ligação dupla na estrutura, condição indispensável para que a transformação descrita ocorra.

Por fim, é importante destacar que a codificação SMARTS empregada apresenta limitações inerentes à sua natureza estrutural. Esse tipo de representação não permite incorporar aspectos estequiométricos das reações, nem estabelecer relações quantitativas entre as concentrações dos reagentes e dos produtos. Da mesma forma, não contempla a definição dos estados físicos das substâncias nem a influência de variáveis externas, como a pressão e a temperatura. Assim, o modelo aplicado restringe-se à transformação estrutural das moléculas, sem considerar parâmetros físico-químicos adicionais.

O número total de produtos únicos obtidos a partir das nove reações de oxidação avaliadas é apresentado na Tabela 2, cujos mecanismos são descritos

detalhadamente no Apêndice A. A denominação “produtos únicos” foi adotada porque, em determinadas situações, moléculas reagentes que apresentam simetria geométrica, observada em relação ao eixo da ligação química principal, podem gerar produtos estruturalmente idênticos. Nesses casos, embora a transformação possa ocorrer em posições equivalentes da molécula, o resultado final é a mesma estrutura química. Por esse motivo, tais produtos foram contabilizados apenas uma vez, evitando duplicidade na quantificação.

Tabela 2 – *Número de produtos únicos produzidos nas reações químicas estudadas considerando as 261 moléculas de reagentes.*

#	Reação principal	Mecanismo	Número de produtos únicos
1	Radical OH	Cadeias aromáticas	250
2		Cadeias alifáticas	10
3	Ozônio	Cadeias alifáticas	8
4		Cadeias aromáticas	338
5	Oxigênio atômico	Cadeias alifáticas(A)	8
6		Cadeias alifáticas(B)	6
7		Cadeias alifáticas(C)	5
8		Alcanos	1048
9		Cadeias aromáticas	189

Fonte: Elaborada pelos autores.

Entre as reações analisadas, a reação de oxigênio atômico, identificada como #8, foi a que gerou o maior número de produtos, considerando exclusivamente reagentes constituídos por cadeias alifáticas abertas. Em contraste, a reação de oxigênio atômico #7, caracterizada pelo ataque a cadeias cíclicas insaturadas com formação de um epóxido, apresentou o menor rendimento em termos de quantidade de produtos distintos formados.

A modelagem das reações químicas que ocorrem na troposfera, conforme apresentada nesta seção, permite a investigação sistemática das diferentes possibili-

dades de transformação estrutural associadas aos processos de oxidação atmosférica. A partir dessa abordagem, torna-se viável gerar produtos específicos de forma automatizada e organizá-los em um novo banco de dados estruturado, ampliando o conjunto de informações disponíveis para análises posteriores.

Com esse conjunto de moléculas geradas, é possível avançar para etapas complementares de avaliação, como o cálculo e a predição de toxicidade (procedimento descrito no tópico seguinte), além de contribuir para estudos relacionados à bioacumulação e à biotransformação (Costa *et al.*, 2008; Tchounwou *et al.*, 2012; Flores-Rojas; Esterhuizen-Londt; Pflugmacher, 2015; Magalhães *et al.*, 2021). A modelagem também fornece subsídios para investigações sobre a formação de aerossóis atmosféricos e de material particulado secundário (Souza *et al.*, 2023b), aspectos diretamente associados à química atmosférica e à qualidade do ar.

Sob uma perspectiva prática, essa abordagem computacional permite estudar reações químicas específicas com significativa economia de tempo e redução no consumo de reagentes, além de eliminar a geração de resíduos laboratoriais. Dessa forma, o procedimento adotado contribui não apenas para a compreensão teórica dos processos atmosféricos, mas também para a otimização de recursos e para práticas alinhadas aos princípios de sustentabilidade.

4.2 Gerador Molecular Computacional (*Molecular Generator*)

Com o objetivo de trabalhar com um conjunto representativo de moléculas orgânicas específicas presentes no diesel combustível, empregou-se um gerador molecular baseado em rede neural generativa, associado a um algoritmo evolucionário multiobjetivo (Silva *et al.*, 2024). Essa abordagem permitiu a geração controlada de estruturas químicas com características semelhantes às de subprodutos do diesel.

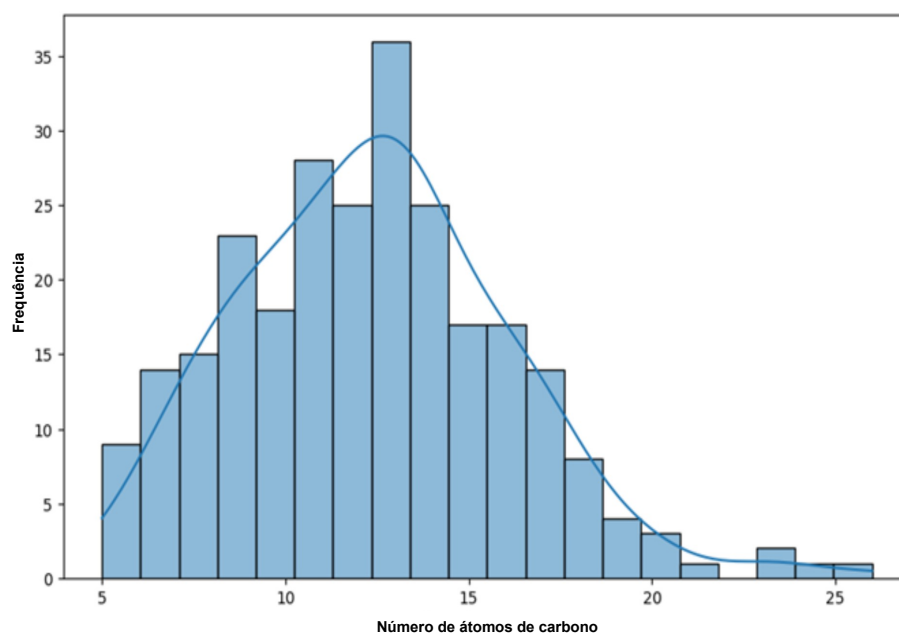
Para orientar o processo de geração molecular, utilizou-se um conjunto de 27 moléculas previamente validadas experimentalmente como alvos de similaridade (Guan *et al.*, 2022). A similaridade estrutural entre as moléculas geradas e as moléculas de referência foi avaliada quantitativamente por meio da distância de Tanimoto, calculada a partir de impressões digitais de Morgan com raio de 2, conforme descrito por Rogers e Hahn (2010).

Cada uma das 27 moléculas de referência foi considerada um objetivo independente no processo de otimização, o que resultou em um problema de 27 dimensões. O algoritmo foi estruturado para maximizar, simultaneamente, a similaridade estrutural em relação a todos esses alvos. Como é característico de algoritmos evo-

lucionários multiobjetivos, o processo não conduz a uma única solução ótima global. Em vez disso, produz-se um conjunto diversificado de soluções que compõem uma aproximação à fronteira de Pareto, representando diferentes compromissos entre os objetivos de similaridade.

A partir da geração final do algoritmo, foram selecionadas 261 estruturas moleculares exclusivas, que compuseram o conjunto de reagentes empregado neste trabalho. As moléculas obtidas atenderam às especificações técnicas do diesel combustível descritas na literatura (Petrobras, 2024; Speight, 2006), particularmente no que se refere ao número de átomos de carbono, à distribuição de aromaticidade e ao intervalo de massa molecular. Esses parâmetros são apresentados nas Figuras 8, 9 e 10.

Figura 8 – *Distribuição do número de átomos de carbono entre as 261 moléculas geradas pelo molecular generator.*

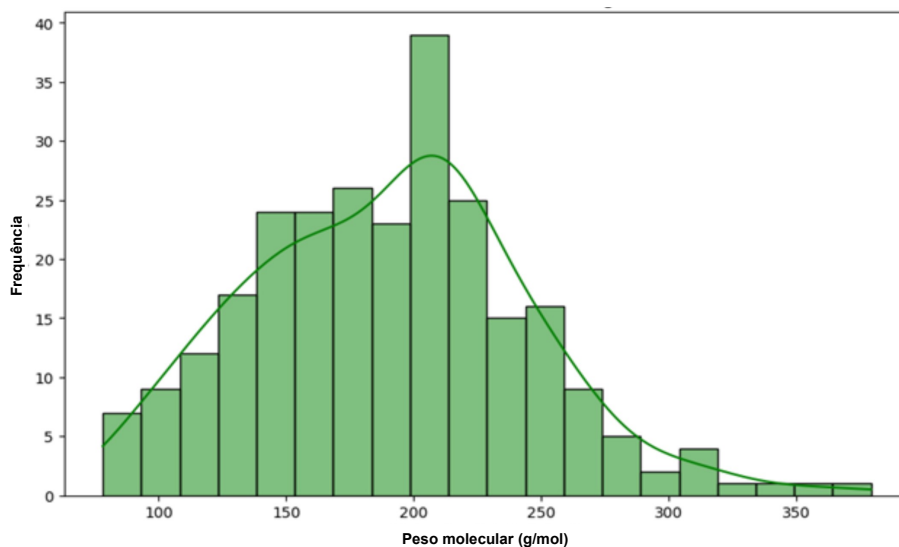


Fonte: Elaborada pelos autores.

Todas as análises apresentadas nestas figuras, incluindo a classificação estrutural direta das moléculas, foram realizadas por meio de funcionalidades específicas da biblioteca RDKit, com programação em Python. Os códigos correspondentes encontram-se disponíveis no repositório https://github.com/Schmidt99/atmospheric_reactions, na rotina denominada “diesel characterization”.

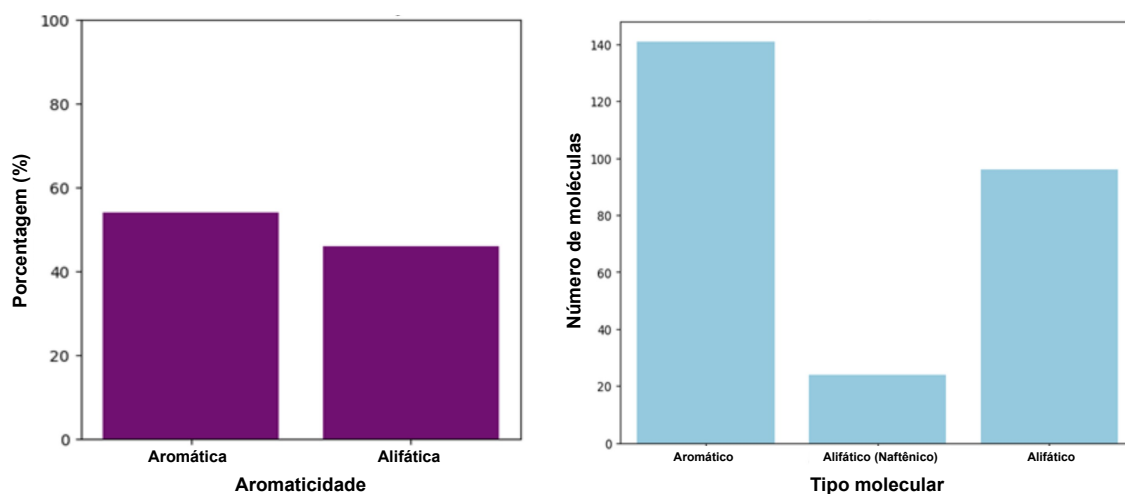
A análise das 261 estruturas moleculares selecionadas a partir do gerador molecular demonstra que o conjunto obtido apresenta características estruturais compatíveis com as especificidades técnicas do diesel combustível. Conforme evidenciado nas Figuras 8 a 10, há uma distribuição coerente do número de átomos de

Figura 9 – *Distribuição do peso molecular das 261 moléculas geradas pelo molecular generator.*



Fonte: Elaborada pelos autores.

Figura 10 – *Distribuição de aromaticidade e classificação das 261 moléculas geradas pelo molecular generator. Os alifáticos da figura à direita incluem uma mistura de alcanos e alcenos.*



Fonte: Elaborada pelos autores.

carbono, concentrada majoritariamente na faixa típica de hidrocarbonetos presentes nesse combustível. De maneira semelhante, a distribuição de massa molecular revela predominância em intervalos consistentes com compostos intermediários e pesados característicos de frações do diesel.

A avaliação da aromaticidade indica a presença tanto de estruturas aromáticas quanto alifáticas, refletindo a diversidade estrutural esperada para esse tipo de combustível. A classificação direta das moléculas confirma essa heterogeneidade

química, evidenciando que o conjunto gerado não se restringe a uma única classe estrutural, mas abrange diferentes perfis moleculares relevantes, como mostrado na Figura 10.

Esses resultados indicam que o algoritmo generativo foi capaz de produzir um conjunto estruturalmente plausível e tecnicamente consistente com as características do diesel, mantendo, simultaneamente, a diversidade molecular. A aproximação da fronteira de Pareto obtida pelo algoritmo multiobjetivo permitiu equilibrar os critérios de similaridade em relação aos 27 alvos de referência, resultando em uma base de dados representativa para as etapas subsequentes do estudo.

Dessa forma, o conjunto molecular gerado constitui uma base de dados adequada para a aplicação de análises preditivas, especialmente na avaliação de propriedades toxicológicas. A caracterização estrutural realizada nesta seção estabelece, portanto, o fundamento necessário para a próxima etapa do trabalho, dedicada à aplicação de modelos de toxicidade às moléculas obtidas.

4.3 Modelos de toxicidade

A modelagem computacional da toxicidade de compostos orgânicos constitui um desafio significativo, principalmente devido à grande diversidade de ensaios biológicos disponíveis para a avaliação da bioatividade. Esses ensaios envolvem diferentes organismos e sistemas biológicos, incluindo animais, linhagens celulares e microrganismos, além de apresentarem variações metodológicas e múltiplas fontes de dados. Essa heterogeneidade torna o processo de construção de modelos preditivos particularmente complexo. Na literatura, observa-se uma ampla variedade de abordagens voltadas tanto à classificação quanto à regressão. Existem modelos consolidados de classificação (MultiCASE Inc., 2025; Cavasotto; Scardino, 2022), bem como modelos de regressão aplicados à predição quantitativa de toxicidade (Roy; Ghosh, 2010; Hendriks; Traas; Huijbregts, 2005; Jiang *et al.*, 2019). Entretanto, essas metodologias apresentam especificidades e limitações inerentes aos seus conjuntos de dados, aos descritores moleculares e aos respectivos domínios de aplicabilidade (Alves; Streit; Pizzolato, 2023; Cronin; Yoon, 2019).

Considerando esse cenário, optou-se, neste trabalho, pela construção de um modelo de classificação baseado nos testes de mutagenicidade de Ames (Benigni, 2019). A escolha desse ensaio justifica-se por sua ampla utilização como indicador preliminar de potencial mutagênico e por sua relevância regulatória. Para o treinamento e validação dos modelos, utilizou-se a base de dados pública ISSSTY (Istituto Superiore di Sanità, 2021), composta por 6.226 moléculas previamente avaliadas ex-

perimentalmente. Dois modelos robustos de classificação foram testados e descritos a seguir.

4.3.1 Modelo Extra Trees Classifier

Os resultados obtidos com o modelo Extra Trees Classifier encontram-se apresentados na Tabela 3. Observa-se que, ao analisar as médias ponderadas das métricas de desempenho, o modelo apresenta resultados consistentes, com valores superiores a 0,80 (80%) para os principais parâmetros de avaliação: acurácia, precisão, recall e f1-score. O valor de ROC AUC (0,9131) indica uma boa capacidade de discriminar entre as classes.

Tabela 3 – *Métricas de avaliação no conjunto de teste do modelo Extra Trees Classifier.*

Classe	Precisão	Recall	f1-score	Acurácia	ROC AUC	Amostras
1	0,80	0,80	0,80			455
2	0,60	0,12	0,19			26
3	0,85	0,88	0,87			685
Média pond.	0,8256	0,8302	0,8241	0,8302	0,9131	1.166

Fonte: Elaborada pelos autores.

A análise individual por classe revela comportamentos distintos entre as classes. Para as classes 1 e 3, os valores de precisão, recall e f1-score situam-se em patamares elevados (aproximadamente entre 0,80 e 0,87), indicando desempenho satisfatório na identificação correta dessas categorias. Entretanto, para a classe 2, observa-se redução significativa nos valores de recall (0,12) e f1-score (0,19), o que sugere maior dificuldade do modelo em reconhecer adequadamente essa classe específica. Essa diferença pode estar associada à menor representatividade dessa categoria no conjunto de teste, que conta com apenas 26 amostras, conforme indicado na última coluna da Tabela 3. Em contraste, as classes 1 e 3 apresentam 455 e 685 amostras, respectivamente, o que evidencia um desequilíbrio na distribuição entre elas.

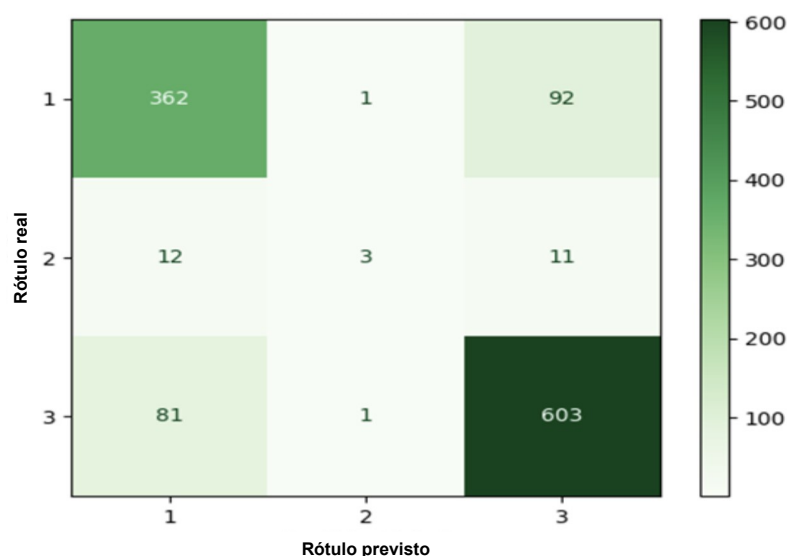
A média ponderada (*weighted average*), calculada considerando a proporção de amostras por classe, apresenta valores elevados e consistentes (acurácia = 0,8302; f1-score = 0,8241), indicando que, de forma global, o modelo apresenta desempenho robusto para o conjunto avaliado.

Complementando a análise das métricas apresentadas na Tabela 3, destaca-se o desempenho do parâmetro ROC AUC, que atingiu 0,9131. Esse resultado pode ser considerado bastante satisfatório, especialmente levando em conta que o conjunto de dados utilizado não é balanceado e que a classe 2 é de difícil predição. De acordo com Benigni (2019), em seu estudo sobre a predição *in silico* da mutagenicidade no teste de Ames, essa classe corresponde a resultados “inconclusivos”, o que, naturalmente, introduz maior incerteza no processo de classificação e impacta negativamente o desempenho do modelo.

Os demais parâmetros de avaliação apresentaram valores próximos entre si, indicando estabilidade no comportamento preditivo global do modelo. Também foi testada a aplicação da Análise de Componentes Principais, *Principal Component Analysis* (PCA), à matriz de dados QSAR, com o objetivo de reduzir a dimensionalidade e, possivelmente, melhorar o desempenho do classificador. No entanto, essa estratégia não resultou no aprimoramento das métricas avaliadas, sugerindo que a representação original dos descritores já continha informações suficientes para o treinamento do modelo Extra Trees.

A Figura 11 apresenta a matriz de confusão obtida no conjunto de teste. Observa-se que os elementos da diagonal principal concentram as classificações corretas do modelo. Mesmo diante da dificuldade associada à classe 2, o algoritmo conseguiu classificar corretamente 3 amostras dessa classe, o que corresponde a 11% do total dessa classe. Observa-se predominância de classificações corretas na diagonal principal, o que evidencia um desempenho global satisfatório.

Figura 11 – Matriz de confusão obtida para o conjunto de teste, no modelo Extra Trees Classifier.



Fonte: Elaborada pelos autores.

As classes 1 (negativo) e 3 (positivo) apresentaram elevado número de

acertos, o que demonstra que o modelo possui boa capacidade discriminativa para compostos com comportamento mutagênico definido. Os principais erros ocorreram entre essas duas classes, indicando possível sobreposição nos padrões estruturais representados pelos descritores moleculares.

Em contrapartida, a classe 2 (inconclusivo) apresentou desempenho inferior, com menor número de classificações corretas. Esse resultado é consistente com o desbalanceamento da base de dados e com a própria natureza ambígua dessa categoria, já discutida na literatura. A dificuldade na predição dessa classe impacta principalmente métricas como recall e f1-score específicos, mas não compromete de forma significativa o desempenho global do modelo.

O valor obtido para o ROC AUC (0,9131) confirma a elevada capacidade discriminativa do modelo, mesmo diante do desbalanceamento da base e da presença de uma classe intermediária. As demais métricas apresentaram comportamento consistente entre si, reforçando a estabilidade da abordagem baseada em árvores de decisão.

A aplicação de PCA aos descritores QSAR foi avaliada como estratégia de redução de dimensionalidade; entretanto, não se observaram melhorias nas métricas de desempenho. Esse resultado sugere que o modelo já consegue lidar adequadamente com a estrutura dimensional original dos dados, possivelmente devido à natureza robusta do algoritmo Extra Trees em relação às correlações entre variáveis.

De forma geral, o Extra Trees Classifier demonstrou desempenho adequado na classificação de mutagenicidade, apresentando boa separação entre compostos negativos e positivos, com concentração limitada na classe 2 (inconclusiva).

Com o objetivo de comparar essa abordagem baseada em árvores com um modelo de aprendizado profundo, foi implementado um modelo de classificação baseado em rede neural com a biblioteca Keras, cujos resultados são apresentados na seção seguinte. .

4.3.2 Modelo de Aprendizado Profundo

Modelos baseados em redes neurais profundas têm apresentado crescimento significativo na área de Química Computacional e modelagem molecular nos últimos anos, especialmente em tarefas envolvendo relações estrutura-atividade e predição de propriedades químicas (Sahu; Mokhade; Bokde, 2023). Com o objetivo de comparar o desempenho do modelo baseado em árvores com uma abordagem de aprendizado profundo, foi implementada uma rede neural com a biblioteca Keras (Gulli; Pal, 2017).

O modelo foi treinado com o mesmo conjunto de treinamento e teste empregado no Extra Trees Classifier, garantindo comparabilidade direta entre as

abordagens. Além disso, foi realizado um processo de otimização de hiperparâmetros com a biblioteca Optuna (Akiba *et al.*, 2019), visando identificar configurações que maximizam o desempenho preditivo. A arquitetura final adotada foi do tipo totalmente conectada, com estrutura 70-128-64-32-3.

Apesar da maior complexidade do modelo, os resultados apresentados na Tabela 4 indicam desempenho inferior ao observado com o Extra Trees Classifier. O modelo apresentou acurácia global de 0,7796 e ROC AUC de 0,8770, ambos inferiores aos valores obtidos anteriormente.

Tabela 4 – *Métricas de avaliação no conjunto de teste do modelo de Aprendizado Profundo.*

Classe	Precisão	Recall	f1-score	Acurácia	ROC AUC	Amostras
1	0,71	0,77	0,74			455
2	0	0	0			26
3	0,83	0,81	0,82			685
Média pond.	0,7648	0,7796	0,7717	0,7796	0,8770	1.166

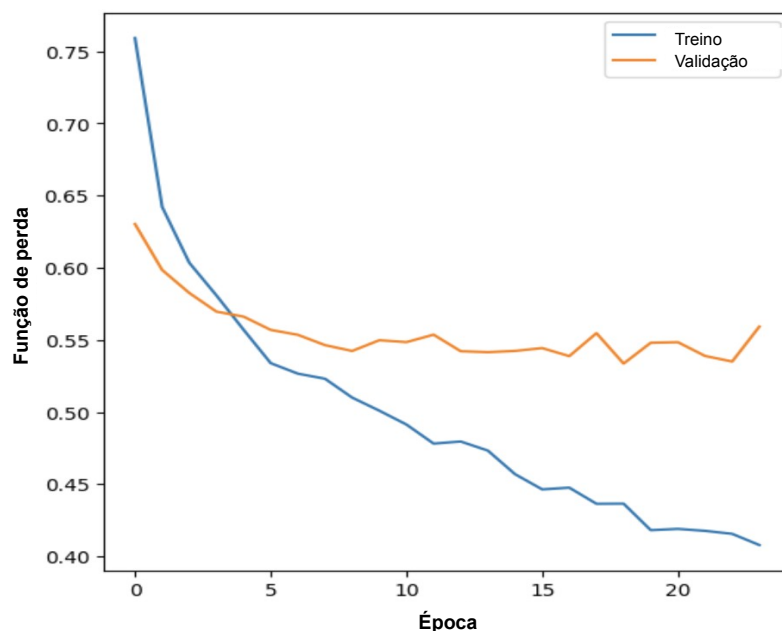
Fonte: Elaborada pelos autores.

Observa-se que o desempenho foi particularmente afetado pela classe 2 (inconclusiva), para a qual a rede neural não conseguiu prever corretamente nenhuma amostra. Esse comportamento evidencia elevada sensibilidade do modelo ao desbalanceamento da base de dados, impactando diretamente as métricas precisão, recall e f1-score dessa categoria. Mesmo para as classes 1 (negativo) e 3 (positivo), os valores de f1-score foram inferiores aos obtidos com o modelo baseado em árvores.

Os resultados da função de perda para os conjuntos de treinamento e validação, apresentados na Figura 12, demonstram comportamento estável ao longo das épocas. O uso do recurso *Early Stopping* impediu o treinamento até o limite máximo de 100 épocas, interrompendo o processo quando não houver melhora adicional no desempenho na validação, o que sugere ausência de sobreajuste significativo.

Observa-se redução contínua da perda de treinamento, enquanto a perda de validação apresenta estabilização precoce com pequenas oscilações ao longo das épocas. A presença de uma diferença moderada entre as curvas indica limitação na capacidade de generalização do modelo, especialmente diante do desbalanceamento da base de treinamento. O uso do recurso *Early Stopping* interrompeu o treinamento

Figura 12 – *Função de perda em relação ao número de épocas considerando o recurso Early Stopping*



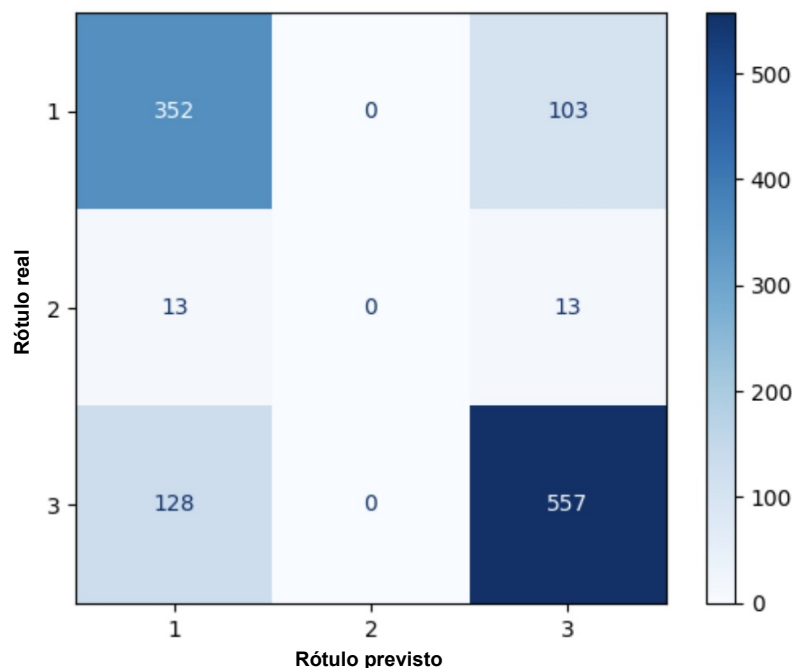
antes do limite máximo de épocas, evitando a intensificação do sobreajuste, embora não tenha sido suficiente para melhorar o desempenho na classe minoritária.

A matriz de confusão apresentada na Figura 13 confirma a redução do desempenho em comparação ao Extra Trees Classifier, com menor número de classificações corretas para as classes 1 e 3 e completa incapacidade de modelagem da classe 2. Esse resultado sugere que, para o conjunto de descritores QSAR utilizados e considerando o tamanho e o desbalanceamento da base, a abordagem baseada em árvores de decisão mostrou-se mais robusta do que a rede neural profunda.

A matriz de confusão evidencia que o modelo concentrou suas previsões nas classes majoritárias (1 e 3), não realizando nenhuma classificação correta para a classe 2. Esse comportamento indica uma forte influência do desequilíbrio na base de treinamento da rede neural, resultando em um viés em direção às classes mais representativas. Como consequência, o modelo apresentou desempenho reduzido em métricas sensíveis à classe minoritária, embora tenha mantido desempenho moderado nas classes predominantes.

Esses resultados sugerem que, para conjuntos de dados com dimensionalidade moderada e distribuição desbalanceada, modelos baseados em árvores podem apresentar maior robustez do que redes neurais profundas, cujo desempenho tende a ser mais sensível à quantidade e à distribuição das amostras disponíveis para cada classe.

Figura 13 – *Matriz de confusão obtida para o conjunto de teste, no modelo de Aprendizado Profundo.*



4.4 Considerações finais do capítulo

Neste capítulo foram apresentadas as etapas de simulação computacional de reações em ambiente Python, a obtenção de descritores moleculares e a aplicação de modelos de aprendizado de máquina para a predição de mutagenicidade. A integração dessas etapas permitiu estruturar um fluxo completo de geração, caracterização e classificação dos compostos analisados.

Entre os modelos avaliados, o Extra Trees Classifier apresentou desempenho superior à rede neural profunda implementada com Keras, demonstrando maior robustez diante do desbalanceamento da base de dados e maior capacidade discriminativa global. A rede neural profunda, embora mais complexa e otimizada por meio de ajuste de hiperparâmetros, mostrou maior sensibilidade à distribuição desigual das amostras, não modelando adequadamente a classe minoritária. Os resultados obtidos evidenciam que a escolha do algoritmo deve considerar as características específicas do conjunto de dados, especialmente a dimensionalidade e o equilíbrio entre as classes. Para o cenário analisado, o modelo baseado em árvores apresentou melhor equilíbrio entre desempenho preditivo, estabilidade e custo computacional.

De forma geral, os procedimentos desenvolvidos neste capítulo demonstram a viabilidade da abordagem proposta para a predição de mutagenicidade, evidenciando a importância da escolha adequada do algoritmo em função das características do conjunto de dados analisado.

Conclusão

A integração de métodos de aprendizado de máquina e de ferramentas de quimioinformática na modelagem de reações químicas demonstrou-se uma abordagem consistente e tecnicamente viável para aplicações em Química Ambiental. Este trabalho evidenciou a possibilidade de simular computacionalmente a formação de produtos de oxidação atmosférica e avaliar, de forma preditiva, seu potencial mutagênico, contribuindo para estratégias de análise de risco ambiental.

Diante da crescente diversidade de contaminantes atmosféricos, abordagens exclusivamente experimentais tornam-se limitadas por custos, tempo e pela geração adicional de resíduos. Nesse contexto, a aplicação de modelos baseados em descritores QSAR permitiu estabelecer relações entre características estruturais moleculares e a atividade mutagênica, demonstrando a relevância da informação estrutural na predição de propriedades toxicológicas específicas.

A utilização da biblioteca RDKit para cálculo de descritores, aliada ao desempenho superior do modelo Extra Trees Classifier em comparação à rede neural profunda avaliada, reforça que a escolha adequada do algoritmo é determinante para o sucesso preditivo, especialmente em bases de dados desbalanceadas. Os resultados obtidos indicam que a abordagem proposta constitui uma ferramenta eficaz para a triagem preliminar de compostos derivados do diesel, contribuindo para a priorização de substâncias potencialmente mais preocupantes do ponto de vista toxicológico.

De forma geral, o estudo demonstra que a combinação entre simulação computacional de reações atmosféricas e modelagem preditiva baseada em aprendizado de máquina representa uma estratégia metodológica robusta para investigação de contaminantes emergentes, alinhando eficiência analítica e sustentabilidade científica.

Diante dos resultados obtidos, também é possível apontar algumas possibilidades de continuidade para este trabalho. Como sugestão para trabalhos futuros, este estudo pode ser ampliado com a inclusão de um número maior de moléculas, sem se limitar às presentes no diesel. Além disso, os modelos de *machine learning* podem ser aprimorados por meio de outras formas de avaliação, como a validação

cruzada, e por meio de testes com diferentes configurações dos algoritmos, visando a resultados mais precisos. Também é possível explorar outros tipos de descritores moleculares ou até integrar dados experimentais já disponíveis na literatura, o que pode tornar as previsões mais confiáveis.

Por fim, a metodologia desenvolvida neste trabalho pode ser aplicada em outros contextos da Química Ambiental, como na previsão de diferentes poluentes e na análise de seus possíveis efeitos à saúde e ao meio ambiente. Dessa forma, o trabalho não apenas contribui para o avanço da área, mas também abre caminhos para novas investigações.

Referências

ABADI, Martín; BARHAM, Paul; CHEN, Jianmin; CHEN, Zhifeng; DAVIS, Andy; DEAN, Jeffrey; DEVIN, Matthieu; GHEMAWAT, Sanjay; IRVING, Geoffrey; ISARD, Michael *et al.* {TensorFlow}: a system for {Large-Scale} machine learning. *In: 12th USENIX symposium on operating systems design and implementation (OSDI 16)*. [S.l.: s.n.], 2016. p. 265–283. 33

AKIBA, Takuya; SANO, Shotaro; YANASE, Toshihiko; OHTA, Takeru; KOYAMA, Masanori. Optuna: A next-generation hyperparameter optimization framework. *In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. [S.l.: s.n.], 2019. p. 2623–2631. 62

AKIMOTO, Hajime. **Atmospheric reaction chemistry**. [S.l.]: Springer, 2016. 50

ALVES, Marcelo S; STREIT, Livia; PIZZOLATO, Tânia Mara. Utilização de modelos in silico para avaliação da toxicidade de resíduos de agrotóxicos, fármacos e metabólitos em águas naturais. **Química Nova**, SciELO Brasil, v. 46, p. 881–889, 2023. 58

ALVES, Vinicius M; BRAGA, Rodolpho C; MURATOV, Eugene N; ANDRADE, Carolina Horta. Quimioinformática: uma introdução. **Química Nova**, SciELO Brasil, v. 41, n. 2, p. 202–212, 2018. 34, 35

AMES, Bruce N; LEE, Frank D; DURSTON, William E. An improved bacterial test system for the detection and classification of mutagens and carcinogens. **Proceedings of the National Academy of Sciences**, v. 70, n. 3, p. 782–786, 1973. 44

ATKINSON, Roger; AREY, Janet. Atmospheric degradation of volatile organic compounds. **Chemical reviews**, ACS Publications, v. 103, n. 12, p. 4605–4638, 2003. 50

BAIRD, Colin; CANN, Michael. **Química ambiental**. 4. ed. Porto Alegre: Bookman, 2011. ISBN 9788577808489. 22, 23, 50

BARBER, Victoria P; GREEN, William H; KROLL, Jesse H. Screening for new pathways in atmospheric oxidation chemistry with automated mechanism generation. **The Journal of Physical Chemistry A**, ACS Publications, v. 125, n. 31, p. 6772–6788, 2021. 19

BELCIC, Ivan; STRYKER, Cole. **What Is a Machine Learning Pipeline?** 2026. <https://www.ibm.com/think/topics/machine-learning-pipeline>. Acessado em: 25 mar. 2026. 34, 35

BENIGNI, Romualdo. Towards quantitative read across: Prediction of Ames mutagenicity in a large database. **Regulatory Toxicology and Pharmacology**, Elsevier, v. 108, p. 104434, 2019. 58, 60

BENIGNI, Romualdo; BATTISTELLI, Chiara Laura; BOSSA, Cecilia; TCHERMENSKAIA, Olga; CRETIAZ, Pierre. New perspectives in toxicological information management, and the role of ISSTOX databases in assessing chemical mutagenicity and carcinogenicity. **Mutagenesis**, v. 28, n. 4, p. 401–409, 2013. 44

BURLAKOVA, Elena B.; SHILOV, Alexander E.; VARFOLOMEEV, Sergey D.; ZAIKOV, Gennady E. **Chemical and Biological Kinetics New Horizons, Volume 1: Chemical Kinetics**. Boca Raton, USA: CRC Press, Taylor & Francis Group, 2005. ISBN 9780429074813. 50

CAREY, Francis A; SUNDBERG, Richard J. **Advanced organic chemistry: part A: structure and mechanisms**. [S.l.]: Springer, 2000. 27, 28

CAVASOTTO, Claudio N.; SCARDINO, Valeria. Machine Learning Toxicity Prediction: Latest Advances by Toxicity End Point. **ACS Omega**, v. 7, n. 51, p. 47536–47546, 2022. Disponível em: <https://doi.org/10.1021/acsomega.2c05693>. 58

CHEMAXON. **Marvin: Chemical Drawing Software**. 2024. <https://chemaxon.com/marvin>. Acessado em: 12 mar. 2026. 43

CHEN, Tianzeng; LIU, Jun; CHU, Biwu; GE, Yanli; ZHANG, Peng; MA, Qingxin; HE, Hong. Combined smog chamber/oxidation flow reactor study on aging of secondary organic aerosol from photooxidation of aromatic hydrocarbons. **Environmental Science & Technology**, ACS Publications, v. 57, n. 37, p. 13937–13947, 2023. 19

CHERKASOV, Artem; MURATOV, Eugene N; FOURCHES, Denis; VARNEK, Alexandre; BASKIN, Igor I; CRONIN, Mark; DEARDEN, John; GRAMATICA, Paola; MARTIN, Yvonne C; TODESCHINI, Roberto *et al.* QSAR modeling: where

have you been? Where are you going to? **Journal of medicinal chemistry**, ACS Publications, v. 57, n. 12, p. 4977–5010, 2014. 34

COSTA, Carla Regina; OLIVI, Paulo; BOTTA, Clarice MR; ESPINDOLA, Evaldo LG. A toxicidade em ambientes aquáticos: discussão e métodos de avaliação. **Química nova**, SciELO Brasil, v. 31, p. 1820–1830, 2008. 55

CRONIN, Mark TD; YOON, Miyoung. Computational methods to predict toxicity. *In: The history of alternative test methods in toxicology*. [S.l.]: Elsevier, 2019. p. 287–300. 58

CRUZ, Lícia PS; MOTA, Elisvan R; CAMPOS, Vânia P; SANTANA, Franciele O; LUZ, Sâmeque R; SANTOS, Daniela F. Inorganic and organic acids in the atmosphere of the urban area of the city of Salvador, Brazil. **Journal of the Brazilian Chemical Society**, SciELO Brasil, v. 30, p. 904–914, 2019. 19

DAI, Simin; CHEN, Xuwu; LIANG, Jie; LI, Xin; LI, Shuai; CHEN, Gaojie; CHEN, Zuo; BIN, Juan; TANG, Yifan; LI, Xiaodong. Response of PM_{2.5} pollution to meteorological and anthropogenic emissions changes during COVID-19 lockdown in Hunan Province based on WRF-Chem model. **Environmental Pollution**, Elsevier, v. 331, p. 121886, 2023. 18

DAYLIGHT CHEMICAL INFORMATION SYSTEMS, INC. **SMARTS — A Language for Describing Molecular Patterns**. 2007. <https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>. *In: Daylight Theory Manual*, Chapter 4. Acessado em: 12 mar. 2026. 39

DEB, Kalyanmoy; JAIN, Himanshu. An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part I: solving problems with box constraints. **IEEE Transactions on Evolutionary Computation**, IEEE, v. 18, n. 4, p. 577–601, 2013. 44

DUARTE, Julio Cesar; OLIVEIRA-FILHO, Antonio GS de; MÁXIMO-CANADAS, Matheus; SOUZA, Rubens C; JR, Itamar Borges. A survey of basic concepts and applications of machine learning to chemistry. **Journal of the Brazilian Chemical Society**, SciELO Brasil, v. 36, n. 8, p. e–20250082, 2025. 35, 36

ELTON, Daniel C; BOUKOUVALAS, Zois; FUGE, Mark D; CHUNG, Peter W. Deep learning for molecular design—a review of the state of the art. **Molecular Systems Design & Engineering**, Royal Society of Chemistry, v. 4, n. 4, p. 828–849, 2019. 34

FLORES-ROJAS, Nelida Cecilia; ESTERHUIZEN-LONDT, Maranda; PFLUGMACHER, Stephan. Antioxidative stress responses in the floating macrophyte *lemna minor* l. with cylindrospermopsin exposure. **Aquatic Toxicology**, v. 169, p. 188–195, 2015. ISSN 0166-445X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0166445X15300849>. 55

GASTWIRTH, Joseph L. The estimation of the lorenz curve and gini index. **The review of economics and statistics**, JSTOR, p. 306–316, 1972. 32

GAULTON, Anna; BELLIS, Louisa J; BENTO, A Patricia; CHAMBERS, Jon; DAVIES, Mark; HERSEY, Anne; LIGHT, Yvonne; MCGLINCHEY, Shaun; MICHALOVICH, David; AL-LAZIKANI, Bissan *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. **Nucleic acids research**, Oxford University Press, v. 40, n. D1, p. D1100–D1107, 2012. 44

GAVIDIA-CALDERÓN, Mario; SCHUCH, Daniel; VARA-VELA, Angel; INOUE, Rita; FREITAS, Edmilson D; ALBUQUERQUE, Taciana Toledo de A; ZHANG, Yang; ANDRADE, Maria de Fatima; BELL, Michelle L. Air quality modeling in the metropolitan area of São Paulo, Brazil: A review. **Atmospheric Environment**, Elsevier, v. 319, p. 120301, 2024. 19

GENTNER, Drew R; JATHAR, Shantanu H; GORDON, Timothy D; BAHREINI, Roya; DAY, Douglas A; HADDAD, Imad El; HAYES, Patrick L; PIEBER, Simone M; PLATT, Stephen M; GOUW, Joost de *et al.* Review of urban secondary organic aerosol formation from gasoline and diesel motor vehicle emissions. **Environmental science & technology**, ACS Publications, v. 51, n. 3, p. 1074–1093, 2017. 19

GÉRON, Aurélien. **Mãos à obra: aprendizado de máquina com Scikit-Learn & TensorFlow**. [S.l.]: Alta Books, 2019. 31, 32, 33, 34, 35

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron; BENGIO, Yoshua. **Deep learning**. [S.l.]: MIT press Cambridge, 2016. 31

GUAN, Yue-Ming; GUAN, Dong; ZHANG, Cheng; YUAN, Sheng-Hua; CAI, Guang-Qing; ZHANG, Lin-Zhou. Diesel molecular composition and blending modeling based on SU-BEM framework. **Petroleum Science**, Elsevier, v. 19, n. 2, p. 839–847, 2022. 44, 55

GULLI, Antonio; PAL, Sujit. **Deep learning with Keras**. [S.l.]: Packt Publishing Ltd, 2017. 33, 61

HASSNER, Alfred; GALLE, James E. Stereochemistry. I. n-phosphorylated aziridines. reaction of 2-iodoalkyl azides with phosphines and phosphites. **Journal of the American Chemical Society**, v. 92, n. 12, p. 3733–3739, 1970. Disponível em: <https://doi.org/10.1021/ja00715a033>. 50

HENDRIKS, A. Jan; TRAAS, Theo P.; HUIJBREGTS, Mark A. J. Critical Body Residues Linked to Octanol-Water Partitioning, Organism Composition, and LC50 QSARs: Meta-analysis and Model. **Environmental Science & Technology**, v. 39, n. 9, p. 3226–3236, 2005. PMID: 15926573. Disponível em: <https://doi.org/10.1021/es048442o>. 58

HUANG, Juan; FENG, Yanli; LI, Jian; XIONG, Bin; FENG, Jialiang; WEN, Sheng; SHENG, Guoying; FU, Jiamo; WU, Minghong. Characteristics of carbonyl compounds in ambient air of shanghai, china. **Journal of Atmospheric Chemistry**, Springer, v. 61, n. 1, p. 1–20, 2008. 36

ISTITUTO SUPERIORE DI SANITÀ. **ISSTOX: Chemical Toxicity Databases**. 2021. <https://www.iss.it/en/isstox>. Acessado em: 13 mar. 2026. 44, 58

JEWELL, Steven P; HOLBROOK, Kenneth A; OLDERSHAW, Geoffrey A. The reaction of atomic oxygen o (3p) with propane. **International Journal of Chemical Kinetics**, Wiley Online Library, v. 13, n. 1, p. 69–84, 1981. 50

_____. The reaction of atomic oxygen o (3p) with isobutane. **International Journal of Chemical Kinetics**, Wiley Online Library, v. 14, n. 5, p. 585–591, 1982. 50

JIANG, Changsheng; YANG, Hongbin; DI, Peiwen; LI, Weihua; TANG, Yun; LIU, Guixia. In silico prediction of chemical reproductive toxicity using machine learning. **Journal of applied toxicology**, Wiley Online Library, v. 39, n. 6, p. 844–854, 2019. 58

JÚNIOR, Aldo Muro. **Proposta de implementação de legislação ambiental, através de políticas de controle da poluição atmosférica, por meio de sistemas de monitoramento passivo**. Tese (Doutorado em Ciências Ambientais) — Universidade Federal de Goiás, Goiânia, 2013. 42, 43

KIM, Hyun; KANG, Dahyun; JUNG, Heon Young; JEON, Jongho; LEE, Jae Young. Review of smog chamber experiments for secondary organic aerosol formation. **Atmosphere**, MDPI, v. 15, n. 1, p. 115, 2024. 19

KRUK, I. Environmental toxicology and chemistry of oxygen species. *In: Reactions and Processes - Part I*. Berlin, Germany: Springer-Verlag, 1998, (The Handbook of Environmental Chemistry, v. 2). 50

LANDRUM, Greg. **RDKit: Open-Source Cheminformatics**. 2006. <https://www.rdkit.org>. Acessado em: 12 mar. 2026. 34, 36, 37

LENZI, Ervim; FAVERO, Luzia Otilia Bortotti. **Introdução à química da atmosfera: ciência, vida e sobrevivência**. Rio de Janeiro: LTC, 2009. ISBN 9788521634843. 22, 23

LIMA, Diego Sarmento Duncan; BORGES, Emilio. Determinação das constantes de velocidade dos mecanismos de formação do ozônio troposférico a partir de reações fotoquímicas envolvendo o acetaldeído como precursor. *In: Anais do XXI Simpósio Brasileiro de Química Teórica (SBQT)*. Galoá, 2021. Acesso em: 10 mar. 2026. Disponível em: <https://proceedings.science/sbqt-2021/trabalhos/determinacao-das-constantes-de-velocidade-dos-mecanismos-de-formacao-do-ozonio-t?lang=pt-br>. 23, 36

LIU, Yanhui; LI, Jiayin; MA, Yufang; ZHOU, Ming; TAN, Zhaofeng; ZENG, Limin; LU, Keding; ZHANG, Yuanhang. A review of gas-phase chemical mechanisms commonly used in atmospheric chemistry modelling. **Journal of Environmental Sciences**, Elsevier, v. 123, p. 522–534, 2023. 19

LUTZ, Mark. **Learning python: Powerful object-oriented programming**. [S.l.]: O'Reilly Media, Inc., 2025. 37

MAGALHÃES, D. P.; FILHO, A. da S. Ferrão; ESPÍNDOLA, E. L. G. Salomão; SILVA, A. C. Lima e; GOMES, M. V.; MENDES, M. F. **Ecotoxicologia: perspectivas para o século XXI**. São Carlos: Rima Editora, 2021. 55

MALLET, Vivien; SPORTISSE, Bruno. Air quality modeling: From deterministic to stochastic approaches. **Computers & Mathematics with Applications**, Elsevier, v. 55, n. 10, p. 2329–2337, 2008. 19

MANAHAN, Stanley E. **Environmental chemistry**. [S.l.]: CRC press, 2022. 50

MARCH, Jerry. **Advanced organic chemistry: reactions, mechanisms, and structure**. [S.l.]: Wiley New York, 1985. 28, 29

MARTINS, Leila Droprinchinski. **Sensibilidade da formação do ozônio troposférico às emissões veiculares na região metropolitana de São Paulo**.

Tese (Doutorado em Meteorologia) — Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Universidade de São Paulo, São Paulo, 2007. 19

MEHTA, Pankaj; BUKOV, Marin; WANG, Ching-Hao; DAY, Alexandre GR; RICHARDSON, Clint; FISHER, Charles K; SCHWAB, David J. A high-bias, low-variance introduction to machine learning for physicists. **Physics reports**, Elsevier, v. 810, p. 1–124, 2019. 31

MEYLAN, William M; HOWARD, Philip H. A review of quantitative structure-activity relationship methods for the prediction of atmospheric oxidation of organic chemicals. **Environmental toxicology and chemistry**, Wiley Periodicals, Inc. Hoboken, v. 22, n. 8, p. 1724–1732, 2003. 19

MILLER, Sharon K. R. Tutorial on atomic oxygen effects and contamination. *In*: NASA GLENN RESEARCH CENTER. **Applied Space Environments Conference**. Huntsville, Alabama, 2017. NASA Technical Reports Server, NTRS ID: 20170006623. Disponível em: <https://ntrs.nasa.gov/citations/20170006623>. 50

MULTICASE INC. **CASE Ultra**. 2025. <https://multicase.com/case-ultra/>. Acessado em: 25 mar. 2026. 58

NIAZI, Sarfaraz K; MARIAM, Zamara. Recent advances in machine-learning-based chemoinformatics: a comprehensive review. **International Journal of Molecular Sciences**, MDPI, v. 24, n. 14, p. 11488, 2023. 19

NKEMDILIM, Mbeledogu Njideka; UZOAMAKA, Paul Roseline; DANIEL, Ugoh; CHIDI, Mbeledogu Kaodilichukwu. An overview of supervised machine learning paradigms and their classifiers. **International Journal of Advanced Engineering, Management and Science**, v. 10, n. 3, p. 24–32, 2024. 34, 35

ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE. **Nove em cada dez pessoas em todo o mundo respiram ar poluído**. maio 2018. Acesso em: 10 mar. 2026. Disponível em: <https://www.paho.org/pt/noticias/1-5-2018-nove-em-cada-dez-pessoas-em-todo-mundo-respiram-ar-poluido>. 21

PARK, Sanggil; HAN, Herim; KIM, Hyungjun; CHOI, Sunghwan. Machine learning applications for chemical reactions. **Chemistry—An Asian Journal**, Wiley Online Library, v. 17, n. 14, p. e202200203, 2022. 19, 20, 27, 33, 34, 35, 36

PEDREGOSA, Fabian; VAROQUAUX, Gaël; GRAMFORT, Alexandre; MICHEL, Vincent; THIRION, Bertrand; GRISEL, Olivier; BLONDEL, Mathieu; PRETTENHOFER, Peter; WEISS, Ron; DUBOURG, Vincent *et al.* Scikit-learn: Machine

learning in Python. **The Journal of machine Learning research**, JMLR. org, v. 12, p. 2825–2830, 2011. 37

PETROBRAS. **Óleo Diesel: Informações Técnicas**. 2024. Versão setembro/2024. Disponível em: <https://petrobras.com.br/documents/2677942/25973428/Manual%20de%20Diesel%202024.pdf>. Acesso em: 25 mar. 2026. 56

RAJKIEWICZ, Maria; TYSZKIEWICZ, Waldemar; WERTEJUK, Zbigniew. **Chemistry and Physics of Complex Materials: Concepts and Applications**. Boca Raton, USA: Apple Academic Press Inc, Taylor & Francis Group, 2014. ISBN 9781482231281. 50

RAMSUNDAR, Bharath; EASTMAN, Peter; WALTERS, Patrick; PANDE, Vijay; LESWING, Karl; WU, Zhenqin. **Deep Learning for the Life Sciences**. O'Reilly Media, 2019. Disponível em: <https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837>. 36

RAMSUNDAR, Bharath; EASTMAN, Peter; WALTERS, Pat; PANDE, Vijay. **Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more**. [S.l.]: O'Reilly Media, 2019. 31, 34

RAPOSO, Alceu; ASSIS, Welligtosn Lopes de; NORTON, Shyanne Raposo. Revisão bibliográfica dos efeitos da poluição do ar sobre a saúde humana e uma breve análise crítica da atual legislação brasileira sobre os padrões de qualidade do ar (CONAMA 491/2018). **Cadernos do Leste**, v. 21, n. 21, 2021. 21

RDKit DEVELOPMENT TEAM. **The RDKit Book**. 2025. Version 2025.09.6. Acessado em: 13 mar. 2026. Disponível em: https://www.rdkit.org/docs/RDKit_Book.html. 53

ROGERS, David; HAHN, Mathew. Extended-connectivity fingerprints. **Journal of chemical information and modeling**, ACS Publications, v. 50, n. 5, p. 742–754, 2010. 55

ROY, Kunal; GHOSH, Gopinath. Exploring QSARs with Extended Topochemical Atom (ETA) indices for modeling chemical and drug toxicity. **Current pharmaceutical design**, Bentham Science Publishers direct, v. 16, n. 24, p. 2625–2639, 2010. 58

SAHU, Santosh Kumar; MOKHADE, Anil; BOKDE, Neeraj Dhanraj. An overview of machine learning, deep learning, and reinforcement learning-based techniques in quantitative finance: recent progress and challenges. **Applied Sciences**, MDPI, v. 13, n. 3, p. 1956, 2023. 33, 61

SEINFELD, John H; PANDIS, Spyros N. **Atmospheric chemistry and physics: from air pollution to climate change**. [S.l.]: John Wiley & Sons, 2016. 25, 30

SHARIFANI, Koosha; AMINI, Mahyar. Machine learning and deep learning: A review of methods and applications. **World Information Technology and Engineering Journal**, v. 10, n. 07, p. 3897–3904, 2023. 31

SILVA, Cleyton Martins da; ARBILLA, Graciela. **Emissões Atmosféricas e Mudanças Climáticas**. Rio de Janeiro: Freitas Bastos, 2022. 200 p. ISBN 978-65-5675-196-2. 24

SILVA, Matheus Muller Pereira Da; ANGELO, Jaqueline Silva; GUEDES, Isabella Alvim; DARDENNE, Laurent Emmanuel. A generative evolutionary many-objective framework: A case study in antimicrobial agent design. *In: Proceedings of the Genetic and Evolutionary Computation Conference Companion*. [S.l.: s.n.], 2024. p. 1623–1630. 55

SILVA, Sam J; EVANS, Mat. **Artificial intelligence and machine learning in atmospheric chemistry**. [S.l.]: ACS Publications, 2024. 330–331 p. 19, 20, 44

SOLOMONS, T. W. Graham; FRYHLE, Craig B.; SNYDER, Scott A. **Química Orgânica**. 13. ed. Rio de Janeiro: LTC, 2024. 640 p. ISBN 9788521638896. 26, 27, 29, 30

SOUZA, Maria Clara R. de; FREITAS, Brenda Ramos; FIGUEIREDO, Ana Paula de S.; VENIAL, Henrique Jordem; CORRADINI, Patricia Gon; SOUZA, Murilo de Oliveira. A Influência dos Compostos Orgânicos Voláteis nas Mudanças Climáticas: Uma breve revisão. **Revista Virtual de Química**, Sociedade Brasileira de Química, v. 15, n. 2, p. 227–240, 2022. 22

_____. _____. **Revista Virtual de Química**, v. 15, n. 2, p. 227–240, 2023. Disponível em: <http://dx.doi.org/10.21577/1984-6835.20220111>. 18

SOUZA, Maria Clara R. de; FREITAS, Brenda Ramos; FIGUEIREDO, Ana Paula de S.; VENIAL, Henrique Jordem; CORRADI, Patricia Gonçalves; SOUZA, Murilo de Oliveira. A influência dos compostos orgânicos voláteis nas mudanças climáticas: Uma breve revisão. **Revista Virtual de Química**, Sociedade Brasileira de Química, v. 15, n. 2, p. 227–240, 2023. Disponível em: <https://rvq.sbq.org.br/pdf/v15n2a04>. 55

SPEIGHT, James G. **The chemistry and technology of petroleum**. [S.l.]: CRC press, 2006. 56

SPIRO, Thomas G.; STIGLIANI, William M. **Química ambiental**. 2. ed. Pearson, 2008. ISBN 9788576051961. Disponível em: <https://www.amazon.com.br/dp/8576051966>. 21, 22, 23, 24, 25, 26, 27, 29

TCHOUNWOU, Paul B; YEDJOU, Clement G; PATLOLLA, Anita K; SUTTON, Dwayne J. Heavy metal toxicity and the environment. **Molecular, clinical and environmental toxicology: volume 3: environmental toxicology**, Springer, p. 133–164, 2012. 55

United Nations Environment Programme. **Actions on Air Quality: A Global Summary of Policies and Programmes to Reduce Air Pollution**. 2021. Acesso em: 11 mar. 2026. Disponível em: <https://www.unep.org/resources/report/actions-air-quality-global-summary-policies-and-programmes-reduce-air-pollution>. 25

VANROSSUM, Guido; DRAKE, Fred L. **The Python language reference**. [S.l.]: Python Software Foundation Amsterdam, The Netherlands, 2010. 37

VENKATRAMAN, Gopinath; GIRIBABU, Nelli; MOHAN, Priyadarshini Sakthi; MUTTIAH, Barathan; GOVINDARAJAN, Venkat Kumar; ALAGIRI, Mani; RAHMAN, Puteri Shafinaz Abdul; KARSANI, Saiful Anuar. Environmental impact and human health effects of polycyclic aromatic hydrocarbons and remedial strategies: A detailed review. **Chemosphere**, Elsevier, v. 351, p. 141227, 2024. 18, 43

VORMITTAG, Evangelina da Motta PA de Araújo; CIRQUEIRA, Samirys Sara Rodrigues; WICHER, Hélio; SALDIVA, Paulo Hilário N. Análise do monitoramento da qualidade do ar no Brasil. **Estudos Avançados**, SciELO Brasil, v. 35, n. 102, p. 7–30, 2021. 18

WANG, Wei-Chun; DAT, Nguyen Duy; CHI, Kai-Hsien; CHANG, Moo-Been. Characterization of pm2.5 and particulate pahs emitted from vehicles via tunnel sampling in different time frames. **Aerosol and Air Quality Research**, Springer Science and Business Media LLC, v. 21, n. 12, p. 210074, 2021. ISSN 2071-1409. Disponível em: <http://dx.doi.org/10.4209/aaqr.210074>. 50

WARK, Kenneth; WARNER, Cecil Francis. **Air pollution: its origin and control**. [S.l.]: Harper and Row Publishers, New York, NY, 1980. 24

WEININGER, David. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. **Journal of chemical information and computer sciences**, ACS Publications, v. 28, n. 1, p. 31–36, 1988. 38, 39

WU, Z.; CHEN, X.; WANG, Z.; CHEN, H.; WANG, Z.; MU, Q.; WU, L.; WANG, W.; TANG, X.; LI, J.; LI, Y.; WU, Q.; WANG, Y.; ZOU, Z.; JIANG, Z. Modeling of polycyclic aromatic hydrocarbons (PAHs) from global to regional scales: model development (IAP-AACM_PAH v1.0) and investigation of health risks in 2013 and 2018 in China. **Geoscientific Model Development**, v. 17, n. 24, p. 8885–8907, 2024. Disponível em: <https://gmd.copernicus.org/articles/17/8885/2024/>. 18

YAO, Qinzuo; SHI, Jianwu; HAN, Xinyu; TIAN, Senlin; HUANG, Jianhong; LI, Yingjie; NING, Ping. Emissions of polycyclic aromatic hydrocarbons in PM_{2.5} emitted from motor vehicles exhaust (PAHs-PM_{2.5}-MVE) under the plateau with low oxygen content. **Atmospheric Environment**, Elsevier, v. 321, p. 120364, 2024. 42

ZHOU, Liyuan; SALVADOR, Christian M; PRIESTLEY, Michael; HALLQUIST, Mattias; LIU, Qianyun; CHAN, Chak K; HALLQUIST, Åsa M. Emissions and secondary formation of air pollutants from modern heavy-duty trucks in real-world traffic — Chemical characteristics using on-line mass spectrometry. **Environmental science & technology**, ACS Publications, v. 55, n. 21, p. 14515–14525, 2021. 18, 19, 30, 36

Apêndice A – Padrões de reação e códigos SMARTS

Este apêndice apresenta os padrões de reação química (ver Tabela 1) montados no software Marvin Sketch e seus correspondentes códigos SMARTS

Reações com radical $\cdot\text{OH}$

Reação #1

Figura 14 – *Ataque nas ligações duplas das cadeias aromáticas*



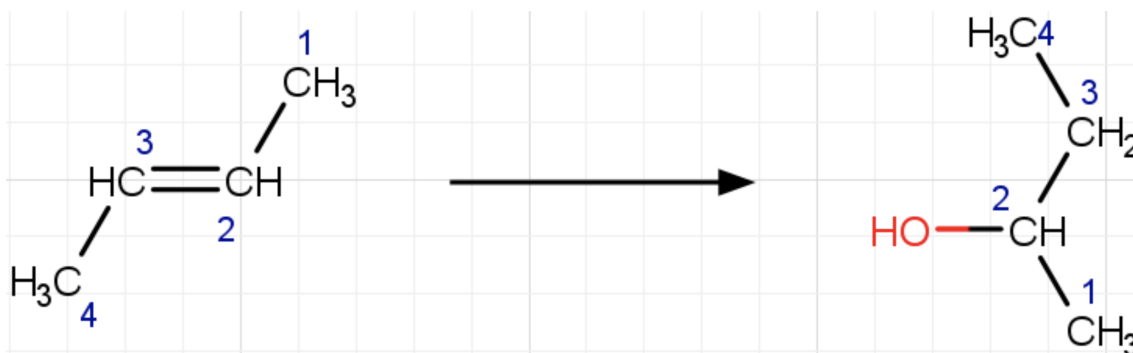
Fonte: Elaborada pelos autores.

Código SMARTS:

```
[c:1]:[c:2]:[c:3]>>[#8]-[c:2](:[c:1]):[c:3]
```

Reação #2

Figura 15 – Ataque às ligações duplas das cadeias alifáticas.



Fonte: Elaborada pelos autores.

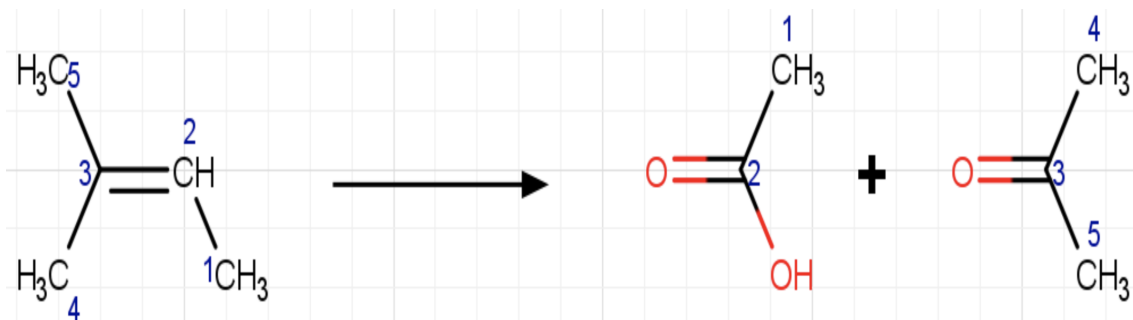
Código SMARTS:

[#6:1]\[#6:2]=[#6:3]\[#6:4]>>[#6:4]-[#6:3]-[#6:2](-[#6:1])-[#8]

Reações com Ozônio

Reação #3

Figura 16 – Ataque às ligações duplas de cadeias alifáticas.



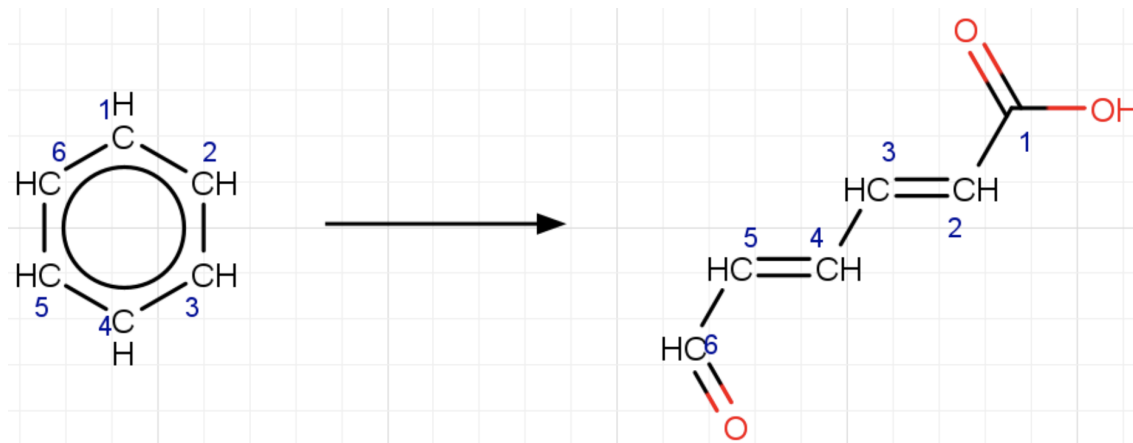
Fonte: Elaborada pelos autores.

Código SMARTS:

[H:5]\[#6:4]=[#6:3](\ [H:1]) [H:2]>>[H:5] [#6:4] (-[#8])=O. [H:1] [#6:3] ([H:2])=O

Reação #4

Figura 17 – *Ataque nas ligações duplas das cadeias aromáticas.*



Fonte: Elaborada pelos autores.

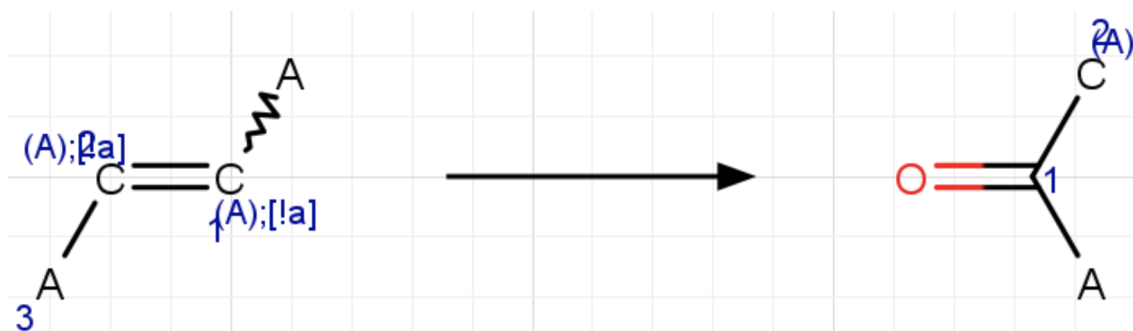
Código SMARTS:

```
[c:1]1[c:2][c:3][c:4][c:5][c:6]1>>[#8]-[#6:1](=O)\[#6:2]=[#6:3]\[#6:4]=[#6:5]\[#6:6]=O
```

Reações com oxigênio atômico

Reações envolvendo cadeias alifáticas: três mecanismos e produtos distintos (Reação #5)

Figura 18 – *Reações com cadeias alifáticas, mecanismo A.*



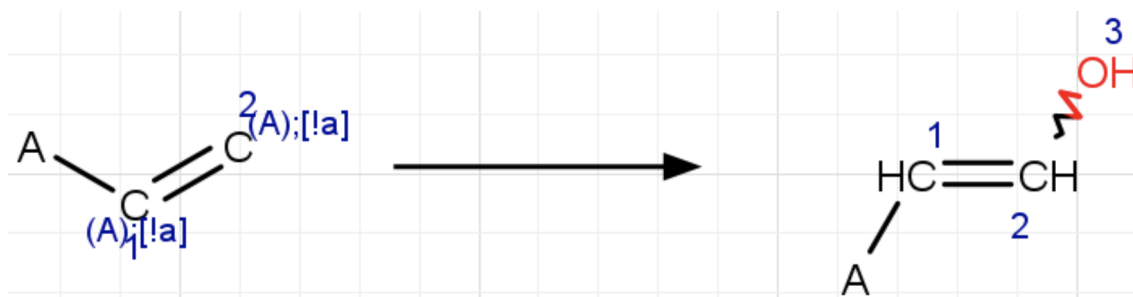
Fonte: Elaborada pelos autores.

Código SMARTS:

```
[*:0]-[C;!a:1]=[C;!a:2]-[*:3]>>[C:2]-[C:1](-[*:0])=O
```

Reação #6

Figura 19 – Reações com cadeias alifáticas, mecanismo B.



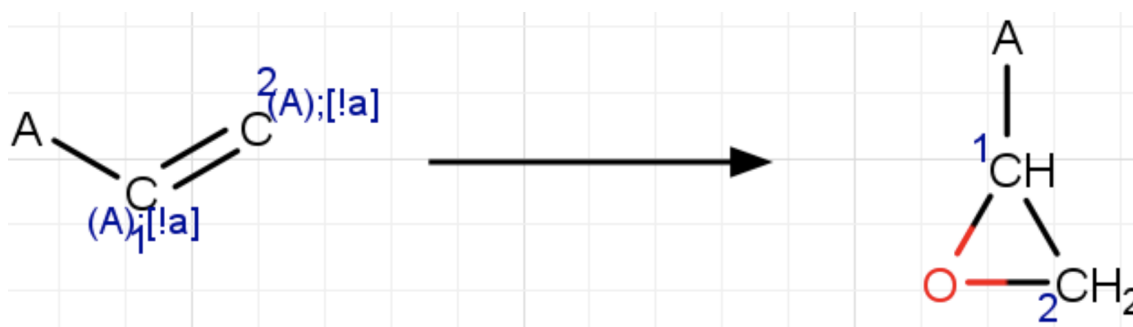
Fonte: Elaborada pelos autores.

Código SMARTS:

[*:0]-[C;!a:1]=[C;!a:2]>>[0:3]-[C:2]=[C:1]-[*:0]

Reação #7

Figura 20 – Reações envolvendo cadeias alifáticas, mecanismo C.



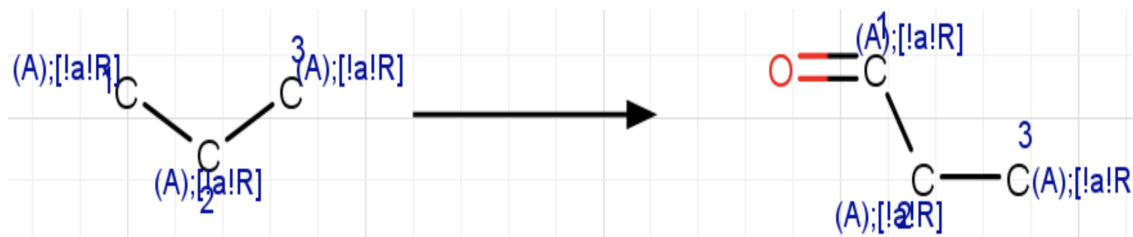
Fonte: Elaborada pelos autores.

Código SMARTS:

[*:0]-[C;!a:1]=[C;!a:2]>>[*:0]-[C:1]-1-[C:2]-[O]-1

Reação #8

Figura 21 – Ataque às ligações simples em alcanos.



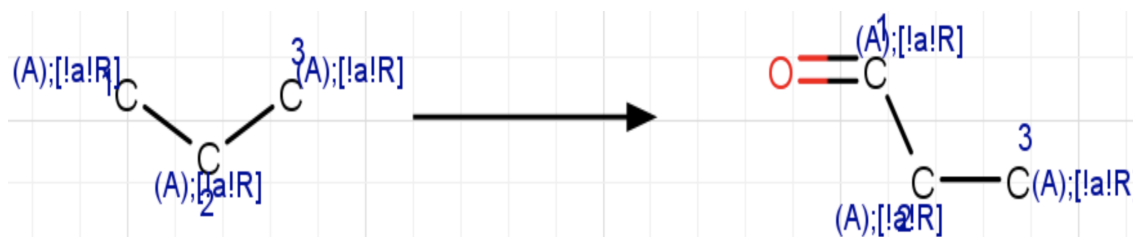
Fonte: Elaborada pelos autores.

Código SMARTS:

```
[C;!a;!R:1]-[C;!a;!R:2]-[C;!a;!R:3]>>[C;!a;!R:3]-
[C;!a;!R:2]-[C;!a;!R:1]=O
```

Reação #8

Figura 22 – Ataque às ligações simples em alcanos.



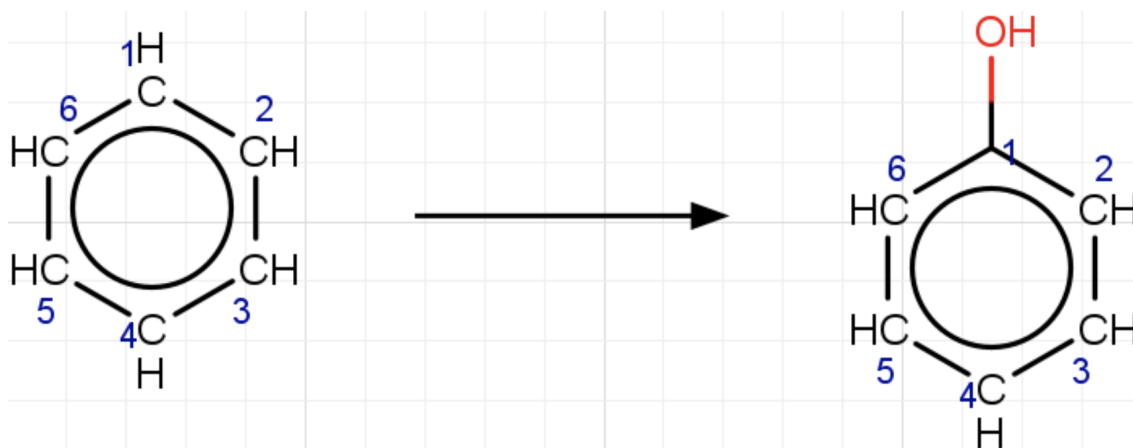
Fonte: Elaborada pelos autores.

Código SMARTS:

```
[C;!a;!R:1]-[C;!a;!R:2]-[C;!a;!R:3]>>[C;!a;!R:3]-
[C;!a;!R:2]-[C;!a;!R:1]=O
```

Reação #9

Figura 23 – *Ataque aos hidrocarbonetos aromáticos.*



Fonte: Elaborada pelos autores.

Código SMARTS:

```
[c:1]1[c:2][c:3][c:4][c:5][c:6]1>>[c:1]1([O])[c:2][c:3][c:4][c:5][c:6]
```