Universidade Federal de Goiás
Programa de Pós-Graduação em Química
Instituto de Química

THERMOPRED: APLICAÇÃO DE INTELIGÊNCIA ARTIFICIAL E QUÍMICA QUÂNTICA NA PREDIÇÃO DE PROPRIEDADES TERMOQUÍMICAS E ESPONTANEIDADE DE REAÇÕES DE DEGRADAÇÃO EM INSUMOS FARMACÊUTICOS ATIVOS

Diullio Pereira dos Santos

Goiânia
2026

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE QUÍMICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC n° 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[ ] Dissertação [x] Tese [ ] Outro*:_____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

Exemplos: Estudo de caso ou Revisão sistemática ou outros formatos.

2. Nome completo do autor

Diullio Pereira dos Santos

3. Título do trabalho

"ThermoPred: AI-Enhanced Quantum Chemistry Data Set and ML Toolkit for Thermochemical Properties of API-Like Compounds and Their Degradants"

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Heibbe Cristhian Benedito De Oliveira**, **Professor do Magistério Superior**, em 24/03/2026, às 16:23, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **6046507** e o código CRC **012A39F7**.

Diullio Pereira dos Santos
Químico Bacharel

THERMOPRED: APLICAÇÃO DE INTELIGÊNCIA ARTIFICIAL E QUÍMICA QUÂNTICA NA PREDIÇÃO DE PROPRIEDADES TERMOQUÍMICAS E ESPONTANEIDADE DE REAÇÕES DE DEGRADAÇÃO EM INSUMOS FARMACÊUTICOS ATIVOS

Orientador:

Prof. Dr. **HEIBBE CRISTHIAN B. DE OLIVEIRA**

Tese apresentada ao Programa de Pós-Graduação em Química, do instituto de química da Universidade Federal de Goiás (UFG), como requisito para obtenção do título de Doutor em Química. Área de concentração: Química. Linha de pesquisa: Química Computacional.

Goiânia
2026

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Santos, Diullio Pereira dos
THERMOPRED: APLICAÇÃO DE INTELIGÊNCIA ARTIFICIAL E QUÍMICA QUÂNTICA NA PREDIÇÃO DE PROPRIEDADES TERMOQUÍMICAS E ESPONTANEIDADE DE REAÇÕES DE DEGRADAÇÃO EM INSUMOS FARMACÊUTICOS ATIVOS [e-books] / Diullio Pereira dos Santos. - 2026.
LX, 60 f.: 2026

Orientador: Prof. Dr. HEIBBE CRISTHIAN BENEDITO DE OLIVEIRA
Tese (Doutorado) - Universidade Federal de Goiás, Instituto de Química (IQ), Programa de Pós-Graduação em Química, Goiânia, 2026.
Bibliografia.
Inclui: siglas, simbolos, lista de figuras, lista de tabelas.

1. Propriedades Termodinâmicas. 2. Rdc 964/2025. 3. Produtos de Degradação. 4. Inteligência Artificial. 5. Xgboost.

I. OLIVEIRA, HEIBBE CRISTHIAN BENEDITO DE, orient. II. Título.

CDU 54

UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE QUÍMICA

ATA DE DEFESA DE TESE

Ata Nº **1** da sessão de Defesa de Tese de **Diullio Pereira dos Santos** que confere o título de **Doutor em Química**, na área de concentração em **Química.**

Aos **vinte e sete dias do mês de fevereiro dois mil e vinte e seis**, a partir das **09h:00**, no **Laboratório 223 do IQ 1**, realizou-se a sessão pública de Defesa de Tese intitulada ***"ThermoPred: AI-Enhanced Quantum Chemistry Data Set and ML Toolkit for Thermochemical Properties of API-Like Compounds and Their Degradants"***. Os trabalhos foram instalados pelo Orientador, Professor Doutor **Heibbe Cristhian Benedito de Oliveira (IQ – UFG)** com a participação dos demais membros da Banca Examinadora: Doutora **Alany Ingrid Ribeiro (Synvia)**; Doutor **Alisson Moraes e Silva (Brainfarma)**, Professor Doutor **Daniel Scalabrini Machado (IQ - UnB)**, Professor Doutor **Guilherme Colherinhas (IF – UFG)**. Durante a arguição, os membros da banca **não fizeram** sugestão de alteração do título do **trabalho**. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Tese, tendo sido o candidato **aprovado** pelos seus membros. Proclamados os resultados pelo Professor Doutor **Heibbe Cristhian Benedito de Oliveira**, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos **vinte e sete dias do mês de fevereiro dois mil e vinte e seis.**

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Heibbe Cristhian Benedito De Oliveira**, **Professor do Magistério Superior**, em 11/03/2026, às 12:06, conforme horário oficial de Brasília, com fundamento no § 3° do art. 4° do Decreto n° 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Alany Ingrid Ribeiro**, **Usuário Externo**, em 11/03/2026, às 13:41, conforme horário oficial de Brasília, com fundamento no § 3° do art. 4° do Decreto n° 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Guilherme Colherinhas De Oliveira**, **Professor do Magistério Superior**, em 11/03/2026, às 14:43, conforme horário oficial de Brasília, com fundamento no § 3° do art. 4° do Decreto n° 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Alisson Moraes E Silva**, **Usuário Externo**, em 12/03/2026, às 11:07, conforme horário oficial de Brasília, com fundamento no § 3° do art. 4° do Decreto n° 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Daniel Francisco Scalabrini Machado**, **Usuário Externo**, em 26/03/2026, às 12:31, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **6040917** e o código CRC **8C646D9A**.

Referência: Processo nº 23070.004315/2026-45 SEI nº 6040917

"A tarefa não é ver o que ninguém ainda viu,
mas pensar o que ninguém ainda pensou,
sobre o que todos veem".
Arthur Schopenhauer

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AD	Applicability Domain (Domínio de Aplicabilidade)
Adam	Adaptive Moment Estimation (Otimizador)
AI	Inteligência Artificial
AMU	Atomic Mass Unit
ANVISA	Agência Nacional de Vigilância Sanitária
API	Interface de Programação de Aplicações
API-like	Moléculas semelhantes a Insumos Farmacêuticos Ativos
ANI	Atomic Neural Network Potentials
B3LYP	Becke, 3-parameter, Lee–Yang–Parr Functional
BOT	Bottom of the Well (referência zero de energia vibracional)
CB	Conjunto Bootstrap (contexto ML)
DCB	Denominação Comum Brasileira
DFT	Density Functional Theory (Teoria do Funcional da Densidade)
EMA	European Medicines Agency (Agência Europeia de Medicamentos)
FDA	Food and Drug Administration
FTIR	Espectroscopia de Infravermelho por Transformada de Fourier
GNN	Graph Neural Network
GPL-3.0	GNU General Public License v3.0
HPC	High Performance Computing
HPLC	High-Performance Liquid Chromatography
ICH	International Council for Harmonisation
IFAs	Insumos Farmacêuticos Ativos
InChI	International Chemical Identifier
InChIKey	Identificador derivado do InChI (forma compacta)
LC-MS	Cromatografia Líquida acoplada à Espectrometria de Massas
MD	Molecular Dynamics (Dinâmica Molecular)
ML	Machine Learning (Aprendizado de Máquina)
MLP	Multilayer Perceptron (Perceptron Multicamadas)
MSE	Mean Squared Error (Erro Quadrático Médio)
OC20	Open Catalyst Project 2020 Dataset
OC22	Open Catalyst Project 2022 Dataset
OMS	Organização Mundial da Saúde
PCA	Principal Component Analysis
PDF	Portable Document Format
PM6	Parametric Method 6

PM7	Parametric Method 7
QbD	Quality by Design
QM9	Quantum Machine Dataset 9
QSAR	Quantitative Structure–Activity Relationship
QSPR	Quantitative Structure–Property Relationship
RDKit	Toolkit de química computacional open-source
RDC	Resolução da Diretoria Colegiada
ReLU	Rectified Linear Unit (Função de ativação)
RF	Random Forest
RMN	Ressonância Magnética Nuclear
RMSE	Root Mean Squared Error (Erro Quadrático Médio da Raiz)
SCF	Self-Consistent Field (Energia Eletrônica Convergida)
SMILES	Simplified Molecular Input Line Entry Specification
SPICE	Sparse Interaction Chemical Energies Dataset
TSV	Tab-Separated Values
UMAP	Uniform Manifold Approximation and Projection
UV–Vis	Espectroscopia Ultravioleta–Visível
XYZ	Formato cartesiano tridimensional para moléculas
xTB	eXtended Tight Binding (Método semiempírico)
ZPE	Zero-Point Energy (Correção vibracional de ponto zero)

LISTA DE SÍMBOLOS

α	Taxa de aprendizado no MLP
b	Viés de neurônio
C_r	Capacidade térmica rotacional
C_V	Capacidade térmica a volume constante
e	Base do logaritmo natural
E	Energia térmica interna
E_t	Energia interna translacional
E_r	Energia interna rotacional
E_V	Energia interna vibracional
E_{SCF}	Energia eletrônica obtida pela solução SCF
$f_t(x_i)$	Modelo adicionado na iteração t (Boosting)
g_i	Gradiente da loss para amostra i
g_0	Multiplicidade de spin eletrônico
h	Constante de Planck
h_i	Hessiano da loss para amostra i
I	Momento de inércia molecular
k_B	Constante de Boltzmann
$\ln Q_{\text{elec}}$	Logaritmo natural da função de partição eletrônica
$\ln Q_{\text{rot}}$	Logaritmo natural da função de partição rotacional
$\ln Q_{\text{trans}}$	Logaritmo natural da função de partição translacional
$\ln Q_{\text{vib}}$	Logaritmo natural da função de partição vibracional
L_t	Função de perda na iteração t
m	Massa molecular
n	Número de amostras ou número de partículas
n_j	Soma ponderada antes da ativação do neurônio
N	Número total de partículas
N_A	Número de Avogadro
O_j	Saída do neurônio j
P	Pressão
q	Função de partição total
q_e	Função de partição eletrônica
q_r	Função de partição rotacional
q_t	Função de partição translacional
q_v	Função de partição vibracional
Q^2_{ext}	Coeficiente de correlação preditivo externo

Q^2_{int}	Coeficiente de correlação preditivo interno
R	Constante dos gases
R^2_{ext}	Coeficiente de determinação externo
R^2_{int}	Coeficiente de determinação interno
RMSE	Root Mean Squared Error
r^2_{ext}	Coeficiente de correlação externo ao quadrado
r^2_{int}	Coeficiente de correlação interno ao quadrado
S	Entropia total
S_t	Entropia translacional
S_r	Entropia rotacional
S_V	Entropia vibracional
T	Temperatura absoluta
UMAP	Uniform Manifold Approximation and Projection
Θ_r	Temperatura rotacional
$\theta_{v,K}$	Temperatura vibracional do modo K
U	Energia interna microscópica
V	Volume
w_j	Peso associado à folha j
x_i	Vetor de entrada da amostra i
y_i	Valor real (verdade de base)
$\hat{y}_i$	Valor previsto pelo modelo
λ	Regularização L2 (XGBoost)
γ	Parâmetro mínimo de ganho para split (XGBoost)
ω_n	Degenerescência eletrônica do nível energético

RESUMO

No presente estudo, foram desenvolvidos modelos de inteligência artificial para prever propriedades termodinâmicas associadas à degradação de insumos farmacêuticos ativos (IFAs), com ênfase na energia livre de Gibbs e na entalpia. A pesquisa está alinhada à RDC nº 964/2025, publicada pela Anvisa, que estabelece requisitos gerais para estudos de degradação forçada e define parâmetros para notificação, identificação e qualificação de produtos de degradação. A nova regulamentação reforça a importância do conhecimento aprofundado do comportamento de degradação dos IFAs em consonância com padrões internacionais.

Foi construído um banco de dados contendo mais de 14 mil estruturas químicas, incluindo IFAs e seus produtos de degradação, selecionados a partir da lista DCB e de bases públicas como o PubChem. As propriedades termodinâmicas foram calculadas no *Gaussian 16* utilizando o método M06-2X/6-31G(d), gerando descritores fisicoquímicos empregados no treinamento dos modelos.

Três algoritmos de aprendizado de máquina — *XGBoost*, *Random Forest* e *Perceptron Multicamadas (MLP)* — foram avaliados utilizando Q^2, R^2 e *RMSE* como métricas de desempenho. O *XGBoost* apresentou o melhor desempenho global, com Q^2 de 0,9947 e *RMSE* de 0,0137 para energia livre de Gibbs, além de resultados igualmente consistentes para entalpia. A validação interna por *StratifiedKFold* e a validação externa confirmaram a robustez e a capacidade de generalização dos modelos, com o *MLP* demonstrando desempenho ligeiramente superior em cenários específicos de predição de entalpia.

As moléculas foram representadas por *Morgan fingerprints* (RDKit), e a similaridade estrutural foi avaliada pela métrica de Tanimoto. A normalização das variáveis termodinâmicas, seguida da reversão para a escala original após a predição, assegurou consistência físico-química aos resultados.

Os resultados evidenciam o potencial da integração entre química computacional e inteligência artificial como ferramenta preditiva para compreensão de mecanismos de degradação e antecipação de propriedades termodinâmicas fundamentais de IFAs. A base de dados estruturada possibilita ainda sua expansão com novos descritores moleculares, níveis teóricos de cálculo e integração com dados experimentais, bem como aplicações em modelagem cinética de degradação, avaliação de estabilidade e suporte a estudos de bioequivalência, contribuindo para decisões mais eficientes no desenvolvimento farmacêutico e no planejamento regulatório.

Palavras-chave: 1.Propriedades termodinâmicas 2.RDC 964/2025 3.Produtos de degradação 4.Inteligência artificial 5.XGBoost 6.Random Forest 7.Perceptron Multicamadas

ABSTRACT

In this study, artificial intelligence models were developed to predict thermodynamic properties associated with the degradation of active pharmaceutical ingredients (APIs), with emphasis on Gibbs free energy and enthalpy. This research is aligned with the Brazilian regulatory guideline RDC No. 964/2025, issued by the National Health Surveillance Agency (Anvisa), which establishes general requirements for forced degradation studies and defines parameters for reporting, identifying, and qualifying degradation products. The updated regulation reinforces the importance of understanding API degradation behavior within a harmonized international framework.

A dataset comprising more than 14,000 molecular structures, including APIs and their degradation products, was assembled based on the Brazilian Common Denominations (DCB) list and public chemical databases such as PubChem. Thermodynamic properties were calculated using Gaussian 16 at the M06-2X/6-31G(d) level of theory, generating physicochemical descriptors for model training.

Three machine learning algorithms — XGBoost, Random Forest, and a Multilayer Perceptron (MLP) — were evaluated using Q^2, R^2, and root mean square error (RMSE). XGBoost achieved the best overall performance, with a Q^2 of 0.9947 and an RMSE of 0.0137 for Gibbs free energy, and similarly strong results for enthalpy. Internal StratifiedK-Fold cross-validation and external validation confirmed the robustness and generalization capacity of the models, with the MLP showing slightly improved performance in specific enthalpy prediction scenarios.

Molecular structures were encoded using Morgan fingerprints (RDKit), and structural similarity was assessed via the Tanimoto coefficient. Normalization and inverse transformation of thermodynamic variables ensured physical consistency of predictions.

These findings highlight the value of integrating computational chemistry and artificial intelligence as a predictive framework for understanding degradation mechanisms and anticipating key thermodynamic properties of APIs. The curated dataset further enables future expansion toward additional molecular descriptors, higher-level theoretical calculations, degradation kinetics modeling, and applications related to stability assessment and bioequivalence studies, supporting more efficient pharmaceutical development and regulatory decision-making.

Keywords: 1.Thermodynamic properties 2.RDC 964/2025 3.Degradation products 4.Artificial intelligence 5.XGBoost 6.Random Forest 7.Multilayer Perceptron

1 INTRODUÇÃO

A estabilidade de medicamentos constitui um dos fundamentos estruturantes da garantia da qualidade farmacêutica, estando diretamente relacionada à manutenção da eficácia terapêutica, da segurança do paciente e da conformidade regulatória ao longo de todo o ciclo de vida do produto. A degradação de um insumo farmacêutico ativo (IFA) pode resultar na redução de potência, na formação de impurezas potencialmente tóxicas ou mutagênicas e na perda de confiabilidade clínica, configurando risco sanitário significativo. Conforme estabelecido pela Organização Mundial da Saúde, a estabilidade corresponde à capacidade do medicamento de manter suas propriedades físicas, químicas, microbiológicas e terapêuticas dentro de limites especificados durante o período de validade [1].

A degradação química de IFAs pode ocorrer por múltiplos mecanismos, incluindo hidrólise, oxidação, fotólise, racemização, ciclizações intramoleculares e rearranjos estruturais, sendo influenciada por fatores ambientais, como temperatura, umidade e luz, bem como por características intrínsecas da molécula e pelo microambiente da formulação [2]. Estudos recentes demonstram que pequenas variações estruturais podem alterar significativamente a susceptibilidade à degradação, modificando rotas reacionais, estabilidade relativa de intermediários e tendências energéticas [3, 4]. A diversidade funcional e estrutural típica dos IFAs amplia ainda mais essa complexidade, tornando a previsão de comportamento degradativo um desafio científico e tecnológico de grande relevância.

O impacto da degradação ultrapassa o domínio teórico. Episódios recentes envolvendo formação de nitrosaminas em medicamentos amplamente utilizados evidenciaram que rotas degradativas não plenamente compreendidas podem resultar em recolhimentos globais, revisões regulatórias extensivas e significativa repercussão sanitária [5, 6]. Esses eventos reforçaram a necessidade de identificar riscos potenciais de forma antecipada, estimulando o desenvolvimento de abordagens mais preditivas e fundamentadas em conhecimento mecanístico.

Nesse contexto, a estabilidade ocupa uma posição estratégica na ciência regulatória internacional. Diretrizes do International Council for Harmonisation (ICH), como o Q1A(R2), estabelecem princípios para estudos de estabilidade sob condições aceleradas e de longa duração [7]. O ICH Q3B define critérios para identificação, qualificação e controle de produtos de degradação em medicamentos terminados [8], enquanto o ICH M7 introduz avaliação baseada em risco para impurezas mutagênicas [9]. Essas diretrizes evidenciam uma expectativa crescente de que os fabricantes detenham conhecimento aprofundado sobre o comportamento químico de seus IFAs, incluindo potenciais rotas degradativas.

No Brasil, essa convergência regulatória também se manifesta de forma clara. A RDC nº 53/2015 estabelece critérios específicos para produtos de degradação de IFAs sintéticos e semissintéticos [10], enquanto a RDC nº 318/2019 atualiza parâmetros para estudos de estabilidade [11]. A recente RDC nº 964/2025 reforça a incorporação de abordagens baseadas em conhecimento e gerenciamento de risco, alinhadas aos princípios de Quality by Design (QbD)

[12].

Adicionalmente, a RDC nº 677/2022, ao estabelecer regras para avaliação de risco, testes confirmatórios e controle de nitrosaminas potencialmente carcinogênicas, reforça a responsabilidade técnica do detentor do registro/fabricante em demonstrar domínio científico sobre seu produto e seu processo. Embora não trate de modelagem termoquímica de forma explícita, a norma exige uma análise sistemática e documentada dos fatores intrínsecos e extrínsecos que podem levar à formação/introdução de nitrosaminas ao longo do ciclo de vida do medicamento, incluindo: rota e impurezas do IFA (e potenciais precursores), excipientes, materiais de embalagem, condições e etapas de fabricação, armazenamento, além de possíveis contribuições de degradação e contaminação cruzada. Nesse sentido, em conjunto com a RDC nº 964/2025, a RDC nº 677/2022 consolida um cenário regulatório que privilegia decisões baseadas em conhecimento mecanístico e gestão de risco, isto é, compreensão das causas, rotas de formação e condições favorecedoras de impurezas, e não apenas a verificação empírica posterior por ensaios de controle.

Historicamente, o atendimento a essas exigências baseia-se predominantemente em métodos experimentais. A cromatografia líquida de alta eficiência (HPLC), frequentemente acoplada à espectrometria de massas (LC–MS), constitui o padrão ouro para detecção e quantificação de impurezas [13]. Estudos de degradação forçada permitem explorar rotas potenciais sob condições extremas [14], enquanto técnicas espectroscópicas, como FTIR e RMN, contribuem para a elucidação estrutural de degradantes [15, 16].

Essas metodologias são indispensáveis e permanecem como requisito regulatório. Contudo, apresentam limitações inerentes. A identificação de produtos de degradação depende da ocorrência efetiva da reação sob condições experimentais, exigindo planejamento detalhado, múltiplas análises e, frequentemente, padrões de referência específicos. Esses procedimentos implicam custos elevados e prazos prolongados [17–19]. Além disso, tratam-se de abordagens essencialmente reativas: detectam produtos após sua formação, não permitindo antecipar sistematicamente quais rotas são termodinamicamente mais favoráveis antes da experimentação.

Nesse cenário, a química computacional surge como ferramenta complementar capaz de investigar propriedades moleculares e perfis energéticos de forma mecanística. A Teoria do Funcional da Densidade (DFT) possibilita estimar energias eletrônicas, geometrias e propriedades termoquímicas relevantes [20, 21]. Parâmetros como energia livre de Gibbs (ΔG) e entalpia (ΔH) são particularmente importantes, pois permitem avaliar espontaneidade e balanço energético de reações químicas, incluindo processos degradativos.

A determinação dessas propriedades requer o cálculo da energia eletrônica associado a correções vibracionais e térmicas fundamentadas na termodinâmica estatística [22]. Funcionais como M06-2X demonstram desempenho robusto para sistemas orgânicos relevantes ao contexto farmacêutico [23, 24], sendo amplamente implementados em softwares como o Gaussian 16 [25]. Entretanto, o custo computacional cresce significativamente com o tamanho molecular, dificultando a aplicação direta a grandes bibliotecas de compostos farmacêuticos.

Paralelamente, o avanço da inteligência artificial tem ampliado a capacidade preditiva em química molecular [26]. Modelos treinados em bases como QM9 e ANI-1 alcançaram elevada precisão na previsão de energias moleculares [27, 28]. Contudo, essas bases concentram-se predominantemente em moléculas pequenas e estruturalmente simples, não refletindo a diversidade funcional e complexidade estrutural típica dos IFAs. Essa limitação compromete a generalização de modelos para o contexto farmacêutico e para a previsão termoquímica de rotas de degradação.

Emerge, portanto, uma lacuna científica e tecnológica: inexistem repositórios termoquímicos dedicados exclusivamente a IFAs e seus produtos de degradação, tampouco modelos preditivos treinados especificamente para estimar parâmetros termodinâmicos associados à estabilidade farmacêutica. Ademais, não se identificam frameworks integrados que combinem geração sistemática de dados quânticos, modelagem preditiva e rastreabilidade compatível com diretrizes regulatórias nacionais.

Diante desse cenário, esta tese propõe uma abordagem integrada baseada na construção de uma base de dados termoquímica especializada e no desenvolvimento de um software científico capaz de predizer propriedades termoquímicas moleculares absolutas, especificamente energia livre de Gibbs (G) e entalpia (H), associadas a moléculas de interesse farmacêutico. A base desenvolvida contém mais de 14.500 estruturas químicas individuais, incluindo IFAs listados na Denominação Comum Brasileira, seus produtos de degradação conhecidos e análogos estruturais.

As propriedades termoquímicas foram obtidas para cada espécie molecular isolada, não para sistemas reacionais completos. Dessa forma, o conjunto de dados e os modelos de aprendizado de máquina desenvolvidos foram treinados exclusivamente com propriedades absolutas moleculares (G e H), e não com variações energéticas de reações específicas (ΔG ou ΔH).

Para aplicação prática em estudos de degradação, o usuário deve definir explicitamente a reação química balanceada, incluindo reagentes e produtos com suas respectivas estequiometrias. O cálculo das variações termodinâmicas da reação é então realizado a posteriori, por meio da combinação estequiometricamente ponderada das propriedades moleculares estimadas para cada espécie envolvida.

Essa estratégia assegura coerência termodinâmica formal, transparência metodológica e maior capacidade de generalização do sistema, uma vez que o modelo não está restrito a um conjunto fixo de reações previamente definidas, mas fundamentado em propriedades moleculares fundamentais. Dessa maneira, torna-se possível avaliar múltiplos cenários degradativos, inclusive aqueles não explicitamente contemplados durante a etapa de treinamento do modelo.

A escolha por modelar propriedades moleculares absolutas, em vez de variações energéticas reacionais diretamente, amplia a robustez da abordagem e evita dependência de um espaço reacional previamente parametrizado, permitindo aplicação flexível a novos sistemas químicos de interesse regulatório.

As propriedades termoquímicas utilizadas para treino, foram obtidas por meio de cál-

culos sistemáticos no nível M06-2X/6-31G(d), incluindo otimizações geométricas e análises vibracionais completas no Gaussian 16 [23, 25]. A padronização metodológica assegura homogeneidade do conjunto de dados e sua adequação para posterior modelagem preditiva.

Com base nesse repositório, foi desenvolvido o software ThermoPred [29], implementado em Python e disponibilizado via interface gráfica Streamlit, integrando base especializada, modelos de aprendizado de máquina e pipelines reprodutíveis. O sistema permite estimar ΔG e ΔH associados a processos de degradação, funcionando como ferramenta preditiva complementar aos métodos experimentais tradicionais.

A proposta não substitui estudos analíticos exigidos pelas diretrizes vigentes, mas acrescenta uma camada mecanística e preditiva que pode auxiliar na priorização de riscos, na avaliação preliminar de rotas degradativas e na fundamentação científica de decisões em estudos de estabilidade. Ao alinhar modelagem computacional com requisitos regulatórios nacionais, especialmente RDC nº 53/2015, RDC nº 964/2025 e RDC nº 677/2022 [10, 12], esta pesquisa contribui para aproximar ciência computacional avançada e prática regulatória, fortalecendo abordagens baseadas em conhecimento.

Assim, esta tese posiciona-se na interface entre termoquímica computacional, inteligência artificial e ciência regulatória aplicada à estabilidade farmacêutica, propondo infraestrutura de dados inédita, metodologia padronizada e ferramenta computacional aberta voltada especificamente ao contexto de IFAs e seus produtos de degradação, em consonância com as demandas científicas e regulatórias contemporâneas.

2 TEORIA E DETALHES COMPUTACIONAIS

Nos últimos anos, a IA passou a ocupar um espaço central na ciência da computação devido à sua capacidade de resolver problemas que são difíceis ou inviáveis por métodos tradicionais. Entre suas diversas aplicações, destacam-se os modelos de aprendizado de máquina, que permitem identificar padrões, fazer previsões e analisar grandes volumes de dados de maneira eficiente. Essas ferramentas têm contribuído de forma decisiva para avanços em áreas como química, saúde, engenharia e desenvolvimento farmacêutico.

Antes de apresentar e discutir os resultados deste trabalho, é importante revisar os principais conceitos teóricos envolvidos na metodologia adotada. Essa parte introdutória explica, de forma clara e gradual, os fundamentos dos cálculos termoquímicos realizados com o software Gaussian, o papel da função erro na avaliação dos modelos e as características dos três algoritmos de inteligência artificial utilizados: *Random Forest* (RF), *Multilayer Perceptron* (MLP) e *XGBoost*. Esses elementos formam a base necessária para compreender o processo de construção, treinamento e análise dos modelos apresentados ao longo desta tese.

2.1 Propriedades Termoquímicas no *Gaussian*

O objetivo desta seção é apresentar de forma clara como o software *Gaussian* calcula as propriedades termoquímicas utilizadas neste trabalho, em especial a entalpia e a energia livre de Gibbs associadas a uma reação química. Embora o texto descreva as equações e procedimentos adotados pelo *Gaussian*, a explicação não se aprofunda nos fundamentos matemáticos dessas expressões. Dessa forma, pressupõe-se que o leitor possua uma compreensão básica dos conceitos de mecânica estatística, como funções de partição, estados acessíveis e contribuições vibracionais, rotacionais e translacionais para a energia das moléculas. A intenção aqui é fornecer uma visão geral suficiente para contextualizar os valores reportados, permitindo que o leitor compreenda de onde surgem as quantidades termodinâmicas utilizadas nas análises subsequentes.

Em cada uma das próximas quatro subseções, serão apresentadas e discutidas as equações utilizadas para calcular as contribuições para a entropia, a energia e a capacidade térmica associadas aos movimentos translacional, eletrônico, rotacional e vibracional das moléculas. Em todos os casos, o ponto de partida é a função de partição específica de cada modo, denotada por $q(V, T)$, que representa o componente correspondente na função de partição total do sistema.

Nesta seção, será apresentada uma visão geral de como essas propriedades termodinâmicas — entropia (S), energia interna (U) e capacidade térmica (C_V) — podem ser derivadas a partir da função de partição.

A função de partição de qualquer componente pode ser utilizada para determinar sua contribuição à entropia S, conforme a relação apresentada em McQuarrie (§7-6, Eq. 7.27). A expressão geral é dada por:

$$S = Nk_B + Nk_B \ln\left(\frac{q(V,T)}{N}\right) + Nk_BT\left(\frac{\partial \ln q}{\partial T}\right)_V. \tag{2.1.1}$$

A forma utilizada pelo *Gaussian* é um caso particular dessa equação. Como o programa fornece valores molares, podemos dividir por $n = N/N_A$ e substituir N_Ak_B pela constante dos gases R. Além disso, o primeiro termo pode ser incorporado ao logaritmo, resultando em:

$$S = R + R\ln(q(V,T)) + RT\left(\frac{\partial \ln q}{\partial T}\right)_V. \tag{2.1.2}$$

Reescrevendo o termo constante dentro do logaritmo, obtemos:

$$S = R\ln(q(V,T)\,e) + RT\left(\frac{\partial \ln q}{\partial T}\right)_V, \tag{2.1.3}$$

e, finalmente, considerando que a função de partição total é o produto das parcelas translacional, eletrônica, rotacional e vibracional ($q = q_tq_eq_rq_v$), chegamos a:

$$S = R\left[\ln(q_tq_eq_rq_v\,e) + T\left(\frac{\partial \ln q}{\partial T}\right)_V\right]. \tag{2.1.4}$$

A energia térmica interna E também pode ser obtida a partir da função de partição, conforme McQuarrie (§3-8, Eq. 3.41):

$$E = Nk_BT^2\left(\frac{\partial \ln q}{\partial T}\right)_V, \tag{2.1.5}$$

e, por fim, a energia interna pode ser usada para determinar a capacidade térmica a volume constante, dada por (McQuarrie, §3-4, Eq. 3.25):

$$C_V = \left(\frac{\partial E}{\partial T}\right)_{N,V}. \tag{2.1.6}$$

As três equações apresentadas (2.1.4, 2.1.5 e 2.1.6) constituem a base para derivar as expressões finais empregadas pelo *Gaussian* no cálculo dos diferentes componentes das grandezas termodinâmicas impressas no arquivo de saída.

2.1.1 Contribuições do Movimento Translacional

A função de partição translacional para um gás ideal é dada em McQuarrie (§4-1, Eq. 4.6) como:

$$q_t = \left(\frac{2\pi mk_BT}{h^2}\right)^{3/2} V. \tag{2.1.1.1}$$

A derivada parcial de $\ln q_t$ em relação à temperatura é:

$$\left(\frac{d \ln q_t}{dT}\right)_V = \frac{3}{2T}, \tag{2.1.1.2}$$

e será usada tanto no cálculo da energia interna translacional E_t quanto no terceiro termo da Eq. 2.1.4.

O segundo termo da Eq. 2.1.4 envolve o volume V, que não é conhecido diretamente. Entretanto, para um gás ideal:

$$PV = NRT = (n/N_A)N_A k_B T,$$

onde:

- N_A é o número de Avogadro;
- k_B é a constante de Boltzmann;
- T é a temperatura absoluta.

de onde se obtém:

$$V = \frac{k_B T}{P}. \tag{2.1}$$

Substituindo esta expressão na função de partição translacional:

$$q_t = \left(\frac{2\pi m k_B T}{h^2}\right)^{3/2} \left(\frac{k_B T}{P}\right), \tag{2.1.1.3}$$

que é exatamente a forma utilizada pelo *Gaussian*. Note que essa substituição não é necessária para obter a derivada da Eq. 2.1.1.2, pois esta foi calculada mantendo V constante.

A entropia translacional, incluindo o termo e proveniente da aproximação de Stirling, é dada por:

$$S_t = R\left[\ln(q_t e) + T\left(\frac{3}{2T}\right)\right], \tag{2.1.1.4}$$

que pode ser simplificada para:

$$S_t = R\left(\ln q_t + 1 + \frac{3}{2}\right). \tag{2.1.1.5}$$

A contribuição translacional para a energia térmica interna é:

$$E_t = N_A k_B T^2 \left(\frac{d \ln q_t}{dT}\right)_V, \tag{2.1.1.6}$$

que, substituindo a Eq. 2.1.1.2, resulta em:

$$E_t = RT^2\left(\frac{3}{2T}\right), \tag{2.1.1.7}$$

levando finalmente a:

$$E_t = \frac{3}{2}RT. \tag{2.1.1.8}$$

2.1.2 Contribuições do Movimento Eletrônico

A função de partição eletrônica usual é [22]:

$$q_{\mathrm{e}} = \sum_{n} \omega_n e^{-\epsilon_n / k_{\mathrm{B}} T}, \tag{2.1.2.1}$$

onde ω_n é a degenerescência do nível de energia, ϵ_n é a energia do n-ésimo nível. O *Gaussian* assume que a primeira energia de excitação eletrônica é muito maior que $k_{\mathrm{B}}T$. Portanto, o primeiro estado excitado e os mais elevados são considerados inacessíveis a qualquer temperatura. Além disso, a energia do estado fundamental é definida como zero. Estas suposições simplificam a função de partição eletrônica para:

$$q_{\mathrm{elec}} = \omega_0, \tag{2.1.2.2}$$

que é simplesmente a multiplicidade de spin eletrônico da molécula. A entropia devido ao movimento eletrônico é:

$$S_{\mathrm{elec}} = k_{\mathrm{B}} \ln q_{\mathrm{elec}}, \tag{2.1.2.3}$$

$$S_{\mathrm{elec}} = k_{\mathrm{B}} \ln g_0. \tag{2.1.2.4}$$

Como não há termos dependentes da temperatura na função de partição, a capacidade térmica eletrônica e a energia térmica interna devido ao movimento eletrônico são ambas zero.

2.1.3 Contribuições do movimento rotacional

A discussão sobre rotação molecular pode ser dividida em vários casos: átomos únicos, moléculas poliatômicas lineares e moléculas poliatômicas não lineares gerais. Vamos detalhar cada um em ordem. Para um único átomo, $q_r = 1$. Como q_r não depende da temperatura, a contribuição da rotação para a energia térmica interna, sua contribuição para a capacidade térmica e sua contribuição para a entropia são todas idênticas a zero.

Para uma molécula linear, a função de partição rotacional é [22]:

$$q_r = \frac{T}{\sigma_r \Theta_r}, \tag{2.1.3.1}$$

onde $\Theta_r = \dfrac{h^2}{8\pi^2 I k_{\mathrm{B}}}$. Aqui, I é o momento de inércia. A contribuição rotacional para a entropia é

$$S_r = R\left(lnq_r + T\left(\frac{dlnq_r}{dT}\right)_V\right), \tag{2.1.3.2}$$

$$S_r = R(lnq_r + 1). \tag{2.1.3.3}$$

A contribuição da rotação para a energia térmica interna é

$$E_r = RT^2 \left(\frac{dlnq_r}{dT} \right)_V , \tag{2.1.3.4}$$

$$E_r = RT^2 \left(\frac{1}{T} \right), \tag{2.1.3.5}$$

$$E_r = RT. \tag{2.1.3.6}$$

e a contribuição para a capacidade térmica é

$$C_r = \left(\frac{\partial E_r}{\partial T} \right)_V \tag{2.1.3.7}$$

$$C_r = R \tag{2.1.3.8}$$

Para o caso geral de uma molécula poliatômica não linear, a função de partição rotacional é [22]:

$$qr = \frac{\pi^{\frac{1}{2}}}{\sigma_r} \left(\frac{T^{\frac{3}{2}}}{(\theta_{r,x}\theta_{r,y}\theta_{r,z})^{\frac{1}{2}}} \right). \tag{2.1.3.9}$$

Agora temos que, $\left(\frac{dlnq}{dT}\right)_V = \frac{3}{2T}$, então a entropia para esta função de partição é

$$S_r = R \left(\ln q_r + T \left(\frac{d \ln q_r}{dT} \right)_V \right), \tag{2.1.3.10}$$

$$S_r = R \left(\ln q_r + \frac{3}{2} \right). \tag{2.1.3.11}$$

Finalmente, a contribuição para a energia térmica interna é

$$E_r = RT^2 \left(\frac{d \ln q_r}{dT} \right)_V , \tag{2.1.3.12}$$

$$E_r = RT^2 \left(\frac{3}{2T} \right), \tag{2.1.3.13}$$

$$E_r = \frac{3}{2} RT, \tag{2.1.3.14}$$

e a contribuição para a capacidade térmica é

$$C_r = \left(\frac{dE_r}{dT} \right)_V , \tag{2.1.3.15}$$

$$C_r = \frac{3}{2} R. \tag{2.1.3.16}$$

A contribuição média para a energia térmica interna de cada grau de liberdade rotacional é $\frac{3}{2}R$, enquanto sua contribuição para C_r é $\frac{R}{2}$.

2.1.4 Contribuições do movimento vibracional

As contribuições para a função de partição, entropia, energia interna e capacidade calorífica a volume constante dos movimentos vibracionais são compostas por uma soma (ou produto) das contribuições de cada modo vibracional, K. Cada um dos modos $3n_atoms{-}6$ (ou $3n_atoms{-}5$ para moléculas lineares) tem uma temperatura vibracional característica, $\theta_{v,K} = hv_K/k_B$. Existem duas maneiras de calcular a função de partição, dependendo da escolha do zero de energia: ou na parte inferior do poço de energia potencial internuclear, ou no primeiro nível vibracional. A escolha depende se as contribuições decorrentes da energia do ponto zero serão ou não calculadas separadamente. Se eles forem calculados separadamente, então o fundo do poço deve ser usado como ponto de referência, caso contrário, o primeiro nível de energia vibracional é a escolha apropriada. Se o fundo do poço (BOT) for escolhido como ponto de referência zero, então a contribuição para a função de partição de um determinado modo vibracional é [22]:

$$q_{v,K} = \left(\frac{e^{\frac{-\theta_{v,K}}{2T}}}{1 - e^{\frac{-\theta_{v,K}}{T}}} \right), \tag{2.1.4.1}$$

e a função de partição vibracional geral é [22]:

$$q_{v,K} = \left(\frac{e^{\frac{-\theta_{v,k}}{2T}}}{1 - e^{\frac{-\theta_{v,k}}{T}}} \right). \tag{2.1.4.2}$$

Por outro lado, se o primeiro nível de energia vibracional for escolhido como sendo o zero de energia (V $= 0$), então a função de partição para cada nível vibracional é

$$q_{v,K} = \frac{1}{1 - e^{\frac{\theta_{v,N}}{T}}}, \tag{2.1.4.3}$$

e a função de partição vibracional geral é

$$q_v = \Pi_K \frac{1}{1 - e^{\frac{-\theta_{v,K}}{T}}}. \tag{2.1.4.4}$$

O Gaussian usa o fundo do poço como o zero de energia (BOT) para determinar as outras grandezas termodinâmicas, mas também imprime a função de partição V $= 0$. Em última análise, a única diferença entre as duas referências é o fator adicional de $\frac{\theta_{v,K}}{2}$, (que é a energia vibracional do ponto zero) na equação para a energia interna E_v. Nas expressões para capacidade calorífica e entropia, esse fator desaparece, pois diferencia em relação à temperatura (T).

A contribuição total de entropia da função de partição vibracional é:

$$S_V = R \left(\ln (q_V) + T \left(\frac{d \ln q}{dT} \right)_V \right), \tag{2.1.4.5}$$

$$S_V = R \left(\ln (q_V) + T \left(\sum_K \frac{\theta_{v,k}}{2T^2} + \sum_K \frac{\left(\frac{\theta_{v,K}}{T^2} \right) e^{\frac{-\theta_{v,K}}{T}}}{1 - e^{\frac{-\theta_{v,K}}{T}}} \right) \right), \tag{2.1.4.6}$$

$$S_V = R(\sum_K \left(\frac{\theta_{v,k}}{2T} + \ln\left(1 - e^{\frac{-\theta_{v,K}}{T}}\right)\right) + T\left(\sum_K \frac{\theta_{v,k}}{2T^2} + \sum_K \frac{\left(\frac{\theta_{v,K}}{T^2}\right) e^{\frac{-\theta_{v,K}}{T}}}{1 - e^{\frac{-\theta_{v,K}}{T}}}\right), \quad (2.1.4.7)$$

$$S_V = R(\sum_K \ln\left(1 - e^{\frac{-\theta_{v,K}}{T}}\right) + \left(\sum_K \frac{\left(\frac{\theta_{v,K}}{T}\right) e^{\frac{-\theta_{v,K}}{T}}}{1 - e^{\frac{-\theta_{v,K}}{T}}}\right), \quad (2.1.4.8)$$

$$S_V = R\sum_K \left(\frac{\frac{\theta_{v,k}}{T}}{e^{\frac{\theta_{v,K}}{T}} - 1} - \ln\left(1 - e^{\frac{-\theta_{v,K}}{T}}\right)\right). \quad (2.1.4.9)$$

Para ir da quarta linha (2.1.4.8) à quinta linha (2.1.4.9) nas equações acima, você deve multiplicar por $\frac{e^{\frac{\theta_{v,k}}{T}}}{e^{\frac{\theta_{v,k}}{T}}}$.

A contribuição para a energia térmica interna resultante da vibração molecular é

$$E_V = R\sum_K \theta_{v,K}\left(\frac{1}{2} + \frac{1}{\frac{e^{\theta_{v,K}}}{T} - 1}\right). \quad (2.1.4.10)$$

Finalmente, a contribuição para a capacidade térmica de volume constante é

$$C_V = R\sum_K e^{\frac{\theta_{v,K}}{T}}\left(\frac{\frac{\theta_{v,K}}{T}}{e^{\frac{-\theta_{v,K}}{T}-1}}\right)^2. \quad (2.1.4.11)$$

Os modos de baixa frequência (definidos abaixo) estão inclusos nos cálculos descritos acima. Alguns destes modos podem ser rotações internas e, portanto, podem precisar ser tratados separadamente, dependendo das temperaturas e das barreiras envolvidas. Para facilitar a correção desses modos, suas contribuições são impressas separadamente, para que possam ser subtraídas. Um modo de baixa frequência no Gaussian é definido como aquele para o qual é provável que mais de cinco por cento de um conjunto de moléculas exista em estados vibracionais excitados à temperatura ambiente. Em outras unidades, isso corresponde a cerca de $625cm^{-1}$, $1,9{\times}1013Hz$, ou uma temperatura vibracional de $900K$.

2.2 Erro Quadrático Médio (MSE)

Uma função de perda no aprendizado de máquina é uma medida da precisão com que um modelo de ML é capaz de prever o resultado esperado, ou seja, a verdade básica. A função de perda terá dois itens como entrada: o valor de saída do nosso modelo e o valor esperado da verdade básica. A saída da função de perda é chamada de perda, que é uma medida de quão bem o modelo se saiu na previsão do resultado. Um valor alto para a perda significa que o modelo teve um desempenho muito ruim. Um valor baixo para a perda significa que o modelo teve um desempenho muito bom.

A seleção da função de perda adequada é crítica para treinar um modelo preciso. Certas funções de perda terão certas propriedades e ajudarão o modelo a aprender de uma maneira específica. Alguns podem dar mais peso aos valores discrepantes, outros à maioria.

O erro quadrático médio (em inglês *Mean Squared Error* - MSE) é talvez a função de perda mais simples e comum. Para calcular o MSE, utiliza-se a diferença entre as previsões do modelo e a verdade básica, eleva-se ao quadrado e calcula-se a média de todo o conjunto de dados. O MSE é definido por:

$$E = \frac{1}{N}\sum_{i=1}^{n}\left(\hat{Y}_i - y_i\right)^2, \tag{2.2.1}$$

sendo N o número de amostras que estamos testando. Plotando a função temos a Figura **??**.

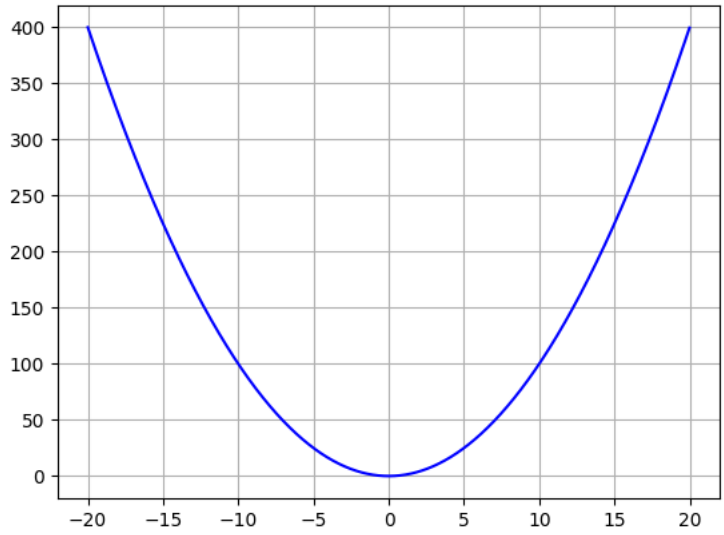

Figura 2.1. Função Erro Quadrático Médio

O MSE garante que o modelo treinado não tenha previsões atípicas com erros enormes, já que o MSE dá maior peso a esses erros devido à quadratura da função.

2.3 XGBoost

Primeiramente, para entender um modelo de *machine learning* precisamos conhecer alguns conceitos básicos [30]. O primeiro deles é o ensemble, uma técnica que combina resultados de modelos distintos para gerar uma predição. A Figura 2.2 demonstra um exemplo de ensemble constituído por duas árvores de decisão.

Para realizar a predição de uma amostra x_i, percorremos cada árvore seguindo um caminho constituído pelas regras de decisão, destacado na imagem em vermelho, até atingir um ponto final. Podemos denominar o caminho de $q\left(x_i\right)$ e o total de folhas que constituem uma

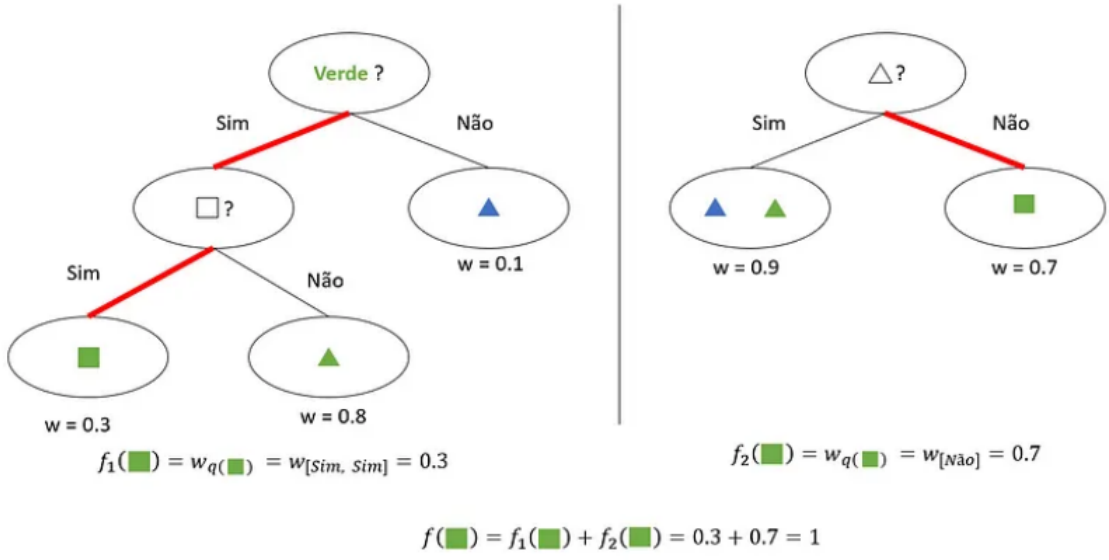

Figura 2.2. Exemplo de um ensemble de duas árvores de decisão

árvore de T. Para cada folha j atribuímos um peso, w_j. O valor final da predição, por sua vez, vai ser igual à soma de todos os valores $f_t\left(w_j\right)$, ou seja, a soma dos pesos das folhas atingidas.

De forma específica, o Boosting é um método de ensemble em que um conjunto t de modelos é treinado sequencialmente, sendo que o modelo t tem como objetivo corrigir os erros do modelo $t-1$.

Considerando que, y_i seja o valor alvo e $\hat{y}_i^t$ a predição do t-ésimo modelo para a amostra x_i, l uma função de erro qualquer (MSE - Mean Squared Error, por exemplo) e n o número total de amostras, o erro do modelo na iteração t é definido como:

$$L_t = \sum_{i=1}^{n} l\left(y_i, \hat{\mathrm{y}}_i^t\right). \tag{2.3.1}$$

Sabemos que o modelo foi construído de forma iterativa, portanto a predição na iteração t é equivalente à predição na iteração $t-1$ somada à predição de um novo modelo, f_t :

$$\hat{\mathrm{y}}_i^t = \hat{\mathrm{y}}_i^{t-1} + f_t\left(x_i\right), \tag{2.3.2}$$

$$L_t = \sum_{i=1}^{n} l\left(y_i, \hat{\mathrm{y}}_i^t + f_t\left(x_i\right)\right), \tag{2.3.3}$$

Adicionando um termo de regularização, ajudando a controlar a complexidade do modelo, temos que:

$$L_t = \sum_{i=1}^{n} l\left(y_i, \hat{\mathrm{y}}_i^t + f_t\left(x_i\right)\right) + \Omega\left(f_t\right) \tag{2.3.4}$$

O objetivo é que a adição de cada árvore seja certeira, obtendo sempre a melhor árvore que irá minimizar o erro. Para isso, consideraremos a função L como um problema de otimiza-

ção, ou seja, queremos encontrar o f_t que minimiza o L. Dependendo da escolha da função de erro, l, essa tarefa pode ser complicada.

Portanto, iremos transformá-la em uma função mais simples utilizando uma série de Taylor. Sabemos que toda função infinitamente diferenciável pode ser escrita da forma:

$$f(a+h) = f(a) + f'(a)h + \frac{1}{2}f''(a)h^2 + \cdots + \frac{1}{n!}\frac{1}{2}f^n(a)h^n \tag{2.3.5}$$

Se interrompermos a série, teremos uma aproximação da função. Neste caso, vamos parar na ordem 2.

$$f(a+h) \simeq f(a) + f'(a)h + \frac{1}{2}f''(a)h^2, \tag{2.3.6}$$

$$a = \hat{\mathrm{y}}_i^{t-1}, \tag{2.3.7}$$

$$h = f_t(x), \tag{2.3.8}$$

$$L_t = \sum_{i=1}^{n}\left[l\left(y_i, \hat{\mathrm{y}}_i^{t-1}\right) + \frac{dl\left(y_i \hat{\mathrm{y}}_i^{t-1}\right)}{d\hat{\mathrm{y}}_i^{t-1}} f_t\left(x_i\right) + \frac{1}{2}\frac{d^2 l\left(y_i, \hat{\mathrm{y}}_i^{t-1}\right)}{d\hat{\mathrm{y}}_i^{t-1\,2}} f_t\left(x_i\right)^2\right] + \Omega\left(f_t\right). \tag{2.3.9}$$

Substituindo as derivadas por g_i (gradiente) e h_i (hessiano), temos que:

$$L_t = \sum_{i=1}^{n}\left[l\left(y_i, \hat{\mathrm{y}}_i^{t-1}\right) + g_i f_t\left(x_i\right) + \frac{1}{2}h_i f_t\left(x_i\right)^2\right] + \Omega\left(f_t\right) \tag{2.3.10}$$

Como o objetivo é encontrar o f_t que minimiza a equação, não precisamos do termo l, visto que ele é constante. Portanto temos que:

$$L_t = \sum_{i=1}^{n}\left[g_i f_t\left(x_i\right) + \frac{1}{2}h_i f_t\left(x_i\right)^2\right] + \Omega\left(f_t\right) \tag{2.3.11}$$

A função erro (loss) deve ser, obrigatoriamente, duplamente diferenciável, passando como parâmetros como calcular o gradiente (derivada de primeira ordem) e o hessiano (derivada de segunda ordem).

Levando em consideração como funcionam as árvores de decisão, vamos reescrever a nossa equação L.

Sabemos que cada amostra x_i vai estar associada a uma folha j. Então, para cada folha, podemos criar um conjunto de índices I_j, em que cada elemento se refere a uma das amostras contidas na folha.

I_j é o conjunto tal que, para todo índice i em I_j, o caminho de decisão q percorrido pela amostra x_i leva à folha j . A resposta do modelo para a amostra x_i será o peso da folha em que x_i está contida. Logo:

$$f_t\left(x_i\right) = \omega_j. \tag{2.3.12}$$

Portanto, reescrevendo alguns termos da equação temos que:

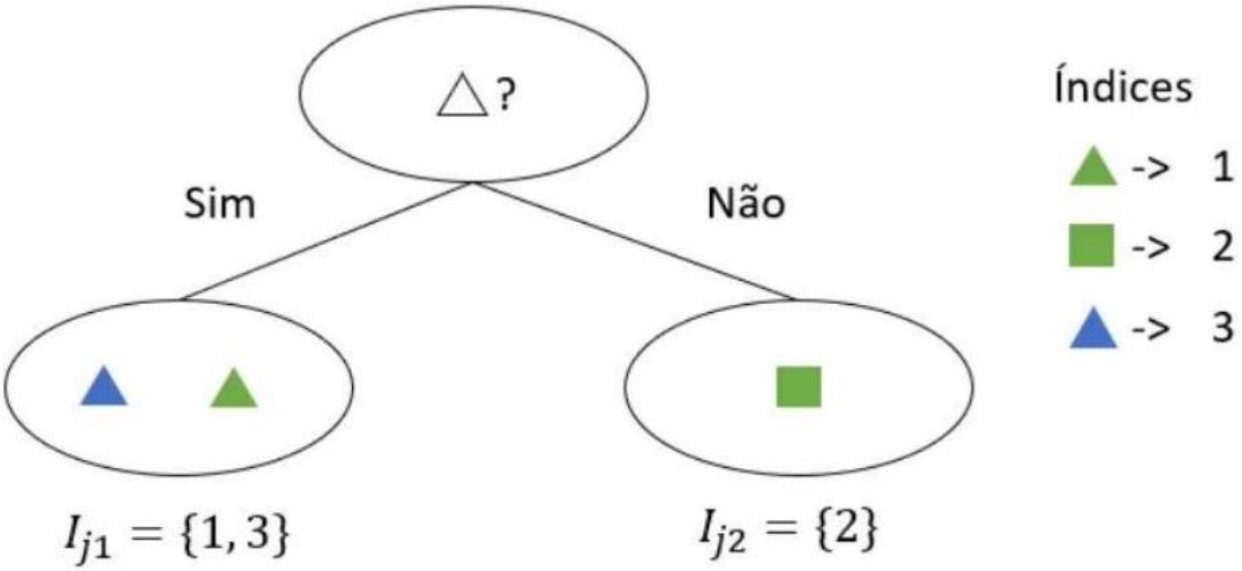

Figura 2.3. Exemplos dos conjuntos índice para cada folha de uma arvore de decisão

$$\sum_{i=1}^{n} g_i f_t(x_i) = \sum_{j=1}^{T} \omega_j \sum_{i \in I_j} gi, \tag{2.3.13}$$

$$\sum_{i=1}^{n} h_i f_t(x_i)^2 = \sum_{j=1}^{T} \omega_j^2 \sum_{i \in I_j} hi. \tag{2.3.14}$$

Substituindo, temos:

$$L_t = \sum_{j=1}^{T} \omega_j \sum_{i \in I_j} g_i + \frac{1}{2} \sum_{j=1}^{T} \omega_j^2 \sum_{i \in I_j} hi + \Omega(f_t). \tag{2.3.15}$$

Expandindo o termo de regularização:

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2, \tag{2.3.16}$$

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} \omega_j^2, \tag{2.3.17}$$

$$L_t = \sum_{j=1}^{T} \omega_j \sum_{i \in I_j} g_i + \frac{1}{2} \sum_{j=1}^{T} \omega_j^2 \sum_{i \in I_j} hi + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} \omega_j^2, \tag{2.3.18}$$

$$L_t = \sum_{j=1}^{T} \left[\sum_{i \in I_j} (g_i) \omega_j + \frac{1}{2} \sum_{i \in I_j} (hi) \omega_j^2 + \frac{\lambda \omega^2}{2} \right] + \gamma T, \tag{2.3.19}$$

$$L_t = \sum_{j=1}^{T} \left[\sum_{i \in I_j} (g_i) \omega_j + \frac{1}{2} \sum_{i \in I_j} (hi + \lambda) \omega_j^2 \right] + \gamma T. \tag{2.3.20}$$

Para otimização do erro, ao invés de se olhar para todas as folhas da árvore, fixa-se uma folha j.

$$L_{t_j} = \sum_{i \in I_j} (g_i)\, \omega_j + \frac{1}{2} \sum_{i \in I_j} (hi + \lambda)\omega_j^2 + \gamma T \tag{2.3.21}$$

O objetivo é encontrar o conjunto de pesos ω que minimiza L. A nossa função erro para uma folha é quadrática, e, portanto, o mínimo é definido como o ponto de inflexão da curva, Figura 2.4, em que a primeira derivada é igual a zero.

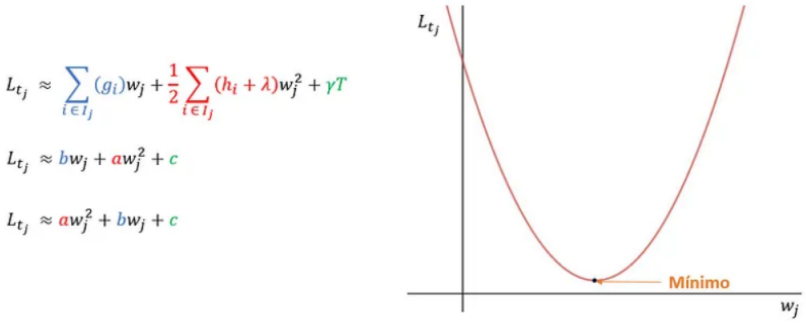

Figura 2.4. Função Erro

$$\frac{dL_{t_j}}{d\omega_j} = \sum_{i \in I_j} g_i + \frac{1}{2} 2 \sum_{i \in I_j} (hi + \lambda)\omega_j = 0. \tag{2.3.22}$$

Isolando ω, temos que:

$$\omega_j = \frac{-\sum_{i \in I_j} g_i}{\sum_{i \in I_j} (hi + \lambda)} \tag{2.3.23}$$

Agora sabe-se qual é a equação que nos dará o peso ótimo de uma folha arbitrária. Logo, se substituirmos na nossa equação L, chegamos em:

$$L_t = -\frac{1}{2} \sum_{j=1}^{T} \frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} (hi + \lambda)} + \gamma T. \tag{2.3.24}$$

Na prática, essa equação será usada para avaliar cada split de uma nova árvore, assim como a entropia ou o coeficiente de Ginni são utilizados tradicionalmente na criação de árvore de decisão.

A cada split são gerados dois nós, o esquerdo (*left*) e o direito (*right*). O ganho do split é definido como a soma do L_1 (*left*) e do L_r (*right*) que correspondem às novas folhas, subtraídos do erro anterior, L_t (lembrando que $T = 1$, já que está sendo olhado apenas para uma folha).

$$\text{Ganho } = L_l + L_r - L_t, \tag{2.3.25}$$

$$\text{Ganho } = -\frac{1}{2}\frac{\left(\sum_{i\in I_{esq}} g_i\right)^2}{\sum_{i\in I_{esq}}(hi+\lambda)} + \gamma - \frac{1}{2}\frac{\left(\sum_{i\in I_{dir}} g_i\right)^2}{\sum_{i\in I_{dir}}(hi+\lambda)} + \gamma + \frac{1}{2}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j}(hi+\lambda)} - \gamma, \quad (2.3.26)$$

$$\text{Ganho } = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_{\text{esq}}} g_i\right)^2}{\sum_{i\in I_{\text{esq}}}(hi+\lambda)} + \frac{\left(\sum_{i\in I_{\text{dir}}} g_i\right)^2}{\sum_{i\in I_{\text{dir}}}(hi+\lambda)} + \frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j}(hi+\lambda)}\right] - \gamma \quad (2.3.27)$$

A partir deste ponto, calcula-se todos os splits possíveis e o que gera maior ganho é escolhido. Das variáveis descritas, temos que, o lambda afeta o peso das folhas. Quanto maior o lambda, menor o valor absoluto do peso. Por essa razão, λ é um parâmetro que controla a complexidade do modelo, visto que evita pesos muito grandes. Mais especificamente, trata-se de regularização L2. O valor de gamma diminui o ganho. Por essa razão, γ é descrito como o valor mínimo para que um split aconteça, visto que um valor menor que γ resultaria em um ganho negativo, o qual nunca é considerado (pois o resultado, na prática, estaria piorando). O valor da soma dos h s de cada nó filho deve ser maior que o valor estabelecido por esse parâmetro para que um split seja realizado. Lembrando que h é dado pela derivada da função de erro (l). Portanto, se o valor de h é baixo, significa que a folha já está ”pura”o suficiente e não deve ser mais particionada.

Neste segmento foi destacado apenas os parâmetros relacionados ao processo de boosting, embora ainda existam outros parâmetros relativos à construção de árvores de decisão (como altura máxima, por exemplo).

2.4 Random Forest

Nesse tópico discutiremos a fim de fornecer uma compreensão holística e intuitiva do algoritmo de aprendizado de máquina supervisionado ”*Random Forests*”[31, 32].

Para começar, precisamos entender o conceito de *Bagging*. É um procedimento geral que pode ser usado para reduzir a variância do nosso modelo. Uma variação maior significa que seu modelo está super ajustado. Certos algoritmos, como arvores de decisão sofrem com alta variância. Por outro lado, as arvores de decisão são sensíveis aos dados nos quais foram treinadas. Se os dados subjacentes forem alterados um pouco, a árvore de decisão resultante poderá ser muito diferente e, como resultado, as previsões do nosso modelo mudarão drasticamente. O *Bagging* oferece uma solução para o problema de alta variância. Pode reduzir sistematicamente o overfitting tomando uma média de várias árvores de decisão. *Bagging* usa amostragem *bootstrap* (amostrar linhas aleatoriamente do conjunto de dados de treinamento com substituição) e finalmente agrega os modelos individuais calculando a média para obter as previsões finais.

Com o *bagging*, é possível desenhar um único exemplo de treinamento mais de uma vez. Isso resulta em uma versão modificada do conjunto de treinamento onde algumas linhas

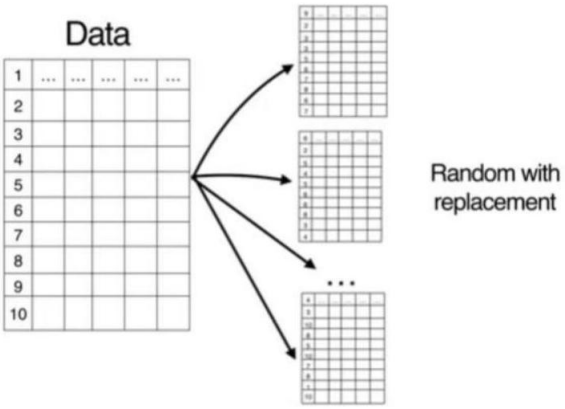

Figura 2.5. Amostragem de inicialização

são representadas diversas vezes e outras estão ausentes. Isso também permite criar dados, que são semelhantes aos dados com os quais você começou. Ao fazer isso, você pode ajustar muitos modelos diferentes, mas semelhantes.

No *bagging* você extrai B amostras com substituição do conjunto de dados original, onde B é um número menor ou igual a n, o número total de amostras no conjunto de treinamento, conforme Figura 2.6.

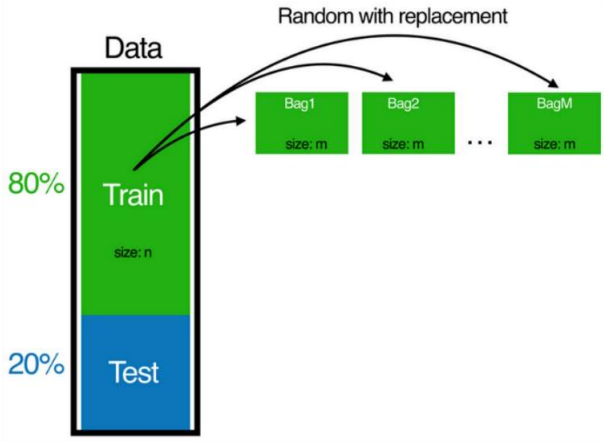

Figura 2.6. Conjunto de dados para treinamento

Na etapa seguinte, temos que treinar árvores de decisão em amostras *bootstrap* recém criadas. Repita a Etapa 1 e a Etapa 2 quantas vezes desejar. Geralmente quanto maior o número de árvores, melhor é o modelo. O número excessivo de árvores pode tornar um modelo complicado e, em última análise, levar a um ajuste excessivo, à medida que seu modelo começa a ver relacionamentos nos dados que não existem em primeiro lugar.

Para gerar uma previsão usando a abordagem de árvores ensacadas (*bagged trees*), você deve gerar uma previsão de cada uma das árvores de decisão e, em seguida, simplesmente calcular a média das previsões para obter uma previsão final. A previsão ensacada ou de conjunto é a previsão média nas árvores inicializadas amostradas. O modelo de árvores ensacadas funciona de maneira muito semelhante ao conselho. Normalmente, quando um conselho precisa tomar

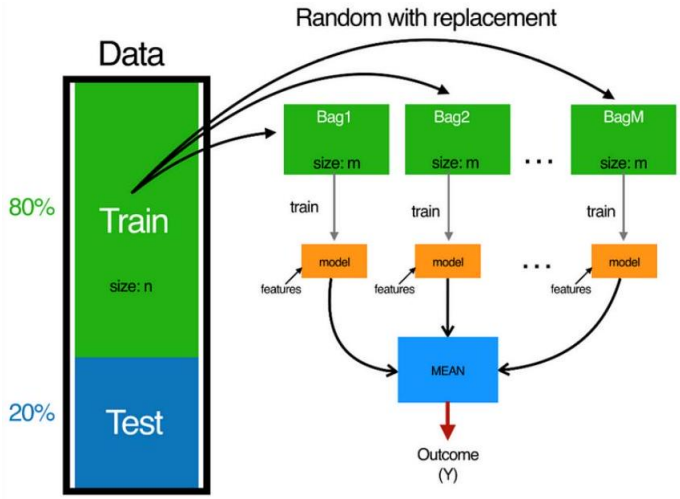

Figura 2.7. Treino em árvores de decisão

uma decisão, ele simplesmente considera a votação majoritária. A opção que obtiver mais votos (digamos, a opção A obteve 100 votos e a opção B obteve 90 votos) é a decisão final do conselho. Da mesma forma, no bagging, quando você está tentando resolver um problema de classificação, você está basicamente obtendo uma votação majoritária em todas as suas árvores de decisão. E, no caso de regressão, simplesmente calculamos uma média de todas as previsões da árvore de decisão. O conhecimento coletivo de um conjunto diversificado de árvores de decisão normalmente supera o conhecimento de qualquer árvore individual. As árvores ensacadas oferecem, portanto, melhor desempenho preditivo.

A floresta aleatória é diferente do *bagging* em apenas um aspecto. Ele usa um algoritmo de aprendizagem em árvore modificado que inspeciona, em cada divisão do processo de aprendizagem, um subconjunto aleatório de recursos. Isso é realizado para evitar a correlação entre as árvores. Suponha que haja um preditor muito forte no conjunto de dados junto com vários outros preditores moderadamente fortes; então, na coleção de árvores ensacadas, a maioria ou todas as nossas árvores de decisão usarão o preditor muito forte para a primeira divisão, portanto, todas as árvores ensacadas serão semelhantes. Consequentemente, todas as previsões das árvores ensacadas serão altamente correlacionadas. Os preditores correlacionados não podem ajudar a melhorar a precisão da previsão. Ao usar um subconjunto aleatório de recursos, o *Random Forest* evita sistematicamente a correlação e melhora o desempenho do modelo. A Figura 2.8 ilustra como funciona o algoritmo *Random Forest*.

Em um problema de classificação, como demonstrado na imagem, os dados de treinamento possuem quatro variáveis: Recurso 1, Recurso 2, Recurso 3 e Recurso 4. Cada modelo individual no algoritmo será treinado em um subconjunto específico desses recursos. Por exemplo, a Árvore de Decisão 1 será treinada com os recursos 1 e 4, a Árvore de Decisão 2 com os recursos 2 e 4, e a Árvore de Decisão 3 com os recursos 3 e 4. Assim, três modelos distintos são gerados, cada um utilizando um subconjunto diferente de recursos. Novos dados de teste são então aplicados a cada um desses modelos, e uma previsão é gerada. A decisão final do

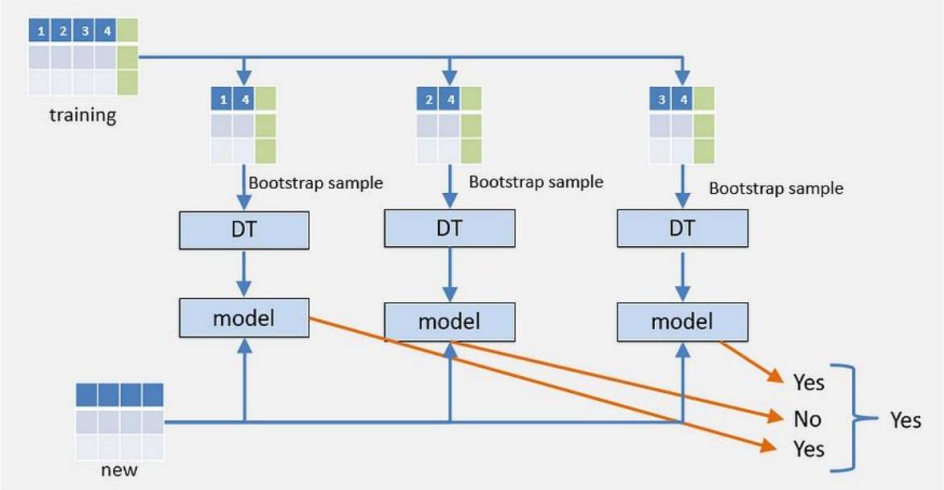

Figura 2.8. Funcionamento do algoritmo Random Forest

algoritmo *Random Forest* será baseada na previsão que obtiver o maior número de votos.

O *Random Forest* é amplamente utilizado entre os algoritmos de aprendizado de conjunto, pois ao usar múltiplas amostras do conjunto de dados original, a variância do modelo final é reduzida. A baixa variância está associada à diminuição do overfitting, um fenômeno que ocorre quando o modelo tenta ajustar variações pequenas e específicas de um conjunto de dados, que representa apenas uma pequena amostra de todos os exemplos possíveis do fenômeno a ser modelado. Ao criar múltiplas amostras aleatórias com substituição a partir do conjunto de treinamento, o efeito do *overfitting* é minimizado.

2.5 Perceptron Multicamadas - MLP

As redes neurais artificiais são baseadas na rede biológica de neurônios, que são responsáveis pelo processamento cerebral. A entrada de um neurônio é formada pelas saídas de vários outros neurônios. A comunicação é realizada por neurotransmissores, que transportam os sinais quimicamente entre os neurônios. Cada neurônio recebe essa comunicação com pesos em cada entrada vinda dos outros neurônios e o neurônio é ativado se a soma ponderada de suas entradas for maior que um limiar, que pode ser definido de diferentes maneiras [33].

O aprendizado ocorre através de modificações constantes nas sinapses que conectam os neurônios, de acordo com a liberação de neurotransmissores. Logo, de acordo com novos eventos, algumas ligações entre neurônios são fortalecidas, enquanto outras são enfraquecidas. O ajuste nas ligações dos neurônios é uma das principais características desse modelo.

Esses modelos aprendem funções matemáticas (dada uma entrada ele aprende a responder saídas específicas, até se tornar uma função) e podem ser usados tanto para regressão quanto para classificação. A disposição desse modelo é similar ao funcionamento do sistema biológico, onde os neurônios podem ser ativados por estímulos de entrada (função de ativação), porém o cérebro possui cerca de 100 bilhões de neurônios, enquanto as RNAs ficam abaixo de mil normalmente.

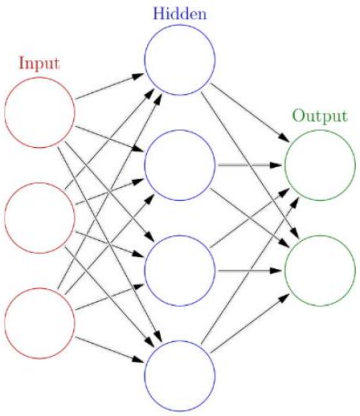

Figura 2.9. Modelo de aprendizado por redes neurais

A entrada do modelo consiste nos valores dos dados, enquanto as arestas representam os pesos associados a cada neurônio. Na camada oculta, são realizados cálculos que resultam na saída, podendo ser uma classificação ou regressão. Cada camada executa o mesmo tipo de cálculo, geralmente baseado na soma ponderada das entradas em cada unidade. O conhecimento adquirido durante o treinamento é armazenado nos pesos dos neurônios. Esse tipo de modelo é frequentemente descrito como uma ”caixa preta”, o que significa que é difícil, quando não impossível, extrair e interpretar o conhecimento adquirido durante o aprendizado.

Trata-se de um modelo de aprendizado supervisionado, no qual são fornecidos dados rotulados com classes, representados numericamente, para que o modelo possa aprender. Posteriormente, ele pode ser usado para classificar novas instâncias ou estimar um valor de saída no caso de problemas de regressão.

Antes do treinamento, existem duas etapas preliminares essenciais: a escolha da arquitetura da rede (número de neurônios e camadas ocultas) e a escolha da função de ativação dos neurônios. O perceptron foi o primeiro e mais simples modelo de rede neural artificial, proposto por Rosenblatt em 1959. Ele é composto por vários neurônios de entrada com valores ponderados e um único neurônio de saída que aplica uma função de ativação, como a função threshold ou a função logística, em sua forma mais simples.

A aprendizagem nos perceptrons consiste em dois elementos, os pesos entre as unidades de entrada e saída e o valor do threshold. O treinamento ocorre da seguinte maneira, os valores dos pesos são inicializados aleatoriamente, geralmente no intervalo de ($-1, 1$). Para cada exemplo de treinamento é calculada a saída observada da rede $o(E)$, se a saída desejada $t(E)$ for diferente da observada então os pesos da rede são ajustados para que elas se aproximem, isso é feito aplicando-se a regra de aprendizado do perceptron. O aprendizado não termina necessariamente após todos os exemplos terem sido usados, é possível repetir o ciclo novamente até que se produza as saídas corretas, chamada de convergência. A regra geral de aprendizagem acontece da seguinte maneira: quando $t(E)$ for diferente de $o(E)$, então é adicionado um Δ_i ao peso de ω_i, onde Δ_i é definido como:

$$\Delta_i = (t(E) - o(E))x_i, \tag{2.5.1}$$

onde, n é a taxa de aprendizado, ou quão rápido o aprendizado acontecerá, é importante dar passos pequenos no aprendizado para garantir que o ponto ideal não seja ignorado; $t(E) - o(E)$ é a diferença entre o valor esperado e o valor obtido, dando a direção do aumento ou diminuição dos pesos; x_i é o valor de entrada do neurônio.

Figura 2.10. À esquerda é possível resolver o problema com uma equação de primeiro grau. No caso da esquerda, é necessária uma equação muito mais complexa

O *perceptron* é um classificador linear, ou seja, é possível classificar instâncias que sejam linearmente separáveis, entretanto alguns anos depois ocorreu o desenvolvimento das MLP Multilayer *perceptron* que permitem aprender conceitos bem mais complexos.

A rede MLP (*Multi-Layer Perceptron*) pode ser vista como uma extensão do *perceptron*, composta por vários neurônios semelhantes. Diferente do *perceptron* simples, a MLP possui pelo menos uma camada oculta entre a camada de entrada e a de saída, permitindo que a rede aprenda representações mais complexas dos dados. Embora seja uma estrutura robusta, o aprendizado na MLP depende do cálculo diferencial. Isso significa que funções como a função de threshold, usada no *perceptron*, não podem ser aplicadas, pois não são diferenciáveis no ponto em questão. Por isso, é necessário utilizar uma função diferenciável, e a função sigmoide é amplamente empregada nesse contexto devido à sua suavidade e diferenciabilidade.

O Perceptron é um tipo especial de rede neural que realiza apenas classificação binária, e no MLP a partir da camada oculta podemos captar novos padrões. É possível aprender funções polinomiais, trigonométricas, parâmetros para sistemas lineares e até sistemas de equações para modelar situações complexas com classificação de imagens. Mas tudo isso tem o preço de gerar um modelo extremamente complexo e, pouco intuitivo para o entendimento de um ser humano. Porém, uma vez que este modelo resolva determinadas tarefas com uma precisão aceitável, é possível delegar determinadas tarefas a este tipo de modelo.

De acordo com o teorema de George Cybenko: Um perceptron multicamadas com função de ativação sigmoide é um aproximador universal para qualquer função de base real. Na prática, podemos usar uma combinação de vários modelos para aprender e automatizar tarefas cada vez mais complexas.

Para definir o modelo e seus parâmetros iremos utilizar a seguinte topologia conforme Figura 2.11.

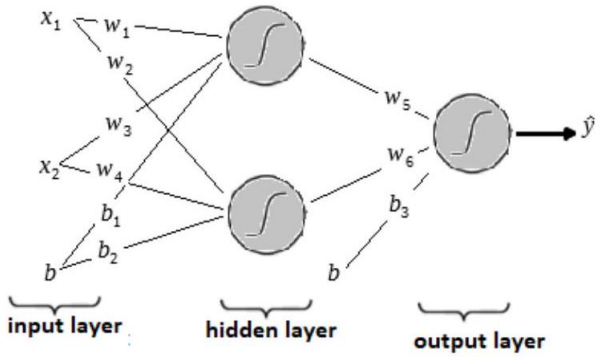

Figura 2.11. Modelo de rede neural com suas respectivas entradas e pesos

Na Figura 2.11 podemos ter as duas entradas x_1 e x_2, e temos o viés b. Uma observação sobre o viés é que na prática ele sempre levará a entrada 1 , o que importa mesmo são os pesos do viés no caso da primeira camada temos b_1 e b_2. Logo em seguida podemos observar também os pesos dos insumos para a primeira camada que vai de ω_1 a ω_4. Após multiplicarmos os pesos e adicionarmos as tendências, aplicamos a função sigmoide. O mesmo processo é feito nas saídas da próxima camada (camada oculta) até gerar as saídas da rede.

Considerando um conjunto de treinamento,

Tabela 2.1. Conjunto de treinamento para rede neural

$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{b}$	label
1	1	1	0
1	2	1	1
2	3	1	0
2	1	1	1

na coluna b temos apenas o número 1 , tornando os cálculos de viés mais intuitivos. Quando inicializamos o treinamento, os pesos são inicializados aleatoriamente. Os pesos das tendências e camadas são chamados de parâmetros de rede. Quando fazemos otimização de rede, o que fazemos é buscar os melhores parâmetros. Mas para isso precisamos dos chamados hiperparâmetros.

O primeiro chamamos de alfa ou taxa de aprendizagem, este parâmetro regula a velocidade com que a rede irá aprender. A próxima é a função de ativação, como por exemplo a sigmoide. A função sigmoide é interessante porque tende a 1 quando x se aproxima do infinito e tende a 0 quando x se aproxima do infinito negativo. Isso nos dá um discriminante 0,5 acima que posso classificar como positivo e abaixo como negativo em uma ampla gama de situações.

$$O_j = \frac{1}{1 + e^{-n_j}} \tag{2.5.2}$$

em que n refere-se ao produto das entradas e tendências pelos pesos.
Na Equação 2.5.3 conseguimos gerar o número n para cada neurônio, a partir da generalização da multiplicação vetorial:

$$n_j = \sum_{i=1}^{n} \omega_i x_i \tag{2.5.3}$$

Para calcular a saída do neurônio destacado em vermelho, utilizamos o vetor de entrada e o vetor de pesos associados a esse neurônio (representados pelas arestas vermelhas). O processo segue a fórmula de soma ponderada, onde cada valor de entrada é multiplicado pelo peso correspondente, e o resultado é somado.

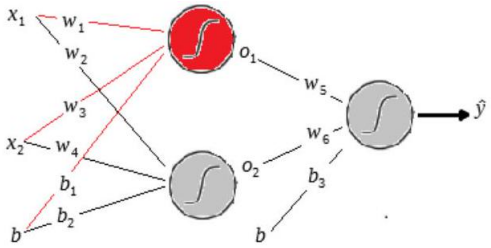

Figura 2.12. Modelo de rede neural com vetor entrada e vetor peso

A função que será otimizada também é chamada de função objetivo ou função de perda, para isso utilizaremos o erro quadrático médio representado na fórmula abaixo, onde o símbolo $\hat{\mathrm{Y}}_i$ refere-se à saída da rede e o y_i refere-se ao rótulo original. O número 2 que multiplica N é para facilitar o cálculo da derivada e N é a quantidade de amostras no conjunto de dados. O objetivo é minimizar esse erro e assim encontrar o melhor conjunto de parâmetros ω_i e b_i.

$$E = \frac{1}{2N} \sum_{i=1}^{n} \left(\hat{Y}_i - y_i \right)^2 \tag{2.5.4}$$

Agora que o erro foi identificado, é possível propagá-lo para as camadas anteriores e, em seguida, ajustar os pesos de acordo com a regra de atualização. Esse processo é conhecido como retropropagação (backpropagation), onde o erro é transmitido da camada de saída em direção às camadas ocultas, e os pesos de cada neurônio são atualizados para minimizar esse erro. A regra de atualização dos pesos segue a fórmula, em que ω_n se refere ao novo peso após a atualização, ω_0 se refere ao peso antigo e α se refere à taxa de aprendizagem.

$$\omega_n = \omega_0 - \alpha \frac{dE}{d\omega_0} \tag{2.5.5}$$

O MLP (*Multilayer Perceptron*) é um modelo amplamente utilizado em diversas aplicações [34], como reconhecimento de fala, reconhecimento de imagem, tradução automática e outras classificações, além de problemas de regressão. No entanto, apesar de seu poder, as MLPs apresentam alguns desafios, como o treinamento lento. À medida que a rede cresce, o número de parâmetros a serem estimados também aumenta, o que pode levar ao aumento do erro

de estimação e ao risco de *overfitting*. Além disso, o desempenho da MLP pode ser sensível à variação dos valores de inicialização dos parâmetros, já que a rede pode ficar presa em mínimos locais, dependendo do ponto de partida na curva de erro. Outro fator que pode impactar o desempenho é a presença de outliers em número ou magnitude elevados.

3 MATERIAIS E MÉTODOS

Esta seção descreve detalhadamente todos os procedimentos metodológicos adotados no desenvolvimento deste trabalho. Primeiramente, apresenta-se o fluxo geral da metodologia empregada, seguido por descrições específicas sobre: (i) a seleção e preparação das moléculas, (ii) os cálculos de química quântica, (iii) a geração de descritores moleculares e construção dos modelos de aprendizado de máquina, e (iv) o pré-processamento dos dados e avaliação dos modelos preditivos.

A Figura 3.1 apresenta uma visão geral do pipeline metodológico implementado neste estudo.

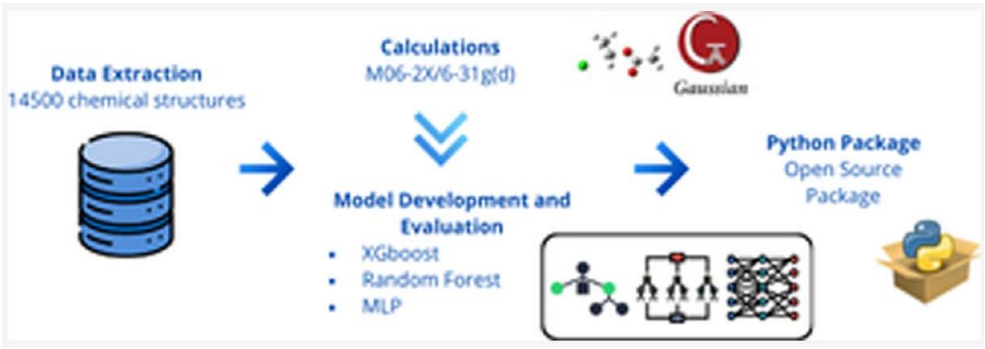

Figura 3.1. Fluxograma geral da metodologia implementada neste trabalho

3.0.1 Seleção e Preparação das Moléculas

A construção do conjunto molecular iniciou-se pela coleta de Insumos Farmacêuticos Ativos (IFAs) e produtos de degradação provenientes de bases de dados públicas reconhecidas, incluindo a lista de Denominações Comuns Brasileiras (DCB) da ANVISA e o banco de dados PubChem.

Além dessas estruturas, moléculas similares foram incluídas com o intuito de expandir o espaço químico representado e incorporar potenciais subprodutos de degradação.

Após a coleta inicial, realizou-se um rigoroso pré-processamento envolvendo:

- remoção de duplicatas por SMILES e InChIKeys;
- padronização das representações químicas;
- verificação da integridade estrutural.

Ao final, obteve-se um conjunto contendo aproximadamente 14.500 moléculas únicas. As estruturas tridimensionais (formato XYZ) foram geradas a partir dos SMILES utilizando a biblioteca *RDKit*[35].

Tabela 3.1. Resumo das etapas de seleção e preparação das moléculas.

Etapa	Descrição
Coleta inicial	DCB (ANVISA) e PubChem
Expansão do conjunto	Moléculas similares a IFAs
Deduplicação	SMILES e InChIKeys
Padronização	Correção estrutural e formatação
Geometrias 3D	Geração via RDKit (XYZ)

3.1 Cálculos de Química Quântica

Os cálculos de estrutura eletrônica foram realizados com o software *Gaussian 16* [25]. Cada molécula foi otimizada geometricamente no nível M06-2X/6-31G(d), seguido de cálculos de frequências harmônicas para garantir que a estrutura correspondia a um mínimo verdadeiro (ausência de frequências imaginárias).

As propriedades termodinâmicas avaliadas a 298,15 K e 1 atm incluem:

- energia eletrônica total;
- entalpia;
- energia livre de Gibbs;
- frequências vibracionais.

As coordenadas otimizadas foram exportadas em formato XYZ para processamento posterior.

3.2 Geração de Descritores e Modelos de Aprendizado de Máquina

A etapa de modelagem utilizou descritores calculados via RDKit e fingerprints do tipo Morgan (raio = 2, 4096 bits), que capturam topologia molecular e subestruturas relevantes.

Foram avaliados três algoritmos supervisionados:

- **XGBoost**: 1000 árvores, profundidade 6, *learning rate* de 0,1 [30];
- **Random Forest**: 500 árvores com profundidade otimizada automaticamente [31];
- **MLP**: arquiteturas 512–256–64, ativação ReLU, otimizador Adam [34].

Todos os modelos foram treinados em cluster de alto desempenho, utilizando as mesmas entradas (*fingerprints*) para garantir comparabilidade.

3.3 Pré-processamento dos Dados e Avaliação dos Modelos

As propriedades alvo Gibbs (G) e Entalpia (H) foram:

1. normalizadas pela maior magnitude observada;
2. reescaladas para o intervalo [0, 1] por *min–max scaling*.

Após a predição, aplicou-se a transformação inversa para retornar à escala original. O conjunto final continha **14.207 moléculas**. Utilizou-se:

- 80% para treinamento (11.365);
- 20% para teste externo (2.842).

A avaliação empregou:

- coeficiente de determinação (R^2);
- coeficiente de predição (Q^2);
- erro quadrático médio da raiz (RMSE).

A validação interna utilizou validação cruzada estratificada em 10 folds.

4 RESULTADOS E DISCUSSÕES

Esta seção apresenta e discute os principais resultados obtidos neste trabalho. Inicialmente, descreve-se o conjunto de dados final utilizado nas análises, seguido de uma avaliação estatística e estrutural por meio de análise exploratória. Em seguida, detalham-se os principais aspectos da implementação computacional, incluindo a organização do pacote Python desenvolvido, bem como suas funcionalidades e fluxo de uso. As subseções a seguir fornecem uma visão completa do comportamento químico, computacional e estatístico do conjunto, fundamentando os modelos preditivos apresentados posteriormente.

4.1 Descrição do Conjunto de Dados

O conjunto completo de dados é disponibilizado na forma de uma tabela química em formato `CSV`, contendo todas as geometrias otimizadas e propriedades moleculares calculadas. O arquivo inclui, entre outros, os seguintes campos: SMILES, entalpia, energia livre de Gibbs, energia eletrônica, frequências vibracionais, coordenadas otimizadas em formato XYZ, capacidade calorífica, massa molar e diversos descritores termodinâmicos. A Tabela 4.1 resume os principais campos presentes na base de dados.

Todos os arquivos `log` brutos provenientes dos cálculos quânticos no nível M06-2X/6-31G(d) também estão disponíveis para download em: http://bit.ly/4jEo7AE.

Tabela 4.1. Descrição dos campos da base de dados.

Campos	**Descrição**
electronic energy (SCF Done)	Energia eletrônica total somada à repulsão nuclear na geometria otimizada (hartree).
zero-point correction (ZPE)	Correção da energia vibracional de ponto zero (hartree).
thermal correction to internal energy	Correção térmica para a energia interna U (hartree).
thermal correction to enthalpy	Correção térmica para a entalpia H (hartree).
thermal correction to Gibbs free energy	Correção térmica para a energia livre de Gibbs G (hartree).
internal energy	Energia interna calculada como $U = E_{\text{SCF}}+$ correção térmica (hartree).
enthalpy	Entalpia calculada como $H = E_{\text{SCF}}+$ correção térmica (hartree).
Gibbs free energy	Energia livre de Gibbs calculada como $G = E_{\text{SCF}}+$ correção térmica (hartree).
heat capacity C_v	Capacidade calorífica a volume constante (cal·mol^{-1}·K^{-1}).
entropy	Entropia (cal·mol^{-1}·K^{-1}).
$\ln Q_{trans}$	Logaritmo natural da função de partição translacional.
$\ln Q_{rot}$	Logaritmo natural da função de partição rotacional.
$\ln Q_{vib}$	Logaritmo natural da função de partição vibracional.
$\ln Q_{elec}$	Logaritmo natural da função de partição eletrônica.
optimized coordinates (XYZ)	Coordenadas cartesianas da geometria convergida.
SMILES molecular mass	Massa molecular derivada da representação SMILES (*atomic mass units* - amu).

4.2 Análise Exploratória do Conjunto de Dados

A análise inicial do conjunto de dados revelou uma ampla diversidade de elementos químicos e grupos funcionais, característica essencial para a construção de modelos termodinâmicos generalizáveis. A Tabela 4.2 apresenta as contagens de elementos atômicos e grupos funcionais identificados.

Os elementos mais abundantes incluem carbono (269 734 ocorrências), oxigênio (67 462) e nitrogênio (37 821), refletindo a composição típica de moléculas bioativas e de produtos de degradação. Halogênios como flúor, cloro e bromo também estão bem representados, mostrando a diversidade estrutural encontrada em APIs e moléculas correlatas.

A análise de grupos funcionais evidencia uma distribuição rica, com destaque para cetonas (11 035), álcoois (10 029), ésteres (6 560), alcenos (5 232) e outros motivos estruturais frequentemente encontrados em moléculas farmacêuticas. Essa diversidade assegura uma cobertura química expressiva, fundamental para o desenvolvimento de modelos preditivos com-

petitivos e robustos.

Tabela 4.2. Distribuição de elementos atômicos e grupos funcionais identificados no conjunto de dados.

Elemento	**Quantidade**
B	16
Br	13
C	269 734
Cl	2 520
F	4 223
N	37 821
O	67 462
P	152
S	3 629
Grupo funcional	**Quantidade**
cetonas	11 035
éteres	1 905
álcoois	10 029
aminas	3 043
alcenos	5 232
ésteres	6 560
alquinos	243
hidrazinas	201

4.2.1 Análise de Redução de Dimensionalidade com UMAP

A redução de dimensionalidade utilizando o algoritmo *Uniform Manifold Approximation and Projection* (UMAP), aplicada às *fingerprints* binárias de Morgan geradas para as moléculas do conjunto de dados, evidenciou de forma clara a diversidade estrutural do conjunto de moléculas do tipo *API-like* desenvolvido neste trabalho. Esse conjunto inclui Insumos Farmacêuticos Ativos (IFAs), produtos de degradação reais e teóricos, bem como impurezas estruturalmente relacionadas, o que resulta em uma ampla cobertura do espaço químico relevante para aplicações farmacêuticas.

No mapa bidimensional gerado pelo UMAP, cada ponto corresponde a uma molécula, e a coloração foi atribuída de acordo com a massa molecular (*MolWeight*). Essa escolha permite visualizar simultaneamente a distribuição estrutural e a variação de massa molecular no espaço projetado. Observa-se que a maior parte das moléculas se concentra em uma região central do mapa, indicando um grau moderado de similaridade estrutural entre uma fração substancial dos compostos. Por outro lado, a presença de aglomerados periféricos e pontos mais isolados indica moléculas estruturalmente distintas, incluindo potenciais produtos de degradação e impurezas singulares, que expandem a cobertura do espaço químico além do *core* estrutural principal.

A Figura 4.1 ilustra o mapa UMAP colorido por massa molecular, destacando tanto a região densa de moléculas estruturalmente semelhantes quanto os domínios mais esparsos, onde se localizam compostos mais exóticos ou menos frequentes em formulações farmacêuticas típicas.

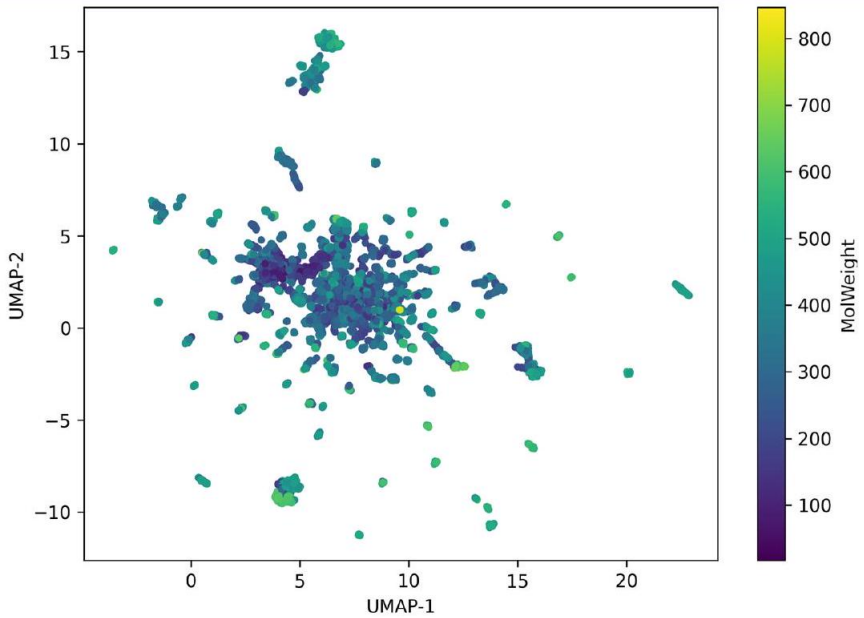

Figura 4.1. Mapa bidimensional obtido por UMAP a partir das *fingerprints* binárias de Morgan do conjunto de dados do tipo *API-like*. Cada ponto representa uma molécula, colorida de acordo com sua massa molecular (*MolWeight*). O conjunto inclui insumos farmacêuticos ativos, produtos de degradação reais e teóricos, além de impurezas estruturalmente relacionadas.

A escolha do UMAP, em detrimento de técnicas lineares como a Análise de Componentes Principais (PCA), deve-se à sua capacidade de preservar tanto a estrutura local quanto a estrutura global dos dados em altas dimensões. Esse aspecto é particularmente importante para *fingerprints* moleculares binárias e esparsas, nas quais as relações de vizinhança são determinadas por padrões de subestruturas químicas compartilhadas entre as moléculas.

Como medida de similaridade, adotou-se o índice de Jaccard, que é equivalente ao coeficiente de Tanimoto para dados binários. Em termos práticos, essa métrica quantifica o quão grande é a interseção entre os bits "ligados" (igual a 1) em comparação à união dos bits ativados nos vetores de *fingerprints*. Dessa forma, moléculas com muitas subestruturas em comum apresentam similaridade elevada, enquanto moléculas estruturalmente distintas exibem similaridade baixa. A combinação entre UMAP e a métrica de Jaccard/Tanimoto permitiu uma representação realista das relações de similaridade no espaço químico, reforçando a adequação do banco de dados para o treinamento de modelos de aprendizado de máquina voltados à predição de propriedades termodinâmicas no contexto farmacêutico.

4.2.2 Avaliação Quantitativa do Desempenho dos Modelos

Para avaliar o desempenho preditivo dos modelos de aprendizado de máquina desenvolvidos neste trabalho, foram utilizadas três métricas estatísticas principais: o coeficiente de correlação quadrático preditivo (Q^2) [36], o coeficiente de determinação (R^2) [37] e o erro quadrático médio da raiz (RMSE) [38]. Em conjunto, essas métricas permitem quantificar, respectivamente, a capacidade preditiva sob validação, a fração da variância explicada pelos modelos e a magnitude média dos erros de predição.

De forma geral, valores elevados de Q^2 e R^2, próximos de 1, indicam que o modelo consegue reproduzir com boa fidelidade as variações dos dados de referência. Por outro lado, valores baixos de RMSE indicam erros médios de predição pequenos na escala da propriedade estudada. Assim, a combinação Q^2 alto, R^2 alto e RMSE baixo é um indicativo robusto de desempenho preditivo satisfatório.

Os resultados resumidos na Tabela 4.3 mostram que, para a predição da energia livre de Gibbs, o modelo XGBoost apresentou o melhor desempenho global. Na validação interna (validação cruzada estratificada em 10 *folds*), foram obtidos $Q^2 = 0{,}997 \pm 0{,}001$, $R^2 = 0{,}9975 \pm 0{,}0001$ e RMSE $= 0{,}009 \pm 0{,}002$. Na validação externa, aplicada a um conjunto de teste independente, o coeficiente de correlação de Pearson ao quadrado (r^2_{ext}) foi de 0,9825, evidenciando forte concordância entre os valores previstos e os valores de referência oriundos dos cálculos de estrutura eletrônica.

Para a predição de entalpia, o modelo *Multi-Layer Perceptron* (MLP) se destacou na validação externa, alcançando $Q^2_{\text{ext}} = 0{,}9751$, RMSE $= 0{,}0299$ e $r^2_{\text{ext}} = 0{,}9876$. Esses resultados indicam alta acurácia preditiva e excelente correlação entre os valores previstos e aqueles obtidos por química quântica, reforçando o potencial do MLP para modelar relações não lineares complexas entre descritores estruturais e propriedades termodinâmicas.

O modelo *Random Forest* também apresentou desempenho consistentemente elevado tanto para energia livre de Gibbs quanto para entalpia, embora de forma geral ligeiramente inferior aos melhores resultados obtidos por XGBoost (para G) e MLP (para H). Ainda assim, sua performance robusta e estável o torna uma alternativa interessante, especialmente quando se deseja um balanço entre desempenho e interpretabilidade dos *features*.

A concordância entre as métricas de validação interna e externa é um ponto central na análise de robustez dos modelos. Em particular, a ausência de queda acentuada no desempenho ao passar do conjunto de treinamento/validação para o conjunto de teste independente indica que não houve sobreajuste significativo (*overfitting*). Isso significa que os modelos aprendem padrões gerais das relações entre estrutura e propriedades, e não apenas memorizam o conjunto de treinamento. Essa característica é fundamental para aplicações práticas, nas quais o objetivo é prever propriedades de novas moléculas ainda não avaliadas computacionalmente ou experimentalmente.

Tabela 4.3. Desempenho preditivo dos modelos XGBoost, Random Forest e MLP para energia livre de Gibbs e entalpia.[a]

Validação externa — energia livre de Gibbs				
Modelo	Q^2_{ext}	R^2_{ext}	RMSE	r^2_{ext}
XGBoost	0.9653	0.9975	0.0353	0.9825
Random Forest	0.9361	0.9880	0.0479	0.9685
MPL	0.9773	0.9911	0.0285	0.9889
Validação interna — energia livre de Gibbs (10-*fold* estratificado)				
Modelo	Q^2_{int}	R^2_{int}	RMSE_{int}	r^2_{int}
XGBoost	$0{,}997 \pm 0{,}001$	$0{,}9975 \pm 0{,}0001$	$0{,}009 \pm 0{,}002$	$0{,}9988 \pm 0{,}0005$
Random Forest	$0{,}988 \pm 0{,}001$	$0{,}9879 \pm 0{,}0001$	$0{,}021 \pm 0{,}001$	$0{,}9944 \pm 0{,}0006$
MLP	$0{,}9911 \pm 0{,}0008$	$0{,}99106 \pm 0{,}00009$	$0{,}00032 \pm 0{,}00003$	$0{,}9958 \pm 0{,}0005$
Validação externa — entalpia				
Modelo	Q^2_{ext}	R^2_{ext}	RMSE	r^2_{ext}
XGBoost	0.9433	0.9663	0.0451	0.9716
Random Forest	0.9363	0.9879	0.0478	0.9687
MLP	0.9751	0.9896	0.0299	0.9876
Validação interna — entalpia (10-*fold* estratificado)				
Modelo	Q^2_{int}	R^2_{int}	RMSE_{int}	r^2_{int}
XGBoost	$0{,}966 \pm 0{,}002$	$0{,}9663 \pm 0{,}0002$	$0{,}0346 \pm 0{,}0008$	$0{,}983 \pm 0{,}001$
Random Forest	$0{,}988 \pm 0{,}001$	$0{,}9879 \pm 0{,}0001$	$0{,}021 \pm 0{,}001$	$0{,}9944 \pm 0{,}0006$
MLP	$0{,}990 \pm 0{,}001$	$0{,}9896 \pm 0{,}0001$	$0{,}00037 \pm 0{,}00004$	$0{,}9948 \pm 0{,}0006$

[a] Resultados de validação interna expressos como média $\pm$ desvio padrão obtidos a partir de validação cruzada estratificada em 10 *folds*. As métricas de validação externa foram calculadas em um conjunto de teste independente.

4.2.3 Gráficos de Paridade e Qualidade das Predições

Os gráficos de paridade (*parity plots*) fornecem uma forma visual direta de comparar os valores previstos pelos modelos com os valores de referência obtidos pelos cálculos de química quântica. A Figura 4.2 apresenta os gráficos de paridade para energia livre de Gibbs e entalpia, considerando os três modelos avaliados (XGBoost, Random Forest e MLP). Todos os valores foram previamente normalizados para o intervalo [0, 1], facilitando a comparação entre propriedades e modelos.

Em cada gráfico, a linha diagonal vermelha representa a situação ideal em que os valores previstos coincidem exatamente com os valores de referência ($r^2 = 1$). Pontos próximos a essa diagonal indicam predições acuradas, enquanto pontos afastados sugerem subpredição ou superpredição.

Para a energia livre de Gibbs (linha superior da Figura 4.2), o modelo XGBoost apresenta o alinhamento mais próximo da diagonal, com dispersão mínima ao longo de todo o intervalo

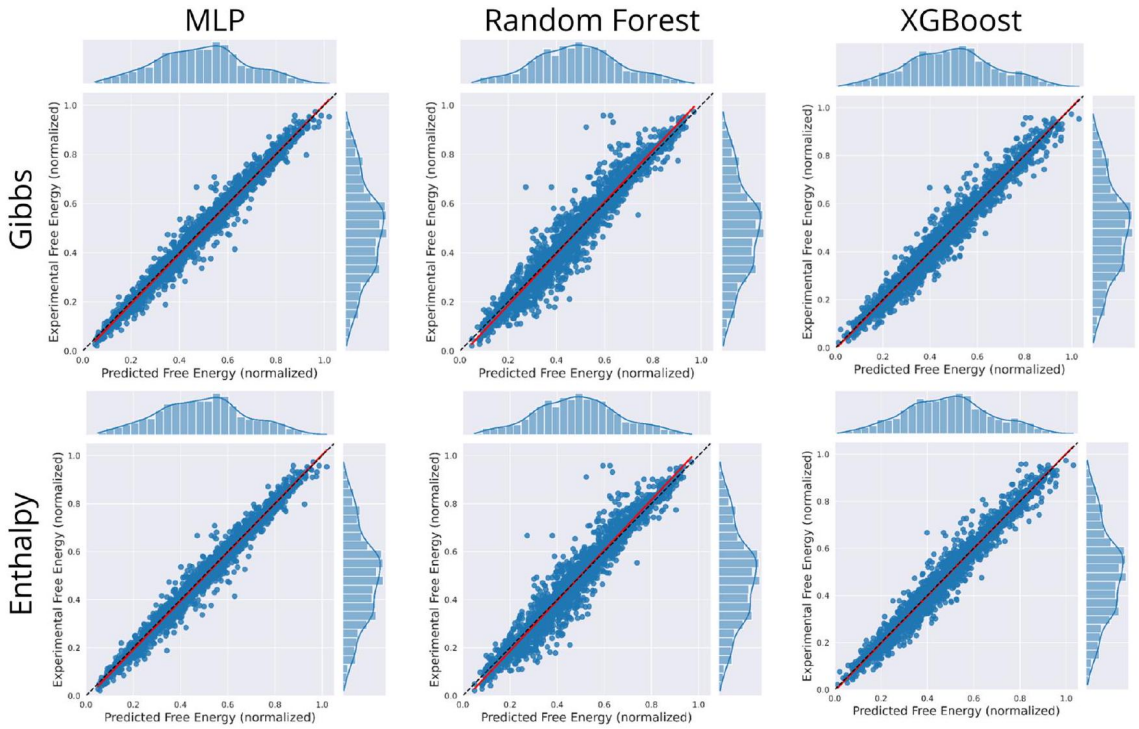

Figura 4.2. Gráficos de paridade comparando os valores previstos e experimentais (normalizados) para energia livre de Gibbs (linha superior) e entalpia (linha inferior), utilizando três modelos de aprendizado de máquina: *Multi-Layer Perceptron* (MLP), *Random Forest* e XGBoost. A linha diagonal vermelha representa a correlação ideal ($r^2 = 1$), na qual os valores previstos e experimentais coincidem. Pontos próximos a essa linha indicam predições acuradas, enquanto desvios revelam sub ou superestimações pelos modelos. Os histogramas posicionados acima e ao lado de cada gráfico de dispersão ilustram, respectivamente, a distribuição dos valores previstos e dos valores experimentais.

normalizado. Esse comportamento reforça sua superioridade na tarefa de prever G. O MLP apresenta um desempenho comparável, mas com ligeiramente maior dispersão em torno da linha de referência, indicando variância um pouco mais elevada nas predições. O modelo Random Forest exibe distribuição mais espalhada, especialmente em valores intermediários, o que sugere menor consistência em comparação aos outros dois modelos.

No caso da entalpia (linha inferior da Figura 4.2), o MLP assume a liderança em termos de qualidade de predição, com pontos fortemente concentrados ao longo da diagonal, demonstrando excelente concordância entre valores previstos e de referência. O XGBoost apresenta desempenho também bastante satisfatório, com boa aderência à linha ideal, ao passo que o Random Forest volta a exibir maior dispersão, particularmente nas extremidades do intervalo.

4.2.4 Domínio de Aplicabilidade dos Modelos

Para garantir que as predições realizadas pelos modelos se situam em uma região de espaço químico bem representada no conjunto de treinamento, foi avaliado o domínio de aplicabilidade (*Applicability Domain*, AD) por meio da similaridade de Tanimoto. Para isso, calculou-se a similaridade de Tanimoto entre cada molécula do conjunto de teste externo e as moléculas do conjunto de treinamento, utilizando as mesmas *fingerprints* de Morgan empregadas na modelagem (raio = 2, 4096 bits).

A Figura 4.3 apresenta a distribuição das similaridades máximas de Tanimoto entre as moléculas do conjunto de teste e o conjunto de treinamento.

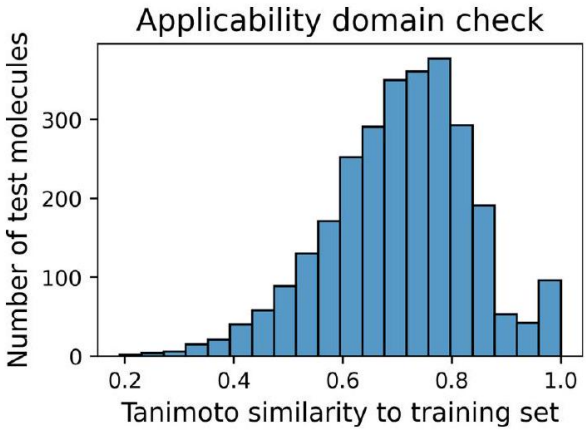

Figura 4.3. Distribuição das similaridades de Tanimoto entre as moléculas do conjunto de teste externo e o conjunto de treinamento, calculadas a partir de *fingerprints* de Morgan (raio = 2, 4096 bits).

Observa-se que a maior parte dos compostos de teste apresenta similaridade de Tanimoto superior a 0,6, com um pico de distribuição entre 0,7 e 0,85. Esse resultado indica que, em geral, as predições são realizadas em uma região do espaço químico bem representada no conjunto de treinamento, reduzindo o risco de extrapolação para regiões estruturalmente muito distintas. Apenas uma fração pequena das moléculas exibe similaridade inferior a 0,4, o que sugere baixa incidência de predições fora do domínio de aplicabilidade.

Essa análise complementa as métricas de desempenho, fornecendo evidência adicional de que os resultados obtidos na validação externa são confiáveis e não decorrentes de extrapolações extremas.

4.2.5 Análise de *Scaffolds* de Bemis–Murcko

A análise dos *scaffolds* de Bemis–Murcko [39] (Figura 4.4) foi realizada com o objetivo de identificar os arcabouços estruturais centrais presentes no conjunto de dados. Essa metodologia isola o núcleo estrutural de cada molécula ao remover substituintes periféricos, preservando apenas o framework básico composto por anéis e ligações de conexão. Dessa forma, é possível avaliar de forma mais objetiva a distribuição dos motivos estruturais fundamentais que caracterizam o espaço químico representado.

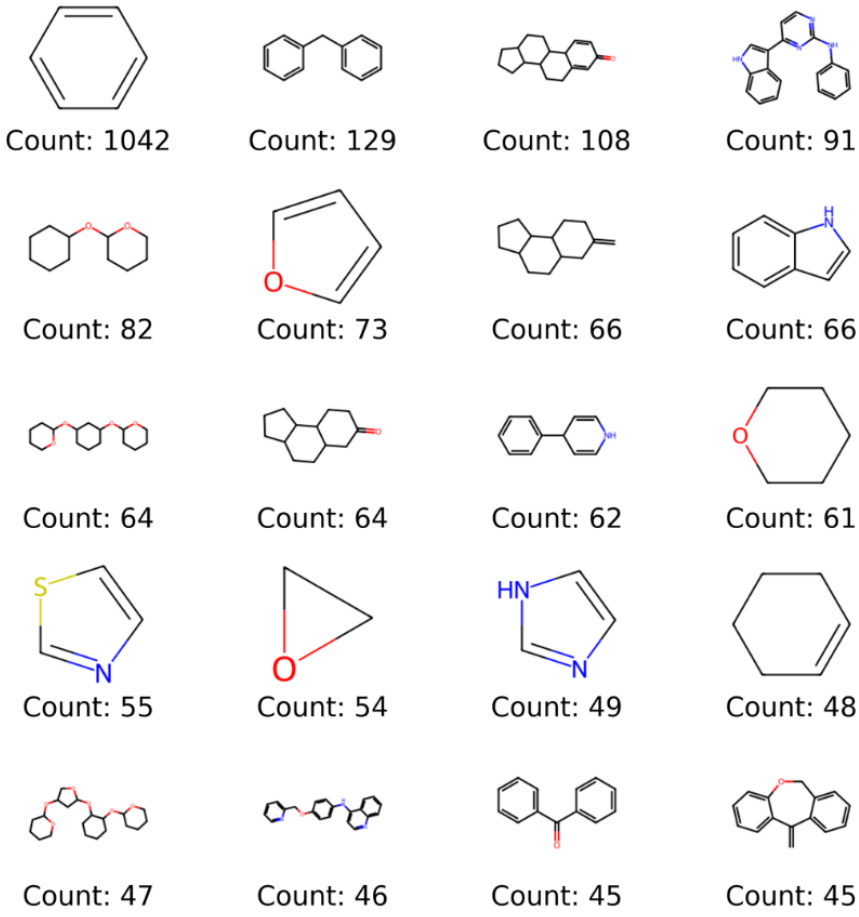

Figura 4.4. Vinte *scaffolds* de Bemis–Murcko mais frequentes identificados no conjunto de dados, destacando a predominância de núcleos aromáticos e heteroaromáticos típicos de moléculas do tipo *API-like*.

Os resultados revelaram a predominância marcante de sistemas aromáticos e heteroaromáticos, refletindo a natureza "API-like" do conjunto de moléculas analisado. O anel benzênico simples foi identificado como o *scaffold* mais frequente, com 1 042 ocorrências, seguido por estruturas bicíclicas fundidas ou conectadas, como bifenila (129 ocorrências) e tetraidro-β-carbolina (108 ocorrências). Esses núcleos são amplamente encontrados em fármacos comercializados, devido à sua capacidade de fornecer rigidez, planaridade e diversidade eletrônica às estruturas moleculares.

Além dos sistemas aromáticos clássicos, observou-se frequência significativa de heterociclos contendo oxigênio, nitrogênio ou enxofre, como morfolina, piperazina e tiazóis. A presença desses sistemas indica a ocorrência abundante de fragmentos farmacofóricos típicos, frequentemente associados a propriedades bioativas, estabilidade térmica e rotas de degradação comuns em moléculas orgânicas complexas.

A diversidade dos *scaffolds* identificados — variando de anéis aromáticos simples a estruturas policíclicas complexas e heterociclos de diferentes tamanhos — demonstra que o conjunto de dados captura tanto motivos estruturais amplamente representados em medicamentos quanto núcleos mais raros, possivelmente associados a produtos de degradação específicos ou impurezas minoritárias. Essa diversidade estrutural complementa a análise de redução de dimensionalidade (UMAP) e o estudo de domínio de aplicabilidade, reforçando que o banco de dados cobre adequadamente regiões amplas e relevantes do espaço químico farmacêutico.

A Figura 4.4 apresenta os 20 *scaffolds* mais frequentes do conjunto, juntamente com sua contagem de ocorrência. O script completo utilizado para realizar essa análise, assim como as demais análises exploratórias e de validação, está disponível em:
https://github.com/jeffrichardchemistry/thermopred.

4.3 Implementação

A implementação desenvolvida neste trabalho foi utilizada para transformar o conjunto teórico e metodológico apresentado nos capítulos anteriores em uma ferramenta computacional funcional, reprodutível e acessível. Essa etapa integra os dados termoquímicos derivados de cálculos de estrutura eletrônica, os modelos de aprendizado de máquina treinados e uma interface programática simplificada, resultando em um pacote Python capaz de prever propriedades termodinâmicas com rapidez e elevada acurácia. A seguir, descrevem-se a estrutura, o funcionamento e os principais aspectos técnicos do pacote *ThermoPred*.

4.3.1 Visão Geral da Implementação

Nesta etapa do trabalho, foi desenvolvido um pacote Python de código aberto denominado *ThermoPred* [29], acompanhado por um conjunto de dados abrangente contendo informações termoquímicas e resultados de cálculos de química quântica para aproximadamente 14 500 moléculas do tipo *API-like* e seus respectivos produtos de degradação. O principal objetivo do pacote é permitir a predição rápida e confiável de energia livre de Gibbs e entalpia a partir de representações moleculares simples (SMILES), simulando com elevada acurácia valores tipicamente obtidos por meio de softwares de química computacional, como Gaussian 16.

Tanto o pacote quanto o conjunto completo de dados estão disponíveis publicamente no repositório oficial do projeto:
https://github.com/jeffrichardchemistry/thermopred.

O *ThermoPred* foi projetado para atuar como uma plataforma aberta e extensível, permitindo a integração em fluxos computacionais de predição termoquímica, automação de estudos de degradação e triagem virtual de moléculas bioativas.

4.3.2 Arquitetura do Pacote Python

O pacote *ThermoPred* foi desenvolvido com o intuito de oferecer uma interface simples, modular e de uso intuitivo para usuários de diferentes níveis de experiência em modelagem molecular. Ele é inteiramente implementado em Python e disponibiliza modelos de aprendizado de máquina previamente treinados — incluindo XGBoost, *Random Forest* e *Multi-Layer Perceptron* (MLP) — otimizados para acurácia e eficiência computacional.

A Figura 4.5 apresenta a estrutura do repositório, organizada de forma a facilitar manutenção, expansão e reprodutibilidade. A pasta *dataset* contém o arquivo principal com as propriedades termoquímicas utilizadas no treinamento. A documentação em *Markdown* e PDF encontra-se na pasta *docs*. O diretório raiz inclui arquivos essenciais para distribuição e instalação do pacote (como `setup.py`, `requirements.txt`, `LICENSE` e `README.md`). O diretório *Thermopred* armazena os módulos responsáveis pelo cálculo das propriedades termoquímicas, além dos subdiretórios contendo os modelos treinados para as propriedades de entalpia e energia livre de Gibbs.

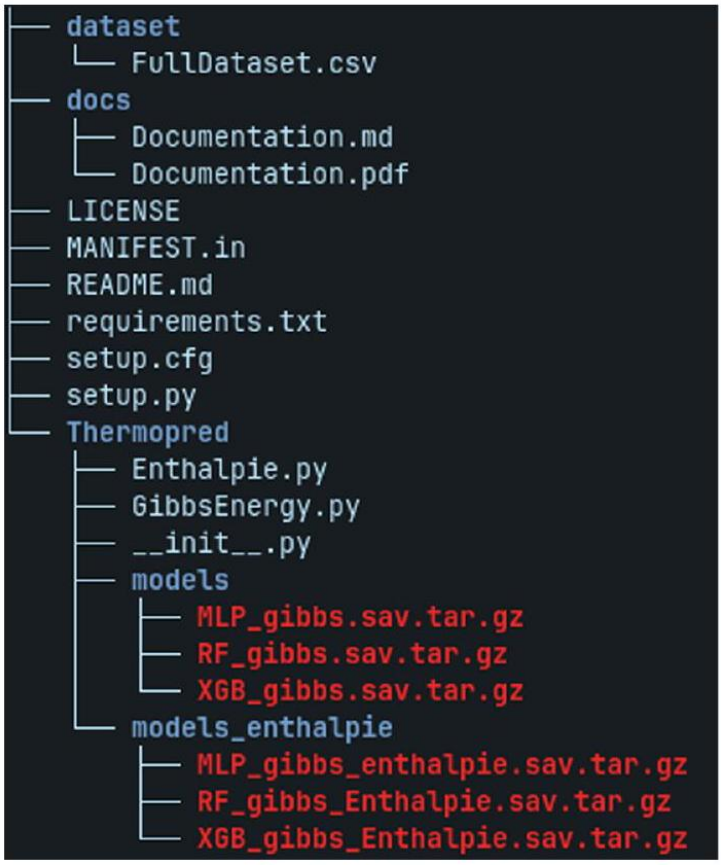

Figura 4.5. Estrutura de diretórios do projeto *ThermoPred*. A pasta *dataset* contém o arquivo de propriedades termoquímicas (`FullDataset.csv`). O diretório *docs* agrega a documentação completa do projeto. Os arquivos `LICENSE`, `MANIFEST.in`, `README.md`, `requirements.txt`, `setup.cfg` e `setup.py` encontram-se no diretório raiz e fornecem configurações de instalação e gerenciamento de dependências. O diretório *Thermopred* armazena os módulos Python responsáveis por calcular propriedades termoquímicas, incluindo `Enthalpie.py` e `GibbsEnergy.py`, bem como os modelos pré-treinados correspondentes.

4.3.3 Funcionalidades e Modo de Uso

O *ThermoPred* fornece uma interface direta para a predição das propriedades termoquímicas investigadas — energia livre de Gibbs e entalpia — a partir de representações SMILES. Para isso, o pacote encapsula modelos de aprendizado de máquina previamente treinados, garantindo predições rápidas, reprodutíveis e compatíveis com unidades e formatos tradicionalmente utilizados em softwares de estrutura eletrônica.

Os modelos foram otimizados para replicar a saída numérica de cálculos quânticos realizados com Gaussian 16, minimizando o desvio entre as predições e os valores de referência. Dessa forma, o *ThermoPred* permite realizar análises termoquímicas em larga escala com custo computacional drasticamente reduzido em comparação a métodos ab initio.

A instalação e execução do pacote podem ser realizadas facilmente em qualquer ambiente Python. Prever energia livre de Gibbs ou entalpia requer apenas importar os módulos adequados e fornecer a string SMILES. O trecho de código apresentado na Figura 4.6 ilustra

esse fluxo de trabalho.

```
from Thermopred.Enthalpie import EnthalpieEnergy
from Thermopred.GibbsEnergy import GibbsFreeEnergy

smiles = 'CN1C=CN(CCCN(c2cc(Cl) ccc2O) c2ccccc2S)CCl'

ee = EnthalpieEnergy()
result_enthalpie = ee.predict(smiles)

gfe = GibbsFreeEnergy()
result_gibbs = gfe.predict(smiles=smiles)
```

Figura 4.6. Exemplo de código Python demonstrando o uso do pacote *ThermoPred* para prever energia livre de Gibbs e entalpia. O usuário fornece a representação SMILES da molécula, e os modelos pré-treinados retornam as propriedades termoquímicas correspondentes.

Por ser totalmente aberto à comunidade, o *ThermoPred* também incentiva contribuições externas, incluindo melhorias na interface, incorporação de novos tipos de descritores, expansão do conjunto de dados e retraining de modelos para cenários específicos. Como resultado, o pacote foi concebido como uma plataforma flexível para estudos de predição termoquímica de moléculas do tipo *API-like* e seus produtos de degradação, atendendo a demandas de triagem virtual, priorização de experimentos e análise mecanística em pesquisa farmacêutica.

4.3.4 Disponibilidade dos Dados e Reprodutibilidade

Todo o conjunto de dados utilizado para o treinamento e validação dos modelos — incluindo energias eletrônicas, entalpias, energias livres de Gibbs, frequências vibracionais e propriedades termoquímicas derivadas — encontra-se disponível de forma aberta no repositório oficial do projeto *ThermoPred*. A base de dados inclui aproximadamente 14 500 moléculas do tipo *API-like* e seus produtos de degradação, e está acompanhada das estruturas moleculares correspondentes em formato SMILES e coordenadas otimizadas.

Os arquivos completos contendo as saídas do Gaussian (arquivos `.log`) utilizados para extrair as propriedades termoquímicas podem ser acessados no repositório hospedado na plataforma HuggingFace:
https://huggingface.co/datasets/diullio/ThermoPredLogs.

O pacote *ThermoPred*, desenvolvido neste trabalho, também está disponível como software livre sob licença GPL-3.0, permitindo auditoria, reprodução e extensão dos métodos apresentados. O código-fonte completo, juntamente com os modelos pré-treinados, documentação e scripts de validação, está acessível em:
https://github.com/jeffrichardchemistry/thermopred.

Para facilitar a experimentação, uma interface interativa baseada em *Streamlit* foi disponibilizada, permitindo que usuários testem as predições de energia livre de Gibbs e entalpia

diretamente no navegador, sem necessidade de instalação prévia de dependências, e está acessível em:
https://thermopred.streamlit.app/

5 CONCLUSÃO

O presente trabalho apresentou o desenvolvimento de uma plataforma integrada para predição termoquímica de moléculas do tipo *API-like*, combinando a construção de um banco de dados quântico de grande escala, o treinamento de modelos de aprendizado de máquina validados e a implementação de um pacote Python de código aberto. A base de dados, composta por mais de 14 500 moléculas incluindo IFAs e seus potenciais produtos de degradação, foi obtida por meio de otimizações geométricas completas no nível M06-2X/6-31G(d) com o software Gaussian 16, seguidas da extração sistemática de energias eletrônicas, correções vibracionais e propriedades termoquímicas derivadas. O conjunto resultante, construído ao longo de aproximadamente quatro meses de processamento contínuo, constitui uma das maiores coleções públicas dedicadas especificamente à termoquímica de moléculas do tipo *API-like* e seus degradantes.

Com base nesse repositório, foram desenvolvidos modelos de aprendizado de máquina baseados em XGBoost, *Random Forest* e *Multi-Layer Perceptron*, configurados para prever energia livre de Gibbs (ΔG) e entalpia (ΔH) no mesmo nível teórico de referência. As análises de validação interna e externa demonstraram desempenho consistente, com elevados valores de Q^2 e R^2 e baixos erros médios (RMSE), indicando alta capacidade de replicação numérica dos resultados obtidos por cálculos quânticos. A avaliação do domínio de aplicabilidade, associada às projeções UMAP do espaço químico, confirmou que os modelos operam majoritariamente em regiões estruturalmente bem representadas pelo conjunto de treinamento, assegurando robustez, estabilidade preditiva e baixo risco de sobreajuste.

Como forma de disponibilizar esses avanços à comunidade científica e ao setor industrial, foi desenvolvido o pacote *ThermoPred*, uma ferramenta modular, reprodutível e de uso intuitivo, capaz de estimar rapidamente propriedades termoquímicas a partir de representações moleculares SMILES. O pacote integra modelos pré-treinados, scripts de validação, documentação técnica e infraestrutura de dados aberta, permitindo sua aplicação em fluxos computacionais automatizados e em cenários práticos, como triagem virtual de moléculas, avaliação preliminar de estabilidade, análise mecanística de rotas degradativas e apoio ao planejamento de estudos regulatórios.

A integração entre um banco de dados quântico especializado, modelos preditivos estatisticamente validados e uma interface computacional acessível representa uma contribuição relevante para a química computacional aplicada ao desenvolvimento farmacêutico. Ao reduzir drasticamente o custo computacional associado à obtenção de propriedades termoquímicas e ao promover transparência e reprodutibilidade por meio da disponibilização aberta de dados e códigos, a plataforma estabelece uma base sólida para a incorporação de abordagens preditivas em estratégias baseadas em conhecimento, em consonância com o cenário regulatório contemporâneo.

Por fim, os resultados obtidos indicam que a integração entre química quântica e inteligência artificial constitui um caminho promissor para ampliar a compreensão mecanística de

processos degradativos em IFAs. Como perspectivas futuras, destacam-se a incorporação de novos descritores moleculares, a expansão do banco de dados para outras classes químicas, o desenvolvimento de modelos multitarefa capazes de prever múltiplas propriedades simultaneamente e a integração com arquiteturas modernas de aprendizado profundo. Essas extensões poderão consolidar o *ThermoPred* como uma plataforma evolutiva para predição termoquímica no contexto farmacêutico, ampliando seu impacto científico e tecnológico.

REFERÊNCIAS

[1] World Health Organization. Who guidelines on stability testing of active pharmaceutical ingredients and finished pharmaceutical products. Technical report, World Health Organization, Geneva, 2018. In: WHO Technical Report Series, No. 1010, Annex 10. Acesso em: 21 nov. 2025.

[2] Steven W. Baertschi, Karen M. Alsante, and Robert A. Reed. *Pharmaceutical stress testing*. CRC Press, Boca Raton, 2 edition, 2013.

[3] Alany I. Ribeiro, Dayvson J. Palmeira, Irwin A. P. Linares, Carlos E. M. Santos, Rodrigo S. Martins, Eder Lorenzato Jr., Angelica Abido, Gustavo G. da Silva, Lillian F. S. Nascimento, and Diogo dos Santos Alves. Predição in silico dos produtos de degradação da cefalexina com o software degradation plot. *ALTOX*, 2023. Acesso em: 21 nov. 2025.

[4] M. Y. Kawamura, D. J. Ponting, C. G. Barber, and et al. Computational mechanistic study on n-nitrosation reaction of secondary amines. *Journal of Molecular Modeling*, 31:303, 2025. Received: 18 July 2025; Accepted: 16 September 2025; Published: 16 October 2025.

[5] Umesh Dobariya, Narendra Chauhan, Himani Patel, and Nidhi Pardeshi. Nitrosamine impurities: origin, control and regulatory recommendations. *International Journal of Drug Regulatory Affairs*, 9:77–80, 2021. Acesso em: 21 nov. 2025.

[6] European Medicines Agency. Questions and answers for marketing authorisation holders/applicants on the chmp opinion for the article 5(3) of regulation (ec) no 726/2004 referral on nitrosamine impurities in human medicinal products. Technical report, European Medicines Agency, Amsterdam, 2020. EMA/409815/2020. Acesso em: 21 nov. 2025.

[7] International Council for Harmonisation. Ich q1a(r2): Stability testing of new drug substances and products. Technical report, ICH, Geneva, 2003. Acesso em: 21 nov. 2025.

[8] International Council for Harmonisation. Ich q3b(r2): Impurities in new drug products. Technical report, ICH, Geneva, 2006. Acesso em: 21 nov. 2025.

[9] International Council for Harmonisation. Assessment and control of dna reactive (mutagenic) impurities in pharmaceuticals to limit potential carcinogenic risk – m7(r2). Ich harmonised guideline, ICH, Geneva, 2023. Acesso em: 21 nov. 2025.

[10] Agência Nacional de Vigilância Sanitária. Resolução da diretoria colegiada - rdc nº 53, de 4 de dezembro de 2015. Technical report, ANVISA, Brasília, DF, 2015. Diário Oficial da União. Acesso em: 21 nov. 2025.

[11] Agência Nacional de Vigilância Sanitária. Resolução da diretoria colegiada - rdc nº 318, de 6 de novembro de 2019. Technical report, ANVISA, Brasília, DF, 2019. Diário Oficial da União nº 216. Acesso em: 21 nov. 2025.

[12] Agência Nacional de Vigilância Sanitária. Resolução da diretoria colegiada - rdc nº 964, de 20 de fevereiro de 2025. Technical report, ANVISA, Brasília, DF, 2025. Diário Oficial da União. Acesso em: 21 nov. 2025.

[13] Somenath Mitra, editor. *Sample preparation techniques in analytical chemistry*, volume 162 of *Chemical Analysis: A Series of Monographs on Analytical Chemistry and its Applications*. John Wiley & Sons, Hoboken, NJ, 2003.

[14] Steven W. Baertschi. *Stress testing of drug substances and drug products: understanding and implementation*. Informa Healthcare, New York, 2011. Acesso em: 21 nov. 2025.

[15] Donald L. Pavia, Gary M. Lampman, George S. Kriz, and James A. Vyvyan. *Introduction to spectroscopy*. Cengage Learning, Stamford, CT, 5 edition, 2015.

[16] Ishvarchandra Parmar, Hemalatha Rathod, and Shayeda Shaik. A review: recent trends in analytical techniques for characterization and structure elucidation of impurities in drug substances. *Indian Journal of Pharmaceutical Sciences*, 83(3):402–415, 2021. Acesso em: 21 nov. 2025.

[17] Surbhi Mehta, Ravi P. Shah, and Saranjit Singh. Strategy for identification and characterization of small quantities of drug degradation products using lc and lc-ms: application to valsartan, a model drug. *Drug Testing and Analysis*, 2(2):82–90, 2010. Acesso em: 21 nov. 2025.

[18] Joseph A. DiMasi, Henry G. Grabowski, and Ronald W. Hansen. Innovation in the pharmaceutical industry: new estimates of r&d costs. *Journal of Health Economics*, 47:20–33, 2016.

[19] Sándor Görög. The workflow of impurity profiling. *Trends in Analytical Chemistry*, 25(8):755–757, 2006.

[20] Robert G. Parr and Weitao Yang. *Density-functional theory of atoms and molecules*. Oxford University Press, 1995.

[21] Walter Kohn and Lu J. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133–A1138, 1965.

[22] Donald A. McQuarrie. *Statistical mechanics*. Harper & Row, New York, 1973.

[23] Yan Zhao and Donald G. Truhlar. The m06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements. *Theoretical Chemistry Accounts*, 120(1–3):215–241, 2008.

[24] Narbe Mardirossian and Martin Head-Gordon. Thirty years of density functional theory in computational chemistry. *Molecular Physics*, 115(19):2315–2372, 2017.

[25] Frisch, M. J. et al. *Gaussian 16 Revision C.01*. Gaussian, Inc., Wallingford, CT, 2016. Referência padrão do software Gaussian.

[26] O. Anatole von Lilienfeld. Quantum machine learning in chemical compound space. *Angewandte Chemie International Edition*, 57(16):4164–4169, 2018. Referência fundamental para união de ML e química computacional. Acesso em: 21 nov. 2025.

[27] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1:140022, 2014.

[28] Justin S. Smith, Olexandr Isayev, and Adrian E. Roitberg. Ani-1: A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Scientific Data*, 4:170193, 2017.

[29] Diullio P. Santos, Jefferson R. Dias-Silva, Luiz H. K. Q. Júnior, and Heibbe C. B. de Oliveira. Thermopred: Ai-enhanced quantum chemistry data set and machine learning toolkit for thermochemical properties of api-like compounds and their degradants. *Journal of Chemical Information and Modeling*, 2025. Application Note. Acesso em: 21 nov. 2025.

[30] Tianqi Chen and Carlos Guestrin. Xgboost: a scalable tree boosting system. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[31] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[32] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. Acesso em: 21 nov. 2025.

[33] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, New York, 2006.

[34] M. W. Gardner and S. R. Dorling. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32(14–15):2627–2636, 1998. Acesso em: 21 nov. 2025.

[35] Greg Landrum. Rdkit: open-source cheminformatics software, 2016. Software disponível em código aberto. Acesso em: 21 nov. 2025.

[36] Gerrit Schüürmann, Ralf Ebert, Jingwen Chen, Bin Wang, and Ralph Kühne. External validation and prediction employing the predictive squared correlation coefficient: test set

activity mean vs training set activity mean. *Journal of Chemical Information and Modeling*, 48:2140–2145, 2008. Acesso em: 21 nov. 2025.

[37] Alessandro Di Bucchianico. *Coefficient of determination (R^2)*. John Wiley & Sons, Hoboken, NJ, 2008. Acesso em: 21 nov. 2025.

[38] T. O. Hodson. Root-mean-square error (rmse) or mean absolute error (mae): when to use them or not. *Geoscientific Model Development*, 15(14):5481–5487, 2022. Acesso em: 21 nov. 2025.

[39] Oliver B. Scott and Edith A. W. Chan. Scaffoldgraph: an open-source library for the generation and analysis of molecular scaffold networks and scaffold trees. *Bioinformatics*, 36(12):3930–3931, 2020. Acesso em: 21 nov. 2025.