Gabriel Marques de Amaral Gravina

Utilização da Inteligência Artificial Explicável (XAI) na Pré-Seleção de Ligantes em Triagem Virtual

Campos dos Goytacazes, RJ

25 de novembro de 2025

Gabriel Marques de Amaral Gravina

Utilização da Inteligência Artificial Explicável (XAI) na Pré-Seleção de Ligantes em Triagem Virtual

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Estadual do Norte Fluminense Darcy Ribeiro como requisito para a obtenção do título de Bacharel em Ciência da Computação, sob orientação do Prof. Dr. João Luiz de Almeida Filho.

Universidade Estadual do Norte Fluminense Darcy Ribeiro – UENF

Centro de Ciência e Tecnologia – CCT

Laboratório de Ciências Matemáticas – LCMAT

Curso de Ciência da Computação

Orientador: Prof. Dr. João Luiz de Almeida Filho

Campos dos Goytacazes, RJ

25 de novembro de 2025

Gabriel Marques de Amaral Gravina

Utilização da Inteligência Artificial Explicável (XAI) na Pré-Seleção de Ligantes em Triagem Virtual

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Estadual do Norte Fluminense Darcy Ribeiro como requisito para a obtenção do título de Bacharel em Ciência da Computação, sob orientação do Prof. Dr. João Luiz de Almeida Filho.

Campos dos Goytacazes, RJ, 25 de novembro de 2025:

Prof. Dr. João Luiz de Almeida Filho
Orientador

Prof. Annabell Del Real Tamariz
Membro da Banca

Prof. Luis Mariano del Val Cura
Membro da Banca

Prof. Jorge Hernandez Fernandez
Convidado

Campos dos Goytacazes, RJ 25 de novembro de 2025

Agradecimentos

Aos meus pais, pelo carinho, incentivo e amor incondicionais. Às minhas irmãs, Gerlene e Roberta, por sempre me apoiarem e quererem o meu melhor. À minha família, por sempre me trazer conforto, alegria e afeto. À minha namorada, por todo o apoio e companheirismo. Aos meus amigos, em especial ao João Bosco e ao João Vítor Fernandes, que sempre estiveram presentes. Ao professor João Luiz, pela orientação, ensinamentos compartilhados e por ter sido um guia fundamental em minha jornada acadêmica. Aos membros da banca, pela disponibilidade e pela participação nesta etapa tão importante da minha vida. Por fim, ao corpo docente, técnico e administrativo da UENF, essenciais para a concretização deste trabalho.

Resumo

Nos últimos anos, algoritmos de Aprendizado de Máquina (ML) têm se tornado cada vez mais precisos. No entanto, esse aumento de precisão trouxe consigo um grande aumento de complexidade e o fenômeno das "caixas-pretas". Com o funcionamento interno opaco e difícil de interpretar, embora possível observar o resultado de uma previsão, tornou-se difícil descobrir o porquê. Em aplicações da Química Computacional, como a Triagem Virtual Baseada em Ligantes (LBVS), essa falta de transparência compromete a confiabilidade de predições usadas na descoberta de fármacos. Sendo assim, este trabalho teve como objetivo desenvolver um pipeline computacional explicável para LBVS aplicado ao alvo SOS1, integrando algoritmos de ML com técnicas de Inteligência Artificial Explicável (XAI). A metodologia envolveu a replicação de um estudo clássico baseado em ECFP, seguida da substituição dessas representações por 210 descritores fisico-químicos contínuos. Três modelos baseados em árvores (Decision Tree, Random Forest e LightGBM) foram treinados e avaliados, e suas decisões interpretadas por SHAP por meio de análises globais e locais. Os resultados mostram que Random Forest obteve o melhor desempenho entre os modelos ($R^2 = 0{,}89$). A aplicação de XAI permitiu identificar quais descritores moleculares afetam as previsões, mostrando de forma explícita como cada atributo contribui positiva ou negativamente para a saída do modelo. Portanto, conclui-se que é possível integrar interpretabilidade ao pipeline LBVS sem perda significativa de desempenho, oferecendo uma abordagem computacional mais transparente, auditável e adequada para apoiar decisões na descoberta de novos inibidores de SOS1.

Palavras-chave: Aprendizado de Máquina, Inteligência Artificial, Inteligência Artificial Explicável, Triagem Virtual.

Abstract

In recent years, Machine Learning (ML) algorithms have become increasingly accurate. However, this improvement in precision has been accompanied by a significant increase in complexity and the emergence of the "black box" phenomenon. Due to opaque and difficult to interpret internal mechanisms, understanding the rationale behind specific predictions has become challenging. In Computational Chemistry applications, such as Ligand-Based Virtual Screening (LBVS), this lack of transparency compromises the reliability of predictions used in drug discovery. Therefore, this work aims to develop an explainable computational pipeline for LBVS applied to the SOS1 target, integrating ML algorithms with Explainable Artificial Intelligence (XAI) techniques. The methodology involves replicating a classic study based on ECFP, followed by the substitution of these representations with 210 continuous physicochemical descriptors. Three tree-based models (Decision Tree, Random Forest, and LightGBM) were trained and evaluated, and their decisions were interpreted using SHAP through global and local analyses. Results indicate that the Random Forest model achieved the best performance ($R^2 = 0.89$). The application of XAI enabled the identification of molecular descriptors affecting predictions, explicitly demonstrating how each attribute contributes positively or negatively to the model output. Consequently, it is concluded that interpretability can be integrated into the LBVS pipeline without significant performance loss, offering a more transparent, auditable, and suitable computational approach to support decision-making in the discovery of new SOS1 inhibitors.

Keywords: Machine Learning, Artificial Intelligence, Explainable Artificial Intelligence, Virtual Screening.

Lista de ilustrações

Figura 1 – Curva ROC representando o desempenho de um classificador genérico. 20
Figura 2 – Exemplo de R^2 obtida no treinamento do modelo RF. 21
Figura 3 – Análise de resíduos do modelo RF. (Esquerda) Resíduos *vs* valores preditos mostrando distribuição homocedástica ao redor de zero. (Direita) Histograma dos resíduos apresentando distribuição aproximadamente normal, indicando ausência de viés sistemático. 41
Figura 4 – SHAP *Beeswarm Plot* mostrando o impacto dos 20 descritores mais importantes. Cada ponto representa uma molécula do conjunto de teste. A cor indica o valor do descritor (vermelho = alto, azul = baixo), e a posição horizontal representa o valor SHAP (impacto na predição). . . 42
Figura 5 – Ranking dos 20 descritores mais importantes baseado no valor SHAP médio absoluto. Descritores no topo da lista têm maior impacto médio nas predições do modelo. 43
Figura 6 – Mapa de calor de correlação entre os 15 descritores mais importantes. Valores próximos de +1 (vermelho) indicam forte correlação positiva, enquanto valores próximos de -1 (azul) indicam correlação negativa. Correlações moderadas a fortes sugerem redundância informacional ou potenciais interações sinérgicas. 44
Figura 7 – SHAP *Waterfall Plot* para uma amostra do conjunto de teste. O gráfico mostra como cada descritor contribui para desviar a predição do valor base (*baseline*) até o valor final predito. Barras vermelhas empurram a predição para cima (aumentam pChEMBL), enquanto barras azuis empurram para baixo. 45

-

Lista de tabelas

Tabela 1 – Ligantes originais com propriedades químicas 17
Tabela 2 – Amostra com *bootstrapping* dos ligantes 17
Tabela 3 – Comparação de desempenho entre *Decision Tree, Random Forest e LightGBM* após treinamento inicial 39
Tabela 4 – Comparação de desempenho entre *Decision Tree, Random Forest e LightGBM* após validação cruzada e *rand_search* 39
Tabela 5 – Comparação de desempenho do modelo *Random Forest* utilizando Descritores Moleculares *versus Fingerprints* 46
Tabela 6 – Correlação de Spearman entre Descritores e pChEMBL (Top 10) . . . 47

Lista de abreviaturas e siglas

AUC	Area Under the Curve (Área Sob a Curva)
DT	Decision Tree (Árvore de Decisão)
ECFP	Extended-Connectivity Fingerprints
LBVS	Ligand-Based Virtual Screening (Triagem Virtual Baseada em Ligantes)
LightGBM	Light Gradient Boosting Machine
LIME	Local Interpretable Model-agnostic Explanations
ML	Machine Learning (Aprendizado de Máquina)
RF	Random Forest (Floresta Aleatória)
ROC	Receiver Operating Characteristic
SAR	Structure-Activity Relationship (Relação Estrutura-Atividade)
SHAP	SHapley Additive exPlanations
SOS1	Son of Sevenless 1
VS	Virtual Screening (Triagem Virtual)
XAI	Explainable Artificial Intelligence (Inteligência Artificial Explicável)

Sumário

1 INTRODUÇÃO . . . 12
1.1 Problemática . . . 13
1.2 Hipótese . . . 13
1.3 Justificativa . . . 14
1.4 Objetivos . . . 14
1.4.1 Objetivos Específicos . . . 14
1.5 Estrutura do Trabalho . . . 15

2 REFERENCIAL TEÓRICO . . . 16
2.1 Termos importantes relacionados à *Machine Learning* . . . 16
2.2 Aprendizado de Máquina - ML . . . 17
2.3 Etapas do Processo de Aprendizado de Máquina . . . 18
2.3.1 Pré-processamento . . . 18
2.3.2 Divisão de Conjuntos de Treino, Validação e Teste . . . 18
2.3.3 Ajuste de Hiperparâmetros e *Randomized Search* . . . 18
2.3.4 Validação e Métricas de Avaliação . . . 19
2.3.5 Tipos de Aprendizado . . . 21
2.4 Inteligência Artificial Explicável - XAI . . . 22
2.5 LIME . . . 23
2.6 SHAP . . . 23
2.6.1 Fundamentos do SHAP . . . 24
2.6.2 SHAP TreeExplainer . . . 25
2.7 Descoberta de Fármacos . . . 25
2.8 Triagem Virtual . . . 26
2.9 Triagem Virtual Baseada em Ligantes (*Ligand-Based Virtual Screening*) . . . 27
2.10 Algoritmos de *Machine Learning* aplicados a VS . . . 27
2.10.1 *Random Forest* . . . 27
2.10.2 *LightGBM* . . . 27
2.11 Inteligência Artificial Explicável na VS . . . 28
2.12 SHAP - *SHapley Additive exPlanations* na VS . . . 28
2.13 *Fingerprints* Moleculares . . . 29
2.13.1 Geração e Tipos de *Fingerprints* Moleculares . . . 29
2.13.2 Geração de Descritores . . . 30

3 METODOLOGIA . . . 31

3.1 Replicação do Estudo de Duo et al. (2024) . . . 31
3.2 Extensões Metodológicas . . . 31
3.3 Aquisição e Preparação dos Dados . . . 32
3.3.1 Base de dados de Ligantes de SOS1 . . . 32
3.3.2 Representação Molecular: Descritores Fisico-Químicos . . . 32
3.3.2.1 Cálculo de Descritores . . . 33
3.3.2.2 Tratamento de Valores Inválidos . . . 33
3.3.3 Divisão dos Dados . . . 34
3.4 Treinamento e Otimização de Modelos . . . 34
3.4.1 Modelo base: *Decision Tree* . . . 34
3.4.2 *Random Forest* . . . 34
3.4.3 *LightGBM (Gradient Boosting)* . . . 35
3.4.4 Otimização de Hiperparâmetros . . . 35
3.4.5 Treinamento do Modelo Final . . . 36
3.4.6 Métricas de Avaliação . . . 36
3.5 Análise Interpretativa com SHAP . . . 36
3.5.1 Importância *Global de Features* . . . 36
3.5.2 Mapa de Correlação de Descritores . . . 36
3.5.3 *Waterfall Plot* (Predições Individuais) . . . 37
3.5.4 Análise de Resíduos . . . 37
3.6 Identificação de Propriedades Moleculares Críticas . . . 37
3.7 Ambiente Computacional . . . 38

4 RESULTADOS . . . 39
4.1 Comparação dos Modelos . . . 39
4.1.1 Desempenho Comparativo . . . 39
4.1.1.1 Análise Crítica do *Decision Tree* . . . 40
4.1.1.2 Superioridade dos Métodos *Ensemble* . . . 40
4.2 Análise Residual . . . 40
4.3 Interpretabilidade dos Modelos via SHAP . . . 41
4.3.1 Importância Global dos Descritores . . . 41
4.3.1.1 Gráfico *Beeswarm* . . . 41
4.3.1.2 Ranking de Importância Média . . . 42
4.3.2 Interações Entre Descritores . . . 43
4.4 Análise de Predições Individuais . . . 45
4.4.1 SHAP *Waterfall Plot* . . . 45
4.5 Comparação entre Descritores e *Fingerprints* . . . 46
4.5.1 Descritores Moleculares (Este Trabalho) . . . 46
4.5.2 Fingerprints Moleculares . . . 46
4.6 Correlação entre Descritores . . . 47

4.7 Síntese dos resultados 47

5 DISCUSSÃO 49

6 CONCLUSÃO 51

REFERÊNCIAS 53

1 Introdução

Com o passar do tempo, os algoritmos de Aprendizado de Máquina (Machine Learning - ML) foram ganhando cada vez mais precisão. Capazes de fornecerem respostas precisas e aplicadas a contextos relevantes, a utilização do ML se tornou imprescindível em diversas áreas. No entanto, esse avanço trouxe consigo um grande desafio. Embora os algoritmos apresentassem alta precisão, isso geralmente ocorria ao custo da interpretabilidade. Métodos como *ensembles*, *gradient boosting* e redes neurais profundas, por exemplo, atuam como grandes "caixas-pretas", dificultando auditoria, validação e compreensão de seus critérios de decisão (LUNDBERG; LEE, 2017; RIBEIRO; SINGH; GUESTRIN, 2016). Essa limitação tornou-se central em aplicações críticas que exigem transparência e rastreabilidade.

Nesse contexto, surgem as técnicas de Inteligência Artificial Explicável (Explainable AI — XAI), desenvolvidas para tornar modelos complexos mais compreensíveis. Entre elas, o SHAP (SHapley Additive exPlanations) destaca-se por fornecer explicações locais e globais baseadas em princípios da teoria dos jogos cooperativos, quantificando a contribuição individual de cada atributo para a predição (LUNDBERG; LEE, 2017). Essa capacidade permite interpretar modelos altamente não lineares sem comprometer seu desempenho.

Um domínio onde a falta de interpretabilidade se torna particularmente problemática é a descoberta de fármacos baseada em Inteligência Artificial. Nesse cenário, modelos de ML lidam com centenas ou milhares de descritores moleculares frequentemente correlacionados, o que aumenta o risco de aprendizado de padrões artificiais ou irrelevantes (JIMÉNEZ-LUNA; GRISONI; SCHNEIDER, 2020). Sem transparência, torna-se difícil verificar se o modelo realmente captura relações químicas relevantes ou apenas memoriza dados.

A urgência por métodos explicáveis é agravada pela própria crise de eficiência da descoberta moderna de fármacos: o desenvolvimento de um medicamento pode levar de 10 a 15 anos, custar cerca de U$ 2,6 bilhões e apresentar taxas de falha superiores a 90% nas fases clínicas (DIMASI; GRABOWSKI; HANSEN, 2016; WOUTERS; MCKEE; LUYTEN, 2020). Esse cenário impulsiona o uso de metodologias computacionais para acelerar a identificação de candidatos promissores. A Triagem Virtual (*Virtual Screening — VS*) surge como abordagem essencial, permitindo filtrar milhões de moléculas in silico com reduções significativas de custo e tempo (WALTERS; STAHL; MURCKO, 1998; LYU et al., 2019).

A integração de ML à VS ampliou esse potencial. Modelos treinados em descrito-

res moleculares e fingerprints — como *Random Forest e LightGBM* — alcançam métricas superiores a 90% na classificação de compostos ativos (WÓJCIKOWSKI; BALLESTER; SIEDLECKI, 2017; SHIMAZAKI; TACHIKAWA, 2022), possibilitando explorar espaços químicos massivos como o ZINC, com mais de 20 bilhões de moléculas (IRWIN; SHOICHET, 2005). Entretanto, esse ganho vem acompanhado de maior opacidade: sem interpretabilidade, os pesquisadores não conseguem validar se as decisões refletem padrões de Relação Estrutura–Atividade (SAR) ou artefatos da base de dados. Além disso, órgãos regulatórios como FDA e ANVISA passaram a exigir maior transparência em modelos aplicados a decisões biomédicas (Food and Drug Administration, 2021), reforçando a necessidade de explicações auditáveis.

Dentro desse cenário, a Triagem Virtual Baseada em Ligantes (LBVS) configura-se como um excelente estudo de caso para aplicação de XAI. Por trabalhar diretamente com descritores estruturais quantificáveis de ligantes conhecidos, a LBVS oferece um ambiente ideal para investigar como métodos explicáveis podem revelar quais atributos governam as predições de modelos de ML. Assim, este trabalho propõe replicar e estender um pipeline LBVS para o alvo oncológico SOS1, integrando técnicas de XAI para identificar e interpretar os descritores moleculares mais relevantes nas decisões do modelo.

1.1 Problemática

Apesar dos avanços em Inteligência Artificial Explicável, ainda não existe um consenso metodológico sobre como aplicar XAI de forma robusta e reprodutível em pipelines de Triagem Virtual Baseada em Ligantes (LBVS). O principal desafio não está apenas em gerar explicações, mas em garantir que elas sejam estáveis, coerentes com as representações moleculares e capazes de revelar falhas estruturais nos modelos, como dependência excessiva de atributos redundantes ou correlações não significativas.

Dessa forma, surge a questão central deste trabalho: como integrar técnicas de Inteligência Artificial Explicável (XAI) a pipelines de ML em LBVS, de modo a revelar os critérios de decisão internos do modelo, aumentar a confiabilidade das predições e permitir auditoria e validação independente?

1.2 Hipótese

H1: A aplicação de XAI ao pipeline LBVS-SOS1 permitirá identificar, de forma robusta, as *features* mais influentes na predição do modelo, demonstrando capacidade do SHAP de gerar explicações consistentes mesmo em cenários de alta dimensionalidade e multicolinearidade.

1.3 Justificativa

Embora os principais *benchmarks* para modelos de ML sejam baseados na precisão, essas métricas de avaliação deixam de fora contextos importantes para sustentar uma tomada de decisão. Tomar decisões de alto risco com base no que um modelo determinou, por exemplo, mesmo tratando-se de um modelo preciso, não é uma decisão prudente. Médicos, economistas, biólogos e engenheiros devem sempre ser capazes de compreender o porquê de uma decisão, invés de apenas confiar cegamente em uma ferramenta.

Para mitigar esse problema, a integração de XAI pode atuar como uma ferramenta de auditoria algorítmica, permitindo identificar atributos redundantes, dependências espúrias e possíveis manifestações de *overfitting* — problemas clássicos de ML que permanecem ocultos em modelos de alta *performance*. A capacidade de revelar o comportamento interno do modelo contribui para a construção de sistemas mais confiáveis, transparentes e reprodutíveis. Além disso, gerar explicações pode ajudar especialistas a reformular hipóteses e estratégias diante do problema.

Já na descoberta de fármacos, a aceleração da descoberta de inibidores de SOS1 — um regulador crítico da via RAS/MAPK — pode gerar impacto direto na disponibilidade de terapias mais eficazes e menos tóxicas. A transparência dos modelos aumenta a confiança pública e regulatória da IA aplicada à saúde.

Portanto, integrar XAI ao pipeline LBVS permitirá o desenvolvimento de modelos não apenas precisos, mas também parcimoniosos, onde a seleção de atributos será orientada por interpretabilidade química, contribuindo para o campo de *explainable drug discovery*.

1.4 Objetivos

O objetivo geral deste trabalho é: desenvolver e validar uma metodologia computacional que integre técnicas de Inteligência Artificial Explicável (XAI) a pipelines de *Machine Learning* aplicados à Triagem Virtual Baseada em Ligantes (LBVS), avaliando a interpretabilidade, a robustez e o comportamento interno dos modelos sem comprometer o desempenho preditivo.

1.4.1 Objetivos Específicos

- Implementar e reproduzir um pipeline LBVS como estudo de caso, utilizando algoritmos de ML (*Random Forest e LightGBM*) aplicados a representações estruturais de moléculas;

- Integrar técnicas de XAI, em especial SHAP, para identificar, ranquear e explicar as *features* mais influentes nas predições dos modelos;
- Produzir visualizações interpretáveis que permitam compreender o comportamento interno do modelo e apoiar a auditoria algorítmica;
- Validar se as explicações obtidas refletem padrões estruturais coerentes com o domínio químico, garantindo que o modelo esteja aprendendo relações generalizáveis e não artefatos da base de dados.

1.5 Estrutura do Trabalho

Este trabalho está organizado em seis capítulos: O Capítulo 2 apresenta o referencial teórico; o Capítulo 3 a metodologia; o Capítulo 4 os resultados obtidos; o Capítulo 5 a discussão; o Capítulo 6 a conclusão.

2 Referencial Teórico

Nesta seção, será apresentada uma revisão crítica dos estudos mais relevantes, cobrindo desde os princípios fundamentais de XAI até suas aplicações na Triagem Virtual Baseada em Ligantes. Além disso, serão explicados termos e conceitos importantes relacionados a ML e LBVS.

2.1 Termos importantes relacionados à *Machine Learning*

- ***Ensemble Learning:*** é uma estratégia de aprendizado de máquina que consiste em utilizar algoritmos de ML para produzir resultados preditivos e, então, combinar as predições dos modelos para obter um resultado preciso (DONG et al., 2020).

- ***Overfitting:*** ocorre quando um modelo de aprendizado de máquina se ajusta excessivamente aos dados de treinamento, capturando não apenas as tendências gerais, mas também o ruído e as flutuações aleatórias, resultando em baixa capacidade de generalização para novos exemplos (GOODFELLOW; BENGIO; COURVILLE, 2016). Em outras palavras, o modelo aprende padrões específicos demais, que não se repetem em dados não vistos, levando a um desempenho inflado em treinamento e degradado em validação ou teste. Algumas estratégias comuns para mitigar *overfitting* incluem dividir o conjunto de dados em k partições, treinando e validando o modelo em combinações distintas para estimar sua variância de forma mais confiável.

- ***Black Box ("Caixa-Preta")***: Embora eficientes, os processos internos de decisão dos modelos de ML não são facilmente interpretáveis por usuários humanos. Logo, não é possível entender com clareza quais as operações internas e os parâmetros que levaram a uma dada resposta. Modelos como redes neurais profundas e *ensembles* complexos (como o RF) são considerados *black boxes*. Ou seja, podem ser definidas por sistemas que escondem sua lógica interna do usuário (GUIDOTTI et al., 2018).

- ***Bootstrapping***: é uma técnica de amostragem que consiste em gerar vários conjuntos de dados artificiais a partir do conjunto original (*resampling*) , selecionando observações com reposição (HENDERSON, 2005). Cada amostra do *bootstrap* tem o mesmo tamanho do *dataset* original, mas por causa da reposição algumas amostras aparecem várias vezes e outras podem não aparecer. Para ilustrar o conceito de *boostrapping*, abaixo encontra-se um exemplo (Tabela 1) com dados de ligantes moleculares:

Tabela 1 – Ligantes originais com propriedades químicas

Ligante	**Peso Molecular (Da)**	**LogP**	**HBA**	**Atividade (classe)**
Lig1	312.4	2.1	5	Inativo
Lig2	285.3	3.0	3	Ativo
Lig3	298.7	1.8	6	Inativo
Lig4	330.1	4.2	2	Ativo

Fonte: o autor (dados simulados).

Abaixo, é exibida uma amostra de *bootstrapping* (com reposição) aplicada sobre o conjunto original de ligantes.

Tabela 2 – Amostra com *bootstrapping* dos ligantes

Ligante	**Peso Molecular (Da)**	**LogP**	**HBA**	**Atividade (classe)**
Lig4	330.1	4.2	2	Ativo
Lig1	312.4	2.1	5	Inativo
Lig2	285.3	3.0	3	Ativo
Lig4	330.1	4.2	2	Ativo

Fonte: o autor (gerado por *bootstrapping*).

Como é possível observar, determinados registros da tabela inicial foram selecionados mais de uma vez na amostra *bootstrap*. Essa repetição é gerada aleatoriamente, garantindo variação.

2.2 Aprendizado de Máquina - ML

O Aprendizado de Máquina, técnica que aprimora o desempenho de um sistema por meio de métodos computacionais capazes de aprender a partir da experiência (ZHOU, 2021), é um subcampo da Inteligência Artificial. Trata-se da utilização de métodos computacionais para identificar padrões em dados e construir modelos capazes de aprender e tomar decisões. Com suas amplas aplicações, modelos de ML se tornaram bastante populares nas mais diversas áreas (SHINDE; SHAH, 2018). Esse avanço é impulsionado, sobretudo, pelo crescimento exponencial da quantidade de dados gerados globalmente. Conforme destacado por Clissa, Lassnig e Rinaldi (2023), a produção mundial de dados atinge dimensões sem precedentes — apenas o CERN (Organização Europeia para a Pesquisa Nuclear), por exemplo, gerou em 2023 mais de 160 PB de dados. Isso reforça a necessidade de ferramentas capazes de processar e extrair informações úteis a partir desses dados.

2.3 Etapas do Processo de Aprendizado de Máquina

O processo de Aprendizado de Máquina é composto por um conjunto de etapas interdependentes que visam transformar dados brutos em conhecimento útil e modelos preditivos eficazes. Essas etapas envolvem desde a coleta, organização e pré-processamento dos dados até a escolha, treinamento e avaliação de algoritmos adequados ao problema em questão.

2.3.1 Pré-processamento

Entre as etapas fundamentais está a etapa do pré-processamento de dados. Essa etapa é crítica, pois a qualidade dos dados influencia diretamente o desempenho do modelo. Conforme apontado por Domingos (2012), o sucesso do aprendizado depende mais da representação dos dados do que da escolha do algoritmo em si.

2.3.2 Divisão de Conjuntos de Treino, Validação e Teste

Para garantir que o modelo aprenda de forma generalizável — e não apenas memorize os dados —, o conjunto original deve ser comumente dividido em subconjuntos de *treinamento*, *validação* e *teste*.

- **Treinamento:** utilizado para ajustar os parâmetros internos do modelo.
- **Validação:** serve para ajustar hiperparâmetros e evitar *overfitting*, fornecendo uma medida de desempenho em dados não vistos durante o treinamento.
- **Teste:** é reservado para a avaliação final do modelo, após a definição de todos os parâmetros, simulando sua aplicação em dados reais.

A prática mais comum é a divisão dos dados em proporções como 70% para treino, 15% para validação e 15% para teste, embora esses valores possam variar conforme o tamanho e a natureza do conjunto de dados (GOODFELLOW; BENGIO; COURVILLE, 2016). Alternativamente, técnicas como *k-fold cross-validation* são amplamente utilizadas para otimizar o uso dos dados disponíveis, reduzindo a variância da estimativa de desempenho.

2.3.3 Ajuste de Hiperparâmetros e *Randomized Search*

Após definir o modelo e os dados, é necessário ajustar seus *hiperparâmetros*, isto é, configurações externas ao processo de aprendizado que controlam seu comportamento

(como taxa de aprendizado, profundidade de árvore ou número de neurônios). Esses parâmetros não são aprendidos automaticamente e precisam ser otimizados para obter o melhor desempenho possível.

Uma das abordagens clássicas para essa tarefa é o *Grid Search*, que realiza uma busca exaustiva sobre todas as combinações possíveis de valores pré-definidos. Entretanto, esse método pode se tornar computacionalmente caro, especialmente quando há muitos hiperparâmetros (LIASHCHYNSKYI; LIASHCHYNSKYI, 2019).

Em contraste, o *Randomized Search* propõe a amostragem aleatória de combinações de hiperparâmetros dentro de distribuições especificadas, reduzindo o custo computacional e frequentemente encontrando resultados comparáveis ou até superiores (BERGSTRA; BENGIO, 2012). Essa técnica de amostragem é amplamente utilizada por equilibrar eficiência e desempenho, sendo especialmente útil em cenários com limitação de recursos computacionais ou com grande número de parâmetros.

2.3.4 Validação e Métricas de Avaliação

A etapa final consiste em avaliar o desempenho do modelo por meio de métricas adequadas ao tipo de problema. Em tarefas de classificação, métricas como *accuracy*, *precision*, *recall* e *F1-score* são comumente empregadas; já em regressão, utilizam-se o erro quadrático médio (MSE) e o coeficiente de determinação (R^2). A escolha da métrica correta é essencial para interpretar corretamente os resultados e compreender as limitações do modelo. Abaixo, encontram-se algumas métricas de avaliação importantes:

- ***Receiver Operating Characteristic*** (ROC): é uma curva que plota a taxa de verdadeiros positivos (TPR) contra a taxa de falsos positivos (FPR) à medida que se varia o limiar de decisão do classificador (HOO; CANDLISH; TEARE, 2017).

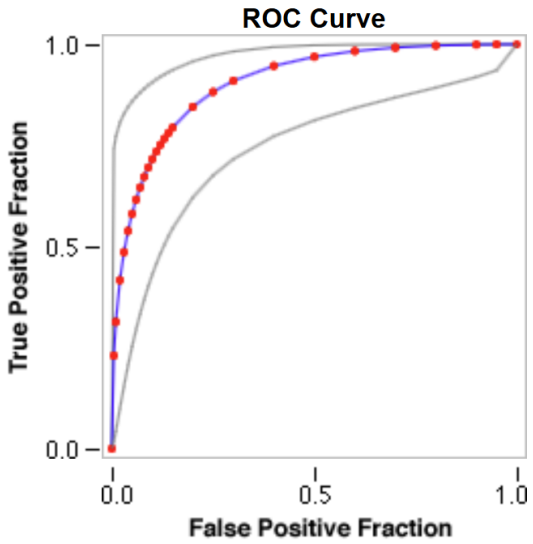


Figura 1 – Curva ROC representando o desempenho de um classificador genérico.

Fonte: o autor.

- ***Area Under the Curve*** (AUC): mede a área sob a curva ROC, fornecendo um valor entre 0,5 (classificação aleatória) e 1,0 (classificador perfeito). Quanto maior a AUC, melhor a capacidade do modelo de distinguir entre as classes positiva e negativa (HOO; CANDLISH; TEARE, 2017). No exemplo da Figura 1, o AUC tem o valor de 0,9055.

- ***Coeficiente de Determinação*** (R^2): é uma métrica que avalia o quanto o modelo explica a variabilidade dos dados observados em relação à média. Seu valor varia entre 0 e 1, onde valores mais próximos de 1 indicam que o modelo possui maior capacidade de prever corretamente as observações. Em outras palavras, o R^2 quantifica a proporção da variância dos dados que é explicada pelo modelo em relação à variância total (CHICCO; WARRENS; JURMAN, 2021).

$$R2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{2.1}$$

  Nessa equação, y_i representa os valores reais, $\hat{y}_i$ os valores previstos pelo modelo e $\bar{y}$ a média dos valores observados. Um R^2 próximo de 1 indica excelente ajuste do modelo aos dados, enquanto valores próximos de 0 sugerem baixo poder preditivo.

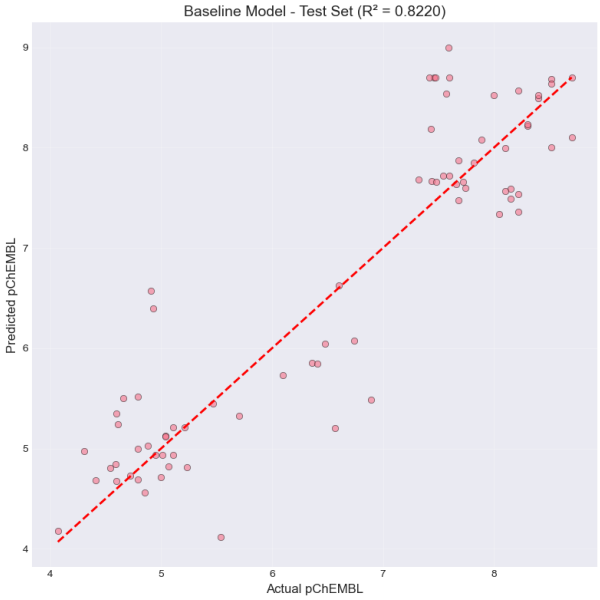


Figura 2 – Exemplo de R^2 obtida no treinamento do modelo RF.

Fonte: o autor.

- **RMSE (*Root Mean Squared Error*)**: Raiz quadrada da média dos erros ao quadrado. Sensível a *outliers* e na mesma unidade que a variável dependente (pChEMBL).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- **MAE (*Mean Absolute Error*)**: Média dos erros absolutos. Mais robusta a *outliers* que RMSE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Onde y_i são os valores reais, $\hat{y}_i$ as predições, $\bar{y}$ a média dos valores reais, e n o número de amostras.

2.3.5 Tipos de Aprendizado

Além das etapas de pré-processamento, os modelos também podem ser classificados em diferentes categorias de aprendizado:

- **Aprendizado Supervisionado**: o modelo é treinado com dados rotulados, onde cada exemplo de entrada está associado a uma saída conhecida. O objetivo é treinar uma função ($\mathbf{f} : \mathbf{X} \rightarrow \mathbf{Y}$) capaz de mapear entradas (X) e saídas (Y), identificando e generalizando padrões para prever resultados em dados não vistos (CUNNINGHAM; CORD; DELANY, 2008). Esse paradigma opera em duas fases:

  - Treinamento: o modelo ajusta seus parâmetros usando pares:
    $(\mathbf{x_1}, \mathbf{y_1}), (\mathbf{x_2}, \mathbf{y_2}), ..., (\mathbf{x_n}, \mathbf{y_n})$, e algoritmos como o *gradient descent* minimizam o erro entre as revisões e *labels* reais;
  - Inferência: o modelo aplica o mapeamento apreendido a novos dados de entrada, gerando novas previsões $\hat{\mathbf{y}}$.

  Além dessas duas fases, existem também duas categorias distintas de previsões.

  - Classificação: previsão de categorias discretas (ex: molécula inativa, ativa);
  - Regressão: previsão de valores contínuos (ex: energia de ligação proteína-ligante).

- **Aprendizado Não Supervisionado**: ao contrário do supervisionado, este modelo utiliza dados não rotulados, identificando padrões em dados sem orientação prévia. É ideal para:

  - Agrupamento (*clustering*): agrega dados em grupos homogêneos (ex: agrupar moléculas por similaridade química);
  - Redução de dimensionalidade: simplifica dados mantendo padrões essenciais.

  Em resumo, pode ser considerado como um método de encontrar padrões em dados (GHAHRAMANI, 2003).

- **Aprendizado por Reforço - *Reinforcement Learning (RL)***: diferentemente das técnicas anteriores, esse paradigma envolve um agente que, ao tomar decisões, recebe recompensas ou penalidades baseadas na eficácia de sua previsão. Conforme decisões são tomadas, o algoritmo se ajusta a fim de minimizar erros.

2.4 Inteligência Artificial Explicável - XAI

Atualmente, a alta acurácia de modelos de ML é frequentemente alcançada por modelos cada vez mais complexos, como redes neurais profundas ou modelos de *ensemble* (LUNDBERG; LEE, 2017). Portanto, para resolver o problema da complexidade e mitigar o problema da caixa-preta, técnicas de XAI foram projetadas para tornar as decisões de modelos de ML transparentes e interpretáveis para humanos. Dessa maneira, explicações locais (explicações do porquê uma decisão específica foi tomada) ou globais (explicações sobre o que o modelo avalia para a tomada de decisões) podem ser geradas para entender o processo de tomada de decisão dos algoritmos. Dentre as técnicas de XAI, destacam-se as técnicas abaixo:

2.5 LIME

Sendo proposto inicialmente por Ribeiro, Singh e Guestrin (2016), o LIME é um modelo genérico de XAI desenvolvido para tornar interpretáveis as predições de qualquer classificador caixa-preta. O LIME gera explicações locais para cada instância x ao perturbar seus atributos e amostras de acordo com sua proximidade. Na fórmula abaixo, é possível entender como o LIME funciona:

$$\xi(x) \;=\; \underset{g \in G}{\arg\min} \Big[L\Big(f, g, \pi_x\Big) \;+\; \Omega(g) \Big] \tag{2.2}$$

onde:

f Modelo "caixa-preta" original;

g Modelo explicável da classe **G** (por exemplo, regressão linear);

$\boldsymbol{\pi}_x$ Peso que reflete a proximidade de amostras perturbadas em torno de x;

$\mathbf{L}(\mathbf{f}, \mathbf{g}, \boldsymbol{\pi}_x)$ Função que mede a infidelidade de **g** em aproximar **f** localmente;

$\boldsymbol{\Omega}(\mathbf{g})$ Termo de regularização que penaliza a complexidade de **g**.

Além disso, os autores conduziram experimentos com usuários (leigos e especialistas) para avaliar se, com o suporte das explicações geradas pelo LIME, era possível escolher com maior confiança qual modelo generaliza melhor em cenários reais. No estudo, os resultados mostraram que em mais de 80% dos casos os participantes identificaram corretamente o classificador mais robusto com base nas explicações locais, comparado a menos de 50% quando não havia explicação disponível. Com isso, concluiu-se que o LIME não apenas aumenta a confiança do usuário nas predições, mas também auxilia na detecção de correlações.

Por fim, Ribeiro, Singh e Guestrin (2016) destacam que, apesar de sua eficácia, o método apresenta limitações em termos de custo computacional para modelos de alta dimensionalidade e de instabilidade em regiões pouco densas do espaço de entrada, sugerindo como trabalhos futuros a investigação de explicadores alternativos e otimizações no pré-processamento das perturbações.

2.6 SHAP

O SHAP (*SHapley Additive exPlanations*) trata-se de um *framework* unificado para interpretar a previsão de modelos. No estudo de Lundberg e Lee (2017), é proposta a perspectiva de que qualquer explicação de uma previsão de um modelo pode ser vista como um modelo em si, denominado "modelo de explicação". Sendo assim, enquanto o

LIME provê uma aproximação linear local para fornecer interpretabilidade, o SHAP utiliza uma abordagem fundamentada na teoria dos jogos cooperativos para fornecer valores de importância de cada atributo, levando a um resultado mais consistente e a explicações mais intuitivas.

2.6.1 Fundamentos do SHAP

A base teórica do método reside no cálculo do valor de Shapley, que distribui o "ganho"total da predição entre as *features* de forma justa. O valor SHAP ϕ_i para uma determinada *feature* i é calculado através da média ponderada das suas contribuições marginais em todas as combinações possíveis de *features* (LUNDBERG et al., 2019).

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right] \tag{2.3}$$

Onde:

- F é o conjunto de todas as *features* de entrada;
- S é um subconjunto de *features* que não inclui i;
- $f_S(x_S)$ representa a predição do modelo utilizando apenas o subconjunto de *features* S (na prática, muitas vezes aproximado pela esperança condicional $E[f(x)|x_S]$);
- O termo $[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)]$ representa a contribuição marginal da *feature* i, ou seja, a diferença na predição causada pela inclusão de i ao conjunto S;
- A fração $\frac{|S|!(|F|-|S|-1)!}{|F|!}$ atua como um fator de ponderação, considerando todas as permutações possíveis em que a *feature* i pode ser adicionada ao modelo.

Tratando-se de uma ferramenta robusta, o SHAP sustenta-se em 3 propriedades principais (LUNDBERG et al., 2019):

1. **Acurácia Local (*Local Accuracy*)**: Esta propriedade é similar ao conceito de "soma zero" ou aditividade. Ela garante que a explicação seja matematicamente precisa para aquela previsão específica. Ou seja, valores SHAP quantificam a contribuição de cada *feature* para a diferença entre a predição de uma amostra específica e o valor médio das predições (*baseline*). Para uma predição $f(x)$, a soma dos valores SHAP deve igualar a saída do modelo:

$$f(x) = \phi_0 + \sum_{j=1}^{M} \phi_j$$

Onde ϕ_0 é o valor base (média das predições do modelo no conjunto de dados) e M é o número de *features*.

2. **Ausência (*Missingness*)**: esta propriedade indica que características que não estão presentes na entrada original não influenciam o resultado. Se uma *feature* x_i tem valor zero ou está ausente, seu valor SHAP ϕ_i deve ser zero.

3. **Consistência (*Consistency*)**: caso um modelo seja alterado de forma que a contribuição marginal de uma *feature* aumente ou permaneça a mesma (independentemente das outras *features*), o valor SHAP dessa variável também deve aumentar ou permanecer inalterado. Isso evita contradições onde uma variável se torna mais importante para o modelo, mas recebe uma atribuição menor de importância.

2.6.2 SHAP TreeExplainer

Para modelos baseados em árvores (*Decision Tree, Random Forest, LightGBM*), utilizou-se o algoritmo `TreeExplainer`. Diferentemente de abordagens agnósticas ao modelo que dependem de amostragem (como o Kernel SHAP), o `TreeExplainer` explora a estrutura interna das árvores de decisão para calcular os valores de Shapley exatos.

A principal vantagem desta abordagem é a redução drástica da complexidade computacional. O cálculo exato dos valores de Shapley pela fórmula clássica possui complexidade exponencial $O(TLM2^M)$, onde T é o número de árvores, L o número máximo de folhas e M o número de *features*, o que torna a aplicação inviável para *datasets* com muitas variáveis. O algoritmo `TreeExplainer` reduz esse custo para um tempo polinomial $O(TLD^2)$, onde D é a profundidade máxima da árvore, viabilizando explicações exatas e consistentes em tempo hábil (LUNDBERG et al., 2019).

2.7 Descoberta de Fármacos

Nesse contexto, a descoberta de fármacos também tem produzido grandes quantidades de dados relacionados a ensaios clínicos e análises computacionais relacionadas a esses experimentos. O *Protein Data Bank* (PDB) (BERMAN et al., 2000), por exemplo, possui mais de 160 000 estruturas tridimensionais (FENG et al., 2020) de macromoléculas (proteínas, ácidos nucleicos) e tem sido uma fonte essencial para mineração de dados estruturais em larga escala. Esse grande volume de dados abre oportunidades significativas para o treinamento de modelos de aprendizado de máquina e *deep learning* voltados à previsão de estrutura, interação proteína-ligante e descoberta de novos alvos terapêuticos, aproveitando padrões que surgem apenas em escala.

Por outro lado, bases como ChEMBL (GAULTON et al., 2012) e ZINC (IRWIN; SHOICHET, 2005) ampliam ainda mais o ecossistema de dados para fármacos. O ChEMBL

é um repositório manualmente curado de moléculas bioativas com propriedades farmacológicas, fornecendo ligações entre compostos, alvos e atividades biológicas (ZDRAZIL et al., 2024). Já o ZINC é uma base que reúne compostos comerciais disponíveis para triagem virtual, o que permite exploração de dezenas de bilhões de moléculas para *docking*, geração de bibliotecas e modelagem de grande escala. Juntas, estas bases fornecem dados em grande volume, contendo diferentes modalidades de dados (estrutura 3D, bioatividade, compostos-ligantes), tornando-as extremamente importantes para o treinamento de modelos de IA multi-modais ou para pipelines de descoberta automatizada de fármacos. Com a expressiva quantidade de informações presentes nessas bases de dados, surge um dos principais métodos de descoberta de fármacos recentes: a Triagem Virtual - (*Virtual Screening* - VS).

Diante desse grande volume de dados estruturais e bioativos, surge a necessidade de métodos computacionais capazes de explorar eficientemente esse espaço químico. É nesse contexto que se insere a Triagem Virtual.

2.8 Triagem Virtual

Enquanto existem centenas de milhares de proteínas e, para cada proteína, centenas de milhares de potenciais ligantes, testar experimentalmente todas as combinações possíveis sempre foi um dos maiores desafios da química e da descoberta de fármacos. Conforme o estudo de Bohacek, McMartin e Guida (1996), estima-se que o número de pequenas moléculas orgânicas potenciais esteja na ordem de 10^{60}, um valor que ilustra a vastidão do espaço químico. Antes dos anos 2000, os químicos eram capazes de testar apenas um número limitado de moléculas por ano, tornando-se inviável testar experimentalmente toda essa diversidade molecular. Dessa maneira, surge a importância de selecionar com precisão quais proteínas e ligantes são mais promissores, para então testá-los em laboratório.

Assim, surge a Triagem Virtual (VS). Uma abordagem *in silico* que utiliza algoritmos computacionais para filtrar grandes bibliotecas de compostos, identificando ligantes promissores antes da fase de ensaios experimentais (WALTERS; STAHL; MURCKO, 1998). Ao ligar compostos a um determinado alvo molecular, determinados organismos (como agentes patogênicos) podem ser afetados e ter sua função inibida. Neste estudo, será empregada uma de suas principais vertentes: a Triagem Virtual Baseada em Ligantes (*Ligand-Based Virtual Screening*, LBVS), que se apoia em informações de compostos bioativos previamente conhecidos para inferir novas moléculas com potencial atividade biológica.

2.9 Triagem Virtual Baseada em Ligantes (*Ligand-Based Virtual Screening*)

Quando não há informaões acerca da estrutura molecular da proteína alvo — geralmente devido à ausência de experimentos para obter a estrutura molecular cristalizada —, a Triagem Virtual Baseada em Ligantes (LBVS) é utilizada. Como a estrutura do alvo é desconhecida, a LBVS utiliza apenas as informações dos ligantes de um determinado alvo molecular para então determinar sua atividade. Deste modo, pode-se encontrar novos ligantes ao avaliar a similaridade entre os ligantes já confirmados experimentalmente e moléculas candidatas(HAMZA; WEI; ZHAN, 2012).

2.10 Algoritmos de *Machine Learning* aplicados a VS

No contexto da descoberta de fármacos, algoritmos de aprendizado de máquina desempenham um papel central na identificação de padrões complexos entre características moleculares e atividades biológicas (MELVILLE; BURKE; HIRST, 2009). Dentre os principais modelos encontrados, destacam-se:

2.10.1 *Random Forest*

Entre esses algoritmos, o *Random Forest* têm obtido um desempenho competitivo (MELVILLE; BURKE; HIRST, 2009), sendo amplamente utilizado em estudos de *drug discovery* para prever afinidades entre ligantes e alvos a partir de descritores moleculares.

Baseado em um conjunto de árvores de decisão, o modelo combina previsões de múltiplas árvores independentes, reduzindo o risco de sobreajuste e aumentando a precisão geral das predições. Em aplicações de LBVS mais recentes (DUO et al., 2024), o *Random Forest* tem mostrado excelente desempenho em tarefas de classificação e regressão, permitindo diferenciar compostos ativos e inativos com base em dados estruturais e físico-químicos, além de fornecer medidas de importância de variáveis úteis para a interpretação dos resultados.

2.10.2 *LightGBM*

O *LightGBM* (KE et al., 2017), por sua vez, representa uma evolução dos métodos baseados em árvores, oferecendo maior eficiência computacional e escalabilidade — fatores essenciais quando se trabalha com bancos de dados massivos como o ChEMBL e o ZINC. Esse algoritmo utiliza uma estratégia de crescimento de árvores *leaf-wise*, priorizando os ramos que produzem maior ganho de informação. Isso resulta em modelos mais

rápidos e precisos, especialmente adequados para triagens virtuais que envolvem milhões de compostos.

Além disso, sua compatibilidade com otimização paralela e uso reduzido de memória o torna ideal para *pipelines* modernos de descoberta de fármacos assistidos por inteligência artificial. Dessa forma, tanto o *Random Forest* quanto o *LightGBM* consolidam-se como ferramentas poderosas no avanço da triagem virtual, contribuindo para a seleção eficiente de candidatos promissores a fármacos com base em dados de larga escala.

2.11 Inteligência Artificial Explicável na VS

Nesta seção, serão apresentadas as aplicações existentes do XAI no contexto da Triagem Virtual.

2.12 SHAP - *SHapley Additive exPlanations* na VS

Na pesquisa de Shimazaki e Tachikawa (2022), os autores propuseram uma estratégia colaborativa entre pontuações químicas simplificadas e técnicas de XAI para refinar a triagem virtual de inibidores da CDK2. A metodologia utilizou dados estruturais tridimensionais de complexos proteína-ligante para prever afinidades de ligação, enquanto modelos de *machine learning* foram empregados para classificar ligantes ativos com base em descritores de interação. O modelo *LightGBM* obteve a melhor *performance* (AUC = 0,93), seguido por SVM (AUC = 0,92) e Random Forest (AUC = 0,89). Técnicas de XAI (Permutation Importance e SHAP) identificaram resíduos-chave no sítio de ligação responsáveis pelas decisões do modelo. No entanto, o score do *docking* isolado mostrou baixa discriminação entre ativos e *decoys*. Além disso, a qualidade das explicações depende da representatividade das *features* e da precisão do *docking*. Validação experimental *in vitro* ainda é necessária para confirmar os achados.

No estudo de Jiménez-Luna, Grisoni e Schneider (2020), foi proposta uma metodologia baseada em redes neurais (chamada de DeltaDelta), voltada para a otimização de compostos em estágios iniciais da descoberta de fármacos. A metodologia consistia em incorporar XAI para permitir que os químicos medicinais entendessem as razões por trás das previsões do modelo, facilitando a interpretação e precisão dos dados. A aplicação das técnicas explicativas permitiu identificar alterações estruturais críticas em moléculas. Entretanto, algumas limitações foram encontradas:

> *Jiménez-Luna, Grisoni e Schneider (2020): Técnicas atuais sofrem de desafios técnicos, dada ao número grande de possíveis explicações e métodos aplicáveis para uma determinada tarefa. A maioria das abordagens não se apresenta como soluções prontas para uso, mas precisam ser adaptadas a cada aplicação individual. Além disso, um conhecimento profundo do domínio do problema é crucial para identificar quais decisões*

> *do modelo exigem explicações adicionais. Logo, a acessibilidade torna-se limitada para usuários menos familiarizados com o assunto.*

2.13 *Fingerprints* Moleculares

Para que um computador possa processar e comparar estruturas químicas em larga escala, é necessário traduzir a informação tridimensional de uma molécula em um formato computacionalmente eficiente. Os *fingerprints* moleculares (ou impressões digitais moleculares) cumprem exatamente essa função.

> *(TODESCHINI; CONSONNI, 2008) The molecular descriptor is the final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into an useful number or the result of some standardized experiment.*

Logo, podem tomar forma de representações abstratas de uma molécula, tipicamente na forma de um vetor binário (uma sequência de zeros e uns), que codificam suas características estruturais e físico-químicas. A premissa fundamental por trás do uso de *fingerprints* na triagem virtual baseada em ligantes (LBVS) é o Princípio da Similaridade Molecular: moléculas com estruturas semelhantes tendem a apresentar atividades biológicas semelhantes. Assim, ao converter as estruturas de ligantes conhecidos e de moléculas candidatas em *fingerprints*, pode-se quantificar essa similaridade e priorizar os candidatos mais promissores.

2.13.1 Geração e Tipos de *Fingerprints* Moleculares

A geração de um *fingerprint* envolve o mapeamento de diferentes características moleculares para posições específicas no vetor binário. Se uma determinada característica está presente na molécula, o bit correspondente no vetor é definido como "1"; caso contrário, é "0". Existem diversas maneiras de gerar esses *fingerprints*, que podem ser amplamente categorizados em:

- *Fingerprints* Circulares (ou *Extended-Connectivity Fingerprints* - ECFPs): Considerados um dos métodos mais eficazes, os ECFPs analisam o ambiente de cada átomo de forma radial. Para cada átomo, o algoritmo identifica as subestruturas vizinhas em raios crescentes (diâmetros de 2, 3, 4 etc.), gerando uma representação altamente detalhada e única da vizinhança de cada átomo (ROGERS; HAHN, 2010).

- *Fingerprints* Baseados em Subestruturas (Chaves Estruturais): Estes são os tipos mais intuitivos de *fingerprints*. Eles utilizam um dicionário predefinido de fragmentos moleculares ou subestruturas. O algoritmo então verifica a presença ou ausência de cada um desses fragmentos na molécula. Um exemplo clássico são as chaves

MACCS (*Molecular ACCess System*), que consistem em uma lista de 166 subestruturas químicas. Cada bit no *fingerprint* MACCS corresponde a uma dessas subestruturas.

2.13.2 Geração de Descritores

Além dos *fingerprints* moleculares, uma forma de obter informações a respeito das moléculas é utilizando o rdkit (LANDRUM et al., 2013) e calculando os descritores moleculares. No trabalho de König e Vellido (2024), foram utilizados descritores do RDkit para treinar um modelo de ML e então explicar quais descritores mais influenciavam as decisões locais e globais.

3 Metodologia

Este capítulo descreve a metodologia computacional adotada para o desenvolvimento, avaliação e interpretação de modelos de aprendizado de máquina aplicados ao problema de Triagem Virtual Baseada em Ligantes (LBVS). Como estudo de caso, a proteína SOS1 foi utilizada para validar o pipeline proposto.

3.1 Replicação do Estudo de Duo et al. (2024)

No estudo original de Duo et al. (2024), os autores utilizaram *Extended Connectivity Fingerprints* (ECFP) como representação molecular e algoritmo *Random Forest* como modelo preditivo. A metodologia consistiu em:

- **Representação molecular**: ECFP4 (raio 3), gerando *fingerprints* binários de 512 bits que codificam a presença/ausência de subestruturas moleculares;
- **Divisão dos dados**: Particionamento aleatório 80/20 para treino e teste;
- **Modelo base**: *Random Forest Regressor* para predição de valores de pChEMBL;
- **Otimização de hiperparâmetros**: Busca sistemática via validação cruzada para seleção dos melhores parâmetros;
- **Métricas de avaliação**: Coeficiente de determinação (R^2), raiz do erro quadrático médio (RMSE) e erro absoluto médio (MAE).

Esta fase de replicação foi implementada seguindo fielmente os protocolos descritos no artigo original, permitindo validar os resultados reportados e estabelecer uma linha de base para comparações subsequentes.

3.2 Extensões Metodológicas

Embora a replicação inicial tenha seguido o protocolo de Duo et al. (2024) utilizando ECFP4, o presente trabalho estende a análise original através de:

1. **Representação alternativa**: Substituição de ECFP por descritores fisico-químicos contínuos calculados pelo RDKit, oferecendo maior interpretabilidade sobre as representações e atributos utilizados pelo modelo;

2. **Comparação de algoritmos**: Aplicação de XAI aos três modelos — *Decision Tree, Random Forest e LightGBM* — permitindo análise comparativa entre diferentes arquiteturas baseadas em árvores.

3. **Interpretabilidade via SHAP**: Incorporação de análise SHAP (*SHapley Additive exPlanations*) para explicar as predições dos modelos e identificar os descritores mais relevantes, aspecto não abordado no estudo original.

4. **Visualizações avançadas**: Desenvolvimento de múltiplas visualizações (*beeswarm plots, waterfall plots*, mapas de calor de interação) para análise detalhada das contribuições de cada *feature* para a predição.

Esta abordagem permite validar os resultados obtidos em Duo et al. (2024) e compreender como diferentes tipos de representações influenciam a interpretabilidade e a estrutura interna do modelo, fornecendo informações relevantes para auditoria e validação computacional das decisões.

3.3 Aquisição e Preparação dos Dados

3.3.1 Base de dados de Ligantes de SOS1

O conjunto de dados utilizado foi obtido do banco de dados ChEMBL, focando em moléculas com atividade experimental reportada contra a proteína SOS1. O *dataset final* contém 375 estruturas químicas únicas, cada uma representada por:

- **Identificador ChEMBL**: Código único para rastreabilidade;
- **Estrutura *SMILES***: Representação linear da estrutura molecular;
- **Valor de atividade**: Medida experimental (IC_{50} ou K_i);
- **pChEMBL**: Valor de atividade convertido para escala logarítmica negativa: pChEMBL = $-\log_{10}(IC_{50}\ [M])$.

A conversão para pChEMBL normaliza a distribuição dos valores de atividade e facilita a modelagem.

3.3.2 Representação Molecular: Descritores Fisico-Químicos

Diferentemente de abordagens baseadas em *fingerprints* moleculares (que codificam presença/ausência de subestruturas), este trabalho utilizou descritores moleculares contínuos calculados pela biblioteca RDKit. Essa escolha foi motivada pela necessidade de

avaliar como diferentes representações estruturadas impactam a interpretabilidade produzida pelas técnicas de XAI.

3.3.2.1 Cálculo de Descritores

Um total de 210 descritores fisico-químicos foram calculados para cada molécula utilizando o módulo `MoleculeDescriptors` do RDKit. Os descritores incluem:

- **Propriedades moleculares básicas**: MolWt, HeavyAtomMolWt, ExactMolWt;
- **Índices topológicos**: Índices de conectividade de Kier-Hall (Chi0, Chi1, Chi2, Chi3, Chi4), índices Kappa (Kappa1, Kappa2, Kappa3), BalabanJ, BertzCT;
- **Propriedades eletrônicas**: Carga parcial máxima/mínima, polarizabilidade;
- **Propriedades de lipofilicidade**: LogP,SlogP_VSA;
- **Propriedades de superfície**: Área de superfície polar topológica (TPSA);
- **Doadores/aceptores de ligação de hidrogênio**: NumHDonors, NumHAcceptors;
- **Flexibilidade molecular**: Número de ligações rotacionáveis (NumRotatableBonds).

3.3.2.2 Tratamento de Valores Inválidos

Durante o cálculo de descritores, algumas moléculas podem gerar valores inválidos (NaN, infinito) devido a limitações computacionais ou características moleculares extremas. Portanto, um pipeline de limpeza foi implementado:

1. **Detecção individual**: Durante o cálculo valores infinitos são convertidos para NaN. Então, cada descritor é verificado quanto a NaN ou infinito, sendo substituído pela média dos descritores da coluna se inválido.
2. **Limitação (*clipping*) de valores extremos**: Descritores com valores absolutos superiores a 10^{10} são limitados a esse valor para prevenir *overflow* numérico.
3. **Conversão de tipo**: Todos os descritores são convertidos para float32 para eficiência computacional e compatibilidade com a biblioteca *scikit-learn*.

Esse processo de limpeza foi crítico para prevenir erros como `ValueError: Input X contains infinity or a value too large for dtype('float32')`, que ocorriam em versões iniciais do código.

3.3.3 Divisão dos Dados

O *dataset* foi dividido em conjuntos de treino (80%, 300 moléculas) e teste (20%, 75 moléculas) utilizando a função `train_test_split` do scikit-learn com `random_state=42` para garantir reprodutibilidade. Essa proporção é comum em problemas de regressão.

A estratificação não foi aplicada por se tratar de um problema de regressão (valores contínuos de *pChEMBL*), no qual não há categorias discretas a balancear, mas a distribuição de valores foi verificada para assegurar representatividade em ambos os conjuntos.

3.4 Treinamento e Otimização de Modelos

Três algoritmos de aprendizado de máquina baseados em árvores de decisão foram avaliados: Decision Tree (DT), *Random Forest* (RF) e LightGBM. Todos foram configurados para tarefas de regressão, predizendo valores contínuos de pChEMBL.

3.4.1 Modelo base: *Decision Tree*

Uma árvore de decisão simples foi treinada como modelo inicial para estabelecer uma referência de desempenho. Hiperparâmetros:

- `max_depth`: Profundidade máxima da árvore;
- `min_samples_split`: Número mínimo de amostras para dividir um nó;
- `min_samples_leaf`: Número mínimo de amostras em cada folha;
- `max_features`: Número máximo de *features* consideradas em cada divisão;
- `random_state=42`: Reprodutibilidade.

3.4.2 *Random Forest*

O RF combina múltiplas árvores de decisão através de *bootstrapping* e amostragem aleatória de *features*, reduzindo variância e *overfitting*. Hiperparâmetros otimizados:

- `n_estimators`: Número de árvores na floresta [100, 200, 300, 500];
- `max_depth`: Profundidade máxima de cada árvore [None, 10, 20, 30];
- `min_samples_split`: [2, 5, 10, 20];
- `min_samples_leaf`: [1, 2, 4];

- `max_features`: ['sqrt', 'log2', None];
- `bootstrap`: [True, False].

3.4.3 *LightGBM (Gradient Boosting)*

LightGBM é um *framework* de *gradient boosting* que constrói árvores de forma sequencial, onde cada nova árvore corrige erros das anteriores. Hiperparâmetros otimizados:

- `n_estimators`: Número de iterações de boosting [50, 100, 200, 300, 500];
- `max_depth`: Profundidade máxima [3, 5, 7, 10, 15, -1], onde -1 indica ausência de limite de profundidade;
- `learning_rate`: Taxa de aprendizado [0.01, 0.05, 0.1, 0.2];
- `num_leaves`: Número máximo de folhas [15, 31, 63, 127];
- `min_child_samples`: Mínimo de amostras para criar folha [5, 10, 20, 30, 50];
- `subsample`: Fração de amostras usadas por árvore [0.6, 0.8, 1.0];
- `colsample_bytree`: Fração de *features* usadas por árvore [0.6, 0.8, 1.0];
- `reg_alpha`: Regularização L1 [0, 0.1, 0.5, 1.0];
- `reg_lambda`: Regularização L2 [0, 0.1, 0.5, 1.0].

3.4.4 Otimização de Hiperparâmetros

Para cada modelo, foi realizada busca aleatória de hiperparâmetros utilizando `RandomizedSearchCV` do scikit-learn:

- **Estratégia de validação**: Validação cruzada repetida com 5 folds, repetida 3 vezes (*RepeatedKFold*);
- **Métrica de otimização**: Coeficiente de determinação (R^2);
- **Número de iterações**: 100 combinações aleatórias de hiperparâmetros;
- **Paralelização**: Uso de todos os núcleos de CPU disponíveis (`n_jobs=-1`).

A validação cruzada repetida aumenta a robustez da seleção de hiperparâmetros, reduzindo a probabilidade de *overfitting* à partição específica dos dados.

3.4.5 Treinamento do Modelo Final

Após identificar os melhores hiperparâmetros via validação cruzada, um modelo final foi treinado utilizando 100% dos dados disponíveis (conjunto completo de 375 moléculas). Essa prática, comum em problemas com *datasets* limitados, maximiza a informação disponível para o modelo que será usado em produção, aproveitando todo o conhecimento dos dados para melhorar a capacidade preditiva.

3.4.6 Métricas de Avaliação

Os modelos foram avaliados utilizando as métricas R^2 (Coeficiente de Determinação) e RMSE, que trata-se da raiz quadrada da média dos erros ao quadrado. Essa etapa de avaliação foi realizada conforme o estudo original com o intuito de comparar o resultado preditivo.

3.5 Análise Interpretativa com SHAP

Para compreender como os modelos estruturados em árvores tomam decisões e quais *features* exercem maior impacto nas predições, foi aplicada o framework SHAP, utilizando o algoritmo *TreeExplainer*.

3.5.1 Importância *Global de Features*

Duas visualizações foram geradas para análise global:

- ***Beeswarm Plot***: Mostra a distribuição dos valores SHAP para cada descritor. Cada ponto representa uma molécula, com a cor indicando o valor do descritor (vermelho = alto, azul = baixo) e a posição horizontal representando o impacto SHAP (positivo ou negativo na predição).
- ***Bar Plot***: Ranking dos descritores por importância média absoluta, facilitando identificação das propriedades mais influentes.

3.5.2 Mapa de Correlação de Descritores

Para investigar redundância e potenciais interações entre descritores, foi construído um mapa de calor (*heatmap*) de correlação entre os 15 descritores mais importantes. Correlações altas ($|\rho| > 0.7$) sugerem redundância informacional, enquanto correlações moderadas podem indicar interações sinérgicas.

3.5.3 *Waterfall Plot* (Predições Individuais)

Para exemplificar como o modelo faz predições individuais, foram gerados *waterfall plots* mostrando a contribuição passo a passo de cada descritor, representando como cada descritor afasta a predição do valor base até o valor final predito. Essa visualização é particularmente útil para:

- Validação de predições específicas;
- Identificação de moléculas com comportamentos atípicos;
- Comunicação de resultados a não-especialistas em ML.

3.5.4 Análise de Resíduos

A qualidade preditiva foi avaliada através de análise de resíduos (diferença entre valores reais e preditos):

- **Gráfico de dispersão**: Resíduos vs valores preditos, para detectar heterocedasticidade ou viés sistemático.
- **Histograma**: Distribuição dos resíduos, verificando normalidade e identificando *outliers*.

Resíduos distribuídos aleatoriamente ao redor de zero indicam que o modelo não apresenta viés sistemático.

3.6 Identificação de Propriedades Moleculares Críticas

A análise SHAP permitiu identificar quais *features* numéricas derivadas da representação molecular exercem maior influência nas predições dos modelos.

- **Descritores com SHAP positivo**: Indicam *features* cujo aumento tende a deslocar a predição para valores maiores de pChEMBL, refletindo impacto positivo na saída do modelo;
- **Descritores com SHAP negativo**: Indicam propriedades que, quando elevadas, diminuem o pChEMBL predito (menor atividade).

Essa informação é diretamente acionável para otimização estrutural de candidatos a fármacos: sabe-se quais propriedades aumentar e quais reduzir para maximizar atividade.

3.7 Ambiente Computacional

Todos os experimentos foram realizados no seguinte ambiente:

- **Sistema Operacional**: macOS;
- **Python**: Versão 3.8;
- **Principais bibliotecas**:
  - RDKit 2022.09.1: Cálculo de descritores moleculares;
  - scikit-learn 1.3.0: Implementação de RF, DT, e validação cruzada;
  - LightGBM 4.0.0: Implementação do gradient boosting;
  - SHAP 0.42.1: Cálculo de valores de Shapley e visualizações;
  - NumPy 1.24.3: Operações numéricas;
  - Pandas 2.0.3: Manipulação de dados tabulares;
  - Matplotlib 3.7.2 e Seaborn 0.12.2: Visualizações;

Todo o código foi desenvolvido em Jupyter Notebooks para facilitar análise exploratória e iteração rápida. A semente aleatória (`random_state=42`) foi fixada para garantir reprodutibilidade total dos resultados.

4 Resultados

Este capítulo apresenta os resultados obtidos com três modelos de aprendizado de máquina — *Decision Tree* (DT), *Random Forest* (RF) e *LightGBM* — utilizando 210 *features* numéricas derivadas de descritores moleculares calculados pelo RDKit. Essas representações fornecem um conjunto contínuo e estruturado de atributos que permitem avaliar comparativamente o comportamento preditivo e a interpretabilidade desses modelos.

4.1 Comparação dos Modelos

Três algoritmos de aprendizado de máquina foram avaliados para predição da atividade biológica (pChEMBL) dos ligantes de SOS1, todos utilizando o mesmo conjunto de 210 descritores moleculares calculados pela biblioteca RDKit.

4.1.1 Desempenho Comparativo

As Tabelas abaixo apresentam as métricas de desempenho dos três modelos avaliados, antes e depois da validação cruzada e *rand_search*:

Tabela 3 – Comparação de desempenho entre *Decision Tree, Random Forest e LightGBM* após treinamento inicial

Modelo	**Treino**		**Teste**	
	R^2	**RMSE**	**R^2**	**RMSE**
Decision Tree	0.9949	0.1125	0.7942	0.6719
Random Forest	0.9832	0.2048	0.8914	0.4881
LightGBM	0.9893	0.1637	0.8002	0.6620

Fonte: Elaborado pelo autor.

Tabela 4 – Comparação de desempenho entre *Decision Tree, Random Forest e LightGBM* após validação cruzada e *rand_search*

Modelo	**Treino**		**Teste**	
	R^2	**RMSE**	**R^2**	**RMSE**
Decision Tree	0.9665	0.2892	0.8218	0.6252
Random Forest	0.9922	0.1398	0.8918	0.4872
LightGBM	0.9947	0.1152	0.8714	0.5311

Fonte: Elaborado pelo autor.

4.1.1.1 Análise Crítica do *Decision Tree*

O modelo DT apresentou desempenho inferior aos *ensembles*, mas ainda assim obteve resultados razoáveis considerando sua simplicidade. Um dos principais problemas foi o *overfitting*, no qual o modelo obteve R^2 de 0.96 no treino e 0.82 no teste. Isso indica uma boa capacidade de generalização, mas ainda contém variância ainda elevada quando comparado aos *ensembles*.

Além disso, pequenas variações nos dados podem gerar árvores estruturalmente distintas, prejudicando a reprodutibilidade. Esses resultados confirmam a limitação bem conhecida de árvores de decisão simples em problemas de regressão com alta dimensionalidade, justificando o uso de métodos *ensemble*.

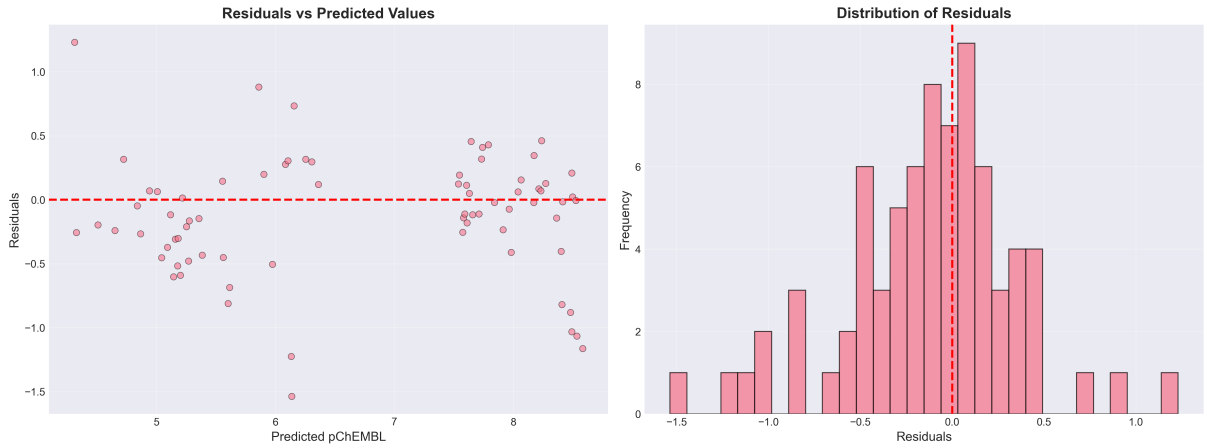
4.1.1.2 Superioridade dos Métodos *Ensemble*

Tanto RF quanto *LightGBM* apresentaram desempenho semelhante no conjunto de teste (R^2 de 0,89 e 0,87, respectivamente). A variação mínima nos resultados indica que ambos os métodos conseguem capturar bem as relações presentes nas *features* numéricas, com o RF obtendo um desempenho um pouco melhor. Ademais, a distância menor entre desempenho de treino e teste no RF (0.99 vs 0.89) quando comparado ao DT (0.96 vs 0.82) demonstra sua superior capacidade de generalização.

Em relação ao *LightGBM*, o desempenho se manteve próximo ao do RF. Tratando-se de modelos de maior complexidade, esse resultado era esperado e está de acordo com os resultados obtidos em Duo et al. (2024).

4.2 Análise Residual

A análise de resíduos permite avaliar a qualidade das predições e identificar possíveis vieses sistemáticos dos modelos. Apresentam-se aqui os resultados para o modelo RF, que demonstrou o melhor desempenho geral.



Fonte: Elaborado pelo autor.

Figura 3 – Análise de resíduos do modelo RF. (Esquerda) Resíduos *vs* valores preditos mostrando distribuição homocedástica ao redor de zero. (Direita) Histograma dos resíduos apresentando distribuição aproximadamente normal, indicando ausência de viés sistemático.

A distribuição dos resíduos centrada em zero e sem padrões aparentes no gráfico de dispersão indica que o modelo não apresenta viés sistemático. A distribuição aproximadamente normal dos resíduos sugere que as premissas estatísticas são razoavelmente atendidas.

4.3 Interpretabilidade dos Modelos via SHAP

Para compreender quais propriedades moleculares são mais relevantes para a predição da atividade biológica (pChEMBL), foi utilizado o SHAP. As análises apresentadas focam no modelo *Random Forest* devido ao seu desempenho superior.

4.3.1 Importância Global dos Descritores

4.3.1.1 Gráfico *Beeswarm*

O gráfico *beeswarm* mostra a distribuição dos valores SHAP para cada descritor molecular, revelando tanto a magnitude quanto a direção do impacto de cada propriedade nas predições.

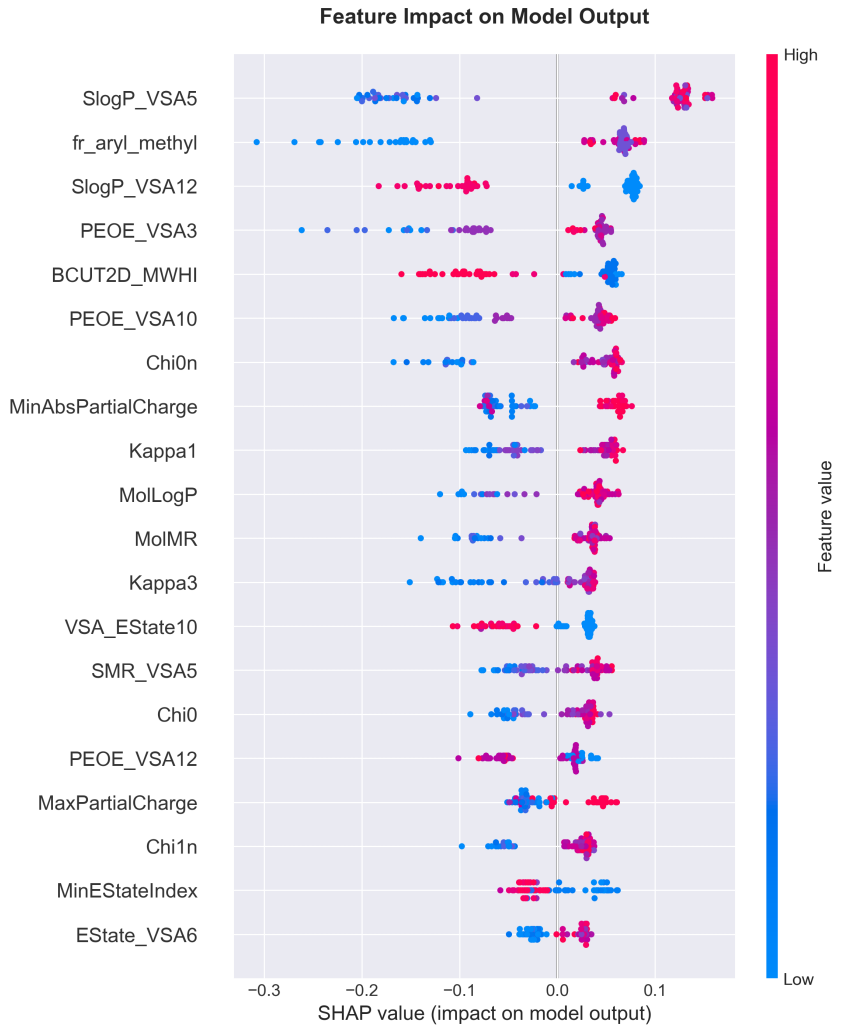


Figura 4 – SHAP *Beeswarm Plot* mostrando o impacto dos 20 descritores mais importantes. Cada ponto representa uma molécula do conjunto de teste. A cor indica o valor do descritor (vermelho = alto, azul = baixo), e a posição horizontal representa o valor SHAP (impacto na predição).

Fonte: o autor.

4.3.1.2 Ranking de Importância Média

O gráfico de barras apresenta a importância média absoluta de cada descritor, facilitando a identificação hierárquica das propriedades mais relevantes.

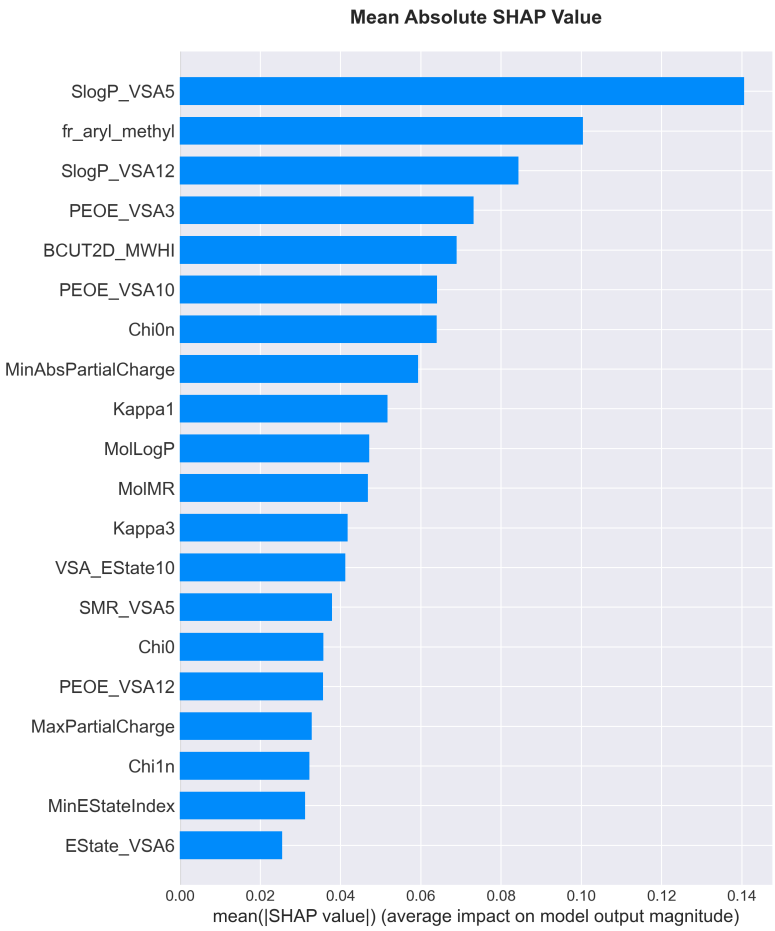
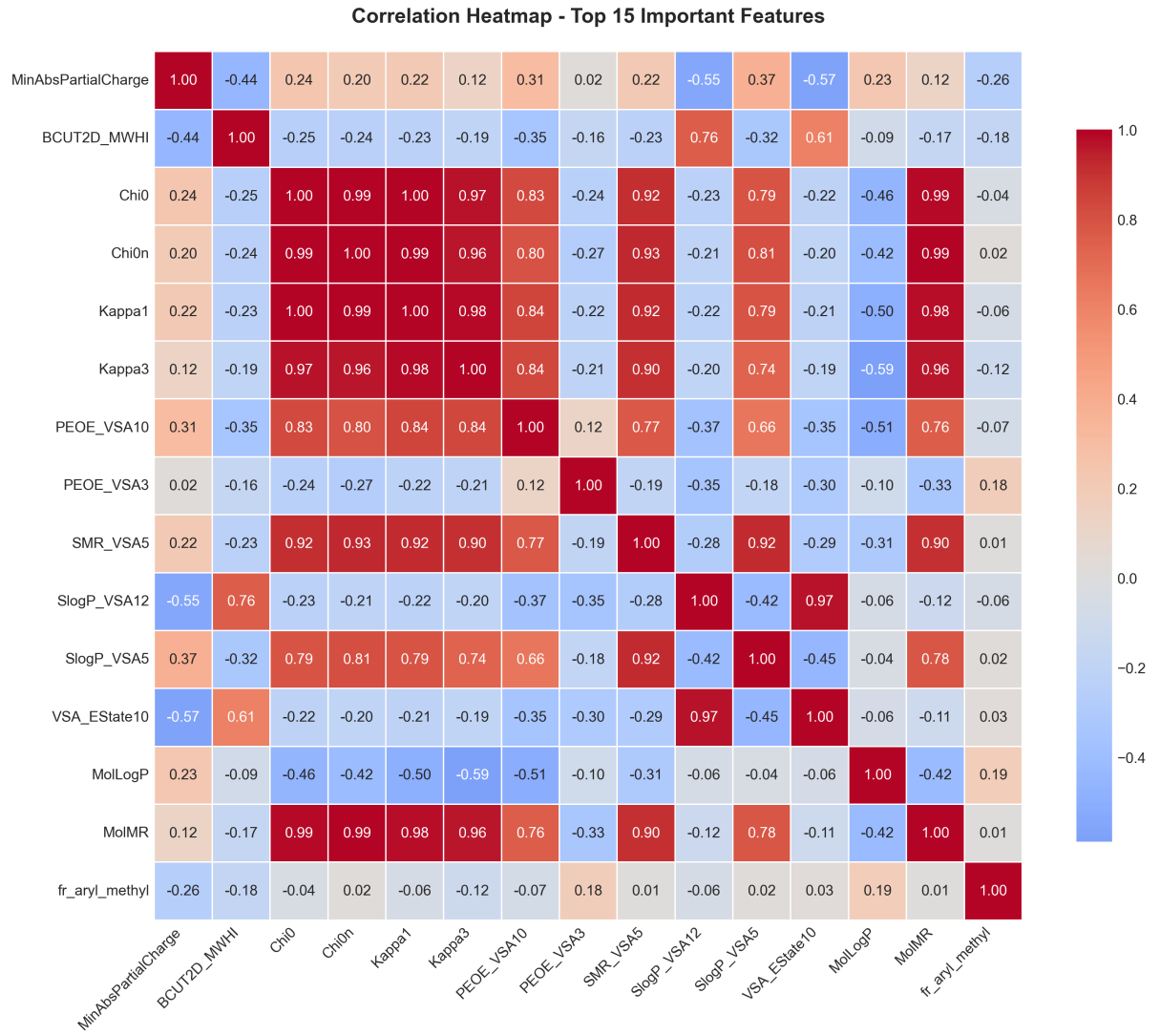


Figura 5 – Ranking dos 20 descritores mais importantes baseado no valor SHAP médio absoluto. Descritores no topo da lista têm maior impacto médio nas predições do modelo.

Fonte: o autor.

4.3.2 Interações Entre Descritores

Para investigar correlações e potenciais interações entre os descritores mais importantes, foi elaborado um mapa de calor de correlação.



Fonte: o autor.

Figura 6 – Mapa de calor de correlação entre os 15 descritores mais importantes. Valores próximos de +1 (vermelho) indicam forte correlação positiva, enquanto valores próximos de -1 (azul) indicam correlação negativa. Correlações moderadas a fortes sugerem redundância informacional ou potenciais interações sinérgicas.

As correlações observadas mostram que diversas *features* presentam redundância estrutural, algo esperado em representações derivadas de descritores calculados. Essa redundância deve ser considerada ao interpretar valores SHAP, pois *features* correlacionadas tendem a compartilhar importância.

4.4 Análise de Predições Individuais

4.4.1 SHAP *Waterfall Plot*

Para exemplificar como o modelo faz predições individuais, abaixo encontra-se o gráfico *waterfall* de uma amostra do conjunto de teste.

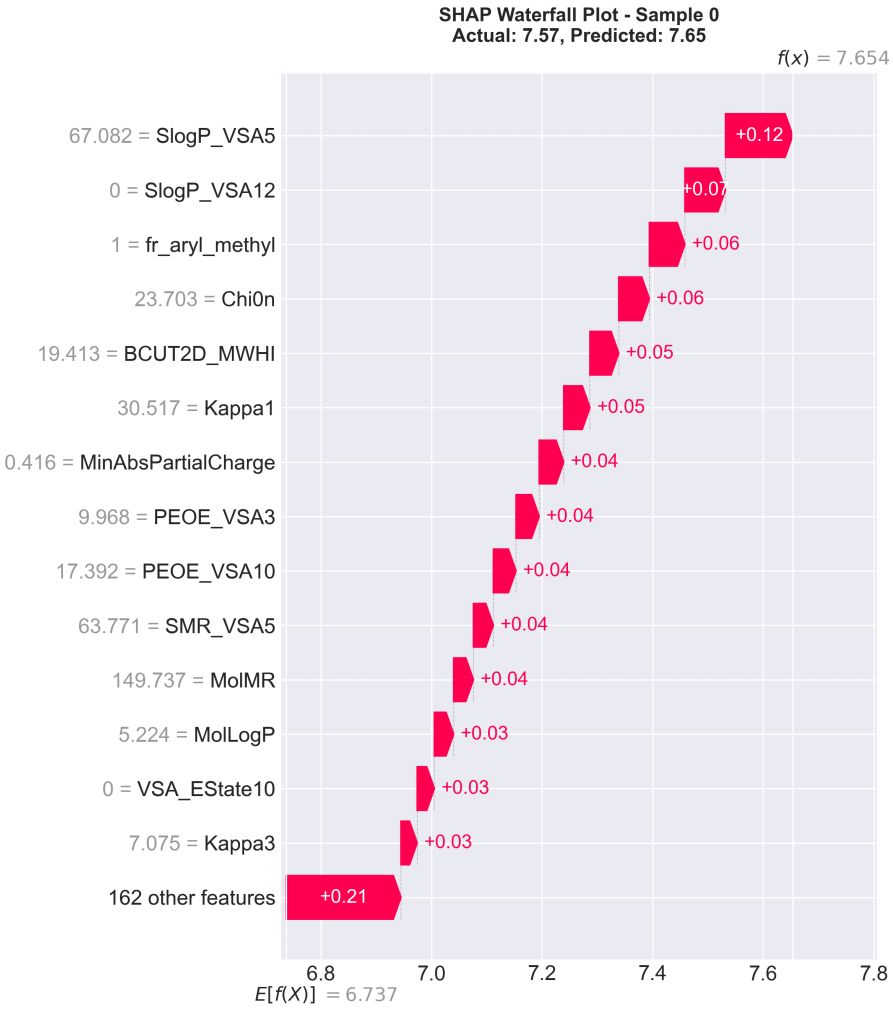


Figura 7 – SHAP *Waterfall Plot* para uma amostra do conjunto de teste. O gráfico mostra como cada descritor contribui para desviar a predição do valor base (*baseline*) até o valor final predito. Barras vermelhas empurram a predição para cima (aumentam pChEMBL), enquanto barras azuis empurram para baixo.

Fonte: o autor.

Este tipo de visualização permite rastrear o raciocínio do modelo para cada predição, identificando quais propriedades moleculares específicas contribuíram positiva ou negativamente para a predição da atividade biológica.

4.5 Comparação entre Descritores e *Fingerprints*

Os resultados apresentados neste trabalho permitem comparação entre duas filosofias distintas de representação molecular:

Tabela 5 – Comparação de desempenho do modelo *Random Forest* utilizando Descritores Moleculares *versus Fingerprints*

Abordagem	**Coeficiente de Determinação (R^2)**	
	Treino	**Teste**
Fingerprints (Morgan/ECFP)	0.9835	**0.9244**
Descritores (RDKit)	**0.9922**	0.8918

Fonte: o autor.

4.5.1 Descritores Moleculares (Este Trabalho)

A análise do SHAP Beeswarm Plot (Figura 4) permite identificar os atributos de maior influência nas predições do modelo. Entre as variáveis mais importantes, destacam-se os descritores topológicos (como Chi0 e Chi1N), indicando que a complexidade estrutural das moléculas desempenha um papel central na definição da atividade. Observa-se também uma forte correlação dos descritores eletrônicos (relacionados à distribuição de carga e polarizabilidade) com o output do modelo.

Além das características estruturais, propriedades físico-químicas mostraram-se determinantes. A lipofilicidade (expressa por descritores como SLogP_VSA5) apresenta um impacto significativo.

4.5.2 Fingerprints Moleculares

Os *fingerprints* moleculares, com destaque para algoritmos como ECFP e Morgan, são amplamente utilizados devido à sua capacidade de capturar a alta dimensionalidade do espaço químico, representando milhares de padrões estruturais simultaneamente. Essa abordagem permite identificar de forma explícita a presença ou ausência de motivos moleculares específicos, facilitando buscas por similaridade química.

Entretanto, essa metodologia apresenta limitações importantes. A principal delas é a interpretação indireta, visto que os bits individuais ativados raramente possuem um significado químico óbvio ou isolado para o pesquisador. Somado a isso, a natureza binária dessas representações impede a captura da intensidade ou quantidade de determinadas propriedades físico-químicas. Por fim, o processo de *hashing* pode resultar em colisões, onde subestruturas químicas distintas são mapeadas para o mesmo bit, introduzindo ruído na representação molecular. Portanto, não é possível obter uma interpretabilidade satisfatória utilizando-os.

4.6 Correlação entre Descritores

Para validar a interpretação complexa fornecida pelos modelos de aprendizado de máquina, foi realizada uma análise estatística direta para identificar quais propriedades físico-químicas apresentam relação monotônica com a atividade biológica (pChEMBL) no conjunto de dados original.

Para isso, utilizou-se o coeficiente de correlação de Spearman (ρ), que avalia relações não lineares (monotônicas) entre as variáveis, sendo mais adequado que a correlação de Pearson para dados biológicos e químicos que frequentemente não seguem uma distribuição estritamente linear.

Esta análise independe do modelo preditivo e serve para destacar quais descritores possuem, isoladamente, maior correlação em relação à variável alvo (*pChembl Value*).

Tabela 6 – Correlação de Spearman entre Descritores e pChEMBL (Top 10)

Descritor	ρ	**p-valor ajustado**	**Significativo**
SlogP_VSA5	0.7442	4.83e-66	Sim
MinAbsPartialCharge	0.6389	2.09e-43	Sim
MaxPartialCharge	0.6375	2.45e-43	Sim
PEOE_VSA14	0.6018	1.29e-37	Sim
VSA_EState10	-0.5971	5.43e-37	Sim
NumHAcceptors	0.5680	6.85e-33	Sim
Chi1v	0.5465	3.55e-30	Sim
VSA_EState9	0.5464	3.55e-30	Sim
TPSA	0.5028	4.45e-25	Sim
PEOE_VSA12	-0.4589	1.25e-20	Sim

Fonte: Elaborado pelo autor.

Comparando estes resultados com a análise de importância de *features* via SHAP, é possível notar uma forte convergência entre as variáveis que o modelo *Random Forest* identificou como mais impactantes para a predição e as variáveis que apresentam alta correlação de Spearman com a atividade biológica.

4.7 Síntese dos resultados

Entre os modelos avaliados, o RF obteve ($R^2 = 0.8918$), seguido pelo *LightGBM* ($R^2 = 0.8714$) e, por último, pelas DT ($R^2 = 0.8218$). Esses resultados reforçam a importância de métodos ensemble na modelagem de dados moleculares de alta dimensionalidade, uma vez que tanto a agregação de múltiplas árvores (no caso do RF) quanto o *boosting sequencial* (no *LightGBM*) contribuíram para mitigar problemas de *overfitting* comuns em modelos baseados em árvores individuais.

Apesar da maior complexidade, os modelos *ensemble* preservaram bons níveis de interpretabilidade. Em particular, o uso de SHAP permitiu identificar descritores críticos para as previsões, como índices topológicos, LogP e TPSA, destacando o papel central que tais propriedades desempenham na determinação da atividade molecular.

Do ponto de vista das implicações para a descoberta de fármacos, os resultados indicam que o RF oferece o melhor equilíbrio entre desempenho preditivo e estabilidade, sendo, portanto, a escolha mais apropriada para este conjunto de dados. Além disso, ficou evidente que descritores moleculares podem orientar a otimização racional de candidatos, permitindo ajustes estruturais guiados por propriedades específicas. Nesse contexto, a validação experimental deve priorizar moléculas que apresentaram valores favoráveis nos descritores mais relevantes identificados pelo modelo. Também se destaca o potencial de abordagens híbridas que combinem descritores e fingerprints, o que pode levar a melhorias adicionais na *performance* preditiva.

Diante dos resultados obtidos, recomenda-se como próximos passos a exploração de modelos híbridos que integrem descritores e *fingerprints*, bem como a expansão contínua do conjunto de dados com novos ligantes de SOS1 para aumentar a robustez dos modelos. Ademais, a metodologia aqui apresentada pode ser estendida para outros alvos terapêuticos, ampliando seu potencial de aplicação em diferentes contextos de descoberta de fármacos.

5 Discussão

Neste trabalho, o algoritmo *Random Forest* (RF) destacou-se com o melhor desempenho preditivo ($R^2 = 0.8918$), superando o *LightGBM* e a *Decision Tree*. A superioridade do RF observada aqui deste algoritmo para o alvo em questão está de acordo com o estudo original de Duo et al. (2024), que também indicaram o *Random Forest* como o modelo mais robusto ($R^2 = 0.918$) para a triagem de inibidores de SOS1. No entanto, a opção por focar a análise nos 210 descritores moleculares contínuos (RDKit) — apesar de uma leve redução na acurácia — permitiu obter explicações diretas das predições para propriedades físico-químicas tangíveis (como lipofilicidade e área de superfície polar).

A escolha entre descritores moleculares e *fingerprints* remete ao debate central sobre a "caixa-preta"na inteligência artificial. Jiménez-Luna, Grisoni e Schneider (2020) ressaltam que, no contexto da química medicinal, frequentemente se tolera uma pequena perda de acurácia em troca de modelos que se alinhem melhor à intuição e às práticas da área. Essa transparência é fundamental para a aceitação do modelo por químicos medicinais, pois é capaz de fornecer predições precisas e justificativas para corroborá-las. Além disso, a XAI é essencial para evitar que o modelo aprenda correlações irrelevantes (JIMÉNEZ-LUNA; GRISONI; SCHNEIDER, 2020).

Conforme discutem Shimazaki e Tachikawa (2022), algoritmos de aprendizado de máquina, por natureza, focam na identificação de correlações estatísticas nos dados, as quais nem sempre correspondem a relações causais ou a princípios fundamentais do domínio. Nesse contexto, a aplicação de técnicas de explicabilidade é crucial para validar se os padrões não lineares capturados pelo modelo representam significados reais extraídos da distribuição dos dados, mitigando o risco do algoritmo basear suas decisões em artefatos estatísticos ou ruído.

Crucialmente, a integração do SHAP validou a premissa de que é possível mitigar a opacidade inerente aos modelos *ensemble*. Esse achado está de acordo com a literatura, a qual destaca que modelos de alta complexidade apresentam interpretabilidade mínima por natureza, dependendo da aplicação de ferramentas de explicação *a posteriori* para fornecer *insights* úteis sobre seu processo de decisão (CENTOFANTI; NEGRI, 2022). A análise confirma que a utilização de tais técnicas de interpretabilidade externa é indispensável para conferir confiabilidade a preditores de caixa-preta, permitindo a validação da lógica interna do algoritmo.

Apesar da robustez do método, o uso de descritores 2D simplifica a complexidade espacial das moléculas. Shimazaki e Tachikawa (2022) indicam que representações baseadas em estruturas 3D capturam detalhes essenciais que escapam aos descritores globais.

Assim, a limitação principal não está no algoritmo de aprendizado, mas na simplificação dos dados de entrada.

6 Conclusão

A metodologia proposta, os resultados obtidos e as conclusões apresentadas neste trabalho permitem avaliar a robustez técnica do pipeline desenvolvido e também a relevância científica das contribuições introduzidas.

A principal inovação metodológica reside na substituição das representações binárias do tipo ECFP4 por descritores contínuos derivados do RDKit, possibilitando aplicar técnicas de interpretabilidade de maneira integrada ao pipeline de LBVS, e não como uma etapa isolada. Essa escolha foi determinante para viabilizar análises detalhadas do comportamento interno dos modelos, algo impraticável com *fingerprints* binários.

Além disso, o trabalho introduziu uma etapa de auditoria de redundância entre descritores, utilizando mapas de correlação combinados com valores SHAP. Esse tipo de auditoria raramente é explorado em LBVS e representa uma camada adicional de confiabilidade sobre a estrutura interna do modelo. Nesse sentido, a estratégia adotada demonstra a consistência da metodologia SHAP ao apresentar resultados alinhados com as correlações observadas. Os resultados mostraram desempenho consistente entre os modelos *ensemble (Random Forest e LightGBM)*, com boa generalização e distribuição de resíduos. No entanto, o caráter mais significativo dos resultados vai além das métricas, pois envolve a interpretação dos modelos.

O presente trabalho identificou quais descritores físico-químicos governam a predição de atividade contra SOS1. Logo, foram identificadas propriedades específicas — como lipofilicidade, área de superfície polar, índices topológicos e cargas parciais — que explicam o aprendizado dos modelos. Tal análise expande o escopo da literatura atual ao detalhar explicitamente as propriedades moleculares que influenciam as predições, preenchendo uma lacuna importante na interpretação de modelos computacionais para este alvo.

Os resultados demonstraram que é possível construir um pipeline LBVS totalmente interpretável sem perda substancial de acurácia. Mesmo com a substituição de representações binárias por descritores contínuos e a incorporação de múltiplas camadas de explicação via SHAP, os modelos mantiveram desempenho competitivo (R^2 próximo a 0.89). Ao mostrar empiricamente que ambos podem coexistir, o trabalho contribui metodologicamente para a consolidação de pipelines explicáveis na descoberta de fármacos.

A análise conjunta entre valores SHAP e mapa de correlação permitiu identificar redundâncias estruturais entre descritores importantes, como clusters de índices topológicos, correlação entre peso molecular e flexibilidade conformacional, e anticorrelações entre descritores eletrônicos e topológicos simples. Assim, o trabalho oferece uma camada de

explicabilidade adicional, fortalecendo a confiabilidade das decisões do modelo.

As conclusões apresentadas são suportadas de forma consistente pelas análises quantitativas e qualitativas. Os modelos *ensemble* mostraram-se superiores aos de árvore única, e a escolha do RF como modelo final está alinhada com os resultados de generalização e estabilidade observados. A relevância dos descritores identificados está de acordo com princípios conhecidos de SAR, reforçando a validade biológica das explicações fornecidas pelo SHAP. Além disso, os resultados sugerem que a integração da interpretabilidade em fluxos de LBVS é viável e desejável. Embora modelos baseados em descritores possam apresentar ligeiras variações de desempenho em relação a *fingerprints* de alta dimensionalidade, o ganho em transparência e auditabilidade é capaz de compensar essa diferença, oferecendo uma ferramenta mais robusta para a tomada de decisão racional.

Portanto, os resultados demonstram que a abordagem adotada expande o escopo analítico das técnicas de LBVS aplicadas à SOS1. Além de corroborar o desempenho preditivo observado no estado da arte, o estudo introduz avanços interpretativos que enriquecem a compreensão dos determinantes moleculares da atividade, servindo de referencial para o desenvolvimento contínuo de inibidores.

Referências

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. *The journal of machine learning research*, JMLR. org, v. 13, n. 1, p. 281–305, 2012.

BERMAN, H. M. et al. The protein data bank. *Nucleic acids research*, Oxford University Press, v. 28, n. 1, p. 235–242, 2000.

BOHACEK, R. S.; MCMARTIN, C.; GUIDA, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Medicinal research reviews*, Wiley Online Library, v. 16, n. 1, p. 3–50, 1996.

CENTOFANTI, T.; NEGRI, F. Exploring the trade-offs in explainable ai: Accuracy vs interpretability. *Annals of Applied Sciences*, v. 3, n. 1, 2022.

CHICCO, D.; WARRENS, M. J.; JURMAN, G. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *Peerj computer science*, PeerJ Inc., v. 7, p. e623, 2021.

CLISSA, L.; LASSNIG, M.; RINALDI, L. How big is big data? a comprehensive survey of data production, storage, and streaming in science and industry. *Frontiers in big Data*, Frontiers Media SA, v. 6, p. 1271639, 2023.

CUNNINGHAM, P.; CORD, M.; DELANY, S. J. Supervised learning. In: *Machine learning techniques for multimedia: case studies on organization and retrieval*. [S.l.]: Springer, 2008. p. 21–49.

DIMASI, J. A.; GRABOWSKI, H. G.; HANSEN, R. W. Innovation in the pharmaceutical industry: new estimates of r&d costs. *Journal of health economics*, Elsevier, v. 47, p. 20–33, 2016.

DOMINGOS, P. A few useful things to know about machine learning. *Communications of the ACM*, ACM New York, NY, USA, v. 55, n. 10, p. 78–87, 2012.

DONG, X. et al. A survey on ensemble learning. *Frontiers of Computer Science*, Springer, v. 14, p. 241–258, 2020.

DUO, L. et al. Discovery of novel sos1 inhibitors using machine learning. *RSC Medicinal Chemistry*, Royal Society of Chemistry, v. 15, n. 4, p. 1392–1403, 2024.

FENG, Z. et al. Impact of the protein data bank across scientific disciplines. *Data science journal*, v. 19, p. 25–25, 2020.

Food and Drug Administration. Based software as a medical device (samd) action plan. *Food and Drug Administration*, p. 2021–06, 2021.

GAULTON, A. et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, Oxford University Press, v. 40, n. D1, p. D1100–D1107, 2012.

GHAHRAMANI, Z. Unsupervised learning. In: *Summer school on machine learning*. [S.l.]: Springer, 2003. p. 72–112.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. http://www.deeplearningbook.org.

GUIDOTTI, R. et al. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, ACM New York, NY, USA, v. 51, n. 5, p. 1–42, 2018.

HAMZA, A.; WEI, N.-N.; ZHAN, C.-G. Ligand-based virtual screening approach using a new scoring function. *Journal of Chemical Information and Modeling*, v. 52, n. 4, p. 963–974, 2012. PMID: 22486340. Disponível em: https://doi.org/10.1021/ci200617d.

HENDERSON, A. R. The bootstrap: a technique for data-driven statistics. using computer-intensive analyses to explore experimental data. *Clinica chimica acta*, Elsevier, v. 359, n. 1-2, p. 1–26, 2005.

HOO, Z. H.; CANDLISH, J.; TEARE, D. *What is an ROC curve?* [S.l.]: BMJ Publishing Group Ltd and the British Association for Accident, 2017. 357–359 p.

IRWIN, J. J.; SHOICHET, B. K. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, ACS Publications, v. 45, n. 1, p. 177–182, 2005.

JIMÉNEZ-LUNA, J.; GRISONI, F.; SCHNEIDER, G. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, Nature Publishing Group UK London, v. 2, n. 10, p. 573–584, 2020.

KE, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, v. 30, 2017.

KÖNIG, C.; VELLIDO, A. Understanding predictions of drug profiles using explainable machine learning models. *BioData Mining*, Springer, v. 17, n. 1, p. 25, 2024.

LANDRUM, G. et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, v. 8, n. 31.10, p. 5281, 2013.

LIASHCHYNSKYI, P.; LIASHCHYNSKYI, P. Grid search, random search, genetic algorithm: a big comparison for nas. *arXiv preprint arXiv:1912.06059*, 2019.

LUNDBERG, S. M. et al. Explainable ai for trees: From local explanations to global understanding. *arXiv preprint arXiv:1905.04610*, 2019.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, v. 30, 2017.

LYU, J. et al. Ultra-large library docking for discovering new chemotypes. *Nature*, Nature Publishing Group UK London, v. 566, n. 7743, p. 224–229, 2019.

MELVILLE, J. L.; BURKE, E. K.; HIRST, J. D. Machine learning in virtual screening. *Combinatorial chemistry & high throughput screening*, Bentham Science Publishers, v. 12, n. 4, p. 332–343, 2009.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?": Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 11351144. ISBN 9781450342322. Disponível em: https://doi.org/10.1145/2939672.2939778.

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

SHIMAZAKI, T.; TACHIKAWA, M. Collaborative approach between explainable artificial intelligence and simplified chemical interactions to explore active ligands for cyclin-dependent kinase 2. *ACS Omega*, v. 7, n. 12, p. 10372–10381, 2022. Disponível em: https://doi.org/10.1021/acsomega.1c06976.

SHINDE, P. P.; SHAH, S. A review of machine learning and deep learning applications. In: IEEE. *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*. [S.l.], 2018. p. 1–6.

TODESCHINI, R.; CONSONNI, V. *Handbook of molecular descriptors*. [S.l.]: John Wiley & Sons, 2008.

WALTERS, W. P.; STAHL, M. T.; MURCKO, M. A. Virtual screening—an overview. *Drug discovery today*, Elsevier, v. 3, n. 4, p. 160–178, 1998.

WÓJCIKOWSKI, M.; BALLESTER, P. J.; SIEDLECKI, P. Performance of machine-learning scoring functions in structure-based virtual screening. *Scientific Reports*, Nature Publishing Group UK London, v. 7, n. 1, p. 46710, 2017.

WOUTERS, O. J.; MCKEE, M.; LUYTEN, J. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *Jama*, American Medical Association, v. 323, n. 9, p. 844–853, 2020.

ZDRAZIL, B. et al. The chembl database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic acids research*, Oxford University Press, v. 52, n. D1, p. D1180–D1192, 2024.

ZHOU, Z.-H. *Machine learning*. Singapore: Springer, 2021. ISBN 9789811519666.