UNIVERSIDADE ESTADUAL PAULISTA – UNESP
Instituto de Biociências, Letras e Ciências Exatas - Campus de São José do Rio Preto

RODRIGO PEREZ VERNA

METODOLOGIA DE REDUÇÃO DIMENSIONAL HIERÁRQUICA COM PREVENÇÃO DE VAZAMENTO DE DADOS PARA PREDIÇÃO DE HOTSPOTS EM INTERAÇÕES PROTEÍNA-PROTEÍNA

São José do Rio Preto
2026

Rodrigo Perez Verna

Metodologia de Redução Dimensional Hierárquica com Prevenção de Vazamento de Dados para Predição de Hotspots em Interações Proteína-Proteína

Dissertação, apresentada à Universidade Estadual Paulista (UNESP), Instituto de Biociências, Letras e Ciências Exatas, São José do Rio Preto/SP, para obtenção do título de Mestre em Ciências da Computação.

Área de Concentração: Inteligência Computacional

Orientador: Profº Dr. Geraldo Francisco Donegá Zafalon

São José do Rio Preto/SP
2026

V529m

Verna, Rodrigo Perez

Metodologia de redução dimensional hierárquica com prevenção de vazamento de dados para predição de hotspots em interações proteína-proteína / Rodrigo Perez Verna. -- São José do Rio Preto, 2026

72 p. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (UNESP), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto

Orientador: Geraldo Francisco Donegá Zafalon

1. Ciência da computação. 2. Bioinformática. 3. Proteínas. 4. Teoria da aprendizagem computacional. 5. Reconhecimento de padrões. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

IMPACTO POTENCIAL DESTA PESQUISA

A contribuição científica e tecnológica deste estudo consiste na proposição de uma metodologia rigorosa para predição computacional de hotspots em interações proteína-proteína, combinando descritores estruturais, aprendizado de máquina e mecanismos de prevenção de vazamento de dados. A solução apresentada amplia a confiabilidade e a capacidade de generalização de modelos preditivos aplicados à bioinformática e à biologia estrutural. A metodologia pode ser aplicada a outros problemas de classificação em dados biológicos de alta dimensionalidade, contribuindo para a produção de conhecimento científico.

POTENTIAL IMPACT OF THIS RESEARCH

The scientific and technological contribution of this study consists of proposing a rigorous methodology for the computational prediction of hotspots in protein-protein interactions, combining structural descriptors, machine learning techniques, and data leakage prevention mechanisms. The proposed solution enhances the reliability and generalization capability of predictive models applied to bioinformatics and structural biology. Furthermore, the proposed approach can be applied to other classification problems involving high-dimensional biological data, contributing to the advancement of scientific knowledge.

RODRIGO PEREZ VERNA

METODOLOGIA DE REDUÇÃO DIMENSIONAL HIERÁRQUICA COM PREVENÇÃO DE VAZAMENTO DE DADOS PARA PREDIÇÃO DE HOTSPOTS EM INTERAÇÕES PROTEÍNA-PROTEÍNA

Dissertação apresentado(a) à Universidade Estadual Paulista (UNESP), Instituto de Biociências, Letras e Ciências Exatas, São José do Rio Preto/SP, para obtenção do título de Mestre em Ciências da Computação.

Área de Concentração: Inteligência Computacional

Data de defesa: 02/03/2026

BANCA EXAMINADORA

Profº Dr. Geraldo Francisco Donegá Zafalon
UNESP – Instituto de Biociências, Letras e Ciências Exatas – Campus de São José do Rio Preto/SP

Profº Dr. Carlos Roberto Valêncio
Universidade Estadual Paulista "Júlio de Mesquita Filho"

Profº Dr. Ângelo Cesar Colombini
Universidade Federal Fluminense

Dedico este trabalho à Maraisa Navarro Verna e Marina Verna, pela paciência, compreensão e apoio nas inúmeras noites de ausência.

AGRADECIMENTOS

À Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP) e ao Instituto de Biociências, Letras e Ciências Exatas (IBILCE), pela infraestrutura e pelo ambiente acadêmico que possibilitaram o desenvolvimento desta pesquisa.

Aos Diretores da Usina São José da Estiva, pelo incentivo à realização deste estudo. Em especial, ao Sr. Nilton José Andreotti Filho, que proporcionou todo o suporte necessário para sua execução.

Ao meu orientador, Prof. Dr. Geraldo Francisco Donegá Zafalon, que me acolheu desde o primeiro momento na universidade, ensinou-me a investigar e questionar, e me conduziu com segurança diante das dúvidas surgidas ao longo da elaboração deste trabalho.

À Embrapa Agricultura Digital, em especial ao Dr. Ivan Mazoni e ao Dr. Inácio Henrique Yano, que me ensinaram a compreender a importância das proteínas para a manutenção da vida e abriram as portas da biologia estrutural. A contribuição de ambos foi fundamental para o desenvolvimento deste trabalho.

Aos membros da banca examinadora, Prof. Dr. Carlos Roberto Valêncio e Prof. Dr. Ângelo Cesar Colombini, cujas contribuições me permitiram enxergar melhorias que tornaram este trabalho mais robusto e relevante.

Aos colegas do Laboratório de Biologia Computacional do Instituto de Biociências, Letras e Ciências Exatas, em especial a Vitória Zanon Gomes e Bruno Rodrigues da Silveira, que muito me auxiliaram nos desafios metodológicos enfrentados durante este estudo.

“Nunca é tarde demais para ser o que você poderia ter sido.”
(ELIOT, George)

RESUMO

As interfaces proteína-proteína representam nanoambientes funcionais críticos para a comunicação molecular e a função biológica. A energia de ligação nessas interfaces concentra-se em poucos resíduos denominados hotspots, cuja identificação experimental é custosa e de difícil escalabilidade. Este trabalho apresenta uma metodologia hierárquica para predição computacional de hotspots, integrando descritores moleculares estruturais do STING_RDB com dados experimentais curados do PPI-HotSpotDB. A metodologia priorizou o controle rigoroso de vazamento de dados em múltiplas camadas: clusterização de sequências por similaridade, particionamento estratificado por proteína e validação cruzada com GroupKFold. A seleção hierárquica de descritores combinou mRMR *in-fold*, teste de Kolmogorov-Smirnov com correção FDR, remoção de multicolinearidade via VIF e seleção final por Boruta, resultando em redução dimensional de 97,6% a partir do conjunto inicial de descritores. Seis modelos foram avaliados sobre 918 resíduos de 158 proteínas. O SVM com kernel RBF alcançou AUC-ROC de 0,821 e sensibilidade de 0,844, enquanto o XGBoost obteve MCC de 0,538 e F1-score de 0,715. A comparação entre pipelines com e sem otimização de hiperparâmetros demonstrou que a qualidade dos descritores selecionados constitui o principal determinante do desempenho preditivo, superando o impacto do ajuste algorítmico. Os resultados demonstram a viabilidade de descritores estruturais para caracterização de determinantes energéticos em interfaces proteína-proteína, conciliando desempenho preditivo, rigor estatístico e interpretabilidade biológica.

Palavras-Chave: hotspots; interações proteína-proteína; aprendizado de máquina; STING_RDB.

ABSTRACT

Protein-protein interfaces represent critical functional nanoenvironments for molecular communication and biological function. Binding energy at these interfaces is concentrated in a few residues termed hotspots, whose experimental identification is costly and difficult to scale. This work presents a hierarchical methodology for computational hotspot prediction, integrating structural molecular descriptors from STING_RDB with curated experimental data from PPI-HotSpotDB. The methodology prioritized rigorous data leakage control through multiple layers: sequence clustering by similarity, protein-stratified partitioning, and GroupKFold cross-validation. The hierarchical descriptor selection combined in-fold mRMR, Kolmogorov-Smirnov test with FDR correction, multicollinearity removal via VIF, and final selection by Boruta, achieving 97.6% dimensional reduction from the initial descriptor set. Six models were evaluated on 918 residues from 158 proteins. SVM with RBF kernel achieved AUC-ROC of 0.821 and sensitivity of 0.844, while XGBoost obtained MCC of 0.538 and F1-score of 0.715. The comparison between pipelines with and without hyperparameter optimization demonstrated that the quality of selected descriptors constitutes the main determinant of predictive performance, surpassing the impact of algorithmic tuning. The results demonstrate the feasibility of structural descriptors for characterizing energetic determinants in protein-protein interfaces, reconciling predictive performance, statistical rigor, and biological interpretability.

Keywords: hotspots; protein-protein interactions; machine learning; STING_RDB.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A evolução das técnicas computacionais possibilita a aceleração de estudos laboratoriais que enfrentam a lentidão dos experimentos e os limites orçamentários para a sua realização (Bajorath, 2002), e diversas áreas de estudo têm se beneficiado de técnicas computacionais, com destaque para a biologia estrutural (Schwede, 2013). Profissionais da área apoiam-se em tecnologias como bancos de dados, processamento de imagem e aprendizado de máquina para analisar, prever, descobrir padrões e, principalmente, entender o funcionamento molecular e estrutural das proteínas (Greener et al., 2022).

As proteínas são macromoléculas compostas por cadeias de aminoácidos, cuja estrutura tridimensional determina suas funções biológicas, atuando como catalisadores enzimáticos, elementos estruturais, moléculas de sinalização e transporte (Alberts et al., 2017). Elas desempenham papéis essenciais em praticamente todos os processos celulares e raramente exercem suas funções isoladamente (Berg; Tymoczko; Stryer, 2012). Além disso, dependem das interações proteína-proteína (PPIs) para exercer funções biológicas complexas, incluindo sinalização celular, regulação metabólica e montagem de complexos macromoleculares.

As PPIs e as interfaces proteicas representam um dos principais nanoambientes funcionais em proteínas e são críticas para a comunicação molecular e para determinar a função biológica (Mazoni; Neshich, 2018). Clackson & Wells (1995) demonstraram que a energia de ligação em complexos proteína-proteína não é uniformemente distribuída na interface, mas está concentrada em poucos resíduos da interface. Esta descoberta desafiou o entendimento vigente de que todos os resíduos da interface contribuem igualmente para a afinidade de ligação. Os resíduos de alta contribuição energética foram denominados pontos de alta atratividade, ou hotspots (Bogan; Thorn, 1998).

A identificação e a caracterização de hotspots tornaram-se, desde então, problemas centrais na biologia estrutural e na bioinformática devido à sua relevância para a compreensão das interações proteína-proteína (Clackson; Wells, 1995; Moreira; Fernandes; Ramos, 2007). Hotspots estão diretamente associados à estabilidade de complexos proteicos e representam alvos estratégicos para engenharia de proteínas e desenvolvimento de fármacos (Keskin; Ma; Nussinov, 2005). Entretanto, sua determinação experimental é custosa, demorada e de difícil escalabilidade, o que limita a disponibilidade de dados experimentais e reforça a necessidade de abordagens computacionais confiáveis (Moreira; Fernandes; Ramos, 2007).

Nesse cenário, os métodos computacionais capazes de descrever com precisão o ambiente estrutural local dos resíduos proteicos tornam-se ferramentas essenciais. Entretanto, muitos desses métodos enfrentam desafios relacionados à generalização para proteínas não observadas durante o treinamento e frequentemente possibilitam o vazamento de dados entre os conjuntos de treino e teste (Park; Marcotte, 2012). Além disso, a interpretabilidade biológica dos modelos permanece limitada, dificultando a compreensão dos determinantes estruturais que possuem

maior sinal preditivo para os hotspots. Os descritores moleculares que capturam propriedades geométricas e físico-químicas do ambiente estrutural local de resíduos oferecem potencial para endereçar essas limitações (Neshich et al., 2003), mas requerem estratégias rigorosas de seleção e validação robusta para garantir a capacidade de generalização do modelo.

Diante desse cenário, o problema de pesquisa desta dissertação consiste em investigar se é possível propor uma metodologia de predição de hotspots em interações proteína-proteína que integre descritores estruturais do STING_RDB com prevenção sistemática de vazamento de dados em múltiplas camadas, e que os descritores selecionados por essa metodologia apresentem convergência com as teorias clássicas de formação de hotspots. Como contribuição principal, esta dissertação propõe e valida uma metodologia hierárquica de seleção de descritores que concilia desempenho preditivo, rigor estatístico e interpretabilidade biológica.

1.1 MOTIVAÇÃO

A identificação precisa de hotspots possui implicações diretas para múltiplas áreas da biotecnologia. Na saúde humana, hotspots representam alvos farmacológicos privilegiados para o desenvolvimento de terapias baseadas em modulação de PPIs, com aplicações em doenças onde interações proteicas são críticas, incluindo câncer, doenças neurodegenerativas e infecções virais (Keskin et al., 2008). No agronegócio, a compreensão de PPIs em plantas abre caminhos para o desenvolvimento de cultivares com maior resistência a pragas e doenças, tolerância a estresses abióticos, e melhor eficiência no uso de nutrientes, contribuindo diretamente para a segurança alimentar e sustentabilidade agrícola (Braun; Gingras, 2012). Entretanto, a identificação *in vitro* de hotspots permanece custosa e de difícil escalabilidade, limitando sua aplicação em larga escala. Métodos computacionais que predizem hotspots com alta precisão e generalização robusta podem acelerar significativamente a descoberta de alvos estratégicos nessas áreas, reduzindo custos e tempo em processos de pesquisa e desenvolvimento. Além disso, modelos interpretáveis que revelam os determinantes estruturais de hotspots contribuem para o entendimento fundamental de PPIs, informando estratégias racionais de engenharia molecular.

1.2 OBJETIVO

Este trabalho busca propor e validar uma metodologia hierárquica de seleção de descritores para predição de hotspots em PPIs, capaz de reduzir dimensionalmente os descritores com prevenção de vazamento de dados e apresentar validação biológica com a teoria clássica de hotspots. Os objetivos específicos incluem:

(i) Integrar os descritores estruturais do STING_RDB com as informações de hotspots experimentais do conjunto curado PPI-Hotspot+PDB^{BM};

(ii) Aplicar normalização intra-proteína para reduzir o viés estrutural entre famílias proteicas;

(iii) Definir e aplicar estratégia de particionamento entre conjuntos de treino e teste com segregação por proteína, de modo a reduzir a dependência estrutural entre amostras e mitigar o vazamento de dados;

(iv) Desenvolver um pipeline hierárquico combinando mRMR *in-fold*, teste Kolmogorov-Smirnov com correção FDR, remoção de multicolinearidade via VIF e seleção final via Boruta;

(v) Treinar e avaliar Random Forest, Support Vector Machines e XGBoost utilizando validação cruzada estratificada por proteínas (GroupKFold);

(vi) Comparar o desempenho entre pipelines com hiperparâmetros padrão e otimizado via GridSearchCV;

(vii) Analisar a explicabilidade dos modelos via SHAP;

(viii) Avaliar a relevância biológica dos descritores selecionados pela metodologia com a bibliografia clássica sobre hotspots.

1.3 JUSTIFICATIVA

Trabalhos anteriores estabeleceram fundamentos importantes para predição computacional de hotspots. Pereira (2012) demonstrou que descritores estruturais baseados em propriedades físico-químicas e geométricas são altamente informativos para discriminar hotspots, alcançando AUC de 0,799. Wang, Liu & Deng (2018) avançaram o campo integrando múltiplas fontes de informação com técnicas de aprendizado profundo, reportando métricas competitivas no conjunto de teste independente (AUC = 0,831, MCC = 0,70). Trabalhos mais recentes como SPOTONE (Preto; Moreira, 2020) e Chen et al. (2024) exploraram abordagens baseadas em *ensemble* e frameworks automatizados de aprendizado de máquina, demonstrando que a integração de descritores estruturais, energéticos e evolutivos com técnicas modernas de aprendizado de máquina melhora significativamente a capacidade preditiva.

Entretanto, análises críticas revelam limitações metodológicas recorrentes. Park & Marcotte (2012) demonstraram que o vazamento de dados é recorrente em predições envolvendo amostras correlacionadas, resultando em superestimação sistemática da capacidade de generalização dos modelos computacionais. No contexto de predição de hotspots, observam-se problemas específicos: Wang, Liu & Deng (2018), apesar do alto desempenho reportado, utilizaram múltiplas mutações das mesmas interfaces proteicas em treino e teste, criando dependência estrutural entre amostras. Assim, o modelo pode aprender características específicas de um complexo e reutilizá-las em mutações relacionadas no conjunto de teste, elevando artificialmente as métricas reportadas. Trabalhos como Pereira (2012), SPOTONE (Preto; Moreira, 2020) e Chen et al. (2024) não reportaram controle explícito de homologia ou estratificação por proteína,

potencialmente permitindo que proteínas relacionadas ou até idênticas apareçam em ambos os conjuntos de treino e teste.

1.4 ORGANIZAÇÃO DO TRABALHO

O Capítulo 2 apresenta a fundamentação teórica que embasa este estudo. O Capítulo 3 apresenta o desenvolvimento do trabalho, subdividido em três eixos principais. O primeiro eixo descreve os materiais e métodos utilizados, incluindo as bases de dados PPI-HotSpotDB e STING_RDB, a estratégia de particionamento com segregação por proteína, o pipeline hierárquico de seleção de descritores e os algoritmos de aprendizado de máquina empregados. O segundo eixo apresenta os resultados obtidos em cada etapa do pipeline, incluindo a redução dimensional, a caracterização dos descritores selecionados, o desempenho dos modelos e a análise de explicabilidade via SHAP. O terceiro eixo discute os resultados à luz da literatura clássica, analisando a relevância biológica dos descritores, comparando o desempenho com trabalhos anteriores e apresentando as limitações do estudo. Por fim, o Capítulo 4 apresenta as conclusões, sumarizando as principais contribuições e sugerindo direções para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os fundamentos teóricos necessários à compreensão do problema abordado neste trabalho. Inicialmente, descrevem-se os princípios bioquímicos que governam a estrutura e a função das proteínas, incluindo as forças de ligação responsáveis pela estabilidade de complexos moleculares. Em seguida, introduz-se o conceito de nanoambiente proteico e sua relevância para a caracterização funcional de resíduos em interfaces proteína-proteína. A definição de hotspots e as teorias clássicas que explicam sua formação são então apresentadas, seguidas pela descrição dos métodos experimentais e computacionais disponíveis para sua identificação. Posteriormente, descrevem-se as bases de dados utilizadas neste estudo, o PPI-HotSpotDB e o STING_RDB, cujos dados experimentais e descritores estruturais constituem a base do pipeline proposto. Por fim, discutem-se os aspectos estatísticos e computacionais relevantes para a modelagem preditiva em dados biológicos, com ênfase no controle de redundância, na validação estruturada e na multicolinearidade, e apresentam-se os trabalhos correlatos que contextualizam a contribuição do presente estudo.

2.1 PROTEÍNAS E ENERGIA LIVRE DE LIGAÇÃO

As proteínas são biomoléculas essenciais aos sistemas vivos que, ao lado de carboidratos, lipídios e ácidos nucleicos, compõem o repertório molecular fundamental dos organismos. Distinguem-se das demais macromoléculas pela diversidade funcional que exercem: atuam como catalisadores enzimáticos, elementos estruturais, moléculas de sinalização, transportadores e agentes de defesa imunológica (Alberts et al., 2017).

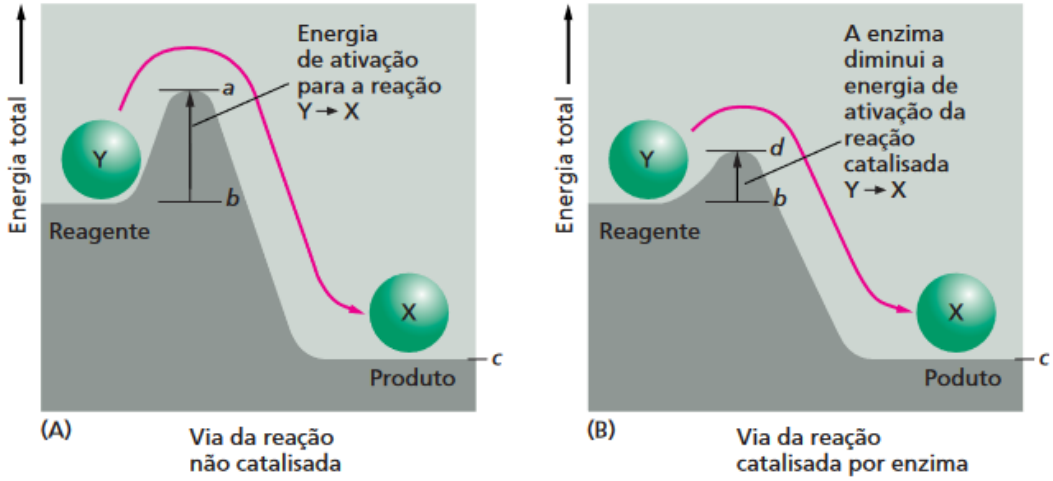
No ambiente intracelular, as moléculas biológicas encontram-se em estados relativamente estáveis. As moléculas precisam de ajuda para ultrapassar uma barreira energética considerável e sofrer uma reação química. Dentro das células, essas barreiras têm sua energia de ativação reduzida através do auxílio de uma classe especializada de proteínas, as enzimas. Cada enzima se liga a uma ou mais moléculas denominadas substratos. Essa ligação reduz consideravelmente a energia de ativação necessária facilitando assim a ocorrência de determinadas interações químicas. Nesse sentido, as enzimas atuam como catalisadores (aceleradores de reações químicas) eficientes, pois reduzem a energia de ativação das reações que catalisam (Alberts et al., 2017).

A energia livre de Gibbs (G) é uma grandeza termodinâmica que quantifica a energia disponível em um sistema para realizar trabalho útil a temperatura e pressão constantes. Em sistemas biológicos, essa grandeza determina a espontaneidade das reações químicas e a estabilidade das interações moleculares (Alberts et al., 2017).

O termo de maior relevância para a bioquímica e a biologia molecular é a variação da energia livre de Gibbs (ΔG), que expressa a diferença de energia livre entre os estados final e inicial de uma reação. Quando $\Delta G < 0$, a reação é termodinamicamente favorável e ocorre espontane-

amente; quando $\Delta G > 0$, a reação requer aporte de energia para prosseguir. No contexto das interações proteína-proteína, ΔG quantifica a estabilidade termodinâmica da associação entre duas proteínas: valores mais negativos indicam maior afinidade de ligação. A variação da energia livre de ligação é expressa em kcal/mol e constitui o parâmetro central para a caracterização energética de interfaces proteicas e, em particular, para a identificação de hotspots, conforme discutido nas seções subsequentes (Alberts et al., 2017).

Figura 1 – Energia de Ativação.



Fonte: Adaptado de Alberts et al. (2017)

Nesse sentido, a predição de reações entre moléculas pode ser executada observando $\Delta\mathbf{G}$, pois é quando ocorre o ganho ou perda de energia livre. Nesse processo, um mol de reagente é convertido em um mol de produto, sob condições padrão (quando as moléculas estão presentes na concentração de 1 mol/L e no pH 7,0).

2.2 FORÇAS DE LIGAÇÃO

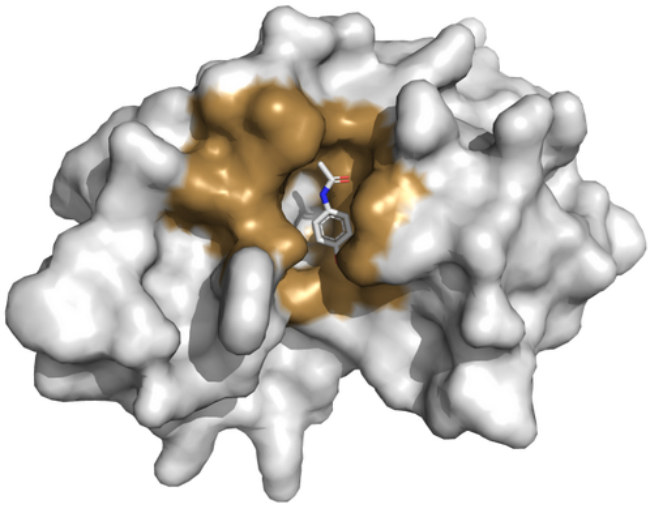
As proteínas possuem a propriedade de se ligar a outras moléculas e a outras proteínas. Nesse sentido, qualquer substância que se liga a uma determinada proteína é classificada como ligante. As ligações proteicas podem ser covalentes ou não covalentes: as covalentes são aquelas em que dois átomos compartilham um par de elétrons e as não covalentes são ligações fracas que incluem interações eletrostáticas, pontes de hidrogênio e força de van der Waals. Além disso, a capacidade de ligação não covalente permite o enovelamento das proteínas em modos diferentes. Essas ligações podem somar-se e criar uma atração forte entre duas moléculas. No momento em que essas moléculas encaixarem-se perfeitamente uma à outra pode haver, entre elas, a formação de muitas ligações não covalentes. Esse processo é denominado mão em luva (quando ocorre o encaixe perfeito entre superfícies proteicas) (Alberts et al., 2017).

A habilidade da ligação seletiva e a alta afinidade ao ligante são dependentes da formação de um conjunto de interações fracas não covalentes. A seguir, listam-se os principais processos de ligação não covalente:

- **Pontes de hidrogênio:** como são polarizadas, duas moléculas de água que estejam adjacentes podem formar uma ligação conhecida como ponte de hidrogênio;
- **Força de Van der Waals:** em um dado instante os elétrons de uma molécula apolar, que estão em constante movimento, passam a ter mais elétrons de um lado do que de outro, tornando a molécula momentaneamente polarizada. Essa indução elétrica polariza a molécula vizinha, ou seja, cria um dipolo induzido;
- **Interações hidrofóbicas:** A estrutura tridimensional da água força os grupos hidrofóbicos a ficarem juntos de modo a minimizar o efeito que eles têm em perturbar a rede de moléculas de água mantida por pontes de hidrogênio. A expulsão da solução proporciona a interação chamada hidrofóbica;
- **Ligação iônica:** interações eletrostáticas entre grupos com cargas opostas, como as pontes salinas formadas entre cadeias laterais de aminoácidos carregados positiva e negativamente.

Essa habilidade de enovelar tomando uma estrutura tridimensional traz para perto, no espaço tridimensional, os resíduos afastados na sequência. Esse processo compõe uma cavidade ou uma fresta que acomoda o ligante. Na Figura 2 observam-se, em marrom, os resíduos formadores de um sítio de ligação. A molécula ligante é representada em bastões (Nelson; Cox, 2022).

Figura 2 – Cavidade que acomoda os ligantes.



Fonte: Adaptado de Alberts et al. (2017)

2.3 O NANOAMBIENTE PROTEICO

O ambiente estrutural local das proteínas constitui um determinante fundamental de sua função biológica, sendo definido pelas propriedades físico-químicas e geométricas dos resíduos de aminoácidos e pelas interações estabelecidas entre eles no espaço tridimensional. Esses resíduos não atuam de forma isolada, mas integram um contexto estrutural coletivo que combina composição local, organização espacial e conectividade com resíduos vizinhos. Esse contexto funcional foi denominado por Neshich *et al.* (Neshich et al., 2015) como *nanoambiente proteico*.

O nanoambiente proteico pode ser compreendido como uma região local ao redor de um resíduo capaz de capturar simultaneamente características intrínsecas do aminoácido e propriedades emergentes do arranjo estrutural circundante. Alterações na composição, densidade, hidrofobicidade ou padrão de contatos nessa região podem modificar substancialmente o comportamento funcional do resíduo, influenciando sua participação em interações moleculares, sua estabilidade estrutural e sua contribuição energética.

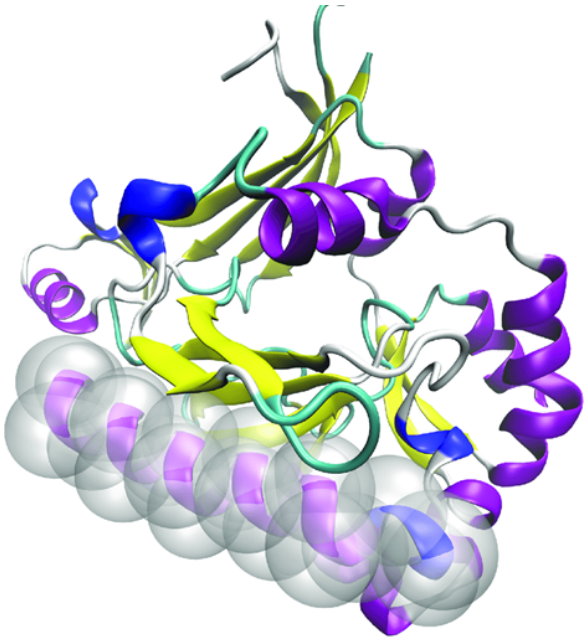


Figura 3 – Representação esquemática do nanoambiente proteico.

Fonte: Adaptado de Mazoni (2018).

Entre os diferentes contextos nos quais os nanoambientes proteicos podem ser definidos, as interfaces proteína-proteína representam um caso particular de interesse. Nessas regiões, resíduos localizados na superfície das proteínas estabelecem contatos diretos responsáveis pela formação de complexos funcionais. A estabilidade e a especificidade dessas interações dependem fortemente de propriedades locais, como complementaridade geométrica, exclusão de solvente, densidade de contatos e organização eletrostática (Bogan; Thorn, 1998).

Estudos experimentais demonstraram que a energia de ligação em interfaces proteína-proteína não é distribuída de forma homogênea ao longo da interface, mas concentrada em um subconjunto

reduzido de resíduos (Clackson; Wells, 1995). Esses resíduos constituem nanoambientes energéticos especializados dentro da interface, nos quais propriedades estruturais e físico-químicas convergem para maximizar a contribuição energética local. A identificação e a caracterização desses resíduos motivaram a definição do conceito de *hotspots*, discutido em detalhe na seção 2.4 seguinte.

A caracterização computacional de nanoambientes associados a interfaces proteína-proteína requer descritores capazes de capturar não apenas propriedades individuais dos resíduos, mas também informações contextuais relacionadas a contatos, densidade estrutural, conectividade e organização espacial local. Abordagens baseadas em nanoambientes estruturais fornecem, portanto, uma base conceitual consistente para a modelagem e a predição de regiões energeticamente relevantes em interfaces proteína-proteína (Neshich et al., 2004).

2.4 HOTSPOTS EM INTERFACES PROTEÍNA-PROTEÍNA

O trabalho de referência conduzido por Clackson e Wells (Clackson; Wells, 1995) demonstrou que a energia de ligação em interfaces proteína-proteína não é distribuída de maneira uniforme entre os resíduos envolvidos, mas concentrada em um subconjunto reduzido da interface. Esses resíduos exercem contribuição desproporcional para a estabilidade do complexo, desempenhando papel dominante na afinidade de ligação, e foram denominados *hotspots*.

Bogan e Thorn (Bogan; Thorn, 1998) formalizaram a definição experimental de hotspots por meio da técnica de *Alanine Scanning Mutagenesis*. Nessa abordagem, resíduos individuais da interface são sistematicamente substituídos por alanina e a variação da energia livre de ligação causada pela mutação ($\Delta\Delta G = \Delta G_{\text{mutante}} - \Delta G_{\text{selvagem}}$) é mensurada. Diferentemente de ΔG, que quantifica a energia livre de ligação do complexo, $\Delta\Delta G$ expressa especificamente a contribuição energética individual de um resíduo para a estabilidade da interface. Resíduos cuja mutação resulta em $\Delta\Delta G \geq 2{,}0$ kcal/mol são classificados como hotspots, estabelecendo um critério operacional amplamente adotado na literatura (Bogan; Thorn, 1998; Moreira; Fernandes; Ramos, 2007). Cabe observar que algumas abordagens utilizam limiares alternativos, como $\Delta\Delta G \geq 1{,}5$ kcal/mol, refletindo diferentes balanços entre sensibilidade e especificidade na definição de *hotspot*. O presente trabalho adota o limiar de 2,0 kcal/mol por ser o critério utilizado na curadoria do conjunto PPI-HotSpotDB (Chen et al., 2022), assegurando consistência entre a definição experimental e os dados de treinamento.

A Figura 4 ilustra a distribuição espacial de hotspots na interface do complexo β-catenina/Tcf4, evidenciando a organização em regiões discretas (*hot regions*) que concentram os resíduos de maior contribuição energética.

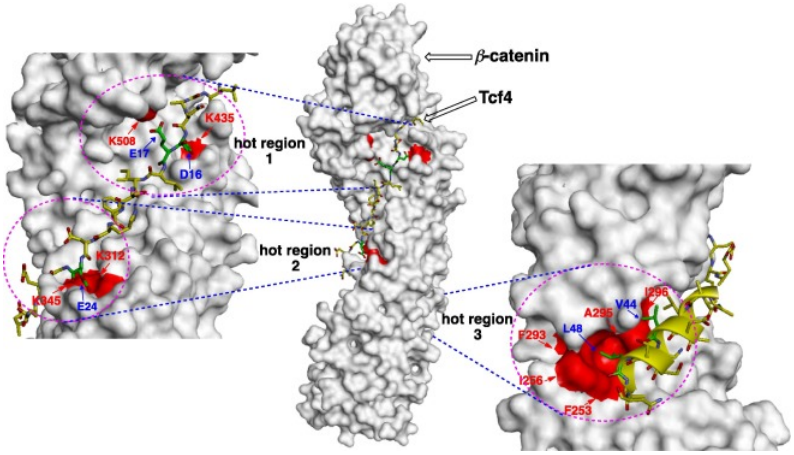


Figura 4 – Distribuição espacial de hotspots na interface do complexo β-catenina/Tcf4. Resíduos em vermelho indicam hotspots experimentais organizados em três regiões discretas (*hot regions*).

Fonte: Adaptado de Guo, Wisniewski & Ji (2014).

Análises experimentais subsequentes indicaram que determinados aminoácidos apresentam maior propensão a atuar como hotspots, com destaque para triptofano, arginina e tirosina. Essa preferência está associada às propriedades físico-químicas de suas cadeias laterais, incluindo hidrofobicidade elevada, capacidade de formar múltiplas ligações de hidrogênio e interações eletrostáticas favoráveis, que contribuem para o fortalecimento das interações na interface (Pereira, 2012).

2.4.1 Teoria do O-Ring

A Teoria do O-Ring, proposta por Bogan e Thorn (Bogan; Thorn, 1998), sugere que hotspots centrais são circundados por um anel de resíduos menos energéticos, frequentemente polares, cuja função principal é promover a exclusão parcial de moléculas de água da interface. Essa organização estrutural reduz a constante dielétrica local e intensifica interações eletrostáticas e ligações de hidrogênio, contribuindo significativamente para a estabilidade do complexo proteína-proteína.

Embora a exclusão de solvente desempenhe papel relevante, os autores destacam que ela não constitui o único determinante da contribuição energética de um resíduo. O tipo, o número e a geometria das interações estabelecidas também exercem influência crítica, indicando que hotspots emergem da combinação integrada de múltiplos fatores estruturais.

2.4.2 Teoria da Dupla Exclusão de Água

Li & Liu (2009a) refinaram a hipótese do O-Ring ao introduzir a Teoria da Dupla Exclusão de Água. Segundo esse modelo, o núcleo do hotspot é caracterizado por um empacotamento denso de resíduos hidrofóbicos, completamente desprovido de moléculas de água, enquanto

uma segunda camada externa atua na manutenção dessa exclusão e na estabilização do ambiente energético local.

Essa teoria fornece uma descrição mais detalhada da organização estrutural dos hotspots e estabelece uma relação direta entre densidade estrutural local, exclusão de solvente e contribuição energética elevada. Além de aprofundar a compreensão dos determinantes físicos das interfaces proteína-proteína, esse modelo oferece bases conceituais relevantes para aplicações em engenharia de proteínas e no desenvolvimento racional de moléculas moduladoras de interações proteína-proteína.

2.4.3 Relevância Científica dos Hotspots

O estudo de hotspots em interfaces proteína-proteína possui relevância central para a biologia estrutural e molecular, uma vez que esses resíduos concentram os determinantes energéticos que governam a formação, a estabilidade e a especificidade das interações proteicas. A compreensão de como poucos resíduos dominam a energia de ligação desafia modelos clássicos baseados em contribuições uniformes e fornece uma visão mais precisa da organização funcional das interfaces.

Além do impacto fundamental na compreensão de processos celulares e vias de sinalização, hotspots representam alvos estratégicos para a engenharia de proteínas e para o desenvolvimento de moléculas capazes de modular interações proteína-proteína (Wells; McClendon, 2007). Embora a maioria dos estudos seja focado em aplicações terapêuticas, os princípios estabelecidos podem ser aplicados ao agronegócio, incluindo o desenho racional de agrodefensivos, a identificação de alvos moleculares em fitopatógenos e pragas agrícolas, e a engenharia de proteínas associadas à resistência a estresses em plantas. Dessa forma, a caracterização estrutural de hotspots constitui tanto um problema científico fundamental quanto uma oportunidade para aplicações biotecnológicas no setor agropecuário.

2.4.4 Métodos experimentais para detecção de hotspots

A detecção experimental de hotspots consiste em medir a variação da energia livre de ligação através da varredura por alaninas, onde os resíduos são substituídos individualmente por alanina. Esse processo é demorado e trabalhoso, pois cada proteína com um aminoácido substituído por alanina (Clackson; Wells, 1995) precisa ser expressa, purificada e analisada separadamente através de experimentos *in vitro*. Devido ao alto custo operacional e à demanda de tempo e recursos laboratoriais, a caracterização *in vitro* em larga escala de hotspots torna-se inviável, motivando o desenvolvimento de métodos computacionais para predição. A literatura apresenta ainda outros métodos como o *shotgun scanning* (Weiss et al., 2000) que utiliza bibliotecas de *phage-display* (técnica de clonagem utilizada em biologia molecular) e o método de substituição binomial. A mutação de resíduos por alanina equivale à remoção dos átomos que estão além do carbono β em outros aminoácidos (Pereira, 2012).

A Figura 5 ilustra o procedimento de varredura por alanina, no qual cada resíduo da sequência nativa é sistematicamente substituído por alanina em experimentos independentes, permitindo quantificar a contribuição energética individual de cada posição.

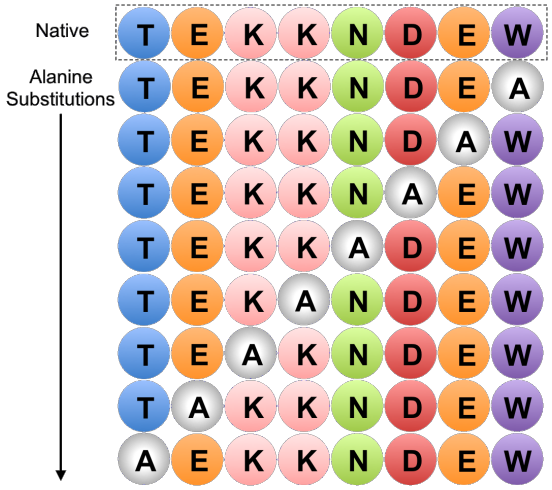


Figura 5 – Representação esquemática da varredura por alanina (*Alanine Scanning Mutagenesis*). Cada resíduo da sequência nativa (topo) é substituído individualmente por alanina (A), gerando variantes independentes para mensuração de $\Delta\Delta G$.

Fonte: Sitko (2013). Licença CC BY-SA 3.0.

2.4.5 Métodos computacionais para predição de hotspots

Os métodos computacionais para predição de hotspots utilizam informações estruturais e energéticas de complexos proteína-proteína com o objetivo de estimar a contribuição individual dos resíduos para a energia de ligação. De forma geral, essas abordagens podem ser classificadas em duas categorias principais: (i) métodos baseados no conhecimento e na análise de dados experimentais (*knowledge-based*) e (ii) métodos baseados em princípios físicos (*physics-based*) (Moreira; Fernandes; Ramos, 2007).

Os métodos *physics-based* representam as primeiras tentativas computacionais sistemáticas de identificar hotspots e buscam reproduzir, de maneira explícita, o processo experimental de mutagênese por alanina. Essas abordagens utilizam simulações em nível atômico para estimar a $\Delta\Delta G$ causada pela mutação de um resíduo para alanina, empregando técnicas como dinâmica molecular, *free energy perturbation* (FEP) e *thermodynamic integration* (TI). Ferramentas como Robetta (Kim; Chivian; Baker, 2004) e FoldX (Schymkowitz et al., 2005) exemplificam essa classe de métodos, apresentando boa correlação com dados experimentais em muitos casos. No entanto, o elevado custo computacional dessas simulações, aliado à necessidade de parametrizações cuidadosas e infraestrutura computacional dedicada, limita sua aplicação em larga escala (Kortemme; Kim; Baker, 2004; Pereira, 2012).

Em contraste, os métodos *knowledge-based* surgiram como alternativas mais eficientes do ponto de vista computacional. Essas abordagens utilizam descritores estruturais e físico-químicos extraídos de complexos proteicos conhecidos, tais como acessibilidade ao solvente, hidrofobici-

dade, densidade de contatos, complementaridade geométrica e propriedades eletrostáticas. Esses descritores são combinados com modelos estatísticos ou de aprendizado de máquina treinados a partir de conjuntos de dados experimentais de mutagênese, nos quais valores de $\Delta\Delta G$ de ligação estão disponíveis (Cho; Kim; Lee, 2009).

Inicialmente, muitos desses métodos empregavam regras heurísticas e modelos estatísticos simples, baseados em limiares fixos para propriedades estruturais. Com o avanço do aprendizado de máquina, técnicas supervisionadas passaram a ser amplamente utilizadas, incluindo árvores de decisão, *random forests*, *support vector machines* e, mais recentemente, métodos baseados em *ensemble* e gradiente. Essas abordagens permitiram capturar relações não lineares entre descritores e contribuição energética, melhorando o desempenho preditivo em comparação com métodos puramente heurísticos.

Apesar dos avanços, métodos *knowledge-based* enfrentam desafios importantes, como a alta dimensionalidade dos descritores, a presença de multicolinearidade e o risco de viés estrutural e vazamento de informação entre conjuntos de treinamento e teste. Essas limitações justificam a adoção, neste trabalho, de um pipeline hierárquico de seleção de descritores com controle explícito de redundância, multicolinearidade e vazamento de dados, cuja formulação e validação são apresentadas no Capítulo 3.

2.5 BANCO DE DADOS DE HOTSPOTS E DESCRITORES PROTEICOS

O desenvolvimento de modelos computacionais para predição de hotspots requer dois componentes fundamentais: informações de hotspots validadas e descritores estruturais capazes de capturar as propriedades físico-químicas e geométricas que distinguem os resíduos energeticamente críticos. Esta seção descreve as fontes de dados empregadas neste estudo: o PPI-HotSpotDB, que fornece anotações experimentais curadas de hotspots derivadas de mutagênese por varredura de alanina, e o STING_RDB, um banco de dados relacional que oferece mais de 2.000 descritores moleculares que caracterizam o nanoambiente de cada resíduo em estruturas proteicas. A integração desses recursos permite a investigação sistemática dos determinantes estruturais subjacentes à formação de hotspots.

2.5.1 PPI-HotSpotDB

O banco de dados PPI-HotSpotDB foi construído a partir de dados experimentais provenientes das bases Alanine Scanning Energetics Database (ASEdb) (Thorn; Bogan, 2001), Structural Kinetic and Energetic Database of Mutant Protein Interactions (SKEMPI 2.0) (Jankauskaitė et al., 2019) e Universal Protein Knowledgebase (UniProtKB) (UniProt Consortium, 2021). Ao todo, ele reúne 4.039 hotspots distribuídos em 1.893 proteínas, abrangendo 173 espécies, com predominância de origem humana. Inclui informações detalhadas sobre cada hotspot, como estrutura secundária, acessibilidade ao solvente, rigidez, mutações associadas, estruturas tridimensionais e vínculos com bases externas, como UniProt e PDB (Chen et al., 2022).

A correspondência entre posições em UniProt e PDB foi validada por meio do Structure Integration with Function, Taxonomy and Sequences (SIFTS) (Velankar et al., 2013), assegurando o mapeamento correto dos resíduos. Além disso, as análises estruturais publicadas demonstram que 53% dos hotspots estão expostos ao solvente e 34% estão parcialmente enterrados. Em relação à estrutura secundária, 37% localizam-se em α-hélices e 30% em folhas-β, indicando predominância nessas regiões. Quanto à rigidez, aproximadamente 75% encontram-se em um desvio padrão da média e 19% são classificados como rigidamente estruturados (Chen et al., 2022).

Cabe destacar que os 4.039 hotspots referem-se a resíduos únicos por proteína, sem considerar que o mesmo hotspot pode ser mapeado em múltiplas estruturas distintas do PDB. O conjunto de dados disponibilizado pelos autores em seu sítio eletrônico possui múltiplas instâncias específicas de hotspots experimentais em diferentes estruturas e cadeias. Essa diferença ocorre porque cada entrada representa uma ocorrência estrutural individual, refletindo variações conformacionais e de resolução. Os autores disponibilizaram em seu sítio eletrônico um recorte tratado chamado `PPI-Hotspot+PDB`$^{\mathrm{BM}}$ [1] específico para utilização em modelos computacionais para predição de hotspots (Chen et al., 2022).

O PPI-HotSpotDB está organizado em quatro coleções principais e um subconjunto:

- **UniProt:** contém dados sobre proteínas e mutações experimentais, com foco funcional e anotação biológica.
- **PDB_separate:** reúne estruturas individuais de cadeias proteicas para análise granular de hotspots.
- **PDB_aggregate:** integra estruturas relacionadas a uma mesma proteína para estudo ampliado em diferentes contextos estruturais.
- **Bound:** abrange estruturas de proteínas em estado ligado, permitindo a análise de hotspots no contexto funcional das PPIs.
- **PPI-Hotspot+PDB$^{\mathbf{BM}}$:** subconjunto curado e padronizado contendo hotspots estruturais, selecionado após redução de redundância e identidade de sequência inferior a 60%, destinado para *benchmark* em modelos de predição de hotspots.

2.5.2 STING_RDB

O STING_RDB [2] representa um avanço significativo na análise de estrutura e sequência de proteínas através de sua implementação em uma coleção de dados armazenada em banco de dados relacional. Foi desenvolvido para enfrentar os desafios comuns em bancos de dados de proteínas, que são frequentemente ruidosos, de alta dimensionalidade, esparsos e redundantes.

[1] Disponível em: https://ppihotspot.limlab.dnsalias.org/scripts/download.php
[2] Disponível em: https://www.cbi.cnptia.embrapa.br/SMS/index_s.html

Esses problemas surgem da complexidade biológica, limitações nos procedimentos experimentais e erros humanos na coleta e curadoria dos dados.

A integração do STING_RDB compreende duas camadas principais: descritores públicos de Sequência e Estrutura de Proteínas (PSS) e descritores STING PSS. A primeira camada consolida dados de bancos de dados públicos como PDB, HSSP, Prosite, UniProt e Protherm, enquanto a segunda camada calcula os descritores STING PSS considerando alguns parâmetros de bancos de dados públicos. A Figura 6 apresenta a organização do STING_RDB (Oliveira et al., 2007).

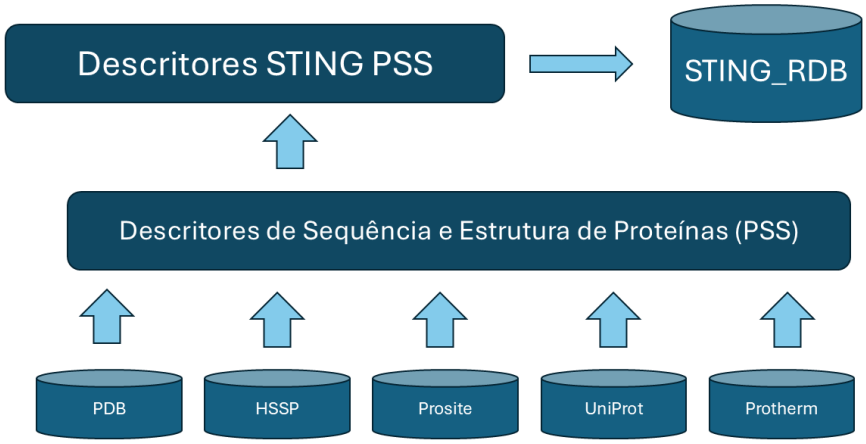


Figura 6 – Fluxo de dados de diferentes fontes para compor o banco de dados STING_RDB. Fonte: Adaptado de Oliveira et al. (2007).

Este banco de dados disponibiliza uma ampla gama de descritores físico-químicos, estruturais e geométricos utilizados para caracterizar resíduos de proteínas em análises de interação e predição de hotspots. Para a seleção de descritores, considerou-se os descritores agrupados atualmente em 120 tabelas, com propriedades físicas, químicas, estruturais e de conectividade (Kuser-Falcão et al., 2007). A seguir, descrevem-se as classes de descritores utilizadas neste estudo:

2.5.2.1 Classes de Descritores Proteicos

Descritores de Estrutura: Englobam os parâmetros fundamentais que definem a estrutura tridimensional das proteínas, incluindo a estrutura primária (sequência), secundária (DSSP, Stride) e terciária (Kabsch; Sander, 1983; Frishman; Argos, 1995; Neshich et al., 2003).

Descritores de Contato: Representam as interações não-covalentes entre resíduos e átomos na estrutura proteica (Neshich et al., 2003; Kuser-Falcão et al., 2007). Os contatos são fundamentais para a estabilidade da proteína e incluem parâmetros como energia de contato, contatos residuais, número de contatos ponderados e contatos não utilizados.

Descritores de Superfície e Cavidades: Caracterizam as regiões de superfície e cavidades das proteínas, que são frequentemente sítios de interação molecular e atividade catalítica. Incluem parâmetros como pockets (bolsões), curvatura da superfície e densidade esponja.

Descritores Físico-Químicos: Fornecem informações sobre as propriedades físicas e químicas das proteínas, como potencial eletrostático, hidrofobicidade e solvatação.

Descritores de Interação: Caracterizam as interações entre proteínas e ligantes, incluindo outras proteínas, ácidos nucleicos e pequenas moléculas.

Descritores Conformacionais: Fornecem informações detalhadas sobre a conformação dos resíduos e suas cadeias laterais. Incluem rotâmeros, orientação de cadeias laterais e distâncias interatômicas.

Descritores Especializados: Englobam parâmetros avançados como descritores de grafos para análise topológica da estrutura proteica, estatísticas do banco de dados e informações sobre permutações distintas.

2.6 ASPECTOS ESTATÍSTICOS E COMPUTACIONAIS RELEVANTES

Dados de domínio biológico apresentam características intrínsecas que os distinguem de outros domínios. São altamente correlacionados e, muitas vezes, derivados das mesmas relações físicas, químicas e evolutivas associadas ao mesmo nanoambiente. No contexto da biologia estrutural, essas características manifestam-se, por exemplo, na organização hierárquica das proteínas, nas restrições geométricas impostas pelo enovelamento da estrutura tridimensional da proteína e na dependência funcional entre os resíduos situados no mesmo nanoambiente (Neshich et al., 2004; Alberts et al., 2017).

Uma consequência direta dessa organização estrutural é a elevada redundância entre amostras. Proteínas homólogas ou variantes estruturais de um mesmo complexo tendem a compartilhar grande parte de sua informação geométrica e funcional, introduzindo dependência estatística entre as observações tratadas como independentes. Descritores extraídos dessas estruturas refletem padrões altamente similares, violando pressupostos fundamentais de muitos métodos de aprendizado de máquina (Park; Marcotte, 2012).

Quando essa dependência estatística não é adequadamente controlada, os modelos computacionais podem capturar padrões específicos de proteínas ou interfaces particulares, em vez de aprender determinantes gerais. Esse fenômeno, conhecido como vazamento de informação (*data leakage*), ocorre quando amostras estruturalmente relacionadas aparecem simultaneamente nos conjuntos de treinamento e teste, permitindo que o modelo reutilize informações implícitas já observadas (Park; Marcotte, 2012). O vazamento pode manifestar-se de forma explícita, quando dados idênticos ou quase idênticos são compartilhados entre treino e teste, ou de forma implícita, quando dependências estruturais ou evolutivas não são devidamente controladas durante o processo de validação.

No contexto específico da predição de hotspots, o vazamento de informação manifesta-se principalmente em três cenários (Park; Marcotte, 2012; Wang; Liu; Deng, 2018):

(i) Múltiplas mutações de alanina scanning realizadas sobre a mesma interface proteica são distribuídas entre conjuntos de treino e teste, permitindo que o modelo aprenda características específicas daquela interface.

(ii) Proteínas homólogas com interfaces estruturalmente conservadas são tratadas como observações independentes, violando a premissa de independência estatística.

(iii) Estruturas cristalográficas redundantes do mesmo complexo molecular, depositadas no PDB sob diferentes códigos de acesso, aparecem em ambos os conjuntos.

Para mitigar esses efeitos, estratégias rigorosas de controle de homologia e particionamento de dados são fundamentais. O uso de algoritmos como CD-HIT (Li; Godzik, 2006) para remoção de sequências redundantes, combinado com esquemas de validação cruzada baseados em agrupamento por proteína (*GroupKFold*), permite garantir que proteínas estruturalmente relacionadas permaneçam isoladas em um único conjunto, preservando a independência estatística necessária para avaliações com maior assertividade do desempenho preditivo. Esse controle rigoroso constitui um requisito fundamental para o desenvolvimento de modelos computacionais na predição de hotspots.

2.6.1 Controle de redundância e validação estruturada

O controle de redundância em nível de sequência, embora necessário, não é suficiente para eliminar completamente a dependência estrutural entre amostras. Proteínas com baixa identidade de sequência podem compartilhar arquiteturas estruturais similares, especialmente em regiões funcionais como interfaces proteína-proteína. Além disso, múltiplos resíduos pertencentes à mesma proteína ou interface tendem a apresentar dependência funcional, mesmo após a redução de redundância por sequência. Dessa forma, estratégias adicionais de validação tornam-se necessárias para garantir a independência efetiva entre os conjuntos de treinamento e teste (Park; Marcotte, 2012).

Para mitigar esse problema, esquemas de validação cruzada baseados em agrupamento (*grouped cross-validation*) têm sido propostos. Nessa abordagem, as amostras são associadas a identificadores de grupo que representam entidades estruturalmente dependentes, como proteínas individuais ou complexos. A divisão em partições é então realizada de forma que todas as amostras pertencentes a um mesmo grupo sejam mantidas exclusivamente em um único subconjunto, seja de treinamento ou de teste. O método GroupKFold formaliza esse princípio ao garantir que grupos inteiros não sejam fragmentados entre diferentes partições da validação cruzada. Dessa forma, a avaliação do modelo passa a refletir sua capacidade de generalizar para grupos completamente não observados durante o treinamento, configurando um cenário de validação mais rigoroso e biologicamente realista (Pedregosa et al., 2011).

Em aplicações de predição de hotspots, a validação estruturada por grupos é particularmente relevante, pois evita que o modelo explore dependências específicas de uma proteína ou interface já observada. Ao impor a separação estrita por proteína, o GroupKFold reduz significativamente o risco de vazamento de informação decorrente da similaridade estrutural e permite uma estimativa mais confiável do desempenho preditivo em proteínas inéditas. Assim, essa estratégia constitui um componente essencial para a construção de modelos robustos e estatisticamente válidos em biologia estrutural computacional (Park; Marcotte, 2012).

2.6.2 Multicolinearidade dos descritores

A multicolinearidade consiste na presença de correlação elevada entre variáveis explicativas de um conjunto de dados, caracterizando situações em que dois ou mais descritores carregam informações redundantes ou altamente dependentes (Belsley; Kuh; Welsch, 1980). Esse fenômeno é particularmente frequente em problemas de biologia estrutural, nos quais múltiplos descritores são derivados de propriedades físicas, químicas e geométricas relacionadas ao mesmo nanoambiente proteico.

No contexto da predição de hotspots, descritores estruturais como área acessível ao solvente, densidade de contatos, hidrofobicidade local, potencial eletrostático e métricas de conectividade estrutural tendem a apresentar correlações significativas entre si. Essa redundância emerge do fato de que tais propriedades são consequências diretas da organização tridimensional da proteína e, portanto, refletem aspectos distintos de um mesmo arranjo estrutural. Como resultado, conjuntos de dados com milhares de descritores frequentemente exibem elevada dependência linear ou quase linear entre variáveis.

A presença de multicolinearidade impõe limitações importantes aos modelos de aprendizado de máquina e às análises estatísticas. Em modelos lineares, a multicolinearidade pode resultar em coeficientes instáveis, elevada variância das estimativas e dificuldade de interpretação dos efeitos individuais dos descritores. Mesmo em modelos não lineares, como árvores de decisão e métodos baseados em *ensemble*, a redundância excessiva pode inflar artificialmente a importância de variáveis correlacionadas, mascarando o real sinal preditivo e dificultando a interpretação biológica dos resultados. Do ponto de vista da interpretação biológica, a multicolinearidade dificulta a identificação de determinantes estruturais realmente relevantes. Quando múltiplos descritores altamente correlacionados são mantidos simultaneamente, torna-se inviável atribuir significado funcional claro a variáveis individuais, uma vez que seu efeito está acoplado ao de outras variáveis redundantes. Assim, o controle explícito da multicolinearidade é um requisito essencial para a construção dos modelos de predição de hotspots (Dormann et al., 2013).

2.7 TRABALHOS CORRELATOS E ESTADO DA ARTE

A predição computacional de hotspots em interfaces proteína-proteína tem sido abordada através de diferentes metodologias. Os primeiros métodos fundamentaram-se em cálculos de

energia livre de ligação. Kortemme & Baker (2002) desenvolveram um protocolo baseado em modelagem por homologia e cálculos energéticos usando Rosetta, demonstrando que mutações computacionais podiam predizer hotspots experimentais com precisão de aproximadamente 80% para mutações que causam grandes mudanças em $\Delta\Delta G$. Essa abordagem, embora muito precisa, é computacionalmente custosa e depende da disponibilidade de estruturas tridimensionais de alta qualidade.

Pereira (2012) foi o primeiro autor a utilizar os descritores STING_RDB com os dados experimentais do ASEdb. A seleção de descritores foi realizada através de etapas de forward e backward selection, reduzindo dimensionalmente os 298 descritores iniciais para os 43 descritores finais. O melhor classificador publicado pelo autor possui 49% de precisão, 90% de sensibilidade, 61% de especificidade e F-score de 64%. O autor destaca que o tratamento para a seleção de descritores auxiliou a conquistar os números reportados.

No trabalho de Wang, Liu & Deng (2018), os autores utilizam 600 descritores iniciais reduzidos dimensionalmente para 26 descritores finais mediante seleção em duas etapas (mRMR seguido de *sequential forward selection*). O melhor modelo (PredHS2), baseado em XGBoost e avaliado no conjunto de teste independente (BID), alcançou acurácia de 87%, precisão de 81%, sensibilidade de 77%, especificidade de 92%, F1-score de 79%, MCC de 0,70 e AUC-ROC de 0,831. Os autores destacam que a seleção estatística de descritores e a integração de propriedades de vizinhança estrutural (Euclidiana e Voronoi) contribuíram para o desempenho reportado.

Abordagens baseadas exclusivamente em informações de sequência também têm sido exploradas. O estudo de Preto & Moreira (2020) abordou o problema da predição de hotspots utilizando 173 descritores derivados exclusivamente da sequência primária, incluindo codificação *one-hot* dos aminoácidos, posição relativa na cadeia, propriedades físico-químicas obtidas da literatura e descritores baseados em janelas deslizantes de diferentes tamanhos. O modelo final, baseado no algoritmo *Extremely Randomized Trees* (ExtraTrees) com ajuste de hiperparâmetros via *grid search* e posterior correção de limiares de probabilidade por tipo de aminoácido, foi avaliado sobre um dataset de 534 resíduos (127 hotspots e 407 non-hotspots) de 53 complexos proteicos. Os resultados no conjunto de teste independente incluem acurácia de 82%, AUROC de 0,83, precisão de 91%, sensibilidade de 82% e F1-score de 85%, indicando desempenho competitivo para um método fundamentado exclusivamente em informações da sequência primária.

Nesse mesmo sentido, Sargsyan & Lim (2024) utiliza *embeddings de Protein Language Models* (PLMs) como descritores de seus modelos. Os autores utilizam o *Evolutionary Scale Model*(ESM-2) (Lin et al., 2022) para compor os descritores de seu trabalho. Diferente dos métodos tradicionais, os autores não aplicam a seleção explícita de descritores: os *embeddings* completos são fornecidos diretamente aos modelos supervisionados, que realizam a seleção implícita por meio de suas estruturas internas de árvore. O melhor desempenho foi obtido com ESM-2 combinado ao XGBoost, com precisão de 72%, sensibilidade de 70% e F-score em 71%.

A Tabela 1 sintetiza as principais características metodológicas e os resultados reportados pelos trabalhos correlatos discutidos nesta seção.

Tabela 1 – Síntese comparativa dos trabalhos correlatos.

Trabalho	Desc.	Seleção	Homologia	Prec.	Rec.	AUC
Pereira (2012)	298/43	Fwd/Bwd	NR	0,49	0,90	0,799
Wang, Liu & Deng (2018)	600/26	mRMR+SFS	NR	0,81	0,77	0,831
Preto & Moreira (2020)	173/173	ExtraTrees	NR	0,91	0,82	0,83
Sargsyan & Lim (2024)	ESM-2	Implícita	NR	0,72	0,70	NR
Chen et al. (2024)	–/4	AutoGluon	CD-HIT 60%	0,76	0,67	NR

Fonte: Elaborada pelo autor.

SFS = Sequential Forward Selection. Desc. = Descritores iniciais/finais. ET = ExtraTrees. NR = Não reportado. – = Não aplicável.

A análise comparativa evidencia duas tendências: a crescente sofisticação dos pipelines de seleção de descritores e a ausência recorrente de controle explícito de homologia na maioria dos trabalhos. Apenas Chen et al. (2024) reportaram controle de redundância via CD-HIT com limiar de 60%, protocolo adotado também no presente trabalho. Essa lacuna metodológica reforça a necessidade de abordagens que integrem seleção rigorosa de descritores com prevenção sistemática de vazamento de dados, conforme proposto nesta dissertação.

3 DESENVOLVIMENTO

Neste capítulo é apresentada a metodologia proposta, que compreende cinco etapas principais: (i) integração e preparação dos dados; (ii) filtragem e controle de qualidade, (iii) seleção hierárquica de descritores, (iv) modelagem preditiva e, ao final, (v) avaliação e interpretabilidade.

A Figura 7 apresenta a visão geral da metodologia proposta. À esquerda, o conjunto PPI-Hotspot+PDB^{BM} (918 resíduos de 158 proteínas) passa por controle de redundância via CD-HIT e particionamento com segregação por proteína, resultando em 742 resíduos para treino e 176 para teste. Ao centro, a integração com o STING_RDB é realizada por meio das chaves estruturais (`pdb_id`, `pdb_chain`, `pdb_resno`), que associam cada resíduo rotulado aos seus 2.419 descritores moleculares. À direita, o pipeline hierárquico de redução dimensional reduz progressivamente os descritores: limpeza inicial (1.455), filtragem ID-like (917), constância intraproteína (567), mRMR *in-fold* (494), teste KS com correção FDR (276), remoção de multicolinearidade via VIF (100), expansão por transformações invariantes (162) e seleção final por Boruta (58 descritores, redução total de 97,6%).

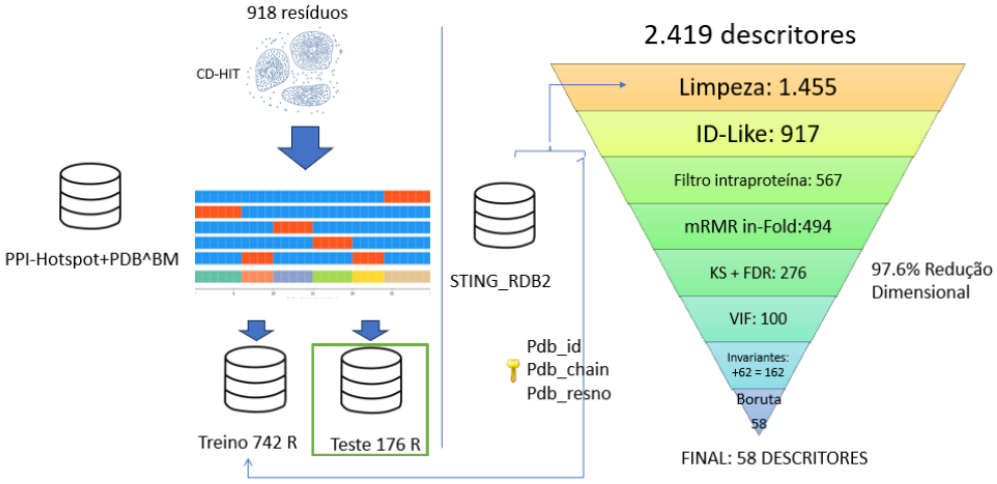


Figura 7 – Visão geral da metodologia proposta. À esquerda, integração do PPI-Hotspot+PDB^{BM} e particionamento por proteína. Ao centro, associação com os descritores do STING_RDB via chaves estruturais. À direita, pipeline hierárquico de redução dimensional (2.419 para 58 descritores, 97,6% de redução).

Fonte: Elaborado pelo autor.

3.1 INTEGRAÇÃO, LIMPEZA E CONTROLE DE QUALIDADE DOS DADOS

Para maximizar a busca de descritores estruturais no banco STING_RDB, optou-se pela utilização do conjunto de dados `PPI-Hotspot+PDB`^{BM}, que compreende 918 resíduos distribuídos em 158 proteínas únicas, com equilíbrio de 45,1% de hotspots e 54,9% de não hotspots.

Antes de aplicar os métodos de seleção de características, realizou-se verificação da curadoria original através de análise de redundância do conjunto de dados por meio do algoritmo CD-HIT (Li; Godzik, 2006). Buscou-se, nessa etapa, garantir a validade da validação cruzada, uma vez que proteínas homólogas distribuídas entre os conjuntos de treino e teste podem causar vazamento de dados, inflando artificialmente as métricas de desempenho.

A Figura 8 ilustra o funcionamento do algoritmo CD-HIT: cada sequência é comparada aos representantes dos clusters existentes e agrupada ao cluster mais similar caso a identidade de sequência supere o limiar definido; caso contrário, a sequência inicia um novo cluster, resultando em um banco de dados não-redundante.

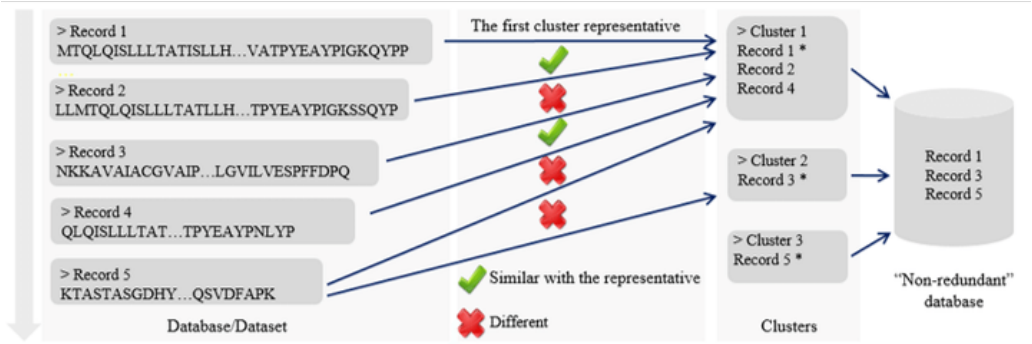


Figura 8 – Paradigma de funcionamento do algoritmo CD-HIT. Sequências são comparadas ao representante de cada cluster e agrupadas por similaridade acima do limiar definido, resultando em um conjunto não-redundante.

Fonte: Adaptado de Chen et al. (2018).

O CD-HIT foi aplicado às 158 proteínas do conjunto **PPI-Hotspot+PDB^{BM}** com limiares de identidade de sequência variando de 90% a 60%, conforme apresentado na Tabela 2.

Tabela 2 – Análise de redundância por CD-HIT.

Ident. (%)	k	Clusters	Sing. (%)
90	5	158	100
80	5	158	100
70	5	158	100
70	4	158	100
60	4	158	100

Fonte: Elaborada pelo autor.

Os resultados demonstram que todas as 158 proteínas constituem sequências únicas, sem agrupamento mesmo no *threshold* mais permissivo de 60% de identidade. Este resultado indica que o conjunto **PPI-Hotspot+PDB^{BM}** apresenta alta diversidade de sequências e ausência de redundância, característica desejável para conjuntos de dados.

(i) **Remoção de descritores com variância zero:** Descritores com valores constantes em todos os resíduos foram identificados e removidos, em grande parte códigos de inserção PDB e metadados normalizados.

(ii) **Filtragem de descritores ID-like:** Descritores com cardinalidade superior a 90% do número de resíduos foram classificados como *ID-like*. Estes dados não carregam informação biológica, distinguem amostras, mas não explicam o fenômeno.

(iii) **Filtragem de descritores constantes por grupo proteico:** Descritores que apresentam variação global mas são constantes dentro de cada proteína individual foram removidos, pois representam características específicas de proteínas ao invés de propriedades generalizáveis de resíduos.

(iv) **Codificação de descritores categóricos:** Aplicou-se *one-hot encoding* para transformar descritores categóricos em formato numérico, preservando a natureza categórica sem introduzir valores artificiais.

3.2 NORMALIZAÇÃO INTRA-PROTEÍNA

Para reduzir viés estrutural entre proteínas de diferentes famílias e tamanhos, aplicou-se transformação z-score calculada independentemente para cada proteína segundo a Equação 1, onde X_{ij} é o valor do descritor j para o resíduo i, $\mu_j^{(p)}$ e $\sigma_j^{(p)}$ são média e desvio-padrão calculados exclusivamente para resíduos da proteína p, e $\epsilon = 10^{-8}$ é constante de estabilização numérica.

$$X_{ij}^{\text{norm}} = \frac{X_{ij} - \mu_j^{(p)}}{\sigma_j^{(p)} + \epsilon} \tag{1}$$

3.3 PARTICIONAMENTO DOS DADOS

Após o processo de limpeza de dados, o conjunto de dados resultante foi particionado em conjuntos de treino e teste utilizando *GroupShuffleSplit* (Pedregosa et al., 2011), com proporção 80% para treino e 20% para teste. O agrupamento foi realizado por proteína individual, utilizando o identificador único de cada proteína (código PDB da cadeia) como critério de grupo. A partição resultou em 742 resíduos de 126 proteínas para treino e 176 resíduos de 32 proteínas para teste.

A escolha do *GroupShuffleSplit* com agrupamento por proteína justifica-se por três razões fundamentais. Primeiro, resíduos pertencentes à mesma proteína compartilham contexto estrutural, composição aminoacídica e propriedades geométricas do nanoambiente, gerando dependência estatística entre as observações. Se resíduos de uma mesma proteína fossem distribuídos entre treino e teste, o modelo poderia explorar essas similaridades intra-proteína para realizar predições no conjunto de teste, configurando vazamento de dados implícito (Park; Marcotte, 2012). Segundo, o cenário realista de aplicação da metodologia é a predição de hotspots em proteínas para as quais não existem dados experimentais de mutagênese. A segregação por proteína simula exatamente esse cenário, avaliando a capacidade do modelo de generalizar para estruturas completamente não observadas durante o treinamento. Terceiro, a proporção 80/20 equilibra dois

requisitos concorrentes: maximizar o volume de dados disponível para treinamento do pipeline hierárquico de seleção de descritores e, simultaneamente, manter um conjunto de teste com número suficiente de proteínas (32) para estimativas minimamente estáveis de desempenho.

O conjunto de teste foi mantido completamente isolado durante todas as etapas subsequentes: seleção de descritores, normalização, treinamento e otimização de hiperparâmetros. Nenhuma informação proveniente das 32 proteínas do teste foi utilizada em qualquer decisão metodológica. Este isolamento rigoroso assegura que as métricas de desempenho reportadas reflitam a genuína capacidade de generalização, e não artefatos decorrentes de contaminação entre conjuntos. O conjunto de teste foi utilizado exclusivamente na etapa final de avaliação, após a conclusão completa do treinamento dos modelos.

3.4 SELEÇÃO HIERÁRQUICA DE DESCRITORES

A seleção de descritores seguiu uma abordagem hierárquica em múltiplas etapas, combinando filtros estatísticos com métodos baseados em importância.

3.4.1 Seleção de Descritores via mRMR In-Fold

Nesta etapa, aplicou-se o algoritmo mRMR (*Minimum Redundancy Maximum Relevance*) (Ding; Peng, 2005), que seleciona os descritores balanceando dois critérios: maximização da relevância em relação à variável-alvo e minimização da redundância entre as variáveis selecionadas. A relevância é quantificada mediante informação mútua, enquanto a redundância é medida pela correlação entre features já selecionadas.

Implementou-se a estratégia *mRMR in-fold*, que realiza seleção independente em cada partição da validação cruzada, diferentemente de abordagens convencionais que aplicam mRMR uma única vez sobre o conjunto completo. O conjunto de treino foi dividido em 5 partições mediante *GroupKFold* (Pedregosa et al., 2011), mantendo proteínas inteiras em cada partição. Para cada iteração, o mRMR foi executado sobre os dados de treino da partição, gerando 5 conjuntos independentes de descritores. Na agregação final, os descritores selecionados em múltiplas partições foram priorizados por demonstrarem robustez preditiva. Esse procedimento mitiga o risco de vazamento de dados e garante sinal preditivo em diferentes subconjuntos dos dados (Saeys; Inza; Larrañaga, 2007).

3.4.2 Filtragem Estatística via Kolmogorov-Smirnov In-Fold

Para validar estatisticamente o sinal preditivo dos descritores selecionados pelo mRMR, aplicou-se o teste de Kolmogorov-Smirnov (KS) (Jr, 1951)com validação cruzada aninhada. O teste KS é um método não-paramétrico que compara distribuições entre dois grupos sem assumir normalidade dos dados (Hollander; Wolfe; Chicken, 2013).

Para cada partição, o teste KS foi aplicado individualmente a cada descritor candidato, comparando as distribuições dos valores daquele descritor entre os grupos de resíduos hotspots e

não-hotspots. O teste foi executado exclusivamente sobre os dados de treino de cada partição. Os valores-p obtidos para todos os descritores em uma mesma iteração foram então submetidos à correção de múltiplos testes pelo método de Benjamini-Hochberg (Benjamini; Hochberg, 1995), controlando a taxa de falsa descoberta (FDR) com limiar de $q \leq 0.10$. Paralelamente ao critério de FDR, foi aplicado um critério complementar de tamanho de efeito, baseado na magnitude da estatística KS, com limiar mínimo de 0.12.

A escolha do limiar de FDR $q \leq 0,10$ justifica-se pelo contexto específico desta etapa da metodologia. O teste KS *in-fold* não constitui a etapa final de seleção, mas um filtro intermediário cuja função é reter descritores com sinal preditivo, delegando a confirmação definitiva de relevância ao algoritmo Boruta na etapa subsequente. Um limiar excessivamente restritivo nesta fase poderia eliminar prematuramente descritores cujo sinal preditivo contribui de forma relevante quando combinado com outros descritores no modelo final. O limiar complementar de tamanho de efeito ($D_{KS} \geq 0{,}12$) foi adotado como salvaguarda contra a rejeição prematura de descritores com diferença de distribuição biologicamente relevante em amostras limitadas. O valor de $0{,}12$ foi estabelecido empiricamente como critério conservador compatível com o tamanho amostral por partição neste estudo, seguindo o princípio geral de que limiares de efeito devem ser calibrados ao contexto da análise (Hollander; Wolfe; Chicken, 2013). Descritores que não atingissem o critério de FDR mas apresentassem magnitude D_{KS} acima desse limiar foram retidos para avaliação na etapa subsequente pelo algoritmo Boruta, que realiza a confirmação definitiva de relevância.

Ao final das 5 iterações de validação cruzada, foi realizada uma etapa de agregação dos resultados. Um descritor foi considerado validado se satisfez o critério de FDR ($q \leq 0.10$) ou o critério de tamanho de efeito (KS ≥ 0.12) (Cohen, 1988) em pelo menos uma das 5 partições. Esse critério de agregação por estabilidade permite identificar descritores que apresentam capacidade discriminativa consistente, mesmo que essa consistência não se manifeste uniformemente em todas as partições dos dados. Descritores validados em ≥ 1 fold foram retidos para as etapas subsequentes.

3.4.3 Remoção de Multicolinearidade via VIF

Para eliminação sistemática da multicolinearidade identificada no conjunto de descritores, empregou-se o Variance Inflation Factor (VIF), uma métrica que quantifica o grau em que a variância de um coeficiente estimado é inflada devido à colinearidade com outros descritores (Kutner et al., 2005). Para cada descritor i, o VIF é calculado como:

$$\text{VIF}_i = \frac{1}{1 - R_i^2} \tag{2}$$

onde R_i^2 representa o coeficiente de determinação obtido ao regredir o descritor i contra todos os demais descritores do conjunto. Um VIF igual a 1 indica ausência de colinearidade, enquanto valores elevados indicam que o descritor pode ser substancialmente explicado pelos demais. Em-

bora não exista um limiar universalmente aceito, valores de VIF superiores a 10 são amplamente reconhecidos como indicativos de multicolinearidade problemática (O'Brien, 2007; Hair et al., 2010).

Neste estudo, o VIF foi aplicado de forma iterativa: em cada ciclo, o descritor com maior VIF acima do limiar de 10 foi removido, e os valores de VIF foram recalculados para os descritores remanescentes. O processo foi repetido até que todos os descritores apresentassem VIF ≤ 10, garantindo a convergência para um conjunto sem multicolinearidade severa. Esta abordagem iterativa é preferível à remoção simultânea, pois a eliminação de um descritor altamente colinear frequentemente reduz o VIF de outros descritores correlacionados, preservando maior quantidade de informação no conjunto final (Dormann et al., 2013).

3.4.4 Transformações Invariantes de Domínio

Nesta etapa, foram criados descritores adicionais projetados para serem invariantes a diferenças entre proteínas. Essas transformações são computadas de forma independente para cada proteína, eliminando as dependências interproteicas que poderiam amplificar o *covariate shift* (Ioffe; Szegedy, 2015):

(i) **Proporções intra-proteína:** Normalizam o valor de cada resíduo pela soma absoluta dos valores na proteína, seguindo princípios estabelecidos para normalização de dados proteômicos quantitativos (Karpievitch; Dabney; Smith, 2012; Välikangas; Suomi; Elo, 2018):

$$X_i^{\text{prop}} = \frac{X_i}{\sum_{k \in p} |X_k| + \epsilon} \tag{3}$$

(ii) **Ranks intra-proteína:** Convertem valores absolutos em posições relativas dentro de cada proteína, uma abordagem robusta a *outliers* e amplamente validada em análises de expressão diferencial (Breitling et al., 2004):

$$X_i^{\text{rank}} = \frac{\text{rank}(X_i, p) - 1}{\max(|p| - 1, 1)} \tag{4}$$

(iii) **Escala Min-Max intra-proteína:** Normalizam para o intervalo [0,1] dentro de cada proteína, preservando as relações relativas entre resíduos conforme metodologias consagradas para dados de alta dimensão (Bolstad et al., 2003):

$$X_i^{\text{minmax}} = \frac{X_i - \min_{k \in p} X_k}{\max_{k \in p} X_k - \min_{k \in p} X_k + \epsilon} \tag{5}$$

(iv) **Razões entre descritores:** Criaram-se razões biologicamente relevantes entre pares de descritores correlacionados, capturando relações proporcionais independentes de escala, uma estratégia eficaz em engenharia de *features* para sequências proteicas (Lim et al., 2022; Bonidia et al., 2022).

3.4.5 Seleção Baseada em Importância via Boruta

Após as etapas de remoção de multicolinearidade e normalização intra-proteína, aplicou-se o algoritmo Boruta (Kursa; Rudnicki, 2010) para identificar os descritores genuinamente relevantes para a discriminação de hotspots. O Boruta é um método de seleção de variáveis que utiliza o algoritmo Random Forest (Breiman, 2001a) como estimador interno para calcular a importância dos descritores, aplicando um critério estatístico para distinguir sinais verdadeiros de ruído.

O algoritmo fundamenta-se na criação de *shadow features*, cópias permutadas aleatoriamente dos descritores originais que destroem qualquer associação genuína com a variável alvo. Em cada iteração, um modelo Random Forest é treinado sobre o conjunto expandido (descritores originais e sombras), e a importância de cada descritor original é comparada estatisticamente com a importância máxima observada entre os *shadow features*. Descritores consistentemente superiores a este limiar adaptativo são classificados como relevantes; aqueles consistentemente inferiores são rejeitados (Kursa; Rudnicki, 2010).

Esta abordagem apresenta vantagens sobre métodos baseados em limiares arbitrários de importância, pois adapta-se automaticamente às características do conjunto de dados e identifica todos os descritores relevantes, não apenas os mais importantes. No presente estudo, o algoritmo foi configurado com Random Forest como estimador base ($n\_estimators = 100$, $max\_depth = 7$), limiar de percentil 95 e máximo de 100 iterações. Apenas os descritores estatisticamente confirmados foram retidos para treinamento dos modelos finais.

3.5 MODELAGEM E AVALIAÇÃO

Esta seção descreve a etapa final do pipeline metodológico, na qual os 58 descritores selecionados são utilizados para construir classificadores capazes de discriminar hotspots de resíduos regulares de interface. Apresentam-se os algoritmos de aprendizado supervisionado empregados, o protocolo de treinamento e validação, as métricas de avaliação adotadas e o modelo baseline utilizado como referência comparativa.

3.5.1 Algoritmos de Classificação

Três algoritmos de aprendizado supervisionado foram selecionados para avaliação, cada um representando uma abordagem arquitetural distinta para o problema de classificação dos hotspots:

(i) **Random Forest (RF):** Método de ensemble baseado em agregação por bootstrap (*bagging*) de múltiplas árvores de decisão (Breiman, 2001a). A combinação por votação majoritária reduz a variância das predições, conferindo robustez a ruído e outliers. Random Forest é particularmente adequado para conjuntos de dados com muitos descritores, pois a seleção aleatória de variáveis em cada nó mitiga os efeitos de correlações residuais entre descritores.

(ii) **Support Vector Machine (SVM):** Algoritmo que busca o hiperplano de separação com margem máxima entre classes (Cortes; Vapnik, 1995). A utilização do kernel RBF (Radial Basis Function) permite capturar fronteiras de decisão não-lineares através do mapeamento implícito dos dados para um espaço de dimensionalidade superior. SVM é eficaz em espaços de alta dimensionalidade e robusto a sobreajuste quando o número de descritores é comparável ao número de amostras.

(iii) **XGBoost:** Implementação otimizada de gradient boosting que constrói árvores de decisão de forma sequencial, onde cada árvore corrige os erros residuais das anteriores (Chen; Guestrin, 2016a). O algoritmo incorpora regularização L1 e L2 nos pesos das folhas, controlando a complexidade do modelo e reduzindo o risco de sobreajuste. XGBoost é reconhecido por seu desempenho competitivo em problemas tabulares com classes desbalanceadas.

A seleção desses três algoritmos permite avaliar o desempenho de abordagens fundamentalmente diferentes: agregação paralela de modelos fracos (RF), maximização de margem em espaço transformado (SVM) e boosting sequencial com regularização (XGBoost). Essa diversidade arquitetural possibilita identificar se os padrões discriminativos de hotspots são capturados de forma consistente independentemente do algoritmo, o que indicaria robustez dos descritores selecionados.

3.5.2 Treinamento e Validação

A avaliação de desempenho empregou validação cruzada estratificada por proteínas via *GroupKFold* com $k = 5$ partições (Pedregosa et al., 2011). A escolha do *GroupKFold* também nesta etapa, justifica-se por duas razões complementares. Primeiro, garante consistência metodológica ao longo de todo o pipeline, evitando que diferentes critérios de particionamento introduzam vieses nas comparações entre etapas. Segundo, assegura que as métricas de validação reflitam o cenário realista de aplicação: predição em proteínas não observadas durante o treinamento. Se a validação cruzada permitisse que resíduos de uma mesma proteína aparecessem simultaneamente em folds de treino e validação, o modelo poderia explorar similaridades intra-proteína, inflando artificialmente as métricas e selecionando hiperparâmetros inadequados para generalização (Park; Marcotte, 2012).

Duas estratégias de otimização foram avaliadas para investigar o impacto do ajuste de hiperparâmetros sobre o desempenho preditivo:

(i) **Pipeline GroupKFold:** Emprega os hiperparâmetros padrão definidos pelas implementações dos classificadores no *scikit-learn* (Pedregosa et al., 2011), sem qualquer ajuste adicional. Esta estratégia prioriza simplicidade, reprodutibilidade e menor risco de sobreajuste, servindo como linha de base para avaliar se a otimização de hiperparâmetros proporciona ganhos substantivos.

(ii) **Pipeline GridSearchCV:** Incorpora *GridSearchCV* aninhado para otimização sistemática de hiperparâmetros, seguindo metodologia semelhante à de Wang, Liu & Deng (2018). O *GridSearchCV* opera internamente ao loop de validação cruzada, testando múltiplas combinações de parâmetros e selecionando aquela com melhor desempenho médio. O uso do *GroupKFold* como estratégia de particionamento dentro do *GridSearchCV* garante que a seleção de hiperparâmetros seja guiada por métricas calculadas em proteínas distintas das utilizadas no treino, evitando a escolha de configurações que apenas memorizem padrões proteína-específicos.

A Figura 9 apresenta a arquitetura dos dois pipelines avaliados. Ambos compartilham os mesmos dados de treino (742 resíduos), o mesmo conjunto de 58 descritores e a mesma estratégia de validação cruzada por proteína (GroupKFold, $k = 5$). A distinção reside exclusivamente na etapa de otimização de hiperparâmetros, presente apenas no Pipeline GridSearchCV (destacada em verde), permitindo isolar o efeito do ajuste paramétrico sobre o desempenho preditivo.

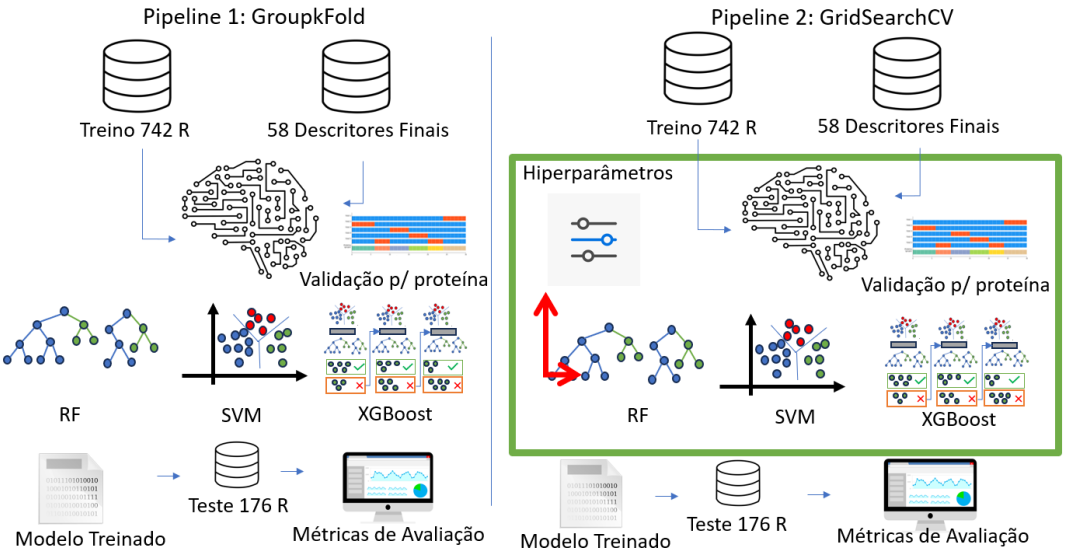


Figura 9 – Comparação entre os dois pipelines de avaliação. Pipeline 1 (GroupKFold): treinamento com hiperparâmetros padrão. Pipeline 2 (GridSearchCV): incorpora busca exaustiva de hiperparâmetros (destacada em verde) antes do treinamento final. Ambos utilizam os mesmos 58 descritores, validação cruzada estratificada por proteína e avaliação no conjunto de teste independente (176 resíduos).

Fonte: Elaborado pelo autor.

A comparação entre as duas estratégias permite quantificar o benefício da otimização de hiperparâmetros quando a metodologia de seleção de descritores já incorpora controles rigorosos de generalização.

3.5.3 Métricas de Avaliação

Métricas clássicas de desempenho para classificação binária com classes moderadamente balanceadas foram adotadas:

- **AUC-ROC:** Área sob a curva ROC, medindo capacidade discriminativa independente do limiar de decisão (Fawcett, 2006).
- **MCC (Matthews Correlation Coefficient):** Métrica balanceada que considera todos os elementos da matriz de confusão, especialmente informativa para datasets com desbalanceamento moderado (Matthews, 1975; Chicco; Jurman, 2020).
- **F1-Score:** Média harmônica entre precisão e sensibilidade (Sokolova; Lapalme, 2009).
- **Precisão e Sensibilidade:** Métricas complementares para análise detalhada do trade-off entre identificação correta de hotspots e minimização de falsos positivos (Sokolova; Lapalme, 2009).

A interpretabilidade dos modelos foi avaliada por meio da análise SHAP (*SHapley Additive exPlanations*) (Lundberg; Lee, 2017). Os valores SHAP são fundamentados na teoria dos jogos cooperativos, calculando a contribuição de cada descritor para predições individuais considerando todas as possíveis combinações de descritores. Esta abordagem fornece notas de importância consistentes e localmente precisos, permitindo identificar os descritores mais relevantes para a caracterização de hotspots e oferecendo explicações tanto globais quanto específicas para cada instância.

3.5.4 Modelo Baseline

Para estabelecer uma referência comparativa e quantificar o ganho proporcionado pela metodologia proposta, desenvolveu-se um modelo baseline utilizando cinco descritores clássicos frequentemente empregados em estudos de predição de hotspots:

- Acessibilidade relativa ao solvente (*Air_acc_rsa_naccess*)
- Razão de área de superfície enterrada (*Air_bsa_ratio_naccess*)
- Energia total de contatos inter-residuais (*Residue_Contacts_total_energy*)
- Densidade atômica local em raio de 5 Å (*Density_Sponge_density_CA_5*)
- Indicador de contato na interface (*Contact_ifr*)

Esses descritores representam os principais determinantes estruturais de hotspots segundo as teorias O-Ring (Bogan; Thorn, 1998) e de dupla exclusão de água (Li; Liu, 2009b). O pré-processamento consistiu em imputação de valores ausentes pela média, seguida de padronização por *StandardScaler*, ambos ajustados exclusivamente no conjunto de treino. A partição dos dados seguiu o mesmo protocolo da metodologia proposta, com divisão estratificada por proteína (80% treino, 20% teste) e validação cruzada via *GroupKFold* com cinco partições. Três algoritmos foram avaliados com hiperparâmetros padrão: Random Forest, SVM-RBF e XGBoost.

A comparação entre o baseline e o pipeline proposto permite isolar a contribuição da engenharia de features e do processo de seleção hierárquica para o desempenho preditivo final.

3.6 RESULTADOS

3.6.1 Redução Dimensional e Seleção de Descritores

O conjunto de dados inicial foi integrado resultando em 2.419 descritores para 918 resíduos do PPI−Hotspot+PDB^{BM}. A Tabela 3 sumariza o impacto de cada etapa da metodologia hierárquica de filtragem.

Tabela 3 – Pipeline hierárquico de redução dimensional.

Etapa	Entrada	Saída	Redução (%)
Limpeza inicial	2.419	1.455	39,8
Filtro ID-like	1.455	917	37,0
Constância intra-proteína	917	567	38,2
mRMR In-Fold	567	494	12,9
KS In-Fold	494	276	44,1
VIF iterativo	276	100	63,8
Transf. invariantes	100	162	–
Boruta	162	58	64,2
Total	2.419	58	97,6

Fonte: Elaborada pelo autor.

A limpeza inicial partiu de 2.419 descritores do STING_RDB e removeu: (i) 246 descritores de conservação evolutiva baseados em alinhamentos múltiplos de sequências (HSSP, entropia, pressão evolutiva); (ii) 12 descritores contendo identificadores únicos de átomos; (iii) 355 descritores com variância zero; (iv) 248 descritores redundantes; e (v) 103 descritores com valores ausentes, resultando em 1.455 descritores para as etapas subsequentes. O critério ID-like (nunique$/N > 0,9$) removeu 538 descritores, incluindo coordenadas atômicas e métricas com valores praticamente únicos por resíduo. O critério de constância intra-proteína removeu 350 descritores que variam entre proteínas mas permanecem constantes dentro de cada estrutura, produzindo 567 descritores.

A seleção mRMR in-fold com GroupKFold reteve 494 descritores presentes em pelo menos 2 folds, com 236 descritores aparecendo em todos os 5 folds. A validação estatística via teste de Kolmogorov-Smirnov (FDR $q \leq 0,1$, $D_{KS} \geq 0,12$) reteve 276 descritores, removendo 218 que não discriminavam adequadamente entre hotspots e não-hotspots.

O procedimento iterativo de remoção por VIF convergiu após 176 iterações, reduzindo de 276 para 100 descritores com $\text{VIF}_{\text{max}} = 9,97$. A Figura Figura 10 ilustra o processo de convergência.

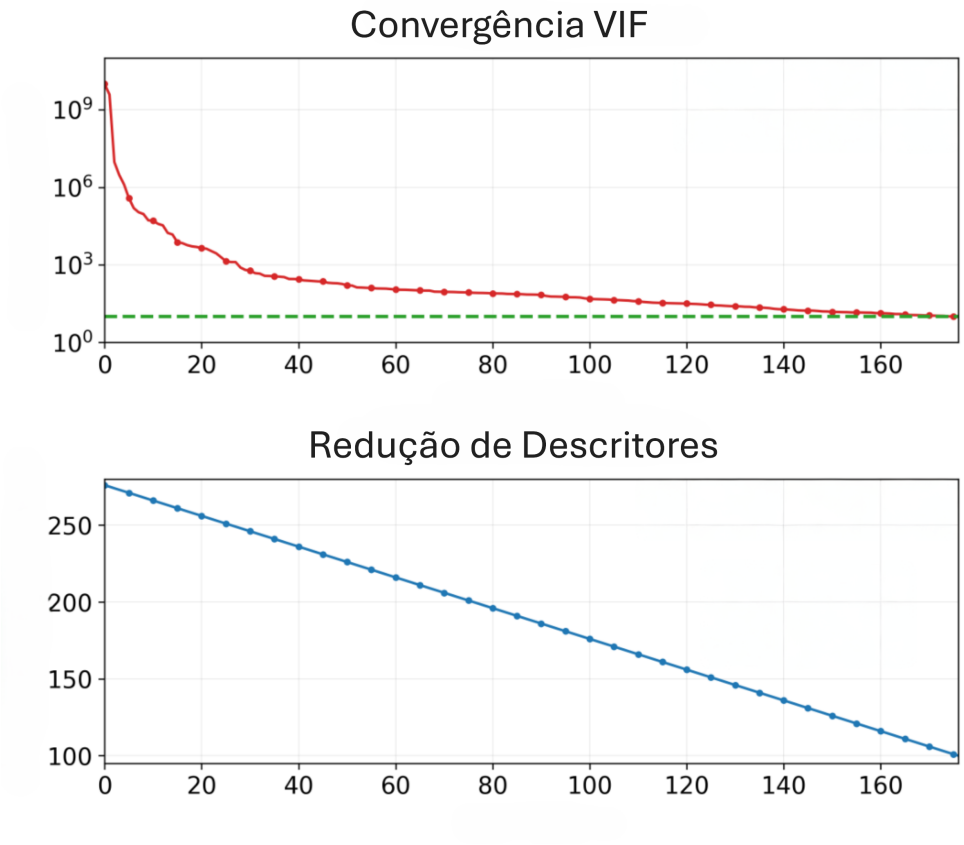


Figura 10 – Convergência do VIF durante remoção iterativa de multicolinearidade.

Fonte: Elaborado pelo autor.

Finalmente, a seleção de descritores Boruta identificou 58 descritores como relevantes, constituindo o conjunto final de descritores para treinamento dos modelos.

3.6.2 Descritores Selecionados pelo Boruta

A criação de descritores invariantes de domínio expandiu o conjunto de 100 para 162 descritores. O algoritmo Boruta convergiu confirmando 58 descritores como estatisticamente relevantes, dos quais 17 (29,3%) são descritores originais e 41 (70,7%) são transformações invariantes de domínio.

A Tabela 4 apresenta os 17 descritores originais em sua ordem de seleção, conforme determinado pelo algoritmo Boruta.

Tabela 4 – Descritores originais selecionados pelo algoritmo Boruta.

Rank	Descritor	Categoria
1	Unused_Contacts_WNA_hydrophobic_uc	Contato
2	Unused_Contacts_GN_hydrophobic_uc	Contato
3	Graph_Descriptor_WNA_rwb	Grafos
4	Cross_Link_Order_WNA_clo_5_30_CB	Interação
5	Unused_Contacts_hydrophobic_uc	Contato
6	Weighted_Contact_Number_WNA_k_5	Contato
7	Density_Sponge_SW_sponge_CA_4	Superfície
8	Graph_Descriptor_GN_bottle_neck	Grafos
9	Cross_Presence_Order_WNA_cpo_35_30	Interação
10	Hydrophobicity_kite_doolittle	Físico-químico
11	Residue_Contacts_WNA_aromatic	Contato
12	Density_Sponge_GN_density_CA_4	Superfície
13	Graph_Descriptor_SW_rwb_9	Grafos
14	Cross_Link_Order_VD_clo_35_15_CB	Interação
15	Cross_Presence_Order_WNA_cpo_35_15	Interação
16	Cross_Link_Order_clo_5_15_CA	Interação
17	Graph_Descriptor_betweenness	Grafos

Fonte: Elaborada pelo autor.

A Tabela 5 apresenta a distribuição categórica do conjunto final dos descritores.

Tabela 5 – Distribuição categórica das features após transformações invariantes.

Categoria	Originais	Transformadas	Total (%)
Interação	5	12	29,3
Contato	5	12	29,3
Físico-químicos	1	9	17,2
Grafos	4	5	15,5
Superfície	2	3	8,6
Total	17	41	100

Fonte: Elaborada pelo autor.

3.6.3 Desempenho dos Modelos de Predição

3.6.3.1 Pipeline GroupKFold

O primeiro pipeline implementou validação cruzada estratificada por grupos (*GroupKFold*, $k = 5$) sem otimização de hiperparâmetros, utilizando parâmetros padrão dos algoritmos. Este protocolo garante que proteínas utilizadas no treinamento não apareçam no conjunto de validação interna, simulando um cenário de predição em proteínas completamente novas. A Tabela 6 apresenta os resultados de validação cruzada para os três algoritmos avaliados.

Tabela 6 – Resultados de validação cruzada (Pipeline GroupKFold).

Modelo	AUC	F1	MCC
Random Forest	0,722 $\pm$ 0,086	0,640 $\pm$ 0,053	0,343 $\pm$ 0,098
SVM	0,671 $\pm$ 0,069	0,602 $\pm$ 0,018	0,266 $\pm$ 0,046
XGBoost	0,716 $\pm$ 0,055	0,626 $\pm$ 0,024	0,304 $\pm$ 0,068

Fonte: Elaborada pelo autor.

Na validação cruzada, o Random Forest apresentou o melhor desempenho médio (AUC = 0,722 $\pm$ 0,086), seguido pelo XGBoost (AUC = 0,716 $\pm$ 0,055) e SVM (AUC = 0,671 $\pm$ 0,069). Entretanto, o Random Forest também apresentou a maior variabilidade entre as partições, com desvio padrão de 0,086 no AUC, indicando sensibilidade à composição específica dos conjuntos de treino e validação. O XGBoost demonstrou maior estabilidade, com desvio padrão de apenas 0,055 no AUC, sugerindo maior robustez às particularidades de cada partição. Essa variabilidade entre partições é esperada quando o número de proteínas é limitado, especialmente se proteínas com características distintivas forem isoladas em uma única partição.

A Tabela 7 apresenta os resultados obtidos no conjunto de teste independente, que não foi utilizado em nenhuma etapa de treinamento ou seleção de descritores.

Tabela 7 – Resultados do Pipeline GroupKFold no conjunto de teste independente.

Modelo	AUC	Precisão	Sensibilidade	F1	MCC
Random Forest	0,818	0,671	0,734	0,702	0,520
SVM	0,729	0,684	0,609	0,645	0,461
XGBoost	0,795	0,651	0,641	0,646	0,446

Fonte: Elaborada pelo autor.

No conjunto de teste independente, o Random Forest alcançou AUC-ROC = 0,818, MCC = 0,520 e F1-Score = 0,702, estabelecendo-se como o melhor modelo desta abordagem. O desempenho superior ao obtido na validação cruzada (+13,3% no AUC) sugere que o conjunto de teste, embora independente, possui características estruturais ligeiramente mais favoráveis à discriminação hotspot/non-hotspot do que a média observada nas partições de validação cruzada. A precisão de 67,1% indica que aproximadamente dois terços dos resíduos preditos como hotspots são verdadeiramente hotspots, enquanto a sensibilidade de 73,4% mostra que o modelo identifica corretamente cerca de três quartos dos hotspots reais presentes no conjunto de teste.

3.6.3.2 Pipeline GridSearchCV

O segundo pipeline incorporou otimização de hiperparâmetros via *GridSearchCV*, mantendo a mesma metodologia de validação cruzada estratificada por grupos. A otimização foi conduzida mediante busca exaustiva em grade pré-definida de hiperparâmetros, avaliando cada combinação através de validação cruzada interna (*GroupKFold*, $k = 5$) sobre o conjunto de treino. A combinação de hiperparâmetros que apresentou o melhor desempenho médio nas 5 partições foi selecionada, e o modelo final foi treinado utilizando as 126 proteínas do conjunto de treino

com esses hiperparâmetros otimizados. A Tabela 8 apresenta os melhores hiperparâmetros identificados para cada algoritmo.

Tabela 8 – Hiperparâmetros otimizados via GridSearchCV.

Parâmetro	Random Forest	SVM	XGBoost
max_depth	10	–	7
n_estimators	100	–	300
min_samples_split	5	–	–
learning_rate	–	–	0,01
C / gamma	–	0,1 / 0,01	–
kernel	–	rbf	–
class_weight	balanced	balanced	–
scale_pos_weight	–	–	1,12

Fonte: Elaborada pelo autor.

A configuração otimizada do Random Forest adotou profundidade máxima limitada em 10 níveis e um número moderado de estimadores de 100 árvores, parâmetros que favorecem a generalização em detrimento do ajuste perfeito aos dados de treino. Para o SVM, o kernel RBF foi selecionado com parâmetros de regularização conservadores (C = 0,1, gamma = 0,01), indicando que o espaço de descritores original já proporciona separabilidade adequada sem necessidade de transformações não-lineares agressivas. O XGBoost convergiu para uma configuração com 300 estimadores e uma taxa de aprendizado baixa (0,01), estratégia que privilegia a construção gradual do ensemble e reduz o risco de vazamento de dados. O parâmetro `scale_pos_weight = 1,12` reflete o leve desbalanceamento entre classes (350 hotspots vs 392 non-hotspots, razão 1,12).

A Tabela 9 apresenta os resultados no conjunto de teste independente.

Tabela 9 – Resultados do Pipeline GridSearchCV no conjunto de teste independente.

Modelo	AUC	Precisão	Sensibilidade	F1	MCC
Random Forest	0,818	0,632	0,750	0,686	0,486
SVM (RBF)	0,821	0,581	0,844	0,688	0,478
XGBoost	0,806	0,671	0,766	0,715	0,538

Fonte: Elaborada pelo autor.

O SVM com kernel RBF obteve AUC-ROC = 0,821, representando o melhor desempenho em capacidade discriminativa entre os modelos avaliados. Nota-se, entretanto, um trade-off entre precisão e sensibilidade: o SVM priorizou sensibilidade elevada (84,4%), capturando a maior fração de hotspots reais, ao custo de precisão reduzida (58,1%), resultando em maior taxa de falsos positivos. O XGBoost apresentou o melhor equilíbrio global, alcançando MCC = 0,538 e F1-Score = 0,715, métricas que ponderam adequadamente tanto verdadeiros positivos quanto verdadeiros negativos. Este desempenho equilibrado do XGBoost sugere que o algoritmo aprendeu a discriminar hotspots de forma mais robusta, sem viés excessivo para nenhuma das classes.

3.6.3.3 Comparação entre Pipelines

A Tabela 10 consolida o desempenho dos melhores modelos de cada abordagem metodológica.

Tabela 10 – Comparação entre pipelines (melhor modelo por métrica).

Pipeline	Modelo	AUC	MCC	Nº de features
GroupKFold	Random Forest	0,818	0,520	58
GridSearchCV	SVM (RBF)	0,821	0,478	58
GridSearchCV	XGBoost	0,806	0,538	58

Fonte: Elaborada pelo autor.

A otimização de hiperparâmetros do Pipeline GridSearchCV resultou em ganho marginal de apenas +0,4% no AUC (de 0,818 para 0,821) em relação ao Pipeline GroupKFold com parâmetros padrão. Este ganho modesto indica que os parâmetros padrão dos algoritmos já proporcionam desempenho próximo ao ótimo para este conjunto de dados, e que o principal determinante da capacidade preditiva reside na qualidade dos 58 descritores selecionados pelo pipeline hierárquico, não nos hiperparâmetros específicos dos modelos. Ambos os pipelines utilizaram exatamente o mesmo conjunto de features (100% de sobreposição), diferindo apenas na estratégia de ajuste de parâmetros.

3.6.3.4 Comparação com Modelo Baseline

Para quantificar o ganho proporcionado pelo pipeline hierárquico de seleção de descritores, os mesmos três algoritmos foram avaliados utilizando apenas cinco descritores clássicos fundamentados nas teorias de formação de hotspots:

- Acessibilidade relativa ao solvente (*Air_acc_rsa_naccess*);
- Razão de área de superfície enterrada (*Air_bsa_ratio_naccess*);
- Energia total de contatos inter-residuais (*Residue_Contacts_total_energy*);
- Densidade atômica local em raio de 5 Å (*Density_Sponge_density_CA_5*);
- Indicador de contato na interface (*Contact_ifr*).

Esses descritores representam as características citadas na literatura como determinantes de hotspots (Moreira et al., 2017; Bogan; Thorn, 1998).

A Tabela 11 apresenta os resultados de validação cruzada do modelo baseline.

Tabela 11 – Resultados de validação cruzada do modelo baseline (5 descritores).

Modelo	AUC	F1	MCC
Random Forest	0,518 ± 0,024	0,448 ± 0,054	0,036 ± 0,028
SVM-RBF	0,505 ± 0,042	0,456 ± 0,044	0,004 ± 0,060
XGBoost	0,510 ± 0,016	0,474 ± 0,057	0,021 ± 0,035

Fonte: Elaborada pelo autor.

Na validação cruzada, o modelo baseline apresentou desempenho próximo ao acaso (AUC ≈ 0,5), indicando ausência de capacidade discriminativa consistente. O valor de MCC próximo a zero (variando entre 0,004 e 0,036) confirma que as predições não são superiores a uma classificação aleatória. Este resultado é particularmente revelador, pois os cinco descritores utilizados são frequentemente citados como os mais relevantes para caracterização de hotspots. A incapacidade desses descritores, quando utilizados isoladamente, de discriminar hotspots sugere que o fenômeno de formação de hotspots não pode ser adequadamente capturado por um conjunto reduzido de características, mesmo que teoricamente fundamentadas.

A Tabela 12 apresenta os resultados no conjunto de teste independente.

Tabela 12 – Resultados do modelo baseline no conjunto de teste independente.

Modelo	AUC	Precisão	Sensibilidade	F1	MCC
Random Forest	0,530	0,478	0,443	0,460	0,077
SVM-RBF	0,426	0,486	0,546	0,515	0,108
XGBoost	0,485	0,398	0,423	0,410	-0,061

Fonte: Elaborada pelo autor.

No conjunto de teste, o desempenho permaneceu próximo ao acaso, com o melhor modelo (Random Forest) alcançando apenas AUC = 0,530. Notavelmente, o XGBoost apresentou MCC negativo (-0,061), indicando que suas predições foram sistematicamente piores que uma classificação aleatória. Este resultado contrasta fortemente com o desempenho observado quando o mesmo algoritmo utiliza os 58 descritores selecionados (AUC = 0,806, MCC = 0,538), evidenciando que a arquitetura do algoritmo de aprendizado é secundária em importância quando comparada à qualidade e relevância das features utilizadas.

A Tabela 13 quantifica os ganhos obtidos pelo pipeline hierárquico em relação ao baseline.

Tabela 13 – Comparação entre baseline e pipeline proposto (conjunto de teste).

Abordagem	Modelo	Features	AUC	MCC	Δ AUC
Baseline	Random Forest	5	0,530	0,077	–
Pipeline GroupKFold	Random Forest	58	0,818	0,520	+54,3%
Pipeline GridSearchCV	SVM-RBF	58	0,821	0,478	+54,9%
Pipeline GridSearchCV	XGBoost	58	0,806	0,538	+52,1%

Fonte: Elaborada pelo autor.

O pipeline proposto demonstrou ganhos substanciais em relação ao baseline: o AUC-ROC aumentou de 0,530 para 0,821 (ganho relativo de +54,9%), enquanto o MCC evoluiu de 0,077

para 0,538 (ganho absoluto de +0,461, representando aumento de 599% em termos relativos). Esses resultados evidenciam a eficácia da estratégia de engenharia de features baseada em descritores invariantes de domínio combinada com o processo hierárquico de seleção.

O desempenho próximo ao acaso do baseline confirma observações prévias na literatura de que descritores isolados, embora fundamentados teoricamente, capturam apenas aspectos parciais do fenômeno de formação de hotspots (Tuncbag; Gursoy; Keskin, 2009). A discriminação efetiva requer a integração de informações complementares sobre múltiplos aspectos do nanoambiente molecular. Os 58 descritores retidos pelo pipeline hierárquico representam essa integração multiescala de informação estrutural.

3.6.4 Análise de Generalização

A capacidade de generalização foi avaliada comparando o desempenho obtido durante validação cruzada com o desempenho observado no conjunto de teste independente. A Tabela 14 apresenta esta comparação para todos os modelos avaliados.

Tabela 14 – Análise de generalização: validação cruzada versus teste independente.

Pipeline	Modelo	Treino (AUC)	Teste (AUC)	Δ (%)
GroupKFold	Random Forest	0,722	0,818	+13,3
GroupKFold	SVM	0,671	0,729	+8,7
GroupKFold	XGBoost	0,716	0,795	+10,9
GridSearchCV	Random Forest	0,738	0,818	+10,9
GridSearchCV	SVM (RBF)	0,720	0,821	+14,1
GridSearchCV	XGBoost	0,732	0,806	+10,1

Fonte: Elaborada pelo autor.

Os modelos apresentaram desempenho no conjunto de teste independente numericamente superior às estimativas de validação cruzada. A comparação permite afirmar que não houve degradação severa de desempenho entre validação cruzada e teste independente. A capacidade preditiva consistente em ambas as avaliações (AUC $> 0{,}72$ na validação cruzada, AUC $> 0{,}80$ no teste independente) é compatível com um modelo que generaliza para proteínas estruturalmente distintas das observadas durante o treinamento, embora a confirmação definitiva desse comportamento requeira avaliação em conjuntos de teste maiores.

As Figuras Figura 11 e Figura 12 apresentam as curvas ROC comparativas obtidas no conjunto de teste independente para ambos os pipelines.

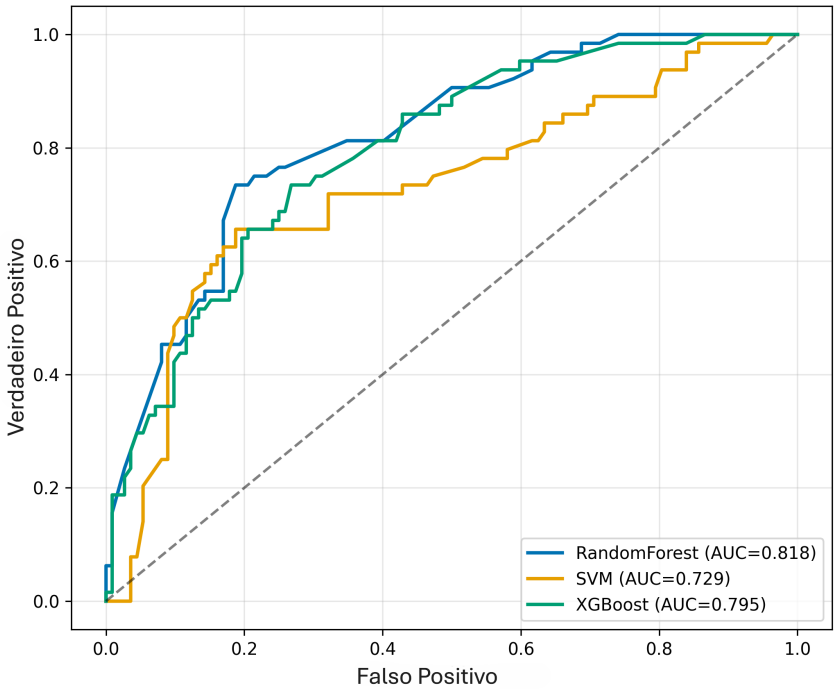


Figura 11 – Curvas ROC no conjunto de teste independente – Pipeline GroupKFold. Random Forest (AUC = 0,818), XGBoost (AUC = 0,795) e SVM (AUC = 0,729).

Fonte: Elaborado pelo autor.

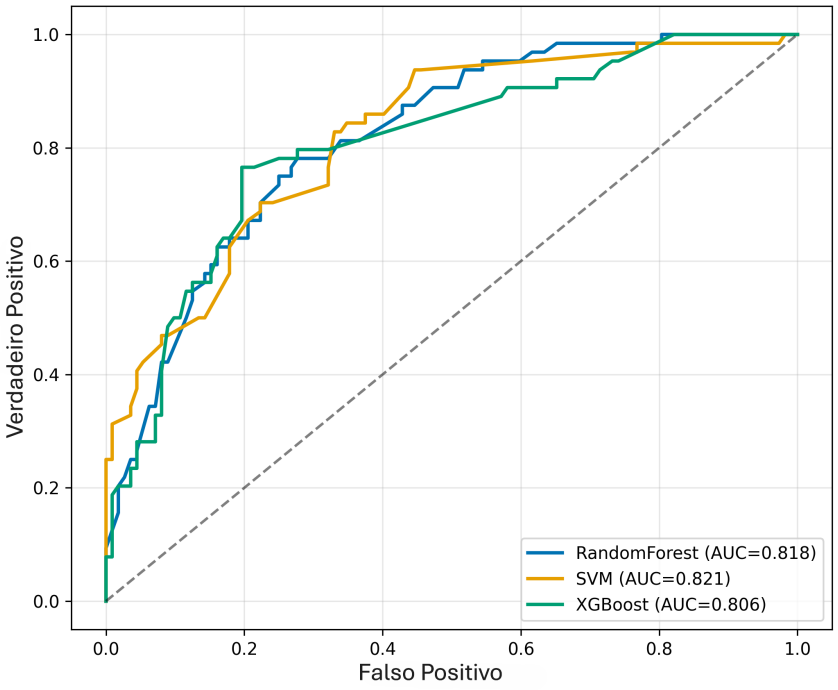


Figura 12 – Curvas ROC no conjunto de teste independente – Pipeline GridSearchCV. SVM (AUC = 0,821), Random Forest (AUC = 0,818) e XGBoost (AUC = 0,806).

Fonte: Elaborado pelo autor.

As curvas ROC demonstram que todos os modelos apresentam capacidade discriminativa substancialmente superior ao classificador aleatório (linha diagonal). A proximidade entre as curvas dos diferentes algoritmos (variação de apenas 9,2 pontos percentuais entre o melhor e o pior modelo) reforça a conclusão de que a qualidade dos descritores selecionados é o principal determinante do desempenho, com impacto secundário da escolha do algoritmo de aprendizado.

3.6.5 Interpretabilidade via Análise SHAP

A análise SHAP (*SHapley Additive exPlanations*) (Lundberg; Lee, 2017) foi aplicada aos modelos finais para quantificar a contribuição individual de cada descritor para as predições. Os valores SHAP decompõem a predição de cada amostra em contribuições aditivas de cada descritor, fundamentados na teoria de jogos cooperativos. Valores SHAP positivos indicam que o descritor contribui para predição de hotspot, enquanto valores negativos favorecem predição de non-hotspot.

3.6.5.1 Descritores Mais Relevantes

A Figura 13 e Figura 14 apresentam a importância média absoluta SHAP para os 20 descritores mais relevantes nos modelos de melhor desempenho de cada pipeline.

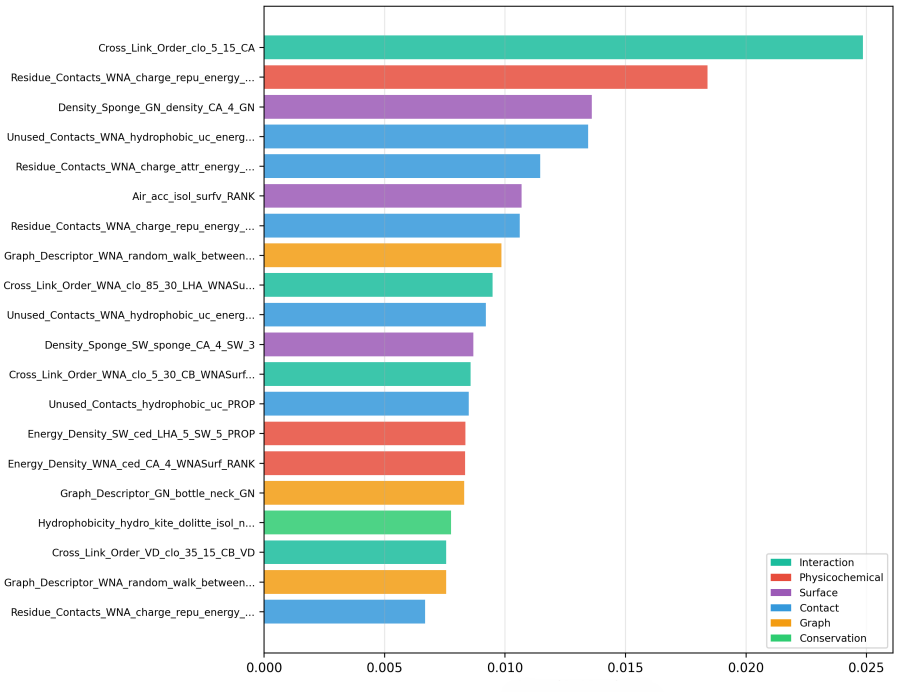


Figura 13 – Importância média absoluta SHAP para os 20 descritores mais relevantes no modelo Random Forest (Pipeline GroupKFold).

Fonte: Elaborado pelo autor.

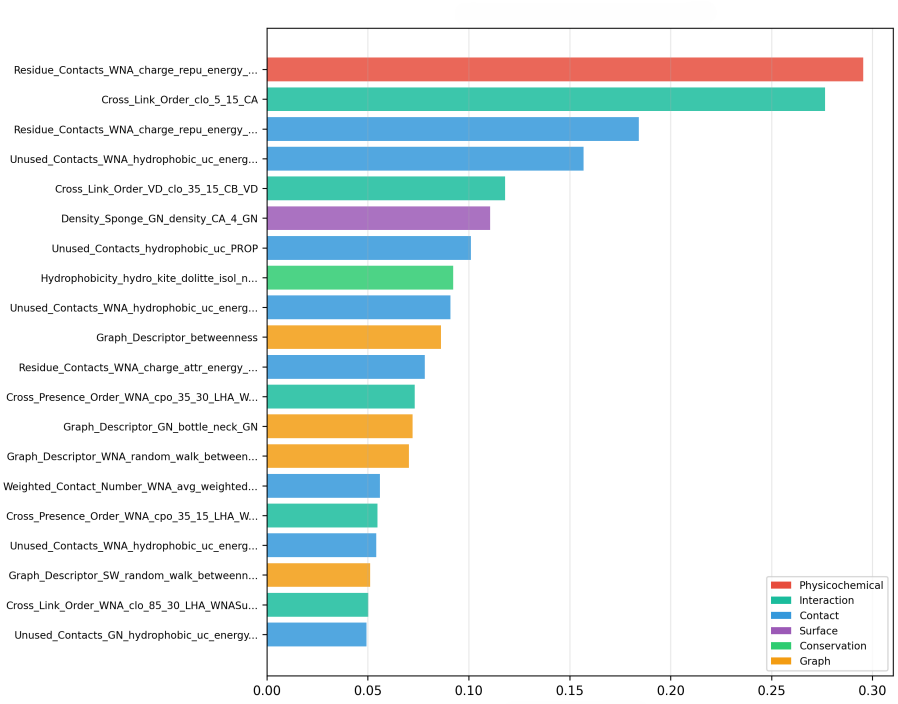


Figura 14 – Importância média absoluta SHAP para os 20 descritores mais relevantes no modelo XGBoost (Pipeline GridSearchCV).

Fonte: Elaborado pelo autor.

A Tabela 15 consolida os cinco descritores de maior importância SHAP para cada modelo.

Tabela 15 – Top-5 features por importância SHAP.

#	GroupKFold (Random Forest)	GridSearchCV (XGBoost)
1	CLO_5_15_CA	RC_charge_repu^{D}
2	RC_charge_repu^{D}	CLO_5_15_CA
3	DS_GN_density_4	RC_charge_repu^{R}
4	UC_hydrophobic^{R}	UC_hydrophobic^{R}
5	RC_charge_attr^{R}	CLO_VD_35_15

Fonte: Elaborada pelo autor. ^{D} = descritor direto, ^{R} = descritor de anel (ring).

Ambos os modelos identificaram CLO_5_15_CA e RC_charge_repu^{D} como os dois descritores mais importantes, embora com ordenação invertida. O descritor CLO_5_15_CA quantifica o fechamento geométrico do nanoambiente em uma esfera de 5-15 Å ao redor dos carbonos-α. Valores elevados deste descritor indicam regiões altamente empacotadas, característica publicada para hotspots na literatura clássica (Bogan; Thorn, 1998). O descritor RC_charge_repu^{D} captura a repulsão eletrostática entre cargas de mesmo sinal no resíduo central, refletindo o custo energético de aproximar grupos carregados em um ambiente espacialmente confinado.

Embora ambos os pipelines tenham utilizado exatamente o mesmo conjunto de 58 descritores selecionados pelo Boruta, os rankings de importância SHAP diferiram: apenas 11 dos 20 descritores mais importantes (55%) apareceram simultaneamente nas top-20 de ambos os

modelos. Esta divergência reflete diferenças nos mecanismos de aprendizado dos algoritmos: Random Forest utiliza agregação de árvores independentes com seleção aleatória de descritores em cada nó, enquanto XGBoost constrói árvores sequencialmente mediante boosting, priorizando descritores que corrigem erros dos estimadores anteriores. Portanto, algoritmos distintos podem identificar padrões preditivos complementares no mesmo conjunto de descritores, resultando em rankings de importância parcialmente divergentes, embora convergindo para o desempenho preditivo similar.

As Figuras Figura 15 e Figura 16 apresentam as matrizes de confusão para os modelos de cada pipeline.

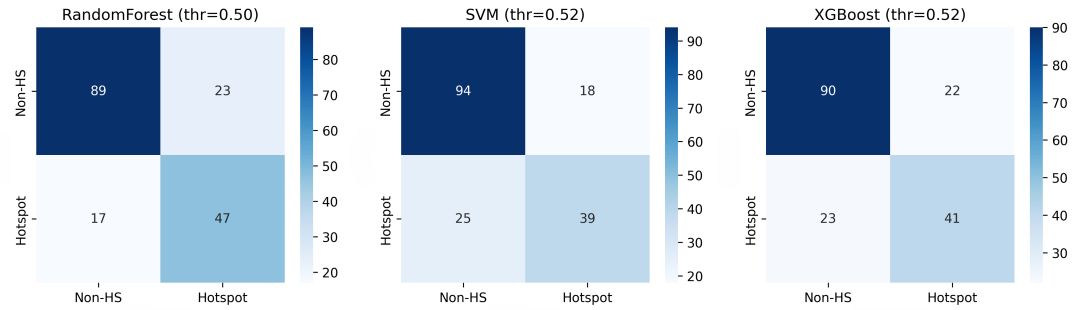


Figura 15 – Matrizes de confusão no conjunto de teste – Pipeline GroupKFold. Random Forest (threshold = 0,50), SVM (threshold = 0,52) e XGBoost (threshold = 0,52).

Fonte: Elaborado pelo autor.

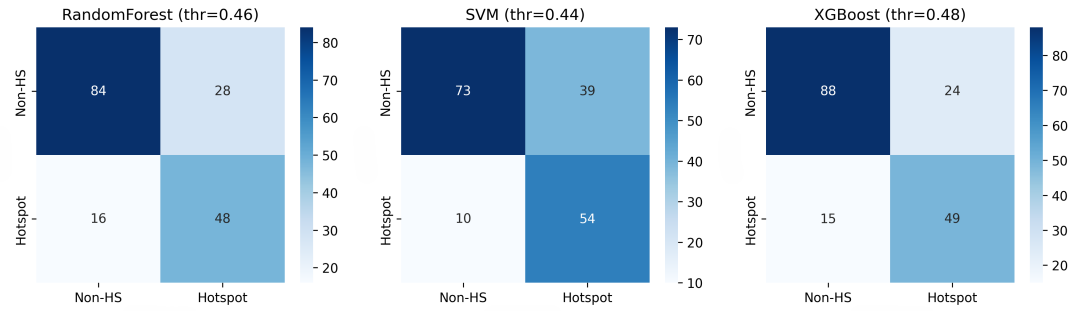


Figura 16 – Matrizes de confusão no conjunto de teste – Pipeline GridSearchCV. Random Forest (threshold = 0,46), SVM (threshold = 0,44) e XGBoost (threshold = 0,48).

Fonte: Elaborado pelo autor.

As matrizes de confusão revelam os padrões de classificação de cada modelo. Considerando o Pipeline GridSearchCV, o SVM apresentou a maior sensibilidade (84,4%), capturando a maioria dos hotspots reais, mas com especificidade moderada (65,2%), resultando em taxa de falsos positivos de 34,8%. O XGBoost apresentou o melhor equilíbrio, com taxas de acerto balanceadas: 76,6% de sensibilidade e 78,6% de especificidade. O Random Forest obteve

valores intermediários, com sensibilidade e especificidade equilibradas em 75,0%. Os thresholds de decisão (variando de 0,44 a 0,52 no GridSearchCV e 0,50 a 0,52 no GroupKFold) foram determinados maximizando o F1-Score em cada modelo.

3.7 DISCUSSÃO

Esta seção examina os resultados obtidos sob quatro perspectivas complementares. Inicialmente, analisam-se os componentes metodológicos que determinaram o desempenho preditivo observado, com ênfase no controle de homologia, normalização intra-proteína e eliminação de multicolinearidade. Em seguida, investiga-se a relevância biológica dos descritores selecionados e sua relação com as teorias clássicas de formação de hotspots. Posteriormente, contextualizam-se os resultados em relação a trabalhos anteriores, considerando diferenças fundamentais em protocolos de validação e controle de redundância. Por fim, discutem-se as limitações inerentes ao estudo, no que concerne ao tamanho do conjunto de dados e ao escopo da inferência estatística empregada.

3.7.1 Fatores de Sucesso da Metodologia Proposta

O desempenho obtido pela metodologia apresentada deriva dos seguintes componentes metodológicos:

- Controle rigoroso de homologia: O conjunto de dados do PPI-HotSpotDB (Chen et al., 2022), curado para redundância pelos autores, foi validado e certificado mediante agrupamento CD-HIT (Li; Godzik, 2006), confirmando ausência de proteínas homólogas entre os 158 complexos retidos. A validação cruzada estratificada por grupos (GroupKFold) complementou este controle, garantindo que proteínas inteiras permaneçam em partições distintas durante o treinamento (Park; Marcotte, 2012). A generalização positiva observada entre validação cruzada e teste independente evidencia a eficácia deste protocolo duplo na prevenção de vazamento de dados.

- **Normalização intra-proteína:** A normalização mitigou o *covariate shift* (Shimodaira, 2000; Sugiyama; Krauledat; Müller, 2007) entre diferentes famílias estruturais, permitindo que o modelo aprenda padrões invariantes de domínio em vez de características específicas de classes proteicas particulares. Esta estratégia de normalização específica por instância tem demonstrado eficácia em reduzir viés de domínio em problemas de bioinformática estrutural (Abbas et al., 2025).

- Remoção iterativa de multicolinearidade: A eliminação mediante critério VIF (*Variance Inflation Factor*) (O'Brien, 2007; Dormann et al., 2013) (176 iterações até $\text{VIF}_{\text{max}} = 9,97$) eliminou redundância multivariada, assegurando que cada descritor retido contribua com

informação única e não-redundante para o modelo preditivo. A remoção de multicolinearidade é particularmente importante em modelos lineares e quando se busca interpretabilidade das contribuições individuais de descritores (Kuhn; Johnson, 2013).

A comparação entre pipelines com e sem otimização de hiperparâmetros demonstra que a qualidade dos descritores selecionados constitui o principal determinante do desempenho preditivo, superando em importância o ajuste fino de parâmetros dos modelos. Este resultado alinha-se com observações da literatura de aprendizado de máquina que evidenciam que *feature engineering* frequentemente proporciona ganhos superiores aos obtidos mediante otimização de hiperparâmetros (Domingos, 2012). A transição do baseline para o pipeline completo representa ganho de +54,9%, enquanto a otimização subsequente de hiperparâmetros contribuiu com apenas +0,4% adicional, evidenciando a desproporcionalidade entre as duas estratégias.

3.7.2 Arquiteturas de Modelos e Padrões de Predição

Os três algoritmos avaliados apresentaram comportamentos preditivos distintos em função de suas arquiteturas de aprendizado:

- Random Forest: Demonstrou robustez excepcional, alcançando AUC = 0,818 em ambos os pipelines independentemente de otimização de hiperparâmetros. Esta estabilidade deriva do método de aprendizado em conjunto baseado em *bootstrap aggregating* (bagging), onde múltiplas árvores de decisão são treinadas em subconjuntos aleatórios dos dados e combinadas por votação majoritária. A agregação de estimadores independentes reduz variância e confere robustez a particularidades do conjunto de treino (Breiman, 2001b).

- SVM com kernel RBF: Apresentou desempenho melhor (AUC variando de 0,729 a 0,821 entre pipelines) comparado aos demais algoritmos, evidenciando elevada sensibilidade à parametrização controlada por C e da largura de banda do kernel γ. Esta dependência crítica reflete a natureza dual do SVM: o hiperparâmetro de regularização C controla o balanço entre maximização da margem geométrica e minimização do erro empírico, enquanto γ define a complexidade da fronteira de decisão mediante mapeamento implícito para espaço de Hilbert de dimensão potencialmente infinita via kernel radial (Vapnik, 2013). A otimização via GridSearchCV identificou configuração conservadora ($C = 0,1$, $\gamma = 0,01$) que maximiza capacidade de generalização mediante regularização forte e kernel de baixa complexidade, resultando no melhor desempenho global (AUC = 0,821, sensibilidade = 84,4%). A diferença de desempenho observada (AUC 0,729 com parâmetros padrão vs 0,821 com parâmetros otimizados) evidencia que a parametrização apropriada de C e γ é crítica para maximizar a capacidade discriminativa do SVM, embora ambas as configurações demonstrem generalização robusta conforme atestado pela performance superior no teste independente comparada à validação cruzada (Scholkopf; Smola, 2018).

- XGBoost: Alcançou o melhor balanceamento global entre classes (MCC = 0,538, F1-Score = 0,715) mediante estratégia de *gradient boosting* sequencial, onde estimadores aditivos são construídos iterativamente para minimizar função de perda via gradiente descendente (Chen; Guestrin, 2016b). A arquitetura incremental do ensemble, na qual cada novo estimador especializa-se na correção de erros residuais dos anteriores, permite aprendizado progressivo e captura de padrões discriminativos sutis. A otimização via GridSearchCV convergiu para configuração conservadora ($max\_depth = 7$, $learning\_rate = 0,01$, $scale\_pos\_weight = 1,12$) que maximiza capacidade de generalização: profundidade limitada das árvores restringe complexidade individual dos estimadores base, enquanto taxa de aprendizado reduzida impõe regularização mediante passos conservadores na direção do gradiente. O hiperparâmetro $scale\_pos\_weight = 1,12$ reflete ajuste para o desbalanceamento moderado entre classes (45,1% hotspots vs 54,9% non-hotspots), ponderando assimetricamente a função de perda para compensar a menor representatividade de hotspots. O equilíbrio entre sensibilidade (76,6%) e especificidade (78,6%) evidencia que o modelo aprendeu padrões discriminativos robustos para ambas as classes, sem viés excessivo para nenhuma delas.

3.7.3 Relevância Biológica e Convergência com Teorias Clássicas

A distribuição dos 58 descritores finais entre categorias estruturais (29,3% interações, 29,3% contatos, 17,2% propriedades físico-químicas, 15,5% grafos, 8,6% superfície) reproduz, de forma independente, determinantes estruturais amplamente estabelecidos na literatura clássica sobre formação de hotspots. Esta convergência demonstra que o pipeline hierárquico de seleção identificou os descritores genuinamente informativos e não apenas descritores estatísticos.

Teoria O-Ring (Bogan; Thorn, 1998): Três descritores de contatos não-utilizados (unused contacts) emergiram entre os mais relevantes: UC_hydrophobic_WNA, UC_hydrophobic_GN e UC_hydrophobic. O conceito de contatos não-utilizados quantifica o potencial de enterramento hidrofóbico ainda não realizado no estado não-ligado, representando superfícies hidrofóbicas expostas ao solvente que tendem a se tornar enterradas após formação do complexo. Este mecanismo está diretamente alinhado com a formulação original da teoria O-Ring, que propõe que hotspots são circundados por anéis de resíduos (*o-ring*) que, embora individualmente não sejam hotspots, fornecem contexto estrutural essencial para a estabilidade da interface. A presença destes descritores valida empiricamente a hipótese de que o potencial de criação de novas interações hidrofóbicas durante a ligação constitui determinante estrutural de hotspots.

Teoria da Dupla Exclusão de Água (Li; Liu, 2009a): Esta teoria postula que hotspots localizam-se em regiões caracterizadas por dupla camada de exclusão de água, associada a elevada densidade atômica local que impede penetração do solvente. Dois descritores de densidade emergiram entre os mais relevantes: DS_SW_sponge_CA_4 e DS_GN_density_CA_4, ambos quantificando densidade atômica em raio de 4 Å ao redor do carbono-α. A análise SHAP destacou DS_GN_density_4 como terceiro descritor mais importante no modelo Random Forest,

evidenciando que compactação estrutural local constitui característica distintiva de hotspots. Estes descritores capturam diretamente o empacotamento geométrico denso responsável pela exclusão de moléculas de água da interface, validando empiricamente a teoria da dupla exclusão e confirmando que a dessolvatação controlada constitui componente energético crucial da formação de hotspots.

3.7.4 Comparação com Trabalhos Anteriores e Diferenciais Metodológicos

As diferenças de desempenho observadas entre este trabalho e estudos anteriores refletem distinções metodológicas fundamentais em protocolos de validação e controle de homologia, não indicando necessariamente superioridade ou inferioridade dos métodos preditivos em si.

Tabela 16 – Comparação de desempenho no conjunto de teste independente.

Trabalho	AUC	Precisão	Sensibilidade	MCC
Literatura				
Pereira (2012)	0,799	0,490	0,900	NR
Wang, Liu & Deng (2018)	0,831	0,81	0,77	0,70
Preto & Moreira (2020)	0,83	0,91	0,82	NR
Chen et al. (2024)	NR	0,760	0,670	NR
Este trabalho - Pipeline GroupKFold				
Random Forest	0,818	0,671	0,734	0,520
SVM-RBF	0,729	0,684	0,609	0,461
XGBoost	0,795	0,651	0,641	0,446
Este trabalho - Pipeline GridSearchCV				
Random Forest	0,818	0,632	0,750	0,486
SVM-RBF	**0,821**	0,581	**0,844**	0,478
XGBoost	0,806	0,671	0,766	**0,538**

Fonte: Elaborada pelo autor.

NR = Não reportado pelos respectivos autores.

- Pereira (2012): Trabalho realizado pelo mesmo grupo de pesquisa no qual o presente estudo se insere. Limitações metodológicas identificadas:

  (i) Seleção de descritores aplicada sobre o conjunto completo antes da divisão treino-teste, introduzindo vazamento de informação ao permitir que o processo de seleção "veja" amostras posteriormente utilizadas para avaliação;

  (ii) Ausência de controle explícito de homologia via clustering CD-HIT ou estratificação por grupos proteicos.

  O presente trabalho corrige estas limitações mediante seleção *in-fold* (features selecionadas independentemente em cada partição da validação cruzada) e controle rigoroso de homologia, resultando em estimativas mais conservadoras porém metodologicamente robustas.

- Wang, Liu & Deng (2018): Reportaram métricas competitivas no conjunto de teste independente (AUC = 0,831, MCC = 0,70) utilizando conjunto de treino com 313 resíduos de 34 complexos proteicos e conjunto de teste (BID) com 127 resíduos de 18 complexos. Limitações metodológicas:

  (i) Ausência de especificação de controle explícito de homologia de sequência ou estrutural;

  (ii) Validação cruzada convencional sem estratificação por grupos proteicos;

  (iii) Seleção de features realizada sobre o conjunto completo antes da partição treino-teste.

  Em conjuntos de dados de PPIs, proteínas homólogas inevitavelmente compartilham padrões estruturais e energéticos, resultando em vazamento de informação quando resíduos de estruturas evolutivamente relacionadas aparecem simultaneamente em treino e teste. O conjunto reduzido de 34 complexos pode conter múltiplas cadeias estruturalmente relacionadas, e a ausência de controle explícito pode inflar artificialmente as métricas reportadas.

- Preto & Moreira (2020): Reportou métricas competitivas no conjunto de teste independente (AUROC = 0,83, Precisão = 0,91, Sensibilidade = 0,82, F1 = 0,85) utilizando abordagem baseada exclusivamente em sequência via *Extremely Randomized Trees*, sem requerer estruturas tridimensionais. Características metodológicas:

  (i) Conjunto de 534 resíduos (127 hotspots, 407 non-hotspots) de 53 complexos proteicos;

  (ii) Estratificação da divisão treino-teste por tipo de aminoácido e classe (hotspot/non-hotspot), sem estratificação por proteína, permitindo que resíduos de um mesmo complexo apareçam simultaneamente em treino e teste;

  (iii) Ajuste *post-hoc* de limiares de probabilidade por aminoácido calibrado sobre o conjunto de teste, comprometendo a independência estrita entre conjuntos;

  (iv) Ausência de controle explícito de redundância via CD-HIT ou ferramenta equivalente além da eliminação de redundância de sequência reportada no trabalho anterior.

  A ausência de estratificação por proteína é particularmente relevante: em um dataset de 534 resíduos distribuídos em apenas 53 complexos, a probabilidade de resíduos do mesmo complexo serem compartilhados entre treino e teste é elevada. Nessa condição, o modelo pode explorar padrões estruturais e composicionais específicos de proteínas já observadas durante o treinamento, inflando as métricas reportadas. Adicionalmente, a calibração de limiares de decisão utilizando informações do conjunto de teste introduz dependência entre o processo de ajuste e a avaliação final, violando o princípio de isolamento estrito do teste.

O presente trabalho evita ambas as situações mediante GroupKFold com segregação por proteína e isolamento completo do conjunto de teste durante todas as etapas do pipeline.

- Chen et al. (2024): Desenvolveram método para identificação de hotspots usando estrutura livre da proteína, validado no maior conjunto de hotspots experimentais até então (PPI-HotspotDB). Características metodológicas:

  (i) 158 proteínas não-redundantes com identidade de sequência <60% (mesmo threshold adotado no presente trabalho);

  (ii) 414 hotspots experimentais e 504 non-hotspots;

  (iii) Framework de machine learning automatizado (AutoGluon) utilizando apenas quatro features (conservação, tipo de aminoácido, SASA, energia de fase gasosa);

  (iv) Reportou Precisão = 0,760 e Sensibilidade = 0,670, com desempenho superior a FTMap e SPOTONE.

  Embora utilize controle de homologia similar ao presente trabalho (60% identidade), PPI-hotspotID utiliza conjunto reduzido de features estruturais e energéticas calculadas sobre a estrutura livre, enquanto o presente trabalho emprega descritores multiescala do STING_RDB calculados sobre complexos. A diferença de desempenho pode refletir tanto a natureza complementar das representações estruturais quanto diferenças nos conjuntos de dados específicos utilizados para validação.

Em síntese, a comparação com trabalhos anteriores revela que a metodologia proposta apresenta desempenho competitivo quando consideradas as diferenças em protocolos de validação. O melhor modelo deste trabalho (SVM-RBF, AUC $= 0{,}821$) apresenta desempenho numericamente próximo ao de Wang, Liu & Deng (2018) (AUC $= 0{,}831$). Contudo, a comparação direta é metodologicamente limitada por três fatores: os conjuntos de teste têm tamanhos distintos (176 resíduos neste trabalho versus 127 no trabalho de Wang; Liu; Deng), nenhum dos dois trabalhos reporta intervalos de confiança para as métricas no teste independente, e os protocolos de validação diferem substancialmente quanto ao controle de homologia e à estratificação por proteína. Sem acesso aos dados brutos dos trabalhos anteriores, qualquer afirmação de equivalência ou superioridade entre os métodos excede o suporte estatístico disponível. O que se pode afirmar com segurança é que o presente trabalho obtém desempenho numericamente comparável sob protocolo de validação mais rigoroso, o que constitui resultado relevante independentemente da comparação pontual de métricas. Em relação a Preto & Moreira (2020) (AUROC = 0,83, Precisão = 0,91), o presente trabalho apresenta desempenho praticamente equivalente em capacidade discriminativa (AUC = 0,821 vs 0,83), porém sob protocolo de validação mais rigoroso: SPOTONE não estratifica por proteína e calibra limiares de decisão sobre o conjunto de teste, enquanto o presente trabalho mantém segregação estrita por proteína e isolamento completo do teste. Quando comparado com Chen et al. (2024), que adota protocolo de controle de redundância equivalente

(CD-HIT 60%), o presente trabalho apresenta desempenho superior em sensibilidade (0,844 vs 0,670) e precisão comparável, sugerindo que os descritores multiescala do STING_RDB capturam informação estrutural complementar às quatro features utilizadas pelo PPI-hotspotID. Dessa forma, a principal contribuição deste trabalho não reside na maximização absoluta de métricas, mas na demonstração de que é possível alcançar desempenho preditivo robusto (AUC > 0,80, MCC > 0,50) sob condições rigorosas de validação que garantem generalização para proteínas estruturalmente distintas daquelas observadas durante o treinamento.

3.7.5 Limitações do Estudo

Tamanho do conjunto de dados: O conjunto de 918 resíduos distribuídos em 158 proteínas é modesto para os padrões de aprendizado de máquina contemporâneo. Esta limitação é inerente à metodologia experimental empregada. Considerando a necessidade de particionamento entre conjuntos de treino (80%) e teste (20%), o volume de treino efetivo (742 resíduos, 126 proteínas) situa-se no limite inferior recomendado para modelagem supervisionada. Este tamanho é adequado para algoritmos como Random Forest, Support Vector Machines e XGBoost, que demonstram bom desempenho em regimes de dados limitados, mas é insuficiente para arquiteturas de aprendizado profundo, que requerem mais exemplos de treinamento para generalização efetiva.

Validação rigorosa versus comparabilidade com literatura: A adoção de controle explícito de homologia mediante agrupamento CD-HIT e validação cruzada estratificada por grupos (GroupKFold) resulta em estimativas mais realistas da capacidade de generalização para proteínas estruturalmente distintas daquelas observadas durante treinamento. Entretanto, este protocolo rigoroso tende a produzir métricas de desempenho mais conservadoras quando comparado a validações convencionais amplamente utilizadas na literatura. Esta discrepância metodológica limita comparações quantitativas diretas com estudos anteriores, que empregam esquemas de validação menos restritivos e, consequentemente, reportam métricas numericamente superiores. A escolha por validação conservadora reflete compromisso com estimativas realistas de aplicabilidade prática, priorizando confiabilidade sobre maximização de métricas.

Apesar das limitações identificadas, os resultados obtidos demonstram que a metodologia proposta é capaz de produzir modelos preditivos com capacidade de generalização comprovada sob condições rigorosas de validação. As limitações apontadas constituem, simultaneamente, direções concretas para trabalhos futuros: a ampliação do conjunto de dados com novos hotspots experimentais, a integração com estruturas preditas por ferramentas como AlphaFold, e o desenvolvimento de protocolos padronizados de validação que permitam comparações mais equitativas entre métodos. Essas perspectivas são detalhadas no capítulo seguinte.

4 CONCLUSÃO

Este trabalho apresenta e valida uma metodologia hierárquica para predição de hotspots em interações proteína-proteína, reduzindo 2.419 descritores moleculares iniciais para 58 descritores finais (redução de 97,6%) mediante pipeline de seleção rigoroso: filtragem estatística, mRMR *in-fold*, validação por teste de Kolmogorov-Smirnov (FDR $\leq 0{,}10$), remoção iterativa de multicolinearidade (VIF < 10) e seleção final por Boruta. Os 58 descritores retidos abrangem cinco categorias complementares: descritores de interação (29,3%), descritores de contato (29,3%), descritores físico-químicos (17,2%), descritores topológicos baseados em grafos (15,5%) e descritores de superfície (8,6%).

Os modelos desenvolvidos alcançaram desempenho robusto no conjunto de teste independente: AUC-ROC entre 0,729 e 0,821, MCC entre 0,446 e 0,538, e F1-Score entre 0,645 e 0,715. O SVM com kernel RBF (Pipeline GridSearchCV) apresentou melhor capacidade discriminativa (AUC = 0,821) e maior sensibilidade (84,4%), enquanto o XGBoost obteve melhor balanceamento global (MCC = 0,538, F1 = 0,715). O Random Forest demonstrou robustez excepcional, alcançando AUC = 0,818 independentemente de otimização de hiperparâmetros, evidenciando que a qualidade dos descritores selecionados constitui o principal determinante do desempenho preditivo, superando o impacto do ajuste algorítmico.

A análise de explicabilidade via SHAP revelou convergência entre os descritores mais relevantes e as teorias clássicas de formação de hotspots: descritores de contatos não-utilizados (teoria O-Ring), descritores de densidade atômica local (dupla exclusão de água) e descritores de centralidade em redes de contatos (análise topológica). Esta convergência independente valida empiricamente as teorias estabelecidas e demonstra que o pipeline identificou descritores genuinamente informativos, não artefatos estatísticos.

O protocolo de validação rigoroso adotado garante estimativas realistas de generalização para proteínas estruturalmente distintas daquelas observadas durante a etapa de treinamento. A ausência de degradação severa entre validação cruzada (AUC $0{,}72$–$0{,}74$) e teste independente (AUC $0{,}80$–$0{,}82$) é compatível com a hipótese de que os modelos generalizam para proteínas estruturalmente distintas das observadas durante o treinamento. A diferença numérica observada entre as duas avaliações é consistente com a variabilidade amostral esperada, dado o tamanho reduzido do conjunto de teste e não constitui, por si só, evidência de ausência de sobreajuste. A confirmação dessa capacidade de generalização requer avaliação em conjuntos de teste mais amplos.

Em relação aos trabalhos anteriores, a metodologia proposta apresenta desempenho superior ao de Chen et al. (2024) (Sensibilidade = 0,844 vs 0,670), único trabalho que adota protocolo de controle de redundância equivalente (CD-HIT 60%), e praticamente equivalente ao de Wang, Liu & Deng (2018) (AUC = 0,821 vs 0,831) e Preto & Moreira (2020) (AUC = 0,821 vs 0,83), cujos protocolos de validação não incluem estratificação por proteína. No caso do SPOTONE,

a calibração de limiares de decisão sobre o conjunto de teste compromete adicionalmente a independência da avaliação.

A dependência atual de estruturas experimentalmente resolvidas limita a aplicabilidade do método a proteínas com dados estruturais depositados em repositórios públicos. A integração com modelos estruturais previstos por AlphaFold (Abramson et al., 2024) permitiria expandir significativamente o escopo de aplicação, viabilizando análise de proteínas pouco caracterizadas experimentalmente. Considerando que o *AlphaFold Protein Structure Database* contém predições para mais de 200 milhões de estruturas, esta integração representaria evolução substancial da metodologia, habilitando triagem estrutural em larga escala e identificação prospectiva de hotspots em proteínas de interesse terapêutico e agronômico ainda não caracterizadas experimentalmente.

Como contribuição central, este trabalho estabelece um protocolo metodológico reprodutível e biologicamente interpretável para a predição de hotspots em interações proteína-proteína, com controle explícito de vazamento de dados aplicado a cenários reais de generalização por proteína.

Como contribuição central, este trabalho estabelece um protocolo metodológico reprodutível e biologicamente interpretável para a predição de hotspots em interações proteína-proteína, com controle explícito de vazamento de dados aplicado a cenários reais de generalização por proteína. O código de implementação do pipeline está disponível mediante solicitação ao autor.

REFERÊNCIAS

ABBAS, K. et al. Graph neural network-based drug-drug interaction prediction. **Scientific Reports**, Nature Publishing Group UK London, v. 15, n. 1, p. 30340, 2025.

ABRAMSON, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. **Nature**, 2024.

ALBERTS, B. et al. **Fundamentos da Biologia Celular-4**. [S.l.]: Artmed Editora, 2017.

BAJORATH, J. Integration of virtual and high-throughput screening. **Nature Reviews Drug Discovery**, v. 1, n. 11, p. 882–894, 2002.

BELSLEY, D. A.; KUH, E.; WELSCH, R. E. **Regression Diagnostics: Identifying Influential Data and Sources of Collinearity**. [S.l.]: John Wiley & Sons, 1980.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley, v. 57, n. 1, p. 289–300, 1995.

BERG, J. M.; TYMOCZKO, J. L.; STRYER, L. **Biochemistry**. 7. ed. New York: W. H. Freeman, 2012.

BOGAN, A. A.; THORN, K. S. Anatomy of hot spots in protein interfaces. **Journal of molecular biology**, Elsevier, v. 280, n. 1, p. 1–9, 1998.

BOLSTAD, B. M. et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. **Bioinformatics**, v. 19, n. 2, p. 185–193, 2003.

BONIDIA, R. P. et al. Mathfeature: feature extraction package for dna, rna and protein sequences based on mathematical descriptors. **Briefings in Bioinformatics**, v. 23, n. 1, p. bbab434, 2022.

BRAUN, P.; GINGRAS, A.-C. History of protein-protein interactions: From egg-white to complex networks. **Proteomics**, v. 12, n. 10, p. 1478–1498, 2012.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

BREIMAN, L. Random forests. **Machine Learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

BREITLING, R. et al. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. **FEBS Letters**, v. 573, n. 1-3, p. 83–92, 2004.

CHEN, Q. et al. Comparative analysis of sequence clustering methods for deduplication of biological databases. **Journal of Data and Information Quality**, ACM, v. 9, n. 3, p. 1–27, 2018.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: ACM. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.], 2016. p. 785–794.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.: s.n.], 2016. p. 785–794.

CHEN, Y. C. et al. Ppi-hotspotdb: database of protein–protein interaction hot spots. **Journal of Chemical Information and Modeling**, ACS Publications, v. 62, n. 4, p. 1052–1060, 2022.

CHEN, Y. C. et al. PPI-hotspot^{ID} : $fordetectingprotein--proteininteractionhotspotsfromthefreepro$.

CHICCO, D.; JURMAN, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. **BMC genomics**, BioMed Central, v. 21, n. 1, p. 1–13, 2020.

CHO, K.-i.; KIM, D.; LEE, D. A feature-based approach to modeling protein–protein interaction hot spots. **Nucleic acids research**, Oxford University Press, v. 37, n. 8, p. 2672–2687, 2009.

CLACKSON, T.; WELLS, J. A. A hot spot of binding energy in a hormone-receptor interface. **Science**, American Association for the Advancement of Science, v. 267, n. 5196, p. 383–386, 1995.

COHEN, J. **Statistical Power Analysis for the Behavioral Sciences**. 2nd. ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, 1995.

DING, C.; PENG, H. Minimum redundancy feature selection from microarray gene expression data. **Journal of Bioinformatics and Computational Biology**, v. 3, n. 02, p. 185–205, 2005.

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.

DORMANN, C. F. et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. **Ecography**, v. 36, n. 1, p. 27–46, 2013.

FAWCETT, T. An introduction to roc analysis. **Pattern recognition letters**, Elsevier, v. 27, n. 8, p. 861–874, 2006.

FRISHMAN, D.; ARGOS, P. Knowledge-based protein secondary structure assignment. **Proteins: Structure, Function, and Bioinformatics**, v. 23, n. 4, p. 566–579, 1995.

GREENER, J. G. et al. A guide to machine learning for biologists. **Nature Reviews Molecular Cell Biology**, v. 23, n. 1, p. 40–55, 2022.

GUO, W.; WISNIEWSKI, J. A.; JI, H. Hot spot-based design of small-molecule inhibitors for protein–protein interactions. **Bioorganic & Medicinal Chemistry Letters**, Elsevier, v. 24, n. 11, p. 2546–2554, 2014.

HAIR, J. F. et al. **Multivariate Data Analysis**. 7. ed. Upper Saddle River, NJ: Pearson, 2010. ISBN 978-0138132637.

HOLLANDER, M.; WOLFE, D. A.; CHICKEN, E. **Nonparametric Statistical Methods**. 3rd. ed. [S.l.]: John Wiley & Sons, 2013. ISBN 978-0-470-38737-5.

IOFFE, S.; SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. **Proceedings of the 32nd International Conference on Machine Learning**, v. 37, p. 448–456, 2015.

JANKAUSKAITĖ, J. et al. Skempi 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. **Bioinformatics**, Oxford University Press, v. 35, n. 3, p. 462–469, 2019.

JR, F. J. M. The kolmogorov-smirnov test for goodness of fit. **Journal of the American statistical Association**, Taylor & Francis, v. 46, n. 253, p. 68–78, 1951.

KABSCH, W.; SANDER, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. **Biopolymers**, v. 22, n. 12, p. 2577–2637, 1983.

KARPIEVITCH, Y. V.; DABNEY, A. R.; SMITH, R. D. Normalization and missing value imputation for label-free lc-ms analysis. **BMC Bioinformatics**, BioMed Central, v. 13, n. Suppl 16, p. S5, 2012.

KESKIN, O. et al. Principles of protein-protein interactions: what are the preferred ways for proteins to interact? **Chemical Reviews**, v. 108, n. 4, p. 1225–1244, 2008.

KESKIN, O.; MA, B.; NUSSINOV, R. Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. **Journal of Molecular Biology**, Elsevier, v. 345, n. 5, p. 1281–1294, 2005.

KIM, D. E.; CHIVIAN, D.; BAKER, D. Protein structure prediction and analysis using the robetta server. **Nucleic Acids Research**, v. 32, n. Web Server issue, p. W526–W531, 2004.

KORTEMME, T.; BAKER, D. A simple physical model for binding energy hot spots in protein-protein complexes. **Proceedings of the National Academy of Sciences**, v. 99, n. 22, p. 14116–14121, 2002.

KORTEMME, T.; KIM, D. E.; BAKER, D. Computational alanine scanning of protein-protein interfaces. **Science's STKE**, American Association for the Advancement of Science, v. 2004, n. 219, p. pl2–pl2, 2004.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. New York, NY: Springer, 2013. ISBN 978-1-4614-6848-6.

KURSA, M. B.; RUDNICKI, W. R. Feature selection with the boruta package. **Journal of statistical software**, v. 36, p. 1–13, 2010. Disponível em: https://www.jstatsoft.org/article/view/v036i11.

KUSER-FALCÃO, P. et al. Sting_rdb: a relational database of structural parameters for protein analysis with support for data warehousing and data mining. **Genetics and Molecular Research**, v. 6, n. 4, p. 911–922, 2007.

KUTNER, M. H. et al. **Applied Linear Statistical Models**. 5th. ed. Boston, MA: McGraw-Hill Education, 2005. ISBN 9780073108742.

LI, J.; LIU, Q. Double water exclusion: a hypothesis refining the O-ring theory for the hot spots at protein interfaces. **Bioinformatics**, Oxford University Press, v. 25, n. 12, p. i254–i260, 2009. ISMB/ECCB 2009 Conference Proceedings.

LI, J.; LIU, Q. The double water exclusion hypothesis and its implications for homology modeling of the major histocompatibility complex. **Bioinformatics**, Oxford University Press, v. 25, n. 12, p. i254–i260, 2009.

LI, W.; GODZIK, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. **Bioinformatics**, Oxford University Press, v. 22, n. 13, p. 1658–1659, 2006.

LIM, H. et al. Evaluation of protein descriptors in computer-aided rational protein engineering tasks and its application in property prediction in sars-cov-2 spike glycoprotein. **Computational and Structural Biotechnology Journal**, v. 20, p. 788–798, 2022.

LIN, Z. et al. **Evolutionary-scale prediction of atomic-level protein structure with a language model**. 2022.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: **Advances in Neural Information Processing Systems**. [s.n.], 2017. v. 30. Disponível em: https://arxiv.org/abs/1705.07874.

MATTHEWS, B. W. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. **Biochimica et Biophysica Acta (BBA)-Protein Structure**, Elsevier, v. 405, n. 2, p. 442–451, 1975.

MAZONI, I. **Análise do nano-ambiente propício para nucleação e manutenção dos elementos da estrutura secundária no contexto estrutural das proteínas funcionais**. 2018. Tese (Doutorado) — [Universidade Estadual de Campinas, Instituto de Biologia], 2018.

MAZONI, I.; NESHICH, G. Dpin: um dicionário dos nanoambientes internos das proteínas e seu potencial para transformação em ativos para a agricultura. In: **Agricultura Digital: Pesquisa, Desenvolvimento e Inovação nas Cadeias Produtivas**. Brasília, DF: Embrapa, 2018. p. 218–232.

MOREIRA, I. et al. **Spoton: high accuracy identification of protein-protein interface hot-spots. Sci Rep 7 (1): 8007**. 2017.

MOREIRA, I. S.; FERNANDES, P. A.; RAMOS, M. J. Hot spots—a review of the protein–protein interface determinant amino-acid residues. **Proteins: Structure, Function, and Bioinformatics**, v. 68, n. 4, p. 803–812, 2007.

NELSON, D. L.; COX, M. M. **Princípios de bioquímica de Lehninger**. [S.l.]: Artmed Editora, 2022.

NESHICH, G. et al. JavaProtein Dossier: a novel web-based data visualization tool for comprehensive analysis of protein structure. **Nucleic Acids Research**, v. 32, n. suppl_2, p. W595–W601, 2004.

NESHICH, G. et al. STING Millennium: a web based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. **Nucleic Acids Research**, v. 31, n. 13, p. 3386–3392, 2003.

NESHICH, I. A. P. et al. Computational biology tools for identifying specific ligand binding residues for novel agrochemical and drug design. **Current Protein and Peptide Science**, Bentham Science Publishers, v. 16, n. 8, p. 701–717, 2015.

O'BRIEN, R. M. A caution regarding rules of thumb for variance inflation factors. **Quality & Quantity**, Springer, v. 41, n. 5, p. 673–690, 2007.

OLIVEIRA, S. et al. Sting_rdb: a relational database of structural parameters for protein analysis with support for data warehousing and data mining. **Genetics and Molecular Research**, v. 6, n. 4, p. 911–922, 2007.

PARK, Y.; MARCOTTE, E. M. Flaws in evaluation schemes for pair-input computational predictions. **Nature Methods**, v. 9, n. 12, p. 1134–1136, 2012.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PEREIRA, J. G. de C. **Caracterização dos aminoácidos da interface proteína-proteína com maior contribuição na energia de ligação e sua predição a partir dos dados estruturais**. 2012. Dissertação (Mestrado) — [Universidade Estadual de Campinas, Instituto de Biologia], 2012.

PRETO, A. J.; MOREIRA, I. S. Spotone: Hot spots on protein complexes with extremely randomized trees via sequence-only features. **International Journal of Molecular Sciences**, v. 21, n. 19, p. 7281, 2020. Disponível em: https://www.mdpi.com/1422-0067/21/19/7281.

SAEYS, Y.; INZA, I.; LARRAÑAGA, P. A review of feature selection techniques in bioinformatics. **Bioinformatics**, Oxford University Press, v. 23, n. 19, p. 2507–2517, 2007.

SARGSYAN, K.; LIM, C. Using protein language models for protein interaction hot spot prediction with limited data. **BMC Bioinformatics**, v. 25, p. 115, 2024. Disponível em: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-024-05737-2.

SCHOLKOPF, B.; SMOLA, A. J. **Learning with kernels: support vector machines, regularization, optimization, and beyond**. [S.l.]: MIT press, 2018.

SCHWEDE, T. Protein modeling: what happened to the "protein structure gap"? **Structure**, v. 21, n. 9, p. 1531–1540, 2013.

SCHYMKOWITZ, J. et al. The foldx web server: an online force field. **Nucleic Acids Research**, v. 33, n. Web Server issue, p. W382–W388, 2005.

SHIMODAIRA, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. **Journal of Statistical Planning and Inference**, Elsevier, v. 90, n. 2, p. 227–244, 2000.

SITKO, A. **Alanine scanning**. 2013. Wikimedia Commons. Disponível em: https://commons.wikimedia.org/w/index.php?curid=23682643. Licença CC BY-SA 3.0. Acesso em: 28 dez. 2024.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. **Information processing & management**, Elsevier, v. 45, n. 4, p. 427–437, 2009.

SUGIYAMA, M.; KRAULEDAT, M.; MÜLLER, K.-R. Covariate shift adaptation by importance weighted cross validation. **Journal of Machine Learning Research**, v. 8, p. 985–1005, 2007.

THORN, K. S.; BOGAN, A. A. Asedb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. **Bioinformatics**, Oxford University Press, v. 17, n. 3, p. 284–285, 2001.

TUNCBAG, N.; GURSOY, A.; KESKIN, O. Hotpoint: hot spot prediction server for protein interfaces. **Nucleic Acids Research**, Oxford University Press, v. 38, n. suppl_2, p. W286–W292, 2009.

UniProt Consortium. Uniprot: the universal protein knowledgebase in 2021. **Nucleic Acids Research**, v. 49, n. D1, p. D480–D489, 2021.

VÄLIKANGAS, T.; SUOMI, T.; ELO, L. L. A systematic evaluation of normalization methods in quantitative label-free proteomics. **Briefings in Bioinformatics**, v. 19, n. 1, p. 1–11, 2018.

VAPNIK, V. **The nature of statistical learning theory**. [S.l.]: Springer science & business media, 2013.

VELANKAR, S. et al. Sifts: Structure integration with function, taxonomy and sequences resource. **Nucleic Acids Research**, Oxford University Press, v. 41, n. D1, p. D483–D489, 2013.

WANG, H.; LIU, C.; DENG, L. Enhanced prediction of hot spots at protein-protein interfaces using extreme gradient boosting. **Scientific reports**, Nature Publishing Group UK London, v. 8, n. 1, p. 14285, 2018.

WEISS, G. A. et al. Rapid mapping of protein functional epitopes by combinatorial alanine scanning. **Proceedings of the National Academy of Sciences**, v. 97, n. 16, p. 8950–8954, 2000.

WELLS, J. A.; MCCLENDON, C. L. Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. **Nature**, Nature Publishing Group UK London, v. 450, n. 7172, p. 1001–1009, 2007.

DADOS CURRICULARES

IDENTIFICAÇÃO	
	Rodrigo Perez Verna Data nasc. 19/05/1982
Nacionalidade	Brasileiro
Nome em citações bibliográficas	Verna, Rodrigo Perez Verna, R. P.
Currículo Lattes	https://lattes.cnpq.br/3242217378780082
ORCID	https://orcid.org/0000-0003-0310-2523
FORMAÇÃO ACADÊMICA	
2023/2026	Ciência da Computação – Mestrado Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), Brasil. Orientador: Geraldo Francisco Donegá Zafalon.
2020/2022	Gestão com Ênfase em Liderança e Inovação – Especialização Fundação Getulio Vargas (FGV), Brasil.
2018/2020	Gestão da Tecnologia da Informação – Graduação Pontifícia Universidade Católica do Paraná (PUCPR), Brasil.
PARTICIPAÇÃO EM EVENTOS CIENTÍFICOS	
VERNA, R. P. Redução Dimensional com Validação Estratificada: Estudo para Seleção de Features em Problemas de Classificação em Biologia Computacional. Trabalho aprovado e apresentado no XIV Workshop do Programa de Pós-Graduação em Ciência da Computação da UNESP (WPPGCC 2025), realizado de 23 a 24 de outubro de 2025, em São José do Rio Preto, SP.	