

UNIVERSIDADE FEDERAL DE VIÇOSA

**PERSEUCPP: uma estratégia de aprendizagem supervisionada para previsão
de peptídeo de penetração celular**

Gustavo de Oliveira Almeida
Magister Scientiae

**VIÇOSA - MINAS GERAIS
2025**

GUSTAVO DE OLIVEIRA ALMEIDA

**PERSEUCPP: uma estratégia de aprendizagem supervisionada para previsão
de peptídeo de penetração celular**

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*.

Orientadora: Sabrina de A. Silveira

**Ficha catalográfica elaborada pela Biblioteca Central da Universidade
Federal de Viçosa - Campus Viçosa**

T

A447p
2025

Almeida, Gustavo de Oliveira, 1997-
PERSEUCPP: uma estratégia de aprendizagem
supervisionada para previsão de peptídeo de penetração celular /
Gustavo de Oliveira Almeida. – Viçosa, MG, 2025.
1 dissertação eletrônica (58 f.): il. (algumas color.).

Orientador: Sabrina de Azevedo Silveira.
Dissertação (mestrado) - Universidade Federal de Viçosa,
Departamento de Informática, 2025.
Referências bibliográficas: f. 55-58.
DOI: <https://doi.org/10.47328/ufvbbt.2025.638>
Modo de acesso: World Wide Web.

1. Aprendizado do computador. 2. Peptídeos penetradores
de células. 3. Bioinformática. I. Silveira, Sabrina de Azevedo,
1983-. II. Universidade Federal de Viçosa. Departamento de
Informática. Programa de Pós-Graduação em Ciência da
Computação. III. Título.

CDD 22. ed. 006.31

GUSTAVO DE OLIVEIRA ALMEIDA

**PERSEUCPP: uma estratégia de aprendizagem supervisionada para previsão
de peptídeo de penetração celular**

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*.

APROVADA: 13 de agosto de 2025.

Assentimento:

Gustavo de Oliveira Almeida
Autor

Sabrina de Azevedo Silveira
Orientadora

Essa dissertação foi assinada digitalmente pelo autor em 08/10/2025 às 12:36:25 e pela orientadora em 14/10/2025 às 14:57:48. As assinaturas têm validade legal, conforme o disposto na Medida Provisória 2.200-2/2001 e na Resolução nº 37/2012 do CONARQ. Para conferir a autenticidade, acesse <https://siadoc.ufv.br/validar-documento>. No campo 'Código de registro', informe o código **9W5B.T38M.8I8A** e clique no botão 'Validar documento'.

Gostaria de iniciar esta seção expressando minha profunda gratidão a todos que, direta ou indiretamente, contribuíram para a realização desta dissertação.

Dedico, primeiramente, este trabalho à minha esposa, Vitória, meu grande amor, que sempre esteve ao meu lado, apoiando minhas decisões e me dando forças. Sem você, nada disso seria possível. Essa conquista, sem dúvida, é nossa.

Aos meus pais, por todo o amor, incentivo e pelas condições que sempre me proporcionaram para que eu pudesse chegar até aqui. Sem vocês, esse momento não existiria.

Dedico também este trabalho à minha orientadora e amiga, professora Sabrina de Azevedo Silveira, pelas incontáveis conversas, ensinamentos e, principalmente, por acreditar em mim nesta empreitada multidisciplinar entre a Física e a Bioinformática.

Agradeço aos meus amigos de trabalho Igor, Juliana, Moruf e Vinícius pela amizade e pelas contribuições fundamentais ao longo da minha trajetória. Sempre estiveram dispostos a me ajudar e esclarecer dúvidas, e sou muito grato por isso.

Por fim, agradeço a todos os(as) professores(as) da Universidade Federal de Viçosa que, ao longo desses dois anos, contribuíram significativamente para minha formação acadêmica e pessoal.

AGRADECIMENTOS

Este trabalho foi realizado com o apoio das seguintes agências de pesquisa brasileiras: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

RESUMO

ALMEIDA, Gustavo de Oliveira, M.Sc., Universidade Federal de Viçosa, agosto de 2025. **PERSEUCPP: uma estratégia de aprendizagem supervisionada para previsão de peptídeo de penetração celular.** Orientadora: Sabrina de Azevedo Silveira.

Peptídeos de penetração celular (CPPs) são ferramentas promissoras para o *delivery* celular, isto é, são capazes de translocar diversos compostos para o interior das células sem causar danos à membrana celular. Esses peptídeos atuam como sistemas eficientes de entrega de fármacos, com baixa citotoxicidade em comparação a peptídeos artificiais. No entanto, a identificação de CPPs por métodos experimentais é custosa e demorada, o que reforça a necessidade de estratégias computacionais por conta da escalabilidade e custo reduzido. Neste trabalho, propomos o PERSEUCPP, uma estratégia baseada em aprendizagem de máquina voltada para a identificação de CPPs e a predição de sua eficiência de penetração. O modelo utiliza descritores derivados de propriedades físico-químicas, estruturais e composição atômica das sequências, sendo treinado com o algoritmo *Extremely Randomized Trees*. Na primeira etapa, o classificador de CPPs foi desenvolvido com um conjunto balanceado de 967 CPPs e 967 não-CPPs, utilizando validação cruzada 10-fold. A validação foi realizada com dois *benchmarks*, alcançando desempenho superior aos métodos do estado da arte. A segunda etapa concentrou-se na predição da eficiência de penetração, sendo treinada com um conjunto balanceado de 140 CPPs e 140 não-CPPs, também validada em um conjunto independente. O modelo de eficiência apresentou resultados competitivos. Além do bom desempenho, o PERSEUCPP se destaca por sua interpretabilidade, oferecendo subsídios biológicos sobre os descritores mais relevantes para a atividade de penetração celular. Os resultados obtidos indicam que o PERSEUCPP pode ser uma ferramenta valiosa para a produção de CPPs aplicáveis em sistemas de entrega de fármacos e pesquisas biomédicas.

Palavras-chave: peptídeos de penetração celular; cpps; aprendizado de máquina; eficiência de penetração celular

ABSTRACT

ALMEIDA, Gustavo de Oliveira, M.Sc., Universidade Federal de Viçosa, August, 2025. **PERSEUCPP: A Supervised Learning Strategy for Predicting Cell-Penetrating Peptides.** Adviser: Sabrina de Azevedo Silveira.

Cell-penetrating peptides (CPPs) are promising tools for cellular delivery, as they can translocate various compounds into cells without causing damage to the cellular membrane. These peptides act as efficient drug delivery systems, exhibiting lower cytotoxicity compared to artificial peptides. However, the experimental identification of CPPs is time-consuming and costly, which highlights the need for computational strategies due to their scalability and reduced costs. In this work, we propose PERSEUCPP, a machine learning-based strategy aimed at identifying CPPs and predicting their penetration efficiency. The model leverages descriptors derived from physicochemical, structural, and atomic composition properties of the sequences and is trained using the Extremely Randomized Trees algorithm. In the first stage, the CPP classifier was developed using a balanced dataset of 967 CPPs and 967 non-CPPs, evaluated through 10-fold cross-validation. Validation was also performed using two benchmark datasets, achieving performance superior to state-of-the-art methods. The second stage focused on predicting penetration efficiency, trained with a balanced set of 140 CPPs and 140 non-CPPs, and validated on an independent dataset. The efficiency model showed competitive results. In addition to its strong performance, PERSEUCPP stands out for its interpretability, offering biological insights into the most relevant descriptors associated with cell-penetrating activity. The results indicate that PERSEUCPP may serve as a valuable tool for the development of CPPs in drug delivery systems and biomedical research.

Keywords: cell penetrating peptide; cpps; machine Learning; uptake efficiency

Lista de ilustrações

Figura 1 – Ilustração: Aminoácidos, peptídeos e proteínas.	16
Figura 2 – Ilustração: Estrutura básica dos aminoácidos. Em azul, está representado o grupo amino (NH_2); em preto, o carbono central (C); em roxo, o grupo carboxila terminal ($-COOH$); em vermelho, o Hidrogênio (H); e em laranja, a cadeia lateral (R), que varia de acordo com cada aminoácido.	18
Figura 3 – Ilustração: aprendizado construtivo. Imagem gerada pelo autor . . .	24
Figura 4 – Detalhamento do PERSEUCPP. Os passos principais são divididos em quatro etapas. 1 Representa a aquisição de dados, as sequência peptídicas são obtidas de diversas fontes, sendo as principais CPPSite2.0 e MLCPP2.0; 2 Representa o cálculo dos descritores: (i) Propriedades físico-químicas, (ii) Padrões de sequências; (iii) Frequências atômicas; 3 Representa a etapa da aprendizado supervisionado, seleção de descritores principais, treino e teste do modelo. 4 Representa o preditor de CPP, isto é, o modelo final disponível para uso via web.	27
Figura 5 – Boxplot dos CPPs e não-CPPs baseado em características físico-químicas no conjunto de dados de treinamento.	32
Figura 6 – Boxplot dos CPPs e não-CPPs baseado nas características normalizadas CKSAAGP de carga positiva no conjunto de dados de treinamento. CP-H representa a interação do grupo de carregado positivamente com o grupo hidrofóbico; CP-CN representa a interação do grupo carregado positivamente com o grupo carregado negativamente; CP-PNC representa a interação do grupo carregado positivamente com o grupo polar não carregado; CP-CP representa a interação entre grupos carregados positivamente	33
Figura 7 – Boxplot dos CPPs e não-CPPs baseado nas características normalizadas CKSAAGP de resíduos polares não carregados no conjunto de dados de treinamento. PNC-H representa a interação do grupo polar não carregado com o grupo hidrofóbico; PNC-CN representa a interação do grupo polar não carregado com o grupo carregado negativamente; PNC-PNC representa a interação entre grupos polares não carregados; PNC-PC representa a interação entre o grupo polar não carregado com o grupo carregado positivamente.	34

Figura 8 – Boxplot dos CPPs e não-CPPs baseado nas características normalizadas CKSAAGP de resíduos com carga negativa no conjunto de dados de treinamento. CN-H representa a interação do grupo carregado negativamente com o grupo hidrofóbico; CN-CN representa a interação dos grupos carregados negativamente; CN-PNC representa a interação do grupo carregado negativamente com o grupo polar não carregado; CN-CP representa a interação do grupo carregado negativamente com grupo carregado positivamente.	35
Figura 9 – Boxplot dos CPPs e não-CPPs baseado nas características normalizadas CKSAAGP hidrofóbicas no conjunto de dados de treinamento. H-H representa a interação entre grupos hidrofóbicos; H-NC representa a interação do grupo hidrofóbico com o grupo carregado negativamente; H-PNC representa a interação do grupo hidrofóbico com o grupo polar não carregado; H-CP representa a interação do grupo hidrofóbico com o grupo carregado negativamente.	35
Figura 10 – Boxplot dos CPPs e não-CPPs baseado nas características normalizadas de composição atômica no conjunto de dados de treinamento. . . .	36
Figura 11 – 20 principais descritores independentes para a classificação de CPPs identificados pelo algoritmo Extremely Randomized Trees. As abreviações PC, PNC e H representam os grupos de aminoácidos: carregados, polares não carregados e hidrofóbicos, respectivamente.	43
Figura 12 – Scatter plot dos CPPs e não-CPPs com base no ponto isoelétrico e na carga líquida no conjunto de dados de treinamento. Os CPPs são predominantemente peptídeos catiônicos, apresentando valores de ponto isoelétrico superiores a 7 em condições fisiológicas, enquanto grande parte dos não-CPPs são peptídeos aniônicos, com valores de ponto isoelétrico inferiores a 7. A linha tracejada em $x = 7$ indica o ponto isoelétrico, destacando onde se inicia a transição de carga.	44
Figura 13 – A razão das frequências dos descritores KG, RN, KN, RG e PG entre os grupos CPP e não-CPP. Cada barra no gráfico representa a razão (Frequência em CPP / Frequência em não-CPP) das frequências de uma característica específica, calculada como a frequência da característica no grupo CPP dividida pela frequência da mesma característica no grupo não-CPP. Valores acima de 1 sugerem que a frequência da característica é maior no grupo CPP. Valores abaixo de 1 sugerem que a frequência da característica é maior no grupo não-CPP.	45

Figura 14 – Razão entre as médias normalizadas dos descritores normnN, normnO, normnH e normnC entre os grupos CPP e não-CPP. Cada barra no gráfico representa a razão (Frequência em CPP / Frequência em não-CPP) da média normalizada de uma característica específica, calculada como a média normalizada da característica no grupo CPP dividida pela média normalizada da mesma característica no grupo não-CPP. Valores acima de 1 sugerem que a média normalizada da característica é maior no grupo CPP. Valores abaixo de 1 indicam que a média normalizada da característica é maior no grupo não-CPP.	46
Figura 15 – Contribuição relativa dos diferentes grupos de descritores na classificação de CPPs. Cada barra representa a contribuição do grupo de descritores, em que dipeptídeos representam grupos de dois aminoácidos e tripeptídeos representam grupos de três. O grupo físico-químico inclui descritores como índice GRAVY, Massa Molecular, Ponto Isoelétrico, Hidrofobicidade e Carga Líquida. A Composição Atômica refere-se aos átomos de nitrogênio, carbono, hidrogênio, oxigênio e enxofre. CKSA-AGP representa a Composição de Pares de Grupos de Aminoácidos com Espaçamento k	47
Figura 16 – Visualização t-SNE de CPPs e não-CPPs. Cada ponto representa um peptídeo descrito por um conjunto de descritores calculados. Os pontos são coloridos com base em sua classificação: azul para CPPs e laranja para não-CPPs. A separação distinta entre os agrupamentos azul e laranja sugere que as características selecionadas diferenciam de forma eficaz os CPPs dos não-CPPs.	48
Figura 17 – Comparação do Ponto Isoelétrico e da Carga Líquida entre as Sequências de Treinamento e as Sequências Classificadas Incorretamente. A apresenta a distribuição dos pontos isoelétricos para as sequências positivas do conjunto de treinamento e as sequências classificadas incorretamente. B mostra a distribuição do ponto isoelétrico para as sequências negativas do treinamento e as sequências classificadas incorretamente. C exibe a distribuição da carga líquida para as sequências positivas do treinamento e as sequências classificadas incorretamente. D representa a distribuição da carga líquida para as sequências negativas do treinamento e as sequências classificadas incorretamente.	50
Figura 18 – Importância de cada descritor atômico do preditor de eficiência. Cada barra representa a importância extraída pelo modelo ERT, em ordem temos: átomos de oxigênio, enxofre, nitrogênio, hidrogênio e carbono respectivamente	53

Lista de tabelas

Tabela 1	–	Propriedades físico-químicas e composição atômica dos aminoácidos . .	15
Tabela 2	–	Descritores pós teste Mann-Whitney	34
Tabela 3	–	Descritores pós análise de importância	36
Tabela 4	–	Comparação de desempenho dos algoritmos de aprendizado de máquina para predição de CPPs utilizando o mesmo conjunto de dados. Apresentamos desempenho de cinco modelos de aprendizado de máquina, SVM, XGB, MLP, RF e ERT, na predição de peptídeos de penetração celular (CPPs).	38
Tabela 5	–	Comparação entre PERSEUCPP e os métodos competidores no dataset independente CPP924	40
Tabela 6	–	Comparação entre PERSEUCPP e os métodos competidores no dataset independente MLCPP2.0	41
Tabela 7	–	Comparação entre PERSEUCPP e os métodos competidores no dataset de validação CPP1708	42
Tabela 8	–	Previsões incorretas no PERSEUCPP. Freq. representa o número de sequências nos subconjuntos CPP e não-CPP, CP são as previsões corretas e IP são as previsões incorretas	50
Tabela 9	–	PERSEUCPP comparado com o estado da arte de predição de eficiência	51
Tabela 10	–	Resultados da comparação do PERSEUCPP proposto. ALL representa todas as características normalizadas, TRIPEP é o grupo de tripeptídeos, DIPEP é o grupo de dipeptídeos, CKSAAGP representa a Composição de Pares de Grupos de Aminoácidos com Espaçamento k , PC é o grupo de características físico-químicas e ATM é o grupo de características de composição atômica	52
Tabela 11	–	Resultados da comparação do PERSEUCPP proposto com as TOP-N características de cada grupo de descritores.	52

Sumário

1	Introdução	12
2	Referencial teórico	15
2.1	Peptídeos de penetração celular	15
2.1.1	Definição, histórico e características fundamentais	16
2.2	Competidores	19
2.2.1	CellPPD	19
2.2.2	CPPpred	20
2.2.3	SkipCPP-Pred	20
2.2.4	BChemRF	21
2.2.5	MLCPP2.0	22
2.2.6	SiameseCPP	22
2.2.7	PractiCPP	24
2.2.8	GraphCPP	25
3	Materiais e métodos	27
3.1	Aquisição de dados	27
3.2	Teste Mann-Whitney	29
3.3	Engenharia de descitores	30
3.4	Seleção e importância de descritores	31
3.5	Métricas de desempenho	37
3.6	Treinamento do modelo	38
4	Resultados e discussões	39
4.1	CPP924 - Conjunto de dados independente	39
4.2	MLCPP2.0 - Conjunto de dados independente	40
4.3	CPP1708 - Conjunto de dados independente	41
4.4	Resumo dos resultados das comparações	41
4.5	Interpretação dos descritores	42
4.6	Predições Incorretas	49
4.7	Classificador de eficiência	51
4.7.1	PERSEUCPP comparado com o estado da arte de eficiência de penetração	51
4.7.2	Interpretação dos descritores	51
5	Conclusões	54
	Referências	55

1 Introdução

A descoberta e o uso clínico de antibióticos figuram entre os maiores avanços da medicina no século XX, tendo reduzido drasticamente as taxas de mortalidade por infecções bacterianas ([Nature Editorial, 2018](#); [GAYNES, 2017](#)). Durante a chamada "Era de Ouro" dos antibióticos, diversas moléculas antibacterianas como Penicilina (1928), Estreptomicina (1943) e Tetraciclina (1950), foram identificadas e incorporadas à prática médica ([Nature Editorial, 2018](#)). No entanto, o uso excessivo e muitas vezes inadequado desses medicamentos contribuiu para o surgimento e disseminação da resistência bacteriana, um dos maiores desafios da saúde pública global na atualidade ([LI; WU et al., 2023](#); [LIU; XU; SLAVEYKOVA, 2023](#)). Estima-se que a resistência bacteriana seja responsável por milhões de mortes anualmente em todo o mundo, afetando gravemente tanto ambientes hospitalares quanto populações expostas por vias ambientais e agroindustriais ([CDC, 2019](#)).

Nesse cenário, cresce a necessidade de desenvolver alternativas terapêuticas eficazes, seguras e que sejam menos suscetíveis aos mecanismos de resistência já conhecidos. Os peptídeos de penetração celular (CPPs), são caracterizados por uma pequena cadeia de aminoácidos (geralmente em torno de 30 resíduos), parcialmente hidrofóbicos, em geral de carga positiva, com alto ponto isoelétrico e ricos em arginina e lisina ([RAMSEY; FLYNN, 2015](#)). Tais peptídeos atuam como sistemas eficientes de entrega intracelular para moléculas terapêuticas, incluindo antibióticos, proteínas e ácidos nucleicos, com baixa citotoxicidade, dessa forma, se tornam ótimos candidatos para apoiar na busca de soluções para a resistência antimicrobiana. ([MULLARD, 2023](#); [GALZITSKAYA, 2023](#)). Um exemplo clássico, é o peptídeo trans-activator of transcription (TAT), derivado do HIV, amplamente reconhecido por sua capacidade de transportar moléculas terapêuticas para o interior das células ([GULSEREN; UFUK et al., 2015](#)). Além disso, o uso de CPPs para transportar antibióticos diretamente para o interior de bactérias resistentes pode reduzir a concentração necessária do fármaco, minimizando efeitos colaterais e diminuindo o risco de surgimento de novas resistências ([RAMSEY; FLYNN, 2015](#)).

Apesar de seu potencial terapêutico, a identificação experimental de CPPs e a avaliação de sua eficiência de penetração celular envolvem procedimentos complexos, demorados e de alto custo. Diante disso, abordagens computacionais baseadas em inteligência artificial (IA) e aprendizado de máquina (ML, do inglês Machine Learning) ([DAINA; ZOETE, 2016](#); [RODRIGUEZ-PÉREZ et al., 2018](#); [GALUCIO; MONTEIRO et al., 2019](#)) têm se consolidado como alternativas atrativas, proporcionando escalabilidade, padronização e reprodutibilidade na descoberta de novos CPPs, sem a necessidade da sintetização e teste de diversos tipos de peptídeos, afim de encontrar alguns com a capacidade de penetração.

Diversas estratégias computacionais foram desenvolvidas ao longo da última década. A plataforma CellPPD (GAUTAM; CHAUDHARY et al., 2013) combinou dados experimentais com métodos de aprendizado de máquina, baseando-se em propriedades como carga, hidrofobicidade e comprimento da cadeia peptídica. Já o modelo CPPpred (HOLTON; POLLASTRI et al., 2013), adotou uma abordagem supervisionada com base em composição de aminoácidos e propriedades físico-químicas, permitindo ajustes de limiares para controle de falsos positivos. Em seguida, o BChemRF (OLIVEIRA; SANTANA et al., 2021) incorporou uma abordagem de ensemble learning com Random Forest, integrando informações estruturais, flexibilidade e dados evolutivos para aumentar a precisão das predições.

Existem também modelos mais sofisticados, como o SkipCPP-Pred (WEI; TANG et al., 2017), que introduziram técnicas como k-skip-n-gram, uma variação dos tradicionais n-grams, que permite capturar relações entre resíduos que não estão necessariamente adjacentes na sequência, ao considerar pares ou grupos de aminoácidos separados por até k posições. Essa abordagem aumenta a capacidade do modelo de identificar padrões espaciais mais complexos ao longo das cadeias peptídicas. Por outro lado, o MLCPP2.0 (MA-NAVALAN; PATRA, 2022) propôs uma arquitetura em duas camadas para prever tanto a presença de CPPs quanto sua eficiência de penetração. Essa última estratégia combinou uma variedade de descritores, de composições simples até codificações mais complexas como CTD (do inglês *Composition, Transition e Distribution*) e DDE (do inglês *Dipeptide Deviation from Expected mean*), e utilizou análise SHAP (*SHapley Additive exPlanations*), uma abordagem baseada na teoria dos jogos que permite interpretar modelos de aprendizado de máquina ao atribuir a cada variável uma contribuição individual para a previsão realizada.

Abordagens mais recentes com redes neurais profundas foram propostas, como o modelo SiameseCPP (ZHANG; WEI et al., 2022), baseado em redes siamesas e aprendizado contrastivo, destacou-se por extrair representações intrínsecas diretamente das sequências, superando o desempenho do MLCPP2.0, mas perdendo em interpretabilidade. Além disso, o modelo PractiCPP (KEXIN; YUANPENG et al., 2024) introduziu o conceito de *hard negative sampling* para lidar com bases de dados altamente desbalanceadas, combinando descritores sequenciais, *Morgan fingerprints* e *embeddings* oriundos do modelo ESM-2. Já o GraphCPP avançou na representação estrutural dos peptídeos por meio de redes neurais em grafos, alcançando resultados superiores ao modelar interações moleculares complexas entre resíduos e propriedades físico-químicas.

Apesar dos avanços notáveis, muitos desses modelos ainda enfrentam limitações quanto à interpretabilidade, à necessidade de grandes volumes de dados de treinamento e à elevada demanda computacional. Além disso, poucos métodos oferecem um equilíbrio satisfatório entre desempenho preditivo e interpretabilidade dos fatores envolvidos.

Diante desse cenário, este trabalho mostra o desenvolvimento do PERSEUCPP, uma estratégia computacional baseada em aprendizado supervisionado, voltada para a predição de CPPs e avaliação de sua eficiência de penetração celular. O modelo utiliza descritores derivados de propriedades físico-químicas, como massa, carga e ponto isolétrico, além de descritores estruturais, que consideram padrões de aminoácidos na cadeia, e descritores atômicos, que quantificam os átomos que compõem as sequências, com ênfase na interpretabilidade e robustez estatística. Ao empregar o algoritmo Extremely Randomized Trees (ERT), bem como estratégias de validação cruzada e testes em bases independentes, busca-se oferecer uma ferramenta interpretável e com elevado desempenho preditivo.

Além do desempenho preditivo, uma contribuição importante do PerseuCPP está em sua interpretabilidade. Ao basear-se em descritores biologicamente significativos (por exemplo, propriedades físicoquímicas, estruturais e atômicas), o PerseuCPP permite a inspeção direta de quais características estão impulsionando uma determinada previsão. Isso nos permite: (i) formular hipóteses sobre o papel de propriedades específicas, como a distribuição de carga ou a composição de aminoácidos, na atividade de CPPs (peptídeos penetrantes de células); (ii) investigar por que certas sequências são classificadas incorretamente e se elas podem, de fato, representar CPPs ainda não descobertos; e (iii) orientar o design racional de peptídeos ao sugerir quais propriedades poderiam ser modificadas para ampliar o potencial de penetração celular. Tais percepções não estão prontamente disponíveis apenas com embeddings de “caixa-preta”, nos quais as fronteiras de decisão subjacentes não são biologicamente interpretáveis. Dessa forma, o PerseuCPP não apenas alcança desempenho preditivo de alta qualidade, como também fornece um arcabouço para conectar previsões computacionais à descoberta e ao design experimental de peptídeos, destacando assim seu impacto para além da simples acurácia de classificação.

Dessa forma, o objetivo deste estudo foi desenvolver e avaliar o modelo PERSEUCPP, verificando sua capacidade de identificar CPPs e prever sua eficiência de penetração, bem como investigar os descritores mais relevantes para essa tarefa, comparando-o com os principais métodos da literatura.

2 Referencial teórico

2.1 Peptídeos de penetração celular

Os aminoácidos são pequenas moléculas orgânicas que constituem as unidades fundamentais das proteínas e peptídeos, cada um apresentando características físico-químicas próprias, como mostrado na Tabela 1. Quando agrupados em cadeias curtas, originam peptídeos, enquanto cadeias mais longas e complexas dão origem às proteínas (ER; LARAIB et al., 2004), veja Figura 1. A principal diferença entre proteínas e peptídeos está no tamanho e na estrutura: peptídeos geralmente possuem até 50 aminoácidos, apresentando estruturas mais simples e flexíveis, enquanto proteínas são formadas por sequências mais longas que frequentemente se doam em conformações tridimensionais complexas e desempenham funções biológicas mais amplas, como catálise enzimática e sinalização celular (ER; LARAIB et al., 2004; IMB, University of Queensland, 2017).

Tabela 1 – Propriedades físico-químicas e composição atômica dos aminoácidos

Aminoácido	Símb.	pI	Polaridade	Hidrofobicidade	nH	nC	nN	nO	nS
Alanina	A	6.0	Não polar	Hidrofóbico	7	3	1	2	0
Arginina	R	10.8	Polar (+)	Hidrofílico	14	6	4	2	0
Asparagina	N	5.4	Polar	Hidrofílico	8	4	2	3	0
Ácido aspártico	D	2.8	Polar (−)	Hidrofílico	7	4	1	4	0
Cisteína	C	5.1	Polar	hidrofóbico	7	3	1	2	1
Ácido glutâmico	E	3.2	Polar (−)	Hidrofílico	9	5	1	4	0
Glutamina	Q	5.7	Polar	Hidrofílico	10	5	2	3	0
Glicina	G	6.0	Não polar	Hidrofóbico	5	2	1	2	0
Histidina	H	7.6	Polar (+)	Hidrofílico	9	6	3	2	0
Isoleucina	I	6.0	Não polar	Hidrofóbico	13	6	1	2	0
Leucina	L	6.0	Não polar	Hidrofóbico	13	6	1	2	0
Lisina	K	9.7	Polar (+)	Hidrofílico	14	6	2	2	0
Metionina	M	5.7	Não polar	Hidrofóbico	11	5	1	2	1
Fenilalanina	F	5.5	Não polar	Hidrofóbico	11	9	1	2	0
Prolina	P	6.3	Não polar	Hidrofóbico	9	5	1	2	0
Serina	S	5.7	Polar	Hidrofílico	7	3	1	3	0
Treonina	T	5.6	Polar	Hidrofílico	9	4	1	3	0
Triptófano	W	5.9	Não polar	Hidrofóbico	12	11	2	2	0
Tirosina	Y	5.7	Polar	hidrofóbico	11	9	1	3	0
Valina	V	6.0	Não polar	Hidrofóbico	11	5	1	2	0

As abreviações utilizadas foram pI = Ponto Isoelétrico, nH = Número de átomos de Hidrogênio, nC = Número de átomos de Carbono, nN = Número de átomos de Nitrogênio, nO = Número de átomos de Oxigênio e nS = Número de átomos de Enxofre. (SCIENTIFIC, 2025)

Os peptídeos de penetração celular (CPPs) representam uma classe de biomoléculas com capacidade de translocar-se através das membranas celulares e de facilitar a

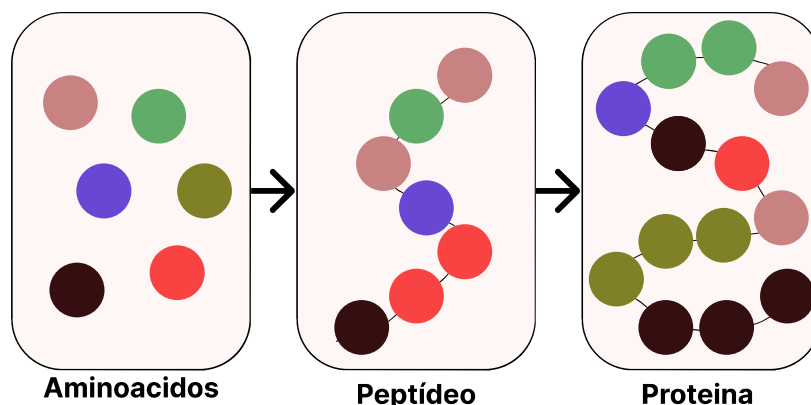


Figura 1 – Ilustração: Aminoácidos, peptídeos e proteínas.

entrega de diversas moléculas terapêuticas (MULLARD, 2023). Desde sua descoberta, os CPPs evoluíram de peptídeos lineares de ocorrência natural para estruturas cíclicas e funcionalizadas, visando superar limitações como a instabilidade metabólica e a baixa biodisponibilidade (PARK; SAJID et al., 2019). Esta otimização contínua é fundamental para a entrega de fármacos, impulsionando a demanda por abordagens computacionais eficientes. A compreensão dos mecanismos de internalização, que variam com a concentração e as propriedades físico-químicas dos peptídeos, é crucial para o design de CPPs. Nesse contexto, a bioinformática emerge como uma ferramenta indispensável, permitindo a predição, caracterização e otimização de CPPs por meio de descritores moleculares avançados. A análise de composições de aminoácidos, dipeptídeos, tripeptídeos, padrões de k-espaçamento, composição atômica e propriedades de transição e distribuição oferece um arcabouço robusto para acelerar a descoberta e o desenvolvimento de CPPs com perfis de eficácia e segurança aprimorados, pavimentando o caminho para novas terapias e diagnósticos.

2.1.1 Definição, histórico e características fundamentais

Os CPPs são altamente diversos em termos de suas propriedades físico-químicas e biológicas. Embora tipicamente curtos, variando de 5 a 50 aminoácidos, eles apresentam variações consideráveis em comprimento, carga, solubilidade e hidrofobicidade (MULLARD, 2023). Uma característica predominante é a presença de uma carga líquida positiva, frequentemente devido à riqueza em aminoácidos básicos como arginina e lisina, o que resulta em um alto ponto isoelétrico. Muitos CPPs também incorporam regiões parcialmente hidrofóbicas, que são essenciais para suas interações com a bicamada lipídica da membrana celular (MULLARD, 2023; GALZITSKAYA, 2023; XIE; BI et al., 2020). A internalização mediada por CPPs é considerada segura e eficiente, uma vez que geralmente não induz perturbações na estrutura da membrana plasmática. A predominância de CPPs catiônicos e seu mecanismo de interação primário, que envolve interações eletrostáticas com as membranas celulares que possuem carga negativa, estabelece uma relação direta

fundamental (MULLARD, 2023; XIE; BI et al., 2020). Essa propriedade inerente explica por que descritores bioquímicos que quantificam a carga ou a composição de aminoácidos básicos (como arginina e lisina) são intrinsecamente preditivos da atividade de CPPs. A compreensão de que a carga positiva é um fator chave para a interação e internalização direciona a seleção e o desenvolvimento de ferramentas bioinformáticas que capturam essa característica, como a carga líquida, a frequência de aminoácidos básicos ou o ponto isoeletrico, tornando-as altamente relevantes e preditivas para a identificação e caracterização de CPPs.

O grupo de descritores CTD (*Composition, Transition e Distribution*) são muito importantes para a classificação dos peptídeos como CPPs, pois capturam padrões globais das propriedades físico-químicas (YUAN; GAO et al., 2022). Para calcular esses descritores, primeiro escolhe-se uma propriedade (como hidrofobicidade, volume, polaridade etc.) e classifica-se os aminoácidos em grupos (tipicamente 3 grupos, por exemplo para polaridade: polar, neutro e hidrofóbico) (DUBCHAK; MUCHNIK et al., 1995). O CTDC (Composition) quantifica a fração de aminoácidos de uma sequência que pertence a cada grupo definido para a propriedade em questão (DUBCHAK; MUCHNIK et al., 1995). O CTDT (Transition) mede a frequência de transições entre dois grupos diferentes ao longo da sequência – em outras palavras, a proporção de junções adjacentes onde um aminoácido de um certo grupo é seguido por outro de grupo distinto (DUBCHAK; MUCHNIK et al., 1995). Já o CTDD (Distribution) captura a distribuição posicional dos aminoácidos de cada grupo na sequência: calcula as posições relativas (ao longo da cadeia) do primeiro aminoácido de um grupo, do 25% quantil, 50% (mediana), 75% e do último aminoácido daquele grupo, retornando cinco percentis que fornecem uma ideia de quão espalhados ou concentrados estão os resíduos de determinado tipo ao longo do peptídeo (DUBCHAK; MUCHNIK et al., 1995).

Combinando todos os valores de composição, transição e distribuição para um conjunto de propriedades (por exemplo, hidrofobicidade, polaridade, acessibilidade ao solvente, carga, etc.), obtém-se um vetor numérico que sumariza características globais da sequência. Em resumo, os descritores CTDC, CTDT e CTDD fornecem, respectivamente, informações sobre a composição de categorias de aminoácidos, frequência de transição entre categorias vizinhas e posições relativas dessas categorias na sequência (DUBCHAK; MUCHNIK et al., 1995; YUAN; GAO et al., 2022)

Um outro descritor muito utilizado é o DDE (do inglês *Dipeptide Deviation from Expected mean*), este quantifica o desvio da frequência observada de cada dipeptídeo (par de aminoácidos adjacentes) em relação à sua frequência esperada em uma sequência, dada a composição de aminoácidos dessa sequência (YUAN; GAO et al., 2022). A ideia é considerar que, se os aminoácidos se combinassem aleatoriamente de acordo com suas proporções globais, cada par teria uma certa probabilidade teórica de ocorrer. Primeiro calcula-se a composição bruta de dipeptídeos (frequência de cada um dos 400 possíveis

pares de aminoácidos) na sequência, depois estima-se a média teórica de ocorrência de cada dipeptídeo com base nas frequências individuais dos dois aminoácidos que o formam (SARAVANAN; GAUTHAM, 2015). Isso pode capturar padrões sequenciais de 2 resíduos não detectáveis pela composição simples, por exemplo, um CPP pode ter ocorrências de “RR”, “KK” muito acima do aleatório, indicando um motivo básico repetido. A inclusão de DDE enriquece o conjunto de atributos com informações de ordem local, complementando os CTD que cobrem ordem global e atributos físicos.

Os vinte aminoácidos proteinogênicos possuem todos uma estrutura básica comum, veja Figura 2, composta pelos elementos carbono (C), hidrogênio (H), oxigênio (O), nitrogênio (N) e enxofre (S) (ZOUBOULIS, 2009). Cada aminoácido contém um grupo amina ($-NH_2$) ligado a um carbono α central, e um grupo carboxila ($-COOH$) terminal; além disso, todos possuem pelo menos um átomo de hidrogênio ligado ao carbono central. A diferença entre eles está na cadeia lateral (grupo R), que varia em estrutura e composição atômica. Algumas cadeias laterais contêm apenas átomos de C e H (por exemplo, alanina, valina, leucina), enquanto outras incluem oxigênio (como serina, treonina), ou nitrogênio (como lisina, arginina) (ZOUBOULIS, 2009). Dois aminoácidos apresentam enxofre (S) em suas cadeias laterais: a cisteína, com um grupo tiol ($-SH$), e a metionina (ZOUBOULIS, 2009), as quantidades atômicas podem ser conferidas na Tabela 1. Essa diversidade de átomos nas cadeias laterais confere propriedades distintas a cada aminoácido, por exemplo, cadeias ricas em N tendem a ser básicas (capazes de protonar e portar carga positiva), cadeias com O tendem a ser polares ou ácidas, e assim por diante. Em resumo, todos os aminoácidos contêm C, H, O, N na sua estrutura, e uma minoria contém S, sendo essas diferenças atômicas fundamentais para as características físico-químicas de cada resíduo (ZOUBOULIS, 2009).

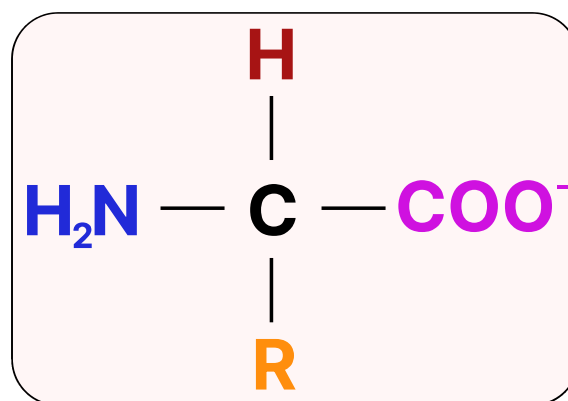


Figura 2 – Ilustração: Estrutura básica dos aminoácidos. Em azul, está representado o grupo amina (NH_2); em preto, o carbono central (C); em roxo, o grupo carboxila terminal ($-COOH$); em vermelho, o Hidrogênio (H); e em laranja, a cadeia lateral (R), que varia de acordo com cada aminoácido.

Dessa forma, as características abordadas acima podem ser empregadas para diferenciar peptídeos que penetram células (CPPs, cell-penetrating peptides) de peptídeos que

não possuem essa capacidade. Sabe-se que muitos CPPs compartilham certos padrões de sequência e composição, por exemplo, tendem a ser catiônicos e anfipáticos, apresentando múltiplos resíduos com carga positiva (como arginina e lisina) e simultaneamente resíduos hidrofóbicos que permitem interações com a bicamada lipídica (MULLARD, 2023). Os descritores CTD podem capturar quantitativamente essas características: um CPP típico muitas vezes exibirá alta composição de grupos de aminoácidos básicos (positivamente carregados) no CTDC e também um perfil de distribuição indicando resíduos positivos tanto nas porções iniciais quanto finais da sequência (refletindo certas sequências sinal) no CTDD. Da mesma forma, uma alta frequência de transições entre resíduos carregados e apolares no CTDT pode refletir a anfipaticidade da sequência, um atributo importante para a translocação através de membranas (YUAN; GAO et al., 2022; MULLARD, 2023). Já o descritor DDE pode evidenciar motivos dipeptídicos enriquecidos em CPPs – por exemplo, pares como arginina-triptofano (RW) ocorrem frequentemente em alguns CPPs conhecidos e seu desvio positivo no vetor DDE indicaria essa super-representação em comparação a peptídeos não penetrantes (SARAVANAN; GAUTHAM, 2015; YUAN; GAO et al., 2022; MULLARD, 2023). Adicionalmente, informações sobre a composição elementar ajudam a interpretar essas tendências: um elevado conteúdo de nitrogênio (proveniente de muitas cadeias laterais básicas) e baixo conteúdo de enxofre (CPPs raramente contêm muitos resíduos de cisteína/metionina, a menos que formem pontes internas) podem ser pistas complementares sobre a capacidade de penetração (ZOUBOULIS, 2009; MULLARD, 2023). Em suma, descritores globais como CTD, padrões locais como DDE e características atômicas dos aminoácidos atuam em conjunto como um conjunto de ‘pegadas digitais’ de um peptídeo, permitindo identificar aqueles com alta probabilidade de funcionar como CPPs.

2.2 Competidores

2.2.1 CellPPD

O modelo CellPPD (GAUTAM; CHAUDHARY et al., 2013) foi um dos primeiros preditores de CPPs baseados em aprendizado de máquina, combinando um conjunto de dados experimentais de peptídeos conhecidos com descritores físico-químicos. Este modelo, desenvolvido por Gautam et al. (2013), utilizou uma máquina de vetor de suporte (SVM) treinada em características como composição de aminoácidos (AAC) e dipeptídeos, padrões binários de sequência e propriedades físico-químicas (PCP) dos peptídeos. Em particular, o CellPPD enfatiza atributos conhecidos por influenciar a entrada celular: carga positiva líquida, hidrofobicidade e comprimento da sequência. Esses fatores refletem tendências observadas experimentalmente, por exemplo, a maioria dos CPPs é constituída de peptídeos curtos (tipicamente 5–30 aminoácidos) com alta carga catiônica e fração considerável de resíduos hidrofóbicos (KAUFFMAN; FUSELIER; OTHERS., 2015). A carga

positiva favorece interações eletrostáticas com membranas celulares negativas, enquanto regiões hidrofóbicas facilitam a translocação através da bicamada lipídica (KAUFFMAN; FUSELIER; OTHERS., 2015). O CellPPD incorpora essas características como variáveis de entrada: peptídeos com maior carga e hidrofobicidade tendem a receber pontuações mais altas de CPP, e há um comprimento ótimo (em torno de 10–20 resíduos), sequências muito curtas podem não conter motivos suficientes, ao passo que sequências muito longas podem ter dificuldade de penetração. Graças a essa abordagem híbrida (dados experimentais + características calculadas), o CellPPD alcançou desempenho robusto, tornando-se uma referência inicial na área. Em suma, o CellPPD mostrou que integrar conhecimentos bioquímicos em algoritmos de ML melhora significativamente a identificação de CPPs.

2.2.2 CPPpred

O CPPpred (HOLTON; POLLASTRI et al., 2013) introduziu um preditor de CPPs que utiliza como entrada principalmente composição de aminoácidos aliada a descritores físico-químicos globais da sequência. Especificamente, o modelo computa frequências relativas de cada aminoácido (AAC) e várias propriedades como peso molecular, ponto isoeletrico (pI) e índices de hidrofobicidade para caracterizar a sequência. Essas características alimentam uma rede neural do tipo N-to-1 (um modelo que sintetiza toda a sequência em uma única saída) que produz uma probabilidade de o peptídeo ser CPP. Um aspecto importante do CPPpred é o ajuste de limiar de decisão para equilibrar sensibilidade e especificidade. Os autores notaram que usar o limiar padrão de 0,5 (ou seja, classificar como CPP se a probabilidade $> 0,5$) resultava em aproximadamente 20% de falsos positivos, enquanto um limiar mais alto, como 0,7, reduzia os falsos positivos para menos de 10%. Os autores enfatizam que determinadas composições de aminoácidos (por exemplo, enriquecimento em Arg/Lys) e propriedades globais (como alto pI e moderada hidrofobicidade) diferenciam CPPs de peptídeos comuns, e incorporou essas informações no modelo. A combinação dos descritores e o ajuste no limiar de decisão fizeram com que CPPpred fosse considerado o estado da arte na época.

2.2.3 SkipCPP-Pred

O SkipCPP-Pred (WEI; TANG et al., 2017) introduziu uma inovação no modo de representar sequências, visando capturar padrões não contíguos importantes para a atividade CPP. O cerne da abordagem é o conceito de k-skip-n-gram adaptativo. Em vez de considerar apenas n-gramas contíguos (subsequências de n resíduos adjacentes), a técnica k-skip-n-gram inclui subsequências de comprimento n onde os elementos podem estar separados por até k posições na cadeia. Em outras palavras, ela incorpora informação de distância no modelo de n-gramas tradicional. Por exemplo, um 2-skip-3-grama considera grupos de três aminoácidos que não precisam estar lado a lado na sequência. Entre dois

deles, pode haver até dois outros aminoácidos no meio. Isso permite identificar padrões mesmo quando os aminoácidos estão um pouco afastados uns dos outros. Assim, padrões relacionados a distâncias curtas na sequência, como aminoácidos básicos repetidos a cada 2 posições, ou pares hidrofóbico-catiônico alternados, podem ser reconhecidos pela representação k-skip-n-gram, mesmo que não formem segmentos contínuos. Essa representação de características é então fundida com um classificador Random Forest no SkipCPP-Pred, construindo um modelo rápido e acurado.

Em suma, o SkipCPP-Pred exemplifica a evolução das técnicas de representação de sequências em bioinformática ao expandir a janela de percepção do modelo além de segmentos consecutivos, ele conseguiu detectar padrões dispersos (mas coordenados) característicos dos CPPs.

2.2.4 BChemRF

O BChemRF-CPPred ([OLIVEIRA; SANTANA et al., 2021](#)) representa um avanço ao integrar aprendizado de conjunto (ensemble) e uma rica variedade de descritores, indo “além de regras químicas” fixas. O nome BChemRF denota "beyond chemical rules framework" com Random Forest. Em sua essência, o modelo emprega um algoritmo Random Forest, um conjunto de múltiplas árvores de decisão treinadas em diferentes amostras/-subconjuntos de atributos, cujo os resultados são combinados e para decidir se um determinado peptídeo é CPP ou não, é feito um esquema de votação ([DONGES, 2024](#)).

Para o modelo de Oliveira, et al (2021), cada árvore se beneficia de um conjunto abrangente de descritores de peptídeo, que abrangem tanto características de sequência quanto de estrutura e evolução. Especificamente, o modelo extrai: (a) propriedades físico-químicas globais (carga, hidrofobicidade média, etc.), (b) composição aminoacídica e padrões de díptídeos, (c) informações estruturais preditas (por exemplo, tendência à hélice/folha, acessibilidade de superfície, presença de estrutura secundária particular), (d) medidas de flexibilidade da cadeia peptídica, e (e) dados evolutivos derivados de alinhamentos (perfis de sequência indicando conservação de cada posição) ([OLIVEIRA; SANTANA et al., 2021](#)).

Ao incorporar simultaneamente todos esses aspectos, o Random Forest aprende a partir de centenas de atributos (os autores relatam uso de 270+ features no total). O resultado desse ensemble alcançou alta acurácia ($\approx 90 - 96\%$) e MCC acima de 0,80 em testes independentes. Em resumo, o BChemRF exemplifica a evolução para modelos híbridos e de ensemble: ele une diversas fontes de informação (sequência + estrutura + filogenia) e utiliza Random Forest para capturar relações complexas entre esses descritores. Essa abordagem melhora a generalização do modelo, por exemplo, mesmo CPPs não-lineares (cíclicos ou modificados) podem ser identificados quando suas características estruturais/flexibilidade se assemelham aos padrões aprendidos.

2.2.5 MLCPP2.0

O MLCPP 2.0 ([MANAVALAN; PATRA, 2022](#)) representa um preditor que simultaneamente identifica CPPs e estima sua eficácia de internalização (“força” de captação). Ele é construído como um modelo de duas camadas (two-layer predictor), também chamado de abordagem de stacking, onde uma camada inicial de modelos gera previsões que alimentam uma segunda camada refinadora. Especificamente, os autores treinaram dezenas de modelos básicos (com diferentes algoritmos de ML e conjuntos de descritores) usando validação cruzada, e então mesclaram as probabilidades previstas por esses modelos em um novo vetor de atributos para cada exemplo. A primeira camada decide se o peptídeo é CPP ou não, e a segunda camada (ativada somente se CPP) prediz se ele possui alta eficiência de internalização, implementando efetivamente um classificador cascata para duas propriedades distintas. Essa arquitetura hierárquica (CPP vs não, seguido de forte vs fraco) reflete o entendimento biológico de que nem todos os CPPs têm o mesmo potencial de entrega, e permitiu ao modelo abordar ambos os problemas de forma integrada. Em termos de descritores de sequência, o MLCPP 2.0 explorou 17 métodos de codificação de atributos, mas destacou-se o uso de CTD e DDE, combinados a outros, no conjunto final otimizado.

Um aspecto importante do MLCPP2.0 foi a interpretabilidade. Os autores aplicaram a análise SHAP (SHapley Additive Explanations) para explicar as contribuições de cada variável na decisão do modelo final. O SHAP baseia-se em teoria de jogos, calculando valores de Shapley que atribuem a cada feature um valor de importância indicando quanto aquela feature aumenta ou diminui a probabilidade predita, em média, considerando combinações possíveis de features ([LUNDBERG; LEE, 2017](#); [RAMASUNDARAM; SOHN et al., 2025](#)).

Em resumo, o MLCPP 2.0 consolidou vários avanços: (1) um modelo empilhado de duas camadas, que melhora acurácia combinando múltiplos algoritmos, (2) um conjunto ampliado de descritores, unindo CTD globais e DDE locais entre outros, para capturar todos os níveis de padrão na sequência, e (3) interpretação pós-hoc via SHAP, garantindo transparência sobre a influência de cada característica na predição.

2.2.6 SiameseCPP

O SiameseCPP ([ZHANG; WEI et al., 2022](#)) representa um desvio marcante dos métodos baseados em descritores manuais, ao empregar redes neurais siamesas e aprendizado contrastivo para extrair automaticamente uma representação intrínseca das sequências de peptídeos. Uma rede siamês consiste de duas (ou mais) sub-redes idênticas (compartilhando pesos) que processam dois inputs em paralelo, produzindo vetores de saída comparáveis, isto é, o objetivo dessa arquitetura não é classificar os dados de entrada, mas sim diferenciar as entradas ([CHOPRA; HADSELL; LECUN, 2005](#)). A ideia é treinar

o sistema não para classificar diretamente cada sequência, mas sim para diferenciar pares de sequências com base em suas características: sequências similares (por exemplo, dois CPPs conhecidos) devem gerar vetores de embedding próximos, enquanto sequências de classes distintas (um CPP e um não-CPP) devem resultar em vetores distantes no espaço de representação, um exemplo ilustrativo desse comportamento pode ser observado na Figura 3. Esse é o cerne do aprendizado contrastivo: o modelo aprende uma função de similaridade entre peptídeos, de forma que CPPs acabem concentrados em uma região do espaço de embeddings, separada dos não-CPPs. No SiameseCPP, isso foi implementado pela primeira vez em predição de CPPs.

Os benefícios dessa abordagem são notáveis. Primeiro, a rede aprende representações contínuas de peptídeos (embeddings) de forma auto-supervisionada pelos rótulos, sem necessidade de pré-definir descritores manuais como AAC, hidrofobicidade, etc. Isso permite capturar padrões complexos e de alto nível, possivelmente combinando motivos de sequência dispersos, propensão estrutural e outras correlações, que humanos poderiam não antecipar facilmente. De fato, o SiameseCPP mostrou que os vetores gerados por sua rede, os embeddings, foram tão informativos que, mesmo ao serem usados em um modelo simples no final, como uma regressão logística ou outro classificador básico, ainda conseguiram superar o desempenho de modelos tradicionais. Isso indica que a rede foi capaz de aprender, sozinha, padrões muito relevantes sobre os CPPs, que não eram facilmente capturados por métodos mais antigos baseados apenas em características manuais.

Por outro lado, existem desafios e considerações nessa abordagem. O treinamento contrastivo requer estratégias de amostragem de pares eficazes, isto é, selecionar pares negativos difíceis (aqueles não-CPPs que mais se parecem com CPPs) é essencial para forçar a rede a refinar sua discriminação, de outra forma ela poderia encontrar um padrão trivial e não generalizar. A dificuldade reside em definir o que são “sequências similares” ou “diferentes” de forma significativa: felizmente, no caso de CPPs, a definição binária de rótulo serve de guia, mas ainda assim o espaço de possíveis pares é enorme, exigindo amostragem cuidadosa. Além disso, as representações aprendidas pela rede são resultados de múltiplas transformações não lineares, e por isso, configuram um método opaco, diferentemente de modelos anteriores nos quais poderíamos apontar “carga” ou “hidrofobicidade” como fatores, aqui cada dimensão do embedding é uma combinação abstrata de características, dificultando a interpretabilidade do modelo. De maneira geral, o SiameseCPP apresentou resultados superiores a seus competidores e marcou a transição para algoritmos que aprendem a característica certa a partir dos dados, ao invés de serem explicitamente informados de quais características observar.

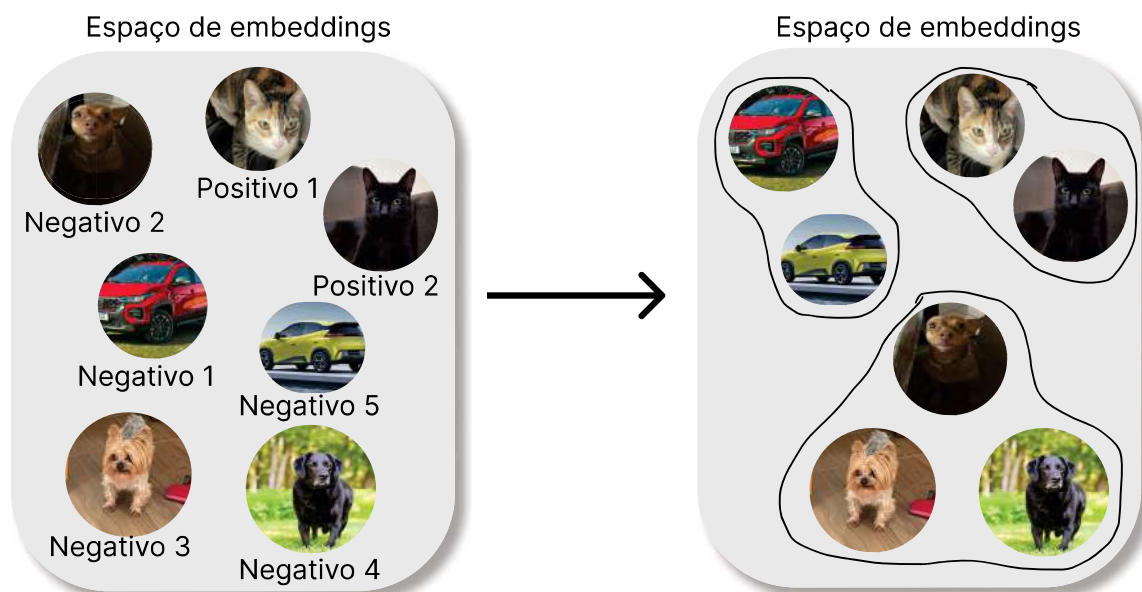


Figura 3 – Ilustração: aprendizado construtivo. Imagem gerada pelo autor

2.2.7 PractiCPP

O PractiCPP (KEXIN; YUANPENG et al., 2024) é uma estratégia computacional de aprendizado profundo projetado para lidar com a situação de extrema desproporção de classes (datasets altamente desbalanceados, onde CPPs positivos são muito raros frente aos negativos). Ele inova em dois aspectos principais: (i) a utilização de *hard negative sampling iterativo* para robustez do treinamento, e (ii) a integração de informações multi-modais sobre os peptídeos, combinando descrições tanto no espaço da sequência quanto no espaço químico/estrutural, incluindo embeddings pretreinados de linguagem de proteínas.

Hard negative sampling é uma técnica que visa refinar as fronteiras de decisão do modelo, expondo-o de forma reforçada a exemplos negativos que são particularmente traiçoeiros (isto é, facilmente confundidos com positivos) (KEXIN; YUANPENG et al., 2024). Dessa forma o PractiCPP, funciona da seguinte forma: durante cada iteração de treinamento, para cada lote de sequências positivas (CPPs) incluídas, seleciona-se um subconjunto de sequências negativas que o modelo atual considera mais propensas a serem CPP, ou seja, aquelas com maior probabilidade predita. Esses negativos "difíceis" são então incorporados no treino junto com os positivos, forçando o classificador a aprender a distingui-los. Esse procedimento faz o modelo “se concentrar” nas fronteiras: se certas sequências não-CPP compartilham características com os CPPs a ponto de enganar o classificador, elas serão elevadas à prioridade de aprendizado, ajustando os parâmetros da rede para separá-las melhor. À medida que o modelo melhora, os hard negatives evoluem, isto é, novas sequências negativas se tornam os casos mais desafiadores e o processo se repete iterativamente até a convergência. Em síntese, o hard negative sampling atua como um “treinador exigente”, desafiando continuamente a rede com os exemplos mais difíceis até que ela generalize bem mesmo nos casos limítrofes. Essa abordagem é fun-

damental quando há um desequilíbrio severo, pois, na sua ausência, o modelo tenderia a rejeitar negativamente a maioria trivial, sem aprimorar a capacidade de distinguir os poucos positivos frente aos negativos mais desafiadores. O PractiCPP demonstrou que o hard negative sampling melhorou significativamente métricas como AUC-PR, focando a precisão justamente onde importa: nos falsos positivos de maior risco.

O segundo pilar do PractiCPP é seu módulo de extração de características e predição (PractiCPP-base), que combina três tipos de descritores de peptídeo complementares. Primeiro, ele extrai descritores sequenciais clássicos da cadeia de aminoácidos, que já discutimos nos modelos anteriores. Segundo, incorpora descritores químicos locais via Morgan fingerprints, que são vetores binários amplamente utilizados em quimioinformática para representar estruturas moleculares (LESORT; DIAZ-RODRIGUEZ et al., 2018). No contexto de peptídeos, isso captura, por exemplo, presença de anéis aromáticos ou fragmento de cadeia contendo lisina seguida de valina. Conforme mencionado pelos autores (KEXIN; YUANPENG et al., 2024), o *Morgan fingerprint* é definido como um descritor do ambiente estrutural local dos átomos. O terceiro tipo de feature são embeddings pré-treinados do modelo de linguagem ESM-2, que é um transformer de última geração treinado em milhões de sequências de proteínas, que produz para qualquer sequência de aminoácidos um vetor de alta dimensão resumindo padrões evolutivos e estruturais aprendidos de forma autônoma. Em conjunto, sequencial + Morgan + ESM-2 cobrem diferentes perspectivas: a sequência provê atributos globais diretos, o *fingerprint* químico local detalha microestruturas e propriedades atômicas, e o *embedding* ESM adiciona intuição estatística de alta ordem baseada em big data de proteínas. O PractiCPP fornece esses três conjuntos de features em um modelo de aprendizado profunda para produzir a predição final de CPP.

A eficácia dessa abordagem multimodal ficou evidente quando PractiCPP superou métodos anteriores em bases desbalanceadas, mostrando robustez e sensibilidade elevadas. Notavelmente, ele conseguiu manter alta detecção de CPPs sem inflar falsos positivos, graças ao refinamento por hard negatives e à informação abrangente fornecida pelos descritores heterogêneos.

2.2.8 GraphCPP

O GraphCPP introduz uma abordagem de Graph Neural Network (GNN) para prever a capacidade de penetração de membrana de peptídeos (ATILLA; BALAZS et al., 2025). Diferentemente dos modelos tradicionais, que se baseiam apenas na sequência linear de aminoácidos ou em descritores moleculares globais, o GraphCPP propõe uma abordagem inovadora ao tratar cada peptídeo como uma estrutura química representada por um grafo molecular. Nessa representação, cada átomo da molécula do peptídeo é convertido em um nó, enquanto as ligações químicas entre os átomos são representadas como

arestas conectando esses nós. Para isso, os peptídeos são convertidos para a notação SMILES isomérica, que captura não apenas a composição atômica da molécula, mas também informações estruturais como quiralidade e estereoquímica. A conversão dos SMILES em grafos é feita utilizando a biblioteca RDKit.

Cada nó do grafo (ou seja, cada átomo) recebe um vetor com 32 características químicas relevantes, como o tipo de átomo, carga formal, hibridização, aromaticidade, número de hidrogênios ligados, quiralidade, entre outras. Essas informações são fundamentais para que a rede neural compreenda a estrutura local de cada átomo e suas interações no contexto da molécula.

Além da estrutura baseada no grafo, o modelo incorpora também os **topological torsion fingerprints**, que são descritores clássicos da química computacional baseados em padrões de ligação entre átomos ao longo da estrutura. Esses fingerprints são processados separadamente e, em seguida, concatenados à representação do grafo gerada pela GNN. O vetor final é então passado por camadas lineares com função de ativação PReLU até chegar na camada de saída, que realiza a predição final sobre a capacidade do peptídeo penetrar na célula.

Em resumo, o GraphCPP representa o estado da arte atual, aproveitando Redes Neurais em Grafo para capturar de forma abrangente as características estruturais e moleculares que governam a capacidade de um peptídeo penetrar células. Isso reflete a evolução contínua da área: dos modelos baseados apenas na sequência passamos a incorporar estrutura, primeiro via descritores e agora via arquiteturas de aprendizado profundo nativamente estruturais, alcançando níveis inéditos de desempenho preditivo e potencializando descobertas de CPPs com propriedades complexas.

3 Materiais e métodos

Nesta seção detalharemos o PERSEUCPP, nossa estratégia de aprendizado supervisionada que utiliza descritores estruturais, físico-químicos e composição atômica para classificar peptídeos de penetração celular. Aqui será detalhado como os dados foram coletados e preparados, o cálculo dos descritores, a etapa de aprendizado supervisionado e a estratégia de validação. A Figura 4 apresenta um resumo dos passos principais do método de predição proposto.

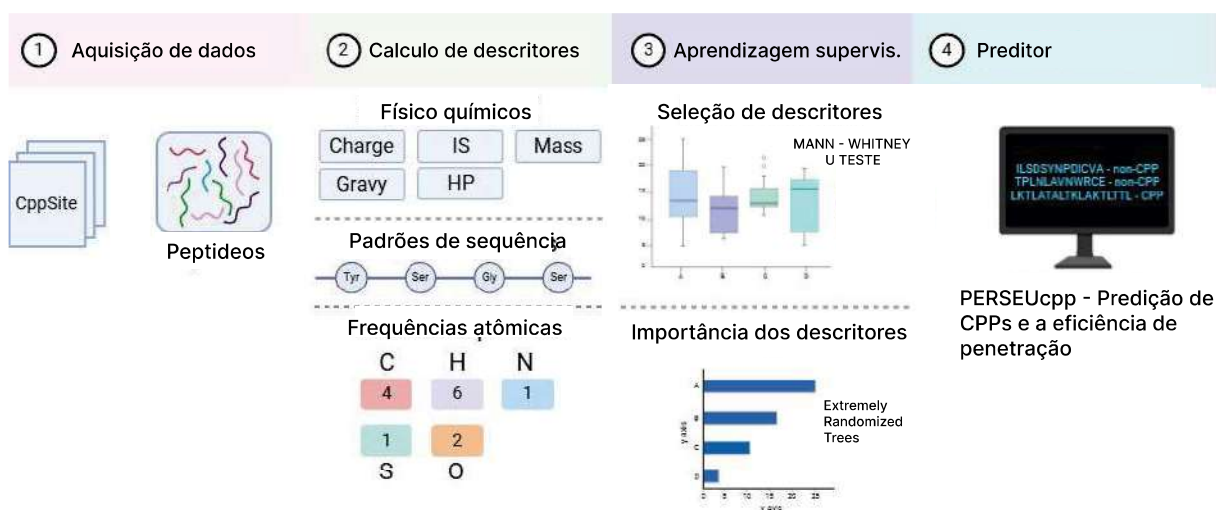


Figura 4 – Detalhamento do PERSEUCPP. Os passos principais são divididos em quatro etapas. **1** Representa a aquisição de dados, as sequência peptídicas são obtidas de diversas fontes, sendo as principais CPPSite2.0 e MLCPP2.0; **2** Representa o cálculo dos descritores: (i) Propriedades físico-químicas, (ii) Padrões de sequências; (iii) Frequências atômicas; **3** Representa a etapa da aprendizado supervisionado, seleção de descritores principais, treino e teste do modelo. **4** Representa o preditor de CPP, isto é, o modelo final disponível para uso via web.

3.1 Aquisição de dados

Dois conjuntos de dados foram utilizados para treinar nosso modelo, um para a classificação dos peptídeos como CPP ou não-CPPs, e outro para classificação da eficiência de penetração. Para validação, demonstração da capacidade de generalização e comparação com os competidores, foram utilizados três datasets independentes. Na literatura existem datasets que são compostos de outros datasets já existentes, abaixo explicamos a composição dos datasets principais e como o nosso foi construído.

1. **CPPSite2.0:** Composto por 1.883 CPPs únicos. O CPPSite2.0 é um conjunto de dados continuamente atualizado de CPPs validados experimentalmente. Ele foi utilizado para aumentar nosso conjunto de dados de treinamento. ([AGRAWAL et al., 2015](#))
2. **Conjunto de dados de treinamento MLCPP2.0:** Originalmente composto por 573 CPPs validados experimentalmente do CPPSite2.0 e 573 não-CPPs onde parte são validados experimentalmente e parte gerados artificialmente por modelos anteriores ([MANAVALAN; PATRA, 2022](#)). Os autores unificaram dados de seis métodos preexistentes para formar um conjunto de treinamento mais confiável. Após remoção de sequências redundantes ($> 85\%$ de identidade), selecionaram 573 CPPs e 573 não-CPPs deste conjunto.
3. **Conjunto de dados de treinamento PERSEUCPP:** Nosso conjunto de dados é composto pelo conjunto de treinamento MLCPP2.0, ampliado com sequências positivas adicionais provenientes do CPPSite2.0. Para balancear o número de novas sequências positivas com negativas, fizemos o download de dados de diversos métodos previamente publicados, já mencionados anteriormente no *Capítulo Referencial Teórico* em 2.2.1-2.2.8, excluímos quaisquer sequências que apareciam nos conjuntos de teste (conjunto independente MLCPP2.0, conjunto independente CPP924, conjunto de teste CPP1708) e removemos duplicatas. O conjunto resultante contém peptídeos naturais com 967 CPPs e 967 não-CPPs, que utilizamos para construir nosso modelo de classificação. É importante ressaltar que as sequências negativas utilizadas em nosso estudo, assim como aquelas empregadas por outros modelos, não são todas validadas experimentalmente. Em muitos casos, os autores geram exemplos negativos artificialmente, utilizando sequências que foram confirmadas experimentalmente como não-CPPs ou assumindo que peptídeos com similaridade estrutural ou funcional a não-CPPs conhecidos também são não-CPPs.
4. **Conjunto de dados de treinamento de eficiência MLCPP2.0:** Composto por 140 peptídeos com baixa e 140 CPPs com alta eficiência de internalização, ambos validados experimentalmente ([MANAVALAN; PATRA, 2022](#)). Este conjunto de dados foi utilizado para construir nosso modelo para a tarefa de classificação da eficiência de internalização de CPPs.
5. **Conjunto de dados independente MLCPP2.0:** Composto por 157 CPPs validados experimentalmente e 2.170 não-CPPs ([MANAVALAN; PATRA, 2022](#)), esse dataset foi construído aplicando os mesmos métodos do conjunto de treino citado anteriormente. Esses dados foram utilizados para comparar o PERSEUCPP com modelos concorrentes.

6. **Conjunto de dados independente de eficiência MLCPP2.0:** Composto por 40 CPPs com baixa e 40 CPPs com alta eficiência de internalização. O objetivo deste conjunto de dados foi comparar o PERSEUCPP com modelos concorrentes ([MANAVAN; PATRA, 2022](#)).
7. **Conjunto de dados independente CPP924:** Um conjunto de dados de alta qualidade composto por 462 CPPs naturais únicos e 462 sequências negativas geradas artificialmente ([WEI; TANG et al., 2017](#)). A similaridade entre quaisquer duas sequências no grupo positivo era inferior a 20%. As sequências negativas foram geradas aleatoriamente, com composição de aminoácidos e comprimento semelhante aos CPPs, e selecionadas de forma a não apresentarem alta identidade com nenhuma sequência positiva. No entanto, como essas sequências não possuem validação experimental, assume-se que não possuem atividade de penetração celular. O objetivo deste conjunto de dados foi comparar o PERSEUCPP com modelos concorrentes.
8. **Conjunto de dados CPP1708:** Composto por 854 CPPs validados experimentalmente e 854 não-CPPs ([ATILLA; BALAZS et al., 2025](#)), o conjunto CPP1708 foi construído a partir da combinação de sequências provenientes de diversos métodos previamente publicados. As sequências passaram por um processo de rotulagem, remoção de duplicatas e padronização estrutural, isto é, as estruturas peptídicas foram convertidas para o formato SMILES isomérico (representação química linear). Para reduzir a redundância, foi aplicado o algoritmo CD-HIT com limiar de 70% de similaridade de sequência, resultando em 854 CPPs únicos e 3.712 não-CPPs. Em seguida, foram selecionadas aleatoriamente 854 sequências negativas para garantir o balanceamento do conjunto. Por fim, o dataset foi dividido em subconjuntos de treinamento (80%), validação (10%) e teste (10%).

3.2 Teste Mann-Whitney

O teste de Mann-Whitney U (MWU) é um teste estatístico não paramétrico usado para comparar duas amostras independentes quando não podemos supor que elas sigam uma distribuição normal ([HART, 2001](#)). Por exemplo, em estudos com tamanho de amostra pequeno ou dados assimétricos, o teste de Mann-Whitney pode fornecer resultados mais confiáveis do que testes paramétricos tradicionais ([HART, 2001](#)), isto é, no caso dos CPPs, possuímos uma base de dados relativamente pequena e diversos descritores, o teste MWU, irá nos dizer quais descritores realmente são úteis para a diferenciação entre os grupos de CPPs e não CPPs.

O teste de Mann-Whitney trabalha com os ranks (postos) dos dados. O procedimento básico é ordenar conjuntamente todos os valores das duas amostras do menor para o maior, ignorando de qual grupo eles vêm ([HART, 2001](#); [TAI; DHALIWAL; BALASU-](#)

BRAMANIAM, 2022)(por exemplo, CPP e não CPP). Em seguida, atribui-se o posto 1 ao menor valor, posto 2 ao segundo menor, e assim por diante, até o valor mais alto (que recebe posto N, onde N é o total de observações nos dois grupos). Depois, calcula-se a soma dos postos de cada grupo (ou a média dos postos em cada grupo). Se as duas amostras vêm de populações semelhantes, espera-se que os postos altos e baixos fiquem misturados entre os grupos, resultando em somas de postos relativamente próximas em ambos os grupos (HART, 2001; TAI; DHALIWAL; BALASUBRAMANIAM, 2022).

A estatística U do teste de Mann-Whitney baseia-se justamente nessa ideia de comparação de postos. Uma forma de entendê-la é contar o número de “vitórias” de um grupo sobre o outro em todas as comparações par a par entre observações dos dois grupos. Por exemplo, para cada par formado por um valor do Grupo A e um do Grupo B, verifica-se qual dos dois é maior. Sob a hipótese nula (nenhuma diferença entre os grupos), esperaríamos que aproximadamente metade das comparações favoreça o Grupo A e metade favoreça o Grupo B. Nesse caso, as somas de postos de A e B seriam próximas. Porém, se um grupo tende a ter valores sistematicamente maiores que o outro, seus dados receberão postos mais altos (próximos do topo da ordenação), enquanto o outro grupo receberá postos mais baixos (HART, 2001; TAI; DHALIWAL; BALASUBRAMANIAM, 2022). Assim, quanto mais distantes estiverem as somas de postos dos dois grupos, mais evidência haverá de diferença entre eles, o que resulta em um p-valor menor. O p-valor representa a probabilidade de observarmos uma diferença tão grande (ou maior) entre os grupos caso a hipótese nula seja verdadeira (isto é, se realmente não houver diferença). Assim, um p-valor pequeno indica que tal diferença seria improvável por acaso, sugerindo que os grupos são, de fato, diferentes.

3.3 Engenharia de descritores

Na primeira parte deste estudo, um conjunto de 8.426 descritores foi calculado e normalizado pelo comprimento da cadeia de cada sequência peptídica. Esses descritores incluíram *Composição Atômica*, *Composição de Dipeptídeos* (DPC), *Composição de Tripeptídeos* (TPC), *Composição de Pares de Grupos de Aminoácidos com Espaçamento k* (CKSAAGP), e *Propriedades Físico-Químicas*, como o índice GRAVY (Grand Average of Hydropathy), Massa Molecular, Ponto Isoelétrico, Hidrofobicidade e Carga Líquida.

Para o DPC, foram computadas todas as combinações possíveis dos 20 aminoácidos tomadas em pares. Também utilizamos o DPC normalizado pelo comprimento do peptídeo. Para o TPC, foram calculadas todas as combinações possíveis dos 20 aminoácidos tomadas em grupos de três. Adicionalmente, o TPC também foi normalizado pelo comprimento do peptídeo.

Para o CKSAAGP, calculamos a frequência de pares de grupos de aminoácidos

separados por K resíduos na sequência peptídica. Os aminoácidos foram agrupados com base em suas propriedades físico-químicas, como carga, polaridade e hidrofobicidade. Esse descritor captura interações proximais entre grupos de aminoácidos que podem influenciar significativamente a estrutura e a função do peptídeo. Além disso, os descritores CKSA-AGP foram normalizados pelo comprimento do peptídeo, garantindo consistência entre sequências de diferentes tamanhos.

As *propriedades físico-químicas*, índice GRAVY, Massa Molecular, Ponto Isoelétrico, Hidrofobicidade e Carga Líquida, foram calculadas utilizando a biblioteca Biopython (COCK et al., 2009). Por fim, a *Composição Atômica* foi determinada a partir das quantidades de carbono, hidrogênio, nitrogênio, oxigênio e enxofre, também normalizadas pelo comprimento da cadeia.

3.4 Seleção e importância de descritores

Devido ao grande número de descritores, uma análise exploratória dos dados foi realizada utilizando o teste estatístico de Mann-Whitney, conforme discutido em 3.2, para verificar se cada característica apresentava diferença significativa (valor de $p < 0,05$) entre os grupos CPP e não-CPP. Para essa análise cada descritor foi calculado e então normalizado pelo tamanho da sequência. Como resultado, o conjunto inicial de 8.426 descritores foi reduzido para 1.137.

A Figura 5 mostra os cinco descritores *físico-químicos* e a diferença entre as classes de CPP e não-CPPs. Todos os descritores serão considerados pois apresentaram um p-value $< 0,05$.

Já para os 16 descritores *CKSAAGP*, da Figura 6-9, temos 9 descritores que demonstraram p-value $< 0,05$ entre CPPs e não-CPPs:

1. Carga positiva e hidrofobia isto é, em uma sequência peptídica, para os CPPs (Figura 6, primeiro gráfico, caixa azul), há grupos de aminoácidos carregados positivamente próximos de grupos hidrofóbicos mais predominantes do que em não-CPPs (Figura 6, primeiro gráfico em laranja);
2. Carga positiva e carga positiva, ou seja, em uma sequência peptídica, para os CPPs (Figura 6, último gráfico, caixa azul), há grupos de aminoácidos carregados positivamente próximos de grupos de aminoácidos carregados positivamente mais predominantes do que em não-CPPs.
3. Polar não carregado e hidrofóbicos, em uma sequência peptídica, para os CPPs (Figura 7, primeiro gráfico, caixa azul), é menos predominante do que não-CPPs;

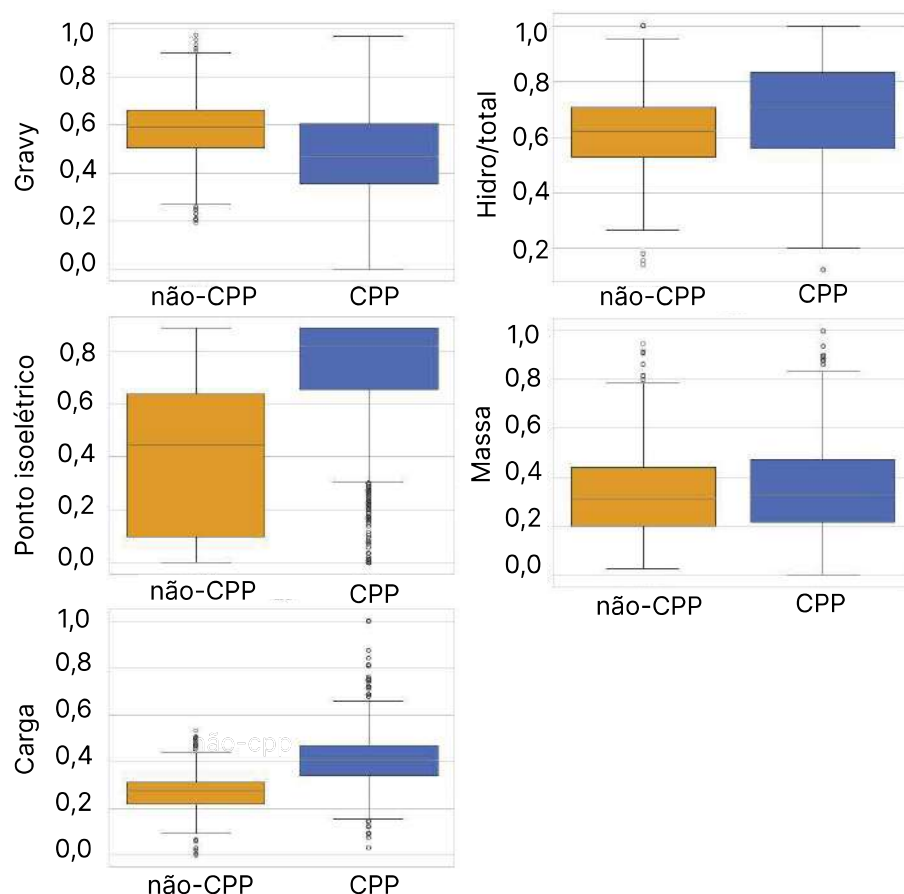


Figura 5 – Boxplot dos CPPs e não-CPPs baseado em características físico-químicas no conjunto de dados de treinamento.

4. Polar não carregado e polar não carregado, em uma sequência peptídica, para os CPPs (Figura 7, terceiro gráfico, caixa azul), é menos predominante do que não-CPPs;
5. Polar não carregado e carga positiva, em uma sequência peptídica, para os CPPs (Figura 7, último gráfico, caixa azul), é mais predominante do que não-CPPs;
6. Carga negativo e hidrofóbicos, em uma sequência peptídica, para os não-CPPs (Figura 8, primeiro gráfico, caixa laranja) é mais predominante do que em CPPs;
7. Hidrofóbicos e negativamente carregados, em uma sequência peptídica, para os não-CPPs (Figura 9, segundo gráfico, caixa laranja) é mais predominante do que em CPPs;
8. Hidrofóbicos e polar não carregador, em uma sequência peptídica, para os não-CPPs (Figura 9, terceiro gráfico, caixa laranja) é mais predominante do que em CPPs;
9. Hidrofóbicos e carga positiva, em uma sequência peptídica, para os CPPs (Figura 9, último gráfico, caixa azul) é mais predominante do que em não CPPs.

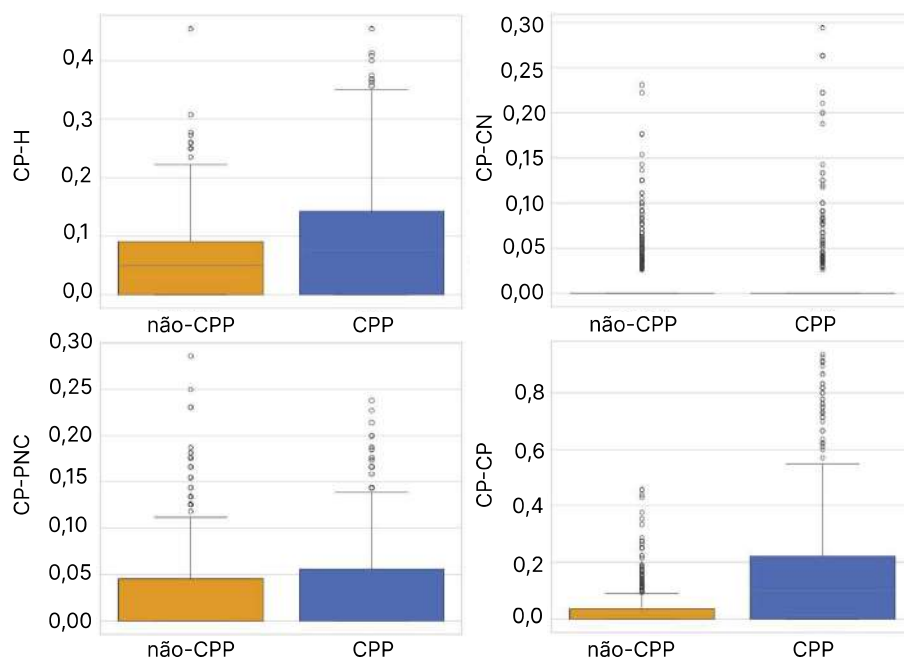


Figura 6 – Boxplot dos CPPs e não-CPPs baseado nas características normalizadas CK-SAAGP de carga positiva no conjunto de dados de treinamento. **CP-H** representa a interação do grupo de carregado positivamente com o grupo hidrofóbico; **CP-CN** representa a interação do grupo carregado positivamente com o grupo carregado negativamente; **CP-PNC** representa a interação do grupo carregado positivamente com o grupo polar não carregado; **CP-CP** representa a interação entre grupos carregados positivamente

Também temos os cinco descritores de composição atômica, que é referente a quantidade de átomos em cada aminoácidos que compõe a cadeia peptídica, como discutido na Seção 2.1. Todas as características atômicas serão incluídas, pois demonstraram $p\text{-value} < 0,05$ e a diferença entre as classes podem ser conferidas na Figura 10.

Para as características sequenciais (dipeptídeos e tripeptídeos) temos 8.400 features, pois estas são compostas de combinações possíveis dos 20 aminoácidos naturais, conforme discutido na Seção 2.1. Assim, teríamos os 8.400 gráficos para apresentar. Para contornar isso, calculamos o $p\text{-value}$ de todos os dipeptídeos e tripeptídeos, selecionando somente aqueles com $p\text{-value} < 0,05$. Dessa forma, o conjunto das características sequenciais foi reduzido de 8.400 para 1.118.

Depois da aplicação dessa estratégia em todos os grupos de descritores mencionados anteriormente, o conjunto de 8.419 features foi reduzido para 1.137, seguindo a distribuição presente na Tabela 2.

Por fim, a partir dos 1.137 descritores, treinamos o modelo utilizando o algoritmo ERT (Extremely Randomized Trees), que permite identificar os descritores com maior influência na classificação de um peptídeo como CPP. O algoritmo gera uma tabela com os graus de importância de cada descritor, cujos valores somam 100%. Com base nesses dados, removemos os descritores cuja contribuição foi igual a zero, resultando na redução

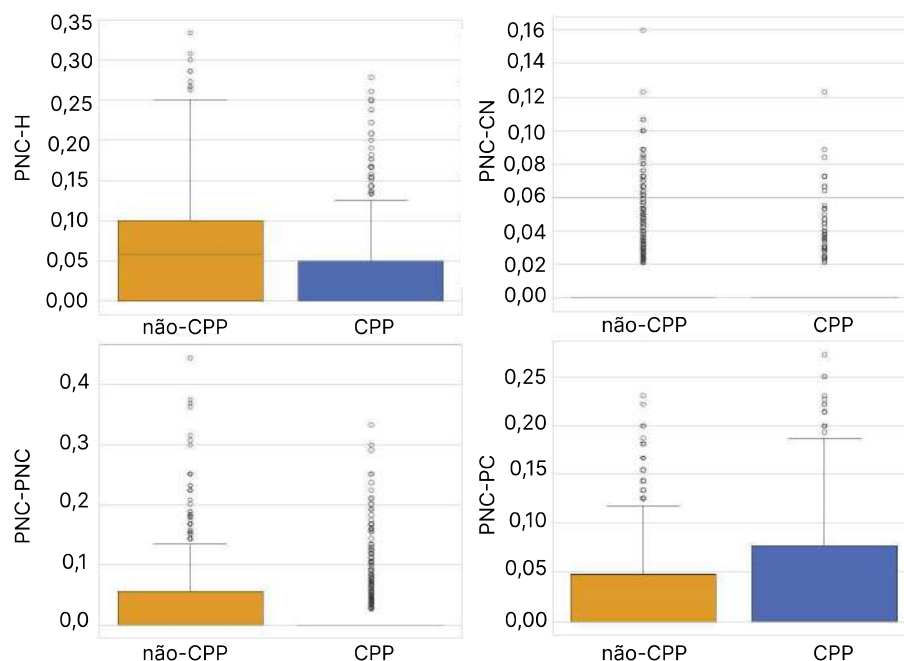


Figura 7 – Boxplot dos CPPs e não-CPPs baseado nas características normalizadas CKSAAGP de resíduos polares não carregados no conjunto de dados de treinamento. **PNC-H** representa a interação do grupo polar não carregado com o grupo hidrofóbico; **PNC-CN** representa a interação do grupo polar não carregado com o grupo carregado negativamente; **PNC-PNC** representa a interação entre grupos polares não carregados; **PNC-PC** representa a interação entre o grupo polar não carregado com o grupo carregado positivamente.

Tabela 2 – Descritores pós teste Mann-Whitney

Descritor	Qtd inicial	Qtd pós MWU
Físico-Químicos	5	5
Comp. Atômicos	5	5
CKSAAGP	16	9
Dipeptídeos	400	310
Tripeptídeos	8.000	808

do conjunto original para 491 descritores, que podem ser conferidos na Tabela 3.

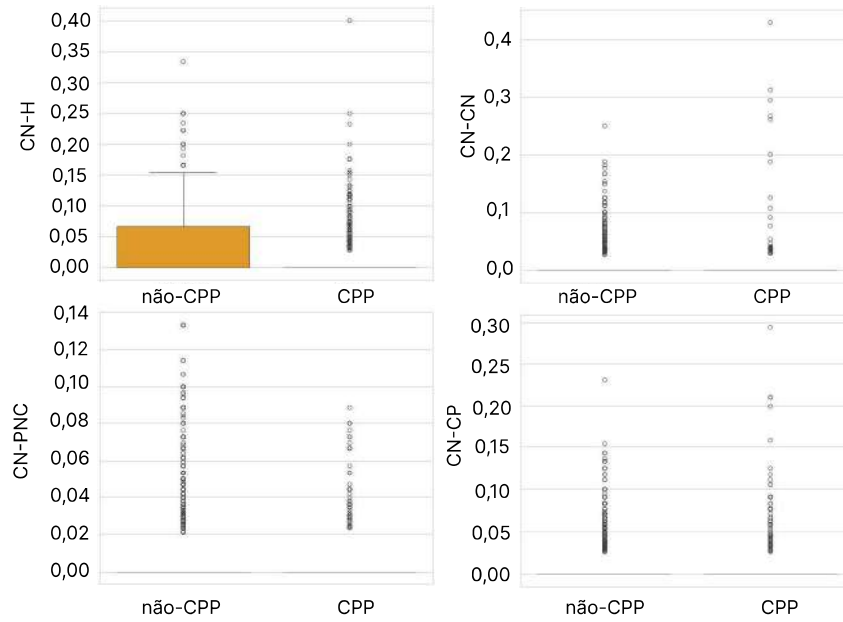


Figura 8 – Boxplot dos CPPs e não-CPPs baseado nas características normalizadas CK-SAAGP de resíduos com carga negativa no conjunto de dados de treinamento. **CN-H** representa a interação do grupo carregado negativamente com o grupo hidrofóbico; **CN-CN** representa a interação dos grupos carregados negativamente; **CN-PNC** representa a interação do grupo carregado negativamente com o grupo polar não carregado; **CN-CP** representa a interação do grupo carregado negativamente com grupo carregado positivamente.

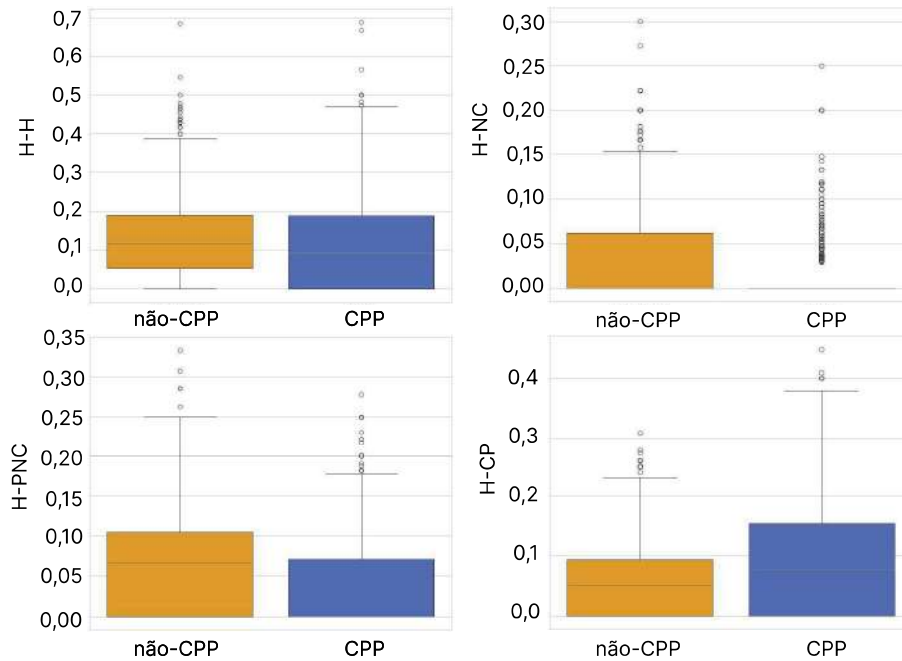


Figura 9 – Boxplot dos CPPs e não-CPPs baseado nas características normalizadas CK-SAAGP hidrofóbicas no conjunto de dados de treinamento. **H-H** representa a interação entre grupos hidrofóbicos; **H-NC** representa a interação do grupo hidrofóbico com o grupo carregado negativamente; **H-PNC** representa a interação do grupo hidrofóbico com o grupo polar não carregado; **H-CP** representa a interação do grupo hidrofóbico com o grupo carregado negativamente.

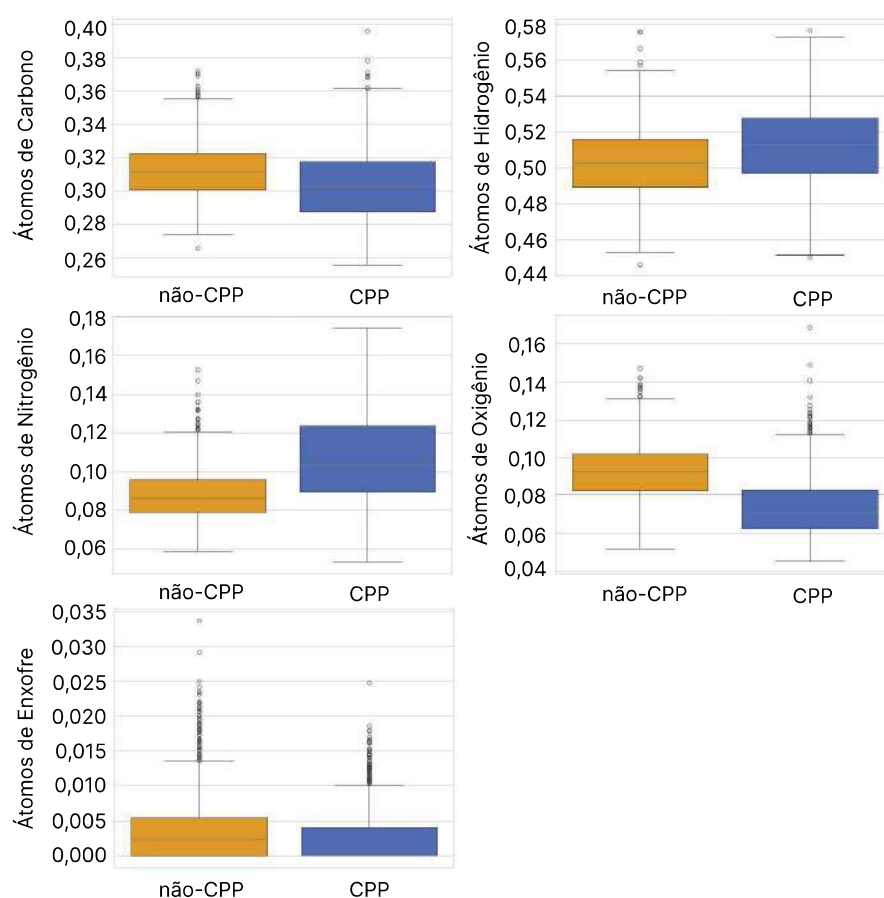


Figura 10 – Boxplot dos CPPs e não-CPPs baseado nas características normalizadas de composição atômica no conjunto de dados de treinamento.

Tabela 3 – Descritores pós análise de importância

Descritor	Quantidade	Importância
Físico-Químicos	5	25,51%
CKSAAGP	3	9,44%
Comp. Atômicos	5	15,41%
Dipeptídeos	285	35,03%
Tripeptídeos	193	14,58%

3.5 Métricas de desempenho

Para avaliar o desempenho dos modelos desenvolvidos e compará-los com abordagens previamente propostas na literatura, foram utilizadas métricas clássicas, essas permitem quantificar a capacidade dos algoritmos em distinguir corretamente entre peptídeos CPPs e não-CPPs, fornecendo uma visão abrangente sobre acertos, erros e equilíbrio nas predições. Todas as métricas dependem que utilizaremos no estudo dependem de quatro variáveis: TP (Verdadeiro Positivo) que representa o número de casos classificados corretamente como positivos, enquanto FN (Falso Negativo) refere-se aos casos incorretamente classificados como negativos, apesar de serem positivos. FP (Falso Positivo) indica o número de casos incorretamente classificados como positivos, sendo na verdade negativos, e TN (Verdadeiro Negativo) representa os casos corretamente classificados como negativos. A seguir, as métricas são apresentadas e descritas:

1. **Sensibilidade (SN)** mede a fração de instâncias positivas corretamente identificadas pelo classificador, variando de 0 a 1. Valores mais altos de SN indicam menos falsos negativos:

$$SN = \frac{TP}{TP + FN} \quad (3.1)$$

2. **Especificidade (SP)** mede a fração de instâncias negativas corretamente identificadas, também variando de 0 a 1, com valores mais altos indicando menos falsos positivos:

$$SP = \frac{TN}{FP + TN} \quad (3.2)$$

3. **Acurácia (ACC)** mede a correção geral das previsões do modelo, também variando de 0 a 1, mas pode não ser adequada para conjuntos de dados desbalanceados:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.3)$$

4. **Coeficiente de Correlação de Matthews (MCC)** é uma métrica robusta para problemas de classificação binária, especialmente útil em conjuntos de dados desbalanceados. Seu valor varia de -1 a +1, onde +1 indica uma predição perfeita, 0 representa uma classificação aleatória e -1 indica uma total discordância entre as predições e os rótulos reais (previsões inversas). O MCC leva em consideração todos os elementos da matriz de confusão (TP, TN, FP e FN), oferecendo uma medida equilibrada mesmo quando as classes têm distribuições muito diferentes:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(FP + TN)(TN + FN)}} \quad (3.4)$$

5. **Área sob a Curva (AUC)** plota a taxa de verdadeiros positivos contra a taxa

de falsos positivos e representa a capacidade do modelo de distinguir entre as classes. Um valor de AUC igual a 0,5 sugere ausência de capacidade discriminativa, enquanto valores crescentes acima de 0,5 indicam desempenho discriminativo progressivamente melhor. Com exceção da ACC, todas as métricas são robustas e apropriadas mesmo em cenários com conjuntos de dados desbalanceados.

3.6 Treinamento do modelo

Para a etapa de treinamento, já com as 491 features determinadas, foram aplicados os algoritmos Extremely Randomized Tree (ERT), Random Forest (RF), XGBoost (XGB), Support Vector Machine (SVM) e Multilayer Perceptron (MLP) ao nosso conjunto de dados de treinamento para identificar aquele que produziria os melhores resultados. Os valores das métricas de cada algoritmo são apresentados na Tabela 4. Realizamos uma validação cruzada de 10 partições (10-fold cross-validation), obtendo um desempenho médio que reflete a estabilidade dos modelos em diferentes configurações de dados de treinamento e teste.

Dessa forma, o modelo escolhido foi o ERT, por apresentar os melhores resultados, com os 491 descritores, que nomeamos de *PERSEUCPP*. Além disso, treinamos um modelo com somente os TOP 20 descritores com o maior valor de importância de classificação e nomeamos de *Perseu20*. A capacidade de classificação e de generalização do *PERSEUCPP* e *Perseu20* serão discutidos mais à frente. Adicionalmente, com o objetivo de capturar informações semânticas de alto nível, ou seja, padrões complexos relacionados à estrutura, função e contexto evolutivo das sequências, geramos representações vetoriais (embeddings) das proteínas por meio de um modelo de linguagem especializado. Para isso, utilizamos o modelo ESM-1b (ESM1-T6-43M-UR50S) (RIVES et al., 2021), que foi treinado em larga escala para aprender a “linguagem” das proteínas. Essas representações foram extraídas para o nosso conjunto principal de treinamento, originando o modelo *PerseuESM*.

Tabela 4 – Comparação de desempenho dos algoritmos de aprendizado de máquina para predição de CPPs utilizando o mesmo conjunto de dados. Apresentamos desempenho de cinco modelos de aprendizado de máquina, SVM, XGB, MLP, RF e ERT, na predição de peptídeos de penetração celular (CPPs).

Modelo	MCC	ACC	SN	SP	AUC	F1
SVM	0,192	0,594	0,725	0,466	0,647	0,540
XGB	0,657	0,826	0,769	0,898	0,901	0,872
MLP	0,686	0,841	0,536	0,901	0,906	0,852
RF	0,740	0,896	0,874	0,877	0,949	0,901
ERT	0,801	0,934	0,868	0,998	0,981	0,926

Nota: Os valores em negrito representam o melhor desempenho entre os competidores.

4 Resultados e discussões

Para demonstrar a habilidade de classificação do nosso modelo, foi realizado um conjunto de experimentos. Primeiro, avaliamos cinco diferentes algoritmos de machine learning (ERT, RF, XGB, SVM e MLP) para determinar qual seria o melhor para construir o PERSEUCPP, os resultados obtidos estão disponíveis na Tabela 4, esclarecendo a escolha do algoritmo ERT. Depois de escolhida a melhor estratégia, comparamos o modelo proposto com outros cinco preditores, incluindo o estado da arte. Nas seções seguintes, serão discutidos os resultados em relação a cada competidor.

4.1 CPP924 - Conjunto de dados independente

Para este tópico, o desempenho do PERSEUCPP foi comparado com cinco preditores de CPPs: CPPred-RF, TargetCPP, StackCPPred, SiameseCPP e PractiCPP. As comparações foram realizadas utilizando o conjunto independente de validação CPP924 para avaliar o desempenho do nosso método em relação aos preditores existentes. Esse conjunto de dados é balanceado entre CPPs e não-CPPs, compreendendo 924 peptídeos naturais, onde 462 CPPs são validados experimentalmente e 462 não-CPPs foram gerados artificialmente conforme discutido na *Seção 3.1*.

O SiameseCPP, conforme discutimos na *Subseção 2.2.6* é uma arquitetura de aprendizado profundo que aprende descritores discriminativos diretamente a partir das sequências peptídicas. O método foi construído com o conjunto de dados CPP924, no qual 80% foi utilizado para treinamento e 20% para teste, sendo então comparado com os concorrentes. Apesar de geralmente apresentar bons resultados, o modelo carece de interpretabilidade e apoio na compreensão das questões biológicas subjacentes aos dados.

De acordo com (RAMEZAN; WARNER; MAXWELL, 2019), construir e testar um modelo de classificação utilizando 80% dos dados para treinamento e 20% para teste é menos robusto do que a validação cruzada, havendo a possibilidade de que os resultados obtidos dessa forma sejam superestimados. No entanto, para uma comparação justa, utilizamos a mesma partição com 20% do conjunto de dados para testar nosso método frente aos concorrentes, já que essa foi a configuração utilizada pelo então estado da arte. A Tabela 5 apresenta a comparação entre o PERSEUCPP e os métodos concorrentes no CPP924. Nossa estratégia apresenta resultados superiores ao SiameseCPP nesse conjunto de dados.

O PractiCPP (KEXIN; YUANPENG et al., 2024) emprega amostragem altamente negativa (hard negative sampling) para aprimorar os limites de decisão e utiliza três tipos de descritores: descritores sequenciais derivados das sequências de aminoácidos, descri-

tores locais obtidos a partir da estratégia chamada *Morgan Finger Print* (que captura o ambiente químico ao redor de cada átomo), e embeddings pré-treinados do modelo de linguagem ESM-2, conforme discutido na Subseção 2.2.7. O PractiCPP supera modelos de última geração como o SiameseCPP e o MLCPP2.0 tanto em conjuntos de dados balanceados quanto desbalanceados. Por outro lado, nosso modelo supera o PractiCPP no conjunto de dados CPP924, conforme mostrado na Tabela 5. No entanto, não foi possível realizar uma comparação justa utilizando o conjunto de dados desbalanceado, pois ele não foi disponibilizado juntamente com o código-fonte no artigo do PractiCPP.

Tabela 5 – Comparação entre PERSEUCPP e os métodos competidores no dataset independente CPP924

Method	MCC	ACC	SN	SP	AUC
PERSEUCPP20	0,745	0,878	0,764	0,844	0,933
PerseuESM	0,869	0,934	0,894	0,926	0,973
CPPred-RF	0,831	0,916	0,905	0,926	-
TargetCPP	0,871	0,935	0,943	0,936	-
StackCPPred	0,890	0,945	0,942	0,948	-
PractiCPP	0,913	0,956	0,943	0,971	-
SiameseCPP	0,923	0,961	0,959	0,964	-
PERSEUCPP	0,940	0,970	0,961	0,978	0,975

Nota: Os valores em negrito representam o melhor desempenho entre os competidores. PERSEUCPP20 representa o modelo treinado somente com as TOP 20 features e PerseuESM representa o modelo com embeddings pré-treinados do modelo de linguagem ESM-2. Os valores de AUC não foram reportados pelos competidores.

4.2 MLCPP2.0 - Conjunto de dados independente

Nesta etapa, o desempenho do PERSEUCPP foi comparado com quatro preditores de CPPs: C2Pred, BchemRF, MLCPP2.0 e SiameseCPP. Para esse fim, foi utilizado o conjunto de validação MLCPP2.0, que também foi empregado como referência pelos quatro métodos mencionados. Esse conjunto de dados é composto por 157 CPPs (validados experimentalmente) e 2.184 não-CPPs (onde maioria foi gerado artificialmente).

Os resultados apresentados na Tabela 6 demonstram que o PERSEUCPP supera os métodos C2Pred, BchemRF, MLCPP2.0 e também o SiameseCPP. Essa superioridade destaca a eficácia do PERSEUCPP e seu potencial de aplicação prática na predição de CPPs.

Tabela 6 – Comparação entre PERSEUCPP e os métodos competidores no dataset independente MLCPP2.0

Method	MCC	ACC	SN	SP	AUC
C2Pred	0.326	0.781	0.790	0.781	0.867
BchemRF	0.467	0.893	0.745	0.908	0.914
MLCPP2.0	0.624	0.934	0.847	0.940	0.928
SiameseCPP	0.652	0.959	0.624	0.983	0.980
PerseuESM	0.794	0.970	0.904	0.973	0.969
PerseuCPP20	0.736	0.974	0.871	0.979	0.972
PerseuCPP	0.854	0.989	0.860	0.994	0.984

Nota: Os valores em negrito representam o melhor desempenho entre os competidores. PERSEUCPP20 representa o modelo treinado somente com as TOP 20 features e PerseuESM representa o modelo com embeddings pré-treinados do modelo de linguagem ESM-2. Os valores de AUC não foram reportados pelos competidores.

4.3 CPP1708 - Conjunto de dados independente

Nesta seção, o desempenho do PERSEUCPP foi comparado com quatro métodos de predição: CellPPD, C2Pred, MLCPP2.0 e GraphCPP. Este conjunto de dados é composto por um conjunto diverso de 90 CPPs e 98 não-CPPs, projetado para testar a capacidade de generalização dos modelos preditivos.

O GraphCPP (ATILLA; BALAZS et al., 2025) utiliza redes neurais gráficas (GNNs) para prever CPPs, modelando as sequências peptídicas como grafos moleculares. Esse método captura relações intrincadas entre os aminoácidos e suas propriedades físico-químicas. Ao integrar descritores como nós e arestas, o GraphCPP fornece embeddings robustos que aumentam a precisão na predição de CPPs.

Os resultados apresentados na Tabela 7 demonstram que o PERSEUCPP apresenta resultados compatíveis ou superiores em comparação com o concorrente de estado da arte, evidenciando sua robusta capacidade de predição mesmo sob condições desafiadoras.

4.4 Resumo dos resultados das comparações

De modo geral, a versão do PERSEUCPP treinada com os 491 descritores tradicionais apresentou, de forma consistente, os melhores desempenhos nos três conjuntos de teste independentes, MLCPP2.0, CPP924 e CPP1708, quando comparada com modelos de última geração e com nossas próprias versões utilizando ESM. Notavelmente, embora a versão baseada em ESM (PerseuESM) tenha apresentado bom desempenho, especialmente nos conjuntos MLCPP2.0 e CPP924, ela não superou o modelo com 491 descritores em nenhum dos três benchmarks. Essa consistência destaca a robustez e a generalidade

Tabela 7 – Comparação entre PERSEUCPP e os métodos competidores no dataset de validação CPP1708

Method	MCC	ACC	PR	SN	F1	AUC
CellPPD	0,0000	0,59	0,0	0,0	0,0	-
C2Pred	0,0540	0,571	0,458	0,256	0,328	0,6500
MLCPP2.0	0,3001	0,667	0,786	0,256	0,386	0,6860
GraphCPP	0,5787	0,795	0,776	0,731	0,752	0,8459
PerseuESM	0,652	0,811	0,979	0,595	0,740	0,853
PERSEUCPP20	0,515	0,754	0,846	0,556	0,671	0,849
PERSEUCPP	0,646	0,811	0,960	0,607	0,744	0,8652

Nota: Os valores em negrito representam o melhor desempenho entre os competidores. PERSEUCPP20 representa o modelo treinado somente com as TOP 20 features e PerseuESM representa o modelo com embeddings pré-treinados do modelo de linguagem ESM.

do modelo PERSEUCPP construído com base em descritores físico-químicos, estruturais e atômicos, que são ao mesmo tempo biologicamente interpretáveis e computacionalmente eficientes. Esses resultados sustentam a decisão de adotar essa configuração como a versão principal do modelo, oferecendo um equilíbrio favorável entre desempenho, interpretabilidade e aplicabilidade prática. Em termos de desempenho geral, o PERSEUCPP apresenta resultados compatíveis ou superiores a todos os métodos previamente propostos.

4.5 Interpretação dos descritores

A primeira parte deste estudo concentrou-se nos efeitos da combinação de características físico-químicas, estruturais e atômicas na classificação de CPPs. O algoritmo Extra Trees foi fundamental nesse processo, fornecendo insights sobre a importância de cada característica. A Figura 11 apresenta os 20 principais descritores, destacando seu papel relevante no processo de classificação.

O *Ponto Isoelétrico* e a *Carga Líquida* são descritores do grupo de propriedades físico-químicas e exercem grande influência na caracterização dos CPPs. Na Figura 12, foi confirmado que os CPPs analisados apresentaram predominantemente valores de ponto isoelétrico superiores a sete ($pI > 7$), sendo, portanto, em sua maioria catiônicos em condições fisiológicas. De acordo com a literatura, a membrana celular é composta principalmente por fosfolipídios, os quais podem apresentar grupos fosfato e grupos substituintes carregados negativamente em suas cabeças polares (ALBERTS et al., 2002), o que facilita a interação dos CPPs com as membranas celulares, especialmente em bactérias, confirmando a relevância dessas características na classificação dos CPPs.

Arginina (R) e lisina (K) estão entre os resíduos relevantes para a eficácia dos peptídeos de penetração celular, uma vez que são catiônicos. CPPs ricos em arginina

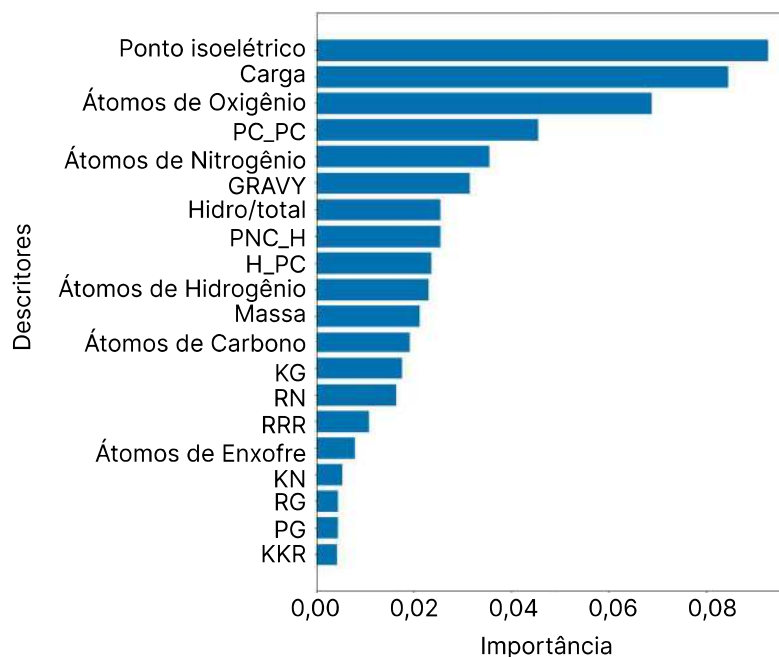


Figura 11 – 20 principais descritores independentes para a classificação de CPPs identificados pelo algoritmo Extremely Randomized Trees. As abreviações PC, PNC e H representam os grupos de aminoácidos: carregados, polares não carregados e hidrofóbicos, respectivamente.

mostram-se promissores para a entrega de macromoléculas terapêuticas, como peptídeos, proteínas e ácidos nucleicos, no interior das células (SCHMIDT; MISHRA et al., 2009). A importância da arginina nos CPPs é evidenciada por seu envolvimento em interações eletrostáticas e sua contribuição para a afinidade de ligação dos CPPs com as membranas celulares (SCHMIDT; MISHRA et al., 2009). Além disso, peptídeos ricos em lisina demonstraram maior atividade de penetração celular em comparação com outros peptídeos (KIM; CHOI et al., 2018). É, portanto, razoável que o modelo tenha identificado a arginina e a lisina como descritores críticos para a classificação, conforme mostrado na Figura 11. Esses aminoácidos se destacam nos dipeptídeos *KG*, *RN*, *KN* e *RG* nas posições 13, 14, 17 e 18 da Figura 11, respectivamente, assim como nos tripeptídeos *RRR* e *KKR* nas posições 15 e 20.

A utilidade da composição de dipeptídeos se estende à predição da localização subcelular de proteínas (LI; GAO, 2020), à identificação de proteínas tóxicas bacterianas (SU et al., 2014) e ao design de peptídeos de penetração celular altamente eficientes (GAUTAM; CHAUDHARY et al., 2013). Em especial na busca por CPPs, a composição de dipeptídeos tem sido utilizada para diferenciar CPPs de não-CPPs (GAUTAM; CHAUDHARY et al., 2013). Como pode ser observado na Figura 13, a razão CPPs/não-CPPs para os dipeptídeos *KG*, *RN*, *KN*, *RG* e *PG* sugere que a presença de *KN* e *PG* é mais comum em CPPs, enquanto *KG*, *RN* e *RG* estão mais presentes em não-CPPs.

Cada resíduo de aminoácido possui uma composição específica de átomos de carbono (nC), hidrogênio (nH), nitrogênio (nN) e oxigênio (nO), e alguns também apresen-

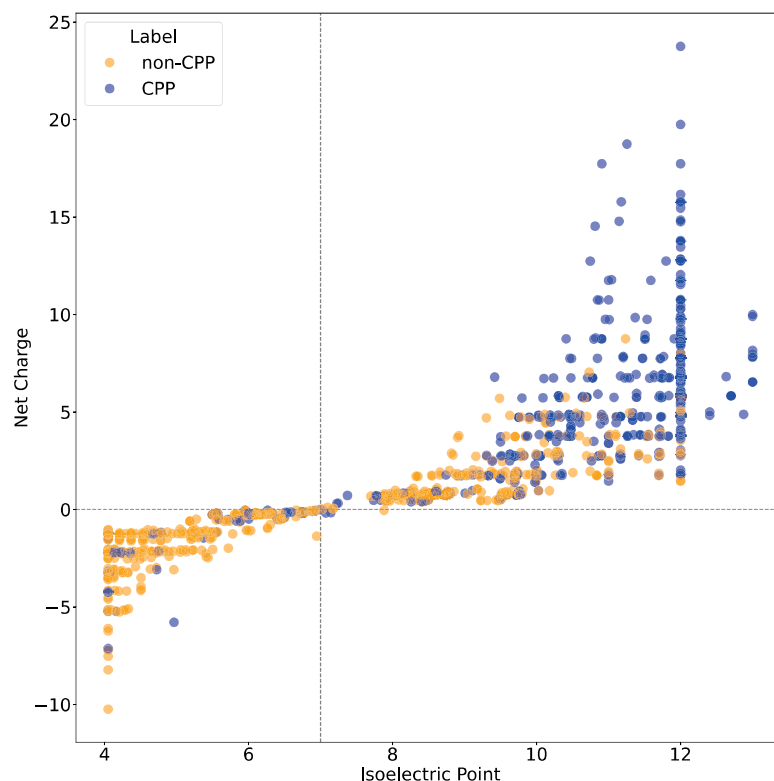


Figura 12 – Scatter plot dos CPPs e não-CPPs com base no ponto isoelétrico e na carga líquida no conjunto de dados de treinamento. Os CPPs são predominantemente peptídeos catiônicos, apresentando valores de ponto isoelétrico superiores a 7 em condições fisiológicas, enquanto grande parte dos não-CPPs são peptídeos aniônicos, com valores de ponto isoelétrico inferiores a 7. A linha tracejada em $x = 7$ indica o ponto isoelétrico, destacando onde se inicia a transição de carga.

tam enxofre (nS), como é o caso da cisteína (C) e da metionina (M). Neste trabalho, as frequências desses átomos normalizada pelo comprimento da cadeia foram utilizadas como descritores. Observa-se na Figura 11 que os descritores *Átomos de Oxigênio* e *Átomos de Nitrogênio* desempenham um papel importante na classificação. A relação entre o número de átomos de nitrogênio e oxigênio e a capacidade de penetração celular dos peptídeos tem sido investigada, destacando a importância desses átomos na potencialização da função peptídica (KALAFATOVIC; GIRALT, 2017; HAO; ZHANG et al., 2022). Especificamente, peptídeos ricos em arginina, que contêm múltiplos átomos de nitrogênio em seus grupos guanidínio, demonstraram capacidade efetiva de penetrar membranas celulares. Isso se deve principalmente à sua habilidade de formar ligações de hidrogênio e interagir eletrostaticamente com componentes da membrana celular carregados negativamente, como fosfolipídios e glicosaminoglicanos (HAO; ZHANG et al., 2022).

Além disso, a presença de átomos de oxigênio em peptídeos específicos, frequentemente na forma de grupos hidroxila ou carbonila, pode contribuir para sua natureza anfipática, a qual é essencial para o processo de translocação através da membrana. Esses grupos facilitam interações com as regiões hidrofílica e hidrofóbica da membrana,

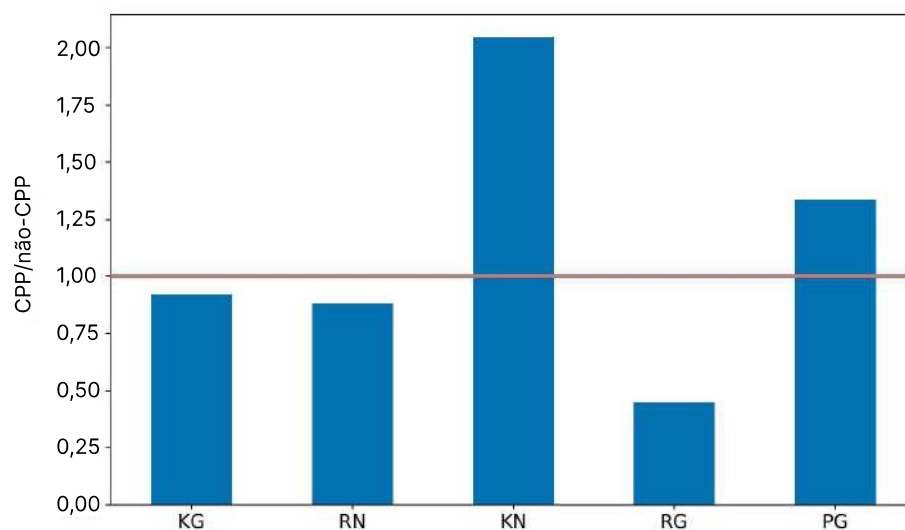


Figura 13 – A razão das frequências dos descritores KG, RN, KN, RG e PG entre os grupos CPP e não-CPP. Cada barra no gráfico representa a razão (Frequência em CPP / Frequência em não-CPP) das frequências de uma característica específica, calculada como a frequência da característica no grupo CPP dividida pela frequência da mesma característica no grupo não-CPP. Valores acima de 1 sugerem que a frequência da característica é maior no grupo CPP. Valores abaixo de 1 sugerem que a frequência da característica é maior no grupo não-CPP.

melhorando, assim, a capacidade do peptídeo de atravessar a bicamada lipídica (KALAFATOVIC; GIRALT, 2017).

Vale destacar que a quantidade de nitrogênio é maior em CPPs quando comparada aos não-CPPs, enquanto a quantidade de oxigênio é superior nos não-CPPs, conforme mostrado na Figura 14. Esse resultado é justificado pelos radicais dos resíduos de aminoácidos, que podem conter grupos ionizáveis, tornando-se positivos pela protonação dos grupos amino (na presença de nitrogênio) ou tornando-se negativos pela desprotonação dos grupos carbonila ou hidroxila (na presença de oxigênio) em pH fisiológico. Dessa forma, peptídeos catiônicos são ricos em nitrogênio, enquanto peptídeos aniônicos são ricos em oxigênio.

Conforme mostrado na Figura 11, os descritores PC_PC , PNC_H e H_PC , que fazem parte do CKSAAGP (Composição de Pares de Grupos de Aminoácidos com Espaçamento k) com distância de aminoácidos $S=1$, também exerceram forte influência na classificação. Essa característica captura a frequência de pares de grupos de aminoácidos separados por um resíduo, refletindo interações proximais que podem influenciar diretamente a estrutura e a função dos peptídeos. Especificamente, interações entre cargas positivas com cargas positivas (PC_PC), grupos polares não carregados com hidrofóbicos (PNC_H) e hidrofóbicos com cargas positivas (H_PC) são descritas na literatura como relevantes para a capacidade de penetração celular (MULLARD, 2023). Essas interações afetam propriedades cruciais como afinidade com a membrana celular, estabilidade con-

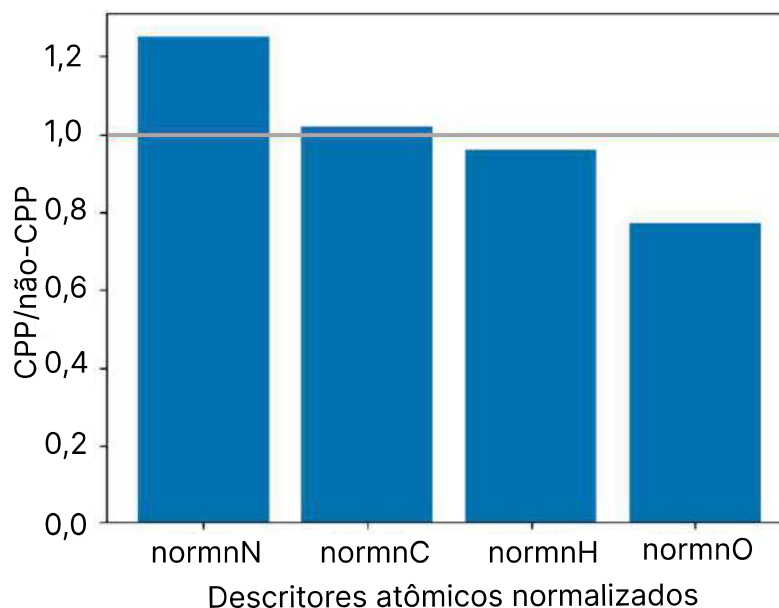


Figura 14 – Razão entre as médias normalizadas dos descritores normnN, normnO, normnH e normnC entre os grupos CPP e não-CPP. Cada barra no gráfico representa a razão (Frequência em CPP / Frequência em não-CPP) da média normalizada de uma característica específica, calculada como a média normalizada da característica no grupo CPP dividida pela média normalizada da mesma característica no grupo não-CPP. Valores acima de 1 sugerem que a média normalizada da característica é maior no grupo CPP. Valores abaixo de 1 indicam que a média normalizada da característica é maior no grupo não-CPP.

formacional e solubilidade, contribuindo significativamente para a eficiência de penetração dos peptídeos.

A Figura 15 ilustra a contribuição relativa dos diferentes grupos de descritores na classificação de CPPs pelo PERSEUCPP. A análise revela que *dipeptídeos* e *tripeptídeos* exercem influência significativa na tarefa de classificação. No entanto, quando analisados isoladamente, ou seja, descritor por descritor, tendem a apresentar pouca influência, o que era esperado, considerando que existem 20^2 dipeptídeos e 20^3 tripeptídeos. Outros descritores apresentam impacto significativo mesmo quando considerados individualmente. Por exemplo, o *ponto isoelétrico*, que faz parte do grupo de descritores físico-químicos, apresenta uma contribuição individual de quase 10%. Além disso, os demais descritores físico-químicos, como *carga líquida*, *hidrofobicidade*, *gravy* e *massa*, também desempenham um papel importante, especialmente para peptídeos catiônicos que interagem com regiões aniônicas dos fosfolípidios na superfície celular (ALBERTS et al., 2002). Adicionalmente, a *composição atômica* mostrou uma contribuição moderada, sugerindo que elementos específicos, como nitrogênio e oxigênio, influenciam as interações eletrostáticas e a formação de ligações de hidrogênio com componentes da membrana (KALAFATOVIC; GIRALT, 2017; HAO; ZHANG et al., 2022). De modo geral, essa análise dos descritores segmentados em categorias indica que cada grupo possui uma contribuição relevante para

a predição de CPPs, ressaltando a importância de descritores que capturem propriedades físico-químicas, composição de resíduos e também algum contexto (como capturado por di e tripeptídeos), uma vez que todos esses aspectos estão relacionados à capacidade de carregar moléculas terapêuticas para dentro da célula.

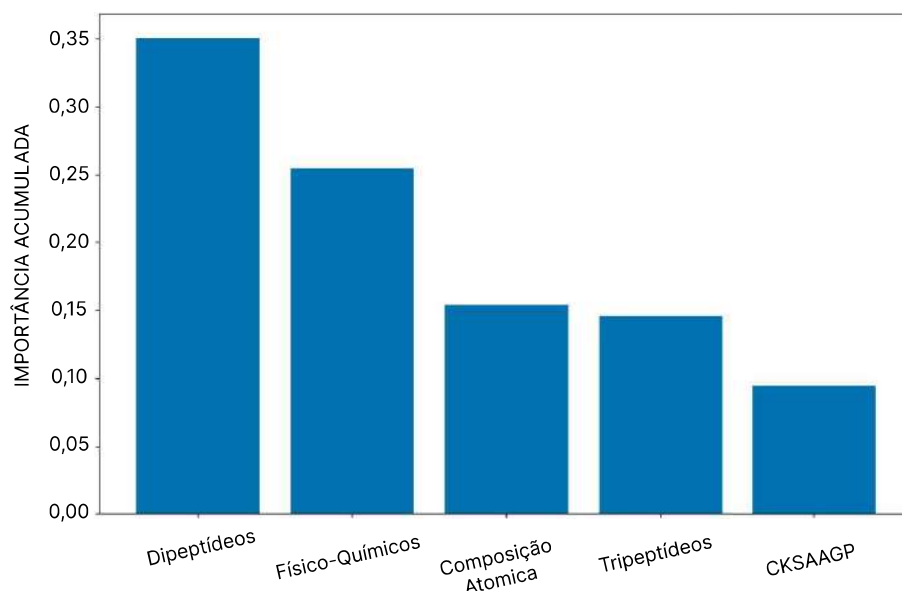


Figura 15 – Contribuição relativa dos diferentes grupos de descritores na classificação de CPPs. Cada barra representa a contribuição do grupo de descritores, em que dipeptídeos representam grupos de dois aminoácidos e tripeptídeos representam grupos de três. O grupo físico-químico inclui descritores como índice GRAVY, Massa Molecular, Ponto Isoelétrico, Hidrofobicidade e Carga Líquida. A Composição Atômica refere-se aos átomos de nitrogênio, carbono, hidrogênio, oxigênio e enxofre. CKSAAGP representa a Composição de Pares de Grupos de Aminoácidos com Espaçamento k .

Por fim, analisamos também a separação dos dados através da visualização t-distributed Stochastic Neighbor Embedding (t-SNE), que é uma estratégia de redução de dimensionalidade utilizada para visualizar dados complexos e de alta dimensão. A análise mostrou uma separação significativa entre CPPs e não-CPPs, conforme apresentado na Figura 16. Todos os 20 principais descritores (TOP20) foram utilizados para gerar essa visualização t-SNE. A combinação desses descritores separa de forma eficaz os CPPs dos não-CPPs, demonstrando seu poder discriminativo. Assim, a combinação dos descritores discutidos acima com as contribuições dos demais 471 descritores, juntamente com o algoritmo ERT, permitiu ao PERSEUCPP alcançar resultados de alta qualidade e superar o estado da arte.

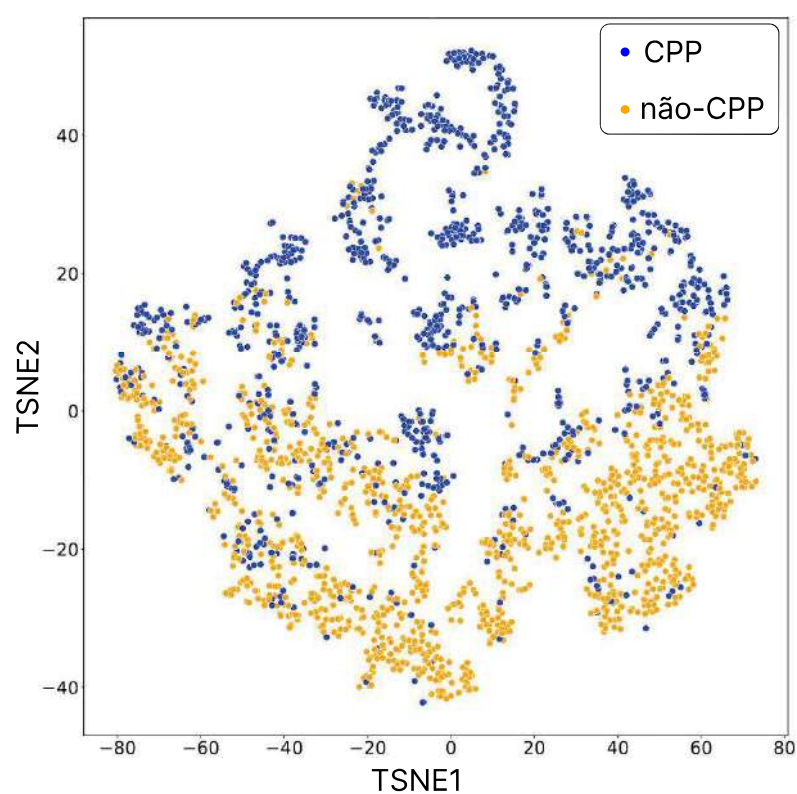


Figura 16 – Visualização t-SNE de CPPs e não-CPPs. Cada ponto representa um peptídeo descrito por um conjunto de descritores calculados. Os pontos são coloridos com base em sua classificação: azul para CPPs e laranja para não-CPPs. A separação distinta entre os agrupamentos azul e laranja sugere que as características selecionadas diferenciam de forma eficaz os CPPs dos não-CPPs.

4.6 Predições Incorretas

Considerando os conjuntos de dados independentes CPP924, MLCPP2.0 e CPP1708, nosso modelo classificou incorretamente 98 de 2.699 sequências, sendo 5 do CPP924, 35 do MLCPP2.0 e 58 do CPP1708. A Tabela 8 mostra o número de previsões incorretas para cada conjunto de dados, considerando CPPs e não-CPPs.

Para ilustrar, discutimos brevemente os descritores ponto isoelétrico e carga líquida, que são os descritores mais importantes quando analisados separadamente. Na Figura 17A, a mediana do ponto isoelétrico das sequências positivas do conjunto de treino (caixa azul rotulada como CPPs) apresenta um valor próximo à mediana do ponto isoelétrico das sequências classificadas erroneamente (caixa laranja rotulada como não-CPPs preditos como CPPs) na Figura 17B. Um comportamento semelhante é observado ao se analisar a carga, pois na Figura 17C a mediana da carga das sequências positivas do treino (caixa azul rotulada como CPPs) está próxima da mediana das sequências classificadas erroneamente (caixa laranja rotulada como não-CPPs preditos como CPPs) na Figura 17D. Assim, aparenta que o que causa confusão em nosso modelo é que as sequências negativas classificadas erroneamente (não-CPPs preditos como CPPs) apresentam descritores semelhantes aos das sequências positivas do conjunto de treino.

Embora modelos de aprendizado de máquina possam ter tendência a serem mais sensíveis a falsos positivos ou falsos negativos, nosso modelo superou todos os concorrentes considerando os conjuntos de dados de referência utilizados na literatura para validação dos resultados. Obtivemos instâncias classificadas incorretamente nos conjuntos independentes CPP924, MLCPP2.0 e CPP1708, incluindo tanto falsos positivos quanto falsos negativos. No entanto, isso não é necessariamente um erro, devido à natureza das sequências negativas. Dada a escassez de dados, muitas dessas sequências foram geradas automaticamente com base em sequências negativas validadas experimentalmente, sendo assim, essas instâncias podem ser sequências positivas reais ainda não determinadas experimentalmente. Nesse caso, nosso método não estaria cometendo um erro de predição, mas sim identificando um CPP ainda não determinado experimentalmente.

Tabela 8 – Previsões incorretas no PERSEUCPP. Freq. representa o número de sequências nos subconjuntos CPP e não-CPP, CP são as previsões corretas e IP são as previsões incorretas

Base	Label	Freq.	CP	IP
CPP924	CPP	92	88	4
	non-CPP	93	92	1
MLCPP	CPP	157	150	7
	non-CPP	2170	2142	28
CPP1708	CPP	79	50	29
	non-CPP	96	67	29

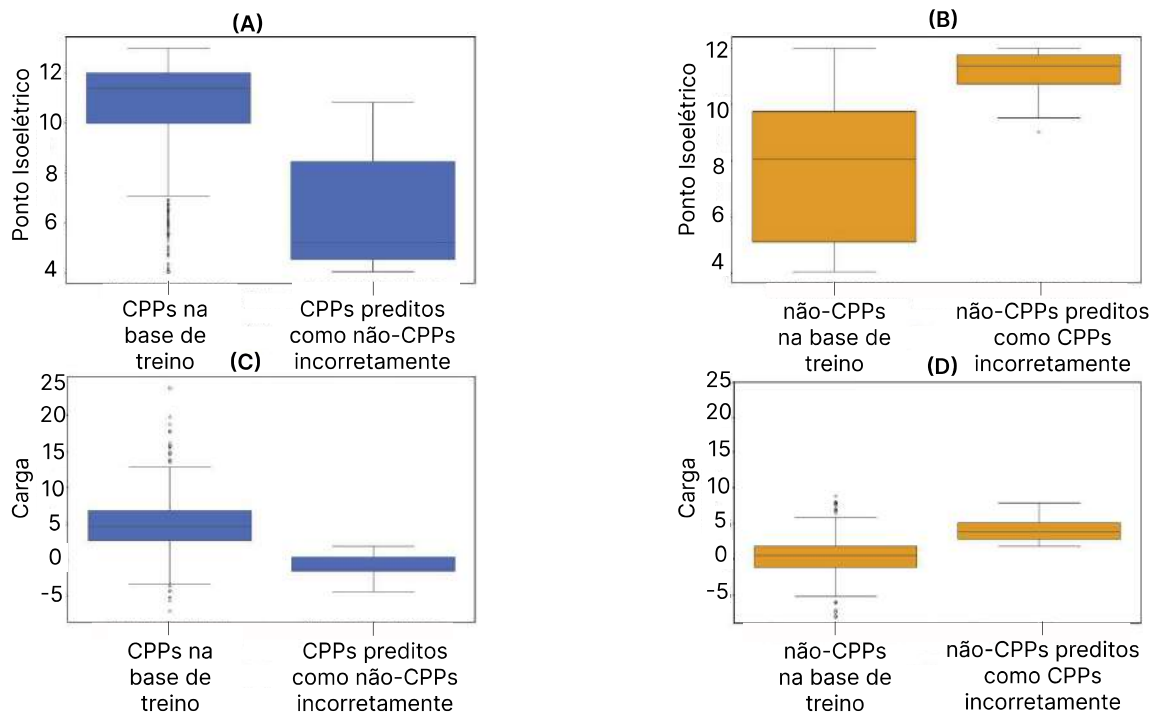


Figura 17 – Comparação do Ponto Isoelétrico e da Carga Líquida entre as Sequências de Treinamento e as Sequências Classificadas Incorretamente. **A** apresenta a distribuição dos pontos isoelétricos para as sequências positivas do conjunto de treinamento e as sequências classificadas incorretamente. **B** mostra a distribuição do ponto isoelétrico para as sequências negativas do treinamento e as sequências classificadas incorretamente. **C** exibe a distribuição da carga líquida para as sequências positivas do treinamento e as sequências classificadas incorretamente. **D** representa a distribuição da carga líquida para as sequências negativas do treinamento e as sequências classificadas incorretamente.

4.7 Classificador de eficiência

A segunda etapa do PERSEUCPP foi responsável por avaliar a eficiência da penetração dos peptídeos. Em outras palavras, além de categorizar um peptídeo como CPP ou não-CPP, o método proposto avaliou quão eficiente o peptídeo é em termos de internalização.

4.7.1 PERSEUCPP comparado com o estado da arte de eficiência de penetração

Para construir o preditor de eficiência do PERSEUCPP, utilizamos a composição atômica (o número de átomos de carbono, hidrogênio, nitrogênio, oxigênio e enxofre normalizado pelo comprimento da cadeia). Para o nosso método, a composição atômica desempenhou um papel importante na melhoria dos resultados. O preditor de eficiência MLCPP2.0, considerado à época o método de estado da arte, foi treinado com 140 sequências de alta efetividade de penetração e 140 de baixa efetividade, ambas validadas experimentalmente. O modelo focou em descritores como Composição de Pares de Grupos de Aminoácidos com Espaçamento k (CKSAAGP), Composição de Dipeptídeos (DPC) e Composição de Aminoácidos (AAC).

O PERSEUCPP obteve sucesso na predição da eficiência dos CPPs, apresentando resultados compatíveis ou superiores em comparação com o estado da arte, conforme apresentado na Tabela 9.

Tabela 9 – PERSEUCPP comparado com o estado da arte de predição de eficiência

Method	MCC	ACC	SN	SP	AUC
MLCPP2.0	0,354	0,677	0,652	0,750	0,708
PERSEUCPP-Atm	0,356	0,726	0,761	0,625	0,690

Nota: Atm indica que o modelo foi treinado somente com os descritores atômicos. Os valores em negrito mostram o melhor desempenho obtido.

4.7.2 Interpretação dos descritores

Como estamos lidando exclusivamente com CPPs e tentando distingui-los entre alta e baixa eficiência, uma variedade de experimentos foi conduzida para avaliar qual grupo de descritores é capaz de melhorar a classificação. Os resultados desses experimentos estão disponíveis nas Tabelas 10-11. Dessa forma, o conjunto de descritores selecionado para treinar o modelo de classificação da eficiência de um CPP foi o de frequências atômicas, pois apresentou os melhores resultados, veja Tabela 9. A Figura 18, fornece o detalhamento da importância de cada característica atômica.

Tabela 10 – Resultados da comparação do PERSEUCPP proposto. ALL representa todas as características normalizadas, TRIPEP é o grupo de tripeptídeos, DIPEP é o grupo de dipeptídeos, CKSAAGP representa a Composição de Pares de Grupos de Aminoácidos com Espaçamento k , PC é o grupo de características físico-químicas e ATM é o grupo de características de composição atômica

Descriptors	MCC	ACC	SN	SP	AUC
ALL Features	0,273	0,451	0,533	0,476	0,501
TRIPEP	-0,092	0,451	0,456	0,437	0,465
DIPEP	0,314	0,645	0,608	0,750	0,692
CKSAAGP	-0,052	0,580	0,695	0,250	0,565
PC	0,256	0,705	0,761	0,625	0,690
PC + ATM	0,144	0,645	0,717	0,437	0,675
PC + ATM + CKSA	0,145	0,645	0,700	0,423	0,631
ATM + CKSA	-0,032	0,564	0,652	0,312	0,559
ATM + DIP	-0,032	0,564	0,652	0,312	0,559
ATM + TRIP	-0,096	0,419	0,391	0,500	0,559
ATM	0,356	0,726	0,761	0,625	0,690

Tabela 11 – Resultados da comparação do PERSEUCPP proposto com as TOP-N características de cada grupo de descritores.

Descriptors	MCC	ACC	SN	SP	AUC
TOP10	0,128	0,548	0,521	0,625	0,522
TOP50	-0,151	0,403	0,391	0,437	0,493
TOP100	-0,054	0,483	0,500	0,437	0,497
TOP150	-0,112	0,435	0,434	0,437	0,493
TOP200	-0,019	0,483	0,478	0,500	0,504
TOP250	0,019	0,516	0,521	0,500	0,513
TOP300	0,003	0,500	0,478	0,562	0,539

Um aspecto importante do desempenho do nosso modelo são suas métricas superiores em sensibilidade (SN) e acurácia (ACC), além de uma AUC comparável à do MLCPP2.0. Nosso modelo alcançou uma SN de 0,761, representando um aumento de 16,72% em relação à SN do MLCPP2.0, que é 0,652. A SN, que mede a capacidade do modelo de prever corretamente instâncias de CPPs eficientes dentro de todas as instâncias positivas existentes, é uma métrica relevante para o nosso modelo, pois garante uma identificação mais precisa de CPPs eficientes. Adicionalmente, nosso modelo apresenta uma AUC de 0,690, valor próximo ao alcançado pelo MLCPP2.0.

Embora nossos resultados para a predição da eficiência de internalização de CPPs sejam promissores, é importante enfatizar que os conjuntos de dados de treinamento e teste para essa tarefa ainda são relativamente pequenos. Sendo assim, recomendamos cautela na interpretação desses resultados ou na formulação de conclusões definitivas até que validações adicionais possam ser realizadas com conjuntos de dados independentes

maiores e mais diversos.

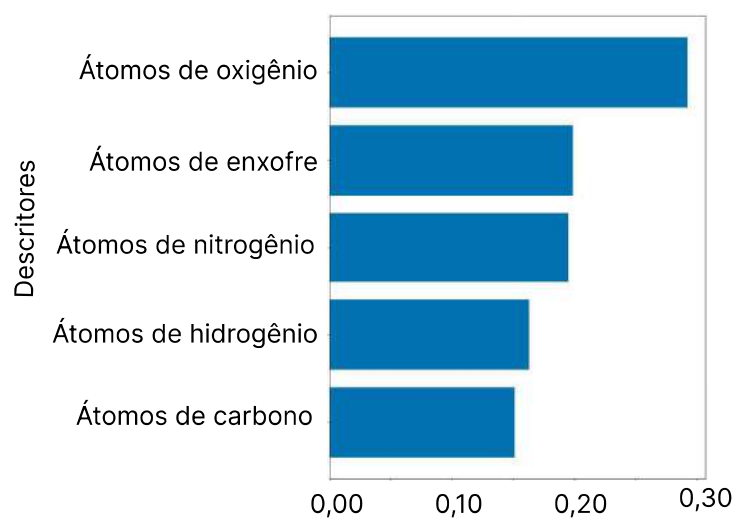


Figura 18 – Importância de cada descritor atômico do preditor de eficiência. Cada barra representa a importância extraída pelo modelo ERT, em ordem temos: átomos de oxigênio, enxofre, nitrogênio, hidrogênio e carbono respectivamente

5 Conclusões

Neste estudo, apresentamos o PERSEUCPP, uma estratégia baseada em aprendizado de máquina desenvolvida especificamente para a predição de peptídeos de penetração celular e da respectiva eficiência de penetração. Embora o algoritmo de base, o Extremely Randomized Trees, já seja conhecido na literatura, a proposta deste trabalho reside na modelagem inovadora que combina um conjunto abrangente e bem fundamentado de descritores com uma abordagem rigorosa de seleção de características. Essa combinação estratégica resultou em um modelo computacional que se destaca tanto pelo alto desempenho preditivo quanto pela capacidade de interpretação biológica dos resultados.

Ao empregar descritores que contemplam propriedades físico-químicas, características atômicas e padrões sequenciais, o PERSEUCPP demonstrou desempenho superior aos métodos de estado da arte em todas as métricas de classificação nos conjuntos independentes MLCPP2.0, CPP924 e CPP1708. A diversidade e a qualidade dos descritores utilizados não apenas contribuíram para a capacidade de predição do modelo, como também favoreceram a extração de interpretações biológicas relevantes, auxiliando na compreensão das características mais determinantes para a penetração celular.

Uma contribuição central deste trabalho é o resgate e valorização de métodos clássicos de aprendizado de máquina, frequentemente preteridos frente a modelos mais complexos, como redes neurais profundas e grafos. Embora métodos modernos tenham grande potencial, sua complexidade computacional e natureza opaca podem limitar sua aplicabilidade em contextos nos quais a eficiência computacional e a interpretabilidade são desejáveis. Ao compararmos o PERSEUCPP com preditores baseados em deep learning e redes neurais de grafos (GNNs), mostramos que métodos clássicos, quando cuidadosamente modelados, podem não apenas competir de forma justa, como também superar soluções mais complexas em desempenho, transparência e custo computacional.

Além disso, o PERSEUCPP possui uma segunda etapa especializada na predição da eficiência de penetração dos CPPs, classificando-os em categorias de alta ou baixa eficiência. Nessa fase, nosso modelo também obteve resultados comparáveis ao principal método da literatura, reforçando sua robustez e aplicabilidade.

Por fim, este trabalho introduz uma estrutura que alia uma boa capacidade preditiva à interpretação biológica, permitindo uma aplicação prática mais ampla. Uma abordagem explicativa orientada por descritores, possibilita não apenas a previsão de CPPs, mas também o entendimento aprofundado de relações estrutura-função, o que pode orientar estudos futuros sobre toxicidade, mecanismos de ação e o design racional de novos peptídeos.

Referências

- AGRAWAL, P. et al. Cppsite 2.0: a repository of experimentally validated cell penetrating peptides. **Nucleic Acids**, v. 44, n. D1, 2015.
- ALBERTS, B. et al. **Molecular Biology of the Cell. 4th Edition. The Lipid Bilayer**. New York: Garland Science, 2002.
- ATILLA, I.; BALAZS, B. et al. Graphcpp: The new state-of-the-art method for cell-penetrating peptide prediction via graph neural networks. **Br J Pharmacol**, v. 182(3), p. 495–509, 2025. ISSN 0893-6080.
- CDC. **Antibiotic Resistance Threats in the United States, 2019**. Atlanta, GA: U.S. Department of Health and Human Services, CDC, 2019.
- CHOPRA, S.; HADSELL, R.; LECUN, Y. Learning a similarity metric discriminatively, with application to face verification. In: **2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)**. [S.l.: s.n.], 2005. v. 1, p. 539–546 vol. 1.
- COCK, P. et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. **Bioinformatics**, v. 25, n. 11, p. 1422–1423, 2009. ISSN 1367-4803.
- DAINA, A.; ZOETE, V. A boiled-egg to predict gastrointestinal absorption and brain penetration of small molecules. **ChemMedChem**, v. 11, p. 1117–1121, 2016.
- DONGES, N. **Random Forest: A Complete Guide for Machine Learning**. 2024. <<https://builtin.com/data-science/random-forest-algorithm>>. Acesso em: 1 jun. 2025.
- DUBCHAK, I.; MUCHNIK, I. et al. Prediction of protein folding class using global description of amino acid sequence. **Proceedings of the National Academy of Sciences**, v. 92, n. 19, p. 8700–8704, 1995.
- ER, S.; LARAIB, U. et al. Amino acids, peptides, and proteins: Implications for nanotechnological applications in biosensing and drug/gene delivery. **Amino Acids, Peptides, and Proteins**, 2004. PubMed Central review. Disponível em: <<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8622868/>>.
- GALUCIO, J. M.; MONTEIRO, E. F. et al. In silico identification of natural products with anticancer activity using a chemo-structural database of brazilian biodiversity. **Computational Biology and Chemistry**, v. 83, p. 107102, 2019.
- GALZITSKAYA, O. V. Enhancing the antimicrobial properties of peptides through cell-penetrating peptide conjugation: A comprehensive assessment. **International Journal of Molecular Sciences**, v. 24, n. 23, p. 16723, 2023.
- GAUTAM, A.; CHAUDHARY, K. et al. In silico approaches for designing highly effective cell penetrating peptides. **Journal of Translational Medicine**, v. 11, p. 74, 2013.
- GAYNES, R. The discovery of penicillin—new insights after more than 75 years of clinical use. **Emerging Infectious Diseases**, v. 23, n. 5, p. 849–853, 2017.

GULSEREN, P.; UFUK, G. et al. Synthesis and cytotoxicity of a tat peptide-doxorubicin conjugate for breast cancer treatment. **Journal of Drug Design and Research**, v. 2, n. 1, p. 1007, 2015.

HAO, M.; ZHANG, L. et al. Membrane internalization mechanisms and design strategies of arginine-rich cell-penetrating peptides. **International Journal of Molecular Sciences**, v. 23, n. 16, p. 9038, 2022. ISSN 1422-0067.

HART, A. Mann-whitney test is not just a test of medians: differences in spread can be important. **BMJ**, v. 323, n. 7309, p. 391–393, 2001.

HOLTON, T. A.; POLLASTRI, G. et al. Cpppred: prediction of cell-penetrating peptides. **Bioinformatics**, v. 29, n. 23, p. 3094–3096, 2013.

IMB, University of Queensland. **Explainer: Peptides vs proteins – what’s the difference?** 2017. <<https://imb.uq.edu.au/article/2017/11/explainer-peptides-vs-proteins-whats-difference>>.

KALAFATOVIC, D.; GIRALT, E. Cell-penetrating peptides: Design strategies beyond primary structure and amphipathicity. **Molecules**, v. 22, n. 11, p. 1929, 2017. ISSN 1420-3049.

KAUFFMAN, W. B.; FUSELIER, T.; OTHERS. Mechanism matters: A taxonomy of cell penetrating peptides. **Trends in Biochemical Sciences**, v. 40, n. 12, p. 749–764, 2015.

KEXIN, S.; YUANPENG, X. et al. Practicpp: a deep learning approach tailored for extremely imbalanced datasets in cell-penetrating peptide prediction. **Bioinformatics**, v. 40, n. 2, p. btae058, 2024.

KIM, G. C.; CHOI, S. H. et al. Photoswitching of cell penetration of amphipathic peptides by control of alpha-helical conformation. **Biomacromolecules**, v. 19, n. 7, p. 2863–2869, 2018. ISSN 1525-7797.

LESORT, T.; DIAZ-RODRIGUEZ, N. et al. State representation learning for control: An overview. **Neural Networks**, v. 108, p. 379–392, 2018. ISSN 0893-6080.

LI, F.; GAO, X. Predicting gram-positive bacterial protein subcellular location by using combined features. **BioMed Research International**, v. 2020, p. 1–8, 2020. ISSN 2314-6133.

LI, S.; WU, Y. et al. Antibiotics degradation by advanced oxidation process (aops): Recent advances in ecotoxicity and antibiotic-resistance genes induction of degradation products. **Chemosphere**, v. 311, p. 136977, 2023.

LIU, W.; XU, Y.; SLAVEYKOVA, V. I. Oxidative stress induced by sub-lethal exposure to copper as a mediator in development of bacterial resistance to antibiotics. **Science of The Total Environment**, v. 860, p. 160516, 2023.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: **Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)**. [S.l.: s.n.], 2017. p. 4768–4777.

- MANAVALAN, B.; PATRA, M. Mlcpp 2.0: An updated cell-penetrating peptides and their uptake efficiency predictor. **Journal of Molecular Biology**, v. 434, n. 11, 2022.
- MULLARD, A. The case for cell-penetrating peptides. **Nature Reviews Drug Discovery**, v. 22, p. 615–616, 2023.
- Nature Editorial. Progress on antibiotic resistance. **Nature**, v. 562, p. 307, 2018.
- OLIVEIRA, E. C. L. de; SANTANA, M. et al. Predicting cell-penetrating peptides using machine learning algorithms and navigating in their chemical space. **Scientific Reports**, v. 11, p. 7628, 2021.
- PARK, S. E.; SAJID, M. I. et al. Cyclic cell-penetrating peptides as efficient intracellular drug delivery vectors. **Molecular Pharmaceutics**, v. 16, p. 1234–1245, 2019.
- RAMASUNDARAM, M.; SOHN, H. et al. A bird's-eye view of the biological mechanism and machine learning prediction approaches for cell-penetrating peptides. **Frontiers in Artificial Intelligence**, Volume 7 - 2024, 2025. ISSN 2624-8212.
- RAMEZAN, C. A.; WARNER, T. A.; MAXWELL, A. E. Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. **Remote Sensing**, v. 11, n. 2, p. 185, 2019. ISSN 2072-4292.
- RAMSEY, J.; FLYNN, N. Cell-penetrating peptides transport therapeutics into cells. **Pharmacology & Therapeutics**, v. 154, p. 78–86, 2015.
- RIVES, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. **Proceedings of the National Academy of Sciences**, v. 118, n. 15, p. e2016239118, 2021.
- RODRIGUEZ-PÉREZ, R. et al. Prediction of compound profiling matrices using machine learning. **ACS Omega**, v. 3, p. 4713–4723, 2018.
- SARAVANAN, V.; GAUTHAM, N. Harnessing computational biology for exact linear b-cell epitope prediction: A novel amino acid composition-based feature descriptor. **OMICS: A Journal of Integrative Biology**, v. 19, n. 10, p. 648–658, 2015.
- SCHMIDT, N.; MISHRA, A. et al. Arginine-rich cell-penetrating peptides. **FEBS Letters**, v. 584, n. 9, p. 1806–1813, 2009. ISSN 0014-5793.
- SCIENTIFIC, T. F. **Amino Acid Physical Properties**. 2025. <<https://www.thermo-fisher.com/br/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/amino-acid-physical-properties.html>>. Acesso em: 6 jul. 2025.
- SU, M. et al. Incorporating amino acids composition and functional domains for identifying bacterial toxin proteins. **BioMed Research International**, v. 2014, p. 1–7, 2014. ISSN 2314-6133.
- TAI, K.; DHALIWAL, J.; BALASUBRAMANIAM, V. Leveraging mann–whitney u test on large-scale genetic variation data for analysing malaria genetic markers. **Malar J**, v. 21, n. 79, 2022.
- WEI, L.; TANG, J. et al. Skipcpp-pred: an improved and promising sequence-based predictor for predicting cell-penetrating peptides. **BMC Genomics**, v. 18, p. 742, 2017.

XIE, J.; BI, Y. et al. Cell-penetrating peptides in diagnosis and treatment of human diseases: From preclinical research to clinical application. **Frontiers in Pharmacology**, v. 11, p. 697, 2020.

YUAN, S.; GAO, D. et al. Ibpred: A sequence-based predictor for identifying ion binding protein in phage. **Comput Struct Biotechnol J**, v. 20, p. 4942–4951, 2022.

ZHANG; WEI et al. Siamesecpp: Prediction of cell-penetrating peptides using siamese networks. **Briefings in Bioinformatics**, v. 23, p. bbac242, 2022.

ZOUBOULIS, C. Amino acids. In: ZHANG, N. K. F. (Ed.). **Skin Aging Handbook**. [S.l.]: William Andrew Publishing, 2009. p. 99–118.