Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 12, n. 1, 2026.

Trabalho apresentado no XLIV CNMAC, Fundação Getulio Vargas - Rio de Janeiro - RJ, 2025.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Integração de Parâmetros Físico-Químicos e *Machine Learning* para a Identificação de Epítopos

Reginaldo J. Silva[1], Andréia S. Santos [2], Mara L. M. Lopes[3]
Departamento de Engenharia Elétrica/UNESP, Ilha Solteira, SP
André L. C. Costa[4]
Departamento de Microbiologia e Imunologia/UNIFAL, Alfenas, MG
Angela L. Moreno[5]
Departamento de Matemática/UNIFAL, Alfenas, MG

Resumo. Este estudo propõe um método para classificar epítopos em três categorias: MHC classe I, MHC classe II e epítopos de células B, utilizando dados do *Immune Epitope Database* (IEDB) e técnicas de aprendizado de máquina. A classificação de epítopos é importante para o desenvolvimento de vacinas e terapias imunológicas, pois permite identificar alvos específicos e prever respostas imunes. Parâmetros físico-químicos foram usados para representar as sequências de aminoácidos. O modelo *Random Forest* (RF) obteve os melhores resultados, com acurácia de 84,32% e F_1-*score* macro de 0,8276, destacando-se na classificação de MHC I (F_1-*score* de 0,9557).

Palavras-chave. Sistema Imune, Imunoterapia, Antígenos, Random Forest, Classificação

1 Introdução

O sistema imune humano é composto por dois eixos complementares: a imunidade inata, que atua como primeira linha de defesa, mediada por barreiras físicas (como pele e mucosas), células fagocíticas (macrófagos, neutrófilos) e moléculas pró-inflamatórias de resposta rápida a patógenos; e a imunidade adaptativa, caracterizada por especificidade e memória imunológica, sustentada por linfócitos B e T. Esta última depende do reconhecimento preciso de epítopos - regiões moleculares críticas derivadas de antígenos (proteínas, lipídios ou carboidratos) capazes de interagir com receptores de linfócitos (BCR/TCR) ou anticorpos [3].

Os epítopos B, que podem ser lineares (sequências contínuas de aminoácidos) ou conformacionais (dependentes da estrutura tridimensional do antígeno), permitem a neutralização direta de patógenos por anticorpos. Já os epítopos T, processados e apresentados por moléculas do Complexo Principal de Histocompatibilidade (MHC), direcionam respostas celulares: epítopos associados a MHC I (expressos em todas as células nucleadas) são reconhecidos por linfócitos T CD8+ citotóxicos, induzindo apoptose de células infectadas ou neoplásicas, enquanto epítopos ligados a MHC II (expressos por células apresentadoras de antígenos, como dendríticas) são reconhecidos por linfócitos T CD4+ auxiliares, modulando a ativação de outras células imunes [13, 15].

Aplicações translacionais exploram a especificidade de epítopos para desenvolver vacinas de próxima geração e terapias personalizadas. Vacinas de mRNA contra SARS-CoV-2, por exemplo,

[1]reginaldo.silva@unesp.br
[2]andreia.faria@unesp.br
[3]mara.lopes@unesp.br
[4]andre.costa@sou.unifal-mg.edu.br
[5]angela.moreno@unifal-mg.edu.br

codificam a proteína spike, cujos epítopos B (domínio RBD) e T (MHC I/II) induzem neutralização viral e imunidade celular duradoura [11]. Em oncologia, algoritmos preditivos identificam neoantígenos tumorais para vacinas personalizadas, enquanto células CAR-T direcionadas a epítopos de superfície (ex.: CD19) demonstram eficácia em neoplasias hematológicas [6, 9]. Em doenças autoimunes, estratégias de tolerância imunológica envolvem modificação de epítopos autorreativos ou expansão de células T reguladoras (Tregs), visando suprimir respostas patogênicas [10, 16].

Avanços tecnológicos, como predição de epítopos via inteligência artificial, estão revolucionando a imunoterapia. Algoritmos de *machine learning* utilizam bibliotecas proteômicas para identificar epítopos imunogênicos, otimizando o desenvolvimento de intervenções precisas [12]. Trabalhos recentes, como os de Vita et al. [15] na classificação de epítopos de MHC-I, Jensen et al. [4] em MHC-II e Jespersen et al. [5] em epítopos de células B, demonstram como métodos computacionais refinam a seleção de alvos terapêuticos. Desta forma, a integração entre imunologia básica e inovação translacional posiciona os epítopos como eixos estratégicos para enfrentar desafios globais em saúde, desde pandemias até resistência terapêutica em oncologia.

Neste contexto, este artigo tem como objetivo principal propor uma abordagem para a classificação de epítopos em três categorias fundamentais: MHC classe I, MHC classe II e epítopos de células B, utilizando dados extraídos do *Immune Epitope Database* (IEDB) e algoritmos de aprendizado de máquina. A classificação adequada desses epítopos é essencial para o desenvolvimento de vacinas, imunoterapias e diagnósticos imunológicos, uma vez que permite a identificação de alvos imunogênicos específicos e a previsão de respostas imunes adaptativas.

2 Metodologia

Os dados utilizados neste estudo foram construídos utilizando os conjuntos de dados de epítopos disponíveis no *Immune Epitope Database* (IEDB) [15], uma das principais fontes de dados imunológicos, que contém informações sobre epítopos de células B, MHC I/II, além de ligantes de células T. Após a obtenção dos dados no IEDB foi realizada uma seleção para garantir a qualidade e relevância dos dados.

Primeiramente, as sequências de aminoácidos que apareciam tanto nos conjuntos de dados de MHC quanto nos de células T foram selecionadas, gerando um conjunto único de MHC. Esse passo é importante pois, as células T reconhecem antígenos apenas quando esses antígenos são processados e apresentados por moléculas do MHC. Essa relação é fundamental para a ativação das células T e, consequentemente, para a resposta imune adaptativa. Deste modo, ao selecionar apenas as sequências comuns aos dois conjuntos de dados, garantem-se epítopos biologicamente relevantes, ou seja, que sejam capazes de interagir tanto com o MHC quanto com os receptores das células T (TCRs).

Em seguida, foram selecionadas apenas as sequências únicas que apareciam tanto no conjunto de MHC quanto no conjunto de células B. Para evitar um desbalanceamento excessivo dos dados, apenas 15.000 sequências de células B foram incluídas no conjunto final. Resultando em 43.975 amostras, sendo 17.933 de MHC classe I, 11.042 de MHC classe II e 15.000 de células B.

A seleção de sequências únicas é importante porque elimina redundâncias nos dados. Essa redundância pode ocorrer devido a diferentes fatores, como contextos experimentais variados ou anotações repetidas. Por exemplo, uma mesma sequência de aminoácidos pode ser reportada múltiplas vezes em estudos distintos, com pequenas variações nas condições experimentais ou nos métodos de detecção. Essas duplicações podem distorcer a representatividade dos dados e afetar a qualidade e confiabilidade do conjunto final.

Com o conjunto de dados construído, as sequências de aminoácidos foram convertidas em valores numéricos, utilizando uma variedade de parâmetros físico-químicos e estruturais. Esses parâmetros foram escolhidos com base em sua relevância para a interação entre epítopos e o sistema imunológico, sendo amplamente utilizados na literatura para caracterizar propriedades biológicas

e funcionais de peptídeos. Neste trabalho, combinamos abordagens inspiradas em dois estudos anteriores: Kozlova et al. [8], que destaca a importância de parâmetros como hidrofobicidade e carga para a classificação de epítopos, e Tayebi, Ali e Patterson [14], que utiliza métricas como entropia de Shannon e similaridade de motivos para análise de sequências imunológicas.

No total, 33 parâmetros foram utilizados, incluindo: índice alifático, grau de hidropaticidade (índice GRAVY), ponto isoelétrico, hidrofobicidade média, percentual de aminoácidos carregados positivamente, percentual de aminoácidos carregados negativamente, percentual de aminoácidos sem carga, comprimento da sequência, distribuição de carga, similaridade de motivos, entropia de Shannon, índice de Simpson e a frequência de aminoácidos específicos (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y). Como o comprimento das sequências é variável, alguns parâmetros como hidrofobicidade, distribuição de carga, os percentuais de aminoácidos carregados positivamente, negativamente e sem carga foram fornecidos diretamente em forma percentual, garantindo que fossem comparáveis entre sequências de diferentes tamanhos. Para os demais atributos, que não estavam em forma percentual, foi aplicada a normalização min-max, conforme a equação (1).

$$x_{\text{norm}} = \frac{x - \min(\boldsymbol{x})}{\max(\boldsymbol{x}) - \min(\boldsymbol{x})}. \tag{1}$$

Essa normalização foi realizada para garantir que todos os parâmetros estivessem na mesma escala, evitando que variáveis com magnitudes maiores afetassem os algoritmos. Após a normalização, o conjunto de dados foi dividido em conjuntos de treinamento e teste utilizando a técnica de *hold-out* na proporção 70/30, com a divisão sendo estratificada para preservar a distribuição das classes em ambos os conjuntos. O conjunto de treinamento (70% dos dados) foi utilizado para a busca e seleção dos melhores parâmetros para os algoritmos de classificação por meio da busca em grade combinada com validação cruzada de 5-*folds*.

A busca em grade é uma técnica exaustiva que testa todas as combinações possíveis de hiperparâmetros pré-definidos para cada algoritmo, com o objetivo de encontrar a configuração que maximiza o desempenho do modelo [1]. Já a validação cruzada (5-*folds*) divide o conjunto de treinamento em 5 partes, utilizando 4 partes para treinamento e 1 para validação, em um processo iterativo que garante que cada *fold* seja usado como conjunto de validação uma vez. Essa técnica é essencial para evitar *overfitting*, proporcionando uma estimativa mais robusta do desempenho do modelo em dados não vistos [2, 7].

Os algoritmos de aprendizado de máquina selecionados para o experimento foram: O *Random Forest* (RF), baseado em *ensemble learning*, emprega múltiplas árvores de decisão treinadas em subconjuntos aleatórios dos dados, promovendo robustez e reduzindo *overfitting* por meio da agregação de predições. O *K-Nearest Neighbors* (KNN) opera sob o princípio de similaridade, classificando instâncias com base na proximidade no espaço de características. O *Gradient Boosting* (GB) utiliza uma estratégia iterativa de otimização, em que modelos subsequentes são ajustados para minimizar os resíduos dos anteriores, resultando em alta capacidade preditiva. O *Support Vector Machine* (SVM) busca um hiperplano ótimo em espaços de alta dimensão para separar os dados. Por fim, o *Multilayer Perceptron* (MLP), uma arquitetura de rede neural artificial, utiliza múltiplas camadas de neurônios interconectados com funções de ativação não lineares possibilitando a captura de padrões complexos [2]. Para a avaliação dos modelos, foram utilizadas métricas como acurácia (ACC), precisão (Pre), sensibilidade (Se), especificidade (Sp), F_1-*score* (F_1) e área sob a curva ROC (AUC). Considerando que se trata de um problema de classificação multiclasse, na qual há mais de duas categorias a serem previstas, foram calculadas as versões macro e micro dessas métricas para garantir uma avaliação mais abrangente do desempenho dos modelos. A abordagem macro calcula a métrica para cada classe individualmente e tira a média simples dos resultados. Já a abordagem micro agrega os acertos e erros de todas as classes, calculando a métrica como se fosse um problema único.

3 Resultados

A Tabela 1 apresenta as métricas tanto macro quanto micro obtidas por cada método. Como os dados são desbalanceados, a métrica de interesse principal é a métrica macro, que trata todas as classes de forma igualitária, garantindo que classes minoritárias não sejam negligenciadas na avaliação. O RF destacou-se como o modelo com melhor desempenho geral, com uma acurácia de 84,32% e um F_1-*score* macro de 0,8276. Além disso, o RF obteve a maior AUC macro de 0,9477, revelando uma excelente capacidade de distinguir entre as classes. O GB e o MLP apresentaram resultados próximos, com acurácias de 83,82% e 83,43% respectivamente, e F_1-*scores* macro de 0,8216 e 0,8180. Esses modelos também demonstraram alta especificidade macro de 92,19% para o GB e 92,09% para o MLP, sugerindo uma boa capacidade de evitar os falsos positivos.

Por outro lado, o KNN e o SVM tiveram desempenhos inferiores com acurácias de 76,93% e 66,60%, respectivamente. A AUC macro do SVM de 0,8345 foi significativamente menor que a dos outros modelos, indicando dificuldades na separação das classes.

Tabela 1: Métricas de Desempenho dos Modelos.

Modelo	ACC	Se		Sp		Pre		F_1-*score*		AUC	
		Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro
RF	0,8432	0,8317	0,8432	0,9247	0,9216	0,8268	0,8432	0,8276	0,8432	0,9477	0,9585
KNN	0,7693	0,7461	0,7693	0,8852	0,8847	0,7469	0,7693	0,7460	0,7693	0,9078	0,9208
SVM	0,6660	0,6502	0,6660	0,8400	0,8330	0,6487	0,6660	0,6442	0,6660	0,8345	0,7434
GB	0,8382	0,8241	0,8382	0,9219	0,9191	0,8204	0,8382	0,8216	0,8382	0,9448	0,9562
MLP	0,8343	0,8232	0,8343	0,9209	0,9172	0,8180	0,8343	0,8180	0,8343	0,9409	0,9541

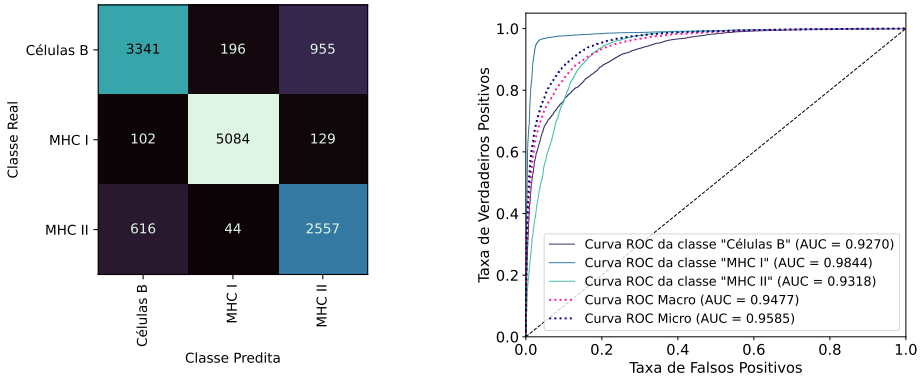
A Tabela 2 por sua vez, apresenta métricas como sensibilidade, especificidade, precisão e F_1-*score* para as categorias de epítopos: Células B (C1), MHC I (C2) e MHC II (C3). O RF novamente se destacou, especialmente na classificação de MHC I, com sensibilidade de 95,65%, especificidade de 96,89%, precisão de 95,49% e F_1-*score* de 0,9557. Esses valores indicam que o RF é altamente eficaz na identificação de MHC I. Para MHC II, o RF também apresentou o melhor desempenho entre os modelos, com F_1-*score* de 0,7457, embora essa métrica seja inferior à obtida para MHC I, indicando maior complexidade nessa classe.

Tabela 2: Desempenho por Classe de Cada Modelo.

Modelo	Classe	Se	Sp	Pre	F_1	Modelo	Classe	Se	Sp	Pre	F_1
	C1	0,7438	0,9158	0,8231	0,7814		C1	0,7522	0,9024	0,8022	0,7764
RF	C2	0,9565	0,9689	0,9549	0,9557	GB	C2	0,9556	0,9685	0,9543	0,9550
	C3	0,7948	0,8895	0,7023	0,7457		C3	0,7644	0,8949	0,7046	0,7333
	C1	0,6699	0,8519	0,7042	0,6866		C1	0,7213	0,9163	0,8194	0,7672
KNN	C2	0,9370	0,9126	0,8808	0,9080	MLP	C2	0,9543	0,9699	0,9563	0,9553
	C3	0,6313	0,8913	0,6558	0,6433		C3	0,7939	0,8764	0,6782	0,7315
	C1	0,4704	0,8023	0,5561	0,5096		C1: Células B				
SVM	C2	0,8578	0,9503	0,9225	0,8890	Rótulos	C2: MHC I				
	C3	0,6223	0,7675	0,4675	0,5339		C3: MHC II				

O GB e o MLP tiveram desempenhos semelhantes ao RF, com F_1-*score*s de 0,9550 e 0,9553 para MHC I, respectivamente. No entanto, para MHC II, o GB obteve um F_1-*score* de 0,7333, enquanto o MLP alcançou 0,7315, ambos ligeiramente inferiores ao RF. Para Células B, o RF também se saiu melhor, com F_1-*score* de 0,7814, comparado a 0,7764 do GB e 0,7672 do MLP. Já o KNN e o

SVM apresentaram dificuldades, especialmente na classificação de MHC II, com F_1-*score* de 0,6433 e 0,5339, respectivamente. Para Células B, o SVM teve o pior desempenho (F_1-*score* de 0,5096), indicando baixa generalização. Na Figura 1 são apresentados a matriz de confusão e a curva ROC para o RF, onde pode-se observar o desempenho satisfatório na classificação da classe MHC I.



(a) Matriz de Confusão. (b) Curva ROC e AUC para cada classe.

Figura 1: Resultados do Classificador Random Forest. Fonte: dos autores.

4 Discussões

Esses resultados tem implicações importantes para o desenvolvimento de vacinas de nova geração. A capacidade de classificar epítopos com alta precisão e rapidez pode acelerar o processo de identificação de candidatos a vacinas, reduzindo tempo e custos. Comparado a métodos padrão disponíveis no IEDB, como o NetMHC e o BepiPred, que utilizam ferramentas separadas para classificação de epítopos, o uso de métodos de *machine learning*, como o RF, pode reduzir o acúmulo de erros associados a cada ferramenta individual. O NetMHC, por exemplo, utiliza redes neurais artificiais para prever a ligação de peptídeos ao MHC, enquanto o BepiPred emprega modelos de Markov para prever epítopos lineares de células B. Embora eficazes, essas ferramentas realizam apenas classificações independentes, enquanto a técnica aqui apresentada realiza a classificação do epítopo tanto em relação à peptídeos ao MHC quanto em relação à células B e, por isso, não possui erros cumulativos. É importante salientar que o classificador proposto analisa o epítopo, não a cadeia completa como seus predecessores.

Embora este trabalho não apresente uma ferramenta que examine profundamente a sequência em busca de epítopos e depois os classifique, é importante ressaltar que isso pode ser feito de maneira relativamente simples. Uma abordagem viável é o uso de janelas deslizantes, em que a sequência de proteínas é dividida em segmentos de tamanho fixo e cada segmento é avaliado como epítopo ou não. Após a identificação, os epítopos podem ser classificados em categorias (células B, MHC I ou MHC II) utilizando os métodos de aprendizado de máquina propostos neste trabalho como o RF, que demonstrou excelente desempenho na classificação de epítopos neste estudo.

A utilização de parâmetros físico-químicos foi fundamental para o sucesso deste trabalho. Esses parâmetros capturam informações importantes sobre as propriedades e estrutura das sequências de aminoácidos como hidrofobicidade, carga, estabilidade térmica e frequência de aminoácidos, que são críticas para a interação dos epítopos com o sistema imunológico. A inclusão desses parâmetros permitiu que os modelos aprendessem padrões complexos e não lineares nos dados, o que contribuiu para a alta precisão dos resultados. Além disso, a normalização desses parâmetros em porcentagem garantiu que sequências de diferentes comprimentos pudessem ser comparadas de forma equitativa, sendo essencial para a classificação de epítopos.

A interpretabilidade dos parâmetros físico-químicos é uma vantagem significativa deste trabalho. Diferentemente de métodos baseados em redes neurais profundas, que muitas vezes funcionam como "caixas pretas", os parâmetros físico-químicos são facilmente interpretáveis e podem fornecer informações valiosas sobre as características dos epítopos. Por exemplo, um alto índice alifático pode indicar maior estabilidade térmica, enquanto um alto grau de hidrofobicidade pode sugerir uma tendência a interagir com ambientes lipídicos. Essa interpretabilidade permite que pesquisadores e desenvolvedores de vacinas entendam melhor as razões por trás das classificações feitas pelos modelos, o que pode auxiliar no desenvolvimento de vacinas e terapias imunológicas.

A importância dos parâmetros físico-químicos é ainda mais evidente quando consideramos que cada sequência de aminoácidos pode ter um comprimento diferente, o que adiciona uma camada adicional de complexidade à classificação. Sequências mais longas podem conter mais informações, mas também podem introduzir ruído ou redundância, enquanto sequências mais curtas podem não capturar todas as características relevantes. A normalização dos parâmetros em porcentagem permite que sequências de diferentes comprimentos sejam comparadas de forma justa, garantindo que o modelo não seja enviesado por diferenças no comprimento das sequências.

5 Considerações Finais

Este trabalho propôs a classificação de epítopos em três categorias: MHC classe I, MHC classe II e epítopos de células B, utilizando dados extraídos do *Immune Epitope Database* (IEDB) e algoritmos de aprendizado de máquina. A classificação adequada desses epítopos é crucial para o desenvolvimento de vacinas, imunoterapias e diagnósticos imunológicos, pois permite a identificação de alvos imunogênicos específicos e a previsão de respostas imunes adaptativas.

Os resultados demonstraram que o modelo de *Random Forest* (RF) destacou-se como o mais eficaz, com uma acurácia de 84,32% e um F_1-*score* macro de 0,8276, superando outros algoritmos como *Gradient Boosting* (GB), *Multilayer Perceptron* (MLP), *K-Nearest Neighbors* (KNN) e *Support Vector Machine* (SVM). O RF mostrou-se particularmente eficiente na classificação de epítopos de MHC I, com um F_1-*score* de 0,9557, e também obteve bons resultados para MHC II e epítopos de células B, embora com menor desempenho em comparação ao MHC I. Esses resultados indicam que o RF é uma ferramenta eficiente para a classificação de epítopos, especialmente em cenários onde a precisão e a capacidade de generalização são essenciais.

A utilização de parâmetros físico-químicos e a normalização desses parâmetros corroboraram para os resultados, permitindo que os modelos aprendessem padrões complexos e não lineares nos dados, além de permitir que sequências de diferentes comprimentos pudessem ser comparadas de forma equitativa. Além disso, a abordagem proposta pode ser facilmente adaptada para incluir técnicas como janelas deslizantes, que permitem a identificação de epítopos em sequências de proteínas mais longas, seguida de sua classificação em categorias específicas. Futuros trabalhos podem expandir o método para identificar epítopos em sequências mais longas.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 e à Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG)

Referências

[1] J. Bergstra e Y. Bengio. "Random Search for Hyper-Parameter Optimization". Em: **Journal of Machine Learning Research** 13.10 (2012), pp. 281–305.

[2] C. M. Bishop. **Pattern Recognition and Machine Learning (Information Science and Statistics)**. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN: 0387310738.

[3] D. S. Chen e I. Mellman. "Elements of cancer immunity and the cancer–immune set point". Em: **Nature** 541.7637 (2017), pp. 321–330. DOI: 10.1038/nature21349.

[4] K. K. Jensen et al. "Improved methods for predicting peptide binding affinity to MHC class II molecules". Em: **Immunology** 154.3 (2018), pp. 394–406. DOI: 10.1111/imm.12889.

[5] M. C. Jespersen et al. "BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes". Em: **Nucleic Acids Research** 45.W1 (2017), pp. W24–W29. DOI: 10.1093/nar/gkx346.

[6] C. H. June et al. "CAR T cell immunotherapy for human cancer". Em: **Science** 359.6382 (2018), pp. 1361–1365. DOI: 10.1126/science.aar6711.

[7] R. Kohavi. "A study of cross-validation and bootstrap for accuracy estimation and model selection". Em: **Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143.

[8] E. E. G. Kozlova et al. "Classification epitopes in groups based on their protein family". Em: **BMC Bioinformatics** 16.19 (2015), pp. S7. DOI: 10.1186/1471-2105-16-S19-S7.

[9] P. A. Ott et al. "A Phase Ib Trial of Personalized Neoantigen Therapy Plus Anti-PD-1 in Patients with Advanced Melanoma, Non-small Cell Lung Cancer, or Bladder Cancer". Em: **Cell** 183.2 (2020), pp. 347–362. DOI: 10.1016/j.cell.2020.08.053.

[10] M. Rosenzwajg et al. "Immunological and clinical effects of low-dose interleukin-2 across 11 autoimmune diseases in a single, open clinical trial". Em: **Annals of the Rheumatic Diseases** 78.2 (2019), pp. 209–217. DOI: 10.1136/annrheumdis-2018-214229.

[11] U. Sahin et al. "COVID-19 vaccine BNT162b1 elicits human antibody and TH1 T cell responses". Em: **Nature** 586.7830 (2020), pp. 594–599. DOI: 10.1038/s41586-020-2814-7.

[12] J. L. Sanchez-Trincado, M. Gomez-Perosanz e P. A. Reche. "Fundamentals and Methods for T- and B-Cell Epitope Prediction". Em: **Journal of Immunology Research** 2017.1 (2017), pp. 2680160. DOI: 10.1155/2017/2680160.

[13] A. Sette e R. Rappuoli. "Reverse Vaccinology: Developing Vaccines in the Era of Genomics". Em: **Immunity** 33.4 (2010), pp. 530–541. DOI: 10.1016/j.immuni.2010.09.017.

[14] Z. Tayebi, S. Ali e M. Patterson. "TCellR2Vec: efficient feature selection for TCR sequences for cancer classification". Em: **PeerJ Computer Science** 10 (2024), pp. e2239. DOI: 10.7717/peerj-cs.2239.

[15] R. Vita et al. "The Immune Epitope Database (IEDB): 2018 update". Em: **Nucleic Acids Research** 47.D1 (2018), pp. D339–D343. DOI: 10.1093/nar/gky1006.

[16] D. C. Wraith. "Therapeutic peptide vaccines for treatment of autoimmune diseases". Em: **Immunology Letters** 122.2 (2009). Special Section: Vaccination Immunology: Prevention and Beyond, pp. 134–136. DOI: 10.1016/j.imlet.2008.11.013.

