Revista Eletrônica de Iniciação Científica em Computação, 2025, 23:1
ISSN: 3085-8461 • doi: 10.5753/reic.2025.6062

Research Paper

A Machine Learning-Guided Approach for a Multi-Epitope HIV Vaccine Design

Pedro Bento [Universidade Federal de Minas Gerais |*pedro.bento@dcc.ufmg.br*]
Yan Aquino [Universidade Federal de Minas Gerais |*yanaquino@dcc.ufmg.br*]
Arthur Buzelin [Universidade Federal de Minas Gerais |*arthurbuzelin@dcc.ufmg.br*]
Pedro B. Rigueira [Universidade Federal de Minas Gerais |*pedrorigueira@dcc.ufmg.br*]
André Gambogi [Universidade Federal de Minas Gerais |*andregambogi@gmail.com*]
Luisa G. Porfírio [Universidade Federal de Minas Gerais |*luisagontijo@gmail.com*]
Italo Doria [Universidade Federal de Minas Gerais |*italofariao@gmail.com*]
Sofia Anunciação [Universidade Federal de Minas Gerais |*softgodoi2002@gmail.com*]
Gabriel Mendes [Universidade Federal de Minas Gerais |*gabriel95edu@gmail.com*]
Raquel Minardi [Universidade Federal de Minas Gerais |*raquelcm@dcc.ufmg.br*]
Adriana Alves Paim [Universidade Federal de Minas Gerais |*dribio@gmail.com*]
Gisele L. Pappa [Universidade Federal de Minas Gerais |*glpappa@dcc.ufmg.br*]
Flavio da Fonseca [Universidade Federal de Minas Gerais |*dafonsecaflavio@gmail.com*]
Wagner Meira Jr. [Universidade Federal de Minas Gerais |*meira@dcc.ufmg.br*]

✉ *Department of Computer Science (DCC), Universidade Federal de Minas Gerais (UFMG), Av. Pres. Antônio Carlos 6627, Pampulha, Belo Horizonte, MG, 31270-901, Brazil.*

Abstract. Developing an effective HIV vaccine remains challenging due to the virus's variability and complex immune responses. We propose a novel multi-epitope vaccine design using machine learning and computational methods to identify conserved, immunodominant epitopes from diverse HIV variants. These epitopes, selected to elicit humoral and cellular responses – targeting CD4+ T cells, CD8+ cytotoxic T cells, and B cells – are incorporated into a chimeric protein delivered via a viral vector to enhance immunity. Our framework integrates epitope selection, in silico physicochemical predictions, 3D construction of the chimeric protein, illustrated in Figure 1, and in vitro analysis, contributing to the development of a broadly protective and durable HIV vaccine.

Keywords: HIV vaccine, multi-epitope design, machine learning, immunoinformatics, chimeric protein

Received: 24 may 2025 • **Accepted:** 28 june 2025 • **Published:** 11 july 2025

1 Introduction

The Human Immunodeficiency Virus (HIV) continues to pose a critical global health burden, with UNAIDS reporting approximately 39 million people living with the virus and 630,000 AIDS-related deaths in 2023. Characterized by its progressive weakening of the immune system, HIV heightens vulnerability to opportunistic infections and culminates in Acquired Immunodeficiency Syndrome (AIDS) if untreated. Despite decades of scientific and financial investment in vaccine development, no preventive or curative vaccine has yet been realized. While advancements in viral structure analysis, immune response mechanisms, and antiretroviral therapies have transformed disease management, prolonging lives and reducing transmission, the absence of a viable vaccine highlights persistent gaps in our immunization strategies. This underscores the urgent need for innovative research to address the unresolved challenges in HIV prevention and eradication.

The development of an effective HIV vaccine has proven challenging over decades of research, with clinical trials demonstrating limited success. The immune response in mammals comprises two primary components: the innate and adaptive responses. The innate response provides the first line of defense through nonspecific mechanisms that activate adaptive immune components, including antibodies (produced by B cells) and T cells. Despite progress, no consensus exists on the ideal immune strategy for an HIV vaccine. Some studies prioritize neutralizing antibodies, which bind and disable viral particles, yet HIV's extreme genetic variability, evidenced by 538 distinct variants documented in GenBank, enables rapid mutation of surface proteins, diminishing antibody efficacy. This adaptability has led many researchers to argue that a robust cellular immune response, particularly driven by cytotoxicT lymphocytes (CD8$^+$ T cells), may be crucial for vaccine efficacy, especially when complemented by CD4$^+$ T cell-mediated immune modulation.T lymphocytes (CD8$^+$ T cells), may be crucial for vaccine efficacy, especially when complemented by CD4$^+$ T cell-mediated immune modulation.

Recent advances in *in silico* vaccine design, particularly during the COVID-19 pandemic, have demonstrated the potential of computational methodologies in immunogen design. Central to this progress are epitopes – the precise molecular signatures that immune cells recognize as foreign threats. Modern vaccines increasingly combine these epitopes into chimeric proteins, artificial structures that mimic natural antigens while targeting multiple viral vulnerabilities. While this approach showed promise during COVID-19 (with candidates like SpinTec Hojo-Souza *et al.* [2024] reaching phase III trials), HIV remains a formidable challenge. Despite genomic databases tracking 538 viral variants and prediction tools like IEDB(Immune Epitope Database) mapping thousands of potential epitopes, no multi-epitope HIV vaccine has entered clinical testing.

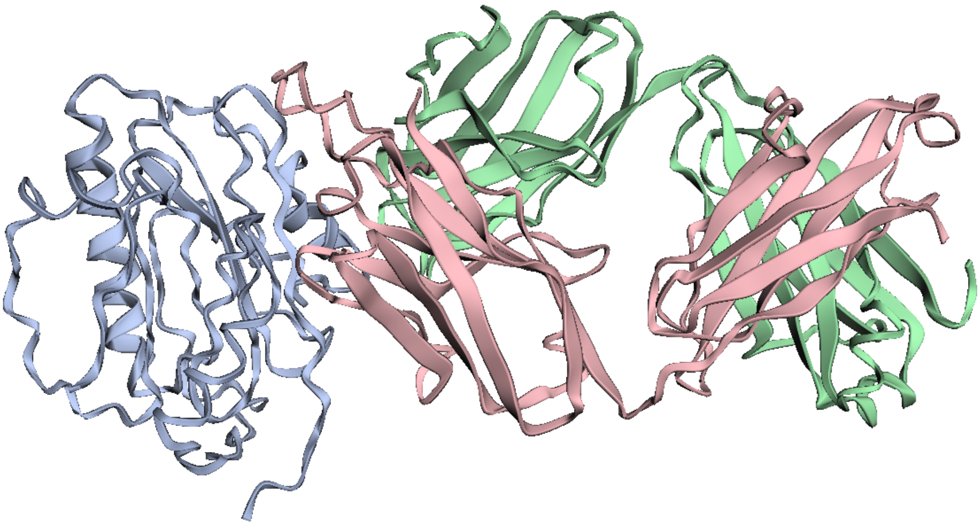

Figure 1. AlphaFold2-generated 3D structure and predicted docking interface of a representative chimera interacting with the antibody.

Our study addresses the challenges in HIV vaccine development by leveraging a machine learning framework to design optimized chimeric vaccines. Rather than relying on random epitope combinations, we systematically identify conserved targets from HIV's most stable proteins and use computational methods to assemble them into efficient configurations. This approach begins by choosing the most conserved epitopes recognized by B cells, CD4$^+$ T cells, and CD8$^+$ T cells. These epitopes are then arranged into stable chimeric proteins using predictive algorithms, with a focus on immune activation and manufacturability. After the initial optimization, a crucial safety and solubility filtering phase ensures that the final vaccine candidates are predictably safe and viable for production. The vaccine constructs are then delivered via MVA(Modified Vaccinia Ankara) viral vectors to enhance immune memory. By targeting multiple immune responses, our strategy aims to provide comprehensive protection against HIV, combining computational precision with biological insights to create a promising vaccine candidate ready for real-world validation. Figure 2 provides an overview of our proposed pipeline.

2 Related Work

The quest for an effective HIV vaccine has been marked by significant challenges, primarily due to the virus's high genetic variability and its ability to evade immune responses Ng'uni *et al.* [2020]; Cohen and Dolin [2013]. Traditional vaccine strategies, which often focus on inducing neutralizing antibodies, have struggled to provide durable protection against the diverse range of circulating HIV strains Gómez *et al.* [2012]. The RV144 trial, while demonstrating modest efficacy, highlighted the potential for achieving some level of protection through vaccination, but subsequent trials have failed to replicate these results. This underscores the necessity for novel approaches that can elicit both robust humoral and cellular immune responses Kaur and Vaccari [2024].

Multi-epitope vaccines have emerged as a promising strategy to address the challenges posed by HIV's genetic diversity Zhang [2018]. By incorporating conserved epitopes from multiple viral strains, these vaccines aim to induce broad and long-lasting protection Zhang [2018]. Computational methods, particularly machine learning and bioinformatics tools, play a crucial role in the design and optimization of multi-epitope vaccines Shen *et al.* [2021]. Tools like the Immune Epitope Database (IEDB) and VaccineDesigner facilitate the identification of potential epitopes and the assembly of multi-epitope constructs.

The integration of structural biology and advanced computational techniques has revolutionized the field of rational vaccine design. Recent innovative approaches, including mRNA technology, have shown promise in inducing broadly neutralizing antibodies Mu *et al.* [2021]. Instead of delivering antigenic proteins directly, mRNA-based vaccines rely on the host's cells to manufacture protein immunogens which serve as targets for antibody and cytotoxic T cell responses Mu *et al.* [2021]. However, the efficacy of these vaccines has been modest, with the best results achieving approximately 30% effectiveness .

Recombinant viral vectors, such as Modified Vaccinia Ankara (MVA), have demonstrated potential in delivering HIV antigens and inducing immune responses Gómez *et al.* [2012]. Additionally, poxvirus vectors have been explored as potential HIV/AIDS vaccines in human trials Gómez *et al.* [2012]. Despite these advances, achieving high efficacy remains a significant challenge, highlighting the ongoing need for comprehensive research and innovative strategies in HIV vaccine development.

3 Methodology

The development of a chimeric multi-epitope vaccine involves the integration of immunodominant epitopes capable of eliciting a robust and coordinated immune response. In our approach, we implemented a specific pipeline to first identify conserved genomic regions, then select those most likely to represent epitopes. Following this, we explored various configurations of these epitopes to optimize their arrangement for maximum efficacy. The optimization process took into account key factors such as solubility, toxicity, allergenicity, and the three-dimensional structure of the potential chimeras.

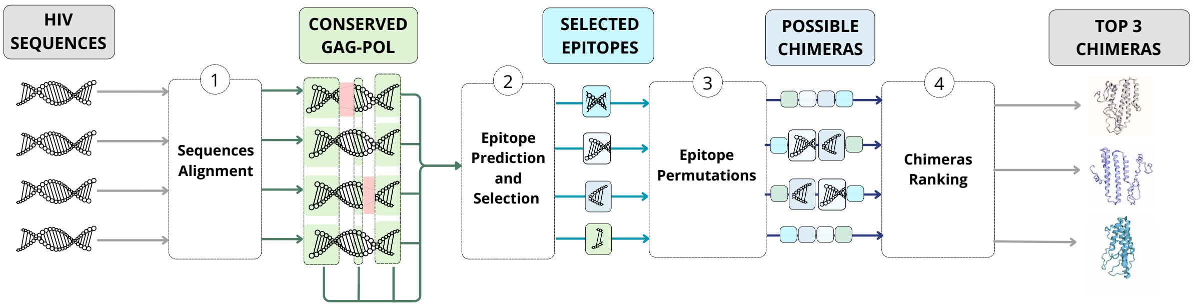

Figure 2. Pipeline for constructing HIV chimeras using conserved epitopes.

This methodology will be detailed in the subsequent subsections.

3.1 Conserved Sequences Identification

One of the greatest challenges in developing an effective HIV vaccine is the virus's high mutation rate across different variants. While certain genomic regions may elicit strong immunogenic responses, extensive variability among viral variants can significantly undermine a vaccine's ability to provide broad and long-lasting protection. This occurs because effective immune responses depend on recognizing previously encountered antigenic sequences.

Consequently, the initial phase of our research involved the analysis of all available HIV genomic sequences deposited in GenBank Benson *et al.* [2012]. The objective was to identify conserved genomic regions across HIV variants sequences that remain stable despite the virus's mutational capability. GenBank covers genomic sequences sampled from a diverse array of geographical locations and collected over extensive time periods, thus ensuring that the identified conserved regions reflect global viral diversity and evolutionary stability. Such an approach allows for the development of the most universal effective vaccine possible, potentially capable of inducing protective immunity to a wide range of geographically and genetically distinct virus types and ensuring sustained efficacy over the years.

To identify conserved genomic regions, all 538 available HIVs sequences were aligned using MAFFT Katoh and Standley [2013], a widely adopted bioinformatics tool known for its efficiency in sequence alignment and identification of conserved regions. MAFFT is particularly effective in processing large datasets, allowing precise detection of conserved genomic segments across multiple viral variants. Following sequence alignment, these conserved regions were identified, selected, and advanced to subsequent phases of the study.

3.2 Epitope Selection

Once the conserved sequences were identified, we now focused on selecting the immunogenic candidates, known as epitopes. This process was guided by two essential criteria: targeting stable proteins and activating the immune system on multiple fronts.

To begin, we chose conserved regions specifically within the *gag* and *pol* proteins. This decision was due to their critical roles in the viral life cycle across HIV variants, which allows them to remain stable despite the virus's mutational capacity. Moreover, *gag* and *pol* derived proteins are usually recognized by the immune system, making them prime candidates for vaccine development. In contrast, earlier vaccine approaches often targeted the *env* protein, which, although immunogenic, is highly variable and can be rapidly altered by the virus, enabling it to evade immune detection. Our strategy, therefore, focuses on more consistent targets, offering a better foundation for long-term protection against HIV.

Additionally, a broad immune response requires the activation of multiple cell types, each playing a critical role in defense and memory. B cells, through antibody production, are essential for neutralizing pathogens and preventing reinfection. CD8$^+$ T cells (cytotoxic T cells) directly target and kill infected cells. CD4$^+$ T cells (helper T cells) assist in orchestrating the immune response by activating both B cells and CD8$^+$ T cells, while also promoting the production of cytokines that enhance immune function. A vaccine capable of activating all three cell types ensures a robust immune defense, offering immediate protection and long-term immunity through memory formation. This coordinated response is crucial for an effective and lasting defense against HIV.

With these considerations in mind, the identification of potential epitopes was conducted using the Immune Epitope Database (IEDB) Vita *et al.* [2024], a widely recognized immunoinformatics resource. This tool enables the separate prediction of epitopes for B cells, CD4$^+$ T cells, and CD8$^+$ T cells. Following this analysis, five epitopes from each cell type were selected, prioritizing those with the highest predicted potential to provoke an immune response. The final 15 epitopes were incorporated into the chimeric protein formation phases. This strategy ensured a balanced activation of the three cell types while avoiding excessive molecular size, a critical factor in maintaining protein solubility.

3.3 Optimized Chimera Construction

In this stage, the 15 selected epitopes were used to form initial chimeric proteins by linking them via flexible peptide linkers, which provide structural mobility and aid in the 3D configuration of the molecules. The objective is to identify the optimal arrangement of the epitopes that enhances their physicochemical properties, including solubility, bioavailability, and minimal cross-reactivity with human antigens. Additionally, the ability of these chimeric proteins to bind antibodies and be presented by MHC-I (on cytotoxic T cells) and MHC-II (on helper T cells) molecules is thoroughly assessed.

To further optimize the protein's 3D structure and stability, the five epitopes from each cell type were grouped together. In this way, we improved the overall structural integrity of the molecule, ensuring that its functionality is preserved and that it can trigger an immune response.

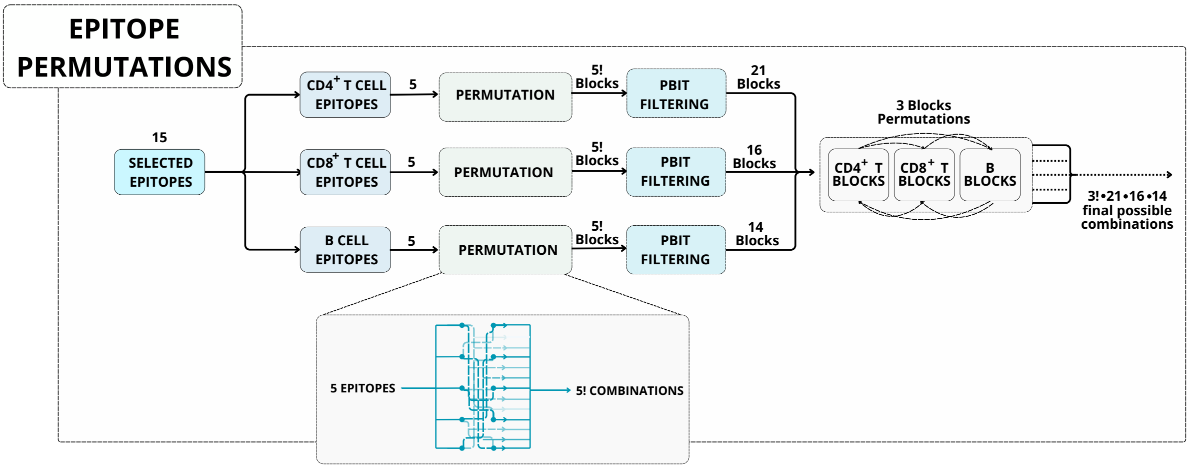

Figure 3. Overview of the optimized chimera construction.

However, this strategy introduced significant combinatorial complexity in determining the optimal arrangement of epitopes within the chimeric protein. With 15 epitopes divided into three blocks, each containing five epitopes specific to B cells, CD4$^+$ T cells, and CD8$^+$ T cells, the total number of possible configurations amounted to:

$$5!^3 \times 3! = 10{,}368{,}000$$

This immense search space rendered an exhaustive computational exploration impractical.

To manage this complexity efficiently, a greedy optimization approach was adopted. First, epitope combinations within each of the three blocks (5! per block) were evaluated individually using the predicted binding immunogenicity threshold (pBIT Chakraborty *et al.* [2024]) score, enabling the identification of the top-ranked configurations. Notably, multiple blocks achieved identical, highest pBIT scores, resulting in a tie for first place. Consequently, from each block, the optimal epitope sequences were selected: 21 for block 1 (B cells), 16 for block 2 (CD4$^+$ T cells), and 14 for block 3 (CD8$^+$ T cells).

After selecting the highest-ranked epitopes order per block, these subsequences were combined to form complete chimeric sequences, resulting in $21 \times 16 \times 14 \times 3! = 28,224$ arrangements. Each resulting chimeric sequence underwent another round of evaluation using the pBIT scoring system to determine the optimal chimeras, ensuring the best consistency of immunogenicity, structural stability, and antigenic properties. This approach reduced computational demands significantly while preserving the rigorous selection of an optimally immunogenic chimera. The entire process is illustrated in Figure 3.

3.4 Safety and Solubility Filtering

Following the selection and optimization of chimeric constructs based on immunogenicity, a crucial filtering phase was conducted to ensure the safety and viability of the designed vaccine candidates. This included assessments of toxicity, allergenicity, and solubility, three essential properties that significantly influence the suitability of a vaccine for further development.

Toxicity Prediction: To evaluate potential toxicity, we employed ToxinPred Gupta *et al.* [2013], a widely used tool for the prediction of toxic peptides. ToxinPred analyzes peptide sequences based on machine learning models trained on experimentally validated toxic and non-toxic peptides. Each chimeric candidate was screened through ToxinPred. After this step, sequences predicted to be toxic were filtered out and excluded from further analysis. This step was critical to minimize the risk of adverse effects in future in vivo applications.

Allergenicity Assessment: Subsequently, we assessed the allergenic potential of the chimeric proteins using AllerTOP v.2Dimitrov *et al.* [2014]. AllerTOP utilizes an alignment-independent method based on amino acid E-descriptors and auto- and cross-covariance transformation of protein sequences into uniform equal-length vectors. After classification, any sequence labeled as a probable allergen was removed from the candidate pool to ensure that the vaccine construct would not elicit harmful hypersensitivity reactions in the host.

Solubility Evaluation: Finally, we evaluated the solubility of the selected chimeric proteins using Protein-SolHebditch *et al.* [2017], a predictive tool that estimates the solubility of recombinant proteins when overexpressed in E. coli. Protein-Sol computes solubility scores based on sequence-derived features. Constructs predicted to be insoluble or below the solubility threshold were also discarded, as solubility is essential for effective expression, purification, and delivery of the vaccine candidate.

Together, these filtering steps ensured that only the most promising, non-toxic, non-allergenic, and soluble candidates advanced to the next stages of validation, reinforcing the safety and feasibility of the proposed chimeric multi-epitope vaccine design.

3.5 Final Candidate Selection and Structural Evaluation:

After the toxicity, allergenicity, and solubility filters, a total of 7,435 chimeric sequences remained. These candidates were re-evaluated using the pBIT scoring method to confirm their immunogenic potential. From this evaluation, 17 top-scoring sequences – tied in first place – were identified for further structural and functional analysis.

These top 17 chimeras were manually evaluated by specialists. The 3D structures of each candidate were predicted using AlphaFold2 Jumper *et al.* [2021], enabling a detailed visualization of their conformational features. Structural models were then analyzed for their potential to effectively interact with key immune components, including MHC class I, MHC class II molecules, and antibodies. This docking analysis was crucial to ensure that the epitopes were appropriately exposed and positioned for recognition by the immune system.

Based on this review and structural screening, three chimeras were selected as the most promising vaccine candidates. These constructs will undergo in vitro evaluation to validate their immunogenic potential.

3.6 Preclinical Evaluation

Once the optimal sequences are identified, the last phase involves *in vitro* testing. The chimeric construct is inserted into a plasmid and transfected into chicken embryo fibroblast (CEF) cells infected with MVA. This step ensures the successful integration of the chimeric gene into the viral vector, followed by the purification of the recombinant virus. The purified virus is then used to perform immunogenicity assessments in murine models, completing an iterative cycle for vaccine refinement.

4 Results

Our machine learning-guided framework led to the successful identification of conserved genomic regions within the gag and pol proteins, derived from the alignment of 538 HIV sequences. These regions formed the basis for robust epitope prediction using the Immune Epitope Database (IEDB), resulting in the selection of 15 high-potential epitopes – five each for B cells, CD4$^+$ T cells cells, and CD8$^+$ T cells.

To optimize their arrangement into chimeric constructs, we employed a greedy approach guided by the predicted binding immunogenicity threshold (pBIT) score. This reduced the initial combinatorial complexity (over 10 million configurations) to a focused subset of 28,224 candidate sequences.

We then applied three key filtering criteria to ensure safety and viability: toxicity (via ToxinPred), allergenicity (via AllerTOP v.2), and solubility (via Protein-Sol). After filtering, 23 chimeras remained that met all criteria.

These 23 sequences were reassessed using pBIT to identify the most immunogenic candidates. The top five were selected for detailed structural modeling using AlphaFold2. Structural assessments focused on epitope exposure and interactions with immune mediators such as MHC class I, MHC class II, and antibodies.

From this final evaluation, **three chimeras** demonstrated optimal structural features and immunogenic profiles. These were chosen as leading candidates for in vitro validation. The selected constructs will be integrated into the Modified Vaccinia Ankara (MVA) vector and tested in murine and CEF cell models to assess their capacity to induce both humoral and cellular immune responses.

5 Conclusion

The identification of three chimeric constructs with strong immunogenic profiles, structural stability, and favorable safety characteristics highlights the potential of multi-epitope strategies for addressing HIV's genetic diversity. The balanced targeting of B cells, CD4, and CD8 T cells offers a promising route toward comprehensive immune activation, which remains a central challenge in HIV vaccine development.

These results reinforce the value of integrating computational methods into early-stage vaccine design, not only to streamline candidate selection but also to improve the precision and efficacy of immune targeting. If validated experimentally, this approach could be extended to other rapidly evolving pathogens, contributing to more adaptable and responsive vaccine platforms in global health.

Acknowledgments

The Acknowledgments section is hidden for anonymity purposes.

References

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2012). Genbank. *Nucleic acids research*, 41(D1):D36–D42. DOI: 10.1093/nar/gks1195.

Chakraborty, S., Askari, M., Barai, R. S., and Idicula-Thomas, S. (2024). Pbitv3: A robust and comprehensive tool for screening pathogenic proteomes for drug targets and prioritizing vaccine candidates. *Protein Science*, 33(2):e4892. DOI: 10.1002/pro.4892.

Cohen, Y. and Dolin, R. (2013). Novel hiv vaccine strategies: overview and perspective. *Ther Adv Vaccines*, 1(3):99–112. DOI: 10.1177/2051013613494535.

Dimitrov, I., Bangov, I., Flower, D. R., and Doytchinova, I. (2014). Allertop v. 2—a server for in silico prediction of allergens. *Journal of molecular modeling*, 20:1–6. DOI: 10.1007/s00894-014-2278-5.

Gupta, S., Kapoor, P., Chaudhary, K., Gautam, A., Kumar, R., Consortium, O. S. D. D., and Raghava, G. P. (2013). In silico approach for predicting toxicity of peptides and proteins. *PloS one*, 8(9):e73957. DOI: 10.1371/journal.pone.0073957.

Gómez, C. E., Perdiguero, B., García-Arriaza, J., and Esteban, M. (2012). Poxvirus vectors as hiv/aids vaccines in humans. *Human Vaccines & Immunotherapeutics*, 8(9):1192–1207. DOI: 10.4161/hv.20778.

Hebditch, M., Carballo-Amador, M. A., Charonis, S., Curtis, R., and Warwicker, J. (2017). Protein–sol: a web tool for predicting protein solubility from sequence. *Bioinformatics*, 33(19):3098–3100. DOI: 10.1093/bioinformatics/btx345.

Hojo-Souza, N. S., de Castro, J. T., Rivelli, G. G., Azevedo, P. O., Oliveira, E. R., Faustino, L. P., Salazar, N., Bagno, F. F., Carvalho, A. F., Rattis, B., *et al.* (2024). Spin-tec: At cell-based recombinant vaccine that is safe, immunogenic, and shows high efficacy in experimental models challenged with sars-cov-2 variants of concern. *Vaccine*, 42(26):126394. DOI: 10.1016/j.vaccine.2024.126394.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., *et al.* (2021). Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589. DOI: 10.1038/s41586-021-03819-2.

Katoh, K. and Standley, D. M. (2013). Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780. DOI: 10.1093/molbev/mst010.

Kaur, A. and Vaccari, M. (2024). Exploring hiv vaccine progress in the pre-clinical and clinical setting: From history to future prospects. *Viruses*, 16(3):368. DOI: 10.3390/v16030368.

Mu, Z., Haynes, B. F., and Cain, D. W. (2021). Hiv mrna vaccines—progress and future paths. *Vaccines*, 9(2). DOI: 10.3390/vaccines9020134.

Ng'uni, T., Chasara, C., and Ndhlovu, Z. M. (2020). Major scientific hurdles in hiv vaccine development: Historical perspective and future directions. *Frontiers in Immunology*, 11:590780. DOI: 10.3389/fimmu.2020.590780.

Shen, J., Liu, F., Tu, Y., and Tang, C. (2021). Finding gene network topologies for given biological function with recurrent neural network. *Nat Commun*, 12(1):3125. DOI: 10.1038/s41467-021-23420-5.

Vita, R., Blazeska, N., Marrama, D., Members, I. C. T., Duesing, S., Bennett, J., Greenbaum, J., De Almeida Mendes, M., Mahita, J., Wheeler, D. K., Cantrell, J. R., Overton, J. A., Natale, D. A., Sette, A., and Peters, B. (2024). The immune epitope database (iedb): 2024 update. *Nucleic Acids Research*, 53(D1):D436–D443. DOI: 10.1093/nar/gkae1092.

Zhang, L. (2018). Multi-epitope vaccines: a promising strategy against tumors and viral infections. *Cell Mol Immunol*, 15(2):182–184. DOI: 10.1038/cmi.2017.92.