Applying generative language models to design new antibody sequences to target Influenza´s hemagglutinin

Carlos Gomes[1][0000-0003-4524-8109], Pedro Moreira[2][0000-0001-7734-1792], Diana Lousa [2][0000-0002-2309-0980] and Miguel Rocha [3][0000-0001-8439-8172]

[1] School of Engineering, University of Minho, Braga, Portugal
[2] António Xavier Institute of Chemical and Biological Technology, University Nova de Lisboa, Oeiras, Portugal
[3] Center of Biological Engineering, University of Minho, Braga, Portugal

Abstract. Influenza is a highly contagious viral disease that has a significant impact on the health of the world's population and has a very fast evolutionary capacity. So, in this project, using various bioinformatics tools such as SAbDab, AlphaFold, ProPythia, and many others, we intend to design new antibody sequences to target Influenza hemagglutinin.

Keywords: Influenza, Hemagglutinin, Generative Language Models, Bioinformatics.

1 Motivation and Objectives

Influenza is a highly contagious viral disease that infects millions of human beings every year. Its symptoms range from mild to severe and can include fever, muscle aches extreme fatigue and serious respiratory complications such as pneumonia. Although most people recover from the flu without serious complications, it still represents a significant public health burden, especially in high-risk groups [1]. Despite advances in the prevention and treatment of Influenza, including the availability of seasonal vaccines and antivirals, the Influenza virus continues to pose a significant challenge to public health due to its ability to mutate rapidly and evade existing immunity[2]. Therefore, it is essential to continue investing in research and development into new prevention and treatment strategies.

So, the main objective of this project is to obtain new antibody sequences to target the hemagglutinin of the Influenza virus.

2 State of the Art

2.1 Influenza

Epidemiology and Transmission. Influenza is a highly contagious viral infection that manifests itself in seasonal epidemics, mainly in winter [1]. The Influenza virus can affect all organs of the body and manifests itself as an acute febrile illness with varying degrees of systemic and respiratory symptoms. The main symptoms include fever, chills, headaches, weakness, redness of the eyes, sore throat, runny nose, and dry cough, and when complications from the infection are severe, they can be life-threatening for high-risk individuals or groups [3].

The main route of transmission is through the inhalation of infectious respiratory particles (large droplet transmission) when an infected person coughs or sneezes. There is also evidence of airborne transmission (small particles transmitted by speech or exhalation) and by fomites [4]. The typical incubation period is 24 to 48 hours. Patients are infectious one to two days before the onset of symptoms and for five to seven days afterwards. Children and immunosuppressed people may experience prolonged viral transmission [5].

For most outpatients, the diagnosis is made clinically, and laboratory confirmation is not necessary. Laboratory tests can be useful in hospitalized patients with suspected flu and in patients for whom a confirmed diagnosis will change treatment decisions. Rapid molecular assays are the preferred diagnostic tests because they can be carried out at the point of care, are highly accurate and have rapid results. Treatment with one of the four approved anti-Influenza drugs can be considered if the patient presents within 48 hours of the onset of symptoms. The benefit of treatment is greatest when antiviral therapy is started within 24 hours of the onset of symptoms. These drugs can reduce the risk of serious complications. There is also the possibility of annual vaccination as a form of flu prevention, and it is recommended for all people over six months of age who have no contraindications [6].

Etiology. Influenza viruses evolve quickly by frequent antigenic variation. Antigenic drift and shift are terms used to describe how the virus mutates and results in new strains. There is a significant change in the virus's genome in antigenic shift resulting in new hemagglutinin (HA) and neuraminidase (NA) protein expression [2]. This means that, despite improvements in prevention, control and case management, the antigenic shift continues to make Influenza a disease transmitted worldwide [7].

Influenza has two surface glycoproteins, Hemagglutinin and Neuraminase [8] .The virus infects the host by binding to the host cell and penetrating the membrane. As will be explored in the next subchapter, hemagglutinin binds to cell surface receptors and initiates the entry of the virus into these cells. Neuraminase is an enzyme that aids viral replication and allows the virus to be released from the host cell. So, viral glycoproteins play an essential role in the virulence and pathogenesis of the Influenza virus [7].

The role of hemagglutinin. The surface of Influenza virions is dominated by HA, which outnumbers NA by five to ten-fold on average (1) [9]. Hemagglutinin can agglutinate red blood cells, and this ability can be attributed to its receptor-binding function. HA binds to sialylated glycan receptors on host cells to initiate viral entry and carries the machinery for membrane fusion [10].

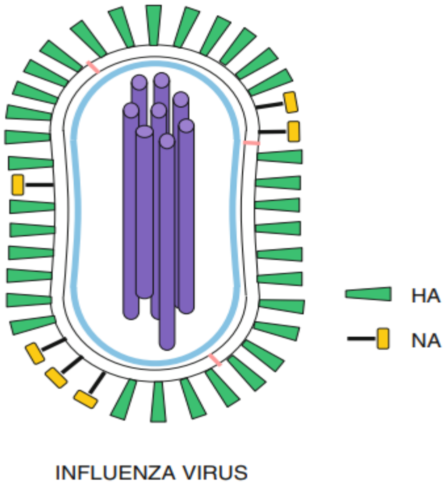

Fig. 1. Cartoon showing the architecture of Influenza virus. HA stands for Hemagglutinin while NA stands for Neuraminidase [11].

Hemagglutinins are found from the virus surface membrane as glycoprotein spikes, and each spike contains three identical subunits formed by two glycopolypeptides. Each of these subunits is divided into four subdomains (2) [12]. It has two main roles in Influenza infections: attaching the virus to the host receptors and promoting membrane fusion between the virus and host membranes. They recognize cell surface glycoconjugates containing sialic acid as receptors, but have limited affinity for them and, consequently, the binding of the virus to cells requires its interaction with several HAs of the virus. The receptor-bound virus is transferred to the endosomes where membrane fusion by HAs is activated at a pH between 5 and 6.5, depending on the virus strain. The fusion activity requires extensive rearrangements in the HA conformation, which include the extrusion of a buried "fusion peptide" to bind to the endosomal membrane, form a bridge to the virus membrane and finally bring the two membranes closer together [12]. Influenza viruses are classified based on their antigenicity, which is determined by their surface glycoproteins. There are four characterized types of flu virus, A, B, C and D. Influenza A and B viruses have two surface glycoproteins, while influenza C and D viruses have only one surface glycoprotein, hemagglutinin-esterase fusion. Both influenza A and B viruses infect humans and can cause severe illness or death. In contrast, Influenza C virus only causes mild symptoms in most cases. Human infection with Influenza D virus has not been observed. Therefore, most Influenza research has been focused on influenza A and B viruses. Between those two viruses, the main difference is that influenza B virus is only found in humans, whereas the primary natural reservoir for influenza A virus is aquatic birds. As a result, influenza A virus usually receives more attention and has been studied more extensively [13].

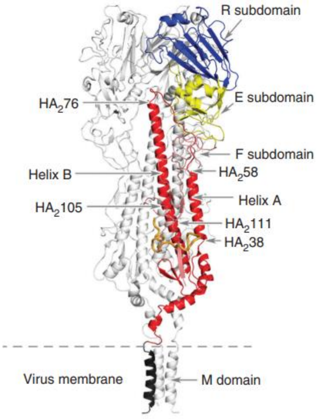

Fig. 2. The polypeptide and subdomain structure of hemagglutinin [12].

2.2 Antibodies and their role in defending against the virus

Antibodies are the basis of the immune response in vertebrates. These proteins form complexes with potentially pathogenic molecules called antigens and inhibit their function or recruit other components of the immune machinery to destroy them [14]. In addition to the biological importance of antibodies, their ability to be raised against an almost unlimited number of molecules has made them useful laboratory tools and increasingly useful as therapeutic agents in humans [15].

As mentioned above, in the case of Influenza, the HA and NA proteins are highly immunogenic, and antibodies directed at both glycoproteins can be isolated after natural infection or vaccination. By binding to the viral surface proteins HA and NA, antibodies can block essential steps in the virus replication cycle, thus limiting the spread of infection. Due to the host's immune pressure and the error-prone RNA polymerase, HA and NA are very plastic and show differences in antigenic properties. For instance, in the case of Influenza A, the virus can split into 18 HA subtypes and 11 NA subtypes [16]. Previously, the protective efficiency of antibodies was measured by their ability to prevent HA binding through the neutralization assay and the hemagglutination inhibition assay [17], and antibodies without these functions did not receive as much attention. However, increasing evidence suggests that non-neutralizing antibodies (nnAbs) can also confer protection through multiple mechanisms without interrupting virus entry or membrane fusion, such as complement activation, increased phagocytosis, targeting of internal viral proteins and activation of crystallizable fragment functions [18].

In addition to the biological importance of antibodies, their ability to be created against an almost unlimited number of molecules has made them useful laboratory tools and increasingly useful as therapeutic agents in humans. This biopharmaceutical application has increased the desire to understand how the antibody's binding, stability and immunogenic properties are determined and how they can be modified [14].
Due to their efficiency, analysis and computational tools are increasingly being used to aid the antibody engineering process. Currently, many of these tools use only antibody data, as opposed to general protein data, as this has been shown to increase performance [19].

2.3 Relevant software, packages, databases and models

In this project, different databases and different software will be used to fulfill the proposed objectives. Starting with the state of the art, we mainly used databases such as PubMed. This is an open-access site that has a wide range of articles on topics ranging from life sciences to bioengineering and is therefore important for searching for scientific evidence [20]. As for obtaining information on the antibodies that will be most useful for this work SAbDab was used. SAbDab is an online antibody structural database that brings together all publicly available antibody structures. The data is enriched with various properties, including experimental information, genetic details, antigen details, and where available the antibody-antigen binding affinity [14]. In addition, the classification of the structures of the complementarity determining regions (CDRs) of antibodies is very important for predicting the structure of antibodies and for their computational design. PyIgClassify is a database and webserver that gives access to CDR structures present in the Protein Data Bank (PDB) [21] and will therefore be important in this work. The Protein Data Bank is a global repository of experimentally determined 3D structures of biological macromolecules, experimental data, and the associated metadata [22].
We will also use protein language models to generate new antibody sequences. These models are deep learning models based on natural language processing methods. They are trained using large sets of protein sequences and find long-range dependencies in a protein sequence [23]. These pre-trained models can predict structure in an unsupervised way either taking as input a single sequence [24] or a multiple sequence alignment [25]. In this work we will be able to use two protein language modeling approaches, Evolutionary Scale Modeling (ESM) or Multiple Sequences Alignments-Transformer (MSA-Transformer).

The MSA Transformer is a protein language model that has been trained on MSAs using the masked language modeling objective, without additional supervised training, unlike ALphafold [26]. ESM, on the other hand, is a model that can effectively simulate the evolution of human antibodies, suggesting mutations that are evolutionarily plausible, despite not providing the model with any information about the target antigen, binding specificity or protein structure [27].

Other tools will be used to extract properties from the results obtained. ProPythia is a Python package that will be used to extract some of the properties. It offers a range of functions for calculating several physicochemical properties and other representations of proteins. It allows the user to pre-process the dataset, manage and modify sequences, clustering, manifold learning, feature selection and dimensionality reduction with a variety of diagrams to facilitate user interpretation. It also allows for the training and optimization of traditional Machine Learning models to make predictions on unseen data and respective feature analysis [28]. AlphaFold is a web service that will also be used in this work. It uses an approach based on deep-learning and a conventional neural network. This technique can predict the distance and torsion distribution of proteins and three-dimensional structure using training schemes of experimentally determined PDB relative structures [29].

3 Methodology

The aim of this project is to obtain new antibody sequences to target the hemagglutinin of the Influenza virus. To achieve this, the following work plan has been drawn up.

First, we will start with a literature review. At the start of the project, it is essential to do a comprehensive literature review. This involves an analysis of studies already done on the interaction of antibodies and Influenza hemagglutinin, as well as bioinformatics and molecular modeling techniques.

Then, we will search for Antibodies to Influenza Hemagglutinin in Databases. This step consists of exploring databases such as SabDab or PylgClassify to identify the best-known antibodies with affinity for the virus [14, 21].

Once the desired antibodies have been selected, it is essential to characterize the epitopes, in other words, the regions of Hemagglutinin to which these antibodies bind, and specify which variant(s) those antibodies are effective with.

Then, we are going to use Antibody Sequences to generate multiple sequence alignments (MSA), an essential technique in the bioinformatic analysis of protein sequences. With this approach, we sought to identify conserved and variable patterns in the antibody sequences, as well as areas of interaction with hemagglutinin. This information is crucial to understanding the diversity and evolution of antibodies in response to the virus.

The next activity consists of using different established language models such as ESM or MSA-transformer, to generate new sequences.

Then, we will analyze the results. Once these sequences have been obtained, they need to be categorized using various processes, such as multiple alignment and clustering, to select the most interesting results.

Finally, ProPythia will be used to extract the physical-chemical properties [28], while AlphaFold2 and/or IgFold will be used to predict their structural structure [29]. Analyzing these properties will allow us to identify the sequences that meet specific criteria so that they can be tested in the laboratory.

References

1. Bridges, C.B., Kuehnert, M.J., Hall, C.B.: Transmission of Influenza: Implications for Control in Health Care Settings. Clinical Infectious Diseases. 37, 1094–1101 (2003). https://doi.org/10.1086/378292/2/37-8-1094-FIG005.GIF.
2. Kim, H., Webster, R.G., Webby, R.J.: Influenza Virus: Dealing with a Drifting and Shifting Pathogen. Viral Immunol. 31, 174–183 (2018). https://doi.org/10.1089/VIM.2017.0141/ASSET/IMAGES/LARGE/FIGURE2.JPEG.
3. Clohisey, S., Baillie, J.K.: Host susceptibility to severe influenza A virus infection. Crit Care. 23, 1–10 (2019). https://doi.org/10.1186/S13054-019-2566-7/TABLES/1.
4. Xu, X., Blanton, L., Elal, A.I.A., Alabi, N., Barnes, J., Biggerstaff, M., Brammer, L., Budd, A.P., Burns, E., Cummings, C.N., Garg, S., Kondor, R., Gubareva, L., Kniss, K., Nyanseor, S., O'Halloran, A., Rolfes, M., Sessions, W., Dugan, V.G., Fry, A.M., Wentworth, D.E., Stevens, J., Jernigan, D.: Update: Influenza Activity in the United States During the 2018–19 Season and Composition of the 2019–20 Influenza Vaccine. MMWR Morb Mortal Wkly Rep. 68, 544–551 (2019). https://doi.org/10.15585/MMWR.MM6824A3.
5. Leekha, S., Zitterkopf, N.L., Espy, M.J., Smith, T.F., Thompson, R.L., Sampathkumar, P.: Duration of Influenza A Virus Shedding in Hospitalized Patients and Implications for Infection Control. Infect Control Hosp Epidemiol. 28, 1071–1076 (2007). https://doi.org/10.1086/520101.
6. Gaitonde, D.Y., Moore, F.C., Morgan, M.K.: Influenza: Diagnosis and Treatment. Am Fam Physician. 100, 751–758 (2019).
7. Javanian, M., Barary, M., Ghebrehewet, S., Koppolu, V., Vasigala, V.K.R., Ebrahimpour, S.: A brief review of influenza virus infection, (2021). https://doi.org/10.1002/jmv.26990.
8. Te Velthuis, A.J.W., Fodor, E.: Influenza virus RNA polymerase: insights into the mechanisms of viral RNA synthesis. Nature Reviews Microbiology 2016 14:8. 14, 479–493 (2016). https://doi.org/10.1038/nrmicro.2016.87.
9. Hutchinson, E.C., Charles, P.D., Hester, S.S., Thomas, B., Trudgian, D., Martínez-Alonso, M., Fodor, E.: Conserved and host-specific features of influenza virion architecture. Nature Communications 2014 5:1. 5, 1–11 (2014). https://doi.org/10.1038/ncomms5816.
10. Maeda, T., Ohnishi, S. ichi: Activation of influenza virus by acidic media causes hemolysis and fusion of erythrocytes. FEBS Lett. 122, 283–287 (1980). https://doi.org/10.1016/0014-5793(80)80457-1.
11. Luo, M.: Influenza virus entry. Adv Exp Med Biol. 726, 201–221 (2012). https://doi.org/10.1007/978-1-4614-0980-9_9.
12. Gamblin, S.J., Vachieri, S.G., Xiong, X., Zhang, J., Martin, S.R., Skehel, J.J.: Hemagglutinin Structure and Activities. Cold Spring Harb Perspect Med. 11, a038638 (2021). https://doi.org/10.1101/CSHPERSPECT.A038638.
13. Wu, N.C., Wilson, I.A.: Influenza Hemagglutinin Structures and Antibody Recognition. (2020). https://doi.org/10.1101/cshperspect.a038778.

14. Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., Shi, J., Deane, C.M.: SAbDab: the structural antibody database. Nucleic Acids Res. 42, D1140–D1146 (2014). https://doi.org/10.1093/NAR/GKT1043.
15. Walsh, G.: Biopharmaceutical benchmarks 2010. Nature Biotechnology 2010 28:9. 28, 917–924 (2010). https://doi.org/10.1038/nbt0910-917.
16. Gao, R., Sheng, Z., Sreenivasan, C.C., Wang, D., Li, F.: Influenza A Virus Antibodies with Antibody-Dependent Cellular Cytotoxicity Function. Viruses 2020, Vol. 12, Page 276. 12, 276 (2020). https://doi.org/10.3390/V12030276.
17. Von Holle, T.A., Anthony Moody, M.: Influenza and antibody-dependent cellular cytotoxicity. Front Immunol. 10, 457028 (2019). https://doi.org/10.3389/FIMMU.2019.01457/BIBTEX.
18. Padilla-Quirarte, H.O., Lopez-Guerrero, D. V., Gutierrez-Xicotencatl, L., Esquivel-Guadarrama, F.: Protective antibodies against influenza proteins. Front Immunol. 10, 461698 (2019). https://doi.org/10.3389/FIMMU.2019.01677/BIBTEX.
19. Choi, Y., Deane, C.M.: Predicting antibody complementarity determining region structures without classification. Mol Biosyst. 7, 3327–3334 (2011). https://doi.org/10.1039/C1MB05223C.
20. Hoogland, M.A.: How Medical Students Discover and Use Medical Information Tools. Med Ref Serv Q. 38, 347–357 (2019). https://doi.org/10.1080/02763869.2019.1661197.
21. Adolf-Bryfogle, J., Xu, Q., North, B., Lehmann, A., Dunbrack, R.L.: PyigClassify: A database of antibody CDR structural classifications. Nucleic Acids Res. 43, D432–D438 (2015). https://doi.org/10.1093/nar/gku1106.
22. Burley, S.K., Berman, H.M., Kleywegt, G.J., Markley, J.L., Nakamura, H., Velankar, S.: Protein Data Bank (PDB): The single global macromolecular structure archive. In: Methods in Molecular Biology. pp. 627–641. Humana Press Inc. (2017). https://doi.org/10.1007/978-1-4939-7000-1_26.
23. Sgarbossa, D., Lupo, U., Bitbol, A.F.: Generative power of a protein language model trained on multiple sequence alignments. Elife. 12, (2023). https://doi.org/10.7554/eLife.79854.
24. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., Fergus, R.: Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A. 118, e2016239118 (2021). https://doi.org/10.1073/PNAS.2016239118/SUPPL_FILE/PNAS.2016239118.SAPP.PDF.
25. Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., Rives, A.: Transformer protein language models are unsupervised structure learners. bioRxiv. 2020.12.15.422761 (2020). https://doi.org/10.1101/2020.12.15.422761.
26. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S.,

Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D.: Highly accurate protein structure prediction with AlphaFold. Nature 2021 596:7873. 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2.

27. Hie, B.L., Shanker, V.R., Xu, D., Bruun, T.U.J., Weidenbacher, P.A., Tang, S., Wu, W., Pak, J.E., Kim, P.S.: Efficient evolution of human antibodies from general protein language models. Nat Biotechnol. 42, 275–283 (2024). https://doi.org/10.1038/s41587-023-01763-2.

28. Sequeira, A.M., Lousa, D., Rocha, M.: ProPythia: A Python package for protein classification based on machine and deep learning. Neurocomputing. 484, 172–182 (2022). https://doi.org/10.1016/j.neucom.2021.07.102.

29. Gutnik, D., Evseev, P., Miroshnikov, K., Shneider, M.: Using AlphaFold Predictions in Viral Research. Current Issues in Molecular Biology 2023, Vol. 45, Pages 3705-3732. 45, 3705–3732 (2023). https://doi.org/10.3390/CIMB45040240.