The 7th International Conference on Interdisciplinary Research
on Computer Science, Psychology, and Education (2023)

Replicating Ford *et al.*'s investigation on the immune consequences and SARS-CoV-2 antibody's binding affinity using an *in silico* approach

Cameron Jones [1Δ], Hannah Zeru[2Δ], Denis Jacob Machado[2*]
[1]Department of Computer Science, Eastern Michigan University, MI, U.S.A.
[2]Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, NC, U.S.A.
[*]Correspondence: dmachado@charlotte.edu.

Abstract

Artificial intelligence (AI) offers innovative approaches for the estimation of protein structures and protein-protein interactions. Notably, scientists have successfully employed AI to swiftly estimate novel SARS-CoV-2 viral proteins and their binding affinities to neutralizing antibodies mere hours after the discovery of new variants. Regrettably, the process of estimating protein structures and their affinities with neutralizing antibodies has not yet been fully automated, necessitating the expertise of skilled practitioners for execution. In this endeavor, we embark on the automation of this process while enhancing its user-friendliness. The instrumental tools employed encompass cutting-edge protein structure prediction algorithms (specifically, AlphaFold2), precise protein binding estimators (HADDOCK), and sophisticated software for parsing, editing, and visualizing three-dimensional protein structures (ChimeraX). This endeavor will serve as a cornerstone in the development of a pipeline utilizing Bash and Python3. By doing so, we concurrently reduce processing time, thereby enabling more expeditious predictions. To showcase the efficacy of our newly automated pipeline, we conduct an in-depth investigation of select SARS-CoV-2 viral proteins (variants XBB.1.5, BJ.1, BM1.1.1, and B.1.1.529) and human-neutralizing antibodies (AZD1061, AZD8895, 58G6, C110, CV38-142, LY-CoV-1404, LY-CoV-555, P5C3, EY6A, and COVOX-150).

Keywords: Deep learning; Spike protein; Bioinformatics; Coronavirus; Pandemic.

1. Introduction

BACKGROUND–In late November 2022, the United States Centers for Diseases Control began tracking a new SARS-CoV-2 variant named XBB.1.5, which accounted for about 3% of infections at the time. By January 2023, XBB.1.5 had grown to represent 30% of all infections [1]. This variant is characterized by 40 mutations in the Spike protein (S), with 22 occurring in the receptor binding domain (RBD) [2]. It has been proposed that XBB.1.5 is a recombinant strain of the virus from BJ.1 and BM.1.1.1, but alternative explanations, such as convergent evolution, are also being considered [3], [4], [5], [6]. Ford et al. [7] investigated XBB.1.5 and related variants (B.1.1.529, BJ.1, and BM.1.1) using *in silico* modeling to predict Spike protein structures and their ability to escape neutralizing antibodies. By predicting these binding affinities, they could determine whether new variants would respond well to current treatments and vaccines.

AIMS–Our goal is to reproduce the study by Ford et al. [7], corroborate their findings, and produce a more granular documentation of their methodology to facilitate automation. To do that, we followed the original study with a few modifications. In the original study, the AI tools used were HADDOCK (version 2.2/2.4) [8], [9], PRODIGY [10], [11], AlphaFold2 (ColabFold-mmseqs2 version) [12], RoseTTAfold [13], [14], and PyMOL (version 1.8) [15]. In the current study, we used a different version of HADDOCK (version 2.4) [8, 9], incorporated ChimeraX (version 1.6.1) [16], and executed AlphaFold2 via two Google Colabs (ColabFold-mmseq2 and AlphaFold colab) [17, 12, 18]. The detailed protocols are in "Methods and Materials" below.

2. Methods and Materials

DATA–Selected nucleotide sequences of the Spike gene from different variants of SARS-CoV-2 were downloaded from public databases [19]. Protein structures validated by

[Δ]These authors share the first authorship and are listed alphabetically.
[*]Corresponding author, Email: dmachado@charlotte.edu.

direct observation (i.e., crystallography) [20] were downloaded from [21]. Antibody data was retrieved from [22], [23]. The input data is summarized in Table 1.

Table 1. This is the full list of neutralizing antibodies and SARS-CoV-2 variants used in this study.

Neutralizing antibodies	SARS-CoV-2 variants
LY-CoV555	XBB.1.5
LY-CoV1404	BM.1.1.1
P5C3	B.1.1.529
COVOX-150	BJ.1
AZD1061	
AZD8895	
C110	
EY6A	
58G6	
CV38-142	

Computational Workflow We modified the workflow from [7] as shown in Figure 1. Viral protein structures corresponding to the Spike gene (S) were predicted using AlphaFold2 v2.3.2 (using either AlphaFold Colab [17, 18] or ColabFold-mmseq2 [12]). Protein-to-protein docking analyses between viral targets and neutralizing antibodies were executed in HADDOCK v2.4 [8, 9]. ChimeraX v1.6.1 [16] was used to manipulate and visualize the results after we encountered technical difficulties executing PyMOL [15] (see "Challenges" below).

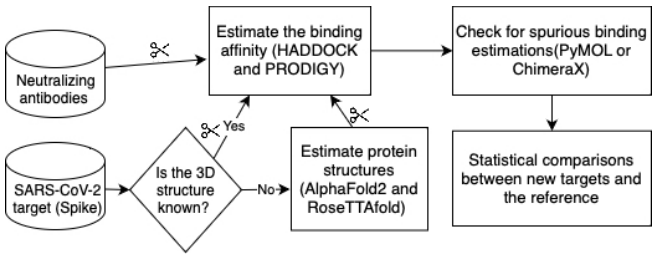

Figure 1. A flowchart summarizing our methodology. The scissors symbol indicates the editing or formatting of structural files in ChimeraX.

CHALLENGES–We encountered and overcame two major challenges in conducting this research.

1. Handling AlphaFold2 executing time in AlphaFold Colab: We learned that the parameters of ColabFold result in a faster and equally reliable prediction, which will inform the Phyloinformatics lab's future research. With AlphaFold Colab, we had an issue running a test SARS-CoV-2 variant. This variant was used in the comparison between the two programs to make sure we had a well-documented older variant, but AlphaFold Colab was unable to run and predict this variant.
2. Downloading and using PyMOL [15]: PyMOL proved difficult to handle by inexperienced users. This allowed us to find alternatives to PyMOL. We learned that ChimeraX has similar functionality and is more user-friendly than PyMOL. Future workflows designed by the Phyloinformatics lab will use ChimeraX instead of PyMOL.

3. Results

ALPHAFOLD–We tested ColabFold and AlphaFold's accuracy and speed. ColabFold was consistently faster than AlphaFold by an average of 60 to 90 minutes, with ColabFold only taking 3 to 7 minutes to produce results. When comparing the predictions for both programs, we found negligible differences between the resulting files. For the test variant, we found that while ColabFold took 87 minutes to fully run, AlphaFold Colab would take 420 minutes to run before crashing. AlphaFold Colab could not handle the test variant, so we could not collect results from both programs.

HADDOCK–Overall, differences between our results and [7] are negligible (for example, see Figure 2). Changes we made to the original pipeline gave us similar results compared to the original study.

The plot in 2 shows how similar our results are to the original study. Like the original publication, HADDOCK scores reported here are negative values instead of positive, which is why the boxplot uses negative numbers. The lower (more negative) the HADDOCK score is, the stronger the binding affinity between that specific neutralizing antibody and SARS-CoV-2 variant is. When comparing the two studies' results, it is apparent that many of the scores are in a similar range, with only the bond between XBB.1.5 and AZD1061 having vastly different old and new scores.

PYMOL AND CHIMERAX–After attempting to download and run both programs (PyMOL and ChimeraX), we found that ChimeraX is easier to install and more user-friendly. ChimeraX was sufficient for all the steps in the workflow used, including analyzing 3D structures and edit-

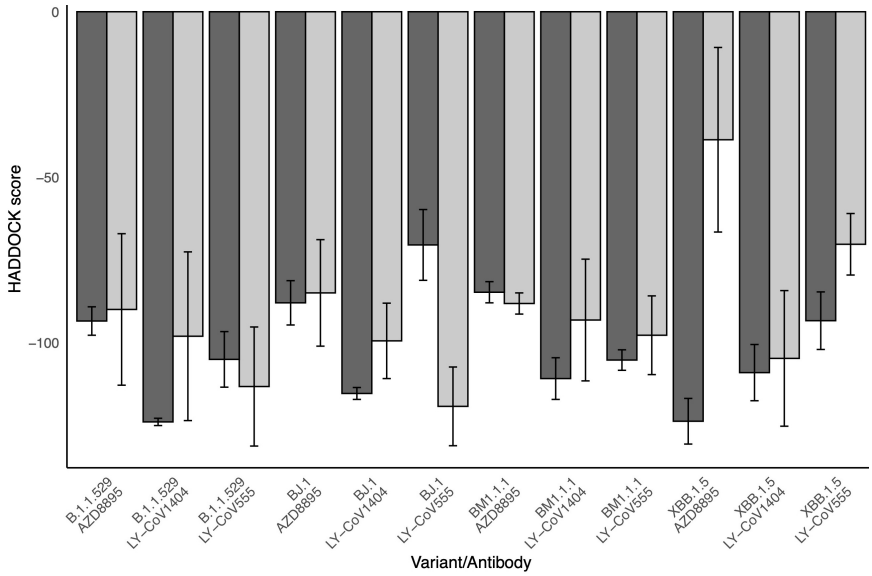

Figure 2. This plot summarizes our results and shows that we (light gray) were able to replicate [7] (dark grey) HADDOCK scores (Y-axis) for each pair of variant and antibody (X-axis).

ing files. Furthermore, we documented how to use ChimeraX instead of PyMOL in future analyses.

We note that installing and executing PyMOL was difficult, making it hard to reproduce [7]. ChimeraX performed the same as PyMOL when needed, making it a better choice for automation.

OTHER OUTCOMES–While replicating the original study, we broke its workflow into granular steps and produced a tutorial to facilitate its replication and automatization. We developed a flowchart and spreadsheets that compare AlphaFold Colab and ColabFold, including the original study's results. Finally, this abstract was produced with a poster (presentation: July 28, 2023, at UNC Charlotte, Charlotte NC) [24].

4. Conclusion

During this study, we engaged with a diverse array of AI tools. Through active participation in classes and discussions within UNC Charlotte's Phyloinformatics lab (phyloinformatics.com), we gained a profound appreciation for their pivotal role in the fields of bioinformatics and phylogenetics. By meticulously replicating the original study, we conducted a comprehensive comparative analysis of various alternative tools, including PyMOL and ChimeraX. Our findings will serve as a foundation for shaping future workflows.

Upon comparing ColabFold and AlphaFold Colab, we ascertained that ColabFold's parameters exhibit superior speed without compromising accuracy when compared to those of AlphaFold Colab. Consequently, we incorporated ColabFold's parameters in forthcoming automation endeavors.

In our evaluation of PyMOL and ChimeraX, we observed that ChimeraX not only offers a more intuitive user experience but it is also equally amenable to seamless integration within an automated workflow, akin to PyMOL. Remarkably, even for files necessitating modification, we found that manipulation within ChimeraX obviated the need for PyMOL entirely.

With regard to the HADDOCK results, we noted that the reported scores must be multiplied by -1 to conform to the standard adopted by other programs, wherein the most negative values signify the strongest binding affinities. This adjustment will be implemented automatically in future automated processes.

In summation, we successfully replicated the original study and meticulously documented our procedural steps. The ability to validate the original findings, even through applying distinct tools and methodologies, underscores the feasibility of automating this pipeline for large-scale drug screening. While some tools employed in this study may be better suited for automation than those in the original study, the work of Ford *et al.* [7] holds immense promise and can be reliably reproduced.

IN SHORT–The following are the main outcomes of our research project.

- ColabFold is faster and not less accurate than AlphaFold colab: we will use its parameters during automation.
- ChimeraX is as easy to implement in a pipeline as PyMOL is. However, ChimeraX was easier to install and manipulate: we will use ChimeraX instead of PyMOL for automation.
- HADDOCK scores have to be multiplied by -1 to report scores in the same standard as other programs where the most negative values are read as the strongest binding affinities: we will automatically adjust HADDOCK scores in the future during automation.
- Original paper's results [7] are reproducible even using slightly different methodologies: this further validates this workflow for large-scale drug screening.

Acknowledgements

This research was performed by participants of the University of North Carolina at Charlotte "REU Site: Smart and Future Computing" (College of Computing and Informatics) funded by the National Science Foundation (NSF; Grant No. 2244424). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors, not the NSF.

Reference

[1] E. Callaway, "Is coronavirus variant XBB.1.5 a global threat," *Nature*, Vol. 613, pp. 222–223, 2023.

[2] M. J. L. e. a. Gangavarapu Karthik, Latif Alaa Abdel, "Outbreak.info genomic reports: Scalable and dynamic surveillance of SARS-CoV-2 variants and mutations," *Nature Methods*, 2023.

[3] R. P. Karyakarte, R. Das, S. Dudhate, J. Agarasen, P. Pillai, P. M. Chandankhede, R. S. Labhshetwar, Y. Gadiyal, M. V. Rajmane, P. P. Kulkarni *et al.*, "Clinical characteristics and outcomes of laboratory-confirmed SARS-CoV-2 cases infected with Omicron subvariants and the XBB recombinant variant," *Cureus*, Vol. 15, No. 2, 2023.

[4] N. Lasrado, A.-r. Collier, J. Miller, N. Hachmann, J. Liu, M. Sciacca, C. Wu, T. Anand, E. Bondzie, J. Fisher *et al.*, "Waning immunity against XBB.1.5 following bivalent mRNA boosters," *bioRxiv*, pp. 2023–01, 2023.

[5] J. Miller, N. P. Hachmann, A.-r. Y. Collier, N. Lasrado, C. R. Mazurek, R. C. Patio, O. Powers, N. Surve, J. Theiler, B. Korber *et al.*, "Substantial neutralization escape by SARS-CoV-2 omicron variants BQ.1.1 and XBB.1," *New England Journal of Medicine*, Vol. 388, No. 7, pp. 662–664, 2023.

[6] Y. Cao, F. Jian, J. Wang, Y. Yu, W. Song, A. Yisimayi, J. Wang, R. An, X. Chen, N. Zhang *et al.*, "Imprinted SARS-CoV-2 humoral immunity induces convergent omicron RBD evolution," *Nature*, Vol. 614, No. 7948, pp. 521–529, 2023.

[7] C. T. Ford, S. Yasa, D. Jacob Machado, R. A. White III, and D. A. Janies, "Predicting changes in neutralizing antibody activity for SARS-CoV-2 XBB.1.5 using *in silico* protein modeling," *Frontiers in Virology*, Vol. 3, p. 1172027, 2023.

[8] C. Dominguez, R. Boelens, and A. M. Bonvin, "HADDOCK: A protein-protein docking approach based on biochemical or biophysical information," *Journal of the American Chemical Society*, Vol. 125, No. 7, pp. 1731–1737, 2003.

[9] G. Van Zundert, J. Rodrigues, M. Trellet, C. Schmitz, P. Kastritis, E. Karaca, A. Melquiond, M. van Dijk, S. De Vries, and A. Bonvin, "The HADDOCK2.2 web server: User-friendly integrative modeling of biomolecular complexes," *Journal of Molecular Biology*, Vol. 428, No. 4, pp. 720–725, 2016.

[10] A. Vangone and A. M. Bonvin, "Contacts-based prediction of binding affinity in protein-protein complexes," *eLife*, Vol. 4, p. e07454, 2015.

[11] L. C. Xue, J. P. Rodrigues, P. L. Kastritis, A. M. Bonvin, and A. Vangone, "PRODIGY: A web server for predicting the binding affinity of protein-protein complexes," *Bioinformatics*, Vol. 32, No. 23, pp. 3676–3678, 2016.

[12] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger, "Colabfold: Making protein folding accessible to all," *Nature Methods*, Vol. 19, No. 6, pp. 679–682, 2022.

[13] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer *et al.*, "Accurate prediction of protein structures and interactions using a three-track

neural network," *Science*, Vol. 373, No. 6557, pp. 871–876, 2021.

[14] I. R. Humphreys, J. Pei, M. Baek, A. Krishnakumar, I. Anishchenko, S. Ovchinnikov, J. Zhang, T. J. Ness, S. Banjade, S. R. Bagde *et al.*, "Computed structures of core eukaryotic protein complexes," *Science*, Vol. 374, No. 6573, p. eabm4805, 2021.

[15] L. Schrödinger and W. DeLano, "PyMOL." [Online]. Available: http://www.pymol.org/pymol

[16] E. F. Pettersen, T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris, and T. E. Ferrin, "UCSF ChimeraX: Structure visualization for researchers, educators, and developers," *Protein Science*, Vol. 30, No. 1, pp. 70–82, 2021.

[17] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, Bates *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, Vol. 596, No. 7873, pp. 583–589, 2021.

[18] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon *et al.*, "Alphafold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models," *Nucleic Acids Research*, Vol. 50, No. D1, pp. D439–D444, 2022.

[19] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, "GenBank," *Nucleic Acids Research*, Vol. 41, No. D1, pp. D36–D42, 2012.

[20] K. H. Nam, "AI-based protein models enhance the accuracy of experimentally determined protein crystal structures," *Frontiers in Molecular Biosciences*, Vol. 10, p. 1208810, 2023.

[21] P. W. Rose, A. Prlić, A. Altunkaya, C. Bi, A. R. Bradley, C. H. Christie, L. D. Costanzo, J. M. Duarte, S. Dutta, Z. Feng *et al.*, "The RCSB protein data bank: Integrative view of protein, gene and 3d structural information," *Nucleic Acids Research*, p. gkw1000, 2016.

[22] M. I. J. Raybould, A. Kovaltsuk, C. Marks, and C. M. Deane, "CoV-AbDab: The coronavirus antibody database," *Bioinformatics*, Vol. 37, No. 5, pp. 734–735, 2021. [Online]. Available: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btaa739/5893556

[23] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic Acids Research*, Vol. 28, No. 1, pp. 235–242, 2000.

[24] J. M. Denis, J. Cameron, and Z. Hannah, "Replicating Ford et al.'s investigation on the immune consequences and antibody binding affinity of SARS-CoV-2 variant XBB.1.5 using an *in silico* approach," 2023. [Online]. Available: https://doi.org/10.5281/zenodo.10068320