ADVANCING INFLUENZA VACCINE DESIGN THROUGH CONSENSUS-BASED APPROACHES: TARGETING H3N2 HA PROTEIN SEQUENCES

by

M H M Mubassir

(Under the Direction of Eva-Maria Strauch)

ABSTRACT

Vaccination is the cost-effective preventive measure for influenza; however, at times with antigenic mismatch, annual revision of seasonal influenza vaccine components due to constant-evolving nature of the virus can potentially occur with the vaccine target HA. This research project endeavors to pioneer advancements in the realm of influenza vaccine design by adopting a consensus-based methodology centered on the HA protein sequence of the H3N2 influenza strain. The approach involves the creation of synthetic HA proteins that represent the diversity of the H3N2 HA population, aiming to enhance cross-reactivity and broaden vaccine coverage. To achieve this, consensus sequences were meticulously generated for the entire HA sequence, as well as the HA head and HA stem regions which were strategically designed according to diverse phylogenetic trees, antigenic clusters, and vaccine strains. A total of 97 distinct sequences were derived, out of which 30 designed sequences were finalized to test *in-vitro*.

INDEX WORDS: Influenza A virus, vaccine, consensus, phylogenetic tree

ADVANCING INFLUENZA VACCINE DESIGN THROUGH CONSENSUS-BASED APPROACHES: TARGETING H3N2 HA PROTEIN SEQUENCES

by

M H M Mubassir

BSc, Bangladesh Agricultural University, Bangladesh, 2012

MSc, Bangladesh Agricultural University, Bangladesh, 2014

MPhil, Universiti Teknologi Malaysia, Malaysia, 2017

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2023

ADVANCING INFLUENZA VACCINE DESIGN THROUGH CONSENSUS-BASED APPROACHES: TARGETING H3N2 HA PROTEIN SEQUENCES

by

M H M Mubassir

Major Professor:	Eva-Maria Strauch
Committee:	Stephen M. Tompkins
	Eileen J. Kennedy

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
December 2023

DEDICATION

To my dearest parents Prof. Dr. A Q M Bazlur Rashid and Akhtar Jahan, my wife Marzia Khatun, my daughter Sarah, my siblings, in-law's family members and Bangladeshi community at the UGA for their unconditional love and patience in my journey. They have been my pillars of support through storms and sunshine, and I cannot imagine these years without them.

ACKNOWLEDGEMENTS

Firstly, I convey my sincere thanks to the Almighty Creator, our Lord Allah (SWT) for making everything possible as there is no power nor movement without His permission. Peace be upon Prophet Muhammad (SAW), the last messenger whose beautiful teachings of peace and prosperity shaped my psychology and life's philosophy.

I am greatly indebted to my beloved family members back home for their unconditional love, support and prayer. I would like to express my heartfelt gratitude for my supervisor Dr. Eva-Maria Strauch for his constant support and guidance throughout my study with patience and enthusiasm. I am fortunate to be able to work with her, without which part of this work would not have been possible. Dr. Strauch was particularly flexible and considerate, and I certainly enjoy the lab environment in which I am surrounded by a bunch of encouraging and warmhearted peers. I am greatly thankful to my committee members Dr. Eileen J. Kennedy and Dr. Stephen M. Tompkins for their tremendous support in my research with lot of constructive advice that reshaped my thesis work.

I want to specially thank my wife Marzia Khatun for her constant motivation and support to keep me going. Without her continuous help, it was impossible to finish the reseach. I am grateful to my daughter Sarah Mubashsarah for her unconditional love. I am grateful to all my lab members including Dr. Raulia Syrlybaeva, Karen Juliana Gonzalez Restrepo, Qingfa Hou, Je Hoon Michael Oh, Laiba, and Ho Suk Lee without whom it was impossible. The suggestions and mentoring of my seniors Dr. Raulia and Dr. Karen regarding research were precious.

I am indebted to friends Masud Parvez, Marzan Sarkar, Nabil Hassan, Ehsan Suez, Asiful Alam, Pritam Sarkar, Avik Das and Himadry Sekhar Das for their unimaginable support in every moment. Special thanks to all the members of BASA, UGA. Their company made my life colorful and very enjoyable here.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS v

LIST OF TABLES ix

LIST OF FIGURES x

CHAPTER

1 INTRODUCTION 1

2 LITERATURE REVIEW 4

Influenza 4

Etiology 5

Structure of influenza virus 6

Influenza A virus (IAV) subtypes 8

1968 pandemic and H3N2 9

H3N2 introduction and evolution 10

Antigenic drift and shift 15

Antigenic drift of H3N2 16

Influenza vaccine design strategies 18

Hemagglutinin (HA) structure and function 20

Novel HA based vaccine design strategies 22

Consensus-based vaccine design strategies 23

Nomenclature for influenza viruses 26

3 MATERIALS AND METHODS 27
Dataset 27
Yearly consensus sequence generation 28
Consensus based on time window clustering 28
Consensus based on phylogenetic tree clustering 29
Consensus based on H3N2 recommended vaccine strains 30
Consensus based on H3N2 antigenic clusters 31
Strategies to address consensus ambiguity 31
Filtering the designs 34
4 RESULTS AND DISCUSSION 36
Yearly consensus sequence generation 36
Consensus based on time window clustering 37
Consensus based on phylogenetic tree clustering 38
Consensus based on H3N2 Recommended Vaccine Strains 54
Consensus based on H3N2 Antigenic Clusters 55
Finalizing the designs for order 57
Addition of Foldon, Histag and Codon optimization 70
5 SUMMARY AND CONCLUSION 74
REFERENCES 76

LIST OF TABLES

Page

Table 2.1: WHO recommended vaccine strains for H3N2 from 1968 to 2000 11

Table 2.2: WHO recommended vaccine strains for H3N2 from 2000 to 2023 13

Table 3.1: Different template information used to address consensus ambiguity 32

Table 4.1: Number of designed sequences for different set of consensus design.....56

Table 4.2: The rationale behind finalizing the designs (batch 1, 16 designs).....57

Table 4.3: The rationale behind finalizing the designs (batch 2, 14 designs).....63

LIST OF FIGURES

Page

Figure 2.1: Different types of influenza virus with their host information 6

Figure 2.2: The configuration of the hemagglutinin (HA) trimer, monomer, and HA2 domain within the H3 subtype of influenza A 21

Figure 3.1: Consensus vaccine design strategies that was incorporated in the study 31

Figure 4.1: Consensus phylogenetic tree for HA entire sequences based on UPGMA 39

Figure 4.2: Consensus phylogenetic tree for HA head sequences based on UPGMA 41

Figure 4.3: Consensus phylogenetic tree for HA stem sequences based on UPGMA 43

Figure 4.4: Consensus phylogenetic tree for HA entire sequences based on neighbor-joining 45

Figure 4.5: Consensus phylogenetic tree for HA head sequences based on neighbor-joining 46

Figure 4.6: Consensus phylogenetic tree for HA stem sequences based on neighbor-joining 48

Figure 4.7: Consensus phylogenetic tree for HA entire sequences based on ML 50

Figure 4.8: Consensus phylogenetic tree for HA head sequences based on ML 52

Figure 4.9: Consensus phylogenetic tree for HA stem sequences based on ML 53

Figure 4.10: Schematic representation of the finalized 16 designs of batch 1 63

Figure 4.11: Schematic representation of the finalized 14 designs of batch 2 69

Figure 4.12: Number of finalized designs under each consensus design categories 70

Figure 4.13: Representative of final design where we added signal peptide, foldon and 6 his-tag with each designed sequence..... 71

CHAPTER 1

INTRODUCTION

Influenza is a respiratory viral disease which can cause severe illness in different age group of human population. An annual infection with influenza viruses may result in approximately 950,000 hospitalizations and 250,000 to 500,000 death every year [1, 2]. Among the various types of influenza, influenza A viruses (IAV) are currently circulating dominantly. All the historical influenza pandemics in human have been caused by IAV [2, 3]. To tackle this, yearly vaccines are accessible, encompassing live attenuated or inactivated virus variations in various formulations. However, these vaccines shows limited effectiveness due to the rapid change in the virus genes known as antigenic drift [4, 5].

Antigenic drift occurs when point mutations are introduced in the virus surface proteins mostly in the hemagglutinin (HA) and neuraminidase (NA) protein [5, 6]. These mutations can alter the binding of antibody with the HA, resulting in escaping the immune responses [7]. Consequently, vaccines must be modified according to projected strains of circulating influenza viruses in the upcoming season [8]. However, if the strains have undergone mutations or different strains are prevalent, the vaccine's protective capacity diminishes. This presents a significant challenge in selecting appropriate strains for the vaccine, complicated further by the need for selection up to nine months prior to the influenza season due to the production and distribution timeline of vaccines [5, 8, 9].

In 1968, a pandemic emerged when a reassortant avian virus H3N2 occurred in human, causing a global outbreak [10-12]. Since being introduced in 1968, H3N2 influenza viruses have

experienced substantial genetic and antigenic modifications, leading to the occurrence of many seasonal epidemics. This evolution is evident in the World Health Organization's recommendation of 46 vaccine strain modifications during this period.

The genetic segment within the influenza HA encodes a glycoprotein that forms an elongated trimer structure on the surface of the viral capsid. This glycoprotein is of central importance as it serves as a principal focus for the immune response via antibodies during instances of influenza infection. The HA protein undergoes cleavage, splitting into two distinct polypeptide chains designated as HA1 and HA2, linked by di-sulfide bonds [13, 14]. Its role involves binding to glycans containing sialic acid within the receptor binding site (RBS) located on the viral surface. This binding action facilitates the virus's attachment to cells in the upper respiratory tract [15]. HA serves as a focal point for the development of seasonal vaccines and other numerous candidates aiming for the creation of universal influenza vaccines [16]. While antibodies against HA's head region are generated through infection or vaccination, they're often strain-specific due to the high mutation rate around the RBS [17]. New vaccine strategies involve sequential immunization with synthetic HA constructs, 'headless' HA constructs, and prime/boost immunization to generate cross-neutralizing antibodies and heterosubtypic immunity [18]. Another strategy, known as consensus-based vaccine design focuses on creating synthetic hemagglutinin (HA) proteins that represent the diversity of the HA population. Consensus based strategies are initially subtype-specific, but could potentially extend to multiple subtypes and a combined multivalent vaccine. Various innovative approaches within the consensus framework aim to enhance cross-reactivity, such as micro-consensus immunogens [19], COBRA [20, 21] and centralized HA genes [22]. The key challenges in consensus design is to remove the potential bias on selecting the population for vaccine design.

This study aims to address the challenges posed by the antigenic drift of H3N2 influenza by exploring consensus-based vaccine design. The approach involves creating synthetic HA proteins that represent the diversity of H3N2 HA variants to enhance cross-reactivity and broaden vaccine coverage. This was achieved by following the protocol developed by our lab for H1N1. We modified the protocol to generate consensus sequences for the entire HA sequence, the HA head and HA stem regions, strategically designed based on various phylogenetic trees, antigenic clusters, and vaccine strains. The goal is to create a universal vaccine strain that covers a wide spectrum of influenza H3N2 diversity. This research has the potential to advance influenza vaccine design, offering more effective and adaptable solutions against the evolving influenza virus.

CHAPTER 2

LITERATURE REVIEW

Influenza

Influenza, a contagious viral disease, affects birds, mammals (including humans), and has been known for over four centuries since its initial documentation in 1580 [23]. It has led to periodic pandemics and numerous seasonal epidemics. Presently, influenza remains a persistent global health concern, periodically drawing attention from the medical community. Influenza pandemics occur when a new, antigenically distinct strain of the virus emerges, resulting from the exchange of gene segments within the virus. This new strain then spreads among humans, particularly in populations lacking immunity.

In the 20th century, there were four major influenza pandemics. They comprised of heavily destructive H1N1 pandemic of 1918 popularly known as Spanish flu which estimated to have caused deaths across the globe ranging from 21 million to more than 26 million [3]. This was followed by H2N2 Asian flu in 1957 and then by H3N2 Hong Kong flu in 1968 finally by H1N1 Russian flu in 1977 [3, 24]. The only influenza belonging to this century happened to be the H1N1 swine-origin flu pandemic of 2009 [25]. It is pertinent here that most influenza pandemics came from non-human sources with aquatic wildfowl being one primary natural reservoir [26].

Influenza pandemics are a real threat, and seasonal influenza contributes to the annual disease burden. CDC estimates that for the 2022-2023 influenza season, since October 1, 2022 through April 30, 2023 there have been roughly 27–54 million cases of illness resulting in 12–26 million medical visits 300 – 650 thousand hospitalizations [27].

It is noted that, this research has been done for the Northern Hemisphere (NH) and Southern Hemisphere (SH), where, winter timing differs between NH and SH. In NH, influenza season duration is from November to April while in the SH occurs between May and October. In areas with tropical or subtropical climates, epidemiology of influenza varies throughout the year and as a result, defining any particular influenza period is difficult [28].

Etiology

Influenza viruses are all categorized under the Orthomyxoviridae family of RNA viruses [29]. Influenza enveloped viruses have a segmented, single negative-strand RNA genome that codes for different surface glycoproteins such as HA (hemagglutinin), causing interaction with cell surface receptors and also leading to virus entry while NA is an enzyme aiding in viral replication and helping in the release of the virus from its host cell. There has been established roles these viral glycoproteins play not only in influenza virulence but pathogenesis too [30-32].

There are four types of Influenza virus, Influenza A (IAV), Influenza B (IBV), Influenza C (ICV), and Influenza D (IDV) (Figure 2.1) [6]. Human population faced influenza epidemics due to either Influenza A or Influenza B [33]. IAV is the most common type during flu seasons which affects human [34]. IBV infections are very contagious. Infections due to IBV may cause serious illness occasionally as well but usually lead to a minor localized outbreak [35]. The severity of IBV is less compared to IAV in flu seasons [29]. ICV viruses can cause mild upper respiratory symptoms in children [36]. With occasional sporadic cases and minor localized outbreaks, ICV can also infect some animals like swine [37]. IDV viruses basically affect pigs and cattle. They have not been reported to cause any infections towards the human race. [38, 39].

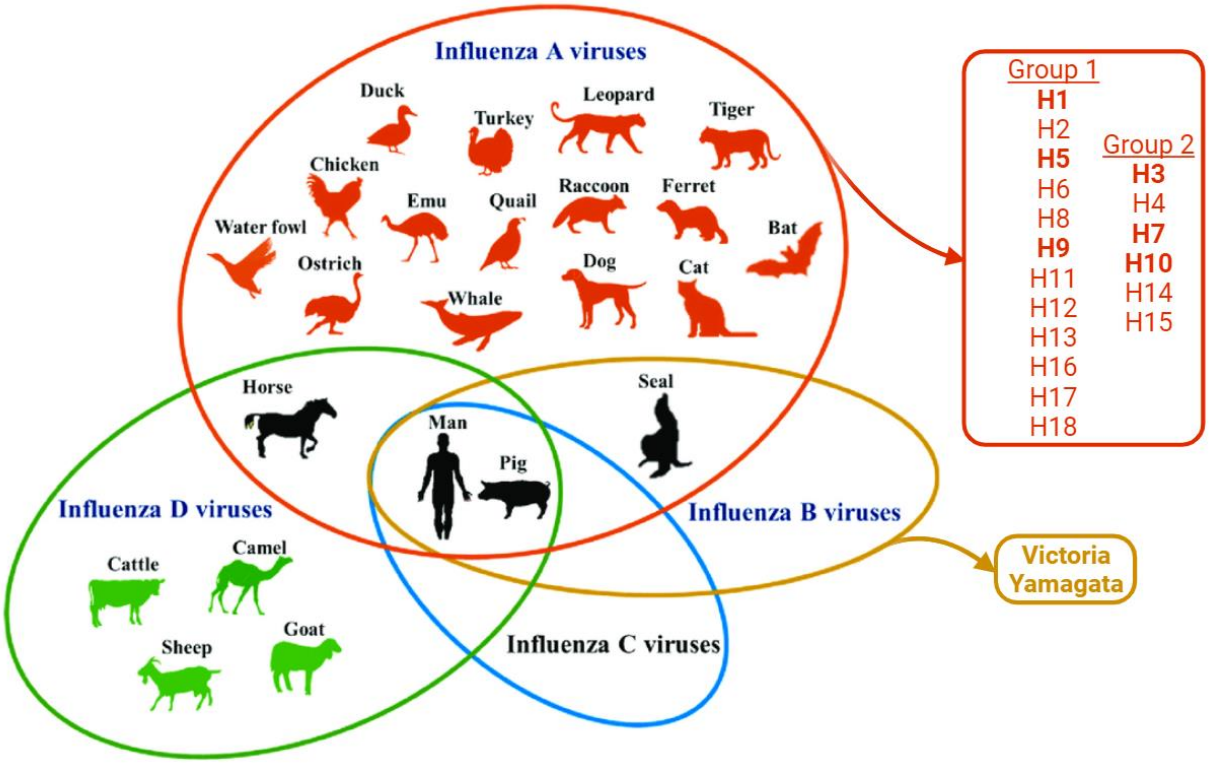

Figure 2.1. Different types of influenza virus with their host information. The bold letters under group 1 and group 2 indicates the subtypes that infects human. Figure adapted and modified from Kuchipudi *et al.* [40]

Structure of influenza virus

Both Influenza A and B viruses possess genomes comprising eight negative-sense, single-stranded viral RNA (vRNA) segments each, while influenza C virus has a genome consisting of seven segments. The segments in both influenza A and B viruses (and influenza C virus) are numbered based on their decreasing lengths [30]. The first three segments of the virus genome are the largest genes. Segment one encodes the polymerase basic two (PB2) protein, which is responsible for 5' cap recognition. Segment two encodes the polymerase basic one (PB1) protein, acting as a transcriptase, and also produces a small pro-apoptotic factor called PB1-F2 through a

second open reading frame. Segment three codes for two polymerase acidic proteins: PA (P3 for ICV and IDV) functioning as an endonuclease, and PA-X [30]. These accessory proteins are crucial for suppressing host defense mechanisms, enhancing virulence, and contributing to the overall pathogenicity of the virus. The fifth segment of the virus genome encodes three of the polymerase proteins (PB2, PB1, and PA) and the nucleocapsid protein, NP. They form a viral complex known as ribonucleoprotein (vRNP) complex which is crucial for the replication and transcription of the viral RNA (vRNA) [41, 42].

Segments four and six, respectively, of IAV and IBV, encode the viral surface glycoproteins, HA and NA. HA plays an important role in mediating the virus binding and internalization into cells while NA plays a critical role in releasing newly formed virus particles from cell surfaces to facilitate virally spreading [43]. Consistent with this, for ICV and IDV, they encode a hemagglutinin-esterase fusion protein on one segment which fuses both HA and NA functions in a single protein [44]. Segments seven and eight of the virus genome encode both, M1 and M2, but in different ways (unspliced mRNA versus spliced mRNA, respectively). Both encode the two viral matrix proteins while encoding the non-structural proteins NS1 and NS2. These proteins have essential roles in various aspects of the viral life cycle and infection process [45, 46]. Proteins M1, NS1, and NS2 have essential functions in facilitating the nuclear export of newly synthesized viral ribonucleoprotein (vRNP) complexes, which are made up of PB2, PB1, PA, and NP. These vRNPs, in conjunction with the primary viral structural protein M1, come together to form virions, which then separate from the cellular membrane as part of the viral assembly process. M2, acting as an ion channel protein, is also involved in this process and can function at different stages of the infection [47-49]. Additionally, the protein NS1 functions as an interferon (IFN) antagonist, suppressing the host's immune response [50].

Influenza A virus (IAV) subtypes

Influenza A virus (IAVs) are common and cause recurrent epidemics to human beings as well as domestic animal species like poultry, pigs, and horses [51, 52]. Influenza A viruses (IAVs) possess two variable glycoproteins, hemagglutinin (HA) and neuraminidase (NA), which are expressed in an active form on the surface of the virion within the host cell. So far, scientists have identified 18 different HA and 11 NA antigenic subtypes within IAV. Altogether, this has resulted in the discovery of over 120 distinct combinations of HA and NA, such as H3N2, H5N1, and H10N8 [53]. Like all RNA viruses, IAVs are characterized by an elevated mutation rate as well as during coinfection the reassortment of viral segments leading to novel strains. Most of the known IAVs occur naturally in birds with aquatic habits. [51, 54]. However like unique HA subtypes (H17 and H18) in several bat species, these might offer important reservoirs for diverse IAVs [55]. Only four IAV subtypes were identified to sustain human-to-human transmission leading to global pandemics: H1N1, H3N2, sporadically detected H1N2 and H2N2 [56].

In the last 2 decades alone, other IAV subtypes of avian origin were identified in humans. These included H6N1, H7N3, H7N7, H9N2, and H10N7 which caused primarily nonfatal symptoms as well as mild acute upper respiratory tract infections [57-59]. A patient died from infection with virulent avian bird flu in a human in the Netherlands 2003 [60, 61]. Conversely, H5N1 and H7N9 strains have been linked to alarmingly high mortality rates in individuals who become infected, despite the fact that these strains do not readily transmit from one human to another [62]. Recently, H10N8 was linked to a fatal pneumonia case [63]. The focus of this study is H3N2 of Influenza A virus.

1968 pandemic and H3N2

In Hong Kong, the influenza normally occurs in two periods, one starts from January and another from July. In July 1968, there was a sudden rise of patients with influenza-like illness (ILI) [11, 64]. This sudden outbreak was the largest one in Hong Kong since the 1957 H2N2 pandemic with about 500,000 ILI cases for July [64]. On July 17, 1968, a new influenza A(H3N2) virus was isolated as a clear distinct antigenic variant of influenza. On 16th August, the WHO issued a warning about this new viral strain [65]. Soon after that, the virus spread rapidly to other countries aided by air travel. The affected places included Thailand, Philippines, Singapore, Taiwan, Malaysia, Vietnam, and India [66].

In the United States, the first isolate of the A(H3N2) virus was obtained from a Marine who returned from Vietnam on September 2 of the same year [67]. Subsequent ILI cases were reported among the students and contacts from the Marine Corps Drill Instructors School in San Diego. Military personnel returning from southeast Asia also experienced outbreaks in Hawaii and Alaska. As a result, there was a surge in surveillance efforts in the United States, and influenza activity sharply increased in October. The initial civilian outbreak in the continental United States was identified in Needles, California, where a substantial portion of the population reported influenza-like illness (ILI). Over the subsequent weeks, other western states and Hawaii also reported outbreaks, and eventually, outbreaks occurred in eastern states as well [67]. The pandemic activity peaked between December 14 and January 11, with most states experiencing increased school absenteeism, and some states faced school and college closures due to the influenza spread. The pandemic's activity started in the western United States and moved eastward.

H3N2 introduction and evolution

Pandemic due to H3N2 is one among the major influenza pandemics that took place in human history. Its starting point was a novel influenza type H3N2 virus development originating in 1968, A/Hong Kong/1/1968 (HK/68) within Hong Kong. This type rapidly traveled across the entire world and became into a worldwide pandemic which caused over 1 million deaths all around the world. There had been no reported previous cases of human infections through any H3N2 virus before this outbreak. It is considered that the H3N2 strain originated from reassortment event between circulating human H2N2 viruses and avian H3N2 influenza viruses, consequently resulting in a novel H3N2 viral strain with ability to infect as well as transmit among humans [68, 69].

The new H3N2 strain was formed by combining the HA and PB1 fragments from avian H3N2 and the NA from the H2N2 pandemic strain of 1957 [70]. From 1968 onward, H3N2 IAVs have been regularly circulating within the human population, giving rise to numerous seasonal epidemics, substantial morbidity, and notable mortality [68]. The hemagglutinin proteins found on the surface of pandemic influenza viruses often exhibit variations compared to their avian predecessors. These differences arise from mutations in the receptor binding site (RBS), which in turn modify the viral receptor specificity. This alteration shifts the preference of the virus from binding primarily to α2,3 linked sialic acids (SAs) to α2,6 SAs. [71]. During the Hong Kong outbreak of 1968, the H3N2 isolates had a specific combination of five amino acid substitutions in their hemagglutinin (HA) protein. These substitutions played a crucial role in enabling the virus to adapt from birds to humans, ultimately leading to its emergence as a pandemic strain [69].

Since 1968, H3N2 influenza virus has constantly evolved mechanisms for evasion of host immune pressures. It achieves this through various mechanisms including the addition of N-

glycosylation sites, changes in the antigenic sites known as antigenic drift, and introduction of charged amino acid substitutions near the receptor binding site (RBS) of the HA protein [72]. These adaptations leading to WHO recommended 46 changes in vaccine strains (Table 2.1 and Table 2.2) to cope up with the emerging novel clades. The mutation that are responsible for H3N2 antigenic drift occurs every 2-5 years [73]. During these periods, antigenic drift can lead to the virus staying within a group of sequence variants that share similar antigenic properties, effectively forming an antigenic clade. [74].

Table 2.1. WHO recommended vaccine strains for H3N2 from 1968 to 2000. Yellow color represents long lasting vaccine strains which were effective for 4 years or more than 4 years.

H3N2 vaccine strain	**1974**	**1975**	**1976**	**1977**	**1978**	**1979**	**1980**	**1981**	**1982**	**1983**	**1984**	**1985**	**1986**	**1987**	**1988**	**1989**	**1990**	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**
A/Port Chalmers/1/1973	3 years																										
A/Victoria/3/75			3 years																								
A/Texas/1/77					3 years																						
A/Bangkok/01/1979							4 years																				
A/Philippines/2/82										4 years																	
A/Mississippi/1/85													2 years														

A/Christchurch/4/1985													2 years												
A/Leningrad/360/1986														2 years											
A/Sichuan/02/87															2 years										
A/Shanghai/11/87																2 years									
A/Guizhou/54/89																	2 years								
A/Beijing/353/89																		3 years							
A/Beijing/32/92																			2 years						
A/Shangdong/9/93																				2 years					
A/Johannesburg/33/94																					2 years				
A/Wuhan/359/95																						3 years			
A/Nanchang/933/95																							2 years		
A/Sydney/5/97																								3 years	

Table 2.2. WHO recommended vaccine strains for H3N2 from 2000 to 2023. Yellow color represents long lasting vaccine strains which were effective for 4 years or more than 4 years.

H3N2 vaccine strain	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**
A/Moscow/10/99	5 years																							
A/Panama/2007/1999	5 years																							
A/Fujian/411/2002					2 years																			
A/Kumamoto/102/2002					2 years																			
A/Wyoming/03/2003					2 years																			
A/Wellington/1/2004						1 year																		
A/New York/55/2004						2 years																		
A/California/7/2004						2 years																		
A/Wisconsin/67/2005							3 years																	
A/Hiroshima/52/2005							3 years																	
A/Brisbane/10/2007									3 years															
A/Uruguay/716/2007										2 years														

A/Wisconsin/15/2009											2 years												
A/Perth/16/2009											3 years												
A/Victoria/210/2009												1 year											
A/Victoria/361/2011													3 years										
A/Texas/50/2012														3 years									
A/Switzerland/9715293/2013																2 years							
A/South Australia/55/2014																2 years							
A/Stockholm/6/2014																2 years							
A/Norway/466/2014																2 years							
A/Hong Kong/4801/2014																	3 years						
A/Singapore/INFIMH-16-0019/2016																			2 years				
A/Switzerland/8060/2017																				1 year			
A/Kansas/14/2017																				2 years			

A/South Australia/34/2019																				1 year			
A/Hong Kong/2671/2019																				2 years			
A/Cambodia/e0826360/2020																					2 years		
A/Darwin/9/2021																						2 years	

Antigenic drift and shift

Influenza viruses increase diversity in their protein sequences through reassortment between different strains, mutations, and genomic evolution, which help them escape the host's immune system and antibody pressure [68]. This evasion of the host immune response is achieved through changes in the antigenicity of HA and NA, known as antigenic drift and antigenic shift [16]. Antigenic drift refers to gradual changes in the viral antigens, whereas antigenic shift involves abrupt changes that can lead to the emergence of new pandemic strains.

Antigenic drift results from point mutations in the viral genome, particularly in essential sites on the virus's surface proteins, with the HA protein being particularly vulnerable to this process. It occurs due to the host's immune response, leading to the emergence of new strains that can escape pre-existing immunity [75]. Antigenic drift drives the continuous evolution of seasonal influenza viruses, requiring regular updates of the influenza vaccine components [76]. Moreover, antigenic drift can modify the HA protein's host receptor binding site, favoring a preference for human sialic acid linkages over avian ones, influencing the virus's host tropism [77]. Other internal

proteins, such as NP, PA, and PB2, contain mutation sites crucial for viral replication [78]. The E627K mutation can convert a nonlethal H5N1 virus isolated from a human into a lethal strain in mice [79]. Additionally, avian H5N1 can become transmissible between ferrets with just five mutations, highlighting the potential for cross-species transmission [80, 81].

Antigenic shift is the second evolutionary mechanism of influenza viruses, achieved through the recombination or reassortment of viral genome segments. This occurs when a cell is co-infected by two different subtypes of viruses, allowing one strain to acquire the genome segment of the other, particularly the HA segment. Influenza viruses' segmented genomes enable this reassortment process [82]. When cells are infected by a mixture of viruses from various host species, such as humans and animals, the viral genome segments can undergo reassortment, leading to the creation of new hybrid strains that may be highly pathogenic and lack preexisting immunity in the human population [83]. However, antigenic shift mostly occurs among influenza viruses of the same genus, with IAV being the most common candidates. While this process generates diverse influenza strains in birds, it is less frequent in human [84]. Bats, pigs, and quails serve as potential "mixing vessels" for reassortment due to their sialic acid receptors being compatible with both mammalian and avian IAVs [85, 86]. Pandemic influenza usually arises from antigenic shift when a new virus strain lacks immunological immunity in the human and at the same time can sustain the transmission. This highlights the importance of antigenic shift in driving the emergence of novel and potentially dangerous influenza strains.

Antigenic drift of H3N2

The HA (Hemagglutinin) gene segment of the influenza virus contains instructions for a glycoprotein that assembles into an elongated trimer structure on the surface of the viral capsid.

This glycoprotein serves as a primary target for the antibody response during influenza infections. The HA protein is cleaved into two chains, HA1 and HA2, which are joined together by multiple di-sulfide bonds [87]. RBS site of the HA binds to sialic acid-containing glycans and facilitates the attachment of it to cells in the upper respiratory tract [71]. Mutations in HA allow influenza viruses to evade antibody neutralization.

Prior research using monoclonal antibodies has pinpointed five specific antigenic sites (A to E) located on the globular head of the HA protein. These sites are the primary targets for these antibodies during the immune response [88, 89]. Between the introduction of the A/H3N2 virus in humans in 1968 and 2003, there were a total of eleven clusters of influenza viruses, each with distinct antigenic properties. These clusters were successively replaced over time [90]. Until recently, it was believed that viruses needed multiple amino acid substitutions in at least two of these sites to form a new antigenic cluster. However, recent studies by Koel and colleagues [91] revealed that major antigenic evolution in these viruses can be attributed to substitutions at only seven amino acid positions. Surprisingly, in seven cluster transitions, only a single substitution in amino acid was adequate to explain the difference in antigenicity, with two and three substitutions causing two and one cluster transitions, respectively [91]. Furthermore, in subsequent periods, genetic clades known as CAL04, WI05, BR07, and PE09 also emerged, forming distinct antigenic clusters. However, the precise genetic factors that determine these clusters are not fully understood [92]. It's worth mentioning that the genetic clades PE09 and 3C.1 were found to be antigenically similar, sharing identical amino acids in the seven positions described by Koel et al. [93].

The mutations that drive transitions between antigenic clusters in the HA gene segment of influenza viruses are situated at specific positions: 145, 155, 156, 158, 159, 189, and 193. These particular sites are positioned on the exposed surface of the HA protein, closely adjacent to the

RBS [93]. Among these mutations, two of them, found at positions 189 and 193, are located near each other on the exposed side of an alpha-helix within antigenic site B. This alpha-helix runs alongside the receptor binding site (RBS) near the trimer interface. The residues at positions 155, 156, 158, and 159 are positioned in a loop adjacent to this helix and have the potential to interact with nearby residues, including residue 193 [91, 93].

These specific amino acid positions play a crucial role in the antigenic evolution of various influenza virus lineages. For instance, mutations at position 159 were associated with reduced antibody binding to ferret antibodies in genetic clade 3C.1 [94]. Similarly, the recent antigenic changes observed in genetic clades 3C.2A and 3C.3A are associated with the mutations Phe159Tyr and Phe159Ser, respectively [95]. It's important to note that the introduction of a potential N-linked glycosylation site at position 158 is also regarded as a significant factor contributing to the antigenic changes observed in the 3C.2A clade [96].

Moreover, these critical amino acid positions are not exclusive to human A/H3N2 viruses; they also play a crucial role in other influenza virus lineages, including human A/H1pdm [97], A/H2N2 [98], avian A/H5N1 [99], equine A/H3N8 [100], and swine A/H1 [101] and A/H3 viruses [102]. Single amino acid mutations at these sites have been shown to enable virus escape from strain-specific antibodies and receptor binding pocket targeting broadly neutralizing monoclonal antibodies [103].

Influenza vaccine design strategies

Despite annual seasonal influenza infections, vaccination remains a crucial strategy for influenza prevention. Currently, vaccines are manufactured in "trivalent" or "quadrivalent" formulations. These vaccines contain components derived from A(H1N1) pdm09, A(H3N2)

influenza A viruses, and either one or two influenza B viruses, which can belong to either the Victoria or Yamagata lineages [104]. In the 2021-2022 influenza season, the United States licensed three vaccine types: inactivated, recombinant, and live-attenuated [105]. The inactivated influenza vaccine can be produced using either the whole inactivated virus (WIV) method or the split virus method [106, 107]. However, WIV vaccines can lead to adverse effects [107]. Recombinant influenza vaccines utilize the baculovirus vector system to express the HA protein on the surfaces of insect cells [107].

The live attenuated vaccine is designed for intranasal administration and is cold-adaptive and temperature-sensitive [108]. Yet, the current vaccines have limitations and they require annual adjustments to match circulating strains due to antigenic evolution. Vaccine strains are selected by the World Health Organization (WHO) for northern and southern hemispheres, leading to variable protection due to mismatches. Additionally, dependency on egg-based production presents challenges, such as longer timelines, limited scalability in pandemics, and poor replication of certain strains [109-111]. This emphasizes the necessity of developing a universal influenza vaccine to enhance pandemic preparedness. Yet, because of the variability in the head domain of the influenza virus, the immune response primarily targets specific strains. Innovative vaccine platforms are required to stimulate both humoral and cellular immune responses, providing broader protection against a range of influenza viruses originating from animal reservoirs. Universal vaccine candidates should aim to offer immunity against viruses of the same subtype (homosubtypic) as well as those of different subtypes (heterosubtypic), encompassing a wider spectrum of influenza strains.

Hemagglutinin (HA) structure and function

HA plays a pivotal role in influenza virus function and antigenicity. As the primary protein on the virus surface, HA mediates viral entry by facilitating receptor binding and subsequently enabling the fusion of the viral membrane with the host cell membrane [112]. It also serves as a key antigen, leading to substantial adaptive evolution within influenza strains [113]. Despite sequence diversity among subtypes, HA retains essential components such as different structural motifs, cleavage site and fusion domain and forms a homotrimer structure on the virion surface. Initially, HA exists as a single polypeptide precursor known as HA0. To become functional and necessary for viral infectivity, it must undergo cleavage by host proteases, resulting in the production of two subunits: HA1 (head) and HA2 (stem). This cleavage process is essential for the maturation of HA. [114]. HA2, which primarily consists of the stalk region and the C-terminus, serves to anchor the protein to the virion envelope. On the other hand, HA1 contains the N-terminal signal peptide and forms the globular head domain (Figure 2.2). This globular head domain contains the receptor binding site, which interacts with sialic acid molecules on the surface of host cells. This interaction is the initial step in the process of viral entry, where the virus attaches to and enters the host cell. [41].

Upon internalization, endosomal acidification prompts a change in the HA conformation, exposing the fusion peptide of the HA2 subunit. This induce the membrane fusion process and later on, release of viral RNA into the host cell [116, 117]. Both the globular head and stems contains antigenic sites, targeted by neutralizing antibodies interfering with HA-sialic acid binding, a process essential for viral entry [118].

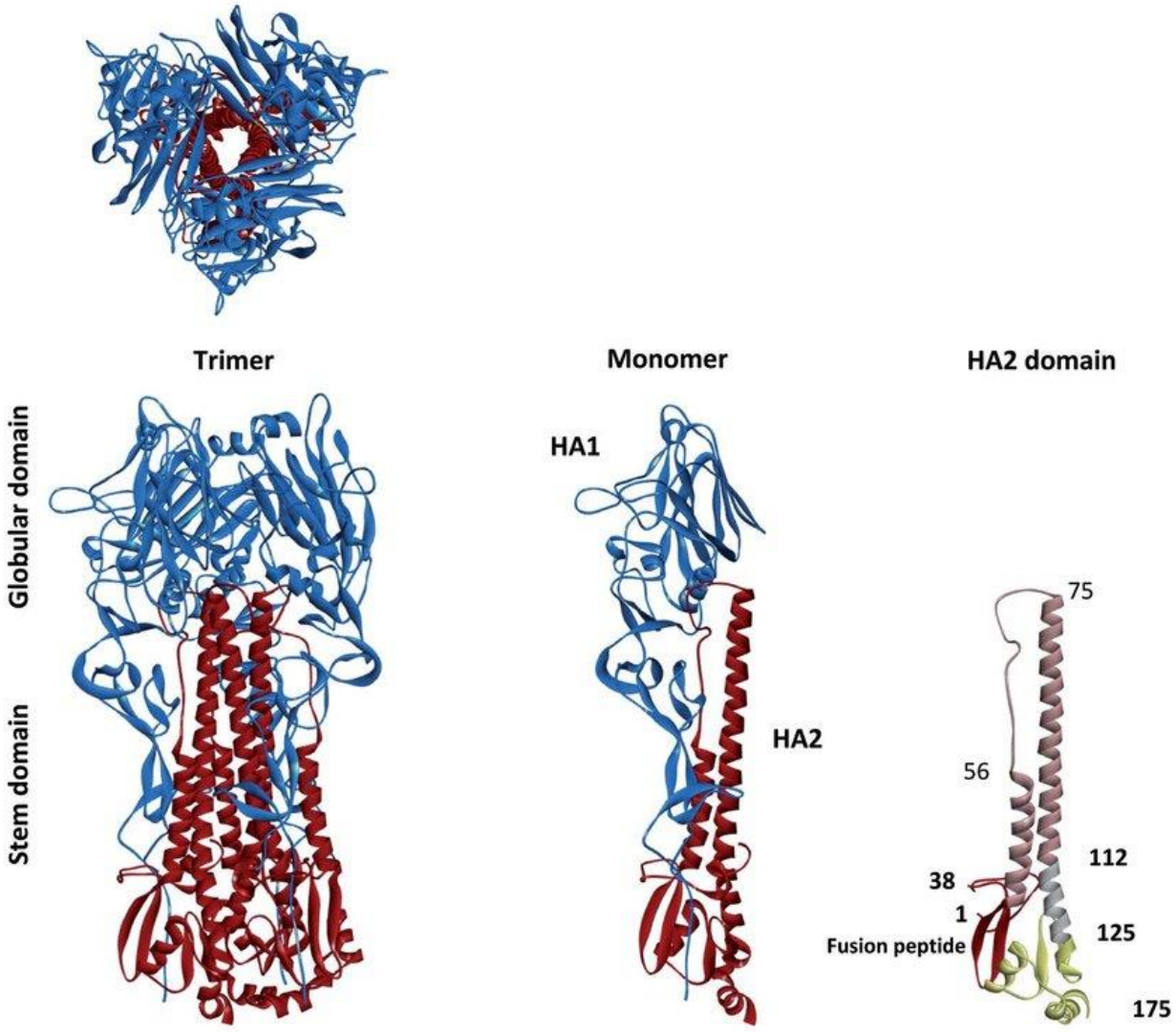

Figure 2.2. The configuration of the hemagglutinin (HA) trimer, monomer, and HA2 domain within the H3 subtype of influenza A. In both the trimer and monomer forms of HA, the prominent immunoglobulin-like domain is depicted in blue, while the less immunoreactive stem domain is represented in red. In HA2 domain, the fusion peptide is colored as red. Figure taken from Kostolanský et al. [115]

Moreover, HA, as a glycoprotein, goes through significant glycosylation [119]. Despite overall architectural conservation among HA from various subtypes, there are differences in

sequence composition and glycosylation patterns, particularly near the the globular head [119]. The RBS of the head is a shallow pocket enclosed by the 130-loop, 190-helix, and 220-loop, with four highly conserved amino acids at its base [17]. In H3N2, N-linked glycosylation sites changes upon human transmission, potentially contributing to viral antigenic drift [119-121]. This insight into HA's structure and function underscores its significance in influenza infection and immune responses.

Novel HA based vaccine design strategies

Antibodies targeting hemagglutinin (HA) play a pivotal role in immunity against influenza infection. The protection against influenza is measured through antibody responses using the HAI assay against HA. This has led to HA being a key target in both current seasonal vaccines and potential universal influenza vaccine candidates [16]. Structural biology advancements have deepened the understanding of HA's structure, allowing for structure-guided vaccine design [16, 122]. Antibodies directed at HA's head region are commonly generated through infection or vaccination and offer protection by hindering viral entry or endocytosis. However, these antibodies are often strain-specific and ineffective against drifted variants [17].

Moreover, broadly neutralizing antibodies targeting the more conserved stem region of HA have been reported. These antibodies, exemplified by CR6261 and F10, exhibit cross-reactivity with multiple HA subtypes [123-126]. Their binding inhibits the fusion process by stabilizing HA in its prefusion state [127]. Other stem-directed antibodies have exhibited distinct inhibitory mechanisms, such as preventing proteolytic cleavage of HA or disrupting viral replication [127]. Novel approaches to vaccine design have emerged, such as sequential immunization is done with different synthetic, chimeric HA constructs where the HA1 domain is from novel subtypes keeping

the stem constant [128]. Prime/boost immunization with these constructs has shown promise in generating cross-neutralizing antibodies and protecting animals from heterologous viral challenges [128]. Furthermore, 'headless' HA constructs [125], presenting only HA2 or the HA stem, have been engineered and demonstrated heterosubtypic immunity in animals [18, 129]. While strain-specific head-targeting antibodies are commonplace, the pursuit of broadly neutralizing antibodies targeting the stem region presents a promising strategy for achieving broader protection against a range of influenza viruses. Novel vaccine designs, including chimeric constructs and 'headless' immunogens, are advancing towards clinical evaluation, raising hopes for more effective and comprehensive influenza protection.

Consensus-based vaccine design strategies

Consensus-based strategies is focus on the entire HA protein with an aim of creating a synthetic HA that represents the diverse HA population. Consensus HA proteins are created by identifying the most frequently occurring amino acid at each position within the HA protein sequence population after aligning multiple sequences. This approach helps to capture the most prevalent genetic features among the variants. These synthetic HA genes are designed computationally. Unlike stalk-directed methods, consensus-based approaches primarily elicit hemagglutination inhibition (HI) antibodies targeting the HA's head region, a recognized protection factor against influenza infection [130, 131]. These approaches are subtype-specific initially, with potential extension to multiple subtypes and the creation of a combined multivalent vaccine. A few of the consensus-based strategies of vaccine design are described below:

Researchers have focused on constructing consensus H5 genes and proteins, demonstrating their potential to induce cross-reactive immune responses. When mice were vaccinated with a DNA plasmid containing a consensus H5 gene, they developed antibodies that could react with

multiple H5 viruses originating from different clades. This suggests that the vaccination induced a cross-reactive immune response capable of targeting a broad spectrum of H5 influenza strains [132]. Moreover, chickens that received vaccination with a consensus H5 protein incorporated into virus-like particles (VLPs) were able to achieve complete protection from a lethal challenge posed by H5 viruses belonging to distinct clades. This demonstrates the effectiveness of the vaccine in providing broad immunity against diverse H5 influenza strains [133]. However, a challenge of consensus immunogens lies in their reliance on all available H5 protein sequences, potentially leading to geographical bias and inaccurate representation of HA diversity [130].

Innovative strategies have emerged within the consensus framework to enhance cross-reactivity. One such approach involves micro-consensus immunogens, where multiple consensus HAs per subtype are developed. These micro-consensus genes, administered as a cocktail, improved cross-reactivity against a diverse range of H3 subtype strains in mice [19]. The cocktail induced robust antibody responses and HA-specific cellular immunity, affording protection against lethal H3N2 challenges [19].

To address the potential bias in consensus design, researchers have pursued centralized HA genes that reside at the phylogenetic tree's central node. This strategy reduces genetic and antigenic disparities among unmatched strains. Weaver et al. [22] developed a centralized H1 gene by selecting HA sequence representatives from each major branch of the phylogenetic tree to construct a consensus. This centralized H1 gene was inserted into an adenovirus vector, yielding better cross-protection against H1 strains compared to traditional vaccines [22]. This approach was extended to create centralized H3 and H5 genes, both demonstrating improved cross-protection [134]. Combining H1, H2, H3, and H5 into a multivalent vaccine did not compromise cross-reactivity [135].

The COBRA (computationally optimized broadly cross-reactive antigen) method offers an alternative strategy to address potential bias in consensus-based vaccine design, aiming to enhance cross-reactivity against various influenza strains. This approach employs multiple rounds of consensus generation to minimize sampling and sequencing bias within the target population. COBRA vaccines have been developed for different influenza subtypes, including H1, H2, H3, H5, and swine H1 [20, 136-140]. The initial COBRA strategy targeted H5 clade 2 viruses, tested in mice, ferrets, chickens, and non-human primates (NHPs) [141-144]. NHP vaccination generated cross-reactive antibodies against multiple clade 2 viruses, as well as clade 1 and 7 influenza viruses [145]. Multiple COBRA H1 immunogens were designed, with a similar new strategy called next-gen COBRA design [146]. Similarly, for the H3 subtype, multiple COBRA immunogens were designed, with some of those showing promising results in ferret [20, 21]. Ferret models demonstrated increased cross-reactive antibody titers against the entire panel of H3 strains in pre-immune ferrets compared to naïve ones [21]. This suggested that COBRA immunogens were effective in boosting cross-reactive immunity in pre-immune adults, with potential implications for children, although further research is needed.

While COBRA strategies show promise, challenges persist. These approaches target the full-length HA protein, eliciting subtype-specific antibodies primarily against the variable head region. Thus, they lack the broad cross-subtype reactivity of stalk-directed strategies. However, these vaccines exhibit enhanced cross-reactivity within subtypes and often induce hemagglutination inhibition (HI) activity. Consensus-based HA vaccines are relatively new, lagging behind stalk-based approaches in human clinical trials.

Additionally, little research has addressed escape mutants from pre-existing immunity induced by consensus vaccines. These subtype-specific vaccines require a multivalent cocktail to

protect against diverse subtypes circulating in humans. Ensuring that a multivalent vaccine maintains efficacy without interference is a critical challenge. Nonetheless, the success of the centralized HA approach in mice suggests promise for inducing robust immune responses against multiple subtypes important for human health. We developed a new protocol to design consensus-based vaccine for different subtypes. The detailed protocol is illustrated in the method section of the thesis.

Nomenclature for influenza viruses

The nomenclature for influenza viruses includes the virus type, the species it was isolated from (if non-human host), the location of isolation, an isolation identifier, the isolate year, and for IAVs, numbering for the HA and NA subtypes. For instance, A/Hong Kong/1/1968 (H3N2) was a human IAV isolated from Hong Kong in 1968, with isolation number 1, HA subtype 3, and NA subtype 2. Another example, A/Swine/Netherlands/3/1980 (H1N1), was a swine IAV isolated from the Netherlands in 1980, with isolation number 3, HA subtype 1, and NA subtype 1.

CHAPTER 3

MATERIALS AND METHODS

Dataset

To acquire the necessary dataset for our study, we retrieved all available full-length amino acid sequences of human HA (H3N2) from the GISAID database (https://gisaid.org) [147] spanning the period from 1968 to May 2023. Due to a download limitation of 20,000 sequences at a time, we employed a strategy based on isolation year to download the sequences. Specifically, we ensured that only complete HA sequences were included in the dataset. The total number of sequences obtained through this process was 114,090.

To ensure data integrity and remove any duplicate entries, we utilized the rmdup tool from seqkit [148]. Additionally, we developed an in-house Python script to truncate the header information associated with each sequence. In this script, we replaced the original sequence names with concise and unique identifiers that contained only the most pertinent information, such as the strain name and the year of isolation. We also employed another Python script to verify the presence of proper header information in each sequence. Any sequences with incomplete or inadequate header information, lacking year information, were eliminated from further analysis.

Following these preprocessing steps, we obtained a refined dataset consisting of 29,238 unique sequences as our primary database for subsequent analyses. Recognizing that the muscle alignment tool (MUSCLE 3.8.31) [149] performs optimally with datasets of less than 10,000 sequences, we further partitioned the database into four subsets using an additional Python script.

Yearly consensus sequence generation

To facilitate the generation of consensus sequences for each year, the sequences were first divided based on their isolation year. Subsequently, each group was aligned using MUSCLE 3.8.31 [149]. To mitigate the bias resulting from an uneven distribution of sequences, those with a sequence identity exceeding 99.65% were eliminated from each group. This filtering process was accomplished using the ProDy python package [150]. Next, a single consensus sequence was estimated for each year employing the Bio python package [151], employing a consensus threshold of 50%. However, for years in which only two sequences remained after the refinement process, both sequences were retained for subsequent analyses, as it was not possible to generate a reliable consensus with only two data points. To create a comprehensive set of consensus sequences, a multi-step approach was implemented, encompassing multiple layers of consensus calculations and diverse sequence-clustering strategies. The design of these sequences was guided by the following methods.

Consensus based on time window clustering

The per-year consensus sequences were utilized to generate two distinct subsets of consensus designs. The first set of consensus sequences encompassed all available sequences spanning the period from 1968 to 2023. The second set, however, focused specifically on sequences from 2015 to 2023. This particular time frame was selected due to the substantial antigenic drift experienced by H3N2 during the 2014-2015 season, resulting in the emergence of a new cluster [94]. Within each subset, multiple unique consensus sequences were generated subsequent to sequence alignment and filtering based on a <99.65% sequence identity criterion.

Consensus based on phylogenetic tree clustering

To enhance the diversity of the consensus sequence set, three distinct sequence-clustering algorithms, namely neighbor-joining [152], UPGMA (Unweighted Pair Group Method with Arithmetic Mean) [153], and maximum likelihood [154], were employed to cluster the per-year consensus sequences. Neighbor-joining and UPGMA clustering were conducted using the MUSCLE tool [149], while maximum likelihood clustering was performed using RAxML 8.2 [155]. Visual representation of each clustering tree was accomplished using iTOL [156], and multiple iterations of consensus calculations were applied to each branch of the tree until obtaining individual sequences. The selection of subgroups for consensus calculations was performed manually, based on the visual inspection of the phylogenetic trees. The specific methods employed to obtain single sequences from different trees are described below.

Phylogenetic-based clustering was conducted on three distinct sets of sequences. The first set comprised full-length per-year consensus sequences, while the second set consisted of truncated per-year consensus sequences, containing only the head region of the HA protein (known as the immunodominant region of HA proteins). The final set involved per-year consensus sequences, encompassing solely the stem region of the HA protein. Prior to clustering, each set of sequences were aligned using MUSCLE [34], and duplicate sequences were removed from consideration. For the RaxML method, the full-length or protein head or protein stem alignments underwent refinement using Gblocks 0.91b [157], which facilitated the removal of extensive gap regions that could potentially disrupt group distribution.

The protein head sequence was defined based on the crystal structure of A/Hong Kong/1/1968 (H3N2) (PDB 4fnk). Specifically, it encompassed the region between the sequences "VQSS" and "GSIP" (residues 43-289, corresponding to PDB 4fnk chain E numbering). On the

other hand, the stem sequence was defined as the region between the sequences "GLFG" and "NNRF" (residues 1-171, corresponding to PDB 4fnk chain F numbering).

Consensus based on H3N2 recommended vaccine strains

To capture the characteristics of H3N2 strains recommended by the World Health Organization (WHO) as vaccine candidates between 1968 and 2023, 46 consensus sequences were calculated. The initial consensus sequence was estimated using all available vaccine strain sequences obtained from the online database at https://www.bv-brc.org/ [158] and GISAID (https://gisaid.org/) [147]. Unlike the previous consensus calculations, which employed a 50% threshold for defining consensus residues, this particular consensus sequence was calculated using a more stringent 45% threshold to yield a more distinct sequence representation.

Considering that the recommended vaccine strains did not cover years before 1974, a historical HA sequence, namely A/Hong Kong/1/1968 (H3N2), was included in the vaccine strains alignment to ensure representation of that time period. Furthermore, another consensus was derived exclusively from vaccine strains that demonstrated protective efficacy for more than four years. These strains included A/Bangkok/01/1979, A/Philippines/2/1982, A/Moscow/10/1999, and A/Panama/2007/1999. The dataset containing the long lasting vaccine strains along with the historical sequences was utilized to calculate the second set of consensus sequences and the third set of consensus sequences was derived from only long lasting vaccine strains.

Consensus based on H3N2 antigenic clusters

Smith's research [90] demonstrated that from the time of the introduction of the A/H3N2 virus in humans in 1968 until 2003, there were eleven identifiable clusters of viruses exhibiting

unique antigenic properties [90]. Subsequently, Fonville et al. [92] conducted a study revealing that viruses belonging to subsequent genetic clades, namely CAL04, WI05, BR07, and PE09, also formed distinct antigenic clusters [92]. Furthermore, in addition to these findings, there are three distinct genetic clades known as 3C.1, 3C.2A, and 3C.3A which showed dominance for the last decades [93, 96]. To capture the representative characteristics of these 18 clusters, we selected a representative sequence from each cluster and generated consensus sequence for these 18 representative sequences.

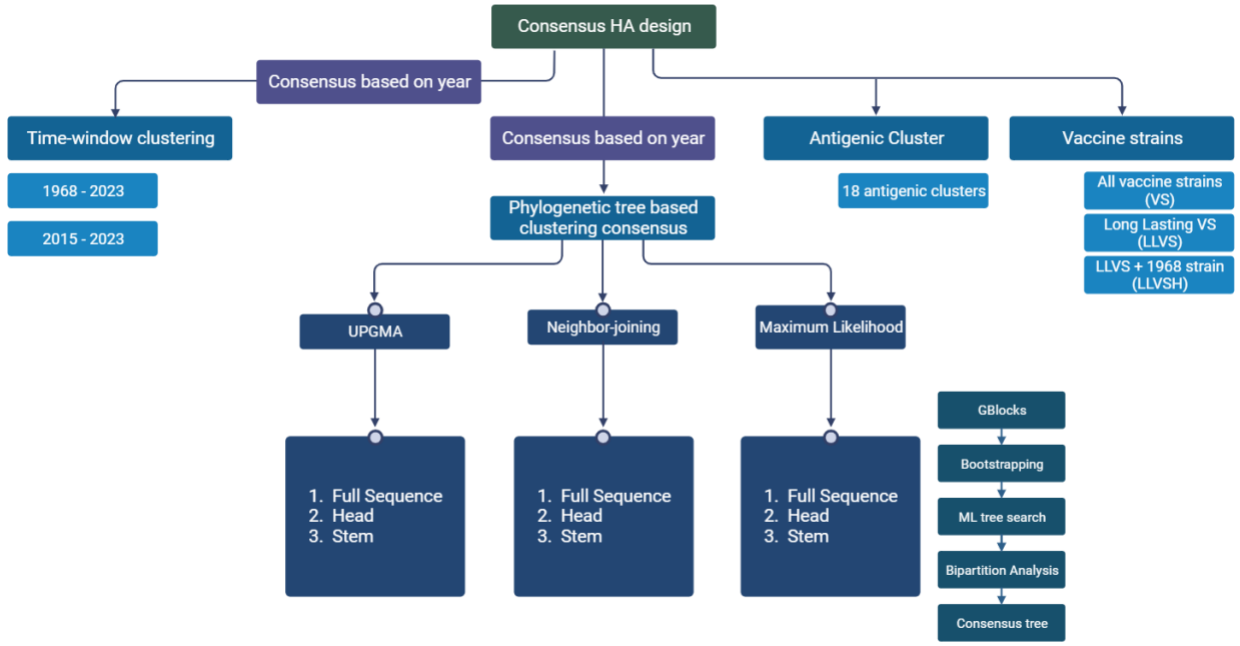

Figure 3.1: Consensus vaccine design strategies that was incorporated in the study

Strategies to address consensus ambiguity

Consensus sequences derived from small datasets often exhibit ambiguity, where multiple amino acids have equal frequencies at a given position. To resolve this ambiguity, we employed both sequence-based and structure-based approaches. In the sequence-based strategy, undefined positions were replaced with the corresponding residue from a vaccine strain sequence that

demonstrated protection for more than four years (specifically A/Bangkok/01/1979, A/Philippines/2/1982, A/Moscow/10/1999, and A/Panama/2007/1999), or by the corresponding residue from the consensus sequences obtained using all recommended H3N2 vaccine strains (as mentioned in the "Consensus based on H3N2 Vaccine Strains" section), and corresponding residue from the consensus sequences obtained using H3N2 representative cluster sequences (as mentioned in the "Consensus based on H3N2 Antigenic Clusters" section) and WHO recommended last vaccine strains (for the year 2022-2023) A/Darwin/9/2021. For the undefined positions we got for both Consensus based on H3N2 Vaccine Strains and Consensus based on H3N2 Antigenic Clusters, we replaced those positions with the A/Michigan/15/2014, (PDB 6bkp) the latest available crystal structure of a vaccine strain and the most prevalent clade currently circulating.

In a similar vein, the structure-based approach encompassed the threading of all designed sequences onto three distinctive structural templates. This process enabled the substitution of ambiguous regions with the corresponding amino acids found in the template structures. The utilized templates details are given in Table 3.1.

Table 3.1. Different template information used to address consensus ambiguity

Template Number	Template name	Detail of the template
1	A/Bangkok/01/1979	Long-lasting vaccine strain (LLVS)
2	A/Philippines/2/1982	Long-lasting vaccine strain (LLVS)
3	A/Moscow/10/1999	Long-lasting vaccine strain (LLVS)
4	A/Panama/2007/1999	Long-lasting vaccine strain (LLVS)

5	Consensus0.45_VS_a_refined1_combined_with_6bkp	Consensus based on vaccine strains where the ambiguous positions (x) were replaced with the corresponding amino acids from A/Michigan/15/2014
6	Consensus0.0_cluster_refined1_combined_with_6bkp	Consensus based on antigenic clusters where the ambiguous positions (x) were replaced with the corresponding amino acids from A/Michigan/15/2014
7	4fnk_chainEF	Sequence of Chain E and Chain F of A/Hong Kong/1/1968 (H3N2) (PDB 4fnk)
8	4we9_VC11	Sequence of Chain E and Chain F of A/Hong Kong/1/1968 (H3N2) (PDB 4fnk)
9	6bkp_last_VS_structure_last_dominant_cluster3c2a	Sequence of A/Michigan/15/2014, the latest available crystal structure of a vaccine strain, and the most prevalent clade (3c2a) currently circulating
10	A/Darwin/9/2021_last_vaccineStrains	This is the last vaccine strain approved by WHO for the year 2022-2023

After replacing the ambiguous positions with the corresponding templates, duplicated sequences were removed using the rmdup tool from seqkit [148]. This approach was also employed to complete sequences that solely contained the HA head or stem regions, as mentioned in the "Consensus based on phylogenetic trees clustering" section. Consequently, certain designed

proteins exhibited a consensus sequence in the head region, while the remaining portion of the protein corresponded to a long-lasting vaccine strain.

Filtering the designs

To finalize and filter the designs, we focused on the sequence diversity specifically in the seven critical antigenic positions near the receptor binding site. Based on the clustering methods and datasets used, the sets of sequences obtained are named as follows:

1. Rax Entire: The entire consensus sequence derived from maximum likelihood clustering performed using the RAxML tool.
2. Rax Head: The consensus sequence of the head region obtained from maximum likelihood clustering performed using the RAxML tool, with the remaining sequence derived from templates.
3. Rax Stem: The consensus sequence of the stem region obtained from maximum likelihood clustering performed using the RAxML tool, with the rest of the sequence derived from templates.
4. UPGMA Entire: The entire consensus sequence obtained from UPGMA clustering, with the rest of the sequence derived from templates.
5. UPGMA Head: The consensus sequence of the head region obtained from UPGMA clustering, with the remaining sequence derived from templates.
6. UPGMA Stem: The consensus sequence of the stem region obtained from UPGMA clustering, with the rest of the sequence derived from templates.
7. NJ Entire: The entire consensus sequence obtained from neighbor-joining clustering, with the rest of the sequence derived from templates.

8. NJ Head: The consensus sequence of the head region obtained from neighbor-joining clustering, with the remaining sequence derived from templates.

9. NJ Stem: The consensus sequence of the stem region obtained from neighbor-joining clustering, with the rest of the sequence derived from templates.

10. Vaccine Strains All: Consensus sequence based on H3N2 recommended vaccine strains.

11. Vaccine Strains LLVS: Consensus sequence based on H3N2 long-lasting vaccine strains.

12. Vaccine Strains LLVSH: Consensus sequence based on H3N2 long-lasting vaccine strains, including the historic A/Hong Kong/1/1968 sequence.

13. Antigenic Cluster: Consensus sequence based on representative sequences from H3N2 antigenic clusters.

14. Time Window All Year: Consensus sequences encompassing all available sequences from 1968 to 2023.

15. Time Window After 2015: Consensus sequences encompassing the time frame from 2015 to May 2023.

These sets of sequences were carefully filtered and selected based on the specific clustering methods and datasets used, with a focus on the diversity within the crucial antigenic positions. In each case, the objective was to obtain 10 sequences for 10 templates.

CHAPTER 4

RESULTS AND DISCUSSION

We successfully acquired a comprehensive dataset for our study by retrieving human HA (H3N2) amino acid sequences from the GISAID database [147], spanning the years from 1968 to May 2023. Overcoming a download limitation of 20,000 sequences at a time, we employed a strategic approach based on isolation year to obtain a total of 114,090 sequences, ensuring that only complete HA sequences were included. To enhance data quality, we employed several preprocessing steps, including duplicate removal and the development of an in-house Python script to optimize sequence headers. This script replaced the original sequence names with concise and informative identifiers containing strain names and isolation years while eliminating sequences with incomplete or inadequate header information. Following these rigorous procedures, we established a refined dataset comprising 29,238 unique sequences as the primary database for our subsequent analyses. Recognizing the limitations of the MUSCLE alignment tool [149] with datasets exceeding 10,000 sequences, we further partitioned the database into four subsets using an additional Python script, facilitating efficient data processing for our research. These preprocessing steps laid the foundation for our subsequent analyses in the results section.

Yearly consensus sequence generation

To remove sample bias, we generated a single consensus sequence for each year from the year 1968 to 2023. To do that, we initially divided the sequences based on their isolation year and then performed alignment. To address sequence distribution bias, sequences with a sequence

identity exceeding 99.65% were removed from each group a single consensus sequence for each year was generated applying a 50% consensus threshold. For all the year we got single consensus representative sequences except for the year 1970, 1981 and 1984. For the year 1970, there was only a single sequence after the filtering process mentioned above and we kept that sequence as a representative for 1970. However, for the year 1981 and 1984 two sequences remained after the refinement and both were retained due to the insufficient data points for reliable consensus generation.

Consensus based on time window clustering

After getting the representative sequences for each year, we generated two distinct sets of consensus sequences: one encompassing all available sequences from 1968 to 2023 and another focusing specifically on sequences from 2015 to 2023, given the significant antigenic drift in H3N2 during the 2014-2015 season [159]. Within each subset, we produced unique consensus sequences following alignment and filtering based on a <99.65% sequence identity criterion.

When deriving consensus sequences encompassing all available sequences spanning the period from 1968 to 2023, we found 11 x positions, with multiple positions falling within the seven critical antigenic positions. When there was no clear consensus amino acid at a particular position within a sequence alignment, the letter "x" was used as a placeholder to indicate the ambiguity or lack of a definitive amino acid assignment for that position. To overcome ambiguity, we used 10 templates as described in the method section “Strategies to Address Consensus Ambiguity”. As the A/Philippines/2/1982 didn’t have the entire HA sequence information, this template was excluded from the design for this case. Thus, we initially obtained 9 designs, and after removing

duplication, we were left with a total of 7 designs for consensus spanning the period from 1968 to 2023.

In the scenario of consensus sequences encompassing the time frame between 2015 and May 2023, we encountered 1 “x” position located within the signal peptide region, but not among the seven critical antigenic positions. We replaced this position with the template 4fnk mentioned in the method section “Strategies to Address Consensus Ambiguity”.

Consensus based on phylogenetic tree clustering

To diversify the consensus sequence set, we employed different distinct sequence-clustering algorithms as mentioned in the method section. These clustering methods were applied to per-year consensus sequences and visual representation of clustering trees was achieved using iTOL [156]. Multiple iterations of consensus calculations were conducted for each tree branch until individual sequences were obtained. Subgroup selection for consensus calculations was guided by manual examination of the phylogenetic trees. Phylogenetic-based clustering was performed on three distinct sets of sequences: full-length per-year consensus sequences, truncated per-year consensus sequences containing only the head region of the HA protein (known as the immunodominant region), and per-year consensus sequences encompassing solely the stem region of the HA protein. The protein head sequence was defined based on the crystal structure of A/Hong Kong/1/1968 (H3N2), spanning residues 43-289, while the stem sequence encompassed residues 1-171, both according to PDB 4fnk chain numbering. Multiple consensus sequences were generated from three which are described below:

UPGMA entire

In the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) tree, encompassing the entire HA sequence, a noteworthy clustering pattern emerged among consensus sequences spanning different years. Specifically, from 1979 to 2002, these sequences were grouped under a common monophyletic group, exhibiting multiple clades and subclades, mirroring a similar arrangement observed for other sequences. In contrast, the years from 1969 to 1978 were clustered under the same common clade, as were the years from 2003 to 2021. Notably, the years 2023 and 2022 formed a distinct clade within the dataset (Figure 4.1).

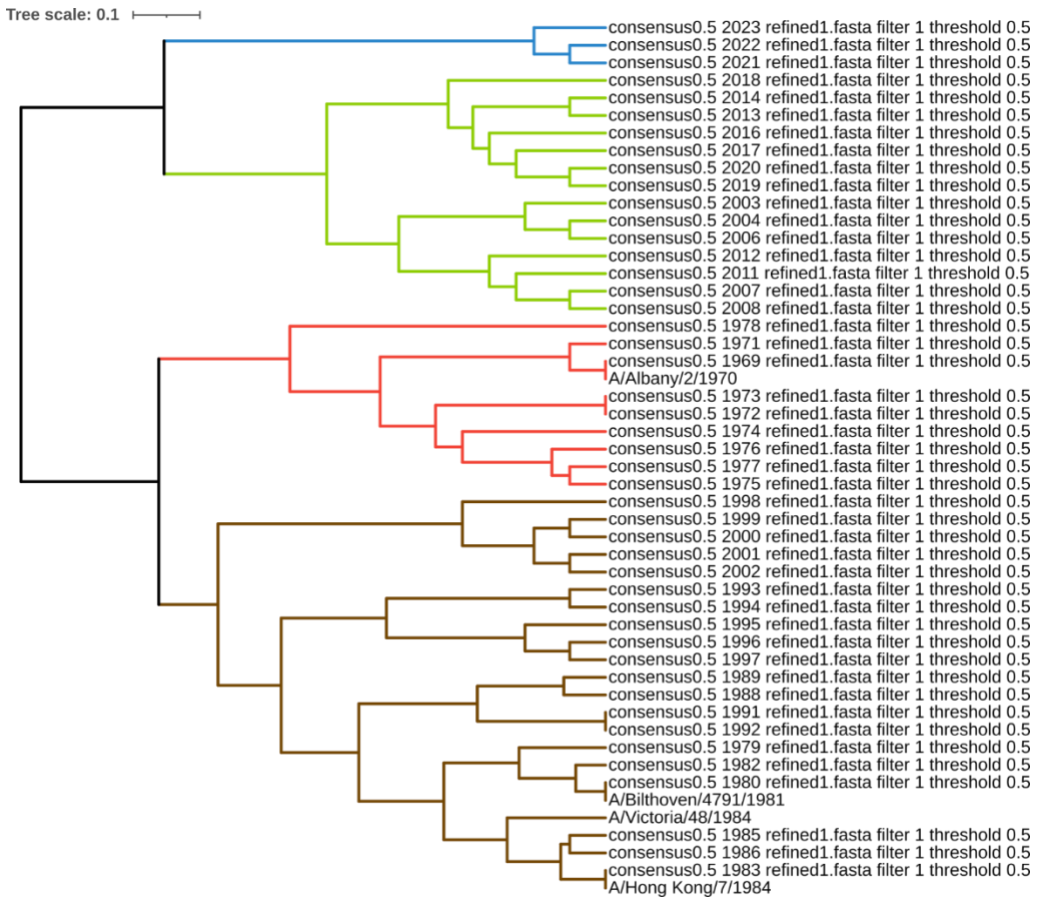

Figure 4.1. Consensus phylogenetic tree for HA entire sequences based on UPGMA. Similar color represents the sequences which fall under the same clade information.

To consolidate these observations into single consensus sequences, we created one consensus sequence for the period spanning 1979 to 2002, another for the years 1969 to 1978, a third for 2003 to 2023, and a fourth for the years 2022 and 2023, which shared the same common clade. Subsequently, we generated a final consensus sequence from this set of four newly generated consensus sequences. It's worth noting that in each step of our process, we conducted sequence alignment for all the sequences within the respective set before proceeding to generate the single consensus sequence from that set. This alignment step was a crucial part of ensuring the accuracy and reliability of our consensus sequence generation process. For this case, one position remained ambiguous and did not correspond to any of the seven critical antigenic positions. After eliminating duplicate sequences, only one sequence remained, where the "x" ambiguous position was replaced with the template sequence 4fnk, as detailed in the methodology section (Strategies to Address Consensus Ambiguity).

UPGMA Head

In our analysis of the phylogenetic tree encompassing the consensus UPGMA head sequences spanning the years 1968 to 2023, we noted distinctions compared to the tree generated using UPGMA entire sequences. In this particular tree, sequences from 1969 to 1992 exhibited a shared clade. However, this set of sequences further diverged into two distinct sub-clades: one encompassing sequences from 1969 to 1978 and another comprising the consensus sequences spanning from 1979 to 1992. Moreover, these two sub-clades were nested within a parental clade, which, in turn, included another set of consensus sequences ranging from 1993 to 2002. Ultimately, all three of these sets were part of a broader parental clade housing a unique collection of consensus sequences spanning from 2003 to 2023 (Figure 4.2).

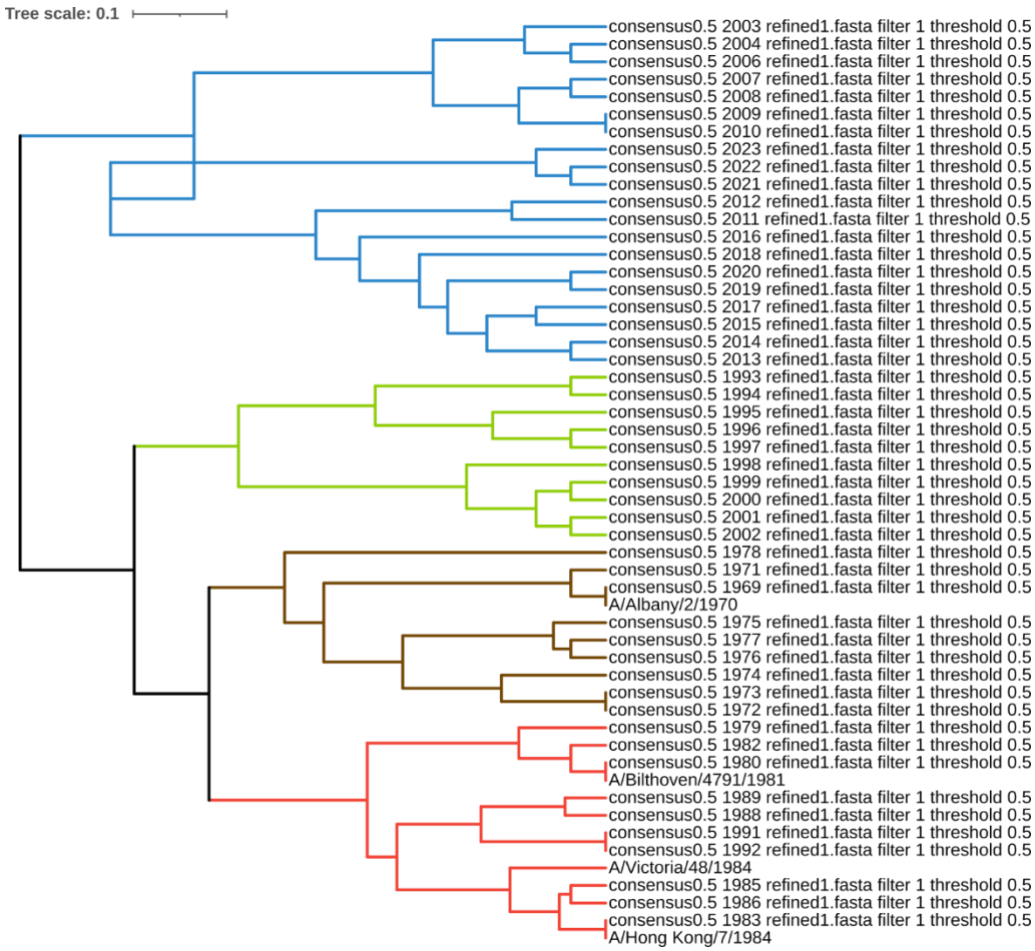

Figure 4.2. Consensus phylogenetic tree for HA head sequences based on UPGMA. Similar color represents the sequences which fall under the same clade information.

We adopted a stepwise approach to generate the final consensus sequence for the head region. Initially, we created a consensus sequence for the set of sequences belonging to the node encompassing the years 1969 to 1978. This consensus sequence was then incorporated into the second set, comprising consensus sequences spanning from 1979 to 1992. Subsequently, we derived a single consensus sequence from this combined set. This process was iterated as the resulting consensus sequence was included in the set of sequences ranging from 1993 to 2002 to generate another consensus sequence. Finally, the consensus sequence from the previous step was

included in the set of consensus sequences spanning from 2003 to 2023 to yield the ultimate single consensus sequence for the head region. Similar to the UPGMA entire, in each step of our process, we conducted sequence alignment for all the sequences within the respective set before proceeding to generate the single consensus sequence from that set.

In the final consensus sequence generated through this procedure, we encountered one ambiguous position denoted as "x," which happened to fall within the critical antigenic positions. To address this ambiguity, we replaced this position with the templates specified in the methodology section. Additionally, for the stem region, sequence information was sourced from different templates. Following a similar strategy, the incomplete sequence of A/Philippines/2/1982 resulted in its exclusion from the design process. Consequently, in this context, we arrived at a total of 9 designs, with the head representing the consensus obtained from the UPGMA tree, and the remaining sequences derived from the specified templates as described in the methodology section (excluding A/Philippines/2/1982).

UPGMA Stem

In our analysis of the UPGMA stem region, we encountered an intriguing tree structure, with the majority of sequences being filtered out due to the <99.65% sequence identity criterion. Notably, for this specific case, the years 1981, 1975, 1972, 1984, 2002, 1986, 1993, and 1998 clustered together within a single node, forming what we refer to as set 1 (Figure 4.3). Similarly, the years 1982, 1970, 2003, 2006, and 2007 clustered as a cohesive clade, constituting set 2. Lastly, the years 1994, 2008, 2018, 2011, and 2021 formed a distinct cluster, representing set 3. To generate the final consensus sequence for the stem region, we employed a stepwise approach akin to the one used for the head region. Initially, we created a consensus sequence for the sequences

within set 1. This consensus sequence was then merged into the second set, comprising the consensus sequences for 1982, 1970, 2003, 2006, and 2007. Subsequently, we derived a single consensus sequence from this combined set. We repeated this process as the resulting consensus sequence was included in the set 3 sequences, ultimately culminating in the generation of the definitive single consensus sequence for the stem region. As with our previous method, in each step of this process, we performed sequence alignment for all the sequences within the respective set before proceeding to generate the single consensus sequence from that set, ensuring the accuracy of our results.

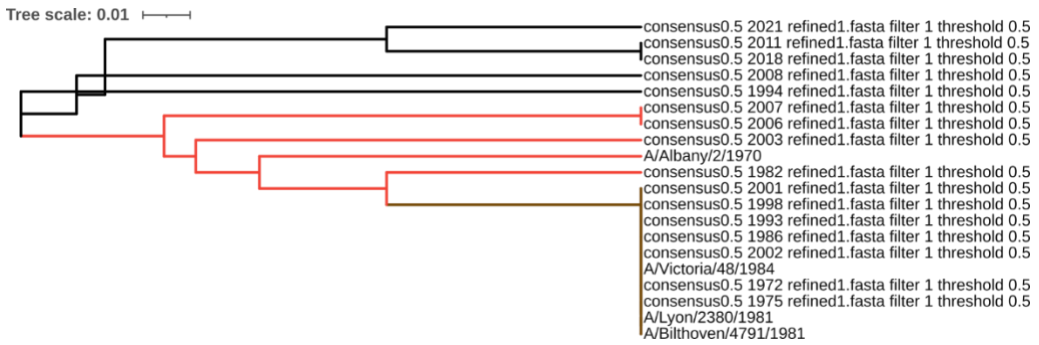

Figure 4.3. Consensus phylogenetic tree for HA stem sequences based on UPGMA. Similar color represents the sequences which fall under the same clade information.

In the UPGMA Stem analysis, it's noteworthy that we did not encounter any ambiguous "x" positions within the stem region. Additionally, we opted to exclude the long-lasting vaccine strains from the template designs due to their tendency to produce inaccurate alignments in MUSCLE. Consequently, we derived a total of six designs for this case. In these designs, the stem region was represented by the consensus obtained from the UPGMA tree, while the remaining sequences were derived from the specified templates as outlined in the methodology section, with the exception of the long-lasting vaccine strains. This approach ensured the accuracy and reliability of our results while providing valuable insights into the stem region of the H3N2 virus.

NJ Entire

In our analysis, we observed differences between the phylogenetic tree obtained from the neighbor-joining (NJ) method based on the entire HA sequence and the tree obtained from UPGMA. In the NJ tree, a distinct clustering pattern emerged: the consensus sequences for the years 1984 to 2023 formed a shared clade, characterized by multiple sub-clades and sub-sub-clades, constituting set 1. In contrast, the sequences spanning from 1968 to 1983 formed a separate and distinct clade, referred to as set 2. Notably, A/Hong Kong/7/1984 stood out as it formed an isolated clade of its own within the tree (Figure 4.4).

To generate the final consensus sequence for NJ Entire, we adopted a systematic approach. Initially, we produced a single consensus sequence for all the sequences within set 1 after performing sequence alignment. Subsequently, we generated another single consensus sequence for all the sequences within set 2. These two consensus sequences were then combined with A/Hong Kong/7/1984, forming set 3. Following the alignment of these three sequences, we generated the ultimate single consensus tree for the consensus NJ entire.

However, it's important to note that in the NJ Entire analysis, we encountered 11 ambiguous "x" positions, some of which were situated within the seven critical antigenic positions. In line with our previous approach for UPGMA, we excluded the template A/Philippines/2/1982 from the design process. As a result, we obtained a total of nine designs in this case, each providing valuable insights into the HA sequence characteristics of H3N2 over time.

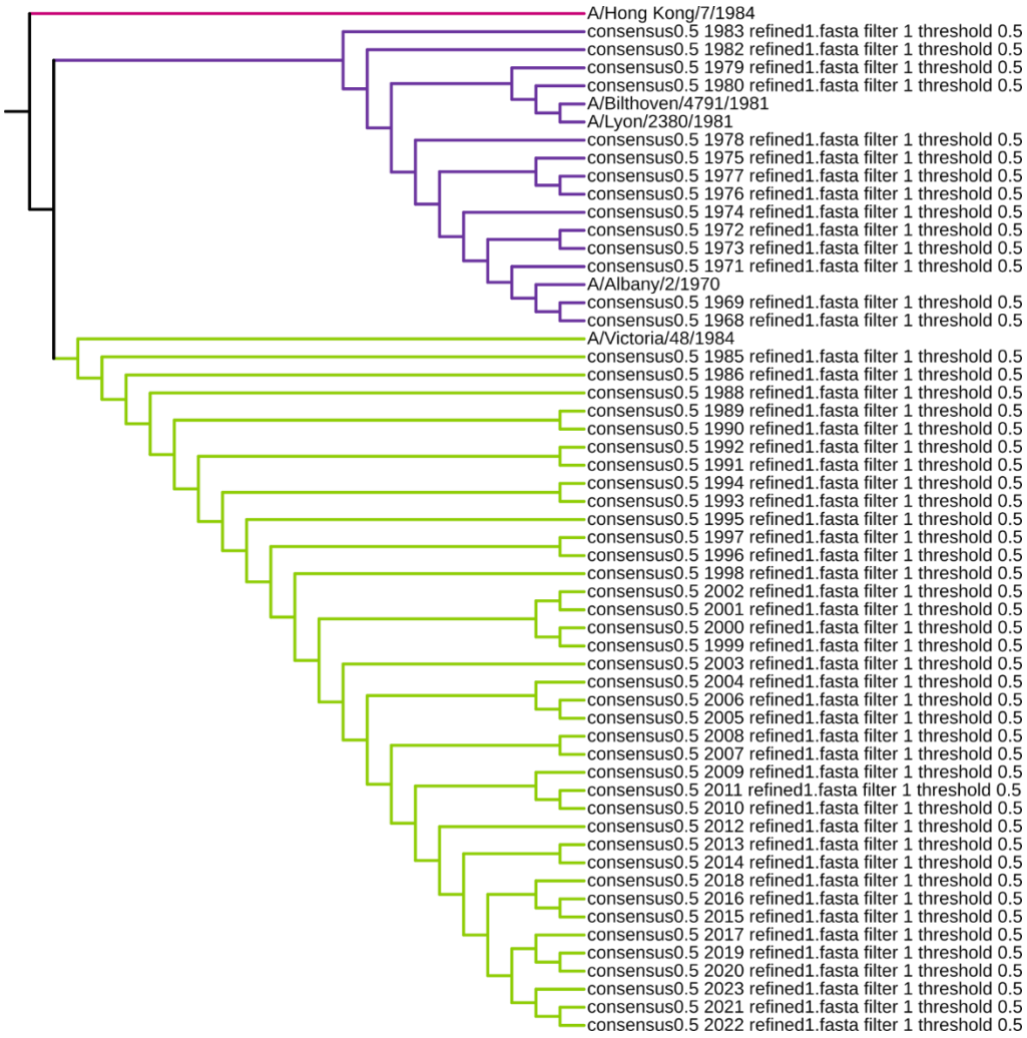

Figure 4.4. Consensus phylogenetic tree for HA entire sequences based on neighbor-joining. Similar color represents the sequences which fall under the same clade information.

NJ Head

In our examination of the phylogenetic tree generated by the neighbor-joining (NJ) method for the head sequences, we observed a nuanced tree structure compared to the NJ tree obtained for entire sequences. Notably, the consensus sequences from the years 1979 to 1992 shared a common clade, denoted as set 1, while the sequences spanning from 1968 to 1978 formed another distinct clade, referred to as set 2 (Figure 4.5). Both of these clades were nested within a larger parental

clade. Additionally, the sequences from 1993 to 2014 consistently followed a similar pattern, clustering into set 3. Intriguingly, the consensus sequences from the years 2018, 2016, and 2015 fell under the same clade, designated as set 5, while the years 2015 to 2023, excluding 2015, 2016, and 2018, formed another shared clade known as set 4.

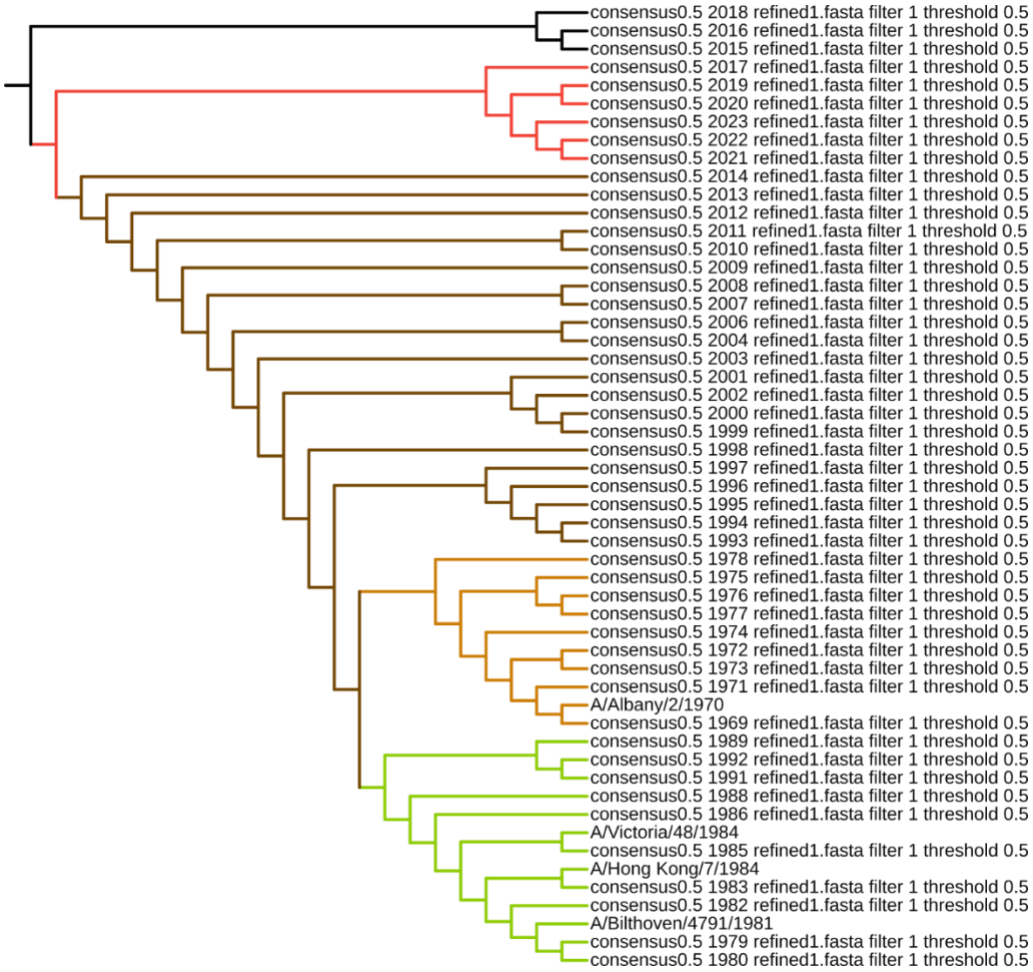

Figure 4.5. Consensus phylogenetic tree for HA head sequences based on neighbor-joining. Similar color represents the sequences which fall under the same clade information.

To generate the final consensus sequence for NJ Head, we employed a systematic approach. Initially, we computed a single consensus sequence for all the sequences from the years 1979 to 1992, categorized as set 1. We then generated another single consensus sequence for the

years 1968 to 1978 after conducting sequence alignment, constituting set 2. Both of these consensus sequences, obtained from set 1 and set 2, were included in the set of sequences spanning from 1993 to 2014, resulting in set 3. Subsequently, we generated a single consensus for set 3 after sequence alignment. This consensus sequence was subsequently incorporated into set 4, where we performed sequence alignment to derive the single consensus sequence for set 4. Finally, this sequence was integrated into set 5 to yield the ultimate single consensus sequence for NJ Head.

In the context of NJ Head, we did encounter an "x" position, which, notably, did not fall within the seven critical antigenic positions. Similar to our previous methodology, we excluded the incomplete sequence of A/Philippines/2/1982 from the design process. Consequently, we arrived at a total of nine designs in this case, with the head region represented by the consensus obtained from the neighbor-joining tree mentioned above, while the remaining sequences were derived from the specified templates as detailed in the methodology section, excluding A/Philippines/2/1982.

NJ Stem

In our analysis of the NJ stem region, we encountered a distinctive tree structure, where the majority of sequences were filtered out due to the stringent <99.65% sequence identity criterion, mirroring our observations in the UPGMA stem analysis. This outcome was expected, given the high conservation of the stem part of the HA protein. Specifically, for this case, a subset of years, including 2018, 2021, 2011, 2008, 2007, 2006, 2003, 2001, 2002, 1998, 1993, 1994, and 1984, clustered together within a single node, forming what we denoted as set 1. Similarly, the years 1972, 1970, 1975, and 1981 formed a cohesive clade, representing set 2. Notably, in the NJ stem tree, the year 1986 stood out as it formed a distinct cluster (Figure 4.6).

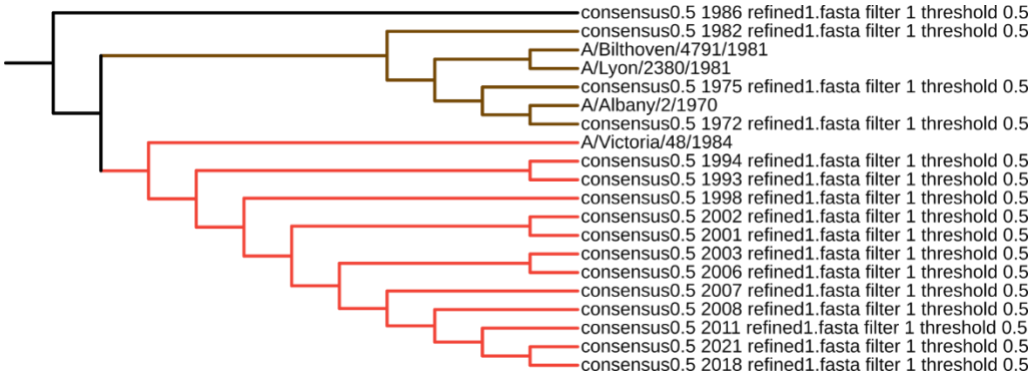

Figure 4.6. Consensus phylogenetic tree for HA stem sequences based on neighbor-joining. Similar color represents the sequences which fall under the same clade information.

To generate the final consensus sequence for the stem region in the NJ analysis, we applied a method analogous to that used for the UPGMA stem. Initially, we created a consensus sequence for the sequences within set 1. This consensus sequence was subsequently integrated into the second set, comprising the consensus sequences for 1972, 1970, 1975, and 1981. We then derived a single consensus sequence from this combined set. This process was iterated as the resulting consensus sequence was included in the set 3 sequences, ultimately culminating in the generation of the definitive single consensus sequence for the stem region. Similar to our previous method, in each step of this process, we conducted sequence alignment for all the sequences within the respective set before proceeding to generate the single consensus sequence from that set, ensuring the accuracy of our results.

It's noteworthy that, in the NJ Stem analysis, we did not encounter any ambiguous "x" positions within the stem region. Similarly to the approach taken in the UPGMA Stem analysis, we excluded all long-lasting vaccine strains from the template designs. Consequently, we obtained a total of six designs for this case, with the stem region represented by the consensus obtained

from the neighbor-joining tree, while the remaining sequences were derived from the templates mentioned in the methodology section, ensuring the reliability and accuracy of our results.

RaxML Entire

In our analysis of the RaxML tree, which encompassed the entire HA sequence, we observed a distinct clustering pattern among consensus sequences spanning different years, which markedly differed from the UPGMA and NJ trees. The tree structure revealed a notable grouping of sequences from 2017 to 2023, excluding 2018, forming a common monophyletic group with multiple clades and subclades, effectively constituting set 1. The consensus sequences from 2015, 2016, and 2018 were grouped together as set 2. Similarly, the years 2010 to 2014 comprised set 3, while the years 2001 to 2009 formed set 4. Set 5 consisted of sequences from 1991 to 2000, and set 6 encompassed sequences from 1978 to 1990. The consensus sequences from 1973 to 1977 constituted set 7, and the years 1968 to 1972 represented the final set 8 (Figure 4.7).

To generate a single consensus sequence from this tree, we followed a systematic approach. Initially, we created a consensus for set 1 and then incorporated it into set 2 to generate another single consensus, as mentioned previously. After obtaining a single consensus sequence, it was included in set 3, and we generated another consensus for those sequences. We repeated this process iteratively until we reached set 8, ultimately resulting in a single consensus sequence that encompassed all the years. Throughout each step of this process, sequences were aligned as described earlier, ensuring the accuracy and reliability of our consensus sequence generation.

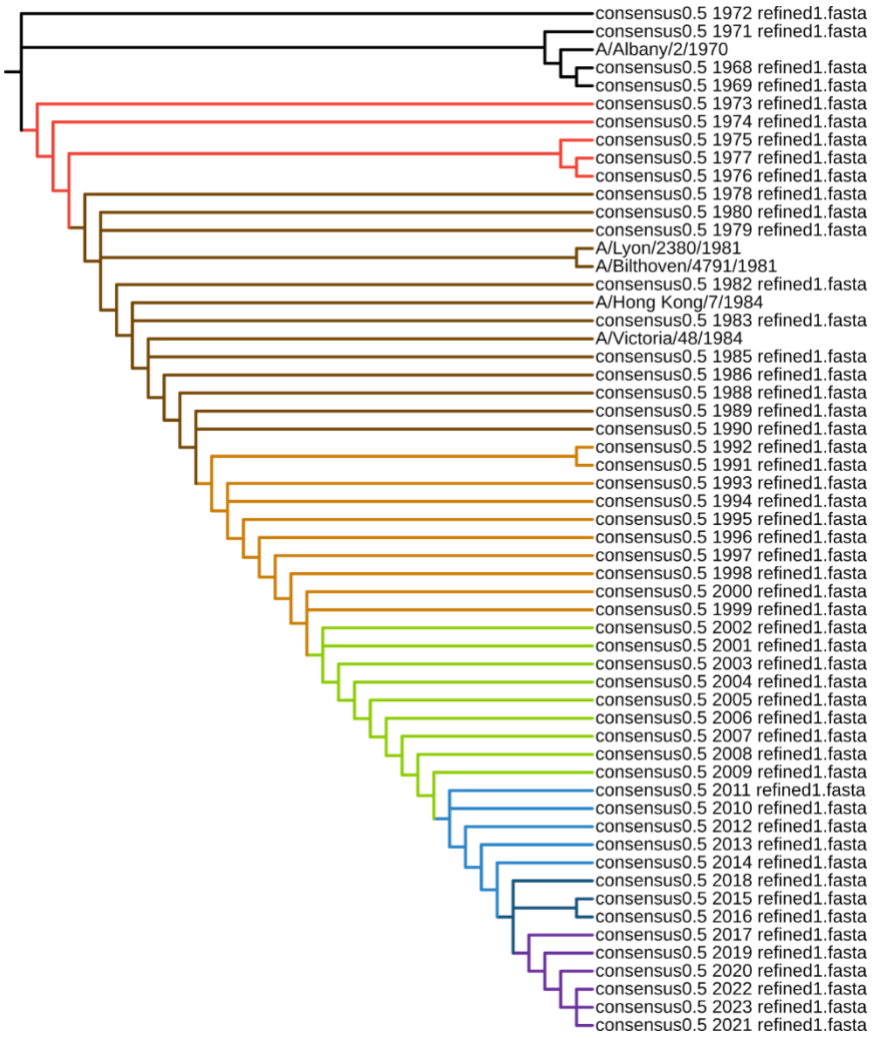

Figure 4.7. Consensus phylogenetic tree for HA entire sequences based on maximum likelihood. Similar color represents the sequences which fall under the same clade information.

Regarding Rax Entire, we identified 5 "x" positions, none of which were located within the seven critical antigenic positions. After eliminating duplicate sequences, only 2 sequences remained, with only one amino acid difference between them, and none of these differences were within the seven critical antigenic positions. Consequently, the sequence was combined with the template 4fnk, and this composite sequence was ultimately selected and retained for further analysis, ensuring the consistency and reliability of our results.

RaxML Head

In our analysis of the RaxML head, we observed a pattern largely similar to that of RaxML entire, with the exception of a few sequences being filtered out due to the stringent <99.65% sequence identity criterion. The tree structure revealed the grouping of sequences from 2017 to 2023, excluding 2018, forming a common monophyletic group with multiple clades and subclades (Figure 4.8), effectively constituting set 1, mirroring the findings in RaxML entire (Figure 4.7). The consensus sequences from 2015, 2016, and 2018 were grouped together as set 2. Similarly, the years 2010 to 2014 comprised set 3, while the years 2001 to 2009 formed set 4. Set 5 consisted of sequences from 1991 to 2000, and set 6 encompassed sequences from 1978 to 1989, with the sequence from 1990 being filtered out due to the <99.65% sequence identity criterion. The consensus sequences from 1973 to 1977 constituted set 7, and the years 1969 to 1972 represented the final set 8, as the consensus sequence from 1968 was filtered out for the same criterion.

To generate a single consensus sequence from this tree, we followed a similar approach as in RaxML entire. Initially, we created a consensus for set 1 and then incorporated it into set 2 to generate another single consensus, as mentioned previously. After obtaining a single consensus sequence, it was included in set 3, and we generated another consensus for those sequences. We repeated this process iteratively until we reached set 8, ultimately resulting in a single consensus sequence that encompassed all the years. Throughout each step of this process, sequences were aligned as described earlier, ensuring the accuracy and reliability of our consensus sequence generation.

Regarding Rax Head, there were 2 x positions, and none of these were in the seven critical antigenic positions. As the sequence of A/Philippines/2/1982 was incomplete, it was excluded from the design. Consequently, there were 9 designs in this case, where the head represented the

consensus obtained from the maximum likelihood tree, and the remaining sequences were derived from the specified templates as detailed in the methodology section (except for A/Philippines/2/1982).

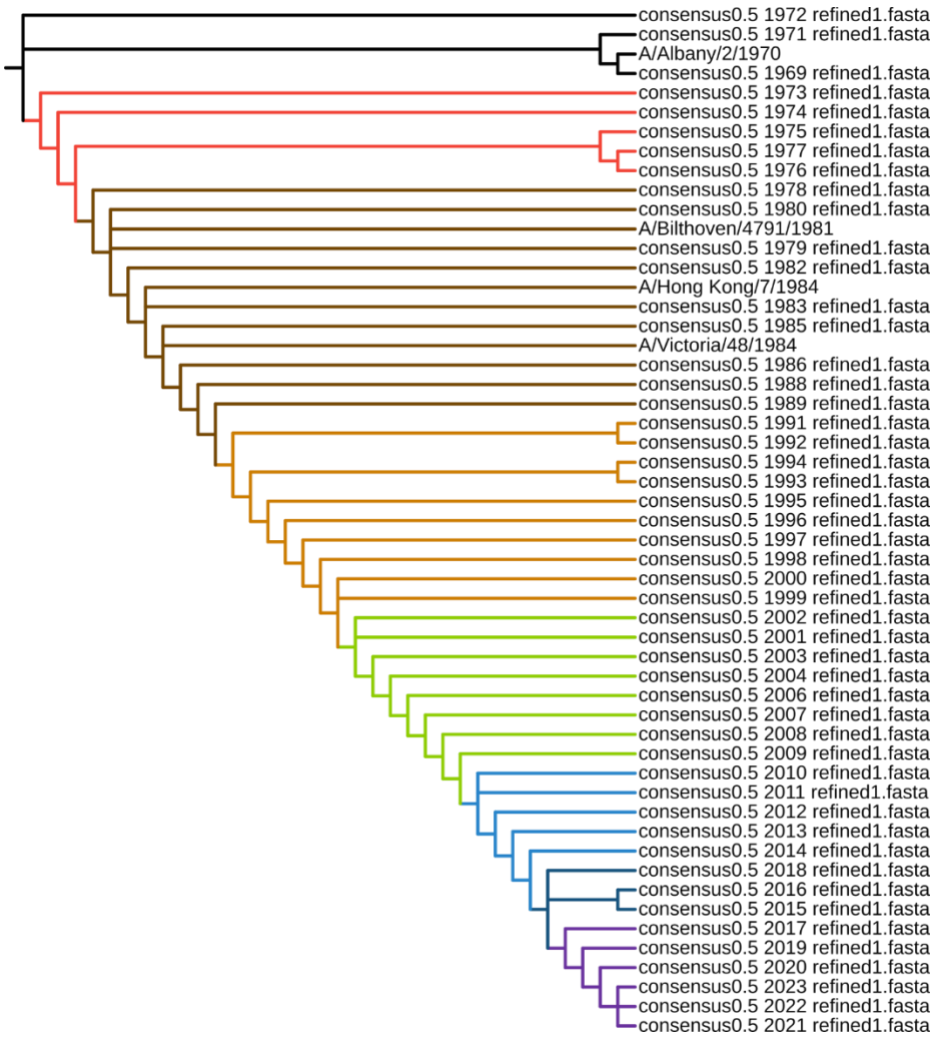

Figure 4.8. Consensus phylogenetic tree for HA head sequences based on maximum likelihood. Similar color represents the sequences which fall under the same clade information.

RaxML Stem

In our analysis of the RaxML stem sequences, we encountered a distinctive tree structure, where the majority of sequences were filtered out due to the stringent <99.65% sequence identity

criterion, mirroring our observations in the UPGMA and NJ stem analyses. This outcome was expected, given the high conservation of the stem part of the HA protein, resulting in less sequence diversity. Specifically, for this case, a subset of years, including 2018, 2021, 2011, 2008, and 2007, clustered together within a single node, forming what we denoted as set 1 (Figure 4.9). Similarly, the years 2006, 2003, 2001, and 2002 formed a cohesive clade, representing set 2. The year 1994, 1998, 1993, 1986, and 1982 were from a common clade, which comprised set 3. Finally, the years 1972, 1970, 1975, and 1981 formed a cohesive clade, representing set 4.

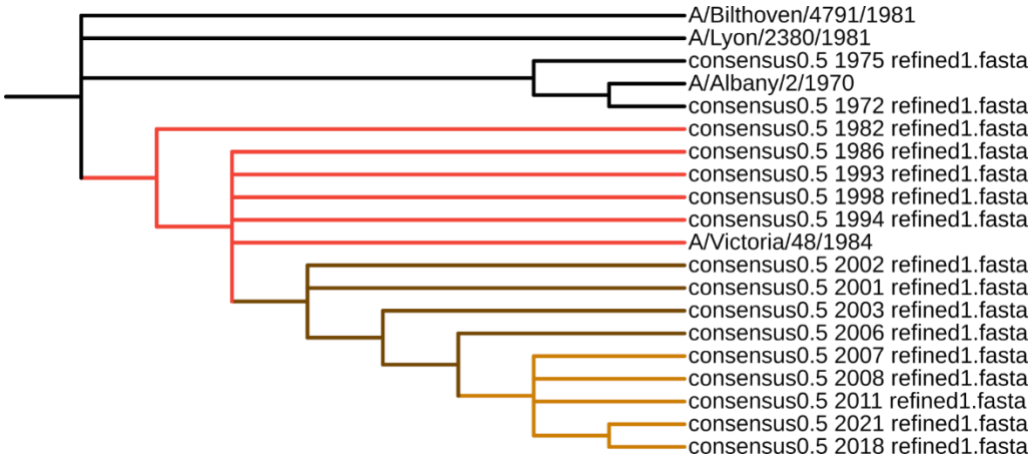

Figure 4.9. Consensus phylogenetic tree for HA stem sequences based on maximum likelihood. Similar color represents the sequences which fall under the same clade information.

To generate the final consensus sequence for the stem region in the RaxML analysis, we applied a method analogous to that used for the UPGMA and NJ stem analyses. Initially, we created a consensus sequence for the sequences within set 1. This consensus sequence was subsequently integrated into the second set, comprising the consensus sequences for 2006, 2003, 2001, and 2002. We then derived a single consensus sequence from this combined set. This process was iterated as the resulting consensus sequence was included in the set 3 sequences, and consequently in the set 4 sequences, ultimately culminating in the generation of the definitive

single consensus sequence for the stem region. Similar to our previous method, in each step of this process, we conducted sequence alignment for all the sequences within the respective set before proceeding to generate the single consensus sequence from that set, ensuring the accuracy of our results.

For Rax Stem, there was a single "x" position in the stem. All long-lasting vaccine strains were removed from the template designs, as including them caused incorrect alignment issues with the muscle algorithm, similar to the UPGMA and NJ stem analyses. As a result, 6 designs were obtained for this case, with the stem representing the consensus obtained from the maximum likelihood tree, and the remaining sequences derived from the template as described for the UPGMA and NJ stem analyses.

Consensus based on H3N2 Recommended Vaccine Strains

In this case, we aimed to capture the distinctive features of H3N2 strains recommended by the World Health Organization (WHO) as vaccine candidates spanning the period from 1968 to 2023. A total of 46 vaccine strain sequences were obtained from the online databases as mentioned in the method section. A notable departure from previous consensus calculations was our use of a stricter 45% threshold to define consensus residues, resulting in a more refined sequence representation. Recognizing that the recommended vaccine strains did not extend to years before 1974, we ensured representation of that earlier period by incorporating a historical HA sequence, specifically A/Hong Kong/1/1968 (H3N2), into the vaccine strains alignment. For the Consensus based on H3N2 recommended vaccine strains, we identified 7 x positions where multiple of those were located in the seven critical antigenic positions. After removing duplicate sequences, we obtained 9 designs, all of which were retained.

Additionally, we generated another consensus sequence exclusively from vaccine strains that demonstrated protective efficacy for more than four years, including A/Bangkok/01/1979, A/Philippines/2/1982, A/Moscow/10/1999, and A/Panama/2007/1999. The dataset encompassing these enduring vaccine strains, along with the historical sequences, allowed us to compute the second set of consensus sequences, while the third set of consensus sequences was exclusively derived from the long-lasting vaccine strains. This comprehensive approach enabled us to capture the evolution and diversity of H3N2 strains recommended for vaccination over the specified timeframe. Regarding the Consensus based on H3N2 long-lasting vaccine strains, we observed 26 x positions, and we successfully obtained 10 designs from 10 templates. In the case of the Consensus based on H3N2 long-lasting vaccine strains and the historic A/Hong Kong/1/1968 sequence, we identified 18 x positions. After removing duplication, we obtained 9 designs. Similar to previous cases, we excluded the A/Philippines/2/1982 template from the design process.

Consensus based on H3N2 Antigenic Clusters

In Smith's research, it was demonstrated that from the introduction of the A/H3N2 virus in humans in 1968 until 2003, there were eleven identifiable clusters of viruses, each exhibiting unique antigenic properties [90]. Subsequently, Fonville et al.'s study revealed that viruses from subsequent genetic clades, specifically CAL04, WI05, BR07, and PE09, also formed distinct antigenic clusters [92]. Additionally, in line with these findings, three distinct genetic clades known as 3C.1, 3C.2A, and 3C.3A have shown dominance in recent decades [93, 96]. To capture the representative characteristics of these 18 clusters, we selected one representative sequence from each cluster based on Bruke’s paper [93] and generated a consensus sequence for these 18 representative sequences. For this case, 4 x positions were generated where multiple of those were

located in the seven critical antigenic positions. We obtained 10 sequences after replacing those with templates. After removing the duplicate sequences, we obtained 7 designs.

Therefore, the total number of designs are 99 which is shown in Table 4.1

Table 4.1: Number of designed sequences for different set of consensus design

Set No	Method	Number of designed sequences
1	Rax Entire	1
2	Rax Head	9
3	Rax Stem	6
4	UPGMA Entire	1
5	UPGMA Head	9
6	UPGMA Stem	6
7	NJ Entire	9
8	NJ Head	9
9	NJ Stem	6
10	Vaccine strains all	9
11	Vaccine strains LLVS	10
12	Vaccine strains LLVSH	9
13	Antigenic cluster	7
14	Time window all year	7
15	Time window after 2015	1

After removing duplicate sequences from the initial set of 99 fasta sequences, a total of 97 unique sequences remained.

Finalizing the designs for order

To finalize the designs for ordering, representative designs were selected from each of the 1-15 sets mentioned above in the previous table. The designs were divided into two batches: the 1st batch consisting of 16 designs (Figure 4.10), and the 2nd batch consisting of 14 designs (Figure 4.11). In some cases, multiple designs were chosen from a single set if there were differences in multiple amino acids or if at least one amino acid differed in the seven important antigenic positions. These selections were made carefully, incorporating multiple rounds of sequence alignment using the muscle algorithm. The rationale behind finalizing the designs is provided in Table 4.2 and Table 4.3.

Table 4.2: The rationale behind finalizing the designs (batch 1, 16 designs)

No	Short identifier	Full header	Identity/Rationale
1	allSeq1	allSeq1_consensus0.5_all_refined0.9965_combined_with_4fnk_chainEF	Set: Consensus of all years from 1968 to 2023 (Set 14) Rationale: The design was generated based on the consensus sequence derived from all available H3N2 vaccine strains spanning the period from 1968 to 2023. The ambiguous positions (x) within the sequence were replaced with the reference structure template A/Hong Kong/1/1968 (PDB ID 4fnk).
2	allSeq2	allSeq2_consensus0.5_all_refined0.9965_combined_with_	Set: Consensus of all years from 1968 to 2023 (Set 14) Rationale: This design was obtained from the consensus sequence representing all H3N2 vaccine strains from 1968 to 2023. The x positions within the sequence were replaced

		A/Darwin/9/20 21_last_vaccin eStrains	with the reference template sequence A/Darwin/9/2021, which corresponds to the most recent vaccine strain.
3	allSeq3	allSeq3_consen sus0.5_all_refi ned0.9965_co mbined_with_6 bkp_last_VS_st ructure_last_do minant_cluster 3c2a	Set: Consensus of all years from 1968 to 2023 (Set 14) Rationale: This design was derived from the consensus sequence encompassing all H3N2 vaccine strains from 1968 to 2023. The x positions within the sequence were replaced with the reference template 6bkp structure, which represents the most prevalent clade (3c2a) currently circulating in humans and is the latest available crystal structure of a vaccine strain.
4	NJentire1	NJentire1_cons ensus0.5_p3_a _refined1_com bined_with_A/ Darwin/9/2021 _last_vaccineSt rains	Set: Single consensus of the entire sequence derived from the neighbor-joining tree (Set 7) Rationale: This design was generated from the single consensus sequence obtained from the neighbor-joining tree analysis. The x positions within the sequence were substituted with the reference template sequence A/Darwin/9/2021, corresponding to the most recent vaccine strain.
5	NJentire2	NJentire2_cons ensus0.5_p3_a _refined1_com bined_with_co	Set: Single consensus of the entire sequence derived from the neighbor-joining tree (Set 7) Rationale: This design was obtained from the single consensus sequence derived from the neighbor-joining tree analysis. The x positions within

		nsensus0.0_clu ster_refined1_c ombined_with_ 6bkp	the sequence were replaced with the consensus sequence template based on antigenic clusters, where the x positions in that template were replaced with A/Michigan/15/2014.
6	NJentire3	NJentire3_cons ensus0.5_p3_a _refined1_com bined_with_6b kp_last_VS_str ucture_last_do minant_cluster 3c2a	Set: Single consensus of the entire sequence derived from the neighbor-joining tree (Set 7) Rationale: This design was obtained from the single consensus sequence derived from the neighbor-joining tree analysis. The x positions within the sequence were replaced with the template A/Michigan/15/2014.
7	raxEntire1	raxEntire1_con sensus0.5_p8_a _refined1_com bined_with_4fn k_chainEF	Set: Single consensus of the entire sequence derived from the RAxML maximum likelihood tree (Set 1) Rationale: This design was obtained from the single consensus sequence derived from the maximum likelihood tree analysis using the RAxML algorithm. The x positions within the sequence were replaced with the reference structure template A/Hong Kong/1/1968 (PDB ID 4fnk).
8	Cluster1	Cluster1_conse nsus0.0_clus_r efined1_combi	Set: Consensus based on H3N2 antigenic cluster representative sequences (Set 13) Rationale: This design was generated from the consensus sequence based on representative sequences from H3N2 antigenic clusters.

		ned_with_4fnk _chainEF	The x positions within the sequence were replaced with the reference structure template A/Hong Kong/1/1968 (PDB ID 4fnk).
9	Cluster2	Cluster2_consensus0.0_clus_refined1_combined_with_A/Darwin/9/2021_last_vaccineStrains	Set: Consensus based on H3N2 antigenic cluster representative sequences (Set 13) Rationale: This design was derived from the consensus sequence based on representative sequences from H3N2 antigenic clusters. The x positions within the sequence were substituted with the reference template sequence A/Darwin/9/2021, corresponding to the most recent vaccine strain.
10	Cluster3	Cluster3_consensus0.0_clus_refined1_combined_with_6bkp_last_VS_structure_last_dominant_cluster3c2a	Set: Consensus based on H3N2 antigenic cluster representative sequences (Set 13) Rationale: This design was obtained from the consensus sequence based on representative sequences from H3N2 antigenic clusters. The x positions within the sequence were replaced with the reference template 6bkp structure, representing the most prevalent clade (3c2a) currently circulating in humans.
11	LLVS1	LLVS1_consensus0.45_LLVS_refined1_combined_with_A/Darwin/9/2021	Set: Consensus based on H3N2 long-lasting vaccine strains (Set 11) Rationale: This design was derived from the consensus sequence based on H3N2 long-lasting vaccine strains. The x positions within the sequence were replaced

		_last_vaccineStrains	with the reference template sequence A/Darwin/9/2021, corresponding to the most recent vaccine strain.
12	LLVS2	LLVS2_consensus0.45_LLVS_refined1_combined_with_6bkp_last_VS_structure_last_dominant_cluster 3c2a	Set: Consensus based on H3N2 long-lasting vaccine strains (Set 11) Rationale: This design was obtained from the consensus sequence based on H3N2 long-lasting vaccine strains. The x positions within the sequence were replaced with the reference template 6bkp structure, representing the most prevalent clade (3c2a) currently circulating in humans.
13	LLVSH1	LLVSH1_consensus0.45_LLVSH_refined1_combined_with_A/Darwin/9/2021_last_vaccineStrains	Set: Consensus based on H3N2 long-lasting vaccine strains and historic A/Hong Kong/1/1968 sequence (Set 12) Rationale: This design was derived from the consensus sequence based on H3N2 long-lasting vaccine strains, incorporating the historic A/Hong Kong/1/1968 sequence. The x positions within the sequence were replaced with the reference template sequence A/Darwin/9/2021, corresponding to the most recent vaccine strain.
14	VS1	VS1_consensus0.45_vs_a_refined1_combine	Set: Consensus of all H3N2 vaccine strains (Set 10) Rationale: This design was obtained from the consensus sequence derived from all available H3N2 vaccine strains. The x positions within the sequence were replaced with the

		d_with_4fnk_c hainEF	reference structure template A/Hong Kong/1/1968 (PDB ID 4fnk).
15	VS2	VS2_consensus 0.45_vs_a_refi ned1_combine d_with_A/Dar win/9/2021_las t_vaccineStrain s	Set: Consensus of all H3N2 vaccine strains (Set 10) Rationale: This design was generated from the consensus sequence derived from all available H3N2 vaccine strains. The x positions within the sequence were replaced with the reference template sequence A/Darwin/9/2021, corresponding to the most recent vaccine strain.
16	VS3	VS3_consensus 0.45_vs_a_refi ned1_combine d_with_6bkp_l ast_VS_structu re_last_domina nt_cluster3c2a	Set: Consensus of all H3N2 vaccine strains (Set 10) Rationale: This design was derived from the consensus sequence encompassing all H3N2 vaccine strains. The x positions within the sequence were replaced with the reference template 6bkp structure, representing the most prevalent clade (3c2a) currently circulating in humans and the latest available crystal structure of a vaccine strain.

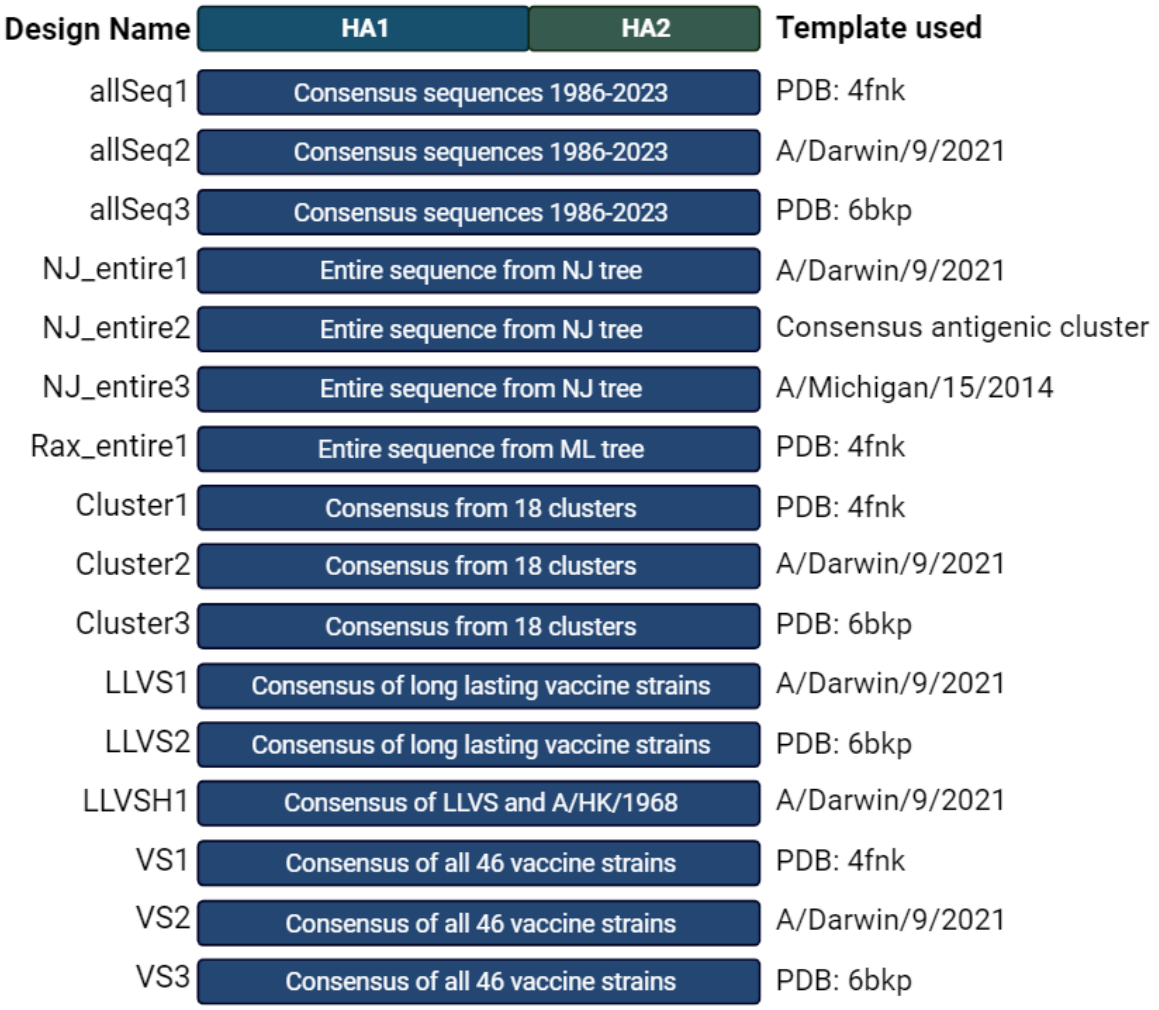

Figure 4.10. Schematic representation of the finalized 16 designs of batch 1

Table 4.3: The rationale behind finalizing the designs (batch 2, 14 designs)

No	Short identifier	Full header	Identity/Rationale
1	upgma_entire1	upgma_entire1_consensus0.5_seq4_refined1_combined_with	Set: Single consensus of the entire sequence derived from the UPGMA tree (Set 4) Rationale: This design was generated from the single consensus sequence obtained from the UPGMA tree

		_4fnk_chainEF	analysis. The x positions within the sequence were replaced with the reference structure template A/Hong Kong/1/1968 (PDB ID 4fnk).
2	upgma_head1	upgma_head1_consensus0.5_p4_a_refined1_combined_with_4fnk_chainEF	Set: Consensus of head region from UPGMA clustering with remaining sequence of 4fnk. (Set 5) Rationale: The consensus sequence of the head region obtained from UPGMA clustering, with the remaining sequence derived from the reference structure template A/Hong Kong/1/1968 (PDB ID 4fnk) template. The x positions were also replaced with the same template.
3	upgma_head9	upgma_head9_consensus0.5_p4_a_refined1_combined_with_6bkp_last_VS_structure_last_dominant_cluster3c2a	Set: Consensus of head region from UPGMA clustering with remaining sequence of 6bkp. (Set 5) Rationale: The consensus sequence of the head region obtained from UPGMA clustering, with the remaining sequence derived from the reference structure template 6bkp structure, representing the most prevalent clade (3c2a) currently circulating in humans and the latest available crystal structure of a vaccine strain. The x positions were also replaced with the same template.
4	nj_head7	nj_head7_consensus0.5_p5_a_refined1_combined_with_co	Set: Consensus of head region from neighbor-joining clustering with remaining sequence of vaccine strains. (Set 8)

		nsensus0.45_V S_a_refined1_c ombined_with_ 6bkp	Rationale: The consensus sequence of the head region obtained from neighbor-joining tree clustering, with the remaining sequence derived from the consensus of vaccine strains. The x positions were also replaced with the same template.
5	nj_head8	nj_head8_cons ensus0.5_p5_a _refined1_com bined_with_4w e9_VC11	Set: Consensus of head region from neighbor-joining clustering with remaining sequence of vaccine strains. (Set 8) Rationale: The consensus sequence of the head region obtained from neighbor-joining tree clustering, with the remaining sequence derived from the reference structure template A/Victoria/361/2011 (PDB ID 4we9) template. The x positions were also replaced with the same template.
6	rax_Head3	rax_Head3_con sensus0.5_p8_a _refined1_com bined_with_A/ Darwin/9/2021 _last_vaccineSt rains	Set: Single consensus of the head sequence derived from the RAxML maximum likelihood tree, where the rest of the sequence is from A/Darwin/9/2021 (Set 2). Rationale: The consensus sequence of the head region obtained from RAxML maximum likelihood tree, with the remaining sequence derived from the reference template A/Darwin/9/2021. The x positions were also replaced with the same template.
7	rax_Head6	rax_Head6_con sensus0.5_p8_a	Set: Single consensus of the head sequence derived from the RAxML maximum likelihood tree, where the rest of the

		_refined1_com bined_with_co nsensus0.0_clu ster_refined1_c ombined_with_ 6bkp	sequence is from consensus of cluster recombined with 6bkp. Rationale: The consensus sequence of the head region obtained from RAxML maximum likelihood tree, with the remaining sequence derived from the reference template of consensus of antigenic clusters. The x positions were also replaced with the same template.
8	upgma_stem1	upgma_stem1_ consensus0.5_p 3_a_refined1_c ombined_with_ 4fnk_chainEF	Set: Consensus of stem region from UPGMA clustering with remaining sequence of 4fnk. (Set 6) Rationale: The consensus sequence of the stem region obtained from UPGMA clustering, with the remaining sequence derived from the reference structure template A/Hong Kong/1/1968 (PDB ID 4fnk) template. The x positions were also replaced with the same template.
9	upgma_stem4	upgma_stem4_ consensus0.5_p 3_a_refined1_c ombined_with_ 6bkp_last_VS_ structure_last_ dominant_clust er3c2a	Set: Consensus of stem region from UPGMA clustering with remaining sequence of 6bkp. (Set 6) Rationale: The consensus sequence of the stem region obtained from UPGMA clustering, with the remaining sequence derived from the reference structure template 6bkp structure, representing the most prevalent clade (3c2a) currently circulating in humans and the latest

			available crystal structure of a vaccine strain. The x positions were also replaced with the same template.
10	nj_stem4	nj_stem4_consensus0.5_p3_a_refined1_combined_with_consensus0.45_VS_a_refined1_combined_with_6bkp	Set: Consensus of stem region from neighbor-joining clustering with remaining sequence of vaccine strains. (Set 9) Rationale: The consensus sequence of the head region obtained from neighbor-joining tree clustering, with the remaining sequence derived from the consensus of vaccine strains. The x positions were also replaced with the same template.
11	nj_stem5	nj_stem5_consensus0.5_p3_a_refined1_combined_with_4we9_VC11	Set: Consensus of stem region from neighbor-joining clustering with remaining sequence of vaccine strains. (Set 9) Rationale: The consensus sequence of the stem region obtained from neighbor-joining tree clustering, with the remaining sequence derived from the reference structure template A/Victoria/361/2011 (PDB ID 4we9) template. The x positions were also replaced with the same template.
12	rax_stem1	rax_stem1_consensus0.5_p4_a_refined1_com	Set: Single consensus of the stem sequence derived from the RAxML maximum likelihood tree, where the rest of the sequence is from A/Darwin/9/2021 (Set 3).

		bined_with_A/ Darwin/9/2021 _last_vaccineSt rains	Rationale: The consensus sequence of the stem region obtained from RAxML maximum likelihood tree, with the remaining sequence derived from the reference template A/Darwin/9/2021. The x positions were also replaced with the same template.
13	rax_stem3	rax_stem3_cco nsensus0.5_p4_ a_refined1_co mbined_with_c onsensus0.0_cl uster_refined1_ combined_with _6bkp	Set: Single consensus of the stem sequence derived from the RAxML maximum likelihood tree, where the rest of the sequence is from A/Darwin/9/2021 (Set 3). Rationale: The consensus sequence of the stem region obtained from RAxML maximum likelihood tree, with the remaining sequence derived from the reference template of consensus of antigenic clusters. The x positions were also replaced with the same template.
14	after2015	after2015_cons ensus0.5_post_ 2015_refined0. 9965.fasta filter 0.9965 threshold 0.5	Set: Consensus of all years from 2015 to May 2023 (Set 15) Rationale: In the scenario of consensus sequences encompassing the time frame between 2015 and May 2023, we encountered 1 x position located within the signal peptide region, but not among the seven critical antigenic positions. We replaced this position with the template 4fnk. This particular time frame was selected due to the substantial antigenic drift experienced by H3N2 during the

	2014-2015 season, resulting in the emergence of a new cluster

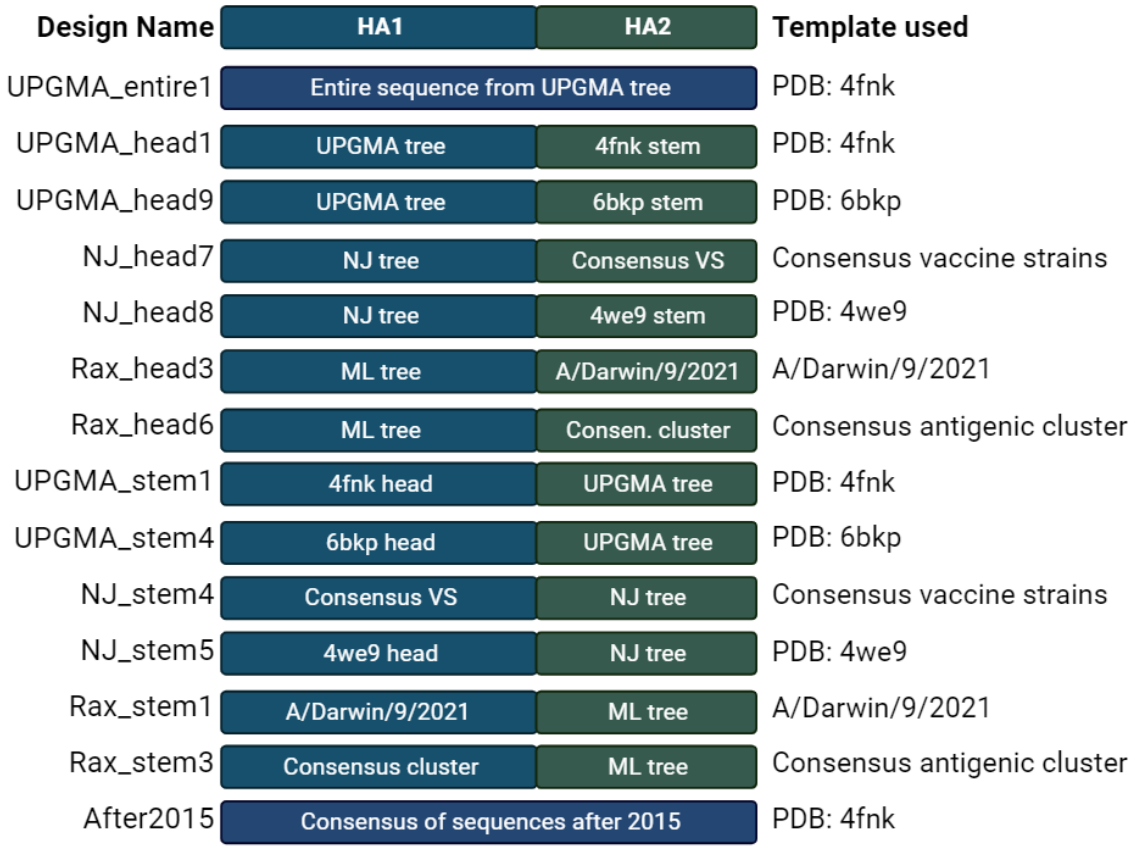

Figure 4.11. Schematic representation of the finalized 14 designs of batch 2

Ultimately, we have successfully distilled a selection of 30 designs from an initial pool of 97 (Figure 4.12), encompassing a diverse range of methodological approaches. These methodologies include time window-based clustering, phylogenetic tree-based clustering (utilizing UPGMA, neighbor-joining, and maximum likelihood methods), antigenic clustering, and vaccine strain considerations. It is noteworthy that our final set of 30 designs has been

meticulously curated to ensure comprehensive coverage of all the aforementioned methodologies. Specifically, within the realm of time window clustering, we have arrived at a final set of 4 designs out of an initial 8. For phylogenetic tree-based clustering, our selections consist of 5 designs from 16 for UPGMA, 7 designs from 24 for neighbor-joining, and 5 designs from 25 in the case of maximum likelihood tree-based approaches. These selections span across different aspects of the HA sequence, encompassing designs related to the entire HA sequence, the HA head, and the HA stem regions. Within the domain of antigenic clustering, we have identified 3 designs from a pool of 7, and for vaccine strains, we have finalized 6 designs from an initial compilation of 28 consensus designs.

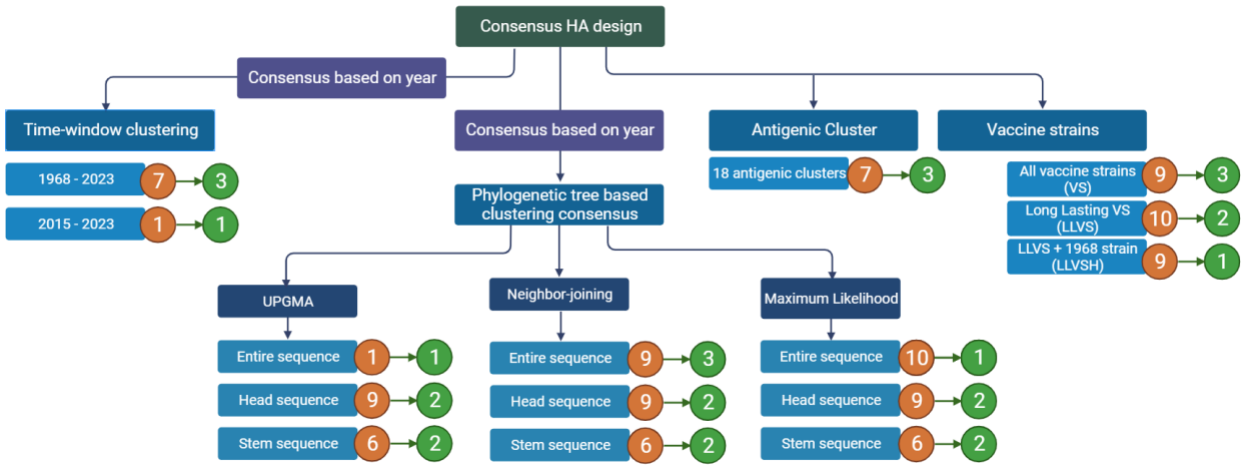

Figure 4.12. Number of finalized designs under each consensus design categories. Number in orange colors represent the initial number of designs and the number in green colors represent the final designs ordered.

Addition of Foldon, Histag and Codon optimization

After finalizing the designs, each sequence was further processed by adding the signal peptide from the reference strain A/Aichi/2-1/1968 at the beginning of the sequence. Additionally,

a foldon domain and a 6x histidine (His) tag were appended at the end of each sequence. The modified sequences are presented below:

Signal peptide
Foldon
6histag

>allSeq1_consensus0.5_all_refined0.9965_combined_with_4fnk_chainEF

MKTIIALSYIFCLALGQDLPGNDNSTATLCLGHHAVPNGTIVKTITNDQIEVTNATELVQS
SSTGEICDSPHQILDGINCTLIDALLGDPQCDGFQNAKWDLFVERPSKYSNCYPYDVPDY
ASLRSLVASSGTLEFINESFNWTGVTQNGTSSACIRGSSNSFFSRLNWLTHLGSKYPALN
VTMPNNEKFDKLYIWGVHHPGTDQDQIFLYAQASGRITVSTKRSQQTVIPNIGSRPRVRG
IPSRISIYWTIVKPGDILLINSTGNLIAPRGYFKIRSGKSSIMRSDAPIGKCISECITPNGSIPN
DKPFQNVNRITYGACPRYVKQNTLKLATGMRNVPEKQTRGIFGAIAGFIENGWEGMVD
GWYGFRHQNSEGTGQAADLKSTQAAIDQINGKLNRLIGKTNEKFHQIEKEFSEVEGRIQ
DLEKYVEDTKIDLWSYNAELLVALENQHTIDLTDSEMNKLFEKTKKQLRENAEDMGNG
CFKIYHKCDNACIGSIRNGTYDHDVYRDEALNNRFQIKGGYIPEAPRDGQAYVRKDGE
WVLLSTFLGHHHHHH

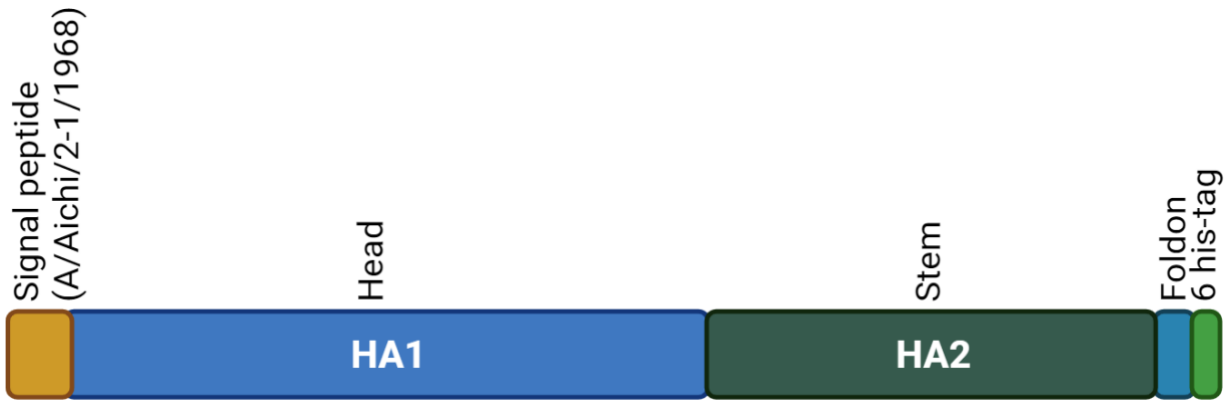

Figure 4.13: Representative of final design where we added signal peptide, foldon and 6 his-tag with each designed sequence. The signal peptide is depicted in yellow, the head region

is denoted in blue, and the stem domain is represented in dark green. Moving towards the C-terminal end, the foldon domain is visualized in cyan, while the 6-histidine (6 His-tag) sequence is displayed in light green.

For the codon optimization process, certain restriction enzymes, namely NdeI [CATATG], NheI [GCTAGC], SacI [GAGCTC], and XhoI [CTCGAG], were excluded from consideration, and the human expression host was selected. Codon optimization was performed using the Thermo Fisher Invitrogen GeneArt Gene Synthesis Services tool. Following the optimization, the sequence modification included the addition of the sequence GCTAGCGCCACC at the beginning of each sequence, and the sequence TAA at the end of each sequence. The final sequence for the Allseq1 design is presented below:

>allSeq1

GCTAGCGCCACCATGAAGACCATCATTGCCCTGAGCTACATCTTCTGTCTGGCCCTC
GGACAGGACCTGCCTGGCAACGATAATAGCACCGCCACACTGTGTCTGGGCCACCA
CGCTGTGCCTAATGGCACCATCGTGAAAACCATCACCAACGACCAGATCGAAGTGA
CCAACGCCACCGAGCTGGTGCAGTCTAGCTCTACCGGCGAGATCTGCGATAGCCCTC
ACCAGATCCTGGACGGCATCAACTGCACCCTGATCGATGCCCTGCTGGGCGACCCTC
AGTGTGACGGATTTCAGAACGCCAAGTGGGACCTGTTCGTGGAACGGCCCAGCAAG
TACAGCAACTGCTACCCCTACGACGTGCCCGATTACGCCAGCCTGAGATCTCTGGTG
GCCAGCTCTGGCACCCTGGAATTCATCAACGAGAGCTTCAACTGGACCGGCGTGAC
CCAGAATGGCACAAGCAGCGCCTGTATCAGAGGCAGCAGCAACAGCTTCTTCAGCA
GACTGAACTGGCTGACCCACCTGGGCAGCAAGTATCCCGCTCTGAACGTGACCATG
CCTAACAACGAGAAGTTCGACAAGCTGTACATCTGGGGCGTGCACCATCCTGGCAC

CGACCAGGATCAGATCTTCCTGTATGCCCAGGCCAGCGGCAGAATCACCGTGTCCAC

CAAAAGATCCCAGCAGACAGTGATCCCCAACATCGGCAGCAGACCTAGAGTGCGGG

GCATCCCTAGCCGGATCAGCATCTACTGGACAATCGTGAAGCCCGGCGACATCCTGC

TGATCAACAGCACCGGCAATCTGATCGCCCCTCGGGGCTACTTCAAGATCAGAAGC

GGCAAGAGCAGCATCATGCGGAGCGACGCCCCTATCGGCAAGTGCATCAGCGAGTG

CATCACCCCTAACGGCAGCATCCCCAACGACAAGCCCTTCCAGAACGTGAACCGGA

TCACCTACGGCGCCTGTCCTAGATACGTGAAGCAGAACACCCTGAAACTGGCCACC

GGCATGAGAAATGTGCCCGAGAAGCAGACCAGAGGCATCTTCGGAGCCATTGCCGG

CTTCATCGAGAACGGCTGGGAAGGCATGGTGGACGGATGGTACGGCTTCAGACACC

AGAACAGCGAAGGCACAGGACAGGCCGCTGACCTGAAATCTACACAGGCCGCCATC

GACCAGATTAACGGCAAGCTGAACCGGCTGATCGGCAAGACCAATGAGAAGTTCCA

CCAGATTGAGAAAGAGTTCAGCGAGGTCGAGGGCAGAATCCAGGACCTTGAGAAAT

ACGTCGAGGACACCAAGATCGACCTGTGGTCCTACAACGCCGAACTGCTGGTGGCC

CTGGAAAACCAGCACACCATCGACCTGACCGACAGCGAGATGAACAAGCTGTTCGA

AAAGACCAAGAAGCAGCTGCGCGAGAACGCCGAGGATATGGGCAACGGCTGCTTTA

AGATCTACCACAAGTGCGACAACGCCTGCATCGGCTCCATCAGAAACGGCACCTAC

GACCACGACGTGTACAGAGATGAGGCCCTGAACAACCGGTTCCAGATCAAAGGCGG

CTACATCCCCGAGGCTCCTAGAGATGGACAGGCCTACGTCAGAAAGGACGGCGAAT

GGGTGCTGCTGAGCACATTCCTGGGACACCACCATCATCACCACTAA

CHAPTER 5

SUMMARY AND CONCLUSION

The elusive ideal influenza vaccine, flexible and universal, that has eluded virology and immunology. Antigenic drift driving this very threat to global public health annually wreaks havoc on the world with thousands hospitalized and tens of thousands dying per year. Our research project for this assignment was aimed at advancing influenza vaccine design by innovative use of consensus-based approaches specifically targeted toward the H3N2 HA protein sequence. We sought primarily a strategy to harness sequences containing consensus, broaden protection afforded against both circulating and emerging H3N2 variants contributing ultimately to a pursuit of a universal influenza vaccine. Our research explores the concept of consensus-based vaccine design as a potential solution to the challenges posed by antigenic drift. This approach involves the creation of synthetic HA proteins that represent the diversity of the HA population, aiming to enhance cross-reactivity and broaden vaccine coverage. Various strategies within the consensus framework, including micro-consensus immunogens, COBRA, and centralized HA genes, hold promise in achieving this goal.

To diversify our consensus sequence set, we employed different sequence-clustering algorithms, as detailed in the method section. These algorithms allowed us to identify distinct phylogenetic clusters within the HA protein sequence, enabling us to select consensus sequences strategically. We applied this approach to full-length HA sequences, truncated sequences containing only the head region (the immunodominant region), and sequences encompassing solely the stem region of the HA protein. We acknowledge that there are still challenges to

overcome, such as potential biases in selecting the population for vaccine design. Moreover, the vaccine that we designed using this strategy needs to be tested both in-vitro and in vivo.

The implications of our work are paramount. We expect, given a consensus-based strategy, that the influenza vaccines will become more effective and adaptable in their capacity. Focusing on the protein sequence H3N2 HA puts an emphasis on broadening protection and insuring continued efficacy against prevalent variants as well as ones to come. This approach could relieve some of the pressure of annual strain selection and conjunction with global efforts for outbreaks.

In summary, our efforts represent an important milestone forward in expanding the breadth of flu protection with new design of influenza vaccine. By making creative use of consensus approaches, we have provided hope that expansion of protection against the H3N2 influenza variants can be possible. As the influenza virus evolves so too must our strategies for defeating it. We look forward to future progress and to the promise of a brighter future where more effective and flexible influenza vaccines better able to protect the world population are available.

REFERENCES

1. Petrova, V.N. and C.A.J.N.R.M. Russell, *The evolution of seasonal influenza viruses.* 2018. **16**(1): p. 47-60.
2. WHO, *Influenza (seasonal) fact sheet*. 2018, World Health Organization Geneva, Switzerland.
3. Hsieh, Y.-C., et al., *Influenza pandemics: past, present and future.* 2006. **105**(1): p. 1-6.
4. Houser, K., K.J.C.h. Subbarao, and microbe, *Influenza vaccines: challenges and solutions.* 2015. **17**(3): p. 295-300.
5. Carrat, F. and A.J.V. Flahault, *Influenza vaccine: the challenge of antigenic drift.* 2007. **25**(39-40): p. 6852-6862.
6. Hampson, A.W. and J.S.J.M.J.o.A. Mackenzie, *The influenza viruses.* 2006. **185**(S10): p. S39-S43.
7. Boni, M.F.J.V., *Vaccination and antigenic drift in influenza.* 2008. **26**: p. C8-C14.
8. Tenforde, M.W., et al., *Effect of antigenic drift on influenza vaccine effectiveness in the United States—2019–2020.* 2021. **73**(11): p. e4244-e4250.
9. Schweiger, B., et al., *Antigenic drift and variability of influenza viruses.* 2002. **191**: p. 133-138.
10. Allen, J.D., T.M.J.H.v. Ross, and immunotherapeutics, *H3N2 influenza viruses in humans: Viral mechanisms, evolution, and evaluation.* 2018. **14**(8): p. 1840-1847.
11. Jester, B.J., T.M. Uyeki, and D.B.J.A.j.o.p.h. Jernigan, *Fifty years of influenza A (H3N2) following the pandemic of 1968.* 2020. **110**(5): p. 669-676.
12. Scholtissek, C., et al., *On the origin of the human influenza virus subtypes H2N2 and H3N2.* 1978. **87**(1): p. 13-20.

13. Bertram, S., et al., *Novel insights into proteolytic cleavage of influenza virus hemagglutinin.* 2010. **20**(5): p. 298-310.

14. Taubenberger, J.K.J.P.o.t.N.A.o.S., *Influenza virus hemagglutinin cleavage into HA1, HA2: no laughing matter.* 1998. **95**(17): p. 9713-9715.

15. Wu, N.C. and I.A.J.V. Wilson, *Structural biology of influenza hemagglutinin: An amaranthine adventure.* 2020. **12**(9): p. 1053.

16. Wei, C.-J., et al., *Next-generation influenza vaccines: opportunities and challenges.* 2020. **19**(4): p. 239-252.

17. Lazniewski, M., et al., *The structural variability of the influenza A hemagglutinin receptor-binding site.* 2018. **17**(6): p. 415-427.

18. Krammer, F.J.C.h. and microbe, *The quest for a universal flu vaccine: headless HA 2.0.* 2015. **18**(4): p. 395-397.

19. Elliott, S.T., et al., *A synthetic micro-consensus DNA vaccine generates comprehensive influenza A H3N2 immunity and protects mice against lethal challenge by multiple H3N2 viruses.* 2018. **29**(9): p. 1044-1055.

20. Allen, J.D. and T.M.J.S.R. Ross, *Next generation methodology for updating HA vaccines against emerging human seasonal influenza A (H3N2) viruses.* 2021. **11**(1): p. 4554.

21. Allen, J.D. and T.M.J.F.i.I. Ross, *Evaluation of next-generation H3 influenza vaccines in ferrets pre-immune to historical H3N2 viruses.* 2021. **12**: p. 707339.

22. Weaver, E.A., et al., *Protection against divergent influenza H1N1 virus by a centralized influenza hemagglutinin.* 2011. **6**(3): p. e18314.

23. Morens, D.M., M. North, and J.K.J.T.L. Taubenberger, *Eyewitness accounts of the 1510 influenza pandemic in Europe.* 2010. **376**(9756): p. 1894-1895.

24. Beveridge, W.I.J.H. and p.o.t.l. sciences, *The chronicle of influenza epidemics.* 1991: p. 223-234.

25. Al Hajjar, S. and K.J.A.o.S.m. McIntosh, *The first influenza pandemic of the 21st century.* 2010. **30**(1): p. 1-10.

26. Blagodatski, A., et al., *Avian influenza in wild birds and poultry: dissemination pathways, monitoring methods, and virus ecology.* 2021. **10**(5): p. 630.

27. [CDC], C.f.D.C.a.P. *2022-2023 Preliminary In-Season Burden Estimate*. 2023 08/03/2023]; Available from: https://www.cdc.gov/flu/about/burden/preliminary-in-season-estimates.htm#:~:text=During%20the%202021%2D2022%20influenza,10%2C000%20hospitalizations%2C%20and%205%2C000%20deaths.

28. Viboud, C., W.J. Alonso, and L.J.P.m. Simonsen, *Influenza in tropical regions.* 2006. **3**(4): p. e89.

29. Javanian, M., et al., *A brief review of influenza virus infection.* 2021. **93**(8): p. 4638-4646.

30. Lamb, R.A. and P.W.J.A.r.o.b. Choppin, *The gene structure and replication of influenza virus.* 1983. **52**(1): p. 467-506.

31. Lakadamyali, M., et al., *Endocytosis of influenza viruses.* 2004. **6**(10): p. 929-936.

32. Ada, G., P.J.C.t.i.m. Jones, and immunology, *The immune response to influenza infection.* 1986: p. 1-54.

33. Lagace-Wiens, P.R., E. Rubinstein, and A.J.C.c.m. Gumel, *Influenza epidemiology—past, present, and future.* 2010. **38**: p. e1-e9.

34. Cheung, T.K. and L.L.J.A.o.t.N.Y.A.o.S. Poon, *Biology of influenza a virus.* 2007. **1102**(1): p. 1-25.

35. Koutsakos, M., et al., *Knowns and unknowns of influenza B viruses.* 2016. **11**(1): p. 119-135.

36. Matsuzaki, Y., et al., *Clinical features of influenza C virus infection in children.* 2006. **193**(9): p. 1229-1235.

37. Ma, W., et al., *The role of swine in the generation of novel influenza viruses.* 2009. **56**(6-7): p. 326-337.

38. Liu, R., et al., *Influenza D virus.* 2020. **44**: p. 154-161.

39. Asha, K. and B.J.J.o.C.M. Kumar, *Emerging influenza D virus threat: what we know so far!* 2019. **8**(2): p. 192.

40. Kuchipudi, S.V. and R.H.J.V.s. Nissly, *Novel flu viruses in bats and cattle:"Pushing the Envelope" of Influenza Infection.* 2018. **5**(3): p. 71.

41. Bouvier, N.M. and P.J.V. Palese, *The biology of influenza viruses.* 2008. **26**: p. D49-D53.

42. Cauldwell, A.V., et al., *Viral determinants of influenza A virus host range.* 2014. **95**(6): p. 1193-1210.

43. Wagner, R., M. Matrosovich, and H.D.J.R.i.m.v. Klenk, *Functional balance between haemagglutinin and neuraminidase in influenza virus infections.* 2002. **12**(3): p. 159-166.

44. Wang, M., M.J.P. Veit, and cell, *Hemagglutinin-esterase-fusion (HEF) protein of influenza C virus.* 2016. **7**(1): p. 28-45.

45. Watanabe, T., et al., *Immunogenicity and protective efficacy of replication-incompetent influenza virus-like particles.* 2002. **76**(2): p. 767-773.

46. Roberts, P.C., R.A. Lamb, and R.W.J.V. Compans, *The M1 and M2 proteins of influenza A virus are important determinants in filamentous particle formation.* 1998. **240**(1): p. 127-137.

47. Ito, T., et al., *Evolutionary analysis of the influenza A virus M gene with comparison of the M1 and M2 proteins.* 1991. **65**(10): p. 5491-5498.

48. Elleman, C. and W.J.V. Barclay, *The M1 matrix protein controls the filamentous phenotype of influenza A virus.* 2004. **321**(1): p. 144-153.

49. Wang, C., et al., *Ion channel activity of influenza A virus M2 protein: characterization of the amantadine block.* 1993. **67**(9): p. 5585-5594.

50. García-Sastre, A., et al., *Influenza A virus lacking the NS1 gene replicates in interferon-deficient systems.* 1998. **252**(2): p. 324-330.

51. Webster, R.G., et al., *Evolution and ecology of influenza A viruses.* 1992. **56**(1): p. 152-179.

52. Yoon, S.-W., et al., *Evolution and ecology of influenza A viruses.* 2014: p. 359-375.

53. Liu, S., et al., *Panorama phylogenetic diversity and distribution of type A influenza virus.* 2009. **4**(3): p. e5022.

54. Bahl, J., et al., *Gene flow and competitive exclusion of avian influenza A virus in natural reservoir hosts.* 2009. **390**(2): p. 289-297.

55. Tong, S., et al., *New world bats harbor diverse influenza A viruses.* 2013. **9**(10): p. e1003657.

56. Rejmanek, D., et al., *Evolutionary dynamics and global diversity of influenza A virus.* 2015. **89**(21): p. 10993-11001.

57. Arzey, G.G., et al., *Influenza virus A (H10N7) in chickens and poultry abattoir workers, Australia.* 2012. **18**(5): p. 814.

58. Yuan, J., et al., *Origin and molecular characteristics of a novel 2013 avian influenza A (H6N1) virus causing human infection in Taiwan.* 2013. **57**(9): p. 1367-1368.

59. Peiris, M., et al., *Human infection with influenza H9N2.* 1999. **354**(9182): p. 916-917.

60. Fouchier, R.A., et al., *Avian influenza A virus (H7N7) associated with human conjunctivitis and a fatal case of acute respiratory distress syndrome.* 2004. **101**(5): p. 1356-1361.

61. Koopmans, M., et al., *Transmission of H7N7 avian influenza A virus to human beings during a large outbreak in commercial poultry farms in the Netherlands.* 2004. **363**(9409): p. 587-593.

62. Tanner, W., et al., *The pandemic potential of avian influenza A (H7N9) virus: a review.* 2015. **143**(16): p. 3359-3374.

63. To, K.K., et al., *Emergence in China of human disease due to avian influenza A (H10N8)–cause for concern?* 2014. **68**(3): p. 205-215.

64. Cockburn, W.C., P. Delon, and W.J.B.o.t.W.H.O. Ferreira, *Origin and progress of the 1968-69 Hong Kong influenza epidemic.* 1969. **41**(3-4-5): p. 343.

65. Chang, W.J.B.o.t.W.h.o., *National influenza experience in Hong Kong, 1968.* 1969. **41**(3-4-5): p. 349.

66. Grais, R.F., J. Hugh Ellis, and G.E.J.E.j.o.e. Glass, *Assessing the impact of airline travel on the geographic spread of pandemic influenza.* 2003. **18**: p. 1065-1072.

67. Sharrar, R.G.J.B.o.t.W.H.O., *National influenza experience in the USA, 1968-69.* 1969. **41**(3-4-5): p. 361.

68. Westgeest, K.B., et al., *Genomewide analysis of reassortment and evolution of human influenza A (H3N2) viruses circulating between 1968 and 2011.* 2014. **88**(5): p. 2844-2857.

69. Van Poucke, S., et al., *Role of substitutions in the hemagglutinin in the emergence of the 1968 pandemic influenza virus.* 2015. **89**(23): p. 12211-12216.

70. Shao, W., et al., *Evolution of influenza a virus by mutation and re-assortment.* 2017. **18**(8): p. 1650.

71. Weis, W., et al., *Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid.* 1988. **333**(6172): p. 426-431.

72. Kobayashi, Y. and Y.J.P.o. Suzuki, *Compensatory evolution of net-charge in influenza A virus hemagglutinin.* 2012. **7**(7): p. e40422.

73. Park, A.W., et al., *Quantifying the impact of immune escape on transmission dynamics of influenza.* 2009. **326**(5953): p. 726-728.

74. Blackburne, B.P., A.J. Hay, and R.A.J.P.p. Goldstein, *Changing selective pressure during antigenic changes in human influenza H3.* 2008. **4**(5): p. e1000058.

75. Park, J.-K., et al., *Pre-existing immunity to influenza virus hemagglutinin stalk might drive selection for antibody-escape mutant viruses in a human challenge model.* 2020. **26**(8): p. 1240-1246.

76. Saunders-Hastings, P.R. and D.J.P. Krewski, *Reviewing the history of pandemic influenza: understanding patterns of emergence and transmission.* 2016. **5**(4): p. 66.

77. Glaser, L., et al., *A single amino acid substitution in 1918 influenza virus hemagglutinin changes receptor binding specificity.* 2005. **79**(17): p. 11533-11536.

78. Subbarao, E.K., W. London, and B.R.J.J.o.v. Murphy, *A single amino acid in the PB2 gene of influenza A virus is a determinant of host range.* 1993. **67**(4): p. 1761-1764.

79. Shinya, K., et al., *PB2 amino acid at position 627 affects replicative efficiency, but not cell tropism, of Hong Kong H5N1 influenza A viruses in mice.* 2004. **320**(2): p. 258-266.

80. Herfst, S., et al., *Airborne transmission of influenza A/H5N1 virus between ferrets.* 2012. **336**(6088): p. 1534-1541.

81. Russell, C.A., et al., *The potential for respiratory droplet–transmissible A/H5N1 influenza virus to evolve in a mammalian host.* 2012. **336**(6088): p. 1541-1547.

82. Zambon, M.C.J.J.o.A.C., *Epidemiology and pathogenesis of influenza.* 1999. **44**(suppl_2): p. 3-9.

83. Treanor, J.J.N.E.J.o.M., *Influenza vaccine—outmaneuvering antigenic shift and drift.* 2004. **350**(3): p. 218-220.

84. Lycett, S.J., F. Duchatel, and P.J.P.T.o.t.R.S.B. Digard, *A brief history of bird flu.* 2019. **374**(1775): p. 20180257.

85. Quammen, D., *Spillover: animal infections and the next human pandemic*. 2012: WW Norton & Company.

86. Morens, D.M. and A.S.J.T.J.o.i.d. Fauci, *The 1918 influenza pandemic: insights for the 21st century.* 2007. **195**(7): p. 1018-1028.

87. Skehel, J.J. and M.D.J.P.o.t.N.A.o.S. Waterfield, *Studies on the primary structure of the influenza virus hemagglutinin.* 1975. **72**(1): p. 93-97.

88. Wilson, I.A. and N.J.J.A.r.o.i. Cox, *Structural basis of immune recognition of influenza virus hemagglutinin.* 1990. **8**(1): p. 737-787.

89. Wiley, D., I. Wilson, and J.J.N. Skehel, *Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation.* 1981. **289**(5796): p. 373-378.

90. Smith, D.J., et al., *Mapping the antigenic and genetic evolution of influenza virus.* 2004. **305**(5682): p. 371-376.

91. Koel, B.F., et al., *Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution.* 2013. **342**(6161): p. 976-979.

92. Fonville, J.M., et al., *Antigenic maps of influenza A (H3N2) produced with human antisera obtained after primary infection.* 2016. **213**(1): p. 31-38.

93. Burke, D.F.J.V., *Structural Consequences of Antigenic Variants of Human A/H3N2 Influenza Viruses.* 2023. **15**(4): p. 1008.

94. Chambers, B.S., et al., *Identification of hemagglutinin residues responsible for H3N2 antigenic drift during the 2014–2015 influenza season.* 2015. **12**(1): p. 1-6.

95. Li, C., et al., *Selection of antigenically advanced variants of seasonal influenza viruses.* 2016. **1**(6): p. 1-10.

96. Jorquera, P.A., et al., *Insights into the antigenic advancement of influenza A (H3N2) viruses, 2011–2018.* 2019. **9**(1): p. 2676.

97. Koel, B.F., et al., *Identification of amino acid substitutions supporting antigenic change of influenza A (H1N1) pdm09 viruses.* 2015. **89**(7): p. 3763-3775.

98. Linster, M., et al., *The molecular basis for antigenic drift of human A/H2N2 influenza viruses.* 2019. **93**(8): p. 10.1128/jvi. 01907-18.

99. Koel, B.F., et al., *Antigenic variation of clade 2.1 H5N1 virus is determined by a few amino acid substitutions immediately adjacent to the receptor binding site.* 2014. **5**(3): p. 10.1128/mbio. 01070-14.

100. Lewis, N., et al., *Antigenic and genetic evolution of equine influenza A (H3N8) virus from 1968 to 2007.* 2011. **85**(23): p. 12742-12749.

101. Lorusso, A., et al., *Genetic and antigenic characterization of H1 influenza viruses from United States swine from 2008.* 2011. **92**(Pt 4): p. 919.

102. Lewis, N.S., et al., *Substitutions near the hemagglutinin receptor-binding site determine the antigenic evolution of influenza A H3N2 viruses in US swine.* 2014. **88**(9): p. 4752-4763.

103. Doud, M.B., J.M. Lee, and J.D.J.N.c. Bloom, *How single mutations affect viral escape from broad and narrow antibodies to H1 influenza hemagglutinin.* 2018. **9**(1): p. 1386.

104. Harding, A.T. and N.S.J.V. Heaton, *Efforts to improve the seasonal influenza vaccine.* 2018. **6**(2): p. 19.

105. Grohskopf, L.A., et al., *Prevention and control of seasonal influenza with vaccines: recommendations of the Advisory Committee on Immunization Practices, United States, 2021–22 influenza season.* 2021. **70**(5): p. 1.

106. Gross, P.A., et al., *A controlled double-blind comparison of reactogenicity, immunogenicity, and protective efficacy of whole-virus and split-product influenza vaccines in children.* 1977. **136**(5): p. 623-632.

107. O'Gorman, W.E., et al., *The Split Virus Influenza Vaccine rapidly activates immune cells through Fcγ receptors.* 2014. **32**(45): p. 5989-5997.

108. Carter, N.J. and M.P.J.D. Curran, *Live attenuated influenza vaccine (FluMist®; fluenz™) a review of its use in the prevention of seasonal influenza in children and adults.* 2011. **71**: p. 1591-1622.

109. de Jong, J.C., et al., *Mismatch between the 1997/1998 influenza vaccine and the major epidemic A (H3N2) virus strain as the cause of an inadequate vaccine-induced antibody response to this strain in the elderly.* 2000. **61**(1): p. 94-99.

110. Skowronski, D., et al., *Estimating vaccine effectiveness against laboratory-confirmed influenza using a sentinel physician network: results from the 2005–2006 season of dual A and B vaccine mismatch in Canada.* 2007. **25**(15): p. 2842-2851.

111. Skowronski, D.M., et al., *Low 2012–13 influenza vaccine effectiveness associated with mutation in the egg-adapted H3N2 vaccine strain not antigenic drift in circulating viruses.* 2014. **9**(3): p. e92153.

112. Skehel, J.J.B., *An overview of influenza haemagglutinin and neuraminidase.* 2009. **37**(3): p. 177-178.

113. Bhatt, S., et al., *The genomic rate of molecular adaptation of the human influenza A virus.* 2011. **28**(9): p. 2443-2451.

114. Steinhauer, D.A.J.V., *Role of hemagglutinin cleavage for the pathogenicity of influenza virus.* 1999. **258**(1): p. 1-20.

115. Kostolanský, F., et al., *Universal anti-influenza vaccines based on viral HA2 and M2e antigens.* 2020. **64**: p. 417-426.

116. Bullough, P.A., et al., *Structure of influenza haemagglutinin at the pH of membrane fusion.* 1994. **371**(6492): p. 37-43.

117. Benton, D.J., et al., *Structural transitions in influenza haemagglutinin at membrane fusion pH.* 2020. **583**(7814): p. 150-153.

118. Skehel, J.J. and D.C.J.A.r.o.b. Wiley, *Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin.* 2000. **69**(1): p. 531-569.

119. Schulze, I.T.J.J.o.I.D., *Effects of glycosylation on the properties and functions of influenza virus hemagglutinin.* 1997. **176**(Supplement_1): p. S24-S28.

120. Vigerust, D.J., et al., *N-linked glycosylation attenuates H3N2 influenza viruses.* 2007. **81**(16): p. 8593-8600.

121. Sun, S., et al., *Glycosylation site alteration in the evolution of influenza A (H1N1) viruses.* 2011. **6**(7): p. e22844.

122. Soema, P.C., et al., *Current and next generation influenza vaccines: Formulation and production strategies.* 2015. **94**: p. 251-263.

123. Ekiert, D.C., et al., *Antibody recognition of a highly conserved influenza virus epitope.* 2009. **324**(5924): p. 246-251.

124. Yassine, H.M., et al., *Use of hemagglutinin stem probes demonstrate prevalence of broadly reactive group 1 influenza antibodies in human sera.* 2018. **8**(1): p. 8628.

125. Valkenburg, S.A., et al., *Stalking influenza by vaccination with pre-fusion headless HA mini-stem.* 2016. **6**(1): p. 22666.

126. Ekiert, D.C. and I.A.J.C.o.i.v. Wilson, *Broadly neutralizing antibodies against influenza virus and prospects for universal therapies.* 2012. **2**(2): p. 134-141.

127. Nath Neerukonda, S., R. Vassell, and C.D.J.V. Weiss, *Neutralizing antibodies targeting the conserved stem region of influenza hemagglutinin.* 2020. **8**(3): p. 382.

128. Krammer, F., et al., *Chimeric hemagglutinin influenza virus vaccine constructs elicit broadly protective stalk-specific antibodies.* 2013. **87**(12): p. 6542-6550.

129. Wohlbold, T.J., et al., *Vaccination with soluble headless hemagglutinin protects mice from challenge with divergent influenza viruses.* 2015. **33**(29): p. 3314-3321.

130. Bullard, B.L. and E.A.J.V. Weaver, *Strategies targeting hemagglutinin as a universal influenza vaccine.* 2021. **9**(3): p. 257.

131. Ellebedy, A. and R.J.V. Webby, *Influenza vaccines.* 2009. **27**: p. D65-D68.

132. Chen, M.-W., et al., *A consensus–hemagglutinin-based DNA vaccine that protects mice against divergent H5N1 influenza viruses.* 2008. **105**(36): p. 13538-13543.

133. Wu, P., et al., *Single dose of consensus hemagglutinin-based virus-like particles vaccine protects chickens against divergent H5 subtype influenza viruses.* 2017. **8**: p. 1649.

134. Webby, R.J. and E.A.J.P.o. Weaver, *Centralized consensus hemagglutinin genes induce protective immunity against H1, H3 and H5 influenza viruses.* 2015. **10**(10): p. e0140702.

135. Lingel, A., B.L. Bullard, and E.A.J.S.r. Weaver, *Efficacy of an adenoviral vectored multivalent centralized influenza vaccine.* 2017. **7**(1): p. 14912.

136. Carter, D.M., et al., *Design and characterization of a computationally optimized broadly reactive hemagglutinin vaccine for H1N1 influenza viruses.* 2016. **90**(9): p. 4720-4734.

137. Reneer, Z.B., et al., *Computationally optimized broadly reactive H2 HA influenza vaccines elicited broadly cross-reactive antibodies and protected mice from viral challenges.* 2020. **95**(2): p. 10.1128/jvi. 01526-20.

138. Wong, T.M., et al., *Computationally optimized broadly reactive hemagglutinin elicits hemagglutination inhibition antibodies against a panel of H3N2 influenza virus cocirculating variants.* 2017. **91**(24): p. 10.1128/jvi. 01581-17.

139. Nuñez, I.A., Y. Huang, and T.M.J.P. Ross, *Next-generation computationally designed influenza hemagglutinin vaccines protect against H5Nx virus infections.* 2021. **10**(11): p. 1352.

140. Skarlupka, A.L., et al., *Computationally optimized broadly reactive vaccine based upon swine H1N1 influenza hemagglutinin sequences protects against both swine and human isolated viruses.* 2019. **15**(9): p. 2013-2029.

141. Giles, B.M. and T.M.J.V. Ross, *A computationally optimized broadly reactive antigen (COBRA) based H5N1 VLP vaccine elicits broadly reactive antibodies in mice and ferrets.* 2011. **29**(16): p. 3043-3054.

142. Giles, B.M., et al., *Antibody breadth and protective efficacy are increased by vaccination with computationally optimized hemagglutinin but not with polyvalent hemagglutinin-based H5N1 virus-like particle vaccines.* 2012. **19**(2): p. 128-139.

143. Ross, T.M., et al., *A computationally designed H5 antigen shows immunological breadth of coverage and protects against drifting avian strains.* 2019. **37**(17): p. 2369-2376.

144. Giles, B.M., et al., *A computationally optimized hemagglutinin virus-like particle vaccine elicits broadly reactive antibodies that protect nonhuman primates from H5N1 infection.* 2012. **205**(10): p. 1562-1570.

145. Crevar, C.J., et al., *Cocktail of H5N1 COBRA HA vaccines elicit protective antibodies against H5N1 viruses from multiple clades.* 2015. **11**(3): p. 572-583.

146. Huang, Y., et al., *Next Generation of computationally optimized broadly reactive ha vaccines elicited cross-reactive immune responses and provided protection against H1N1 virus infection.* 2021. **9**(7): p. 793.

147. Shu, Y. and J.J.E. McCauley, *GISAID: Global initiative on sharing all influenza data–from vision to reality.* 2017. **22**(13): p. 30494.

148. Shen, W., et al., *SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation.* 2016. **11**(10): p. e0163962.

149. Edgar, R.C.J.N.a.r., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* 2004. **32**(5): p. 1792-1797.

150. Bakan, A., L.M. Meireles, and I.J.B. Bahar, *ProDy: protein dynamics inferred from theory and experiments.* 2011. **27**(11): p. 1575-1577.

151. Cock, P.J., et al., *Biopython: freely available Python tools for computational molecular biology and bioinformatics.* 2009. **25**(11): p. 1422.

152. Saitou, N., M.J.M.b. Nei, and evolution, *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* 1987. **4**(4): p. 406-425.

153. Sneath, P.H. and R.R. Sokal, *Numerical taxonomy. The principles and practice of numerical classification*. 1973.

154. Le Cam, L.J.I.S.R.R.I.d.S., *Maximum likelihood: an introduction.* 1990: p. 153-171.

155. Stamatakis, A.J.B., *RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.* 2014. **30**(9): p. 1312-1313.

156. Letunic, I. and P.J.B. Bork, *Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation.* 2007. **23**(1): p. 127-128.

157. Castresana, J.J.O.v.a.a.h.m.c.c.e.c.G.s.h., *Gblocks, v. 0.91 b.* 2002.

158. Olson, R.D., et al., *Introducing the bacterial and viral bioinformatics resource center (BV-BRC): a resource combining PATRIC, IRD and ViPR.* 2023. **51**(D1): p. D678-D689.

159. Flannery, B., et al., *Enhanced genetic characterization of influenza A (H3N2) viruses and vaccine effectiveness by genetic group, 2014–2015.* 2016. **214**(7): p. 1010-1019.