Rational Vaccine Design by Reverse & Structural Vaccinology and Ontology

by

Edison Ong

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2021

Doctoral Committee:

Associate Professor Yongqun He, Chair
Professor Denise Kirschner
Professor Kayvan Najarian
Associate Professor Zhenhua Yang
Professor Yang Zhang

Edison Ong

edong@umich.edu

ORCID iD: 0000-0002-5159-414X

Acknowledgments

First of all, I would like to acknowledge my advisor Dr. Yongqun (Oliver) He, for his admirable mentorship. He trained me to be a motivated and well-organized person throughout my pursuit of the doctoral program. I appreciate the opportunity to work in his laboratory and contribute to scientific research. With his support, I designed and completed many projects, resulting in multiple publications. I feel thankful that he showed me how to be independent and creative to all the possibilities as a scientist. My special thanks go to the former Dr. He laboratory members, Dr. Sirarat Sarntivijai, Dr. Asiyah Yu Lin, Dr. Rebecca Racz, and Zuoshuang Xiang, for their supports and advice that are valuable throughout my dissertation work.

The committee member's help and guidance have maximized my potential to finish the dissertation. They supported my interdisciplinary research involving machine learning, structural biology, microbiology, immunology, and epidemiology.

My sincere thanks go to Dr. Zhenhua Yang, and she provided me many grateful supports. She showed her enthusiasm for science and gave me indispensable advice for my project and future career. Dr. Yang offered me many collaboration opportunities in tuberculosis related projects. In particular, Dr. Yang introduced me to work with the Michigan Department of Health & Human Services (MDHHS) to process Mycobacterium tuberculosis whole-genome sequence data. I also want to thank Dr. Marty Soehnlen and the MDHHS laboratory staff for their warm supports.

I would like to recognize the invaluable assistance that Dr. Yang Zhang has given to me on structural biology and its application in vaccine design. Dr. Zhang has provided much valuable advice that is critical to my dissertation work. I would also like to thank Dr. Xiaoqiang Huang and Mr. Robin Pearce for their huge contribution to my structural vaccinology work.

I wish to show my gratitude to Dr. Kayvan Najarian and Dr. Denise Kirschner for their insightful advice in machine learning and tuberculosis studies. Dr. Najarian offered his knowledgeable machine learning experience and guided me throughout my design of machine learning-based reverse vaccinology work. Dr. Kirschner provided many positive and warm suggestions that facilitate understanding in immunology and professional networking in the field.

I would like to express my gratitude to the Department of Computational Medicine and Bioinformatics and Unit for Laboratory Animal Medicine (ULAM), the advisors and staff, Dr. Margit Burmeister, Dr. Maureen Sartor, Dr. Robert C. Dysko, and Dr. Jeffrey de Wet, and Julia Eussen. I appreciate the financial support from the National Institutes of Health (NIH) Integrated Network-based Cellular Signatures (LINCS), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Kidney Precision Medicine Project (KPMP), National Institute of Allergy and Infectious Diseases (NIAID) Immunology Database and Analysis Portal (ImmPort), and University of Michigan Rackham Predoctoral Fellowship and ULAM bridge fund. I also appreciate the European Bioinformatics Institute's support and funding during my summer internship in the United Kingdom in 2017.

I am grateful that Dr. Giovanni Paternostro and Dr. Carlos Piermarocchi offered me a research programmer position after my undergraduate degree. With my valuable bioinformatics research experience throughout my two years of work, I eventually pursued a doctoral degree in Bioinformatics.

The experience of the whole Ph.D. and writing the dissertation is not easy for me, and I appreciate the help from all the people around me. I would like to show my deepest gratitude to my wife, Mei U, who has been a great accompany in my journey. We met each other in kindergarten, and it was her encouragement leading to my passion for biology. We have been supporting, encouraging, and educating each other from high school to university. She has been, and always will be, my Sirius in the night. I would also thank my aunts, Elisa and Kate, who fully supported me when I decided to apply for a doctoral degree. They provided me with everything needed to succeed in my academic and career. Finally, I would like to dedicate my dissertation to my mother, Emily, who inspired my scientific curiosity. She always gave me numerous supports and precious memory in my childhood. I wish she would be here to see me finish the doctoral degree. Still, I am assured that she will be proud of her boy being a diligent and dedicated person.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

RV	reverse vaccinology
SV	structural vaccinology
PAg	protective antigen
VF	virulence factor
CD	cluster of differentiation
TCR	T-cell receptor
MHC	Major histocompatibility complex
MenB	Group B *meningococcus*
SCL	subcellular localization
BPAgs	bacterial protective antigens
TMH	transmembrane α-helices
TMB	transmembrane β-barrel
AP	adhesin probability
COG	Clusters of Orthologous Groups
G+	Gram-positive
G-	Gram-negative
GO	Gene Ontology
BP	biological process
MF	molecular function

CC	cellular component
ML	machine learning
SMOTE	Synthetic Minority Over-sampling technique
LR	logistic regression
SVM	support vector machine
KNN	k-nearest neighbor
RF	random forest
XGB	extreme gradient boosting
N5CV	nested five-fold cross-validation
LOPOV	leave-one-pathogen-out validation
ROC	receiver operating characteristics
PR	precision-recall
WF1	weighted F1-score
MCC	Matthew’s correlation coefficient
COVID-19	Coronavirus Disease 2019
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2
HCoV	human coronaviruses
S	spike protein
RBD	receptor-binding domain
NTD	N-terminal domain
N	nucleocapsid protein
M	membrane protein

E	envelope protein
VLP	virus-like particles
ACE2	angiotensin-converting enzyme 2
MHC	major histocompatibility complex
SASAr	solvent accessible surface area ratio
ECS	epitope content score
HSS	human similarity score
JSD	Jensen-Shannon divergence
EEU	EvoDesign energy unit
MD	molecular dynamics
HPI	host-pathogen interaction

Abstract

Vaccination is one of the most successful public health interventions in modern medicine. However, it is still challenging to develop effective vaccines against many infectious diseases such as tuberculosis, HIV, and malaria. There are challenges in integrating the high volume, variety, and variability of vaccine-related data and rationally designing effective and safe vaccines efficiently. In my thesis study, I systematically and comprehensively analyzed manually annotated protective vaccine antigens in the Protegen database and identified these protective antigens' enriched patterns. I then created Vaxign-ML, a novel machine learning-based reverse vaccinology method based on the curated Protegen data for rational vaccine design. Vaxign-ML was used to successfully predict vaccine antigens for tuberculosis and Coronavirus Disease 2019 (COVID-19). I also developed a new structural vaccinology design program that optimizes COVID-19 spike glycoprotein as a vaccine candidate for enhanced vaccine protection via T cell epitope engineering. The vaccine antigens selected and optimized by Reverse and Structural Vaccinology in this dissertation are subjected to future experimental verification. Furthermore, I created a community-based Ontology of Host-Pathogen Interactions (OHPI), which served as a platform to semantically represent the interactions between host and virulence factors that are also protective antigens. I developed the Vaccine Investigation Ontology (VIO) for standardized metadata representation for high throughput vaccine OMICS data analysis. Overall, my thesis research aims to uncover protective antigen patterns, create methods/tools to effectively develop vaccines against infectious diseases of public health significance, and strengthen our understanding of vaccine protection mechanisms. These works can be further expanded and

integrated with other technologies such as epitope prediction, molecular epidemiology, and high-throughput sequencing to build the foundation of precision vaccinology.

Chapter 1 Introduction

Civilization has been associated with epidemics and pandemics throughout entire human history. The first recorded pandemic can be traced back to the plague of Athens in 430 B.C., which killed about two-thirds of the population in the cities and eventually led to the downfall of the Athenian empire (Cartwright and Biddiss 2000). The bubonic plague caused three recorded worldwide pandemics: the first Plague of Justinian in 540-590 A.D. killed millions; the second "black death" took at least 20 million Europeans' life in the years 1347-1352; and its third and final appearance in a pandemic from 1894 to 1950 claimed 15 million lives (Byrne 2008; Cartwright and Biddiss 2000; Echenberg 2002). As we progressively advance our understanding of diseases (e.g., plague, measles, cholera, and leprosy), more and better interventions are being created to cure and prevent these contagious and yet lethal diseases. One of the most successful medical interventions ever introduced in modern medicine is the vaccine.

The earliest written record of the vaccine usage in China can be traced back to the fifteenth century, where variolation was used to prevent the fatal childhood disease smallpox (variola virus) (L. Zhang 1709). One of the variolation methods described was to apply fresh pustule or squama from the sick child to the nostril of the healthy child to cause a more benign smallpox infection and prevent them from getting the disease again. The variolation was then started to be widely adopted in Europe from 1710 (Cartwright and Biddiss 2000). However, such an inoculation method was not always successful with a 2-3% mortality rate and could cause the spreading of the disease. In 1796, Edward Jenner introduced cowpox inoculation to prevent smallpox infection. The term cowpox inoculation is referred to as vaccination (variolae vaccinae means smallpox of the cow) and was the first vaccine developed against infectious disease. Since then, the vaccine is one of the most significant medical inventions for public health and has been the crucial component to effectively and efficiently control an epidemic or pandemic. The 1957 flu pandemic, caused by a new influenza A (H2N2) virus, claimed an estimated 1.1 million lives in 39 countries worldwide before a vaccine was developed to effectively contain the pandemic (Viboud et al. 2015). Vaccines also led to the global smallpox eradication (Belongia and Naleway 2003) and elimination of Polio in most countries (Bahl et al. 2018) and saved 122 million children's lives since 1990 by reducing childhood death caused by infectious diseases. It is estimated that for every dollar spent on childhood immunization, the society receives 44 dollars of economic benefits (Gates and Gates 2017).

1.1 Challenges in Vaccine-preventable Diseases and Outbreaks

However, it is still difficult to develop vaccines against many infectious diseases of global public health importance, such as tuberculosis. Tuberculosis (TB) epidemic has an estimated 10 million cases and claimed over 1.4 million lives in 2019 (World Health Organization 2020a). The End TB Strategy initiated by the World Health Organization aims to halt the global TB epidemic by 2035 and reduce TB incidence by 90% between 2015 and 2035 (Uplekar et al. 2015). The live-attenuated *Mycobacterium bovis* vaccine designated Bacillus Calmette-Guerin (BCG) vaccine is the only TB vaccine currently licensed. However, it provided minimal protection against adult pulmonary TB and the reactivation of latent TB infection (LTBI) (Fine 1995). Approximately a quarter of the world's population is estimated to be latently infected (Houben and Dodd 2016). Therefore, a safe and effective TB vaccine to prevent primary infection of *Mycobacterium tuberculosis* (MTB) and the reactivation of LTBI is a key to address the challenge facing global TB elimination (Izzo 2017).

Besides the already existed infectious diseases, human also continuously faces the threat of emerging diseases. In 2003, the SARS disease caused by the SARS-associated coronavirus (SARS-CoV) infected over 8,000 people worldwide and was contained in the summer of 2003 (Lu et al. 2020). The MERS disease infected more than 2,000 people, caused by the MERS-associated coronavirus (MERS-CoV), and was first reported in Saudi Arabia and spread to several other countries since 2012 (Chan et al. 2015). The emerging Coronavirus Disease 2019 (COVID-19) pandemic poses a massive crisis to global public health, and WHO declared the COVID-19 as a pandemic on March 11, 2020. The causative agent of COVID-19 is SARS-CoV-2, which shares a high sequence identity with SARS-CoV (Lai et al. 2020). As of November 19, 2020, this on-going COVID-19 pandemic caused over 55 million infection cases and over one million deaths globally. With the advance of transportation and globalization, we face a huge challenge of a potential epidemic and even pandemic in the future. Advance methods are in earnestly demand to develop vaccines quickly and effectively, as a response to the ever-increasing threat of existing and emerging infectious diseases.

1.2 Immunity and Vaccines

Immunity is the ability to distinguish “self” and “non-self” material to eliminate the “non-self” material (Delves et al. 2016). This “non-self” material is often referred to as antigens, which are the parts of molecules (e.g., proteins) from the disease-causing microorganisms (also known as pathogens). Human has developed a sophisticated system of interacting with cells to identify antigens and remove the “non-self” substances and pathogens. Immunity can be classified into innate immunity and adaptive immunity. Innate immunity is non-specific immunity that is antigen-independent and mostly mounts an immediate but short-living immune response against the “non-self” objects. On the other hand, adaptive immunity, which is the major mechanism of vaccines, is antigen-specific immune responses and often involves the production of antibodies by the B-lymphocytes (also known as humoral immunity) and specific immune cells, including T-lymphocytes (also known as cell-mediated immunity). Adaptive immunity also produces memory immune cells (B and/or T cells) so that a re-exposure to the same pathogen will trigger a more robust and rapid immune response or secondary immunity. This type of immunity can last for months, years, or often a lifetime. The vaccine takes advantage of the adaptive immunity by mounting a primary immunity and stimulates the production and formation of such memory immune cells without causing the disease.

The classic definition of vaccines is the preparation of live-attenuated or killed/inactivated microorganisms (e.g., bacteria and viruses) administered to produce or artificially increase immunity to a particular disease (S A Plotkin, Orenstein, and Offit 2012). After Edward Jenner created the first vaccine, Louis Pasteur generalized vaccines to preventable diseases other than smallpox (Stanley A. Plotkin 2005). He discovered that the chicken cholera bacteria lost its disease-causing properties (virulence) after a few generations in culture. The weakened form of the pathogen is called live-attenuated vaccines, which can teach the host immune system to fight the infection without suffering severe symptoms. In the meantime, Daniel Salmon and Theobald Smith created the first killed/inactivated vaccine for cholera. The inactivated vaccine can be prepared by killing the pathogens using physical (e.g., heat) or chemical (e.g., formalin) treatments. But the “inactivated” pathogens can still protect against infections (Stanley A. Plotkin 2005). The preparation of live-attenuated and inactivated vaccines corresponds to the conventional vaccine development strategy, a very time-consuming process.

With the advance of technologies, the modern vaccine classification includes vaccines containing proteins, polysaccharides, or nucleic acids (DNA/RNA) of the pathogens (disease-causing microorganisms) that are delivered as single entities, as part of complex (e.g., nano and virus-like) particles, or as living attenuated phages or vectors to induce specific immune responses that inactivate, destroy, or suppress the pathogen (S A Plotkin, Orenstein, and Offit 2012). The subunit vaccines are made from purified proteins or polysaccharides of microorganisms. Recombinant vaccines and nucleic acid vaccines are produced using DNA/RNA derived (directly or indirectly) from an organism that codes for a protective protein, which can be in the format of a purified expressed protein (recombinant subunit vaccine) or carried by a vector (recombinant vector vaccine). These types of vaccines only make up a portion of the pathogens and are usually safer than live-attenuated and inactivated vaccines. However, the immune-inducing ability, or immunogenicity, of subunit and recombinant vaccines are often weaker and require adjuvants to boost the immune responses. An essential step in developing subunit and recombinant vaccines is selecting the vaccine antigen candidate, usually the protein(s) or the nucleic acid expressing the protein(s). In this thesis, the prediction and selection of vaccine antigen candidates (Chapters 3 and 4) and the optimization of these candidates' immunogenicity (Chapter 5) were explored and investigated.

1.3 Reverse Vaccinology

The conventional method of this selection is performed via *in vitro* screening in the laboratory. However, this method is resource-consuming, and that not all the pathogens can be cultured, and not all the proteins can be purified for testing. Since the early 1990s, the advance of high-throughput sequencing technology has fostered an innovative genome-based vaccine design approach, termed Reverse Vaccinology (RV) (Rappuoli 2000). The first RV study identified vaccine candidates against the meningitis B disease from the whole genome sequences of the disease-causing bacteria (Pizza et al. 2000). Within a relatively short period, this study has led to the licensing of the meningitis B vaccine, Bexsero®, in the United States since 2015 (Folaranmi et al. 2015). The first RV study's great success has led to many RV prediction programs (Dalsass et al. 2019).

The existing open-source RV prediction programs could be characterized based on the algorithmic approaches or input feature types (Figure 1-1). The algorithmic approaches include rule-based filtering (or decision tree-like) and machine learning (ML) classification methods. The first publicly available rule-based filtering RV program is NERVE (Vivona, Bernante, and Filippini 2006), available as a standalone software program. In 2010, Vaxign was developed as the first web-based filtering RV program with additional analyses (e.g., adhesin probability and similarity to host) (He, Xiang, and Mobley 2010). Vaxign has been applied to predict vaccine candidates for more than ten pathogenic bacteria such as *Helicobacter pylori* (Navarro-Quiroz et al. 2018), *Acinetobacter baumannii* (Singh et al. 2016), *Mycobacterium spp.* (Hossain et al. 2017). From 2013 to 2017, two additional filtering-based RV programs, Jenner-predict server (Jaiswal et al. 2013) and VacSol (Rizwan et al. 2017) were also created and included more bioinformatics analyses, including conserved domains and biological pathways. However, all these currently available rule-based filtering RV programs use only biological features as the data input, and the prediction performance of all RV programs are not satisfactory.

With the advance of machine learning (ML) and the accumulation of vaccine data over the past decades, there is a need to develop the next-generation ML-based RV tool. Machine learning, as defined by a pioneer in artificial intelligence Arthur Samuel, is the "field of study that gives computers the ability to learn without being explicitly programmed" (Samuel 2000; Kohavi and Provost 1998). The ability to make predictions or classifications is achieved by the ML algorithms building a model based on the input data (also known as training data). ML classification has also been applied in many biomedical studies, including protein structure prediction (W. Zheng et al. 2019; Alquraishi 2019), drug discovery (Tran et al. 2014; Vamathevan et al. 2019), as well as vaccine candidate prediction in RV (Dalsass et al. 2019). VaxiJen was the first ML classification RV program published in 2007 (Doytchinova and Flower, 2007). Bowman et al. and Heinson et al. improved the work of VaxiJen by extending the training data of VaxiJen and revising the ML algorithm (Bowman et al. 2011; Heinson et al. 2017). A key difference between VaxiJen and the Bowman-Heinson method was that VaxiJen used physicochemical features of the input proteins while the later program used biological features.

Significant effort has been made to enhance ML-based RV prediction performance, but there is still much room for improvement. All existing methods use either biological properties or physicochemical properties of the proteins in the training data. In Chapter 2, the relation between the protectiveness of BPAgs and biological properties was investigated, and significant correlations were reported for properties such as subcellular localization, adhesin probability, and peptide signaling (E. Ong, Wong, and He 2017). On the other hand, physicochemical properties were reported to have significantly associated with BPAg protection (Mayers et al. 2003). ML models trained with physicochemical properties data showed high BPAg prediction accuracy (Dalsass et al. 2019). Therefore, a combination of biological and physicochemical properties is likely to enhance the prediction performance of BPAg prediction further. Second, there is a lack of high-quality benchmarking datasets, and the performance of various ML-based RV software programs has not been systematically evaluated. The training data from VaxiJen and the Bowman-Heinson method only included PAgs with supporting experimental evidence. The negative samples were randomly selected from non-homologous proteins to the PAgs. Such random under-sampling may not reflect the real distribution of PAgs in the pathogen proteomes. Also, proteins with low sequence similarity to the known PAgs can potentially still induce protective immune responses. Besides, VaxiJen and Bowman-Heinson model was not evaluated using an external independent dataset. In Chapter 3, the ML-based Vaxign program, Vaxign-ML, was developed and systematically evaluated with existing RV tools (including rule-based and ML-based methods) via a high-quality benchmarking dataset. Vaxign-ML was trained on the BPAgs with their biological and physicochemical features annotated. The BPAgs and the non-protective proteins were first carefully checked for homology to ensure training data quality. Three evaluation steps, including nested five-fold cross-validation, leave-one-pathogen-out validation, and independent benchmarking, were implemented. Vaxign-ML demonstrated superior predictive performance to all existing RV methods. In Chapter 4, Vaxign-ML was also applied to predict vaccine candidates for COVID-19 vaccine development.

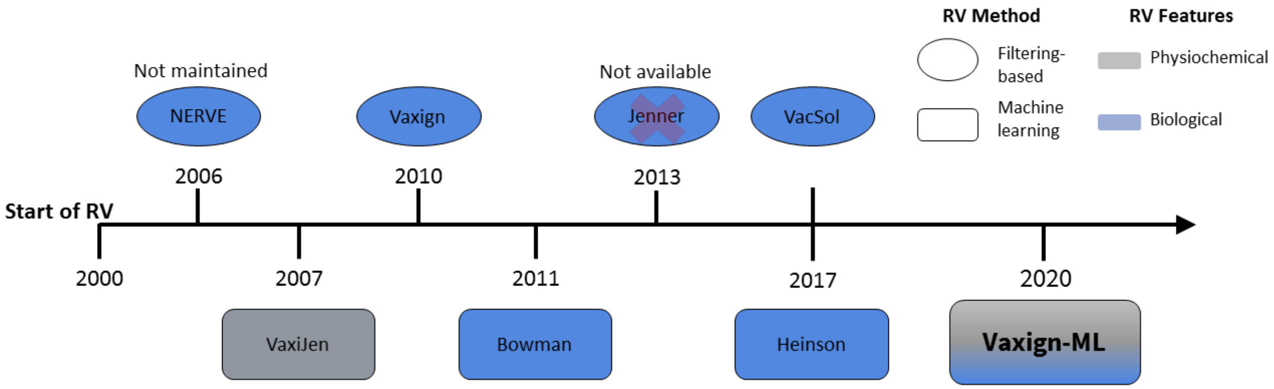


Figure 1-1 Reverse Vaccinology (RV) tools development timeline.

All the existing open-source RV tools are listed. Each can be categorized based on i) type of RV software and ii) RV software interface. The oval frame represents the filtering-based, and square frame represents the machine learning (ML) based RV tools. The background color indicates whether the methods utilizing biological features (grey) and/or blue physicochemical properties (blue) of the input proteins. In 2020, Vaxign-ML was created as an ML-based RV tool that incorporates both the input proteins' biological and physicochemical properties and provides terminal and web interfaces.

1.4 Structural Vaccinology

Structural vaccinology (SV) also emerges as a revolutionary vaccine design method to engineer vaccine candidates based on 3D structure. The first proof-of-concept study enhanced the immunogenicity of the fusion (F) glycoprotein of respiratory syncytial virus (RSV) by fixing the conformation-dependent neutralization-sensitive epitopes (B. S. Graham, Modjarrad, and McLellan 2015). RSV is a leading cause of infant mortality and adult morbidity, but there is currently no licensed RSV vaccine. The F glycoprotein contributes to the membrane fusion of RSV and the host cell and is a primary target for vaccine development. For decades, researchers have been using the post-fusion F glycoprotein as a vaccine candidate, but it does not provide protection in challenge studies. An investigation into the conformational rearrangement of this protein between its metastable pre-fusion and stable post-fusion identified a change of its epitope content in these two conformations (Figure 1-2) (McLellan et al. 2013). Epitopes are the specific units of an antigen recognized by the immune system. The less stable form of the pre-fusion F glycoprotein has more and better epitopes. Therefore, a vaccine candidate that fixed the F glycoprotein in its pre-fusion conformation induced a more potent neutralizing antibody response. The discovery of this structure-based approach has revolutionized vaccine development.

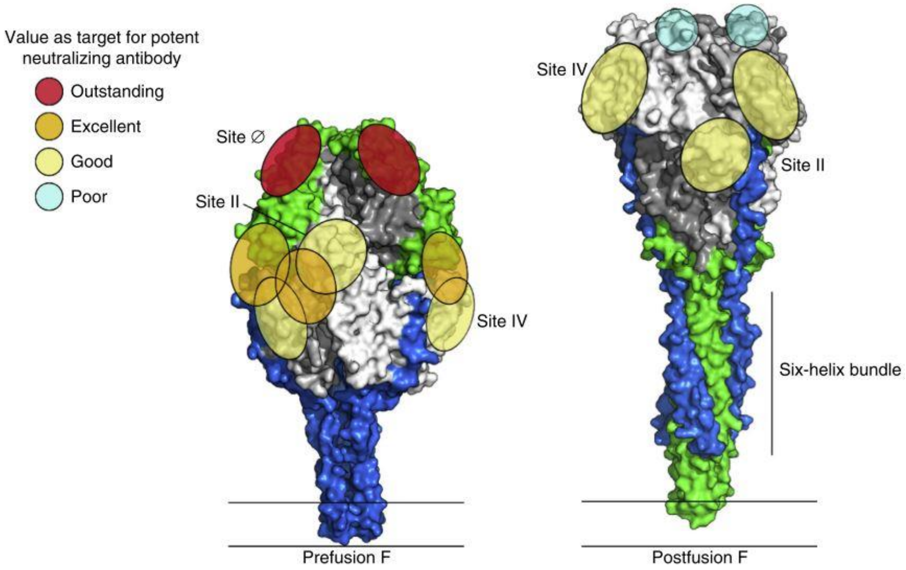


Figure 1-2 Surface representation of respiratory syncytial virus (RSV) fusion (F) glycoprotein.

The RSV F glycoprotein exists in pre-fusion and post-fusion conformations. The alteration between two conformations occurs as part of the membrane fusion of the virus entry to the host cell. Both conformations consist of a different set of epitopes, but the most neutralization-inducing epitopes are only present in pre-fusion form, which is relatively less stable than the post-fusion F glycoprotein. This figure is reprinted from (B. S. Graham, Modjarrad, and McLellan 2015) published in *Current Opinion in Immunology* with permission from Elsevier.

In light of the RSV vaccine development's success, structural vaccinology has been applied to design vaccines for other pathogens, particularly SARS-CoV-2. The COVID-19 pandemic has sparked an unprecedented race to develop a safe and effective vaccine to contain the widespread outbreak. Similar to the RSV, the spike (S) glycoprotein of the SARS-CoV-2 plays a crucial role in mediating virus entry, and many computational studies utilizing reverse vaccinology and immuno-informatics reported the S protein to be a promising vaccine antigen (E. Ong, Wong, Huffman, and He 2020; Grifoni, Sidney, et al. 2020; Enayatkhani et al. 2020). Clinical studies also identified anti-S protein neutralizing antibodies in patients recovered from COVID-19 (F. Wu et al. 2020; L. Ni et al. 2020; Cao et al. 2020). Therefore, S protein has been the primary target of many vaccines currently in clinical trials. Since the cryo-EM structure of the S protein (Wrapp, Wang, et al. 2020) and the neutralizing antibodies that bind to the S protein (Barnes et al. 2020; Wrapp, De Vlieger, et al. 2020) were determined, structural vaccinology approaches have been applied to optimize the S protein structure as a vaccine candidate. For example, Henderson et al. controlled the S protein's receptor-binding domain (RBD) domain between the "up" and "down" configurations to induce immunogenicity (Henderson et al. 2020). On the other hand, structural modifications were also performed on the native S protein to stabilize the S protein in its pre-fusion form (Bos et al. 2020), a strategy similar to the RSV vaccine development.

However, these modifications have focused on the humoral immunity to induce neutralizing antibodies targeting the SARS-CoV-2 S protein. Studies have also shown the importance of the cluster of differentiation 4 (CD4) T cell response in controlling SARS-CoV-2 infection and possible pre-existing immunity in healthy individuals without exposure to SARS-CoV-2 (Grifoni, Weiskopf, et al. 2020; Bert et al. 2020; Braun et al. 2020). T cell response can be primarily categorized into CD4 and CD8 T cell responses. CD4 is a glycoprotein located at the surface of immune cells, including T cells. It serves as a co-receptor for the T-cell receptor (TCR) to interact with the epitopes presented by the major histocompatibility complex (MHC) class II molecule. CD4 T cells (often referred to as T-helper cells) play a major role in releasing signals to aid both humoral and cell-mediated responses and the induction of long-term memory. CD8 is also a glycoprotein predominantly found on cytotoxic T cells' surface and plays a major role in the cell-mediated response. CD8 T cell interacts with cells presenting epitope bound to MHC class I molecule and induces the programmed cell death of the presenting cells (Delves et al. 2016). A successful COVID-19 vaccination is likely linked to a robust and long-term humoral response to the SARS-CoV-2 S protein with the help of CD4 T cells.

Contrary to the enhancement of immunogenicity for vaccine candidates, there is another branch of structural vaccinology that aims to reduce the immunogenic property of therapeutic proteins and avoid auto-immune response. This auto-immune response is an adverse event of the therapeutic proteins that cause the immune system to target “self” contrary to the “non-self” material. The process of reducing the immunogenicity of therapeutic proteins is referred to as “deimmunization”. There are many reported methods for deimmunizing therapeutic proteins. For example, EpiSweep was developed to reduce the epitope content of *Staphylococcus simulans* lysostaphin, which is an effective staphylococcal bacteriocin to treat drug-resistant *Staphylococcus aureus* infection (Blazanovic et al. 2015). Baker et al. used the Rosetta suite to deimmunize the fluorescent reporter protein super-folder GFP and *Pseudomonas* exotoxin A by reducing the number of MHC-II restricted CD4 T cell epitopes while preserving the proteins’ function (King et al. 2014). The deimmunization of a protein involved reducing MHC-II restricted CD4 T cell epitopes to reduce the proteins’ immunogenicity.

If we flip how the deimmunization process works, the induction of MHC-II restricted CD4 T cell epitopes can be applied to vaccine design to enhance the immunogenicity of the vaccine candidate. The addition of a single epitope to induce stronger immune responses has also been applied to develop H7N9 vaccines. The H7N9 hemagglutinin (HA) vaccines elicited non-neutralizing antibody responses in clinical trials (Mulligan et al. 2014; Guo et al. 2014). Rudenko et al. reported fewer CD4 T cell epitopes found in H7N9 HA than the seasonal H1 and H3 HA proteins (Rudenko et al. 2016). Based on this finding, Wada et al. improved the H7N9 vaccine by introducing a known H3 immunogenic epitope to the H7 HA protein without perturbing its conformation, which resulted in an over 4-fold increase in the HA-binding antibody response (Wada et al. 2017). Therefore, in Chapter 5, I developed a structural vaccinology approach to rationally design the SARS-CoV-2 S protein by generating thousands of stable S protein variants without perturbing the protein's surface conformation to maintain the same B cell epitope profile. In the meantime, mutations were introduced to the residues buried inside the S protein so that more MHC-II restricted CD4 T cell epitopes would be added into the newly designed S protein to potentially induce a stronger immune response.

1.5 Vaccine-informatics and Ontology

With the exponential growth of data accumulated in vaccine informatics, there is a challenge to translate the high volume, variety, and variability of vaccine-related data in the era of “big-data”. The Vaccine Investigation and Online Information Network (VIOLIN) (He et al. 2014) is the most comprehensive database collected and curated vaccine-related information. However, such a vaccine-focused database often consists of data targeting different aspects of vaccine mechanisms such as virulence factors and protective antigens. Virulence factors (VFs) are molecules that allow microbial pathogens to overcome host defense mechanisms and cause disease in a host. A total of 5,304 VFs supported by experimental evidence (e.g., loss or reduction of pathogenicity in the host after the VF gene mutation) are curated and stored in the Victors database (Sayers et al. 2019). On the other hand, some of these VFs are also used as PAgs for vaccine development due to their important roles in pathogenicity and protective antigenicity. Protegen contains a set of 590 protective antigens (PAgs) over 100 infectious diseases caused by pathogens (bacteria, viruses, and parasites) and non-infectious diseases, including cancers and allergies (B. Yang et al. 2011). These PAgs are manually collected and curated from the literature with supporting experimental evidence (e.g., protection assay against a challenge or immune response assay correlates with protection). There is rich data for vaccine research, but it requires an organization of these vaccine-related data in a computationally tractable manner to predict vaccine candidates and provide insights into the mechanistic drivers of vaccine protection.

Ontology has emerged to be a feasible approach to integrate and synthesize knowledge from data. Ontology is a computer- and human-interpretable representation of the entities and the relations among objects. Ontology can facilitate the integration of vaccine-related data from distinct domains (e.g., host vs. pathogen, gene vs. protein) and capture the connections among these data within the VIOLIN database to represent knowledge. A set of ontology development and visualization tools is needed to develop ontologies efficiently (E. Ong et al. 2017; He, Xiang, et al. 2018; Z. Xiang et al. 2010). In Chapter 6, these tools were implemented to create two ontologies, Ontology of Host-Pathogen Interactions (OHPI) and Vaccine Investigation Ontology (VIO). These ontologies were then applied to standardize and analyze the vaccine-related data from the VIOLIN database. The relations among the VIOLIN data were also defined and modeled in these ontologies to facilitate data integration and analysis.

1.6 Dissertation Outline

Overall, my thesis research aims to uncover protective antigen patterns, create methods/tools to effectively develop vaccines against infectious diseases of public health significance, and strengthen our understanding of vaccine protection mechanisms (summarized in Figure 1-3). In Chapter 2, a systematic analysis was conducted to identify a significant correlation between the protectiveness of vaccine candidates and these proteins' biological properties. In Chapter 3, a novel machine learning-based reverse vaccinology tool, Vaxign-ML, was developed to select vaccine candidates, and in chapter 4, Vaxign-ML is applied to predict COVID-19 vaccine candidates. In Chapter 5, a novel structural vaccinology tool is created to optimize the structure of the COVID-19 vaccine candidate, spike glycoprotein, for better vaccine protection. In Chapter 6, two ontologies, Vaccine Investigation Ontology (VIO) and Ontology of Host-Pathogen Interactions (OHPI) were created to facilitate vaccine data integration to advance our understanding of vaccine protection mechanism.

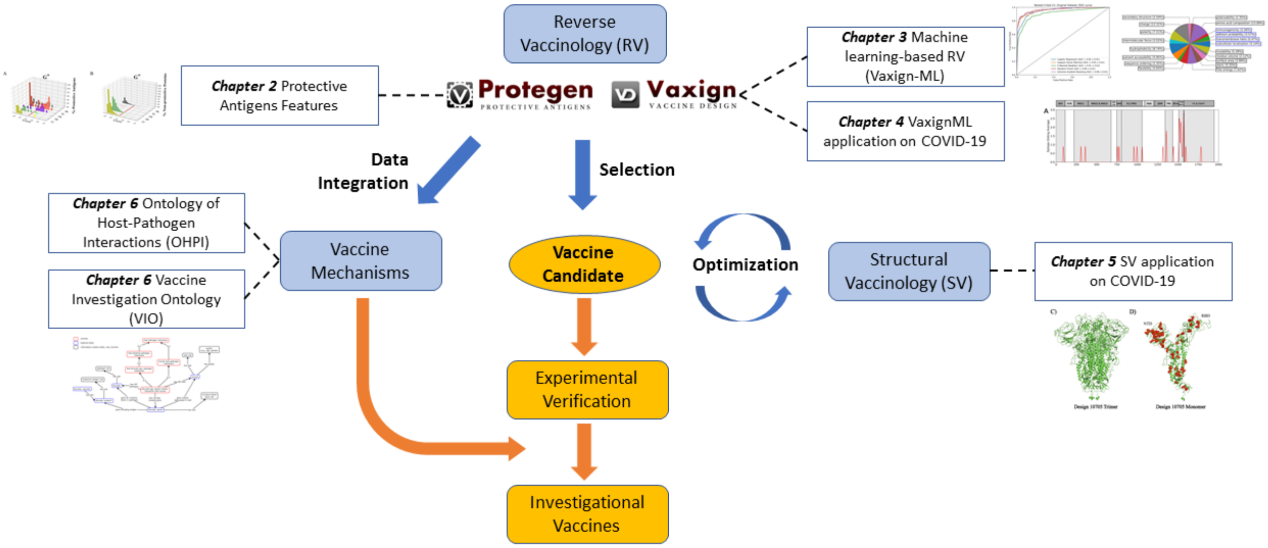

Reverse Vaccinology (RV)
Chapter 2 Protective Antigens Features
Protegen
PROTECTIVE ANTIGENS
Vaxign
VACCINE DESIGN
Chapter 3 Machine learning-based RV (Vaxign-ML)
Chapter 4 VaxignML application on COVID-19
Data Integration
Selection
Chapter 6 Ontology of Host-Pathogen Interactions (OHPI)
Chapter 6 Vaccine Investigation Ontology (VIO)
Vaccine Mechanisms
Vaccine Candidate
Optimization
Structural Vaccinology (SV)
Chapter 5 SV application on COVID-19
Experimental Verification
Investigational Vaccines

Figure 1-3 Dissertation Overview.

A systematic analysis of protective antigens (Chapter 2) led to the development of a novel machine learning-based reverse vaccinology (RV) tool Vaxign-ML (Chapter 3). It was applied to predict COVID-19 vaccine candidates (Chapter 4). The predicted candidates from the RV could be subject to structural design to optimize for immunogenicity (Chapter 5). On the other hand, two ontologies, Ontology of Host-Pathogen Interactions (OHPI) and Vaccine Investigation Ontology (VIO), were created to study the host-pathogen and host-vaccine interactions (Chapter 6). All the presented works in this dissertation are based on computational predictions and require experimental verification, as highlighted by orange boxes in the figure.

Chapter 2 Identification of New Features from Known Bacterial Protective Vaccine Antigens Enhances Rational Vaccine Design

2.1 Abstract

With many protective vaccine antigens reported in the literature and verified experimentally, how to use the knowledge mined from these antigens to support rational vaccine design and study the underlying design mechanism remains unclear. To address the problem, systematic bioinformatics analysis was performed on 291 Gram-positive and Gram-negative bacterial protective antigens with experimental evidence manually curated in the Protegen database. The bioinformatics analyses evaluated the subcellular localization, adhesin probability, peptide signaling, transmembrane α-helix and β-barrel, conserved domain, Clusters of Orthologous Groups, and Gene Ontology functional annotations. Here we showed the critical role of adhesins and subcellular localization, peptide signaling, in predicting secreted extracellular or surface-exposed protective antigens, with mechanistic explanations supported by functional analysis. We also found a significant negative correlation of transmembrane α-helix to antigen protectiveness in Gram-positive and -negative pathogens. In contrast, a positive correlation of transmembrane β-barrel was observed in Gram-negative pathogens. The commonly less focused cytoplasmic and cytoplasmic membrane proteins could be potentially predicted with other selection criteria such as adhesin probability and functional analysis. This study's significant findings can support rational vaccine design and enhance our understanding of vaccine design mechanisms.

2.2 Introduction

Vaccination is considered the most effective medical intervention ever introduced in modern medicine (Rappuoli et al. 2014). It has prevented 103 million cases of infectious diseases in the United States since 1924 (van Panhuis et al. 2013). However, it is still challenging to develop safe and effective vaccines against many infectious diseases, including tuberculosis, HIV, and malaria (WHO 2014). The emerging reverse vaccinology (RV) addresses the challenge through rational vaccine design by predicting vaccine antigen based on bioinformatics analysis of pathogen genomes (Rappuoli 2000; Adu-Bobie et al. 2003). The first application of RV in Group B *meningococcus* (MenB) vaccine development predicted 350 surface-exposed proteins from MenB, and the following experiments verified 25 of them capable of inducing bactericidal antibodies (Pizza et al. 2000). This finding led to the approval of the first MenB vaccine, Bexsero, for use in Europe (Vernikos and Medini 2014) and the United States (Folaranmi et al. 2015). The success of Bexsero is a milestone for rational vaccine design, and RV has also been applied in vaccine prediction against other challenging pathogens such as *Mycobacterium tuberculosis* (Baldwin et al. 2016).

Many selection criteria have been applied to vaccine antigen prediction, but a deep understanding of their usage rationale is still missing. The initial RV study of MenB vaccine prediction used the subcellular localization (SCL) as a primary selection criterion (Pizza et al. 2000). Humoral immunity is vital to the host protection against MenB, and the protective antigens (PAgs) inducing antibody response are primarily located in the extracellular or outer membrane. However, vaccine antigens' preference in specific SCL varies across different pathogens, and SCL might not be equivalently critical for those pathogens against which cell-mediated immunity plays a significant role. Another frequently used criterion is the number of transmembrane α-helices (TMH) due to the difficulty in isolating proteins with more than one TMH (He et al. 2010).

However, it is unclear whether the number of TMH and transmembrane β-barrel (TMB) of a protein correlates with vaccine protection. Adhesins are also crucial to pathogen invasion into host cells (Ribet and Cossart 2015), but adhesin probability (AP) usage has not been widely appreciated. Other criteria include signal peptides, conserved domains, and biological function analysis (He et al. 2010) have been used in different RV tools (e.g., NERVE (Vivona, Bernante, and Filippini 2006), Vaxign (He, Xiang, and Mobley 2010), and Jenner-predict server (Jaiswal et al. 2013)). Machine learning techniques are also applied to vaccine design studies (Bowman et al. 2011; Goodswen, Kennedy, and Ellis 2013). However, the significance and association of the above criteria with the protectiveness of bacterial PAgs are still lacking. The identification of such association is essential to improve vaccine antigen prediction and design studies.

This study aims to systematically analyze known bacterial PAgs reported in the literature and identify underlying design mechanisms for better rational vaccine prediction. Our study uses PAgs collected from Protegen with antigen information and experimental protection evidence manually annotated from peer-reviewed articles (B. Yang et al. 2011). The significance and association of these Protegen PAgs are analyzed using bioinformatics tools for SLC (Yu et al. 2010), AP (Sachdeva et al. 2005), signal peptide (Petersen et al. 2011), TMH (Krogh et al. 2001). and TMB (Bigelow et al. 2004), conserved domains (Punta et al. 2012), Clusters of Orthologous Groups (COG) (Tatusov et al. 2000), and Gene Ontology (GO) (Blake et al. 2015). This report provides a systematic analysis of protein properties and biological functions associated with known bacterial PAgs to support future rational vaccine prediction and design.

2.3 Methods

2.5.1 Protective Antigens and Background Pan-proteome Non-Protective Protein Sequences

PAgs in G+ and G- bacteria with supporting experimental evidence were downloaded from the Protegen database. The most common experimental evidence is the protection results against virulent bacterial challenges in laboratory animal models. Reported assay results that correlate to protection or immune responses are also considered. Using the G+ and G- pathogen information provided along with the PAgs from Protegen, all protein-coding sequences of these pathogens were downloaded from the UniProt database (The UniProt Consortium 2008). The taxonomy IDs reported in Protegen were queried against UniProt for possible pan-proteome sequences. The detail of taxonomy ID mapping between the reported G+ and G-pathogens from Protegen and their corresponding pan-proteome in Uniprot is available in Table 2-1. By merging all the pan-proteome protein sequences from UniProt, we obtained the background proteome for two groups used in this study: G+ and G- pathogen background proteomes. There is no curated dataset of non-protective G+ and G- proteins available in the literature. The non-protective protein datasets were generated by applying similar strategies reported in previous vaccine design studies (Doytchinova and Flower 2007; Bowman et al. 2011; El-Manzalawy, Dobbs, and Honavar 2012). Specifically, the G+ and G- pan-proteomes downloaded from UniProt were first aligned to Protegen PAg sequences using BLAST (Camacho et al. 2009). Then sequences that shared similar homology with the Protegen PAgs (E-value less than or equal to 10 and have a shared percent identity of 10%) were removed from the datasets. All the remaining sequences within the datasets were considered as non-protective proteins throughout the entire study. The non-protective proteins generated in this study only provide an estimated survey of the true non-protective datasets, and some non-protective proteins included in this study could have never been tested for the protective capacity.

2.5.2 Protein Properties Computations

In this paper, 5 types of protein properties were computed: (i) SCL, (ii) AP, (iii) signal peptide, (iv) TMH and (v) TMB.

For SCL computation, all sequences were computed for tentative SCL locations by running through the PSORTb v3.0 program (Yu et al. 2010). Briefly, PSORTb uses a Bayesian network to integrate different SCL location prediction modules such as SVM, SCL-BLAST, and motif-based modules. The program predicts and assigns a score for each possible SLC locations of the input sequence, and the location with the highest score is returned. In this study, the default setting was used besides specifying the G+ or G- of input sequences.

The AP of all sequences was computed using the SPAAN program with a default setting (Sachdeva et al. 2005). SPAAN calculates the probability of being adhesin for an input sequence using a neural network with five features, including amino acid frequencies, multiplet frequencies, dipeptide frequencies, charge composition, and hydrophobic composition. Sachdeva et al. reported 89% sensitivity and 100% specificity when the cutoff value AP $\geq$ 0.51 was used (Sachdeva et al. 2005), and therefore the same threshold was applied in this study.

Prediction of signal protein secretion of all sequences was calculated by SignalP 4.1 standalone version (Petersen et al. 2011), which is built solely on a neural network to discriminate signal peptides from transmembrane regions. The discrimination score (D-score) computed by SignalP provides value for protein secretion. The SignalP D-score threshold value of 0.45 for G+ and 0.51 for G- provides the best sensitivity in signal peptides detection. In this study, the suggested cutoff values were used, and the default configuration was applied besides specifying the G+ or G- of input sequences.

The TMH was computed using TMHMM 2.0 (Krogh et al. 2001) with default settings, and the number of TMH of the input G+ and G- pathogen sequences were reported. In brief, the tool uses a hidden Markov model to predict the transmembrane state of the input sequences, and Krogh et al. reported 97-98% prediction sensitivity (Krogh et al. 2001).

The TMB was computed using the PROFtmb tool, which is also a hidden Markov model-based prediction program (Bigelow et al. 2004). Only TMB of G- pathogen sequences were computed because classical G+ bacteria do not contain β-barrel membrane proteins (Wimley 2003). Based on the performance evaluation of the PROFtmb on discriminating transmembrane versus non-transmembrane β-barrel using the whole protein dataset by Bigelow et al. (Bigelow et al. 2004), a cut-off of ≥ 0.6 accuracy was chosen in order to achieve a balance with coverage.

2.5.3 Protein Sequence Properties Computations

The PAg sequences, non-protective protein, and background proteome sequences were functionally annotated with (i) Pfam conserved domains, (ii) COG functional classifications, and (iii) GO BP, MF, and CC terms.

The PfamScan tool was used to annotate the conserved domains in all PAg, non-protective proteins, and background proteomes. The sequences were aligned using the downloaded Pfam-A domain hidden Markov models (Punta et al. 2012).

The sequences of all PAgs were scanned for COG clusters using HMMER with the hidden Markov models downloaded from the EggNog 4.5 database (Huerta-Cepas et al. 2016). Each input sequence was initially assigned with one ENOG identifier, then mapped to the corresponding COG cluster. For background proteomes and non-protective proteins, the COG cluster identifiers were retrieved directly from the UniProt database.

The PAg sequences were submitted to the Argot2 web server for GO annotation prediction (Falda et al. 2012). The GO information of non-protective proteins and background proteomes was directly downloaded from the UniProt database.

2.5.4 Statistical Analysis

Unless specified, the statistical significance of the association between reported PAgs and computed protein properties, including SCL, AP, signal peptide, TMH, and TMB were calculated using one-way Fisher's exact test since we were only interested in the over-representation of properties in PAgs only. For the ad-hoc analysis of specific property (e.g., SCL prediction), the significance of individual sub-property (e.g., individual SCL locations such as extracellular, cell wall, cytoplasmic membrane, and cytoplasm in G+ bacteria) were further examined by performing one vs. other Fisher's exact test. The resulting p-value was adjusted by applying Bonferroni correction.

The over-representation of conserved domains, COG clusters, and GO BP, MF, CC terms among Protegen PAgs were tested using Fisher's exact test and adjusted using Benjamini–Hochberg–Yekutieli procedure. In addition, the significant (adjusted p-value ≤ 0.05) GO terms (BP, MF, CC) were visualized in a hierarchical format using GOfox (E. Ong and He 2015). GOfox laid out GO terms using the internal hierarchical GO structure simplification algorithm since GO enrichment analysis tends to generate an extensive list of enriched GO terms (E. Ong and He 2015).

2.4 Results

Three sets of data were collected and generated for the bioinformatics analysis. Our study specifically analyzed frequently used PAg prediction features, including SCL, AP, signal peptide, TMH and TMB, conserved domain, and biological function analysis.

2.3.1 Collection of Protective Vaccine Antigens, Background, and Non-protective proteins

After removal of identical sequences, the curated Protegen dataset contained 81 and 210 non-redundant vaccine PAgs from 14 Gram-positive (G+) and 34 Gram-negative (G-) bacteria, respectively (Table 2-1). The corresponding pan-proteomes of these G+ and G- pathogens were downloaded from the UniProt database (The UniProt Consortium 2008) as the background proteomes, which included 39,397 G+ and 73,371 G- peptide sequences. A set of non-protective proteins were selected from background proteome as described in the Method section and other RV studies (Doytchinova and Flower 2007; Bowman et al. 2011; El-Manzalawy, Dobbs, and Honavar 2012; Goodswen, Kennedy, and Ellis 2013), and contained 4,954 G+ and 5,478 G- pathogen peptide sequences.

2.3.2 Subcellular Localization (SCL) Analysis

Our analysis found that 44.4% and 19.8% of PAgs in G+ bacteria are located in extracellular space and cell wall, respectively (Figure 2-1 A). In comparison, only 1.7% and 1.2% of the G+ non-protective proteins were extracellular and cell wall proteins, respectively (Figure 2-1 B). Our statistical analysis showed a significant over-representation of PAgs in these two SCLs (p-value < 0.01). In G- bacteria, 15.7%, 30.0%, and 8.1% of PAgs were extracellular, outer membrane, and periplasmic proteins, respectively (Figure 2-1 D). Compared to the corresponding SCL proportions in G- non-protective proteins (0.4%, 0.4%, and 0.9%) (Figure 2-1 E), these three locations were significantly over-represented in PAgs (p-value < 0.01). In non-protective proteins, most proteins (78.3% in G+ and 67.7% in G-) were localized in the cytoplasmic or cytoplasmic membrane (Figure 2-1 B&E), but these two SCL locations also accounted for 26.8% G+ and 31.1% G- of the reported PAgs (Figure 2-1 A&D).

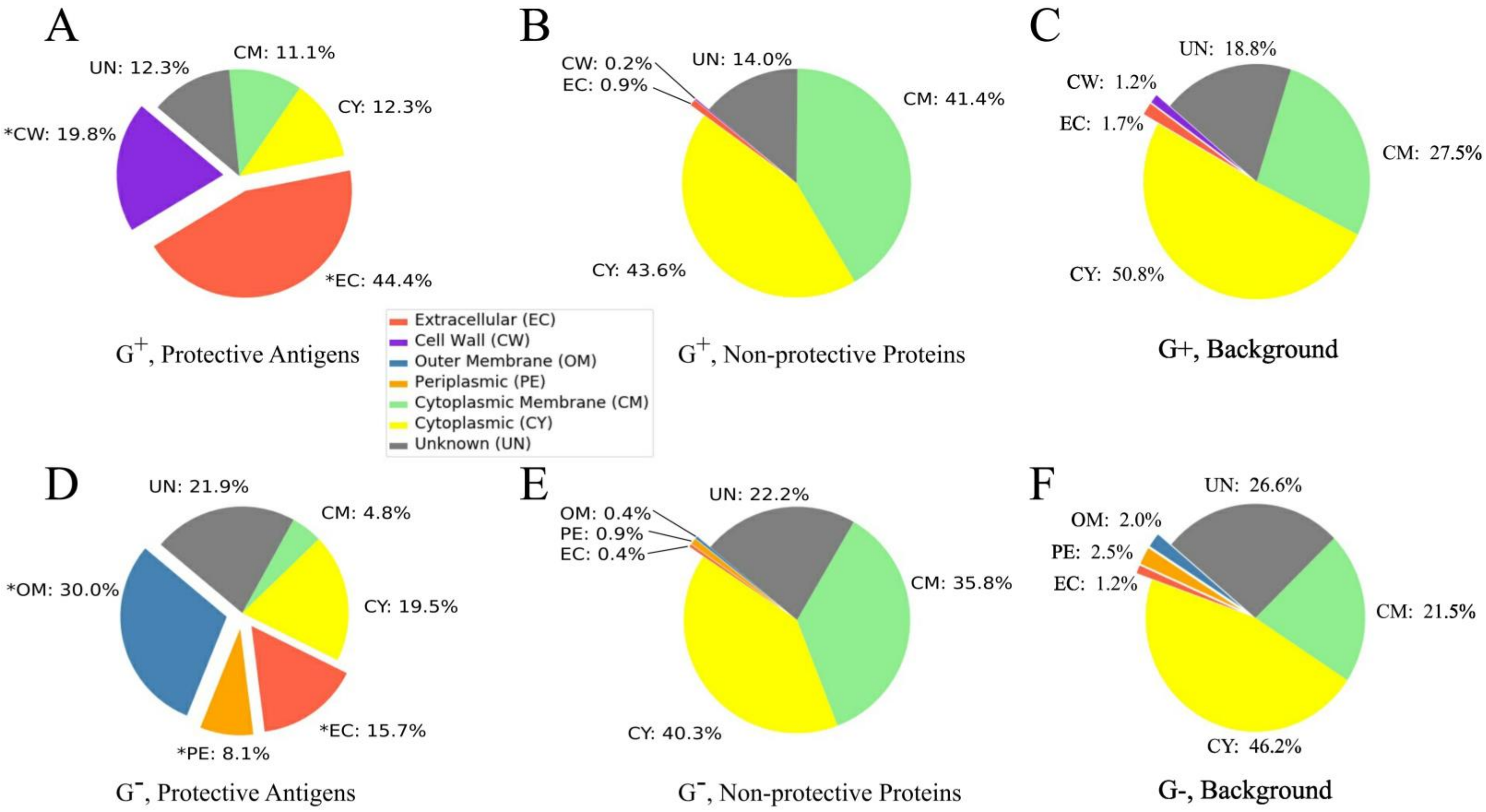


Figure 2-1 Subcellular localization profiles.

The bacterial PAgs (A&D) showed significant enrichment (p-value < 0.01) at different cellular locations: extracellular for both Gram+ (G+) and Gram- (G-) bacteria; cell wall for G+; outer membrane and periplasm for G-, when compared to the non-protective proteins (B&E). The background proteome subcellular localization distribution (C&F) was similar to the non-protective proteins.

* indicates significant (p-value < 0.01) over-representation of PAgs' subcellular localization prediction compared to non-protective proteins.

Table 2-1 A list of Gram-positive and Gram-negative bacteria used to analyze significant features associated with protective antigens.

	Pathogen Taxonomy ID	Pathogen Name	Uniprot Pan-proteome ID	Protein Counts	Protective Antigen Counts
Gram-Positive Bacterium	1773	Mycobacterium tuberculosis	UP000001584	3,993	27
	1392	Bacillus anthracis	UP000000594	5,493	15
	1313	Streptococcus pneumoniae	UP000000586	2,030	13
	1491	Clostridium botulinum	UP000001986	3,590	7
	1336	Streptococcus equi	UP000001368	1,851	6
	1280	Staphylococcus aureus	UP000008816	2,889	6
	1314	Streptococcus pyogenes	UP000000750	1,690	5
	1311	Streptococcus agalactiae	UP000001415	3,240	4
	1485	Clostridium tetani	UP000001412	2,415	2
	1639	Listeria monocytogenes	UP000000817	2,844	2
	1717	Corynebacterium diphtheriae	UP000002198	2,265	1
	1781	Mycobacterium marinum	UP000001190	5,418	1
	1648	Erysipelothrix rhusiopathiae	UP000007944	1,679	1
	1765	Mycobacterium bovis	UP000001584	3,993	1
Gram-Negative Bacterium	234	Brucella spp.	UP000002719	3,023	28
	632	Yersinia pestis	UP000000815	3,909	26
	487	Neisseria meningitidis	UP000000425	2,001	21
	83334	Escherichia coli	UP000000625	4,306	19
	727	Haemophilus influenzae	UP000000579	1,707	14
	520	Bordetella pertussis	UP000047656	3,783	14
	83555	Chlamydophila abortus	UP000000431	895	10
	83560	Chlamydia muridarum	UP000000431	895	10
	197	Campylobacter jejuni	UP000000799	1,623	10
	210	Helicobacter pylori	UP000000429	1,553	9
	263	Francisella tularensis	UP000001174	1,528	9
	590	Salmonella spp.	UP000000625	4,306	7
	715	Actinobacillus pleuropneumoniae	UP000001432	2,004	7
	83558	Chlamydophila pneumoniae	UP000000431	895	6
	620	Shigella	UP000002716	3,897	6
	287	Pseudomonas aeruginosa	UP000002438	5,563	4
	28450	Burkholderia pseudomallei	UP000000605	5,717	4
	139	Borrelia burgdorferi	UP000001807	1,290	4
	636	Edwardsiella tarda	UP000001485	3,686	4
	83554	Chlamydophila psittaci	UP000014824	1,714	3
	747	Pasteurella multocida	UP000000809	2,015	3

	780	Rickettsia spp	UP000002480	834	3
	666	Vibrio cholerae	UP000036184	4,527	3
	160	Treponema pallidum	UP000000811	1,028	3
	777	Coxiella burnetii	UP000002671	1,815	2
	645	Aeromonas salmonicida	UP000000756	4,121	2
	171	Leptospira spp.	UP000001408	3,676	2
	633	Yersinia pseudotuberculosis	UP000000815	3,909	1
	2096	Mycoplasma gallisepticum	UP000001418	761	1
	393305	Yersinia enterocolitica	UP000000815	3,909	1
	55601	Vibrio anguillarum (Listonella anguillarum)	UP000006800	3,722	1
	2099	Mycoplasma hyopneumoniae	UP000000548	671	1
	738	Haemophilus parasuis	UP000006743	2,002	1
	813	Chlamydia trachomatis	UP000000431	895	1

To confirm the SCL analysis results, we also analyzed signal peptides using SignalP (Petersen et al. 2011), which predicted the presence of signal sequences of most synthesized proteins designated to secretory pathways. The distribution histograms of the calculated score for PAgs, non-protective proteins, and background proteomes were plotted (Figure 2-2). The signal peptide scores of extracellular (both G+ and G-) or surface-exposed proteins (cell wall for G+ and outer membrane for G-) showed that a large fraction of PAgs was predicted to be secreted signal peptides.

2.3.3 Adhesin Probability (AP) Analysis

Adhesins are proteins critical for bacterial pathogens to invade host cells and cause infections (Ribet and Cossart 2015). Over half of the PAgs could be identified with AP (56.8% of G+ and 52.8% of G-) using the suggested cutoff of no less than 0.51 (Sachdeva et al. 2005). The AP of proteins with different SCLs also had different patterns (Figure 2-3). Specifically, comparing PAgs (Figure 2-3 B&E) and non-protective proteins (Figure 2-3 C&F), PAgs with SCL locations other than cytoplasmic membrane and cytoplasm generally showed an increasing trend in AP. There were 87.5% G+ PAgs in the cell wall and 82.5% G- PAgs in the outer membrane that was also adhesins, compared to 37.5% G+ and 20% G- non-protective proteins in the cell wall and outer membrane, respectively (Figure 2-3). This high preference of surface-exposed proteins (cell wall for G+ and outer membrane for G-) with high AP was significant (p-value < 0.01, Figure 2-3) and illustrated the importance of SCL and AP as two significant criteria in vaccine design. Additionally, 90.0% and 54.3% of the PAgs in G+ and G- bacteria with unknown SCL were in fact predicted to be adhesins. Therefore, utilizing AP with SCL could potentially overcome the limitation of excluding “Unknown” SCL and avoid inaccuracy generated by individual SCL prediction tool. For PAgs located at the cytoplasmic membrane and cytoplasm, the computed AP also showed different patterns between G+ and G- (Figure 2-3 B&E). G+ PAgs in cytoplasmic membrane were more likely adhesins (77.8%) while in G-, only 20.0 % were adhesins.

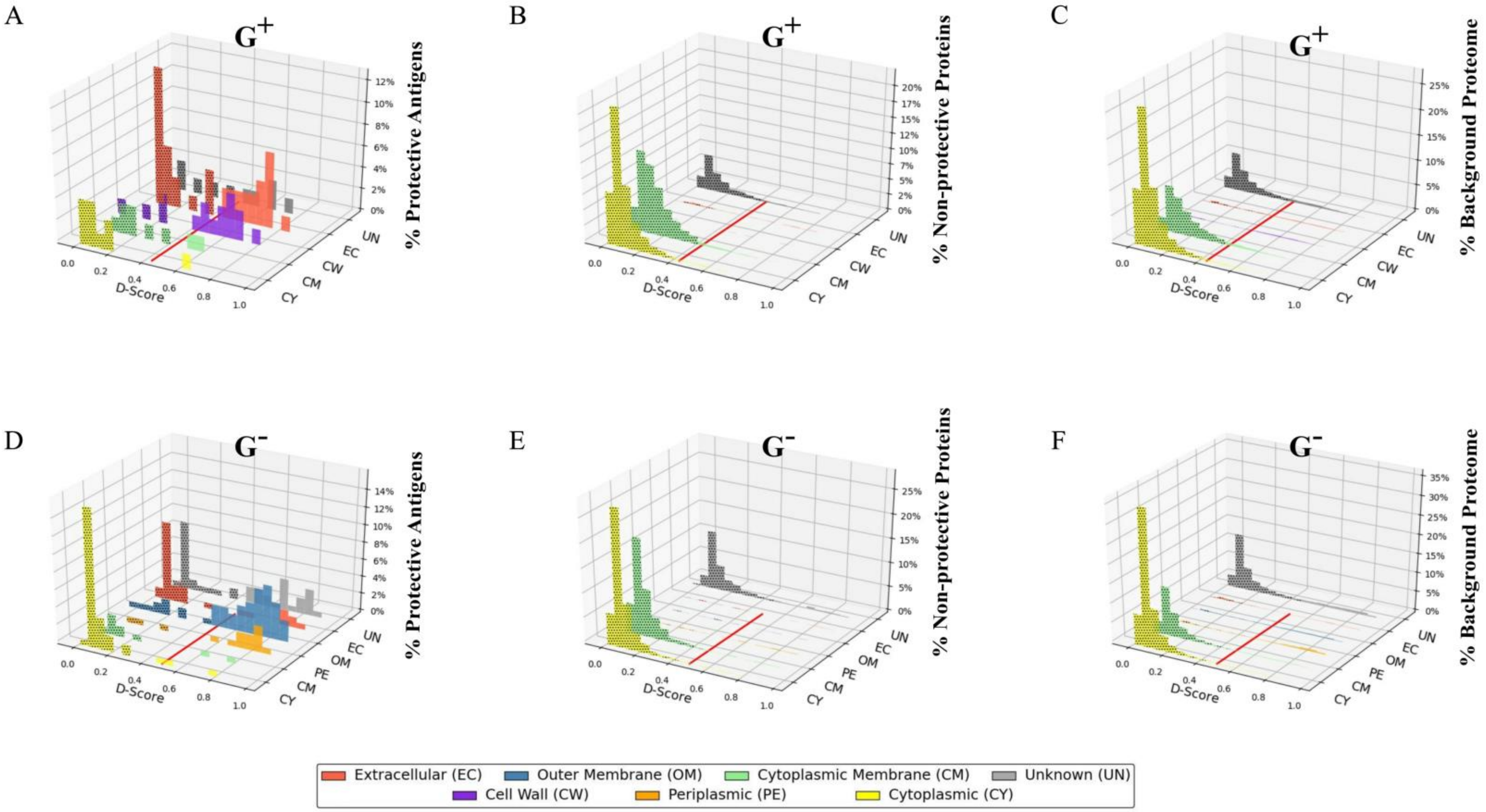


Figure 2-2 SignalP score distribution of protective antigens, non-protective proteins, and background proteome. Protective antigens (A&D) showed significantly more signaling peptide predictions in subcellular locations, including cell wall for Gram-positive, outer membrane and Periplasmic for Gram-negative), and extracellular and unknown locations for both Gram-positive and -negative bacteria in comparison to the non-protective proteins (B&E) and background proteome (C&F).

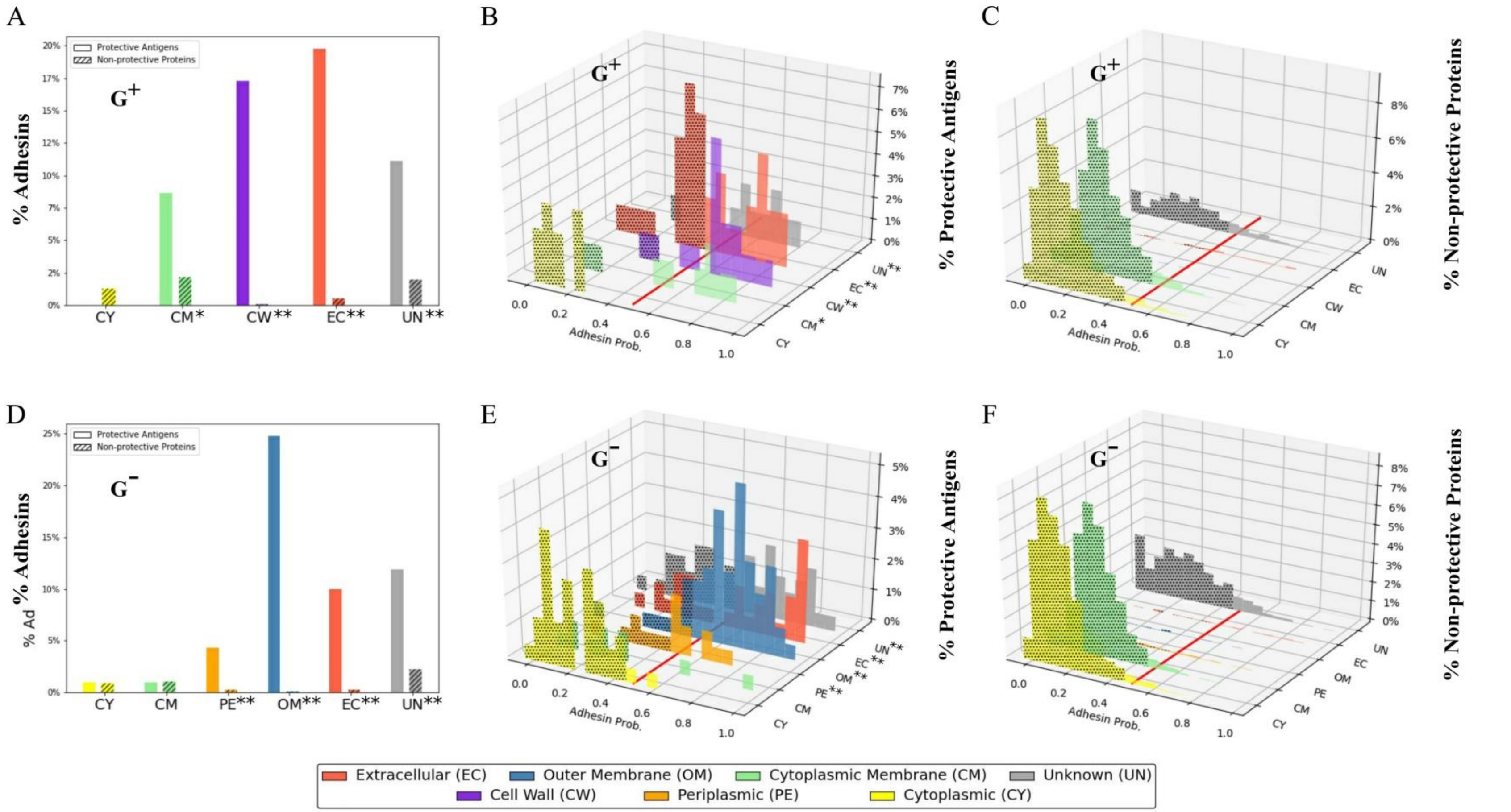


Figure 2-3 Profiles of adhesin probabilities of protective antigens and non-protective proteins with different subcellular localizations.

The top three subfigures (A-C) show Gram+ (G+) pathogens and the bottom three show Gram- (G-) pathogens. Specifically, the first column (A & D) represents the overall percentages of adhesin probabilities. The second column (B & E) and third column (C & F) show adhesin probability distributions of protective antigens (PAgs) and non-protective proteins, respectively. The red line in (B, C, E, F) indicates an adhesin probability cutoff of no less than 0.51. Overall, PAgs have significantly higher (p-value < 0.01) percentages in extracellular (G+ & G-), cell wall (G+), periplasm, and outer membrane (G-). Interestingly, the cytoplasmic membrane PAgs in G+ is also significant (p-value < 0.05) when coupled with adhesin probability, which might be associated with the induction of cell-mediated immunity.

* and ** indicates significant over-representation of PAgs' adhesin probabilities at different subcellular localizations compared to non-protective proteins with p-values < 0.05 and p-value < 0.01, respectively.

2.3.4 Transmembrane α-helix (TMH) and β-barrel (TMB)

We analyzed and compared the TMH profiles between PAgs and non-protective proteins. Specifically, none of the PAgs located at the cell wall (G+), outer membrane, or periplasm (G-) had more than one TMH (Figure 2-4). There were two G- PAgs with more than 10 TMH (lipoprotein signal peptidases in *Brucella melitensis* and L-lactate permease in *Neisseria meningitides*). The β-barrel analysis was only performed for G- pathogens because classical G+ bacteria do not contain β-barrel membrane proteins (Wimley 2003). Using the probability cutoff of 0.60, our study found that 12.9% of Gram-negative PAgs predicted to have TMB compared to less than 0.001% in non-protective proteins (Figure 2-4).

2.3.5 Conserved Domain Analysis

Conserved domains represent functional units in proteins, and some domains are more frequently associated with PAgs (He and Xiang 2012; Jaiswal et al. 2013). Our analysis identified eight conserved domains that were only frequently found among reported PAgs (Table 2-2). These domains included autotransporter beta-domain, outer membrane protein beta-barrel domain, fimbrial protein, TonB-dependent receptor plug domains, OmpH-like outer membrane protein, extended signal peptide of type V secretion system, ABC transporter, and Extended Signal Peptide of Type V secretion system. The top two frequently found Pfam-A conserved domains among reported PAgs were β-barrel domains, which support the positive selection of TMB in PAg prediction. In addition, proteins with over-represented conserved domains were more likely related to bacteria's pathogenesis, including pathogen colonization and invasion. They, therefore, could be used as a good indicator of PAg prediction.

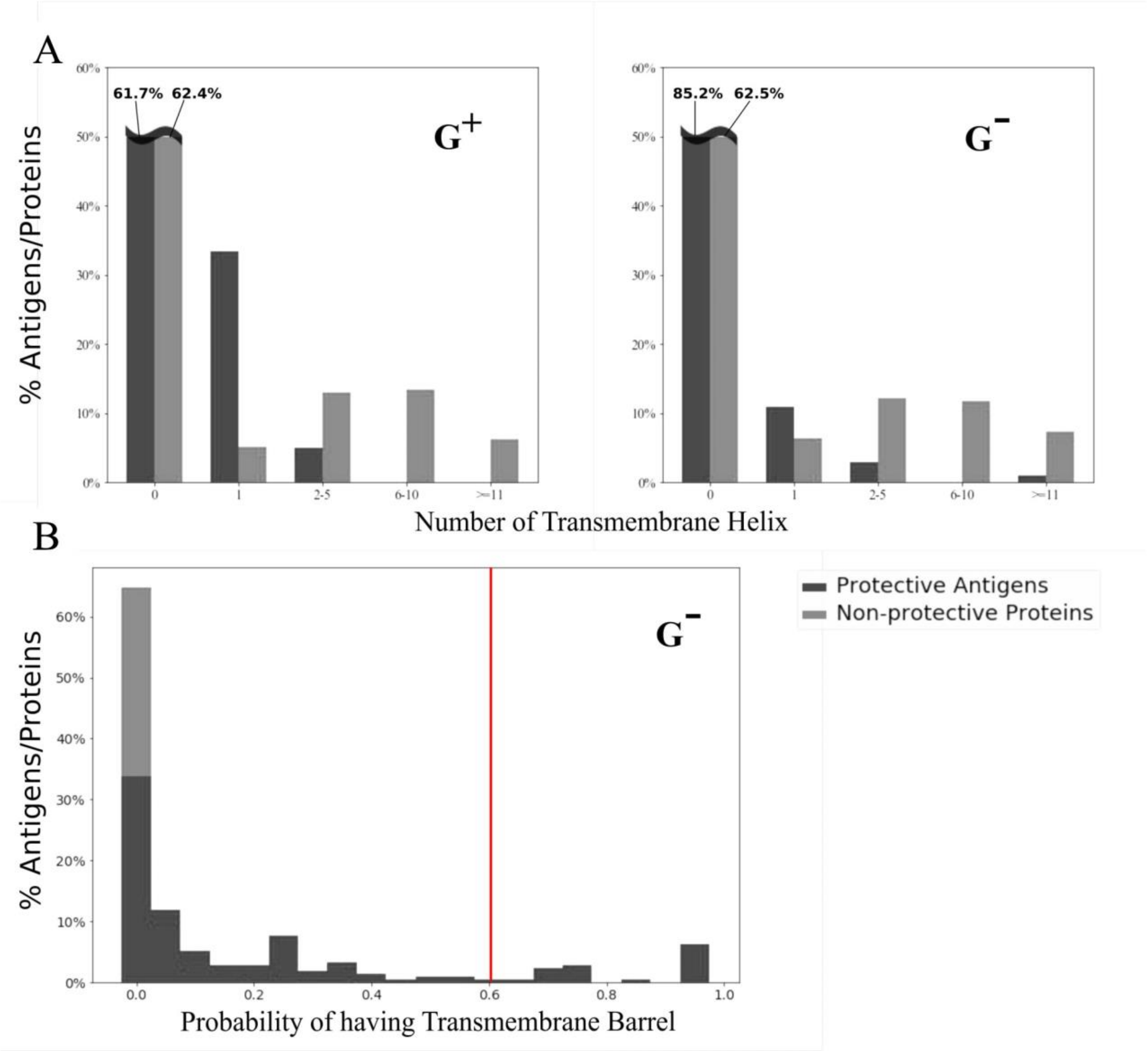


Figure 2-4. Transmembrane α-helix and β-barrel profiles in protective antigens.

Compared to non-protective proteins, there were much higher percentages of PAgs with zero or one transmembrane α-helix (A). For transmembrane β-barrel (B), only 2 (0.0004%) out of all non-protective proteins had a probability higher than the designated cutoff (indicated as a vertical line).

Table 2-2 Frequent conserved domains among reported protective antigens.

Pfam domain description	Protective antigen count
Autotransporter beta-domain	11
Outer membrane protein beta-barrel domain	10
Fimbrial protein	10
ATPase family associated with various cellular activities (AAA)	9
TonB-dependent Receptor Plug Domain	8
Outer membrane protein (OmpH-like)	5
ABC transporter	5
Extended Signal Peptide of Type V secretion system	5

2.3.6 Functional Analysis

The functional annotations were analyzed using the COG and GO. COG includes 26 functional clusters (Tatusov et al. 2000). Our COG analysis of PAgs identified 16 COG functional categories that were significantly enriched (adjusted p-value < 0.05) in PAgs (Figure 2-5). Four COG clusters cell wall/membrane envelope biogenesis, cell motility, signal transduction mechanisms, and extracellular structures were notably enriched in PAgs.

We also analyzed enriched GO terms from the three GO branches: biological process (BP), molecular function (MF), and cellular component (CC) (Blake et al. 2015). Eighteen GO BP terms were found significantly enriched (adjusted p-value < 0.05) in bacterial PAgs, including 'pathogenesis' as the most significantly enriched term among PAgs in bacterial pathogens (Figure 2-6). BPs related to pathogen invasion (e.g., cell adhesion and proteolysis) and terms related to the transporter (e.g., transmembrane transport) were significantly over-represented among PAgs. Twenty GO MF terms were significantly enriched (adjusted p-value < 0.05), including those related to invasion (e.g., peptidase activity) and transportation (e.g., transferase activity and receptor activity). Fifteen GO CC terms were significantly enriched (adjusted p-value < 0.05). In agreement with the SCL prediction results, extracellular or surface-exposed CC terms were significantly over-represented among reported PAgs. In addition, CC terms that were related to bacterial colonization and invasion within the host, such as bacterial-type flagellum filament, pilus, host cell part, host cell plasma membrane, and host cell junction, were also enriched, suggesting PAgs' role in the interactions between bacteria and the host cells.

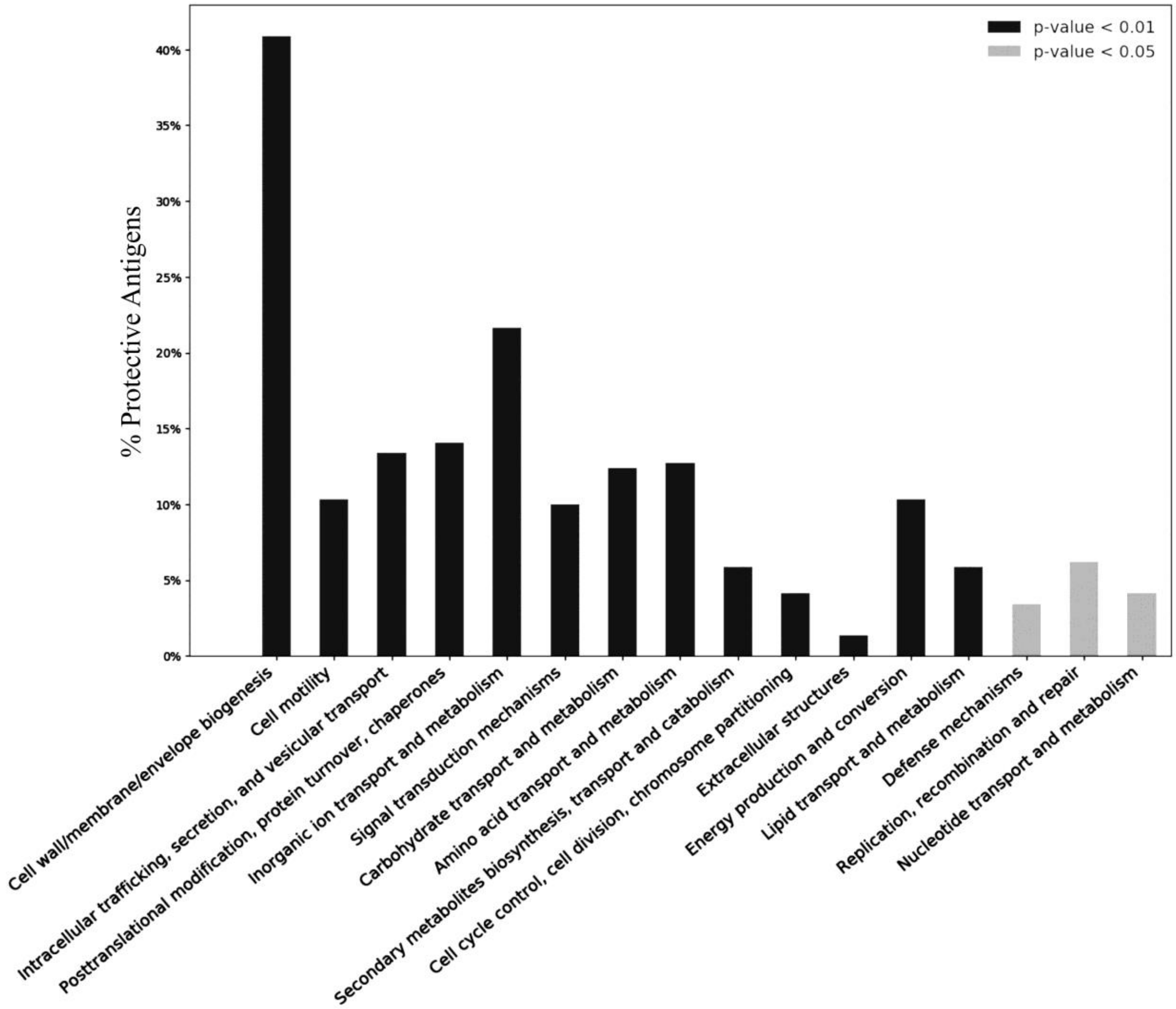


Figure 2-5 Over-represented Clusters of Orthologous Groups clustering profiles among reported protective antigens.

Over 40% of the reported protective antigens (PAgs) belong to the cluster cell wall/membrane/envelop biogenesis, which agrees with common knowledge of using surface-exposed proteins as a key criterion in vaccine antigen prediction. Other Clusters of Orthologous Groups (COG) clusters related to pathogen motility, secretion, signal transduction, and transportation are also significantly enriched in PAgs in comparison to non-protective proteins. The significant over-representation of PAgs' COG clusters compared to non-protective proteins is colored with grey (p-value < 0.05) and black (p-value < 0.01).

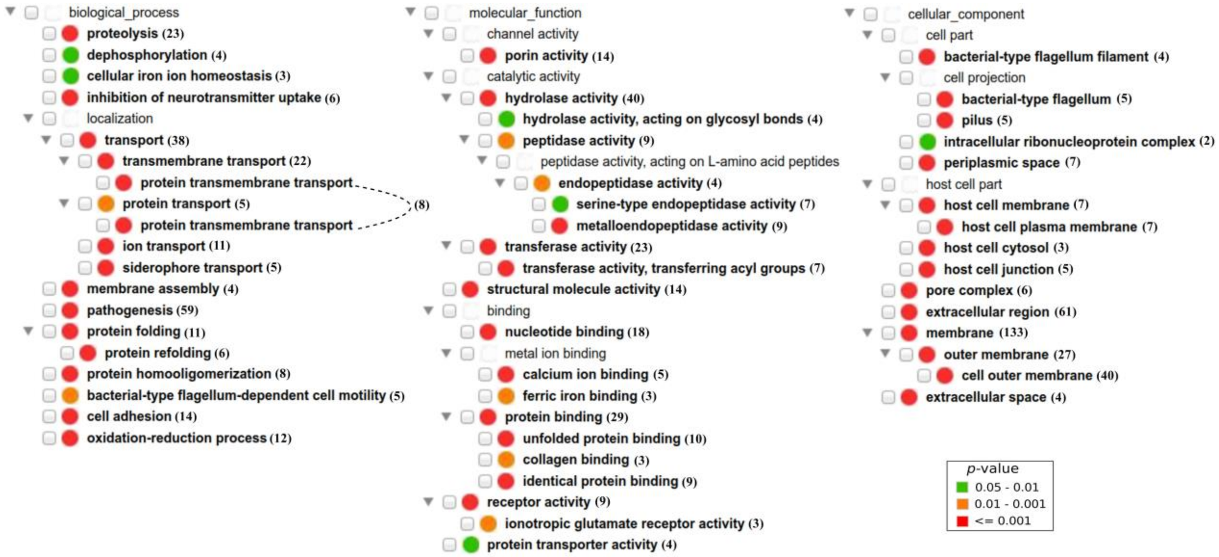

biological_process
proteolysis (23)
dephosphorylation (4)
cellular iron ion homeostasis (3)
inhibition of neurotransmitter uptake (6)
localization
transport (38)
transmembrane transport (22)
protein transmembrane transport
protein transport (5)
protein transmembrane transport
(8)
ion transport (11)
siderophore transport (5)
membrane assembly (4)
pathogenesis (59)
protein folding (11)
protein refolding (6)
protein homooligomerization (8)
bacterial-type flagellum-dependent cell motility (5)
cell adhesion (14)
oxidation-reduction process (12)
molecular_function
channel activity
porin activity (14)
catalytic activity
hydrolase activity (40)
hydrolase activity, acting on glycosyl bonds (4)
peptidase activity (9)
peptidase activity, acting on L-amino acid peptides
endopeptidase activity (4)
serine-type endopeptidase activity (7)
metalloendopeptidase activity (9)
transferase activity (23)
transferase activity, transferring acyl groups (7)
structural molecule activity (14)
binding
nucleotide binding (18)
metal ion binding
calcium ion binding (5)
ferric iron binding (3)
protein binding (29)
unfolded protein binding (10)
collagen binding (3)
identical protein binding (9)
receptor activity (9)
ionotropic glutamate receptor activity (3)
protein transporter activity (4)
cellular_component
cell part
bacterial-type flagellum filament (4)
cell projection
bacterial-type flagellum (5)
pilus (5)
intracellular ribonucleoprotein complex (2)
periplasmic space (7)
host cell part
host cell membrane (7)
host cell plasma membrane (7)
host cell cytosol (3)
host cell junction (5)
pore complex (6)
extracellular region (61)
membrane (133)
outer membrane (27)
cell outer membrane (40)
extracellular space (4)
p-value
🟩 0.05 - 0.01
🟧 0.01 - 0.001
🟥 <= 0.001

Figure 2-6 Over-represented gene ontology biological process, molecular function, and cellular component profiles among reported protective antigens.

The number next to each Gene Ontology (GO) term indicates the number of protective antigens (PAgs) with the corresponding GO functional annotation. Similar to Clusters of Orthologous Groups clustering, GO terms related to pathogen motility, secretion, signal transduction, and transportation are also significantly enriched in PAgs compared to non-protective proteins. The GO cellular component terms also supported the high preference of extracellular, surface-exposed (cell wall in Gram-positive and outer membrane in Gram-negative) and periplasmic (Gram-negative) PAgs. The significant over-representation of PAgs' GO terms compared to non-protective proteins is color-coded following the legend in the lower right corner.

2.5 Discussion

Although extensive research has been conducted, modern vaccine research and development still faces challenges of rapid and accurate development of vaccines in response to major infectious diseases (e.g., tuberculosis (WHO 2014)), outbreaks (e.g., Ebola and Zika virus (Leligdowicz et al. 2016; Saiz et al. 2016)), and new drug-resistant pathogens (Kling et al. 2014). Our efforts to develop vaccines using traditional methods have not been successful in addressing these challenges. Effective vaccine development's future success relies on robust and rational vaccine design, including reverse and structural vaccinology (Rappuoli et al. 2014), and our more in-depth understanding of vaccination mechanisms. Our comprehensive bioinformatics study analyzed important vaccine design criteria by systematically studying and comparing bacterial PAgs and non-protective proteins, including various protein properties and biological functions. This study's summarized characteristics are used explicitly for bacterial model PAg prediction and might not hold true for viral or parasitic pathogens. The results of this study confirmed and provided details on the usage of these prediction criteria, including SCL, AP, signal peptides, TMH and TMB, conserved domains, and biological function annotations, for RV prediction against bacterial pathogens. Most importantly, our results suggested new insights towards rational vaccine prediction and design.

In accordance with secreted extracellular or surface-exposed antigens commonly known to be PAgs, our study observed the differences among the SCL profiles of G+ and G- bacterial PAgs (Figure 2-1). In terms of extracellular proteins, G+ bacterial PAgs had a much higher percentage (44%) being PAgs than G- bacterial PAgs (15.7%). We also found a strong correlation between the presence of secretory signal peptides and PAgs. Approximately half of the PAgs (over 45% in both G+ and G-) were predicted to be signal peptides (Figure 2-2).

Coupling the selection of SCL and signal peptides, particularly in G+ bacterial pathogens, pose a viable option for a more precise PAg prediction. On the other hand, 19.8% of cell wall proteins in G+ and 30.0% outer membrane proteins in G- bacteria were surface-exposed PAgs (Figure 2-1 A&D). The G+ bacterial PAgs showed a higher preference in extracellular proteins, while both G+ and G- bacterial PAgs shared similar proportions as surface-exposed proteins.

Moreover, 8.1% G- PAgs were in the periplasm, a subcellular location that vaccine researchers often ignore due to lack of direct interaction with the host immune cells. The percentage of periplasmic PAgs was significant (p-value < 0.05, Figure 2-1 C) and was also supported by the over-represented GO terms (Figure 2-6). G- bacterial periplasmic proteins can be possibly released extracellularly after being packed within outer membrane vesicles and induce strong immune responses (Collins 2011; Godlewska et al. 2016). These periplasmic proteins can potentially be a good source of PAg candidates when coupling with other selection criteria such as functional analysis.

Our study results highlight AP's importance and its effect on improving RV prediction when combined with SCL. Adhesin is critical for bacterial invasion and can induce strong immune responses (Ribet and Cossart 2015). Adhesins can also function as enzymes and mediate a prominent part of bacterial pathogenesis (S. Patel, Mathivanan, and Goyal 2017). The majority of vaccine design studies do not incorporate AP in their selection pipeline (Doytchinova and Flower 2007; Bowman et al. 2011; Goodswen, Kennedy, and Ellis 2013; Pizza et al. 2000), and AP as a selection criterion is currently underused and poorly investigated in the vaccine development field. Our study managed to identify over 50% of the PAgs with AP as the only criterion. By addressing the importance of adhesin playing an essential role in vaccine development, we hope to promote the AP as a viable option in future vaccine design studies.

The functional analysis of adhesive PAgs in our study proposes a mechanistic explanation of their roles in pathogen colonization and invasion. Cell motility is one of the essential steps in host colonization and invasion. The bacterial movement requires structures such as flagellum and pillus for cell adhesion and colonization (Ramos, Rumbo, and Sirard 2004). Cell motility related COG clusters and GO terms were significantly enriched (Figure 2-5 and 2-6). Pilli is composed of fimbrial and other proteins(Ramos, Rumbo, and Sirard 2004), and the Pfam domain 'fimbrial protein' was highly conserved among the reported PAgs (Table 2-2). GO BP term proteolysis and GO MF terms peptidase activity (Figure 2-6) were also significant in the functional analysis. For instance, *Yersinia pestis* can produce the surface protease to mediate invasion into host endothelial cells (La et al. 2001). The pili, fimbri, and protease mentioned earlier can part of the various adhesins' architectures (S. Patel, Mathivanan, and Goyal 2017). Given these critical roles of adhesins, more investigations of adhesins as potential PAgs and how they induce protective immunity are much deserved.

Our study showed two distinct correlation patterns of the PAgs protectiveness to the TMH and TMB. The TMH is more abundant in cytoplasmic or inner membranes, and the TMB type is more likely located in bacterial outer membranes (Schulz 2002). Our study confirmed that TMH proteins with more than one TMH were not typically used for vaccine development (He et al. 2010) (Figure 2-4 A). Two exceptions with more than 10 TMHs were *Brucella* lipoprotein signal peptidase and *Neisseria meningitides* L-lactate permease. *Brucella* lipoprotein signal peptidase is a known virulence factor involved in lipopolysaccharides biosynthesis (Zygmunt et al. 2006). The *N. meningitides* L-lactate permease is a protein required by *N. meningitides* during bacteraemic infection, which can induce protective immunity in systemic *meningococcal* infection (Sun et al. 2005). Different from TMH, our study indicated that the presence of TMB was associated with significantly higher portions (p-value < 0.01 from the chi-squared test) of G-PAgs (Figure 2-4 B). In particular, none of the G- outer membrane non-protective proteins was predicted to have TMB. Our results suggested the use of TMH as a negative and TMB as a positive selection criterion in future vaccine development.

Although not usually considered as PAgs, large portions (26.8% G+ and 31.1% G-) of cytoplasmic and cytoplasmic membrane proteins were found to be PAgs (Figure 2-1 A, D). Compared to a much larger size of cytoplasmic and cytoplasmic membrane non-protective proteins, this fraction of PAgs was not significant. However, the ignorance of proteins located at these two SCLs might hinder effective PAg prediction productivity. Cytoplasmic and cytoplasmic membrane proteins might not induce humoral immune response due to their SCLs, but these proteins often can be potent inducers of cell-mediated immunity. For example, the cytoplasmic catalase-peroxidase protein in *Mycobacterium tuberculosis* contributes to intracellular survival within the host macrophage by protecting against reactive oxygen species (Ng et al. 2004) and can induce protective immunity (Z. Li et al. 1999). How to accurately predict cytoplasmic PAgs remains a big challenge, but it can be potentially addressed using multiple features such as AP, conserved domains, COG clusters, and GO terms. In particular, G+ PAgs showed significant over-representation in the cytoplasmic membrane (p-value < 0.05) when coupled with AP prediction. Conserved domains have been reported as a viable option in the PAgs prediction (Jaiswal et al. 2013). In our study, many conserved domains were frequently found among PAgs, which might link to essential bacterial biological functions (e.g., TonB-dependent receptor plug domain). As a strategy in antibiotics resistance is the bacterial efflux pumps (Jessica M. A. Blair, Mark A. Webber, Alison J. Baylay 2015), TonB-dependent receptor is a G- bacterial protein responsible for the transportation of large ion complex and have been identified as a potent vaccine PAgs (Z. Ni et al. 2017). The over-represented COG clusters and GO terms among the reported PAgs suggested a viable alternative to overcome the challenge of identifying cytoplasmic and cytoplasmic membrane PAgs and complement current vaccine prediction studies.

This study's findings can be translated into a predictive framework with different approaches to improve existing methods and achieve better identification and validation of novel PAgs. Even though traditional rule-based prediction has been successful in multiple studies (Pizza et al. 2000; Baldwin et al. 2016) and also applied in many tools (He, Xiang, and Mobley 2010; Jaiswal et al. 2013; Vivona, Bernante, and Filippini 2006), this type of “all-or-nothing” selection might fail to capture the inter-relation among different criteria (Goodswen, Kennedy, and Ellis 2013). For example, a potential cytoplasmic or cytoplasmic membrane PAg would be immediately discarded from a study that includes surface-exposing SCL as one of the criteria. As indicated in one of our findings, the cytoplasmic or cytoplasmic membrane PAg could be predicted by incorporating other criteria such as AP, conserved domains, and biological functions. A combinatory strategy has been proposed as a natural solution that assigns each criterion with weight and synthesizes multiple criteria in a composite way, such as weighted metrics (Lopera-Madrid et al. 2017). Candidate proteins with low scores in a set of rules could still achieve a reasonable score and are compensated by another set of selection criteria. Another advanced technique is to apply machine learning methods such as support vector machine, random forest, and neural network as described in previous studies (Bowman et al. 2011; Goodswen, Kennedy, and Ellis 2013; El-Manzalawy, Dobbs, and Honavar 2012; He and Xiang 2012). Even though the machine learning-based prediction can overcome the “all-or-nothing” scenario, these methods have not captured all the significant features reported in this study. For example, AP and conserved domains are not implemented in current ML-based prediction (Bowman et al. 2011; Goodswen, Kennedy, and Ellis 2013; El-Manzalawy, Dobbs, and Honavar 2012) except the preliminary study by Xiang & He (He and Xiang 2012), and none of these studies incorporated TMB and biological functional analysis into their prediction pipeline. The additional features given from our findings showed promising improvement on current machine learning methods.

Based on the discoveries reported in this study, we plan to explore the possibility of integrating these significant criteria, including MHC-epitope binding and structure on protein selection, to predict vaccine candidates and improve our Vaxign software program (He, Xiang, and Mobley 2010). Even though our analysis focused on the bacterial model, some criteria such as AP, signal peptide, transmembrane proteins, pathogenesis-related conserved domains, and biological functions can be extended to viral or parasitic PAgs prediction after further verification and analysis. A better understanding of the association between individual criterion and PAgs and the inter-relation among different criteria will provide new opportunities for more accurate and rational vaccine design, leading to better prevention and control of various infectious diseases.

2.6 Acknowledgement

I appreciate the support of the US National Institutes of Health (1R01AI081062). The result presented in this chapter is published in (E. Ong, Wong, and He 2017). Edison Ong (EO) and Yongqun He (YH) conceived and designed the study; EO and Mei U Wong (MW) collected protective antigens and background proteome sequences; EO performed bioinformatics analyses of the protein properties and functional annotations; EO, MW, and YH prepared the manuscript.

All authors participated in result interpretation, paper editing, discussion, and approved the paper publication.

Chapter 3 Vaxign-ML: Supervised Machine Learning Reverse Vaccinology Model for Improved Prediction of Bacterial Protective Antigens

3.1 Abstract

Reverse vaccinology (RV) is a milestone in rational vaccine design, and machine learning (ML) has been applied to enhance RV prediction accuracy. However, ML-based RV still faces challenges in prediction accuracy and program accessibility. This study presents Vaxign-ML, a supervised ML classification to predict bacterial protective antigens. To identify the best ML method with optimized conditions, five ML methods were tested with biological and physiochemical features extracted from well-defined training data. Nested five-fold cross-validation and leave-one-pathogen-out validation were used to ensure unbiased performance assessment and the capability to predict vaccine candidates against a new emerging pathogen. The best performing model, Vaxign-ML, was compared to three publicly available RV programs with a high-quality benchmark dataset. Vaxign-ML showed superior performance in predicting bacterial protective antigens. Vaxign-ML is deployed in a publicly available web server.

3.2 Introduction

As the most successful medical intervention in modern medicine, vaccination is still facing the considerable difficulty of developing safe and effective vaccines against many infectious diseases such as tuberculosis, HIV, and malaria (WHO 2014). The advance of high-throughput sequencing technology has fostered an innovative genome-based vaccine design approach in the early 1990s, termed Reverse Vaccinology (RV) (Rappuoli 2000). The first RV study identified meningococcal protein vaccine candidates using the whole genome sequences of Group B meningococcus. This study selected and verified 28 immunogenic proteins using a bioinformatics approach followed by experimental validation (Pizza et al. 2000). The Bexsero vaccine, formulated using 5 out of the 28 protein candidates, has been licensed in Europe and the United States (Vernikos and Medini 2014; Folaranmi et al. 2015).

The great success of the first RV study has led to many RV prediction programs (Dalsass et al. 2019). The currently reported open-source RV programs could be characterized based on the algorithmic approaches or input feature types. The algorithmic approaches include rule-based filtering and machine learning (ML) classification methods. NERVE, the first publicly available rule-based filtering RV program, is a standalone software published in 2006 (Vivona, Bernante, and Filippini 2006). Four years later, the first web-based filtering RV program, Vaxign, was developed similar to NERVE but with additional analyses (He, Xiang, and Mobley 2010). Vaxign has been applied in vaccine design studies against more than ten pathogenic bacteria such as *Helicobacter pylori* (Navarro-Quiroz et al. 2018), *Acinetobacter baumannii* (Singh et al. 2016), *Mycobacterium spp.* (Hossain et al. 2017). Following NERVE and Vaxign, two other filtering-based RV programs, Jenner-predict server (Jaiswal et al. 2013) and VacSol (Rizwan et al. 2017) were published (Jaiswal et al. 2013; Rizwan et al. 2017). All these currently available rule-based filtering RV programs use only biological features as the data input.

ML classification has also been used for RV prediction. VaxiJen was the first ML classification RV program published in 2007 (Doytchinova and Flower, 2007). Bowman et al. and Heinson et al. extended the training data of VaxiJen and revised the ML algorithm (Bowman et al. 2011; Heinson et al. 2017). Their final training data, termed 200BPA, consisted of 200 bacterial protective antigens (BPAgs) and 200 non-protective proteins. The non-protective proteins were selected if it had no homology (BLASTp E-value ≤ 10E-3) to the BPAgs. A major difference between VaxiJen and Bowman-Heinson was that VaxiJen used physicochemical features of the input proteins while the later program used biological features. The second lineage of ML-based RV prediction program originated from the development of ANTIGENpro in 2010 (Magnan et al. 2010). ANTIGENpro collected protective antigens (PAgs) from the literature combined with the positive and negative samples tested via protein microarrays probed with sera from naive, exposed, and vaccinated individuals. Rahman et al. revised the algorithmic method of ANTIGENpro and developed Antigenic using the same training data (Rahman et al. 2019). Both ANTIGENpro and Antigenic used physicochemical features as the data input. The authors of these two papers argued that proteins being able to elicit a significant antibody response could be considered "protective antigens". However, data collected based on antibody responses did not guarantee to be protective. Such data lack the results from protection assays in at least laboratory animal models. More importantly, antibody production does not capture cell-mediated immunity, which is often an essential protective immune mechanism. For example, *Brucella* vaccine RB51-induced protection is purely based on cell-mediated immunity, and its induced antibody response does not offer any observed protection (Jimenez de Bagues et al. 1994).

All of the ML RV programs mentioned earlier were not designed to predict eukaryotic vaccine candidates. Goodswen et al. developed the first and the only ML RV targeting eukaryotic pathogens to our best knowledge (Goodswen, Kennedy, and Ellis 2013). However, due to the lack of reported eukaryotic PAgs, proteins which are surface-exposed and have at least one T-cell epitopes were treated as positive in this study. These collected data might lack supporting experimental evidence. Furthermore, a protein with epitopes does not mean that this protein can elicit protective immune responses (Flower et al. 2010). An independent resource, Protegen, had manually collected 590 protective antigens over 100 infectious diseases caused by pathogens (bacteria, viruses, and parasites) and non-infectious diseases, including cancers and allergies (B. Yang et al. 2011). Each of these collected protective antigens can elicit a protective immune response, which has been experimentally verified by at least one laboratory animal model. A preliminary ML RV study trained on the Protegen data reported high PAgs prediction accuracy (He and Xiang, 2012). Protegen has doubled the number of annotated pathogen PAgs since its initial release in 2011.

Although a significant effort has been made to enhance the RV prediction with ML, there are still many obstacles in ML-based RV prediction. First, all currently available programs use either biological properties or physicochemical properties for input protein sequence annotations. Previous studies reported that the protectiveness of BPAgs was significantly correlated to biological properties (E. Ong, Wong, and He 2017) and physicochemical properties (Mayers et al. 2003). Studies using ML algorithms trained on data with physicochemical properties annotated also showed high BPAg prediction accuracy. Therefore, the relations of BPAgs to biological and physicochemical properties deserved a more in-depth analysis and should be combined with annotated proteins in the training data for better BPAg prediction. Secondly, the quality of the benchmarking datasets varied in current reported studies. As mentioned above, the testing data used to evaluate ANTIGENpro and Antigenic was primarily based on the antibody responses and might miss the cell-mediated immune responses. Therefore, the dataset of ANTIGENpro and Antigenic was excluded from this study. Finally, the 200BPA data from VaxiJen and Bowman-Heinson only included PAgs with supporting experimental evidence. The negative samples were randomly selected from non-homologous proteins to the PAgs. The random under-sampling may not reflect the real distribution of PAgs in the proteome. Besides, VaxiJen and Bowman-Heinson model were not evaluated using external independent dataset.

In the current Vaxign-ML study, epitope information was not incorporated into the pipeline. The prediction of epitopes has been an active area of vaccine design. The IEDB database and IEDB-AR resources (Fleri et al. 2017) provide comprehensive T cell and B cell epitope query, prediction, and analysis tools. However, the epitopes' prediction is dependent on the host information (e.g., MHC alleles and antibodies). The training dataset in Vaxign-ML consisted of experimentally verified protective antigens manually annotated from studies in over ten host species. Therefore, the prediction of BPAgs in Vaxign-ML did not take host species into account, and the T cell or B cell epitope predictions were not included. Current epitope-based BPAg prediction methods such as iVAX (L Moise et al. 2015) often depend on the epitopes' frequency or density located on the protein. However, such epitope measurement may not necessarily translate into protective immune responses. Despite the uncertainty of correspondence between epitope and protective immune response, we implemented an epitope-based method using IEDB epitope prediction tools. The performance of Vaxign-ML and the other four BPAg prediction methods was to the epitope-based method in this study.

This paper presented a systematic evaluation of a supervised ML classification RV program trained on the Protegen BPAgs with their biological and physicochemical features annotated. The BPAgs and the non-protective proteins were first carefully checked for homology to ensure training data quality. Three data resampling strategies were applied to the original data due to imbalance classes in the training data. Nested five-fold cross-validation and leave-one-pathogen-out validation were used to evaluate five supervised ML algorithms with feature selection and hyperparameter optimization. The best performing model, termed Vaxign-ML, was benchmarked using a curated external independent dataset and demonstrated superior predictive performance.

3.3 Methods

The overall project workflow is described in Figure 3-1. In brief, positive samples and negative samples were downloaded and processed from Protegen and Uniprot (B. Yang et al. 2011; The UniProt Consortium 2008). The biological and physicochemical features for these protein sequences were annotated using publicly available bioinformatics software. Four data resampling strategies and Five supervised ML classification algorithms were trained and evaluated. The best model's performance, named Vaxign-ML, was compared to four BPAg prediction methods and one epitope-based method using a curated external independent dataset.

3.5.1 Data Preparation

BPAgs with supporting experimental evidence were downloaded from the Protegen database (B. Yang et al. 2011). As of 2019-07-31, Protegen included 584 BPAgs from 50 Gram+ and Gram-pathogenic bacteria. BPAgs with sequence similarity over 30% were considered homologous proteins, a commonly accepted threshold for homologous proteins (Pearson 2013), and removed from the study to avoid potential bias. The final positive samples in the original data consisted of 397 BPAgs. A set of “gold-standard” non-protective proteins does not exist. Therefore, the training dataset's negative samples were selected based on its sequence dis-similarity to the BPAgs, as described in previous ML-based protective antigen prediction studies (Bowman *et al.*, 2011; Doytchinova and Flower, 2007; Heinson *et al.*, 2017). The whole pathogen proteomes of the 50 pathogenic bacteria were downloaded from the Uniprot database (The UniProt Consortium 2008). Any pathogen proteins with sequence similarity less than 30% to the BPAgs were kept, and homologous proteins were also removed.

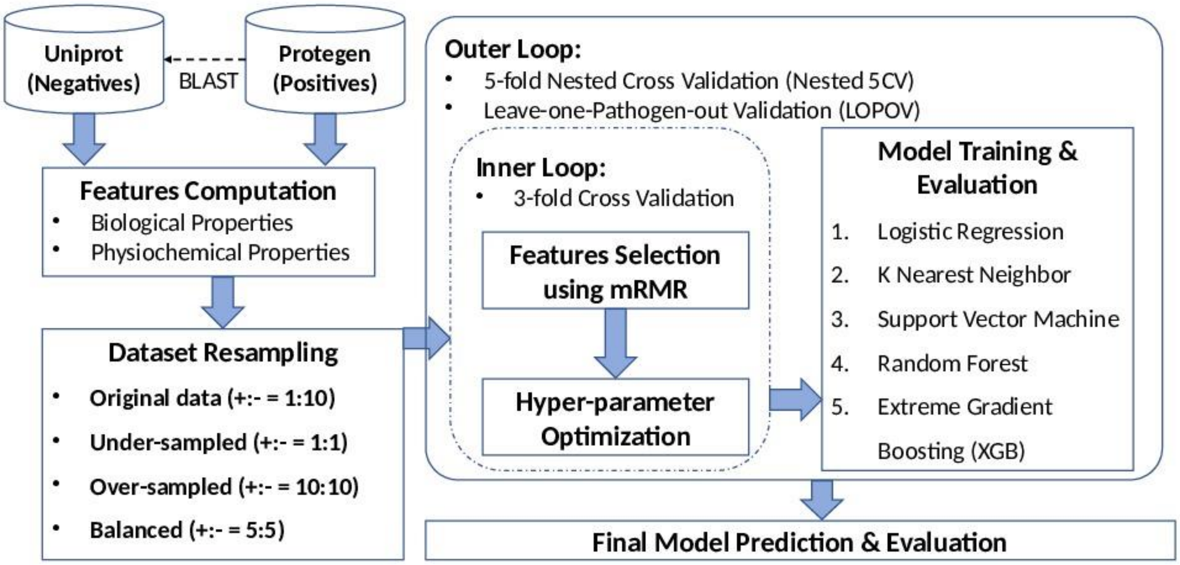

Uniprot (Negatives)
BLAST
Protegen (Positives)
Features Computation
Biological Properties
Physiochemical Properties
Dataset Resampling
Original data (+:- = 1:10)
Under-sampled (+:- = 1:1)
Over-sampled (+:- = 10:10)
Balanced (+:- = 5:5)
Outer Loop:
5-fold Nested Cross Validation (Nested 5CV)
Leave-one-Pathogen-out Validation (LOPOV)
Inner Loop:
3-fold Cross Validation
Features Selection using mRMR
Hyper-parameter Optimization
Model Training & Evaluation
1. Logistic Regression
2. K Nearest Neighbor
3. Support Vector Machine
4. Random Forest
5. Extreme Gradient Boosting (XGB)
Final Model Prediction & Evaluation

Figure 3-1 Overall Vaxign-ML workflow.

This flowchart depicted the entire process to train and evaluate machine learning-based reverse vaccinology models. See main text for details.

Vaxign-ML used two categories of features for each of the protein sequences: biological features and physicochemical features. Biological features including the Gram(+/-) stain, subcellular localization (Yu et al. 2010), adhesin probability (Sachdeva et al. 2005), transmembrane helix (Krogh et al. 2001), signal peptide (Petersen et al. 2011), and immunogenicity (Fleri et al. 2017) were computed using publicly available bioinformatics software programs. On the other hand, the analyzed physicochemical features include the compositions, transitions, and distributions (Dubchak et al. 1995), quasi-sequence-order (Chou 2000), Moreau-Broto auto-correlation (Lin and Pan, 2001; Feng and Zhang, 2000), and Geary auto-correlation (Sokal and Thomson, 2006) of various physicochemical properties such as charge, hydrophobicity, polarity, and solvent accessibility, etc. (S. A. K. Ong et al. 2007). A total of 509 biological and physicochemical features were annotated for each of the protein sequences in the original data.

The original data were imbalanced and had a dimension of 4,367 samples (positive-to-negative classes ratio = 1:10) and 509 features. In order to study the effect of class imbalance, three data resampling strategies were implemented. Firstly, the negative samples in the original data were randomly sampled without replacement (positive-to-negative classes ratio = 1:1). Secondly, the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al. 2002) was applied to the original data to increase the number of positive samples (positive-to-negative classes ratio = 10:10). Finally, the balanced resampling strategy was applied by combining both under-sampling and over-sampling strategies. The negative samples in the original training data were randomly sampled without replacement to have five times the size of positive samples. Then, the positive samples were over-sampled using SMOTE (positive-to-negative ratio = 5:5).

3.5.2 Supervised machine learning classification

Five supervised ML classification algorithms were used in this study, including logistic regression (LR), support vector machine (SVM), k-nearest neighbor (KNN), random forest (RF) (Pedregosa et al. 2012), and extreme gradient boosting (XGB) (T. Chen and Guestrin 2016). The best performing model was trained and named “Vaxign-ML”. The output of Vaxign-ML is the percentile rank score from the final ML classification model, termed “protegenicity”.

A nested five-fold cross-validation (N5CV) was applied to evaluate all supervised ML classification models. The training data, including original data, under-sampled, over-sampled, and balanced, were randomly split into five parts while preserving the percentage of positive and negative samples. One important note is that the data resampling was only performed after the N5CV splitting to avoid duplicated positive samples in both training and testing data. Among the five parts, four parts were for training. Feature selection with mRMR (Ding and Peng 2003) and hyperparameter optimization were applied before training all classification models. The remaining part was used as the testing set for model evaluation.

In order to determine whether the discriminative power of the prediction models depended on the immunogenic potential in the Protegen dataset rather than sequence dis-similarity, the negative dataset was randomly split into two sets with sequence identity less than 30%. The same N5CV was applied to confirm that the discriminative performance depended on the PAg potential in the Protegen database rather than sequence dis-similarity.

To have a more unbiased estimation of the classification performance and to mimic the situation where vaccine candidates would be needed for a new emerging pathogen, a leave-one-pathogen-out validation (LOPOV) was implemented. Ten tested pathogens included four Gram+ pathogens (*Mycobacterium tuberculosis, Staphylococcus aureus, Streptococcus pneumoniae,* and *Streptococcus pyogenes*) and six Gram- pathogens (*Helicobacter pylori, Neisseria meningitidis,* *Brucella abortus, Escherichia coli*, *Yersinia pestis*, *Haemophilus parasuis*). The positive and negative samples from these ten pathogens were held out as testing sets. The remaining samples were used for training. Data sampling, feature selection, and hyperparameter optimization were applied before training all classification models similar to the N5CV.

3.5.3 Benchmarking and evaluation with an independent dataset

A curated external independent dataset was created to benchmark the best performing model (Vaxign-ML). Dalsass et al. collected a list of 100 bacterial protective antigens termed 100BPA (Dalsass et al. 2019). However, 100BPA only includes positive samples. Another dataset consisted of 200 positive and 200 negative samples (200BPA) was initially created for the development of the VaxiJen program (Doytchinova and Flower, 2007) and later extended by Bowman et al. and Heinson et al. (Bowman et al. 2011; Heinson et al. 2017). Both 100BPA and 200BPA were combined and used for benchmarking. To ensure this external independent dataset's quality, all positive samples in 100BPA and 200BPA were checked against the BPAgs in Protegen. Any duplicated positive samples were then removed. Meanwhile, the negative samples in 200BPA were also evaluated to ensure no supporting experimental evidence from the literature as our initial effort to address the “true negative” dataset for ML-based BPAgs prediction. The final curated external independent dataset consisted of 131 positive and 118 negative samples, named “iBPA”.

Vaxign-ML was compared to four BPAg candidate prediction programs: Vaxign (He, Xiang, and Mobley 2010); VaxiJen (Doytchinova and Flower 2007); Heinson-Bowman (Heinson et al. 2017; Bowman et al. 2011); Antigenic (Rahman et al. 2019), and one epitope-based prediction method using IEDB-AR epitope prediction tools (Dhanda et al. 2019). For Vaxign prediction, we used two suggested criteria: surface-exposed proteins (subcellular localization in the cell wall, outer membrane, or extracellular space) and adhesin probability > 0.51. The recommended cut-off (0.5) was used for BPA prediction by VaxiJen. For the Heinson-Bowman method, a nested cross-validated SVM prediction model was tested with the iBPA dataset annotated by the top ten significant biological properties (Heinson et al. 2017). Vaxign-ML had major differences compared to the Heinson-Bowman method, including the quality of training data, selection of ML algorithms, data resampling methods, and annotated features. For Antigenic, the default settings and cut-off values were used to call BPAgs from the iBPA dataset. An epitope-based prediction method was implemented by thresholding the percentile ranking of the epitope frequency in the iBPA dataset, compared to 10,000 randomly selected background proteins. The epitope frequency of a protein was calculated by summing the top 1% predicted MHC-I restricted epitopes and top 10% predicted MHC-II restricted epitopes across a set of reference set alleles (Greenbaum et al. 2011; Weiskopf et al. 2013) using the IEDB-AR epitope prediction tools (Dhanda et al. 2019). A percentile ranking threshold of 58% was used after optimizing the true positive rate and false positive rate. Proteins in the iBPA dataset with epitope frequency above 58% percentile rank compared to the random background were considered to have significant immunogenic potential.

The receiver operating characteristics (ROC) curve, precision-recall (PR) curve, weighted F1-score (WF1), and Matthew's correlation coefficient (MCC) were computed for both N5CV and LOPOV. An additional evaluation was performed for the LOPOV. For the benchmarking of the Vaxign-ML with iBPA, the precision, recall, weighted F1, and MCC metrics were calculated. Finally, the protegenicity scores of 20 proteins from five *M. tuberculosis* (MTB) vaccines undergoing clinical trials and the licensed DPT vaccine (*Corynebacterium diphtheriae, Bordetella pertussis,* and *Clostridium tetani*) were calculated.

3.4 Results

3.3.1 Effect of data resampling strategies on the classification

All ML classification algorithms performed worse when trained on under-sampled and over-sampled data than original or balanced data (Table 3-1). When evaluating the performance of different data resampling strategies based on AUROC (Figure 3-2), almost all ML classification algorithms had high values (AUROC = [0.89, 0.96]) except KNN (AUROC = 0.76). Since the data resampling step was only performed on the training data during the Nested-5CV but not the testing data, the AUPRC, WF1, and MCC metrics were less prone to the imbalanced classes of the data than AUROC. All ML algorithms trained on under-sampled and over-sampled data consistently had lower AUPRC (Figure 3-3), WF1, and MCC. The balanced data did not significantly improve the performance of the ML algorithms used in this study. Furthermore, the MCC values of the SVM and RF trained on balanced data were dramatically reduced, which indicated high degrees of over-fitting in these two ML models. KNN algorithm was more sensitive to the sample class ratio changes because all four metrics of the models trained on under-sampled, over-sampled, and balanced data were lower than the original data. Although the LR and XGB trained on balanced data had slightly higher AUPRC, these two ML models had lower WF1 and MCC when trained on original data. Therefore, balancing the positive and negative samples did not significantly improve the BPAg prediction.

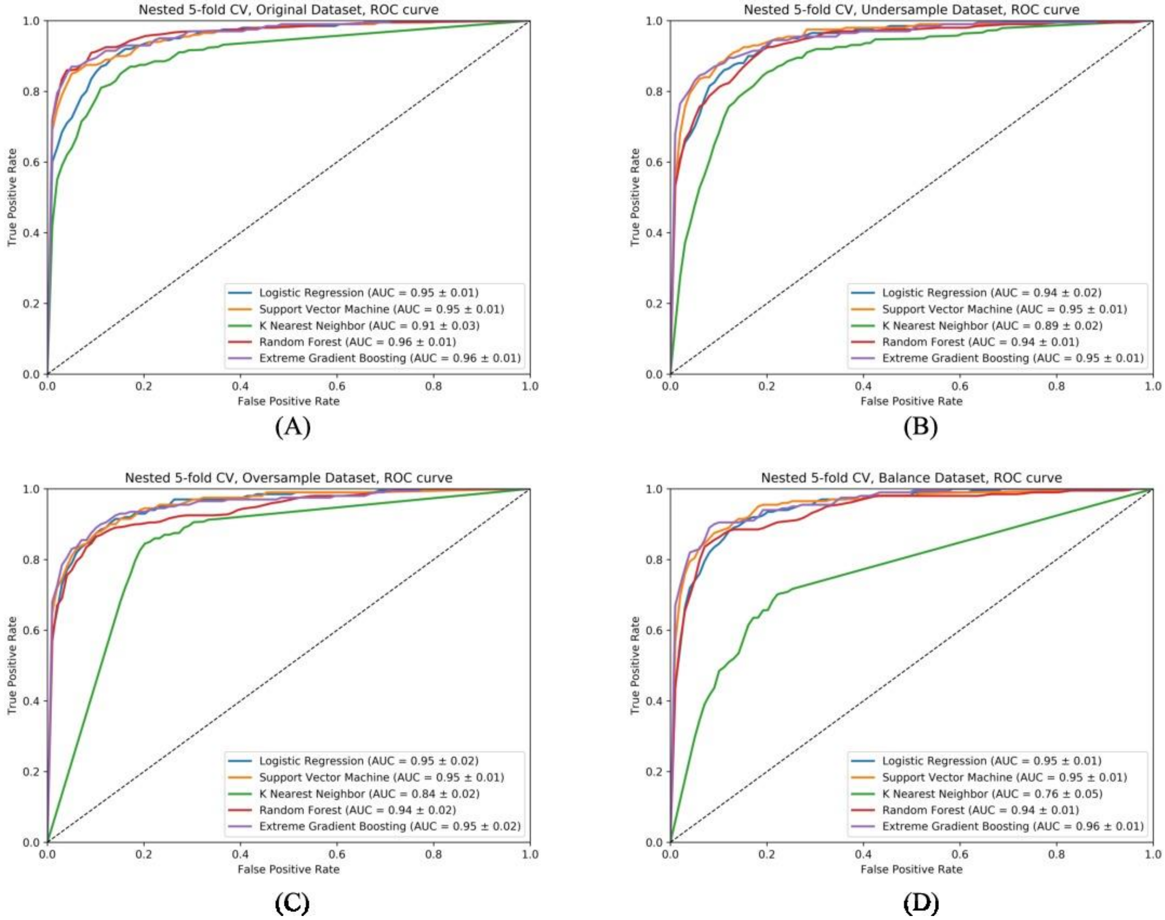


Figure 3-2 The average ROC curves of five machine learning algorithms in nested five-fold cross-validation with different data resampling strategies.

The average ROC curves of five machine learning algorithms, logistic regression, support vector machine, k nearest neighbor, random forest, and extreme gradient boosting, were evaluated by the nested five-fold cross-validation with (A) original data; (B) under-sampled data; (C) over-sampled data; and (D) balanced data. The k nearest neighbor algorithm was more prone to over-fitting as the resampling the dataset (in particular, oversampled and balanced) drastically reduce the prediction performance.

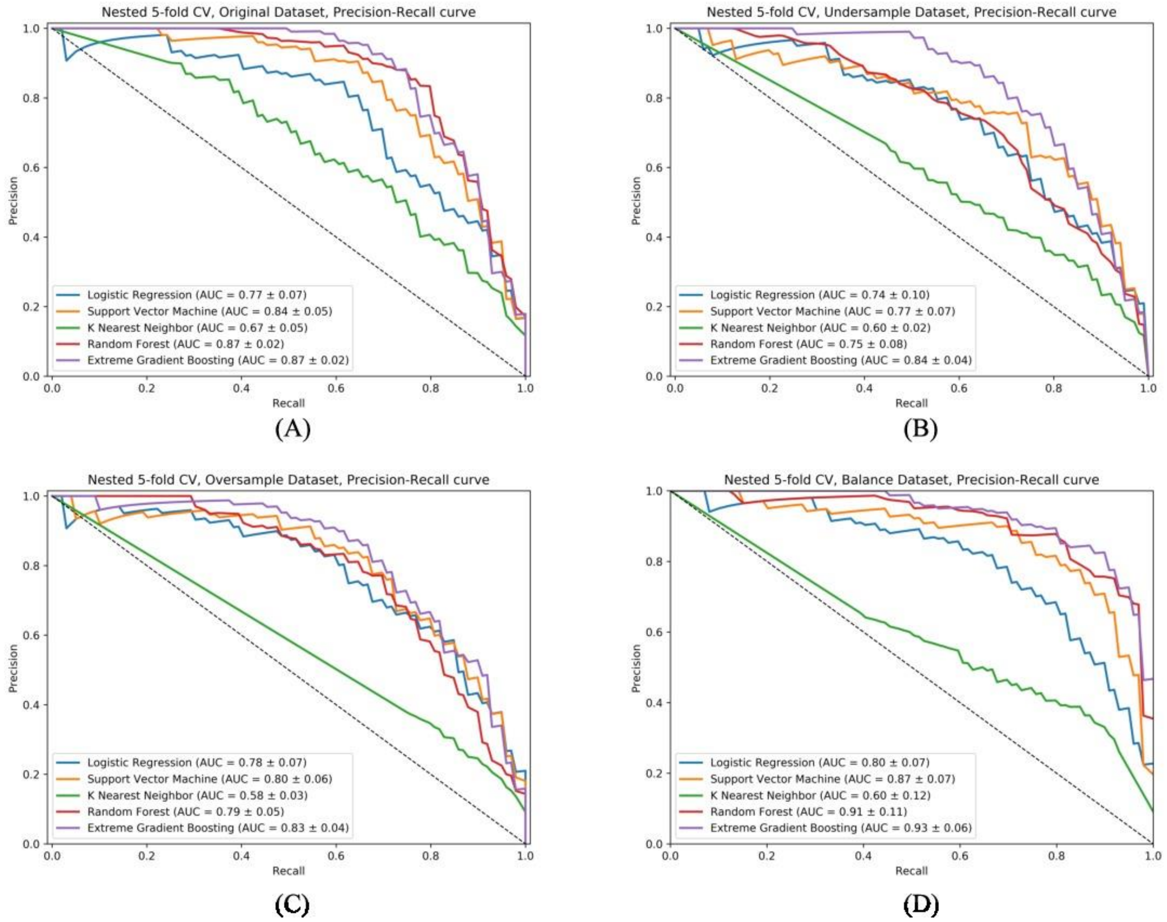


Figure 3-3 The average precision-recall curves of five machine learning algorithms in nested five-fold cross-validation with different data resampling strategies.

The average precision-recall curves of five machine learning algorithms, logistic regression, support vector machine, k nearest neighbor, random forest, and extreme gradient boosting, were evaluated by the nested five-fold cross-validation with (A) original data; (B) under-sampled data; (C) over-sampled data; and (D) balanced data. Comparing to other four algorithms, the k nearest neighbor algorithm had the lowest precision-recall performance. More importantly, the k nearest neighbor algorithm was more prone to over-fitting as the resampling the dataset (in particular, oversampled and balanced) drastically reduce the prediction performance.

Table 3-1 Nested five-fold cross-validation evaluation metrics of five machine learning algorithms trained using four data resampling methods.

Original Data				Under-sampled				Over-sampled				Balanced			
AUROC	AUPRC	WF1	MCC	AUROC	AUPRC	WF1	MCC	AUROC	AUPRC	WF1	MCC	AUROC	AUPRC	WF1	MCC
Logistic Regression (LR)															
0.95	0.77	0.93 (±0.02)	0.60	0.94	0.75	0.93 (±0.02)	0.60	0.95	0.78	0.93 (±0.03)	0.63	0.95	0.8	0.91 (±0.03)	0.58
Support Vector Machine (SVM)															
0.95	0.84	0.96 (±0.006)	0.76	0.95	0.77	0.92 (±0.02)	0.58	0.95	0.8	0.94 (±0.01)	0.67	0.95	0.87	0.87 (±0.002)	0.03
K-Nearest Neighbor (KNN)															
0.91	0.67	0.93 (±0.01)	0.59	0.89	0.6	0.83 (±0.04)	0.43	0.84	0.58	0.83 (±0.01)	0.41	0.76	0.6	0.87 (±0.01)	0.31
Random Forest (RF)															
0.96	0.87	0.96 (±0.005)	0.76	0.94	0.75	0.93 (±0.03)	0.58	0.94	0.79	0.95 (±0.01)	0.69	0.94	0.91	0.87 (±0.003)	0.06
Extreme Gradient Boosting (XGB)															
0.96	0.87	0.96 (±0.02)	0.79	0.95	0.84	0.95 (±0.006)	0.71	0.95	0.83	0.95 (±0.002)	0.72	0.96	0.93	0.95 (±0.009)	0.72

3.3.2 Extreme gradient boosting as the best performing model

In N5CV, the XGB model consistently had the highest performance compared to the other four ML algorithms when trained on four different data resampling methods (Table 3-1). Three models, including XGB trained on original data (XGB-original) and balanced data (XGB-balance), and random forest trained on original data had the highest area under ROC curve (AUROC). XGB-original had the highest WF1 and MCC while XGB-balance had the highest area under PRC (AUPRC). Both XGB-original and XGB-balance were evaluated with the LOPOV and had similar AUROCs for the ten pathogens held out in LOPOV and the average of these pathogens (Figure 3-4 A&B). However, the XGB-original had a higher average AUPRC (Figure 3-4 C&D), WF1, and MCC (Table 3-2). Therefore, the best performing XGB-original was selected as the final BPAg prediction model and termed Vaxign-ML for benchmarking.

On a separate note, the N5CV results of five ML prediction models to discriminate two sets of randomly selected dis-similar non-BPAgs were approximately equivalent to random prediction (Figure 3-5). The discriminative power of the current BPAg prediction pipeline was indeed dependent on the immunogenic potential in the Protegen dataset rather than sequence dis-similarity.

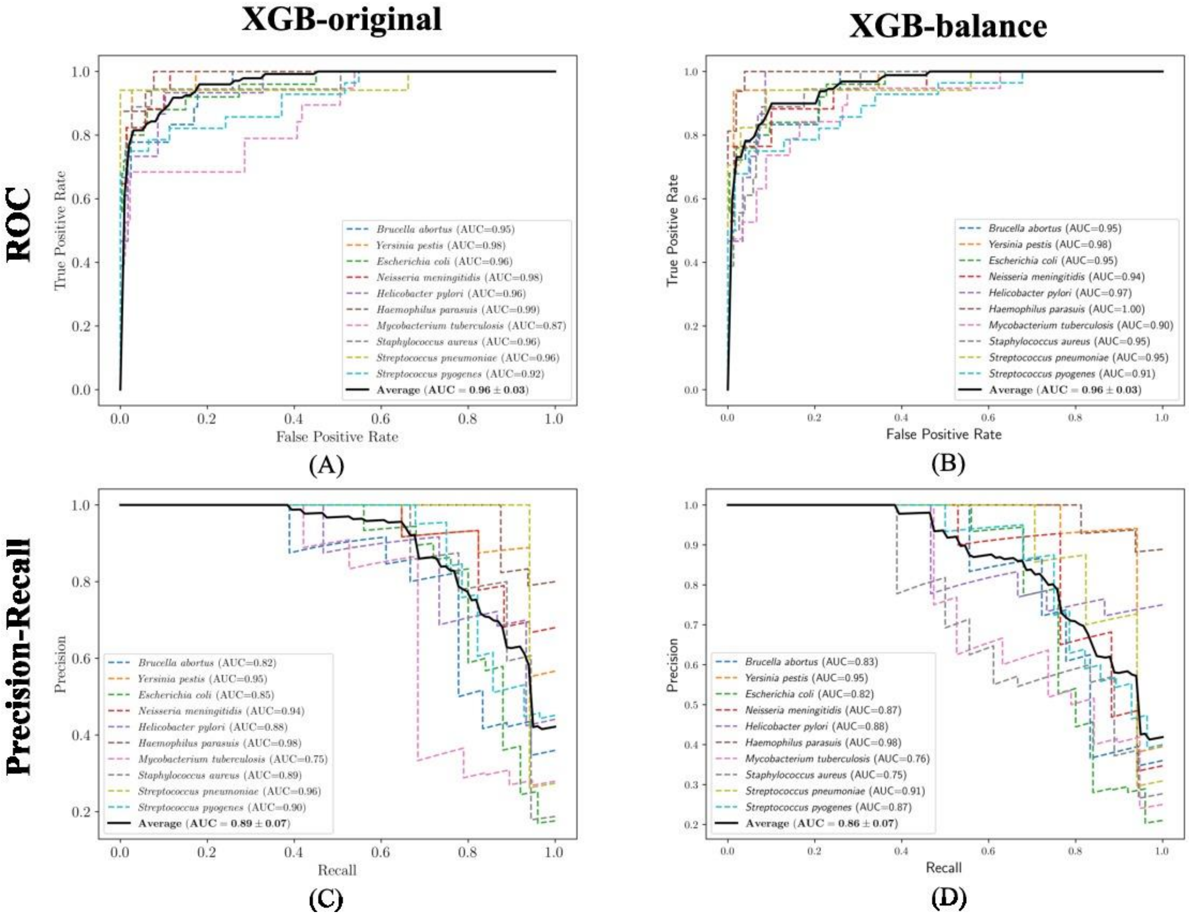


Figure 3-4 Leave-one-pathogen-out validation (LOPOV) of the two best performing models, XGB-original and XGB-balance.

The top row and bottom row plotted the ROC and precision-recall (PR) curves, respectively. Ten pathogens were tested in the LOPOV (color dash lines), as shown in the legend. The average of the ROC and PR curves were also plotted (black line). The average Area Under ROC curve (AUROC) of both XGB-original (A) and XGB-balance (B) performed equally well. However, XGB-original had a higher average Area Under PR curve (AUPRC) than the XGB-balance. Additionally, XGB-balance had lower performance when tested on the *M. tuberculosis*, *S. aureus*, while XGB-original had lower performance than the former.

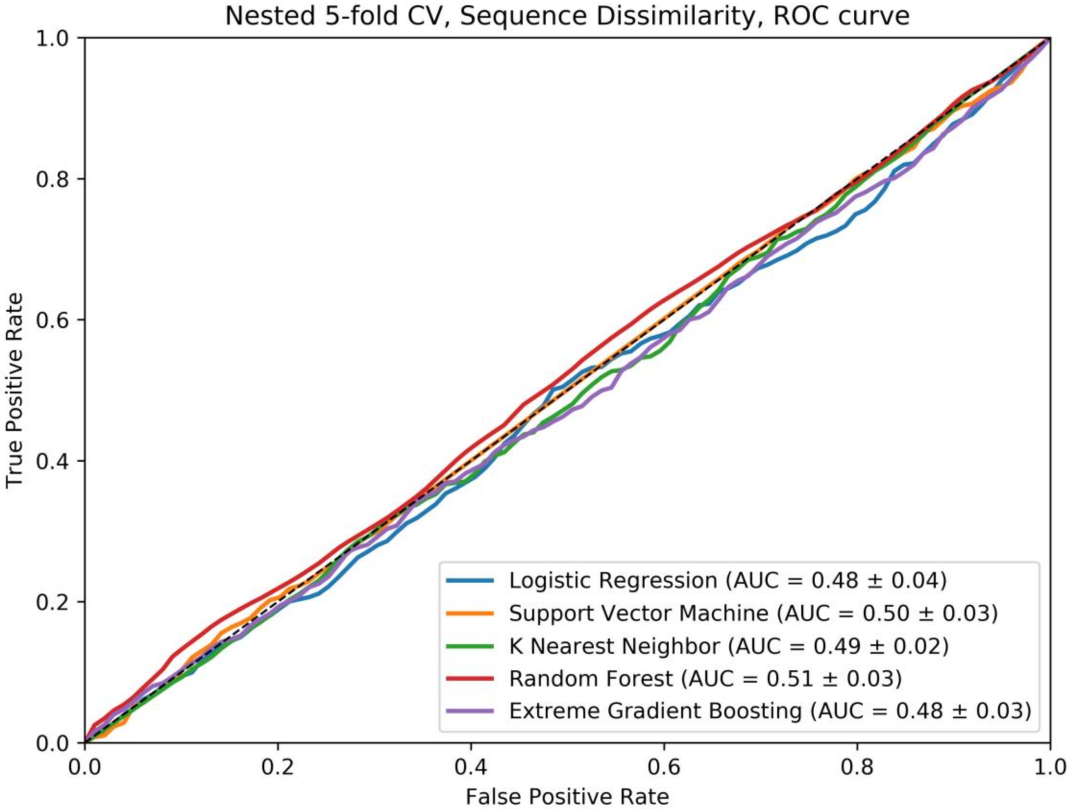


Figure 3-5 The ROC curves of the five machine learning algorithms to discriminate two sets of non-antigen proteins with less than 30% sequence identity.

The ROC curves of the discrimination predicted by five machine learning algorithms between two sets of non-antigen proteins with less than 30% sequence identity were evaluated by the nested five-fold cross-validation. The area under ROC and PRC were approximately 0.5, which is equivalent to random prediction. Therefore, the discriminative power of the presented Vaxign-ML was indeed dependent on the immunogenic potential of the protective antigens collected in the Protegen database rather than sequence dis-similarity.

Table 3-2 Leave-one-pathogen-out evaluation metrics of five machine learning algorithms trained using four data re-sampling methods.

Original Data				Under-sampled				Over-sampled				Balanced			
AUROC	AUPRC	WF1	MCC	AUROC	AUPRC	WF1	MCC	AUROC	AUPRC	WF1	MCC	AUROC	AUPRC	WF1	MCC
Logistic Regression (LR)															
0.94	0.82	0.86 (±0.07)	0.58	0.95	0.83	0.80 (±0.13)	0.56	0.94	0.82	0.84 (±0.07)	0.59	0.95	0.83	0.77 (±0.12)	0.51
Support Vector Machine (SVM)															
0.96	0.88	0.91 (±0.04)	0.70	0.95	0.87	0.88 (±0.05)	0.64	0.96	0.85	0.90 (±0.03)	0.67	0.97	0.88	0.91 (±0.03)	0.71
K-Nearest Neighbor (KNN)															
0.91	0.77	0.90 (±0.04)	0.65	0.9	0.71	0.71 (±0.09)	0.43	0.84	0.67	0.74 (±0.05)	0.44	0.81	0.62	0.76 (±0.07)	0.45
Random Forest (RF)															
0.95	0.89	0.93 (±0.03)	0.75	0.95	0.83	0.90 (±0.04)	0.67	0.96	0.84	0.91 (±0.03)	0.68	0.95	0.84	0.91 (±0.03)	0.69
Extreme Gradient Boosting (XGB)															
0.96	0.89	0.94 (±0.02)	0.77	0.94	0.85	0.89 (±0.04)	0.67	0.96	0.87	0.93 (±0.03)	0.75	0.96	0.87	0.92 (±0.03)	0.73

3.3.3 Biological and physicochemical features in Vaxign-ML

The mRMR feature selection and hyperparameter optimization steps in Vaxign-ML suggested an optimal set of 180 features. The biological features, including subcellular localization, adhesin probability, transmembrane helix, and immunogenicity score, are frequently used in filtering-based vaccine prediction programs (e.g., NERVE and Vaxign). However, in Vaxign-ML, these features only accounted for 11.4% of the importance in the final XGB model (Figure 3-6).

The pathogen’s Gram(+/-) stain was excluded from the Vaxign-ML due to its lack of contribution to the outcome. Although the physicochemical properties are often difficult to be interpreted during vaccine design, these features accounted for 88.6% of the importance. In particular, amino acid composition (13.1%), charge (12.3%), hydrophobicity (8.7%), free energy (7.6%), and polarity (7.5%) were the top five important categories of the physicochemical properties in the Vaxign-ML model (Figure 3-6).

3.3.4 Benchmarking Vaxign-ML

The final constructed model, Vaxign-ML, was benchmarked and compared to currently publicly available BPAg candidate prediction programs (Vaxign, VaxiJen, and Antigenic), Heinson-Bowman, and epitope-based method. The performance of these programs was evaluated based on the iBPA dataset described in section 3.5.3. Vaxign-ML had the highest performance in three out of four metrics (recall = 0.81, WF1 = 0.76 and MCC = 0.51) among all methods (Table 3-3). The rule-based Vaxign program had a higher precision value (0.79) than the Vaxign-ML (0.75), likely due to the more restrictive rules in the Vaxign program. Vaxign-ML also out-performed the ML-based Heinson-Bowman method, suggesting that the enhancement of Vaxign-ML in terms of training data quality, annotated features, and ML algorithms the BPAg prediction.

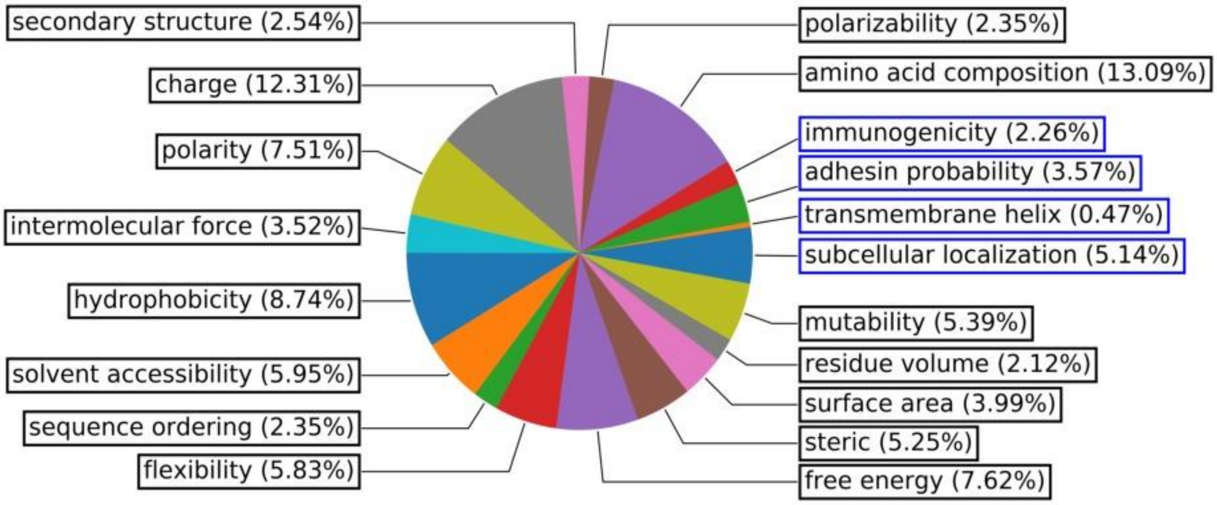


Figure 3-6 Categories of features and its weight in the Vaxign-ML model.

Features in blue and black boxes are biological and physicochemical features, respectively.

Table 3-3 Benchmarking of Vaxign-ML compared to publicly available programs and epitope-based method.

	Recall	Precision	WF1	MCC
Vaxign-ML	0.81	0.75	0.76	0.51
Heinson-Bowman	0.72	0.69	0.68	0.37
VaxiJen	0.69	0.68	0.66	0.32
Vaxign	0.32	0.79	0.56	0.27
Antigenic	0.50	0.52	0.49	-0.02
Epitope-based	0.63	0.65	0.62	0.24

Abbreviation: WF1: weighted F1 score; MCC: Matthew’s correlation coefficient.

Finally, the epitope-based prediction method had lower performance than Vaxign-ML, Heinson-Bowman, and VaxiJen across all four metrics in the context of BPAg prediction.

3.3.5 Vaxign-ML predicting current clinical trial or licensed vaccines

As a final validation, Vaxign-ML was used to calculate and rank the corresponding protegenicity scores of five clinical trial MTB vaccines and one licensed DPT vaccine (*C. diphtheriae, B. pertussis,* and *C. tetani*) (Table 3-4). The protegenicity score was the percentile rank score generated by Vaxign-ML trained on the entire original data. A total of 20 proteins were included in these six vaccines, and all of them had a predicted protegenicity score of over 90%. In other words, these 20 proteins were ranked in the top 10% of best BPAg candidates by the Vaxign-ML.

Table 3-4 Vaxign-ML prediction of five MTB vaccines currently in the clinical trial and one licensed DPT vaccine.

Vaccine	Protein	Protegenicity Score (%)
Mycobacterium tuberculosis		
H1, H4, H56	Ag85B	95.21
H1, H56	ESAT-6	94.89
H4	EsxH	90.91
H56	Rv2660	91.23
M72	PPE18	92.05
	PepA	94.28
ID93	EsxW	90.95
	PPE42	91.89
	EsxV	91.53
	Rv1813	91.09
Bordetella pertussis		
Pertussis vaccine	Pertussis toxin subunit 1	94.07
	Pertussis toxin subunit 2	91.53
	Pertussis toxin subunit 3	90.91
	Pertussis toxin subunit 4	91.37
	Pertussis toxin subunit 5	91.62
	Filamentous hemagglutinin	98.9
	Pertactin autotransporter	95.35
	Fimbrial protein	95.99
Corynebacterium diphtheriae		
Diphtheria vaccine	Diphtheria toxin	97.73
Clostridium tetani		
Tetanus vaccine	Tetanus toxin	99.79

3.5 Discussion

Overall, Vaxign-ML showed superior performance in BPAg prediction compared to all other BPAg prediction methods. Our study also demonstrated the significance of both biological and physicochemical properties in ML-based RV prediction. Finally, the results of Vaxign-ML highlighted the critical role of physicochemical properties and might have an implication in structural vaccinology.

Our study showed that Vaxign-ML (extreme gradient boosting trained on original data with mRMR feature section and hyperparameter optimization) was the best performing supervised ML classification model with an unbiased N5CV and LOPOV validations. The LOPOV validation also assessed how well the model could predict BPAgs when encountering a new emerging pathogen. The benchmarking of Vaxign-ML using a curated external independent dataset suggested the superior performance of Vaxign-ML to its predecessor with the highest recall, weighted F1 score, and Matthew's correlation coefficient. Notably, the iBPA dataset was derived and curated from the VaxiJen program. Vaxign-ML was trained on the Protegen dataset and did not encounter any samples in the iBPA, and yet Vaxign-ML had better predictive power than VaxiJen. Although the preceding rule-based Vaxign program missed many potential candidates (recall = 0.32, Table 3-3), the rule-based RV method had better potential in filtering out non-protective proteins, as demonstrated by the highest precision value among all four programs being studied. A combination of Vaxign-ML, followed by filtering similar to Vaxign, might be a future direction to enhance the predictive performance further.

Vaxign-ML is the first RV method that incorporates both biological and physicochemical properties. Historically, the biological and physicochemical features had been treated as two isolated silos in the field of BPAgs prediction. Several ML RV studies predicted BPAgs based on the physicochemical properties of the input proteins (Doytchinova and Flower, 2007; Rahman et al., 2019; Magnan et al., 2010). In this paper, all the physicochemical features were grouped into one category to better interpret individual properties (Figure 3-6). Mayers et al. reported that known protein vaccine antigens had distinct characteristics in amino acid composition, hydrophobicity, flexibility, and mutability (Mayers et al., 2003), which accounted for 13.1%, 8.7%, 5.8%, and 5.4% of the Vaxign-ML feature importance respectively. Polarity (7.5%) and charge (12.3%) had an important implication in vaccine design. Studies showed that antibody-antigen interfaces are likely polar (Hebditch and Warwicker, 2019). On the other hand, highly negatively charged vaccines often possess limited cell uptake ability, whereas highly positively charged vaccines exert significant cytotoxicity (Zhang et al., 2018). Positively charged nanoparticles induce a more robust and systemic antibody response in a recent nano-based vaccine delivery study (Fromen et al., 2015). Finally, free energy is an essential factor in the structural design of the chimeric subunit vaccine (Nazarian et al., 2012), as well as describing the binding between epitope and major histocompatibility complex (Patronov and Doytchinova, 2013).

The significance of biological property profiles in BPAgs (Ong et al., 2017) had been utilized by both rule-based RV programs (He et al., 2010; Jaiswal et al., 2013; Vivona et al., 2006; Rizwan et al., 2017) and supervised ML BPAg classifications (Bowman et al., 2011; Heinson et al., 2017). Vaxign-ML took substantial consideration into the biological properties, including subcellular localization, adhesin probability, and immunogenicity. However, some biological features (e.g., Gram stain and transmembrane helix) might not be significantly associated with protectiveness and were considered for practical reasons (pathogen characterization and efficacy in recombinant protein isolation) (He et al., 2010). The biological features of the protein sequences in the training data are predictions and are dependent on the performance of the corresponding bioinformatics tools. Although some of these bioinformatics tools already included physicochemical properties in the prediction pipelines, these properties were utilized to address specific scientific questions (e.g., subcellular location prediction, signal peptide). In Vaxign-ML, these biological features of attributed to 11.4% of the importance in BPAg prediction and could be the key factor leading to better prediction performance by Vaxign-ML compared to VaxiJen and Antigenic.

Currently, Vaxign-ML does not consider the epitopes and structure in the prediction model. The comparison of Vaxign-ML and the epitope-based method, which was benchmarked using the iBPA dataset (Table 3-3), showed that Vaxign-ML had better BPAg prediction than the epitope-based method. Epitope prediction does not necessarily correlate with the immune protection due to the host diversity, amino acid properties, epitope location, and the co-evolution between pathogen and host immune system (Halling-Brown et al. 2008; 2009). Undoubtedly, epitopes still play a role in antibody and cell-mediated immunity. The integration of BPAg prediction with epitope identification and antigen structural analysis will be investigated in the future.

3.6 Acknowledgement

I appreciate the support of the US National Institutes of Health (1R01AI081062). The result presented in this chapter is published in (E. Ong, Wang, Wong, Seetharaman, et al. 2020). Edison Ong (EO) and Yongqun He conceived and designed the study; EO, Haihe Wang, Mei U Wong, Meenakshi Seetharaman, and Ninotchka Valdez collected and curated data; EO performed bioinformatics analyses, tools development, and prepared the initial manuscript. All

authors participated in result interpretation, paper editing, discussion, and approved the paper publication.

Chapter 4 COVID-19 Coronavirus Vaccine Design Using Reverse Vaccinology and Machine Learning

4.1 Abstract

To ultimately combat the emerging COVID-19 pandemic, it is desired to develop an effective and safe vaccine against this highly contagious disease caused by the SARS-CoV-2 coronavirus. Our literature and a clinical trial survey showed that the whole virus, as well as the spike (S) protein, nucleocapsid (N) protein, and membrane (M) protein, have been tested for vaccine development against SARS and MERS. However, these vaccine candidates might lack the induction of complete protection and have safety concerns. We then applied the Vaxign and the newly developed machine learning-based Vaxign-ML reverse vaccinology tools to predict COVID-19 vaccine candidates. Our Vaxign analysis found that the N protein from the SARS-CoV-2 N protein sequence is conserved with SARS-CoV and MERS-CoV but not from the other four human coronaviruses causing mild symptoms. By investigating the entire proteome of SARS-CoV-2, six proteins, including the S protein and five non-structural proteins (nsp3, 3CL-pro, and nsp8-10), were predicted to be adhesins, which are crucial to the viral adhering and host invasion. The S, nsp3, and nsp8 proteins were also predicted by Vaxign-ML to induce high protective antigenicity. Besides the commonly used S protein, the nsp3 protein has not been tested in any coronavirus vaccine studies and was selected for further investigation. The nsp3 was found to be more conserved among SARS-CoV-2, SARS-CoV, and MERS-CoV than among 15 coronaviruses infecting humans and other animals. The protein was also predicted to contain promiscuous MHC-I and MHC-II T-cell epitopes, and the predicted linear B-cell

epitopes were found to be localized on the surface of the protein. Our predicted potential vaccine targets have the potential for effective and safe COVID-19 vaccine development. We also propose that an “Sp/Nsp cocktail vaccine” containing a structural protein(s) (Sp) and a non-structural protein(s) (Nsp) would stimulate effective complementary immune responses.

4.2 Introduction

The emerging Coronavirus Disease 2019 (COVID-19) pandemic poses a massive crisis to global public health. As of March 11, 2020, there were 118,326 confirmed cases and 4,292 deaths, according to the World Health Organization (WHO), and WHO declared the COVID-19 as a pandemic on the same day. On May 12, WHO reported 4,088,848 confirmed COVID-19 cases and 283,153 deaths globally, showing a dramatic increase in terms of case and death numbers. The causative agent of the COVID-19 disease is the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Coronaviruses can cause animal diseases such as avian infectious bronchitis caused by the infectious bronchitis virus (IBV) and pig transmissible gastroenteritis caused by a porcine coronavirus (Perlman and Netland, 2009). Bats are commonly regarded as the natural reservoir of coronaviruses, which can be transmitted to humans and other animals after genetic mutations. There are seven known human coronaviruses, including the novel SARS-CoV-2. Four of them (HCoV-HKU1, HCoV-OC43, HCoV-229E, and HCoV-NL63) have been circulating in the human population worldwide and cause mild symptoms (Cabeça, Granato, and Bellei 2013). Coronavirus became prominent after Severe acute respiratory syndrome (SARS) and Middle East Respiratory Syndrome (MERS) outbreaks. In 2003, the SARS disease caused by the SARS-associated coronavirus (SARS-CoV) infected over 8,000 people worldwide and was contained in the summer of 2003 (Lu et al. 2020). SARS-CoV-2 and SARS-CoV share high sequence identity (Lai et al. 2020). The MERS disease infected more than

2,000 people, which is caused by the MERS-associated coronavirus (MERS-CoV) and was first reported in Saudi Arabia and spread to several other countries since 2012 (Chan et al. 2015).

Great efforts have been made to develop and manufacture COVID-19 vaccines, and these efforts in pushing the vaccine clinical trials are phenomenal. Coronaviruses are positively-stranded RNA viruses with their genome packed inside the nucleocapsid (N) protein and enveloped by the membrane (M) protein, envelope (E) protein, and the spike (S) protein (F. Li 2016). While many coronavirus vaccine studies targeting different structural proteins were conducted, most of these efforts eventually ceased soon after the outbreak of SARS and MERS. With the recent COVID-19 pandemic outbreak, it is urgent to resume the coronavirus vaccine research. There were only three SARS-CoV and six MERS-CoV vaccine clinical trials, and extensive effort has been made to develop COVID-19 vaccines in response to the current pandemic. Well established vaccines targeting pathogens other than SARS-CoV-2 are also under investigation, such as measles (NCT04357028) and BCG (NCT04327206), which may induce strong immune responses and provide non-specific protective effects against SARS-CoV-2 infection (Redelman-Sidi 2020).

There are two primary design strategies for coronavirus vaccine development: the usage of the whole virus or genetically engineered vaccine antigens that can be delivered through different formats. The whole virus vaccines include inactivated (See et al. 2006) or live-attenuated vaccines (R. L. Graham et al. 2012; Fett et al. 2013). The two live attenuated SARS vaccines mutated the exoribonuclease and envelop protein to reduce the virulence and/or replication capability of the SARS-CoV. Recent works also showed promising development of three types of SARS-CoV-2 vaccines, including inactivated whole virus vaccine (Gao et al. 2020), RNA vaccine (McKay et al. 2020), and virus-like particles (VLP) vaccine (Zha et al.

2020). Overall, the whole virus vaccines can induce a strong immune response and protect against coronavirus infections. Genetically engineered vaccines that target specific coronavirus proteins are often used to improve vaccine safety and efficacy. The coronavirus antigens such as S protein, N protein, and M protein can be delivered as recombinant DNA vaccine and viral vector vaccine.

As the most superficial and protrusive protein of the coronaviruses, S protein plays a crucial role in mediating virus entry. In the SARS and MERS vaccine development, the full-length S protein and its S1 subunit (which contains receptor binding domain) have been frequently used as the vaccine antigens due to their ability to induce neutralizing antibodies that prevent host cell entry and infection. As the immediate response to the on-going pandemic, the first testing in humans of the mRNA-based vaccine targeting the S protein of SARS-CoV-2 (ClinicalTrials.gov Identifier: NCT04283461) started on March 16, 2020.

From experimentally identified immune responses induced by coronavirus vaccines, we found evidence of the protective roles of both antibody and cell-mediated immunity (Bisht et al. 2004; J. Zhao et al. 2016). The protective role of the neutralizing antibody to coronavirus S protein has been demonstrated by the experimental result that a passive transfer of the serum from mice immunized with MVA/S to naïve mice reduced the replication of challenged SARS-CoV in the respiratory tract (Bisht et al. 2004). Here the MVA/S is the highly attenuated modified vaccinia virus Ankara (MVA) containing the gene encoding full-length SARS-CoV S protein. The antibodies developed in the mice immunized with MVA/S could also bind to the S1 domain of S and neutralize SARS-CoV in vitro. Passive transfer of anti-S neutralizing antibody also offered protection against SARS-CoV (Traggiai et al. 2004). However, antibody responses in patients previously infected with respiratory viruses, including SARS-CoV and MERS-CoV, tend to be short-lived (Channappanavar, Zhao, and Perlman 2014). Instead, T cell responses are often long-lived by targeting conserved proteins and showed to have a significant correlation in protective immunity against influenza virus infection (Wilkinson et al. 2012). SARS-CoV-specific memory T cells but not antibody-producing B cells could be detected in patients six years after SARS-CoV infection (Tang et al. 2011). A further study showed that respiratory tract memory CD4+ T cells specific for an epitope of the nucleocapsid (N) protein of SARS-CoV provided protection against virulent challenge with SARS-CoV and MERS-CoV (J. Zhao et al. 2016). CD8+ T cells were also found to be crucial for the clearance of SARS-CoV and MERS-CoV infections (J. Zhao, Zhao, and Perlman 2010; Coleman et al. 2017). Therefore, our vaccine prediction would target those viral antigens with the ability to induce protective neutralizing antibody and/or T cell responses.

However, the current coronavirus vaccines, including S protein-based vaccines, might have issues in the lack of inducing complete protection and possible safety concerns (Roper and Rehm 2009; De Wit et al. 2016). Most existing SARS/MERS vaccines were reported to induce neutralizing antibodies and partial protection against the viral challenges in animal models. A recent study reported that adenovirus vaccine vector encoding full-length MERS-CoV S protein (ChAdOx1 MERS) showed protection upon MERS-CoV challenge in rhesus macaques (van Doremalen et al. 2020). Nonetheless, it is desired for a COVID-19 vaccine to induce complete protection or sterile immunity. Moreover, it has become increasingly clear that multiple immune responses, including those induced by humoral or cell-mediated immunity, are responsible for correlates of protection than antibody titers alone (Stanley A. Plotkin 2020). Both killed SARS-CoV whole virus vaccine and adenovirus-based recombinant vector vaccines expressing S or N proteins induced neutralizing antibody responses but did not provide complete protection in the animal model (See et al. 2008). A study has shown increased liver pathology in the vaccinated ferrets immunized with modified vaccinia Ankara-S recombinant vaccine (Weingartl et al. 2004). The safety and efficacy of these vaccination strategies have not been fully tested in human clinical trials, but safety could be a major concern. Therefore, novel strategies are needed to enhance the efficacy and safety of COVID-19 vaccine development.

In recent years, the development of vaccine design has been revolutionized by reverse vaccinology (RV), which aims to first identify promising vaccine candidates through bioinformatics analysis of the pathogen genome. RV has been successfully applied to vaccine discovery for pathogens such as Group B meningococcus and led to the license Bexsero vaccine (Folaranmi et al. 2015). Among current RV prediction tools (He et al. 2010; Dalsass et al. 2019), Vaxign is the first web-based RV program (He, Xiang, and Mobley 2010) and has been used to predict vaccine candidates against different bacterial and viral pathogens (Z. A. Xiang and He 2013; Singh et al. 2016; Navarro-Quiroz et al. 2018). Recently we have also developed a machine learning approach called Vaxign-ML to enhance prediction accuracy (E. Ong, Wang, Wong, Seetharaman, et al. 2020).

In this study, we first surveyed the existing coronavirus vaccine development status, and then applied the Vaxign and Vaxign-ML RV approaches to predict COVID-19 protein candidates for vaccine development. We identified six possible adhesins, including the structural S protein and five other non-structural proteins, and three of them (S, nsp3, and nsp8 proteins) were predicted to induce high protective immunity. The S protein was predicted to have the highest protective antigenicity score, and it has been extensively studied as the target of coronavirus vaccines by other researchers. The sequence conservation and immunogenicity of the multi-domain nsp3 protein, which was predicted to have the second-highest protective antigenicity score yet, was further analyzed in this study. Based on the predicted structural S protein and non-structural proteins (including nsp3) using reverse vaccinology and machine learning, we proposed and discussed a cocktail vaccine strategy for rational COVID-19 vaccine development.

4.3 Methods

4.3.1 Vaxign and Vaxign-ML Reverse Vaccinology Prediction

The SARS-CoV-2 sequence was obtained from NCBI. All the proteins of six known human coronavirus strains, including SARS-CoV, MERS-CoV, HCoV-229E, HCoV-OC43, HCoV-NL63, and HCoV-HKU1 were extracted from Uniprot proteomes (The UniProt Consortium 2008). The full proteomes of these seven coronaviruses were then analyzed using the Vaxign reverse vaccinology pipeline (He, Xiang, and Mobley 2010; E. Ong, Wang, Wong, Seetharaman, et al. 2020). The Vaxign program predicted serval biological features, including adhesin probability (Sachdeva et al. 2005), transmembrane helix (Krogh et al. 2001), orthologous proteins (L. Li, Stoeckert, and Roos 2003), protein functions (He, Xiang, and Mobley 2010), and Vaxign-ML protegenicity score (E. Ong, Wang, Wong, Seetharaman, et al. 2020).

The Vaxign-ML protegenicity score was calculated following a similar methodology described in the Vaxign-ML. In brief, the positive samples in the training data included 397 bacterial and 178 viral protective antigens (PAgs) recorded in the Protegen database (B. Yang et al. 2011) after removing homologous proteins with over 30% sequence identity. There were 4,979 negative samples extracted from the corresponding pathogens’ Uniprot proteomes (The UniProt Consortium 2008) with sequence dis-similarity to the PAgs, as described in previous studies (Bowman et al. 2011; Doytchinova and Flower 2007; Heinson et al. 2017). Homologous proteins in the negative samples were also removed. The proteins in the resulting dataset were annotated with biological and physicochemical features. The biological features included adhesin probability (Sachdeva et al. 2005), transmembrane helix (Krogh et al. 2001), and immunogenicity (Fleri et al. 2017). The physicochemical features included the compositions, transitions, and distributions (Dubchak et al. 1995), quasi-sequence-order (Chou 2000), Moreau-Broto auto-correlation(Feng and Zhang, 2000; Lin andPan, 2001), and Geary auto-correlation (Sokal and Thomson 2006) of various physicochemical properties such as charge, hydrophobicity, polarity, and solvent accessibility (S. A. K. Ong et al. 2007). Five supervised ML classification algorithms, including logistic regression, support vector machine, k-nearest neighbor, random forest (Pedregosa et al. 2012), and extreme gradient boosting (XGB) (T. Chen and Guestrin 2016) were trained on the annotated proteins dataset. The performance of these models was evaluated using a nested five-fold cross-validation (N5CV) based on the area under receiver operating characteristic curve, precision, recall, weighted F1-score, and Matthew's correlation coefficient. The best performing XGB model was selected to predict the protegenicity score of all SARS-CoV-2 isolate Wuhan-Hu-1 (GenBank ID: MN908947.3) proteins, downloaded from NCBI. The protegenicity score is the percentile rank score from the Vaxign-ML classification model. A protein with higher protegenicity score is considered as stronger vaccine candidate with higher utility toward protection. In addition, using the protegenicity score of 0.9 as a threshold resulted in the highest prediction performance with weighted F1-score = 0.94 in N5CV.

4.3.2 Phylogenetic Analysis

The protein nsp3 was selected for further investigation. The nsp3 proteins of 14 coronaviruses besides SARS-CoV-2 were downloaded from the Uniprot (Table 4-1). Multiple sequence alignment of these nsp3 proteins was performed using MUSCLE (Edgar 2004) and visualized via

SEAVIEW (Gouy, Guindon, and Gascuel 2010). The phylogenetic tree was constructed using PhyML (Lefort, Longueville, and Gascuel 2017), and the amino acid conservation was estimated by the Jensen-Shannon Divergence (JSD) (Capra and Singh, 2007). The JSD score was also used to generate a sequence conservation line using the nsp3 protein sequences from 4 or 13 coronaviruses.

4.3.3 Immunogenicity Analysis

The immunogenicity of the nsp3 protein was evaluated by the prediction of T cell MHC-I and MHC-II, and linear B cell epitopes. For T cell MHC-I epitopes, the IEDB consensus method was used to predicting promiscuous epitopes binding to 4 out of 27 MHC-I reference alleles with consensus percentile ranking less than 1.0 score (Fleri et al. 2017). For T cell MHC-II epitopes, the IEDB consensus method was used to predicting promiscuous epitopes binding to more than half of the 27 MHC-II reference alleles with a consensus percentile ranking less than 10.0. The MHC-I and MHC-II reference alleles covered a wide range of human genetic variations representing the majority of the world population (Greenbaum et al. 2011; Weiskopf et al. 2013). The linear B cell epitopes were predicted using the BepiPred 2.0 with a cutoff of 0.55 score (Jespersen et al. 2017). Linear B cell epitopes with at least ten amino acids were mapped to the predicted 3D structure of SARS-CoV-2 nsp3 protein visualized via PyMol (Schrödinger 2015). The predicted count of T cell MHC-I and MHC-II epitopes and the predicted score of linear B cell epitopes were computed as the sliding averages with a window size of ten amino acids. The nsp3 protein 3D structure was predicted using C-I-Tasser (W. Zheng et al. 2019) available in the Zhang Lab webserver (https://zhanglab.ccmb.med.umich.edu/C-I-TASSER/2019-nCov/).

Table 4-1 The full proteome and nsp3 protein IDs of 15 coronaviruses used in this study.

Proteome ID	Protein ID	Organism	Organism Taxon ID
UP000000354	P0C6X7	Human SARS coronavirus (SARS-CoV) (Severe acute respiratory syndrome coronavirus)	694009
UP000171868	T2B9U0	Middle East respiratory syndrome-related coronavirus	1335626
UP000006716	P0C6X1	Human coronavirus 229E (HCoV-229E)	11137
UP000001985	P0C6X4	Human coronavirus HKU1 (isolate N5) (HCoV-HKU1) (Strain: Isolate N5)	443241
UP000007552	P0C6X6	Human coronavirus OC43 (HCoV-OC43)	31631
UP000008573	P0C6X5	Human coronavirus NL63 (HCoV-NL63)	277944
UP000000835	Q98VG9	Feline coronavirus (strain FIPV WSU-79/1146) (FCoV) (Strain: FIPV WSU-79/1146)	33734
UP000006717	P0C6Y1	Avian infectious bronchitis virus (strain Beaudette) (IBV) (Strain: Beaudette)	11122
UP000007192	P0C6X9	Murine coronavirus (strain A59) (MHV-A59) (Murine hepatitis virus) (Strain: A59)	11142
UP000001440	P0C6Y5	Porcine transmissible gastroenteritis coronavirus (strain Purdue) (TGEV) (Strain: Purdue)	11151
UP000007451	P0C6W4	Bat coronavirus HKU5 (BtCoV) (BtCoV/HKU5/2004)	694008
UP000006576	P0C6W5	Bat coronavirus HKU9 (BtCoV) (BtCoV/HKU9)	694006
UP000006574	P0C6W3	Bat coronavirus HKU4 (BtCoV) (BtCoV/HKU4/2004) (Strain: B04f)	694007
UP000113079	P0C6W0	Bat coronavirus 512/2005 (BtCoV) (BtCoV/512/2005)	693999
UP000007450	P0C6W2	Bat coronavirus HKU3 (BtCoV) (SARS-like coronavirus HKU3)	442736

4.4 Results

4.4.1 SARS-CoV-2 N protein sequence is conserved with the N protein from SARS-CoV and MERS-CoV

We first used the Vaxign analysis framework (He, Xiang, and Mobley 2010; E. Ong, Wang, Wong, Seetharaman, et al. 2020) to compare the full proteomes of seven human coronavirus strains (SARS-CoV-2, SARS-CoV, MERS-CoV, HCoV-229E, HCoV-OC43, HCoV-NL63, and HCoV-HKU1). The proteins of SARS-CoV-2 were used as the seed for the pan-genomic comparative analysis. The Vaxign pan-genomic analysis reported only the N protein in SARS-CoV-2 having high sequence similarity among the more severe form of coronavirus (SARS-CoV and MERS-CoV) while having low sequence similarity among the more typically mild HCoV-229E, HCoV-OC43, HCoV-NL63, and HCoV-HKU1. The sequence conservation suggested the potential of N protein as a candidate for the cross-protective vaccine against SARS and MERS. The N protein was also evaluated and used for vaccine development. As a protein inside the viral envelope, the N protein packs the coronavirus RNA to form the helical nucleocapsid in virion assembly. This protein is more conserved than the S protein and was reported to induce a humoral and cellular immune response against coronavirus infections (P. Zhao et al. 2005). A conserved CD4+ T cell epitope in the SARS-CoV N was also found important for the induction of protection against the challenge of SARS-CoV or MERS-CoV (J. Zhao et al. 2016). However, a study also showed the linkage between N protein and severe pneumonia or other serious liver failures, suggesting N protein-induced pathogenesis and possible adverse effects caused by N protein-derived vaccines (Yasui et al. 2008).

4.4.2 Six adhesive proteins in SARS-CoV-2 identified as potential vaccine targets

The Vaxign RV analysis predicted six SARS-CoV-2 proteins (S protein, nsp3, 3CL-PRO, and nsp8-10) as adhesive proteins (Table 4-1). Adhesin plays a critical role in the virus adhering to the host cell and facilitating the virus entry to the host cell (Ribet and Cossart 2015), which has a significant association with the vaccine-induced protection (E. Ong, Wong, and He 2017). In SARS-CoV-2, S protein was predicted to be adhesin, matching its primary role in virus entry. The structure of SARS-CoV-2 S protein was determined (Wrapp, Wang, et al. 2020) and reported to contribute to the host cell entry by interacting with the angiotensin-converting enzyme 2 (ACE2) (Letko, Marzi, and Munster 2020). Besides S protein, the other five predicted adhesive proteins were all non-structural proteins. In particular, nsp3 is the largest non-structural protein of SARS-CoV-2 comprises various functional domains (Lei, Kusov, and Hilgenfeld 2018).

Table 4-2 Vaxign-ML Prediction and adhesin probability of all SARS-CoV-2 proteins.

Protein			Vaxign-ML Score	Adhesin Probability
orf1ab	nsp1	Host translation inhibitor	79.312	0.297
	nsp2	Non-structural protein 2	89.647	0.319
	nsp3	Non-structural protein 3	**95.283***	**0.524**[#]
	nsp4	Non-structural protein 4	89.647	0.289
	3CL-PRO	Proteinase 3CL-PRO	89.647	**0.653**[#]
	nsp6	Non-structural protein 6	89.017	0.320
	nsp7	Non-structural protein 7	89.647	0.269
	nsp8	Non-structural protein 8	**90.349***	**0.764**[#]
	nsp9	Non-structural protein 9	89.647	**0.796**[#]
	nsp10	Non-structural protein 10	89.647	**0.769**[#]
	RdRp	RNA-directed RNA polymerase	89.647	0.229
	Hel	Helicase	89.647	0.398
	ExoN	Guanine-N7 methyltransferase	89.629	0.183
	NendoU	Uridylate-specific endoribonuclease	89.647	0.254
	2'-O-MT	2'-O-methyltransferase	89.647	0.421
	S	Surface glycoprotein	**97.623***	**0.635**[#]
	ORF3a	ORF3a	66.925	0.383
	E	envelope protein	23.839	0.234
	M	membrane glycoprotein	84.102	0.282
	ORF6	ORF6	33.165	0.095
	ORF7	ORF7a	11.199	0.451
	ORF8	ORF8	31.023	0.311
	N	nucleocapsid phosphoprotein	89.647	0.373
	ORF10	ORF10	6.266	0.0

* denotes Vaxign-ML predicted vaccine candidate.

[#] denotes predicted adhesin.

4.4.3 Three adhesin proteins were predicted to induce strong protective immunity

The recently published Vaxign-ML pipeline was applied to compute the protegenicity (protective antigenicity) score and predict the induction of protective immunity by a vaccine candidate (E. Ong, Wang, Wong, Seetharaman, et al. 2020). Vaxign-ML predicts the protegenicity score using an optimized supervised machine learning model with manually annotated training data consisted of bacterial and viral protective antigens. These protective antigens were tested to be protective in at least one animal challenge model (B. Yang et al. 2011). The performance of the Vaxign-ML models was evaluated (Table 4-2 and Figure 4-1), and the best performing model had a weighted F1-score and Matthew's correlation coefficient of 0.94 and 0.66, respectively, in nested cross-validation. Using the optimized Vaxign-ML model, we predicted three proteins (S protein, nsp3, and nsp8) as vaccine candidates with significant protegenicity scores (Table 4-1). The S protein was predicted to have the highest protegenicity score, which is consistent with the experimental observations reported in the literature. The nsp3 protein is the second most promising vaccine candidate besides S protein. There was currently no study of nsp3 as a vaccine target. The structure and functions of this protein have various roles in coronavirus infection, including replication and pathogenesis (immune evasion and virus survival) (Lei, Kusov, and Hilgenfeld 2018). Therefore, we selected nsp3 for further investigation, as described below.

Table 4-3 Nested five-fold cross-validation evaluation metrics of five machine learning algorithms.

Models	Precision	Recall	Weighted F1	MCC
Logistic Regression	0.541	0.366	0.886 (±0.02)	0.370
Support Vector Machine	0.902	0.483	0.932 (±0.01)	0.633
K Nearest Neighbor	0.489	0.552	0.895 (±0.006)	0.458
Random Forest	0.949	0.403	0.923 (±0.008)	0.593
Extreme Gradient Boosting	0.801	0.600	0.939 (±0.008)	0.663

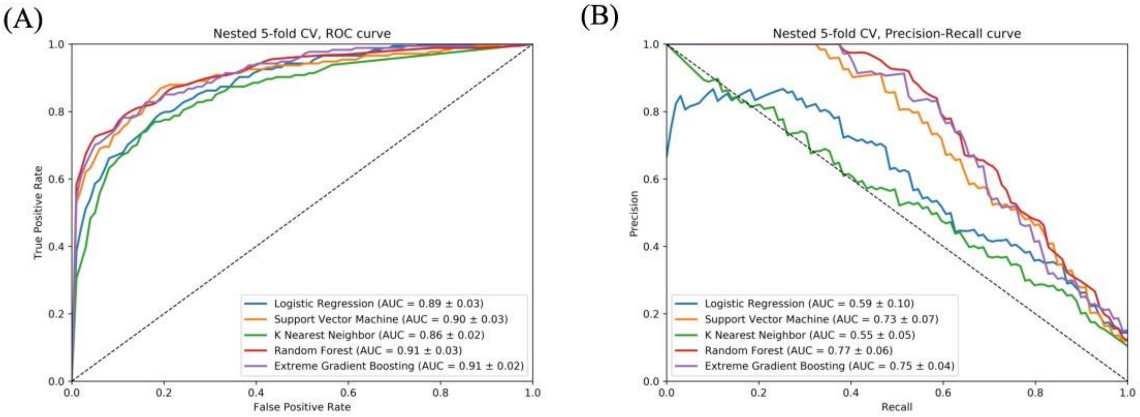


Figure 4-1 The average ROC and precision-recall curves of five machine learning algorithms in nested five-fold cross-validation.

Vaxign-ML virus model (A) Receiver operating characteristic (ROC) curve and (B) precision-recall curve of the nested five-fold cross-validation. The average ROC curves in nested five-fold cross-validation of five machine learning algorithms (logistic regression, support vector machine, k nearest neighbor, random forest, and extreme gradient boosting). While both logistic regression (LR) and k nearest neighbor (KNN) methods had relatively good AUROC curves, KNN and LR had low precision and recall, as indicated in Table 4-3. In particular, the behavior near the upper left corner by the fact that although recall in precision-recall curve can only decrease monotonically, the precision, as the ratio between true positvies and predicted positives (true positivies + false positives), does not necessarily decrease monotonically when increasing the threshold. and LR had an “n” shape curve near zero recall.

4.4.4 Nsp3 as a vaccine candidate

The multiple sequence alignment and the resulting phylogeny of nsp3 protein showed that this protein in SARS-CoV-2 was more closely related to the human coronaviruses SARS-CoV and MERS-CoV, and bat coronaviruses BtCoV/HKU3, BtCoV/HKU4, and BtCoV/HKU9. We studied the genetic conservation of nsp3 protein (Figure 4-2 A) in seven human coronaviruses and eight coronaviruses infecting other animals (Table 4-1). The five human coronaviruses, SARS-CoV-2, SARS-CoV, MERS-CoV, HCoV-HKU1, and HCoV-OC43, belong to the beta-coronavirus while HCoV-229E and HCoV-NL63 belong to the alpha-coronavirus. The HCoV-HKU1 and HCoV-OC43, as the human coronavirus with mild symptoms clustered together with murine MHV-A59. The more severe form of human coronavirus SARS-CoV-2, SARS-CoV, and MERS-CoV grouped with three bat coronaviruses BtCoV/HKU3, BtCoV/HKU4, and BtCoV/HKU9.

When evaluating the amino acid conservations relative to the functional domains in nsp3, all protein domains, except the hypervariable region (HVR), macro-domain 1 (MAC1), and beta-coronavirus-specific marker βSM, showed higher conservation in SARS-CoV-2, SARS-CoV, and MERS-CoV (Figure 4-2 B). The amino acid conservation between the major human coronavirus (SARS-CoV-2, SARS-CoV, and MERS-CoV) was plotted and compared to all 15 coronaviruses used to generate the phylogenetic of nsp3 protein along with the corresponding SARS-CoV domains (Figure 4-2 B).

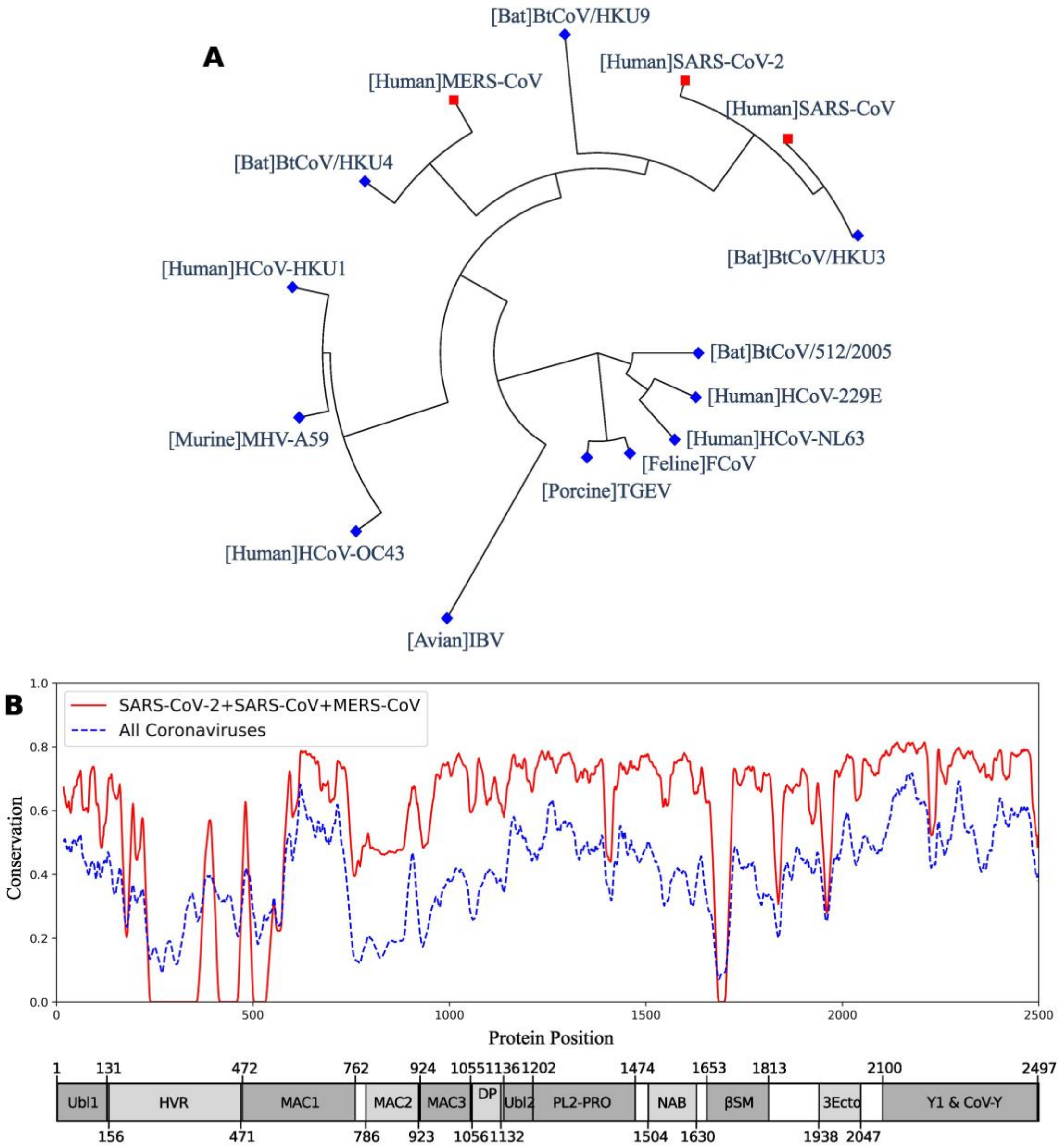


Figure 4-2 The phylogeny and sequence conservation of coronavirus nsp3.

(A) Phylogeny of 15 strains based on the nsp3 protein sequence alignment and phylogeny analysis. (B) The conservation of nsp3 among different coronavirus strains. The red line represents the conservation among the four strains (SARS-CoV, SARS-CoV-2, MERS, and BtCoV-HKU3). The blue line was generated using all the 15 strains. The bottom part represents the nsp3 peptides and their sizes. The phylogenetically close four strains have more conserved nsp3 sequences than all the strains being considered.

The immunogenicity of nsp3 protein in terms of T cell MHC-I & MHC-II and linear B cell epitopes was also investigated. There were 28 (Table 4-4) and 42 (Table 4-5) promiscuous epitopes predicted to bind the reference MHC-I & MHC-II alleles, which covered the majority of the world population, respectively. In terms of linear B cell epitopes, there were 14 epitopes with BepiPred scores over 0.55 and had at least ten amino acids in length (Table 4-6). The 3D structure of SARS-CoV-2 protein was plotted and highlighted with the T cell MHC-I & MHC-II and linear B cell epitopes (Figure 4-3). The predicted B cell epitopes were more likely located on the surface of the nsp3 protein. Most of the predicted MHC-I & MHC-II epitopes were embedded inside the protein. The sliding averages of T cell MHC-I & MHC-II and linear B cell epitopes were plotted with respect to the tentative SARS-CoV-2 nsp3 protein domains using SARS-CoV nsp3 protein as a reference (Figure 4-4). The ubiquitin-like domain 1 and 2 (Ubl1 and Ubl2) are only predicted to have MHC-I epitopes. The Domain Preceding Ubl2 and PL2-PRO (DPUP) domain had only predicted MHC-II epitopes. The PL2-PRO contained both predicted MHC-I and MHC-II epitopes, but not B cell epitopes. In particular, the TM1, TM2, and AH1 were predicted helical regions with high T cell MHC-I and MHC-II epitopes(Rothbard and Taylor 1988). The TM1 and TM2 are transmembrane regions passing the endoplasmic reticulum (ER) membrane. The HVR, MAC2, MAC3, nucleic-acid binding domain (NAB), βSM, Nsp3 ectodomain; (3Ecto), Y1, and CoV-Y domain contained predicted B cell epitopes. Finally, the Vaxign RV framework also predicted two regions (position 251-260 and 329-337) in the MAC1 domain of the nsp3 having high sequence similarity to the human mono-ADP-ribosyltransferase PARP14 (NP_060024.2).

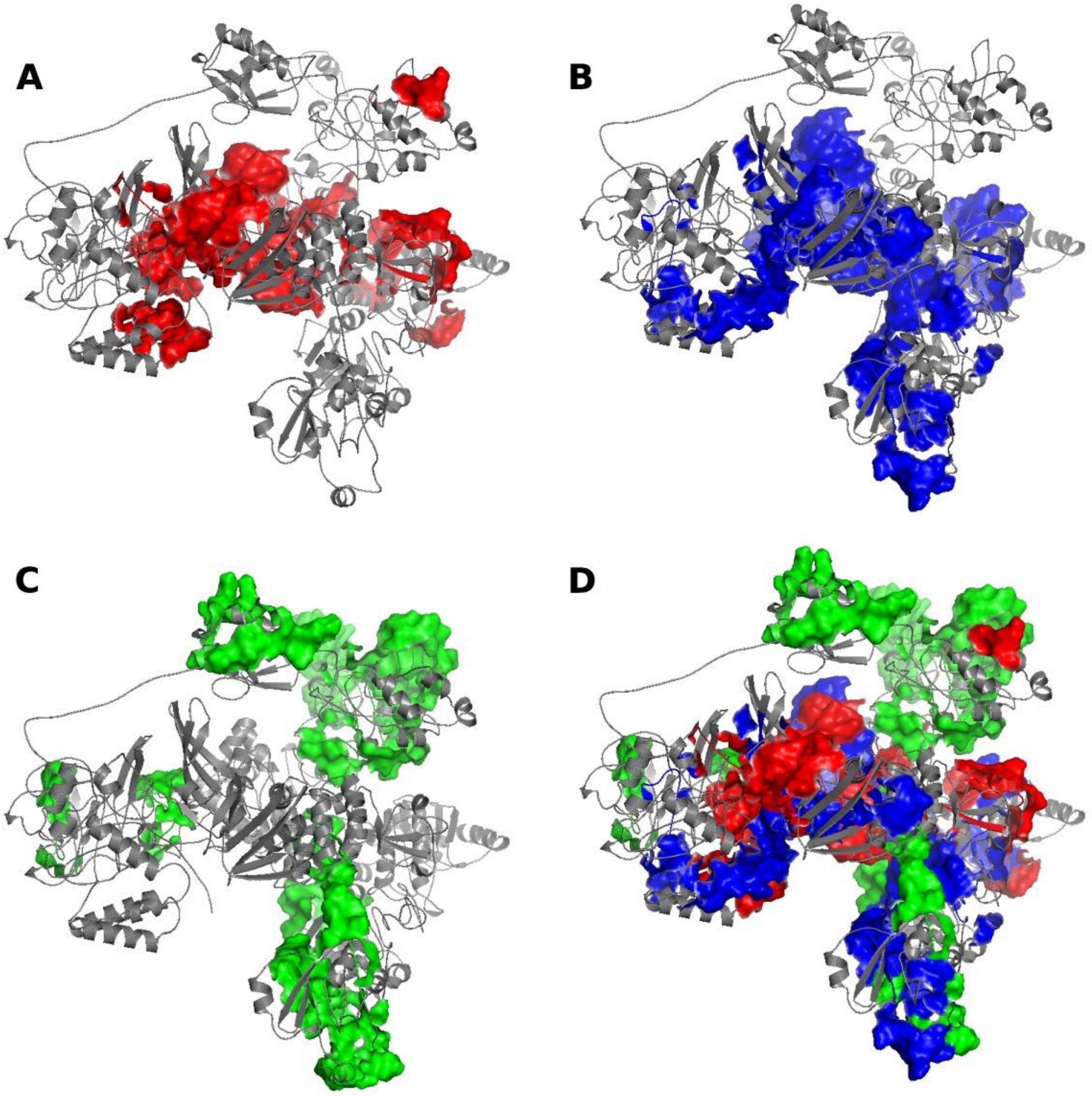


Figure 4-3 Predicted 3D structure of nsp3 protein with highlighted epitopes.

Predicted 3D structure of nsp3 protein highlighted with (A) MHC-I T cell epitopes (red), (B) MHC-II (blue) T cell epitopes, (C) linear B cell epitopes (green), and the merged epitopes. The B cell epitopes are more exposed on the protein surface, while the T cell MHC-I and MHC-II epitopes are more located within the protein.

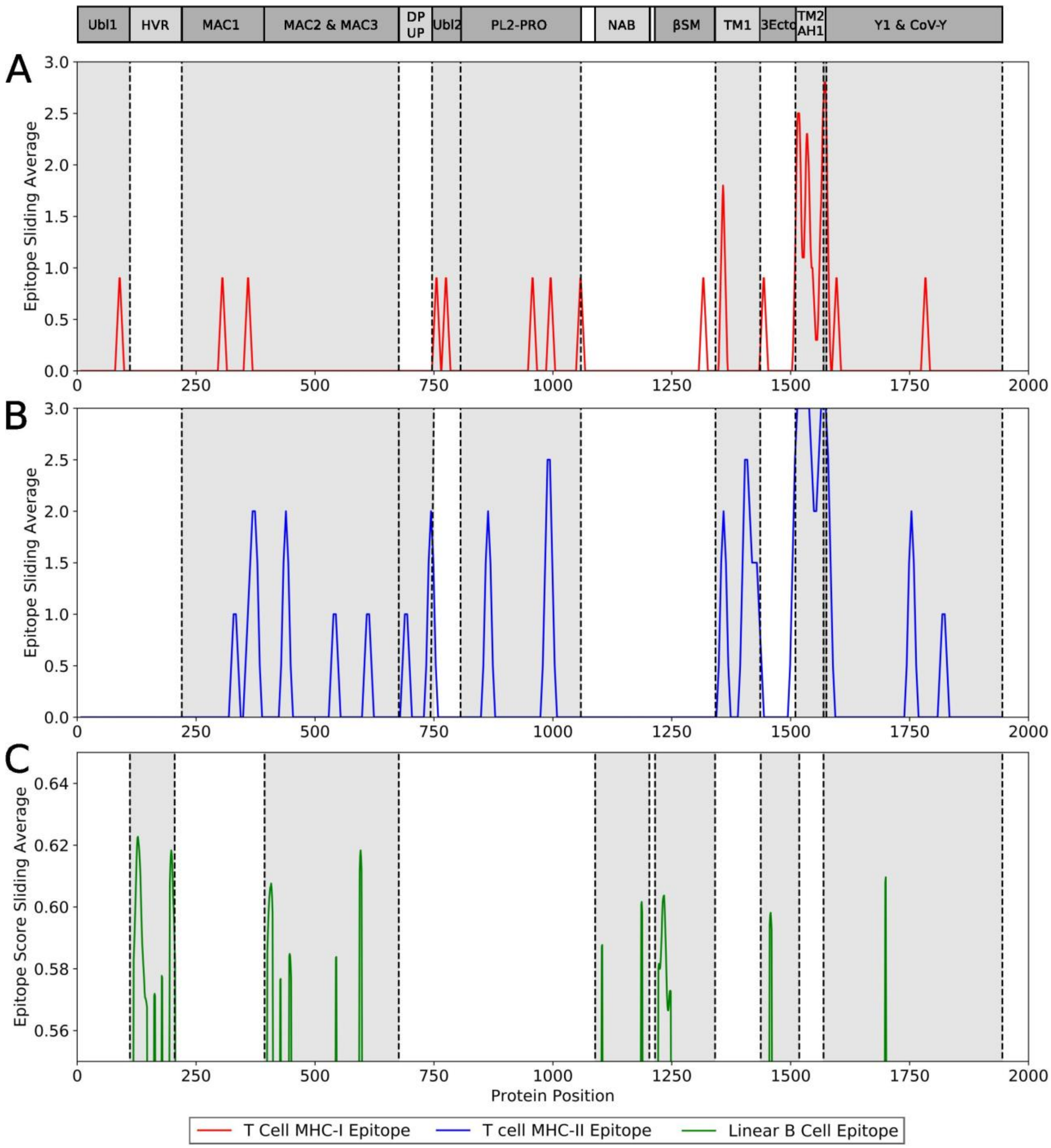


Figure 4-4 Immunogenic region of nsp3 between SARS-CoV-2 and the four conservation strains.

(A) MHC-I (red) T cell epitope (B) MHC-II (blue) T cell epitope (C) linear B cell epitope (green).

Table 4-4 Predicted promiscuous T cell MHC-I epitopes of SARS-CoV-2 nsp3.

Epitope	Start	End	Allele
STNVTIATY	1455	1465	HLA-A*26:01,HLA-B*15:01,HLA-A*30:02,HLA-A*01:01
RMYIFFASF	1564	1574	HLA-A*23:01,HLA-A*24:02,HLA-A*32:01,HLA-B*08:01,HLA-B*15:01
AEWFLAYIL	1507	1517	HLA-B*44:02,HLA-A*32:01,HLA-B*44:03,HLA-B*40:01
MSNLGMPSY	1436	1446	HLA-A*30:02,HLA-B*35:01,HLA-A*01:01,HLA-B*58:01,HLA-B*15:01
LVAEWFLAY	1505	1515	HLA-B*35:01,HLA-A*26:01,HLA-B*15:01,HLA-A*01:01
ILFTRFFYV	1514	1524	HLA-A*02:01,HLA-A*02:06,HLA-A*02:03,HLA-B*08:01
MMSAPPAQY	988	998	HLA-B*15:01,HLA-A*03:01,HLA-A*30:02,HLA-B*35:01
VMYMGTLSY	950	960	HLA-A*11:01,HLA-A*03:01,HLA-A*30:02,HLA-A*01:01,HLA-A*32:01,HLA-B*15:01
KENSYTTTI	1051	1061	HLA-A*32:01,HLA-B*44:03,HLA-B*44:02,HLA-B*40:01
WSMATYYLF	82	92	HLA-A*23:01,HLA-A*24:02,HLA-B*53:01,HLA-B*58:01,HLA-A*32:01,HLA-B*57:01,HLA-B*15:01
AIMQLFFSY	1527	1537	HLA-A*11:01,HLA-A*26:01,HLA-A*30:02,HLA-A*32:01,HLA-B*15:01,HLA-B*44:03
FFASFYYVW	1568	1578	HLA-A*23:01,HLA-A*24:02,HLA-B*53:01,HLA-B*58:01
LAAVNSVPW	1309	1319	HLA-B*35:01,HLA-B*57:01,HLA-B*58:01,HLA-B*53:01
MPYFFTLLL	1351	1361	HLA-B*35:01,HLA-B*53:01,HLA-B*51:01,HLA-B*08:01,HLA-B*07:02
LAAIMQLFF	1525	1535	HLA-B*51:01,HLA-B*35:01,HLA-B*53:01,HLA-B*58:01
STCMMCYKR	1589	1599	HLA-A*11:01,HLA-A*33:01,HLA-A*31:01,HLA-A*68:01
YIFFASFYY	1566	1576	HLA-A*11:01,HLA-A*26:01,HLA-A*03:01,HLA-A*30:02,HLA-B*35:01,HLA-A*01:01,HLA-B*15:01,HLA-A*68:01
QMAPISAMV	1555	1565	HLA-A*68:02,HLA-A*02:01,HLA-A*02:06,HLA-A*02:03
SAMVRMYIF	1560	1570	HLA-A*32:01,HLA-B*57:01,HLA-B*35:01,HLA-B*08:01
RTNVYLAVF	352	362	HLA-B*57:01,HLA-B*15:01,HLA-A*32:01,HLA-B*58:01
MSMTYGQQF	768	778	HLA-B*35:01,HLA-B*53:01,HLA-B*58:01,HLA-B*57:01,HLA-B*15:01
RTIKVFTTV	748	758	HLA-A*68:02,HLA-A*02:06,HLA-A*32:01,HLA-B*58:01
YMPYFFTLL	1350	1360	HLA-A*24:02,HLA-A*02:01,HLA-A*02:06,HLA-A*02:03
LAYILFTRF	1511	1521	HLA-B*35:01,HLA-B*53:01,HLA-B*51:01,HLA-B*58:01,HLA-B*15:01
QLFFSYFAV	1530	1540	HLA-A*68:02,HLA-A*02:01,HLA-A*02:06,HLA-A*02:03
YVNTFSSTF	1776	1786	HLA-A*32:01,HLA-A*26:01,HLA-B*15:01,HLA-B*35:01
HFISNSWLM	1539	1549	HLA-A*23:01,HLA-A*26:01,HLA-B*35:01,HLA-A*24:02
HVVGPNVNK	298	308	HLA-A*11:01,HLA-A*03:01,HLA-A*30:01,HLA-A*68:01

Table 4-5 Predicted promiscuous T cell MHC-I epitopes of SARS-CoV-2 nsp3.

Epitope	Start	End	Allele
ISNSWLMWLIINLVQ	1541	1557	HLA-DPA1*03:01,HLA-DRB1*03:01,HLA-DQA1*01:02,HLA-DRB1*04:01,HLA-DQB1*06:02,HLA-DRB1*08:02,HLA-DRB1*15:01,HLA-DQB1*05:01,HLA-DPB1*04:01,HLA-DRB1*11:01,HLA-DPA1*01:03,HLA-DQA1*04:01,HLA-DPB1*04:02,HLA-DRB1*12:01,HLA-DQB1*04:02,HLA-DRB1*04:05,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPA1*01,HLA-DRB5*01:01,HLA-DPB1*02:01,HLA-DRB1*01:01,HLA-DQA1*01:01
LAYILFTRFFYVLGL	1511	1527	HLA-DPA1*03:01,HLA-DRB1*03:01,HLA-DRB1*08:02,HLA-DRB1*15:01,HLA-DQB1*05:01,HLA-DPB1*04:01,HLA-DRB1*11:01,HLA-DPA1*01:03,HLA-DPB1*14:01,HLA-DPB1*04:02,HLA-DPB1*05:01,HLA-DRB1*12:01,HLA-DRB1*07:01,HLA-DRB3*01:01,HLA-DRB1*04:05,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPA1*01,HLA-DRB5*01:01,HLA-DPB1*02:01,HLA-DRB4*01:01,HLA-DQA1*01:01
AAIMQLFFSYFAVHF	1526	1542	HLA-DPA1*03:01,HLA-DRB1*04:01,HLA-DRB1*08:02,HLA-DRB1*15:01,HLA-DQB1*05:01,HLA-DPB1*04:01,HLA-DRB1*11:01,HLA-DPA1*01:03,HLA-DPB1*14:01,HLA-DPB1*04:02,HLA-DRB1*12:01,HLA-DPB1*05:01,HLA-DRB1*07:01,HLA-DRB3*01:01,HLA-DRB1*04:05,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPA1*01,HLA-DRB5*01:01,HLA-DPB1*02:01,HLA-DRB4*01:01,HLA-DQA1*01:01
AMVRMYIFFASFYYV	1561	1577	HLA-DPA1*03:01,HLA-DRB1*04:01,HLA-DRB1*08:02,HLA-DRB1*15:01,HLA-DQB1*05:01,HLA-DPB1*04:01,HLA-DRB1*11:01,HLA-DPA1*01:03,HLA-DPB1*14:01,HLA-DQA1*05:01,HLA-DPB1*04:02,HLA-DRB1*12:01,HLA-DPB1*05:01,HLA-DRB1*09:01,HLA-DRB3*01:01,HLA-DRB1*04:05,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPA1*01,HLA-DQB1*02:01,HLA-DRB5*01:01,HLA-DPB1*02:01,HLA-DQA1*01:01
YIFFASFYYVWKSYV	1566	1582	HLA-DPB1*04:02,HLA-DRB1*12:01,HLA-DPB1*05:01,HLA-DPA1*03:01,HLA-DRB1*15:01,HLA-DRB1*07:01,HLA-DQB1*05:01,HLA-DRB5*01:01,HLA-DPB1*02:01,HLA-DPB1*04:01,HLA-DPB1*14:01,HLA-DRB3*01:01,HLA-DRB1*11:01,HLA-DPA1*01:03,HLA-DRB1*08:02,HLA-DRB1*04:05,HLA-DQA1*01:01,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPA1*01
EETKFLTENLLLYID	426	442	HLA-DQA1*05:01,HLA-DPB1*04:02,HLA-DRB1*13:02,HLA-DQB1*02:01,HLA-DPB1*05:01,HLA-DPA1*03:01,HLA-DRB1*12:01,HLA-DRB1*03:01,HLA-DQB1*03:02,HLA-DPB1*02:01,HLA-DPB1*04:01,HLA-DPB1*14:01,HLA-DRB3*01:01,HLA-DRB1*11:01,HLA-DPA1*01:03,HLA-DRB1*04:01,HLA-DQA1*03:01,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPA1*01
MPYFFTLLLQLCTFT	1351	1367	HLA-DPA1*03:01,HLA-DRB1*04:01,HLA-DRB1*08:02,HLA-DRB1*15:01,HLA-DQB1*05:01,HLA-DPB1*04:01,HLA-DRB1*11:01,HLA-DPA1*01:03,HLA-DPB1*14:01,HLA-DQA1*05:01,HLA-DPB1*04:02,HLA-DRB1*12:01,HLA-DPB1*05:01,HLA-DRB1*04:05,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPA1*01,HLA-DQB1*02:01,HLA-DRB5*01:01,HLA-DPB1*02:01,HLA-DQA1*01:01
IIIWFLLLSVCLGSL	1411	1427	HLA-DPA1*03:01,HLA-DRB1*03:01,HLA-DRB1*04:01,HLA-DRB1*08:02,HLA-DRB1*15:01,HLA-DQB1*05:01,HLA-DPB1*04:01,HLA-DRB1*11:01,HLA-DPB1*14:01,HLA-DQA1*05:01,HLA-DPB1*04:02,HLA-DRB1*12:01,HLA-DRB1*04:05,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPA1*01,HLA-DQB1*02:01,HLA-DRB5*01:01,HLA-DRB4*01:01,HLA-DRB1*01:01,HLA-DQA1*01:01
VAEWFLAYILFTRFF	1506	1522	HLA-DQB1*03:02,HLA-DPA1*03:01,HLA-DRB1*04:01,HLA-DRB1*08:02,HLA-DRB1*15:01,HLA-DQB1*05:01,HLA-DPB1*04:01,HLA-DRB1*11:01,HLA-DPA1*01:03,HLA-DPB1*14:01,HLA-DQA1*04:01,HLA-DQA1*05:01,HLA-DPB1*04:02,HLA-DPB1*05:01,HLA-DRB1*12:01,HLA-DRB1*07:01,HLA-DQB1*04:02,HLA-DRB1*04:05,HLA-DQA1*03:01,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPA1*01,HLA-DQB1*02:01,HLA-DRB5*01:01,HLA-DPB1*02:01,HLA-DRB4*01:01,HLA-DQA1*01:01
CKSAFYILPSIISNE	531	547	HLA-DRB1*12:01,HLA-DRB1*15:01,HLA-DRB1*03:01,HLA-DRB1*07:01,HLA-DQB1*05:01,HLA-DRB1*09:01,HLA-DRB5*01:01,HLA-DPB1*02:01,HLA-DRB1*01:01,HLA-DRB1*11:01,HLA-DRB1*04:01,HLA-DRB1*04:05,HLA-DQA1*01:01,HLA-DPA1*02:01,HLA-DPA1*01:03,HLA-DPB1*14:01,HLA-DRB1*08:02

QQESPFVMMSAPPAQ	981	997	HLA-DRB3*02:02,HLA-DRB1*13:02,HLA-DRB1*03:01,HLA-DRB1*15:01,HLA-DRB1*09:01,HLA-DRB5*01:01,HLA-DQB1*04:02,HLA-DRB3*01:01,HLA-DQA1*01:02,HLA-DRB1*01:01,HLA-DRB1*04:01,HLA-DRB1*11:01,HLA-DRB1*04:05,HLA-DQB1*06:02,HLA-DRB1*08:02,HLA-DQA1*04:01
LFFSYFAVHFISNSW	1531	1547	HLA-DPB1*04:02,HLA-DPB1*05:01,HLA-DPA1*03:01,HLA-DRB1*15:01,HLA-DRB1*07:01,HLA-DQB1*05:01,HLA-DRB5*01:01,HLA-DPB1*04:01,HLA-DPB1*02:01,HLA-DPB1*14:01,HLA-DRB3*01:01,HLA-DRB1*11:01,HLA-DPA1*01:03,HLA-DRB1*04:01,HLA-DRB1*04:05,HLA-DQA1*01:01,HLA-DPA1*02:01,HLA-DRB1*08:02,HLA-DPB1*01:01,HLA-DPA1*01
FVMMSAPPAQYELKH	986	1002	HLA-DRB3*02:02,HLA-DRB1*13:02,HLA-DRB1*12:01,HLA-DRB1*03:01,HLA-DRB1*15:01,HLA-DRB1*09:01,HLA-DRB5*01:01,HLA-DQB1*04:02,HLA-DRB3*01:01,HLA-DRB1*01:01,HLA-DRB1*04:01,HLA-DRB1*11:01,HLA-DRB1*04:05,HLA-DRB1*08:02,HLA-DQA1*04:01
LMWLIINLVQMAPIS	1546	1562	HLA-DRB1*13:02,HLA-DQB1*03:02,HLA-DPA1*03:01,HLA-DRB1*03:01,HLA-DQA1*01:02,HLA-DRB1*04:01,HLA-DQB1*06:02,HLA-DRB1*08:02,HLA-DRB1*15:01,HLA-DQB1*05:01,HLA-DRB1*11:01,HLA-DPB1*14:01,HLA-DQA1*05:01,HLA-DPB1*04:02,HLA-DRB1*12:01,HLA-DRB1*07:01,HLA-DRB1*04:05,HLA-DQA1*03:01,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DQB1*02:01,HLA-DRB5*01:01,HLA-DRB4*01:01,HLA-DRB1*01:01,HLA-DQA1*01:01
LMCQPILLLDQALVS	1746	1762	HLA-DQA1*05:01,HLA-DPB1*04:02,HLA-DRB1*13:02,HLA-DQB1*02:01,HLA-DRB1*12:01,HLA-DPA1*03:01,HLA-DPA1*01,HLA-DRB1*03:01,HLA-DRB1*15:01,HLA-DPB1*04:01,HLA-DRB3*01:01,HLA-DRB4*01:01,HLA-DRB1*11:01,HLA-DRB1*04:01,HLA-DRB1*01:01,HLA-DRB1*04:05,HLA-DPA1*02:01,HLA-DPB1*14:01,HLA-DRB1*08:02
TAFGLVAEWFLAYIL	1501	1517	HLA-DQB1*03:02,HLA-DPA1*03:01,HLA-DRB1*15:01,HLA-DQB1*05:01,HLA-DPB1*04:01,HLA-DPA1*01:03,HLA-DPB1*14:01,HLA-DQA1*04:01,HLA-DQA1*05:01,HLA-DPB1*04:02,HLA-DPB1*05:01,HLA-DRB1*07:01,HLA-DQB1*04:02,HLA-DQA1*03:01,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPA1*01,HLA-DQB1*02:01,HLA-DRB5*01:01,HLA-DPB1*02:01,HLA-DQA1*01:01
FAVHFISNSWLMWLI	1536	1552	HLA-DRB3*02:02,HLA-DPA1*03:01,HLA-DRB1*03:01,HLA-DQA1*01:02,HLA-DRB1*04:01,HLA-DQB1*06:02,HLA-DRB1*08:02,HLA-DRB1*15:01,HLA-DQB1*05:01,HLA-DPB1*04:01,HLA-DRB1*11:01,HLA-DPA1*01:03,HLA-DPB1*14:01,HLA-DPB1*04:02,HLA-DRB1*07:01,HLA-DRB3*01:01,HLA-DRB1*04:05,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPA1*01,HLA-DRB5*01:01,HLA-DPB1*02:01,HLA-DQA1*01:01
SFNYLKSPNFSKLIN	1396	1412	HLA-DRB3*02:02,HLA-DPB1*05:01,HLA-DRB1*03:01,HLA-DRB1*07:01,HLA-DRB1*15:01,HLA-DRB1*09:01,HLA-DRB5*01:01,HLA-DPB1*04:01,HLA-DRB3*01:01,HLA-DRB1*11:01,HLA-DRB1*04:01,HLA-DRB1*08:02,HLA-DRB1*04:05,HLA-DRB1*01:01,HLA-DPA1*02:01,HLA-DPB1*14:01,HLA-DPA1*01
CYLATALLTLQQIEL	856	872	HLA-DQB1*03:02,HLA-DPA1*03:01,HLA-DQA1*01:02,HLA-DRB1*04:01,HLA-DQB1*06:02,HLA-DPB1*04:01,HLA-DRB1*11:01,HLA-DPA1*01:03,HLA-DPB1*14:01,HLA-DQA1*04:01,HLA-DQA1*05:01,HLA-DPB1*04:02,HLA-DPB1*05:01,HLA-DQB1*04:02,HLA-DQA1*03:01,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPA1*01,HLA-DQB1*02:01,HLA-DPB1*02:01,HLA-DRB4*01:01
VCTNYMPYFFTLLLQ	1346	1362	HLA-DPB1*04:02,HLA-DPB1*05:01,HLA-DPA1*03:01,HLA-DRB1*15:01,HLA-DQB1*05:01,HLA-DPB1*04:01,HLA-DPB1*02:01,HLA-DPB1*14:01,HLA-DRB3*01:01,HLA-DPA1*01:03,HLA-DRB1*04:01,HLA-DRB1*04:05,HLA-DQA1*01:01,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPA1*01
DKLVSSFLEMKSEKQ	366	382	HLA-DPB1*04:02,HLA-DPA1*03:01,HLA-DRB1*03:01,HLA-DRB1*15:01,HLA-DQB1*05:01,HLA-DRB5*01:01,HLA-DPB1*04:01,HLA-DPB1*02:01,HLA-DPB1*14:01,HLA-DRB1*11:01,HLA-DRB1*04:01,HLA-DPA1*01:03,HLA-DRB1*04:05,HLA-DQA1*01:01,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPA1*01
DKNLYDKLVSSFLEM	361	377	HLA-DRB3*02:02,HLA-DPB1*04:02,HLA-DPB1*05:01,HLA-DPA1*03:01,HLA-DRB1*15:01,HLA-DRB1*07:01,HLA-DRB1*09:01,HLA-DQB1*04:02,HLA-DPB1*04:01,HLA-DPB1*02:01,HLA-DRB1*01:01,HLA-DPA1*01:03,HLA-DRB1*08:02,HLA-DRB1*04:05,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPB1*14:01,HLA-DPA1*01,HLA-DQA1*04:01

VYYSQLMCQPILLLD	1741	1757	HLA-DPB1*04:02,HLA-DPA1*03:01,HLA-DRB1*03:01,HLA-DQB1*05:01,HLA-DPB1*04:01,HLA-DPB1*02:01,HLA-DRB4*01:01,HLA-DRB1*11:01,HLA-DPA1*01:03,HLA-DRB1*04:01,HLA-DRB1*04:05,HLA-DQA1*01:01,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPB1*14:01,HLA-DPA1*01,HLA-DRB1*01:01
GARFYFYTSKTTVAS	601	617	HLA-DRB3*02:02,HLA-DPB1*04:02,HLA-DPB1*05:01,HLA-DPA1*03:01,HLA-DPA1*01,HLA-DRB1*15:01,HLA-DRB1*07:01,HLA-DRB1*09:01,HLA-DRB5*01:01,HLA-DPB1*02:01,HLA-DPB1*04:01,HLA-DPB1*14:01,HLA-DRB1*11:01,HLA-DRB1*04:01,HLA-DPA1*01:03,HLA-DRB1*04:05,HLA-DRB1*01:01,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DRB1*08:02
FTRFFYVLGLAAIMQ	1516	1532	HLA-DPA1*03:01,HLA-DRB1*03:01,HLA-DRB1*04:01,HLA-DRB1*08:02,HLA-DRB1*15:01,HLA-DQB1*05:01,HLA-DPB1*04:01,HLA-DRB1*11:01,HLA-DPA1*01:03,HLA-DPB1*14:01,HLA-DQA1*04:01,HLA-DQA1*05:01,HLA-DPB1*04:02,HLA-DRB1*12:01,HLA-DPB1*05:01,HLA-DRB1*07:01,HLA-DRB1*09:01,HLA-DQB1*04:02,HLA-DQB1*03:01,HLA-DRB1*04:05,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPA1*01,HLA-DQB1*02:01,HLA-DRB5*01:01,HLA-DPB1*02:01,HLA-DRB1*01:01,HLA-DQA1*01:01
EVITFDNLKTLLSLR	731	747	HLA-DRB3*02:02,HLA-DPB1*04:02,HLA-DPB1*05:01,HLA-DPA1*03:01,HLA-DRB1*03:01,HLA-DRB3*01:01,HLA-DRB1*11:01,HLA-DRB1*04:01,HLA-DRB1*01:01,HLA-DRB1*04:05,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPB1*14:01,HLA-DRB1*08:02
FCLEASFNYLKSPNF	1391	1407	HLA-DRB3*02:02,HLA-DPB1*05:01,HLA-DPA1*01,HLA-DRB1*03:01,HLA-DRB1*07:01,HLA-DRB1*09:01,HLA-DRB5*01:01,HLA-DPB1*02:01,HLA-DPB1*04:01,HLA-DRB1*11:01,HLA-DRB1*04:01,HLA-DPA1*01:03,HLA-DRB1*04:05,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DRB1*08:02
MAPISAMVRMYIFFA	1556	1572	HLA-DPB1*05:01,HLA-DRB1*15:01,HLA-DRB1*03:01,HLA-DRB1*04:05,HLA-DRB1*09:01,HLA-DRB5*01:01,HLA-DPB1*02:01,HLA-DRB3*01:01,HLA-DQA1*01:02,HLA-DRB4*01:01,HLA-DRB1*11:01,HLA-DPA1*01:03,HLA-DRB1*04:01,HLA-DQB1*06:02,HLA-DPA1*02:01,HLA-DRB1*08:02
CLGSLIYSTAALGVL	1421	1437	HLA-DQA1*05:01,HLA-DRB3*02:02,HLA-DRB1*13:02,HLA-DQB1*02:01,HLA-DRB1*12:01,HLA-DRB1*03:01,HLA-DRB1*07:01,HLA-DRB1*15:01,HLA-DRB1*09:01,HLA-DRB5*01:01,HLA-DPB1*02:01,HLA-DQB1*03:01,HLA-DPB1*14:01,HLA-DRB1*01:01,HLA-DRB1*04:01,HLA-DRB1*11:01,HLA-DRB1*04:05,HLA-DPA1*01:03,HLA-DPA1*02:01,HLA-DPB1*01:01
DNLKTLLSLREVRTI	736	752	HLA-DPB1*04:02,HLA-DRB1*12:01,HLA-DPA1*03:01,HLA-DRB1*03:01,HLA-DRB1*15:01,HLA-DRB4*01:01,HLA-DRB1*11:01,HLA-DRB1*04:01,HLA-DRB1*01:01,HLA-DRB1*04:05,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPB1*14:01,HLA-DRB1*08:02
INLVQMAPISAMVRM	1551	1567	HLA-DRB3*02:02,HLA-DQA1*05:01,HLA-DRB1*13:02,HLA-DRB1*12:01,HLA-DRB1*03:01,HLA-DRB1*15:01,HLA-DRB1*07:01,HLA-DRB1*09:01,HLA-DRB5*01:01,HLA-DQB1*04:02,HLA-DQB1*03:01,HLA-DRB4*01:01,HLA-DQA1*01:02,HLA-DRB1*01:01,HLA-DRB1*11:01,HLA-DRB1*04:01,HLA-DRB1*04:05,HLA-DQB1*06:02,HLA-DRB1*08:02,HLA-DQA1*04:01
FKWDLTAFGLVAEWF	1496	1512	HLA-DQA1*05:01,HLA-DQB1*02:01,HLA-DQB1*03:02,HLA-DPB1*05:01,HLA-DRB1*03:01,HLA-DQB1*05:01,HLA-DRB1*09:01,HLA-DQB1*04:02,HLA-DPB1*04:01,HLA-DRB3*01:01,HLA-DQA1*01:01,HLA-DRB1*04:01,HLA-DRB1*04:05,HLA-DQA1*03:01,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPA1*01,HLA-DQA1*04:01
LTENLLLYIDINGNL	431	447	HLA-DPB1*04:02,HLA-DRB1*13:02,HLA-DRB1*12:01,HLA-DQB1*03:02,HLA-DPA1*03:01,HLA-DRB1*03:01,HLA-DRB1*15:01,HLA-DQB1*05:01,HLA-DQA1*03:01,HLA-DRB3*01:01,HLA-DRB1*04:01,HLA-DRB1*04:05,HLA-DQA1*01:01,HLA-DPA1*02:01,HLA-DPB1*01:01
YVLGLAAIMQLFFSY	1521	1537	HLA-DRB1*03:01,HLA-DQA1*01:02,HLA-DRB1*04:01,HLA-DQB1*06:02,HLA-DRB1*08:02,HLA-DRB1*15:01,HLA-DPB1*04:01,HLA-DRB1*11:01,HLA-DPA1*01:03,HLA-DPB1*14:01,HLA-DQA1*04:01,HLA-DQA1*05:01,HLA-DPB1*05:01,HLA-DRB1*12:01,HLA-DRB1*09:01,HLA-DQB1*04:02,HLA-DRB3*01:01,HLA-DQB1*03:01,HLA-DRB1*04:05,HLA-DPA1*02:01,HLA-DPA1*01,HLA-DQB1*02:01,HLA-DRB5*01:01,HLA-DPB1*02:01,HLA-DRB4*01:01

WADNNCYLATALLTL	851	867	HLA-DQA1*05:01,HLA-DPB1*04:02,HLA-DQB1*02:01,HLA-DPB1*05:01,HLA-DPA1*03:01,HLA-DRB1*07:01,HLA-DPB1*02:01,HLA-DPB1*04:01,HLA-DQA1*01:02,HLA-DRB1*11:01,HLA-DPA1*01:03,HLA-DRB1*04:01,HLA-DQB1*06:02,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPB1*14:01,HLA-DPA1*01
NQHEVLLAPLLSAGI	321	337	HLA-DPB1*04:02,HLA-DRB1*12:01,HLA-DPA1*03:01,HLA-DRB1*15:01,HLA-DRB1*03:01,HLA-DRB1*09:01,HLA-DPB1*02:01,HLA-DPB1*14:01,HLA-DRB1*11:01,HLA-DRB1*04:01,HLA-DRB1*01:01,HLA-DRB1*04:05,HLA-DPA1*01:03,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DRB1*08:02
VLSTFISAARQGFVD	1811	1827	HLA-DRB3*02:02,HLA-DQB1*03:02,HLA-DRB1*03:01,HLA-DRB1*07:01,HLA-DRB1*09:01,HLA-DRB5*01:01,HLA-DQB1*04:02,HLA-DRB3*01:01,HLA-DQA1*01:02,HLA-DRB1*11:01,HLA-DRB1*04:01,HLA-DQB1*06:02,HLA-DQA1*03:01,HLA-DRB1*08:02,HLA-DQA1*04:01
SFYYVWKSYVHVVDG	1571	1587	HLA-DPB1*04:02,HLA-DPA1*03:01,HLA-DRB1*15:01,HLA-DRB1*03:01,HLA-DRB1*07:01,HLA-DRB1*09:01,HLA-DRB5*01:01,HLA-DPB1*04:01,HLA-DPB1*02:01,HLA-DRB3*01:01,HLA-DRB1*11:01,HLA-DPA1*01:03,HLA-DRB1*08:02,HLA-DRB1*04:05,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPB1*14:01,HLA-DPA1*01
TKYLVQQESPFVMMS	976	992	HLA-DRB3*02:02,HLA-DQA1*05:01,HLA-DRB1*13:02,HLA-DQB1*02:01,HLA-DRB1*12:01,HLA-DRB1*03:01,HLA-DRB1*15:01,HLA-DRB1*04:05,HLA-DRB1*09:01,HLA-DRB5*01:01,HLA-DRB3*01:01,HLA-DRB4*01:01,HLA-DQA1*01:02,HLA-DRB1*11:01,HLA-DRB1*04:01,HLA-DRB1*01:01,HLA-DQB1*06:02,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPB1*14:01
EHFIETISLAGSYKD	681	697	HLA-DPB1*04:02,HLA-DRB1*12:01,HLA-DPA1*03:01,HLA-DRB1*03:01,HLA-DRB1*07:01,HLA-DRB1*04:05,HLA-DRB5*01:01,HLA-DPB1*02:01,HLA-DQA1*01:02,HLA-DRB1*11:01,HLA-DRB1*04:01,HLA-DPA1*01:03,HLA-DQB1*06:02,HLA-DPA1*02:01,HLA-DPB1*14:01,HLA-DRB1*08:02
KSPNFSKLINIIIWF	1401	1417	HLA-DPB1*04:02,HLA-DRB1*12:01,HLA-DPA1*03:01,HLA-DPB1*05:01,HLA-DRB1*15:01,HLA-DRB1*07:01,HLA-DRB5*01:01,HLA-DRB3*01:01,HLA-DRB4*01:01,HLA-DRB1*11:01,HLA-DRB1*04:01,HLA-DRB1*04:05,HLA-DPA1*02:01,HLA-DPB1*01:01
VRTNVYLAVFDKNLY	351	367	HLA-DRB1*13:02,HLA-DPB1*05:01,HLA-DRB1*03:01,HLA-DRB1*07:01,HLA-DQB1*05:01,HLA-DRB5*01:01,HLA-DPB1*02:01,HLA-DPB1*04:01,HLA-DPB1*14:01,HLA-DRB1*11:01,HLA-DPA1*01:03,HLA-DQA1*01:01,HLA-DPA1*02:01,HLA-DPB1*01:01,HLA-DPA1*01

Table 4-6 Predicted linear B cell epitopes in nsp3 protein using BepiPred 2.0.

Epitope	Start	End	Length
EDEEEGDCEEEEFEPSTQYEYGTEDDYQGKPLEFGATS	111	148	38
EEEQEEDWLDDD	154	165	12
VGQQDGSEDNQ	170	180	11
IVEVQPQLEMELTPVVQTIEV	187	207	21
EVKPFITESKPSVEQRKQDDK	392	412	21
EEVTTTLEETK	419	429	11
YIDINGNLHPDSAT	438	451	14
YILPSIISNEK	536	546	11
RKYKGIKIQEGVVD	586	599	14
DLVPNQPYPNA	1095	1105	11
NATNKATYKPNT	1178	1189	12
DAQGMDNLACEDLKPVSEEVVENPTIQKDVLECNVK	1214	1249	36
YREGYLNSTNVTIA	1448	1461	14
GQKTYERHSLS	1691	1701	11

4.5 Discussion

Our prediction of the potential SARS-CoV-2 antigens, which could induce protective immunity, provides a timely analysis for the vaccine development against COVID-19. Currently, most coronavirus vaccine studies use the whole inactivated or attenuated virus or target the structural proteins such as the spike (S) protein, nucleocapsid (N) protein, and membrane (M) protein. But the inactivated or attenuated whole virus vaccine might cause strong adverse events. On the other hand, vaccines targeting the structural proteins induce a robust immune response (P. Zhao et al. 2005; Shi et al. 2006; Al-Amri et al. 2017). In some studies, these structural proteins, including the S and N proteins, were reported to associate with the pathogenesis of coronavirus (Yasui et al. 2008; Glansbeek et al. 2002) and might raise safety concerns (Weingartl et al. 2004). Recently, the epitopes of the SARS-CoV-2 were computationally predicted and evaluated by sequence homology analysis of SARS-CoV and MERS-CoV epitopes (Grifoni, Sidney, et al. 2020). Following this study, the predicted T cell MHC-I and MHC-II epitopes of SARS-CoV-2 were experimentally evaluated using the "megapools" approach, and both CD4+ and CD8+ responses were detected (Grifoni, Weiskopf, et al. 2020). The present work is complementary but not overlapping with the recent reports. Our study applied state-of-the-art Vaxign reserve vaccinology (RV) and Vaxign-ML machine learning strategies to the entire SARS-CoV-2 proteomes, including both structural and non-structural proteins for vaccine candidate prediction. Our results indicate, for the first time, that many non-structural proteins could be used as potential vaccine candidates.

The SARS-CoV-2 S protein was identified by our Vaxign and Vaxign-ML analysis as the most favorable vaccine candidate. First, the Vaxign RV framework predicted the S protein as a likely adhesin, which is consistent with the role of S protein for the invasion of host cells.

Second, our Vaxign-ML predicted that the S protein had a high protective antigenicity score. These results confirmed the role of S protein as the important target of COVID-19 vaccines. However, targeting only the S protein may induce high serum-neutralizing antibody titers but cannot induce complete protection (See et al. 2008). In addition, HCoV-NL63 also uses S protein and employs the angiotensin-converting enzyme 2 (ACE2) for cellular entry, despite markedly weak pathogenicity (Hofmann et al. 2005). This suggests that the S protein is not the only factor determining the infection level of a human coronavirus. Thus, alternative vaccine antigens may be considered as potential targets for COVID-19 vaccines.

Among the five non-structural proteins being predicted as potential vaccine candidates, the nsp3 protein was predicted to have the second-highest protective antigenicity score, adhesin property, promiscuous MHC-I & MHC-II T cell epitopes, and B cell epitopes. The nsp3 is the largest non-structural protein that includes multiple functional domains related to viral pathogenesis(Lei, Kusov, and Hilgenfeld 2018). The multiple sequence alignment of nsp3 also showed higher sequence conservation in most of the functional domains in SARS-CoV-2, SARS-CoV, and MERS-CoV, than in all 15 coronavirus strains (Figure 4-2 B). Besides the nsp3 protein, our study also predicted four additional non-structural proteins (3CL-pro, nsp8, nsp9, and nsp10) as possible vaccine candidates based on their adhesin probabilities, and the nsp8 protein was also predicted to have a significant protective antigenicity score.

However, these predicted non-structural proteins (nsp3, 3CL-pro, nsp8, nsp9, and nsp10) are not part of the viral structural particle, and all the current SARS/MERS/COVID-19 vaccine studies target the structural (S/M/N) proteins. Although structural proteins are commonly used as viral vaccine candidates, non-structural proteins correlate to vaccine protection. The non-structural protein NS1 was found to induce protective immunity against infections by flaviviruses (Salat et al. 2020). Since NS1 is not part of the virion, antibodies against NS1 have no neutralizing activity but some exhibit complement-fixing activity (Schlesinger, Brandriss, and Walsh 1985). However, passive transfer of anti-NS1 antibody or immunization with NS1 conferred protection (Gibson, Schlesinger, and Barrett 1988). The anti-NS1 antibody could also reduce viral replication by complement-dependent cytotoxicity of infected cells, block NS1-induced pathogenic effects, and attenuate NS1-induced disease development during the critical phase (H. R. Chen, Lai, and Yeh 2018). Finally, NS1 is not a structural protein and the anti-NS1 antibody will not induce antibody-dependent enhancement (ADE), which is a virulence factor and a risk factor causing many adverse events (H. R. Chen, Lai, and Yeh 2018). In addition to the induction of antibody responses, non-structural proteins of viruses could induce virus-specific T cells, especially cytotoxic T lymphocytes, that are important to control viral infection. The non-structural proteins of the hepatitis C virus were reported to induce HCV-specific vigorous and broad-spectrum T-cell responses (Ip et al. 2014). The non-structural HIV-1 gene products were also shown to be valuable targets for prophylactic or therapeutic vaccines (Cafaro et al. 2019). Therefore, it is reasonable to hypothesize that the SARS-CoV-2 non-structural proteins (e.g., nsp3) are possible vaccine targets, which might induce cell-mediated or humoral immunity necessary to prevent viral invasion and/or replication.

The SARS-CoV-2 nsp3 protein was recently reported to account for the virus-specific T cell response. Grifoni et al. showed that the three major structural (S/M/N) proteins accounted for 59% of the total CD4+ T cell response in COVID-19 recovered patients, while other non-structural proteins, including nsp3, also accounted for the response (Grifoni, Weiskopf, et al. 2020). In addition, SARS-CoV-2-reactive CD4+ T cells could be detected in a large portion of unexposed individuals, suggesting cross-reactive T cell recognition between SARS-CoV-2 and the other coronaviruses that only cause the common cold. In our study, the nsp3 protein showed sequence conservation among the 15 coronaviruses, and particularly, the protein shared higher similarity among the more severe form of coronavirus (SARS-CoV, MERS-CoV, and SARS-CoV-2) (Figure 4-2). The preexisting immunity against the mild human coronaviruses might offer cross-protection to the SARS-CoV-2 infected individuals(Grifoni, Weiskopf, et al. 2020). In spite of that, none of the non-structural proteins have been evaluated as vaccine candidates, and the feasibility of these proteins as vaccine targets are subject to further experimental verification.

Besides immunogenicity, safety is also an important factor of a successful COVID-19 vaccine. One of the safety issues of COVID-19 vaccines might occur due to vaccine delivery (e.g., vectors, adjuvants, formulation doses, or route of administration), which cannot be evaluated by the machine learning approach presented in this study. In addition, the nsp3 and other viral adhesive proteins with sequence homology to the host cell adhesion molecules might also cause auto-reactivity with self-antigen or induce T regulatory, leading to low responsiveness of the host to the virus. By applying Vaxign and epitope predictions, our study found that the MAC1 domain of nsp3 protein shares sequence homology with the human mono-ADP-ribosyltransferase PARP14, and there is no predicted T cell MHC-I, MHC-II, and linear B cell epitopes within the aligned region.

In addition to vaccines expressing a single or a combination of structural proteins, here we propose an "Sp/Nsp cocktail vaccine" as an effective strategy for COVID-19 vaccine development. A typical cocktail vaccine includes more than one antigen to cover different aspects of protection (Sealy et al. 2009; Millet et al. 1993). The licensed Group B meningococcus Bexsero vaccine, which was developed via reverse vaccinology, contains three protein antigens (Folaranmi et al. 2015). To develop an efficient and safe COVID-19 cocktail vaccine, an "Sp/Nsp cocktail vaccine", which mixes a structural protein(s) (Sp, such as S protein) and a non-structural protein(s) (Nsp, such as nsp3) could induce more favorable protective immune responses than vaccines expressing a structural protein(s). Current COVID-19 vaccines mostly target the S protein with various types of delivery systems (such as recombinant virus vectors), and none of the non-structural proteins has not been used. The benefit of a cocktail vaccine strategy could induce immunity that can protect the host against not only the S-ACE2 interaction and viral entry to the host cells but also protect against the accessary non-structural adhesin proteins (e.g., nsp3), which might also be vital to the viral entry and replication. The usage of more than one antigen allows us to reduce the volume of each antigen and thus to reduce the induction of adverse events. Nonetheless, the potential and safety of the proposed "Sp/Nsp cocktail vaccine" strategy need to be experimentally validated.

For rational COVID-19 vaccine development, it is critical to understand the fundamental host-coronavirus interaction and protective immune mechanism (Roper and Rehm, 2009). Such understanding may not only provide us guidance in terms of antigen selection but also facilitate our design of vaccine formulations. For example, an important foundation of our prediction in this study is based on our understanding of the critical role of adhesin as a virulence factor as well as a protective antigen. The choice of DNA vaccine, recombinant vaccine vector, and another method of vaccine formulation is also deeply rooted in our understanding of pathogen-specific immune response induction. Different experimental conditions may also affect results (He et al. 2014; E. Ong et al. 2019). Therefore, it is crucial to understand the underlying molecular and cellular mechanisms for rational vaccine development.

4.6 Acknowledgement

I appreciate the support of the US National Institutes of Health (1R01AI081062). The result presented in this chapter is published in (E. Ong, Wong, Huffman, and He 2020). Edison Ong (EO) and Yongqun He (YH) designed the study. EO, Mei U Wong (MW), and Anthony Huffman collected the data. EO performed bioinformatics analysis. EO, MW, and YH wrote the initial manuscript. All authors performed result interpretation, discussed, and reviewed the manuscript.

Chapter 5 Computational Design of SARS-CoV-2 Spike Glycoproteins to Increase Immunogenicity by T Cell Epitope Engineering

5.1 Abstract

The development of effective and safe vaccines is the ultimate way to efficiently stop the ongoing COVID-19 pandemic, which is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Built on the fact that SARS-CoV-2 utilizes the association of its Spike (S) protein with the human Angiotensin-converting enzyme 2 (ACE2) receptor to invade host cells, we computationally redesigned the S protein sequence to improve its immunogenicity and antigenicity. Toward this purpose, we extended an evolutionary protein design algorithm, EvoDesign, to create thousands of stable S protein variants that perturb the core protein sequence but keep the surface conformation and B cell epitopes. The T cell epitope content and similarity scores of the perturbed sequences were calculated and evaluated. Out of 22,914 designs with favorable stability energy, 301 candidates contained at least two pre-existing immunity-related epitopes and had promising immunogenic potential. The benchmark tests showed that, although the epitope restraints were not included in the scoring function of EvoDesign, the top S protein design successfully recovered 31 out of the 32 major histocompatibility complex (MHC) -II T cell promiscuous epitopes in the native S protein, where two epitopes were present in all seven human coronaviruses. Moreover, the newly designed S protein introduced nine new MHC-II T cell promiscuous epitopes that do not exist in the wildtype SARS-CoV-2. These results demonstrated a new and effective avenue to enhance a target protein's immunogenicity using

rational protein design, which could be applied for new vaccine design against COVID-19 and other human viruses.

5.2 Introduction

The current Coronavirus Disease 2019 (COVID-19) pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has resulted in over 77 million confirmed cases and over 1.7 million deaths globally as of December 24, 2020, according to the World Health Organization (World Health Organization 2020b). Tremendous efforts have been made to develop effective and safe vaccines against this viral infection. The Pfizer-BioNTech BNT162b2 mRNA vaccine showed 95% effectiveness in preventing COVID-19 (Polack et al. 2020), and the Moderna mRNA-1273 induced strong immune responses among recipients between the age of 18 and 55 during phase III clinical trials (Anderson et al. 2020). The FDA issued an Emergency Use Authorization for the mRNA-1273 and BNT162b2 mRNA vaccines in record time. On the other hand, the Inovio INO-4800 DNA vaccine not only showed protection from the viral infection in rhesus macaques but was also reported to induce long-lasting memory (A. Patel et al. 2020). In addition to these two vaccines, there are over a hundred COVID-19 vaccines currently in clinical trials, including other types of vaccines such as the Oxford-AstraZeneca adenovirus-vectored vaccine (ChAdOx1 nCoV-19) (Folegatti et al. 2020), CanSino's adenovirus type-5 (Ad5)-vectored COVID-19 vaccine (F. C. Zhu et al. 2020), and Sinovac's absorbed COVID-19 (inactivated) vaccine (ClinicalTrials.gov Identifier: NCT04456595). Among all the vaccines, a vast majority of them select the spike glycoprotein as their primary target.

The SARS-CoV-2 Spike (S) protein is a promising vaccine target, and many clinical studies reported anti-S protein neutralizing antibodies in COVID-19 recovered patients (Grifoni, Weiskopf, et al. 2020). After the SARS outbreak in 2003 (Lu et al. 2020), clinical studies reported neutralizing antibodies targeting the SARS-CoV S protein (Temperton et al. 2005; Chan et al. 2015), which was selected as the target of vaccine development (Shim et al. 2010; Z. Y. Yang et al. 2004). Since SARS-CoV-2 shares a high sequence identity with SARS-CoV (Zhou et al. 2020), and both viruses utilize the attachment of the S protein to the human angiotensin-converting enzyme 2 (ACE2) receptor to invade host cells, neutralization of the SARS-CoV-2 S protein could induce protection in COVID-19 vaccine development (Tay et al. 2020). Many computational studies utilizing reverse vaccinology and immuno-informatics reported the S protein to be a promising vaccine antigen (E. Ong, Wong, Huffman, and He 2020; Grifoni, Sidney, et al. 2020; Enayatkhani et al. 2020), and clinical studies identified anti-S protein neutralizing antibodies in patients that have recovered from COVID-19 (F. Wu et al. 2020; L. Ni et al. 2020; Cao et al. 2020). The cryo-EM structures of the S protein (Wrapp, Wang, et al. 2020) and the neutralizing antibodies that bind to the S protein (Barnes et al. 2020; Wrapp, De Vlieger, et al. 2020) were determined. Besides neutralizing antibodies, studies have also shown the importance of the CD4 T cell response in the control of SARS-CoV-2 infection and possible pre-existing immunity in healthy individuals without exposure to SARS-CoV-2 (Grifoni, Weiskopf, et al. 2020; Bert et al. 2020; Braun et al. 2020). Kalita et al. has proposed a multi-peptide subunit-based epitope vaccine that is comprised of B cell, helper T cell, and cytotoxic T cell epitopes (Kalita et al. 2020). Overall, successful vaccination is likely linked to a robust and long-term humoral response to the SARS-CoV-2 S protein, which could be further enhanced by the rational structural design of the protein.

Structural vaccinology has shown success in improving vaccine candidates' immunogenicity through protein structural modification. The first proof-of-concept was achieved by fixing the conformation-dependent neutralization-sensitive epitopes on the fusion glycoprotein of the respiratory syncytial virus (McLellan et al. 2013). A similar strategy has been applied to SARS-CoV-2 to conformationally control the S protein's receptor-binding domain (RBD) domain between the "up" and "down" configurations to induce immunogenicity (Henderson et al. 2020). Besides the structure-based rational design approaches, the directed evolution is classified as an irrational method and has also been widely used in diverse fields, such as enzyme engineering (Hall 1978), protein-RNA interaction (Morozova, Myers, and Shamoo 2006), and COVID-19 therapeutic strategies (Padhi et al. 2020). The directed evolution usually first applies random mutagenesis to generate a large pool of variants, followed by screening for candidates with the preferred properties using high-throughput strategies. The advantage of directed evolution is that it works well without structural information. However, once a high-quality protein structure can be obtained either from the experimental determination or computational prediction, the structure-based approach is more suitable as it can efficiently explore a much larger sequence/conformational space using computer programs, yielding a few potential candidates for further screening and validation, which is more time-, money-, and labor-saving compared to a typical directed evolution process.

In this study, we extended structural vaccinology to a structure-based computational design of the SARS-CoV-2 S protein. Briefly, we used a protein design approach, EvoDesign (Pearce et al. 2019), to generate multiple stable S protein variants without perturbing the surface amino acids to maintain the same B cell epitope profile. Meanwhile, we introduced mutations to the residues buried inside the S protein so that more major histocompatibility complex (MHC)-II T cell epitopes would be added into the newly designed S protein to potentially induce a stronger immune response. Finally, we evaluated the computationally designed protein candidates and compared them to the native S protein.

5.3 Methods

5.3.1 Computational redesign of SARS-CoV-2 S protein

Figure 5-1 illustrates the workflow for redesigning the SARS-CoV-2 S protein to improve its immunogenic potential for vaccine design. The full-length structure model (1,273 amino acids for the S monomer) of SARS-CoV-2 S assembled by C-I-TASSER (C. Zhang et al. 2020) was used as the template for fixed-backbone protein sequence design using EvoDesign (Pearce et al. 2019). Although the cryo-EM structure for SARS-CoV-2 S is available (PDB ID: 6VSB) (Wrapp, Wang, et al. 2020), it contains a large number of missing residues, and therefore, the full-length C-I-TASSER model was used for S protein design instead. The C-I-TASSER model used the cryo-EM density map to assemble the individual domain models and to refine the structure. The model showed a high similarity to the cryo-EM structure with a TM-score (Y. Zhang and Skolnick 2005) of 0.87 and root-mean-square deviation (RMSD) of 3.4 Å in the commonly aligned regions, indicating a good model quality. The residues in the S protein were categorized into three groups: core, surface, and intermediate (X. Huang, Pearce, and Zhang 2020b), according to their solvent accessible surface area ratio (SASAr). Specifically, SASAr is defined as the ratio of the absolute SASA of a residue in the structure to the maximum area of the residue in the GXG state (Tian, Huang, and Zhu 2015), where X is the residue of interest and G is a glycine residue. The most extended GXG conformation measures the maximum exposure degree of the residue X in the solvated environment taking into account the local protein backbone. The SASAr ratios were calculated using the ASA web-server (http://cib.cf.ocha.ac.jp/bitool/ASA/), where the maximum area of each of the 20 canonical amino acid residues is provided. The core and surface residues were defined by us as those with SASAr <5% and >25%, respectively, while the other residues were regarded as intermediate.

Since the surface residues may be involved in the interactions with other proteins (e.g., the formation of the S homotrimer, S-ACE2 complex, and S-antibody interaction) and may partially constitute the B cell epitopes, these residues were excluded from design, and more rigorously, their side-chain conformations were kept constant as well.

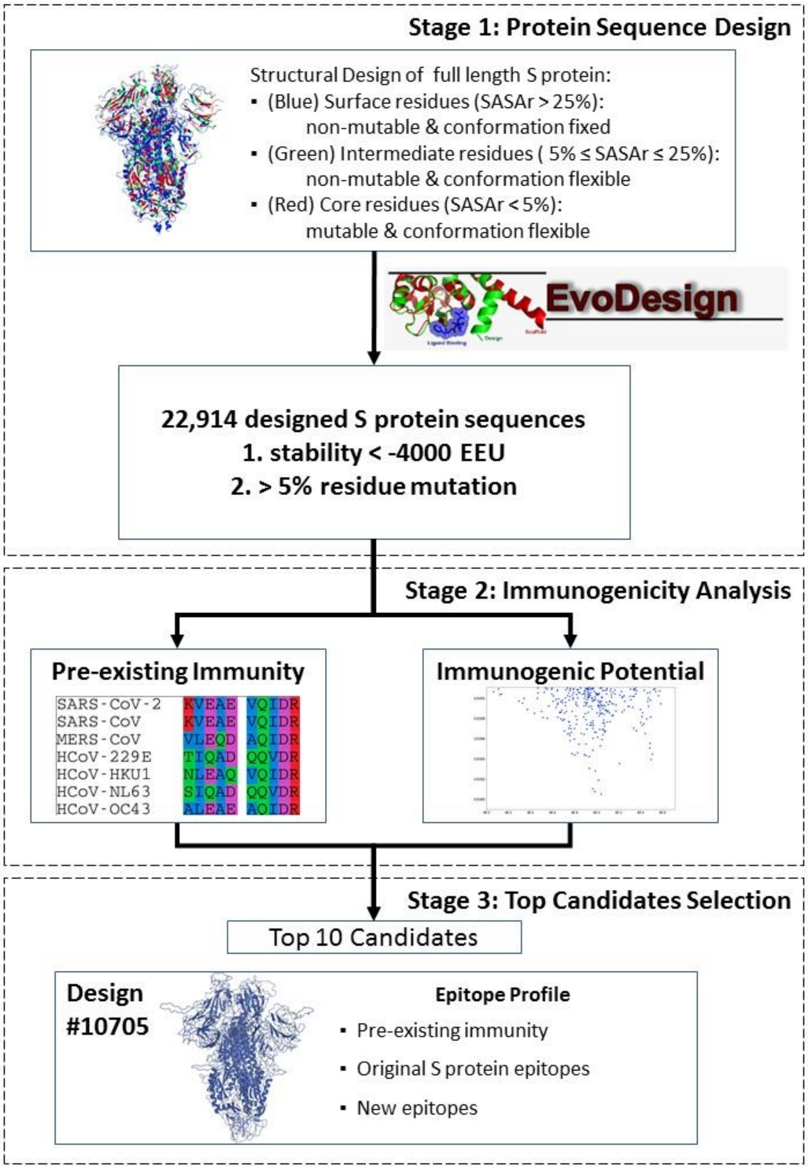

Stage 1: Protein Sequence Design
Structural Design of full length S protein:
▪ (Blue) Surface residues (SASAr > 25%):
non-mutable & conformation fixed
▪ (Green) Intermediate residues (5% ≤ SASAr ≤ 25%):
non-mutable & conformation flexible
▪ (Red) Core residues (SASAr < 5%):
mutable & conformation flexible
EvoDesign
22,914 designed S protein sequences
1. stability < -4000 EEU
2. > 5% residue mutation
Stage 2: Immunogenicity Analysis
Pre-existing Immunity
SARS-CoV-2 KVEAE VQIDR
SARS-CoV KVEAE VQIDR
MERS-CoV VLEQD AQIDR
HCoV-229E TIQAD QQVDR
HCoV-HKU1 NLEAQ VQIDR
HCoV-NL63 SIQAD QQVDR
HCoV-OC43 ALEAE AQIDR
Immunogenic Potential
Stage 3: Top Candidates Selection
Top 10 Candidates
Design #10705
Epitope Profile
▪ Pre-existing immunity
▪ Original S protein epitopes
▪ New epitopes

Figure 5-1 The workflow for designing and screening immunogenicity-enhanced SARS-CoV-2 S proteins.

The procedure started by defining the full-length SARS-CoV-2 native S protein into surface, intermediate, and core residues. This information was then fed into EvoDesign to generate structurally stable designs that introduce mutations to the core residues while keeping the surface conformation unchanged. The output design candidates from EvoDesign were then evaluated based on their immunogenic potential. The top ten candidates were also compared and evaluated in comparison to the native S protein.

Additionally, the residues that may form B cell epitopes reported by Grifoni et al. (Grifoni, Sidney, et al. 2020) were also fixed. The remaining core residues were subjected to design, allowing amino acid substitution, whereas the intermediate residues were repacked with conformation substitution. Specifically, 243, 275, and 755 residues were designed, repacked, and fixed, respectively. The 243 designable core residues were also compared to the global S protein mutations (global frequencies > 0.001) recorded in the GISAID database (as of December 7, 2020) (Korber et al. 2020; Elbe and Buckland-Merrett 2017). These residues were also evaluated for their intrinsic disorder predisposition based on the reported disorder regions in the DisProt database (Hatos et al. 2020). The corresponding Jensen-Shannon Divergence (JSD) scores (higher scores indicate greater conservation) of these core residues residing within the disordered regions were reported (Capra and Singh 2007). During protein design, the evolution term in EvoDesign was turned off as this term would introduce evolutionary constraints on the sequence simulation search, which were not needed for this design (X. Huang, Pearce, and Zhang 2020a); therefore, only the physical energy function, EvoEF2 (X. Huang, Pearce, and Zhang 2020b), was used for design scoring to broaden sequence diversity and help to identify more candidates with increased immunogenicity. In previous studies, EvoEF2 has been appropriately utilized to model the binding interactions between the SARS-CoV-2 S-RBD and a large number of ACE2 orthologs to identify the zoonotic origin of this novel coronavirus (X. Huang et al. 2020) and to design multiple anti-SARS-CoV-2 peptide therapeutics (X. Huang, Pearce, and Zhang 2020a). We performed 20 independent design simulations and collected all the simulated sequence decoys. A total of 5,963,235 sequences were obtained, and the best-scoring sequence had stability energy of -4100.97 EvoEF2 energy unit (EEU). A set of 22,914 non-redundant sequences that were within a 100 EEU window of the lowest energy and had >5% of the design residues mutated were retained for further analysis (Fig. 1). We also utilized another popular protein design software, Rosetta (Leaver-Fay et al. 2011), to generate 1000 low-energy S variants using the "*fixbb*" protocol due to lower computational efficiency. The same surface-intermediate-core criterion was applied to the Rosetta protein design process. The EvoDesign and Rosetta designs were then analyzed and compared to examine the advantages/limitations of EvoDesign designs.

5.3.2 MHC-II T cell epitope prediction and epitope content score calculation

The full-length S protein sequence was divided into 15-mers with ten amino-acid overlaps. For each 15-mer, the T cell MHC-II promiscuous epitopes were predicted using NetMHCIIpan v3.2 (Jensen et al. 2018). In brief, the percentile ranks of an epitope binding to each of the seven MHC-II alleles (i.e., HLA-DRB1*03:01, HLA-DRB1*07:01, HLA-DRB1*15:01, HLA-DRB3*01:01, HLA-DRB3*02:02, HLA-DRB4*01:01, and HLA-DRB5*01:01) were calculated, where the percentile rank was generated by comparing the 15-mer predicted binding affinity to the MHC-II molecule against that of a large set of similarly sized peptides randomly selected from the SWISS-PROT database (Dhanda et al. 2019). An epitope was considered a promiscuous epitope if the median percentile rank was ≤ 20% by binding the 15-mer to any of the seven MHC-II alleles (Fleri et al. 2017). The selection of these seven MHC-II alleles aimed to predict the dominant MHC-II T cell epitopes across different ethnicities and HLA polymorphisms (Paul et al. 2015). The MHC-II promiscuous epitopes of the native SARS-CoV-2 S protein (QHD43416) predicted using this method were also validated and compared to the dominant T cell epitopes mapped by Grifoni et al. (Grifoni, Sidney, et al. 2020). In brief, Grifoni et al. mapped the experimentally verified SARS-CoV T cell epitopes reported in the Immune Epitope Database (IEDB) database, which includes experimentally verified T cell MHC-II epitope data, to the SARS-CoV-2 S protein based on sequence homology and reported as the dominant T cell epitopes (Fleri et al. 2017). The epitope content score (ECS) for a full-length S protein was calculated as the average value of the median percentile ranks for all the 15-mers spanning the whole sequence.

5.3.3 Human epitope similarity and human similarity score calculation

The human proteome included 20,353 reviewed (Swiss-Prot) human proteins downloaded from Uniprot (as of July 1, 2020) (The UniProt Consortium 2008). A total of 261,908 human MHC-II T cell promiscuous epitopes were predicted, as described above. The human epitope similarity between a peptide of interest (e.g., a peptide of the S protein) and a human epitope was then calculated using a normalized peptide similarity metric proposed by Frankild et al. (Frankild et al. 2008). In brief, the un-normalized peptide similarity score, $A(x, y)$, was first determined by the BLOSUM35 matrix (Henikoff and Henikoff 1992) for all the positions between a target peptide (y) and a human epitope (x), which was subsequently normalized using the minimum and maximum similarity scores for the human epitope (Equation 1). The maximum and minimum similarity scores were determined from a range of similarity scores between a human epitope and all of its possible amino acid substitutions. Finally, the maximum normalized similarity score of a 15-mer peptide was calculated by comparing it to all the predicted human MHC-II T cell promiscuous epitopes. The human similarity score (HSS) of the full-length S protein was calculated by averaging the human epitope similarity of all the 15-mers.

$$S(x, y) = \frac{A(x,y) - A_{min}^{x}}{A_{max}^{x} - A_{min}^{x}} \qquad (1)$$

5.3.4 Pre-existing immunity evaluation of the designed proteins

The pre-existing immunity of the designed proteins was evaluated and compared to that of the native S protein of seven human coronaviruses (HCoVs) (i.e., SARS-CoV-2, SARS-CoV, MERS-CoV, HCoV-229E, HCoV-OC43, HCoV-NL63, and HCoV-HKU1). The sequences of the seven HCoV S proteins were downloaded from Uniprot (The UniProt Consortium 2008) (Table 5-1), and the MHC-II T cell epitopes were predicted as described above. The conserved epitopes were determined by the IEDB epitope clustering tool (Dhanda et al. 2019) and aligned using SEAVIEW (Gouy, Guindon, and Gascuel 2010).

5.3.5 Foldability assessment of the designed proteins

Since EvoDesign only produces a panel of mutated sequences, it is important to examine if the designed sequences can fold into the desired structure that the native S protein adopts. To examine their foldability, we used C-I-TASSER to model the structure of the designed sequences, where the structural similarity between the native and designed S proteins was assessed by TM-score (Y. Zhang and Skolnick 2004). Here, C-I-TASSER is a recently developed protein structure prediction program, which constructs full-length structure folds by assembling fragments threaded from the PDB, under the guidance of deep neural-network learning-based contact maps (Y. Li, Zhang, et al. 2019; Y. Li, Hu, et al. 2019). The ectodomain of the S homotrimers and the functional domains including the N-terminal domain (NTD), receptor-binding domain (RBD), fusion peptide (FP), heptapeptide repeat sequence 1 (HR1), and connector domain (CD) (Y. Huang et al. 2020; Henderson et al. 2020) were visualized via PyMOL (Schrödinger 2015). Sequence logo plots for the top ten and worst ten S protein designs were also generated (Crooks et al. 2004). The multiple sequence alignment of the top four EvoDesign S protein candidates with balanced ECS and HSS were aligned to the native S protein using SEAVIEW (Gouy, Guindon, and Gascuel 2010).

Table 5-1 Seven human coronavirus S proteins.

Spike Protein ID	Organism	Organism Taxon ID
P59594	Human SARS coronavirus (SARS-CoV) (Severe acute respiratory syndrome coronavirus)	694009
R9UQ53	Middle East respiratory syndrome-related coronavirus	1335626
P15423	Human coronavirus 229E (HCoV-229E)	11137
Q0ZME7	Human coronavirus HKU1 (isolate N5) (HCoV-HKU1) (Strain: Isolate N5)	443241
P36334	Human coronavirus OC43 (HCoV-OC43)	31631
Q6Q1S2	Human coronavirus NL63 (HCoV-NL63)	277944
P0DTC2	Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) (SARS-CoV-2)	2697049

5.3.6 Molecular dynamics (MD) simulation

The extracellular domain (amino acids 1-1146) of the trimeric wildtype S and the top design (i.e., Design-10705, see Results) was subjected to MD simulation using GROMACS (Abraham et al. 2015) with the CHARMM36 force field (Best et al. 2012). The initial full-length (amino acids 1-1273) trimers were built by C-I-TASSER and residues 1147-1273 were deleted; glycosylation was not considered during structure modeling and the simulation. In each simulation case, a dodecahedron box was constructed with a distance of 10 Å from the solute, and TIP3P (Jorgensen et al. 1983) water molecules were filled into the box. The system was then neutralized by the addition of an appropriate number of Na^+ or Cl^- ions. After the system was assembled, energy minimization was carried out using steepest descent minimization with a maximum force of 10 kJ/mol. The system was then equilibrated at 300 K using 100 ps NVT simulations and 100 ps NPT simulations with position restraints (1000 kJ/mol) on the heavy atoms of the protein. After the two equilibration phases, the system was well-equilibrated at the desired temperature and pressure. Unconstrained production MD was then carried out at 300 K for 50 ns as suggested in similar MD simulation studies (Q. Li, Huang, and Zhu 2014; Xue, Huang, and Zhu 2019). The LINCS (Hess et al. 1997) algorithm was used to restrain the hydrogen bonds. Non-bonded interactions were truncated at 12 Å, and the Particle Mesh Ewald (Essmann et al. 1995) method was utilized for long-range electrostatic interactions. The velocity-rescaling thermostat (Bussi, Donadio, and Parrinello 2007) and Parrinello-Rahman barostat (Parrinello and Rahman 1981) were used to couple the temperature and pressure, respectively. About 25000 snapshots were saved with a time interval of 2 fs and utilized for further analysis using the built-in GROMACS command-line tools.

5.4 Results

The epitope content score (ECS) and human similarity score (HSS) of the S proteins from seven HCoV strains (severe HCoV: SARS-CoV-2, SARS-CoV, and MERS-CoV; mild HCoV: HCoV-229, HCoV-HKU1, HCoV-NL63, and HCoV-OC43) were computed. The ECS for the severe HCoV S proteins (mean=49.3, standard deviation (SD)=24.7) was significantly different (p = 0.0016, Mann-Whitney) from that of the mild ones (mean=45.8, SD=24.5). In terms of HSS, the severe HCoV S proteins (mean=0.640, SD=0.03) tended to be less self-like (p = 0.097, Mann-Whitney) than the mild ones (mean=0.642, SD=0.03). Overall, it was shown that both ECS and HSS might be used as indicators of the immunogenic potential of the designed S proteins.

On the other hand, previous studies suggested the potential role of pre-existing immunity in fighting COVID-19 (Grifoni, Weiskopf, et al. 2020; Bert et al. 2020; Braun et al. 2020). Therefore, the predicted MHC-II T cell promiscuous epitopes of the SARS-CoV-2 S protein were compared to those from the other six HCoVs. There were two SARS-CoV-2 predicted MHC-II T cell promiscuous epitopes, which were also present on all of the seven HCoV S proteins (Figure 5-2) and could be potentially linked to pre-existing immunity. Therefore, the designs were subsequently filtered based on the availability of these pre-existing immunity-related epitopes (Figure 5-1). In particular, the SARS-CoV-2 promiscuous epitope S816-D830 overlapped with the dominant B cell epitope F802-E819 reported by Grifoni et al. (Grifoni, Sidney, et al. 2020).

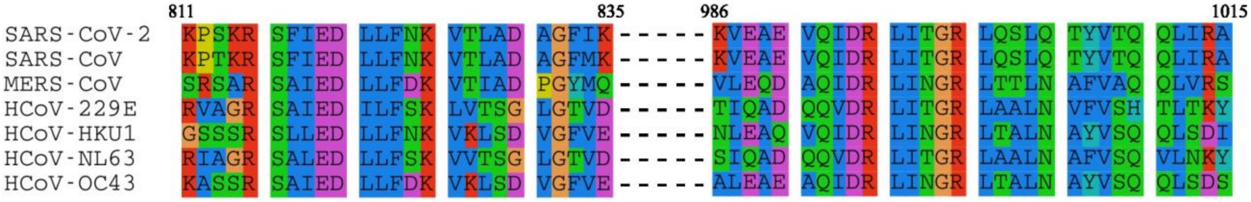

811
835
986
1015
SARS-CoV-2 KPSKR SFIED LLFNK VTLAD AGFIK ----- KVEAE VQIDR LITGR LQSLQ TYVTQ QLIRA
SARS-CoV KPTKR SFIED LLFNK VTLAD AGFMK ----- KVEAE VQIDR LITGR LQSLQ TYVTQ QLIRA
MERS-CoV SRSAR SAIED LLFDK VTIAD PGYMQ ----- VLEQD AQIDR LINGR LTTLN AFVAQ QLVRS
HCoV-229E RVAGR SAIED ILFSK LVTSG LGTVD ----- TIQAD QQVDR LITGR LAALN VFVSH TLTKY
HCoV-HKU1 GSSSR SLLED LLFNK VKLSD VGFVE ----- NLEAQ VQIDR LINGR LTALN AYVSQ QLSDI
HCoV-NL63 RIAGR SALED LLFSK VVTSG LGTVD ----- SIQAD QQVDR LITGR LAALN AFVSQ VLNKY
HCoV-OC43 KASSR SAIED LLFDK VKLSD VGFVE ----- ALEAE AQIDR LINGR LTALN AYVSQ QLSDS

Figure 5-2 The two pre-existing immunity-related SARS-CoV-2 MHC-II T cell promiscuous epitopes.

The first SARS-CoV-2 promiscuous epitope is located within residues 816-830 (indexed by SARS-CoV-2).

EvoDesign generated a total of 22,914 low-energy S protein designs, in which 243 core residues were subjected to substitution (see Methods for details). As SARS-CoV-2 has been reported to have substantial mutations in its genome (Padhi and Tripathi 2020), it is important to compare the EvoDesign mutations to the natural mutations (global mutation frequency of >0.001) reported by GISAID (Table 5-2). There were two EvoDesign core residues (D80 and S98) that also had natural mutations, and these two core residues had different mutation rates in EvoDesign in comparison to the natural infection. Specifically, for the D80 core residue, the natural mutation frequency (D80Y) was 0.005, but the EvoDesign mutations, D80N, D80A, and D80S, had frequencies of 0.149, 0.106, and 0.003, respectively. For S98, the natural mutation (S98F) frequency was 0.007, but EvoDesign mutations S98T and S98A had frequencies of 0.837 and 0.008, respectively. To further investigate whether the EvoDesign candidates' mutations were related to the intrinsic disorder predisposition, the 243 core residues were aligned to the DisProt database (Hatos et al. 2020) and the aligned residues' sequence conservations were evaluated. Twenty core residues were annotated as intrinsically disordered in DisProt, but these residues showed relatively high levels of conservation, with JSD scores ranging from 0.76 to 0.85 (Table 5-3), in the top 10 designs.

Table 5-2 Global S protein mutations compared with the EvoDesign core residues' substitution frequencies.

Position	Original Residue	Reported Mutation	Global Frequency	EvoDesign Residue Def.	EvoDesign Substitution & Frequency
80	D	Y	0.005	Designable (core residue)	unmutated: 0.741 D80N: 0.149 D80A: 0.106 D80S: 0.003
98	S	F	0.007	Designable (core residue)	unmutated: 0.155 S98T: 0.837 S98A: 0.008
5	L	F	0.013	Fixed	---
18	L	F	0.095	Fixed	---
21	R	I	0.007	Fixed	---
68	I	-	0.019	Fixed	---
69	H	-	0.019	Fixed	---
70	V	I	0.019	Fixed	---
144	Y	-	0.005	Fixed	---
176	L	F	0.003	Fixed	---
215	D	H	0.003	Fixed	---
222	A	V	0.194	Fixed	---
253	D	G	0.003	Fixed	---
262	A	S	0.011	Fixed	---
272	P	L	0.007	Fixed	---
439	N	K	0.016	Fixed	---
453	Y	F	0.004	Fixed	---
477	S	N	0.063	Fixed	---
501	N	Y	0.005	Fixed	---
570	A	D	0.004	Fixed	---
583	E	D	0.007	Fixed	---
614	D	G	0.895	Fixed	---
626	A	S	0.004	Fixed	---
655	H	Y	0.004	Fixed	---
681	P	H	0.005	Fixed	---
688	A	V	0.004	Fixed	---
716	T	I	0.004	Fixed	---
723	T	I	0.004	Fixed	---
936	D	Y	0.005	Fixed	---
982	S	A	0.004	Fixed	---
1073	K	N	0.004	Fixed	---
1118	D	H	0.004	Fixed	---
1163	D	Y	0.007	Fixed	---
1167	G	V	0.006	Fixed	---
1263	P	L	0.003	Fixed	---

Table 5-3 The intrinsic disorder predisposition of the EvoDesign core residues and their corresponding conservation scores.

DisProt disorder regions	Top 10 EvoDesign conservation score
67	0.84041
72	0.81773
75	0.79401
76	0.76281
77	0.76185
79	0.81535
80	0.80062
142	0.81402
143	0.84579
145	0.84866
250	0.76164
259	0.83607
260	0.81371
261	0.81059
673	0.8344
851	0.80869
1241	0.80855
1242	0.75916
1248	0.82732
1253	0.80943

Among the 22,914 designs with relatively low (favorable) stability energy, 19,063 candidates that contained the two pre-existing immunity-related epitopes were ranked based on ECS and HSS (Figure 5-3). Using the ECS and HSS of the native SARS-CoV-2 S as the cutoff, we obtained 301 candidates with better immunogenic potential (i.e., lower ECS and HSS) (Fig 3B). Ten candidates with balanced ECS and HSS were selected and evaluated (Table 5-4). The EvoDesign energy and sequence identity of all designs were plotted, and the top 10 designs were highlighted (Figure 5-4). The S protein variants generated by EvoDesign had consistently better ECS in comparison to the Rosetta designs, although the latter had a better HSS score (Figure 5-5). All 1000 Rosetta designs had higher ECS (thus worse immunogenic potential) than the native S, whereas EvoDesign was able to produce a few designs with both lower ECS and HSS, affirming EvoDesign's ability to design vaccine candidates with better T cell promiscuous epitopes.

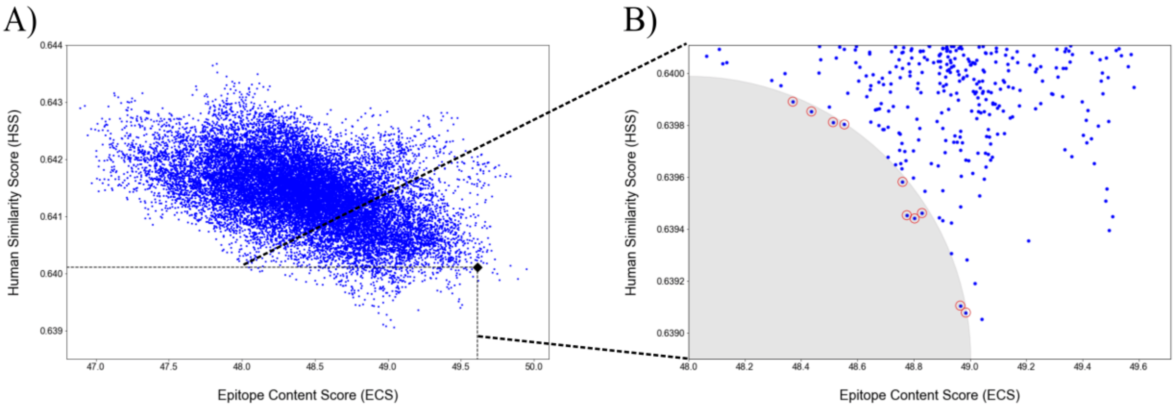


Figure 5-3 The epitope content score (ECS) and human similarity score (HSS) for designed S proteins.

(A) All 22,914 designs. Each design is shown as a blue dot, whereas the native SARS-CoV-2 S was plotted as a black diamond marker. The dashed-line box defines the 301 candidates with both lower ECS and HSS scores than the native. (B) The shaded area contains the top ten candidates (highlighted by red circles) with balanced ECS and HSS scores.

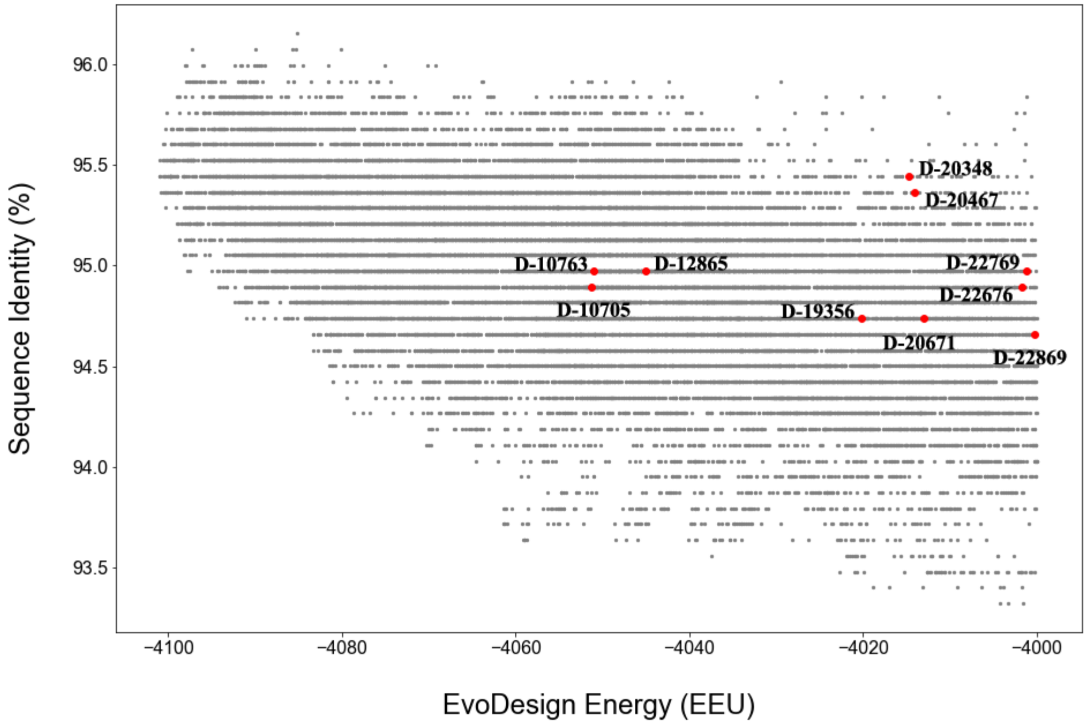


Figure 5-4 The EvoDesign energy and sequence identity for designed S proteins.

The top ten EvoDesign S protein variants were highlighted in the EvoDesign energy vs. sequence identity. The best design Design-10705 with optimizaed immunogenicity had moderate sequence identity and energy stability.

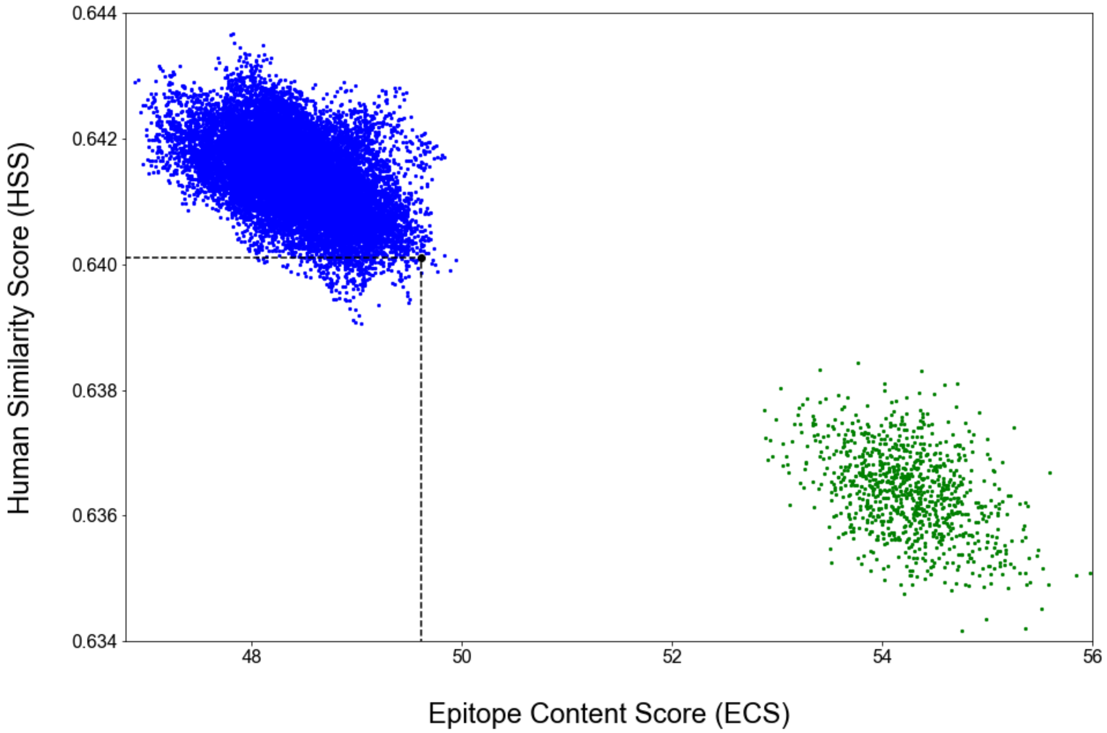


Figure 5-5 The comparison of epitope content score (ECS) and human similarity score (HSS) for designed S proteins between EvoDesign and Rosetta.

All the 1000 Rosetta designs had higher ECS (thus worse immunogenic potential) than the native S, whereas EvoDesign was able to produce a few designs with both lower ECS and HSS, and hence, better predicted immunogenicity for vaccine development.

Table 5-4 Summary of the features for the top 10 EvoDesign S protein candidates.

Design ID	PEC	REC [a]	ECS	HSS	EE (EEU)	RMSD (Å) [b]	TM-score [b]	SI (%)
10705	40	31	48.78	0.6394	-4051.21	3.45	0.931	94.9
10763	40	31	48.80	0.6394	-4051.04	3.06	0.944	95.0
12865	40	31	48.76	0.6396	-4044.99	3.14	0.939	95.0
19356	41	30	48.44	0.6399	-4020.14	3.12	0.929	94.7
20348	38	30	48.99	0.6390	-4014.74	3.33	0.929	95.4
20467	38	30	48.97	0.6391	-4014.10	4.32	0.901	95.4
20671	37	28	48.83	0.6395	-4013.03	3.36	0.940	94.7
22676	36	28	48.37	0.6399	-4001.70	3.35	0.939	95.0
22769	38	28	48.51	0.6398	-4001.11	3.27	0.937	95.0
22869	38	28	48.55	0.6398	-4000.23	3.24	0.919	94.7
Native	32	--	49.61	0.6401	--	--	--	--

PEC: Promiscuous Epitope Count; REC: Recovered Epitope Count; ECS: Epitope Content Score; HSS: Human Similarity Score; EE: EvoDesign Energy (in EEU, EvoEF2 energy unit); RMSD: Root Mean Square Deviation; TM: TM-score; SI: Sequence identity.

[a]: The number of predicted promiscuous epitopes in designs that overlap with those in the native S protein.

[b]: The RMSD and TM-score compared to the C-I-TASSER model of the native S protein.

The multiple sequence alignment of the top four candidates showed that 88 of the 243 core residues were mutated at least once (Figure 5-6). There were 32 core residues substituted to the same amino acids (R34T, V62I, I100M, R102Q, C136T, V143T, Y145S, E191V, T250A, Y279F, R328K, V341I, V350S, W353A, D420A, Y423M, C432V, S438V, V512I, T523N, T599L, S673T, N777T, S875A, T881I, L916Y, C1043A, F1052H, S1055A, C1241G, S1242G, C1248T) in all top four designs. Additionally, the ten top and ten worst designs were also plotted to infer functionally important mutations to enhance immunogenicity (Figure 5-7). Specifically, there were 12 core residues in both the top-scoring and worst-scoring designs that were substituted to the same amino acids in comparison to the native S protein (V62I, C136T, V143T, Y145S, E191V, R328K, V341I, D420A, C432V, T599L, S1055A, C1241G). In particular, two remained unmutated in the top-scoring designs but were mutated in the worst-scoring designs (Y265W and V267T), suggesting mutations of these two residues might result in reduced immunogenicity.

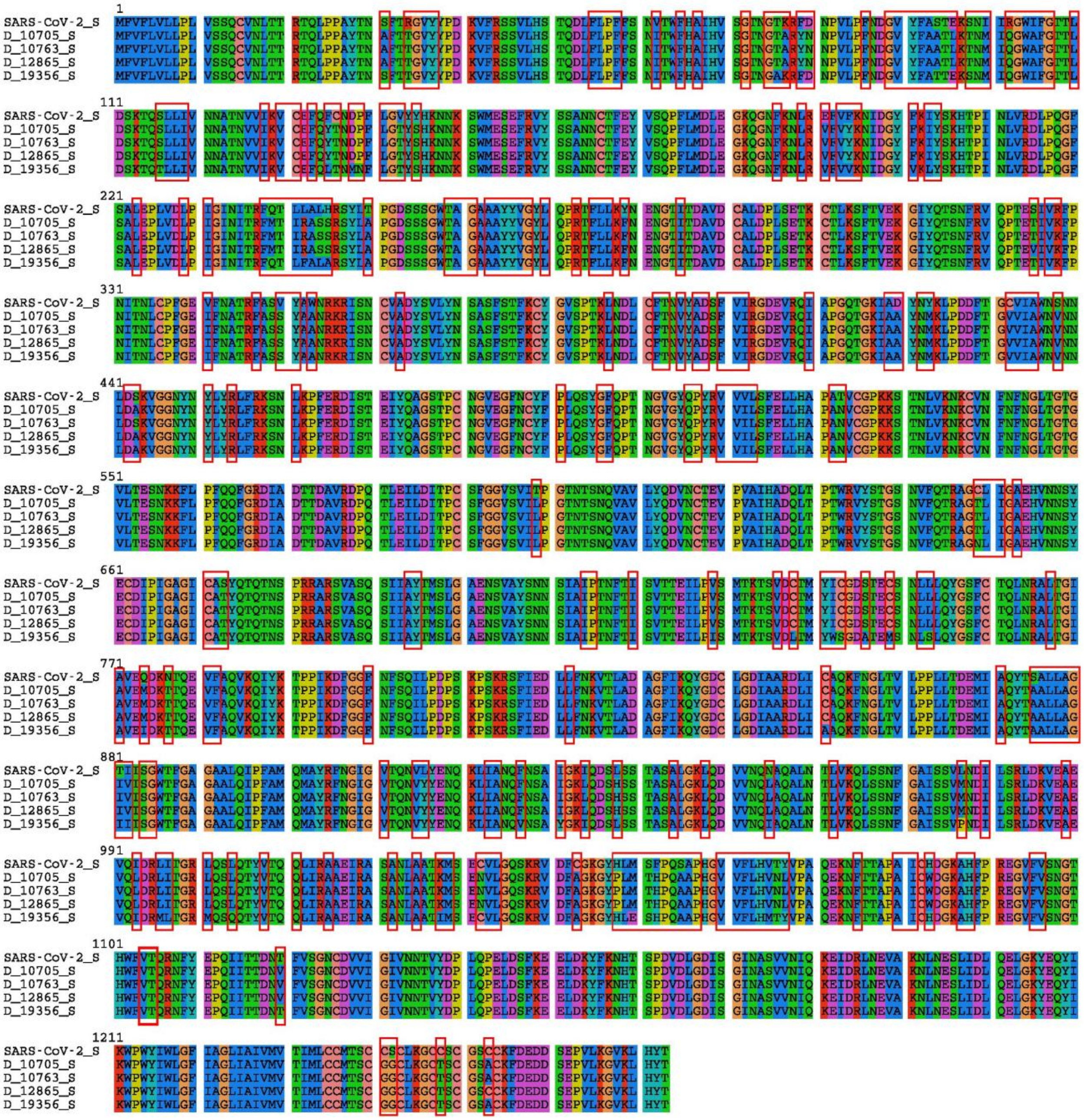


Figure 5-6 The multiple sequence alignment of the top four designed S proteins in comparison to the native SARS-CoV-2 S protein.

The four EvoDesign S proteins (Design-10705, 10763, 12865, and 19356) were selected based on their high structural similarity to the native S protein and promising immunogenic potential (in terms of promiscuous epitope count, ECS, and HSS scores). The solid red boxes highlight the core residues that were subjected to mutations by EvoDesign.

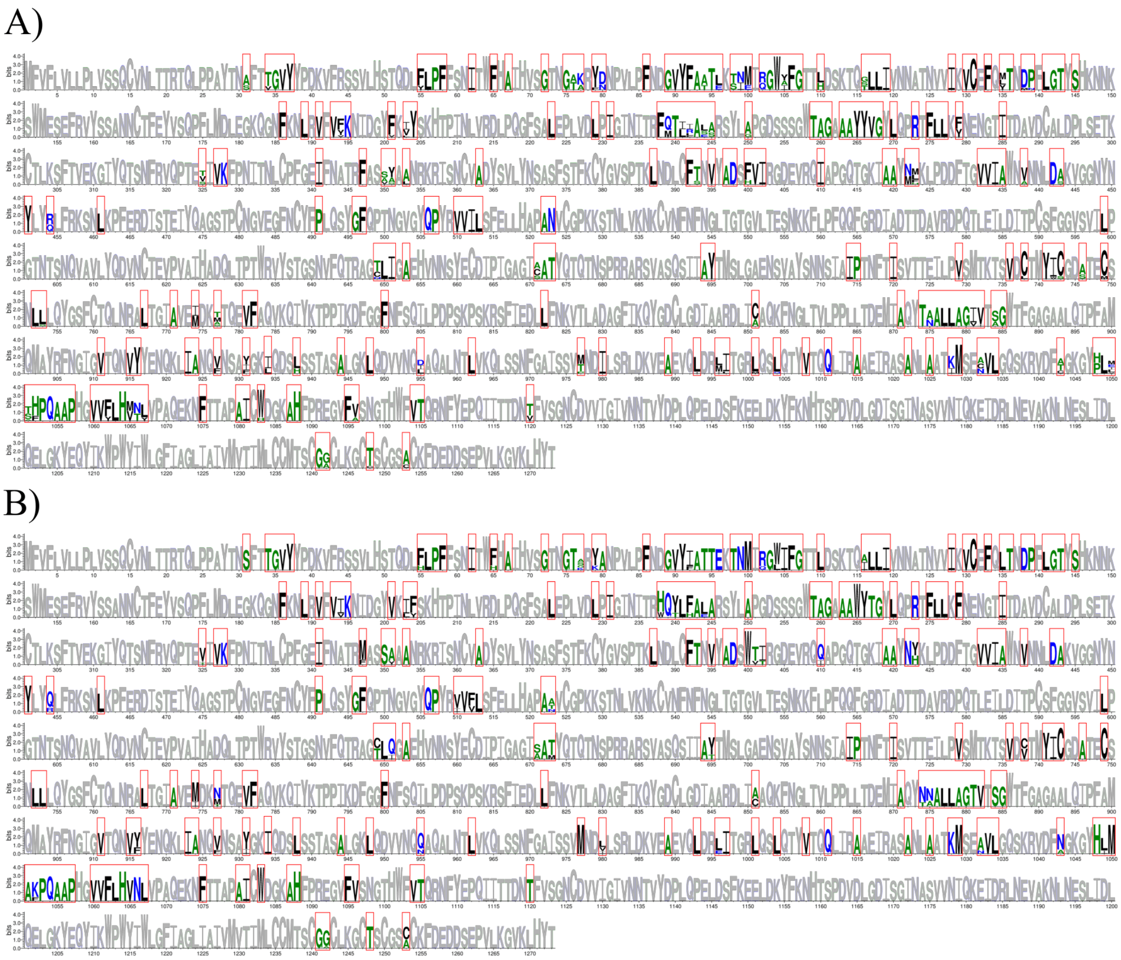


Figure 5-7 The sequence logo plot of (A) top 10 versus (B) worst 10 S protein designs. There were 12 core residues in both the top-scoring and worst-scoring designs substituted to the same amino acids in comparison to the native S protein. In particular, two core residues remained unmutated in the top-scoring designs but mutated in the worst-scoring designs (Y265W and V267T), suggesting mutations on these two residues might result in reduced immunogenicity.

Design-10705 was overall the best candidate with high structural similarity to the native S protein and good immunogenic potential (in terms of promiscuous epitope count, ECS, and HSS scores) amongst the top ten candidates. The candidate Design-10705 had a 93.9% sequence identity to the native S protein with a TM-score of 0.931 and an RMSD of 3.45 Å to the C-I-TASSER model of the native S protein. The homo-trimer 3D structure of Design-10705 was visualized and compared to the S protein C-I-TASSER and cryo-EM structural models (Figure 5-8). In terms of immunogenicity, it had the second-highest number of promiscuous epitopes. Table 5-5 shows the complete MHC-II T cell epitope profile of Design-10705. There were 32 predicted promiscuous epitopes in the native S protein (Table 5-6), and 31 of them were recovered in Design-10705. The two pre-existing immunity-related epitopes, V991-Q1005 and S816-D830, were both recovered in the new design. Besides these two epitopes, there were 19 epitopes identical to the native S protein epitopes, while ten epitopes had at least one mutation in Design-10705. Compared with the native S protein, the only missing MHC-II epitope in design 10705 was V911-N926, which was predicted to have reduced binding affinity to HLA-DRB1*03:01 and HLA-DRB4*01:01. Critically, this design introduced nine new MHC-II T cell promiscuous epitopes, which could potentially induce a stronger immune response with minimal perturbation compared to the native S protein.

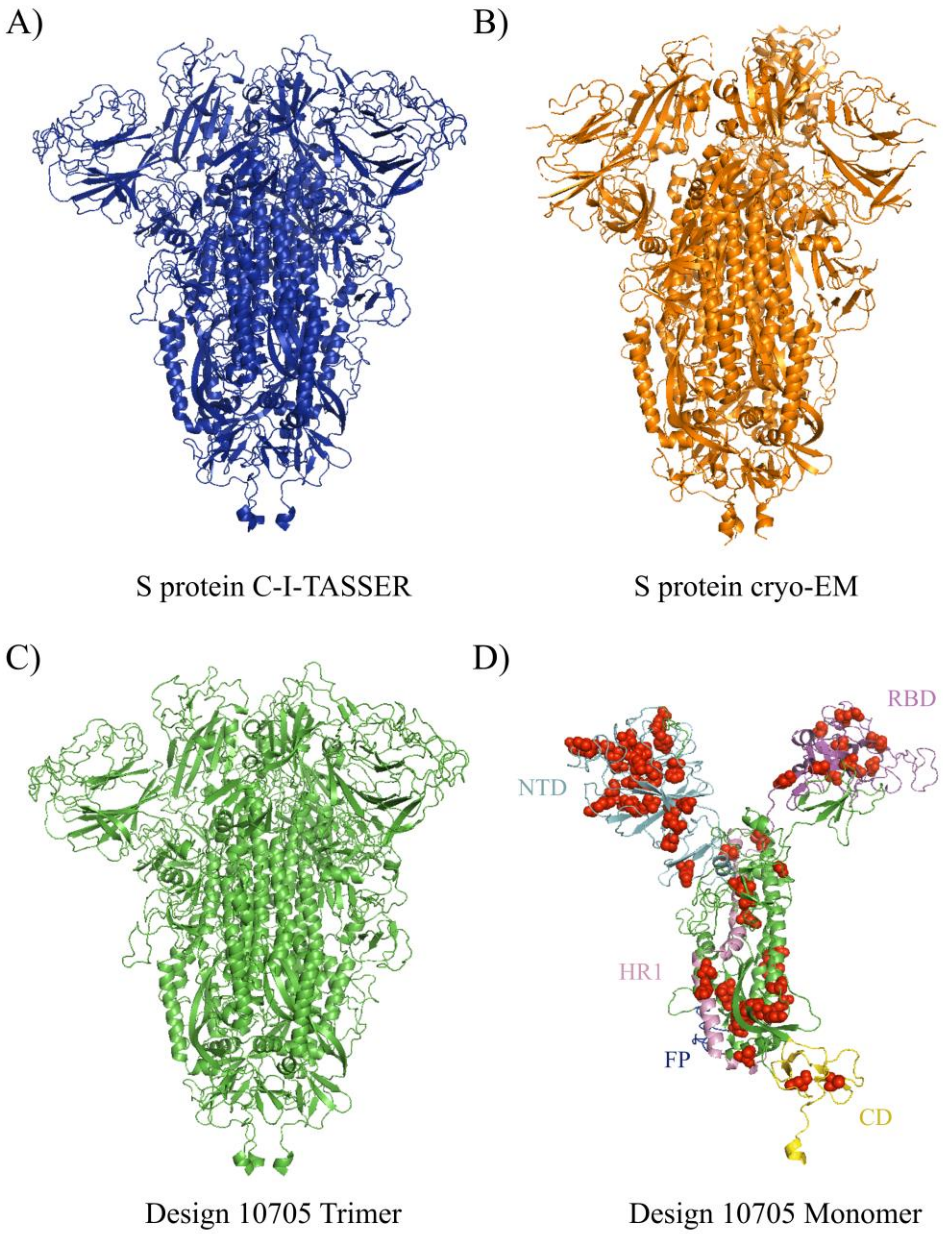


Figure 5-8 The 3D structures of A) C-I-TASSER S protein trimer, B) cryo-EM trimer, C) Design-10705 trimer, and D) Design-10705 monomer.

The ectodomain of Design-10705 was modeled using C-I-TASSER. Both the homo-trimer and monomer of Design-10705 are rendered. The NTD, RBD, HR1, FP, and CD domains are also highlighted in the Design-10705 monomer. The mutations introduced in Design-10705 are shown in red spheres.

Table 5-5 The predicted promiscuous MHC-II T cell epitopes of top EvoDesign S protein candidate.

Epitope	Epitope Comment	Start	End	Median Percentile Rank	Binding Alleles
VQLDRLITGRLQSLQ	Pre-existing Immunity Related Epitopes	991	1005	17	DRB1*03:01;DRB1*15:01;DRB4*01:01;DRB5*01:01
SFIEDLLFNKVTLAD		816	830	16	DRB1*03:01;DRB1*15:01;DRB3*01:01;DRB3*02:02;DRB4*01:01
VYYPDKVFRSSVLHS	Original S Protein Identical Epitopes	36	50	11	DRB1*03:01;DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB3*02:02
KVFRSSVLHSTQDLF		41	55	17	DRB1*07:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
SLLIVNNATNVVIKV		116	130	6.5	DRB1*03:01;DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB3*02:02;DRB4*01:01
EFRVYSSANNCTFEY		156	170	18	DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB3*02:02
FKIYSKHTPINLVRD		201	215	14	DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB5*01:01
SVLYNSASFSTFKCY		366	380	18	DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB3*02:02
YLYRLFRKSNLKPFE		451	465	5.7	DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
SIIAYTMSLGAENSV		691	705	4.7	DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
YGSFCTQLNRALTGI		756	770	19	DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB5*01:01
LLFNKVTLADAGFIK		821	835	17	DRB1*03:01;DRB1*07:01;DRB3*01:01;DRB3*02:02
CAQKFNGLTVLPPLL		851	865	19	DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB5*01:01
GAALQIPFAMQMAYR		891	905	18	DRB1*07:01;DRB1*15:01;DRB4*01:01;DRB5*01:01
IPFAMQMAYRFNGIG		896	910	3.7	DRB1*03:01;DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
QMAYRFNGIGVTQNV		901	915	19	DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB3*02:02
TLVKQLSSNFGAISS		961	975	14	DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB4*01:01
TYVTQQLIRAAEIRA		1006	1020	20	DRB1*07:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
QLIRAAEIRASANLA		1011	1025	12	DRB1*03:01;DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
AEIRASANLAATKMS		1016	1030	7.9	DRB1*03:01;DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
REGVFVSNGTHWFVT		1091	1105	9.4	DRB1*03:01;DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB3*02:02;DRB5*01:01
LPFFSNITWFHAIHV	Original S protein Mutated Epitopes	56	70	7.1	DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB3*02:02;DRB5*01:01
VFVYKNIDGYFKIYS		191	205	13	DRB1*03:01;DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB3*02:02;DRB5*01:01
IGINITRFMTIRASS		231	245	6.2	DRB1*03:01;DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB4*01:01;DRB5*01:01

TRFMTIRASSRSYLA	New Epitopes	236	250	1.2	DRB1*03:01;DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
YVGYLQPRTFLLKFN		266	280	12	DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
SNFRVQPTETIVKFP		316	330	14	DRB1*07:01;DRB3*01:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
IFNATRFASSYAANR		341	355	13	DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB5*01:01
RFASSYAANRKRISN		346	360	17	DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB5*01:01
VILSFELLHAPANVC		511	525	14	DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
KLIANQFNSAIGKLQ		921	935	17	DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
NITWFHAIHVSGTNG		61	75	20	DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB5*01:01
FNDGVYFAATLKTNM		86	100	14	DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB5*01:01
GKQGNFKNLRVFVYK		181	195	13	DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
LVDLPIGINITRFMT		226	240	20	DRB1*03:01;DRB3*01:01;DRB3*02:02;DRB4*01:01
GVVIAWNVNNLDAKV		431	445	11	DRB1*03:01;DRB1*15:01;DRB3*01:01;DRB3*02:02;DRB4*01:01
TDEMIAQYTAALLAG		866	880	19	DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB4*01:01
VVNQLAQALNTLVKQ		951	965	19	DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB4*01:01
GAISSVMNDILSRLD		971	985	20	DRB1*03:01;DRB3*01:01;DRB3*02:02;DRB4*01:01
VFLHVNLVPAQEKNF		1061	1075	16	DRB1*03:01;DRB1*07:01;DRB3*02:02;DRB4*01:01;DRB5*01:01

Table 5-6 The predicted MHC-II T cell promiscuous epitopes of the native SARS-CoV-2 S protein.

Epitope	Start	End	Median Percentile Rank	MHC-II Alleles
SLLIVNNATNVVIKV	116	130	6.5	DRB1*03:01;DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB3*02:02;DRB4*01:01
TRFQTLLALHRSYLT	236	250	2.9	DRB1*03:01;DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
REGVFVSNGTHWFVT	1091	1105	9.4	DRB1*03:01;DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB3*02:02;DRB5*01:01
IGINITRFQTLLALH	231	245	17	DRB1*03:01;DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
RFASVYAWNRKRISN	346	360	8.2	DRB1*03:01;DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
QLIRAAEIRASANLA	1011	1025	12	DRB1*03:01;DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
IPFAMQMAYRFNGIG	896	910	3.7	DRB1*03:01;DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
VFNATRFASVYAWNR	341	355	9.3	DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB5*01:01
AEIRASANLAATKMS	1016	1030	7.9	DRB1*03:01;DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
QMAYRFNGIGVTQNV	901	915	19	DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB3*02:02
LPFFSNVTWFHAIHV	56	70	7.3	DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB3*02:02;DRB5*01:01
SVLYNSASFSTFKCY	366	380	18	DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB3*02:02
SIIAYTMSLGAENSV	691	705	4.7	DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
KLIANQFNSAIGKIQ	921	935	18	DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB4*01:01
SFIEDLLFNKVTLAD	816	830	16	DRB1*03:01;DRB1*15:01;DRB3*01:01;DRB3*02:02;DRB4*01:01
TLVKQLSSNFGAISS	961	975	14	DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB4*01:01
YLYRLFRKSNLKPFE	451	465	5.7	DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
SNFRVQPTESIVRFP	316	330	9.9	DRB1*03:01;DRB1*07:01;DRB3*01:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
VYYPDKVFRSSVLHS	36	50	11	DRB1*03:01;DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB3*02:02
EFVFKNIDGYFKIYS	191	205	13	DRB1*03:01;DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB3*02:02;DRB5*01:01
TYVTQQLIRAAEIRA	1006	1020	20	DRB1*07:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
FKIYSKHTPINLVRD	201	215	14	DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB5*01:01
EFRVYSSANNCTFEY	156	170	18	DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB3*02:02
YVGYLQPRTFLLKYN	266	280	13	DRB1*07:01;DRB1*15:01;DRB3*01:01;DRB4*01:01;DRB5*01:01
VQIDRLITGRLQSLQ	991	1005	15	DRB1*03:01;DRB1*15:01;DRB4*01:01;DRB5*01:01
GAALQIPFAMQMAYR	891	905	18	DRB1*07:01;DRB1*15:01;DRB4*01:01;DRB5*01:01
YGSFCTQLNRALTGI	756	770	19	DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB5*01:01
KVFRSSVLHSTQDLF	41	55	17	DRB1*07:01;DRB3*02:02;DRB4*01:01;DRB5*01:01
LLFNKVTLADAGFIK	821	835	17	DRB1*03:01;DRB1*07:01;DRB3*01:01;DRB3*02:02
CAQKFNGLTVLPPLL	851	865	19	DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB5*01:01
VTQNVLYENQKLIAN	911	925	16	DRB1*03:01;DRB1*15:01;DRB3*01:01;DRB3*02:02
VVLSFELLHAPATVC	511	525	19	DRB1*07:01;DRB1*15:01;DRB3*02:02;DRB4*01:01;DRB5*01:01

One concern is that the top design, Design-10705, might lose the desired structure and protein function due to reduced stability caused by redesigning the core regions. To examine this concern, 50-ns MD simulations were carried out to compare the stability and flexibility of Design-10705 and the native S. As shown in Figure 5-9 A, Design-10705 and the wildtype showed similar RMSD shifts for both the backbones and side-chains after convergence at about 30 ns. The root-mean-square fluctuation (RMSF) measurement showed that the two proteins exhibited similar fluctuating profiles and thus comparable flexibility (Figure 5-9 B). Moreover, the radius of gyration and solvent-accessible surface area (SASA) as a function of simulation time were also analyzed for the two proteins. Design-10705 showed a slightly smaller radius of gyration (Figure 5-9 C) and a smaller SASA than the wildtype S (Figure 5-9 D), indicating that Design-10705 had a slightly more compact structure. Taken together, Design-10705 is expected to be sufficiently stable with the desired biological function (e.g., increased immunogenicity).

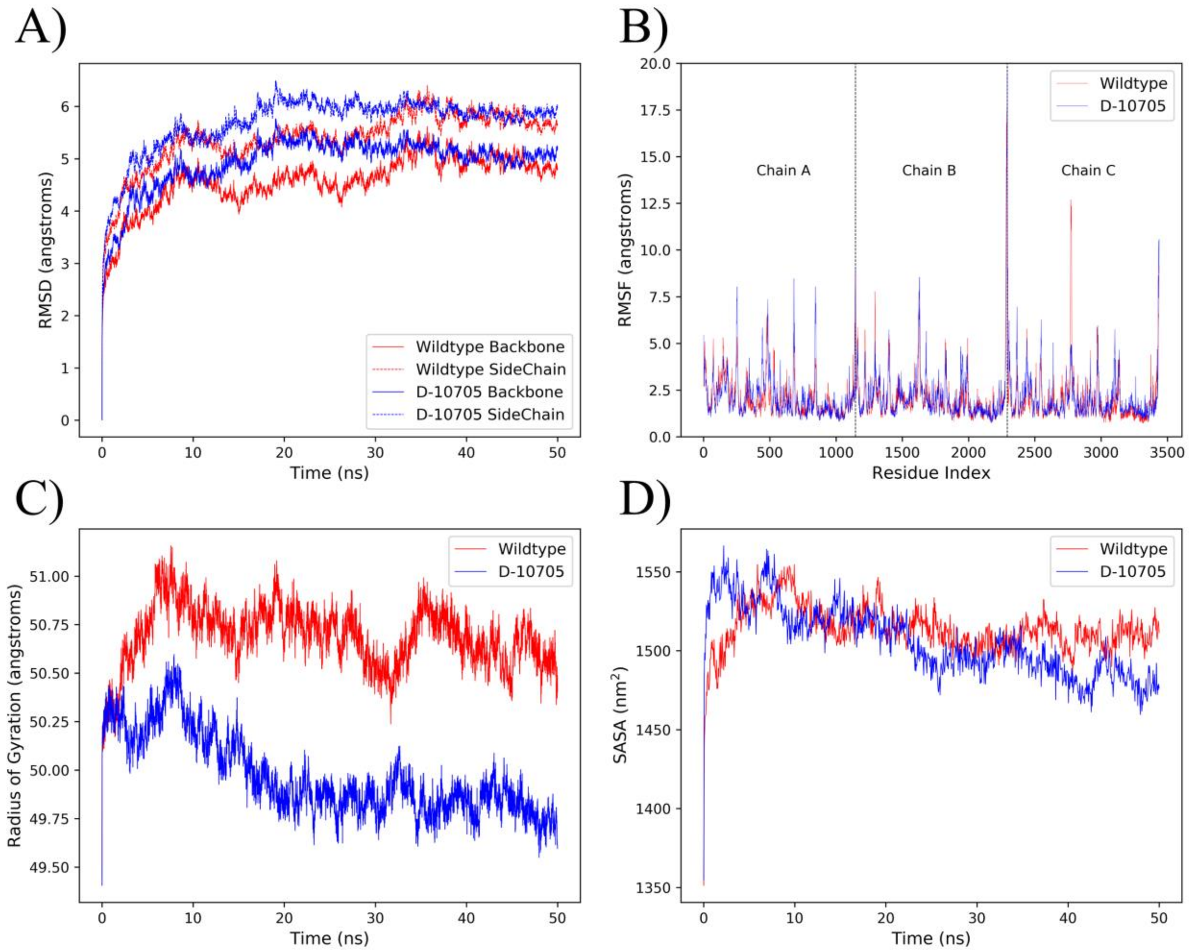


Figure 5-9 Analysis of molecular dynamics simulation results for the wildtype S and Design-10705.

Design-10705 is denoted as D-10705 in the plot. (A) Root-mean-square deviation (RMSD) for the wildtype S and D-10705 backbone and side-chains. (B) Root-mean-square fluctuation (RMSF) for all the residues in the trimeric protein. The three chains are separated by black dashed lines (Chain A: amino acids 1-1146; Chain B: 1147-2292; Chain C: 2293-3438). (C) The radius of gyration. (D) Solvent-accessible surface area (SASA).

5.5 Discussion

The subunit, DNA, and mRNA vaccines are typically considered to be safer but often induce weaker immune responses than the live-attenuated and inactivated vaccines. Although the addition of adjuvant or better vaccination strategies can compensate for the immunogenicity, the addition of new epitopes to the antigen provides an alternative way to induce stronger immune responses (Wada et al. 2017; Hewitt et al. 2019). During the protein design process, we applied design constraints so that the surface conformation, and in particular, B cell epitopes of the designed S protein variants were unchanged. For the designed S proteins with at least 5% of the core residues mutated, the immunogenicity potential of these candidates was evaluated and was structurally compared to the native S protein. The top candidate (Design-10705) recovered 31 out of 32 MHC-II promiscuous epitopes, and the two pre-existing immunity-related epitopes (V991-Q1005 and S816-D830) were present in the design. In addition to the 31 recovered epitopes, Design-10705 also introduced nine new MHC-II promiscuous epitopes with the potential to induce stronger CD4 T cell response. MD analysis of Design-10705 and the native S protein showed that the two proteins shared similar stability and flexibility (Fig 6). Overall, the newly designed S protein should preserve the native S protein's structure and function with enhanced immunogenicity.

The concept of manipulating epitopes to decrease the immunogenicity has been applied to therapeutic proteins. King at el. disrupted the MHC-II T cell epitopes in GFP and *Pseudomonas* exotoxin A using the Rosetta protein design protocol (King et al. 2014; Fleishman et al. 2011). The EpiSweep program was also applied to structurally redesign bacteriolytic enzyme lysostaphin as an anti-staphylococcal agent with reduced immunogenicity to the host (Blazanovic et al. 2015; Choi et al. 2017). In this study, a similar strategy, but to improve immunogenicity, was applied to redesign the SARS-CoV-2 S protein as an enhanced vaccine candidate; specifically, we aimed to increase immunogenicity by introducing more MHC-II T cell promiscuous epitopes to the protein without reducing the number of B cell epitopes.

The addition of epitopes to induce stronger immune responses has been previously applied to develop H7N9 vaccines. The H7N9 hemagglutinin (HA) vaccine-elicited non-neutralizing antibody responses in clinical trials (Mulligan et al. 2014; Guo et al. 2014). Rudenko et al. reported that there were fewer CD4 T cell epitopes found in H7N9 HA in comparison to the seasonal H1 and H3 HA proteins (Rudenko et al. 2016). Based on this finding, Wada et al. improved the H7N9 vaccine by introducing a known H3 immunogenic epitope to the H7 HA protein without perturbing its conformation, which resulted in an over 4-fold increase in the HA-binding antibody response (Wada et al. 2017). However, the number of epitopes is not the only factor that influences protective immunity. Studies have reported that CD8 T cell epitopes might induce regulatory T cell responses (Calis, de Boer, and Keşmir 2012; Frankild et al. 2008), and pathogens adapted to include CD4 and CD8 epitopes with high similarity to human peptides as a means to suppress host immunity for its survival (Leonard Moise et al. 2013). Therefore, we examined the significance of ECS and HSS in the context of mild versus severe forms of HCoV infection and then utilized these two scores to evaluate the designed S protein candidates.

The computational design of the SARS-CoV-2 S protein could be coupled with some other structural modifications for a more rational structure-based vaccine design. The present study aims to introduce new epitopes to the S protein while keeping the surface residues unchanged to minimize the structural change of the designed proteins, and according to the protein structure prediction results, the designed candidates were predicted to be structurally similar to the native S protein (Table 1 & Fig 5). The structural modifications performed on the native S protein, such as stabilizing the protein in its prefusion form (Bos et al. 2020), or fixing the RBD in the “up” or “down” state, could still be applied to the final candidate in this study. The combination of these structural vaccinology techniques into the current pipeline could further enhance the immunogenicity of the S protein as a vaccine target. However, a major limitation of the present study is the wet-lab experimental validation of the designed proteins. First, the newly designed protein sequences need to be folded properly with a structure comparable to that of the native S protein. Second, the capability of the newly added epitopes for binding MHC-II molecules and subsequently inducing immune responses needs to be validated. Finally, these candidates should be tested for their protectiveness and safety in animal models.

Overall, this study presents a strategy to improve the immunogenicity and antigenicity of a vaccine candidate by manipulating the MHC-II T cell epitopes through computational protein design. In the current settings, the immunogenicity evaluation was carried out after the standard protein design simulations with EvoDesign. In the future, the assessment of the immunogenic potential could be incorporated into the protein design process so that the sequence decoy generated at each step will be guided by balancing both the protein stability and immunogenicity. Moreover, with proper prior knowledge of known epitopes (e.g., both MHC-I and MHC-II from the pathogen proteome), it is also possible to create a chimeric protein, which integrates epitopes from antigens other than the target protein.

5.6 Acknowledgement

I appreciate the support of the National Institute of Allergy and Infectious Diseases (AI081062). The result presented in this chapter is published in (E. Ong et al. 2021). Yang Zhang and Yongqun He conceived and designed the project. Edison Ong (EO) and Xiaoqiang Huang performed studies on protein design and analyses. Robin Pearce participated in the discussion.

EO drafted the manuscript. All authors performed result interpretation, edited, and approved the manuscript.

Chapter 6 Ontology-based Vaccine Data Integration and Analysis

6.1 Abstract

Infectious diseases, acquired through pathogenic microbial agents, remain among the most common and fatal threats to human health throughout the world. The host-pathogen interaction (HPI) and vaccine-host interaction are the keys to understand the mechanisms of infectious diseases and vaccine protection, which require in-depth knowledge synthesized data capturing various aspects of vaccination. The Ontology of Host-Pathogen Interaction (OHPI), a community-based biomedical ontology in the domain of HPIs, was developed to integrate and analyze the virulence factors and protective antigens data stored in the Victors and Protegen databases. The Vaccine Investigation Ontology (VIO) was developed and applied to systematically classify the different variables and relations among these variables. VIO was used to integrate and analyze the differential gene expression and biological pathways from two Yellow Fever vaccines. Overall, the OHPI supports the knowledge representation and analysis of the interactions between host cells (or genes) and pathogen proteins serving as virulence factors or protective antigens. VIO standardizes the metadata types in vaccine investigation studies and the semantic relations among these metadata types. The combination of OHPI, VIO, and bioinformatics tools based on these ontologies provides a robust framework for integrative knowledge generation, modeling, and storage of the heterogeneous vaccine-related data, leading to a fundamental understanding of the underlying mechanisms of vaccine immunity.

6.2 Introduction

Infectious diseases, acquired through pathogenic microbial agents, remain among the most common and fatal threats to human health throughout the world. The World Health Organization estimated that infectious and parasitic diseases caused 9.31 million deaths in 2015, accounting for 16.5% of total global mortality (WHO 2016). There is still a critical need to develop more effective preventative and therapeutic measures against various infections. As one of the most significant inventions in modern medicine, vaccination has been used to efficiently protect humans against many infectious diseases and improve human health. Vaccines are also being developed against cancer (Schlom et al. 2010), allergy (Huggins and Looney 2004), and many other non-infectious diseases (Lynch and Mills 2012; Nicholas, Odumosu, and Langridge 2011). However, our efforts to develop vaccines to protect against diseases have not always been successful. The future success of effective vaccine development relies on a deep understanding of protective vaccine-induced immune mechanisms against different diseases. The protective mechanism can be better understood with a systematic analysis of high throughput data being generated in the vaccine domain.

It has been a considerable challenge to systematically, logically represent, and integrate various vaccine-related databases and study the underlying host-pathogen interaction (HPI) mechanism. Infectious disease is the result of an interactive relationship between a pathogen and its host, and the study of HPIs is crucial in understanding microbial pathogenesis and host immune mechanisms. Extracted from peer-reviewed publications, several databases, including PHIDIAS (Z. Xiang, Tian, and He 2007), PHISTO (Durmuş Tekir et al. 2013), and PHI-base (Urban et al. 2015), have been developed to store host-pathogen interaction data. Virulence factors (VFs) are the key elements of HPIs, which allow microbial pathogens to overcome host defense mechanisms and cause diseases in the host. As the central part of the PHIDIAS (Z. Xiang, Tian, and He 2007), Victors (http://www.phidias.us/victors) is a specific database comprised of genes experimentally observed to be necessary for virulence (Sayers et al. 2019). The Victors database includes over 5,000 VFs from different bacteria, viruses, parasites, and fungi, which are pathogenic to animals and humans. Host vaccine-induced immune factors (vaximmutor) are also annotated and curated in the VaximmutorDB database within the VIOLIN system (http://www.violinet.org/vaximmutordb, to be submitted). Many VFs have been proven to be useful protective antigens (PAgs). Protegen (http://www.violinet.org/protegen) is a database that stores over 1,000 PAgs (B. Yang et al. 2011; He and Xiang 2012). By comparing Victors VFs and Protegen PAgs, we were able to identify VFs that are also PAgs used in different vaccines. Bioinformatics analyses of both the Victors and Protegen reveal unique and overlapping biological properties between the VFs and PAgs (Sayers et al. 2019). Systematic identification and analysis of the VFs, PAgs, and vaximmutors would enhance our understanding of how HPIs are involved in the host protective immunity and develop new measures against infectious diseases. These databases have rich but sparse information of HPIs focusing on different aspects, and ontology can be used to integrate the available HPI data better and enhance the understanding of vaccine mechanisms.

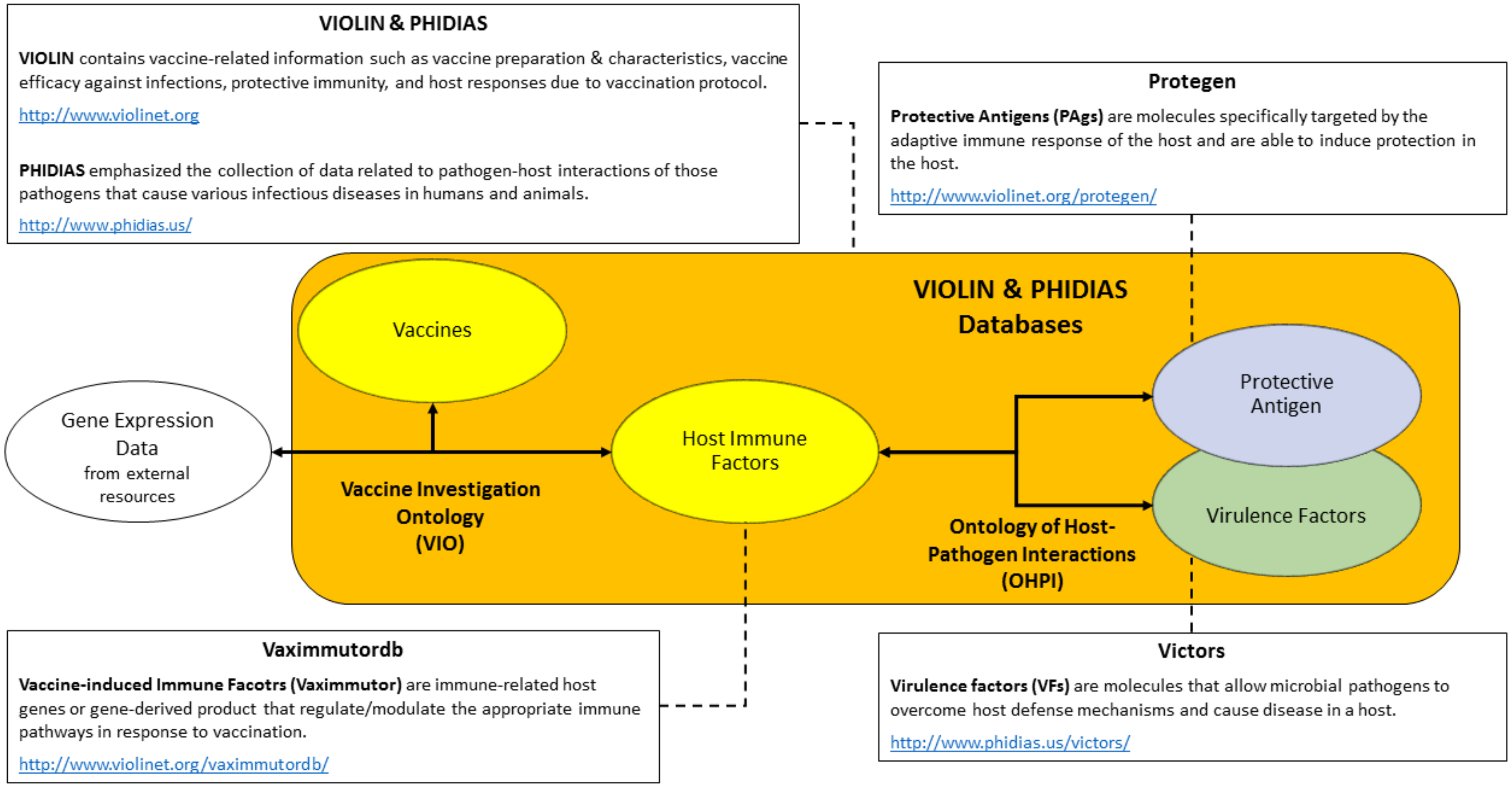
Another bottleneck in high-throughput vaccine-host interaction studies is that inconsistent experimental results were frequently generated even with similar experimental designs. A typical example is a gene-level host immune responses induced by the live attenuated Yellow Fever vaccine 17D (YF-17D) from various gene expression studies. The live attenuated YF-17D (Theiler and Smith 1937) and the sub-strains derived from the original 17D strain (Gardner and Ryman 2010) are widely used for vaccination against Yellow Fever infections.

These vaccine strains can induce strong and effective protective immune responses in vaccinated humans (Pulendran 2009; Roukens and Visser 2008). As a result, YF-17D has become an excellent model to study general host responses induced by vaccinations, and many differentially expressed genes have been reported in YF-17D-vaccinated human subjects. However, these studies reported different results even though similar experimental designs were used. For example, three studies used human subjects who were all vaccinated with YF-17D or YF-VAX (made with a specific YF-17D strain) but generated overlapping but a quite different gene expression profiles (Gaucher et al. 2008; Querec et al. 2009; Scherer et al. 2007).

To address the above-mentioned challenges, the ontology provides a feasible and robust way to integrate heterogeneous vaccine-related data and standardize experimental conditions. Ontology offers an ideal platform to properly and robustly solve the critical issue of different but overlapping results from studies on the same scientific question. Basically, ontology standardizes the representation of entities and relations among entities in a specific domain using human- and computer-interpretable format. Such standardization is important since experimental studies are often reported using inconsistent vocabulary and incomplete representation, often resulting in non-reproducible outcomes. The ontology usage can solve the issues in the standardized experimental and data representation from different studies. Given the nature of ontology, such standardization can also be understood by computers and so useful for data sharing. In addition to standardization, ontology also provides a hierarchical structure and logical relations among different entities, supporting advanced reasoning and data analysis. Several vaccine investigation-related ontologies exist. The Vaccine Ontology (VO) represents vaccine-related entities, such as vaccines, vaccine components, vaccinations, host responses to vaccines, and the relations among these entities (He et al. 2009; Özgür et al. 2011; Y. Lin and He 2012). The

Ontology of Biological and Clinical Statistics (OBCS) is a community-based ontology of statistics in the biological and clinical domains (J. Zheng et al. 2016). The community-based Ontology for Biomedical Investigations (OBI) targets to represent various biomedical investigation components shared by different biomedical communities (Brinkman et al. 2010).

In this chapter, two ontologies were developed and applied to standardize, integrate, and analyze host-pathogen interaction knowledge and vaccine-host interaction investigation data (Figure 6-1). For the host-pathogen interaction study, the Ontology of Host-Pathogen Interaction (OHPI) was introduced to model related HPI information from the Victors VFs database (Sayers et al. 2019) and the Protegen PAgs database (B. Yang et al. 2011). For the vaccine-host study, the Vaccine Investigation Ontology (VIO) was first developed to classify different variables and the relations among these variables in the vaccine investigation studies. Then VIO was applied to standardize and analyze the host responses induced by the Yellow Fever vaccine YF-17D and its sub-strains in two studies (Gaucher et al. 2008; Querec et al. 2009). Overall, the VIO and OHPI were designed to support advanced knowledge representation, integration, sharing, and analysis among these vaccine databases.


VIOLIN & PHIDIAS

VIOLIN contains vaccine-related information such as vaccine preparation & characteristics, vaccine efficacy against infections, protective immunity, and host responses due to vaccination protocol.

http://www.violinet.org

PHIDIAS emphasized the collection of data related to pathogen-host interactions of those pathogens that cause various infectious diseases in humans and animals.

http://www.phidias.us/

Protegen

Protective Antigens (PAgs) are molecules specifically targeted by the adaptive immune response of the host and are able to induce protection in the host.

http://www.violinet.org/protegen/

Vaximmutordb

Vaccine-induced Immune Facotrs (Vaximmutor) are immune-related host genes or gene-derived product that regulate/modulate the appropriate immune pathways in response to vaccination.

http://www.violinet.org/vaximmutordb/

Victors

Virulence factors (VFs) are molecules that allow microbial pathogens to overcome host defense mechanisms and cause disease in a host.

http://www.phidias.us/victors/

Figure 6-1 Ontology-based framework to integrate vaccine-related data for vaccine-host and host-pathogen interaction studies.

6.3 Methods

6.3.1 OHPI ontology development

The OHPI development followed the OBO Foundry ontology development principles (Smith et al. 2007) and the eXtensible Ontology Development (XOD) strategy (He, Xiang, et al. 2018). The OHPI development started with the XOD1 and XOD2 principles, which reused, imported, and aligned terms and relations from existing ontologies. There were two ontology tools used to import existing ontologies into OHPI. First, the ODK was used to generate import files for Basic Formal Ontology (BFO) (Arp, Smith, and Spear 2016), Ontology of Biomedical Investigations (OBI) (Bandrowski et al. 2016), Chemical Entities of Biological Interest (ChEBI) (Hastings et al. 2013), Protein Ontology (PR) (Natale et al. 2014), Disease Ontology (DOID) (Kibbe et al. 2015), Uberon multi-species anatomy ontology (UBERON) (Mungall et al. 2012), Ontology of General Medical Science (OGMS) ("The Ontology for General Medical Science (OGMS)" n.d.), Gene Ontology (GO) (Blake et al. 2015), Information Artefact Ontology (IAO) ("The Information Artifact Ontology (IAO)" n.d.), and Relation Ontology (RO) (Smith et al. 2005). These ontologies were then imported using ODK for its capability to automatically check new versions of the importing ontologies whenever a new OHPI release was built.

Besides using ODK to import the above-mentioned ontologies, we also used Ontofox that supports additional features such as the inclusion of all the children terms and the hierarchy extraction with computed intermediates (Z. Xiang et al. 2010). Specifically, the import files were generated by submitting a CURL request to the Ontofox web service for the following ontologies: Cell Ontology (CL) (Diehl et al. 2016), Cell Line Ontology (CLO) (Sarntivijai et al. 2014), Interaction Network Ontology (INO) (Hur et al. 2015), Infectious Disease Ontology (IDO) (Cowell and Smith 2010), Brucellosis Ontology (IDOBRU) (Y. Lin, Xiang, and He 2011), NCBI organismal classification (NCBITaxon) (Federhen 2012), Vaccine Ontology (VO) (He et al. 2009) and Ontology of Genes and Genomes (OGG) (He, Liu, and Zhao 2014). Currently, the update of the importing ontologies' versions using Ontofox relies on the Ontofox internal Virtuoso database. In other words, the importing ontologies might not be up-to-date when a new OHPI release is built, depending on the update frequency of the Ontofox database. Therefore, the existing ontologies in OHPI were imported and aligned to the top-level ontology BFO by both ODK and Ontofox. Additionally, all imported ontologies were selected to reuse existing ontological entities that are relevant to the OHPI. For example, PR and OGG were imported for the virulence factor genes and proteins. NCBITaxon, DOID were imported to include pathogenic organisms and their associated diseases.

In addition to reusing existing ontologies, OHPI also applied the XOD3 design pattern strategy to modify existing and add new ontology terms and relations automatically. The VFs and their related information, including NCBI Gene identifiers, PubMed references, host and pathogen NCBITaxonomy identifiers, and HPIs, were extracted and downloaded from the Victors database (Sayers et al. 2019). The PAgs and their corresponding experimentally verified vaccines were also extracted and downloaded from the Protegen database (B. Yang et al. 2011). All of these data were stored in tabular format and modeled in OHPI using Ontorat (Z. Xiang et al. 2015) with the design pattern defined in Figure 3. New terms that were not related to HPIs were assigned new identifiers using the prefix “OHPI_” followed by automatically incremented seven-digit numbers starting from one. The HPI terms were also assigned new identifiers using the prefix “OHPI_” and automatically incremented seven-digit numbers starting from “9000001”.

The Protégé OWL editor (http://protege.stanford.edu/) was used for the manual term editing and verification. OHPI was quality checked for consistency and integrity using the ODK SPARQL queries. The final OHPI was then built using the ODK. Different sub-versions were provided, including ohpi-base, ohpi-merge, and ohpi-full in three different formats OWL, OBO, and JSON. The ohpi-base only included terms, relations, and annotation defined by the OHPI and removed all the import ontologies. The ohpi-merge merged all the import ontologies into one single file. The ohpi-full merged and also used ELK reasoner (Bail et al. 2013) to infer relationships in OHPI.

6.3.2 VIO ontology development

As an extension of the Vaccine Ontology (VO) (He et al. 2009), the Vaccine Investigation Ontology (VIO) was developed by following the eXtensible Ontology Development (XOD) principles (He, Xiang, et al. 2018). Specifically, a list of vaccine investigation-related terms available in VO was initially identified. Ontofox (Z. Xiang et al. 2010) was then used to extract this list of terms and other relevant information (including logical axioms and annotations) from VO and imported into VIO. Additionally, many OBCS and OBI terms related to vaccine investigation were also imported into VIO using Ontofox. Since VO, OBCS, and OBI all follow the Open Biomedical Ontology (OBO) Foundry ontology development principles (Smith et al. 2007) and use the same upper-level ontology, Basic Formal Ontology (BFO) (Arp, Smith, and Spear 2016), these terms coming from different ontologies were efficiently and seamlessly aligned to each other in VIO. The resulting VIO was manually edited and checked using the Protégé OWL editor.

The microarray data sets reported in Gaucher et al. 2008 and Querec et al. 2009 are available through the GEO (Barrett et al. 2013) under series accession numbers GSE13699 and

GSE13486, respectively. The raw data set from Scherer et al. 2007 was not available from GEO or the paper supplemental material files and was excluded from this study. GEO2R (Barrett et al. 2013) was used to analyze the two microarray datasets as reported in Gaucher et al. 2008 and Querec et al. 2009. In brief, GEO2R applies log2 transformation if the expression values of the given GEO dataset are not in log space, and then performs differential expression analysis using Linear Models for Microarray Analysis (LIMMA) (Smyth 2004). The resulting p-values were adjusted for multiple comparisons using the false discovery rate (FDR). The GEO2R results for the two microarray datasets were exported and compared for overlapping using a Venn diagram. The same cut-off (adjusted p-value based on FDR < 0.05 and log fold change less than -1.3 fold or greater than 1.3 fold) for identifying significant results was applied. For the gene-level comparison, gene symbols were updated to official gene symbols using the DAVID Gene ID Conversion Tool (https://david.ncifcrf.gov/conversion.jsp) (D. W. Huang, Lempicki, and Sherman 2009). All genes analyzed in this study were mapped to their corresponding Entrez Gene IDs using the DAVID Gene ID Conversion. The Gene Ontology (GO) and pathway enrichment analyses of the original study were performed based on the original list of differentially expressed genes. The DAVID bioinformatics resources (D. W. Huang, Lempicki, and Sherman 2009) was used to analyze the similarities and differences of different GO terms enriched in the original analysis or the standardized re-analysis of the two microarray datasets.

The performance of the standardized re-analysis was estimated by the identification of shared significant GO biological processes between the two microarray datasets. The hierarchical structure of significantly enriched GO terms and their related ancestor terms were also visualized and analyzed using GOfox (http://gofox.hegroup.org) (E. Ong and He 2015). By integrating and extending features from two popular ontology programs, Ontofox (Z. Xiang et al. 2010) and

Ontobee (E. Ong et al. 2017), the GOfox web program is able to generate full or simplified hierarchical GO subsets to classify and display enriched GO terms and their ancestor terms. By considering the multiple inheritances of GO, the GOfox includes a simplified hierarchical classification method that outputs a GO hierarchical structure among enriched GO terms and their minimal upper-level ancestor terms in a user-friendly interactive visualization scheme. In addition, we also used the Reactome pathway analysis tool (Fabregat et al. 2016) to analyze enriched pathways in the Reactome pathway knowledgebase. Both GO biological processes, and Reactome pathway enrichment analysis applied adjusted p-value based on FDR < 0.05 as the significance cut-offs.

6.4 Results

6.4.1 Modeling Host-Pathogen Interactions with OHPI

OHPI was developed as a biomedical ontology to support the data representation, integration, and analysis of the VFs, HPIs, and PAgs stored in the Victors and Protegen databases. OHPI reuses a subset of virulence factor genes from the OGG ontology (He, Liu, and Zhao 2014), and all of them were assigned the role 'virulence factor gene role' (OHPI_0000089) and annotated with experimental evidence from at least one peer-reviewed article. Each VF gene was linked to a pathogen organism via the OHPI "gene as virulence factor in pathogen" (OHPI_0000003) object property (Figure 6-2). This object property represents a relation between a gene and an organism, where the gene is a virulence factor, and the organism is a pathogen, and the mutant of the gene for the pathogen is attenuated in the host. Each VF gene was also linked to at least one host organism, cell, or cell line cell via the OHPI 'gene mutant attenuated in host' (OHPI_0000007) object property or its descendants. These object properties represent relations between a gene and a host organism, cell, or cell line cell where the microbial mutant lacking the gene is attenuated in the host compared to the wild-type microbe.

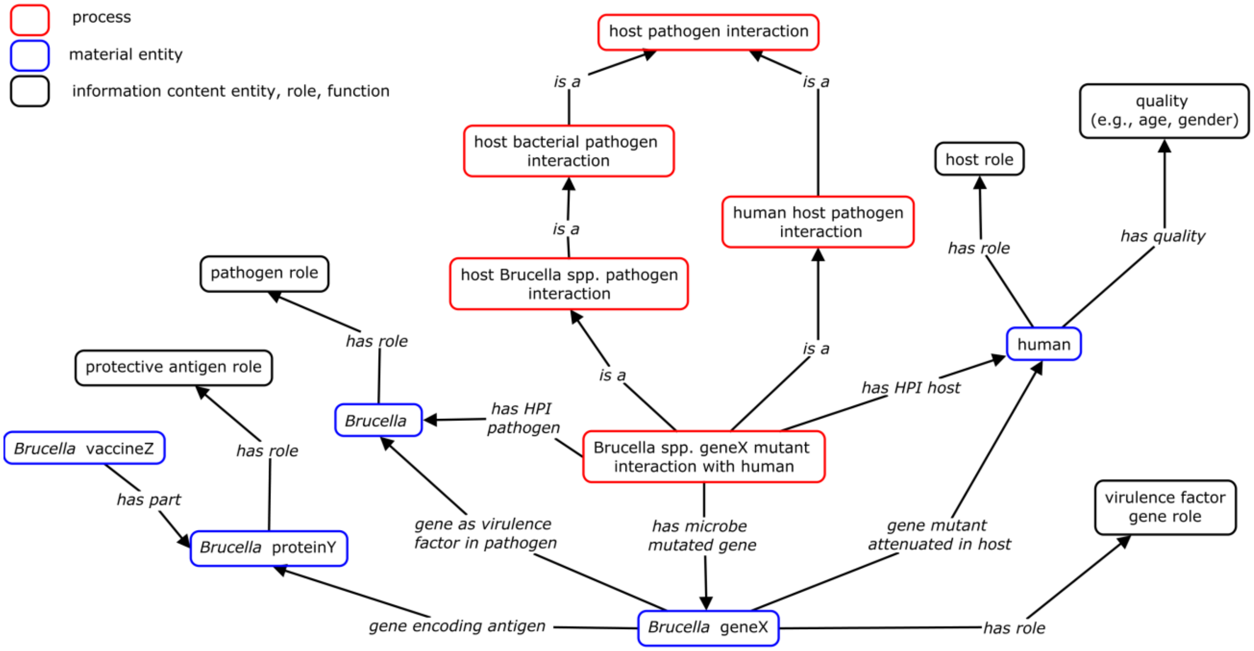

process
material entity
information content entity, role, function
host pathogen interaction
is a
is a
host bacterial pathogen interaction
quality (e.g., age, gender)
host role
is a
human host pathogen interaction
has role
has quality
pathogen role
host Brucella spp. pathogen interaction
has role
protective antigen role
is a
is a
human
has HPI host
Brucella
has HPI pathogen
Brucella vaccineZ
has role
Brucella spp. geneX mutant interaction with human
has part
virulence factor gene role
gene as virulence factor in pathogen
has microbe mutated gene
gene mutant attenuated in host
Brucella proteinY
gene encoding antigen
Brucella geneX
has role

Figure 6-2 Ontology of Host-Pathogen Interaction design pattern.

This example used human (host) and *Brucella* (pathogen) to illustrate the design pattern modeling the host-pathogen interaction available in the OHPI. Besides the virulence factor role and the host-pathogen interaction relations, OHPI also models the protective antigen role and its associating vaccine.

The HPIs extracted from the Victors database were also modeled and represented in OHPI (Figure 6-2). There are two variables in the ‘host pathogen interaction’ (OHPI_0000001) class: host and pathogen. In OHPI modeling, these two variables were treated independently to generate the top-level HPI branches. The host branch includes classes generated by varying the host organisms, such as ‘human host pathogen interaction’ (OHPI_0000100) and ‘mouse host pathogen interaction’ (OHPI_0000103). This branch of HPIs can be further expanded from organism level to the cell or cell line cell, such as ‘human cell host pathogen interaction’ (OHPI_0000101) and ‘human cell line cell host pathogen interaction’ (OHPI_0000102). On the other, the pathogen branch contains classes generated by varying the pathogen organisms into ‘host bacterial pathogen interaction’ (OHPI_0000006), ‘host viral pathogen interaction’ (OHPI_0000010), ‘host parasitic pathogen interaction’ (OHPI_0000011), and ‘host fungal pathogen interaction’ (OHPI_0000012). Then the pathogen branch is further expanded into individual species such as ‘host *Brucella* spp. pathogen interaction’ (OHPI_0000014), ‘host M. tuberculosis interaction’ (OHPI_0000020), and ‘host Influenza virus pathogen interaction’ (OHPI_0000069).

With the top-level host pathogen interaction hierarchy defined as described above, all HPIs stored in the Victors database were automatically generated with design patterns with the appropriate hosts (organism, cell, or cell line cell) and pathogen organisms. For example, the interaction ‘*Brucella* spp. *virB9* mutant interaction with human HeLa cell’ (OHPI_9000744) had two asserted axioms:

‘has microbe mutated gene’ some ‘BRA0061’

‘has HPI host cell line cell’ some ‘HeLa cell’

The gene ‘BRA0061’ (OGG_3001164498) is a VF gene in *Brucella suis* 1330, as defined in Victors and OGG, respectively. The ‘HeLa cell’ (CLO_0003684) is an immortalized human epithelial cell. Therefore, during the OHPI building process, all these Victors host pathogen interactions were inferred automatically by the reasoner as the child term of both ‘human cell line cell host pathogen interaction’ (OHPI_0000102) and ‘host *Brucella* spp. pathogen interaction’ (OHPI_0000014).

In addition to the VFs and their related information from the Victors database, we also extracted PAgs, which are annotated as VFs at the same time, from the Protegen database. Since the VFs in Victors are annotated as genes while the PAgs are annotated as proteins, such relations were model using the OHPI object property ‘gene encoding antigen’ (OHPI_0000090). For example, the VF gene ‘fimH’ (OGG_3000948847) from *Escherichia coli* str. K-12 substr. MG1655 had the following axiom:

‘gene encoding antigen’ some ‘FimH’

The ‘FimH’ (VO_0010987) is the protective antigen protein and, at the same time, linked to a research vaccine ‘*E. coli* FimH with CFA and then IFA’ (VO_0001168) via an axiom:

‘has part’ some ‘FimH’

Through the modeling of both Victors and Protegen databases, OHPI represents 4,428 VFs and 2,063 host pathogen interactions from 82 pathogens with experimental evidence tested on nine host organisms stored in the Victors database (Table 6-1). Among the VFs, 52 were also encoded protective antigen proteins tested in 17 research vaccines, and that the mutants of these VFs become less virulence inside a host organism or host cells. The source code of OHPI is available on the GitHub website: https://github.com/OHPI/ohpi.

Table 6-1 Ontology of Host-Pathogen Interaction and vaccine-related statistics.

	Bacteria	**Virus**	**Parasite**	**fungi**
# Pathogens	47	23	7	5
VFs	4,127	57	21	223
HPIs	2,027	36	0	0
PAgs	42	9	1	0
# Vaccines	16	1	0	0

Note: Only bacterial and viral HPIs were included in the current OHPI ontology.

6.4.2 Modeling Vaccine Investigation Data with VIO

VIO focuses on the vaccine investigation, especially on defining and standardizing metadata types in various vaccine investigation studies. Most variables in the three Yellow Fever studies (Gaucher et al. 2008; Querec et al. 2009; Scherer et al. 2007) were modeled in the VIO design pattern and standardized in the data re-analysis process pipeline. These variables included data transformation method, human genome annotation version, significant gene identification method such as LIMMA, LIMMA version, and GO version used for GO enrichment analysis. To a certain extent, studies with different experimental settings can be considered as permutations to the host immune system and can be used to better understand the immune response mechanisms induced by the vaccine immunization. Therefore, controlling these experimental conditions is not necessary to understand the contributions of different variables to the final observed immune response outcomes. Instead, we can carefully dissect and identify the similarities and dissimilarity among these variables from different experimental studies. VIO was then applied to the re-analysis of two Yellow Fever vaccine studies with controlled conditions (Gaucher et al. 2008; Querec et al. 2009).

When integrating the two VF-Vax vaccine studies (Gaucher et al. 2008 and Querec et al. 2009) with the conditions defined by VIO (Figure 6-3), there were 554 and 126 significantly differentiated genes in the Gaucher and Querec, respectively (Figure 6-4 A). When comparing these two significant gene lists, there were 465, 89, 37 genes found to be unique in Gaucher, shared by both studies, and unique in Querec, respectively. When summarizing the genes to GO biological processes, our re-analysis identified more consistency between the two studies. When comparing the reported GO terms in the two original studies (Gaucher et al. 2008; Querec et al. 2009), only four enriched GO biological process terms were shared, and twenty terms were found to be different. However, our re-analysis standardized by the VIO modeling had seven enriched GO biological process terms being shared by the two studies (Figure 6-4 B) and provided more consistent GO enrichment results.

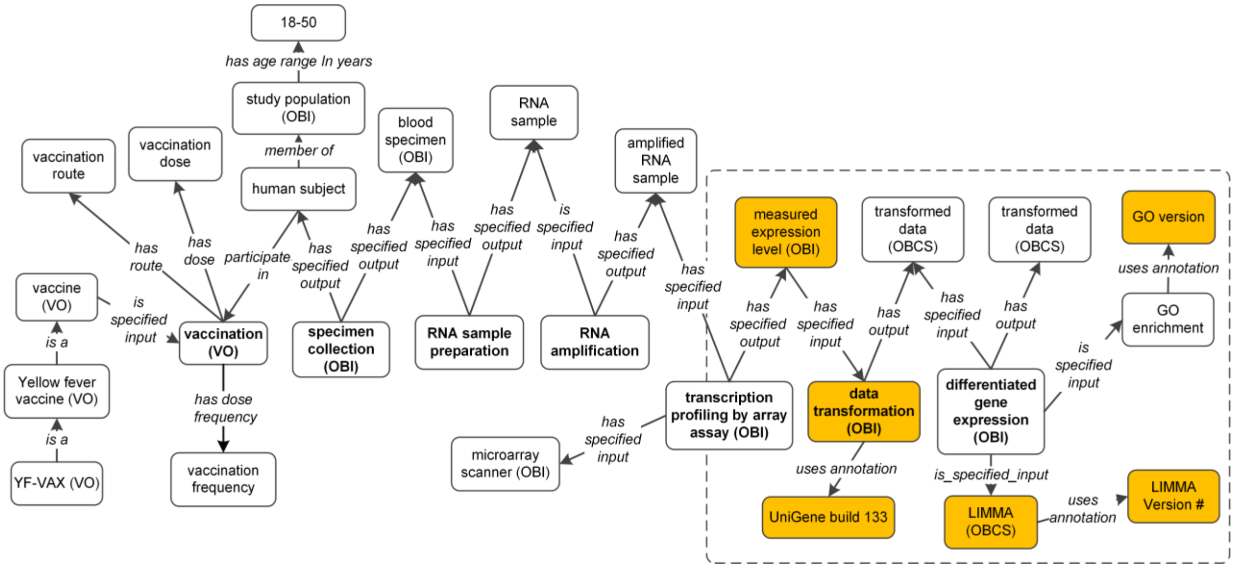


Figure 6-3 Vaccine Investigation Ontology design pattern to model the YF-VAX vaccination studies.

The boxed section includes different components that are related to data processing and analyses.

The brown-colored boxes are examples of variables changeable in our data re-analysis.

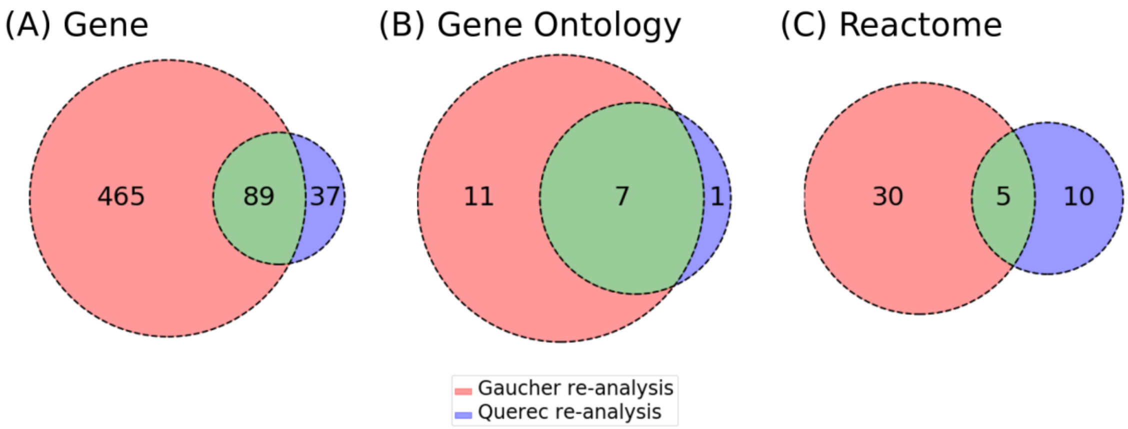
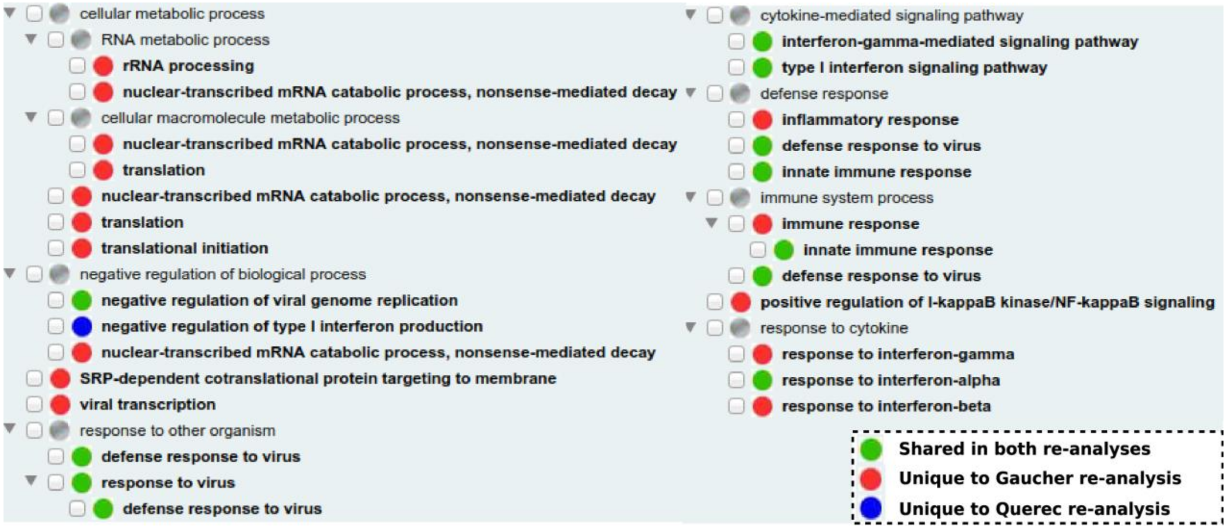


Figure 6-4 Comparison of the significant (A) genes, (B) biological processes, and (C) Reactome pathways in the re-analysis of YF-Vax studies.

Venn diagram illustrating the comparison of significant (adjusted p-value based on FDR < 0.05) (A) differentially expressed genes, (B) Gene Ontology biological process terms, (C) Reactome pathways between the re-analysis of the gene expression profile of VF-Vax vaccination from Gaucher et al. 2008 and Querec et al. 2009 studies.

It is possible that the non-overlapped GO terms have closer relations in terms of the GO hierarchical structure. For example, these non-overlapped GO terms might share the same parents, siblings, or children terms. To test this hypothesis, we applied the GOfox GO visualization tool (E. Ong and He 2015) to visualize the significant GO terms based on their hierarchical structure (Figure 6-5). The shared enriched GO terms (with green color circles) are focused on categories including responses to viruses, cytokine-mediated signaling pathways, and defense response. Interestingly, responses to three types (alpha, beta, and gamma) of interferon cytokines are identified in the story. The response to interferon-alpha is shared between both re-analyses. However, responses to interferon-beta and interferon-gamma are significantly enriched in only Gaucher re-analysis (with red circles). The only GO term unique to Querec re-analysis is negative regulation of type I interferon production (with blue circle). How different interferon signaling pathways get involved in the protective immunity against Yellow Fever deserves further investigation. Several GO terms under cellular and RNA macromolecule metabolic processes were enriched only in Gaucher re-analysis, suggesting more general metabolic processes were detected in re-analysis of Gaucher than Querec. This study demonstrated that the hierarchical visualization of the enriched GO terms provides more useful information than plain lists of enriched GO terms.

The home page and the source code of VIO are publicly available from the GitHub website: https://github.com/vaccineontology/VIO. VIO has been deposited to the Ontobee website: http://www.ontobee.org/ontology/VIO, and BioPortal: http://bioportal.bioontology.org/ontologies/VIO.



- ▼ □ ● cellular metabolic process
  - ▼ □ ● RNA metabolic process
    - □ 🔴 **rRNA processing**
    - □ 🔴 **nuclear-transcribed mRNA catabolic process, nonsense-mediated decay**
  - ▼ □ ● cellular macromolecule metabolic process
    - □ 🔴 **nuclear-transcribed mRNA catabolic process, nonsense-mediated decay**
    - □ 🔴 **translation**
  - □ 🔴 **nuclear-transcribed mRNA catabolic process, nonsense-mediated decay**
  - □ 🔴 **translation**
  - □ 🔴 **translational initiation**
- ▼ □ ● negative regulation of biological process
  - □ 🟢 **negative regulation of viral genome replication**
  - □ 🔵 **negative regulation of type I interferon production**
  - □ 🔴 **nuclear-transcribed mRNA catabolic process, nonsense-mediated decay**
- □ 🔴 **SRP-dependent cotranslational protein targeting to membrane**
- □ 🔴 **viral transcription**
- ▼ □ ● response to other organism
  - □ 🟢 **defense response to virus**
  - ▼ □ 🟢 **response to virus**
    - □ 🟢 **defense response to virus**
- ▼ □ ● cytokine-mediated signaling pathway
  - □ 🟢 **interferon-gamma-mediated signaling pathway**
  - □ 🟢 **type I interferon signaling pathway**
- ▼ □ ● defense response
  - □ 🔴 **inflammatory response**
  - □ 🟢 **defense response to virus**
  - □ 🟢 **innate immune response**
- ▼ □ ● immune system process
  - ▼ □ 🔴 **immune response**
    - □ 🟢 **innate immune response**
  - □ 🟢 **defense response to virus**
- □ 🔴 **positive regulation of I-kappaB kinase/NF-kappaB signaling**
- ▼ □ ● response to cytokine
  - □ 🔴 **response to interferon-gamma**
  - □ 🟢 **response to interferon-alpha**
  - □ 🔴 **response to interferon-beta**

🟢 **Shared in both re-analyses**
🔴 **Unique to Gaucher re-analysis**
🔵 **Unique to Querec re-analysis**

Figure 6-5 Hierarichal display of significantly enriched GO biological process terms from the re-analysis of YF-Vax vaccine studies.

Circles colored with green, red, and blue represent GO terms shared in both re-analyses, unique to Gaucher et al. 2008 and unique to Querec et al. 2009, respectively.

6.5 Discussion

Overall, this chapter illustrated the development of the two ontologies, Vaccine Investigation Ontology (VIO) and Ontology of Host-Pathogen Interactions (OHPI), and demonstrated their applications on the standardization, integration, and analysis of host-pathogen and vaccine-host interaction studies. Both VIO and OHPI development followed the state-of-the-art eXtensible Ontology Development (XOD) principles, which support ontology reuse, alignment, design pattern usage, and community extensibility (He, Xiang, et al. 2018). The application of VIO to identify and standardize different variables in vaccine investigation studies provided a feasible way to integrate and compare published results from different vaccine host response studies. The creation of OHPI provides a valuable platform to integrate existing vaccine databases (specifically, virulence factor database Victors and protective antigen database Protegen) and one step forward to a more integrative host-pathogen interaction analysis. Since virulence factors (VFs) are utilized by the pathogens to infect host cells, an effective prevention method against the infectious pathogens is to evaluate VFs with the potential as protective antigens (PAgs). An ontology-based approach powered by the OHPI and VIO can improve our understanding of microbial pathogenesis and host immunity better to support effective vaccine research.

The VIO ontology provides a way to standardize the representation of minimal information standards and metadata representation for vaccine investigations, including both experimental and analytic parts. Our VIO modeling identified variables involved in raw data processing, data transformation, and statistical analyses (Figure 6-3). The re-analysis of two Yellow Fever vaccine studies standardized by VIO modeling found that the gene lists differed a lot between the two studies, while the GO enrichment results were more consistent (Figure 6-4).

This suggests that although the specific significantly differentiated genes might differ given different conditions, they participate in similar or related biological processes. Furthermore, our GOfox analyses showed that even the GO terms might show differences, the hierarchical structure comparison between the two sets of results showed that the different GO terms could often be aligned under the same ancestor GO terms (Figure 6-5). The identification of these hierarchical structures makes it better to understand the underlying molecular mechanisms.

Different from this study, where many data processing and analysis-related variables exist, a previous meta-analysis of *Brucella* vaccine protection study shows only one data-related variable (i.e., protection or not) (Todd et al. 2013). The *Brucella* meta-analysis study focuses on the effects of different experimental conditions toward the same vaccine protection efficiency. In that case, the data analysis is simple, but the roles of different experimental conditions can be determined. In total, the *Brucella* vaccine protection study identified approximately 20 experimental variables whose variations may change the protection outcomes. One major difference between these two types of vaccine investigations is that the *Brucella* vaccine protection study includes a step of virulent pathogen challenge, while the Yellow Fever vaccine study does not have the challenge step.

Not only in the vaccine domain, the challenge of standardizing and integrating homogenous data also exists in other biomedical domains and can be caused by experimental or analytical factors in the metadata. For example, the fields of cancer prognosis and prediction (Kourou et al. 2015), stem cell differentiation and aging (Muller-Sieburg et al. 2012), lung disease (Erb-Downward et al. 2011) all face the challenge. There are various sources of errors and inconsistencies associated with different high throughput technologies such as the microarray technology (Jaksik et al. 2015), flow cytometry (Cossarizza et al. 2017), and RNA-seq (Conesa et al. 2016). This study represents an effective ontology-based effort to solve the critical issue of different but overlapping results from studies on the same scientific question. In addition, the development of OHPI and VIO by following the state-of-the-art strategy and ensure that the ontology is open and logically well-formed to enable interoperability to ontologies in other biomedical domains. The interoperability can further solve the critical issue of data heterogeneity and inconsistency in interdisciplinary studies.

6.6 Acknowledgement

I appreciate the support by the National Institute of Allergy and Infectious Diseases (AI081062) awarded to Yongqun He (YH). Guanming Wu (GW) was awarded the National Human Genome Research Institute of the National Institutes of Health under Award Number U41HG003751. The results presented in this chapter is published in (E. Ong et al. 2019) and presented in the 2019 International Conference of Biomedical Ontologies (ICBO) (E. Ong and He 2019). Edison Ong (EO), GW, and YH conceived and designed the project. Peter Sun and Kimberly Berke performed data collection. EO, Jie Zheng, and YH contributed to ontology development. EO, GW, and YH performed the data analyses. EO and YH drafted the manuscript and presentation. All authors performed result interpretation, edited, and approved the manuscript and presentation.

Chapter 7 Summary and Discussion

In my dissertation work, I explored reverse vaccinology (RV) and structural vaccinology (SV) to select tentative vaccine candidates and support the optimization of such candidates. Vaxign-ML is a machine learning (ML)-based RV prediction tool that facilitates vaccine candidate selection with high accuracy. An SV strategy that utilizes the evolutionary-based protein design program, EvoDesign, is created to design vaccine candidate variants with enhanced immunogenicity. Besides, two ontologies, the Ontology of Host-Pathogen Interactions (OHPI) and Vaccine Investigation Ontology (VIO), were created to support the standardization, integration, and analysis of the heterogeneous vaccine data available in the Vaccine Investigation and Online Information Network (VIOLIN). The combination of vaccine development and ontology-based data analysis strategies provide a great avenue to not only augment our understanding of how pathogen and vaccines interact with the host immune system but also enhance our capability to quickly develop safe and effective vaccines. In the following sections, I summarized and discussed the dissertation work combining RV, SV, and ontology to support rational vaccine design, and how my work can be the foundation of future precision vaccinology studies.

7.1 Summary

As the starting point of my dissertation work, Chapter 2 describes a comprehensive bioinformatics study to analyze important vaccine design criteria by systematically studying and comparing bacterial protective antigens (PAgs) and non-protective proteins, including various protein properties and biological functions (E. Ong, Wong, and He 2017). The results of this study confirmed and provided details on the usage of these biological properties, such as subcellular localization, transmembrane helix, and adhesin probability, to be applied in both filtering-based and ML-based RV prediction of PAgs. In particular, these findings supported the creation of a new ML-based Vaxign program, Vaxign-ML, described in Chapter 3.

Vaxign-ML utilized biological and physicochemical properties computed from a high-quality Protegen database and showed superior predictive performance compared to existing RV tools (E. Ong, Wang, Wong, Seetharaman, et al. 2020). Protegen is a public PAg database that has continuously curated over thousands of PAgs supported by experimental evidence (vaccination-challenge assay of animal models) over the past decade (B. Yang et al. 2011). In Chapter 4, Vaxign-ML was also applied to predict COVID-19 vaccine antigen candidates (E. Ong, Wong, Huffman, and He 2020), with the SARS-CoV-2 spike (S) glycoprotein being the top candidate followed by the non-structural protein 3 (nsp3). The S protein is the primary target of most COVID-19 vaccines, including the Pfizer (Polack et al. 2020) and Moderna (Anderson et al. 2020) mRNA vaccines with high reported efficacy in Phase 3 clinical trials. On the other hand, the nsp3 protein predicted by Vaxign-ML and contained the Papain-Like protease (PLpro) sub-domain (Shin et al. 2020). PLpro was reported to play a critical role in the SARS-CoV-2 evasion mechanism against host antiviral immune responses, and inhibition of PLpro impaired the virus-induced cytopathogenic effect, maintained the antiviral interferon pathway and reduced viral replication in infected cells.

However, the current Vaxign-ML only elaborates on the pathogen proteins' properties and does not incorporate epitope information. Epitopes play a role in antibody and cell-mediated immunity, and the prediction of epitopes has been an active area of vaccine design. A comprehensive set of T cell and B cell epitope query, prediction, and analysis tools is available (Fleri et al. 2017). There are also epitope-based PAg prediction methods such as iVAX (L Moise et al. 2015) that utilize the frequency or density of the epitopes located on the protein. These epitope prediction methods can be streamlined into the Vaxign-ML pipeline to offer host-specific vaccine candidate prediction with better performance. In related work, I applied the prediction of epitopes and population coverage to evaluate five *Mycobacterium tuberculosis* (MTB) vaccines (E. Ong, He, and Yang 2020). Using computational approaches, I (i) predicted the capacity of the epitopes to be presented by the HLA molecules, (ii) predicted the promiscuity of the predicted epitopes based on a reference set of alleles and supertype alleles, and (iii) estimated the population coverage for ten protein antigens (Mtb39a, Mtb32a, Ag85B, ESAT-6, TB10.4, Rv2660, Rv3619, Rv2608, Rv3620, and Rv1813) constituting five MTB subunit vaccines (M72, H1, H4, H56, and ID93) that are currently in clinical trials. Our prediction showed that the ID93 vaccine was predicted to have the best potential for preventing both active and latent MTB infection (E. Ong, He, and Yang 2020). The study demonstrates the value of the computational approaches to pre-clinical evaluation of novel subunit vaccines. Future studies incorporating both pathogen and human variations in the context of epitope prediction can further extend the Vaxign and Vaxign-ML frameworks to augment our knowledge and improve vaccine antigen selection.

In Chapter 5, the computational design of the vaccine antigen was applied to the SARS-CoV-2 spike (S) protein to improve the immunogenicity and antigenicity of a vaccine candidate by manipulating the MHC-II T cell epitopes (E. Ong et al. 2021). The study aims to introduce new epitopes to the S protein while keeping the surface residues unchanged to minimize the structural change of the designed proteins. In the current settings, the immunogenicity evaluation was carried out after the standard protein design simulations with EvoDesign. In the future, the assessment of the immunogenic potential could be incorporated into the protein design process so that the sequence decoy generated at each step will be guided by balancing both the protein stability and immunogenicity. Moreover, with proper prior knowledge of known epitopes (e.g., both MHC-I and MHC-II from the pathogen proteome), it is also possible to create a chimeric protein, which integrates epitopes from antigens other than the target protein. Nonetheless, this SV strategy could be coupled with other structural modifications for a more rational structure-based vaccine design. The structural modifications performed on the native S protein, such as stabilizing the protein in its prefusion form (Bos et al. 2020) could still be integrated to design the S protein. The combination of these structural vaccinology technologies into the Vaxign framework could further enhance our capability to select candidates with better vaccine potential quickly. However, a major limitation of the current vaccine antigen design is the lack of experimental verification, which should be followed in future studies.

In Chapter 6, I described the creation of two ontologies, OHPI and VIO, to standardize and integrate vaccine data available among different databases. The OHPI integrates data from three databases: Protegen for protective antigens, Victors for the virulence factors related to the pathogen, and VaximmutorDB for host immune factors induced by vaccination. By integrating these three databases, OHPI models the host pathogen interaction data and can be applied with machine learning to predict vaccine antigens in future vaccine design studies. On the other hand, the VIO is created to model the variables associated with vaccine investigation studies. The data and meta-data of these studies are stored in the VIOLIN and PHIDIAS systems, such as vaccine preparation, efficacy, and protocol. VIO is also used in a study to model meta-data of gene expressions from public repositories such as GEO and ImmPort, and can help us to understand the mechanism of protection in the future. In the future, OHPI and VIO with other ontologies can provide a framework to support vaccine-related data integration and be utilized by machine learning to inform future vaccine design.

7.2 Future Direction of the Vaxign framework

The comprehensive Vaxign framework, which includes ML-based RV prediction, SV-based antigen optimization, epitope prediction, and population coverage assessment, as presented in this dissertation, may generate new knowledge about vaccine candidates that cannot easily be obtained from pre-clinical *in-vitro* and animal studies, or clinical trials conducted in a limited number of populations. A future workflow that streamlines (i) vaccine antigen candidate prediction by Vaxign-ML or potential extended work; (ii) candidate evaluation based on coverage of MHC-I and MHC-II supertype alleles, epitope promiscuity, and immunogenicity; (iii) antigen candidates optimized for immunogenicity by SV-guided design. With the accumulation of PAgs in the literature, it is also feasible to apply deep learning to improve the RV-based antigen selection process further. The population coverage of the vaccine candidates can also be assessed computationally based on known allele frequencies reported for the population of concern before entering the clinical trial. Finally, these computationally predicted candidates should be followed up and verified by *in-vitro* or *in-vivo* experiments. The computational approaches may also bridge the pre-clinical studies and clinical development of vaccine candidates, which is an important gap in current vaccine development. The integration of bioinformatics and computational approaches with traditional vaccine development tools presents the best opportunity for rapid development of effective and safe vaccines in the era of precision medicine.

7.3 The Promise of Precision Vaccinology

Precision medicine is broadly defined as the delivering personalized treatments to individual patients, or “the right drug for the right patient at the right time” (Abrahams 2008). The same concept could be applied to vaccinology, from precision medicine to precision vaccinology. The practice of precision vaccinology depends on the high-throughput technologies to acquire detailed molecular phenotypes of humans and derive nuanced descriptions of disease, and support advanced vaccine discovery. A key component of precision vaccinology is the ability to obtain large amounts of molecular data through high-throughput sequencing technologies to investigate the underlying mechanisms, and it has been applied to study the repertoire of B cell receptors (BCRs) and T cell receptors (TCRs) (Chiffelle et al. 2020). The recombination of the genes encoding for BCRs and TCRs results in a massive pool of repertoire for these two receptors with an approximation of at least 10^{12} (Briney et al. 2019) and 10^{15} (Nikolich-Žugich, Slifka, and Messaoudi 2004) unique BCRs and TCRs, respectively, in humans. Due to the extreme diversity of the human immune repertoire, researchers could have only studied a tiny fraction of the complete antigen receptor repertoire before the era of next-generation sequencing (NGS) using non-sequencing experimental protocols such as hybridization-based methods (Bernardin et al. 2003; Baum and McCune 2006). The advent of NGS technology allowed researchers to study the immune repertoire at a much superior depth than the previous decades. For example, NGS has been the driving force of the identification and development of the broadly neutralizing antibodies against Human Immunodeficiency Virus (HIV) infection (X. Wu et al. 2011; J. Zhu, Wu, et al. 2013; J. Zhu, Ofek, et al. 2013). Single-cell sequencing method also reveals the diversity in clonal expansion of TCR repertoire in antiretroviral therapy treated HIV patients (Gantner et al. 2020).

Besides analyzing the host immune repertoire, NGS can also be applied to study pathogenicity and pathogen-host interactions of existing and emerging infectious diseases. The whole-genome sequencing (WGS) was applied to monitor multidrug-resistant tuberculosis in Austria, Romania and Germany in 2014 (Fiebig et al. 2017). In a collaboration project with the Michigan Department of Community Health (MDCH), a sequencing pipeline was built to process and analyze the MTB whole-genome sequence data (https://github.com/e4ong1031/MDHHS_TB_WGS). The genome sequence of the SARS-CoV-2 has dramatically speeded up the development of diagnostic tools, drug discovery, and vaccine development to control the COVID-19 pandemic. On the other hand, the dual RNA-seq protocol (Westermann, Gorski, and Vogel 2012; Westermann et al. 2016) also facilitates the in-depth analysis of the mechanisms behind host-pathogen interactions. Typically, the pathogen and host cells are separated before the sequencing by, for example, physical separation of a centrifuge. However, in dual RNA-seq, the total isolated RNA of the pathogen infected cells is sequenced at the same time and separated later on during the mapping process. This technology enables researchers to monitor the changes of gene expression profiles in both the host and pathogen under different experimental conditions. Overall, the immune repertoire "big-data" not only improves our system-level understanding of immunology, infectious diseases, and vaccinations, but also build the foundation of precision vaccinology.

The generation of this massive molecular data is definitely a valuable resource to decipher the mechanism of the immune system, but, at the same time, such heterogeneous data also requires an infrastructure to ensure the data to be Findable, Accessible, Interoperable, and Reusable (FAIR) (Stall et al. 2019). To achieve these objectives, ontology can serve as the ideal platform for data FAIRness, and it has been applied to precision medicine. The creation of the

Kidney Tissue Atlas Ontology (KTAO) (He, Steck, et al. 2018) and Ontology of Precision Medicine (OPMI) (He et al. 2019), along with the vaccine-informatics research in my dissertation work, are one of the pioneering work to integrate the clinical, histopathological, and molecular data (Figure 7-1) generated in the Kidney Precision Medicine Project (KPMP) consortium (E. Ong, Wang, Schaub, O'Toole, et al. 2020).

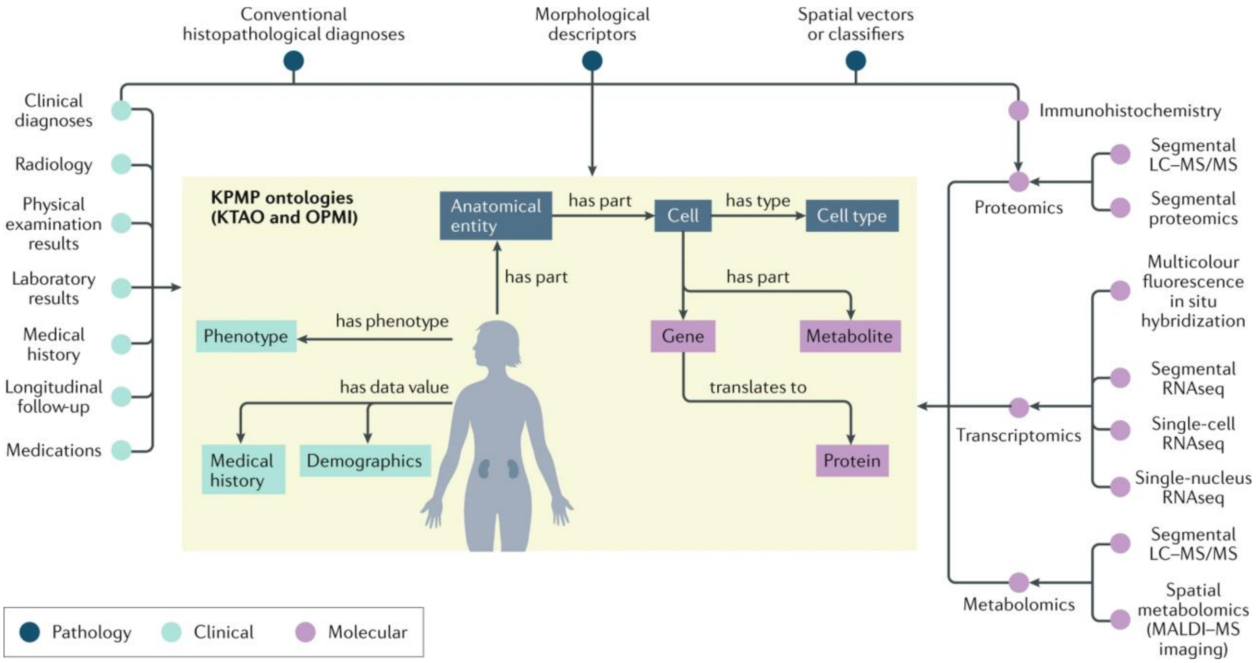


Figure 7-1 The KPMP ontology framework for supporting data representation, integration and analysis.

Clinical, pathology, and molecular data collected from Kidney Precision Medicine Project (KPMP) recruitment sites and tissue interrogation sites will be deposited in the KPMP Kidney Tissue Atlas. Different types of data (clinical, pathology, and molecular) feed into the KPMP ontology environment. Two KPMP ontologies, the Kidney Tissue Atlas Ontology (KTAO) and the Ontology of Precision Medicine Investigation (OPMI), provide a semantic framework for modeling relationships between the heterogeneous data in the atlas. LC-MS/MS, liquid chromatography-tandem mass spectrometry; MALDI-MS, matrix-assisted laser desorption/ionization-mass spectrometry; RNAseq, RNA sequencing. This figure is reprinted from (E. Ong, Wang, Schaub, O'Toole, et al. 2020) published in *Nature Reviews Nephrology* under the license of Creative Commons Attribution License (CC BY).

A preliminary ontology infrastructure has already been implemented for precision vaccinology. The integrative Vaccine Investigation and Online Information Network (VIOLIN) (He et al. 2014) has cumulated large amounts of vaccine-related data. The Vaccine Ontology (VO) is created to represent vaccine-related entities, such as vaccines, vaccine components, vaccinations, host responses to vaccines, and the relations among these entities available in the VIOLIN (He et al. 2009; Özgür et al. 2011; Y. Lin and He 2012). Protegen (http://www.violinet.org/protegen) is a VIOLIN sub-database that stores over 1,000 PAgs (B. Yang et al. 2011; He and Xiang 2012), and it is a key component in the development of vaccine prediction tools. Another useful resource is the Pathogen-Host Interaction Data Integration and Analysis System (PHIDIAS) (Z. Xiang, Tian, and He 2007). The Victors database is a sub-database of PHIDIAS and includes over 5,000 VFs from different bacteria, viruses, parasites, and fungi, which are pathogenic to animals and humans (Sayers et al. 2019). The creation of Ontology of Host-Pathogen Interaction (OHPI) and Vaccine Investigation Ontology (VIO) described in Chapter 6 is the first attempt to utilize ontology to support advanced knowledge representation, integration, sharing, and analysis among these vaccine databases. Last but not least, ontologies have been applied with machine learning to predict protein-protein interactions and gene-disease associations (Smaili, Gao, and Hoehndorf 2018; 2019). Using the precision vaccinology knowledge formalized within the ontologies, which is curated by experienced domain experts, the information may be used as a priori in machine learning-based prediction. The combination of a comprehensive vaccine design framework, massive immune repertoire sequencing data, and an ontology-powered data integration infrastructure can serve as the foundation for precision vaccinology.

Bibliography

Abraham, Mark James, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindah. 2015. "Gromacs: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers." *SoftwareX* 1–2: 19–25.

Abrahams, Edward. 2008. "Right Drug-Right Patient-Right Time: Personalized Medicine Coalition." *Clinical and Translational Science*. https://doi.org/10.1111/j.1752-8062.2008.00003.x.

Adu-Bobie, Jeannette, Barbara Capecchi, Davide Serruto, Rino Rappuoli, and Mariagrazia Pizza. 2003. "Two Years into Reverse Vaccinology." *Vaccine* 21 (7–8): 605–10. https://doi.org/10.1016/S0264-410X(02)00566-2.

Al-Amri, Sawsan S., Ayman T. Abbas, Loai A. Siddiq, Abrar Alghamdi, Mohammad A. Sanki, Muhanna K. Al-Muhanna, Rowa Y. Alhabbab, Esam I. Azhar, Xuguang Li, and Anwar M. Hashem. 2017. "Immunogenicity of Candidate MERS-CoV DNA Vaccines Based on the Spike Protein." *Scientific Reports* 7: 44875. https://doi.org/10.1038/srep44875.

Alquraishi, Mohammed. 2019. "AlphaFold at CASP13." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btz422.

Anderson, Evan J., Nadine G. Rouphael, Alicia T. Widge, Lisa A. Jackson, Paul C. Roberts, Mamodikoe Makhene, James D. Chappell, et al. 2020. "Safety and Immunogenicity of SARS-CoV-2 MRNA-1273 Vaccine in Older Adults." *New England Journal of Medicine* 383: 2427–38.

Arp, Robert, Barry Smith, and Andrew D. Spear. 2016. *Building Ontologies with Basic Formal Ontology*. *Building Ontologies with Basic Formal Ontology*. https://doi.org/10.7551/mitpress/9780262527811.001.0001.

Bahl, Sunil, Pankaj Bhatnagar, Roland W. Sutter, Sigrun Roesel, and Michel Zaffran. 2018. "Global Polio Eradication – Way Ahead." *Indian Journal of Pediatrics*. https://doi.org/10.1007/s12098-017-2586-8.

Bail, Samantha, Birte Glimm, Rafael Gon, and Ernesto Jim. 2013. "ELK Reasoner: Architecture and Evaluation." In *Proceedings of the 2nd OWL Reasoner Evaluation Workshop (ORE 2013)*. Vol. 858. CEUR Workshop Proceedings.

Baldwin, Susan L., Valerie A. Reese, Po Wei D Huang, Elyse A. Beebe, Brendan K. Podell, Steven G. Reed, and Rhea N. Coler. 2016. "Protection and Long-Lived Immunity Induced by the ID93/GLA-SE Vaccine Candidate against a Clinical Mycobacterium Tuberculosis Isolate." *Clinical and Vaccine Immunology* 23 (2): 137–47. https://doi.org/10.1128/CVI.00458-15.

Bandrowski, Anita, Ryan Brinkman, Mathias Brochhausen, Matthew H. Brush, Bill Bug, Marcus C. Chibucos, Kevin Clancy, et al. 2016. “The Ontology for Biomedical Investigations.” *PLoS One* 11 (4): e0154556. https://doi.org/10.1371/journal.pone.0154556.

Barnes, Christopher O., Anthony P. West, Kathryn E. Huey-Tubman, Magnus A.G. Hoffmann, Naima G. Sharaf, Pauline R. Hoffman, Nicholas Koranda, et al. 2020. “Structures of Human Antibodies Bound to SARS-CoV-2 Spike Reveal Common Epitopes and Recurrent Features of Antibodies.” *Cell*. https://doi.org/10.1016/j.cell.2020.06.025.

Barrett, Tanya, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, et al. 2013. “NCBI GEO: Archive for Functional Genomics Data Sets - Update.” *Nucleic Acids Research* 41 (D1): 991–95. https://doi.org/10.1093/nar/gks1193.

Baum, Paul D., and Joseph M. McCune. 2006. “Direct Measurement of T-Cell Receptor Repertoire Diversity with AmpliCot.” *Nature Methods* 3 (11): 895–901. https://doi.org/10.1038/nmeth949.

Belongia, Edward A., and Allison L. Naleway. 2003. “Smallpox Vaccine: The Good, the Bad, and the Ugly.” *Clinical Medicine & Research*. https://doi.org/10.3121/cmr.1.2.87.

Bernardin, Flavien, Laurence Doukhan, Alcira Longone-Miller, Patrick Champagne, Rafick Sekaly, and Eric Delwart. 2003. “Estimate of the Total Number of CD8+ Clonal Expansions in Healthy Adults Using a New DNA Heteroduplex-Tracking Assay for CDR3 Repertoire Analysis.” *Journal of Immunological Methods* 274 (1–2): 159–75. https://doi.org/10.1016/S0022-1759(02)00514-8.

Bert, Nina Le, Anthony T. Tan, Kamini Kunasegaran, Christine Y. L. Tham, Morteza Hafezi, Adeline Chia, Melissa Hui Yen Chng, et al. 2020. “SARS-CoV-2-Specific T Cell Immunity in Cases of COVID-19 and SARS, and Uninfected Controls.” *Nature*. https://doi.org/10.1038/s41586-020-2550-z.

Best, Robert B., Xiao Zhu, Jihyun Shim, Pedro E.M. Lopes, Jeetain Mittal, Michael Feig, and Alexander D. MacKerell. 2012. “Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone φ, ψ and Side-Chain X1 and X2 Dihedral Angles.” *Journal of Chemical Theory and Computation* 8: 3257–73.

Bigelow, Henry R., Donald S. Petrey, Jinfeng Liu, Dariusz Przybylski, and Burkhard Rost. 2004. “Predicting Transmembrane Beta-Barrels in Proteomes.” *Nucleic Acids Research* 32 (8): 2566–77. https://doi.org/10.1093/nar/gkh580.

Bisht, Himani, Anjeanette Roberts, Leatrice Vogel, Alexander Bukreyev, Peter L. Collins, Brian R. Murphy, Kanta Subbarao, and Bernard Moss. 2004. “Severe Acute Respiratory Syndrome Coronavirus Spike Protein Expressed by Attenuated Vaccinia Virus Protectively Immunizes Mice.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (17): 6641–46. https://doi.org/10.1073/pnas.0401939101.

Blake, J. A., K. R. Christie, M. E. Dolan, H. J. Drabkin, D. P. Hill, L. Ni, D. Sitnikov, et al. 2015. “Gene Ontology Consortium: Going Forward.” *Nucleic Acids Research* 43 (D1): D1049–56. https://doi.org/10.1093/nar/gku1179.

Blazanovic, Kristina, Hongliang Zhao, Yoonjoo Choi, Wen Li, Regina S. Salvat, Daniel C. Osipovitch, Jennifer Fields, et al. 2015. “Structure-Based Redesign of Lysostaphin Yields

Potent Antistaphylococcal Enzymes That Evade Immune Cell Surveillance." *Molecular Therapy - Methods and Clinical Development* 2: 15021.

Bos, Rinke, Lucy Rutten, Joan E.M. van der Lubbe, Mark J.G. Bakkers, Gijs Hardenberg, Frank Wegmann, David Zuijdgeest, et al. 2020. "Ad26 Vector-Based COVID-19 Vaccine Encoding a Prefusion-Stabilized SARS-CoV-2 Spike Immunogen Induces Potent Humoral and Cellular Immune Responses." *Npj Vaccines* 5 (1). https://doi.org/10.1038/s41541-020-00243-x.

Bowman, Brett N., Paul R. McAdam, Sandro Vivona, Jin X. Zhang, Tiffany Luong, Richard K. Belew, Harpal Sahota, et al. 2011. "Improving Reverse Vaccinology with a Machine Learning Approach." *Vaccine* 29 (45): 8156–64. https://doi.org/10.1016/j.vaccine.2011.07.142.

Braun, Julian, Lucie Loyal, Marco Frentsch, Daniel Wendisch, Philipp Georg, Florian Kurth, Stefan Hippenstiel, et al. 2020. "SARS-CoV-2-Reactive T Cells in Healthy Donors and Patients with COVID-19." *Nature*. https://doi.org/10.1038/s41586-020-2598-9.

Briney, Bryan, Anne Inderbitzin, Collin Joyce, and Dennis R. Burton. 2019. "Commonality despite Exceptional Diversity in the Baseline Human Antibody Repertoire." *Nature* 566 (7744): 393–97. https://doi.org/10.1038/s41586-019-0879-y.

Brinkman, Ryan R, Mélanie Courtot, Dirk Derom, Jennifer M Fostel, Yongqun He, Phillip Lord, James Malone, et al. 2010. "Modeling Biomedical Experimental Processes with OBI." *Journal of Biomedical Semantics* 1 (Suppl 1): 1–11.

Bussi, Giovanni, Davide Donadio, and Michele Parrinello. 2007. "Canonical Sampling through Velocity Rescaling." *Journal of Chemical Physics* 126: 014101.

Byrne, J P. 2008. *Encyclopedia of Pestilence, Pandemics, and Plagues*. Encyclopedia of Pestilence, Pandemics, and Plagues. Greenwood Press. https://books.google.com/books?id=32l7tAEACAAJ.

Cabeça, Tatiane K., Celso Granato, and Nancy Bellei. 2013. "Epidemiological and Clinical Features of Human Coronavirus Infections among Different Subsets of Patients." *Influenza and Other Respiratory Viruses* 7 (6): 1040–47. https://doi.org/10.1111/irv.12101.

Cafaro, Aurelio, Antonella Tripiciano, Orietta Picconi, Cecilia Sgadari, Sonia Moretti, Stefano Buttò, Paolo Monini, and Barbara Ensoli. 2019. "Anti-Tat Immunity in HIV-1 Infection: Effects of Naturally Occurring and Vaccine-Induced Antibodies against Tat on the Course of the Disease." *Vaccines* 7 (3): 99. https://doi.org/10.3390/vaccines7030099.

Calis, Jorg J.A., Rob J. de Boer, and Can Keşmir. 2012. "Degenerate T-Cell Recognition of Peptides on MHC Molecules Creates Large Holes in the T-Cell Repertoire." *PLoS Computational Biology* 8 (3): e1002412.

Camacho, C, G Coulouris, V Avagyan, N Ma, J Papadopoulos, K Bealer, and T L Madden. 2009. "BLAST plus: Architecture and Applications." *BMC Bioinformatics* 10 (421): 1. https://doi.org/Artn 421\nDoi 10.1186/1471-2105-10-421.

Cao, Yunlong, Bin Su, Xianghua Guo, Wenjie Sun, Yongqiang Deng, Linlin Bao, Qinyu Zhu, et al. 2020. "Potent Neutralizing Antibodies against SARS-CoV-2 Identified by High-Throughput Single-Cell Sequencing of Convalescent Patients' B Cells." *Cell* 182 (1): 73–

84.

Capra, John A., and Mona Singh. 2007. "Predicting Functionally Important Residues from Sequence Conservation." *Bioinformatics* 23 (15): 1875–82. https://doi.org/10.1093/bioinformatics/btm270.

Cartwright, Frederick Fox., and Michael D Biddiss. 2000. "Disease & History." *Disease and History*. Stroud: Sutton Pub. file://catalog.hathitrust.org/Record/004106535.

Chan, Jasper F.W., Susanna K.P. Lau, Kelvin K.W. To, Vincent C.C. Cheng, Patrick C.Y. Woo, and Kwok Yung Yue. 2015. "Middle East Respiratory Syndrome Coronavirus: Another Zoonotic Betacoronavirus Causing SARS-like Disease." *Clinical Microbiology Reviews* 28 (2): 465–522.

Channappanavar, Rudragouda, Jincun Zhao, and Stanley Perlman. 2014. "T Cell-Mediated Immune Response to Respiratory Coronaviruses." *Immunologic Research* 59 (1–3): 118–28. https://doi.org/10.1007/s12026-014-8534-z.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. "SMOTE: Synthetic Minority Over-Sampling Technique Nitesh." *Journal of Artificial Intelligence Research* 16 (1): 321–57. https://doi.org/10.1613/jair.953.

Chen, Hong Ru, Yen Chung Lai, and Trai Ming Yeh. 2018. "Dengue Virus Non-Structural Protein 1: A Pathogenic Factor, Therapeutic Target, and Vaccine Candidate." *Journal of Biomedical Science* 25 (1): 58. https://doi.org/10.1186/s12929-018-0462-0.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Augu:785–94. https://doi.org/10.1145/2939672.2939785.

Chiffelle, Johanna, Raphael Genolet, Marta AS Perez, George Coukos, Vincent Zoete, and Alexandre Harari. 2020. "T-Cell Repertoire Analysis and Metrics of Diversity and Clonality." *Current Opinion in Biotechnology* 65: 284–95. https://doi.org/10.1016/j.copbio.2020.07.010.

Choi, Yoonjoo, Deeptak Verma, Karl E Griswold, and Chris Bailey-Kellogg. 2017. "EpiSweep: Computationally Driven Reengineering of Therapeutic Proteins to Reduce Immunogenicity While Maintaining Function." In *Computational Protein Design*, edited by Ilan Samish, 375–98. New York, NY: Springer New York.

Chou, Kuo-Chen. 2000. "Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Order Effect." *Biochemical and Biophysical Research Communications* 278 (2): 477–83. https://doi.org/10.1006/bbrc.2000.3815.

Coleman, Christopher M., Jeanne M. Sisk, Gabor Halasz, Jixin Zhong, Sarah E. Beck, Krystal L. Matthews, Thiagarajan Venkataraman, Sanjay Rajagopalan, Christos A. Kyratsous, and Matthew B. Frieman. 2017. "CD8+ T Cells and Macrophages Regulate Pathogenesis in a Mouse Model of Middle East Respiratory Syndrome." *Journal of Virology* 91 (1). https://doi.org/10.1128/jvi.01825-16.

Collins, Brenda S. 2011. "Gram-Negative Outer Membrane Vesicles in Vaccine Development." *Discovery Medicine* 12 (62): 7–15. http://www.ncbi.nlm.nih.gov/pubmed/21794204.

Conesa, Ana, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michal Wojciech Michał Wojciech Szcześniak, et al. 2016. “A Survey of Best Practices for RNA-Seq Data Analysis.” *Genome Biology* 17 (1): 13. https://doi.org/10.1186/s13059-016-0881-8.

Cossarizza, Andrea, Hyun Dong Chang, Andreas Radbruch, Mübeccel Akdis, Immanuel Andrä, Francesco Annunziato, Petra Bacher, et al. 2017. “Guidelines for the Use of Flow Cytometry and Cell Sorting in Immunological Studies.” *European Journal of Immunology* 47 (10): 1584–1797. https://doi.org/10.1002/eji.201646632.

Cowell, Lindsay Grey, and Barry Smith. 2010. “Infectious Disease Ontology.” In *Infectious Disease Informatics*, 373–95. Springer New York. https://doi.org/10.1007/978-1-4419-1327-2_19.

Crooks, Gavin E., Gary Hon, John Marc Chandonia, and Steven E. Brenner. 2004. “WebLogo: A Sequence Logo Generator.” *Genome Research* 14 (6): 1188–90. https://doi.org/10.1101/gr.849004.

Dalsass, Mattia, Alessandro Brozzi, Duccio Medini, and Rino Rappuoli. 2019. “Comparison of Open-Source Reverse Vaccinology Programs for Bacterial Vaccine Antigen Discovery.” *Frontiers in Immunology* 10 (February): 1–12. https://doi.org/10.3389/fimmu.2019.00113.

Delves, P J, S J Martin, D R Burton, and I M Roitt. 2016. *Roitt’s Essential Immunology*. Essentials. Wiley. https://books.google.com/books?id=QZWDDQAAQBAJ.

Dhanda, Sandeep Kumar, Swapnil Mahajan, Sinu Paul, Zhen Yan, Haeuk Kim, Martin Closter Jespersen, Vanessa Jurtz, et al. 2019. “IEDB-AR: Immune Epitope Database—Analysis Resource in 2019.” *Nucleic Acids Research* 47 (W1): W502–6.

Diehl, Alexander D., Terrence F. Meehan, Yvonne M. Bradford, Matthew H. Brush, Wasila M. Dahdul, David S. Dougall, Yongqun He, et al. 2016. “The Cell Ontology 2016: Enhanced Content, Modularization, and Ontology Interoperability.” *Journal of Biomedical Semantics* 7 (1): 44. https://doi.org/10.1186/s13326-016-0088-7.

Ding, C., and H. Peng. 2003. “Minimum Redundancy Feature Selection from Microarray Gene Expression Data.” *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003* 3 (2): 185–205. https://doi.org/10.1109/CSB.2003.1227396.

Doremalen, Neeltje van, Elaine Haddock, Friederike Feldmann, Kimberly Meade-White, Trenton Bushmaker, Robert Fischer, Atsushi Okumura, et al. 2020. “A Single Dose of ChAdOx1 MERS Provides Broad Protective Immunity against a Variety of MERS-CoV Strains.” BioRxiv [Preprint]. 2020. https://www.biorxiv.org/content/10.1101/2020.04.13.036293v1.

Doytchinova, Irini a, and Darren R Flower. 2007. “VaxiJen: A Server for Prediction of Protective Antigens, Tumour Antigens and Subunit Vaccines.” *BMC Bioinformatics* 8: 4. https://doi.org/10.1186/1471-2105-8-4.

Dubchak, Inna, Ilya Muchnik, Stephen R. Holbrook, and Sung Hou Kim. 1995. “Prediction of Protein Folding Class Using Global Description of Amino Acid Sequence.” *Proceedings of the National Academy of Sciences of the United States of America* 92 (19): 8700–8704. https://doi.org/10.1073/pnas.92.19.8700.

Durmuş Tekir, Saliha, Tunahan Çakır, Emre Ardiç, Ali Semih Sayılırbaş, Gökhan Konuk, Mithat Konuk, Hasret Sarıyer, et al. 2013. “PHISTO: Pathogen-Host Interaction Search Tool.” *Bioinformatics (Oxford, England)* 29 (10): 1357–58. https://doi.org/10.1093/bioinformatics/btt137.

Echenberg, Myron. 2002. “Pestis Redux: The Initial Years of the Third Bubonic Plague Pandemic, 1894-1901.” *Journal of World History : Official Journal of the World History Association* 13 (2): 429–49. https://doi.org/10.1353/jwh.2002.0033.

Edgar, Robert C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucleic Acids Research* 32 (5): 1792–97. https://doi.org/10.1093/nar/gkh340.

El-Manzalawy, Yasser, Drena Dobbs, and Vasant Honavar. 2012. “Predicting Protective Bacterial Antigens Using Random Forest Classifiers.” In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, 426–433. https://doi.org/10.1145/2382936.2382991.

Elbe, Stefan, and Gemma Buckland-Merrett. 2017. “Data, Disease and Diplomacy: GISAID’s Innovative Contribution to Global Health.” *Global Challenges* 1 (1): 33–46. https://doi.org/10.1002/gch2.1018.

Enayatkhani, Maryam, Mehdi Hasaniazad, Sobhan Faezi, Hamed Guklani, Parivash Davoodian, Nahid Ahmadi, Mohammad Ali Einakian, Afsaneh Karmostaji, and Khadijeh Ahmadi. 2020. “Reverse Vaccinology Approach to Design a Novel Multi-Epitope Vaccine Candidate against COVID-19: An in Silico Study.” *Journal of Biomolecular Structure and Dynamics*, 1–16.

Erb-Downward, John R., Deborah L. Thompson, Meilan K. Han, Christine M. Freeman, Lisa McCloskey, Lindsay A. Schmidt, Vincent B. Young, et al. 2011. “Analysis of the Lung Microbiome in the ‘Healthy’ Smoker and in COPD.” *PLoS ONE* 6 (2). https://doi.org/10.1371/journal.pone.0016384.

Essmann, Ulrich, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen. 1995. “A Smooth Particle Mesh Ewald Method.” *The Journal of Chemical Physics* 103 (19): 8577–93.

Fabregat, Antonio, Konstantinos Sidiropoulos, Phani Garapati, Marc Gillespie, Kerstin Hausmann, Robin Haw, Bijay Jassal, et al. 2016. “The Reactome Pathway Knowledgebase.” *Nucleic Acids Research* 44 (D1): D481–87. https://doi.org/10.1093/nar/gkv1351.

Falda, Marco, Stefano Toppo, Alessandro Pescarolo, Enrico Lavezzo, Barbara Di Camillo, Andrea Facchinetti, Elisa Cilia, Riccardo Velasco, and Paolo Fontana. 2012. “Argot2: A Large Scale Function Prediction Tool Relying on Semantic Similarity of Weighted Gene Ontology Terms.” *BMC Bioinformatics* 13 (Suppl 4): S14. https://doi.org/10.1186/1471-2105-13-S4-S14.

Federhen, S. 2012. “The NCBI Taxonomy.” *Nucleic Acids Res.* 40 (D1): D136--D143. https://doi.org/10.1093/nar/gkr1178.

Feng, Z. P., and C. T. Zhang. 2000. “Prediction of Membrane Protein Types Based on the Hydrophobic Index of Amino Acids.” *Journal of Protein Chemistry* 19 (4): 269–75. https://doi.org/10.1023/A:1007091128394.

Fett, C., M. L. DeDiego, J. A. Regla-Nava, L. Enjuanes, and S. Perlman. 2013. “Complete Protection against Severe Acute Respiratory Syndrome Coronavirus-Mediated Lethal Respiratory Disease in Aged Mice by Immunization with a Mouse-Adapted Virus Lacking E Protein.” *Journal of Virology* 87 (12): 6551–59. https://doi.org/10.1128/jvi.00087-13.

Fiebig, Lena, Thomas A Kohl, Odette Popovici, Margarita Mühlenfeld, Alexander Indra, Daniela Homorodean, Domnica Chiotan, et al. 2017. “A Joint Cross-Border Investigation of a Cluster of Multidrug-Resistant Tuberculosis in Austria, Romania and Germany in 2014 Using Classic, Genotyping and Whole Genome Sequencing Methods: Lessons Learnt.” *Euro Surveillance : Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* 22 (2). https://doi.org/10.2807/1560-7917.ES.2017.22.2.30439.

Fine, Paul EM. 1995. “Variation in Protection by BCG: Implications of and for Heterologous Immunity.” *The Lancet* 346 (8986): 1339–45. https://doi.org/10.1016/S0140-6736(95)92348-9.

Fleishman, Sarel J., Andrew Leaver-Fay, Jacob E. Corn, Eva Maria Strauch, Sagar D. Khare, Nobuyasu Koga, Justin Ashworth, et al. 2011. “Rosettascripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite.” *PLoS ONE* 6 (6): e20161.

Fleri, Ward, Sinu Paul, Sandeep Kumar Dhanda, Swapnil Mahajan, Xiaojun Xu, Ward Fleri, Bjoern Peters, and Alessandro Sette. 2017. “The Immune Epitope Database and Analysis Resource in Epitope Discovery and Synthetic Vaccine Design.” *Frontiers in Immunology* 8 (MAR): 1–16. https://doi.org/10.3389/fimmu.2017.00278.

Flower, Darren R., Isabel K. MacDonald, Kamna Ramakrishnan, Matthew N. Davies, and Irini A. Doytchinova. 2010. “Computer Aided Selection of Candidate Vaccine Antigens.” *Immunome Research* 6 (SUPPL. 2): 1–16. https://doi.org/10.1186/1745-7580-6-S2-S1.

Folaranmi, T, L Rubin, S W Martin, M Patel, and J R MacNeil. 2015. “Use of Serogroup B Meningococcal Vaccines in Persons Aged≥ 10 Years at Increased Risk for Serogroup B Meningococcal Disease: Recommendations of the Advisory Committee on Immunization Practices, 2015.” *MMWR Morb Mortal Wkly Rep* 64 (22): 608–12.

Folegatti, Pedro M, Katie J Ewer, Parvinder K Aley, Brian Angus, Stephan Becker, Sandra Belij-Rammerstorfer, Duncan Bellamy, et al. 2020. “Safety and Immunogenicity of the ChAdOx1 NCoV-19 Vaccine against SARS-CoV-2: A Preliminary Report of a Phase 1/2, Single-Blind, Randomised Controlled Trial.” *The Lancet*. https://doi.org/10.1016/S0140-6736(20)31604-4.

Frankild, Sune, Rob J. de Boer, Ole Lund, Morten Nielsen, and Can Kesmir. 2008. “Amino Acid Similarity Accounts for T Cell Cross-Reactivity and for ‘Holes’ in the T Cell Repertoire.” *PLoS ONE* 3 (3): e1831.

Gantner, Pierre, Amélie Pagliuzza, Marion Pardons, Moti Ramgopal, Jean Pierre Routy, Rémi Fromentin, and Nicolas Chomont. 2020. “Single-Cell TCR Sequencing Reveals Phenotypically Diverse Clonally Expanded Cells Harboring Inducible HIV Proviruses during ART.” *Nature Communications* 11 (1): 1–9. https://doi.org/10.1038/s41467-020-17898-8.

Gao, Qiang, Linlin Bao, Haiyan Mao, Lin Wang, Kangwei Xu, Minnan Yang, Yajing Li, et al.

2020. “Rapid Development of an Inactivated Vaccine for SARS-CoV-2.” BioRxiv [Preprint]. 2020. https://www.biorxiv.org/content/10.1101/2020.04.17.046375v1.

Gardner, Christina L., and Kate D. Ryman. 2010. “Yellow Fever: A Reemerging Threat.” *Clinics in Laboratory Medicine*. https://doi.org/10.1016/j.cll.2010.01.001.

Gates, Bill, and Melinda Gates. 2017. “Warren Buffett’s Best Investment.” 2017. https://www.gatesnotes.com/2017-Annual-Letter?WT.mc_id=02_15_2017_00_AL2017Twitter_GF-TW_&WT.tsrc=GFTW.

Gaucher, Denis, René Therrien, Nadia Kettaf, Bastian R. Angermann, Geneviéve Boucher, Abdelali Filali-Mouhim, Janice M. Moser, et al. 2008. “Yellow Fever Vaccine Induces Integrated Multilineage and Polyfunctional Immune Responses.” *Journal of Experimental Medicine* 205 (13): 3119–31. https://doi.org/10.1084/jem.20082292.

Gibson, Carl A., Jacob J. Schlesinger, and Alan D.T. Barrett. 1988. “Prospects for a Virus Non-Structural Protein as a Subunit Vaccine.” *Vaccine* 6 (1): 7–9. https://doi.org/10.1016/0264-410X(88)90004-7.

Glansbeek, H. L., B. L. Haagmans, E. G. Te Lintelo, H. F. Egberink, V. Duquesne, A. Aubert, M. C. Horzinek, and P. J.M. Rottier. 2002. “Adverse Effects of Feline IL-12 during DNA Vaccination against Feline Infectious Peritonitis Virus.” *Journal of General Virology* 83 (1): 1–10. https://doi.org/10.1099/0022-1317-83-1-1.

Godlewska, Renata, Maciej Kuczkowski, Agnieszka Wyszyńska, Joanna Klim, Katarzyna Derlatka, Anna Woźniak-Biel, Elżbieta K. Jagusztyn-krynicka, et al. 2016. “Evaluation of a Protective Effect of in Ovo Delivered Campylobacter Jejuni OMVs.” *Applied Microbiology and Biotechnology* 100 (20): 8855–64. https://doi.org/10.1007/s00253-016-7699-x.

Goodswen, Stephen J, Paul J Kennedy, and John T Ellis. 2013. “A Novel Strategy for Classifying the Output from an in Silico Vaccine Discovery Pipeline for Eukaryotic Pathogens Using Machine Learning Algorithms.” *BMC Bioinformatics* 14 (1): 315. https://doi.org/10.1186/1471-2105-14-315.

Gouy, Manolo, Stéphane Guindon, and Olivier Gascuel. 2010. “Sea View Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building.” *Molecular Biology and Evolution* 27 (2): 221–24.

Graham, Barney S., Kayvon Modjarrad, and Jason S. McLellan. 2015. “Novel Antigens for RSV Vaccines.” *Current Opinion in Immunology* 35: 30–38. https://doi.org/10.1016/j.coi.2015.04.005.

Graham, Rachel L., Michelle M. Becker, Lance D. Eckerle, Meagan Bolles, Mark R. Denison, and Ralph S. Baric. 2012. “A Live, Impaired-Fidelity Coronavirus Vaccine Protects in an Aged, Immunocompromised Mouse Model of Lethal Disease.” *Nature Medicine* 18 (12): 1820. https://doi.org/10.1038/nm.2972.

Greenbaum, Jason, John Sideny, Jolan Chung, Christian Brander, Bjoern Peter, and Alessandro Sette. 2011. “Functinal Classification of Class II Human Leukocyte Antigen (HLA) Molecules Reveals Seven Different Supertypes and a Surprising Degree of Repertoire Sharing across Supertypes.” *Immunogenetics* 63 (6): 325–35. https://doi.org/10.1007/s00251-011-0513-0.Functional.

Grifoni, Alba, John Sidney, Yun Zhang, Richard H. Scheuermann, Bjoern Peters, and Alessandro Sette. 2020. “A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2.” *Cell Host and Microbe* 27 (4): 671-680.e2.

Grifoni, Alba, Daniela Weiskopf, Sydney I Ramirez, Jose Mateus, Jennifer M Dan, Carolyn Rydyznski Moderbacher, Stephen A Rawlings, et al. 2020. “Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals.” *Cell* 181 (7): 1489–1501.

Guo, Li, Xi Zhang, Lili Ren, Xuelian Yu, Lijuan Chen, Hongli Zhou, Xin Gao, et al. 2014. “Human Antibody Responses to Avian Influenza A(H7N9) Virus, 2013.” *Emerging Infectious Diseases* 20 (2): 192–200.

Hall, B. G. 1978. “Experimental Evolution of a New Enzymatic Function. II. Evolution of Multiple Functions for EBG Enzyme in E. Coli.” *Genetics*.

Halling-Brown, Mark, Clare E. Sansom, Matthew Davies, Richard W. Titball, and David S. Moss. 2008. “Are Bacterial Vaccine Antigens T-Cell Epitope Depleted?” *Trends in Immunology* 29 (8): 374–79. https://doi.org/10.1016/j.it.2008.06.001.

Halling-Brown, Mark, Raheel Shaban, Dan Frampton, Clare E. Sansom, Matthew Davies, Darren Flower, Melanie Duffield, Richard W. Titball, Vladimir Brusic, and David S. Moss. 2009. “Proteins Accessible to Immune Surveillance Show Significant T-Cell Epitope Depletion: Implications for Vaccine Design.” *Molecular Immunology* 46 (13): 2699–2705. https://doi.org/10.1016/j.molimm.2009.05.027.

Hastings, Janna, Paula de Matos, Adriano Dekker, Marcus Ennis, Bhavana Harsha, Namrata Kale, Venkatesh Muthukrishnan, et al. 2013. “The ChEBI Reference Database and Ontology for Biologically Relevant Chemistry: Enhancements for 2013.” *Nucleic Acids Research* 41 (Database issue): D456-63. https://doi.org/10.1093/nar/gks1146.

Hatos, András, Borbála Hajdu-Soltész, Alexander M. Monzon, Nicolas Palopoli, Lucía Álvarez, Burcu Aykac-Fas, Claudio Bassot, et al. 2020. “DisProt: Intrinsic Protein Disorder Annotation in 2020.” *Nucleic Acids Research* 48 (D1): D269–76. https://doi.org/10.1093/nar/gkz975.

He, Yongqun, Fang Chen, Richard Scheuermann, Thomas Todd, Ryan Brinkman, Lesley Colby, Melanie Courtot, et al. 2009. “VO: Vaccine Ontology.” *Nature Precedings*, no. October: 1–2. https://doi.org/10.1038/npre.2009.3552.1.

He, Yongqun, Yue Liu, and Bin Zhao. 2014. “OGG: A Biological Ontology for Representing Genes and Genomes in Specific Organisms.” In *CEUR Workshop Proceedings*, 1327:13–20.

He, Yongqun, Edison Ong, Jennifer Schaub, Frederick Dowd, John F. O’Toole, Anastasios Siapos, Christian Reich, et al. 2019. “OPMI: The Ontology of Precision Medicine and Investigation and Its Support for Clinical Data and Metadata Representation and Analysis.” In *International Conference on Biomedical Ontology*.

He, Yongqun, Rebecca Racz, Samantha Sayers, Yu Lin, Thomas Todd, Junguk Hur, Xinna Li, et al. 2014. “Updates on the Web-Based VIOLIN Vaccine Database and Analysis System.” *Nucleic Acids Research* 42 (D1): 1124–32. https://doi.org/10.1093/nar/gkt1133.

He, Yongqun, Rino Rappuoli, Anne S De Groot, Robert T. Chen, Anne S. De Groot, and Robert T. Chen. 2010. “Emerging Vaccine Informatics.” *Journal of Biomedicine and Biotechnology* 2010: 1–26. https://doi.org/10.1155/2010/218590.

He, Yongqun, Becky Steck, Edison Ong, Laura Mariani, Chrysta Lienczewski, Ulysses J Balis, Matthias Kretzler, Jonathan Himmelfarb, John F Bertram, and Evren U Azeloglu. 2018. “KTAO: A Kidney Tissue Atlas Ontology to Support Community-Based Kidney Knowledge Base Development and Data Integration.” In *International Conference on Biomedical Ontology*.

He, Yongqun, and Zuoshuang Xiang. 2012. “Bioinformatics Analysis of Bacterial Protective Antigens in Manually Curated Protegen Database.” *Procedia in Vaccinology* 6: 3–9. https://doi.org/10.1016/j.provac.2012.04.002.

He, Yongqun, Zuoshuang Xiang, and Harry L T Mobley. 2010. “Vaxign: The First Web-Based Vaccine Design Program for Reverse Vaccinology and Applications for Vaccine Development.” *Journal of Biomedicine and Biotechnology* 2010: 1–15. https://doi.org/10.1155/2010/297505.

He, Yongqun, Zuoshuang Xiang, Jie Zheng, Yu Lin, James A. Overton, and Edison Ong. 2018. “The EXtensible Ontology Development (XOD) Principles and Tool Implementation to Support Ontology Interoperability.” *Journal of Biomedical Semantics*. https://doi.org/10.1186/s13326-017-0169-2.

Heinson, Ashley I., Yawwani Gunawardana, Bastiaan Moesker, Carmen C. Denman Hume, Elena Vataga, Yper Hall, Elena Stylianou, et al. 2017. “Enhancing the Biological Relevance of Machine Learning Classifiers for Reverse Vaccinology.” *International Journal of Molecular Sciences* 18 (2): 312. https://doi.org/10.3390/ijms18020312.

Henderson, Rory, Robert J Edwards, Katayoun Mansouri, Katarzyna Janowska, Victoria Stalls, Sophie M C Gobeil, Megan Kopp, et al. 2020. “Controlling the SARS-CoV-2 Spike Glycoprotein Conformation.” *Nature Structural & Molecular Biology* 27 (10): 925–33. https://doi.org/10.1038/s41594-020-0479-4.

Henikoff, S., and J. G. Henikoff. 1992. “Amino Acid Substitution Matrices from Protein Blocks.” *Proceedings of the National Academy of Sciences of the United States of America* 89 (22): 10915–19.

Hess, Berk, Henk Bekker, Herman J.C. Berendsen, and Johannes G.E.M. Fraaije. 1997. “LINCS: A Linear Constraint Solver for Molecular Simulations.” *Journal of Computational Chemistry* 18: 1463–72.

Hewitt, Joshua S., Anbu K. Karuppannan, Swan Tan, Phillip Gauger, Patrick G. Halbur, Priscilla F. Gerber, Anne S. De Groot, Leonard Moise, and Tanja Opriessnig. 2019. “A Prime-Boost Concept Using a T-Cell Epitope-Driven DNA Vaccine Followed by a Whole Virus Vaccine Effectively Protected Pigs in the Pandemic H1N1 Pig Challenge Model.” *Vaccine* 37 (31): 4302–9.

Hofmann, Heike, Krzysztof Pyrc, Lia Van Der Hoek, Martina Geier, Ben Berkhout, and Stefan Pöhlmann. 2005. “Human Coronavirus NL63 Employs the Severe Acute Respiratory Syndrome Coronavirus Receptor for Cellular Entry.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (22): 7988–93.

https://doi.org/10.1073/pnas.0409465102.

Hossain, Md Saddam, Abul Kalam Azad, Parveen Afroz Chowdhury, and Mamoru Wakayama. 2017. “Computational Identification and Characterization of a Promiscuous T-Cell Epitope on the Extracellular Protein 85B of Mycobacterium Spp. For Peptide-Based Subunit Vaccine Design.” *BioMed Research International* 2017. https://doi.org/10.1155/2017/4826030.

Houben, Rein M.G.J., and Peter J. Dodd. 2016. “The Global Burden of Latent Tuberculosis Infection: A Re-Estimation Using Mathematical Modelling.” *PLoS Medicine*. https://doi.org/10.1371/journal.pmed.1002152.

Huang, Da Wei, Richard a Lempicki, and Brad T Sherman. 2009. “Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources.” *Nature Protocols* 4 (1): 44–57. https://doi.org/10.1038/nprot.2008.211.

Huang, Xiaoqiang, Robin Pearce, and Yang Zhang. 2020a. “De Novo Design of Protein Peptides to Block Association of the SARS-CoV-2 Spike Protein with Human ACE2.” *Aging* 12 (12): 11263–76.

———. 2020b. “EvoEF2: Accurate and Fast Energy Function for Computational Protein Design.” *Bioinformatics* 36 (4): 1135–42.

Huang, Xiaoqiang, Chengxin Zhang, Robin Pearce, Gilbert S. Omenn, and Yang Zhang. 2020. “Identifying the Zoonotic Origin of SARS-CoV-2 by Modeling the Binding Affinity between the Spike Receptor-Binding Domain and Host ACE2.” *Journal of Proteome Research* 19 (12): 4844–56.

Huang, Yuan, Chan Yang, Xin feng Xu, Wei Xu, and Shu wen Liu. 2020. “Structural and Functional Properties of SARS-CoV-2 Spike Protein: Potential Antivirus Drug Development for COVID-19.” *Acta Pharmacologica Sinica* 41 (9): 1141–49. https://doi.org/10.1038/s41401-020-0485-4.

Huerta-Cepas, Jaime, Damian Szklarczyk, Kristoffer Forslund, Helen Cook, Davide Heller, Mathias C. Walter, Thomas Rattei, et al. 2016. “EGGNOG 4.5: A Hierarchical Orthology Framework with Improved Functional Annotations for Eukaryotic, Prokaryotic and Viral Sequences.” *Nucleic Acids Research* 44 (D1): D286–93. https://doi.org/10.1093/nar/gkv1248.

Huggins, Jennifer L., and R. John Looney. 2004. “Allergen Immunotherapy.” *American Family Physician*. https://doi.org/10.2165/00003495-199040040-00002.

Hur, Junguk, Arzucan Özgür, Zuoshuang Xiang, and Yongqun He. 2015. “Development and Application of an Interaction Network Ontology for Literature Mining of Vaccine-Associated Gene-Gene Interactions.” *Journal of Biomedical Semantics* 6 (1): 1–10. https://doi.org/10.1186/2041-1480-6-2.

Ip, Peng Peng, Annemarie Boerma, Joke Regts, Tjarko Meijerhof, Jan Wilschut, Hans W. Nijman, and Toos Daemen. 2014. “Alphavirus-Based Vaccines Encoding Nonstructural Proteins of Hepatitis c Virus Induce Robust and Protective T-Cell Responses.” *Molecular Therapy* 22 (4): 881–90. https://doi.org/10.1038/mt.2013.287.

Izzo, Angelo A. 2017. “Tuberculosis Vaccines — Perspectives from the NIH/NIAID

Mycobacteria Vaccine Testing Program." *Current Opinion in Immunology* 47: 78–84. https://doi.org/10.1016/j.coi.2017.07.008.

Jaiswal, Varun, Sree Krishna Chanumolu, Ankit Gupta, Rajinder S Chauhan, and Chittaranjan Rout. 2013. "Jenner-Predict Server: Prediction of Protein Vaccine Candidates (PVCs) in Bacteria Based on Host-Pathogen Interactions." *BMC Bioinformatics* 14 (1): 211. https://doi.org/10.1186/1471-2105-14-211.

Jaksik, Roman, Marta Iwanaszko, Joanna Rzeszowska-Wolny, and Marek Kimmel. 2015. "Microarray Experiments and Factors Which Affect Their Reliability." *Biology Direct*. https://doi.org/10.1186/s13062-015-0077-2.

Jensen, Kamilla Kjærgaard, Massimo Andreatta, Paolo Marcatili, Søren Buus, Jason A. Greenbaum, Zhen Yan, Alessandro Sette, Bjoern Peters, and Morten Nielsen. 2018. "Improved Methods for Predicting Peptide Binding Affinity to MHC Class II Molecules." *Immunology* 154 (3): 394–406.

Jespersen, Martin Closter, Bjoern Peters, Morten Nielsen, and Paolo Marcatili. 2017. "BepiPred-2.0: Improving Sequence-Based B-Cell Epitope Prediction Using Conformational Epitopes." *Nucleic Acids Research* 45 (W1): W24–29. https://doi.org/10.1093/nar/gkx346.

Jessica M. A. Blair, Mark A. Webber, Alison J. Baylay, David O. Ogbolu & Laura J. V. Piddock. 2015. "Molecular Mechanisms of Antibiotic Resistance." *Nature Reviews Microbiology* 13 (1): 42–51. https://doi.org/10.1039/c0cc05111j.

Jimenez de Bagues, M. P., P. H. Elzer, S. M. Jones, J. M. Blasco, F. M. Enright, G. G. Schurig, and A. J. Winter. 1994. "Vaccination with Brucella Abortus Rough Mutant RB51 Protects BALB/c Mice against Virulent Strains of Brucella Abortus, Brucella Melitensis, and Brucella Ovis." *Infection and Immunity* 62 (11): 4990–96.

Jorgensen, William L., Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. 1983. "Comparison of Simple Potential Functions for Simulating Liquid Water." *The Journal of Chemical Physics* 79: 926–35. https://doi.org/10.1063/1.445869.

Kalita, Parismita, Aditya K. Padhi, Kam Y.J. Zhang, and Timir Tripathi. 2020. "Design of a Peptide-Based Subunit Vaccine against Novel Coronavirus SARS-CoV-2." *Microbial Pathogenesis* 145. https://doi.org/10.1016/j.micpath.2020.104236.

Kibbe, Warren A, Cesar Arze, Victor Felix, Elvira Mitraka, Evan Bolton, Gang Fu, Christopher J Mungall, et al. 2015. "Disease Ontology 2015 Update: An Expanded and Updated Database of Human Diseases for Linking Biomedical Knowledge through Disease Data." *Nucleic Acids Research* 43 (Database issue): D1071-8. https://doi.org/10.1093/nar/gku1011.

King, Chris, Esteban N. Garza, Ronit Mazor, Jonathan L. Linehan, Ira Pastan, Marion Pepper, and David Baker. 2014. "Removing T-Cell Epitopes with Computational Protein Design." *Proceedings of the National Academy of Sciences* 111 (23): 8577–82.

Kling, Heather M., Gerard J. Nau, Ted M. Ross, Thomas G. Evans, Krishnendu Chakraborty, Kerry M. Empey, and Jo Anne L Flynn. 2014. "Challenges and Future in Vaccines, Drug Development, and Immunomodulatory Therapy." *Annals of the American Thoracic Society* 11: S201–10. https://doi.org/10.1513/AnnalsATS.201401-036PL.

Kohavi, Ron, and Foster Provost. 1998. "Glossary of Terms." *Machine Learning* 30 (2–3): 271–

74.

Korber, Bette, Will M. Fischer, Sandrasegaram Gnanakaran, Hyejin Yoon, James Theiler, Werner Abfalterer, Nick Hengartner, et al. 2020. “Tracking Changes in SARS-CoV-2 Spike: Evidence That D614G Increases Infectivity of the COVID-19 Virus.” *Cell* 182 (4): 812-827.e19. https://doi.org/10.1016/j.cell.2020.06.043.

Kourou, Konstantina, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. 2015. “Machine Learning Applications in Cancer Prognosis and Prediction.” *Computational and Structural Biotechnology Journal*. https://doi.org/10.1016/j.csbj.2014.11.005.

Krogh, Anders, Björn Larsson, Gunnar von Heijne, and Erik L.L Sonnhammer. 2001. “Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes.” *J Mol Biol* 305 (3): 567–80. https://doi.org/10.1006/jmbi.2000.4315.

La, Kaarina, Maini Kukkonen, Timo K Korhonen, Kaarina Lähteenmäki, Maini Kukkonen, and Timo K Korhonen. 2001. “The Pla Surface Protease/Adhesin of *Yersinia Pestis* Mediates Bacterial Invasion into Human Endothelial Cells.” *FEBS Letters* 504 (1–2): 69–72. https://doi.org/10.1016/S0014-5793(01)02775-2.

Lai, Chih-Cheng, Tzu-Ping Shih, Wen-Chien Ko, Hung-Jen Tang, and Po-Ren Hsueh. 2020. “Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and Coronavirus Disease-2019 (COVID-19): The Epidemic and the Challenges.” *International Journal of Antimicrobial Agents* 55 (3). https://doi.org/10.1016/j.ijantimicag.2020.105924.

Leaver-Fay, Andrew, Michael Tyka, Steven M Lewis, Oliver F Lange, James Thompson, Ron Jacak, Kristian Kaufman, et al. 2011. “ROSETTA3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules.” *Methods in Enzymology* 487: 545–74.

Lefort, Vincent, Jean Emmanuel Longueville, and Olivier Gascuel. 2017. “SMS: Smart Model Selection in PhyML.” *Molecular Biology and Evolution* 34 (9): 2422–24. https://doi.org/10.1093/molbev/msx149.

Lei, Jian, Yuri Kusov, and Rolf Hilgenfeld. 2018. “Nsp3 of Coronaviruses: Structures and Functions of a Large Multi-Domain Protein.” *Antiviral Research*. Elsevier B.V. https://doi.org/10.1016/j.antiviral.2017.11.001.

Leligdowicz, Aleksandra, William A. Fischer, Timothy M. Uyeki, Thomas E. Fletcher, Neill K. J. Adhikari, Gina Portella, Francois Lamontagne, et al. 2016. “Ebola Virus Disease and Critical Illness.” *Critical Care* 20 (1): 217. https://doi.org/10.1186/s13054-016-1325-2.

Letko, Michael, Andrea Marzi, and Vincent Munster. 2020. “Functional Assessment of Cell Entry and Receptor Usage for SARS-CoV-2 and Other Lineage B Betacoronaviruses.” *Nature Microbiology* 5 (4): 562–69. https://doi.org/10.1038/s41564-020-0688-y.

Li, Fang. 2016. “Structure, Function, and Evolution of Coronavirus Spike Proteins.” *Annual Review of Virology* 3: 237–61. https://doi.org/10.1146/annurev-virology-110615-042301.

Li, L., Christian J. Stoeckert, and David S. Roos. 2003. “OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes.” *Genome Research* 13 (9): 2178–89. https://doi.org/10.1101/gr.1224503.

Li, Qing, Xiaoqiang Huang, and Yushan Zhu. 2014. “Evaluation of Active Designs of Cephalosporin C Acylase by Molecular Dynamics Simulation and Molecular Docking.” *Journal of Molecular Modeling* 20: 2314.

Li, Yang, Jun Hu, Chengxin Zhang, Dong Jun Yu, and Yang Zhang. 2019. “ResPRE: High-Accuracy Protein Contact Prediction by Coupling Precision Matrix with Deep Residual Neural Networks.” *Bioinformatics* 35: 4647–55.

Li, Yang, Chengxin Zhang, Eric W. Bell, Dong Jun Yu, and Yang Zhang. 2019. “Ensembling Multiple Raw Coevolutionary Features with Deep Residual Neural Networks for Contact-Map Prediction in CASP13.” *Proteins: Structure, Function and Bioinformatics* 87: 1082–91.

Li, Zhongming, Angela Howard, Cynthia Kelley, Giovanni Delogu, Frank Collins, and Sheldon Morris. 1999. “Immunogenicity of DNA Vaccines Expressing Tuberculosis Proteins Fused to Tissue Plasminogen Activator Signal Sequences.” *Infection and Immunity* 67 (9): 4780–86.

Lin, Yu, and Yongqun He. 2012. “Ontology Representation and Analysis of Vaccine Formulation and Administration and Their Effects on Vaccine Immune Responses.” *Journal of Biomedical Semantics* 3 (1). https://doi.org/10.1186/2041-1480-3-17.

Lin, Yu, Zuoshuang Xiang, and Yongqun He. 2011. “Brucellosis Ontology (IDOBRU) as an Extension of the Infectious Disease Ontology.” *Journal of Biomedical Semantics* 2 (1). https://doi.org/10.1186/2041-1480-2-9.

Lin, Zong, and Xian Ming Pan. 2001. “Accurate Prediction of Protein Secondary Structural Content.” *Protein Journal* 20 (3): 217–20. https://doi.org/10.1023/A:1010967008838.

Lopera-Madrid, Jaime, Jorge E. Osorio, Yongqun He, Zuoshuang Xiang, L. Garry Adams, Richard C. Laughlin, Waithaka Mwangi, et al. 2017. “Safety and Immunogenicity of Mammalian Cell Derived and Modified-Vaccinia Ankara Vectored African Swine Fever Subunit Antigens in Swine.” *Veterinary Immunology and Immunopathology* 185: 20–33. https://doi.org/10.1016/j.vetimm.2017.01.004.

Lu, Roujian, Xiang Zhao, Juan Li, Peihua Niu, Bo Yang, Honglong Wu, Wenling Wang, et al. 2020. “Genomic Characterisation and Epidemiology of 2019 Novel Coronavirus: Implications for Virus Origins and Receptor Binding.” *The Lancet* 395 (10224): 565–74.

Luabeya, Angelique Kany Kany, Benjamin M.N. N Kagina, Michele D. Tameris, Hennie Geldenhuys, Soren T. Hoff, Zhongkai Shi, Ingrid Kromann, et al. 2015. “First-in-Human Trial of the Post-Exposure Tuberculosis Vaccine H56: IC31 in Mycobacterium Tuberculosis Infected and Non-Infected Healthy Adults.” *Vaccine* 33 (33): 4130–40. https://doi.org/10.1016/j.vaccine.2015.06.051.

Lynch, Marina A., and Kingston H.G. Mills. 2012. “Immunology Meets Neuroscience - Opportunities for Immune Intervention in Neurodegenerative Diseases.” *Brain, Behavior, and Immunity*. https://doi.org/10.1016/j.bbi.2011.05.013.

Magnan, Christophe N., Michael Zeller, Matthew A. Kayala, Adam Vigil, Arlo Randall, Philip L. Felgner, and Pierre Baldi. 2010. “High-Throughput Prediction of Protein Antigenicity Using Protein Microarray Data.” *Bioinformatics* 26 (23): 2936–43. https://doi.org/10.1093/bioinformatics/btq551.

Mayers, Carl, Melanie Duffield, Sonya Rowe, Julie Miller, Bryan Lingard, Sarah Hayward, and Richard W. Titball. 2003. "Analysis of Known Bacterial Protein Vaccine Antigens Reveals Biased Physical Properties and Amino Acid Composition." *Comparative and Functional Genomics* 4 (5): 468–78. https://doi.org/10.1002/cfg.319.

McKay, Paul F., Kai Hu, Anna K. Blakney, Karnyart Samnuan, Clement R. Bouton, Paul Rogers, Krunal Polra, et al. 2020. "Self-Amplifying RNA SARS-CoV-2 Lipid Nanoparticle Vaccine Induces Equivalent Preclinical Antibody Titers and Viral Neutralization to Recovered COVID-19 Patients." BioRxiv [Preprint]. 2020. https://www.biorxiv.org/content/10.1101/2020.04.22.055608v1.

McLellan, Jason S., Man Chen, M. Gordon Joyce, Mallika Sastry, Guillaume B. E. Stewart-Jones, Yongping Yang, Baoshan Zhang, et al. 2013. "Structure-Based Design of a Fusion Glycoprotein Vaccine for Respiratory Syncytial Virus." *Science (New York, N.Y.)* 342 (6158): 592–98.

Millet, P., G. H. Campbell, A. J. Sulzer, K. K. Grady, J. Pohl, M. Aikawa, and W. E. Collins. 1993. "Immunogenicity of the Plasmodium Falciparum Asexual Blood-Stage Synthetic Peptide Vaccine SPf66." *American Journal of Tropical Medicine and Hygiene* 48 (3): 424–31. https://doi.org/10.4269/ajtmh.1993.48.424.

Moise, L, a Gutierrez, F Kibria, R Martin, R Tassone, R Liu, F Terry, B Martin, and a S De Groot. 2015. "IVAX: An Integrated Toolkit for the Selection and Optimization of Antigens and the Design of Epitope-Driven Vaccines." *Hum Vaccin Immunother* 11 (9): 2312–21. https://doi.org/10.1080/21645515.2015.1061159.

Moise, Leonard, Andres H Gutierrez, Chris Bailey-kellogg, Frances Terry, Qibin Leng, Karim M Abdel Hady, Nathan C Verberkmoes, et al. 2013. "The Two-Faced T Cell Epitope: Examining the Host-Microbe Interface with JanusMatrix." *Human Vaccines & Immunotherapeutics* 9 (7): 1577–86.

Morozova, N., J. Myers, and Y. Shamoo. 2006. "Protein-RNA Interactions: Exploring Binding Patterns with a Three-Dimensional Superposition Analysis of High Resolution Structures." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btl470.

Muller-Sieburg, Christa E., Hans B. Sieburg, Jeff M. Bernitz, and Giulio Cattarossi. 2012. "Stem Cell Heterogeneity: Implications for Aging and Regenerative Medicine." *Blood*. https://doi.org/10.1182/blood-2011-12-376749.

Mulligan, Mark J., David I. Bernstein, Patricia Winokur, Richard Rupp, Evan Anderson, Nadine Rouphael, Michelle Dickey, et al. 2014. "Serological Responses to an Avian Influenza A/H7N9 Vaccine Mixed at the Point-of-Use with MF59 Adjuvant a Randomized Clinical Trial." *JAMA - Journal of the American Medical Association* 312 (14): 1409–19.

Mungall, Christopher J, Carlo Torniai, Georgios V Gkoutos, Suzanna E Lewis, and Melissa A Haendel. 2012. "Uberon, an Integrative Multi-Species Anatomy Ontology." *Genome Biology* 13 (1): R5. https://doi.org/10.1186/gb-2012-13-1-r5.

Natale, Darren a., Cecilia N. Arighi, Judith a. Blake, Carol J. Bult, Karen R. Christie, Julie Cowart, Peter D'Eustachio, et al. 2014. "Protein Ontology: A Controlled Structured Network of Protein Entities." *Nucleic Acids Research* 42 (D1): 415–21. https://doi.org/10.1093/nar/gkt1173.

Navarro-Quiroz, Elkin, Roberto Navarro-Quiroz, Pierine España-Puccini, José Luis Villarreal, Anderson Díaz Perez, Cecilia Fernandez Ponce, Jorge Bilbao, Lucy Vasquez, and Dary Luz Mendoza. 2018. "Prediction of Epitopes in the Proteome of Helicobacter Pylori." *Global Journal of Health Science* 10 (7): 148. https://doi.org/10.5539/gjhs.v10n7p148.

Ng, Vincent H., Jeffery S. Cox, Alexandra O. Sousa, John D. MacMicking, and John D. McKinney. 2004. "Role of KatG Catalase-Peroxidase in Mycobacterial Pathogenisis: Countering the Phagocyte Oxidative Burst." *Molecular Microbiology* 52 (5): 1291–1302. https://doi.org/10.1111/j.1365-2958.2004.04078.x.

Ni, Ling, Fang Ye, Meng-Li Cheng, Yu Feng, Yong-Qiang Deng, Hui Zhao, Peng Wei, et al. 2020. "Detection of SARS-CoV-2-Specific Humoral and Cellular Immunity in COVID-19 Convalescent Individuals." *Immunity* 52 (6): 971–77.

Ni, Zhaohui, Yan Chen, Edison Ong, and Yongqun He. 2017. "Antibiotic Resistance Determinant-Focused Acinetobacter Baumannii Vaccine Designed Using Reverse Vaccinology." *International Journal of Molecular Sciences* 18 (2): 458. https://doi.org/10.3390/ijms18020458.

Nicholas, Dequina, Oludare Odumosu, and William H.R. Langridge. 2011. "Autoantigen Based Vaccines for Type 1 Diabetes." *Discovery Medicine*.

Nikolich-Žugich, Janko, Mark K. Slifka, and Ilhem Messaoudi. 2004. "The Many Important Facets of T-Cell Repertoire Diversity." *Nature Reviews Immunology* 4 (2): 123–32. https://doi.org/10.1038/nri1292.

Ong, Edison, and Yongqun He. 2015. "GOfox: Semantics-Based Simplified Hierarchical Classification and Interactive Visualization to Support GO Enrichment Analysis." In *Proceedings of the International Conference on Biomedical Ontology*.

———. 2019. "Ontology of Host-Pathogen Interactions and Its Usage in Identification and Analysis of Virulence Factors Also Serving the Role of Vaccine Protective Antigens." In *International Conference on Biomedical Ontology*.

Ong, Edison, Yongqun He, and Zhenhua Yang. 2020. "Epitope Promiscuity and Population Coverage of Mycobacterium Tuberculosis Protein Antigens in Current Subunit Vaccines under Development." *Infection, Genetics and Evolution* 80 (August 2019): 104186. https://doi.org/10.1016/j.meegid.2020.104186.

Ong, Edison, Xiaoqiang Huang, Robin Pearce, Yang Zhang, and Yongqun He. 2021. "Computational Design of SARS-CoV-2 Spike Glycoproteins to Increase Immunogenicity by T Cell Epitope Engineering." *Computational and Structural Biotechnology Journal* 19 (December): 518–29. https://doi.org/10.1016/j.csbj.2020.12.039.

Ong, Edison, Peter Sun, Kimberly Berke, Jie Zheng, Guanming Wu, and Yongqun He. 2019. "VIO: Ontology Classification and Study of Vaccine Responses given Various Experimental and Analytical Conditions." *BMC Bioinformatics* 20 (21): 1–10. https://doi.org/10.1186/s12859-019-3194-6.

Ong, Edison, Haihe Wang, Mei U Wong, Meenakshi Seetharaman, Ninotchka Valdez, and Yongqun He. 2020. "Vaxign-ML: Supervised Machine Learning Reverse Vaccinology Model for Improved Prediction of Bacterial Protective Antigens." *Bioinformatics* 36 (10): 3185–91. https://doi.org/10.1093/bioinformatics/btaa119.

Ong, Edison, Lucy L. Wang, Jennifer Schaub, John F. O'Toole, Becky Steck, Avi Z. Rosenberg, Frederick Dowd, et al. 2020. "Modelling Kidney Disease Using Ontology: Insights from the Kidney Precision Medicine Project." *Nature Reviews Nephrology*. https://doi.org/10.1038/s41581-020-00335-w.

Ong, Edison, Mei U Wong, and Yongqun He. 2017. "Identification of New Features from Known Bacterial Protective Vaccine Antigens Enhances Rational Vaccine Design." *Frontiers in Immunology* 8 (October): 1–11. https://doi.org/10.3389/fimmu.2017.01382.

Ong, Edison, Mei U Wong, Anthony Huffman, and Yongqun He. 2020. "COVID-19 Coronavirus Vaccine Design Using Reverse Vaccinology and Machine Learning." *Frontiers in Immunology* 11: 1581.

Ong, Edison, Zuoshuang Xiang, Bin Zhao, Yue Liu, Yu Lin, Jie Zheng, Chris Mungall, Mélanie Courtot, Alan Ruttenberg, and Yongqun He. 2017. "Ontobee: A Linked Ontology Data Server to Support Ontology Term Dereferencing, Linkage, Query and Integration." *Nucleic Acids Research* 45 (D1): D347–52. https://doi.org/10.1093/nar/gkw918.

Ong, Serene A.K., Hong Huang Lin, Yu Zong Chen, Ze Rong Li, and Zhiwei Cao. 2007. "Efficacy of Different Protein Descriptors in Predicting Protein Functional Families." *BMC Bioinformatics* 8: 1–14. https://doi.org/10.1186/1471-2105-8-300.

Özgür, Arzucan, Zuoshuang Xiang, Dragomir R. Radev, and Yongqun He. 2011. "Mining of Vaccine-Associated IFN-γ Gene Interaction Networks Using the Vaccine Ontology." *Journal of Biomedical Semantics*. https://doi.org/10.1186/2041-1480-2-S2-S8.

Padhi, Aditya K., Parismita Kalita, Kam Y. J. Zhang, and Timir Tripathi. 2020. "High Throughput Designing and Mutational Mapping of RBD-ACE2 Interface Guide Non-Conventional Therapeutic Strategies for COVID-19." *BioRxiv [Preprint]*. https://doi.org/10.1101/2020.05.19.104042.

Padhi, Aditya K., and Timir Tripathi. 2020. "Can SARS-CoV-2 Accumulate Mutations in the S-Protein to Increase Pathogenicity?" *ACS Pharmacology and Translational Science*. https://doi.org/10.1021/acsptsci.0c00113.

Panhuis, Willem G. van, John Grefenstette, Su Yon Jung, Nian Shong Chok, Anne Cross, Heather Eng, Bruce Y. Lee, et al. 2013. "Contagious Diseases in the United States from 1888 to the Present." *New England Journal of Medicine* 369 (22): 2152–58. https://doi.org/10.1056/NEJMms1215400.

Parrinello, M., and A. Rahman. 1981. "Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method." *Journal of Applied Physics* 52 (12): 7182–90.

Patel, Ami, Jewell Walters, Emma L. Reuschel, Katherine Schultheis, Elizabeth Parzych, Ebony N. Gary, Igor Maricic, et al. 2020. "Intradermal-Delivered DNA Vaccine Provides Anamnestic Protection in a Rhesus Macaque SARS-CoV-2 Challenge Model." *BioRxiv [Preprint]*. https://www.biorxiv.org/content/10.1101/2020.07.28.225649v1.

Patel, Seema, Nithya Mathivanan, and Arun Goyal. 2017. "Bacterial Adhesins, the Pathogenic Weapons to Trick Host Defense Arsenal." *Biomedicine & Pharmacotherapy* 93: 763–71. https://doi.org/10.1016/j.biopha.2017.06.102.

Paul, Sinu, Cecilia S. Lindestam Arlehamn, Thomas J. Scriba, Myles B.C. Dillon, Carla Oseroff, Denise Hinz, Denise M. McKinney, et al. 2015. "Development and Validation of a Broad Scheme for Prediction of HLA Class II Restricted T Cell Epitopes." *Journal of Immunological Methods* 422: 28–34.

Pearce, Robin, Xiaoqiang Huang, Dani Setiawan, and Yang Zhang. 2019. "EvoDesign: Designing Protein–Protein Binding Interactions Using Evolutionary Interface Profiles in Conjunction with an Optimized Physical Energy Function." *Journal of Molecular Biology* 431 (13): 2467–76.

Pearson, William R. 2013. "An Introduction to Sequence Similarity ('homology') Searching." *Current Protocols in Bioinformatics* 42: 3–1. https://doi.org/10.1002/0471250953.bi0301s42.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2012. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30. https://doi.org/10.1007/s13398-014-0173-7.2.

Penn-Nicholson, Adam, Michele Tameris, Erica Smit, Tracey A. Day, Munyaradzi Musvosvi, Lakshmi Jayashankar, Julie Vergara, et al. 2018. "Safety and Immunogenicity of the Novel Tuberculosis Vaccine ID93 + GLA-SE in BCG-Vaccinated Healthy Adults in South Africa: A Randomised, Double-Blind, Placebo-Controlled Phase 1 Trial." *The Lancet Respiratory Medicine* 6 (4): 287–98. https://doi.org/10.1016/S2213-2600(18)30077-8.

Perlman, Stanley, and Jason Netland. 2009. "Coronaviruses Post-SARS: Update on Replication and Pathogenesis." *Nature Reviews Microbiology* 7 (6): 439–50. https://doi.org/10.1038/nrmicro2147.

Petersen, Thomas Nordahl, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2011. "SignalP 4.0: Discriminating Signal Peptides from Transmembrane Regions." *Nature Methods* 8 (10): 785–86. https://doi.org/10.1038/nmeth.1701.

Pizza, Mariagrazia, Vincenzo Scarlato, Vega Masignani, Marzia Monica Giuliani, Beatrice Aricò, Maurizio Comanducci, Gary T Jennings, et al. 2000. "Identification of Vaccine Candidates Against Serogroup B Meningococcus by Whole-Genome Sequencing." *Science* 287 (5459): 1816–20. https://doi.org/10.1126/science.287.5459.1816.

Plotkin, S A, W A Orenstein, and P A Offit. 2012. *Vaccines*. ClinicalKey 2012. Elsevier/Saunders. https://books.google.com/books?id=hoigDQ6vdDQC.

Plotkin, Stanley A. 2005. "Vaccines: Past, Present and Future." *Nature Medicine*. https://doi.org/10.1038/nm1209.

———. 2020. "Updates on Immunologic Correlates of Vaccine-Induced Protection." *Vaccine* 38 (9): 2250–57. https://doi.org/10.1016/j.vaccine.2019.10.046.

Polack, Fernando P, Stephen J Thomas, Nicholas Kitchin, Judith Absalon, Alejandra Gurtman, Stephen Lockhart, John L Perez, et al. 2020. "Safety and Efficacy of the BNT162b2 MRNA Covid-19 Vaccine." *The New England Journal of Medicine,* December. https://doi.org/10.1056/NEJMoa2034577.

Pulendran, Bali. 2009. "Learning Immunology from the Yellow Fever Vaccine: Innate Immunity to Systems Vaccinology." *Nature Reviews Immunology*. https://doi.org/10.1038/nri2629.

Punta, M, Pc Coggill, Ry Eberhardt, J Mistry, J Tate, C Boursnell, N Pang, et al. 2012. “The Pfam Protein Families Databases.” *Nucleic Acids Res 40: D290-D301*. 30 (1): 1–12. https://doi.org/10.1093/nar/gkp985.

Querec, Troy D., Rama S. Akondy, Eva K. Lee, Weiping Cao, Helder I. Nakaya, Dirk Teuwen, Ali Pirani, et al. 2009. “Systems Biology Approach Predicts Immunogenicity of the Yellow Fever Vaccine in Humans.” *Nature Immunology* 10 (1): 116–25. https://doi.org/10.1038/ni.1688.

Rahman, M. Saifur, Md Khaledur Rahman, Sanjay Saha, M. Kaykobad, and M. Sohel Rahman. 2019. “Antigenic: An Improved Prediction Model of Protective Antigens.” *Artificial Intelligence in Medicine* 94 (May 2018): 28–41. https://doi.org/10.1016/j.artmed.2018.12.010.

Ramos, Hugo Cruz, Martin Rumbo, and Jean Claude Sirard. 2004. “Bacterial Flagellins: Mediators of Pathogenicity and Host Immune Responses in Mucosa.” *Trends in Microbiology* 12 (11): 509–17. https://doi.org/10.1016/j.tim.2004.09.002.

Rappuoli, Rino. 2000. “Reverse Vaccinology.” *Curr Opin Microbiol* 3: 445–50. https://doi.org/10.1016/S1369-5274(00)00119-3.

Rappuoli, Rino, Mariagrazia Pizza, Giuseppe Del Giudice, and Ennio De Gregorio. 2014. “Vaccines, New Opportunities for a New Society.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (34): 12288–93. http://www.pnas.org/content/111/34/12288.full?sid=d1498d3b-c95a-4988-9998-a80d15a735d7.

Redelman-Sidi, Gil. 2020. “Could BCG Be Used to Protect against COVID-19?” *Nature Reviews Urology* 17: 316–17. https://doi.org/10.1038/s41585-020-0325-9.

Ribet, David, and Pascale Cossart. 2015. “How Bacterial Pathogens Colonize Their Hosts and Invade Deeper Tissues.” *Microbes and Infection* 17 (3): 173–83. https://doi.org/10.1016/j.micinf.2015.01.004.

Rizwan, Muhammad, Anam Naz, Jamil Ahmad, Kanwal Naz, Ayesha Obaid, Tamsila Parveen, Muhammad Ahsan, and Amjad Ali. 2017. “VacSol: A High Throughput in Silico Pipeline to Predict Potential Therapeutic Targets in Prokaryotic Pathogens Using Subtractive Reverse Vaccinology.” *BMC Bioinformatics* 18 (1): 1–7. https://doi.org/10.1186/s12859-017-1540-0.

Roper, Rachel L., and Kristina E. Rehm. 2009. “SARS Vaccines: Where Are We?” *Expert Review of Vaccines* 8 (7): 887–98. https://doi.org/10.1586/erv.09.43.

Rothbard, J B, and W R Taylor. 1988. “A Sequence Pattern Common to T Cell Epitopes.” *The EMBO Journal* 7 (1): 93–100. https://doi.org/10.1002/j.1460-2075.1988.tb02787.x.

Roukens, Anna H., and Leo G. Visser. 2008. “Yellow Fever Vaccine: Past, Present and Future.” *Expert Opinion on Biological Therapy* 8 (11): 1787–95. https://doi.org/10.1517/14712598.8.11.1787.

Rudenko, Larisa, Irina Isakova-Sivak, Anatoly Naykhin, Irina Kiseleva, Marina Stukova, Mariana Erofeeva, Daniil Korenkov, Victoria Matyushenko, Erin Sparrow, and Marie Paule Kieny. 2016. “H7N9 Live Attenuated Influenza Vaccine in Healthy Adults: A Randomised, Double-Blind, Placebo-Controlled, Phase 1 Trial." *The Lancet Infectious Diseases* 16 (3): 303–10.

Sachdeva, Gaurav, Kaushal Kumar, Preti Jain, and Srinivasan Ramachandran. 2005. "SPAAN: A Software Program for Prediction of Adhesins and Adhesin-like Proteins Using Neural Networks." *Bioinformatics* 21 (4): 483–91. https://doi.org/10.1093/bioinformatics/bti028.

Saiz, Juan Carlos, Angela Vazquez-Calvo, Ana B. Blazquez, Teresa Merino-Ramos, Estela Escribano-Romero, and Miguel A. Martín-Acebes. 2016. "Zika Virus: The Latest Newcomer." *Frontiers in Microbiology* 7 (APR): 1–19. https://doi.org/10.3389/fmicb.2016.00496.

Salat, Jiri, Kamil Mikulasek, Osmany Larralde, Petra Pokorna Formanova, Ales Chrdle, Jan Haviernik, Jana Elsterova, et al. 2020. "Tick-Borne Encephalitis Virus Vaccines Contain Non-Structural Protein 1 Antigen and May Elicit NS1-Specific Antibody Responses in Vaccinated Individuals." *Vaccines* 8 (1): 81. https://doi.org/10.3390/vaccines8010081.

Samuel, A. L. 2000. "Some Studies in Machine Learning Using the Game of Checkers." *IBM Journal of Research and Development*. https://doi.org/10.1147/rd.441.0206.

Sarntivijai, Sirarat, Yu Lin, Zuoshuang Xiang, Terrence F Meehan, Alexander D Diehl, Uma D Vempati, Stephan C Schürer, et al. 2014. "CLO: The Cell Line Ontology." *Journal of Biomedical Semantics* 5 (1): 37. https://doi.org/10.1186/2041-1480-5-37.

Sayers, Samantha, Li Li, Edison Ong, Shunzhou Deng, Guanghua Fu, Yu Lin, Brian Yang, et al. 2019. "Victors: A Web-Based Knowledge Base of Virulence Factors in Human and Animal Pathogens." *Nucleic Acids Research* 47 (D1): D693–700. https://doi.org/10.1093/nar/gky999.

Scherer, Christina A., Charles L. Magness, Kathryn V. Steiger, Nicholas D. Poitinger, Christine M. Caputo, Douglas G. Miner, Patricia L. Winokur, et al. 2007. "Distinct Gene Expression Profiles in Peripheral Blood Mononuclear Cells from Patients Infected with Vaccinia Virus, Yellow Fever 17D Virus, or Upper Respiratory Infections." *Vaccine* 25 (35): 6458–73. https://doi.org/10.1016/j.vaccine.2007.06.035.

Schlesinger, J. J., M. W. Brandriss, and E. E. Walsh. 1985. "Protection against 17D Yellow Fever Encephalitis in Mice by Passive Transfer of Monoclonal Antibodies to the Nonstructural Glycoprotein Gp48 and by Active Immunization with Gp48." *Journal of Immunology* 135 (4): 2805–9.

Schlom, Jeffrey, Matteo Vergati, Chiara Intrivici, Ngar Yee Huen, and Kwong Y. Tsang. 2010. "Strategies for Cancer Vaccine Development." *Journal of Biomedicine and Biotechnology*. https://doi.org/10.1155/2010/596432.

Schrödinger, LLC. 2015. "The PyMol Molecular Graphics System, Version~1.8." 2015. https://pymol.org.

Schulz, Georg E. 2002. "The Structure of Bacterial Outer Membrane Proteins." *Biochimica et Biophysica Acta* 1565 (2): 308–17. https://doi.org/S0005273602005771 [pii].

Sealy, Robert, Karen S. Slobod, Patricia Flynn, Kristen Branum, Sherri Surman, Bart Jones, Pamela Freiden, Timothy Lockey, Nanna Howlett, and Julia L. Hurwitz. 2009. "Preclinical and Clinical Development of a Multi-Envelope, DNA-Virus-Protein (D-V-P) HIV-1

Vaccine." *International Reviews of Immunology* 28 (1–2): 49–68. https://doi.org/10.1080/08830180802495605.

See, Raymond H., Martin Petric, David J. Lawrence, Catherine P.Y. Mok, Thomas Rowe, Lois A. Zitzow, Karuna P. Karunakaran, et al. 2008. "Severe Acute Respiratory Syndrome Vaccine Efficacy in Ferrets: Whole Killed Virus and Adenovirus-Vectored Vaccines." *Journal of General Virology* 89 (9): 2136–46. https://doi.org/10.1099/vir.0.2008/001891-0.

See, Raymond H., Alexander N. Zakhartchouk, Martin Petric, David J. Lawrence, Catherine P.Y. Mok, Robert J. Hogan, Thomas Rowe, et al. 2006. "Comparative Evaluation of Two Severe Acute Respiratory Syndrome (SARS) Vaccine Candidates in Mice Challenged with SARS Coronavirus." *Journal of General Virology* 87 (3): 641–50. https://doi.org/10.1099/vir.0.81579-0.

Shi, Shu Qun, Jing Pian Peng, Yin Chuan Li, Chuan Qin, Guo Dong Liang, Li Xu, Ying Yang, Jin Ling Wang, and Quan Hong Sun. 2006. "The Expression of Membrane Protein Augments the Specific Responses Induced by SARS-CoV Nucleocapsid DNA Immunization." *Molecular Immunology* 43 (11): 1791–98. https://doi.org/10.1016/j.molimm.2005.11.005.

Shim, Byoung Shik, Sung Moo Park, Ji Shan Quan, Dhananjay Jere, Hyuk Chu, Man K. Song, Dong W. Kim, et al. 2010. "Intranasal Immunization with Plasmid DNA Encoding Spike Protein of SARS-Coronavirus/Polyethylenimine Nanoparticles Elicits Antigen-Specific Humoral and Cellular Immune Responses." *BMC Immunology* 11: 65.

Shin, Donghyuk, Rukmini Mukherjee, Diana Grewe, Denisa Bojkova, Kheewoong Baek, Anshu Bhattacharya, Laura Schulz, et al. 2020. "Papain-like Protease Regulates SARS-CoV-2 Viral Spread and Innate Immunity." *Nature*. https://doi.org/10.1038/s41586-020-2601-5.

Singh, Ravinder, Nisha Garg, Geeta Shukla, Neena Capalash, and Prince Sharma. 2016. "Immunoprotective Efficacy of Acinetobacter Baumannii Outer Membrane Protein, FilF, Predicted In Silico as a Potential Vaccine Candidate." *Frontiers in Microbiology* 7 (February): 158. https://doi.org/10.3389/fmicb.2016.00158.

Smaili, Fatima Zohra, Xin Gao, and Robert Hoehndorf. 2018. "Onto2Vec: Joint Vector-Based Representation of Biological Entities and Their Ontology-Based Annotations." In *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bty259.

———. 2019. "OPA2Vec: Combining Formal and Informal Content of Biomedical Ontologies to Improve Similarity-Based Prediction." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bty933.

Smith, Barry, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, et al. 2007. "The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration." *Nature Biotechnology* 25 (11): 1251–55. https://doi.org/10.1038/nbt1346.

Smith, Barry, Werner Ceusters, Bert Klagges, Jacob Köhler, Anand Kumar, Jane Lomax, Chris Mungall, Fabian Neuhaus, Alan L Rector, and Cornelius Rosse. 2005. "Relations in Biomedical Ontologies." *Genome Biology* 6 (5): R46. https://doi.org/10.1186/gb-2005-6-5-r46.

Smyth, Gordon K. 2004. "Linear Models and Empirical Bayes Methods for Assessing

Differential Expression in Microarray Experiments." *Statistical Applications in Genetics and Molecular Biology* 3 (1). https://doi.org/10.2202/1544-6115.1027.

Sokal, Robert R., and Barbara A. Thomson. 2006. "Population Structure Inferred by Local Spatial Autocorrelation: An Example from an Amerindian Tribal Population." *American Journal of Physical Anthropology* 129 (1): 121–31. https://doi.org/10.1002/ajpa.20250.

Stall, Shelley, Lynn Yarmey, Joel Cutcher-Gershenfeld, Brooks Hanson, Kerstin Lehnert, Brian Nosek, Mark Parsons, Erin Robinson, and Lesley Wyborn. 2019. "Make Scientific Data FAIR." *Nature*. https://doi.org/10.1038/d41586-019-01720-7.

Suliman, Sara, Angelique Kany Kany Luabeya, Hennie Geldenhuys, Michele Tameris, Soren T. Hoff, Zhongkai Shi, Dereck Tait, et al. 2019. "Dose Optimization of H56:IC31 Vaccine for Tuberculosis-Endemic Populations a Double-Blind, Placebo-Controlled, Dose-Selection Trial." *American Journal of Respiratory and Critical Care Medicine* 199 (2): 220–31. https://doi.org/10.1164/rccm.201802-0366OC.

Sun, Yaohui, Yanwen Li, Rachel M. Exley, Megan Winterbotham, Catherine Ison, Harry Smith, and Christoph M. Tang. 2005. "Identification of Novel Antigens That Protect against Systemic Meningococcal Infection." *Vaccine* 23 (32): 4136–41. https://doi.org/10.1016/j.vaccine.2005.03.015.

Tang, Fang, Yan Quan, Zhong-Tao Xin, Jens Wrammert, Mai-Juan Ma, Hui Lv, Tian-Bao Wang, et al. 2011. "Lack of Peripheral Memory B Cell Responses in Recovered Patients with Severe Acute Respiratory Syndrome: A Six-Year Follow-Up Study." *The Journal of Immunology* 186 (12): 7264–68. https://doi.org/10.4049/jimmunol.0903490.

Tatusov, R L, M Y Galperin, D A Natale, and E V Koonin. 2000. "The COG Database: A Tool for Genome-Scale Analysis of Protein Functions and Evolution." *Nucleic Acids Research* 28 (1): 33–36. https://doi.org/10.1093/nar/28.1.33.

Tay, Matthew Zirui, Chek Meng Poh, Laurent Rénia, Paul A. MacAry, and Lisa F.P. Ng. 2020. "The Trinity of COVID-19: Immunity, Inflammation and Intervention." *Nature Reviews Immunology* 20 (6): 363–74.

Temperton, Nigel J., Paul K. Chan, Graham Simmons, Maria C. Zambon, Richard S. Tedder, Yasuhiro Takeuchi, and Robin A. Weiss. 2005. "Longitudinally Profiling Neutralizing Antibody Response to SARS Coronavirus with Pseudotypes." *Emerging Infectious Diseases* 11 (3): 411–16.

"The Information Artifact Ontology (IAO)." n.d. Accessed September 30, 2019. https://github.com/information-artifact-ontology/IAO.

"The Ontology for General Medical Science (OGMS)." n.d. Accessed September 30, 2019. https://github.com/OGMS/ogms.

The UniProt Consortium. 2008. "The Universal Protein Resource (UniProt)." *Nucleic Acids Research* 36 (35): D193-197.

Theiler, Max, and Hugh H. Smith. 1937. "The Use of Yellow Fever Virus Modified by in Vitro Cultivation for Human Immunization." *Journal of Experimental Medicine* 65 (6): 787–800. https://doi.org/10.1084/jem.65.6.787.

Tian, Ye, Xiaoqiang Huang, and Yushan Zhu. 2015. "Computational Design of Enzyme–Ligand

Binding Using a Combined Energy Function and Deterministic Sequence Optimization Algorithm." *Journal of Molecular Modeling* 21 (8): 191.

Todd, T E, O Tibi, Y Lin, S Sayers, D N Bronner, Z Xiang, and Y He. 2013. "Meta-Analysis of Variables Affecting Mouse Protection Efficacy of Whole Organism Brucella Vaccines and Vaccine Candidates." *BMC Bioinformatics* 14 Suppl 6 (Suppl 6): S3. https://doi.org/10.1186/1471-2105-14-s6-s3.

Traggiai, Elisabetta, Stephan Becker, Kanta Subbarao, Larissa Kolesnikova, Yasushi Uematsu, Maria Rita Gismondo, Brian R. Murphy, Rino Rappuoli, and Antonio Lanzavecchia. 2004. "An Efficient Method to Make Human Monoclonal Antibodies from Memory B Cells: Potent Neutralization of SARS Coronavirus." *Nature Medicine* 10 (8): 871–75. https://doi.org/10.1038/nm1080.

Tran, Trish P, Edison Ong, Andrew P Hodges, Giovanni Paternostro, and Carlo Piermarocchi. 2014. "Prediction of Kinase Inhibitor Response Using Activity Profiling, in Vitro Screening, and Elastic Net Regression." *BMC Systems Biology* 8 (1): 74. https://doi.org/10.1186/1752-0509-8-74.

Uplekar, Mukund, Diana Weil, Knut Lonnroth, Ernesto Jaramillo, Christian Lienhardt, Hannah Monica Dias, Dennis Falzon, et al. 2015. "WHO's New End TB Strategy." *The Lancet* 385 (9979): 1799–1801. https://doi.org/10.1016/s0140-6736(15)60570-0.

Urban, Martin, Rashmi Pant, Arathi Raghunath, Alistair G. Irvine, Helder Pedro, and Kim E. Hammond-Kosack. 2015. "The Pathogen-Host Interactions Database (PHI-Base): Additions and Future Developments." *Nucleic Acids Research* 43 (D1): D645–55. https://doi.org/10.1093/nar/gku1165.

Vamathevan, Jessica, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, et al. 2019. "Applications of Machine Learning in Drug Discovery and Development." *Nature Reviews Drug Discovery*. https://doi.org/10.1038/s41573-019-0024-5.

Vernikos, George, and Duccio Medini. 2014. "Bexsero® Chronicle." *Pathogens and Global Health* 108 (7): 305–16. https://doi.org/10.1179/2047773214Y.0000000162.

Viboud, Cécile, Lone Simonsen, Rodrigo Fuentes, Jose Flores, Mark A. Miller, and Gerardo Chowell. 2015. "Global Mortality Impact of the 1957-1959 Influenza Pandemic." *Journal of Infectious Diseases*. https://doi.org/10.1093/infdis/jiv534.

Vivona, Sandro, Filippo Bernante, and Francesco Filippini. 2006. "NERVE: New Enhanced Reverse Vaccinology Environment." *BMC Biotechnology* 6: 35. https://doi.org/10.1186/1472-6750-6-35.

Wada, Yamato, Arnone Nithichanon, Eri Nobusawa, Leonard Moise, William D. Martin, Norio Yamamoto, Kazutaka Terahara, et al. 2017. "A Humanized Mouse Model Identifies Key Amino Acids for Low Immunogenicity of H7N9 Vaccines." *Scientific Reports* 7 (1): 1–11.

Weingartl, H., M. Czub, S. Czub, J. Neufeld, P. Marszal, J. Gren, G. Smith, et al. 2004. "Immunization with Modified Vaccinia Virus Ankara-Based Recombinant Vaccine against Severe Acute Respiratory Syndrome Is Associated with Enhanced Hepatitis in Ferrets." *Journal of Virology* 78 (22): 12672–76. https://doi.org/10.1128/jvi.78.22.12672-12676.2004.

Weiskopf, Daniela, Michael A. Angelo, Elzinandes L. de Azeredo, John Sidney, Jason A. Greenbaum, Anira N. Fernando, Anne Broadwater, et al. 2013. “Comprehensive Analysis of Dengue Virus-Specific Responses Supports an HLA-Linked Protective Role for CD8+ T Cells.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (22): E2046-53. https://doi.org/10.1073/pnas.1305227110.

Westermann, Alexander J., Konrad U. Förstner, Fabian Amman, Lars Barquist, Yanjie Chao, Leon N. Schulte, Lydia Müller, Richard Reinhardt, Peter F. Stadler, and Jörg Vogel. 2016. “Dual RNA-Seq Unveils Noncoding RNA Functions in Host–Pathogen Interactions.” *Nature* 529 (7587): 496–501. https://doi.org/10.1038/nature16547.

Westermann, Alexander J., Stanislaw a. Gorski, and Jörg Vogel. 2012. “Dual RNA-Seq of Pathogen and Host.” *Nature Reviews. Microbiology* 10 (9): 618–30. https://doi.org/10.1038/nrmicro2852.

WHO. 2014. “WHO | MDG 6: Combat HIV/AIDS, Malaria and Other Diseases.” *WHO*. http://www.who.int/topics/millennium_development_goals/diseases/en/.

———. 2016. “Disease Burden and Mortality Estimates 2000-2016.” *WHO*. https://www.who.int/healthinfo/global_burden_disease/estimates/en/index1.html.

Wilkinson, Tom M., Chris K.F. Li, Cecilia S.C. Chui, Arthur K.Y. Huang, Molly Perkins, Julia C. Liebner, Rob Lambkin-Williams, et al. 2012. “Preexisting Influenza-Specific CD4 + T Cells Correlate with Disease Protection against Influenza Challenge in Humans.” *Nature Medicine* 18 (2): 274–80. https://doi.org/10.1038/nm.2612.

Wimley, William C. 2003. “The Versatile Beta-Barrel Membrane Protein.” *Current Opinion in Structural Biology* 13 (4): 404–11. https://doi.org/10.1016/S0959-440X(03)00099-X.

Wit, Emmie De, Neeltje Van Doremalen, Darryl Falzarano, and Vincent J. Munster. 2016. “SARS and MERS: Recent Insights into Emerging Coronaviruses.” *Nature Reviews Microbiology* 14 (8): 523–34. https://doi.org/10.1038/nrmicro.2016.81.

World Health Organization. 2020a. “Global Tuberculosis Report 2020.” https://www.who.int/tb/publications/global_report/en/.

———. 2020b. “WHO Coronavirus Disease (COVID-19) Dashboard.” 2020. https://covid19.who.int/.

Wrapp, Daniel, Dorien De Vlieger, Kizzmekia S. Corbett, Gretel M. Torres, Nianshuang Wang, Wander Van Breedam, Kenny Roose, et al. 2020. “Structural Basis for Potent Neutralization of Betacoronaviruses by Single-Domain Camelid Antibodies.” *Cell* 181 (5): 1004–15.

Wrapp, Daniel, Nianshuang Wang, Kizzmekia S Corbett, Jory A Goldsmith, Ching-Lin Hsieh, Olubukola Abiona, Barney S Graham, and Jason S McLellan. 2020. “Cryo-EM Structure of the 2019-NCoV Spike in the Prefusion Conformation.” *Science (New York, N.Y.)* 367 (6483): 1260–63.

Wu, Fan, Aojie Wang, Mei Liu, Qimin Wang, Jun Chen, Shuai Xia, Yun Ling, et al. 2020. “Neutralizing Antibody Responses to SARS-CoV-2 in a COVID-19 Recovered Patient Cohort and Their Implications.” *MedRxiv [Preprint]*. https://doi.org/10.2139/ssrn.3566211.

Wu, Xueling, Tongqing Zhou, Jiang Zhu, Baoshan Zhang, Ivelin Georgiev, Charlene Wang, Xuejun Chen, et al. 2011. “Focused Evolution of HIV-1 Neutralizing Antibodies Revealed by Structures and Deep Sequencing.” *Science* 333 (6049): 1593–1602. https://doi.org/10.1126/science.1207532.

Xiang, Zuoshuang Allen, and Yongqun Oliver He. 2013. “Genome-Wide Prediction of Vaccine Targets for Human Herpes Simplex Viruses Using Vaxign Reverse Vaccinology Human Herpes Simplex (HSV) Viruses” 14 (Suppl 4): 1–10.

Xiang, Zuoshuang, Mélanie Courtot, Ryan R Brinkman, Alan Ruttenberg, and Yongqun He. 2010. “OntoFox: Web-Based Support for Ontology Reuse.” *BMC Research Notes* 3 (June): 175. https://doi.org/10.1186/1756-0500-3-175.

Xiang, Zuoshuang, Yuying Tian, and Yongqun He. 2007. “PHIDIAS: A Pathogen-Host Interaction Data Integration and Analysis System.” *Genome Biology* 8 (7). https://doi.org/10.1186/gb-2007-8-7-r150.

Xiang, Zuoshuang, Jie Zheng, Yu Lin, and Yongqun He. 2015. “Ontorat: Automatic Generation of New Ontology Terms, Annotations, and Axioms Based on Ontology Design Patterns.” *Journal of Biomedical Semantics* 6: 4. https://doi.org/10.1186/2041-1480-6-4.

Xue, Jing, Xiaoqiang Huang, and Yushan Zhu. 2019. “Using Molecular Dynamics Simulations to Evaluate Active Designs of Cephradine Hydrolase by Molecular Mechanics/Poisson-Boltzmann Surface Area and Molecular Mechanics/Generalized Born Surface Area Methods.” *RSC Advances* 9 (24): 13868–77.

Yang, Brian, Samantha Sayers, Zuoshuang Xiang, and Yongqun He. 2011. “Protegen: A Web-Based Protective Antigen Database and Analysis System.” *Nucleic Acids Research* 39 (SUPPL. 1): 1073–78. https://doi.org/10.1093/nar/gkq944.

Yang, Zhi Yong, Wing Pui Kong, Yue Huang, Anjeanette Roberts, Brian R. Murphy, Kanta Subbarao, and Gary J. Nabel. 2004. “A DNA Vaccine Induces SARS Coronavirus Neutralization and Protective Immunity in Mice.” *Nature* 428 (6982): 561–64.

Yasui, Fumihiko, Chieko Kai, Masahiro Kitabatake, Shingo Inoue, Misako Yoneda, Shoji Yokochi, Ryoichi Kase, et al. 2008. “Prior Immunization with Severe Acute Respiratory Syndrome (SARS)-Associated Coronavirus (SARS-CoV) Nucleocapsid Protein Causes Severe Pneumonia in Mice Infected with SARS-CoV.” *The Journal of Immunology* 181 (9): 6337–48. https://doi.org/10.4049/jimmunol.181.9.6337.

Yu, Nancy Y., James R. Wagner, Matthew R. Laird, Gabor Melli, S??bastien Rey, Raymond Lo, Phuong Dao, et al. 2010. “PSORTb 3.0: Improved Protein Subcellular Localization Prediction with Refined Localization Subcategories and Predictive Capabilities for All Prokaryotes.” *Bioinformatics* 26 (13): 1608–15. https://doi.org/10.1093/bioinformatics/btq249.

Zha, Lisha, Hongxin Zhao, Mona O Mohsen, Liang Hong, Yuhang Zhou, Chuankai Yao, Lijie Guo, et al. 2020. “Development of a COVID-19 Vaccine Based on the Receptor Binding Domain Displayed on Virus-like Particles.” BioRxiv [Preprint]. January 1, 2020. https://www.biorxiv.org/content/10.1101/2020.05.06.079830v2.

Zhang, Chengxin, Wei Zheng, Xiaoqiang Huang, Eric W. Bell, Xiaogen Zhou, and Yang Zhang. 2020. “Protein Structure and Sequence Reanalysis of 2019-NCoV Genome Refutes Snakes as Its Intermediate Host and the Unique Similarity between Its Spike Protein Insertions and

HIV-1." *Journal of Proteome Research* 19 (4): 1351–60.

Zhang, Lu. 1709. *Zhang Shi Yi Shu*. Open Collections Program at Harvard University. Contagion. China: s.n.]. https://id.lib.harvard.edu/curiosity/contagion/36-990061701170203941.

Zhang, Yang, and Jeffrey Skolnick. 2004. "Scoring Function for Automated Assessment of Protein Structure Template Quality." *Proteins: Structure, Function and Genetics* 57: 702.

———. 2005. "TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score." *Nucleic Acids Research* 33 (7): 2302–9.

Zhao, Jincun, Jingxian Zhao, Ashutosh K. Mangalam, Rudragouda Channappanavar, Craig Fett, David K. Meyerholz, Sudhakar Agnihothram, Ralph S. Baric, Chella S. David, and Stanley Perlman. 2016. "Airway Memory CD4+ T Cells Mediate Protective Immunity against Emerging Respiratory Coronaviruses." *Immunity* 44 (6): 1379–91. https://doi.org/10.1016/j.immuni.2016.05.006.

Zhao, Jincun, Jingxian Zhao, and Stanley Perlman. 2010. "T Cell Responses Are Required for Protection from Clinical Disease and for Virus Clearance in Severe Acute Respiratory Syndrome Coronavirus-Infected Mice." *Journal of Virology* 84 (18): 9318–25. https://doi.org/10.1128/jvi.01049-10.

Zhao, Ping, Jie Cao, Lan Juan Zhao, Zhao Lin Qin, Jin Shan Ke, Wei Pan, Hao Ren, Jian Guo Yu, and Zhong Tian Qi. 2005. "Immune Responses against SARS-Coronavirus Nucleocapsid Protein Induced by DNA Vaccine." *Virology* 331 (1): 128–35. https://doi.org/10.1016/j.virol.2004.10.016.

Zheng, Jie, Marcelline R. Harris, Anna Maria Masci, Yu Lin, Alfred Hero, Barry Smith, and Yongqun He. 2016. "The Ontology of Biological and Clinical Statistics (OBCS) for Standardized and Reproducible Statistical Analysis." *Journal of Biomedical Semantics* 7 (1): 1–13. https://doi.org/10.1186/s13326-016-0100-2.

Zheng, Wei, Yang Li, Chengxin Zhang, Robin Pearce, S. M. Mortuza, and Yang Zhang. 2019. "Deep-Learning Contact-Map Guided Protein Structure Prediction in CASP13." *Proteins: Structure, Function and Bioinformatics* 87 (12): 1149–64. https://doi.org/10.1002/prot.25792.

Zhou, Peng, Xing Lou Yang, Xian Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao Rui Si, et al. 2020. "A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin." *Nature* 579 (7798): 270–73.

Zhu, Feng Cai, Yu Hua Li, Xu Hua Guan, Li Hua Hou, Wei Wen Juan Wang, Jing Xin Li, Shi Po Wu, et al. 2020. "Safety, Tolerability, and Immunogenicity of a Recombinant Adenovirus Type-5 Vectored COVID-19 Vaccine: A Dose-Escalation, Open-Label, Non-Randomised, First-in-Human Trial." *The Lancet* 395 (10240): 1845–54.

Zhu, Jiang, Gilad Ofek, Yongping Yang, Baoshan Zhang, Mark K. Louder, Gabriel Lu, Krisha McKee, et al. 2013. "Mining the Antibodyome for HIV-1-Neutralizing Antibodies with next-Generation Sequencing and Phylogenetic Pairing of Heavy/Light Chains." *Proceedings of the National Academy of Sciences of the United States of America* 110 (16): 6470–75. https://doi.org/10.1073/pnas.1219320110.

Zhu, Jiang, Xueling Wu, Baoshan Zhang, Krisha McKee, Sijy O'Dell, Cinque Soto, Tongqing Zhou, et al. 2013. "De Novo Identification of VRC01 Class HIV-1-Neutralizing Antibodies by next-Generation Sequencing of B-Cell Transcripts." *Proceedings of the National Academy of Sciences of the United States of America* 110 (43). https://doi.org/10.1073/pnas.1306262110.

Zygmunt, Michel S., Sue D. Hagius, Joel V. Walker, and Philip H. Elzer. 2006. "Identification of Brucella Melitensis 16M Genes Required for Bacterial Survival in the Caprine Host." *Microbes and Infection* 8 (14–15): 2849–54. https://doi.org/10.1016/j.micinf.2006.09.002.