國立臺灣大學生命科學院生化科學研究所

碩士論文

Institute of Biochemical Sciences
College of Life Science
National Taiwan University
Master Thesis

人類冠狀病毒 229E 的宿主識別與抗原漂變之分子基礎

Molecular basis of host recognition and antigenic drift of human coronavirus 229E

蔡淯璽

Yu-Xi Tsai

指導教授：徐尚德 博士

Advisor: Shang-Te Danny Hsu, Ph.D.

中華民國 112 年 7 月

July 2023

doi:10.6342/NTU202301899

CONTENTS

doi:10.6342/NTU202301899

doi:10.6342/NTU202301899

doi:10.6342/NTU202301899

中文摘要

新冠病毒（COVID-19）大流行已經導致全球超過 600 萬人死亡，促使人們進行大量研究以對抗嚴重急性呼吸綜合症冠狀病毒 2 型（SARS-CoV-2）。相比之下，已經發現了超過 50 年且只引起輕微症狀的最早識別的季節性冠狀病毒 HCoV-229E 在學界受到較少的關注。在本研究中，我們首先探討 HCoV-229E 之親緣關係。我們選擇了八個病毒株來依時間順序展示基因演化。結果顯示於 HCoV-229E 棘蛋白（S）的受體結合區域（RBD）和 N 端區域（NTD）存在許多突變。另外，在 N-糖基化序列中有六個突變，表明糖基組成亦存在變異。接著，我們專注於其中兩個病毒株，即 P100E 與 Seattle，以分別代表“舊病毒株”和“新病毒株”。與 P100E 棘蛋白序列相比，Seattle 病毒株棘蛋白中被觀察到了共 72 個點突變。通過模擬預測發現這些突變有可能對蛋白質結構產生破壞性影響，此猜想在對 Seattle 病毒株棘蛋白進行負染色的電子顯微鏡（NSEM）觀察時得到了確認，其中正常折疊的棘蛋白數量大幅減少。通過液相色譜-串聯質譜（LC-MS/MS）進行的 N-糖基分析揭示了明顯的糖基組成差異，P100E 在 N62 和 N930 位置表現出更多的高甘露糖型糖基，而 Seattle 株則含有四分之一以上的複雜型糖基。我們假設 Seattle 株棘蛋白的結構完整性受損，導致較多的複雜型醣基存在。為了驗證此假設，我們引入了 GlycoSHIELD 以定量分析醣基化的三維分子結構，進而提供與結構和演化相關的證據。另一方面，目前已知的 229E S PDB 結構僅有無法讓受體結合的 RBD-all-down 狀態。為了瞭解其宿主識別機制，我們選用相對穩定的 P100E S 進行 cryo-EM 分析，研究 229E S 與人類胺肽酶 N （hAPN）形成之複合物。根據計算出的庫倫電位圖，該複合物由兩個採用了未曾報導過之 RBD-up 構型的 229E 棘蛋白與 hAPN 同源二聚體結合，後者為細胞表面常見的寡聚化狀態。為了更接近生理狀態，我們再次利用 GlycoSHIELD 結合所有醣基化分析結果和 cryo-EM 解析之複合物結構，構建一個完全醣基化的模型。該模型提供了有關 HCoV-229E 病毒入侵體內細胞的分子機制。本研究建立的分析流程提供了一種有效而全面的方法，

doi:10.6342/NTU202301899

用於探究抗原漂變、醣基化和分子結構之間的交互作用，並且能適用於其他蛋白質。此外，它亦可作為一系統性的分析策略，以預測蛋白質表面潛在的廣泛性中和抗體（bnAbs）抗原決定位。

關鍵字：季節性冠狀病毒 HCoV-229E、抗原漂變、N-連接醣基化、醣基化修飾之遮蔽效應、冷凍電子顯微鏡、Python 程式設計

doi:10.6342/NTU202301899

ABSTRACT

The COVID-19 pandemic has resulted in over six million deaths worldwide, prompting extensive research efforts to combat SARS-CoV-2. In contrast, HCoV-229E, the earliest-identified seasonal coronavirus that has been present for over 50 years and only causes mild symptoms, receiving comparatively less attention from scientific studies. In this study, we initially examined the phylogeny of HCoV-229E and selected eight strains to demonstrate the genetic evolution chronologically. Significant mutations were identified, particularly in the receptor binding domain (RBD) and N-terminal domain (NTD) of the HCoV-229E spike protein (S). Moreover, six mutations were found in the N-glycosylation sequon, indicating variations in glycan profile. We then took P100E strain and Seattle strain to represent the "old strain" and the "new strain". A total of 72 point mutations were observed in the Seattle strain compared to the P100E sequence. *In-silico* ΔΔG prediction suggested destabilizing effects of the mutations, which were later confirmed by negative staining EM (NSEM) of the Seattle strain, where the number of well-folded S proteins decreased. N-glycan analysis via LC-MS/MS revealed marked differences in the glycan composition, with P100E showing a higher abundance of high-mannose type glycans at N62 and N930, while Seattle exhibited over ¼ of the complex glycans. We hypothesized that the structural integrity of the Seattle strain was compromised, leading to the presence of further processed glycans. To investigate this hypothesis, GlycoSHIELD was used to quantitatively analyze the glycosylated 3D molecular architectures, providing structurally and evolutionarily relevant results. In addition, all the current 229E S PDB structures only display the RBD-all-down state, which prevents the receptor binding. To uncover the molecular basis of its host recognition, we conducted cryo-electron microscopy (cryo-EM) analysis of P100E S in complex with human aminopeptidase N (hAPN), revealing a structure consisting of two 229E spikes adopting an unprecedented RBD-up conformation,

doi:10.6342/NTU202301899

bound to a hAPN homodimer. To emulate the physiological state, we utilized GlycoSHIELD again to combine all glycosylation profiles and the cryo-EM-solved complex structure, constructing a fully glycosylated model anchored on a lipid membrane. This model provides insights into HCoV-229E virus entry *in vivo*. The established workflow serves as an efficient and comprehensive approach to investigate the interplay between antigenic drift, glycosylation, and molecular structure, with potential application to other proteins. It also provides a systematic way to probe the possible targets for broadly neutralizing antibodies (bnAbs).

Keywords：Seasonal coronavirus HCoV-229E, Antigenic drift, N-linked glycosylation, glycan shielding, Cryo-EM, Python programming

doi:10.6342/NTU202301899

LIST OF FIGURES

doi:10.6342/NTU202301899

doi:10.6342/NTU202301899

doi:10.6342/NTU202301899

LIST OF TABLES

doi:10.6342/NTU202301899

ABBREVIATIONS

HCoV: human coronavirus

HCoV-229E or “229E” for brevity: Human coronavirus 229E

HCoV-NL63 or “NL63” for brevity: Human coronavirus NL63

PEDV: Porcine epidemic diarrhea virus

hAPN: human aminopeptidase N

ACE2: angiotensin-converting enzyme 2

DPP4: dipeptidyl peptidase 4

S protein: spike protein

NTD: N-terminal domain

CTD: C-terminal domain

RBD: receptor-binding domain

HexNAc: N-acetylhexosamine

Hex: hexose

Fuc: fucose

Neu5Ac: 5-N-acetyl neuraminic acid

9-O-Ac-Sia: 9-O-acetylated sialic acid

NSEM: negative staining electron microscopy

Cryo-EM: cryogenic electron microscopy

LC: liquid chromatography

MS: mass spectrometry

MD simulation: molecular dynamic simulation

T: trypsin

C: chymotrypsin

aLP: alpha-lytic protease

doi:10.6342/NTU202301899

PSM: peptide-spectrum match

SASA: solvent-accessible surface area

ddG/ΔΔG: delta delta Gibbs free energy

COM: center of mass

2D: two dimensional

3D: three dimensional

nAb: neutralizing antibody

bnAb: broadly neutralizing antibody

doi:10.6342/NTU202301899

Chapter 1 INTRODUCTION

1.1 Human coronaviruses (HCoVs)

Coronaviruses (CoVs) are a diverse group of positive-strand RNA viruses with enveloped structures. They belong to the *Coronavirinae* subfamily, which is part of the *Coronaviridae* family under the *Nidovirales* order. CoVs can be classified into four different genera: alpha, beta, gamma, and delta[1]. The term "corona" derives from Latin and means "crown," referring to the distinctive crown-like structures observed on the virus membrane under an electron microscope[2]. Apart from infecting various animals, these viral pathogens are known for their ability to cross species barriers, leading to outbreaks of respiratory diseases in humans. Since the mid-1960s when the first human coronaviruses (HCoVs) were discovered, a total of seven distinct HCoVs have been identified: HCoV-229E (1965), HCoV-OC43 (1967), SARS-CoV (2002), HCoV-NL63 (2004), HCoV-HKU1 (2005), MERS-CoV (2012), and SARS-CoV-2 (2019) [1]. Over the past two decades, three of these HCoVs have caused global outbreaks, posing significant threats to human health. The severe acute respiratory syndrome (SARS) outbreak originated in the Guangdong province of China in 2002 and subsequently spread worldwide, resulting in over 8,000 reported cases with a mortality rate of 10% (https://www.cdc.gov/sars/surveillance/absence.html). Similarly, the Middle East respiratory syndrome (MERS) outbreak emerged in Saudi Arabia in 2012 and later spread to other countries in the Middle East, as well as South Korea in East Asia. The MERS outbreak led to approximately 2,600 reported cases with a mortality rate of 35% (https://www.who.int/emergencies/disease-outbreak-news/item/2022-DON363). More recently, SARS-CoV-2, a closely related virus to SARS-CoV, emerged and caused the coronavirus disease 2019 (COVID-19) pandemic, infecting millions of people worldwide. Various variants of SARS-CoV-2 continue to emerge. It is worth noting that unlike the

doi:10.6342/NTU202301899

aforementioned HCoVs, which primarily affect the lower respiratory system, the remaining HCoVs, commonly cause mild upper respiratory tract infections[3]. While rare severe pneumonia cases have been reported, most individuals infected with these HCoVs experience recovery without medical intervention.

doi:10.6342/NTU202301899

1.2 Structural characteristics of spike (S) protein

The RNA genome of HCoVs encodes four distinct structural proteins: spike (S), envelope (E), membrane (M), and nucleocapsid (N) proteins. Among these, the entry of HCoVs into host cells primarily relies on the S protein, which facilitates host attachment and membrane fusion. The S protein is also the primary target of neutralizing antibodies produced by the host's immune response[4].

The S protein is a trimeric class I fusion protein consisting of three protomers. Each protomer contains two subunits: S1, responsible for receptor attachment, and S2, responsible for membrane fusion. The S1 subunit of the S protein can be further divided into distinct domains, namely the NTD, C-terminal domain (CTD), SD1, and SD2 subdomains[5]. Both the NTD and CTD can serve as the RBD to recognize the host cell receptors (**Table 1.2-1**). Generally, the NTD interacts with sugar-based co-receptors, such as non-modified 5-N-acetyl neuraminic acid (Neu5Ac) in the case of MERS-CoV S protein or 9-O-acetylated sialic acid (9-O-Ac-Sia) in the case of HCoV-HKU1 S protein. On the other hand, the CTD interacts with protein-based receptors, such as angiotensin-converting enzyme 2 (ACE2) for SARS-CoV-1 and SARS-CoV-2 S proteins or dipeptidyl peptidase 4 (DPP4) for MERS-CoV S protein. Cryo-EM analysis has revealed significant conformational changes in the RBD (i.e., the CTD of the S protein) of SARS-CoV-1, SARS-CoV-2, and MERS-CoV S proteins, transitioning between "RBD-up" and "RBD-down" states to regulate the accessibility of receptor-binding motifs (RBMs) for receptor interaction[6-8]. Notably, these distinct conformational changes have not been observed in other published structures of HCoVs adopting the "RBD-down" conformation[9-12], suggesting that the RBD requires structural rearrangement prior to receptor recognition for viral entry (e.g., HCoV-NL63 and HCoV-229E). Consequently, the S1 subunit of the S protein plays a crucial role in engaging host cell receptors, while the receptor specificity

doi:10.6342/NTU202301899

of its functional domains determines the host range and tissue tropism.

Primarily, the S protein requires proteolytic priming at the S1/S2 site by specific cellular proteases. This priming releases the structural constraint between the S1 and S2 subunits, subsequently initiating a cascade of conformational changes that lead to the insertion of the fusion peptide (FP) into the host cell membrane[13]. Apart from the RBDs of the S1 subunit, the tissue and cell-type specificity of the protease also contribute to viral tropism. Following FP insertion, the heptad repeat regions (HR1 and HR2) downstream of each protomer assemble into a six-helix bundle, facilitating conformational changes that drive the fusion of viral and host cell membranes, allowing the transfer of the viral genetic material into the host cell.

doi:10.6342/NTU202301899

Table 1.2-1. Cellular receptors of HCoVs.

HCoV genera	HCoV	Receptor binding domain	Cellular (co)receptor
α-CoVs	HCoV-229E	S1-CTD	hAPN[a]
	HCoV-NL63	S1-CTD	hACE2[b]
β-CoVs	SARS-CoV-1	S1-CTD	hACE2[b]
	SARS-CoV-2	S1-CTD	hACE2[b]
	MERS-CoV	S1-NTD	Neu5Ac[c]
		S1-CTD	hDPP4[d] (CD26)
	HCoV-HKU1	S1-NTD	9-O-Ac-Sia[e]
	HCoV-OC43	S1-NTD	9-O-Ac-Sia[e]

a. APN: aminopeptidase N; b. ACE2: angiotensin-converting enzyme 2; c. Neu5Ac: N-acetylneuraminic acid; d. DPP4: dipeptidyl peptidase 4; e. 9-O-Ac-Sia: 9-O-acetyl-sialic acid

doi:10.6342/NTU202301899

1.3 Structural characteristics of human aminopeptidase N (hAPN)

Human aminopeptidase N (hAPN), also known as CD13, is a transmembrane metalloprotease enzyme that plays a crucial role in various physiological and pathological processes, such as pain sensation, blood pressure regulation, tumor angiogenesis and metastasis, immune cell chemotaxis, sperm motility, cell-cell adhesion, and coronavirus entry[14]. This protein is primarily expressed on the surface of cells, including endothelial cells, fibroblasts, and epithelial cells, and is the most extensively studied member in the M1-family of metalloenzymes[14, 15].

The structural features of hAPN contribute significantly to its functional versatility and enzymatic activity. hAPN consists of multiple domains that work in harmony to execute its catalytic function and interact with various substrates and binding partners. The primary structure of hAPN includes a signal peptide, a large extracellular domain, a single transmembrane helix, and a short cytoplasmic tail[16].

The extracellular domain of hAPN is the most extensive region of the protein and consists of three distinct subdomains: a zinc-binding catalytic domain, a protease-associated domain, and a stalk region. The catalytic domain, located near the N-terminus, contains a zinc-binding motif (HEXXH) and plays a crucial role in the enzyme's aminopeptidase activity[17]. This domain is responsible for the cleavage of amino acids from the N-terminus of peptides and proteins.

The protease-associated domain, situated in the middle of the extracellular region, is responsible for dimerization and interacts with other proteins to form functional complexes[16]. It plays a significant role in protein-protein interactions and cell signaling processes[16]. The Ser/Thr rich stalk region connects the catalytic and protease-associated domains, providing flexibility and stability to the overall protein structure.

The transmembrane domain of hAPN consists of a single hydrophobic helix that

doi:10.6342/NTU202301899

anchors the protein to the cell membrane. This region plays a critical role in maintaining the enzyme's orientation and stability within the lipid bilayer, allowing it to interact with substrates and signaling molecules present in the extracellular environment.

Finally, the short cytoplasmic tail of hAPN serves as a link between the transmembrane domain and intracellular signaling pathways, and the exact functions of the cytoplasmic tail are still being explored.

In conclusion, the structural features of hAPN, including the extracellular catalytic and protease-associated domains, the transmembrane helix, and the cytoplasmic tail, are essential for its enzymatic activity, substrate specificity, and functional interactions. Understanding the intricate details of APN's structure can provide valuable insights into its diverse biological functions and potentially lead to the development of novel therapeutic approaches for various diseases and disorders[18].

doi:10.6342/NTU202301899

1.4 Protein N-linked glycosylation

Protein N-glycosylation is a major alteration that involves the addition of N-glycans to specific locations on secretory and membrane-bound glycoproteins known as the NxS/T sequon (N: asparagine, x ≠ proline, S/T: serine or threonine). OST transfer sugars from a lipid-linked precursor, dolichol-P-P-GlcNAc2Man9Glc3, onto the sequons of newly generated nascent peptide chains as they pass through the translocon in the endoplasmic reticulum (ER). However, due to conformational restrictions or other causes, N-glycan transfer may not always occur. As a result, the NxS/T sequons are thought to be potential N-glycosylation sites, and more experimental data is needed to validate the presence of N-glycans. It is important to note that N-glycan addition takes place during protein translation on partially folded polypeptides. This process is also known as "co-translational modification" (CTM) [19].

N-glycosylation modifies glycoproteins by adding non-charged, bulky, and hydrophilic groups, which help maintain solubility and aid to appropriate protein folding. Furthermore, N-glycans play an important function in the ER's quality control process. Aberrant proteins are degraded as a result of ER-associated degradation[20]. Well-folded glycoproteins are delivered to the Golgi apparatus after passing through the ER-protein quality control (ERQC) systems. Glycans are further modified in the Golgi by a succession of glycosidases and glycosyltransferases, resulting in the particular glycosylation patterns of proteins. N-glycans are broadly categorized into three types: high-mannose, hybrid, and complex-type glycans. Because of the wide variety of processing procedures involved in glycan maturation, site-specific N-glycans on proteins may contain an amalgam of glycoforms. The length, number of branches, composition, and core changes of these glycoforms can all vary.

doi:10.6342/NTU202301899

1.5 Cryo-EM, mass spectrometry, and MD simulation for S glycoproteins

Cryo-EM single-particle analysis is a robust biophysical technique that delivers near-atomic resolution insights into the structures of biological macromolecules and biomolecular assemblies. The numerous structural stages of the dynamic biological process are now more clearly displayed, thanks to advancements in electron microscopes as well as the development of algorithms and software that contribute to determining local heterogeneity. However, the atomic coordinates of residues with poor local resolution in EM-maps are tricky to specify due to intrinsic protein dynamics, resulting in missing loops or domains in the final reconstructed EM structures. In contrast, liquid chromatography-mass spectrometry (LC-MS) is a widely used technology for studying glycoproteins. The composition and distribution of N-glycans on specific protein sites are determined by site-specific N-glycopeptide analysis. Furthermore, the relationship between the various N-glycosylation stages and the accessibility of glycan processing enzymes provides a structural understanding of the protein's immediate environment. As a result, cryo-EM and LC-MS are complementing techniques that meet various needs for S glycoprotein analysis.

Although cryo-EM and MS can offer structural and N-glycan information on S glycoproteins, respectively, neither can fully characterize the dynamic glycans' effects on protein surfaces, which is required for glycoproteins to avoid the host immune system. As a result, molecular dynamic (MD) simulation may be the best tool for providing a different perspective on the structural and glycan dynamics of the S protein. However, MD simulation of full-length S glycoproteins is typically computationally expensive and requires supercomputer experience. Thus, glycoSHIELD[21], an open-source pipeline that allows non-expert users to model glycan conformer arrays onto any static protein structure in order to quantify the N-glycan shielding effect on the protein surfaces, may

doi:10.6342/NTU202301899

be another technique for shedding light on the dynamics of these N-glycans.

doi:10.6342/NTU202301899

1.6 Antigenic drift of CoVs

The ongoing COVID-19 pandemic caused by SARS-CoV-2 has highlighted the importance of understanding the evolutionary mechanisms that shape the antigenic properties of these viruses. One such mechanism, known as antigenic drift, plays a crucial role in the continuous adaptation of CoVs to the host immune system, impacting viral fitness, immune evasion, and the development of effective vaccines.

Antigenic drift refers to the gradual accumulation of genetic variations within the genes encoding the surface proteins of a virus, particularly those involved in eliciting host immune responses. In the case of CoVs, the S protein is a primary target for the host immune system. However, due to its exposure to immune pressure, the S protein undergoes antigenic drift, leading to the emergence of new viral variants capable of evading preexisting immunity[22].

The antigenic drift of CoVs is driven by several factors, including intrinsic error-prone replication mechanisms, host immune responses, and interspecies transmission events. The RNA-dependent RNA polymerase (RdRp) of CoVs lacks proofreading activity, leading to a relatively high mutation rate during replication[23-26]. While the majority of these mutations are detrimental and lead to viral fitness reduction, some confer a selective advantage by altering antigenic properties, allowing escape from immune recognition. The host immune response, particularly humoral immunity mediated by neutralizing antibodies, places strong selective pressure on CoVs. As individuals develop an immune response against the dominant circulating strain, the virus may acquire mutations within the epitopes targeted by antibodies or increase the steric hindrance surrounding the epitopes via gaining new N-glycosylation sequon[22], enabling immune evasion and leading to the emergence of new viral variants. This continuous adaptation can result in reduced efficacy of neutralizing antibodies and pose challenges

doi:10.6342/NTU202301899

for the long-term control of viral infections. In addition to intrinsic and immune-driven factors, interspecies transmission events between animals and humans provide opportunities for the introduction of novel CoVs into the human population[27]. These zoonotic events often involve viruses with antigenically distinct S proteins, which can further complicate the immune response dynamics by introducing completely new epitopes to which humans may have little or no preexisting immunity.

COVID-19 has spurred a myriad of researches on the antigenic drift of SARS-CoV-2 S protein along with its structural and functional ramifications[28]. In contrast, the seasonal coronavirus, HCoV-229E, receives scarce attention due to its common association with the mild symptoms, albeit holds a comparatively more extensive evolutionary history[29, 30]. Only recently did Eguia et al. [30] and Xiang et al. [31] systematically analyze the antigenic drift and the epitopes of HCoV-229E S proteins, respectively. The results suggest a rapid viral evolution and the ensuing immune escape reminiscent of SARS-CoV-2.

Understanding the antigenic drift of CoVs has crucial implications for public health interventions, including the development of effective vaccines and therapeutics. As SARS-CoV-2 and HCoV-229E continue to evolve, monitoring and characterizing antigenic drift will be essential for identifying emerging variants of concern and updating vaccine formulations accordingly. Additionally, insights gained from studying the antigenic drift of CoVs can inform our understanding of other viral pathogens and contribute to the development of more broadly protective vaccines and antiviral strategies.

In conclusion, antigenic drift plays a central role in the evolution of CoVs and has profound implications for viral fitness, immune evasion, and vaccine development. By comprehensively examining the underlying mechanisms driving antigenic drift, we can gain valuable insights into the dynamic interplay between the virus and the host immune

doi:10.6342/NTU202301899

system, enabling more effective control strategies against emerging and reemerging CoVs.

doi:10.6342/NTU202301899

1.7 Specific aims

- Phylogenetic investigation of HCoV-229E S proteins to identify the distant strains for the biophysical assays and structural analysis.
- Quantitatively and visually describe the effect of antigenic drift on the protein stability, functional competence, and N-linked glycosylation profile.
- Select a qualified strain for cryo-EM analysis of HCoV-229E S-hAPN complex structure, aiming to unveil the unknown molecular basis of the host recognition of HCoV-229E.
- Combine the above experiments to propose a comprehensive model depicting the potential mechanism for HCoV-229E cell entry *in vivo*.

doi:10.6342/NTU202301899

Chapter 2 MATERIAL AND METHODS

2.1 Phylogenetic analysis and antigenic drift analysis

The procedure for the analyses was as described previously by Eguia et al. [30] The HCoV-229E NCBI Virus accessions used here were an updated version as of October 11th, 2022, which is available on https://github.com/coco0981568491/Master_Thesis/blob/master/HCoV-229E_accessions/20221011_NCBI_Virus_229E_accessions.csv.

2.2 Estimating HCoV-229E S protein stability via *in silico* ΔΔG prediction

According to this review[32], utilizing multiple delta delta Gibbs Free Energy (ΔΔG) predictors is a rough, yet effective approach to arrive at a more reliable conclusion about the influence of mutations on protein stability. Here we used five well-established and tested tools to predict the ΔΔG of several HCoV-229E strain pairs. The tools we used were MAESTRO[33] (command-line version), INPS-3D[34], INPS sequence[35] (abbreviated as "INPS-Seq*"), DDGun with input PDB structure[36] (abbreviated as "DDGun-3D"), and DDGun with input sequence[36] (abbreviated as "DDGun-Seq*"). The following strain pairs were analyzed: GenBank: DQ243972 (1984) to GenBank: AAG48592.1 (7CYC), GenBank: DQ243972 (1984) to GenBank: AAK32191.1 (P100E strain), GenBank: DQ243972 (1984) to GenBank: ABB90519.1 (1992), GenBank: DQ243972 (1984) to GenBank: ABB90520.1 (2001), GenBank: DQ243972 (1984) to GenBank: AIW52753.1 (2008), GenBank: DQ243972 (1984) to GenBank: APT69883.1 (SC1073, Seattle strain), GenBank: DQ243972 (1984) to GenBank: APT69856.1 (SC677), and P100E strain to Seattle strain.

All sequences were in FASTA format and downloaded from National Center for Biotechnology Information (NCBI). For protein structure-based predictions, i.e., the tools

doi:10.6342/NTU202301899

named with “3D”, the sequences of the corresponding strains were submitted to SWISS-MODEL[37] for homology modeling with the reported cryo-EM structure of HCoV-229E (PDB ID: 6U7H) as the reference to generate the required PDBs. Finally, all results from the aforementioned tools were listed in **Table 3.2-1.**

2.3 Construct design

Sino Biological Inc. synthesized the genes that encode for the ectodomain of the S protein of the HCoV-229E Seattle strain, whereas the genes for the ectodomain of the S protein of the HCoV-229E P100E strain were provided as a gift by the California Institute of Technology (Caltech). In addition, the genes encoding the ectodomain of hAPN were purchased from Addgene (https://www.addgene.org/) with the construct name "pLEX307-APN-puro" under the plasmid number #158454. The DNA sequences of HCoV-229E Seattle strain S (residues 1-1111) and hAPN ectodomain (residues 66-967) were cloned into a modified mammalian cell expression vector pcDNA3.4-TOPO (Invitrogen, USA), which contains a T4 fibritin foldon trimerization domain followed by a c-Myc sequence and a hexa-repeat histidine (His 6) tag. The plasmid containing HCoV-229E P100E strain S sequence from Caltech, which comprises a T4 fibritin foldon trimerization domain followed by a nona-repeat histidine (His 9) tag and a AviTag™, was unmodified and directly used a construct for cell transfection. To modify the HCoV-229E P100E strain S, its original signal peptide was substituted at the N-terminus with the μ-phosphatase signal sequence (MGILPSPGMPALLSLVSLLSVLLMG). On the other hand, Ig kappa (Igk) leader sequence was used for hAPN construct. As for HCoV-229E Seattle strain S, the native signal peptide was maintained. All S proteins were stabilized with dual proline (2P) mutation (^{871}TI^{872}→^{871}PP^{872} for P100E S, ^{869}TI^{870}→^{869}PP^{870} for Seattle S with 2 deletions).

doi:10.6342/NTU202301899

2.4 Protein expression and purification

The plasmids of HCoV-229E P100E strain S, HCoV-229E Seattle strain S, and hAPN ectodomain were transiently transfected into Expi293F cells with ExpiFectamine 293 transfection kit by following its protocols in the user guide. The transfected cells were incubated at 37 ºC with 8% CO_2 for six days. After that, the cells were centrifuged at 4000 rpm for 30 min, then the supernatant containing the secreted form of S proteins and hAPN ectodomain in the medium was harvested and filtered through a 0.22 μm cutoff membrane (Satorius, France). Furthermore, the medium was incubated with HisPur Cobalt Resin (Thermo Fisher Scientific, U. S. A.) in binding buffer (20 mM Tris-HCl (pH 8.0), 150 mM NaCl, and 5 mM imidazole) at 4 ºC overnight. The resin was then transferred into the PD-10 column and iteratively washed with wash buffer (20 mM Tris-HCl (pH 8.0), 300 mM NaCl, and 10 mM imidazole). The target proteins were subsequently eluted by elution buffer (20 mM Tris-HCl (pH 8.0), 300 mM NaCl, and 150 mM imidazole). After elution, the proteins were concentrated and subsequently subjected to additional purification using a size exclusion chromatography (SEC) column (Superose 6 increase 10/300 GL; GE Healthcare, U.S.A.), utilizing a running buffer consisting of 50 mM Tris-HCl (pH 7.6), 150 mM NaCl, and 0.02% NaN3. The protein concentrations were measured using a UV-Vis spectrometer (Nano-photometer N60, IMPLEN, Germany) by analyzing the UV absorbance at 280 nm.

2.5 Differential scanning fluorimeter (DSF)

The S proteins of HCoV-229E P100E strain and HCoV-229E Seattle strain were analyzed by Tycho NT.6 (NanoTemper Technologies). Fresh samples were diluted ten-fold in buffers having distinct pH values, which included 100 mM sodium acetate (pH 4.0 and pH 5.0), MES (pH 6.0), HEPES (pH 7.0), and Tris (pH 7.6 and pH 8.6). All buffers were

doi:10.6342/NTU202301899

supplemented with 150 mM NaCl. The final concentration of the samples was adjusted to fall within the suggested value range by the machine manual. The temperature ramping span was set from 35ºC to 95ºC at a scanning rate of 30 ºC/min, and the resulting data were processed by the internal software from NanoTemper.

2.6 Negative staining electron microscopy (NSEM) analysis

Negative-stained grids were prepared for the S proteins of HCoV-229E P100E strain and HCoV-229E Seattle strain as well as the hAPN ectodomain, for negative staining electron microscopy (NSEM). To prepare the grids for NSEM, the carbon-coated grids were initially glow-discharged with 25 mA for 30 seconds. Next, 4 μL of fresh samples at a concentration of 50 μg/mL were applied to the grid surface, and incubated for 1 minute to facilitate sample absorption. After that, the grids were blotted with a sheet of filter paper and stained with 4 μL of 0.2% uranyl formate (UF) for 1 minute, then blotted again and air-dried for one day. Images were collected by FEI Tecnai G2-F20 electron microscope at 200 keV (FEI, the Netherlands). An imaging magnification of 50000x was employed, which corresponded to a pixel size of 1.7 Å. All datasets were processed by CryoSPARC v4.1.1. Following the import of micrographs, patch-CTF estimation was conducted, and subsequent particle-picking was carried out utilizing the "blob picker" function. The picked particles were extracted from the micrographs with a box size of 224 pixels, which were later cleaned up by applying iterative rounds of 2D classification. The particles from selected 2D classes were used for ab-initio 3D reconstruction and heterogeneous refinement. Then, the particles from the best 3D class were applied to non-uniform (NU) refinement without symmetry (C1) to generate the final NSEM map. The resulting NSEM maps were visualized and analyzed by UCSF-ChimeraX[38].

doi:10.6342/NTU202301899

2.7 Biolayer interferometry (BLI)

The experimental procedure was as described previously[39], with the exception that human aminopeptidase N, the receptor of HCoV-229E, was immobilized onto High Precision Streptavidin (SAX) biosensors (Sartorius) instead of ACE2. Additionally, the SARS-CoV-2 S proteins were replaced with those of HCoV-229E P100E and Seattle strain.

2.8 Cryo-EM grid preparation and Data collection

To confirm the binding between the HCoV-229E S protein and its receptor, hAPN, the SEC-purified HCoV-229E S protein solution was initially incubated with hAPN for 1 hour at 4 ºC. Subsequently, a second round of SEC purification was performed to isolate only the fractions that indicated the formation of 229E-hAPN complexes. The fractions containing the complexes were collected and concentrated to 3.5 mg/mL for ensuing cryo-EM analysis, from which 3 μL of the complex solution was applied onto 300-mesh Quantifoil R1.2/1.3 holey carbon grids. The grids were glow-discharged at 20 mA for 30 s. After a 30-second waiting time, the grids were blotted for 2.5 s at 4 ºC with 100% humidity, then vitrified by Vitrobot Mark IV (ThermoFisher Scientific, U. S. A.). Data collection was performed on a 300-keV Titan Krios transmission electron microscope (TEM) equipped with a Gatan K3 direct detector (Gatan) in the super-resolution mode using EPU 2.10 software (Thermo Fisher Scientific). Movies were collected with a defocus range between -1.5 and -2 μm and a magnification of 81000 x, resulting in a pixel size of 1.08 Å. A total dose of 50 $e^-/Å^2$ was distributed over 50 frames with an exposure time of 2.5 s.

2.9 Image processing and 3D reconstruction

A total of 7961 super-resolution movies were motion-corrected on the fly by

doi:10.6342/NTU202301899

MotionCor2_1.4.4[40] during the data collection. The motion-corrected movies were later imported to CryoSPARC v4.1.1[41] to perform successive image analysis. Contrast transfer function (CTF) estimation was conducted by patch-CTF estimation function in CryoSPARC v4.1.1[41]. Afterwards, all the CTF-corrected micrographs were subjected to particle picking with parameters of 150 Å "Minimum particle diameter" and 400 Å "Maximum particle diameter". A total of 2,663,525 picked particles were Fourier-cropped to a box size of 410 pixels after the particle extraction with the 820-pixels box size (2x2-binned) and then applied to iterative rounds of 2D classification for filtering out junk particles. The remaining 253,164 particles were used to perform *ab-initio* reconstruction with C1 symmetry, followed by heterogeneous refinement to generate three distinct classes. Of note, among the three generated 3D classes, only one class resembled the theoretical shape of a 229E-hAPN complex. Therefore, all the following analysis was based on this single class while the other two being discarded. The processing procedure to this stage had already singled out the desired structure target, 229E-hAPN complex, thus the corresponding particles were re-extracted with a box size of 600 pixels (1.36x1.36-binned) to retain as much of the original structural information as possible with limited computing resource. Note that extracting particles by their full box size (un-binned) would usually be the ideal situation. However, due to the computational infrastructure used, a trade-off of 600 pixels was deemed the most feasible option between the resolution limit and computing power. Next, the re-extracted particles were applied to non-uniform (NU) refinement to generate an initial map and its corresponding mask. NU-refinement takes the disordered or flexible regions of a structure into account[42], thus a higher resolution of 4.33 Å was achieved compared to 5.91 Å from the conventional homogeneous refinement of the single class. Since the resulting map of the 229E HCoV S-hAPN complex structure exhibited a highly dynamic nature of the two attached 229E

doi:10.6342/NTU202301899

HCoV S proteins, which led to a significant loss of local resolution, local refinement was incorporated to deal with this issue. The following 6 local refinements were performed (the names indicate the refined regions): mono-hAPN + singe S, single S, mono-hAPN + single RBD, hAPN dimer + single RBD + hinge (the peptide linking the RBD region and the S1 subunit of S protein), mono-hAPN, and single RBD + hinge. All the local refinements were carried out with C1 symmetry to account for the non-symmetrical characteristics of the 229-hAPN complex. Finally, all 6 focus-refined cryo-EM Coulomb potential maps along with the NU-refined full complex map were imported to Phenix[43] to generate a composite map by a built-in function called "combine_focused_maps". The resulting composite map was used for the following model building and structural refinement.

2.10 Model building and refinement

The initial model of HCoV-229E-hAPN complex was generated based on the previously reported cryo-EM structure of HCoV-229E S (PDB ID: 6U7H) [44] and X-ray crystallography structure of HCoV-229E RBD in complex with hAPN (PDB ID: 6U7G) [44]. The atomic coordinates were segregated into individual domains and manually rigid-body-fitted into the Phenix composite map from the previous step using UCSF-ChimeraX[38]. The fitted model was examined and manually adjusted via Coot[45] and the derived model served as a starting point for further structural optimization in finer details. Next, the model was processed by real-space refinement in Phenix[43] with the default parameters. To enhance the correlation between the model and the cryo-EM composite map, multiple rounds of manual optimization were performed in Coot, followed by real-space refinement in Phenix using the same set of parameters. To this stage, the model still contained several steric clashes even with the per-residue geometric assessment by Coot

doi:10.6342/NTU202301899

being acceptable. This was likely due to the compromised local resolution stemming from the dynamic nature of 229E-hAPN complex. To tackle this, ISOLDE[46] was utilized to perform an MD-facilitated flexible fitting for steric clashes minimization. The whole complex was first added hydrogens and simulated by ISOLDE overnight to equilibrate the system, followed by manual checking and tweaking of the individual domains to clear off all the clash warnings reported by ISOLDE. Multiple rounds of adjustment and simulation in ISOLDE were further conducted to procure the final HCoV-229E S-hAPN complex model. Of note, the complex model to this stage was not glycosylated. To build the N-linked glycans onto the model, CHARMM-GUI[47] PDB Reader & Manipulator was used to attach glycan molecules to the asparagine side-chains containing the N-glycosylation sequon. The geometry and the steric clashes of the glycosylated model were again analyzed by Coot[45] and manually refined. To assess the glycosylated model, a comprehensive cryo-EM validation tool in Phenix was employed as a final step. UCSF-ChimeraX and Pymol 2.3.4 (Schrodinger Inc. U.S.A.) were utilized for the structural visualization and rendering of structural representations.

2.11 In-gel proteolytic digestion of the proteins (Digestion protocol A)

3 μg of recombinant S protein were separated by gel electrophoresis and visualized the proteins by staining with One-Step Blue (Biotium, U.S.A.). The gel slices were diced into small pieces (1 mm^3) and treated with 10 mM dithioerythritol (DTE) in 25 mM ammonium bicarbonate at 37 °C for 1 hour, followed by alkylation with 50 mM iodoacetamide (IAM) in 25 mM ammonium bicarbonate for 1 hour at room temperature in the dark. The gel pieces were then de-stained with 25% acetonitrile (ACN) in 25 mM ammonium bicarbonate at 37 °C for 15 minutes, repeating the process until the gel pieces were fully de-stained. Next, the gel pieces were dehydrated with 100% ACN, and the

doi:10.6342/NTU202301899

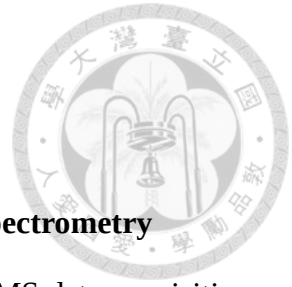
remaining acetonitrile was removed using a Speed Vac, leaving behind dried gel pieces. The dried gel pieces were then treated with sequencing grade trypsin at an enzyme-to-substrate ratio of 1:50 at 37 °C overnight, followed by chymotrypsin at an enzyme-to-substrate ratio of 1:30 at 37 °C on the next day. The digested products were extracted by adding d_2H_2O and sonicating for 15 minutes, and then collected in a clean tube. This extraction process was repeated twice more, replacing d_2H_2O with 1% formic acid (FA) and 50% ACN/1% FA each time. The extracted products were then combined and dried down by Speed Vac. The resulting samples were dissolved in 0.1% FA, cleaned up using ZipTip C18 (Merck Millipore, Ireland), and dried by Speed Vac again for further experiments.

2.12 In-solution proteolytic digestion of the proteins (Digestion protocol B)

Protocol B was derived from the Method described previously[48]. Recombinant S protein and hAPN ectodomain, at a quantity of 5 µg, were subjected to reduction with 10 mM DTE and 6 M Urea in 25 mM ammonium bicarbonate, at a temperature of 37 °C for 1 hour. They were subsequently treated with 50 mM IAM in 25 mM ammonium bicarbonate, in the absence of light, at room temperature for 1 hour. Following this, the concentration of DTE was increased to 50 mM, and the buffer was changed to 25 mM ammonium bicarbonate via Amicon Ultra-0.5, 10 kDa (Merck Millipore, Ireland). The digested products were then exposed overnight to either sequencing grade trypsin and chymotrypsin (T + C) at an enzyme-to-substrate ratio of 1:30 and a temperature of 37 °C, or alpha-lytic protease (aLP) at an enzyme-to-substrate ratio of 1:30 and a temperature of 37 °C. The resulting digested products were further processed by diluting them with 1% FA to a final concentration of 0.1% and cleaning them up using ZipTip C18 (Merck Millipore, Ireland). Finally, the samples were dried down by Speed Vac for use in

doi:10.6342/NTU202301899

subsequent experiments.



2.13 Glycopeptide analysis by liquid chromatography-mass spectrometry

The method previously described[48] was utilized to conduct LC-MS/MS data acquisition on all S and hAPN protein samples, without implementing electron-transfer/higher-energy collision dissociation (EThcD).

2.14 Glycopeptide identification and quantification

The LC-MS/MS raw files (.RAW) were subjected to N-glycopeptide identification using Byonic v3.9.6. The identification parameters were set to search against the sequences of individual S and hAPN proteins with fully specific cleavages at residues determined by the digestion enzymes (e.g., F, K, L, R, W, Y residues for trypsin plus chymotrypsin digested samples, and A, S, T, V residues for alpha-lytic protease digested samples). Up to 2 missed cleavages were allowed in sample digestion. The "Both: CID & HCD" option of the fragmentation type was selected, with a precursor mass tolerance of 5 ppm and a fragment mass tolerance of 10 ppm. The modifications included cysteine carbamidomethylation (+57.0215 Da, at C), methionine oxidation (+15.9949 Da, at M), asparagine and glutamine deamidation (+0.9840 Da, at N, Q). The built-in N-glycan library of "182 human" was used for N-glycopeptide identification. The identification results were exported into Excel files for further processing. The Byonic search results and the MS raw files were utilized in the Byologic module of the Byos suite (v3.11, Protein Metrics Inc., USA) for N-glycopeptide quantification based on the peak areas of extracted ion chromatograms. The final quantification results were also exported into Excel files for later processing.

The Excel files generated from Byonic and Byologic were further analyzed using in-

doi:10.6342/NTU202301899

house Python scripts (available on GitHub at https://github.com/coco0981568491/Master_Thesis/tree/master/Glycopeptide_identification_and_quantification). In brief, positive peptide-spectrum matches (PSMs) were filtered based on the score > 200 and PEP2D < 0.001 criteria. The glycopeptides with identical features, such as N-glycosylation site, sequence, glycan composition, and calculated m/z, were summed by the number of PSMs. The highest-scoring match was considered as a "unique glycopeptide", and its peak area under the extracted ion chromatograms (XICAUC) was recorded as a quantified entry in different N-glycan category figures. Note that any glycopeptide containing more than two NxS/T sequons could not be quantified by our workflow.

2.15 Representative Glycoforms selection of the proteins

To describe the experimentally observed N-glycosylation patterns on the pre-fusion structures of the S proteins as well as the ectodomain of hAPN, it was necessary to select representative glycoforms for specific N-glycosylation sites. Based on the results of glycopeptide analysis described in **2.12**, the most abundant glycoforms of the dominant glycan types (e.g., high-mannose, hybrid, complex, and unoccupied) of each N-glycosylation site under different digestion protocols as well as enzymatic treatments quantified by the peak areas under the extracted ion chromatograms (XICAUC) were all listed in **Tables 3.5-1, 3.6-1, 3.7-1.** The site-specific representative glycoforms were then selected according to the following principles.

To begin with, we established the biological replicates across various experimental batches by considering the variations in the protein digestion protocols and enzymatic treatments. Specifically, we identified samples that underwent identical enzymatic treatments as biological duplicates. For instance, the Ingel-TC and Insol-TC samples of

doi:10.6342/NTU202301899

HCoV-229E Seattle strain S protein, the Insol-TC duplicates of HCoV-229E P100E strain S protein, and the Insol-TC duplicates of hAPN ectodomain were considered as biological replicates. The next step involved choosing the most representative glycoforms based on the biological duplicates with the greatest coverage of N-glycosylation sites. Particularly, we selected Insol-TC #2 of HCoV-229E P100E strain S protein, Insol-TC of HCoV-229E Seattle strain S protein, and Insol-TC #1 of hAPN ectodomain as the most representative glycoforms. In addition, the number of the covered N-glycosylation sites was used to decide the priority of the experimental data. That is, the one with the highest coverage would be the first priority, the medium coverage the second, and so on. The third step involved completing the missing glycoforms by utilizing data from other experimental batches in accordance with the aforementioned data priority. Lastly, any N-glycosylation sites that remained unidentified by our glycopeptide analysis, such as those involving multiple sequons within a glycopeptide resulting in the loss of N-glycosylation profile, were supplemented with alternative Man5 structures. These structures only represented the potential spatial occupancy of glycans and were utilized for modeling purposes. For clarity, the representative glycoforms of each protein for model building were all listed in **Tables 3.5-2, 3.6-2, 3.7-2.** With this, we could further construct experiment-based glycosylated models by CHARMM-GUI[47] as well as GlycoSHIELD[49].

2.16 Atomic model of fully M5 glycosylated proteins by CHARMM-GUI

This was for the schematics showing the site-specific high-mannose content of HCoV-229E P100E strain S, HCoV-229E Seattle strain S, and hAPN ectodomain (**Figs. 3.5-3. B, 3.6-3. B, 3.7-3. B**). The glycans here were solely for demonstrating certain spatial occupancy; thus, only Man5 was used to simplify the glycan building process on CHARMM-GUI. The high-mannose content from high to low was indicated, instead,

doi:10.6342/NTU202301899

with the colors “forest”, orange, and magenta, respectively. The building procedure is briefly described below.

To generate the initial coordinate files for CHARMM-GUI glycosylated model building, the previously published[44] cryo-EM structure of HCoV-229E S (PDB ID: 7CYC) was utilized for P100E strain, whereas for Seattle strain, SWISS-MODEL[37] was used to perform homology modeling based on its protein sequence with 7CYC as the reference structure. In addition, we opted for our own structure for the hAPN ectodomain, considering the differences between the cryo-EM-solved and X-ray-solved models. The discrepancy in the molecular structure stemming from these two distinct experimental methods was elaborated in the RESULTS section. Of note, since 7CYC and our hAPN models contained missing residues due to the poor cryo-EM map quality in certain areas, we had to fill those in by using the built-in tool, “Fit Loop by Rama Search”, in Coot[45] before proceeding to CHARMM-GUI.

Once all the initial coordinate files were ready without any missing residues, we submitted them to CHARMM-GUI PDB Reader & Manipulator provided within the input generator, by which Man5 molecules were attached to all potential N-glycosylation sites, i.e., all asparagine residues located in the NxS/T sequon. To strike a balance between the CHARMM-GUI computation time and the efficiency of this workflow, we only processed 5 N-sites at a time to avoid a waiting time lasting several hours. This strategy applied to all three proteins for glycan modeling.

Finally, all the Man5 fully glycosylated coordinate files were downloaded from CHARMM-GUI server. The glycans on the protein model were manually inspected and adjusted one by one to rectify any steric clashes with the protein structure or the neighboring glycans. In theory, if we processed all N-sites at once, the whole system would be fully glycosylated, and energy minimized to avoid any steric clashes. However,

doi:10.6342/NTU202301899

the complete processing of this approach would be extremely time-consuming. As a trade-off, the former strategy was adopted, followed by manual tweaking of the attached glycans.

The above-mentioned colors, i.e., “forest”, “orange”, and “magenta”, were subsequently applied to the resulting glycosylated models to illustrate the high-mannose content, which was further visualized and rendered via UCSF-ChimeraX[38].

2.17 Glycan shield modeling with GlycoSHIELD

The versatile pipeline for modeling the glycan conformer ensemble onto static protein structures for visualizing and analyzing the glycan shielding effect on protein surfaces were described previously[49]. In the script "GlycoSHIELD-0.1.py", the aforementioned protein models were provided as input PDBs. N-glycosylation sites and the corresponding glycans to be grafted on S proteins were defined in input files based on the representative glycoforms (**Tables 3.5-2, 3.6-2, 3.7-2.**). CG (coarse-grained) mode and the recommended threshold value of 3.5 Å were used. The option "--shuffle-sugar" was also applied. Next, due to the multiple N-glycosylation sites, the script "GlycoTRAJ.py" was executed to merge the output files from "GlycoSHIELD-0.1.py", which were the .XTC and .PDB formats of the glycan frames of each N-glycosylation site on protein structures, into a single trajectory with all grafted glycans. In principle, the smallest number of the grafted glycan frames within all the N-glycosylation sites was applied in the option "--maxframe" of "GlycoTRAJ.py" to remove excessive frames, then the output files of trajectory with static protein and glycan conformer ensemble were exported. Finally, the script "GlycoSASA-0.1.py" was carried out to analyze the glycan shielding effect on protein surfaces, calculating the solvent-accessible surface area (SASA) of un-glycosylated and glycosylated proteins ($SASA_{nogly}$ and $SASA_{gly}$, respectively). The

doi:10.6342/NTU202301899

glycan shielding effect (here followed the original term used in the reference[49]; $SASA_{rel}$ means “relative SASA remained exposed in the presence of the glycans”) was defined as the following function:

$$SASA_{rel} = (SASA_{nogly} - SASA_{gly}) / SASA_{nogly}$$

A shielding value of 1 indicated a complete contribution from glycan shielding, whereas a shielding value of 0 indicated either no shielding from the glycans or a complete shielding from the protein structure (i.e., steric hindrance of the protein obstructs GROMACS[50] SASA probing, which was specifically annotated with an occupancy value of 0 in the output PBDs). The option "--probelist" was set to 0.75 nm, emulating the radius of the average size of hypervariable loops in antibody complementarity determining region (CDR) described previously[51]. The maximum shielding values per residue were written into the b-factor column of the output PDBs. The visualizations of the glycan shielding effect on protein surfaces were accomplished by UCSF-ChimeraX[38].

2.18 Atomic model building of final HCoV-229E virus infection schematic by CHARMM-GUI, GlycoSHIELD, and AlphaFold2

The glycan building procedure using GlycoSHIELD[49] was already described in the previous section. This procedure was applied to the cryo-EM-solved coordinate file of HCoV-229E-S-hAPN complex, with all missing residues being filled in by “Fit Loop by Rama Search” tool in Coot[45], and N-sites being glycosylated based on LC-MS/MS results. The unoccupied sites were modeled with alternative Man5 glycans to depict their spatial occupancy. The output file contained an experiment-based fully glycosylated complex model that was ready for further processing.

Since hAPN proteins are usually anchored in the cellular membrane, forming a homodimer that is responsible for a myriad of physiological functions, such as peptide

doi:10.6342/NTU202301899

metabolism, cell motility and adhesion, HCoV-229E virus entry, and signal transduction and regulation, it is essential to incorporate the membrane structure into our complex model. To achieve this, CHARMM-GUI Membrane Builder[52] was utilized to construct a sheet of lipid bilayer surrounding the cryo-EM-solved hAPN dimer, followed by built-in energy minimization and systematic equilibrium. For HCoV-229E viral surface, LipidWrapper[53] was used to generate a 3-dimensional curved virion lipid membrane modeled upon a cryo-electron tomography (cryo-ET) image of a influenza virus to mimic the topology of HCoV-229E. Moreover, the stalk region and the transmembrane domain (from residue 1034 to residue 1173) of HCoV-229E was modeled using AlphaFold2[54] Google Colab scripts available on https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb. Then the full-length S proteins of HCoV-229E P100E strain were manually inserted into the generated virion lipid membrane, which was neither energy minimized nor equilibrated, yet crudely capture the possible theme of HCoV-229E cell entry event *in vivo*.

Nevertheless, while the resulting fully glycosylated, membrane-anchored HCoV-229E-S-hAPN complex model did not undergo rigorous systematic equilibration and energy minimization, the overall structure and post-translational modifications (PTMs) were buttressed by cryo-EM and LC-MS/MS glycan identification and quantification and were modeled to scale, providing molecular insights into the potential mechanism of virus infection.

2.19 Batch processing and calculation of RBD-up angle of currently available cryo-EM structures of SARS-CoV-2 S protein on Protein Data Bank (PDB)

As of April 3rd, 2023, there were 976 PDB entries solved by cryo-EM that related to

doi:10.6342/NTU202301899

SARS-CoV-2 structures (the database updates constantly. This is merely the number recorded during the analysis). These PDB files resulted from the entry filtering on Protein Data Bank (https://www.rcsb.org/) with "Source Organism Taxonomy Name (Full Lineage)" set to "SARS-CoV-2", "Experimental Method" set to "ELECTRON MICROSCOPY", and "Polymer Entity Type" set to "Protein" in Advanced Search Query Builder to exclude as many non-spike entries as possible before proceeding to in-depth data sorting and analysis. The entries used for the following analysis are available on GitHub (https://github.com/coco0981568491/Master_Thesis/blob/master/Batch_RBD-up_angle_calculation/rcsb_pdb_ids.txt).

Next, we batch downloaded all PBD entries in Crystallographic Information File format (.cif) based on the IDs listed in the above-mentioned text file via a shell script provided by Protein Data Bank (https://github.com/coco0981568491/Master_Thesis/blob/master/Batch_RBD-up_angle_calculation/batch_download_PDB.sh). Since the previous Advanced Search Query Builder settings could not exclude all the non-spike entries, an in-house Python script named "PDB_preprocessor.py" was written to iterate through all entries to inspect the chain composition and the sequence length. To simplify the following analysis, we narrowed down the chain names to A, B and C. Additionally, the coordinate files were also checked for their trimerization and S1/S2 subunits. These criteria were to ensure the common features of spike proteins were present. The protein sequences of the remaining "qualified" entries were subsequently examined for the characteristic residue types and mutations to accomplish SARS-CoV-2 variant identification (for the detailed classification algorithm, please refer to https://github.com/coco0981568491/Master_Thesis/blob/master/Batch_RBD-up_angle_calculation/PDB_preprocessor.py). To this stage, 640 entries remained for the

doi:10.6342/NTU202301899

ensuing geometric analysis.

To batch calculate the RBD-up angle of numerous structures, another Python script named "get_RBD_angles.py" was written to accelerate the process. The rationale behind the analysis is briefly summarized as follows: starting residues of the receptor-binding domain (RBD) of two arbitrary chains of the S model were chosen, and the PyMOL 2.3.4 "get_coords" command (https://pymolwiki.org/index.php/Get_coords) was used to record the coordinates of these residues. The coordinates of the starting residue of RBD on the third chain were obtained by using the "get_coords" command. The coordinates of the two starting residues were subtracted by those of the third to generate 2 vectors. The cross product of these 2 vectors was calculated to obtain the normal vector of the plane consisting of the three starting residues of RBDs. The center of mass (COM) of all three RBDs was found by using the PyMOL "centerofmass" command (https://pymolwiki.org/index.php/Centerofmass). The coordinates of the corresponding starting residues were subtracted from the coordinates of the three COMs of RBDs to obtain three vectors representing the RBD conformation of each chain. The chain-specific angles between these vectors and the normal vector derived above were calculated (for more details, please refer to https://github.com/coco0981568491/Master_Thesis/blob/master/Batch_RBD-up_angle_calculation/get_RBD_angles.py). This workflow was iteratively applied to all 640 entries via the aforementioned Python script. The results were automatically exported to an Excel spreadsheet for later analysis. During the process, 137 entries were excluded due to missing coordinates of the starting residues of RBDs. The remaining 503 entries marked the end of the pipeline, and the calculated RBD-up angles were further sorted and visualized using tools written in a Jupyter Notebook named "RBD_Angle_PostProcessing_andOtherTools.ipynb"

doi:10.6342/NTU202301899

(https://github.com/coco0981568491/Master_Thesis/blob/master/Batch_RBD-up_angle_calculation/RBD_Angle_PostProcessing_andOtherTools.ipynb).

For the visual explanation on the definition and calculation of the RBD-up angles, please refer to **Fig. A20.** in **APPENDIX**.

doi:10.6342/NTU202301899

Chapter 3 RESULTS

3.1 Phylogenetic analysis of HCoV-229E variants

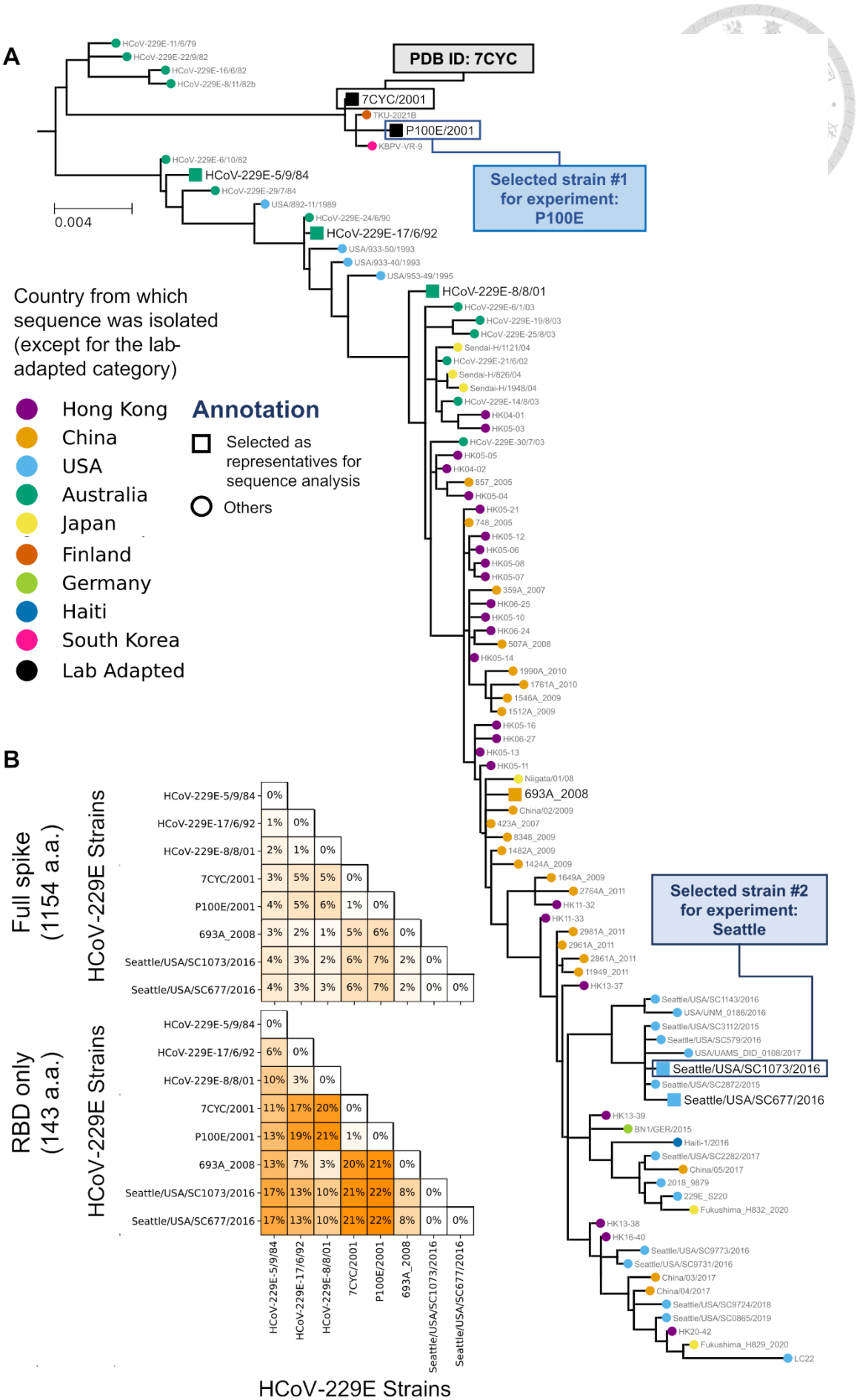
To explore the impact of antigenic drift on the biophysical characteristics and the protein structure of the HCoV-229E spike, we first needed to examine its phylogeny to identify the strains for later experiments. NCBI Virus (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/) was used to gather the up-to-date protein sequences of the spikes of HCoV-229E strains, which were analyzed with the methods proposed by Eguia et al. [30]. Since GenBank: AF344189 and GenBank: AF304460 contain the spike protein sequences of the cryo-EM structures, PDB ID:6U7H[44] and PDB ID:7CYC[11], respectively. The python scripts deposited by the group on GitHub (https://github.com/jbloomlab/CoV_229E_antigenic_drift) were slightly tweaked to incorporate GenBank: AF344189 and GenBank: AF304460 into the dataset for a more comprehensive analysis. A total of 96 sequences were retained for the final inference of the phylogenetic tree of HCoV-229E spikes whereof the earliest strain dated back to 1979 (**Fig. 3.1-1. A**, and the Excel table of the retained sequences for the phylogenetic tree available on GitHub: https://github.com/coco0981568491/Master_Thesis/blob/master/HCoV-229E_accessions/20221011_NCBI_Virus_229E_accessions.csv). The two lab-adapted sequences, GenBank: AF344189 (labeled as “P100E/2001”) and GenBank: AF304460 (labeled as “7CYC/2001”), were grouped to the clade containing GenBank: OK662398 (labeled as “TKU-2021B”) and MZ712010 (labeled as “KBPV-VR-9”), while the rest of the phylogenetic tree structure exhibited the similar “clock-like” and “ladder-like” characteristics mentioned by Eguia et al. [30].

Following the pipeline established by Eguia et al. [30], we chose the same five strains (GenBank: DQ243972, DQ243976, DQ243977, KM055556, KY369909), plus GenBank:

doi:10.6342/NTU202301899

AF344189, AF304460, and KY369913, as the chronological representatives of the evolution indicated by the black texts and square shapes in the phylogenetic tree, whereof the sequences were used to compute the pairwise protein sequence divergence as did by Eguia et al. [30]. The results were visualized by the diagonal heatmaps shown in **Fig. 3.1-1. B**. Note that the protein sequence divergence computed over the receptor-binding domain (RBD) is generally higher than that over the entire sequence, indicating intensive mutations within the RBD. Furthermore, the highest pairwise sequence divergence happened between GenBank: AF344189 and KY369913 (labeled as "Seattle/USA/SC1073/2016") as well as GenBank: AF344189 and KY369909 (labeled as "Seattle/USA/SC677/2016"). We then selected the former pair to carry out the later experiments. The two selected strains are annotated with blue and dark blue text boxes in the phylogenetic tree for clarity.

doi:10.6342/NTU202301899



doi:10.6342/NTU202301899

Figure 3.1-1. Phylogenetic analysis of HCoV-229E variants and the selected strains for the following analyses.

(A) A maximum-likelihood inference using IQ-TREE[55] was used to create a phylogenetic tree of 229E spikes. The tips of the tree are colored according to the country of origin for each virus isolate (except for the lab-adapted one), with black text and square shapes used to indicate the spikes that were used in the experiments. (B) The divergence in protein sequences of the spikes utilized in the experiments was calculated by considering either the full sequence or only the receptor-binding domain (RBD). The divergence was quantified as the Levenshtein distance between the amino acid sequences, divided by the number of sites.

doi:10.6342/NTU202301899

3.2 Antigenic drift and protein stability prediction

To show the distribution of the mutations described in **RESULTS 3.1** on the HCoV-229E spike protein, we performed multiple sequence alignment on these eight selected strains via Clustal Omega[56], and submitted the alignment file to ESPript3.0[57] with PDB ID:7CYC[11] as the reference structure. The results revealed the extensive mutations in receptor-binding domain (RBD) (**Figs. 3.2-1. A, 3.2-2. A**), corroborating the raised sequence divergence shown in the diagonal heatmap (**Fig. 3.1-1. B**). Moreover, several mutations are concerned with the canonical sequon of N-linked glycosylation (NxS/T), which is a common post-translational modification (PTM) on proteins, that is, these mutations have the potential to alter the glycan composition of the spikes (**Fig. 3.2-1. B**). The residue conservation score from multiple sequence alignment was mapped onto the reference structure (PDB ID: 7CYC), with the color ramping from red to white to represent incrementing identity of the protein sequence (**Fig. 3.2-2. A**). RBD and domain A were two less conserved regions where several red patches were present, indicating a lower conservation score.

To illustrate the distribution of all the mutations related to N-linked glycosylation, GenBank: AF344189 and KY369913 were chosen for their highest sequence divergence, and the side chains of the mutated sites were styled in “sphere” by using ChimeraX[38]. Due to the missing protein structure around Asn 22, only five out of six mutations pertaining to N-linked glycosylation, namely D51N, D111N, T176N, K488N, and N714K, were shown and colored in tomato (#FF6347), while the other 66 mutations were colored in yellow, making it 72 mutations in total (**Fig. 3.2-2. B**). In line with the conservation score distribution in **Fig. 3.2-2. A**, most of the mutations occur in RBD and domain A, while others can be found in the central helices that might affect the structural packing of the spike protein.

doi:10.6342/NTU202301899

The Omicron variant of SARS-CoV-2 bears more than 30 mutations within the spike protein, which modify the structural properties and the biochemical characteristics to various extents[58, 59]. Now we want to see if the 72 mutations engender any difference between GenBank: AF344189 and KY369913.

The mutations on protein tend to have certain impact on its stability, and a common measure to quantify the effect is to calculate the unfolding free energy difference between the wild type and mutant protein ($\Delta\Delta G = \Delta G_{wildtype} - \Delta G_{mutant}$). Several computational tools have been developed to facilitate *in-silico* ΔΔG quantification, and the mutational effect is usually classified into stabilizing, destabilizing, or neutral, based on its yielded value[33, 35, 36, 60, 61]. Albeit the tremendous endeavors invested in improving the performance of ΔΔG prediction, most of the current predictors are far from ideal in terms of their correctness and reliability[32]. Moreover, due to the difficulty in measuring experimentally the unfolding free energy difference on a large scale, the empirical datasets of ΔΔG are often found to be limited and unbalanced toward the destabilizing mutations, which might give rise to the prediction bias seen in the tools assessed by Marabotti, A., et al. [32]. A rough, yet effective approach to make up for the unreliability of the tools is to reach a consensus prediction by combining the results from several predictors, as suggested by Marabotti, A., et al. [32]. Here we chose three online tools, DDGun (includes both sequence-based and structure-facilitated predicting methods) [36], INPS-MD (includes the original sequence-based INPS[35] and the newly-added INPS-3D) [34], and MAESTRO (command-line version based on a structure-facilitated method) [62], to quantify the effect of the antigenic drift on the previously selected HCoV-229E strains. For all the prediction results, please refer to **Table 3.2-1.** The mutations between the following strain pairs: DQ243972 and AF304460 (denoted as “1984 to 7CYC”), DQ243972 and AF344189 (denoted as “1984 to P100E”), DQ243972 and DQ243976 (denoted as “1984 to 1992”), DQ243972 and DQ243977

doi:10.6342/NTU202301899

(denoted as “1984 to 2001”), DQ243972 and KM055556 (denoted as “1984 to 2008”), DQ243972 and KY369913 (denoted as “1984 to Seattle”), DQ243972 and KY369909 (denoted as “1984 to SC677”), and AF344189 and KY369913 (denoted as “P100E to Seattle”), served as the inputs for the aforementioned online predictors to assess the influence of the mutations. As for the structure-facilitated methods, PDB ID: 7CYC[11] was submitted as the reference protein structure. The prediction results were visualized by the bar chart in **Fig. 3.2-2. C** and grouped by the above-mentioned strain pairs.

According to Marabotti, A., et al. [32], the predicted ΔΔG values that fall within the range ± 0.5 kcal/mol should be considered unreliable due to their proximity to the experimental error[63]. This standard was applied to all the predictors, and only the values outside the range were classified as either “destabilizing” (ΔΔG > 0.5 kcal/mol, signified with the light red region in **Fig. 3.2-2. C**) or “stabilizing” (ΔΔG < 0.5 kcal/mol, signified with the light green region in **Fig. 3.2-2. C**) mutations, while the rest were deemed “neutral” (signified with the light grey region in **Fig. 3.2-2. C**) as the effect of the mutations remained undetermined or insignificant. Despite the apparent variations in the ΔΔG values from different predictors, the overall trend still showed an increasing destabilizing effect as the time interval between the strains lengthens, which is most evident among the MAESTRO[62] predictions. Furthermore, “P100E and Seattle” was the sole strain pair predicted to bear the destabilizing mutations by all the tools. Therefore, we chose them for the later *in-vitro* experiments. For convenience, we will refer to GenBank: AF344189 as “P100E strain” and GenBank: KY369913 as “Seattle strain”. It is worth noting that the antibody neutralizing assay is not included in this study to directly prove the existence of immune escape between P100E and Seattle strains. Nevertheless, Eguia et al. [30] demonstrated that the neutralizing activity of human sera is lower against “future” viruses than those that elicited the immunity, and the immune escape was indeed

doi:10.6342/NTU202301899

present between HCoV-229E-5/9/84 and Seattle/USA/SC677/2016 (**Fig. 3.1-1. A**). Additionally, based on the phylogenetic tree, P100E strain belongs to an even older clade dated back to year 1979, and the sequence of Seattle strain (Seattle/USA/SC1073/2016) is highly similar to that of Seattle/USA/SC677/2016. Thus, we deduced that the numerous mutations between P100E and Seattle strains should also lead to immune escape, i.e., one of the salient ramifications of antigenic drift.

doi:10.6342/NTU202301899

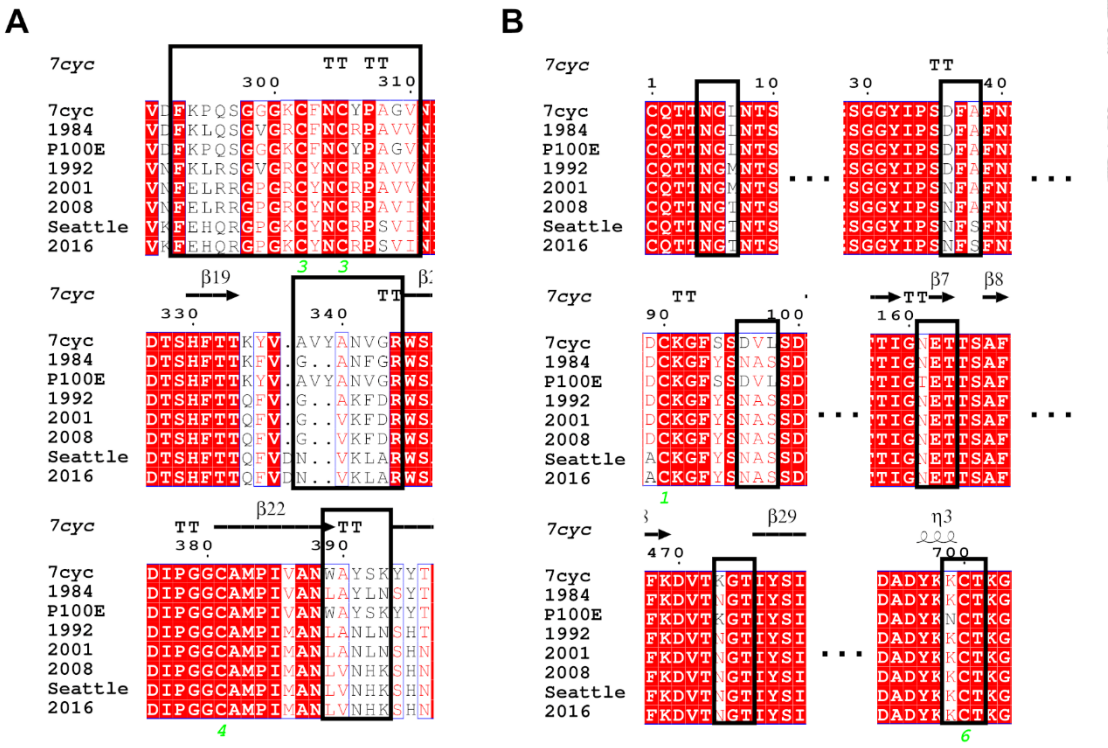


Figure 3.2-1. Multiple sequence alignment of the eight selected HCoV-229E strains using ESPript3.0.

The eight selected strains were listed on the left side of the multiple sequence alignment result. The mutations related to the three receptor binding loops of RBD and the ones located in the N-linked glycosylation sequon (N-X-S/T, X is not proline) were delineated by the black boxes in (A) and (B), respectively. The full multiple sequence alignment profile can be found in **Fig. A1**.

doi:10.6342/NTU202301899

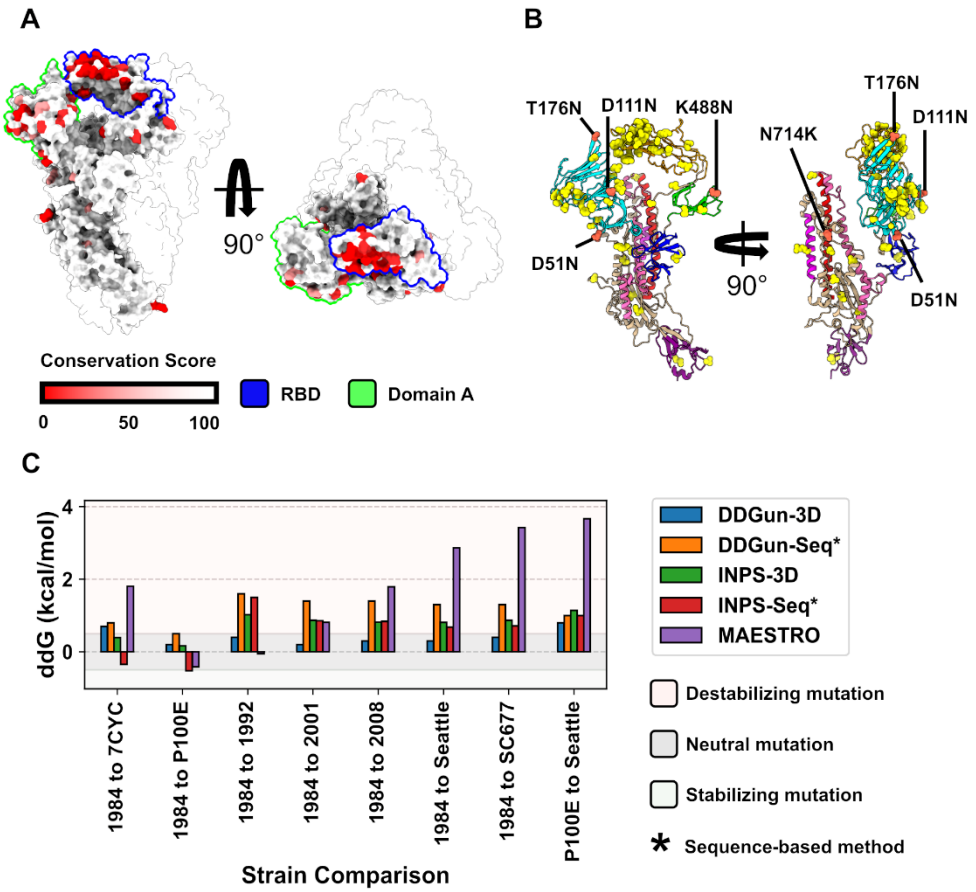


Figure 3.2-2. The distribution of the mutations and their predicted effects on the protein stability.

(A) The conservation score derived from the multiple sequence alignment in **Fig. A1** was mapped onto the cryo-EM molecular structure of HCoV-229E S (PDB ID: 7CYC[11]). The protein surface was colored in a gradient from red to white, indicating a low to high conservation score. For clarity, the receptor binding domain (RBD) and domain A were delineated in color blue and lime, respectively. (B) There are 2 deletions and 72 point mutations from P100E strain to Seattle strain. The mutations were mapped onto PDB ID: 7CYC[11] structure and colored in yellow. 5 out of the 6 mutations related to N-linked glycosylation sequon were colored in tomato (#FF6347) and annotated by their corresponding protein residue names. The missing mutation was residue 22, which is not structurally available in 7CYC[11] due to the compromised cryo-EM map quality. (C) Pairwise delta delta Gibbs free energy (ddG) predictions were made for the selected eight strains using five *in-silico* predictors, with the sequence-based algorithm marked by an asterisk. The background of the plot was colored in light pink, light grey, and light green to respectively show the effect of the mutations on the protein stability.

doi:10.6342/NTU202301899

Table 3.2-1. HCoV-229E S protein stability estimation via ΔΔG servers and command-line tool.

PROTEINS	SERVERS	ddG	PREDICTION	NOTES
DQ243972 (1984) to 7CYC	MAESTRO	1.81	destabilizing	confidence: 0.659045
DQ243972 (1984) to 7CYC	INPS-3D	-0.39	neutral	
DQ243972 (1984) to 7CYC	INPS-Seq*	0.35	neutral	
DQ243972 (1984) to 7CYC	DDGun-Seq*	-0.80	destabilizing	
DQ243972 (1984) to 7CYC	DDGun-3D	-0.70	destabilizing	
DQ243972 (1984) to P100E	MAESTRO	-0.42	neutral	confidence: 0.679160
DQ243972 (1984) to P100E	INPS-3D	-0.17	neutral	
DQ243972 (1984) to P100E	INPS-Seq*	0.52	stabilizing	
DQ243972 (1984) to P100E	DDGun-Seq*	-0.50	neutral	
DQ243972 (1984) to P100E	DDGun-3D	-0.20	neutral	
DQ243972 (1984) to 1992	MAESTRO	-0.06	neutral	confidence: 0.740918
DQ243972 (1984) to 1992	INPS-3D	-1.02	destabilizing	
DQ243972 (1984) to 1992	INPS-Seq*	-1.50	destabilizing	
DQ243972 (1984) to 1992	DDGun-Seq*	-1.60	destabilizing	
DQ243972 (1984) to 1992	DDGun-3D	-0.40	neutral	
DQ243972 (1984) to 2001	MAESTRO	0.81	destabilizing	confidence: 0.732311
DQ243972 (1984) to 2001	INPS-3D	-0.87	destabilizing	
DQ243972 (1984) to 2001	INPS-Seq*	-0.86	destabilizing	
DQ243972 (1984) to 2001	DDGun-Seq*	-1.40	destabilizing	
DQ243972 (1984) to 2001	DDGun-3D	-0.20	neutral	
DQ243972 (1984) to 2008	MAESTRO	1.79	destabilizing	confidence: 0.696628
DQ243972 (1984) to 2008	INPS-3D	-0.82	destabilizing	
DQ243972 (1984) to 2008	INPS-Seq*	-0.84	destabilizing	
DQ243972 (1984) to 2008	DDGun-Seq*	-1.40	destabilizing	
DQ243972 (1984) to 2008	DDGun-3D	-0.30	neutral	
DQ243972 (1984) to SC1073 (Seattle)	MAESTRO	2.87	destabilizing	confidence: 0.685937
DQ243972 (1984) to SC1073 (Seattle)	INPS-3D	-0.81	destabilizing	
DQ243972 (1984) to SC1073 (Seattle)	INPS-Seq*	-0.68	destabilizing	
DQ243972 (1984) to SC1073 (Seattle)	DDGun-Seq*	-1.30	destabilizing	
DQ243972 (1984) to SC1073 (Seattle)	DDGun-3D	-0.30	neutral	
DQ243972 (1984) to SC677	MAESTRO	3.43	destabilizing	confidence: 0.623254
DQ243972 (1984) to SC677	INPS-3D	-0.87	destabilizing	
DQ243972 (1984) to SC677	INPS-Seq*	-0.71	destabilizing	

doi:10.6342/NTU202301899

PROTEINS	SERVERS	ddG	PREDICTION	NOTES
DQ243972 (1984) to SC677	DDGun-Seq*	-1.30	destabilizing	
DQ243972 (1984) to SC677	DDGun-3D	-0.40	neutral	
P100E to Seattle *seq-based	MAESTRO	3.67	destabilizing	Full-L S trimer
P100E to Seattle *seq-based	INPS-3D	-1.14	destabilizing	S monomer chain A (sum > 0.5/N)
P100E to Seattle *seq-based	INPS-Seq*	-1.00	destabilizing	FASTA (sum >0.5/N)
P100E to Seattle *seq-based	DDGun-Seq*	-1.00	destabilizing	FASTA (sum >0.5/N)
P100E to Seattle *seq-based	DDGun-3D	-0.80	destabilizing	S monomer chain A

doi:10.6342/NTU202301899

3.3 Expression and validation of HCoV-229E S proteins and hAPN ectodomain

Surface glycosylation is a crucial and species-specific[64] post-translational modification (PTM) responsible for protein folding and function[65]. Besides, some mutations between P100E and Seattle are associated with the N-liked-glycosylation sequon as mentioned before. We thus decided to express the spike proteins of both P100E and Seattle strain via Expi293™ Expression System to emulate the human glycosylation pattern on the protein surface. To simplify the target system, only the ectodomains (residues 1-1113 for P100E; residue 1-1111 for Seattle with 2 deletions)-of the spikes were produced as fusion proteins with a C-terminal T4 fibritin foldon trimerization domain[66] **(Figs. 3.3-1 B-C**). Additionally, dual-proline mutation (2P) was introduced into both sequences (^{871}TI^{872}→^{871}PP^{872} for P100E S, ^{869}TI^{870}→^{869}PP^{870} for Seattle S with 2 deletions) to stabilize the pre-fusion conformation[67, 68]. Please refer to **Fig. 3.3-1** for the detailed schematic of the construct design. Utilizing the His-tags in the constructs, the soluble spikes produced by Expi293™ cells were collected from the cell medium, and preliminarily purified with HisPur™ cobalt resin, followed by further purification with size-exclusion chromatography (SEC) using Superose® 6 Increase 10/300 GL column. The SEC profiles of Expi293™ expressed spike glycoproteins are shown in the top two rows of **Fig. 3.3-2**. The estimated molecular weight of both spike monomers is around 128 kDa based on their pure amino acid sequences in a single chain, making the total molecular weight of the spike trimers 384 kDa. Nevertheless, the elution volumes of both spike trimers exceeded that of the SEC standard marker of 669 kDa, which indicates the potential contribution to the overall molecular weight from protein glycosylation. Notwithstanding the cone-shaped structure of the spikes that might complicate the relationship between the elution volume and the molecular weight, the spike proteins have been widely reported to be highly glycosylated[44, 69, 70], which may also suggest the added

doi:10.6342/NTU202301899

molecular weight from the extensive glycans. On the right side of the SEC profiles are the corresponding SDS-PAGEs for each spike protein. In agreement with the increased molecular weight observed in SEC, both spike monomers located above the marker indicating 170 kDa, with smeared bands implying the heterogeneity of the glycans.

Unlike SARS-CoV-2, the receptor of the HCoV-229E S is human aminopeptidase N (hAPN), instead of ACE2. Since the ultimate goal of the protein expression is to prepare the needed materials for the following structural analysis aiming to shed light on the interactions between HCoV-229E S and its receptor, the ectodomain of hAPN (residues 66-967) was also expressed, purified, and crudely validated according to the above-mentioned procedure (the bottom row in **Fig. 3.3-2**).

Another way to assess the protein integrity is to test its thermal stability via differential scanning fluorimeter (DSF). The protein will be subjected to a thermal ramping from 35 ºC to 95 ºC to disrupt the molecular structure (**section 2.5**). The DSF profile can provide a fingerprint of the potential unfolding procedure of the tested protein. Here we conducted DSF experiment on the S proteins of both P100E and Seattle strains. The results are shown in **Fig. 3.3-3.** Although both spike proteins belong to the same virus, the DSF profiles exhibit different overall patterns. The P100E strain might underwent a two-step unfolding process, while the Seattle strain only underwent one. Since the structural alteration during heating could be a complex and multi-step process, a more straightforward method to directly inspect the protein structure is warranted. For this, we resorted to negative staining electron microscopy (NSEM).

To prepare the samples for NSEM, the eluted fractions matching the main peaks of SEC profiles (denoted with the light grey background in **Fig. 3.3-2**) were collected and concentrated to about 0.05 mg/mL. The representative NSEM micrographs of both strains as well as the yielded 2D classification results from CryoSPARC[41] were shown in **Fig.**

doi:10.6342/NTU202301899

3.3-4. A. The P100E spike particles in the left micrograph were more compact and intact compared to those of the Seattle strain in the right micrograph, which also manifested in the plunge in the number of the selected particles per micrograph of Seattle relative to that of P100E (**Fig. 3.3-4. B**). The discrepancy between the two strains in the structural integrity was made extra evident by the selected 2D classes shown under the NSEM micrographs, as many spikes of Seattle formed an unusually elongated shape while those of P100E displayed a typical cone shape seen in the pre-fusion conformation of a HCoV-229E spike protein[11, 44]. To collate the structure of our spike protein to that of the latest PDB entry, the selected 2D classes of P100E were used to generate ab-initio models, which were later filtered and refined by heterogeneous refinement and non-uniform refinement in CryoSPARC[41, 42]. The processed Coulomb potential map was exported, and rigid body docked to the molecular model of PDB ID: 7CYC (**Fig. 3.3-4. C**). Albeit the low resolution of NSEM map, which is inherently limited by the grain size of the negative stain, the general shape and the domains of the spike were still enclosed by the map. In addition, the low-contoured NSEM map showed the upper part of T4 fibritin foldon trimerization domain, whereof the molecular structure was absent in PDB ID: 7CYC due to its dynamics that made it rather challenging for cryo-EM. Of note, although most of the observed structure of Seattle strain S proteins was seemingly damaged compared to that of P100E strain S proteins, there were still particles constituting a single 2D class exhibiting the common spike protein features. This might imply either a post-translational protein structure distortion or a mixture of well-folded and unfolded products generated during the protein biosynthesis. Here we hypothesized that the compromised stability of Seattle strain S proteins was due to the extensive mutations in the protein sequence. Nevertheless, more experiments are needed to pinpoint the underlying cause of the dissimilar protein stability of P100E and Seattle strains.

doi:10.6342/NTU202301899

As for hAPN, NSEM was also employed to examine its protein structure. The results are shown in **Fig. 3.3-5.** The common states of hAPN are apo-form and homodimer[16], which were both found in the 2D classification results of its NSEM micrographs. However, 2D classes of trimer formation were also observed. This might be an artifact stemming from the relatively concentrated sample used in this batch (**Fig. 3.3-5. A**). Just as the spike proteins, the calculated Coulomb potential map of hAPN was rigid body fitted to the previously reported X-ray structure (PDB ID: 6U7G). The main features were enclosed, validating the identity and the structural integrity of hAPN (**Fig. 3.3-5. B**).

Lastly, to guarantee that the expressed proteins are not only structurally correct but also functionally competent, biolayer interferometry (BLI) was deployed to assess the binding affinity between the S proteins of P100E and Seattle strains and the ectodomain of hAPN. The resulted BLI sensorgrams are shown in **Fig. 3.3-6.**, in comparison with the binding affinity of SARS-CoV-1-ACE2 pair. In line with the deformed protein structure of Seattle strain S proteins, the BLI signal of which was significantly weaker relative to that of P100E strain S proteins (**Figs. 3.3-6. A-B**), indicating the disrupted structure indeed took its toll on the protein functionality, and that the binding signal might come from the remaining intact Seattle spike proteins observed in the single 2D class of NSEM (**Fig. A12.**).

doi:10.6342/NTU202301899

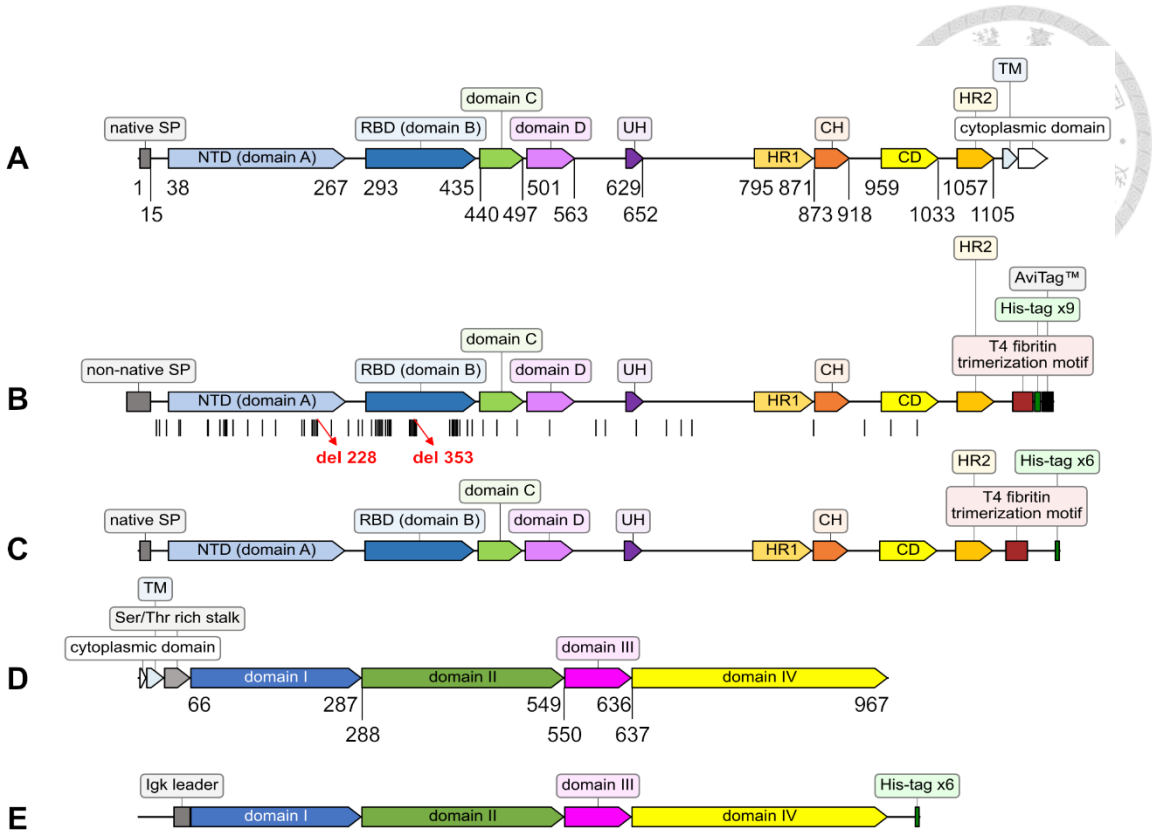


Figure 3.3-1. Construct design of HCoV-229E P100E strain, Seattle strain, and hAPN ectodomain.

The essential domains of HCoV-229E S and human aminopeptidase N (hAPN) are shown in (A) and (D), respectively. Both are membrane proteins. (B) shows the construct design for HCoV-229E P100E strain S, which only includes the ectodomain of the S protein, with a T4 fibritin trimerization motif followed by a nona-repeat histidine (His 9) tag to facilitate protein purification. As for HCoV-229E Seattle strain S ectodomain, a His 6 tag was used instead to server the same function, and the construct design is as shown in (C). Finally, (E) shows the construct for hAPN ectodomain, where the IgK leader sequence is used as the signal peptide for the expression of the soluble form of the protein. Of note, in (B), the 72 point mutations of the Seattle strain were marked with black vertical lines on the protein sequence of P100E strain, and the 2 deletions were annotated with the corresponding residue numbers in red.

doi:10.6342/NTU202301899

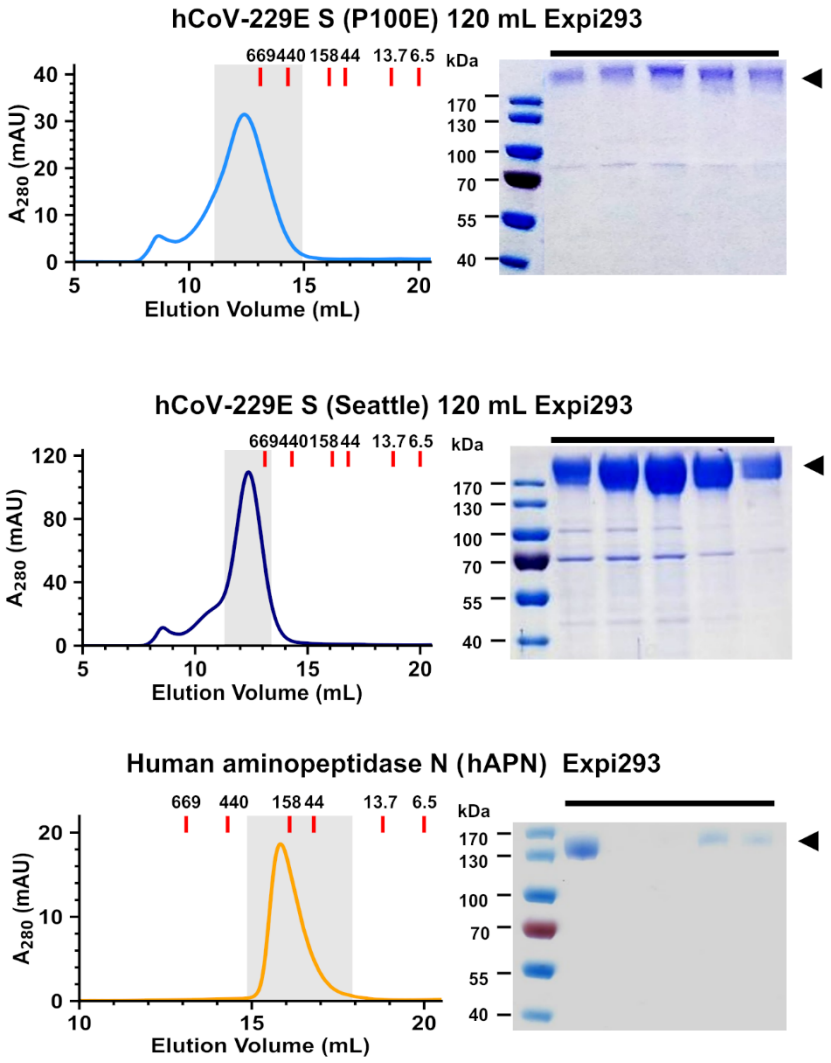


Figure 3.3-2. Size exclusion chromatography (SEC) and electrophoresis analysis (SDS-PAGE) of HCoV-229E P100E strain, Seattle strain, and hAPN ectodomain. Size-exclusion chromatograms of HCoV-229E P100E strain S, HCoV-229E Seattle strain S, and hAPN ectodomain are shown from top to bottom. The collected elution volume range was annotated by grey color. The individual fractions were further analyzed by a 4-12% gradient SDS-PAGE (right panel), no which the black arrows indicate the positions of the corresponding proteins.

doi:10.6342/NTU202301899

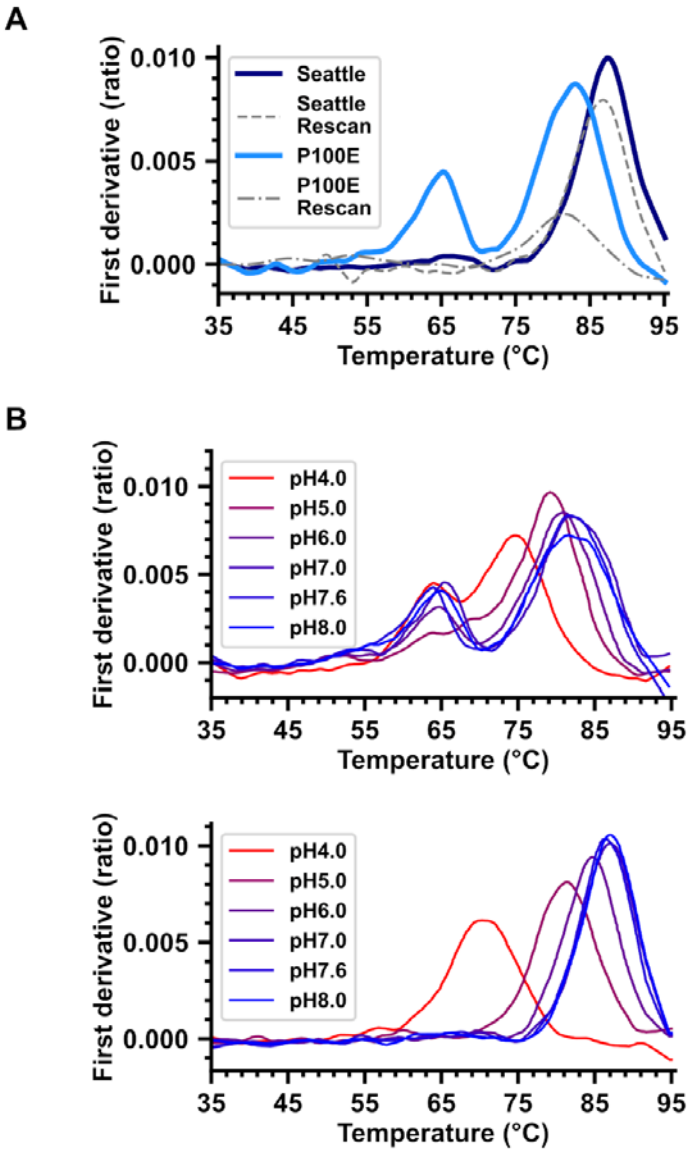


Figure 3.3-3. Thermal stabilities of HCoV-229E P100E strain S protein and Seattle strain S protein by differential scanning fluorimeter (DSF).

(A) DSF profiles of HCoV-229E P100E strain S (light blue) and Seattle strain S (dark blue) proteins. (B) DSF profiles of P100E and Seattle strains as a function of pH values from top to bottom.

doi:10.6342/NTU202301899

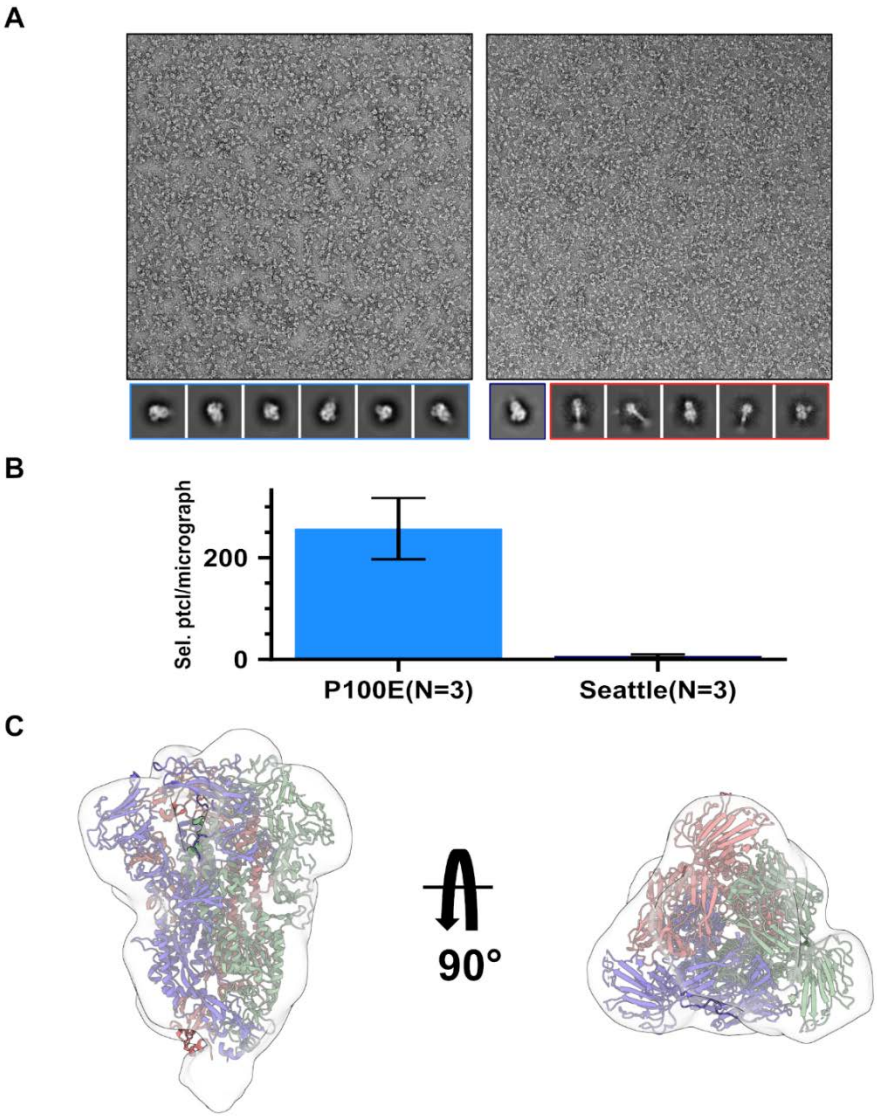


Figure 3.3-4. Negative-stain EM (NSEM) of HCoV-229E P100E strain S protein and Seattle strain S protein.

(A) shows the representative micrographs of NSEM and the selected 2D classes of picked particles of P100E (left) and Seattle (right) strains. For Seattle strains, the deformed S proteins are outlined by the red box. The results of NSEM data processing were further quantified in (B) by plotting the strains against the number of the selected particles per micrograph, which revealed a stark contrast in the structural integrity of the S between the two strains (N=3). (C) The molecular structure of PDB ID:7CYC was rigid body fitted into the derived Coulomb potential map from NSEM datasets of P100E strain.

doi:10.6342/NTU202301899

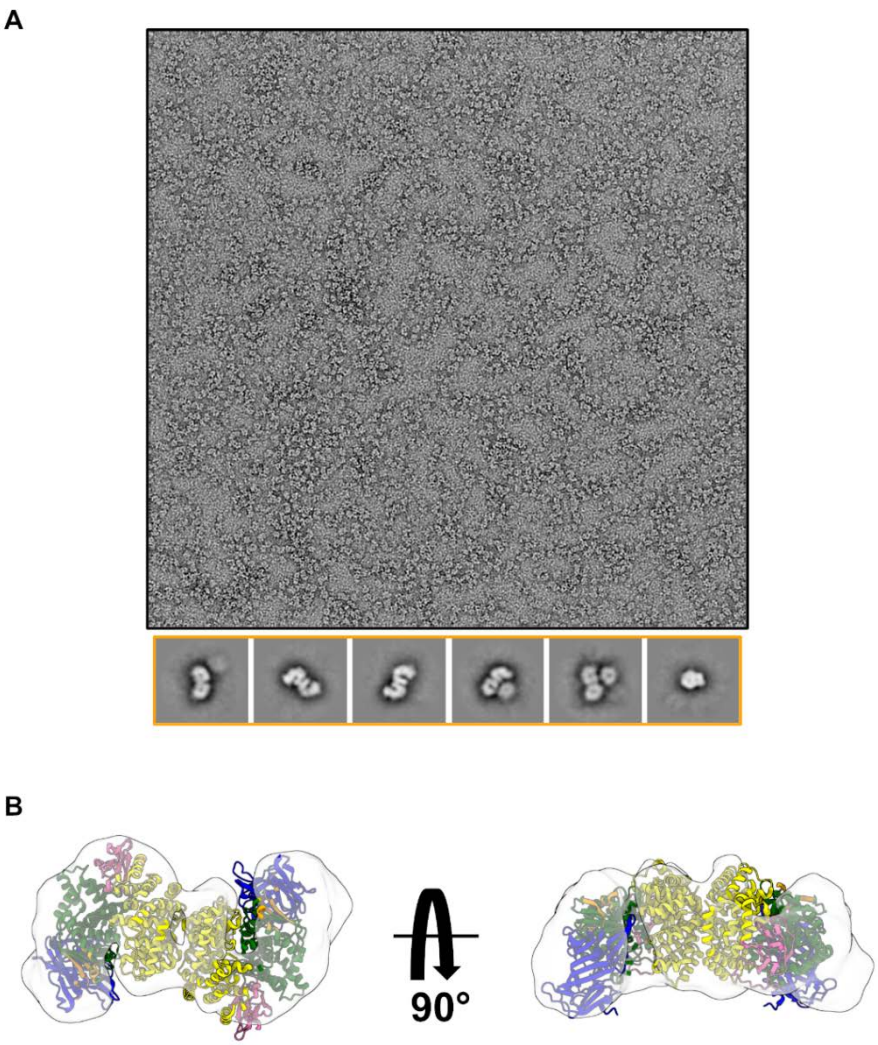


Figure 3.3-5. Negative-stain EM (NSEM) of human aminopeptidase N (hAPN).
(A) shows the representative micrographs of NSEM and the selected 2D classes of picked particles of hAPN ectodomain. (B) The molecular structure of PDB ID:6U7G was rigid body fitted into the derived Coulomb potential map from NSEM datasets of hAPN ectodomain.

doi:10.6342/NTU202301899

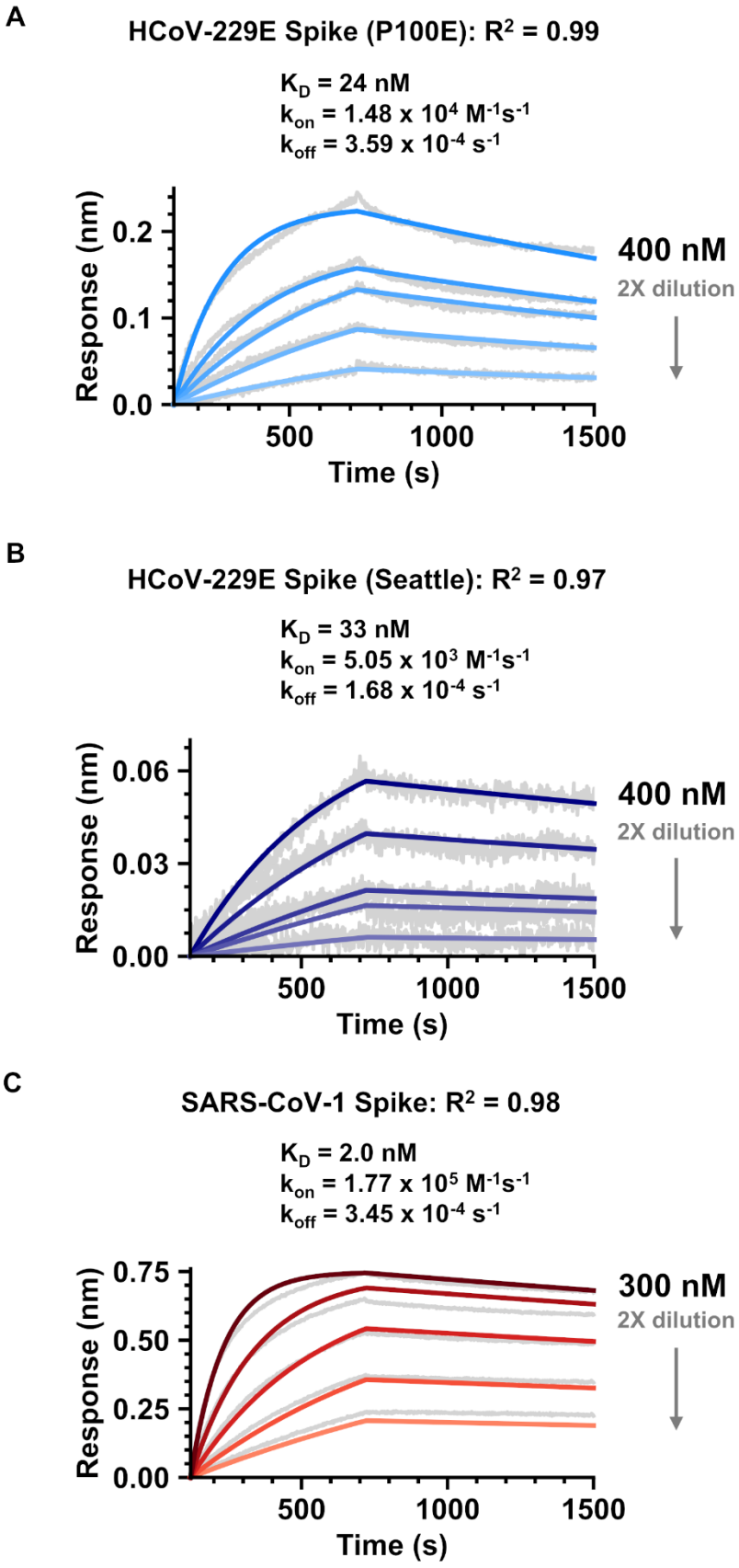


Figure 3.3-6. Biolayer interferometry (BLI) to examine the binding affinity between HCoV-229E strains and hAPN.

doi:10.6342/NTU202301899

BLI sensorgrams of P100E strain S (A) and Seattle strain S (B) binding to hAPN, and SARS-CoV-1 S (C) binding to ACE2. hAPN and ACE2 were immobilized on the sensor tip. The highest concentrations used for the three S proteins are indicated on the right in grey bold font. The remaining four concentration gradients were prepared from the highest concentration by 2-fold serial dilution for independent BLI binding assays. The dissociation constant (K_D), k_{on}, and k_{off} derived from global fitting of the sensorgrams are shown above, with the light-grey experimental signals overlaid with the fitted data shown in colored lines.

doi:10.6342/NTU202301899

3.4 The workflow of N-glycopeptide analysis

The overall workflow is illustrated in **Fig. 3.4-1**. To summarize, the different digestion protocols (e.g., Ingel versus Insol digestion), the enzyme treatments (e.g., trypsin, chymotrypsin, and alpha-lytic protease), were carried out on different S recombinant proteins and hAPN ectodomain to generate glycopeptide samples. These samples were then analyzed using liquid chromatography with tandem mass spectrometry (LC-MS/MS). The resulting MS raw files were searched using Byonic for N-glycopeptide identification, and the Byologic module of the Byos suite was used for N-glycopeptide quantification, followed by the processing of a python script (https://github.com/coco0981568491/Master_Thesis/blob/master/Glycopeptide_identification_and_quantification/BBP_PostProcessing.ipynb) to efficiently remove the PSMs unqualified for our criteria and to sort out the "unique N-glycopeptide" for each site. The accepted data were categorized into different glycan types to illustrate the processed status of each N-glycosylation site. Representative glycoforms for each N-glycosylation site were selected based on the most abundant glycoforms from the glycopeptide analysis, and these observed glycoforms were mapped onto the corresponding protein structures using CHAEMM-GUI. This mapping revealed the relative positions of the N-linked glycans on the protein surfaces.

The N-glycosylation patterns were visualized in different formats. Firstly, the normalized bar charts of the individual batches of experiments with distinct digestion protocols as well as enzyme treatments were shown in **APPENDIX Figs. A2-A10**, in which the "unique glycopeptides" were classified into the previously described[71] N-glycan categories, including high-mannose type glycan series (Man9 to Man5), afucosylated and fucosylated hybrid type glycans (Hybrid and Fhybrid), as well as complex type glycan series according to the number of antennae and the modification of

doi:10.6342/NTU202301899

core fucose (A1 to FA4/FA3B). Furthermore, the results corresponding to the selected representative glycoforms were sorted and presented in **Figs 3.5-1**, **3.6-1**, and **3.8-1**. Secondly, to showcase the variation and reproducibility between different batches of experiments, normalized bar charts illustrating the overall batch-to-batch glycosylation patterns were presented in **Figs 3.5-2**, **3.6-2**, and **3.8-2**, in which the "unique glycopeptides" were classified into another N-glycan categories previously described[48], showing the N-glycosylation processed degree from low to high (Man9 to Man5 and 3HexNAc to 8HexNAc, without considering the number of Hex, Fuc, and NeuAc). Thirdly, a simplified pie chart version classifying the "unique glycopeptides" into the high-mannose type (Man5-Man9), hybrid type (3HexNAc, without considering the number of Hex, Fuc, and NeuAc), and complex type (4-8HexNAc, without considering the number of Hex, Fuc, and NeuAc), which presented the results corresponding to the selected representative glycoforms, were shown in **Figs 3.5-3 A**, **3.6-3 A**, and **3.8-3 A**. Finally, the high-mannose content of the selected representative glycoforms on the CHARMM-GUI-derived fully glycosylated S protein models was presented in **Figs 3.5-3 B**, **3.6-3 B**, and **3.8-3 B**, demonstrating the proportion of high-mannose glycans in these glycoforms.

doi:10.6342/NTU202301899

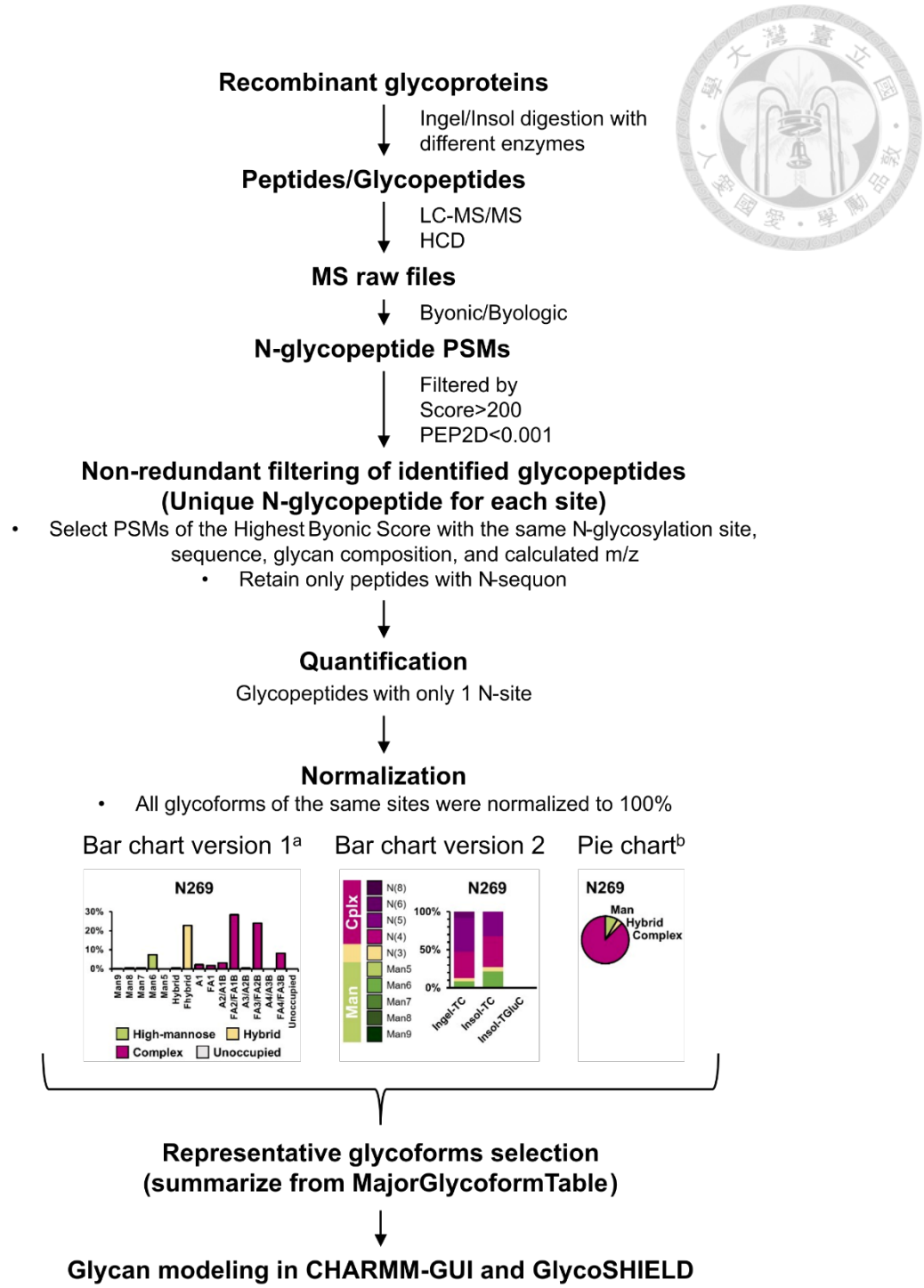


Figure 3.4-1. The workflow of N-glycopeptide analysis.

[a]. The experimental results of each batch, as well as the final sorted-out results corresponding to the representative glycoforms, were presented in both versions 1 and 2 of the bar chart. [b]. The pie chart depicted solely the experimental results that corresponded to the representative glycoforms.

doi:10.6342/NTU202301899

3.5 Site specific N-glycopeptide analysis of HCoV-229E P100E strain S

A pivotal aspect of protein structure is surface glycosylation, especially for HCoVs in which the glycans aid not just folding but immune escape[69, 72]. As previously mentioned in **section 3.2**, six mutations were found located within the N-linked glycosylation sequon when performing the sequence alignment between P100E strain S and Seattle strain S, indicating the variation in glycosylation profile throughout the evolution. To delve deeper into this topic, liquid chromatography with tandem mass spectrometry (LC-MS/MS) was employed to investigate the disparity of N-linked glycosylation between P100E and Seattle strain S proteins. Here, we first discuss the analysis for the former.

P100E S protein contains 30 theoretical N-glycosylation sites based on the NxS/T sequon prediction. To determine site-specific glycosylation status, different digestion protocols, including Ingel-TC (protocol A described in **2.11**) and Insol-TC (protocol B described in **2.12**), were utilized for glycopeptide sample preparation. The prepared samples would then be further analyzed by LC-MS/MS as well as the processing of MS search engines and the data sorting by Python scripts described in **sections 2.13-15.**

The N-linked glycans on HCoV-229E P100E strain S mostly belong to the high-mannose type, including sites N62, N98, N171, N220, N326-N518, N568-N663, N714, N930, and N1061-N1096 (**Figs. 3.5-1., 3.5-2.**). It has been reported that the steric hindrance from the protein structure or the neighboring glycans can hinder the enzymatic processing and trimming of Man9 inside the endoplasmic reticulum (ER) [73]. We hypothesized that the abundant high-mannose type glycans discovered on P100E S protein were partially due to the densely distributed N-sites on its surface. Additionally, as visualized in **Fig. 3.5-3. B**, N62 and N930 sit inside the cavities created by the closed RBD and the S2 central helices, respectively. This might contribute to the limited accessibility from the glycan processing enzymes, leading to the presence of high-

doi:10.6342/NTU202301899

mannose type glycans. In contrast, a combination of high-mannose/hybrid and complex-type glycans was observed at N147, N243, and N671 sites (**Fig. 3.5-2**). Notably, N147 site exhibited the highest prevalence of complex-type glycans, supported by three experimental repetitions. This suggests a less obstructed environment for enzymatic processing in the Golgi apparatus. It is worth noting that our workflow could not quantify glycopeptides containing two NxS/T sequons. While Byonic could assign glycans to these peptides and provide information about the overall glycan configuration, the distribution of glycans between the two N-glycosylation sites on a single peptide chain remained unknown. As a result, these N-sites are denoted as "N/A" similar to the unidentified ones and are excluded from the subsequent discussion.

doi:10.6342/NTU202301899

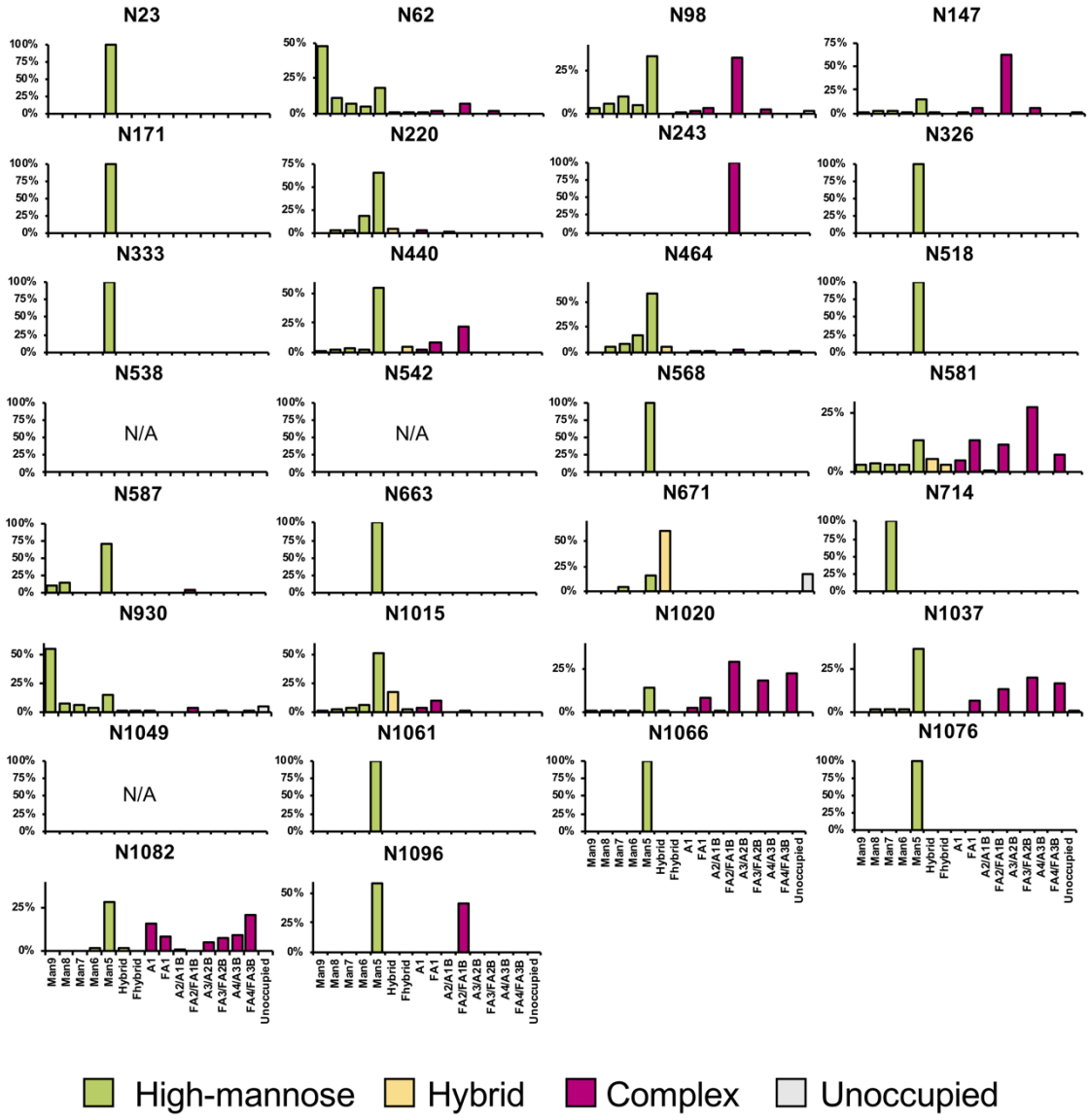


Figure 3.5-1. Quantification of the site-specific N-linked glycosylation analysis of the HCoV-229E P100E strain S protein.

The bar charts provide an overview of the relative quantities of N-glycans identified through LC-MS/MS. The charts only display results corresponding to the selected representative glycoforms, including overall data from Insol-TC #2 and sites N23, N98, N243, N581, N1015, N1020, and N1096 from Insol-aLP. N538, N542, and N1049 are not available. As previously described[74], the N-glycan categories are high-mannose type glycan series (Man9 to Man5, shown in green), afucosylated and fucosylated hybrid type glycans (Hybrid and Fhybrid, shown in yellow), and complex type glycan series based on the number of antennae and the modification of core fucose (A1 to FA4/FA3B, shown in red-purple). For more detailed information about individual digestion protocols, please refer to **Figs. A2-A4.**

doi:10.6342/NTU202301899

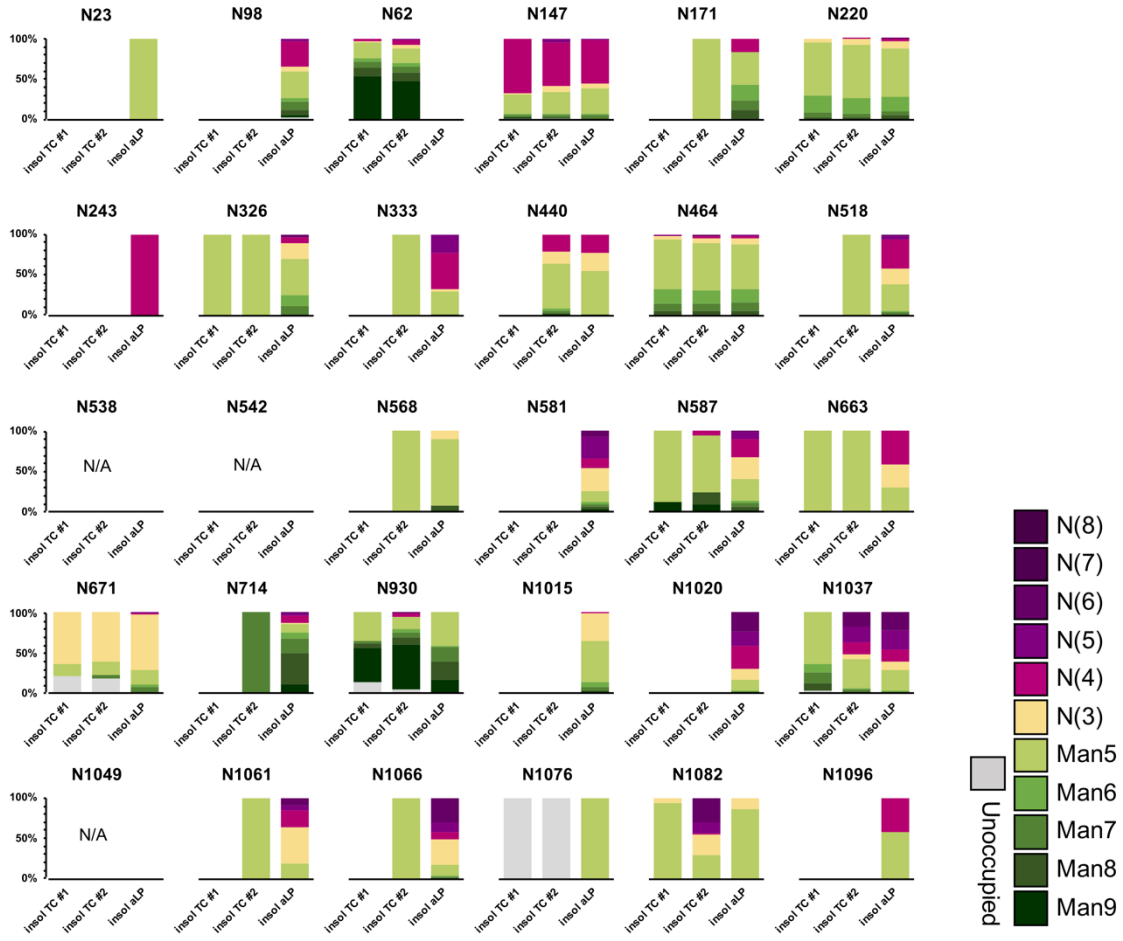


Figure 3.5-2. Overall site-specific N-linked glycosylation pattern of the HCoV-229E P100E strain S protein.

The glycosylation pattern of various digestion protocols is compared in the following sequence from left to right: in-solution digestion with trypsin plus chymotrypsin, in-solution digestion with trypsin plus chymotrypsin repeat #2, and in-solution digestion with alpha lytic protease. The bar charts depict the relative quantities of N-glycans that were identified using LC-MS/MS. On the right side, the color scheme is presented for the different N-glycan types, which corresponds to the degree of N-glycosylation, ranging from low (Man9) to high (N8). The definition of N-glycan types was previously described[48].

doi:10.6342/NTU202301899

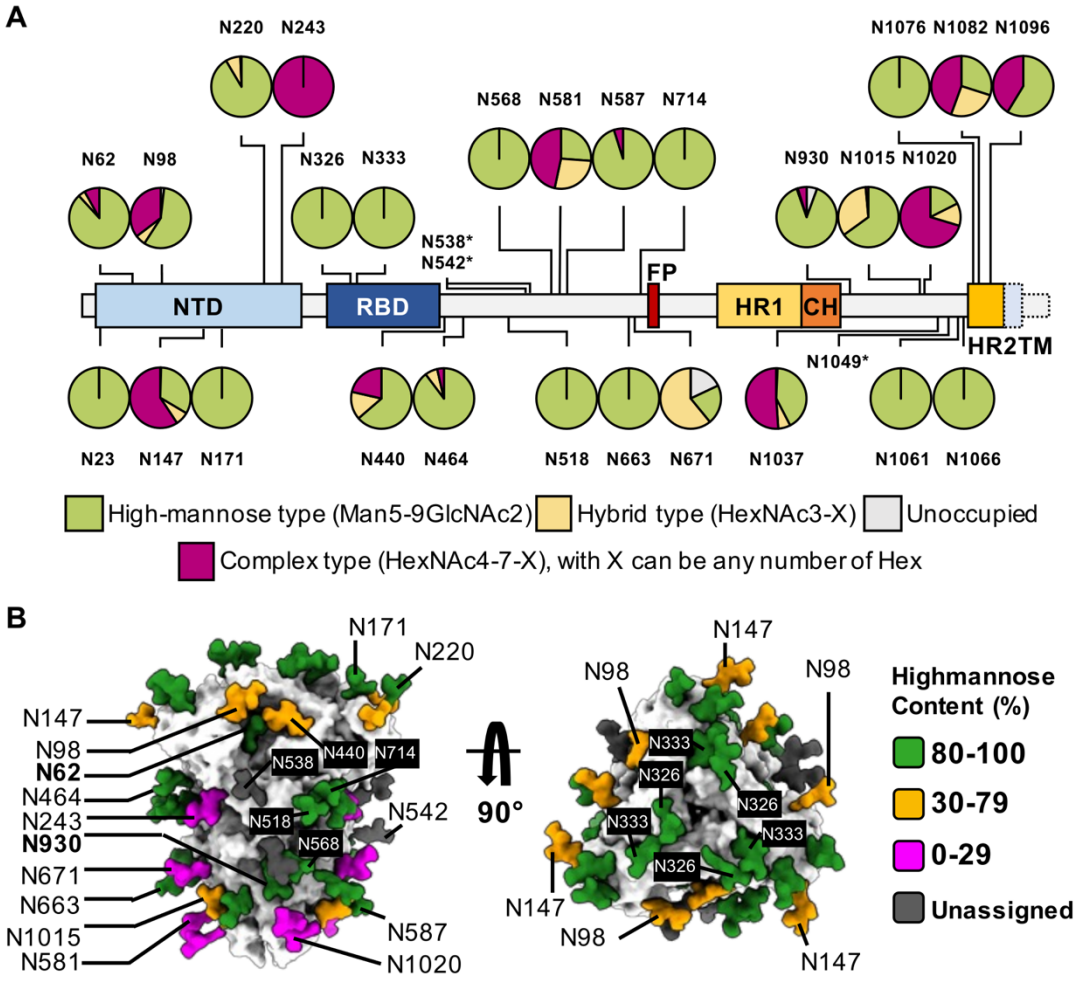


Figure 3.5-3. Structural mapping of N-linked representative glycoforms on HCoV-229E P100E strain S protein.

(A) The schematic on top shows the key domains of HCoV-229E P100E strain S protein. The pie charts present a summary of the relative quantities of N-glycans identified by LC-MS/MS, corresponding to the selected representative glycoforms with simplified categories, as defined in the legend. N-sites that were not identified by glycopeptide analysis are marked with an asterisk, and regions not in the constructs are represented with dashed lines. NTD, N-terminal domain; RBD, receptor-binding domain; FP, fusion peptide; HR1/HR2, heptad repeat 1/2; CH: central helix; TM, transmembrane domain. (B) The experimentally observed glycoforms (**Table 3.5-2**) were modeled onto the structures of HCoV-229E S (PDB ID: 7CYC) using CHARMM-GUI. The glycans are colored based on high-mannose content as defined in the legend, and the unassigned sites modeled with alternative glycan, Man5, are colored in dark grey.

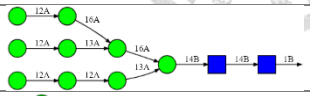
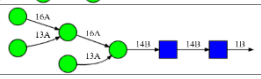
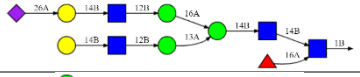
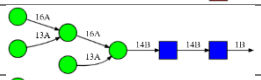
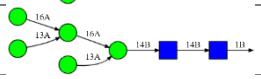
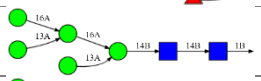
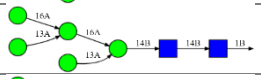
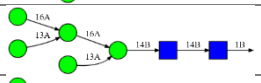
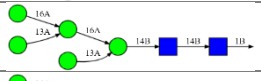
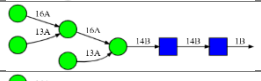
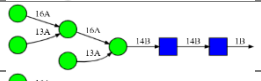
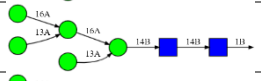
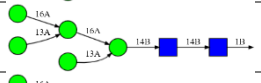
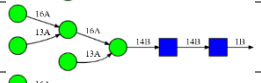
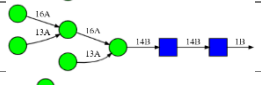
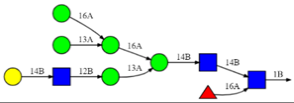
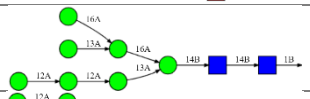
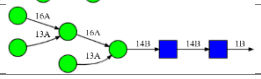
doi:10.6342/NTU202301899

Table 3.5-1. The predominant glycoforms at each N-glycosylation site of HCoV-229E P100E strain S proteins from various digestion protocols.

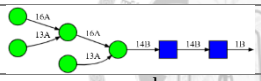
Digestion protocol		Insol-TC #1	Insol-TC #2	insol-aLP
Sequon	**N-site**	**Major Glycoform**		
NTS	**23**	N/A	N/A	HexNAc(2)Hex(5)
NTS	**62**	HexNAc(2)Hex(9)	HexNAc(2)Hex(9)	N/A
NGT	**98**	N/A	N/A	HexNAc(2)Hex(5)
NNT	**147**	HexNAc(4)Hex(5)Fuc(1) NeuAc(1)	HexNAc(4)Hex(5)Fuc(1) NeuAc(1)	HexNAc(4)Hex(5)Fuc(1) NeuAc(1)
NTT	**171**	N/A	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)
NVT	**220**	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)
NVS	**243**	N/A	N/A	HexNAc(4)Hex(5)Fuc(1)
NIT	**326**	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)
NET	**333**	N/A	HexNAc(2)Hex(5)	HexNAc(2)Hex(4)
NVT	**440**	N/A	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)
NDT	**464**	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)
NFT	**518**	N/A	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)
NGT	**538**	N/A	N/A	N/A
NCT	**542**	N/A	N/A	N/A
NVS	**568**	N/A	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)
NLS	**581**	N/A	N/A	HexNAc(2)Hex(5)
NWT	**587**	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)
NVS	**663**	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)
NLS	**671**	HexNAc(3)Hex(6)Fuc(1)	HexNAc(3)Hex(6)Fuc(1)	HexNAc(2)Hex(5)
NCT	**714**	N/A	HexNAc(2)Hex(7)	HexNAc(2)Hex(8)
NGT	**930**	HexNAc(2)Hex(9)	HexNAc(2)Hex(9)	HexNAc(2)Hex(5)
NVT	**1015**	N/A	N/A	HexNAc(2)Hex(5)
NIS	**1020**	N/A	N/A	HexNAc(2)Hex(5)
NKT	**1037**	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)
NYT	**1049**	N/A	N/A	N/A
NQT	**1061**	N/A	HexNAc(2)Hex(5)	HexNAc(3)Hex(3)Fuc(1)
NLT	**1066**	N/A	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)
NKS	**1076**	Unoccupied	Unoccupied	HexNAc(2)Hex(5)
NYT	**1082**	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)
NST	**1096**	N/A	N/A	HexNAc(2)Hex(5)
Coverage		12/30	20/30	26/30

doi:10.6342/NTU202301899

Table 3.5-2. Glycoforms of HCoV-229E P100E strain S protein selected as representatives for later GlycoSHIELD modeling.

N-site	Glycan composition	Graph Representation
23	N/A[b]	N/A[b]
62	HexNAc(2)Hex(9)	
98	HexNAc(2)Hex(5)	
147	HexNAc(4)Hex(5)Fuc(1)NeuAc(1)	
171	HexNAc(2)Hex(5)	
220	HexNAc(2)Hex(5)	
243	HexNAc(4)Hex(5)Fuc(1)	
326	HexNAc(2)Hex(5)	
333	HexNAc(2)Hex(5)	
440	HexNAc(2)Hex(5)	
464	HexNAc(2)Hex(5)	
518	HexNAc(2)Hex(5)	
538	HexNAc(2)Hex(5)[a]	
542	HexNAc(2)Hex(5)[a]	
568	HexNAc(2)Hex(5)	
581	HexNAc(2)Hex(5)	
587	HexNAc(2)Hex(5)	
663	HexNAc(2)Hex(5)	
671	HexNAc(3)Hex(6)Fuc(1)	
714	HexNAc(2)Hex(7)	
930	HexNAc(2)Hex(9)	
1015	HexNAc(2)Hex(5)	

doi:10.6342/NTU202301899

N-site	Glycan composition	Graph Representation
1020	HexNAc(2)Hex(5)	
1037	N/A[b]	N/A[b]
1049	N/A[b]	N/A[b]
1061	N/A[b]	N/A[b]
1066	N/A[b]	N/A[b]
1076	N/A[b]	N/A[b]
1082	N/A[b]	N/A[b]
1096	N/A[b]	N/A[b]

[a]. To model the structure, Man5 was used as an alternative glycoform when the glycopeptide analysis did not identify the N-site or the glycoform was unassigned by our workflow.

[b]. The glycans were not modelled since the residues were not available in PDB ID:7CYC.

doi:10.6342/NTU202301899

3.6 Site specific N-glycopeptide analysis of HCoV-229E Seattle strain S and a comparative discussion on the glycosylation profiles of the two strains

Seattle S protein contains 34 theoretical N-glycosylation sites based on the NxS/T sequon prediction. To determine site-specific glycosylation status, different digestion protocols, including Ingel-TC (protocol A described in **2.11**) and Insol-TC (protocol B described in **2.12**), were utilized for glycopeptide sample preparation. The prepared samples would then be further analyzed by LC-MS/MS as well as the processing of MS search engines and the data sorting by Python scripts described in **sections 2.13-15.**

The N-linked glycans on HCoV-229E Seattle strain S mostly belong to the high-mannose type, including sites N98, N171, N220, N438, N462, N486, N516, N566, N585, N1074, and N1094 (**Figs. 3.6-1., 3.6-2.**). The rest of the N-sites exhibit a more processed glycan composition such as the hybrid or complex type (**Fig. 3.6-2**). Notably, in S1 subunit, N147 remained as a highly processed N-site, similar to that observed on P100E strain. This suggests a less obstructed environment for enzymatic processing in the Golgi apparatus.

Since the N-linked glycosylation analysis was carried out in parallel on these two strains, we can collate the results to glean more information on the interplay between the protein structural integrity and the glycosylation profile. A comparison between **Fig. 3.5-3. B** and **Fig. 3.6-3 B** shows a generally higher degree of glycan enzymatic processing on the Seattle strain S protein. Especially in the N-sites surrounding the top portion of the S1 subunit, the originally medium percentages of the high-mannose content (colored in orange) were replaced with low values (colored in magenta). This might be the result of the less compact structure of Seattle strain S protein, which is consistent with the observation revealed by NSEM (**Fig. 3.3-4. A**). Moreover, according to the MS quantification results on sites N62 and N930 of the P100E strain S protein, the most

doi:10.6342/NTU202301899

abundant glycan type is Man9, which takes up around 50% of the total glycan composition. On the other hand, the most plentiful glycan type shifts to Man5 on the Seattle strain S, and the overall glycan profile moves towards the hybrid and complex types (**Fig. 3.6-4.**), supporting once again the hypothesis on the higher enzyme accessibility in the disrupted molecular structure of the Seattle strain S.

To further visualize the interaction between the glycan processing enzyme and the glycoprotein, class I human ER 1,2-α-mannosidase (PDB ID: 5KIJ) was docked onto the surface of HCoV-229E S (PDB ID: 7CYC) and aligned to the terminal D1, D2 or D3 arms of the Man9 moiety[75] at position N62 and N930 (**Fig. 3.6-5.**). The severe steric clashes between the mannosidase and the spike protein suggest that the Seattle strain S might either undergo a significant conformational alteration around these two sites or suffer from a certain degree of structural damage to create sufficient space for enzyme accommodation and functioning.

Here we provided a perspective on the potential relationship between the protein structure and the glycan processing. Nonetheless, whether the antigenic drift of HCoV-229E from the P100E strain to the Seattle strain directly gives rise to the aforementioned disparity in the structural stability and the glycosylation pattern warrants further analysis.

doi:10.6342/NTU202301899

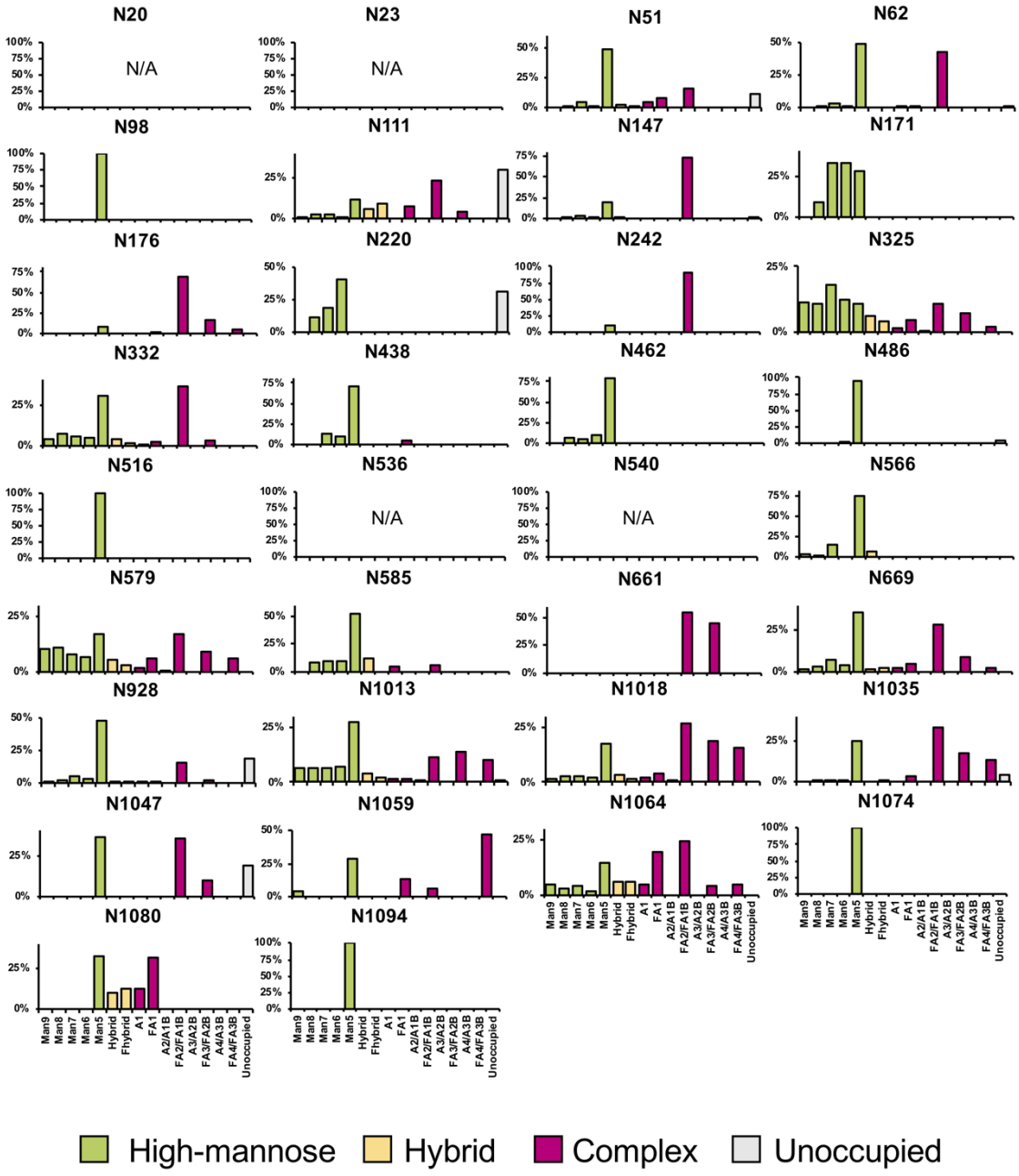


Figure 3.6-1. Quantification of the site-specific N-linked glycosylation analysis of the HCoV-229E Seattle strain S protein.

The bar charts provide an overview of the relative quantities of N-glycans identified through LC-MS/MS. The charts only display results corresponding to the selected representative glycoforms, including overall data from Insol-TC and sites N171, N176, N332, N566, N579, N585, N661, N669, N1013, N1018, N1047, N1059, N1064, N1074, and N1094 from Insol-aLP, as well as site N1080 from ingel-TC. N20, N23, N536, and N540 are not available. As previously described[74], the N-glycan categories are high-mannose type glycan series (Man9 to Man5, shown in green), afucosylated and fucosylated hybrid type glycans (Hybrid and Fhybrid, shown in yellow), and complex

doi:10.6342/NTU202301899

type glycan series based on the number of antennae and the modification of core fucose (A1 to FA4/FA3B, shown in red-purple). For more detailed information about individual digestion protocols, please refer to **Figs. A5-A7.**

doi:10.6342/NTU202301899

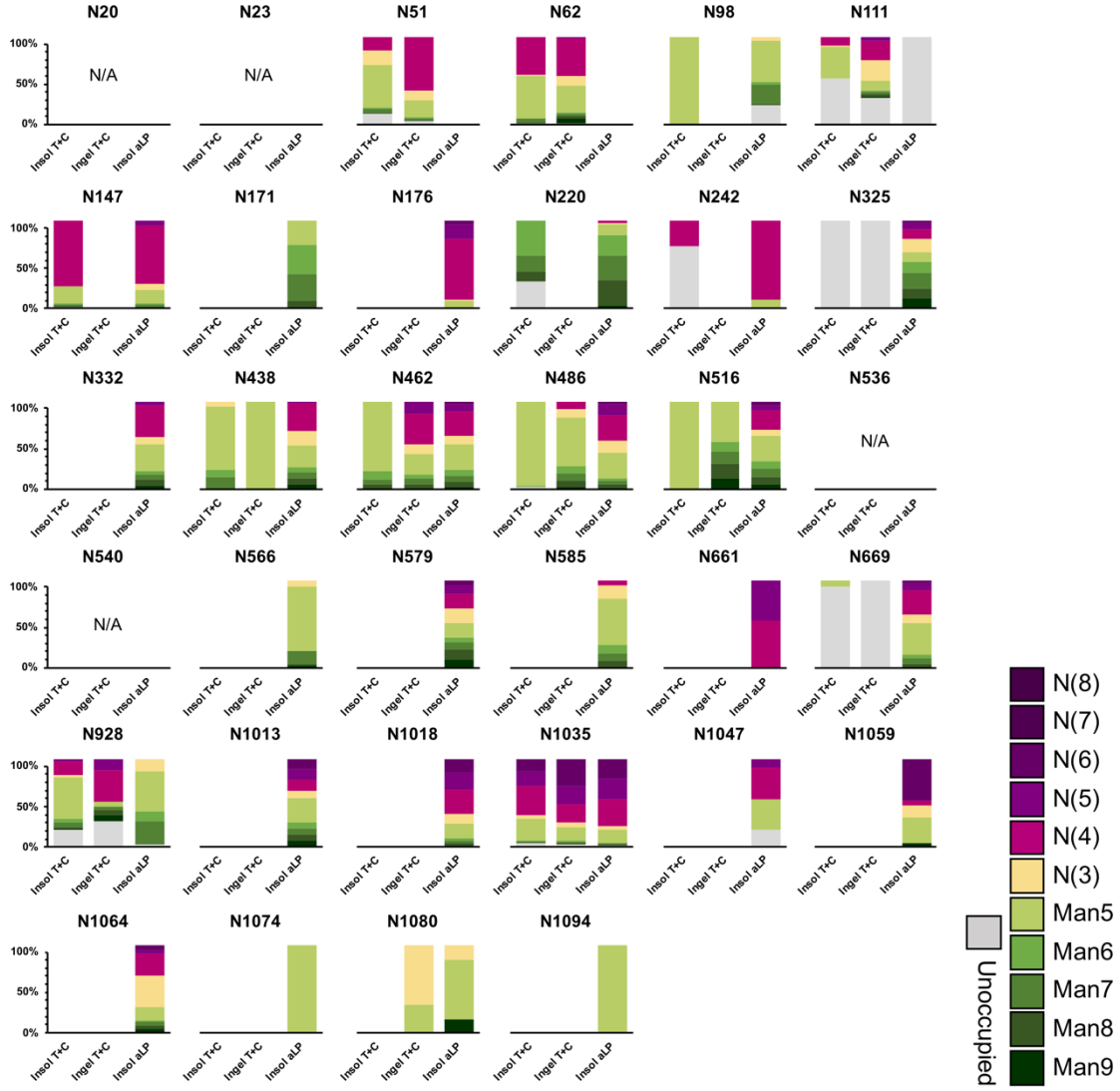


Figure 3.6-2. Overall site-specific N-linked glycosylation pattern of the HCoV-229E Seattle strain S protein.

The glycosylation pattern of various digestion protocols is compared in the following sequence from left to right: in-solution digestion with trypsin plus chymotrypsin, in-gel digestion with trypsin plus chymotrypsin, and in-solution digestion with alpha lytic protease. The bar charts depict the relative quantities of N-glycans that were identified using LC-MS/MS. On the right side, the color scheme is presented for the different N-glycan types, which corresponds to the degree of N-glycosylation, ranging from low (Man9) to high (N8). The definition of N-glycan types was previously described[48].

doi:10.6342/NTU202301899

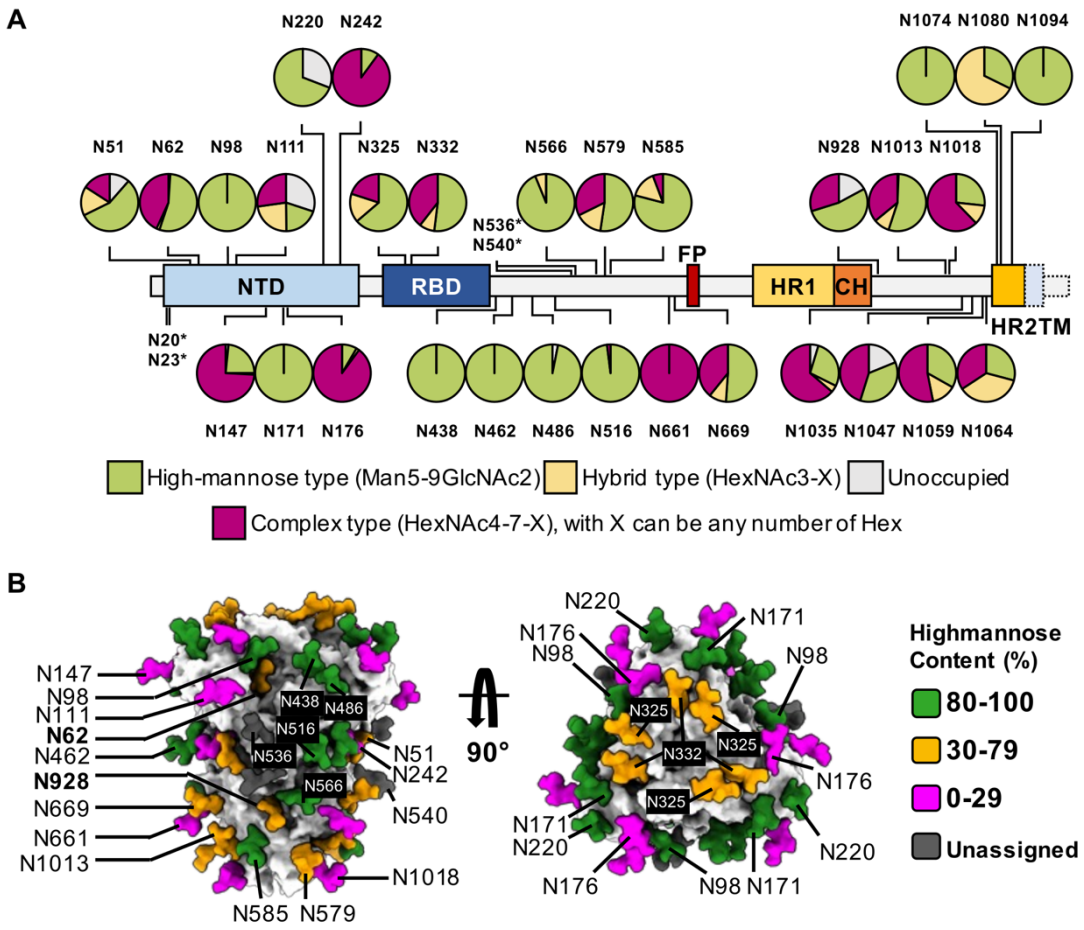


Figure 3.6-3. Structural mapping of N-linked representative glycoforms on HCoV-229E Seattle strain S protein.

(A) The schematic on top shows the key domains of HCoV-229E Seattle strain S protein. The pie charts present a summary of the relative quantities of N-glycans identified by LC-MS/MS, corresponding to the selected representative glycoforms with simplified categories, as defined in the legend. N-sites that were not identified by glycopeptide analysis are marked with an asterisk, and regions not in the constructs are represented with dashed lines. NTD, N-terminal domain; RBD, receptor-binding domain; FP, fusion peptide; HR1/HR2, heptad repeat 1/2; CH: central helix; TM, transmembrane domain. (B) The experimentally observed glycoforms (**Table 3.6-2**) were modeled onto the structures of HCoV-229E S (PDB ID: 7CYC) using CHARMM-GUI. The glycans are colored based on high-mannose content as defined in the legend, and the unassigned sites modeled with alternative glycan, Man5, are colored in dark grey.

doi:10.6342/NTU202301899

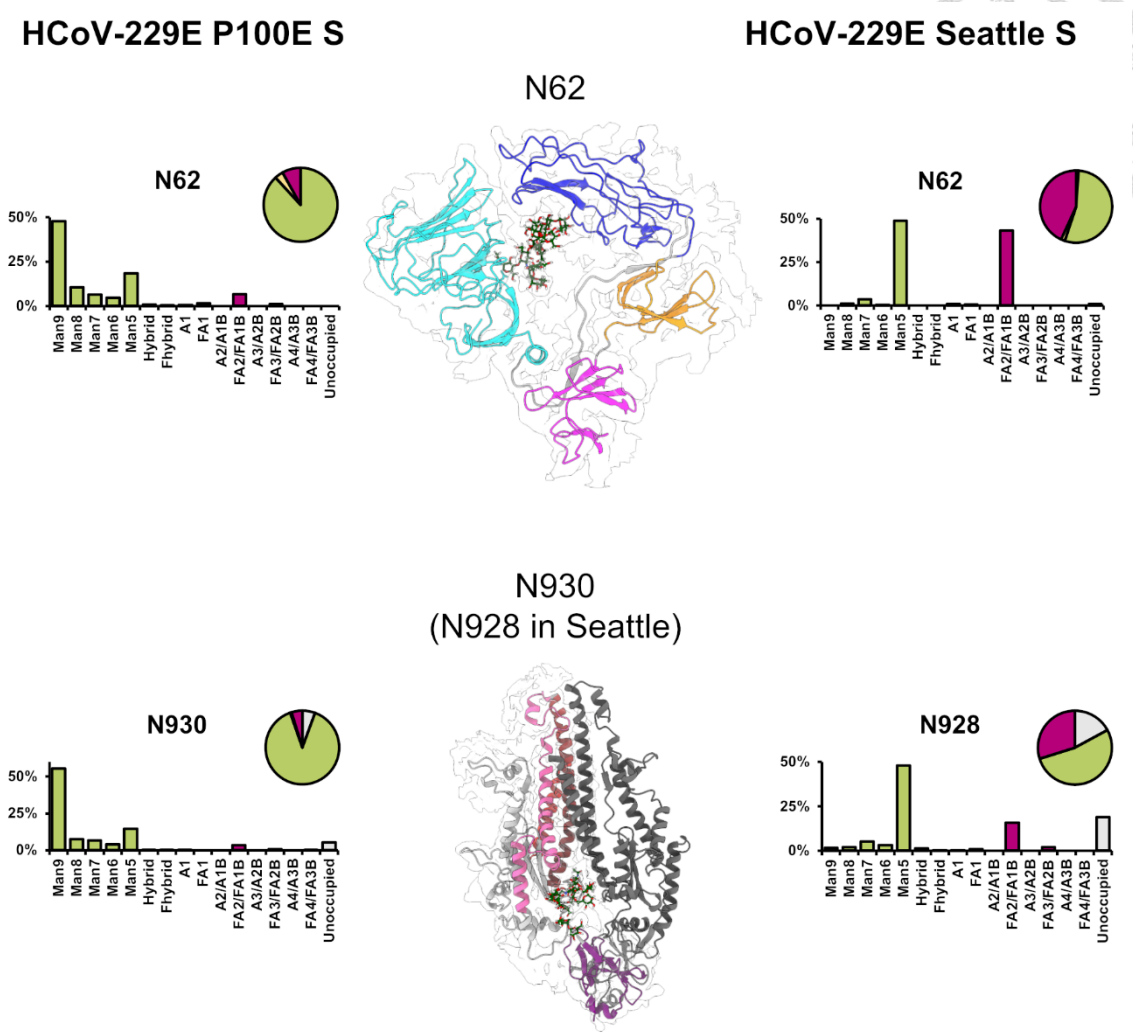


Figure 3.6-4. The stark contrast in high-mannose content of N62 and N930 (N928 in Seattle) between HCoV-229E P100E and Seattle strain S proteins.

Among all the N-sites analyzed by LC-MS/MS, N62 and N930 (N928 in Seattle strain due to the 2 deletions) presented the most significant difference in the N-linked glycosylation profile between P100E and Seattle strains. The site-specific quantitative summaries of the glycosylation analysis are shown in bar and pie charts described as in **Figs. 3.5-1**, **3.5-3**, **3.6-1**, and **3.6-3**. In the middle, the partial molecular structures surrounding N62 and N930/N928 were extracted from PDB ID: 7CYC, with Man9 glycans attached to the corresponding asparagine residues. NTD (domain A), RBD, domain C, domain D, UH, HR1, CH, and CD, are colored in cyan, blue, orange, magenta, grey, hot pink, dark red, and purple, respectively.

doi:10.6342/NTU202301899

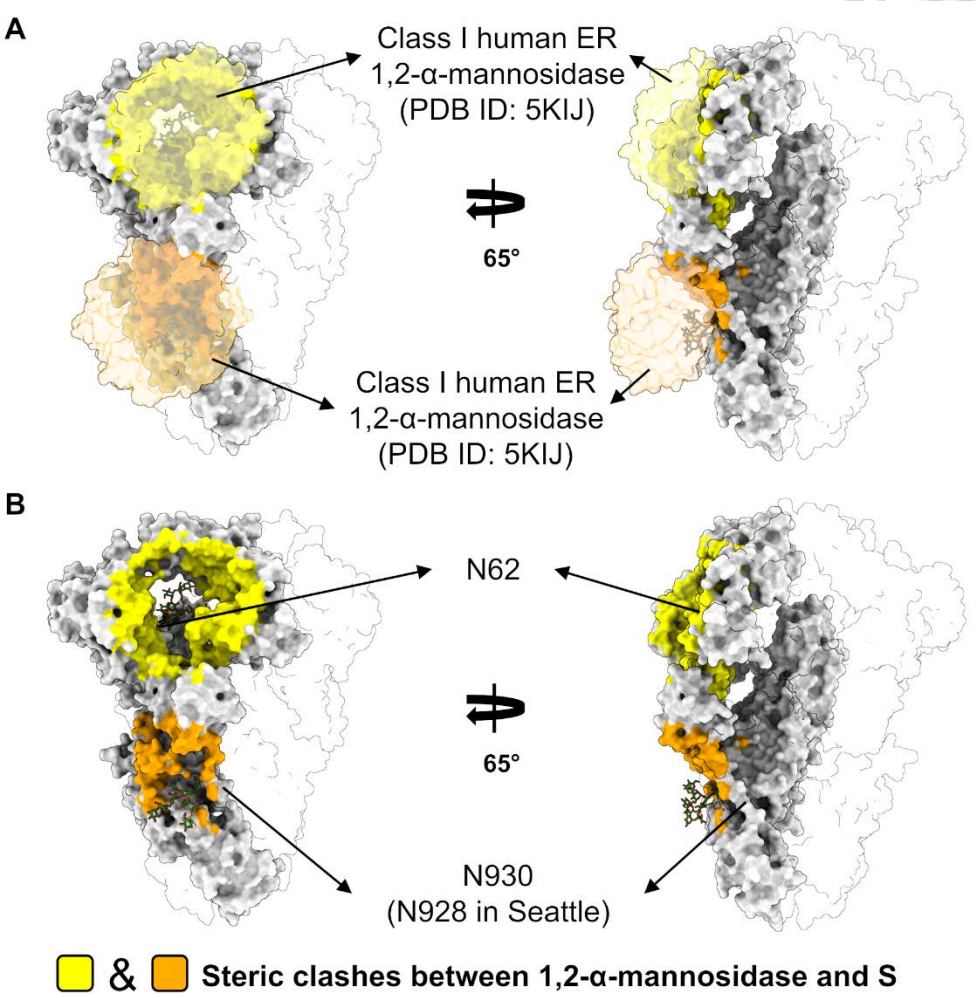


Figure 3.6-5. Steric hindrance of the intact HCoV-229E S protein prevents class I human ER 1,2-α-mannosidase from Man9 trimming.

Significant clashes were observed when docking the 1,2-α-mannosidase structure onto the surface of HCoV-229E S (PDB ID: 7CYC) and aligning it on the terminal D1, D2 or D3 arms of the Man9 moiety at position N62 and N930/N928. The colored surfaces (yellow and orange) indicate regions of the S protein that are within 2Å of the 1,2-α-mannosidase after docking the terminal mannoses on Man9 into the enzyme pocket. These clashes impede the enzyme access to its glycan substrates, and therefore, are predicted to prevent glycan processing. In (A), two 1,2-α-mannosidases were docked onto the S protein, with the molecular structure colored in semi-transparent yellow and orange. Whereas in (B), the molecular models were removed to show the steric clashes underneath.

doi:10.6342/NTU202301899

Table 3.6-1. The predominant glycoforms at each N-glycosylation site of HCoV-229E Seattle strain S proteins from various digestion protocols.

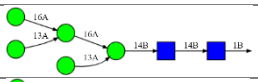
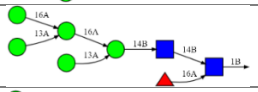
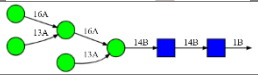
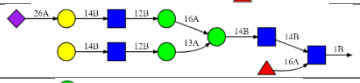
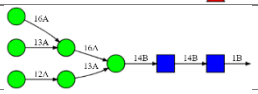
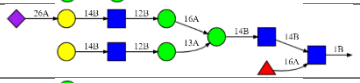
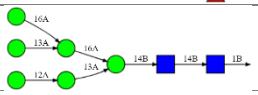
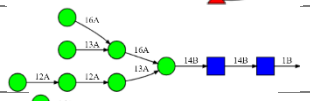
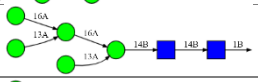
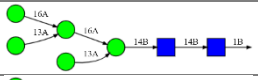
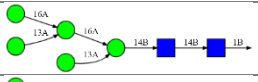
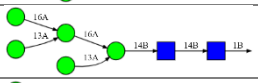
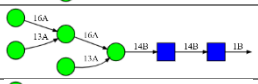
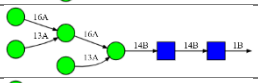
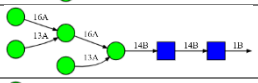
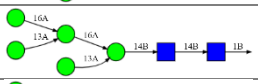
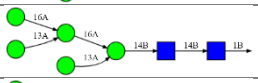
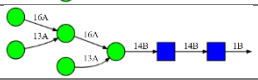
Digestion protocol		Ingel-TC	Insol-aLP	insol-TC
Sequon	N-site	Major Glycoform		
NGT	20	N/A	N/A	N/A
NTS	23	N/A	N/A	N/A
NFS	51	HexNAc(4)Hex(5)Fuc(1)	N/A	HexNAc(2)Hex(5)
NTS	62	HexNAc(4)Hex(5)Fuc(1)	N/A	HexNAc(2)Hex(5)Fuc(1)
NGT	98	N/A	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)
NAS	111	HexNAc(4)Hex(5)Fuc(1)	Unoccupied	Unoccupied
NNT	147	N/A	HexNAc(4)Hex(5)Fuc(1) NeuAc(1)	HexNAc(4)Hex(5)Fuc(1) NeuAc(1)
NTT	171	N/A	HexNAc(2)Hex(6)	N/A
NET	176	N/A	HexNAc(4)Hex(5)Fuc(1) NeuAc(1)	N/A
NVT	220	N/A	HexNAc(2)Hex(8)	HexNAc(2)Hex(6)
NVS	242	N/A	HexNAc(4)Hex(5)Fuc(1)	Unoccupied
NIT	325	Unoccupied	HexNAc(2)Hex(7)	Unoccupied
NET	332	N/A	HexNAc(2)Hex(5)	N/A
NVT	438	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)
NDT	462	HexNAc(4)Hex(5)Fuc(1)	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)
NGT	486	HexNAc(2)Hex(5)	HexNAc(4)Hex(5)Fuc(1) NeuAc(1)	HexNAc(2)Hex(5)
NFT	516	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)
NGT	536	N/A	N/A	N/A
NCT	540	N/A	N/A	N/A
NVS	566	N/A	HexNAc(2)Hex(5)	N/A
NLS	579	N/A	HexNAc(2)Hex(5)	N/A
NWT	585	N/A	HexNAc(2)Hex(5)	N/A
NVS	661	N/A	HexNAc(5)Hex(6)Fuc(1) NeuAc(1)	N/A
NLS	669	Unoccupied	HexNAc(2)Hex(5)	Unoccupied
NGT	928	HexNAc(4)Hex(5)Fuc(1) NeuAc(1)	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)
NVT	1013	N/A	HexNAc(2)Hex(5)	N/A
NIS	1018	N/A	HexNAc(4)Hex(5)Fuc(1) NeuAc(1)	N/A
NKT	1035	HexNAc(6)Hex(7)Fuc(1) NeuAc(1)	HexNAc(4)Hex(5)Fuc(1) NeuAc(1)	HexNAc(4)Hex(5)Fuc(1) NeuAc(1)
NYT	1047	N/A	HexNAc(4)Hex(5)Fuc(1)	N/A
NQT	1059	N/A	HexNAc(6)Hex(7)Fuc(1)	N/A
NLT	1064	N/A	HexNAc(3)Hex(4)Fuc(1)	N/A
NKS	1074	N/A	HexNAc(2)Hex(5)	N/A

doi:10.6342/NTU202301899

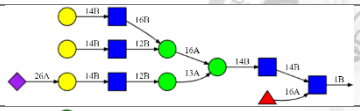
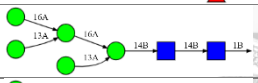
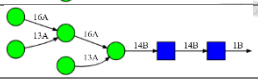
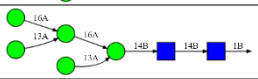
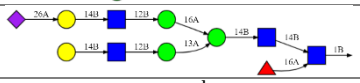
Digestion protocol		Ingel-TC	Insol-aLP	insol-TC
Sequon	**N-site**	**Major Glycoform**		
NYT	**1080**	HexNAc(3)Hex(4)Fuc(1)	HexNAc(2)Hex(5)	N/A
NST	**1094**	N/A	HexNAc(2)Hex(5)	N/A
Coverage		12/34	28/34	15/34

doi:10.6342/NTU202301899

Table 3.6-2. Glycoforms of HCoV-229E Seattle strain S protein selected as representatives for later GlycoSHIELD modeling.

N-site	Glycan composition	Graph Representation
20	N/A[b]	N/A[b]
23	N/A[b]	N/A[b]
51	HexNAc(2)Hex(5)	
62	HexNAc(2)Hex(5)Fuc(1)	
98	HexNAc(2)Hex(5)	
111	HexNAc(4)Hex(5)Fuc(1)	
147	HexNAc(4)Hex(5)Fuc(1)NeuAc(1)	
171	HexNAc(2)Hex(6)	
176	HexNAc(4)Hex(5)Fuc(1)NeuAc(1)	
220	HexNAc(2)Hex(6)	
242	HexNAc(4)Hex(5)Fuc(1)	
325	HexNAc(2)Hex(7)	
332	HexNAc(2)Hex(5)	
438	HexNAc(2)Hex(5)	
462	HexNAc(2)Hex(5)	
486	HexNAc(2)Hex(5)	
516	HexNAc(2)Hex(5)	
536	HexNAc(2)Hex(5)[a]	
540	HexNAc(2)Hex(5)[a]	
566	HexNAc(2)Hex(5)	
579	HexNAc(2)Hex(5)	
585	HexNAc(2)Hex(5)	

doi:10.6342/NTU202301899

N-site	Glycan composition	Graph Representation
661	HexNAc(5)Hex(6)Fuc(1)NeuAc(1)	
669	HexNAc(2)Hex(5)	
928	HexNAc(2)Hex(5)	
1013	HexNAc(2)Hex(5)	
1018	HexNAc(4)Hex(5)Fuc(1)NeuAc(1)	
1035	N/A[b]	N/A[b]
1047	N/A[b]	N/A[b]
1059	N/A[b]	N/A[b]
1064	N/A[b]	N/A[b]
1074	N/A[b]	N/A[b]
1080	N/A[b]	N/A[b]
1094	N/A[b]	N/A[b]

[a]. To model the structure, Man5 was used as an alternative glycoform when the glycopeptide analysis did not identify the N-site or the glycoform was unassigned by our workflow.

[b]. The glycans were not modelled since the residues were not available in PDB ID:7CYC.

doi:10.6342/NTU202301899

3.7 Site specific N-glycopeptide analysis of human aminopeptidase N

Full-Length hAPN protein contains 11 theoretical N-glycosylation sites based on the NxS/T sequon prediction. However, here we only expressed the ectodomain, i.e., from residue 66 to 967, making it 10 N-sites in total. To determine site-specific glycosylation status, different digestion protocols, including Ingel-TC (protocol A described in **2.11**) and Insol-TC (protocol B described in **2.12**), were utilized for glycopeptide sample preparation. The prepared samples would then be further analyzed by LC-MS/MS as well as the processing of MS search engines and the data sorting by Python scripts described in **sections 2.13-15.**

Based on our glycopeptide analysis and quantification workflow described in **3.4**, an in-house Python script was used to concatenate and sort the data frames from Byonic and Byos, generating a new data frame containing the quantified N-glycosylation information for the later visualization. Nevertheless, if the signal-to-noise ratio presented in Byonic output files is not ideal, Byos might have difficulty quantifying the corresponding glycopeptides[76, 77], leading to the missing Byos outputs in the concatenated data frame. This problem happened to sites N128, N319, and N573 of hAPN ectodomain, rendering these sites unable to be quantified and plotted in **Figs. 3.7-1., 3.7-2.** Thus, including sites N681 and N735, which are excluded at the early stage of the data processing due to the low scores (< 200) and high PEP2D values (> 0.001), half of the N-sites on the hAPN ectodomain cannot be quantified and assigned a dominant glycoform following our workflow (**Figs. 3.7-1., 3.7-2., 3.7-3.**).

Although the data quality is not ideal, judging from **Fig. 3.7-2.**, sites N234 and N265 exhibit Man5-dominant and A3/A2B complex-type dominant glycoforms, respectively. Interestingly, N234 and its attached glycan is in close proximity to the 229E S binding sites, as indicated in **Fig. 3.8-5.**, which could potentially contribute to the

doi:10.6342/NTU202301899

receptor association of HCoV-229E. Due to the lack of references, more experiments are warranted to support this hypothesis.

doi:10.6342/NTU202301899

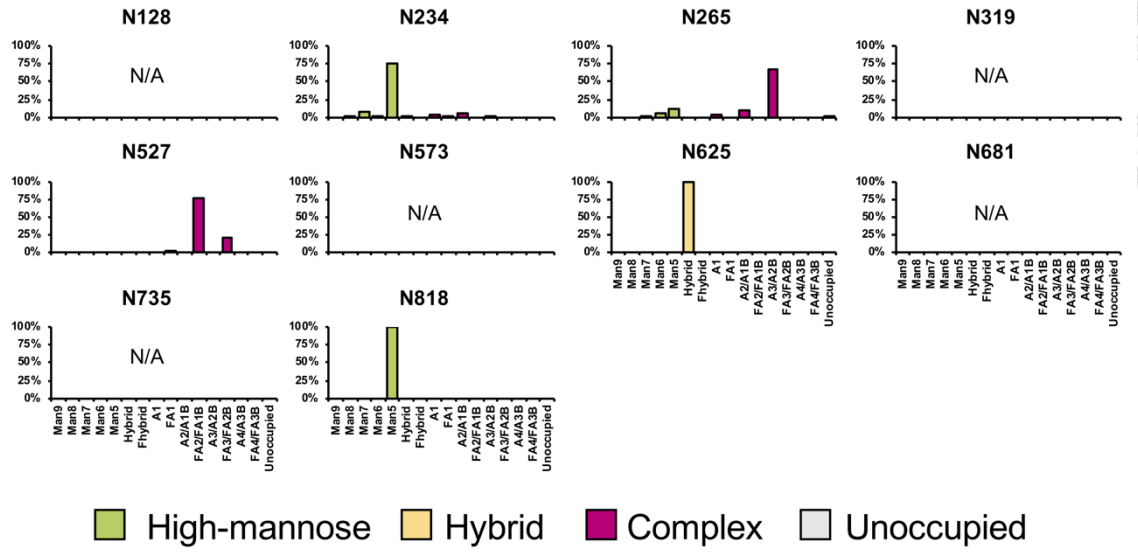


Figure 3.7-1. Quantification of the site-specific N-linked glycosylation analysis of the human aminopeptidase N (hAPN) ectodomain.

The bar charts provide an overview of the relative quantities of N-glycans identified through LC-MS/MS. The charts only display results corresponding to the selected representative glycoforms, including overall data from Insol-TC #1 and site N818 from Insol-TC #2, as well as site N527 from insol aLP. N128, N319, N573, N681, and N735 are not available. As previously described[74], the N-glycan categories are high-mannose type glycan series (Man9 to Man5, shown in green), afucosylated and fucosylated hybrid type glycans (Hybrid and Fhybrid, shown in yellow), and complex type glycan series based on the number of antennae and the modification of core fucose (A1 to FA4/FA3B, shown in red-purple). For more detailed information about individual digestion protocols, please refer to **Figs. A8-A10.**

doi:10.6342/NTU202301899

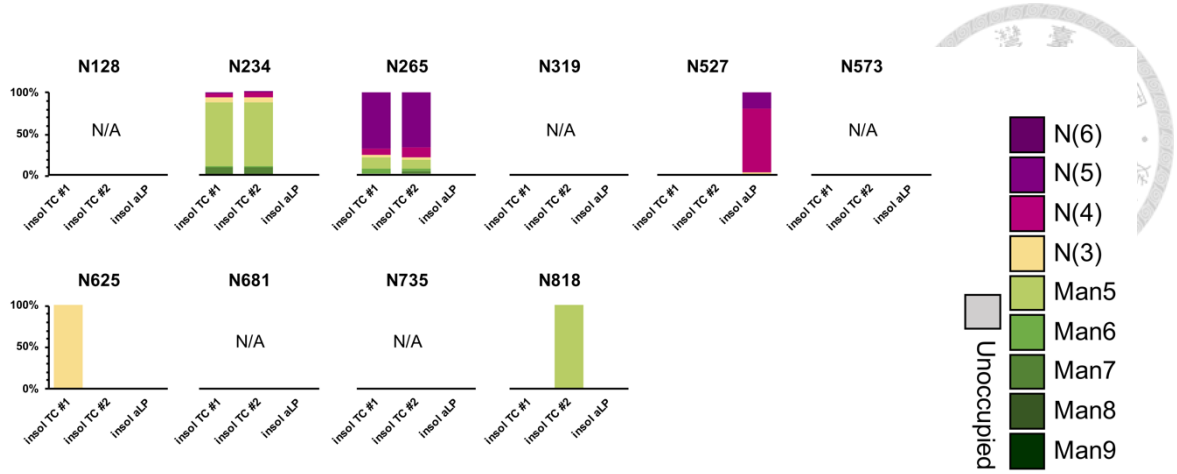


Figure 3.7-2. Overall site-specific N-linked glycosylation pattern of the human aminopeptidase N (hAPN) ectodomain.

The glycosylation pattern of various digestion protocols is compared in the following sequence from left to right: in-solution digestion with trypsin plus chymotrypsin, in-solution digestion with trypsin plus chymotrypsin repeat #2, and in-solution digestion with alpha lytic protease. The bar charts depict the relative quantities of N-glycans that were identified using LC-MS/MS. On the right side, the color scheme is presented for the different N-glycan types, which corresponds to the degree of N-glycosylation, ranging from low (Man9) to high (N8). The definition of N-glycan types was previously described[48].

doi:10.6342/NTU202301899

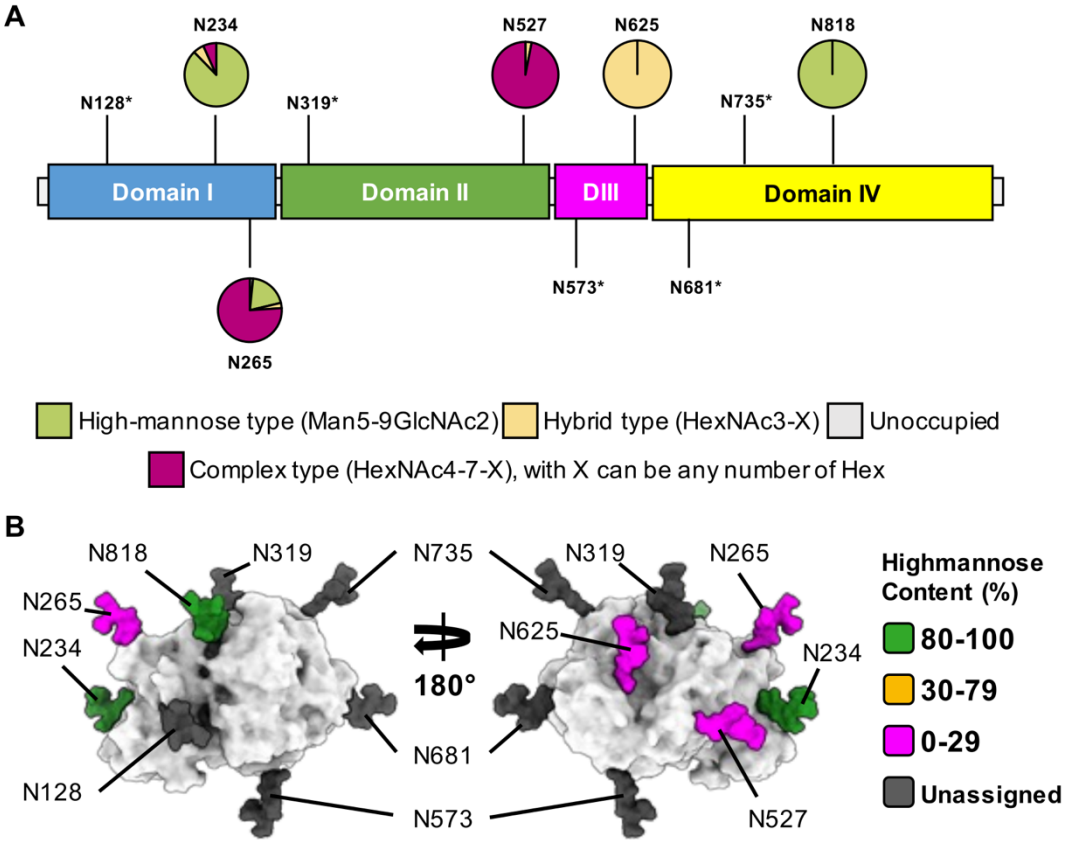


Figure 3.7-3. Structural mapping of N-linked representative glycoforms on human aminopeptidase N (hAPN) ectodomain.

(A) The schematic on top shows the key domains of human aminopeptidase N (hAPN) ectodomain. The pie charts present a summary of the relative quantities of N-glycans identified by LC-MS/MS, corresponding to the selected representative glycoforms with simplified categories, as defined in the legend. N-sites that were not identified by glycopeptide analysis are marked with an asterisk. DIII, domain III. (B) The experimentally observed glycoforms (**Table 3.7-2**) were modeled onto the structure of hAPN ectodomain (PDB ID: 6U7G) using CHARMM-GUI. The glycans are colored based on high-mannose content as defined in the legend, and the unassigned sites modeled with alternative glycan, Man5, are colored in dark grey.

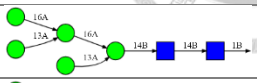
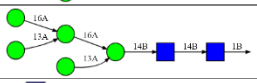
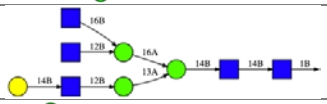
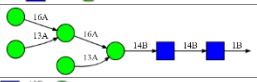
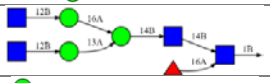
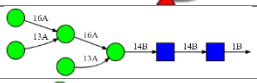
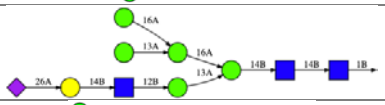
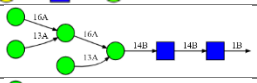
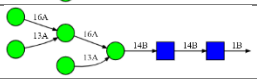
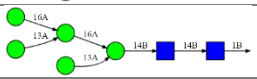
doi:10.6342/NTU202301899

Table 3.7-1. The predominant glycoforms at each N-glycosylation site of human aminopeptidase N (hAPN) ectodomain from various digestion protocols.

Digestion protocol		Insol-TC #1	Insol-TC #2	insol-aLP
Sequon	**N-site**	**Major Glycoform**		
NSS	**42**	Not expressed		
NYT	**128**	N/A	N/A	N/A
NIT	**234**	HexNAc(2)Hex(5)	HexNAc(2)Hex(5)	N/A
NVT	**265**	HexNAc(5)Hex(4)	HexNAc(5)Hex(5)	N/A
NVT	**319**	N/A	N/A	N/A
NRS	**527**	N/A	N/A	HexNAc(4)Hex(3)Fuc(1)
NVT	**573**	N/A	N/A	N/A
NVT	**625**	HexNAc(3)Hex(6)NeuAc(1)	N/A	N/A
NNT	**681**	N/A	N/A	N/A
NNT	**735**	N/A	N/A	N/A
NAT	**818**	N/A	HexNAc(2)Hex(5)	N/A
Coverage		3/10	3/10	1/10

doi:10.6342/NTU202301899

Table 3.7-2. Glycoforms of human aminopeptidase N (hAPN) ectodomain selected as representatives for later GlycoSHIELD modeling.

N-site	Glycan composition	Graph Representation
42	Not expressed	
128	HexNAc(2)Hex(5)[a]	
234	HexNAc(2)Hex(5)	
265	HexNAc(5)Hex(4)	
319	HexNAc(2)Hex(5)[a]	
527	HexNAc(4)Hex(3)Fuc(1)	
573	HexNAc(2)Hex(5)[a]	
625	HexNAc(3)Hex(6)NeuAc(1)	
681	HexNAc(2)Hex(5)[a]	
735	HexNAc(2)Hex(5)[a]	
818	HexNAc(2)Hex(5)	

[a]. To model the structure, Man5 was used as an alternative glycoform when the glycopeptide analysis did not identify the N-site or the glycoform was unassigned by our workflow.

doi:10.6342/NTU202301899

3.8 The quantification and visualization of glycan shielding effects using MD-based GlycoSHIELD

Extensive N-linked glycosylation on the spike protein surfaces has been reported on alphacoronaviruses (αCoVs), such as the human-infecting HCoV-229E[11, 44] and HCoV-NL63[69], as well as the swine-infecting PEDV[78, 79]. In the studies for HCoV-229E, the glycosylation information solely came from the observation on cryo-EM Coulomb potential map for the glycan-like protrusions. However, this approach is inherently limited by the inability of cryo-EM solving the dynamic structures, which is the case for most of the glycan moieties above the core (penta-saccharide, $Man_3GlcNAc_2$) and certain areas of the spike protein. This usually results in the "truncated" form of the cryo-EM-based glycan structures, where the map quality of the branches is often rendered substandard by their dynamics. Unless the glycans are found inside a rather enclosed environment, which restricts the glycan processing and activeness, leading to a relatively homogeneous composition and rigid molecular state in favor of cryo-EM. This exception occurred to site N62 that was buried within a cavity formed by the NTD and RBD of HCoV-229E S, where an apparent protrusion on Coulomb potential map indicated a highly ordered glycan, as described previously[44]. The glycosylation profile of this N-site was later proved to be Man9-dominant by our own N-glycopeptide analysis of P100E strain (**Figs. 3.4-1., 3.5-1.**). On the other hand, HCoV-NL63 and PEDV were investigated with a complementary method via combining mass spectrometry and cryo-EM/ET to procure a complete overview on both glycosylation and protein structure[69, 79]. Nonetheless, all studies mentioned above were incapable of demonstrating the dynamic nature of the glycans on the reported molecular models. Moreover, since the time scale of the glycan movements falls within nanoseconds while that of the receptor/antibody association ranges from microseconds to milliseconds, the glycans seen from the

doi:10.6342/NTU202301899

molecules targeting the spike protein should exhibit a shielding effect[80], rather than the static structures commonly seen on the cryo-EM models. Also, as mentioned in **section 3.5**, this glycan shield of CoVs serves to accomplish immune escape, and thus playing a crucial role in virus evolution and competence[69, 72]. Hence, it is imperative to estimate the shielding effect of N-linked glycans and quantify the surface areas that are effectively protected by these glycans on a three-dimensional (3D) model. A common way to achieve this is to leverage MD simulation, a rather matured and well-developed field, to construct an *in-silico* system to emulate the physiological conditions. Several MD studies have conducted on the fully glycosylated SARS-CoV-2 spike proteins and unveiled in detail the biologically relevant molecular mechanisms that are typically averaged out in other experimental methods[80-82]. Nevertheless, MD simulation tends to be computationally expensive, which makes it inaccessible to non-expert users. Furthermore, due to the time scale of the most biological processes, extensive simulation duration is often needed to observe the target phenomena, let alone the required repetitions to validate the reliability of the simulation. Given all these technical challenges, it is imaginable that a pipeline capable of striking a balance between computing power and the conventional MD standards would be a game changer in the field of structural biology. In this study, we integrated several in-house Python scripts into an open-source modelling tool called GlycoSHIELD[83] to semi-automatically quantify and visualize the shielding effect of the N-linked glycans on protein surfaces of S and hAPN ectodomains (**Fig. 3.8-1.**). The detailed definitions and related equations are described in **section 2.17**.

As previously elaborated in **section 3.6**, there exists a disparity in the overall N-linked glycosylation profile between the P100E and Seattle strains. Here we employed GlycoSHIELD to again address this topic to bring in the three-dimensional perspective for depicting the interplay between the surface glycans and the protein structure. **Figs.**

doi:10.6342/NTU202301899

3.8-2. and **3.8-3.** respectively show the LC-MS/MS-based densely glycosylated model and quantified glycan shielding effect of the two strains resulted from our Python-facilitated GlycoSHIELD pipeline. The variation analysis described by Eguia et al. [30] of the 96 HCoV-229E S sequences mentioned in **section 3.1** was also incorporated to compare with the $SASA_{rel}$ distribution of the old strain, P100E. Based on the results, several less shielded protein areas correspond to the regions of higher sequence variations (indicated by darker shades of red), especially in the NTD and the three receptor binding loops of RBD (**Fig. 3.8-2.**). This relationship is made further evident in the plot of per-residue quantification of the shielding effect (**Fig. 3.8-4.**) and is in line with the discussion of **section 3.2** and **Fig. 3.2-2. A**. Additionally, since the glycan canopy is known to shield the protein from the selection pressure of the host immune system, and that the receptor binding sites should be free of the steric hindrance from the glycans, the results here proved that GlycoSHIELD is effective in capturing the influence of glycan dynamics while demanding a computing resource that is sustainable on a desktop machine.

As for hAPN, the glycan shielding effect was also analyzed via the same procedure and the results are shown in **Fig. 3.8-5.,** of which the per-residue summary was plotted in **Fig. 3.8-6.** In **Fig. 3.8-5.**, the 229E S binding region[84] was further delineated with red curves to mark the $SASA_{rel}$ within and the glycan ensembles in the surroundings. Notwithstanding the $SASA_{rel}$ of the binding sites is not markedly low in glycan shielding, windows of less-shielded residues can still be observed in **Fig. 3.8-6.** (annotated by two red downward arrows), which correspond to the interacting regions between the hAPN H-site[84] and the binding loops 1 to 3 of HCoV-229E RBD. It is worth noting that the $SASA_{tot}$ of the 229E S binding sites on hAPN surface is significantly higher compared to that of the hAPN binding sites on both P100E and Seattle strains (**Fig. 3.8-4.**). Due to the lack of the experimental evidence on the interactions between hAPN surface glycans and

doi:10.6342/NTU202301899

the binding residues, whether the glycans contributing to the shielding effect facilitate the molecular association warrants further research. This also brings about the topic on the limitations of our GlycoSHIELD analysis discussed in **Chapter 4**.

doi:10.6342/NTU202301899

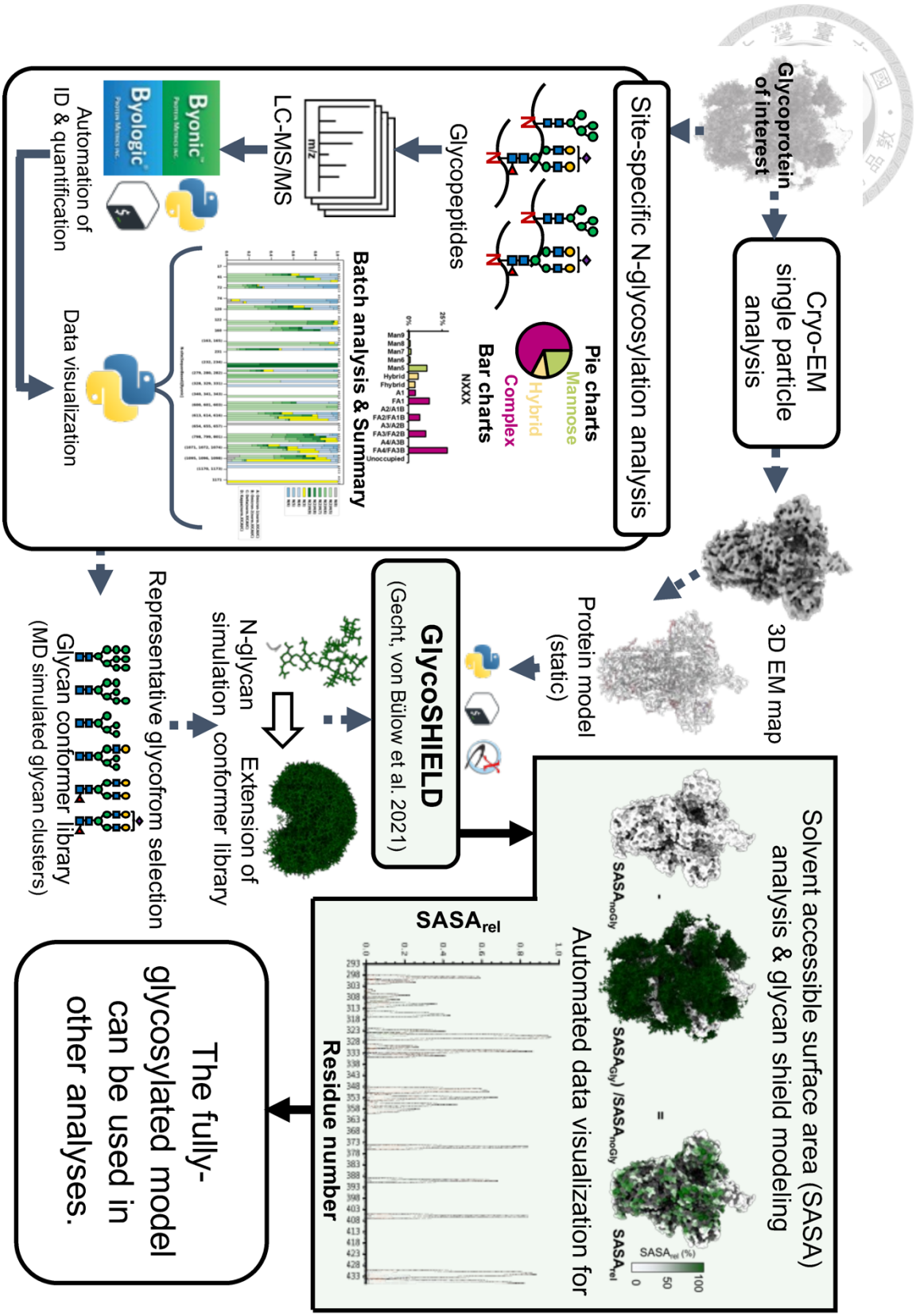


Figure 3.8-1. The process of generating realistic glycan shield using GlycoSHIELD.

doi:10.6342/NTU202301899

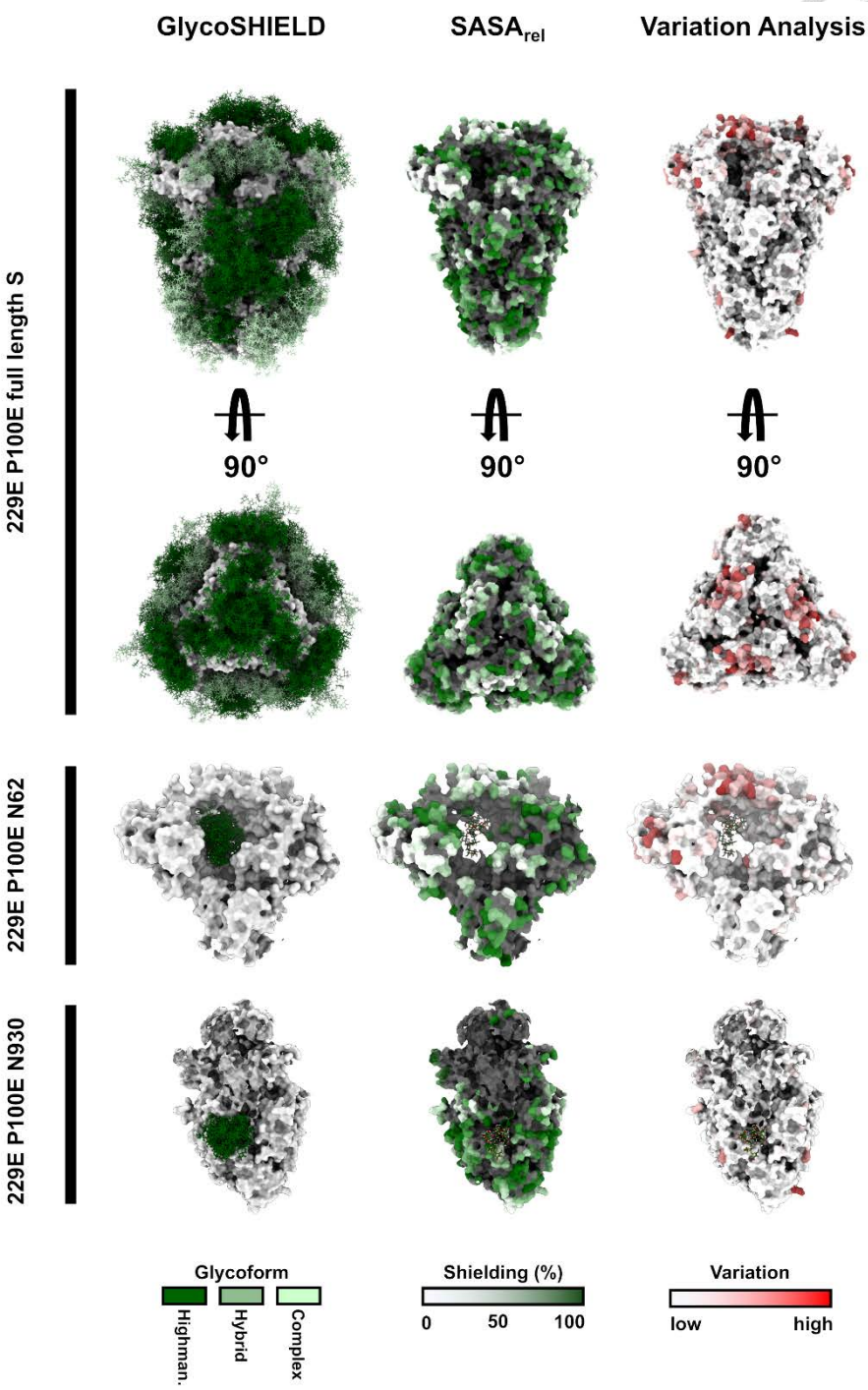


Figure 3.8-2. The glycan shielding effect of HCoV-229E P100E strain S protein.
The strain name and the molecular range of the model are as shown on the left side. The types of the analyses are as listed on the top panel. Left-most column: The experimentally observed representative glycoforms reconstructed on the structure of HCoV-229E S protein (PDB ID: 7CYC) by GlycoSHIELD, for which about 20 glycan conformers are displayed on each N-site. The experimentally observed glycans are colored in forest, green, and light green, to respectively show the high-mannose, hybrid, and complex type glycans. Middle column: The glycan shielding effect of S protein was analyzed by

doi:10.6342/NTU202301899

GlycoSHIELD, and the 3D-heatmap of the normalized relative solvent accessible surface area ($SASA_{rel}$) were calculated under the probe size of 0.75 nm, representing the surface areas that are prone to receptor binding and antibody recognition. The darker the shade of green, the higher the shielding. The un-probed surfaces are colored in dark grey. Right-most column: The results from the variation analysis as previously described[30] were mapped onto the surface of PDB ID: 7CYC to show the distribution of the highly mutated residues. The darker the shade of red, the higher the variation. The overall pipeline for GlycoSHIELD was described previously[21], and the same color scheme and data processing procedure were applied to N62 and N930 analyses.

doi:10.6342/NTU202301899

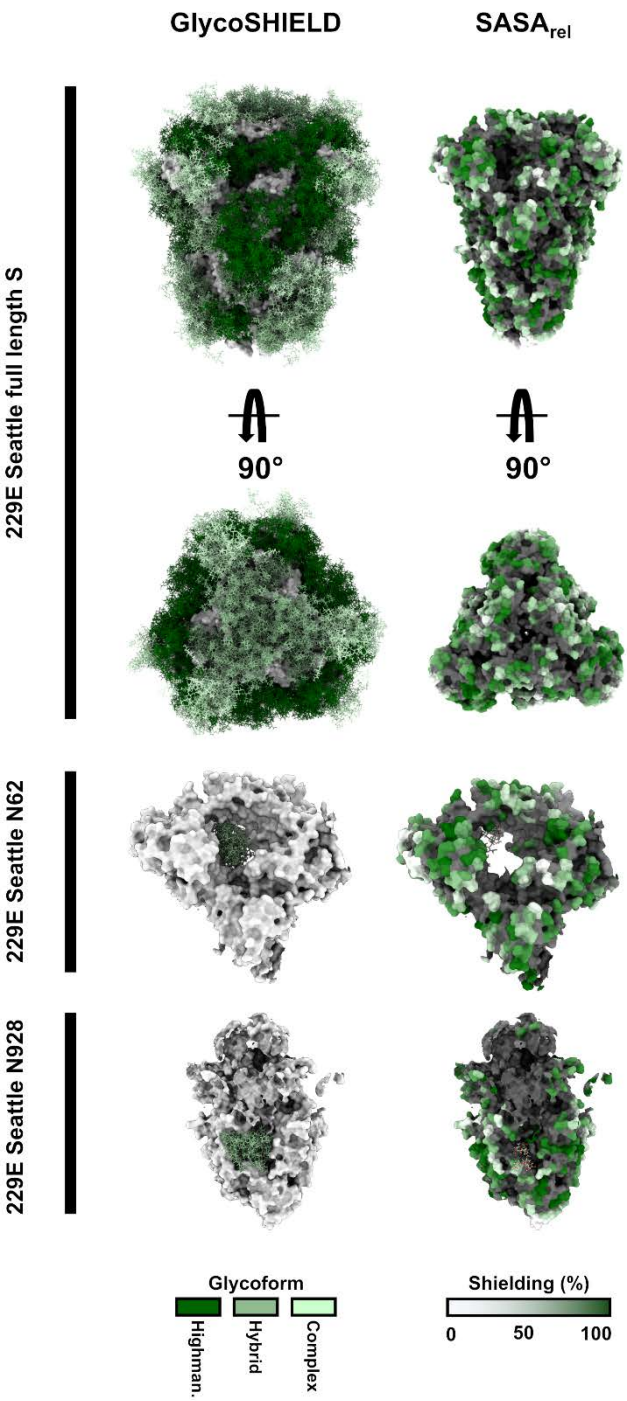


Figure 3.8-3. The glycan shielding effect of HCoV-229E Seattle strain S protein.
The strain name and the molecular range of the model are as shown on the left side. The types of the analyses are as listed on the top panel. Left column: The experimentally observed representative glycoforms reconstructed on the structure of HCoV-229E S protein (PDB ID: 7CYC) by GlycoSHIELD, for which about 20 glycan conformers are displayed on each N-site. The experimentally observed glycans are colored in forest, green, and light green, to respectively show the high-mannose, hybrid, and complex type glycans. Right column: The glycan shielding effect of S protein was analyzed by

doi:10.6342/NTU202301899

GlycoSHIELD, and the 3D-heatmap of the normalized relative solvent accessible surface area ($SASA_{rel}$) were calculated under the probe size of 0.75 nm, representing the surface areas that are prone to receptor binding and antibody recognition. The darker the shade of green, the higher the shielding. The un-probed surfaces are colored in dark grey. The overall pipeline for GlycoSHIELD was described previously[21], and the same color scheme and data processing procedure were applied to N62 and N928 analyses.

doi:10.6342/NTU202301899

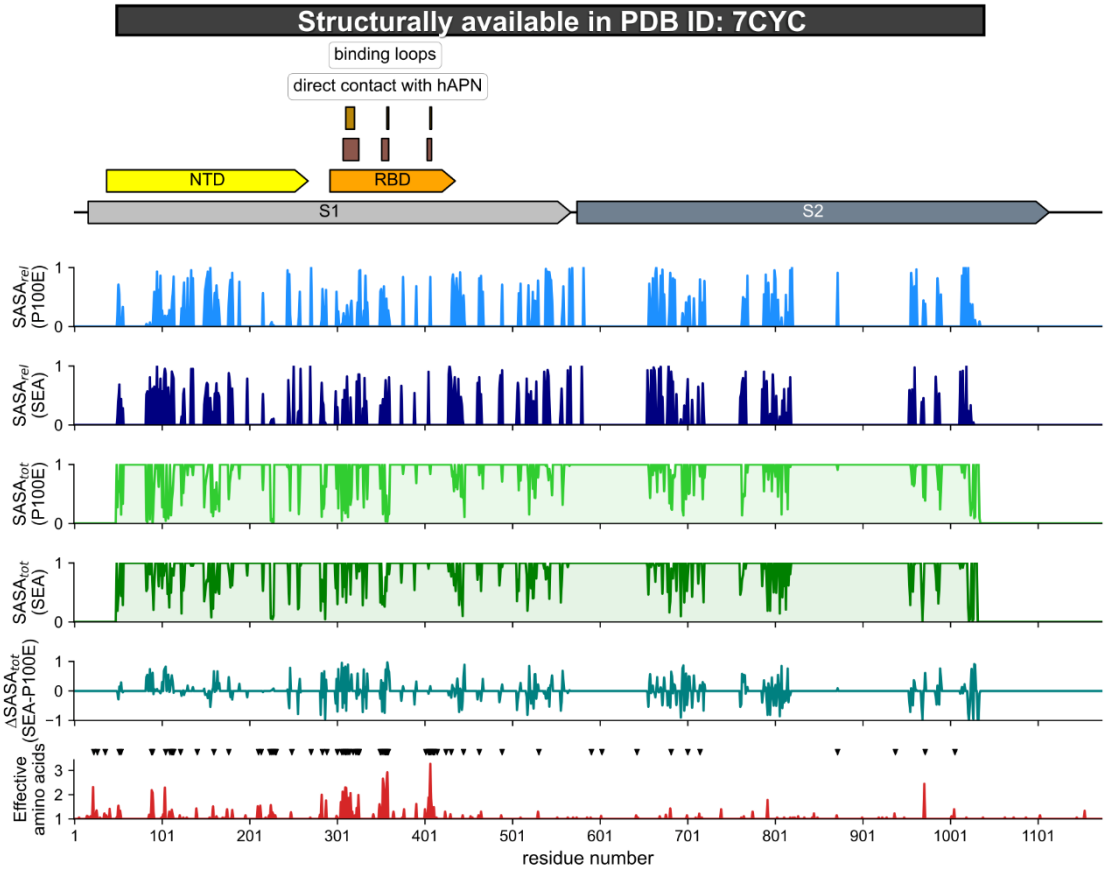


Figure 3.8-4. The summary of GlycoSHIELD analysis on P100E and Seattle strains. The schematic of the HCoV-229E S protein is shown at the top. It illustrates the N-terminal domain (NTD), also known as domain A, and the RBD, also known as domain B, within the S1 subunit. The three loops in the RBD that bind to hAPN are indicated by rectangles colored in #8C564B, while the residues directly contacting hAPN within the binding loops were further indicated by rectangles colored in #B8860B. Below the schematic is an array of 6 subplots showing relative SASA of P100E and Seattle strains, total SASA of P100E and Seattle strains, the difference in the total SASA of both strains, and the sequence variability across the alignment of HCoV-229E S proteins in **Fig. 3.1-1**. The level of variability at a site is measured by the effective number of amino acids present[85], where a value of one suggests complete conservation, while higher values indicate greater sequence variability. In the last subplot, the 72 point mutations of the Seattle strain are annotated with black downward arrows.

doi:10.6342/NTU202301899

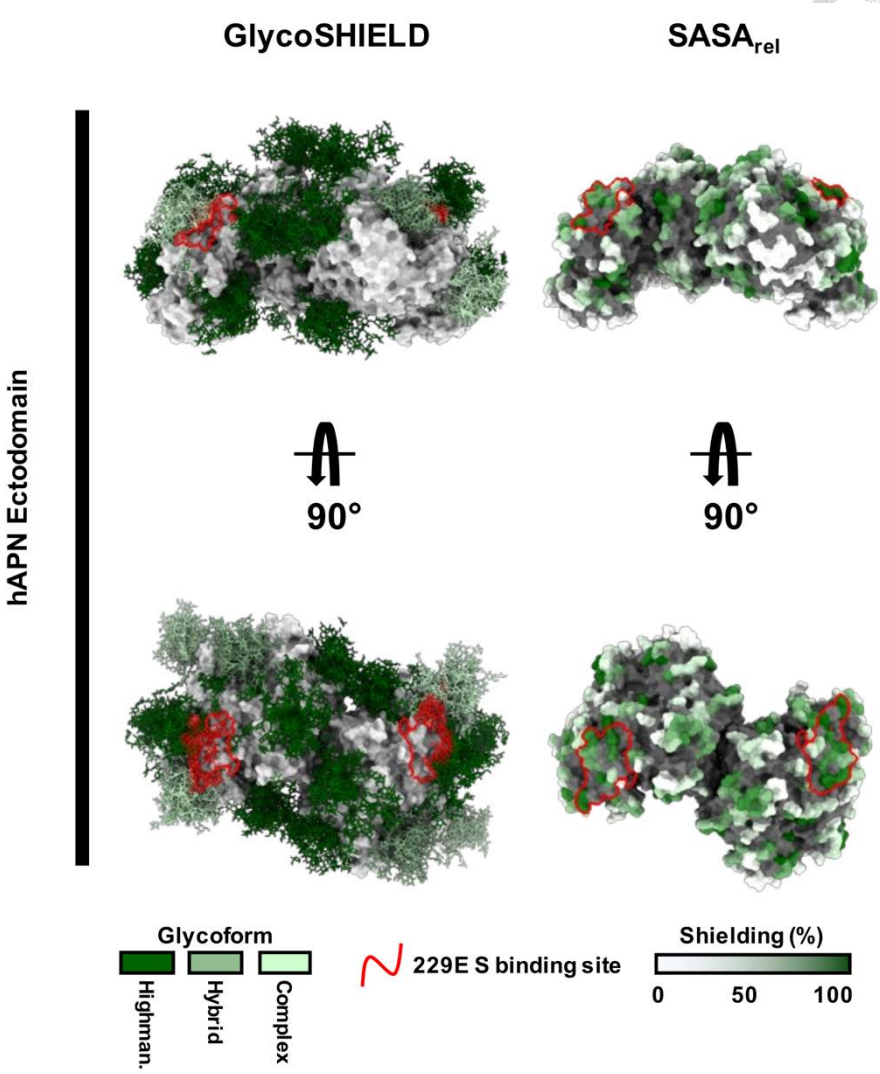


Figure 3.8-5. The glycan shielding effect of hAPN ectodomain.

The molecular range of the model is as shown on the left side. The types of the analyses are listed on the top panel. Left column: The experimentally observed representative glycoforms reconstructed on the structure of hAPN ectodomain (PDB ID: 6U7G) by GlycoSHIELD, for which about 20 glycan conformers are displayed on each N-site. The experimentally observed glycans are colored in forest, green, and light green, to respectively show the high-mannose, hybrid, and complex type glycans. Right column: The glycan shielding effect of hAPN protein was analyzed by GlycoSHIELD, and the 3D-heatmap of the normalized relative solvent accessible surface area ($SASA_{rel}$) were calculated under the probe size of 0.75 nm, representing the surface areas that are prone to 229E S binding and other molecular interactions. The darker the shade of green, the higher the shielding. The un-probed surfaces are colored in dark grey. The reported 229E S binding sites[84] are delineated with red curves on the protein surfaces. The overall pipeline for GlycoSHIELD was described previously[21].

doi:10.6342/NTU202301899

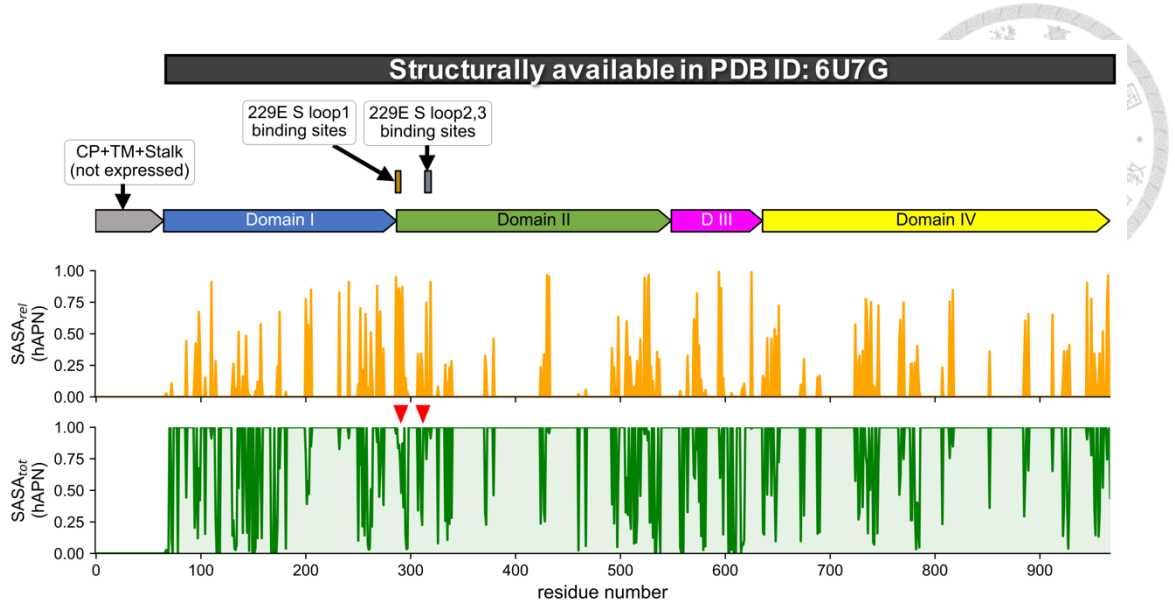


Figure 3.8-6. The summary of GlycoSHIELD analysis on hAPN ectodomain.
The schematic of hAPN protein domains is shown at the top. It includes the cytoplasmic domain (CP), transmembrane domain (TM), stalk region, and domains I-IV. Note that the CP, TM, and stalk regions were not expressed, which leads to the data absence in the plots. The 229E S binding sites on hAPN (H-site) [84] are indicated by two rectangles above the domain schematic. Below the schematic are two subplots showing relative SASA and total SASA of hAPN ectodomain. The corresponding residue range of 229E S binding sites in the total SASA plot is further annotated by two red downward arrows, revealing less-shielded windows for molecular interaction.

doi:10.6342/NTU202301899

3.9 Cryo-EM structure of HCoV-229E P100E S in complex with hAPN

As mentioned in **Chapter 1**, all the currently available cryo-EM structures of HCoV-229E S are in RBD-all-down state, which creates a significant steric hindrance that prevents the association of its receptor, hAPN. To unveil the unknown molecular mechanism of HCoV-229E host recognition, we resorted to cryo-EM to shed light on the formation of HCoV-229E S-hAPN complex. According to the predictions and experiments in **section 3.2** and **3.3**, the overall stability and binding affinity to hAPN of the P100E strain are superior to those of the Seattle strain. Therefore, we chose the former to proceed to cryo-EM analysis. Of note, since our goal is to examine the structure of HCoV-229E S-hAPN complex, we had to firstly ensure that the prepared sample concentration was suitable for the complex formation. Based on the K_D derived from the BLI assay (**Fig. 3.3-6. A**), the theoretical concentration of HCoV-229E P100E strain S proteins would be 24 nM when half of the S binding sites on hAPN are occupied in the system equilibrium. However, according to **Fig. 3.3-5.**, hAPN tends to associate with each other, forming numerous aggregation-like clumps shown in the NSEM micrographs. This might compromise the working concentration contributing to the complex formation. Besides, considering the previously reported cryo-EM protocols for spike proteins, the most commonly used sample concentrations were usually above 0.5 mg/mL[11, 39, 79], and the high concentration can facilitate the particle identification during the initial screening of the cryo-EM squares. Therefore, a higher concentration of 3.5 mg/mL of the sample mixture with a 1 to 2 molar ratio between P100E S protein and hAPN ectodomain was employed to maximize the fully occupied complex formation.

In practice, the SEC-purified P100E S proteins (**Fig. 3.3-2.**, top row) were mixed with the SEC-purified hAPN ectodomains (**Fig. 3.3-2.**, bottom row), followed by an incubation at 4ºC for 1 hour to ensure an equilibrated system. Next, the mixture was

doi:10.6342/NTU202301899

injected into Superose 6 increase 10/300 GL column to identify the putative complex fractions (**Fig. 3.9-1.**, indicated by light grey background), which were collected and further concentrated to 3.5 mg/mL for cryo-EM grid preparation. The sample preparation and the data collection/processing procedure for cryo-EM are described in detail in **sections 2.8** and **2.9**. The derived Coulomb potential map of the complex is shown in **Fig. 3.9-2.**, in which three selected levels of the contour map are exhibited to illustrate a binding mode involving two RBD-up P100E spike proteins and a hAPN homodimer (**Figs. A15, A16, Table 3.9-1.**). This RBD-up conformation is never-before-seen among the previously reported human αCoVs[11, 44, 69], which provides a novel molecular basis for the host recognition of HCoV-229E.

To compare with the currently available molecular models of both HCoV-229E S and hAPN ectodomain, the overall scale of the corresponding partial structures of our cryo-EM complex were measured with the built-in tool in ChimeraX[38]. The measurements are as described in **Fig. 3.9-3. B** and **C**. Note that the vertical length of our HCoV-229E S structure is shorter in comparison with that of PDB ID: 6U7H[44] due to the inferior map quality around the connector domain (CD). Also, the down-state RBDs of P100E S proteins are poorly resolved, where the map is too fragmented to build reliable and connected protein backbone. We hypothesized that the presence of abundant hAPN ectodomains induce the conformational change in some of the down-state RBDs, leading to a mixture of 1-up, 2-up, and 3-up modes, which has been found in SARS-CoV-2 spike proteins[86-88]. This potential heterogeneity in the conformation could average out the comparatively dominant signals from the 1-up group, worsening the local resolution of the pertaining areas. Although the down-state RBDs were not qualified for model building, the single erected RBD was resolved via combining the NU-refinement of the full complex and the focused refinement of the RBD region (**Fig. A15**). An angular

doi:10.6342/NTU202301899

discrepancy of 60 degrees was found between the downward and upward RBDs measured by the built-in tool in ChimeraX[38] (**Figs. 3.9-3. B** and **3.9-4. A**). To procure a comprehensive view on the flexibility of the spike protein RBD, a systematic analysis on RBD-up angles of all currently available SARS-CoV-2 S entries on Protein Data Bank was carried out with in-house Python scripts (**section 2.19**, **Figs. A18., A19., A20.**). The results are summarized in **Fig. 3.9-5.**, where the RBD-up angle of P100E S induced by hAPN association is similar to those of ACE2-induced movements of SARS-CoV-2 RBDs. As for hAPN, our structure was aligned with PDB ID: 6U7G[44] by domain IV (**Fig. 3.9-3. A**), where a 12-degree angular disparity was observed between the centers of mass (COMs) of domain I (**Fig. 3.9-3. A**) regions of our cryo-EM structure and that of the PDB model. It has been reported that hAPN can adopt either open (inactive) or closed (active) conformation to regulate its enzymatic activity[16], and that HCoV-229E S proteins are able to associate with either state unlike the interaction between PRCoV S and pAPN[84]. This finding is in line with our proposed model in which hAPN takes on a structure akin to the open form described by Wong, et al[16]. Moreover, this slight difference in the bending angle of the domain IV give rise to a distance alteration between the N-terminal residues of hAPN ectodomain illustrated in **Fig. 3.9-4. C**, which has been linked to signal transduction within the cytosol[15, 16]. The aforementioned structural divergence of hAPN might stem from the distinct experimental methods in that X-ray crystallography tend to yield a more compact structure from the highly ordered crystal formation, while the native-like environment of cryo-EM renders the molecule flexible and expanded. Whether this hypothesis directly engender the above observation warrants further experimental evidence.

Another goal of cryo-EM is to investigate the surface glycosylation of HCoV-229E S-hAPN. The N-glycosylation profiles of both P100E S and hAPN ectodomain were

doi:10.6342/NTU202301899

already elaborated in **sections 3.5** and **3.7**, here we compare the site-specific glycans detected by the two experimental methods, MS and cryo-EM (**Table 3.9-2.**). Six (N62, N98, N243, N440, N518, and N671) out of 30 N-sites of P100E S and two (N265 and N625) out of 10 N-sites of hAPN ectodomain were observed by both methods. Due to the above-mentioned compromised local resolution of the complex, the cryo-EM based N-glycans built on the final molecular model (map contour = 18 sigma) all lack the branching above the common core (penta-saccharide, $Man_3GlcNAc_2$). Intriguingly, the N-glycan attaching to site N62 of P100E S, which boasts a cryo-EM map covering the glycan moieties beyond the tri-mannose core as previously described[44], remains to be the most well-resolved one, albeit the mediocre map quality of the complex. For N62, the first three saccharides ($ManGlcNAc_2$) of the common core were constructed and refined based on the methods defined in **section 2.10**, while for the rest of the N-sites, only the first two or fewer saccharides could be built under a map contour level of 18 sigma. Since 6U7H is in RBD-all-down state[44], which encloses the N62 glycan, we hypothesized that the more fragmented cryo-EM map of our N62 glycan is owing to the reduced restriction of the RBD-up conformation. Besides, the poor data quality of hAPN MS results as mentioned in **section 3.7** could diminish the agreement between the N-glycan observations of the two experimental approaches.

After establishing this workflow of combining site-specific N-glycosylation analysis and cryo-EM for a more comprehensive understanding of the glycoproteins, we can utilize it to investigate the working mechanism of the known antibodies, as well as to predict the potential neutralizing sites. An epitope mapping of HCoV-229E S was published by Xiang, et al. [31], in which the antibodies C04 and S11 were found to target the distal end of 229E RBD; antibody D12 targeted the NTD of 229E S, and F12 was the first S2-directed neutralizing antibody against human αCoVs. The corresponding

doi:10.6342/NTU202301899

epitopes are mapped onto the $SASA_{rel}$ model of the 229E S-hAPN complex generated by GlycoSHIELD[83] (**Fig. 3.9-6.**). Just as described[31], the epitopes of NTD-targeting D12 and RBD-targeting C04/S11 require a RBD-up conformation to be exposed, and all the epitopes include less shielded areas colored in light green or white.

Finally, to recapitulate all the findings from **section 3.5** to **3.9**, a comprehensive modelling approach as described in **section 2.18** was employed to generate the realistic, membrane-bound HCoV-229E S-hAPN model (**Fig. 3.9-7.**), offering insights into the molecular mechanism of virus cell entry *in vivo*.

doi:10.6342/NTU202301899

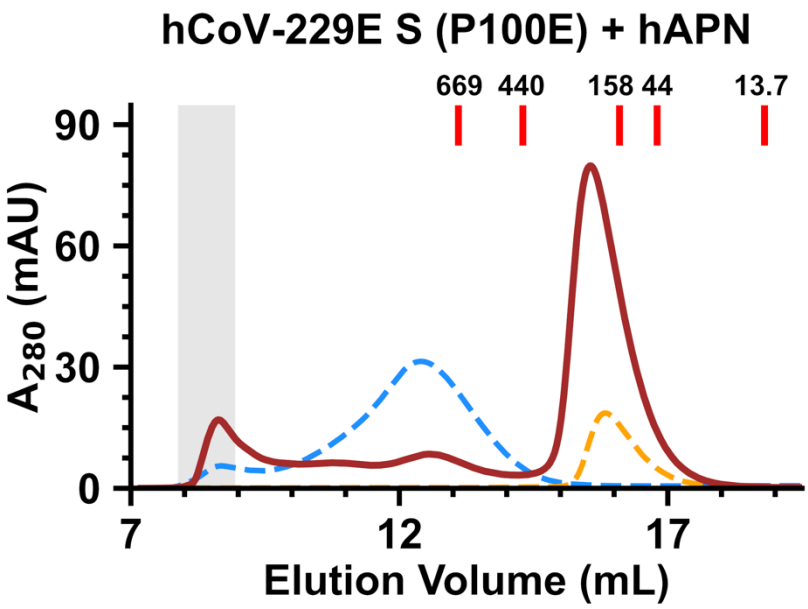


Figure 3.9-1. SEC profile of the mixture of HCoV-229E S and hAPN ectodomain. The SEC-purified 229E S protein was incubated with SEC-purified hAPN at 4ºC for 1 hour. Then the mixture was injected into Superose 6 increase 10/300 GL column to identify the putative complex fractions (indicated by the light grey background in the plot), which were later collected and concentrated to 3.5 mg/mL for cryo-EM grid preparation. For comparison, the apo-form SEC profiles of both HCoV-229E P100E strain S and hAPN ectodomain (from **Fig. 3.3-2.** with the same color scheme) were aligned onto the SEC profile of the mixture to demonstrate the substantial shift in the elution volume.

doi:10.6342/NTU202301899

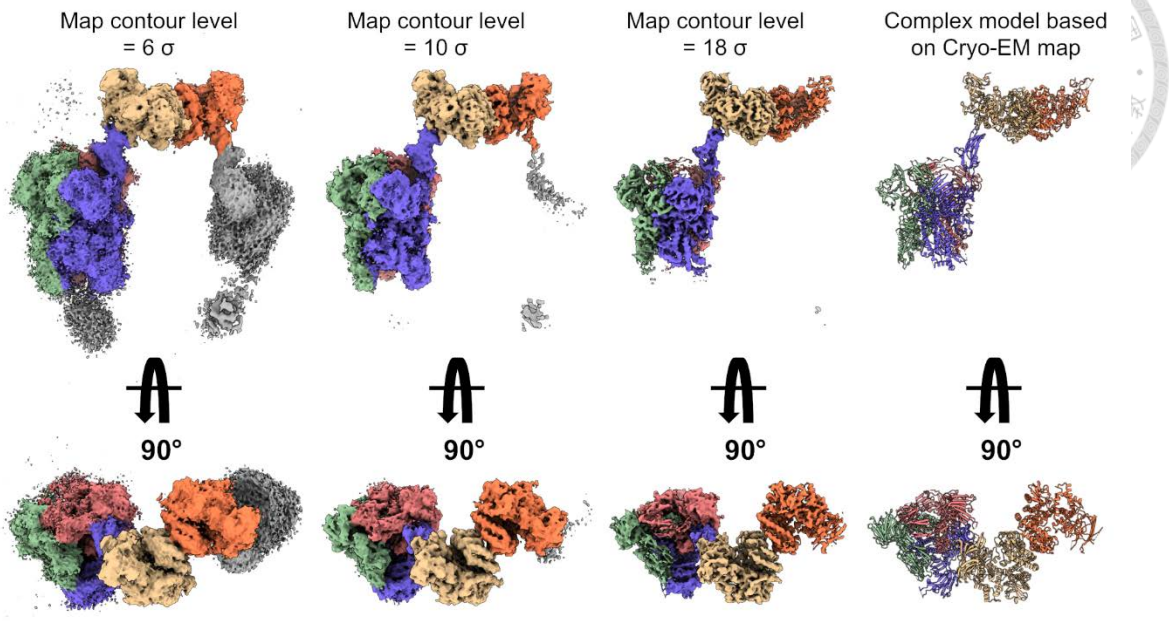


Figure 3.9-2. The low contour cryo-EM map of HCoV-229E P100E S-hAPN complex revealed two S proteins bound to a hAPN homodimer.

The cryo-EM Coulomb potential map of the HCoV-229E P100E S-hAPN complex is represented by three contour levels. From left to right, the levels are 6 sigma, 10 sigma, and 18 sigma, respectively. The molecular model on the far right corresponds to the 18-sigma contour level, where the HCoV-229E S protein adopts an RBD-up conformation to bind with a hAPN homodimer. Additionally, based on the 6-sigma contour level map, there appears to be another potential S protein bound to hAPN, resulting in a complex consisting of two S proteins in the RBD-up conformation attached to a hAPN homodimer.

doi:10.6342/NTU202301899

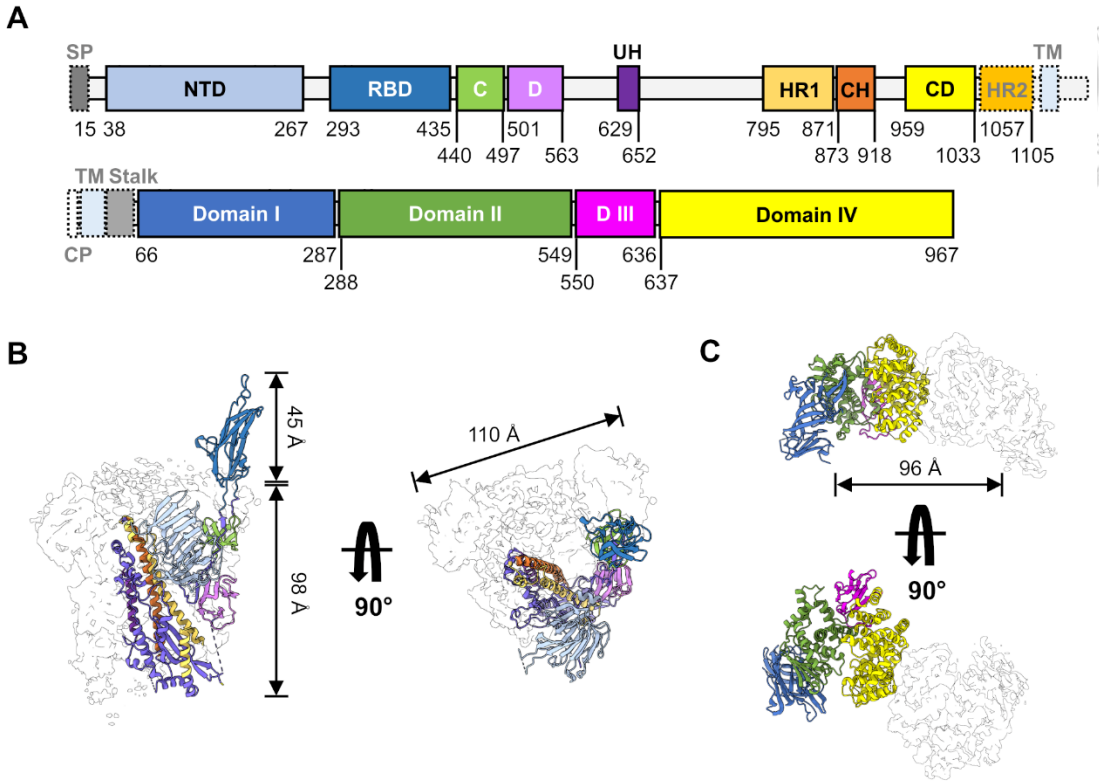


Figure 3.9-3. The molecular structure of HCoV-229E P100E S protein and hAPN ectodomain derived from cryo-EM Coulomb potential map.

(A) is a schematic of the expressed HCoV-229E P100E S protein and hAPN ectodomain. The residue range of each key domain is annotated below the schematic. The un-expressed regions are indicated by the dashed lines. (B)(C) The molecular structures of 229E P100E strain S and hAPN ectodomain were built based on the cryo-EM map described in **Fig. 3.10-1**. The color scheme is the same as (A), indicating the corresponding structures of the key domains in the schematic. The scales of the molecules are further illustrated with the two headed arrows and the distances in angstrom calculated by PyMOL built-in tool.

doi:10.6342/NTU202301899

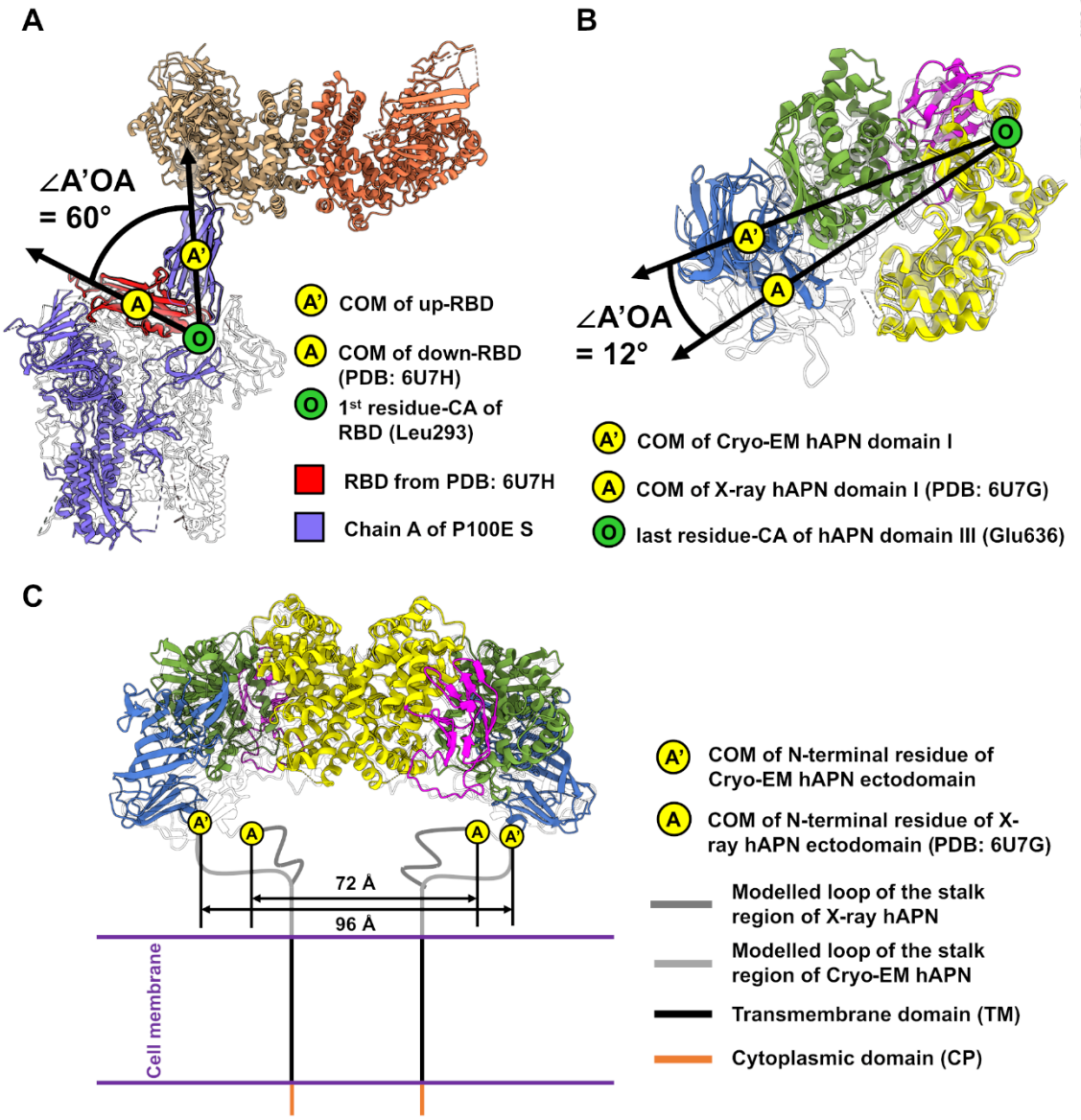


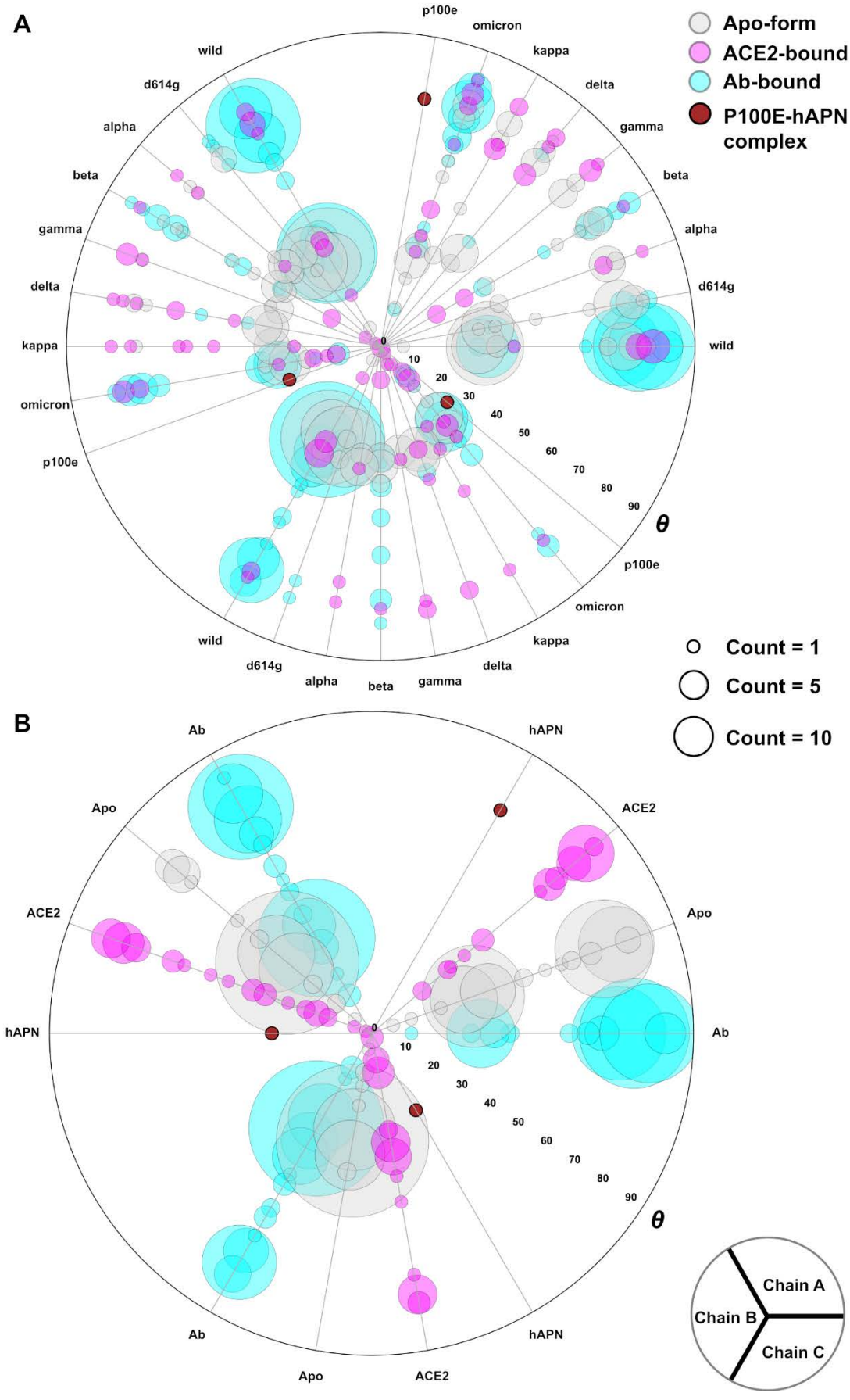
Figure 3.9-4. A comparison between the molecular structures solved by cryo-EM and the previously reported models.

The C-alpha atoms chosen for the angle and distance calculations are as described in the legends. (A) A comparison between the solved cryo-EM structure of HCoV-229E S protein and the previously reported cryo-EM model (PDB ID: 6U7H). The molecular model proposed in this thesis is the first-ever RBD-up conformation observed among the human αCoVs, resulting in a 60-degree angle between the center of mass (COM) of the two RBDs. (B) A comparison between the solved cryo-EM structure of hAPN ectodomain and the previously reported X-ray model (PDB ID: 6U7G). There is 12-degree angular discrepancy between the center of mass (COM) of the two domain Is. (C) hAPN can adopt either open form (inactive) or closed form (active) to regulate the enzymatic activity for peptide processing[16]. The solved cryo-EM structure of hAPN was aligned with the X-ray model (PDB ID: 6U7G, closed form) to show the comparatively wider distance between

doi:10.6342/NTU202301899

the COMs of the N-terminal residues. The stalk and the transmembrane regions were constructed to-scale.

doi:10.6342/NTU202301899



doi:10.6342/NTU202301899

Figure 3.9-5. RBD-up angle of HCoV-229E P100E S protein compared to that of all SARS-CoV-2 S proteins available on Protein Data Bank (PDB).

The molecular compositions (apo-form, ACE2-bound, Ab-bound, and P100E-hAPN complex with corresponding colors), PDB entry count (discrete representative count numbers 1, 5, and 10), and the angular representation of the protein chains (from A to C chain) are described in the legends located in the top-right, middle-right, and bottom-right, respectively. (A) The analysis of the RBD-up angles of all currently available PDB structures of SARS-CoV-2 S proteins in comparison with that of the solved cryo-EM structure of HCoV-229E S. This figure was plotted against the variant names of the S proteins, with each variant separated by 10 degrees. (B) The analysis of the RBD-up angles of all currently available PDB structures of SARS-CoV-2 S proteins in comparison with that of the solved cryo-EM structure of HCoV-229E S. This figure was plotted against the molecular compositions, with each state separated by 20 degrees. Please refer to **Figs. A18-19.** for detailed programming flowcharts and the related GitHub URLs of the Python scripts.

doi:10.6342/NTU202301899

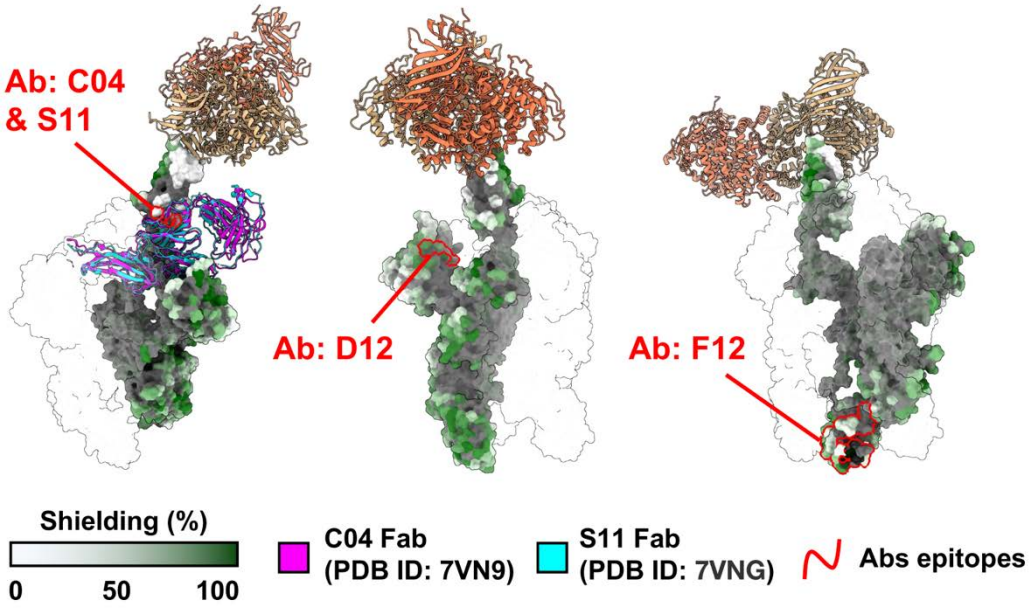


Figure 3.9-6. Antibody epitope mapping of HCoV-229E P100E S onto the cryo-EM-derived complex structure.

PDB ID: 7VN9 and 7VNG reported previously[31] were aligned onto the RBD region of the solved cryo-EM complex structure. The epitopes of the Fabs are delineated with red curves. Of note, for C04 and S11, the epitope of the adjacent RBD is indicated to avoid the obstruction from the crystal structures. The shielding effect of the glycans on the complex was quantified by GlycoSHIELD[83] and visualized with surface heatmap as described in **RESULTS 3.9**. The names of the Fabs and the color scheme are as indicated by the red labels and the legend in the bottom panel. Of note, only the chain A of the S protein is shown here for clarity.

doi:10.6342/NTU202301899

Table 3.9-1. Cryo-EM data collection parameters, refinement and validation report for HCoV-229E P100E-hAPN complex.

	P100E-hAPN complex	mono-hAPN + singe S	single S	mono-hAPN + single RBD	hAPN dimer + single RBD + hinge	mono-hAPN	single RBD + hinge	P100E-hAPN complex
Job type	NU-refinement	local refinemet	local refinemet	local refinemet	local refinemet	local refinemet	local refinemet	composite map[a]
Data collection and processing								
Magnification	81000x	81000x	81000x	81000x	81000x	81000x	81000x	81000x
Voltage (keV)	300	300	300	300	300	300	300	300
Electron exposure (e–/Å^2)	50	50	50	50	50	50	50	50
Defocus range (μm)	-1.5~-2.0	-1.5~-2.0	-1.5~-2.0	-1.5~-2.0	-1.5~-2.0	-1.5~-2.0	-1.5~-2.0	-1.5~-2.0
Pixel size (Å)	1.08	1.08	1.08	1.08	1.08	1.08	1.08	1.08
Symmetry imposed	C1	C1	C1	C1	C1	C1	C1	C1
Initial particle number	253,164	112,541	112,541	112,541	112,541	112,541	112,541	112,541
Final particle number	112,541	112,541	112,541	112,541	112,541	112,541	112,541	112,541
Map resolution (Å)	4.37	3.71	3.39	3.57	3.61	3.73	3.23	4.10
FSC threshold	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143
Refinement								
Initial model (PDB code)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	6U7H (for S), 6U7G (for hAPN)
Model resolution (Å)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	3.60
FSC threshold	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.143
Map sharpening B factor (Å^2)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	108.5
Model composition								
Non-hydrogen atoms	N/A	N/A	N/A	N/A	N/A	N/A	N/A	33,410
Protein residues	N/A	N/A	N/A	N/A	N/A	N/A	N/A	4,170
Ligands	N/A	N/A	N/A	N/A	N/A	N/A	N/A	37
B factors (Å^2)								
Protein	N/A	N/A	N/A	N/A	N/A	N/A	N/A	85.78
Ligand	N/A	N/A	N/A	N/A	N/A	N/A	N/A	89.67
RMSD								
Bond lengths (Å)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.011
Bond angles (°)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.923
Validation								
MolProbity score	N/A	N/A	N/A	N/A	N/A	N/A	N/A	1.92
Clashscore	N/A	N/A	N/A	N/A	N/A	N/A	N/A	5.31
Poor rotamers (%)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	2.27

doi:10.6342/NTU202301899

Ramachandran plot								
Favored (%)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	94.61
Allowed (%)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	5.04
Outliers (%)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.34

[a]. The composite map was generated from 6 local refinement jobs using Phenix[89].

doi:10.6342/NTU202301899

Table 3.9-2. A summary of N-glycosylation sites of 229E S-hAPN complex observed experimentally through mass spectrometry and cryo-EM.

P100E S protein			hAPN ectodomain		
N-sites	**Analytical method**		**N-sites**	**Analytical method**	
	MS	Cryo-EM		MS	Cryo-EM
23	O	X	**128**	X	X
62	O	O	**234**	O	X
98	O	O	**265**	O	O
147	O	X	**319**	X	O
171	O	X	**527**	O	X
220	O	X	**573**	X	X
243	O	O	**625**	O	O
326	O	X	**681**	X	O
333	O	X	**735**	X	X
440	O	O	**818**	O	X
464	O	X			
518	O	O			
538	X	O			
542	X	X			
568	O	X			
581	O	X			
587	O	X			
663	O	X			
671	O	O			
714	O	X			
930	O	X			
1015	O	X			
1020	O	X			
1037	O	X			
1049	X	X			
1061	O	X			
1066	O	X			
1076	O	X			
1082	O	X			
1096	O	X			

doi:10.6342/NTU202301899

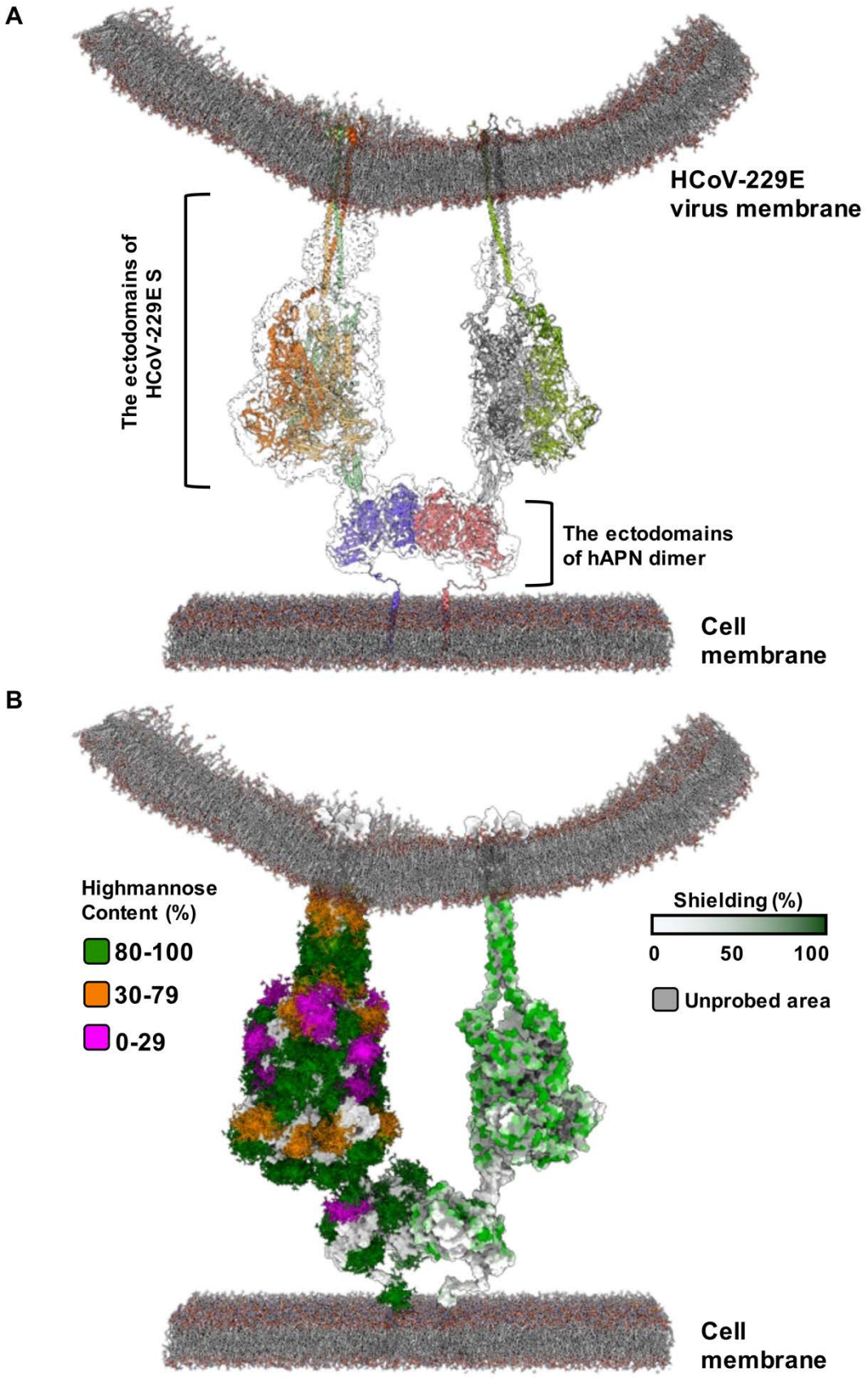


Figure 3.9-7. The proposed model of HCoV-229E S attaching to cell-surface hAPN to perform virus entry.

doi:10.6342/NTU202301899

A comprehensive modeling approach, which is described in detail in section **2.18**, was adopted to generate the realistic 3D schematic. (A) The cryo-EM derived Coulomb potential map is represented in transparent envelope to show its fitting to the corresponding molecular structure. A low map contour of 5 sigma is used to show the stalk and the neighboring spike protein. Extra map noises are cleaned by ChimeraX[38] "Hide Dust" tool for clarity. (B) The glycan ensembles and the $SASA_{rel}$ heatmap generated from MD-based tool, GlycoSHIELD[83], are shown for the left and right parts of the 229E-S-hAPN complex, respectively. The color scheme for the high-mannose content is shown in the top-left legend, while that of $SASA_{rel}$ heatmap is on the top-right.

doi:10.6342/NTU202301899

Chapter 4 DISCUSSION

4.1 A look into the distinct DSF profile of HCoV-229E Seattle strain and its potential link to the general structural features of the spike proteins

According to **Fig. 3.3-3. A**, the overall pattern of the Seattle strain DSF profile is different from that of the P100E strain, with the latter exhibiting an additional prominent peak roughly between 55°C and 70°C. Interestingly, when the protein samples were re-heated (the rescan data drawn in dashed lines), this additional peak disappeared and rendered the profiles of the both strains similar. Nonetheless, all the following experiments, including NSEM, BLI assays, and N-glycosylation analysis with MS, consistently revealed distinct properties between the two strains. Thus, we hypothesized that DSF was just an early indication of a fundamental discrepancy in the protein structures of the P100E and Seattle spikes.

To test this hypothesis, a DSF-mimicking experiment was carried out to find the potential link between the S structure and the DSF profile pattern (**Fig. 4.1-1.)**. In brief, the spike proteins of the selected HCoVs (NL63, SARS, MERS, and HKU1) were aliquoted and subjected to room temperature, 65°C, and 95°C 3-min incubation, respectively. Afterwards, all samples were equilibrated back to room temperature for NSEM grid preparation. The three temperatures emulate the initial protein state, the first peak of the first derivative of 350 nm/330 nm fluorescence intensity ratio, and the end of the DSF heating process. The spike protein structure from each status was crudely examined by NSEM. The results of this DSF simulation experiment are shown in **Fig. 4.1-2.**

Based on the profiles in **Fig. 4.1-2. A**, all spikes except for that of NL63 underwent certain conformational change before 65°C, leading to one or several peaks in the first derivative curves. Moreover, akin to **Fig. 3.3-3. A**, the rescan profiles of the selected

doi:10.6342/NTU202301899

spikes lose all these additional peaks (repeated two times; colored in grey and dark grey). This hypothesis on the DSF profile-related structural alteration was further visualized by NSEM analysis in **Fig. 4.1-2. B**, where six representative 2D classes processed by CryoSPARC[41] were shown to capture the overall molecular integrity of the spike proteins. Similar to the DSF profiles, all the spikes underwent drastic structural alteration after the 3-min incubation of 65°C except for NL63, and as the temperature rose, most spikes started to adopt an elongated shape reminiscent of its post-fusion form[90, 91]. In contrast, for NL63, these novel elongated structures were already observed in the initial state, possibly contributing to the almost unchanged DSF patterns from Heating #1 to Rescan #2 in **Fig. 4.1-2. A**.

To test the reproducibility of the DSF results in **Fig. 3.3-3. A**, the S of P100E and Seattle strains were heated again followed by one rescanning. Also, the S protein of another αCoV, porcine epidemic diarrhea virus (PEDV), was assayed in parallel to investigate the hypothesis above. The results are shown in **Fig. 4.1-3.**, where the single-peaked DSF profile of the Seattle strain S corresponds to the elongated protein structures presented in the NSEM 2D classes, similar to the case for NL63.

In summary, DSF could serve as an effective preliminary screening to ensure a well-folded S protein before proceeding to more advanced experiments. And the distinct peaks in the profile might indicate a shared multi-step unfolding procedure among the S variants[92]. Based on the experimental results and the inflection temperatures (Ti), we suggest that the higher Ti (ranging from 75-85°C) might originated from the shared T4 fibritin foldon in our protein expression constructs (**Fig. 3.3-1. B-C**), whereof the reemerging DSF rescanning profiles are indicative of the melting temperature and the quickly refolded nature of T4 fibritin reported previously[93-95]. While the unfolding of T4 fibritin can be reversed, the thermal damage to the rest of the S protein structure should

doi:10.6342/NTU202301899

be irreversible according to the permanently altered DSF profiles and the loss of the lower Ti. The remaining challenge is to identify the underlying cause for the initially disrupted S protein structures of HCoV-229E Seattle strain and HCoV-NL63.

doi:10.6342/NTU202301899

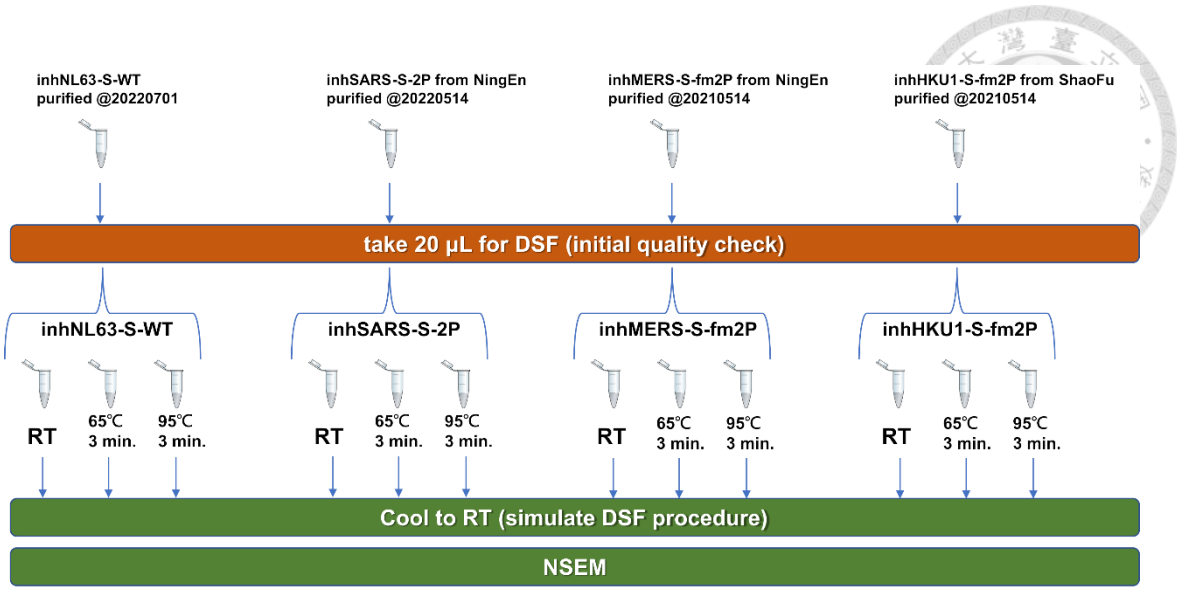


Figure 4.1-1. The experimental workflow to simulate DSF procedure.

doi:10.6342/NTU202301899

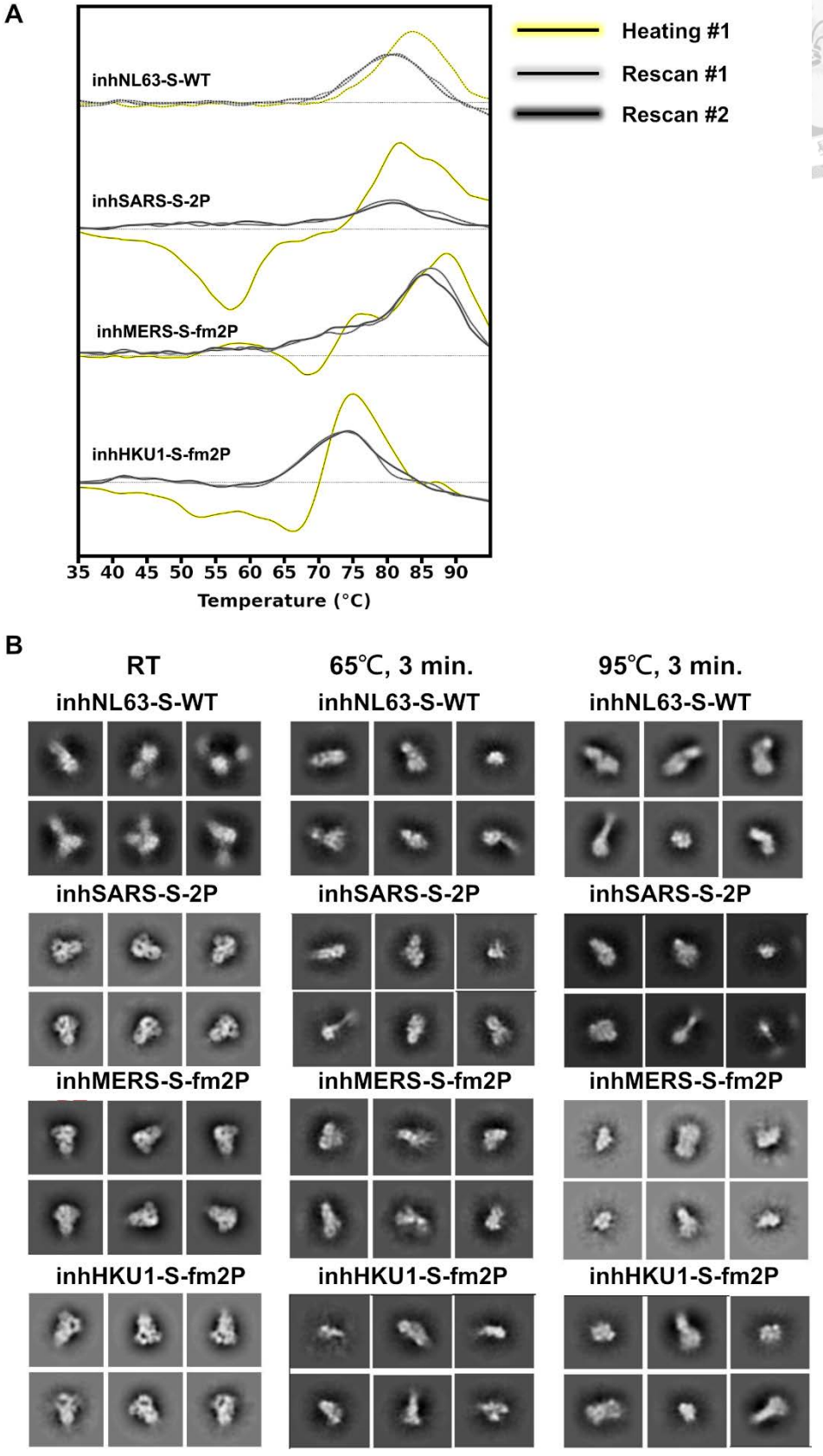


Figure 4.1-2. The results of DSF simulation experiment from four selected spike proteins.

doi:10.6342/NTU202301899

(A) DSF profiles of the selected spike proteins. From top to bottom: in-house wild-type HCoV-NL63 S, in-house SARS-CoV-1 S with dual proline mutation, in-house MERS-CoV S with dual proline and furin cleavage site mutations, and in-house HCoV-HKU1 S with dual proline and furin cleavage site mutations. The DSF experiment was done by firstly heating the spike proteins to disrupt the molecular structure, followed by re-heating for two times (rescan #1 and #2) to look for the remaining folded structures, which may stem from the potential thermal resilience, of the samples. (B) The representative 2D classification results processed by CryoSPARC[41] from the NSEM micrographs of the selected proteins subjected to different temperature treatments listed on the top panel. The NSEM data processing procedure is described in **MATERIAL AND METHODS 2.6**.

doi:10.6342/NTU202301899

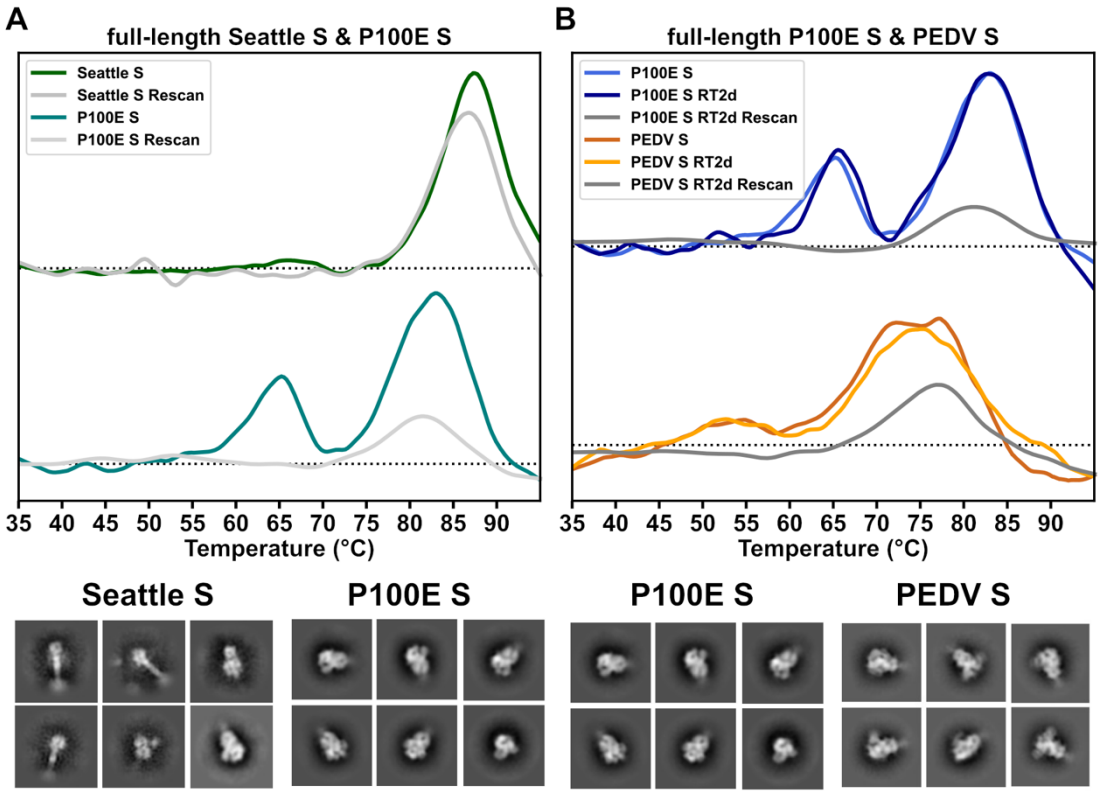


Figure 4.1-3. The potential relationship between the S protein structure and the DSF profile.

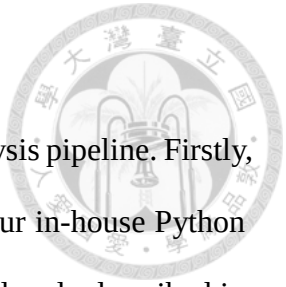
The DSF profiles of HCoV-229E P100E strain S, Seattle strain S, and Porcine epidemic diarrhea virus (PEDV) S were paired with the corresponding NSEM representative 2D classes to show the possible link between the S protein structure and its unfolding procedure. The color scheme of the line art is described by the top-left legend of both (A) and (B).

doi:10.6342/NTU202301899

4.2 The discrepancy in N-linked glycosylation profile of HCoV-229E strains is potentially pertinent to TAMP (trimer-associated mannose patch) that was first described in HIV-1 envelope protein (Env)

In **section 3.6**, we have discussed the discrepancy in the N-glycosylation profile between the spike proteins of the P100E and Seattle strain, and we hypothesized the further processed glycan distribution of the latter was due to the compromised structural integrity observed via NSEM. This relationship between the glycosylation and the protein structure was first described in HIV-1 envelope (Env) protein[73, 96, 97], where the N-glycan profiles of monomeric Env, uncleaved pseudo-trimeric Env, and native-like Env were analyzed and compared in parallel. As elaborated in this review[73], glycan analyses conducted on both virion-associated Env and recombinant trimers reveal a higher oligo-mannose content compared to Env monomers. Within the structure of a naturally folded trimer, interactions between glycans and proteins, as well as glycans and other glycans, are believed to contribute to the emergence of an additional feature known as the "trimer-associated mannose patch" (TAMP). In practice, the TAMP does not always arise from a complete transition in the glycan processing state, going entirely from complex to oligo-mannose, but rather signifies an increase in the prevalence of under-processed glycans. The formation of the TAMP relies on cleavage-induced, native-like envelope folding and is not observed in un-cleaved or improperly assembled trimers. These descriptions are in line with our observations on the Seattle S structure and its MS results (**sections 3.3, 3.5, 3.6**). Intriguingly, the NSEM images of the soluble uncleaved pseudo-trimeric Env shown in the review are eerily similar to those of the Seattle strain spike proteins, supporting the idea that the splayed-out Seattle S allows more mannosidase accessibility.

doi:10.6342/NTU202301899



4.3 The limitations of our glycopeptide analysis pipeline

The following are the three major limitations of our glycopeptide analysis pipeline. Firstly, glycopeptides with multiple N-sites are excluded automatically via our in-house Python script before proceeding to quantification, for which the reason was already described in section **3.5**. Even with the overall decent data quality of both P100E and Seattle strain N-glycosylation analyses, sites N538 and N542 in P100E, i.e., N536 and N540 in Seattle, were still prevented from being quantified since they were located on the same glycopeptide. Although alpha-lytic protease provided the orthogonal specificity for trypsin and chymotrypsin, these sites could not be separated. Manual inspection of the raw LC-MS/MS signals or a new set of digesting enzymes will be needed in this case, especially when the N-sites affected are potentially involved in essential physiological processes. Secondly, with a mediocre data quality seen in hAPN dataset, most data entries representing the identified glycopeptides by the search engines will be filtered out by the criteria described in **2.14**., leaving certain N-sites with single data entry. When normalized, the single entry will be visualized as 100% in our bar and pie charts, creating a rather skewed and inflated impression on the glycoform composition, as in the cases of N625 and N818 on hAPN ectodomain (**Figs. 3.7-1,2,3**). This problem was further complicated when we strived to construct hAPN glycoSHIELD model based on the results from N-glycopeptide analysis, where for each N-site, a single major glycofrom was chosen to generate the glycan ensemble. This inevitably gave rise to an over simplified view of the glycan canopy with numerous potential glycoforms being underrepresented. Moreover, the abundance of the identified glycopeptides of each N-site might vary significantly, not only between different experimental samples but the same protein. Whether the results of N-glycopeptide analysis from disparate batches and the corresponding glycoSHIELD models can be combined and discussed inside the same system, e.g., our final complex

doi:10.6342/NTU202301899

model (**Fig. 3.9-6**), merits further investigation. Lastly, O-glycosylation cannot be analyzed by our current pipeline. This could be one of the key features to incorporate into our future computational workflow. Since core-1 and core-2 structures were reported to occupy threonine (T678) near the furin cleavage site of SARS-CoV-2 S protein[98], automating O-glycopeptide analysis via Python programming might shed light on the influence of O-glycosylation on proteolytic activation in an accelerated and systematic manner.

doi:10.6342/NTU202301899

4.4 The limitations of our cryo-EM analysis

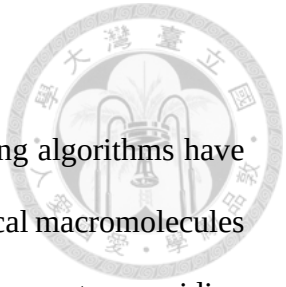
In section **3.9**, we have pointed out that the down-form RBDs of 229E S proteins might be induced by the surrounding hAPNs, whereof the increased conformational dynamics led to an inferior map quality. Nevertheless, with the advent of novel cryo-EM data processing tools, such as 3D variability analysis[99] and 3DFlex[100], we should be able to further single out particle groups with distinct conformations and even explore the continuous molecular heterogeneity. Unfortunately, we could not carry out these analyses due to some technical issues mainly stemming from the upper limit of GPU memory. Regardless, the results will enable us to delve deeper into the intrinsic propensity of 229E S RBDs to adopt the up conformation. It has been reported that SARS-CoV-2 S RBDs are capable of taking on up conformation sans the presence of receptors or antibodies[101, 102]. In contrast, both HCoV-229E and NL63 S RBDs showed solely the down-form as apo-form proteins[11, 44, 69]. This intriguing feature of the latter was also observed on the intact virions resolved by cryo-ET[103]. It would be interesting to see if the conformation of 229E S proteins on the intact virions is also down-form dominant. Together with our single particle analysis, these data will offer a more comprehensive perspective on the inter-genus and intra-genus disparity in the inherent RBD dynamics of human CoVs.

Another limitation of our cryo-EM analysis is the lack of lipid membrane. Although we strived to emulate the realistic physiological composition and constructed the complex model to scale (**Fig. 3.9-6.**), the absence of both viral and cellular membranes grants the molecules extra degrees of freedom, compromising the reliability of the related discussion (**Figs. 3.9-2,3,6**). Just as mentioned above, cryo-ET could be an optimal approach to see if our observations were experimental artifacts. The spatial distribution and the pertinent statistics of the S proteins could as well validate our proposed binding stoichiometry of HCoV 229E S-hAPN complex model.

doi:10.6342/NTU202301899

Finally, due to the limited resolution, the binding interface between HCoV 229E S and hAPN cannot be confidently determined, and the molecular interactions were mainly based on the previous X-ray crystallography-derived model[44]. Since the binding interface is of great importance when it comes to the adaptive evolution of the virus and the rational immunogen design, we will expand on this topic in section **4.6**.

doi:10.6342/NTU202301899



4.5 The limitations of our glycoSHIELD analysis

The advancements in electron microscopy (EM) and image-processing algorithms have made it possible to study the movements of specific regions in biological macromolecules through cryo-EM. However, cryo-EM has limitations when it comes to providing complete structural information on N-glycans. While certain parts of the glycan structures, such as the two GlcNAc units and the core fucose, may be visible in EM maps, the remaining portions of the glycan trees are difficult to define due to their highly dynamic nature. In contrast, MS offers a different approach by providing detailed information about the composition and distribution of individual glycan types at specific sites. Therefore, cryo-EM and MS complement each other as powerful tools for studying glycoproteins. To achieve a more comprehensive understanding of N-glycans on target proteins, researchers have combined the advantages of cryo-EM and MS to generate fully glycosylated models using CHARMM-GUI[47] (**Figs. 3.5-3 B, 3.6-3 B, and 3.7-3 B**).

However, it is important to note that protein-bound glycans are highly dynamic and act as a shield to evade the host immune system[69, 80, 104]. The static fully glycosylated models generated earlier may not accurately capture the dynamic behavior of glycans and their shielding effects. To address this, another modeling approach called GlycoSHIELD[83] was used to calculate the solvent-accessible surface area (SASA) of the protein under the influence of different glycan conformations. This method quantitatively describes the impact of glycans. However, there are two limitations in this pipeline. Firstly, the cryo-EM maps often have low local resolution, resulting in regions with undefined atomic coordinates in published PDB structures. Although these missing regions were manually reconstructed using Coot[45], the static models with the derived shielding effect only provide a snapshot of the dynamic proteins, disregarding their natural local motions and substantial conformational changes. Secondly, while glycans are typically known for

doi:10.6342/NTU202301899

their role in immune evasion[69, 80, 104], it has been reported that broadly neutralizing antibodies (bnAbs) can target the mannose patch on the envelope spike of HIV-1[105]. This suggests that glycans can also be recognized as epitopes by antibodies. Although most antibody structures aligned with the GlycoSHIELD-derived ΔSASA models in this study corresponded to surfaces with low or moderate shielding scores (**Fig. 3.9-6**), the potential exceptions cannot be overlooked.

To summarize, for the discovery of the novel bnAbs, the integrative pipeline established in this study should be used not only for displaying the overwhelming heterogeneity of the N-glycan composition on the 3D protein model, but for searching the conserved homogeneity of the mannose patches throughout the extensive antigenic drift of the HCoVs.

doi:10.6342/NTU202301899

4.6 HCoV-229E S epitope mapping and rational immunogen design

In section **4.4**, we have mentioned that HCoV 229E S is low in the intrinsic propensity to adopt RBD up-conformation based on the previous cryo-EM apo-form structures[11, 44]. Recently, Xiang, et al. [31] conducted the epitope mapping of HCoV 229E S and the results suggested that NTD is antigenic dominant, which is in stark contrast to previous studies on SARS-CoV, MERS-CoV and SARSCoV-2 S proteins, wherein their RBDs instead of NTDs are most frequently targeted by nAbs[106, 107]. In the same year, Shi, et al. [108] also verified the antigenicity of NTD, on which a novel nAb, 5H10, was further characterized. The stronger antigenicity of NTD might be a manifestation of the down-form dominant structural distribution of the apo-form HCoV 229E S protein, since most RBD related epitopes are hidden in this state. Interestingly, it has been reported that domain A (NTD) of HCoV-HKU1 S protein binds to a sialoglycan-based primary receptor to trigger the open-state of domain B (akin to RBD of 229E S) via a allosteric inter-domain crosstalk[109]. Additionally, for the other human alphacoronavirus, HCoV-NL63, heparan sulfate was found to associate with domain 0 (a duplicate of domain A), wherein a positively charged patch might putatively mediate the binding interaction[69]. Taken together, 229E S protein could possibly utilize an NTD-binding co-receptor to increase the dynamic of RBD, boosting the probability of capturing the cell surface hAPN. Whether this initial structural priming exists before proceeding to the RBD-up conformation in our complex model warrants future research.

Since the epitope mapping of 229E S is available, we can incorporate the known epitopes and the corresponding nAbs into our glycoSHIELD analysis summary (**Fig. 4.6-1.**). 5H10 binds to NTD E1 motif (residues 147-167), with F159 being indispensable[108]. Shi, at al. also found that mutations at position 159 gradually appeared over time and completely abolished the neutralizing ability of 5H10, supporting the notion that position

doi:10.6342/NTU202301899

159 may be under selective pressure during the human epidemic[108]. Although the major cause of the abolishment could be the loss of π-π interaction between 5H10 heavy chain and residue 159[108], the results of our glycoSHIELD total shielding effect analysis are also in line with the findings and bring in the glycosylation perspective as another contributing factor. The total shielding percentage of P100E is around 33% at position 159, and an obvious plunge can be seen in **Fig. 4.6-1.** On the contrary, the percentage of Seattle jumped to 100%, rendering the site inaccessible to the antibody. The effective amino acid value of position 159 calculated from the 96 HCoV 229E variants' S sequences (**Fig. 3.1-1. A**) is about 1.5, corroborating its gradual mutations throughout the virus evolution. The epitopes of D12 (residues 206-215) were found to be the "NTD antigenic supersite" [31]. Nonetheless, the corresponding total shielding percentages of both P100E and Seattle are 100%, except for site 215 experiencing 57% and 53%, respectively. Since insect cell sf9 was used for antibody characterization as previously described[31], we hypothesize that the profound shielding effect estimated by our glycoSHIELD workflow might come from the more heterogeneous and bulkier glycoforms of Expi293 reservoir. C04 and S11 are reminiscent of the cryptic epitopes of SARS-CoV-2[110-113], targeting the distal end of RBD that is only exposed when the adjacent RBD adopts the up-state. Thus, **Fig. 4.6-1** is unable to illustrate the related shielding effect due to the all-down RBDs of 7CYC. It is worth noting that both sites 418 and 419 are conserved across 96 S sequences, which is beneficial for maintaining the neutralizing activity. Lastly, F12 was claimed to be the first identified S2-directed nAb against alpha-HCoVs[31]. Based on our GlycoSHIELD analysis, the epitopes boast both broad windows of low total shielding and mostly conserved sequence ranging from residue 1000 to 1033.

Even though HCoV-229E is constantly evolving, and the strong antigenic regions, e.g., RBD and NTD, are also extremely variable in protein sequences, we can now take

doi:10.6342/NTU202301899

advantage of our pipeline established in this study to systematically probe for the epitope candidates. To be stringent, the probing criteria are as follows: 1) effect amino acids value should equal to one, guaranteeing a conserved residue from year 1979 to 2022 to minimize the chance of immune escape. 2) the total shielding percentage should be less than 50%, leaving enough space for antibody interaction. **Table 4.6-1.** shows all residues meeting these two criteria, and some of them are already included in the known nAbs[31, 108].

In summary, our pipeline can easily be extrapolated to other proteins of interest, and with the help of Python programming, we can automatically process sequences in large scale in search of the potential epitopes, providing critical information for rational immunogen design.

doi:10.6342/NTU202301899

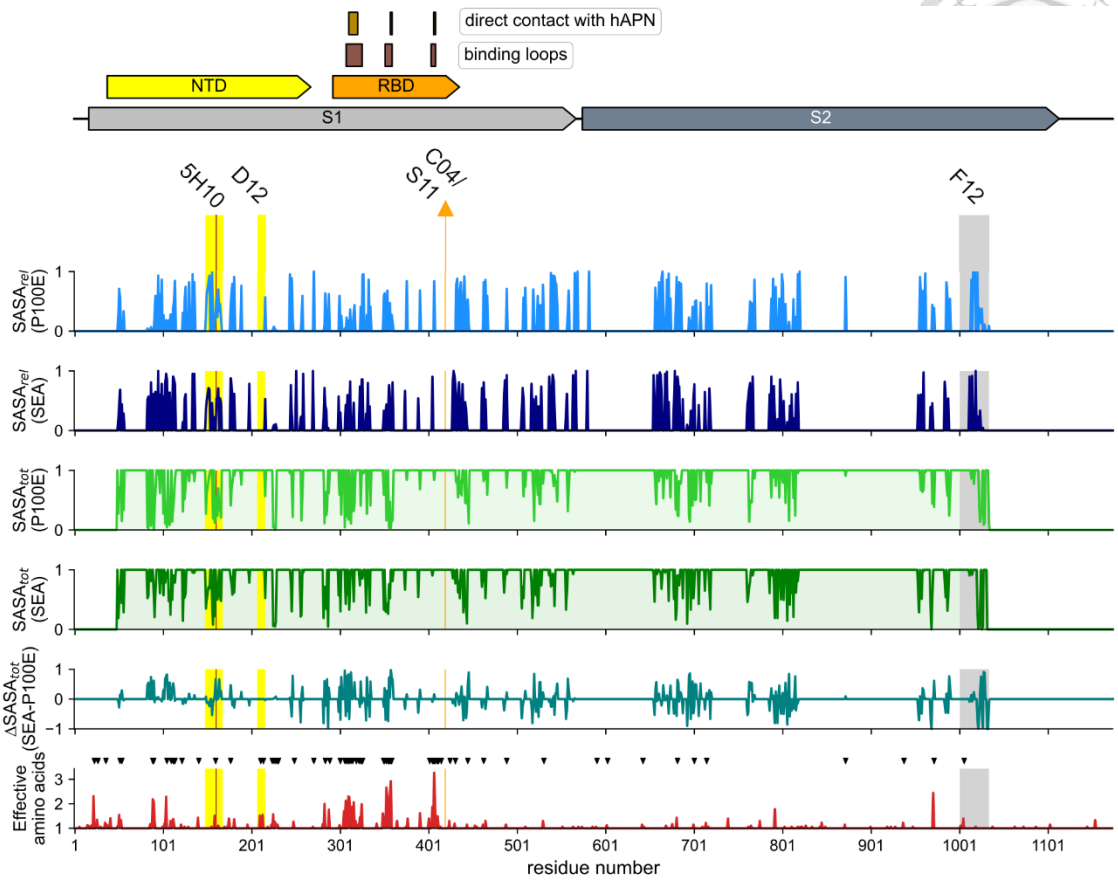


Figure 4.6-1. Mapping the known 229E neutralizing epitopes to glycoSHIELD summary.

The glycoSHIELD summary plot is the same as **Fig. 3.8-4.**, except for the annotated epitopes of the known 229E S neutralizing antibodies. The epitopes of the two NTD-binding nAbs, 5H10[108] and D12[31], are residues 147-167 and 206-215, respectively. The regions are marked by the yellow background of the subplots. For 5H10, the reported indispensable residue F159[108] is further indicated with the brown line within the yellow region. The epitopes of the two RBD-binding nAbs, C04 and S11, are both 418-419[31], which are marked by the orange background with an upward arrow. The epitopes of the novel S2-binding nAb, F12[31], are marked by the grey background.

doi:10.6342/NTU202301899

Table 4.6-1. Conserved potential antibody epitopes.

residue num.	P100E total shielding (%)	Seattle total shielding (%)	effective aa.	region	known nAb
55	33	29	1	NTD	
106	18	47	1	NTD	
108	28	24	1	NTD	
110	14	50	1	NTD	
122	31	14	1	NTD	
148	27	35	1	NTD	
156	19	8	1	NTD	
158	12	23	1	NTD	5H10
161	24	50	1	NTD	
165	21	48	1	NTD	
226	0	5	1	NTD	
256	16	24	1	NTD	
315	32	47	1	RBD	
441	46	48	1		
443	44	10	1		
679	34	16	1	S2	
969	44	46	1	S2	
1022	39	4	1	S2	
1023	14	33	1	S2	F12
1025	37	0	1	S2	
1033	9	0	1	S2	

doi:10.6342/NTU202301899

REFERENCES

(1) Zhao, X.; Ding, Y.; Du, J.; Fan, Y. 2020 update on human coronaviruses: One health, one world. *Med Nov Technol Devices* **2020**, 8, 100043. DOI: 10.1016/j.medntd.2020.100043.

(2) Sturman, L. S.; Holmes, K. V. The Molecular Biology of Coronaviruses. In *Advances in Virus Research*, Lauffer, M. A., Maramorosch, K. Eds.; Vol. 28; Academic Press, **1983**; pp 35-112.

(3) Liu, D. X.; Liang, J. Q.; Fung, T. S. Human Coronavirus-229E, -OC43, -NL63, and -HKU1 (Coronaviridae). In *Encyclopedia of Virology*, **2021**; pp 428-440.

(4) Jiang, S.; Hillyer, C.; Du, L. Neutralizing Antibodies against SARS-CoV-2 and Other Human Coronaviruses. (1471-4981 (Electronic)). From **2020** May.

(5) Millet, J. K.; Jaimes, J. A.; Whittaker, G. R. Molecular diversity of coronavirus host cell entry receptors. *FEMS Microbiol Rev* **2021**, *45* (3). DOI: 10.1093/femsre/fuaa057.

(6) Gui, M.; Song, W.; Zhou, H.; Xu, J.; Chen, S.; Xiang, Y.; Wang, X. Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding. *Cell Res* **2017**, *27* (1), 119-129. DOI: 10.1038/cr.2016.152.

(7) Pallesen, J.; Wang, N.; Corbett, K. S.; Wrapp, D.; Kirchdoerfer, R. N.; Turner, H. L.; Cottrell, C. A.; Becker, M. M.; Wang, L.; Shi, W.; et al. Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. *Proc Natl Acad Sci U S A* **2017**, *114* (35), E7348-E7357. DOI: 10.1073/pnas.1707304114.

(8) Yang, T.-J.; Yu, P.-Y.; Chang, Y.-C.; Chang, N.-E.; Tsai, Y.-X.; Liang, K.-H.; Draczkowski, P.; Lin, B.; Wang, Y.-S.; Chien, Y.-C.; et al. Structure-activity relationships of B.1.617 and other SARS-CoV-2 spike variants. *bioRxiv* **2021**, 2021.2009.2012.459978. DOI: 10.1101/2021.09.12.459978.

(9) Kirchdoerfer, R. N.; Cottrell, C. A.; Wang, N.; Pallesen, J.; Yassine, H. M.; Turner, H. L.; Corbett, K. S.; Graham, B. S.; McLellan, J. S.; Ward, A. B. Pre-fusion structure of a human coronavirus spike protein. *Nature* **2016**, *531* (7592), 118-121. DOI: 10.1038/nature17200.

(10) Zhang, K.; Li, S.; Pintilie, G.; Chmielewski, D.; Schmid, M. F.; Simmons, G.; Jin, J.; Chiu, W. A 3.4-A cryo-electron microscopy structure of the human coronavirus spike trimer computationally derived from vitrified NL63 virus particles. *QRB Discov* **2020**, *1*, e11. DOI: 10.1017/qrd.2020.16.

(11) Song, X.; Shi, Y.; Ding, W.; Niu, T.; Sun, L.; Tan, Y.; Chen, Y.; Shi, J.; Xiong, Q.;

doi:10.6342/NTU202301899

Huang, X.; et al. Cryo-EM analysis of the HCoV-229E spike glycoprotein reveals dynamic prefusion conformational changes. *Nat Commun* **2021**, *12* (1), 141. DOI: 10.1038/s41467-020-20401-y.

(12) Wang, C.; Hesketh, E. L.; Shamorkina, T. M.; Li, W.; Franken, P. J.; Drabek, D.; van Haperen, R.; Townend, S.; van Kuppeveld, F. J. M.; Grosveld, F.; et al. Antigenic structure of the human coronavirus OC43 spike reveals exposed and occluded neutralizing epitopes. *Nat Commun* **2022**, *13* (1), 2921. DOI: 10.1038/s41467-022-30658-0.

(13) Tang, T.; Bidon, M.; Jaimes, J. A.; Whittaker, G. R.; Daniel, S. Coronavirus membrane fusion mechanism offers a potential target for antiviral development. *Antiviral Res* **2020**, *178*, 104792. DOI: 10.1016/j.antiviral.2020.104792.

(14) Mina-Osorio, P. The moonlighting enzyme CD13: old and new functions to target. *Trends in molecular medicine* **2008**, *14* (8), 361-371.

(15) Chen, L.; Lin, Y.-L.; Peng, G.; Li, F. Structural basis for multifunctional roles of mammalian aminopeptidase N. **2012**, *109* (44), 17966-17971. DOI: doi:10.1073/pnas.1210123109.

(16) Wong, A. H.; Zhou, D.; Rini, J. M. The X-ray crystal structure of human aminopeptidase N reveals a novel dimer and the basis for peptide processing. *The Journal of biological chemistry* **2012**, *287* (44), 36804-36813. DOI: 10.1074/jbc.M112.398842

(17) Lipscomb, W. N.; Sträter, N. Recent advances in zinc enzymology. *Chemical Reviews* **1996**, *96* (7), 2375-2434.

(18) Lu, C.; Amin, M. A.; Fox, D. A. CD13/Aminopeptidase N Is a Potential Therapeutic Target for Inflammatory Disorders. *J Immunol* **2020**, *204* (1), 3-11. DOI: 10.4049/jimmunol.1900868

(19) Breitling, J.; Aebi, M. N-linked protein glycosylation in the endoplasmic reticulum. *Cold Spring Harb Perspect Biol* **2013**, *5* (8), a013359. DOI: 10.1101/cshperspect.a013359.

(20) Vembar, S. S.; Brodsky, J. L. One step at a time: endoplasmic reticulum-associated degradation. *Nat Rev Mol Cell Biol* **2008**, *9* (12), 944-957. DOI: 10.1038/nrm2546.

(21) Gecht, M.; von Bülow, S.; Penet, C.; Hummer, G.; Hanus, C.; Sikora, M. GlycoSHIELD: a versatile pipeline to assess glycan impact on protein structures. *bioRxiv* **2022**, 2021.2008.2004.455134. DOI: 10.1101/2021.08.04.455134.

(22) Yewdell, J. W. Antigenic drift: Understanding COVID-19. *Immunity* **2021**, *54* (12),

doi:10.6342/NTU202301899

2681-2687. DOI: 10.1016/j.immuni.2021.11.016

(23) Moeller, N. H.; Shi, K.; Demir, Ö.; Belica, C.; Banerjee, S.; Yin, L.; Durfee, C.; Amaro, R. E.; Aihara, H. Structure and dynamics of SARS-CoV-2 proofreading exoribonuclease ExoN. *Proc. Natl. Acad. Sci.* U. S. A. **2022**, *119* (9), e2106379119.

(24) Sanjuán, R.; Nebot, M. R.; Chirico, N.; Mansky, L. M.; Belshaw, R. Viral mutation rates. *Journal of virology* **2010**, *84* (19), 9733-9748.

(25) Jenkins, G. M.; Rambaut, A.; Pybus, O. G.; Holmes, E. C. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *Journal of molecular evolution* **2002**, *54*, 156-165.

(26) Drake, J. W.; Holland, J. J. Mutation rates among RNA viruses. *Proc. Natl. Acad. Sci.* U. S. A. **1999**, *96* (24), 13910-13913.

(27) Chan, J. F.; To, K. K.; Tse, H.; Jin, D. Y.; Yuen, K. Y. Interspecies transmission and emergence of novel viruses: lessons from bats and birds. *Trends Microbiol* **2013**, *21* (10), 544-555. DOI: 10.1016/j.tim.2013.05.005

(28) Yuan, M.; Huang, D.; Lee, C.-C. D.; Wu, N. C.; Jackson, A. M.; Zhu, X.; Liu, H.; Peng, L.; van Gils, M. J.; Sanders, R. W.; et al. Structural and functional ramifications of antigenic drift in recent SARS-CoV-2 variants. *Science* **2021**, *373* (6556), 818-823. DOI: doi:10.1126/science.abh1139.

(29) Hamre, D.; Procknow, J. J. J. P. o. t. s. f. e. b.; medicine. A new virus isolated from the human respiratory tract. **1966**, *121* (1), 190-193.

(30) Eguia, R. T.; Crawford, K. H. D.; Stevens-Ayers, T.; Kelnhofer-Millevolte, L.; Greninger, A. L.; Englund, J. A.; Boeckh, M. J.; Bloom, J. D. A human coronavirus evolves antigenically to escape antibody immunity. *PLoS Pathog* **2021**, *17* (4), e1009453. DOI: 10.1371/journal.ppat.1009453

(31) Xiang, J.; Su, J.; Lan, Q.; Zhao, W.; Zhou, Y.; Xu, Y.; Niu, J.; Xia, S.; Qi, Q.; Sidhu, S.; et al. Antigenic mapping reveals sites of vulnerability on α-HCoV spike protein. *Communications Biology* **2022**, *5* (1), 1179. DOI: 10.1038/s42003-022-04160-8.

(32) Marabotti, A.; Del Prete, E.; Scafuri, B.; Facchiano, A. Performance of Web tools for predicting changes in protein stability caused by mutations. *BMC Bioinformatics* **2021**, *22* (7), 345. DOI: 10.1186/s12859-021-04238-w.

(33) Laimer, J.; Hofer, H.; Fritz, M.; Wegenkittl, S.; Lackner, P. MAESTRO - multi agent stability prediction upon point mutations. *BMC Bioinformatics* **2015**, *16* (1), 116. DOI: 10.1186/s12859-015-0548-6.

(34) Savojardo, C.; Fariselli, P.; Martelli, P. L.; Casadio, R. INPS-MD: a web server to

doi:10.6342/NTU202301899

predict stability of protein variants from sequence and structure. *Bioinformatics (Oxford, England)* **2016**, *32* (16), 2542-2544. DOI: 10.1093/bioinformatics/btw192 %J Bioinformatics (acccessed 11/23/2022).

(35) Fariselli, P.; Martelli, P. L.; Savojardo, C.; Casadio, R. INPS: predicting the impact of non-synonymous variations on protein stability from sequence. *Bioinformatics (Oxford, England)* **2015**, *31* (17), 2816-2821. DOI: 10.1093/bioinformatics/btv291 %J Bioinformatics (acccessed 11/23/2022).

(36) Montanucci, L.; Capriotti, E.; Frank, Y.; Ben-Tal, N.; Fariselli, P. DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations. *BMC Bioinformatics* **2019**, *20* (14), 335. DOI: 10.1186/s12859-019-2923-1.

(37) Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A. P.; Rempfer, C.; Bordoli, L.; et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* **2018**, *46* (W1), W296-w303. DOI: 10.1093/nar/gky427

(38) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Meng, E. C.; Couch, G. S.; Croll, T. I.; Morris, J. H.; Ferrin, T. E. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Prot. Sci.* **2021**, *30* (1), 70-82. DOI: 10.1002/pro.3943

(39) Yang, T.-J.; Yu, P.-Y.; Chang, Y.-C.; Liang, K.-H.; Tso, H.-C.; Ho, M.-R.; Chen, W.-Y.; Lin, H.-T.; Wu, H.-C.; Hsu, S.-T. D. Effect of SARS-CoV-2 B.1.1.7 mutations on spike protein structure and function. *Nat. Struc. Mol. Biol.* **2021**, *28* (9), 731-739. DOI: 10.1038/s41594-021-00652-z.

(40) Zheng, S. Q.; Palovcak, E.; Armache, J.-P.; Verba, K. A.; Cheng, Y.; Agard, D. A. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nature Methods* **2017**, *14* (4), 331-332. DOI: 10.1038/nmeth.4193.

(41) Punjani, A.; Rubinstein, J. L.; Fleet, D. J.; Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods* **2017**, *14* (3), 290-296. DOI: 10.1038/nmeth.4169.

(42) Punjani, A.; Zhang, H.; Fleet, D. J. Non-uniform refinement: adaptive regularization improves single-particle cryo-EM reconstruction. *Nature Methods* **2020**, *17* (12), 1214-1221. DOI: 10.1038/s41592-020-00990-8.

(43) Liebschner, D.; Afonine, P. V.; Baker, M. L.; Bunkoczi, G.; Chen, V. B.; Croll, T. I.; Hintze, B.; Hung, L.-W.; Jain, S.; McCoy, A. J.; et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix.

doi:10.6342/NTU202301899

Acta Crystallogr. D **2019**, *75* (10), 861-877. DOI: doi:10.1107/S2059798319011471.

(44) Li, Z.; Tomlinson, A. C. A.; Wong, A. H. M.; Zhou, D.; Desforges, M.; Talbot, P. J.; Benlekbir, S.; Rubinstein, J. L.; Rini, J. M. The human coronavirus HCoV-229E S-protein structure and receptor binding. *eLife* **2019**, *8*, e51230. DOI: 10.7554/eLife.51230.

(45) Emsley, P.; Cowtan, K. J. A. c. s. D. b. c. Coot: model-building tools for molecular graphics. **2004**, *60* (12), 2126-2132.

(46) Croll, T. I. J. A. C. S. D. S. B. ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. **2018**, *74* (6), 519-530.

(47) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. J. J. o. c. c. CHARMM−GUI: a web−based graphical user interface for CHARMM. **2008**, *29* (11), 1859-1865.

(48) Kuo, C. W.; Yang, T. J.; Chien, Y. C.; Yu, P. Y.; Hsu, S. D.; Khoo, K. H. Distinct shifts in site-specific glycosylation pattern of SARS-CoV-2 spike proteins associated with arising mutations in the D614G and Alpha variants. *Glycobiology* **2022**, *32* (1), 60-72. DOI: 10.1093/glycob/cwab102

(49) Gecht, M.; von Bülow, S.; Penet, C.; Hummer, G.; Hanus, C.; Sikora, M. GlycoSHIELD: a versatile pipeline to assess glycan impact on protein structures. *BioRxiv* **2021**, 2021.2008. 2004.455134.

(50) Bekker, H.; Berendsen, H.; Dijkstra, E.; Achterop, S.; Vondrumen, R.; Vanderspoel, D.; Sijbers, A.; Keegstra, H.; Renardus, M. Gromacs-a parallel computer for molecular-dynamics simulations. In *4th international conference on computational physics (PC 92)*, 1993; World Scientific Publishing: pp 252-256.

(51) Grant, O. C.; Montgomery, D.; Ito, K.; Woods, R. J. Analysis of the SARS-CoV-2 spike protein glycan shield reveals implications for immune recognition. *Scientific Reports* **2020**, *10* (1), 14991. DOI: 10.1038/s41598-020-71748-7.

(52) Jo, S.; Lim, J. B.; Klauda, J. B.; Im, W. CHARMM-GUI Membrane Builder for Mixed Bilayers and Its Application to Yeast Membranes. *Biophysical Journal* **2009**, *97* (1), 50-58. DOI: 10.1016/j.bpj.2009.04.013 (acccessed 2023/05/06).

(53) Durrant, J. D.; Amaro, R. E. LipidWrapper: an algorithm for generating large-scale membrane models of arbitrary geometry. *PLoS Comput Biol* **2014**, *10* (7), e1003720. DOI: 10.1371/journal.pcbi.1003720

(54) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583-589. DOI:

doi:10.6342/NTU202301899

10.1038/s41586-021-03819-2.

(55) Nguyen, L.-T.; Schmidt, H. A.; von Haeseler, A.; Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **2014**, *32* (1), 268-274. DOI: 10.1093/molbev/msu300 %J Molecular Biology and Evolution (acccessed 1/2/2023).

(56) Madeira, F.; Pearce, M.; Tivey, A. R. N.; Basutkar, P.; Lee, J.; Edbali, O.; Madhusoodanan, N.; Kolesnikov, A.; Lopez, R. Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic acids research* **2022**, gkac240. DOI: 10.1093/nar/gkac240 PubMed.

(57) Gouet, P.; Courcelle, E.; Stuart, D. I.; M√©toz, F. ESPript: analysis of multiple sequence alignments in PostScript. *Bioinformatics (Oxford, England)* **1999**, *15* (4), 305-308. DOI: 10.1093/bioinformatics/15.4.305 %J Bioinformatics (acccessed 11/28/2022).

(58) Zhao, Z.; Zhou, J.; Tian, M.; Huang, M.; Liu, S.; Xie, Y.; Han, P.; Bai, C.; Han, P.; Zheng, A.; et al. Omicron SARS-CoV-2 mutations stabilize spike up-RBD conformation and lead to a non-RBM-binding monoclonal antibody escape. *Nat. Comm.* **2022**, *13* (1), 4958. DOI: 10.1038/s41467-022-32665-7.

(59) McCallum, M.; Czudnochowski, N.; Rosen, L. E.; Zepeda, S. K.; Bowen, J. E.; Walls, A. C.; Hauser, K.; Joshi, A.; Stewart, C.; Dillen, J. R.; et al. Structural basis of SARS-CoV-2 Omicron immune evasion and receptor engagement. **2022**, *375* (6583), 864-868. DOI: doi:10.1126/science.abn8652.

(60) Dehouck, Y.; Grosfils, A.; Folch, B.; Gilis, D.; Bogaerts, P.; Rooman, M. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics (Oxford, England)* **2009**, *25* (19), 2537-2543. DOI: 10.1093/bioinformatics/btp445 %J Bioinformatics (acccessed 11/25/2022).

(61) Rodrigues, C. H.; Pires, D. E.; Ascher, D. B. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Research* **2018**, *46* (W1), W350-W355. DOI: 10.1093/nar/gky300 %J Nucleic Acids Research (acccessed 11/23/2022).

(62) Laimer, J.; Hiebl-Flach, J.; Lengauer, D.; Lackner, P. MAESTROweb: a web server for structure-based protein stability prediction. *Bioinformatics (Oxford, England)* **2016**, *32* (9), 1414-1416. DOI: 10.1093/bioinformatics/btv769 %J Bioinformatics (acccessed 11/23/2022).

(63) Khan, S.; Vihinen, M. J. H. m. Performance of protein stability predictors. **2010**, *31*

doi:10.6342/NTU202301899

(6), 675-684.

(64) Raju, T. S.; Briggs, J. B.; Borge, S. M.; Jones, A. J. S. Species-specific variation in glycosylation of IgG: evidence for the species-specific sialylation and branch-specific galactosylation and importance for engineering recombinant glycoprotein therapeutics. *Glycobiology* **2000**, *10* (5), 477-486. DOI: 10.1093/glycob/10.5.477 (acccessed 2/6/2023).

(65) Xu, C.; Ng, D. T. W. Glycosylation-directed quality control of protein folding. *Nature Reviews Molecular Cell Biology* **2015**, *16* (12), 742-752. DOI: 10.1038/nrm4073.

(66) Tao, Y.; Strelkov, S. V.; Mesyanzhinov, V. V.; Rossmann, M. G. J. S. Structure of bacteriophage T4 fibritin: a segmented coiled coil and the role of the C-terminal domain. **1997**, *5* (6), 789-798.

(67) Pallesen, J.; Wang, N.; Corbett, K. S.; Wrapp, D.; Kirchdoerfer, R. N.; Turner, H. L.; Cottrell, C. A.; Becker, M. M.; Wang, L.; Shi, W.; et al. Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. **2017**, *114* (35), E7348-E7357. DOI: doi:10.1073/pnas.1707304114.

(68) Kirchdoerfer, R. N.; Wang, N.; Pallesen, J.; Wrapp, D.; Turner, H. L.; Cottrell, C. A.; Corbett, K. S.; Graham, B. S.; McLellan, J. S.; Ward, A. B. Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. *Scientific Reports* **2018**, *8* (1), 15701. DOI: 10.1038/s41598-018-34171-7.

(69) Walls, A. C.; Tortorici, M. A.; Frenz, B.; Snijder, J.; Li, W.; Rey, F. A.; DiMaio, F.; Bosch, B.-J.; Veesler, D. Glycan shield and epitope masking of a coronavirus spike protein observed by cryo-electron microscopy. *Nat. Struc. Mol. Biol.* **2016**, *23* (10), 899-905. DOI: 10.1038/nsmb.3293.

(70) Watanabe, Y.; Allen, J. D.; Wrapp, D.; McLellan, J. S.; Crispin, M. Site-specific glycan analysis of the SARS-CoV-2 spike. *Science* **2020**, *369* (6501), 330. DOI: 10.1126/science.abb9983.

(71) Watanabe, Y.; Berndsen, Z. T.; Raghwani, J.; Seabright, G. E.; Allen, J. D.; Pybus, O. G.; McLellan, J. S.; Wilson, I. A.; Bowden, T. A.; Ward, A. B. Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. *Nat. Comm.* **2020**, *11* (1), 2688.

(72) Zhang, Y.; Zhao, W.; Mao, Y.; Chen, Y.; Wang, S.; Zhong, Y.; Su, T.; Gong, M.; Du, D.; Lu, X. J. M.; et al. Site-specific N-glycosylation characterization of recombinant SARS-CoV-2 spike proteins. **2021**, 100058.

(73) Behrens, A.-J.; Crispin, M. J. C. o. i. s. b. Structural principles controlling HIV

doi:10.6342/NTU202301899

envelope glycosylation. **2017**, *44*, 125-133.

(74) Watanabe, Y.; Berndsen, Z. T.; Raghwani, J.; Seabright, G. E.; Allen, J. D.; Pybus, O. G.; McLellan, J. S.; Wilson, I. A.; Bowden, T. A.; Ward, A. B.; et al. Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. *Nat Commun* **2020**, *11* (1), 2688. DOI: 10.1038/s41467-020-16567-0.

(75) Pritchard, L. K.; Spencer, D. I. R.; Royle, L.; Bonomelli, C.; Seabright, G. E.; Behrens, A.-J.; Kulp, D. W.; Menis, S.; Krumm, S. A.; Dunlop, D. C.; et al. Glycan clustering stabilizes the mannose patch of HIV-1 and preserves vulnerability to broadly neutralizing antibodies. *Nat. Comm.* **2015**, *6* (1), 7479. DOI: 10.1038/ncomms8479.

(76) Roushan, A.; Wilson, G. M.; Kletter, D.; Sen, K. I.; Tang, W.; Kil, Y. J.; Carlson, E.; Bern, M. Peak Filtering, Peak Annotation, and Wildcard Search for Glycoproteomics. *Mol Cell Proteomics* **2021**, *20*, 100011. DOI: 10.1074/mcp.RA120.002260

(77) Fang, P.; Ji, Y.; Oellerich, T.; Urlaub, H.; Pan, K. T. Strategies for Proteome-Wide Quantification of Glycosylation Macro- and Micro-Heterogeneity. *Int J Mol Sci* **2022**, *23* (3). DOI: 10.3390/ijms23031609

(78) Wrapp, D.; McLellan, J. S. The 3.1-Angstrom Cryo-electron Microscopy Structure of the Porcine Epidemic Diarrhea Virus Spike Protein in the Prefusion Conformation. *J Virol* **2019**, *93* (23). DOI: 10.1128/jvi.00923-19

(79) Huang, C.-Y.; Draczkowski, P.; Wang, Y.-S.; Chang, C.-Y.; Chien, Y.-C.; Cheng, Y.-H.; Wu, Y.-M.; Wang, C.-H.; Chang, Y.-C.; Chang, Y.-C.; et al. In situ structure and dynamics of an alphacoronavirus spike protein by cryo-ET and cryo-EM. *Nat. Comm.* **2022**, *13* (1), 4877. DOI: 10.1038/s41467-022-32588-3.

(80) Casalino, L.; Gaieb, Z.; Goldsmith, J. A.; Hjorth, C. K.; Dommer, A. C.; Harbison, A. M.; Fogarty, C. A.; Barros, E. P.; Taylor, B. C.; McLellan, J. S.; et al. Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Central Science* **2020**, *6* (10), 1722-1734. DOI: 10.1021/acscentsci.0c01056.

(81) Choi, Y. K.; Cao, Y.; Frank, M.; Woo, H.; Park, S.-J.; Yeom, M. S.; Croll, T. I.; Seok, C.; Im, W. Structure, Dynamics, Receptor Binding, and Antibody Binding of the Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein in a Viral Membrane. *Journal of Chemical Theory and Computation* **2021**, *17* (4), 2479-2487. DOI: 10.1021/acs.jctc.0c01144.

(82) Woo, H.; Park, S.-J.; Choi, Y. K.; Park, T.; Tanveer, M.; Cao, Y.; Kern, N. R.; Lee, J.; Yeom, M. S.; Croll, T. I.; et al. Developing a Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein Model in a Viral Membrane. *The Journal of Physical Chemistry B* **2020**, *124* (33), 7128-7137. DOI: 10.1021/acs.jpcb.0c04553.

doi:10.6342/NTU202301899

(83) Gecht, M.; von Bülow, S.; Penet, C.; Hummer, G.; Hanus, C.; Sikora, M. J. b. GlycoSHIELD: a versatile pipeline to assess glycan impact on protein structures. **2021**.

(84) Wong, A. H. M.; Tomlinson, A. C. A.; Zhou, D.; Satkunarajah, M.; Chen, K.; Sharon, C.; Desforges, M.; Talbot, P. J.; Rini, J. M. Receptor-binding loops in alphacoronavirus adaptation and evolution. *Nat. Comm.* **2017**, *8* (1), 1735. DOI: 10.1038/s41467-017-01706-x.

(85) Echave, J.; Wilke, C. O. Biophysical Models of Protein Evolution: Understanding the Patterns of Evolutionary Sequence Divergence. *Annual review of biophysics* **2017**, *46*, 85-103. DOI: 10.1146/annurev-biophys-070816-033819

(86) Wang, Q.; Meng, F.; Xie, Y.; Wang, W.; Meng, Y.; Li, L.; Liu, T.; Qi, J.; Ni, X.; Zheng, S. In Silico Discovery of Small Molecule Modulators Targeting the Achilles' Heel of SARS-CoV-2 Spike Protein. *ACS Cen. Sci.* **2023**, *9* (2), 252-265.

(87) Lv, Z.; Deng, Y.-Q.; Ye, Q.; Cao, L.; Sun, C.-Y.; Fan, C.; Huang, W.; Sun, S.; Sun, Y.; Zhu, L. Structural basis for neutralization of SARS-CoV-2 and SARS-CoV by a potent therapeutic antibody. *Science* **2020**, *369* (6510), 1505-1509.

(88) Wang, P.; Casner, R. G.; Nair, M. S.; Yu, J.; Guo, Y.; Wang, M.; Chan, J. F.-W.; Cerutti, G.; Iketani, S.; Liu, L. A monoclonal antibody that neutralizes SARS-CoV-2 variants, SARS-CoV, and other sarbecoviruses. *Emerging Microbes & Infections* **2022**, *11* (1), 147-157.

(89) Adams, P. D.; Grosse-Kunstleve, R. W.; Hung, L.-W.; Ioerger, T. R.; McCoy, A. J.; Moriarty, N. W.; Read, R. J.; Sacchettini, J. C.; Sauter, N. K.; Terwilliger, T. C. PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **2002**, *58* (11), 1948-1954. DOI: doi:10.1107/S0907444902016657.

(90) Fan, X.; Cao, D.; Kong, L.; Zhang, X. Cryo-EM analysis of the post-fusion structure of the SARS-CoV spike glycoprotein. *Nat. Comm.* **2020**, *11* (1), 3618.

(91) Tai, L.; Zhu, G.; Yang, M.; Cao, L.; Xing, X.; Yin, G.; Chan, C.; Qin, C.; Rao, Z.; Wang, X. Nanometer-resolution in situ structure of the SARS-CoV-2 postfusion spike protein. *Proc. Natl. Acad. Sci.* U. S. A. **2021**, *118* (48), e2112703118.

(92) Arruda, H. R. S.; Lima, T. M.; Alvim, R. G. F.; Victorio, F. B. A.; Abreu, D. P. B.; Marsili, F. F.; Cruz, K. D.; Marques, M. A.; Sosa-Acosta, P.; Quinones-Vega, M.; et al. Conformational stability of SARS-CoV-2 glycoprotein spike variants. *iScience* **2023**, *26* (1), 105696. DOI: 10.1016/j.isci.2022.105696

(93) Bhardwaj, A.; Walker-Kopp, N.; Wilkens, S.; Cingolani, G. Foldon-guided self-

doi:10.6342/NTU202301899

assembly of ultra-stable protein fibers. *Prot. Sci.* **2008**, *17* (9), 1475-1485. DOI: 10.1110/ps.036111.108

(94) Frank, S.; Kammerer, R. A.; Mechling, D.; Schulthess, T.; Landwehr, R.; Bann, J.; Guo, Y.; Lustig, A.; Bächinger, H. P.; Engel, J. Stabilization of short collagen-like triple helices by protein engineering. *J Mol Biol* **2001**, *308* (5), 1081-1089. DOI: 10.1006/jmbi.2001.4644

(95) Güthe, S.; Kapinos, L.; Möglich, A.; Meier, S.; Grzesiek, S.; Kiefhaber, T. Very Fast Folding and Association of a Trimerization Domain from Bacteriophage T4 Fibritin. *J. Mol. Biol.* **2004**, *337* (4), 905-915. DOI: https://doi.org/10.1016/j.jmb.2004.02.020.

(96) Behrens, A.-J.; Harvey, D. J.; Milne, E.; Cupo, A.; Kumar, A.; Zitzmann, N.; Struwe, W. B.; Moore, J. P.; Crispin, M. Molecular Architecture of the Cleavage-Dependent Mannose Patch on a Soluble HIV-1 Envelope Glycoprotein Trimer. **2017**, *91* (2), e01894-01816. DOI: doi:10.1128/JVI.01894-16.

(97) Pritchard, Laura K.; Vasiljevic, S.; Ozorowski, G.; Seabright, Gemma E.; Cupo, A.; Ringe, R.; Kim, Helen J.; Sanders, Rogier W.; Doores, Katie J.; Burton, Dennis R.; et al. Structural Constraints Determine the Glycosylation of HIV-1 Envelope Trimers. *Cell Reports* **2015**, *11* (10), 1604-1613. DOI: https://doi.org/10.1016/j.celrep.2015.05.017.

(98) Sanda, M.; Morrison, L.; Goldman, R. N- and O-Glycosylation of the SARS-CoV-2 Spike Protein. *Analytical Chemistry* **2021**, *93* (4), 2003-2009. DOI: 10.1021/acs.analchem.0c03173.

(99) Punjani, A.; Fleet, D. J. 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *J. Struc. Biol.* **2021**, *213* (2), 107702. DOI: https://doi.org/10.1016/j.jsb.2021.107702.

(100) Punjani, A.; Fleet, D. J. 3DFlex: determining structure and motion of flexible proteins from cryo-EM. *Nature Methods* **2023**, *20* (6), 860-870. DOI: 10.1038/s41592-023-01853-8.

(101) Walls, A. C.; Park, Y. J.; Tortorici, M. A.; Wall, A.; McGuire, A. T.; Veesler, D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **2020**, *181* (2), 281-292.e286. DOI: 10.1016/j.cell.2020.02.058

(102) Xu, C.; Wang, Y.; Liu, C.; Zhang, C.; Han, W.; Hong, X.; Wang, Y.; Hong, Q.; Wang, S.; Zhao, Q.; et al. Conformational dynamics of SARS-CoV-2 trimeric spike glycoprotein in complex with receptor ACE2 revealed by cryo-EM. *Sci Adv* **2021**, *7* (1). DOI: 10.1126/sciadv.abe5575

doi:10.6342/NTU202301899

(103) Chmielewski, D.; Wilson, E.; Pintilie, G.; Zhao, P.; Chen, M.; Schmid, M.; Simmons, G.; Wells, L.; Jin, J.; Singharoy, A.; et al. Integrated analyses reveal a hinge glycan regulates coronavirus spike tilting and virus infectivity. *Research square* **2023**. DOI: 10.21203/rs.3.rs-2553619/v1

(104) Yang, T.-J.; Chang, Y.-C.; Ko, T.-P.; Draczkowski, P.; Chien, Y.-C.; Chang, Y.-C.; Wu, K.-P.; Khoo, K.-H.; Chang, H.-W.; Hsu, S.-T. D. Cryo-EM analysis of a feline coronavirus spike protein reveals a unique structure and camouflaging glycans. *Proc. Natl. Acad. Sci.* U. S. A. **2020**, *117* (3), 1438. DOI: 10.1073/pnas.1908898117.

(105) Doores, K. J. The HIV glycan shield as a target for broadly neutralizing antibodies. **2015**, *282* (24), 4679-4691. DOI: https://doi.org/10.1111/febs.13530.

(106) Harvey, W. T.; Carabelli, A. M.; Jackson, B.; Gupta, R. K.; Thomson, E. C.; Harrison, E. M.; Ludden, C.; Reeve, R.; Rambaut, A.; Consortium, C.-G. U. SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology* **2021**, *19* (7), 409-424.

(107) Conlan, M.; Tan, H.-X.; Esterbauer, R.; Kelly, A.; Kent, S. J.; Wheatley, A. K.; Lee, W. S. Monoclonal antibodies against human coronavirus NL63 spike. *bioRxiv* **2023**, 2023.2003.2012.532265. DOI: 10.1101/2023.03.12.532265.

(108) Shi, J.; Shi, Y.; Xiu, R.; Wang, G.; Liang, R.; Jiao, Y.; Shen, Z.; Zhu, C.; Peng, G. Identification of a Novel Neutralizing Epitope on the N-Terminal Domain of the Human Coronavirus 229E Spike Protein. **2022**, *96* (4), e01955-01921. DOI: doi:10.1128/jvi.01955-21.

(109) Pronker, M. F.; Creutznacher, R.; Drulyte, I.; Hulswit, R. J. G.; Li, Z.; Kuppeveld, F. J. M. v.; Snijder, J.; Lang, Y.; Bosch, B.-J.; Boons, G.-J.; et al. Sialoglycan binding triggers spike opening in a human coronavirus. *bioRxiv* **2023**, 2023.2004.2020.536837. DOI: 10.1101/2023.04.20.536837.

(110) Zhan, W.; Tian, X.; Zhang, X.; Xing, S.; Song, W.; Liu, Q.; Hao, A.; Hu, Y.; Zhang, M.; Ying, T. Structural study of SARS-CoV-2 antibodies identifies a broad-spectrum antibody that neutralizes the omicron variant by disassembling the spike trimer. *Journal of virology* **2022**, *96* (16), e00480-00422.

(111) Zhou, D.; Duyvesteyn, H. M.; Chen, C.-P.; Huang, C.-G.; Chen, T.-H.; Shih, S.-R.; Lin, Y.-C.; Cheng, C.-Y.; Cheng, S.-H.; Huang, Y.-C. Structural basis for the neutralization of SARS-CoV-2 by an antibody from a convalescent patient. *Nature structural & molecular biology* **2020**, *27* (10), 950-958.

(112) Yuan, M.; Wu, N. C.; Zhu, X.; Lee, C.-C. D.; So, R. T.; Lv, H.; Mok, C. K.; Wilson, I. A. A highly conserved cryptic epitope in the receptor binding domains of SARS-

doi:10.6342/NTU202301899

CoV-2 and SARS-CoV. *Science* **2020**, *368* (6491), 630-633.

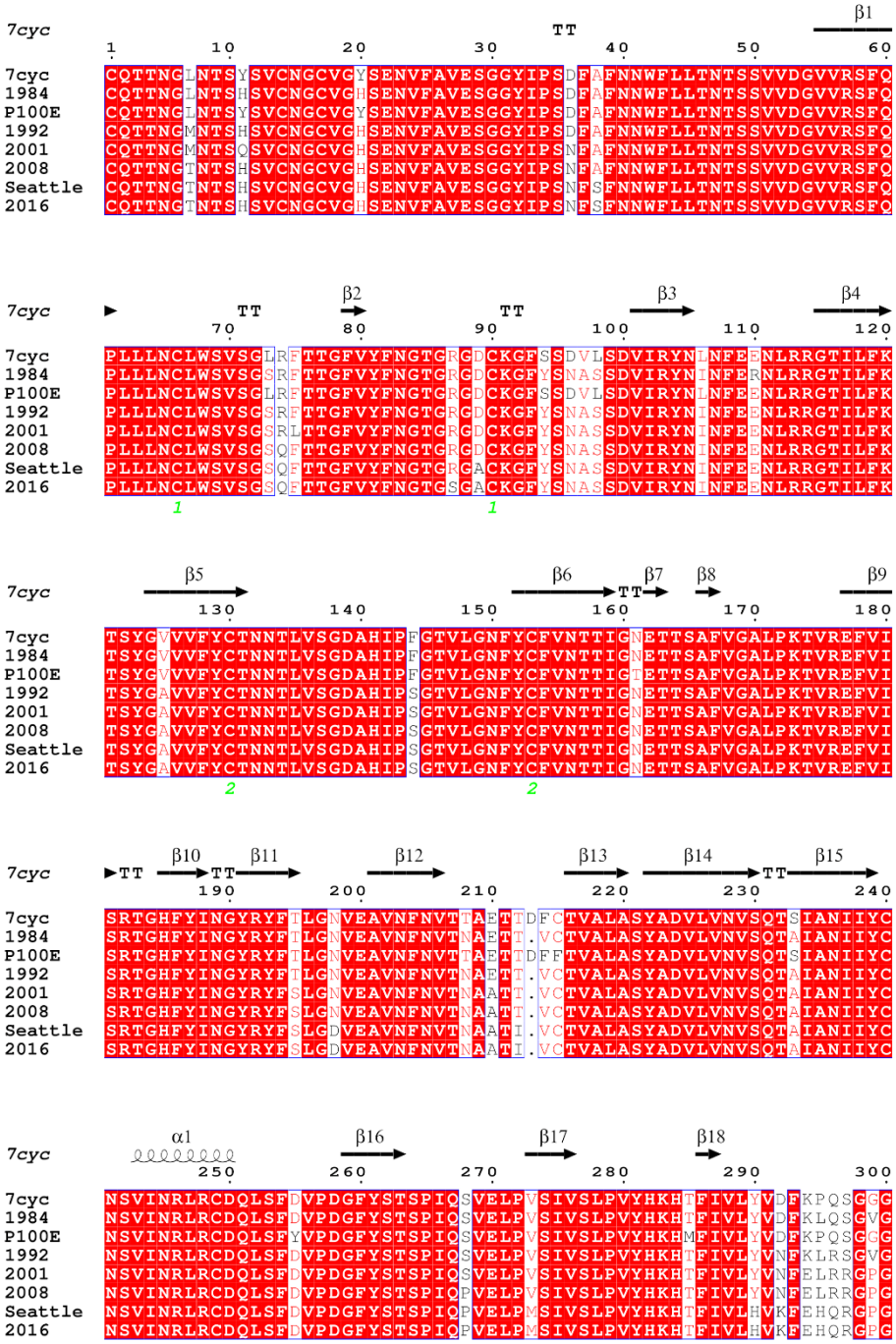
(113) Piccoli, L.; Park, Y.-J.; Tortorici, M. A.; Czudnochowski, N.; Walls, A. C.; Beltramello, M.; Silacci-Fregni, C.; Pinto, D.; Rosen, L. E.; Bowen, J. E. Mapping neutralizing and immunodominant sites on the SARS-CoV-2 spike receptor-binding domain by structure-guided high-resolution serology. *Cell* **2020**, *183* (4), 1024-1042. e1021.

(114) Asarnow, D.; Palovcak, E.; Cheng, Y. asarnow/pyem: UCSF pyem v0. 5. Zenodo. 2019.

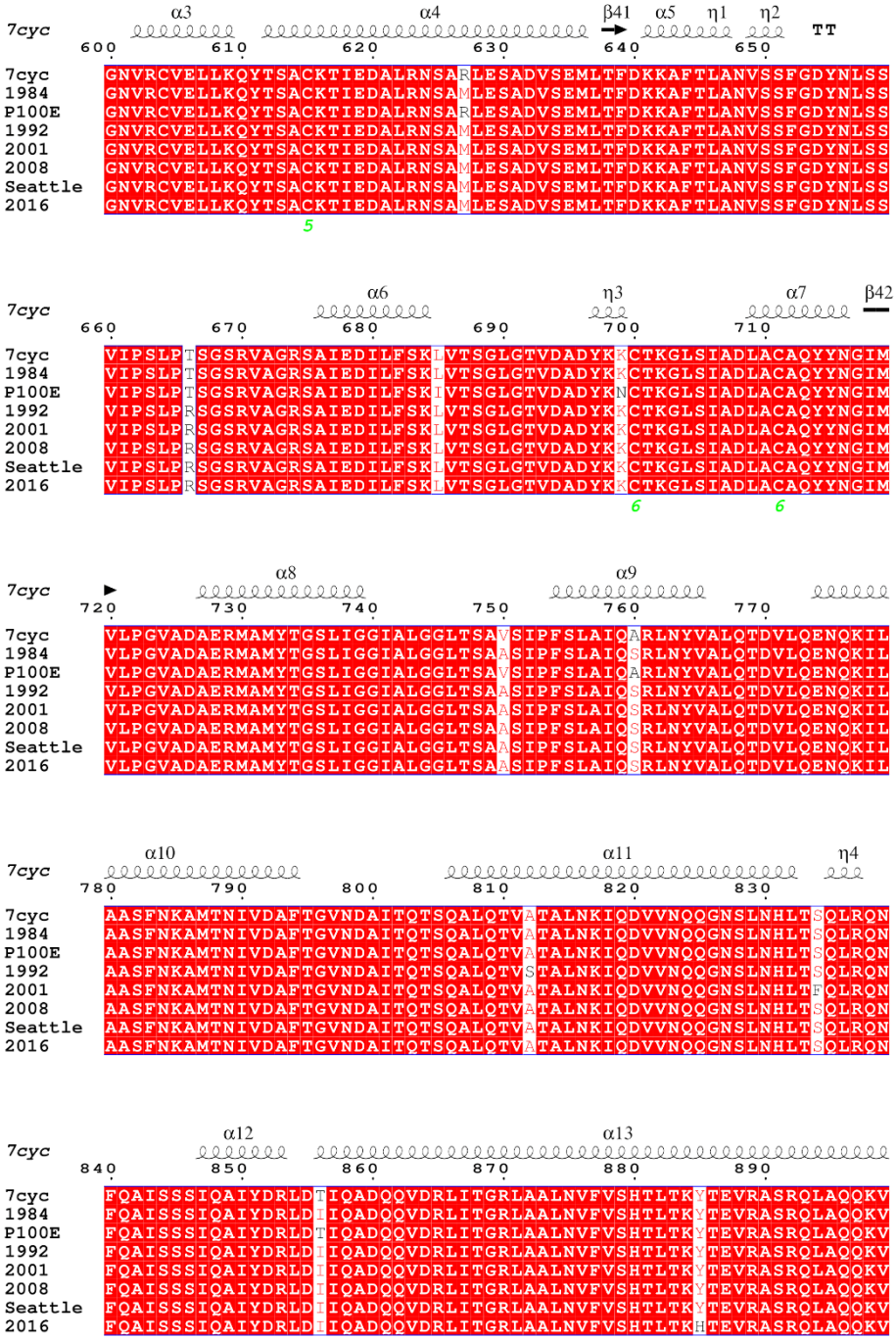
doi:10.6342/NTU202301899

APPENDIX

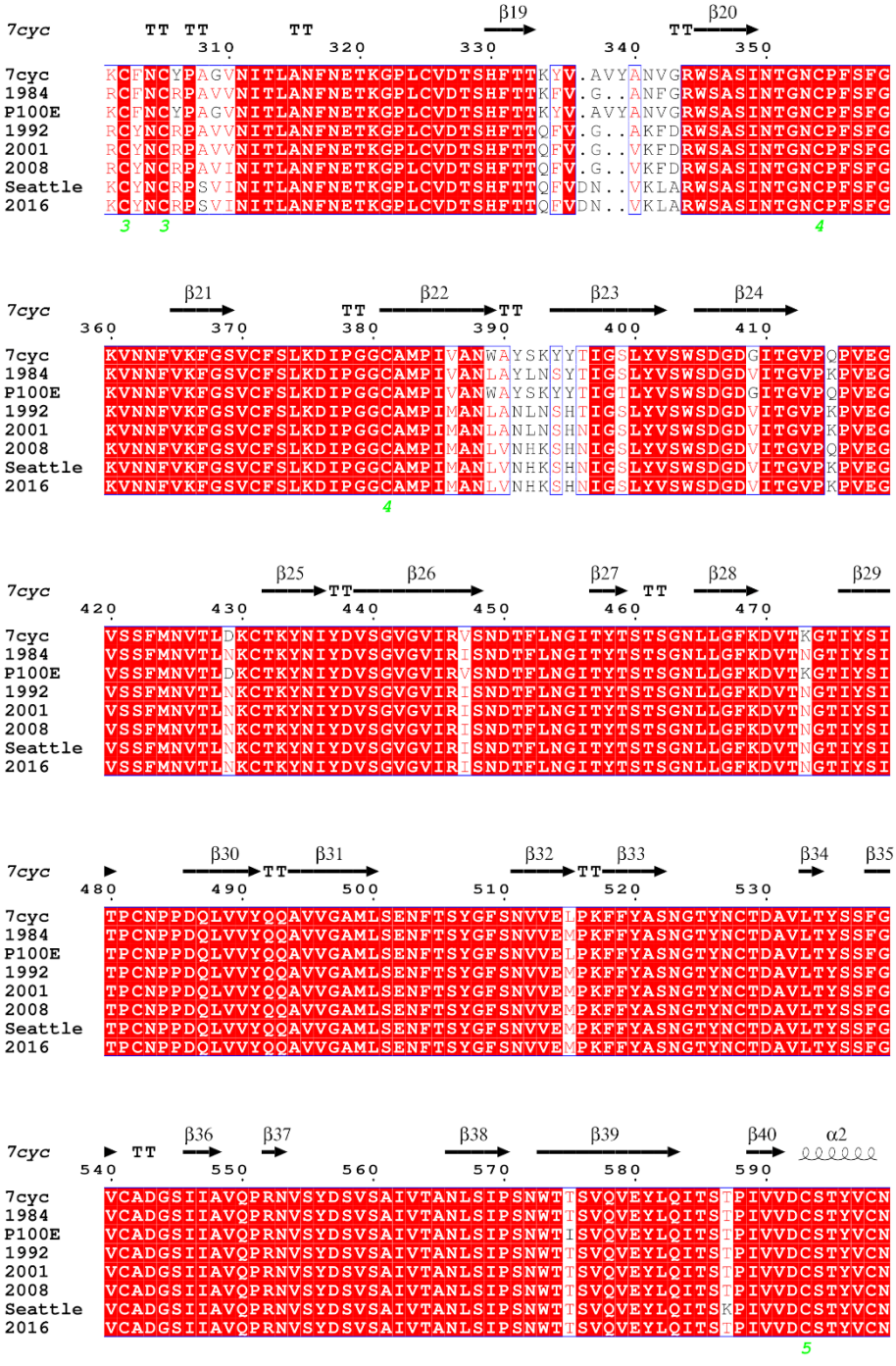
doi:10.6342/NTU202301899



doi:10.6342/NTU202301899



doi:10.6342/NTU202301899



doi:10.6342/NTU202301899

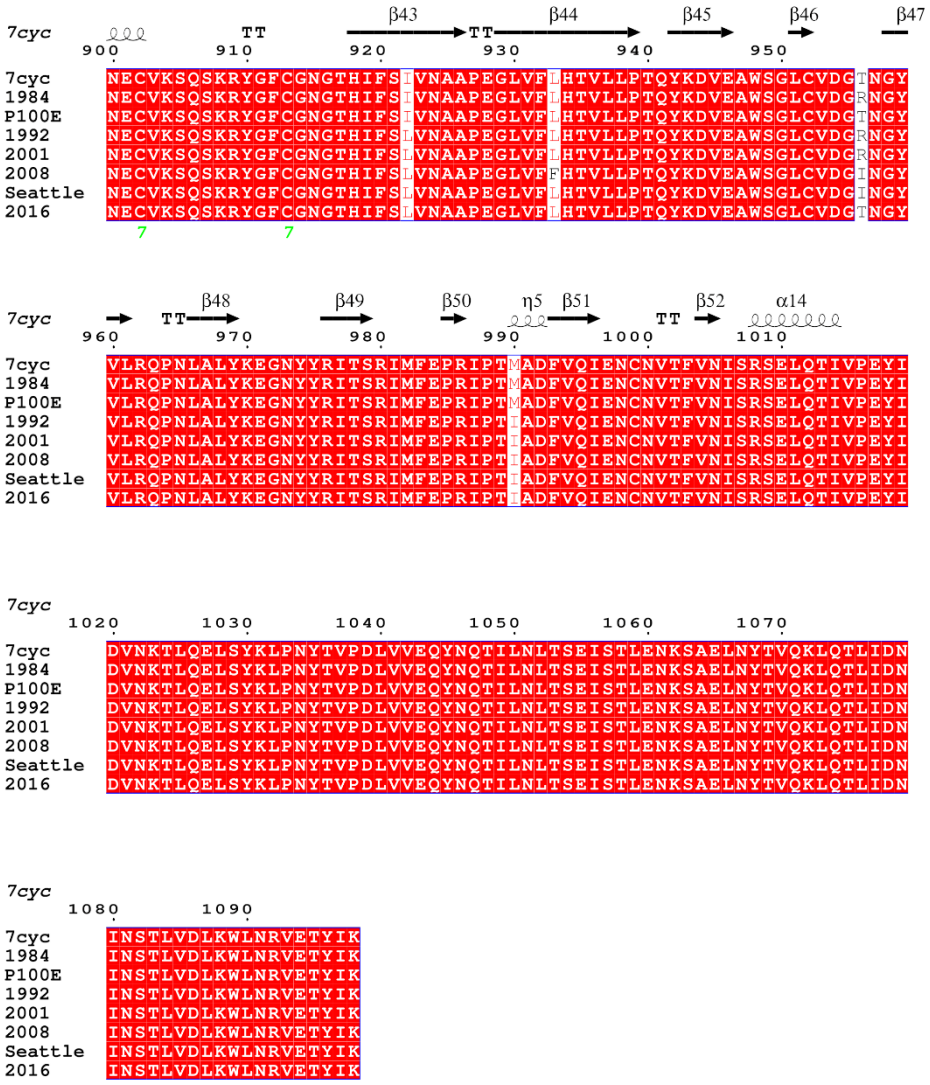


Figure A1. Multiple sequence alignment of selected HCoV-229E strains using ESPript3.

doi:10.6342/NTU202301899

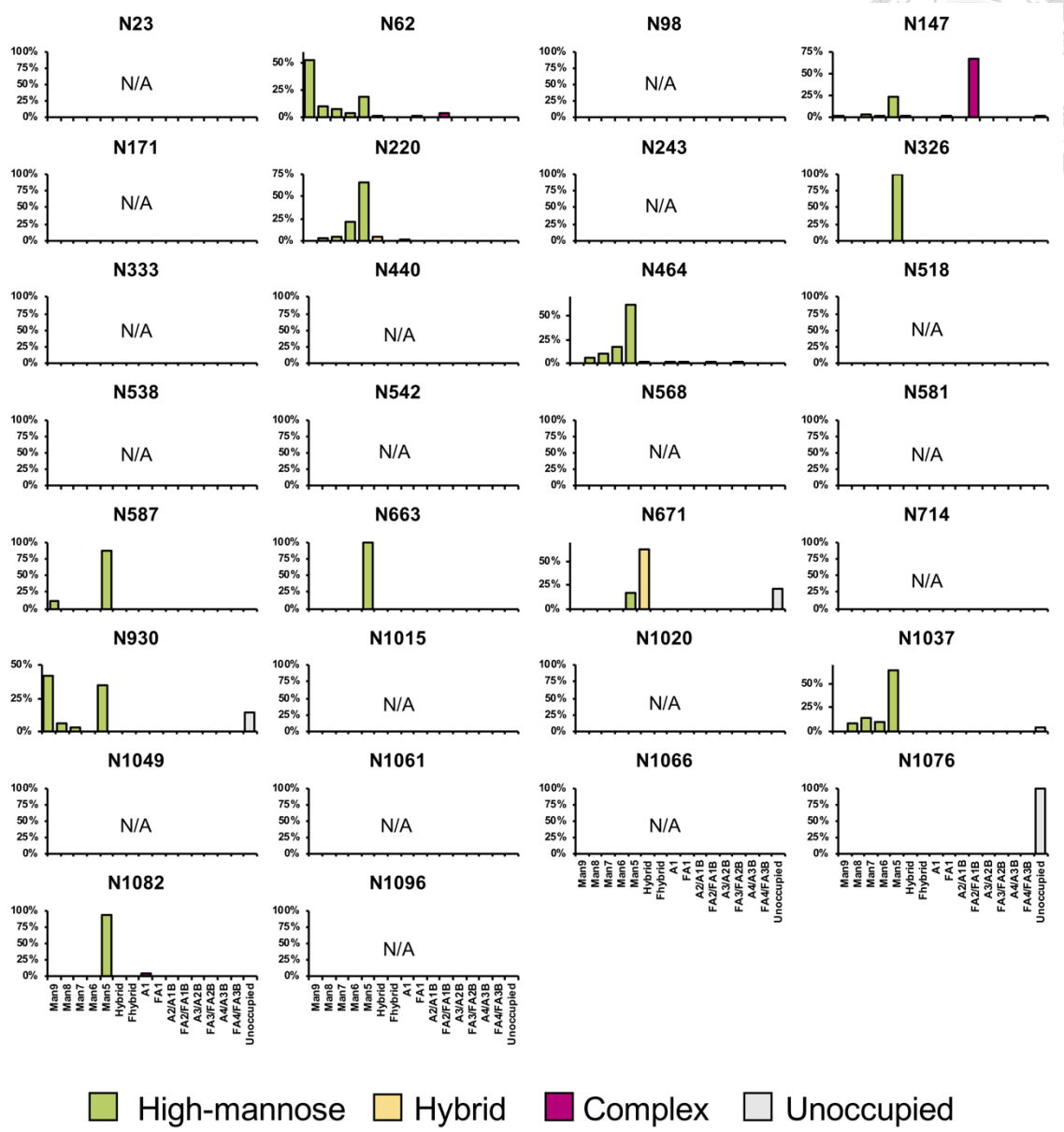


Figure A2. Quantification of the site-specific N-linked glycosylation analysis of the HCoV-229E P100E strain S protein (Insol-TC #1).

doi:10.6342/NTU202301899

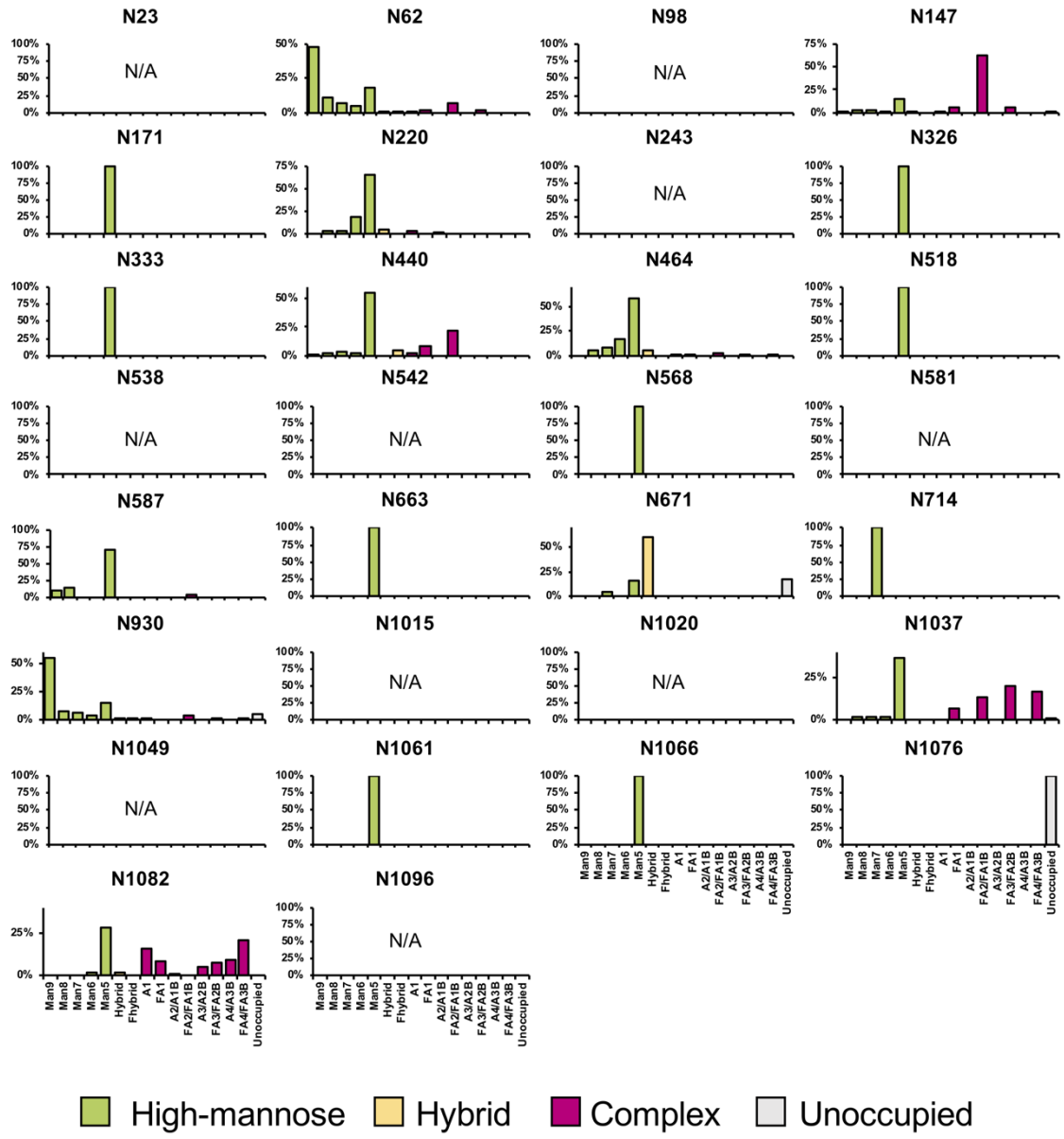


Figure A3. Quantification of the site-specific N-linked glycosylation analysis of the HCoV-229E P100E strain S protein (Insol-TC #2).

doi:10.6342/NTU202301899

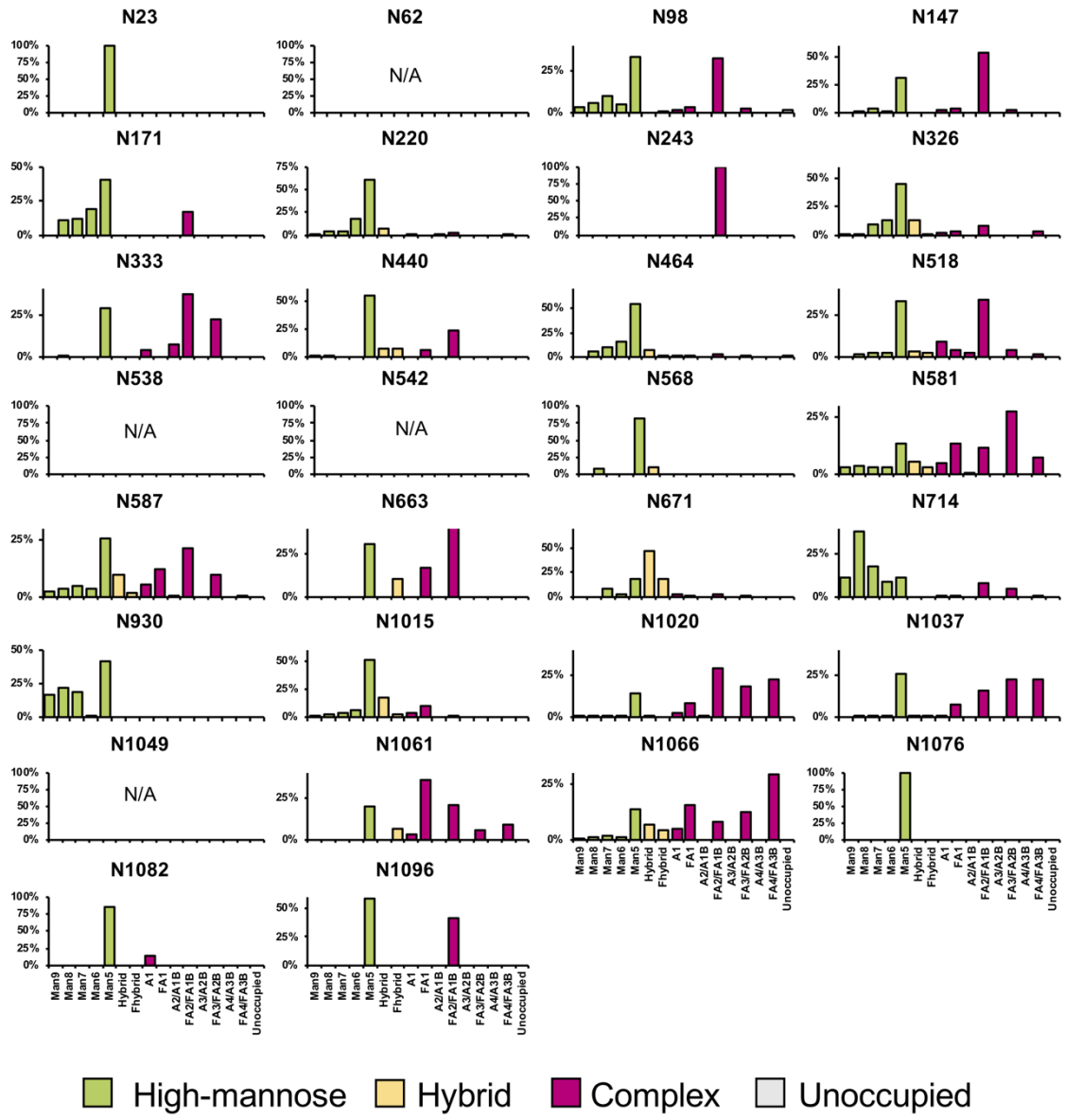


Figure A4. Quantification of the site-specific N-linked glycosylation analysis of the HCoV-229E P100E strain S protein (Insol-aLP).

doi:10.6342/NTU202301899

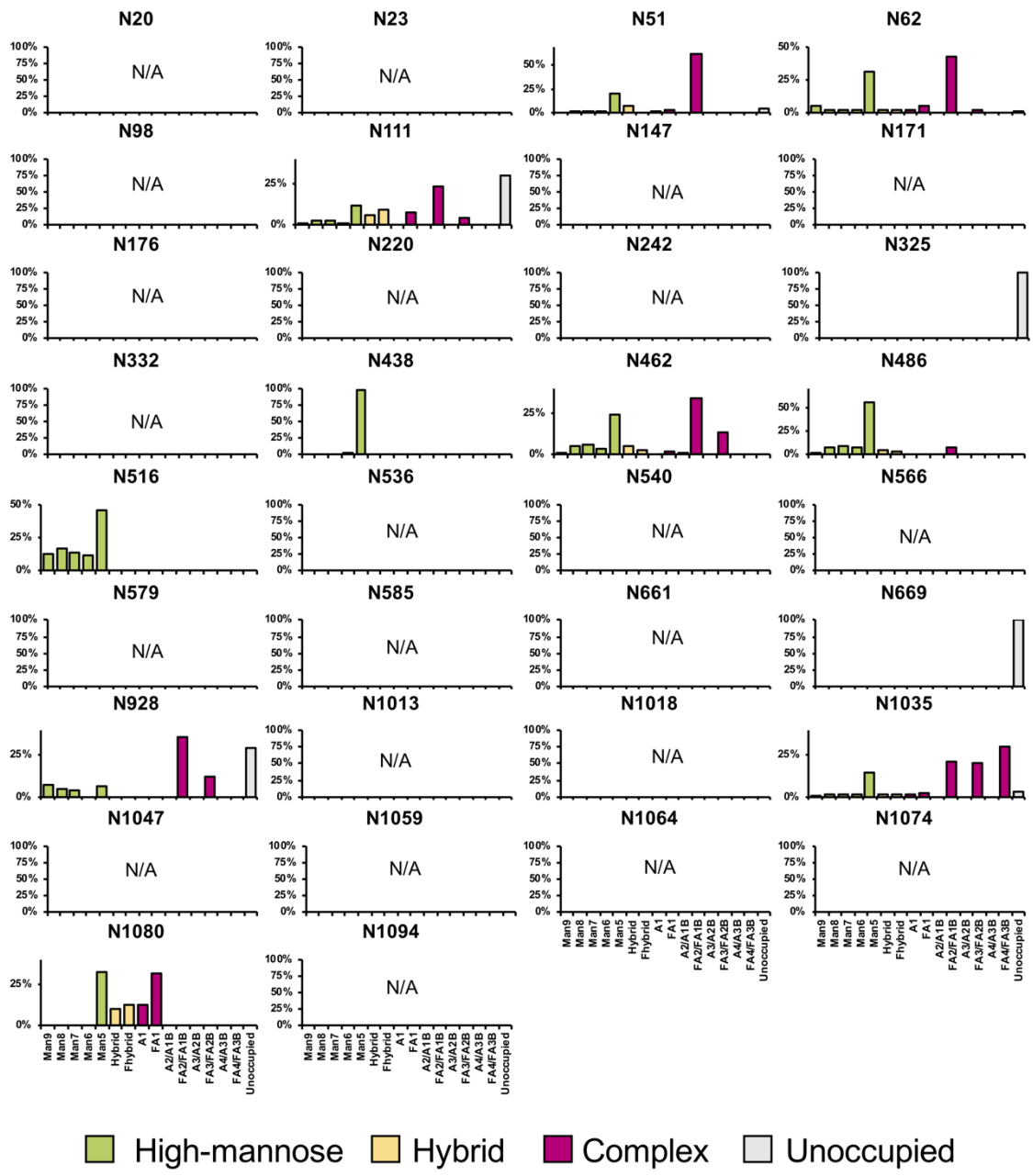


Figure A5. Quantification of the site-specific N-linked glycosylation analysis of the HCoV-229E Seattle strain S protein (Ingel-TC).

doi:10.6342/NTU202301899

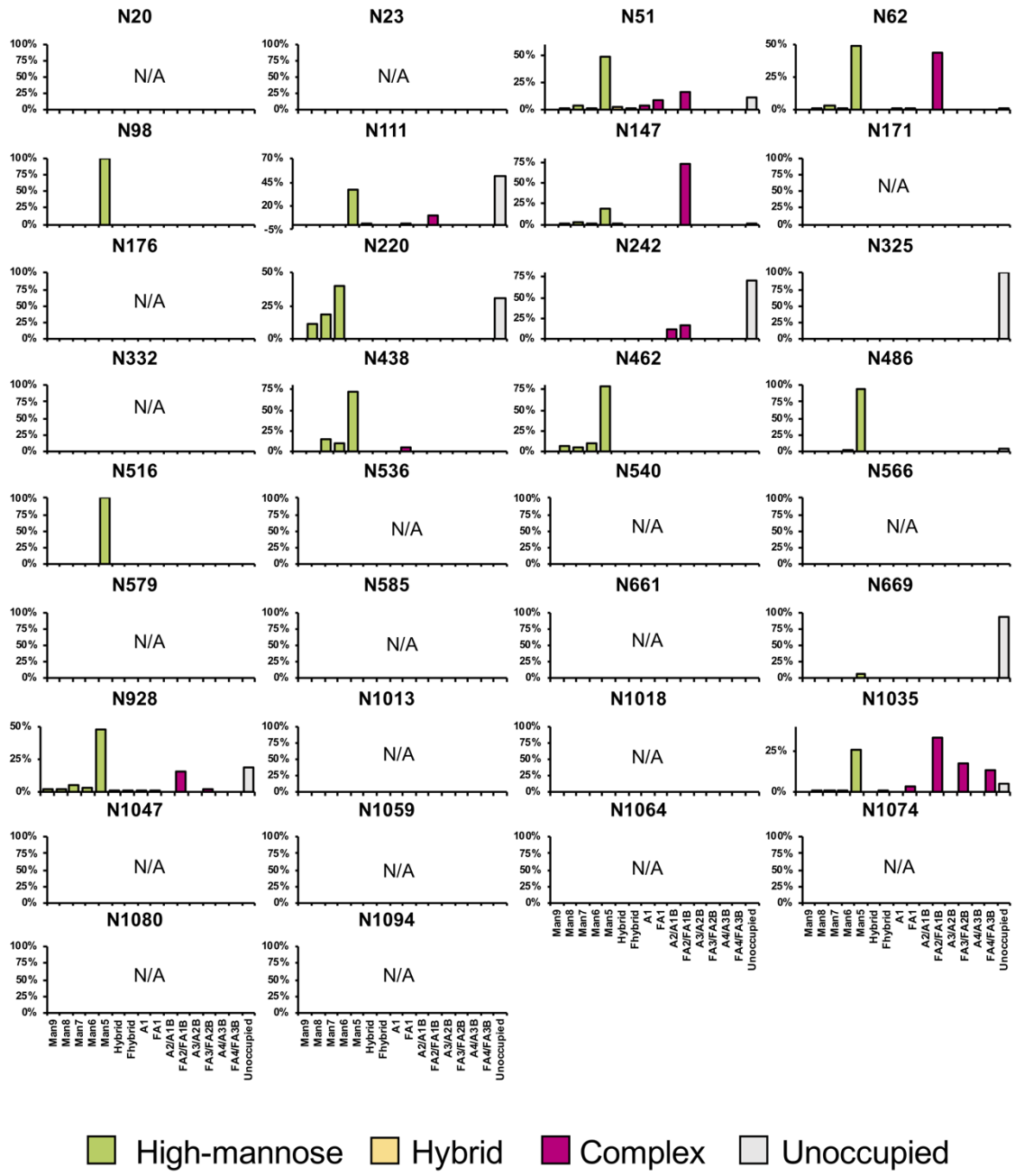


Figure A6. Quantification of the site-specific N-linked glycosylation analysis of the HCoV-229E Seattle strain S protein (Insol-TC).

doi:10.6342/NTU202301899

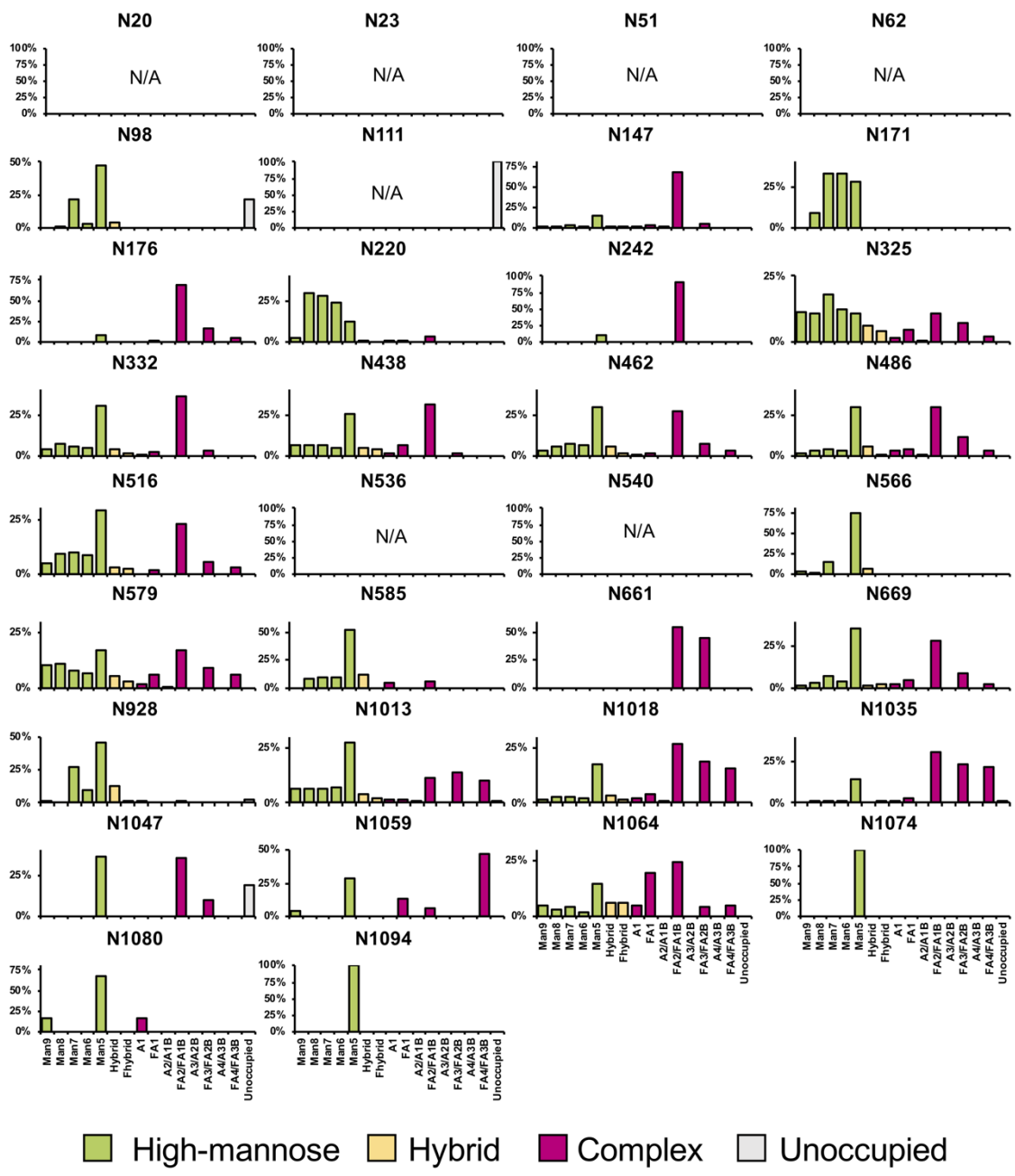


Figure A7. Quantification of the site-specific N-linked glycosylation analysis of the HCoV-229E Seattle strain S protein (Insol-aLP).

doi:10.6342/NTU202301899

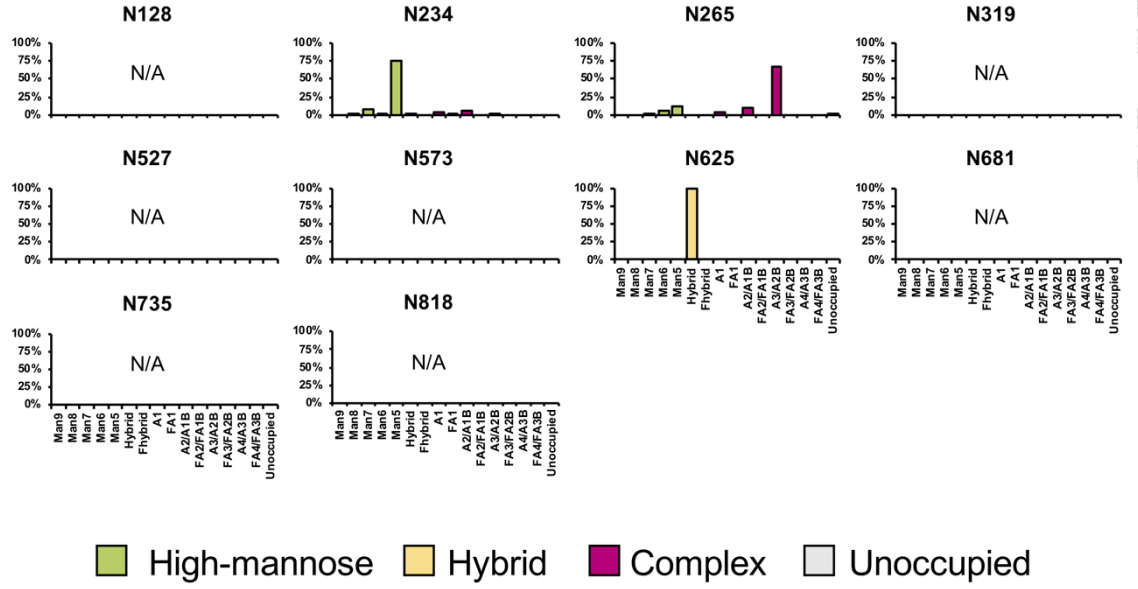


Figure A8. Quantification of the site-specific N-linked glycosylation analysis of the human aminopeptidase N (hAPN) ectodomain (Insol-TC #1).

doi:10.6342/NTU202301899

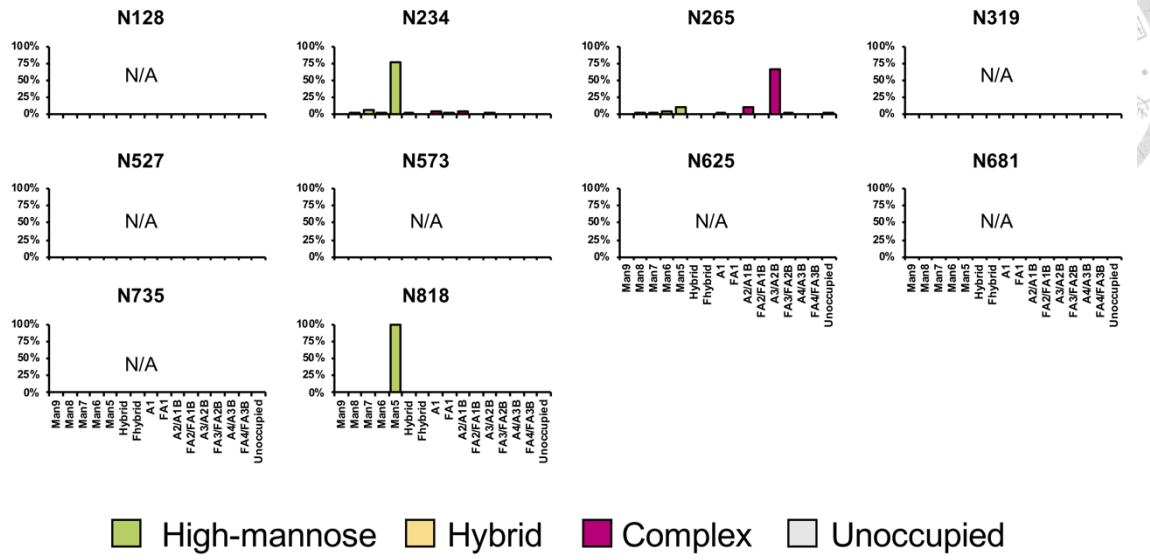


Figure A9. Quantification of the site-specific N-linked glycosylation analysis of the human aminopeptidase N (hAPN) ectodomain (Insol-TC #2).

doi:10.6342/NTU202301899

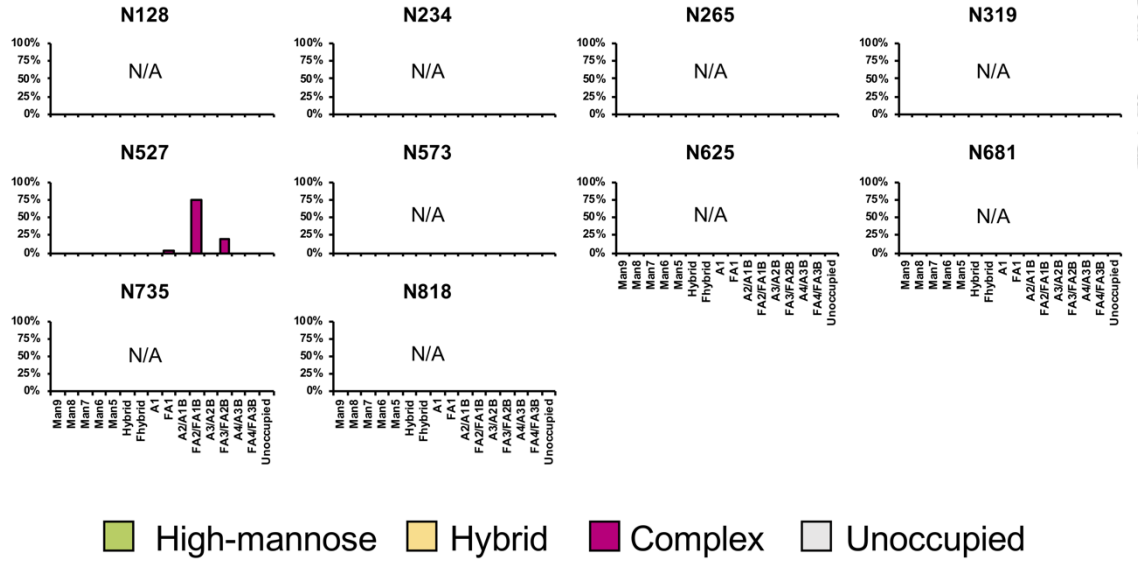


Figure A10. Quantification of the site-specific N-linked glycosylation analysis of the human aminopeptidase N (hAPN) ectodomain (Insol-aLP).

doi:10.6342/NTU202301899

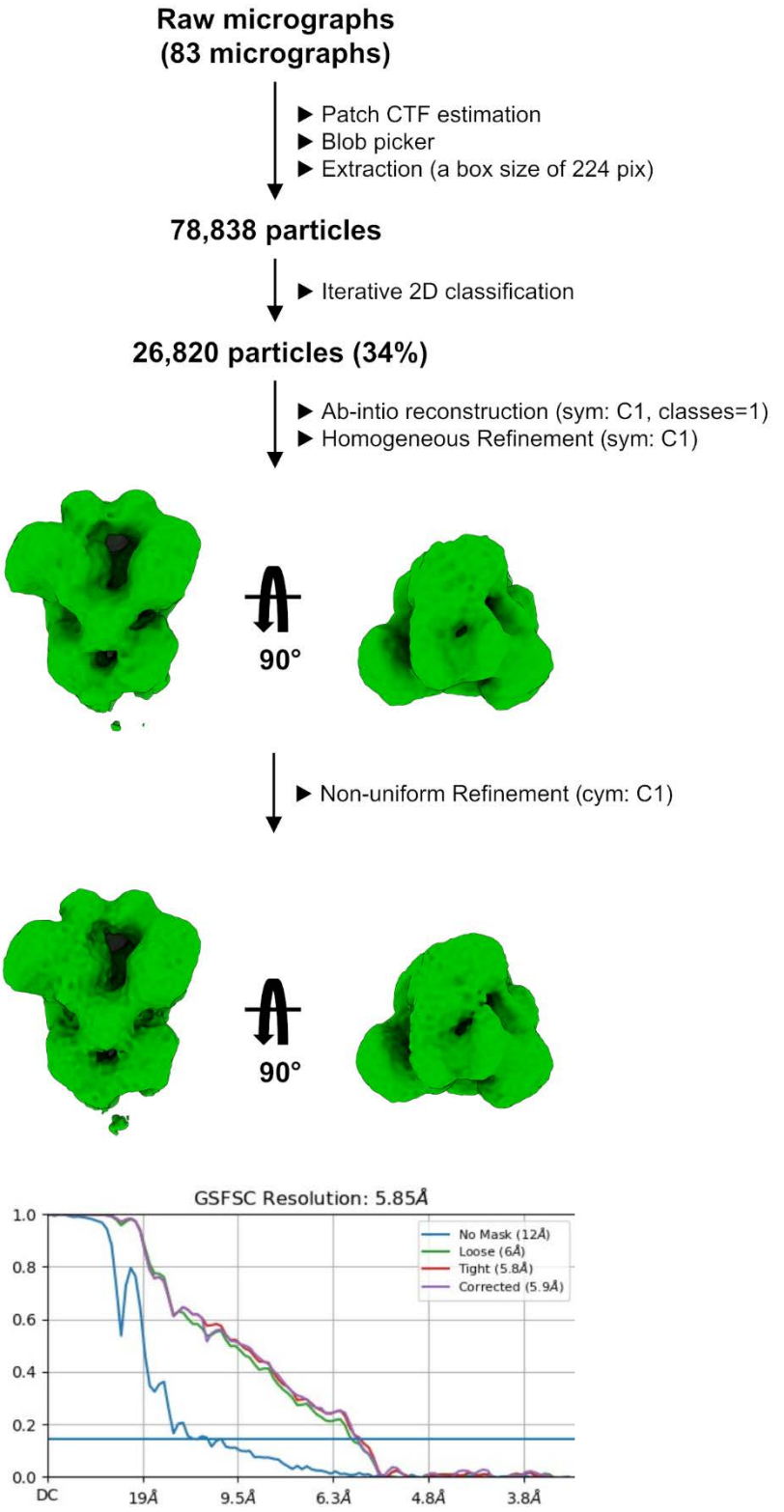


Figure A11. NSEM data processing workflow of HCoV-229E P100E S protein.

doi:10.6342/NTU202301899

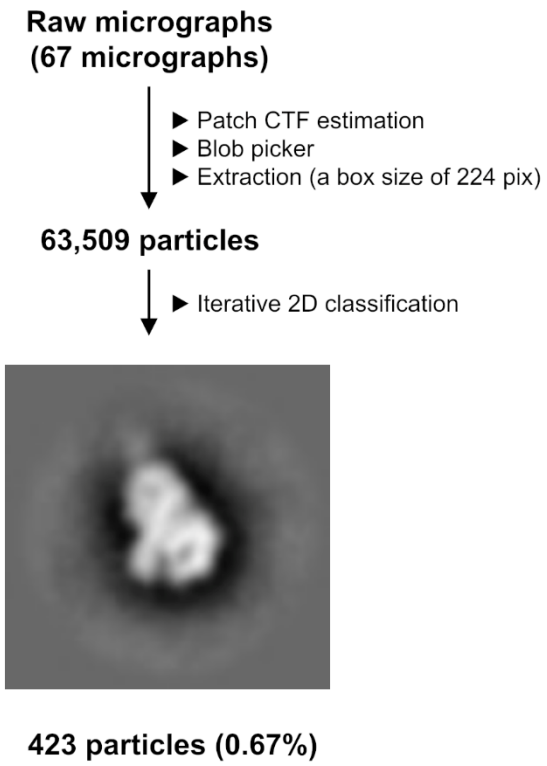


Figure A12. NSEM data processing workflow of HCoV-229E Seattle S protein.

doi:10.6342/NTU202301899

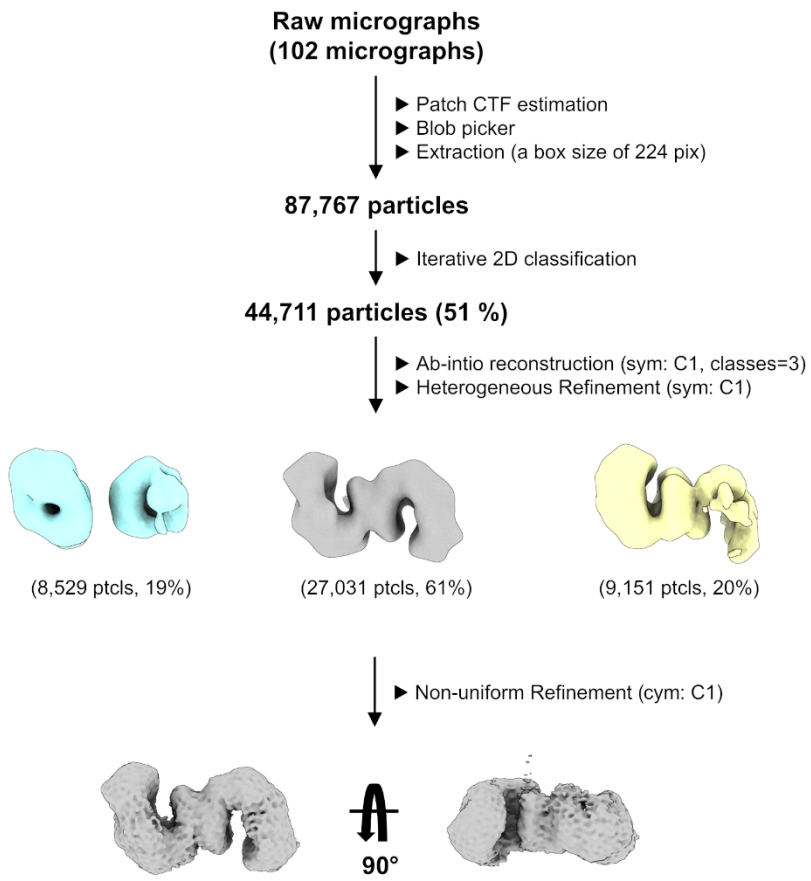


Figure A13. NSEM data processing workflow of hAPN ectodomain.

doi:10.6342/NTU202301899

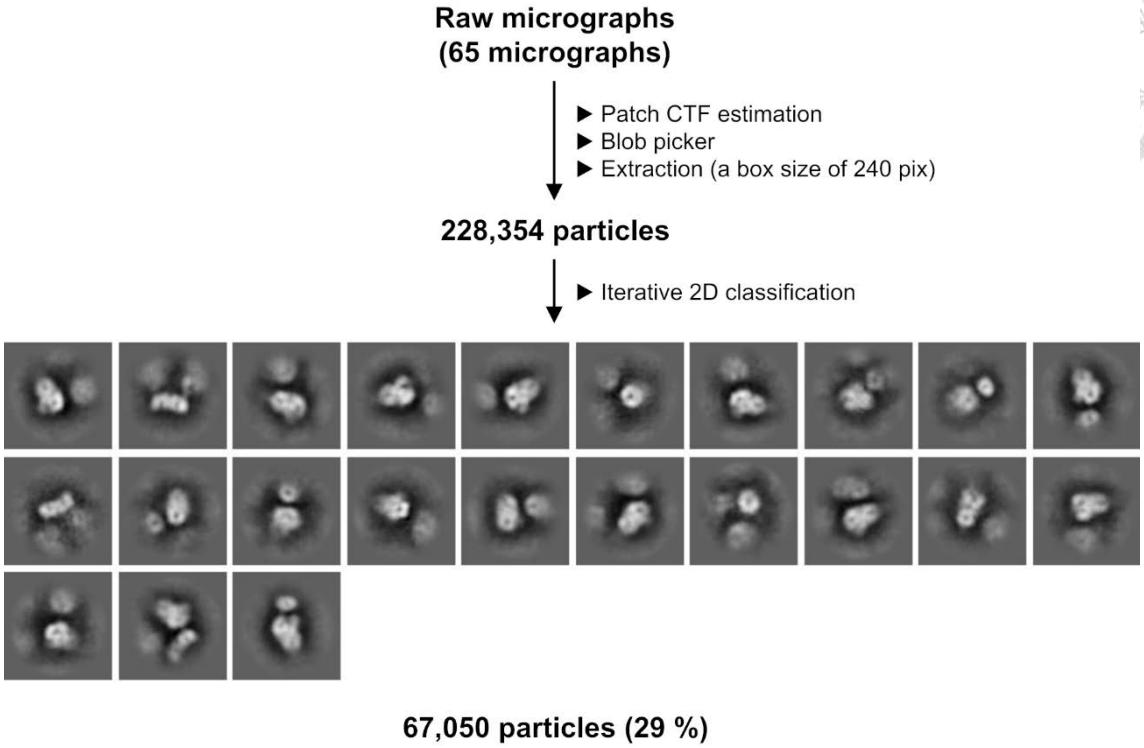


Figure A14. NSEM data processing workflow of P100E S-hAPN complex.

doi:10.6342/NTU202301899

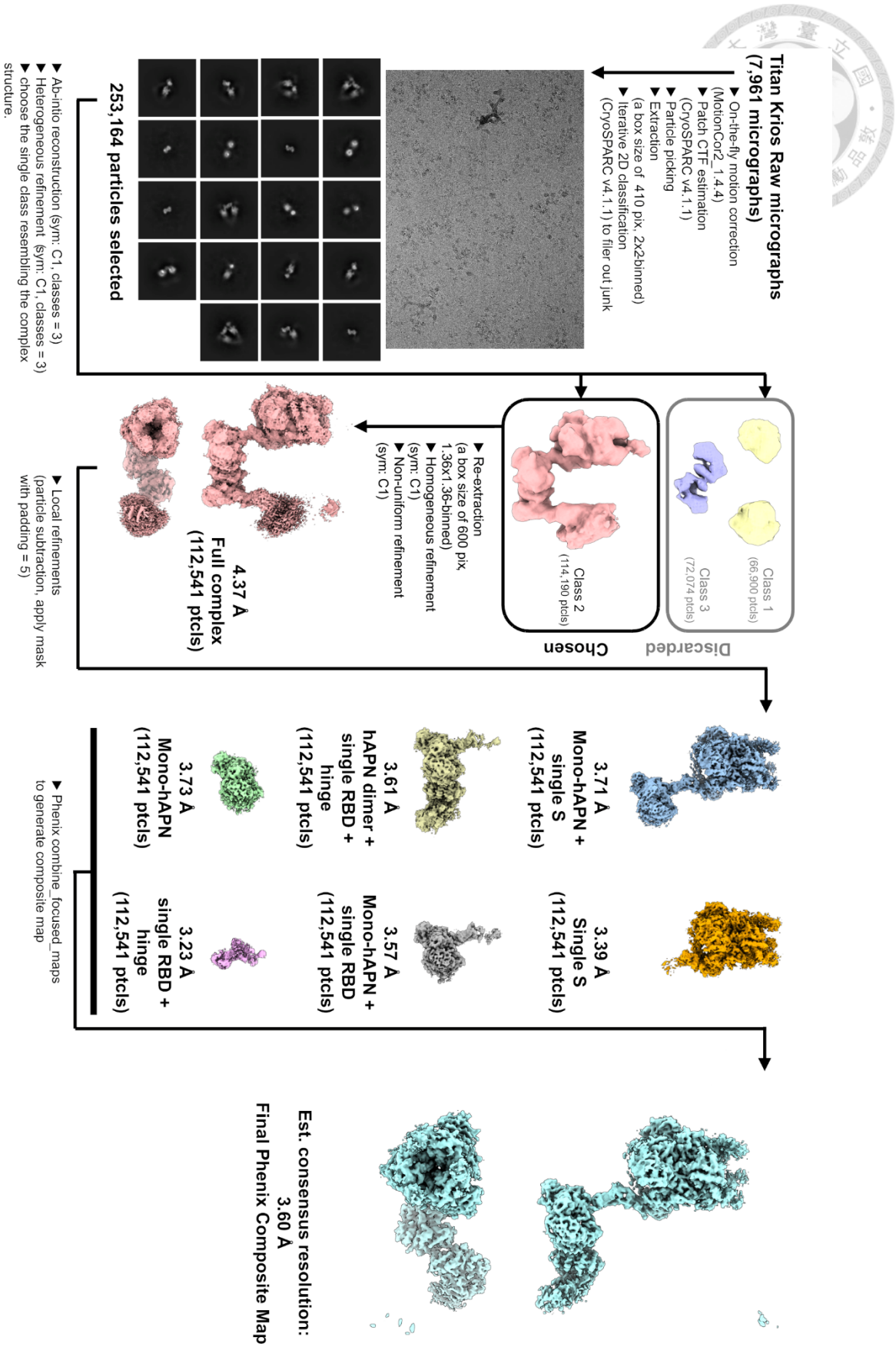
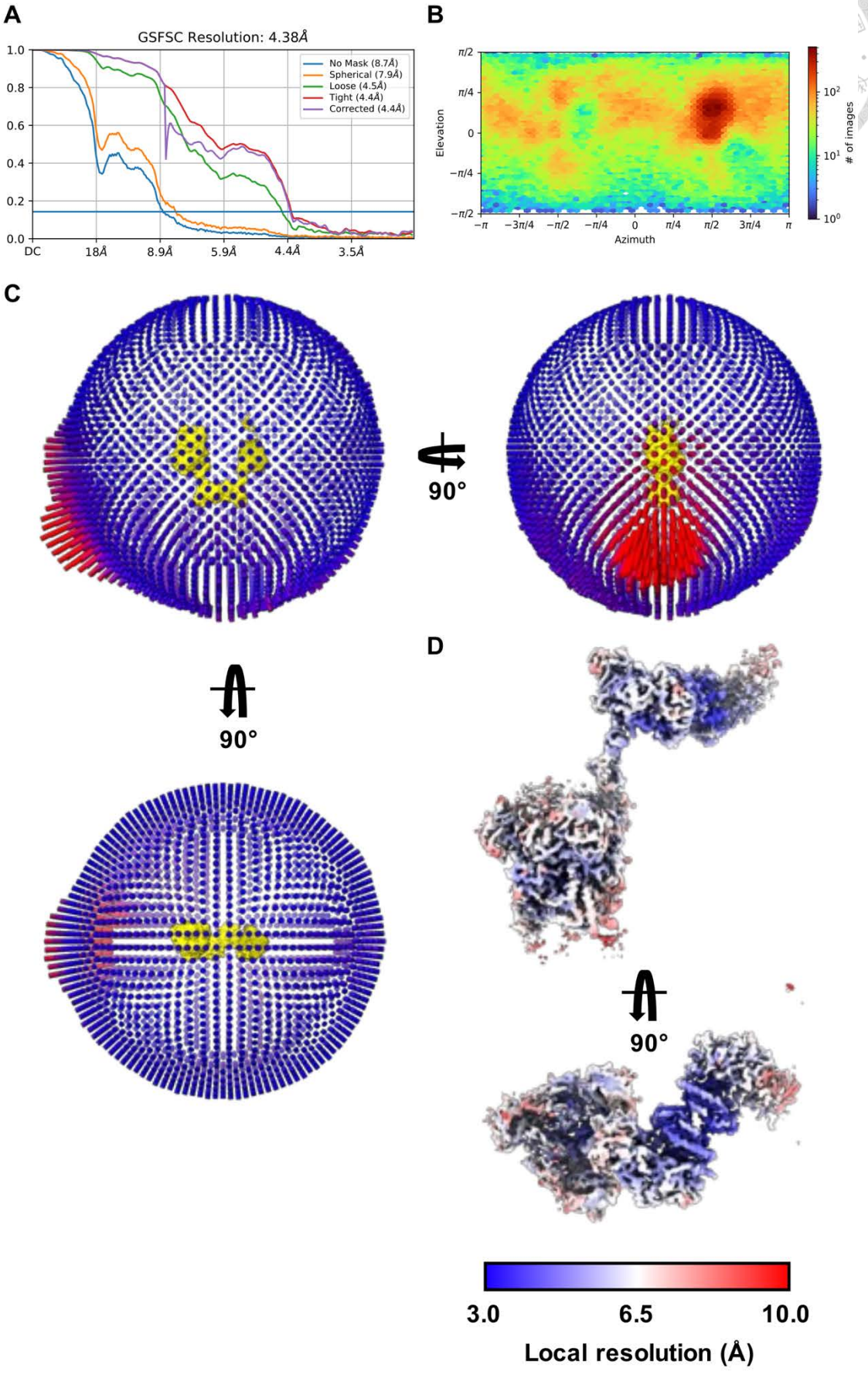


Figure A15. Cryo-EM data processing workflow of P100E S-hAPN complex

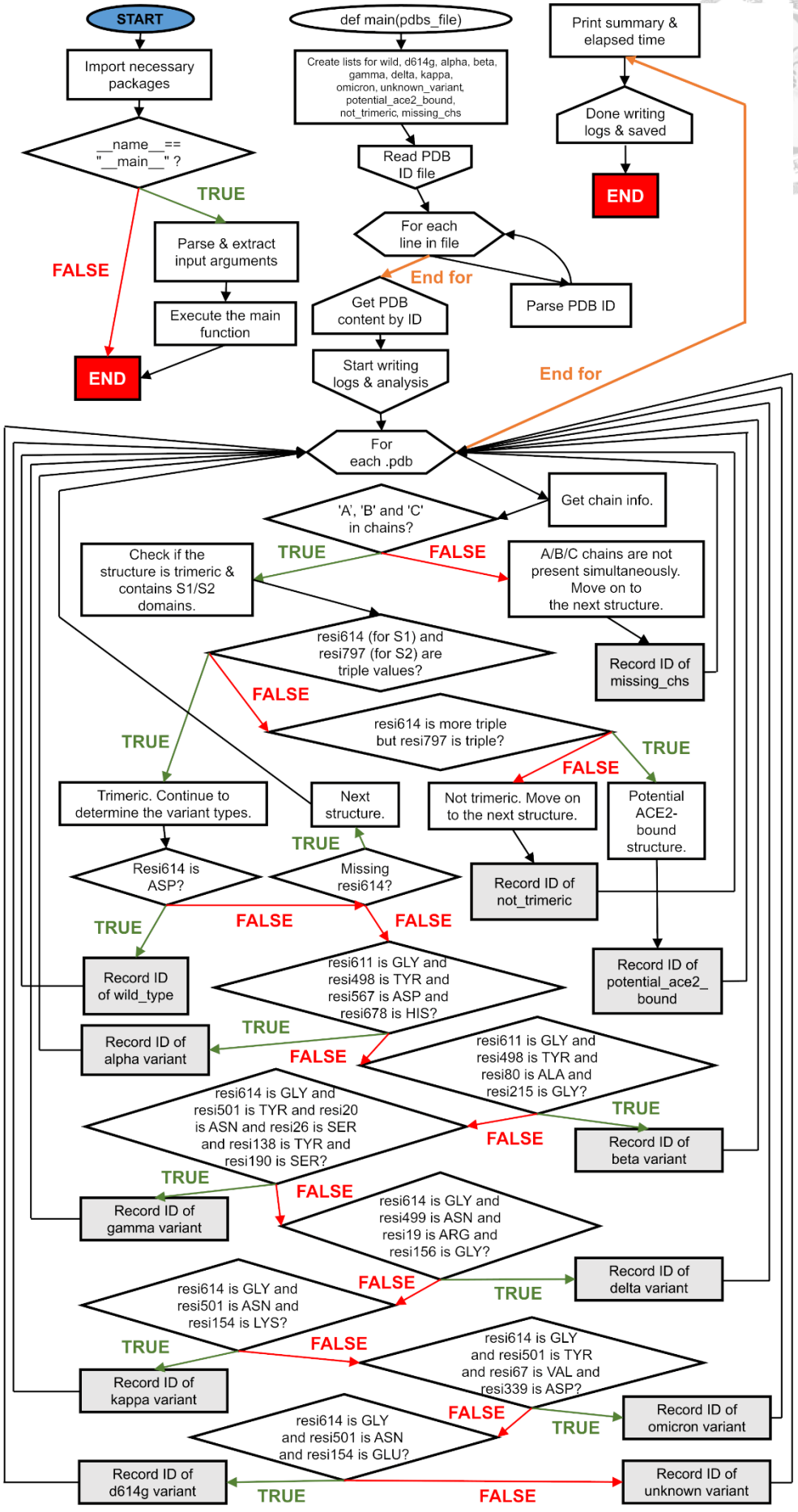
doi:10.6342/NTU202301899



doi:10.6342/NTU202301899

Figure A16. Cryo-EM data analysis and validation of P100E S-hAPN complex. (A) Gold-standard Fourier shell correlation (GSFSC) curves generated from the final NU-refinement of the complex in CryoSPARC v 4.1.1. (B) Particle angular distribution heatmap generated from the final NU-refinement of the complex in CryoSPARC v 4.1.1. (C) Euler angle distribution of the particles in the final reconstruction of the complex generated from PyEM[114]. (D) Local resolution distribution of the final complex EM map. The local resolution values are indicated in the color key below.

doi:10.6342/NTU202301899



doi:10.6342/NTU202301899

Figure A17. Programming flowchart of PDB_preprocessor.py.
The script is available on https://github.com/coco0981568491/Master_Thesis/blob/master/Batch_RBD-up_angle_calculation/PDB_preprocessor.py.

doi:10.6342/NTU202301899

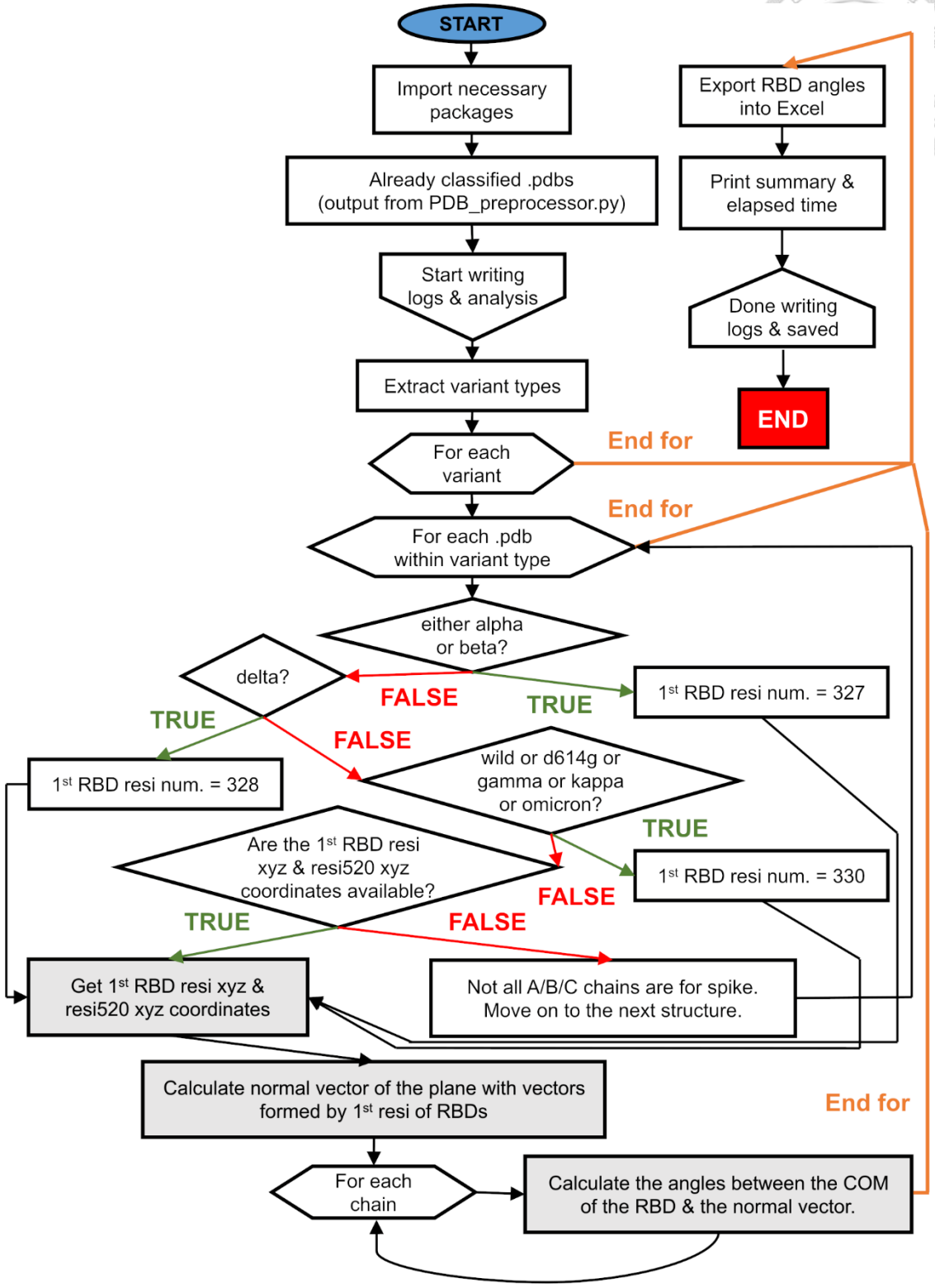


Figure A18. Programming flowchart of get_RBD_angles.py.

The script is available on https://github.com/coco0981568491/Master_Thesis/blob/master/Batch_RBD-up_angle_calculation/get_RBD_angles.py

doi:10.6342/NTU202301899

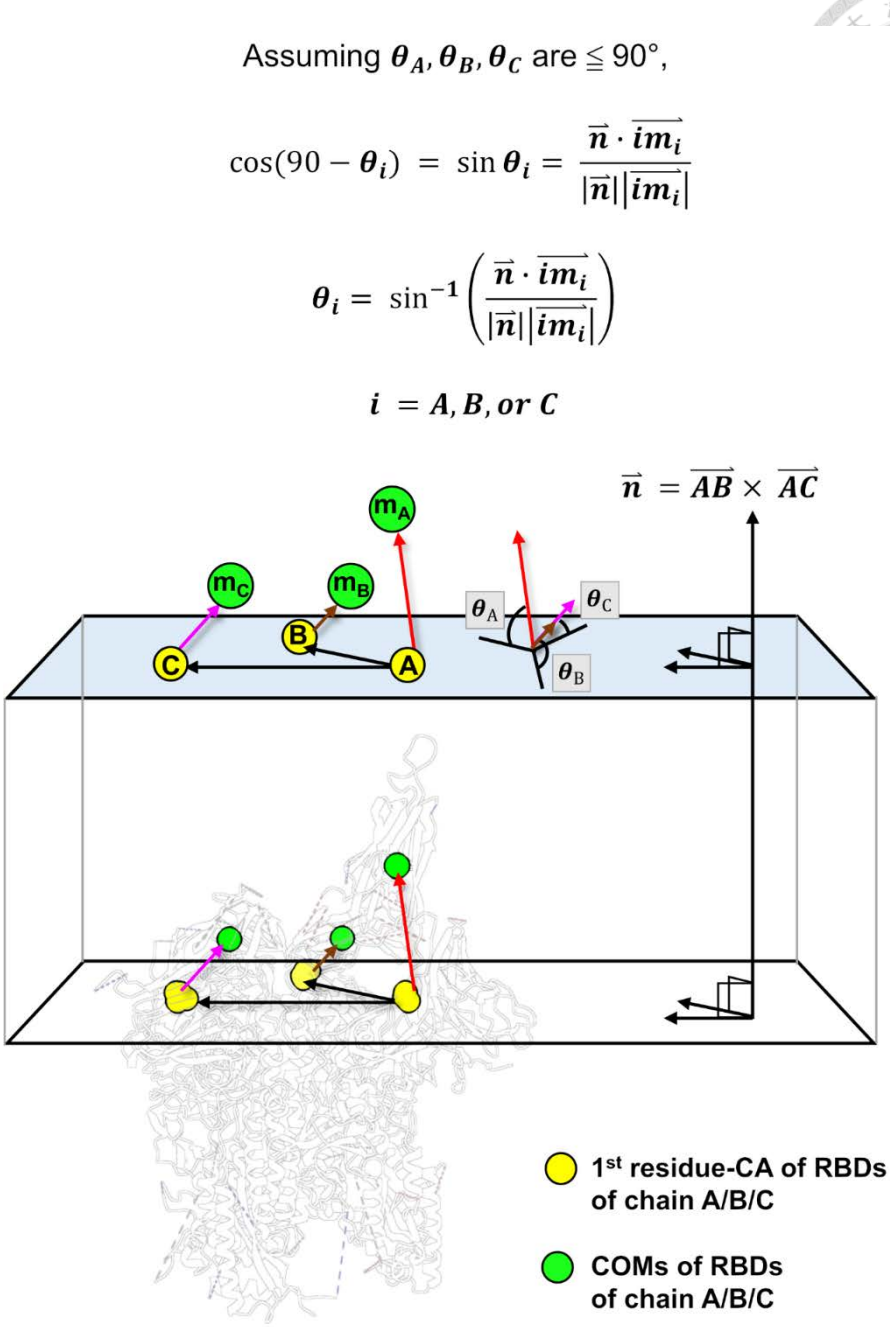


Figure A19. RBD-up angle definition and calculation.

The final polar projection plot for RBD angle distribution was based on the *θi* calculated here.

doi:10.6342/NTU202301899