

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Leveraging multiscale computational chemistry methods in basic biology and drug discovery

Permalink

<https://escholarship.org/uc/item/0zt868qc>

Author

Seitz, Christian

Publication Date

2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Leveraging multiscale computational chemistry methods in basic biology and drug discovery

A dissertation submitted in partial satisfaction of the

requirements for the degree

Doctor of Philosophy

in

Chemistry

with a specialization in

Multi-Scale Biology

by

Christian Gabriel Seitz

Committee in charge:

Professor Rommie E. Amaro, Co-Chair

Professor J. Andrew McCammon, Co-Chair

Professor Simpson Joseph

Professor Andrew McCulloch

Professor Wei Wang

Professor John Weare

2022

The dissertation of Christian Gabriel Seitz is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

III

DEDICATION

*To me, for not letting myself give up
and to my people, for believing in me when I didn't*

EPIGRAPH

He stopped, for saying the truth aloud was unendurable. It was not shame that stopped him, but fear, the same fear. He knew now why this tranquil life in sea and sunlight on the rafts seemed to him like an after-life or a dream, unreal. It was because he knew in his heart that reality was empty: without life or warmth or color or sound: without meaning. There were no heights or depths. All this lovely play of form and light and color on the sea and in the eyes of men, was no more than that: a playing of illusions on the shallow void.

They passed, and there remained the shapelessness and the cold. Nothing else.

Sparrowhawk was looking at him, and he had looked down to avoid that gaze. But there spoke in Arren unexpectedly a little voice of courage or of mockery: it was arrogant and pitiless, and it said, "Coward! Coward! Will you throw even this away?"

So he looked up, with a great effort of his will, and met his companion's eyes.

Sparrowhawk reached out and took his hand in a hard grasp, so that both by eye and by flesh they touched. He said Arren's true name, which he had never spoken: "Lebannen." Again he said it: "Lebannen, this is. And thou art. There is no safety, and there is no end. The word must be heard in silence; there must be darkness to see the stars. The dance is always danced above the hollow place, above the terrible abyss."

Arren clenched his hands and bent his forehead down till it pressed against Sparrowhawk's hand. "I failed you," he said. "I will fail you again and fail myself. I have not strength enough!"

"You have strength enough." The mage's voice was tender, but beneath tenderness was that same hardness that had risen in the depths of Arren's own shame, and mocked him. "What you love, you will love. What you undertake, you will complete. You are a fulfiller of hope; you are to be relied on."

Ursula K. LeGuin, *The Farthest Shore* (1972)

TABLE OF CONTENTS

Dissertation Approval Page	III
Dedication	IV
Epigraph	V
Table of Contents	VII
List of Figures	X
List of Tables	XV
Acknowledgments	XVII
Vita	XXI
Abstract of the Dissertation	XXIV
Chapter 1: Introduction	1
1.1 Overview	1
1.2 Elastic network models	2
1.3 Brownian dynamics	3
1.4 Molecular dynamics	3
1.5 Markov state models	4
1.6 Density functional theory	5
References	6
Part I: Elastic network models	
Chapter 2: Architecture and self-assembly of the jumbo bacteriophage nuclear shell	9
2.1 Abstract	9
2.2 Main	9
2.3 Discussion	27
2.4 Methods	29
2.5 Data availability	44
2.6 Acknowledgements	45
2.7 Author contributions	45
2.8 Competing interests	46
2.9 Extended data figures and tables	46

References	83
 Part II: Brownian dynamics, molecular dynamics and Markov state models	
Chapter 3: Multiscale simulations examining glycan shield effects on drug binding to influenza neuraminidase	89
3.1 Abstract	89
3.2 Statement of significance	89
3.3 Introduction	90
3.4 Methods	91
3.5 Results	99
3.6 Discussion	106
3.7 Conclusion	115
3.8 Author contributions	116
3.9 Acknowledgements	116
3.10 Supporting material	118
References	133
 Chapter 4: Breathing and tilting: mesoscale simulations illuminate influenza glycoprotein vulnerabilities	140
4.1 Abstract	140
4.2 Introduction	140
4.3 Results	145
4.4 Discussion	165
4.5 Materials and methods	171
4.6 Associated content	174
4.7 Author information	174
4.8 Acknowledgements	175
References	177
4.9 Supporting information	184
Supplementary references	242
 Chapter 6: Cell-scale Markov state models characterize the structural and dynamical basis of Group 1 influenza fusion inhibition	246
5.1 Abstract	246
5.2 Introduction	247
5.3 Results and discussion	250
5.4 Conclusions	258
5.5 Methods	258
5.6 Acknowledgments	259
5.7 Author contributions	260

5.8 Supplementary Information	261
References	268

Part III: Density functional theory

Chapter 6: Data for molecular dynamics simulations of Escherichia coli cytochrome bd oxidase with the Amber force field	271
6.1 Abstract	271
6.2 Specifications table	272
6.3 Data description	273
6.4 Experimental design, materials and methods	274
6.5 CRediT authorship contribution statement	278
6.6 Declaration of competing interest	279
6.7 Acknowledgements	279
References	280

LIST OF FIGURES

Figure 2.1	<i>In situ</i> tomography and subtomogram analysis of the 201phi2-1 phage nucleus	10
Figure 2.2	<i>In vitro</i> cryo-EM structure of the 201phi2-1 phage nuclear shell protein chimallin	13
Figure 2.3	Flexibly attached N- and C-terminal segments mediate self-assembly of the chimallin shell ...	17
Figure 2.4	Electrostatics and pores of the chimallin shell	20
Figure 2.5	Structural conservation of chimallin in the distantly related <i>E. coli</i> jumbo phage Goslar	24
Figure 2.6	<i>In situ</i> cryoFIB-ET of 201phi2-1-infected <i>P. chlororaphis</i> cells and subtomogram analysis	47
Figure 2.7	<i>In vitro</i> tomography of 201phi2-1 chimallin void peak	49
Figure 2.8	Single-particle reconstruction of the <i>in vitro</i> 201phi2-1 chimallin cubic assembly	50
Figure 2.9	Partial homology of chimallin fold topology to known structures	52
Figure 2.10	Interactions of NTS and CTS segments with the chimallin core	54
Figure 2.11	SEC-MALS of 201phi2-1 chimallin truncations	55
Figure 2.12	GFP-chimallin incorporation into the nuclear shell in 201phi2-1-infected <i>P. chlororaphis</i> and truncation mutant growth curves	57
Figure 2.13	Analysis of flexibility by Gaussian network models and subtomogram analysis	59
Figure 2.14	Size and hydrophobicity profiles of the four-fold pores	60
Figure 2.15	<i>In situ</i> cryoFIB-ET of Goslar infected APEC2248 cells and subtomogram analysis	62
Figure 2.16	Single-particle reconstruction of the <i>in vitro</i> Goslar chimallin cubic assembly	63
Figure 2.17	Unidentified spherical bodies present in jumbo phage-infected cell populations and speculative models	65

Figure 2.18	Alignment of 201phi2-1 and phiKZ chimallin proteins	67
Figure 3.1	Structure of the NA head	94
Figure 3.2	Glycan shielding on NA	99
Figure 3.3	Clustered glycans on monomer A	100
Figure 3.4	The superposition of each glycan onto the static NA structure shows the conformational variation of the glycans	101
Figure 3.5	Association rates of each ligand to the primary and secondary sites	103
Figure 3.6	A sample successful trajectory, showing that the ligand explores significant space before finding the binding site	110
Figure 3.7	Ligand binding mechanisms to neuraminidase for monovalent binders (A) and multivalent binders (B)	115
Figure 3.8	2D structures of the ligands used in this study	118
Figure 3.9	Comparison of our computed association rates with previously published experimental association rates	119
Figure 3.10	Structure of oseltamivir	120
Figure 3.11	Structure of sialic acid	121
Figure 3.12	Structure of zanamivir	122
Figure 3.13	Clustered glycans on monomer B	127
Figure 3.14	Clustered glycans on monomer C	127
Figure 3.15	Clustered glycans on monomer D	128
Figure 4.1	All-atom glycosylated model of influenza A virion	143
Figure 4.2	NA head tilting motion	146
Figure 4.3	HA ectodomain tilting motion	151

Figure 4.4	Breathing motion of the HA head domain	155
Figure 4.5	HA breathing reveals the FluA-20 cryptic epitope	158
Figure 4.6	HA/NA interplay	163
Figure 4.7	HA and NA interplay at the basis of the viral egress process	169
Figure 4.8	All-atom glycosylated model of H1N1-Mich2015 influenza A virion	203
Figure 4.9	Aligned sequences of HA in H1N1-Shan2009 (UniProtKB accession code: F2YI86) and H1N1-Mich2015 (UniProtKB accession code: A0A144YDV8) as modeled in this work	204
Figure 4.10	Aligned sequences of NA in H1N1-Shan2009 (UniProtKB accession code: F2YI86) and H1N1-Mich2015 (UniProtKB accession code: A0A0X9QTS2) as modeled in this work	205
Figure 4.11	Schematic representation of the glycosylation procedure	206
Figure 4.12	Glycosylation of the 708 monomeric HAs (top) and 120 monomeric NAs (bottom) included in the H1N1-Shan2009 whole-virion model	207
Figure 4.13	List of amino acid point mutations in H1N1-Mich2015 with respect to H1N-Shan2009 for HA (left panel) and NA (right panel)	208
Figure 4.14	Glycan variation in H1N1-Mich2015 with respect to H1N1-Shan2009 for HA (A) and NA (B). A new N-linked glycan is introduced at position N179 within the HA head	209
Figure 4.15	Glycosylation of the 708 monomeric HAs (top) and 120 monomeric NAs (bottom) included in the H1N1-Mich2015 whole-virion model	210
Figure 4.16	Average RMSD profiles for HA (A) and NA (B) as calculated over the course of H1N1-Shan2009 (black line) and H1N1-Mich2015 (gray line) whole-virion MD simulations	211
Figure 4.17	Average RMSF profiles for HA (A) and NA (B) as calculated over the course of H1N1-Shan2009 (black line) and H1N1-Mich2015 (gray line) whole-virion MD simulations	212
Figure 4.18	Maximum NA head-tilt angle observed in H1N1-Shan2009 and H1N1-Mich2015 NA tetramers	213
Figure 4.19	The implied timescale plots for the MSM of the NA head tilting in (A) H1N1-Shan2009 and (B) H1N1-Mich2015	213

Figure 4.20	The CK test plots for the MSM of the NA head tilting in (A) H1N1-Shan2009 and (B) H1N1-Mich2015	214
Figure 4.21	Isolation and characterization of NA-directed mAb NDS.1	215
Figure 4.22	Negative-stain EM analysis of N2 A/Darwin/9/2021 NA tetramer in complex with Fab NDS.1	215
Figure 4.23	NS-EM reconstruction model of N2 A/Indiana/10/2011 NA tetramer in complex with NDS.1 Fab and 1G01 Fab	216
Figure 4.24	The NA head-underside epitope	217
Figure 4.25	The implied timescale plots for the MSM of the HA ectodomain tilting in (A) H1N1-Shan2009 and (B) H1N1-Mich2015	217
Figure 4.26	The CK test plots for the MSM of the HA ectodomain tilting in (A) H1N1-Shan2009 and (B) H1N1-Mich2015	218
Figure 4.27	Analysis of the correlation between HA ectodomain tilting and HA protein confinement	220
Figure 4.28	HA anchor epitope accessibility	222
Figure 4.29	The implied timescale plots for the MSM of the HA head breathing in (A) H1N1-Shan2009 and (B) H1N1-Mich2015	222
Figure 4.30	The CK test plots for the MSM of the HA head breathing in (A) H1N1-Shan2009 and (B) H1N1-Mich2015	223
Figure 4.31	HA head trimer interface epitope accessibility	225
Figure 4.32	RMSF of HA and NA transmembrane domain (TMD) COM and ectodomain COM. The initial frame was used as a reference for RMSF calculation	226
Figure 4.33	Analysis of glycoprotein inter-connections for HA and NA	227
Figure 4.34	Analysis of glycoprotein macro-cluster formation	228
Figure 4.35	Analysis of glycoprotein connections involving glycans	230
Figure 4.36	Analysis of NA head breathing	232

Figure 5.1	Arbidol binding site	249
Figure 5.2	Individual HA trajectories are extracted from the viral coat simulation and are then featurized for analysis by the construction of an MSM	255
Figure 5.3	Representative structures of the two major macrostates with mean first-passage times shown	256
Figure 5.4	View of the primary pocket and pocket extension	257
Figure 5.5	Free energy landscape	261
Figure 5.6	CK test of the H1N1-Shan2009 unglycosylated MSM	262
Figure 5.7	CK test of the H1N1-Shan2009 glycosylated MSM	262
Figure 5.8	CK test of the H1N1-Mich2015 glycosylated MSM	263
Figure 5.9	Representative structures of the four major macrostates of the unglycosylated H1N1-Shan2009 model with mean first-passage times shown	264
Figure 5.10	Ligand interaction diagram of arbidol in its binding pocket	265
Figure 5.11	The total kinetic variance of the MSM features for each MSM (2009u = H1N1-Shan2009 unglycosylated model, 2009g = H1N1-Shan2009 glycosylated model, 2015g = H1N1-Mich2015 glycosylated model)	266
Figure 5.12	a , Implied timescale plot for the H1N1-Shan2009 unglycosylated mode. b , Implied timescale plot for the H1N1-Shan2009 glycosylated model. c , Implied timescale plot for the H1N1-Mich2015 glycosylated model	267

LIST OF TABLES

Table 2.1	<i>In situ</i> cryoFIB-ET data collection and reconstruction statistics	69
Table 2.2	Cryo-EM data collection, reconstruction, and refinement statistics	70
Table 2.3	Summary of <i>in vitro</i> 201phi2-1 chimallin protomer interfaces	71
Table 2.4	Summary of a 201phi2-1 chimallin protomer's polar interactions <i>in vitro</i>	72
Table 2.5	Gaussian network model hinge residues for 201phi2-1 chimallin	73
Table 2.6	Pore-facing residues during 201phi2-1 chimallin simulations	75
Table 2.7	Bacteriophage chimallin proteins	76
Table 2.8	Summary of <i>in vitro</i> Goslar chimallin protomer interfaces	78
Table 2.9	Summary of a Goslar chimallin protomer's polar interactions <i>in vitro</i>	79
Table 2.10	Bacterial strains and plasmids used in this study	81
Table 3.1	Glycan structures from the Glyprot web server	93
Table 3.2	Ions used in the implicit solvent in the BD trajectories	118
Table 3.3	Grid spacing for each system	118
Table 3.4	b radius and q radius values for each system	118
Table 3.5	Primary site atom pairs for oseltamivir	123
Table 3.6	Primary site atom pairs for sialic acid	124
Table 3.7	Primary site atom pairs for zanamivir	125
Table 3.8	Secondary site atom pairs for sialic acid	126
Table 3.9	Secondary site atom pairs for oseltamivir	126
Table 3.10	Secondary site atom pairs for zanamivir	126
Table 3.11	Association rate data for oseltamivir	129
Table 3.12	Association rate data for oseltamivir	130
Table 3.13	Association rate data for sialic acid boat conformation	131

Table 3.14	Association rate data for sialic acid chair conformation	132
Table 4.1	Tilt angles (degree) of ten tilted and untilted structures (microstates) extracted from the MSM of the NA head tilting	233
Table 4.2	Tilt angles (degree) of ten tilted and untilted structures (microstates) extracted from the MSM of the HA ectodomain tilting	234
Table 4.3	The table reports the distance between the COM of the HA monomer head (residues 66 to 286) and the COM of L α Hs' apical residue (419) calculated for open and closed structures (microstates) extracted from the MSM of the HA head breathing	235
Table 4.4	P(A B) for HA ectodomain tilting	236
Table 4.5	P(A B) for HA head breathing	237
Table 4.6	P(A B) for NA head tilting	238
Table 4.7	P(B A) for HA ectodomain tilting	239
Table 4.8	P(B A) for HA head breathing	240
Table 4.9	P(B A) for NA head tilting	241

ACKNOWLEDGMENTS

This world is enveloped in chaos. Chaos is adiaphorous. It is morally ambiguous. It is neither good nor bad, nor is it predictable. But through this chaos, some amazing events can happen if the right conditions line up. This series of seemingly unconnected, serendipitous events was set in motion by my parents. My mom, who gave up her career, clipped coupons and worked at a pizza shop to teach me and propel me forward. My dad, who worked in a cereal factory during the day and painted houses at night to put food on the table. Their intense sacrifices gave me an education and allowed me to attend college, and I would like to thank both of my parents for that.

I would like to thank the admissions committee at Elon University for giving me a chance when other universities did not think I was capable of succeeding, and I would like to thank several professors there. Dr. Dan Wright, who called this terrified freshman into his office for a one-on-one meeting and told me to take chemistry seriously: You identified potential in me and spoke it into existence. I would not be a scientist today if it were not for you (but Krispy Kreme is still better than Dunkin Donuts any day of the week). Dr. Karl Sienerth, my academic advisor, who guided me on a path through a chemistry degree I never planned to do. Dr. Joel Karty, who gave me an ambitious research project and helped me through years of troubleshooting: thank you. The three-headed monster (☺) of Dr. Olivia Choplin, Dr. Sophie Adamson and Dr. Sarah Glasco: on paper you taught me French, but in reality you fought for me. You didn't let me give up. You spoke out in my defense when you didn't need to. Dr. Janet Myers and Sarah Lentz: Every PhD student has a superpower. You taught me how to harness my writing, and make it my superpower. And to Dr. Leo Lambert and the Lumen Prize committee: You helped me believe in myself when life, and people in it, were telling me that I did not bELONg. Dr. Mai-Thi Nguyen-Kim, and her wonderful team of Dr. Jannik Borghs and Dr. Jan Wallenborn in Aachen: You showed me how to love science. Dr. Bill Goddard, Dr. Soo-Kyung Kim and the Amgen committee at Caltech: You showed me how to love research. Simone Jasper, for always cheering me on through ~everything~ (and cheering on our basketball team together). Monica Evans: I'm glad I have someone I can process generational traumas with

(among other things in our delightfully long phone calls). Nikita Fyodorov: You always wanted to spread joy to the world – you have certainly spreaded it to me. Rev. Courtney Davis Shoemaker and Rev. Julie Tonnesen: You gave me space to grieve when I so desperately needed it, and in turn gave me space to heal. These people I thank and admire. Some of them I talk to frequently, others I do not. But I keep all of them in my heart.

I would like to thank the admissions committee at UCSD, who, unlike some others, believed I went to a reputable school and thus was eligible for a PhD. And I would like to thank numerous people in San Diego. Dr. Andy McCammon, who is always praying on my success. Dr. Rommie Amaro, who showed that you can be empathetic and a scientist. Thank you two for giving me a chance to live out my dreams when others did not. The Amaro and McCammon labs, specifically, Abby Dommer, Robert Konecny, Ozlem Demir, Jeff Wagner, and Dan Mermelstein, and the IT staff of Bryan Hill and Kevin Smith, for their support. My committee of Dr. Andrew McCulloch, Dr. Wei Wang, Dr. Simpson Joseph and Dr. John Weare, for helping me look at science through different lenses. My NSF GRFP reviewers, who selected me for recognition out of a sea of choices, you forcefully told me that I am valid.

Alec Rachkov: What a guy, your faith in me and support of me has never wavered. You still belong in the National Archives as a national treasure. From all those epic nights on the dance floor under neon lights to simply laying on your couch falling asleep while watching tennis, you make everything fun. With you, I am myself, and that is a revolutionary thing. Aashish Shivkumar: You always know just what I need, and have been there when I needed you, especially all those late night dinners unpacking life the many times I didn't want to see or be around anyone. Some people take so much energy but you take none – it's easy to be with you. I notice the actions of you two and how much you show that you love me: I treasure the words you say, and you always impress me just by being yourself. You keep my heart crying with joy that I have found you two, and I hope to keep you as close as possible for as long as I live. Aleks Radakovic, the og techno Serb, I'm glad we got weird together. You opened an entire new world for me, and showed me how to love music and to love tennis. Your advice I treasure more than electrum. Aayush Shah, always

looking on the bright side and bringing a laugh to my belly, you were a luminous star in those dark pandemic days – thanks for getting me out of my cave and for helping me through adulthood. Ilker Deveci, your radical honesty inspires me to be honest to myself and to set boundaries (such as firmly insisting that I need these four gallons of milk). Constanza Torres-Paris, you have shown me resilience, and you instill that strength in me. Kartik Rallapalli, I hope we keep dancing through life ☺. Alex Demetriades, I’m glad I’m as bad of an influence on you as you are on me – whenever you’re around I know I’ll have a fun time. Zied Gaieb, you took me under your wing and protected me when I was vulnerable – I’ve told you so many times but you don’t know how much you helped (especially scoring goals). Sean Dunphy, I’m glad we found each other on a rock in the middle of nowhere (but hummingbirds still live in trees, stop telling me I’m wrong). Stanley Buczynski, you are a ball of fun. Invite me over when you buy your anti-aircraft gun. And Vedran Markota, you are a homie (even when you refuse to listen to my grand ideas of you being a standup comedian or when you send me TikToks in Croatian that I don’t understand). My soccer groups and my rave group (more plur than sus), you helped me relax and come back to earth, and visit a fantasy world without stress or anxiety. Without being in my shoes, these people are not capable of comprehending just how much they have helped me, have carried me, have embraced me, have shown me *love*. My puny thanks will never repay them for just how *much* they have loved me. But I remember what they have done, and will never forget them.

Most of my PhD was funded through grants for myself, with copious help from others. I have been funded from the UCSD Interfaces program NIH T32 grant (T32EB009380) and an NSF GRFP grant (DGE-1650112).

Chapter 2, in full, is a modified reprint of the material as it appears in “Laughlin, T.G.*, Deep, A.*, Prichard, A.M., Seitz, C., Gu, Y., Enustun, E., Suslov, S., Khanna, K., Birkholz, E.A., Armbruster, E., McCammon, J.A., Amaro, R.E., Pogliano, J., Corbett, K.D., Villa, E. Architecture and self-assembly of the jumbo bacteriophage nuclear shell. *Nature*, **2022**, 608. DOI: 10.1038/s41586-022-05013-4.” The dissertation author was a co-investigator and co-author of this work.

Chapter 3, in full, is a modified reprint of the material as it appears in “Seitz, C., Casalino, L., Konecny, R., Huber, G., Amaro, R.E., McCammon, J.A. Multiscale Simulations Examining Glycan Shield Effects on Drug Binding to Influenza Neuraminidase. *Biophys. J.*, **2020**, 119 (11), pp 2275-2289. **DOI:** 10.1016/j.bpj.2020.10.024.” The dissertation author was the primary investigator and co-author of this work.

Chapter 4, in full, is a modified reprint of the material as it appears in “Seitz, C.*, Kochanek, S.*, Durrant, J., Casalino, L., Amaro, R.E. Cell-scale Markov state models characterize the structural and dynamical basis of Group 1 influenza fusion inhibition. In review.” The dissertation author was a primary co-investigator and co-author of this work.

Chapter 5, in full, is a modified reprint of the material as it appears in “Casalino, L., Seitz, C., Lederhofer, J., Tsybovsky, Y.; Wilson, I.A.; Kanekiyo, M., Amaro, R.E. Breathing and tilting: the flexibility of influenza glycoproteins reveals their vulnerabilities. *ACS Central Science*, **2022**, in press.” The dissertation author was a co-investigator and co-author of this work.

Chapter 6, in full, is a modified reprint of the material as it appears in “Ahn, S.-H.*, Seitz, C*, Cruzeiro, V.D.W., McCammon, J.A., Götz, A.W. Data for molecular dynamics simulations of Escherichia coli cytochrome bd oxidase with the Amber force field. *Data in Brief*, **2021**, 38. **DOI:** 10.1016/j.dib.2021.107401.” The dissertation author was a primary co-investigator and co-author of this work.

*denotes shared primary authorship

VITA

2016	Bachelor of Science, Elon University
2016 – 2017	Teaching Assistant, University of California San Diego
2020	Master of Science, University of California San Diego
2022	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

Seitz, C.G., Zhang, H., Mo, Y., Karty, J.M. Why Do Enolate Anions Favor O-Alkylation in the Gas Phase? The Roles of Resonance and Inductive Effects in the Gas-Phase S_N2 Reaction between the Acetaldehyde Enolate Anion and Methyl Fluoride. *J. Org. Chem.*, **2016**, 81 (9), pp 3711-3719. **DOI:** 10.1021/acs.joc.6b00351

Seitz, C., Casalino, L., Konecny, R., Huber, G., Amaro, R.E., McCammon, J.A. Multiscale Simulations Examining Glycan Shield Effects on Drug Binding to Influenza Neuraminidase. *Biophys. J.*, **2020**, 119 (11), pp 2275-2289. **DOI:** 10.1016/j.bpj.2020.10.024

Ahn, S.-H.*, **Seitz, C***, Cruzeiro, V.D.W., McCammon, J.A., Götz, A.W. Data for molecular dynamics simulations of Escherichia coli cytochrome bd oxidase with the Amber force field. *Data in Brief*, **2021**, 38. **DOI:** 10.1016/j.dib.2021.107401

Laughlin, T.G.*, Deep, A.*, Prichard, A.M., **Seitz, C.**, Gu, Y., Enustun, E., Suslov, S., Khanna, K., Birkholz, E.A., Armbruster, E., McCammon, J.A., Amaro, R.E., Pogliano, J., Corbett, K.D., Villa, E. Architecture and self-assembly of the jumbo bacteriophage nuclear shell. *Nature*, **2022**, 608. **DOI:** 10.1038/s41586-022-05013-4

Takahashi, Y., Bosmans, K., Hsu, P.-K., Paul, K., Yeh, C.-Y., Wang, Y.-S., Yarmolinsky, D., Sierla, M., Vahisalu, T., **Seitz, C.**, McCammon, J.A., Kangasjarvi, J., Kollist, H., Zhang, L., Trac, T., Schroeder, J.I. Stomatal CO₂/bicarbonate Sensor Consists of Two Interacting Protein Kinases, Raf-like HT1 and non-kinase-activity requiring MPK12/MPK4. *Sci. Adv.*, **2022**, 8, (49). **DOI:** 10.1126/sciadv.abq616

Seitz, C.*, Kochanek, S.*, Durrant, J., Casalino, L., Amaro, R.E. Cell-scale Markov state models characterize the structural and dynamical basis of Group 1 influenza fusion inhibition. *In revisions at Cell CHEM*

Casalino, L., **Seitz, C.**, Lederhofer, J., Tsybovsky, Y.; Wilson, I.A.; Kanekiyo, M., Amaro, R.E. Breathing and tilting: the flexibility of influenza glycoproteins reveals their vulnerabilities. *In press at ACS Central Science*

Shivkumar, A.*, Berg, K.*, Sibucão, K.*, Leriche, G., Espinoza, C., Dozier, L., Patrick, G., Gaieb, Z., **Seitz, C.**, Amaro, R.E., Park, H.-H., Hoe, H.-S., Wozniak, J., Gonzalez, D., Yang, J. Nootropic Benzothiazoles Promote Dendritic Spine Formation by Targeting Fascin-1. *Submitted to Nature Chemical Biology*

*denotes shared primary authorship

FIELD OF STUDY

Major Field: Chemistry and Biochemistry

Track: Biochemistry and Biophysics

Specialization: Multi-Scale Biology

Studies in Computational and Theoretical Chemistry

Professor Rommie E. Amaro

Professor J. Andrew McCammon

ABSTRACT OF THE DISSERTATION

Leveraging multiscale computational chemistry methods in basic biology and drug discovery

by

Christian Gabriel Seitz

Doctor of Philosophy in Chemistry with a specialization in Multi-Scale Biology

University of California San Diego, 2022

Professor Rommie E. Amaro, Co-Chair
Professor J. Andrew McCammon, Co-Chair

Computers have held an increasingly important place in chemical and biological research. Non-specialists may presume that computers can perform a small number of specialized techniques, but in reality they may be used on a wide range of biological scales to answer disparate questions in biology and chemistry. This body of work exemplifies that, using scales ranging from coarse grained nodes representing amino acids at the cell scale down to electron density clouds. Specifically, the projects here use elastic network models to investigate basic bacterial biology, Brownian dynamics and molecular dynamics to investigate influenza, and quantum mechanics to investigate tuberculosis. In this work we find that the protein sheets of *chimallin* in *Pseudomonas chlororaphis* cells infected with jumbo phage 201phi2-1 display a significant amount of flexibility and have several pores which could transport ions or nucleotides. On a different project we found that glycans can inhibit drug binding to influenza, and that the glycoproteins in influenza can access protein conformations amenable to cellular exocytosis and endocytosis. We also find that the flexibility of these glycoproteins opens new epitopes and drug binding sites that cannot be found from their crystal structures. Finally, we develop force field parameters for cytochrome *bd*-type quinol oxidase, which is a promising drug target for *Mycobacterium tuberculosis* and *Escherichia coli*.

Chapter 1

Introduction

1.1 Overview

Biological systems are composed of some combination of proteins, nucleic acids, water, ions, lipids, carbohydrates, and others. In particular, humans contain a dazzling variety of proteins, some of which are well studied and some of which are not. Knowledge about these proteins allows one to better understand a disparate variety of life functions, from how our muscles contract to how we perceive light to how food is digested and much, much more. Proteins are made up of amino acids, with covalent bonds linking these residues together. These bonds allow for a wide range of protein flexibility, when needed. Some proteins must change conformation upon receiving stimuli, others must lock in a specific conformation for an enzymatic process to occur, and others may exist in a broad range of conformations instead of simply existing in one state or another. Understanding these protein dynamics allows us to understand how biological life cycles can continue.

An understanding of protein dynamics is needed to comprehend how life can exist, but it is also needed for understanding how life can end. Many diseases are caused by aberrant proteins whose mutations cause transportation or catalytic cycles to fail. Other diseases are caused by invading bacteria or viruses which infect and kill our cells. Being able to “fix” misregulated proteins or block invading proteins similarly requires an understanding of protein dynamics. Once one understands how these proteins function, one can begin to target them through drug discovery, to modify their function.

The work here touches disparate aspects of protein dynamics, and is organized into three parts. Part I, entitled “**Elastic network models**”, contains material from one publication to understand how techniques such as elastic network models and molecular dynamics can be used to explore protein flexibility and solvent/solute passage through protein assemblages. Coarse grained methods such as this are required in situations where the system size makes more exact techniques infeasible. Even so, we report the largest system ever used in elastic

network modeling through one of the most popular elastic network model software programs. In Part II, “**Brownian dynamics, molecular dynamics and Markov state models**”, we reduce our size scale and use Brownian dynamics, molecular dynamics and Markov state models to investigate small molecule binding and glycoprotein flexibility in influenza. These are more exact methods where we can ask and answer more detailed questions with smaller systems, dealing with glycans’ involvement in ligand binding, protein stability and protein dynamics. In Part III, “**Density functional theory**”, we reduce our size scale even further and use density functional theory to create molecular dynamics force field parameters for a newly-discovered bacterial protein. This requires extremely precise calculations to be made, where errors in the assigned electronic charge to an atom can cause large errors in calculations.

These chapters, and the work therein, are collaborative studies. Specific author contributions are detailed within each chapter. The rest of this introduction provides an overview of the techniques used in the chapters.

1.2 Elastic network models

Proteins come in a wide variety of shapes that beget a wide variety of functions; understandably the interplay between these shapes and functions is governed by their dynamics. These dynamics engender a wide variety of conformational changes, some large and some small. Moreover, the speeds at which the proteins can change cover many timescales, from the order of nanoseconds to the order of hours and longer. The timescales, and the corresponding protein conformational transitions, limit the methods used to study them and the computing power required. One way to reach longer timescale motions that we used in this work is to simplify the system we are studying. Computer simulation, and more specifically normal mode analysis, reduces proteins, composed of amino acids which are in turn composed of atoms which are composed of electrons, protons and neutrons, to a system of atomic centers interacting by a simplified molecular dynamics (MD) potential energy function.¹⁻² Elastic network models simplify these models by reducing the potential energy function to harmonic spring potentials; i.e. a set of springs connecting the alpha carbons of nearby residues.³⁻⁵ Anisotropic network models (ANM) then advance this further by introducing a more detailed model for local interactions.^{2, 6} This allows for estimation of longer timescale dynamics, at the expense of glossing over short timescale motions. A different adaptation of elastic network models are Gaussian network models (GNM).⁷⁻¹⁰ In the GNM scheme we

assume that fluctuations are isotropic, not anisotropic, which means that the magnitude of the fluctuations are independent of the direction of measurement. These models also assume that motion follows a Gaussian distribution, which allows a decomposition of the interparticle interaction matrix to reduce from three dimensions to two, along with a change in the interparticle potential energy function. This results in GNMs being better at predicting long timescale motions, their size and cross-correlations, but they cannot predict the direction of motion since it is only a two dimensional analysis. In practice, one can use GNMs for measuring which regions of a protein display correlated motions, but one cannot tell the direction of those motions.¹¹⁻¹³

1.3 Brownian dynamics

Brownian dynamics simulates the diffusion of ligands, and can be used to measure their (diffusive) association rates to proteins. In this regime the protein may remain stationary; the diffusing ligand displays translational and rotational motion as a particle, but - in rigid models - no bonds are allowed to change length or angle. With these limitations, Brownian dynamics can generate association rates much faster than can be acquired through the much more computationally costly molecular dynamics. These association rates can be used for understanding if some proteins or ligands diffuse/associate faster than others, and thus would be more likely to bind, be it protein-protein binding or protein-ligand binding.¹⁴ However, MD provides a host of other analysis functions apart from generating association rates.

1.4 Molecular dynamics

MD is a method of propagating particles (usually atoms in small molecules or proteins) over time to investigate conformational changes, chemical interactions, and much more.¹⁵⁻¹⁸ In MD, there are several different ensembles that can be used while propagating motion:

Ensemble	Description	Thermodynamic function
Microcanonical ensemble	Isolated system, no energy or mass transfer. Also known as NVE ensemble.	$S(U, V, N)$ entropy as a function of energy, volume and number of particles.

Canonical ensemble	System in contact with a heat bath. $A(T, V, N)$ Helmholtz free energy Energy transfer but no mass as a function of temperature, transfer. Also known as NVT volume and number of particles. ensemble.
Grand canonical ensemble	System in contact with a heat and $Z(\mu, V, T)$ other physical quantities particle bath. Energy and mass as a function of chemical potential, transfer. Also known as μVT volume and temperature. ensemble.
Isothermal-isobaric ensemble	System under constant $\Delta(N, P, T)$ function of number of temperature and pressure. Energy particles, pressure and and volume transfer but no mass temperature. transfer. Also known as NPT ensemble.

NPT is useful because it allows for direct comparison with experimental data. Additionally, it is useful when the “correct” volume of the simulation is not known, such as in initiating simulations of solvated proteins. MD is used for a wide variety of purposes, from examining protein¹⁹⁻²⁰ or nucleic acid²¹⁻²² stability to investigating protein-ligand²³⁻²⁵ or protein-protein²⁶ interactions to measuring bulk sheet properties²⁷⁻²⁸ to exploring protein conformational transitions²⁹⁻³⁰ to tracking redox cycles³¹⁻³² and many more. However, MD is limited in that bonds cannot be formed or broken with the most commonly used potential energy functions. Additionally, MD applied to proteins is sometimes limited by the available computing power, which may not be sufficient to see some protein conformational transitions, and certainly not see the rate at which these transitions happen. This can be accomplished with Markov state models (MSMs).

1.5 Markov state models

Rare conformational transitions can sometimes be seen through MD, and sometimes these transitions cannot be seen. However, one can break up MD trajectories into components encompassing the transition, and the

rebuild the transition and its associated kinetics through MSMs. This type of analysis can extract longer timescale dynamics from MD trajectories while using the same data. This is based off the assumption that one can recover the probability that a transition will happen given enough instances where this condition is met; knowing this probability then allows one to know the speed at which this condition will be met (i.e. the speed at which a conformational transition will occur). This allows one to create conformational pathways and see the speed that the transitions between these pathways occurs at.

1.6 Density functional theory

Quantum mechanics (QM) offers a more accurate description of the energetics of conformational changes than the force fields models normally used in MD, but it comes with a correspondingly significant leap in computational costs. Thus it can be used for small molecules but cannot easily be used on proteins or other larger systems. Inside the QM regime, we used density functional theory (DFT). This is a method of computing the electron density present in molecules, and predicting how this density will determine the molecular properties, be it structure, bonding, or electrostatics. Considering the detailed treatment of electrons, DFT has many uses in studying electron transfer pathways, electronic structure of materials³³⁻³⁴, electron orbital properties (such as hydrogen bonding)³⁵, molecule reactivity preferences³⁶, absorption/emission spectra³⁷⁻³⁸, and many more. One use for DFT is for determining force field parameters that can be used in MD simulations, which would then allow for a host of other questions to be answered.

References

1. Bauer, Jacob A.; Pavlović, Jelena; Bauerová-Hlinková, Vladena, Normal Mode Analysis as a Routine Part of a Structural Investigation. *Molecules* **2019**, *24* (18).
2. López-Blanco, José Ramón; Miyashita, Osamu; Tama, Florence; Chacón, Pablo, Normal Mode Analysis Techniques in Structural Biology. *eLS* **2014**.
3. Bahar, Ivet; Lezon, Timothy R.; Yang, Lee-Wei; Eyal, Eran, Global Dynamics of Proteins: Bridging Between Structure and Function. *Annual Review of Biophysics* **2010**, *39*, 23-42.
4. Lezon, Timothy R.; Shrivastava, Indira H.; Yang, Zheng; Bahar, Ivet, Elastic Network Models For Biomolecular Dynamics: Theory and Application to Membrane Proteins and Viruses. *Handbook on Biological Networks* **2009**, 129-158.
5. Leioatts, Nicholas; Romo, Tod D.; Grossfield, Alan, Elastic Network Models Are Robust to Variations in Formalism. *Journal of Chemical Theory and Computation* **2012**, *8* (7), 2424-2434.
6. Atilgan, Ali Rana; Durrell, Stewart R.; Jernigan, Robert L.; Demirel, Melik C.; Keskin, Özlem; Bahar, Ivet, Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophys J* **2001**, *80*, 505-515.
7. Hu, Guang, Identification of Allosteric Effects in Proteins by Elastic Network Models. In *Allostery: Methods and Protocols*, Paola, Luisa di; Giuliani, Alessandro, Eds. Humana: New York, NY, 2021; Vol. 2253, pp 21-35.
8. Bahar, Ivet; Cheng, Mary Hongying; Lee, Ji Young; Kaya, Cihan; Zhang, She, Structure-Encoded Global Motions and Their Role in Mediating Protein-Substrate Interactions. *Biophys J* **2015**, *109* (6), 1101-1109.
9. Erman, Burak, The Gaussian Network Model: Precise Predictions of Residue Fluctuations and Application to Binding Problems. *Biophys J* **2006**, *91*, 3589-3599.
10. Tekpinar, Mustafa; Yildirim, Ahmet, Only a Subset of Normal Modes is Sufficient to Identify Linear Correlations in Proteins. *Journal of Chemical Information and Modeling* **2018**, *58* (9), 1947-1961.
11. Khairallah, Afrah; Ross, Caroline J.; Bishop, Özlem Tastan, Probing the Structural Dynamics of the Plasmodium falciparum Tunneling-Fold Enzyme 6-Pyruvoyl Tetrahydropterin Synthase to Reveal Allosteric Drug Targeting Sites. *Frontiers in Molecular Biosciences* **2020**, *7*.
12. Luca, Andrea Di; Kaila, Ville R. I., Global collective motions in the mammalian and bacterial respiratory complex I. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **2018**, *1859* (5).
13. Liang, Zhongjie; Hu, Junchi; Yan, Wenying; Jiang, Hualiang; Hu, Guang; Luo, Cheng, Deciphering the role of dimer interface in intrinsic dynamics and allosteric pathways underlying the functional transformation of DNMT3A. *Biochimica et Biophysica Acta (BBA) - General Subjects* **2018**, *1862* (7), 1667-1679.
14. Sadiq, S. Kashif; Chicharro, Abraham Muñoz; Friedrich, Patrick; Wade, Rebecca C., Multiscale Approach for Computing Gated Ligand Binding from Molecular Dynamics and Brownian Dynamics Simulations. *Journal of Chemical Theory and Computation* **2021**, *17* (12), 7912-7929.
15. Braun, Efrem; Gilmer, Justin; Mayes, Heather B.; Mobley, David L.; Monroe, Jacob I.; Prasad, Samarjeet; Zuckerman, Daniel M., Best Practices for Foundations in Molecular Simulations. *Living Journal of Computational Molecular Science* **2019**, *1* (1).
16. Jr., Freddie R. Salsbury, Molecular dynamics simulations of protein dynamics and their relevance to drug discovery. *Curr Opin Pharmacol* **2010**, *10*, 738-744.
17. Zwier, Matthew C.; Chong, Lilian T., Reaching biological timescales with all-atom molecular dynamics simulations. *Curr Opin Pharmacol* **2010**, *10*, 745-752.
18. Amaro, Rommie E.; Mulholland, Adrian J., Bridging Chemical and Biological Complexity in the Search for Cures: Multiscale Methods in Drug Design. *Nature Reviews Chemistry* **2018**, *2*.
19. F.Coêlho, Danilo; Ferraz, Matheus V. F.; Marques, Ernesto T. A.; Lins, Roberto D.; Viana, Isabelle F. T., The influence of biotinylation on the ability of a computer designed protein to detect B-cells producing anti-HIV-1 2F5 antibodies. *Journal of Molecular Graphics and Modeling* **2019**, *93*.
20. Shehata, Mohamed; Timucin, Emel; Venturini, Alessandro; Sezerman, Osman Uğur, Understanding thermal and organic solvent stability of thermoalkalophilic lipases: insights from computational predictions and experiments. *Journal of Molecular Modeling* **2020**, *26* (122).
21. Rallapalli, Kartik L.; Komor, Alexis C.; Paesani, Francesco, Computer simulations explain mutation-induced effects on the DNA editing by adenine base editors. *Science Advances* **2020**, *6* (10).

22. Rallapalli, Kartik L.; Ranzau, Brodie L.; Ganapathy, Kaushik R.; Paesani, Francesco; Komor, Alexis C., Combined Theoretical, Bioinformatic, and Biochemical Analyses of RNA Editing by Adenine Base Editors. *CRISPR Journal* **2022**, 5 (2).
23. Lai, Thanh T.; Eken, Yiğitcan; Wilson, Angela K., Binding of Per- and Polyfluoroalkyl Substances to the Human Pregnane X Receptor. *Environ Sci Technol* **2020**, 54 (24), 15986-15995.
24. Gaieb, Zied; Lo, David D.; Morikis, Dimitrios, Molecular Mechanism of Biased Ligand Conformational Changes in CC Chemokine Receptor 7. *Journal of Chemical Information and Modeling* **2016**, 56 (9), 1808-1822.
25. Almeida, Nuno M. S.; Eken, Yiğitcan; Wilson, Angela K., Binding of Per- and Polyfluoro-alkyl Substances to Peroxisome Proliferator-Activated Receptor Gamma. *ACS Omega* **2021**, 6 (23), 15103-15114.
26. Takahashi, Yohei; Bosmans, Krystal C.; Hsu, Po-Kai; Paul, Karnelia; Seitz, Christian; Yeh, Chung-Yueh; Wang, Yuh-Shuh; Yarmolinksy, Dmitry; Sierla, Maija; Vahisalu, Triin; McCammon, J. Andrew; Kangasjärvi, Jaako; Zhang, Li; Kollist, Hannes; Trac, Thien; Schroder, Julian I., Stomatal CO₂/bicarbonate sensor consists of two interacting protein kinases, Raf-like HT1 and non-kinase-activity activity requiring MPK12/MPK4. *Science Advances* **2022**, 8 (49).
27. Cruzeiro, Vinícius Wilian D.; Galib, Mirza; Limmer, David T.; Götz, Andreas W., Uptake of N₂O₅ by aqueous aerosol unveiled using chemically accurate many-body potentials. *Nature Communications* **2022**, 13 (1266).
28. Shehata, Mohamed; Unlu, Aise; Sezerman, Ugur; Timucin, Emel, Lipase and Water in a Deep Eutectic Solvent: Molecular Dynamics and Experimental Studies of the Effects of Water-In-Deep Eutectic Solvents on Lipase Stability. *Journal of Physical Chemistry B* **2020**, 124 (40), 8801-8810.
29. Gaieb, Zied; Morikis, Dimitrios, Conformational heterogeneity in CCR7 undergoes transitions to specific states upon ligand binding. *Journal of Molecular Graphics and Modelling* **2017**, 74, 352-358.
30. Gaieb, Zied; Morikis, Dimitrios, Detection of Side Chain Rearrangements Mediating the Motions of Transmembrane Helices in Molecular Dynamics Simulations of G Protein-Coupled Receptors. *Computational and Structural Biotechnology Journal* **2017**, 15, 131-137.
31. Cruzeiro, Vinícius Wilian D.; Feliciano, Gustavo Troiano; Roitberg, Adrian E., Exploring Coupled Redox and pH Processes with a Force-Field-Based Approach: Applications to Five Different Systems. *J Am Chem Soc* **2020**, 142 (8), 3823-3835.
32. Cruzeiro, Vinícius Wilian D.; Roitberg, Adrian E., Multidimensional Replica Exchange Simulations for Efficient Constant pH and Redox Potential Molecular Dynamics. *Journal of Chemical Theory and Computation* **2019**, 15 (2), 871-881.
33. Pasumarthi, Viswanath; Liu, Taifeng; Dupuis, Michel; Li, Can, Charge carrier transport dynamics in W/Mo-doped BiVO₄: first principles-based mesoscale characterization. *Journal of Materials Chemistry A* **2019**, 7, 3054-3065.
34. Yang, Weizhun; Ramadan, Sherif; Orwenyo, Jared; Kakeshpour, Tayeb; Diaz, Thomas; Eken, Yigitcan; Sanda, Miloslav; Jackson, James E.; Wilson, Angela K.; Huang, Xuefei, Chemoenzymatic synthesis of glycopeptides bearing rare N-glycan sequences with or without bisecting GlcNAc. *Chemical Science* **2018**, 9, 8194-8206.
35. Mohan, Mood; Viswanath, Pasumarthi; Banerjee, Tamal; Goud, Vaibhav V., Multiscale modelling strategies and experimental insights for the solvation of cellulose and hemicellulose in ionic liquids. *Molecular Physics* **2018**, 116 (15-16), 2108-2128.
36. Seitz, Christian G.; Zhang, Huaiyu; Mo, Yirong; Karty, Joel M., Why Do Enolate Anions Favor O-Alkylation over C-Alkylation in the Gas Phase? The Roles of Resonance and Inductive Effects in the Gas-Phase S_N2 Reaction between the Acetaldehyde Enolate Anion and Methyl Fluoride. *J Org Chem* **2016**, 81 (9), 3711-3719.
37. Turan, Haydar Taylan; Eken, Yiğitcan; Marazzi, Marco; Pastore, Mariachiara; Aviyente, Viktorya; Monari, Antonio, Assessing One- and Two-Photon Optical Properties of Boron Containing Arenes. *Journal of Physical Chemistry C* **2016**, 120 (32), 17916-17926.
38. Cruzeiro, Vinícius Wilian D.; Wildman, Andrew; Li, Xiaosong; Paesani, Francesco, Relationship between Hydrogen-Bonding Motifs and the 1b₁ Splitting in the X-ray Emission Spectrum of Liquid Water. *Journal of Physical Chemistry Letters* **2021**, 12 (16), 3996-4002.

Part I

Elastic network models

Chapter 2

Architecture and self-assembly of the jumbo bacteriophage nuclear shell

2.1 Abstract

Bacteria encode myriad defenses that target the genomes of infecting bacteriophage, including restriction-modification and CRISPR-Cas systems.¹ In response, one family of large bacteriophages employs a nucleus-like compartment to protect their replicating genomes by excluding host defense factors.²⁻⁴ However, the principal composition and structure of this compartment remain unknown. Here, we find that the bacteriophage nuclear shell assembles primarily from one protein, termed chimallin. Combining cryo-electron tomography of nuclear shells in bacteriophage-infected cells and cryo-electron microscopy of a minimal chimallin compartment *in vitro*, we show that chimallin self-assembles as a flexible sheet into closed micron-scale compartments. The architecture and assembly dynamics of the chimallin shell suggest mechanisms for its nucleation and growth, and its role as a scaffold for phage-encoded factors mediating macromolecular transport, cytoskeletal interactions, and viral maturation.

2.2 Main

Over billions of years of conflict with bacteriophages (phages), plasmids, and other mobile genetic elements, bacteria have evolved an array of defensive systems to target and destroy foreign nucleic acids.¹ Phages have in turn evolved mechanisms, including anti-restriction and anti-CRISPR proteins, that counter specific bacterial defense systems.⁵⁻⁷ We recently showed that a family of “jumbo phages” – named for their large genomes (typically >200 kb) and large virion size (~145 nm capsids, ~200 nm contractile tails) – assemble a selectively-permeable, protein-based shell that encloses the replicating viral genome and is associated with a unique phage life cycle, represented in (Fig. 2.1a).^{2, 8} This micron-scale nucleus-like compartment, termed the “phage nucleus”, forms *de novo* upon infection and grows with the replicating viral DNA. Meanwhile, phage proteins are synthesized in the host cell cytoplasm. These include PhuZ, a phage-encoded tubulin homolog that assembles into filaments that treadmill to transport empty capsids from their assembly sites at the host cell membrane to

the surface of the phage nucleus. Capsids dock to the phage nucleus shell and are filled with viral DNA before detaching and completing assembly with phage tails. Mature particles are released by host cell lysis. In contrast to other characterized anti-restriction systems, the phage nuclear shell renders jumbo phages broadly immune to DNA-targeting host restriction systems, including CRISPR-Cas, throughout infection by serving as a physical barrier between the viral DNA and host nucleases.³⁻⁴

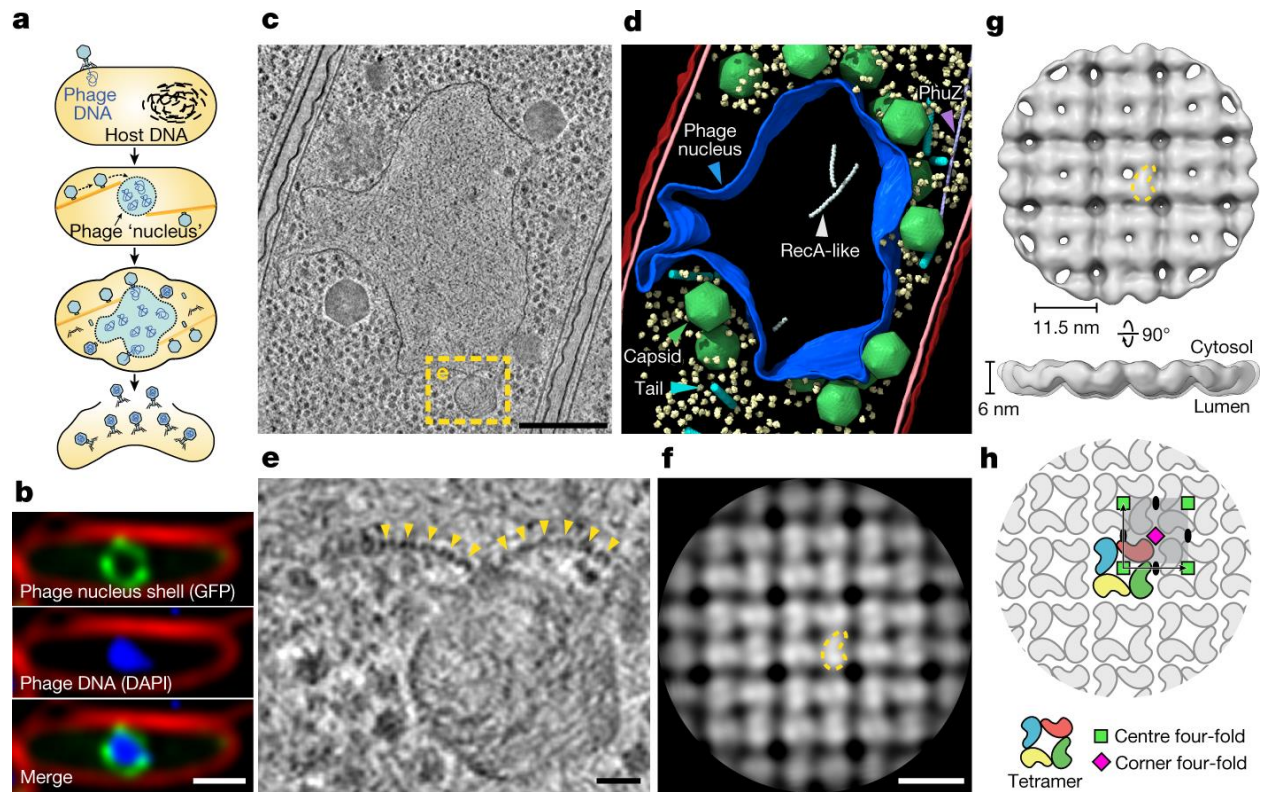


Figure 2.1 | In situ tomography and subtomogram analysis of the 201phi2-1 phage nucleus.

a, Schematic of the jumbo phage infection cycle. **b**, Fluorescence microscopy of a 201phi2-1-infected *P. chlororaphis* cell at 45 mpi ($n = 5$ independent experiments). Phage nucleus shell component, gp105, (green) is tagged with GFP, phage DNA (blue) is stained with DAPI, and the outer cell membrane (red) is stained with FM4-64. **c**, Tomographic slice of a phage nucleus in a 201phi2-1-infected *P. chlororaphis* cell at 50-60 mpi. **d**, Segmentation of the tomogram in **c**. Outer and inner bacterial membranes are burgundy and pink, respectively. The phage nucleus is colored blue. Phage capsids and tails are green and cyan, respectively. PhuZ and recA-like protein filaments are light purple and white, respectively. A subset of five-hundred host ribosomes are shown in pale-yellow. **e**, Enlarged view of the boxed region in **c**. Yellow arrows point to the repetitive feature of the phage nucleus perimeter. **f**, Slice of the cytosolic face of the subtomogram average of the repetitive feature in the phage nucleus perimeter with a comma-shaped subunit outlined in yellow. **g**, Cytosolic and side views of the shell subtomogram average isosurface with a single subunit outlined in yellow. **h**, Schematic representation of the *p442*-like arrangement of chimallin protomers. Scale bars: **b**:1 μm , **c**:250 nm, **e**:25 nm, **f**:10 nm.

While we previously showed that the phage nuclear shell incorporates at least one abundant phage-encoded protein,^{2, 8} the overall composition and architecture of this structure are still largely unknown. Also unclear is

how these phages address the challenges arising from the separation of transcription and translation, specifically the need for directional transport of mRNA out of the phage nucleus and transport of DNA processing enzymes into the phage nucleus.² Finally, how genomic DNA produced in the phage nucleus is packaged into capsids assembled in the cytosol is also unknown.⁹

Architecture of the phage nuclear shell

To gain insight into the native architecture of the phage nuclear shell, we performed focused ion-beam milling coupled with cryo-electron tomography (cryoFIB-ET) of *Pseudomonas chlororaphis* 200-B cells (henceforth, *P. chlororaphis*) infected with jumbo phage 201phi2-1 (317 kb genome) at 50-60 minutes post infection (mpi, typical time to lysis is around 90 mpi) (**Fig. 2.1, Fig. 2.6a-i**). The observed phage nuclei were pleomorphic compartments devoid of ribosomes, bounded by a ~6 nm thick proteinaceous shell (**Fig. 2.1b-d, Fig. 2.6**). Close inspection of the compartment perimeter revealed repeating doublets of globular densities with ~11.5 nm spacing, suggesting that the shell consists of a single layer of proteins in a repeating array (**Fig. 2.1e, Fig. 2.6j**). Furthermore, we occasionally captured face-on views that exhibit a square lattice of densities with the same repeat spacing, reinforcing the idea of a repeating array of protomers (**Fig. 2.6k**).

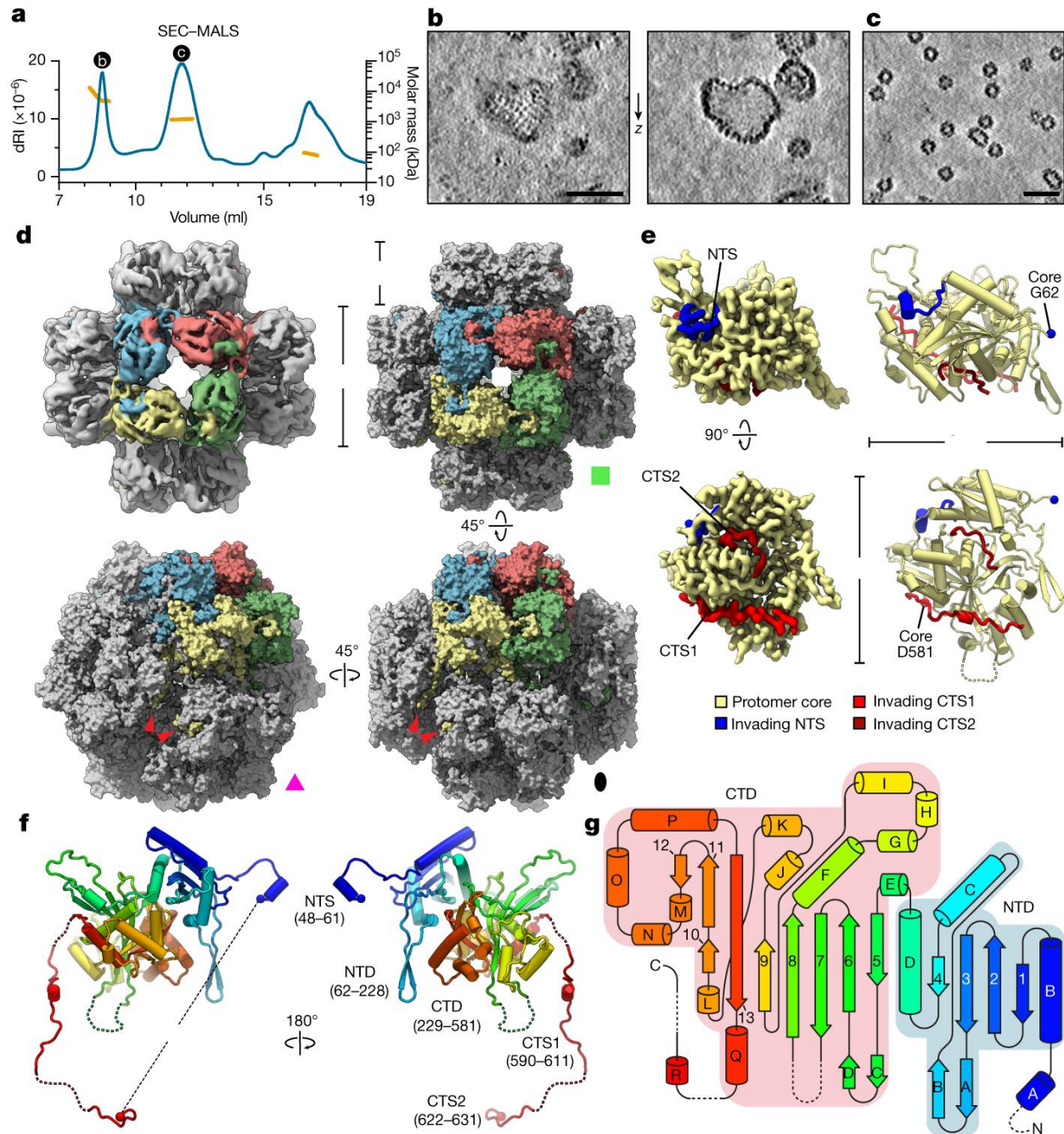
Using subtomogram analysis to average over low-curvature regions from eight separate 201phi2-1 nuclei, we obtained a ~24 Å resolution reconstruction of the phage nuclear shell (**Fig. 2.1f,g, Fig. 2.6h-o & Table 2.1**). The reconstruction reveals the shell as a quasi-*p4*, or square, lattice (quasi because of variable curvature). The repeating unit is an 11.5 x 11.5 nm square tetramer approximately 6 nm thick, with internal four-fold rotational symmetry. These units form a square lattice with a second 4-fold rotational symmetry axis at the corner of each square unit and 2-fold axes on each side (the *p442* wallpaper group; **Fig. 2.1h, Fig. 2.6m,o**). The four individual protomer densities within each 11.5 x 11.5 nm unit measure ~6 x 6 x 7.5 nm, dimensions consistent with a ~70 kDa protein. Thus, the phage nuclear shell appears to be predominantly composed of a single protein component arranged in a square lattice.

Chimallin is the principal shell protein

We previously showed that the abundant and early-expressed 201phi2-1 protein gp105 becomes integrated into the nuclear shell.² Furthermore, 201phi2-1 gp105 has a molecular weight of 69.5 kDa (631 amino acids), consistent with the size of an individual protomer density from our *in situ* cryo-ET map. Along with the high apparent compositional homogeneity of the shell as observed by cryo-ET, these data support 201phi2-1 gp105 as the principal component of the phage nuclear shell. Homologs of 201phi2-1 gp105 are encoded by a large set of jumbo phage that infect diverse bacteria including *Pseudomonas*, *Vibrio*, *Salmonella*, and *Escherichia coli*, but these proteins bear no detectable sequence homology to any other proteins. Because of its role in protecting the phage genome against host defenses, we term this protein chimallin (ChmA) after the *chimalli*, a shield carried by ancient Aztec warriors.¹⁰

To understand the structure and assembly mechanisms of chimallin, we expressed and purified 201phi2-1 chimallin from *E. coli*. Size-exclusion chromatography and multi-angle light scattering (SEC-MALS) of purified chimallin indicated a mixture of oligomeric states including monomers, well-defined assemblies of approximately 1.2 MDa, and larger heterogeneous species ranging from 4 to 13 MDa (mean=6.9 MDa; **Fig. 2.2a**). Cryo-ET of the largest species revealed pleomorphic, closed compartments with near-identical morphology to the phage nuclear shell we observe *in situ* (**Fig. 2.2b**, **Fig. 2.7**). Analysis of the smaller, more defined assemblies by cryo-ET revealed a near-homogeneous population of cubic assemblies with a diameter of ~22 nm, and a minor population of rectangular assemblies with dimensions ~22 x 33.5 nm (**Fig. 2.2c**).

Figure 2.2 | *In vitro* cryo-EM structure of the 201phi2-1 phage nuclear shell protein chimallin. **a**, Size-exclusion coupled to multi-angle light scattering (SEC-MALS) analysis of purified 201phi2-1 chimallin. From left to right, the measured molar mass of three peaks are 6.9 MDa (range from 4-13 MDa), 1.2 MDa, and 87 kDa. **b,c**, Z-slices from tomograms of samples from the correspondingly labeled SEC-MALS peaks in **a**. Full field of view of **b** supplied as Extended Data Fig. 2. **d**, Top-left, O-symmetrized reconstruction of the chimallin cubic assembly viewed along the 4-fold axis. The protomers of one 4-fold face are colored. Top-right, surface representation of the chimallin cubic assembly model viewed along the 4-fold axis. Bottom-right and bottom-left, views of the model along the 2- and 3-fold axes, respectively. Red arrows point to the C-terminal segments of the yellow protomer. **e**, Localized asymmetric reconstruction of the chimallin protomer (left) and cartoon model (right). Invading N- and C-terminal segments from neighboring protomers are colored blue (NTS), red (CTS1), and burgundy (CTS2). Resolved core protomer termini are shown as spheres. **f**, Rainbow colored cartoon model of the Chimallin protomer conformation in the cubic assembly. Resolved N- and C-termini are shown as spheres. Domains and segments are labelled. Unresolved linkers are shown as dashed lines. **g**, A rainbow colored fold diagram of chimallin (blue at N-terminus, red at C-terminus) with α -helices labelled alphabetically and β -strands labelled numerically. The N- and C-terminal domains are highlighted blue and red, respectively. Dashed lines indicate unresolved loops. Scale bars: **b,c**: 50 nm.



We next acquired cryo-electron microscopy (cryo-EM) data and performed single-particle analysis (SPA) of the defined chimallin assemblies (**Fig. 2.8**). Two-dimensional class averages revealed that each cubic particle consists of six chimallin tetramers (24 protomers, 1.67 MDa) arranged to form a minimal closed compartment with apparent octahedral (O, 432) symmetry (**Fig. 2.8a**). Similarly, the minor population of $\sim 22 \times 33.5$ nm

rectangular particles are assemblies of ten chimallin tetramers (40 protomers, 2.78 MDa) with apparent D4 symmetry (**Fig. 2.8g**). Each tetrameric unit in these assemblies is an 11.5 x 11.5 nm square, in line with our *in situ* subtomogram analysis of the phage nuclear shell. Three-dimensional reconstruction of the cubic particles with enforced O symmetry resulted in a ~4.4 Å density map with distorted features (**Fig. 2.2d**), likely arising from inherent plasticity of these assemblies breaking symmetry. Localized reconstruction of the square faces from each particle with C4 symmetry resulted in an improved density map at ~3.4 Å. Further reduction of the structure to focus on an individual protomer resulted in the highest quality map at ~3.1 Å, enabling atomic modelling of the chimallin protomer (**Fig. 2.2e,f, Tables 2.2-2.4**).

Chimallin folds into a compact two-domain core with extended N- and C-terminal segments (**Fig. 2.2f,g**). The N-terminal domain (residues 62-228) shows little structural homology to any characterized protein, adopting an $\alpha+\beta$ fold that is topologically similar only to an uncharacterized protein from *Enterococcus faecalis* (**Fig. 2.9a,b**). The C-terminal domain (residues 229-581) adopts a GCN5-related N-acetyltransferase fold most similar to that of *E. coli* AtaT and related tRNA acetylating toxins (PDB: 6AJM, RMSD 4.2 Å for 269 aligned Ca atom pairs) (**Fig. 2.9c-e**).¹¹ While chimallin lacks the acetyltransferase active site residues of AtaT and related toxins, this structural similarity suggests that the jumbo phage nuclear shell may have evolved from a bacterial toxin-antitoxin system.

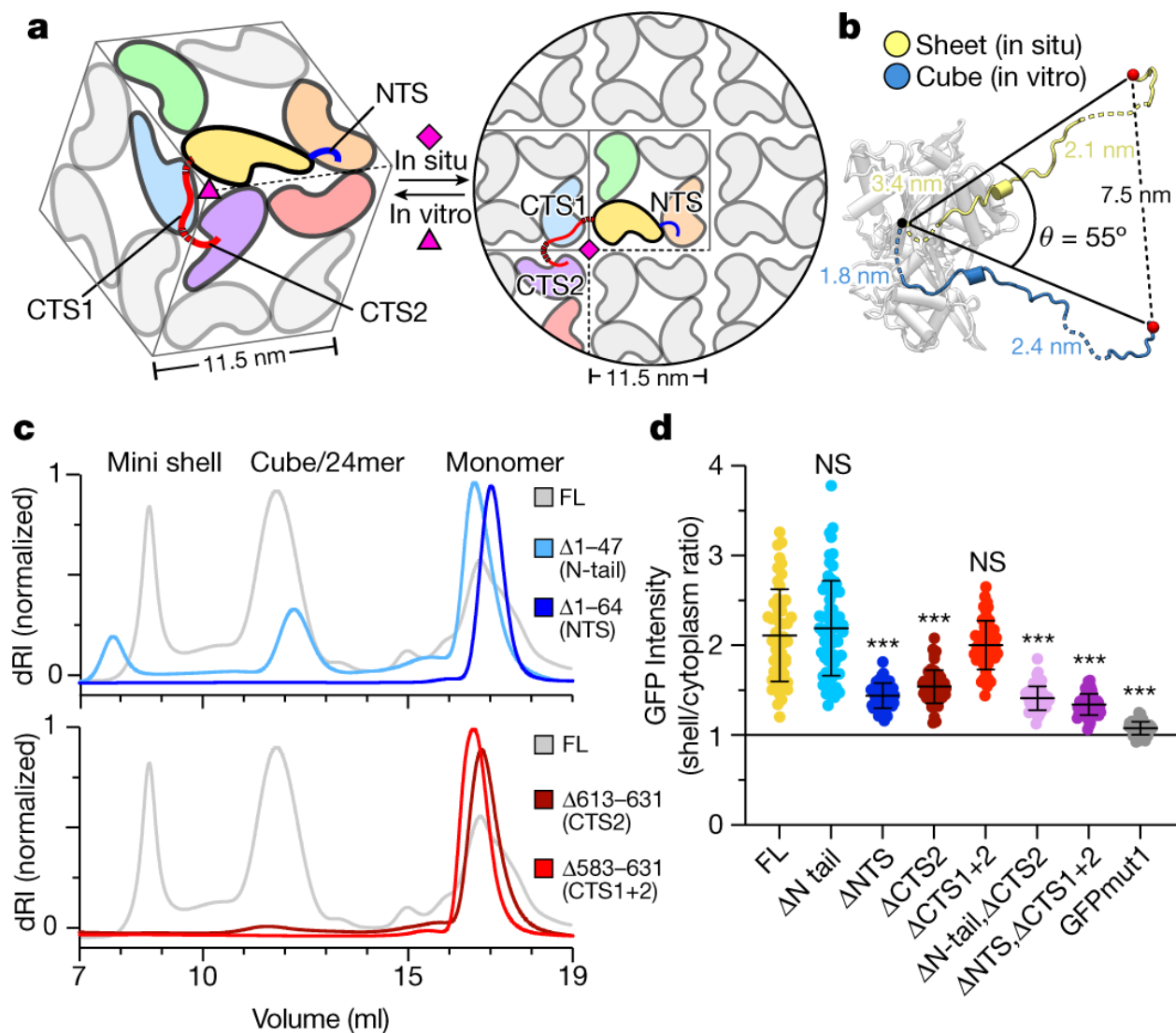
Atomic models of the C4-symmetric chimallin tetramer and quasi-O symmetric cubic assemblies reveal the molecular basis for chimallin self-assembly. The chimallin protomer map contains three interacting peptide segments from the N- and C-termini of neighboring protomers (**Fig. 2.2d-f**). Using the maps for the tetrameric face and full cubic assembly, we built models for the higher-order chimallin oligomers to understand the subunit interconnectivity. The N-terminal interacting segment (NTS, residues 48-61) of each protomer extends counterclockwise (as viewed from outside the cube) and docks against a neighboring protomer's N-terminal domain within a given face of the cubic assembly, thus establishing intra-tetramer connections. Meanwhile, two extended segments of the C-terminus (CTS1: residues 590-611, and CTS2: residues 622-631) establish inter-tetramer interactions. While the linkers between the C-terminal domain and CTS1 (residues 582-589) and

between CTS1 and CTS2 (residues 612-621) are unresolved in our maps, we could confidently infer the path of each protomer's C-terminus within the cubic assembly. CTS1 extends from one protomer to a neighboring protomer positioned counter-clockwise around the three-fold symmetry axis of the cube, and CTS2 further extends counterclockwise to the third subunit around the same axis (**Fig. 2.2d**). Notably, the binding of CTS1 to the chimallin C-terminal domain resembles the interaction of the antitoxin AtaR with the AtaT acetyltransferase toxin (**Fig. 2.9e**),¹¹ further hinting that the phage nuclear shell could have evolved from a bacterial toxin-antitoxin system. Compared to the tightly-packed chimallin tetramers mediated by the well-ordered NTS region, the length and flexibility of the linkers between the chimallin C-terminal domain, CTS1, and CTS2 suggest that flexible inter-tetramer packing enables chimallin to assemble into structures ranging from a flat sheet to the observed cubic assembly.

Chimallin self-assembly and dynamics

To investigate the interconnectivity of chimallin protomers in the context of the phage nucleus, we docked copies of the high-resolution chimallin tetramer model into the *in situ* cryo-ET map. The chimallin tetramers fit well into our ~24 Å resolution cryo-ET map of the shell without clashes and with an overall map-model correlation coefficient (CC) of 0.56 (**Fig. 2.10a**). To accommodate a flat sheet structure, the three-fold symmetry axis at each corner of the cubic assembly must be “unfolded” into a four-fold symmetry axis (**Fig. 2.3a**). Since CTS1 and CTS2 mediate interactions across this symmetry axis in the cubic assembly, this change requires that CTS1 rotate ~55° relative to the chimallin protomer core, and that CTS2 rotate the same amount relative to CTS1 (**Fig. 2.3b**). In the resulting sheet, each chimallin C-terminus extends counterclockwise to contact the two neighboring subunits in the new four-fold symmetry axis at the corners of the tetrameric units. The distances spanned by each disordered linker are similar in the flat sheet and the cubic assembly (**Fig. 2.3b**). Thus, both N- and C-terminal interacting segments contribute to shell self-assembly, with the C-terminus in particular likely imparting significant structural plasticity to the phage shell while maintaining its overall integrity.

Figure 2.3 | Flexibly attached N- and C-terminal segments mediate self-assembly of the chimallin shell. a, Relationship protomer packing in the cubic/24mer assemblies (left) and flat sheet model (right). One protomer is colored yellow with NTS in blue and CTS1/CTS2 in red. Protomers interacting directly with this focal protomer are colored orange, green, blue, purple, and red. Non-interfacing protomers are gray. Red arrows indicate locations of unresolved linkers (red dashed lines), and pink symbols indicate 3- or 4-fold symmetry axes. **b,** Comparison of chimallin C-terminus conformation in the *in vitro* sheet (yellow) and *in situ* cube (blue). Distances spanned by each disordered segment (CTD-CTS1: residues 582-589; CTS1-CTS2: residues 612-621) in the two models are noted. **(c)** SEC-MALS of N- and C-terminal truncation mutants (Δ N-tail= Δ 1-47 (residues 48-631 present); Δ NTS= Δ 1-64 (residues 65-631 present); Δ CTS2= Δ 613-631 (residues 1-611 present); Δ CTS1+2= Δ 583-631 (residues 1-582 present). dRI: differential refractive index, see **Fig. 2.10** for molar mass measurements by SEC-MALS. **(d)** Relative incorporation of GFPmut1-chimallin variants into the 201phi2-1 phage nucleus of infected *P. chlororaphis* cells. Incorporation is calculated as the ratio of GFP fluorescence/pixel in the shell vs. outside the shell (**Fig. 2.11** for details). Data are the mean $\pm$ s.d. Statistical analysis performed was an unpaired t-test between a given variant and full-length (FL, n = 67 cells) ‘***’ indicates a calculated P-value < 0.0001: Δ N-tail (residues 48-631, n = 51, P = 0.4131) ; Δ NTS (residues 65-631, n = 53, P < 0.0001); Δ CTS1 (residues 1-611, n = 54, P < 0.0001); Δ CTS1+2 (residues 1-582, n = 50, P = 0.1884) ; Δ N-tail, Δ CTS2 (residues 48-611, n = 58, P < 0.0001); Δ NTS, Δ CTS1+2 (residues 65-582, n = 63, P < 0.0001); GFPmut1 (n = 60, P < 0.0001). Threshold for significance was Bonferroni-corrected to P < 0.007 to account for multiple hypothesis testing.



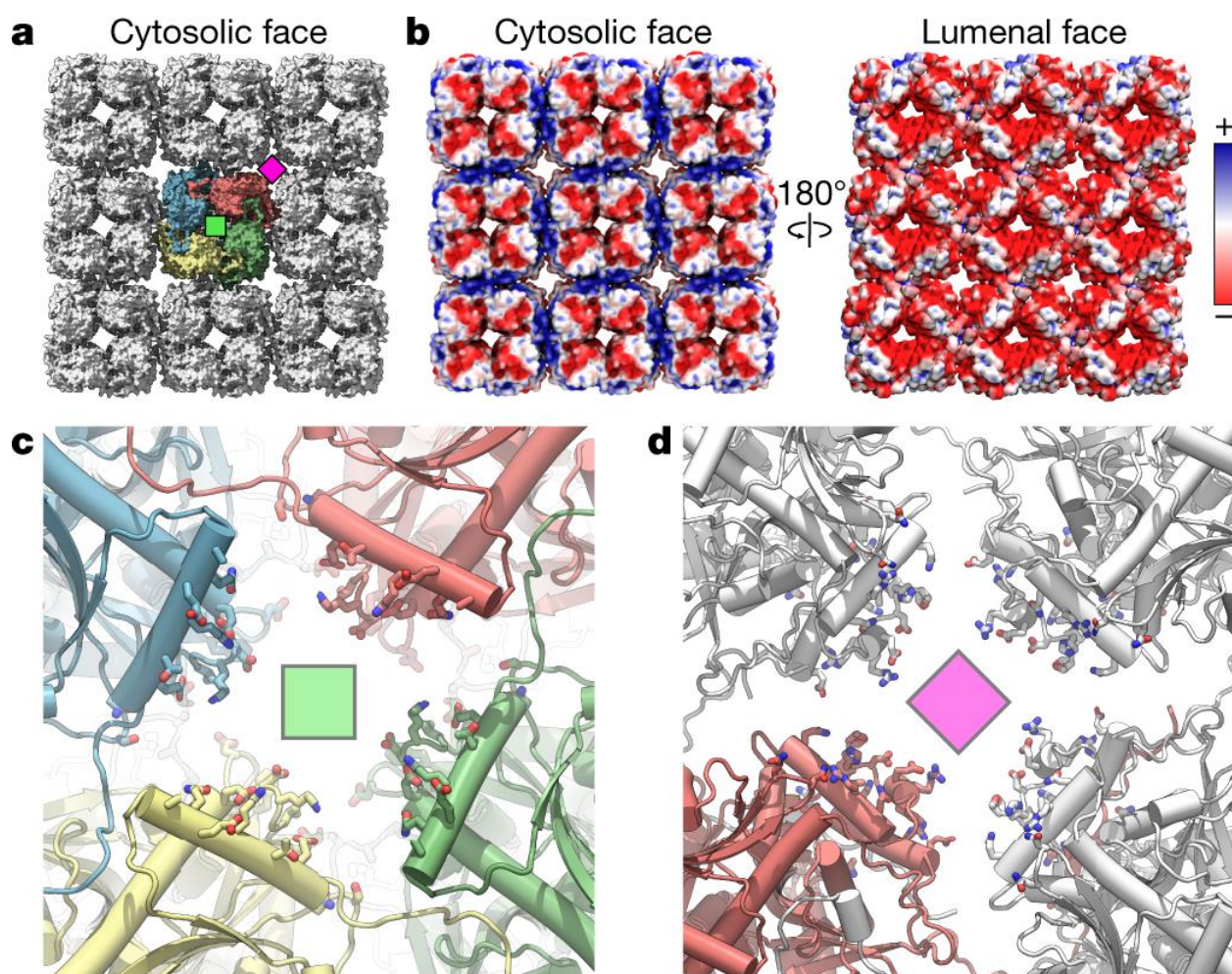
We next assessed the importance of the NTS and CTS regions for chimallin self-assembly both *in vitro* and *in vivo*. *In vitro*, deletion of either the NTS or CTS (CTS2 alone or CTS1+CTS2) completely disrupted self-assembly as measured by SEC-MALS (**Fig. 2.3c**, **Fig. 2.11**). We expressed the same truncations in phage 201phi2-1-infected *P. chlororaphis* cells, and measured incorporation of GFP-chimallin into the nuclear shell assembled by the phage-encoded full-length chimallin. We observed shell assembly in all cases, and found that deletion of the NTS, CTS, or both partially compromised incorporation into the shell (**Fig. 2.3d**, **Fig. 2.12a**). Consistent with this finding, expression of truncated chimallin did not strongly affect propagation of phage 201phi2-1 as measured by bacterial growth curves (**Fig. 2.12b**). Overall, these data support the idea that

chimallin NTS and CTS regions are important for efficient self-assembly of the chimallin shell, and show that self-assembly of full-length chimallin is robust even in the presence of chimallin protomers lacking these regions.

Flexibility of the chimallin shell

The phage nucleus shields the viral genome in a manner similar to a viral capsid. However, unlike viral capsids, chimallin does not tightly interact with the encapsulated DNA. Indeed, estimation of the electrostatics of chimallin indicate that both cytosolic and lumenal faces are negatively charged (**Fig. 2.4a,b**). The negative character of the phage nuclear shell likely mitigates interactions with the enclosed DNA, thereby keeping the genetic material accessible for transcription, replication, and capsid packaging.

Figure 2.4 | Electrostatics and pores of the chimallin shell. **a**, Surface model of the 3x3 chimallin tetramer lattice viewed from the cytosol with the central tetramer colored. A ‘center’ four-fold is indicated by a green square and ‘corner’ four-fold by a magenta square. **b**, Surface model of the cytosolic and luminal faces of the chimallin lattice colored by relative electrostatic potential. **c**, Cytosolic views of the center four-fold pore cartoon model with pore-facing residues (**Table 2.8**) shown as sticks. The center pores ($n = 9$) exhibit an average volume of $798 \pm 81 \text{ nm}^3$ over the course of 300-ns molecular dynamics simulations ($n = 5$, **Fig 2.13**). **d**, Same as in **c**, for the corner pore. The core pores ($n = 4$) exhibit an average volume of $1429 \pm 227 \text{ nm}^3$ over the course of simulations ($n = 5$ independent simulations, **Fig. 2.13**).



Again, in contrast to viral capsids and other protein-based organelles such as bacterial microcompartments, which form regular assemblies with defined facets, the phage nuclear shell adopts a highly irregular morphology (**Fig. 2.1c**, **Fig. 2.6b-i**). To identify the forms of conformational heterogeneity within the chimallin lattice, we performed an elastic network model analysis of a 3x3 tetramer chimallin sheet. This analysis not only indicated hinging along tetramer boundaries, which would lead to the cubic arrangement observed *in vitro*, but also hinging within a given tetramer (**Fig. 2.13a-f**). Prompted by this analysis, we performed focused classification with our *in situ* data which revealed three distinct classes of the central chimallin tetramer (**Fig. 2.6n,p** & **2.13j**). The predominant class shows a flat sheet, while two minor classes show the central unit raised (convex) or lowered (concave) by ~1 nm compared to surrounding units. Docking the tetramer model into maps representing concave, flat, and convex subpopulations indicated that the best fit is to the convex class (model-map CC = 0.64), and the

worst fit is to the concave class (model-map CC = 0.51). Moreover, close inspection of the density representing a chimallin protomer in the different classes revealed that each protomer rotates $\sim 25^\circ$ between the concave and convex classes, with the inner side of the tetramer pinching inward in the convex class (**Fig. 2.13i,j**). The tetramer model derived from the cubic assembly, which represents a highly convex state, shows a further $\sim 25^\circ$ inward tilt of each protomer. These observations suggest that the chimallin tetramer itself is flexible, with the C-terminal domains on the shell's inner face rotating inward in convex regions of the shell, and outward in concave regions. Thus, the morphology and flexibility of the phage nuclear shell likely derives from both intra- and inter-tetramer motions.

To analyze the inter-tetramer motions in the *in situ* data beyond the subtomograms that we extracted from relatively flat regions of the nuclear shell, we examined the flexibility between chimallin tetramers in *in situ* nuclear shells by calculating the curvature of an annotated surface in a tomogram and estimating the angle between tetramers. Our analysis showed that chimallin inter-tetramer conformations can range from slightly concave, i.e., -35° between neighboring tetramers, to highly convex, i.e., up to 80° between neighboring tetramers, close to the maximum bend of 90° observed in the purified sample (**Fig. 2.13k,l**). This enables the contorted shape observed in the cryo-ET data that presumably stems from addition of chimallin protomers to the lattice at multiple random locations, without distributing the strain throughout the lattice, rather than at a single privileged site.

Narrow pores in the chimallin shell

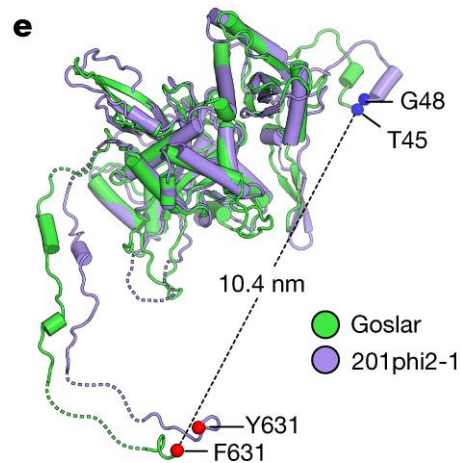
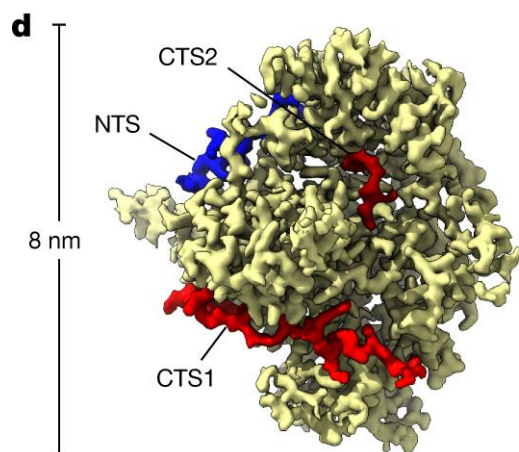
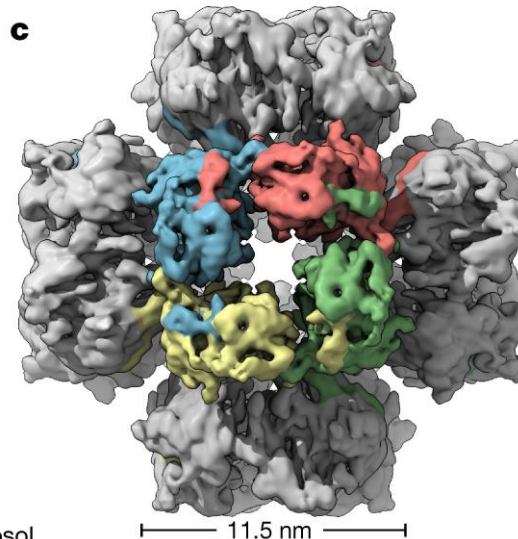
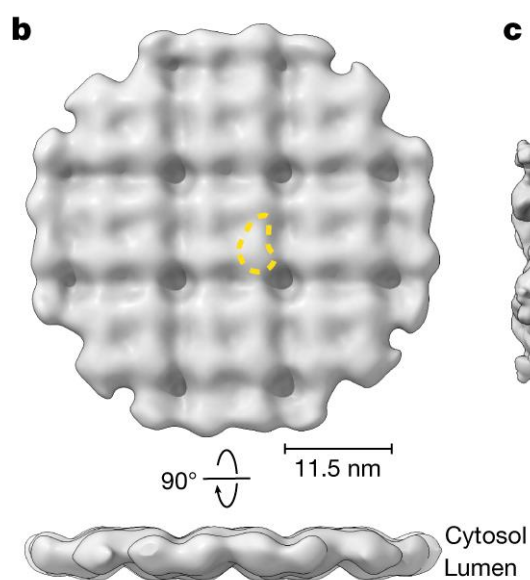
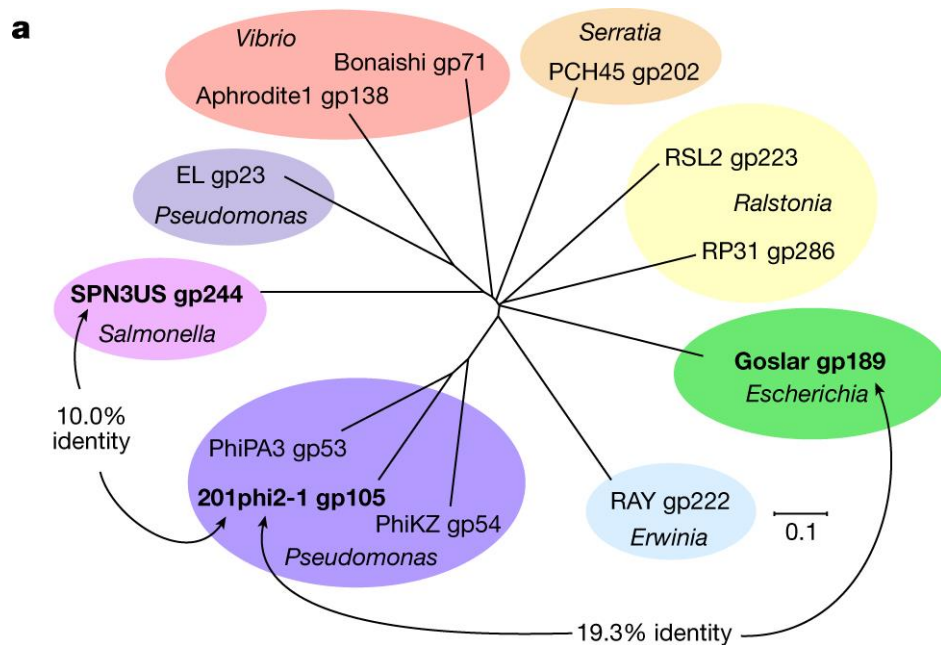
Like the eukaryotic nucleus, the phage nucleus separates transcription and translation.² Thus, the phage nuclear shell must accommodate trafficking of mRNA out of the nucleus, and that of specific proteins into the nucleus. To determine whether the pores at the two four-fold symmetry axes of the chimallin lattice could serve as conduits for macromolecule transport, we used all-atom molecular dynamics to simulate the motions of a 3x3 flat sheet of chimallin tetramers (**Fig. 2.4**). Assessing the variability in these pores through five separate 300 ns simulations, we found that the restrictive diameters of both the center and corner four-fold pores are ~ 1.4 nm on average, varying throughout the simulations from ~ 0 nm (i.e., closed) to as wide as 2.3 nm (**Fig. 2.14 & Table 2.6**). These data strongly suggest that the pores are too small to accommodate most folded proteins. However,

this pore size is sufficient to enable exchange of metabolites, nucleotides, and amino acids, and potentially large enough to support export of single-stranded mRNA molecules. An intriguing model for mRNA export is suggested by prior findings on viral capsid-resident RNA polymerases, which physically dock onto the inner face of the capsid and extrude mRNA co-transcriptionally through ~1.2-nm wide pores.¹²⁻¹³ Notably, chimallin-encoding jumbo phages have been shown to encode distinctive multi-subunit RNA polymerases, suggesting co-evolution of chimallin and transcriptional machinery in this family.¹⁴ Further study will be required to determine the protein(s) and sites responsible for protein import, as well as whether mRNA is directly extruded through the chimallin lattice pores.

Nuclear shell architecture is conserved

We recently discovered that the *E. coli* jumbo bacteriophage Goslar (237 kb genome) assembles a nuclear shell morphologically similar to those observed in the *Pseudomonas* phages 201phi2-1, PhiPA3, and PhiKZ.¹⁵ Goslar encodes a divergent homolog of 201phi2-1 chimallin (gp189, 631 amino acids), with 19.3% overall sequence identity between the two proteins (**Fig. 2.5a**). We performed cryoFIB-ET on *E. coli* cells infected with Goslar at mid-infection (see **Methods**), followed by subtomogram analysis of phage nuclei (**Fig. 2.15**). The resulting ~30 Å resolution reconstruction showed striking overall similarity to the structure of the 201phi2-1 nuclear shell, with a square grid of 11.5 x 11.5 nm units (**Fig. 2.5b, Fig. 2.15h,i**).

Figure 2.5 | Structural conservation of chimallin in the distantly related *E. coli* jumbo phage Goslar. **a**, Unrooted phylogenetic tree of chimallin homologs. Homologs are listed as phage and gene product (gp) numbers (see **Table 2.7**). Groups based on proximity are colored and host genus is labelled in italics (Scale bar: 0.1 substitutions/position). **b**, *In situ* subtomogram reconstruction of the Goslar chimallin shell. A comma-shaped protomer is marked by a yellow dashed outline and cytosolic and lumenal faces indicated. **c**, O-symmetrized map of the Goslar chimallin cubic assembly viewed along the 4-fold axis. **d**, Localized asymmetric reconstruction of the Goslar chimallin protomer. Invading N- and C-terminal segments from neighboring protomers are colored blue (NTS), red (CTS1), and burgundy (CTS2). **e**, Superposition of the Goslar (green) and 201phi2-1 (purple) coordinate models for the cube confirmation of the protomers **e**. Resolved termini are shown as spheres for the protomers in **e**. The root mean square deviations is 1.8 Å for the aligned protomers.



We next purified Goslar chimallin and characterized its self-assembly by SEC-MALS. Like 201phi2-1 chimallin, Goslar chimallin forms a mixture of monomers, defined assemblies of ~1.75 MDa, and large aggregates of ~10 MDa (**Fig. 2.16a**). Cryo-EM analysis of the 1.75 MDa assemblies revealed cubic particles ~22 nm in diameter, paralleling those formed by 201phi2-1 chimallin (**Fig. 2.5c, Fig. 2.16b**). We obtained a 4.2 Å resolution structure of the overall Goslar chimallin assembly, and used a similar localized reconstruction procedure as for the 201phi2-1 chimallin to obtain a 2.6 Å resolution structure of the Goslar chimallin tetramer, and a 2.3 Å resolution structure of a single protomer (**Fig. 2.5c,d, Fig. 2.16c-g**). Overall, Goslar chimallin shows high structural homology to 201phi2-1 chimallin despite the low overall sequence identity, with a Ca RMSD of 1.8 Å within a protomer, and 4.8 Å over an entire C4-symmetric tetramer (**Fig. 2.5e,f**). The NTS, CTS1, and CTS2 segments of Goslar chimallin show near-identical interactions to neighboring protomers compared to 201phi2-1 chimallin with the exception of CTS2, which is shorter in Goslar than in 201phi2-1 and shows a distinct set of interactions (**Fig. 2.10e-g, Tables 2.8 & 2.9**).

Recently, transmission electron micrographs of cryo-preserved lysates from *Salmonella* cells infected with the jumbo phage SPN3US revealed an unidentified square lattice structure with a 13.5 nm periodicity.¹⁶ We identified a diverged chimallin homolog (gp244) in SPN3US that shares only 10% identity with 201phi2-1 chimallin (**Fig. 2.5a**). The similar dimensions and overall morphology of the SPN3US lattice compared to the 201phi2-1 and Goslar nuclear shells strongly suggest that these structures are composed of chimallin. Further, during revision of this manuscript, a preprint reported a cryo-EM map of a chimallin sheet reconstituted *in vitro* from *Pseudomonas* jumbo phage PhiPA3 (gp53, 52.2% sequence identity to 201phi2-1 gp105).¹⁷ The PhiPA3 chimallin possesses a similar overall structure to 201phi2-1 and Goslar at the protomer level, but predominantly assembled *in vitro* as an oblique (p2) lattice in contrast to the predominantly square (p4) lattice observed in cells.¹⁷ Thus, together with our findings on 201phi2-1 and Goslar, these data show that despite extremely low sequence conservation, chimallin proteins from diverged jumbo phages exhibit high structural conservation at both the level of an individual protomer and overall nuclear shell architecture.

2.3 Discussion

Here, we detail the molecular architecture of the jumbo phage nuclear shell, a self-assembling, micron-scale proteinaceous compartment that segregates transcription and translation, largely excluding protein transport yet allowing selective protein import and mRNA export. We find that the shell is primarily composed of a single protein, termed chimallin, which self-assembles through its extended N- and C-termini into a closed compartment. The nuclear shell is a square quasi-lattice that effectively balances integrity and flexibility, erecting a physical barrier between the replicating phage genome and host defenses including restriction enzymes and CRISPR-Cas nucleases.

Chimallin is the first and most abundant protein produced upon phage 201phi2-1 infection of a host cell.² A key question is how chimallin specifically nucleates around the injected phage genome. While it does self-assemble *in vitro* at high concentration, chimallin does not form a phage nucleus by itself when overexpressed in uninfected cells,^{2, 8} suggesting that the phage encodes nucleation factor(s) that promote assembly of the shell around its own genome. These factors may be produced alongside chimallin in the infected cell, or alternatively may be injected along with the phage DNA upon initial infection. Thus, there is likely a window of time during the initial infection in which the phage genome is not protected by the nuclear shell. How does the phage protect itself during that time? We have consistently observed unidentified spherical bodies (USBs) in cells infected with either 201phi2-1 or Goslar, with internal density consistent with tightly packed DNA (**Fig. 2.17**). In 201phi2-1-infected cells, USBs average 201 nm in diameter (internal volume $\sim 4.25 \times 10^{-3} \mu\text{m}^3$), while in Goslar-infected cells, they average 182 nm in diameter (internal volume $\sim 3.16 \times 10^{-3} \mu\text{m}^3$). The internal volume of USBs in 201phi2-1-infected cells is 1.34x that of Goslar-infected cells, closely matching the 1.33x ratio between the genome sizes of two phages (317 kb for 201phi2-1 vs. 237 kb for Goslar). For both 201phi2-1 and Goslar, the internal volume of USBs is ~ 3.3 - 4.4 times that of their capsids, suggesting that if each USB contains one phage genome, the DNA is less densely packed in USBs compared to the capsid. Using subtomogram analysis, we did not find USBs to have the same surface as the nuclear shell made of a single layer of chimallin; rather, the boundary's density is most consistent with a lipid bilayer (**Fig. 2.17f-l**). Notably, similarly sized compartments have been observed by thin-section TEM of jumbo phage PhiKZ and SPN3US infecting *P. aeruginosa* and

Salmonella, respectively.¹⁸⁻¹⁹ Further work will be required to determine whether USBs represent unproductive infections that failed to pierce the inner membrane of the host, or instead represent a mechanism to protect a phage's genome early in infection, prior to chimallin production and nuclear shell assembly.

Another unresolved question is how the phage nuclear shell grows concomitantly with phage genome replication. We propose that shell growth is accomplished by incorporating soluble chimallin subunits into the lattice through a presumably isoenthalpic transfer of the N- and C-terminal segments that bind the chimallin interfaces on the lattice to the new chimallin subunits, effectively breaking and subsequently resealing the existing lattice. The high propensity of chimallin to self-assemble, and the protein's abundance in infected cells, likely ensures that host defense factors have little opportunity to access the replicating phage genome during its growth.

We and others have shown that specific phage proteins are actively imported into the phage nucleus, while other proteins – including host defense nucleases – are excluded.³⁻⁴ Given that the pores of the chimallin lattice are not large enough to allow passage of most folded proteins, these data suggest that the phage encodes minor shell components that mediate specific, directional transport of proteins and potentially phage-encoded mRNAs through the protein barrier. Similarly, specific proteins either associated with or integrated into the shell likely mediate interactions with the phage-encoded tubulin homolog PhuZ to position the phage nucleus² and enable phage capsids to dock on the shell surface for genome packaging.⁹ Intriguingly, the prohead protease of the jumbo phage PhiKZ has been shown to cleave chimallin (gp54) between its C-terminal domain and CTS1 (**Fig. 2.18**),²⁰ suggesting that proteolytic processing of chimallin may contribute to capsid docking and filling by locally disrupting lattice integrity.

The jumbo phage nucleus is a striking example of convergent evolution to solve a problem – isolation of a genome from the surrounding cell contents – previously thought to have evolved only once in the history of life. Here, we have shown how the phage-encoded chimallin protein (ChmA) self-assembles into an effective nuclear-cytosolic barrier. This work sets the stage for future identification of minor shell components (ChmB, ChmC, etc.) that manage shell nucleation and growth, mediate nuclear-cytoplasmic transport, and direct key steps in the

phage life cycle including cytoskeletal interactions and genome packaging. Finally, the structural elucidation of the principal component of the phage nuclear shell opens the possibility to the design of engineered protein-based compartments with sophisticated functions that span nanometer to micron scales.

2.4 Methods

Bacterial and phage growth conditions

Pseudomonas chlororaphis 200-B²¹ was grown on HA media at 30°C as previously described. 201phi2-1 lysates were collected as previously described with minor modifications.² Briefly, 0.5 mL from a dense *P. chlororaphis* culture grown in HA liquid media (HA with no agar added) was infected with 10 µL serial dilutions of high-titer 201phi2-1 lysate (10^{11} - 10^{12} pfu/mL), incubated for 15 minutes at room temperature, mixed with 4.5 mL of HA 0.35% top agar, poured over HA plates, and incubated overnight at 30°C. Then, 5 mL phage buffer was added to web lysis plates the following day and incubated at room temperature for 5 hours. The phage lysate was collected by aspiration, cell debris was pelleted by centrifugation at 3220 x g for 10 minutes, and the resulting clarified phage lysate was stored at 4°C. *Escherichia coli* strain APEC 2248 was obtained from the DSMZ in Germany and grown in LB media at 37°C. Goslar lysate (10^{10} pfu/mL) was sent by Johannes Wittmann at the DSMZ and stored at 4°C. All bacterial strains used in this study are listed in **Table 2.10**.

Plasmid construction and expression

The pHERD-30T plasmid was used for expressing GFP-tagged full-length and truncated Chimallin in *P. chlororaphis*.²²⁻²³ All constructs were designed with GFPmut1 fused to the N-terminus and synthesized by GenScript. The plasmids were electroporated into *P. chlororaphis*, and 25 µg/mL gentamicin sulfate was used for selection of colonies. All plasmids used in this study are listed in **Table 2.10**.

Live cell fluorescence microscopy and image analysis

P. chlororaphis cells (5 µL of OD₆₀₀=0.6) were inoculated on imaging pads in weller microscope slides. The imaging pads were composed of 1% agarose, 25% HA broth, 25 µg/mL gentamicin sulfate, 2 µg/mL FM4-64, 0.2 µg/mL DAPI, and 1% arabinose to induce expression. The slides were incubated at 30°C for 3 hours in a humid chamber, and 10 µL of undiluted high-titer 201phi2-1 lysate was added to the pads to infect the cells 60 minutes before imaging. Samples were imaged using the DeltaVision Elite deconvolution microscope (Applied

Precision, Issaquah, WA, USA). Images were deconvolved using the aggressive algorithm in the DeltaVision softWoRx program-v6.5.2.

All image analysis was performed on images prior to deconvolution. Average protein incorporation into the 201phi2-1 phage nucleus structure was determined using FIJI by measuring the mean gray value of the ring of GFP intensity that denotes the phage nucleus and the mean gray value of cytoplasmic GFP outside of this ring. The ratio of mean gray values of this ring GFP to cytoplasmic GFP was calculated as the average incorporation. Representative cells were chosen from each dataset, and a 3D graph of normalized GFP intensity was generated in MATLAB 2019a. Statistical analyses were performed using Prism-v9.3 (GraphPad Software).

Growth Curves

Culture of *P. chlororaphis* transformed with either an empty vector or one of the chimallin constructs were grown in HA broth containing 25 µg/mL gentamicin sulfate to an OD₆₀₀ between 0.6 and 0.8 and then back-diluted to an OD₆₀₀ of 0.1 in media containing 1% arabinose in 96-well plates. Serial 10-fold dilutions of 201phi2-1 lysate were added and growth monitored at OD₆₀₀ at 10-minute intervals for 8 hours with continuous shaking at 30°C. All growth curves were performed in duplicate, with duplicate wells (a total of four wells averaged per data point).

Cryo-electron microscopy of in situ samples and image acquisition

For grid preparation of 201phi2-1 infections, host bacterial cells were infected on agarose pads as previously described² for 50-60 minutes. For grid preparation of Goslar infections, 10 agarose pads (1% agarose, 25% LB) were prepared in well slides and spotted with 10 µl of *E. coli* (APEC 2248) cells at an OD₆₀₀ of ~0.35 then incubated at 37°C for 1.5 hours in a humidior. 10µl of Goslar lysate from the DSMZ was added to each pad. At approximately 30 mpi, a portion of the infected cells were collected at room temperature and delivered for plunge-freezing. Infected cells were collected by the addition of 25 µl of 25% LB to each pad and gentle scraping with the bottom of an eppendorf tube followed by aspiration. A portion of the collection was aliquoted, the remainder was centrifuged at 6000 x g for 45 seconds, resuspended with 0.25x volume of the supernatant, and a portion of that was diluted 1:1 in supernatant. The remaining cells incubated on pads at 37°C until 90 mpi, at which point they were assessed for productive infections by light-microscopy. Plunging of samples began 20-

30 minutes after removal from 37°C), which significantly slows infection progression. Since phage nuclei were observed in this sample after cryo-FIB-ET, the sample was suitable for the analyses performed in this study.

A volume of 4-7 μl of cells were deposited on R2/1 Cu 200 grids (Quantifoil) that had been glow-discharged for 1 min at 0.19 mbar and 20 mA in a PELCO easiGlow device shortly before use. Grids were mounted in a custom-built manual plunging device (Max Planck Institute of Biochemistry, Germany) and excess liquid blotted with filter paper (Whatman #1) from the backside of the grid for 5-7 seconds prior to freezing in a 50:50 ethane:propane mixture (Airgas) cooled by liquid nitrogen.

Grids were mounted into modified Autogrids (Thermo Fisher Scientific) compatible with cryo-focus ion beam milling. Samples were loaded into an Aquilos 2 cryo-focused ion beam/scanning electron microscope (TFS) and milled to generate lamellae approximately $\sim 150\text{-}250$ nm thick as previously described.²⁴

Lamellae were imaged using a Titan Krios G3 transmission electron microscope (Thermo Fisher Scientific) operated at 300 kV configured for fringe-free illumination and equipped with a K2 directed electron detector (Gatan) mounted post Quantum 968 LS imaging filter (Gatan). The microscope was operated in EFTEM mode with a slit-width of 20 eV and using a 70 μm objective aperture. Automated data acquisition was performed using SerialEM-v3.8b11²⁵ and all images were collected using the K2 in counting mode.

For lamellae of 201phi2-1-infected *P. chlororaphis*, tilt-series were acquired at a 3.46 Å pixel size over a nominal range of $\pm 51^\circ$ in 3° steps with a grouping 2 using a dose-symmetric scheme²⁶ with a per-tilt fluence of $1.8\text{ e}^- \cdot \text{Å}^{-2}$ and total of about $120\text{ e}^- \cdot \text{Å}^{-2}$ per tilt-series. Nine tilt-series were acquired with a realized defocus range of -4.5 to $-6\text{ }\mu\text{m}$ along the tilt-axis. An additional two data sets of six tilt-series each were collected at a pixel size of 4.27 Å with nominal tilt ranges of $\pm 50^\circ$ and $\pm 60^\circ$ in 2° steps with a grouping 2 using a dose-symmetric scheme with a per-tilt fluence of $1.8\text{-}2.0\text{ e}^- \cdot \text{Å}^{-2}$ and total of about $100\text{-}110\text{ e}^- \cdot \text{Å}^{-2}$ per tilt-series.

For lamellae of Goslar-infected APEC 2248, tilt-series were acquired at a 4.27 Å pixel size over a nominal range of $\pm 56^\circ$ in 2° steps with a grouping 2 using a dose-symmetric scheme with a per-tilt fluence of $2.6\text{ e}^- \cdot \text{Å}^{-2}$ and total of about $150\text{ e}^- \cdot \text{Å}^{-2}$ per tilt-series. Twenty-one tilt-series were acquired with a realized defocus range of -5 to $-6\text{ }\mu\text{m}$ along the tilt-axis.

Image processing and subtomogram analysis of in situ cryo-electron microscopy data of the phage nucleus

All tilt-series pre-processing was performed using Warp-v1.09 unless otherwise specified.²⁷ Tilt-movies were corrected for whole-frame motion and aligned via patch-tracking using Etomo (IMOD-v4.10.28).²⁸ Tomograms were reconstructed with the deconvolution filter for visualization and manual picking in 3dmod (IMOD-v4.10.28). All subsequently reported resolution estimates are based on the 0.143-cutoff criterion of the Fourier shell correlations between masked, independently refined half-maps using high-resolution noise-substitution to mitigate masking artifacts.²⁹

First, for the 201phi2-1-infected *P. chlororaphis* dataset collected at 3.46 Å per pixel, subtomogram averaging of the *P. chlororaphis* host cell ribosomes was performed in order to improve initial tilt-series alignments using the recently developed multi-particle framework, M³⁰ (**Fig. 2.6n**). A set of 400 particles were manually picked across the tomograms, extracted at 20 Å per pixel, and aligned in RELION-v3.1.1 to generate an initial reference.³¹⁻³² This data-derived reference was used for template-matching against 20 Å per pixel tomograms at a sampling rate of 15°. Template-matched hits were curated in Cube (<https://github.com/dtegunov/cube>), ultimately resulting in 17,169 particle picks. The initial particle set was extracted at 10 Å per pixel and subjected to Class3D, after which 11,148 particles were selected for further analysis. Refine3D of the curated particle set reached the binned Nyquist limit of 20 Å. The refined particles were imported into M-v1.09 at 3.46 Å per pixel. Three iterations of refinement were performed starting with image-warp and particle poses, then incorporating refinement of stage angles and volume-warp, and finally including individual tilt-movie alignment. This procedure resulted in a ribosome reconstruction at an estimated resolution of about 11 Å. Further refinement of the particles in RELION yielded a reconstruction at an estimated resolution of about 10 Å. Neither additional attempts at 3D-classification nor multi-particle refinement lead to an improved ribosome reconstruction.

New tomograms were reconstructed at 20 Å per pixel using the ribosome alignment metadata. The perimeters of the 201phi2-1 phage nuclei were coarsely traced in these updated tomograms using 3dmod.³² Traces were converted into surface models using custom MATLAB (MathWorks, v2019a) scripts and built-in Dynamo-v1.1.514 functions.³³ Points were sampled every 4 nm along the surface models, oriented normal to the surface (i.e., positive-Z towards the cytosol), and extracted from normalized, CTF-corrected tomograms at 10 Å per pixel in a 480 Å side-length box using Dynamo.³³ Initial orientations of the 66,887 extracted particles were

curated using the Place Object plugin³⁴ for UCSF-Chimera-v1.15³⁵ and incorrectly oriented particles were manually flipped.

To generate an initial reference, a subset of 17,622 particles from two tomograms were subjected to reference-free alignment in Dynamo-v1.1.514 for several iterations. For this procedure, no point-group symmetry was enforced, alignment was limited to 40 Å by an *ad hoc* low-pass filter each interaction, and the out-of-plane searches were restricted to prevent flipping of sidedness. The alignment converged to yield a reconstruction conforming to an apparent square (*p*4, 442) lattice. Analysis of particle positions and orientations using Place Objects³⁴ and “neighbor plots”³⁶ were consistent with reconstructed average and indicated a spacing between approximately 11.5 nm between congruent 4-fold axes.

The initial reference was subsequently used to align the entire dataset for a single iteration in Dynamo. For this step, alignment was limited to 40 Å, the out-of-plane searches were restricted to prevent flipping of sidedness, C4 symmetry enforced, and a box-wide by 240 Å cylindrical alignment mask applied. Inspection of particle positions and orientations using Place Objects³⁴ and neighbor plots³⁶ were again consistent with a square lattice-like arrangement. To deal with the initial over-sampling, particle duplicates were identified as those within 9 nm center-to-center distance of another particle and the one with the lower cross-correlation to the reconstruction removed, which resulted in 21,165 retained particles. In addition, a geometry-based cleaning step was performed to remove particles with less than three neighbors within 10 to 13 nm, which resulted in 8,454 retained particles.

The curated particle set was split into approximately equal half-sets on a per-tomogram basis, converted to the STAR file format using the dynamo2m-v0.2.2 package,³⁷ and re-extracted into a 480 Å side-length box at 5 Å per pixel in Warp for use in RELION. A round of Refine3D was performed using a 40 Å lowpass filtered reference, C4 symmetry, local-searches starting at 3.7°, and a box-wide soft-shape mask. This resulted in a reconstruction at an estimated resolution of 24 Å for the 8,454 particle set.

Classification without alignment was performed using a 320 Å spherical mask and C4 symmetry, to promote convergence from the relatively low particle count. This differentiated three distinct classes corresponding to “concave” (2,033 particles), “flat” (4,475 particles), and “convex” (945 particles) states of the central tetramer, along with a noisy class (1,001 particles). The particles corresponding to the interpretable classes were re-extracted into a 320 Å box at 5 Å per pixel and subjected to 3D auto-refinement as described for the consensus

reconstruction. The estimated resolutions for the lowered, intermediate, and raised classes were 20 Å, 18 Å, and 23 Å, respectively. Refinement in M of either the consensus particle set or the three aforementioned classes above did not yield notable improvements in the reconstructions. This may be attributed to prior refinement of the tilt-series alignment, the most resolution-limiting factor, using the host ribosomes.³⁰

The twelve tilt-series of 201phi2-1-infected *P. chlororaphis* dataset collected at 4.27 Å per pixel were pre-processed and host ribosomes averaged as described above. The ribosome reconstruction from above was low-pass filtered to 40 Å and used for template-matching in Warp-v1.09, and curated in Cube to yield 47,469 particle positions. Ribosomes were extracted at 10 Å per pixel and subjected to reference-free 3D classification in RELION-v3.1.1 which resulted in 15,782 subtomograms. Masked auto-refinement resulted in a reconstruction with an estimated resolution of 28 Å. Refinement of tilt-series parameters in M improved the resolution to about 20 Å (not shown).

For the 201phi2-1 nucleus in the 4.27 Å per pixel dataset, we were unable to completely resolve the quasi-lattice register using either an *ab initio* reference or the reconstruction from above as determined by neighbor plots. Thus, this dataset was solely included in the analysis of the unidentified spherical bodies and not further subtomogram averaging.

The tilt-series of Goslar-infected APEC 2248 were pre-processed and host ribosomes averaged similarly as described above. An initial host ribosome reference was generated from 400 randomly selected particles and used for template-matching in Warp-v1.09, and curated in Cube to yield 98,981 particle positions. Ribosomes were extracted at 10 Å per pixel and subjected to reference-free 3D classification in RELION-v3.1.1 which distinguished 70S and 50S classes containing 46,056 and 3,710 particles, respectively. Particles were re-extracted at 6 Å per pixel and subjected to 3D auto-refinement to yield 20 Å and 14 Å for the 50S and 70S classes, respectively. Refinement of tilt-series parameters in M, followed by an additional round of 3D auto-refinement at 4.27 Å per pixel resulted in 12 Å and 8.54 Å (Nyquist-limit of the data) for the 50S and 70S, respectively.

For the Goslar nucleus, the nuclei perimeters were traced from 6 tomograms and used to extract over-sampled points normal to the surface, which resulted in 42,416 initial particles. A subset of 10,512 particles were used to generate an *ab initio* reference in C1 as described above. The initial Goslar reference presented a similar spacing

(~11.5 nm) and apparent C4 symmetry as the 201phi2-1 reconstruction, however the average converged on the opposite 4-fold axis. For ease of subsequent analysis, the center of the Goslar initial reference was shifted to match 201phi2-1. Alignment of the entire dataset was performed as described above and distance-based cleaning post-alignment resulted in 4,501 particles for further processing. Alignment and reconstruction in RELION enforcing C4 symmetry resulted in a reconstruction with an estimated resolution of 27 Å. Similar to the 201phi2-1, 3D classification of the consensus refinement without alignment separated the data into classes in which the central tetramer appeared “concave” (2,802) or “convex” (1,699). Refinement of these classes resulted in reconstruction with estimated resolutions of 20 Å and 25 Å for the lowered and raised classes, respectively.

Analysis of unidentified spherical bodies

Unidentified spherical bodies (USBs) were manually identified and their maximal apparent diameters measured from their line intensity profiles in 20 Å per pixel tomograms using FIJI.³⁸ In order to assess whether the surfaces of these compartments possessed an underlying structure, we attempted subtomogram analysis of the compartment surfaces essentially as described in.^{37, 39} We were unable to obtain a reconstruction exhibiting a regular underlying structure as assessed both visually and by neighbor plots. However, despite their differing exterior membrane, the interior density of the USBs is visually consistent with nucleic acid like that of the interior of the phage nucleus.

Segmentation and visualization of in situ tomography data

Segmentation of host cell membranes and the phage nucleus perimeter was performed on 20 Å per pixel tomograms by first coarsely segmenting using TomoSegMemTV⁴⁰ followed by manual patching with Amira-v6.7 (TFS). For the purposes of segmentation, phage capsids, tails, PhuZ, and RecA-like particles were subjected to a coarse subtomogram averaging procedure using particles sampled at 20 Å per pixel. For capsids, all particles were manually picked. Reference-generation and alignment of capsids was performed by enforcing icosahedral symmetry with Relion-v3.1.1³¹⁻³² (despite the capsids possessing C5 symmetry) in order to promote convergence from the low number of particles. For the phage tails, the start and end points along the filament axis were defined manually and used to seed over-sampled filament models in Dynamo-v1.1.514.^{33, 41} An initial reference for the tail was generated using Dynamo-v1.1.514 from two full-length tails with clear polarity. The resulting reference displayed apparent C6 symmetry, which was enforced for the alignment of all tails from a given

tomogram using Dynamo-v1.1.514 and Relion-v3.1.1. Similar to the phage tails, the PhuZ and RecA-like filaments were picked and refined but without enforcing symmetry. We do not report resolution claims for these averages and solely use them for display purposes in segmentations. Duplicate particles were removed and final averages were placed back in the reference-frame of their respective tomograms using *dynamo_table_place*. For clarity, a random subset of 500 ribosomes were selected for display in the segmentation.

Surface curvature estimates from segmentation

The segmentation of the phage nucleus, depicted in Figure 1d, sampled at 2 nm/pixels, was used to estimate the principal curvature of the shell using PyCurv.⁴² PyCurv was run with default parameters and a hit radius of 3 pixels. For visualization purposes, the principal curvature values (κ_1 , κ_2) were converted from a radius of curvature (r , nm⁻¹) to an angle (θ , degrees) using the following formula:

$$\theta = 2 \arctan\left(\frac{s}{2r}\right)$$

where r is the radius of curvature (inverse of κ_n) and s is for the side-length of the polygon circumscribed, taken as 5.75 nm.

Protein expression and purification

Full-length Chimallin from bacteriophages 201phi2-1 (gp105; NCBI Accession YP_001956829.1) and Goslar (gp189; NCBI Accession YP_009820873.1) were cloned with an N-terminal TEV protease-cleavable His₆-tag using UC Berkeley Macrolab vector 2-BT (Addgene #29666). Truncations and other modified constructs were cloned by PCR mutagenesis and isothermal assembly, and inserted into the same vector. Proteins were expressed in *E. coli* Rosetta2 pLysS (Novagen) by growing cells to $A_{600}=0.8$, inducing expression with 0.3 mM IPTG, then growing cells at 20°C for 16-18 hours. Cells were harvested by centrifugation and resuspended in buffer A (50 mM Tris pH 7.5, 10 mM Imidazole, 300 mM NaCl, 10% glycerol, and 2 mM β -mercaptoethanol), then lysed by sonication and the lysate cleared by centrifugation. Protein was purified by Ni²⁺ affinity method. The purified proteins were centrifuged briefly to settle down the floating particles (visible large assemblies and aggregated proteins). The proteins were dialyzed into buffer B (20 mM Tris pH 7.5, 250 mM NaCl, 2 mM β -mercaptoethanol) and the N-terminal histidine tag was cleaved using TEV protease with overnight incubation at

4°C. The retrieved tagless proteins were further purified for homogeneity through Superose 6 Increase 10/300 GL column (Cytiva) in buffer B. The quality of purified proteins was verified by SDS-PAGE analysis.

For analysis by size exclusion chromatography coupled to multi-angle light scattering (SEC-MALS), a 100 μ L sample of protein at 2 mg/mL was passed over a Superose 6 Increase 10/300 GL column (Cytiva) in buffer B. Light scattering and refractive index profiles were collected by miniDAWN TREOS and Optilab T-rEX detectors (Wyatt Technology), respectively, and molecular weight was calculated using ASTRA v. 8 software (Wyatt Technology).

Cryo-electron microscopy of in vitro samples and image acquisition

For grid preparation, freshly purified recombinant 201phi2-1 chimallin was collected from size-exclusion chromatography (estimated concentration of 4 μ M of the monomer, 0.3 mg/mL). Immediately prior to use, R2/2 Cu 300 grids (Quantifoil) were glow-discharged for 1 min at 0.19 mbar and 20 mA in a PELCO easiGlow device. Sample was applied to a grid as a 3.2 μ L drop in the environmental chamber of a Vitrobot Mark IV (Thermo Fisher Scientific) held at 16°C and 100% humidity. Upon application of the sample, the grid was blotted immediately with filter paper for 3 seconds prior to plunging into a 50:50 ethane:propane mixture cooled by liquid nitrogen. Grids were mounted into standard AutoGrids (Thermo Fisher Scientific) for imaging. Grids for recombinant Goslar chimallin protein were prepared similarly, but with the modification that the sample was concentrated to approximately 33 μ M of the monomer (2.5 mg/mL) prior to plunge-freezing. The void peaks from each purification were frozen similarly at the eluted concentration after dilution 1:1 with 6 nm BSA-tracer gold (Electron Microscopy Sciences).

All samples were imaged using a Titan Krios G3 transmission electron microscope (Thermo Fisher Scientific) operated at 300 kV configured for fringe-free illumination and equipped with a K2 directed electron detector (Gatan) mounted post Quantum 968 LS imaging filter (Gatan). The microscope was operated in EFTEM mode with a slit-width of 20 eV and using a 70 μ m objective aperture. Automated data acquisition was performed using SerialEM-v3.8b11²⁵ and all images were collected using the K2 in counting mode.

For the 201phi2-1 24-mer sample, tilt-series were acquired using a pixel size of 1.376 Å with a per-tilt fluence of 4.7 e⁻·Å⁻² using a dose-symmetric scheme²⁵ from +/- 51° in 3° steps and a grouping 3, resulting in a fluence of 164.5 e⁻·Å⁻² per tilt-series. In total 4 tilt-series were collected with a realized defocus of -2.5 to -4 μ m along

the tilt-axis. Movies for single-particle analysis were recorded at a pixel size of 1.075 Å with fluence of 42.6 e⁻·Å⁻² distributed uniformly over 40 frames. Automated data acquisition was performed using image-shift with active beam-tilt compensation to acquire 9 movies per hole per stage movement. In total 4,192 movies were acquired with a realized defocus range of -0.1 to -1.5 µm.

For the Goslar 24mer sample, movies for single-particle analysis were recorded at a pixel size of 0.8452 Å with fluence of 40 e⁻·Å⁻² distributed uniformly over 44 frames. Automated data acquisition was again performed using image-shift with active beam-tilt compensation to acquire 10 movies per hole per stage movement. In total, 3921 movies were acquired with a realized defocus range of -0.1 to -1.5 µm.

For void peak samples, tilt-series were acquired similarly to that of the 201phi2-1 24mer sample but using a pixel size of 1.752 Å and tilt-range of +/- 60°.

Image processing of in vitro cryo-electron microscopy data

All movie pre-processing was performed using Warp-v1.09 unless otherwise specified.²⁷ Tilt-movies of the 201phi2-1 chimallin were corrected for whole-frame motion and aligned via patch-tracking using Etomo (IMOD-v4.10.28).²⁸ Tomograms were reconstructed with the deconvolution filter for visualization and manual picking of subtomograms using 3dmod (IMOD-v4.10.28).⁴³ A total of 203 manually picked subtomograms and their corresponding 3D-CTF volumes were reconstructed with a 288 Å side-length. Subtomograms were aligned and averaged initially in C1 by reference-free refinement as implemented in RELION-v3.1.1³¹⁻³² to an estimated resolution of 22 Å. The C1 reconstruction displayed features consistent with a cubic assembly of the 201phi2-1 chimallin protomers. Thus, an additional round of refinement using the C1 reconstruction as a reference and enforcing O point-group symmetry improved the estimated resolution to 18 Å.

For the single-particle 201phi2-1 chimallin data, movies were motion-corrected with exposure-weighting and initial CTF parameters estimated using 5x5 grids. Micrographs were culled by thresholding for an estimated defocus in the range of 0.3-1.5 µm and CTF-fit resolutions better than 6 Å resulting in 4,098 micrographs for further processing. An initial set of 140,782 particle positions were picked with BoxNet2 (Warp-v1.09)²⁷ using a model re-trained on 20 manually curated micrographs and using a threshold of 0.95. Particle images were extracted using a 396 Å side-length. All further processing was performed using RELION-v3.1.1³² unless otherwise specified. A single round of reference-free 2D-classification was performed and the 128,798 particle

images assigned to the averages displaying internal features were selected for further processing. At this stage, analysis of the 2D averages suggested the presence of 4-, 3-, and 2-fold symmetry axes, consistent with a cubic arrangement of the chimallin protomers in the particles. Thus, we subjected the particle images to 3D-refinement using the subtomogram average obtained from above as an initial reference lowpass filtered to 35 Å and O point-group symmetry enforced, which resulted in a reconstruction at an estimated resolution of 4.2 Å. However, the reconstruction did not display features consistent with this resolution estimate (e.g., β -strands were not separated). The high apparent point-group symmetry and distribution of 2D class averages did not support the inflated resolution being due to a preferred orientation. Partitioning particles into half-sets by micrograph did not change the estimated resolution of reconstruction, indicating the inflated estimate was not due to splitting identical or adjacent particles across the half-sets. In addition, extensive 3D-classification with and without symmetry enforced did not yield distinct classes. Therefore, the possibility of quasi-symmetry was investigated by performing localized reconstructions of sub-structures within the particles. To reduce computational burden, the apparent O symmetry was first partially expanded to C4 using *relion_particle_symmetry_expand* to fully expand to C1 before removing redundant image replicates (noting that redundant views of the 4-fold axes possess the same last two Euler angles) to yield 772,788 sub-particles. Refinement of the partially expanded particles while enforcing C4 point-group symmetry and using a soft shape mask resulted in a reconstruction with an estimated resolution of 3.6 Å with notably improved features. Re-centering and re-extraction using a 245 Å side-length followed by refinement improved the estimated resolution to 3.6 Å. CTF refinement⁴⁴ (per particle defocus, per micrograph astigmatism, beam-tilt, and trefoil) and Bayesian polishing⁴⁵ successively improved the resolution further to 3.5 Å and 3.4 Å, respectively. A round of 2D-classification without alignment was performed to remove particles assigned to empty or poorly resolved classes, which yielded a set of 664,363 sub-particles and no change in the estimated resolution upon re-running 3D-refinement. Although the reconstruction substantially improved through this procedure, the C4 map still exhibited distorted density (e.g., elongated helices). Attempts at 3D-classification did not separate distinct classes. Thus, a localized reconstruction was performed focused on the individual chimallin protomer in C1. Again, to reduce computational burden, before expanding the symmetry to C1 another round of Bayesian polishing was performed in which the sub-particle images were extracted using a 354 Å side-length and premultiplied by their CTF before cropping in real-space

to a 200 Å side-length. After another round of 3D-refinement enforcing C4 point-group symmetry, the data was expanded to C1 which resulted in 2,657,452 sub-particles and refined to an estimated resolution of 3.3 Å. The Bayesian polishing job was re-run to extract sub-particles at the full box size and without premultiplication by their CTF for import into cryoSPARC-v3.2.⁴⁶ A single round of local non-uniform refinement⁴⁷ was performed in C1 using a user-supplied static mask, marginalization, and FSC noise-substitution options, which lead to a final reconstruction of the 201phi2-1 chimallin monomer at an estimated resolution of 3.1 Å.

The Goslar chimallin single-particle data were pre-processed similarly to the 201phi2-1 chimallin data, which after initial thresholding resulted in 2889 micrographs for further processing. Initial particle positions were identified using the 201phi2-1 chimallin-trained BoxNet2 (Warp-v1.09)²⁷ model with a threshold of 0.1, which resulted in 289,387 picks. Particles were extracted using a 400 Å side-length and subjected to iterative rounds of 2D-classification and sub-selection, which resulted in 78,532 particles used for initial 3D-refinement. The Goslar chimallin particles exhibited the same quasi-symmetry as the 201phi2-1 chimallin described above, thus were processed using the same localized reconstruction procedure. The quasi-O, quasi-C4, and C1 reconstructions yielded estimated resolutions of 4.0 Å, 2.6 Å, and 2.4 Å, respectively. The quasi-C4 and C1 reconstructions within RELION³² were performed on particle images that were extracted using a 470 Å side-length, premultiplied by their CTF, and cropped in real-space to a 200 Å side-length. The final C1 reconstruction was performed in cryoSPARC-v3.2⁴⁶ as described above, which led to a final reconstruction of the Goslar chimallin monomer at an estimated resolution of 2.3 Å from 1,407,340 sub-particle images.

All resolution estimates are based on the 0.143-cutoff criterion of the Fourier shell correlations between masked independently refined half-maps using high-resolution noise-substitution to mitigate masking artifacts.²⁹ Local resolution estimates were computed using RELION with default parameters. Resolution anisotropy for the C1 reconstructions were assessed using the 3DFSC⁴⁸ web server which reported sphericity values of 0.963 and 0.994 for the 201phi2-1 and Goslar maps, respectively.

Void peak tilt-series were processed similarly to the 201phi2-1 24-mer tilt-series, but using the gold-fiducials for alignment instead of patch-tracking in Etomo.²⁸

Coordinate model building and refinement

Initial monomer models were generated via the DeepTracer web server⁴⁹ followed by manual building in COOT-v0.9.1⁵⁰ and subjected to real-space refinement in PHENIX-v1.19.2.⁵¹ To generate tetramer models, monomer models were rigid-body docked into the C4 maps using UCSF Chimera-v1.15³⁵ and the N-terminal segments joined to the appropriate protomer cores. To generate 24mer models, tetramer models were rigid-body docked into the hexahedral maps and the C-terminal segments reassigned to the appropriate protomer cores. To ensure robust refinement, tetramer and 24mer structures were refined with C4 or O non-crystallographic symmetry (NCS) constraints and reference-model restraints based on high-resolution monomer structures. Isotropic atomic displacement parameters were refined against the respective unsharpened maps. All models were validated using MolProbity⁵² and EMRinger⁵³ (**Table 2.2**). EMRinger scores for the 201phi2-1 24mer, tetramer, and monomer models were 0.46, 2.86, and 2.39, respectively. EMRinger scores for the Goslar 24mer, tetramer, and monomer models were 0.92, 3.10, and 3.59, respectively.

Interface analysis

Interface analysis for the cubic assemblies to identify interacting residues and to calculate buried surface area was performed using the ePISA-v1.52⁵⁴ and CaPTURE⁵⁵ web servers.

Nine-tetramer sheet modelling

Nine chimallin tetramers were arranged in a flat sheet (3x3) structure by fitting in the consensus subtomogram average. Assuming the unfolding of the cubical assembly to create a flat sheet structure, the interacting C-terminal segments in the corner three-fold axis were re-assigned in a four-fold symmetry axis to the corresponding protomer. The missing residues between the C-terminal domain and C-terminal segment were built manually in COOT-v0.9.1 ensuring no clash with other modelled atoms (taking the flat sheet model in consideration).⁵⁰ The missing loop region in a protomer (residues 307-319) was built using the DaReUS-Loop web server.⁵⁶ This modelled chain (residues 45-612) was used to re-create the flat-sheet structure by applying symmetry. Finally in this flat sheet model, the protruding C-terminal segments of peripheral protomers were trimmed and twelve interacting segments in the periphery were included in the final model (48 chains).

Protonation state assignment and electrostatics estimates

The electrostatic surface representation was generated with the APBS-v3.0.0⁵⁷ using the AMBER99 force field⁵⁸ and a pH of 7.5 for assigning protonation states using PROPKA-v3.4.0⁵⁹ through PDB2PQR-v3.4.0.⁶⁰

Elastic Network Models

Elastic network models⁶¹⁻⁶³ are a subset of normal mode analysis.⁶⁴⁻⁶⁵ Here we used anisotropic network models (ANM)⁶⁶ and Gaussian network models (GNM).⁶⁷⁻⁶⁹ Both of these models simplify the protein structure into a series of nodes, with an internode potential energy function governing node motion. To look at it another way, each mode is an eigenvector whose corresponding eigenvalue is the frequency of that motion in the model; lower frequencies correspond to dynamics that best describe the structure's intrinsic motions. ProDy (version 1.0) is a software program enabling calculation of ANM and GNM modes,⁷⁰⁻⁷¹ which we used in this study. We created 20,412 nodes for the ANM and GNM calculations, which is the largest number of nodes ever used in ProDy.

The five lowest frequency GNM modes accounted for 76% of the overall variance. Considering we do not need to use all ENM modes to capture the system's dynamics⁷² we selected these five GNM modes and the five lowest frequency ANM modes to use in our models. The GNM's Kirchoff matrix was built with a pairwise interaction cutoff distance of 10 Å and a spring constant of 1.0, while the ANM's Hessian matrix used a pairwise interaction cutoff distance of 15 Å and a spring constant of 1.0. The ANM structural ensemble movies used an RMSD difference of 25 Å from the original conformation to display the protein sheet's flexibility.

Molecular dynamics simulations

Simulations were performed using the 9-tetramer chimallin sheet model. This structure was protonated and placed in a water box through Amber's tleap module.⁷³ The system was neutralized with Na⁺ using a 12-6 ion model.⁷⁴⁻⁷⁵ The CUDA version 10.1 implementation⁷⁶⁻⁷⁸ of Amber 20 was used.⁷⁹ The water model used was OPC⁷⁶ with the Amber 19ffsb force field.⁸⁰ The resulting system, including the protein and waterbox, contained 1,729,704 atoms. Energy minimization was performed for a total of 10,000 cycles using a combination of

steepest descent and conjugate gradient methods⁷⁶ while the heavy atoms were restrained with a force constant of 10.0 kcal/(mol x Å²). Next, the system was slowly heated from 10.0 K to 300.0 K over 4 ns before stabilizing at 300.0 K for the next 6 ns using the NVT ensemble with a Langevin thermostat with a friction coefficient (collision frequency) of $\gamma=5.0 \text{ ps}^{-1}$ and the heavy atoms restrained with a force constant of 1.0 kcal/(mol x Å²). Equilibration was performed in the NPT ensemble for 20 ns, using a timestep of 2 fs and the SHAKE algorithm, constraining bonds involving hydrogens.⁸¹ The equilibration temperature was set at 300.0 K with a Langevin thermostat with a friction coefficient (collision frequency) of $\gamma=1.0 \text{ ps}^{-1}$ ⁸²⁻⁸³ and the pressure set to 1 bar with a Berendsen barostat⁸⁴ with relaxation time constant $\tau=1.0 \text{ ps}^{-1}$ and a heavy atom restraint with a force constant of 0.1 kcal/(mol x Å²). Periodic boundary conditions were enforced with the van der Waals interaction cutoff at 8 Å, while long-range interactions were treated with the Particle mesh Ewald algorithm.⁸⁵ After equilibration, the system was cloned into five replicates for the production runs, still set at 300.0 K in the NPT ensemble. Each was run for 300 ns, resulting in 1.5 μs of total sampling.

The resulting molecular dynamics trajectories were analyzed through CPPTRAJ-v.25.6⁸⁶ and MDTraj-v1.9.4⁸⁷. In particular, RMSD was calculated through MDTraj. This was done by calculating the RMSD of the Ca atoms for all tetramers, as well just the central tetramer, from each trajectory and averaging the results.

Pore Analysis

Pore annotation was performed using CHAP-v0.9.1⁸⁸ was used for all other analyses. The free energy and solvent density plots were averaged between physiologically identical pores across all simulation replicates. The intertetramer (“corner four-fold”) pore in the upper-left quadrant contained two frames that caused CHAP to crash; these frames were removed before averaging after consultation with the CHAP developers. Considering we still averaged 1502 frames x 4 pores - 2 bad frames = 6006 frames for the intertetramer pores, we do not feel that this removal causes any difference in our conclusions.

Structure visualization and figure generation

Density maps, coordinate models, and simulation trajectories were visualized and figures generated with PyMOL-v2.5 (Schrödinger L., DeLano W., 2021), UCSF Chimera-v1.15,³⁵ ChimeraX-v1.2.5,⁸⁹ and VMD-1.9.4a35.⁹⁰

2.5 Data availability

Cryo-EM density maps have been deposited in the Electron Microscopy Data Bank. Subtomogram averaging maps have the accession numbers: EMD-25221 (201phi2-1, consensus), EMD-25220 (201phi2-1, concave), EMD-25222 (201phi2-1, flat), EMD-25223 (201phi2-1, convex), EMD-25183 (*P. chlororaphis*, 70S), EMD-25229 (Goslar, consensus), EMD-25262 (Goslar, concave), EMD-25358 (Goslar, convex), EMD-25359 (APEC2248, 70S), EMD-25360 (APEC2248, 50S). Single-particle maps have the accession numbers: EMD-25393 (201phi2-1, O), EMD-25391 (201phi2-1, C4), EMD-25392 (201phi2-1, C1), EMD-25393 (201phi2-1, D4) EMD-25394 (Goslar, O), EMD-25395 (Goslar, C4), and EMD-25395 (Goslar, C1). Coordinate models have been deposited in the RCSB Protein Data Bank with the accession numbers 7SQQ (201phi2-1, O), 7SQR (201phi2-1, C4), 7SQS (201phi2-1 C1 monomer), 7SQT (Goslar, O), 7SQU (Goslar, C4), and 7SQV (Goslar, C1). Raw cryo-EM data have been deposited with the Electron Microscopy Public Image Archive with accession codes: EMPIAR-10859 (*in situ* 201phi2-1 tilt-series), EMPIAR-10860 (*in situ* Goslar tilt-series), EMPIAR-10862 (*in vitro* 201phi2-1 frame-series), and EMPIAR-10863 (*in vitro* Goslar frame-series). Genbank IDs for protein sequences used in this study are provided in **Table 2.7**. All other data are available upon request to the corresponding author(s).

Code availability

The programs used in this study are previously published with the sources and citations indicated in the Methods section.

2.6 Acknowledgements

All electron microscopy data were collected at the UCSD Cryo-Electron Microscopy Facility, which was built and equipped with funds from UCSD and an initial gift from the Agouron Institute. We thank the UCSD Physics Computing for computational support. We thank V. Lam and J. Hutchings for advice on sample preparation and subtomogram averaging, respectively. We thank J. Krieger and I. Bahar at the University of Pittsburgh for discussions on elastic network models and expanding ProDy to accommodate the size of the chimallin sheet model. We thank J. Whittman at DMSZ for the gift of Goslar lysates. The authors thank members of the Pogliano, Villa, Corbett, and Amaro labs for helpful discussions and feedback. The authors acknowledge funding from the National Institutes of Health grants R01GM129245 (to J.P. and E.V.), R35 GM144121 (to K.D.C.), and R01GM031749 (to J.A.M.), as well as from the National Science Foundation grants CHE060073 (to R.E.A) and DBI 1920374 (to E.V.). T.L. is a Simons Foundation Awardee of the Life Sciences Research Foundation. C.S. is supported by a National Science Foundation Graduate Research Fellowship (DGE-1650112). E.V. is a Howard Hughes Medical Institute Investigator. Molecular graphics and analyses were performed in part with UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases.

Chapter 2, in full, is a modified reprint of the material as it appears in “Laughlin, T.G.*, Deep, A.*, Prichard, A.M., Seitz, C., Gu, Y., Enustun, E., Suslov, S., Khanna, K., Birkholz, E.A., Armbruster, E., McCammon, J.A., Amaro, R.E., Pogliano, J., Corbett, K.D., Villa, E. Architecture and self-assembly of the jumbo bacteriophage nuclear shell. *Nature*, **2022**, 608.” The dissertation author was a co-investigator and co-author of this work.

2.7 Author contributions

Conceptualization: TGL, AD, CS, REA, JP, KDC, EV; Methodology: TGL, AD, JP, KDC, EV; Validation: TGL, AD, CS, REA, JP, KDC, EV; Formal Analysis: TGL, AD, AMP, CS, REA, JP, KDC, EV; Investigation: TGL, AD, AMP, CS, YG, EE, SS, KK, EAB, EA; Data Curation: TGL, AD, CS; Writing - Original Draft: TGL, AD, CS, REA, KDC, EV; Writing- Review & Editing: AMP, YG, EE, SS, KK, EAB, EA, JAM, JP;

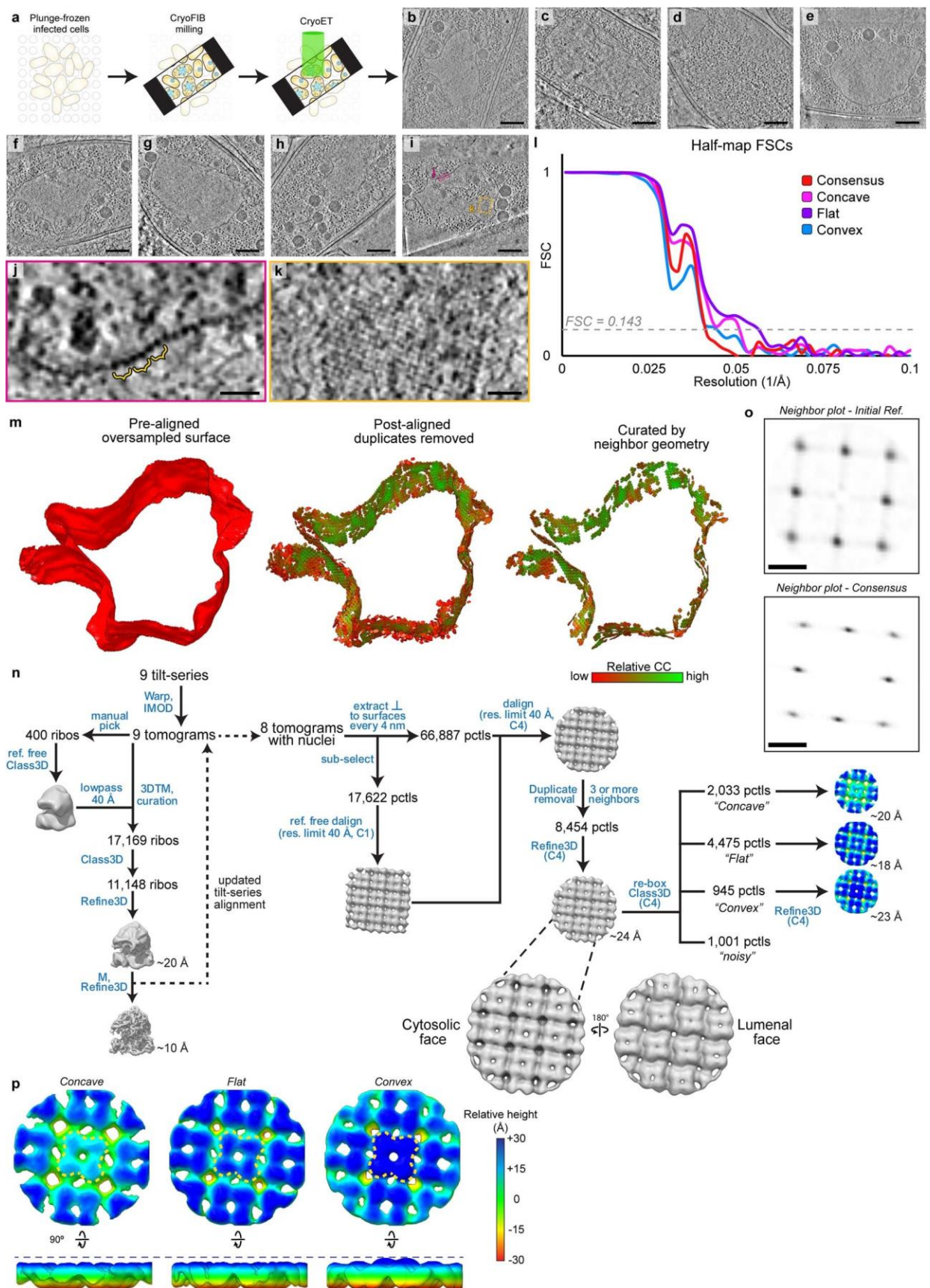
Visualization: TGL, AD, AMP, CS, KDC, EV; Supervision: JAM, REA, JP, KDC, EV; Funding acquisition: TGL, CS, REA, JAM, JP, KDC, EV.

2.8 Competing interests

The authors declare no competing interests.

2.9 Extended data figures and tables

Figure 2.6 | *In situ* cryoFIB-ET of 201phi2-1-infected *P. chlororaphis* cells and subtomogram analysis. **a**, Schematic of cryoFIB-ET workflow. **b-i**, Slices of the eight 201phi2-1 nucleus-containing tomograms used in this study. **j** Enlarged view of the colored boxed region in **i**. Exemplar doublets are indicated by yellow braces. **k** Enlarged view of the correspondingly colored boxed region in **i** which shows a square mesh-like texture corresponding to the square lattice. **l**, Half-map Fourier shell correlation (FSC) curves for the 201phi2-1 subtomogram reconstructions. **m**, Example over-sampling and subtomogram curation strategy using lattice plots for the tomogram shown in Figure 1c. **n**, Schematic of the subtomogram averaging workflow. Enlarged views of the consensus average with cytosolic and lumenal faces indicated. **o**, Neighbor plot of the initial (top), asymmetrically aligned reference. Neighbor plot of the symmetrized consensus (bottom) refinement. **p**, Enlarged views of the resolved classes colored by relative height. The central tetramer is denoted by a yellow, dashed line for each class. Scale bars: **b-i**: 250 nm, **j,k**: 25 nm, **o**: 10 nm.



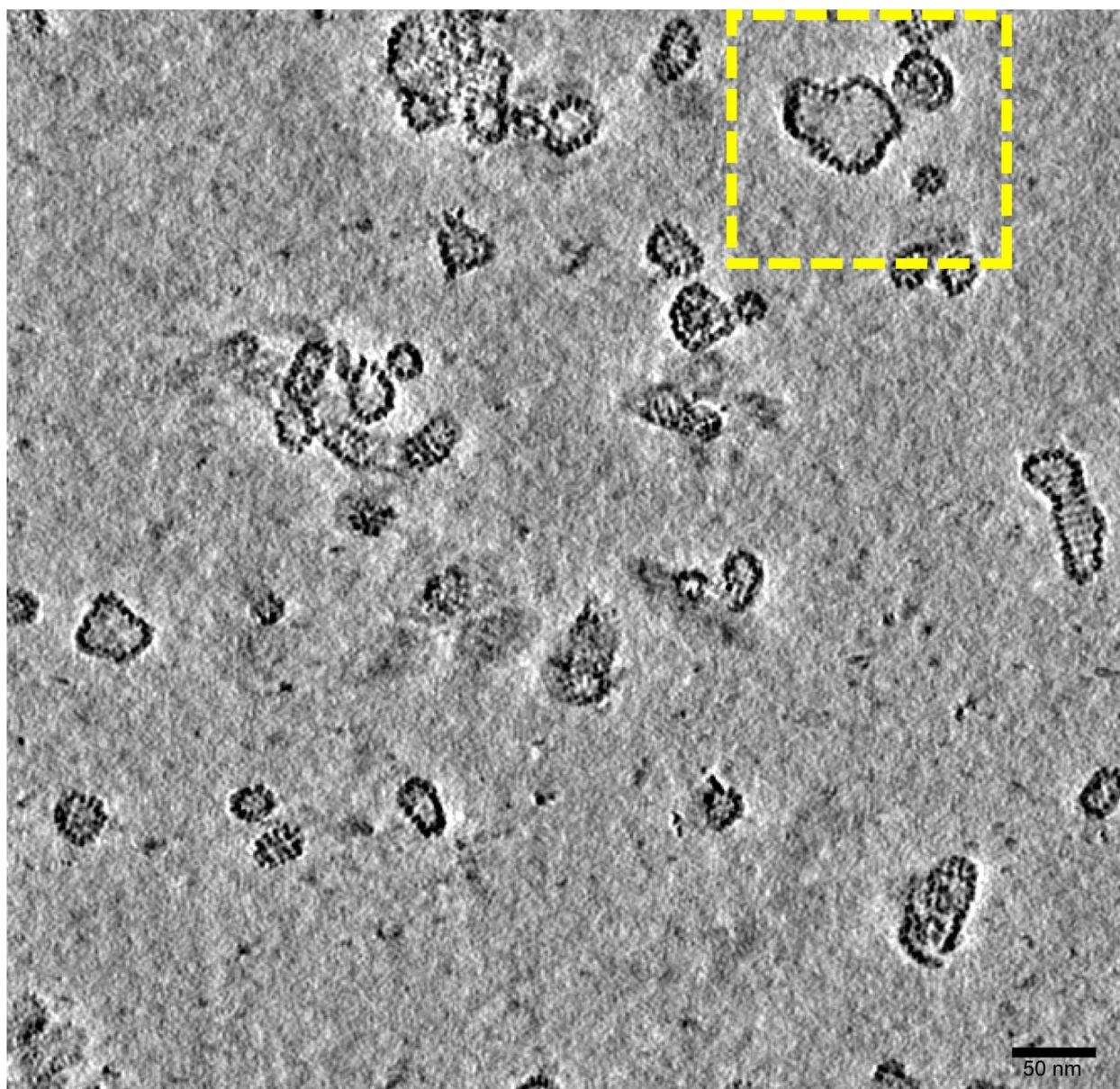


Figure 2.7 | In vitro tomography of 201phi2-1 chimallin void peak. Slice through the tomogram of the 201phi2-1 chimallin SEC size-exclusion chromatography void peak. Region marked by a dashed yellow box is used in Figure 2b. Scale bar is 50 nm.

Figure 2.8 | Single-particle reconstruction of the *in vitro* 201phi2-1 chimallin cubic assembly. **a**, Exemplar micrograph and 2D class averages. **b**, Schematic of the localized reconstruction workflow. **c**, Unsharpened density map views centered on helix B (residues 68-84) at progressive stages of the localized reconstruction process. Final view of the C1 map shown with a fitted coordinate model. **d**, Fourier shell correlation (FSC) curves for the half-maps at progressive stages of the localized reconstruction process (red, yellow, and blue), histogram of local resolution estimates for the C1 reconstruction (light blue). **e**, C1 reconstruction filtered and colored by local resolution. **f**, (left) 2D class average of the minor (517 particles) species of elongated, quasi-D4 assemblies. (right) Orthogonal views of the D4-symmetrized single particle reconstruction. Scale bars: **a**: 50 nm, **f**: 10 nm.

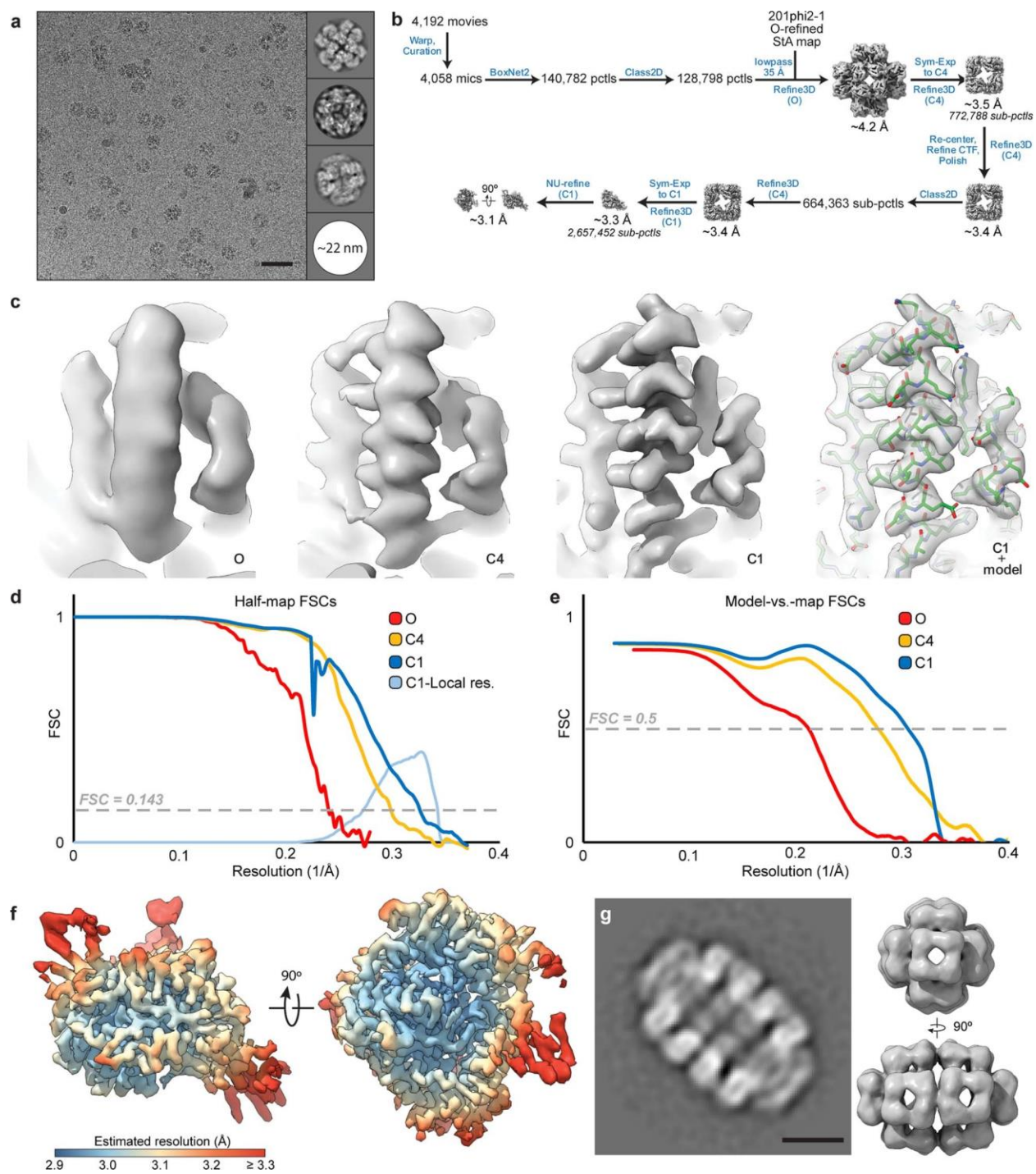
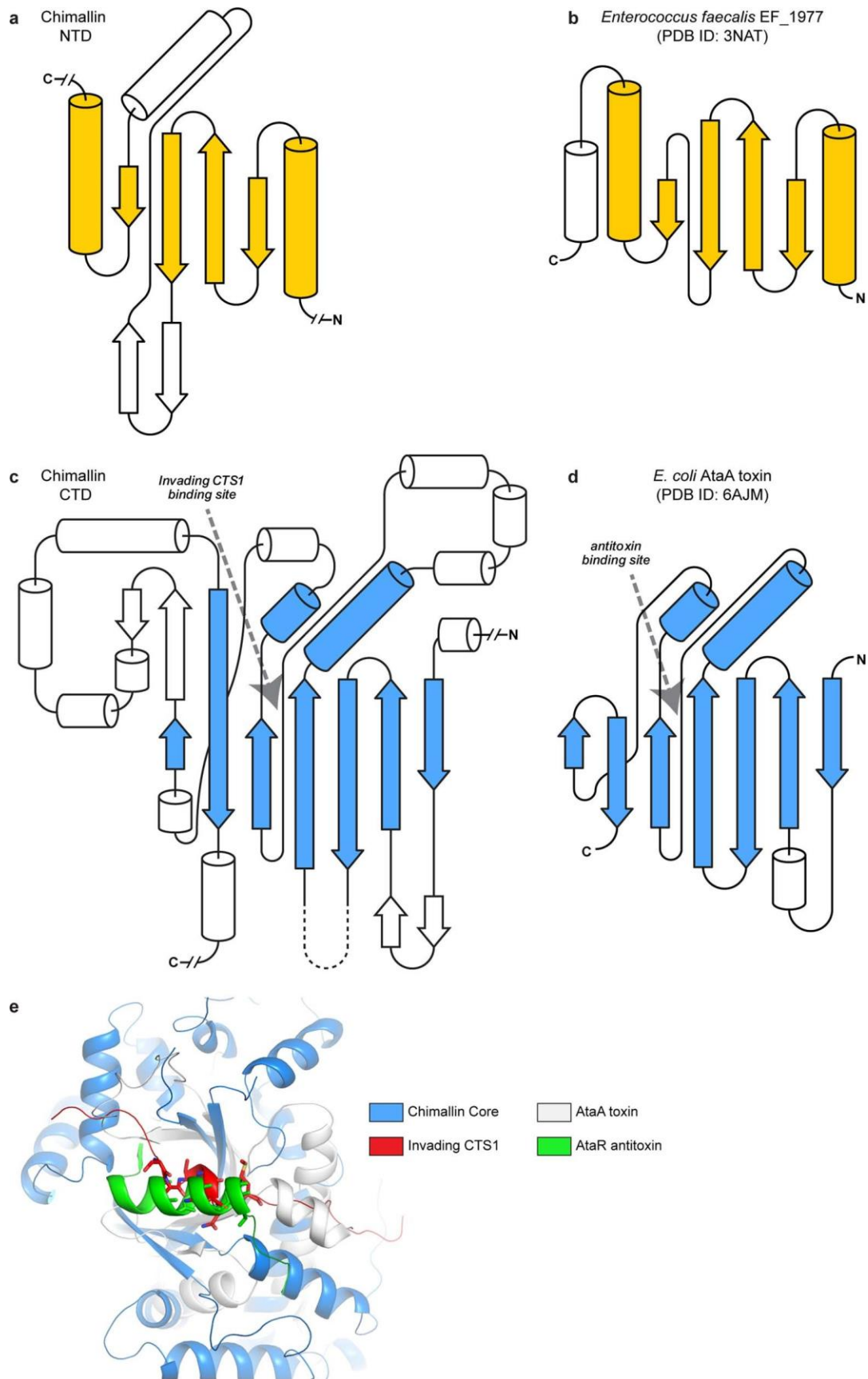


Figure 2.9 | Partial homology of chimallin fold topology to known structures. **a**, Topology of the 201phi2-1 chimallin N-terminal domain (NTD, residues 62-228). **b**, Topology of *E. faecalis* EF_1977 (PDB ID: 3NAT), the closest structural relative of the chimallin A NTD. The root mean square deviation (RMSD) between chimallin NTD and 3NAT coordinate models is 4.6 Å over 97 aligned Ca atoms. Homologous secondary structure elements are colored in yellow. **c**, Topology of the 201phi2-1 chimallin C-terminal domain (CTD, residues 229-581). **d**, Topology of the *E. coli* AtaT tRNA-acetylating toxin (PDB ID: 6AJM)¹¹. The root mean square deviation (RMSD) between chimallin CTD and 6AJM coordinate models is 4.2 Å over 269 aligned Ca atoms. Homologous secondary structure elements are colored in blue. **e**, Structural overlay of the chimallin CTD (blue) and AtaT (white; PDB ID 6AJM), showing the similarity in binding site for the chimallin CTS1 segment (red) and the antitoxin AtaR (green).



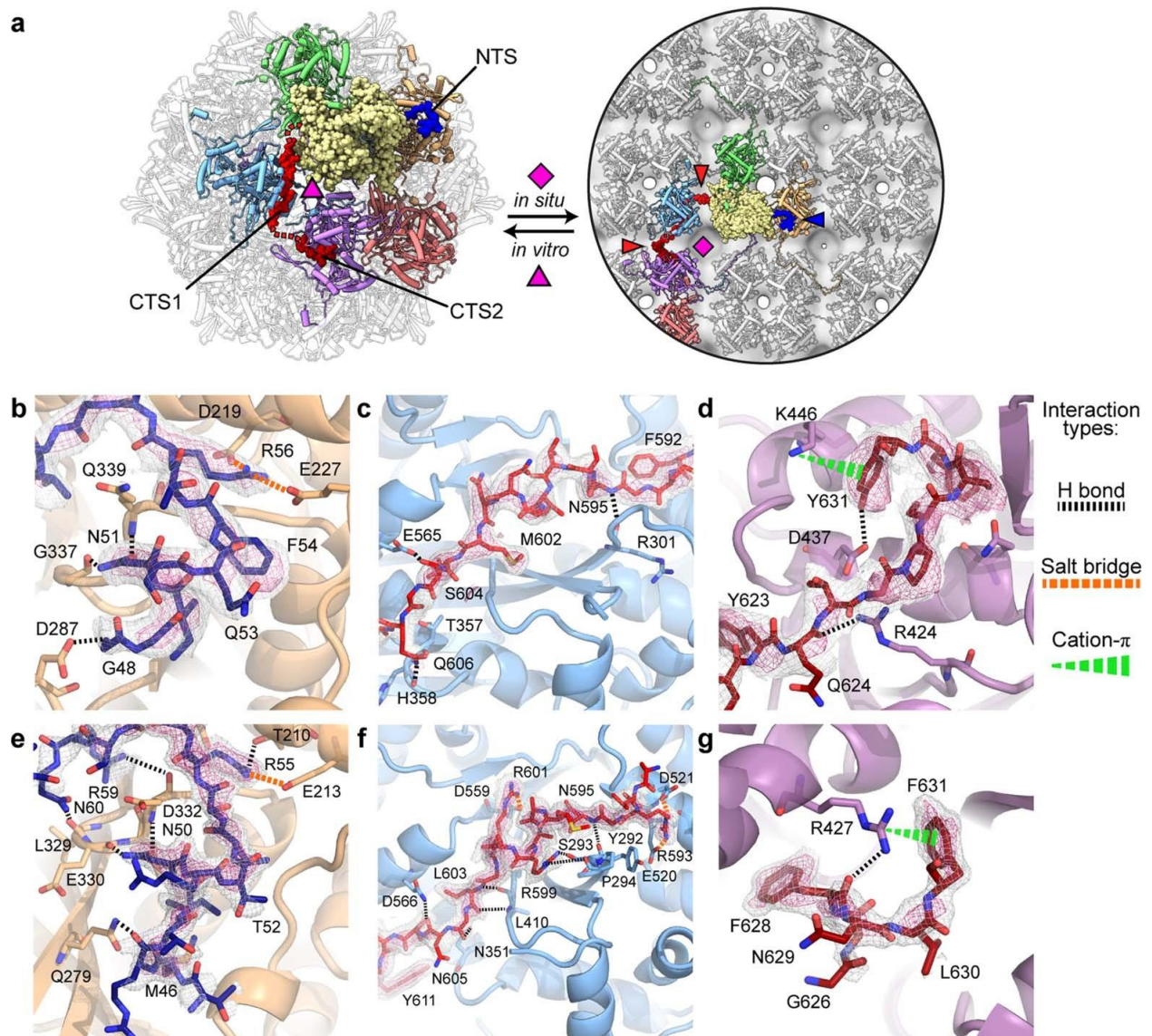


Figure 2.10 | Interactions of NTS and CTS segments with the chimallin core. **a**, Relationship of 201phi2-1 chimallin protomer packing in the cube (left) and flat sheet model (right). One protomer is shown as spheres and colored yellow with its NTS in blue and CTS1/CTS2 in red. Protomers that interact directly with this central protomer are colored. Non-interfacing protomers are in white. The flat sheet model is docked within the 201phi2-1 consensus subtomogram average map shown as transparent grey. Red arrows point to locations of unresolved linkers (red dashed lines), and pink symbols indicate 3- or 4-fold symmetry axes. **b-d**, Close-ups of the 201phi2-1 coordinate model around the binding sites for NTS (**b**), CTS1 (**c**), and CTS2 (**d**). **(e-g)** Close-ups of the Goslar coordinate model around the binding sites for NTS (**e**), CTS1 (**f**), and CTS2 (**g**). For all panels, cryo-EM density map is shown as a mesh at high (pink) and low (grey) contours. Polar interactions are depicted by the symbols indicated in the key at the far right.

Figure 2.11 | SEC-MALS of 201phi2-1 chimallin truncations. **a**, Domain diagram of 201phi2-1 chimallin (top), with truncations tested by SEC-MALS (bottom). **b-h**, SEC-MALS analysis of full-length 201phi2-1 chimallin (**b**) and truncated constructs lacking the N-tail (**c**), NTS (**d**), CTS2 (**e**), CTS1+CTS2 (**f**), N-tail+CTS2 (**g**), or NTS + CTS1/2 (**h**). For panels b-h, differential refractive index (dRI) shows protein concentration (blue curves), and yellow points indicate measured molecular weight. Average molecular weight for each peak is shown.

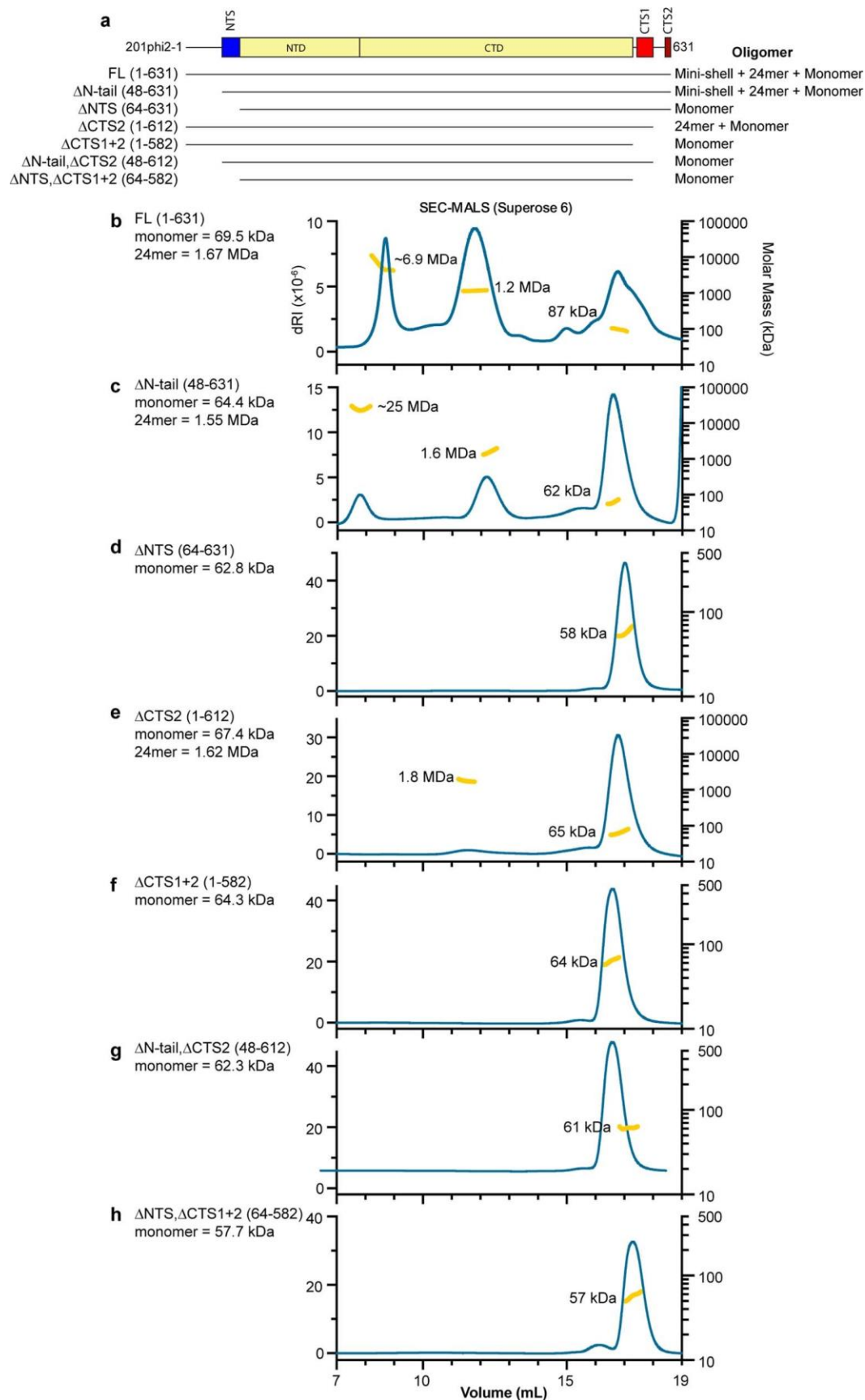
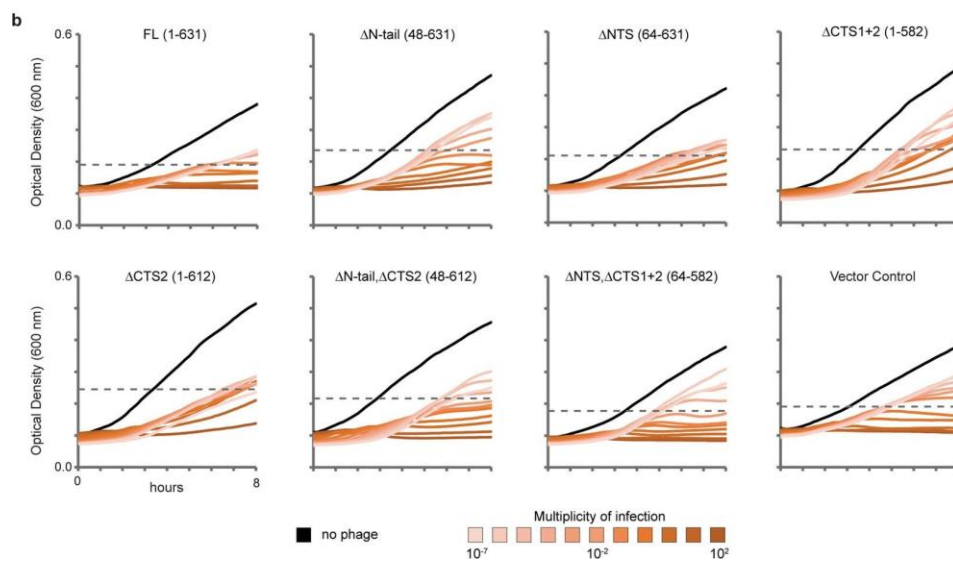
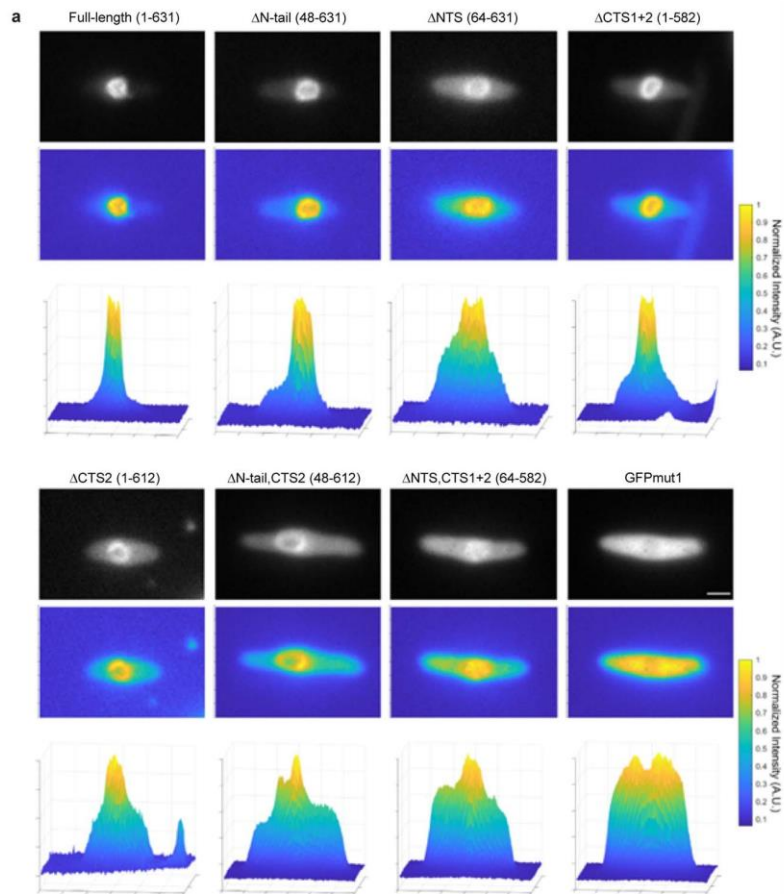


Figure 2.12 | GFP-chimallin incorporation into the nuclear shell in 201phi2-1-infected *P. chlororaphis* and truncation mutant growth curves. **a**, Raw microscopy images of representative cells expressing GFP-chimallin and infected with 201phi2-1 60 minutes post-infection (mpi) showing GFP fluorescence with associated 3D graphs showing normalized GFP fluorescence intensity within these cells from a top and side view. GFPmut1 was expressed without fusion to chimallin as a negative control and shows no incorporation. Growth curves for *P. chlororaphis* expressing the indicated 201phi2-1 chimallin truncation mutant (or empty vector control) and challenged with either no phage (black line) or increasing multiplicity of infection of 201phi2-1 (color key at the bottom) over a period of 8 hours. Dashed grey-line indicates the half of the maximal optical density at 600 nm achieved by the no phage control in each experiment. Curves are the average of four replicates (n = 4) of each condition. Scale bar: **a**: 1 μm .



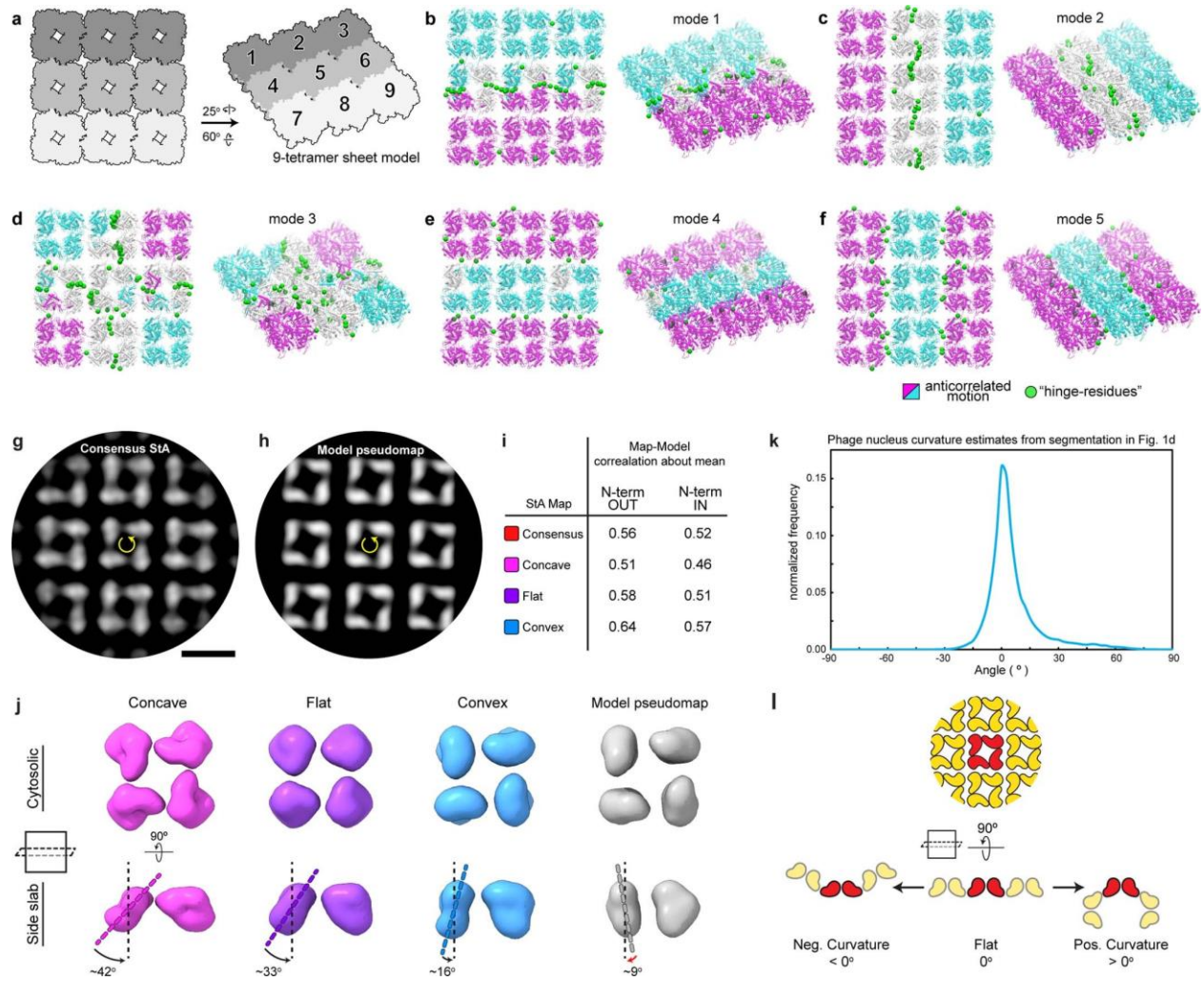
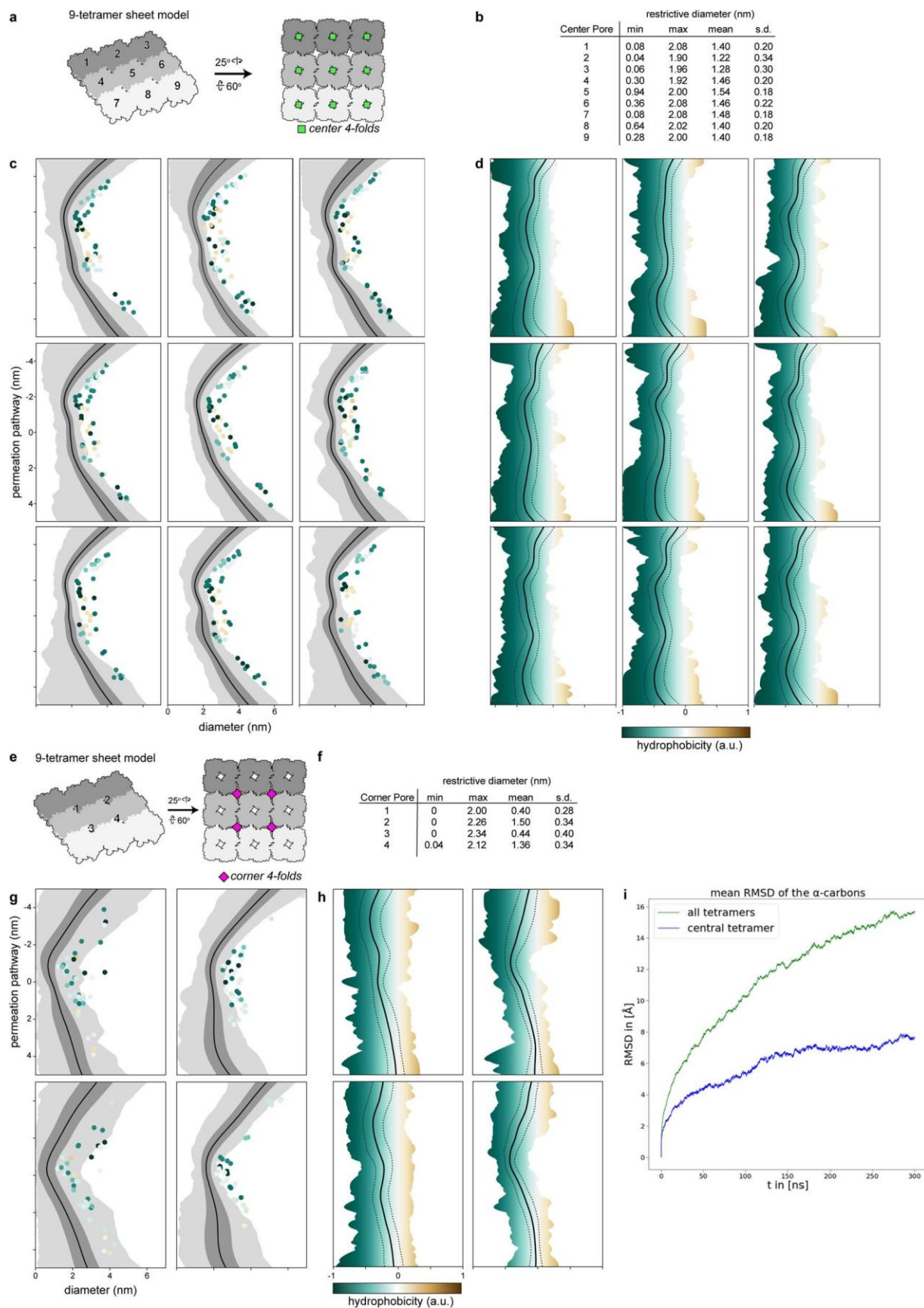


Figure 2.13 | Analysis of flexibility by Gaussian network models and subtomogram analysis. **a**, Cytosolic and tilted view schematic of 3x3 tetramer sheet model. **b-f**, Cartoon sheet model colored by results of Gaussian Network Model modes 1 through 5, respectively. Regions are colored according to the directional correlation of motion: positive (cyan), negative (magenta), and near-zero (white). “Hinge-residues” are depicted as green spheres. A list of the hinge-residues for each mode is in **Table 2.5**. **g**, Slice through the consensus subtomogram average for the 201phi2-1 nuclear shell, with the four-fold axis defining the central tetramer noted. **h**, Equivalent view of panel **g** from a pseudomap generating by fitting tetramers into the consensus subtomogram average. **i**, Model-map correlation coefficient about the mean (CC) for a tetramer model fit into the consensus subtomogram average and the three subclasses (concave, flat, and convex) within either the N-termini facing the cytosol (OUT) or lumen (IN). **j**, Two views of the concave, flat, and convex subclasses, compared to a pseudomap generated from the tetramer model. Dotted lines indicate the orientation of one monomer in each map. Denoted angles are with respect to the perpendicular. The arc arrow is red for the pseudomap to denote its opposite direction compared to the subtomogram average maps. **k**, Angles between tetramers in the phage nucleus lattice, derived from surface curvature estimates for phage nucleus segmentation in Figure 1d. **l**, Schematic of example manifestations of lattice curvature. The positive curvature shown on the right represents the 90° angle seen in the *in vitro* cubic assembly.

Figure 2.14 | Size and hydrophobicity profiles of the four-fold pores. **a**, Schematic of the 9-tetramer sheet model with the center four-fold pores marked with green squares. **b**, Pore diameter summary statistics for the nine center four-folds denoted in **a** over the course of the averaged 300-ns simulations ($n = 5$). **c**, Diameter profiles for each pore. The permeation pathway from top (negative values) to bottom (positive values) corresponds with cytosol to lumen. Solid black lines denote the mean diameter, dark gray shading $\pm$ one standard deviation, and light grey shading the range. Dots indicate pore-facing residues and are colored by hydrophobicity. **d**, Hydrophobicity profiles for each pore. Solid black lines denote the mean hydrophobicity, dashed lines $\pm$ one standard deviation, and shaded regions mark the range. **e**, Schematic of the 9-tetramer sheet model with the corner four-fold pores marked with pink squares. **f**, Pore diameter summary statistics for the four corner four-folds denoted in **a** over the course of the averaged 300-ns simulations ($n = 5$). **g**, Same as **c** for the corner four-fold pores. **h**, Same as **d** for the corner four-fold pores. **i**, Mean root mean square deviation (RMSD) of the alpha-carbons in the 3x3 tetramer sheet model over the course of the simulations for all alpha-carbons (green) and for those just within the central tetramer (blue). The central tetramer is embedded in a physiological environment, flanked by other tetramers. The edge tetramers continue to display an increasing RMSD since they are not connected to adjacent tetramers. Our analysis in the text stems from the central pore and corner pores formed by this tetramer.



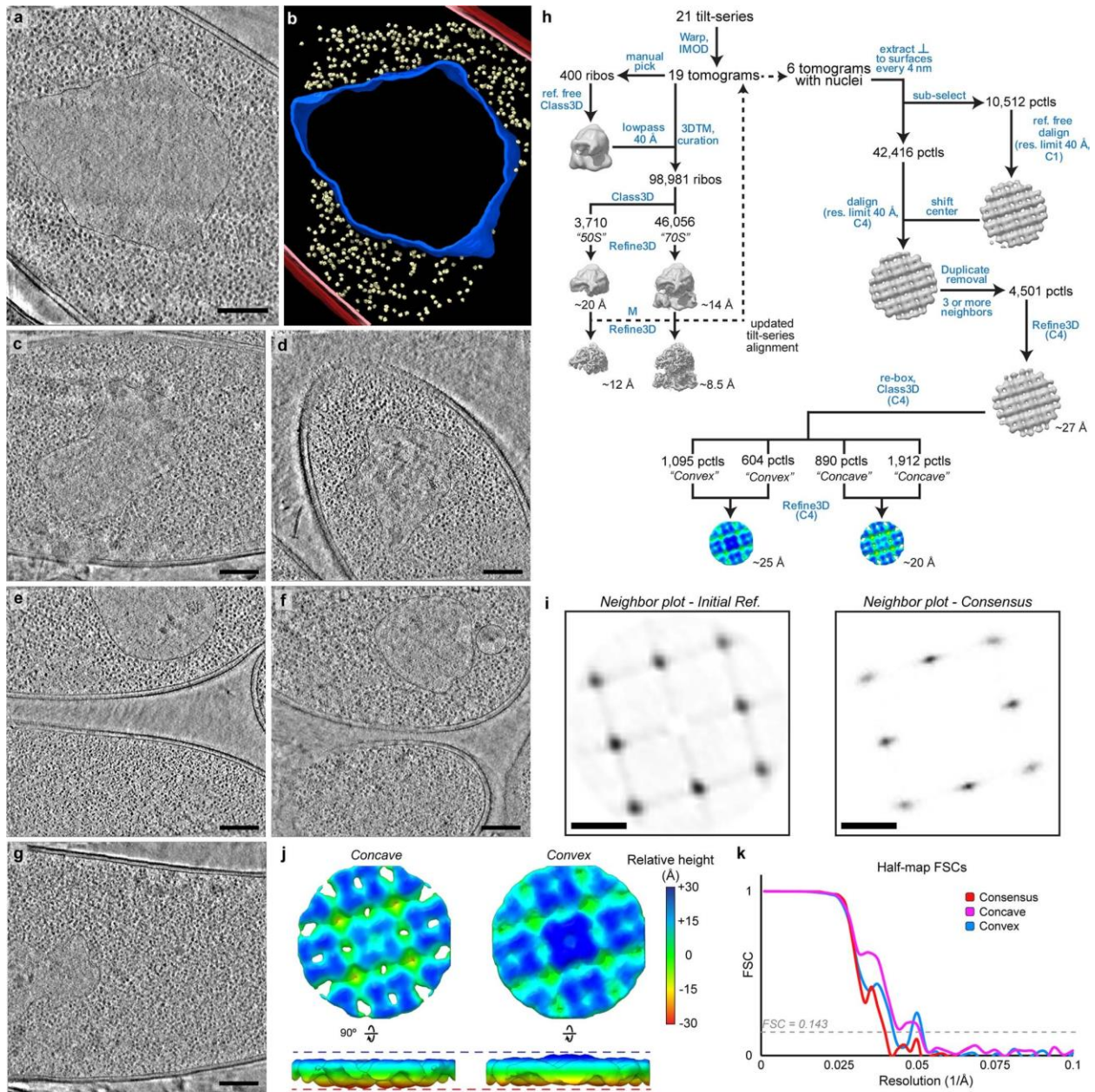
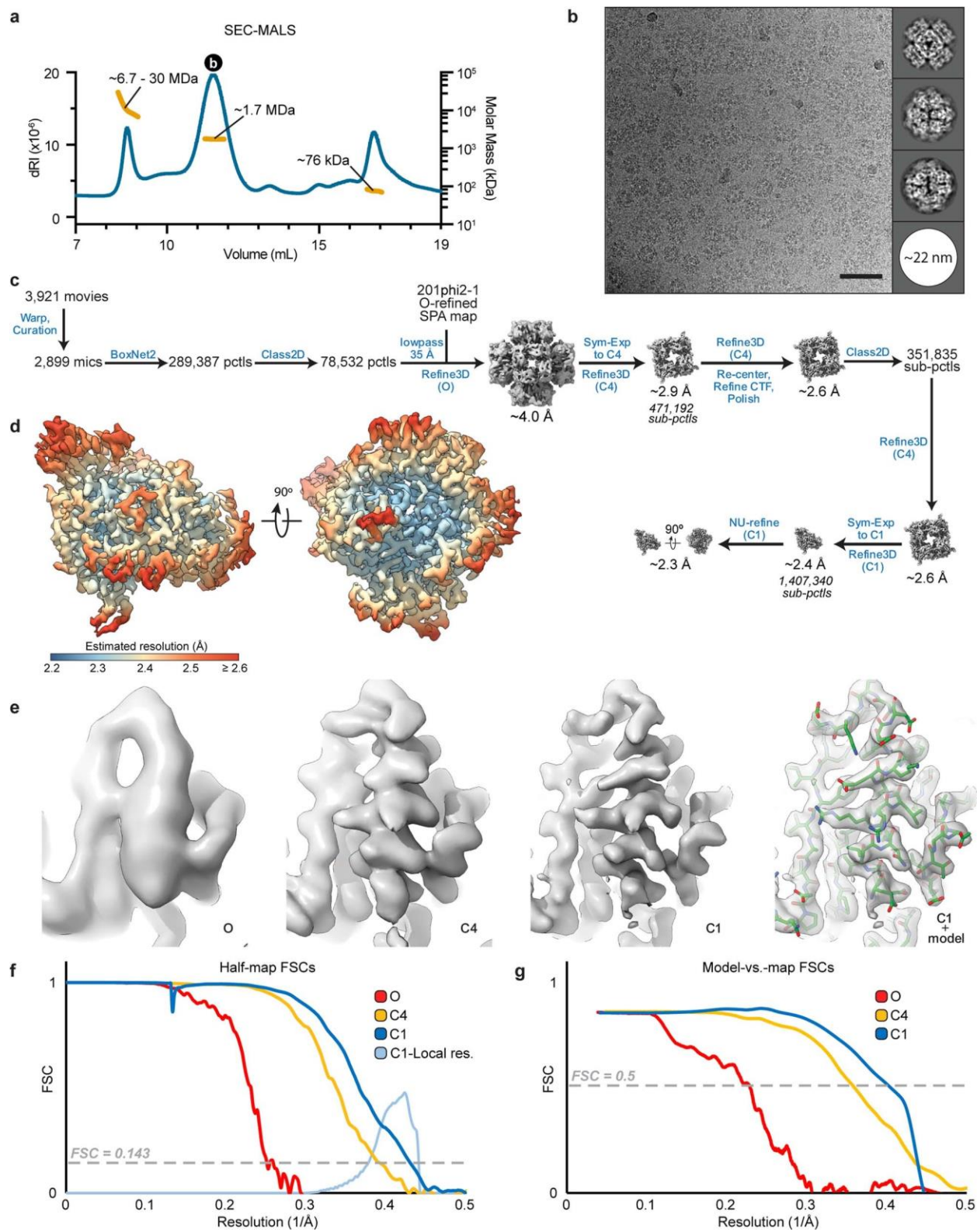


Figure 2.15 | *In situ* cryoFIB-ET of Goslar infected APEC2248 cells and subtomogram analysis. **a**, Tomographic slice of a Goslar-nucleus and **b** the corresponding segmentation model. Outer and inner bacterial membranes are burgundy and pink, respectively. The phage nucleus is colored blue and host ribosomes are colored pale yellow. Five-hundred randomly selected 70S ribosomes are placed for clarity. **c-g**, A slice from each of the Goslar-nucleus containing tomograms used for subtomogram averaging in this study. The cells were plunged at effectively ~20-30 mpi, thus too early to observe virion assembly. **h**, Schematic of the subtomogram averaging workflow. **i**, Neighbor plots of the asymmetrically aligned initial reference (left) and symmetrized consensus refinement (right). **j**, Enlarged views of the resolved classes colored by relative height. **k**, Half-map Fourier shell correlation (FSC) curves for the subtomogram reconstructions.

Figure 2.16 | Single-particle reconstruction of the *in vitro* Goslar chimallin cubic assembly. **a**, Size-exclusion coupled to multi-angle light scattering (SEC-MALS) analysis of purified, full-length Goslar chimallin. **b**, Exemplar micrograph and 2D class averages. **c**, Schematic of the localized reconstruction workflow. **d**, C1 reconstruction filtered and colored by local resolution estimates. **e**, Unsharpened density map views centered on helix B (residues 64-78) at progressive stages of the localized reconstruction process. Final view of the C1 map shown with a fitted coordinate model. **f,g**, Fourier shell correlation (FSC) curves for the half-maps and against corresponding models at progressive stages of the localized reconstruction process (red, yellow, and blue), histogram of local resolution estimates for the C1 reconstruction (light blue), and the C1 model-vs-map FSC curve (black). Scale bar: **b**: 50 nm.



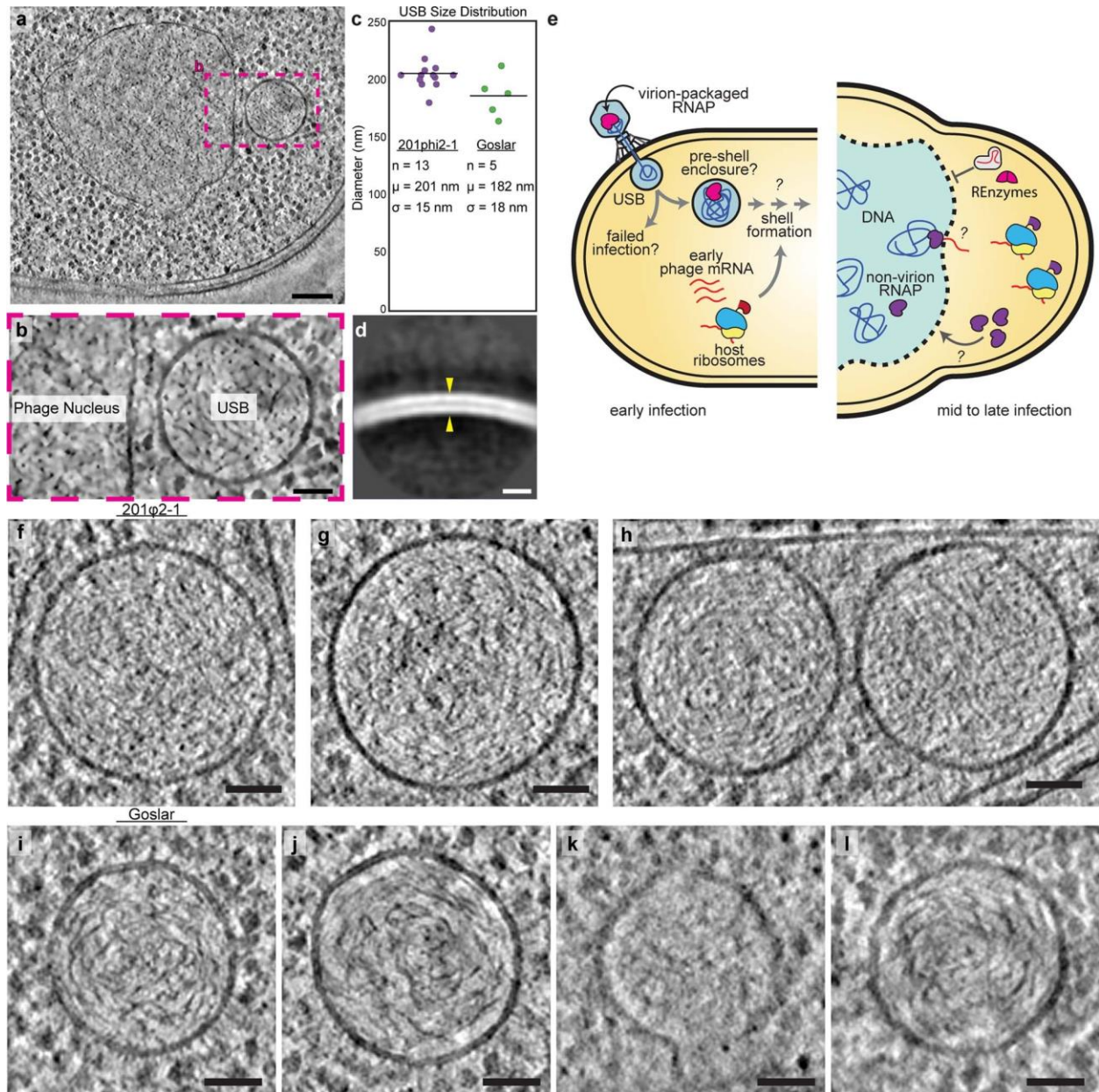


Figure 2.17 | Unidentified spherical bodies present in jumbo phage-infected cell populations and speculative models. **a**, Tomographic slice of Goslar-infected APEC2248 cell containing a bonafide phage nucleus, as well as an unidentified spherical body (USB). **b**, Enlarged view of the phage nucleus and USB from the region boxed in **a**. **c**, Plot of the apparent maximal diameter distributions for 201phi2-1 (purple) and Goslar (green) USBs with the summary statistics listed. **d**, Subtomogram average of the USBs picked from the Goslar dataset. Yellow arrow pointing to putative membrane leaflets. **e**, Left, model of USBs as the previously proposed pre-shell/nucleus enclosure of the phage DNA³. Right, schematic summary of structural models in this work: (i) exclusion of host nucleases by small chimallin pore sizes, (ii) possible extrusion of phage mRNA via these pores, and (iii) implication of additional shell components to enable uptake of specific phage proteins into the phage nucleus. **f-h**, Gallery of USBs observed in tomograms of 201phi2-1-infected cell populations. **i-l**, Gallery of USBs observed in tomograms of Goslar-infected cell populations. Scale bars: **a**: 150 nm, **b**: 50 nm, **d**: 10 nm, **f-l**: 50 nm.

2.10 Supplementary information

This file contains Figures 2.18 and Tables 2.1-2.10. SI Figure shows an alignment for the 201phi2-1 and PhiKZ chimallin. Table contents: cryo-EM map statistics, coordinate model statistics, protein-protein interface statistics, hinge residues, pore-facing residues, sequences used to construct the tree in Figure 2.5a, and the bacterial strains and plasmids used in this study.

Figure 2.18 | Alignment of 201phi2-1 and phiKZ chimallin proteins. Alignment of 201phi2-1 gp105 and phiKZ gp54 protein sequences. Secondary structure and domain annotation are shown above the alignment and are based on the 201phi2-1 structure. Dashed lines in the domain annotation indicate unresolved segments. The phiKZ prohead protease (gp175) cleavage-site in gp54 is indicated by a yellow arrow. Cleavage-site based on that identified by Weintraub *et al.*²⁰.

Table 2.1 | *In situ* cryoFIB-ET data collection and reconstruction statistics

	201phi2-1 infected <i>P. chlororaphis</i>					Goslar infected <i>E. coli</i> (APEC2248)				
	chimallin					chimallin				
	70S (EMD- 25183)	Consensus (EMD- 25221)	Concave (EMD- 25220)	Flat (EMD- 25222)	Convex (EMD- 25223)	50S (EMD- 25360)	70S (EMD- 25359)	Consensus (EMD- 25229)	Concave (EMD- 25262)	Convex (EMD- 25358)
Data collection										
Magnification			33,000						42,000	
Voltage (kV)						300				
Defocus range (µm)			-4.5 to -6						-5 to -6	
Pixel size (Å)			3.46						4.27	
Electron exposure (e ⁻ /Å ²)			120						150	
Tilt-range/step (°)			±51 / 3						±56 / 2	
Tilt-scheme						dose-symmetric, grouping 2				
Processing										
Symmetry imposed	C1	C4	C4	C4	C4	C1	C1	C4	C4	C4
Initial particle images (no.)	17,169	66,887		8,454			98,981	42,416		4,501
Final particle images (no.)	11,148	8,454	2,033	4,475	945	3,710	46,056	4,501	2,802	1,699
Map resolution (Å)	10	24	20	18	23	12	8.5	27	20	25
FSC threshold						0.143				

Table 2.2 | Cryo-EM data collection, reconstruction, and refinement statistics

	201phi2-1 gp105			Goslar gp189		
	24mer (EMD- 25390) (PDB 7SQQ)	tetramer (EMD- 25391) (PDB 7SQR)	monomer (EMD- 25392) (PDB 7SQS)	24mer (EMD- 25394) (PDB 7SQT)	tetramer (EMD- 25395) (PDB 7SQU)	monomer (EMD- 25396) (PDB 7SQV)
Data collection and processing						
Magnification		130,000			165,000	
Voltage (kV)			300			
Electron exposure (e ⁻ /Å ²)		43			40	
Defocus range (μm)			-0.1 to -1.5			
Pixel size (Å)		1.075			0.8452	
Symmetry imposed	O	C4	C1	O	C4	C1
Initial particle images (no.)	140,782	772,788	2,657,452	289,387	471,192	1,407,340
Final particle images (no.)	128,798	664,363	2,657,452	78,532	351,835	1,407,340
Map resolution (Å)	4.2	3.4	3.1	4.0	2.6	2.3
FSC threshold			0.143			
Map resolution range (Å)	4.0 to 5.1	3.3 to 4.4	2.9 to 4.8	3.7 to 6.0	2.5 to 3.4	2.3 to 3.4
Refinement						
Initial model used (PDB code)		<i>de novo</i>			7SQS	
Model resolution (Å)	4.7	3.6	3.3	4.4	2.8	2.5
FSC threshold			0.5			
Model composition						
Non-hydrogen atoms	103,632	17,272	4,376	107,136	17,856	4,471
Protein residues	13,296	2,216	560	13,752	2,292	573
Ligands	0	0	0	0	0	7
B factors (Å ²)						
Protein	408.17	125.69	108.00	318.81	97.77	77.87
Ligand (waters)	n.a.	n.a.	n.a.	n.a.	n.a.	71.61
R.m.s. deviations						
Bond lengths (Å)	0.003	0.003	0.002	0.010	0.010	0.001
Bond angles (°)	0.633	0.632	0.469	0.699	0.698	0.380

Table 2.2 | Cryo-EM data collection, reconstruction, and refinement statistics (continued)

Validation						
MolProbity score	1.63	1.56	1.31	1.16	0.94	0.87
Clashscore	8.23	6.82	3.59	2.13	1.16	1.36
Poor rotamers (%)	0	0	0.22	1.25	1.04	0.63
Ramachandran plot						
Favored (%)	96.89	96.89	97.08	97.53	97.53	99.29
Allowed (%)	3.11	3.11	2.92	2.47	2.47	0.71
Disallowed (%)	0	0	0	0	0	0

Table 2.3 | Summary of *in vitro* 201phi2-1 chimallin protomer interfaces

Chain-1	Atoms	Residues	Chain-2	Atoms	Residues	Area (Å ²)	H bonds	Salt bridge	Cation- π
A	28	9	F	28	9	307	0	0	0
A	130	34	H	147	43	1253	7	6	0
A	147	43	B	130	34	1253	7	6	0
A	198	53	D	205	56	1945	14	0	1
A	205	56	G	198	53	1945	14	0	1

Table 2.4 | Summary of a 201phi2-1 chimallin protomer's polar interactions *in vitro*

Chain-1	Resi	Resn	Atom	Chain-2	Resi	Resn	Atom	Dist (Å)	Type		
A	B	48	GLY	N	H	A	287	ASP	OD2	3.6	H bond
A	B	51	ASN	NE2	H	A	337	GLY	O	3.2	H bond
A	B	51	ASN	OD1	H	A	339	GLN	N	3.4	H bond
A	B	56	ARG	O	H	A	339	GLN	NE2	2.6	H bond
A	B	59	ASN	N	H	A	86	ASP	O	3.5	H bond
A	B	102	GLN	OE1	H	A	281	ASN	ND2	2.3	H bond
A	B	163	SER	OG	H	A	282	GLU	OE2	3.1	H bond
A	B	56	ARG	NE	H	A	219	ASP	OD1	2.5	Salt bridge
A	B	56	ARG	NH1	H	A	227	GLU	OE1	3.4	Salt bridge
A	B	56	ARG	NH1	H	A	227	GLU	OE2	3.7	Salt bridge
A	B	56	ARG	NH2	H	A	227	GLU	OE1	3.9	Salt bridge
A	B	56	ARG	NH2	H	A	227	GLU	OE2	2.9	Salt bridge
A	B	56	ARG	NH2	H	A	219	ASP	OD1	3.3	Salt bridge
A	G	319	THR	OG1	D	A	319	THR	N	2.9	H bond
A	G	406	GLN	OE1	D	A	306	GLN	N	3.9	H bond
A	G	424	ARG	NH2	D	A	624	GLN	O	3.3	H bond
A	G	430	GLN	OE1	D	A	629	SER	N	3.8	H bond
A	G	430	GLN	OE1	D	A	629	SER	OG	2.8	H bond
A	G	437	ASP	OD2	D	A	631	TYR	OH	2.8	H bond
A	G	592	PHE	O	D	A	516	TRP	NE1	2.3	H bond
A	G	595	ASN	N	D	A	301	ARG	O	3.5	H bond
A	G	604	SER	N	D	A	565	GLU	OE2	3	H bond
A	G	604	SER	OG	D	A	565	GLU	OE2	3.4	H bond
A	G	605	GLY	N	D	A	565	GLU	OE2	3.6	H bond
A	G	606	GLN	NE2	D	A	358	HIS	O	3.1	H bond
A	G	607	ALA	N	D	A	357	THR	OG1	3.8	H bond
A	G	610	HIS	N	D	A	578	ILE	O	3.1	H bond
A	G	446	LYS	NZ	D	A	631	TYR	-	5.8	Cation- π

Table 2.5 | Gaussian network model hinge residues for 201phi2-1 chimallin

resi resn	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5	total
45 ALA	1	2	2	4	4	13
46 GLY	5	5	4	-	-	14
47 THR	-	1	1	-	2	4
50 ILE	-	-	1	-	-	1
51 ASN	1	-	-	-	-	1
60 VAL	-	-	1	-	-	1
62 GLY	2	3	2	-	-	7
63 GLY	4	3	3	-	-	10
64 ASP	-	-	1	-	-	1
66 ARG	-	-	1	-	-	1
97 ASP	-	-	1	-	-	1
106 SER	-	-	1	-	-	1
132 ASN	-	-	1	-	-	1
138 LYS	-	-	1	-	-	1
161 VAL	-	-	1	-	-	1
163 SER	-	-	1	-	-	1
166 TYR	-	-	1	-	-	1
239 GLN	-	-	1	-	-	1
241 GLU	-	-	1	-	-	1
242 THR	-	-	1	-	-	1
243 LEU	-	-	1	-	-	1
244 ALA	-	-	1	-	-	1
271 GLN	-	-	1	-	-	1
276 ASN	1	1	1	-	-	3
277 GLY	3	2	2	-	-	7
278 GLN	1	2	1	-	-	4
279 GLN	1	1	1	-	-	3
280 GLU	-	-	1	-	-	1
284 TYR	-	1	1	-	-	2
285 GLU	3	2	2	-	-	7
286 THR	3	3	3	-	-	9
289 LYS	-	-	1	-	-	1
336 ASP	-	-	1	-	-	1
337 GLY	-	-	1	-	-	1
358 HIS	-	-	1	-	-	1
368 PRO	-	-	1	-	-	1
370 THR	2	1	2	-	-	5
371 GLY	1	1		-	-	2
372 VAL	3	2	2	-	-	7
373 ALA	-	1	1	-	-	2
375 ASP	-	-	1	-	-	1
396 ALA	-	-	1	-	-	1
397 ALA	-	-	1	-	-	1
400 ASP	-	-	1	-	-	1
441 GLY	-	-	-	-	1	1
442 PRO	-	-	-	-	1	1
446 LYS	-	-	1	-	-	1
526 ASN	-	-	1	-	-	1
538 ASN	-	-	1	-	-	1
541 ARG	-	-	1	-	-	1
543 TYR	-	-	1	-	-	1
545 GLY	-	-	1	-	-	1
546 SER	-	-	1	-	-	1
568 ASP	-	-	1	-	-	1
580 MET	-	-	1	-	-	1
582 ASN	-	-	-	-	1	1
583 THR	-	-	-	4	1	5
584 SER	1	2	1	3	5	12

Table 2.5 | Gaussian network model hinge residues for 201phi2-1 chimallin (continued)

resi resn	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5	total
585VAL	-	-	1	-	-	1
586LEU	-	-	-	5	5	10
587ASN	-	-	-	-	1	1
588SER	1	1	2	-	-	4
589GLY	-	-	1	-	-	1
596SER	-	1	-	1	-	2
606GLN	-	-	1	-	-	1
607ALA	1	-	1	1	-	3
611SER	-	-	1	-	-	1
612ALA	4	1	6	4	10	25
total	38	36	78	22	311	

Table 2.6 | Pore-facing residues during 201phi2-1 chimallin simulations

Center 4-fold Corner 4-fold
resi resn countresi resn count

69 GLU	2	117 ASN	3
72 THR	6	119 GLN	2
76 ARG	9	303 PRO	1
79 GLU	9	304 GLN	1
80 GLN	6	305 ALA	1
83 ALA	5	307 ALA	1
131 ASN	1	309 THR	1
133 SER	1	311 PHE	1
196 ASP	2	312 PRO	1
198 VAL	9	317 VAL	1
200 LYS	9	319 THR	1
201 ASP	8	320 PRO	2
203 ARG	1	414 PRO	1
203 GLU	3	415 ASN	1
372 VAL	4	445 GLN	1
373 ALA	1	452 ILE	1
374 LYS	8	453 ARG	2
375 ASP	2	456 ASN	1
377 LYS	1	464 GLU	4
378 ASP	1	465 ARG	2
391 ARG	5	466 PHE	1
395 LYS	4	468 ASP	3
480 GLY	1	469 HIS	1
488 PHE	9	470 THR	4
490 GLY	8	471 THR	4
491 ASP	4	472 GLN	2
493 LYS	6	562 LYS	2
509 ASP	1	569 ARG	1
509 GLU	8	573 GLU	1
511 ASN	1	596 SER	1
546 SER	5	600 ASN	3
547 THR	5	601 ASN	1
		602 MET	1
		604 SER	2
		606 GLN	1
		607 ALA	1

Table 2.7 | Bacteriophage chimallin proteins

Genbank ID	Protein ID	Scientific Name
QDB70465.1	gp052	<i>Aeromonas</i> phage CF8
QDJ97041.1	gp044	<i>Aeromonas</i> phage D3
QDH47031.1	gp211	<i>Aeromonas</i> phage LAh10
QDJ96807.1	gp048	<i>Aeromonas</i> phage PS1
BCM95275.1	gp239	<i>Burkholderia</i> phage FLC6
YP_008125763.1	gp023	<i>Cronobacter</i> phage CR5
YP_009821902.1	gp002	<i>Edwardsiella</i> phage pEt-SU
YP_009005014.1	gp219	<i>Erwinia</i> phage Ea35-70
QQO90522.1	gp038	<i>Erwinia</i> phage pEa_SNUABM_43
YP_009010075.1	gp022	<i>Erwinia</i> phage PhiEaH1
YP_007237854.1	gp204	<i>Erwinia</i> phage phiEaH2
YP_009290657.1	gp039	<i>Erwinia</i> phage vB_EamM_Asesino
YP_009292105.1	gp248	<i>Erwinia</i> phage vB_EamM_Caitlin
YP_009292727.1	gp253	<i>Erwinia</i> phage vB_EamM_ChrisDB
YP_009278336.1	gp024	<i>Erwinia</i> phage vB_EamM_EarlPhillipIV
YP_009293007.1	gp244	<i>Erwinia</i> phage vB_EamM_Huxley
YP_009278639.1	gp265	<i>Erwinia</i> phage vB_EamM_Kwan
YP_009283513.1	gp024	<i>Erwinia</i> phage vB_EamM_Phobos
ANH52002.1	gp222	<i>Erwinia</i> phage vB_EamM_RAY
YP_009612772.1	gp028	<i>Erwinia</i> phage vB_EamM_RisingSun
YP_009606010.1	gp299	<i>Erwinia</i> phage vB_EamM_Simmy50
QKN85862.1	gp240	<i>Escherichia</i> phage vB_EcoM_EC001
* YP_009820873.1 gp189		<i>Escherichia</i> phage vB_EcoM_Goslar
YP_006383413.1	gp133	<i>Halocynthia</i> phage JM-2012 (infects <i>Vibrio</i>)
QPB09376.1	gp281	<i>Klebsiella</i> phage Miami
QGH71899.1	gp036	<i>Klebsiella</i> phage N1M2
QQV92022.1	gp083	<i>Klebsiella</i> phage vB_KpM_FBKp24
CAD5236077.1	gp088	<i>Klebsiella</i> phage vB_KvM-Eowyn
YP_009853356.1	gp004	<i>Photobacterium</i> phage PDCC-1
QMP24145.1		<i>Proteus</i> phage 10
QNN97444.1		<i>Proteus</i> phage 7
* YP_001956829.1 gp105		<i>Pseudomonas</i> phage 201phi2-1
* YP_418056.1	gp023	<i>Pseudomonas</i> phage EL
ANM44831.1	gp073	<i>Pseudomonas</i> phage KTN4

Table 2.7 | Bacteriophage chimallin proteins (continued)

Genbank ID	Protein ID	Scientific Name
YP_009608965.1	gp050	<i>Pseudomonas</i> phage Noxifer
YP_004957964.1	gp057	<i>Pseudomonas</i> phage OBP
BBI55859.1	gp186	<i>Pseudomonas</i> phage PA02
QBX32206.1		<i>Pseudomonas</i> phage PA1C
YP_009617618.1	gp330	<i>Pseudomonas</i> phage PA7
ARV76713.1	gp082	<i>Pseudomonas</i> phage Phabio
* YP_009217136.1	gp054	<i>Pseudomonas</i> phage phiPA3
QBJ02586.1	gp056	<i>Pseudomonas</i> phage Psa21
YP_009619844.1	gp304	<i>Pseudomonas</i> phage SL2
QOV07924.1		<i>Pseudomonas</i> phage vB_PaeM_kmuB
QEM41783.1		<i>Pseudomonas</i> phage vB_PaeM_PS119XW
* NP_803620.1	gp054	<i>Pseudomonas</i> virus phiKZ
YP_009598981.1	gp133	<i>Ralstonia</i> phage RP12
* BAW19549.1	gp286	<i>Ralstonia</i> phage RP31
YP_009208033.2	gp229	<i>Ralstonia</i> phage RSF1
* YP_009213072.1	gp223	<i>Ralstonia</i> phage RSL2
QNI20738.1		<i>Salmonella</i> phage JN03
QOC54489.1	gp029	<i>Salmonella</i> phage pSal-SNUABM-04
* YP_009153316.1	gp244	<i>Salmonella</i> phage SPN3US
AWN08967.1	gp006	<i>Salmonella</i> phage STsAS
QKE54546.1	gp230	<i>Salmonella</i> phage vB_SalM_SA002
YP_009833990.1	gp310/311/312	<i>Serratia</i> phage 2050HW
YP_009849136.1	gp040	<i>Serratia</i> phage Moabite
* QFP93061.1		<i>Serratia</i> phage PCH45
YP_009843316.1	gp214	<i>Vibrio</i> phage 2 TSL-2019
* YP_009622192.1	gp065	<i>Vibrio</i> phage Aphrodite1
* AXH70812.1	gp071	<i>Vibrio</i> phage BONAISHI
YP_009599417.1	gp130	<i>Vibrio</i> phage pTD1
AQT28149.1	gp208	<i>Vibrio</i> phage pVa-21
YP_009847864.1	gp128	<i>Vibrio</i> phage USC-1
QNI21050.1		<i>Vibrio</i> phage vB_pir03
QAX96061.1		<i>Vibrio</i> phage vB_VmeM-Yong XC31
YP_009626053.1	gp077	<i>Vibrio</i> phage VP4B
QNR51901.1	gp210	<i>Xanthomonas</i> phage Xoo-sp14

Table 2.8 | Summary of *in vitro* Goslar chimallin protomer interfaces

Chain-1	Atoms	Residues	Chain-2	Atoms	Residues	Area (Å ²)	H bonds	Salt bridge	Cation- π
A	11	3	F	11	3	91	0	2	0
A	164	38	H	210	62	1779	18	11	1
A	210	62	B	164	38	1779	18	11	1
A	206	54	D	210	58	2024	16	6	3
A	210	58	G	206	56	2024	16	6	3

Table 2.9 | Summary of a Goslar chimallin protomer's polar interactions *in vitro*

Chain-1	Resi	Resn	Atom	Chain-2	Resi	Resn	Atom	Dist (Å)	Type	
A	F	140	HIS	NE2	F	A 427	ASP	OD2	3.2	Salt bridge
A	B	46	MET	O	H	A 279	GLN	NE2	3.5	H bond
A	B	50	ASN	ND2	H	A 330	GLU	O	3.4	H bond
A	B	50	ASN	OD1	H	A 332	ASP	N	3.5	H bond
A	B	55	ARG	NE	H	A 206	ASN	OD1	3.1	H bond
A	B	55	ARG	NH2	H	A 210	THR	OG1	3.2	H bond
A	B	56	ARG	NH1	H	A 107	SER	OG	3.7	H bond
A	B	56	ARG	NH2	H	A 209	GLN	OE1	3.4	H bond
A	B	56	ARG	NH2	H	A 107	SER	OG	3.6	H bond
A	B	57	ILE	N	H	A 205	ASP	OD2	3.1	H bond
A	B	57	ILE	O	H	A 202	SER	OG	3.6	H bond
A	B	59	ARG	NH1	H	A 332	ASP	O	3.7	H bond
A	B	59	ARG	NH2	H	A 331	LEU	O	3.1	H bond
A	B	60	ASN	ND1	H	A 329	LEU	O	3.3	H bond
A	B	97	SER	O	H	A 275	TYR	OH	3.2	H bond
A	B	130	PRO	O	H	A 278	GLN	NE2	2.7	H bond
A	B	152	ASN	ND2	H	A 276	ASP	O	3.3	H bond
A	B	152	ASN	OD1	H	A 276	ASP	N	3.3	H bond
A	B	154	ASN	ND2	H	A 273	SER	O	3.3	H bond
A	B	55	ARG	NH1	H	A 213	GLU	OE1	3.8	Salt bridge
A	B	55	ARG	NH1	H	A 213	GLU	OE2	3.1	Salt bridge
A	B	55	ARG	NH2	H	A 213	GLU	OE1	3.4	Salt bridge
A	B	56	ARG	NE	H	A 205	ASP	OD1	3.1	Salt bridge
A	B	56	ARG	NH1	H	A 85	ASP	OD2	3.3	Salt bridge
A	B	56	ARG	NH2	H	A 205	ASP	OD1	3.1	Salt bridge
A	B	134	ARG	NE	H	A 276	ASP	OD2	3.9	Salt bridge
A	B	134	ARG	NH2	H	A 274	ASP	OD1	3.5	Salt bridge
A	B	134	ARG	NH2	H	A 274	ASP	OD2	3.2	Salt bridge
A	B	134	ARG	NH2	H	A 276	ASP	OD2	3.6	Salt bridge
A	B	147	ARG	NH1	H	A 276	ASP	OD2	3.6	Salt bridge
A	B	56	ARG	CZ	H	A 86	PHE	-	5.6	Cation- π
A	G	399	TYR	OH	D	A 300	ASN	N	3.5	H bond
A	G	403	GLN	NE2	D	A 297	PRO	O	3.4	H bond
A	G	427	ARG	NH2	D	A 628	PHE	O	2.5	H bond
A	G	593	ARG	NH2	D	A 517	GLU	O	4.7	H bond
A	G	595	ASN	N	D	A 292	TYR	O	2.9	H bond

Table 2.9 | Summary of a Goslar chimallin protomer's polar interactions *in vitro* (continued)

Chain-1	Resi	Resn	Atom	Chain-2	Resi	Resn	Atom	Dist (Å)	Type		
A	G	595	ASN	O	D	A	293	SER	OG	2.9	H bond
A	G	599	ARG	NE	D	A	293	SER	OG	3.6	H bond
A	G	599	ARG	NH2	D	A	294	PRO	O	3.3	H bond
A	G	599	ARG	NH2	D	A	293	SER	OG	3.7	H bond
A	G	603	LEU	N	D	A	410	LEU	O	2.7	H bond
A	G	603	LEU	O	D	A	410	LEU	N	3.6	H bond
A	G	604	GLY	O	D	A	351	ASN	ND2	3.5	H bond
A	G	605	ASN	O	D	A	566	GLN	NE2	2.8	H bond
A	G	610	ILE	O	D	A	579	ILE	N	2.9	H bond
A	G	611	TYR	OH	D	A	348	ALA	O	3.2	H bond
A	G	612	ALA	N	D	A	579	ILE	O	3.4	H bond
A	G	593	ARG	NE	D	A	521	ASP	OD1	3.0	Salt bridge
A	G	593	ARG	NH1	D	A	520	GLU	OE1	3.5	Salt bridge
A	G	593	ARG	NH2	D	A	521	ASP	OD1	3.6	Salt bridge
A	G	593	ARG	NH2	D	A	520	GLU	OE1	2.6	Salt bridge
A	G	601	ARG	NH1	D	A	559	ASP	OD1	3.4	Salt bridge
A	G	601	ARG	NH1	D	A	559	ASP	OD2	3.1	Salt bridge
A	G	427	ARG	CZ	D	A	631	PHE	-	4.3	Cation- π
A	G	593	ARG	CZ	D	A	292	TYR	-	5.2	Cation- π
A	G	599	ARG	CZ	D	A	407	TYR	-	5.8	Cation- π

Table 2.10 | Bacterial strains and plasmids used in this study

Bacterial strain	Source
<i>Pseudomonas chlororaphis</i> 200-B	⁸⁸
<i>Escherichia coli</i> APEC 2248	DSMZ (dsmz.de) #103255
<i>Escherichia coli</i> Rosetta2 pLysS	EMD Millipore

Table 2.10 | Bacterial strains and plasmids used in this study (continued)

Plasmid	Host strain	Source
pHERD30-T	<i>P. chlororaphis</i>	⁸⁹
pHERD30-T + GFPmut1	<i>P. chlororaphis</i>	⁹⁰
pHERD30-T + 201phi2-1 GFP-chimallin full-length	<i>P. chlororaphis</i>	this study
pHERD30-T + 201phi2-1 GFP-chimallin 48-631	<i>P. chlororaphis</i>	this study
pHERD30-T + 201phi2-1 GFP-chimallin 64-631	<i>P. chlororaphis</i>	this study
pHERD30-T + 201phi2-1 GFP-chimallin 1-611	<i>P. chlororaphis</i>	this study
pHERD30-T + 201phi2-1 GFP-chimallin 1-581	<i>P. chlororaphis</i>	this study
pHERD30-T + 201phi2-1 GFP-chimallin 48-611	<i>P. chlororaphis</i>	this study
pHERD30-T + 201phi2-1 GFP-chimallin 64-581	<i>P. chlororaphis</i>	this study
2-BT (His6-TEV site fusion expression vector)	<i>E. coli</i> Rosetta2 pLysS	Addgene (addgene.org) #29666
2-BT + 201phi2-1 chimallin FL	<i>E. coli</i> Rosetta2 pLysS	this study
2-BT + 201phi2-1 chimallin 48-631	<i>E. coli</i> Rosetta2 pLysS	this study
2-BT + 201phi2-1 chimallin 64-631	<i>E. coli</i> Rosetta2 pLysS	this study
2-BT + 201phi2-1 chimallin 1-612	<i>E. coli</i> Rosetta2 pLysS	this study
2-BT + 201phi2-1 chimallin 1-582	<i>E. coli</i> Rosetta2 pLysS	this study
2-BT + 201phi2-1 chimallin 48-612	<i>E. coli</i> Rosetta2 pLysS	this study
2-BT + 201phi2-1 chimallin 64-582	<i>E. coli</i> Rosetta2 pLysS	this study
2-BT + Goslar chimallin full-length	<i>E. coli</i> Rosetta2 pLysS	this study

References

1. Koonin, Eugene V.; Makarova, Kira S.; Wolf, Yuri I., Evolutionary Genomics of Defense Systems in Archaea and Bacteria. *Annu Rev Microbiol* **2017**, *71*, 233-261.
2. Chaikerasak, Vorrapon; Nguyen, Katrina; Khanna, Kanika; Brilot, Axel F.; Erb, Marcella L.; Coker, Joanna K. C.; Vavilina, Anastasia; Newton, Gerald L.; Buschauer, Robert; Pogliano, Kit; Villa, Elizabeth; Agard, David A.; Pogliano, Joe, Assembly of a nucleus-like structure during viral replication in bacteria. *Science* **2017**, *355* (6321), 194-197.
3. Mendoza, Senén D.; Nieweglowska, Eliza S.; Govindarajan, Sutharsan; Leon, Lina M.; Berry, Joel D.; Tiwari, Anika; Chaikerasak, Vorrapon; Pogliano, Joe; Agard, David A.; Bondy-Denomy, Joseph, A bacteriophage nucleus-like compartment shields DNA from CRISPR nucleases. *Nature* **2020**, *577*.
4. Malone, Lucia M.; Warring, Suzanne L.; Jackson, Simon A.; Warnecke, Carolin; Gardner, Paul P.; Gummy, Laura F.; Fineran, Peter C., A jumbo phage that forms a nucleus-like structure evades CRISPR–Cas DNA targeting but is vulnerable to type III RNA-based immunity. *Nature Microbiology* **2020**, *5*, 48-55.
5. Rostøl, Jakob T.; Marraffini, Luciano, (Ph)ighting Phages: How Bacteria Resist Their Parasites. *Cell Host & Microbe* **2019**, *25* (2), 184-194.
6. Hampton, Hannah G.; Watson, Bridget N. J.; Fineran, Peter C., The arms race between bacteria and their phage foes. *Nature* **2020**, *577*, 327-336.
7. Li, Yuping; Bondy-Denomy, Joseph, Anti-CRISPRs go viral: The infection biology of CRISPR-Cas inhibitors. *Cell Host & Microbe* **2021**, *29* (5), 704-714.
8. Chaikerasak, Vorrapon; Nguyen, Katrina; Egan, MacKennon E.; Erb, Marcella L.; Vavilina, Anastasia; Pogliano, Joe, The Phage Nucleus and Tubulin Spindle Are Conserved among Large Pseudomonas Phages. *Cell Reports* **2017**, *20* (7), 1563-1571.
9. Chaikerasak, Vorrapon; Khanna, Kanika; Nguyen, Katrina T.; Sugie, Joseph; Egan, MacKennon E.; Erb, Marcella L.; Vavilina, Anastasia; Nonejuie, Poochit; Nieweglowska, Eliza; Pogliano, Kit; Agard, David A.; Villa, Elizabeth; Pogliano, Joe, Viral Capsid Trafficking along Treadmilling Tubulin Filaments in Bacteria. *Cell* **2019**, *177* (7), 1771-1780.
10. Pohl, John, *Aztec, Mixtec and Zapotec Armies*. Osprey Publishing: 1991.
11. Yashiro, Yuka; Yamashita, Seisuke; Tomita, Kozo, Crystal Structure of the Enterohemorrhagic Escherichia coli AtaT-AtaR Toxin-Antitoxin Complex. *Structure* **2019**, *27* (3), 476-484.
12. Ilca, Serban L.; Kotecha, Abhay; Sun, Xiaoyu; Poranen, Minna M.; Stuart, David I.; Huiskonen, Juha T., Localized reconstruction of subunits from electron cryomicroscopy images of macromolecular complexes. *Nature Communications* **2015**, *6* (8843).
13. Ding, Ke; Celma, Cristina C.; Zhang, Xing; Chang, Thomas; Shen, Wesley; Atanasov, Ivo; Roy, Polly; Zhou, Z. Hong, In situ structures of rotavirus polymerase in action and mechanism of mRNA transcription and release. *Nature Communications* **2019**, *10* (2216).
14. Iyer, Lakshminarayan M.; Anantharaman, Vivek; Krishnan, Arunkumar; Burroughs, A. Maxwell; Aravind, L., Jumbo Phages: A Comparative Genomic Overview of Core Functions and Adaptions for Biological Conflicts. *Viruses* **2021**, *13* (1).
15. Birkholz, Erica A.; Laughlin, Thomas G.; Armbruster, Emily; Suslov, Sergey; Lee, Jina; Wittmann, Johannes; Corbett, Kevin D.; Villa, Elizabeth; Pogliano, Joe, A cytoskeletal vortex drives phage nucleus rotation during jumbo phage replication in E. coli. *Cell Reports* **2022**, *40* (7).
16. Heymann, J. Bernard; Wang, Bing; Newcomb, William W.; Wu, Weimin; Winkler, Dennis C.; Cheng, Naiqian; Reilly, Erin R.; Hsia, Ru-Ching; Thomas, Julie A.; Steven, Alasdair C., The Mottled Capsid of the Salmonella Giant Phage SPN3US, a Likely Maturation Intermediate with a Novel Internal Shell. *Viruses* **2020**, *12* (9).
17. Nieweglowska, Eliza S.; Brilot, Axel F.; Méndez-Moran, Melissa; Baek, Minkyung; Li, Junrui; Kokontis, Claire; Cheng, Yifan; Baker, David; Bondy-Denomy, Joe; Agard, David A., The ϕ PA3 Phage Nucleus is Enclosed by a Self-Assembling, 2D Crystalline Lattice. *bioRxiv* **2022**.
18. Danilova, Yana A.; Belousova, Viktoriia V.; Moiseenko, Andrey V.; Vishnyakov, Innokentii E.; Yakunina, Maria V.; Sokolova, Olga S., Maturation of Pseudo-Nucleus Compartment in P. aeruginosa, Infected with Giant phiKZ Phage. *Viruses* **2020**, *12* (10).

19. Reilly, Erin R.; Abajorga, Milky K.; Kiser, Cory; Redzuan, Nurul Humaira Mohd; Haidar, Zein; Adams, Lily E.; Diaz, Randy; Pinzon, Juliana A.; Hudson, André O.; Black, Lindsay W.; Hsia, Ru-Ching; Weintraub, Susan T.; Thomas, Julie A., A Cut above the Rest: Characterization of the Assembly of a Large Viral Icosahedral Capsid. *Viruses* **2020**, *12* (7).
20. Weintraub, Susan T.; Redzuan, Nurul Humaira Mohd; Barton, Melissa K.; Amin, Nur Amira Md; Desmond, Maxim I.; Adams, Lily E.; Ali, Bazla; Pardo, Sammy; Molleur, Dana; Wu, Weimin; Newcomb, William W.; Osier, Michael V.; Black, Lindsay W.; Steven, Alasdair C.; Thomas, Julie A., Global Proteomic Profiling of Salmonella Infection by a Giant Phage. *J Virol* **2019**, *93* (5).
21. Serwer, Philip; Hayes, Shirley J.; Zaman, Shehnaz; Lieman, Karen; Rolando, Mandy; Hardies, Stephen C., Improved isolation of undersampled bacteriophages: finding of distant terminase genes. *Virology* **2004**, *329* (2), 412-424.
22. Qiu, Dongru; Damron, F. Heath; Mima, Takehiko; Yu, Herbert P. Schweizer
Hongwei D., PBAD-Based Shuttle Vectors for Functional Analysis of Toxic and Highly Regulated Genes in Pseudomonas and Burkholderia spp. and Other Bacteria. *Appl Environ Microbiol* **2008**, *74* (23).
23. Nguyen, Katrina T.; Sugie, Joseph; Khanna, Kanika; Egan, MacKennon E.; Birkholz, Erica A.; Lee, Jina; Beierschmitt, Christopher; Villa, Elizabeth; Pogliano, Joe, Selective transport of fluorescent proteins into the phage nucleus. *PLoS One* **2021**, *16* (6).
24. Lam, Vinson; Villa, Elizabeth, Practical Approaches for Cryo-FIB Milling and Applications for Cellular Cryo-Electron Tomography. In *cryoEM*, Gonen, Tamir; Nannenga, Brent L., Eds. Springer: Humana, New York, 2021; Vol. 2215.
25. Mastronarde, David N., Automated electron microscope tomography using robust prediction of specimen movements. *J Struct Biol* **2005**, *152* (1), 36-51.
26. Hagen, Wim J. H.; Wan, William; Briggs, John A. G., Implementation of a cryo-electron tomography tilt-scheme optimized for high resolution subtomogram averaging. *J Struct Biol* **2017**, *197* (2), 191-198.
27. Tegunov, Dmitry; Cramer, Patrick, Real-time cryo-electron microscopy data preprocessing with Warp. *Nat Meth* **2019**, *16*, 1146-1152.
28. Mastronarde, David N.; Held, Susannah R., Automated tilt series alignment and tomographic reconstruction in IMOD. *J Struct Biol* **2017**, *197* (2), 102-113.
29. Chen, Shaoxia; McMullan, Greg; Faruqi, Abdul R.; Murshudov, Garib N.; Short, Judith M.; Scheres, Sjors H. W.; Henderson, Richard, High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. *Ultramicroscopy* **2013**, *135*, 24-35.
30. Tegunov, Dmitry; Xue, Liang; Dienemann, Christian; Cramer, Patrick; Mahamid, Julia, Multi-particle cryo-EM refinement with M visualizes ribosome-antibiotic complex at 3.5 Å in cells. *Nat Meth* **2021**, *18*, 186-193.
31. Bharat, Tanmay A. M.; Russo, Christopher J.; Löwe, Jan; Passmore, Lori A.; Scheres, Sjors H. W., Advances in Single-Particle Electron Cryomicroscopy Structure Determination applied to Sub-tomogram Averaging. *Structure* **2015**, *23* (9), 1743-1753.
32. H.W.Scheres, Sjors, RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol* **2012**, *180* (3), 519-530.
33. Castaño-Díez, Daniel; Kudryashev, Mikhail; Stahlberg, Henning, Dynamo Catalogue: Geometrical tools and data management for particle picking in subtomogram averaging of cryo-electron tomograms. *J Struct Biol* **2017**, *197* (2), 135-144.
34. Qu, Kun; Glass, Bärbel; Doležal, Michal; Schur, Florian K. M.; Murciano, Brice; Rein, Alan; Rumlová, Michaela; Ruml, Tomáš; Kräusslich, Hans-Georg; Briggs, John A. G., Structure and architecture of immature and mature murine leukemia virus capsids. *Proceedings of the National Academy of Science* **2018**, *115* (50), E11751-E11760.
35. Goddard, Thomas D.; Huang, Conrad C.; Ferrin, Thomas E., Visualizing density maps with UCSF Chimera. *J Struct Biol* **2007**, *157* (1), 281-287.
36. Kovtun, Oleksiy; Leneva, Natalya; Bykov, Yury S.; Ariotti, Nicholas; Teasdale, Rohan D.; Schaffer, Miroslava; Owen, Benjamin D. Engel
David. J.; Briggs, John A. G.; Collins, Brett M., Structure of the membrane-assembled retromer coat determined by cryo-electron tomography. *Nature* **2018**, *561*, 561-564.

37. Burt, Alister; Gaifas, Lorenzo; Dendooven, Tom; Gutsche, Irina, A flexible framework for multi-particle refinement in cryo-electron tomography. *PLoS Biol* **2021**, 19 (8).
38. Schindelin, Johannes; Arganda-Carreras, Ignacio; Frise, Erwin; Kaynig, Verena; Longair, Mark; Pietzsch, Tobias; Preibisch, Stephan; Rueden, Curtis; Saalfeld, Stephan; Schmid, Benjamin; Tinevez, Jean-Yves; White, Daniel James; Hartenstein, Volker; Eliceiri, Kevin; Tomancak, Pavel; Cardona, Albert, Fiji: an open-source platform for biological-image analysis. *Nat Meth* **2012**, 9, 676-682.
39. Scaramuzza, Stefano; Castaño-Díez, Daniel, Step-by-step guide to efficient subtomogram averaging of virus-like particles with Dynamo. *PLoS Biol* **2021**, 19 (8).
40. Martinez-Sanchez, Antonio; Garcia, Inmaculada; Asano, Shoh; Lucic, Vladan; Fernandez, Jose-Jesus, Robust membrane detection based on tensor voting for electron tomography. *J Struct Biol* **2014**, 186 (1), 49-61.
41. Castaño-Díez, Daniel; Kudryashev, Mikhail; Arheit, Marcel; Stahlberg, Henning, Dynamo: A flexible, user-friendly development tool for subtomogram averaging of cryo-EM data in high-performance computing environments. *J Struct Biol* **2012**, 178 (2), 139-151.
42. Salfer, Maria; Collado, Javier F.; Baumeister, Wolfgang; Fernández-Busnadiego, Rubén; Martínez-Sánchez, Antonio, Reliable estimation of membrane curvature for cryo-electron tomography. *PLoS Computational Biology* **2020**, 16 (8).
43. Kremer, James R.; Mastronarde, David N.; McIntosh, J. Richard, Computer Visualization of Three-Dimensional Image Data Using IMOD. *J Struct Biol* **1996**, 116 (1), 71-76.
44. Zivanov, Jasenko; Nakane, Takanori; Scheres, Sjors H. W., Estimation of high-order aberrations and anisotropic magnification from cryo-EM data sets in RELION-3.1. *International Union of Crystallography Journal* **2020**, 7 (2), 253-267.
45. Zivanov, Jasenko; Nakane, Takanori; Scheres, Sjors H. W., A Bayesian approach to beam-induced motion correction in cryo-EM single-particle analysis. *International Union of Crystallography Journal* **2019**, 6 (1), 5-17.
46. Punjani, Ali; Rubinstein, John L.; Fleet, David J.; Brubaker, Marcus A, cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Meth* **2017**, 14, 290-296.
47. Punjani, Ali; Zhang, Haowei; Fleet, David J., Non-uniform refinement: adaptive regularization improves single-particle cryo-EM reconstruction. *Nat Meth* **2020**, 17.
48. Tan, Yong Zi; Baldwin, Philip R; Davis, Joseph H; Williamson, James R; Potter, Clinton S; Carragher, Bridget; Lyumkis, Dmitry, Addressing preferred specimen orientation in single-particle cryo-EM through tilting. *Nat Meth* **2017**, 14, 793-796.
49. Pfab, Jonas; Phan, Nhut Minh; Si, Dong, DeepTracer for fast de novo cryo-EM protein structure modeling and special studies on CoV-related complexes. *Proceedings of the National Academy of Sciences* **2020**, 118 (2).
50. Emsley, Paul; Lohkamp, Bernhard; Scott, William G.; Cowtan, Kevin, Features and development of Coot. *Acta Crystallographica Section D* **2010**, 66 (4), 486-501.
51. Afonine, Pavel V.; Poon, Billy K.; Read, Randy J.; Sobolev, Oleg V.; Terwilliger, Thomas C.; Urzhumtsev, Alexandre; Adams, Paul D., Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallographica Section D* **2018**, D74, 531-544.
52. Williams, Christopher J.; Headd, Jeffrey J.; Moriarty, Nigel W.; Prisant, Michael G.; Videau, Lizbeth L.; Deis, Lindsay N.; Verma, Vishal; Keedy, Daniel A.; Hintze, Bradley J.; Chen, Vincent B.; Jain, Swati; Lewis, Steven M.; III, W. Bryan Arendall; Snoeyink, Jack; Adams, Paul D.; Lovell, Simon C.; Richardson, Jane S.; Richardson, David C., MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci* **2018**, 27 (1), 293-315.
53. Barad, Benjamin A; Echols, Nathaniel; Wang, Ray Yu-Ruei; Cheng, Yifan; DiMaio, Frank; Adams, Paul D; Fraser, James S, EMRinger: side chain-directed model and map validation for 3D cryo-electron microscopy. *Nat Meth* **2015**, 12, 943-946.
54. Krissinel, Evgeny; Henrick, Kim, Inference of Macromolecular Assemblies from Crystalline State. *J Mol Biol* **2007**, 372 (3), 774-797.
55. Gallivan, Justin P.; Dougherty, Dennis A., Cation- π interactions in structural biology. *Proceedings of the National Academy of Sciences* **1999**, 96 (17), 9459-9464.
56. Karami, Yasaman; Rey, Julien; Postic, Guillaume; Murail, Samuel; Tufféry, Pierre; Vries, Sjoerd J de, DaReUS-Loop: a web server to model multiple loops in homology models. *Nucleic Acids Res* **2019**, 47 (W1), W423-W428.

57. Jurrus, Elizabeth; Engel, Dave; Star, Keith; Monson, Kyle; Brandi, Juan; Felberg, Lisa E.; Brookes, David H.; Wilson, Leighton; Chen, Jiahui; Liles, Karina; Chun, Minju; Li, Peter; Gohara, David W.; Dolinsky, Todd; Konecny, Robert; Koes, David R.; Nielsen, Jens Erik; Head-Gordon, Teresa; Geng, Weihua; Krasny, Robert; Wei, Guo-Wei; Holst, Michael J.; McCammon, J. Andrew; Baker, Nathan A., Improvements to the APBS biomolecular solvation software suite. *Protein Sci* **2017**, *27*, 112-128.
58. Wang, Junmei; Cieplak, Piotr; Kollman, Peter A., How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry* **2000**, *21*, 1049-1074.
59. Olsson, Mats H. M.; Søndergaard, Chresten R.; Rostkowski, Michal; Jensen, Jan H., PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *Journal of Chemical Theory and Computation* **2011**, *7* (2), 525-537.
60. Dolinsky, Todd J.; Nielsen, Jens E.; McCammon, J. Andrew; Baker, Nathan A., PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *Nucleic Acids Res* **2004**, *32* (2), W665-W667.
61. Bahar, Ivet; Lezon, Timothy R.; Yang, Lee-Wei; Eyal, Eran, Global Dynamics of Proteins: Bridging Between Structure and Function. *Annual Review of Biophysics* **2010**, *39*, 23-42.
62. Lezon, Timothy R.; Shrivastava, Indira H.; Yang, Zheng; Bahar, Ivet, Elastic Network Models For Biomolecular Dynamics: Theory and Application to Membrane Proteins and Viruses. *Handbook on Biological Networks* **2009**, 129-158.
63. Leioatts, Nicholas; Romo, Tod D.; Grossfield, Alan, Elastic Network Models Are Robust to Variations in Formalism. *Journal of Chemical Theory and Computation* **2012**, *8* (7), 2424-2434.
64. Bauer, Jacob A.; Pavlović, Jelena; Bauerová-Hlinková, Vladena, Normal Mode Analysis as a Routine Part of a Structural Investigation. *Molecules* **2019**, *24* (18).
65. López-Blanco, José Ramón; Miyashita, Osamu; Tama, Florence; Chacón, Pablo, Normal Mode Analysis Techniques in Structural Biology. *eLS* **2014**.
66. Atilgan, Ali Rana; Durrell, Stewart R.; Jernigan, Robert L.; Demirel, Melik C.; Keskin, Özlem; Bahar, Ivet, Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophys J* **2001**, *80*, 505-515.
67. Hu, Guang, Identification of Allosteric Effects in Proteins by Elastic Network Models. In *Allostery: Methods and Protocols*, Paola, Luisa di; Giuliani, Alessandro, Eds. Humana: New York, NY, 2021; Vol. 2253, pp 21-35.
68. Bahar, Ivet; Cheng, Mary Hongying; Lee, Ji Young; Kaya, Cihan; Zhang, She, Structure-Encoded Global Motions and Their Role in Mediating Protein-Substrate Interactions. *Biophys J* **2015**, *109* (6), 1101-1109.
69. Erman, Burak, The Gaussian Network Model: Precise Predictions of Residue Fluctuations and Application to Binding Problems. *Biophys J* **2006**, *91*, 3589-3599.
70. Bakan, Ahmet; Meireles, Lidio M.; Bahar, Ivet, ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics* **2011**, *27* (11), 1575-1577.
71. Bakan, Ahmet; Dutta, Anindita; Mao, Wenzhi; Liu, Ying; Chennubhotla, Chakra; Lezon, Timothy R.; Bahar, Ivet, Evol and ProDy for bridging protein sequence evolution and structural dynamics. *Bioinformatics* **2014**, *30* (18), 2681-2683.
72. Tekpinar, Mustafa; Yildirim, Ahmet, Only a Subset of Normal Modes is Sufficient to Identify Linear Correlations in Proteins. *Journal of Chemical Information and Modeling* **2018**, *58* (9), 1947-1961.
73. Case, David A.; Aktulga, H. Metin; Belfon, Kellon; Ben-Shalom, Ido Y.; Brozell, Scott R.; Cerutti, David S.; III, Thomas E. Cheatham; Cisneros, G. Andrés; Cruzeiro, Vinícius Wilian D.; Darden, Tom A.; Duke, Robert E.; Giambasu, George; Gilson, Michael K.; Gohlke, Holger; Götz, Andreas W.; Harris, Robert; Izadi, Saeed; Izmailov, Sergei A.; Jin, Chi; Kasavajhala, Koushik; Kaymak, Mehmet C.; King, Edward; Kovalenko, Andriy; Kurtzman, Tom; Lee, Tai Sung; LeGrand, Scott; Li, Pengfei; Lin, Charles; Liu, Jian; Luchko, Tyler; Luo, Ray; Machado, Matias; Man, Viet; Manathunga, Madushanka; Merz, Kenneth M.; Miao, Yinglong; Mikhailovskii, Oleg; Monard, Gerald; Nguyen, Hai; O'Hearn, Kurt A.; Onufriev, Alexey; Pan, Feng; Pantano, Sergio; Qi, Ruxi; Rahnamoun, Ali; Roe, Daniel R.; Roitberg, Adrian; Sagui, Celeste; Schott-Verdugo, Stephan; Shen, Jana; Simmerling, Carlos L.; Skrynnikov, Nikolai R.; Smith, Jamie; Swails, Jason; Walker, Ross C.; Wang, Jinan; Wei, Haixin; Wolf, Romain M.; Wu, Xiongwu; Xue, Yi; York, Darrin M.; Zhao, Shiji; Kollman, Peter A. *AMBER 2021*, University of California, San Francisco, 2021.
74. Li, Pengfei; Merz, Kenneth M., Taking into Account the Ion-Induced Dipole Interaction in the Nonbonded Model of Ions. *Journal of Chemical Theory and Computation* **2014**, *10* (1), 289-297.
75. Li, Pengfei; Song, Lin Frank; Merz, Kenneth M., Systematic Parameterization of Monovalent Ions Employing the Nonbonded Model. *Journal of Chemical Theory and Computation* **2015**, *11* (4), 1645-1657.

76. Salomon-Ferrer, Romelia; Götz, Andreas W.; Poole, Duncan; LeGrand, Scott; Walker, Ross C., Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *Journal of Chemical Theory and Computation* **2013**, 9 (9), 3878-3888.
77. Götz, Andreas W.; Williamson, Mark J.; Xu, Dong; Poole, Duncan; Grand, Scott Le; Walker, Ross C., Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *Journal of Chemical Theory and Computation* **2012**, 8 (5), 1542-1555.
78. Grand, Scott Le; Götz, Andreas W.; Walker, Ross C., SPFP: Speed without compromise—A mixed precision model for GPU accelerated molecular dynamics simulations. *Computer Physics Communications* **2013**, 184 (2), 374-380.
79. Izadi, Saeed; Anandakrishnan, Ramu; Onufriev, Alexey V., Building Water Models: A Different Approach. *Journal of Physical Chemistry Letters* **2014**, 5 (21), 3863-3871.
80. Tian, Chuan; Kasavajhala, Koushik; Belfon, Kellon A. A.; Raguette, Lauren; Huang, He; Migués, Angela N.; Bickel, John; Wang, Yuzhang; Pincay, Jorge; Wu, Qin; Simmerling, Carlos, ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *Journal of Chemical Theory and Computation* **2020**, 16 (1), 528-552.
81. Ryckaert, Jean-Paul; Ciccotti, Giovanni; Berendsen, Herman J. C., Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *Journal of Computational Physics* **1977**, 23, 327-341.
82. Loncharich, Richard J.; Brooks, Bernard R.; Pastor, Richard W., Langevin dynamics of peptides: The frictional dependence of isomerization rates of N-acetylalanine-N'-methylamide. *Biopolymers* **1992**, 32, 523-535.
83. Pastor, Richard W.; Brooks, Bernard R.; Szabo, Attila, An analysis of the accuracy of Langevin and molecular dynamics algorithms. *Molecular Physics* **1988**, 65 (6), 1409-1419.
84. Berendsen, Herman J. C.; Postma, Johan P. M.; Gunsteren, Wilfred F. van; Nola, Alfredo Di; Haak, Jan R., Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics* **1984**, 81 (8), 3684-3690.
85. Darden, Tom; York, Darrin M.; Pedersen, Lee G., Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *Journal of Chemical Physics* **1993**, 98 (12), 10089-10092.
86. Roe, Daniel R.; Thomas E. Cheatham, III, PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of Chemical Theory and Computation* **2013**, 9 (7), 3084-3095.
87. McGibbon, Robert T.; Beauchamp, Kyle A.; Harrigan, Matthew P.; Klein, Christoph; Swails, Jason M.; Hernández, Carlos X.; Schwantes, Christian R.; Wang, Lee-Ping; Lane, Thomas J.; Pande, Vijay S., MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys J* **2015**, 109 (8), 1528-1532.
88. Klesse, Gianni; Rao, Shanlin; Sansom, Mark S.P.; Tucker, Stephen J., CHAP: A Versatile Tool for the Structural and Functional Annotation of Ion Channel Pores. *J Mol Biol* **2019**, 431 (17), 3353-3365.
89. Pettersen, Eric F.; Goddard, Thomas D.; Huang, Conrad C.; Meng, Elaine C.; Couch, Gregory S.; Croll, Tristan I.; Morris, John H.; Ferrin, Thomas E., UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* **2021**, 30 (1), 70-82.
90. Humphrey, William; Dalke, Andrew; Schulten, Klaus, VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **1996**, 14 (1), 33-38.

Part II
Brownian dynamics, molecular dynamics and Markov state
models

Chapter 3

Multiscale simulations examining glycan shield effects on drug binding to influenza neuraminidase

3.1 Abstract

Influenza neuraminidase is an important drug target. Glycans are present on neuraminidase, and are generally considered to inhibit antibody binding via their glycan shield. In this work we studied the effect of glycans on the binding kinetics of antiviral drugs to the influenza neuraminidase. We created all-atom *in silico* systems of influenza neuraminidase with experimentally-derived glycoprofiles consisting of four systems with different glycan conformations and one system without glycans. Using Brownian dynamics simulations, we observe a two- to eight-fold decrease in the rate of ligand binding to the primary binding site of neuraminidase due to the presence of glycans. These glycans are capable of covering much of the surface area of neuraminidase, and the ligand binding inhibition is derived from glycans sterically occluding the primary binding site on a neighboring monomer. Our work also indicates that drugs preferentially bind to the primary binding site (i.e. the active site) over the secondary binding site, and we propose a binding mechanism illustrating this. These results help illuminate the complex interplay between glycans and ligand binding on the influenza membrane protein neuraminidase.

3.2 Statement of Significance

The influenza glycoprotein neuraminidase is the target for three FDA-approved influenza drugs in the US. However, drug resistance and low drug effectiveness merits further drug development towards neuraminidase, which is hindered by our limited understanding of glycan effects on ligand binding. Generally, drug developers do not include glycans in their development pipelines. Here, we show that even though glycans can reduce drug binding towards neuraminidase, we recommend future drug development work to focus on strong binders with a long lifetime. Furthermore, we examine the binding competition between the primary and secondary binding

sites on neuraminidase, leading us to propose a new, to the best of our knowledge, multivalent binding mechanism.

3.3. Introduction

It has been long appreciated that glycans on influenza membrane proteins help shield the virus from the host immune system's antibodies ¹⁻⁷. Unrecognized glycosylation differences can also attenuate influenza vaccines ⁸. In one study, glycans were shown to reduce epitope accessibility and drug binding to receptor proteins ⁹. Glycans can clearly influence antibody binding due to their presence in the antibody binding site. However, it remains to be seen whether this glycan shielding and glycoprofile variability is also a concern for influenza drugs, recognizing that these drugs are smaller than human antibodies, and the fact that glycans present themselves near, but not directly inside, the catalytic sites. Currently there are three FDA-approved influenza neuraminidase (NA) antivirals in the US: Tamiflu (oseltamivir), Relenza (zanamivir) and Rapivab (peramivir), all of which have lingering questions over their efficacy, side effects, and drug resistance ¹⁰⁻¹¹. This necessitates the need for further drug development against influenza ¹².

Drug developers have many hurdles to clear when designing a new influenza drug: classical ADMET characteristics, clinical trials and governmental regulations, among others. What is not often considered is the viral glycosylation state. The glycosylation state is the assemblage of glycans, linkages of sugars found on the surface of about half of all proteins ¹³. Influenza contains N-linked glycosylation sites, defined by the Asn-X-Ser/Thr sequon, where X can be anything besides proline ¹⁴. This leads to the so-called glycan shield, where glycans on the protein surface are capable of accessing much of the protein's surface area, and potentially shielding it from outside interactions ¹⁵⁻²². Among their many biological functions, glycans play a crucial, but complex role in viral infection ²³. One salient example of glycan function in influenza is how they help the virus evade the immune system ¹⁻⁷. Furthermore, glycans are capable of affecting receptor binding in influenza ^{5, 7, 24-27}.

Traditionally, glycans have been difficult to study due to their flexibility and heterogeneity. Most of the glycan characterization studies are done through mass spectrometry, which can yield highly variable glycoprofile data, such as differences in the degree of post-translational modifications, sequon occupancy, and type of glycan, for different strains of influenza ²⁸⁻³³. Similarly, glycan occupancy levels are not consistent across studies, even

when using the same cell line and strain of influenza³⁴⁻³⁵ These discrepancies may arise from differences in system setup, sample preparation, cell culturing and/or analysis method, which increases the difficulty in determining the transferability of experimental glycan results. Though not well understood, the number and position of the glycosylation sites on influenza can change over time as a result of antigenic drift³⁶⁻³⁹. This increases the glycoprofile variability, effectuating irregular but significant changes in the glycan shield over time.

Considering the variability and immune evasion function of the glycan shield, it remains to be seen what effect this shield has on small-molecule antiviral drug binding to viral surface proteins. Previous work has shown that, depending on the viral strain and receptor mimetic used, removing viral glycans can improve binding to cell receptor mimetics⁴⁰⁻⁴². Other studies have shown that these viral glycans decrease binding of other cell receptor mimetics^{27, 43-46}. Regardless, antiviral drugs will be much smaller than a receptor mimetic, and it is not clear whether this size difference means antiviral drugs will still be affected by the viral glycans. An earlier study by Kasson and Pande, using 100 ns molecular dynamics (MD) simulations, showed reduced binding of α 2-3-sialyllactose trisaccharides to hemagglutinin due to glycans⁴³. A recent review concluded that the viral glycosylation state should be considered when designing small molecule antivirals⁴⁷.

Focusing on how small molecule antivirals are affected by the glycan shield, we combine results from distinct BD and MD simulations into an integrated multiscale simulation study. We have utilized BD to estimate the rates of binding of small molecules to the primary (i.e. active/catalytic) and secondary (i.e. hemadsorption) binding sites of influenza neuraminidase in glycosylated and unglycosylated states. We see that the glycan shield is capable of moderately inhibiting drug association to the primary binding site of NA on the order of two to eight times. Small molecule association is faster to the primary binding site than the secondary binding site. Ligand binding is independent between the primary and secondary sites – the presence of one site does not influence binding at the other site. Overall, this work provides insights into the impact of glycans on small-molecule binding to NA.

3.4 Methods

In this study, we use Brownian dynamics (BD), which has been previously used to simulate protein-small molecule association⁴⁸⁻⁵¹. Specifically, it has also been used to simulate the association of small molecules to

influenza neuraminidase ⁵²⁻⁵⁴. BD makes the implicit assumption that long-range electrostatics and stochastic collisions with solvent molecules are the driving forces behind protein-ligand binding ⁵⁵. Therefore, it is an efficient method to simplify binding to describe electrostatically-influenced diffusion. Using BD allows for a reduction in system complexity and a focus on specific modulations of ligand association.

To assess whether glycans affect small molecule binding to NA, we created an *in silico* NA model using the strain of influenza A virus, A/Viet Nam/ 1203/2004(H5N1) and tetrameric PDB 2HTY, with Uniprot ID Q6DPL2 ⁵⁶. Building on this structure, we generated five NA constructs: (i) one unglycosylated model; (ii) one glycosylated model with web server-derived glycan conformations; (iii) three glycosylated models, each with unique, biologically-relevant glycan conformations derived from all-atom MD simulations that were based on (ii) as the starting structure. Finally, we ran BD simulations using these models to examine binding characteristics of oseltamivir, zanamivir and sialic acid. To note, the BD input and results files are provided on Github (<https://github.com/cgseitz>).

Setup of the unglycosylated model

The unglycosylated model was built using an avian H5N1 strain and was used as a basis for the other models. We picked this strain of influenza because it contains a glycosylation site at N146, a member of the 150-loop that hangs over the primary binding site, as shown in **Figure 3.1**. This close proximity provides a good test of whether glycans were capable of interfering with ligand association to influenza neuraminidase. As the BD simulations used here keep bonds rigid, it was necessary to select ligand conformations that represented a bound state and protein conformations that represented an open state, to properly approximate the initial binding contact. Thus, we selected a crystallized apo head region of the strain mentioned above (PDB 2HTY) ⁵⁶. The stalk region has not been crystallized for any influenza NA and is unlikely to influence ligand association due to its large distance from the distal binding sites, so it was not modelled. The crystallized calcium ions were retained throughout, while the crystallized glycan fragments were removed ⁵⁷.

2HTY was crystallized with a Y171H mutation (PDB numbering), which was reversed for this project through PyMOL ⁵⁸. The histidine rotamer was chosen to be the one with the highest occurrence in proteins, according to PyMOL. The crystal structure contained a broken backbone between P169 and N170 which was fixed through Schrodinger Maestro; subsequently residues 168-171 (on each side of the fixed bond) were minimized through

Maestro ⁵⁹. The same procedure was done for the broken backbone between V411 and Q412: the bond was created and residues 410-413 were minimized. This refitting was done for each monomer in the tetramer. The pH was set to pH 6.4, as this was done in the reference k_{on} experiments ⁶⁰⁻⁶¹. Using this pH, protonation states on the neuraminidase were assigned using PROPKA ⁶². The protonation assignments were done through the PDB2PQR server ⁶³. Partial charges on the protein were assigned according to the AMBER99 force field ⁶⁴. Parameterizing the glycans needed special treatment as there are not glycan parameters in the AMBER99 force field. We used the GLYCAM_06h-1 parameters as these would be consistent with the AMBER99 force field ⁶⁵.

Setup of the glycosylated NA model with web server-derived glycan conformations

To build the first glycosylated construct (with web server-derived glycan conformations), the unglycosylated NA structure was uploaded to the Glyprot server, and three representative glycans were added to each NA monomer, for a total of 12 glycans on the NA homotetramer ⁶⁶. Though there is experimental variability in glycosylation site occupancy, we decided to place a glycan in each glycosylation site to see the maximum potential effect the glycoprofile can have on ligand association. Considering most of the human H5N1 transmission came directly from avian sources, the glycans used to model this structure came from an avian (hen egg) source for growing these glycans ³⁰. Additionally, this dataset is the only one containing structures experimentally found on influenza NA ³⁰. We chose representative glycans from this dataset, however we note that both larger and smaller glycans will exist in nature; these size differences may slightly affect the results presented here. The exact glycans were selected as shown in **Table 3.1**.

Table 3.1 | Glycan structures from the Glyprot web server. The “Glycan structure” entries came from experimental results ³⁰. These structures consist of N-acetylglucosamine (GlcNAc), mannose (Man), N-acetylhexosamine (HexNAc) and hexose (Hex). HexNAc and Hex were interpreted according to their corresponding “Glyprot identifier” and the structures shown in **Figure 3.1**.

#	Glycosylation site	Glycan structure	Glyprot identifier
1	N88	(GlcNAc) ₂ (Man) ₃ (Hex) ₃	gly_8792.pdb
2	N146	(GlcNAc) ₂ (Man) ₃ (HexNAc) ₂ (Hex) ₂	gly_9196.pdb
3	N234	(GlcNAc) ₂ (Man) ₃ (HexNAc) ₁ (Hex) ₂	gly_8582.pdb

To better diversify our system, three glycosylation sites (termed as site #1, site #2 and site #3) present on each monomer were linked to three different glycan types. Importantly, the four monomers (termed as monomer A,

monomer B, monomer C, and monomer D) of our homotetrameric NA model were symmetrically glycosylated, meaning that sites #1, #2 and #3 were populated with the same glycan across monomers.

This resulted in glycans A1-3, B1-3, C1-3, and D1-3, where all glycans linked at site #1 have identical structures (but not necessarily identical conformations), all glycans linked at site #2 have identical structures, and all glycans linked at site #3 glycans have identical structures. The glycans selected are shown in **Figure 3.1**.

Figure 3.1

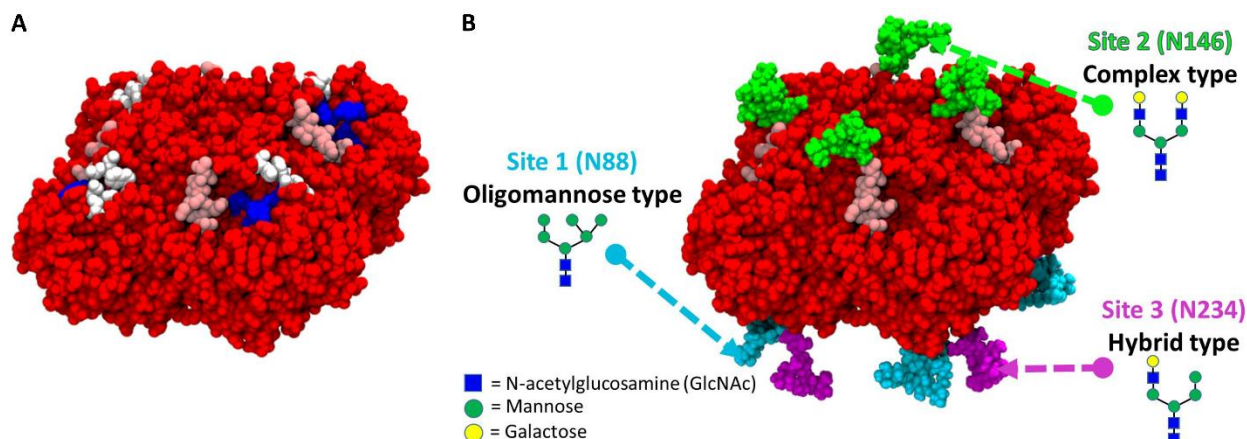


Figure 3.1 | Structure of the NA head. (A) The relation of the binding sites to the 150 loop and each other are shown. The NA protein is in red, the 150-loop is in pink, the primary binding site is in blue, and the secondary binding site in white. (B) The glycan structures and types used in this study are shown in relation to the 150 loop. The glycan on top of each monomer is attached to the 150 loop. Three unique glycan structures and types were used in the simulations. The font color of the glycosylation site refers to the corresponding glycan structure on the protein.

Molecular dynamics simulations

Starting with the structure containing web server-derived glycan conformations, MD was then used to generate representative glycan conformations, with the assumption that MD would provide realistic conformations of glycans within a microsecond's worth of sampling⁶⁷⁻⁶⁸. The first step was porting the structure with web server-derived glycans into CHARMM-GUI to prepare the structure for MD⁶⁹⁻⁷⁴. The disulfide bonds were taken from Uniprot ID Q6DPL2. The system was embedded into a box described with explicit water molecules using the TIP3P model⁷⁵. An ion model was used as described previously⁷⁶. The full system had a size of 299,732 atoms. An ionic solution of 0.15 M NaCl was used, and the CHARMM36 all-atom additive force fields were used for

the protein and the glycans ⁷⁷. Molecular dynamics simulations were run using GPU-accelerated AMBER18 with an NPT ensemble ⁷⁸⁻⁷⁹. The system was initially minimized for a total of 5000 cycles using a combination of steepest descent and conjugant gradient methods ⁷⁸⁻⁷⁹. Equilibration in an NPT ensemble was performed for 125 ps, using a timestep of 2 fs and the SHAKE algorithm to constrain all bonds involving hydrogen ⁸⁰. The equilibration temperature was set at 298 K and regulated through a Langevin dynamics thermostat ⁸¹⁻⁸². The pressure was fixed at 1 bar through a Monte Carlo barostat ⁸³. These simulations were run using Extreme Science and Engineering Discovery Environment (XSEDE), specifically the Comet supercomputer housed at the San Diego Supercomputer Center ⁸⁴. Periodic boundary conditions were used with a non-bonded short-range interaction cutoff of 12 Å and force-based switching at 10 Å. Particle mesh Ewald was used for the long-range electrostatic interactions ⁸⁵. For the production runs, the temperature was set at 310.15 K ⁶¹. After equilibration, this system was cloned into 50 identical replicates. Each one was run in parallel for 20 ns each with a unique starting velocity, totaling 1 µs of sampling.

Glycan clustering and setup of the glycosylated NA models with representative glycan conformations

Once the MD simulations finished, the trajectory of each glycan was concatenated independently of the rest of the system. Each of these individual glycan trajectories were then clustered using GROMACS-based GROMOS clustering with an RMSD cutoff of 2.5 Å ⁸⁶. This number was chosen so the three most populated clusters would represent at least 50% of the total glycan conformations in each of the simulations. The central structure, defined as the structure with the smallest average RMSD from all other members of the cluster, from the top cluster of each of the 12 glycans was then selected; the pyranose ring from the reducing end of the glycan was then aligned to the analogous pyranose ring of the corresponding glycan from the Glyprot-glycosylated structure as this should be the most stable part of the glycan ⁸⁷. The Glyprot glycans were removed and the glycans from the MD simulations were attached through Schrodinger Maestro, to create a new NA system with each glycosylation site inhabited by the central structure of the most representative conformation from the MD simulations. This was then repeated for the second and third most representative glycan clusters from the MD simulations.

Ligand setup

The sialic acid structures used were drawn from PDB 1MWE, which crystallized the boat conformation in the active site and the chair conformation in the secondary site ⁸⁸. The chair conformation of sialic acid was

crystallized with a missing carboxylate group, which was added through Schrodinger Maestro to model an energetically-favorable gauche conformation. Zanamivir was extracted from the 3CKZ crystal structure ⁶⁰. Oseltamivir was extracted from the 3CL0 crystal structure ⁶⁰. A 2D comparison of these ligands can be seen in **Figure 3.8**, showing their structural similarities; we note that all mentions in this study of oseltamivir pertain to Tamiflu's active metabolite oseltamivir carboxylate. These ligands were then uploaded to the PRODRG server to add hydrogens ⁸⁹. Charges according to the AMBER99 force field were added through the PDB2PQR server ⁶³⁻⁶⁴.

Brownian dynamics (binding pairs and simulations)

BD simulations were run using Browndye ⁹⁰. Even though the MD was run with the AMBER18 force field and the BD was run with the AMBER99 force field, we assume these to be sufficiently independent steps and the slight force field differences should not appreciably affect the results, especially as our k_{on} numbers are relative, not absolute. The charges for the protein and ligands were reassigned according to the AMBER99 force field ⁶⁴. The temperature was set to 310.15 K, which was the temperature for the referenced k_{on} experiments ⁶⁰⁻⁶¹. The ions used are shown in **Table 3.2**. These ions were selected to mimic the ion and buffer concentration of the reference k_{on} experiments ⁶⁰⁻⁶¹.

The experimental assay used 5 mM $CaCl_2$ and 32.5 mM MES buffer ⁶⁰. The Ca^{2+} and Cl^- concentrations were simply calculated by finding their ionic strengths. MES buffer is prepared with Na^+ ; the concentrations of the buffer and Na^+ at pH 6.4 were calculated with the Henderson-Hasselbalch equation ⁹¹⁻⁹². This resulted in an overall ionic strength of 0.039 M. The calcium, chlorine and sodium van der Waals radii were taken from the literature ^{64, 93}. The MES radius was determined by building it in Schrodinger Maestro and measuring it in VMD ⁹⁴.

APBS was used to create the electrostatic grids needed by Browndye for these simulations ⁹⁵. The grid spacings are listed in **Table 3.3**.

The solvent dielectric was set to 78 while the protein dielectric was set to 4. Desolvation forces were turned off. The Debye length, determined from the concentration and charges of the ions in the solution, was set to 15.7 Å. In Browndye, the b radius is defined as the starting radius for the ligand trajectories, at a distance where the force between the protein and ligand is independent of orientation. This distance is determined from the hydrodynamic

center of the receptor. Because of the different glycan conformations used, the b radius differed slightly between systems. If a ligand reaches what is known as the q radius, the trajectory either ends as a non-association or is restarted from the b radius according to Browndye's algorithm. The q radius is defined as 1.1 times the b radius distance. The b radius ranged between 109 Å and 112 Å depending on the system, and the q radius ranged from 120 Å to 123 Å. The exact b and q radius values for each system are shown in **Table 3.4**.

BD simulations were run on all five NA models generated (i.e. unglycosylated, glycosylated with web server-derived glycans, and the three systems with MD-derived glycan conformations). These simulations totaled 10 million trajectories for each ligand/binding site pair, consistently giving reproducible rates within the small level of error reported and resulting in 600 million trajectories total. Reproducible rates will be obtained by having a binding probability of around one in a million trajectories; we found we could roughly obtain these probabilities by using 10 million trajectories for each ligand/binding site as has been reported previously⁵³. This number of trajectories produced error values comparable to those seen in the reference experimental studies, as seen in **Figure 3.9**. For systems where we saw at least one binding event, the number of binding events ranged from two to 889 (see Supporting Material for details).

Brownian dynamics reaction criteria

BD simulations using Browndye requires the creation of reaction criteria, consisting of a list of protein-ligand atom pairs and a cutoff distance. If any three of these pairs simultaneously came closer than the cutoff distance, we assume the ligand will associate. The cutoff distance was empirically determined to be 3.228 Å; this distance approximately yielded the experimental k_{on} rates for both oseltamivir and zanamivir⁶⁰. There are no other experimental k_{on} rates towards the primary site of H5N1, and no referenced rates at all for the secondary site. The referenced k_{on} experiments were done with glycans attached to NA and measured to the full tetramer; this was confirmed in personal correspondence with the corresponding author (Stephen Martin of the MRC National Institute for Medical Research, correspondence on July 21, 2018). Considering that the reaction criteria and reaction distance were created for oseltamivir and no significant changes were made before applying them to zanamivir, we can safely assume that they are generalizable for sialic acid, an analog of both oseltamivir and zanamivir (**Figure 3.8**).

The protein-ligand atom pairs were taken from crystal structures of ligands in the primary and secondary sites of neuraminidase for each monomer, and simulations were run for the full tetramer. The primary binding site was determined according to the crystallized binding pocket for our strain of neuraminidase ⁵⁶. This pocket is noted to have a surface area of 941.3 Å² and a volume of 574.8 Å³ ⁹⁶. The secondary site contacts were determined from a structure of influenza A/tern/Australia/G70C/75 ⁸⁸. However, all the secondary site residues are conserved between that strain and the strain used in our simulations. The combined site simulations are defined as simulations with criteria allowing for association to either the primary or secondary site; it is simply a simulation run with a concatenation of the binding criteria for these sites.

In this work, we define binding site contacts to be those protein-ligand contacts seen in crystal structures. From these contacts, we created protein-ligand atom pairs in Browndye to determine when a reaction has occurred in our BD trajectories. There are seven primary binding site contacts reported between oseltamivir and the 3CL0 crystal structure ⁶⁰. These binding site contacts are reported in **Table 3.5** and **Figure 3.10**. There are five primary binding site contacts reported between sialic acid and the 1MWE crystal structure; all five of these are analogous to those seen for oseltamivir ⁸⁸. The binding site contacts for sialic acid are registered in **Table 3.6** and **Figure 3.11**. There is one primary binding site contact reported between zanamivir and the 3CKZ crystal structure; this one is analogous to one seen in oseltamivir ⁶⁰. The binding site contacts from oseltamivir were transferred to zanamivir retaining the one contact seen in the 3CKZ crystal structure and are reported in **Table 3.7** and **Figure 3.12**. Using the structural similarities of sialic acid and zanamivir to oseltamivir, analogous primary binding site atom pairs were created so that each ligand had seven primary protein-ligand atom pairs.

There are five secondary binding site contacts reported between sialic acid and the 1MWE crystal structure ⁸⁸. These contacts are reported in **Table 3.8**. There are no published reports of crystal structures of oseltamivir or zanamivir in the secondary binding site, so five analogous secondary binding site protein-ligand atom pairs were created for oseltamivir (**Table 3.9**) and zanamivir (**Table 3.10**) to match those seen in sialic acid, so that each ligand had five secondary binding site protein-ligand atom pairs.

3.5 Results

Glycan shield and individual glycan clusters

To pare down the data from 1 μ s of cumulative MD sampling and pick out biologically-relevant glycan conformations, we clustered each glycan from the MD simulations. The glycan trajectories were extracted and affixed on the static NA crystal structure, to reveal the conformational space they can access (**Figure 3.2**). Visualizing these glycan trajectories on the NA structure gives a qualitative representation of how much volume and surface area the glycans are capable of accessing. Keeping in mind the primary and secondary binding sites are located just beneath the glycans (**Figure 3.1**), the size and flexibility of the glycans here shows that they have the capability to “shield” the binding sites from ligand association.

Figure 3.2

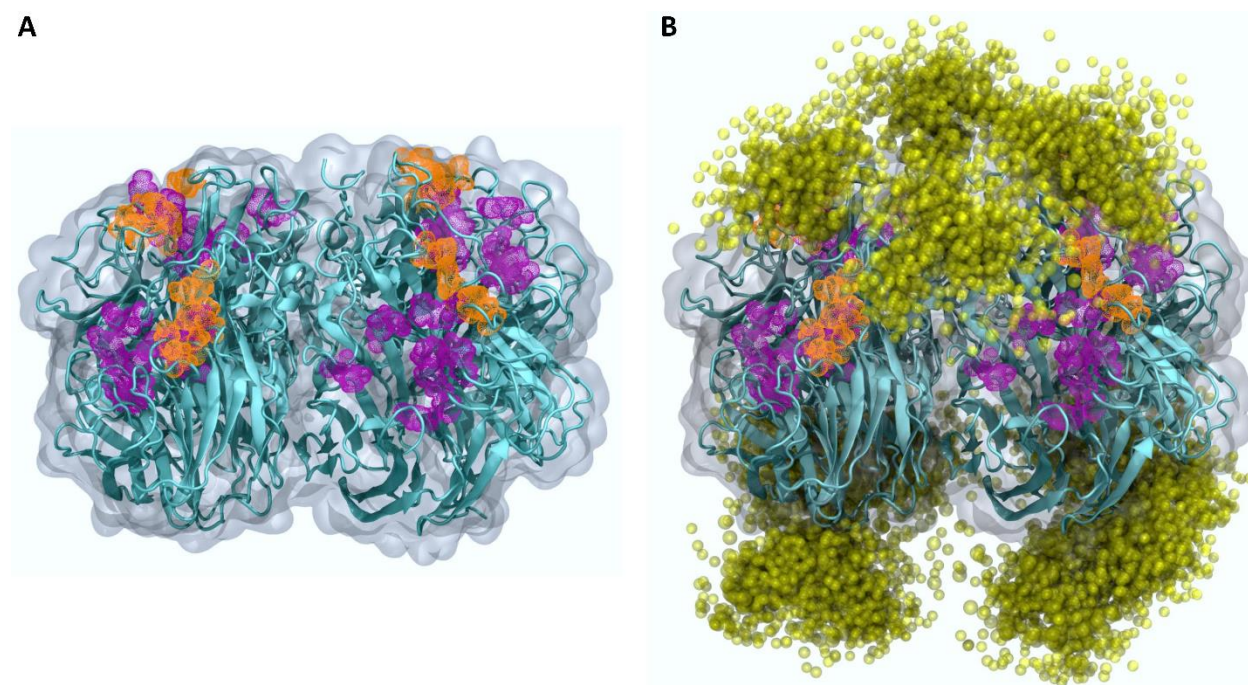


Figure 3.2 | Glycan shielding on NA. (A) Molecular representation of the unglycosylated NA's homotetrameric head showing accessible binding sites. The static structure of NA is in teal. The primary binding sites are in purple, and the secondary binding sites are in orange. (B) Molecular representation of the glycosylated NA's homotetrameric head showing much of its surface area shielded. The static structure of NA is in teal, the primary binding sites are in purple, and the secondary binding sites are in orange. The glycan shielding density is represented with a yellow cloud of spheres using multiple layered glycan conformations from the MD simulations. Each sphere approximates a monosaccharide. A total of 10 frames, using a stride of 100, were selected from the 1000 frames-long trajectory obtained upon concatenation of all the MD simulations.

The three most representative clusters for each glycan were extracted from the MD simulations. The central structure from each cluster was compared with the conformation generated from Glyprot. These clusters show some conformational diversity, but none show a particularly similar conformation to the Glyprot structure. However, the third glycan in each monomer shows a markedly decreased conformational diversity compared to the other two monomers. The clustering results from each monomer show the same trends; the results from monomer A are shown in **Figure 3.3**, while the results from monomer B (**Figure 3.13**), monomer C (**Figure 3.14**), and monomer D (**Figure 3.15**) are shown in the Supporting Material.

Figure 3.3

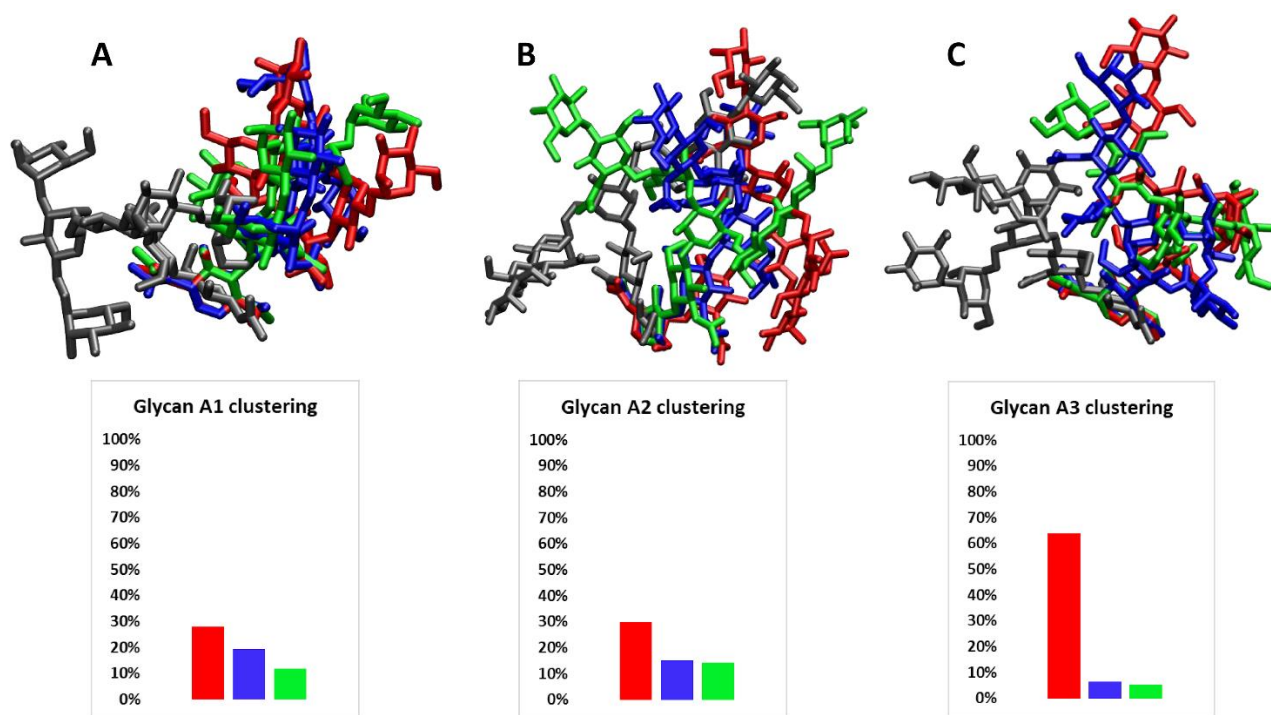


Figure 3.3 | Clustered glycans on monomer A. The glyprot structure is in gray, while the other colors represent clusters from the MD simulations.

The glycans on the top of the NA head are all situated directly over the primary binding site on their own monomer. Conformations extracted from the MD simulations universally bent towards the secondary binding site located on the adjacent monomer, leaving the primary binding site closest to them accessible (**Figure3. 4**).

Figure 3.4

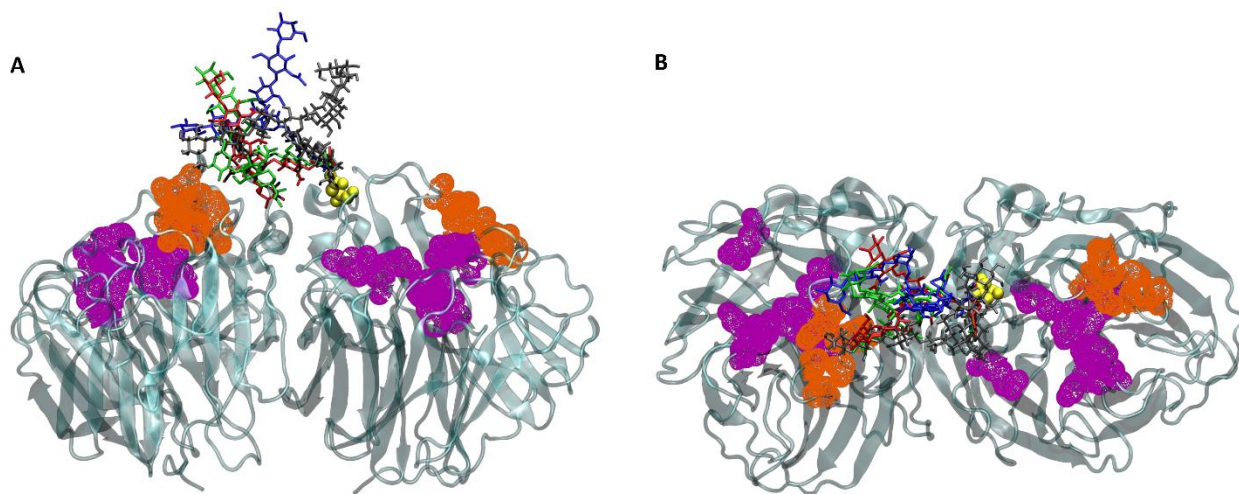


Figure 3.4 | The superposition of each glycan onto the static NA structure shows the conformational variation of the glycans. (A) Side view and (B) top view of the glycosylated NA model with MD-derived glycan conformations. For ease of viewing, only two monomers are shown here, with the glycans only coming from the monomer on the right in panel (A) and panel (B). The yellow glycosylation site N88 and the attached glycans reside in the same monomer. The glycans bend away from the binding sites on their monomer towards the binding sites on the neighboring monomer. This is seen for each monomer. The primary binding sites are in purple and the secondary binding sites are in orange. The linkage between the glycans and the protein is in yellow. The NA structure is in teal. The Glyprot conformation is in gray, the first conformation from the MD simulations is in orange, the second conformation is in blue, and the third conformation is in green.

Association rates of oseltamivir, zanamivir and sialic acid

To be confident in our computed association rates, we first needed to benchmark our system against experimental results. We created empirically-derived system criteria for the association of oseltamivir to the primary binding site of glycosylated NA, as described in the methods. After matching the experimental association rate with oseltamivir, the same parameters were applied to zanamivir. These are the only two experimental association rates for H5N1 NA.

Subsequently, we investigated the association of oseltamivir and zanamivir to the primary sites of glycosylated NA, obtaining association rates of 2.52 ± 0.21 / $\mu\text{M}\cdot\text{s}$ for oseltamivir and 0.47 ± 0.09 / $\mu\text{M}\cdot\text{s}$ for zanamivir. These are in agreement with the experimentally-measured rates of 2.52 ± 0.21 / $\mu\text{M}\cdot\text{s}$ and 0.95 ± 0.21 / $\mu\text{M}\cdot\text{s}$, respectively, as visualized in **Figure 3.9**⁶⁰. Considering the experimental systems were glycosylated, we had to pick one glycan conformation to use for computing these benchmarks in our glycosylated system; for reproducibility we chose the conformation generated from the Glyprot server. We note that choosing a different

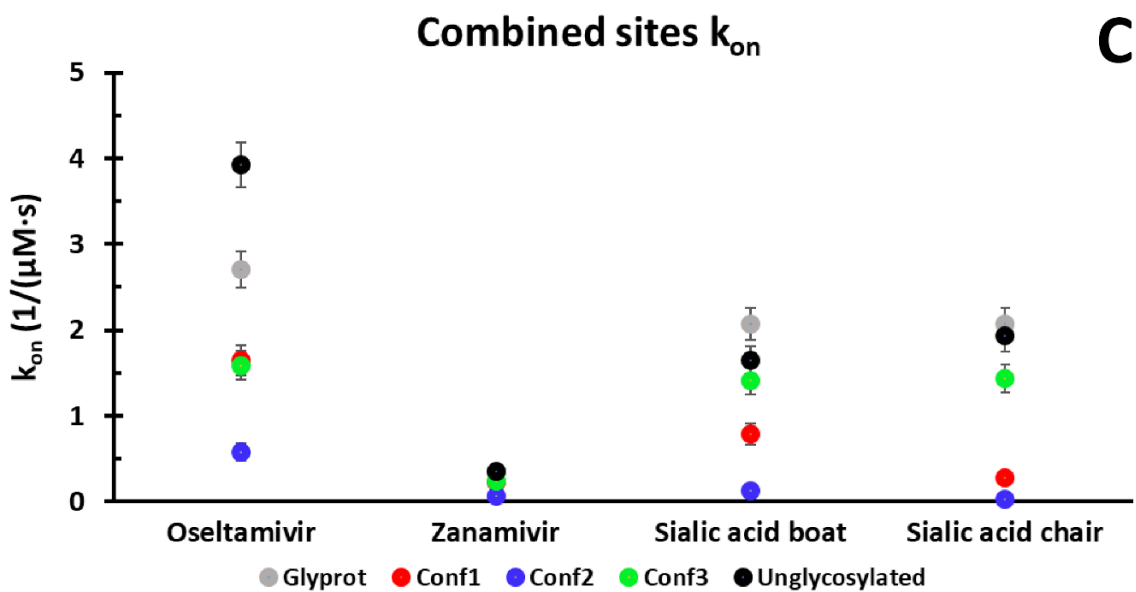
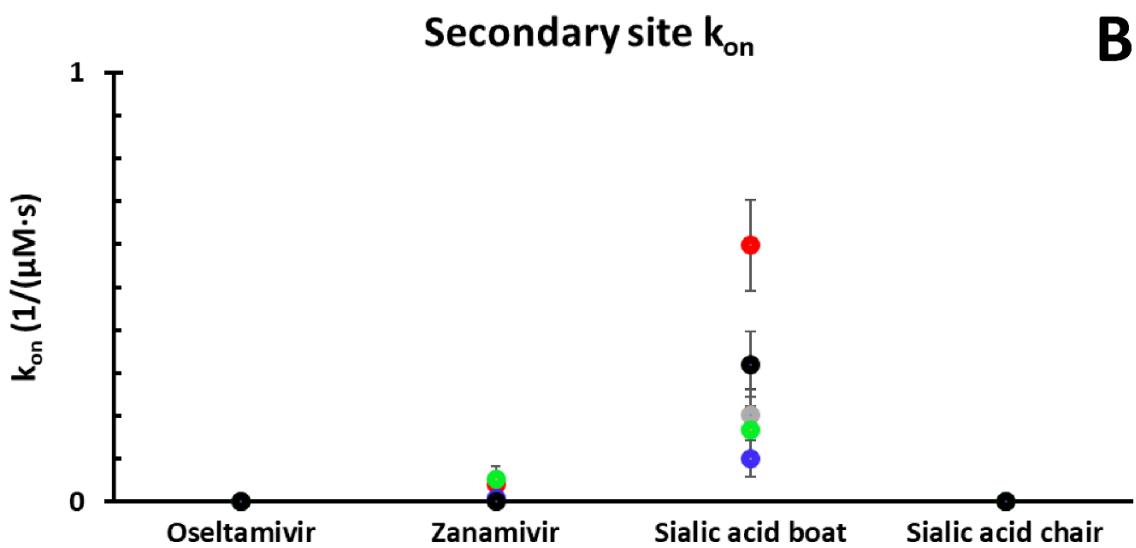
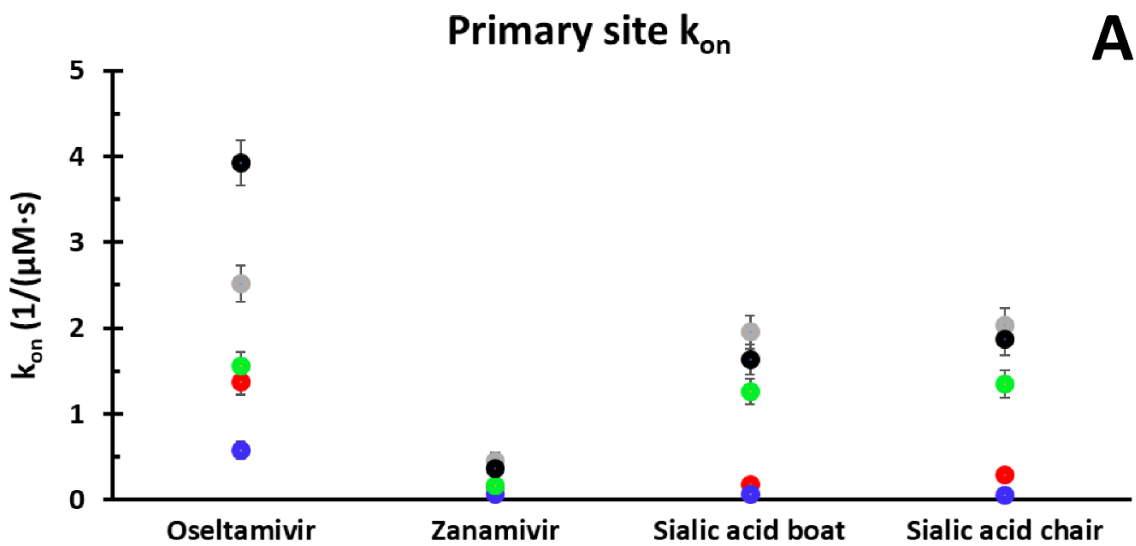
conformation for our computed benchmark would change the absolute association rates by a scaling factor, but the trends would remain the same.

Since the predicted k_{on} for oseltamivir and zanamivir both matched up well with the experimental rates, the system proved to be transferable to ligand analogs for the primary site. We then applied the same criteria to two different conformations of sialic acid, boat and chair, to probe if the association rate was dependent on conformation. This was done in addition to analyzing how association rate was modulated by different functional groups, via comparisons of ligand analogs such as oseltamivir, zanamivir and sialic acid.

With the binding criteria set up, we calculated the association rates of each of the ligands to the primary site (**Figure 3.5A**). These results show two important findings. First, there is not a large difference in association rates between the system with Glyprot glycans and the unglycosylated system. This shows that a glycan may adopt a conformation where it does not inhibit ligand binding much at all. The second finding is that the glycans from the MD simulations all show a moderate level of inhibition, more than the system with Glyprot glycans. This shows that biologically-relevant glycan conformations will likely exhibit a moderate level of inhibition towards ligand binding. Combining the first and second finding discussed in this paragraph, glycans are capable of perturbing ligand binding to NA.

Figure 3.5

Figure 3.5 | Association rates of each ligand to the primary and secondary sites. Conf1 is the glycan structure from the most populated cluster from the MD simulations. Conf2 is from the second most populated cluster, and conf3 is from the third most populated cluster. The association rates using glycans structures downloaded from Glyprot are shown in gray. The association rates using structures derived from the MD simulations are in bright, colorful shades whereas the others are in grayscale. The association rates without using any glycans are shown in black. (A) The glycan structures from the MD simulations show a moderate association rate inhibition to the primary binding site irrespective of ligand chosen. (B) Little association is seen to the secondary binding site. Note the different y-axis used to be able to see the small amount of binding. (C) Association rates of trajectories run with either the primary site or secondary site as the trajectory end point. Similar to (A), the glycans structures from the MD simulations in (C) show a moderate inhibition of ligand association. The raw data for this figure is seen in Table S10 (oseltamivir), Table S11 (zanamivir), Table S12 (sialic acid boat conformation), and Table S13 (sialic acid chair conformation).



There are no experimental association rates for ligands to the secondary site, so criteria were chosen based off of crystal structure data and discussed in the methods. Only sialic acid has been crystallized in the secondary site of avian NA, so binding site criteria for the secondary site were extracted from that structure and used to create the criteria for oseltamivir and zanamivir, as discussed in the methods ⁸⁸. A previous BD study suggested that oseltamivir can bind to the avian NA secondary site ⁵². A follow-up NMR study also suggested that the oseltamivir binds to the avian NA secondary site ⁹⁷. However, a more recent experimental study disagreed with these findings and did not see oseltamivir binding to the avian NA secondary site ⁹⁸. Considering the disagreement with oseltamivir binding to the secondary site, we decided to test this and secondary site binding for zanamivir as well. The computed association rates towards the secondary site show a markedly different story than those to the primary site (**Figure 3.5B**). None of the ligands exhibited noticeable binding towards the secondary site, with the exception of the boat conformation of sialic acid. Even with this conformation, there is no consistent trend when compared to primary site binding. Although the boat conformation sialic acid displays a small amount of binding, the chair conformation does not show binding. These results show that we can differentiate between these two sialic acid conformations at the BD level of theory.

Finally, trajectories were run where the ligand could associate to either the primary site or the secondary site (**Figure 3.5C**). Intriguingly, the results are essentially a concatenation of the rates seen for the primary and secondary sites individually. Considering the low level of secondary site binding, the trends here are the same as seen for the primary site.

As can be seen in **Figure 3.8**, there is a formal charge difference between the ligands: Sialic acid contains a formal charge of -1 while oseltamivir and zanamivir are neutral. Running test BD trajectories without charge treatment (results not shown), we saw analogous results to those seen in **Figure 3.5**. This meant that only the sterics of the systems affected binding, not electrostatics. Clearly one or a few of the structural differences between the ligands play outsized roles in affecting the association rates. In this work we did not further probe which exact atoms in the ligands will change the association rates.

3.6 Discussion

The glycan conformational flexibility underpinning the glycan shield

Biologically, the influenza replication cycle is propagated through NA recognizing and cleaving sialic acid. This study compares the interplay between that molecular recognition process and NA's aforementioned glycan shielding capabilities. This interplay is simplified here by approximating ligand binding as a diffusion-governed association process, modulated by protein electrostatics.

Previous studies have shown that viral proteins can exhibit a degree of glycosylation large enough to partially protect a variety of viruses from immune system antibodies; this is termed the viral glycan shield^{18, 21, 99-101}. From static structures one can envision the shielding that glycans can provide, but a dynamic representation better depicts the steric barrier encountered by immune system antibodies and drugs¹⁰². In our single NA protein, we see that glycans are capable of covering most of the NA surface area, as shown in **Figure 3.2**. This is consistent with studies explaining how the influenza glycan shield can cloak the influenza virion from the immune system⁵⁻⁷. The glycans can access a large volume, allowing for a considerable shielding potential. However, it is worthwhile to note that influenza glycoproteins are usually not as extensively glycosylated as on some other viral proteins, such as the HIV envelope protein or the SARS-CoV-2 S protein^{15, 17, 103-105}. The exact H5N1 construct prepared here contains a glycosylation site at N146. This is part of the 150 loop that borders the primary binding site (**Figure 3.1**). The representation in **Figure 3.2** shows that the glycans present at site N146 on each monomer have the combined capability to cover both NA binding sites, potentially thwarting the binding of small molecules. The results shown here display a moderate inhibitory effect due to glycans, but this effect would likely not be present in proteins whose glycans only reside far from the ligand binding sites, i.e. if the setup in **Figure 3.1** only contained the glycans at site 1 and site 3 on the bottom of the NA head. When examining the effect of glycan conformation on binding inhibition, the Glyprot glycans display a fairly vertical conformation. On the other hand, the glycans from the MD simulations bend backwards, away from the primary binding site on their own monomer and towards the secondary binding site of the adjacent monomer, as shown in **Figure 3.4**. Interestingly, this bend appears to be enough to inhibit primary site binding.

It has been previously shown that specific chemical modifications on the glycans can significantly change their flexibility¹⁰⁶⁻¹⁰⁸. It has also been hypothesized that glycan flexibility plays a role in protein-receptor binding

equilibria⁸⁷. Considering the scale of biological interactions that glycans participate in, it is likely that they would exploit their flexibility to facilitate these interactions. However, the glycan environment, and nearby steric clashes would conceivably affect this flexibility as well, introducing competing effects. Revisiting the input NA structure in **Figure 3.1**, we hypothesized that the glycan on top of each NA monomer (the oligomannose type glycans linked to site #1) would achieve a higher degree of flexibility than the two on the bottom of each monomer (the complex and hybrid type glycans linked to sites #2 and #3, respectively). Our reasoning was that these two may find steric restrictions on their flexibility, and that the placement on the glycan on the NA head would be more important than the type of glycan examined. Our results show this is not quite the case. The clusters in **Figure 3.3**, **Figure 3.13**, **Figure 3.14**, and **Figure 3.15**, show that, similar to the complex-type glycans (A-D2), the oligomannose-type glycans (A-D1) were quite flexible even though they were situated near the hybrid-type glycans (A-D3) on the bottom of the NA surface; this large degree of conformational freedom is backed up by previous work specifying that this flexibility is driven by the mannose(4)- α (1-3)-mannose(3) and the mannose(5)- α (1-6)-mannose(3) linkages¹⁰⁹. These are the linkages connecting the chitobiose glycan “stalk” to the two glycan “branches”. Finally, the hybrid-type glycans showed noticeably less conformational flexibility than either the oligomannose-type glycans or the complex-type glycans. Overall, the type of glycan and its specific linkages seemed to govern its flexibility more than potential nearby steric clashes. This agrees with previous work showing that unless there is a direct steric clash, inter-residue hydrogen bonds may have a larger effect governing glycan conformations^{106, 109}.

Glycan effects on association rates

The results shown in **Figure 3.5** are consistent with diffusion controlled reactions, and show relatively high association rates. The space explored is consistent with the random walk nature of diffusion. The randomness of the ligand trajectories (from Brownian motion) and the small sizes of the ligands considered here minimize the effects of the glycans on binding. The rates for each ligand are mostly of similar orders of magnitude, with or without glycosylation. However, the glycan structures from the MD simulations show a moderate inhibition compared to the unglycosylated NA structure and the NA structure with glycan structures taken directly from the Glyprot web server. The extent of this inhibition ranges from a factor of about two to eight.

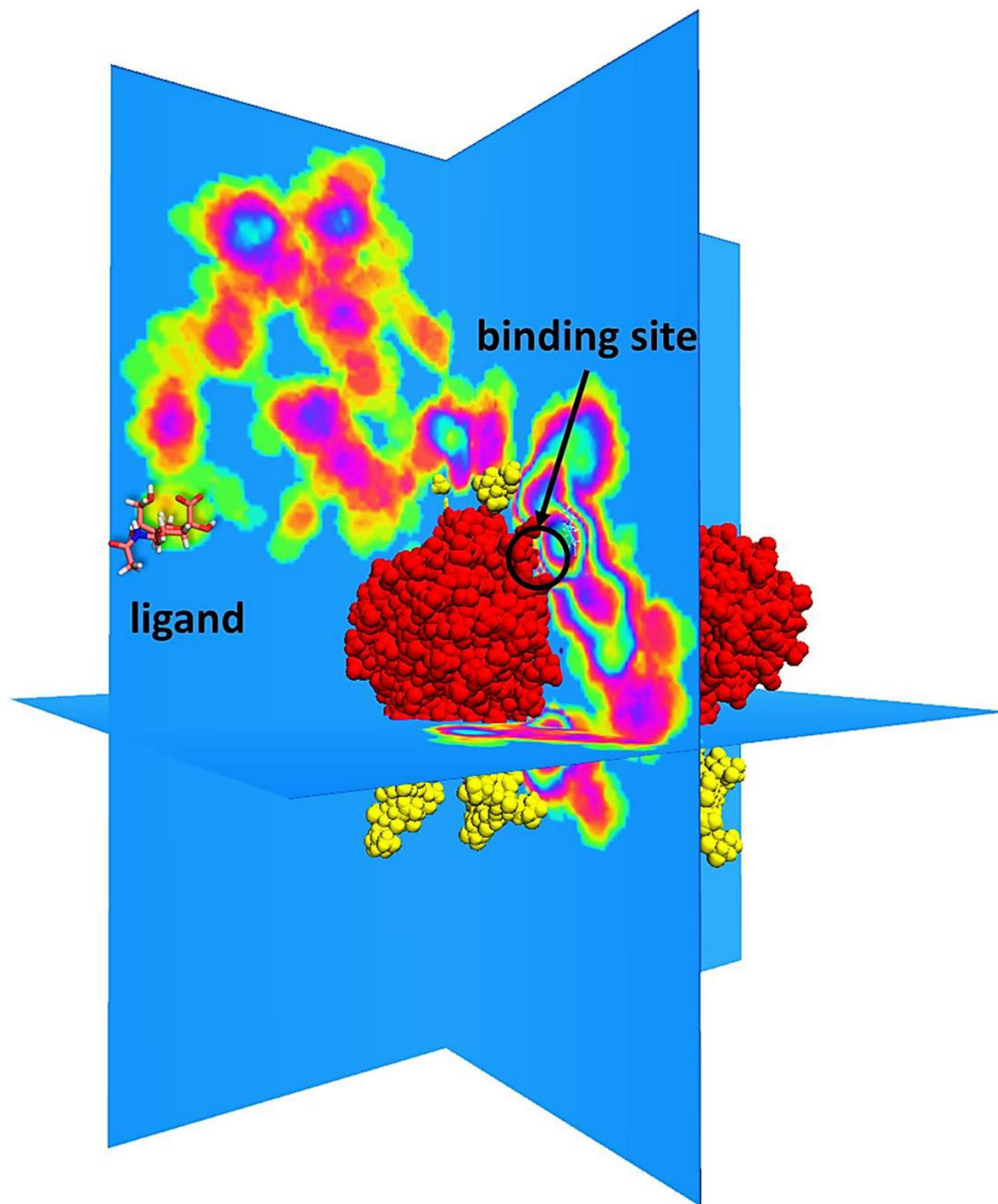
In general, glycans can decrease binding activity of viral proteins ^{3, 42, 44, 110}. Due to their bulk and proximity to the primary ligand binding site, we hypothesized that, irrespective of conformation, the presence of glycans, particularly those near the binding sites, could substantially reduce ligand binding and removing these glycans would restore binding. What we found was a more nuanced picture. The NA constructs with glycan conformations from the Glyprot server showed similar binding rates to unglycosylated constructs. However, more realistic glycan conformations, extracted from the MD simulations, showed a moderate but noticeable decrease in association rate, k_{on} , on the order of two to eight times. One may naturally question whether glycans would have the same effect on dissociation rate, k_{off} . One previous study testing antibody binding to cancer cells showed that antibody binding was relatively insensitive to the presence of glycans, indicating a similar dampening of k_{on} and k_{off} due to the presence of glycans ⁹. In this study mentioned, the overall equilibrium constant K_D changed by less than a factor of two irrespective of the presence or absence of glycans ⁹. However, a different study done in the influenza membrane protein hemagglutinin showed that trimming the glycans from a standard length seen in HEK293 cells to a single monosaccharide decreases the equilibrium constant K_D by a factor of two to 48, depending on the receptor mimic used ⁴². This meant that the k_{on} and the k_{off} were not affected in the same way by the presence of glycans ⁴². Glycans are present in the antibody binding sites of both of the studies mentioned above; this is in contrast to our system where glycans are situated near the catalytic sites, but not directly inside them. With this in mind, it seems likely that the slight slowing of binding small ligands by the glycans would be similarly reflected in a slight slowing of release, so that the equilibrium constants for binding these molecules are relatively insensitive to the presence of glycans. In effect, this is because the presence of glycans near the binding sites should not change the ΔG in accordance with the Gibbs relationship. We hypothesize that our observed decrease in association rate is due to the glycans at glycosylation site N146 (site #2) as only those glycans are capable of sterically inhibiting the binding sites (**Figure 3.2**), and we assume the glycans at sites N88 (site #1) and N234 (site #3) do not impair binding. Taking the inhibition results discussed here with a different binding study using larger ligands for influenza NA, there appears to be a size dependence on this inhibitory potential: smaller ligands are not as affected as larger ligands ⁴³. The key points here are that small molecules are not seriously impeded from binding by the glycans; future drug discovery efforts can be

focused on the development of strong binders with correspondingly long lifetimes of binding. Modeling studies focused on small inhibitors are likely to be helpful, even when glycans are not included.

The results seen in **Figure 3.5** highlight the importance of using biologically-relevant glycan conformations relaxed on the protein structure as opposed to simply generating a glycan conformation and attaching it to the protein. Though this study did use static structures as per the BD setup, we would expect similar trends if this study were repeated using a dynamic MD environment since our BD trajectories already used the most highly-accessed glycan conformations gleaned from extensive MD sampling. Moreover, a study using mixed BD-MD simulations analyzing the association of oseltamivir and zanamivir to NA actually showed a less accurate k_{on} rate than our coarser study using only BD ⁵⁴. We can rationalize that the slower binding kinetics seen in our systems with biologically-relevant glycan conformations (**Figure 3.5**) are due to the ligands having to maneuver around the glycans, even after running into them, and then continuing with the trajectory until reaching the binding site. This type of maneuverability can be seen in **Figure 3.6**.

Figure 3.6

Figure 3.6 | A sample successful trajectory, showing that the ligand explores significant space before finding the binding site. This sample NA contains glycans from the Glyprot server. Sialic acid (not to scale) is superimposed on the left side of the image for clarity. Because of the space explored, it makes sense that a blockage in part of the path won't significantly affect the association rate. Here, the protein is in red and the glycans are in yellow. The blue planes are used for creating subsections of 3D space into 2D space. The colors on the planes indicate how often the ligand has spent in that 2D space, with the lime green inside magenta circles being the most occupied.



Ligand binding at the primary and secondary sites

We generated BD trajectories that could end with the ligand binding to the primary site (**Figure 3.5A**), the secondary site (**Figure 3.5B**), or either site (**Figure 3.5C**) on any monomer. Using this setup, we were able to

differentiate binding between the primary and secondary sites, and in fact found an additive binding mode when examining both sites concurrently. By simply adding up the association rates observed for the primary site (**Figure 3.5A**) to the analogous simulation run to the secondary site (**Figure 3.5B**), the association rate to both sites (**Figure 3.5C**) can be roughly obtained. We do not see any evidence of a further increase in association rate using both sites, showing that the presence of a proximal binding site does not influence association rate, either for the primary site or the secondary site.

Our primary site binding results show two conclusions supported by literature. In **Figure 3.5A** we see that oseltamivir associates faster than zanamivir, as has been seen in experimental kinetics studies⁶⁰. Moreover, we see faster binding of oseltamivir than sialic acid. This is qualitatively in agreement with an NMR study showing that oseltamivir outcompetes $\alpha(2,3)$ -sialyllactose in binding to the avian NA active site⁹⁷. It is not immediately clear which atoms on the ligands drive their binding differences.

Ligand binding to the secondary site has not been extensively studied, but it does not appear to have catalytic activity¹¹¹⁻¹¹². Focusing on the secondary site, our results show three important findings. We first see that binding to the secondary site is slower than to the primary site, if binding is seen at all (**Figure 3.5**). We do not see secondary site binding for oseltamivir and very little for zanamivir, though this may be as they are at the lower detection limit of our method. Furthermore, we see that sialic acid binds faster to the secondary site than oseltamivir, which is in agreement with one study showing that $\alpha(2,3)$ -sialyllactose outcompetes oseltamivir for binding to the avian NA secondary site⁹⁷. A more recent study goes further and does not show any binding of oseltamivir to the avian NA secondary site⁹⁸. However, we caution that a small amount of drug binding, likely only with zanamivir, may occur to the secondary site, as seen with zanamivir bound in the secondary site in the unpublished crystal structure PDB 2CML, and also seen in **Figure 3.5B**. Secondly, in the small amount of secondary site binding seen (**Figure 3.5B**), glycans are actually capable of enhancing or inhibiting binding, foreshadowing the complex role glycans play in ligand binding. Finally, there appears to be a small conformational dependence on association rate, but this is only seen towards the secondary site (**Figure 3.5B**). We used two different conformations of sialic acid for these binding studies. The boat conformation was crystallized in the active site and the chair conformation was crystallized in the secondary site. In our results we see the sialic acid chair conformation actually shows fractionally higher binding to the primary site than the boat

conformation (**Figure 3.5A**). Conversely, only the boat conformation shows binding to the secondary site; the chair conformation does not register binding at all (**Figure 3.5B**). However, we caution that these results may be because sialic acid was crystallized in a different strain of avian NA than we used in our studies. Taken together, these results show that the exact ligand conformation upon approach to the binding site may not match the crystallized binding pose, but the results we present here do not permit us to explore this note or further explain a conformational dependence on binding.

Comparing the association rates in **Figure 3.5** one may naturally query the competition in association rates between the primary and secondary sites. We see faster association to the primary site than the secondary site, which is not in agreement with two previous BD simulation studies⁵²⁻⁵³. However, the methodology of our study differs from these two studies, and from this we can unify the difference. Those BD studies showed that ligands reach a distance of 7.5 Å away from the secondary site faster than to the primary site. We then show that ligands reach a distance of 3.228 Å away from the primary site faster than the secondary site, though we would like to note that our paper and the Sung et al. paper assigned charges for the BD trajectories according to the AMBER99 force field and the Amaro et al. paper assigned charges according to the CHARMM36 force field⁵²⁻⁵³. Taken together, the secondary site appears to contain stronger long-range electrostatics to draw in ligands, but when the ligands approach the binding sites and sterics come into play, it appears to be more favorable for ligands to move closer to the primary site than the secondary site, assuming the rigid body approximations applied herein. Considering the fact that the realistic substrates NA encounters will exhibit multivalent binding, one previous study showed that the secondary site improved avian NA enzymatic activity in removing sialic acid both from soluble macromolecular substrates and from cells¹¹³. Another study confirmed that the binding in the secondary site improved catalytic activity against multivalent substrates¹¹⁴. Other previous studies have suggested that the secondary site enhances the overall NA catalytic activity by binding substrates and bringing them close to the catalytic primary site^{111, 113-116}. Taking the studies above with our results, we postulate that multivalent cleavage will occur in a stepwise manner (**Figure 3.7**). The first association event of the multivalent substrate, such as sialylated cell surface receptors, will bind to the primary site, and then to the secondary site. After sialidase cleavage occurs in the primary site, the cleaved glycan branch will dissociate. Then the sialylated glycan branch bound in the secondary site will be transferred to the primary site, as suggested previously^{111, 113-115}. After this

passage, cleavage will again occur, and the full glycan will be released, finishing the enzymatic cycle. This mechanism is in disagreement with a previously proposed mechanism, which postulates that both binding sites will not be bound simultaneously ¹¹⁵. However, we feel there is a greater body of literature suggesting that binding both sites simultaneously increases catalytic activity. We note that our proposed binding mechanism may be muddled in the case of multivalent ligands with viral glycans situated near the binding sites; in this case, the glycans may sterically inhibit multivalent binding, slowing down enzymatic activity and attenuating the replication cycle. In the case of monovalent binders, such as the inhibitors oseltamivir and zanamivir, we show in **Figure 3.5** that association will happen to the primary site faster than to the secondary site. This appears to be biologically viable considering that previous studies have showed that the secondary site activity has no effect on enzymatic activity for monovalent substrates ^{98, 117-119}. As the primary site is the main site of enzymatic activity, it is reasonable to assume that ligands would preferentially bind to the primary site over the secondary site; reducing transfers of ligands between the binding sites would ostensibly increase catalytic activity and efficiency. Taken together, abolishing the secondary site in avian NA will not affect monovalent substrates such as influenza drugs as these associate faster to the primary site anyways, which our results confirm. To exposit this a different way, influenza drugs will preferentially block primary site binding over secondary site binding.

Figure 3.7

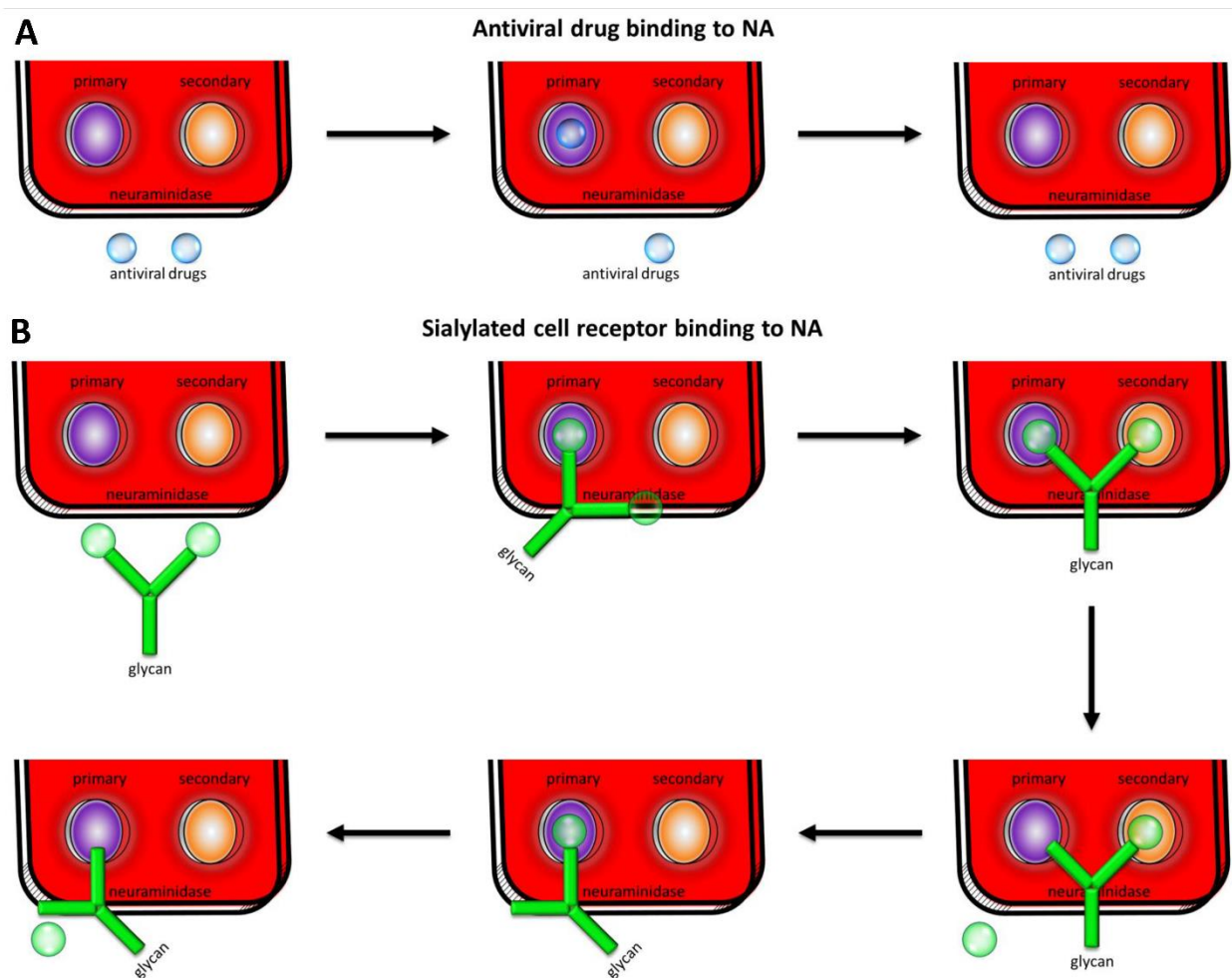


Figure 3.7 | Ligand binding mechanisms to neuraminidase for monovalent binders (A) and multivalent binders (B). A free monovalent binder will associate to the primary site over the secondary site (A). This monovalent binder will release from the primary site before a second monovalent binder will associate to the secondary site. Glycans, with their sialic acid tips, are an example of a multivalent binder (B). Similar to the monovalent binders, the first multivalent binding event will occur to the primary site. Next, the second sialic acid tip binds to the secondary site. With both sites bound, the sialic acid in the primary site is cleaved and released. The sialic acid bound in the secondary site is then transferred to the primary site. Finally, the second sialic acid is cleaved and released, and the enzymatic cycle is complete.

3.7 Conclusion

In this work, we created NA systems with varying glycan conformations, and also without the presence of glycans. These glycans are capable of covering much of the surface area of NA. Their conformational flexibility is dependent on their glycan type, not necessarily their spatial position. The glycosylated systems showed moderate inhibition of ligands to the primary binding site. Finally, we propose a new binding mechanism for multivalent binders to NA, such as cell surface receptors. These results have implications for future drug

development, the overall understanding of glycans, and the NA enzymatic mechanism. Much sustained effort has gone into developing NA inhibitors, and will continue to do so in the future. Measuring the binding of a potential drug is an important step in the drug discovery process. However, most drug discovery efforts have not taken into account viral glycans. Neglecting this effect can lead to a surprising drop in drug binding⁹. Our work shows that glycans can have an inhibitory effect on influenza NA primary site binding. There have already been a number of studies using multivalent binders as NA antivirals¹²⁰⁻¹²⁵. With the results shown here, we recommend future work on multivalent NA drugs, to focus on developing strong binders with a long lifetime, regardless of the presence or absence of glycans. With the detection limitations of our study, we cannot conclude how glycans affect secondary site binding, although we believe binding to the secondary site will be slower than binding to the primary site (**Figure 3.55**). However, it follows from these results that glycans could evoke a secondary site binding inhibition similar to the primary site. In summary, this work examines glycan inhibition on drug binding, compares the drug binding interplay between two binding sites, and proposes a new mechanism of ligand binding to NA.

3.8 Author Contributions

C.S., R.E.A., and J.A.M. designed the experiments and performed analysis. C.S. ran the simulations and wrote the manuscript with contributions from all authors. L.C., R.K., and G.H. provided critical support necessary for the simulations to run. All authors approved the final version of the manuscript.

3.9 Acknowledgements

C.S. would like to thank Zied Gaieb for valuable discussions. This work was supported in part by grants from the National Institutes of Health, USA (NIH grant T32EB009380 to C.S. and GM031749 to J.A.M. and G.H.). This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1650112 to C.S. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) Comet at the San Diego Supercomputer Center

(SDSC) through allocation csd373 to R.E.A. R.E.A. has equity interest in and is a cofounder of and on the scientific advisory board of Actavalon.

Chapter 3, in full, is a modified reprint of the material as it appears in “Seitz, C., Casalino, L., Konecny, R., Huber, G., Amaro, R.E., McCammon, J.A. Multiscale Simulations Examining Glycan Shield Effects on Drug Binding to Influenza Neuraminidase. *Biophys. J.*, **2020**, *119* (11), pp 2275-2289.” The dissertation author was a co-investigator and co-author of this work.

3.10 Supporting Material

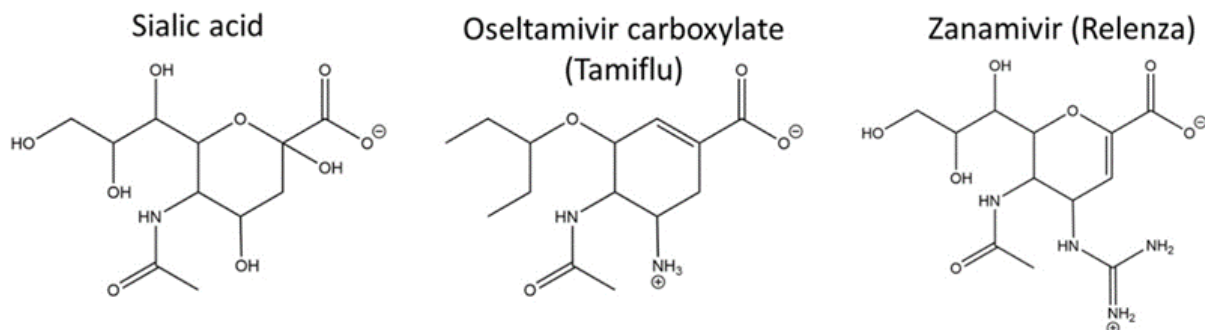


Figure 3.8 | 2D structures of the ligands used in this study.

Table 3.2 | Ions used in the implicit solvent in the BD trajectories.

Ion	Charge	Concentration	Radius
Ca ²⁺	2.000	0.005 M	1.7131 Å
Cl ⁻	-1.000	0.010 M	1.7350 Å
Na ⁺	1.000	0.024 M	1.8680 Å
MES ⁻	-1.000	0.024 M	4.0070 Å

Table 3.3 | Grid spacing for each system. The grid spacing is defined as the fine grid length divided by the number of grid points.

System	Grid spacing
Unglycosylated NA	0.61 x 0.61 x 0.66 Å
NA with Glyprot glycans	0.80 x 0.80 x 0.83 Å
NA with MD conf1 glycans	0.59 x 0.59 x 0.68 Å
NA with MD conf2 glycans	0.60 x 0.61 x 0.70 Å
NA with MD conf3 glycans	0.60 x 0.61 x 0.70 Å

Table 3.4 | b radius and q radius values for each system. The distance is the distance away from the hydrodynamic center of the protein.

System	b radius	q radius
Unglycosylated NA	109.047 Å	119.952 Å
NA with Glyprot glycans	109.310 Å	120.241 Å
NA with MD conf1 glycans	111.999 Å	123.199 Å
NA with MD conf2 glycans	110.382 Å	121.420 Å
NA with MD conf3 glycans	110.098 Å	121.108 Å

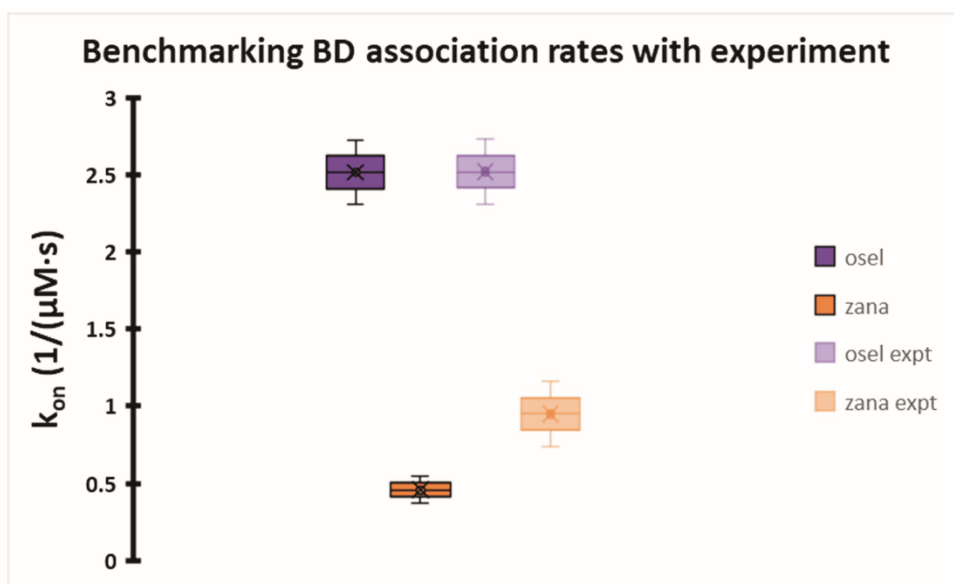


Figure 3.9 | Comparison of our computed association rates with previously published experimental association rates. The computed oseltamivir rate (purple) is within the level of error of the experimental oseltamivir rate (faded purple). The computed zanamivir rate (orange) is similar to the experimental zanamivir rate (faded orange).

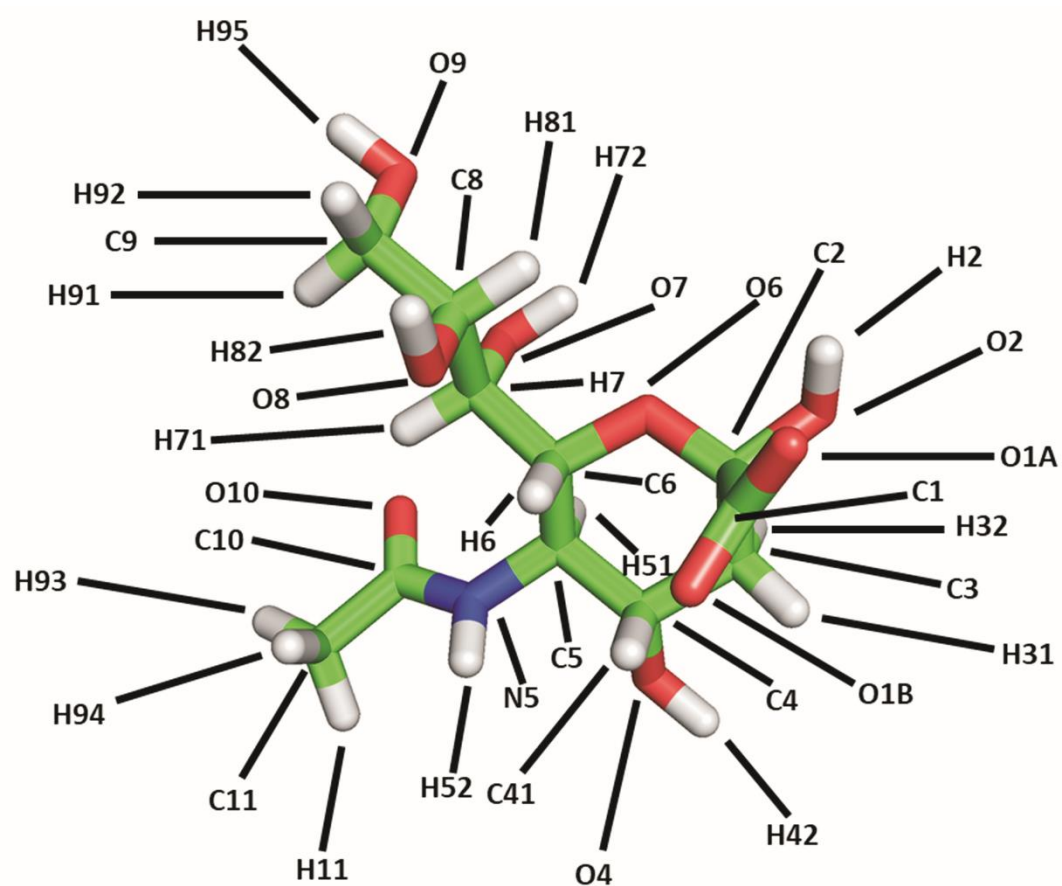
Overall, the primary site binding parameters were transferred from oseltamivir to zanamivir, correcting for the replacement of oseltamivir's isopentanetriol ring substituent with zanamivir's 1,2,3-pentanetriol ring substituent, as seen in **Figure 3.8**. They were then transferred to sialic acid, accounting for the methanetriamine ring substituent (**Figure 3.8**). The secondary site binding parameters were determined for sialic acid before being transferred to oseltamivir and zanamivir. These choices were made due to the availability of crystal structures.

The primary binding site contacts for oseltamivir are presented in **Table 3.4**. The oseltamivir isopentanetriol group is replaced by the sialic acid 08 and 09 oxygen atoms in binding pairs for both the primary and secondary sites (**Table 3.6** and **Table 3.9**). This isopentanetriol group in oseltamivir appears to form hydrophobic contacts with E276 of the neuraminidase instead of the hydrogen bonding that zanamivir and sialic acid shows, so these hydrophobic contacts were approximated as having contacts to the C91, C8 and C82 atoms as shown above. The zanamivir (**Table 3.5**) binding pairs are based off of those for oseltamivir.

The 1MWE crystal structure shows the tryptophan's conjugated benzene ring having an interaction with the C11 atom of sialic acid in the secondary binding site. We approximated this putative hydrophobic interaction by

In summary, the primary and secondary site binding pairs for these ligands are set to be analogs of each other. The NA binding pair atoms were kept constant across each ligand for consistency.





Sialic acid

Figure 3.11 | Structure of sialic acid.

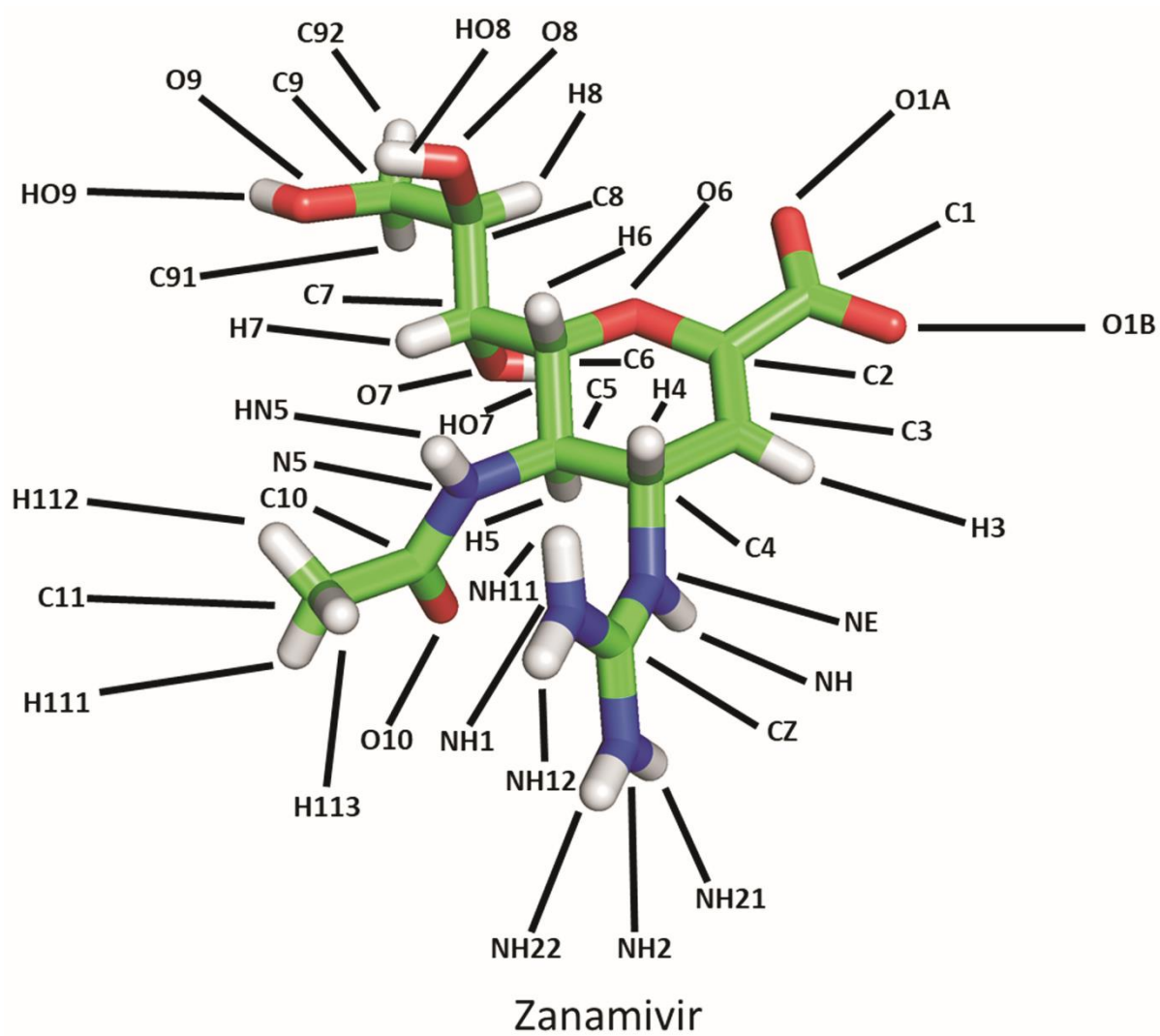


Figure 3.12 | Structure of zanamivir.

Table 3.5 | Primary site atom pairs for oseltamivir.

Neuraminidase residue	Residue atom	Oseltamivir atom
R118	N η_1	O1A
R118	N η_1	O1B
R118	N η_2	O1A
R118	N η_2	O1B
R152	N η_1	O10
R152	N η_2	O10
E276	C β	C8
E276	C β	C82
E276	C β	C91
R292	N η_1	O1A
R292	N η_1	O1B
R292	N η_2	O1A
R292	N η_2	O1B
Y347	O η	O1A
Y347	O η	O1B
R371	N η_1	O1A
R371	N η_1	O1B
R371	N η_2	O1A
R371	N η_2	O1B
R371	N ϵ	O1B
R371	N ϵ	O1A
Y406	O η	O1A
Y406	O η	O1B

Table 3.6 | Primary site atom pairs for sialic acid.

Neuraminidase residue	Residue atom	Sialic acid atom
R118	N η_1	O1A
R118	N η_1	O1B
R118	N η_2	O1A
R118	N η_2	O1B
R152	N η_1	O10
R152	N η_2	O10
E276	O ϵ_1	O8
E276	O ϵ_2	O8
E276	O ϵ_1	O9
E276	O ϵ_2	O9
R292	N η_1	O1A
R292	N η_1	O1B
R292	N η_2	O1A
R292	N η_2	O1B
Y347	O η	O1A
Y347	O η	O1B
R371	N η_1	O1A
R371	N η_1	O1B
R371	N η_2	O1A
R371	N η_2	O1B
R371	N ϵ	O1B
R371	N ϵ	O1A
Y406	O η	O1A
Y406	O η	O1B

Table 3.7 | Primary site atom pairs for zanamivir.

Neuraminidase residue	Residue atom	Zanamivir atom
R118	N η_1	O1A
R118	N η_1	O1B
R118	N η_2	O1A
R118	N η_2	O1B
R152	N η_1	O10
R152	N η_2	O10
E276	O ϵ_1	O8
E276	O ϵ_2	O8
E276	O ϵ_1	O9
E276	O ϵ_2	O9
R292	N η_1	O1A
R292	N η_1	O1B
R292	N η_2	O1A
R292	N η_2	O1B
Y347	O η	O1A
Y347	O η	O1B
R371	N η_1	O1A
R371	N η_1	O1B
R371	N η_2	O1A
R371	N η_2	O1B
R371	N ϵ	O1B
R371	N ϵ	O1A
Y406	O η	O1A
Y406	O η	O1B

Table 3.8 | Secondary site atom pairs for sialic acid.

Neuraminidase residue	Residue atom	Sialic acid atom
S367	OG	O1A
S367	OG	O1B
S370	CA	C91
S370	CA	C8
S370	CA	C82
S372	OG	N5
W403	CD2	C11
W403	CZ2	C11
W403	CZ3	C11
K432	CG	O8
K432	CG	O9

Table 3.9 | Secondary site atom pairs for oseltamivir.

Neuraminidase residue	Residue atom	Oseltamivir atom
S367	OG	O1A
S367	OG	O1B
S370	CA	C91
S370	CA	C8
S370	CA	C82
S372	OG	N5
W403	CD2	C11
W403	CZ2	C11
W403	CZ3	C11
K432	CG	C91
K432	CG	C8
K432	CG	C82

Table 3.10 | Secondary site atom pairs for zanamivir.

Neuraminidase residue	Residue atom	Zanamivir atom
S367	OG	O1A
S367	OG	O1B
S370	CA	C91
S370	CA	C8
S370	CA	C82
S372	OG	N5
W403	CD2	C11
W403	CZ2	C11
W403	CZ3	C11
K432	CG	O8
K432	CG	O9

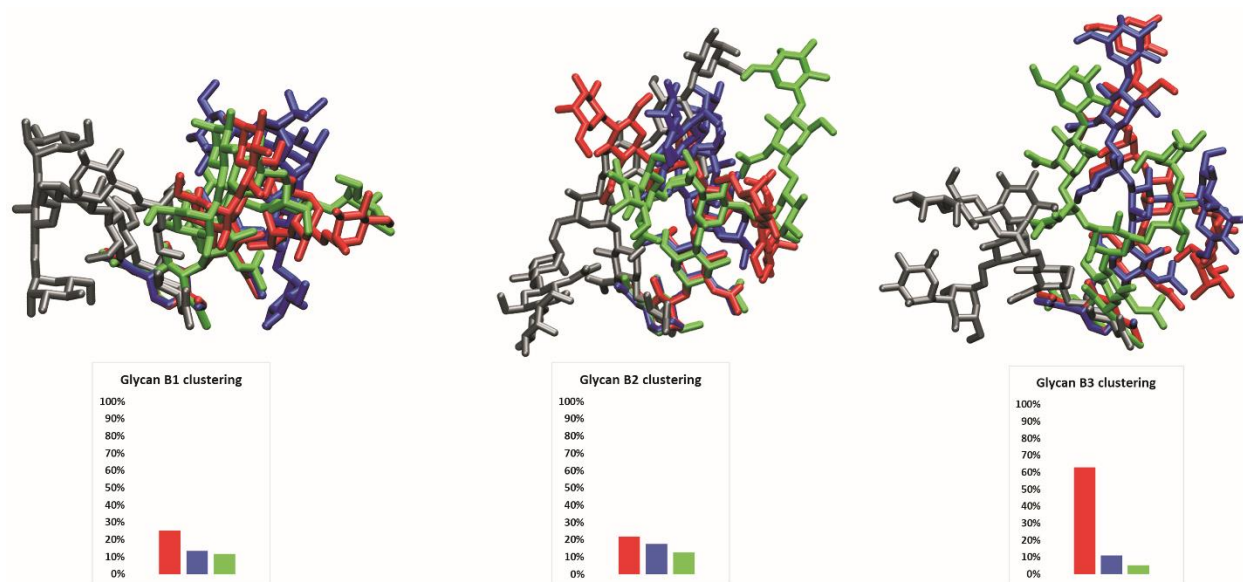


Figure 3.13 | Clustered glycans on monomer B. The Glyprot structure is in gray, while the other colors represent clusters from the MD simulations.

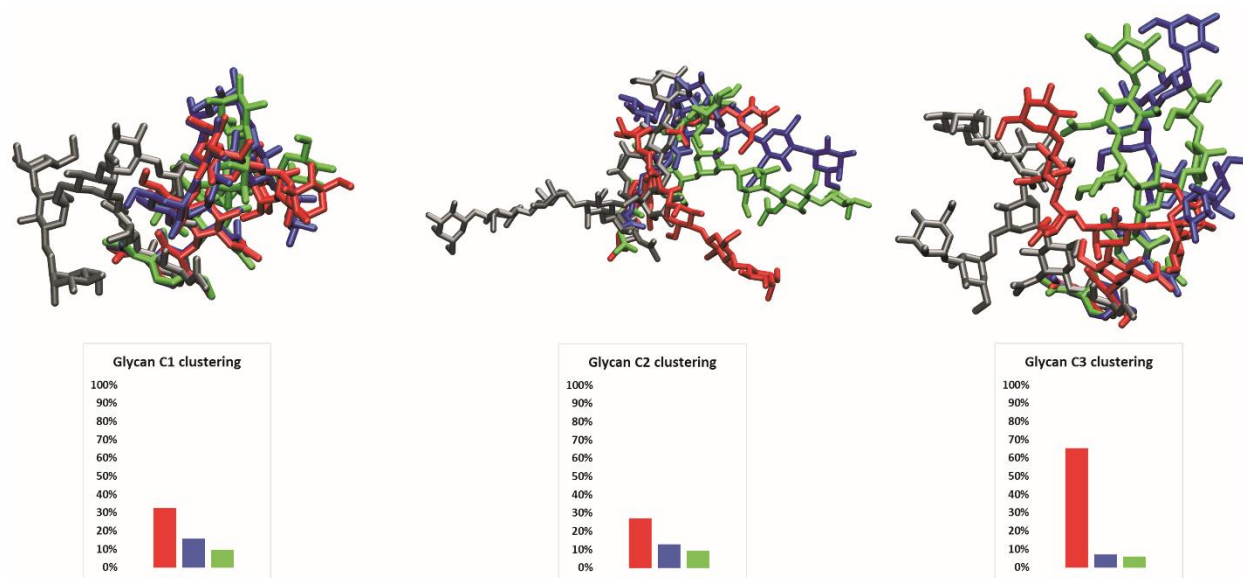


Figure 3.14 | Clustered glycans on monomer C. The Glyprot structure is in gray, while the other colors represent clusters from the MD simulations.

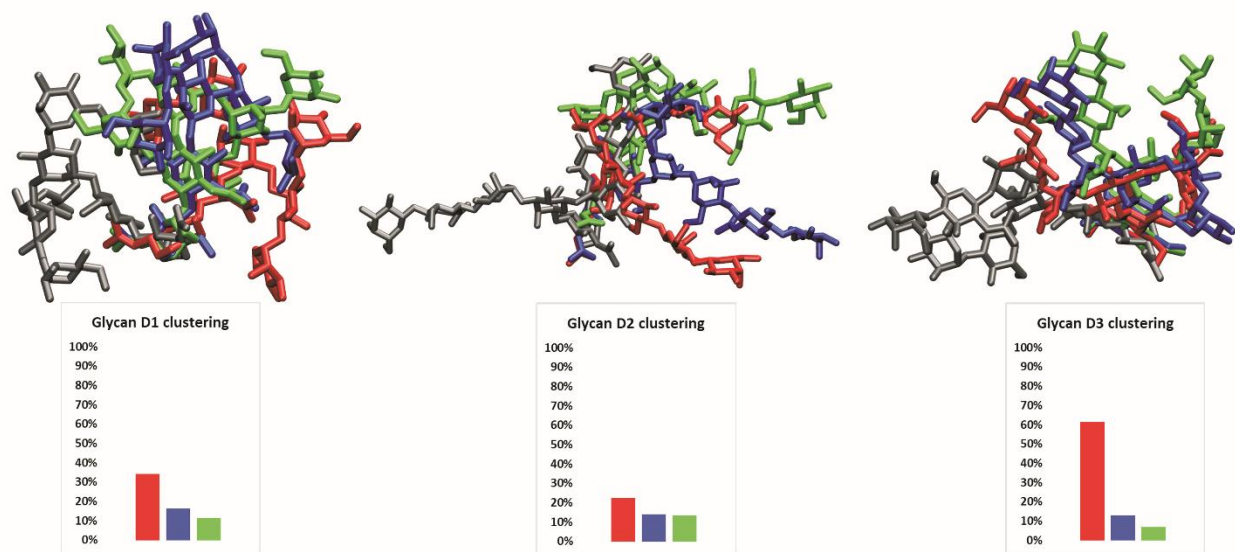


Figure 3.15 | Clustered glycans on monomer D. The Glyprot structure is in gray, while the other colors represent clusters from the MD simulations.

Table 3.11 | Association rate data for oseltamivir. The files containing this data are provided.

System	k _{on}			Probability			Total binding events
	low	mean	high	low	mean	high	
Primary site							
Unglycosylated NA	3.66/μM·s	3.93/μM·s	4.19/μM·s	8.28 x 10 ⁻⁵	8.88 x 10 ⁻⁵	9.48 x 10 ⁻⁵	888
NA with Glyprot glycans	2.31/μM·s	2.52/μM·s	2.73/μM·s	5.24 x 10 ⁻⁵	5.72 x 10 ⁻⁵	6.20 x 10 ⁻⁵	572
NA with MD conf1 glycans	1.22/μM·s	1.38/μM·s	1.53/μM·s	2.68 x 10 ⁻⁵	3.03 x 10 ⁻⁵	3.38 x 10 ⁻⁵	303
NA with MD conf2 glycans	0.47/μM·s	0.57/μM·s	0.68/μM·s	1.06 x 10 ⁻⁵	1.29 x 10 ⁻⁵	1.52 x 10 ⁻⁵	129
NA with MD conf3 glycans	1.39/μM·s	1.55/μM·s	1.72/μM·s	3.13 x 10 ⁻⁵	3.50 x 10 ⁻⁵	3.87 x 10 ⁻⁵	350
Secondary site							
Unglycosylated NA	0.00/μM·s	0.00/μM·s	0.00/μM·s	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0
NA with Glyprot glycans	0.00/μM·s	0.00/μM·s	0.00/μM·s	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0
NA with MD conf1 glycans	0.00/μM·s	0.00/μM·s	0.00/μM·s	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0
NA with MD conf2 glycans	0.00/μM·s	0.00/μM·s	0.00/μM·s	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0
NA with MD conf3 glycans	0.00/μM·s	0.00/μM·s	0.00/μM·s	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0
Combined sites							
Unglycosylated NA	3.67/μM·s	3.93/μM·s	4.19/μM·s	8.29 x 10 ⁻⁵	8.89 x 10 ⁻⁵	9.49 x 10 ⁻⁵	889
NA with Glyprot glycans	2.49/μM·s	2.70/μM·s	2.92/μM·s	5.65 x 10 ⁻⁵	6.15 x 10 ⁻⁵	6.65 x 10 ⁻⁵	615
NA with MD conf1 glycans	1.47/μM·s	1.64/μM·s	1.82/μM·s	3.24 x 10 ⁻⁵	3.62 x 10 ⁻⁵	4.00 x 10 ⁻⁵	362
NA with MD conf2 glycans	0.48/μM·s	0.58/μM·s	0.68/μM·s	1.07 x 10 ⁻⁵	1.30 x 10 ⁻⁵	1.53 x 10 ⁻⁵	130
NA with MD conf3 glycans	1.42/μM·s	1.59/μM·s	1.75/μM·s	3.19 x 10 ⁻⁵	3.57 x 10 ⁻⁵	3.95 x 10 ⁻⁵	357

Table 3.12 | Association rate data for zanamivir. The files containing this data are provided.

System	k _{on}			Probability			Total binding events
	low	mean	high	low	mean	high	
Primary site							
Unglycosylated NA	0.28/μM·s	0.36/μM·s	0.44/μM·s	0.66 x 10 ⁻⁵	0.84 x 10 ⁻⁵	1.02 x 10 ⁻⁵	84
NA with Glyprot glycans	0.37/μM·s	0.46/μM·s	0.55/μM·s	0.85 x 10 ⁻⁵	1.06 x 10 ⁻⁵	1.27 x 10 ⁻⁵	106
NA with MD conf1 glycans	0.10/μM·s	0.16/μM·s	0.21/μM·s	0.23 x 10 ⁻⁵	0.35 x 10 ⁻⁵	0.47 x 10 ⁻⁵	35
NA with MD conf2 glycans	0.03/μM·s	0.07/μM·s	0.10/μM·s	0.07 x 10 ⁻⁵	0.15 x 10 ⁻⁵	0.23 x 10 ⁻⁵	15
NA with MD conf3 glycans	0.11/μM·s	0.17/μM·s	0.22/μM·s	0.26 x 10 ⁻⁵	0.38 x 10 ⁻⁵	0.50 x 10 ⁻⁵	38
Secondary site							
Unglycosylated NA	0.00/μM·s	0.00/μM·s	0.00/μM·s	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0
NA with Glyprot glycans	0.00/μM·s	0.00/μM·s	0.00/μM·s	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0
NA with MD conf1 glycans	0.01/μM·s	0.04/μM·s	0.07/μM·s	0.03 x 10 ⁻⁵	0.09 x 10 ⁻⁵	0.15 x 10 ⁻⁵	9
NA with MD conf2 glycans	0.00/μM·s	0.01/μM·s	0.02/μM·s	-0.01 x 10 ⁻⁵	0.02 x 10 ⁻⁵	0.05 x 10 ⁻⁵	2
NA with MD conf3 glycans	0.02/μM·s	0.05/μM·s	0.08/μM·s	0.05x 10 ⁻⁵	0.12 x 10 ⁻⁵	0.19 x 10 ⁻⁵	12
Combined sites							
Unglycosylated NA	0.27/μM·s	0.35/μM·s	0.43/μM·s	0.63 x 10 ⁻⁵	0.81 x 10 ⁻⁵	0.99 x 10 ⁻⁵	81
NA with Glyprot glycans	0.27/μM·s	0.35/μM·s	0.43/μM·s	0.64 x 10 ⁻⁵	0.82 x 10 ⁻⁵	1.00 x 10 ⁻⁵	82
NA with MD conf1 glycans	0.16/μM·s	0.22/μM·s	0.29/μM·s	0.36 x 10 ⁻⁵	0.50 x 10 ⁻⁵	0.64 x 10 ⁻⁵	50
NA with MD conf2 glycans	0.03/μM·s	0.07/μM·s	0.10/μM·s	0.07 x 10 ⁻⁵	0.15 x 10 ⁻⁵	0.23 x 10 ⁻⁵	15
NA with MD conf3 glycans	0.17/μM·s	0.24/μM·s	0.30/μM·s	0.40 x 10 ⁻⁵	0.55 x 10 ⁻⁵	0.70 x 10 ⁻⁵	55

Table 3.13 | Association rate data for sialic acid boat conformation. The files containing this data are provided.

System	k _{on}			Probability			Total binding events
	low	mean	high	low	mean	high	
Primary site							
Unglycosylated NA	1.46/μM·s	1.63/μM·s	1.81/μM·s	3.24 x 10 ⁻⁵	3.62 x 10 ⁻⁵	4.00 x 10 ⁻⁵	362
NA with Glyprot glycans	1.76/μM·s	1.95/μM·s	2.14/μM·s	3.93 x 10 ⁻⁵	4.35 x 10 ⁻⁵	4.77 x 10 ⁻⁵	435
NA with MD conf1 glycans	0.12/μM·s	0.18/μM·s	0.24/μM·s	0.27 x 10 ⁻⁵	0.39 x 10 ⁻⁵	0.51 x 10 ⁻⁵	39
NA with MD conf2 glycans	0.03/μM·s	0.06/μM·s	0.10/μM·s	0.07 x 10 ⁻⁵	0.14 x 10 ⁻⁵	0.21 x 10 ⁻⁵	14
NA with MD conf3 glycans	1.11/μM·s	1.26/μM·s	1.41/μM·s	2.46 x 10 ⁻⁵	2.79 x 10 ⁻⁵	3.12 x 10 ⁻⁵	279
Secondary site							
Unglycosylated NA	0.24/μM·s	0.32/μM·s	0.40/μM·s	0.54 x 10 ⁻⁵	0.71 x 10 ⁻⁵	0.88 x 10 ⁻⁵	71
NA with Glyprot glycans	0.14/μM·s	0.20/μM·s	0.26/μM·s	0.31 x 10 ⁻⁵	0.45 x 10 ⁻⁵	0.58 x 10 ⁻⁵	45
NA with MD conf1 glycans	0.49/μM·s	0.60/μM·s	0.70/μM·s	1.06 x 10 ⁻⁵	1.29 x 10 ⁻⁵	1.52 x 10 ⁻⁵	129
NA with MD conf2 glycans	0.06/μM·s	0.10/μM·s	0.14/μM·s	0.13 x 10 ⁻⁵	0.22 x 10 ⁻⁵	0.31 x 10 ⁻⁵	22
NA with MD conf3 glycans	0.11/μM·s	0.17/μM·s	0.22/μM·s	0.25 x 10 ⁻⁵	0.37 x 10 ⁻⁵	0.49 x 10 ⁻⁵	37
Combined sites							
Unglycosylated NA	1.47/μM·s	1.64/μM·s	1.81/μM·s	3.26 x 10 ⁻⁵	3.64 x 10 ⁻⁵	4.02 x 10 ⁻⁵	364
NA with Glyprot glycans	1.88/μM·s	2.07/μM·s	2.26/μM·s	4.18 x 10 ⁻⁵	4.61 x 10 ⁻⁵	5.04 x 10 ⁻⁵	461
NA with MD conf1 glycans	0.67/μM·s	0.79/μM·s	0.91/μM·s	1.44 x 10 ⁻⁵	1.70 x 10 ⁻⁵	1.96 x 10 ⁻⁵	170
NA with MD conf2 glycans	0.08/μM·s	0.13/μM·s	0.18/μM·s	0.18 x 10 ⁻⁵	0.29 x 10 ⁻⁵	0.40 x 10 ⁻⁵	29
NA with MD conf3 glycans	1.25/μM·s	1.40/μM·s	1.56/μM·s	2.74 x 10 ⁻⁵	3.10 x 10 ⁻⁵	3.45 x 10 ⁻⁵	310

Table 3.14 | Association rate data for sialic acid chair conformation. The files containing this data are provided.

System	k _{on}			Probability			Total binding events
	low	mean	high	low	mean	high	
Primary site							
Unglycosylated NA	1.68/μM·s	1.87/μM·s	2.05/μM·s	3.69 x 10 ⁻⁵	4.09 x 10 ⁻⁵	4.49 x 10 ⁻⁵	409
NA with Glyprot glycans	1.84/μM·s	2.04/μM·s	2.23/μM·s	4.07 x 10 ⁻⁵	4.49 x 10 ⁻⁵	4.91 x 10 ⁻⁵	449
NA with MD conf1 glycans	0.22/μM·s	0.29/μM·s	0.36/μM·s	0.46 x 10 ⁻⁵	0.62 x 10 ⁻⁵	0.78 x 10 ⁻⁵	62
NA with MD conf2 glycans	0.02/μM·s	0.06/μM·s	0.09/μM·s	0.05 x 10 ⁻⁵	0.12 x 10 ⁻⁵	0.19 x 10 ⁻⁵	12
NA with MD conf3 glycans	1.19/μM·s	1.34/μM·s	1.50/μM·s	2.60 x 10 ⁻⁵	2.94 x 10 ⁻⁵	3.28 x 10 ⁻⁵	294
Secondary site							
Unglycosylated NA	0.00/μM·s	0.00/μM·s	0.00/μM·s	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0
NA with Glyprot glycans	0.00/μM·s	0.00/μM·s	0.00/μM·s	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0
NA with MD conf1 glycans	0.00/μM·s	0.00/μM·s	0.00/μM·s	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0
NA with MD conf2 glycans	0.00/μM·s	0.00/μM·s	0.00/μM·s	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0
NA with MD conf3 glycans	0.00/μM·s	0.00/μM·s	0.00/μM·s	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0.00 x 10 ⁻⁵	0
Combined sites							
Unglycosylated NA	1.75/μM·s	1.93/μM·s	2.13/μM·s	3.84 x 10 ⁻⁵	4.25 x 10 ⁻⁵	4.66 x 10 ⁻⁵	425
NA with Glyprot glycans	1.87/μM·s	2.06/μM·s	2.26/μM·s	4.11 x 10 ⁻⁵	4.54 x 10 ⁻⁵	4.97 x 10 ⁻⁵	454
NA with MD conf1 glycans	0.20/μM·s	0.27/μM·s	0.34/μM·s	0.43 x 10 ⁻⁵	0.58 x 10 ⁻⁵	0.73 x 10 ⁻⁵	58
NA with MD conf2 glycans	0.01/μM·s	0.03/μM·s	0.06/μM·s	0.02 x 10 ⁻⁵	0.07 x 10 ⁻⁵	0.12 x 10 ⁻⁵	7
NA with MD conf3 glycans	1.27/μM·s	1.43/μM·s	1.59/μM·s	2.77 x 10 ⁻⁵	3.12 x 10 ⁻⁵	3.47 x 10 ⁻⁵	312

References

1. James E. Crowe, Jr., Antibody Determinants of Influenza Immunity. *J Infect Dis* **2019**, 2019 (Supplement_1), S21-S29.
2. Watanabe, Yasunori; Bowden, Thomas A.; Wilson, Ian A.; Crispin, Max, Exploitation of glycosylation in enveloped virus pathobiology. *Biochimica et Biophysica Acta - General Subjects* **2019**, 1863 (10), 1480-1497.
3. Abe, Yasuhiro; Takashita, Emi; Sugawara, Kanetsu; Matsuzaki, Yoko; Muraki, Yasushi; Hongo, Seiji, Effect of the Addition of Oligosaccharides on the Biological Activities and Antigenicity of Influenza A/H3N2 Virus Hemagglutinin. *J Virol* **2004**, 78 (18), 9605-9611.
4. Schulze, Irene T., Effects of Glycosylation on the Properties and Functions of Influenza Virus Hemagglutinin. *J Infect Dis* **1997**, 176 (Supplement_1), S24-S28.
5. Das, Suman R.; Hensley, Scott E.; David, Alexandre; Schmidt, Loren; Gibbs, James S.; Puigbò, Pere; Ince, William L.; Bennink, Jack R.; Yewdell, Jonathan W., Fitness costs limit influenza A virus hemagglutinin glycosylation as an immune evasion strategy. *Proceedings of the National Academy of Sciences USA* **2011**, 108 (51), E1417-E1422.
6. Reading, Patrick C.; Tate, Michelle D.; Pickett, Danielle L.; Brooks, Andrew G., Glycosylation as a Target for Recognition of Influenza Viruses by the Innate Immune System. *Current Topics in Innate Immunity* **2007**, 598, 279-292.
7. Tate, Michelle D.; Job, Emma R.; Deng, Yi-Mo; Gunalan, Vithiagarar; Maurer-Stroh, Sebastian; Reading, Patrick C., Playing Hide and Seek How Glycosylation of the Influenza Virus Hemagglutinin Can Modulate the Immune Response to Infection. *Viruses* **2014**, 6 (3), 1294-1316.
8. Zost, Seth J.; Parkhouse, Kaela; Gumina, Megan E.; Kim, Kangchon; Perez, Sebastian Diaz; Wilson, Patrick C.; Treanor, John J.; Sant, Andrea J.; Cobey, Sarah; Hensley, Scott E., Contemporary H3N2 influenza viruses have a glycosylation site that alters binding of antibodies elicited by egg-adapted vaccine strains. *Proceedings of the National Academy of Sciences USA* **2017**, 114 (47), 12578-12583.
9. Peiris, Diluka; Spector, Alexander F.; Lomax-Browne, Hannah; Azimi, Tayebbeh; Ramesh, Bala; Loizidou, Marilena; Welch, Hazel; Dwek, Miriam V., Cellular glycosylation affects Herceptin binding and sensitivity of breast cancer cells to doxorubicin and growth factors. *Scientific Reports* **2017**, 7.
10. Jefferson, Tom; Doshi, Peter, Multisystem failure: the story of anti-influenza drugs. *BMJ* **2014**, 348 (348), 1-3.
11. Hussein, Mazhar; Galvin, Henry D.; Haw, Tatt Y.; Nutsford, Ashley N.; Husain, Matloob, Drug resistance in influenza A virus: the epidemiology and management. *Infection and Drug Resistance* **2017**, 10, 121-134.
12. Saladino, Raffaele; Barontini, Maurizio; Crucianelli, Marcello; Nencioni, Lucia; Sgarbanti, Rossella; Palamara, Anna Teresa, Current Advances in Anti-Influenza Therapy. *Curr Med Chem* **2010**, 17 (20), 2101-2140.
13. Defaus, Sira; Gupta, Preeti; Andreu, David; Gutiérrez-Gallego, Ricardo, Mammalian protein glycosylation – structure versus function. *Analyst* **2014**, 139, 2944-2967.
14. Bause, Ernst, Structural requirements of N-glycosylation of proteins. *Biochem J* **1983**, 209 (2), 331-336.
15. Crispin, Max; Ward, Andrew B.; Wilson, Ian A., Structure and Immune Recognition of the HIV Glycan Shield. *Annual Reviews of Biophysics* **2018**, 47, 499-523.
16. Watanabe, Yasunori; Allen, Joel D.; Wrapp, Daniel; McLellan, Jason S.; Crispin, Max, Site-specific glycan analysis of the SARS-CoV-2 spike. *Science* **2020**.
17. Yasunori Watanabe, Zachary T. Berndsen, Jayna Raghwan, Gemma E. Seabright, Joel D. Allen, Oliver G. Pybus, Jason S. McLellan, Ian A. Wilson, Thomas A. Bowden, Andrew B. Ward, Max Crispin, Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. *Nature Communications* **2020**, 11.
18. Sommerstein, Rami; Flatz, Lukas; Remy, Melissa M.; Malinge, Pauline; Magistrelli, Giovanni; Fischer, Nicolas; Sahin, Mehmet; Berghaler, Andreas; Igonet, Sebastien; Meulen, Jan ter; Rigo, Dorothée; Meda, Paolo; Rabah, Nadia; Coutard, Bruno; Bowden, Thomas A.; Lambert, Paul-Henri; Siegrist, Claire-Anne; Pinschewer, Daniel D., Arenavirus Glycan Shield Promotes Neutralizing Antibody Evasion and Protracted Infection. *PLoS Pathogens* **2015**, 11 (11).
19. Falkowska, Emilia; Kajumo, Francis; Garcia, Edie; Reinus, John; Dragic, Tatjana, Hepatitis C Virus Envelope Glycoprotein E2 Glycans Modulate Entry, CD81 Binding, and Neutralization. *J Virol* **2007**, 81 (15), 8072-8079.

20. Szakonyi, Gerda; Klein, Michael G.; Hannan, Jonathan P.; Young, Kendra A.; Ma, Runlin Z.; Asokan, Rengasamy; Holers, V. Michael; Chen, Xiaojiang S., Structure of the Epstein-Barr virus major envelope glycoprotein. *Nat Struct Mol Biol* **2006**, *13* (11), 996-1001.
21. Walls, Alexandra C.; Tortorici, M. Alejandra; Frenz, Brandon; Snijder, Joost; Li, Wentao; Rey, Félix A; DiMaio, Frank; Bosch, Berend-Jan; Veesler, David, Glycan shield and epitope masking of a coronavirus spike protein observed by cryo-electron microscopy. *Nat Struct Mol Biol* **2016**, *23*, 899-905.
22. Doores, Katie J., The HIV glycan shield as a target for broadly neutralizing antibodies. *FEBS J* **2015**, *282* (24), 4679-4691.
23. Breedam, Wander Van; Pohlmann, Stefan; Favoreel, Herman W.; Groot, Raoul J. de; Nauwynck, Hans J., Bitter-sweet symphony: glycan-lectin interactions in virus biology. *FEMS Microbiol Rev* **2014**, *38* (4), 598-632.
24. Tsuchiya, Emi; Sugawara, Kanetsu; Hongo, Seiji; Matsuzaki, Yoko; Muraki, Yasushi; Li, Zhu-Nan; Nakamura, Kiyoto, Effect of addition of new oligosaccharide chains to the globular head of influenza A/H2N2 virus haemagglutinin on the intracellular transport and biological activities of the molecule. *J Gen Virol* **2002**, *83*, 1137-1146.
25. Tsuchiya, Emi; Sugawara, Kanetsu; Hongo, Seiji; Matsuzaki, Yoko; Muraki, Yasushi; Li, Zhu-Nan; Nakamura, Kiyoto, Antigenic structure of the haemagglutinin of human influenza A/H2N2 virus. *J Gen Virol* **2001**, *82*, 2475-2484.
26. Kim, Jin Il; Lee, Ilseob; Park, Sehee; Hwang, Min-Woong; Bae, Joon-Yong; Lee, Sangmoo; Heo, Jun; Park, Mee Sook; García-Sastre, Adolfo; Park, Man-Seong, Genetic Requirement for Hemagglutinin Glycosylation and Its Implications for Influenza A H1N1 Virus Evolution. *J Virol* **2013**, *87* (13), 7539-7549.
27. Alymova, Irina V.; York, Ian A.; Air, Gillian M.; Cipollo, John F.; Gulati, Shelly; Baranovich, Tatiana; Kumar, Amrita; Zeng, Hui; Gansebom, Shane; McCullers, Jonathan A., Glycosylation changes in the globular head of H3N2 influenza hemagglutinin modulate receptor binding without affecting virus virulence. *Scientific Reports* **2016**, *6*.
28. Chen, Wentian; Zhong, Yaogang; Su, Rui; Qi, Huicai; Deng, Weina; Sun, Yu; Ma, Tianran; Wang, Xilong; Yu, Hanjie; Wang, Xiurong; Li, Zheng, N-glycan profiles in H9N2 avian influenza viruses from chicken eggs and human embryonic lung fibroblast cells. *J Virol Methods* **2017**, *249*, 10-20.
29. Khatri, Kshitij; Klein, Joshua A.; White, Mitchell R.; Grant, Oliver C.; Leymarie, Nancy; Woods, Robert J.; Hartshorn, Kevan L.; Zaia, Joseph, Integrated Omics and Computational Glycobiology Reveal Structural Basis for Influenza A Virus Glycan Microheterogeneity and Host Interactions. *Mol Cell Proteomics* **2016**, *15* (6), 1895-1912.
30. Blake, Thomas A.; Williams, Tracie L.; Pirkle, James J.; Barr, John R., Targeted N-Linked Glycosylation Analysis of H5N1 Influenza Hemagglutinin by Selective Sample Preparation and Liquid Chromatography/Tandem Mass Spectrometry. *Anal Chem* **2009**, *81* (8), 3109-3118.
31. Schwarzer, Jana; Rapp, Erdmann; Hennig, René; Genzel, Yvonne; Jordan, Ingo; Sandig, Volker; Reichl, Udo, Glycan analysis in cell culture-based influenza vaccine production: Influence of host cell line and virus strain on the glycosylation pattern of viral hemagglutinin. *Vaccine* **2009**, *27* (32).
32. An, Yanming; Rininger, Joseph A.; Jarvis, Donald L.; Jing, Xianghong; Ye, Zhiping; Aumiller, Jared J.; Eichelberger, Maryna; Cipollo, John F., Comparative Glycomics Analysis of Influenza Hemagglutinin (H5N1) Produced in Vaccine Relevant Cell Platforms. *J Proteome Res* **2013**, *12*, 3707-3720.
33. Parsons, Lisa M.; An, Yanming; Vries, Robert P. de; Haan, Cornelis A. M. de; Cipollo, John F., Glycosylation Characterization of an Influenza H5N7 Hemagglutinin Series with Engineered Glycosylation Patterns: Implications for Structure-Function Relationships. *J Proteome Res* **2016**, *16* (2), 398-412.
34. Liu, Yan-Jun; Wu, Shiaw-Lin; Love, Kerry R.; Hancock, William S., Characterization of Site-Specific Glycosylation in Influenza A Virus Hemagglutinin Produced by Spodoptera frugiperda Insect Cell Line. *Anal Chem* **2017**, *89* (20), 11036-11043.
35. Zhang, Sheng; Sherwood, Robert W.; Yang, Yong; Fish, Tara; Chen, Wei; McCardle, James A.; Jones, R. Mark; Yusibov, Vidadi; May, Eliel Ruiz; Rose, Jocelyn K. C.; Thannhauser, Theodore W., Comparative characterization of the glycosylation profiles of an influenza hemagglutinin produced in plant and insect hosts. *Proteomics* **2012**, *12*, 1269-1288.
36. Blackburne, Benjamin P.; Hay, Alan J.; Goldstein, Richard A., Changing Selective Pressure during Antigenic Changes in Human Influenza H3. *PLoS Pathogens* **2008**, *4* (5), 1-6.

37. Altman, Meghan O.; Angel, Matthew; Košík, Ivan; Trovão, Nidia S.; Zost, Seth J.; Gibbs, James S.; Casalino, Lorenzo; Amaro, Rommie E.; Hensely, Scott E.; Nelson, Martha I.; Yewdell, Jonathan W., Human Influenza A Virus Hemagglutinin Glycan Evolution Follows a Temporal Pattern to a Glycan Limit. *mBio* **2019**, *10* (2).
38. Sun, Shisheng; Wang, Qinzhe; Zhao, Fei; Chen, Wentian; Li, Zheng, Glycosylation Site Alteration in the Evolution of Influenza A (H1N1) Viruses. *PLoS One* **2011**, *6* (7).
39. Gao, Jin; Couzens, Laura; Burke, David F.; Wan, Hongquan; Wilson, Patrick; Memoli, Matthew J.; Xu, Xiyang; Harvey, Ruth; Wrammert, Jens; Ahmed, Rafi; Taubenberger, Jeffery K.; Smith, Derek J.; Fouchier, Ron A. M.; Eichelberger, Maryna C., Antigenic Drift of the Influenza A(H1N1)pdm09 Virus Neuraminidase Results in Reduced Effectiveness of A/California/7/2009 (H1N1pdm09)-Specific Antibodies. *mBio* **2019**, *10* (2).
40. Jayaraman, Akila; Koh, Xiaoying; Li, Jing; Raman, Rahul; Viswanathan, Karthik; Shriver, Zachary; Sasisekharan, Ram, Glycosylation at Asn91 of H1N1 haemagglutinin affects binding to glycan receptors. *Biochem J* **2012**, *444* (3), 429-435.
41. Stevens, James; Blixt, Ola; Chen, Li-Mei; Donis, Ruben O.; Paulson, James C.; Wilson, Ian A., Recent Avian H5N1 Viruses Exhibit Increased Propensity for Acquiring Human Receptor Specificity. *J Mol Biol* **2008**, *381* (5), 1382-1394.
42. Wang, Cheng-Chi; Chen, Juine-Ruey; Tseng, Yung-Chieh; Tsu, Che-Hsiung; Hung, Yu-Fu; Chen, Shih-Wei; Chen, Chin-Mei; Khoo, Kay-Hooi; Cheng, Ting-Jen; Cheng, Yih-Shyun E.; Jan, Jia-Tsong; Wu, Chung-Yi; Ma, Che; Wong, Chi-Huey, Glycans on influenza hemagglutinin affect receptor binding and immune response. *Proceedings of the National Academy of Sciences USA* **2009**, *106* (43), 18137-18142.
43. Kasson, Peter M.; Pande, Vijay S., Structural Basis for Influence of Viral Glycans on Ligand Binding by Influenza Hemagglutinin. *Biophys J* **2008**, *95* (7), L48-L50.
44. Sun, Xiangjie; Jayaraman, Akila; ManiPrasad, Pavithra; Raman, Rahul; Houser, Katherine V.; Pappas, Claudia; Zeng, Hui; Sasisekharan, Ram; Katz, Jacqueline M.; Tumpey, Terrence M., N-Linked Glycosylation of the Hemagglutinin Protein Influences Virulence and Antigenicity of the 1918 Pandemic and Seasonal H1N1 Influenza A Viruses. *J Virol* **2013**, *87* (15), 8756-8766.
45. Xu, Qi; Wang, Weijia; Cheng, Xing; Zengel, James; Jin, Hong, Influenza H1N1 A/Solomon Island/3/06 Virus Receptor Binding Specificity Correlates with Virus Pathogenicity, Antigenicity, and Immunogenicity in Ferrets. *J Virol* **2010**, *84* (10), 4936-4945.
46. Owen, Rachel E.; Yamada, Eriko; Thompson, Catherine I.; Phillipson, Louisa J.; Thompson, Clare; Taylor, Elizabeth; Zambon, Maria; Osborn, Helen M. I.; Barclay, Wendy S.; Borrow, Persephone, Alterations in Receptor Binding Properties of Recent Human Influenza H3N2 Viruses Are Associated with Reduced Natural Killer Cell Lysis of Infected Cells. *J Virol* **2007**, *81* (20), 11170-11178.
47. Kasson, Peter, Receptor Binding by Influenza Virus: Using Computational Techniques To Extend Structural Data. *Biochemistry* **2012**, *51* (12), 2359-2365.
48. Elcock, Adrian H., Molecular Simulations of Diffusion and Association in Multimacromolecular Systems. *Methods Enzymol* **2004**, *383*, 166-198.
49. Gabdoulline, Razif R.; Wade, Rebecca C., Bimolecular diffusion association. *Curr Opin Struct Biol* **2002**, *12* (2), 204-213.
50. Ermak, Donald L.; McCammon, J. Andrew, Brownian dynamics with hydrodynamic interactions. *Journal of Chemical Physics* **1978**, *69* (1352).
51. Huber, Gary A.; McCammon, J. Andrew, Brownian Dynamics Simulations of Biological Molecules. *Trends in Chemistry* **2019**, *1* (8), 727-738.
52. Sung, Jeffrey C.; Wynsberghe, Adam W. Van; Amaro, Rommie E.; Li, Wilfred W.; McCammon, J. Andrew, Role of Secondary Sialic Acid Binding Sites in Influenza N1 Neuraminidase. *J Am Chem Soc* **2010**, *132* (9), 2883-2885.
53. Amaro, Rommie E.; Jeong, Pek U.; Huber, Gary; Dommer, Abigail; Steven, Alasdair C.; Bush, Robin M.; Durrant, Jacob D.; Votapka, Lane W., A Computational Assay that Explores the Hemagglutinin/Neuraminidase Functional Balance Reveals the Neuraminidase Secondary Site as a Novel Anti-Influenza Target. *ACS Central Science* **2018**, *4* (11), 1570-1577.
54. Zeller, Fabian; Luitz, Manuel P.; Bomblies, Rainer; Zacharias, Martin, Multiscale Simulation of Receptor–Drug Association Kinetics: Application to Neuraminidase Inhibitors. *Journal of Chemical Theory and Computation* **2017**, *13* (10), 5097-5105.

55. Noskov, Sergey Yu.; Lim, Carmay, Free Energy Decomposition of Protein-Protein Interactions. *Biophys J* **2001**, *81* (2), 737-750.
56. Russell, Rupert J.; Haire, Lesley F.; Stevens, David J.; Collins, Patrick J.; Lin, Yi Pu; Blackburn, G. Michael; Hay, Alan J.; Gamblin, Steven J.; Skehel, John J., The Structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design. *Nature* **2006**, *443*, 45-49.
57. Chong, Andrew K. J.; Pegg, Michael S.; Itzstein, Mark von, Influenza Virus Sialidase: Effect of Calcium on Steady-State Kinetic Parameters. *Biochim Biophys Acta* **1991**, *1077* (1), 65-71.
58. Schrödinger, LLC *The PyMOL Molecular Graphics System, Version 1.8*, 2015.
59. Schrödinger, LLC *Maestro*, New York, NY, 2020.
60. Collins, Patrick J.; Haire, Lesley F.; Lin, Yi Pu; Liu, Junfeng; Russell, Rupert J.; Walker, Phillip A.; Skehel, John J.; Martin, Stephen R.; Hay, Alan J.; Gamblin, Steven J., Crystal structures of oseltamivir-resistant influenza virus neuraminidase mutants. *Nature* **2008**, *453*.
61. Vries, Erhard van der; Collins, Patrick J.; Vachieri, Sebastien G.; Xiong, Xiaoli; Liu, Junfeng; Walker, Philip A.; Haire, Lesley F.; Hay, Alan J.; Schutten, Martin; Osterhaus, Albert D. M. E.; Martin, Steve R.; Boucher, Charles A. B.; Skehel, John J.; Gamblin, Steve J., H1N1 2009 Pandemic Influenza Virus: Resistance of the I223R Neuraminidase Mutant Explained by Kinetic and Structural Analysis. *PLoS Pathogens* **2012**, *8* (9), 1-8.
62. Olsson, Mats H. M.; Søndergaard, Chresten R.; Rostkowski, Michal; Jensen, Jan H., PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *Journal of Chemical Theory and Computation* **2011**, *7* (2), 525-537.
63. Dolinsky, Todd J.; Nielsen, Jens E.; McCammon, J. Andrew; Baker, Nathan A., PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations *Nucleic Acids Res* **2004**, *32* (2), W665-W667.
64. Wang, Junmei; Cieplak, Piotr; Kollman, Peter A., How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry* **2000**, *21*, 1049-1074.
65. Kirschner, Karl N.; Yongye, Austin B.; Tschampel, Sarah M.; González-Outeiriño, Jorge; Daniels, Charlisa R.; Foley, B. Lachele; Woods, Robert J., GLYCAM06: A generalizable biomolecular force field. Carbohydrates. *Journal of Computational Chemistry* **2008**, *29* (4), 622-655.
66. Böhm, Michael; Böhne-Lang, Andreas; Frank, Martin; Loss, Alexander; Rojas-Macias, Miguel A; Lütkeke, Thomas, Glycosciences.DB: an annotated data collection linking glycomics and proteomics data (2018 update). *Nucleic Acids Res* **2019**, *47* (D1), D1195-D1201.
67. Zwier, Matthew C.; Chong, Lilian T., Reaching biological timescales with all-atom molecular dynamics simulations. *Curr Opin Pharmacol* **2010**, *10*, 745-752.
68. Taha, Hashem A.; Richards, Michele R.; Lowary, Todd L., Conformational Analysis of Furanoside-Containing Mono- and Oligosaccharides. *Chemical Reviews* **2013**, *113*, 1851-1876.
69. Jo, Sunhwan; Kim, Taehoon; Iyer, Vidyashankara G.; Im, Wonpil, CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry* **2008**, *29* (11), 1859-1865.
70. Brooks, Bernard R.; III, Charles L. Brooks; Jr., Alexander D. Mackerell; Nilsson, Lennart; Petrella, Robert J.; Roux, Benoît; Won, Youngdo; Archontis, Giorgios; Bartels, Christian; Boresch, Stefan; Caffisch, Amadeo; Caves, Leo; Cui, Qiang; Dinner, Aaron R.; Feig, Michael; Fischer, Stefan; Gao, Jiali; Hodošček, Milan; Im, Wonpil; Kuczera, Krzysztof; Lazaridis, Themis; Ma, Jianpeng; Ovchinnikov, Victor; Paci, Emanuele; Pastor, Richard W.; Post, Carol Beth; Pu, Jing Zhi; Schaefer, Michael; Tidor, Bruce; Venable, Richard M.; Woodcock, H. Lee; Wu, Xiongwu; Yang, Wei; York, Darrin M.; Karplus, Martin, CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry* **2009**, *30* (10), 1545-1614.
71. Lee, Jumin; Cheng, Xi; Swails, Jason M.; Yeom, Min Sun; Eastman, Peter K.; Lemkul, Justin A.; Wei, Shuai; Buckner, Joshua; Jeong, Jong Cheol; Qi, Yifei; Jo, Sunhwan; Pande, Vijay S.; Case, David A.; III, Charles L. Brooks; Jr., Alexander D. MacKerell; Klauda, Jeffery B.; Im, Wonpil, CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *Journal of Chemical Theory and Computation* **2016**, *12* (1), 405-413.
72. Jo, Sunhwan; Song, Kevin C.; Desaire, Heather; Jr., Alexander D. MacKerell; Im, Wonpil, Glycan reader: Automated sugar identification and simulation preparation for carbohydrates and glycoproteins. *Journal of Computational Chemistry* **2011**, *32* (14), 3135-3141.

73. Park, Sang-Jun; Lee, Jumin; Patel, Dhilon S.; Ma, Hongjing; Lee, Hui Sun; Jo, Sunhwan; Im, Wonpil, Glycan Reader is improved to recognize most sugar types and chemical modifications in the Protein Data Bank. *Bioinformatics* **2017**, 33 (19), 3051-3057.
74. Park, Sang-Jun; Lee, Jumin; Qi, Yifei; Kern, Nathan R; Lee, Hui Sun; Jo, Sunhwan; Joung, In Suk; Joo, Keehyung; Lee, Jooyoung; Im, Wonpil, CHARMM-GUI Glycan Modeler for modeling and simulation of carbohydrates and glycoconjugates. *Glycobiology* **2019**, 29 (4), 320-331.
75. Jorgensen, William L.; Chandrasekhar, Jayaraman; Madura, Jeffry D., Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics* **1983**, 79 (2), 926-935.
76. Beglov, Dmitrii; Roux, Benoît, Finite representation of an infinite bulk system: Solvent boundary potential for computer simulations. *Journal of Chemical Physics* **1994**, 100 (12), 9050-9063.
77. Best, Robert B.; Zhu, Xiao; Shim, Jihyun; Lopes, Pedro E. M.; Mittal, Jeetain; Feig, Michael; Alexander D. MacKerell, Jr., Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ , ψ and Side-Chain χ_1 and χ_2 Dihedral Angles. *Journal of Chemical Theory and Computation* **2012**, 8 (9), 3257-3273.
78. Salomon-Ferrer, Romelia; Götz, Andreas W.; Poole, Duncan; LeGrand, Scott; Walker, Ross C., Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *Journal of Chemical Theory and Computation* **2013**, 9 (9), 3878-3888.
79. Case, David A.; Ben-Shalom, Ido Y.; Brozell, Scott R.; Cerutti, David S.; Thomas E. Cheatham, III; Cruzeiro, Vinícius Wilian D.; Darden, Tom A.; Duke, Robert E.; Ghoreishi, Delaram; Gilson, Michael K.; Gohlke, Holger; Götz, Andreas W.; Greene, D'Artagnan; Harris, Robert; Homeyer, Nadine; Izadi, Saeed; Kovalenko, Andriy; Kurtzman, Tom; Lee, Tai Sung; LeGrand, Scott; Li, Pengfei; Lin, Charles; Liu, Jian; Luchko, Tyler; Luo, Ray; Mermelstein, Daniel J.; Merz, Kenneth M.; Miao, Yinglong; Monard, Gérald; Nguyen, Crystal; Nguyen, Hai; Omelyan, Igot; Onufriev, Alexey; Pan, Feng; Qi, Ruxi; Roe, Daniel R.; Roitberg, Adrian; Sagui, Celeste; Schott-Verdugo, Stephan; Shen, Jana; Simmerling, Carlos L.; Smith, Jamie; Salomon-Ferrer, Romelia; Swails, Jason; Walker, Ross C.; Wang, Junmei; Wei, Haixin; Wolf, Romain M.; Wu, Xiongwu; Xiao, Li; York, Darrin M.; Kollman, Peter A., AMBER 2018. **2018**.
80. Ryckaert, Jean-Paul; Ciccotti, Giovanni; Berendsen, Herman J. C., Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamcis of n-Alkanes. *Journal of Computational Physics* **1977**, 23, 327-341.
81. Loncharich, Richard J.; Brooks, Bernard R.; Pastor, Richard W., Langevin dynamics of peptides: The frictional dependence of isomerization rates of N -acetylalanyl-N '-methylamide. *Biopolymers* **1992**, 32, 523-535.
82. Pastor, Richard W.; Brooks, Bernard R.; Szabo, Attila, An analysis of the accuracy of Langevin and molecular dynamics algorithms. *Molecular Physics* **1988**, 65 (6), 1409-1419.
83. Chow, Kim-Hung; Ferguson, David M., Isothermal-isobaric molecular dynamics simulations with Monte Carlo volume sampling. *Computer Physics Communications* **1995**, 91, 283-289.
84. Towns, John; Cockerill, Timothy; Dahan, Maytal; Foster, Ian; Gaither, Kelly; Grimshaw, Andrew; Hazlewood, Victor; Lathrop, Scott; Lifka, Dave; Peterson, Gregory D.; Roskies, Ralph; Scott, J. Ray; Wilkins-Diehr, Nancy, XSEDE: Accelerating Scientific Discovery. *Computing in Science & Engineering* **September-October 2014**, 16, 62-74.
85. Darden, Tom; York, Darrin M.; Pedersen, Lee G., Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *Journal of Chemical Physics* **1993**, 98 (12), 10089-10092.
86. Abraham, Mark James; Murtola, Teemu; Schulz, Roland; Páll, Szilárd; Smith, Jeremy C.; Hess, Berk; Lindahl, Erik, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, 1-2, 19-25.
87. Xu, Dong; Newhouse, E. Irene; Amaro, Rommie E.; Pao, Hsing C.; Cheng, Lily S.; Markwick, Phineus R.L.; McCammon, J. Andrew; Li, Wilfred W.; Arzberger, Peter W., Distinct Glycan Topology for Avian and Human Sialopentasaccharide Receptor Analogues upon Binding Different Hemagglutinins: A Molecular Dynamics Perspective. *J Mol Biol* **2009**, 387 (2), 465-491.
88. Varghese, Joseph N.; Colman, Peter M.; Donkelaar, Albertus van; Blick, Tony J.; Sahasrabudhe, Anjali; McKimm-Breschkin, Jennifer L., Structural evidence for a second sialic acid binding site in avian influenza neuraminidases. *Proceedings of the National Academy of Sciences USA* **1997**, 94, 11808-11812.
89. Schüttelkopf, Alexander W.; Aalten, Daan M. F. van, PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallographica Section D* **2004**, 60 (8), 1355-1363.

90. Huber, Gary A.; McCammon, J. Andrew, Browndye: A Software Package for Brownian Dynamics. *Computer Physics Communications* **2010**, *181* (11), 1896-1905.
91. Guldberg, Cato M.; Waage, Peter, Studier i affiniteten. *CM Forhandlinger: Videnskabs-Selskabet i Christiana* **1864**, 35.
92. Guldberg, Cato M.; Waage, Peter, Über die chemische Affinität. *Journal für Praktische Chemie* **1879**, *19* (69).
93. Åqvist, Johan, Ion-water interaction potentials derived from free energy perturbation simulations. *Journal of Physical Chemistry* **1990**, *94*, 8021-8024.
94. Humphrey, William; Dalke, Andrew; Schulten, Klaus, VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **1996**, *14* (1), 33-38.
95. Jurrus, Elizabeth; Engel, Dave; Star, Keith; Monson, Kyle; Brandi, Juan; Felberg, Lisa E.; Brookes, David H.; Wilson, Leighton; Chen, Jiahui; Liles, Karina; Chun, Minju; Li, Peter; Gohara, David W.; Dolinsky, Todd; Konecny, Robert; Koes, David R.; Nielsen, Jens Erik; Head-Gordon, Teresa; Geng, Weihua; Krasny, Robert; Wei, Guo-Wei; Holst, Michael J.; McCammon, J. Andrew; Baker, Nathan A., Improvements to the APBS biomolecular solvation software suite. *Protein Sci* **2017**, *27*, 112-128.
96. Sun, Xiaoman; Li, Qing; Wu, Yan; Wang, Mingyang; Liu, Yue; Qi, Jianxun; Vavricka, Christopher J.; Gao, George F., Structure of Influenza Virus N7: the Last Piece of the Neuraminidase "Jigsaw" Puzzle. *J Virol* **2014**, *88* (16), 9197-9207.
97. Lai, Jimmy C. C.; Garcia, Jean-Michel; Dyason, Jeffrey C.; Böhm, Raphael; Madge, Paul D.; Rose, Faith J.; Nicholls, John M.; Peiris, J. S. Malik; Haselhorst, Thomas; Itzstein, Mark von, A Secondary Sialic Acid Binding Site on Influenza Virus Neuraminidase: Fact or Fiction? *Angewandte Chemie International Edition* **2012**, *51* (9), 2221-2224.
98. Benton, Donald J.; Wharton, Stephen A.; Martin, Stephen R.; McCauley, John W., Role of Neuraminidase in Influenza A(H7N9) Virus Receptor Binding. *J Virol* **2017**, *91* (11).
99. Wei, Xiping; Decker, Julie M.; Wang, Shuyi; Hui, Huxiong; Kappes, John C.; Wu, Xiaoyun; Salazar-Gonzalez, Jesus F.; Salazar, Maria G.; Kilby, J. Michael; Saag, Michael S.; Komarova, Natalia L.; Nowak, Martin A.; Hahn, Beatrice H.; Kwong, Peter D.; Shaw, George M., Antibody neutralization and escape by HIV-1. *Nature* **2003**, *422*, 307-312.
100. Machiels, Bénédicte; Lété, Céline; Guillaume, Antoine; Mast, Jan; Stevenson, Philip G.; Vanderplasschen, Alain; Gillet, Laurent, Antibody Evasion by a Gammaherpesvirus O-Glycan Shield. *PLoS Pathogens* **2011**, *7* (11).
101. Helle, François; Duverlie, Gilles; Dubuisson, Jean, The Hepatitis C Virus Glycan Shield and Evasion of the Humoral Immune Response. *Viruses* **2011**, *3* (10), 1909-1932.
102. Casalino, Lorenzo; Gaieb, Zied; Dommer, Abigail C.; Harbison, Aoife M.; Fogarty, Carl A.; Barros, Emilia P.; Taylor, Bryn C.; Fadda, Elisa; Amaro, Rommie E., Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Central Science* **2020**.
103. Kwong, Peter D.; Wyatt, Richard; Robinson, James; Sweet, Raymond W.; Sodroski, Joseph; Hendrickson, Wayne A., Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature* **1998**, *393*, 648-659.
104. Wyatt, Richard; Sodroski, Joseph, The HIV-1 Envelope Glycoproteins: Fusogens, Antigens, and Immunogens. *Science* **1998**, *280* (5371), 1884-1888.
105. Kumar, Swatantra; Maurya, Vimal K.; Prasad, Anil K.; Bhatt, Madan L. B.; Saxena, Shailendra K., Structural, glycosylation and antigenic variation between 2019 novel coronavirus (2019-nCoV) and SARS coronavirus (SARS-CoV). *Virusdisease* **2020**, *31* (1), 13-21.
106. Nishima, Wataru; Miyashita, Naoyuki; Yamaguchi, Yoshiki; Sugita, Yuji; Re, Suyong, Effect of Bisecting GlcNAc and Core Fucosylation on Conformational Properties of Biantennary Complex-Type N-Glycans in Solution. *Journal of Physical Chemistry B* **2012**, *116* (29), 8504-8512.
107. Landström, Jens; Widmalm, Göran, Glycan flexibility: insights into nanosecond dynamics from a microsecond molecular dynamics simulation explaining an unusual nuclear Overhauser effect. *Carbohydr Res* **2010**, *345* (2), 330-333.
108. Re, Suyong; Nishima, Wataru; Miyashita, Naoyuki; Sugita, Yuji, Conformational flexibility of N-glycans in solution studied by REMD simulations. *Biophysical Reviews* **2012**, *4*, 179-187.
109. Harbison, Aoife M.; Brosnan, Lorna P.; Fenlon, Keith; Fadda, Elisa, Sequence-to-structure dependence of isolated IgG Fc complex biantennary N-glycans: a molecular dynamics study. *Glycobiology* **2019**, *29* (1), 94-103.

110. Ohuchi, Masanobu; Ohuchi, Reiko; Feldmann, Anke; Klenk, Hans-Dieter, Regulation of receptor binding affinity of influenza virus hemagglutinin by its carbohydrate moiety. *J Virol* **1997**, 71 (11), 8377-8384.
111. Du, Wenjuan; Guo, Hongbo; Nijman, Vera S.; Doedt, Jennifer; Vries, Erhard van der; Lee, Joline van der; Li, Zeshi; Boons, Geert-Jan; Kuppeveld, Frank J. M. van; Vries, Erik de; Matrosovich, Mikhail; Haan, Cornelis A. M. de, The 2nd sialic acid-binding site of influenza A virus neuraminidase is an important determinant of the hemagglutinin-neuraminidase-receptor balance. *PLoS Pathogens* **2019**, 15 (6).
112. Air, Gillian M., Influenza virus–glycan interactions. *Current Opinion in Virology* **2014**, 7, 128-133.
113. Uhlenendorff, Jennifer; Matrosovich, Tatyana; Klenk, Hans-Dieter; Matrosovich, Mikhail, Functional significance of the hemadsorption activity of influenza virus neuraminidase and its alteration in pandemic viruses. *Arch Virol* **2009**, 154 (6), 945-957.
114. Dai, Meiling; McBride, Ryan; Dortmans, Jos C. F. M.; Peng, Wenjie; Bakkers, Mark J. G.; Groot, Raoul J. de; Kuppeveld, Frank J. M. van; Paulson, James C.; Vries, Erik de; Haan, Cornelis A. M. de, Mutation of the Second Sialic Acid-Binding Site, Resulting in Reduced Neuraminidase Activity, Preceded the Emergence of H7N9 Influenza A Virus. *J Virol* **2017**, 91 (9).
115. Durrant, Jacob D.; Kochanek, Sarah E.; Casalino, Lorenzo; Jeong, Pek U.; Dommer, Abigail C.; Amaro, Rommie E., Mesoscale All-Atom Influenza Virus Simulations Suggest New Substrate Binding Mechanism. *ACS Central Science* **2020**, 6 (2), 189-196.
116. Du, Wenjuan; Dai, Meiling; Li, Zeshi; Boons, Geert-Jan; Peeters, Ben; Kuppeveld, Frank J. M. van; Vries, Erik de; Haan, Cornelis A. M. de, Substrate Binding by the Second Sialic Acid-Binding Site of Influenza A Virus N1 Neuraminidase Contributes to Enzymatic Activity. *J Virol* **2018**, 92 (20).
117. Hausmann, Jürgen; Kretzschmar, Evelyn; Garten, Wolfgang; Klenk, Hans-Dieter, N1 neuraminidase of influenza virus A/FPV/Rostock/34 has haemadsorbing activity. *J Gen Virol* **1995**, 76, 1719-1728.
118. Kobasa, Darwyn; Rodgers, M. Elizabeth; Wells, Krisna; Kawaoka, Yoshihiro, Neuraminidase hemadsorption activity, conserved in avian influenza A viruses, does not influence viral replication in ducks. *J Virol* **1997**, 71 (9).
119. Webster, Robert G.; Air, Gillian M.; Metzger, D. W.; Colman, Peter M.; Varghese, Joseph N.; Baker, Ashley T.; Laver, W. Graeme, Antigenic structure and variation in an Influenza Virus N9 Neuraminidase. *J Virol* **1987**, 61 (9), 2910-2916.
120. Honda, Takeshi; Yoshida, Shuku; Arai, Masami; Masuda, Takeshi; Yamashita, Makoto, Synthesis and anti-Influenza evaluation of polyvalent sialidase inhibitors bearing 4-guanidino-Neu5Ac2en derivatives. *Bioorg Med Chem Lett* **2002**, 12 (15), 1929-1932.
121. Macdonald, Simon J. F.; Cameron, Rachel; Demaine, Derek A.; Fenton, Rob J.; Foster, Graham; Gower, David; Hamblin, J. Nicole; Hamilton, Stephanie; Hart, Graham J.; Hill, Alan P.; Inglis, Graham G. A.; Jin, Betty; Jones, Haydn T.; McConnell, Darryl B.; McKimm-Breschkin, Jennifer; Mills, Gail; Nguyen, Van; Owens, Ian J.; Parry, Nigel; Shanahan, Stephen E.; Smith, Donna; Watson, Keith G.; Wu, Wen-Yang; Tucker, Simon P., Dimeric Zanamivir Conjugates with Various Linking Groups Are Potent, Long-Lasting Inhibitors of Influenza Neuraminidase Including H5N1 Avian Influenza. *J Med Chem* **2005**, 48 (8), 2964-2971.
122. Weight, Alisha K.; Haldar, Jayanta; Cienfuegos, Luis Álvarezde; Gubareva, Larisa V.; Tumpey, Terrence M.; Chen, Jianzhu; Klivanov, Alexander M., Attaching Zanamivir to a Polymer Markedly Enhances Its Activity Against Drug-resistant Strains of Influenza A Virus. *J Pharm Sci* **2011**, 100 (3), 831-835.
123. Lee, Chia Min; Weight, Alisha K.; Haldar, Jayanta; Wang, Ling; Klivanov, Alexander M.; Chen, Jianzhu, Polymer-attached zanamivir inhibits synergistically both early and late stages of influenza virus infection. *Proceedings of the National Academy of Sciences USA* **2012**, 109 (50), 20385-20390.
124. Yang, Zhao-Liang; Zeng, Xiong-Fei; Liu, Hai-Peng; Yu, Qun; Meng, Xin; Yan, Zhong-Li; Fan, Zhen-Chuan; Xiao, Hai-Xia; Iyer, Suri S.; Yang, Yang; Yu, Peng, Synthesis of multivalent difluorinated zanamivir analogs as potent antiviral inhibitors. *Tetrahedron Lett* **2016**, 57 (24), 2579-2582.
125. Zhao, Teng-Fei; Qin, Hai-Juan; Yu, Yao; Yang, Mei-Bing; Chang, Hao; Guo, Na; He, Yun; Yang, Yang; Yu, Peng, Multivalent zanamivir-bovine serum albumin conjugate as a potent influenza neuraminidase inhibitor. *J Carbohydr Chem* **2017**, 36 (4-6), 235-246.

Chapter 4

Breathing and tilting: mesoscale simulations illuminate influenza glycoprotein vulnerabilities

4.1 Abstract

Influenza virus has resurfaced recently from inactivity during the early stages of the COVID-19 pandemic, raising serious concerns about the nature and magnitude of future epidemics. The main antigenic targets of influenza virus are two surface glycoproteins, hemagglutinin (HA) and neuraminidase (NA). Whereas the structural and dynamical properties of both glycoproteins have been studied previously, the understanding of their plasticity in the whole-virion context is fragmented. Here, we investigate the dynamics of influenza glycoproteins in a crowded protein environment through mesoscale all-atom molecular dynamics simulations of two evolutionary-linked glycosylated influenza A whole-virion models. Our simulations reveal and kinetically characterize three main molecular motions of influenza glycoproteins: NA head tilting, HA ectodomain tilting, and HA head breathing. The flexibility of HA and NA highlights antigenically relevant conformational states, as well as facilitates the characterization of a novel monoclonal antibody, derived from human convalescent plasma, that binds to the underside of the NA head. Our work provides previously unappreciated views on the dynamics of HA and NA, advancing the understanding of their interplay and suggesting possible strategies for the design of future vaccines and antivirals against influenza.

4.2 Introduction

Influenza is a single-stranded, negative-sense RNA virus that contains two membrane-embedded glycoproteins: hemagglutinin (HA) and neuraminidase (NA).¹⁻⁴ The viral envelope is also dotted with transmembrane proton channels named M2.^{5,6} HA and NA are involved in a complex interplay whose combined functions include viral entry, progeny release, and evading host immune pressure.⁷⁻⁹ These processes are, in part, modulated by glycans, whose number and position can vary from year to year owing to the relentless evolution of the influenza virus.¹⁰⁻

¹³ As a part of influenza seasonal antigenic drift, the introduced missense mutations may directly alter the

antigenicity of HA¹⁰ and NA¹³ epitopes or sometimes result in the gain or loss of a glycosylation site,¹⁰ thus leading to shielding or unmasking of nearby epitopes. Antigenic drift means that vaccines targeting influenza need to be regularly updated to match the currently predominant strains of the virus.^{14,15} Despite HA being the immunodominant protein,^{16,17} NA also represents a potential vaccine target;¹⁸ knowledge of accessible and currently utilized epitopes is crucial in retaining influenza vaccine efficacy and in designing new vaccines. This is particularly relevant as the scientific community marches toward a universal influenza vaccine that is not affected by antigenic drift.^{16,17,19}

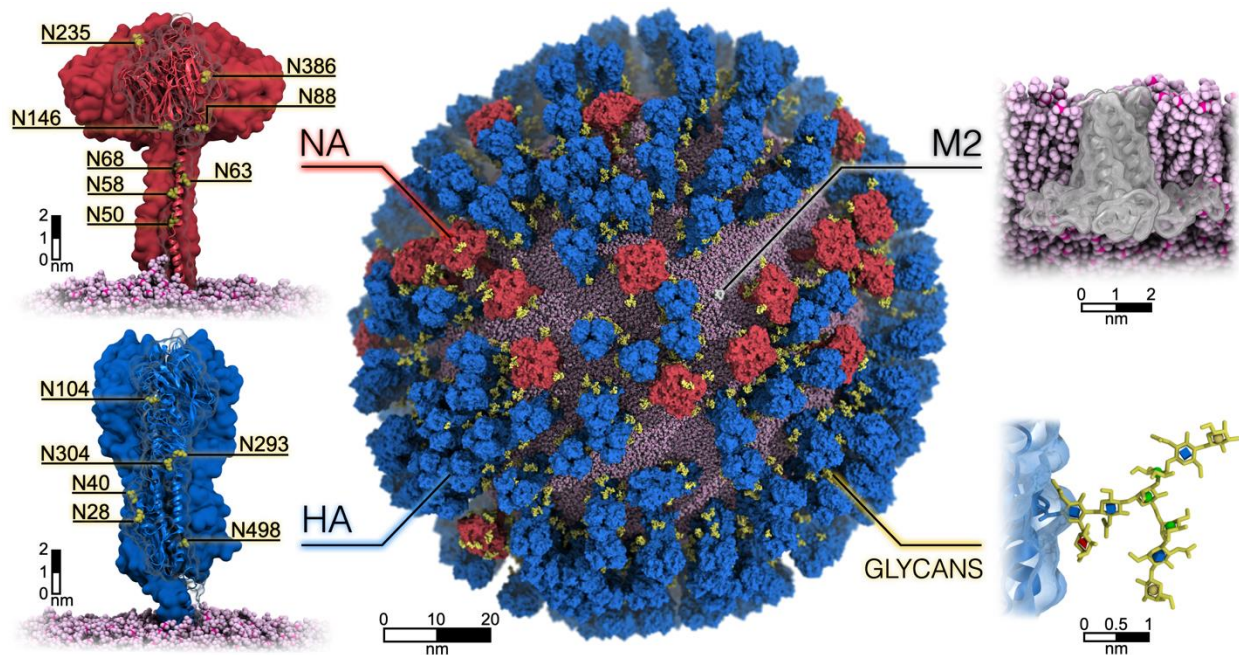
Antigenic drift also affects the HA/NA interplay, which in turn is relevant in the context of influenza endocytosis²⁰ and exocytosis (or viral budding).^{21–23} The understanding of these processes is far from complete. As just one example, it was previously thought that NA played no role in endocytosis or exocytosis.²⁴ Next, it was assumed that HA was involved in receptor binding and was not essential for viral budding,²⁵ whereas NA was only involved in exocytosis by cleaving terminal host cell sialic acid moieties hydrogen bound to HA. More recently, it has been recognized that the role these two proteins play in cell entry and exit is more nuanced. NA can play a role in receptor binding^{26–28} while HA can initiate viral budding.²¹ This egress process appears to be the rate-limiting step in viral replication, with only ~10% of budded viral particles being fully released,²² and is the major target of current antivirals. Considering the weak binding capabilities of both HA and NA, multivalent interactions are needed for both cell entry and exit.^{26,29–32} Although the exact valency for influenza cell entry and exit is still not known,³⁰ sialic acid has been determined to be the cell receptor recruitment group involved in these multivalent interactions.^{33–36}

One potential reason for our limited understanding of the specifics of HA and NA function may be their poorly understood native dynamics *in vivo*. Whole virion protein dynamics are extremely challenging to study experimentally; current techniques such as surface affixation or aggregated kinetics assays are difficult to interpret due to their non-native environment or virion pleiomorphy, respectively. Cryo-electron tomography (cryoET) images can capture native virions, but protein flexibility diminishes the resolution of the resulting images. Instead, cryo-electron microscopy (cryoEM) has been successful in characterizing the plastic behavior of the individual glycoprotein ectodomains, showing, for example, the existence of tilted HA conformations,³⁷ or “breathing” HA heads^{38–43} and open NA heads.⁴⁴ Some of the above-mentioned difficulties can be allayed

through computational studies, which can create specific, pre-defined mesoscale models and examine protein dynamics in the mesoscale.^{45–49} Whole-virion scale simulations, whether coarse-grained^{50–52} or all-atom,^{45,53–62} can provide a detailed look at protein dynamics in a more realistic, crowded protein environment than simulations of discrete proteins. Additionally, such large simulations naturally afford statistical sampling for proteins of interest, i.e., an influenza virion will contain a few hundred HA proteins and a few dozen NA proteins, enabling the application of statistical analysis techniques, such as Markov state models (MSM), to quantify conformational transitions.^{45,63–65}

Building on our previous work,⁴⁵ we created two glycosylated whole-virion models accounting for two evolutionary-linked influenza A viruses, namely, A/swine/Shandong/N1/2009 (H1N1) (**Fig. 4.1**), hereafter referred to as H1N1-Shan2009, and A/45/Michigan/2015 (H1N1), hereafter referred to as H1N1-Mich2015 (**Fig. 4.8**). We then performed all-atom MD simulations of these massive systems—tallying 161 million atoms including explicit water and ions—to examine the conformational plasticity of influenza glycoproteins in a crowded environment and its impact on their antigenic properties and dynamic interplay.

Figure 4.1 | All-atom glycosylated model of influenza A virion. The center panel shows the all-atom model of the influenza H1N1-Shan2009 virion, where the neuraminidase (NA) and hemagglutinin (HA) glycoproteins, and the M2 ion channels, are depicted with red, blue, and white surfaces, respectively. N-linked glycans are shown in yellow. The spherical lipid envelope is represented with pink van der Waals (vdW) spheres. Magnified views of NA (top left) and HA (bottom left) are provided, where asparagine residues within each N-linked glycosylation sequon are shown with yellow vdW spheres. A cross-section, magnified view of an M2 ion channel is represented in the top-right corner. In the bottom right corner, a magnified view of a glycan is displayed, where the glycan is shown with yellow sticks, and the different monosaccharides are highlighted using 3D-SNFG representation (blue cube for GlcNAc, yellow sphere for galactose, red cone for fucose, and green spheres for mannose). Hydrogen atoms have been hidden for clarity.



As a result, our simulations indicate three main molecular motions underlying the exceeding flexibility of the influenza glycoproteins: NA head tilting, HA ectodomain tilting, and HA breathing. Using a Markov state model analysis framework, we characterize the kinetics of these motions. Concomitant with these three motions, the vulnerabilities of the glycoproteins are uncovered, including a novel epitope located on the underside of the NA globular head. This epitope becomes directly accessible to a human monoclonal antibody, termed NDS.1, when NA head tilting occurs. In addition, we show how cryptic or usually occluded epitopes become transiently accessible during HA breathing and HA ectodomain tilting, respectively. Finally, we characterize the interplay between glycoproteins, breaking down the number and type of inter-glycoprotein contacts and measuring the extent of glycoprotein “clumping” occurring on the virion surface. Although not many differences are observed between H1N1-Shan2009 and H1N1-Mich2015, the loss of glycan N386 within the NA head domain in the H1N1-Mich2015 remarkably reduces NA propensity to engage with neighboring glycoproteins.

Overall, our findings advance the understanding of HA/NA functional balance as affected by the antigenic drift, shedding light on the viral egress process. We propose that the synergy of the three identified motions is critical for the exocytosis cycle. Moreover, our work provides previously unappreciated views into the antigenicity of the influenza glycoproteins by accounting for their dynamics in the crowded environment of the virion surface.

This information suggests that locking HA and NA in transiently sampled vulnerable states can be pursued as a viable strategy for the development of the next flu vaccines and antiviral drugs.

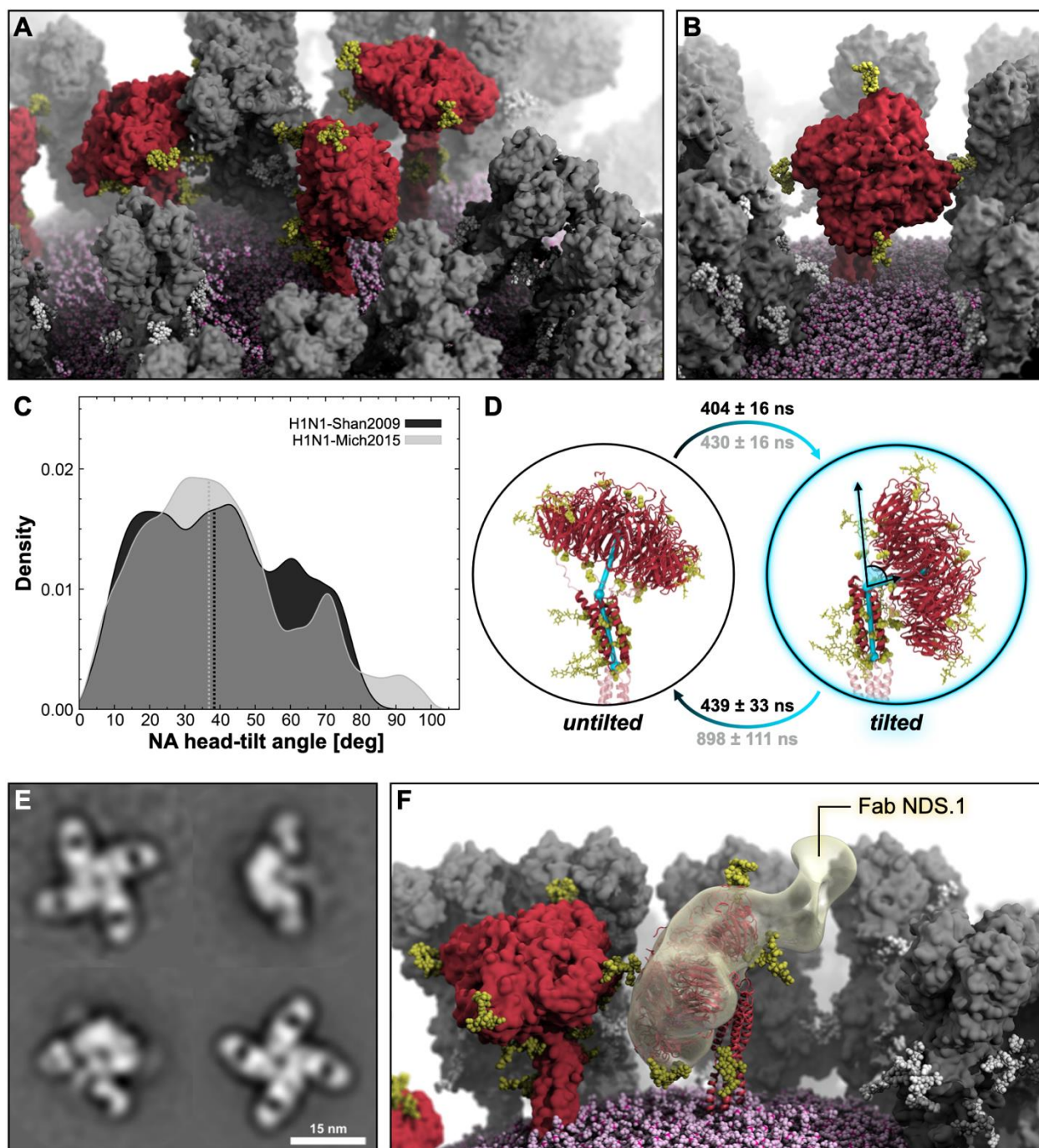
4.3 Results

All-atom MD simulations of H1N1-Shan2009 and H1N1-Mich2015 glycosylated whole-virion models were conducted for 442 and 425 ns, respectively. Our models include 30 NA homotetramers, yielding 13.3 μ s (442 ns $\times$ 30 tetramers for H1N1-Shan2009) and 12.8 μ s (425 ns $\times$ 30 tetramers for H1N1-Mich2015) of tetrameric aggregate sampling, and 236 HA homotrimers, accounting for 104.3 μ s (442 ns $\times$ 236 trimers for H1N1-Shan2009) and 100.3 μ s (425 ns $\times$ 30 trimers for H1N1-Mich2015) of trimeric aggregate sampling. Moreover, 11 M2 ion channels are embedded in the viral envelope. A detailed description of system setup, mutation of H1N1-Shan2009 into H1N1-Mich2015, glycosylation process, and MD simulations are provided in the Supporting Information (SI) Material and Methods section and shown in **Figs. 4.9-4.15**. Analysis of the average Root Mean Square Deviation (RMSD) values calculated for HA and NA evinces that the conformation of both glycoproteins keeps changing during the simulations with respect to the initial geometry (**Fig. 4.16**). More stability is attained only towards the end, where 4-7 Å and 4-11 Å RMSD ranges are observed for HA and NA, respectively. This behavior underlies the presence of dynamic interplay among the glycoproteins during the simulations facilitated by the remarkable flexibility of the glycoprotein ectodomains.

Extensive tilting exposes neuraminidase head underside to immune recognition

The NA globular head sits on top of the stalk and is formed by six β -sheets arranged into a propeller-like structure.⁶⁶ The first β -sheet is connected to the stalk helix by an unstructured, 14-residue-long linkage (residues 82–95) that acts as a hinge point and shows to be highly flexible during the simulations, as evinced by the Root Mean Square Fluctuations (RMSF) analysis (**Fig. 4.17**). In the tetrameric arrangement, the impact of this structural feature is two-fold: uncoupling the motions of the head and the stalk and augmenting the head's conformational freedom. Our whole-virion simulations show that the NA head undergoes an extensive tilt motion, exhibiting extraordinary flexibility (**Fig. 4.2A-B**).

Figure 4.2 | NA head tilting motion. **A)** A snapshot from the MD simulation of the influenza H1N1-Shan2009 virion capturing different NAs (red surface) with varying degrees of head tilt. HAs are represented with a gray surface, whereas N-glycans linked to NA or HA are shown with yellow or white vdW spheres, respectively. The lipid envelope is depicted using pink vdW spheres. **B)** Close-up view of a NA tetramer exhibiting a high degree of head tilt. **C)** Distribution of NA head-tilt angle obtained from the MD simulation of the H1N1-Shan2009 (black) and H1N1-Mich2015 (gray) virions. The distribution is shown as a kernel density. **D)** Two-state MSM of NA head tilting with representative structures from the MD simulation of the H1N1-Shan2009 virion. The mean first-passage times between the states, including the respective standard deviation, are reported rounded to the nearest one for H1N1-Shan2009 (black text) and H1N1-Mich2015 (gray text). NA head-tilt angle is highlighted with cyan cylinders, spheres, and lines drawn within the molecular representation of the NA. NA residues used for angle calculation are shown with opaque red cartoons. N-glycans are represented with yellow sticks, whereas glycosylation sequons are illustrated with yellow vdW spheres. **E)** NS-EM 2D class averages of recombinant A/Darwin/9/2021 N2 in complex with Fab NDS.1 (scale bar: 15 nm) **F)** The density of a single NDS.1 Fab bound to A/Darwin/9/2021 N2 is fitted onto a tilted NA head (red cartoons) as captured in a snapshot from the H1N1-Shan2009 whole-virion simulation. The density is shown with a transparent yellow surface. The surrounding HA and NA are represented in gray and red surfaces, respectively, whereas N-glycans are depicted with yellow vdW spheres.



Remarkably, the NA head can tilt $> 90^\circ$ relative to the stalk axis, as shown in **Fig. 4.2B**. We calculated the tilt angle formed by the stalk's principal axis and the vector joining the top of the stalk with the center of mass (COM) of the tetrameric head for all 30 NA tetramers of the two virion models, at each frame of their respective simulations. Next, we plotted the two sets of angle values as different distributions (**Fig. 4.2C**). The tilt angle is

displayed in **Fig. 4.2D**. Interestingly, despite the differences in amino acid sequences, NA tetramers in H1N1-Shan2009 and H1N1-Mich2015 exhibit a similar extent of head tilt, with median values at 38.6° and 37.2°, respectively. A one-to-one correspondence between the two strains is also observed, with a few exceptions, for the maximum head tilt achieved by each NA tetramer during the simulations (**Fig. 4.18**). Values of head-tilt angle slightly larger than 90° are allowed by the flexibility of the hinge region between the head and the stalk. As the NA tetramer head tilts, at least one of the four monomer heads will move closer to the stalk. Incorporating this geometric criterion into a set of two features (see Material and Methods in the SI, and **Figs. 4.19, 4.20**) and exploiting the large cumulative sampling generated by the many NA tetramers present on the virion surface, we built a two-state MSM to characterize the kinetics of the NA head tilting motion (**Fig. 4.2D**). As a result, for H1N1-Shan2009, the stationary distributions for the untilted and tilted states, which provide an understanding of the weight of each state in the MSM, are 0.51 and 0.49, respectively. The average times required for the transitions between the two states to first occur, referred to as mean first-passage times (MFPTs), are comparable: 404 ± 16 ns for the untilted-to-tilted motion and 439 ± 33 ns for the opposite, tilted-to-untilted transition. Instead, for H1N1-Mich2015, the stationary distributions are 0.44 (untilted) and 0.56 (tilted), with MFPTs of 430 ± 16 ns (untilted-to-tilted) and 898 ± 111 ns (tilted-to-untilted). These findings evince a similar propensity to tilt for the NA heads of both strains, showing that they can tilt $> 75^\circ$ within a few hundred nanoseconds, whereas, in the case of H1N1-Mich2015, it takes longer for them to return to their upright state (see **Table 4.1** for the tilt angles calculated for the macrostates extracted from the MSM).

Aiming to investigate the immunogenic implications of the NA head tilting motion, we studied an H3N2 influenza convalescent donor and discovered a human monoclonal antibody, termed NDS.1, which recognizes the underside of the NA head. NA-specific B cells were isolated from peripheral blood mononuclear cells from the donor using recombinant N2 NA head by flow cytometry (**Fig. 4.21**), while the recombinant antibody NDS.1 was synthesized by using immunoglobulin sequences recovered from the corresponding B cell. Biolayer interferometry (BLI) assays show that Fab of NDS.1 antibody possesses relatively high affinity (10^{-8} M) to two distinct N2 NA tetramers tested, namely A/Wisconsin/672005 and A/Darwin/9/2021 (**Fig. S14**). Negative-stain electron microscopy (NS-EM) analysis and its 3D reconstruction reveal that NDS.1 Fab recognizes an epitope residing on the underside of the NA head (**Fig. 4.2E, Fig. 4.22**). This was further confirmed by analyzing the

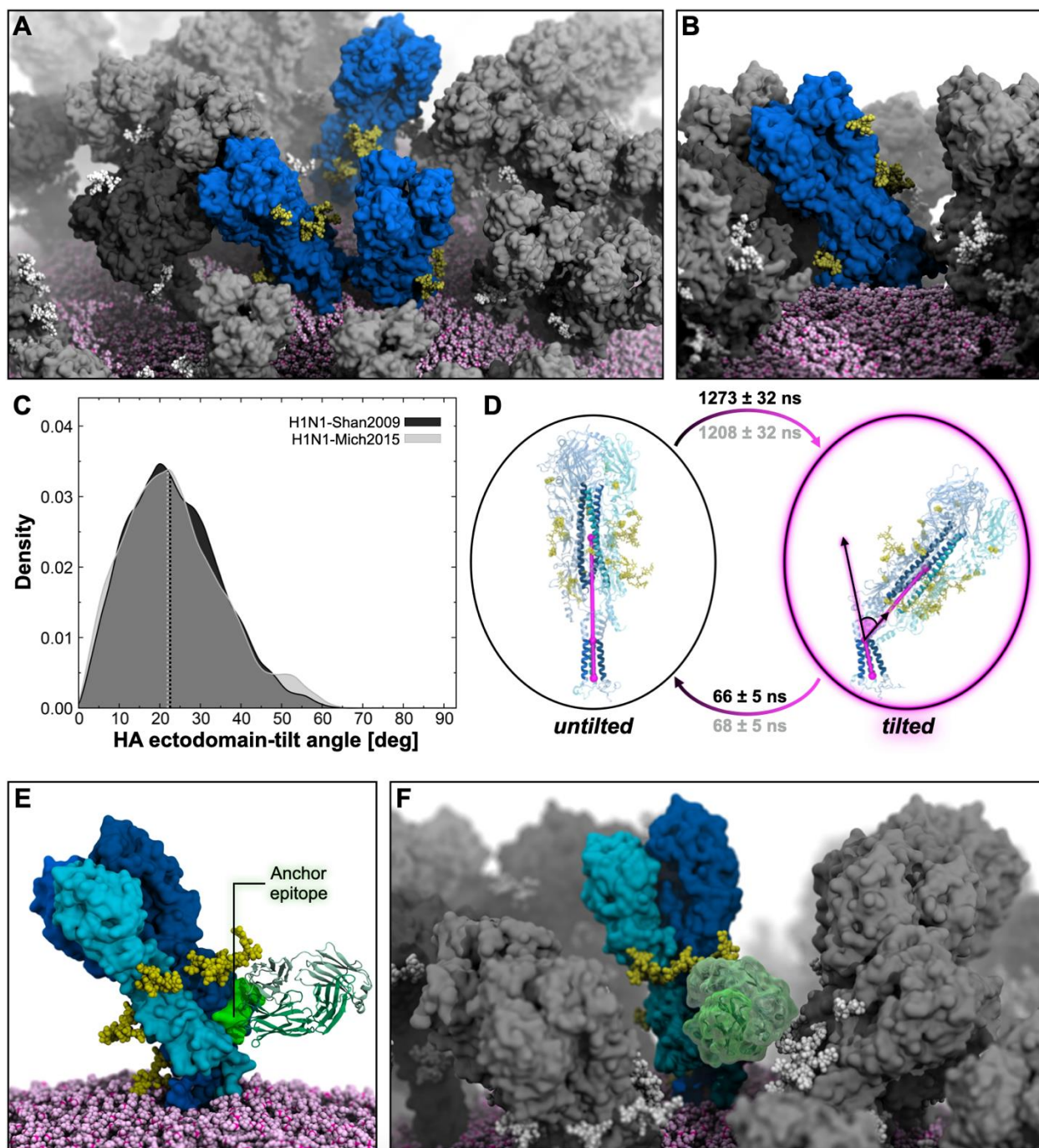
ternary complex of N2-subtype NA, NDS.1 Fab, and 1G01 Fab by NS-EM (**Fig. 4.23**), a recently discovered broadly cross-reactive antibody that binds the catalytic site of NA.⁶⁷ We then fitted the conformation of one tilted NA tetramer head obtained from the H1N1-Shan2009 whole-virion simulation into the density map of the A/Darwin/9/2021 N2 NA tetramer head bound to a single NDS.1 Fab. As shown in **Fig. 4.2F**, the NA head underside region targeted by NDS.1 becomes directly available for immune recognition as the head tilts. Although the NA head underside is solvent accessible in the untilted state, it is not a facile target since the antibody and/or B cell receptors on B cells must reach this region from underneath with a slightly upward approach angle. Instead, the tilt exposes the head underside of at least one monomer and changes the antibody approach angle to downward, thus allowing NDS.1 Fab to gain direct access to the underside epitope. As evinced from the fitted density depicted in **Fig. 4.2F**, the binding of NDS.1 appears sterically compatible with the N-glycans linked to the N1-subtype NA head, i.e., N88, N146, N235, and N386 (for H1N1-Shan2009). A close-up view of the portion of the NA head targeted by NDS.1 is shown in **Fig. 4.24**, where a mock antibody was fitted into the density of NDS.1 to examine a putative binding mode.

Hemagglutinin ectodomain tilting motion facilitates the approach of anchor epitope-directed antibodies

The HA ectodomain protrudes ~15 nm from the viral membrane, to which it is anchored via a coiled-coil transmembrane domain (TMD).³⁷ As also indicated by the average RMSF for HA (**Fig. 4.17**), a flexible hinge region encompassing residues 504–528 connects the HA ectodomain to the TMD, increasing the plasticity of the ectodomain itself.³⁷ Our simulations show considerable flexibility of the HA ectodomain, which exhibits a profuse tilting motion relative to the TMD axis (**Fig. 4.3A** and **Fig. 4.3B**). We measured the tilt angle formed by the TMD principal axis and the vector joining the residues at the top of the TMD and the center-of-mass (COM) of the ectodomain's long α -helices (L α Hs) for all 236 HA trimers from the two virion models, at each frame of their respective simulations. The distributions for the ectodomain-tilt angles are presented in **Fig. 4.3C**, whereas the calculated tilt angle is depicted in **Fig. 4.3D**. The distribution profiles obtained for H1N1-Shan2009 and H1N1-Mich2015 are very similar in shape, with median values of 22.6° and 22.3°, respectively. We note that the calculated tilt angle does not always coincide with the same extent of tilting relative to the axis normal to the membrane plane since the TMD also moves and tilts within the lipid bilayer. Remarkably, a few HA trimers display ectodomain-tilt angle values larger than 50° over the course of the simulations (**Fig. 4.3C**). These values

are in striking agreement with cryoEM experiments examining full-length HA in detergent micelles, where an ectodomain-tilt angle of 52° was reported.³⁷

Figure 4.3 | HA ectodomain tilting motion. **A)** A snapshot from the MD simulation of the influenza H1N1-Shan2009 virion capturing different HAs with varying degrees of ectodomain tilt (blue surface). NAs are represented with a dark gray surface, while other HAs are represented with a gray surface. N-glycans linked to HAs or NAs are shown with yellow or white vdW spheres, respectively. The lipid envelope is depicted using pink vdW spheres. **B)** Close-up view of an HA trimer exhibiting a high degree of ectodomain tilt. **C)** Distribution of HA ectodomain-tilt angle obtained from the MD simulation of the H1N1-Shan2009 (black) and H1N1-Mich2015 (gray) virions. The distribution is shown as a kernel density. **D)** Two-state MSM of HA ectodomain tilting with representative structures from the MD simulation of the H1N1-Shan2009) virion. The mean first-passage times between the states, including the respective standard deviation, are reported rounded to the nearest one for H1N1-Shan2009 (black text) and H1N1-Mich2015 (gray text). HA ectodomain-tilt angle is highlighted with magenta cylinders, spheres, and lines drawn within the molecular representation of the HA. HA residues used for angle calculation are shown with opaque blue cartoons. N-glycans are represented with yellow sticks, whereas glycosylation sequons are illustrated with yellow vdW spheres. **E)** Molecular representation (side view) of a tilted HA trimer exposing one “anchor” epitope, highlighted with a shiny green surface, to FISW84 Fab (green cartoons, PDB ID: 6HJQ³⁷) for direct binding. HA protomers are represented with surfaces colored with different shades of blue, whereas N-linked glycans are shown with yellow vdW spheres. The viral membrane is depicted with pink vdW spheres. **F)** Molecular representation (frontal view) of FISW84 Fab (green transparent surface) bound to the same HA trimer shown in panel E but in the context of the crowded environment of the H1N1-Shan2009 virion, where other HAs are represented with a gray surface, NAs with a dark gray surface, and N-linked glycans with white vdW spheres.



To gauge the transition time scales between the untitled and the tilted states, we built a two-state MSM of the HA ectodomain tilting using a combination of head–stalk distance criteria as features (**Fig. 3D**) (see Extended Material and Methods in the SI and **Figs. 4.25, 4.26**). The resulting stationary distributions are strongly shifted toward the untitled state, with values of 0.94 (untitled) and 0.06 (tilted) for both H1N1-Shan2009 and H1N1-Mich2009 (see **Table 4.2** for the tilt angles calculated for the macrostates extracted with MSM). The MFPTs are

also comparable between the two strains: 1273 ± 32 ns (H1N1-Shan2009) and 1208 ± 32 ns (H1N1-Mich2015) for the untilted-to-tilted transition; 66 ± 5 ns (H1N1-Shan2009) and 68 ± 5 ns (H1N1-Mich2015) for the tilted-to-untilted transition. This analysis reveals that, on average, longer time scales are needed by the ectodomain to considerably tilt ($> 50^\circ$) relative to the TMD, whereas the untilted position is restored much faster. We then investigated whether the microsecond-long MFPTs estimated for the untilted-to-tilted transition could be somehow imposed by the crowded membrane environment, which may lead to a form of protein confinement. We note that, although influenza viruses are densely covered with glycoprotein spikes,⁶⁸ HA's and NA's ectodomains are not restricted to a small region by a fixed physical barrier and so are not subjected to protein confinement by definition.⁶⁹ In fact, they are not even fully enclosed by neighboring glycoproteins in most cases. In addition, surrounding glycoproteins are often at a loose distance, allowing HA's and NA's ectodomains to move and wiggle through the gaps. However, when at a very close distance from each other, i.e., when interacting, glycoprotein crowding can provide a confining boundary owing to the fact that the glycoproteins are tethered to the membrane. To inquire whether this type of confinement could have an impact on the HA ectodomain tilting motion, we calculated the correlation between the HA ectodomain-tilt angle swing range exhibited by the 236 HAs and the respective average number of connections made with surrounding glycoproteins during the simulations, with the latter used as an estimate of the average confinement. As a result, we found only a weak negative correlation, with Pearson correlation coefficients (r) equal to -0.23 for H1N1-Shan2009 and -0.15 for H1N1-Mich2015 (Fig. S20). Hence, the microsecond-long kinetics extrapolated for the HA ectodomain tilting motion does not seem to be strictly imposed by this specific form of protein confinement. Consistently, almost all the isolated HAs, i.e., the ones that do not interact with any neighboring glycoproteins during the dynamics, do not exhibit the full untilted-to-tilted transition in the sub-microsecond time scale explored by our simulations (**Fig. 4.27**).

The gradual tilting of the ectodomain has a profound impact on the accessibility of the so-called “anchor” epitope located on the lower portion of the HA stalk in the proximity of the viral membrane. As this epitope is known as a target for broadly neutralizing antibodies,^{37,70} we docked the anchor-binding FISW84 Fab³⁷ onto both an untilted and a tilted conformation assumed by an HA trimer throughout the virion simulations, examining the respective poses. When the HA is untilted and upright, the antibody must approach the anchor epitope with an

upward angle (**Fig. 4.28**). Due to the trimer symmetry, accessibility to all three epitopes is hampered by the presence of the lipid bilayer underneath (**Fig. 4.28**). On the contrary, when HA flexes onto the membrane, the angle of approach to at least one of the three anchor epitopes on the HA trimer gradually shifts from upward to lateral or slightly downward (**Fig. 4.3E, Fig. 4.28**), facilitating antibody binding. Next, we docked the FISW84 Fab fragment³⁷ onto a tilted HA in the context of the crowded virion environment. As shown in **Fig. 4.3F**, the antibody can directly target the anchor epitope, even in the presence of neighboring glycoproteins.

Breathing of hemagglutinin head reveals a cryptic epitope

The immunodominant domain of HA is the globular head,^{39,71} spanning residues 66 to 286 and comprising the receptor-binding site (RBS) and the vestigial esterase domain.⁷² In the trimer assembly, the HA1 subunits and the respective head domains are tightly packed in a “closed” state around and atop the HA2 L α Hs. However, upon binding to sialylated receptors during the infection process, HA heads, followed by the HA1 subunits, swing away from HA2, making room for the necessary conformational changes leading to membrane fusion.⁷³ We note that in this work HA was modeled in its uncleaved form (HA0). In our whole-virion constructs, all of the HA trimers are initially in the closed state, with the heads interacting with each other atop the L α Hs. However, over the course of the virion simulations, some of the HA trimers start “breathing,” i.e., reversibly transitioning from the closed state to a partially “open” state where the heads transiently move away from each other (**Fig. 4.4A–B**).

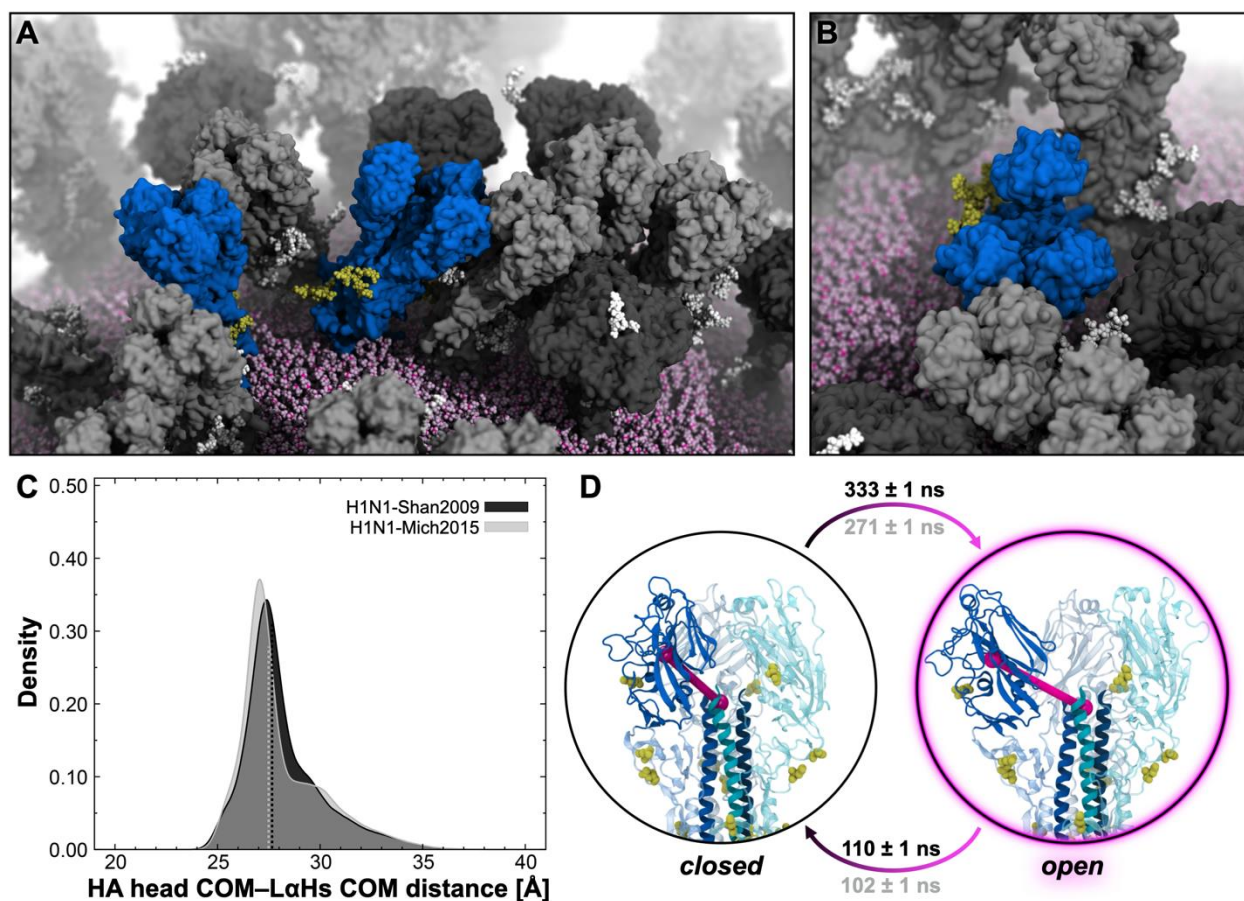


Figure 4.4 | Breathing motion of the HA head domain. **A**) A snapshot from the MD simulation of the H1N1-Shan2009 virion capturing different HAs with varying degrees of head breathing motions (blue surface). NAs are represented with a dark gray surface, whereas other HAs are represented with a gray surface. N-glycans linked to HAs or NAs are shown with yellow or white vdW spheres, respectively. The lipid envelope is depicted using pink vdW spheres. **B**) Close-up, top view of an open HA trimer where the head domains undergo extensive breathing motion. **C**) Value distribution of the distance between the center-of-mass (COM) of the HA head and the COM of the long α -helices' apical residues (L α Hs) obtained from the MD simulation of the H1N1-Shan2009 (black) and H1N1-Mich2015 (gray) virions. The distribution is shown as a kernel density. Dashed lines indicate the median values of the respective distribution. **D**) Two-state MSM of HA head-breathing motion with representative structures from the MD simulation of the H1N1-Shan2009 virion. The mean first-passage times between the states, including the respective standard deviation, are reported rounded to the nearest one for H1N1-Shan2009 (black text) and H1N1-Mich2015 (gray text). Monitored HA head-L α Hs distance is depicted with magenta cylinders drawn within the molecular representation of the HA. HA residues of the head domain and L α Hs are shown with opaque blue cartoons. N-glycans are represented with yellow sticks, whereas glycosylation sequons are illustrated with yellow vdW spheres.

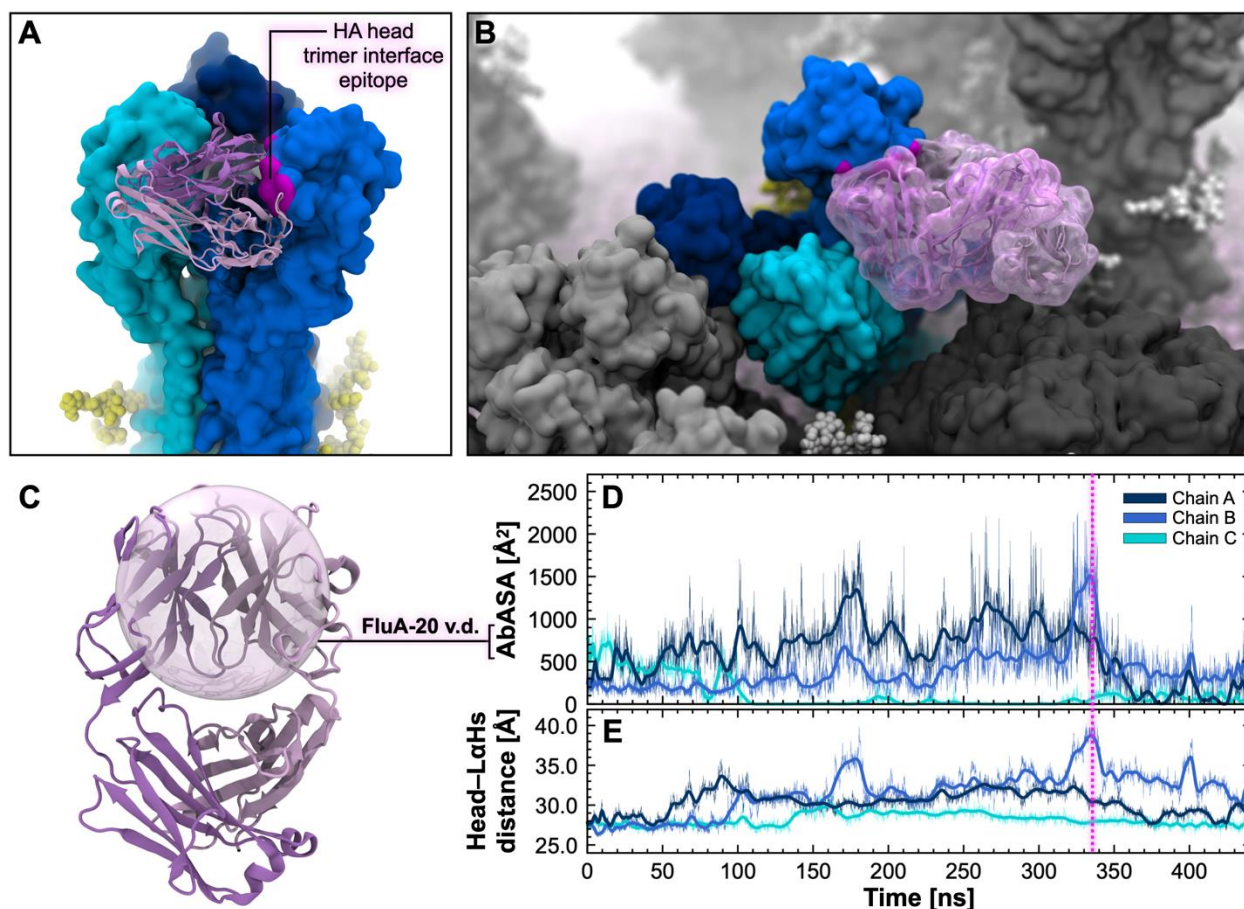
We examined the extent of the HA head breathing motion for all the 236 HA trimers of the two virion models by measuring the distance between the COM of each monomeric HA head and the COM of the apical residues of the three L α Hs. The distributions of the calculated values are presented in **Fig. 4.4C**, whereas the monitored distance is illustrated in **Fig. 4.4D**. Analogously to the NA head tilting and HA ectodomain tilting motions,

H1N1-Shan2009 and H1N1-Mich2015 display a similar extent of HA breathing, exhibiting almost overlapping right-skewed distributions with median values of 27.5 Å and 27.7 Å, respectively (**Fig. 4.4C**). These values coincide with a closed state. When the HA head starts breathing, its distance from the LαHs becomes larger than 30 Å, reaching maximum values of 41.0 Å and 42.2 Å for H1N1-Shan2009 and H1N1-Mich2015, respectively. The two-state MSM constructed for the HA head breathing motion allows us to quantify the kinetics of the closed-to-open and open-to-closed transitions. In this case, we used one feature, defined as the smallest mean distance of the HA heads to each other (see Material and Methods in the SI, and **Figs. 4.29, 4.30**). For H1N1-Shan2009, a stationary distribution of 0.75 was estimated for the closed state, whereas 0.25 for the open state. For H1N1-Mich2015, a stationary distribution of 0.73 was calculated for the closed state, while 0.27 for the open state. The MFPTs are comparable between the two strains: 331 ± 1 ns (H1N1-Shan2009) and 271 ± 1 ns (H1N1-Mich2015) for the closed-to-open transition, whereas 110 ± 1 ns (H1N1-Shan2009) and 102 ± 1 ns (H1N1-Mich2015) for the open-to-closed transition. The opening motion is slower on average than the return to a closed conformation, which is approximately three times faster, pinpointing the open state as a short-lived state. The values of the distance between the HA head and the LαHs for the macrostates extracted with MSM are reported in **Table 4.3**.

Recently, a novel, extremely conserved epitope within the HA head was discovered,³⁹ providing further opportunities for the development of a universal vaccine against influenza.⁷⁴ Promisingly, FluA-20, a broadly protective human antibody, was found to target this cryptic epitope in the head interface region with high breadth and potency against many influenza A virus subtypes.³⁹ Similar to a previously discovered epitope buried in the HA head trimer interface,⁷⁵ the FluA-20 epitope is completely occluded when the HA is in the closed state (**Fig. S24**), necessitating an opening, or breathing, of the HA head to become accessible.³⁹ Our simulations show that the FluA-20 cryptic epitope becomes transiently accessible during HA head breathing (**Fig. 4.5A–B**). To characterize whether the HA head opening is compatible with FluA-20 binding to the cryptic epitope, we extracted the most open conformation of one HA trimer from the H1N1-Shan2009 simulation, and we docked FluA-20 onto it. Strikingly, FluA-20 could be fitted onto the cryptic epitope with no apparent clashes with neighboring monomers (**Fig. 4.5A and Fig. 4.31**). This pose is similar to the one disclosed in a recent structural study.⁴³ Interestingly, the FluA-20 binding to the open HA head is also compatible with the crowded virion

environment (**Fig. 4.5B**). We next estimated the accessibility of the cryptic epitope to FluA-20. FluA-20 variable domains can be fitted into a sphere of radius 18.6 Å (**Fig. 4.5C**), which we used as a probe to calculate the antibody accessible surface area (AbASA) of the cryptic epitope during the simulation (**Fig. 4.5D**). As a test case, we considered an HA trimer that showed extensive breathing. Together with the AbASA profile, we also present the time evolution profile of the distance between the three HA head COMs and the L α Hs COM (**Fig. 4.5E**).

Figure 4.5 | HA breathing reveals the FluA-20 cryptic epitope. **A)** Molecular representation of an HA trimer in the open state extracted from the H1N1-Shan2009 simulations (side view). The three HA monomers are depicted with surfaces colored with different shades of blue. FluA-20 Fab heavy and light chains are represented with pink and purple cartoons, respectively. FluA-20 Fab structure was taken from PDB ID: 6OC3.³⁹ The cryptic epitope targeted by FluA-20 is highlighted with a magenta surface. N-linked glycans are shown with yellow vdW spheres. **B)** Molecular representation of the same open HA trimer shown in panel A, in the context of the crowded virion environment. NAs are represented with a dark gray surface, whereas other HAs with a gray surface. N-glycans linked to HAs or NAs are shown with yellow or white vdW spheres, respectively. The lipid envelope is depicted using pink vdW spheres. **B)** Close-up view of an open HA trimer where the head domains undergo extensive breathing motion. **C)** Molecular representation of the FluA-20 human antibody, where the heavy and light chains are depicted with pink and purple cartoons, respectively. The 18.6 Å radius probe sphere comprising FluA-20 variable domains is overlaid as a transparent surface. **D)** The time evolution profile of the AbASA of the FluA-20 epitope is shown for the three HA chains illustrated in panels A and B. A magenta dashed line indicates the frame shown in panel A and B. **E)** The time evolution profile of the distance between the COM of the HA head and the COM of the L α Hs' apical residues is shown for the three HA chains illustrated in panels A and B.



The respective epitopes in chains A, B, and C in the trimeric HA head show different accessibility according to the extent of breathing. Chains A and B in the head display the most extensive breathing, and the maximum epitope accessibility occurs in correspondence to the full opening of chain B in the head (**Fig. 4.5D–E**). The snapshot corresponding to the maximum head opening of chain B was used to dock FluA-20 onto the respective epitope (**Fig. 4.5A–B**). Another interesting consideration is that the heads appear to breathe asymmetrically, opening or closing irrespective of each other (**Fig. 4.5E**). They can fleetingly open and suddenly rotate back to a close position. Interestingly, possibly due to the presence of nearby glycoproteins, the head of chain C does not breathe, keeping the respective FluA-20 cryptic epitope fully occluded.

The dynamic interplay between glycoproteins

Throughout the simulation of both virus strains, HA and NA are characterized by exceptional flexibility of their extra-virion functional domains. Among others, we have highlighted the NA head tilting motion, the HA ectodomain tilting motion, and the breathing of HA heads. Although these motions relate to individual

glycoproteins, they occur in a crowded subcellular environment that is realistically reproduced in our MD simulations with an atomic level of detail. This is one of the main advantages of performing simulations at the “whole-virion” regime rather than at the “single-protein” regime. Embedded in the viral membrane are, in fact, 236 HA trimers and 30 NA tetramers that do more than wiggling and jiggling and whose interplay and balance are at the basis of viral fitness.⁷⁶ We, therefore, examined the intricate yet fundamental interplay between HA and NA glycoproteins in our mesoscale simulations. The glycoproteins can move across the lipid bilayer and can twist, bend, tilt and interact with each other using their extra-virion functional domains. Even though transmembrane proteins in model membranes, such as ours, usually partition into unrealistic liquid disordered domains,⁷⁷ we see protein clustering, which is also seen biologically.^{78,79} As proteins usually do not form these sorts of clusters, or quinary interactions, without a purpose,^{80,81} these patches are then likely to have functional significance in viral entry and/or egress.^{78,82,83} To investigate the motion of the glycoproteins’ TMD within the viral membrane, we measured the RMSF of the TMD COM of every HA and NA over the course of the simulations with respect to the starting frame. Similarly, we also computed the RMSF of the ectodomain COM of every HA and NA. We then compared the resulting distributions of RMSF values (**Fig. 4.32**). For both strains, this analysis evinces that HA and NA do not vastly translate through the membrane in the microsecond timescales (**Fig. 4.32**), whereas their ectodomains show higher mobility than the TMDs since they can tilt or breathe. However, we cannot rule out the possibility that HA and NA could translate more over longer timescales. With this regard, we note that the interior of the virion models simulated here is filled with water molecules and that the matrix (M1) proteins lining the internal side of the POPC lipid bilayer were not modeled.

Next, we analyzed the number of one-to-one connections established by each glycoprotein with the surrounding glycoproteins. A connection is formed when two glycoproteins move within 5 Å of each other, considering only the protein/glycan heavy atoms of their ectodomains (see Material and Methods in the SI for additional details). The schematic and the analysis are shown in **Fig. 4.6A**. We found that each glycoprotein can connect with up to five other glycoproteins at once. Interestingly, the fraction of isolated glycoproteins (i.e., glycoproteins without a connection to any other glycoprotein) decreased over the course of the simulations from 34.2% to 12.4% in H1N1-Shan2009 and from 36.8% to 16.5% in H1N1-Mich2009. The glycoproteins tend to interact with each other forming new connections rapidly. The most considerable growth is in the fraction of glycoproteins forming

two connections, 17.3% to 34.2% (H1N1-Shan2009) and 16.2% to 34.2% (H1N1-Mich2015), followed by the fraction of glycoproteins forming three connections, 5.6% to 13.5% (H1N1-Shan2009) and 4.1% to 11.3% (H1N1-Mich2015). More rarely, HA and NA interconnect with four or five glycoproteins simultaneously. Considering that few quinary interactions last into the μ s timescale,⁸⁴ our results suggest dynamic fluctuations in the degree of protein clumping: within just one microsecond, large protein clumps can form and dissipate. From the analysis of conditional probabilities (see Materials and Methods in the SI for additional details), it emerges that the inter-protein connectivity does not affect the probability of tilting or breathing of NA and HA (**Tables 4.4–4.6**). On the contrary, it appears that extensive HA ectodomain tilting and, to some extent, HA head breathing as well, increase the probability of forming one or more connections for HA, boosting the interplay (**Tables 4.13 and 4.14**). A less defined behavior is instead observed by NA, where for certain ranges of values of head tilt, the interplay is boosted, whereas for other ranges is instead reduced (**Table 4.9**). Despite the point mutations and the glycan addition/deletion within HA/NA, the overall glycoprotein interplay has not dramatically changed from H1N1-Shan2009 to H1N1-Mich2015. However, appreciable differences are observed in the case of NA in H1N1-Mich2015, where the loss of glycan N386 located at the edge of the head appears to have slowed down its propensity to form new connections (**Fig. 4.33**). The dynamic evolution of the pattern of glycoprotein inter-connections for H1N1-Shan2009 and H1N1-Mich2015 is shown.

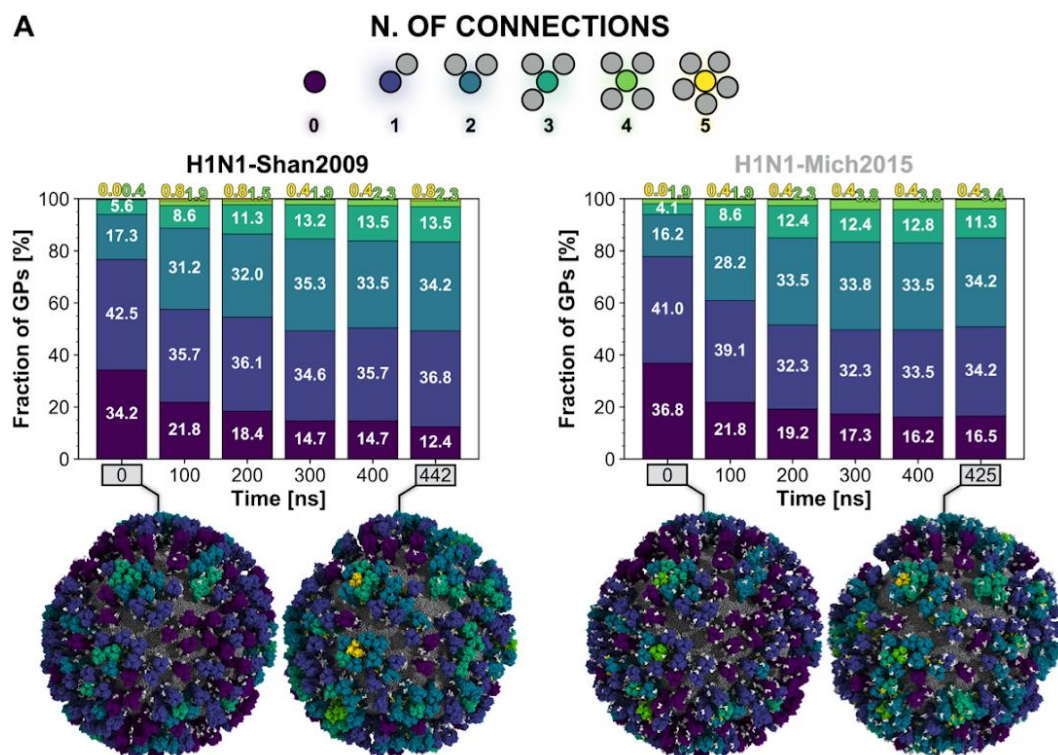
A consequence of influenza glycoproteins clumping together is the formation of large macro-clusters of mixed HA/NA when multiple small clusters come into close contact, i.e., less than 5 Å from each other (**Fig. 4.33**). Congregations of HAs and local clusters of NAs have been previously reported through cryoET experiments.⁶⁸ Here, we observed the formation of macro-clusters comprising as many as 42 (H1N1-Shan2009) and 34 (H1N1-Mich2015) mixed HA/NA glycoproteins at the same time (**Fig. 4.33**). Larger macro-clusters are more frequent and more stable in H1N1-Shan2009 than in H1N1-Mich2015. This is again primarily due to the missing N386 glycan in H1N1-Mich2015 NAs, which leads to the disruption of the continuous pattern of connections between glycoproteins. The dynamic evolution of the glycoprotein macro-clusters for H1N1-Shan2009 and H1N1-Mich2015 is shown.

Subsequently, we sought to dissect the interaction modes between glycoproteins, i.e., how inter-connections are formed. There are three ways a glycoprotein can interact with another one: using i) protein residues only; ii)

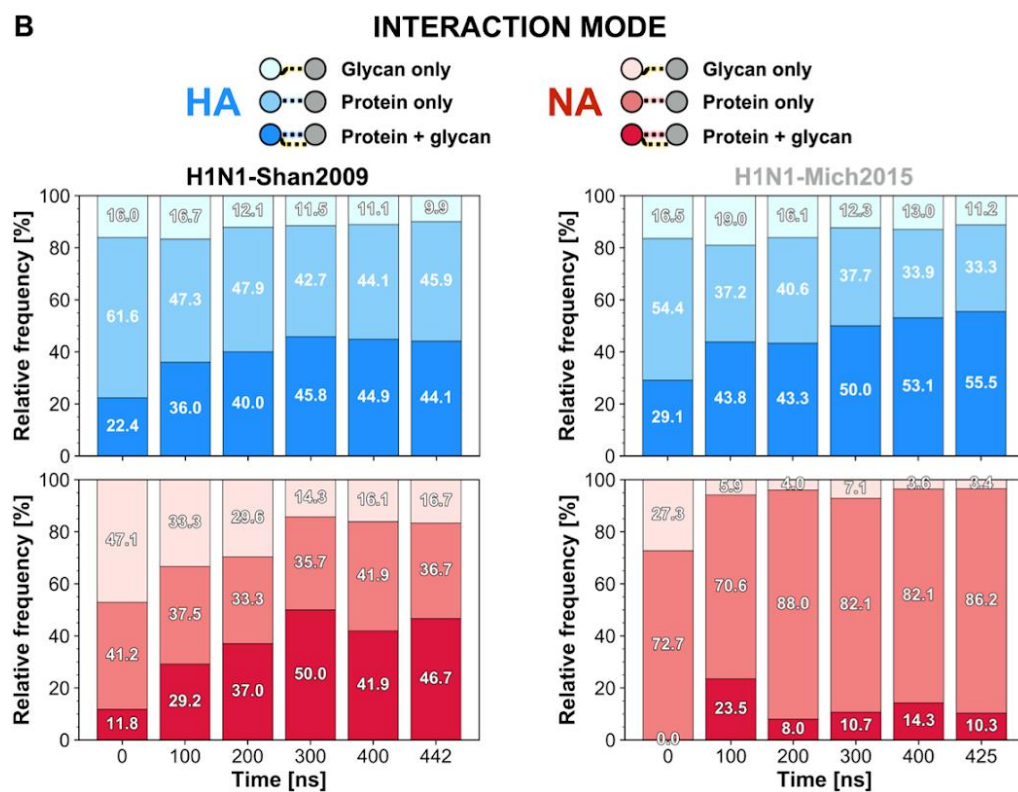
glycan residues only; iii) protein and glycan residues at the same time. The schematic and the analysis are shown in **Fig. 4.6B**. N-glycans play an important role in modulating the HA/NA interplay as they are largely involved in the glycoprotein interactions. The number of connections formed by HA involving glycans progressively increases during the simulations (**Fig. 4.6B**). At the end of the H1N1-Shan2009 simulation, 54.0% of connections made by HAs entail glycans, either alone (9.1%) or concomitant with protein residues (44.1%). Intriguingly, the introduction of glycan N179 within the HA head in H1N1-Mich2015 augments the participation of glycans in connections formed by HA to 66.7% (glycan only 11.2%, concomitant with protein 55.5%) (**Fig. 4.6B**). A similar trend is also observed for NA in H1N1-Shan2009. Instead, the opposite scenario takes place in the case of NA in H1N1-Mich2015, owing to the loss of glycan N386 within the head. While at the end of the H1N1-Shan2009 simulation, glycans are involved in 63.4% of connections formed by NA (16.7% alone, 46.7% with protein), in H1N1-Mich2015, only 13.7% of NA connections include glycans (3.4% alone, 10.3% with protein) (**Fig. 4.6B**). This finding highlights the critical role of NA glycan N386 in H1N1-Shan2009 as a molecular sensor to boost the glycoprotein-glycoprotein interplay. The relative frequency of connections involving glycans alone progressively decreases during the simulations in favor of interactions involving glycans and protein simultaneously and occurring only after the glycoproteins are brought in proximity to each other (**Fig. 4.6B**). Finally, an in-depth analysis of which glycans are mostly used by HA and NA to form connections reveals that NA preferably uses head glycans while HA mostly utilizes stalk glycans (**Fig. 4.35**). The relative frequency of interacting NA head glycans considerably drops in H1N1-Mich205 due to the loss of N386. Instead, the relative frequency of interacting HA head glycans considerably increases in H1N1-Mich205 due to the addition of N179 glycan (**Fig. 4.35**).

Figure 4.6 | HA/NA interplay. A) The pattern of connections (0 to 5) made by either HA or NA (circles colored in viridis palette) with surrounding glycoproteins (gray circles) is represented in the schematic shown on top of the panel. Below are the stacked bar plots representing the fraction (%) of glycoproteins establishing a certain number of connections with surrounding glycoproteins along the simulation of H1N1-Shan2009 (left) and H1N1-Mich2015 (right). Percentages rounded to the first decimal digit are reported. Molecular representations of the virions at the initial and last frame of the simulations are shown under the plots. The glycoproteins are depicted with a surface colored according to the number of connections made by the respective glycoprotein at that frame. Interacting glycans are highlighted with yellow vdW spheres, non-interacting glycans with white vdW spheres. **B)** The legend for the interaction mode between glycoproteins (either HA or NA) is schematically represented on top of the panel. Below are the stacked bar plots representing the relative frequency (%) of connections made by the glycoproteins and involving glycan only (lighter shade), protein only or protein and glycan (darker shade) simultaneously along the simulation of H1N1-Shan2009 (left) and H1N1-Mich2015 (right). A blue gradient for the three categories is adopted for HA, whereas a red gradient is used for NA. Percentages rounded to the first decimal digit are reported.

A



B



4.4 Discussion

A critical determinant of influenza A infection, replication, and transmissibility is the functional balance between HA and NA.^{8,76,85–87} One fascinating aspect of the influenza viral cycle is that the two glycoproteins compete, with opposite actions, for the same substrate, i.e., the terminal sialic acid (SA) residue presented by host cell-linked glycans. While HA mediates viral entry by binding to SA, NA modulates the release of the viral progeny by cleaving SA moieties from both the sialylated receptors and HA.⁷⁶ In a scenario where the virus fitness depends on the delicate, antigenic drift-dependent equilibrium between HA binding avidity and NA enzymatic activity,⁸ substantial knowledge gaps persist in the understanding of the HA/NA dynamic interplay taking place at the level of the virus surface. To shed light on this crucial aspect and its antigenic implications, we performed massive, all-atom whole-virion MD simulations of two evolutionarily linked influenza A strains, H1N1-Shan2009 (**Fig. 4.1**) and H1N1-Mich2015 (**Fig. 4.8**). Maintaining the accuracy of the atomic detail, we implemented a multiscale computational protocol that exploits different modeling tools, analysis methods, and algorithms for crossing spatial scales (from molecular to subcellular) and exploring events occurring in different temporal scales. Our simulations provided previously unseen dynamic views of the plastic conformational landscape of HA and NA ectodomains, revealing three main extensive motions: NA head tilting (**Fig. 4.2**), HA ectodomain tilting (**Fig. 4.3**), and HA head breathing (**Fig. 4.4**). While the last two motions find evidence in previous experimental work,^{37,43,74} to the best of our knowledge, no work has yet detailed the flexibility of the NA head relative to the stalk. In a recent study by Ellis et al.,⁴⁴ the authors investigated the plasticity of a recombinant NA tetramer, revealing an open tetrameric assembly where the four heads move away from each other, giving the NA a clover-like appearance. As shown in **Fig. 4.36**, in our simulations, we did not observe the breathing of the monomeric head domains, most likely due to insufficient sampling and/or inconsistencies between our model and the experimental construct. Instead, the NA tetrameric head tilts as a single body as much as $> 90^\circ$ relative to the stalk (**Fig. 4.2C** and **Fig. 4.18**). Although not discussed in Ellis et al.,⁴⁴ cryoEM micrographs included therein hint at tilted head conformations potentially compatible with poses extracted from our simulations, thus indirectly supporting our findings. The tilting of prefusion HA trimer has been previously mentioned in a theoretical study⁸⁸ and more recently demonstrated by cryoEM experiments probing it in detergent micelles.³⁷ Additionally, the predisposition to assume tilted orientations is a common property shared

by class I fusion viral glycoproteins such as HIV Env⁸⁹ and SARS-CoV-2 spike.^{53,90} The HA head breathing was initially supported by smFRET experiments⁹¹ and then indirectly reported concomitant with the discovery of a cryptic epitope located at the interface between monomers in the head.^{38,39,41,43,92} Breathing has been reported for other viral glycoproteins in flavivirus⁹³ and HIV.⁹⁴ Not only are our simulations in agreement with the experiments,⁴³ but they are the first of their kind to show the HA breathing or flexing on the viral membrane in the context of the whole virion. The probability densities of NA head-tilt angle, HA ectodomain-tilt angle, and HA head–L α His distance (**Figs. 4.2C, 4.3C, 4.4C**, respectively) indicate that open HA head (HA head–L α His distance > 30 Å) or tilted HA (> 50°) and NA (> 75°) conformations are rarely sampled during the simulations, whereas closed, untilted, or slightly tilted conformations are most likely. We also characterized the kinetics of all three motions, building a two-state MSM for each. MFPTs of HA head-breathing and HA ectodomain tilting motions (**Figs. 4.3D, 4.4D**) indicate that transitioning to an open HA trimer or a tilted HA ectodomain is slower than recovering a closed or an upright conformation. Instead, MFPTs of the NA head tilting motion pinpoint that the transition to a tilted state is faster than the reverse one only for H1N1-Mich2015.

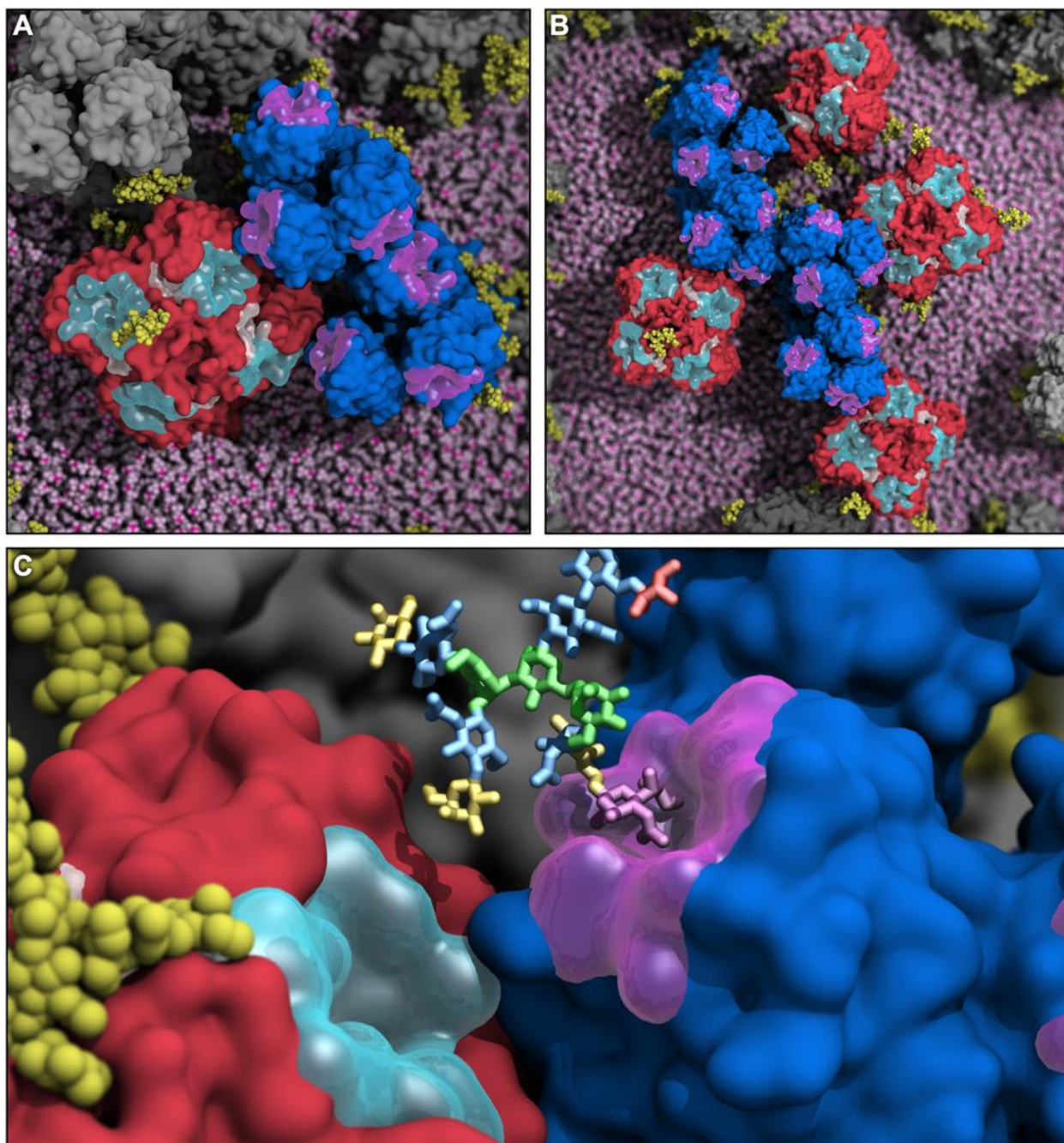
The high-dimensional mobility of the HA and NA ectodomains proffers wide-ranging implications. From an immunogenic point of view, the glycoproteins' accentuated conformational plasticity observed in our simulations has unveiled their vulnerabilities, exposing epitopes that otherwise would not be accessible or would be sterically occluded. As discussed above, tilted, open, or partially open conformations are transient and, most importantly, reversible. Stabilizing these fleeting states can be used as a strategy to lock the glycoproteins in vulnerable orientations that can be harnessed for vaccine development and the design of antiviral drugs. This could be achieved through amino acid substitutions imparting a specific tertiary/quaternary structure to the glycoprotein.⁹⁵ The list of immunogenically relevant states revealed by our simulations includes a previously uncharacterized NA with a tilted head. We identified a novel epitope located on the underside of an N2-subtype NA (**Fig. 4.2E, F** and **Figs. 4.22, 4.23, 4.24**) head that becomes directly accessible for immune recognition during head tilting. Importantly, the underside epitope is conserved in the NA of contemporary H3N2 viruses, unlike the more accessible epitopes surrounding the catalytic site where the H3N2 viruses acquired an additional glycosylation site⁹⁶ resulting in resistance to a subset of antibodies targeting the catalytic site. This reaffirms the importance of NA, and potentially of its head underside, as a key target for vaccine and drug discovery.

Next, the flexibility of the HA ectodomain affects the approachability of a conserved epitope located at the juxta-membrane region of the HA stalk (**Fig. 4.3E–F**), also referred to as the anchor epitope. We showed that as the HA bends towards the membrane, the angle of approach to at least one of the anchor epitopes drastically changes, adjusting to a favorable orientation for antibody binding. Our simulations complement previous structural experiments that reported several broadly neutralizing antibodies targeting the anchor epitope,^{37,97} confirming that the epitope is occluded by the underneath membrane when the HA is untilted (**Fig. 4.28**). Trapping the HA ectodomain in a tilted state could bolster the immune response toward the anchor epitope. Finally, our simulation unveiled how the cryptic epitope located at the protomer interface between the HA heads^{38,39,92} becomes transiently exposed when HA breathes. Most of the HA heads remain in the closed conformation during the dynamics and rarely open. When breathing occurs, the monomers fleetingly open at different times and rapidly rotate back to a closed position. We could not create an MSM with all three monomers breathing symmetrically, meaning that all three monomers breathing symmetrically did not occur often enough to be appropriately sampled, and thus if it occurs at all, will be at a lower probability than a single monomer breathing. Nonetheless, we demonstrated that reversible breathing of at least one monomer is enough to accommodate a broadly protective antibody, such as FluA-20,³⁹ at the interface epitope.

From a functional point of view, the conformational plasticity of influenza glycoproteins' ectodomains would mediate the virus' ability to traverse the cell surface.^{9,98,99} This plasticity could also facilitate gaining access to SA moieties on the host cell surface, either during the viral entry or egress processes. Tilting and breathing may help HA engage the flexible sialylated receptors in a dynamic and crowded environment. Head tilting makes NA like a weedwhacker on a rotating axis, capable of trimming terminal SA moieties from the host cell with a 180° swing range. Improved substrate binding may also indirectly increase the catalytic efficiency of NA, which was shown to be enzymatically active only as a tetramer.¹⁰⁰ In both cases, the plasticity of the ectodomains helps accommodate the hypermobility of the polysaccharide chains presenting the targeted SA residues. Consequently, the functional HA/NA balance also depends on the HA/NA dynamic interplay,^{8,76} i.e., the interactive set of interactions, conformational changes, and collective motions by which the glycoproteins affect each other's dynamics and orientation on the virion surface. Structural factors such as NA stalk length, HA:NA stoichiometry, or the spatial distribution of the glycoproteins within the viral membrane, also contribute to fine-tuning the

HA/NA balance.⁸⁵ Yet, influenza virus is not a static system. As portrayed in the results section, the HA/NA interplay is indeed dynamic because the glycoproteins fleetingly tilt, bend, “breathe”, and move in such a way that the pattern of their interactions continuously changes over time (**Fig. 4.6** and **Figs. 4.33, 4.34**). Glycoproteins that were not contacting any other glycoproteins at the beginning of the simulations ended up interacting with one, two, or more glycoproteins, forming clusters (**Fig. 4.7A**) and larger macro-clusters (**Fig. 4.7B**). Our simulations provided unprecedented views on the HA/NA interplay, where the SA binding sites on HA/NA move in proximity to each other upon concomitant tilting or breathing motions (**Fig. 4.7A** and **Fig. 4.7B**). Hence, NA and HA can compete for the same substrate (**Fig. 4.7C**), as suggested by a recent single-molecule force spectroscopy study.²⁶ Owing to its head flexibility, NA can reach multiple HA binding sites at the same time and contend for the same sialylated cell receptors (**Fig. 4.7C**). When influenza virus is in the proximity of the host cell sialoglycan receptors, HA and NA establish multivalent interactions, i.e., multiple, non-covalent weak interactions, with the terminal sialic acid residues, leading to a firm attachment.³⁰ This process allows virus internalization but can also modulate progeny release.³⁰ Although the number of multivalent HA/NA-cell receptor interactions needed for a firm attachment is still unknown,^{29,30} knowledge of the large degree of HA/NA ectodomain flexibility provided by our simulations may improve future predictions in this area.

Figure 4.7 | HA and NA interplay at the basis of the viral egress process. A-B) Molecular representation of snapshots retrieved from the simulation of H1N1-Shan2009 where HA and NA clump together, leading to the formation of clusters (**A**) and macro-clusters (**B**). The same representation scheme as panel A is adopted. The lipid envelope is depicted using pink vdW spheres. **C)** Close-up molecular representation of a snapshot from the simulation of H1N1-Shan2009 where HA binding site (semi-transparent magenta surface) and NA active site (semi-transparent cyan surface) are in proximity to each other, possibly competing for the same substrate. A complex glycan terminating with sialic acid, mimicking a putative host-cell receptor and represented with colored sticks, was docked into the HA binding site based on the pose of an HA substrate crystalized in PDB ID: 3UBE.¹⁰¹ The glycan residues are colored according to the Symbol Nomenclature for Glycans (SNFG),¹⁰² where N-Acetylglucosamine is colored in blue, mannose in green, fucose in red, galactose in yellow, and sialic acid in purple. HA is shown with a blue surface, whereas NA with a red surface. N-glycans linked to HA and NA are highlighted with yellow vdW spheres.



Glycans are fundamental post-translational modifications of viral envelope proteins,^{103,104} including HA and NA. As such, they exert a profound impact on enveloped virus biology and so on the overall fitness of the virus.^{10,105} Among their many roles, glycans modulate evasion of the immune system by shielding antigenic epitopes from antibody recognition,^{10,106–110} tune the interaction and affinity with the host cell,¹¹¹ affect to a certain extent the binding of HA¹¹² and NA¹¹³ to sialylated receptors, or can even contribute to protein folding¹¹⁴ and HA

cleavage.¹¹⁵ Here, we show that N-linked glycans seem to play a role in enhancing HA/NA interplay. Indeed, glycans form interactions with neighboring glycoproteins, acting as antennas. We note that the impact of glycans could be even more extensive than described in this work since the glycoproteins are not fully glycosylated in our virion models (see Material and Methods for further details). Despite the inter-glycoprotein interaction patterns being similar in H1N1-Shan2009 and H1N1-Mich2015 (**Fig. 4.6A**), substantial differences exist between the two strains in how the interactions are made (**Fig. 4.6B**, **Fig. 4.35**). Especially in the case of NA, the loss of glycan N386 in H1N1-Mich2015 within the head has reduced the ability of NA to form new connections, pinpointing possible impact on the structural/functional balance with HA during infection.

Overall, our massive whole-virion MD simulations provide previously unappreciated views on the influenza virus dynamics, highlighting the remarkable flexibility of the glycoproteins' ectodomains. We characterized reversible conformational transitions, such as NA head tilting, HA ectodomain tilting, and HA head breathing, which also revealed potential vulnerabilities of influenza glycoproteins to the immune system that are conserved across two evolutionarily related H1N1 strains, specifically H1N1-Shan2009 and H1N1-Mich2015. We identified a novel epitope located on the underside face of the NA head, further attesting the NA as a key antigenic target. Finally, our work substantially advances the knowledge of the HA/NA dynamic interplay in the context of viral entry and egress processes, ushering in unprecedented views on glycoprotein cooperativity. Altogether, while expanding knowledge on the molecular determinants of infection, we envision that our work will contribute to a viable strategy for the development of a universal or more effective vaccine against influenza virus.

4.5 Materials and methods

Computational methods summary

Initial coordinates for the all-atom whole-virion models of H1N1-Shan2009 and H1N1-Mich2015 were derived from the last frame of all-atom MD simulations of the unglycosylated H1N1-Shan2009 whole-virion model previously performed by us and thoroughly described in Durrant et al.⁴⁵ This model includes 236 HA trimers, 30 NA tetramers, and 11 M2 proton channels embedded in a semi-spherical lipid bilayer.⁴⁵ The M1 matrix proteins, coating the inner leaflet of the lipid bilayer, and the ribonucleoproteins contained inside of the virion

were not included in the model. Instead, the virion interior was filled in with water molecules. The shape of the virion and the distribution of the glycoproteins in the viral membrane were adopted from a point model of the virion exterior provided by collaborators,^{45,116} which was in turn derived from a cryoET map of the influenza virus pleiomorphy.⁶⁸ H1N1-Mich2015 was generated from H1N1-Shan2009 upon modeling the point mutations within HA (15 per monomer) and NA (14 per monomer) by means of NAMD's PSFGEN¹¹⁷ and the CHARMM36 amino acid topologies.¹¹⁸ From this initial point, the main challenge was the glycosylation of the HA trimers and NA tetramers. In H1N1-Shan2009, HA and NA present six and eight N-linked sequons on each monomer, respectively. H1N1-Mich2015 introduced an additional glycosylation site within the HA head at position N179, whereas glycan N386 within the NA head was deleted. Building upon the coordinates derived from Durrant et al.,¹¹⁹ glycosylation of HA and NA was carried out with the doGlycans tool.¹²⁰ The glycoprofile was based on available glycomics data.^{11,121–125} Since glycans were added to an unglycosylated construct resulting from 69.74 ns of MD,⁴⁵ many sequons ended up not being glycosylated either because sterically hindered by adjacent residues, occluded by neighboring glycoproteins, or not fully exposed to the solvent. The resulting, glycosylated H1N1-Shan2009 and H1N1-Mich2015 whole virion constructs, comprising 236 glycosylated HA homo-trimers, 30 glycosylated NA homo-tetramers, and 11 M2 homo-tetramers embedded in the POPC lipid bilayer, were parameterized using NAMD's PSFGEN¹¹⁷ and CHARMM36 all-atom additive force fields for protein,¹¹⁸ lipids,^{126,127} glycans,¹²⁸ and ions.¹²⁹ As mentioned above, explicit TIP3 water molecules¹³⁰ located inside and outside of the lipid bilayer as well as Ca^{2+} ions bound to the NA heads, were retained from the original unglycosylated system.⁴⁵ The number of Na^+ and Cl^- ions was adjusted to neutralize the outstanding charge, with ionic strength set at 150 mM. The total number of atoms of the final systems is 160,919,594 for H1N1-Shan2009 and 160,981,954 for H1N1-Mich2015, with an orthorhombic periodic cell of $\sim 114 \text{ nm} \times \sim 119 \text{ nm} \times \sim 115 \text{ nm}$. All-atom MD simulations were performed on the ORNL TITAN supercomputer and the NSF Blue Waters supercomputer, respectively, using the CUDA memory-optimized version of NAMD 2.13¹³¹ and CHARMM36 force fields.^{118,126–128} Upon minimization, heating, and equilibration, the systems were submitted to productive MD simulations in NPT conditions using the Langevin thermostat¹³² and the Nosé-Hoover Langevin piston^{133,134} to achieve pressure (1.01325 bar) and temperature (298 K) control. One continuous replica was performed for each system using a timestep of 2 fs. Non-bonded interactions (van der

Waals (vdW) and short-range electrostatic) were calculated at each timestep using a cutoff of 12 Å and a switching distance of 10 Å. All simulations were performed using periodic boundary conditions, employing the particle-mesh Ewald method¹³⁵ with a grid spacing of 2.1 Å to evaluate long-range electrostatic interactions every three timesteps. SHAKE algorithm¹³⁶ was adopted to keep the atomic bonds involving hydrogens fixed. Frames were saved every 30,000 steps (60 ps). For H1N1-Shan2009, we collected a total of 441.78 ns (7363 frames) continuous productive MD. For H1N1-Mich2015, we collected a total of 424.98 ns (7083 frames) continuous productive MD. All the analyses, including RMSD, RMSF, NA head-tilt angle, HA ectodomain-tilt angle, HA head breathing, AbASA, and glycoprotein interactions were all performed with Visual Molecular Dynamics (VMD) software¹³⁷ using in-house developed scripts. Figure panels and movies of the whole-virion simulations were rendered using VMD.¹³⁷ MSMs were created with the PyEmma2 software program.¹³⁸ For this purpose, individual NA tetrameric (30) and HA trimeric (236) trajectories were generated from the whole-virion MD simulations of H1N1-Shan2009 and H1N1-Mich2015, respectively, and loaded in the MSM workflow. We note that with our treatment, only a subset of system coordinates is explicitly accounted for, whereas other components of the entire set of the system coordinates, such as, for example, the inter-protein connectivity, are only implicitly included, potentially introducing non-Markovian terms to the dynamics described by our MSMs. In this regard, we verified that the inter-protein connectivity does not alter the probability of achieving a certain extent of tilting or breathing by HA and NA (see Materials and Methods section in the SI, and Tables S4–S9 for more details). To build the MSMs presented here, we used one to two features, whose dimensionalities were then reduced and/or data reformatted through TICA.^{139–141} After discretization of the trajectories, we clustered them through k-means clustering.¹⁴² We then created Bayesian MSMs and validated them through Chapman-Kolmogorov tests.¹⁴³ Next, we used Robust Perron Cluster Center Analysis (PCCA++)¹⁴⁴ to assign microstates to corresponding macrostates, and finally extracted 10 representative trajectory frames from the microstate with the highest probability assignment to its macrostate. Averaging the tilt angles from those 10 frames showed clear conformational transitions, allowing us to confidently use their associated MFPTs. A full description of computational methods is provided in the SI.

Experimental materials and methods summary

Monoclonal antibody (mAb) NDS.1 was isolated from an H3N2 influenza convalescent human donor by flow cytometry-based single B cell sorting with the N2 NA probe (A/Wisconsin/67/2005). Recombinant mAb NDS.1 was produced by transient transfection of expression vectors encoding the antibody's heavy and light chains and purified by affinity chromatography using the protein A resin. Fab NDS.1 was generated by proteolytic digestion using Lys-C endoprotease. Binding kinetics of the Fab NDS.1 were measured by biolayer interferometry using the Octet HTX instrument. Negative-stain electron microscopy reconstruction model of N2 NA–Fab NDS.1 complex was generated with a final dataset of 2,304 particles. A full description of experimental materials and methods is provided in the SI.

4.6 Associated content

Supporting Information

Materials and Methods, supplementary **figures 4.8 to 4.36**, supplementary **tables 4.1 to 4.9**, captions for supplementary data, and supplementary references (PDF)

4.7 Author information

Author Contributions

L.C. constructed the influenza whole-virion models, performed MD simulations, designed and performed simulation analyses, prepared data for MSM analyses, and created figures and movies. C.S. designed and performed MSM analyses, created related plots, and helped design HA and NA glycoprotein for computational modeling. J.L. isolated and produced anti-NA antibody, performed BLI experiments, created respective figure panels, and wrote related results and methods. Y.T. performed negative-stain EM experiments. M.K. designed and oversaw experiments. L.C., C.S., and R.E.A. wrote the original draft and all the authors contributed to reviewing and editing. I.A.W. mentored L.C. and provided critical feedback throughout. R.E.A. designed, oversaw, and secured resources for the research project.

Funding

L.C. was funded by a Visible Molecular Cell Consortium fellowship. C.S. is supported by a grant from the National Institutes of Health (T32EB009380) and from the National Science Foundation Graduate Research

Fellowship (DGE-1650112). Computing time on the Oak Ridge National Lab Titan supercomputer was provided through INCITE BIP160. Computing time on the NSF Blue Waters supercomputer was provided through NSF OAC-1811685. This work was supported in part by Intramural Research Program of the Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, and federal funds from the Frederick National Laboratory for Cancer Research, National Institutes of Health, under contract number HHSN261200800001E, and by Leidos Biomedical Research, Inc. I.A.W. was supported in part by NIH AI150885 and NIH CIVIC 75N93019C00051.

Notes

The authors declare no competing interests.

Data availability. Jupyter-notebooks for the MSM analysis of NA head tilting, HA ectodomain tilting, and HA head breathing along with PDB files of representative NA/HA conformations extracted for each state are made available in the Supporting Information as Data S1. The density file of the 3D map visualization of Fab NDS.1 in complex with A/Darwin/9/2021 at 24.1 Å resolution is made available in the Supporting Information as Data S2. Final snapshots from the whole-virion simulations will be made available on <https://amarolab.ucsd.edu/>. Full simulation data (30 TB), including trajectories and analysis scripts, will be made available upon request.

4.8 Acknowledgements

L.C. would like to thank Dr. Zied Gaieb and Dr. Lane Votapka for useful discussions on simulation analysis, and Dr. Jacob Durrant for the help he provided for system setup and simulation. C.S. would like to thank Dr. Emilia Pecora de Barros and Dr. Sarah Kochanek for useful discussions on MSM theory and use, and Moritz Hoffman for help with MSM feature selection. We thank Dr. James C. Phillips and Dr. John E. Stone for the incredible help and support with NAMD and VMD. We thank Dr. Meghan Altman and Dr. Pascal Gagneux for useful discussions on the glycan evolution pattern in the influenza virus. We thank Dr. Huan-Xiang Zhou for useful discussions on protein crowding. We thank Dr. Daniel Zuckerman and Dr. Frank Noé for useful discussions on MSM theory.

Chapter 4, in full, is a modified reprint of the material as it appears in “Casalino, L., Seitz, C., Lederhofer, J., Tsybovsky, Y.; Wilson, I.A.; Kanekiyo, M., Amaro, R.E. Breathing and tilting: the flexibility of

influenza glycoproteins reveals their vulnerabilities. *In press at ACS Central Science*” The dissertation author was a co-investigator and co-author of this work.

References

- (1) Wu, N. C.; Wilson, I. A. Influenza Hemagglutinin Structures and Antibody Recognition. *Cold Spring Harb Perspect Med* **2020**, *10* (8), a038778.
- (2) Wilks, S.; de Graaf, M.; Smith, D. J.; Burke, D. F. A Review of Influenza Haemagglutinin Receptor Binding as It Relates to Pandemic Properties. *Vaccine* **2012**, *30* (29), 4369–4376.
- (3) Skehel, J. J.; Wiley, D. C. Receptor Binding and Membrane Fusion in Virus Entry: The Influenza Hemagglutinin. *Annu Rev Biochem* **2003**, *69*, 531–569.
- (4) Air, G. M. Influenza Neuraminidase. *Influenza Other Respir Viruses* **2012**, *6* (4), 245.
- (5) Pielak, R. M.; Chou, J. J. Influenza M2 Proton Channels. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2011**, *1808* (2), 522–529.
- (6) Pinto, L. H.; Lamb, R. A. The M2 Proton Channels of Influenza A and B Viruses *. *Journal of Biological Chemistry* **2006**, *281* (14), 8997–9000.
- (7) Wagner, R.; Matrosovich, M.; Klenk, H. D. Functional Balance between Haemagglutinin and Neuraminidase in Influenza Virus Infections. *Rev Med Virol* **2002**, *12* (3), 159–166.
- (8) Kosik, I.; Yewdell, J. W. Influenza Hemagglutinin and Neuraminidase: Yin–Yang Proteins Coevolving to Thwart Immunity. *Viruses* **2019**, *11* (4), 346.
- (9) de Vries, E.; Du, W.; Guo, H.; de Haan, C. A. M. Influenza A Virus Hemagglutinin–Neuraminidase–Receptor Balance: Preserving Virus Motility. *Trends Microbiol* **2020**, *28* (1), 57–67.
- (10) Altman, M. O.; Angel, M.; Košík, I.; Trovão, N. S.; Zost, S. J.; Gibbs, J. S.; Casalino, L.; Amaro, R. E.; Hensley, S. E.; Nelson, M. I.; et al. Human Influenza A Virus Hemagglutinin Glycan Evolution Follows a Temporal Pattern to a Glycan Limit. *mBio* **2019**, *10* (2), e00204-19.
- (11) Sun, S.; Wang, Q.; Zhao, F.; Chen, W.; Li, Z. Glycosylation Site Alteration in the Evolution of Influenza A (H1N1) Viruses. *PLoS One* **2011**, *6* (7), e22844.
- (12) Blackburne, B. P.; Hay, A. J.; Goldstein, R. A. Changing Selective Pressure during Antigenic Changes in Human Influenza H3. *PLoS Pathog* **2008**, *4* (5), e1000058.
- (13) Gao, J.; Couzens, L.; Burke, D. F.; Wan, H.; Wilson, P.; Memoli, M. J.; Xu, X.; Harvey, R.; Wrammert, J.; Ahmed, R.; et al. Antigenic Drift of the Influenza A(H1N1)Pdm09 Virus Neuraminidase Results in Reduced Effectiveness of A/California/7/2009 (H1N1pdm09)-Specific Antibodies. *mBio* **2019**, *10* (2), 1–17.
- (14) Carrat, F.; Flahault, A. Influenza Vaccine: The Challenge of Antigenic Drift. *Vaccine* **2007**, *25* (39–40), 6852–6862.
- (15) Boni, M. F. Vaccination and Antigenic Drift in Influenza. *Vaccine* **2008**, *26* (Suppl. 3), C8–C14.
- (16) Zost, S. J.; Wu, N. C.; Hensley, S. E.; Wilson, I. A. Immunodominance and Antigenic Variation of Influenza Virus Hemagglutinin: Implications for Design of Universal Vaccine Immunogens. *J Infect Dis* **2019**, *219* (Supplement_1), S38–S45.
- (17) Bullard, B. L.; Weaver, E. A. Strategies Targeting Hemagglutinin as a Universal Influenza Vaccine. *Vaccines* **2021**, *9* (3), 257.
- (18) Creytens, S.; Pascha, M. N.; Ballegeer, M.; Saelens, X.; de Haan, C. A. M. Influenza Neuraminidase Characteristics and Potential as a Vaccine Target. *Front Immunol* **2021**, *12*, 786617.
- (19) Nachbagauer, R.; Feser, J.; Naficy, A.; Bernstein, D. I.; Gupitill, J.; Walter, E. B.; Berlanda-Scorza, F.; Stadlbauer, D.; Wilson, P. C.; Aydiillo, T.; et al. A Chimeric Hemagglutinin-Based Universal Influenza Virus Vaccine Approach Induces Broad and Long-Lasting Immunity in a Randomized, Placebo-Controlled Phase I Trial. *Nat Med* **2020**, *27* (1), 106–114.
- (20) Lakadamyali, M.; Rust, M. J.; Zhuang, X. Endocytosis of Influenza Viruses. *Microbes Infect* **2004**, *6* (10), 929–936.
- (21) Rossman, J. S.; Lamb, R. A. Influenza Virus Assembly and Budding. *Virology* **2011**, *411* (2), 229–236.
- (22) Nayak, D. P.; Balogun, R. A.; Yamada, H.; Zhou, Z. H.; Barman, S. Influenza Virus Morphogenesis and Budding. *Virus Res* **2009**, *143* (2), 147–161.
- (23) Schmitt, A. P.; Lamb, R. A. Influenza Virus Assembly and Budding at the Viral Budozone. *Adv Virus Res* **2005**, *64*, 383–416.
- (24) Liu, C.; Maryna, †; Eichelberger, C.; Compans, R. W.; Air, G. M. Influenza Type A Virus Neuraminidase Does Not Play a Role in Viral Entry, Replication, Assembly, or Budding. *J Virol* **1995**, *69* (2), 1099–1106.

- (25) Nayak, D. P.; Hui, E. K. W.; Barman, S. Assembly and Budding of Influenza Virus. *Virus Res* **2004**, *106* (2), 147–165.
- (26) Reiter-Scherer, V.; Cuellar-Camacho, J. L.; Bhatia, S.; Haag, R.; Herrmann, A.; Lauster, D.; Rabe, J. P. Force Spectroscopy Shows Dynamic Binding of Influenza Hemagglutinin and Neuraminidase to Sialic Acid. *Biophys J* **2019**, *116* (6), 1037–1048.
- (27) Du, R.; Cui, Q.; Rong, L. Competitive Cooperation of Hemagglutinin and Neuraminidase during Influenza A Virus Entry. *Viruses* **2019**, *11* (5), 458.
- (28) Yang, J.; Liu, S.; Du, L.; Jiang, S. A New Role of Neuraminidase (NA) in the Influenza Virus Life Cycle: Implication for Developing NA Inhibitors with Novel Mechanism of Action. *Rev Med Virol* **2016**, *26* (4), 242–250.
- (29) Overeem, N. J.; van der Vries, E.; Huskens, J.; Overeem, N. J.; Huskens, J.; van der Vries, E. A Dynamic, Supramolecular View on the Multivalent Interaction between Influenza Virus and Host Cell. *Small* **2021**, *17* (13), 2007214.
- (30) Müller, M.; Lauster, D.; Wildenauer, H. H. K.; Herrmann, A.; Block, S. Mobility-Based Quantification of Multivalent Virus-Receptor Interactions: New Insights into Influenza A Virus Binding Mode. *Nano Lett* **2019**, *19* (3), 1875–1882.
- (31) Cuellar-Camacho, J. L.; Bhatia, S.; Reiter-Scherer, V.; Lauster, D.; Liese, S.; Rabe, J. P.; Herrmann, A.; Haag, R. Quantification of Multivalent Interactions between Sialic Acid and Influenza A Virus Spike Proteins by Single-Molecule Force Spectroscopy. *J Am Chem Soc* **2020**, *142* (28), 12181–12192.
- (32) Sieben, C.; Kappel, C.; Zhu, R.; Wozniak, A.; Rankl, C.; Hinterdorfer, P.; Grubmüller, H.; Herrmann, A. Influenza Virus Binds Its Host Cell Using Multiple Dynamic Interactions. *Proc Natl Acad Sci U S A* **2012**, *109* (34), 13626–13631.
- (33) Varki, A. Sialic Acids as Ligands in Recognition Phenomena. *The FASEB Journal* **1997**, *11* (4), 248–255.
- (34) Varki, N. M.; Varki, A. Diversity in Cell Surface Sialic Acid Presentations: Implications for Biology and Disease. *Laboratory Investigation* **2007**, *87* (9), 851–857.
- (35) Wiley, D. C.; Skehel, J. J. The Structure and Function of the Hemagglutinin Membrane Glycoprotein of Influenza Virus. *Annu Rev Biochem* **2003**, *56*, 365–394.
- (36) Suzuki, Y.; Ito, T.; Suzuki, T.; Robert E. Holland, Jr.; Chambers, T. M.; Kiso, M.; Ishida, H.; Kawaoka, Y. Sialic Acid Species as a Determinant of the Host Range of Influenza A Viruses. *J Virol* **2000**, *74* (24), 11825.
- (37) Benton, D. J.; Nans, A.; Calder, L. J.; Turner, J.; Neu, U.; Lin, Y. P.; Ketelaars, E.; Kallewaard, N. L.; Corti, D.; Lanzavecchia, A.; et al. Influenza Hemagglutinin Membrane Anchor. *Proc Natl Acad Sci U S A* **2018**, *115* (40), 10112–10117.
- (38) Watanabe, A.; McCarthy, K. R.; Kuraoka, M.; Schmidt, A. G.; Adachi, Y.; Onodera, T.; Tonouchi, K.; Caradonna, T. M.; Bajic, G.; Song, S.; et al. Antibodies to a Conserved Influenza Head Interface Epitope Protect by an IgG Subtype-Dependent Mechanism. *Cell* **2019**, *177* (5), 1124–1135.e16.
- (39) Bangaru, S.; Lang, S.; Schotsaert, M.; Vandervlen, H. A.; Zhu, X.; Kose, N.; Bombardi, R.; Finn, J. A.; Kent, S. J.; Gilchuk, P.; et al. A Site of Vulnerability on the Influenza Virus Hemagglutinin Head Domain Trimer Interface. *Cell* **2019**, *177* (5), 1136–1152.e18.
- (40) Bajic, G.; Maron, M. J.; Adachi, Y.; Onodera, T.; McCarthy, K. R.; McGee, C. E.; Sempowski, G. D.; Takahashi, Y.; Kelsoe, G.; Kuraoka, M.; et al. Influenza Antigen Engineering Focuses Immune Responses to a Subdominant but Broadly Protective Viral Epitope. *Cell Host Microbe* **2019**, *25* (6), 827–835.e6.
- (41) Zost, S. J.; Dong, J.; Gilchuk, I. M.; Gilchuk, P.; Thornburg, N. J.; Bangaru, S.; Kose, N.; Finn, J. A.; Bombardi, R.; Soto, C.; et al. Canonical Features of Human Antibodies Recognizing the Influenza Hemagglutinin Trimer Interface. *J Clin Invest* **2021**, *131* (15), e146791.
- (42) McCarthy, K. R.; Lee, J.; Watanabe, A.; Kuraoka, M.; Robinson-Mccarthy, L. R.; Georgiou, G.; Kelsoe, G.; Harrison, S. C. A Prevalent Focused Human Antibody Response to the Influenza Virus Hemagglutinin Head Interface. *mBio* **2021**, *12* (3), e01144-21.
- (43) Zhu, X.; Han, J.; Sun, W.; Puente-Massaguer, E.; Yu, W.; Palese, P.; Krammer, F.; Ward, A. B.; Wilson, I. A. Influenza Chimeric Hemagglutinin Structures in Complex with Broadly Protective Antibodies to the Stem and Trimer Interface. *Proc Natl Acad Sci U S A* **2022**, *119* (21), 1–8.

- (44) Ellis, D.; Lederhofer, J.; Acton, O. J.; Tsybovsky, Y.; Kephart, S.; Yap, C.; Gillespie, R. A.; Creanga, A.; Olshefsky, A.; Stephens, T.; et al. Structure-Based Design of Stabilized Recombinant Influenza Neuraminidase Tetramers. *Nat Commun* **2022**, *13* (1), 1–16.
- (45) Durrant, J. D.; Kochanek, S. E.; Casalino, L.; Jeong, P. U.; Dommer, A. C.; Amaro, R. E. Mesoscale All-Atom Influenza Virus Simulations Suggest New Substrate Binding Mechanism. *ACS Cent Sci* **2020**, *6* (2), 189–196.
- (46) Tarasova, E.; Nerukh, D. All-Atom Molecular Dynamics Simulations of Whole Viruses. *Journal of Physical Chemistry Letters* **2018**, *9* (19), 5805–5809.
- (47) Huber, R. G.; Marzinek, J. K.; Holdbrook, D. A.; Bond, P. J. Multiscale Molecular Dynamics Simulation Approaches to the Structure and Dynamics of Viruses. *Prog Biophys Mol Biol* **2017**, *128*, 121–132.
- (48) Chavent, M.; Duncan, A. L.; Sansom, M. S. P. Molecular Dynamics Simulations of Membrane Proteins and Their Interactions: From Nanoscale to Mesoscale. *Curr Opin Struct Biol* **2016**, *40*, 8–16.
- (49) Marrink, S. J.; Corradi, V.; Souza, P. C. T.; Ingólfsson, H. I.; Tieleman, D. P.; Sansom, M. S. P. Computational Modeling of Realistic Cell Membranes. *Chem Rev* **2019**, *119* (9), 6184–6226.
- (50) Reddy, T.; Shorthouse, D.; Parton, D. L.; Jefferys, E.; Fowler, P. W.; Chavent, M.; Baaden, M.; Sansom, M. S. P. Nothing to Sneeze At: A Dynamic and Integrative Computational Model of an Influenza A Virion. *Structure* **2015**, *23* (3), 584–597.
- (51) Yu, A.; Pak, A. J.; He, P.; Monje-Galvan, V.; Casalino, L.; Gaieb, Z.; Dommer, A. C.; Amaro, R. E.; Voth, G. A. A Multiscale Coarse-Grained Model of the SARS-CoV-2 Virion. *Biophys J* **2021**, *120* (6), 1097–1104.
- (52) Reddy, T.; Sansom, M. S. P. The Role of the Membrane in the Structure and Biophysical Robustness of the Dengue Virion Envelope. *Structure* **2016**, *24* (3), 375–382.
- (53) Casalino, L.; Dommer, A. C.; Gaieb, Z.; Barros, E. P.; Sztain, T.; Ahn, S. H.; Trifan, A.; Brace, A.; Bogetti, A. T.; Clyde, A.; et al. AI-Driven Multiscale Simulations Illuminate Mechanisms of SARS-CoV-2 Spike Dynamics. *Int J High Perform Comput Appl* **2021**, *35* (5), 432–451.
- (54) Perilla, J. R.; Schulten, K. Physical Properties of the HIV-1 Capsid from All-Atom Molecular Dynamics Simulations. *Nat Commun* **2017**, *8* (1), 1–10.
- (55) Freddolino, P. L.; Arkhipov, A. S.; Larson, S. B.; McPherson, A.; Schulten, K. Molecular Dynamics Simulations of the Complete Satellite Tobacco Mosaic Virus. *Structure* **2006**, *14* (3), 437–449.
- (56) Tarasova, E.; Farafonov, V.; Khayat, R.; Okimoto, N.; Komatsu, T. S.; Taiji, M.; Nerukh, D. All-Atom Molecular Dynamics Simulations of Entire Virus Capsid Reveal the Role of Ion Distribution in Capsid's Stability. *Journal of Physical Chemistry Letters* **2017**, *8* (4), 779–784.
- (57) Andoh, Y.; Yoshii, N.; Yamada, A.; Fujimoto, K.; Kojima, H.; Mizutani, K.; Nakagawa, A.; Nomoto, A.; Okazaki, S. All-Atom Molecular Dynamics Calculation Study of Entire Poliovirus Empty Capsids in Solution. *J Chem Phys* **2014**, *141* (16), 165101.
- (58) Tarasova, E.; Korotkin, I.; Farafonov, V.; Karabasov, S.; Nerukh, D. Complete Virus Capsid at All-Atom Resolution: Simulations Using Molecular Dynamics and Hybrid Molecular Dynamics/Hydrodynamics Methods Reveal Semipermeable Membrane Function. *J Mol Liq* **2017**, *245*, 109–114.
- (59) Miao, Y.; Johnson, J. E.; Ortoleva, P. J. All-Atom Multiscale Simulation of Cowpea Chlorotic Mottle Virus Capsid Swelling. *Journal of Physical Chemistry B* **2010**, *114* (34), 11181–11195.
- (60) Hadden, J. A.; Perilla, J. R.; Schlicksup, C. J.; Venkatakrishnan, B.; Zlotnick, A.; Schulten, K. All-Atom Molecular Dynamics of the HBV Capsid Reveals Insights into Biological Function and Cryo-EM Resolution Limits. *Elife* **2018**, *7*, e32478.
- (61) Farafonov, V. S.; Nerukh, D. MS2 Bacteriophage Capsid Studied Using All-Atom Molecular Dynamics. *Interface Focus* **2019**, *9* (3).
- (62) Dommer, A.; Casalino, L.; Kearns, F.; Rosenfeld, M.; Wauer, N.; Ahn, S.-H.; Russo, J.; Oliveira, S.; Morris, C.; Bogetti, A.; et al. #COVIDisAirborne: AI-Enabled Multiscale Computational Microscopy of Delta SARS-CoV-2 in a Respiratory Aerosol. *bioRxiv* **2021**, 2021.11.12.468428.
- (63) Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *J Am Chem Soc* **2018**, *140* (7), 2386–2396.
- (64) Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything You Wanted to Know about Markov State Models but Were Afraid to Ask. *Methods* **2010**, *52* (1), 99–105.

- (65) Chodera, J. D.; Noé, F. Markov State Models of Biomolecular Conformational Dynamics. *Curr Opin Struct Biol* **2014**, 25, 135–144.
- (66) McAuley, J. L.; Gilbertson, B. P.; Trifkovic, S.; Brown, L. E.; McKimm-Breschkin, J. L. Influenza Virus Neuraminidase Structure and Functions. *Front Microbiol* **2019**, 10 (JAN), 39.
- (67) Stadlbauer, D.; Zhu, X.; McMahon, M.; Turner, J. S.; Wohlbold, T. J.; Schmitz, A. J.; Strohmeier, S.; Yu, W.; Nachbagauer, R.; Mudd, P. A.; et al. Broadly Protective Human Antibodies That Target the Active Site of Influenza Virus Neuraminidase. *Science (1979)* **2019**, 366 (6464), 499–504.
- (68) Harris, A.; Cardone, G.; Winkler, D. C.; Heymann, J. B.; Brecher, M.; White, J. M.; Steven, A. C. Influenza Virus Pleiomorphy Characterized by Cryoelectron Tomography. *Proc Natl Acad Sci U S A* **2006**, 103 (50), 19123–19127.
- (69) Zhou, H. X.; Rivas, G.; Minton, A. P. Macromolecular Crowding and Confinement: Biochemical, Biophysical, and Potential Physiological Consequences*. <https://doi.org/10.1146/annurev.biophys.37.032807.125817> **2008**, 37, 375–397.
- (70) Guthmiller, J. J.; Han, J.; Utset, H. A.; Li, L.; Lan, L. Y. L.; Henry, C.; Stamper, C. T.; McMahon, M.; O'Dell, G.; Fernández-Quintero, M. L.; et al. Broadly Neutralizing Antibodies Target a Haemagglutinin Anchor Epitope. *Nature* **2021**, 602 (7896), 314–320.
- (71) Xu, R.; Ekiert, D. C.; Krause, J. C.; Hai, R.; Crowe, J. E.; Wilson, I. A. Structural Basis of Preexisting Immunity to the 2009 H1N1 Pandemic Influenza Virus. *Science (1979)* **2010**, 328 (5976), 357–360.
- (72) Russell, C. J.; Belser, J. A.; Pulit-Penalosa, J. A.; Sun, X. Hemagglutinin Stability and Its Impact on Influenza A Virus Infectivity, Pathogenicity, and Transmissibility in Avians, Mice, Swine, Seals, Ferrets, and Humans. *Viruses* **2021**, 13 (5), 746.
- (73) Brian Dyer, R.; Eller, M. W. Dynamics of Hemagglutinin-Mediated Membrane Fusion. *Proc Natl Acad Sci U S A* **2018**, 115 (35), 8655–8657.
- (74) Wu, Y.; Gao, G. F. “Breathing” Hemagglutinin Reveals Cryptic Epitopes for Universal Influenza Vaccine Design. *Cell* **2019**, 177 (5), 1086–1088.
- (75) Lee, J.; Boutz, D. R.; Chromikova, V.; Joyce, M. G.; Vollmers, C.; Leung, K.; Horton, A. P.; DeKosky, B. J.; Lee, C. H.; Lavinder, J. J.; et al. Molecular-Level Analysis of the Serum Antibody Repertoire in Young Adults before and after Seasonal Influenza Vaccination. *Nat Med* **2016**, 22 (12), 1456–1464.
- (76) Byrd-Leotis, L.; Cummings, R. D.; Steinhauer, D. A. The Interplay between the Host Receptor and Influenza Virus Hemagglutinin and Neuraminidase. *Int J Mol Sci* **2017**, 18 (7), 1541.
- (77) Parton, D. L.; Tek, A.; Baaden, M.; Sansom, M. S. P. Formation of Raft-Like Assemblies within Clusters of Influenza Hemagglutinin Observed by MD Simulations. *PLoS Comput Biol* **2013**, 9 (4), e1003034.
- (78) Takeda, M.; Leser, G. P.; Russell, C. J.; Lamb, R. A. Influenza Virus Hemagglutinin Concentrates in Lipid Raft Microdomains for Efficient Viral Fusion. *Proc Natl Acad Sci U S A* **2003**, 100 (25), 14610–14617.
- (79) Leser, G. P.; Lamb, R. A. Lateral Organization of Influenza Virus Proteins in the Budozone Region of the Plasma Membrane. *J Virol* **2017**, 91 (9), e02104-16.
- (80) Cohen, R. D.; Pielak, G. J. Electrostatic Contributions to Protein Quinary Structure. *J Am Chem Soc* **2016**, 138 (40), 13139–13142.
- (81) Mu, X.; Choi, S.; Lang, L.; Mowray, D.; Dokholyan, N. v.; Danielsson, J.; Oliveberg, M. Physicochemical Code for Quinary Protein Interactions in Escherichia Coli. *Proc Natl Acad Sci U S A* **2017**, 114 (23), E4556–E4563.
- (82) Leser, G. P.; Lamb, R. A. Influenza Virus Assembly and Budding in Raft-Derived Microdomains: A Quantitative Analysis of the Surface Distribution of HA, NA and M2 Proteins. *Virology* **2005**, 342 (2), 215–227.
- (83) Veit, M.; Thaa, B. Association of Influenza Virus Proteins with Membrane Rafts. *Adv Virol* **2011**, 2011, 370606.
- (84) Rickard, M. M.; Zhang, Y.; Gruebele, M.; Pogorelov, T. v. In-Cell Protein-Protein Contacts: Transient Interactions in the Crowd. *Journal of Physical Chemistry Letters* **2019**, 10 (18), 5667–5673.
- (85) Du, R.; Cui, Q.; Rong, L. Competitive Cooperation of Hemagglutinin and Neuraminidase during Influenza A Virus Entry. *Viruses* **2019**, 11 (5), 458.
- (86) Diederich, S.; Berhane, Y.; Embury-Hyatt, C.; Hisanaga, T.; Handel, K.; Cottam-Birt, C.; Ranadheera, C.; Kobasa, D.; Pasick, J. Hemagglutinin-Neuraminidase Balance Influences the Virulence Phenotype of a Recombinant H5N3 Influenza A Virus Possessing a Polybasic HA 0 Cleavage Site. *J Virol* **2015**, 89 (21), 10724–10734.

- (87) Yen, H. L.; Liang, C. H.; Wu, C. Y.; Forrest, H. L.; Ferguson, A.; Choy, K. T.; Jones, J.; Wong, D. D. Y.; Cheung, P. P. H.; Hsu, C. H.; et al. Hemagglutinin-Neuraminidase Balance Confers Respiratory-Droplet Transmissibility of the Pandemic H1N1 Influenza Virus in Ferrets. *Proc Natl Acad Sci U S A* **2011**, *108* (34), 14264–14269.
- (88) Isin, B.; Doruker, P.; Bahar, I. Functional Motions of Influenza Virus Hemagglutinin: A Structure-Based Analytical Approach. *Biophys J* **2002**, *82* (2), 569–581.
- (89) Rantalainen, K.; Berndsen, Z. T.; Antanasijevic, A.; Schiffner, T.; Zhang, X.; Lee, W. H.; Torres, J. L.; Zhang, L.; Irimia, A.; Copps, J.; et al. HIV-1 Envelope and MPER Antibody Structures in Lipid Assemblies. *Cell Rep* **2020**, *31* (4), 107583.
- (90) Turoňová, B.; Sikora, M.; Schürmann, C.; Hagen, W. J. H.; Welsch, S.; Blanc, F. E. C.; von Bülow, S.; Gecht, M.; Bagola, K.; Hörner, C.; et al. In Situ Structural Analysis of SARS-CoV-2 Spike Reveals Flexibility Mediated by Three Hinges. *Science* **2020**, *370*, 203–208.
- (91) Das, D. K.; Govindan, R.; Nikić-Spiegel, I.; Krammer, F.; Lemke, E. A.; Munro, J. B. Direct Visualization of the Conformational Dynamics of Single Influenza Hemagglutinin Trimers. *Cell* **2018**, *174* (4), 926–937.e12.
- (92) Wu, Y.; Gao, G. F. “Breathing” Hemagglutinin Reveals Cryptic Epitopes for Universal Influenza Vaccine Design. *Cell* **2019**, *177* (5), 1086–1088.
- (93) Dowd, K. A.; Pierson, T. C. The Many Faces of a Dynamic Virion: Implications of Viral Breathing on Flavivirus Biology and Immunogenicity. *Annu Rev Virol* **2018**, *5*, 185–207.
- (94) Ozorowski, G.; Pallesen, J.; de Val, N.; Lyumkis, D.; Cottrell, C. A.; Torres, J. L.; Copps, J.; Stanfield, R. L.; Cupo, A.; Pugach, P.; et al. Open and Closed Structures Reveal Allostery and Pliability in the HIV-1 Envelope Spike. *Nature* **2017**, *547* (7663), 360–363.
- (95) Byrne, P. O.; McLellan, J. S. Principles and Practical Applications of Structure-Based Vaccine Design. *Curr Opin Immunol* **2022**, *77*, 102209.
- (96) Wan, H.; Gao, J.; Yang, H.; Yang, S.; Harvey, R.; Chen, Y. Q.; Zheng, N. Y.; Chang, J.; Carney, P. J.; Li, X.; et al. The Neuraminidase of A(H3N2) Influenza Viruses Circulating since 2016 Is Antigenically Distinct from the A/Hong Kong/4801/2014 Vaccine Strain. *Nature Microbiology* **2019**, *4* (12), 2216–2225.
- (97) Guthmiller, J. J.; Han, J.; Utset, H. A.; Li, L.; Lan, L. Y. L.; Henry, C.; Stamper, C. T.; McMahon, M.; O’Dell, G.; Fernández-Quintero, M. L.; et al. Broadly Neutralizing Antibodies Target a Haemagglutinin Anchor Epitope. *Nature* **2021**, *602* (7896), 314–320.
- (98) Ziebert, F.; Kulić, I. M. How Influenza’s Spike Motor Works. *Phys Rev Lett* **2021**, *126* (21), 218101.
- (99) Sakai, T.; Nishimura, S. I.; Naito, T.; Saito, M. Influenza A Virus Hemagglutinin and Neuraminidase Act as Novel Motile Machinery. *Scientific Reports* **2017**, *7* (1), 1–11.
- (100) Air, G. M. Influenza Neuraminidase. *Influenza Other Respir Viruses* **2012**, *6* (4), 245.
- (101) Xu, R.; McBride, R.; Nycholat, C. M.; Paulson, J. C.; Wilson, I. A. Structural Characterization of the Hemagglutinin Receptor Specificity from the 2009 H1N1 Influenza Pandemic. *J Virol* **2012**, *86* (2), 982–990.
- (102) Varki, A.; Cummings, R. D.; Aebi, M.; Packer, N. H.; Seeberger, P. H.; Esko, J. D.; Stanley, P.; Hart, G.; Darvill, A.; Kinoshita, T.; et al. Symbol Nomenclature for Graphical Representations of Glycans. *Glycobiology* **2015**, *25* (12), 1323–1324.
- (103) Watanabe, Y.; Berndsen, Z. T.; Raghvani, 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), 1–10.
- (104) Watanabe, Y.; Bowden, T. A.; Wilson, I. A.; Crispin, M. Exploitation of Glycosylation in Enveloped Virus Pathobiology. *Biochimica et Biophysica Acta (BBA) - General Subjects* **2019**, *1863* (10), 1480–1497.
- (105) Schulze, I. T. Effects of Glycosylation on the Properties and Functions of Influenza Virus Hemagglutinin. *J Infect Dis* **1997**, *176* (Supplement_1), S24–S28.
- (106) Tate, M. D.; Job, E. R.; Deng, Y. M.; Gunalan, V.; Maurer-Stroh, S.; Reading, P. C. Playing Hide and Seek: How Glycosylation of the Influenza Virus Hemagglutinin Can Modulate the Immune Response to Infection. *Viruses* **2014**, *6* (3), 1294–1316.
- (107) Job, E. R.; Deng, Y.-M.; Barfod, K. K.; Tate, M. D.; Caldwell, N.; Reddiex, S.; Maurer-Stroh, S.; Brooks, A. G.; Reading, P. C. Addition of Glycosylation to Influenza A Virus Hemagglutinin Modulates Antibody-Mediated Recognition of H1N1 2009 Pandemic Viruses. *The Journal of Immunology* **2013**, *190* (5), 2169–2177.

- (108) Sun, X.; Jayaraman, A.; Maniprasad, P.; Raman, R.; Houser, K. v.; Pappas, C.; Zeng, H.; Sasisekharan, R.; Katz, J. M.; Tumpey, T. M. N-Linked Glycosylation of the Hemagglutinin Protein Influences Virulence and Antigenicity of the 1918 Pandemic and Seasonal H1N1 Influenza A Viruses. *J Virol* **2013**, *87* (15), 8756–8766.
- (109) Medina, R. A.; Stertz, S.; Manicassamy, B.; Zimmermann, P.; Sun, X.; Albrecht, R. A.; Uusi-Kerttula, H.; Zagordi, O.; Belshe, R. B.; Frey, S. E.; et al. Glycosylations in the Globular Head of the Hemagglutinin Protein Modulate the Virulence and Antigenic Properties of the H1N1 Influenza Viruses. *Sci Transl Med* **2013**, *5* (187), 187ra70.
- (110) 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 Cent Sci* **2020**, *6* (10), 1722–1734.
- (111) Miller, N. L.; Clark, T.; Raman, R.; Sasisekharan, R. Glycans in Virus-Host Interactions: A Structural Perspective. *Front Mol Biosci* **2021**, *8*, 666756.
- (112) Wang, C. C.; Chen, J. R.; Tseng, Y. C.; Hsu, C. H.; Hung, Y. F.; Chen, S. W.; Chen, C. M.; Khoo, K. H.; Cheng, T. J.; Cheng, Y. S. E.; et al. Glycans on Influenza Hemagglutinin Affect Receptor Binding and Immune Response. *Proc Natl Acad Sci U S A* **2009**, *106* (43), 18137–18142.
- (113) Seitz, C.; Casalino, L.; Konecny, R.; Huber, G.; Amaro, R. E.; McCammon, J. A. Multiscale Simulations Examining Glycan Shield Effects on Drug Binding to Influenza Neuraminidase. *Biophys J* **2020**, *119* (11), 2275–2289.
- (114) Daniels, R.; Kurowski, B.; Johnson, A. E.; Hebert, D. N. N-Linked Glycans Direct the Cotranslational Folding Pathway of Influenza Hemagglutinin. *Mol Cell* **2003**, *11* (1), 79–90.
- (115) Deshpande, K. L.; Fried, V. A.; Ando, M.; Webster, R. G. Glycosylation Affects Cleavage of an H5N2 Influenza Virus Hemagglutinin and Regulates Virulence. *Proc Natl Acad Sci U S A* **1987**, *84* (1), 36–40.
- (116) Amaro, R. E.; Jeong, P. U.; Huber, G.; Dommer, A.; Steven, A. C.; Bush, R. M.; Durrant, J. D.; Votapka, L. W. A Computational Assay That Explores the Hemagglutinin/Neuraminidase Functional Balance Reveals the Neuraminidase Secondary Site as a Novel Anti-Influenza Target. *ACS Cent Sci* **2018**, *4* (11), 1570–1577.
- (117) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J Comput Chem* **2005**, *26* (16), 1781–1802.
- (118) Huang, J.; Mackerell, A. D. CHARMM36 All-Atom Additive Protein Force Field: Validation Based on Comparison to NMR Data. *J Comput Chem* **2013**, *34* (25), 2135–2145.
- (119) 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.
- (120) Danne, R.; Poojari, C.; Martinez-Seara, H.; Rissanen, S.; Lolicato, F.; Róg, T.; Vattulainen, I. DoGlycans-Tools for Preparing Carbohydrate Structures for Atomistic Simulations of Glycoproteins, Glycolipids, and Carbohydrate Polymers for GROMACS. *J Chem Inf Model* **2017**, *57* (10), 2401–2406.
- (121) Cruz, E.; Cain, J.; Crossett, B.; Kayser, V. Site-Specific Glycosylation Profile of Influenza A (H1N1) Hemagglutinin through Tandem Mass Spectrometry. *Hum Vaccin Immunother* **2017**, *14* (3), 508–517.
- (122) Khatri, K.; Klein, J. A.; White, M. R.; Grant, O. C.; Leymarie, N.; Woods, R. J.; Hartshorn, K. L.; Zaia, J. Integrated Omics and Computational Glycobiology Reveal Structural Basis for Influenza a Virus Glycan Microheterogeneity and Host Interactions. *Molecular and Cellular Proteomics* **2016**, *15* (6), 1895–1912.
- (123) Liu, Y. J.; Wu, S. L.; Love, K. R.; Hancock, W. S. Characterization of Site-Specific Glycosylation in Influenza A Virus Hemagglutinin Produced by Spodoptera Frugiperda Insect Cell Line. *Anal Chem* **2017**, *89* (20), 11036–11043.
- (124) She, Y. M.; Farnsworth, A.; Li, X.; Cyr, T. D. Topological N-Glycosylation and Site-Specific N-Glycan Sulfation of Influenza Proteins in the Highly Expressed H1N1 Candidate Vaccines. *Sci Rep* **2017**, *7* (1), 10232.
- (125) Blake, T. A.; Williams, T. L.; Pirkle, J. L.; Barr, J. R. Targeted N-Linked Glycosylation Analysis of H5N1 Influenza Hemagglutinin by Selective Sample Preparation and Liquid Chromatography/Tandem Mass Spectrometry. *Anal Chem* **2009**, *81* (8), 3109–3118.
- (126) Klauda, J. B.; Venable, R. M.; Freites, J. A.; O'Connor, J. W.; Tobias, D. J.; Mondragon-Ramirez, C.; Vorobyov, I.; MacKerell, A. D.; Pastor, R. W. Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types. *Journal of Physical Chemistry B* **2010**, *114* (23), 7830–7843.
- (127) Klauda, J. B.; Monje, V.; Kim, T.; Im, W. Improving the CHARMM Force Field for Polyunsaturated Fatty Acid Chains. *Journal of Physical Chemistry B* **2012**, *116* (31), 9424–9431.

- (128) Guvench, O.; Hatcher, E.; Venable, R. M.; Pastor, R. W.; MacKerell, A. D. CHARMM Additive All-Atom Force Field for Glycosidic Linkages between Hexopyranoses. *J Chem Theory Comput* **2009**, 5 (9), 2353–2370.
- (129) Beglov, D.; Roux, B. Finite Representation of an Infinite Bulk System: Solvent Boundary Potential for Computer Simulations. *J Chem Phys* **1998**, 100 (12), 9050.
- (130) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J Chem Phys* **1998**, 79 (2), 926.
- (131) Phillips, J. C.; Hardy, D. J.; Maia, J. D. C.; Stone, J. E.; Ribeiro, J. v.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W.; et al. Scalable Molecular Dynamics on CPU and GPU Architectures with NAMMD. *J Chem Phys* **2020**, 153 (4), 044130.
- (132) Brünger, A.; Brooks, C. L.; Karplus, M. Stochastic Boundary Conditions for Molecular Dynamics Simulations of ST2 Water. *Chem Phys Lett* **1984**, 105 (5), 495–500.
- (133) Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant Pressure Molecular Dynamics Algorithms. *J Chem Phys* **1998**, 101 (5), 4177.
- (134) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. Constant Pressure Molecular Dynamics Simulation: The Langevin Piston Method. *J Chem Phys* **1998**, 103 (11), 4613.
- (135) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *J Chem Phys* **1998**, 98 (12), 10089.
- (136) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J Comput Phys* **1977**, 23 (3), 327–341.
- (137) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J Mol Graph* **1996**, 14 (1), 33–38.
- (138) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J. H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J Chem Theory Comput* **2015**, 11 (11), 5525–5542.
- (139) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; de Fabritiis, G.; Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *J Chem Phys* **2013**, 139 (1), 015102.
- (140) Molgedey, L.; Schuster, H. G. Separation of a Mixture of Independent Signals Using Time Delayed Correlations. *Phys Rev Lett* **1994**, 72 (23), 3634.
- (141) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J Chem Theory Comput* **2013**, 9 (4), 2000–2009.
- (142) Bock, H.-H. Clustering Methods: A History of k-Means Algorithms. **2007**, 161–172.
- (143) Prinz, J. H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov Models of Molecular Kinetics: Generation and Validation. *J Chem Phys* **2011**, 134 (17), 174105.
- (144) Röblitz, S.; Weber, M. Fuzzy Spectral Clustering by PCCA+: Application to Markov State Models and Data Classification. *Adv Data Anal Classif* **2013**, 7 (2), 147–179.

4.9 Supporting information

Mesoscale all-atom modeling of the influenza virus

In the work presented here, two glycosylated, whole-virion models of the H1N1 influenza A virus were built, simulated, and analyzed. The first corresponds to the strain A/swine/Shandong/N1/2009(H1N1) (hereafter referred to as H1N1-Shan2009), whereas the second one accounts for the A/45/Michigan/2015(H1N1) strain (hereafter referred to as H1N1-Mich2015). Both models are based on the unglycosylated H1N1-Shan2009 whole-virion model previously built by us and presented in Durrant et al.¹ The main difference lies in the addition of N-linked glycans to HA and NA glycoproteins, which substantially improved the biological relevance of the simulated constructs. Although we refer to Durrant et al.¹ for a detailed description of the original modeling of the individual protein, assembly procedure of the whole-virion model, and molecular dynamics (MD) simulations of both the individual proteins (HA, NA, M2) and the whole-virion model, a summary is also provided in the following paragraph.

The unglycosylated, H1N1-Shan2009 whole-virion model constructed and simulated by Durrant et al.¹ comprises 236 HA trimers, 30 NA tetramers, and 11 M2 ion channels embedded in a 3-palmitoyl-2-oleoyl-dglycero-1-phosphatidylcholine (POPC) quasi-spherical lipid bilayer.¹ HA, NA and M2 were individually modeled and separately equilibrated (HA and NA) through MD simulations before being incorporated into the viral membrane.¹ We note that in Durrant et al.¹ glycans were not modeled. LipidWrapper² was used to create the spherical lipid bilayer starting from a large planar POPC-bilayer model generated with the CHARMM-GUI server.^{3,4} As specified in ref. ¹, slight differences in the leaflet densities are magnified by several times in the mesoscale systems, possibly creating large instabilities during simulation.⁵ Although a bilayer made of only POPC represents an oversimplification of the actual membrane composition,⁶ it has the advantage of minimizing such instabilities during simulations. This was the main reason behind the decision to build a POPC-only lipid bilayer. The M1 matrix proteins, coating the inner leaflet of the lipid bilayer, and the ribonucleoproteins contained inside of the virion were not modeled. As described in Amaro et al.⁷, the shape of the virion and the distribution of the glycoproteins in the viral membrane were adopted from a point model of the virion exterior provided by collaborators, which was in turn derived from a cryo-electron tomography (cryoET) map of the influenza virus pleiomorphy.⁸ Positioning of the glycoproteins and insertion of M2 channels in the lipid bilayer

according to the point model was done with PyMolecule.^{9,10} Full procedure, including handling of clashes, is described in refs. ^{1,7}. The whole-virion model was solvated inside and outside of the lipid bilayer with explicit TIP3P water molecules,¹¹ and ionized at 150 mM concentration with Na⁺/Cl⁻, tallying 160,653,271 atoms. All-atom MD simulations were performed on the Blue Waters supercomputer using NAMD2.10¹² and CHARMM36 all-atom additive force fields for protein,¹³ lipids,^{14,15} and ions.¹⁶

In the next paragraphs, we detail the system setup and all-atom MD simulations of the glycosylated H1N1-Shan2009 and H1N1-Mich2015 whole-virion models presented in this work.

System setup of the glycosylated H1N1-Shan2009 whole-virion model

UniProtKB accession codes for the sequences of HA and NA modeled in H1N1-Shan2009 are F2YI86 and F2YI87, respectively. HA was modeled in its uncleaved form (HA0), with residues ranging from 1 to 566. NA residues range from 1 to 469. We note that for HA and NA we adopted the absolute numbering system that starts with the first residue of the sequence in the corresponding UniProtKB entry. Sequences are reported in Fig. S2 and Fig. S3. The M2 construct¹ was based on the 2L0J structure,¹⁷ tallying 41 residues (from 22 to 62) since N-terminal and C-terminal tails were not modeled.

Initial coordinates for the setup of the glycosylated whole-virion model of H1N1-Shan2009 were derived from the last frame of the all-atom MD simulation of the unglycosylated whole-virion model previously performed by us and briefly described in the previous paragraph.¹ Specifically, we used the last frame obtained from the third simulation branch, which is depicted in green in the schematic reported in the supplementary information file of ref. ¹. At that point of the simulation, corresponding to 69.74 ns, lipid-bilayer instabilities had been already resolved and the system had been steadily simulated without major issues. Our goal was to improve the simulated system by adding N-linked glycans to HA and NA while preserving the relaxed, MD-derived morphology of the virion, thus facilitating a smooth continuation of the simulation endeavor. For this reason, we retained the MD-derived coordinates of all the atoms of the system, including water molecules and ions, and then performed glycosylation of HA and NA accordingly (**Fig. 4.11**). Specific oligo-mannose, hybrid, and complex N-linked glycans were identified to be linked to HA and NA as described in the following paragraphs.

Each H1N1-Shan2009 HA monomer exhibits six N-linked sequons (**Fig. 4.12**), of which one (N104) is located within the head and five (N28, N40, N293, N304, N498) within the stalk,¹⁸ for a total of 18 potential N-linked

glycans per trimer. It must be noted that N27 is also a potential glycosylation site; however, as reported in previous studies,^{18,19} it is not glycosylated as N-X-T sequon at residue 28 is preferentially glycosylated over the N-X-S sequence at residue 27. N293 is located on the edge between the head and the stalk, but it is considered part of the stalk.^{20,21} Several glycomics studies were used as a reference to build the glycosylation profile of HA,^{18,22–24} which is shown in **Fig. 4.12**. Instead, each H1N1-Shan2009 NA monomer presents eight N-linked sequons (**Fig. 4.12**), of which four are located within the head (N88, N146, N235, N386) and four within the stalk (N50, N58, N63, N68),¹⁸ for a total of 32 potential N-linked glycans per tetramer. Several glycomics studies were used as a reference to build the glycosylation profile of NA,^{18,19,25} which is displayed in **Fig. 4.12**.

Upon defining the respective glycosylation profiles, N-glycans were iteratively added to all the 236 HA trimers and 30 NA tetramers using the doGlycans tool.²⁶ VMD,²⁷ NAMD,²⁸ GROMACS²⁹ and the tools therein contained were also used in the glycosylation workflow for pulling each glycoprotein from the virion, assessing sequon accessibility, renaming residues, local minimization, and parameterization. HA and NA N-linked sequons are overall all highly occupied.^{22,30,31} However, since glycans were added to a construct resulting from 69.74 ns of MD,¹ many sequons occurred to be sterically hindered by adjacent residues or not fully exposed to the solvent as a result of the previous dynamics. Moreover, since the viral membrane was originally densely studded with unglycosylated HAs and NAs,^{7,8} adding glycans to sequons located at the interface between adjacent glycoproteins was often not possible. For these reasons, all the sequons were subjected to a preliminary screening to evaluate their accessibility. The accessible surface area (ASA) of the asparagine side chain's ND2 atom was calculated to determine whether a sequon was accessible in our initial model. Both scenarios, where HD21 or HD22 were removed from ND2, were considered (**Fig. 4.11A**). ASA was calculated with the *measure sasa* command implemented in VMD,²⁷ combined with in-house scripts. A sphere of radius 10 Å, approximating the size of the basic (GlcNAc)₂(Man)₃ N-core present in all glycans (**Fig. 4.11B**), was used as a probe for the ASA screening. If the ASA of the putative asparagine's ND2 atom was larger than 10 Å², then a glycan would be tentatively linked (**Fig. 4.11C**). Clashes with surrounding amino acids or other glycans were subsequently examined and solved, first by performing a local minimization with NAMD2.13²⁸ and then by eliminating the clashing glycan if the clashes could not be solved with minimization. Once the glycosylated glycoproteins were positioned back in the viral membrane, possible clashes arising between the glycans and the neighboring

(glycosylated) glycoproteins were examined, resulting in the depletion of the clashing glycans. This further reduced the total number of glycans that were added. A total of 1521 glycans (29.2% of all the available sequons), among which 1277 were linked to HA (30.0%) and 244 to NA (25.4%) (**Fig. 4.12**). We note that we could only populate 29 out of the 708 available N104 sequons within the HA head. This low number is due to the overall limited accessibility, in our system, of residue N104, which is occluded by neighboring residues. A similar scenario related to the modeling of HA's N104 glycosylation site is also reported in ref. ¹⁸.

Next, we adjusted the orientation of the 11 M2 ion channels within the lipid bilayer, positioning the N-ter end extravirion and the C-ter end intravirion,^{32,33} since they had been oriented flipped in the original system.¹ No alterations in the composition and the coordinates of the POPC lipid bilayer were done with respect to the original system.¹ The resulting glycosylated H1N1-Shan2009 virion construct, comprising 236 glycosylated HA homo-trimers, 30 glycosylated NA homo-tetramers, and 11 M2 homo-tetramers embedded in the POPC lipid bilayer, was parameterized using NAMD's PSFGEN²⁸ and CHARMM36 all-atom additive force fields for protein,¹³ lipids,^{14,15} glycans,³⁴ and ions.¹⁶ As mentioned above, explicit TIP3 water molecules¹¹ located inside and outside of the lipid bilayer, Na⁺ and Cl⁻ ions, as well as Ca²⁺ ions bound to the NA heads, were retained from the original unglycosylated system.¹ The addition of glycans altered the overall electric charge of the solute, thus requiring tweaking of the total number of Na⁺ and Cl⁻ ions to ensure electrical neutrality. These modifications did not impact the overall ionic strength, which remained at ~150 mM. The following minimization steps solved eventual clashes between newly added glycans and pre-existing water molecules and ions. The total number of atoms for the final system is 160,919,594, with an orthorhombic periodic cell of 1142 Å × 1194 Å × 1155 Å. After generating suitable files for mesoscale simulations with NAMD,³⁵ the system was ready to undergo all-atom MD simulations.

System setup of the glycosylated H1N1-Mich2015 whole-virion model

UniProtKB accession codes for the sequences of HA and NA modeled in H1N1-Mich2015 are A0A144YDV8 and A0A0X9QTS2, respectively. HA was modeled in its uncleaved form (HA0), with residues ranging from 1 to 566. NA residues range from 1 to 469. We note that for HA and NA, we adopted the absolute numbering system that starts with the first residue of the sequence in the corresponding UniProtKB entry. Sequences are

reported in Fig. S2 and Fig. S3. The M2 construct¹ was based on the 2L0J structure,¹⁷ tallying 41 residues (from 22 to 62) since N-terminal and C-terminal tails were not modeled.

Building on the same initial coordinates obtained from the MD simulation of the unglycosylated virion,¹ the system setup for H1N1-Mich2015 was conducted in the same way as described for H1N1-Shan2009. The main challenge was introducing the H1N1-Mich2015 point mutations occurring in HA and NA. All the mutations within HA (15 per monomer) and NA (14 per monomer) were modeled onto the initial construct using NAMD's PSFGEN²⁸ *mutate* command and the CHARMM36 amino acid topologies to build side chain rotamers.¹³ The mutations are listed and displayed in Fig. S6. No mutations were introduced to M2 ion channels, which were treated and positioned in the lipid bilayer in the same way as in H1N1-Shan2009. The resulting, mutated models of HA and NA were then subjected to the same glycosylation procedure adopted for H1N1-Shan2009 and described above.

S179N mutation introduced a second glycan within the HA head (**Fig. 4.13**), in addition to the already existent N104. Therefore, each H1N1-Mich2015 HA monomer exhibits seven N-linked sequons (**Fig. 4.15**), of which two (N104, N179) are located within the head and five (N28, N40, N293, N304, N498) within the stalk,¹⁸ for a total of 21 potential N-linked glycans per trimer. The same glycoprofile as in H1N1-Shan2009 was adopted.^{18,22–24} The new glycan linked to N179 was modeled as a complex tri-antennary (Fig. S8). N386K mutation ablated the N386 sequon within the NA head (**Fig. 4.14**). Therefore, each H1N1-Mich2015 NA monomer presents seven N-linked sequons (**Fig. 4.15**), of which three are located within the head (N88, N146, N235) and four within the stalk (N50, N58, N63, N68),¹⁸ for a total of 28 potential N-linked glycans per tetramer. The same glycoprofile as in H1N1-Shan2009 was adopted except for the deleted N386 glycan (**Fig. 4.15**).^{18,19,25}

A total of 1791 glycans (30.9 % of all the available sequons) were added to H1N1-Mich2015, among which 1626 were linked to HA (32.8%) and 165 to NA (19.6%) (**Fig. 4.15**). We note that also for H1N1-Mich2015 we could populate a low number of N104 sequons within the HA head (33 of 708). Instead, we could populate 351 sequons out of 708 available (49.6%) for the new N179 glycan within the HA head (**Fig. 4.15**).

Similar to H1N1-Shan2009, the resulting glycosylated H1N1-Mich2015 virion construct was parameterized using NAMD's PSFGEN²⁸ and CHARMM36 all-atom additive force fields for protein,¹³ lipids,^{14,15} glycans,³⁴ and ions.¹⁶ As for H1N1-Shan2009, explicit TIP3P water molecules¹¹ located inside and outside of the lipid

bilayer, Na^+ and Cl^- ions, as well as Ca^{2+} ions bound to the NA heads, were retained from the original unglycosylated system.¹ The amino acid mutations and the addition of glycans altered the overall electric charge of the solute, thus requiring further tweaking of the total number of Na^+ and Cl^- ions to establish electrical neutrality. These modifications did not impact the overall ionic strength, which remained at ~ 150 mM. The following minimization steps solved eventual clashes involving mutated residues or between newly added glycans and pre-existing water molecules and ions. The total number of atoms for the final system is 160,981,954, with an orthorhombic periodic cell of $1138 \text{ \AA} \times 1189 \text{ \AA} \times 1150 \text{ \AA}$. After generating suitable files for mesoscale simulations with NAMD,³⁵ the system was ready to undergo all-atom MD simulations.

All-atom molecular dynamics simulations

All-atom MD simulations of H1N1-Shan2009 and H1N1-Mich2015 were performed on the TITAN supercomputer at Oak Ridge National Laboratory (ORNL) and the NSF Blue Waters supercomputer at the University of Illinois at Urbana-Champaign, respectively, using the CUDA memory-optimized version of NAMD 2.13.¹² The same simulation protocol was adopted for both H1N1-Shan2009 and H1N1-Mich2015. CHARMM36 all-atom additive force fields were used to describe proteins,¹³ lipids,^{14,15} glycans,³⁴ and ions.¹⁶ Simulations were performed in explicit water using the TIP3P model.¹¹

Preliminary tests were initially run to assess the stability of the systems, familiarize with the computing facilities, and debug possible errors. Then, the systems were initially minimized in two consecutive cycles of 15,000 steps, each using the NAMD's conjugate gradient energy approach. During the first cycle, the position of protein backbone atoms and of POPC P atoms within $15 \text{ \AA}$ of the M2 ion channels was kept harmonically restrained at $100 \text{ kcal/mol/\AA}^2$. We note that the orientation of M2 ion channels was modified with respect to our previous work,¹ thus requiring more caution. We also remark that the initial coordinates for water molecules, ions, lipids, and proteins (except for M2) were retained from a previous MD simulation as described above,¹ thus being already equilibrated. During the second minimization cycle, the harmonic restraints on the protein backbone atoms were released, whereas the harmonic restraints on the POPC P atoms within $15 \text{ \AA}$ of M2 ion channels were decreased to $10 \text{ kcal/mol/\AA}^2$. After minimization, the systems were gradually heated up with no restraints from 0 K to 298 K in 220 ps using isothermal–isobaric conditions (NPT ensemble). The temperature was

controlled using the Langevin thermostat³⁶ with a damping coefficient of 5/ps, whereas the Nosé-Hoover Langevin piston^{37,38} was used to maintain the pressure at 1.01325 bar. A time-step of 2 fs was used. Next, the systems were equilibrated for 0.5 ns under the NPT ensemble. Temperature (298 K) and pressure (1.01325 bar) control were achieved again using the Langevin thermostat³⁶ with a damping coefficient of 5/ps, and the Nosé-Hoover Langevin piston.^{37,38} During this phase, harmonic restraints at 100 kcal/mol/Å² were applied again to POPC P atoms within 15 Å of M2 ion channels and gradually released to 10 kcal/mol/Å².

Finally, the systems were submitted to productive MD simulations in NPT conditions using the same barostat and thermostat as above to achieve pressure (1.01325 bar) and temperature (298 K) control. Langevin damping coefficient was set to 1/ps during productive MD. One continuous replica was performed for each system using a time-step of 2 fs. To preserve the overall geometry of the virion models in the absence of the interior component, we applied a harmonic position restraint of 20 kcal/mol/Å² to every 20th inner-leaflet P atom. A similar protocol was adopted in our previous virion simulation,¹ where instead every 10th POPC head group was kept fixed. Non-bonded interactions (van der Waals and short-range electrostatic) were calculated at each timestep using a cutoff of 12 Å and a switching distance of 10 Å. All simulations were performed using periodic boundary conditions, employing the particle-mesh Ewald method³⁹ with a grid spacing of 2.1 Å to evaluate long-range electrostatic interactions every three timesteps. SHAKE algorithm⁴⁰ was adopted to keep the atomic bonds involving hydrogens fixed with the NAMD option *rigidbonds all*. Frames were saved every 30,000 steps (60 ps).

For H1N1-Shan2009, we collected a total of ~441.78 ns (7363 frames) continuous productive MD. Simulations were performed with the CUDA memory-optimized version of NAMD 2.13¹² running on 4,096 physical nodes (28,672 processors) on the TITAN supercomputer, with an average benchmark of ~13.83 ns/day. A total of ~17 TB of data were generated, including stripped trajectories.

For H1N1-Mich2015, we collected a total of 424.98 ns (7083 frames) continuous productive MD. Simulations were performed with the CUDA memory-optimized version of NAMD 2.13¹² running on 4,096 physical nodes (114,688 processors) on the Blue Waters supercomputer, with an average benchmark of ~13.78 ns/day. A total of ~16 TB of data was generated including stripped trajectories.

Root Mean Squared Deviation (RMSD) analysis of HA and NA

For both strains, each of the 236 HA trimeric and 30 NA tetrameric trajectories was separately aligned using as a reference the respective initial coordinates of all the backbone atoms (C, N, O, C α). Then, for each of the HA trimers/NA tetramers, we calculated the RMSD of the backbone atoms using the *measure rmsd* command implemented in VMD.²⁷ Subsequently, we computed the mean and standard deviation of the RMSD values obtained at each frame across all the 236 HA trimers / 30 NA tetramers, generating, respectively, the averaged RMSD profiles for HA and NA (**Fig. 4.16**).

Root Mean Squared Fluctuation (RMSF) analysis of HA and NA residues

We performed the calculation of C α -atom RMSF to probe the average mobility of HA and NA residues. Before the calculation of C α -atom RMSF, rotation and translation motions of each individual HA trimer and NA tetramer were removed through targeted, specific alignments. In detail, two sets of aligned trajectories, based on different residue masks, were generated for each of the 236 HA trimers and 30 NA tetramers of both strains. For each of the HA trimers, the first set of alignment was performed using as a reference the respective initial coordinates of all the backbone atoms (C, N, O, C α) of the ectodomain region (residues 18 to 503) of the respective glycoprotein, whereas the second set was based on the region comprising the transmembrane domain (TMD) (residues 529 to 549). The former served for the calculation of RMSF of residues 1 to 503, the latter for the calculation of RMSF of residues 504 to 566. For each of the NA tetramers, the first set of alignment was performed using as a reference the respective initial coordinates of all the backbone atoms (C, N, O, C α) of the head region (residues 96 to 461) of the respective glycoprotein, whereas the second set was based on the region comprising the stalk and the TMD (residues 7 to 81). The former served for the calculation of RMSF of residues 96 to 469, the latter for the calculation of RMSF of residues 1 to 95. This procedure was performed due to peculiar structures of HA and NA that present domains undergoing tilting motions during the dynamics, i.e., HA ectodomain tilting and NA head tilting. If we had performed a single alignment based on the backbone atoms of the whole structure, the extensive motions of the HA ectodomain and NA head, which are larger in size and account for most of the residues, would have biased the RMSF values of the complementary region encompassing the hinge linker (both for HA and NA), the stalk (for NA), the TMD, and the cytosolic tail.

For each trajectory of both sets and for each monomer, we calculated the RMSF of the C α atoms of the residues belonging to the respective region using the *measure rmsf* command implemented in VMD.²⁷ All the frames were used for the calculation. We then combined the RMSF values obtained for the two regions into a single RMSF profile for each monomer. Finally, we computed the mean and standard deviation of the RMSF values obtained for each C α atom across all the 708 HA monomers / 120 NA monomers, generating the respectively averaged RMSF profiles for HA and NA (**Fig. 4.17**).

Positional fluctuations of HA's and NA's TMD and ectodomain center of masses

For both H1N1-Shan2009 and H1N1-Mich2015, we calculated the RMSF of HA's and NA's TMD center of mass (COM) and ectodomain COM using the first frame as a reference. The analysis was performed using in-house tcl scripts within the VMD framework.²⁷ We note that the aim of this analysis was estimating the average positional fluctuations of the TMD COM and ectodomain COM of each glycoprotein within the virion, therefore only the global rotation and translation motions of the whole virion were removed and not those of each individual glycoprotein within the virion. Global rotation and translation motions of the whole virion coordinates during dynamics were removed by aligning the trajectory onto the initial coordinates of all the protein backbone atoms (C, N, O, C α). The COM of HA TMD was defined using the coordinates of backbone atoms of residues 529 to 549. The COM of HA ectodomain was defined using the coordinates of backbone atoms of residues 1 to 528. The COM of NA TMD was defined using the coordinates of backbone atoms of residues 7 to 29. The COM of NA ectodomain was defined using the coordinates of backbone atoms of residues 30 to 469.

At each frame of the whole virion trajectory, we calculated, for each glycoprotein, the squared positional fluctuation of both COMs with respect to their initial position (reference). Next, always for each glycoprotein, we averaged the squared fluctuations of the COMs and calculated the square root of the averages, obtaining the respective RMSF value.

NA head-tilt angle

The NA head-tilt angle was calculated for all 30 NA tetramers at each frame of both virion trajectories using in-house tcl scripts within the VMD framework.²⁷ The NA head-tilt angle is defined by the principal axis of the tetrameric upper stalk region (residues 50 to 78) and the vector joining residue 78 at the top of the stalk with the COM of the tetrameric head (residues 91 to 469). The distribution of NA head-tilt angle values was computed through kernel density estimation using Gaussian kernels with gnuplot plotting program.⁴¹

HA ectodomain-tilt angle

The HA ectodomain-tilt angle was calculated for all 236 HA trimers at each frame of both virion trajectories using in-house tcl scripts within the VMD framework.²⁷ The HA ectodomain-tilt angle is defined by the principal axis of the trimeric TMD (residues 529 to 549) and the vector joining residue 529 at the top of the TMD with the COM of the trimeric ectodomain's long α -helices (LaHs) (residues 419 to 471). The distribution of HA ectodomain-tilt angle values was computed through kernel density estimation using Gaussian kernels with gnuplot plotting program.⁴¹

HA head breathing

The extent of HA head breathing was estimated by calculating the distance between the COM of each monomeric HA head (residues 66 to 286) and the COM of the apical residues of the three LaHs. The apical residue of each LaH corresponds to residue 419 and the COM was calculated based on the three residues 419. This distance was calculated for all 708 HA monomers at each frame of both virion trajectories using in-house tcl scripts within the VMD framework.²⁷ The distribution of HA head–LaHs distance values was computed through kernel density estimation using Gaussian kernels with gnuplot plotting program.⁴¹

NA head breathing

The extent of NA head breathing was estimated by calculating the inter-protomer distances between the four NA head COMs (residues 96 to 461). Considering the four NA heads within each tetramer, a total of six distances have been monitored for each NA tetramer (**Fig. 4.34**), at each frame of both virion trajectories, using in-house

tcl scripts within the VMD framework.²⁷ The distribution of the values of the inter-protomer distances was computed through kernel density estimation using Gaussian kernels with gnuplot plotting program.⁴¹

FluA-20 epitope accessible surface area (ASA)

Accessible surface area (ASA) of the FluA-20⁴² epitope, located on the HA head domain trimer interface, was calculated using the *measure sasa* VMD command,²⁷ an implementation based on the Shrake and Rupley algorithm,⁴³ along with in-house tcl scripts to handle data. The following HA residues were considered for the epitope: residues 109, 230, 232-237, 243. An 18.6 Å probe radius, which approximately encloses the FluA-20 Fab's variable domains (**Fig. 4.5** in the main text), was used. ASA evaluations were conducted on the HA trimer showing the highest degree of head domain breathing, at every frame of the trajectory.

Evaluation of glycoprotein–glycoprotein interactions

For both virion simulations, interactions between glycoproteins were evaluated using the *measure contacts* implemented in VMD²⁷ along with in-house tcl scripts to handle data.

For each glycoprotein, we first pre-computed a list of neighbor glycoproteins using a cutoff of 75 Å. Then, always for each glycoprotein, we assessed the number of connections formed with any of the glycoproteins included in its respective neighbor list every 10th frame of the virion trajectory. At each evaluation, a “connection” was considered formed and counted as one when any protein or glycan heavy atoms of the screened glycoprotein were detected to be within 5 Å of any protein or glycan heavy atoms of the screened neighbor glycoprotein. Only heavy atoms belonging to the ectodomains were considered for the analysis. Importantly, if glycoprotein “A” was, for example, found to form a connection with glycoprotein “B”, that connection was counted one time for glycoprotein “A” and one time also for glycoprotein “B”. Relative fractions of glycoproteins forming respectively 0, 1, 2, 3, 4, or 5 connections at each evaluation were subsequently computed with respect to the total number of glycoproteins (266). Results are shown in **Fig. 4.7** for all the glycoproteins, i.e., HAs + NAs, in the main text and in **Fig. 4.31** for HA and NA.

Within the same workflow, we then evaluated the formation of macro-clusters of glycoproteins. At each assessment, we calculated the number of glycoproteins that were part of a macro-cluster using the same cutoff

of 5 Å between protein/glycan heavy atoms. If glycoprotein “A” was, for example, found to form a connection with glycoprotein “B,” and, glycoprotein “B” was in turn found to establish at the same time a connection with glycoprotein “C,” the three glycoproteins A, B, and C were considered as part of a macro-cluster of 3, regardless of the presence of a direct connection between A and C. Results are shown in **Fig. 4.32**.

Next, we evaluated the nature of the connections formed by each glycoprotein. The connections were broken down into three categories depending on the residues used by each glycoprotein to form such connections: i) “glycan-only” if only glycan residues were used to form the connection; ii) “protein-only” if only protein residues were used to form the connection; “protein and glycan” if both protein and glycan residues were used at the same time to interact with the screened neighbor glycoprotein. Relative frequencies of “glycan-only”, “protein-only”, and “protein and glycan” types of connection were subsequently computed with respect to the total number of counted connections. Results are shown in **Fig. 4.7** in the main text.

Finally, if glycans were involved in a connection between two glycoproteins, for each glycoprotein we counted the number of stalk/head glycans involved in the connection, respectively. Relative frequencies of the interacting stalk/head glycans were subsequently calculated for HA and NA with respect to the respective total number of stalk/head glycans in both strains. Results are shown in **Fig. 4.33**.

Evaluation of conditional probabilities between HA and NA tilting/breathing and inter-protein connectivity

For each motion investigated in this work, namely NA head tilting, HA ectodomain tilting, and HA head breathing, we calculated the conditional probability that a glycoprotein (either NA or HA) exhibits a certain degree of tilting/breathing during the dynamics given that it forms a connection with one or more neighboring glycoproteins. Here, the conditional event, i.e., B, the formation of one or more connections during the dynamics, should account for the inter-protein connectivity. Therefore, with this analysis, we investigate whether the inter-protein connectivity somehow affects the probability of achieving a certain degree of tilting/breathing. In order to calculate such conditional probability, we used the following formula:

$$P(A|B) = P(A \cap B) / P(B)$$

Where:

- $P(A)$ is the probability that a glycoprotein (HA or NA) tilts/breathes by a certain amount during the dynamics with respect to the initial frame.
- $P(B)$ is the probability that the glycoprotein forms one or more connections during the dynamics. $P(B)$ must be > 0 .
- $P(A \cap B)$ is the probability at which the events A and B occur together.
- $P(A|B)$ is the probability of event A occurring if event B has happened.

$P(A)$, $P(B)$ and $P(A \cap B)$ were determined by counting the number of frames where events A, B, and $A \cap B$ occurred, respectively, and dividing it by the total number of screened frames. H1N1-Shan2009 and H1N1-Mich2015 trajectories were analyzed with a stride of 10. To assess whether event A occurred or not for HA ectodomain tilting and NA head tilting, we evaluated, for each HA/NA, the tilting angle at the examined frame with respect to the tilt angle at frame 0 and computed the difference, namely $(\text{tilt-angle_frame}_n - \text{tilt-angle_frame}_0)$. If the obtained difference value was within a certain (positive) range, then event A would be counted as occurred for the evaluated range of values (see **Tables 4.4–4.9**). Instead, to assess whether event A occurred or not for HA head breathing, we evaluated, for each HA trimer, the breathing distance of each monomer at the screened frame with respect to the respective breathing distance at frame 0 and computed the difference, namely $(\text{breathing-distance_frame}_n - \text{breathing-distance_frame}_0)$. If the obtained difference value for at least one monomer within the trimer was within a certain (positive) range, then event A would be counted as occurred for the evaluated range of values (see Tables S4–S9). To assess whether event B occurred or not we evaluated the number of connections made by the screened glycoprotein at each analyzed frame. If the number of connections was equal or larger than one, then event B would be counted as occurred. To assess whether event $A \cap B$ occurred or not, we evaluated whether event A and B occurred at the same time or not. If true, then event $A \cap B$ would be counted as occurred.

The absolute error of $P(A)$ was calculated as $P(A) / \sqrt{(\text{Count of number of times A was observed}) - 1}$. The same was done for the absolute errors of $P(B)$ and $P(A \cap B)$.

The absolute error of $P(A|B)$ was calculated as $E_R(A|B) \times P(A|B)$, where $E_R(A|B)$ is the relative error of $P(A|B)$.

The relative error of $P(A|B)$, namely $E_R(A|B)$, was calculated as $\sqrt{E_R(A)^2 + E_R(B)^2 + E_R(A \cap B)^2}$, where $E_R(A)$ is the relative error of $P(A)$, $E_R(B)$ is the relative error of $P(B)$, and $E_R(A \cap B)$ is the relative error of $P(A \cap B)$.

The relative error of $P(A)$, namely, $E_R(A)$ was calculated as $E(A) / P(A)$. The same was done to calculate $E_R(B)$ and $E_R(A \cap B)$.

Next, we also estimated in a similar way the opposite conditional probability ($P(B|A)$) that a glycoprotein (either NA or HA) forms a connection with one or more glycoproteins, given that it exhibits a certain degree of tilting/breathing during the dynamics. Here, the conditional event is A, i.e., the degree of tilting/breathing achieved during the dynamics. Therefore, with this analysis, we investigate whether tilting/breathing affects the probability of HA/NA forming one or more connections. In order to calculate such conditional probability, we used the following formula:

$$P(B|A) = P(A \cap B) / P(A)$$

Where:

- $P(B)$ is the probability that the glycoprotein forms one or more connections during the dynamics.
- $P(A)$ is the probability that a glycoprotein (HA or NA) tilts/breathes by a certain amount during the dynamics with respect to the initial frame. $P(A)$ must be > 0 .
- $P(A \cap B)$ is the probability at which the events A and B occur together.
- $P(B|A)$ is the probability of event B occurring if event A has happened.

Events A, B, and $A \cap B$ were counted as above, and errors were computed as described above.

Markov State Models (MSMs)

The first step was the preparation of individual NA tetrameric (30) and HA trimeric (236) trajectories, which were generated from the whole-virion MD simulations of H1N1-Shan2009 and H1N1-Mich2015, respectively. Glycans were not included in the generated trajectories because a common topology, i.e., equal for all the NA tetramers and HA trimers, respectively, was necessary to load all the trajectories in the MSM workflow.

However, the impact of glycans on the glycoproteins remains accounted for as the trajectories were generated from the simulations of the glycosylated whole-virions.

NA head tilting. The MD simulation trajectories were imported into the PyEMMA2 software⁴⁴ for MSM analysis. The MSM of the NA head tilting was created as follows. We used two features; the first was defined as the shortest distance between the COM of the head and the COM of the center of the stalk, while the second feature was defined as the shortest distance between the COM of the head and the COM of the bottom of the stalk. The exact residues defining these COM selections are given in the accompanying Jupyter notebook (Data S1). With these two features, we ran a time-lagged independent component analysis (TICA) as is commonly used on MSMs with multiple features^{45–47} using a lag time of 100 steps. Our step size is 0.06, so 100 steps are equivalent to 6 ns. We ran k-means clustering⁴⁸ of the two features, selecting 300 clusters. From our implied timescale plots^{49,50} we selected an MSM lag time of 250 steps (**Fig. 4.19**). We then created Bayesian Markov state models, selecting two macrostates for each MSM. Using these parameters, we created MSMs and validated our two-state systems through Chapman-Kolmogorov (CK) tests^{51–54} at a confidence interval of 95% (**Fig. 4.20**). The estimated and predicted lines overlapped out to 750 steps (3x the MSM lag time number), which meant these MSMs are statistically valid.⁵¹ We then used Robust Perron Cluster Center Analysis (PCCA++) to cluster the first two eigenvectors of the MSM transition matrices to cluster the states.⁵⁵ The first macrostate in the H1N1-Shan2009 strain MSM (untilted) had a stationary distribution of 51.4% of the total population, while the second macrostate (tilted) had a distribution of 48.6%. The first macrostate in the H1N1-Mich2015 strain MSM (untilted) had a stationary distribution of 44.3% of the total population, while the second macrostate (tilted) had a distribution of 55.7%. Finally, we extracted NA structures from the two macrostates per strain by creating a 99.99999% probability cutoff for the H1N1-Shan2009 strain macrostates and a 99.99999% probability cutoff for the H1N1-Mich2015 strain macrostates. This means that the NA structures (microstates) that we selected have a >99.99999% probability of residing in the macrostate state that we think they are in through the PCCA++ fuzzy spectral clustering method. Confirmation of which microstates corresponded to which conformations was performed through the MDTraj software program.⁵⁶

HA ectodomain tilting. The MD simulation trajectories were imported into the PyEMMA2 software⁴⁴ for MSM analysis. The MSM of the HA ectodomain tilting was created as follows. We used two features, defined as the

distance between the COM of the head and the center of the HA stalk, and the distance between the HA monomer head tips and the bottom of the stalk. The exact residues defining these COM selections are given in the accompanying Jupyter notebook. With these two features, we ran TICA⁴⁵⁻⁴⁷ using a lag time of 25 steps. Our step size is 0.06, so 25 steps are equivalent to 1.5 ns. We ran k-means clustering⁴⁸ of the three features, selecting 300 clusters. From our implied timescale plots^{49,50} we selected an MSM lag time of 50 steps (**Fig. 4.24**). We then created Bayesian Markov state models, selecting two macrostates for each MSM. Using these parameters, we created MSMs and validated our two-state systems through CK tests⁵¹⁻⁵⁴ at a confidence interval of 95% (Fig. S18). The estimated and predicted lines overlapped out to 150 steps (>3x the MSM lag time number), which meant these MSMs are statistically valid.⁵¹ We then used PCCA++ to cluster the first two eigenvectors of the MSM transition matrices to cluster the states.⁵⁵ The first macrostate in the H1N1-Shan2009 strain MSM (tilted) had a stationary distribution of 6.3% of the total population, while the second macrostate (tilted) had a distribution of 93.7%. The first macrostate in the H1N1-Mich2015 strain MSM (tilted) had a stationary distribution of 6.2% of the total population, while the second macrostate (untilted) had a distribution of 93.8%. Finally, we extracted HA structures from the two macrostates per strain by creating a 99.999999% probability cutoff for all macrostates. This means that the HA structures (microstates) that we selected have a 99.999999% probability of residing in the macrostate state that we think they are in through the PCCA++ fuzzy spectral clustering method. Confirmation of which microstates corresponded to which conformations was performed through the MDTraj software program.⁵⁶

HA head breathing. The MD simulation trajectories were imported into the PyEMMA2 software⁴⁴ for MSM analysis. The HA head breathing MSM was created as follows. We used one feature, defined as the smallest mean distance of the monomers to each other. The exact residues defining this feature are given in the accompanying Jupyter notebook. With this one feature, we ran TICA⁴⁵⁻⁴⁷ using a lag time of 100 steps to properly format the data, even though the one dimension cannot be reduced, obviously. Our step size is 0.06, so 100 steps are equivalent to 6 ns. We ran k-means clustering⁴⁸ of the features, selecting 300 clusters. Running TICA and k-means clustering should not be necessary for an MSM with one feature, but we ran these anyways for continuity. From our implied timescale plots^{49,50} we selected an MSM lag time of 500 steps (**Fig. 4.27**). We then created Bayesian Markov state models, selecting two macrostates for each MSM. Using these parameters,

we created MSMs and validated our two-state systems through CK tests^{51–54} at a confidence interval of 95% (**Fig. 4.28**). The estimated and predicted lines overlapped out to 1250 steps ($>2.5\times$ the MSM lag time number), which meant these MSMs are statistically valid.⁵¹ We then used PCCA++ to cluster the first two eigenvectors of the MSM transition matrices to cluster the states.⁵⁵ The first macrostate in the H1N1-Shan2009 strain MSM (open) had a stationary distribution of 24.6% of the total population, while the second macrostate (closed) had a distribution of 75.4%. The first macrostate in the H1N1-Mich2015 strain MSM (open) had a stationary distribution of 26.7% of the total population, while the second macrostate (closed) had a distribution of 73.3%. Finally, we extracted HA structures from the two macrostates per strain by creating a 99.99999% probability cutoff for the H1N1-Shan2009 strain macrostates and a 99.99999% probability cutoff for the H1N1-Mich2015 strain macrostates. This means that the HA structures (microstates) that we selected have a $>99.99999\%$ probability of residing in the macrostate state that we think they are in through the PCCA++ fuzzy spectral clustering method. Confirmation of which microstates corresponded to which conformations was performed through the MDTraj software program.⁵⁶

Figures and movies of the whole-virion simulations

Figure panels and movies of the whole-virion simulations were rendered from the trajectories using VMD GPU-accelerated ray tracing.^{27,57}

Single-cell sorting, immunoglobulin amplification, sequencing and mAb production

Cryopreserved PBMC samples were thawed and stained with monoclonal antibodies against human CD3, CD14, CD56, CD19, CD21, CD27, CD38, IgM and IgG (BioLegends, BD BioSciences, Beckam Coulter). The NA probe N2 (A/Wisconsin/67/2005) was conjugated with either AX488 or AX647 fluorochrome according to the manufacturer's protocol (Microscale Protein Labeling Kit, Thermo Fisher Scientific) prior to use in flow cytometry. Non-B cells and dead cells were excluded with CD3, CD14, CD56, and aqua (live-dead) staining in the same fluorescent channel. Memory B cells were further gated on CD38^{low}, CD27^{high}, IgM- and IgG+ and NA-binding memory B cells were single-cell sorted into a 96-well plate (Bio-Rad) using an FACS Aria Instrument (BD Immunocytometry Systems). Plates were subjected to reverse transcription PCR and cDNA

were used for PCR to amplify immunoglobulin (Ig) heavy and light chain genes as described previously.^{58,59} PCR products were sequenced by Sangar sequencing. Ig heavy and light chain sequences were synthesized and cloned into IgG1 heavy and kappa backbone expression vector (GenScript). Antibody NDS.1 was expressed recombinantly by transient transfection of Ig expression plasmids in Expi293 cells with ExpiFectamine (Thermo Fisher Scientific). Supernatant was harvested 5 days post transfection and antibody was purified by using Protein A Sepharose (Thermo Fisher Scientific).

Biolayer interferometry (BLI)

BLI experiments were performed by using the Octet HTX instrument (Sartorius). HIS1K biosensors (Sartorius) were hydrated in PBS prior to use. Recombinant N2 NA tetramers derived from A/Wisconsin/67/2005 and A/Darwin/9/2021 were immobilized on HIS1K biosensors through their hexahistidine tags. After brief equilibration in assay buffer (25 mM Tris pH 8.0, 150 mM NaCl, 1% BSA), the biosensors were dipped into a two-fold dilution series of NDS.1 Fab for 5 min. Starting concentration of NDS.1 was 800 nM. Biosensors were then dipped in the assay buffer to allow Fab to dissociate from NA for 10 min. All assay steps were performed at 30°C with agitation set at 1,000 rpm. Baseline correction was carried out by subtracting the measurements recorded for a sensor loaded with the NA in the same buffer with no Fab. Data analysis and curve fitting were done with the Octet analysis software (version 11). Experimental data were fitted with the binding equations describing a 1:1 (Langmuir model) interaction.

Negative-stain electron microscopy

Samples were diluted to 0.02 mg/ml with buffer containing 10 mM HEPES, pH 7.0, and 150 mM NaCl, adsorbed to glow-discharged carbon-coated copper grids for 10 s, washed with three drops of the same buffer and negatively stained with 0.75% uranyl formate. Micrographs were recorded at a nominal magnification of 100,000 (corresponding to a pixel size of 2.2 Å) on an FEI Tecnai T20 electron microscope equipped with a 2k x 2k Eagle CCD camera and operated at 200 kV. Data collection was performed with SerialEM.⁶⁰ Particles were selected from micrographs using in-house written software (YT, unpublished) and extracted into 144 x 144-pixel boxes. After 2D classification, initial models were obtained and refined using Relion 3.1⁶¹ with C4 symmetry

enforced. The final datasets contained 2,304 and 4,368 particles for NDS.1 Fab and NDS.1/1G01 Fabs, respectively. The corresponding resolutions determined using the Fourier shell correlation threshold of 0.5 was 24.1 and 24 Å. UCSF Chimera⁶² was used for docking and map visualization.

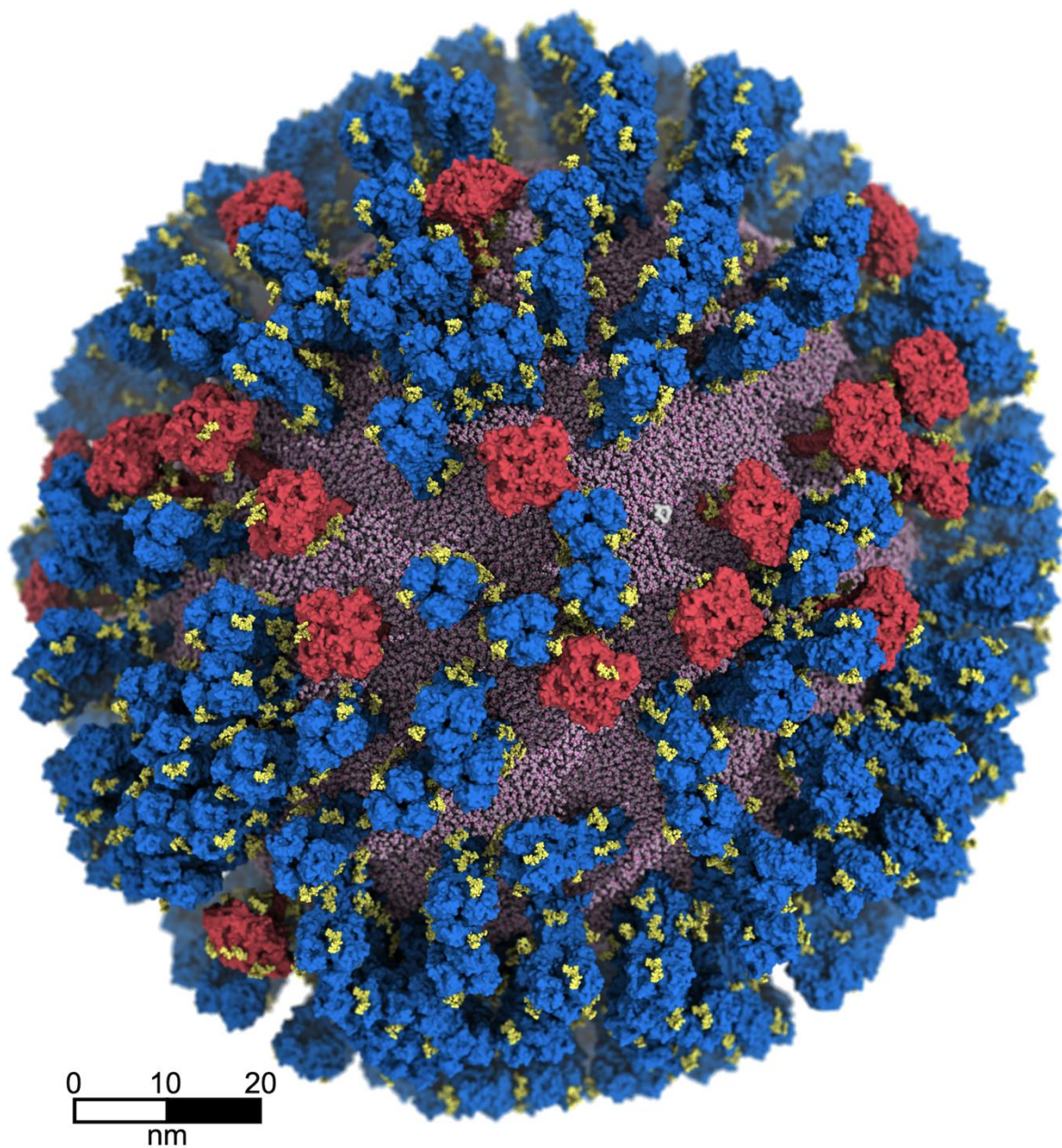


Figure 4.8 | All-atom glycosylated model of H1N1-Mich2015 influenza A virion. Neuraminidase (NA) and hemagglutinin (HA) glycoproteins, and the M2 ion channels, are depicted with red, blue, and white surfaces, respectively. N-linked glycans are shown in yellow. The spherical lipid envelope is represented with pink van der Waals (vdW) spheres.

Score	Expect	Method	Identities	Positives	Gaps
941 bits(2431)	0.0	Compositional matrix adjust.	455/469(97%)	466/469(99%)	0/469(0%)
<i>Mich-2015</i>	N-t	TMD			
	1	13	34	39	43
<i>Mich-2015</i>	MNPNQKIITIGSICMTIGMANLILQIGNIISIWVSHSIQIGNQSIETCNQSVITYENNT				60
<i>Shan-2009</i>	MNPNQKIITIGSVCMTIGMANLILQIGNIISIWVSHSIQIGNQSIETCNQSVITYENNT				60
<i>Mich-2015</i>	61 WVNQTYVNIISNTNFAAGQSVVSVKLAGNSSLCPVSGWAIYSKDNSVRIGSKGDVFFVIREP				120
<i>Shan-2009</i>	WVNQTYVNIISNTNFAAGQSVVSVKLAGNSSLCPVSGWAIYSKDNSVRIGSKGDVFFVIREP				120
<i>Mich-2015</i>	121 FISCSPLECRTFFLTQGALLNDKHSNGTIKDRSPYRTLMSCPIGEVPSPYNSRFESVAWS				180
<i>Shan-2009</i>	FISCSPLECRTFFLTQGALLNDKHSNGTIKDRSPYRTLMSCPIGEVPSPYNSRFESVAWS				180
<i>Mich-2015</i>	181 ASACHDGINWLTIGISGPDNGAVAVLKYNIGIITDTIKSWRNNILRTQESECACVNGSCFT	200			240
<i>Shan-2009</i>	ASACHDGINWLTIGISGPDNGAVAVLKYNIGIITDTIKSWRNNILRTQESECACVNGSCFT				240
<i>Mich-2015</i>	241 IMTDGPSDGGQASYKIFRIEKGKIIKSVEMKAPNYHYEECSYCPDSSEITCVCRDNWHGSN	248	264	269	300
<i>Shan-2009</i>	VMTDGPSNGQASYKIFRIEKGKIVKSVEMNAPNYHYEECSYCPDSSEITCVCRDNWHGSN				300
<i>Mich-2015</i>	301 RPWVSFNQNLLEYQMGYICSGVFGDNPRPNDKTGSCGPVSSNGANGVKGFSFKYGNVWIG	314	321		360
<i>Shan-2009</i>	RPWVSFNQNLLEYQIGYICSGIFGDNPRPNDKTGSCGPVSSNGANGVKGFSFKYGNVWIG				360
<i>Mich-2015</i>	361 RTKSISSRKGFEMIWDPNGWTGTDNKF SIKQDIVGINEWSGYSGSFVQHPELTGLDCIRP	368	386		420
<i>Shan-2009</i>	RTKSISSRNGFEMIWDPNGWTGTDNKF SIKQDIVGINEWSGYSGSFVQHPELTGLDCIRP				420
<i>Mich-2015</i>	421 CFWVELIRGRPEENTIWTSGSSISFCGVNSDTVGWSWPDGAELPFTIDK	432			469
<i>Shan-2009</i>	CFWVELIRGRPKENTIWTSGSSISFCGVNSDTVGWSWPDGAELPFTIDK				469
					C-t

Figure 4.10 | Aligned sequences of NA in H1N1-Shan2009 (UniProtKB accession code: F2YI86) and H1N1-Mich2015 (UniProtKB accession code: A0A0X9QTS2) as modeled in this work. N-linked glycosylation sites are highlighted with blue squares, whereas mutation sites with yellow squares. Alignment was performed with BLAST.⁶³

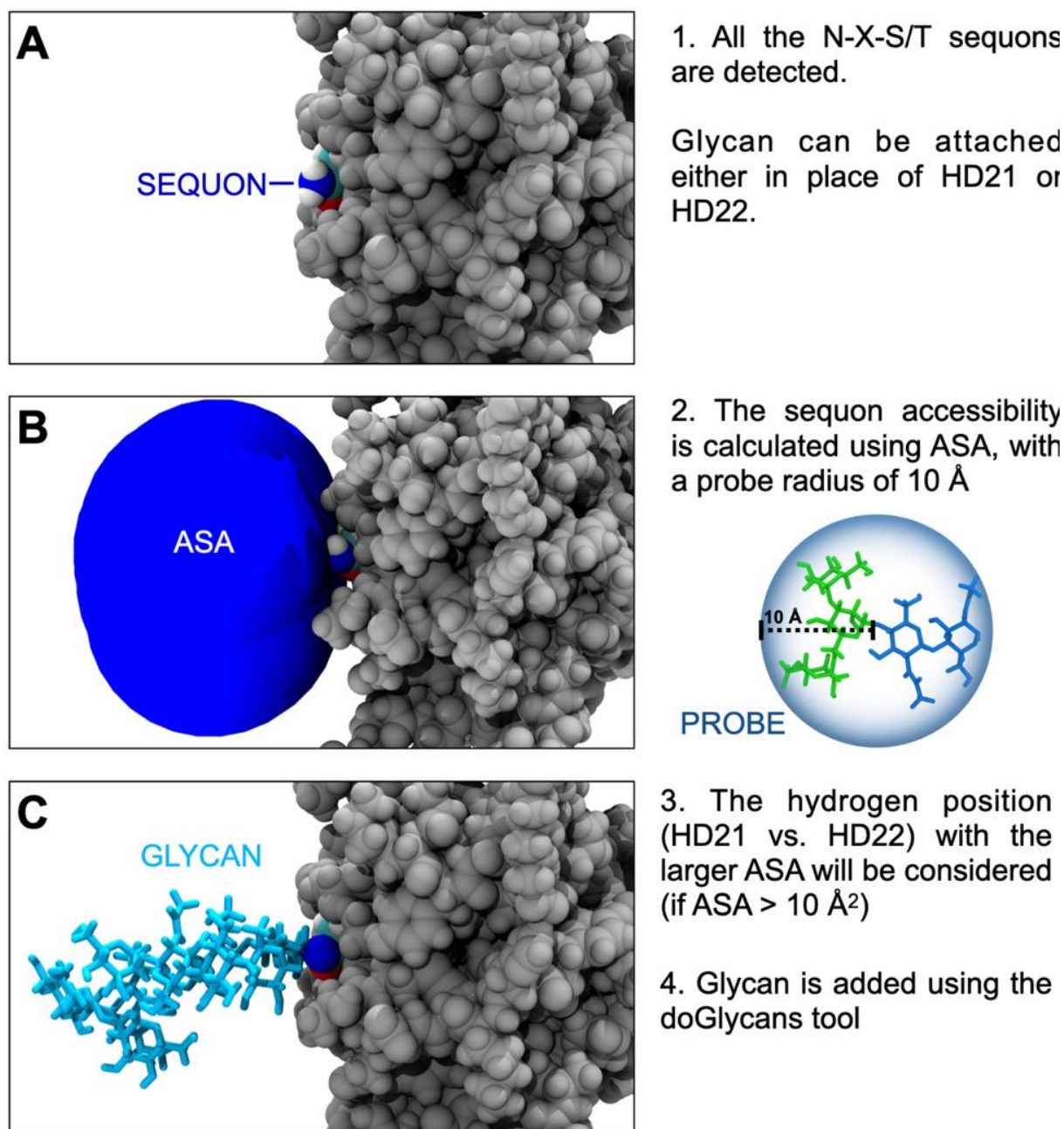
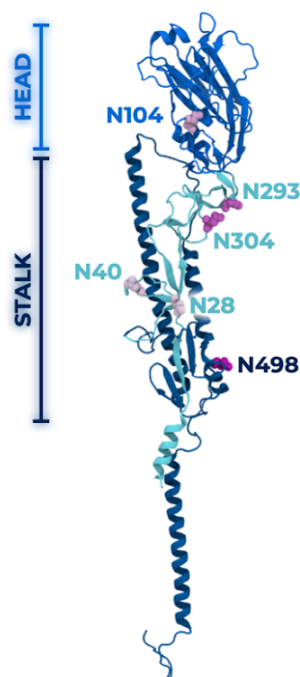
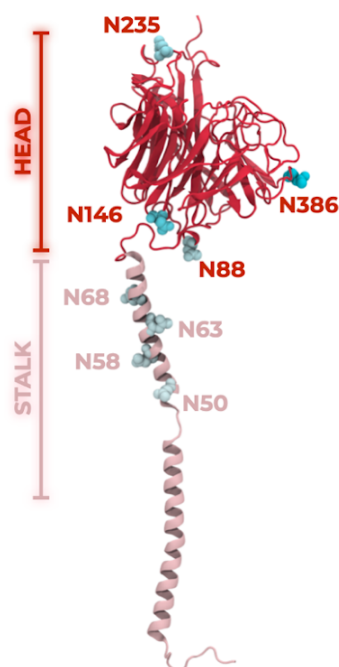


Figure 4.11 | Schematic representation of the glycosylation procedure. **A)** N-X-S/T N-linked sequons are selected. **B)** Although all sequons are overall highly occupied as reported by previous studies,^{22,30,31} in this work we evaluated the accessibility of each selected sequon by calculating the accessible surface area (ASA) of the asparagine side chain's ND2 atom using a probe radius of 10 Å. This was done since the initial coordinates of the whole-virion construct are derived from previous MD simulation of the unglycosylated whole-virion construct.¹ **C)** Addition of glycan.



HA

Seuon	Region	Structure	#	%
N28	Stalk		428/708	60.5%
N40	Stalk		299/708	42.2%
N104	Head		29/708	4.1%
N293	Stalk		36/708	5.1%
N304	Stalk		95/708	13.4%
N498	Stalk		390/708	55.1%
TOTAL			1277/4248	30%



NA

Seuon	Region	Structure	#	%
N50	Stalk		30/120	25.0%
N58	Stalk		0/120	0%
N63	Stalk		38/120	31.7%
N68	Stalk		10/120	8.3%
N88	Head		37/120	30.8%
N146	Head		16/120	13.3%
N235	Head		32/120	26.7%
N386	Head		81/120	67.5%
TOTAL			244/960	25.4%

Figure 4.12 | Glycosylation of the 708 monomeric HAs (top) and 120 monomeric NAs (bottom) included in the H1N1-Shan2009 whole-virion model. The position of N-linked glycosylation sites on HA (top) and NA (bottom) monomeric structure is shown on the left panels. The structure, total number, and relative frequency (%) of glycans added at each N-linked site are summarized on the right panels.

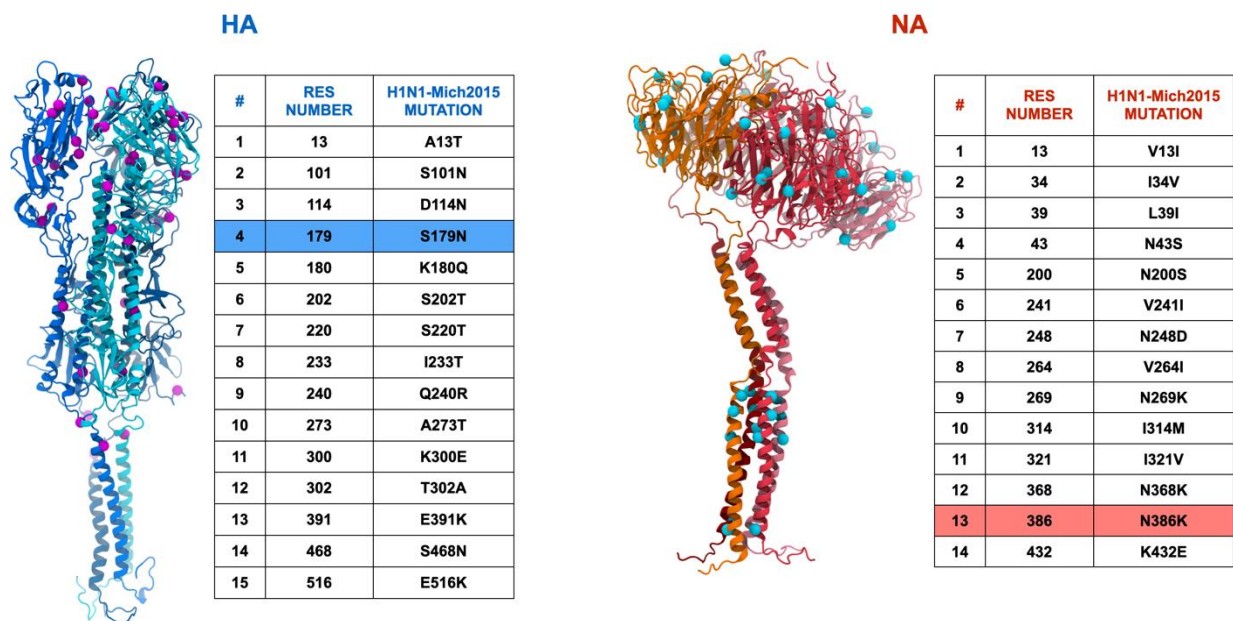


Figure 4.13 | List of amino acid point mutations in H1N1-Mich2015 with respect to H1N-Shan2009 for HA (left panel) and NA (right panel). The addition of N-linked sequon within HA head at position N179 is highlighted in blue. Deletion of N-linked sequon within NA head at position N386 highlighted in red. The location of the point mutations within HA (left) and NA (right) structures is highlighted with magenta and cyan beads, respectively.

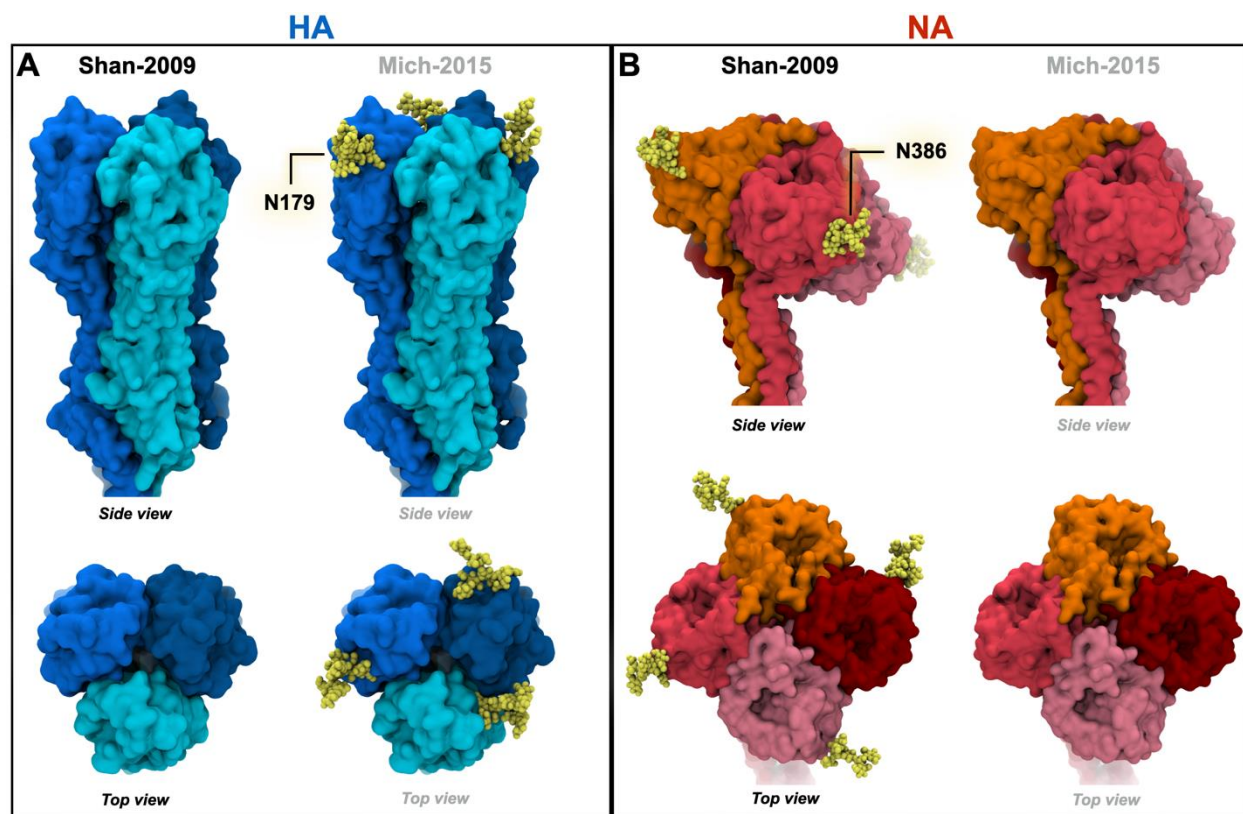
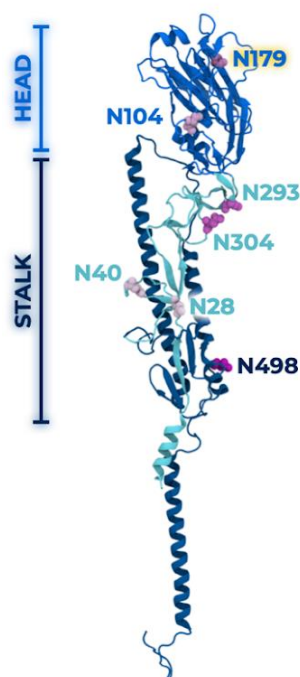
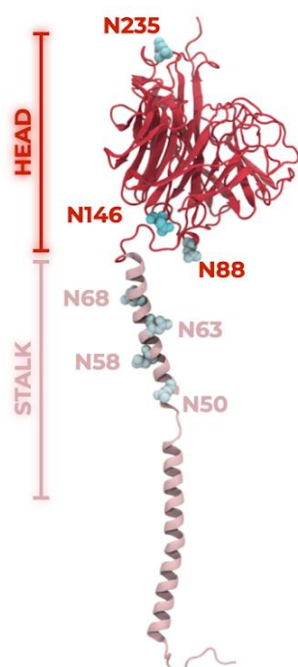


Figure 4.14 | Glycan variation in H1N1-Mich2015 with respect to H1N1-Shan2009 for HA (**A**) and NA (**B**). A new N-linked glycan is introduced at position N179 within the HA head (**A**). N-linked glycan at position N386 within the NA head is instead deleted (**B**). The glycan is illustrated with yellow van der Waals spheres, whereas HA and NA are represented with cyan and red surfaces, respectively.



HA

Sequon	Region	Structure	#	%
N28	Stalk		420/708	59.3
N40	Stalk		288/708	40.7%
N104	Head		33/708	4.7%
N179	Head		351/708	49.6%
N293	Stalk		39/708	5.5%
N304	Stalk		102/708	14.4%
N498	Stalk		393/708	55.5%
TOTAL			1626/4956	32.8%



NA

Sequon	Region	Structure	#	%
N50	Stalk		31/120	25.8%
N58	Stalk		0/120	0%
N63	Stalk		40/120	33.3%
N68	Stalk		11/120	9.2%
N88	Head		36/120	30.0%
N146	Head		18/120	15.0%
N235	Head		29/120	24.2%
N386	—	—	—	—
TOTAL			165/840	19.6%

Figure 4.15 | Glycosylation of the 708 monomeric HAs (top) and 120 monomeric NAs (bottom) included in the H1N1-Mich2015 whole-virion model. The position of N-linked glycosylation sites on the HA (top) and NA (bottom) monomeric structure is shown on the left panels. The structure, total number, and relative frequency (%) of glycans added at each N-linked site are summarized on the right panels.

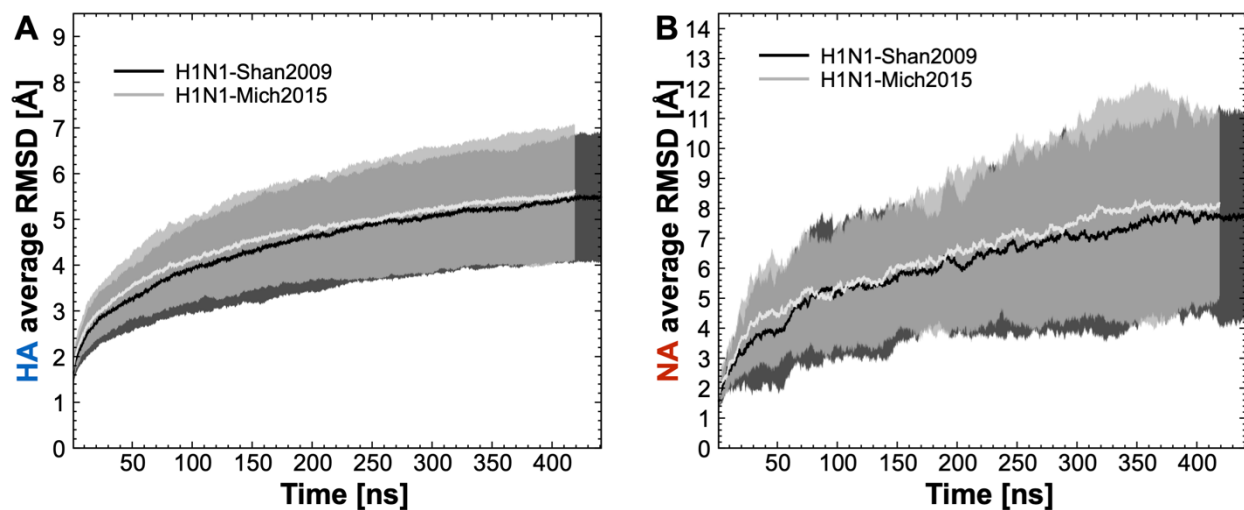


Figure 4.16 | Average RMSD profiles for HA (A) and NA (B) as calculated over the course of H1N1-Shan2009 (black line) and H1N1-Mich2015 (gray line) whole-virion MD simulations. RMSD values have been averaged across the 236 HA trimers and 30 NA tetramers, respectively. Standard deviation values are plotted with a semi-transparent filled area.

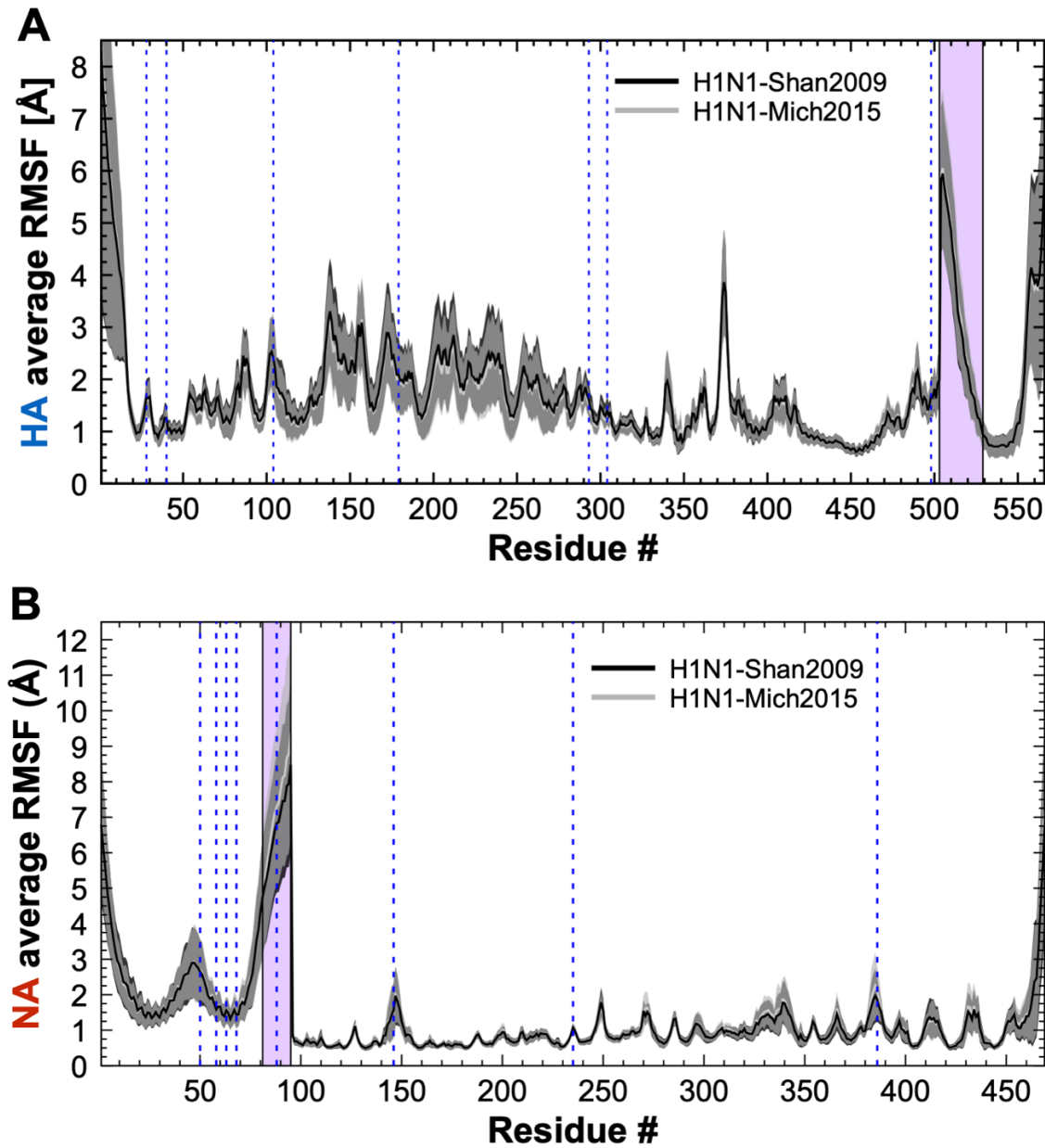


Figure 4.17 | Average RMSF profiles for HA (A) and NA (B) as calculated over the course of H1N1-Shan2009 (black line) and H1N1-Mich2015 (gray line) whole-virion MD simulations. RMSF values have been averaged across the 708 HA and 120 NA monomers, respectively. Standard deviation values are plotted with a semi-transparent filled area. Highlighted with a transparent purple rectangle is the flexible hinge region for HA and NA facilitating HA ectodomain tilting and NA head tilting, respectively. The left side of the purple rectangle in the HA RMSF profile demarcates the two alignment HA regions used for RMSF calculation. The left side of the purple rectangle in the HA RMSF profile demarcates the two alignment regions used for RMSF calculation.

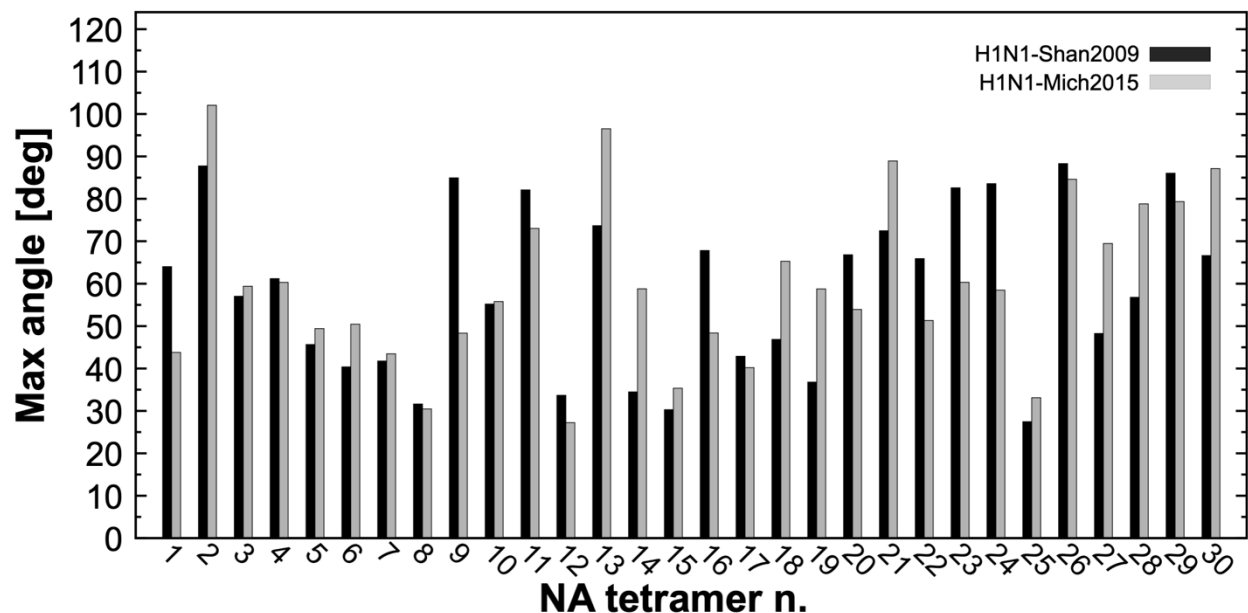


Figure 4.18 | Maximum NA head-tilt angle observed in H1N1-Shan2009 and H1N1-Mich2015 NA tetramers. For each NA tetramer (numbered from #1 to #30), the maximum NA head-tilt angle values (degree) calculated during the simulations of H1N1-Shan2009 and H1N1-Mich2015 are shown with black and gray bars, respectively.

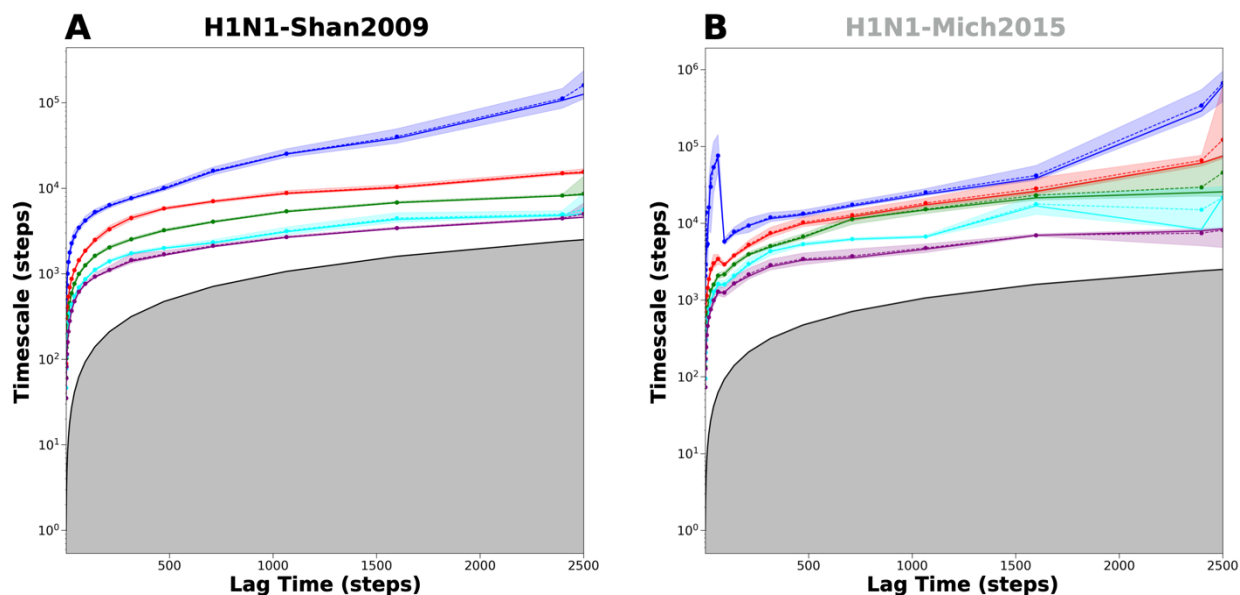


Figure 4.19 | The implied timescale plots for the MSM of the NA head tilting in (A) H1N1-Shan2009 and (B) H1N1-Mich2015. The shading on each of the implied timescales represents statistical uncertainties calculated through Bayesian sampling. From these figures, the lag time was selected to be 250 steps and two macrostates were selected. The gray region represents forbidden space, which is faster than the lag time; transitions in this region cannot be resolved.

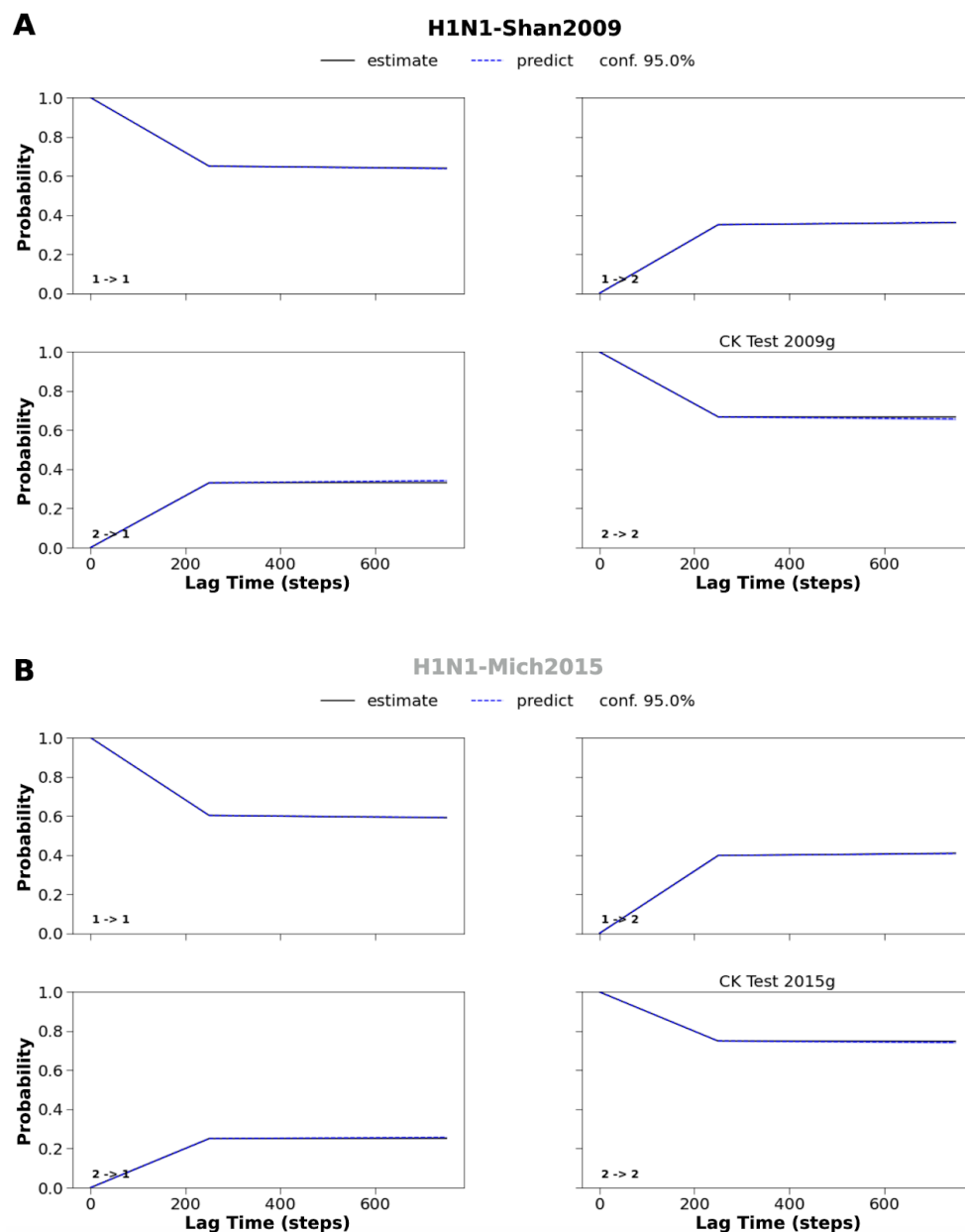


Figure 4.20 | The CK test plots for the MSM of the NA head tilting in **(A)** H1N1-Shan2009 and **(B)** H1N1-Mich2015. The lag time was selected to be 250 steps, with two macrostates selected. The solid line (the “estimate”) represents the transition probabilities of the model created at a lag time of 250 steps, with these probabilities then estimated at multiples of the lag time. The dotted line (the “prediction”) represents predicted probabilities of the same model independently created at each of these multiples of the lag time. Dark blue shading represents a 95% confidence interval for the predicted probabilities. Since the estimates of the MSM model that we used overlap with the predicted probabilities of the model at 3x the lag time, this model satisfies Markovianity, passes the CK test and can be used to extract transition kinetics. The numbers in the lower left corners represent macrostate transitions; thus “1 → 1” represents a transition from macrostate 1 to macrostate 1.

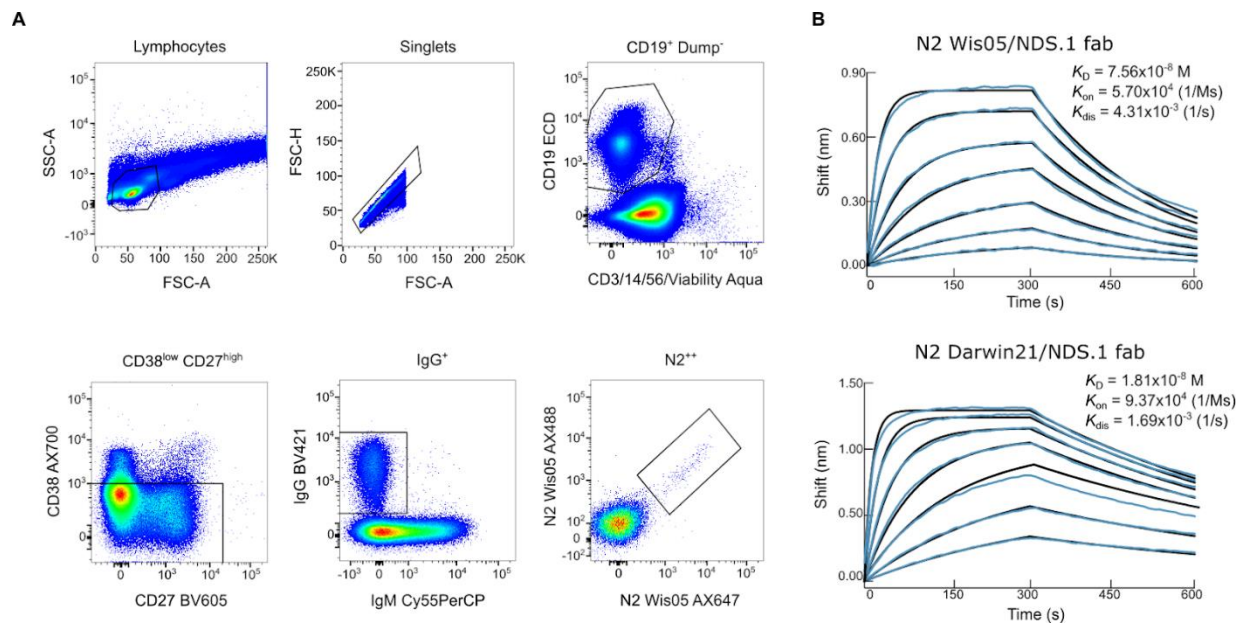


Figure 4.21 | Isolation and characterization of NA-directed mAb NDS.1. **(A)** Gating strategy of NA-specific memory B-cells by flow cytometry. **(B)** Binding kinetics of NDS.1 Fab to N2-Wis05 and N2-Darwin21 by biolayer interferometry. Experimental data (blue traces) were fitted (black lines) with the binding equations describing a 1:1 interaction.

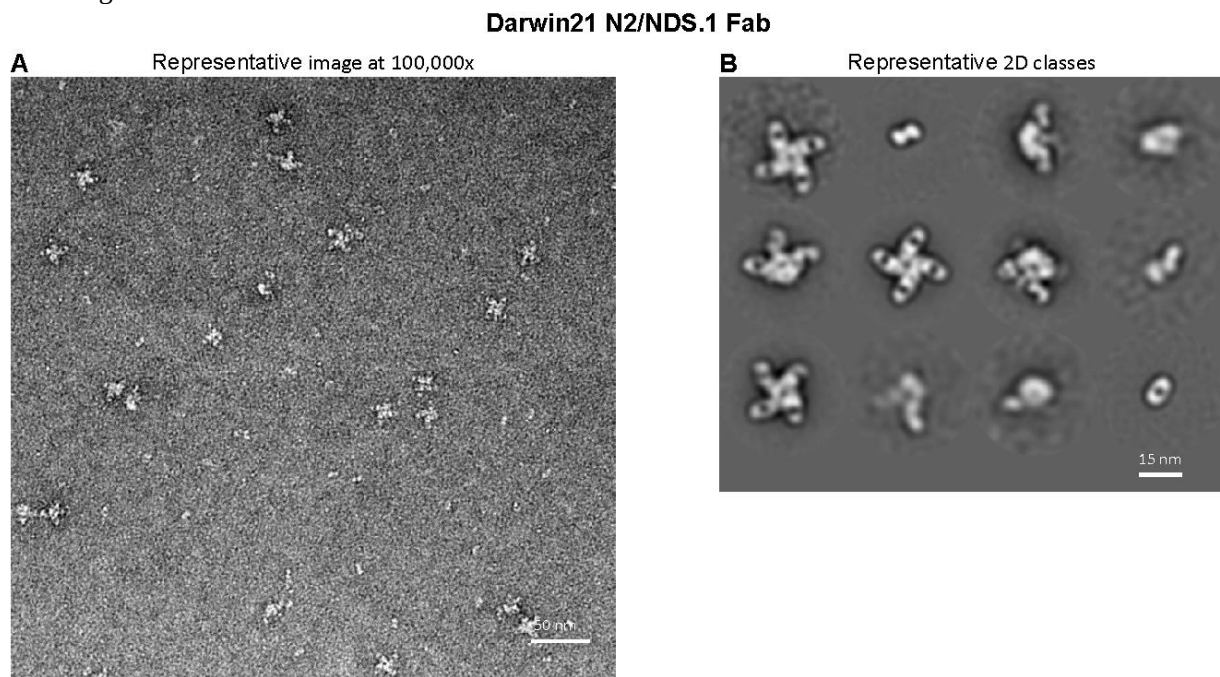


Figure 4.22 | Negative-stain EM analysis of N2 A/Darwin/9/2021 NA tetramer in complex with Fab NDS.1 **(A)** Representative raw NS-EM image and **(B)** 2D class averages of N2 A/Darwin/9/2021 NA in complex with Fab NDS.1 (scale bar: 50 nm and 15 nm).

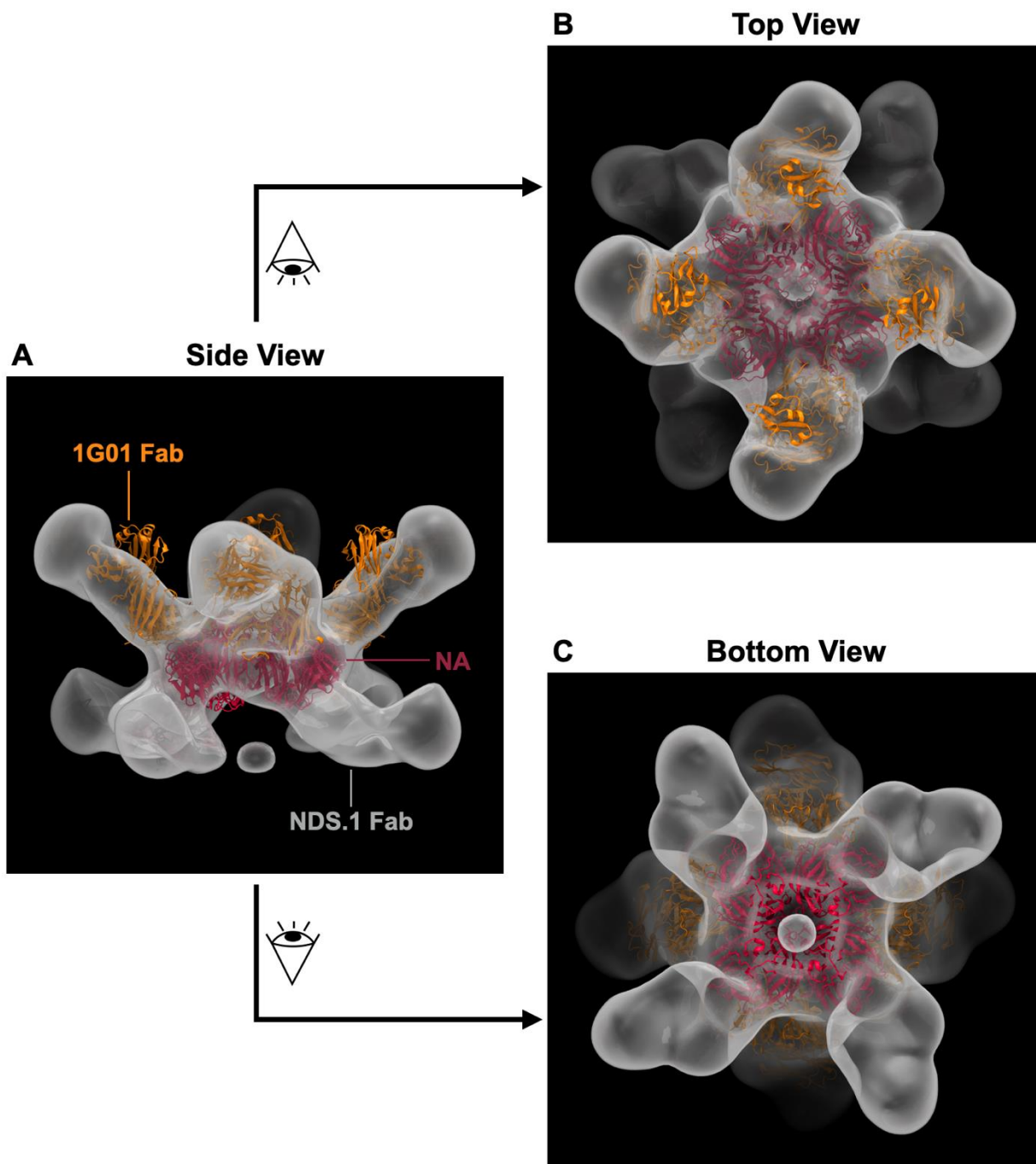


Figure 4.23 | NS-EM reconstruction model of N2 A/Indiana/10/2011 NA tetramer in complex with NDS.1 Fab and 1G01 Fab. Side view (**A**), top view (**B**) and bottom view (**C**) of the ternary complex. Crystal structure of NA-1G01 (PDB ID: 6Q23⁶⁴) is fitted into the NS-EM density map. The final model was generated using a dataset containing 4,368 particles.

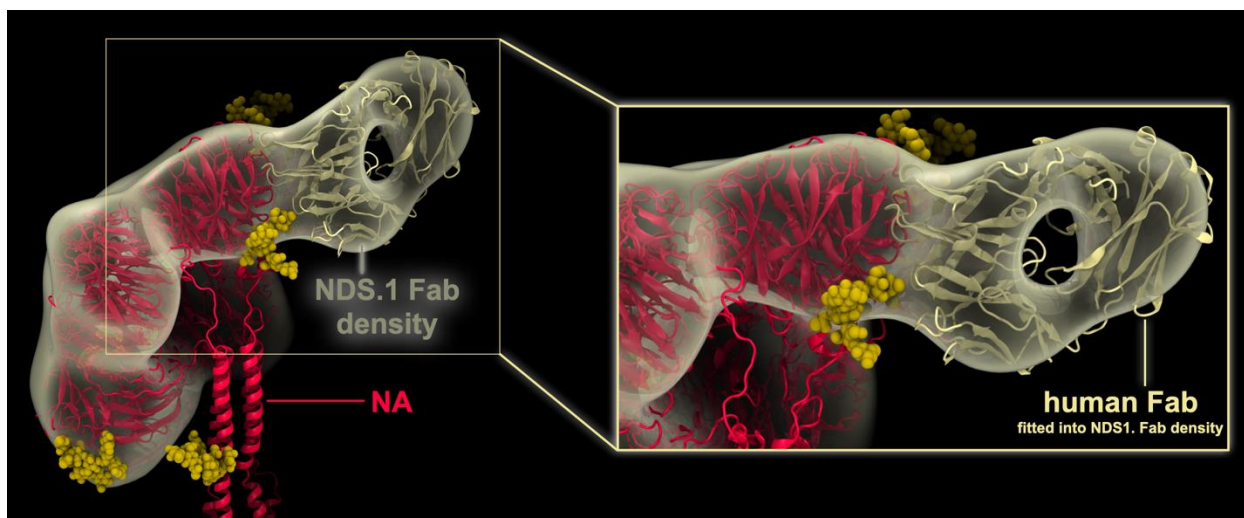


Figure 4.24 | The NA head-underside epitope. A mock Fab, namely the crystal structure of the human immunoglobulin fragment Fab New (IgG1, λ) (PDB ID: 7FAB⁶⁵), together with one conformation of the NA with a tilted head extracted from the simulation of H1N1-Shan2009, was fitted into the NS-EM density map of the N2 A/Indiana/10/2011 NA tetramer in complex with NDS.1 Fab. NA is represented with red cartoons, whereas the human immunoglobulin fragment Fab is shown with yellow cartoons. The NS-EM density is depicted with a transparent yellow surface. Glycans are shown with yellow vdW spheres.

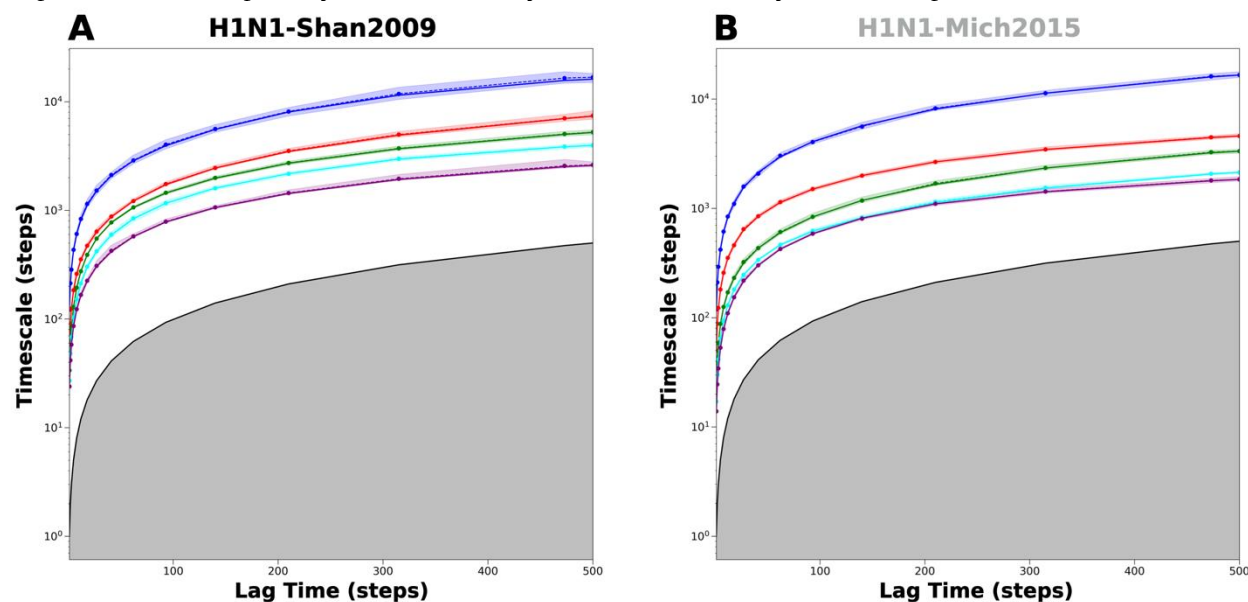


Figure 4.25 | The implied timescale plots for the MSM of the HA ectodomain tilting in (A) H1N1-Shan2009 and (B) H1N1-Mich2015. The shading on each of the implied timescales represents statistical uncertainties calculated through Bayesian sampling. From these figures, the lag time was selected to be 50 steps and two macrostates were selected. The gray region represents forbidden space, which is faster than the lag time; transitions in this region cannot be resolved.

Figure 4.26 | The CK test plots for the MSM of the HA ectodomain tilting in **(A)** H1N1-Shan2009 and **(B)** H1N1-Mich2015. The lag time was selected to be 50 steps, with two macrostates selected. The solid line (the “estimate”) represents the transition probabilities of the model created at a lag time of 50 steps, with these probabilities then estimated at multiples of the lag time. The dotted line (the “prediction”) represents predicted probabilities of the same model independently created at each of these multiples of the lag time. Dark blue shading represents a 95% confidence interval for the predicted probabilities. Since the estimates of the MSM model that we used overlap with the predicted probabilities of the model at 3x the lag time, this model satisfies Markovianity, passes the CK test and can be used to extract transition kinetics. The numbers in the lower left corners represent macrostate transitions; thus “1 -> 1” represents a transition from macrostate 1 to macrostate 1.

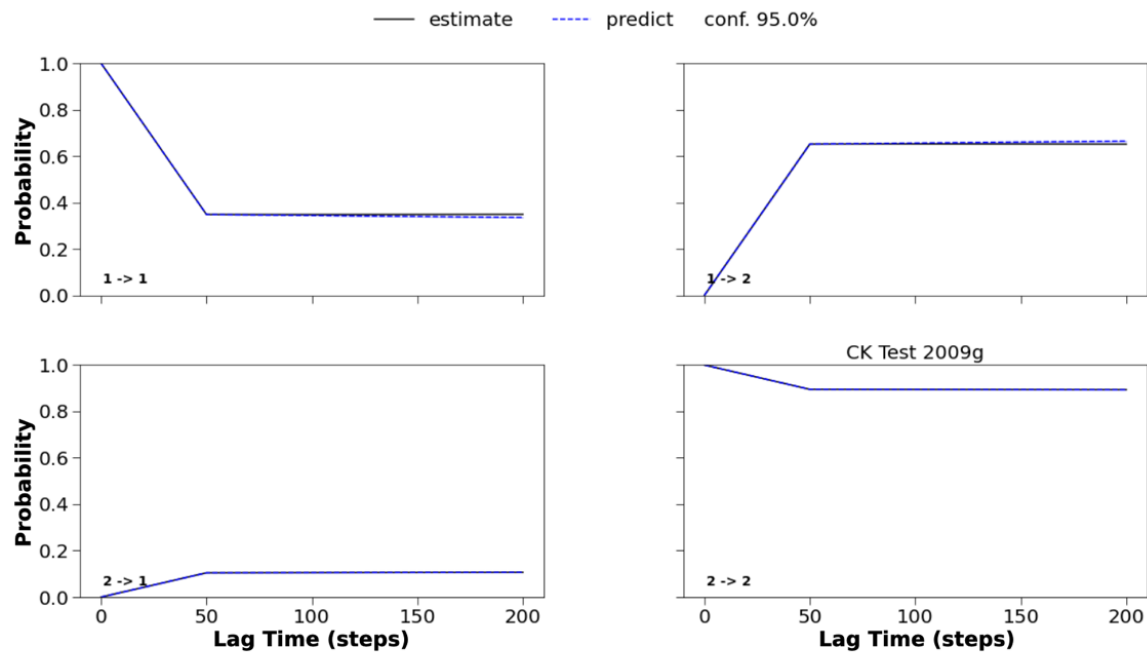
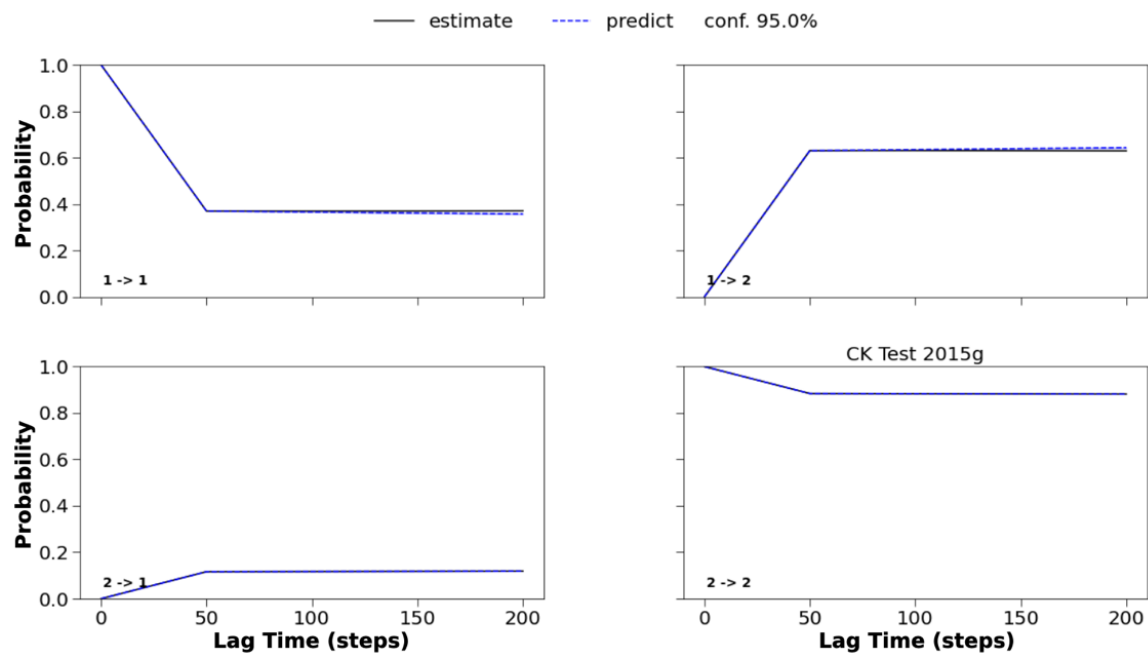
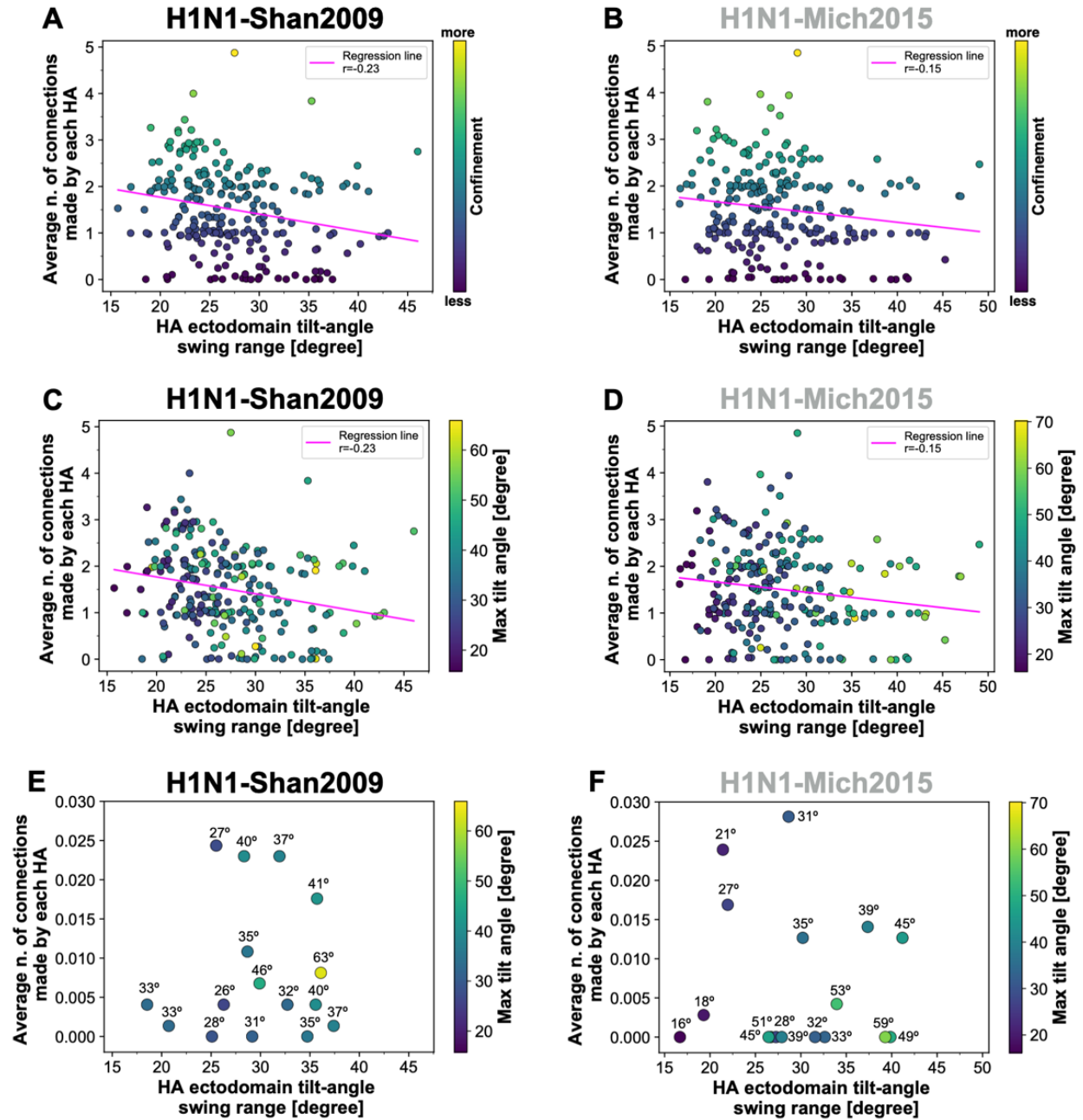
A**H1N1-Shan2009****B****H1N1-Mich2015**

Figure 4.27 | Analysis of the correlation between HA ectodomain tilting and HA protein confinement. The correlation between the HA ectodomain tilt-angle swing range (*x-axis*) and the average number of connections made by each HA (*y-axis*) is illustrated in the scatter plots **(A)** and **(B)** for H1N1-Shan2009 and H1N1-Mich2015, respectively. The tilt-angle swing range is calculated for each HA ectodomain as (*max_tilt-angle_value* – *min_tilt-angle_value*) exhibited during the respective dynamics. The tilt angle and the connections made by each HA with neighboring glycoproteins are evaluated as described in the Material and Methods section. The regression line is shown in magenta. Each point is colored according to the extent of protein confinement using the viridis palette: the higher the average number of connections made by each HA the more the respective confinement. The same plots are replicated in panels **(C)** and **(D)** where instead each point is colored according to the maximum value of tilting angle exhibited during the simulations by the respective HA ectodomain. Plots **(E)** and **(F)** are a subsection of plots **(C)** and **(D)**, respectively, where only the 0.000–0.0030 range is considered on the y-axis. These plots show the extent of tilting for the HA glycoproteins that remain isolated during the simulations, i.e., the ones that do not interact with any neighboring glycoproteins. The value of the maximum tilt angle achieved during the dynamics is reported for each HA.



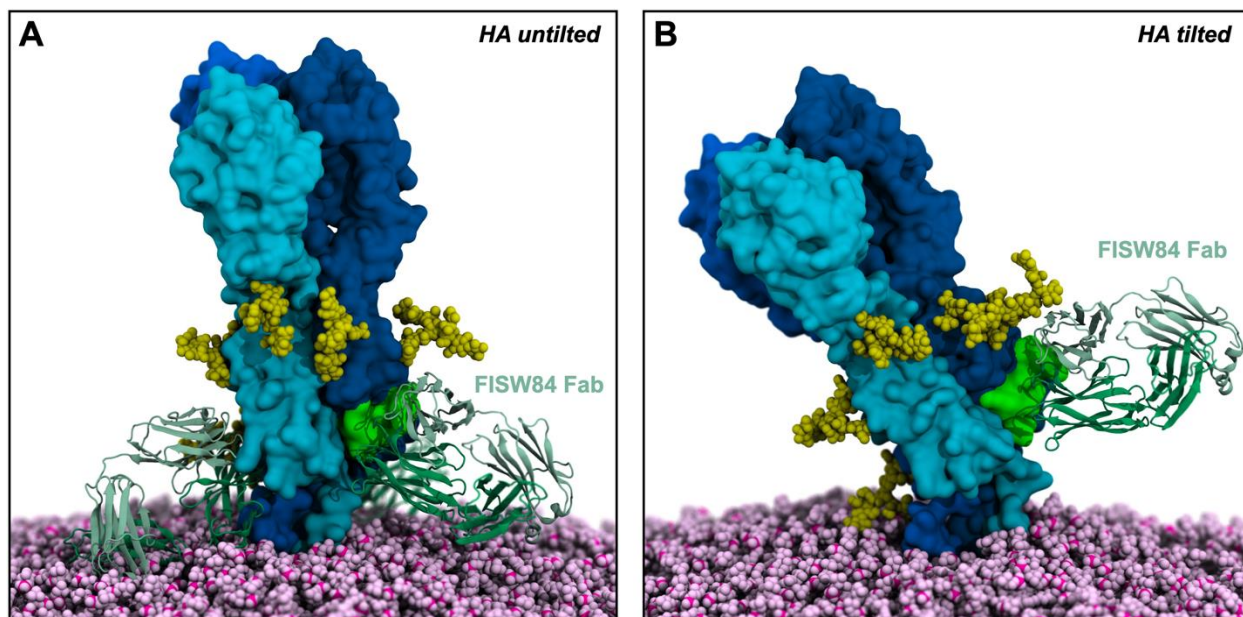


Figure 4.28 | HA anchor epitope accessibility. **A)** When HA is untitled, the FISW84 Fab⁶⁶ must approach the anchor epitope with an upward angle. FISW84 is depicted with green cartoons, whereas the three HA monomers are depicted with surfaces colored with different shades of blue. The anchor epitope is highlighted with a green surface. N-glycans are shown with yellow vdW spheres, whereas the POPC lipids are displayed with pink vdW spheres. **B)** When HA flexes onto the membrane, the angle of approach to at least one epitope gradually shifts from upward to lateral or slightly downward, facilitating FISW84 Fab binding.

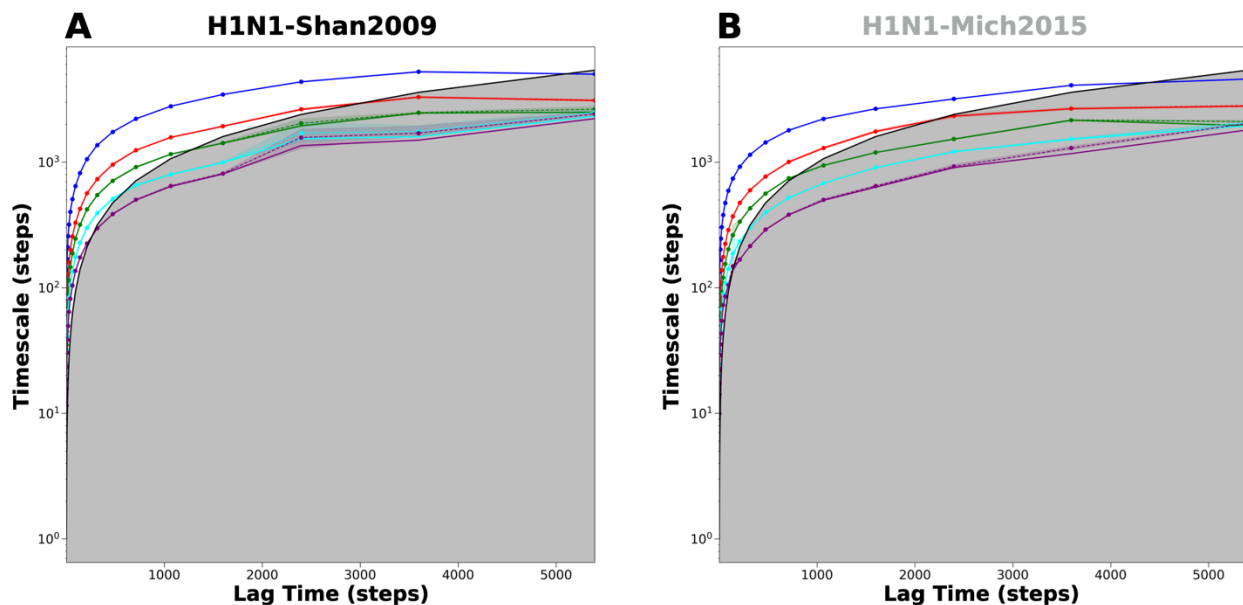
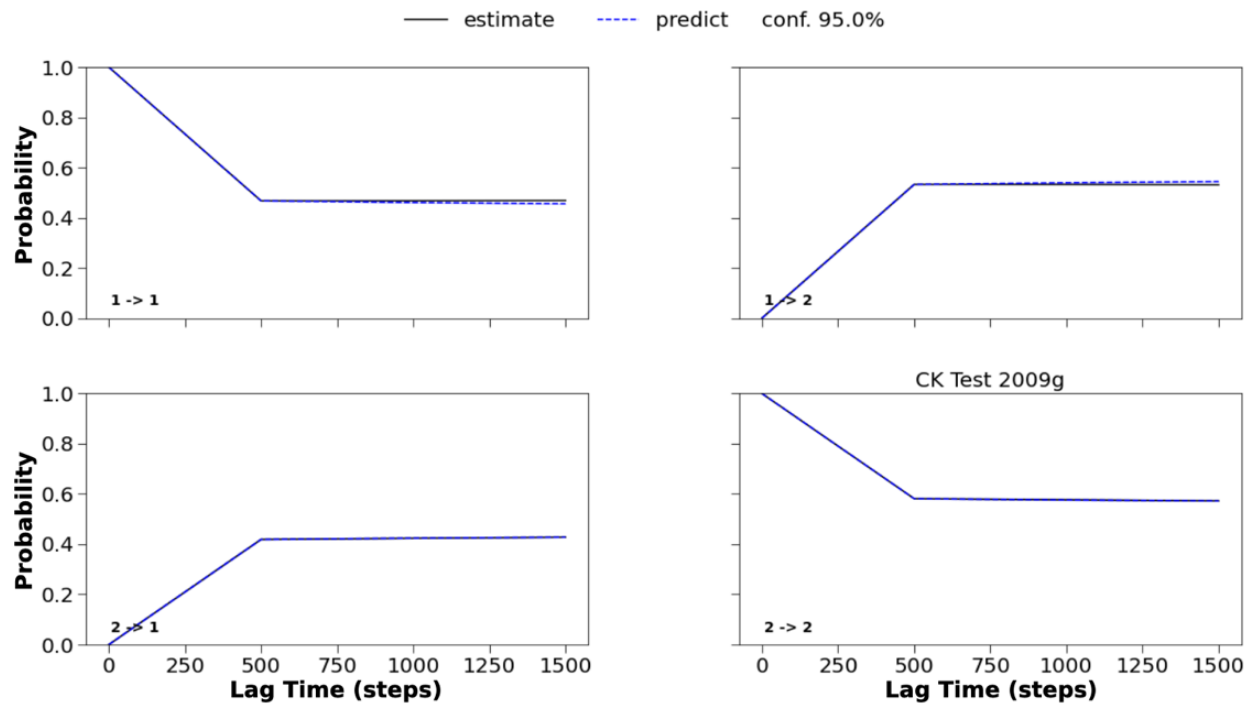
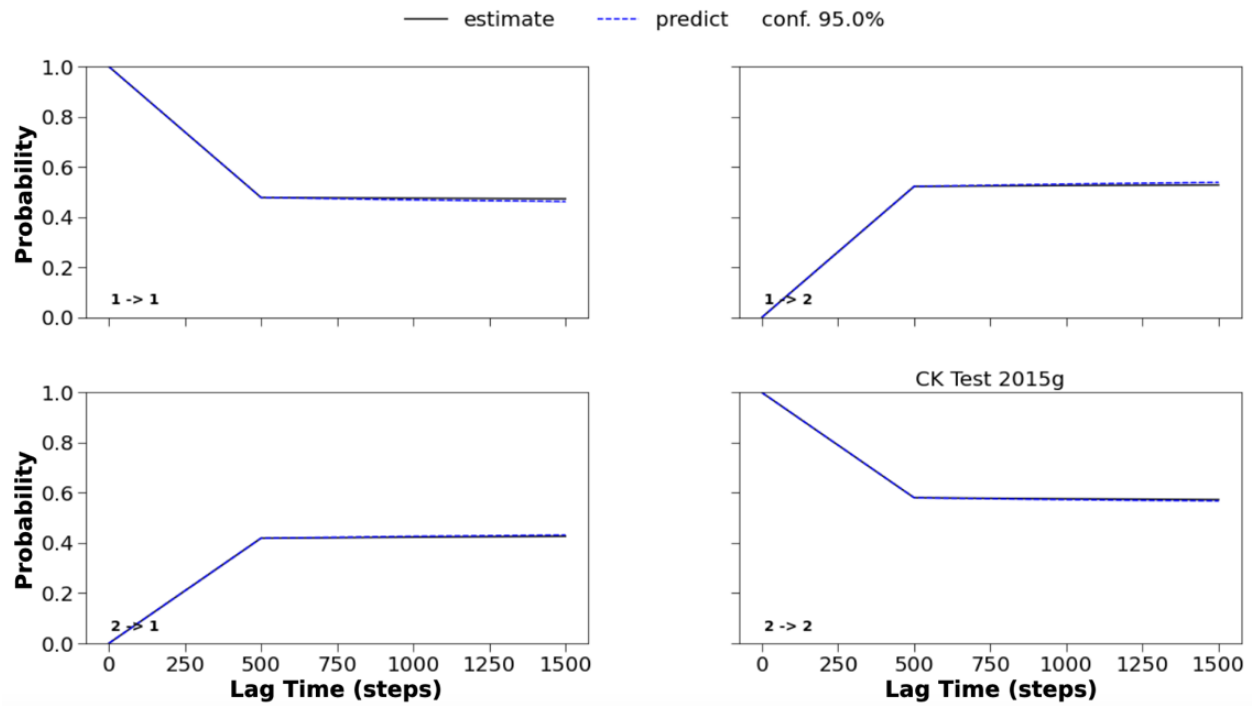


Figure 4.29 | The implied timescale plots for the MSM of the HA head breathing in **(A)** H1N1-Shan2009 and **(B)** H1N1-Mich2015. The shading on each of the implied timescales represents statistical uncertainties calculated through Bayesian sampling. From these figures, the lag time was selected to be 500 steps and two macrostates were selected. The gray region represents forbidden space, which is faster than the lag time; transitions in this region cannot be resolved.

Figure 4.30 | The CK test plots for the MSM of the HA head breathing in **(A)** H1N1-Shan2009 and **(B)** H1N1-Mich2015. The lag time was selected to be 500 steps, with two macrostates selected. The solid line (the “estimate”) represents the transition probabilities of the model created at a lag time of 500 steps, with these probabilities then estimated at multiples of the lag time. The dotted line (the “prediction”) represents predicted probabilities of the same model independently created at each of these multiples of the lag time. Dark blue shading represents a 95% confidence interval for the predicted probabilities. Since the estimates of the MSM model we used overlap with the predicted probabilities of the model at 3x the lag time, this model satisfies Markovianity, passes the CK test and can be used to extract transition kinetics. The numbers in the lower left corners represent macrostate transitions; thus “1 -> 1” represents a transition from macrostate 1 to macrostate 1.

A**H1N1-Shan2009****B****H1N1-Mich2015**

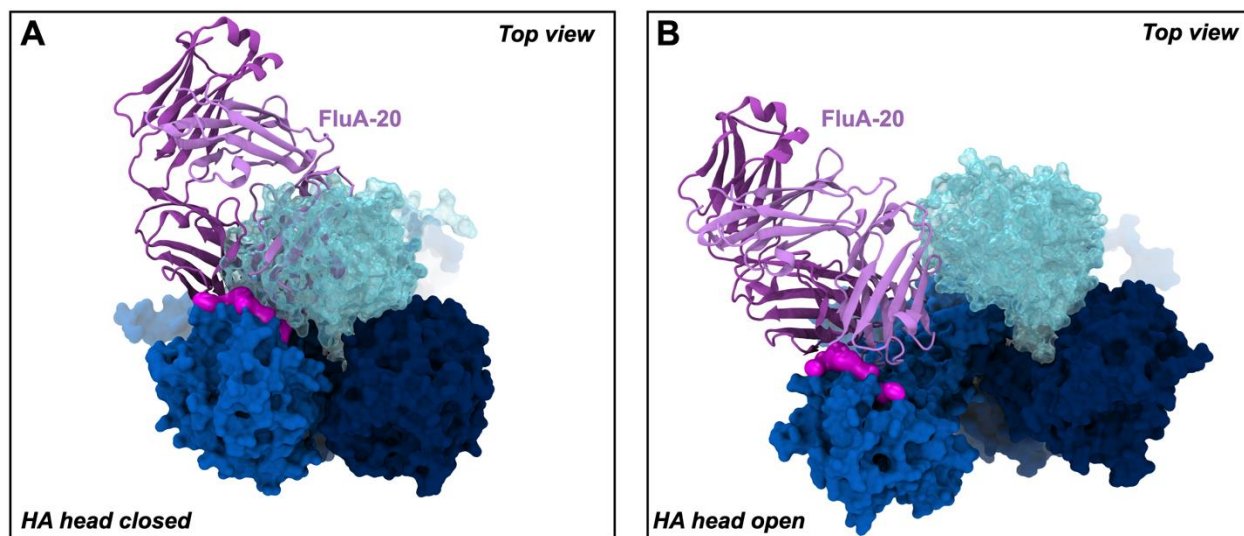


Figure 4.31 | HA head trimer interface epitope accessibility. **A)** When HA heads are closed, the FluA-20 antibody⁴² cannot access the epitope buried in the head trimer interface. FluA-20 is depicted with purple cartoons, whereas the three HA monomers are depicted with surfaces colored with different shades of blue. The FluA-20 epitope is highlighted with a magenta surface on one HA head. **B)** When one HA breathes, the FluA-20 epitope becomes accessible allowing antibody binding.

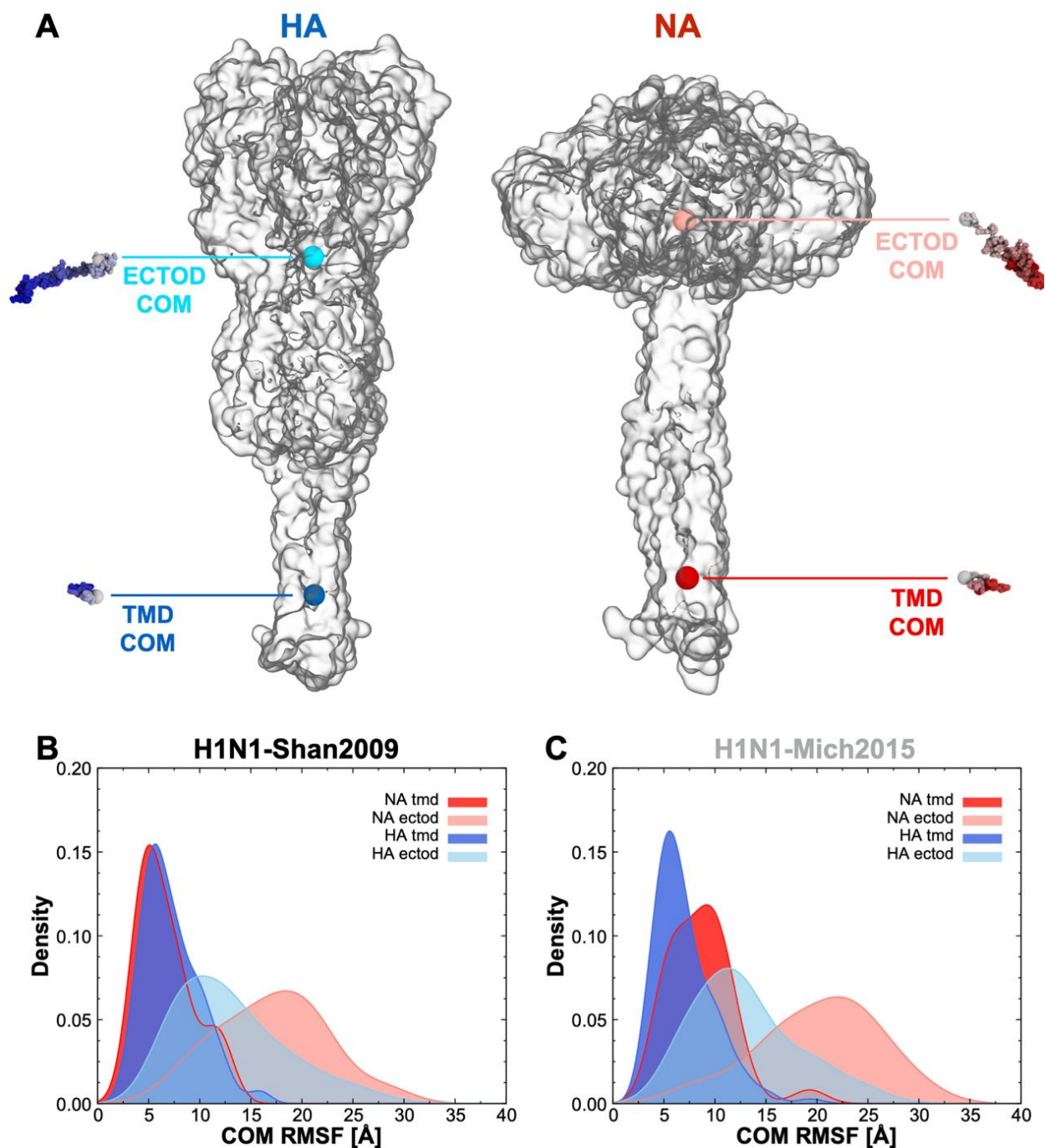


Figure 4.32 | RMSF of HA and NA transmembrane domain (TMD) COM and ectodomain COM. The initial frame was used as a reference for RMSF calculation. **A**) Molecular representation of an HA trimer (left) and a NA tetramer (right) using a transparent surface. TMD COM and ectodomain COM of HA and NA are indicated with cyan and red spheres, respectively. As an example, the position of TMD COM and ectodomain COM is displayed with a sphere every 10th frame of the MD simulation for a representative HA trimer and a representative NA tetramer selected from H1N1-Shan2009. The spheres are colored according to simulation time using a white-to-blue gradient for HA and a white-to-red gradient for NA. **B**) Kernel density distribution profiles of RMSF values calculated for the HA (236 trimers) and NA (30 tetramers) TMD COMs and ectodomain COMs in H1N1-Shan2009. **C**) Kernel density distribution profiles of RMSF values calculated for the HA (236 trimers) and NA (30 tetramers) TMD COMs and ectodomain COMs in H1N1-Mich2015.

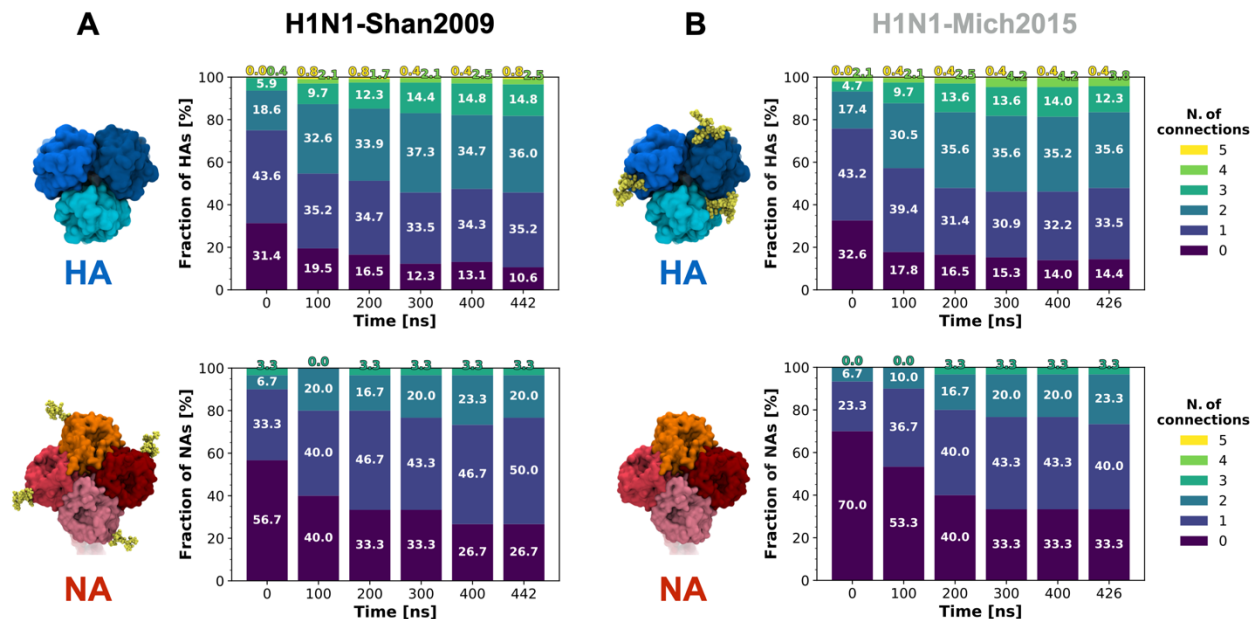
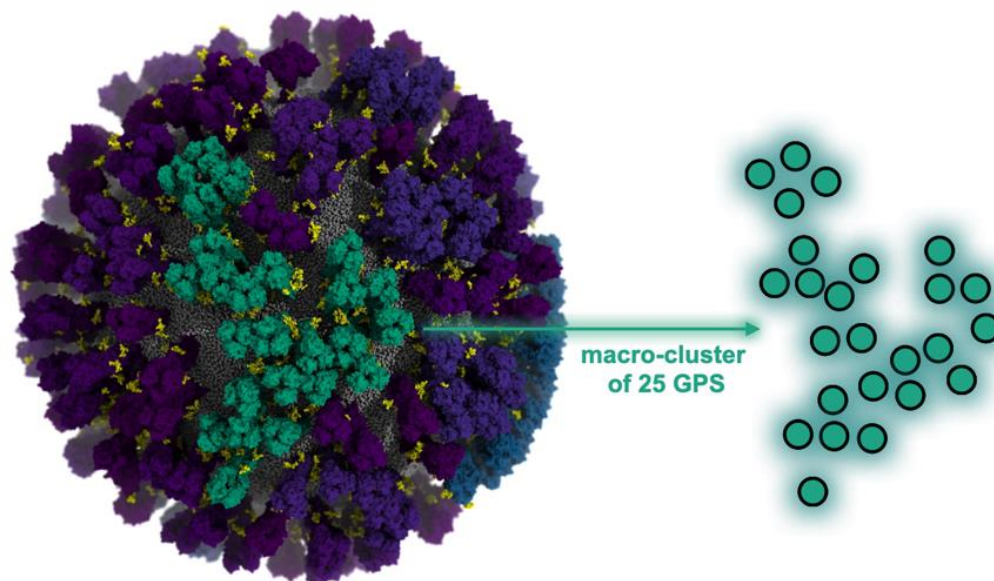


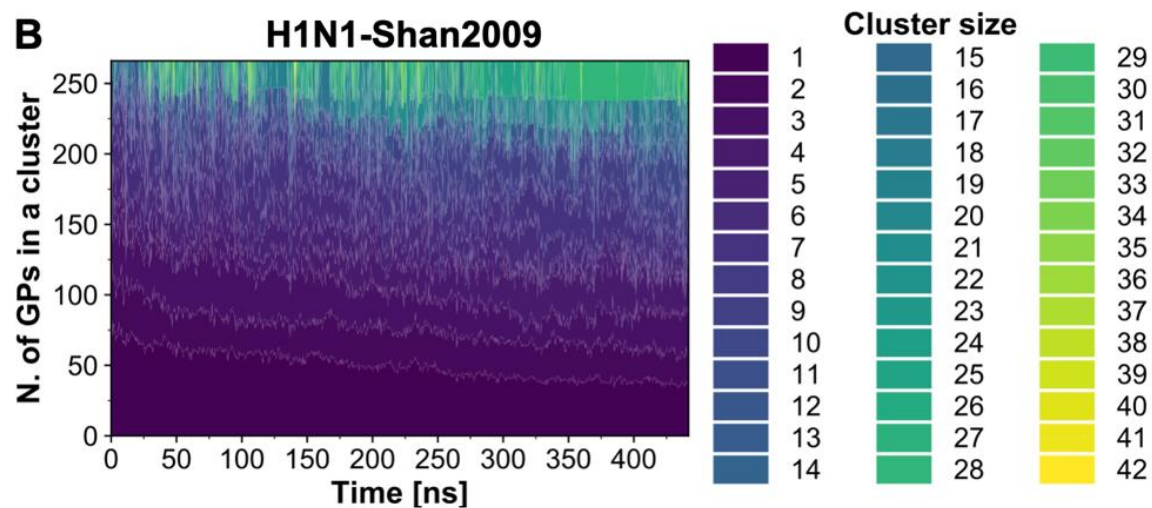
Figure 4.33 | Analysis of glycoprotein inter-connections for HA and NA. Stacked bar plots representing the fraction (%) of glycoproteins (HA top panels, NA bottom panels) establishing a certain number of connections (0 to 5, accordingly colored using the viridis palette) with surrounding glycoproteins along the simulation of (A) H1N1-Shan2009 and (B) H1N1-Mich2015. Percentages rounded to the first decimal digit are reported. A molecular representation of the HA head and NA head is also provided for both strains on the left side of each bar plot, highlighting with yellow vdW spheres the glycan that is added/deleted in H1N1-Mich2015: +1 glycan (N179) in H1N1-Mich2015 HA head and -1 glycan (N386) in H1N1-Mich2015 NA head.

Figure 4.34 | Analysis of glycoprotein macro-cluster formation. **A)** Molecular representation of the H1N1-Shan2009 virion at a frame of the simulation where a macro-cluster of 25 glycoproteins (GPs) is formed. The glycoproteins are depicted with a surface colored using a viridis palette according to the size of the macro-cluster that they belong to. The pattern of interactions of the macro-cluster of 25 GPs is highlighted using colored circles in the schematic on the right. **B-C)** Stacked plots representing the number of GPs involved in a macro-cluster of a certain size along the simulation of **(B)** H1N1-Shan2009 and **(C)** H1N1-Mich2015. Macro-clusters are colored using the *viridis* palette according to their size. The color gradient used for H1N1-Mich2015 (cluster size range: 0 to 34) has been normalized to the color gradient of H1N1-Shan2009 (cluster size range: 0 to 42).

A



B



C

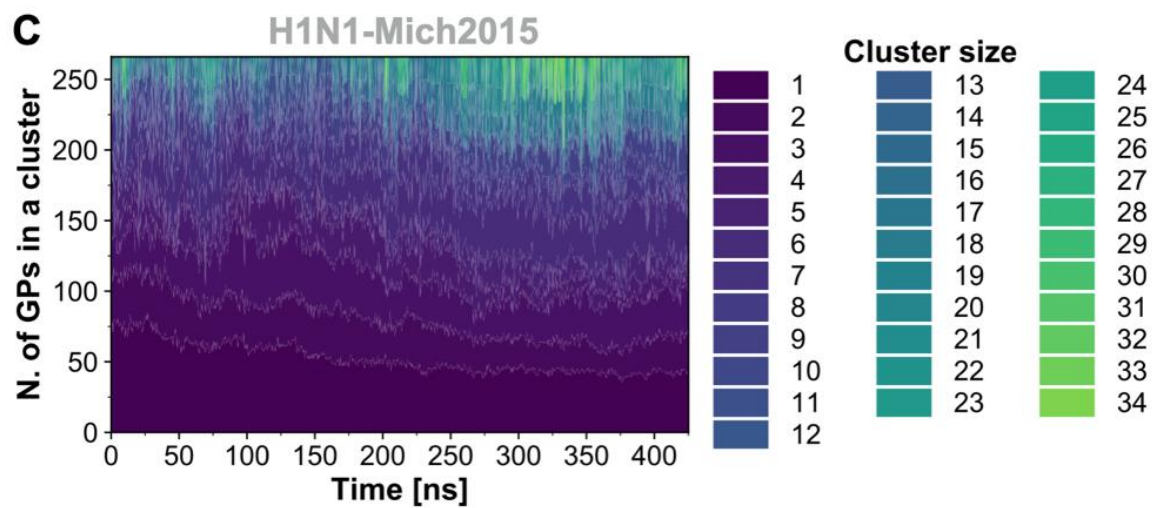
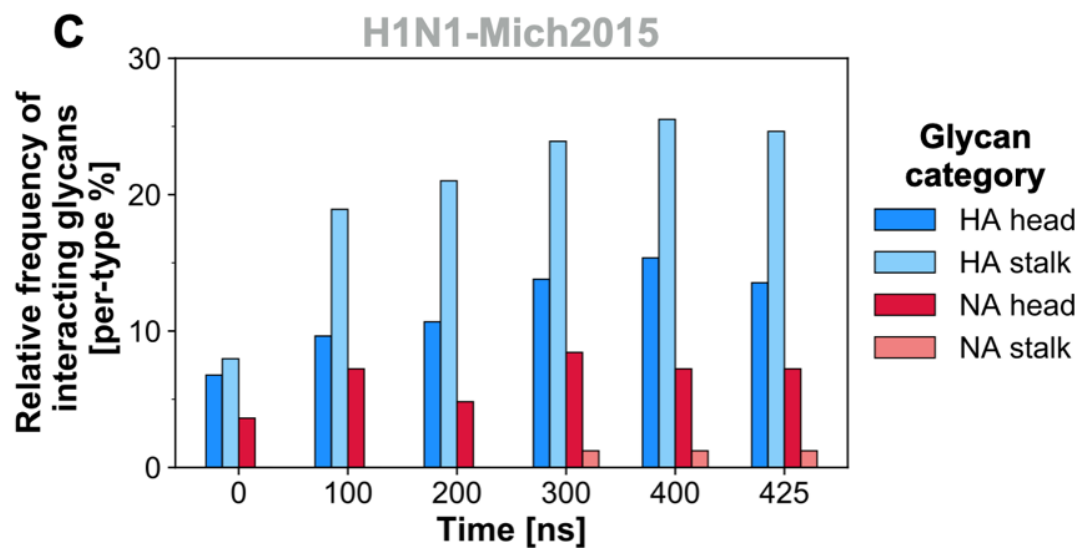
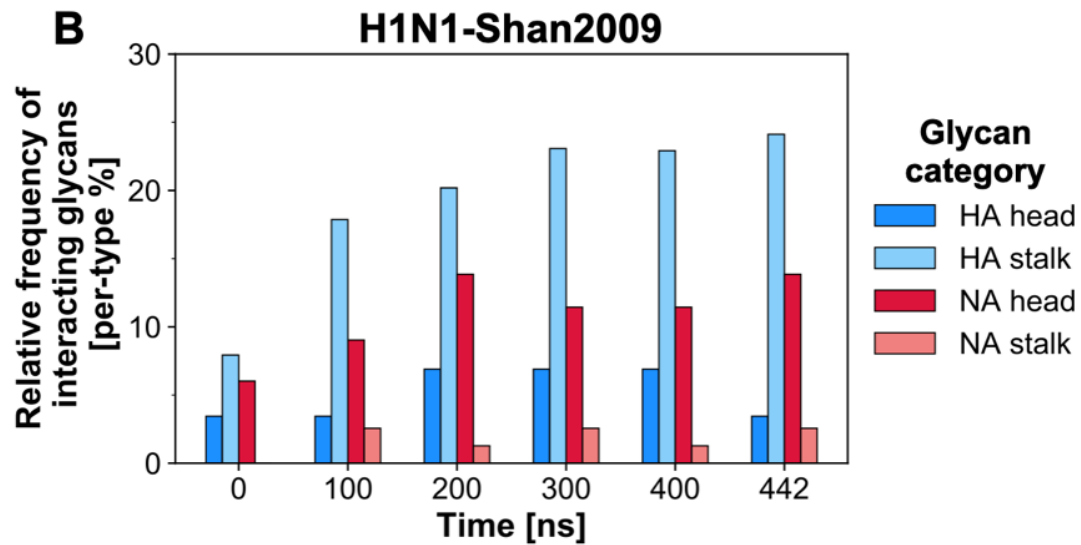
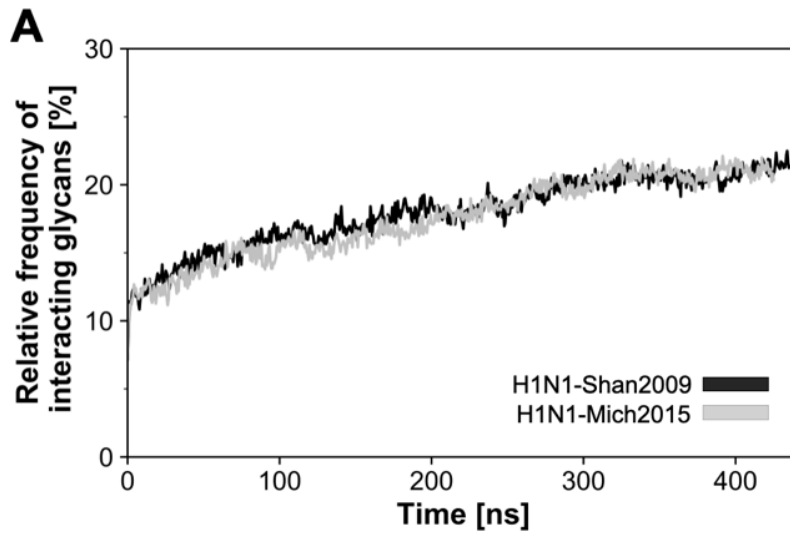


Figure 4.35 | Analysis of glycoprotein connections involving glycans. **A)** Time evolution (ns) of the relative frequency (%) of interacting glycans, i.e., glycans participating in glycoprotein inter-connections. The relative frequency of interacting glycan was calculated with respect to the total number of glycans. **B)** Grouped bar plot depicting the time evolution (ns) of the relative frequency (%) of interacting glycans in H1N1-Shan2009, where glycans have been divided into “head” and “stalk” categories according to their position, both for HA (shades of blue) and NA (shades of red). The relative frequency for each category of interacting glycans was calculated with respect to the total number of glycans of the same category. **C)** Grouped bar plot depicting the time evolution (ns) of the relative frequency (%) of interacting glycans in H1N1-Mich2015, where glycans have been divided into “head” and “stalk” categories according to their position, both for HA (shades of blue) and NA (shades of red). The relative frequency for each category of interacting glycans was calculated with respect to the total number of glycans of the same category.



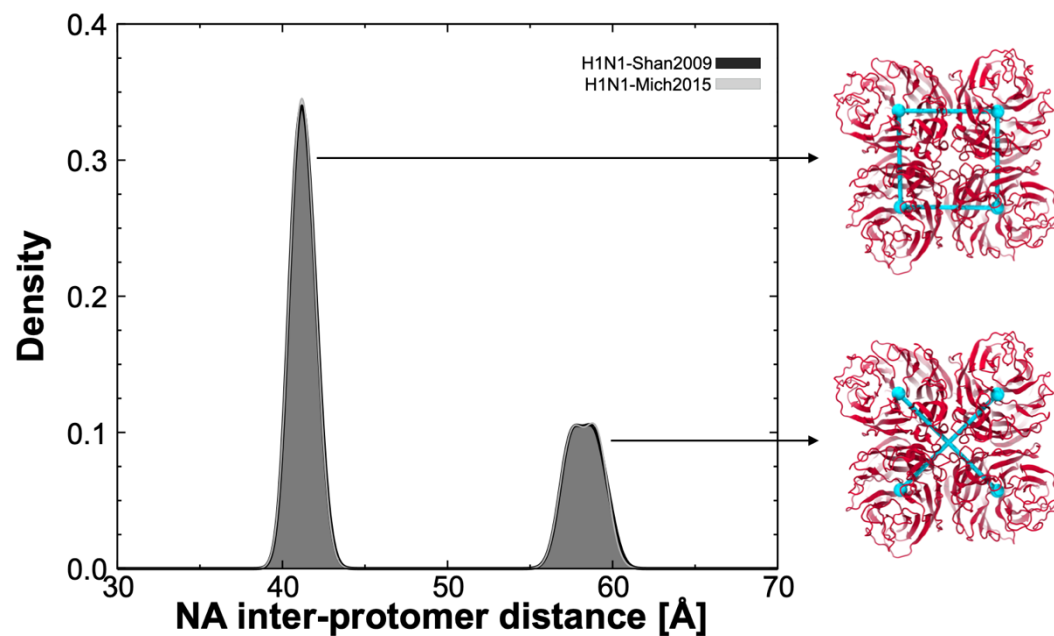


Figure 4.36 | Analysis of NA head breathing. Distribution of inter-protomer distances between the four NA head COMs in H1N1-Shan2009 (black) and H1N1-Mich2015 (gray) virions. The distribution is shown as a kernel density. On the right, a top view molecular representation of the NA head using red cartoons is provided, with the monitored distances between NA head COMs highlighted with cyan cylinders.

Supplementary Tables

Table 4.1 | Tilt angles (degree) of ten tilted and untilted structures (microstates) extracted from the MSM of the NA head tilting. Note: the average is the average of the ten selected microstates (ten selected conformational states and their respective angles). This will be representative of the average tilt angle of all microstates assigned to a macrostate of interest.

NA HEAD-TILT ANGLES [°]				
	H1N1-Shan2009		H1N1-Mich2015	
Structure #	<u>Untilted</u>	<u>Tilted</u>	<u>Untilted</u>	<u>Tilted</u>
#1	18.9°	81.9°	13.5°	90.0°
#2	20.3°	75.6°	19.7°	90.7°
#3	24.0°	76.7°	13.1°	88.1°
#4	22.9°	81.6°	14.4°	87.4°
#5	15.2°	79.5°	21.1°	89.7°
#6	23.6°	75.2°	14.4°	91.3°
#7	17.5°	74.2°	14.4°	87.0°
#8	22.4°	75.8°	20.7°	86.4°
#9	17.5°	78.1°	18.4°	89.6°
#10	32.5°	74.1°	22.1°	92.0°
average	21.5°	77.3°	17.2°	89.2°

Table 4.2 | Tilt angles (degree) of ten tilted and untilted structures (microstates) extracted from the MSM of the HA ectodomain tilting. Note: the average is the average of the ten selected microstates (ten selected conformational states and their respective angles). This will be representative of the average tilt angle of all microstates assigned to a macrostate of interest.

HA ECTODOMAIN-TILT ANGLES [°]				
	H1N1-Shan2009		H1N1-Mich2015	
Structure #	<u>Untilted</u>	<u>Tilted</u>	<u>Untilted</u>	<u>Tilted</u>
#1	8.1°	57.0°	19.1°	66.0°
#2	9.6°	57.3°	7.3°	61.9°
#3	10.0°	54.5°	19.8°	65.1°
#4	12.2°	57.3°	18.9°	61.3°
#5	5.9°	55.8°	5.9°	62.4°
#6	7.8°	56.9°	22.9°	66.2°
#7	21.1°	58.8°	6.2°	59.1°
#8	7.4°	58.2°	10.0°	61.8°
#9	9.8°	57.0°	14.8°	57.7°
#10	4.7°	55.3°	19.8°	67.9°
average	9.7°	56.8°	14.5°	62.9°

Table 4.3 | The table reports the distance between the COM of the HA monomer head (residues 66 to 286) and the COM of L α Hs' apical residue (419) calculated for open and closed structures (microstates) extracted from the MSM of the HA head breathing. The distances indicated here are the largest seen in each microstate, i.e., the largest of the three distances existing within the conformation of the HA trimer represented by the microstate. This applies also to the “closed” microstates, where the reported distance is still the largest distance seen in those microstates. Note: the average is the average of the ten selected microstates (ten selected conformational states and their respective angles). This will be representative of the average tilt angle of all microstates assigned to a macrostate of interest.

HA HEAD COM–L α Hs' apical residue COM DISTANCE [Å]				
	H1N1-Shan2009		H1N1-Mich2015	
Structure #	<u>Closed</u>	<u>Open</u>	<u>Closed</u>	<u>Open</u>
#1	27.3 Å	38.8 Å	27.3 Å	35.9 Å
#2	27.4 Å	37.8 Å	26.9 Å	42.2 Å
#3	28.1 Å	39.1 Å	27.0 Å	37.4 Å
#4	27.7 Å	38.7 Å	27.0 Å	36.6 Å
#5	27.6 Å	37.8 Å	26.9 Å	35.7 Å
#6	27.6 Å	36.7 Å	27.5 Å	37.3 Å
#7	27.4 Å	36.5 Å	26.9 Å	39.3 Å
#8	28.3 Å	36.7 Å	27.1 Å	35.6 Å
#9	28.3 Å	37.8 Å	27.2 Å	39.3 Å
#10	28.3 Å	36.0 Å	27.0 Å	36.5 Å
average	27.8 Å	37.6 Å	27.1 Å	37.6 Å

Table 4.4 | $P(A|B)$ for HA ectodomain tilting, where event A is the achievement of a certain tilt-angle swing range by the HA ectodomain during the dynamics with respect to the initial frame (calculated as $tilt_angle_frame_n - tilt_angle_frame_0$, with only positive values counted), whereas event B is the formation of one or more connections during the dynamics.

EVENTS	H1N-Shan2009		H1N-Mich2015	
	P(A)	P(A B)	P(A)	P(A B)
A: $0^\circ \leq \text{HA tilt-angle swing range} < 10^\circ$ B = 1 or more connections	$45.2 \pm 0.2 \%$	$44.6 \pm 0.3 \%$	$45.8 \pm 0.2 \%$	$45.7 \pm 0.3 \%$
A: $10^\circ \leq \text{HA tilt-angle swing range} < 20^\circ$ B = 1 or more connections	$19.2 \pm 0.1 \%$	$19.8 \pm 0.2 \%$	$17.6 \pm 0.1 \%$	$18.5 \pm 0.2 \%$
A: $20^\circ \leq \text{HA tilt-angle swing range} < 30^\circ$ B = 1 or more connections	$3.9 \pm 0.0 \%$	$4.3 \pm 0.1 \%$	$4.4 \pm 0.1 \%$	$4.5 \pm 0.1 \%$
A: $30^\circ \leq \text{HA tilt-angle swing range}$ B = 1 or more connections	$0.3 \pm 0.0 \%$	$0.4 \pm 0.0 \%$	$0.6 \pm 0.0 \%$	$0.6 \pm 0.0 \%$

Table 4.5 | P(A|B) for HA head breathing, where event A is the achievement of a certain extent of breathing by at least one of the HA head monomer within the same HA trimer during the dynamics with respect to the initial frame (calculated as *breathingdistance_frame_n* – *breathingdistance_frame_0*, with only positive values counted), whereas event B is the formation of one or more connections during the dynamics.

EVENTS	H1N-Shan2009		H1N-Mich2015	
	P(A)	P(A B)	P(A)	P(A B)
A: $1.0 \text{ \AA} \leq \text{breathing of one HA head} < 3.0 \text{ \AA}^\circ$ B: 1 or more connections	$46.9 \pm 0.2 \%$	$46.7 \pm 0.3\%$	$46.2 \pm 0.2 \%$	$46.1 \pm 0.3\%$
A: $3.0 \text{ \AA} \leq \text{breathing of one HA head} < 5.0 \text{ \AA}^\circ$ B: 1 or more connections	$11.2 \pm 0.1 \%$	$11.5 \pm 0.1 \%$	$11.0 \pm 0.1 \%$	$10.9 \pm 0.1 \%$
A: $5.0 \text{ \AA} \leq \text{breathing of one HA head} < 7.0 \text{ \AA}^\circ$ B: 1 or more connections	$2.2 \pm 0.0 \%$	$2.4 \pm 0.1 \%$	$2.8 \pm 0.0 \%$	$2.8 \pm 0.1 \%$
A: $7.0 \text{ \AA} \leq \text{breathing of one HA head}$ B: 1 or more connections	$0.2 \pm 0.0 \%$	$0.2 \pm 0.0 \%$	$0.3 \pm 0.0 \%$	$0.3 \pm 0.0 \%$

Table 4.6 | $P(A|B)$ for NA head tilting, where event A is the achievement of a certain tilt-angle swing range by the NA head during the dynamics with respect to the initial frame (calculated as $tilt_angle_frame_n - tilt_angle_frame_0$, with only positive values counted), whereas event B is the formation of one or more connections during the dynamics.

EVENTS	H1N-Shan2009		H1N-Mich2015	
	P(A)	P(A B)	P(A)	P(A B)
A: $0^\circ \leq \text{NA head tilt-angle swing range} < 10^\circ$ B: 1 or more connections	$32.1 \pm 0.4 \%$	$34.6 \pm 0.7 \%$	$29.9 \pm 0.4 \%$	$27.5 \pm 0.6 \%$
A: $10^\circ \leq \text{NA head tilt-angle swing range} < 20^\circ$ B: 1 or more connections	$18.9 \pm 0.3\%$	$18.0 \pm 0.5 \%$	$14.3 \pm 0.3 \%$	$18.8 \pm 0.5 \%$
A: $20^\circ \leq \text{NA head tilt-angle swing range} < 20^\circ$ B: 1 or more connections	$6.0 \pm 0.2 \%$	$4.9 \pm 0.2 \%$	$6.5 \pm 0.2 \%$	$8.2 \pm 0.3 \%$
A: $30^\circ \leq \text{NA head tilt-angle swing range} < 40^\circ$ B: 1 or more connections	$2.4 \pm 0.1 \%$	$2.5 \pm 0.2 \%$	$6.9 \pm 0.2 \%$	$8.0 \pm 0.3 \%$
A: $40^\circ \leq \text{NA head tilt-angle swing range}$ B: 1 or more connections	$1.8 \pm 0.1 \%$	$2.0 \pm 0.2 \%$	$1.2 \pm 0.1 \%$	$0.9 \pm 0.1 \%$

Table 4.7 | $P(B|A)$ for HA ectodomain tilting, where event A is the achievement of a certain tilt-angle swing range by the HA ectodomain during the dynamics with respect to the initial frame (calculated as $tilt_angle_frame_n - tilt_angle_frame_0$, with only positive values counted), whereas event B is the formation of one or more connections during the dynamics.

EVENTS	H1N-Shan2009		H1N-Mich2015	
	P(B)	P(B A)	P(B)	P(B A)
A: $10^\circ \leq \text{HA tilt-angle swing range} < 20^\circ$ B: 1 or more connections	$82.9 \pm 0.2 \%$	$81.8 \pm 0.5 \%$	$83.4 \pm 0.2 \%$	$83.1 \pm 0.5 \%$
A: $10^\circ \leq \text{HA tilt-angle swing range} < 20^\circ$ B: 1 or more connections	$82.9 \pm 0.2 \%$	$85.5 \pm 0.7 \%$	$83.4 \pm 0.2 \%$	$87.8 \pm 0.8 \%$
A: $20^\circ \leq \text{HA tilt-angle swing range} < 30^\circ$ B: 1 or more connections	$82.9 \pm 0.2 \%$	$91.1 \pm 1.6 \%$	$83.4 \pm 0.2 \%$	$85.5 \pm 1.5 \%$
A: $30^\circ \leq \text{HA tilt-angle swing range}$ B: 1 or more connections	$82.9 \pm 0.2 \%$	$99.8 \pm 5.8 \%$	$83.4 \pm 0.2 \%$	$95.1 \pm 4.4 \%$

Table 4.8 | P(B|A) for HA head breathing, where event A is the achievement of a certain extent of breathing by at least one of the HA head monomer within the same HA trimer during the dynamics with respect to the initial frame (calculated as *breathingdistance_frame_n* – *breathingdistance_frame_0*, with only positive values counted), whereas event B is the formation of one or more connections during the dynamics.

EVENTS	H1N-Shan2009		H1N-Mich2015	
	P(A)	P(B A)	P(A)	P(B A)
A: $1.0 \text{ \AA} \leq \text{breathing of one HA head} < 3.0 \text{ \AA}^\circ$ B: 1 or more connections	$82.9 \pm 0.2 \%$	$82.5 \pm 0.5 \%$	$83.4 \pm 0.2 \%$	$83.2 \pm 0.5 \%$
A: $3.0 \text{ \AA} \leq \text{breathing of one HA head} < 5.0 \text{ \AA}^\circ$ B: 1 or more connections	$82.9 \pm 0.2 \%$	$85.2 \pm 0.9 \%$	$83.4 \pm 0.2 \%$	$82.8 \pm 0.9 \%$
A: $5.0 \text{ \AA} \leq \text{breathing of one HA head} < 7.0 \text{ \AA}^\circ$ B: 1 or more connections	$82.9 \pm 0.2 \%$	$91.5 \pm 2.2 \%$	$83.4 \pm 0.2 \%$	$84.5 \pm 1.8 \%$
A: $7.0 \text{ \AA} \leq \text{breathing of one HA head}$ B: 1 or more connections	$82.9 \pm 0.2 \%$	$97.1 \pm 7.4 \%$	$83.4 \pm 0.2 \%$	$81.8 \pm 5.3 \%$

Table 4.9 | $P(B|A)$ for NA head tilting, where event A is the achievement of a certain tilt-angle swing range by the NA head during the dynamics with respect to the initial frame (calculated as *tilt-angle_frame_n* – *tilt-angle_frame_0*, with only positive values counted), whereas event B is the formation of one or more connections during the dynamics.

EVENTS	H1N-Shan2009		H1N-Mich2015	
	P(B)	P(B A)	P(B)	P(B A)
A: $0^\circ \leq \text{NA head tilt-angle swing range} < 10^\circ$ B: 1 or more connections	$65.8 \pm 0.5 \%$	$71.1 \pm 1.4 \%$	$56.6 \pm 0.4 \%$	$52.1 \pm 1.2 \%$
A: $10^\circ \leq \text{NA head tilt-angle swing range} < 20^\circ$ B: 1 or more connections	$65.8 \pm 0.5 \%$	$62.5 \pm 1.6 \%$	$56.6 \pm 0.5 \%$	$74.3 \pm 2.2 \%$
A: $20^\circ \leq \text{NA head tilt-angle swing range} < 20^\circ$ B: 1 or more connections	$65.8 \pm 0.5 \%$	$54.3 \pm 2.6 \%$	$56.6 \pm 0.5 \%$	$70.8 \pm 3.0 \%$
A: $30^\circ \leq \text{NA head tilt-angle swing range} < 40^\circ$ B: 1 or more connections	$65.8 \pm 0.5 \%$	$67.1 \pm 4.6 \%$	$56.6 \pm 0.5 \%$	$65.5 \pm 2.8 \%$
A: $40^\circ \leq \text{NA head tilt-angle swing range}$ B: 1 or more connections	$65.8 \pm 0.5 \%$	$73.5 \pm 5.6 \%$	$56.6 \pm 0.5 \%$	$45.0 \pm 5.2 \%$

Supplementary References

- (1) Durrant, J. D.; Kochanek, S. E.; Casalino, L.; Jeong, P. U.; Dommer, A. C.; Amaro, R. E. Mesoscale All-Atom Influenza Virus Simulations Suggest New Substrate Binding Mechanism. *ACS Cent Sci* **2020**, 6 (2), 189–196.
- (2) 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.
- (3) Jo, S.; Lim, J. B.; Klauda, J. B.; Im, W. CHARMM-GUI Membrane Builder for Mixed Bilayers and Its Application to Yeast Membranes. *Biophys J* **2009**, 97 (1), 50.
- (4) Jo, S.; Kim, T.; Im, W. Automated Builder and Database of Protein/Membrane Complexes for Molecular Dynamics Simulations. *PLoS One* **2007**, 2 (9), e880.
- (5) Vermaas, J. v.; Mayne, C. G.; Shinn, E.; Tajkhorshid, E. Assembly and Analysis of Cell-Scale Membrane Envelopes. *J Chem Inf Model* **2022**, 62 (3), 602–617.
- (6) Ivanova, P. T.; Myers, D. S.; Milne, S. B.; McClaren, J. L.; Thomas, P. G.; Brown, H. A. Lipid Composition of Viral Envelope of Three Strains of Influenza Virus – Not All Viruses Are Created Equal. *ACS Infect Dis* **2015**, 1 (9), 399–452.
- (7) Amaro, R. E.; Jeong, P. U.; Huber, G.; Dommer, A.; Steven, A. C.; Bush, R. M.; Durrant, J. D.; Votapka, L. W. A Computational Assay That Explores the Hemagglutinin/Neuraminidase Functional Balance Reveals the Neuraminidase Secondary Site as a Novel Anti-Influenza Target. *ACS Cent Sci* **2018**, 4 (11), 1570–1577.
- (8) Harris, A.; Cardone, G.; Winkler, D. C.; Heymann, J. B.; Brecher, M.; White, J. M.; Steven, A. C. Influenza Virus Pleiomorphy Characterized by Cryoelectron Tomography. *Proc Natl Acad Sci U S A* **2006**, 103 (50), 19123–19127.
- (9) Durrant, J. D.; McCammon, J. A. AutoClickChem: Click Chemistry in Silico. *PLoS Comput Biol* **2012**, 8 (3), e1002397.
- (10) Ropp, P.; Friedman, A.; Durrant, J. D. Scoria: A Python Module for Manipulating 3D Molecular Data. *J Cheminform* **2017**, 9 (1), 1–7.
- (11) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J Chem Phys* **1998**, 79 (2), 926.
- (12) Phillips, J. C.; Hardy, D. J.; Maia, J. D. C.; Stone, J. E.; Ribeiro, J. v.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W.; et al. Scalable Molecular Dynamics on CPU and GPU Architectures with NAMD. *J Chem Phys* **2020**, 153 (4), 044130.
- (13) Huang, J.; Mackerell, A. D. CHARMM36 All-Atom Additive Protein Force Field: Validation Based on Comparison to NMR Data. *J Comput Chem* **2013**, 34 (25), 2135–2145.
- (14) Klauda, J. B.; Venable, R. M.; Freites, J. A.; O'Connor, J. W.; Tobias, D. J.; Mondragon-Ramirez, C.; Vorobyov, I.; MacKerell, A. D.; Pastor, R. W. Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types. *Journal of Physical Chemistry B* **2010**, 114 (23), 7830–7843.
- (15) Klauda, J. B.; Monje, V.; Kim, T.; Im, W. Improving the CHARMM Force Field for Polyunsaturated Fatty Acid Chains. *Journal of Physical Chemistry B* **2012**, 116 (31), 9424–9431.
- (16) Beglov, D.; Roux, B. Finite Representation of an Infinite Bulk System: Solvent Boundary Potential for Computer Simulations. *J Chem Phys* **1998**, 100 (12), 9050.
- (17) Sharma, M.; Yi, M.; Dong, H.; Qin, H.; Peterson, E.; Busath, D. D.; Zhou, H. X.; Cross, T. A. Insight into the Mechanism of the Influenza A Proton Channel from a Structure in a Lipid Bilayer. *Science (1979)* **2010**, 330 (6003), 509–512.
- (18) Sun, S.; Wang, Q.; Zhao, F.; Chen, W.; Li, Z. Glycosylation Site Alteration in the Evolution of Influenza A (H1N1) Viruses. *PLoS One* **2011**, 6 (7), e22844.
- (19) Blake, T. A.; Williams, T. L.; Pirkle, J. L.; Barr, J. R. Targeted N-Linked Glycosylation Analysis of H5N1 Influenza Hemagglutinin by Selective Sample Preparation and Liquid Chromatography/Tandem Mass Spectrometry. *Anal Chem* **2009**, 81 (8), 3109–3118.
- (20) Kirkpatrick, E.; Qiu, X.; Wilson, P. C.; Bahl, J.; Krammer, F. The Influenza Virus Hemagglutinin Head Evolves Faster than the Stalk Domain. *Scientific Reports* **2018**, 8 (1), 1–14.
- (21) Pralow, A.; Hoffmann, M.; Nguyen-Khuong, T.; Pioch, M.; Hennig, R.; Genzel, Y.; Rapp, E.; Reichl, U. Comprehensive N-Glycosylation Analysis of the Influenza A Virus Proteins HA and NA from Adherent and Suspension MDCK Cells. *FEBS J* **2021**, 288 (16), 4869–4891.

- (22) Cruz, E.; Cain, J.; Crossett, B.; Kayser, V. Site-Specific Glycosylation Profile of Influenza A (H1N1) Hemagglutinin through Tandem Mass Spectrometry. *Hum Vaccin Immunother* **2017**, *14* (3), 508–517.
- (23) Khatri, K.; Klein, J. A.; White, M. R.; Grant, O. C.; Leymarie, N.; Woods, R. J.; Hartshorn, K. L.; Zaia, J. Integrated Omics and Computational Glycobiology Reveal Structural Basis for Influenza a Virus Glycan Microheterogeneity and Host Interactions. *Molecular and Cellular Proteomics* **2016**, *15* (6), 1895–1912.
- (24) Liu, Y. J.; Wu, S. L.; Love, K. R.; Hancock, W. S. Characterization of Site-Specific Glycosylation in Influenza A Virus Hemagglutinin Produced by Spodoptera Frugiperda Insect Cell Line. *Anal Chem* **2017**, *89* (20), 11036–11043.
- (25) She, Y. M.; Farnsworth, A.; Li, X.; Cyr, T. D. Topological N-Glycosylation and Site-Specific N-Glycan Sulfation of Influenza Proteins in the Highly Expressed H1N1 Candidate Vaccines. *Sci Rep* **2017**, *7* (1), 10232.
- (26) Danne, R.; Poojari, C.; Martinez-Seara, H.; Rissanen, S.; Lolicato, F.; Róg, T.; Vattulainen, I. DoGlycans-Tools for Preparing Carbohydrate Structures for Atomistic Simulations of Glycoproteins, Glycolipids, and Carbohydrate Polymers for GROMACS. *J Chem Inf Model* **2017**, *57* (10), 2401–2406.
- (27) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J Mol Graph* **1996**, *14* (1), 33–38.
- (28) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J Comput Chem* **2005**, *26* (16), 1781–1802.
- (29) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25.
- (30) Li, J.; Liu, S.; Gao, Y.; Tian, S.; Yang, Y.; Ma, N. Comparison of N-Linked Glycosylation on Hemagglutinins Derived from Chicken Embryos and MDCK Cells: A Case of the Production of a Trivalent Seasonal Influenza Vaccine. *Appl Microbiol Biotechnol* **2021**, *105* (9), 3559–3572.
- (31) Liu, Y. J.; Wu, S. L.; Love, K. R.; Hancock, W. S. Characterization of Site-Specific Glycosylation in Influenza A Virus Hemagglutinin Produced by Spodoptera Frugiperda Insect Cell Line. *Anal Chem* **2017**, *89* (20), 11036–11043.
- (32) Schnell, J. R.; Chou, J. J. Structure and Mechanism of the M2 Proton Channel of Influenza A Virus. *Nature* **2008**, *451* (7178), 591–595.
- (33) Pielak, R. M.; Chou, J. J. Influenza M2 Proton Channels. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2011**, *1808* (2), 522–529.
- (34) Guvench, O.; Hatcher, E.; Venable, R. M.; Pastor, R. W.; MacKerell, A. D. CHARMM Additive All-Atom Force Field for Glycosidic Linkages between Hexopyranoses. *J Chem Theory Comput* **2009**, *5* (9), 2353–2370.
- (35) *Preparing large MD simulations with VMD.* <https://www.ks.uiuc.edu/Research/vmd/minitutorials/largesystems/> (accessed 2022-06-01).
- (36) Brünger, A.; Brooks, C. L.; Karplus, M. Stochastic Boundary Conditions for Molecular Dynamics Simulations of ST2 Water. *Chem Phys Lett* **1984**, *105* (5), 495–500.
- (37) Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant Pressure Molecular Dynamics Algorithms. *J Chem Phys* **1998**, *101* (5), 4177.
- (38) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. Constant Pressure Molecular Dynamics Simulation: The Langevin Piston Method. *J Chem Phys* **1998**, *103* (11), 4613.
- (39) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *J Chem Phys* **1998**, *98* (12), 10089.
- (40) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J Comput Phys* **1977**, *23* (3), 327–341.
- (41) Williams, T.; Kelley, C.; Bersch, C.; Bröker, H.-B.; Campbell, J.; Cunningham, R.; Denholm, D.; Elber, G.; Fearick, R.; Grammes, C.; et al. *Gnuplot 5.4 An Interactive Plotting Program*; 1986.
- (42) Bangaru, S.; Lang, S.; Schotsaert, M.; Vandervan, H. A.; Zhu, X.; Kose, N.; Bombardi, R.; Finn, J. A.; Kent, S. J.; Gilchuk, P.; et al. A Site of Vulnerability on the Influenza Virus Hemagglutinin Head Domain Trimer Interface. *Cell* **2019**, *177* (5), 1136–1152.e18.
- (43) Shrake, A.; Rupley, J. A. Environment and Exposure to Solvent of Protein Atoms. Lysozyme and Insulin. *J Mol Biol* **1973**, *79* (2), 351–371.

- (44) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J. H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J Chem Theory Comput* **2015**, *11* (11), 5525–5542.
- (45) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; de Fabritiis, G.; Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *J Chem Phys* **2013**, *139* (1), 015102.
- (46) Molgedey, L.; Schuster, H. G. Separation of a Mixture of Independent Signals Using Time Delayed Correlations. *Phys Rev Lett* **1994**, *72* (23), 3634.
- (47) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J Chem Theory Comput* **2013**, *9* (4), 2000–2009.
- (48) Hochbaum D.; Shmoys D. A Best Possible Heuristic for the K-Center Problem. *Mathematics of Operations Research* **1985**, *10* (2), 180–184.
- (49) Swope, W. C.; Pitera, J. W.; Suits, F. Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory. *Journal of Physical Chemistry B* **2004**, *108* (21), 6571–6581.
- (50) Trendelkamp-Schroer, B.; Wu, H.; Paul, F.; Noé, F. Estimation and Uncertainty of Reversible Markov Models. *J Chem Phys* **2015**, *143* (17), 174101.
- (51) Prinz, J. H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov Models of Molecular Kinetics: Generation and Validation. *J Chem Phys* **2011**, *134* (17), 174105.
- (52) Metzler, R. Generalized Chapman-Kolmogorov Equation: A Unifying Approach to the Description of Anomalous Transport in External Fields. *Phys Rev E* **2000**, *62* (5), 6233.
- (53) Chapman, B. S. On the Brownian Displacements and Thermal Diffusion of Grains Suspended in a Non-Uniform Fluid. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **1928**, *119* (781), 34–54.
- (54) Kolmogoroff, A. Über Die Analytischen Methoden in Der Wahrscheinlichkeitsrechnung. *Math Ann* **1931**, *104* (1), 415–458.
- (55) Röblitz, S.; Weber, M. Fuzzy Spectral Clustering by PCCA+: Application to Markov State Models and Data Classification. *Adv Data Anal Classif* **2013**, *7* (2), 147–179.
- (56) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L. P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys J* **2015**, *109* (8), 1528–1532.
- (57) Stone, J. E.; Vandivort, K. L.; Schulten, K. GPU-Accelerated Molecular Visualization on Petascale Supercomputing Platforms. *Proceedings of the 8th International Workshop on Ultrascale Visualization* **2013**, 6.
- (58) Doria-Rose, N. A.; Bhiman, J. N.; Roark, R. S.; Schramm, C. A.; Gorman, J.; Chuang, G.-Y.; Pancera, M.; Cale, E. M.; Ernandes, M. J.; Louder, M. K.; et al. New Member of the V1V2-Directed CAP256-VRC26 Lineage That Shows Increased Breadth and Exceptional Potency. *J Virol* **2015**, *90* (1), 76–91.
- (59) Tiller, T.; Meffre, E.; Yurasov, S.; Tsuiji, M.; Nussenzweig, M. C.; Wardemann, H. Efficient Generation of Monoclonal Antibodies from Single Human B Cells by Single Cell RT-PCR and Expression Vector Cloning. *J Immunol Methods* **2008**, *329* (1–2), 112–124.
- (60) Mastronarde, D. N. Automated Electron Microscope Tomography Using Robust Prediction of Specimen Movements. *J Struct Biol* **2005**, *152* (1), 36–51.
- (61) Scheres, S. H. W. RELION: Implementation of a Bayesian Approach to Cryo-EM Structure Determination. *J Struct Biol* **2012**, *180* (3), 519–530.
- (62) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera—a Visualization System for Exploratory Research and Analysis. *J Comput Chem* **2004**, *25* (13), 1605–1612.
- (63) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J Mol Biol* **1990**, *215* (3), 403–410.
- (64) Stadlbauer, D.; Zhu, X.; McMahon, M.; Turner, J. S.; Wohlbold, T. J.; Schmitz, A. J.; Strohmeier, S.; Yu, W.; Nachbagauer, R.; Mudd, P. A.; et al. Broadly Protective Human Antibodies That Target the Active Site of Influenza Virus Neuraminidase. *Science* (1979) **2019**, *366* (6464), 499–504.
- (65) Saul, F. A.; Poljak, R. J. Crystal Structure of Human Immunoglobulin Fragment Fab New Refined at 2.0 Å Resolution. *Proteins: Structure, Function, and Bioinformatics* **1992**, *14* (3), 363–371.

(66) Benton, D. J.; Nans, A.; Calder, L. J.; Turner, J.; Neu, U.; Lin, Y. P.; Ketelaars, E.; Kallewaard, N. L.; Corti, D.; Lanzavecchia, A.; et al. Influenza Hemagglutinin Membrane Anchor. *Proc Natl Acad Sci U S A* **2018**, *115* (40), 10112–10117.

Chapter 5

Cell-scale Markov state models characterize the structural and dynamical basis of Group 1 influenza fusion inhibition

5.1 Abstract

Influenza virus infection continues to be a major healthcare issue and the combination of the strain-specificity of vaccines with the increased circulation of therapeutic resistant strains drives the need for new approaches for the prevention and treatment of infection. Arbidol is a broad-spectrum antiviral with demonstrated activity against both group 1 and 2 influenza hemagglutinin (HA) subtypes. While the arbidol binding site has been determined for group 2 HAs, it remains unidentified for group 1 HAs. Further, based on differences between group 1 and group 2 HAs, it is unknown whether or not arbidol binds in a group-specific manner. From all-atom, explicitly solvated simulations of the viral coat of the A/swine/Shandong/N1/2009(H1N1) and A/45/Michigan/2015(H1N1) strains of influenza, we extracted individual HA trajectories and analyzed them with the construction of Markov state models. Doing so allows us to investigate the structural dynamics and conformational transitions of the HA glycoprotein in the context of the crowded viral surface. From our Markov state models, we present the first evidence of a druggable pocket in a group 1 HA that is able to accommodate arbidol. This druggable site is located in an analogous position to the known group 2 arbidol binding site, suggesting that the binding of arbidol to influenza HA is not group specific. We also see a full helix turn that opens the arbidol binding site which can only be accomplished with the presence of glycans; the absence of glycans leads to a reduction of accessible conformational states. The insights obtained here illustrate the necessity of glycans in helix integrity and dynamics, and provide a structural basis for the mechanism of action for arbidol in a group 1 HA, thus facilitating the development of the next generation of influenza therapeutics.

5.2 Introduction

Influenza virus infection (IVI) is a major healthcare concern, with 3-5 million cases of severe disease reported and 300,000 to 500,000 deaths worldwide each year¹. Currently, vaccination is the primary strategy for the prevention and control of IVI. However, as a result of the rapid mutation of the virus, vaccine protection is strain-specific and a new vaccine must be re-engineered annually². Treatment of IVI is limited to two classes of FDA-approved drugs: matrix protein 2 inhibitors and neuraminidase inhibitors. However, many circulating strains of influenza exhibit resistance to one or more of these inhibitors³. The combination of strain-specific vaccines with the increased circulation of therapeutic resistant strains drives the need for a novel approach for the prevention and treatment of IVI⁴⁻⁵.

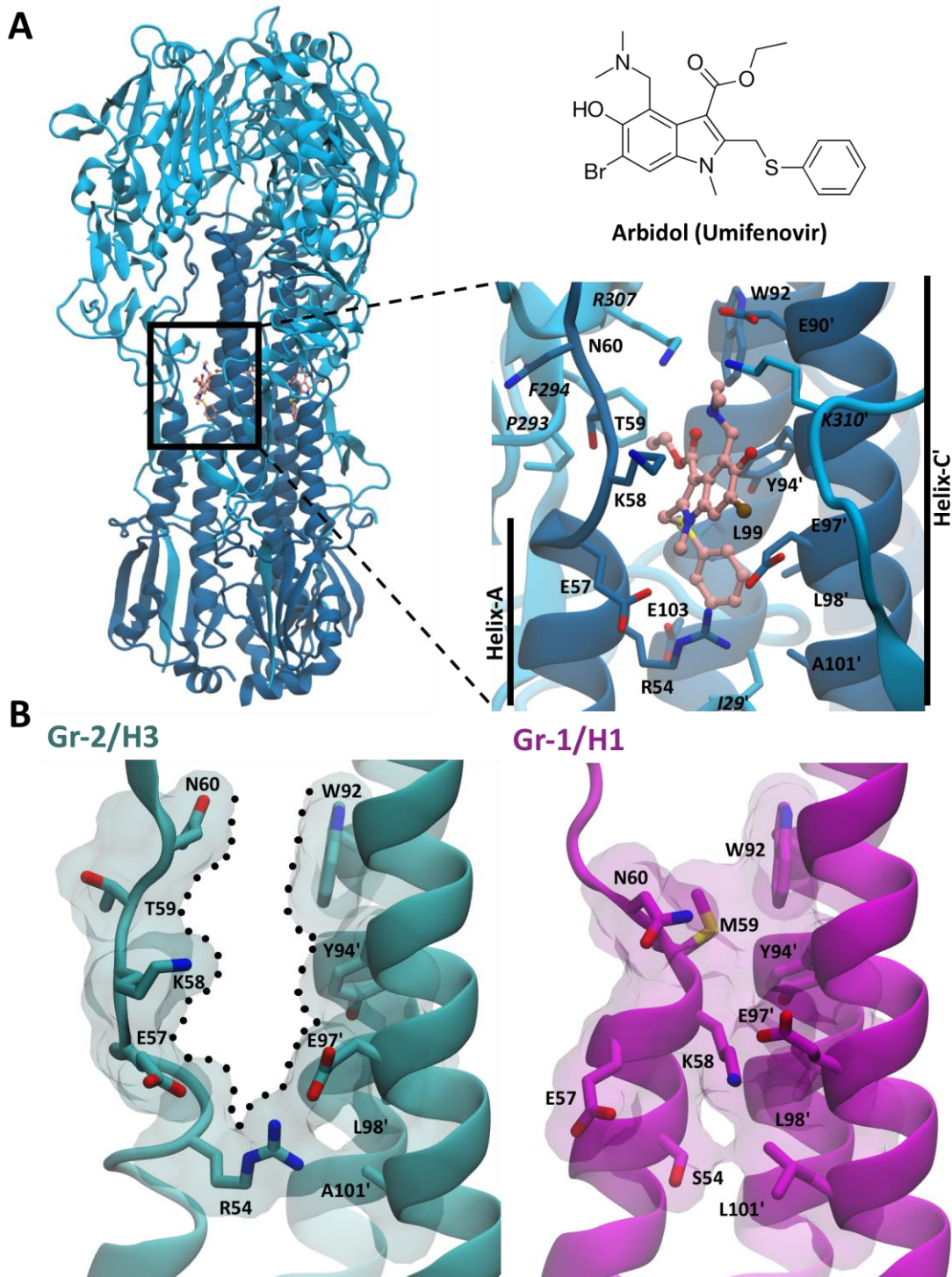
Structural understanding of the influenza virus and its components is critical for the development of novel prophylactic and therapeutic approaches. The influenza virus surface is decorated with the glycoproteins hemagglutinin (HA) and neuraminidase (NA). HA, which is present in the greatest numbers, is responsible for host cell binding and fusion. Following binding to the sialic acid recognition element, present on the host cell, the virus enters the cell by endocytosis. Within the endosome, a drop in pH triggers a large-scale rearrangement of HA to enable fusion of the viral and host cell membranes, allowing the release of viral ribonucleoproteins into the host cell cytoplasm for viral replication⁶. The unique structure of HA facilitates its function in the viral life cycle. The HA protein is trimeric, comprised of a head region (HA1) which contains the receptor binding site and a stem region (HA1/HA2) which contains the fusion machinery⁷. Whereas the head region of HA is highly variable, the HA stem region is largely conserved. Targeting this highly conserved and functionally critical stem region is therefore an enticing approach for the treatment of IVI.

One promising IVI treatment that targets the HA stem region is arbidol, a broad-spectrum antiviral with demonstrated activity against multiple strains of influenza A virus. Prior to the identification of the arbidol binding site, experiments demonstrated that arbidol binds group 1 HAs with higher affinity than group 2 HAs⁸. Recently, x-ray crystal structures of arbidol in complex with two group 2 HAs (H3 and H7) reveal that it binds in a hydrophobic cavity in the stem region between adjacent protomers⁹. The binding site is composed of A, C, and C' alpha-helices from HA2 ("C" refers to the adjacent protomer) and short C-terminal loops and an N-terminal beta-hairpin from HA1 (**Figure 5.1A**). To date, no structures of arbidol in complex with a group 1 HA

have been solved; potentially owing to structural differences between group 1 and group 2 HAs near the arbidol binding site (**Figure 5.1B**). Upon examination of the apo group 1 and 2 crystal structures, two major differences are evident: (1) the extra turn of helix A between residues 57-60, and (2) the presence of a salt bridge between residues K58-E97', both serving to occlude the arbidol binding site. These structural differences in conjunction with differential binding affinities make it unclear if arbidol binds in a group-specific manner, hindering its further development as a specific and optimized influenza therapeutic. Therefore, structural and dynamical insights into the potential of group 1 HA to bind arbidol are critical for the development of the next generation of influenza therapeutics.

Here we report propose that group 1 HA can accommodate arbidol; thus arbidol binding is not group specific between group 1 and group 2 HAs. Furthermore, we find an extension to the previously characterized arbidol pocket, which could be exploited by arbidol derivatives or other larger inhibitors. Finally, we see that the open conformation of the arbidol pocket in glycosylated HA contains a full helix turn, while unglycosylated HA unwinds and cannot fully turn, indicating that glycans play a role in HA helix stabilization.

Figure 5.1 | Arbidol binding site. a, Arbidol binds in a hydrophobic cavity between two adjacent protomers of HA. Residues with side chains that face towards the binding site are shown as sticks, arbidol is represented in ball and stick. Residues from HA1 (light blue) are italicized, residues from HA2 (dark blue) are written in standard font. **b,** Comparison of apo H3 (teal, PDB ID: 4FNK¹⁰) and H1 (magenta, PDB ID: 1RU7¹¹) HAs reveals two main structural differences: (1) the extra turn of helix-A between residues 57-60, and (2) the presence of the K58-E97' salt bridge in H1, that occlude the arbidol binding site.



5.3 Results and discussions

Markov state model construction. Recent work in our lab has led to the construction of atomically detailed, three-dimensional physical models of multiple variants of the influenza viral coat using integrative modeling

that combines high-resolution x-ray crystallography and homology modeling of the membrane-bound glycoproteins with cryo-electron tomography (**Figure 5.2**)¹⁰⁻¹¹. The systems are some of the largest simulated to date, each ~120 nm in diameter with ~160 million atoms. The A/swine/Shandong/N1/2009(H1N1) model, which we will refer to as H1N1-Shan2009, was selected to be studied in greater detail and was simulated for ~120 ns using all-atom molecular dynamics (MD) simulations¹⁰. This virion was then glycosylated, and split into two branches. One was simulated for an additional 450 ns, while the other was mutated to the A/45/Michigan/2015(H1N1) strain, hereafter referred to as the H1N1-Mich2015 model, and also simulated for another 450 ns¹¹. From each simulation, 234 individual HA trajectories have been extracted resulting in a total of ~29 μ s of unglycosylated HA for the H1N1-Shan2009 model, and ~105 μ s of glycosylated HA simulation for both the H1N1-Shan2009 and the H1N1-Mich2015 models. In this study, we present the first structural and kinetic basis for the group 1 HA arbidol binding site, identified and characterized by the construction of cell-scale, ensemble based Markov state models (MSMs)^{10, 12}. MSMs are a powerful tool enabling the study of long time-scale dynamics from many shorter, independent trajectories¹³⁻¹⁴, and as such are well-suited for our system. Cell-scale ensemble-based MSMs differ from the standard approach in that, in place of utilizing many distinct copies of single protein simulations, we instead draw our statistical sampling directly from the many copies of each substituent as simulated in the ‘mesoscape’¹⁵. This approach then enables the theoretically rigorous extraction of metastable protein states and the long timescale dynamics between these states; thus, the effects of protein crowding¹⁶⁻¹⁹ and correlated effects of neighboring components²⁰⁻²¹ are taken into account, which may be necessary to see rare biological events²².

Breaking the K58-E97’ salt bridge opens the arbidol pocket. We created MSMs for the unglycosylated and the glycosylated H1N1-Shan2009 model, and for the H1N1-Mich2015 model. To make for direct comparisons, we mapped the time-lagged independent component analysis (TICA) space of the glycosylated models onto the unglycosylated model (**Fig. 5.5**). We also kept the lag times consistent across models (**Fig. 5.6-5.8**). In each model, we found two major metastable states, representing open and closed conformations, differing primarily in the presence or absence of the K58-E97’ salt bridge (**Fig. 5.3**). For clarity, we define metastable states as relatively uniform macrostates clustered with high confidence through the standard MSM workflow from

discretized MD simulations, from which one can extract individual states (protein conformations) – this is a different meaning than when “metastable” is normally associated with viral proteins. The closed conformations in each MSM show a K58-E97’ salt bridge. The open conformations in each MSM show an absence of this salt bridge, but the “degree of openness” differs. In the unglycosylated H1N1-Shan2009 MSM, we see a clear breaking of the salt bridge. In the glycosylated models, the broken salt bridge is taken further, and we see helix-A and helix-C’ turning away in opposite directions until K58 and E97’ are $\sim 180^\circ$ from each other. As this is a larger conformational change than we see with the unglycosylated construct, it is unsurprising that the MFPTs show us it takes longer to create this “very open” conformation and longer to rebuild the salt bridge than the smaller conformational changes seen in the unglycosylated construct. The free energy landscapes for the glycosylated models fully encompass the unglycosylated model, but also show several islands that were not accessed in the unglycosylated model (**Fig. 5.5**).

Lack of glycans causes helix A to lose structural stability and unwind.

We naturally wondered why the unglycosylated model did not reproduce the large conformational changes seen in the glycosylated models. To investigate this, we created another MSM just of the unglycosylated H1N1-Shan2009 model. Without having to keep a consistent TICA space and lag time with the two glycosylated models, we could extend the lag time in the hopes of seeing a larger conformational change. This enabled us to see four major metastable states with exchange times on the order of microseconds, again differing primarily in the presence or absence of the K58-E97’ salt bridge but also in the conformation of residues 57-60 in helix-A (**Fig. S5**). These four states (representative structures shown in **Fig. S5**) represent diverse conformations, referred to as crystallographic closed (yellow), irregular (green), semi-open (blue), and open (red). The crystallographic closed state includes structures with crystal-like conformations where helix-A is unwound after K58 and the K58-E97’ salt bridge is present, completely occluding the arbidol pocket. In the irregular state, helix-A is kinked between residues 51-54. In the semi-open state, helix-A partially unwinds beginning at residue 57 and the arbidol pocket remains largely occluded by the K58-E97’ salt bridge. Finally, in the open state, helix-A is unwound after residue 56, and the K58-E97’ salt bridge is broken.

With respect to the transitions between the four macrostates, the crystallographic closed, irregular, and open states are well connected; however, the semi-open state is isolated and is only accessible from the open state (**Fig. 5.9**). Transitions between these states involve the folding or unfolding of helix-A and the breaking or forming of the K58-E97' salt bridge or some combination thereof. Since we see no transitions between the semi-open state and closed state, this suggests that the unwinding helix-A is coupled with the breaking of the salt bridge. Similarly, helix rewinding is linked to salt bridge formation. Analyzing the flux from the closed to the open state, the observed mean first-passage time (MFPT) from closed to open is $\sim 10 \mu\text{s}$ and from open to closed is $\sim 4 \mu\text{s}$. The ~ 2.5 -fold greater MFPT of the closed to open transition can be attributed to the stability resulting from both the presence of a salt bridge and the additional secondary structure.

Arbidol pocket structure and binding are not group specific. In the open structure of all of these MSMs, a primary pocket, formed by the same residues as the group 2 arbidol binding site, as well as an extended pocket beyond the primary binding site, have been identified (**Fig. 5.4**). The primary pocket is predominantly nonpolar, consistent with the group 2 binding site. The group 1 and group 2 pockets primarily differ in the conformation of loop-B. The C-alpha RMSD of all pocket residues is $2.1 \text{ \AA}$, whereas the C-alpha RMSD of loop-B (residues 56-60) is $4.4 \text{ \AA}$. The primary contribution to the ~ 2 -fold increase in RMSD can be attributed to the differing orientations of K58 and N60. In group 1, K58 is pointed inward and N60 is at the opening of the pocket; in group 2, K58 is at the opening of the pocket and N60 points inward. To assess the potential of the pocket to accommodate arbidol, rigid receptor docking was performed using Schrodinger's Glide²³. The resulting docked pose of arbidol is in good agreement with the crystallographic pose, retaining the relative positioning of the indole and thiophenol rings. The limitations of rigid-receptor docking preclude precise determination of the pose or interactions of arbidol within the H1 pocket, and therefore will be the topic of future work.

We also investigated the arbidol pocket volume across HA groups using POVME²⁴. The volume of a representative group 1 crystal structure (PDB ID: 1RU7) was calculated to be $483 \text{ \AA}^3$, while that of a representative group 2 crystal structure (PDB ID: 4FNK) was $3768 \text{ \AA}^3$. Considering apo group 1 HA contains a closed arbidol pocket (Fig. 1) while the group 2 pocket is open, this is unsurprising. We then extracted a representative structure from the "open" macrostate of the H1N1-Shan2009 glycosylated construct (a group 1

HA), and calculated its pocket volume to be 3790 Å³. Thus group 1 HA is capable of achieving the crystallographic open pocket volume of group 2 HA structures.

The arbidol pocket contains an extension amenable to future drug discovery work. In H1, the secondary pocket is an extension of the primary pocket and forms between the C, C', and C'' alpha-helices involving residues 104-106, 102'-106', and 102''-106''. In contrast to the primary pocket, which is predominantly nonpolar, the extended pocket is primarily composed of charged residues. The arbidol phenol ring docked into the arbidol pocket forms hydrophobic contacts with the edge of the pocket, but one could modify the thiophenol ring or potentially the ester group to access the arbidol pocket extension (**Fig. 5.10**). Computational solvent mapping with FTMap was used to identify putative ligand binding ‘hotspots’ on an open H1 structure²⁵. Three smaller hotspots are located within the primary pocket, while the largest hotspot is located in the extended pocket. These results are consistent with the characterization of the primary pocket as predominantly nonpolar and the extended pocket as predominantly charged. This extended pocket hotspot is also seen when FTMap is used on a representative structure from the viral coat simulations, which predicts polar, hydrophilic probes to bind there. Intriguingly, this pocket is not readily available in the crystal structures we analyzed. In a representative group 1 HA (PDB ID 1RU7) the pocket extension volume was calculated to be 354 Å³, while for a representative group 2 HA (PDB ID 4FNK) this pocket was 536 Å³. However, a representative structure extracted from the H1N1-2009Shan simulations had a pocket extension volume of 1178 Å³; thus the arbidol pocket extension may not be immediately apparent in crystal structures, but opens in a dynamic protein. We therefore highlight the potential of the extended pocket for forming stronger, more specific contacts than are available within the primary pocket. As such, the identification of this novel secondary pocket invites the possibility to develop more selective and potent antivirals via optimization of current compounds to exploit these novel specific interactions.

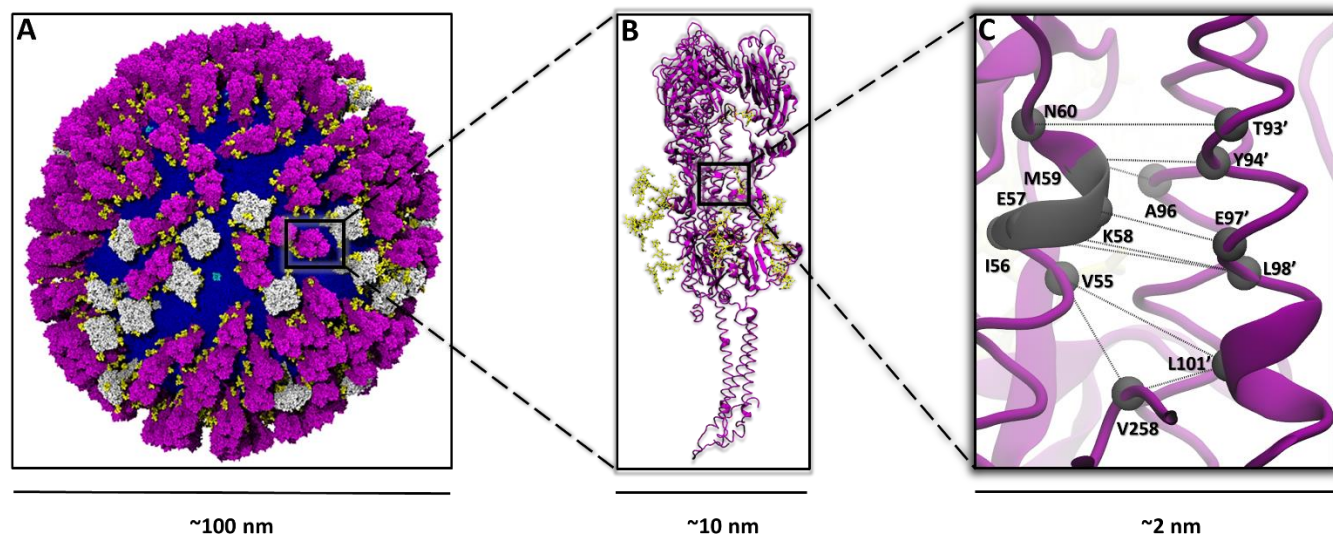


Figure 5.2 | Individual HA trajectories are extracted from the viral coat simulation and are then featured for analysis by the construction of an MSM. a, All-atom model of the H1N1 viral coat with HA in magenta, NA in white, and M2 ion channels in light blue. **b,** 234 HA trimer trajectories are extracted from the viral coat simulation, resulting in a total of ~105 μ s of HA simulation. **c,** 8 pairwise distances between alpha carbons of residues constituting the arbidol binding site as well as the phi and psi angles of residues 57-60 in helix-A were selected as the input features for model construction.

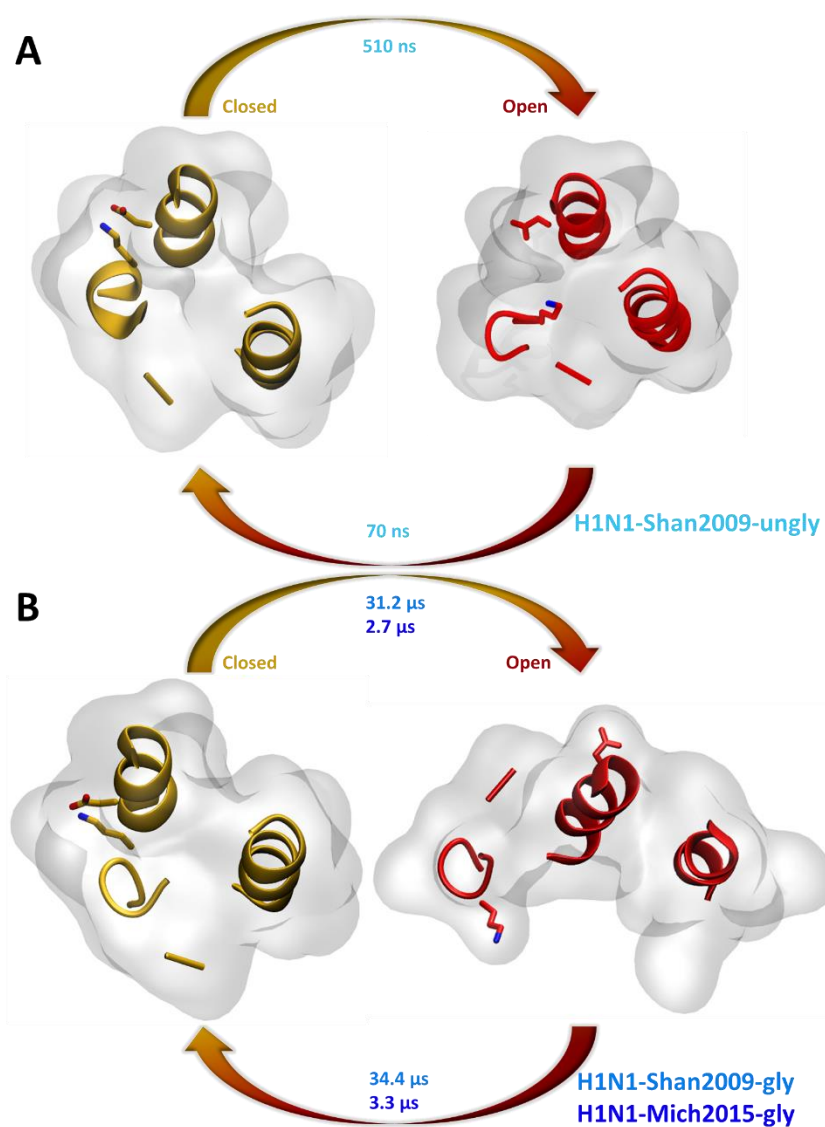


Figure 5.3 | Representative structures of the two major macrostates with mean first-passage times shown. Closed macrostates are yellow, open macrostates are red. **a**, The H1N1-Shan2009 unglycosylated model shows a breaking of the K58-E97' salt bridge. **b**, The glycosylated models show this broken salt bridge, and also show further turns by helix-A and helix-C. The H1N1-Shan2009 glycosylated model is shown here, which is nearly identical to the open and closed states seen in the H1N1-Mich2015 glycosylated model.

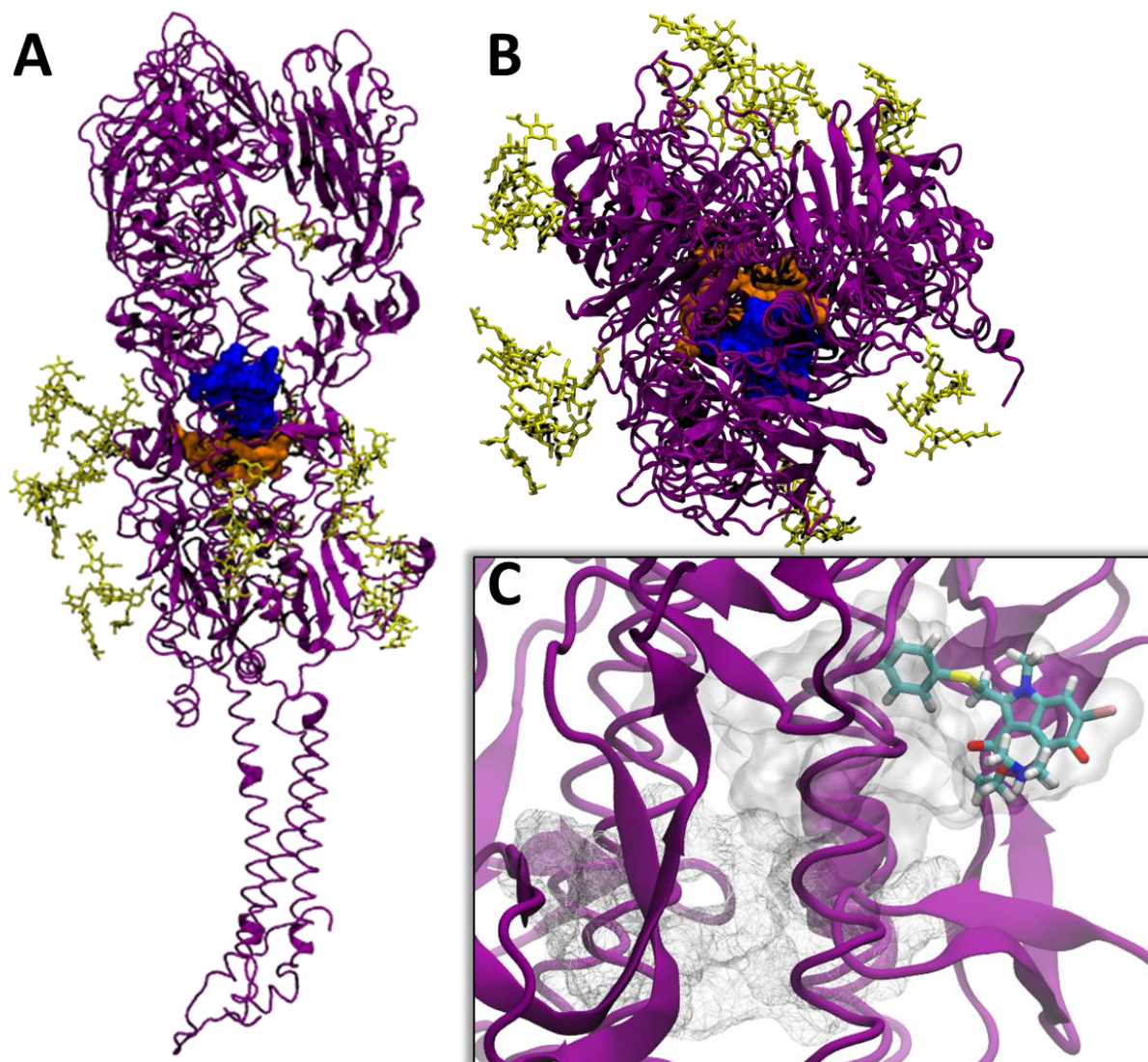


Figure 5.4 | View of the primary pocket and pocket extension. **a**, Side view of the H1N1-Shan2009 model, showing the HA protein (purple), glycans (yellow), crystallographic arbidol pocket (blue) and arbidol pocket extension (orange). **b**, Top down view of the H1N1-Shan2009 model. The arbidol pocket extension is a bit below, and slightly to the side, of the arbidol pocket. **c**, Arbidol docked into the arbidol pocket. The arbidol pocket is shown with a smooth transparent surface, while the pocket extension is shown with a wireframe surface. In contrast to the primary binding site, which is largely nonpolar, the secondary site is constituted of mainly charged residues.

5.4 Conclusions

In order to determine whether or not arbidol binding is group specific, we present an analysis of influenza viral coat MD simulations by construction of cell-scale MSMs to determine whether or not arbidol binding is group specific, and to investigate the effect of glycans on this putative binding site in a crowded protein environment. Helix A bounds part of the binding site; removing glycans on the HA structure causes this helix to unwind which restricts its range of motion. This finding underlines the importance of glycosylating proteins as necessary to faithfully capture structural dynamics in vivo. Furthermore, we identify the probable group 1 arbidol binding site at the same position as the group 2 binding site. This pocket not only shares the same location, but is also consistent in terms of polarity, shape, and size between group 1 and group 2 HAs. Subsequent rigid receptor docking demonstrates the capacity for arbidol to be accommodated within this pocket. In addition to the identification of this primary binding site, we identify and characterize a novel extension of this pocket. Solvent hotspot mapping highlights both the primary pocket and the pocket extension as druggable hotspots. Ultimately, identification and characterization of the primary pocket and pocket extension may promote the development of more specific and potent antiviral compounds.

5.5 Methods

System preparation and simulation

Complete system preparation and simulation details are outlined in ref¹⁰ and in ref¹¹.

Markov state models

All MSMs were constructed from the complete-coat simulations with trajectory frames taken every 0.02 ns. To capture the dynamics of the binding pocket in the MSM, 9 pairwise distances between alpha carbons of residues constituting the arbidol binding site as well as the phi and psi angles of residues 57-60 in helix-A were selected as the input features for model construction. We note that the pairwise distances explained more of the variance in the trajectories, and thus played a bigger role in MSM construction, than the torsional angles (**Fig. 5.11**). Dimensionality reduction was performed using time-lagged independent component analysis (TICA)²⁶⁻²⁸ and trajectory frames were clustered into 300 microstates by k-means clustering²⁹ in TICA space, as implemented in the PyEMMA software package³⁰. When creating MSMs of the first three models (unglycosylated H1N1-

Shan2009, glycosylated H1N1-Shan2009, glycosylated H1N1-Mich2009) a TICA lag time of 0.5 ns was selected for model construction; after examining our implied timescale plots³¹⁻³² we then selected an MSM lag time of 2 ns (**Fig. 5.12**). When extracting just the unglycosylated H1N1-Shan2009 model and creating an MSM with longer lag times, we used a TICA lag time of 20 ns and an MSM lag time of 20 ns. All resulting models were validated by the Chapman-Kolmogorov (CK) test³³⁻³⁸ (**Fig. 5.6-5.8**). We also made sure the CK test was valid out to thrice the MSM lag time, to ensure statistical validity³³. PCCA++³⁹ was used to cluster the states of all MSMs, while we also created a hidden Markov state model (HMM) was used for coarse-graining the MSM into metastable states for the extended lag time H1N1-Shan2009 model. In the short lag time unglycosylated H1N1-Shan2009 model, we identified 5 metastable states, three of which had a population of less than 0.3% and were thus excluded from further analysis. The two glycosylated models showed two metastable states each. In the long lag time unglycosylated H1N1-Shan2009 model, six metastable states were identified. Two of these states, both of which represent closed pocket conformations, had a total combined population of less than 1% and were excluded from further analysis.

5.6 Acknowledgements

This work was funded in part by the Director's New Innovator Award Program NIH DP2 OD007237 and the National Biomedical Computation Resource (NBCR) through NIH P41 GM103426 to REA. Computing support on the NSF Blue Waters Petascale Computer was provided through NSF OAC-1811685. We thank the NSF for access to the TACC Stampede supercomputer through CHE060073N to REA. C.S. acknowledges support from the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1650112. S.E.K. acknowledges support from the NIH Molecular Biophysics Training Program (T32-GM008326).

Chapter 5, in full, is a modified reprint of the material as it appears in "Seitz, C.*, Kochanek, S.*, Durrant, J., Casalino, L., Amaro, R.E. Cell-scale Markov state models characterize the structural and dynamical basis of Group 1 influenza fusion inhibition. *In revisions at Cell CHEM*" The dissertation author was a co-investigator and co-author of this work.

5.7 Author Contributions

REA designed and oversaw the research project. REA, CS and SEK performed the simulation analysis. JD and LC created trajectories. CS and SEK created the figures. REA, CS and SEK wrote the paper.

5.8 Supplementary Information

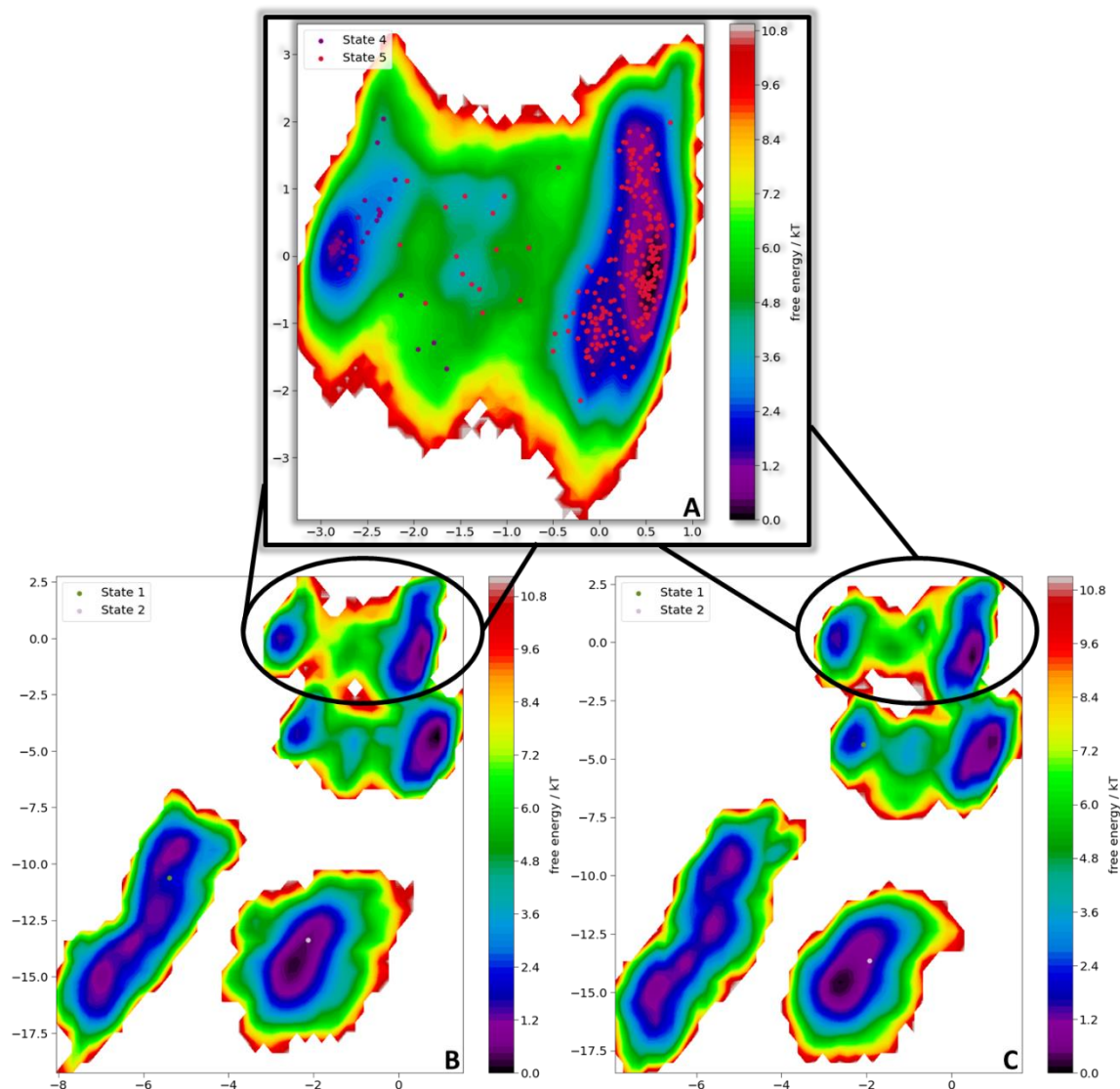


Figure 5.5 | a, The free energy landscape for the H1N1-Shan2009 unglycosylated model. State 4 is an open arbidol site conformation and state 5 is closed. States 1-3 had population percentages under 0.3% and so were excluded from analysis. Note how this free energy landscape can be found in the upper right hand corner of the following two free energy landscapes – this is possible as they were created in the same TICA space, allowing for direct comparison. **b,** The free energy landscape for the H1N1-Shan2009 glycosylated model. State 1 is a closed arbidol site conformation and state 2 is open. **c,** The free energy landscape for the H1N1-Mich2015 glycosylated model. State 1 is an open arbidol site conformation and state 2 is closed. Note how for all three pairs, the closed arbidol site macrostates are in a lower free energy well than the open macrostates.

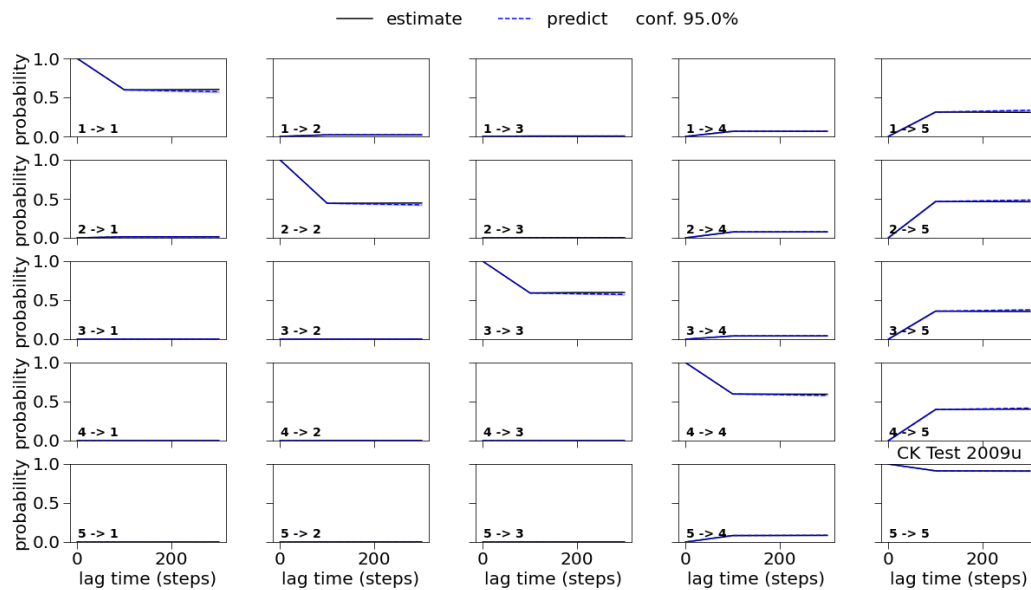


Figure 5.6 | CK test of the H1N1-Shan2009 unglycosylated MSM. With an MSM lag time of 100 steps, this validates.

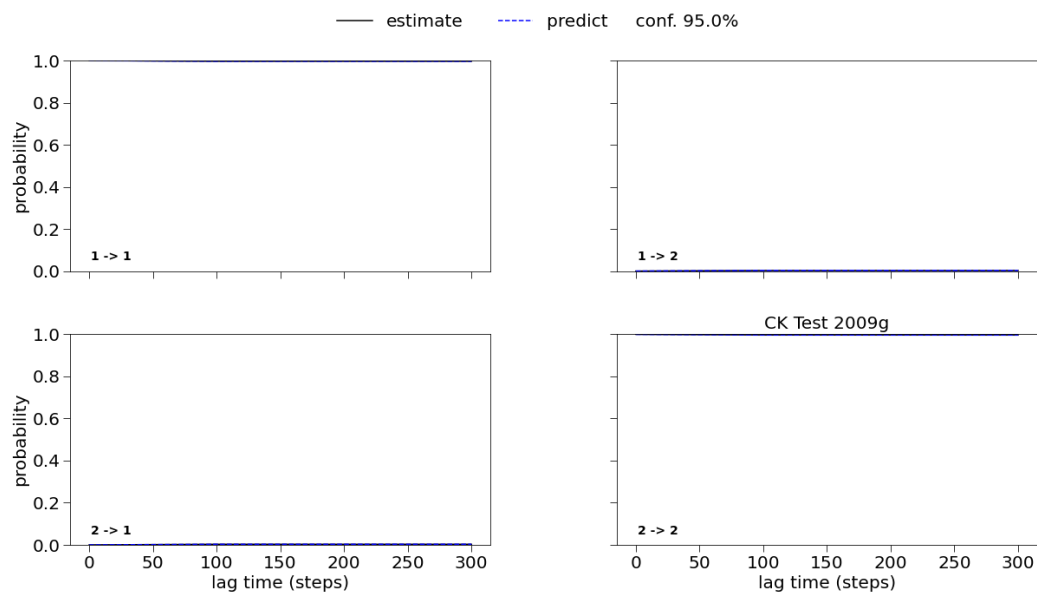


Figure 5.7. CK test of the H1N1-Shan2009 glycosylated MSM. With an MSM lag time of 100 steps, this validates.

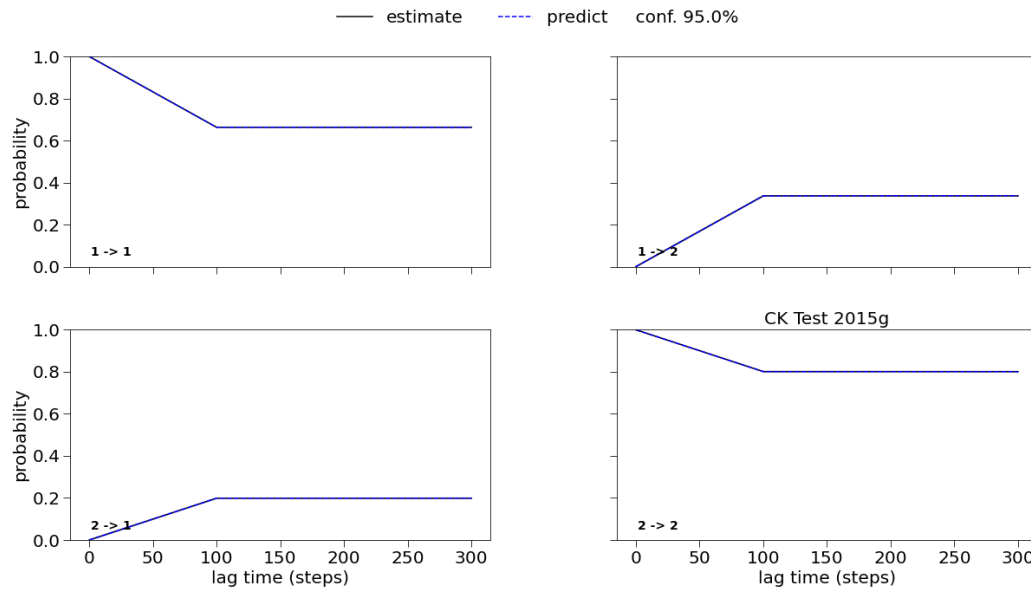


Figure 5.8 | CK test of the H1N1-Mich2015 glycosylated MSM. With an MSM lag time of 100 steps, this validates.

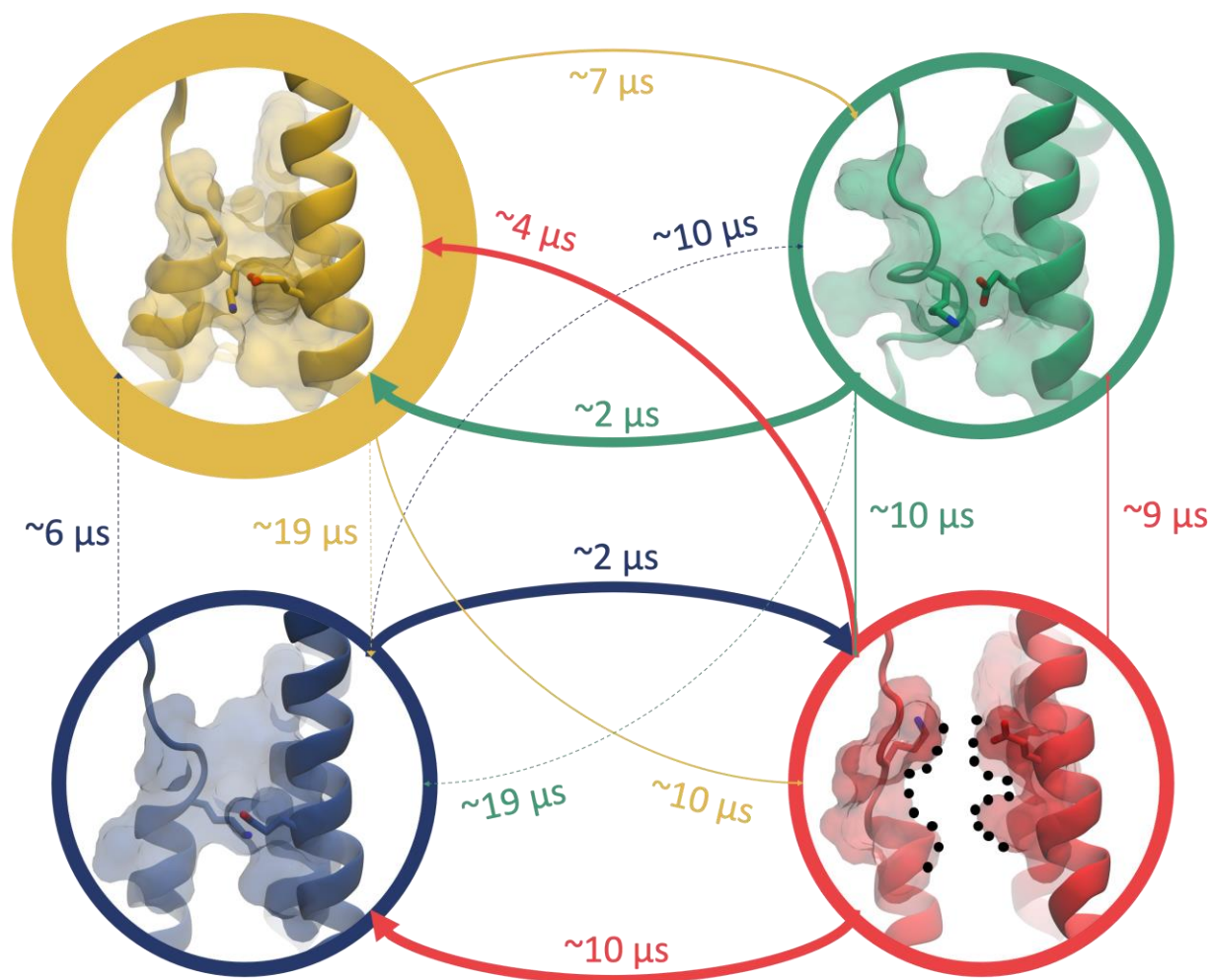


Figure 5.9 | Representative structures of the four major macrostates of the unglycosylated H1N1-Shan2009 model with mean first-passage times shown. Macrostates are colored as follows: crystallographic closed (yellow), irregular (green), semi-open (blue), open (red). The width of the colored outline of each state is proportional to its population and the width of each arrow is proportional to the transition probability. Arrows representing transitions are colored by the state from which they originate. (Note: MFPTs indicated are calculated for all potential pathways.)

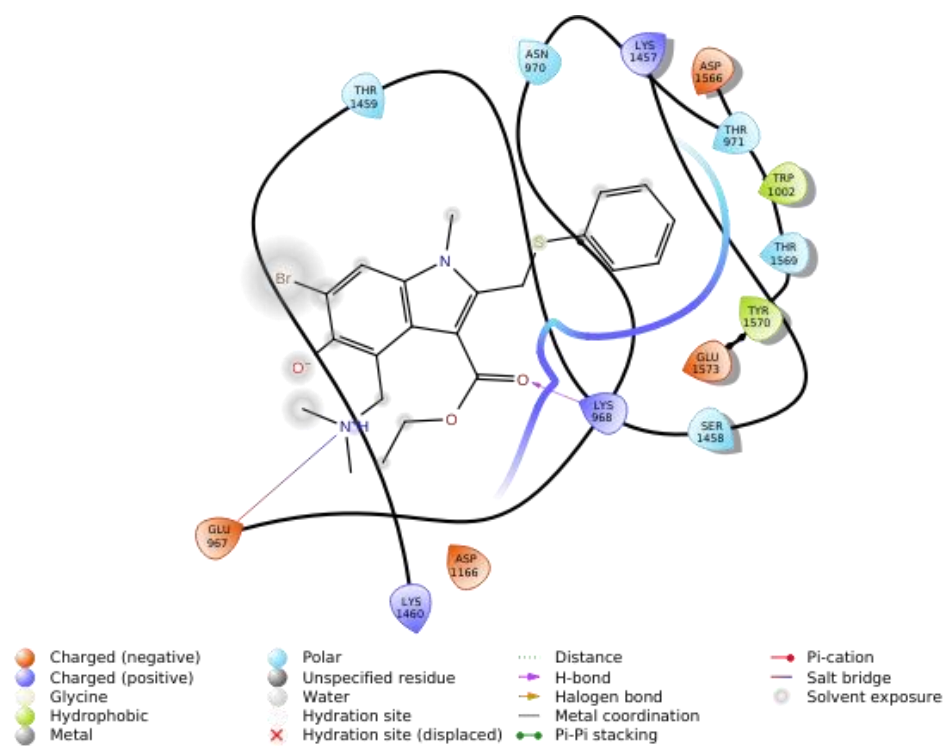


Figure 5.10 | Ligand interaction diagram of arbidol in its binding pocket. The thiophenol or ester groups present opportunities for modification to access the arbidol pocket extension.

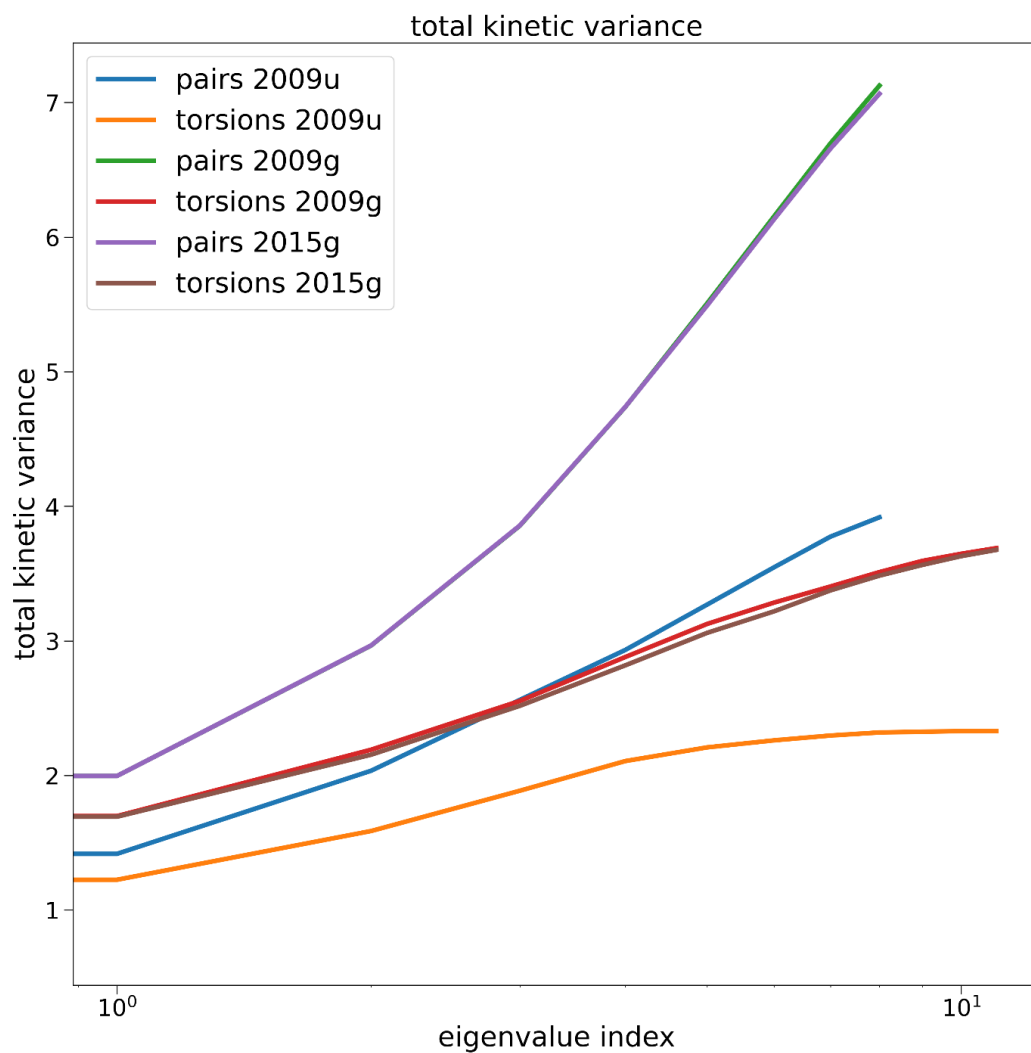


Figure 5.11 | The total kinetic variance of the MSM features for each MSM (2009u = H1N1-Shan2009 unglycosylated model, 2009g = H1N1-Shan2009 glycosylated model, 2015g = H1N1-Mich2015 glycosylated model). The features that show higher variance explain more of the data, and thus are more influential and necessary in MSM creation.

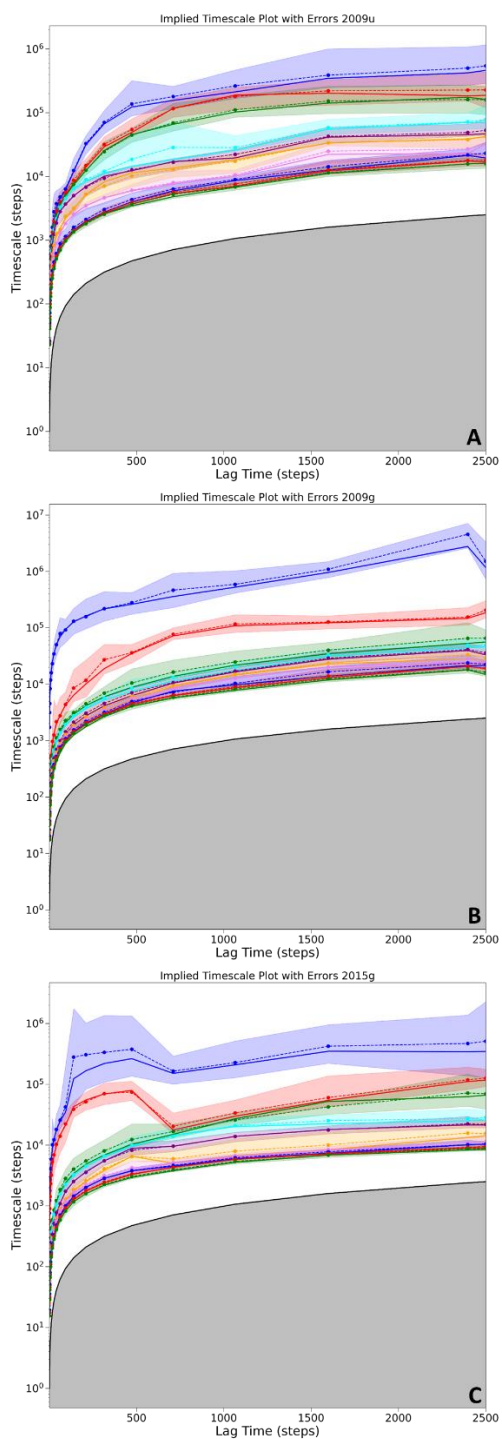


Figure 5.12 | **a**, Implied timescale plot for the H1N1-Shan2009 unglycosylated mode. **b**, Implied timescale plot for the H1N1-Shan2009 glycosylated model. **c**, Implied timescale plot for the H1N1-Mich2015 glycosylated model.

References

1. Lambert, Linda C.; Fauci, Anthony S., Influenza Vaccines for the Future. *N Engl J Med* **2010**, 363 (21), 2036-2044.
2. Fiore, Anthony E.; Uyeki, Timothy M.; Broder, Karen; Finelli, Lyn; Euler, Gary L.; Singleton, James A.; Iskander, John K.; Wortley, Pascale M.; Shay, David K.; Bresee, Joseph S.; Cox, Nancy J., Prevention and control of influenza with vaccines: recommendations of the Advisory Committee on Immunization Practices. *Morb Mortal Wkly Rep* **2010**, 59 (RR08), 1-62.
3. Clerq, Erik De, Antiviral agents active against influenza A viruses. *Nature Reviews Drug Discovery* **2006**, 5, 1015-1025.
4. Samson, Mélanie; Pizzorno, Andrés; Abed, Yacine; Boivin, Guy, Influenza virus resistance to neuraminidase inhibitors. *Antivir Res* **2013**, 98 (2), 174-185.
5. Dong, Guoying; Peng, Chao; Luo, Jing; Wang, Chengmin; Han, Le; Wu, Bin; Ji, Guangju; He, Hongxuan, Adamantane-resistant influenza A viruses in the world (1902-2013): Frequency and distribution of M2 gene mutations. *PLoS One* **2015**.
6. Skehel, John J.; Wiley, Don C., Receptor Binding and Membrane Fusion in Virus Entry: The Influenza Hemagglutinin. *Annu Rev Biochem* **2000**, 69, 531-569.
7. Wilson, Ian A.; Skehel, John J.; Wiley, Don C., Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 Å resolution. *Nature* **1981**, 289.
8. Brancato, Virginia; Peduto, Antonella; Wharton, Stephen; Martin, Stephen; More, Vijaykumar; Mola, Antonia Di; Massa, Antonio; Perfetto, Brunella; Donnarumma, Giovanna; Schiraldi, Chiara; Tufano, Maria Antonietta; Rosa, Mario de; Filosa, Rosanna; Hay, Alan, Design of inhibitors of influenza virus membrane fusion: Synthesis, structure-activity relationship and in vitro antiviral activity of a novel indole series. *Antivir Res* **2013**, 99 (2), 125-135.
9. Kadam, Rameshwar U.; Wilson, Ian A., Structural basis of influenza virus fusion inhibition by the antiviral drug Arbidol. *Proceedings of the National Academy of Sciences* **2017**, 114 (2), 206-214.
10. Durrant, Jacob D.; Kochanek, Sarah E.; Casalino, Lorenzo; Jeong, Pek U.; Dommer, Abigail C.; Amaro, Rommie E., Mesoscale All-Atom Influenza Virus Simulations Suggest New Substrate Binding Mechanism. *ACS Central Science* **2020**, 6 (2), 189-196.
11. Casalino, Lorenzo; Seitz, Christian; Lederhofer, Julia; Tsybovsky, Yaroslav; Wilson, Ian A.; Kanekiyo, Masaru; Amaro, Rommie E., Breathing and tilting: mesoscale simulations illuminate influenza glycoprotein vulnerabilities. *bioRxiv* **2022**.
12. Amaro, Rommie E.; Mulholland, Adrian J., Bridging Chemical and Biological Complexity in the Search for Cures: Multiscale Methods in Drug Design. *Nature Reviews Chemistry* **2018**, 2.
13. Husic, Brooke E.; Pande, Vijay S., Markov State Models: From an Art to a Science. *J Am Chem Soc* **2018**, 140, 2386-2396.
14. Chodera, John D; Noe, Frank, Markov state models of biomolecular conformational dynamics. *Curr Opin Struct Biol* **2014**, 25, 135-144.
15. Johnson, Graham T.; Autin, Ludovic; Al-Alusi, Mostafa; Goodsell, David S.; Sanner, Michel F.; Olson, Arthur J., cellPACK: a virtual mesoscope to model and visualize structural systems biology. *Nat Meth* **2015**, 12, 85-91.
16. Rickard, Meredith M.; Zhang, Yi; Gruebele, Martin; Pogorelov, Taras V., In-Cell Protein-Protein Contacts: Transient Interactions in the Crowd. *Journal of Physical Chemistry Letters* **2019**, 10 (18), 5667-5673.
17. Nawrocki, Grzegorz; Wang, Po-hung; Yu, Isseki; Sugita, Yuji; Feig, Michael, Slow-Down in Diffusion in Crowded Protein Solutions Correlates with Transient Cluster Formation. *Journal of Physical Chemistry B* **2017**, 121 (49), 11072-11084.
18. Bülow, Sören von; Siggel, Marc; Linke, Max; Hummer, Gerhard, Dynamic cluster formation determines viscosity and diffusion in dense protein solutions. *Proceedings of the National Academy of Sciences* **2019**, 116 (20), 9843-9852.
19. McGuffee, Sean R.; Elcock, Adrian H., Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm. *PLoS Computational Biology* **2010**, 6 (3).

20. Ando, Tadashi; Skolnick, Jeffrey, Crowding and hydrodynamic interactions likely dominate in vivo macromolecular motion. *Proceedings of the National Academy of Sciences* **2010**, 107 (43), 18457-18462.
21. Mereghetti, Paolo; Wade, Rebecca C., Atomic Detail Brownian Dynamics Simulations of Concentrated Protein Solutions with a Mean Field Treatment of Hydrodynamic Interactions. *Journal of Physical Chemistry B* **2012**, 116 (29), 8523-8533.
22. Turoňová, Beata; Sikora, Mateusz; Schürmann, Christoph; Hagen, Wim J.H.; Welsch, Sonja; Blanc, Florian E.C.; Bülow, Sören von; Gecht, Michael; Bagola, Katrin; Hörner, Cindy; Zandbergen, Ger van; Landry, Jonathan; Azevedo, Nayara Trevisan Doimo de; Mosalaganti, Shyamal; Schwarz, Andre; Covino, Roberto; Mühlebach, Michael D; Hummer, Gerhard; Locker, Jacomine Krijnse; Beck, Martin, In situ structural analysis of SARS-CoV-2 spike reveals flexibility mediated by three hinges. *Science* **2020**, 370 (6513), 203-208.
23. Schrödinger, LLC *Maestro*, New York, NY, 2020.
24. Wagner, Jeffrey R.; Sørensen, Jesper; Hensley, Nathan; Wong, Celia; Zhu, Clare; Perison, Taylor; Amaro, Rommie E., POVME 3.0: Software for Mapping Binding Pocket Flexibility. *Journal of Chemical Theory and Computation* **2017**, 13 (9), 4584-4592.
25. Kozakov, Dima; Grove, Laurie E; Hall, David R; Bohnuud, Tanggis; Mottarella, Scott E; Luo, Lingqi; Xia, Bing; Beglov, Dmitri; Vajda, Sandor, The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nature Protocols* **2015**, 10 (5).
26. Pérez-Hernández, Guillermo; Paul, Fabian; Giorgino, Toni; Fabritiis, Gianni De; Noé, Frank, Identification of slow molecular order parameters for Markov model construction. *Journal of Chemical Physics* **2013**, 139 (1).
27. Molgedey, Lutz; Schuster, Heinz G., Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters* **1994**, 72 (23), 3634-3637.
28. Schwantes, Christian R.; Pande, Vijay S., Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *Journal of Chemical Theory and Computation* **2013**, 9 (4), 2000-2009.
29. Steinhaus, Hugo, Sur la division des corps matériels en parties. *Bulletin de l'Académie Polonaise des Sciences, Classe 3* **1956**, 4.
30. Scherer, Martin K.; Trendelkamp-Schroer, Benjamin; Paul, Fabian; Pérez-Hernández, Guillermo; Hoffmann, Moritz; Plattner, Nuria; Wehmeyer, Christoph; Prinz, Jan-Hendrik; Noé, Frank, PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *Journal of Chemical Theory and Computation* **2015**, 11, 5525-5542.
31. Trendelkamp-Schroer, Benjamin; Wu, Hao; Paul, Fabian; Noé, Frank, Estimation and uncertainty of reversible Markov models. *Journal of Chemical Physics* **2015**, 143 (17).
32. Swope, William C.; Pitera, Jed W.; Suits, Frank, Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory. *Journal of Physical Chemistry B* **2004**, 108 (21), 6571-6581.
33. Prinz, Jan-Hendrik; Wu, Hao; Sarich, Marco; Keller, Bettina; Senne, Martin; Held, Martin; Chodera, John D.; Schütte, Christof; Noé, Frank, Markov models of molecular kinetics: Generation and validation. *Journal of Chemical Physics* **2011**, 134 (17).
34. Chapman, Sydney, On the Brownian displacements and thermal diffusion of grains suspended in a non-uniform fluid. *Proceedings of the Royal Society A* **1928**, 119 (781), 34-54.
35. Kolmogoroff, Andrey, Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. *Mathematische Annalen* **1931**, 104 (1), 415-458.
36. Metzler, Ralf, Generalized Chapman-Kolmogorov equation: A unifying approach to the description of anomalous transport in external fields. *Physical Review E* **2000**, 62 (5).
37. Bachelier, Louis, Théorie de la spéculation. *Annales scientifiques de l'École Normale Supérieure* **1900**, 3 (17), 21-86.
38. Smoluchowski, Marian von, Über Brownsche Molekularbewegung unter Einwirkung äußerer Kräfte und deren Zusammenhang mit der verallgemeinerten Diffusionsgleichung. *Annalen der Physik* **1916**, 353 (24), 1103-1112.
39. Röblitz, Susanna; Weber, Marcus, Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification. *Advances in Data Analysis and Classification* **2013**, 7.

Part III
Density functional theory

Chapter 6

Data for molecular dynamics simulations of *Escherichia coli* cytochrome *bd* oxidase with the Amber force field

6.1 Abstract

Cytochrome *bd*-type quinol oxidase is an important metalloenzyme that allows many bacteria to survive in low oxygen conditions. Since *bd* oxidase is found in many prokaryotes but not in eukaryotes, it has emerged as a promising bacterial drug target. Examples of organisms containing *bd* oxidases include the *Mycobacterium tuberculosis* (*Mtb*) bacterium that causes tuberculosis (TB) in humans, the *Vibrio cholerae* bacterium that causes cholera, the *Pseudomonas aeruginosa* bacterium that contributes to antibiotic resistance and sepsis, and the *Campylobacter jejuni* bacterium that causes food poisoning. *Escherichia coli* (*E. coli*) is another organism exhibiting the cytochrome *bd* oxidase. Since it has the highest sequence identity to *Mtb* (36%) and we are ultimately interested in finding drug targets for TB, we have built parameters for the *E. coli* *bd* oxidase (Protein Data Bank ID number: 6RKO) that are compatible with the all-atom Amber ff14SB force field for molecular dynamics (MD) simulations. Specifically, we built parameters for the three heme cofactors present in all species of bacterial cytochrome *bd*-type oxidases (heme b558, heme b595, and heme d) along with their axial ligands. This data report includes the parameter and library files that can be used with Amber's LEaP program to generate input files for MD simulations using the Amber software package. We also provide the PDB data files of the initial model both by itself and solvated with TIP3P water molecules and counterions.

6.2 Specifications table

Subject	Chemistry, Biology
Specific subject area	Computational Chemistry, Biophysics
Type of data	Amber OFF library file, Amber parameter files, PDB files, XYZ files
How data were acquired	Density functional theory calculations with Gaussian 16; model preparation with Maestro (Schrödinger), CHARMM-GUI membrane builder, AmberTools, and MCPB.py; molecular dynamics simulations with Amber20
Data format	Raw PDB coordinate files, Raw XYZ coordinate files, Raw Amber OFF library file, and Raw Amber parameter files
Parameters for data collection	Starting geometries based on the cryo-EM structure from Protein Data Bank ID number: 6RKO
Description of data collection	Geometry optimizations, frequency calculations, and electrostatic potential (ESP) charge derivation of the three heme cofactors with axial amino acid ligands with BP86 and B3LYP functionals with empirical dispersion; molecular dynamics minimization, heating, and equilibration with Amber ff14SB, Generalized Amber force field (GAFF), Lipid 14 force field, TIP3P water molecules, Joung/Cheatham counterions, and new parameters for the heme cofactors
Data source location	University of California San Diego, La Jolla, CA, USA
Data accessibility	Data are within this article and available on https://doi.org/10.5281/zenodo.4742814 ¹
Value of the Data	

- We report new force field parameters for the three heme cofactors present in all cytochrome *bd*-type oxidases, including heme d, which has not been previously reported in the literature.
- The new parameters will be useful to investigate cytochrome *bd* oxidase using molecular dynamics (MD) simulations.
- By using the structural ensembles of *bd* oxidase obtained from MD simulations, we can carry out virtual screening for inhibitors, which could potentially lead to new drugs for a variety of bacteria, such as tuberculosis (TB), cholera, *Pseudomonas aeruginosa*, *Campylobacter jejuni* and other prokaryotes.

6.3 Data description

The data provided in this article can be used for running molecular dynamics (MD) simulations of *Escherichia coli* (*E. coli*) cytochrome *bd* oxidase (Protein Data Bank ID number: 6RKO).² The structure provided in this publication in the PDB file format (6rko_final.pdb, protein.pdb, protein.rst7) has been modelled to contain some of the missing residues in the experimental cryo-EM structure (Protein Data Bank ID number: 6RKO). Specifically, residues 240–248 and 334–339 in chain A were modelled with the Maestro software from Schrödinger.³ Also, missing atoms in the palmitoyl-oleoyl-phosphatidylcholine (POPC) lipids were added back with the aid of the CHARMM-GUI membrane builder.⁴ The structure of this system after solvation with TIP3P water molecules and counterions is also provided (protein_solv.pdb, protein_solv.rst7), as well as the structure of this system after minimization, heating and equilibration (protein_solv_equil.pdb, protein_solv_equil.rst7). All library and parameter files compatible with the Amber ff14SB force field needed to generate this system are also available (heme.lib, heme.frcmod, mcpbpy.frcmod, ubiquinone.prepi, ubiquinone.frcmod), along with the Amber LEaP script (tleap.in) that builds the complete system, yielding the parameter/topology and coordinate files for molecular dynamics simulations with Amber (protein.frcmod, protein_solv.frcmod). Finally, the optimized geometry of the three heme cofactors are provided in the XYZ file format (heme_b558.xyz, heme_b595.xyz, and heme_d.xyz). The repository includes all these data files and no other additional files are included.

6.4 Experimental design, materials and methods

6.4.1. Model preparation (PDB data file generation)

The initial atomic coordinates were taken from the Protein Data Bank ID number: 6RKO,² which is the 2.7 Å resolution cryo-EM structure of *Escherichia coli* (*E. coli*) cytochrome *bd* oxidase. Using the Maestro software from Schrödinger,³ the hydrogens were added to the residues, and the protonation states of the residues were assigned using PROPKA⁵ at pH 7.5, same as in Ref.². The placement of the hydrogens for the heme cofactors and the protonation states of several residues referred to in Ref.², which are residues nearby the heme cofactors that are involved in the proton and oxygen pathways, were matched to Ref.². The positions of the water molecules present in the cryo-EM structure were kept fixed. In addition, the missing residues 240–248 and 334–339 in chain A were modelled using Maestro, and the missing atoms in the palmitoyl-oleoyl-phosphatidylcholine (POPC) lipids were added back using the CHARMM-GUI membrane builder.⁴ For the solvated system, TIP3P water molecules⁶ and counterions (Na⁺, Cl[−]) were added to neutralize the system and to set the ionic strength to 0.10 M, same as in Ref.². The water molecules were added so that there would be at least 15 Å between the protein and the simulation box boundaries.

The atomic coordinate data of the initial structure with the aforementioned preparation steps (protein.pdb) is provided, as well as of the solvated structure (protein_solv.pdb). In addition, the atomic coordinates of the system after equilibration (protein_solv_equil.pdb) is provided. The equilibrated structure was obtained by first minimizing the system for 10,000 steps with the backbone atoms and heavy atoms of the heme cofactors restrained with a force constant of 10 kcal/(mol × Å²). Then, the system was gradually heated from 10 to 300 K over 4 ns and kept at 300 K for 6 ns in the NVT ensemble (using a Langevin thermostat with a friction coefficient γ of 5.0 ps^{−1}) with the backbone atoms and heavy atoms of the heme cofactors restrained with a force constant of 1.0 kcal/(mol × Å²). Finally, the system was equilibrated for 10 ns in the NPT ensemble at 300 K and 1 bar (using a Langevin thermostat with a friction coefficient γ of 1.0 ps^{−1} and a Berendsen barostat with a time constant τ of 1.0 ps) with the backbone atoms and heavy atoms of the heme cofactors restrained with a force constant of 0.1 kcal/(mol × Å²). The root-mean-square deviation (RMSD) of all non-hydrogen atoms from the starting structure for the equilibration run plateaued during the equilibration run. All of these steps were executed using the Amber20 MD software package.⁷ The Amber LEaP script that builds the initial dry and the

solvated structures (tleap.in) is provided in this publication. This script uses the Amber ff14SB force field⁸ for the standard residues, TIP3P water molecules⁶, Joung/Cheatham ion parameters⁹ for the counterions (Na⁺, Cl⁻), Generalized Amber force field (GAFF)¹⁰ for ubiquinone, Amber Lipid 14 force field¹¹ for the POPC lipids, and the force field parameters for non-standard residues taken from Ref.¹² (heme.frcmod) or derived using MCPB.py¹³ (mcpbpy.frcmod). More details for the MCPB.py derivation will be provided in the following sections.

6.4.2. Geometry optimization

Geometry optimizations were performed on the three heme cofactors (heme b558, heme b595, and heme d) using Gaussian 16.¹⁴ For all three heme cofactors, the side chains of the axial amino acid ligands were retained and the C α atom was turned into a methyl group. The C α and C β atoms were kept fixed to preserve the axial ligands' placement within the protein. For heme b595, four carbons in the pyrrole rings were also fixed. This was done because heme b595 has only one axial ligand.

We followed the *as isolated* mixed valence A1 state of the enzyme from Fig. S10 of Ref.² to determine the charges and multiplicities of the three heme cofactors. All of the propionates, except for the one in heme d, were protonated during the geometry optimizations so that the partial charges for all atoms, including the hydrogen in the hydroxyl group, could be obtained for the propionates. When making the final structure files, however, all of the propionates, except for one of the propionates of heme b558, were deprotonated since they coordinate with positively charged residues nearby in the cryo-EM structure.² Heme b558 is a low-spin heme (Fe3⁺) with neutral axial ligands (Met 393.A and His 186.A), so its overall charge was 0 and spin multiplicity was 2. Heme b595 is high-spin heme (Fe3⁺) with a deprotonated (thus negatively charged -1) coordinating axial ligand (Glu 445.A), making the overall charge -2 and spin multiplicity 6. While transitioning through state A1, heme d releases a water molecule, gets reduced by an internal electron redistribution, and quickly binds a dioxygen O₂ molecule. We modelled the A1 state without a bound dioxygen molecule. The unbound state contains a high-spin heme Fe2⁺, and is the fully reduced state of the enzyme, i.e., one step in the reaction cycle before dioxygen coordinates to heme d to form state A1 (not shown in Fig. S10 of Ref.²). This unbound state has a -1 charge and a spin multiplicity of 5.

The initial geometry optimizations for heme b558, heme b595 and heme d were performed with the BP86 functional and the Los Alamos National Laboratory effective core potential double zeta basis set (LANL2DZ)¹⁵ for the iron and Ahlrichs (Karlsruhe) split valence with polarization Gaussian basis set (def2-SVP)¹⁶ for the rest of the atoms. Density fitting was also used for the initial geometry optimization to speed up the calculations.¹⁷ Then, each system was submitted to another geometry optimization with the B3LYP functional with the same basis sets. After this, another round of optimization was performed with the Ahlrichs (Karlsruhe) valence triple-zeta with polarization basis set (def2-TZVP) for all atoms. The final optimization was done with the same B3LYP functional with def2-TZVP basis set with polarizable continuum solvent (PCM)¹⁸⁻¹⁹ at dielectric constant of 4; this dielectric constant is common for most proteins and enzymes. All these calculations contained GD3BJ empirical dispersion.²⁰

Gaussian calculations were executed using the Comet supercomputer at the San Diego Supercomputer Center (SDSC), part of the Extreme Science and Engineering Discovery Environment (XSEDE).²¹ As a result, heme b595 and heme d became more planar (with a dihedral angle of -1.2° and -1.3°, respectively, formed by the central nitrogen atoms in each heme in the final optimized structures) compared to the initial cryo-EM structure (-3.0° and -2.5°, respectively), in agreement with our expectations. Heme b558 retained its planarity (dihedral angle of 0.1°) in comparison to the experimental structure (dihedral angle value of 0.0°). The coordinate data for these geometry optimized structures are provided in this publication (heme_b558.xyz, heme_b595.xyz, heme_d.xyz).

6.4.3. Derivation of new force field parameters

The force field parameters solely pertaining to heme b558, heme b595, and heme d except for the spirolactone (γ -lactone) ring part were taken from Ref.¹² and Ref.²², which report general force field parameters for heme cofactors and for heme b in cytochrome c oxidase, respectively. Heme d, unlike the other two heme cofactors, contains an unusual spirolactone ring and parameters for this part of the porphyrin ring were generated using the Antechamber package with GAFF parameters in Amber20,²³ as was done in Ref.²⁴. These force field parameters for the heme cofactors, along with parameters generated with GAFF, are listed in heme.frcmod.

The key parameters involving the metal ions were obtained with the Python-based metal center parameter builder (MCPB.py)¹³ and are listed in mcpbpy.frcmod. Since MCPB.py requires geometry optimization and force

constant calculation of the atoms involved with the metal center, we used the B3LYP functional, LANL2DZ basis set for the iron, def2-SVP basis set for the rest of the atoms, and the implicit solvent model PCM with a dielectric constant of 4.0. The input structures were taken from the final geometry optimizations described in the previous section. In the force field parameters predicted by MCPB.py, all bond force constants between the iron and its ligating atoms were smaller than 200 kcal/(mol $\times$ Å²) and the equilibrium bond distances were smaller than 2.8 Å. Most angle force constants involving the iron were smaller than 100 kcal/(mol $\times$ radian²) with the equilibrium angle values greater than 100°, except for several of them for heme b558 (force constants greater than 100 kcal/(mol $\times$ radian² or equilibrium angle values less than 100°) and a few for heme b595 and heme d (equilibrium angle values less than 100°). The potential barriers were zero for all of the dihedrals involving the iron. The Lennard-Jones radius obtained for the iron was greater than 1 Å for all three heme cofactors.

6.4.4. Point charges derivation

The atomic charges of the three heme cofactors and their axial ligands were derived from the structures obtained at the final geometry optimizations described in the previous subsection. The point charges were also computed at the same level of theory. We computed the electrostatic potential (ESP) on a molecular surface grid using the Merz-Singh-Kollman scheme²⁵⁻²⁶ in Gaussian. The ESP output was then used to generate restrained electrostatic potential (RESP) charges using Amber20's Antechamber package²³. The RESP charge calculation and fitting step was part of MCPB.py. We observed that the RESP charges of the metal ions in each heme were smaller than their oxidation state, as expected.²⁷ Since the axial ligands have charges different from standard Amber ff14SB charges, the axial ligands were incorporated as part of the appropriate heme cofactor system (HB1 for heme b558, HB2 for heme b595, HDD for heme d) and their backbones were separated out as different residues (HIO for histidine, MTO for methionine, GUO for glutamic acid) as done in Ref.²⁸. Note that the propionates in the heme cofactors are not included, i.e., they are not part of the residues HB1, HB2, and HDD, and instead are separated out as different residues (PRP for protonated propionate and PRN for deprotonated propionate). The parametrization for the residues PRP and PRN were taken from Ref.²⁸. Finally, in order to make each heme cofactor system equal to an integer charge, the infinitesimal remaining charge was distributed to all carbon atoms in each system. All of these residues are listed in heme.lib.

6.4.5. Setting up your own system

In order to set up the Amber input files of your own system with the novel heme parametrization reported in this work, users can repeat the same steps that we did for *Escherichia coli* (*E. coli*), which we summarize next. First, we prepared the PDB file, starting from the one publicly available under the ID 6RKO at the Protein Data Bank, by modelling the missing residues and arranging the heme cofactors with the axial ligands, propionates, and axial ligands' backbones (HB1, HB2, HDD, PRN, PRP, HIO, MTO, and GUO) renamed and/or redistributed accordingly to the description in the previous subsection. This PDB file is available as part of this publication: 6rko_final.pdb. Next, we run the Amber LEaP script (tleap.in), which loads the new force field parameters for the heme cofactor systems (heme.lib, heme.frcmod, mcpbpy.frcmod) and GAFF parameters for ubiquinone (ubiquinone.prepi, ubiquinone.frcmod), and generates the topology and coordinates Amber's input file. As one can see in the tleap.in file provided in this publication, it is necessary that the user explicitly specifies the bonds between the propionates and the heme cofactor systems (HB1, HB2 and HDD) and the bonds between these heme cofactor systems with the backbone atoms of the axial ligands (HIO, MTO, and GUO). The Amber LEaP script creates the structure and topology files for the dry (protein.pdb, protein.prmtop, protein.rst7) and solvated (protein_solv.pdb, protein_solv.prmtop, protein_solv.rst7) systems. In case the user would like to change the protonation state of a given propionate residue, this can be done by simply modifying the residue name of this residue in 6rko_final.pdb from PRN to PRP, or vice-versa; PRP corresponds to the protonated propionate, and PRN the deprotonated propionate.

6.5. CRediT authorship contribution statement

Surl-Hee Ahn: Conceptualization, Investigation, Data curation, Writing – original draft. **Christian Seitz:** Investigation, Data curation, Writing – review & editing. **Vinícius Wilian D. Cruzeiro:** Methodology, Validation, Data curation, Writing – review & editing. **J. Andrew McCammon:** Writing – review & editing, Supervision, Funding acquisition. **Andreas W. Götz:** Methodology, Validation, Writing – review & editing, Supervision, Funding acquisition.

6.6 Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

6.7 Acknowledgments

S.A. acknowledges support from NIH GM31749 and University of California San Diego. C.S. acknowledges support from the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1650112. A.W.G. would like to thank NIH for financial support through award R01 GM100934. We also thank Prof. Kurt Krause of the Department of Biochemistry, University of Otago, New Zealand, Özlem Demir, and Pengfei Li for valuable discussions. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) Comet at the San Diego Supercomputer Center (SDSC), which is supported by National Science Foundation grant number ACI-1548562 and through allocation ddp358 to J.A.M. and allocation TG-CHE130010 to A.W.G.

Chapter 6, in full, is a modified reprint of the material as it appears in “Ahn, S.-H.*, Seitz, C*, Cruzeiro, V.D.W., McCammon, J.A., Götz, A.W. Data for molecular dynamics simulations of Escherichia coli cytochrome bd oxidase with the Amber force field. *Data in Brief*, **2021**, 38.” The dissertation author was a co-investigator and co-author of this work.

References

1. Ahn, Surl-Hee; Seitz, Christian; Cruzeiro, Vinícius Wilian D.; McCammon, J. Andrew; Götz, Andreas W., Data for molecular dynamics simulations of Escherichia coli cytochrome bd oxidase with the Amber force field. *Data in Brief* **2021**, 38.
2. Safarian, Schara; Hahn, Alexander; Mills, Deryck J.; Radloff, Melanie; Eisinger, Martin; Nikolaev, Anton; Meier-Credo, Jakob; Melin, Frederic; Miyoshi, Hideto; Gennis, Robert B.; Sakamoto, Junshi; Langer, Julian D.; Hellwig, Petra; Kühlbrandt, Werner; Michel, Harmut, Active site rearrangement and structural divergence in prokaryotic respiratory oxidases. *Science* **2019**, 366 (6461), 100-104.
3. Schrödinger, LLC *Schrödinger release 2020-3: Maestro*, New York, NY, 2020.
4. Jo, Sunhwan; Lim, Joseph B.; Klauda, Jeffery B.; Im, Wonpil, CHARMM-GUI Membrane Builder for Mixed Bilayers and Its Application to Yeast Membranes. *Biophys J* **2009**, 97 (1), 50-58.
5. Olsson, Mats H. M.; Søndergaard, Chresten R.; Rostkowski, Michal; Jensen, Jan H., PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *Journal of Chemical Theory and Computation* **2011**, 7 (2), 525-537.
6. Jorgensen, William L.; Chandrasekhar, Jayaraman; Madura, Jeffery D., Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics* **1983**, 79 (2), 926-935.
7. Case, David A.; Aktulga, H. Metin; Belfon, Kellon; Ben-Shalom, Ido Y.; Brozell, Scott R.; Cerutti, David S.; Thomas E. Cheatham, III; Cisneros, G. Andrés; Cruzeiro, Vinícius Wilian D.; Darden, Tom A.; Duke, Robert E.; Giambasu, George; Gilson, Michael K.; Gohlke, Holger; Götz, Andreas W.; Harris, Robert; Izadi, Saeed; Izmailov, Sergei A.; Jin, Chi; Kasavajhala, Koushik; Kaymak, Mehmet C.; King, Edward; Kovalenko, Andriy; Kurtzman, Tom; Lee, Tai Sung; LeGrand, Scott; Li, Pengfei; Lin, Charles; Liu, Jian; Luchko, Tyler; Luo, Ray; Machado, Matias; Man, Viet; Manathunga, Madushanka; Merz, Kenneth M.; Miao, Yinglong; Mikhailovskii, Oleg; Monard, Gerald; Nguyen, Hai; O'Hearn, Kurt A.; Onufriev, Alexey; Pan, Feng; Pantano, Sergio; Qi, Ruxi; Rahnamoun, Ali; Roe, Daniel R.; Roitberg, Adrian; Sagui, Celeste; Schott-Verdugo, Stephan; Shen, Jana; Simmerling, Carlos L.; Skrynnikov, Nikolai R.; Smith, Jamie; Swails, Jason; Walker, Ross C.; Wang, Jinan; Wei, Haixin; Wolf, Romain M.; Wu, Xiongwu; Xue, Yi; York, Darrin M.; Zhao, Shiji; Kollman, Peter A., AMBER 2020. **2020**.
8. Maier, James A.; Martinez, Carmenza; Kasavajhala, Koushik; Wickstrom, Lauren; Hauser, Kevin E.; Simmerling, Carlos, ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation* **2015**, 11 (8), 3696-3713.
9. Joung, In Suk; III, Thomas E. Cheatham, Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *Journal of Physical Chemistry B* **2008**, 112 (30), 9020-9041.
10. Wang, Junmei; Wolf, Romain M.; Caldwell, James W.; Kollman, Peter A.; Case, David A., Development and testing of a general amber force field. *Journal of Computational Chemistry* **2004**, 25 (9), 1157-1174.
11. Dickson, Callum J.; Madej, Benjamin D.; Skjerve, Åge A.; Betz, Robin M.; Teigen, Knut; Gould, Ian R.; Walker, Ross C., Lipid14: The Amber Lipid Force Field. *Journal of Chemical Theory and Computation* **2014**, 10 (2), 865-879.
12. Capece, Luciana; Boechi, Leonardo; Perissinotti, Laura L.; Arroyo-Mañez, Pau; Bikiel, Damián E.; Smulevich, Giulietta; Marti, Marcelo A.; Estrin, Dario A., Small ligand-globin interactions: Reviewing lessons derived from computer simulation. *Biochimica et Biophysica Acta - Proteins and Proteomics* **2013**, 1834, 1722-1738.
13. Li, Pengfei; Kenneth Merz, Jr., MCPB.py: A Python Based Metal Center Parameter Builder. *Journal of Chemical Information and Modeling* **2016**, 56 (4), 599-604.
14. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams, Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16 Rev. C.01*, Wallingford, CT, 2016.

15. Hay, P. Jeffrey; Wadt, Willard R., Ab initio effective core potentials for molecular calculations. Potentials for the transition metal atoms Sc to Hg. *Journal of Chemical Physics* **1985**, 82, 270-283.
16. Weigend, Florian; Ahlrichs, Reinhart, Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics* **2005**, 7, 3297-3305.
17. Eichkorn, Karin; Treutler, Oliver; Öhm, Holger; Häser, Marco; Ahlrichs, Reinhart, Auxiliary basis sets to approximate Coulomb potentials. *Chemical Physics Letters* **1995**, 240 (4), 283-290.
18. Tomasi, Jacopo; Persico, Maurizio, Molecular Interactions in Solution: An Overview of Methods Based on Continuous Distributions of the Solvent. *Chemical Reviews* **1994**, 94 (7), 2027-2094.
19. Tomasi, Jacopo; Mennucci, Benedetta; Cammi, Roberto, Quantum Mechanical Continuum Solvation Models. *Chemical Reviews* **2005**, 105 (8), 2999-3094.
20. Grimme, Stefan; Ehrlich, Stephan; Goerigk, Lars, Effect of the damping function in dispersion corrected density functional theory. *Journal of Computational Chemistry* **2011**, 32 (7), 1456-1465.
21. Towns, John; Cockerill, Timothy; Dahan, Maytal; Foster, Ian; Gaither, Kelly; Grimshaw, Andrew; Hazlewood, Victor; Lathrop, Scott; Lifka, Dave; Peterson, Gregory D.; Roskies, Ralph; Scott, J. Ray; Wilkins-Diehr, Nancy, XSEDE: Accelerating Scientific Discovery. *Computing in Science & Engineering* **September-October 2014** 16, 62-74.
22. Yang, Longhua; Skjevik, Åge A.; Du, Wen-Ge Han; Noodleman, Louis; Walker, Ross C.; Götz, Andreas W., Data for molecular dynamics simulations of B-type cytochrome c oxidase with the Amber force field. *Data in Brief* **2016**, 8, 1209-1214.
23. Wang, Junmei; Wang, Wei; Kollman, Peter A.; Case, David A., Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling* **2006**, 25 (2), 247-260.
24. Weininger, Philipp; Weiss-Greiler, Petra; Snor, Walter; Viernstein, Helmut; Wolschann, Peter, Molecular dynamics simulations and quantum chemical calculations on b-cyclodextrin–spironolactone complex. *Journal of Inclusion Phenomena and Macrocyclic Chemistry* **2007**, 57, 29-33.
25. Singh, U. Chandra; Kollman, Peter A., An approach to computing electrostatic charges for molecules. *Journal of Computational Chemistry* **1984**, 5 (2), 129-145.
26. Besler, Brent H.; Jr., Kenneth M. Merz; Kollman, Peter A., Atomic charges derived from semiempirical methods. *Journal of Computational Chemistry* **1990**, 11 (4).
27. Henriques, João; Costa, Paulo J.; Calhorda, Maria José; Machuqueiro, Miguel, Charge Parametrization of the DvH-c3 Heme Group: Validation Using Constant-(pH,E) Molecular Dynamics Simulations. *Journal of Physical Chemistry B* **2013**, 117 (1), 70-82.
28. Cruzeiro, Vinícius Wilian D.; Amaral, Marcos S.; Roitberg, Adrian E., Redox potential replica exchange molecular dynamics at constant pH in AMBER: Implementation and validation. *Journal of Chemical Physics* **2018**, 149 (072338).