UNIVERSITY OF CALIFORNIA SAN DIEGO

Molecular Simulation to Explore Targets in Context

A dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy

in

Chemistry

by

Sarah Elizabeth Kochanek

Committee in Charge:

Professor Rommie E. Amaro, Co-Chair
Professor J. Andrew McCammon, Co-Chair
Professor Padmini Rangamani
Professor Susan Taylor
Professor Elizabeth Villa
Professor Joel Yuen-Zhou

2019

The Dissertation of Sarah Elizabeth Kochanek is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Co-Chair

_____
Co-Chair

University of California San Diego

2019

DEDICATION

For my father, Gary Kochanek.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

To my advisors, Rommie Amaro and Andy McCammon: thank you for your unwavering support in my scientific and professional pursuits. I am eternally grateful for your mentorship.

To my doctoral committee, Professors Joel Yuen-Zhou, Elizabeth Villa, Susan Taylor, and Padmini Rangamani: thank you for support, critique, and advice throughout my graduate career.

To Patti Craft and Teri Simas: thank you for your support and organization. Without the two of you my graduate school experience would not have been as smooth and seamless.

To Özlem Demir: thank you for your mentorship, from my first project through my last, your expertise has proved invaluable and helped shape me into the scientist I have become.

To my fellow lab members, especially Lorenzo Casalino and Jacob Durrant: thank you for sharing your expertise and advice, my work would not be as strong without your input.

To my friends: you have been a constant light through the highs and the lows and I am grateful to have formed such meaningful relationships.

To Benjamin Jagger: thank you for your support, both personal and professional. You have enriched the graduate school experience, and I am endlessly grateful for your partnership.

Chapter 1, in full, has been submitted for publication as it may appear in "Durrant, J. D.†, Kochanek, S. E.†, Casalino, L., Ieong, P. U. & Amaro, R. E. Mesoscale All-Atom Influenza Virus Simulations Suggest New Substrate Binding Mechanism. *ACS Central Science*". The dissertation author was a primary coinvestigator and author of this work.

Chapter 2, in full, has been submitted for publication as it may appear in "Kochanek, S. E. & Amaro, R. E. On the Structural Basis for Group 1 Influenza Inhibition by Arbidol. *Chem*". The dissertation author was the primary investigator and author of this work.

Chapter 3, in full, has been submitted for publication as it may appear in "Jagger, B. R.†, Kochanek, S. E.†, Haldar, S., Amaro, R. E. & Mulholland, A. J. Multiscale Simulation Approaches to Drug-Protein Binding. *Current Opinion in Structural Biology*". The dissertation author was a primary coinvestigator and author of this work.

VITA

2012 - 2015	Undergraduate Research Fellow Evanseck Lab, Duquesne University
2015	Bachelor of Science in Chemistry, Minor in Mathematics Duquesne University
2015 - 2019	Graduate Research Fellow Amaro and McCammon Labs, University of California San Diego
2018	Graduate Student Intern Modeling and Informatics, Merck Research Laboratories
2019	Doctor of Philosophy in Chemistry University of California San Diego

PUBLICATIONS

Durrant, J. D.†, **Kochanek, S. E.**†, Casalino, L., Ieong, P. U. & Amaro, R. E. Mesoscale All-Atom Influenza Virus Simulations Suggest New Substrate Binding Mechanism. *Submitted*

Kochanek, S. E. & Amaro, R. E. On the Structural Basis for Group 1 Influenza Inhibition by Arbidol. *Submitted*

Jagger, B. R.†, **Kochanek, S. E.**†, Haldar, S., Amaro, R. E. & Mulholland, A. J. Multiscale Simulation Approaches to Drug-Protein Binding. *Submitted*

ABSTRACT OF THE DISSERTATION

Molecular Simulation to Explore Targets in Context

by

Sarah Elizabeth Kochanek

Doctor of Philosophy in Chemistry

University of California San Diego, 2019

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

As the field of computer-aided drug discovery matures, it has become evident that viewing a molecular target independent of its biological context is often an oversimplification. To this end, our lab has developed an all-atom model of an influenza virus particle. In this dissertation, I present analysis of molecular dynamics simulations of this virus particle by Markov state modeling. I begin with the analysis of influenza neuraminidase, which includes construction of a Markov state model to characterize the catalytic site dynamics. This Markov state model provides a quantitative framework for comparison to dynamics of isolated influenza neuraminidase simulations to validate

the virus particle model. Next, I identify the probable group 1 influenza hemagglutinin binding site for the antiviral Arbidol. Here, a Markov state model is constructed from the virus particle simulations to characterize the dynamics of the proposed binding site. Finally, I conclude with a recent review outlining multiscale simulation approaches to drug-protein binding, for which Markov state modeling is included.

Chapter 1

Mesoscale All-Atom Influenza Virus Simulations Suggest New Substrate Binding Mechanism

1.1 Abstract

Influenza virus circulates in human, avian, and swine hosts, causing seasonal epidemic and occasional pandemic outbreaks. Influenza neuraminidase, a viral surface glycoprotein, has two sialic acid binding sites. The catalytic (primary) site, which also binds inhibitors such as oseltamivir carboxylate, is responsible for cleaving the sialic acid linkages that bind viral progeny to the host cell. In contrast, the functional annotation of the secondary site remains unclear. Here, we better characterize these two sites through the development of an all-atom, explicitly solvated model of the pandemic influenza A H1N1 2009 viral envelope, containing ~160 million atoms. Molecular dynamics simulations of this crowded subcellular environment, coupled with Markov state model theory, provide a novel framework for studying realistic molecular systems at the mesoscale and allow us to quantify the kinetics of the 150-loop transition between the open and

closed states. An analysis of chloride ion occupancy along the neuraminidase surface implies a potential new role for the neuraminidase secondary site, wherein the terminal sialic acid residues of the linkages may bind before transfer to the primary site where enzymatic cleavage occurs. Altogether, our work breaks new ground for molecular simulation in terms of the size, complexity, and methodological analyses of the simulated components, as well as provides fundamental insights into the understanding of substrate recognition processes for this vital influenza drug target, suggesting a new strategy for the development of anti-influenza therapeutics.

1.2 Introduction

Influenza virus infection is responsible for millions of deaths worldwide each year. The Center for Disease Control estimates that pandemic influenza A H1N1 2009 (pH1N1) virus infection affected 60.8 million people, resulting in 12,468 casualties in the United States alone[1,2]. Along with others, this strain dramatically contributes to yearly epidemics, continuously fueling the concerns about the emergence of a new pandemic strain. In addition, the increasingly widespread resistance to antiviral medications is compounding this threat[3], thus requiring the development of novel approaches for the prevention and treatment of influenza virus infection. One such strategy is to target the viral surface glycoprotein neuraminidase (NA), which promotes viral progeny release from the host cell by cleaving terminal sialic acid residues[4–6]. Previous work has identified the importance of characterizing the dynamics of the NA catalytic site for drug design[7–12], understanding mechanisms of antiviral resistance[13], and deciphering the mechanisms underlying substrate binding[14–17].

The catalytic (primary, 1°) site of NA is highly flexible, in part due to the adjacent 150- and 430-loops (residues 147–152 and 429–433, respectively, N2 numbering)[11,14,18]. The

significance of this flexibility is highlighted by structural comparison of the phylogenetically distinct group-1 (N1, N4, N5, and N8) and group-2 (N2, N3, N6, N7, and N9) NAs, which illustrates that opening of the 150-loop in the group-1 structures leads to the formation of the so-called 150-cavity[12] that can bind compounds with increased specificity and potency[10]. However, crystal structures of pH1N1 NA (pN1) reveal that, unlike all other group-1 NAs, its 150-loop is closed and therefore no 150-cavity is present[19]. In contrast, previous investigations utilizing MD simulations have found that the 150-loop of pN1 is in the open state ~60-65% of the time[18,20].

NA also contains a secondary (2°) sialic acid binding site adjacent to the catalytic site. This site was first identified as a hemadsorption site in avian-origin influenza NAs[21–25] and was not initially believed to be present in swine-origin strains due to non-conservation of critical residues at this site[23,26]. However, more recent studies provide support for the presence of a 2° site in swine-origin influenza NAs, including pN1[16,17]. The precise mechanism by which this 2° site functions remains unclear, however a number of studies have demonstrated its role in receptor binding[27–30] and catalytic efficiency[27,28]. In addition, previous Brownian dynamics (BD) simulations of single glycoproteins and various substrates suggested that both endogenous ligands and the drug oseltamivir carboxylate bind faster to the 2° site than the 1° site (*i.e.*, the k_{on} rate is 2- and 7-fold higher for the N1 and N2 2° site, respectively, vs. the corresponding 1° site)[17]. Finally, the 2° site has recently been identified as a target for a novel influenza virus inhibitor[31,32], further highlighting the need to understand its role in viral infectivity.

To study the 1° and 2° sites in the context of the viral surface, we used integrative modeling to construct a fully atomistic model of the pH1N1 viral envelope (Figure 1.1A). This was built using high-resolution crystallographic structures of individual glycoproteins (~1.9 – 2.6 Å resolution)[19,33] that were spatially positioned according to a lower-resolution cryo-electron

tomography (cryo-ET, ~16-20 Å resolution) map of a viral particle[34]. Our viral envelope construct includes 30 NA tetramers (120 monomers) and 236 hemagglutinin (HA) trimers (708 monomers) embedded in a phospholipid bilayer, with realistic density and patterning taken directly from the cryo-ET. The entire pH1N1 all-atom system modeled here amounts to ~160 million atoms (fully solvated) and is ~115 nm in diameter. As such, it is amongst the largest biophysical system yet studied with all-atom molecular dynamics (MD).

Over the past decade, studies of viruses at the molecular and coarse-grained (CG) levels have given unique insights into these systems, complementing and extending available experimental data by providing highly detailed models at never-before-seen scales, as well as suggesting testable biological hypotheses (predictions)[35,36]. Work by Schulten and coworkers established the first explicitly solvated atomic MD simulation of an intact virus, the satellite tobacco mosaic virus (~17 nm diameter, ~1 million atoms, 50 ns dynamics), in 2006[37]. Zink and Grubmuller in 2009 used steered MD to explore the dynamics of the explicitly solvated icosahedral shell of the southern bean mosaic virus (~36 nm diameter, ~4.5 million atoms, 100 ns)[38]. In 2010, Ayton and Voth developed and simulated an implicitly solvated CG representation of the immature HIV-1 virion (~125 nm diameter, 280,000 CG particles)[39]. In 2012, Larsson and coworkers simulated with explicitly solvated all-atom MD the satellite tobacco necrosis virus (~17 nm diameter, ~1.2 million atoms, ~1 us)[39], and Roberts et al. developed a fully atomic poliovirus (~30 nm diameter, ~2.8-4 million atoms, 50 ns)[40] . In 2013, Schulten and coworkers built and simulated a fully atomic representation of the HIV capsid (~70 nm diameter, ~64 million atoms, ~100 ns)[41], and Andoh et al. in 2014 simulated an all-atom poliovirus capsid (~30 nm diameter, ~6.5 million atoms, ~200 ns)[42]. Samson and colleagues in 2015 reported an explicitly solvated CG influenza virus simulation (~80 nm diameter, 5 million particles) [43]. In two separate studies in 2016, Reddy

and Samson[44], and Bond, Verma and coworkers[45], reported coarse grained simulations of the Dengue viral membrane (~50 nm diameter, ~1 million particles). Aside from the rich structural, dynamical, and biophysical insights that these studies each provided, the investigations have collectively pushed the capabilities of molecular simulation, often relying on the world's fastest and most advanced computing architectures.

The work reported here breaks new ground in molecular simulation as the first explicitly solvated atomic-scale simulation of a fully lipid-enveloped viral coat (~120 nm diameter, ~160 million particles, ~120 ns). To further characterize the structural dynamics of the viral envelope and its glycoproteins, we combined our mesoscale all-atom MD simulations with Markov state model (MSM) theory[46–48], thus enabling the extraction of long-timescale (*e.g.*, microseconds) individual glycoprotein dynamics in a crowded environment from the short timescale MD (*e.g.*, nanoseconds) of the fully intact viral surface. The accuracy and utility of MSMs have been demonstrated by experimental validation for many use cases including protein-protein binding, small-molecule binding kinetics, and protein-folding rate prediction[49–51]. Correspondingly, the approach reported here, which relies on the many copies of single glycoproteins present within a biologically accurate environment, provides a novel methodological advance for extracting long time-scale dynamics from short simulations through the powerful MSM theoretical framework at subcellular and cellular scales.

Here, we quantitatively compare the alterations in dynamics when moving from the single-protein scale simulations to a significantly more complex biological scene. By exploiting the whole pH1N1 viral envelope treated entirely with atomic resolution, this study provides unprecedented insights into the two sialic acid binding sites of NA (*e.g.*, 1° and 2°). Our mesoscale atomic simulations suggest that the NA 1° site is even more flexible than previously appreciated and

provide the first rigorous kinetic characterization of the 150-loop dynamics. Furthermore, our work suggests that the 2° site, which is more solvent exposed and, in some strains, has a higher k_{on} rate than the 1° site[17], may be responsible for initially capturing sialic acid residues, which are then electrostatically guided to the 1° site for enzymatic cleavage. Within this context, our mesoscale simulations unveil an unprecedented cooperative interplay between the two sites that further illuminates the process of sialic acid/oseltamivir carboxylate recognition and the 2° site functional annotation. This fundamental discovery may be used as a rationale for the development of novel anti-influenza small molecule therapeutics targeting NA.

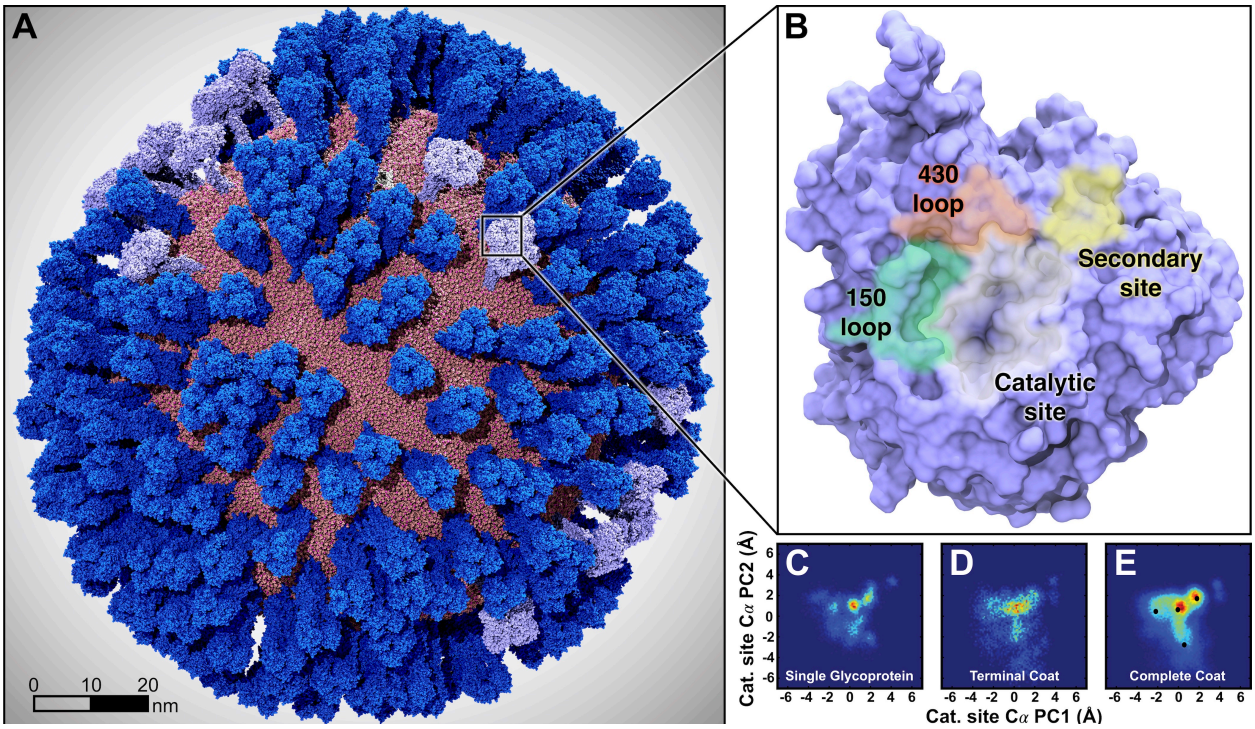

Figure 1.1 Mesoscale simulations enhance conformational sampling of the viral glycoproteins.
(A) A fully intact all-atom model of the influenza A H1N1 2009 (pH1N1) virion, containing over 160 million atoms, shown without explicit water molecules, was simulated with all-atom molecular dynamics simulations. Hemagglutinin (HA) glycoproteins shown in royal (dark) blue, neuraminidase (NA) glycoproteins shown in ice (light) blue. **(B)** Top view of a single NA monomer in surface representation with the catalytic site (white), secondary site (yellow), 150-loop (red) and 430-loop (green) highlighted. **(C-E)** Principal Component Analysis (PCA) was performed by considering the motions of the C_α atoms of 19 1°-pocket residues. PCA histograms were independently normalized so the bins containing the minimum and maximum number of points were blue and red, respectively. **(C)** PCA analysis of the four monomers sampled during a single-NA-tetramer simulation ("single glycoprotein"). **(D)** PCA analysis of the 120 monomeric trajectories extracted during the last 8.33 ns of the "terminal-coat" simulation. **(E)** PCA analysis of all 120 monomeric trajectories extracted from the "complete-coat" simulation.

1.3 Influenza Virus All-Atom Simulations

We performed 121 ns of viral coat simulation production dynamics using NAMD2[52] and CHARMM36 all-atom additive force fields[53]. This complete viral envelope ("complete-coat") simulation included 30 NA tetramers, yielding 14.5 μs of monomeric simulation (121 ns × 30 tetramers × 4 monomers/tetramer). Each glycoprotein structure used to build the initial complete-coat system was taken from a fully-equilibrated microseconds-long individual simulation (see Supporting Information, SI, for materials and methods). To broaden conformational sampling and more efficiently use supercomputer resources, we forked the complete-coat simulation twice. All simulations were run on the Blue Waters petascale supercomputer, using 114,688 processors, equivalent to 16,384 Blue Waters nodes or 4,096 physical nodes. The simulation averaged 25.57 steps/sec. Frames were written every 10,000 steps (20 ps), ultimately occupying 11.66 terabytes of disk space. Data analysis drew upon conformations extracted at equally-spaced timepoints from these trajectories. More details about model building, MD simulations and physical properties of the lipid bilayer (Figure 1.7) are reported in the SI.

To explore the flexibility of the 1° pocket, we concatenated the MD trajectories of all 120 NA monomers and calculated the principal components of 19 pocket-lining NA residues by considering their C_α atoms (heatmap, Figure 1.1C-E). We selected these 19 residues because they are homologous to those within 5 Å of the crystallographic oseltamivir carboxylate from the 2HU4 structure[12]. To judge whether mesoscale simulations enhance conformational sampling, we compared the complete-coat simulations (Figure 1.1E) to five simulations of isolated NA tetramers embedded in small lipid-bilayer patches, described in a previous work ("single glycoprotein", Figure 1.1C)[54]. Collectively, these individual simulations sampled 1.0 μs of monomeric dynamics (5 simulations × 50 ns/simulation × 1 NA tetramer × 4 monomers/tetramer), while the complete-

coat simulations sampled 14.5 μs of monomeric dynamics. To further improve comparison, we next considered only the terminal 8.33 ns of the complete-coat simulation ("terminal-coat"), which is equivalent to 1.0 μs of monomeric dynamics (1 simulation × 8.33 ns/simulation × 30 NA tetramer × 4 monomers/tetramer) (Figure 1.1D). In all the cases, the motions of the C_α atoms of the same 19 residues were projected onto the first two principal components of the complete-coat NA trajectories, and the resulting heatmaps were compared (Figure 1.1C-E). Strikingly, the principal component analysis (PCA) of the NA active-site residues indicates that the complete-coat simulation more thoroughly explored the conformational landscape, even after controlling for total simulation time.

To better study the 1° site conformations sampled by the complete-coat simulation, we applied k-means clustering to the PCA points of Figure 1.1E. Visual inspection of cluster centroids (represented by black dots in Figure 1.1E, structures in Figure 1.8) corresponding to four representative 1° site conformations reveals that R292 and R371, two key residues known to interact with the sialic acid carboxylate group, are the most flexible. In contrast, the carboxylate-stabilizing R152 residue moves outward in only one of the four conformations. Other pocket-lining residues such as R118 and D151, which previous works suggest may play a role in the molecular mechanisms of oseltamivir resistance[13], are also relatively flexible in the apo state.

The PCA and clustering analysis demonstrates that the complete-coat simulation sampled rare, distinct 1° pocket states (Figure 1.1E) not evident in the single-glycoprotein (Figure 1.1C) or terminal-coat (Figure 1.1D) simulations. This enhanced conformational sampling may simply be a product of the large number of NA copies blanketed across the viral surface; however, we do expect some effects from the viral surface environment including long-range electrostatic forces and glycoprotein-glycoprotein interactions that only the complete-coat simulation can capture.

1.4 pN1 Catalytic Site Dynamics

To explore the dynamics of the catalytic site, we analyzed the volume of the 1° and contiguous adjacent pockets over the course of the 120 monomeric NA simulations (Figure 1.2)[55]. We compared the volume of each frame to the volume of our starting crystal structure pN1 (PDB ID: 3NSS[19] with a closed 150-cavity, as well as with a structure of a non-pandemic N1 (PDB ID: 2HTY[12] with an open 150-cavity. From this analysis we find that the volume of the 1° site and adjacent regions ranges from 0 to 4440 $Å^3$, with an average of 1536 $Å^3$. By comparison, crystal structures of N1 with and without an open 150-cavity have volumes of 800 and 1088 $Å^3$, respectively (PDB IDs 3NSS and 2HTY). This indicates that the volume and depth of the catalytic and adjacent sites are remarkably increased over what has been observed in crystal structures (Figure 1.2C). Contributing to this additional cavity volume and depth are two novel subpockets near residues G351 and E227, buried deep inside – but contiguous with – the 1° site (Figure 1.2B). The ability of these new sites to bind small molecules was assessed with FTMap, a server for mapping binding hot spots in macromolecules[56], which suggested that the G351 sub-pocket can accommodate small-molecule ligands. Similar to the 150-cavity and 430-cavity, the G351- and E227-adjacent subpockets may provide new ligand-binding opportunities.

The volumetric and dynamical properties of the 1° site revealed in our simulations suggest that NA can accommodate many structurally distinct and complex sialoglycan receptors as part of the host-cell recognition process. Indeed, human glycans are vastly diverse in both their sugar composition and configuration (*e.g.*, long, short, biantennary, triantennary, etc)[57]. As such, transient deepening and broadening of the 1° NA site may allow the glycoprotein to accommodate bulkier (*e.g.*, long, bi/triantennary) and longer glycan receptors. Given that our simulations model the entire virion coat, it may be that full pocket opening only occurs in a crowded viral-surface

environment. Alternatively, our simulations may capture full pocket opening because the viral coat includes many replicates of individual glycoproteins, enabling extensive conformational sampling. Using the intramolecular distance between the 150- and 430-loops as a metric for 150-cavity formation, we constructed a two-state MSM to estimate the timescales of 150-loop opening and closing motions in the complete-coat simulations. Ultimately, we find the stationary distribution (equilibrium probabilities) of the open and closed states to be similar (0.53 and 0.47, respectively). Correspondingly, the time to transition between the two states (*i.e.*, the mean first-passage time, MFPT) is also roughly equal (39 ± 15 ns for open to closed and 29 ± 11 ns from closed to open), indicating that loop opening and closing occur at similar rates (Figure 1.3). MSM calculations are detailed in the SI and Figure 1.9, Figure 1.10, Figure 1.11, and Figure 1.12.

To understand the impact of sampling the loop dynamics in a crowded viral environment, we also constructed an MSM using structures extracted from simulations of isolated NAs embedded in planar bilayers patches[54] using the same protocol. Again, we found that the populations of the open and closed states are highly similar (0.61 and 0.39, respectively), though less than in the complete-coat simulations, and that the MFPTs between the states are overlapping with those from the complete-coat model (open to closed and closed to open after 50 ± 96 ns and 72 ± 44 ns, respectively). The 150-loop dynamics of both the complete-coat and single glycoprotein simulations are nearly equivalent, though the error associated with the MSM model constructed from the viral-envelope simulations is much smaller, likely due to the increased simulation time. This comparison indicates that 150-cavity dynamics are not influenced by the crowded environment of the viral envelope, an expected result given that this pocket faces inwards, towards the neighboring three monomers within the single tetramer, rather than being oriented outwards.

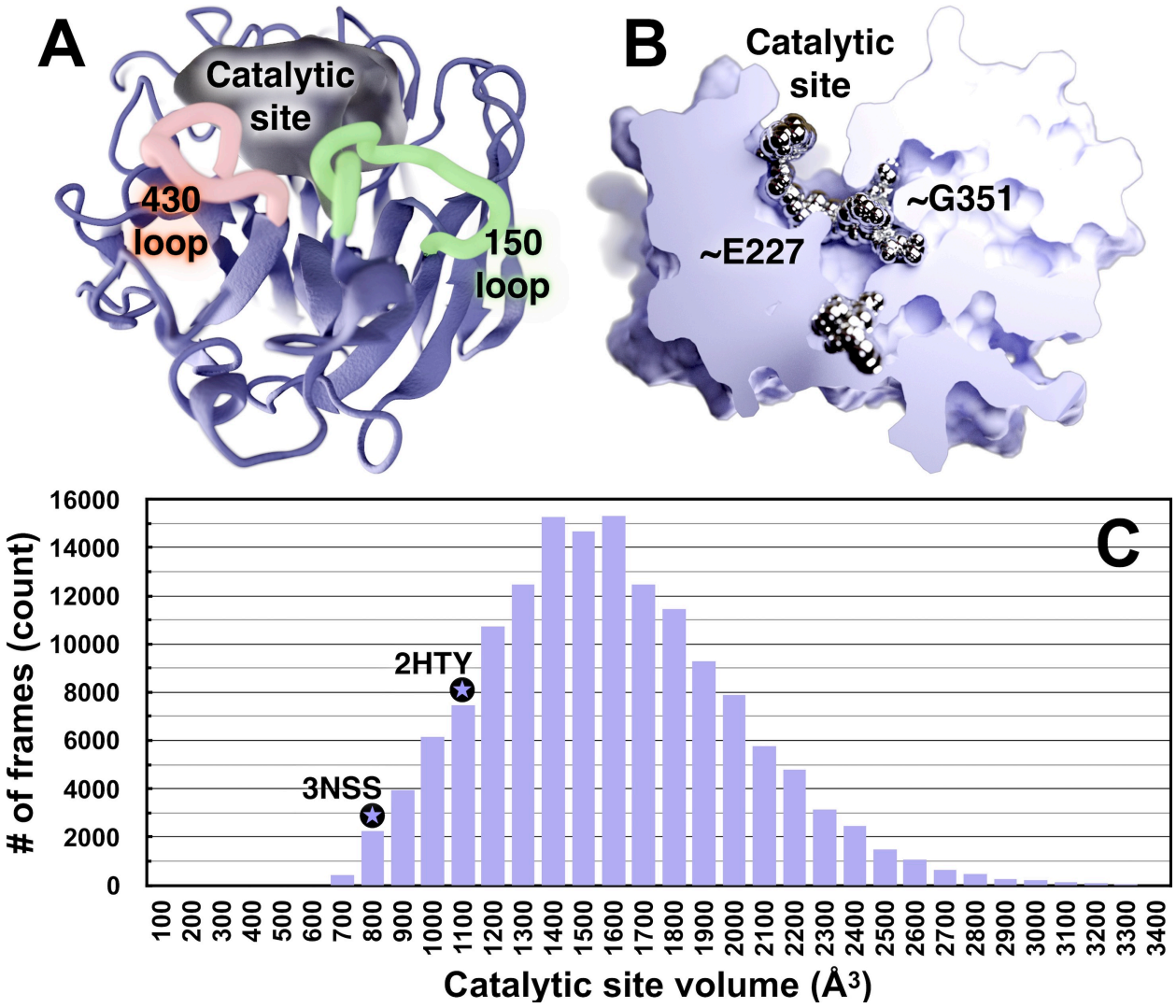

Figure 1.2 Volumetric and ligand binding "hot spot" analyses of the 1° active site.
(A) NA is shown in blue ribbon, and the pocket volume is filled with semi-transparent gel. The 1° catalytic site, 430-loop, and 150-loop are visible. **(B)** NA is shown as solid, and ligand-binding hotspots are metallic. A portion of the surface-rendered protein was removed to facilitate visualization of internal cavities. This NA conformation has a notably open G351 pocket, which has a high propensity to bind ligands (metallic spheres). **(C)** Histogram of the NA catalytic-site volumes sampled during the MD simulations. As reference, the volumes of the same active-site cavity from two crystal structures are indicated with black-circled stars. X-ray structure 3NSS[19] (pH1N1 with a closed 150-cavity) has a volume of 800 Å^3; x-ray structure 2HTY[12] (H5N1 with an open 150-cavity) has a pocket volume of 1088 Å^3. The simulated-pocket volumes range from 0 to 4440 Å^3; the average pocket volume is 1536 Å^3.

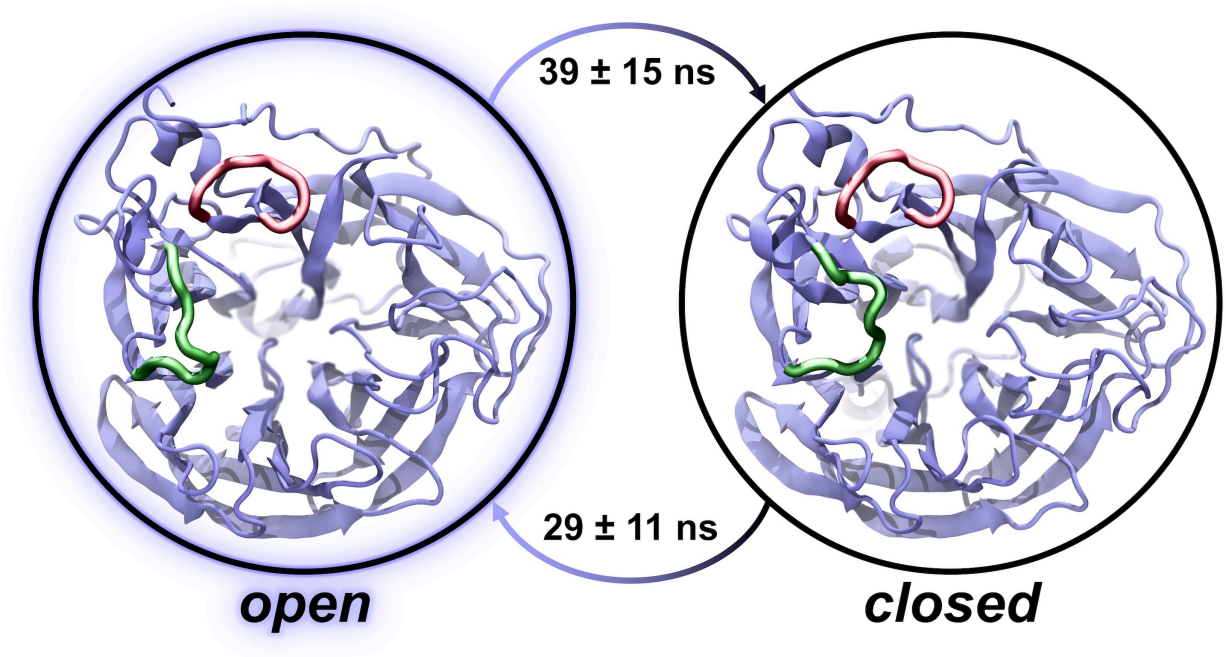

Figure 1.3 Two-state MSM with representative structures from the complete-coat simulation.
The equilibrium populations of the open and closed states are approximately equal in both the complete-coat and isolated-NA simulations. Correspondingly, the mean first-passage times between the states are approximately equal. 150-loop and 430-loop are represented with green and red thicker ribbons, respectively.

1.5 Secondary binding site: functional annotation

We note that all FDA-approved NA inhibitors, as well as the endogenous ligand sialic acid, contain negatively charged carboxylate groups. Considering the hypothesis that the 2° site contributes to catalytic efficiency by recruiting and keeping substrates within close proximity to the catalytic site[27,28], and given that prior BD simulations indicate that substrates bind faster to the 2° site than the 1° site[17], we postulate that sialic acid first binds to the more solvent exposed 2° site. Subsequently, the electrostatics of the NA surface guides the substrates to the 1° enzymatic site.

Although sialic acid substrates were not included in the complete-coat simulation, we propose that the negatively charged chloride anions in the bulk solvent surrounding the NA monomers serve as a rough surrogate for negatively charged ligand moieties that may associate with the glycoprotein surface. To identify regions favorable to chloride occupancy, we concatenated the 120 monomeric NA simulations and aligned by the alpha-carbons of the 1° site. The chloride atoms were binned into 3,375,000 voxels (0.67 Å x 0.67 Å x 0.67 Å each). We focused on voxels containing chloride counts greater than three standard deviations above the mean. Strikingly, our simulations reveal that a volume of high chloride occupancy connects the 1° and 2° sialic acid binding sites (Figure 1.4). This path is wide enough to allow negatively charged small molecules such as sialic acid or oseltamivir carboxylate to move from the 2° site to the 1° active site. Additional regions of high chloride density are depicted in Figure 1.13. This result provides evidence that the two sites may act cooperatively, supporting the work of Lai et al.[16], which confirmed that pN1 has a 2° site that can bind sialic acid.

This discovery suggests a biophysical mechanism for the previously uncharacterized 2° site. Sialic acid receptors may first bind the 2° site before being transferred to the 1° sialidase site

(Figure 1.5). We propose that our chloride distribution analysis is well suited for studying these possible mechanisms of molecular transfer. Unlike a simple electrostatic map, our analysis accounts for both electrostatic and steric factors, as well as for the conformational dynamics sampled over all 14.5 μs of monomeric simulation in the context of the whole-virus environment. In addition, this proposed "bind and transfer" mechanism is in good agreement with prior experimental results and proposed mechanisms[27].

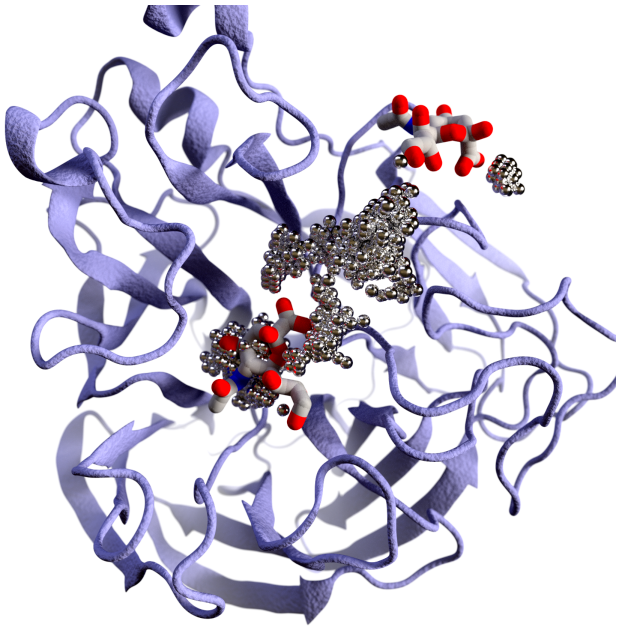

Figure 1.4 Chlorine anion distribution within the NA binding sites.
The chlorine anion distribution shows the pathway between the 1° and 2° sites. NA is drawn in iceblue cartoon. Selected regions of high chloride occupancy are illustrated as dotted silver bubbles. Additional regions are shown in Figure S2. Two sialic acids (PDB ID 1MWE) are superimposed in the catalytic (center) and 2° (upper right) sites for reference[23].

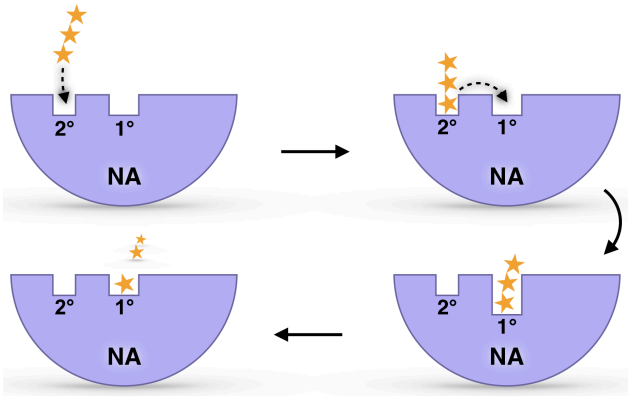

Figure 1.5 The predicted sialic acid binding mechanism involving the NA secondary site.
Yellow stars represent a sialic-acid-containing glycan substrate. Blue half circles represent NA. The 1° active site and 2° site are labeled 1° and 2°, respectively.

1.6 Conclusions

Our work suggests a novel NA binding mechanism wherein a sialic acid containing substrate (*e.g.*, a glycan receptor) first binds the 2° site, as predicted by earlier BD simulations[17]. After binding, the substrate is transferred to the catalytic site via electrostatic interactions. Finally, the catalytic site cleaves the terminal sialic acid substrate. In other words, a budding viral particle might use the 2° site to first attract the sialic acid-tipped receptors before these are cleaved within the catalytic-site, ultimately allowing viral escape from the infected host-cell surface. Moreover, MSM and volumetric analyses have served to further expand the functional annotation of the 1° site and surrounding regions, disclosing exceptional deepening and broadening dynamical properties of the catalytic pocket. Our findings might be exploited to design novel multi-pronged inhibitors capable of reaching these NA additional cavities unveiled in our multiscale simulations. Taken together, this information provides fundamental insights into the understanding of sialic acid/oseltamivir carboxylate recognition, suggesting new strategies for the development of NA inhibitors.

1.7 Acknowledgements

Chapter 1, in full, has been submitted for publication as it may appear in "Durrant, J. D.[†], Kochanek, S. E.[†], Casalino, L., Ieong, P. U. & Amaro, R. E. Mesoscale All-Atom Influenza Virus Simulations Suggest New Substrate Binding Mechanism. *ACS Central Science*". The dissertation author was a primary coinvestigator and author of this work.

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, and the Center for Research Computing at the University of Pittsburgh for a computing allocation to JDD. The authors also thank Dr. Robert Malmstrom for providing sample MSM-building scripts and useful discussions.

1.8 Supplementary Information

1.8.1 Preliminary single-glycoprotein neuraminidase molecular dynamics simulations

To generate glycoprotein structures for use in the larger virion-coat simulations, we first performed microsecond simulations of single NA and hemagglutinin (HA) glycoproteins embedded in small lipid-bilayer patches. The single-glycoprotein NA simulations are described in ref. [54], but the HA simulations have never been published. The remainder of this section describes these simulations in detail.

We created a homology model of the influenza A HA extra-virion domain with Schrödinger's Prime module, using the sequence of the Shandong 2009 H1N1 strain (accession number F2YI86) and the 2WR0[58] structure as a template. Protonation states were assigned using PDB2PQR[59] with PROPKA[60–63] at pH 7.0. The homology model was minimized and geometry-optimized using Schrödinger's Maestro suite.

To model the HA transmembrane domain, we searched the Protein Data Bank[64] for analogous extended trimeric alpha-helical bundles. The 2WPQ structure[65] was identified as a good candidate. Modeled alpha helices with the appropriate HA sequence were then aligned to the helices of the 2WPQ structure. Inter-virion domains, likely unstructured, were added to the modeled transmembrane domain. The HA extra-virion domain with appropriate disulfide bonds was positioned relative to the transmembrane and inter-virion domains using VMD[66]. This

complete HA structure was then embedded in a lipid bilayer, as described in ref. [54]. To resolve steric clashes, any lipid or membrane component within 3.0 Å of the protein was removed using PyMolecule, a beta version of the now-published Scoria Python package[67] for easily manipulating 3D molecular data. More details on the system building of the single glycoproteins can be found in ref. [54].

1.8.2 Building the virion-coat model from a simplified cryoelectron-tomography "point model"

Experimental collaborators provided us with a simple point model of the influenza exterior derived from electron tomography [12]. Surface points represented the lipid-covered virion surface. The NA and HA "spikes" were represented as lines that protruded from the viral surface.

Positioning the glycoprotein models.

Five distinct conformations of HA and NA, respectively, were extracted from the single-glycoprotein (GP) microsecond simulations using RMSD clustering[68], as implemented in the GROMACS computer package[69]. See ref. [54] for details. These GP conformations were programmatically positioned at the appropriate locations on the cryoelectron-tomography-derived point model using PyMolecule[67,70].

The point model suggested that some GPs were in very close proximity, making steric clashes unavoidable. To resolve these clashes, a multi-step process was used. First, GPs that were sufficiently distant from their neighbors were fixed. Using a Monte-Carlo method, the remaining GPs were allowed to randomly "jump" to adjacent regions on the virion surface. With every random jump, the distances between neighboring GPs were again evaluated, and any GP sufficiently distant from its new neighbors was also fixed. Additionally, with each jump, each GP

had a 10% chance of being returned to its original location, assuming that location was not occupied by another GP. This ensured that GPs did not wander too far from their initial, experimentally determined locations.

The ultimate goal of fixing every GP was, unfortunately, not achievable using the above method alone. Consequently, during the course of the Monte Carlo procedure, the distance cutoffs were gradually relaxed. This gradual relaxing kept the number of potential steric clashes to a minimum but did introduce the possibility that some GPs would clash.

Following the Monte Carlo procedure, we performed a pairwise distance comparison between the heavy atoms of all neighboring GPs to correct these occasional clashes. Two GPs were said to clash if the distance between any of their heavy atoms was less than 1.0 A. Those GPs that clashed with the greatest number of their neighbors were deleted first, followed by the less egregious offenders. Ultimately, it was only necessary to eliminate 10-15% of all the GPs to resolve all GP-GP steric clashes.

Inserting M2 channels.

For completeness sake, we randomly positioned eleven identical models of the M2 transmembrane domain, derived from the 2L0J structure, at viral-surface regions that were not occupied by GPs. The published M2 structures available at the time of model construction did not allow us to create a full-sequence M2 homology model. Subsequent examination of the model (after extensive simulation) also revealed that the channels had the incorrect orientation. Fortunately, these channels are sparsely distributed. They are also too distant from the GP ectodomains to impact NA and HA sialic-acid binding sites on nanosecond timescales.

Generating the lipid bilayer.

LipidWrapper[71], a PyMolecule-based program designed for creating large-scale lipid-bilayer models of arbitrary geometry, was used to carpet the entire asymmetrical surface of the virion point model with phosphatidylcholine (POPC) molecules. LipidWrapper used a large planar lipid-bilayer model as input, generated using the CHARMM-GUI server[72,73]. Lipid residues that came within 3 Å of any protein were deleted.

1.8.3 Complete virion-coat all-atom simulations

As mentioned above, the NA and HA glycoproteins used to populate the viral surface were themselves taken from microsecond-long MD simulations and so were extensively pre-equilibrated. In contrast, equilibrating the atomistic, mesoscale bilayer proved challenging. By subjecting the viral-coat model to iterative rounds of minimization, simulation, and hole patching, we ultimately obtained a bilayer that was fairly stable and equilibrated. The viral-surface model did not include the virion interior. To maintain the overall geometry of the model in the absence of structure-supporting interior components, we also fixed the positions of every 10^{th} inner-leaflet lipid head group.

All simulations were run on 114,688 processors, equivalent to 16,384 Blue Waters nodes or 4,096 physical nodes using NAMD[52] and CHARMM36 all-atom additive force fields[53]. Over the course of the simulations, NAMD performed 252 benchmark calculations, which averaged 25.57 steps/sec (51.14 fs of simulation per sec of real time). Frames were written to the simulation trajectories every 10,000 steps (20 ps), ultimately occupying 11.66 terabytes of disk space.

After 40.14 ns, the productive simulation was forked into two daughter simulations that continued for 30.60 and 15.16 ns, respectively. The second daughter simulation was then also forked into simulations that lasted 20.70 and 14.44 ns, respectively. In one of the instances, a single

hole developed in the virion bilayer after roughly 20 ns. This hole comprised roughly 0.2% of the entire virion surface area and so had a minimal effect on complete-coat dynamics. It was repaired prior to launching the 14.44-ns simulation. Taken together, these simulations constitute 121.04 ns total. Data analysis drew upon equally spaced conformations extracted from these trajectories.

1.8.4 Lipid Analysis

Phospholipid curvature values were calculated by extracting the coordinates of the phosphorous atoms of the headgroups and fitting a sphere to surrounding lipids within a 200 Å radius (Figure 1.7). The inverse radius of the fitted sphere gives the curvature value. The sphere fit was calculated using the least squares algorithm available from NumPy [74]. The area per lipid was estimated at each lipid coordinate by dividing the number of lipids within the 200 Å radius by the surface area of the spherical cap as calculated from the radius of curvature.

1.8.5 Volumetric and FTMap analysis

After aligning all 120 NA monomeric trajectories by the active-site C_α, we used POVME to measure the pocket shapes and volumes across all trajectories (grid spacing 2.0) [55]. The frame with the largest G351-adjacent pocket was fed into FTMap to predict druggable hotspots[56]. We performed the same FTMap analysis with an NA crystal structure (PDB ID: 2HU4)[12].

1.8.6 Markov state models

MSMs of the 150-cavity were constructed from both the complete-coat simulations and the isolated NA simulations with trajectory frames taken every 0.02 ns and 0.05 ns, respectively. The minimum distance between I149 and P431 was selected as the input feature for model construction.

Trajectory frames were clustered into 300 microstates by k-means clustering, as implemented in the PyEMMA software package[75]. A lag time of 10 ns was selected for model construction, and the resulting models were validated by the Chapman-Kolmogorov (CK) test (Figure 1.9 and Figure 1.10). PCCA++ in conjunction with a hidden Markov state model (HMM) (lag time of 2 ns) was used for coarse-graining the MSMs into two metastable states, with validation by the CK test (Figure 1.11 and Figure 1.12). Confidence intervals were calculated using Bayesian hidden Markov state models (BHMMs) corresponding to the described HMMs.

1.8.7 Chloride anion analysis

We concatenated all 1°-site-aligned NA monomers, together with nearby chloride ions, into one trajectory. A box with dimensions 100 Å x 100 Å x 100 Å, centered on the 1° sialic-acid binding site, was divided into 3,375,000 voxels (0.67 Å x 0.67 Å x 0.67 Å each). We then tallied the number of chloride anions falling within each voxel, regardless of the associated monomer or frame. To identify regions that favor chloride occupancy, we applied a high-pass filter, retaining only voxels with chloride counts greater than three standard deviations above the mean (Figure 1.13).

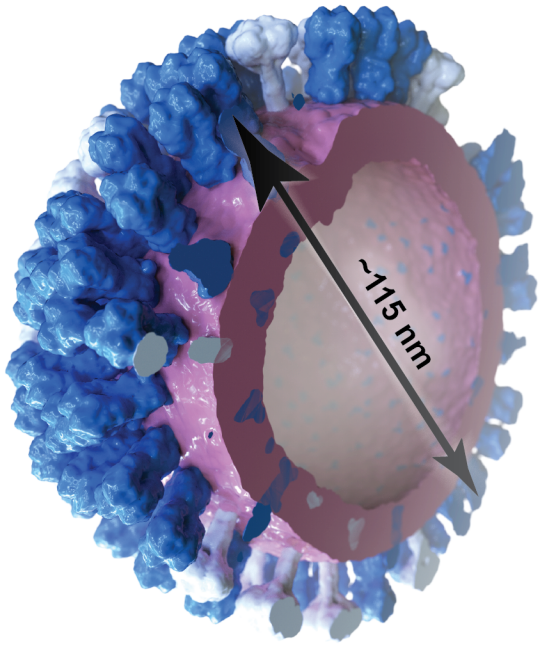

Figure 1.6 Cross section of the influenza viral envelope model.
The phospholipid bilayer is colored in pink, and the two glycoproteins hemagglutinin and neuraminidase are colored in blue and light gray, respectively.

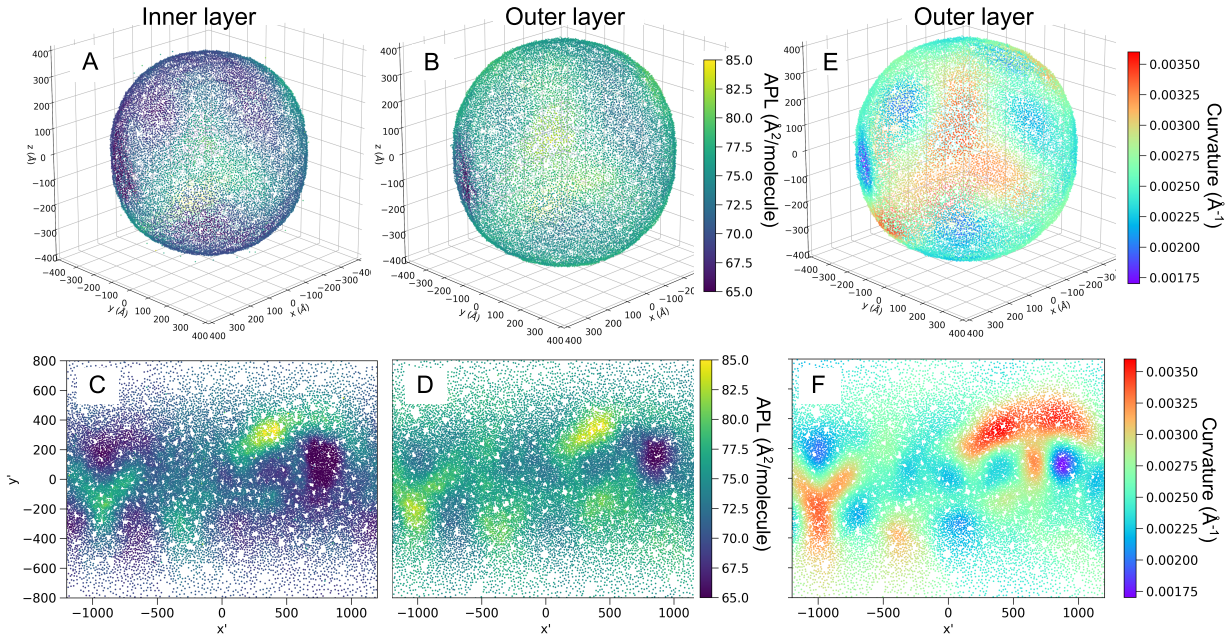

Figure 1.7 3D and Mercator projection plots of area per lipid and lipid curvature values.
Plots of area per lipid (A-D) or curvature (E-F) values associated with the phospholipids at time t = 70ns. The phospholipid headgroups are represented via 3D plot (A-B) or Mercator projection (C-D). The left and right columns plot the lipids from the inner or outer layer of the phospholipid bilayer, respectively. E-F) Curvature values associated with the outer layer of the bilayer are plotted via 3D plot or Mercator projection, respectively.

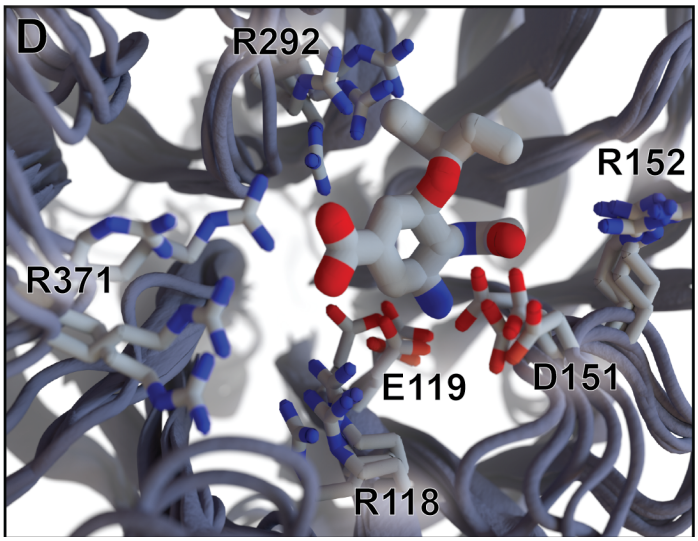

Figure 1.8 An illustration of four representative 1°-pocket conformations taken from the complete-coat simulation.
Key residues are shown as licorice. An oseltamivir molecule taken from 2HU4 has been superimposed on the structure for reference but was not included in the simulations.

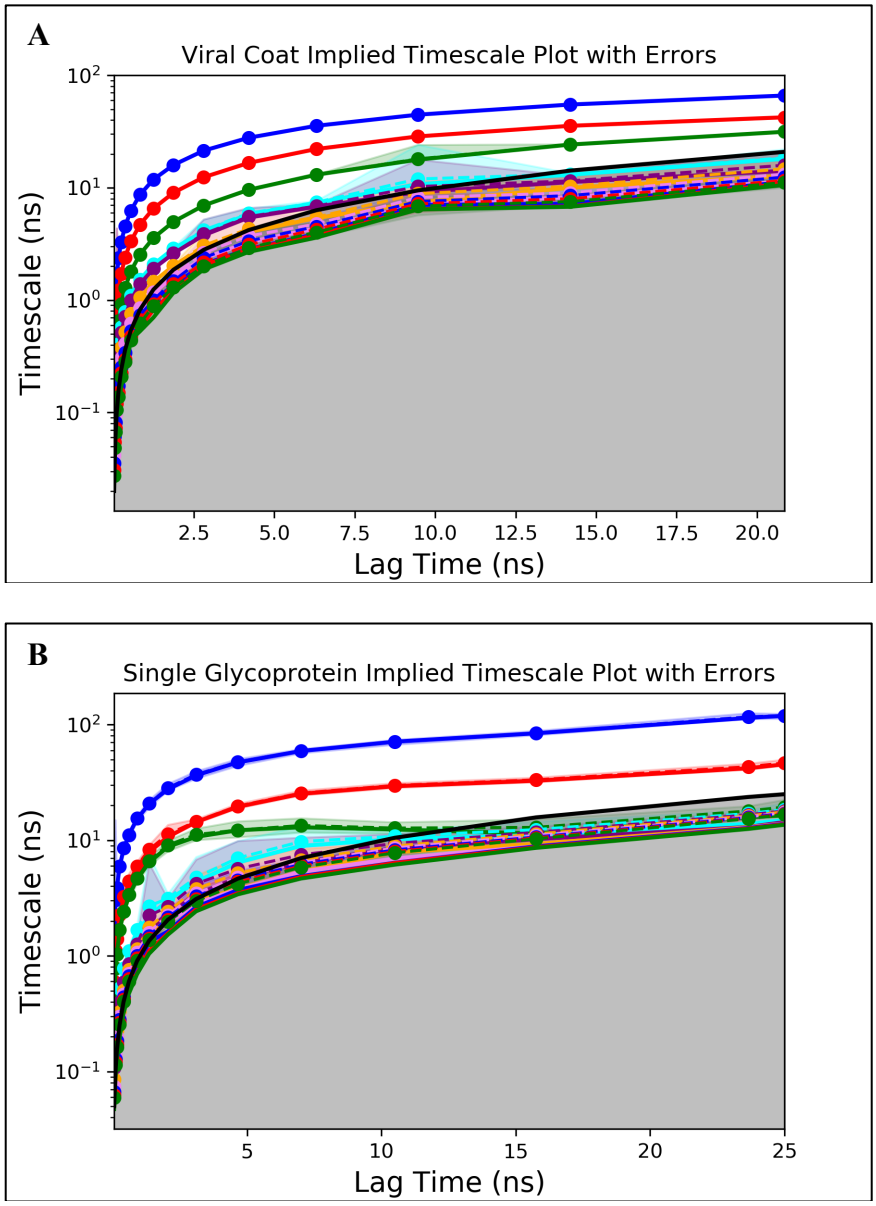

Figure 1.9 Implied time scale plots with errors.
Calculated for MSMs from the (A) viral coat and (B) single glycoprotein simulations.

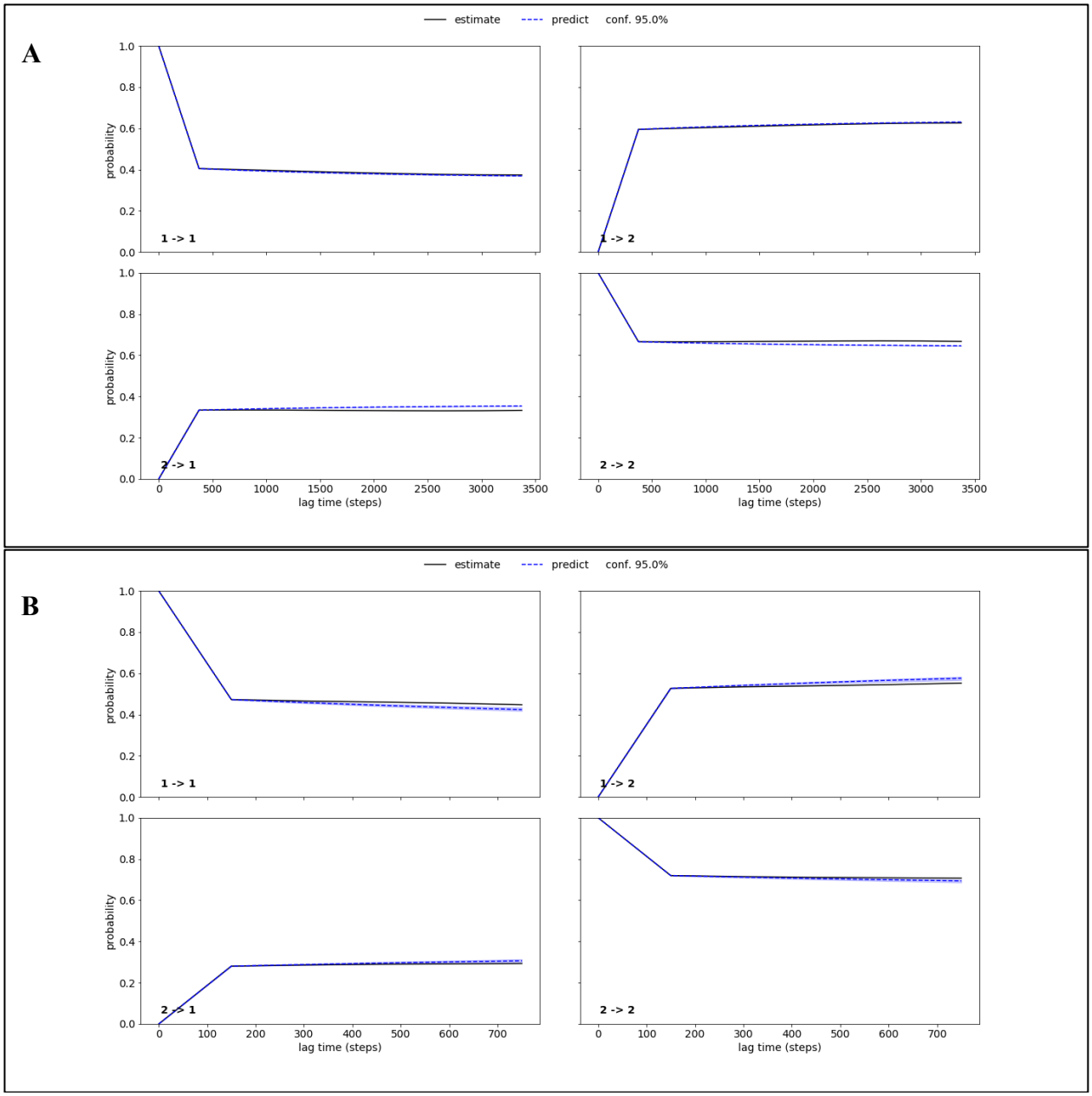

Figure 1.10 Chapman-Kolmogorov test.
Calculated for (**A**) viral coat and (**B**) isolated NA MSMs.

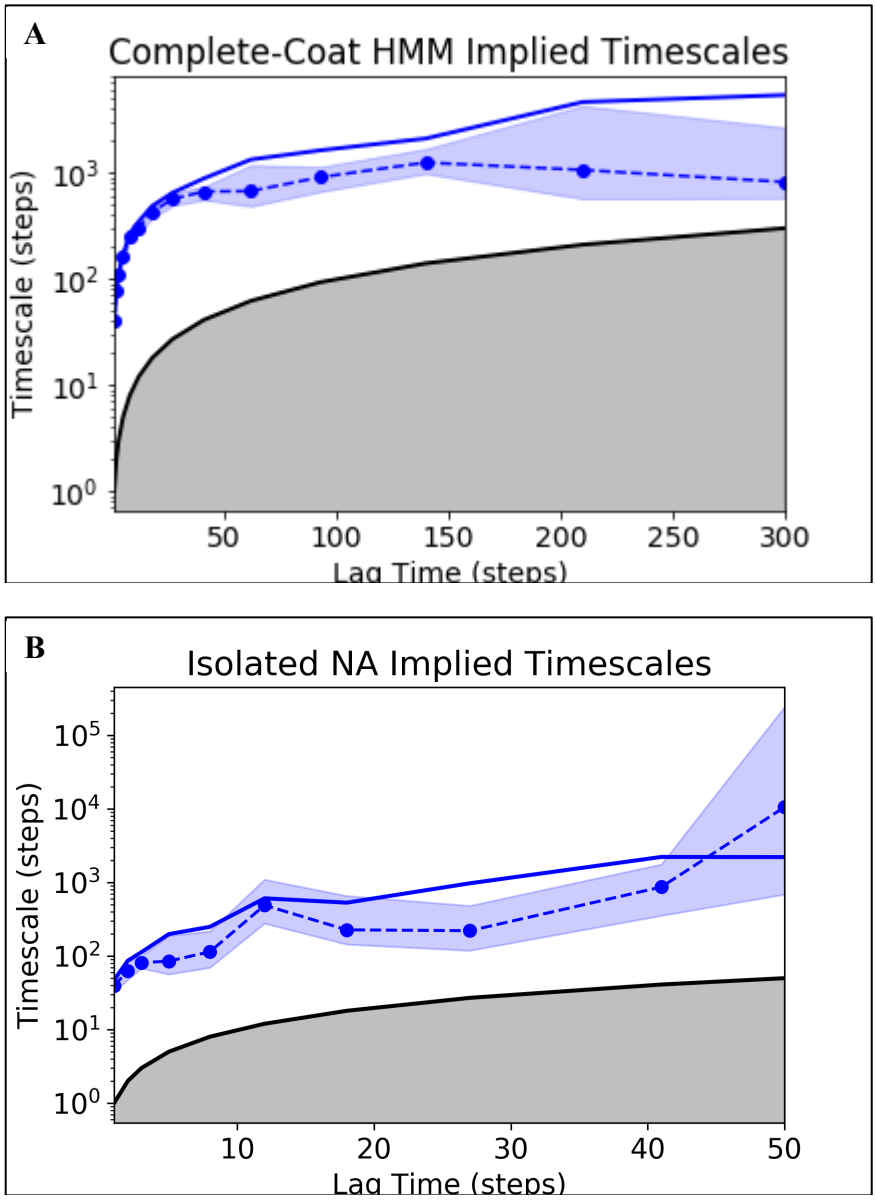

Figure 1.11 Implied time scale plots with errors.
Calculated for the HMMs from the (**A**) viral coat and (**B**) single glycoprotein simulations.

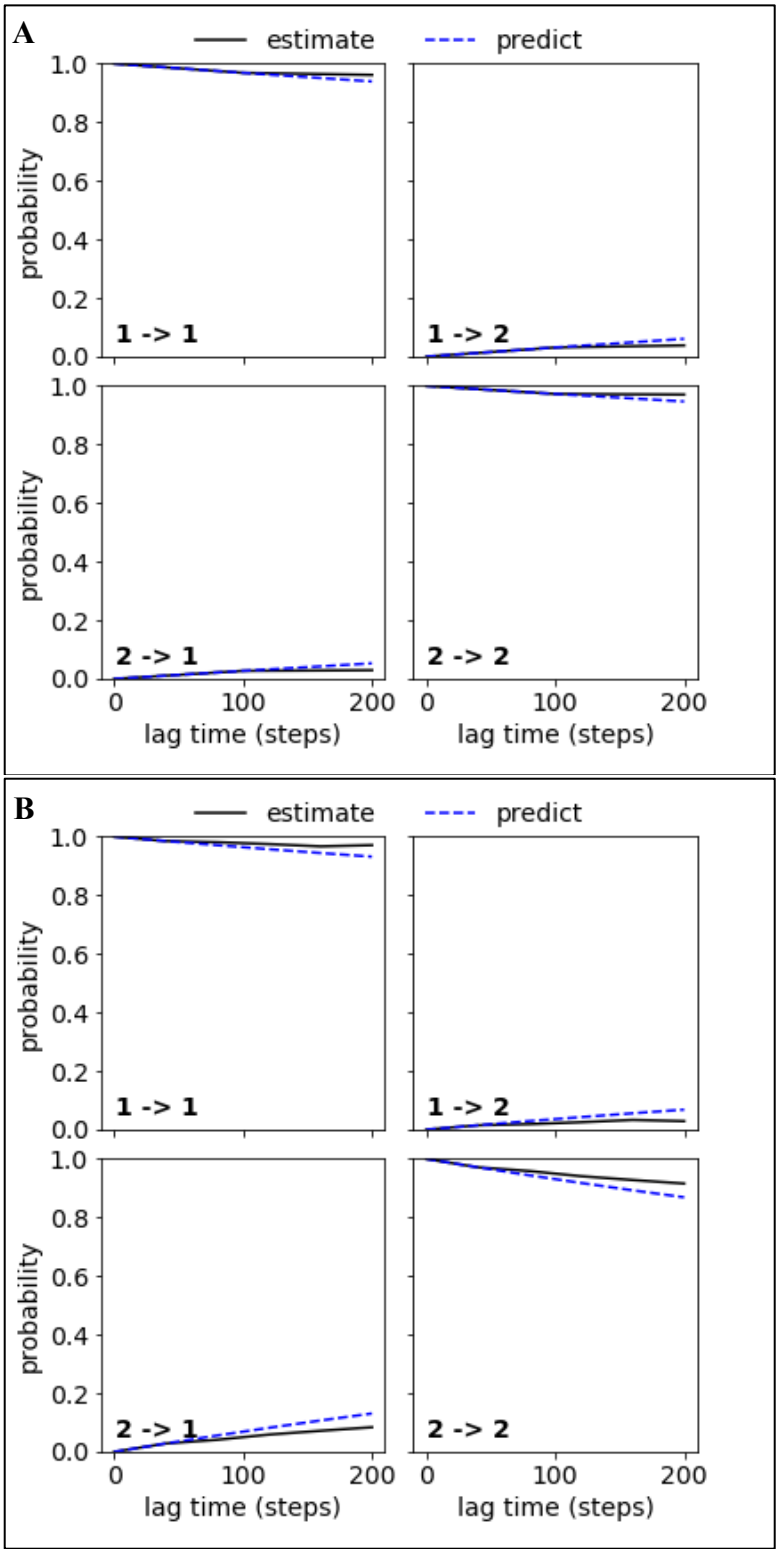

Figure 1.12 Chapman-Kolmogorov test.
Calculated for (A) viral coat and (B) isolated NA HMMs.

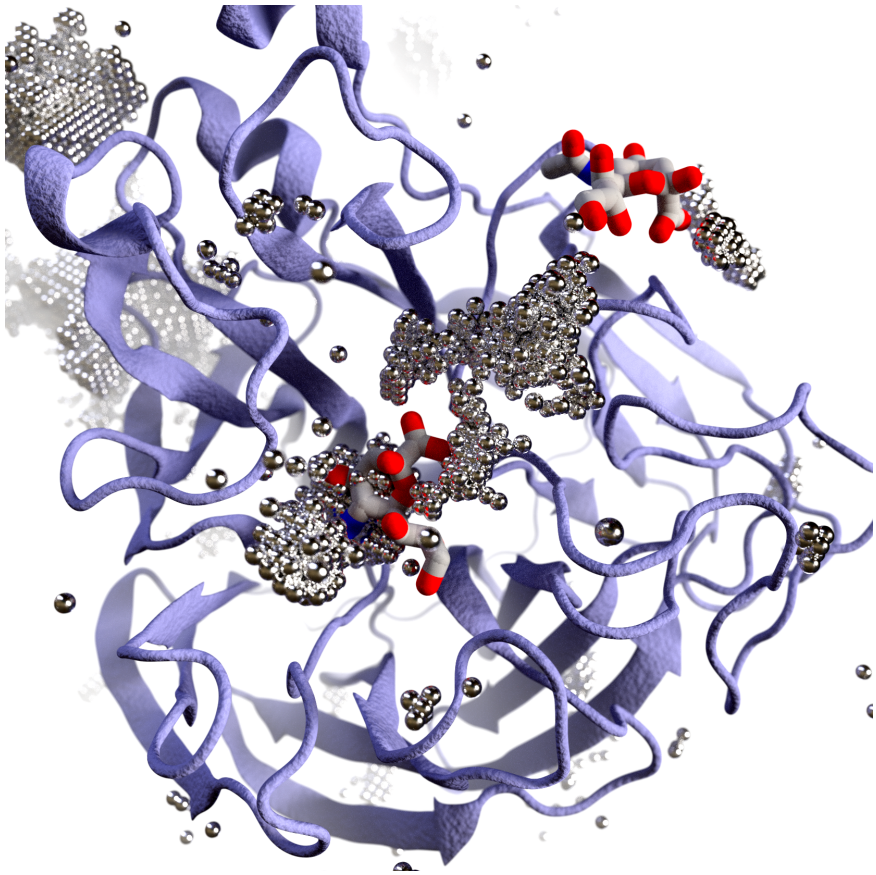

Figure 1.13 Chlorine anion distribution.
Regions of high-chloride occupancy beyond those depicted in Figure 4 are included. NA is drawn in iceblue cartoon. Regions of high chloride occupancy are illustrated as dotted black bubbles. Two sialic acids are superimposed in the 1° and 2° sites for reference (PDB ID 1MWE)[23].

Chapter 2

On the Structural Basis of Group 1 Influenza Fusion Inhibition by Arbidol

2.1 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 subtypes. While the Arbidol binding site has been determined for group 2 hemagglutinins, it remains unidentified for group 1 hemagglutinins. Further, based on differences between group 1 and group 2 hemagglutinins, it is unknown whether or not Arbidol binds in a group-specific manner. Integrative modeling techniques combining cryo-electron tomography with x-ray crystallography, homology modeling, and protein-protein docking have led to the construction of an all-atom model of the 2009 H1N1 influenza viral coat. Using the petascale computing machine Blue Waters, we have performed molecular dynamics simulations of this model. From these simulations, individual hemagglutinin trajectories have been extracted and

analyzed with the construction of a Markov state model. From our Markov state model, we present evidence of a druggable pocket that is able to accommodate Arbidol. Further, 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 hemagglutinin is not group specific. Structural insights obtained here will facilitate the development of the next generation of influenza therapeutics.

2.2 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 year [76]. 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 annually [77]. 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 inhibitors [3]. 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 IVI [78,79].

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 recognition element, sialic acid, 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 replication [80]. 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 machinery [81]. 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 HAs [82]. 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 protomers [83]. 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 2.1A). 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 2.1B). 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.

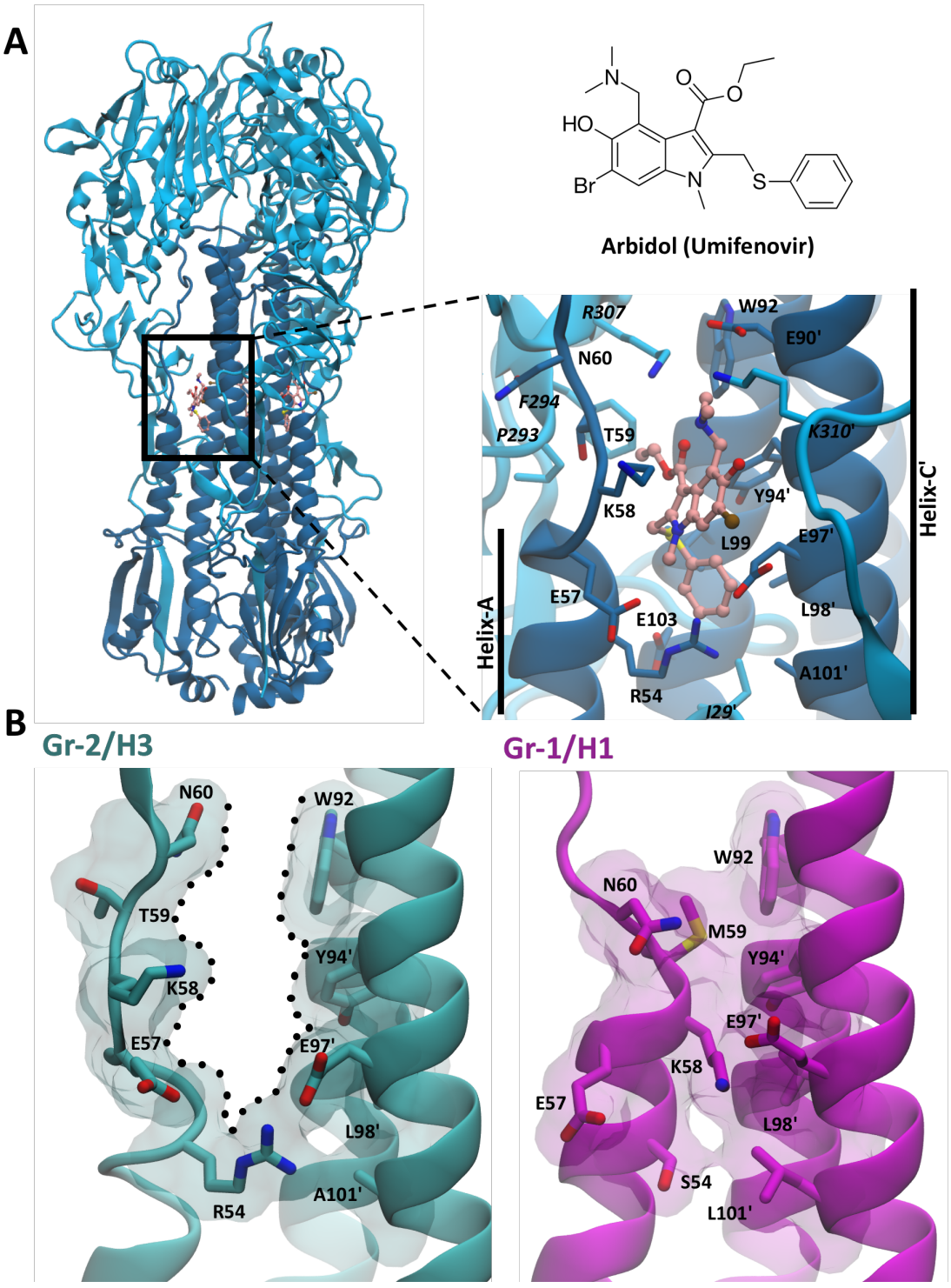

Figure 2.1 Arbidol binding site.
(A) Arbidol binds in a hydrophobic cavity between two adjacent protomers of HA. Residues with side chains that face towards the binding site are shown as sticks, Arbidol is represented in ball and stick. Residues from HA1 (light blue) are italicized, residues from HA2 (dark blue) are written in standard font. (B) Comparison of apo H3 (teal, PDB ID 4FNK [84]) and H1 (magenta, PDB ID 1RU7 [85]) 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.

2.3 Results

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, homology modeling, and protein-protein docking of the membrane-bound glycoproteins with cryo-electron tomography (Figure 2.2) [86]. The systems are some of the largest simulated to date, each ~100 nm in diameter with ~160 million atoms. The 2009 H1N1 model was selected to be studied in greater detail and was simulated for ~120 ns using all-atom molecular dynamics (MD) simulations. From this simulation, 234 individual HA trajectories have been extracted resulting in a total of ~28 μs of HA simulation. In this study, we present the first structural and kinetic basis for the probable group 1 HA Arbidol binding site, identified and characterized by the construction of a cell-scale, ensemble based Markov state model (MSM) [87]. MSMs are a powerful tool enabling the study of long time-scale dynamics from many shorter, independent trajectories [48,88], 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 'mesocape' [89]. This approach then enables the theoretically rigorous extraction of long timescale dynamics of particular cellular components in the context of biological complexity; thus, the effects of protein crowding and correlated effects of neighboring components are taken into account.

From the MSM, four major metastable states with exchange times on the order of microseconds are identified, differing primarily in the conformation of residues 57-60 in helix-A and the presence or absence of the K58-E97' salt bridge (Figure 2.3). These four states (representative structures shown in Figure 2.3) 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 (Figure 2.3). 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 ~10 μs and from open to closed is ~4 μs. 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.

In the open structure, 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 (Figure 2.4). 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 Å, whereas the C-alpha RMSD of loop-B (residues 56-60) is 4.4 Å. 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 Glide [90]. 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.

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. Computational solvent mapping with FTMap was used to identify putative ligand binding 'hotspots' on the open H1 structure [91]. 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 and extended pockets as predominantly nonpolar and charged, respectively. 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.

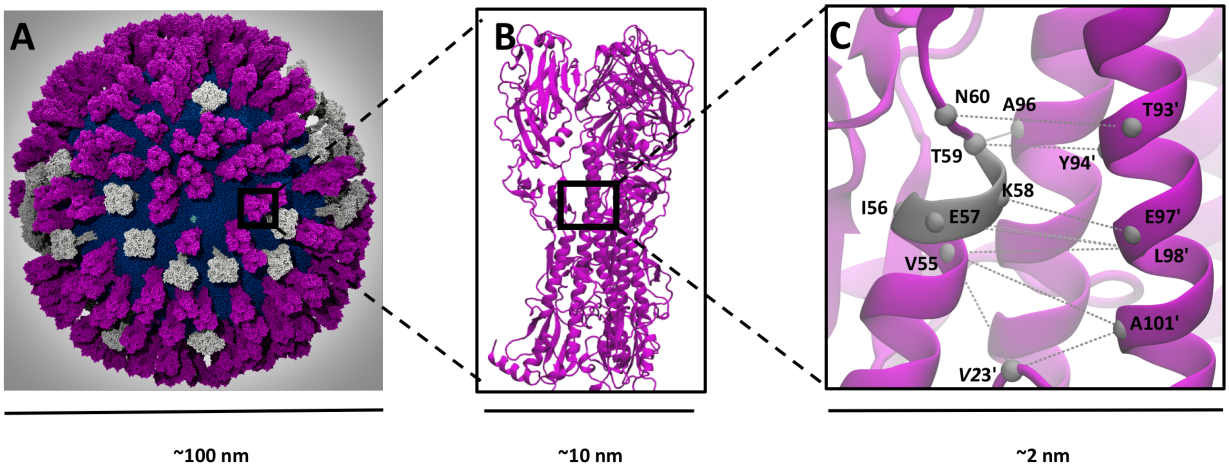

Figure 2.2 Individual HA trajectories are extracted from the viral coat simulation and are then featurized for analysis by the construction of an MSM.
(A) All-atom model of the H1N1 viral coat with HA in magenta, NA in white, and M2 ion channels in light blue. (B) 234 HA trimer trajectories are extracted from the viral coat simulation, resulting in a total of ~28 μs 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.

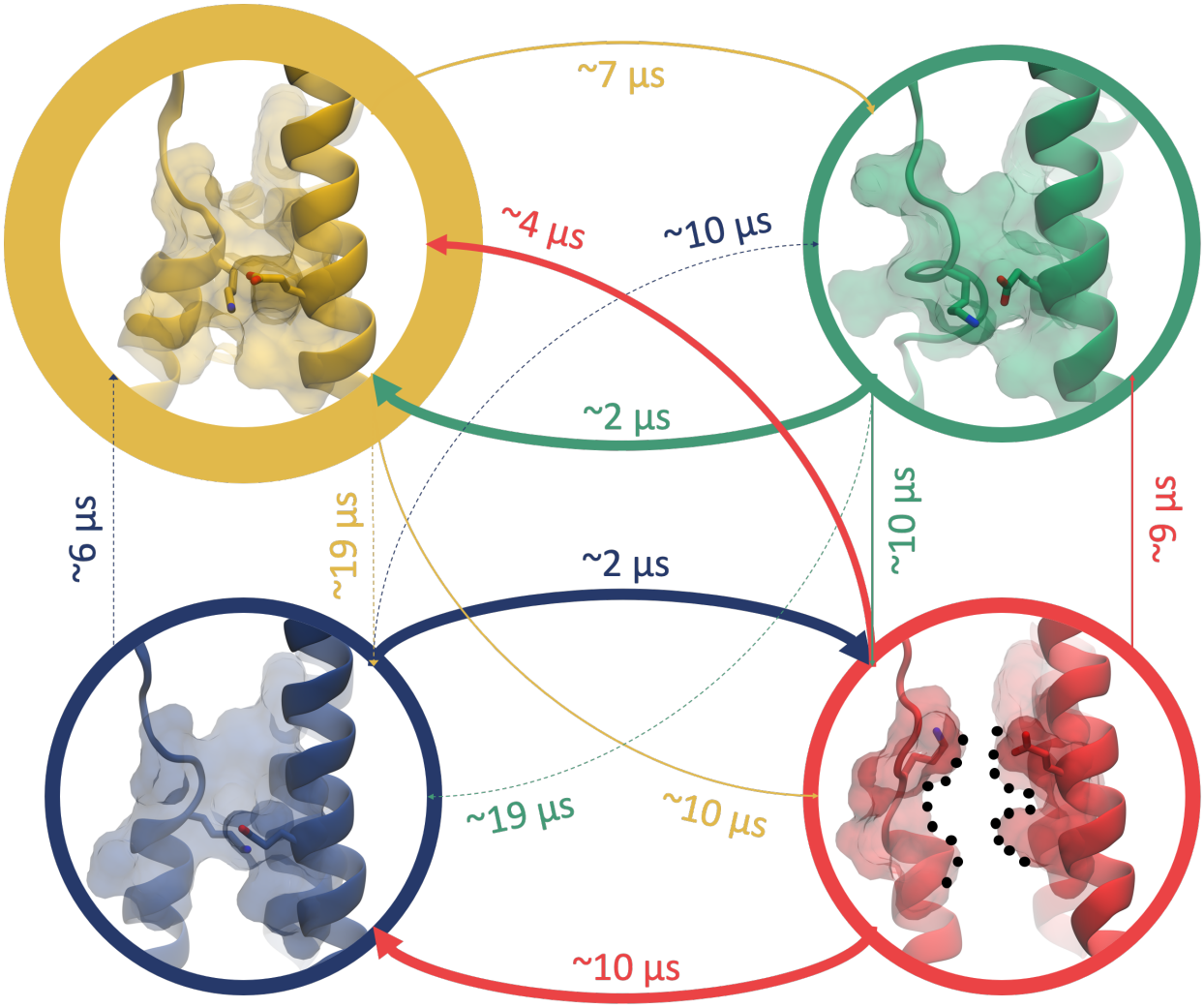

Figure 2.3 Representative structures of the four major macrostates with mean first-passage times shown. Macrostates are colored as follows: crystallographic closed (yellow), irregular (green), semi-open (blue), open (red). The width of the colored outline of each state is proportional to its population 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.)

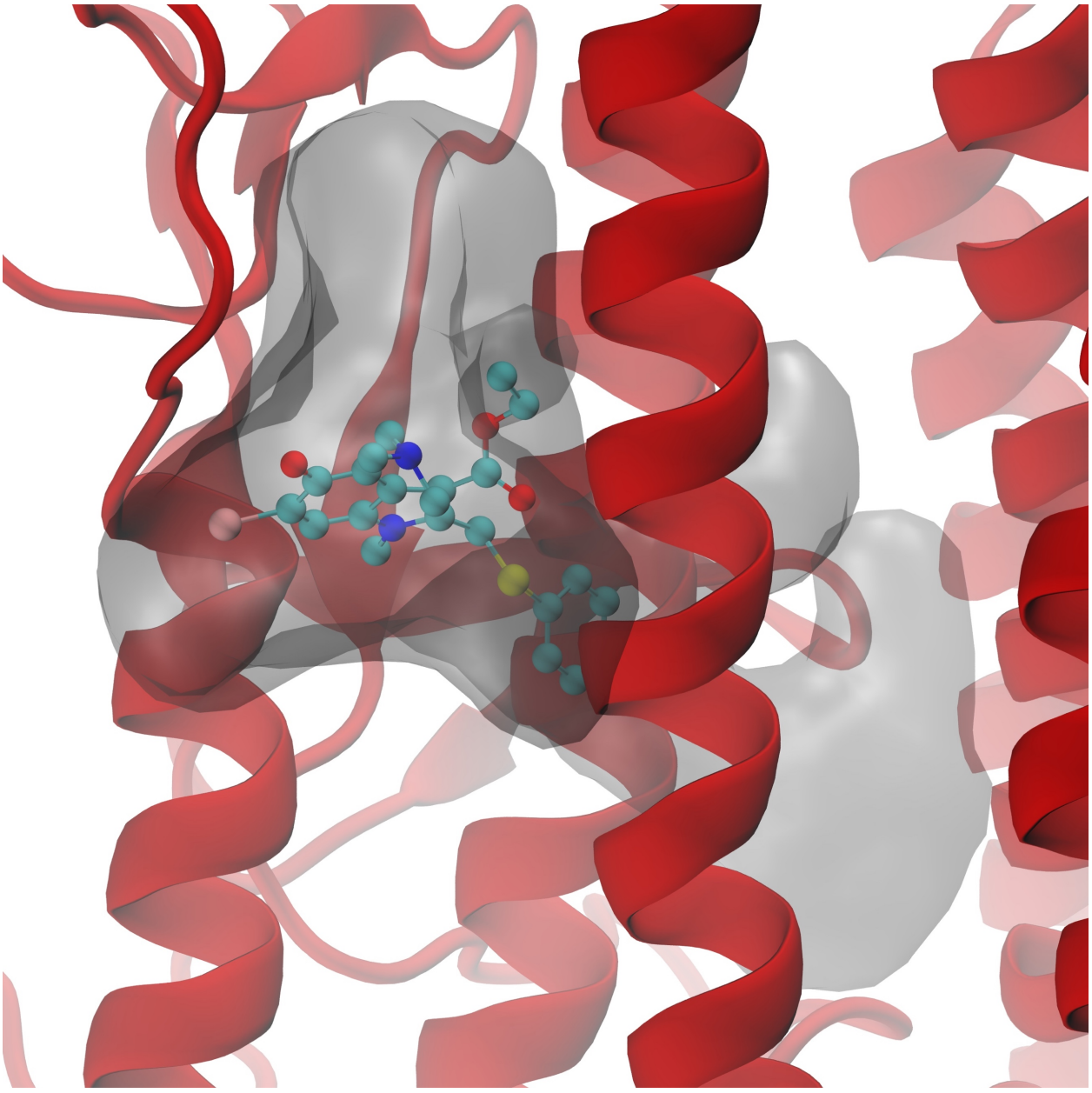

Figure 2.4 View of the primary pocket and pocket extension.
A, C, and C" helices shown in red cartoon; residues constituting the primary and secondary pockets shown in red and pink surface, respectively; docked Arbidol shown in cyan carbon ball and stick. In contrast to the primary binding site, which is largely nonpolar, the secondary site is constituted of mainly charged residues.

2.4 Discussion

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 an MSM. By this analysis we identify the probable group 1 Arbidol binding site at the same position as the group 2 binding site. Further, these two pockets not only share the same location, but also are consistent in terms of polarity, shape, and size. Rigid receptor docking demonstrates the capacity for Arbidol to be accomodated 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 make way for the development of more specific and potent antiviral compounds.

2.5 Methods

2.5.1 System preparation and simulation

The influenza viral coat model was prepared and simulated according to the procedure outlined in ref. [86].

2.5.2 Markov state models

An MSM of the Arbidol site was 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, 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. Dimensionality reduction was performed using time-lagged independent component analysis (TICA) [92] and trajectory frames were clustered into 300 microstates by k-

means clustering in TICA space, as implemented in the PyEMMA software package [75]. A lag time of 30 ns was selected for model construction, and the resulting models were validated by the Chapman-Kolmogorov (CK) test (Figure 2.5 and Figure 2.6). PCCA++ in conjunction with a hidden Markov state model (HMM) was used for coarse-graining the MSM into metastable states. 6 metastable states are identified. Two of these states, both of which represent closed pocket conformations, have total combined population of less than 1% and are excluded from further analysis.

2.6 Acknowledgements

Chapter 2, in full, has been submitted for publication as it may appear in "Kochanek, S. E. & Amaro, R. E. On the Structural Basis for Group 1 Influenza Inhibition by Arbidol. *Chem*". The dissertation author was the primary investigator and author of this work.

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. S.E.K. also acknowledges support from the NIH Molecular Biophysics Training Program (T32-GM008326).

2.7 Supplementary Information

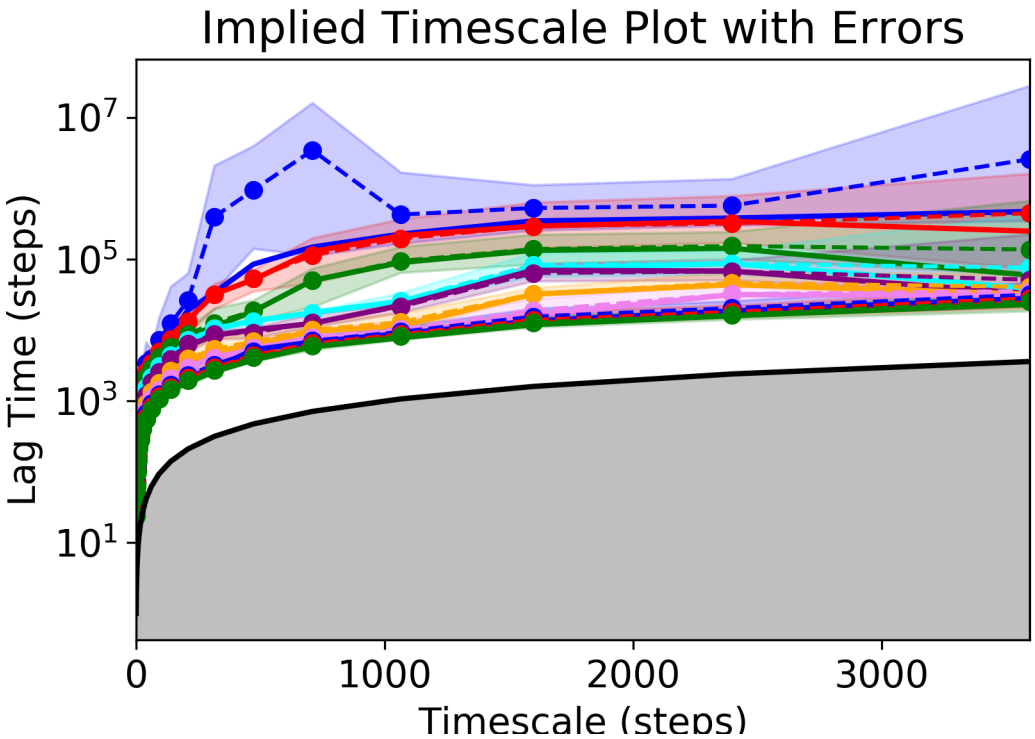

Figure 2.5 Implied timescale plot for HA MSM.

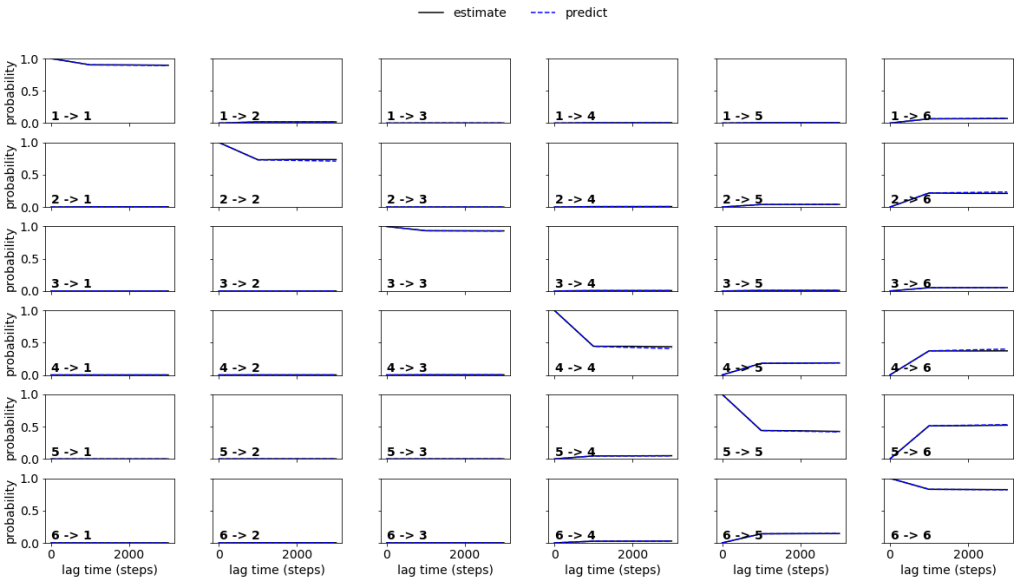

Figure 2.6 Chapman-Kolmogorov test for HA MSM.

Chapter 3

Multiscale Simulations Approaches to Model Drug-Protein Binding

3.1 Abstract

Simulations can provide detailed insight into molecular processes involved in drug action, such as protein-ligand binding, and so contribute to drug design and development. A range of different processes with a large range of length and timescales, may be involved, and understanding these different scales typically requires different types of simulation method. Ideally, simulations should be able to connect across scales, to analyse and predict how changes at one scale affect or play out on another. Multiscale simulation methods, which combine different levels of treatment, are an emerging frontier with great potential in this area. Here we review multiscale frameworks of various types, and some applications to biomolecular systems.

3.2 Introduction

Protein-ligand interactions are integral to coordinating the complex functions of cellular activity. Such interactions include the binding of signaling molecules, enzyme substrates, toxins,

regulating factors, or other proteins to the protein of interest [93–99]. Of particular interest for pharmaceutical development is the binding of drug molecules that mimic, inhibit, or modulate native protein-ligand interactions for therapeutic effect. Molecular simulations are increasingly involved in drug discovery pipelines in understanding protein-drug binding interactions, and also have the potential to reduce cost and time of drug discovery associated with synthesizing and experimentally testing many compounds [100,101]. Simulations can be used as screens during the hit identification phase and also provide insight for lead optimization, as well as to analyze drug resistance [102,103]. Developments in computer architecture, such as GPUs, and the promise of exascale computing power are transforming the range and scope of biomolecular simulations [104]. Simulations can reveal molecular mechanisms and analyze them in a level of detail and dynamic resolution beyond the reach of experiment. A particular focus is in understanding and predicting drug binding and kinetics [105,106]. Increasingly it is clear that the biological activity of many drugs depends on the rates of association or dissociation from their targets, rather than their binding affinity [93–96,107].

Simulations face conflicting challenges in this arena, with a tension between the need to address long timescales for some relevant processes (large-scale conformational changes, macromolecular association, and beyond to changes in organelles and cells) and the requirement for accurate description of molecular interactions and reactions for reliable predictions. Methods exists to simulate biomolecules at different length scales, ranging from quantum mechanical electronic structure calculations to atomistic, coarse-grained, and continuum models, that have well-established domains of applicability and can provide useful predictions of biologically relevant properties when applied with simulation techniques to sample underlying dynamics [108]. There is great potential impact to be realized through the combination of these different types of

methods, e.g. in predicting the higher-level effects of changes at the molecular level by connecting across the scales. There is further promise in leveraging and connecting to experimental, genetic, and other biological data.

Here, we review emerging multiscale methods for studying protein-ligand binding, relevant to drug design and development for small-molecules therapeutics. Multiscale techniques bridge spatial and/or temporal scales, coupling together two or more different types of modelling approaches, with varying degrees of the 'tightness' of coupling (Figure 3.1). Two or more different levels of representation may coexist within a single simulation, a switch between independent levels can be triggered when a critical configuration or milestone is reached , or sampling at a lower, computationally cheaper level can enhance and/or be informed by calculations at a higher level, with information being directly passed between different types of simulation. Here, these different simulation methods are roughly ordered in decreasing 'coarseness'. This coarseness is also closely correlated with the cost of such a simulation for a given system and timescale.

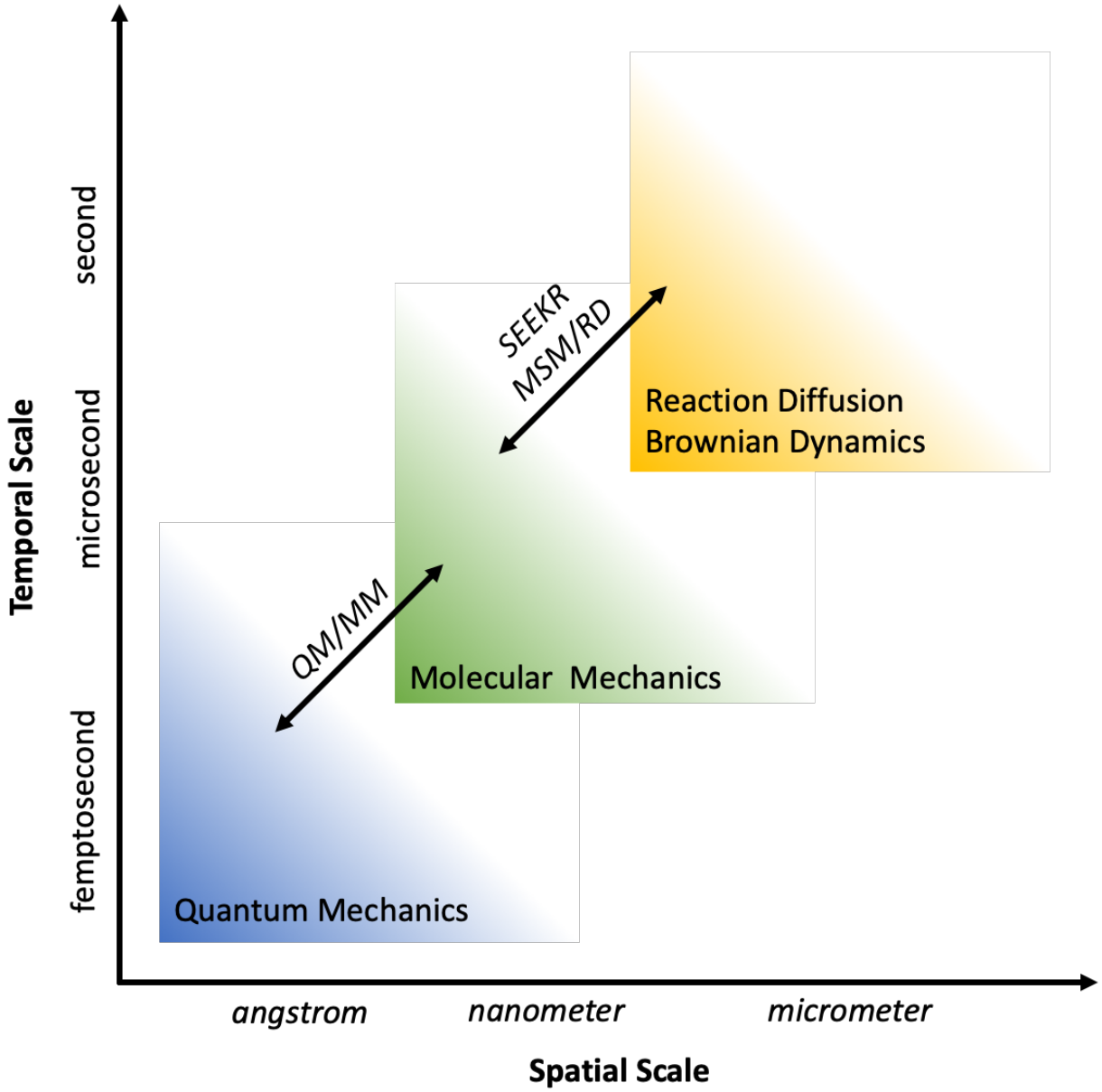

Figure 3.1 Depiction of the approximate spatial and temporal scales of various simulation approaches as well as the multiscale methods used to bridge the gaps between individual methods.
Molecular Mechanics methods underly most biomolecular dynamics simulations and can thus use atomistic representations or coarse grained models for larger systems and longer time scales.

3.3 Main

3.3.1 Combining molecular dynamics and Brownian dynamics for binding

Drug-protein binding typically occurs in two main steps: first, non-specific and long-range electrostatic interactions driving initial association and location of the binding site, followed by specific short-range interactions at the binding site (where specificity of interactions is key). This immediately suggests that the whole binding and recognition process can be simulated by a combination of different methods to simulate these two aspects, and indeed this has inspired multiscale approaches. The diffusion-limited rate constant of two molecules approximated as spheres in aqueous solution is on the order of 10^9-10^{10} $M^{-1}s^{-1}$ [109,110]. However, as a result of molecular and hydrodynamic interactions, crowding, geometric constraints of binding sites, and gating effects, the observed rate constants of ligand binding occur over a much broader range, 10^3 to 10^{10} $M^{-1}s^{-1}$. Brownian dynamics (BD) simulations solve the Langevin equation in the overdamped limit, and often achieve decreased computational cost by neglecting internal degrees of freedom and describing solvent implicitly with a dielectric constant [111]. As such, BD simulations are well-suited for studying the long-range interactions that dictate ligand association, particularly the electrostatic steering involved in forming the initial ligand-protein encounter complex. BD simulations were combined with all-atom molecular dynamics (MD) simulations in one of the earliest multiscale approaches to protein-ligand binding [112]. Chang *et al*. presented a multiscale approach to model binding pathways of ligands to HIV-1 protease that involves initial coarse-grained BD simulations followed by all-atom MD simulations initiated from snapshots of the BD trajectories [113]. The low cost of the coarse-grained BD simulations allows for extensive sampling of multiple association pathways. The pathway data generated from these simulations serves as

the starting point for multiple follow up studies, providing detailed descriptions of the drug binding process [114,115].

3.3.2 Milestoning/ SEEKR for calculations of association and dissociation rates k_{on} and k_{off}

Milestoning is a technique used to subdivide, simulate, and statistically reconnect regions of simulation space called milestones [116–120]. This approach is highly amenable to a multiscale simulation framework because simulations at each milestone are fully independent, as well as agnostic to the simulation type used, because only transition probabilities and transition times between any pair of milestones are required to compute the quantities of interest (e.g., binding or unbinding rates). There are multiple examples using milestoning with MD simulations to study ligand binding [121,122]. In particular, Simulation Enabled Estimation of Kinetic Rates (SEEKR) is a multiscale milestoning simulation technique that directly combines MD and BD simulations to calculate k_{on}, k_{off}, and the free energy of protein-ligand binding, with a focus on small molecule drugs [123–125]. Fully atomistic MD simulations are used in regions close to the binding site, where molecular flexibility and atomistic detail are essential, and rigid-body BD simulations are used in the regions farther away from the binding site, where molecular flexibility is less important, affording a dramatic savings in computational cost (Figure 3.2). SEEKR reduces the compute time required for calculations via an enhancement in sampling of rare events due to statistical bootstrapping, and it is 'embarrassingly' parallel, as each independent milestone can be simulated concurrently. SEEKR has been shown to effectively rank-order ligands by both k_{off} and binding free energy for the biosynthetic receptor system, β-cyclodextrin [125]. SEEKR has also been employed to calculate k_{on}, k_{off}, and the binding free energy for the well-studied model system trypsin with the noncovalent binder, benzamidine [124]. In addition, the binding and unbinding rates

calculated with this method are well-suited to be used as parameters for larger-scale phenomenological and diffusion-based models, furthering the multiscale nature of the SEEKR methodology.

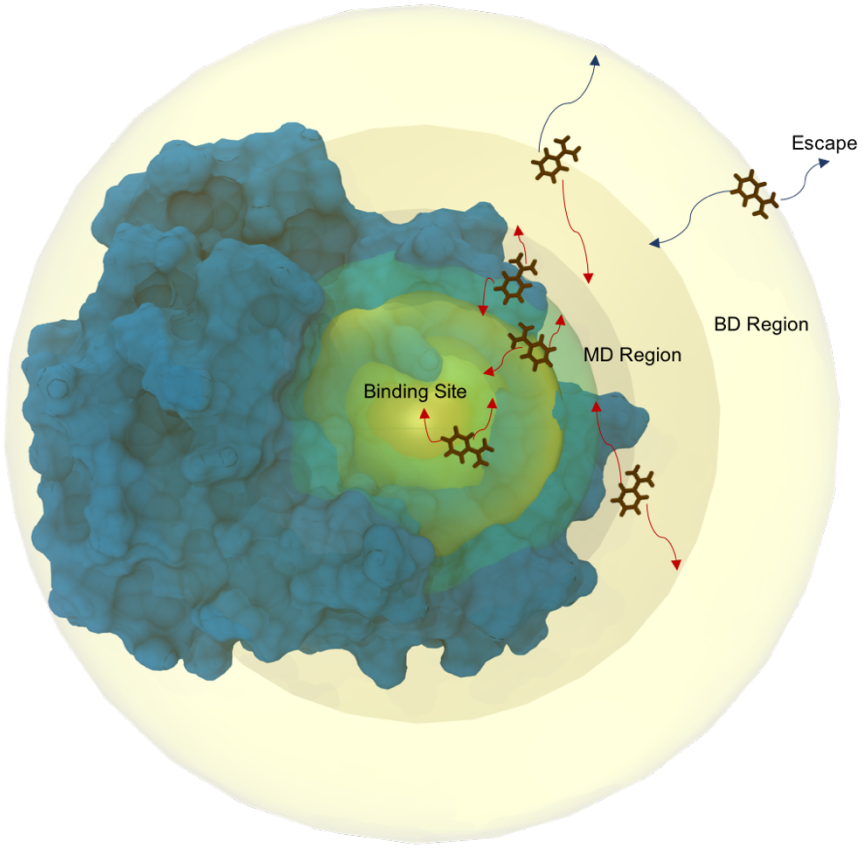

Figure 3.2 SEEKR is designed for calculations of ligand receptor binding kinetics in multiscale framework using molecular dynamics and Brownian dynamics.
Regions closest to the binding site are simulated with atomistic MD and the regions furthest away is simulated using rigid body BD. Ligands are placed on each spherical milestone and only simulated until an adjacent milestone is touched. Arrows in red represent MD trajectories and blue arrows represent BD trajectories. Statistics from each of the independent simulations are then combined to estimate association and dissociation rates as well as binding affinity.

3.3.3 Markov State Models and Reaction-Diffusion Methods

Markov state models (MSMs) describe memoryless transitions between different states of a system, enabling the study of long-timescale dynamics from short-timescale MD simulations of biomolecular systems [47,126,127]. Although MSMs alone can be used to investigate some instances of drug-protein binding [49,128–130], often this is not suitable as it is still limited to relatively small biomolecules and complexes and is limited by the high cost of MD simulations and the timescales involved. In contrast to MD, particle-based reaction-diffusion (RD) simulations are able to model larger and more complex systems but lack producing atomic level detail. By combining MD-based MSMs with RD simulations in a technique called MSM/RD, processes such as drug-protein binding can be modeled at large time- and lengthscales, while conserving atomistic details [131]. More specifically, MSM/RD can be used to model intracellular dynamics and describe diffusion, association, and dissociation on the cellular scale. Dibak *et al*. demonstrate the utility of MSM/RD approaches for biomolecular systems with application to carbon monoxide diffusion into the heme cavity of myoglobin. Extensions of this methodology which incorporate more complex cellular environments have the potential to become a powerful tool for studying off-target effects of drug molecules. The MSM/RD framework is also highly generalizable, with the potential be incorporated into many of the existing powerful RD tools [132].

3.3.4 Enhanced Sampling Simulations

Enhanced sampling simulations (usually based on atomistic MD) are increasingly used to study and predict drug binding kinetics. There is potential to apply enhanced sampling methods in multiscale formwork to extend their scope. For example, they can be applied with coarse-grained model linked to atomistic to extend across scales and at a higher level of detail can be linked to

more rigorous molecular representations based on quantum mechanical electronic structure calculations.

Atomistic molecular dynamics simulations can provide good descriptions of biomolecular interactions, and with free energy approaches can analyze determinants of binding affinity. However, even with emerging supercomputer resources, timescales are limited to the nanosecond-microsecond regime, so it is in general difficult to simulate multiple binding events unless a biasing or enhanced sampling method is applied to accelerate the process or focus sampling on a desired region of phase space. Many such methods exist, and can be applied with different potential functions (atomistic, coarse-grained, QM/MM, etc). An example of an enhanced sampling simulations approach is the calculation of residence times τ ($\tau = 1/k_{off}$) using τ-random accelerated molecular dynamics (τ-RAMD) for a diverse set of inhibitors of an important cancer target, the human N-terminal domain of heat shock protein 90α (N-HSP90). The τ-RAMD method relies on generating a random force which allows exit of the ligands within a short simulation time. τ-RAMD gives an excellent correlation between computed residence time (τ_{comp}) and measured τ_{expt} values for 78% of the compounds [107].

Metadynamics (MetaD) simulations [133] of various types are being increasingly widely used in drug discovery, e.g. prediction of binding kinetics and exploration of ligand binding or unbinding pathways [134,135]. MetaD sampling techniques reply on choosing appropriate collective variables (CV) to describe the slow degrees of freedom and it's a non-trivial task to achieve. Many recent developments have been evolved so far to address this issue [136]. Bernetti *et al*. showed that multiscale MetaD (a combination of MD, MSM and Path-CV) is capable to estimate the binding free energy of antagonist alprenolol to its β_2-adrenergic receptor [137]. Another development relies on the variational approach to conformational dynamics within the metadynamics (VAC-MetaD)

framework which address the slow degrees freedom of the systems under study [138]. Brotzakis *et al*. used this approach to trypsin-benzamidine complex and determined the absolute free energy of binding (ΔG_{bind}) as well as the unbinding rate (k_{off}) at low computational cost [139].

The accuracy of the predictions from such simulations, and ligand binding free energy calculations generally, may now be limited in many cases by the potential functions used: typical, atomistic 'molecular mechanics' forcefields do not allow for changes in electronic polarization and are limited in their description of electrostatics and dispersion effects [140]. Coarse-grained representations (which e.g. group together atoms in an amino acid sidechain), while allowing long simulations of large molecular systems, do not provide detailed descriptions of molecular interactions and by their nature cannot discriminate fine details. Calculations using quantum mechanics can in principle provide accurate descriptions of molecular structure, interactions and reactions, but while encouraging progress has been made in extending their range of applicability (e.g. to whole small proteins), their high computational demands means that they allow only very limited sampling [141,142].

3.3.5 Combining QM Methods with Atomistic Representations

Multiscale approaches provide a route to combine the accuracy of high-level methods with the need for significant sampling. Quantum mechanics/molecular mechanics (QM/MM) methods are a paradigm of multiscale modelling. They combine a electronic structure description of a small region with a simpler empirical MM (usually atomistic) representation of the surroundings (e.g. protein, solvent). The QM treatment can provide an accurate description of the electrostatics and polarization of the high-level region, and model chemical reactivity, e.g. to study covalent reactivity *in situ* [103]. As a ligand binds to a protein, its environment changes significantly, e.g. from

fully solvated to buried within the protein. This may cause significant changes in the polarization of the ligand preferably when a charged residue is close to ligand, affecting its interactions, but the effects of such changes are not included in invariant charge MM models. The effects of electronic polarization changes on binding kinetics can investigated by combining QM/MM free energy calculations with enhanced sampling simulation of binding (Figure 3.3). An example is the anticancer drug imatinib binding to c-Src kinase [143], combining metadynamics simulation of binding with QM/MM free energy corrections at critical points along the (un)binding pathway. In essence, the free energy change for changing from a MM to a QM treatment of the ligand is calculated, for the bound complex, for the transition state (TS), and for unbound ligand in solution, by replica exchange Monte Carlo simulations using a Metropolis-Hastings-Warshel algorithm [144]. The results show that there is a significant difference in the free energy between these environments when compared with MM ones, and polarization has the effect of increasing the off rate, bringing it closer to the experimental value.

QM/MM methods allow chemical reactions to be investigated in proteins and are now widely applied in modelling enzyme-catalyzed reactions, and increasingly in other relevant areas such as covalent inhibition and prediction of drug metabolism. In this last area, they can be combined with coarse-grained and atomistic molecular dynamics simulations. An example is a multiscale protocol combining simulations at these three different levels applied to metabolism of the anticoagulant *R*-warfarin by cytochrome P450 3A4 [145]. A model of the membrane-bound enzyme was assembled and investigated by coarse-grained MD, and this level of theory was also applied to investigate association with warfarin in the membrane. The coarse-grained models were converted into an atomistic simulation to investigate drug binding within the enzyme, and these simulations were in turn converted to QM/MM models to investigate oxidation R-warfarin in the

active site. The simulations showed important effects of the membrane, e.g. on the channels controlling access to the active site and gating residues. This multiscale coarse-grained-atomistic-QM/MM protocol is applicable to other membrane-bound enzymes.

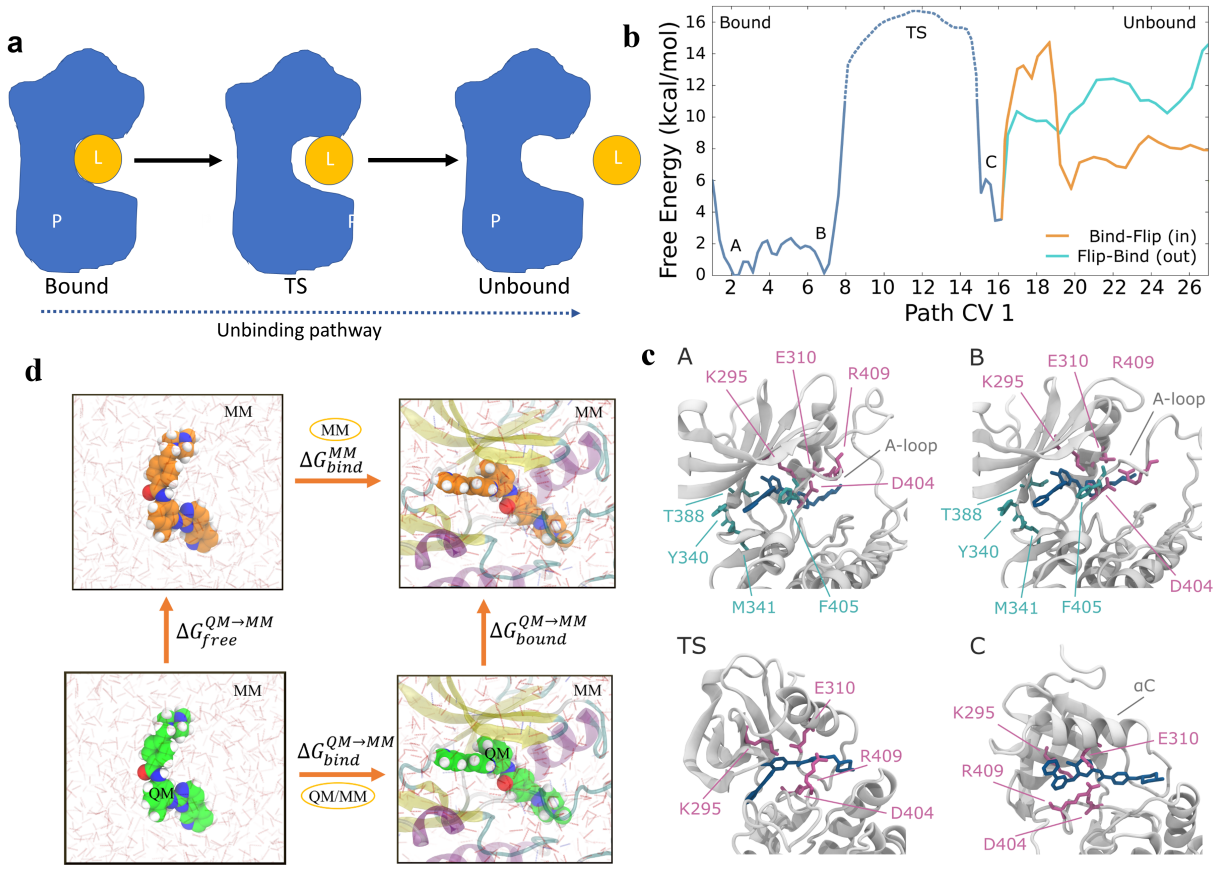

Figure 3.3 A practical Multiscale simulation approach to modeling drug-protein binding kinetics combing atomistic metadynamics simulations and QM/MM free energy calculations.
(A) A schematic representation of the unbinding pathway of the drug from its target showing relevant states: the bound protein-ligand complex, TS and unbound state. (B) The one-dimensional free energy profile calculated using a combination of Parallel tempering metadynamics (PT-MetaD) and Path-CVs. The profile shows that the drug unbinds though a TS with a barrier height of ~16.0 kcal/mol. (C) The representative structures/snapshots found in each minimum on the free energy profile. d. The free energy cycle used to calculate the MM to QM/MM free energy difference. The $\Delta G_{free}^{QM \rightarrow MM}$ and $G_{bound}^{QM \rightarrow MM}$ represents the correction free energy of the ligand in the bulk as well as in bound state, respectively.

3.4 Conclusions and Outlook

Multiscale biomolecular simulation methods are emerging and developing rapidly, promising increased insight and impact in drug development. Multiscale simulation methods connect across one or more scales to investigate, for example, how changes at one level drive or affected by changes at another, and by doing so bring a new depth of mechanistic understanding and unprecedented level of predictive power to drug discovery. Drug action is multiscale, in that molecular-level changes lead to macroscopic changes in biological systems, and in thus the process small molecule binding to a receptor of interest leads to changes across many levels. Here, we have focused on methods for understanding such binding and unbinding processes; these methods will contribute to understanding how they exert their effects at higher levels of biological organization. The challenges involved are many and varied, reflecting the complexity of biological systems and the dynamics of drug targets. As the examples reviewed briefly here show, significant progress has been made in integrating different types of simulation methods to link across diverse time- and length- scales. They have provided insight into the factors determining drug association rates and residence times, and the causes of drug resistance. Together with more expansive studies carried out on larger datasets, continued improvements to force fields, and connections to quantum mechanical approaches, the scope and power of multiscale simulation will certainly increase. Detailed comparison with experiment is essential in developing and testing such methods, which in turn will inform experimental design and analysis, and data engineering. Furthermore, additional potential will likely be realized through the incorporation of experimental and genetic data in fully integrative biological simulation methods [146].

3.5 Acknowledgements

Chapter 3, in full, has been submitted for publication as it may appear in "Jagger, B. R.†, Kochanek, S. E.†, Haldar, S., Amaro, R. E. & Mulholland, A. J. Multiscale Simulation Approaches to Drug-Protein Binding. *Current Opinion in Structural Biology*". The dissertation author was a primary coinvestigator and author of this work.

S.H. and A.J.M thank EPSRC for funding [grant nos. EP/M015378/1, EP/M022609/1], and the the Advanced Computing Research Centre (ACRC), University of Bristol - http://www.bris.ac.uk/acrc/ for computer time. B.R.J., S.E.K., and R.E.A. acknowledge support from the National Bio-medical Computation Resource (NBCR) NIH P41- GM103426. B.R.J. and S.E.K. also acknowledge support from the NIH Molecular Biophysics Training Program (T32-GM008326). The authors declare the following competing financial interest(s): R.E.A. is a co-founder of, on the scientific advisory board of, and has equity interest in Actavalon, Inc.

References

1. Centers for Disease Control and Prevention. Summary of the 2017-2018 Influenza Season. **11,** 7–10 (2018).

2. Wang, X., Jiang, H., Wu, P., Uyeki, T. M., Feng, L., Lai, S., Wang, L., Huo, X., Xu, K., Chen, E., Wang, X., He, J., Kang, M., Zhang, R., Zhang, J., Wu, J., Hu, S., Zhang, H., Liu, X., Fu, W., Ou, J., Wu, S., Qin, Y., Zhang, Z., Shi, Y., Zhang, J., Artois, J., Fang, V. J., Zhu, H., Guan, Y., Gilbert, M., Horby, P. W., Leung, G. M., Gao, G. F., Cowling, B. J. & Yu, H. Epidemiology of avian influenza A H7N9 virus in human beings across five epidemics in mainland China, 2013–17: an epidemiological study of laboratory-confirmed case series. *Lancet Infect. Dis.* **17,** 822–832 (2017).

3. De Clercq, E. Antiviral agents active against influenza A viruses. *Nat. Rev. Drug Discov.* **5,** 1015–1025 (2006).

4. Hamilton, B. S., Whittaker, G. R. & Daniel, S. Influenza virus-mediated membrane fusion: determinants of hemagglutinin fusogenic activity and experimental approaches for assessing virus fusion. *Viruses* **4,** 1144–1168 (2012).

5. von Itzstein, M. The war against influenza: discovery and development of sialidase inhibitors. *Nat Rev Drug Discov* **6,** 967–974 (2007).

6. Gamblin, S. J. & Skehel, J. J. Influenza hemagglutinin and neuraminidase membrane glycoproteins. *J Biol Chem* **285,** 28403–28409 (2010).

7. An, J., Lee, D. C., Law, A. H., Yang, C. L., Poon, L. L., Lau, A. S. & Jones, S. J. A novel small-molecule inhibitor of the avian influenza H5N1 virus determined through computational screening against the neuraminidase. *J Med Chem* **52,** 2667–2672 (2009).

8. Feng, E., Shin, W. J., Zhu, X., Li, J., Ye, D., Wang, J., Zheng, M., Zuo, J. P., No, K. T., Liu, X., Zhu, W., Tang, W., Seong, B. L., Jiang, H. & Liu, H. Structure-based design and synthesis of C-1- and C-4-modified analogs of zanamivir as neuraminidase inhibitors. *J Med Chem* **56,** 671–684 (2013).

9. Wen, W. H., Wang, S. Y., Tsai, K. C., Cheng, Y. S. E., Yang, A. S., Fang, J. M. & Wong, C. H. Analogs of zanamivir with modified C4-substituents as the inhibitors against the group-1 neuraminidases of influenza viruses. *Bioorganic Med. Chem.* **18,** 4074–4084 (2010).

10. Rudrawar, S., Dyason, J. C., Rameix-Welti, M. A., Rose, F. J., Kerry, P. S., Russell, R. J., van der Werf, S., Thomson, R. J., Naffakh, N. & von Itzstein, M. Novel sialic acid derivatives lock open the 150-loop of an influenza A virus group-1 sialidase. *Nat Commun* **1,** 113 (2010).

11. Amaro, R. E., Minh, D. D. L., Cheng, L. S., Lindstrom Jr., W. M., Olson, A. J., Lin, J.-H.,

Li, W. W. & McCammon, J. A. Remarkable loop flexibility in avian influenza N1 and its implications for antiviral drug design. *J. Am. Chem. Soc.* **129,** (2007).

12. Russell, R. J., Haire, L. F., Stevens, D. J., Collins, P. J., Lin, Y. P., Blackburn, G. M., Hay, A. J., Gamblin, S. J. & Skehel, J. J. The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design. *Nature* **443,** 45–49 (2006).

13. Woods, C. J., Malaisree, M., Pattarapongdilok, N., Sompornpisut, P., Hannongbua, S. & Mulholland, A. J. Long time scale GPU dynamics reveal the mechanism of drug resistance of the dual mutant I223R/H275Y neuraminidase from H1N1-2009 influenza virus. *Biochemistry* **51,** 4364–4375 (2012).

14. Amaro, R. E., Cheng, X., Ivanov, I., Xu, D. & McCammon, J. A. Characterizing loop dynamics and ligand recognition in human- and avian-type influenza neuraminidases via generalized born molecular dynamics and end-point free energy calculations. *J. Am. Chem. Soc.* **131,** (2009).

15. Hausmann, J., Kretzschmar, E., Garten, W. & Klenk, H.-D. N1 neuraminidase of influenza virus A / FPV / Rostock / 34 has haemadsorbing activity. *J. Gen. Virol.* **76,** 1719–1728 (1995).

16. Lai, J. C. C., Garcia, J. M., Dyason, J. C., Böhm, R., Madge, P. D., Rose, F. J., Nicholls, J. M., Peiris, J. S. M., Haselhorst, T. & Von-Itzstein, M. A secondary sialic acid binding site on influenza virus neuraminidase: Fact or fiction? *Angew. Chemie - Int. Ed.* **51,** 2221–2224 (2012).

17. Sung, J. C., Van Wynsberghe, A. W., Amaro, R. E., Li, W. W. & Andrew McCammon, J. Role of secondary sialic acid binding sites in influenza N1 neuraminidase. *J. Am. Chem. Soc.* **132,** (2010).

18. Amaro, R. E., Swift, R. V., Votapka, L., Li, W. W., Walker, R. C. & Bush, R. M. Mechanism of 150-cavity formation in influenza neuraminidase. *Nat. Commun.* **2,** (2011).

19. Li, Q., Qi, J., Zhang, W., Vavricka, C. J., Shi, Y., Wei, J., Feng, E., Shen, J., Chen, J., Liu, D., He, J., Yan, J., Liu, H., Jiang, H., Teng, M., Li, X. & Gao, G. F. The 2009 pandemic H1N1 neuraminidase N1 lacks the 150-cavity in its active site. *Nat. Struct. Mol. Biol.* **17,** 1266–1268 (2010).

20. Han, N. & Mu, Y. Plasticity of 150-Loop in Influenza Neuraminidase Explored by Hamiltonian Replica Exchange Molecular Dynamics Simulations. *PLoS One* **8,** e60995 (2013).

21. Laver, W. G., Colman, P. M., Webster, R. G., Hinshaw, V. S. & Air, G. M. Influenza virus neuraminidase with hemagglutinin activity. *Virology* **137,** 314–323 (1984).

22. Webster, R. G., Air, G. M., Metzger, D. W., Colman, P. M., Varghese, J. N., Baker, a T. & Laver, W. G. Antigenic structure and variation in an influenza virus N9 neuraminidase. *J. Virol.* **61,** 2910–6 (1987).

23. Varghese, J. N., Colman, P. M., van Donkelaar, A., Blick, T. J., Sahasrabudhe, A. & McKimm-Breschkin, J. L. Structural evidence for a second sialic acid binding site in avian influenza virus neuraminidases. *Proc Natl Acad Sci U S A* **94,** 11808–11812 (1997).

24. Air, G. M. Influenza neuraminidase. *Influenza Other Respi. Viruses* **6,** 245–256 (2012).

25. Streltsov, V. A., Pilling, P., Barrett, S. & McKimm-Breschkin, J. L. Catalytic mechanism and novel receptor binding sites of human parainfluenza virus type 3 hemagglutinin-neuraminidase (hPIV3 HN). *Antiviral Res.* **123,** 216–223 (2015).

26. Kobasa, D., Rodgers, M. E., Wells, K. & Kawaoka, Y. Neuraminidase hemadsorption activity, conserved in avian influenza A viruses, does not influence viral replication in ducks. *J. Virol.* **71,** 6706–6713 (1997).

27. Uhlendorff, J., Matrosovich, T., Klenk, H.-D. & Matrosovich, M. Functional significance of the hemadsorption activity of influenza virus neuraminidase and its alteration in pandemic viruses. *Arch. Virol.* **154,** 945–957 (2009).

28. Dai, M., McBride, R., Dortmans, J. C. F. M., Peng, W., Bakkers, M. J. G., de Groot, R. J., van Kuppeveld, F. J. M., Paulson, J. C., de Vries, E. & de Haan, C. A. M. Mutation of the Second Sialic Acid-Binding Site, Resulting in Reduced Neuraminidase Activity, Preceded the Emergence of H7N9 Influenza A Virus. *J. Virol.* **91,** (2017).

29. Du, W., Dai, M., Li, Z., Boons, G.-J., Peeters, B., van Kuppeveld, F. J. M., de Vries, E. & de Haan, C. A. M. Substrate Binding by the Second Sialic Acid-Binding Site of Influenza A Virus N1 Neuraminidase Contributes to Enzymatic Activity. *J. Virol.* **92,** (2018).

30. Du, W., Guo, H., Nijman, V. S., Doedt, J., van der Vries, E., van der Lee, J., Li, Z., Boons, G.-J., van Kuppeveld, F. J. M., de Vries, E., Matrosovich, M. & de Haan, C. A. M. The 2nd sialic acid-binding site of influenza A virus neuraminidase is an important determinant of the hemagglutinin-neuraminidase-receptor balance. *PLOS Pathog.* **15,** e1007860 (2019).

31. Ivinson, K., Deliyannis, G., McNabb, L., Grollo, L., Gilbertson, B., Jackson, D. & Brown, L. E. Salivary Blockade Protects the Lower Respiratory Tract of Mice from Lethal Influenza Virus Infection. *J. Virol.* **91,** (2017).

32. Gilbertson, B., Ng, W. C., Crawford, S., McKimm-Breschkin, J. L. & Brown, L. E. Mouse Saliva Inhibits Transit of Influenza Virus to the Lower Respiratory Tract by Efficiently Blocking Influenza Virus Neuraminidase Activity. *J. Virol.* **91,** (2017).

33. Xu, R., Ekiert, D., Krause, J., Hai, R., Crowe, J. & Wilson, I. Structural Basis of Preexisting Immunity to the 2009 H1N1 Pandemic Influenza Virus. *Science (80-.).* **328,** 357–360 (2010).

34. 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.* **103,** 19123–19127 (2006).

35. Hadden, J. A. & Perilla, J. R. All-atom virus simulations. *Curr. Opin. Virol.* **31,** 82–91 (2018).

36. 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.* **128,** 121–132 (2017).

37. Freddolino, P. L., Arkhipov, A. S., Larson, S. B., McPherson, A. & Schulten, K. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure* **14,** 437–449 (2006).

38. Zink, M. & Grubmüller, H. Mechanical properties of the icosahedral shell of southern bean mosaic virus: A molecular dynamics study. *Biophys. J.* **96,** 1350–1363 (2009).

39. Ayton, G. S. & Voth, G. A. Multiscale computer simulation of the immature HIV-1 virion. *Biophys. J.* **99,** 2757–2765 (2010).

40. Roberts, J. A., Kuiper, M. J., Thorley, B. R., Smooker, P. M. & Hung, A. Investigation of a predicted N-terminal amphipathic α-helix using atomistic molecular dynamics simulation of a complete prototype poliovirus virion. *J. Mol. Graph. Model.* **38,** 165–173 (2012).

41. Zhao, G., Perilla, J. R., Yufenyuy, E. L., Meng, X., Chen, B., Ning, J., Ahn, J., Gronenborn, A. M., Schulten, K., Aiken, C. & Zhang, P. Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* **497,** 643–646 (2013).

42. 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.* **141,** 165101 (2014).

43. 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* **23,** 584–597 (2015).

44. Reddy, T. & Sansom, M. S. P. The Role of the Membrane in the Structure and Biophysical Robustness of the Dengue Virion Envelope. *Structure* **24,** 375–382 (2016).

45. Marzinek, J. K., Holdbrook, D. A., Huber, R. G., Verma, C. & Bond, P. J. Pushing the Envelope: Dengue Viral Membrane Coaxed into Shape by Molecular Simulations. *Structure* **24,** 1410–1420 (2016).

46. Pande, V. S., Beauchamp, K. & Bowman, G. R. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **52,** 99–105 (2010).

47. Prinz, J. H., Wu, H., Sarich, M., Keller, B., Senne, M., Held, M., Chodera, J. D., Schtte, C. & Noé, F. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.* **134,** 174105 (2011).

48. Chodera, J. D. & Noé, F. Markov state models of biomolecular conformational dynamics.

Curr. Opin. Struct. Biol. **25,** 135–144 (2014).

49. Plattner, N. & Noé, F. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. *Nat. Commun.* **6,** 7653 (2015).

50. Plattner, N., Doerr, S., De Fabritiis, G. & Noé, F. Complete protein-protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nat. Chem.* **9,** 1005–1011 (2017).

51. Voelz, V. A., Bowman, G. R., Beauchamp, K. & Pande, V. S. Molecular Simulation of ab Initio Protein Folding for a Millisecond Folder NTL9(1-39). *J. Am. Chem. Soc.* **132,** 1526–1528 (2010).

52. Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kal??, L. & Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26,** 1781–1802 (2005).

53. Huang, J. & MacKerell, A. D. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J. Comput. Chem.* **34,** 2135–2145 (2013).

54. Durrant, J. D., Bush, R. M. & Amaro, R. E. Microsecond Molecular Dynamics Simulations of Influenza Neuraminidase Suggest a Mechanism for the Increased Virulence of Stalk-Deletion Mutants. *J. Phys. Chem. B* **120,** (2016).

55. Durrant, J. D., Votapka, L., Sørensen, J. & Amaro, R. E. POVME 2.0: An enhanced tool for determining pocket shape and volume characteristics. *J. Chem. Theory Comput.* **10,** (2014).

56. Ngan, C. H., Bohnuud, T., Mottarella, S. E., Beglov, D., Villar, E. A., Hall, D. R., Kozakov, D. & Vajda, S. FTMAP: extended protein mapping with user-selected probe molecules. *Nucleic Acids Res* **40,** W271-5 (2012).

57. Marino, K., Bones, J., Kattla, J. J. & Rudd, P. M. A systematic approach to protein glycosylation analysis: a path through the maze. *Nat Chem Biol* **6,** 713–723 (2010).

58. Liu, J., Stevens, D. J., Haire, L. F., Walker, P. A., Coombs, P. J., Russell, R. J., Gamblin, S. J. & Skehel, J. J. Structures of receptor complexes formed by hemagglutinins from the Asian Influenza pandemic of 1957. *Proc Natl Acad Sci U S A* **106,** 17175–17180 (2009).

59. Dolinsky, T. J., Czodrowski, P., Li, H., Nielsen, J. E., Jensen, J. H., Klebe, G. & Baker, N. A. PDB2PQR: Expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **35,** W522-5 (2007).

60. Olsson, M. H., Sondergaard, C. R., Rostkowski, M. & Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J Chem Theory Comput* **7,** 525–537 (2011).

61. Sondergaard, C. R., Olsson, M. H., Rostkowski, M. & Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *J Chem Theory Comput* **7,** 2284–2295 (2011).

62. Li, H., Robertson, A. D. & Jensen, J. H. Very fast empirical prediction and rationalization of protein pKa values. *Proteins Struct. Funct. Bioinforma.* **61,** 704–721 (2005).

63. Bas, D. C., Rogers, D. M. & Jensen, J. H. Very fast prediction and rationalization of pKa values for protein–ligand complexes. *Proteins Struct. Funct. Bioinforma.* **73,** 765–783 (2008).

64. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).

65. Hartmann, M. D., Ridderbusch, O., Zeth, K., Albrecht, R., Testa, O., Woolfson, D. N., Sauer, G., Dunin-Horkawicz, S., Lupas, A. N. & Alvarez, B. H. A coiled-coil motif that sequesters ions to the hydrophobic core. *Proc Natl Acad Sci U S A* **106,** 16950–16955 (2009).

66. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **14,** 33–38 (1996).

67. Ropp, P., Friedman, A. & Durrant, J. D. Scoria: a Python module for manipulating 3D molecular data. *J Cheminform* **9,** 52 (2017).

68. Daura, X., Gademann, K., Jaun, B., Seebach, D., van Gunsteren, W. F. & Mark, A. E. Peptide folding: When simulation meets experiment. *Angew. Chemie-International Ed.* **38,** 236–240 (1999).

69. Christen, M., Hunenberger, P. H., Bakowies, D., Baron, R., Burgi, R., Geerke, D. P., Heinz, T. N., Kastenholz, M. A., Krautler, V., Oostenbrink, C., Peter, C., Trzesniak, D. & Van Gunsteren, W. F. The GROMOS software for biomolecular simulation: GROMOS05. *J. Comput. Chem.* **26,** 1719–1751 (2005).

70. Durrant, J. D. & McCammon, J. A. Autoclickchem: Click chemistry in silico. *PLoS Comput. Biol.* **8,** e1002397–e1002397 (2012).

71. Durrant, J. D. & Amaro, R. E. LipidWrapper: an algorithm for generating large-scale membrane models of arbitrary geometry. *PLoS Comput. Biol.* **10,** e1003720 (2014).

72. 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.* **97,** 50–58 (2009).

73. Jo, S., Kim, T. & Im, W. Automated builder and database of protein/membrane complexes for molecular dynamics simulations. *PLoS One* **2,** (2007).

74. Oliphant, T. & Millma, J. k. A guide to NumPy. *Trelgol Publ.* 22–30 (2006). at <https://www.numpy.org/>

75. 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.* **11,** 5525–5542 (2015).

76. Lambert, L. C. & Fauci, A. S. Influenza vaccines for the future. *N. Engl. J. Med.* **363,** 2036–2044 (2010).

77. Fiore, A. E., Uyeki, T. M., Broder, K., Finelli, L., Euler, G. L., Singleton, J. A., Iskander, J. K., Wortley, P. M., Shay, D. K., Bresee, J. S. & Cox, N. J. Prevention and control of influenza with vaccines: recommendations of the Advisory Committee on Immunization Practices (ACIP), 2010. *MMWR. Recomm. Rep.* **59,** 1–62 (2010).

78. Samson, M., Pizzorno, A., Abed, Y. & Boivin, G. Influenza virus resistance to neuraminidase inhibitors. *Antiviral Res.* **98,** 174–185 (2013).

79. Dong, G., Peng, C., Luo, J., Wang, C., Han, L., Wu, B., Ji, G. & He, H. Adamantane-resistant influenza a viruses in the world (1902-2013): Frequency and distribution of M2 gene mutations. *PLoS One* **10,** 1–20 (2015).

80. Skehel, J. J. & Wiley, D. C. Receptor Binding and Membrane Fusion in Virus Entry: The Influenza Hemagglutinin. *Annu. Rev. Biochem.* **69,** 531–569 (2000).

81. Wilson, I. A., Skehel, J. J. & Wiley, D. C. Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 A resolution. *Nature* **289,** 366–373 (1981).

82. Brancato, V., Peduto, A., Wharton, S., Martin, S., More, V., Di Mola, A., Massa, A., Perfetto, B., Donnarumma, G., Schiraldi, C., Tufano, M. A., de Rosa, M., Filosa, R. & Hay, A. Design of inhibitors of influenza virus membrane fusion: Synthesis, structure-activity relationship and in vitro antiviral activity of a novel indole series. *Antiviral Res.* **99,** 125–135 (2013).

83. Kadam, R. U. & Wilson, I. A. Structural basis of influenza virus fusion inhibition by the antiviral drug Arbidol. *Proc. Natl. Acad. Sci. U. S. A.* **114,** 206–214 (2017).

84. Ekiert, D. C., Kashyap, A. K., Steel, J., Rubrum, A., Bhabha, G., Khayat, R., Lee, J. H., Dillon, M. A., O'Neil, R. E., Faynboym, A. M., Horowitz, M., Horowitz, L., Ward, A. B., Palese, P., Webby, R., Lerner, R. A., Bhatt, R. R. & Wilson, I. A. Cross-neutralization of influenza A viruses mediated by a single antibody loop. *Nature* **489,** 526–532 (2012).

85. Gamblin, S. J., Haire, L. F., Russell, R. J., Stevens, D. J., Xiao, B., Ha, Y., Vasisht, N., Steinhauer, D. A., Daniels, R. S., Elliot, A., Wiley, D. C. & Skehel, J. J. The Structure and Receptor Binding Properties of the 1918 Influenza Hemagglutinin. *Science (80-.).* **303,** 1838–1842 (2004).

86. Durrant, J. D., Kochanek, S. E., Casalino, L., Ieong, P. U., Dommer, A. & Amaro, R. E. Mesoscale All-Atom Influenza Virus Simulations Suggest New Substrate Binding Mechanism. *ACS Cent. Sci. Submitt.* (2019).

87. Amaro, R. E. & Mulholland, A. J. Multiscale methods in drug design bridge chemical and biological complexity in the search for cures. *Nat. Rev. Chem.* **2,** 0148 (2018).

88. Husic, B. E. & Pande, V. S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **140,** 2386–2396 (2018).

89. Johnson, G. T., Autin, L., Al-Alusi, M., Goodsell, D. S., Sanner, M. F. & Olson, A. J. cellPACK: a virtual mesoscope to model and visualize structural systems biology. *Nat. Methods* **12,** 85 (2014).

90. Schrödinger Release 2018-3: Glide. (2018).

91. Kozakov, D., Grove, L. E., Hall, D. R., Bohnuud, T., Mottarella, S. E., Luo, L., Xia, B., Beglov, D. & Vajda, S. The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat. Protoc.* **10,** 733–755 (2015).

92. 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.* **139,** 15102 (2013).

93. Copeland, R. A., Pompliano, D. L. & Meek, T. D. Drug–Target Residence Time and Its Implications for Lead Optimization. *Nat. Rev. Drug Discov.* **5,** 730–739 (2006).

94. Tummino, P. J. & Copeland, R. A. Residence Time of Receptor−Ligand Complexes and Its Effect on Biological Function. *Biochemistry* **47,** 5481–5492 (2008).

95. Copeland, R. A. The Drug-Target Residence Time Model: A 10-Year Retrospective. *Nat. Rev. Drug Discov.* **15,** 87–95 (2016).

96. Swinney, D. C. Opinion: Biochemical Mechanisms of Drug Action: What Does it Take for Success? *Nat. Rev. Drug Discov.* **3,** 801–808 (2004).

97. Gao, M. & Skolnick, J. The distribution of ligand-binding pockets around protein-protein interfaces suggests a general mechanism for pocket formation. *Proc. Natl. Acad. Sci.* **109,** 3784–3789 (2012).

98. Schiebel, J., Gaspari, R., Wulsdorf, T., Ngo, K., Sohn, C., Schrader, T. E., Cavalli, A., Ostermann, A., Heine, A. & Klebe, G. Intriguing role of water in protein-ligand binding studied by neutron crystallography on trypsin complexes. *Nat. Commun.* **9,** 3559 (2018).

99. Du, X., Li, Y., Xia, Y.-L., Ai, S.-M., Liang, J., Sang, P., Ji, X.-L. & Liu, S.-Q. Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods. *Int. J. Mol. Sci.* **17,** 144 (2016).

100. De Vivo, M., Masetti, M., Bottegoni, G. & Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **59,** 4035–4061 (2016).

101. Amaro, R. E. & Mulholland, A. J. Multiscale methods in drug design bridge chemical and

biological complexity in the search for cures. *Nat. Rev. Chem.* **2,** 148 (2018).

102. Woods, C. J., Malaisree, M., Pattarapongdilok, N., Sompornpisut, P., Hannongbua, S. & Mulholland, A. J. Long Time Scale GPU Dynamics Reveal the Mechanism of Drug Resistance of the Dual Mutant I223R/H275Y Neuraminidase from H1N1-2009 Influenza Virus. *Biochemistry* **51,** 4364–4375 (2012).

103. Callegari, D., Ranaghan, K. E., Woods, C. J., Minari, R., Tiseo, M., Mor, M., Mulholland, A. J. & Lodola, A. L718Q mutant EGFR escapes covalent inhibition by stabilizing a non-reactive conformation of the lung cancer drug osimertinib. *Chem. Sci.* **9,** 2740–2749 (2018).

104. Lee, C. T. & Amaro, R. E. Exascale Computing: A New Dawn for Computational Biology. *Comput. Sci. Eng.* **20,** 18–25 (2018).

105. Bruce, N. J., Ganotra, G. K., Kokh, D. B., Sadiq, S. K. & Wade, R. C. New approaches for computing ligand – receptor binding kinetics. *Curr. Opin. Struct. Biol.* **49,** 1–10 (2018).

106. Bernetti, M., Masetti, M., Rocchia, W. & Cavalli, A. Kinetics of Drug Binding and Residence Time. *Annu. Rev. Phys. Chem.* **70,** 143–171 (2019).

107. Kokh, D. B., Amaral, M., Bomke, J., Grädler, U., Musil, D., Buchstaller, H.-P., Dreyer, M. K., Frech, M., Lowinski, M., Vallee, F., Bianciotto, M., Rak, A. & Wade, R. C. Estimation of Drug-Target Residence Times by τ-Random Acceleration Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **14,** 3859–3869 (2018).

108. Huggins, D. J., Biggin, P. C., Dämgen, M. A., Essex, J. W., Harris, S. A., Henchman, R. H., Khalid, S., Kuzmanic, A., Laughton, C. A., Michel, J., Mulholland, A. J., Rosta, E., Sansom, M. S. P. & van der Kamp, M. W. Biomolecular simulations: From dynamics and mechanisms to computational assays of biological activity. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **9,** e1393 (2019).

109. Song, Y., Zhang, Y., Shen, T., Bajaj, C. L., McCammon, J. A. & Baker, N. A. Finite Element Solution of the Steady-State Smoluchowski Equation for Rate Constant Calculations. *Biophys. J.* **86,** 2017–2029 (2004).

110. Schreiber, G. Kinetic studies of protein-protein interactions. *Curr. Opin. Struct. Biol.* **12,** 41–7 (2002).

111. Northrup, S. H., Allison, S. A. & McCammon, J. A. Brownian Dynamics Simulation of Diffusion-Influenced Bimolecular Reactions. *J. Chem. Phys.* **80,** 1517–1524 (1984).

112. Luty, B. A., El Amrani, S. & McCammon, J. A. Simulation of the bimolecular reaction between superoxide and superoxide dismutase: synthesis of the encounter and reaction steps. *J. Am. Chem. Soc.* **115,** 11874–11877 (1993).

113. Chang, C.-E. A., Trylska, J., Tozzini, V. & Andrew McCammon, J. Binding Pathways of Ligands to HIV-1 Protease: Coarse-grained and Atomistic Simulations. *Chem. Biol. Drug Des.* **69,** 5–13 (2007).

114. Huang, Y. M., Kang, M. & Chang, C. A. Switches of hydrogen bonds during ligand-protein association processes determine binding kinetics. *J. Mol. Recognit.* **27,** 537–548 (2014).

115. Huang, Y. M., Raymundo, M. A. V., Chen, W. & Chang, C. A. Mechanism of the Association Pathways for a Pair of Fast and Slow Binding Ligands of HIV-1 Protease. *Biochemistry* **56,** 1311–1323 (2017).

116. Faradjian, A. K. & Elber, R. Computing Time Scales from Reaction Coordinates by Milestoning. *J. Chem. Phys.* **120,** 10880–10889 (2004).

117. Vanden-Eijnden, E., Venturoli, M., Ciccotti, G. & Elber, R. On the assumptions underlying milestoning. *J. Chem. Phys.* **129,** 174102 (2008).

118. Májek, P. & Elber, R. Milestoning without a Reaction Coordinate. *J. Chem. Theory Comput.* **6,** 1805–1817 (2010).

119. Bello-Rivas, J. M. & Elber, R. Exact milestoning. *J. Chem. Phys.* **142,** 094102 (2015).

120. Maragliano, L., Vanden-Eijnden, E. & Roux, B. Free Energy and Kinetics of Conformational Transitions from Voronoi Tessellated Milestoning with Restraining Potentials. *J. Chem. Theory Comput.* **5,** 2589–2594 (2009).

121. Yu, T.-Q., Lapelosa, M., Vanden-Eijnden, E. & Abrams, C. F. Full Kinetics of CO Entry, Internal Diffusion, and Exit in Myoglobin from Transition-Path Theory Simulations. *J. Am. Chem. Soc.* **137,** 3041–3050 (2015).

122. Bucci, A., Yu, T.-Q., Vanden-Eijnden, E. & Abrams, C. F. Kinetics of O 2 Entry and Exit in Monomeric Sarcosine Oxidase via Markovian Milestoning Molecular Dynamics. *J. Chem. Theory Comput.* **12,** 2964–2972 (2016).

123. Votapka, L. W. & Amaro, R. E. Multiscale Estimation of Binding Kinetics Using Brownian Dynamics, Molecular Dynamics and Milestoning. *PLOS Comput. Biol.* **11,** e1004381 (2015).

124. Votapka, L. W., Jagger, B. R., Heyneman, A. L. & Amaro, R. E. SEEKR: Simulation Enabled Estimation of Kinetic Rates, A Computational Tool to Estimate Molecular Kinetics and Its Application to Trypsin–Benzamidine Binding. *J. Phys. Chem. B* **121,** 3597–3606 (2017).

125. Jagger, B. R., Lee, C. T. & Amaro, R. E. Quantitative Ranking of Ligand Binding Kinetics with a Multiscale Milestoning Simulation Approach. *J. Phys. Chem. Lett.* acs.jpclett.8b02047 (2018). doi:10.1021/acs.jpclett.8b02047

126. Pande, V. S., Beauchamp, K. & Bowman, G. R. *Everything you wanted to know about Markov State Models but were afraid to ask*. *Methods* **52,** 99–105 (2010).

127. Chodera, J. D. & Noé, F. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.* **25,** 135–144 (2014).

128. Silva, D.-A., Bowman, G. R., Sosa-Peinado, A. & Huang, X. A Role for Both Conformational Selection and Induced Fit in Ligand Binding by the LAO Protein. *PLoS Comput. Biol.* **7,** e1002054 (2011).

129. De Sancho, D., Kubas, A., Wang, P.-H., Blumberger, J. & Best, R. B. Identification of Mutational Hot Spots for Substrate Diffusion: Application to Myoglobin. *J. Chem. Theory Comput.* **11,** 1919–1927 (2015).

130. Kubas, A., Orain, C., De Sancho, D., Saujet, L., Sensi, M., Gauquelin, C., Meynial-Salles, I., Soucaille, P., Bottin, H., Baffert, C., Fourmond, V., Best, R. B., Blumberger, J. & Léger, C. Mechanism of O2 diffusion and reduction in FeFe hydrogenases. *Nat. Chem.* **9,** 88–95 (2017).

131. Dibak, M., del Razo, M. J., De Sancho, D., Schütte, C. & Noé, F. MSM/RD: Coupling Markov state models of molecular kinetics with reaction-diffusion simulations. *J. Chem. Phys.* **148,** 214107 (2018).

132. Roberts, E., Stone, J. E. & Luthey-Schulten, Z. Lattice microbes: High-performance stochastic simulation method for the reaction-diffusion master equation. *J. Comput. Chem.* **34,** 245–255 (2013).

133. Valsson, O., Tiwary, P. & Parrinello, M. Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. *Annu. Rev. Phys. Chem.* **67,** 159–184 (2016).

134. Cavalli, A., Spitaleri, A., Saladino, G. & Gervasio, F. L. Investigating Drug–Target Association and Dissociation Mechanisms Using Metadynamics-Based Algorithms. *Acc. Chem. Res.* **48,** 277–285 (2015).

135. Morando, M. A., Saladino, G., D'Amelio, N., Pucheta-Martinez, E., Lovera, S., Lelli, M., López-Méndez, B., Marenchino, M., Campos-Olivas, R. & Gervasio, F. L. Conformational Selection and Induced Fit Mechanisms in the Binding of an Anticancer Drug to the c-Src Kinase. *Sci. Rep.* **6,** 24439 (2016).

136. Invernizzi, M. & Parrinello, M. Making the Best of a Bad Situation: A Multiscale Approach to Free Energy Calculation. *J. Chem. Theory Comput.* **15,** 2187–2194 (2019).

137. Bernetti, M., Masetti, M., Recanatini, M., Amaro, R. E. & Cavalli, A. An Integrated Markov State Model and Path Metadynamics Approach To Characterize Drug Binding Processes. *J. Chem. Theory Comput.* acs.jctc.9b00450 (2019). doi:10.1021/acs.jctc.9b00450

138. McCarty, J. & Parrinello, M. A variational conformational dynamics approach to the selection of collective variables in metadynamics. *J. Chem. Phys.* **147,** 204109 (2017).

139. Brotzakis, Z. F., Limongelli, V. & Parrinello, M. Accelerating the Calculation of Protein–Ligand Binding Free Energy and Residence Times Using Dynamically Optimized Collective Variables. *J. Chem. Theory Comput.* **15,** 743–750 (2019).

140. Rizzi, A., Murkli, S., McNeill, J. N., Yao, W., Sullivan, M., Gilson, M. K., Chiu, M. W., Isaacs, L., Gibb, B. C., Mobley, D. L. & Chodera, J. D. Overview of the SAMPL6 host–guest binding affinity prediction challenge. *J. Comput. Aided. Mol. Des.* **32,** 937–963 (2018).

141. Sisto, A., Stross, C., van der Kamp, M. W., O'Connor, M., McIntosh-Smith, S., Johnson, G. T., Hohenstein, E. G., Manby, F. R., Glowacki, D. R. & Martinez, T. J. Atomistic non-adiabatic dynamics of the LH2 complex with a GPU-accelerated ab initio exciton model. *Phys. Chem. Chem. Phys.* **19,** 14924–14936 (2017).

142. Lever, G., Cole, D. J., Lonsdale, R., Ranaghan, K. E., Wales, D. J., Mulholland, A. J., Skylaris, C.-K. & Payne, M. C. Large-Scale Density Functional Theory Transition State Searching in Enzymes. *J. Phys. Chem. Lett.* **5,** 3614–3619 (2014).

143. Haldar, S., Comitani, F., Saladino, G., Woods, C., van der Kamp, M. W., Mulholland, A. J. & Gervasio, F. L. A Multiscale Simulation Approach to Modeling Drug–Protein Binding Kinetics. *J. Chem. Theory Comput.* **14,** 6093–6101 (2018).

144. Woods, C. J. & Mulholland, A. J. Multiscale modelling of biological systems. *Chem. Model.* **5,** 13–50 (2008).

145. Lonsdale, R., Rouse, S. L., Sansom, M. S. P. & Mulholland, A. J. A Multiscale Approach to Modelling Drug Metabolism by Membrane-Bound Cytochrome P450 Enzymes. *PLoS Comput. Biol.* **10,** e1003714 (2014).

146. Olsson, S., Wu, H., Paul, F., Clementi, C. & Noé, F. Combining experimental and simulation data of molecular processes via augmented Markov models. *Proc. Natl. Acad. Sci.* **114,** 8265–8270 (2017).