Solving the Mechanism of Na^+/H^+ Antiporters Using Molecular Dynamics Simulations

by

David L. Dotson

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2016 by the
Graduate Supervisory Committee:

Oliver Beckstein, Chair
Sefika Banu Ozkan
Robert Ros
Wade Van Horn

ARIZONA STATE UNIVERSITY

December 2016

ABSTRACT

Na^+/H^+ antiporters are vital membrane proteins for cell homeostasis, transporting Na^+ ions in exchange for H^+ across the lipid bilayer. In humans, dysfunction of these transporters are implicated in hypertension, heart failure, epilepsy, and autism, making them well-established drug targets. Although experimental structures for bacterial homologs of the human Na^+/H^+ have been obtained, the detailed mechanism for ion transport is still not well-understood. The most well-studied of these transporters, *Escherichia coli* NhaA, known to transport 2 H^+ for every Na^+ extruded, was recently shown to bind H^+ and Na^+ at the same binding site, for which the two ion species compete. Using molecular dynamics simulations, the work presented in this dissertation shows that Na^+ binding disrupts a previously-unidentified salt bridge between two conserved residues, suggesting that one of these residues, Lys300, may participate directly in transport of H^+. This work also demonstrates that the conformational change required for ion translocation in a homolog of NhaA, *Thermus thermophilus* NapA, thought by some to involve only small helical movements at the ion binding site, is a large-scale, rigid-body movement of the core domain relative to the dimerization domain. This elevator-like transport mechanism translates a bound Na^+ up to 10 Å across the membrane. These findings constitute a major shift in the prevailing thought on the mechanism of these transporters, and serve as an exciting launchpad for new developments toward understanding that mechanism in detail.

To my family and to my friends, past and present

ACKNOWLEDGMENTS

My Ph.D. has been a long and often rocky road, and I could not have finished it without the people in my life. The most impactful of these was my advisor, Oliver Beckstein, who in his infinite grace allowed me to join the lab after my very tumultuous first year of graduate school. I was very fortunate to have in him a mentor that simultaneously managed to push me to do great work while allowing me great freedom in *how* I performed that work. Our relationship through the years has been of great mutual benefit, advancing both the science the lab produces and the lab itself in its most formative years. I am proud and honored to have been a part of that period in his career.

I also want to thank my office colleagues, in particular Avishek Kumar, Sean Seyler, Allan Friesen, Daniel Martin, Varda Faghir Hagh, Paul Campitelli, and Salman Seyedi, not only for the intellectual support and growth our interactions have inspired, but also the emotional support I so often needed to keep going. In academia, we have a tendency to focus so closely on the science that we can forget that people make it happen, and that life is often more complicated than it appears in the lab. The friendships forged over these years are made of hard stuff, since it was hard times that made them.

I should thank the institutions that enabled this work, in particular the Department of Physics at Arizona State University, in which I have been employed. Financial support came from a variety of sources. I am particularly grateful to the Department of Education for the GAANN Fellowship, which generously funded me in my first year, as well as to Dr. Tianwei Jing for endowing the Molecular Imaging Corporation Fellowship, which I received in 2015. I also acknowledge travel funding from the NSF, and most recently grant funding from the NIH.

Finally, none of this would have been possible without the incredible support of my family and friends. I am extremely fortunate to have had the support network I needed to see this through.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Chapter 1

INTRODUCTION

In all cells, active transport of ions against an electrochemical gradient across the cell membrane is mediated by integral membrane transporter proteins[1–3]. For many such proteins, we possess knowledge of both the overall function ("what they do") and the primary amino acid sequence, but for most we do not know how they are structured or how they fulfill their function ("how they do it"). That is, we do not know, thoroughly, the molecular mechanism of transport.

One such family of transporters is the Na^+/H^+ exchangers, which in mammals mediate the outward movement of protons (H^+) in exchange for sodium (Na^+) ions[4]. These transporters are important in regulating intracellular pH, sodium levels, and cell volume, and they also play roles in control of the cell cycle, cell proliferation, cell migration and vesicle trafficking[4,5]. In humans, these transporters are also linked to various diseases, such as hypertension, heart failure, epilepsy, and autism making them well-established drug targets[4,6,7]. It is understanding the detailed functional mechanism of these transporters that the work detailed in this dissertation is focused on.

1.1 The Early History of Na^+/H^+ Antiporters

The effort to understand the mechanism of Na^+/H^+ antiporters started over 40 years ago. This history is rich with experimental findings of progressively greater detail, often tracking with the development of new methods. In 1974, Ian West and Peter Mitchell discovered that the addition of Na^+ to a suspension of *Escherichia coli* resulted in a change in pH[8]. They concluded that the membrane of *E. coli* contained

antiporter proteins that exchanged H^+ for Na^+ ions, allowing the cell to maintain its preferred internal concentration of Na^+ [8]. Two years later, the pH range over which antiport activity in *E. coli* occurs was observed by Etana Padan and coworkers to start at pH 6.5, reaching maximum activity at pH 8.5 [9].

In 1987, the gene encoding the transporter in *E. coli*, *nhaA*, was isolated [10], and in the next year the gene was sequenced [11]. Finally, in 1991, the protein encoded by the gene was successfully purified [12]. This paved the way for detailed studies of the activity of the transporter under highly controlled conditions in proteoliposomes. The maximum transport rate of the transporter, now termed NhaA, was found to be ~ 1500 s^{-1} [12]. Its electrogenicity, long unknown and postulated to be electroneutral [8], was also established as 1 Na^+ extruded per 2 H^+ transferred into the cell [13].

By this time, the ability to produce and purify NhaA mutants from its known sequence enabled the design and execution of functional studies aimed at identifying the residues vital for function. In 1995, it was found that three aspartates, highly conserved across orthologs of NhaA, were vital for activity of the transporter [14]. Additional functional studies suggested the role of many other residues as part of a "pH sensor," responsible for regulating the activity of the transporter in response to low pH conditions [15,16].

1.2 Experimental Structures Offer New Insights Into *E. coli* NhaA

The first structural data for NhaA, a two-dimensional projection at 4.0 Å resolution, was produced using electron cryomicroscopy in 1999 [17]. The end of the 20th century was also accompanied by the first three-dimensional structure of NhaA, a 7 Å-resolution map of the transmembrane helices [18]. This gave, for the first time, a view of the NhaA structural fold, revealing 12 transmembrane helices organized into two distinct domains (the dimerization interface, and the "core" domain), each

comprised of 6 helices. The protein was observed to span ∼45 Å, the approximate thickness of the biological membrane bilayer.

Functional studies continued to shed light on the detailed behavior of NhaA. It was found that residues on helix IV were important for transport, suggesting a role as part of the putative "pH sensor" [19]. Another study discovered that cysteine replacements of certain residues, creating crosslinks within the protein, abolished transport [20]. This suggested that a conformational change was required for transport, a feature disrupted by the crosslinks. Other cysteine replacement studies revealed new details, including that a conserved lysine (Lys300) is vital for transport [21].

In 2005, the first atomic-resolution X-ray crystallographic structure of *E. coli* NhaA was solved [22]. Obtained at 3.45 Å resolution at pH 4 (below the observed active pH of the transporter), this structure, shown in Figure 1.1, revealed that the conserved aspartates 163 and 164 reside near the center of the protein, forming the likely binding site for H^+ and Na^+. However, no bound Na^+ was resolved in the structure, and this resolution is insufficient to resolve protonation states of the aspartates. Two half-helices, crossing over each other at the binding site as disordered loops, together create a balanced electrostatic environment in the low-dielectric center of the membrane. These helices, it was postulated, would undergo small rearrangements in the presence of bound ions, resulting in alternating access of the binding site from one side of the membrane to the other and facilitating transport [22].

In 2007, the group at D.E. Shaw Research published a molecular dynamics study of the NhaA structure suggesting a detailed mechanism for ion translocation [23]. In this study, the behavior of a pre-bound Na^+ ion to a protonated Asp164 was found to be dependent on the protonation state of Asp163. It also suggested that small movements in the binding site, precipitated by the protonation state of Asp163, were

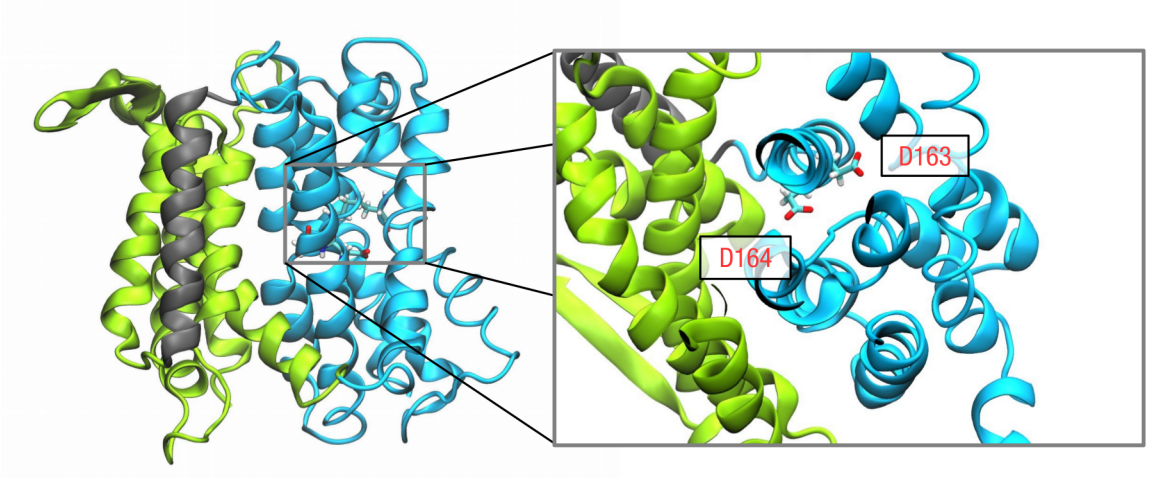


Figure 1.1: The original crystal structure of *E. coli* NhaA [22]. The structure features twelve transmembrane helices divided into two domains, the dimerization domain (green) and the core domain (cyan). The conserved aspartates Asp163 and Asp164 are at the center of core domain.

responsible for ion translocation. A full mechanism of transport for Na^+ and H^+ was proposed based on these observations.

By 2009, the prevailing hypothesis was that the two conserved aspartates were responsible for binding and transporting the two H^+ required for the transport cycle, though it remained unclear where Na^+ binds. Additional studies sought to more finely examine the behavior of the putative "pH sensor", and determine the precise mechanism responsible for the pH sensitivity of NhaA [24–26].

1.3 Recent Studies Reveal a Single Binding Site for Na^+/Li^+ and H^+ in NhaA

In 2011, a detailed electrophysiological study showed that the measured activity of NhaA in the presence of varying pH and Na^+ gradients was consistent with competitive binding; that is, both Na^+ and H^+ bind to the same location in the protein, competing at all times for the same binding site [27]. It was shown that even in the presence of a pH as low as 5, previously thought well-below the active cutoff for the protein, activity could be observed under a symmetric Na^+ gradient. This study

made clear that the concept of a distinct "pH sensor" was not necessary to explain the pH sensitivity of the protein.

In 2012, an isothermal titration calorimetry study combined with mutational analysis revealed strong evidence for the binding site residues of Li^+ in NhaA[28]. Mutation of the conserved aspartates, Asp163, Asp164, and Asp133, along with the well-conserved Thr132, were found to greatly disrupt Li^+ binding. It was also found that the conserved lysine, Lys300, was important for pH regulation, as mutation to arginine alkaline-shifted the pH profile of the transporter.

1.4 Understanding the Molecular Mechanism: Our Work

By 2013, there still remained many unanswered questions about the mechanism of NhaA, and Na^+/H^+ antiporters more broadly. The residues binding H^+ were postulated to be Asp163 and Asp164, but there existed no direct evidence for this assertion[29]. The residues coordinating the binding of Na^+ were also unknown, though evidence suggested that Asp163, Asp164, Asp133, and Thr132 were somehow important[14,28]. Finally, the conformational change, and how it was coupled to binding, was still unknown, though many hypothesized that the conformational change was almost certainly a small rearrangement of the internal half-helices that mediated alternating-access of the binding site[23,24,30].

It is from this historical context that our work proceeds. In 2013, we published our first study on a Na^+/H^+ antiporter: *Thermus thermophilus* NapA, an archaeal homolog of NhaA[31] (see Chapter 3). We presented a crystal structure of NapA in an outward-facing (periplasmic-facing) conformation, which when compared to the existing inward-facing structure of NhaA[22] suggested that both proteins might undergo large-scale ($\sim$10 Å) conformational changes as part of their transport cycle. This was followed in 2014 by our new inward-facing structure of NhaA[32], revealing that

the conserved lysine (Lys300) actually forms a salt bridge with Asp163, a feature not present in the original structure (see Chapter 4). Furthermore, molecular dynamics simulations showed that this salt bridge is broken upon spontaneous binding of Na^+ to Asp164, offering new insights into the detailed transport mechanism, and suggesting that Lys300 may directly participate in transport as a H^+ carrier. Since then, this has been further corroborated using constant-pH molecular dynamics simulations[33]. Most recently, in 2016, we obtained a new inward-facing crystal structure of NapA, showing unambiguously that the transporter undergoes a large (∼10 Å) elevator-like translocation of its core domain as part of its transport cycle[34] (see Chapter 5). The simulations performed in this study also gave a detailed view into Na^+ coordination in the binding site.

The work we have done has greatly influenced the current thinking on the mechanism of Na^+/H^+ antiporters, establishing in particular the likelihood of large, elevator-like domain movements as an expected feature for these transporters[35–37]. The discovery of the salt-bridge between Asp163 and Lys300 in NhaA has also challenged the hypothesis that the two aspartates, Asp164 and Asp163, are the proton carriers, and has shed new light on the role of the lysine, which has long been understood to be important for transport. We are continuing to build on these results, in particular working to quantify the strength of binding as a function of conformation and charge state in NapA, for which we have experimental structures in the inward- and outward-facing states (see Chapter 6).

1.5 A Note on Content

It is important to mention that the work presented in this dissertation is highly collaborative, often spanning at least three research groups across two continents. As computationalists specializing in molecular dynamics simulation, we collaborate

closely with experimentalists to validate our simulation results against measured quantities, and likewise this validation helps to corroborate the observations made in experiment. It would not have been possible to do the work described here without this kind of tight collaboration. To make clear my contributions to the work, each chapter features an introduction explaining the context and significance of the results, as well as my role in obtaining them.

A special mention should be given to the last two chapters of this work, Chapters 7 and 8. Often doing novel computational work requires the development of new tools, and these chapters detail some of the most important results of my efforts in software engineering. These efforts, too, are highly collaborative, which ensures the quality and wide applicability of the tools we build.

Chapter 2

BACKGROUND

Though our work is highly collaborative, our approaches to studying these transporters stand on their own terms as highly valuable instruments of science. The role we play in comparison to our experimentalist collaborators is close to the traditional role of the theorist, seeking to understand experimental results in the context of broader knowledge. However, this role, that of the computationalist, in truth lies somewhere in between the experimentalist and the theorist. We almost-exclusively perform calculations from theoretical models, but these calculations produce data, not clear answers, that must themselves be analyzed as one would do for experimental results. Sitting at the intersection of experiment and theory, our work functions as a bridge helping both to progress.

In line with this theme, this chapter details the theoretical model we use to understand Na^+/H^+ antiporters in the abstract (Section 2.1), the core computational method we use in an attempt to probe that model (Section 2.2), and the techniques we apply (and their limitations) to connect what we observe to experimental results (Section 2.3). There exists a vast literature on all three of these components individually, but we limit our view here to the context required to understand our work detailed in the chapters that follow.

2.1 Free Energy Transduction in Transporters

To understand fully the molecular mechanism of Na^+/H^+ antiporters, we must answer at least three questions:

1. Where and how do Na^+ and H^+ bind?

2. What does the transporter conformational transition look like, in detail?

3. How is ion binding (1), coupled to the conformational transition (2)?

These questions can often be studied separately, as they are represented at different points along the transport cycle (Figure 2.1). Practically all of our efforts are aimed at more finely defining answers to these questions, and much of the work detailed in this dissertation has been focused on answering questions (1), Chapters 4 and 6, and (2), Chapters 3 and 5. We have so far left question (3) largely unaddressed, though it has been suggested from studies of many other transporters that the disruption of interdomain salt bridges by binding may be key to this process[37].

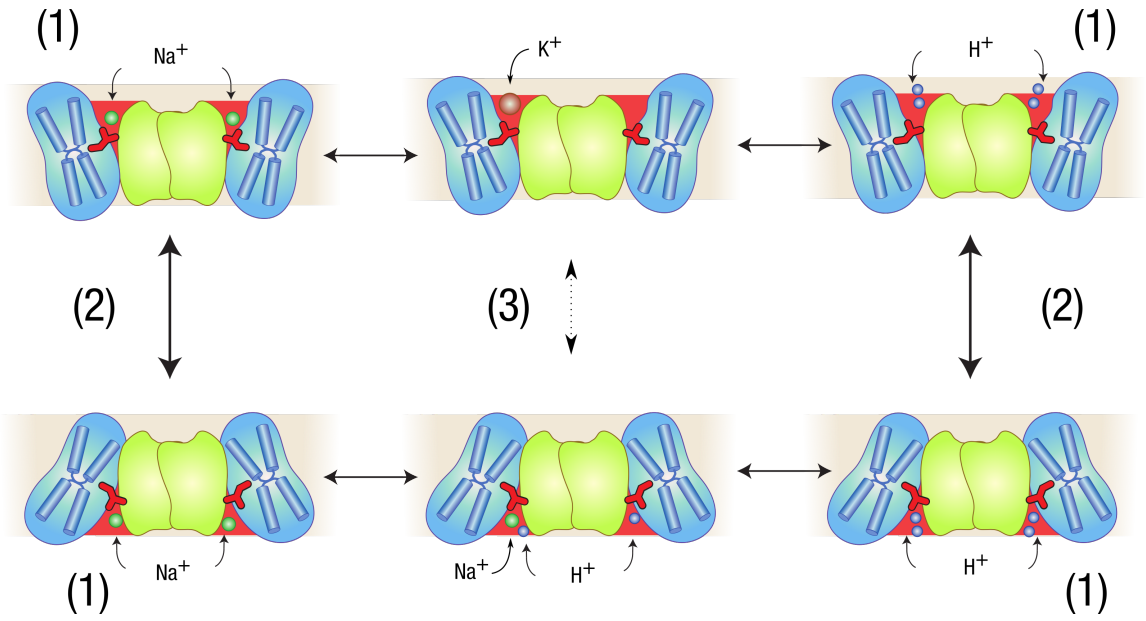


Figure 2.1: A cartoon representation of the transport cycle of NhaA/NapA. The states on the right represent the states during which $2H^+$ are bound and translocation can occur. The states on the left represent the same for $1Na^+$. The states in the middle are binding configurations that should translocate very rarely or not at all. Many of these states and transitions are relevant for understanding (1) where and how the ions bind, (2) what the conformational change looks like, and (3) how ion binding is coupled to the conformational change.

A core part of this work, summarized briefly in Section 1.4 and detailed further in Chapters 4 and 6, is the hypothesis that the conserved lysine in NhaA (Lys300) and NapA (Lys305) participates directly in the transport cycle as a H^+ carrier. The

larger theoretical context in which we can cast this hypothesis, as well as the three questions given above, is the mathematical framework for understanding free energy transduction popularized by Terrell Hill[38]. Given this hypothesis, and the transport cycle that it proposes (Figure 4.10), we can construct a model of this mechanism as a kinetic diagram (Figure 2.2).

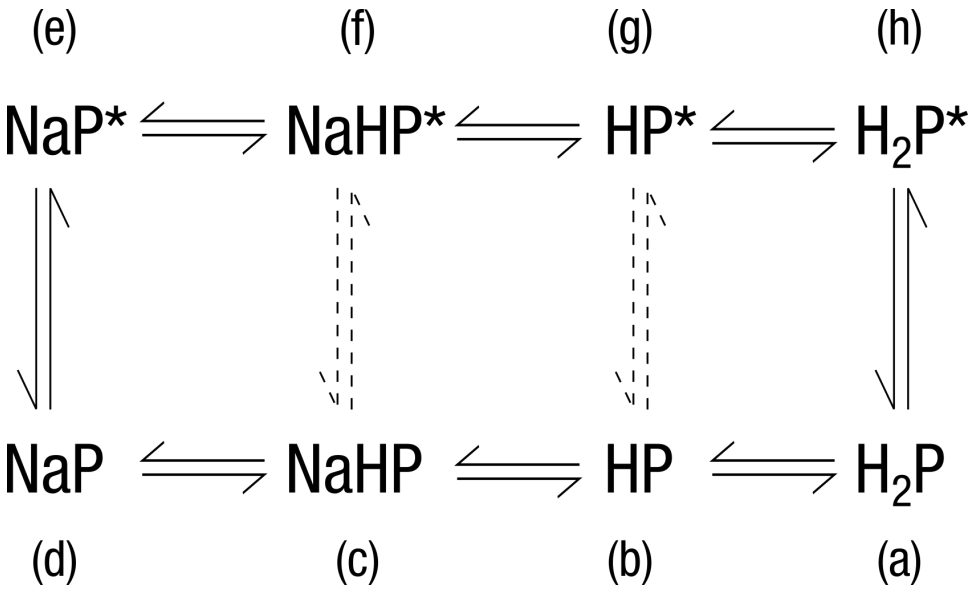


Figure 2.2: The hypothetical transport cycle in which the conserved lysine is a carrier of H^+. Based upon work presented in subsequent chapters, the hypothesis that the conserved lysine is a carrier of H^+ gives this cycle for transport. **P** is the protein in the inward-facing conformation, while **P*** is the protein in the outward-facing conformation. In this cycle, **Na** binding must occur for the second proton **H** to be released (from the lysine). Solid lines indicate rates of transition between states that should occur to a high degree, while transitions with dashed lines should occur very infrequently or not at all. Each state is labeled with a letter (a)-(h) used for reference in the text.

Compared to the general transport cycle shown in Figure 2.1, Figure 2.2 makes explicit our hypothesis for the order of events for H^+ and Na^+ binding in both the outward- and inward-facing conformations. The solid lines connecting each state indicate rates of transition between states we expect to observe with appreciable flux, the precise values of which will depend on the populations observed for each state, which further depend on the concentrations of H^+ and Na^+ on the inside and outside of the membrane and the binding affinity of the transporter for each ion in each state. The dashed lines, by contrast, indicate transitions that we expect to observe

rarely or not at all, in this case conformational changes that result in a cycle that does not preserve the $2H^+/1Na^+$ stoichiometry observed for these transporters, often amounting to wasteful leaks of the H^+ gradient with no net transport of Na^+.

The H^+ gradient is the driving force for Na^+ transport in these antiporters, and the free energy available from the translocation of H^+ across the membrane from the outside to the inside of the cell can be written in terms of its chemical potential[38,39]:

$$X_{H^+} = \mu_{i,H^+} - \mu_{o,H^+} = k_B T \ln \frac{[H^+]_i}{[H^+]_o} + F\Delta\psi \tag{2.1}$$

where F is Faraday's constant, $\Delta\psi$ is the membrane potential (~ -135 mV in *E. coli*[40]), and $[H^+]_i$ ($[H^+]_o$) is the concentration of H^+ on the inside (outside) of the membrane. Likewise, we can write the free energy available from the translocation of Na^+ across the membrane from the outside to the inside, along its physiological gradient:

$$X_{Na^+} = \mu_{i,Na^+} - \mu_{o,Na^+} = k_B T \ln \frac{[Na^+]_i}{[Na^+]_o} + F\Delta\psi \tag{2.2}$$

From these potentials we can calculate the overall driving force for the transport of $2H^+$ into the membrane in exchange for the extrusion of $1Na^+$ outward as:

$$\begin{aligned} X_C &= 2X_{H^+} - X_{Na^+} \\ &= k_B T \ln \frac{[H^+]_i^2 [Na^+]_o}{[H^+]_o^2 [Na^+]_i} + F\Delta\psi \end{aligned} \tag{2.3}$$

and we see that even with a pH gradient of 0 ($[H^+]_i = [H^+]_o$), the electrogenicity of the transport cycle (+1 inward) can still result in a driving force that can overcome the free energy cost of moving Na^+ outward against its own concentration gradient.

Either at equilibrium or at steady state (with a net driving force, X_C), each state is populated by the transporter with some probability, P_{state}. The transition rate

from a state i to j, α_{ij} is related to the (net) transition flux between these states, J_{ij} as[38]:

$$J_{ij} = \alpha_{ij} P_i - \alpha_{ji} P_j \tag{2.4}$$

These transition fluxes can be used to calculate the operational flux of each ion across the membrane[38]. Using the letters (a-h) in Figure 2.2, we see that we obtain for the inward fluxes:

$$J_{\mathrm{H}^+} = J_{ha} + J_{gb} + J_{fc} \tag{2.5}$$

$$J_{\mathrm{Na}^+} = J_{fc} + J_{ed} \tag{2.6}$$

Finally, it is important to note that the state probabilities can be used directly to calculate the free energy differences between states, moving from state i to j, as:

$$\Delta G_{ij} = -k_B T \ln \frac{P_j}{P_i} \tag{2.7}$$

In particular, we see from Figure 2.2 that the binding free energy of Na^+ in the inward-facing conformation would be calculated as:

$$\Delta G_b = -k_B T \ln \frac{P_d + P_c}{P_b + P_a} \tag{2.8}$$

It is this quantity we seek to calculate for NapA in Chapter 6.

Although experimentally obtaining rates for the operational flux containing the contributions from all cycles is possible with measurements of transport activity[38], it can be difficult to experimentally measure individual transition fluxes or free energy differences between states as arranged in Figure 2.2. However, using simulation approaches, we can (in principle) probe a transport cycle at any level of detail, with

the caveat often being whether or not we can sufficiently sample the usually high-dimensional space to obtain converged results. One of these approaches in particular, molecular dynamics simulation, allows for atomically-detailed modelling of the dynamics of macromolecules such as NapA in a hydrated lipid-membrane environment. The use of molecular dynamics, and our application of this technique toward quantifying aspects of the transport cycle discussed here, are detailed in Sections 2.2 and 2.3.

2.2 Molecular Dynamics Simulation

Molecular dynamics (MD) simulations of proteins are an attempt to model the behavior of these macromolecules to atomic detail. Making use of classical dynamics and empirically-derived forces between all the atoms in the system, an MD simulation functions as a question, asking, repeatedly: given the position of each atom, where will each atom be positioned next? When this question is asked many millions, or perhaps billions of times, we obtain a trajectory from which we can begin to derive statistics. Using statistics from multiple simulation trajectories, testing different conditions, we can construct a study. Such studies often form the answers to questions of biological interest.

This section discusses the algorithmic details of the simulations utilized in the studies we present in this document. Because our focus is on Na^+/H^+ antiporters, the content is restricted somewhat to those details that were important for simulating these types of systems.

2.2.1 Simulating membrane proteins

Na^+/H^+ antiporters are integral membrane proteins, embedded in the lipid bilayer that separates the inside of the cell from its surroundings. A cartoon rendering

of a *Thermus thermophilus* NapA dimer in an atomistic membrane, surrounded by atomistic water and free NaCl is shown in Figure 2.3. This is representative of most simulation systems employed in the studies detailed in later chapters.

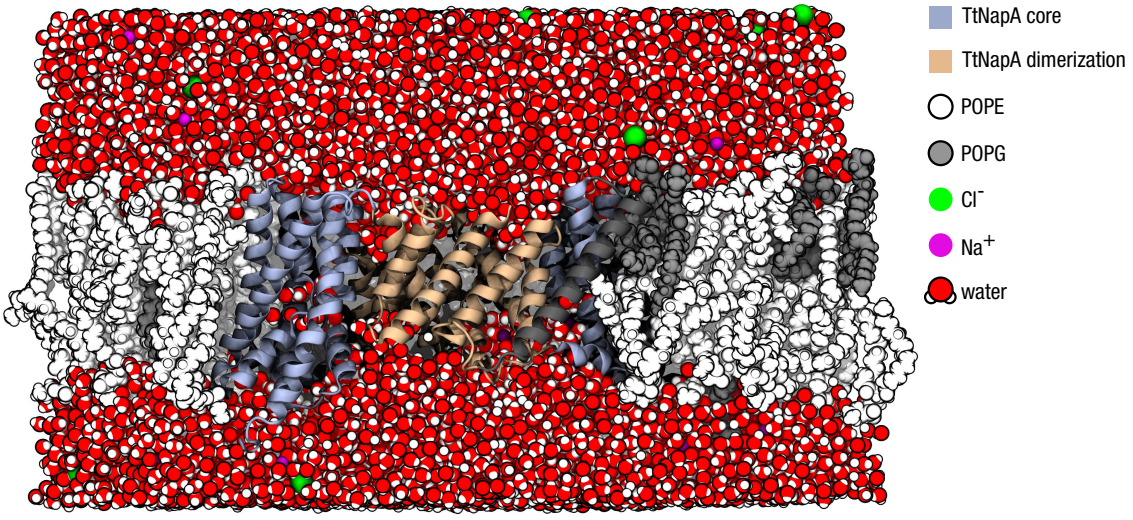


Figure 2.3: The inward-facing NapA dimer in a lipid bilayer composed of ∼4:1 POPE/POPG with free NaCl concentration of ∼250 mM. Total system size is ∼130,000 atoms.

The plasma membrane of *Escherichia coli* in which NhaA resides is composed of ∼ 75% 1-palmitoyl-2-oleoylphosphatidylethanolamine (POPE) and ∼ 20% 1-palmitoyl-2-oleoylphosphatidylglycerol (POPG)[41], and this ∼4:1 ratio of these lipid types is used for the membranes of all of our simulation systems using the CHARMM36[42–45] force field (Chapters 3, 5, and 6). Earlier simulations using the OPLS-AA force field[46–48] used simpler united-atom (non-atomistic) POPC lipid membranes[49] This explicit membrane is self-assembled around the protein using a coarse-grained-to-atomistic approach[50], embedding the protein among the lipids in a relatively unbiased fashion.

Water is modeled explicitly for simulations using the CHARMM36 force field using the CHARMM TIP3P water model[42], which is required for self-consistency of the parameters. For OPLS simulations, the TIP4P water model was used[51]. Each

simulation box featured a free NaCl concentration of ∼250 mM, simulating salt stress and helping to improve the sampling of spontaneous binding events to the protein. Protein dimer systems consisted of about 120,000—140,000 atoms in an orthorhombic simulation cell.

For most simulation systems the starting configuration of the dimer itself is often taken from an available crystal structure. This ensures that the dynamics simulated match as closely as possible the dynamics of the real protein, as relaxation of a non-native protein structure to a native configuration in a simulation may take far longer than the simulation time itself. Before production MD is performed, all structures are energy minimized using a steepest-descent algorithm, possibly followed by position-restrained MD to allow the surroundings to relax around the structure.

All simulations presented in this work were performed using GROMACS, an open-source molecular dynamics engine[52]. These simulations also employ periodic boundary conditions, in which interactions at one end of the system cross over the boundary to the other side. This avoids surface boundary effects.

2.2.2 Integrating Newton's Second Law

Molecular dynamics simulations are *classical*. Despite simulating atoms, there are no explicit quantum-mechanical calculations being performed during an MD simulation. However, the interactions of these atoms with each other are empirically parameterized from quantum-mechanical calculations, with the core idea being that the dynamics relevant to the behavior of molecules are encapsulated in these empirical parameters when integrated with Newton's Second Law ($\mathbf{F} = m\mathbf{a}$). The forms of these parameters are detailed in Sections 2.2.3 and 2.2.4.

Integration of Newton's Second Law is performed numerically using a *leap-frog* integrator[53,54], defined as:

$$\mathbf{v}(t + \frac{1}{2}\Delta t) = \mathbf{v}(t - \frac{1}{2}\Delta t) + \frac{\Delta t}{m}\mathbf{F}(t) \tag{2.9}$$

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \Delta t \mathbf{v}(t + \frac{1}{2}\Delta t) \tag{2.10}$$

This integrator uses velocities from the previous half-timestep and the forces of the current timestep to calculate the velocities of the next half-timestep. It then uses these, as well as the positions at the current time, $\mathbf{r}(t)$, to calculate the new positions of all atoms at the next timestep. The integrator is time-reversible[53], and with the use of P-LINCS for constraining bonds to hydrogen atoms[55] (or SETTLE[56] for water molecules), a timestep of 2 fs is typical.

For simulations performed in Chapter 6 for performing free energy calculations, a stochastic dynamics (Langevin) integrator[54] was used[57]. This has the form[54]:

$$\mathbf{v}' = \mathbf{v}(t - \frac{1}{2}\Delta t) + \frac{1}{m}\mathbf{F}(t)\Delta t \tag{2.11}$$

$$\Delta\mathbf{v} = -\alpha\mathbf{v}'(t + \frac{1}{2}\Delta t) + \sqrt{\frac{k_B T}{m}(1 - \alpha^2)}\mathbf{r}_i^G \tag{2.12}$$

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \left(\mathbf{v}' + \frac{1}{2}\Delta\mathbf{v}\right)\Delta t \tag{2.13}$$

$$\mathbf{v}(t + \frac{1}{2}\Delta t) = \mathbf{v}' + \Delta\mathbf{v} \tag{2.14}$$

$$\alpha = 1 - \exp^{-\gamma\Delta t} \tag{2.15}$$

where γ is the friction constant for energy dissipation, and $\mathbf{r}_i^G$ is the Gaussian noise with $\mu = 0$ and $\sigma = 1$ used to perturb the dynamics and function as a thermostat.

2.2.3 Bonded interactions

The forces between atoms that are bonded are represented by three separate types of interactions: bonds, angles, and dihedrals. The potentials of these forces are described here.

The bonded interactions between neighboring atoms in the same molecule are represented by a harmonic potential, with k_{ij}^b parameterized to match as closely as possible the force profile between the atoms observed from computational quantum-mechanical calculations. The form of these interactions between bonded atoms i and j is[54,58]:

$$V_b(r_{ij}) = \frac{1}{2}k_{ij}^b(r_{ij} - b_{ij})^2 \tag{2.16}$$

Similarly, the angle between every triplet of bonded atoms i, j, and k in a molecule is maintained by a harmonic potential on this angle[54,58]:

$$V_a(\theta_{ijk}) = \frac{1}{2}k_{ijk}^\theta(\theta_{ijk} - \theta_{ijk}^0)^2 \tag{2.17}$$

Finally, torsions around a bond, defined by four bonded atoms i, j, k, and l, are modeled with the potential[54,58]:

$$V_d(\psi_{ijkl}) = \sum_{n=0}^{N} C_n(\cos(\psi))^n \tag{2.18}$$

Torsions differ from the other bonded potentials in that they are periodic, expressed as an expansion of cosines with varying numbers of terms.

2.2.4 Non-bonded interactions

.

The forces between atoms that are not bonded are of two varieties: the Van der Waals interaction and the Coulomb interaction. The Van der Waals interaction is modeled using a Lennard-Jones[59] potential:

$$V_{\mathrm{LJ}}(\mathbf{r}_{ij}) = 4\epsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right) \tag{2.19}$$

where $\sigma_{ij} = 2^{-1/6} r_m$, with r_m is the distance of the minimum of the potential, and ϵ_{ij} determines the depth. For determining σ_{ij} and ϵ_{ij} for each pair of atom types, we often use Lorentz-Berthelot rules[54,58]:

$$\sigma_{ij} = \frac{1}{2}(\sigma_{ii} + \sigma_{jj}) \tag{2.20}$$

$$\epsilon_{ij} = (\epsilon_{ii}\epsilon_{jj})^{1/2} \tag{2.21}$$

where ϵ_{ii} and σ_{ii} are the established parameters for an atom i with another atom of the same type. Because this is a short-range interaction that decays as r^{-6} in large r, often for efficiency it is not calculated for pairs of atoms that are beyond some cutoff distance, often 1.2 nm. For this case the force is switched to zero at the cut-off to avoid artifacts.

The Coulomb interaction between two charged atoms is calculated as:

$$V_{\mathrm{C}}(\mathbf{r}_{ij}) = \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{r_{ij}} \tag{2.22}$$

Because this is a long-range interaction that decays as r^{-1}, it is important that the interaction is calculated for pairs of charged atoms that are far away. For a whole simulation system, this requires the computation of $N(N-1)/2$ interactions, which at $O(N^2)$ scaling comes at incredible cost for large systems. Because our simulation systems use periodic boundary conditions, however, the Coulomb interactions

are themselves periodic. We avoid explicitly calculating all of these interactions at long ranges beyond a cutoff using particle-mesh Ewald summation[60,61], which allows for the calculation of long-range contributions to the potential using fast-Fourier transforms.

2.2.5 Temperature and pressure coupling

To simulate a system in the NPT ensemble, it is necessary to simulate the coupling of the system to a bath that allows both heat and momentum exchange. There exists no perfect algorithm for doing this without disturbing the dynamics of the atoms in a particularly-artificial way, but the algorithms we employ do maintain the proper ensemble.

For maintaining the temperature of the system (usually $T = 310$ K), we often employ a stochastic velocity-rescaling thermostat that functions by modifying the kinetic energy distribution according to[54,62]:

$$\mathrm{d}K = (K_0 - K)\frac{\mathrm{d}t}{\tau_T} + 2\sqrt{\frac{KK_0}{N_f}}\frac{\mathrm{d}W}{\sqrt{\tau_T}} \tag{2.23}$$

where τ_T is related to the time constant of the temperature coupling, which itself corresponds to the time between rescalings in the simulation[54]:

$$\tau = 2C_V\tau_T/N_{df}k_B \tag{2.24}$$

where N_{df} is the number of degrees of freedom of the system and C_V is the system's heat capacity.

For the simulations performed in Chapter 6, the stochastic dynamics integrator (discussed above in Section 2.2.2) functions as a thermostat.

For maintaining the desired system pressure, we utilize a Parinello-Rahman barostat[63], which includes an equation of motion for the vectors of the simulation box itself, $\mathbf{b}$, as[54]:

$$\frac{\mathrm{d}^2\mathbf{b}}{\mathrm{d}t^2} = V\mathbf{W}^{-1}\mathbf{b}'^{-1}(\mathbf{P} - \mathbf{P}_{\mathrm{ref}}) \tag{2.25}$$

where $\mathbf{P}$ is the current pressure, $\mathbf{P}_{\mathrm{ref}}$ is the reference pressure (usually chosen as 1.0 bar), $\mathbf{W}$ is the strength of the coupling, and V is the current box volume. This barostat also changes the equations of motion for the particles by adding an additional term:

$$\frac{\mathrm{d}^2\mathbf{r}_i}{\mathrm{d}t^2} = \frac{\mathbf{F}_i}{m_i} - \mathbf{M}\frac{\mathrm{d}\mathbf{r}_i}{\mathrm{d}t} \tag{2.26}$$

$$\mathbf{M} = \mathbf{b}^{-1}\left[\mathbf{b}\frac{\mathrm{d}\mathbf{b}'}{\mathrm{d}t} + \frac{\mathrm{d}\mathbf{b}}{\mathrm{d}t}\mathbf{b}'\right]\mathbf{b}'^{-1} \tag{2.27}$$

For membrane protein systems, we use a semi-isotropic barostat in which the scaling in the x and y directions are linked, but the z direction is treated independent of these. This ensures that the membrane, situated in the xy-plane, is treated isotropically in these two dimensions.

2.3 Connecting MD to Experiment

It is a fundamental tenet of science that theory, and likewise the results of simulations built on theory, must be validated against experimental measurements. We therefore do our best to connect what we observe in simulations of proteins to what has been measured experimentally for them, as this serves to validate observations made in simulations that are simply not accessible to experiment. This is often a challenge, however, due to the difficulty of sampling the timescales on which many proteins function, which for Na^+/H^+ antiporters extends from hundreds of microseconds to

milliseconds. By contrast, state of the art simulation approaches for fully-atomistic membrane-transporter systems using commodity hardware in large clusters, and even national high-performance supercomputing resources, can currently achieve sampling in the tens of microseconds. It is simply not possible, presently, to examine a full transport cycle such as that in Figure 2.2 directly with brute-force simulations.

However, it is possible to take a targeted approach, performing simulations aimed at quantifying specific portions of a transport cycle. As an example, we can choose to focus on the question of Na^+ binding, as observed in states (c), (d), (e), and (f) of Figure 2.2. Performing equilibrium simulations of the NapA dimer in both outward- and inward-facing conformations, we can directly observe spontaneous binding of Na^+ to Asp157 (Figure 2.4). In principle, we could use these binding events to estimate the rates α_{bc}, α_{cb} and α_{gf}, α_{fg}. But using this brute-force approach, even as focused as it is on the question of Na^+ binding, we *still* are not able to sufficiently sample binding and unbinding events to estimate these rates.

Then, what can we do? We can take an even more targeted approach, choosing to perform simulations meant only to calculate a specific quantity from this cycle instead of equilibrium simulations that happen to observe some of the events of interest in small numbers. This is the approach taken in Chapter 6, in which we perform hundreds of simulations, most of them sampling an unphysical state of the system, to obtain converged values for the binding free energy, ΔG_b, of Na^+ in the states (c), (d), (e), and (f) in Figure 2.2. These free energies can be used to calculate the apparent binding affinity, often measured as its inverse, K_d, we would expect to see given the proposed transport mechanism. These can be compared directly to experimental results, which validates not only the simulations themselves but also the hypothesis they were performed to test.

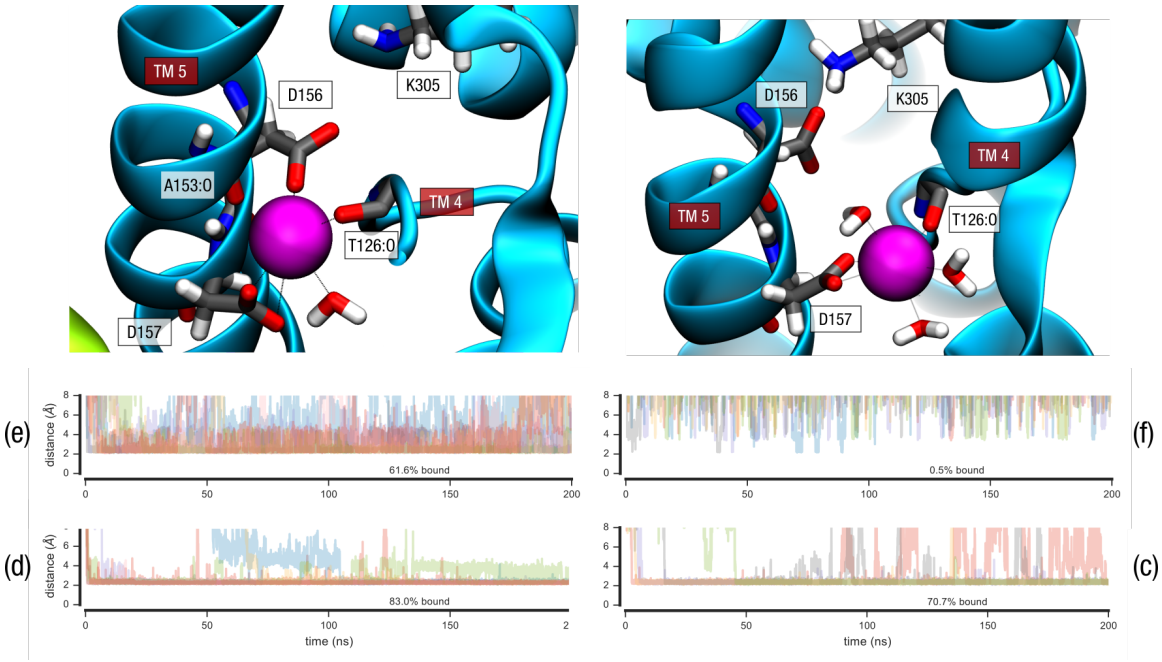


Figure 2.4: Spontaneous binding behavior of Na^+ to NapA for four states of the transport cycle For the states (c), (d), (e), and (f) in Figure 2.2, we show the binding behavior of Na^+ from equilibrium simulations of NapA, with timeseries showing the distance of the nearest Na^+ to the nearest carboxyl oxygen of either D157 or D156 from multiple simulations. Although there appear to be many binding events for each state, there are not enough to predict well-converged binding rates. The rendered images show how the ion is generally coordinated at the binding site in simulations of the protonation states for K305 represented here, with the states on the left representing the deprotonated (uncharged) state, and those on the right representing the protonated (charged) state.

Chapter 3

A TWO-DOMAIN ELEVATOR MECHANISM FOR SODIUM/PROTON ANTIPORT

This chapter is a reprint of the journal article, Lee, C., Kang, H.J., von Ballmoos, C., Newstead, S., Uzdavinys, P., **Dotson, D.L.**, Iwata, S., Beckstein, O., Cameron, A.D., and Drew, D. (2013). A two-domain elevator mechanism for sodium/proton antiport. Nature 501, 573-577. This work revealed the first crystal structure for *Thermus thermophilus* NapA, a structural homolog of *Escherichia coli* NhaA, in the outward-facing conformation. It was in this work that we first proposed a large-scale, elevator-like mechanism of ion transport for this class of proteins, in which the core domain translates and rotates by up to 10 Å during the transport cycle to allow alternating access to the binding site.

My contribution to this work was the performance and analysis of the molecular dynamics simulations.

ABSTRACT

Sodium/proton (Na^+/H^+1) antiporters, located at the plasma membrane in every cell, are vital for cell homeostasis[29]. In humans, their dysfunction has been linked to diseases, such as hypertension, heart failure and epilepsy, and they are well-established drug targets[4]. The best understood model system for Na^+/H^+ antiport is NhaA from *Escherichia coli*[12,29], for which both electron microscopy and crystal structures are available[17,18,22]. NhaA is made up of two distinct domains: a core domain and a

dimerization domain. In the NhaA crystal structure a cavity is located between the two domains, providing access to the ion-binding site from the inward-facing surface of the protein[22,29]. Like many Na^+/H^+ antiporters, the activity of NhaA is regulated by pH, only becoming active above pH 6.5, at which point a conformational change is thought to occur[64]. The only reported NhaA crystal structure so far is of the low pH inactivated form[22]. Here we describe the active-state structure of a Na^+/H^+ antiporter, NapA from *Thermus thermophilus*, at 3 Å resolution, solved from crystals grown at pH 7.8. In the NapA structure, the core and dimerization domains are in different positions to those seen in NhaA, and a negatively charged cavity has now opened to the outside. The extracellular cavity allows access to a strictly conserved aspartate residue thought to coordinate ion binding[23,28,29] directly, a role supported here by molecular dynamics simulations. To alternate access to this ion-binding site, however, requires a surprisingly large rotation of the core domain, some 20° against the dimerization interface. We conclude that despite their fast transport rates of up to 1,500 ions per second[12], Na^+/H^+ antiporters operate by a two-domain rocking bundle model, revealing themes relevant to secondary-active transporters in general.

3.1 Introduction

Na^+/H^+ antiporters are secondary active transporters that are conserved across all biological kingdoms to maintain the internal pH, cell volume and sodium concentration of the cell[29], a mechanism first proposed by West and Mitchell[8]. Na^+/H^+ antiporters are members of the large monovalent cation proton antiporter (CPA) superfamily that includes, among others, the CPA1 and CPA2 clades[4]. It is generally thought that Na^+/H^+ antiporters from the CPA1 clade catalyse electroneutral sodium-proton exchange (SLC9A1–9 in mammals (also known as NHE1-9)), whereas CPA2 members are thought to be electrogenic (SLC9B1-2 in mammals (also known

as NHA1-2))[4], with stoichiometries of $2H^+$:$1Na^+$ and $3H^+$:$2Na^+$ ions reported[13,65]. Like all secondary active transporters, Na^+/H^+ antiporters are thought to operate by an alternating access mechanism; however, the different conformational states of the transport cycle have yet to be determined. The inward-facing structure of the well-characterized bacterial CPA2 protein *E. coli* NhaA is the only representative crystal structure[22].

Using fluorescent-based methods[66,67], we screened members of the CPA2 clade[4] for their suitability for structural studies. NapA from *T. thermophilus*, which has 21% sequence identity to human NHA2, was thus identified (see Methods and Supplementary Information Figures 3.5a and 3.6). Although the overall sequence homology to the better characterized *E. coli* NhaA is low, <15% identity, residues identified to be important for transport in NhaA and mammalian homologues are nonetheless well conserved (Figure 3.6). Using isolated inside-out membrane vesicles it was previously shown that *T. thermophilus* NapA is active above pH 6, with maximum activity for sodium at pH 8[68], similar to *E. coli* NhaA[69]. To measure activity in an isolated system, we co-reconstituted purified NapA and *E. coli* F1F0 ATP synthase into liposomes (see Methods). After establishment of a pH gradient by the addition of ATP, proton efflux was monitored in response to Na^+ or Li^+ addition (Figure 3.1a, 3.1b). In this experimental set-up, the apparent Michaelis constant (Km) values for Na^+ or Li^+ were determined to be 4.0 ± 0.3 (mean $\pm$ s.d.) or 0.41 ± 0.04 mM, respectively, similar to the affinities from inside-out vesicles[68] (Figure 3.1c). In experiments with NapA proteoliposomes trapped with a water-soluble pH sensitive dye, dissipation of the membrane potential ($\Delta\psi$) stimulated exchange activity in the presence of a Na^+ gradient, confirming the electrogenic nature of NapA (Fig. 3.1d, bottom). Furthermore, in the absence of a H^+ or Na^+ gradient, NapA transport activity was solely driven by $\Delta\psi$ (Figure 3.1d, top). Taken together, *T. thermophilus* NapA has a similar

antiport profile to *E. coli* NhaA, and consequently functional and structural studies of the two proteins can complement one another.

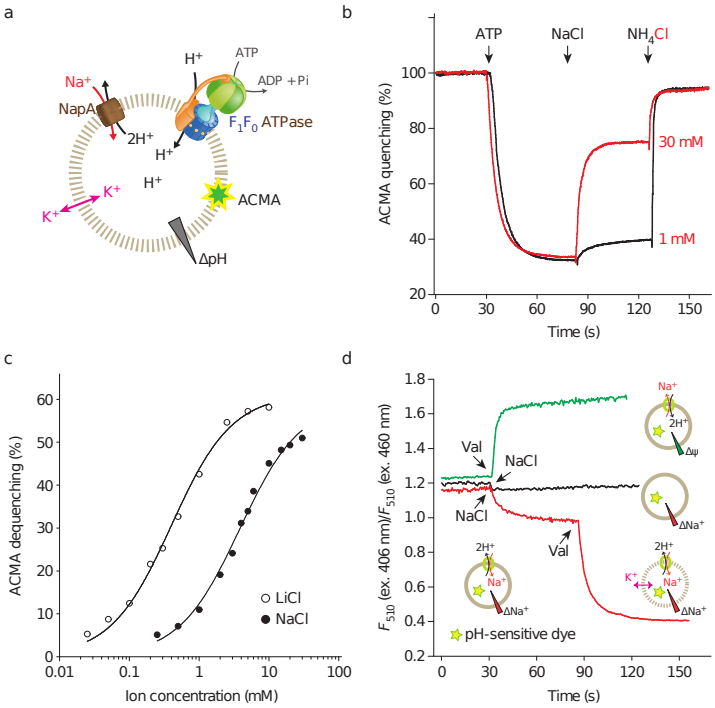


Figure 3.1: NapA Na^+/H^+ transport activity is electrogenic. **a**, Experimental setup for determination of Na^+/Li^+ affinity and electrogenicity (the most likely stoichiometry is shown). The ATP synthase and NapA are co-reconstituted in liposomes. Free K1 diffusion by valinomycin suppresses the effect of $\Delta\psi$. ACMA, 9-amino-6-chloro-2-methoxyacridine. **b**, Representative ACMA fluorescence traces for Na^+/H^+ antiporter activity. ATP-driven H^+ pumping establishes a ΔpH (acidic inside) as monitored by quenching of fluorescence. H^+ efflux is initiated by the addition of NaCl/LiCl, and further NH_4Cl addition collapses the proton gradient. **c**, Apparent binding affinity for Na^+ (filled circles) and Li^+ (open circles) in NapA (pH 7.5). **d**, Na^+/H^+ antiporter activity monitored with liposome entrapped pH sensitive fluorophore pyranine. The fluorescence was recorded at 510 nm (excitation 406 and 460 nm) and the ratio of the two values was plotted against time. Na^+-gradient-(,120 mV)-driven Na^+/H^+ exchange was initiated by the addition of NaCl (red). Addition of valinomycin at 90 s leads to further H^+ efflux; this releases the inhibitory membrane potential established during the first transport phase. No Na^+-gradient-driven transport was observed in liposomes without NapA (black). $\Delta\psi$-driven Na^+/H^+-exchange transport was initiated by establishing a K^+/valinomycin diffusion potential (green). No activity was observed in liposomes without NapA (data not shown). ex., excitation; *F*, fluorescence.

3.2 Results

The structure of NapA at pH 7.8 was solved by multiple isomorphous replacement with anomalous scattering in combination with multi-crystal averaging. The highest resolution data correspond to a triple mutant of NapA, in which three cysteine

residues, which have no apparent effect on functional activity, were introduced to facilitate phasing (see Methods, Supplementary Tables 3.1 and 3.2 and Supplementary Figure 3.5b). The structure was refined at a resolution of 3 Å , R_{factor} value of 22.3% and R_{free} value of 24.8% (Supplementary Table 3.1, Supplementary Figure 3.7 and Methods). Electron density maps at 3.7 Å of wild-type NapA at pH 9.0 show no clear structural differences between the wild-type and the triple cysteine mutant.

NapA is built from 13 transmembrane (TM) helices with an $\mathrm{N}_{\text{out}} - C_{\text{in}}$ topology (Figure 3.2 and Supplementary Figure 3.8). Relative to NhaA, it has an additional helix at the amino terminus. To facilitate comparison to NhaA, we refer to the first helix as TM-1 (Figure 3.2a and Supplementary Figure 3.8). TM-1 to TM5 and TM7 to TM12 are topologically similar but oppositely orientated in the plane of the membrane (Supplementary Figure 3.4). These six-transmembrane-helice topology inverted repeats intertwine to form a core (ion-translocation) and dimerization (interface) domain, and are linked together by TM6 (Figure 3.2b and Supplementary Figure 3.4). The NapA and NhaA[22] structures are very similar (Figure 3.2b). As there is a change in the position of the core relative to the dimerization domain, however, the similarity is best seen when the two domains are superposed separately; root mean squared deviation (r.m.s.d.) of 1.8 Å for 134 out of 148 pairs of C_{α} atoms of the core domain transmembrane helices, and 1.9 Å for 62 out of 88 of the dimerization domain transmembranes (Supplementary Figure 3.9a, 3.9b). The change in position of the core relative to the dimerization domain, which may reflect the difference in the pH at which the transporters were crystallized, gives rise to a large negatively charged cavity that is now open to the outside in contrast to the inward-facing funnel seen in NhaA (Figures 3.2b and 3.3b). The interaction between TM2 and TM4 and TM5 tightly closes the cytoplasmic side of the cavity.

NapA, like other bacterial and mammalian Na^+/H^+ antiporters[18,70,71] purifies as a dimer and is clearly dimeric in the structure with an extensive interface burying a surface area of 1,800 $Å^2$ (Figure 3.3 and Supplementary Figure 3.5c). In molecular dynamics simulations, the dimer sits entirely within a model membrane bilayer (Figure 3.3c). The backbone r.m.s.d. from the starting conformation of each monomer increased to about 2.0 Å over the time course of the simulation (Supplementary Figure 3.10), indicating a slow relaxation of the crystal structure in the native-like membrane environment. The dimer has a crystallographic two-fold operator approximately parallel to the transmembrane helices, and it is largely made up of tight hydrophobic helix-helix packing between TM21 on one monomer and TM7 on the other. There are also contacts between the ends of TM2 and TM9 (Figure 3.3a, 3.3b). NapA lacks the β-hairpin domain that makes most of the contact between protomers at the extracellular membrane surface in NhaA[22,24]. The dimer interface in NapA more closely resembles the dimer interface modelled in the 7 Å electron crystallography structure of NhaP1[72], a CPA1 member from *Methanocaldococcus jannaschii* (formerly *Methanococcus jannaschii*). NhaP1 also has 13 transmembrane helices built up from two six-transmembrane topology repeat units.

The substrate-binding cavity is open to the extracellular side. The cavity begins above the dimerization interface and funnels between the dimerization and core domains, ending in the middle of the membrane (Figure 3.3). It is considerably more open than the inward-facing cavity of NhaA[22] and is negatively charged, being lined with glutamate residues. Although there is a similar distribution of charged residues in NhaA and NapA, few, including those that have been predicted to be involved in pH sensing[29], are conserved between the two proteins (Supplementary Figure 3.11). Near the base of the cavity are two highly conserved aspartates, Asp 156 and Asp 157, located on TM5 (Supplementary Figures 3.6 and 3.11). The residues equiva-

lent to these aspartates in NhaA, Asp 163 and Asp 164 (Supplementary Figures 3.6 and 3.11), probably coordinate ion binding based on their position, conservation with mammalian Na^+/H^+ antiporters[4], phenotypes of mutants[29,73], isothermal titration calorimetry (ITC) experiments[28] and molecular dynamics simulations[23]. In NapA, mutation of either residue to alanine[68] or asparagine abolishes transport activity completely (Supplementary Figure 3.5d and Supplementary Table 3.2). With NapA crystals grown at pH 7.8, the aspartate residues are likely to be deprotonated. Consistent with this, Asp 157 is orientated towards the centre of the cavity rather than hydrogen bonding with the backbone of TM4, as seen for Asp 164 in NhaA, an interaction that requires Asp 164 to be protonated (Supplementary Figure 3.9c). Using solid-state membrane electrophysiology it was previously shown that NhaA transports cations equally well in either direction and the transport activity profile fits a simple H^+ versus Na^+ kinetic binding model to a single common site[27]. To investigate whether Na^+ ions would bind as expected, we carried out equilibrium molecular dynamics simulations of outward-facing NapA in a model membrane bilayer. Simulations were carried out with both Asp 157 and Asp 156 deprotonated, as they are likely to be in the crystal structure, and also as combinations of their neutral and charged forms. With both aspartates charged, Na^+ ions spontaneously entered the negatively charged extracellular cavity to bind to Asp 157 (Figure 3.3c and Supplementary Figure 3.12). Na^+ ions were concentrated at Asp 157 (Supplementary Figure 3.12) and multiple distinct binding and unbinding events could be observed, which is qualitatively consistent with weak binding. By contrast, Na^+ ion binding was not observed when Asp 157 was protonated and was markedly reduced when Asp 156 was neutral (Supplementary Figure 3.12). In molecular dynamics simulations of inward-facing NhaA, Na^+ binds to the equivalent aspartate (Asp 164)[23], which is positioned at the base of the cytoplasmic cavity.

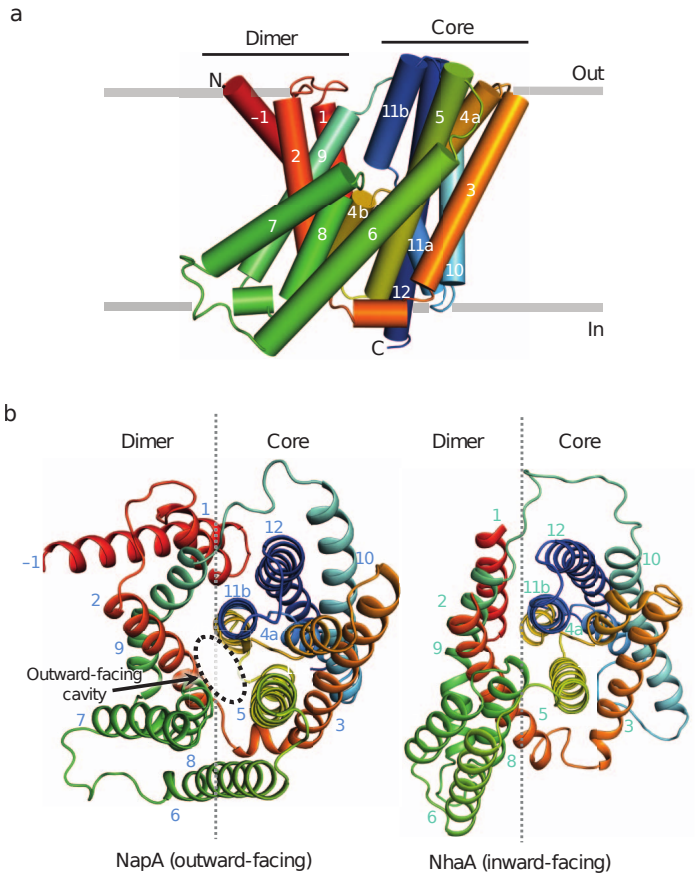


Figure 3.2: Outward-facing NapA structure. **a**, Cartoon representation of NapA as viewed in the plane of the membrane. Transmembrane helices -1 to 12 have been coloured from red at the amino terminus to blue at the carboxy terminus, with the position of the membrane depicted in grey. **b**, Outward-facing NapA structure (left) and inward-facing NhaA structure (right) as viewed from the extracellular side with the removal of the β-hairpin domain located between transmembrane helices 1 and 2 from NhaA to facilitate visual comparison. The dimerization and core domain boundaries are represented by a grey line.

Asp 156 and Asp 157 on TM5 are located next to the point at which the two antiparallel discontinuous helices, TM4a-b and TM11a-b, cross over in the core domain (Supplementary Figure 3.9c). Although discontinuous helices are a common feature for ion-binding in transporters[74], the antiparallel crossing over of transmembrane helices is a unique feature of the NhaA fold[22,75]. It results in the positive dipole ends of TM4a and TM11a facing one another and likewise the negative dipole ends of TM4b and TM11b. In NhaA, the dipoles are proposed to be neutralized by the side

chains of an aspartate (Asp 133; TM4a-b) and a lysine (Lys 300; TM10), respectively (Supplementary Figure 3.9c). The lysine (Lys 305) is retained in NapA but the aspartate is replaced by a serine, as it is in human NHA2 (Supplementary Figures 3.9c and 3.11). From the structure, however, it appears the side chain of Glu 333 from TM11b can take a similar position to the carboxylate of Asp 133 (Supplementary Figures 3.9c and 3.11). The positions of these two residues are pseudo-symmetrically related; the functional significance of this swap is unclear. Although the mutation of Glu 333 to alanine affected the apparent affinity for Li^+ only slightly (<3-fold), the affinity for Na^+ was severely decreased (>15-fold), which is in agreement with the results obtained from an Asp 133 to alanine mutation in the NhaA protein[15] (Supplementary Table 3.2). By contrast, the mutation of the highly conserved Lys 305 to alanine severely decreased both Na^+ and Li^+ affinity (>20-fold; Supplementary Table 3.2). It was recently speculated that Lys 300 in NhaA could have more than a stabilizing role and that it could also form part of the pH activation mechanism, as mutation to arginine changed the pH at which NhaA becomes active[28]. In the NapA structure, Lys 305 forms a salt bridge with Asp 156 (Supplementary Figures 3.9c and 3.12a). Notably, we have recently observed a similar interaction in a different crystal form of NhaA at low pH (O.B. et al., manuscript in preparation). We would expect, however, that after cation binding to Asp 156 and Asp 157 the salt bridge is disrupted. As such, the interaction observed here supports the role of lysine in pH activation.

Previously, it was proposed that the exceptionally fast transport of Na^+/H^+ antiporters would primarily involve local rearrangements of the finely electrostatically balanced discontinuous helices in the core domain[22,29]. However, the major structural difference between NapA and NhaA is in the position of the core domain in relation to the dimerization domain (Supplementary Figures 3.9a, 3.9b). With reference to the dimerization interface, which seems likely to remain stable during trans-

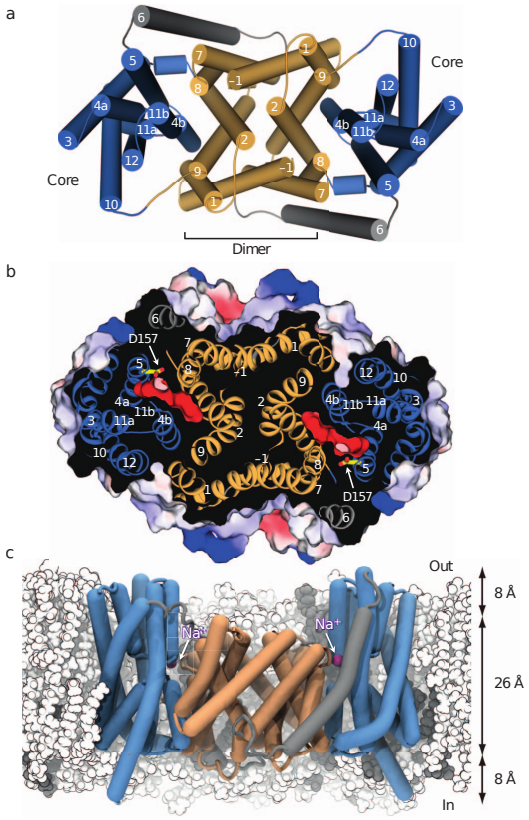


Figure 3.3: Structure of NapA dimer and proposed Na^+/(Li^+)-binding site. **a**, Cartoon representation of the NapA dimer structure, with transmembrane helices for one monomer labelled in cyan and the other in yellow. The transmembrane helices of the dimerization domains are coloured in pale orange, and the core ion-translocation domains are in sky blue, connected together by TM6 shown in grey. **b**, Electrostatic surface representation showing the location of the negatively charged extracellular cavity as a section through the protein from above (same orientation as in **a**). The proposed ion-binding aspartate, Asp 157, is illustrated as a stick model. **c**, In molecular dynamics simulations, the NapA dimer is stable in a 4 POPE (white):1 POPG (grey) lipid membrane (POPE, 1-palmitoyl-2-oleoylphosphatidylethanolamine; POPG, 1-palmitoyl-2-oleoylphosphatidylglycerol). Sodium ions (magenta spheres) spontaneously bind to Asp 157 at the bottom of the outward-facing cavity. The approximate thickness of the hydrophobic core and headgroup regions of the membrane are indicated.

port across the membrane[24,76], the core domain in the NhaA structure is rotated by 21° relative to the core of NapA (Figures 3.4a, 3.4b). This large rotation of the core domain closes the cavity seen on the outside of NapA and opens the cytoplasmic funnel on the inside, as observed in NhaA (Supplementary Videos 1 and 2). During this process, the two cation binding aspartates, which are in line with the tip of TM8 in the dimerization domain of the outward-facing NapA structure, are shifted 10 Å towards the cytoplasmic surface of the transporter (Figure 3.4a). This elevator

movement of a substrate-binding domain, in this case to carry Na^+ (Li^+) ions from one side of the membrane to the other in exchange for protons, resembles that of the transport mechanism seen in the glutamate transporter GltPh[77]. A two-domain transport mechanism was also recently predicted in NhaA, on the basis of a consideration of the two symmetry-related inverted repeats as well as elastic network models and biochemical cross-linking[78]. Such a mechanism was also proposed for the bile acid sodium symporter $ASBT_{NM}$[75], which is a structural homologue of NhaA and NapA.

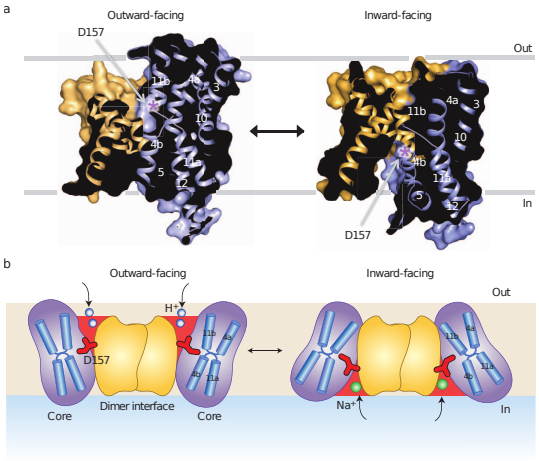


Figure 3.4: Alternating access model of sodium-proton antiport. **a**, Surface representation showing a section through the outward-facing NapA structure (left) and inward-facing NapA model (right) (see Methods and Supplementary Videos 1 and 2). The position of Asp 157 is denoted by an asterisk, and the helices have been coloured as in Figure 3.3. For the sake of clarity, only one molecule is shown. **b**, Schematic of the proposed transport mechanism that illustrates the conformational changes with the core moving against the dimerization domain. Protons (shown in blue) bind to the core domain in the outward-facing state (left) causing it to switch to the inward-facing state (right). On the inside, protons are exchanged for sodium (green) and the core domain moves back to the outside. Asp 157 (shown in red), which is crucial for binding both one of the protons and the sodium ion, moves approximately 10 Å during this process. Other residues involved in ion binding are not shown.

3.3 Conclusions

In summary, the structure of NapA is consistent with a single ion-translocation site mechanism, of which the strictly conserved Asp 157 (Asp 164 in NhaA) is ideally positioned for binding ions in outward- and inward-facing states (Figure 3.4b). To provide

alternating access to the aspartate, however, requires a surprisingly large movement of the core domain that twists around the dimerization interface, in essence creating a two-domain rocking bundle model. Although further structures are needed to clarify how ion binding and release is coupled to these global changes, the NapA structure provides a clear example that the extent of the conformational change may not necessarily correlate with the rate of transport or the size of the substrate transported, as previously assumed. Thus, the NapA structure also reveals fresh mechanistic insights relevant to all types of ion-coupled transporters.

3.4 Methods

NapA was cloned into a cleavable green fluorescent protein (GFP)-His_8 fusion vector pWaldoGFPe[66]. The fusion protein was expressed in *E. coli*, solubilized in 1% dodecyl-b-D-maltopyranoside (DDM), and purified to homogeneity in either DDM or 1% nonyl-b-D-maltopyranoside. The NapA proteoliposome uptake assay was modified as previously described[12]. Crystals were grown at either pH 7.8 or 9.0 by the vapour diffusion method. Data were collected on beamlines I02 and I03 at the Diamond Light Source or ID 23-1 and 23-2 beamlines at the European Synchrotron Radiation Facility. The protein was derivatized before crystallization by incubation with 2.5 mM mercury acetate. The NapA structure was solved by multiple isomorphous replacement with anomalous scattering in combination with multi-crystal averaging and refined at a resolution of 3 Å. Molecular dynamics simulations were performed as described in Section 3.4.7.

3.4.1 NapA sequence

Thermus thermophilus NapA sequence (Uniprot accession number Q72IM4); residues progressively substituted to cysteine are underlined, and additional C-terminal residues retained after TEV cleavage are shown in bold (see next section for cloning details).

MHGAEHLLEIFYLLLAAQVMAFIFKRLNQPVVIGEVLAGVLVGPALLGLVHEGEILEFLA
ELGAVFLLFMVGLETRLKDILAVGKEAFLVAVLGVALPFLGGYLYGLEIGFETLPALFLG
TALVATSVGITARVLQELGVLSRPYSRIILGAAVIDDVLGLIVLAVVNGVAETGQVEVGA
ITRLIVLSVVFVGLAVFLSTLIARLPLERLPVGSPLGFALALGVGMAALAASIGLAPIVG
AFLGGMLLSEVREKYRLEEPIFAIESFLAPIFFAMVGVRLELSALASPVVLVAGTVVTVI
AILGKVLGGFLGALTQGVRSALTVGVGMAPRGEVGLIVAALGLKAGAVNEEEYAIVLFMV
VFTTLFAPFALKPLIAWTERERAAKE**GSENLYFQ**

3.4.2 Expression screening, mutagenesis, protein purification and characterization

NhaA homologues were cloned as green fluorescent protein (GFP)-His_8 fusions into the vector pWaldoGFPe[66], as fluorescence from the C-terminal GFP fusion is a reliable reporter of membrane-integrated expression[79]. The monodispersity of expressed fusions were screened in a number of different detergents by fluorescence-detection size-exclusion chromatography[80]. NapA from *T. thermophilus* was selected as a suitable candidate showing stability in a wide range of detergents including the harsh detergent *N*-dodecyl-*N*,*N*-dimethylamine-*N*-oxide[67]. Expression levels of the protein were initially low in standard culture conditions, but improved significantly using MemStar, which is a new strategy for boosting expression of membrane proteins in *E. coli* (C.L. et al., manuscript in preparation). In brief, Lemo21(DE3) cells[81] were grown at 37°C in PASM-5052 media[82], with and without selenomethionine

incorporation, and induced with 0.4 mM isopropyl-β-D-thiogalactoside (IPTG) at an absorbance (A_{600nm}) of 0.5 for overnight incubation at 25°C.

Wild-type NapA and mutants generated by Quickchange protocol (Agilent Technologies) were purified essentially as previously described[83]. Membranes were isolated from 5-l *E. coli* cultures and solubilized in 1% dodecyl-β-D-maltopyranoside (DDM; Generon) for 2 h in buffer containing 1×PBS, 150 mM NaCl and 10 mM imidazole. The suspension was cleared by ultracentrifugation at 120,000g for 1 h. The sample was mixed with 1 ml of Ni-NTA Superflow resin (Qiagen) per 1 mg of GFP-His_8 and incubated for 2 h at 4°C. Slurry was loaded onto a glass Econo-Column (Bio-Rad) and washed in 1×PBS buffer containing 0.1% DDM, 150 mM NaCl and 20 mM imidazole for 20 column volumes. Bound material was washed for a further 20 column volumes in the same buffer containing 50 mM imidazole. The NapA-GFP-His_8 fusion was eluted in two column volumes of 1×PBS buffer containing 0.6% nonyl-β-D-maltopyranoside (NM; Generon), 150 mM NaCl and 250 mM imidazole. The eluted protein was dialysed overnight in the presence of stoichiometric amounts of His_6-tagged tobacco etch virus protease in 1.5 l of buffer containing 20 mM Tris-HCl, pH 7.5, 150 mM NaCl and 0.5% NM. Dialysed sample was passed through a 5 ml Ni-NTA His-Trap column (GE Healthcare), and the flow-through containing NapA was collected. Protein was concentrated to 10 mg ml^{-1} using concentrators with a relative molecular mass cut-off of 100 kilodaltons (kDa), and was loaded onto a Superdex 200 10/300 gel filtration column (GE Healthcare) equilibrated in 20 mM Tris-HCl, pH 7.5, 150 mM NaCl and 0.45% NM (Anatrace). The protein peak was collected and concentrated to 10 mg ml^{-1} for crystallization.

Purified NapA was loaded onto a Superdex 200 10/300 size exclusion column (GE Healthcare) coupled to a Viscotek TDAmax tetra detector array (Malvern) with GPCmax solvent pump and integrated auto-sampler, using the OmniSEC software

for data analysis. The SEC-UV/LS/refractive index system was equilibrated in 20 mM Tris-HCl, pH 7.5, 150 mM NaCl and 0.03% DDM at a flow rate of 0.3 ml min^{-1}. Standard gel filtration molecular weight markers were used for calibration and all proteins were analysed under the same experimental conditions. Data were collected from the refractive index, right angle LS (RALS) and UV_{280nm} detectors. The oligomeric state of NapA from the NapA-detergent micelle was calculated using methods described previously[84].

3.4.3 NapA transport activity

In this set-up (Figure 3.1a), the described Na^+/H^+ antiport with inverted membrane vesicles was mimicked[12,68]. For this, we co-reconstituted *E. coli* F_1F_0 ATP synthase and *T. thermophilus* NapA in the same vesicles. Although ATP synthase under these conditions has been shown to orient unidirectionally[85], NapA most probably has a heterogeneous orientation, which might affect apparent K_m values. Furthermore, a mixture of liposomes containing none, one or two enzymes is expected. Only the liposomes containing ATP synthase will lead to initial ACMA quenching, but only those containing both ATP synthase and NapA will lead to dequenching after addition of the coupling ion. Furthermore, it is expected that a new equilibrium is reached after every addition, which might influence the extent of dequenching. This is demonstrated by the addition of NH_4Cl at the end of every measurement, in which any remaining ΔpH is dissipated. In brief, purified NapA wild-type and mutants were co-reconstituted with purified ATP synthase from *E. coli* with an ~2:1 molar ratio (NapA:ATP synthase) in MME buffer (10 mM MOPS-NaOH, pH 7.5, 2.5 mM $MgCl_2$, 100 mM KCl) as described for ATP synthase[85,86]. Typically, 50 μl proteoliposomes were diluted into 1.5 ml MME buffer containing 3 nM ACMA and 140 nM valinomycin. Fluorescence was monitored at 480 nm using an excitation wavelength

of 410 nm in a fluorescence spectrophotometer (Cary Eclipse, Agilent Technologies). An outward-directed pH gradient (acidic inside) was established by the addition of 2 mM ATP, as followed by a change in ACMA fluorescence. After ∼2 min equilibration, the activity of NapA wild-type and mutants thereof was assessed by the dequenching of ACMA fluorescence after addition of the indicated concentrations of NaCl or LiCl. Addition of 20 mM NH_4Cl leads to near complete dequenching. Each experiment was performed in triplicate.

3.4.4 Na^+/H^+ antiport activity by NapA is electrogenic

In this set-up, we followed ΔpNa or $\Delta\psi$ driven H^+ transport as a consequence of electrogenic Na^+/H^+ exchange activity. In this sensitive assay, liposomes containing the highly soluble and membrane impermeable pH sensitive dye pyranine were used to follow H^+ influx or efflux. A Na^+ gradient was established by the addition of NaCl, whereas an electrical membrane potential was established with a K^+/valinomycin diffusion potential.

Reconstitution of NapA into liposomes containing pyranine was essentially performed as described[87]. In brief, to a 500 μl liposome (40-80 nm) suspension (soy bean lipids, type II, SIGMA, 20 mg ml^{-1}) in buffer A (10 mM MOPS-PO_4, pH 7.5), 45 μml cholate (20% stock solution) and 17 μl NapA (50 μM, purification buffer) was added and incubated for 30 min at room temperature with occasional mild mixing. The cholate was removed via a PD-10 gel filtration column (GE Healthcare) equilibrated with buffer A and the proteoliposomes in the void volume were collected (∼1.2 ml). They were diluted to 8 ml with buffer A, collected via ultracentrifugation (200,000g, 4°C, 30 min) and resuspended in 250 ml buffer A. Then, 125 μl of proteoliposomes was mixed with 1 mM pyranine (0.1 M stock solution) and the desired Na_2SO_4 and K_2SO_4 concentrations, frozen in liquid nitrogen, thawed in water and

briefly sonicated in a bath type sonicator (2 × 5 s). The freeze/sonication procedure was repeated once. The external pyranine was subsequently removed via a prepacked G25 gel filtration column (GE Healthcare) and the proteoliposomes, equilibrated in buffer A, were collected from the void volume of the column.

Pyranine fluorescence measurements monitoring pH changes on the inside of the proteoliposomes were performed as described[87]. Typically, an amount of 20 ml liposomes containing the desired Na^+ and K^+ concentrations was mixed with 2.5 ml buffer containing the same buffer with the appropriate salt concentrations. After 30 s, exchange activity was either initiated by the addition of 50 mM Na^+ to the outside (0.5 mM Na^+ on the inside), establishing an inwardly-directed Na^+ gradient (∼120 mV), and leading to H^+ efflux. Accordingly, in a system in the absence of Na^+ gradient (50 mM Na^+ on both sides), but in the presence of a K^+ gradient (100 mM K^+ inside, 1 mM K^+ outside liposomes), addition of valinomycin (10 nM) established a membrane potential of ∼120 mV (inside negative) driving H^+ influx (and Na^+ efflux).

3.4.5 Crystallization and preliminary screening

Crystals were grown at 20°C using the hanging drop vapour diffusion method. A 1 μl aliquot of pure protein was mixed 1:1 with reservoir solution containing 0.001 M zinc sulphate, 0.05 M HEPES, pH 7.8 and 22-36% PEG 600 (mutants) or 0.05 M glycine, pH 9.0, 0.05 M magnesium acetate, 26% PEG 400 (wild type). Crystals appeared overnight and reached maximum size after 3-4 days. For specified crystals, coverslips were transferred for overnight incubation with reservoir solution containing 2% increments of PEG 400. Dehydrated crystals in 30-36% PEG 400 were finally soaked with 1 μl reservoir solution containing 1% NM and 40% PEG 400 followed by flash freezing in liquid nitrogen before data collection.

Data were collected at the European Synchrotron Radiation Facility and Diamond Light Source. Most of the crystals were triclinic, however, very occasionally crystals with different space groups were observed. The highest resolution data were collected from an orthorhombic crystal of the triple mutant that had been reannealed on the beamline. This crystal was grown with 0.025% dichloromethane (Hampton Research) as an additive.

3.4.6 Structure determination

To obtain phases cysteine mutants were introduced into the protein to enable derivatization with mercury. Three positions were chosen (Met 20, Val 166 and Val 326) and single, double and triple mutants created as described above. Mercury-derivatized protein was prepared by incubation of the protein at 20°C for 1 h with 2.5 mM mercury acetate. The structure was solved using MIRAS from four triclinic crystals (native, mercury derivatized and selenomethionine) as shown in Supplementary Table 1. Data were processed using the Xia2 pipeline[88] to XDS[89], with further processing using the CCP4 suite of programs[90]. Heavy atom sites were located from anomalous difference Patterson maps of the double mutant using the program RSPS[91]. Phases were calculated and refined in SHARP[92]. Further mercury and selenomethionine sites were located in residual maps. The crystals contain four molecules in the asymmetric unit. Non-crystallographic symmetry operators were determined from the heavy atom positions and averaging was carried out in DM[93] with a mask calculated in O[94]. This map was then used as a search model for molecular replacement in Phaser[95] for the $P2_1$ and $C222_1$ data sets. On the basis of the operators obtained, multi-crystal averaging in DMMULTI[93] using the P1 (4 molecules in au), $P2_1$ (2 molecules in au) and $C222_1$ (1 molecule in au) data sets was carried out. This gave maps of sufficient quality to see all 13 helices of the NapA

subunit. Model building was carried out in O[94] and COOT[96] and was facilitated by the positions of the mercury and selenomethionine peaks. Refinement of the atomic coordinates and individual B-factors using the $C222_1$ data at 2.9 Å was carried out in PHENIX[97]. Secondary structure restraints were applied and refinement was interspersed with rebuilding using O and COOT. Peaks in anomalous difference maps indicated the presence of two zinc ions bound to the protein on the periplasmic surface to residues at the N terminus of TM-1 and the loop between TM1 and TM2. The final refinement statistics are shown in Supplementary Table 3.1.

Maps were calculated based on the final refined model for the data from the wild type selenomethionine derivatized crystals (Supplementary Table 3.1). At 3.7 Å resolution no clear differences were observed and the positions of the peaks in the anomalous difference maps were consistent with the positions of the methionines in the structure (data not shown).

Superpositions were carried out in Lsqman[98] such that all matching Cα pairs were less than 3.8 Å apart after superposition. Figures were drawn using Pymol[99] except those showing electron density, which were made using CCP4mg[100]. The inward-facing model of NapA was created by superposing the core and dimerization domains of NapA onto the corresponding domains of NhaA separately. The only adjustments that were made to the model were to the polypeptide chain between the two domains. No modifications were made to alleviate the clashes between the loop between TM-1 and TM1 and TM4a. The video was made in Lsqman[98] by morphing between the outward-facing crystal structure and the inward-facing model displayed using Pymol[98].

3.4.7 Molecular dynamics simulations

Molecular dynamics simulations of the NapA dimer in a mixed POPE/POPG bilayer were carried out with the Gromacs simulation package[101], either version 4.5.5 or a development version of 4.6. All simulations employed the CHARMM force field including CMAP[42,43] with the original TIP3P water model[51] and updated CHARMM parameters for POPE and POPG lipids (CHARMM)[45], as implemented in Gromacs[101]. The ratio of POPE to POPG molecules was about 4:1 to approximate the major components of the *E. coli* membrane. We used a multi-scale approach to embed the dimer into the membrane[50]. The protein was first simulated in a coarse grained representation[102] and the membrane was allowed to self-assemble around the protein from a random mixture of lipids and water in the simulation box[103]. After 200 ns simulation with an integration time step of 20 fs the bilayer had assembled around the NapA dimer. In the second step of the multi-scale approach, the system was converted to the CHARMM atomistic representation with the CG2AT protocol[50] and the original crystal structure inserted in place of the back-translated protein. The simulation system consisted of an orthorhombic simulation box of size 114 Å × 114 Å × 91 Å containing 118,832 atoms in 768 protein residues, 215 POPC and 57 POPG lipids, 180 Na^+, 109 Cl^- ions and 24,575 water molecules. The free NaCl concentration was about 250 mM in all simulations to approximate the outward-open transporter facing an environment of increased salt stress.

Most titratable residues were predicted by PROPKA[104] to be in their default charge states at pH 7.8 and were simulated as such (including deprotonated Asp 156 and Asp 157, with predicted pK_a as 3.2 and 5.7). The pK_a of the buried Lys 305 was predicted to be 9.5, with its charged form partially stabilized by a salt bridge with the (charged) Asp 156. Therefore, Lys 305 was simulated in its protonated (positively

charged) form. For a number of residues the predicted value was within 1 unit of the environmental pH value we wished to simulate and for those residues we rationalized our choices as follows: The pK_a of Glu 333 was predicted as 7.0. Glu 333 might help to stabilize the helix dipoles of helices TM4a and TM11a (similar to Asp 133 in NhaA) but such a charge-dipole interaction is not encoded in the empirical rules of the PROPKA algorithm[104]. Furthermore, simulations clearly showed that Glu 333 protrudes into the outward facing cavity and is solvated. We, therefore, adopted a charged Glu 333. His 6 and His 51, which reside on the surface and do not seem to have any specific mechanistic role, had predicted pKa of 7.8 and 7.1 and were modelled in their dissociated (neutral) form.

Equilibrium molecular dynamics simulations were performed with periodic boundary conditions at constant temperature $T = 323$ K and pressure $P = 1 \times 10^5$ Pa using the velocity rescaling algorithm for the thermostat (time constant 0.1 ps)[62] and semi-isotropic Parrinello-Rahman barostat (time constant 5.0 ps, compressibility 4.6 $\times$ 10^{-10} Pa^{-1}, coupling every 10 steps)[63]. Long range corrections for energy and pressure were applied[55]. Lennard-Jones interactions were switched off between 8 Å and 12 Å, and electrostatic interactions were handled by the SPME method[105] in which Coulomb interactions were computed in real space up to a cut-off of 12 Å and long range interactions beyond the cut-off were calculated in reciprocal space with fast Fourier transforms on a grid with spacing 1.2 Å and fourth order splines for fitting of the charge density. The grid-based neighbour list was updated every five steps to a distance of 14 Å. Bonds to hydrogen atoms were constrained with the P-LINCS algorithm[55] or SETTLE (for water molecules)[56]. The classical equations of motions were integrated with a leap frog integrator and a time step of 2 fs. Conformations were saved every 1 ps for analysis.

The simulation protocol included an initial energy minimization of the atomistic system and a 1 ns equilibrium simulation during which the protein heavy atoms were restrained with a harmonic force with force constant of 1,000 kJ mol^{-1} nm^{-2}. An initial unrestrained simulation of the dimer with Asp 156 and Asp 157 in their default protonation state was run for 100 ns. Eight additional 100 ns simulations were performed in three sets (two simulations in set 1, five in set 2, one in set 3). The starting configuration for each set was generated from the last frame of the initial simulation by exchanging any sodium ion within 3 Å of Asp 157 or Asp 156 with a random bulk water molecule. Repeats in each set always differed by the seed of the random number generator, thus leading to differing initial assignments of velocities and generation of independent trajectories through the stochastic component of the velocity rescaling thermostat[62], as seen from the different r.m.s.d. time series in Supplementary Figure 3.10.

Additional simulations were performed to assess the influence of the protonation state of the conserved residues Asp 156 and Asp 157 on sodium binding as performed previously for NhaA[23]. Two independent 100 ns simulations were performed for each of (1) Asp 156 deprotonated (negatively charged) and Asp 157 protonated (neutral), (2) Asp 156 protonated and Asp 157 deprotonated, and (3) both Asp 156 and Asp 157 protonated. Charge states were modified with the Gromacs tool pdb2gmx[101]. These six simulations used the 100 ns frame of the initial simulation (default charge states) with any sodium ions near the aspartates exchanged with a bulk water molecule as a well-equilibrated starting conformation. In each case, a 1 ns position restraint simulation (as above) was followed by a 100 ns production equilibrium simulation.

Simulations were analysed with MDAnalysis[106] and Gromacs tools[101]. To calculate the sodium density, data from all nine simulations with the deprotonated aspartates were used at 1 ps intervals. Ion binding and unbinding to each protomer

appeared to be independent so that data for both protomers were combined by superpositioning trajectories of both protomer A and protomer B on the coordinates of protomer A from the start of the initial simulation. The density was calculated by histogramming sodium coordinates in cubic volume elements at a resolution of 1 Å in a fixed coordinate system defined by the initial coordinates of protomer A. The thicknesses of bilayer regions were calculated from the distributions of the headgroup phosphate and acyl chain atoms along the membrane normal, using trajectories superpositioned on the dimer.

Images showing simulation data were prepared with VMD[107] and the Bendix plugin for curved helices[108] or UCSF Chimera[109,110].

Acknowledgements

We are grateful to D. Slotboom for critical reading of the manuscript and N.-J. Hu for assistance in data collection. Data were collected at the European Synchrotron Radiation Facility and Diamond Light Source, with excellent assistance from beamline scientists. This work was funded by grants from the Medical Research Council (MRC grant G0900990 to A.D.C. and D.D.), the Swedish Research Council (to C.v.B. and D.D.) and the BBSRC (BB/G02325/1 to S.I.). The authors are grateful for the use of the Membrane Protein Laboratory funded by the Wellcome Trust (grant 062164/Z/00/Z) at the Diamond Light Source Limited and The Centre for Biomembrane Research (CBR), supported by the Swedish Foundation for Strategic Research. Computer simulations were partially run on XSEDE resources (grant TG-MCB120151 to O.B.). C.L. was a recipient of a BBSRC-funded PhD scholarship, H.J.K. a Human Frontiers Science Program (HFSP) postdoctoral fellowship, and D.D. acknowledges the support from The Royal Society through the University Research Fellow (URF) scheme.

Author Contributions

A.D.C. and D.D. designed the project. Cloning, expression screening, protein purification and crystallization were carried out by C.L. and D.D. with assistance from H.J.K., S.N., S.I. and A.D.C. Data collection and structural determination were carried out by C.L., D.D. and A.D.C. Experiments for functional analysis were designed by C.v.B. and D.D. and carried out by C.v.B., C.L., P.U. and D.D. Molecular dynamics simulations were carried out by D.L.D. and O.B. A.D.C. and D.D. wrote the manuscript with contributions from C.L., H.J.K., C.v.B. and O.B.

Author Information

The coordinates and the structure factors for NapA have been deposited in the Protein Data Bank under accession code 4BWZ. Reprints and permissions information is available at `www.nature.com/reprints`. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.D.C. (a.cameron@warwick.ac.uk) or D.D. (d.drew@imperial.ac.uk; ddrew@dbb.su.se).

3.5 Supplementary Information

	Native (triple)	Native (triple)	Hg (double)	Hg (triple)	Se (single)	Se (wild type)	Hg (triple)
Data collection							
Wavelength (Å)	0.873	0.873	1.008	0.873	0.979	0.978	1.005
Space group	$C222_1$	P1	P1	P1	P1	P1	$P2_1$
Cell dimensions							
Beamline	ID23_2	ID23_2	I03	ID23_2	ID23_1	I24	I03
a, b, c (Å)	73.8, 82.2, 191.8	79.4, 95.5, 105.2	78.8, 95.3, 103.8	77.2, 95.3, 101.7	79.5, 95.1, 104.2	79.5, 93.7, 106.0	55.0, 204.1, 64.2
α, β, γ (°)	90.0, 90.0, 90.0	77.9, 76.0, 80.9	77.5, 75.7, 80.3	77.1, 77.8, 79.5	77.7, 76.3, 80.8	77.6, 76.2, 81.0	90.0, 90.0, 90.0
Resolution (Å)	63.9–3.0 (3.06-2.98)[a]	76.6–4.0 (4.09–3.99)	99.0–4.0 (4.15–4.02)	92.0–4.9 (5.0–4.87)	32.0–3.5 (3.62–3.52)	30.0–3.7 (3.76–3.70)	102.0–4.4 (4.56–4.44)
R_{merge}(%)	8.3 (95.0)	6.6 (85.8)	7.5 (98.4)	10.2 (67.8)	7.5 (115.6)	11.3 (99.7)	7.8 (98.2)
$I/\sigma(I)$	18.8 (2.8)	13.6 (2.1)	11.2 (2.3)	12.1 (3.6)	19.8 (3.0)	21.7 (1.4)	5.9 (2.9)
Completeness (%)	99.9 (100)	98.3 (97.3)	97.8 (95.3)	98.8 (98.1)	97.7 (97.8)	99.3 (99.1)	96.9 (96.2)
Redundancy	11.0 (9.3)	4.6 (4.6)	7.0 (6.8)	8.6 (8.6)	15.3 (15.7)	3.5 (3.5)	3.3 (3.4)
Phasing power[b]							
Isomorphous			0.59	0.63	0.29		
Anomalous			0.75	0.76	0.21		
Refinement							
Resolution (Å)	2.98						
No. reflections	22,911						
R_{work}/ R_{free}[c]	22.3/ 24.8						
No. atoms							
Protein	2824						
R.m.s deviations							
Bond lengths (Å)	0.008						
Bond angles (°)	1.151						
Ramachandran plot							
Outliers (%)[d]	0.8						

Table 3.1: Data Collection, Phasing and Refinement Statistics.
[a]Values in parentheses refer to data in the highest resolution shell
[b]Calculated to 3.5 Å;
[c]Based on 5% of reflections.
[d]From Molprobity [111]

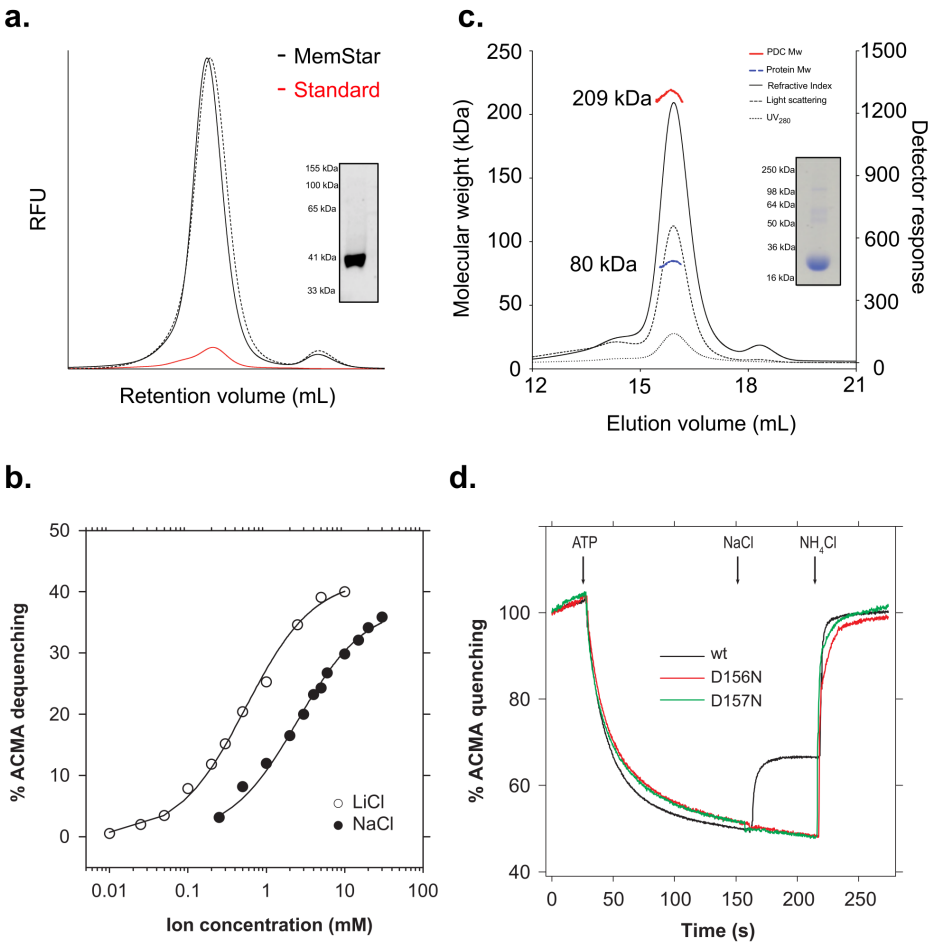


Figure 3.5: Characteristics of the NapA protein. **a**, Fluorescence-detection size-exclusion chromatography (FSEC) traces of LDAO (dotted line) or DDM (solid line) solubilised membranes, which were isolated from cells grown in MemStar (see Methods). For reference an FSEC trace of DDM solubilised membranes isolated from C43(DE3) cells grown in LB (red trace) is shown. In-gel fluorescence of NapA-GFP-His_8 expression is shown in the inset. **b**, Apparent binding affinity for Na^+ (closed circles) and Li^+ (open circles) in NapA at pH 7.5 for the triple cysteine mutant (Met20Cys, Val166Cys, Val326Cys) mutant, K_M Na^+ = 2.55 ± 0.17 mM and Li^+ = 0.54 ± 0.04 mM. **c**, Static light-scattering of purified NapA in DDM. The molecular weight corresponding to the NapA-detergent complex (red line) and NapA only (blue line) is as shown (the predicted Mw of NapA is 40.8 kDa (Q72IM4) and the purity of the protein used for analysis is illustrated by the Coomassie-stained SDS-gel (inset). **d**, Proteoliposomes containing D156N (red trace), D157N (green trace) and wild type (black trace). NapA mutants show no transport activity with the addition of saturating 100 mM Na^+ (red and green traces) or 50 mM Li^+ (data not shown).

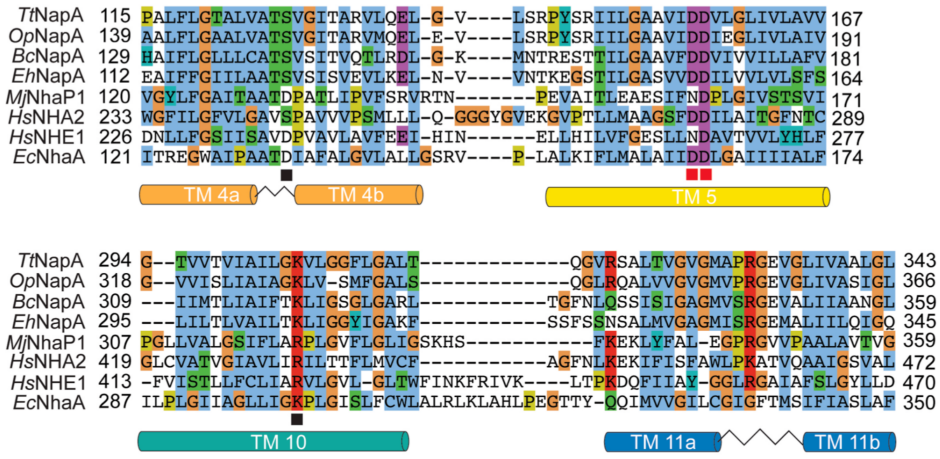


Figure 3.6: Sequence comparison of NapA to human and bacterial Na^+/H^+ antiporters. Sequence alignment of *Thermus thermophilus* NapA (UniProt: Q72IM4) to *Oceanithermus profundus* NapA (E4U6Q4; 65% sequence identity), *Bacillus cereus* NapA (C2MWQ1; 35% sequence identity), *Enterococcus hirae* NapA (P26235; 30% sequence identity), *Methanococcus jannaschii* NhaP1 (Q60362; 23% sequence identity), *Homo sapiens* Nha2 (SLC9B2; Q86UD5; 21% sequence identity), *Homo sapiens* Nhe1 (SLC9A1; P19634; (¡15% sequence identity), and *E. coli* NhaA (P13738; ¡15% sequence identity) using ClustalW (`www.ebi.ac.uk/clustalw/`) and MAFFT in Jalview. Although sequence identity is calculated using pairwise alignment to full length NapA, for sake of clarity, only TMs corresponding to 4, 5, 10 and 11 in NhaA are shown as these TMs harbor conserved charged residues known to be important for function [29]. Red squares indicate the predicted aspartate ion-binding residues and the black squares indicate ionisable residues located between the two charged dipoles at the cross-over of TMs 4 and 11 in NhaA.

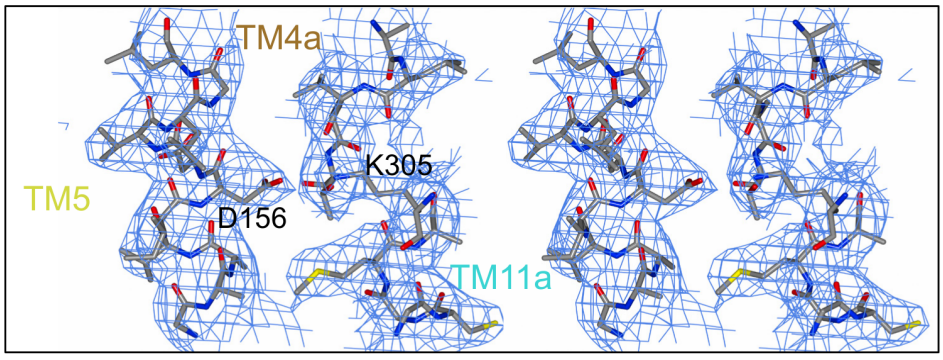


Figure 3.7: Typical electron density. The stereo view shows the region around Asp156. The 2mFo-DFc map has been calculated using phases derived from the final model and contoured at 0.06 e/A^3 (0.6σ).

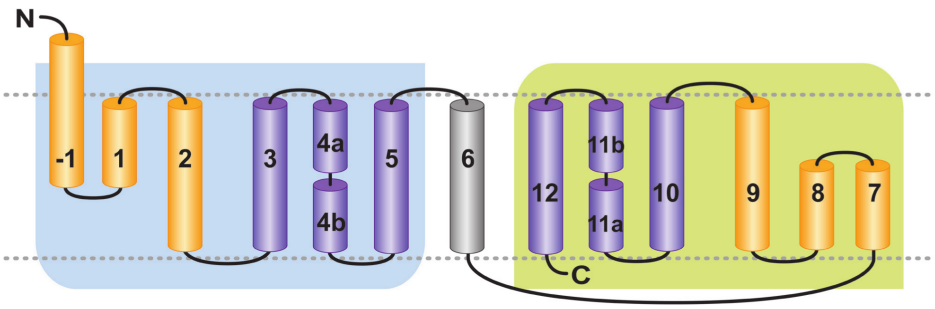


Figure 3.8: NapA topology. Coloured rods depict TMs -1 to 12 that form two 6-TM topology-inverted repeats, -1 to 5 and 7 to 12, connected together by TM6 (grey). The structural repeats intertwine to form a Dimerisation domain (pale orange rods) and a Core domain (blue-purple rods).

NapA	K_M Na^+	K_M Li^+
Wild type	4.0 ± 0.3	0.41 ± 0.04
M20C, V166C, V326C	2.9 ± 0.5	0.54 ± 0.04
D156N,	-	-
D156A*	-	-
D157N,	-	-
D157A*	-	-
K305A	-	>50
E333A	>50	1.3 ± 0.3

Table 3.2: Characterisation of NapA mutants. Apparent affinities were calculated from triplicate measurements (see Methods for details).
***The data were taken from Furrer *et. al.*[68].

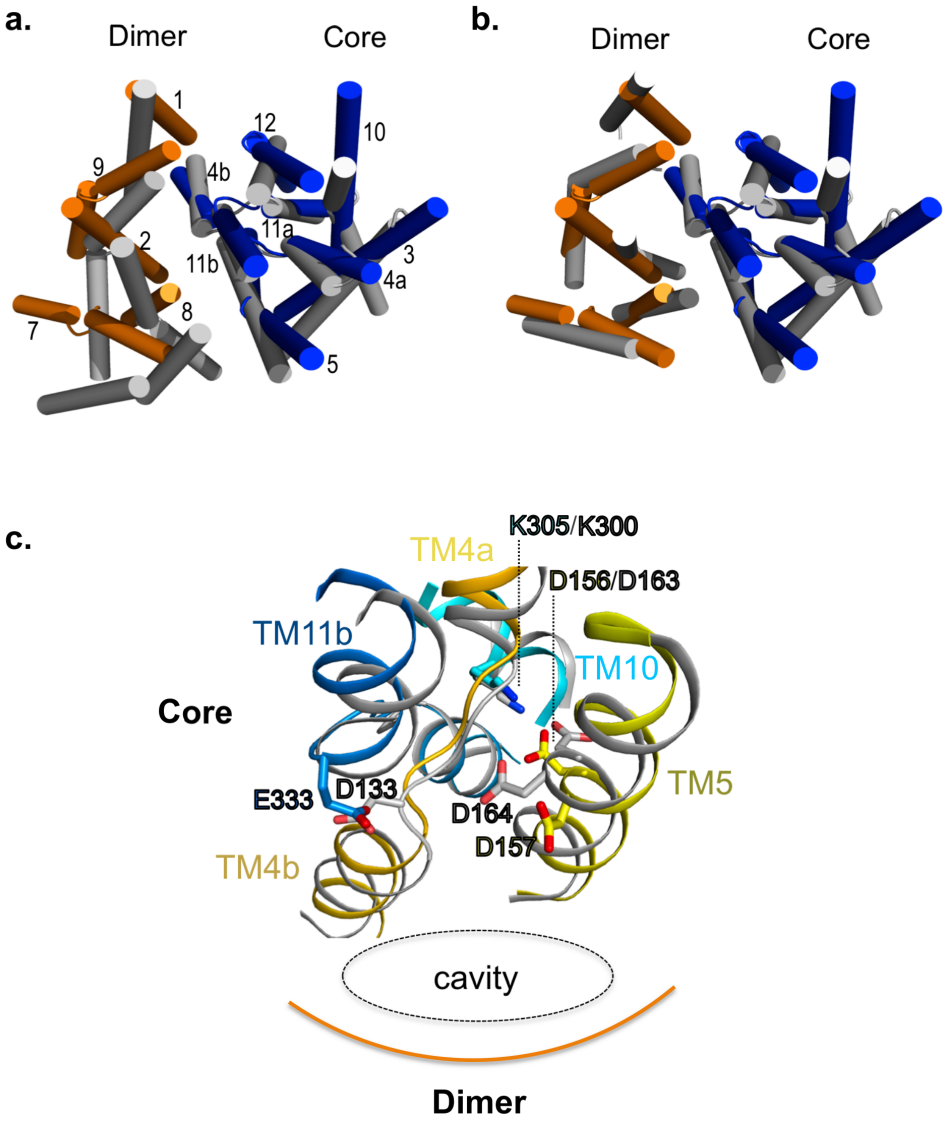


Figure 3.9: Structural comparison of NapA with NhaA. **a**, NhaA structure in grey (4AU5, O.B *et al.* manuscript in preparation) superimposed onto NapA shown in blue for Core and light orange for Dimerisation domain. TMs -1 and 6 have been omitted for clarity. **b**, As in **a**. except with the Dimerisation and Core domains of NhaA superimposed separately onto NapA. **c**, The charged residues discussed in the text (E333/D133, K305/K300, D156/D163, D157/D164) are shown as colour sticks for NapA and grey sticks for NhaA.

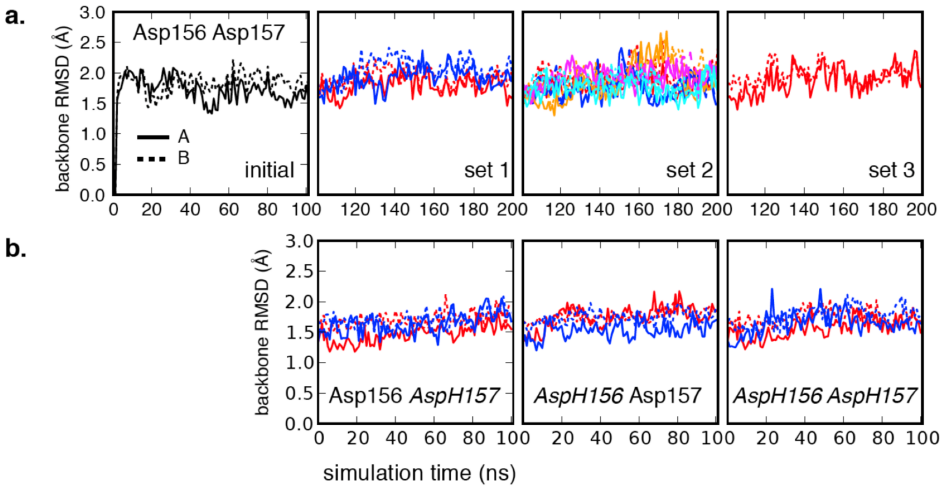


Figure 3.10: Backbone r.m.s.d. of MD simulations. The root mean square deviation of the backbone atoms from the starting frame of initial simulation (crystal structure in the membrane after energy minimisation and 1 ns MD with strong position restraints on all heavy protein atoms) was calculated by superpositioning the whole dimer on the reference structure at each trajectory frame. The panels show the r.m.s.d. for protomer A (solid line) and protomer B (dashed line). **a**. Simulations with Asp156 and Asp157 in their default protonation states (deprotonated, i.e. negatively charged): initial simulation and the three sets of simulations 1-3 that started from the last frame of the initial simulation. **b**. Simulations with alternative protonation states (two repeats each). Asp156 AspH157: charged Asp156 and protonated i.e. neutral Asp157. AspH156 Asp157: neutral Asp156 and charged Asp157. AspH156 AspH157: both aspartate residues are neutral.

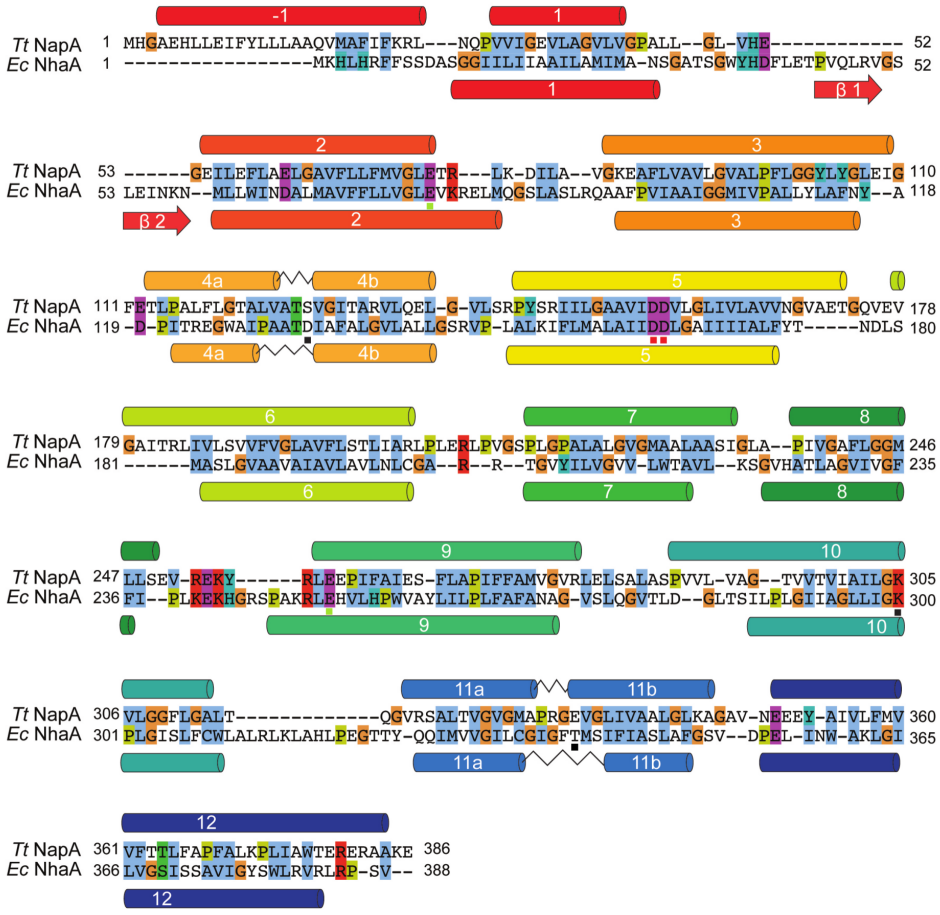


Figure 3.11: Sequence alignment of NapA to well characterised *E. coli* NhaA. Sequence alignment of NapA and NhaA using the PROMALS3D server with the NhaA and NapA structures and manual adjustment in JalView. The α-helices in NapA are depicted as coloured filled boxes from red to blue as in Figure 3.2. Aspartate residues, D156/157 (NapA) and D163/164 (NhaA), likely to coordinate sodium/lithium are labeled with a red square. Ionisable residues between discontinuous TM assembly E333/K305 (NapA) and D133/K300 (NhaA) are labeled with a black square and conserved charged residues E74/E258 (NapA) and E78/E252 (NhaA), thought to be part of pH sensing [29], are shown with a light green square. Other residues predicted to be involved in pH sensing are not conserved.

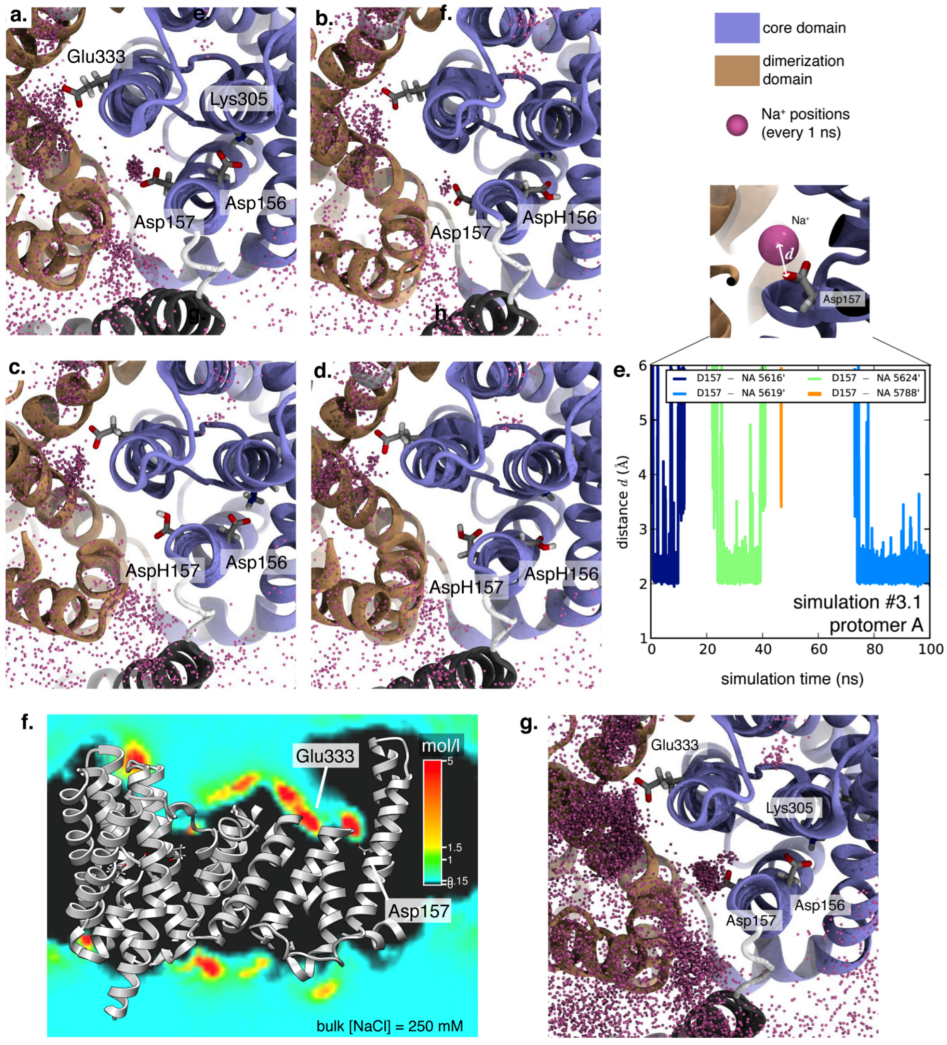


Figure 3.12: MD simulations of Na^+ binding to NapA. **a-d**. View from the extracellular side on a putative sodium binding site near Asp157, with data from 200 ns of MD for simulations with differing protonation states of Asp156 and Asp157. Although the aspartate sidechains display multiple rotameric states no substantial, larger conformational changes are observed over the 1.7 μs of aggregated simulation time (see also Supplementary Figure 3.10). Ion positions (small magenta spheres) were overlaid at 1-ns intervals in the context of a representative conformation of the protein (ribbons) with specific side chains highlighted as sticks. Data from both protomers were overlaid so that the images represent 400 ns of sampling of sodium positions. **a**. With default protonation states (Asp156 and Asp157 deprotonated and hence charged) multiple Na^+ binding/dissociation events were observed; see **e** for discussion of these events and g showing 0.9 μs of data. **b**. For protonated Asp156 (neutral) and charged Asp157 only a single, short binding event of 25 ns duration was observed. **c**. No binding events were observed for Asp156 charged and Asp157 neutral. **d**. Neutral Asp156 and Asp157 also do not show any Na^+ binding. **e** Representative time series of spontaneous sodium binding and dissociation events and definition of the ion-Asp distance. The distance d between Na^+ ions and the closest O_δ atom in Asp157 on each protomer is shown for protomer A in simulation #3.1 (from set 3). Binding events are indicated by prolonged residence at distances < 2.5 Å. Binding and dissociation of sodium ions was observed in almost all 100-ns simulations with Asp156 and Asp157 charged: For each *simulation #* we state the number of binding/dissociation events for each protomer (A, B): *initial* 1/1, 2/1; #1.1 1/0, 1/1; #1.2 1/1, 3/2; #2.1 2/1, 1/1; #2.2 0/0, 2/1; #2.3 2/1, 0/0; #2.4 1/0, 2/2; #2.5 1/0, 1/0; #3.1 3/2 (shown here), 2/2. **f** and **g**. Simulations with Asp156 and Asp157 deprotonated (charged), using data from two protomers over a total of 0.9 μs of simulated time (equivalent to 1.8 μs of sampling of the density around a single protomer). **f**. Cut through the Na^+ density in the context of the dimer structure. The approximate position of Glu333 is indicated, together with Asp157. **g**. Ion positions are shown as in **a** but from 900 ns of trajectory data. High sodium density regions near Asp157 and Glu333 are clearly visible. (Figure **f** was created in UCSF Chimera while **a-d** and **g** were made in VMD and rendered with Tachyon.)

Chapter 4

CRYSTAL STRUCTURE OF THE SODIUM-PROTON ANTIPORTER NHAA DIMER AND NEW MECHANISTIC INSIGHTS

This chapter is a reprint of the journal article, Lee, C., Yashiro, S., **Dotson, D.L.**, Uzdavinys, P., Iwata, S., Sansom, M.S.P., Ballmoos, C. von, Beckstein, O., Drew, D., and Cameron, A.D. (2014). Crystal structure of the sodium-proton antiporter NhaA dimer and new mechanistic insights. J Gen Physiol 144, 529-544. This work introduced a new crystal structure of inward-facing *Escherichia coli* NhaA, obtained at low-pH but crystallized as a physiological dimer. The new structure reveals a previously-unidentified salt-bridge between Asp163 and Lys300, two highly-conserved residues at the putative Na^+/H^+ binding site. We show that the original structure[22] incorrectly assigned the placement of Lys300 due to poor electron density in the vicinity of transmembrane-helix 10, resulting in a register shift that placed the lysine a full 10 Å from its proper position.

Remarkably, all-atom molecular dynamics simulations show that for this new structure, spontaneous binding of Na^+ to the conserved aspartates, Asp164 and Asp163, subsequently breaks the salt bridge formed between Asp163 and Lys300. This information, combined with heuristic pK_a estimates[112] of the titrateable residues obtained from simulation frames, suggests that Lys300 may be directly involved in ion transport as a H^+ carrier. Given these observations, we propose a new mechanism of transport for NhaA and similar transporters.

My contribution to this work was the performance and analysis of the molecular dynamics simulations, and I am a joint first author on this paper. This work first

ABSTRACT

Sodium-proton antiporters rapidly exchange protons and sodium ions across the membrane to regulate intracellular pH, cell volume, and sodium concentration. How ion binding and release is coupled to the conformational changes associated with transport is not clear. Here, we report a crystal form of the prototypical sodium-proton antiporter NhaA from *Escherichia coli* in which the protein is seen as a dimer. In this new structure, we observe a salt bridge between an essential aspartic acid (Asp163) and a conserved lysine (Lys300). An equivalent salt bridge is present in the homologous transporter NapA, but not in the only other known crystal structure of NhaA, which provides the foundation of most existing structural models of electrogenic sodium-proton antiport. Molecular dynamics simulations show that the stability of the salt bridge is weakened by sodium ions binding to Asp164 and the neighboring Asp163. This suggests that the transport mechanism involves Asp163 switching between forming a salt bridge with Lys300 and interacting with the sodium ion. pK_a calculations suggest that Asp163 is highly unlikely to be protonated when involved in the salt bridge. As it has been previously suggested that Asp163 is one of the two residues through which proton transport occurs, these results have clear implications to the current mechanistic models of sodium-proton antiport in NhaA.

4.1 Introduction

Electrogenic sodium-proton antiporters use the proton gradient to drive sodium ions out of the cell, usually under conditions of salt stress at alkaline pH[69]. NhaA from *Escherichia coli* is the prototypical electrogenic sodium-proton antiporter, with

two protons transported for every sodium ion[13]. For a transporter, NhaA is extremely fast, exchanging up to 1,500 ions/s[12]. Its activity is regulated by pH: it is inactive at acidic pH, partially active between pH 6.0 and 7, and fully active at pH 8.0[69]. The crystal structure of NhaA was determined at a resolution of 3.45 Å in an inward-facing conformation at low pH where the protein is inactive[22]. NhaA is made up of two five-transmembrane (TM) topology-inverted repeats that intertwine to form two distinct structural domains: a core (translocation) domain and a dimer (interface) domain[22,29] (Figure 4.1). In the crystal structure, a deep cytoplasmic-facing cavity, containing many negatively charged residues, is located between the core and dimer domains. The core domain is characterized by two discontinuous helices that cross over at the center of the protein (Figure 4.1). The two well-conserved aspartates Asp163 and Asp164 are located near the crossover region. Sodium ion binding to Asp163 and Asp164 at high pH is believed to elicit a switch of the transporter from the inward- to outward-facing conformation, although the details of the mechanism are unclear[24,28,29,64].

Recently, the structure of another sodium-proton antiporter (NapA) was solved at high pH in the outward-facing conformation[113]. In this new structure, the equivalent residue to Asp164 (Asp157 in NapA) was found to be accessible from a large cavity to the outside, rather than the inside as seen in NhaA. Furthermore, in the NapA structure, the equivalent residue to Asp163 in NhaA interacts with a well-conserved lysine (Lys305 in NapA, Lys300 in NhaA), which is thought to be important in charge neutralization and pH activation[22,28]. Here, we describe a new crystal form of NhaA that, like the previous NhaA structure, is in an inward-facing conformation at low pH, but shows the protein packing as a dimer, which is the physiological oligomeric state for sodium-proton antiporters in general[72,113–115]. We show that the salt-bridge interaction observed in the outward-facing structure of NapA is also evident between

Asp163 and Lys300 residues in the new inward-facing NhaA structure. To probe this salt-bridge interaction, further molecular dynamics (MD) simulations were performed in a model membrane bilayer, systematically changing the protonation states of charged residues in the ion-binding site.

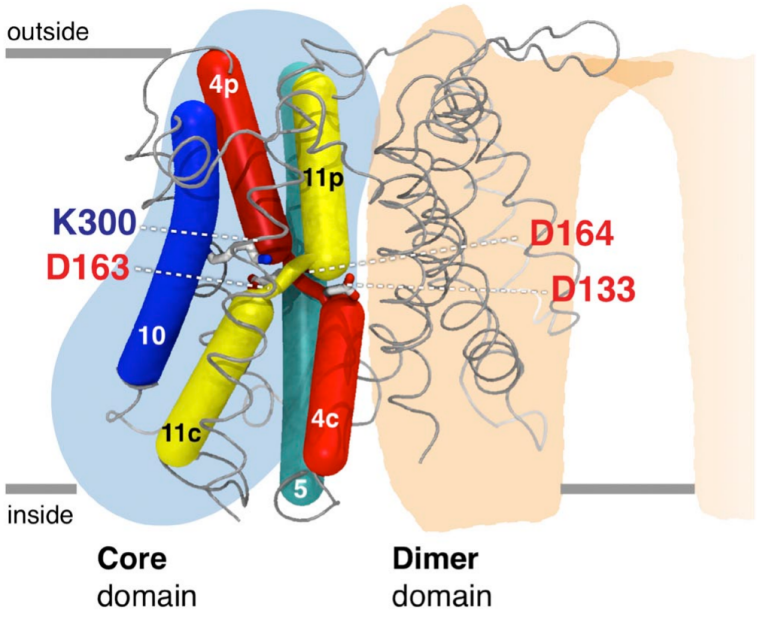


Figure 4.1: Schematic diagram of the NhaA structure. The core and dimer domains are illustrated with blue and beige shadowing, respectively. TM 4 (red) and TM 11 (yellow) are discontinuous and cross over in the center of the protein. Asp133 and Lys300 have been proposed to neutralize the positively and negatively charged helix dipoles of the discontinuous helices. Asp163 and Asp164 are thought to interact with the sodium ion. Coordinates are from Protein Data Bank accession number 1ZCD [22].

4.2 Materials and Methods

4.2.1 Expression, purification, and stabilization of NhaA

NhaA-GFP-His_8 was obtained from the previously constructed membrane protein GFP fusion library [116]. The NhaA-GFP-His_8 fusion protein was overexpressed and purified as described previously [67]. To confirm the sequence assignment of TM 10, Leu296 was mutated to methionine using the QuickChange II XL Site-Directed Mutagenesis kit (Agilent Technologies) in a variant of NhaA containing two stabilizing

point mutations (A109T and Q277G) as shown below. The A109T, Q277G construct is referred to as the "double mutant," and A109T, Q277G, L296M is referred to as the "triple mutant." Expression of the mutants was performed using MemStar[117]. In brief, the *E. coli* strain pLemo(DE3) was cultured in the auto-induction medium for selenomethionine labeling (PASM-5052) but also included an IPTG induction step at mid-log phase. Cultures were harvested after o/n growth at 25°C and were supplemented with kanamycin at 50 μg/ml throughout[81,82].

4.2.2 Preparation of stable mutant of NhaA

The L296M mutation was introduced into a variant of NhaA that had been stabilized by the introduction of two mutations. These stabilizing point mutants were identified by screening ∼100 random mutations of NhaA. Mutations were generated by error-prone PCR using the GeneMorph II Random Mutagenesis kit (Agilent Technologies). NhaA mutants were selected based on expression levels (>1 mg/L based on GFP fluorescence) and those that had >50% extraction efficiency in 2% wt/vol octyl-β-D-thioylglucoside (OG) from total membranes, at a concentration of 3 mg/ml, after 1-h incubation at 4°C. This procedure was performed because we observed a correlation between the stability of membrane proteins, as judged by their unfolding rate in LDAO at 40°C[67], and their solubilization efficiency in OG.

4.2.3 Functional characterization of the thermostabilized NhaA mutant

NhaA wild-type and the thermostabilized mutant were co-reconstituted with purified ATP synthase from *E. coli* with an ∼2:1 molar ratio (NhaA/ATP synthase) in MME buffer (10 mM MOPS-NaOH, pH 8.5, 2.5 mM $MgCl_2$, and 100 mM KCl) as described previously[113]. Typically, 50 μl proteoliposomes were diluted into 1.5 ml MME buffer containing 3 nM 9-aino-6-chloro-2-methoxyacridine (ACMA) and 140 nM

valinomycin. Fluorescence was monitored at 480 nm using an excitation wavelength of 410 nm in a fluorescence spectrophotometer (Cary Eclipse; Agilent Technologies). An outward-directed pH gradient (acidic inside) was established by the addition of 2 mM ATP, as followed by a change in ACMA fluorescence. After an ∼2-min equilibration, the activity of NhaA wild-type and thermostabilized mutant was assessed by the dequenching of ACMA fluorescence after the addition of the indicated concentrations of NaCl or LiCl. The addition of 20 mM NH_4Cl leads to near complete dequenching. By measuring ACMA dequenching by the addition of 10 mM NaCl or LiCl at pH 6.5, 7.0, 7.5, 8.0, 8.5, and 9.0, respectively, the effect of pH to NhaA wild-type and thermostabilized mutant activity was assessed. Each experiment was performed in triplicate.

4.2.4 Crystallization

Crystals of wild-type NhaA and the triple mutant were grown by mixing equal volumes of a protein solution containing 8 mg/ml of pure protein in 20 mM sodium citrate, pH 3.5, 150 mM NaCl, 0.03% high α-*n*-dodecyl-β-D-malotoside (α-DDM), and 1% heptylthiol-β-D-glucoside with reservoir solution containing 0.1 M sodium citrate, pH 3.5, 0.1 M LiS04, and 26% PEG 400. For the triple mutant, 1% Facade-EM (Avanti Polar Lipids, Inc.) was included as an additive. The crystals were dehydrated by transferring the cover slides sequentially to wells with 2% increments of PEG 400 to a final concentration of 32%. Crystals were soaked with 1 μl of reservoir solution containing 1% high α-DDM and 40% PEG 400 and flash frozen in liquid nitrogen before data collection.

4.2.5 Data collection, processing, and refinement

Wild-type NhaA. Data were collected at Diamond Light Source beamline I24 and were processed and scaled using HKL2000[118]. Molecular replacement was performed using Phaser[95] as part of the CCP4 package[90], with the previously published monomeric structure as the search model (Protein Data Bank accession no. 1ZCD[22]). Inspection of the solution showed two dimers to be present in the asymmetric unit. Refinement was performed against data extending to 3.7 Å using the PHENIX package[97] starting from a model in which the β hairpin (Pro45 to Asn58) and other surface loops were omitted. The maps were improved by averaging the four molecules of the asymmetric unit using the RAVE package[119] with B-factor sharpening as described by DeLaBarre and Brunger[120]. Rebuilding was performed in O[94]. The structure was refined with one B factor for each residue TLS[121], grouped into chains, secondary structure restraints, and noncrystallographic symmetry (NCS) restraints between the four molecules of the asymmetric unit. During refinement, the register of TM 10 was moved by one turn of the helix. This improved the fit of the helix to the density and the geometry of the residues in this region. To confirm the assignment, Leu296 on TM 10 was changed to a methionine and data were collected from selenomethionine derivatized crystals (see next section). As this dataset was of higher resolution than the wild type, the model was refined first against these data (see below) before being refined against the wild-type data in a final round of refinement with dihedral restraints to the higher resolution model.

NhaA triple mutant. Data were collected on beamline I03 at Diamond Light Source. They were processed with XDS[89] through the xia2 pipeline[88]. Refinement was performed using the PHENIX package[97] as described above. Refinement was performed against data extending to a resolution of 3.5 Å as for the wild type, except

that F' and F" were given for the selenium atoms and the refinement was performed against the anomalous pairs. In the last rounds of refinement, NCS restraints were only applied over the two dimers. This resulted in slightly better R and R-free.

Coordinates and structure factors have been deposited in the Protein Data Bank under accession number 4AU5 for the wild- type structure and 4ATV for the triple mutant.

4.2.6 MD simulations

MD simulations of the NhaA dimer and a monomer in a 1-palmitoyl-2-oleoyl-phosphatidylcholine (POPC) bilayer were performed with the Gromacs 4.5.3 or 4.6.1 simulation packages[101]. All simulations used the OPLS-AA force field[46–48] with the TIP4P water model[51] and OPLS-UA parameters for POPC lipids[49] (provided by M. Ulmschneider and available from the Lipidbook force field repository[122]). A POPC bilayer was studied as a generic model of a native-like membrane environment instead of other lipid compositions, as the POPC lipid parameters have been shown to perform well in very long ($\geq$1-μs) simulations[123,124]. POPC is a reasonable choice as a model membrane because its membrane thickness is comparable to that of the POPE-rich membrane[125–127] of *E. coli*[128,129]. We used a multi-scale approach to embed the dimer into the membrane[50]. The protein was first simulated in a coarse-grained representation, and the membrane was allowed to self-assemble around the protein from a random mixture of lipids and water in the simulation box[103]. In this way, lipids could accumulate in the dimer interface in an unbiased fashion, only driven by the thermodynamics of the partitioning of the random mixture into a water phase and a lipid/membrane phase. After a 200-ns simulation with an integration time step of 20 fs, the bilayer had assembled around either the NhaA dimer or a monomer (based on chain A of the crystal structure). In the second step of the multi-scale

approach, the systems were converted to the OPLS-AA atomistic representation with the CG2AT protocol[50], and the original crystal structure was inserted in place of the back-translated protein. The dimer simulation system consisted of an orthorhombic simulation box of size of 121 × 121 × 93 Å containing 112,700 atoms in 748 protein residues, 354 lipids, 49 Na^+ and 55 Cl^- ions, and 27,546 water molecules. The monomer system measured 75 × 75 × 90 Å and contained 52,332 atoms (374 protein residues, 135 POPC lipids, 17 Na^+ and 20 Cl^- ions, and 9,875 water molecules). The approximate free NaCl concentration was 100 mM in all simulations.

Equilibrium MD simulations were performed with varying protonation states of Asp133, Asp163, Asp164, and Lys300 (see Table 4.1). The choice of protonation state for other residues was guided by PROPKA[104] based on their pK_a, at pH 7. Asp133 is believed to stabilize the helix dipoles of TMs 4a and 11a, but such a charge-dipole interaction is not encoded in the empirical rules of the PROPKA algorithm[104]. Therefore, we disregarded the predicted value of 7.1 and adopted a charged Asp133, but also performed simulations with Asp133 in its neutral form. Glu82 and Glu252 had predicted pK_a values >8, but in the structure and in simulations, they face the water-filled entrance funnel so the charged default states were selected instead of the protonated forms.

MD simulations were performed with periodic boundary conditions at constant temperature T = 310 K and pressure P = 1 bar using the velocity-rescaling algorithm for the thermostat (time constant of 0.1 ps)[62] and semi-isotropic weak coupling for the barostat (time constant of 1.0 ps; compressibility of 4.6 × 10^{-5} bar^{-1})[130]. Long-range corrections for energy and pressure were applied[101]. Lennard-Jones interactions were cut off at 10 Å, whereas electrostatic interactions were handled by the smooth particle mesh Ewald (SPME) method[105] that computes Coulomb interactions in real space up to a cutoff of 10 Å and long-range interactions beyond the

cutoff in reciprocal space with fast Fourier transforms on a grid with spacing 1.2 Å and fourth-order splines for fitting of the charge density. Bonds to hydrogen atoms were constrained with the P-LINCS algorithm[101] or SETTLE (for water molecules)[56]. The grid-based neighbor list was updated every five steps. The classical equations of motions were integrated with a leapfrog integrator and a time step of 2 fs. Conformations were saved every 1 ps for analysis. The simulation protocol included an initial energy minimization of the atomistic system and a 3-ns equilibrium simulation during which the protein heavy atoms were restrained with a harmonic force constant of 1,000 kJ mol^{-1} nm^{-2}. Simulations were performed as detailed in Table 4.1 with at least three repeats of each main simulation. To change protonation states of residues, the Gromacs tool pdb2gmx was used[101], which also rebuilds hydrogens as needed. When a simulation was continued from a previous simulation, another energy minimization and 3-ns equilibrium MD with positional restraints on the protein heavy atoms were performed after changing protonation states. These simulations are marked with an asterisk in Table 4.1. Repeat simulations started from the last frame of a position restrained simulation were run with different seeds for the random number generator so that the differing initial velocity distributions and the stochastic component of the thermostat[62] would generate independent trajectories. In total, we simulated the dimer system for >10 μs.

4.2.7 Analysis and estimation of pK_a values

MD trajectories were analyzed with MDAnalysis[106] and Gromacs tools[101]. The distance d of the Asp163-Lys300 salt bridge was calculated as the minimum distance between the two carboxyl $O_{\delta 1}$ and $O_{\delta 2}$ atoms with the amine N_ζ at each time step; similarly, ion-carboxyl group distances were also calculated as the minimum distance to the two oxygen atoms. After the simulations, the pK_a values of titratable residues

Simulation name[a]	Assembly	Charge state				Starting structure[b]	Run length
		D133	D163	D164	K300		µs
S1/1	dimer	—	—	0	+	xtal*	1.1
S1/2						S1/1@0.1 µs	0.2
S1/3						S1/1@0.1 µs	0.2
S2/1	dimer	—	—	—	+	xtal*	1.0
S2/2						xtal	1.0
S2/3						xtal	1.0
S2/1.1–1.5	monomer					S2/1	5×0.1
S2/2.1–2.2						4AU5*	2×0.1
S3/1	dimer	—	0	0	+	S2/1*	0.1
S3/2						S2/1	1.0
S3/3						xtal*	0.1
S4/1	dimer	—	—	—	0	S2/1*	1.0
S4/2						S4/1@0.1 µs	1.0
S4/3						S4/1@0.1 µs	0.2
S5/1	dimer	0	—	0	+	xtal*	1.1
S5/2						S5/1@0.1 µs	0.2
S5/3						S5/1@0.1 µs	0.2
S6/1	dimer	0	—	—	+	S2/1*	0.1
S6/2						xtal*	0.1
S7	dimer	0	0	0	+	xtal*	0.1
S8	dimer	0	—	—	0	S2/1*	0.1

Table 4.1: MD simulations. Asterisks denote simulations that were preceded by energy minimization and a 3-ns position restraint MD; simulations without an asterisk were repeats starting from the same initial system conformation as the starred one but with varied initial velocity distribution. The dimer simulation contained 112,700 atoms and the monomer contained 52,332. All simulations were performed with Gromacs 4.6.1 except simulations S2/1, S2/1.1-1.5, and S2/2.1-2.2, which were run in Gromacs 4.5.3. Judging from the RMSD and secondary structure analysis (not depicted), the simulations behaved in the same manner regardless of software version. In total, 10.5 μs of MD simulations was performed. xtal, crystal structure of wild-type NhaA, which was deposited in the Protein Data Bank under accession number 4AU5 after minor refinements; 4AU5, wild-type NhaA as deposited in the Protein Data Bank; S1/1, last frame of simulation S1/1, etc., or frame at 0.1 μs ("@0.1μs").
[a]Simulations are identified by the protonation states and resulting charge states of Asp133 (D133), Asp163 (D163), Asp164 (D164), and Lys300 (K300), using identifiers S1-S8. Repeat simulations are indicated with a serial number after the identifier.
[b]Starting structure denotes the source for the initial input structure for the simulation.

were estimated with PROPKA 3.1[131] using snapshots of the MD simulations that were sampled every 1 ns. Na^+ ions within 6 Å of the protein were taken into account because preliminary calculations showed that the presence of ions close to titratable residues can shift the pK_a by up to 2 units. Data were typically split according to (a) the selected protonation states in the simulation and (b) the state of the Asp163-Lys300 salt bridge. The salt bridge was considered formed if $d < 4$ Å and broken if $d \geq 4$ Å. This value is consistent with work by Kumar and Nussinov[132,133] who considered a salt bridge to exist with $d < 4$ Å between any N-O pair.

The pK_a time series data from the individual protomers A and B in each simulation of the dimer and from repeats of individual simulations were aggregated. Distributions of pK_a values were modeled with a Gaussian kernel density estimator whereby the kernel width was chosen according to Scott's criterion[134] as implemented in the `scipy.stats.gaussian_kde()` function from the SciPy package. Distributions were plotted as violin plots[135] as implemented in the Seaborn package[136]. To assess the effect of the salt bridge on the pK_a values of neighboring residues, the distribution $f(\Delta \mathrm{p}K_a)$ of the difference between pK_a from frames with a broken salt bridge minus pK_a from frames with the intact salt bridge was computed: $f(\Delta \mathrm{p}K_a) = \int dw\, f^{\mathrm{broken}}_{\mathrm{p}K_a}(w) f^{\mathrm{intact}}_{\mathrm{p}K_a}(w - \Delta \mathrm{p}K_a)$, where the distributions $f^{\mathrm{broken}}_{\mathrm{p}K_a}$ and $f^{\mathrm{intact}}_{\mathrm{p}K_a}$ of the pK_a for the intact and broken salt bridge were computed from the data as Gaussian kernel density estimates as detailed above. The ΔpK_a shifts aggregated over all simulations were used to compare the influence on the calculations of the inclusion of Na^+ ions, the distance defining whether the salt bridge was formed or broken and the effect of changing the surface dielectric constant $\epsilon_{\mathrm{surface}}$.

Figures were prepared with PyMOL (Schrödinger, Inc.), CCP4mg[100], and VMD[107], using the Bendix plugin for smoothly bent helices[108]. Superpositions were performed in Lsqman[98], such that all matching Cα pairs were <3.8 Å apart after superposition.

4.2.8 Online supplemental material

The supplemental Discussion gives a more in-depth analysis of the pK_a estimations. Table 4.4 shows the solubilization efficiency of the various NhaA constructs that were screened in searching for a more stable construct. Table 4.5 shows the root-mean-square deviation (RMSD) in positions of the Cα atoms among the structures. Figure 4.11 shows an additional diagram illustrating the difference in the position of Lys300 in this structure relative to the previously published structure (Protein Data Bank ac-

cession no. 1ZCD) and provides a stereo diagram of the electron density in this region, complementing Figure 4.2. Figure 4.12 presents further analysis of the behavior of the protein and lipids during simulation S2/1. Figures 4.13-4.17 show the analyses of the repeated simulations in different protonation states in a similar fashion to the summary presented in Figure 4.7. Figure 4.18 shows the variation in the distance between Asp163 and Lys300 plotted over all simulations. The pK_a values of Asp163, Asp164, Lys300, and Asp133 are also plotted as a function of time as each of the simulations progressed. Figure 4.19 shows a distribution of pK_a values for these four residues when the salt bridge is considered intact and when broken as shown in Figure 4.8, but each set of simulations is plotted separately. Figure 4.20 shows the shift in pK_a of these residues when the salt bridge breaks. Videos 1 and 2 show the first 250 ns of the MD simulation S2/1 for protomers A and B, respectively. Online supplemental material is available at `http://www.jgp.org/cgi/content/full/jgp.201411219/DC1`.

4.3 Results

4.3.1 Structure determination of the NhaA dimer

Crystals of wild-type NhaA were obtained at low pH using protein prepared as a GFP fusion construct[66,67,79]. These crystals contain two dimers in the asymmetric unit. The structure was solved by molecular replacement and refined at 3.7-Å resolution with noncrystallographic restraints between the four molecules of the asymmetric unit (Table 4.2). In the previous crystal structure of NhaA, TM helix 10 (TM 10), which contains the conserved Lys300, was reported to be difficult to build into the electron density[74]. During refinement of the NhaA dimer, we observed that the fit could be improved by adjusting the sequence assignment for TM 10 so that it starts at

residue 287 and extends to 313, rather than from residue 290 to 316, as was modeled previously[22] (Figures 4.2 and 4.11).

4.3.2 Verification of sequence assignment in a thermostabilized form of NhaA

After repositioning TM 10, Lys300 is now within hydro- gen-bonding distance to Asp163 (Figure 4.3). To verify the modified sequence assignment of TM 10, Leu296 was substituted to methionine and selenomethionine derivatized protein produced to obtain anomalous difference maps that would show the exact position of the introduced methionine. However, initial attempts at obtaining well-diffracting selenomethionine derivatized crystals were unsuccessful. To improve crystal quality, a thermostabilized NhaA mutant construct was used instead (Figure 4.4). The NhaA mutant was generated while developing methodology that can rapidly improve the stabilization of membrane proteins by mutagenesis, in particular, when no high affinity ligands are available. In brief, an error-prone PCR library was generated and NhaA mutants were screened for those that could be extracted with >50% efficiency in 2% OG (see section 4.2 and Table 4.4), a benchmark based on our observation that many of the most detergent-stable membrane proteins[67] had >50% extraction efficiency in 2% OG. By incorporating the Leu296Met substitution into an NhaA thermostabilized mutant (Ala109Thr, Gln277Gly), data were obtained extending to a resolution of 3.5 Å after minimal optimization of the crystals (Table 4.2). Importantly, the thermostable mutant has the same pH-dependent profile and similar transport activity to the wild-type protein (Figure 4.4). Anomalous difference maps for the mutant showed peaks in the same position as the methionines in the monomeric wild-type structure[22], consistent with the two proteins having similar conformations. In addition, there was another peak exactly at the position of the introduced methionine for

the Leu296Met mutation in the dimer structure (Figure 4.2C). Thus, we were able to confirm that the reassignment of TM 10 was correct.

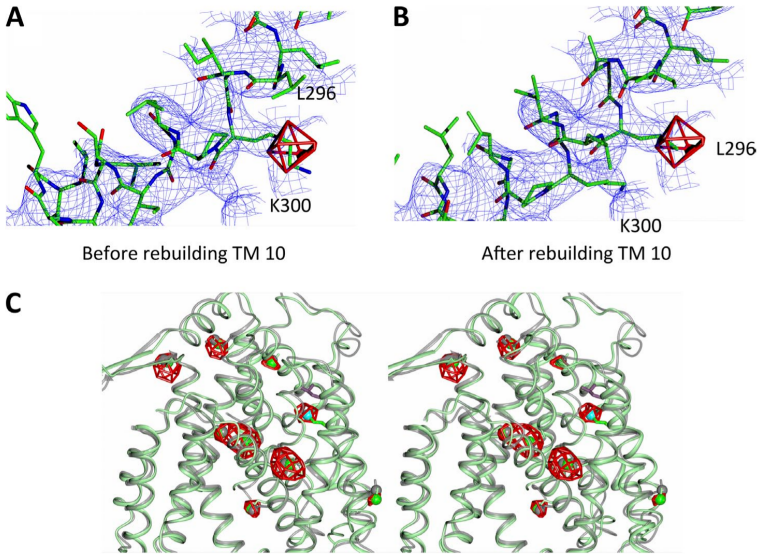


Figure 4.2: Electron density. (A) Electron density of TM 10. The 2mFoDFc map shown in blue (contoured at 1.5 σ) was calculated with phases derived from the model before reassigning the sequence of TM 10 (shown) and averaged over the four molecules of the asymmetric unit. The anomalous difference map shown in red has been calculated from the selenomethionine-derivatized triple mutant and contoured at 3.6 σ. (B) The same maps as in A but with the structure of the wild-type protein (refined before the data of the triple mutant were collected). (C) Stereo view of the superposition of the final refined structure of the triple mutant (green) on that of the published monomeric structure (gray) [22]. The anomalous difference map is shown as in A. Met296 is in cyan, and Leu296 from the published structure is in wine red. Overall, there is good correspondence in the position of the helices and methionines between the two structures, in agreement with the similar activity of the mutant to wild type (Figure 4.4). An enlarged view of TM 10 is shown in Figure 4.11.

The NhaA mutant structure was refined to an R factor of 28.7% and a corresponding R-free of 31.2% (Table 4.2). In both the wild-type and mutant NhaA structures, the associated maps are of reasonable quality (Figures 4.2, 4.5, and 4.11), especially after averaging of the four molecules of the asymmetric unit, although TM 4a, which is thought to move upon binding sodium ions [24], was difficult to model. Additional electron density at the dimer interface could not be assigned to protein residues. We have tentatively modeled this by sulfate ions and the detergent α-DDM. At the resolution of the data, there are no obvious differences to the wild-type structure, which was subsequently refined to an R factor of 31.8% and an R-free of 34.2%, with restraints to the higher resolution mutant structure (Table 4.2).

Crystal	Wild type	Triple mutant
Wavelength (Å)	0.9778	0.9793
Space group	$P2_1$	$P2_1$
Resolution (Å)	29.6-3.69 (3.80-3.70)[a]	56.5-3.5 (3.54-3.50)[a]
Cell dimensions	a = 115.8 Å; b = 100.6 Å; c = 141.6 Å; β = 97.0°	a = 115.8 Å; b = 99.4 Å; c = 140.2 Å; β = 97.4°
Number of measured reflections	117,235	313,049
Number of unique reflections	34,273	37,951
Completeness (%)	98.1 (90.0)	94.6 (69.5)
Redundancy	3.4 (3.0)	8.2 (7.2)
$I/\sigma(I)$	11.7 (1.2)	22.0 (1.4)
R_{merge} (%)	9.8 (82.0)	4.4 (111.4)
R factor (%)	31.8	28.7
R-free[b] (%)	34.2	31.3
RMSD from ideal values		
Bond lengths (Å)	0.010	0.004
Bond angles (°)	1.28	1.48
Ramachandran plot outliers[c] (%)	0.7	0.7

Table 4.2: Data collection and refinement statistics. [a]Values in parentheses refer to data in the highest resolution shell.
[b]Based on 5% of the reflections
[c]From MolProbity [111].

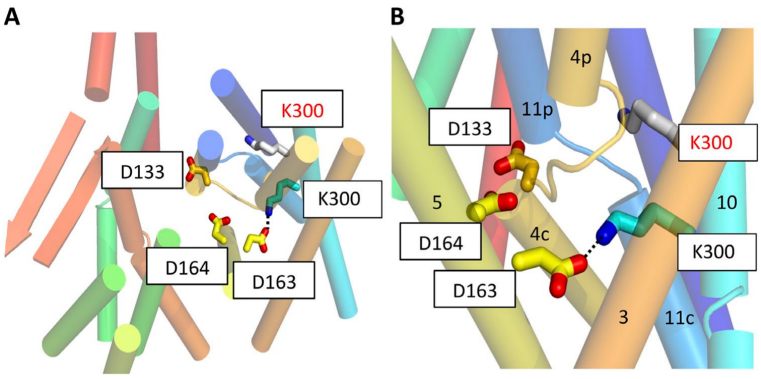


Figure 4.3: Position of Lys300 on TM 10. (A) Cartoon representation of the structure viewed from the periplasmic side of the membrane. The charged residues, Asp133 (TM 4b), Asp163 (TM 5) and Asp164 (TM 5), and Lys300 (TM 10), in the new structure are shown as colored sticks. The position of Lys300 in the previously published structure (Protein Data Bank accession no. 1ZCD) is shown in gray (red text). The structure has been colored from red at the N terminus through to blue at the C terminus. Loop regions except for the breaks between the discontinuous helices have been omitted for clarity. (B) As A, but from the side and zoomed in. The dashed line represents the salt bridge between Asp163 and Lys300.

4.3.3 The crystal structure of the NhaA dimer

The two dimers of the asymmetric unit are almost identical and superimpose with an RMSD of 0.3 Å (Table 4.5). The density is well defined for the β sheet, although some of the side chains on the periplasmic side are less ordered (Figure 4.5A). The overall structure of each subunit in the dimer is very similar to the published

monomeric structure with an RMSD of 1 Å for 358 out of 374 Cα atoms (Figure 4.5 and Table 4.5). The largest conformational difference lies in the position of the 14 amino acid β hairpins linking TMs 1 and 2, which protrude out along the membrane plane and form the predominant dimer interaction. Relative to the monomeric crystal structure, the hairpins twist by $\sim$15° to form a flat four-stranded antiparallel β sheet with residues 45-51 involved in hydrogen-bonding interactions between the strands (Figure 4.5C).

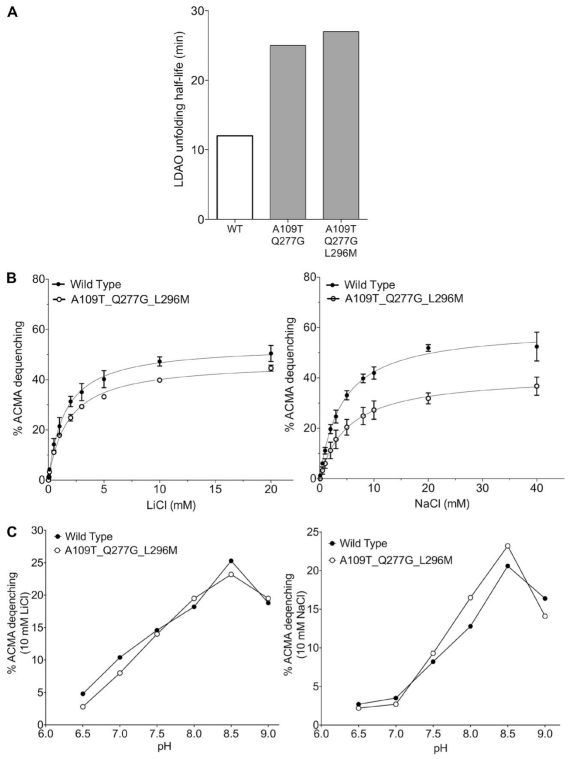


Figure 4.4: Stabilization, characterization, and crystallization of the NhaA mutant. (A) The NhaA double mutant (A109T and Q277G) and NhaA triple mutant (A109T, Q277G, L296M) are more stable in detergent as shown by the longer unfolding half-life ($t_{1/2}$) in LDAO at 40°C. (B) The ATP synthase and NhaA wild-type and mutants were co-reconstituted in liposomes. ATP-driven proton pumping establishes a ΔpH (acidic inside) as monitored by the quenching of 9-amino-6-chloro-2-fluorescence (ACMA). Proton efflux is initiated by the addition of increasing concentrations of NaCl/LiCl, and apparent ion-binding affinities for NhaA wild type (closed circle) and mutant (open circle) at pH 8.5 were calculated: $K_{\mathrm{M}}Na^+$ wild type (mean $\pm$ SD): 1.8 $\pm$ 0.2; $K_{\mathrm{M}}Na^+$ mutant: 1.6 $\pm$ 0.1; $K_{\mathrm{M}}Li^+$ wild type: 4.1 $\pm$ 0.5; $K_{\mathrm{M}}Li^+$ mutant: 4.1 $\pm$ 0.3. (C) pH dependence of NhaA $Na^+(Li^+)$-H^+ antiporter activity for wild type (closed circle) and mutant (open circle) were measured in proteoliposomes by the level of ACMA dequenching as in B at the indicated pH values after the addition of saturating NaCl/LiCl at pH 8.5; all experiments were repeated in triplicate and representative traces are shown.

The arrangement of the subunits in the dimer is very similar to the cryo-electron microscopy (EM) model (Protein Data Bank accession no. 3FI1)[18,24], which, in turn, is consistent with electron spin resonance[115] and cross-linking distance measurements[137]. The exact conformation of the β strands is, however, different. In the structure derived from the cryo-EM data, the β sheet is much more curved, with the tips of the hairpins $\sim$11 Å above the membrane surface[24], placing them parallel, but 7 Å above the β sheet that we observe in the dimeric crystal structure (Figure 4.5D). In this bent position, there are fewer hydrogen-bonding interactions between strands than in the dimer crystal structure.

The interfacial TM preceding the β hairpins and the alternating position of charged and noncharged residues of the four-stranded antiparallel β sheet creates a dimer interface that is amphipathic, with polar residues pointing toward the periplasm and hydrophobic residues toward the protein or the dimer interface (Figure 4.5C). Lys57, at the C terminus of the β hairpin, is the only residue that breaks this trend by pointing into the protein where it makes a salt bridge with Asp65. The Lys57-Asp65 pairing has been proposed to be important, as replacement of either Lys57 or Asp65 by cysteine results in a protein with an apparent K_m for Na^+ that is four- or 10-fold higher than wild type, respectively[26,137]. These residues are 9 Å apart in the model derived from cryo-EM maps.

Other than the interaction between the β hairpins, there are few direct contacts between monomers. Trp258 of TM 9 makes a bridge to TM 7, and Arg204 of TM 7 may interact with the amide oxygen of Val254. It is likely that lipids would fill the space between the protomers in vivo, and in fact, in the crystal structure a detergent molecule has been modeled in extra density at this position.

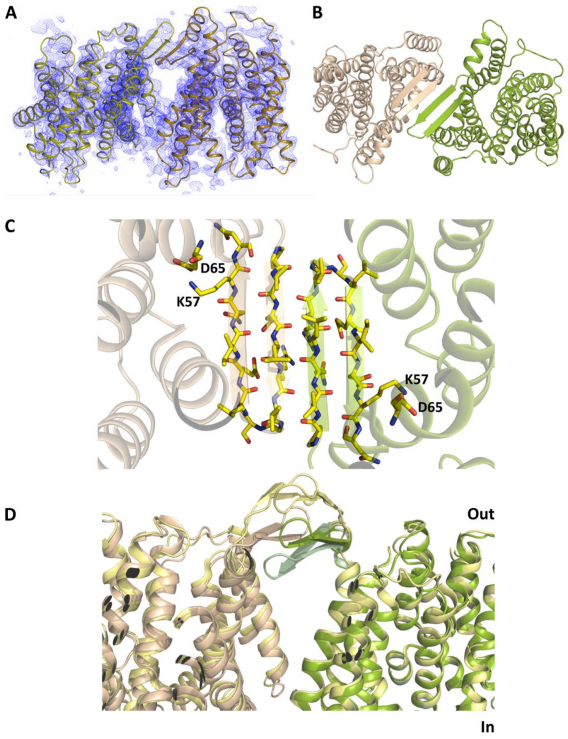


Figure 4.5: Position of β hairpins in the NhaA dimer. (A) 2mFo-DFc electron density averaged over the four molecules of the asymmetric unit (contoured at 1.5 σ). The map was calculated directly after molecular replacement using a search model where the β hairpins and loops had been omitted. (B) Cartoon representation of the NhaA dimer viewed from the periplasmic-facing side of the membrane, with the two protomers shown in light brown and green, respectively. (C) The 14-amino acid β hairpins from each promoter, shown in light brown and green, form a four-stranded antiparallel β sheet, as viewed from the cytoplasmic side of the membrane. The residues facing the membrane are hydrophobic except for Lys57. (D) Comparison of the positions of the β hairpins in each of the determined NhaA structures: crystal structure of the dimer (this work; light brown and green), monomeric crystal structure [22] (pale green), and dimer from cryo-EM [24] (yellow).

4.3.4 MD simulations of the NhaA dimer

To gain further insight into the position of the NhaA dimer in the membrane, the dimer was simulated in a POPC model membrane for 1 μs. As described in Section 4.2, a multi-scale approach [50,103] was used that provides sufficient sampling of lipid-protein interactions in the absence of any knowledge of the location of the protein in the membrane. The Cα RMSD of each monomer relative to the crystal structure increased to $\sim$3.4 Å over the time course of a 1-μs simulation (Figure 4.12A). The RMSD of the TM helices alone stabilized around 2.8-3.0 Å, whereas mobile loops contributed substantially to the observed differences from the crystal

structure (Figure 4.12, B and C). Secondary structure elements such as the β sheet were retained over the whole simulation, although a few regions are less structured: in particular, TM 4a, which was difficult to build in the crystal structure, and to a lesser degree 4b; TM 5 near Asp163 and Asp164 (usually when sodium binding was also observed); and TM 10 near Lys300. NhaA, simulated either as a dimer or as a monomer, sits entirely within the membrane, including the β hairpins (Figure 4.6). It does not extend beyond the lipid head groups, as shown by a density profile along the membrane normal (Figure 4.12D). The periplasmic face of the β sheet was fully exposed to solvent and confined the periplasmic interfacial lipids.

4.3.5 MD simulations of sodium binding with different protonation states of the key residues

The roles of the critical charged residues (Asp164, Asp163, Lys300, and Asp133) have been investigated previously by a combination of biochemical and biophysical experiments and MD simulations[14,19,21–23,25,28]. As the outcome of the MD simulations are likely to be affected by the proximity of Lys300 to Asp163 (distance of 2.5 Å in this structure compared with 12 Å in the published structure), we sought to use our new structure in MD simulations to investigate (a) how the protonation states of these four critical residues affect sodium ion binding, and (b) how the protonation state and presence of sodium ions affect the salt-bridge interaction. MD simulations were performed in the presence of sodium ions at a NaCl concentration of ∼100 mM with the residues in different protonation states (Table 4.1).

First, we examined the most likely situation in the low pH crystal structure, with Asp164 protonated (neutral) and Asp163 deprotonated (negatively charged) because it interacts with the positively charged Lys300 (see Section 4.2). In simulations with these protonation states, no sodium ion binding was observed after 1 μs of

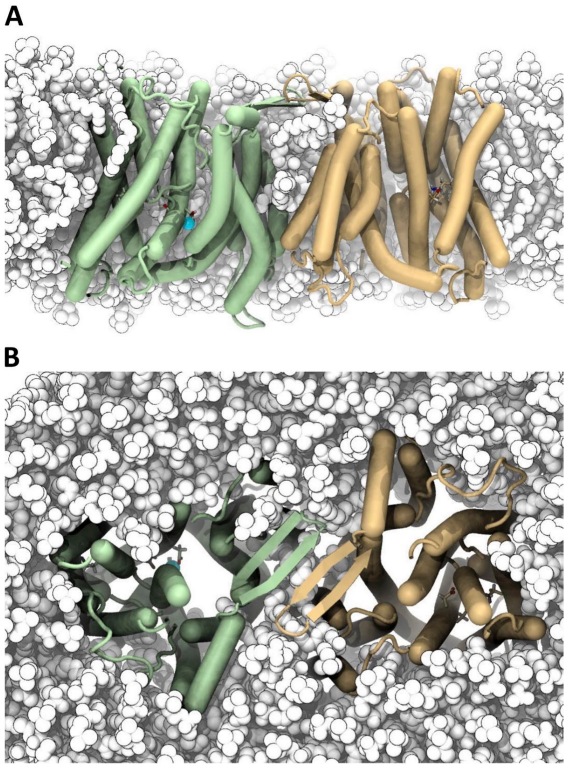


Figure 4.6: MD simulation of the NhaA dimer in a model membrane bilayer. (A) The NhaA dimer is stable in a POPC membrane. Snapshots show the structure after an ∼1-μs MD (simulation S2/1). The periplasmic space is at the top. Protomers are colored wheat and green, and POPC lipids are white. A bound sodium ion is shown as a cyan sphere. Residues Asp163, Asp164, and Lys300 are partially visible in a stick representation. (B) View from the periplasm.

sampling (Figure 4.7A). The Asp163-Lys300 salt bridge remained intact with an average distance of ∼2.4 Å (Figure 4.7B). Similar results were obtained in two separate repeat simulations of 100 ns (Figure 4.13).

Next, we modeled Asp164 in a deprotonated state, as would be expected at high pH where the protein is active. Three independent simulations of 1 mus of the dimer and seven shorter simulations of 100 ns of just the monomer were performed (Figures 4.7, 4.14, and 4.15). In all but one protomer of the dimer simulations and one of the monomer repeat simulations, sodium ions were seen to spontaneously enter the vestibule and bind to Asp164. In two cases, Asp163 switched from forming a salt-bridge interaction with Lys300 to interacting with the sodium ion. On the time scales

simulated here, the salt bridge between Asp163 and Lys300 typically persisted during the whole length of the simulation, as also observed for the simulations with Asp164 protonated, unless it was destabilized by the presence of a sodium ion (Figures 4.7, 4.14, and 4.15). With both Asp163 and Asp164 modeled in their protonated (neutral) form, no sodium ion binding was observed during 1 μs of MD and two additional 100-ns simulations (Figure 4.16).

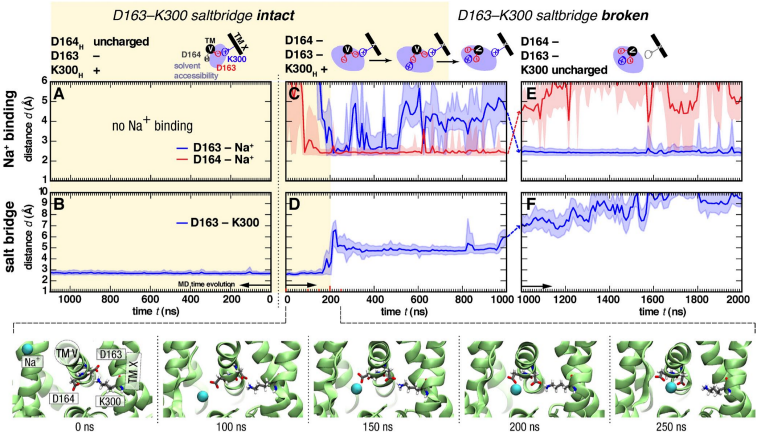


Figure 4.7: Sodium ion binding and salt-bridge stability in MD simulations with different protonation states of Asp163, Asp164, and Lys300. (Left column; A and B) Asp163 deprotonated and Asp164 and Lys300 protonated (simulation S1/1, protomer B). (Middle column; C and D) Asp163 and Asp164 deprotonated and Lys300 protonated (simulation S2/1, protomer B). (Right column; E and F) Asp163, Asp164, and Lys300 deprotonated (simulation S4/1, protomer B). (Top row; A, C, and E) Distances between the closest sodium ion and Asp163 or Asp164 are plotted as a function of time. Spontaneous Na^+ binding to Asp164 was observed when both aspartates were deprotonated. (C) Continuation of the simulation with Lys300 deprotonated (a 3-ns equilibration simulation with position restraints on all heavy protein atoms is symbolized by dashed lines between panels) leads to a rapid change in the Na^+-binding mode toward closer interaction with Asp163. (Bottom row; B, D, and F) Distance of the closest Asp163 carboxyl group from the N-amino group of Lys300. Distances <4 Å are indicative of a stable salt-bridge interaction (yellow shaded area), whereas those ≥4 Å are considered a weak or broken salt bridge. Binding of Na^+ to Asp164 destabilizes the salt bridge. (D) Lines show data averaged over blocks of 10 ns, with fluctuations in the data indicated as shaded regions encompassing the lower 5 and upper 95% percentile. The snapshots show the Na^+ binding event with subsequent rupture of the salt bridge (cytoplasmic view along the axis of helix TM 5). Videos 1 and 2 show this simulation. Repeat simulations (see Table 4.1) are shown in Figures 4.13-4.17.

We next investigated the effect of changing Asp133 and Lys300 to their neutral states. In the above simulations, Asp133 was deprotonated. When it was protonated (neutral), no spontaneous sodium binding was observed (in one 1.1-μs and two 100-ns simulations; not depicted), indicating that perhaps Asp133 is important to capture ions. Lys300 is unlikely to be deprotonated when involved in a salt-bridge interaction. However, it is possible that disruption of the salt bridge by sodium ion binding enables

deprotonation of the lysine (see next section). The simulations also suggest that Lys300 is water accessible, as shown in Figures 4.14 and 4.16. To investigate the consequences of a neutral Lys300 in this situation, the endpoint of the simulation shown in Figure 4.7 (C and D) with a bound sodium ion and the salt bridge broken was continued with Lys300 modeled as depro- tonated (Figure 4.7, E and F). The sodium ion transferred from being bound by Asp164 to binding to Asp163 while the distance between Asp163 and Lys300 increased further, a behavior reproduced in two repeat simulations (Figure 4.17).

4.3.6 Estimations of pK_a for the charged residues

To obtain estimates of the pK_a of the residues putatively involved in sodium ion and proton binding, we used the program PROPKA 3.1[131]. Figure 4.18C shows the pK_a of the residues in the structures along the time course of the simulation S2/1. While the salt bridge remains intact, the pK_a of the four residues remains fairly constant. Upon breaking, however, the pK_a of Lys300 drops markedly and that of Asp163 increases. As the heuristic methods used in PROPKA can be quite sensitive to small changes in the environment of the residues, we decided to study the distributions of pK_a values of the residues from structures from all of the simulations separated into those where the salt bridge was intact and those where it had broken (Figures 4.8 and 4.19). The largest effect is seen for Lys300, where the pK_a of Lys300 is lowered by approximately two to three pH units to pH $\sim$8.5 (Figures 4.8, 4.19, and 4.20, and Table 4.3). A detailed assessment of the effect of breaking the salt bridge together with a sensitivity analysis of the pK_a calculations are given in the supplemental Discussion.

4.4 Discussion

The crystal structure of the NhaA dimer is consistent with the 7-Å cryo-EM structure derived from 2-D crystals grown under more native-like conditions, although a difference in the position of the β sheet is observed. In the crystal structure, the amphipathic β sheet is flat and sits in the plane of the membrane, with one surface exposed to solvent and the other shielding lipids (Figures 4.5C and 4.6). In this position, unlike in the structure from cryo-EM, Lys57 and Asp65 are close enough to form a salt bridge, an interaction that is reported to be physiologically significant [26]. Biochemical data indicate that the monomer is the functional unit of the protein, with dimerization probably enhancing its stability in the membrane [70,138]. In agreement with this, we do not observe significant differences within the TM region of the dimeric structure compared with the previous monomeric structure. The amphipathic β sheet may help to anchor the dimer domain as the core domain undergoes large conformational changes during the transition between outward- and inward-facing states, as suggested by the recent structure of NapA [31].

TM 10 was reportedly difficult to build in determining the original crystal structure of NhaA [74]. Based on the density observed in the wild-type dimer structure, we revised the sequence assignment of this helix and subsequently confirmed its reassignment using anomalous scattering of a selenomethionine residue that replaced the naturally occurring Leu296. This process appears to have been greatly facilitated by the use of a thermostabilized construct of NhaA. In the modified structure, Lys300 is displaced 10 Å from its previously modeled position to lie within hydrogen-bonding distance of Asp163, similar to the interaction reported for the active pH NapA structure [31] (Figure 4.9). Although a water-mediated interaction between Lys300 and

Simulation[a]	Asp133	Asp163	Asp164	Lys300
S1	-0.4 ± 0.7	0.8 ± 0.7	-0.1 ± 0.7	-0.5 ± 0.8
S2	-1.5 ± 1.8	0.6 ± 1.3	-0.4 ± 1.6	-1.9 ± 1.6
S4	-1.0 ± 1.1	1.9 ± 1.2	0.2 ± 1.4	-2.1 ± 1.2
S3	-1.4 ± 0.8	3.0 ± 1.5	0.6 ± 1.6	-3.7 ± 1.4
S5	-0.1 ± 0.7	0.6 ± 0.9	0.5 ± 0.7	-1.1 ± 0.8
S6 + S8	0.4 ± 2.1	2.1 ± 2.1	-0.0 ± 1.6	-2.8 ± 1.5
S7	-0.1 ± 0.6	0.3 ± 0.9	1.3 ± 1.2	-1.2 ± 0.8
All data combined[b]	-1.5 ± 1.9	1.5 ± 1.8	-1.3 ± 2.4	-2.5 ± 1.4
All data combined, Na^+ excluded from calculation[c]	-1.3 ± 1.5	2.6 ± 1.9	-1.0 ± 2.1	-2.4 ± 1.4
All data combined, salt-bridge cutoff 3.5 Å[d]	-1.4 ± 1.9	1.4 ± 1.8	-1.3 ± 2.4	-2.5 ± 1.5
All data combined, $\varepsilon_{surface} = 80$[e]	-1.4 ± 2.0	1.5 ± 1.9	-1.0 ± 2.5	-2.5 ± 1.5

Table 4.3: Estimated pK_a shifts due to breaking of the salt-bridge Lys300-Asp163 [a]For simulation names, see Table 4.1. Data were aggregated for all repeats and protomers A and B. Protein conformations were sampled every 1 ns. Na^+ ions were included within 6 Å of the protein, and the salt bridge was considered broken if the distance was >4 Å (except where noted).
[b]Distributions of pK_a and ΔpK_a were recomputed from all trajectory frames of all simulations (i.e., simulated with differing protonation states), assuming that each frame was an independent sample of the protein conformation. This can lead to distributions that are broader than those for individual simulation sets with means that are not just a simple average of individual means.
[c]Na^+ ions were not included in the PROPKA 3.1 calculation.
[d]Salt bridge was considered broken if the distance was *ge*3.5 Å.
[e]The dielectric constant near the protein surface (corresponding to the solvent-exposed residues) is 160 in the standard PROPKA 3.1 parametrization. A value of 80 was tested, as suggested for calculations on NMR ensembles [131].

Asp163 was predicted for NhaA at high pH [25], a direct interaction observed already at low pH means we must reconsider the most likely steps of the transport cycle.

In MD simulations, sodium ions bind spontaneously to both Asp164 and Asp163, consistent with recent biochemical and biophysical experiments suggesting that both of these residues are likely to coordinate sodium ion binding [28]. During our simulations, the stability of the salt bridge is clearly affected by the interaction of the sodium ions with the aspartate residues. Asp163 either maintains its interaction with Lys300 or moves toward the positively charged sodium ion, thereby breaking the salt bridge. This suggests a competition between the lysine and sodium ion to bind to Asp163. The TM-embedded Lys300 is important for both activity and pH regulation [21]. Of the various mutations that have been made, replacement with arginine and histidine, where the positive charge is maintained, causes the least perturbation [28]. Nevertheless, even these mutations show a marked drop in activity and a substantial decrease

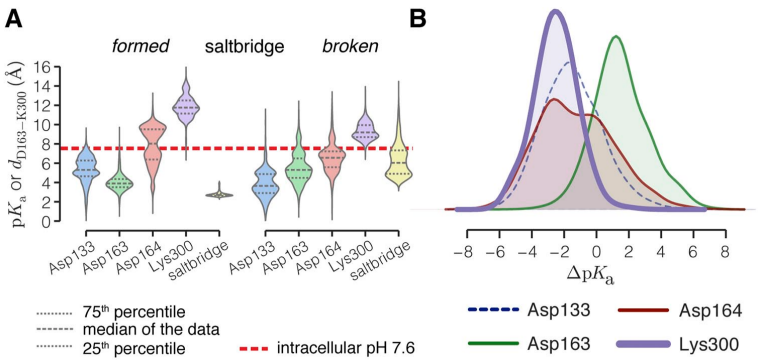


Figure 4.8: Effect of breaking the Asp163-Lys300 salt bridge on the pK_a of conserved residues. (A) Distributions of pK_a values estimated with PROPKA 3.1[131] from all MD simulations shown as violin plots. pK_a values were calculated from snapshots extracted at 1-ns intervals, with sodium ions included within 6 Å of the protein. For reference, a dotted line has been drawn at pH 7.6. The percentiles of the data are shown as broken lines inside the distributions, which were computed as Gaussian kernel density estimates (see Section 4.2). Data were split depending on the state of the salt bridge (formed if the distance is <4 Å, and broken if it is ≥4 Å; the salt bridge distance distribution is also shown in yellow). The multimodal distribution of Asp164 when the salt bridge is formed is a result of data from simulations S1 (Figure 4.19A) during which its solvent accessibility is much smaller than during other simulations (see also Figure 4.13, M-O). (B) Distributions of pK_a shifts (ΔpK_a) when the salt bridge is broken (data aggregated over all MD simulations). The pK_a of Lys300 is downshifted by -2.5 ± 1.4, and thus the charged form is destabilized. Shifts of the other residues are not significantly different from 0 (also see Table 4.3).

in the affinity (K_M) for sodium ions (100- and 30-fold for arginine and histidine, respectively), and lysine at this position is clearly preferred.

NhaA transports two protons into the cytoplasm for every sodium ion exported. There has been some discussion in the literature about which residues are likely to form the proton-binding sites. The presence of the Asp163-Lys300 salt bridge necessarily affects this debate. Asp164 is strictly conserved from bacteria to mammals and seems very likely to bind one of the protons[27,28,31]. The second proton has been suggested to bind to either Asp133 or Asp163[23,139]. Of the two, Asp163 seems the more likely: it is essential for activity[14,28], conserved in the electrogenic transporters, and replaced by asparagine in the electroneutral antiporters[140,141]. Asp133 is not essential for activity[19,28], is generally less well conserved, and is retained in some electroneutral sodium-proton exchangers[140]. However, the estimation of pK_a by PROPKA 3.1 (Figures 4.8 and 4.19) suggests that Asp163 is unlikely to be protonated when involved in a salt bridge to Lys300. As a salt bridge is observed in both

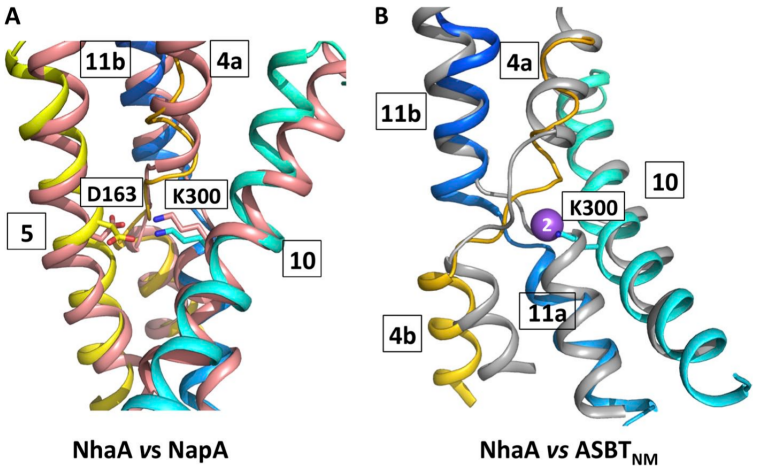


Figure 4.9: Comparison of NhaA with NapA and ASBT$_{NM}$. (A) Structural comparison of NhaA and NapA. Cartoon representation of TMs 4, 5, 10, and 11 of NhaA superimposed on the same helices of NapA [31]. Residue numbering is shown for NhaA. K300 in NhaA corresponds to K305 in NapA, and D163 in NhaA corresponds to D156 in NapA. The helices of the TMs in NhaA have been colored as in Fig. 3, and NapA is colored salmon. (B) Structural comparison of NhaA and ASBTNM. Cartoon representation of TMs 4, 10, and 11 of NhaA superimposed on TMs 4, 8, and 9 of ASBTNM [75]. The helices of the TMs in NhaA have been colored as in A, and ASBTNM is colored gray. The sodium Na2 site in ASBT$_{NM}$ is depicted as a purple sphere, and Lys300 residue in NhaA is shown as a stick model.

the inward-facing NhaA structure and the outward-facing structure of the homologous NapA, it seems that if Asp163 is involved in the translocation of the proton, this will only be transiently. Another candidate for proton transport, which has not been considered up to now, is the lysine. Lysines have been proposed to be involved in proton translocation in other membrane transporters [142,143]. Without putting too much emphasis on the absolute values calculated by PROPKA, the analysis of the pK_a of Lys300 would suggest that it would be protonated when the salt bridge was present but is more likely to lose its proton when disrupted (Figures 4.8 and 4.20, and Table 4.3; see also supplemental Discussion). In the electroneutral transporters, the lysine is replaced by an arginine [140,141]; the pK_a of arginines is known to be much less altered by the environment than that of a lysine [144,145]. In our simulations, the protonation state of Lys300 clearly affects the interaction of the sodium ion with the protein, whereby a deprotonated Lys300 facilitates association of the two conserved aspartates with the ion. It is also noteworthy that the amino group of the

repositioned lysine is located at the same position as one of the sodium ions in the sodium-dependent bile acid symporter ASBTNM, a structural homologue of NhaA and NapA[75] (Figure 4.9B). A similar comparison was used to support the role of a lysine in proton transfer in another family of transporters[142]. Given the delicate positioning of charged residues at this site, however, detailed biochemical and structural information needs to be gathered before any firm conclusions can be drawn.

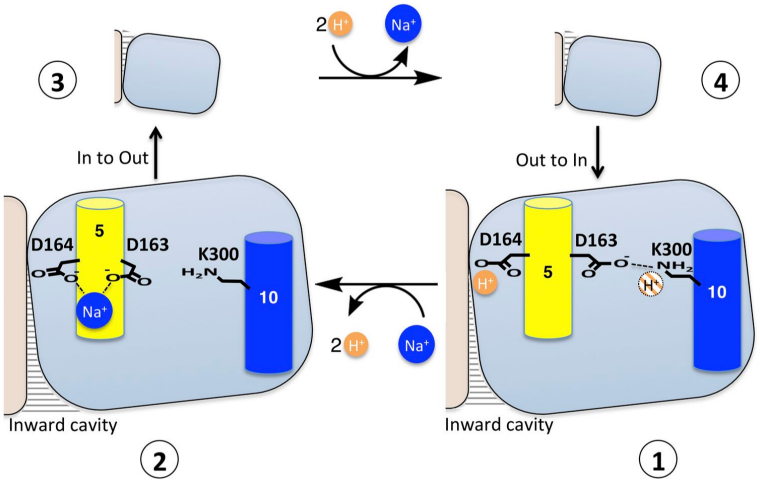


Figure 4.10: Proposed schematic model of NhaA transport. The transport mechanism occurs through a reorientation of the protein, alternately exposing a cavity containing Asp164 to the intracellular or the periplasmic space. The core domain is depicted as a blue square, and the panel domain is in beige. (1) The probable situation in the solved NhaA structure at low pH where the protein will be protonated. One of the protons (in orange) is likely to be bound to Asp164, and we hypothesize that the second proton (orange hashed) will interact with Lys300. Sodium ions (blue) will compete with protons for binding to Asp164, causing the Asp163-Lys300 salt bridge to break and possibly Lys300 to lose a proton (2). The sodium ion-bound protein will then switch to the outward-facing state (3). Upon release of the sodium ion, the protein will be reprotonated and the salt bridge between Lys300 and Asp163 reformed (4). The protein will then switch back to the inward-facing conformation.

Combining this new structural information on NhaA with the recent structure of NapA allows us to postulate a mechanism whereby the interaction between Asp163 and Lys300, seen in both inward- and outward-facing states, must be taken into consideration. At low pH, as seen in the crystal structure of NhaA, Asp164 is likely to be protonated and Asp163 is involved in a salt-bridging interaction with Lys300 (Figure 4.10, 1). As the pH becomes more alkaline, sodium ions will successfully compete with protons for binding to Asp164, in line with a recent electrophysiology

study showing that the pH dependence of NhaA can be explained by the competition of sodium ions and protons to bind to the same site (Fig. 4.10, 2)[27]. At this point, as demonstrated by the MD simulations, Asp163 can switch from interacting with Lys300 to contribute to binding the sodium ion. The substrate-bound form of the transporter would then switch to the outward-facing conformation where the sodium ion can be released and the salt bridge can reform, as seen in the structure of NapA[31] (Figure 4.10; 3 and 4). In the diagram, we also show the possible deprotonation and protonation of Lys300, as discussed above. Upon proton binding to Asp164, the transporter would then switch back to the inward-facing conformation.

In summary, the structure of the NhaA dimer here and in particular the repositioning of TM 10 provides critical new information that must be incorporated into the mechanism of electrogenic sodium-proton antiport. Clearly, the crystal structure of the sodium-bound form of NhaA or a homologue, in combination with further biochemical, biophysical, and MD simulation studies, is required to elucidate how binding and unbinding of ions is ultimately coupled to the conformational changes observed during transport.

Acknowledgements

We are grateful to Etana Padan, Carola Hunte, and Gunnar von Heijne for useful discussions. The authors are grateful for the use of the Membrane Protein Laboratory funded by the Wellcome Trust (WT089809) at the Diamond Light Source, where data were collected. Computer simulations were carried out on the A2C2 Saguaro supercomputer at Arizona State University.

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant OCI-1053575 (allocation MCB130177 to O. Beckstein). This work was supported with grants from

The Royal Society through the University Research Fellow scheme (to D. Drew), the Medical Research Council (MRC_G0900990(91997) to A.D. Cameron and D. Drew), the Swedish Research Council (to C. von Ballmoos and D. Drew), and the Biotechnology and Bio- logical Sciences Research Council (BB/G023425/1 to S. Iwata). D. Drew acknowledges the support from EMBO through the EMBO Young Investigator Program. Part of the work was funded by the ERATO Iwata Human Receptor Crystallography Project, Japan Science and Technology Agency and the EU (European Drug Initiative for Channels and Transporters, EDICT grant 201924).

The authors declare no competing financial interests.

4.5 Supplementary Information

4.5.1 Estimates for pK_a values

Accurate calculation of pK_a values is a computationally very challenging task. Here we use a fast heuristic scheme as implemented in PROPKA 3.1[131] to semi-quantitatively estimate the pK_a values of the conserved residues Asp133, Asp163, Asp164, and Lys300. Although not necessarily exact, the pK_a estimates are nevertheless thought to report on the chemical environment around titratable residues, and in particular, relative changes in the chemical environment will be reflected in shifts in the pK_a that are at least qualitatively indicative of likely protonation-state changes. pK_a values were calculated from frames of all MD trajectories at 1-ns intervals. PROPKA 3.1 can take ligands into account, so any Na^+ ion within 6 Å of the protein was included in the calculation. The resulting time series (Figure 4.18, B-D) shows fluctuations >2-3 pK_a units, which indicates that the pK_a is sensitive to the exact molecular conformation. However, in aggregate, a more consistent picture emerges, as shown by the distributions of pK_a values (Figure 4.19). When the data

are grouped by the state of the Asp163-Lys300 salt bridge (for distance <4 Å, the salt bridge is considered formed or intact, whereas at ≥4 Å, it is considered broken; see Figure 4.18A), the effect of the salt bridge on the pK_a becomes apparent. In general, breakage of the salt bridge reduces the pK_a of Lys300 and raises the pK_a of Asp163. These are predicted to be around 10.5 and 4, respectively, when the salt bridge is intact. The pK_a of Asp164 is predicted to be in the range of 6 to 10 (Figure 4.19), with the high values only appearing in simulations S1 (Figure 4.19A), when the solvent accessibility of Asp164 is decreased compared with the other simulations (Figure 4.13, M-O). The pK_a of Asp133 is typically shifted down, but this is unlikely to have any substantial effect within the normal operating range of NhaA (pH 6.5-8.5) because the pK_a of Asp133 with the salt bridge intact is already around 5.

These absolute pK_a estimates have a high level of uncertainty. Relative changes, however, might be a better indicator for the effect of changes in the chemical environment of a titratable residue. We therefore focus on the shift in pK_a when the salt bridge is broken, ΔpK_a. The distribution of ΔpK_a is calculated from the pK_a distributions with and without the salt bridge (see Section 4.2 in the main text) and is shown in Figure 4.20. The means and standard deviations of these distributions allow for quick comparisons between different states (see Table 4.3 in the main text). The effect of the salt bridge is clearly strongest for Lys300. The loss of the negatively charged carboxylate moiety of Asp163 reduces the pK_a of Lys300 by around 2 units (see Table 4.3, in particular simulations S2 and S4). It is the only residue for which the difference distribution (over all MD simulations) differs from 0 (i.e., no effect) by more than one standard deviation, indicating a significant stabilizing influence of the salt bridge on the charge state of Lys300. With a base pK_a on the order of 10.5 when the salt bridge is intact, such a shift brings Lys300 within the operating pH range of NhaA. Asp163 is shifted upwards but to a lesser degree (around 1.5 units), and

because its base pK_a is low (∼4), such a shift is less likely to result in a protonation-state change. Repeating the pK_a calculations while ignoring all sodium ions in the calculation showed that Asp163 is the only residue whose charge state appears to be substantially stabilized by a Na^+ ion (by about -0.9) when all data are considered in aggregate (see Table 4.3 in the main text); for all the other residues, the difference to the calculation including Na^+ was ≤0.3.

4.5.2 Sensitivity of the pK_a calculations

The effect of including Na^+ ions in the calculation was important, as shown by a stabilization of Asp163 by ∼0.9; analysis of individual snapshots showed that a Na^+ ion could contribute up to 2 pK_a units. (Cl^- ions were never close to the residues of interest and could be excluded.) The calculations were insensitive to using a slightly different criterion for the salt-bridge existence (a cutoff of 3.5 Å instead of 4 Å, as chosen in Figure 4.18A), which resulted in a difference of 0.1 or less (Table 4.3 in the main text). Similarly, changing the surface (or solvent) dielectric constant $\varepsilon_{\text{surface}}$ from its default value of 160 in the standard PROPKA 3.1 parametrization to a value of 80 only had a very small effect of ≤0.3, as seen in Table 4.3 in the main text. The change in the dielectric constant was motivated by the suggestion that ensembles of structures derived by NMR (and by extension, MD) might show fewer surface salt bridges, and thus a lower $\varepsilon_{\text{surface}}$, more similar to the solvent dielectric constant, might be appropriate to evaluate the Coulomb contribution to the pK_a [131].

In summary, the calculated distributions of pK_a values are robust against relatively drastic parameter changes, as long as ions are explicitly take into account. The absolute accuracy of these values will be assessed by more sophisticated approaches in the future.

4.5.3 Movies from MD simulations

The first 250 ns of the 1-μs MD simulation trajectory S2/1 is shown in Videos 1 and 2. The coordinates were processed with a low-pass filter that removes rapid motions on the sub-ns timescale using the Gromacs tool `g_filter`. The simulations were performed with explicit solvent and membrane, but solvent, membrane, and all ions except for an ion participating in a binding event were omitted for clarity.

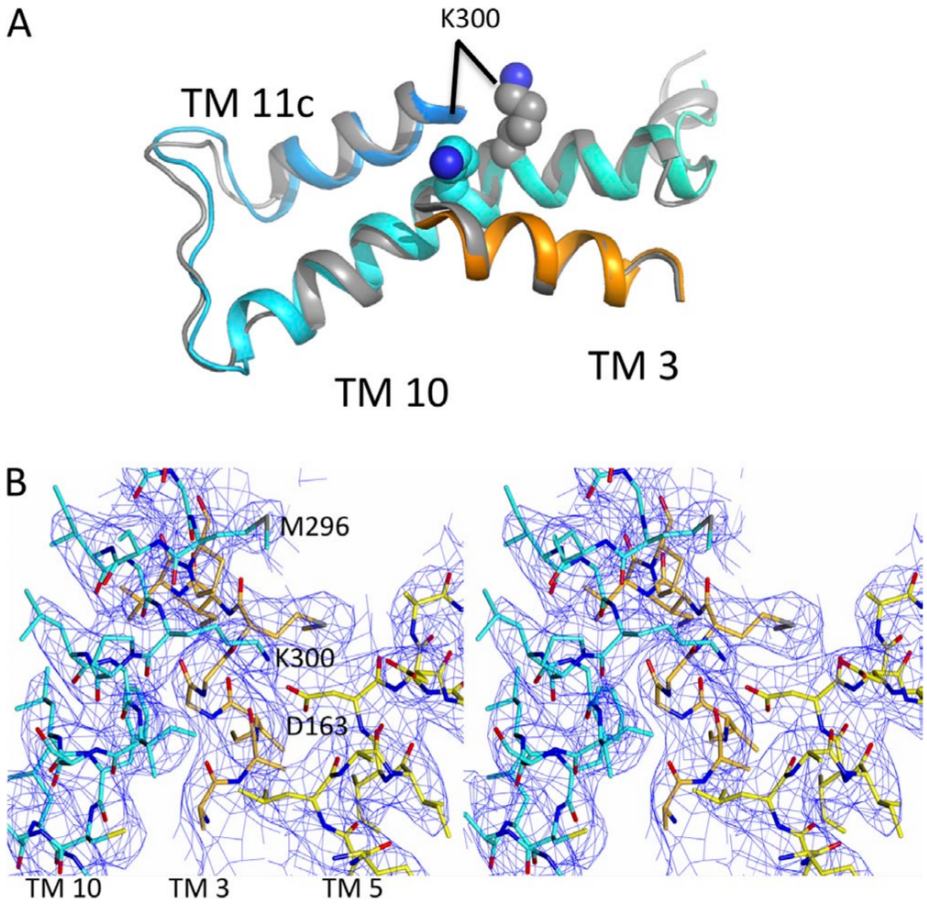


Figure 4.11: Sequence assignment of TM 10. (A) Superposition of TM 10 in cyan from the wild-type dimer structure on that of the same region from the published monomeric structure [22] (gray). The superposition was performed over all Cα atoms of the protomer. The Cα atom of Lys300 is displaced by one turn of the helix between the two structures. (B) Stereo diagram showing the 2mFo-DFc density associated with the refined structure of the triple mutant in the region of TM 10 and the Lys300-Asp163 salt bridge. TM 3 has orange carbon atoms, TM 5 has yellow, and TM 10 has cyan.

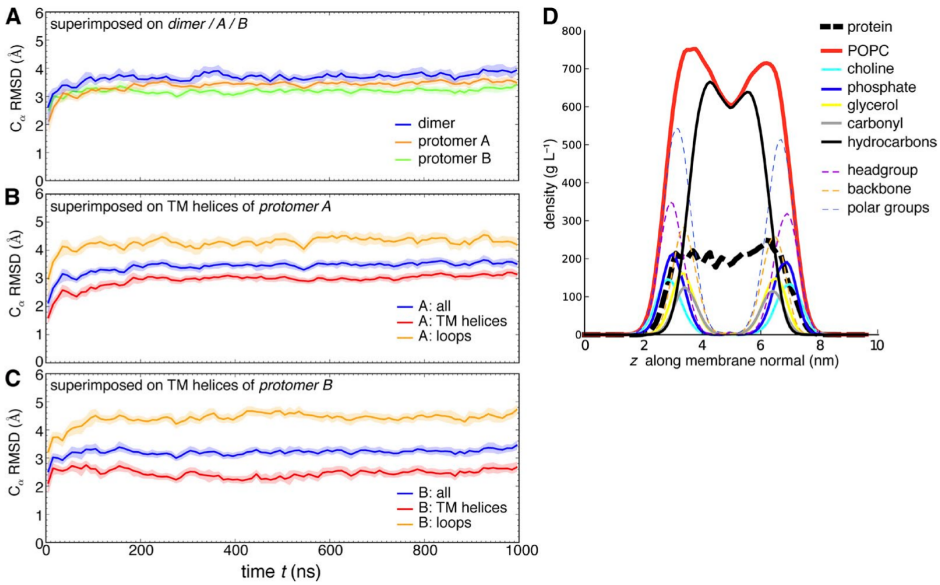


Figure 4.12: Analysis of proteins and lipids in a 1-μs simulation of the dimer (simulation S2/1). (A-C) RMSD of the NhaA dimer in membrane bilayer simulations. The coordinate root-mean-square distance of protein Cα atoms from the crystal structure was plotted as a function of time. Light bands indicate fluctuations between the 5th and 95th percentile around the average. (A) Blue, the dimer was superimposed on the crystal structure, and the RMSD was calculated for the whole dimer; orange, RMSD for protomer A only (superimposed on protomer A of the crystal structure); green, protomer B only. (B) RMSD of components of protomer A: blue, the whole protomer; red, only the TM helices (as defined by Hunte et al. [22]); orange, only the loops between the TM helices for protomer A superimposed on the TM helices. (C) RMSD of components of protomer B. (D) Lipid-protein density profile from MD simulation. The whole system was centered on the membrane. The density for components of the system was computed as averages in the plane of the membrane along the membrane normal z in 100 slices of thickness $\Delta z = 0.093$ nm. The periplasmic leaflet is marked by the peak around $z = 3$ nm, and the cytosolic leaflet is located at $\sim$6.5 nm. The protein is embedded in the membrane with the surfaces of NhaA lying inside the lipid head group region of the membrane. The graphs for protein and POPC are the totals for each group of molecules. Each POPC lipid molecule contains polar groups and the hydrocarbon core. Polar groups consist of the head groups (choline, phosphate) and the backbone (glycerol, carbonyl). Data for S2/1 are typical for other simulations not shown here.

Solubilization efficiency of NhaA constructs

Detergent solubilization efficiency in 2% (wt/vol) OG	Number of NhaA constructs
10–20%	20[a]
20–30%	17
30–40%	6
>50%	2

Table 4.4: Solubilization efficiency of NhaA constructs. [a]Extraction efficiency of wild-type NhaA is between 10 and 20% in 2% OG.

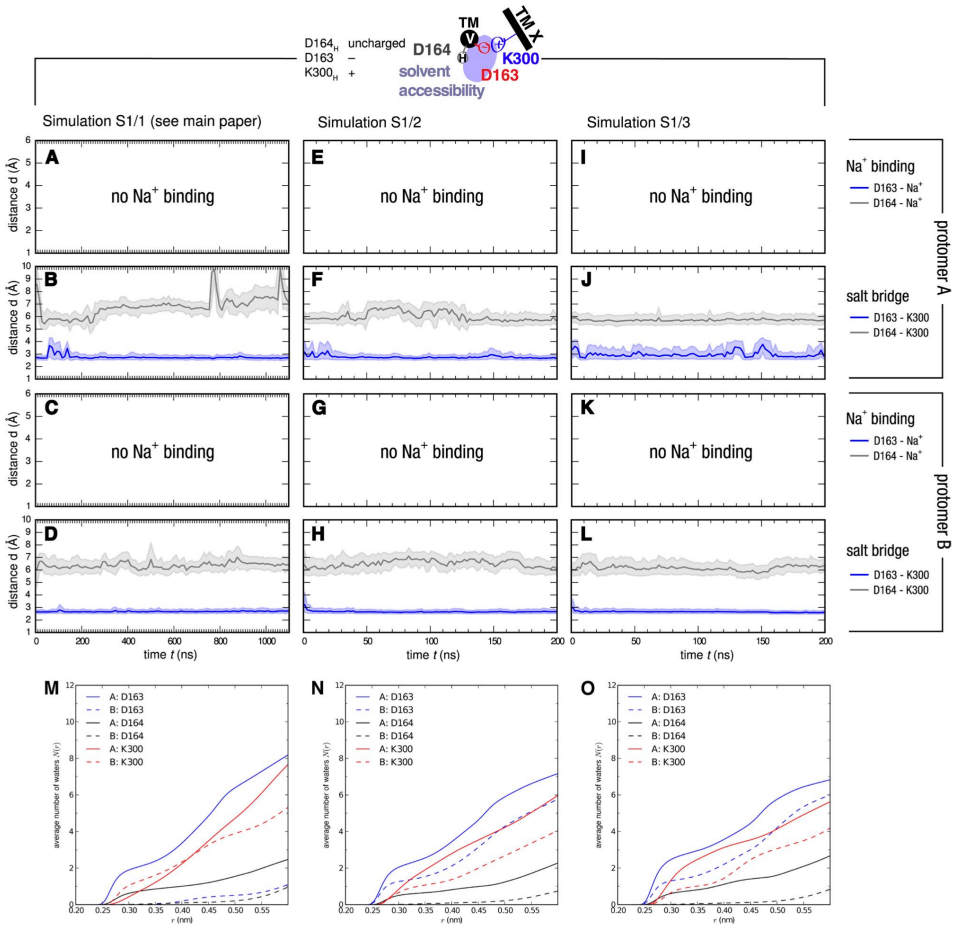


Figure 4.13: Analysis of simulations S1 (charged Asp163, neutral Asp164, charged Lys300). Results are shown for each protomer (labeled A and B). (Top; A-L) Distances between sodium ions and aspartate oxygen atoms (no binding observed in any simulations) and salt bridge distances between Lys300, Asp163, and Asp164. (Bottom; M-O) Solvent accessibility of Asp163, Asp164, and Lys300, measured as the average number of water molecules as a function of distance from the aspartate carboxyl oxygens or the lysine amine nitrogen. Cartoon images illustrate the key characteristics observed in the simulations.

RMSD in positions of C_α atoms among the structures

Chain	Monomer				Dimer	
	Chain B	Chain C	Chain D	1ZCD[a]	Dimer CD	1FI1[b]
Chain A[c]	0.3	0.0	0.3	1.0 358/374[d]	—	—
Dimer AB	—			—	0.3	1.2 702/752[d]

Table 4.5: RMSD in positions of Cα atoms among the structures [a]Previously published monomeric x-ray structure [22].
[b]Model from Cryo-EM [24].
[c]NhaA triple mutant.
[d]n/m: n residues out of m residues matched using the algorithm as described in Section 4.2 in the main text.

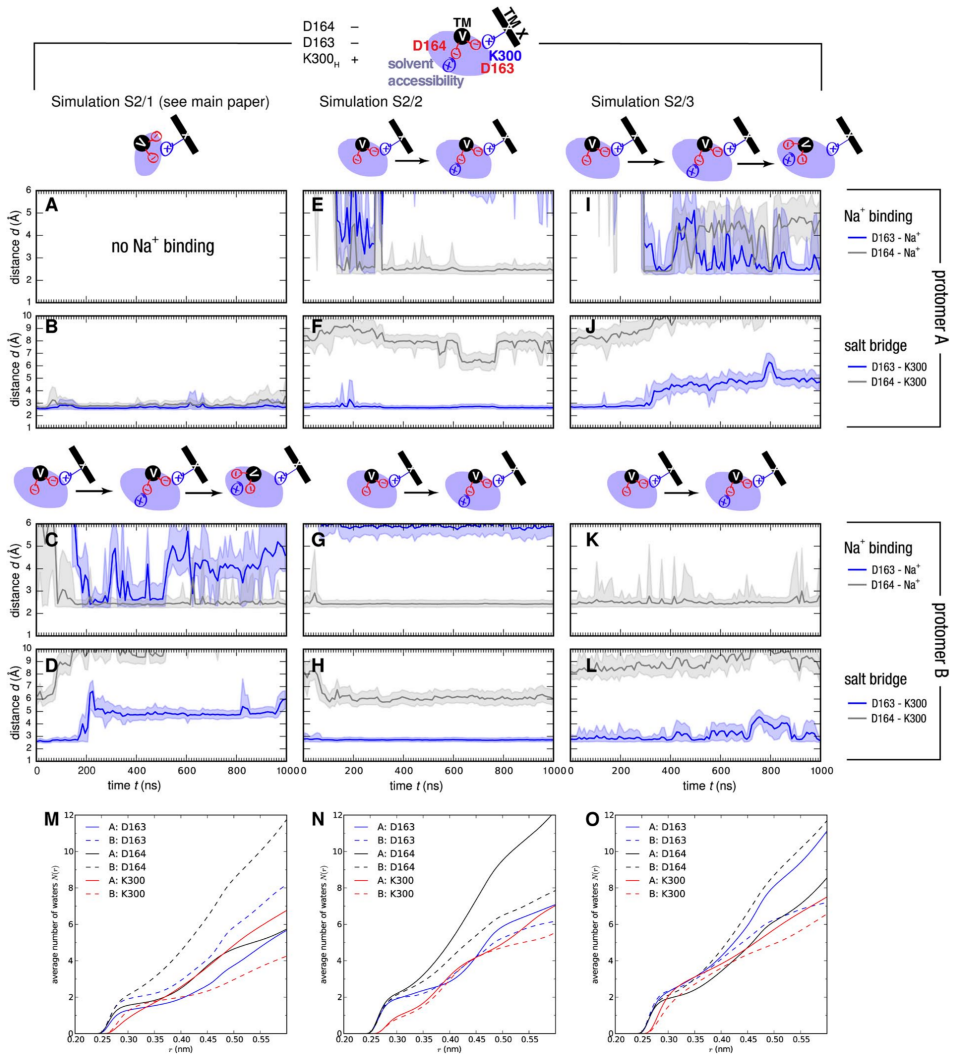


Figure 4.14: Analysis of simulations S2 (charged Asp163, charged Asp164, charged Lys300). Results are shown for each protomer (labeled A and B). Simulation S2/1 (protomer B) is also described in the main text and repeated here for convenience. (Top; A-L) Distances between sodium ions and aspartate oxygen atoms and salt-bridge distances between Lys300, Asp163, and Asp164. (Bottom; M-O) Solvent accessibility of Asp163, Asp164, and Lys300, measured as the average number of water molecules as a function of distance from the aspartate carboxyl oxygens or the lysine amine nitrogen. Cartoon images illustrate the key characteristics observed in the simulations.

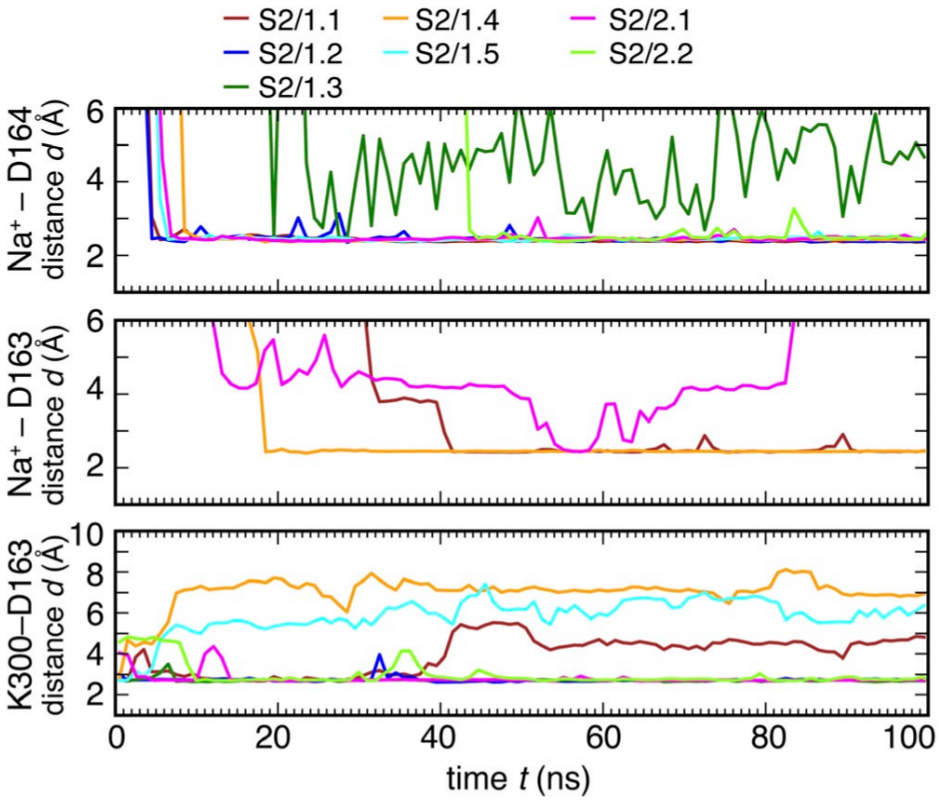


Figure 4.15: Monomer repeat simulations S2 (charged Asp163, charged Asp164, charged Lys300). Seven repeat simulations of the monomer system were analyzed in the same way as the dimer simulation (see Figure 4.14 for details). Time series data were averaged in windows of length of 1 ns; fluctuations are not shown but are of similar magnitude as those in Figure 4.14. (Top) Closest distance between a sodium ion and the δ-carboxyl oxygens of Asp164. A sodium ion binds to Asp164 in six of the repeats. In repeat 1.3, an ion enters the vestibule and stays in the vicinity of Asp164 but has not bound tightly by the end of the 100-ns simulation. (Middle) Closest distance between a sodium ion and Asp163. The broken Lys300-Asp163 salt bridge does not reform once a sodium ion is bound to Asp164 (repeat 1.5) or Asp164 and Asp163 (repeat simulations 1.1 and 1.4). The ion is not seen to bind to both Asp164 and Asp163 while the salt bridge is intact. (Bottom) Salt-bridge distance Lys300-Asp163. The salt bridge eventually breaks in three out of seven repeats (1.1, 1.4, and 1.5).

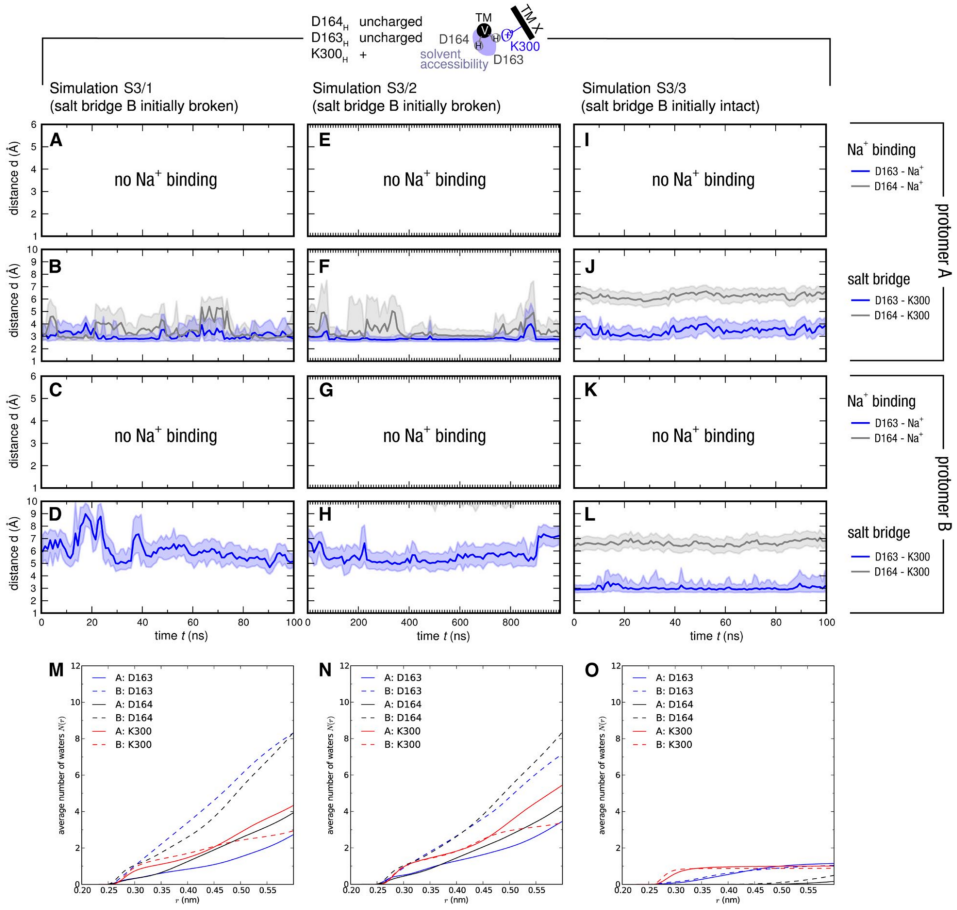


Figure 4.16: Analysis of simulations S3 (neutral Asp163, neutral Asp164, charged Lys300): Results are shown for each protomer (labeled A and B). Simulations S3/1 and S3/2 were initialized from configurations in which the salt bridge in protomer B was already broken (see Table 4.1 in the main text), whereas S3/3 had the salt bridges in both protomer A and B intact. (Top; A-L) Distances between sodium ions and aspartate oxygen atoms and salt-bridge distances between Lys300 (K300), Asp163 (D163), and Asp164 (D164). (Bottom; M-O) Solvent accessibility of Asp163, Asp164, and Lys300, measured as the average number of water molecules as a function of distance from the aspartate carboxyl oxygens or the lysine amine nitrogen. Cartoon images illustrate the key characteristics observed in the simulations.

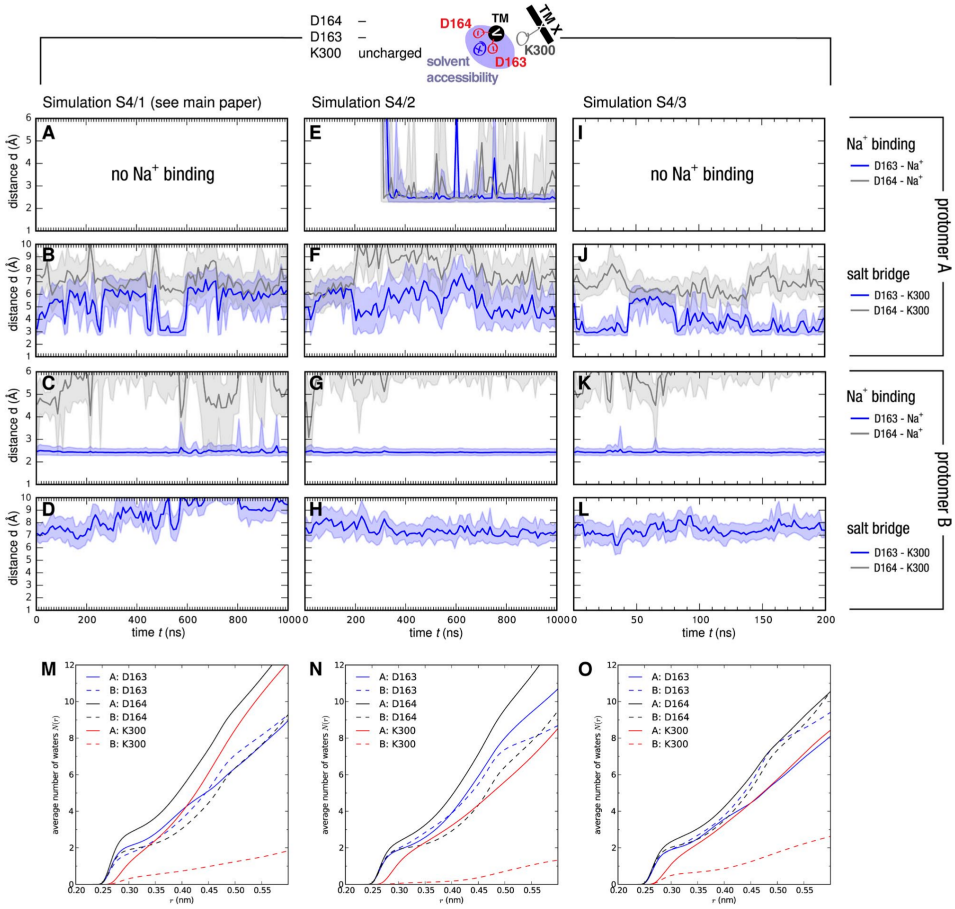


Figure 4.17: Analysis of simulations S4 (charged Asp163, charged Asp164, neutral Lys300). Results are shown for each protomer (labeled A and B). (Top; A-L) Distances between sodium ions and aspartate oxygen atoms and salt-bridge distances between Lys300 (K300), Asp163 (D163), and Asp164 (D164). (Bottom; M-O) Solvent accessibility of Asp163, Asp164, and Lys300, measured as the average number of water molecules as a function of distance from the aspartate carboxyl oxygens or the lysine amine nitrogen. Cartoon images illustrate the key characteristics observed in the simulations.

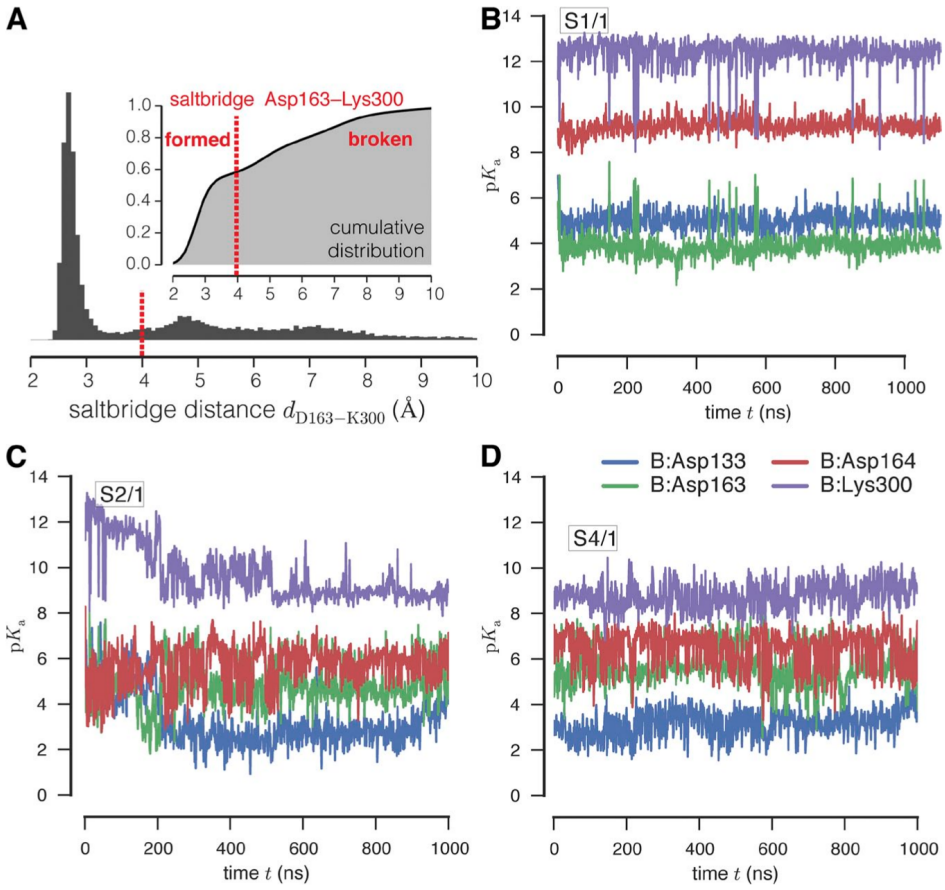


Figure 4.18: Raw data for pK_a value analysis. (A) Histogram of salt-bridge Asp163-Lys300 distances over all simulations. (Inset) Cumulative distribution. The salt bridge was considered formed for $d < 4$ Å and broken at ≥ 4 Å. All further analysis is fairly insensitive to the exact value because changing the cutoff by ± 0.5 Å only changes the number of included snapshots by less than $\sim$5%, as seen from the cumulative distribution. (See also Table 4.3 in the main text for supporting results from a sensitivity analysis.) (B) pK_a calculated by PROPKA 3.1 for frames sampled every 1 ns from the S1/1 trajectory, protomer B. No ion binding occurs, and the salt bridge remains intact. (C) S2/1 simulation, protomer B. Ion binding to Asp164 occurs between 20 and 100 ns, and the salt bridge breaks around 200 ns, resulting in a decrease in the pK_a of Lys300. (D) S4/1 simulation, protomer B. The salt bridge remains broken with a Na^+ shared between Asp164 and Asp163.

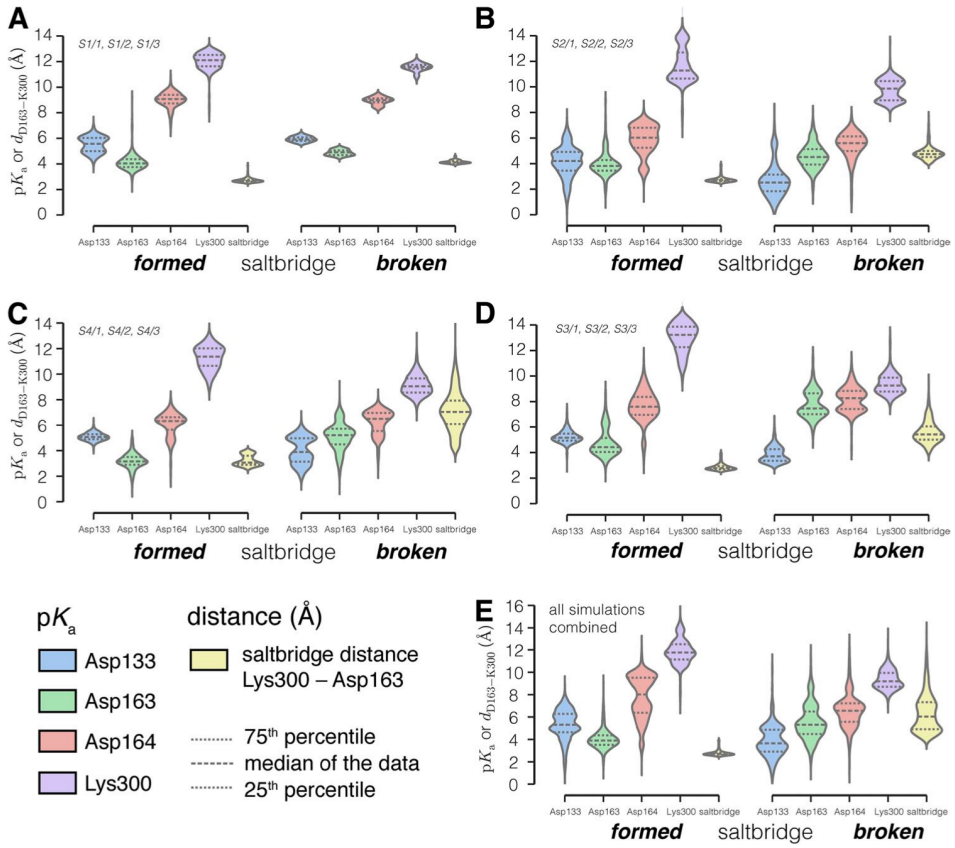


Figure 4.19: Analysis of pK_a values by simulation and salt-bridge state. Distributions of pK_a values are shown as violin plots for the four conserved residues of interest, grouped by (a) simulated protonation states and (b) state of the salt bridge (formed or broken). Distributions are scaled to have the same maximum width. The 25th, 50th (median), and 75th percentiles of the data are indicated by dashed lines. The distribution of the salt-bridge distance (yellow) is also included on the same scale (in Å) and can be seen to be < 4 Å (formed) and ≥ 4 Å (broken). (A) Simulations S1 (Asp163 deprotonated and Asp164 and Lys300 protonated). (B) Simulations S2 (Asp163 and Asp164 deprotonated and Lys300 protonated). (C) Simulations S4 (Asp163, Asp164, and Lys300 deprotonated). (D) Simulations S3 (Asp163, Asp164, and Lys300 protonated). (E) Data from all simulations shown in Table 4.1 in the main text were combined and analyzed in aggregate.

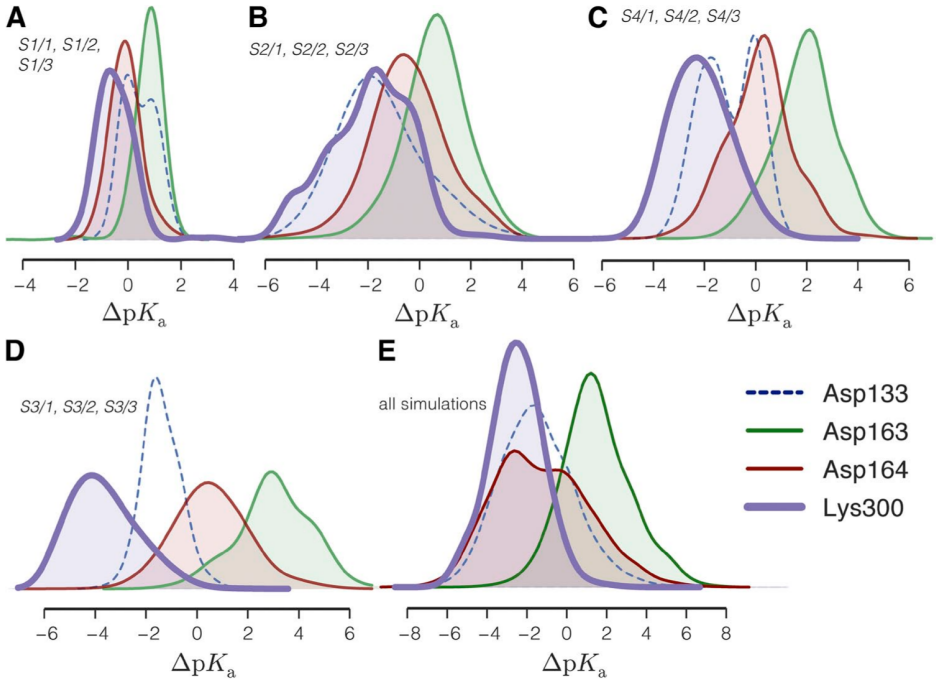


Figure 4.20: Shift in pK_a caused by breaking of the Asp163-Lys300 salt bridge. The distribution of the shift ΔpK_a = $pK_a^{\text{salt bridge broken}}$ - $pK_a^{\text{salt bridge intact}}$ quantifies how the chemical environment around the titratable residues changes as a consequence of breaking the salt bridge (and concomitant Na^+ binding). (A) Simulations S1 (Asp163 deprotonated and Asp164 and Lys300 protonated). (B) Simulations S2 (Asp163 and Asp164 deprotonated and Lys300 protonated). (C) Simulations S4 (Asp163, Asp164, and Lys300 deprotonated). (D) Simulations S3 (Asp163, Asp164, and Lys300 protonated). (E) All simulations listed in Table 4.1 in the main text analyzed in aggregate.

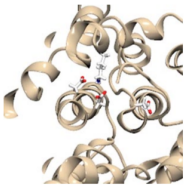


Figure 4.21: (Video 1) NhaA, protomer A, 0-250 ns from simulation S2/1. Stable salt-bridge network between Lys300 and Asp163/Asp164. The conformation of the network is essentially unchanged over the course of the whole simulation, and hence only the first 250 ns are shown. Video: `http://movie.rupress.org/video/10.1085/jgp.201411219/video-1`

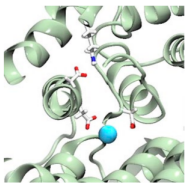


Figure 4.22: (Video 2) NhaA, protomer B, 0-250 ns from simulation S2/1. A sodium ion spontaneously binds to Asp133 and Asp164. Asp133 moves away, leaving the ion with Asp164. The salt bridge Lys300-Asp163 breaks and Asp163 binds the Na^+ ion together with Asp164, preventing the reformation of the salt bridge. These interactions remain until the end of the 1-μs simulation, even though only the first 250 ns are shown. Video: `http://movie.rupress.org/video/10.1085/jgp.201411219/video-2`

Chapter 5

CRYSTAL STRUCTURES REVEAL THE MOLECULAR BASIS OF ION TRANSLOCATION IN SODIUM/PROTON ANTIPORTERS

This chapter is a reprint of the journal article, Coincon, M., Uzdavinys, P., Nji, E., **Dotson, D.L.**, Winkelmann, I., Abdul-Hussein, S., Cameron, A.D., Beckstein, O., and Drew, D. (2016). Crystal structures reveal the molecular basis of ion translocation in sodium/proton antiporters. Nat Struct Mol Biol 23, 248-255. This work revealed an inward-facing crystal structure of *Thermus thermophilus* NapA at 3.7 Å resolution, obtained through disulfide-locking of cysteine mutants. This structure, along with a new outward-facing NapA structure obtained from lipidic-cubic phase (LCP) crystals, clearly show that the core domain of NapA undergoes a large conformational change during the transport cycle, displacing the bound Na^+ a full 10 Å between the inward- and outward-facing states. Functional experiments of cysteine mutants reconstituted in proteoliposomes demonstrate unambiguously that disulfide-locking abolishes transport, but that transport for cysteine mutants can be restored with addition of a reducing agent.

Molecular dynamics simulations demonstrate the stability of the inward-facing structure, both as a disulfide-locked mutant and wild-type, and the outward-facing LCP structure. We also showed that it is the core domain, as opposed to the dimerization domain, that translocates most relative to the membrane during the transition between conformation. It was also demonstrated that both structures are accessible to Na^+ for spontaneous binding, but that binding is transient for the outward-facing structure.

This work clearly establishes that NapA undergoes a large, not a small, conformational change to transport ions. Although not yet proven, it is believed that a similar mechanism applies to *Escherichia coli* NhaA. My contribution to this work was the performance and analysis of all molecular dynamics simulations. This work first appeared in Nature Structural and Molecular Biology, Copyright © 2016 the Authors.

ABSTRACT

To fully understand the transport mechanism of Na^+/H^+ exchangers, it is necessary to clearly establish the global rearrangements required to facilitate ion translocation. Currently, two different transport models have been proposed. Some reports have suggested that structural isomerization is achieved through large elevator-like rearrangements similar to those seen in the structurally unrelated sodium-coupled glutamate-transporter homolog GltPh. Others have proposed that only small domain movements are required for ion exchange, and a conventional rocking-bundle model has been proposed instead. Here, to resolve these differences, we report atomic-resolution structures of the same Na^+/H^+ antiporter (NapA from *Thermus thermophilus*) in both outward- and inward-facing conformations. These data combined with cross-linking, molecular dynamics simulations and isothermal calorimetry suggest that Na^+/H^+ antiporters provide alternating access to the ion-binding site by using elevator-like structural transitions.

5.1 Introduction

Na^+/H^+ exchangers (NHE1-NHE9 in mammals) mediate the outward movement of protons (H^+) in exchange for sodium (Na^+) or lithium (Li^+) ions[4,5]. These

transporters are important in regulating intracellular pH, sodium levels and cell volume, and have roles in control of the cell cycle, cell proliferation, cell migration and vesicle trafficking[4,5]. Na^+/H^+ exchangers belong to the monovalent cation/proton antiporter superfamily (CPA1 and CPA2) and are important drug targets because their dysfunction is linked to a variety of diseases, including cancer and cardiovascular pathophysiological disorders[5,146]. The Na^+/H^+ antiporter from *Escherichia coli* (NhaA), a bacterial homolog of the NHEs, was the first crystal structure of a CPA superfamily member to be determined[22]. The structure revealed that Na^+/H^+ exchangers are made up of a central dimer domain and a core (ion-translocation) domain[22,32]. The core domain contains two antiparallel discontinuous helices, transmembrane (TM) domains 4 and 11, which cross over near the center of the protein. The ion-binding aspartate[22,28,31,32,147], a feature ubiquitous among prokaryotic and eukaryotic Na^+/H^+ antiporters[4], is located on TM5, near the TM4-TM11 crossover. In the crystal structure of NhaA[22], a negatively charged cavity is formed at the interface of the two domains and provides access to the ion-binding aspartate from the cytoplasm. It has been proposed that the two half helices in the core domain may undergo small rearrangements that close access to the inside cavity and open the binding site to the outside[22,23]. However, recent comparisons to an outward-facing crystal structure of a homologous Na^+/H^+ exchanger, NapA from *T. thermophilus*[31], have suggested that substrate translocation may instead occur by a large outward movement of the core domain itself. Because the size of the oligomerization surface in NapA is extensive ($\sim$1,800 $Å^2$), the core domain has been modeled to move against a dimer domain that is fixed, thus yielding elevator-like movements similar to those seen in the trimeric sodium-coupled glutamate-transporter homolog GltPh[31,77].

Although an elevator-like model has been proposed, a detailed mechanism could not previously be resolved because NhaA and NapA have relatively low sequence iden-

tity and show some structural differences[22,31,32], particularly in the dimer domain. Indeed, these movements have been questioned in a more recent crystallographic study of the archaeal Na^+/H^+ antiporter NhaP1 from *Methanococcus jannaschii* (MjNhaP1), which has a topology similar to that of NapA[30]. MjNhaP1 has been solved in an inward-facing conformation by X-ray crystallography, and the structure of the same protein in an outward-facing state has been revealed through ~6 Å (in plane)- and 14 Å (z direction)-resolution EM maps obtained from two-dimensional (2D) crystals[30]. On the basis of a comparison of these two structures, the core domain has been presumed to undergo much smaller movements, thus ruling out an elevator model and instead supporting the conceptually different rocking-bundle mechanism[30].

Because NhaA, NapA and NhaP1 proteins share low sequence identity to one another[31], the extent of rearrangements may differ between these proteins. Even so, the type of alternating-access model used by these and other Na^+/H^+ exchangers should be conserved. It seems that the current ambiguities are likely to be resolvable only by the determination of equivalent or highly homologous Na^+/H^+ exchanger structures at near-atomic resolution in both outward- and inward-facing states. To confirm the structural basis of ion translocation in Na^+/H^+ exchangers, we therefore set out to obtain a crystal structure of NapA in an inward-facing conformation.

5.2 Results

5.2.1 Elevator-like movements of NapA can occur in a lipid bilayer

We hypothesized that if NapA undergoes elevator-like rearrangements, it should be possible to trap an inward-facing conformation by engineering a disulfide bond between a residue at position 31 in the dimer domain and a residue at position 130 in the core domain (Figure 5.1a); in the outward-facing NapA structure, these residues

are located ∼10 Å apart. To test this model, we constructed V31C and I130C mutants of wild-type NapA, which is cysteineless, expressed them in *E. coli* and purified them at active pH (Methods). Incubation of each of the single V31C and I130C mutants with methoxypolyethylene glycol 5000 maleimide (maleimide-PEG-5k) showed a clear species with higher electrophoretic mobility, thus indicating that both cysteine residues were solvent accessible (Figure 5.1b and Supplementary Data Set 1). In contrast, incubation of maleimide-PEG-5k with the purified double-cysteine V31C I130C mutant showed little PEGylated product, thus indicating that the same cysteine residues were no longer solvent accessible, presumably because they had formed a disulfide bond (Figure 5.1b and Supplementary Data Set 1). Consistently with disulfide-bond formation, reactivity to a sulfhydryl-binding dye was also significantly decreased in the double-cysteine mutant compared to each of the single-cysteine mutants (Methods, Supplementary Figure 5.7a and Supplementary Data Set 1).

To determine whether V31C and I130C residues are able to form a disulfide bond in a membrane environment, we co-reconstituted each of the purified cysteine mutants with *E. coli* F_oF_1 ATP synthase into liposomes (Online Methods and Supplementary Figure 5.7b). After establishing a pH gradient by the addition of ATP, we monitored proton efflux in response to Na^+ or Li^+ addition in the presence of valinomycin and potassium. In the presence of DTT, single- and double-cysteine mutants showed robust transporter activity with apparent Na^+ and Li^+ affinities (K_m) similar to that of wild type (Figure 5.1 and Table 5.1). Likewise, after removal of DTT, wild-type NapA and each of the single-cysteine mutants showed similar antiport activities (Figure 5.1c,d,g). In contrast, removal of DTT from proteoliposomes containing the double-cysteine mutant abolished ∼90% of its transport activity (Figure 5.1e,g). Re-addition of DTT restored antiport activity to starting levels (Fig. 5.1e,g). Thus, a disulfide bond between V31C in the dimer domain and I130C in the core domain can

spontaneously form in a membrane environment, in agreement with an elevator-like transport model[31].

Biochemical characterization of NapA mutants

	K_m (mM) at pH 8.0	
NapA	Na^+	Li^+
WT	2.08 ± 0.12	0.56 ± 0.02
V31C	1.48 ± 0.14	0.27 ± 0.04
I130C	6.14 ± 0.83	0.55 ± 0.06
V31C I130C	4.74 ± 0.38	0.64 ± 0.08
V31C I130C[a]	4.89 ± 0.65	
I55C	2.20 ± 0.10	0.44 ± 0.03
V71C L141C	6.46 ± 0.47	0.49 ± 0.04
R133E	9.33 ± 1.56	2.64 ± 0.50
E35R	>50	3.22 ± 0.87
R133E E35R	5.87 ± 0.35	1.22 ± 0.20
K344A	14.60 ± 2.00	2.31 ± 0.32

Table 5.1: Biochemical characterization of NapA mutants. The apparent Na^+ and Li^+ K_m affinities were calculated with a range of ten ion concentrations that were fitted by nonlinear regression with data from two technical repeats (values reported are the mean ± s.e.m. of the fit). For the double-cysteine mutants, the transport kinetics were measured in the presence of 5 mM DTT.
[a]The apparent affinity K for V31C I130C was also measured for Na^+ at pH 8.5.

To provide further support for an elevator model, we further engineered an SDS-resistant NapA dimer by covalently linking the two protomers together by substituting I55 with cysteine (Figure 5.1a, Supplementary Figure 5.7c and Supplementary Data Set 1). Consistently with the dimer domain being fixed during transport, the covalently linked dimer showed wild type-like antiport activity under either nonreducing or reducing conditions (Figure 5.1g, Table 5.1 and Supplementary Figure 5.7d).

5.2.2 The inward-facing crystal structure of NapA

We obtained crystals of the cross-linked V31C I130C mutant in the absence of any additional oxidizing reagents and in similar crystallization conditions to those of the wild-type NapA protein at active pH (Methods). We solved the structure of the cross-linked V31C I130C mutant by molecular replacement at 3.7-Å resolution (Figure 5.2a, Table 5.2 and Supplementary Figure 5.8a). The electron density maps revealed that

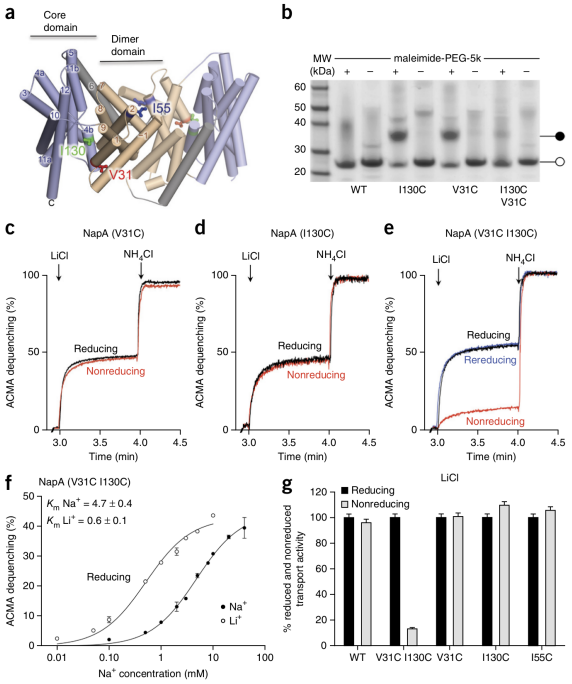


Figure 5.1: Disulfide trapping of NapA in an inward-facing conformation. (a) Cartoon representation of the outward-facing NapA structure (in which one of the protomers is faded). Blue, core (ion-translocation) domain; wheat, dimer domain; gray, linker helix (TM6). Residues used for cysteine mutagenesis are depicted as sticks and are color-coded (red, V31; blue, I55; green, I130). (b) The degree of cysteine reactivity of wild-type NapA (WT) and single- and double-cysteine mutants (open circle) to maleimide-PEG-5k (black circle). MW, molecular weight. (c-e) Typical 9-aino-6-chloro-2-methoxyacridine (ACMA) fluorescence traces of proteoliposomes containing single- and double-cysteine NapA mutants in the presence of DTT (reducing, black trace), after the removal of DTT (nonreducing, red trace) and after the re-addition of DTT (rereducing, blue trace). Each experiment was independently repeated three times; data from one representative experiment are shown. The ΔpH gradient established between 0 and 3 min is not shown (Supplementary Figure 5.7b). (f) Apparent K_m for Na^+ (filled circles) and Li^+ (open circles) of the double-cysteine mutant under reducing conditions. The apparent Na^+ and Li^+ K_m affinities were calculated with a range of ion concentrations that were fitted by nonlinear regression with data from two technical replicates (bars show the range of two data points, and the values reported are the mean ± s.e.m. of the fit; additional data in Table 5.1). (g) Quantification of antiport activity after the addition of LiCl in the presence (reducing) or absence (nonreducing) of DTT (error bars, s.e.m.; $n = 3$ experiments).

a disulfide bond forms between residues C31 and C130 (Figure 5.2b). In molecular dynamics (MD) simulations (Supplementary Table 5.3) the backbone r.m.s. deviation (r.m.s.d.) increased to less than 3 Å over 1 μs, thus indicating that the disulfide-trapped structure is stable in a model membrane bilayer (Supplementary Figure 5.9a). We also observed similar r.m.s.d. values and r.m.s. fluctuations in MD after removing the disulfide by replacing C31 and C130 with wild-type residues (Supplementary Figure 5.9a,b). Furthermore, the disulfide-trapped NapA structure superimposed well onto the inward-facing crystal structure of MjNhaP1 (Supplementary Figure 5.9c);

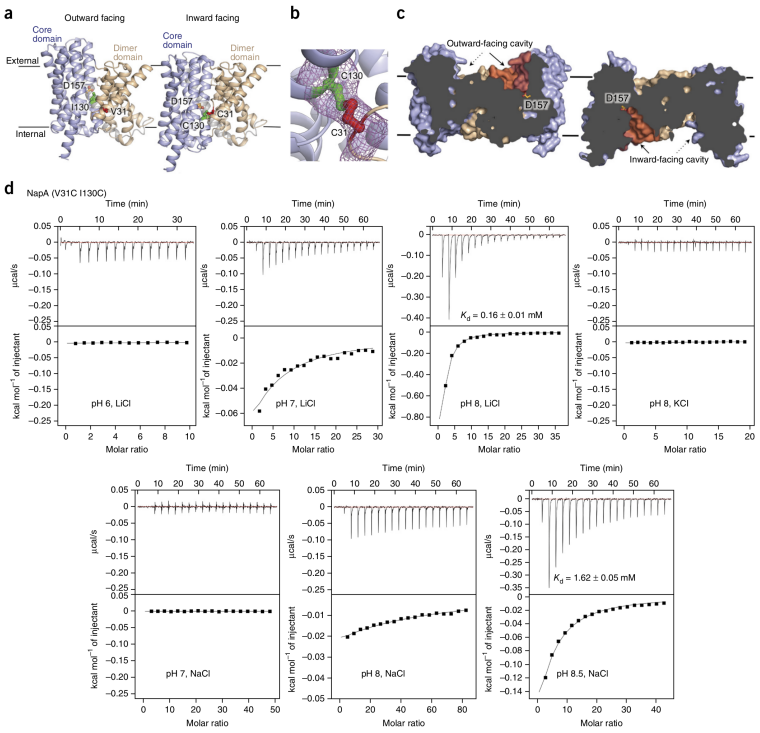


Figure 5.2: The disulfide-trapped structure of NapA in an inward-facing conformation. (a) NapA monomer (blue, core; wheat, dimer) shown in the outward-facing conformation (left, PDB 4BWZ) and inward-facing conformation (right, V31C I130C). Positions of the V31 and I130 residues substituted with cysteine (left) are shown as red and green sticks, respectively. The ion-binding aspartate D157 is shown as orange sticks. (b) Simulated-annealing $F_o - F_c$ omit map at 3.5σ showing the disulfide bond formed between residues C31 and C130 in the inward-facing structure. (c) Surface representation of the NapA structures clipped through the ion-binding sites in either the outward-facing (left) or inward-facing (right) conformation. The strictly conserved ion-binding aspartate D157 is shown as orange sticks and labeled. (d) Representative ITC thermograms from one experiment obtained by successive addition of Li^+, K^+ or Na^+ to the V31C I130C mutant under nonreducing conditions (Methods). Integrated heats with baseline-subtracted (red line) inverted power data are shown in top panels. Fit of a single binding site with 1:1 stoichiometry is shown in bottom panels (black line). K_d values are reported as the mean ± s.e.m. of the fit.

the disulfide bond between residues C31 and C130 appeared to have caused only a local conformational difference at the end of TM1. Strikingly, the strictly conserved ion-binding residues in NapA (D157) and MjNhaP1 (D161) were positioned within only ∼1 Å of one another (Supplementary Figure 5.9c). As expected, the outward-facing cavity in NapA had closed, and a vestibule had opened to the inside (Figure 5.2c). Highly conserved residues were clearly surface exposed by the opening of the inward-facing cavity (Supplementary Figure 5.9d).

The above discussion suggests that the introduction of the disulfide bond does not excessively perturb the NapA structure. To further corroborate this possibility, we

Data collection and refinement statistics

	LCP	V31C I130C
Data collection		
Space group	$C222_1$	$C222_1$
Cell dimensions		
a, *b*, *c* (Å)	56.58, 94.09, 147.28	77.71,84.64,201.78
α, β, γ (°)	90, 90, 90	90, 90, 90
Resolution (Å)	47.0–2.3 (2.38–2.3)[a]	33.6–3.7 (3.83–3.7)
R_{merge}	0.182 (1.036)	0.090 (2.788)
$CC_{1/2}$	0.972 (0.735)	0.999 (0.575)
CC^*	0.993 (0.921)	0.999 (0.854)
$I / \sigma I$	12.2 (1.6)	13.3 (1.1)
Completeness (%)	99 (100)	100 (100)
Redundancy	4.5 (4.5)	12.7 (13.3)
Refinement		
Resolution (Å)		
No. reflections	17,750	7,386
R_{work} / R_{free}	0.206 / 0.218	0.292 / 0.325
No. atoms		
Protein	2,828	2,823
Ligand	147	–
Water	76	–
B factors		
Protein	45.5	197
Ligand	72.8	–
Water	50.4	–
r.m.s. deviations		
Bond lengths (Å)	0.005	0.005
Bond angles (°)	0.96	0.81
Number of TLS group	–	1

Table 5.2: Data collection and refinement statistics. [a]Values in parentheses are for highest-resolution shell.

carried out isothermal titration calorimetry (ITC). The disulfide-locked protein was able to bind Li^+ and Na^+ ions with affinities (K_d) in a similar range as their apparent K_m values at pH 8 and 8.5, respectively (Figure 5.2d and Table 5.1). Furthermore, NapA, as with all Na^+/H^+ antiporters, shows strict pH-dependent transport activity[68,69]. Consistently with this, we observed that its ion binding was also highly sensitive to pH (Figure 5.2d). MD simulations of the inward-facing conformation further demonstrated how bulk Na^+ ions are able to access the strictly conserved ion-binding aspartate D157 from the cytoplasm (Supplementary Figure 5.10a). In the simula-

tions, D157 typically interacts with Na^+ in a bidendate configuration, whereby both side chain oxygen atoms of the aspartate are within less than 3 Å of the ion (Supplementary Figure 5.10a-e and Supplementary Table 5.4). The manner of coordination of Na^+ was similar to how thallium binds to the equivalent aspartate in the inward-facing NhaP structure from *Pyrococcus abyssi*[147]. In MD simulations, we found that no dimer domain residues interact with the Na^+ ion. This result is consistent with the observation that mutation of the poorly conserved dimer-domain glutamate, which has unexpectedly been observed to coordinate Tl^+ in *P. abyssi* NhaP, is not essential for transport[147]. We conclude that the disulfide-trapped structure of NapA is able to bind substrate and is in an inward-facing conformation similar to that seen in homologous Na^+/H^+ antiporter structures[22,30,32,147].

5.2.3 Outward-facing NapA crystal structure in lipidic mesophase

Because the inward-facing crystal structures of NapA and MjNhaP1 are so similar (Supplementary Figure 5.9c), the different mechanism proposed is probably a result of the differences observed between the outward-facing crystal structures. Indeed, the 2D structure of MjNhaP1 based on the EM maps has been reported to be less open to the outside than the outward-facing crystal structure of NapA[30,31]. It is possible that crystallization of NapA in detergent may have contributed to the conformational differences observed between these two proteins[31].

To confirm that the outward-facing crystal structure of NapA represents a state that can also be populated in a lipid-bilayer, we engineered a different set of cysteine mutants between V71 in the dimer domain and L141 in the core domain; these residues are located ∼15 Å apart in the inward-facing structure but come within 5 Å of one another in the outward-facing structure (Supplementary Figure 5.7e). Indeed, in the absence of DTT, a disulfide bond formed with ∼90% efficiency between C71 and C141

residues in both detergent solution and proteoliposomes (Supplementary Figure 5.7f,g and Supplementary Data Set 1). In the presence of DTT, however, the double V71C L141C mutant showed similar antiport activity to that of wild-type NapA (Table 5.1 and Supplementary Figure 5.7g). Finally, we proceeded to combine 71C and 130C mutants and, in contrast to the previous pairings, we modeled them to remain at least ∼10 Å apart in the two conformations to prevent them from spontaneously forming a disulfide bond (Supplementary Figure 5.7e). In proteoliposomes containing the 71C 130C mutant, we observed little change in antiport activity upon the removal of DTT (Supplementary Figure 5.7h).

To provide additional support for the physiological relevance of the outward-facing NapA crystal structure, we further sought to obtain a structure of NapA at active pH by using the lipidic-cubic phase (LCP) method (Figure 5.3 and Table 5.2). Unlike the NapA crystal structure obtained in detergent, the structure refined at 2.3-Å resolution in lipids showed bilayer-type crystal packing (Figure 5.3b). Superimposition of the new outward-facing LCP crystal structure with that obtained in detergent [31] showed that the core domain nonetheless adopts the same overall conformation (Figure 5.3a). The main difference between the two outward-facing structures is that the half helices 11b and 4a are tilted ∼6 Å further away from the cavity in the LCP structure, thus making it more open (Figure 5.3a). This difference might be due to the involvement of these helices in crystal packing (Figure 5.3b). In MD simulations, half helices 11b and 4a fluctuated between the positions seen in both outward-facing NapA crystal structures, thus indicating that they are very mobile (Figure 5.3c). The open vestibule in the LCP structure provides an open path for Na^+ ions to bind to D157 from the extracellular solvent (Supplementary Figure 5.10f-i). Although binding events were infrequent and short (Supplementary Figure 5.10j), they nevertheless showed ion-

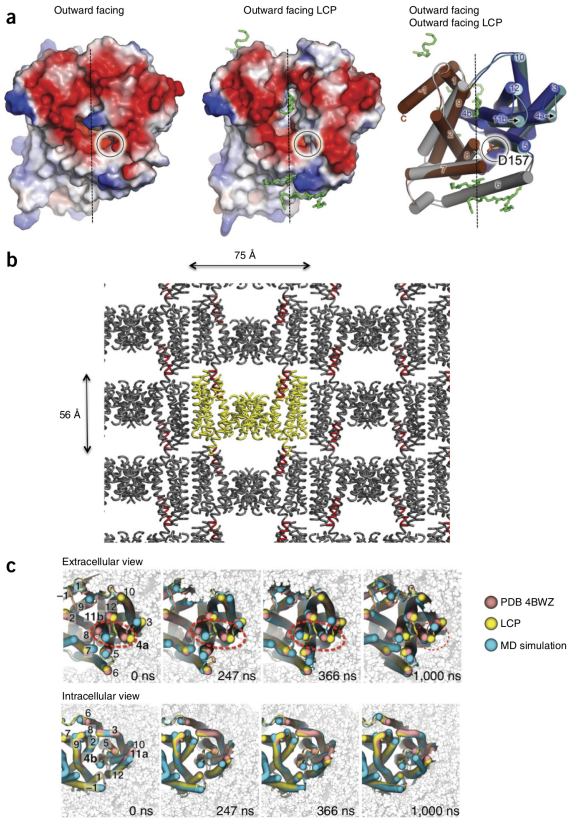


Figure 5.3: LCP structure of NapA supports the physiological positioning of the outward-facing conformation. (a) Electrostatic surface representation of the NapA protomer from the outward-facing structure obtained by vapor diffusion (left) and LCP (middle), as viewed from the extracellular side. Cartoon representation of the NapA outward-facing structure obtained by vapor diffusion (gray, dimer; blue, core) superimposed on the outward-facing LCP structure (brown, dimer; teal, core) (right). The position of the ion-binding site is circled, and additional and nonprotein density at the core-dimer domain interface is modeled as the LCP lipid MAG7.7 (green sticks). (b) Ribbon representation of the outward- facing NapA structure crystallized in the lipidic mesophase. One NapA dimer structure is shown in yellow, and TM11b, which is likely to form crystal-packing contacts, is shown in red. (c) View on the extracellular face (top) and intracellular face (bottom) of NapA. Conformations from a 1-μs MD simulation (cyan) of the LCP structure (yellow) and the detergent structure (PDB 4BWZ, pink) as described in Methods. A diverse range of snapshots displaying the mobility of half helices TM4a and 11b are shown at varying time points (0 ns to 1,000 ns) and are highlighted by a red dotted ellipse.

coordination geometry similar to that observed in the inward-facing conformation (Supplementary Figure 5.10a-e and Supplementary Table 5.4).

5.2.4 An elevator alternating-access mechanism for Na^+/H^+ antiport

Because the dimer domain is held in place by its interaction with the other protomer, it seems reasonable to assume that the core domain moves against the dimer domain rather than that the dimer moves against the core. To test this assumption, we performed MD simulations of the outward- and inward-facing NapA crystal structures in a model membrane bilayer (Supplementary Table 5.3). Whereas the vertical position (z coordinate) of the dimer domain in the membrane differed by only ∼2 Å between outward- and inward-facing conformation, the core domain moved ∼6 Å relative to the membrane and ∼7 Å relative to the position of the dimer domain

(Supplementary Figure 5.11). Thus, the core domain is able to move freely against a dimer domain that can anchor the position of the transporter in the membrane.

Analysis of a morph between the outward- and inward-facing NapA crystal structures showed that whereas the dimer domain 'breathes out' only sideways, the core domain undergoes a large rotation and a translation (Methods and Supplementary Videos 1 and 2). During this rearrangement, the Cα atom of the ion-binding aspartate (D157) is relocated $\sim$8 Å in the vertical direction (Supplementary Figure 5.8b,c). The side chain of D157 further rotates some 55°, which is necessary for the residue to access the inward-facing vestibule (Figure 5.2c and Supplementary Figure 5.8c). The distance between the D157 carboxylate groups in the two opposite-facing states is $\sim$11 Å (Figure 5.4b and Supplementary Figure 5.2b). This relocation of the ion-binding site spans a distance roughly equivalent to the thickness of the hydrophobic dimer-domain interface (Figure 5.4a). The elevator-like movement of the core against the dimer domain is facilitated by long flexible loops that are located between TMs 9 and 10 on the extracellular side and between TMs 6 and 7 on the intracellular side (Methods and Supplementary Figure 5.12). These hinge regions contain a number of glycine residues, G193, G213, G277 and G294, that may facilitate the unwinding of the helices at their connecting periphery. The hinge regions were clearly highlighted when the higher-resolution outward-facing (LCP) NapA structure was represented as temperature factors (Supplementary Figure 5.12b). These hinge regions also showed high per-residue fluctuations in MD simulations (Supplementary Figure 5.9b).

Between the outward- and inward-facing NapA conformations, there is remarkably little structural rearrangement within the core and dimer domain themselves, because separate superimposition of the domains yielded an r.m.s.d. of 0.5 Å for 124 pairs of Cα atoms in the dimer domain and $\sim$1.0 Å for 185 pairs of Cα atoms in the core domain (Figure 5.4c). Within the core domain, residues located on the mobile TM4b

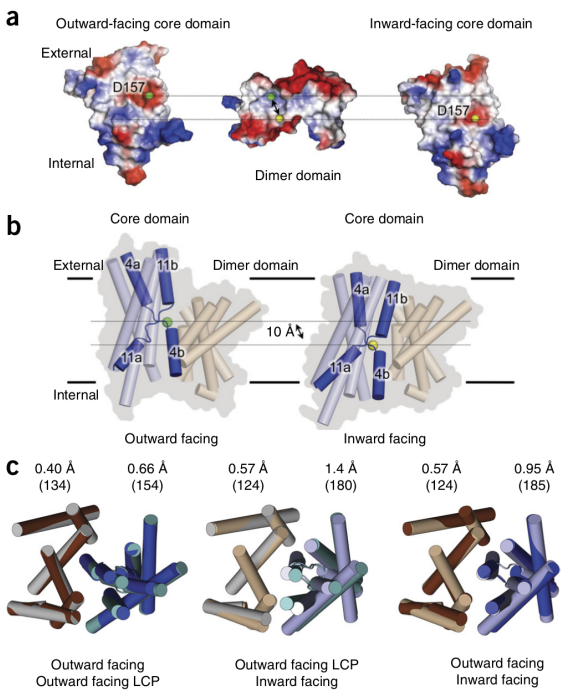


Figure 5.4: An elevator-like alternating-access mechanism for Na^+/H^+ antiport. (a) The outward-facing NapA structure was separated into core and dimer domains, and the electrostatic surface of the core (left) and dimer (middle) domain are rendered. The view shows the domain surfaces that face one another at the interface. The electrostatic surface was further rendered for the inward-facing conformation (right) and is shown on the basis of its relative position to the dimer domain. The ion-binding aspartate D157 is represented as a green and yellow sphere for the outward-facing (left) and inward-facing (right) conformations, respectively. Dashed lines represent the position of the aspartate in each of the two conformations (middle). (b) Cartoon representation highlighting the movement of the core domain (dark blue) between outward-facing (left) and inward-facing (right) conformations. The ion-binding aspartate D157 is shown as a sphere in the outward-facing (green) and inward-facing (yellow) conformations. (c) The core and dimer domains from the different NapA crystal structures are superimposed separately onto each other. Left, outward-facing detergent structure (dark brown, dimer; dark blue, core) superposed on the outward-facing LCP structure (gray, dimer; teal, core). Middle, outward-facing LCP structure (gray, dimer; teal, core) superposed on the inward-facing NapA structure (light orange, dimer; light blue, core). Right, outward-facing detergent structure (dark brown, dimer; dark blue, core) superposed on the inward-facing NapA structure (light orange, dimer; light blue, core). The respective Cα r.m.s.d. of each superimposition is shown.

and TM11b half helices make most of the contacts to the dimer domain; these contacts are extensive enough to close off the cavity to the outside and inside (Methods and Figure 5.5a,b). The small size of the occluding interface was clearly observable when TM4b and TM11b half helices were omitted in a surface representation of NapA (Figure 5.5a). Positively charged residues K344 in TM11b and R133 in TM4b may help to stabilize the alternate conformations by interacting with negatively charged residues in the dimer domain (Figure 5.5b). Indeed, a K344A mutant had a clear negative effect on transport activity, as did the mutation of the intracellular R133 E35 salt bridge between TM4b and TM1, which was rescued by a charge-swapped mutant (Figure 5.5c and Table 5.1). As extensively shown in previous studies, ion binding causes the rearrangement of half helices in the core domain[30,64,148]. It is likely that in the presence of Na^+ (or Li^+), the mobile TM4b and TM11b helices move to further

accommodate the substrate so that they no longer interact with the dimer domain, thus enabling the core to move in an elevator-like fashion as described. Interestingly, in the outward-facing LCP structure, a number of acyl chains, modeled as LCP lipids, are located between the TM4b and TM11b half helices and the dimer domain (Figure 5.3a and Supplementary Figure 5.13a). It is plausible that in vivo lipids may facilitate the elevator-like structural transitions of the core domain. Such an idea has also been proposed to facilitate transport-domain movements in the glutamate-transporter homolog GltPh[149].

5.3 Discussion

Na^+/H^+ antiporters have been extensively studied since West and Mitchell showed, in the 1970s, that H^+ and Na^+ transport is strictly coupled[8]. With exchange activity on the milli- to microsecond timescales, Na^+/H^+ antiporters represent one of the fastest-acting known classes of secondary active transporters[12,30,147]. At active pH, biophysical and biochemical data support an antiport mechanism whereby Na^+ (or Li^+) ions compete with protons for a single binding site[27,28,150,151], which is made up of a strictly conserved aspartic acid and several other polar residues[4,5,22,28,31,32,147,151].

The structural basis for the ion-exchange mechanism has been unclear. From a comparison of the inward-facing structure of NhaA and the outward-facing structure of NapA, we have suggested that both function through an elevator-like transport mechanism[31]. Elevator mechanisms were first described for the sodium-coupled glutamate-transporter homolog GltPh[77] and have since been proposed for various other transporters[152–156] including bacterial sodium-coupled bile acid-transporter homologs, which have the same fold as the Na^+/H^+ antiporters[75,157]. The characteristic of this mechanism is that one domain undergoes a large essentially rigid-body movement against the other to carry the substrate across the membrane. This differentiates

the mechanism from the rocking-bundle mechanism, wherein the two domains move around the bound substrate (Figure 5.6). Thus, an elevator-like mechanism requires a large conformational change and a vertical displacement of the substrate-binding site.

A comparison of the outward- and inward-facing structures of MjNhaP1 has shown only a small change[30]. Indeed, when the dimer domains of the two structures are superimposed, the Cα atom of Asp161 is displaced vertically by only 1.5 Å (Supplementary Figure 5.13b), which is much smaller than the 10 Å predicted from our previous model[31]. In this study, we trapped NapA in an inward-facing conformation through the introduction of a disulfide bond and solved a crystal structure similar to that of inward-facing MjNhaP1[30]. We further showed that the disulfide-locked inward-facing NapA mutant is able to bind Na^+ and Li^+ ions in a pH-specific manner and measured K_d values similar to the respective apparent K_m values. A comparison of this structure with the outward-facing structure of NapA (determined from crystals grown in either detergent or lipid) confirmed and refined the large elevator-like structural transitions. The difference between the two studies lies predominantly in the conformation of the outward-facing structure: the conformation derived for MjNhaP1 from the low-resolution EM maps is much less open than that of NapA[30]. Because side chains cannot be modeled at this resolution, it is not possible to ascertain whether the ion-binding site in MjNhaP1 is accessible to the outside. Further, because this structure was determined at inactive pH[30], the structure of the ion-binding site and possibly the conformation of the core domain may have been affected. Indeed, pH-dependent structural differences have been reported in Na^+/H^+ antiporters including MjNhaP1[64,76,147,158].

It seems likely that the overall rearrangements described here for NapA will be similar in other Na^+/H^+ exchangers, because we were able to predict the inward-

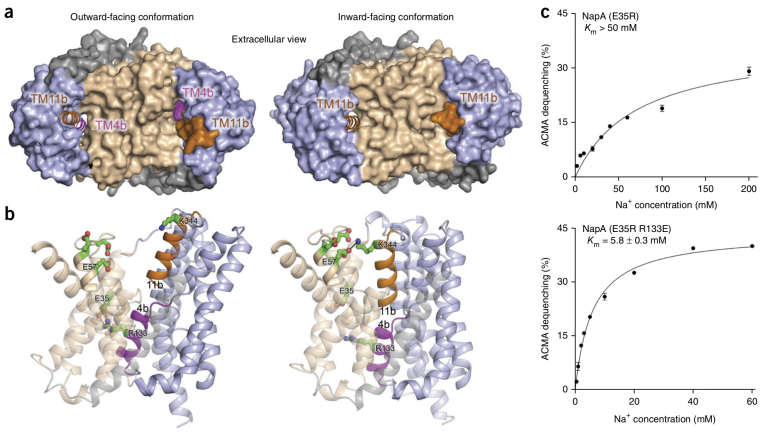


Figure 5.5: The cavity-closing interactions formed between the dimer and core domains. (a) The NapA surface (blue, core; wheat, dimer) shown in either the inward-facing (left) or outward-facing (right) conformation. Half helices TM11b (orange) and TM4b (magenta) are further illustrated in cartoon in the left protomer of each dimer. (b) Ribbon representation of NapA (blue, core; wheat, dimer) in inward-facing (left) or outward-facing (right) conformation. The alternating ionic interactions between the dimer domain and core helices TM11b (orange) and TM4b (magenta) are highlighted (green sticks). (c) Kinetic analysis of a salt-bridge swap between residues E35R in TM11b and R133E in the dimer domain. The apparent (Na^+) K_m affinities were calculated with a range of ion concentrations that were fitted by nonlinear regression with data from two technical replicates (error bars show the range of two data points, and values reported are the mean ± s.e.m. of the fit; additional data in Table 5.1).

facing conformation of NapA with reasonable accuracy by using the inward-facing NhaA crystal structure[22]. Indeed, the extent of core movement is very similar to the elevator-like rearrangements described for bacterial sodium-coupled bile acid-transporter homologs, although they operate at much slower turnover rates[157].

Because a Na^+- or Li^+-bound structure of a Na^+/H^+ antiporter remains to be determined, intermediate and possibly occluded ion-bound states have yet to be experimentally resolved. It seems likely that additional structural changes not yet seen in crystal structures or MD simulations are likely to take place. These states will be important for understanding the coupling between conformational states and transport rates. What is known is that numerous studies including cysteine accessibility analysis, tryptophan quenching and 2D crystallography support the local movement of the half helices at active pH and/or upon the addition of substrate[24,148,160]. We propose that these ion- and/or proton-induced rearrangements will be a prerequisite for the large structural changes observed here and, as seen in NapA, may further

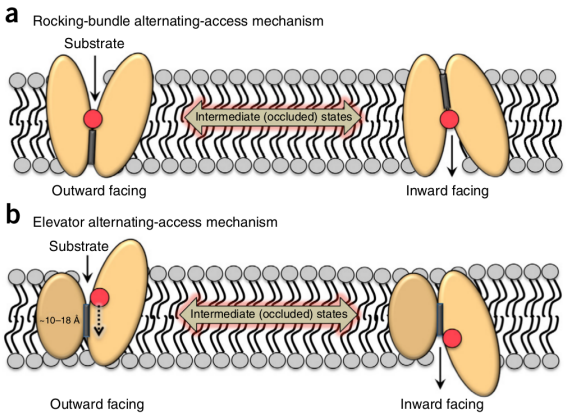


Figure 5.6: Schematic illustrating the conceptual differences between rocking-bundle and elevator alternating-access mechanisms. (a) In the rocking-bundle alternating-access mechanism, the substrate (red circle) binds between two similarly sized domains (light orange) near the center of the membrane, thereby causing domain rearrangement around the substrate. (b) In the elevator alternating-access mechanism, the substrate (red circle) binds predominantly or exclusively to only one of the domains (light orange), which is then carried across the membrane against the fixed oligomerization domain (dark orange). The vertical displacement of the substrate between the two alternative conformations (relative to the scaffold domain) is likely to be in the range of ∼10-18 Å, as seen in NapA6 and GltPh [77,159] structures. For the sake of simplicity, intermediate and occluded states formed between the two major outward- and inward-facing conformations are not shown.

require the breakage of interdomain salt bridges. It is possible that the core domain itself does not readily move against the hydrophobic 'barrier' of the dimer-domain interface surface in the absence of substrate, because of its 'nonfilled' negatively charged ion-binding surface (Figure 5.4a).

Although more studies are clearly required to elucidate the finer coupling of ion binding with the global conformational changes described here, we nonetheless uncovered new atomic-level insights into how an ion exchanger can use large, elevator-like structural transitions to translocate ions across a lipid bilayer.

Accession codes

Coordinates and structure factors for NapA have been deposited in the Protein Data Bank under accession codes PDB 5BZ2 (inward facing) and PDB 5BZ3 (LCP outward facing).

Acknowledgments

We are grateful to G. Verdon for discussions and comments. Data were collected at Diamond Light Source with excellent assistance from beamline scientists. This work was supported by the Swedish Research Council (D.D.) and the Knut and Alice Wallenberg Foundation (D.D.). The authors are grateful for the use of the Membrane Protein Laboratory supported by the Wellcome Trust UK (grant 062164/Z/00/Z) at the Diamond Light Source Limited and the Centre for Biomembrane Research supported by the Swedish Foundation for Strategic Research. Computer simulations were partially run on the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by US National Science Foundation grant OCI-1053575 (allocation TG-MCB130177 to O.B.). M.C. was supported as a Wenner-Gren postdoctoral fellow, and D.D. is supported as a European Molecular Biology Organization (EMBO) Young Investigator.

Author contributions

D.D. designed the project. Cloning, expression screening, protein purification and crystallization of inward-facing NapA were carried out by M.C. and P.U. LCP crystallization of outward-facing NapA was carried out by E.N. Data collection and structural determination were carried out by M.C. and D.D. with assistance from E.N. and A.D.C. Experiments for functional analysis were carried out by P.U., I.W., S.A.-H. and M.C. MD simulations were carried out by D.L.D. and O.B. D.D. wrote the manuscript with contributions from all authors.

Competing Financial Interests

The authors declare no competing financial interests.

5.4 Methods

5.4.1 NapA sequence

T. thermophilus NapA sequence (UniProt Q72IM4); residues substituted to cysteine in the disulfide-trapped inward-facing NapA structure are underlined, and additional C-terminal residues retained after TEV cleavage are shown in bold.

MHGAEHLLEIFYLLLAAQVMAFIFKRLNQPVVIGEVLAGVLVGPALLGLVHEGEILEFLA
ELGAVFLLFMVGLETRLKDILAVGKEAFLVAVLGVALPFLGGYLYGLEIGFETLPALFLG
TALVATSVGITARVLQELGVLSRPYSRIILGAAVIDDVLGLIVLAVVNGVAETGQVEVGA
ITRLIVLSVVFVGLAVFLSTLIARLPLERLPVGSPLGFALALGVGMAALAASIGLAPIVG
AFLGGMLLSEVREKYRLEEPIFAIESFLAPIFFAMVGVRLELSALASPVVLVAGTVVTVI
AILGKVLGGFLGALTQGVRSALTVGVGMAPRGEVGLIVAALGLKAGAVNEEEYAIVLFMV
VFTTLFAPFALKPLIAWTERERAAKEG**SENLYFQ**

5.4.2 Protein expression and purification

Wild-type NapA was previously cloned into the vector pWaldoGFPe, which contains a TEV-cleavable C-terminal GFP-His_8 tag, and mutants were generated from this vector[31] with the QuikChange protocol (Agilent Technologies). Wild-type NapA and mutants were transformed into the *E. coli* strain Lemo21(DE3)[81] and overexpressed with the MemStar protocol[117].

Membranes containing overexpressed fusions were isolated from 6-l *E. coli* cultures and solubilized in 1% (w/v) dodecyl-β-D-maltopyranoside (DDM; Generon) for 1 h in buffer containing 1 $\times$ PBS and 150 mM NaCl. The suspension was cleared by ultracentrifugation at 200,000g for 1 h. The cleared supernatant was incubated with 1 ml of Ni-NTA Superflow resin (Qiagen) per 1 mg of GFP-His_8 and incubated for 3 h at 4°C after addition of 10 mM imidazole. Resin slurry was loaded onto a glass

Econo-Column (Bio-Rad) and washed in 1 × PBS buffer containing 0.1% DDM and 150 mM NaCl for 3 × 10 column volumes at 10, 20 and 45 mM imidazole concentrations, respectively. The NapA-GFP-His_8 fusion was eluted in two column volumes of 1 × PBS buffer containing 0.6% nonyl-β-D-maltopyranoside (NM; Generon), 150 mM NaCl and 250 mM imidazole. The eluted protein was dialyzed overnight in the presence of stoichiometric amounts of His_6-tagged TEV protease containing 5 mM DTT, in 3 l of buffer containing 20 mM Tris-HCl, pH 7.5, 150 mM NaCl and 0.5% NM. Dialyzed samples were passed through a 5-ml Ni-NTA His-Trap column (GE Healthcare), and the flow through containing NapA was collected. Protein was concentrated with concentrators with a relative molecular-mass cutoff of 100 kDa. To reoxidize all proteins to a similar level for functional studies, the purified proteins were incubated with 1 mM $CuSO_4$ for 30 min prior to loading onto a Superdex 200 10/300 gel-filtration column (GE Healthcare) equilibrated in buffer containing 20 mM Tris-HCl, pH 7.5, 150 mM NaCl and 0.03% DDM. For structural investigation, the protein was not incubated with $CuSO_4$ but was instead loaded directly onto a Superdex 200 10/300 gel-filtration column (GE Healthcare) equilibrated in buffer containing 20 mM Tris-HCl, pH 7.5, 150 mM NaCl and 0.6% NM.

5.4.3 Na^+/H^+ cysteine accessibility

To probe for cysteine accessibility, 5 μg of either purified wild-type NapA or cysteine mutants thereof was incubated with 5 mM maleimide-PEG-5K (methoxypolyethylene glycol maleimide; Sigma) at room temperature for 45 min. Incubated proteins were analyzed for reactivity on the basis of a shift in size after electrophoresis on 12% NuPAGE gels (Invitrogen) and staining with Coomassie blue. Cysteine accessibility was further assessed by incubation of 10 μg of either purified wild-type NapA or cysteine mutants thereof with 2 μg of CPM (7-diethylamino-3-(4'-

maleimidylphenyl)-4-methylcoumarin; Life Technologies) dye at 40°C for 15 min. Samples were subjected to SDS-PAGE, and in-gel fluorescent-band intensity (blue light: 480 nm) was used to quantify the level of free sulfhydryl groups. Original images of gels used in this study can be found in Supplementary Data Set 1.

5.4.4 Na^+/H^+ antiport activity

Na^+/H^+ antiport activity in proteoliposomes was carried out as described previously[31]. In brief, l-α-phosphatidylcholine lipids (Sigma) were diluted to 10 mg/ml with 10 mM MOPS, pH 6.5, 5 mM $MgCl_2$, and 100 mM KCl (MMK) at pH 6.5 and vortexed until mixtures were uniform. Lipids underwent eight cycles of freeze-thawing and were stored at 080°C until use. Prior to reconstitution, liposomes were extruded through polycarbonate filters (200 nm) and destabilized by the addition of sodium cholate to a final concentration of 0.65% (v/v). Either 100 μg of purified NapA or cysteine mutants thereof were co-reconstituted together with 100 μg of purified *E. coli* F_oF_1 ATP synthase in MMK buffer (10 mM MOPS-NaOH, pH 8.0, 5.0 mM $MgCl_2$, and 100 mM KCl) as described previously[31,85,86]. Typically, 50 μL of proteoliposomes was diluted into 1.5 mL MMK buffer containing 2.5 nM 9-aino-6-chloro-2-methoxyacrid- ine (ACMA) and 130 nM valinomycin. Fluorescence was monitored at 480 nm with an excitation wavelength of 410 nm in a fluorescence spectrophotometer (Cary Eclipse, Agilent Technologies). An outward-directed pH gradient (acidic inside) was established by the addition of final 0.2 mM (v/v) ATP, as monitored on the basis of change in ACMA fluorescence. After $\sim$3 min equilibration, the activity of wild-type NapA and mutants thereof was assessed by the dequenching of ACMA fluorescence after addition of the indicated concentrations of NaCl or LiCl. Addition of 20 mM NH_4Cl led to near-complete dequenching. The apparent affinity K_m for Na^+ or Li^+ was measured at the indicated concentrations in the presence or

absence of 5 mM DTT. The data were fitted from duplicate measurements to the Michaelis-Menten equation by nonlinear regression with GraphPad Prism software. The K_m values reported are the mean ± s.e.m. of the fit.

To assess the propensity for disulfide-bond formation, antiport activity for wild-type NapA and single- and double-cysteine mutants were first assessed in MMK buffer containing 5 mM DTT essentially as described above, after the addition of saturating 10 mM $LiCl_2$. DTT was removed from proteoliposomes containing either wild-type NapA or cysteine mutants by their passage through two G-25 columns (GE Healthcare). Antiport activities under these nonreducing conditions were reassessed by using the equivalent amount of proteoliposomes by the addition of saturating, 10 mM $LiCl_2$ in MMK buffer without DTT. To confirm that the loss of antiport activity for the double-cysteine mutant was because of disulfide-bond formation rather than a loss of sample, DTT was further added back to proteoliposomes when it had been removed. Each experiment was performed three times from three independent purifications. The data shown represent the mean ± s.e.m. from three technical replicates from one of three independent experiments.

5.4.5 Isothermal calorimetry

All measurements were made on a Micro-200 ITC (MicroCal, Malvern), with an experimental setup similar to that previously described for NhaA [28]. 150 μl of the double-cysteine V31C I130C mutant at 300 μM was diluted to 2 mL in buffer containing 50 mM Bis-Tris-propane, 150 mM KCl, and 0.03% DDM at the desired pH, and was concentrated to ∼150 μl with concentrators with a relative molecular-weight cutoff of 100 kDa. This washing step was repeated a minimum of six times. The last flow through was used to dilute a 2 M stock solution of LiCl or NaCl and the protein to the concentrations used for the runs. Disulfide-bond formation was

confirmed after these steps by the maleimide-PEG shift assay and ACMA transport assay both before and after the ITC run. Protein at 90-150 *mu*M was loaded into the sample cell, and 15 mM LiCl or 40 mM NaCl was loaded into the injection syringe. The system was equilibrated to 20°C with a stirring speed of 600 r.p.m. Titration curves were initiated by a 0.5-μl injection and were followed by 2-*mu*l or 2.8-*mu*l injections every 200 s. Background corrections were obtained by injecting LiCl or NaCl into buffer and buffer into protein with the same parameters. ORIGIN 7 was used to integrate, correct and normalize the heat for each injection and fit the data to a single-site binding isotherm with a fixed protein/ligand stoichiometry of 1, excluding the peak from the first injection.

5.4.6 Crystallization and diffraction

Crystals of the double-cysteine V31C I130C mutant were grown at 20°C with the hanging-drop vapor-diffusion method. 1.2 μl of protein at 15 mg/ml was mixed 1:1 with reservoir solution containing 50 mM glycine, pH 9.5, 0.1 M NaCl, and 34-38% PEG 300. Crystals were dehydrated in 36-40% PEG 300 (2% increments overnight) and subsequently flash frozen in liquid nitrogen before data collection.

For crystallization of NapA in lipidic mesophases, the triple-cysteine mutant (M20C V166C V326C), which has previously been used for determining the outward-facing structure (PDB 4BWZ) and has further been shown to have wild type-like activity[31], was purified and concentrated to 30 mg/mL and mixed with 7.7 MAG (lot 141MG(7.7)-15, Avanti) with a coupled syringe-mixing device to form the cubic phase at a protein solution/lipid ratio of 1:1 (w/w). Crystallization trials were set up by dispensing 50 nL cubic phase onto a 96-well Laminex glass plate (MD11-50, Molecular Dimensions), which was then covered with 800 nL of precipitant solution 0.1 M MES, pH 6.5, 0.1 M NaSCN, 40% (v/v) PEG 400, with a Mosquito LCP robot

(TTP Labtech). Plates were sealed with a Laminex glass cover (MD11-52, Molecular Dimensions) and were stored at 20°C.

5.4.7 Structure determination

Data were collected at Diamond Light Source. Diffraction data were indexed and integrated with XDS[89] and scaled with Aimless[161]. Initial phases were obtained by molecular replacement in Phenix[97] (autoMR) with the dimer domain of NapA (PDB 4BWZ model) and the core domain as separate search models. Refinement was carried out in autoBUSTER (`http://www.globalphasing.com/buster/`)/Phenix[97], and manual rebuilding was carried out in Coot[96]. One TLS group was used during the refinement of the inward-facing conformation. The quality of the model was assessed by MolProbity[162] throughout this process. All residues from final models were found in allowed region of the Ramachandran plot, and both structures were ranked in the 100th percentile by the MolProbity server.

5.4.8 Structure analysis

Superimposition, figures and movies were generated with PyMOL (`http://www.pymol.org/`). The program O[94] was used to calculate the angle through which the core domain of the inward-facing structure needed to rotate to be superposed on the same domain of the inward-facing structure after the respective structures had been superposed on their dimer domains. For this analysis, the domains were first superposed in LSQMAN[98], such that all matching Cα pairs were less than 3.8 Å apart after superposition. To visualize the positional sequence conservation in NapA outward- and inward-facing structures, the ConSurf server was used with default settings[163].

5.4.9 Molecular dynamics simulations

All-atom, explicit solvent MD simulations were performed with Gromacs 4.6.5[52] and the CHARMM36 force field for the protein with the CMAP correction[42–44], ions and lipids[45] together with the CHARMM TIP3P water model[42]. The NapA dimer was simulated in an ∼4:1 1-palmitoyl-2-oleoylphosphatidylethanolamine (POPE)/1-palmitoyl-2-oleoylphosphatidylglycerol (POPG) bilayer. Simulations of the outward-facing structure were based on the LCP crystal structure. Simulations were also performed with the inward-facing disulfide-linked V31C I130C structure and an inward-facing model of the wild-type protein that was obtained by changing C31 back to V31 and C130 to I130. Simulations were run for 1 μs. Wild-type simulations were independently repeated twice for 0.3 μs, starting from the same initial system conformation but with differing random number seeds. All simulations are summarized in Supplementary Table 5.3.

Protein structures were embedded into the model membrane with a coarse-grained self-assembly protocol[103] and subsequent conversion of the lipids to an all-atom representation[50], as described for our previous simulations of NapA[31]. The total system consisted of about 120,000 atoms in an orthorhombic simulation cell with a free NaCl concentration of ∼250 mM. Most titratable residues were simulated in their default charge states, as predicted by PROPKA 3.1[131] for pH 7.8, including charged forms of residues D156 and K305 near the ion-binding site. The binding-site residue D157 was simulated in its charged form in the inward-facing and outward-facing conformation, despite a prediction of a neutral charge state in the outward-facing conformation, because our previous work had suggested that it would be charged (deprotonated) when a sodium/proton antiporter binds a Na^+ ion[31,32].

Equilibrium MD simulations were performed after energy minimization and ∼15 ns of equilibration with position restraints (harmonic force constant 1,000 kJ mol^{-1} nm^{-2} on all heavy protein atoms). All simulations were carried out under periodic boundary conditions at constant temperature (T = 310 K) and pressure (P = 1 bar). A velocity rescaling thermostat[62] was used with a time constant of 1 ps, and three separate temperature-coupling groups were used for protein, lipids and solvent. A Parrinello-Rahman barostat[63] with time constant 5 ps and compressibility 4.6 × 10^{-5} bar^{-1} was used for semi-isotropic pressure coupling. The Verlet neighbor list with a cutoff of 1.2 nm was updated every 20 steps. Coulomb interactions were calculated with the smoothed particle mesh Ewald (SPME) method[105] with a real-space cutoff of 1.2 nm, and interactions beyond the cutoff were calculated in reciprocal space with a fast Fourier transform on a grid with 0.12-nm spacing and fourth-order spline interpolation. The Lennard-Jones potential was shifted to zero at a cutoff of 1.2 nm, without applying any dispersion corrections[164]. Bonds to hydrogen atoms were constrained with the P-LINCS algorithm[55] (with an expansion order of four and two LINCS iterations) or SETTLE[56] (for water molecules). The classical equations of motions were integrated with the leapfrog algorithm with a time step of 2 fs.

Nonbonded parameters for lipid simulations can be critical, and artifacts have been reported with incorrect choices of these parameters[164]. We carefully selected a set of parameters compatible with the Verlet neighbor-list scheme in Gromacs (which is required for running on GPUs and for high performance with the CHARMM TIP3P water model on conventional hardware) while remaining close to the original values of the CHARMM force field. The simulation parameters were validated with simulations of pure POPC (used for validation instead of POPG) and POPE bilayers of 288 lipids and 50 water molecules per lipid for ∼250-ns length, which yielded areas per lipid within 1 $Å^2$ (POPC) and 3 $Å^2$ (POPE) of the original CHARMM36 values[45] (data

not shown). Furthermore, our values differed from the experimental areas per lipid by no more than 2 $Å^2$, thus indicating that our parameter settings in Gromacs produced results of similar quality to the native simulations in CHARMM.

5.4.10 Analysis of molecular dynamics simulations

MD simulations were analyzed with code based on MDAnalysis[106] (`http://www.mdanalysis.org/`) and MDSynthesis (`https://github.com/datreant/MDSynthesis/`). The movement of domains was assessed in the coordinate system associated with the membrane. In the MD simulations, the average normal of the bilayer was parallel to the z axis of the fixed simulation coordinate system, and hence we performed all calculations with the z coordinates of the centers of mass of the core and dimer domain and the bilayer itself. Probability distributions $f_Z(z)$ of a coordinate Z were calculated as Gaussian kernel density estimates (KDE) from the raw simulation data (obtained at 1-ps intervals). The kernel width of the KDE was chosen according to Scott's criterion as implemented in the `scipy.stats.gaussian_kde()` function from the SciPy package (`http://www.scipy.org/`). The joint difference distribution of the random variable $\Delta Z = Z_{\mathrm{OF}} - Z_{\mathrm{IF}}$ was calculated as the convolution $f(\Delta Z) = \int dz\, f_Z^{\mathrm{OF}}(z)\, f_Z^{\mathrm{IF}}(z - \Delta Z)$. In order to calculate the z shift of the core domain relative to the dimer domain between the two crystal structures, it was still necessary to orient the two structures relative to the membrane. We performed the orientation of the crystal structures in two different ways. (i) The dimer domain was superimposed on frames from the MD simulations, and distributions of the z position and the relative shift were calculated as described, yielding $\Delta Z = 7.2 \pm 1.1$ Å (mean ± s.d.). (ii) The dimer domain was superimposed on the oriented outward-facing NapA structure (PDB 4BWZ) as provided by the Orientation of Proteins in the Membrane database (OPM, `http://opm.phar.umich.edu/`)[165], which uses an electrostatic model and a

simple low-dielectric model for the membrane. This static picture yielded $\Delta Z = 7.1$ Å, in full agreement with the analysis based on the orientation of the NapA dimer in the explicit membrane in the MD simulations.

Binding of Na^+ ions was assessed with a distance criterion: any Na^+ ion within 3 Å of any carboxyl oxygen atom of either D157 or D156 was considered to be bound; i.e., an ion was considered to be bound if it contained one of the aspartate carboxylates in its first coordination shell. The probability of observing at least one ion within 3 Å of any of the O_δ of D157 or D156, P_{bound} (shown in Supplementary Figure 5.10 and Supplementary Table 5.4) was calculated as the bound time divided by the total simulated time. To better estimate P_{bound}, we independently repeated simulations an additional two times for 300 ns. However, small but infrequent conformational changes such as movements of parts of helices 4 and 11 affect binding-site accessibility and binding (for example, the absence of any binding in protomer B in the inward-facing conformation in the 1-μs simulation, as shown in Supplementary Table 5.4). Such infrequent events are not well sampled, even in 1 μs of simulation. We therefore decided to weigh all simulations equally for the purpose of computing averages of Pbound over different initial conditions (Supplementary Figure 5.10) and therefore used only the first 300 ns of the 1-μs simulations together with the full 300 ns of the two repeats. For the analysis of coordination geometry, all trajectory frames were used that contained a bound Na^+ ion. The ion-oxygen radial distribution function was computed from all oxygen atoms (protein and water) within 10 Å of the ion. Per-residue contributions (in which all water molecules count as identical) to the first coordination-shell number (oxygen atoms within 3 Å of the ion) were computed as the total residence time of the oxygen in the first solvation shell divided by the total time that an ion was bound. Images showing simulation data were prepared with VMD[107] and the Bendix plugin[108].

5.5 Supplementary Information

conformation	sequence	run length (µs)	repeats	total (µs)
inward facing	wild type	1.0	1	1.0
	wild type	0.3	2	0.6
	V31C I130C	1.0	1	1.0
outward facing	wild type	1.0	1	1.0
	wild type	0.3	2	0.6

Table 5.3: Molecular Dynamics Simulations.

Conformation	Repeat number	Length (ns)	Protomer	P_{bound}
Outward-facing	1	1000	A	15%
			B	3%
	2	300	A	0%
			B	0%
	3	300	A	0%
			B	0%
Inward-facing	1	1000	A	59%
			B	0%
	2	300	A	94%
			B	80%
	3	300	A	98%
			B	85%
Inward-facing (V31C-I130C)	1	1000	A	79%
			B	92%

Table 5.4: Observation of bound Na^+ from MD simulations in multiple conformations. The probability for observing at least one ion within 3 Å of any of the Oδ of D157 or D156 (P_{bound}) was calculated as the fraction of the bound time relative to the total length of the trajectory. Simulations were started without a Na^+ ion within less than 10 Å distance of D157 so all observed binding is spontaneous. K305, D156, and D157 were simulated in their charged state.

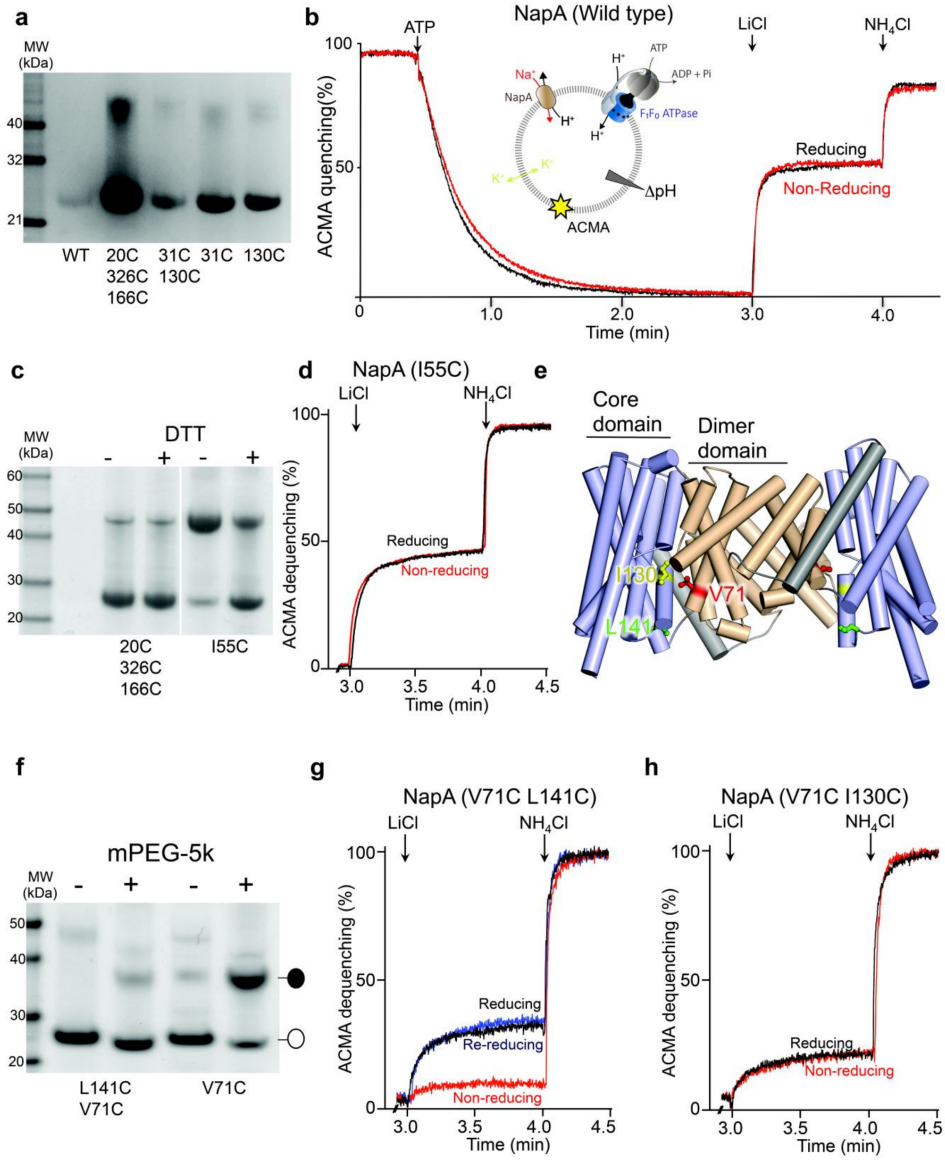


Figure 5.7: Assessing disulfide-bond formation in NapA cysteine mutants. a. Reactivity of purified wild type NapA and single, double, and triple cysteine mutants thereof to the N-[4-(7-diethylamino-4-methyl-3-coumarinyl)phenyl]-maleimide (CPM) dye (see Methods). b. Representative ACMA fluorescence traces of NapA wild type antiport activity in proteoliposomes as outlined in the schematic (insert adapted from Lee et. al., 2013 [31]). The addition of ATP (0.5-3 min), NaCl/LiCl (3rd min) and NH_4Cl (4th min) is indicated by arrows and labeled. Black trace (presence of DTT; reducing) and the red trace (after DTT removal; non-reducing). c. SDS-PAGE analysis of a triple cysteine mutant of NapA (used previously for phasing/structure determination, Lee et. al., 2013 [31]) and the dimer domain mutant I55C in the presence or absence of DTT as labeled. d. Representative Li^+ catalysed antiport activity of the I55C mutant under either reducing (black trace) or non-reducing (red trace) conditions. e. Cartoon representation of the inward-facing NapA structure showing the positions of V71 (red sticks) in the dimer domain (wheat) and I130 (yellow sticks) and L141 (green sticks) in the core domain (blue). f. Reactivity of purified single and double cysteine mutants V71C and V71C L141C (open circle) to mPEG-5K (black circle). g-h. Representative ACMA fluorescence traces for V71C L141C and V71C I130C mutants in the presence of DTT (reducing, black trace), after DTT removal (non-reducing, red trace) and after re-addition of DTT (reducing, blue trace).

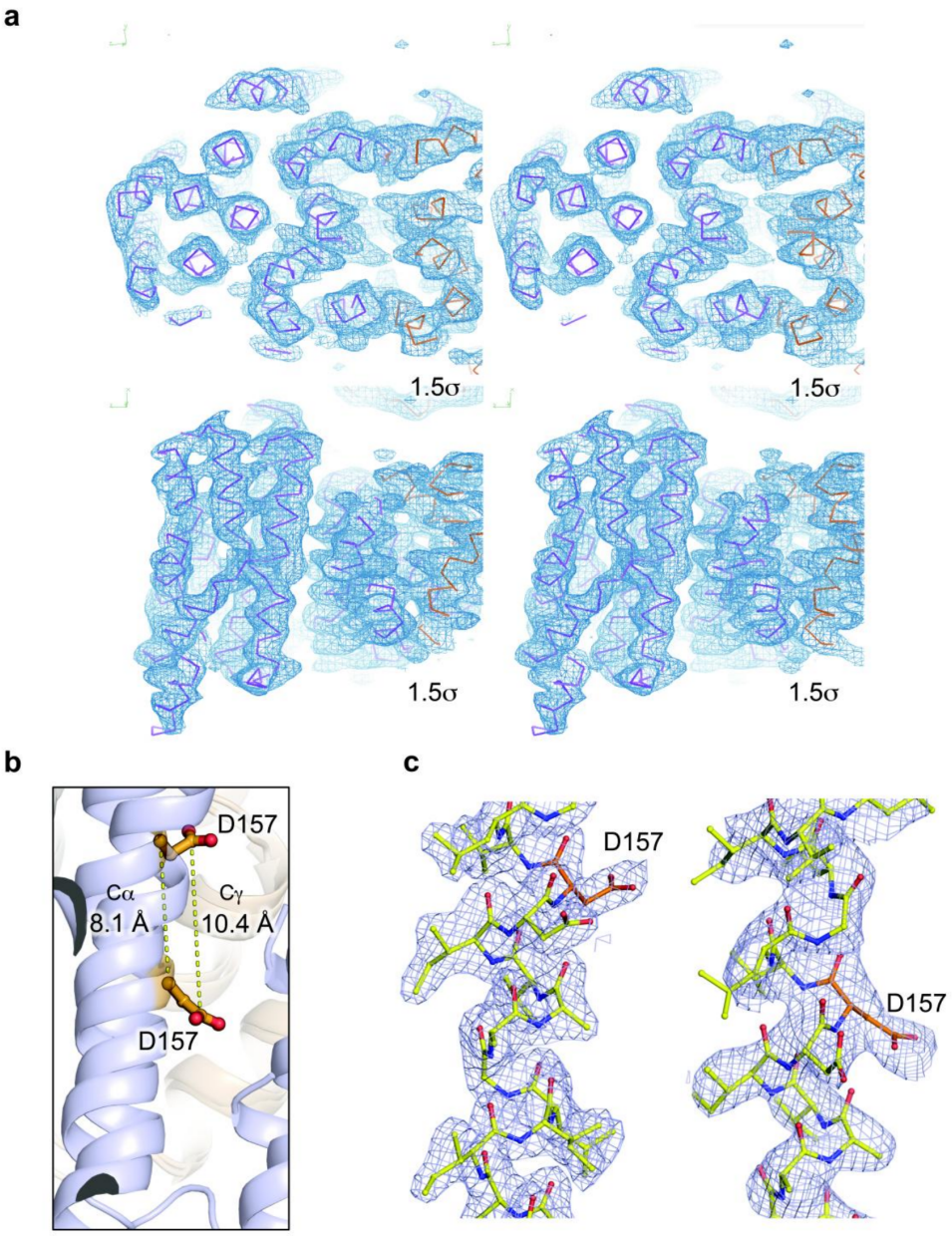


Figure 5.8: Electron density maps of the disulfide-locked inward-facing NapA structure and TM5 in both conformations. a. Stereoview showing $2F_oF_c$ electron density calculated after the final refinement (contoured at 1.5σ level) with the final inward-facing structure overlaid (purple core; orange dimer) from the extracellular side (top) and from the side-view (bottom). b. NapA structures in the inward-facing and outward-facing conformation as superimposed on their respective dimer domains. The ion-binding D157 (orange) displacement between the two states is shown between the position of their Cα atom or their Cγ atom. c. $2F_oF_c$ electron density calculated after the final refinement (contoured at 1.0σ level) around TM5 in the outward-facing (left panel) and inward-facing (right panel) conformation.

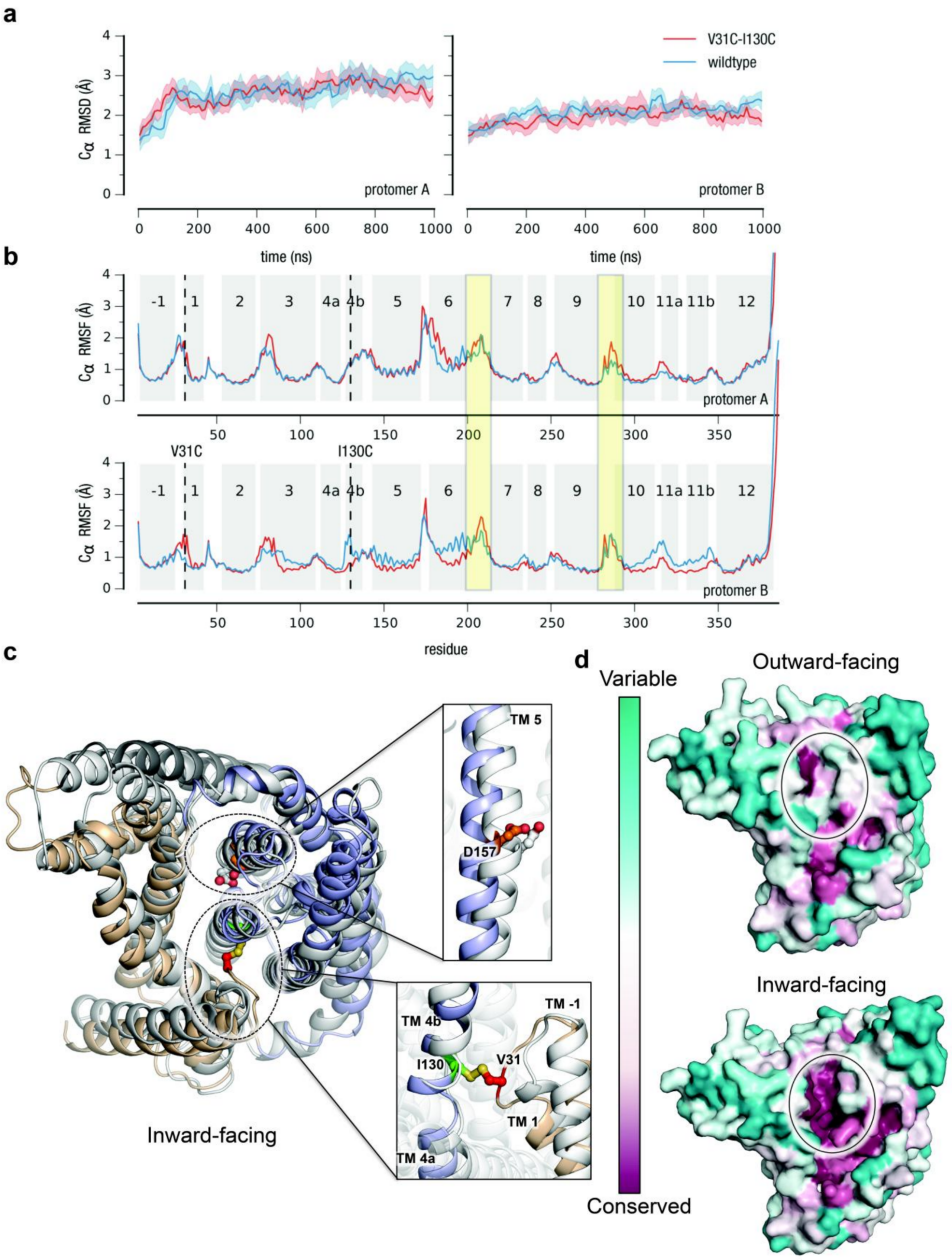
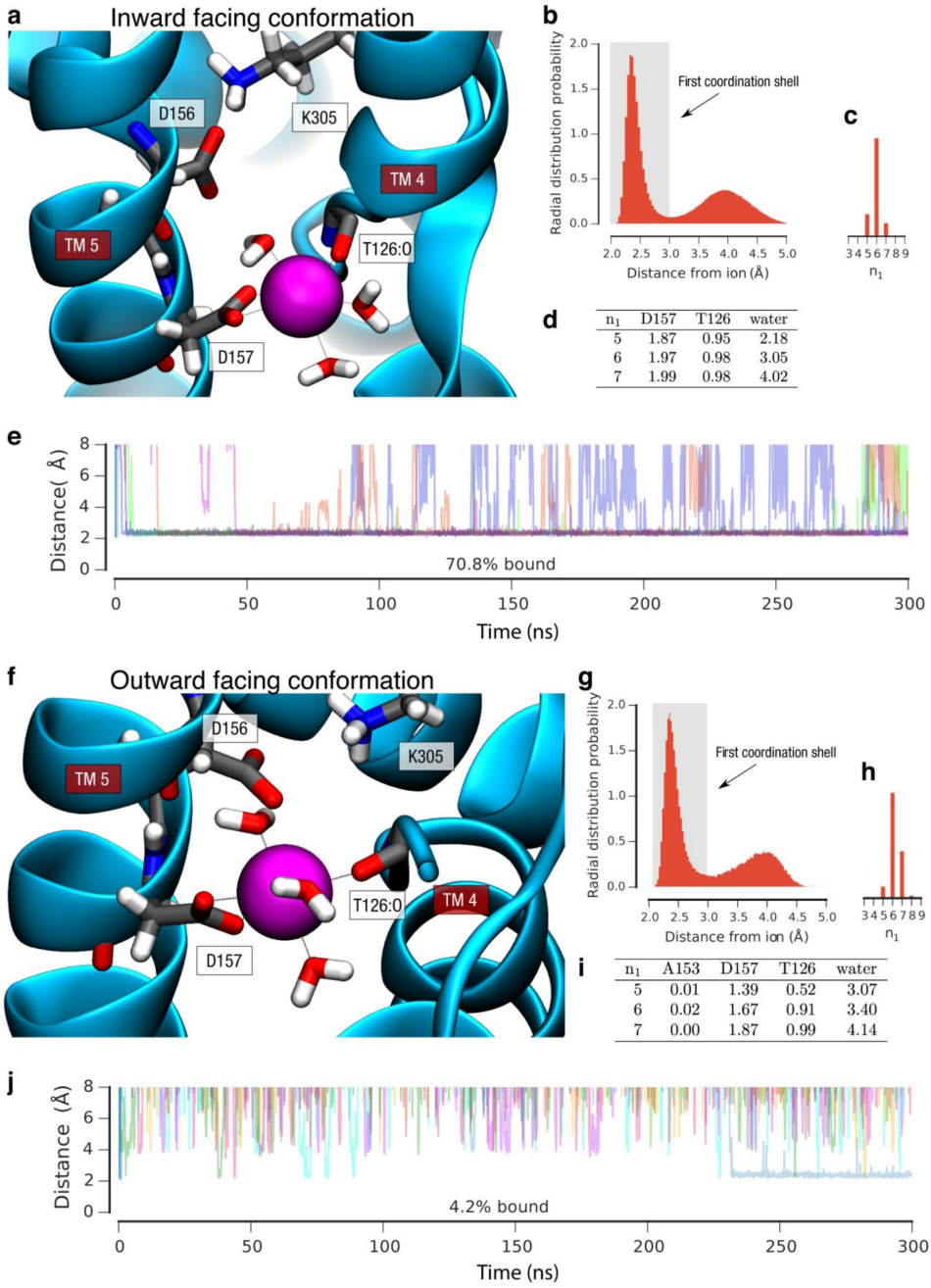


Figure 5.9: Analysis of the disulfide-trapped structure of NapA in an inward-facing conformation. a. MD simulations of the disulfide cross-linked crystal structure V31C I130C (red) and the wild-type structure (blue), which was modelled from the cross-linked structure, were compared for protomer A and B. Cα r.m.s.d as a function of time, after r.m.s.d-fitting to the crystal structure. Data were averaged over 10-ns windows and bands indicate the 95% quantile of the data. b. Per-residue root mean square fluctuations (r.m.s.f.), relative to the average of all MD conformations. Dashed lines indicate the positions of the cross-linking residues C31 and C130. Gray boxes show the locations of helices TM -1 to 12. Two yellow boxes indicate putative hinge regions between dimer and core domain. Two independent repeat simulations of the wild type of 300 ns length (not shown) display the same behavior with r.m.s.d values between 2 Å and 2.5 Å and very similar r.m.s.f. profiles. c. Superimposition of disulfide-trapped NapA (wheat dimer, light-blue core) and inward-facing MjNhaP1 (grey; PDB 4CZB) structures across both core and dimer domains evenly. Top inset shows the relative positioning of the ion-binding aspartate residue D157 in NapA (orange stick) and equivalent aspartate D161 in MjNhaP1 (grey stick). Bottom inset shows the region around the disulfide formed between C130 (green stick) and C31 (red stick) residues in NapA and the equivalent helices in MjNhaP1. d. Cytoplasmic view showing the highly conserved residues in NapA that becomes accessible upon the cavity opening to the inside in the inward-facing disulfide-trapped structure (below), as compared to the published outward-facing structure (above).



n_1	D157	T126	water
5	1.87	0.95	2.18
6	1.97	0.98	3.05
7	1.99	0.98	4.02

n_1	A153	D157	T126	water
5	0.01	1.39	0.52	3.07
6	0.02	1.67	0.91	3.40
7	0.00	1.87	0.99	4.14

Figure 5.10: Ion coordination as seen in MD simulations. a-e: Inward facing conformation. a. Representative snapshot of a Na^+ ion (magenta) in the putative binding site. b. Distribution of Na^+-oxygen distances with oxygen atoms from any protein residues or water molecules; only bound Na^+ ions were considered, i.e., ions within 3 Å of the carboxyl oxygen atoms of D157 or D156. Oxygen atoms within 3 Å radius form the first coordination shell of the Na^+ ion. c. Distribution of observed Na^+ coordination numbers n_1 (the number of oxygens in the first coordination shell). d. Average contribution of oxygen ligands to n_1 for the most frequently observed coordination numbers. e. Time series of the distance of any Na^+ ion to the closest oxygen atom in the carboxyl moiety of D157 or D156. Data for protomer A and B from the first 300 ns of three independent simulations were overlaid and shown in different colors. On average, an ion was bound for 70.8% of the simulation time. f-j: Outward facing conformation. The same quantities as in a-e are displayed; however, on average an ion was only bound for 4.2% of the simulation time (j).

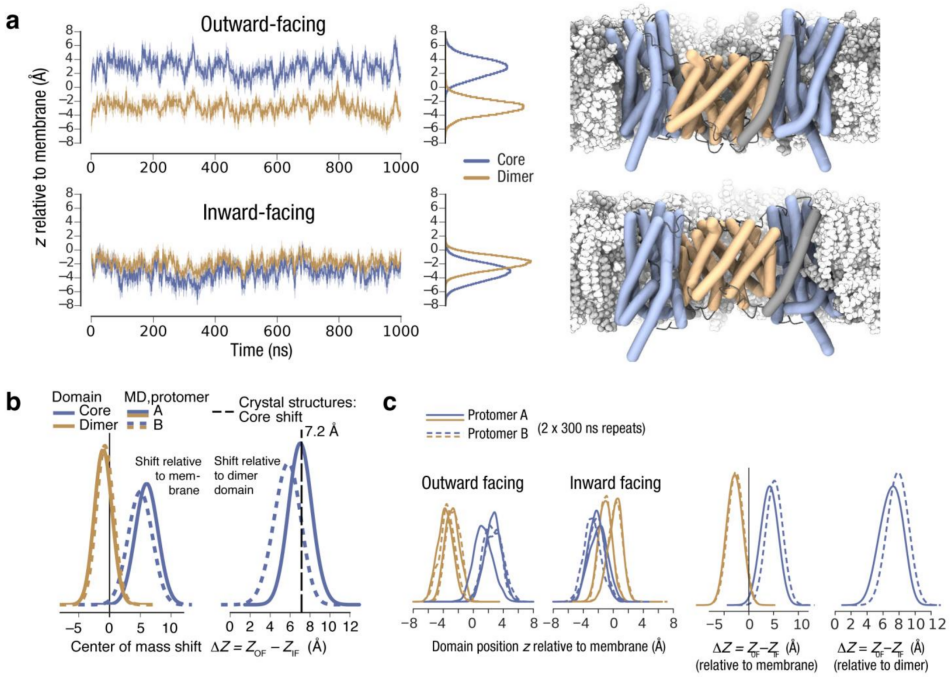


Figure 5.11: Positions of the core and dimer domains in MD simulations. a. Data from 1 μs MD simulations of the outward facing (top) and inward facing (bottom) conformation. Time series and histograms of the z component of the center of mass of the core (blue) and dimer (orange) domain (left panels) indicate equilibrium fluctuations of the domains relative to the instantaneous center of mass of the lipid bilayer. Only protomer A is shown; time series for protomer B are similar. Snapshots from MD simulations show the relative positions of the domains (blue, orange) relative to the membrane (white POPE and gray POPG lipids). b. The z component of the centre of mass of the core and dimer domain in equilibrium MD simulations of outward facing (OF) and inward facing (IF) NapA was used to characterize domain motions. Protomer A (chain A, solid line) and protomer B (chain B, broken line) are shown separately. Left: distribution of the difference in z coordinate relative to the instantaneous center of mass of the lipid bilayer between OF and IF simulations. Right: distribution of the difference in z coordinate relative to the instantaneous center of mass of the dimer domain between OF and IF simulations. The dashed vertical line indicates the value obtained from comparing the LCP OF structure with the disulfide-linked IF structure. c. Distributions of domain positions from two additional independent repeat simulations of 300 ns length.

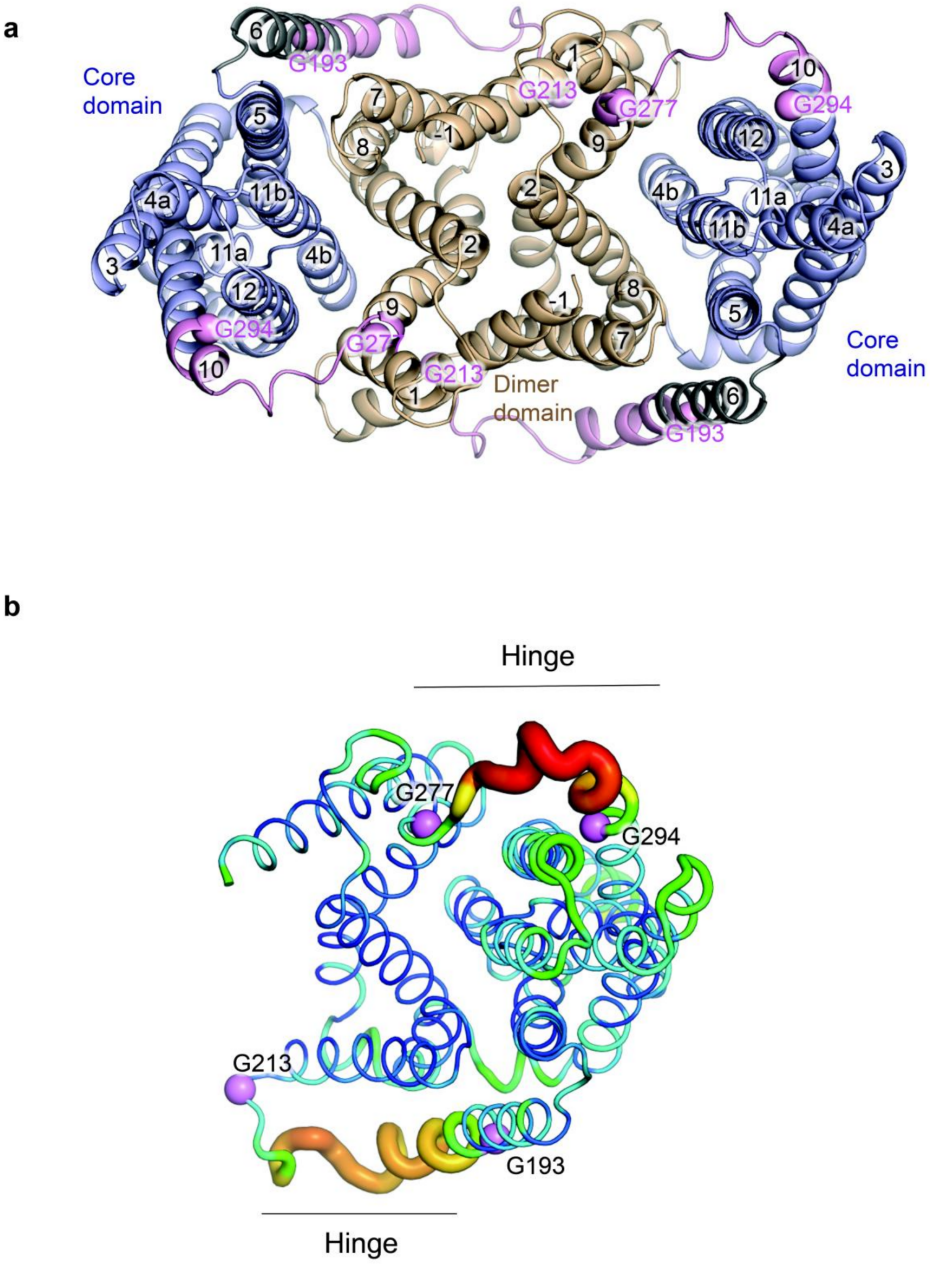


Figure 5.12: The mobile hinge regions that link the core and dimer domains in NapA. a. Ribbon representation of outward-facing NapA (dimer light orange; core light-blue). The linking helix TM6 is shown in grey and the mobile connecting hinge regions shown in pink. The position of flanking hinge glycine residues are labelled and shown as pink spheres. b. Ribbon representation of the high-resolution outward-facing (LCP) NapA structure rendered by b-factors scale (rainbow colour scale and tube diameter). The position of flanking hinge glycine residues are labelled and shown as pink spheres.

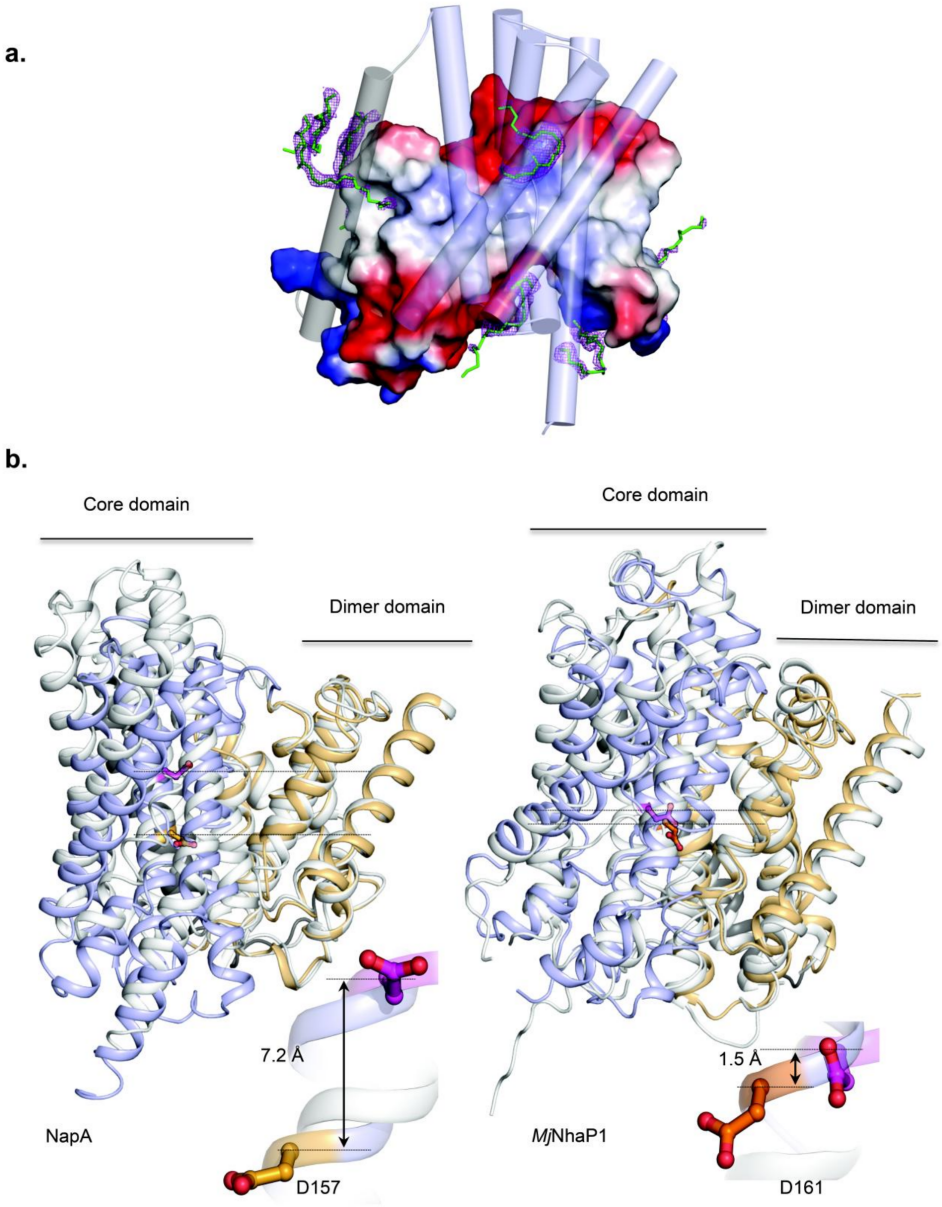


Figure 5.13: Location of bound (LCP) lipid and comparison of the extent of core movements between NapA and MjNhaP1 structures. a. Outward-facing NapA, as viewed on a right-angle from the membrane plane, is represented as an electrostatic surface for the dimer domain and as transparent cartoon helices in blue for the core domain. Non-protein density ($2F_oF_c$ omit map at 1.5σ level) at core-dimer domain interface was fitted as modified monolein MAG7.7 (green sticks). b. The inward-facing structures of NapA (left: dimer light-orange and core blue) and MjNhaP1 (right: dimer light-orange and core blue; PDB 4CZB) were superimposed on the dimer domain of their respective outward-facing structures (shown in white; PDB 4D0A MjNhaP1 and PDB 5BZ3 NapA). The vertical distance between the respective ion-binding aspartates D157 in NapA and D161 in MjNhaP1 are shown as an inset below.

Chapter 6

QUANTIFYING THE STRENGTH OF NA^+ BINDING BETWEEN CONFORMATIONS IN NAPA

This chapter details work that is not yet published. It presents a study of the binding free energy of Na^+ to *Thermus thermophilus* NapA, in both its inward- and outward-facing conformations[34], for a variety of protonation states of the binding-site residues. Our intention is to quantify the strength of binding in these states, with a view toward comparing directly to experimentally-obtained apparent binding affinities (apparent K_a).

To accomplish this, we perform alchemical binding free energy calculations using data from hundreds of independent simulations. This effort drove the development of new software infrastructure, enabling us to perform many simulations at large-scale throughput across multiple compute resources (see Section 7.4) with relative ease.

We show that binding of Na^+ to the outward-facing state of NapA appears substantially weaker than binding to the inward-facing state, implying that the apparent binding affinity for Na^+ is asymmetric across conformations. This is consistent with and advantageous for the physiological role of the transporter.

This work was entirely carried out by me, and is considered to be in preparation for publication.

6.1 Introduction

Transporters such as NhaA and NapA are driven by the electrochemical gradients of their substrates, in this case Na^+ and H^+[69]. Under physiological conditions, it is the free energy available from the electrochemical gradient of H^+ that drives extrusion

of Na^+ against its own gradient[69]. Furthermore, the existing evidence suggests that both substrates competitively bind to a common site in these transporters[27].

It is now well-established that Na^+ binds to D157 in NapA (D164 in NhaA)[22,28,31,32,34,147]. A question that remains unanswered is how strongly, in quantitative terms, Na^+ binds to these transporters, and if the strength of binding varies by conformation and protonation state of the binding site residues. In light of the hypothetical model of proton binding in which the conserved lysine, K305 in NapA (K300 in NhaA), participates directly as a proton transporter[32,33] (Figure 4.10), it is important that these configurations can be quantitatively connected to experimental measurements.

Experimentally-obtained binding affinities ($K_\mathrm{d} = K_\mathrm{a}^{-1}$) are of particular interest, since these can be connected directly to binding free energies as $\Delta G^\circ_\mathrm{bind} = \beta^{-1} \ln(K_d V^\circ)$[166]. Apparent binding affinities for both Li^+ and Na^+ have been experimentally measured for cysteine-locked mutants of NapA in the inward-facing conformation[34]. These affinities alone do not say anything about where the protons bind, but using free energies calculated from simulations, we can test the hypothesis that the lysine binds protons while also quantifying the strength of sodium binding.

To obtain binding free energies, we employ a so-called *alchemical* simulation approach. Because this approach requires many, perhaps hundreds, of expensive simulations, we restricted our focus in this study to NapA, for which we possess experimental structures of both inward- and outward-facing conformations[31,34].

6.2 Methods

The hypothetical mechanism in which the conserved lysine, K305, serves as a proton carrier is shown in Figure 4.10. The shorthand charge state names (S1, S2, and S3) are defined in Table 6.1. Under physiological conditions in the inward-facing (IF) conformation Na^+ binds to the charged aspartate D164 (state IFS2). The

name	charge state
S1	D157(0), D156(−), K305(+)
S2	D157(−), D156(−), K305(+)
S4	D157(−), D156(−), K305(0)

Table 6.1: The charge states of NapA that are represented in the transport mechanism in which K305 transports H^+. We use shorthand names S1, S2, and S4 to refer to these throughout the text.

binding of Na^+ reduces the pK_a of K305 and it loses its proton, becoming neutral and disrupting the salt bridge it forms with D156. This state (IFS4) allows the core domain to move relative the the dimerization domain, and the transporter eventually switches to the outward-facing (OF) conformation. In this state (OFS4) the bound Na^+ is eventually released, increasing the pK_a of K305, resulting in its protonation (state OFS2). Protonation of D157 (OFS1) allows the core domain to move relative to the dimerization domain, and the transporter eventually switches back to the inward-facing (IF) conformation (IFS1). The cycle completes when D157 loses its proton (state IFS2).

To quantify the strength of ion binding in this cycle, we must compute estimates of the binding free energy for each of the configurations shown in Figure 4.10. Because only four of these six states naturally bind Na^+ in simulation (S2 and S4, in both conformations), we began with these. To compute estimates for the S1 states, we will need to employ a different strategy (see Section 6.4).

6.2.1 The alchemical binding pathway

The free energy difference (ΔG) between two states, such as A) a Na^+ ion and a protein, not bound, and B) a Na^+ ion and a protein, bound, is independent of the path taken to go from one to the other. A free energy is a state function, so to calculate the difference in free energies between two states in a thermodynamic system, one can take any route, in principle, between them. In practice, it is important to take

a route that one can obtain converged results from with as little computation as possible, even if that route is non-physical.

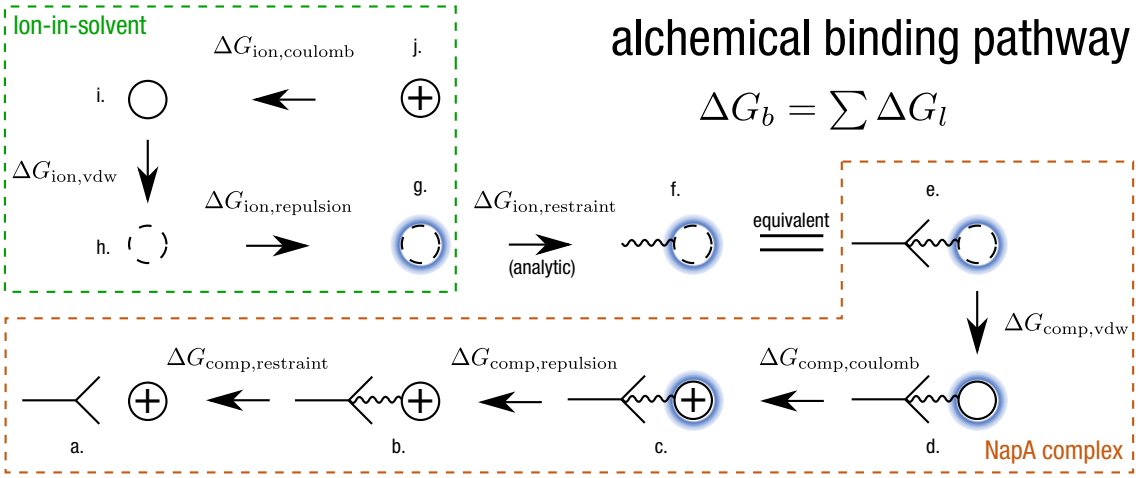


Figure 6.1: The alchemical pathway chosen for calculating the absolute binding free energy of Na^+. This pathway requires simulations performed with the full NapA complex (orange), as well as simulations of only the ion in NaCl solution (green).

To compute free energies of binding for Na^+ to the binding site of NapA, we choose a non-physical, *alchemical* pathway for calculating ΔG_b. This pathway is shown in Figure 6.1. Although the pathway is shown with arrows pointing in the direction required to calculate the free energy of binding (ΔG_b), it is easier to discuss it in the opposite direction for unbinding ($\Delta G_u = -\Delta G_b$). Starting in the lower-left part of Figure 6.1, (a) we begin with an Na^+ ion naturally bound to NapA at D157, as seen in equilibrium MD simulations. We then (b) add a restraint to the ion (see Section 6.2.3) to keep it in place once we decouple its interactions with the surroundings later in the pathway. Next, we (c) add a repulsion on the restrained ion that only acts on the other Na^+ ions in the system (see Section 6.2.2), so that another ion does not bind to the protein once the restrained ion has been decoupled from everything else; this is needed to compute a true single-ion free energy. Finally, we (d) remove the Coulomb interaction from the ion, and then (e) the Van der Waals interaction. At this point

the ion is no longer interacting with the protein, and only weakly interacting with the other Na^+ in the system.

The remainder of the cycle is performed with simulations of the ion in water, as part of an NaCl solution. The decoupled ion with repulsion in (e) is equivalent to the decoupled ion with repulsion in solution, still spatially-restrained (f). We can calculate the cost of removing the restraint in the isotropic conditions of the ion in solution analytically (see Section 6.5.1), resulting in (g) a free ion with a repulsion toward the other Na^+ in the system. We then (h) remove the repulsion to give a completely decoupled ion. Finally, we (i) add back the ion's normal Van der Waals interactions with the rest of the system, and then (j) return its Coulomb interaction as well. The end points of this cycle give (a) the bound state, and (j) the unbound state for the system. This "double-annihilation" approach is now well-established in the field for calculating absolute binding free energies[166,167]. The Na^+ repulsion, however, is a novel feature, and the way it is introduced has a minimal impact on the resulting free energies (See Sections **??** and 6.3).

Each leg of this pathway features a change in the system Hamiltonian, and over each leg we can obtain a reasonably-converged ΔG_l. A given leg represents many simulation windows, each performed with a Hamiltonian at a point along the linear interpolation between the Hamiltonians on each end of the leg:

$$H_{l\lambda} = (1 - \lambda)H_{l0} + \lambda H_{l1} \tag{6.1}$$

Note, however, that for switching off the Van der Waals interaction using a soft-core Lennard-Jones potential[168], which we use in this study, $H_{l\lambda}$ is not a simple linear interpolation as given in Equation 6.1. For each leg enough windows must be chosen, each with enough simulation sampling, to obtain converged results (see

Section 6.2.5) for ΔG_l. These can then be combined directly to obtain the total free energy of binding, ΔG_b:

$$\Delta G_b = \Delta G_{\text{ion}} + \Delta G_{\text{comp}} \tag{6.2}$$

where we have for ΔG_{ion}, the total free energy difference calculated from the simulation legs featuring the ion in solution:

$$\Delta G_{\text{ion}} = \Delta G_{\text{ion,coulomb}} + \Delta G_{\text{ion,vdw}} + \Delta G_{\text{ion,repulsion}} + \Delta G_{\text{ion,restraint}} \tag{6.3}$$

and for ΔG_{comp}, the total free energy difference calculated from the simulation legs featuring the full complex:

$$\Delta G_{\text{comp}} = \Delta G_{\text{comp,vdw}} + \Delta G_{\text{comp,coulomb}} + \Delta G_{\text{comp,repulsion}} + \Delta G_{\text{comp,restraint}} \tag{6.4}$$

6.2.2 Choosing the sodium repulsion

The alchemical pathway we chose features a leg in which a repulsion is applied to the special Na^+ ion that affects only the other Na^+ ions in the system. This was necessary to avoid the case in which another Na^+ ion binds to the protein when the interactions of the special Na^+ are switched off, since this would be sampling a bound state once more. This repulsion was chosen to be strong enough to keep the other Na^+ ions in the system as far away from the special Na^+ as they would natively be in equilibrium MD simulations when the special Na^+ is fully interacting, but no stronger so as to minimally impact the behavior of the system and come at little free energy cost.

To create this repulsion, we chose a $\sigma_{ij} = 0.75$ nm (see Section 2.2) for the Lennard-Jones interaction between the special Na^+ and the other Na^+ ions in the system. This repulsion allows a closest approach of Na^+ to the special Na^+ of about 6-7 Å in solution, which is consistent with the ∼5-7 Å closest approach observed for bound Na^+ in IFS2 NapA simulations. For reference, the value of σ_{ij} for sodium under the CHARMM36 force field is by default 0.25137 nm.

6.2.3 Choosing sodium restraints

In addition to adding a repulsion to the special Na^+, the pathway also requires that the special Na^+ be restrained within the binding site so that it does not drift away as its interactions with the surroundings are switched off. The choice of such restraints are important, as they must be simple enough to allow the analytic calculation of the $G_{\text{ion,restraint}}$ cost of their removal[166,167].

To minimize the increase in configuration space sampled by the fully-decoupled ion, we chose to apply both a distance and angle restraint. We distance restrained the ion with a harmonic potential ($U_d = \frac{1}{2}k_r(r - r_0)^2$) to a distance $r_0 = 2.75$ Å away from the Cγ carbon of D157, with force constant $k_r = 16,000.0$ kJ/nm^2. The angle restraint was applied to the angle formed by (special Na^+)-Cγ-Cβ on D157. This was a harmonic potential ($U_\theta = \frac{1}{2}k_\theta(\theta - \theta_0)$) with neutral angle $\theta_0 = 0.0$ rad and force constant $k_\theta = 23.0$ kJ/(mol rad^2).

These parameters were chosen to match as closely as possible the behavior of the ion when fully interacting and bound to D157. Choosing such parameters is not necessary for correct results in the resulting free energy calculation, but they assist in quicker convergence. The distance parameters were chosen by numerically optimizing the Bhattacharyya distance[169] between the distribution obtained for a choice of parameters k_r and r_0 ($P(r)dr \propto r^2 e^{-\frac{\beta k_r}{2}(r-r_0)^2}dr$) and the distance distribution observed

for binding to D157 with no restraint. The same approach was used for choosing the parameters k_θ and θ_0, using the distribution of θ ($P(\theta)d\theta \propto \sin(\theta)e^{-\frac{\beta k_\theta}{2}(\theta-\theta_0)^2}d\theta$) for such a potential compared to that observed from simulations. A range of parameter choices for k_r and k_θ were then empirically tested with production simulations, with 7 choices for each, giving $7 \times 7 = 49$ simulations.

While $\Delta G_{\text{comp,restraint}}$, the cost of adding the restraint to the ion while in complex with the protein, must be calculated from a leg of simulations, the cost of removing the restraint later in the pathway ($\Delta G_{\text{ion,restraint}}$) must be removed analytically as it is generally not possible to obtain a converged value from simulations[166,167]. The method for calculating $\Delta G_{\text{ion,restraint}}$ given our chosen restraints is detailed in Section 6.5.1.

6.2.4 Calculating ΔG_l with MBAR

Each leg of the alchemical pathway shown in Figure 6.1 features many simulation windows (from 3 to 41), the data from which can then be used to calculate a ΔG_l. A variety of methods are available for doing this[168,170], each varying in the observables used and the efficiency with which one can expect to obtain converged values[171]. The method we chose was the multistate Bennett acceptance ratio (MBAR)[172], an unbiased estimator that makes use of the data collected from all windows in a pathway leg to obtain low-variance estimates of the $\Delta G_{l\{m,n\}}$ between each pair of windows, (m, n).

The MBAR estimator is a set of estimating equations for the dimensionless free energies[172] that must be solved self-consistently for all $\hat{f}_m$ in the leg of windows:

$$\hat{f}_m = -k_B T \ln \sum_{j=1}^{K} \sum_{n=1}^{N_j} \frac{\exp[-u_m(\mathbf{x}_{jn})]}{\sum_{k=1}^{K} N_k \exp[\hat{f}_k - u_k(\mathbf{x}_{jn})]} \tag{6.5}$$

where the free energy difference going from window m to n in leg l is obtained with:

$$\Delta G_{l\{m,n\}} = k_B T(\hat{f}_n - \hat{f}_m) \tag{6.6}$$

The $u_m(\mathbf{x}_{jn})$ are the reduced potentials under the Hamiltonian of window m for each configuration $\mathbf{x}_{jn}$ sampled by window j, for which there are N_j. The reduced potential is defined as[172]:

$$u_m(\mathbf{x}) \equiv \beta_m[U_m(\mathbf{x}) + p_m V(\mathbf{x})] \tag{6.7}$$

We utilized the reference implementation of MBAR featured in the `pymbar` Python package (`https://github.com/choderalab/pymbar`). Because MBAR assumes independent samples for calculating uncertainties, we analyzed the timeseries of reduced potentials across all simulations and found that the distribution of statistical inefficiencies ($g \equiv 1 + 2\tau$, where τ is the correlation time) from these suggested a sampling frequency of 20 ps for each trajectory, as well as discarding the first 5 ns of data as equilibration.

6.2.5 How many simulations of each leg, and for how long?

It was not clear *a priori* how many windows to perform for each leg of the alchemical pathway, and how much simulation sampling would be required for each. An initial run for the IFS2 state was performed in which each leg amounting to a switching of a particular type of parameter (adding the restraint, adding the repulsion, switching off the Coulomb interaction, switching off the Van der Waals interaction) was subdivided into 21 windows, corresponding to a $\Delta\lambda$ between windows of 0.05. This initial run revealed that obtaining reasonably converged values (within 0.5 kJ/mol) for adding the restraint could be obtained with only 10 ns of simulation time for each

window, using only 11 windows ($\Delta\lambda = 0.1$). The initial run also revealed that adding the repulsion came with a very low free energy cost ($\sim 2 \times 10^{-4}$ kJ/mol), so only 3 windows ($\Delta\lambda = 0.5$) of 10 ns each were performed for each state as a check that this remained the case.

For the Coulomb and Van der Waals switching legs, it was found that convergence of ΔG_l to within 1 kJ/mol still was not reached after 50 ns of simulation time using 21 windows per leg. See Section 6.3 for convergence analysis of production runs.

6.2.6 Workflow automation

The production runs of the four complex states (IFS2, IFS4, OFS2, OFS4) and the ion-in-solvent amounted to 486 independent simulations. Each of these simulations, in particular those for the complex, required significant compute resources, totalling nearly 2 million CPU hours. Performing this amount of computation in a reasonable time using multiple compute resources, spread across so many simulations, would not have been possible without workflow automation. For this we used Fireworks[173], and in particular the workflows for performing MD simulations provided by `mdworks` (see Section 7.4).

6.2.7 Molecular dynamics parameters

All-atom, explicit solvent MD simulations were performed with Gromacs 5.1.4[52] and the CHARMM36 force field for the protein with the CMAP correction[42–44], ions and lipids[45] together with the CHARMM TIP3P water model[42].

All simulations were carried out under periodic boundary conditions. The Verlet neighbor list with a cutoff of 1.2 nm was updated every 20 steps. Coulomb interactions were calculated with the smoothed particle mesh Ewald (SPME) method[105] with a real-space cutoff of 1.2 nm, and interactions beyond the cutoff were calculated in

reciprocal space with a fast Fourier transform on a grid with 0.12-nm spacing and fourth-order spline interpolation. The Lennard-Jones force was switched to zero at a cutoff of 1.2 nm, without applying any dispersion corrections[164]; these parameters were previously found to give reasonable behavior for lipids[34] and match closely with the behavior expected under CHARMM. Bonds to hydrogen atoms were constrained with the P-LINCS algorithm[55] (with an expansion order of four and two LINCS iterations) or SETTLE[56] (for water molecules). For alchemical switching of the Lennard-Jones interactions, a soft-core potential was used with $\alpha = 0.5$, $\sigma = 0.3$, and a λ power of 1.

The NapA complex

The NapA dimer was simulated in an ∼4:1 1-palmitoyl-2-oleoylphosphatidylethanolamine (POPE)/1-palmitoyl-2-oleoylphosphatidylglycerol (POPG) bilayer. Simulations of the outward-facing structure were based on the LCP crystal structure (PDB 5BZ3), while the inward-facing structure (PDB 5BZ2) was backmutated from the cysteine-linked structure by changing C31 back to V31 and C130 to I130[34]. Both starting structures were obtained from frames binding Na^+ after >500 ns of simulation time.

The total system consisted of about 130,000 atoms in an orthorhombic simulation cell with a free NaCl concentration of ∼250 mM. Most titratable residues were simulated in their default charge states, as predicted by PROPKA 3.1[131] for pH 7.8, including the charged form of residue D156 at the ion-binding site. The charge state of K305 is dependent on the protonation state being sampled (S2: charged; S4: neutral). Since we only want to sample the binding free energy of a single ion to a single protomer, one of the NapA protomers was given a protonated D157 to avoid sampling a bound state. In the other protomer, the binding-site residue D157 was simulated in its charged form in the inward-facing and outward-facing conformation

to enable binding of Na^+ [31,32,34]. This protomer is the one used for interacting with the special Na^+ ion that is altered along the alchemical pathway.

Each MD simulation was performed after energy minimization and ∼100 ps of pre-production equilibration with a 0.1 fs time step. Production MD was performed with a 2 fs time step. A Langevin integrator was used for temperature control at $T = 310$ K, with the friction coefficient for each particle computed as mass/0.1 ps [57] for each of the protein, lipids, and solvent taken as separate temperature-groups. A Parrinello-Rahman barostat [63] with time constant 5 ps and compressibility 4.6 × 10^{-5} bar^{-1} was used for semi-isotropic pressure coupling.

The ion in solution

The ion in solution system consisted of about 4,900 atoms in a rhombic dodecahedral simulation cell with 5 pairs of Na^+ and Cl^- ions, including the special Na^+ that is altered along the alchemical pathway, and 1,630 water molecules, giving a free NaCl concentration of ∼170 mM.

Each MD simulation was performed after energy minimization and ∼1 ns of pre-production equilibration with a 0.1 fs time step. Production MD was performed with a 2 fs time step. All simulations were run for a total of 50 ns. A Langevin integrator was used for temperature control at $T = 310$ K, with the friction coefficient for each particle computed as mass/0.1 ps [57]. A Parrinello-Rahman barostat [63] with time constant 5 ps and compressibility 4.6 × 10^{-5} bar^{-1} was used for isotropic pressure coupling.

6.3 Results

Calculating the free energies of binding, ΔG_b of Na^+, for each of the four states IFS2, IFS4, OFS2, and OFS4 using an alchemical pathway requires the calculation

of individual free energies ΔG_l for each leg in that pathway. The particular pathway we have chosen includes simulations of the whole NapA complex with a bound ion for some legs, while for other legs we only simulate Na^+ in solvent.

In addition to presenting the resulting absolute binding free energies, ΔG_b, for each state (Section 6.3.3), we also detail here the free energies of each leg, ΔG_l. Of particular importance are the empirical checks of each ΔG_l for convergence.

6.3.1 Hydration free energy of sodium

The ΔG_l for each of the ion-in-solvent legs of the pathway (Figure 6.1) are given in Table 6.2. The simulations for all of these legs extended to 50 ns, with each leg subdivided into 21 windows ($\Delta lambda = 0.05$).

leg	ΔG_l (kJ/mol)
coulomb	382.7 ± 0.2
vdw	-8.81 ± 0.05
repulsion	0.258 ± 0.005
restraint*	17.84
sum	392.0 ± 0.2

Table 6.2: Converged ΔG_l values for each of the ion-in-solvent legs of the alchemical pathway, in the direction of solvation to vacuum. Uncertainties are the standard deviations of each ΔG_l as reported by MBAR[171], and do not necessarily imply convergence (see Figure 6.2).
*The value for the restraint leg has no uncertainty because it was calculated analytically (see Section 6.5.1 and Equation 6.11).

Each of these values were checked for convergence graphically by using increasing amounts of simulation data. These convergence checks are shown in Figure 6.2. We see that for each of these legs, the estimates given in Table 6.2 are well-converged to within 1 kJ/mol or less.

It should also be noted that the Coulomb and Van der Waals decoupling legs of the pathway examined here together amount to a calculation of the solvation free energy of Na^+. The value we have obtained ($\Delta G_{\text{solv}} = -(\Delta G_{\text{ion,coulomb}} + \Delta G_{\text{ion,vdw}})$), $\Delta G_{\text{solv}} =$

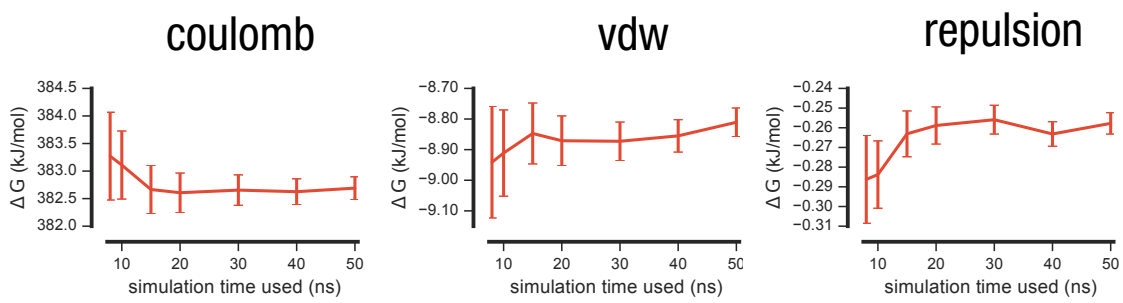


Figure 6.2: Convergence of ion-in-solvent legs. For each of the coulomb, vdw, and repulsion legs of the alchemical pathway, we plot the calculated ΔG_l using increasing amounts of data from each simulation window. The first 5 ns of all simulations was discarded, in all cases, so e.g. the point at 30 ns uses 25 ns of simulation data from each simulation in the leg. All simulations were sampled at 20 ps intervals. For each of these legs, we observe convergence in 50 ns of simulation time to within 1 kJ/mol.

-373.9 ± 0.2 kJ/mol, compares favorably with those observed experimentally[174], including -365[175], -371[176], and -372[177] kJ/mol.

6.3.2 Binding to the complex

The ΔG_l for each of the complex legs of the pathway(Figure 6.1) for each conformation (IF, OF) and each charge state (S2, S4) are given in Table 6.3. We observe that across all states, the cost of adding the repulsion to the bound ion is practically unmeasurable. The next smallest contribution to ΔG_{comp} comes from removing the restraint, which ranges from -2 to -13 kJ/mol. Adding back the Van der Waals interaction to the ion comes at a cost of $\sim$20 kJ/mol for each state, while the largest contribution to ΔG_{comp} by far comes from switching on the Coulomb interaction, which is a gain of $\sim$450 to $\sim$500 kJ/mol, depending on the state.

Convergence plots for each state are given in Figure 6.3. The repulsion legs consistently show a value of $\sim$ 0.0 kJ/mol across all sttes, and the convergence plot serves as a check that this remains true. For the restraint legs, we see variation in $<.5$ kJ/mol, suggesting that with only 10 ns of time for each simulation (amounting to 5 ns per window; see Section 6.2) for each window we can obtain converged results.

conformation	charge state	leg	ΔG_l (kJ/mol)
IF	S2	vdw	20.83 ± 0.04
		coulomb	-453.4 ± 0.1
		repulsion	0.014 ± 0.003
		restraint	-2.35 ± 0.03
	S4	vdw	17.82 ± 0.04
		coulomb	-518.9 ± 0.1
		repulsion	-0.0004 ± 0.0001
		restraint	-11.79 ± 0.05
OF	S2	vdw	21.93 ± 0.04
		coulomb	-445.4 ± 0.1
		repulsion	-0.00003 ± 0.00003
		restraint	-3.09 ± 0.06
	S4	vdw	23.59 ± 0.04
		coulomb	-446.6 ± 0.1
		repulsion	0.0002 ± 0.0002
		restraint	-13.0 ± 0.1

Table 6.3: ΔG_l values for each of the complex legs of the alchemical pathway, in the direction toward the bound state. Uncertainties are the standard deviations of each ΔG_l as reported by MBAR [171], and do not necessarily imply convergence (see Figure 6.3).

The picture is more complicated for the Van der Waals and Coulomb legs. We see variation of $\sim$1.5 kJ/mol for the Van der Waals legs, with what looks like convergence to within $\sim$.5 kJ/mol using 65 ns of data for each simulation (with two independent repeats for each window, giving 120 ns of total data for $\Delta G_{\text{comp,vdw}}$ calculation). For the Coulomb legs, it is not clear that our reported values are converged to within 1 kJ/mol, even with 65 ns for each simulation. All simulations are being extended to 80ns to test convergence but even these preliminary results demonstrate what appear to be nearly converged absolute binding free calculations. It should be noted that binding free energies accurate to 1 kcal/mol (4.2 kJ/mol) are considered "chemically accurate" [178].

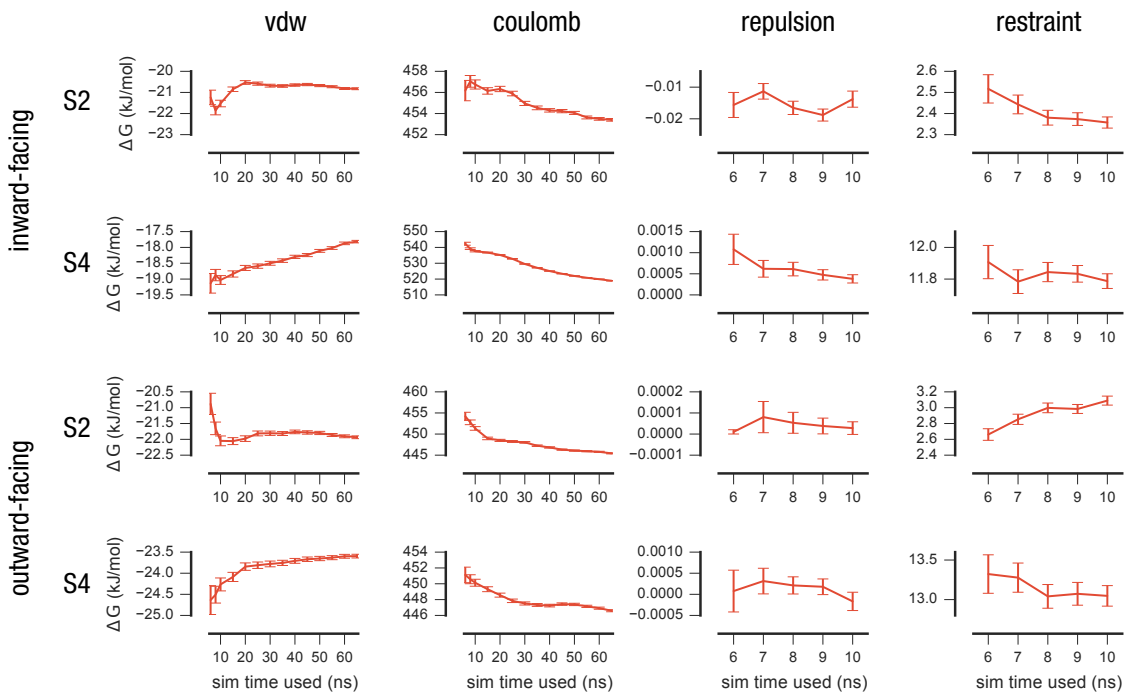


Figure 6.3: Convergence of complex legs. For each of the vdw, coulomb, repulsion and restraint legs of the alchemical pathway, we plot the calculated ΔG_l using increasing amounts of data from each simulation window. The first 5 ns of all simulations was discarded, in all cases, so e.g. the point at 30 ns uses 25 ns of simulation data from each simulation in the leg. All simulations were sampled at 20 ps intervals. For each of these legs, except for the coulomb legs, we observe convergence to within 1 kJ/mol with the simulation time performed.

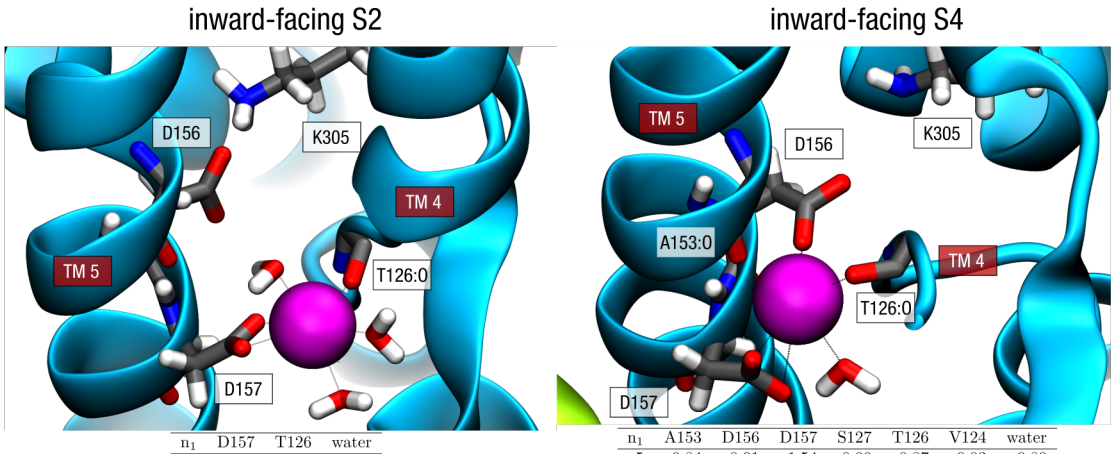
6.3.3 Absolute binding free energies

From the ΔG_l for each leg, we can calculate the absolute binding free energy of Na^+ for each of the conformations and charge states we examined (Equation 6.2). These are listed in Table 6.4.

conformation	charge state	ΔG_{bind} (kJ/mol)
IF	S2	-43.6 ± 0.3
	S4	-122.8 ± 0.3
OF	S2	-35.1 ± 0.3
	S4	-44.4 ± 0.3

Table 6.4: Free energy of binding of Na^+ to NapA from alchemical free energy calculations. Uncertainties are propagated from the uncertainties given for each ΔG_l as reported by MBAR[171]. These values are not yet corrected for artifacts due to charged particles under periodic boundary conditions[179].

We see that between the two conformations, inward-facing (IF) and outward-facing (OF), there is a clear preference for Na^+ binding to the inward-facing state. This is consistent with the physiological role for this transporter, as it is advantageous for transport of Na^+ out of the cell if the affinity for Na^+ is weaker in the OF state[180].



n_1	D157	T126	water
5	1.87	0.95	2.18
6	1.97	0.98	3.05
7	1.99	0.98	4.02

n_1	A153	D156	D157	S127	T126	V124	water
5	0.84	0.91	1.54	0.00	0.87	0.03	0.80
6	0.49	0.88	1.41	0.03	0.77	0.19	2.23
7	0.35	0.94	1.30	0.10	0.76	0.24	3.33

Figure 6.4: Binding site coordination of Na^+ (magenta) in equilibrium MD without restraints. The ion is more tightly coordinated by protein residues in the S4 state than in the S2 state. Tables give the time-averaged contribution number of coordinating oxygen atoms from each residue, sampled over equilibrium MD.

We also see that the ion is bound much more strongly in the S4 charge state than in S2, at least for the IF conformation. This is to be expected, since removal of the charge on K305 frees D156 to interact with the ion, coordinating it more tightly in the binding site (see Figure 6.4). What remains unclear is why, physically, the binding free energy for the OFS4 state is comparable to both IFS2 and OFS2. This will need to be explored in greater detail.

6.4 Discussion

The results of this study so far show that a clear difference exists in the binding affinity of Na^+ between the inward- and outward-facing states of NapA, as well as the primary importance of the charge state of K305 for this value. Convergence of

the largest component of the binding free energies, the $\Delta G_{\text{comp,coulomb}}$ contributions from the Coulomb-switching legs, to within 1 kJ/mol are not yet guaranteed, but the differences between the states studied are robust against further variations in these values.

Before further conclusions can be drawn, however, there are a number of additional considerations. First, to calculate apparent binding affinities (apparent K_{a}), we must take into account the S1 charge state, in which D157 is protonated (see Figure 4.10. Obtaining the binding free energy for this state in the same way as for S2 and S4 is not possible, as Na^+ does not bind on its own. It would thus be difficult, and perhaps impossible, to obtain converged values for $\Delta G_{\text{comp,restraint}}$ in the way we have done here, as the window with $\lambda = 0$ (no restraint), the ion will leave the binding pocket. Another approach, such as alchemically changing the protonation state of D157 after applying the restraint to the ion, is possible, but features other considerations such as adding and removing whole charges (see below).

Second, once we possess converged ΔG_b values for each of the relevant charge states, combining these to estimate an apparent K_{a} requires knowing the relative probability of each state. In principle these can be obtained to some accuracy using data from CpHMD simulations[33,181], but these simulations are expensive to perform. We are exploring how accurately we can expect to obtain relative probabilities using less expensive methods, such as PROPKA[112].

Third, because we are annihilating a whole non-zero charge in our chosen pathway (it returns later, but in a different simulation system), we must take into account finite volume effects due to the way long-range electrostatics interactions are performed using particle-mesh Ewald summation[179]. A correction is needed that at the very least requires Poisson-Boltzmann calculations for the protein complex. Our

reported absolute binding free energies in Table 6.4 do not currently include any such correction, and it is not clear how much of an impact this will have on each.

Finally, it should be noted that presence of the repulsion with respect to other Na^+ ions added to the decoupled ion removes the strict equivalence between points (e) and (f) in Figure 6.1, since the ion is still interacting (weakly) with at least some of the surrounding system. We are investigating to what degree this interaction affects the calculated ΔG_b, as well as if a correction can be applied to account for it. Because the free energy cost of the repulsion in all cases is close to 0.0 kJ/mol, however, the mean interaction with the other Na^+ is negligible; thus, a correction for this interaction is probably also negligible.

Having examined the binding free energy for Na^+ of NapA as a function of conformation and charge state, we intend to perform the same analysis to quantify binding strength for other proteins for which we have structures, in particular EcNhaA [22,31] and PaNhaP. With an view toward comparing to experimental binding affinities, we also plan to do obtain binding free energy of Li^+ for these systems, as experimental results are considerably more precise for this ion.

6.5 Supplemental Information

6.5.1 Analytic cost of removing restraint on ion in solution

We need to calculate the cost of removing the restraint, but it is not possible to get this from simulation results. We can, however, calculate this analytically. We need to obtain [167]:

$$\Delta G_r = -kT \ln \frac{Z_P Z_L}{Z_{CL}} \tag{6.8}$$

where Z_P is the partition function of the protein, which in general we cannot calculate, Z_L is the partition function of the free ligand (noninteracting with the protein or the rest of the system), and Z_{CL} is the partition function of the protein and restrained ligand taken together. For ΔG_r to be easily calculable, then Z_{CL} must be separable as:

$$Z_{CL} = Z_P \times Z_r \times \tilde{Z}_L \tag{6.9}$$

where Z_r is the partition function of the restraint itself, and $\tilde{Z}_L$ is the partition function of the internal degrees of freedom of the ligand. For a point particle, this is $\tilde{Z}_L = 1$. The reason this is $\tilde{Z}_L$ and not Z_L is because external degrees of freedom for the ligand are accounted for in the protein and the restraint. Also, note then:

$$Z_L = V\tilde{Z}_L \tag{6.10}$$

since the partition function for the external degrees of freedom of a noninteracting point particle is simply the volume of the space it can explore.

So for ΔG we get:

$$\Delta G_r = -kT \ln \frac{Z_P Z_L}{Z_P Z_r \tilde{Z}_L} = -kT \ln \frac{V}{Z_r} \tag{6.11}$$

and so we only need Z_r, which we can at least express analytically:

$$\begin{aligned} Z_r &= \int_0^{2\pi} \int_0^{\pi} \int_0^{\infty} dr\, r\, d\theta\, r\, \sin(\theta)\, d\phi\, e^{-\beta(U_\theta(\vec{r}) + U_r(\vec{r}))} \\ &= 2\pi \int_0^{\pi} e^{-\beta U_\theta(\theta)}\, \sin(\theta)\, d\theta \int_0^{\infty} r^2\, e^{-\beta U_r(r)}\, dr \end{aligned} \tag{6.12}$$

And where we have for the restraints:

$$U_\theta(\theta) = \frac{1}{2}k_\theta(\theta - \theta_0)^2 \tag{6.13}$$

$$U_r(r) = \frac{1}{2}k_r(r - r_0)^2 \tag{6.14}$$

Using our chosen restraint parameters, k_r = 16,000.0 kJ/(mol nm^2), r_0 = 0.275 nm, k_θ = 23.0 kJ/(mol rad^2), θ_0 = 0.0 rad, as well as our system temperature T = 310 K and the standard volume 1.6605 nm^3, we can directly calculate the free energy contribution of the restraint as -17.84 kJ/mol. We use the standard volume here to account for scaling our system to the standard state, so our calculated ΔG_b is a standard state free energy[167].

Chapter 7

SOFTWARE

It would have been very difficult, and arguably impossible, to apply the methods we have used to probe scientific questions without good software tools. In many ways, software is the great enabler for our science, often making what was once complex, routine, and allowing us to quickly test multiple hypotheses. Over the course of my doctoral work, I have spent significant time and effort crafting some of the tools that are now routinely used in our lab to do interesting science. This work has propelled not only our work, but continues to ripple outward from our lab into the work of others in the field. In this way software has a multiplicative effect when done well.

Since software engineering has been such a key part of my doctoral research, it would be a mistake to not include mention of some of the projects that enabled it. Some of these are original developments of my own that have blossomed into collaborative projects used by many others around the world, while others are projects that already existed but to which I have made both contributions to and extensive use of.

I would also be remiss if I failed to mention the significant role *open source* software in particular has played in my work. Open source software is publicly developed and freely distributed, making it particularly flexible to the needs of the scientific community, easily deployable, and transparent in its functionality[182]. During the course of my work I was fortunate to be in a lab environment that is friendly to this model of software development, and this is reflected in our choice of tools.

The Python (`https://www.python.org/`) programming language is at the core of our research apparatus. The software stack with which we do our science is largely

written in this language (Figure 7.1), allowing us great flexibility as new tools can be made to work on top of old ones to address problems of increasing complexity and scale. In this way, as computing power has increased to allow ever larger amounts of data to be generated, so too have our tools scaled upward to take advantage of it.

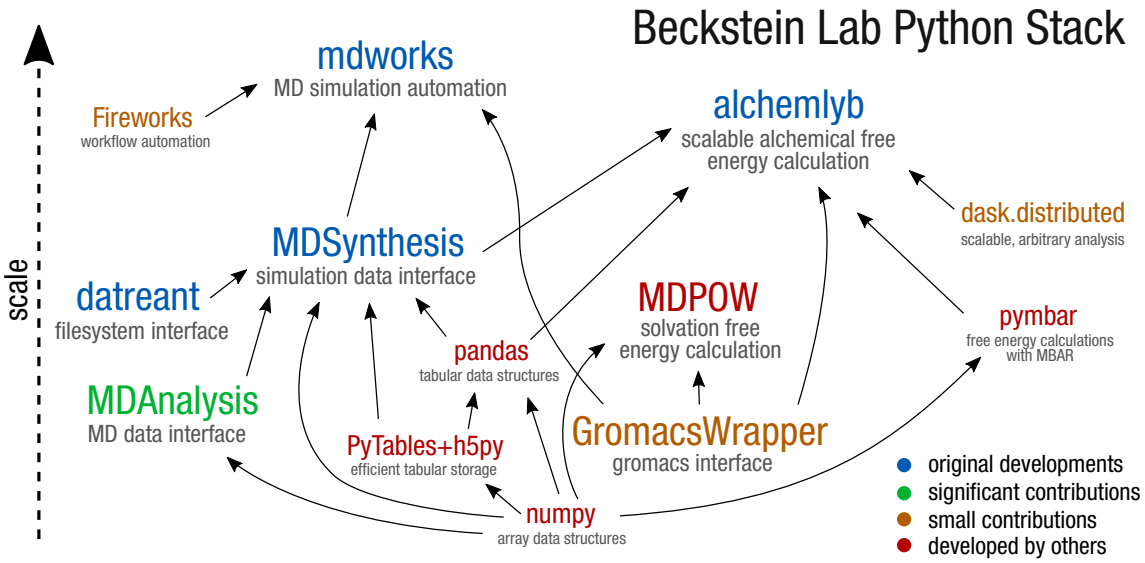


Figure 7.1: Dependency diagram of the scientific Python software stack that our lab operates on. As computing power has increased, new tools are built on top of existing ones to scale upward with larger data requirements. My contributions to each package are indicated qualitatively by color.

My contributions to this stack are shown visually in Figure 7.1. As my needs for managing and analyzing larger and more heterogeneous datasets grew, I designed and built new tools on top of the existing stack. This stack continues to be refined and it evolves as we ask new research questions of ever increasing scale.

Some of the projects detailed here (`datreant`, MDSynthesis, `mdworks`) were created and largely developed by me, often out of some need that was unaddressed by anything else at the time. `datreant` and MDSynthesis have been in development for several years and have an active community of developers working on them. `mdworks` is by contrast rather new and unpolished, but is evolving into a standalone package that is usable by others. Others, such as MDAnalysis, already existed and have an

active developer community, of which I have been a part. All are publicly developed and open source.

For each project, I give a brief description here of the scope and purpose. For some of these I also give a simplified example showing how they are used in practice, giving a flavor for what they enable us to do. In all cases, I strive to design software with clean interfaces and clear abstractions, so that all of these pieces can be used in new ways, often together.

7.1 datreant

Probing scientific questions with simulations is often an organic process, punctuated by structured studies that ask specific scientific questions. In all cases there are choices made about system composition and size, starting configuration, and simulation parameters, with many of these often varied to test hypotheses or perform specific calculations (e.g. alchemical free energies). A consequence of this is that simulation data isn't easily shoehorned into common database solutions that would normally be used to make flexible and scalable analysis more feasible. Instead, raw simulation trajectories and derived datasets are stored in a variety of fileformats in the filesystem, often scattered in ways that are historical and idiosyncratic. This scheme is more flexible for storage than most database solutions, but comes at the cost of making analysis more tedious and time-consuming. Furthermore, the existing ecosystem of tools often assumes the data are in fileformats common to the field, of which there are many, so database solutions would not be usable for all types of analyses.

This organic and relatively unstructured nature of simulation data, as shown for example in Figure 7.2, can be a barrier when trying to make sense of it *post hoc*. Analysis code ends up being written with fragile, hard-coded paths to trajectory

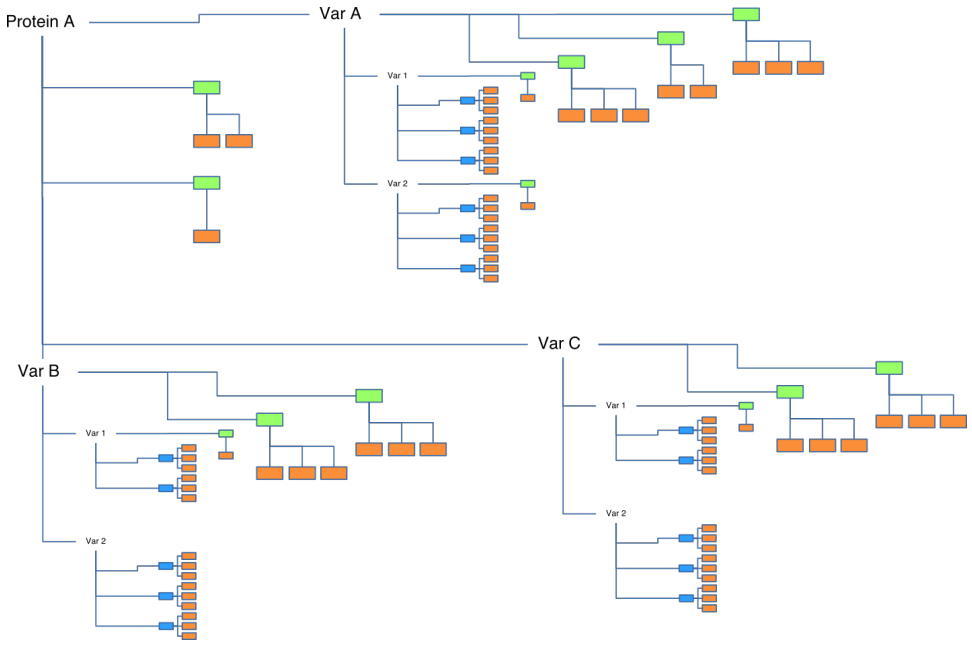


Figure 7.2: Scientific research proceeds organically. The filesystem tree in which data is organized reflects this, with studies on different proteins, different parameter choices, etc. organized by naming convention within the branches of the tree.

data, and researchers spend significant time moving around the filesystem looking for what they need. This slows down the act of asking questions of the data, and means iteration on questions that often don't bear fruit is slow and painful.

`datreant`[183] is a Python package that makes it easy to programmatically work with the contents of the filesystem using idiomatic Python style. In particular it features **Treants**: specially marked directories with distinguishing characteristics that can be discovered, queried, and filtered. Treants can be manipulated individually and in aggregate, with mechanisms for granular access to the directories and files in their trees. By way of Treants, `datreant` adds a lightweight abstraction layer to the filesystem, allowing researchers to focus more on *what* is stored and less on *where*. This greatly reduces the tedium of storing, retrieving, and operating on datasets of interest, no matter how they are organized.

Detailed information on `datreant`, as well as the hub for its developer and user community, can be found at `http://datreant.org`, as well as in Chapter 8. A detailed example showing how `datreant` is used in practice is given later in this chapter (Section 7.3).

7.2 MDAnalysis

MDAnalysis[106,184] is a Python library that serves as an interface to the myriad fileformats in common use in the molecular dynamics community. It provides an important abstraction layer, allowing users to write code to analyze molecular dynamics trajectories without concern for the details of the underlying file format. This saves an incredible amount of time, and also makes the entire scientific Python stack directly usable for working on MD data.

I am a core developer of MDAnalysis, primarily contributing over the past two years to efforts at making the external behavior of MDAnalysis consistent (i.e., designing a cleaner API). This culminated in my work, alongside fellow core developer Richard Gowers, to refactor the library's topology system. We spent several weeks prototyping and iterating on a new design that featured many performance improvements, both in terms of speed and memory usage, while also improving the internal consistency of topologies and cleaning up the user-facing API. This was a heavy effort that is almost complete, and serves as my greatest contribution to the project to date.

More information on MDAnalysis can be found at `http://mdanalysis.org`, as well as in (Michaud-Agrawal et al.[106]) and (Richard J. Gowers et al.[184]).

7.3 MDSynthesis

`datreant` is a general-purpose library, with appeal to researchers in all fields. **MDSynthesis**[183] (`mdsynthesis.readthedocs.org`) expands the repertoire of a Tre-

ant to include a convenience interface to molecular dynamics data, internally using MDAnalysis[106,184] for this purpose. In the same way that `datreant` provides a Pythonic abstraction layer to the filesystem, MDSynthesis offers an abstraction layer to simulation data, individually and in aggregate. This abstraction is extremely valuable, as it makes it possible to access the data of whole collections of simulations at any level of detail necessary in programmatic and highly parallelizable ways. In addition to being useful for interactive work, this abstraction also makes complex automation (e.g. `mdworks`, Section 7.4) and sophisticated analysis far easier to build and execute.

7.3.1 Introspecting individual simulations with Sims

The core data structure in MDSynthesis is the **Sim**. This is a Treant that represents the data for a single simulation, which by convention is often all stored in its own directory in the filesystem. Simulation data is more than just the raw molecular dynamics trajectory; it is also all the files that encode the parameter choices, the starting configuration, any pre-production steps that went into creating it, log files giving runtime details, and more, as shown for example in Figure 7.3.

A Sim offers a Pythonic interface for easily working with any of this data. A directory containing simulation data (e.g. `'NapA_0'`) can be directly manipulated as a Sim object. In Python, we can do:

```
>>> import mdsynthesis as mds
>>> napa = mds.Sim('NapA_0/')
```

And this will create a `Sim` object for the directory `'NapA_0'` that exists in the filesystem. This object can be used to introspect the directory directly:

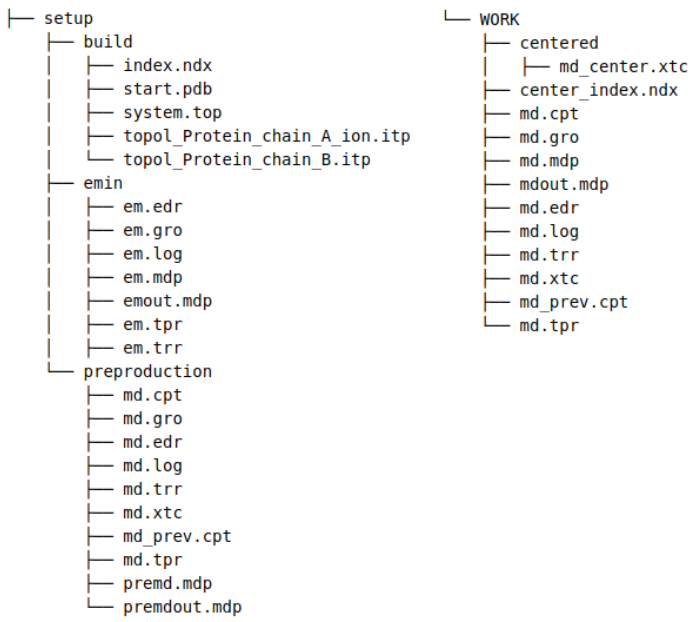


Figure 7.3: An MD simulation is not just the trajectory file(s); it is also the files that encode its parameter choices, preproduction steps, and more.

```
>>> napa.draw(depth=1)
NapA_0/
 +-- WORK/
 +-- setup/
 +-- dists/
 +-- Treant.f91a79df-c223-4789-b222-5744cf7gjfdk.json
 +-- angles/
```

A Sim is a Treant, and what makes a Treant distinct from a normal directory is its **state file**. This file, here `Treant.f91a79df-c223-4789-b222-5744cf7gjfdk.json`, serves to mark the directory as a Treant, making it discoverable. It also includes a unique identifier (uuid) for that Treant, and internally it stores metadata elements to distinguish it, such as tags and categories. These can be used to filter and group Treants and Sims when treated in aggregate (see Section 7.3.2).

We can use the `Sim` object to select out directories in its filesystem tree by name as `Tree` objects, and furthermore introspect those:

```
>>> napa['setup'].draw(depth=1)
setup/
 +-- preproduction/
 +-- build/
 +-- emin/
```

Any file can also be examined. For example, we can work with a Gromacs GRO structure file as a `Leaf` object. We can then use this to directly read its contents:

```
>>> leaf = napa['setup/emin/em.gro']
>>> leaf
<Leaf: 'NapA_0/setup/emin/em.gro'>

>>> print(leaf.read(size=160))
NapA inward-facing
132506
    1GLY      N    1   5.627   5.740   3.317
    1GLY     H1    2   5.613   5.806   3.237
    1GLY     H2    3   5.567   5.769   3.398
```

In addition to working with its filesystem contents individually, a Sim can be used to do work on files as collections just as easily. We can, for example, extract files whose paths match a *globbing* pattern:

```
# grab up existing files in the tree matching globbing pattern
>>> napa['setup'].glob('*/*.itp').relpaths
['NapA_0/setup/build/topol_Protein_chain_A_ion.itp',
 'NapA_0/setup/build/topol_Protein_chain_B.itp']
```

Or extract `Leaf` objects by name from a number of directories at the same time:

```
# get Leaf objects "md.gro" for each tree inside "setup"
>>> gros = tree.trees.loc['md.gro']
>>> gros.relpaths
['NapA_0/setup/build/md.gro',
 'NapA_0/setup/emin/md.gro',
 'NapA_0/setup/preproduction/md.gro']
```

The ability to work at the level of individual objects such as directories and files and to just as easily abstract to the level of whole collections is a common theme in `datreant`, and by extension MDSynthesis. This is a design pattern that

makes analyses of all types easier to perform with less time, and often makes parallel computation with tools such as `dask.distributed` trivial.

7.3.2 Working with Sims in aggregate for ensemble analytics

Just as one can work on files and directories in an individual Sim as aggregates, one can work on many, perhaps hundreds of Sims, as aggregates as well. Sims spread throughout a filesystem tree can be gathered up, and Sim metadata such as tags and categories can be used to filter and split them into aggregate units of interest. If the directory `'sims'` contains many Sims, we can gather them all up with:

```
>>> b = dtr.discover('sims/')
>>> len(b)
49
```

This gives a `Bundle`, an ordered set of the `Sim` objects found within `'sims'`. It provides convenient mechanisms for working with these Sims as a single logical unit. For example, suppose we want to examine the angle between three atoms with time, previously collected and stored in CSV files within each Sim's directory structure:

```
>>> b[:3].loc['angles'].draw(depth=1)
angles/
 +-- sna_cg_cb/
 +-- sna_cg_cb.csv
angles/
 +-- sna_cg_cb/
 +-- sna_cg_cb.csv
angles/
 +-- sna_cg_cb/
 +-- sna_cg_cb.csv
```

These simulations were performed with varying parameters, with these parameters affecting what is observed for this angle. We can group the Sims by the value of one of these parameters, k_θ (`cth`), which we stored in each as a *category*. Among the simulations we tested 7 different values:

```
>>> set(b.categories['cth'])
{9.84841905,
 13.1312254,
 16.41403175,
 19.6968381,
 22.97964445,
 26.2624508,
 27.84115237}
```

We can easily histogram these angles as a function of this parameter, aggregating the data across simulations with the same parameter value (Figure 7.4). `Bundle.categories.groupby` returns key-value pairs giving the parameter values as keys and a `Bundle` of Sims with that value as values:

```
import pandas as pd
import matplotlib.pyplot as plt

fig = plt.figure(figsize=(7, 4))
ax = fig.add_subplot(1,1,1)

groups = b.categories.groupby('cth')

for cth in sorted(groups.keys()):
    s = pd.concat(
        [pd.read_csv(csv.relpath, header=None, index_col=0)
         for csv in groups[cth].loc['angles/sna_cg_cb.csv']])[1]

    s.plot.hist(bins=150, histtype='step', ax=ax, normed=True,
                label="cth: {}".format(cth))

ax.legend(loc='upper left')
```

With mechanisms like `Bundle.categories.groupby`, we can operate on simulation data at a high-level, treating whole sets of simulations as meta-datasets from which we can quickly extract new insights. By making it relatively easy to work with what can often be many terabytes of simulation data spread over tens or hundreds of trajectories, MDSynthesis greatly reduces the time it takes to iterate on new ideas toward answering real biological questions.

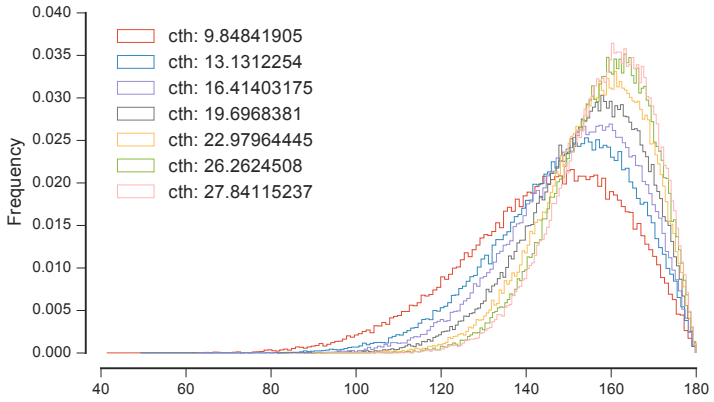


Figure 7.4: Distribution of atom angles aggregated across groups of simulations according to k_θ (`cth`).

7.3.3 Raw trajectory data access via MDAnalysis

The features of MDSynthesis detailed so far are actually features of `datreant`, which MDSynthesis is built around. In addition to these, a `Sim` offers additional conveniences for working with simulation data directly using MDAnalysis as the interface of choice. In particular, a `Sim` can store a `Universe` definition:

```
>>> import MDAnalysis as mda
>>> GRO = 'NapA_0/WORK/md.gro'
>>> XTC = 'NapA_0/WORK/md.xtc'

>>> s = mds.Sim('NapA_0')
>>> s.universe = mda.Universe(GRO, XTC)
>>> s.universe
<Universe with 132506 atoms>
```

And once defined, this can be used by the `Sim` in any other Python process thereafter through the `Sim.universe` property. This is useful because for different simulations the files that compose the trajectory may be organized differently or vary widely in number. Code that must extract data from many different Sims directly from the trajectories can be written without having to specify *which* files to use.

In a similar vein, a `Sim` can also store `AtomGroup` definitions. These are named grouping of atoms from the `Universe`, and once defined they can be recalled with only the name:

```
>>> s = mds.Sim('NapA_0')
>>> s.atomselections.add('water', 'same residue as name OW')
>>> s.atomselections.create('water')
<AtomGroup with 76704 atoms>
```

This introduces an abstraction layer between the definition of the selection, which may vary depending on the system, and what the selection *is*, in this case the water molecules. It makes it possible to write code that does the same thing across many simulation systems, without having to deal directly with all the variations among the simulations. This is an important requirement for scaling out analysis quickly.

In this way and others, MDSynthesis provides an extremely powerful abstraction layer for the large variety and volume of data that we can routinely collect. This tool is often used directly and interactively in Jupyter notebooks (`http://jupyter.org/`) to rapidly explore and make sense of this data, but it can also be used by automated tools to assist in handling data while it is being generated. We use `mdworks` to do exactly this (Section 7.4).

7.4 mdworks

MDSynthesis and `datreant` serve to make the manipulation of the filesystem easier and less error-prone with useful abstractions. This makes analysis easier, but the abstraction of a `Sim` also gives a convenient unit on which automated workflows can be designed to operate. Such systems can even be used to *execute* the simulations themselves, using the `Sim` object to manage and process the data as one might do manually.

Workflow automation systems are becoming increasingly common in scientific computing. Whereas some fields, such as bioinformatics[185], have long used such systems to perform processing on data, the simulation community has been slow to adopt these approaches for doing simulations. The reasons for this are not explored here, but to expand the number of simulations we are able to perform with very high throughput, in particular for doing free energy calculations, we chose to pursue using Fireworks[173].

7.4.1 A molecular dynamics workflow

Fireworks is a general-purpose framework written in Python for building directed acyclic graphs (DAGs) that describe a set of tasks that must be performed, and their dependencies with one another. A graph of this type is a *workflow*, and each node in the graph can be executed by a worker process running on local or remote machines, as needed.

To leverage this system for executing molecular dynamics, I built `mdworks` for constructing self-propagating simulation workflows that can be customized for different simulation needs. The core function of this library produces such a workflow from a single `Sim` object:

```
import mdsynthesis as mds
from mdworks.general import make_md_workflow

sim = mds.Sim('NapA_0')
workflow = make_md_workflow(sim=sim,
                            archive=sim['WORK'].abspath,
                            stages='/home/david/.fireworks/stages.yaml',
                            md_category='md',
                            files=['md.tpr', 'md.cpt'])
```

The resulting workflow object fully describes the dependency graph (Figure 7.5) of tasks that must be performed to execute the simulation, either locally or on a remote

high-performance computing (HPC) resource. All tasks except for the `md` task are performed on the archival machine (usually a fileserver or storage server), from which the necessary files are staged to all compute resources. The remote resources execute the `md` task, and once complete the archival machine will pull the data from the resource it executed on, clean up the files on that resource, and then determine whether or not the entire run is complete based on the number of simulation steps desired. If the simulation should continue, an entire new set of tasks with the same dependency graph is attached, and the process continues.

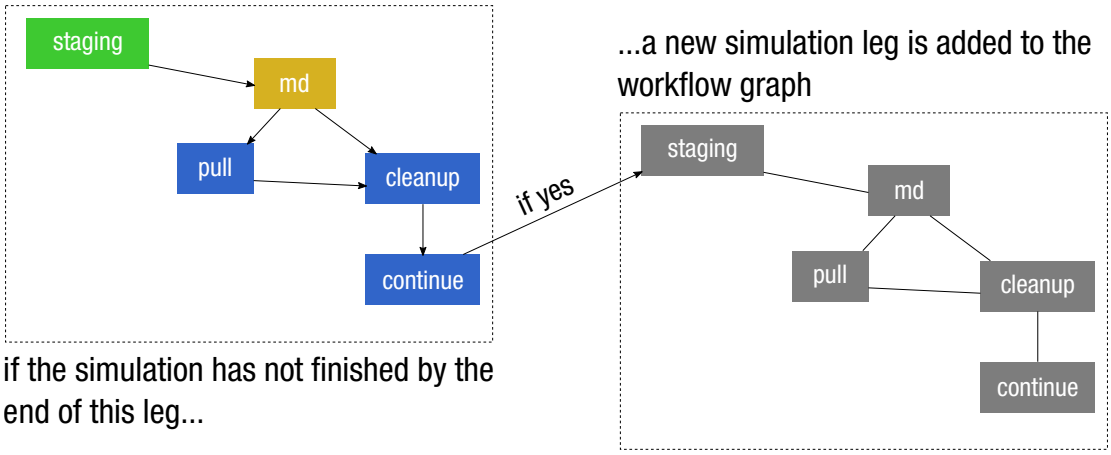


Figure 7.5: The core workflow task graph for molecular dynamics simulation execution under `mdworks`. This workflow self-propagates if the simulation has yet to complete.

In addition to generating these basic workflow graphs, `mdworks` has mechanisms for grafting additional tasks to certain nodes. The `pull` task, for example, can be followed by a post-run postprocessing workflow of arbitrary complexity, allowing each segment of the simulation to immediately be prepared for analysis. An arbitrary post-production postprocessing workflow can also be attached to the `continue` task, to be executed when the simulation has completed. The workflow can also be attached to other workflows, for example a pre-production workflow for generating the simulation system in the first place.

An example of a workflow we have used in production is shown in Figure 7.6. This workflow constructs a simulation system from a template in several pre-production tasks, including an energy minimization of the starting configuration. It then performs a pre-production step using a self-propagating workflow, which continues extending the workflow until the pre-production simulation has finished. After this, the production simulation is prepared using the end of the pre-production run as a starting point, and production MD begins thereafter. This production workflow features several post-run post-processing steps, producing a trajectory in which the protein is centered, updating the Sim's `Universe` definition, and extracting the reduced potentials needed for MBAR (see Section 6.2.4).

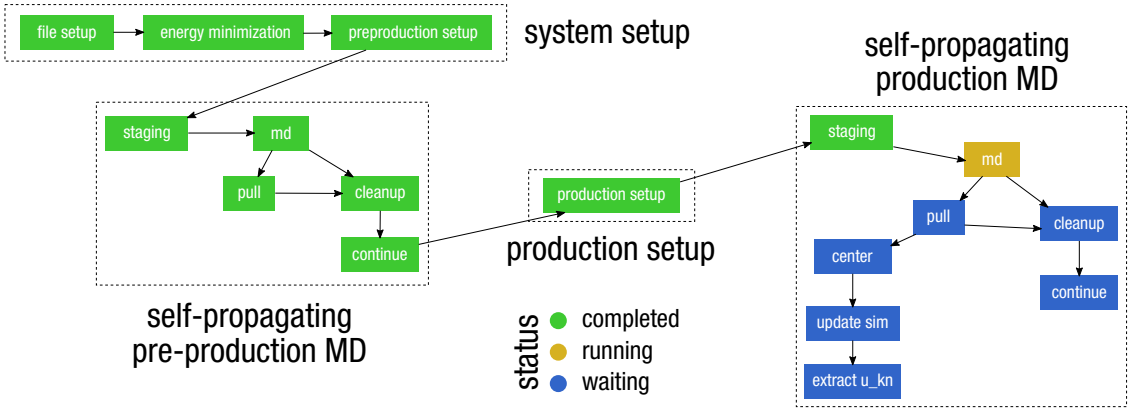


Figure 7.6: A workflow used for creating a simulation system from a template, executing pre-production and production molecular dynamics, and performing per-run post-production steps. This is the dynamic workflow used to perform alchemical binding simulations of the NapA transporter.

The result of casting simulations into workflows, and using MDSynthesis Sims as the unit on which they operate, is that the act of generating raw simulation data requires practically no tedious user input. Workflows can be programmatically generated, and upon being fed to Fireworks, these are then carried out by launcher processes running on different machines, with some launching to HPC queues. The labor shifts from the researcher to the machines, allowing for a high volume of data

to be collected across multiple compute resources. A researcher must only deal with task failures, which are far outnumbered by tasks completing without issue.

7.4.2 Scaling out across multiple clusters

The design of the core molecular dynamics workflow (Figure 7.5) of `mdworks` allows many simulations to be executed across multiple compute resources, with each segment of any given simulation perhaps running on a different resource. This flexibility allows for an *ad hoc* dynamic load-balancing of work across these resources. A compute resource with a saturated queue, not letting many jobs through, means that those jobs are available to go to the queue of another compute resource. Furthermore, with simulation segments as short as 4 hours, individual simulations experience high *churn*: simulations that ran on a fast resource for one segment may run on a slow resource next, and vice-versa. Hundreds of simulations are able to make forward progress each day, and it is no more difficult to manage than running a few of them.

Using four compute resources in this way (Stampede, Saguaro, Hulkster, and our own workstation queue), we have performed nearly 400 NapA alchemical binding simulations (see Chapter 6) simultaneously under this scheme, each numbering over 130,000 atoms in system size. Over the course of 15 days, we collected over 13 μs of trajectory data, totalling nearly 6 TB in raw data, performing nearly 20,000 simulation segments and consuming 1.1 million CPU hours. This kind of throughput, shown graphically in Figure 7.7, would simply be impossible to achieve manually.

Workflow automation of simulation execution under `mdworks` has allowed us to perform hundreds of simulations, totalling tens of thousands of simulation segments, across four independent compute resources, with less time and effort than ever before. As a Python package, `mdworks` is still in early development, but already it is showing itself to be an effective means for doing simulation going forward. Further develop-

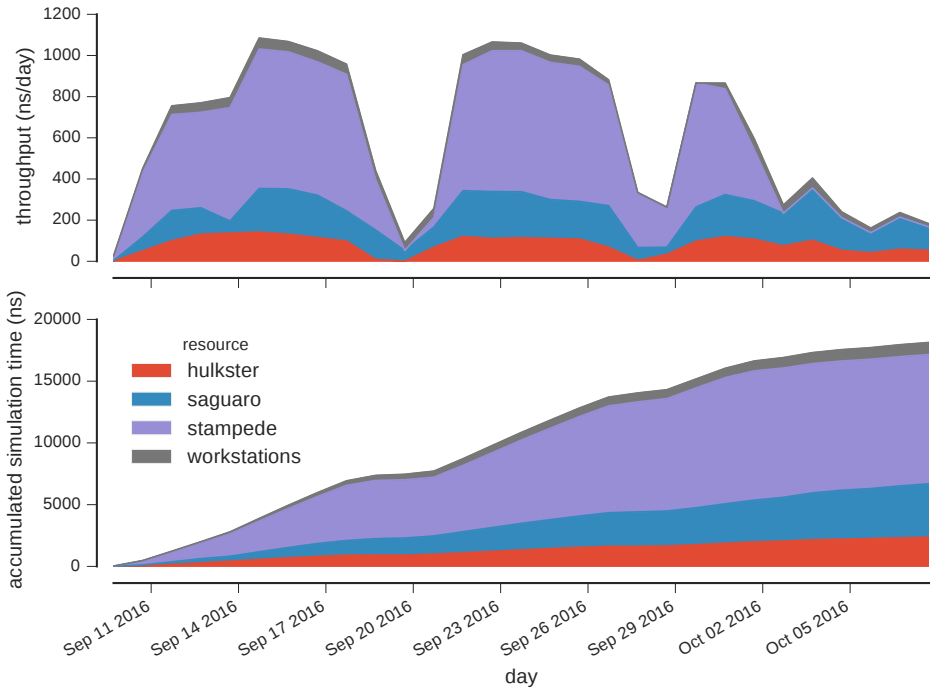


Figure 7.7: Throughput per resource. Simulation throughput (top), measured in nanoseconds per day, and accumulated simulation time (bottom), measured in nanoseconds, over the course of approximately one month and split by the contribution of each resource.

ment will make it easier to deploy, perhaps making its use the norm for performing production simulations across the lab and elsewhere.

Chapter 8

DATREANT: PERSISTENT, PYTHONIC TREES FOR HETEROGENEOUS DATA

This chapter is a reprint of the conference proceeding, **Dotson, D.L.**, Seyler, S.L., Linke, M., Gowers, R.J., and Beckstein, O. (2016). datreant: persistent, Pythonic trees for heterogeneous data. In Proceedings of the 15th Python in Science Conference, S. Benthall, and S. Rostrup, eds. pp. 51–56. This work showcased the functionality and applicability of `datreant`, a Python package that gives a flexible, Pythonic interface to the filesystem. `datreant` is a collaborative project with several core developers, but I am the original creator of the package and lead developer of the project. The package forms an important core of the computational stack used in the Beckstein Lab. See Chapter 7 for more information on the role `datreant` plays in this scope.

I wrote the manuscript, with some assistance from Sean Seyler and Oliver Beckstein. The implementation and design of the package it describes is largely my own, with influence and contributions from others, in particular Max Linke. This work first appeared in the Proceedings of the 15th Python in Science Conference,

ABSTRACT

In science the filesystem often serves as a *de facto* database, with directory trees being the zeroth-order scientific data structure. But it can be tedious and error prone to work directly with the filesystem to retrieve and store heterogeneous datasets. **da-**

treant makes working with directory structures and files Pythonic with **Treants**: specially marked directories with distinguishing characteristics that can be discovered, queried, and filtered. Treants can be manipulated individually and in aggregate, with mechanisms for granular access to the directories and files in their trees. Disparate datasets stored in any format (CSV, HDF5, NetCDF, Feather, etc.) scattered throughout a filesystem can thus be manipulated as meta-datasets of Treants. **datreant** is modular and extensible by design to allow specialized applications to be built on top of it, with MDSynthesis as an example for working with molecular dynamics simulation data. `http://datreant.org/`

8.1 Introduction

In many scientific fields, especially those analyzing experimental or simulation data, there is an existing ecosystem of specialized tools and file formats which new tools must work around. Consequently, specialized database systems may be unsuitable for data management and storage. In these cases the filesystem ends up serving as a *de facto* database, with directory trees the zeroth-order data structure for scientific data. This is particularly true for fields centered around simulation: simulation systems can vary widely in size, composition, rules, parameters, and starting conditions. And with ever-increasing computational power, it is often necessary to store intermediate results from large amounts of simulation data so that they may be accessed and explored interactively.

These problems make data management difficult, and ultimately serve as a barrier to answering scientific questions. To address this, we present `datreant`, a Pythonic interface to the filesystem. `datreant` deals primarily in **Treants**: specially marked directories with distinguishing characteristics that can be discovered, queried, and filtered. Treants can be manipulated individually and in aggregate, with mechanisms

for granular access to the directories and files in their trees. By way of Treants, `datreant` adds a lightweight abstraction layer to the filesystem, allowing researchers to focus more on *what* is stored and less on *where*. This greatly reduces the tedium of storing, retrieving, and operating on datasets of interest, no matter how they are organized.

8.2 Treants as Filesystem Manipulators

The central object of `datreant` is the `Treant`. A Treant is a directory in the filesystem that has been specially marked with a **state file**. A `Treant` is also a Python object. We can create a `Treant` with:

```
>>> import datreant.core as dtr
>>> t = dtr.Treant('maple')
>>> t
<Treant: 'maple'>
```

This creates a directory `maple/` in the filesystem (if it did not already exist), and places a special state file inside which stores the Treant's state. This file also serves as a flagpost indicating that this is more than just a directory:

```
> ls maple
Treant.1dcbb3b1-c396-4bc6-975d-3ae1e4c2983a.json
```

The name of this file includes the type of Treant to which it corresponds, as well as the `uuid` of the Treant, its unique identifier. The state file contains all the information needed to generate an identical instance of this `Treant`, so that we can start a separate Python session and immediately use the same Treant there:

```
# python session 2
>>> import datreant.core as dtr
>>> t = dtr.Treant('maple')
>>> t
<Treant: 'maple'>
```

Making a modification to the `Treant` in one session is immediately reflected by the same `Treant` in any other session. For example, a `Treant` can store any number of descriptive tags to differentiate it from others. We can add tags in the first Python session:

```
# python session 1
>>> t.tags.add('syrup', 'plant')
>>> t.tags
<Tags(['plant', 'syrup'])>
```

And in the other Python session, the same `Treant` with the same tags is visible:

```
# python session 2
>>> t.tags
<Tags(['plant', 'syrup'])>
```

Internally, advisory locking is done to avoid race conditions, making a `Treant` multiprocessing-safe. A `Treant` can also be moved, either locally within the same filesystem or to a remote filesystem, and it will continue to work as expected.

8.2.1 Introspecting a Treant's Tree

A `Treant` can be used to introspect and manipulate its filesystem tree. We can, for example, work with directory structures rather easily:

```
>>> data = t['a/place/for/data/']
>>> data
<Tree: 'maple/a/place/for/data/'>
```

This `Tree` object points to a path in the Treant's own tree, but it need not necessarily exist. We can check this with:

```
>>> data.exists
False
```

This behavior is by design for `Tree` objects (as well as `Leaf` objects; see below). We want to be able to work freely with paths without creating filesystem objects for each, at least until we are ready.

We can make a `Tree` exist in the filesystem easily enough:

```
>>> data.makedirs()
```

and if we also make another directory, too:

```
>>> t['a/place/for/text/'].makedirs()
<Tree: 'maple/a/place/for/text/'>
```

we now have:

```
>>> t.draw()
maple/
 +-- Treant.1dcbb3b1-c396-4bc6-975d-3ae1e4c2983a.json
 +-- a/
     +-- place/
         +-- for/
             +-- data/
             +-- text/
```

Accessing paths in this way returns `Tree` and `Leaf` objects, which refer to directories and files, respectively. These paths need not point to directories or files that actually exist, but they can be used to create and work with these filesystem elements. It should be noted that creating a `Tree` does *not* create a `Treant`. Treants are considered special enough to warrant having a state file with metadata, and making every directory a Treant would make them less useful.

We can, for example, easily store a Pandas[186] DataFrame somewhere in the tree for reference later:

```
>>> import pandas as pd
>>> df = pd.DataFrame(pd.np.random.randn(3, 2),
                      columns=['A', 'B'])
>>> data = t['a/place/for/data/']
>>> data
<Tree: 'maple/a/place/for/data/'>
>>> df.to_csv(data['random_dataframe.csv'].abspath)

# take a look at the contents of `data`
>>> data.draw()
data/
+-- random_dataframe.csv
```

and we can introspect the file directly:

```
>>> csv = data['random_dataframe.csv']
>>> csv
<Leaf: 'maple/a/place/for/data/random_dataframe.csv'>

# this should look like a CSV file
>>> print(csv.read())
,A,B
0,-0.573730932177663,-0.08857033924376226
1,0.03157276797041359,-0.1097921690694506
2,-0.2080757315892524,0.6825003213837373
```

Using `Treant`, `Tree`, and `Leaf` objects, we can work with the filesystem Pythonically without giving much attention to precisely *where* these objects live within that filesystem. This becomes especially powerful when we have many directories/files we want to work with, possibly in many different places.

8.3 Aggregation and Splitting on Treant Metadata

What makes a `Treant` distinct from a `Tree` is its **state file**. This file stores metadata that can be used to filter and split `Treant` objects when treated in aggregate. It also serves as a flagpost, making Treant directories discoverable.

If we have many more Treants, perhaps scattered about the filesystem:

```
>>> for path in ('an/elm/', 'the/oldest/oak',
...              'the/oldest/tallest/sequoia'):
...
...     # make a Treant in filesystem at path
...     dtr.Treant(path)
```

we can gather them up with `datreant.core.discover`:

```
>>> b = dtr.discover('.')
>>> b
<Bundle([<Treant: 'oak'>, <Treant: 'sequoia'>,
         <Treant: 'maple'>, <Treant: 'elm'>])>
```

A `Bundle` is an ordered set of `Treant` objects. This collection gives convenient mechanisms for working with Treants as a single logical unit. For example, it exposes a few basic properties for directly accessing its member data:

```
>>> b.relpaths
['the/oldest/oak/',
 'the/oldest/tallest/sequoia/',
 'maple/',
 'an/elm/']

>>> b.names
['oak', 'sequoia', 'maple', 'elm']
```

A `Bundle` can be constructed in a variety of ways, most commonly using existing `Treant` instances or paths to Treants in the filesystem.

We can use a `Bundle` to subselect Treants in typical ways, including integer indexing and slicing, fancy indexing, boolean indexing, and indexing by name. But in addition to these, we can use metadata features such as **tags** and **categories** to filter and group Treants as desired.

8.3.1 Filtering Treants with tags

Tags are individual strings that describe a Treant. Setting the tags for each of our Treants separately:

```
>>> b['maple'].tags = ['syrup', 'furniture', 'plant']
>>> b['sequoia'].tags = ['huge', 'plant']
>>> b['oak'].tags = ['for building', 'plant', 'building']
>>> b['elm'].tags = ['firewood', 'shady', 'paper',
                     'plant', 'building']
```

we can now work with these tags in aggregate:

```
# will only show tags present in *all* members
>>> b.tags
<AggTags(['plant'])>

# will show tags present among *any* member
>>> b.tags.any
{'building',
 'firewood',
 'for building',
 'furniture',
 'huge',
 'paper',
 'plant',
 'shady',
 'syrup'}
```

and we can filter on them. For example, getting all Treants that are good for construction work:

```
# gives a boolean index for members with this tag
>>> b.tags['building']
[True, False, False, True]

# we can use this to index the Bundle itself
>>> b[b.tags['building']]
<Bundle([<Treant: 'oak'>, <Treant: 'elm'>])>
```

or getting back Treants that are both good for construction *and* used for making furniture by giving tags as a list:

```
# a list of tags serves as an *intersection* query
>>> b[b.tags[['building', 'furniture']]]
<Bundle([])>
```

which in this case none of them are.

Other tag expressions can be constructed using tuples (for *or/union* operations) and sets (for a *negated intersection*), and nesting of any of these works as expected:

```
# we can get a *union* by using a tuple
>>> b[b.tags['building', 'furniture']]
<Bundle([<Treant: 'maple'>, <Treant: 'oak'>,
         <Treant: 'elm'>])>

# we can get a *negated intersection* by using a set
>>> b[b.tags[{'building', 'furniture'}]]
<Bundle([<Treant: 'sequoia'>, <Treant: 'maple'>,
         <Treant: 'oak'>, <Treant: 'elm'>])>
```

Using tag expressions, we can filter to Treants of interest from a `Bundle` counting many, perhaps hundreds, of Treants as members. A common workflow is to use `datreant.core.discover` to gather up many Treants from a section of the filesystem, then use tags to extract only those Treants one actually needs.

8.3.2 Splitting Treants on categories

Categories are key-value pairs that provide another mechanism for distinguishing Treants. We can add categories to each Treant:

```
# add categories to individual members
>>> b['oak'].categories = {'age': 'adult',
                           'type': 'deciduous',
                           'bark': 'mossy'}
>>> b['elm'].categories = {'age': 'young',
                           'type': 'deciduous',
                           'bark': 'smooth'}
>>> b['maple'].categories = {'age': 'young',
                             'type': 'deciduous',
                             'bark': 'mossy'}
>>> b['sequoia'].categories = {'age': 'old',
                               'type': 'evergreen',
                               'bark': 'fibrous',
                               'home': 'california'}

# add value 'tree' to category 'plant'
# for all members
>>> b.categories.add({'plant': 'tree'})
```

and we can access categories for individual Treants:

```
>>> seq = b['sequoia'][0]
>>> seq.categories
<Categories({'home': 'california',
             'age': 'old',
             'type': 'evergreen',
             'bark': 'fibrous',
             'plant': 'tree'})>
```

The aggregated categories for all members in a `Bundle` are accessible via `Bundle.categories`, which gives a view of the categories with keys common to *every* member Treant:

```
>>> b.categories
<AggCategories({'age': ['adult', 'young',
                        'young', 'old'],
                'type': ['deciduous', 'deciduous',
                         'deciduous', 'evergreen'],
                'bark': ['mossy', 'smooth',
                         'mossy', 'fibrous'],
                'plant': ['tree', 'tree',
                          'tree', 'tree']})>
```

Each element of the list associated with a given key corresponds to the value for each member, in member order. Using `Bundle.categories` is equivalent to `Bundle.categories.all`; we can also access categories present among *any* member:

```
>>> b.categories.any
{'age': ['adult', 'young', 'young', 'old'],
 'bark': ['mossy', 'smooth', 'mossy', 'fibrous'],
 'home': [None, None, None, 'california'],
 'type': ['deciduous', 'deciduous',
          'deciduous', 'evergreen']}
```

Members that do not have a given key will have `None` as the corresponding value in the list. Accessing values for a list of keys:

```
>>> b.categories[['age', 'home']]
[['adult', 'young', 'young', 'old'],
 [None, None, None, 'california']]
```

or a set of keys:

```
>>> b.categories[{'age', 'home'}]
{'age': ['adult', 'young', 'young', 'old'],
 'home': [None, None, None, 'california']}
```

returns, respectively, a list or dictionary of lists of values, where the list for a given key is in member order. Perhaps the most powerful feature of categories is the `groupby` method, which, given a key, can be used to group specific members in a `Bundle` by their corresponding category values. If we want to group members by their `'bark'`, we can use `groupby` to obtain a dictionary of members for each value of `'bark'`:

```
>>> b.categories.groupby('bark')
{'fibrous': <Bundle([<Treant: 'sequoia'>])>,
 'mossy': <Bundle([<Treant: 'oak'>,
                   <Treant: 'maple'>])>,
 'smooth': <Bundle([<Treant: 'elm'>])>}
```

Say we would like to get members grouped by both their `'bark'` and `'home'`:

```
>>> b.categories.groupby({'bark', 'home'})
{('fibrous', 'california'):
    <Bundle([<Treant: 'sequoia'>])>}
```

We get only a single member for the pair of keys `('fibrous', 'california')` since `'sequoia'` is the only Treant having the `'home'` category.x Categories are useful as labels to denote the types of data that a Treant may contain or how the data were obtained. By leveraging the `groupby` method, one can extract Treants by selected categories without having to explicitly access each member. This feature can be particularly powerful in cases where many Treants have been created and categorized to handle incoming data over an extended period of time; one can quickly gather any data needed without having to think about low-level details.

8.4 Treant Modularity with Attachable Limbs

`Treant` objects manipulate their tags and categories using `Tags` and `Categories` objects, respectively. These are examples of `Limb` objects: attachable components which serve to extend the capabilities of a `Treant`. While `Tags` and `Categories` are attached by default to all `Treant` objects, custom `Limb` subclasses can be defined for additional functionality.

`datreant` is a namespace package, with the dependency-light core components included in `datreant.core`. The dependencies of `datreant.core` include backports of standard library modules such as `pathlib` and `scandir`, as well as lightweight modules such as `fuzzywuzzy` and `asciitree`.

`datreant.core` remains lightweight because other packages in the `datreant` namespace can have any dependencies they require. One such package is `datreant.data`, which includes a set of convenience `Limb` objects for storing and retrieving Pandas and NumPy[187] datasets in HDF5 using PyTables and h5py internally.

We can attach a `Data` limb to a `Treant` with:

```
>>> import datreant.data
>>> t = dtr.Treant('maple')
>>> t.attach('data')
>>> t.data
<Data([])>
```

and we can immediately start using it to store e.g. a Pandas `Series`:

```
>>> import numpy as np
>>> sn = pd.Series(np.sin(
...     np.linspace(0, 8*np.pi, num=200)))
>>> t.data['sinusoid'] = sn
```

and we can get it back just as easily:

```
>>> t.data['sinusoid'].head()
0    0.000000
1    0.125960
2    0.249913
3    0.369885
4    0.483966
dtype: float64
```

Looking at the directory structure of `"maple"`, we see that the data was stored in an HDF5 file under a directory corresponding to the name we stored it with:

```
>>> t.draw()
maple/
 +-- sinusoid/
 |   +-- pdData.h5
 +-- Treant.1dcbb3b1-c396-4bc6-975d-3ae1e4c2983a.json
```

What's more, `datreant.data` also includes a corresponding `AggLimb` for `Bundle` objects, allowing for automatic aggregation of datasets by name across all member `Treant` objects. If we collect and store similar datasets for each member in our `Bundle`:

```
>>> b = dtr.discover('.')
>>> b
<Bundle([<Treant: 'oak'>, <Treant: 'sequoia'>,
         <Treant: 'maple'>, <Treant: 'elm'>])>

# we want to make each dataset a bit different
>>> b.categories['frequency'] = [1, 2, 3, 4]
>>> for mem in b:
...     freq = mem.categories['frequency']
...     mem.data['sinusoid'] = pd.Series(np.sin(
...         freq * np.linspace(0, 8*np.pi, num=200)))
```

then we can retrieve all of them into a single, multi-index Pandas `Series`:

```
>>> sines = b.data.retrieve('sinusoid', by='name')
>>> sines.groupby(level=0).head()
sequoia  0    0.000000
         1    0.125960
         2    0.249913
         3    0.369885
         4    0.483966
oak      0    0.000000
         1    0.369885
         2    0.687304
         3    0.907232
         4    0.998474
maple    0    0.000000
         1    0.249913
         2    0.483966
         3    0.687304
         4    0.847024
elm      0    0.000000
         1    0.483966
         2    0.847024
         3    0.998474
         4    0.900479
dtype: float64
```

which we can use for aggregated analysis, or perhaps just pretty plots (Figure 8.1).

```
>>> for name, group in sines.groupby(level=0):
...     s = group.reset_index(level=0, drop=True)
...     s.plot(legend=True, label=name)
```

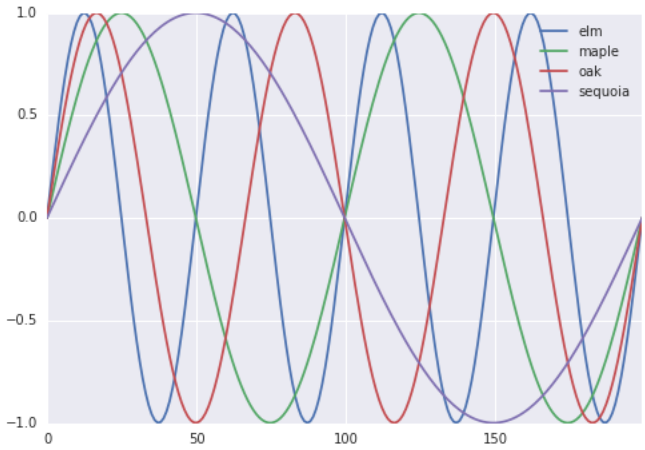


Figure 8.1: Plot of sinusoidal toy datasets aggregated and plotted by source Treant

The `Data` limb stores Pandas and NumPy objects in the HDF5 format within a Treant's own tree. It can also store arbitrary (but pickleable) Python objects as pickles, making it a flexible interface for quick data storage and retrieval. However, it ultimately serves as an example for how `Treant` and `Bundle` objects can be extended to do complex but convenient things.

8.5 Using Treants as the Basis for Dataset Access and Manipulation with the PyData Stack

Although it is possible to extend `datreant` objects with limbs to do complex operations on a Treant's tree, it isn't necessary to build specialized interfaces such as these to make use of the extensive PyData stack. `datreant` fundamentally serves as a Pythonic interface to the filesystem, bringing value to datasets and analysis results by making them easily accessible now and later. As data structures and file formats change, `datreant` objects can always be used in the same way to supplement the way these tools are used.

Because each Treant is both a Python object and a filesystem object, they work remarkably well with distributed computation libraries such as dask.distributed[188] and workflow execution frameworks such as Fireworks[173]. Treant metadata features such as tags and categories can be used for automated workflows, including back-ups and remote copies to external compute resources, making work on datasets less imperative and more declarative when desired.

8.6 Building Domain-Specific Applications on datreant

Built-in `datreant.core` objects are general-purpose, while packages like `datreant.data` provide extensions to these objects that are more specific. But it is possible, and very useful, for domain-specific applications to define their own domain-specific `Treant`

subclasses, with tightly-coupled limbs for domain-specific needs. Not only do objects such as `Bundle` work just fine with `Treant` subclasses and custom `Limb` classes; they are designed explicitly with this need in mind.

The first example of a domain-specific package built around `datreant` is MDSynthesis, a module that enables high-level management and exploration of molecular dynamics simulation data. MDSynthesis gives a Pythonic interface to molecular dynamics trajectories using MDAnalysis[106,184], giving the ability to work with the data from many simulations scattered throughout the filesystem with ease. This package makes it possible to write analysis code that can work across many varieties of simulation, but even more importantly, MDSynthesis allows interactive work with the results from hundreds of simulations at once without much effort.

8.6.1 Leveraging molecular dynamics data with MDSynthesis

MDSynthesis defines a `Treant` subclass called a `Sim`. A `Sim` features special limbs for storing an MDAnalysis `Universe` definition and custom atom selections within its state file, allowing for painless recall of raw simulation data and groups of atoms of interest.

As an example of effectively using `Sims`, say we have 50 biased molecular dynamics simulations that sample the conformational change of the ion transport protein NhaA[32] from the inward-open to outward-open state (Figure 8.2). Let's also say that we are interested in how many hydrogen bonds exist at any given time between the two domains as they move past each other. These `Sim` objects already exist in the filesystem, each having a `Universe` definition already set to point to its unique trajectory file(s).

We can use the MDAnalysis `HydrogenBondAnalysis` class to collect the data for each `Sim` using `Bundle.map` for process parallelism, storing the results using the `datreant.data` limb:

```
import mdsynthesis as mds
import MDAnalysis.analysis.hbonds as hbonds
import pandas as pd
import seaborn as sns

b = mds.discover('NhaA_i2o_transitions')

def get_hbonds(sim):
    dimerization = sim.atomselections['dimer']
    core = sim.atomselections['core']

    hb = hbonds.HydrogenBondAnalysis(
            sim.universe, dimerization, core)
    hb.run()
    hb.generate_table()

    sim.data['hbonds'] = pd.DataFrame(hb.table)

# process parallelism provided internally
# with `multiprocessing`
b.map(get_hbonds, processes=16)
```

Then we can retrieve the datasets in aggregate using the `Bundle` `datreant.data` limb and visualize the result (Figure 8.3):

```
df = b.data.retrieve('hbonds', by='name')

counts = df['distance'].groupby(df.index).count()
counts.index = pd.MultiIndex.from_tuples(
                        counts.index)
counts.index = counts.index.droplevel(0)

sns.jointplot(counts.index, counts, kind='hexbin')
```

By making it relatively easy to work with what can often be many terabytes of simulation data spread over tens or hundreds of trajectories, MDSynthesis greatly reduces the time it takes to iterate on new ideas toward answering real biological questions.

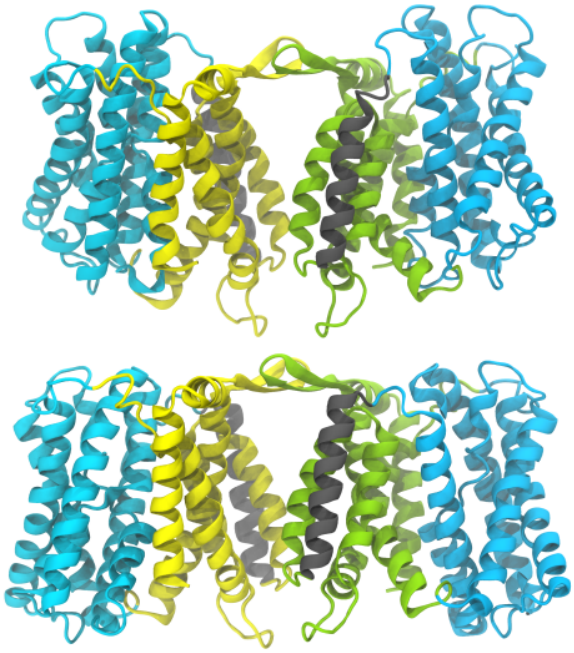


Figure 8.2: A cartoon rendering of an outward-open model (top) and an inward-open crystallographic structure (PDB ID: 4AU5 [32]) (bottom) of *Escherichia coli* NhaA

8.7 Final Thoughts

`datreant` is a young project that started as a domain-specific package for working with molecular dynamics data, but has quickly morphed into a powerful, general-purpose tool for managing and manipulating filesystems and the data spread about them. The dependency-light `datreant.core` package is pure Python, BSD-licensed, and openly developed, and the `datreant` namespace is designed to support useful extensions to the core objects. It is the hope of the authors that `datreant` continues to grow in a way that benefits the wider scientific community, smoothing the common pain point of data glut and filesystem management.

8.8 Acknowledgements

DLD was in part supported by a Molecular Imaging Fellowship from the Department of Physics at Arizona State University. SLS was supported in part by a Wally

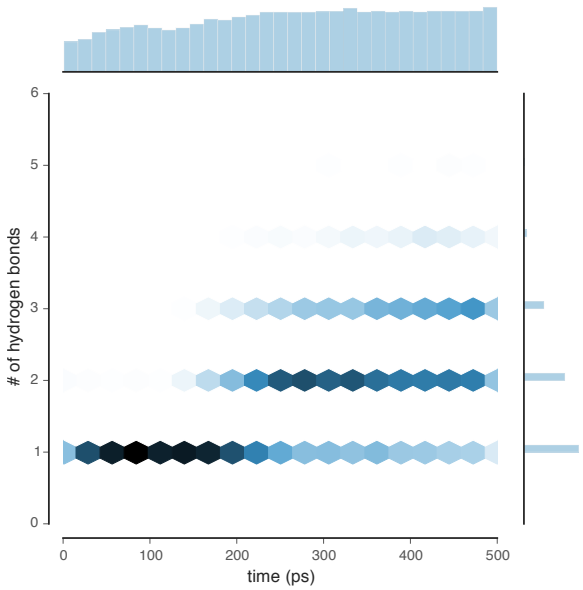


Figure 8.3: The number of hydrogen bonds between the core and dimerization domain during a conformational transition between the inward-open and outward-open state of EcNhaA

Stoelzel Fellowship from the Department of Physics at Arizona State University. ML was supported by the Max Planck Society. RG was supported by BBSRC grant BB/J014478/1. OB was supported in part by grant ACI-1443054 from the National Science Foundation; computational resources for OB's work were in part provided by the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575 (allocation MCB130177 to OB).

Chapter 9

CONCLUSIONS

Since first being discovered in 1974 by West and Mitchell[8], our knowledge of Na^+/H^+ antiporters has come far. After discovery of the *nhaA* gene[10], subsequent sequencing[11], and purification of *Escherichia coli* NhaA[12], functional studies offered a quantitative view of the pH-dependence of NhaA, as well as its electrogenicity (1 Na^+ / 2 H^+). Highly-conserved residues vital for transport, including Asp163, Asp164, Asp133, Thr132, and Lys300 were identified, but their role in the transport mechanism remained elusive[14]. The first crystal structure of NhaA[22] revealed, for the first time, the spatial arrangement of these residues, as well as the organization of the protein into two distinct domains, a dimerization interface and a core domain featuring many of the vital residues. This structure paved the way for studies of the conformational change required for translocation[23,25,26,189], and also gave context for the many functional studies seeking to determine more clearly the Na^+ and H^+ binding sites. A detailed electrophysiological study of NhaA showed that the pH-dependent activity of NhaA could be accounted for by competitive binding between Na^+ and H^+, suggesting that Na^+ and H^+ bind to the same location in the protein[27].

Extending from this context, the work we have detailed in the preceeding chapters has greatly forwarded our understanding of Na^+/H^+ antiport. The outward-facing *Thermus thermophilus* NapA structure[31] offered the first suggestion that the conformational change required for ion translocation in these transporters was large, a $\sim$10 Å translation of the entire core domain relative to the dimerization domain (Chapter 3). The new inward-facing NhaA structure[32] revealed a previously-unidentified salt bridge between the highly-conserved Asp163 and Lys300 residues, and molecular dy-

namics simulations revealed that spontaneous binding of Na^+ to Asp164 breaks the salt bridge. This, along with the presence of the same salt bridge in the outward-facing NapA structure[31], suggested a direct role for Lys300 in the transport mechanism as a H^+ carrier (Chapter 4). Since then, this has been further corroborated using constant-pH molecular dynamics simulations[33]. Our most recent published work, which presented a new inward-facing NapA structure obtained through disulfide-locking of a cysteine mutant, established unambiguously that the core domain undergoes a large-scale, elevator-like translocation relative to the dimerization domain during the transport cycle, as opposed to small movements proposed by others[23,30] (Chapter 5).

Our published work has greatly influenced the conversation on Na^+/H^+ antiporters[35–37], and we continue to work to find answers to remaining questions. Operating under the hypothesis that the lysine is a proton carrier, becoming deprotonated during the transport cycle[31,33], we have examined the strength of Na^+ binding for NapA as a function of conformation and charge state using alchemical binding free energy calculations. These have revealed that binding of Na^+ is weaker in the outward- than in the inward-facing state, and that the binding affinity of Na^+ is highly dependent on the charge state of the conserved lysine (Lys305). This asymmetry is physiologically advantageous for the protein[180], since it must release the Na^+ to the outside under most conditions, but more work needs to be done to extract correct absolute binding free energies, and ultimately apparent binding affinities, from these calculations (Chapter 6).

Finally, we have shown that software has not played a small role in our efforts to understand the molecular mechanism of Na^+/H^+ antiport. We presented software packages we have developed to probe questions of ever-increasing scale and complexity (Chapter 7), and how thoughtful and collaborative design of these tools is key to

making them usable in new and flexible ways. In particular, we detailed the design and use of `datreant`[183], a Python package giving a powerful interface to the filesystem, that forms an important core of our software stack (Chapter 8).

The work we have done, and the work we continue to do is moving the needle toward fully understanding the mechanism of Na^+/H^+ antiport. There remain unanswered questions, including how binding of $2H^+$ or $1Na^+$ is coupled to the translocation of the core domain, and questions with answers but incomplete evidence, such as where the H^+ bind. But given the progress over the 40-year effort to understand these transporters, we are confident that the solutions will come with continued work.

[1] Forrest L R, Krämer R and Ziegler C, 2011 The structural basis of secondary active transport mechanisms. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1807** 167–188.

[2] Schuetz J D, Swaan P W and Tweedie D J, 2014 The role of transporters in toxicity and disease. *Drug Metabolism and Disposition* **42** 541–545.

[3] Shi Y, 2013 Common folds and transport mechanisms of secondary active transporters. *Annual Review of Biophysics* **42** 51–72.

[4] Brett C L, Donowitz M and Rao R, 2005 Evolutionary origins of eukaryotic sodium/proton exchangers. *Am J Physiol Cell Physiol* **288** C223–C239.

[5] Fuster D G and Alexander R T, 2013 Traditional and emerging roles for the SLC9 Na+/H+ exchangers. *Pflügers Archiv - European Journal of Physiology* **466** 61–76.

[6] Kondapalli K C, Hack A, Schushan M, Landau M, Ben-Tal N and Rao R, 2013 Functional evaluation of autism-associated mutations in NHE9. *Nature Communications* **4** 2510.

[7] Kondapalli K C, Prasad H and Rao R, 2014 An inside job: how endosomal Na+/H+ exchangers link to autism and neurological disease. *Frontiers in Cellular Neuroscience* **8** 172.

[8] West I C and Mitchell P, 1974 Proton/sodium ion antiport in Escherichia coli. *Biochem. J* **144** 87–90.

[9] Padan E, Zilberstein D and Rottenberg H, 1976 The proton electrochemical gradient in Escherichia coli cells. *European Journal of Biochemistry* **63** 533–541.

[10] Goldberg E B, Arbel T, Chen J, Karpel R, Mackie G A, Schuldiner S and Padan E, 1987 Characterization of a Na+/H+ antiporter gene of Escherichia coli. *Proceedings of the National Academy of Sciences* **84** 2615–2619.

[11] Karpel R, Olami Y, Taglicht D, Schuldiner S and Padan E, 1988 Sequencing of the gene ant which affects the Na+/H+ antiporter activity in Escherichia coli. *Journal of Biological Chemistry* **263** 10408–10414.

[12] Taglicht D, Padan E and Schuldiner S, 1991 Overproduction and purification of a functional Na+/H+ antiporter coded by nhaA (ant) from Escherichia coli. *Journal of Biological Chemistry* **266** 11289–11294.

[13] Taglicht D, Padan E and Schuldiner S, 1993 Proton-sodium stoichiometry of NhaA, an electrogenic antiporter from Escherichia coli. *The Journal of Biological Chemistry* **268** 5382–5387.

[14] Inoue H, Noumi T, Tsuchiya T and Kanazawa H, 1995 Essential aspartic acid residues, Asp-133, Asp-163 and Asp-164, in the transmembrane helices of a Na+/H+ antiporter (NhaA) from Escherichia coli. *FEBS Letters* **363** 264–268.

[15] Noumi T, Inoue H, Sakurai T, Tsuchiya T and Kanazawa H, 1997 Identification and Characterization of Functional Residues in a Na+/H+ Antiporter (NhaA) from Escherichia coli by Random Mutagenesis. *Journal of Biochemistry* **121** 661–670.

[16] Rimon A, Gerchman Y, Kariv Z and Padan E, 1998 A Point Mutation (G338S) and Its Suppressor Mutations Affect Both the pH Response of the NhaA-Na+/H+ Antiporter as Well as the Growth Phenotype of Escherichia coli. *Journal of Biological Chemistry* **273** 26470–26476.

[17] Williams K A, Geldmacher-Kaufer U, Padan E, Schuldiner S and Kühlbrandt W, 1999 Projection structure of NhaA, a secondary transporter from Escherichia coli, at 4.0 Å resolution. *The EMBO Journal* **18** 3558–3563.

[18] Williams K A, 2000 Three-dimensional structure of the ion-coupled transport protein NhaA. *Nature* **403** 112–115.

[19] Galili L, Rothman A, Kozachkov L, Rimon A and Padan E, 2002 Trans Membrane Domain IV Is Involved in Ion Transport Activity and pH Regulation of the NhaA-Na+/H+ Antiporter of Escherichia coli. *Biochemistry* **41** 609–617.

[20] Galili L, Herz K, Dym O and Padan E, 2004 Unraveling Functional and Structural Interactions between Transmembrane Domains IV and XI of NhaA Na+/H+ Antiporter of Escherichia coli. *Journal of Biological Chemistry* **279** 23104–23113.

[21] Kozachkov L, Herz K and Padan E, 2007 Functional and Structural Interactions of the Transmembrane Domain X of NhaA, Na+/H+ Antiporter of Escherichia coli, at Physiological pH. *Biochemistry* **46** 2419–2430.

[22] Hunte C, Screpanti E, Venturi M, Rimon A, Padan E and Michel H, 2005 Structure of a Na+/H+ antiporter and insights into mechanism of action and regulation by pH. *Nature* **435** 1197–202.

[23] Arkin I T, Xu H, Jensen M O, Arbely E, Bennett E R, Bowers K J, Chow E, Dror R O, Eastwood M P, Flitman-Tene R, Gregersen B A, Klepeis J L, Kolossváry I, Shan Y and Shaw D E, 2007 Mechanism of Na+/H+ Antiporting. *Science* **317** 799–803.

[24] Appel M, Hizlan D, Vinothkumar K R, Ziegler C and Kühlbrandt W, 2009 Conformations of NhaA, the Na/H exchanger from Escherichia coli, in the pH-activated and ion-translocating states. *Journal of molecular biology* **386** 351–65.

[25] Olkhova E, Kozachkov L, Padan E and Michel H, 2009 Combined computational and biochemical study reveals the importance of electrostatic interactions between the "pH sensor" and the cation binding site of the sodium/proton antiporter NhaA of *Escherichia coli*. *Proteins: Structure, Function, and Bioinformatics* **76** 548–559.

[26] Herz K, Rimon A, Olkhova E, Kozachkov L and Padan E, 2010 Transmembrane Segment II of NhaA Na+/H+ Antiporter Lines the Cation Passage, and Asp65 Is Critical for pH Activation of the Antiporter. *Journal of Biological Chemistry* **285** 2211–2220.

[27] Mager T, Rimon A, Padan E and Fendler K, 2011 Transport mechanism and pH regulation of the Na+/H+ antiporter NhaA from Escherichia coli: an electrophysiological study. *The Journal of Biological Chemistry* **286** 23570–23581.

[28] Maes M, Rimon A, Kozachkov-Magrisso L, Friedler A and Padan E, 2012 Revealing the ligand binding site of NhaA Na+/H+ antiporter and its pH dependence. *The Journal of Biological Chemistry* **287** 38150–7.

[29] Padan E, 2008 The enlightening encounter between structure and function in the NhaA Na+/H+ antiporter. *Trends Biochem Sci* **33** 435–443.

[30] Paulino C, Wöhlert D, Kapotova E, Yildiz Ö and Kühlbrandt W, 2014 Structure and transport mechanism of the sodium/proton antiporter MjNhaP1. *eLife* e03583.

[31] Lee C, Kang H J, von Ballmoos C, Newstead S, Uzdavinys P, Dotson D L, Iwata S, Beckstein O, Cameron A D and Drew D, 2013 A two-domain elevator mechanism for sodium/proton antiport. *Nature* **501** 573–577.

[32] Lee C, Yashiro S, Dotson D L, Uzdavinys P, Iwata S, Sansom M S P, Ballmoos C v, Beckstein O, Drew D and Cameron A D, 2014 Crystal structure of the sodium-proton antiporter NhaA dimer and new mechanistic insights. *The Journal of General Physiology* **144** 529–544.

[33] Huang Y, Chen W, Dotson D L, Beckstein O and Shen J, 2016 Mechanism of pH-dependent activation of the sodium-proton antiporter NhaA. *Nature Communications* **7** 12940.

[34] Coincon M, Uzdavinys P, Nji E, Dotson D L, Winkelmann I, Abdul-Hussein S, Cameron A D, Beckstein O and Drew D, 2016 Crystal structures reveal the molecular basis of ion translocation in sodium/proton antiporters. *Nature Structural & Molecular Biology* **23** 248–255.

[35] Ryan R M and Vandenberg R J, 2016 Elevating the alternating-access model. *Nature Structural & Molecular Biology* **23** 187–189.

[36] Mulligan C, Fenollar-Ferrer C, Fitzgerald G A, Vergara-Jaque A, Kaufmann D, Li Y, Forrest L R and Mindell J A, 2016 The bacterial dicarboxylate transporter VcINDY uses a two-domain elevator-type mechanism. *Nature Structural & Molecular Biology* **23** 256–263.

[37] Drew D and Boudker O, 2016 Shared Molecular Mechanisms of Membrane Transporters. *Annual Review of Biochemistry* **85** 543–572.

[38] Hill T L, 2012 *Free energy transduction and biochemical cycle kinetics* (Springer Science & Business Media).

[39] Padan E, Zilberstein D and Schuldiner S, 1981 pH homeostasis in bacteria. *Biochimica et Biophysica Acta* **650** 151–166.

[40] Krulwich T A, Sachs G and Padan E, 2011 Molecular aspects of bacterial pH sensing and homeostasis. *Nat Rev Microbiol* **9** 330–343.

[41] Denning E J and Beckstein O, 2013 Influence of lipids on protein-mediated transmembrane transport. *Chemistry and Physics of Lipids* **169** 57–71.

[42] MacKerell Jr A D, Bashford D, Bellott M, Dunbrack Jr R L, Evanseck J D, Field M J, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau F, Mattos C, Michnick S, Ngo T, Nguyen D, Prodhom B, Reiher W, Roux B, Schlenkrich M, Smith J, Stote R, Straub J, Watanabe M, Wiokiewicz-Kuczera J, Yin D and Karplus M, 1998 All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B* **102** 3586–3616.

[43] MacKerell A D, Feig M and Brooks C L, 2004 Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *Journal of Computational Chemistry* **25** 1400–1415.

[44] Best R B, Zhu X, Shim J, Lopes P E, Mittal J, Feig M and MacKerell Jr A D, 2012 Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *Journal of Chemical Theory and Computation* **8** 3257–3273.

[45] Klauda J B, Venable R M, Freites J A, O'Connor J W, Tobias D J, Mondragon-Ramirez C, Vorobyov I, MacKerell Jr A D and Pastor R W, 2010 Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. *The Journal of Physical Chemistry B* **114** 7830–7843.

[46] Rizzo R C and Jorgensen W L, 1999 OPLS all-atom model for amines: resolution of the amine hydration problem. *Journal of the American Chemical Society* **121** 4827–4836.

[47] Kaminski G A, Friesner R A, Tirado-Rives J and Jorgensen W L, 2001 Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *The Journal of Physical Chemistry B* **105** 6474–6487.

[48] Jensen K P and Jorgensen W L, 2006 Halide, ammonium, and alkali metal ion parameters for modeling aqueous solutions. *Journal of Chemical Theory and Computation* **2** 1499–1509.

[49] Ulmschneider J P and Ulmschneider M B, 2009 United Atom Lipid Parameters for Combination with the Optimized Potentials for Liquid Simulations All-Atom Force Field. *Journal of Chemical Theory and Computation* **5** 1803–1813.

[50] Stansfeld P J and Sansom M S, 2011 From Coarse Grained to Atomistic: A Serial Multiscale Approach to Membrane Protein Simulations. *Journal of Chemical Theory and Computation* **7** 1157–1166.

[51] Jorgensen W L, Chandrasekhar J, Madura J D, Impey R W and Klein M L, 1983 Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79** 926–935.

[52] Pronk S, Páll S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, Shirts M R, Smith J C, Kasson P M, van der Spoel D, Hess B and Lindahl E, 2013 GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* btt055.

[53] Hockney R, Goel S and Eastwood J, 1974 Quiet high-resolution computer models of a plasma. *Journal of Computational Physics* **14** 148–158.

[54] Abraham M, Hess B, van Der Spoel D and Lindahl E, 2016 Gromacs user manual version 5.1.4. *University of Groningen, The Netherlands* .

[55] Hess B, 2008 P-LINCS: A parallel linear constraint solver for molecular simulation. *Journal of Chemical Theory and Computation* **4** 116–122.

[56] Miyamoto S and Kollman P A, 1992 Settle: An analytical version of the shake and rattle algorithm for rigid water molecules. *J. Comput. Chem* **13** 952–962.

[57] Kenney I M, Beckstein O and Iorga B I, 2016 Prediction of cyclohexane-water distribution coefficients for the SAMPL5 data set using molecular dynamics simulations with the OPLS-AA force field. *Journal of Computer-Aided Molecular Design* 1–14.

[58] Leach A R, 2001 *Molecular modelling: principles and applications* (Pearson Education).

[59] Lennard J and Jones I, 1924 On the Determination of Molecular Fields, in *Proc. R. Soc. London*, vol. 106, 441–477.

[60] Ewald P P, 1921 Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik* **369** 253–287.

[61] Darden T, York D and Pedersen L, 1993 Particle mesh Ewald: An $N \log(N)$ method for Ewald sums in large systems. *The Journal of Chemical Physics* **98** 10089–10092.

[62] Bussi G, Donadio D and Parrinello M, 2007 Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **126** 014101.

[63] Parrinello M and Rahman A, 1981 Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **52** 7182–7190.

[64] Kozachkov L and Padan E, 2013 Conformational changes in NhaA Na+/H+ antiporter. *Molecular Membrane Biology* **30** 90–100.

[65] Pinner E, Padan E and Schuldiner S, 1994 Kinetic properties of NhaB, a Na+/H+ antiporter from Escherichia coli. *Journal of Biological Chemistry* **269** 26274–26279.

[66] Drew D, Lerch M, Kunji E, Slotboom D J and de Gier J W, 2006 Optimization of membrane protein overexpression and purification using GFP fusions. *Nature Methods* **3** 303–313.

[67] Sonoda Y, Newstead S, Hu N J, Alguel Y, Nji E, Beis K, Yashiro S, Lee C, Leung J, Cameron A D, Byrne B, Iwata S and Drew D, 2011 Benchmarking membrane protein detergent stability for improving throughput of high-resolution x-ray structures. *Structure* **19** 17 – 25.

[68] Furrer E M, Ronchetti M F, Verrey F and Pos K M, 2007 Functional characterization of a NapA Na+/H+ antiporter from Thermus thermophilus. *FEBS Letters* **581** 572–578.

[69] Padan E, Tzubery T, Herz K, Kozachkov L, Rimon A and Galili L, 2004 NhaA of Escherichia coli, as a model of a pH-regulated Na+/H+ antiporter. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1658** 2–13.

[70] Rimon A, Tzubery T and Padan E, 2007 Monomers of the NhaA Na+/H+ Antiporter of Escherichia coli Are Fully Functional yet Dimers Are Beneficial under Extreme Stress Conditions at Alkaline pH in the Presence of Na+ or Li+. *Journal of Biological Chemistry* **282** 26810–26821.

[71] Hisamitsu T, Ammar Y B, Nakamura T Y and Wakabayashi S, 2006 Dimerization Is Crucial for the Function of the Na+/H+ Exchanger NHE1. *Biochemistry* **45** 13346–13355, pMID: 17073455.

[72] Goswami P, Paulino C, Hizlan D, Vonck J, Yildiz Ö and Kühlbrandt W, 2010 Structure of the archaeal Na+/H+ antiporter NhaP1 and functional role of transmembrane helix 1. *The EMBO Journal* **30** 439–449.

[73] Kuwabara N, Inoue H, Tsuboi Y, Nakamura N and Kanazawa H, 2004 The Fourth Transmembrane Domain of the Helicobacter pylori Na+/H+ Antiporter NhaA Faces a Water-filled Channel Required for Ion Transport. *Journal of Biological Chemistry* **279** 40567–40575.

[74] Screpanti E and Hunte C, 2007 Discontinuous membrane helices in transport proteins and their correlation with function. *Journal of Structural Biology* **159** 261–267.

[75] Hu N J, Iwata S, Cameron A D and Drew D, 2011 Crystal structure of a bacterial homologue of the bile acid sodium symporter ASBT. *Nature* **478** 408–411.

[76] Vinothkumar K R, Smits S H and Kühlbrandt W, 2005 pH-induced structural change in a sodium/proton antiporter from Methanococcus jannaschii. *The EMBO Journal* **24** 2720–2729.

[77] Reyes N, Ginter C and Boudker O, 2009 Transport mechanism of a bacterial homologue of glutamate transporters. *Nature* **462** 880–885.

[78] Schushan M, Rimon A, Haliloglu T, Forrest L R, Padan E and Ben-Tal N, 2012 A Model-Structure of a Periplasm-facing State of the NhaA Antiporter Suggests the Molecular Underpinnings of pH-induced Conformational Changes. *Journal of Biological Chemistry* **287** 18249–18261.

[79] Drew D E, von Heijne G, Nordlund P and de Gier J W L, 2001 Green fluorescent protein as an indicator to monitor membrane protein overexpression in Escherichia coli. *FEBS Letters* **507** 220–224.

[80] Kawate T and Gouaux E Fluorescence-Detection Size-Exclusion Chromatography for Precrystallization Screening of Integral Membrane Proteins. *Structure* **14** 673–681.

[81] Wagner S, Klepsch M M, Schlegel S, Appel A, Draheim R, Tarry M, Högbom M, van Wijk K J, Slotboom D J, Persson J O and de Gier J W, 2008 Tuning Escherichia coli for membrane protein overexpression. *Proceedings of the National Academy of Sciences* **105** 14371–14376.

[82] Studier F W, 2005 Protein production by auto-induction in high-density shaking cultures. *Protein Expression and Purification* **41** 207 – 234.

[83] Drew D, Newstead S, Sonoda Y, Kim H, von Heijne G and Iwata S, 2008 GFP-based optimization scheme for the overexpression and purification of eukaryotic membrane proteins in Saccharomyces cerevisiae. *Nat. Protocols* **3** 784–798.

[84] Slotboom D J, Duurkens R H, Olieman K and Erkens G B, 2008 Static light scattering to characterize membrane proteins in detergent solution. *Methods* **46** 73 – 82, New Methods in Membrane Protein Research.

[85] Wiedenmann A, Dimroth P and Von Ballmoos C, 2009 Functional asymmetry of the F_0 motor in bacterial ATP synthases. *Molecular Microbiology* **72** 479–490.

[86] Ishmukhametov R R, Galkin M A and Vik S B, 2005 Ultrafast purification and reconstitution of His-tagged cysteine-less Escherichia coli F_1F_o ATP synthase. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1706** 110 – 116.

[87] Wiedenmann A, Dimroth P and von Ballmoos C, 2008 $\Delta\psi$ and ΔpH are equivalent driving forces for proton transport through isolated F_0 complexes of ATP synthases. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1777** 1301 – 1310.

[88] Winter G, 2010 *xia2*: an expert system for macromolecular crystallography data reduction. *Journal of Applied Crystallography* **43** 186–190.

[89] Kabsch W, 2010 *XDS*. *Acta Crystallographica Section D* **66** 125–132.

[90] Collaborative Computational Project Number 4, 1994 The *CCP4* suite: programs for protein crystallography. *Acta Crystallographica Section D* **50** 760–763.

[91] Knight S D, 2000 *RSPS* version 4.0: a semi-interactive vector-search program for solving heavy-atom derivatives. *Acta Crystallographica Section D* **56** 42–47.

[92] Bricogne G, Vonrhein C, Flensburg C, Schiltz M and Paciorek W, 2003 Generation, representation and flow of phase information in structure determination: recent developments in and around *SHARP* 2.0. *Acta Crystallographica Section D* **59** 2023–2030.

[93] Cowtan K, 1994 DM: an automated procedure for phase improvement by density modification. *Joint CCP4 and ESF-EACBM newsletter on protein crystallography* **31** 34–38.

[94] Jones T and Kjeldgaard M, 1997 Electron-density map interpretation, in *Macromolecular Crystallography Part B*, vol. 277 of *Methods in Enzymology*, 173 – 208 (Academic Press).

[95] McCoy A J, Grosse-Kunstleve R W, Adams P D, Winn M D, Storoni L C and Read R J, 2007 Phaser crystallographic software. *Journal of Applied Crystallography* **40** 658–674.

[96] Emsley P and Cowtan K, 2004 *Coot*: model-building tools for molecular graphics. *Acta Crystallographica Section D* **60** 2126–2132.

[97] Adams P D, Afonine P V, Bunkóczi G, Chen V B, Davis I W, Echols N, Headd J J, Hung L W, Kapral G J, Grosse-Kunstleve R W, McCoy A J, Moriarty N W, Oeffner R, Read R J, Richardson D C, Richardson J S, Terwilliger T C and Zwart P H, 2010 *PHENIX*: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D* **66** 213–221.

[98] Kleywegt G J and Jones T A, 1994 A super position. *ESF/CCP4 Newsletter* **31** 14.

[99] Delano W L, 2002, The PyMOL Molecular Graphics System, `http://www.pymol.org`.

[100] Potterton L, McNicholas S, Krissinel E, Gruber J, Cowtan K, Emsley P, Murshudov G N, Cohen S, Perrakis A and Noble M, 2004 Developments in the *CCP4* molecular-graphics project. *Acta Crystallographica Section D* **60** 2288–2294.

[101] Hess B, Kutzner C, Van Der Spoel D and Lindahl E, 2008 GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation* **4** 435–447.

[102] Bond P J, Wee C L and Sansom M S, 2008 Coarse-Grained Molecular Dynamics Simulations of the Energetics of Helix Insertion into a Lipid Bilayer. *Biochemistry* **47** 11321–11331.

[103] Scott K A, Bond P J, Ivetac A, Chetwynd A P, Khalid S and Sansom M S, 2008 Coarse-grained MD simulations of membrane protein-bilayer self-assembly. *Structure* **16** 621–630.

[104] Li H, Robertson A D and Jensen J H, 2005 Very fast empirical prediction and rationalization of protein pKa values. *Proteins: Structure, Function, and Bioinformatics* **61** 704–721.

[105] Essmann U, Perera L, Berkowitz M L, Darden T, Lee H and Pedersen L G, 1995 A smooth particle mesh Ewald method. *The Journal of Chemical Physics* **103** 8577–8593.

[106] Michaud-Agrawal N, Denning E J, Woolf T B and Beckstein O, 2011 MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry* **32** 2319–2327.

[107] Humphrey W, Dalke A and Schulten K, 1996 VMD: visual molecular dynamics. *Journal of Molecular Graphics* **14** 33–38.

[108] Dahl A C E, Chavent M and Sansom M S, 2012 Bendix: intuitive helix geometry analysis and abstraction. *Bioinformatics* **28** 2193–2194.

[109] Pettersen E F, Goddard T D, Huang C C, Couch G S, Greenblatt D M, Meng E C and Ferrin T E, 2004 UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **25** 1605–1612.

[110] Goddard T D, Huang C C and Ferrin T E, 2007 Visualizing density maps with UCSF Chimera. *Journal of Structural Biology* **157** 281–287.

[111] Chen V B, Arendall W B, Headd J J, Keedy D A, Immormino R M, Kapral G J, Murray L W, Richardson J S and Richardson D C, 2010 Molprobity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography* **66** 12–21.

[112] Olsson M H M, Søndergaard C R, Rostkowski M and Jensen J H, 2011 PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *Journal of Chemical Theory and Computation* **7** 525–537.

[113] Lee C, Kang H J, von Ballmoos C, Newstead S, Uzdavinys P, Dotson D L, Iwata S, Beckstein O, Cameron A D and Drew D, 2013 A two-domain elevator mechanism for sodium/proton antiport. *Nature* **501** 573–577.

[114] Fafournoux P, Noel J and Pouyssegur J, 1994 Evidence that Na+/H+ exchanger isoforms NHE1 and NHE3 exist as stable dimers in membranes with a high degree of specificity for homodimers. *Journal of Biological Chemistry* **269** 2589–2596.

[115] Hilger D, Jung H, Padan E, Wegener C, Vogel K P, Steinhoff H J and Jeschke G, 2005 Assessing oligomerization of membrane proteins by four-pulse DEER: pH-dependent dimerization of NhaA Na+/H+ antiporter of E. coli. *Biophysical Journal* **89** 1328–1338.

[116] Daley D O, Rapp M, Granseth E, Melén K, Drew D and Von Heijne G, 2005 Global topology analysis of the escherichia coli inner membrane proteome. *Science* **308** 1321–1323.

[117] Lee C, Kang H J, Hjelm A, Qureshi A A, Nji E, Choudhury H, Beis K, de Gier J W and Drew D, 2014 MemStar: A one-shot Escherichia coli-based approach for high-level bacterial membrane protein production. *FEBS letters* **588** 3761–3769.

[118] Otwinowski Z and Minor W, 1997 Processing of X-ray diffraction data collected in oscillation mode. *Macromol Crystallogr Part A* **276** 307–326.

[119] Kleywegt G, Zou J Y, Kjeldgaard M and Jones T, 2006 Around O, in *International Tables for Crystallography Volume F: Crystallography of biological macromolecules*, 353–356 (Springer).

[120] DeLaBarre B and Brunger A T, 2006 Considerations for the refinement of low-resolution crystal structures. *Acta Crystallographica Section D: Biological Crystallography* **62** 923–932.

[121] Winn M, Isupov M and Murshudov G N, 2001 Use of TLS parameters to model anisotropic displacements in macromolecular refinement. *Acta Crystallographica Section D: Biological Crystallography* **57** 122–133.

[122] Domański J, Stansfeld P J, Sansom M S and Beckstein O, 2010 Lipidbook: a public repository for force-field parameters used in membrane simulations. *The Journal of Membrane Biology* **236** 255–258.

[123] Ulmschneider M B, Doux J P, Killian J A, Smith J C and Ulmschneider J P, 2010 Mechanism and kinetics of peptide partitioning into membranes from all-atom simulations of thermostable peptides. *Journal of the American Chemical Society* **132** 3452–3460.

[124] Ulmschneider J P, Smith J C, White S H and Ulmschneider M B, 2011 In silico partitioning and transmembrane insertion of hydrophobic peptides under equilibrium conditions. *Journal of the American Chemical Society* **133** 15487–15495.

[125] Rappolt M, Hickel A, Bringezu F and Lohner K, 2003 Mechanism of the lamellar/inverse hexagonal phase transition examined by high resolution X-ray diffraction. *Biophysical Journal* **84** 3111–3122.

[126] Murzyn K, Róg T and Pasenkiewicz-Gierula M, 2005 Phosphatidylethanolamine-phosphatidylglycerol bilayer as a model of the inner bacterial membrane. *Biophysical Journal* **88** 1091–1103.

[127] Kučerka N, Tristram-Nagle S and Nagle J F, 2006 Structure of fully hydrated fluid phase lipid bilayers with monounsaturated chains. *The Journal of Membrane Biology* **208** 193–202.

[128] Raetz C R, 1986 Molecular genetics of membrane phospholipid synthesis. *Annual Review of Genetics* **20** 253–291.

[129] Cronan J E, 2003 Bacterial membrane lipids: where do we stand? *Annual Reviews in Microbiology* **57** 203–224.

[130] Berendsen H J, Postma J v, van Gunsteren W F, DiNola A and Haak J, 1984 Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **81** 3684–3690.

[131] Søndergaard C R, Olsson M H M, Rostkowski M and Jensen J H, 2011 Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pK_a Values. *Journal of Chemical Theory and Computation* **7** 2284–2295.

[132] Kumar S and Nussinov R, 1999 Salt bridge stability in monomeric proteins. *Journal of Molecular Biology* **293** 1241–1255.

[133] Kumar S and Nussinov R, 2002 Relationship between ion pair geometries and electrostatic strengths in proteins. *Biophysical Journal* **83** 1595–1612.

[134] Scott D W, 2015 *Multivariate density estimation: theory, practice, and visualization* (John Wiley & Sons).

[135] Hintze J L and Nelson R D, 1998 Violin plots: a box plot-density trace synergism. *The American Statistician* **52** 181–184.

[136] Waskom M *et al.*, 2014 Seaborn: statistical data visualization .

[137] Herz K, Rimon A, Jeschke G and Padan E, 2009 Beta-Sheet-dependent Dimerization Is Essential for the Stability of NhaA Na+/H+ Antiporter. *Journal of Biological Chemistry* **284** 6337–6347.

[138] Mager T, Braner M, Kubsch B, Hatahet L, Alkoby D, Rimon A, Padan E and Fendler K, 2013 Differential Effects of Mutations on the Transport Properties of the Na+/H+ Antiporter NhaA from Escherichia coli. *Journal of Biological Chemistry* **288** 24666–24675.

[139] Padan E, Kozachkov L, Herz K and Rimon A, 2009 NhaA crystal structure: functional-structural insights. *Journal of Experimental Biology* **212** 1593–1603.

[140] Landau M, Herz K, Padan E and Ben-Tal N, 2007 Model Structure of the Na+/H+ Exchanger 1 (NHE1): Functional and Clinical Implications. *Journal of Biological Chemistry* **282** 37854–37863.

[141] Schushan M, Xiang M, Bogomiakov P, Padan E, Rao R and Ben-Tal N, 2010 Model-Guided Mutagenesis Drives Functional Studies of Human NHA2, Implicated in Hypertension. *Journal of Molecular Biology* **396** 1181–1196.

[142] Shaffer P L, Goehring A, Shankaranarayanan A and Gouaux E, 2009 Structure and mechanism of a Na+ independent amino acid transporter. *Science* **325** 1010–1014.

[143] Efremov R G and Sazanov L A, 2011 Structure of the membrane domain of respiratory complex I. *Nature* **476** 414–420.

[144] Harms M J, Schlessman J L, Sue G R and E B G M, 2011 Arginine residues at internal positions in a protein are always charged. *Proceedings of the National Academy of Sciences* **108** 18954–18959.

[145] Isom D G, Castañeda C A, Cannon B R and García-Moreno B, 2011 Large shifts in pK_a values of lysine residues buried inside a protein. *Proceedings of the National Academy of Sciences* **108** 5260–5265.

[146] Verma V, Bali A, Singh N and Jaggi A S, 2015 Implications of sodium hydrogen exchangers in various brain diseases. *Journal of Basic and Clinical Physiology and Pharmacology* **26** 417–426.

[147] Wöhlert D, Kühlbrandt W and Yildiz Ö, 2014 Structure and substrate ion binding in the sodium/proton antiporter PaNhaP. *eLife* **3** e03579.

[148] Rimon A, Kozachkov-Magrisso L and Padan E, 2012 The unwound portion dividing helix IV of NhaA undergoes a conformational change at physiological pH and lines the cation passage. *Biochemistry* **51** 9560–9569.

[149] Akyuz N, Georgieva E R, Zhou Z, Stolzenberg S, Cuendet M A, Khelashvili G, Altman R B, Terry D S, Freed J H, Weinstein H, Boudker O and Blanchard S C, 2015 Transport domain unlocking sets the uptake rate of an aspartate transporter. *Nature* **518** 68–73.

[150] Călinescu O, Paulino C, Kühlbrandt W and Fendler K, 2014 Keeping It Simple, Transport Mechanism and pH Regulation in Na+/H+ Exchangers. *Journal of Biological Chemistry* **289** 13168–13176.

[151] Călinescu O, Danner E, Böhm M, Hunte C and Fendler K, 2014 Species differences in bacterial NhaA Na+/H+ exchangers. *FEBS Letters* **588** 3111–3116.

[152] Cao Y, Jin X, Levin E J, Huang H, Zong Y, Quick M, Weng J, Pan Y, Love J, Punta M, Brukhard R, Hendrickson W A, Javitch J A, Rajashankar K R and Zhou M, 2011 Crystal structure of a phosphorylation-coupled saccharide transporter. *Nature* **473** 50–54.

[153] Luo P, Yu X, Wang W, Fan S, Li X and Wang J, 2015 Crystal structure of a phosphorylation-coupled vitamin C transporter. *Nature Structural & Molecular Biology* **22** 238–241.

[154] Johnson Z L, Cheong C G and Lee S Y, 2012 Crystal structure of a concentrative nucleoside transporter from Vibrio cholerae at 2.4 Å. *Nature* **483** 489–493.

[155] Bolla J R, Su C C, Delmar J A, Radhakrishnan A, Kumar N, Chou T H, Long F, Rajashankar K R and Edward W Y, 2015 Crystal structure of the Alcanivorax borkumensis YdaH transporter reveals an unusual topology. *Nature Communications* **6** 6874.

[156] Su C C, Bolla J R, Kumar N, Radhakrishnan A, Long F, Delmar J A, Chou T H, Rajashankar K R, Shafer W M and Edward W Y, 2015 Structure and function of Neisseria gonorrhoeae MtrF illuminates a class of antimetabolite efflux pumps. *Cell Reports* **11** 61–70.

[157] Zhou X, Levin E J, Pan Y, McCoy J G, Sharma R, Kloss B, Bruni R, Quick M and Zhou M, 2014 Structural basis of the alternating-access mechanism in a bile acid transporter. *Nature* **505** 569–573.

[158] Paulino C and Kühlbrandt W, 2014 pH- and sodium-induced changes in a sodium/proton antiporter. *eLife* **3**, 00000.

[159] Yernool D, Boudker O, Jin Y and Gouaux E, 2004 Structure of a glutamate transporter homologue from Pyrococcus horikoshii. *Nature* **431** 811–818.

[160] Kozachkov L and Padan E, 2011 Site-directed tryptophan fluorescence reveals two essential conformational changes in the Na+/H+ antiporter NhaA. *Proceedings of the National Academy of Sciences* **108** 15769–15774.

[161] Evans P R and Murshudov G N, 2013 How good are my data and what is the resolution? *Acta Crystallographica Section D: Biological Crystallography* **69** 1204–1214.

[162] Davis I W, Murray L W, Richardson J S and Richardson D C, 2004 MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Research* **32** W615–W619.

[163] Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T and Ben-Tal N, 2005 ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Research* **33** W299–W302.

[164] Piggot T J, Aǵel P and Khalid S, 2012 Molecular dynamics simulations of phosphatidylcholine membranes: a comparative force field study. *Journal of Chemical Theory and Computation* **8** 4593–4609.

[165] Lomize M A, Lomize A L, Pogozheva I D and Mosberg H I, 2006 OPM: orientations of proteins in membranes database. *Bioinformatics* **22** 623–625.

[166] Mobley D L, Chodera J D and Dill K A, 2006 On the use of orientational restraints and symmetry corrections in alchemical free energy calculations. *The Journal of Chemical Physics* **125** 084902.

[167] Boresch S, Tettinger F, Leitgeb M and Karplus M, 2003 Absolute Binding Free Energies: A Quantitative Approach for Their Calculation. *The Journal of Physical Chemistry B* **107** 9535–9551.

[168] Christ C D, Mark A E and van Gunsteren W F, 2010 Basic ingredients of free energy calculations: A review. *Journal of Computational Chemistry* **31** 1569–1582.

[169] Bhattachayya A, 1943 On a measure of divergence between two statistical population defined by their population distributions. *Bulletin Calcutta Mathematical Society* **35** 99–109.

[170] Michel J and Essex J W, 2010 Prediction of protein–ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. *Journal of Computer-Aided Molecular Design* **24** 639–658.

[171] Shirts M R and Pande V S, 2005 Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration. *The Journal of Chemical Physics* **122** 144107.

[172] Shirts M R and Chodera J D, 2008 Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of Chemical Physics* **129** 124105.

[173] Jain A, Ong S P, Chen W, Medasani B, Qu X, Kocher M, Brafman M, Petretto G, Rignanese G M, Hautier G, Gunter D and Persson K A, 2015 FireWorks: a dynamic workflow system designed for high-throughput applications. *Concurrency and Computation: Practice and Experience* **27** 5037–5059.

[174] Hummer G, Pratt L R and García A E, 1996 Free Energy of Ionic Hydration. *The Journal of Physical Chemistry* **100** 1206–1215.

[175] Marcus Y, 1991 Thermodynamics of solvation of ions, Part 5. *Journal of the Chemical Society, Faraday Transactions* **87** 2995–2999.

[176] Friedman H and Krishnan C, 1973, In water, a comprehensive treatise.

[177] Conway B, 1978 The evaluation and use of properties of individual ions in slution. *Journal of Solution Chemistry* **7** 721–770.

[178] Lim N M, Wang L, Abel R and Mobley D L, 2016 Sensitivity in Binding Free Energies Due to Protein Reorganization. *Journal of Chemical Theory and Computation* .

[179] Rocklin G J, Mobley D L, Dill K A and Hünenberger P H, 2013 Calculating the binding free energies of charged species based on explicit-solvent simulations employing lattice-sum methods: An accurate correction scheme for electrostatic finite-size effects. *The Journal of Chemical Physics* **139** 184103.

[180] Tanford C, 1983 Translocation pathway in the catalysis of active transport. *Proceedings of the National Academy of Sciences* **80** 3701–3705.

[181] Lee M S, Salsbury F R and Brooks C L, 2004 Constant-pH molecular dynamics using continuous titration coordinates. *Proteins: Structure, Function, and Bioinformatics* **56** 738–752.

[182] Prlić A and Procter J B, 2012 Ten Simple Rules for the Open Development of Scientific Software. *PLoS Comput Biol* **8**.

[183] David L. Dotson, Sean L. Seyler, Max Linke, Richard J. Gowers and Oliver Beckstein, 2016 datreant: persistent, Pythonic trees for heterogeneous data, in *Proceedings of the 15th Python in Science Conference* (edited by Sebastian Benthall and Scott Rostrup), 51 – 56.

[184] Richard J. Gowers, Max Linke, Jonathan Barnoud, Tyler J. E. Reddy, Manuel N. Melo, Sean L. Seyler, Jan Domański, David L. Dotson, Sébastien Buchoux, Ian M. Kenney and Oliver Beckstein, 2016 MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations, in *Proceedings of the 15th Python in Science Conference* (edited by Sebastian Benthall and Scott Rostrup), 98 – 105.

[185] Goecks J, Nekrutenko A and Taylor J, 2010 Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* **11** R86.

[186] McKinney W, 2010 Data Structures for Statistical Computing in Python , in *Proceedings of the 9th Python in Science Conference* (edited by S van der Walt and J Millman), 51 – 56.

[187] van der Walt S, Colbert S C and Varoquaux G, 2011 The numpy array: A structure for efficient numerical computation. *Computing in Science Engineering* **13** 22–30.

[188] Matthew Rocklin, 2015 Dask: Parallel Computation with Blocked algorithms and Task Scheduling, in *Proceedings of the 14th Python in Science Conference* (edited by Kathryn Huff and James Bergstra), 130 – 136.

[189] Olkhova E, Padan E and Michel H, 2007 The Influence of Protonation States on the Dynamics of the NhaA Antiporter from Escherichia coli. *Biophysical Journal* **92** 3784–3791.