Unidirectional Single-File Transport of Full-Length Proteins Through a Nanopore

Luning Yu,[1] Xinqi Kang,[2] Fanjun Li,[4] Behzad Mehrafrooz,[6] Amr Makhamreh,[2] Ali Fallahi,[2] Aleksei Aksimentiev,[5] Min Chen,[4] Meni Wanunu[1, 2, 3]*

Department of [1]Physics, [2]Bioengineering and [3]Chemistry and Chemical Biology, Northeastern University, Boston, Massachusetts 02115, United States

[4]Department of Chemistry, University of Massachusetts at Amherst, Amherst, Massachusetts 01003, United States

[5]Department of Physics and [6]Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States

*E-mail: wanunu@neu.edu

Abstract:

Nanopore technology offers long, accurate sequencing of an DNA or RNA strand via enzymatic ratcheting of the strand through a nanopore in single nucleotide steps, producing stepwise modulations of the nanopore ion current. In contrast to nucleic acids, their daughter molecules, proteins, have neutral peptide backbones and side chains of varying charges. Further, proteins have stable secondary and higher order structures that obstruct protein linearization required for single file nanopore transport. Here, we describe a general approach for realizing unidirectional transport of proteins through a nanopore that neither requires the protein to be uniformly charged nor a pull from a biological enzyme. At high concentrations of guanidinium chloride, we find full-length proteins to translocate unidirectionally through an α-hemolysin nanopore in a polymer-based membrane, provided that one of the protein ends is decorated with a short anionic peptide. Molecular dynamics simulations show that such surprisingly steady protein transport is driven by a giant electro-osmotic effect caused by binding of guanidinium cations to the inner surface of the nanopore. We show that ionic current signals produced by protein passage can be used to distinguish two biological proteins and the global orientation of the same protein (N-to-C vs. C-to-N terminus) during the nanopore transport. With the average transport rate of one amino acid per 10 μs, our method may enable direct enzyme-free protein fingerprinting or perhaps even sequencing when combined with a high-speed nanopore reader instrument.

Introduction

Living systems use complex molecules to encode and carry out their essential functions, with nucleic acids being a central information carrier across generations and within species. A tremendous level of scientific and technological efforts has been made to develop new methods for high-throughput and long-read genomic sequencing[1]. In particular, efforts over the past 20 years have led to major breakthroughs in single-molecule sequencing[2], in which individual DNA is sequenced by either real-time replication monitoring[3] or by direct readout of a strand by passing it through a detector[4]. Single-molecule methods offer higher-quality sequence mapping[5-9], uncovering "dark genomic matter"[10] that is undetectable with conventional sequencing methods, resolving pathological short tandem repeats[11,12], and elucidating genetic variants and epigenetic modifications[13-15].

Although DNA provides a blueprint for the types of molecules that a cell can produce, ultimately, cell phenotypes are difficult to predict from genomic information alone. Transcriptome analysis provides insight into the types of mRNA molecules that the cell produces, although ultimately, the proteome, which is defined as the types and quantities of cellular proteins that are produced, defines a cell's phenotype. There are $>10^4$ types of proteins expressed in a human cell, although proteome complexity extends far beyond the canonical exome: various isoforms[16,17] and post-translational modifications (PTMs)[18,19] flavor a cell's phenotype and determine its functionality. A molecular understanding of the roles of these diverse phenotypes requires quantitative methods for protein counting, sequencing, and discrimination among various isoforms and PTMs. In mass spectrometry (MS), the most common bottom-up approach involves fragmenting proteins to smaller peptides, quantifying them, then reconstructing a model proteome from those fragment identities. The high sensitivity of MS to minute quantities of protein permits single-cell proteomics[20], although MS falls short of providing a complete view of proteomes because low peptide ionization rates and other limitations result in only a few percent sampling efficiencies[21,22]. This unmet need has summoned alternative approaches that can deliver a more complete single-cell proteome from a sample of interest without extensive fragmentation. The search for single-molecule signatures that report on the identity of a peptide and/or protein has been vibrant of late[23-26]. As summarized in recent reviews[27,28], approaches for single-molecule protein sequencing include pushing the limits of MS, single-molecule fluorescence, electronic detection, and a combination of these detection methods.

Nanopores have recently gained substantial recognition in genomics by demonstrating ultralong-read DNA and RNA sequencing. The basic principle of nanopore sensing involves passing a molecule through a nanometer-scale pore in an impermeable membrane. Applying voltage across the membrane surrounded by an electrolyte solution produces a steady current of ionic species across the membrane, through the nanopore, which provides a baseline current signal that is monitored using an amplifier. When a DNA or RNA strand is threaded into the nanopore, the ion current through the pore is reduced by an amount related to the sequence of the nucleic acid fragment residing at the pore constriction. Since DNA and RNA molecules have a uniformly charged phosphate backbone, the application of voltage across the pore results in a nearly constant force[29] that pulls the DNA strand through the pore. Enzyme-mediated regulation[30] of strand motion ensures base-by-base feeding of the strand into the pore constriction in ~ms-intervals, during which the strand is taut in the pore due to the electric force acting on it. Attempts to directly read DNA or RNA sequence during voltage-mediated translocation (i.e., enzyme-free)

failed because DNA translocation speeds (0.1-1 μs/nucleotide)[31] are too fast to properly register a stable current level, even with very high bandwidth measurements in solid-state nanopores[32-35].

Adapting nanopore sequencing principle to protein sequencing is challenging for several reasons: First, proteins are composed of a chemically more diverse set of building blocks – while DNA and RNA are composed of only 4 nucleotides (excluding modifications), proteins are made up of at least 20 types of amino acids (aa's), and frequently some of these aa's contain PTMs. Second, proteins cannot be chemically replicated in the same way as DNA or RNA molecules are replicated using polymerases and/or reverse transcriptases. Third, proteins fold into stable and complex tertiary and quaternary structures that are stabilized by multiple types of attractive forces that obstruct protein linearization, a prerequisite to being fed through a pore single file. Finally, in contrast to the uniformly charged nucleic acids, the peptide backbone of a protein is not charged, which means that the transmembrane voltage does not necessarily generates a force large enough to pull the strand taut in the pore for the ionic current readout.

Various approaches for protein translocation through narrow nanopores have been described, with "narrow" being defined as a pore with too small of a constriction to allow transport of a folded peptide or a folded protein (<1.5 nm diameter). Enzyme-based unfolding and translocation of large proteins has been demonstrated using ClpXP as a motor to unfold and to pull the protein through α-hemolysin[36,37]. Genetically-encoding proteins in an organism with a charged tail and an ssrA domain for ClpXP recognition allows multiplexed quantification of expressed proteins[38]. More recently, peptide−DNA conjugates were fed through a pore *via* ratcheting or unwinding through the MspA nanopore using a phi29 DNA polymerase[39] or a Hel308 DNA helicase[40], respectively. Various enzyme-free approaches in which voltage-induced[41-44] or temperature-induced[45] unfolding have been demonstrated, although most commonly, chemical denaturation has been the method of choice. Unfolding using sodium dodecyl sulfate (SDS)[24] and guanidinium chloride (GdmCl)[46] allowed protein detection using narrow solid-state pores, and both urea and GdmCl have been used to study protein denaturation using α-hemolysin and aerolysin pores[47-50]. These studies have shown that the frequency of detected events sharply increases as denaturant concentrations approach the unfolding transition, which suggests enhanced access of the unfolded protein to the nanopore constriction. Some evidence of protein translocations through pores was gathered by preparing a DNA-protein chimera and using real-time PCR to amplify the DNA portion of the chimera from the trans chamber[50]. These studies also reported pronounced variability of the dwell time and current blockade data, suggesting strong protein/pore interactions that result in stick/slip motion, presumably due to an insufficient pulling force to linearize residual protein secondary structure.

Here we have developed a stable platform for linearization and steady transport of full-length proteins through a nanopore reader. Our platform uniquely combines a robust geometric aperture that supports high voltage measurements with a bilayer polymer-based membrane that withstands high denaturant concentrations. Incorporating into the bilayer an α-hemolysin nanopore, which maintains its integrity at high GdmCl concentrations[51], enables steady voltage-driven translocation of proteins through the pore. When a negatively charged tail is present on a protein molecule, end-threading into the pore (from either the N- or C-terminus) is facilitated, but the high GdmCl concentration is essential for the transport of proteins with a constant rate. Molecular dynamics simulations find GdmCl to promote electroosmotic flow directed toward the positive electrode, which drives protein transport. Measured mean protein velocities are independent of the orientation of entry (N- vs. C-terminus) and are slow (~10 μs/amino acid) as

compared to typical nucleic acid transport times (~1 μs/nucleotide[52]). Machine-learning tools can easily classify a protein molecule in a binary mixture on-the-fly with >94% classification accuracy. Finally, we show that N-terminus and C-terminus transport of the same protein leads to opposite signal trends. Our approach paves the way for single-molecule protein readout using a range of nanopores.

Results and Discussion:

Our experimental setup is designed to transport unfolded protein molecules through a nanopore reader. In **Figure 1A** we depict a cut-away view through a flow cell that contains top (*cis*) and bottom (*trans*) buffer reservoirs, each equipped with an electrode. In the Polytetrafluoroethylene (PTFE)-based fluidic cell is affixed a silicon (Si) chip that contains a central opening, atop which a wedge-on-pillar (WOP) membrane support is housed. The WOP support chip comprises a 20-μm-thick SU-8 photoresist layer that houses a 100 μm diameter central aperture (details were published elsewhere[53]), and is designed to provide a stable membrane support for long-duration (~15 hours) experiments[51]. The membranes are made by pipetting a lipid or polymer solution in organic solvent over the aperture, followed by repetitive bubbling of air over the film to thin the membrane down to a bilayer, which is determined by monitoring the membrane's electrical capacitance. A side cross-section of the WOP aperture (**Figure 1B**) shows a poly(1,2-butadiene)-b-poly(ethylene oxide) (PBD_n-PEO_m) block-copolymer bilayer that spans the aperture, as well as the inserted α-hemolysin channel, which is the nanopore used in our experiments. Also depicted is our use of high GdmCl concentrations to mediate protein unfolding and translocation. Current vs. voltage curves for single α-hemolysin channels at different buffer conditions used in this study are shown in **Figure 1C**. All curves exhibit significant asymmetry, with higher current amplitudes at positive voltages. The impact of GdmCl on noise in α-hemolysin is moderate: The 10 kHz bandwidth noise values at 300 mV are 7.2 pA for 2.5 M KCl and 10.7 pA for 1 M KCl + 2.0 M GdmCl, respectively. Power spectra at 0 mV and 300 mV for these two buffers (see inset) reveal a small increase in the noise in the low-to-intermediate frequency regime (<5 kHz), possibly due to Gdm^+ binding to the pore. The proteins used in our experiments are summarized in **Figure 1D**, which shows the net unfolded protein charge (at pH 7.5)[54] and length (number of aa's) of each protein, as well as pKa-based graphical profiles of the charge[54] and relative volume[55] of each aa residue. We have chosen variants of maltose-binding protein (MBP) in its monomeric form (denoted as either MBP-D10 or D10-MBP depending on the C- or N-terminus attachment of the D_{10} tail) and its dimeric form (diMBP-D10, with a GGSG linker between two MBP monomers). As example of another protein, we used green-fluorescent protein (GFP), which we chose in part because its stable β-barrel structure makes this protein notoriously difficult to unfold. Anionic D_{10} aspartate tails were added to the proteins in order to direct protein capture, as per an earlier report[43]. The impact of the tail on protein threading is evident in **Figure 1E** and **1F**, where current traces for wild-type MBP (WT-MBP) and MBP-D10 at 175 mV and 1 M KCl, 2 M GdmCl buffer are shown. While in both experiments protein concentration was 350 nM, the capture rates for MBP-D10 (9.3 $s^{-1}\mu M^{-1}$) were ~80% higher than for WT-MBP (5.2 $s^{-1}\mu M^{-1}$), and further, WT-MBP events exhibited a broad range of amplitudes and fast translocations, whereas MBP-D10 events were deep and with most events being uniform in duration, with nearly 79% of the events forming a tight distribution characterized by a ~85% fractional blockade and a dwell time between 1 to 10 ms. Such reproducible dwell times and current amplitudes suggests an efficient and deterministic translocation process, mediated by the insertion of the D_{10} tail.

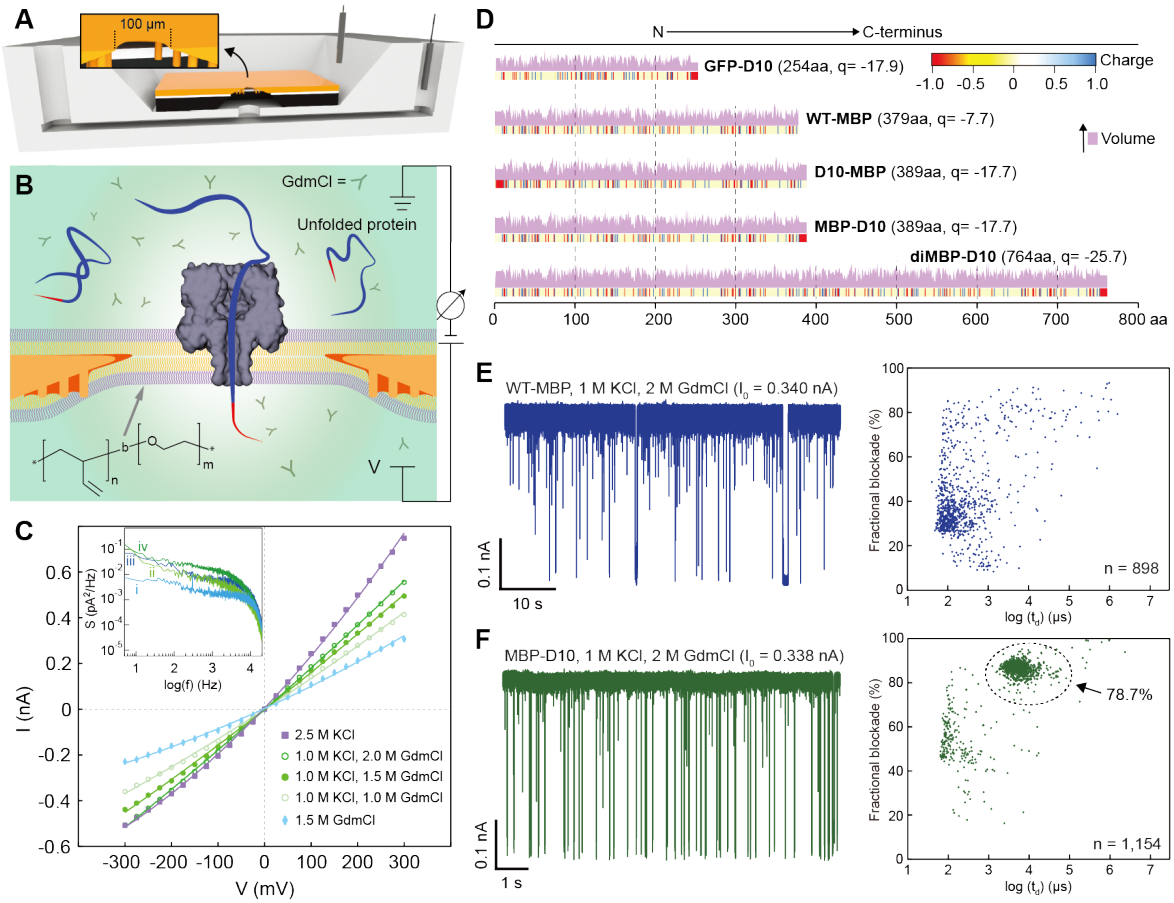

Figure 1. Enzyme-free translocation of full-length proteins through nanopores. **A)** Cartoon model of our fluidic cell that houses a Si-chip (black) upon which an SU-8 wedge-on-pillar (WOP) aperture is suspended[53]. **B)** Schematic side view of a suspended PBD_n-PEO_m block-copolymer bilayer (inset shows polymer structure) suspended on the WOP aperture (orange depicts lipid solvent), with a α-hemolysin channel inserted into the bilayer. Guanidinium chloride (GdmCl) present in the buffer unfolds protein molecules. Ag/AgCl electrodes placed in the top (*cis*) and bottom (*trans*) chambers enable trans-pore ion current measurements using an amplifier. **C)** Current-voltage measurements for a single α-hemolysin channel under different buffer conditions (V is applied on *trans* chamber). **D)** Graphical representations of the formal charges at pH 7.5 for the different protein constructs used in this study in their unfolded state (solvent accessible). Above each charge graph is a plot that represents the relative amino acid volumes (pink). **E)** Current vs. time trace (left) and scatter plot (right) showing fractional blockade vs. dwell time for wild-type maltose binding protein (WT-MBP), obtained at V = 175 mV, [WT-MBP] = 0.35 µM. **F)** Current vs. time trace (left) and scatter plot (right) showing fractional blockade vs. dwell time for MBP with a C-terminus aspartate tail (MBP-D10), obtained at V = 175 mV, [MBP-D10] = 0.35 µM. Open pore current values (I_o) are indicated next to each trace. Fraction of detectable events in dashed circle on scatter plot is indicated.

Influence of GdmCl on complete protein unfolding. Since proteins have complex secondary structures, often with extensive intramolecular interactions, a pre-requisite for single-file protein translocation is neutralization of intramolecular interactions. Bulk measurements suggest that MBP has an unfolding midpoint at 1.0 M GdmCl at room temperature, and further, that MBP is fully denatured in ~1.2 M GdmCl at room temperature[56,57]. In **Figure 2A-C**, we present current

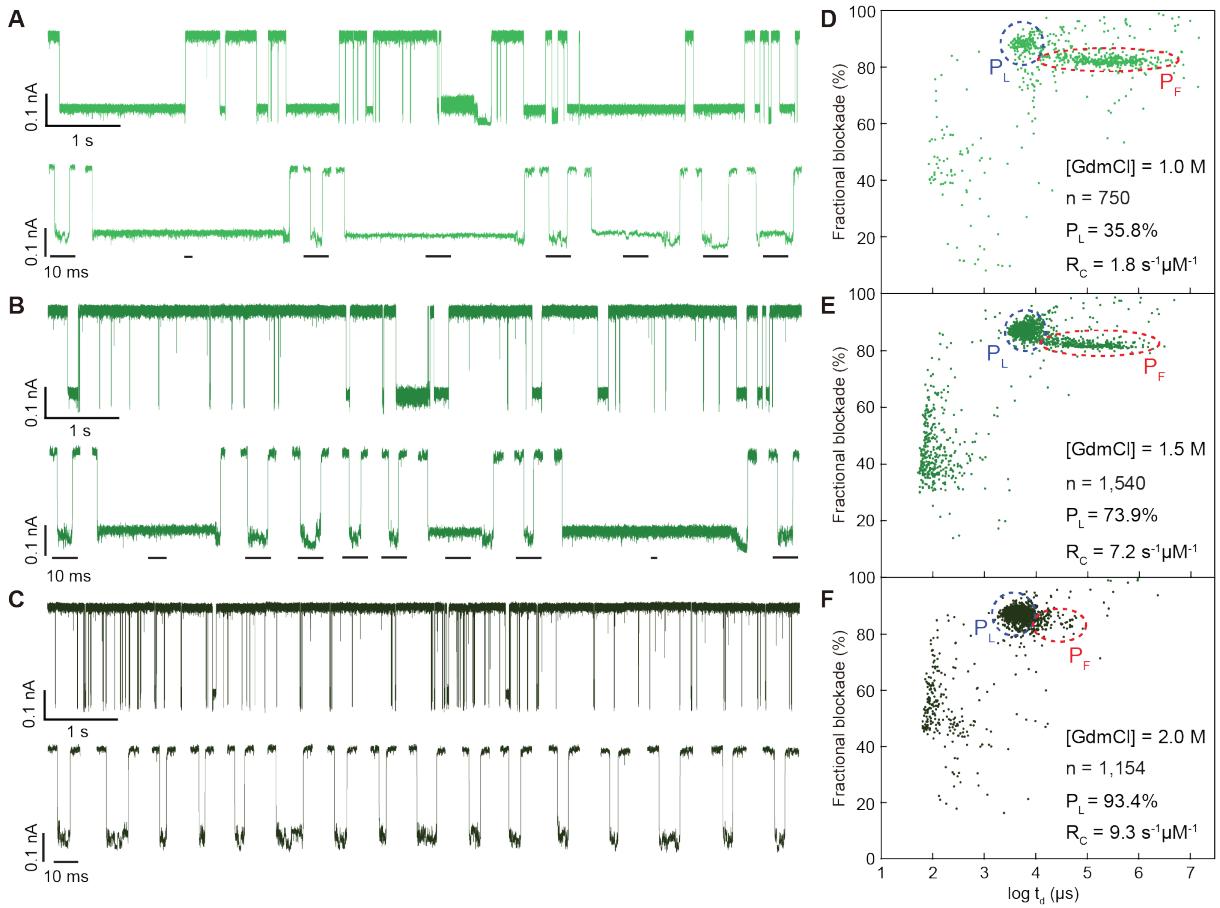

Figure 2. Effect of [GdmCl] on MBP-D10 interactions with α-hemolysin. **A-C)** Continuous current vs. time traces for MBP-D10 in 1.0 M, 1.5 M and 2.0 M GdmCl, with the open pore currents I_0 = 0.260, 0.308, 0.352 nA, respectively. Under each continuous trace we show random events selected at 10 s intervals. All time-scale bars for the selected events are 10 ms. **D-F)** Fractional current blockade vs. dwell time scatter plots for MBP-D10 in 1.0 M, 1.5 M and 2.0 M GdmCl buffer (+1 M KCl). Dashed red and blue ovals show populations that correspond to P_F (partially folded) and P_L (linear, or unfolded) states of MBP-D10, respectively. [MBP-D10] = 0.7 μM for 1.0 M GdmCl and 0.35 μM for 1.5, 2.0 M GdmCl experiments.

recordings of a α-hemolysin channel after the addition of MBP-D10 to the *cis* chamber in 1.0 M, 1.5 M, and 2.0 M GdmCl, respectively. Two populations comprising relatively short (few-ms) and very long (up to 1 s) events are seen in the traces. Further, as the GdmCl concentration increases, the occurrence of long events disappears, and the events are very uniform in duration for the 2.0 M GdmCl case (see **Figure 2C**). **Figure 2D-F** shows scatter plots of the fractional current blockades and the dwell times for the three corresponding experimental conditions shown in **A-C**. In accordance with the traces, the scatter plots reveal two main distributions, highlighted with dashed red and blue dashed circles, in addition to a very "fast" population at ~100 μs which most likely corresponds to collisions of protein molecules with the pore entrance. We attribute the long-spread, long-lived events (red dashed circle) as population P_F, which contains events in which the protein is partly folded, and we ascribe the tight, shorter-lived population P_L (blue dashed circle) to completely linear (or unfolded) proteins, for reasons that are described in the next paragraph. Since these two populations are generally well-resolved, we have quantified the percentage of linear protein events P_L for experiments in various GdmCl concentrations and at

different voltages, where P_L is defined as the fraction $P_L = {}^{\#}P_L/({}^{\#}P_F+{}^{\#}P_L)$, ${}^{\#}P_X$ is the number of events in population X (see SI, **Figure S12**). As seen in **Figure 2D-F**, P_L increases with the GdmCl concentration from ~36% at 1.0 M GdmCl to >93% at 2.0 M GdmCl. The existence of one population at 2.0 M GdmCl suggests that the protein has only one form, and that most likely form, at this high denaturant concentration is the unfolded form.

Evidence of voltage-driven protein transport. Despite the noticeable impact of GdmCl on protein denaturation, concluding that populations P_F and P_L represent partly folded and fully linear proteins demands further evidence. In **Figure 3A, B**, we present dwell time distributions (log-scale) at increasing voltages in the range of 75-300 for MBP-D10 and diMBP-10, respectively, at a denaturant concentration of 1.5 M GdmCl. In the 75-175 mV voltage range, a significant fraction

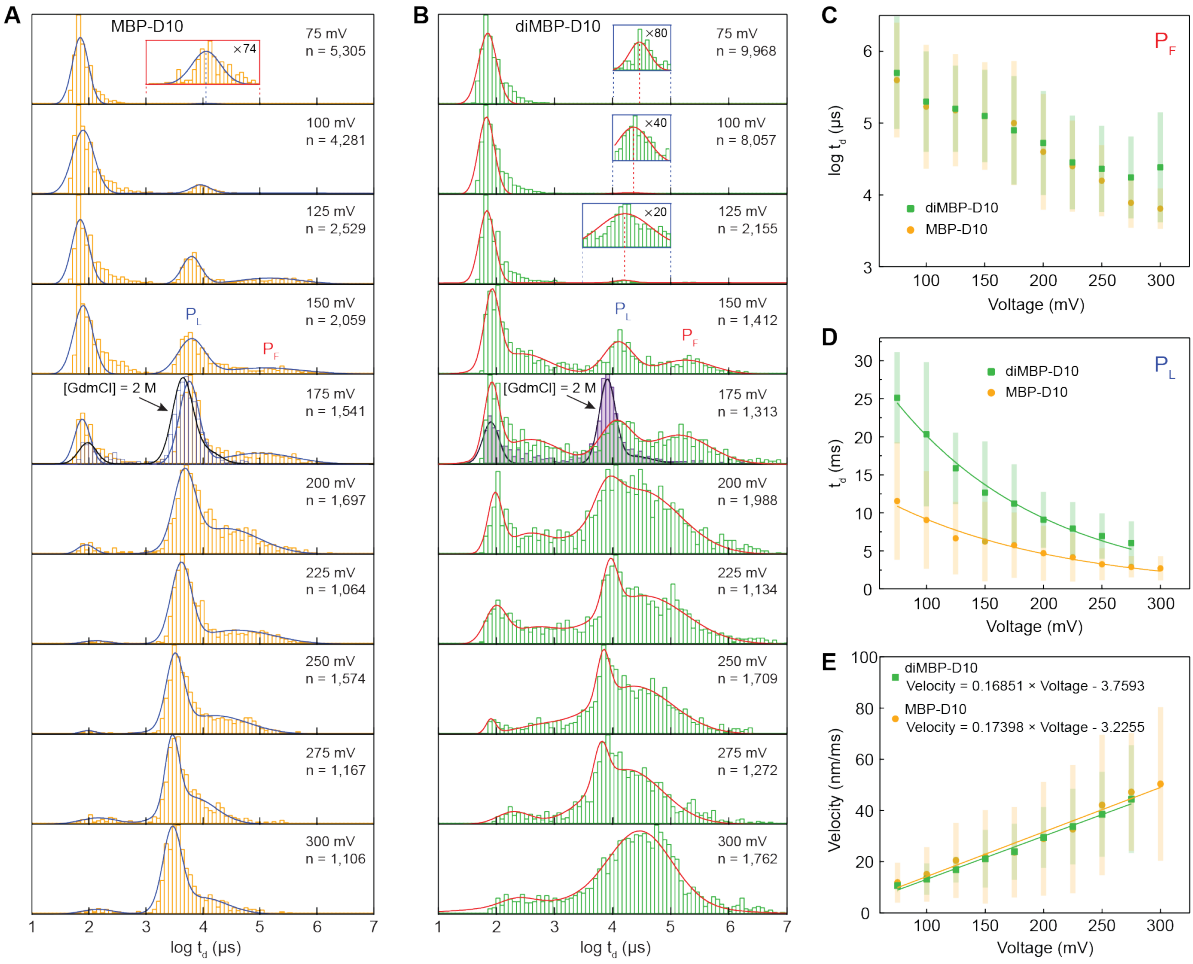

Figure 3. Translocation dynamics of monomer and dimer MBP proteins. **A)** Dwell time histograms of MBP-D10 (0.35 μM) at 75-300 mV and log-normal fits (1.5 M GdmCl, 1 M KCl), **B)** Dwell time histograms of diMBP-D10 (0.35 μM) at 75-300 mV and log-normal fits (1.5 M GdmCl, 1 M KCl). For **A** and **B**, the additional distributions at 175 mV (blue, purple) correspond to MBP-D10 (n=1155) and diMBP-D10 (n=1151) at 2 M GdmCl, 1 M KCl. **C, D)** Mean dwell time vs. voltage for the P_F and P_L populations of MBP-D10 and diMBP-D10, respectively (error bars represent the FWHM of the distribution fits). In **D**, curves are exponential fits to the data. **E)** Protein transport velocities calculated from estimated protein contour length and observed dwell times as a function of applied voltage (error bars are based on the dwell time distribution widths, shown in **panel D**).

of events have ~100 μs dwell times, and as voltage increases further, a greater fraction of the events have longer dwell times in the range of 1-1000 ms, concentrated in two populations that correspond to P_F and P_L. Strikingly, the position of the P_L population shifts monotonically toward faster dwell times as voltage is increased for both monomeric and dimeric MBP (top to bottom, in panels **A** and **B**, respectively). To highlight the impact of higher GdmCl concentrations on complete protein unfolding, we show dwell time distributions at 175 mV for 2 M GdmCl for both monomeric and dimeric MBP (blue and purple histograms, respectively). In **Figure 3C, D** we present mean dwell times for the P_F and P_L populations as a function of voltage, based on the data in **A**, **B** (error bars represent the widths of the distribution fits). We find that the P_F population has long dwell times for both MBP-D10 and diMBP-10 at all voltages, however, that these dwell times decrease with increasing voltage from ~1 s to ~10 ms, consistent with translocation. Similarly, dwell times for population P_L decrease with increasing voltage. However, the trend in dwell times (shown in linear scale) appears milder and more regular than the trend for P_F, and remarkably, dwell times for the dimeric MBP are about a factor of 2 longer than for monomeric MBP. In **Figure 3E,** a related protein "velocity" plot calculated by dividing the contour lengths (0.34 nm per amino acid in MBP-D10 and diMBP-D10) by the dwell times (from **panel D**) of the proteins, shows that velocity is linearly dependent on voltage for both monomeric and dimeric MBP, but interestingly, independent of the protein length. These striking results demonstrate that protein translocation proceeds with a constant speed and further, that the speed is proportional to voltage, similar to ssDNA translocation through α-hemolysin[58-60]. Finally, a plot of the event rates for these experiments (See **SI, Figure S13**) reveals a low-voltage regime characterized by high frequency short-lived collisions and an exponentially increasing capture rate at higher voltages (V > 150 mV), which suggests an entropic barrier for capture[61].

Electroosmotic flow drives protein transport. To identify the molecular mechanism enabling voltage-driven transport of unfolded proteins in the presence of Gdm^+, we built several all-atom models of the experimental systems each containing a α-hemolysin nanopore, a patch of a lipid membrane, and electrolyte solution (see **Figure 4A**). Given that the transmembrane transport through α-hemolysin was previously found to be rather insensitive to the type of the lipid bilayer membrane it was embedded into[62], we did not attempt to build an all-atom model of the polymer membrane used in our experiments and instead used a model of a well-studied synthetic lipid bilayer. Each system was first equilibrated using the all-atom MD method[63] and then simulated under a 200 mV transmembrane bias of either polarity. **Figure 4B** illustrates an outcome of such simulation for the 1.5 M GdmCl/1 M KCl electrolyte system at +200 mV bias. The ionic current through the nanopore (the slope of the charge versus time plot) is found to be carried by Cl^- ions, whereas the current carried by either K^+ or Gdm^+ was close to zero. For the same polarity of the transmembrane bias, strong anion selectivity was also observed in the simulation of the 1.5 M GdmCl systems, whereas in the 2.5M KCl system, the anion selectivity was less extreme (see **Figure 4C** and **SI Figure S15**). Under a reverse bias polarity (-200 mV) the ionic current remained anion-selective for the systems containing Gdm^+, but not for the 2.5 M KCl system, where both K^+ and Cl^- fluxes were of similar magnitude (see **Figure 4C**). Comparing the simulated ionic current to experiments**, SI Table 3**, we find that MD simulations systematically overestimate the ionic current magnitude, which can be largely explained by the differences in the simulated and experimental bulk electrolyte conductivities, **SI Table 4**. Notably, the simulations reproduced the sign and the magnitude of the ionic current rectification ratio for the three electrolyte systems studied.

Consistent with the observed ion selectivity, the α-hemolysin systems were found to exhibit a strong electro-osmotic effect at a positive polarity of the transmembrane bias and, to a lesser degree, at a negative polarity in the presence of GdmCl, **Figure 4D** and **SI Figure S16**. First, we note that the electrolyte conditions were found to have a minor effect on the distribution of the electrostatic potential within the nanopore (**SI Figure S17**), consistent with a prior study[64]. In the case of GdmCl electrolyte, the ion selectivity and the electro-osmotic effect were found to originate from binding of Gdm^+ to the inner nanopore surface. According to **Figure 4E** and **SI Figure S18**, Gdm^+ accumulates in the following three regions: near the *trans* exit of the nanopore stem, above the stem-vestibule junction, and near the *cis* exit from the α-hemolysin vestibule. In the same three regions, individual Gdm^+ were observed to remain bound to the nanopore surface for considerable (> 5 ns) intervals of time (see **Figure 4E**). Gdm^+ binding to the inner nanopore surface was found to mildly correlate with the local charge of the nanopore (see **Figure 4E**). At the same time, the local concentrations of cation and anion within each section of the nanopore volume remained largely similar (**SI Figure S19**), satisfying the local electroneutrality condition.

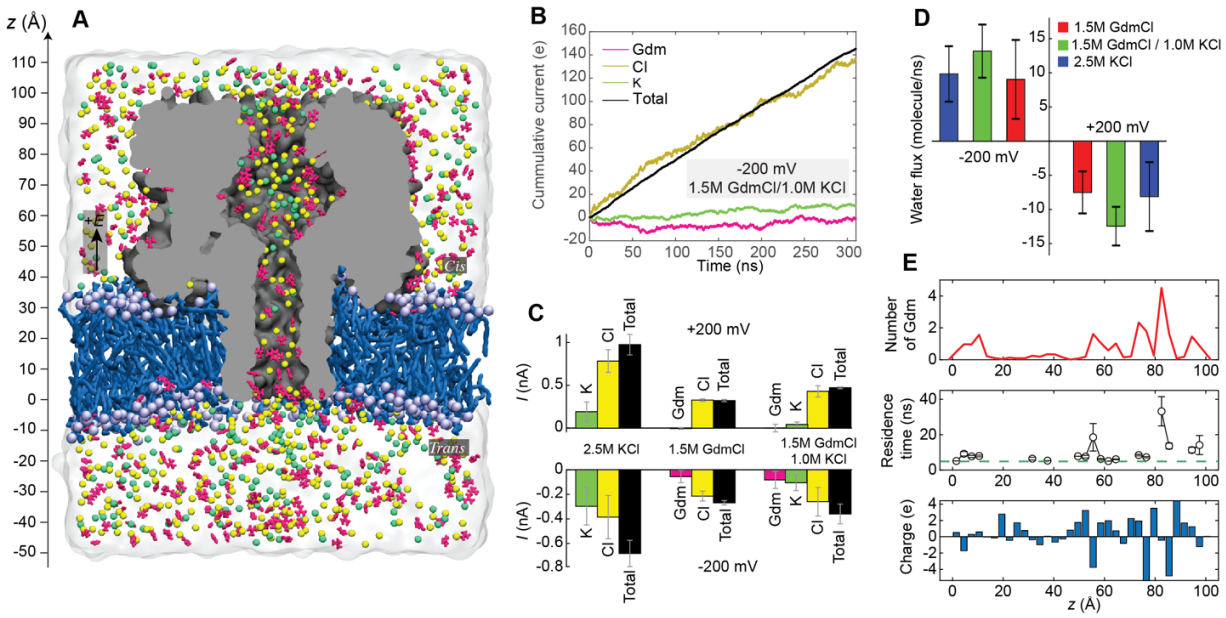

Figure 4. Simulated ion and water flux through α-hemolysin. **A)** All-atom model of α-hemolysin (gray, cut-away molecular surface) embedded in a lipid bilayer membrane (blue) submerged in GdmCl/KCl electrolyte (white transparent surface), with ions shown in pink (Gdm^+), yellow (Cl^-) and green (K^+). **B)** Total charge carried by ion species through the transmembrane pore of α-hemolysin over the course of an MD simulation of the 1.5 M GdmCl/1.0 M KCl system under a +200 mV bias. The slope of each line corresponds to the average current carried. **C, D)** Average ionic current (C) and water flux (D) through α-hemolysin as a function of the transmembrane bias polarity and the electrolyte composition. Negative values indicate transport in the direction opposite that of the *z*-axis. Each average value was computed by splitting the corresponding MD trajectory into 30 ns fragments, finding the average value for each fragment and averaging over the fragments. The error bars show the standard error of the average**. E)** The number of Gdm^+ ions within 3 Å of the nanopore inner surface (top) and the average residence time of the Gdm^+ ions (middle) along the transmembrane pore of α-hemolysin. The *z*-axis is defined in **panel A**. Contacts between Gdm^+ and α-hemolysin lasting less than 5 ns (dashed green line) were discarded from the residence time analysis. The bottom panel shows the local charge of the nanopore inner surface computed, in each 3 Å bin, as the total charge of protein atoms located within 3 Å of water molecules filling the nanopore.

In the mixed electrolyte system (GdmCl/KCl), the local K^+ concentration was considerably (factor 3 or more) lower than that of Gdm^+. Thus, binding of Gdm^+ to the inner nanopore surface rendered the nanopore surface positively charged. That charge is compensated by the more mobile Cl^- ions that carried the majority of the ionic current through the nanopore and produced a strong electro-osmotic effect.

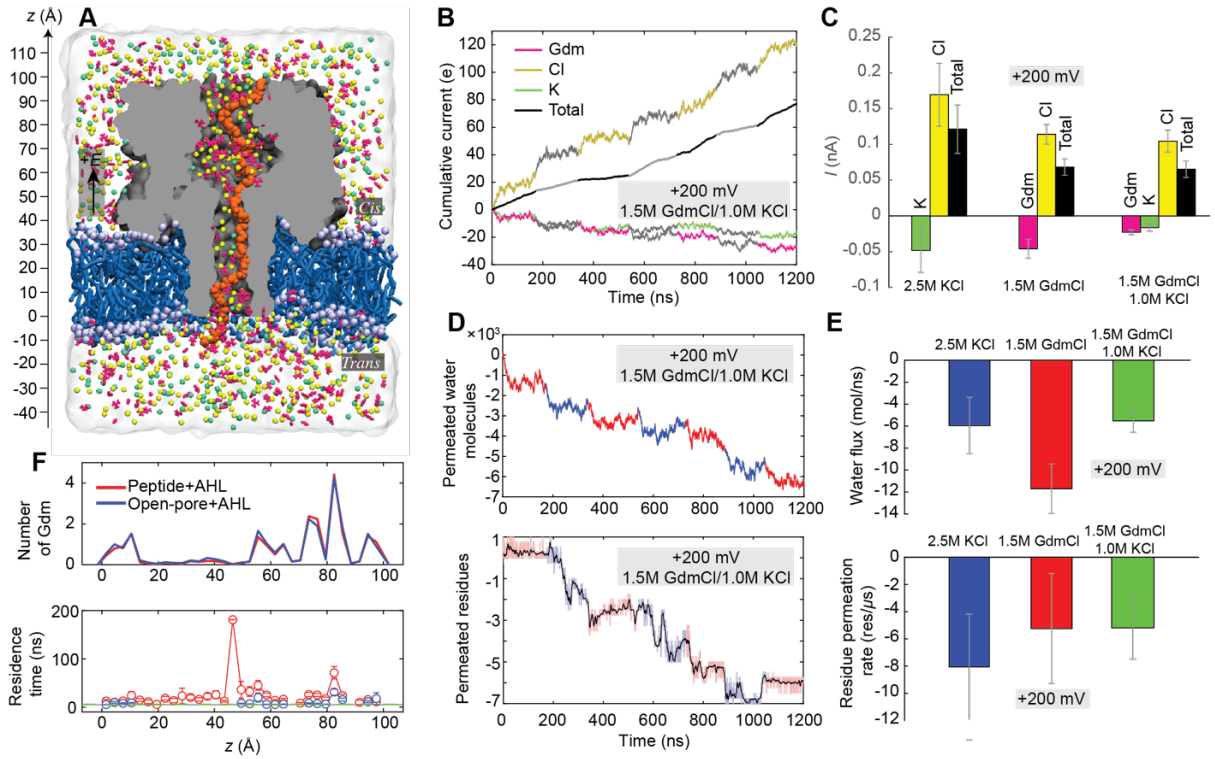

Figure 5. Simulated peptide transport through α-hemolysin. **A)** All-atom model of α-hemolysin (gray) containing a 52 amino acid fragment of the MBP protein (orange), embedded in a lipid membrane (blue) and submerged the GdmCl/KCl electrolyte mixture. **B)** Total charge carried by the ion species through the α-hemolysin nanopore occupied by MBP fragments in seven independent MD simulations carried out under a +200 mV bias. Each trace is shown using two alternating colors to indicate data from the seven MD trajectories differing by the sequence and conformation of the MBP fragment, see **SI Table S5**. The traces were added consecutively to appear as a continuous permeation trace. The slope of each lines indicates the average current carried. **C)** Average ionic current through the α-hemolysin/MBP systems for the three electrolyte conditions at +200 mV bias. The average and the standard error were determined considering the average current from each replica simulation based on an independent measurement. **D)** The number of amino acid residues (top) and water molecules (bottom) permeated through the α-hemolysin constriction (residues 111, 113, and 147) under a +200 mV bias for the 1.5 M GdmCl/1.0 M KCl electrolyte condition. Negative values indicate downward transport in the *z*-direction. Results for the seven independent simulations are shown using alternating colors. **E)** The rate of MBP translocation (top) and the water flux (bottom) averaged over the seven simulations for each ionic condition. **F)** The number of Gdm ions within 3 Å of the nanopore inner surface (top) and the average Gdm^+ residence time (bottom) along the transmembrane pore of α-hemolysin. The *z*-axis is defined in panel A. Contacts between Gdm^+ and α-hemolysin lasting less than 5 ns (dashed green line) were discarded from the residence time analysis. Colors indicate data for open (blue) and MBP-occupied (red) α-hemolysin systems.

Further MD simulations showed that Gdm^+ induced electro-osmotic effect persists even when the nanopore of α-hemolysin is blocked by an unfolded peptide chain. Using the phantom pore method[65], we constructed seven simulation systems for each electrolyte condition, each system containing a different 52-residue fragment of MBP (**SI Table S5**); **Figure 5A** illustrates one such system. Each system was simulated under a +200 mV bias for approximately 150 ns, the simulation trajectories were analyzed with regard to ion, water and peptide transport through the transmembrane nanopore. **Figure 5B** shows that, for the mixed electrolyte conditions, the ionic current selectivity became even stronger than in the equivalent open pore system: the cations were observed to flow in the direction opposite to that prescribed by the electric field. Similar ionic selectivity was observed for all three electrolyte conditions, **Figure 5C** and **SI Figure S20**. Consistent with the ion selectivity data, we observed strong electro-osmotic effects in all three systems, **Figure 5D, E** and **SI Figure S21**.

To determine if the observed electro-osmotic effect can produce unidirectional transport of an unfolded protein through the nanopore, we computed the number of residues translocated through the nanopore constriction as a function of simulation time, **Figure 5D, E** and **SI Figure S22**. For all three systems, we found small but statistically significant displacements of the peptide in the direction of the water flux, with the average rate of 5.2±2.3 amino acid residues per μs, although this estimate must be taken with caution as it also includes the effect of polypeptide chains relaxation from their initially stretched configurations. The simulated peptide translocation rate was found to exhibit insignificant correlation with the electrical charge of the peptide, **SI Figure S23**, although the peptide translocation data were too noisy to disentangle the electrokinetic and electrophoretic contributions to the transport. In the presence of the peptide, the Gdm^+ ions were found to bind to the same spots at the inner surface of α-hemolysin as in the open pore simulations, **Figure 5F** (top) and **SI Figure S24**, however, the binding duration increased dramatically, in particular, at the stem side of the stem-vestibule junction, **Figure 5F** (bottom). In the mixed electrolyte system (GdmCl/KCl), K^+ ions were almost excluded from the α-hemolysin stem, **SI Figure S25**.

The MD simulations paint a picture where Gdm^+ binding to the inner surface of α-hemolysin generates a strong electro-osmotic flow that produces unidirectional protein transport through the nanopore. In this case, protein transport speeds should be governed predominantly by the flow, rather than the protein sequence. In **Figure 6A** we present dwell-time distributions for GFP-D10 (254 aa), MBP-D10 (389 aa), D10-MBP (389 aa), and diMBP-D10 (764 aa), all at 175 mV applied voltage and with 2 M GdmCl denaturant concentrations. The distributions reveal a correlation between dwell times and protein lengths, however, not with the protein sequence (GFP and MBP have very different sequences). In addition, there is no dependence of the protein transport time on its orientation of entry (N-to-C vs. C-to-N terminus), which differs from the orientation dependence of DNA transport through α-hemolysin[65]. The continuous lines in the plots are optimal fits of the dwell-time distributions to the 1D Fokker-Planck equation[66-69], which yields diffusion coefficients D and drift velocities v (shown in the insets for all experiments). Drift velocities for all molecules were in the range of 0.031 – 0.04 nm/μs, which, given the extended backbone distance between amino acids in a protein chain (0.34 nm)[70], translates to a mean residence time of ~10 μs per amino acid in the pore. This mean velocity for a protein chain is roughly an order of magnitude slower than single-stranded DNA transport through α-hemolysin (0.15 nm/μs) [58]. This points to a stark contrast between protein translocation and DNA/RNA translocation, which yields different electromotive forces and pore/polymer interactions. In **Figure 6B** we show fractional current blockades vs. dwell times for MBP-D10 (orange dots) and D10-MBP (black dots). To a

large extent, there is an overlap in the D10-MBP and MBP-D10 dwell time distributions, with the exception that the MBP protein with its N-terminus D_{10} tail (D10-MBP) is less effective at being captured by the pore, as indicated by many collisions which exhibit shorter dwell times (~100 μs) and lower current blockades. However, it is noteworthy that protein translocation from either direction proceeds with the same speed, in contrast to transport of DNA[65].

Protein-specific current signals. We first analyzed the two cases where N-terminus vs. C-terminus entry of MBP were studied. In order to extract an "average shape" for MBP-D10 and D10-MBP events, their barycenters (Fréchet means) were computed using the Soft Dynamic Time Warping metric[71] (Supplementary Note 2). The result of the barycenter computation is a smooth curve, representing the centroid or the "essence" of the translocation events in the dataset. The barycenters $\vec{x}_{MBPD10}$ and $\vec{x}_{D10MBP}$ are shown in **Figure 6C,** with background traces (black)

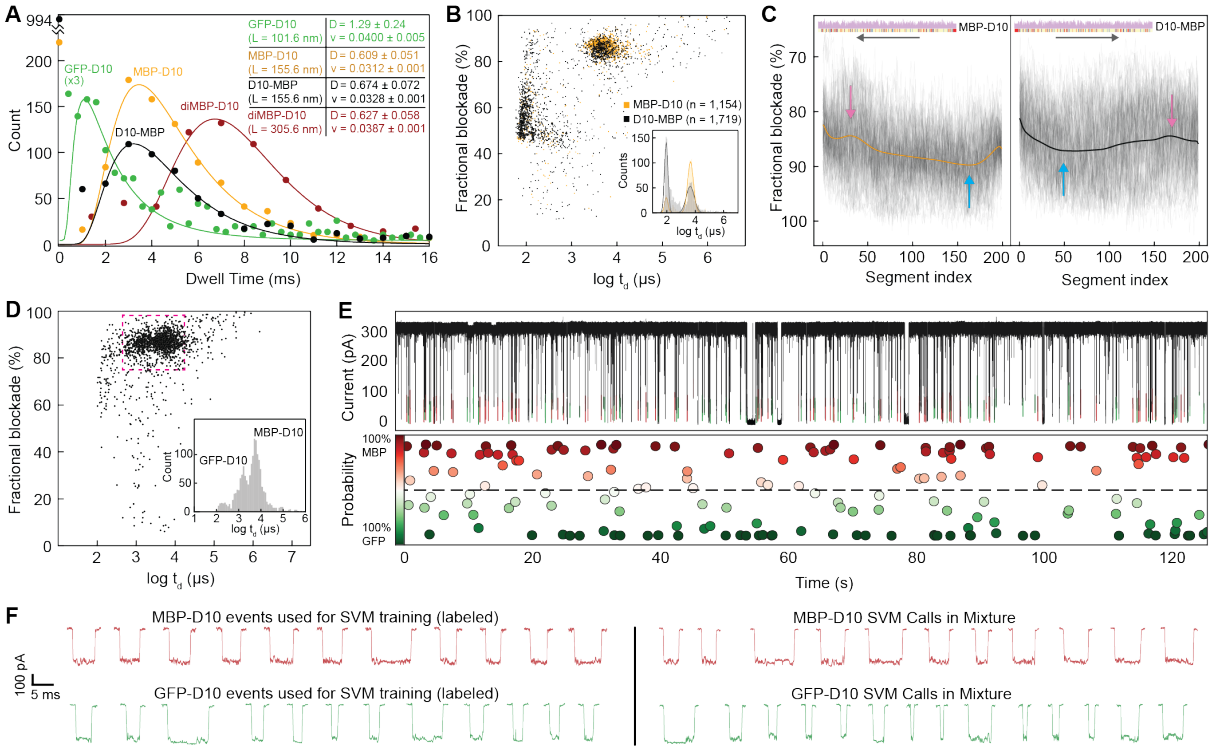

Figure 6. Single-molecule fingerprinting of different proteins. **A)** Dwell time histograms for GFP-D10, MBP-D10, D10-MBP and diMBP-D10 with mean diffusion coefficients ($nm^2/\mu s$) and velocities (nm/μs), determined from fits to the 1D Fokker-Planck equation[66-68]. **B)** Fractional current blockade vs. dwell time scatter plots and dwell time histograms for MBP-D10 and D10-MBP. **C)** Barycenter of 623 MBP-D10 translocation events (orange) and 478 D10-MBP translocation events (black) with 100 events resampled to a fixed length and superimposed in the background, respectively. **D)** Fractional current blockade plot for 0.35 μM MBP-D10 and 0.35 μM GFP-D10 mixture, showing two partially overlapping distributions based on dwell time (dashed red box). **E)** Post-training SVM classification results for MBP-D10 (red) and GFP-D10 (green) mixture experiment with a probability classification estimate associated with each translocation event (see SI, Section 6). **F)** Event traces from pure MBP-D10 (red) and GFP-D10 (green) experiments used for training SVM model, where 51 input features from each event were used to construct SVM decision boundaries (left). Traces of unlabeled translocation events from mixture experiment that were classified by trained SVM model (right). All experiments were performed in 1.0 M KCl, 2.0 M GdmCl, 10 mM Tris, pH 7.5 with 175 mV bias applied to the *trans* chamber.

comprising 100 of their respective resampled events. These curves show a smooth trend of how the events of each protein type tend to progress on average. Both samples distinctly show opposite locations for local maxima and minima (pink and blue arrows, respectively). Whether due to the proteins secondary structure in the pore, or purely due to sequence variation, the opposing current blockage trends are a strong indication of directional protein translocation.

Finally, to investigate whether the signal properties are reproducible and comprise sufficient information to distinguish different proteins, a support vector machine (SVM) was trained on pure GFP-D10 and MBP-D10 and then used to detect protein type from single-pulse measurements in a mixture experiment (1:1 GFP-D10 to MBP-D10 ratio). **Figure 6D** shows a scatter plot of the fractional blockade versus dwell for the mixture experiments, with the histogram showing partial overlap in the dwell time between the two proteins. To accurately classify the proteins irrespectively of dwell times (which significantly overlap), decision boundaries of the SVM classifier were generated based on 51 signal features from each event, where half of the features are the first 25 Fourier coefficients of the signal and the rest are time-domain features. We used a balanced dataset of 381 GFP-D10 events and 381 MBP-D10 events with dwell times ranging between 500 us and 20 ms for SVM training and testing. The model was trained on 80% of the events and was subsequently tested on the remaining 20% of events. Based on the confusion matrix of the test set (Figure S26), the model had a classification accuracy of 93.4%, where 70/74 (94.5%) and 73/79 (93.2%) of GFP-D10 and MBP-D10 events were correctly called, respectively (see SI Section 6 for details). Applying this validated model on 1,142 unlabeled events from the 1:1 mixture experiment, the SVM classified 559 of the events as MBP-D10 (49%) and 583 as GFP-D10 (51%). **Figure 6E** shows the SVM classification results on one of the mixture experiment files, with the confidence score for each call labeled below. Example events from the training sets and mixture classification results are provided in **Figure 6F**. The obtained ~1:1 event ratio is in excellent agreement with the population sizes obtained from fitting the dwell time histograms of the mixture to the sum of two functions that match the pure GFP-D10 and MBP-D10 dwell time histograms (see SI, Supplementary Note 1 and Figure S14), which yields a 53/47 GFP-D10 to MBP-D10 ratio.

Conclusions

We have presented here a plausible system for enzyme-free single-file protein transport through nanopores, mediated by electro-osmosis that is generated through a chemical agent in the buffer. Guanidinium has two important effects on this process: first, by unfolding the protein guanidinium linearizes the protein in preparation for transport, and second, by binding to the hemolysin pore guanidinium induces an electro-osmotic flow that drives protein transport. Nanopore-based single-molecule protein sequencing, which relies on moving a protein chain through a pore detector, requires keeping an unfolded protein taut at the pore in order to reliably measure a set of amino acids in the chain at any given time. Our findings here present a new paradigm for both unfolding and generating a stretching driving force using a single agent, GdmCl. We envision the use of this effect in future nanopore-based single-molecule protein sequencing applications. For example, high-bandwidth measurements can be combined with GdmCl-mediated protein transport through a higher-resolution pore (such as MspA, which maintains its functionality in 2 M GdmCl[72]) or asymmetric buffer conditions in which protein unfolding is mediated in one chamber, electro-osmotic forces are used to keep the protein taut at the pore, and enzyme-mediated motion on another chamber of the pore is used to move the protein through the pore in a discrete manner. Moreover, recent work has shown that the use of different anions (such as nitrate) in the buffer

has an effect of reducing the noise as compared to chloride[73], which could potentially further improve the signal.

Methods

Polymer bilayer painting and nanopore measurement. The chip with 100 μm SU-8 wedge-on-pillar aperture was mounted to our custom designed fluidic cell, sealing properly to separate *cis* and *trans* chambers. Both sides of the aperture were pretreated by 4 mg/ml poly(1,2-butadiene)-b-poly(ethylene oxide) (PBD_{11}–PEO_8) block-copolymer (Polymer Source) dissolved in hexane. After the hexane evaporated, the aperture was coated by a dry and thin polymer layer. The *cis* and *trans* chambers were filled with GdmCl electrolyte (all contain 1 M KCl, 10 mM Tris, pH 7.5), with a pair of Ag/AgCl electrodes (connected to an Axon 200B patch-clamp amplifier) inserted in. Polymer membrane was painted across the aperture using 8 mg/ml polymer dissolved in decane, waiting for at least 60 mins until the polymer membrane got thin enough (~ 60 – 80 pF membrane capacitance for 100 μm aperture and 500 μm thick Si layer). 0.5 μl of 50 μg/ml α-hemolysin (Sigma-Aldrich) was added to the *cis* chamber, waiting until a single pore insertion current signal was observed. Denatured protein sample (incubated in GdmCl before using) was added to *cis* chamber and mixed gently with pipet. Current signals were collected in 250 kHz sampling rate and 100 kHz low pass filter and analyzed by Pythion software. All oligonucleotides were purchased from Integrated DNA Technologies (IDT).

Cloning of the GFP and MBP constructs. The N-terminal 10-aspartate MBP (D10-MBP), C-terminal 10-aspartate MBP (MBP-D10), and C-terminal 10-aspartate GFP (GFP-D10) constructs were obtained by mutagenesis polymerase chain reaction (PCR) using pT7-MBP or pRSETB-GFP as the template plasmid. All primers (Eurofins MWG Operon) used in this study are listed in Table S2. The PCR reaction mixtures were subjected to DpnI digestion for 3 h at 37 °C to degrade the template plasmids. The digested samples were then transformed to chemically competent *E. coli* DH5α cells. The desired mutant plasmids were isolated from colonies and verified by DNA sequencing.

The C-terminal MBP-D10 dimer construct (diMBP-D10) was generated as follows: the first mutagenesis PCR was performed using pT7-hisMBP as the template to remove the stop codon and add a flexible linker ggsg to the C-terminus of the MBP gene. The PCR products were digested with DpnI and transformed into *E. coli* DH5α cells, which resulted in a plasmid pT7-hisMBPggsg containing the hindIII and SfbI restriction sites right after the ggsg linker gene The second PCR was performed with pT7-MBP as the template to introduce HindIII and SfbI cutting sites at the two ends of the MBP gene, and add a D10 at the c-terminal to the MBP fragment. The PCR products and the plasmid pT7-hisMBPgsgg were digested with HindIII and SfbI and ligated by T4 ligase. The ligated products were transformed to chemically competent *E. coli* DH5α cells. The mutant plasmid pT7-diMBP-D10 was verified by enzyme digestion and DNA sequencing.

Expression and purification of GFP and MBP proteins. GFP and MBP protein variants (Table S1) were expressed and purified by using similar protocols. Briefly, plasmids were transformed into chemically competent BL21(DE3) *E. coli* cells. The cells were grown in 1 L of LB medium at 37 °C until the OD600 reached 0.6 and induced with 0.5 mM isopropyl β-D-1-thiogalactopyranoside. The temperature was then decreased to 16°C for overnight expression. Cells were harvested by centrifugation at 13000 RPM for 25 min. The cell pellets were used for protein purification or frozen at −20 °C for future use. To purify proteins, cells were resuspended in 50 ml of 50 mM Tris-HCl (pH 8.0), 150 mM NaCl buffer and lysed via sonication. The lysate

was centrifuged at 13000 RPM for 25 min. The supernatant was filtered through a 0.22 μm syringe filter (CELLTREAT Scientific Products) and then loaded to a Ni-NTA affinity column (ThermoFisher scientific) equilibrated with buffer 50mM Tris-HCl (pH 8.0), 150 mM NaCl. MBP-D10, D10-MBP and diMBP-D10 were eluted in buffer 50 mM Tris-HCl (pH 8.0), 150 mM NaCl, 150 mM imidazole. GFP-D10 was eluted in buffer 50 mM Tris-HCl (pH 8.0), 150mM NaCl, 20 mM imidazole. After Ni-NTA chromatography, MBP-D10, D10-MBP and GFP-D10 exhibited more than 95% purity on SDS-PAGE while the eluted diMBP-D10 fraction contained multiple low-molecular impurity bands. To remove these impurity proteins, the eluted samples were run on a preparative 12 % SDS-PAGE. The band containing the full-length diMBP-D10 was cut out and the protein was extracted from the gel with buffer 50mM Tris-HCl (pH8.0), 8M urea by incubating the gel and the extraction buffer at room temperature for overnight. Supernatant containing the protein was collected by centrifuging the samples at 13000 RPM for 30 min. Protein concentrations of all samples were determined by A280 with Nanodrop and stored at −80 °C for future use.

MD simulation. All MD simulations were performed using the molecular dynamics program NAMD2[74], a 2 femtosecond integration timestep, CHARMM36[75,76] force field, and a custom non-bonded fix (NBFIX) corrections for K, Cl and Gdm ions[77,78]. In all simulations, periodic boundary conditions were applied for all the three main directions, SETTLE algorithm[79] was used to keep covalent bonds involving hydrogen atoms in water molecules, and RATTLE algorithm[80] restrained all other covalent bonds involving hydrogen atoms. The particle-mesh Ewald[81] method was employed for long-range electrostatic interactions computed over a 1.2 Å space grid and the van der Waals and short-range electrostatic forces were calculated using a cutoff of 12 Å and a switching distance of 10 Å. Multiple time stepping [82] was used to calculate nonbonded Lennard-Jones interactions at every time step and full electrostatics at every second-time step.

The all-atom models of *α*-hemolysin suspended in a lipid bilayer membrane were built using CHARMM-GUI[83,84]. The initial structural model of *α*-hemolysin was taken from the Protein Data Bank (PDB ID: 7AHL)[85]. After adding missing atoms and aligning the primary principal axis of the protein with the *z*-axis, the protein structure was merged with a 15×15 nm^2 patch of a pre-equilibrated 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) lipid bilayer. The protein-lipid complex was then solvated in a rectangular volume of ~78,500 pre-equilibrated TIP3P water molecules[86]. Gdm^+, K^+, and Cl^- ions were added at random positions corresponding to target ionic concentrations. Additional charges were introduced to neutralize the system. Each final system was 15×15×18 nm^3 in volume and contained approximately 300,000 atoms. Upon assembly, the systems were initially equilibrated using the default CHARMM-GUI's protocol. Specifically, the systems were subjected to energy minimization for 10,000 steps using the conjugate gradient method. Next, to relax lipid tails and protein side chains, a 2.5 ns pre-equilibration simulation was run while restraining the protein backbones and lipid head groups. This step was followed by a 25 ns unrestrained simulation in the NPT (constant number of particles, pressure, and temperature) ensemble using the Nosé-Hoover Langevin piston pressure control[87]. In all simulations, the temperature was maintained at 298.15 K by coupling all non-hydrogen atoms to a Langevin thermostat with 1 ps^{-1} damping constant.

The atomic coordinates of the maltose-binding protein (MBP) were obtained from the Protein Data Bank (entry 1JW4)[88]. The missing hydrogen atoms were added using the psfgen plugin of VMD[89]. The protonation state of each titratable residue was determined using PROPKA[90,91] according to the experimental pH conditions (7.5 pH). Next, the protein was split into 7 peptide fragments producing six 53-residue and one 52-residue peptides. The N-terminal of each peptide was

terminated with a neutral acetyl group (ACE patch) whereas the C-terminal was terminated with an N-methyl group (CT3 patch). Each peptide was stretched using constant velocity SMD in vacuum, followed by 5 ns equilibration in a 1.5 M GdmCl solution. During the 150 ps SMD run, the C-terminal of the peptide was kept fixed while the N-terminal was coupled to a dummy particle by means of a harmonic potential (k_{spring} = 7 kcal/(mol Å^2)) and the dummy particle was pulled with a constant velocity of 1 Å/ps. At the end of the equilibration step, each peptide fragment had a contour length of approximately 167 Å, ~3.16 Å per residue. Next, we used the phantom-pore method[65] to convert the geometrical shape of the α-hemolysin nanopore to a mathematical surface. To fit the stretched peptide into the α-hemolysin pore, the phantom pore surface was initially made to represent a nanopore that was 1.4 times wider than the pore of *α*-hemolysin. During a 2 ns simulation, the phantom pore was gradually shrunk to match the shape of the *α*-hemolysin nanopore while all atoms of the peptide and all ions laying outside of the potential were pushed toward the center of the nanopore using a constant 50 pN force. At the end of the simulation, the final configuration of each peptide fragment, along with all guanidinium ions residing within 3 Å of any peptide atom, was placed inside the pre-equilibrated *α*-hemolysin system having the peptide's backbone approximately aligned with the nanopore axis. Prior to the production runs, each system was equilibrated for 10 ns in the NPT ensemble at 1.0 bar and 298.15 K with all C_α atoms of the α-hemolysin protein restrained to the crystallographic coordinates.

All production simulations were carried out in the constant number of particles, volume, and temperature ensemble (NVT) under a constant external electric field applied normal to the membrane, producing a ±200 mV transmembrane bias. To maintain the structural integrity of the nanopore, all C_α atoms of the protein were restrained to same coordinates as in the last frame of the equilibration trajectory using harmonic potentials with spring constants of 69.5 pN/nm. The ionic currents were calculated as described previously[63]. The quantify protein translocation, we defined the number of residues translocated as the number of non-hydrogen backbone atoms passing below the α-hemolysin constriction divided by the total number of non-hydrogen backbone atoms in one residue. The constriction's z-coordinate was defined by the center of mass of the backbone atoms of residues 111, 113, and 147. The concentration profile and guanidinium binding analyses were carried out using in-house VMD scripts. All MD trajectories were visualized using VMD[89]. The electrostatic potential in open-pore systems was calculated using a previously described method[63], implemented in the PMEpot Plugin of VMD, where the instantaneous distributions of the electrostatic potential are averaged over the respective MD trajectory.

Data analysis. All data parsing (excluding DTW and SVM) were performed using the Pyth-ion package (https://github.com/wanunulab/Pyth-Ion) and figures were generated using Igor. DTW and SVM analyses were conducted via a Jupyter notebook python script, tslearn[92], SciKit-Learn[93], and a modified version of the PyPore[94] nanopore data analysis library. The Jupyter notebook and associated files are available on GitHub (https://github.com/wanunulab/protein-gd). A detailed description of the DTW and SVM analyses is provided in **SI Section 6**.

Supporting Information Available

Data analysis details, example datasets, and methods description.

Acknowledgment

We thank Nikolai Slavov for helpful discussions regarding protein sequencing. We acknowledge funding from the National Institutes of Health grant # HG11087 (MW); GM115442 (MC); the

National Science Foundatoin grant DMR-1827346 (AA). The supercomputer time was provided through the XSEDE allocation grant (MCA05S028) and the Leadership Resource Allocation MCB20012 on Frontera of the Texas Advanced Computing Center.

References

1 Shendure, J. *et al.* DNA sequencing at 40: past, present and future. *Nature* **550**, 345-353, doi:10.1038/nature24286 (2017).

2 Ameur, A., Kloosterman, W. P. & Hestand, M. S. Single-Molecule Sequencing: Towards Clinical Applications. *Trends in Biotechnology* **37**, 72-85, doi:https://doi.org/10.1016/j.tibtech.2018.07.013 (2019).

3 Eid, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **323**, 133, doi:10.1126/science.1162986 (2009).

4 Derrington, I. M. *et al.* Nanopore DNA sequencing with MspA. *Proceedings of the National Academy of Sciences* **107**, 16060-16065, doi:10.1073/pnas.1001831107 (2010).

5 English, A. C. *et al.* Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLOS ONE* **7**, e47768, doi:10.1371/journal.pone.0047768 (2012).

6 Chaisson, M. J. P. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608-611, doi:10.1038/nature13907 (2015).

7 Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology* **17**, 239, doi:10.1186/s13059-016-1103-0 (2016).

8 Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* **36**, 338-345, doi:10.1038/nbt.4060 (2018).

9 Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79-84, doi:10.1038/s41586-020-2547-7 (2020).

10 Johnson, J. M., Edwards, S., Shoemaker, D. & Schadt, E. E. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends in Genetics* **21**, 93-102, doi:10.1016/j.tig.2004.12.009 (2005).

11 Tsai, Y.-C. *et al.* Amplification-free, CRISPR-Cas9 Targeted Enrichment and SMRT Sequencing of Repeat-Expansion Disease Causative Genomic Regions. *bioRxiv*, 203919, doi:10.1101/203919 (2017).

12 Doi, K. *et al.* Rapid detection of expanded short tandem repeats in personal genomics using hybrid sequencing. *Bioinformatics* **30**, 815-822, doi:10.1093/bioinformatics/btt647 (2014).

13 Flusberg, B. A. *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods* **7**, 461-465, doi:10.1038/nmeth.1459 (2010).

14 Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods* **14**, 407-410, doi:10.1038/nmeth.4184 (2017).

15 Gouil, Q. & Keniry, A. Latest techniques to study DNA methylation. *Essays in Biochemistry* **63**, 639-648, doi:10.1042/EBC20190027 (2019).

16 Smith, L. M. *et al.* Proteoform: a single term describing protein complexity. *Nature Methods* **10**, 186-187, doi:10.1038/nmeth.2369 (2013).

17 Bogaert, A., Fernandez, E. & Gevaert, K. N-Terminal Proteoforms in Human Disease. *Trends in Biochemical Sciences* **45**, 308-320, doi:10.1016/j.tibs.2019.12.009 (2020).

18 Tolsma, Thomas O. & Hansen, Jeffrey C. Post-translational modifications and chromatin dynamics. *Essays in Biochemistry* **63**, 89-96, doi:10.1042/ebc20180067 (2019).

19 Conibear, A. C. Deciphering protein post-translational modifications using chemical biology tools. *Nature Reviews Chemistry* **4**, 674-695, doi:10.1038/s41570-020-00223-8 (2020).
20 Slavov, N. Single-cell protein analysis by mass spectrometry. *Current Opinion in Chemical Biology* **60**, 1-9, doi:https://doi.org/10.1016/j.cbpa.2020.04.018 (2021).
21 Specht, H. & Slavov, N. Transformative Opportunities for Single-Cell Proteomics. *J Proteome Res* **17**, 2565-2571, doi:10.1021/acs.jproteome.8b00257 (2018).
22 Specht, H. *et al.* Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome Biology* **22**, 50, doi:10.1186/s13059-021-02267-5 (2021).
23 Zhao, Y. *et al.* Single-molecule spectroscopy of amino acids and peptides by recognition tunnelling. *Nature Nanotechnology* **9**, 466-473, doi:10.1038/nnano.2014.54 (2014).
24 Kennedy, E., Dong, Z., Tennant, C. & Timp, G. Reading the primary structure of a protein with 0.07 nm3 resolution using a subnanometre-diameter pore. *Nature Nanotechnology* **11**, 968-976, doi:10.1038/nnano.2016.120 (2016).
25 Swaminathan, J. *et al.* Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. *Nature Biotechnology* **36**, 1076-1082, doi:10.1038/nbt.4278 (2018).
26 van Ginkel, J. *et al.* Single-molecule peptide fingerprinting. *Proceedings of the National Academy of Sciences* **115**, 3338-3343, doi:10.1073/pnas.1707207115 (2018).
27 Restrepo-Pérez, L., Joo, C. & Dekker, C. Paving the way to single-molecule protein sequencing. *Nature Nanotechnology* **13**, 786-796, doi:10.1038/s41565-018-0236-6 (2018).
28 Alfaro, J. A. *et al.* The emerging landscape of single-molecule protein sequencing technologies. *Nature Methods* **18**, 604-617, doi:10.1038/s41592-021-01143-1 (2021).
29 Keyser, U. F. *et al.* Direct force measurements on DNA in a solid-state nanopore. *Nature Physics* **2**, 473-477, doi:10.1038/nphys344 (2006).
30 Cherf, G. M. *et al.* Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. *Nature Biotechnology* **30**, 344-348, doi:10.1038/nbt.2147 (2012).
31 Branton, D. *et al.* The potential and challenges of nanopore sequencing. *Nature Biotechnology* **26**, 1146-1153, doi:10.1038/nbt.1495 (2008).
32 Venta, K. *et al.* Differentiation of Short, Single-Stranded DNA Homopolymers in Solid-State Nanopores. *ACS Nano* **7**, 4629-4636, doi:10.1021/nn4014388 (2013).
33 Niedzwiecki, D. J. *et al.* Observing Changes in the Structure and Oligomerization State of a Helical Protein Dimer Using Solid-State Nanopores. *ACS Nano* **9**, 8907-8915, doi:10.1021/acsnano.5b02714 (2015).
34 Chien, C.-C., Shekar, S., Niedzwiecki, D. J., Shepard, K. L. & Drndić, M. Single-Stranded DNA Translocation Recordings through Solid-State Nanopores on Glass Chips at 10 MHz Measurement Bandwidth. *ACS Nano* **13**, 10545-10554, doi:10.1021/acsnano.9b04626 (2019).
35 Shekar, S. *et al.* Wavelet Denoising of High-Bandwidth Nanopore and Ion-Channel Signals. *Nano Letters* **19**, 1090-1097, doi:10.1021/acs.nanolett.8b04388 (2019).
36 Nivala, J., Marks, D. B. & Akeson, M. Unfoldase-mediated protein translocation through an alpha-hemolysin nanopore. *Nat Biotechnol* **31**, 247-250, doi:10.1038/nbt.2503 (2013).
37 Nivala, J., Mulroney, L., Li, G., Schreiber, J. & Akeson, M. Discrimination among Protein Variants Using an Unfoldase-Coupled Nanopore. *ACS Nano* **8**, 12365-12375, doi:10.1021/nn5049987 (2014).
38 Cardozo, N. *et al.* Multiplexed direct detection of barcoded protein reporters on a nanopore array. *bioRxiv*, 837542, doi:10.1101/837542 (2019).
39 Yan, S. *et al.* Single Molecule Ratcheting Motion of Peptides in a Mycobacterium smegmatis Porin A (MspA) Nanopore. *Nano Letters*, doi:10.1021/acs.nanolett.1c02371 (2021).

40 Brinkerhoff, H., Kang, A. S. W., Liu, J., Aksimentiev, A. & Dekker, C. Infinite re-reading of single proteins at single-amino-acid resolution using nanopore sequencing. *bioRxiv*, 2021.2007.2013.452225, doi:10.1101/2021.07.13.452225 (2021).
41 Stefureac, R., Long, Y.-t., Kraatz, H.-B., Howard, P. & Lee, J. S. Transport of α-Helical Peptides through α-Hemolysin and Aerolysin Pores. *Biochemistry* **45**, 9172-9179, doi:10.1021/bi0604835 (2006).
42 Movileanu, L. Squeezing a single polypeptide through a nanopore. *Soft Matter* **4**, 925-931, doi:10.1039/B719850G (2008).
43 Rodriguez-Larrea, D. & Bayley, H. Multistep protein unfolding during nanopore translocation. *Nat Nanotechnol* **8**, 288-295, doi:10.1038/nnano.2013.22 (2013).
44 Rosen, C. B., Bayley, H. & Rodriguez-Larrea, D. Free-energy landscapes of membrane co-translocational protein unfolding. *Commun Biol* **3**, 160, doi:10.1038/s42003-020-0841-4 (2020).
45 Payet, L. *et al.* Thermal unfolding of proteins probed at the single molecule level using nanopores. *Anal Chem* **84**, 4071-4076, doi:10.1021/ac300129e (2012).
46 Cressiot, B. *et al.* Protein Transport through a Narrow Solid-State Nanopore at High Voltage: Experiments and Theory. *ACS Nano* **6**, 6236-6243, doi:10.1021/nn301672g (2012).
47 Oukhaled, G. *et al.* Unfolding of Proteins and Long Transient Conformations Detected by Single Nanopore Recording. *Physical Review Letters* **98**, 158101, doi:10.1103/PhysRevLett.98.158101 (2007).
48 Pastoriza-Gallego, M. *et al.* Dynamics of unfolded protein transport through an aerolysin pore. *J Am Chem Soc* **133**, 2923-2931, doi:10.1021/ja1073245 (2011).
49 Merstorf, C. *et al.* Wild Type, Mutant Protein Unfolding and Phase Transition Detected by Single-Nanopore Recording. *ACS Chemical Biology* **7**, 652-658, doi:10.1021/cb2004737 (2012).
50 Pastoriza-Gallego, M. *et al.* Evidence of Unfolded Protein Translocation through a Protein Nanopore. *ACS Nano* **8**, 11350-11360, doi:10.1021/nn5042398 (2014).
51 Yu, L. *et al.* Stable polymer bilayers for protein channel recordings at high guanidinium chloride concentrations. *Biophysical Journal* **120**, 1537-1541, doi:https://doi.org/10.1016/j.bpj.2021.02.019 (2021).
52 Meller, A. & Branton, D. Single molecule measurements of DNA transport through a nanopore. *Electrophoresis* **23**, 2583-2591, doi:10.1002/1522-2683(200208)23:16<2583::Aid-elps2583>3.0.Co;2-h (2002).
53 Kang, X., Alibakhshi, M. A. & Wanunu, M. One-Pot Species Release and Nanopore Detection in a Voltage-Stable Lipid Bilayer Platform. *Nano Letters* **19**, 9145-9153, doi:10.1021/acs.nanolett.9b04446 (2019).
54 Haynes, W. M. *CRC Handbook of Chemistry and Physics*. (CRC Press, 2016).
55 Perkins, S. J. Protein volumes and hydration effects. *European Journal of Biochemistry* **157**, 169-180, doi:https://doi.org/10.1111/j.1432-1033.1986.tb09653.x (1986).
56 Liu, G. P., Topping, T. B., Cover, W. H. & Randall, L. L. Retardation of folding as a possible means of suppression of a mutation in the leader sequence of an exported protein. *Journal of Biological Chemistry* **263**, 14790-14793, doi:https://doi.org/10.1016/S0021-9258(18)68107-4 (1988).
57 Sheshadri, S., Lingaraju, G. M. & Varadarajan, R. Denaturant mediated unfolding of both native and molten globule states of maltose binding protein are accompanied by large deltaCp's. *Protein Sci* **8**, 1689-1695, doi:10.1110/ps.8.8.1689 (1999).
58 Meller, A., Nivon, L. & Branton, D. Voltage-Driven DNA Translocations through a Nanopore. *Physical Review Letters* **86**, 3435-3438, doi:10.1103/PhysRevLett.86.3435 (2001).
59 Meller, A. & Branton, D. Single molecule measurements of DNA transport through a nanopore. *ELECTROPHORESIS* **23**, 2583-2591, doi:https://doi.org/10.1002/1522-2683(200208)23:16<2583::AID-ELPS2583>3.0.CO;2-H (2002).

60 Hornblower, B. *et al.* Single-molecule analysis of DNA-protein complexes using nanopores. *Nature Methods* **4**, 315-317, doi:10.1038/nmeth1021 (2007).
61 Henrickson, S. E., Misakian, M., Robertson, B. & Kasianowicz, J. J. Driven DNA Transport into an Asymmetric Nanometer-Scale Pore. *Physical Review Letters* **85**, 3057-3060, doi:10.1103/PhysRevLett.85.3057 (2000).
62 Korchev, Y. E. *et al.* Staphylococcus aureus alpha-toxin-induced pores: Channel-like behavior in lipid bilayers and patch clamped cells. *The Journal of Membrane Biology* **143**, 143-151, doi:10.1007/BF00234660 (1995).
63 Aksimentiev, A. & Schulten, K. Imaging α-Hemolysin with Molecular Dynamics: Ionic Conductance, Osmotic Permeability, and the Electrostatic Potential Map. *Biophysical Journal* **88**, 3745-3761, doi:https://doi.org/10.1529/biophysj.104.058727 (2005).
64 Bhattacharya, S. *et al.* Rectification of the Current in α-Hemolysin Pore Depends on the Cation Type: The Alkali Series Probed by Molecular Dynamics Simulations and Experiments. *The Journal of Physical Chemistry C* **115**, 4255-4264, doi:10.1021/jp111441p (2011).
65 Mathé, J., Aksimentiev, A., Nelson, D. R., Schulten, K. & Meller, A. Orientation discrimination of single-stranded DNA inside the α-hemolysin membrane channel. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 12377-12382, doi:10.1073/pnas.0502947102 (2005).
66 Schrödinger, E. Zur theorie der fall-und steigversuche an teilchen mit brownscher bewegung. *Physikalische Zeitschrift* **16**, 289-295 (1915).
67 Cox, D. & Miller, H. (London, 1965).
68 Ammenti, A. & Cecconi, F. Marconi UMB and Vulpiani A. *J. Phys. Chem. B* **2009**, 113 (2009).
69 Ling, D. Y. & Ling, X. S. On the distribution of DNA translocation times in solid-state nanopores: an analysis using Schrödinger's first-passage-time theory. *Journal of Physics: Condensed Matter* **25**, 375102 (2013).
70 Yang, G. *et al.* Solid-state synthesis and mechanical unfolding of polymers of T4 lysozyme. *Proceedings of the National Academy of Sciences* **97**, 139, doi:10.1073/pnas.97.1.139 (2000).
71 Cuturi, M. & Blondel, M. in *Proceedings of the 34th International Conference on Machine Learning* Vol. 70 (eds Precup Doina & Teh Yee Whye) 894--903 (PMLR, Proceedings of Machine Learning Research, 2017).
72 Pavlenok, M., Yu, L., Herrmann, D., Wanunu, M. & Niederweis, M. Control of subunit stoichiometry in single-chain MspA nanopores. *bioRxiv*, 2021.2009.2025.461773, doi:10.1101/2021.09.25.461773 (2021).
73 Ouldali, H. *et al.* Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore. *Nat Biotechnol* **38**, 176-181, doi:10.1038/s41587-019-0345-2 (2020).
74 Phillips, J. C. *et al.* Scalable molecular dynamics on CPU and GPU architectures with NAMD. *The Journal of Chemical Physics* **153**, 044130, doi:10.1063/5.0014475 (2020).
75 Klauda, J. B. *et al.* 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, doi:10.1021/jp101759q (2010).
76 MacKerell, A. D., Feig, M. & Brooks, C. L. Improved Treatment of the Protein Backbone in Empirical Force Fields. *Journal of the American Chemical Society* **126**, 698-699, doi:10.1021/ja036959e (2004).
77 Yoo, J. & Aksimentiev, A. Improved Parametrization of Li+, Na+, K+, and Mg2+ Ions for All-Atom Molecular Dynamics Simulations of Nucleic Acid Systems. *The Journal of Physical Chemistry Letters* **3**, 45-50, doi:10.1021/jz201501a (2012).

78 Yoo, J. & Aksimentiev, A. New tricks for old dogs: improving the accuracy of biomolecular force fields by pair-specific corrections to non-bonded interactions. *Physical Chemistry Chemical Physics* **20**, 8432-8449, doi:10.1039/C7CP08185E (2018).
79 Miyamoto, S. & Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry* **13**, 952-962, doi:https://doi.org/10.1002/jcc.540130805 (1992).
80 Andersen, H. C. Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. *Journal of Computational Physics* **52**, 24-34, doi:https://doi.org/10.1016/0021-9991(83)90014-1 (1983).
81 Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **98**, 10089-10092, doi:10.1063/1.464397 (1993).
82 Tuckerman, M., Berne, B. J. & Martyna, G. J. Reversible multiple time scale molecular dynamics. *The Journal of Chemical Physics* **97**, 1990-2001, doi:10.1063/1.463137 (1992).
83 Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry* **29**, 1859-1865, doi:https://doi.org/10.1002/jcc.20945 (2008).
84 Lee, J. *et al.* CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *Journal of Chemical Theory and Computation* **12**, 405-413, doi:10.1021/acs.jctc.5b00935 (2016).
85 Song, L. *et al.* Structure of staphylococcal α-hemolysin, a heptameric transmembrane pore. *Science* **274**, 1859-1865 (1996).
86 Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* **79**, 926-935 (1983).
87 Martyna, G. J., Tobias, D. J. & Klein, M. L. Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics* **101**, 4177-4189, doi:10.1063/1.467468 (1994).
88 Duan, X. & Quiocho, F. A. Structural Evidence for a Dominant Role of Nonpolar Interactions in the Binding of a Transport/Chemosensory Receptor to Its Highly Polar Ligands. *Biochemistry* **41**, 706-712, doi:10.1021/bi015784n (2002).
89 Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **14**, 33-38, doi:https://doi.org/10.1016/0263-7855(96)00018-5 (1996).
90 Li, H., Robertson, A. D. & Jensen, J. H. Very fast empirical prediction and rationalization of protein pKa values. *Proteins: Structure, Function, and Bioinformatics* **61**, 704-721, doi:https://doi.org/10.1002/prot.20660 (2005).
91 Søndergaard, C. R., Olsson, M. H. M., Rostkowski, M. & Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *Journal of Chemical Theory and Computation* **7**, 2284-2295, doi:10.1021/ct200133y (2011).
92 Tavenard, R. *et al.* Tslearn, A Machine Learning Toolkit for Time Series Data. *J. Mach. Learn. Res.* **21**, 1-6 (2020).
93 Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825-2830 (2011).
94 Schreiber, J. & Karplus, K. Analysis of nanopore data using hidden Markov models. *Bioinformatics* **31**, 1897-1903, doi:10.1093/bioinformatics/btv046 (2015).