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

Luning Yu,[1] Xinqi Kang,[2] Fanjun Li,[4] Behzad Mehrafrooz,[5] 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, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States

# Table of Contents

a. **Figure S14:** Current vs. time trace for MBP-D10 without GdmCl at V = 175 mV
b. **Figure S15:** $P_L$ percentage vs. voltage plots for MBP-D10 and diMBP-D10 in different GdmCl concentrations
c. **Figure S16:** Concentration-normalized capture rates as a function of voltage for MBP-D10 and diMBP-D10 in different GdmCl concentrations
d. **Figure S17:** Dwell time histograms for pure MBP-D10, pure GFP-D10 and mixture of them
e. **Supplementary note 1:** Estimate of the event ratio in the MBP-D10/GFP-D10 mixture

Section 5: MD simulation

a. **Table S3:** Fragmentation of the maltose-binding protein (MBP) into seven peptides for MD simulations.
b. **Figure S18:** MD simulation of α-hemolysin systems containing fragments of MBP.
c. **Figure S19:** Simulated electro-osmotic flow in α-hemolysin systems containing fragments of MBP
d. **Figure S20:** Gdm binding to the inner surface of α-hemolysin in MD simulations of the α-hemolysin systems containing MBP fragments
e. **Figure S21:** Local concentration of ions along the transmembrane nanopore observed in MD simulations of the α-hemolysin systems containing fragments of MBP
f. **Figure S22:** Simulated transport of MBP fragments
g. **Figure S23:** Number of residues confined within the stem region of α-hemolysin

Section 6: Machine learning analysis

a. **Figure S24:** Confusion matrix of model used on mixture data
b. **Figure S25:** Comparison between event traces for pure MBP-D10 or GFP-D10 and event traces classified by trained SVM model
c. **Figure S26:** SVM classification accuracy as a function of class sample ratio
d. **Table S4:** The classification accuracy of the labeled events from each pure file after removing it from the dataset prior to SVM training and testing.
e. **Figure S27:** Classification Accuracy after comparing MBP-D10 and GFP-D10 baseline power spectral density
f. **Table S5:** Sample size and event parameters used for $softDTW_\gamma$comparison of MBP-D10 and D10-MBP
g. **Supplementary note 2:** Soft-DTW Analysis
h. **Supplementary note 3:** SVM Analysis

## Section 1: Information on proteins in this study

**Table S1.** Amino acid sequences of the protein constructs in this study

| |
|---|
| **N-terminal D10 MBP (D10-MBP)**<br>MDDDDDDDDDDKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTVEHPDKLEEKFPQVA<br>ATGDGPDIIFWAHDRFGGYAQSGLLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALS<br>LIYNKDLLPNPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWPLIAADGGYAFKYENGKYD<br>IKDVGVDNAGAKAGLTFLVDLIKNKHMNADTDYSIAEAAFNKGETAMTINGPWAWSNIDTSKV<br>NYGVTVLPTFKGQPSKPFVGVLSAGINAASPNKELAKEFLENYLLTDEGLEAVNKDKPLGAV<br>ALKSYEEELAKDPRIAATMENAQKGEIMPNIPQMSAFWYAVRTAVINAASGRQTVDEALKDA<br>QTRITKHHHHHHHH |
| **C-terminal D10 MBP (MBP-D10)**<br>MHHHHHHHHKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTVEHPDKLEEKFPQVAAT<br>GDGPDIIFWAHDRFGGYAQSGLLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIY<br>NKDLLPNPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWPLIAADGGYAFKYENGKYDIK<br>DVGVDNAGAKAGLTFLVDLIKNKHMNADTDYSIAEAAFNKGETAMTINGPWAWSNIDTSKVN<br>YGVTVLPTFKGQPSKPFVGVLSAGINAASPNKELAKEFLENYLLTDEGLEAVNKDKPLGAVAL<br>KSYEEELAKDPRIAATMENAQKGEIMPNIPQMSAFWYAVRTAVINAASGRQTVDEALKDAQT<br>RITKDDDDDDDDDD |
| **C-terminal D10 dimer MBP (diMBP-D10)**<br>MHHHHHHHHKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTVEHPDKLEEKFPQVAAT<br>GDGPDIIFWAHDRFGGYAQSGLLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIY<br>NKDLLPNPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWPLIAADGGYAFKYENGKYDIK<br>DVGVDNAGAKAGLTFLVDLIKNKHMNADTDYSIAEAAFNKGETAMTINGPWAWSNIDTSKVN<br>YGVTVLPTFKGQPSKPFVGVLSAGINAASPNKELAKEFLENYLLTDEGLEAVNKDKPLGAVAL<br>KSYEEELAKDPRIAATMENAQKGEIMPNIPQMSAFWYAVRTAVINAASGRQTVDEALKDAQT<br>RITKGGSGKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTVEHPDKLEEKFPQVAATGD<br>GPDIIFWAHDRFGGYAQSGLLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNK<br>DLLPNPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWPLIAADGGYAFKYENGKYDIKDV<br>GVDNAGAKAGLTFLVDLIKNKHMNADTDYSIAEAAFNKGETAMTINGPWAWSNIDTSKVNYG<br>VTVLPTFKGQPSKPFVGVLSAGINAASPNKELAKEFLENYLLTDEGLEAVNKDKPLGAVALKS<br>YEEELAKDPRIAATMENAQKGEIMPNIPQMSAFWYAVRTAVINAASGRQTVDEALKDAQTRIT<br>KDDDDDDDDDD |
| **C-terminal D10 GFP (GFP-D10)**<br>MHHHHHHSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVP<br>WPTLVTTFAYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGD<br>TLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADH<br>YQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITHGMDELYKDDDD<br>DDDDDD |

**Table S2.** List of primers for cloning GFP and MBP constructs

| Desired construct | Template plasmid | Primers |
|---|---|---|
| N-terminal 10-aspartate MBP (D10-MBP) | pT7-MBPhis | Forward: 5'-GATGATGACGACGATGATGATAAAATCGAAGAAGGTAAAC-3' |
| | | Reverse: 5'-ATCGTCGTCATCATCGTCATCATCCATATGTATATCTCCTTC-3' |
| C-terminal 10-aspartate MBP (MBP-D10) | pT7-hisMBP | Forward: 5'-GATGATGACGACGATGATGATTAAAAGCTTGGATCC-3' |
| | | Reverse: 5'-ATCGTCGTCATCATCGTCATCATCCTTGGTGATACGAGTC-3' |
| C-terminal 10-aspartate GFP (GFP-D10) | pRSETB-GFP | Forward: 5'-CGATGATGACGACGATGATGATTAAAAGCTTGATCCGGCTGCTAACAAAGCCCGAAAG-3' |
| | | Reverse: 5'-CATCGTCGTCATCATCGTCATCATCTTTGTATAGTTCATCCATG-3' |
| MBP-D10 with HindIII and SbfI cut sites | pT7-hisMBP-D10 | Forward: 5'-ATGATGAAGCTTAAAATCGAAGAAGGTAAAC-3' |
| | | Reverse: 5'-TTCGGACCTGCAGGTTAATCATCATCGTCG-3' |
| MPB with GS-linker at C-terminus, no stop codon | pT7-hisMBP | Forward: 5'-AGGGTAGTAAGCTTGATCCGGCTGCTAACAAAG-3' |
| | | Reverse: 5'-TCAAGCTTACTACCCTTGGTGATACGAGTCTG-3' |

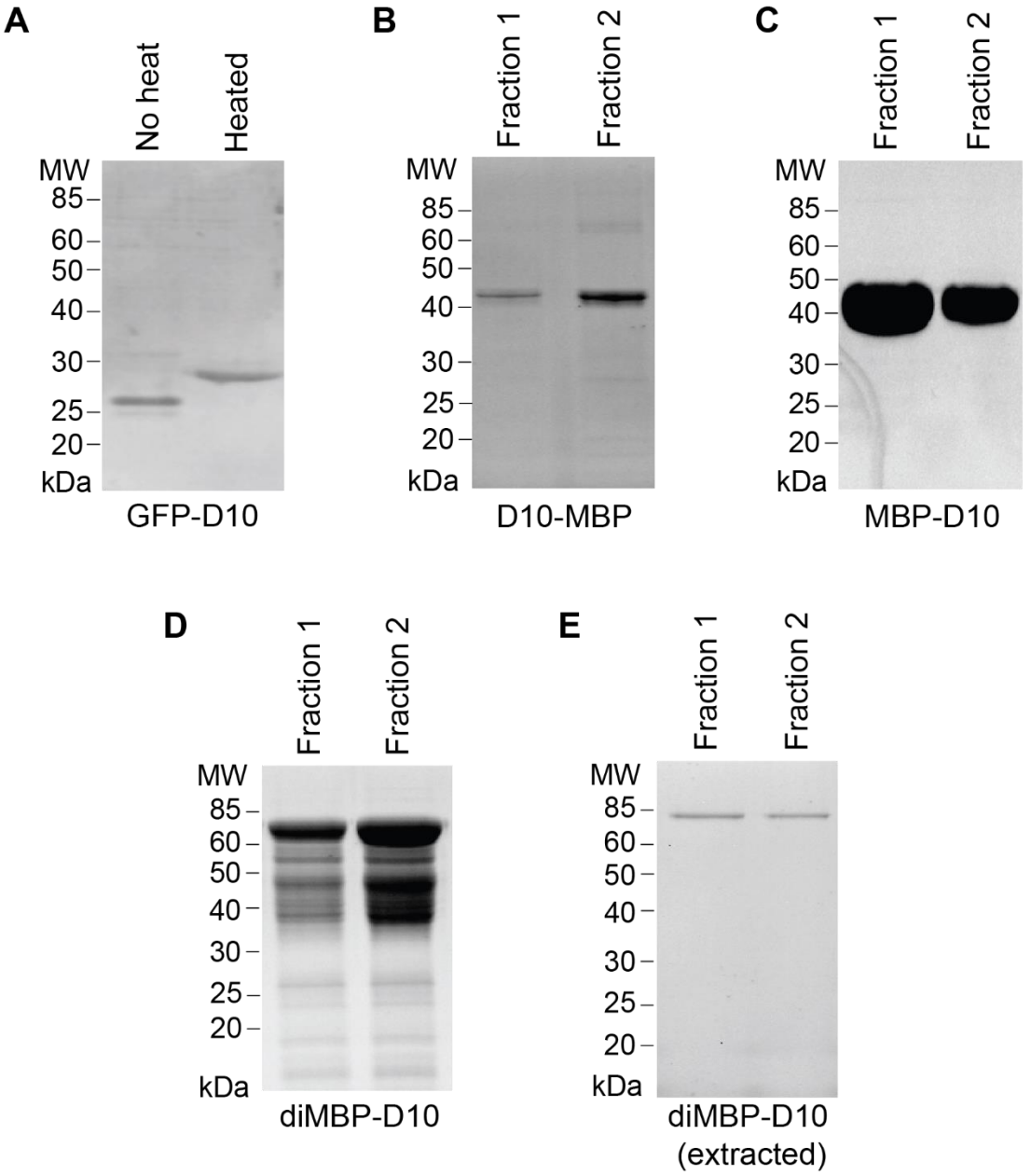


**Figure S1.** 12% SDS-PAGE analysis of MBP and GFP proteins. **A)** Same fraction of C-terminus GFP-D10 after Ni-NTA chromatography purification. Left: no heat; right: heat at 95°C in 50 mM Tris-HCl (pH 8.0), 150 mM NaCl, 20 mM imidazole sample buffer for 15 min. Two fractions of **B)** N-terminus D10-MBP, **C)** C-terminus MBP-D10 and **D)** dimer C-terminus diMBP-D10 after Ni-NTA chromatography purification. **E)** Two representative fractions of diMBP-D10 after gel extraction purification.

## Section 2: Extra data and analysis for MBP-D10

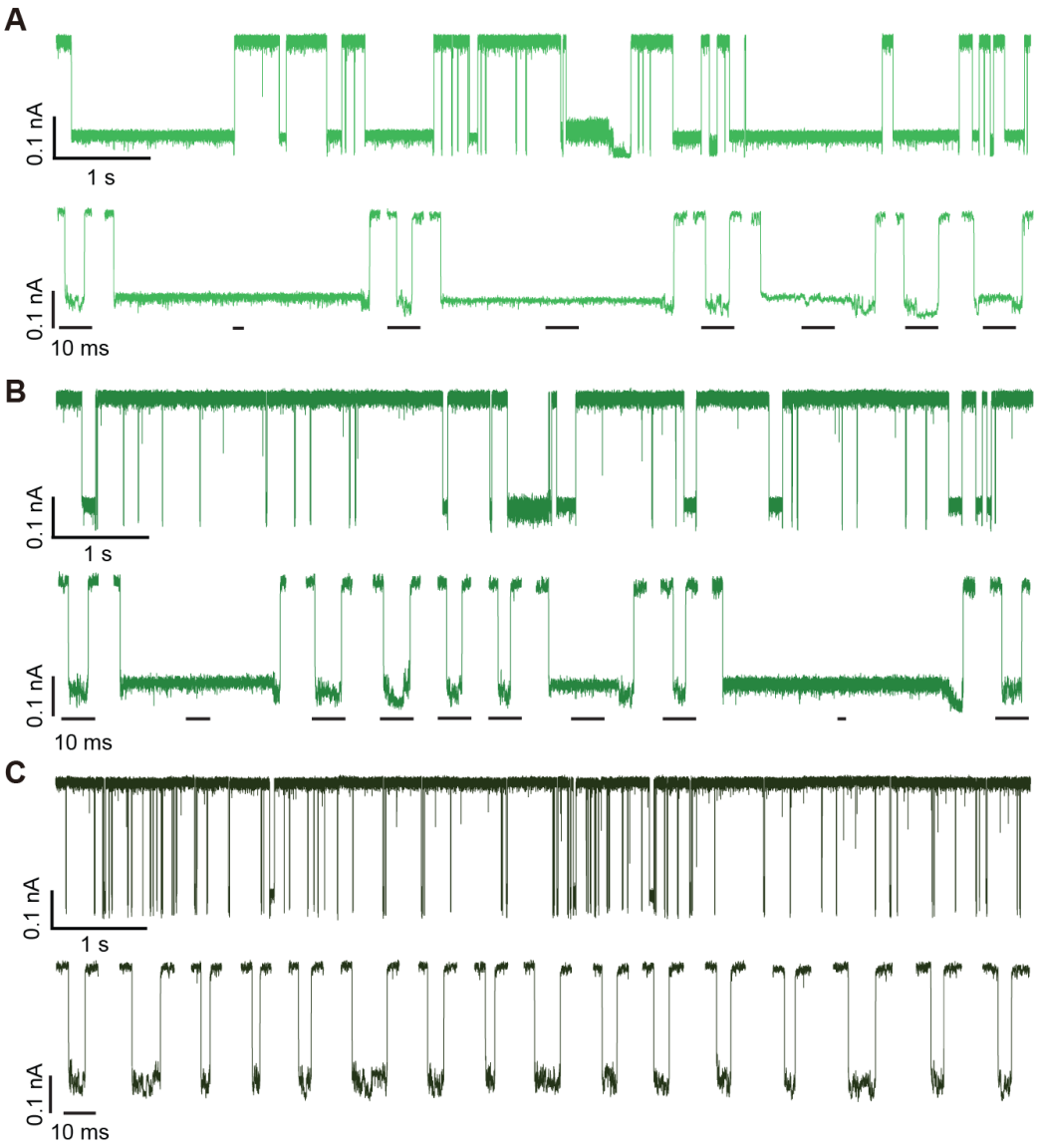


**Figure S2.** Continuous current vs. time traces for MBP-D10 in 1.0 M, 1.5 M and 2.0 M GdmCl, with the open pore currents $I_0$ = 0.260, 0.308, 0.352 nA, respectively. Under each continuous trace we show random events selected at 10 s intervals. All time-scale bars for the selected events are 10 ms. 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 **panel C**).

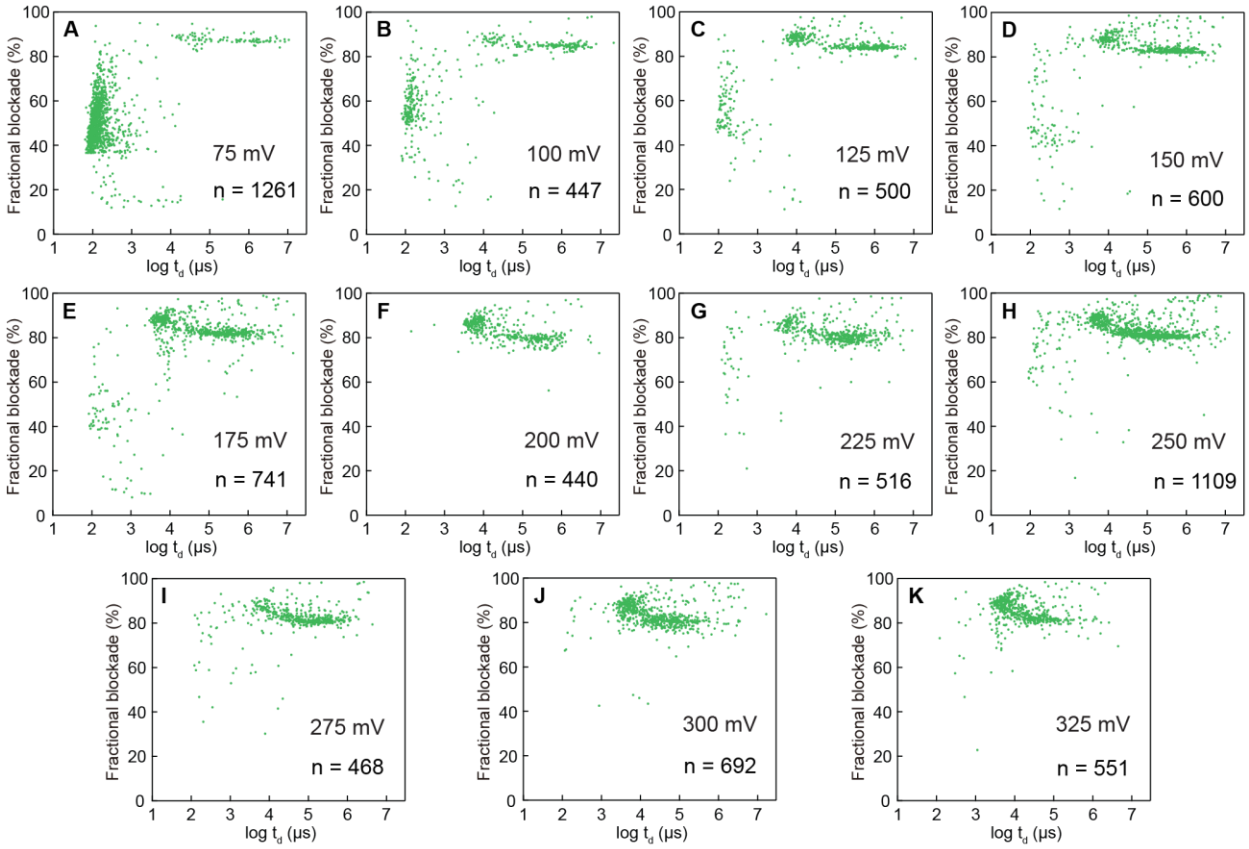


**Figure S3.** Fractional current blockade vs. dwell time scatters for MBP-D10 (0.7 µM) in 1.0 M GdmCl (1 M KCl, 10 mM Tris, pH 7.5), V = 75 – 325 mV.

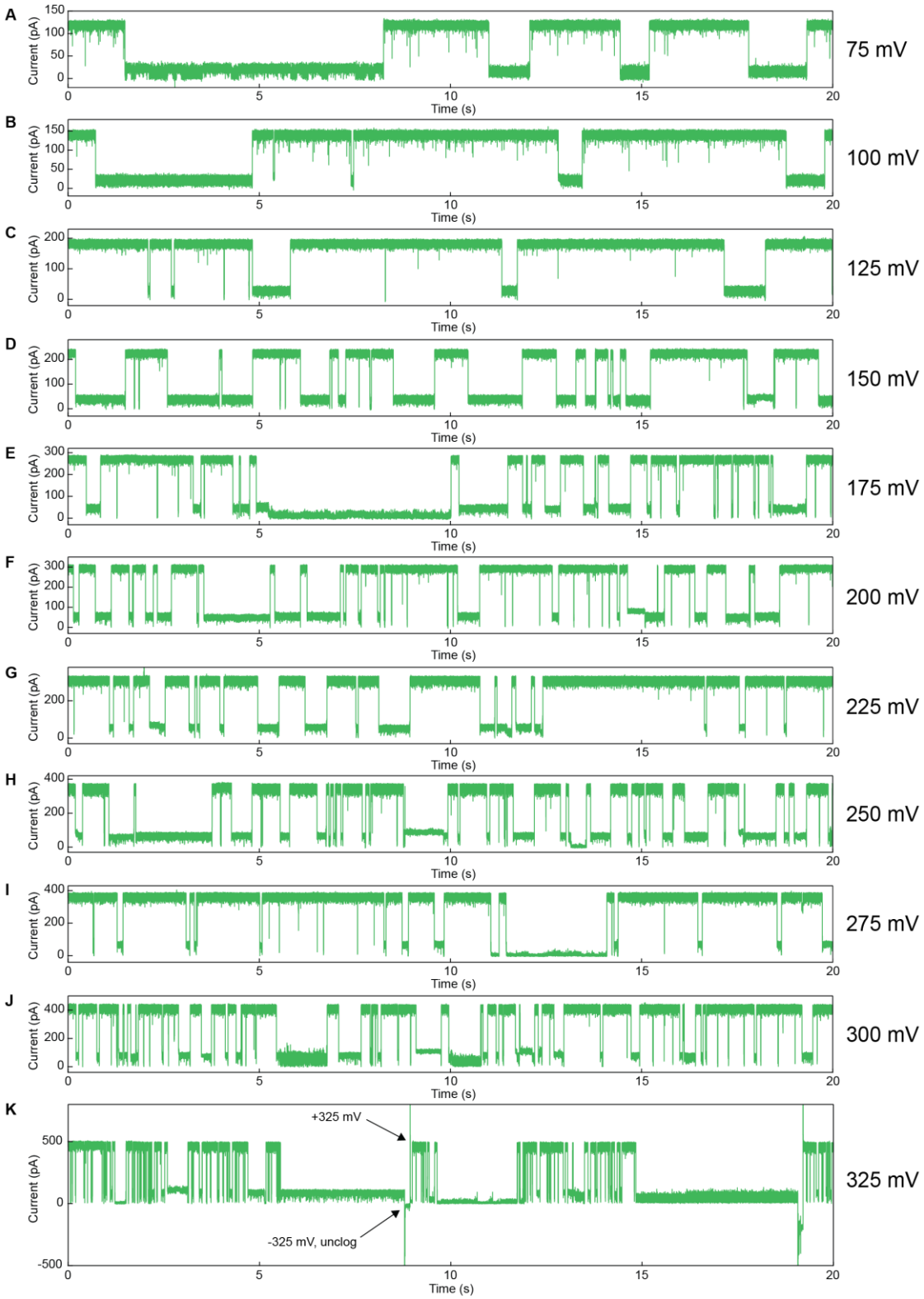


**Figure S4.** Current vs. time traces for MBP-D10 (0.7 µM) in 1.0 M GdmCl (1 M KCl, 10 mM Tris, pH 7.5), V = 75 – 325 mV, where panel K shows that when pore get clogged, switching voltage may help to unclog it.

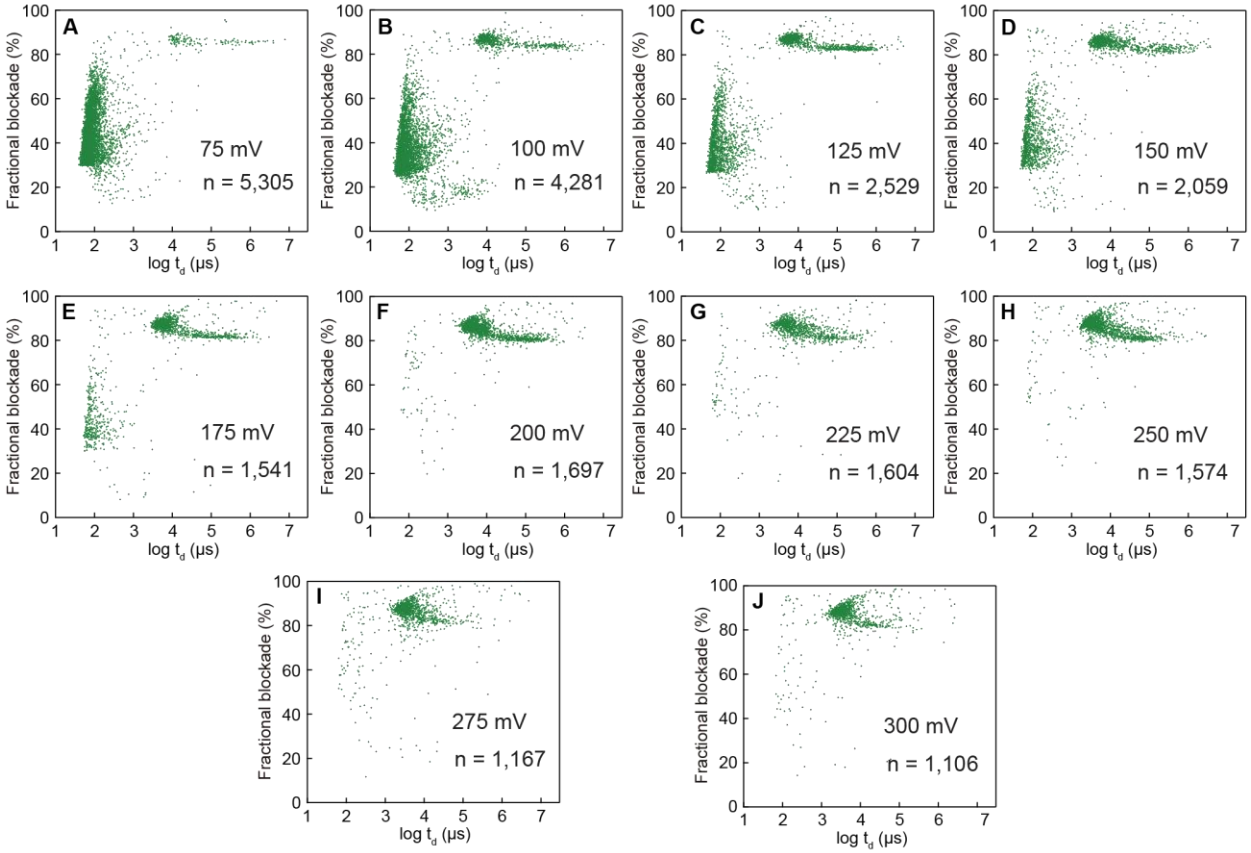


**Figure S5.** Fractional current blockade vs. dwell time scatters for MBP-D10 (0.35 μM) in 1.5 M GdmCl (1 M KCl, 10 mM Tris, pH 7.5), V = 75 – 300 mV.

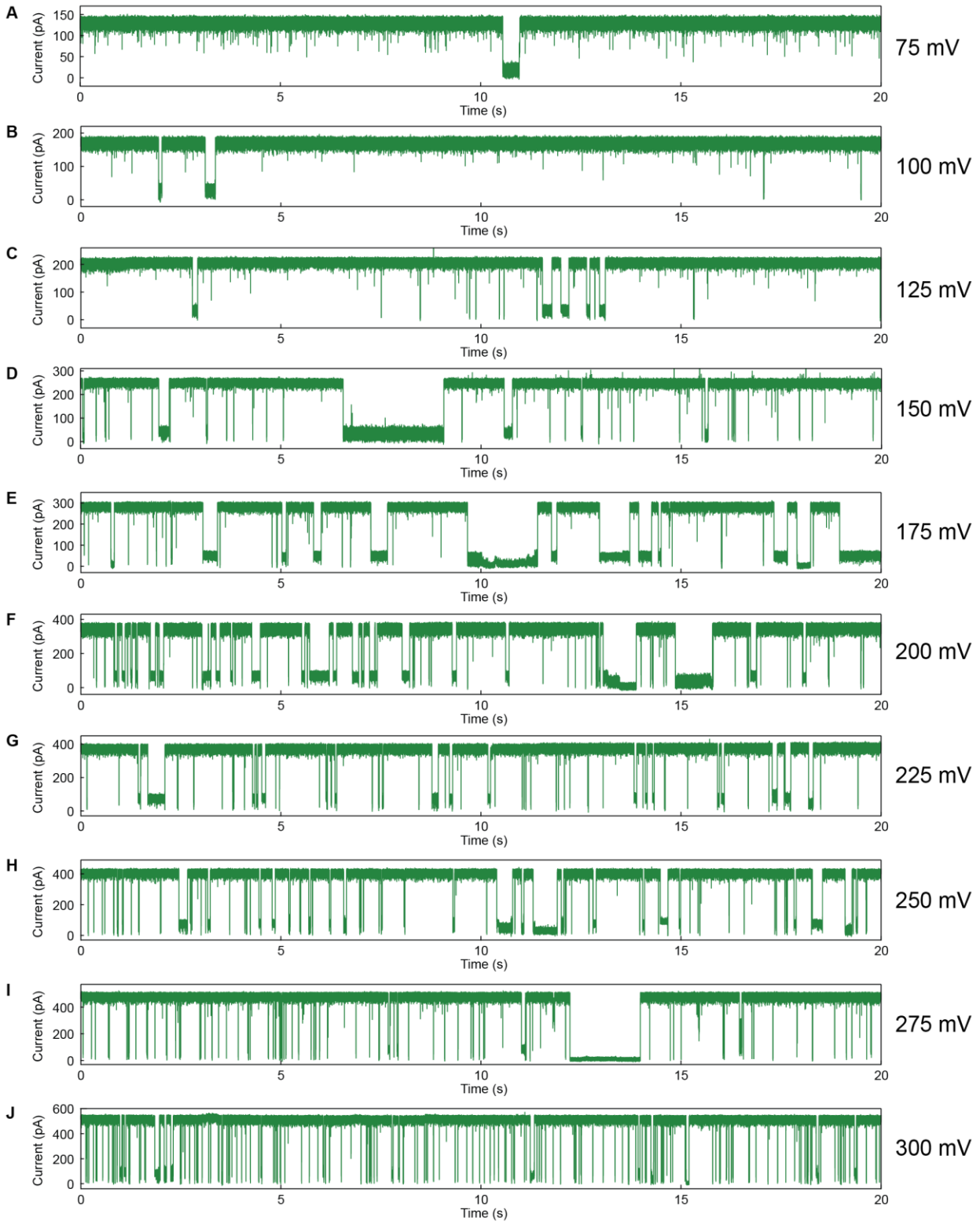


**Figure S6.** Current vs. time traces for MBP-D10 (0.35 µM) in 1.5 M GdmCl (1 M KCl, 10 mM Tris, pH 7.5), V = 75 – 300 mV.

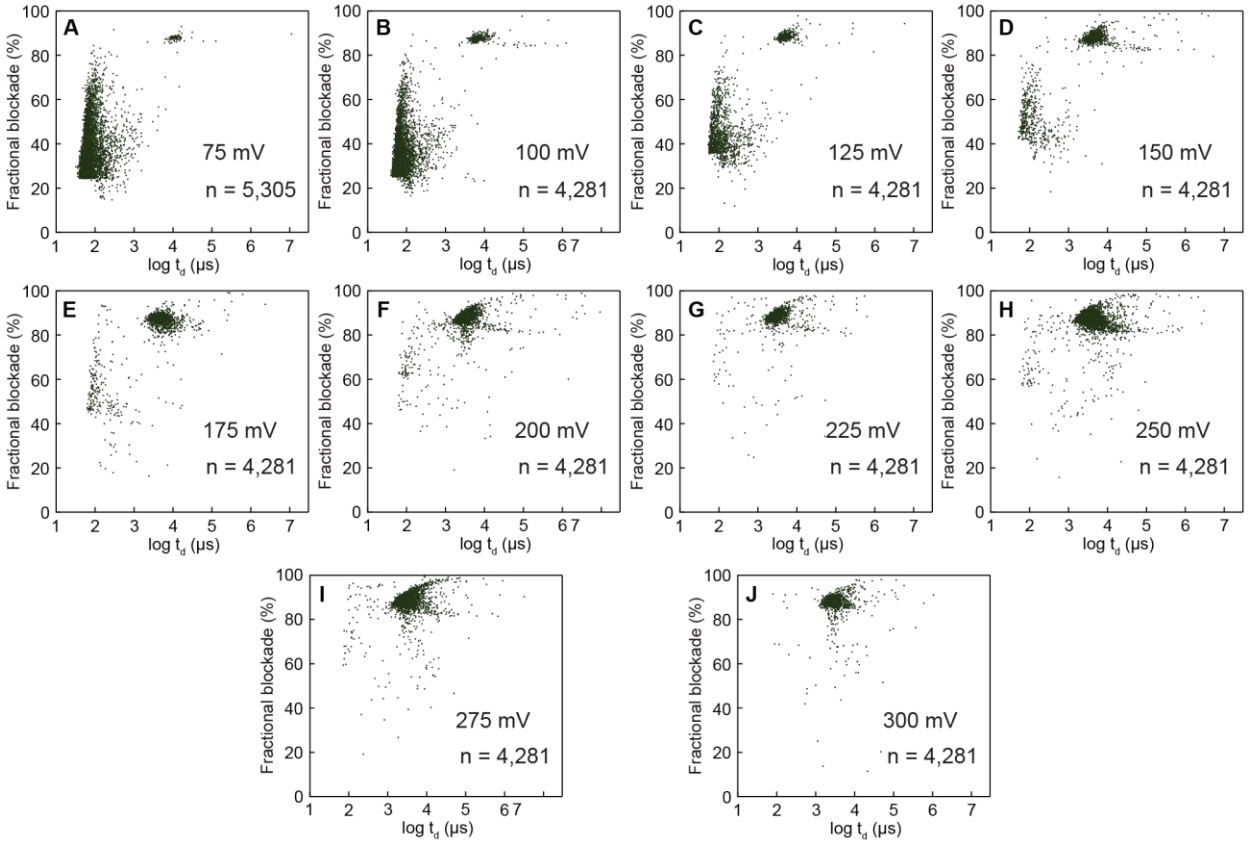


**Figure S7.** Fractional current blockade vs. dwell time scatters for MBP-D10 (0.35 µM) in 2.0 M GdmCl (1 M KCl, 10 mM Tris, pH 7.5), V = 75 – 300 mV.

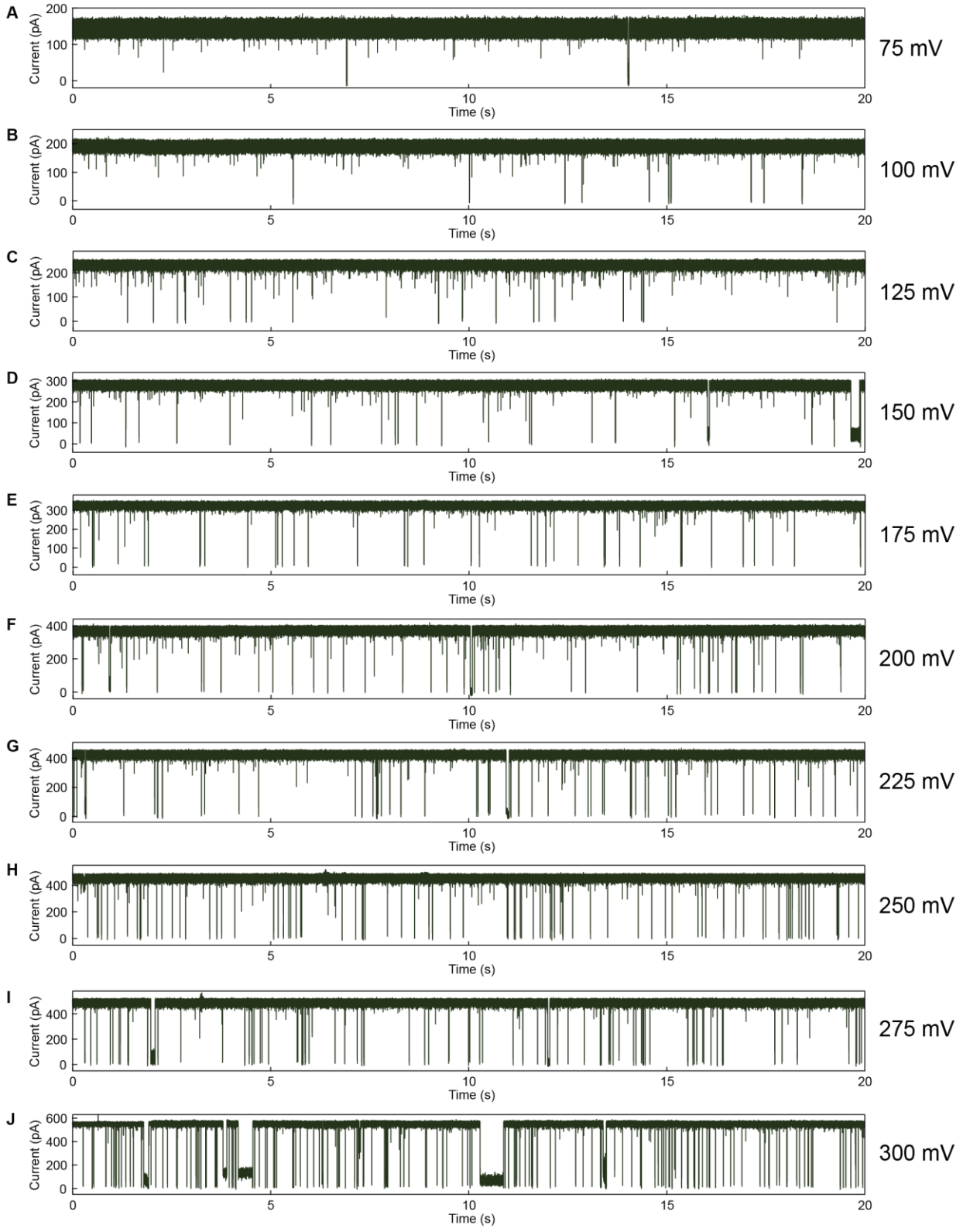


**Figure S8.** Current vs. time traces for MBP-D10 (0.35 µM) in 2.0 M GdmCl (1 M KCl, 10 mM Tris, pH 7.5), V = 75 – 300 mV.

## Section 3: Extra data and analysis for diMBP-D10

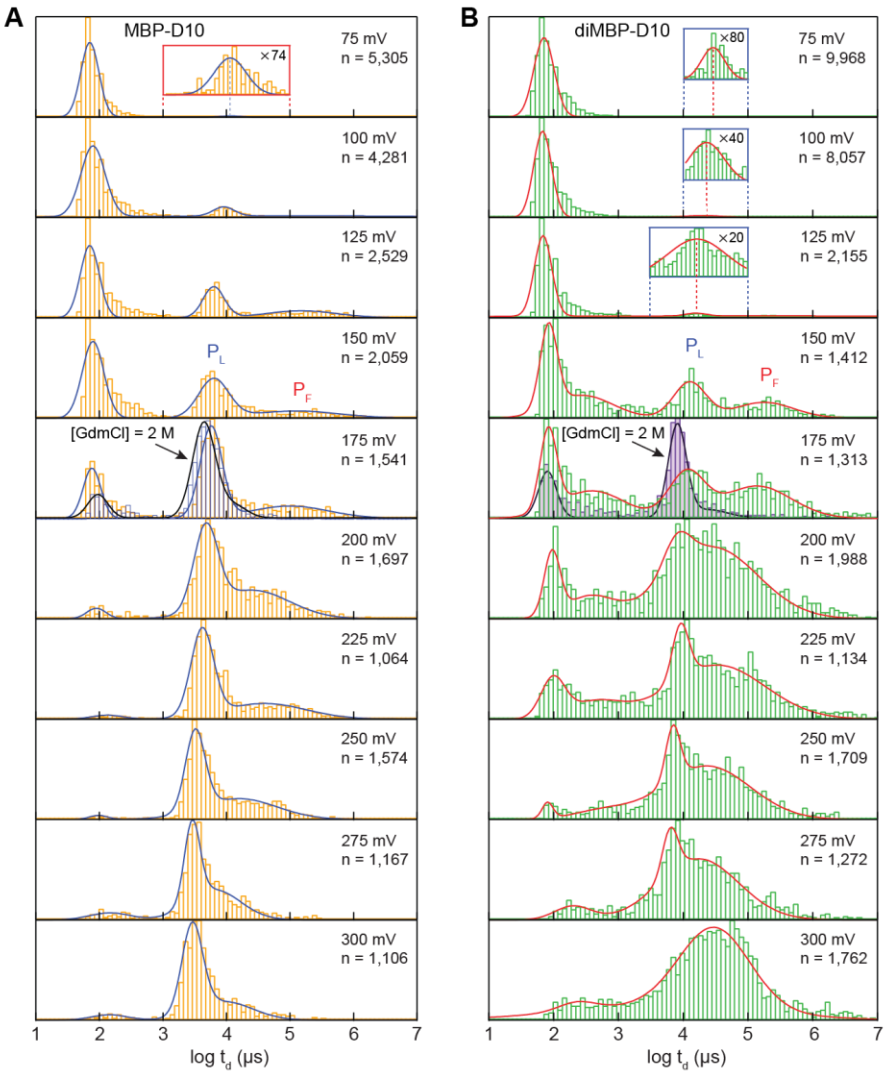


**Figure S9.** Dwell time histograms of **A)** MBP-D10 (0.35 μM) and **B)** diMBP-D10 (0.35 μM) at 75-300 mV and log-normal fits (1.5 M GdmCl, 1 M KCl). In the 75-175 mV voltage range, a significant fraction of events have ~100 μs dwell times, and as voltage increases further, a greater fraction of the 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). To highlight the impact of higher GdmCl concentrations on complete protein unfolding, we show dwell time distributions at 175 mV in 2 M GdmCl (1 M KCl) for both monomeric (n=1155) and dimeric (n=1151) MBP (blue and purple histograms, respectively).

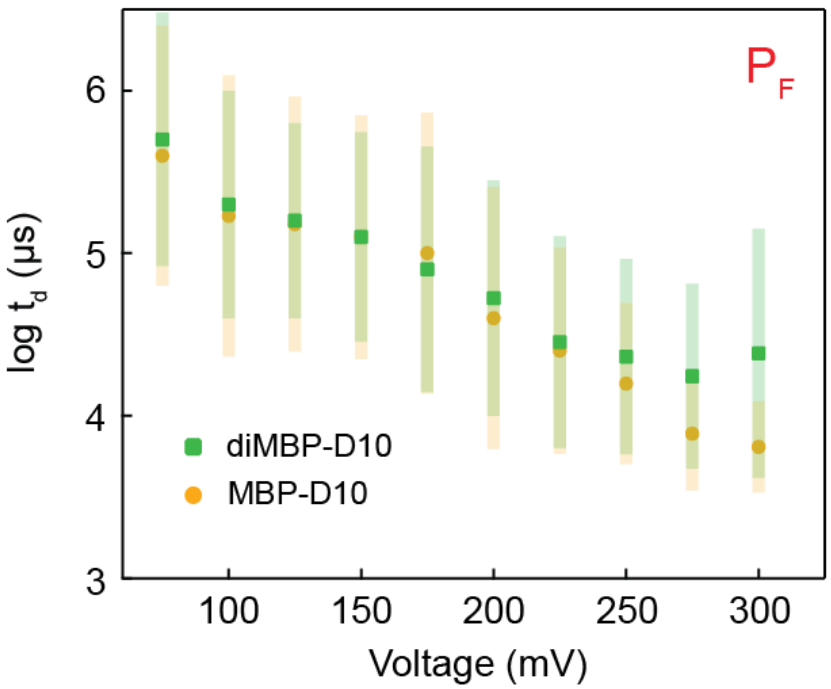


**Figure S10.** Mean dwell time vs. voltage for the $P_F$ populations of MBP-D10 and diMBP-D10, respectively (error bars represent the FWHM of the distribution fits). Buffer: 1 M KCl, 1.5 M GdmCl, 10 mM Tris, pH 7.5.

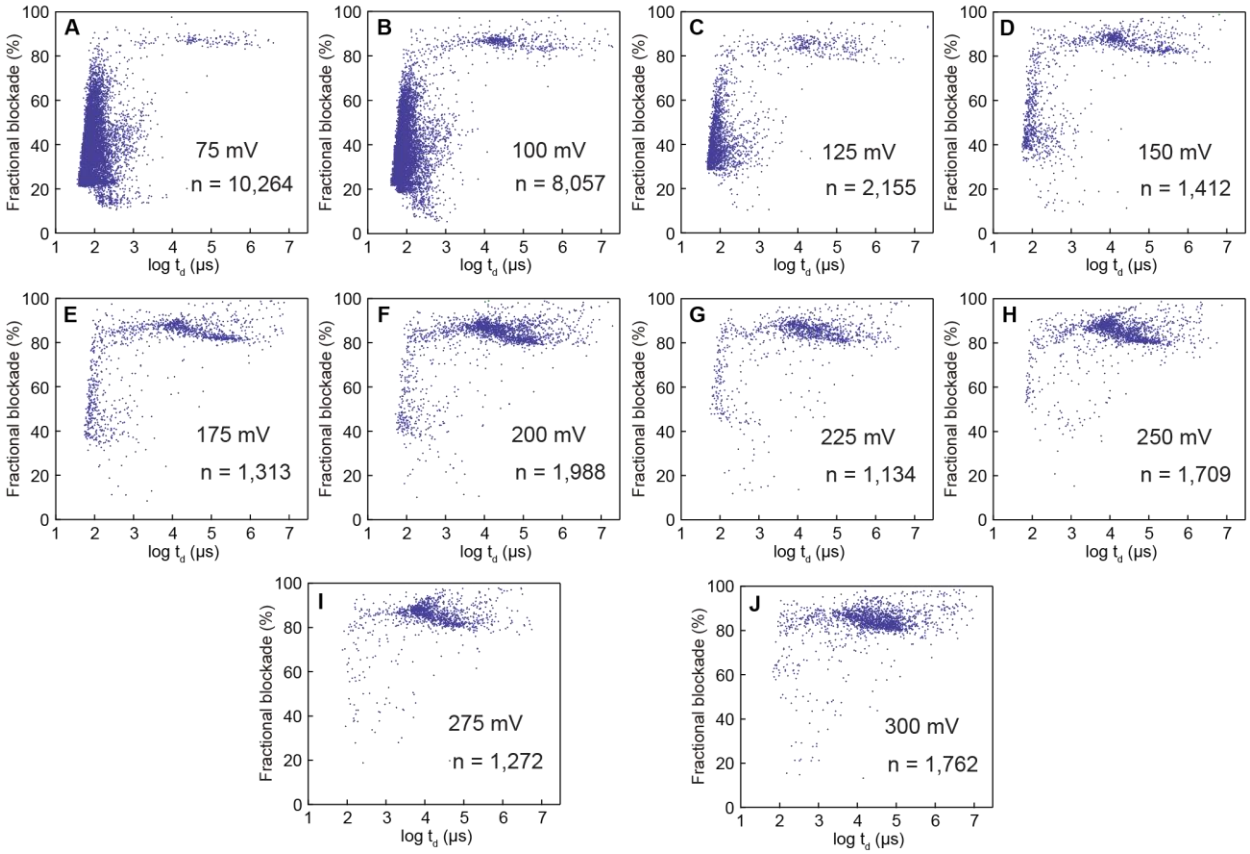


**Figure S11.** Fractional current blockade vs. dwell time scatters for diMBP-D10 (0.35 µM) in 1.5 M GdmCl (1 M KCl, 10 mM Tris, pH 7.5), V = 75 – 300 mV.

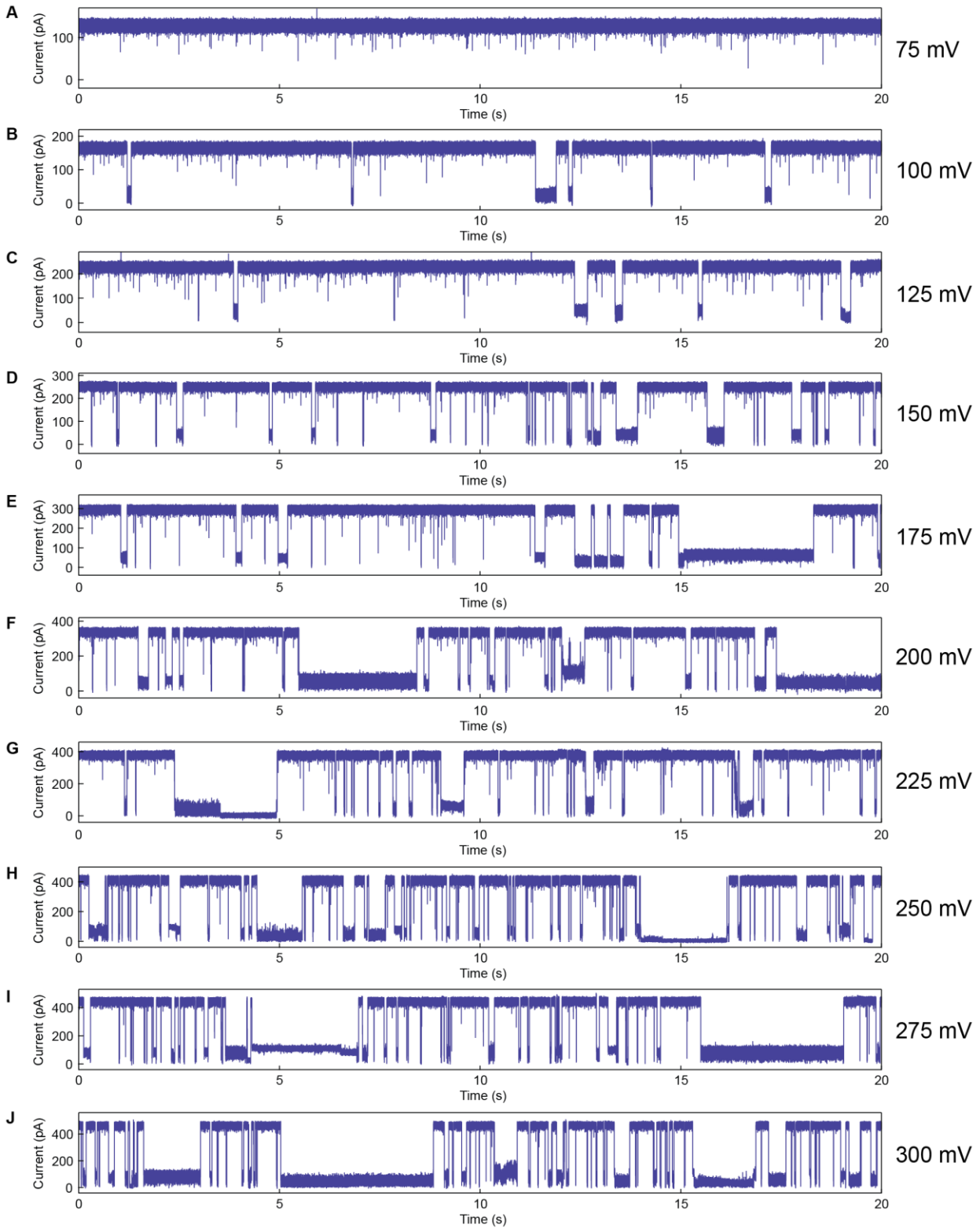


**Figure S12.** Current vs. time traces for diMBP-D10 (0.35 µM) in 1.5 M GdmCl (1 M KCl, 10 mM Tris, pH 7.5), V = 75 – 300 mV.

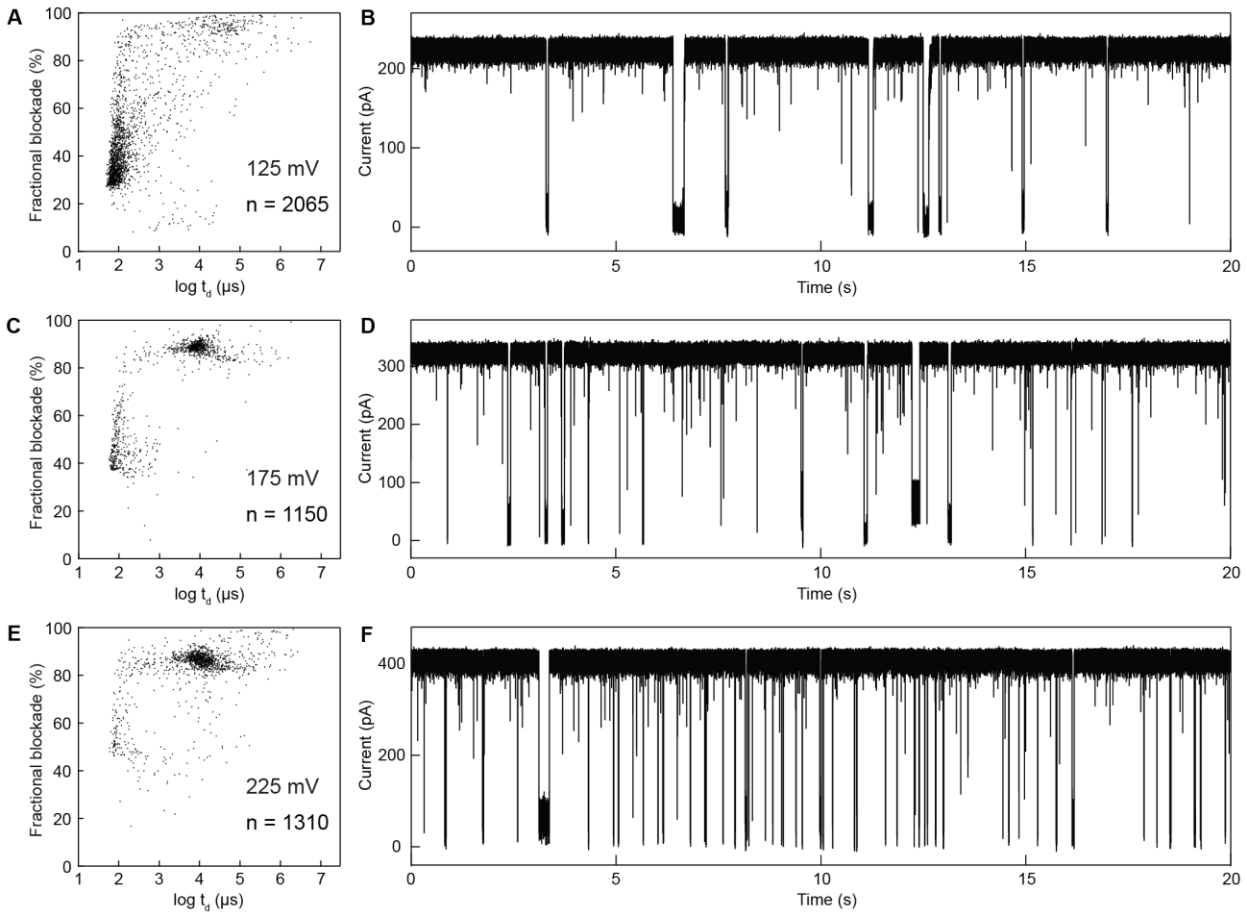


**Figure S13.** Fractional current blockade vs. dwell time scatters (left) and current vs. time traces (right) for diMBP-D10 (0.35 μM) in 2.0 M GdmCl (1 M KCl, 10 mM Tris, pH 7.5), V = 125, 175, 225 mV.

## Section 4: Other experimental data

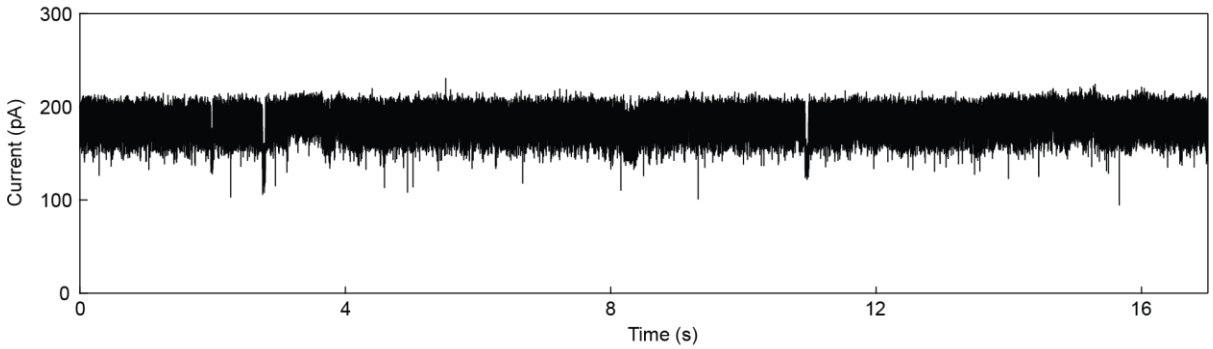


**Figure S14.** Current vs. time trace for for MBP-D10 (0.35 µM) without GdmCl in KCl, 10mM Tris, pH 7.5, V = 175 mV.

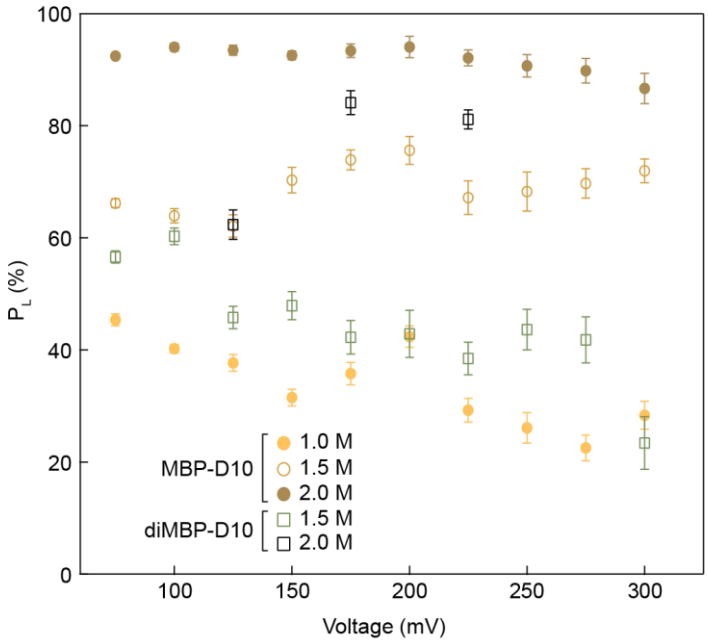


**Figure S15.** $P_L$ percentage vs. voltage for MBP-D10 (0.35 µM) in 1.0, 1.5, 2.0 M GdmCl and diMBP-D10 (0.35 µM) in 1.5, 2.0 M GdmCl (all buffers include 1 M KCl, 10 mM Tris, pH 7.5).

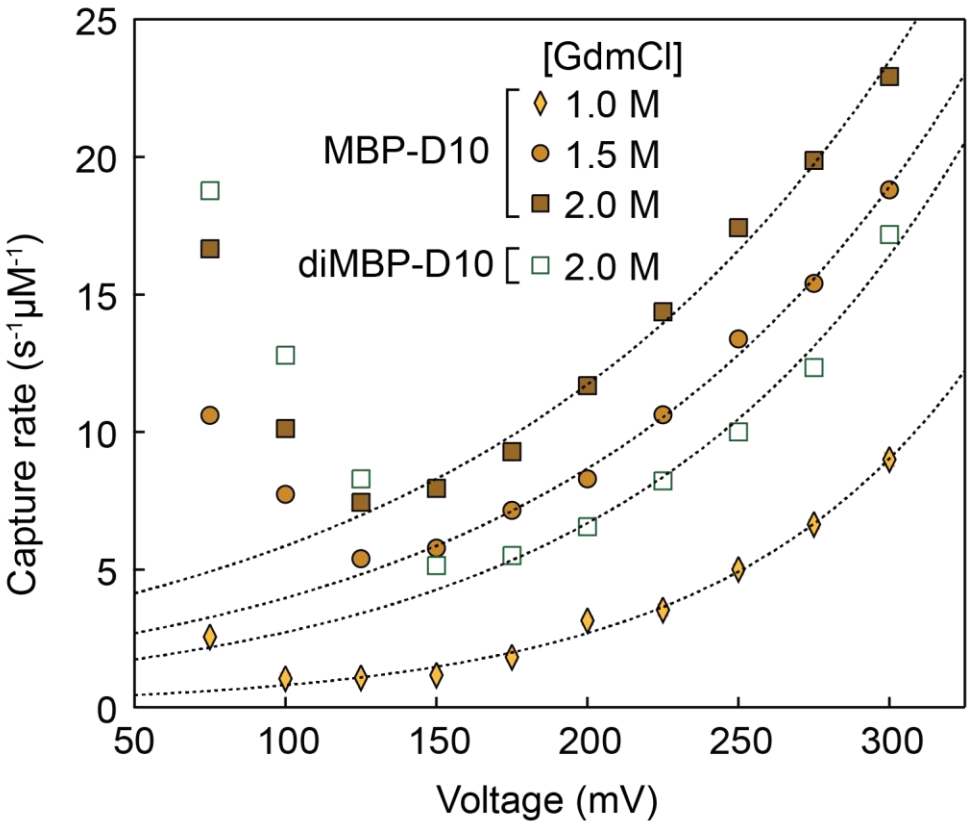


**Figure S16.** Concentration-normalized capture rates as a function of voltage for MBP-D10 (in 1.0 M, 1.5 M and 2.0 M GdmCl + 1 M KCl) and diMBP-D10 (in 1.5 M GdmCl + 1 M KCl). Error bars are smaller than the marker sizes in all cases. Dashed lines are exponential fits to the data in the range shown, where low voltage data was excluded (since the high capture rates mainly come from collisions between analytes and α-hemolysin vestibule).

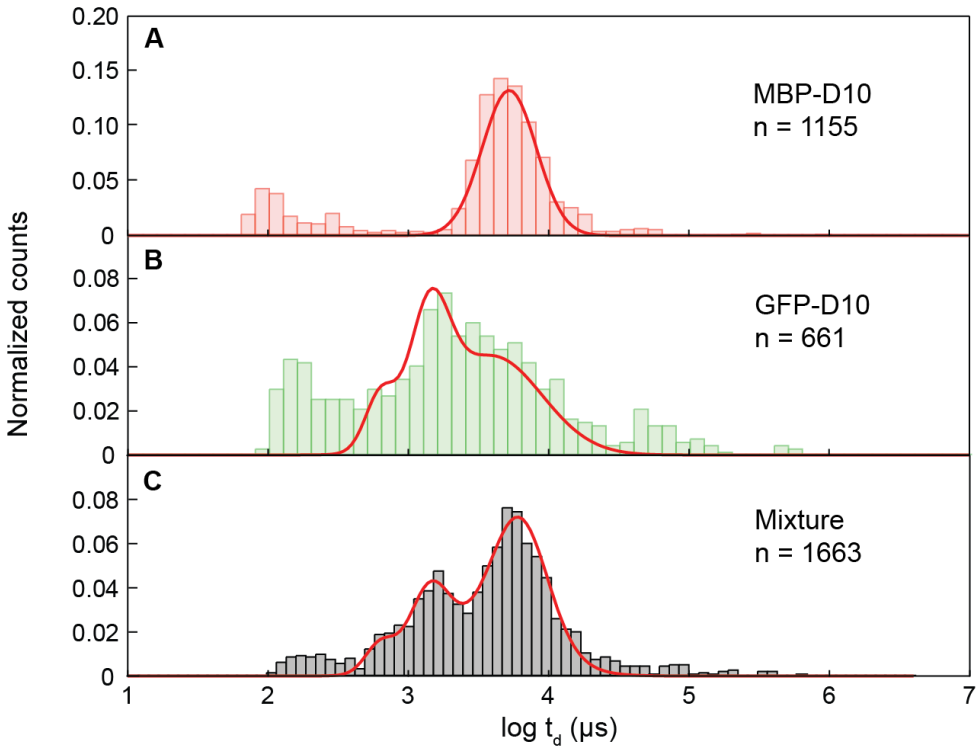


**Figure S17.** Dwell time histograms in 2 M GdmCl (1 M KCl, 10 mM Tris, pH 7.5) for **A)** 0.35 μM MBP-D10, **B)** 0.35 μM GFP-D10 and **C)** mixture of 0.35 μM MBP-D10 and 0.35 μM GFP-D10 with fitting lines.

### Supplementary Note 1: Estimate of the event ratio in the MBP-D10/GFP-D10 mixture

Only considering the translocation events with dwell time range $2.7 < \log t_d < 4.3$ (not considering the collisions and extremely long events), the histogram in **Figure S17, panel A** is fitted with multi-gaussian function (**Formula S1**). Similarly, since the GFP-D10 dwell time distribution is more complex, a triple-gaussian function (**Formula S2**) is used to fit the histogram in **Figure S17, panel B**. Since MBP-D10 and GFP-D10 events are independent of each other, **Formula S3** can be used to fit the histogram in **Figure S17, panel C** for the MBP-D10 and GFP-D10 mixture, where coefficients a and b represent the ratios of MBP-D10 and GFP-D10 events respectively. After fitting, a = 0.53 and b = 0.47 is obtained for the mixture, which shows that the MBP-D10 : GFP-D10 event frequency ratio is ~ 1:1.

$$f_1(x) = A\ exp\left[-\left(\frac{x-X}{W}\right)^2\right] \quad \textbf{(S1)}$$

$$f_2(x) = A_1 exp\left[-\left(\frac{x-X_1}{W_1}\right)^2\right] + A_2 exp\left[-\left(\frac{x-X_2}{W_2}\right)^2\right] + A_3 exp\left[-\left(\frac{x-X_3}{W_3}\right)^2\right] \quad \textbf{(S2)}$$

$$f(x) = a \times f_1(x) + b \times f_2(x) \quad \textbf{(S3)}$$

# Section 5: MD simulations

**Table S3.** Fragmentation of the maltose-binding protein (MBP) into seven peptides for MD simulations. Negatively charged resides are highlighted in blue and positively charged in red. The amino acid sequence of MBP was taken from the Protein Data Bank (PDB ID: 1JW4).

| Fragment | Residue ID | Sequence | Charge (e) |
|---|---|---|---|
| 1 | 1 to 53 | KIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTVEHPDKLEEKFPQVAAT | -1 |
| 2 | 54 to 106 | GDGPDIIFWAHDRFGGYAQSGLLAEITPDKAFQDKLYPFTWDAVRYNGKLIAY | -2 |
| 3 | 108 to 159 | PIAVEALSLIYNKDLLPNPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWP | -1 |
| 4 | 160 to 212 | LIAADGGYAFKYENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMNADTDYSI | -2 |
| 5 | 213 to 265 | AEAAFNKGETAMTINGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFVGVLSAG | 1 |
| 6 | 266 to 318 | INAASPNKELAKEFLENYLLTDEGLEAVNKDKPLGAVALKSYEEELAKDPRIA | -4 |
| 7 | 319 to 370 | ATMENAQKGEIMPNIPQMSAFWYAVRTAVINAASGRQTVDEALKDAQTRITK | 1 |

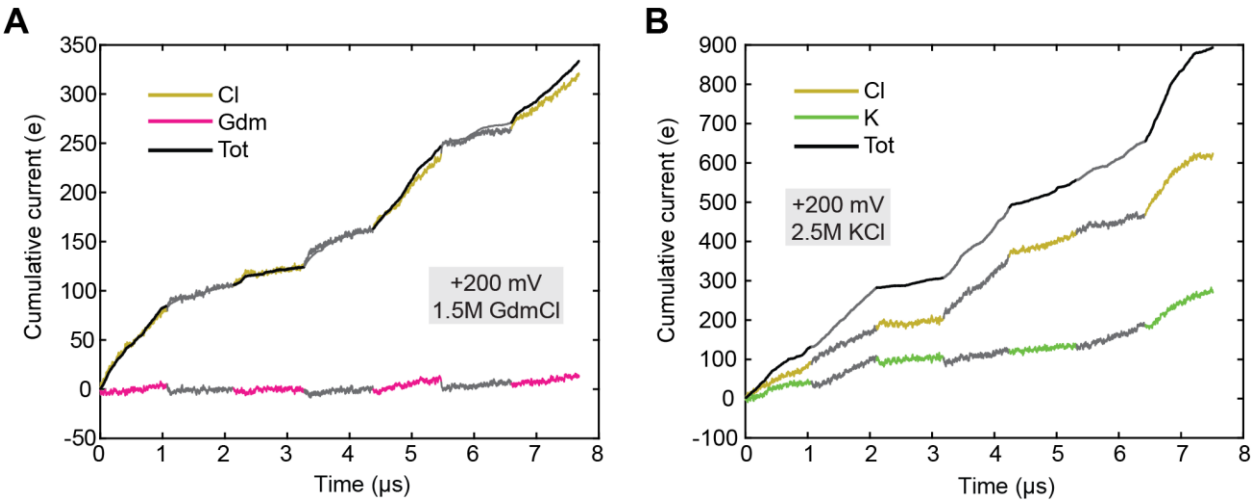


**Figure S18.** MD simulation of α-hemolysin systems containing fragments of MBP. Total charge carried by the ion species through the α-hemolysin nanopore is plotted versus simulation time for seven independent MD simulations carried out under a +200 mV bias for 1.5 M GdmCl **(A)** and 2.5 M KCl electrolyte **(B)** conditions. 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. The traces were added consecutively to appear as a continuous permeation trace.

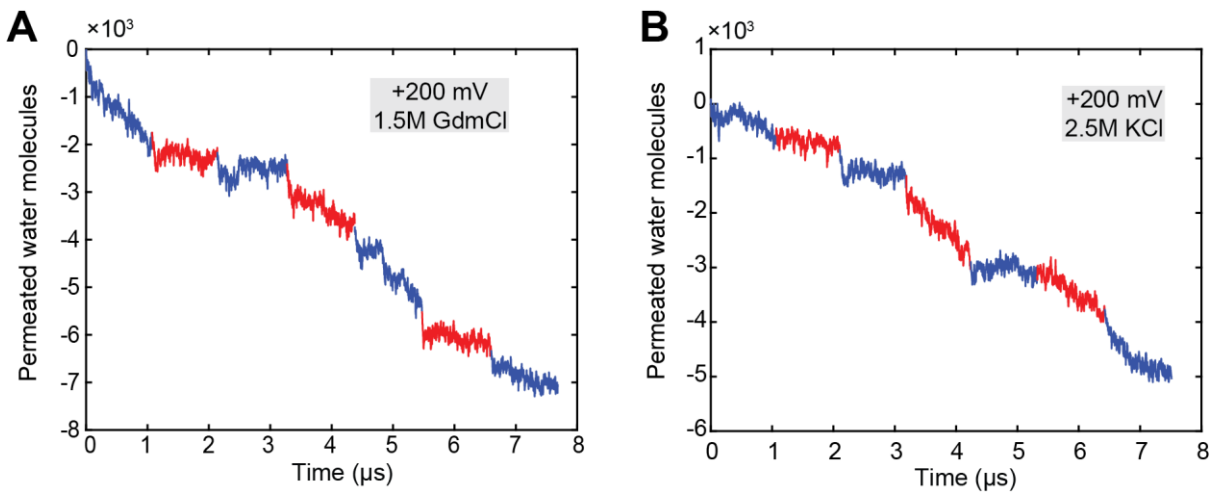


**Figure S19.** Simulated electro-osmotic flow in α-hemolysin systems containing fragments of MBP. The number water molecules permeated through the α-hemolysin constriction (residues 111, 113, and 147) is plotted as a function of simulation time for MD simulations carried out under a +200 mV bias for the 1.5 M GdmCl **(A)** and 2.5 M KCl **(B)** electrolyte conditions. Negative values indicate transport in the direction opposite that of the *z*-axis (defined in **Figure 3A**). Results for the seven independent simulations are shown using alternating colors. The traces were added consecutively to appear as a continuous permeation trace.

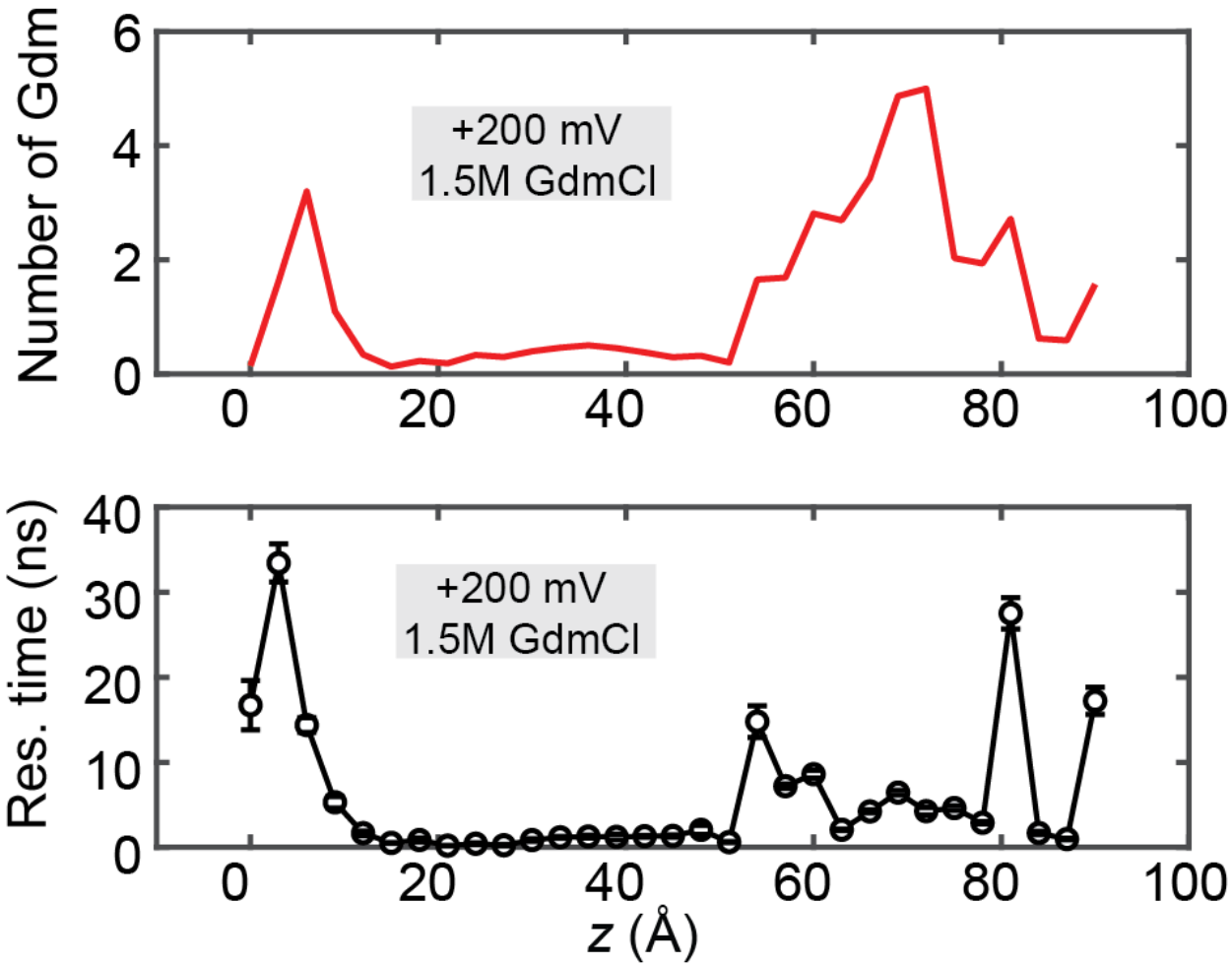


**Figure S20.** Gdm binding to the inner surface of α-hemolysin in MD simulations of the α-hemolysin systems containing MBP fragments. The top row shows the number of Gdm ions located within 3 Å of the nanopore inner surface. The bottom row shows the average residence time of the Gdm ions near the α-hemolysin nanopore surface. The *z*-axis is defined in **Figure 3A**. The data were averaged over the seven replica simulations at the same electrolyte condition.

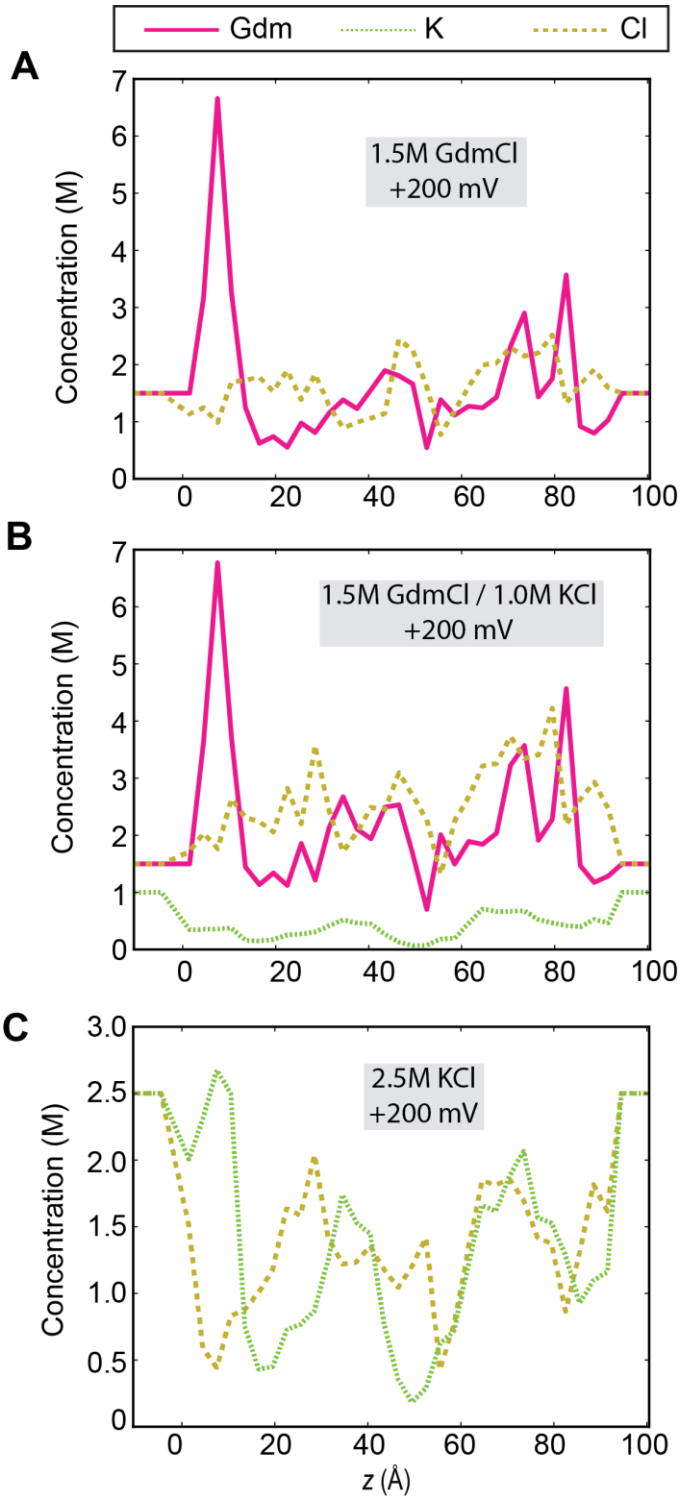


**Figure S21.** Local concentration of ions along the transmembrane nanopore observed in MD simulations of the α-hemolysin systems containing fragments of MBP. For each electrolyte condition, the concentrations were averaged the seven replica simulations in 3 Å bin along the nanopore axis (the *z* axis). The electrolyte and applied bias conditions are specified in each panel.

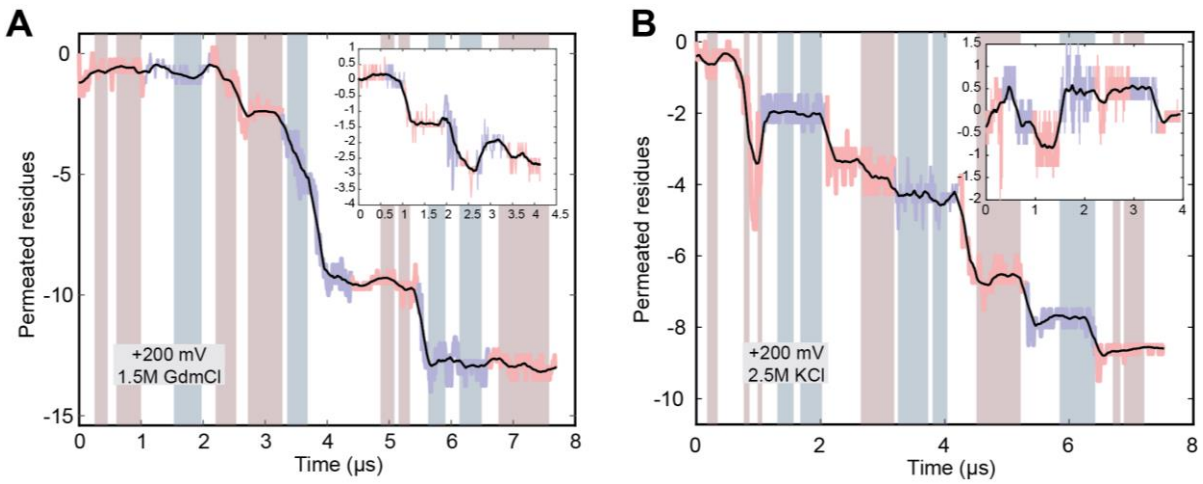


**Figure S22.** Simulated transport of MBP fragments. The number of amino acid residues permeated through the α-hemolysin constriction (residues 111, 113, and 147) is plotted as a function of simulations for seven independent MD simulations carried out under a +200 mV bias for the 1.5 M GdmCl **(A)** and 2.5 M KCl **(B)** electrolyte conditions. Negative values indicate transport in the direction opposite that of the *z*-axis (defined in **Figure 3A**). Results for the seven independent simulations are shown using alternating colors. The traces were added consecutively to appear as a continuous permeation trace. Highlights indicate the parts of the trajectories where the peptide density within the stem (8<z<45 Å) of α-hemolysin is constant; the inset shows consecutive addition of the highlighted regions.

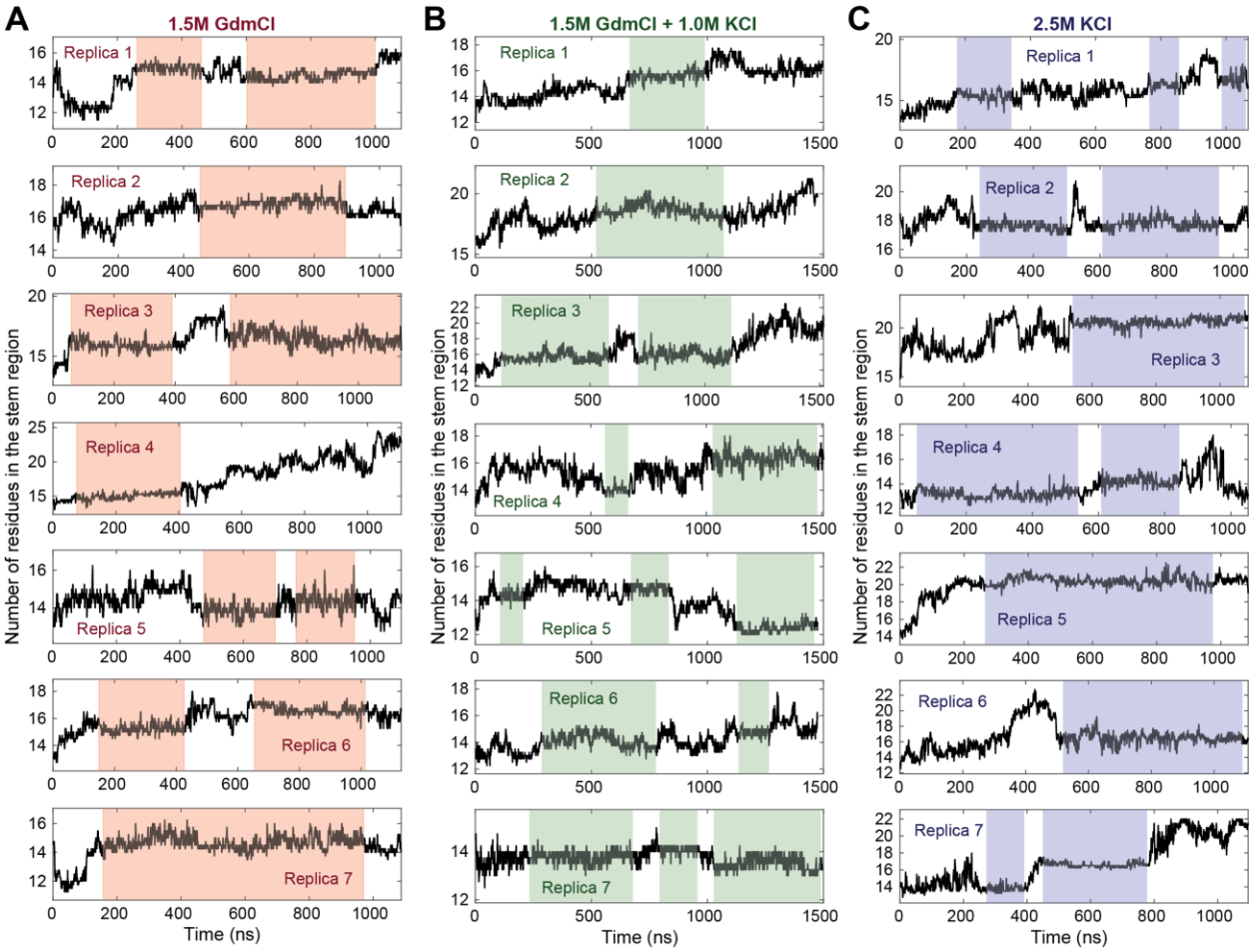


**Figure S23.** Number of residues confined within the stem region of α-hemolysin nanopore. Panels A, B and C correspond to simulations carried out at 1.5 M GdmCl **(A)**, 1.5 M GdmCl / 1.0 M KCl **(B)** and 2.5 M KCl **(C)** electrolyte solutions. Each plots shows data for one independent simulation. The highlights indicate the parts of the trajectories where the peptide density is nearly constant. These regions were used to calculate the peptide translocation rate displayed in **Figure 3H**.

## Section 6: Machine learning analysis

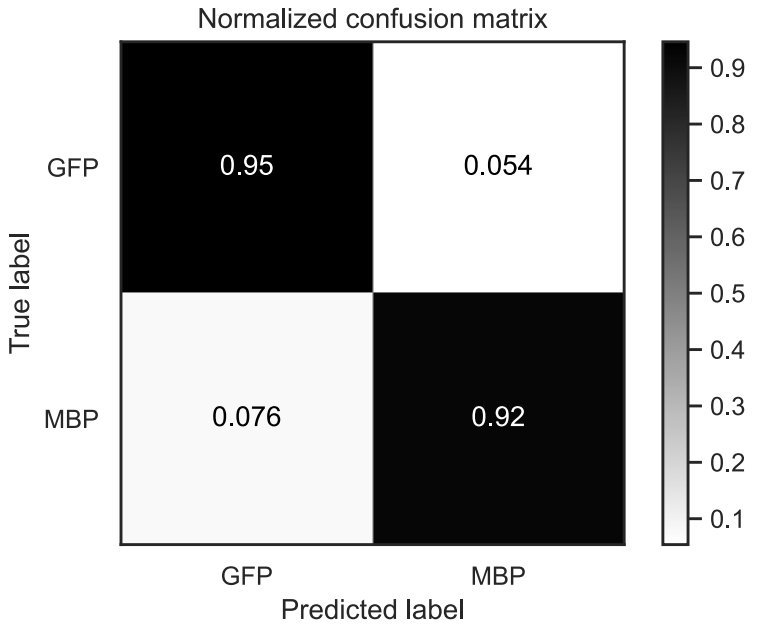


**Figure S24.** Confusion matrix of model used on mixture data. Shown here is the confusion matrix of the SVM model generated from labeled data that was later implemented for classification of unlabeled events in mixture experiments. After training, the model correctly predicted 70 out of 74 GFP-D10 (94.5%) events and 73 out of 79 MBP-D10 (92.4%) events, producing a classification accuracy of 93.4%.

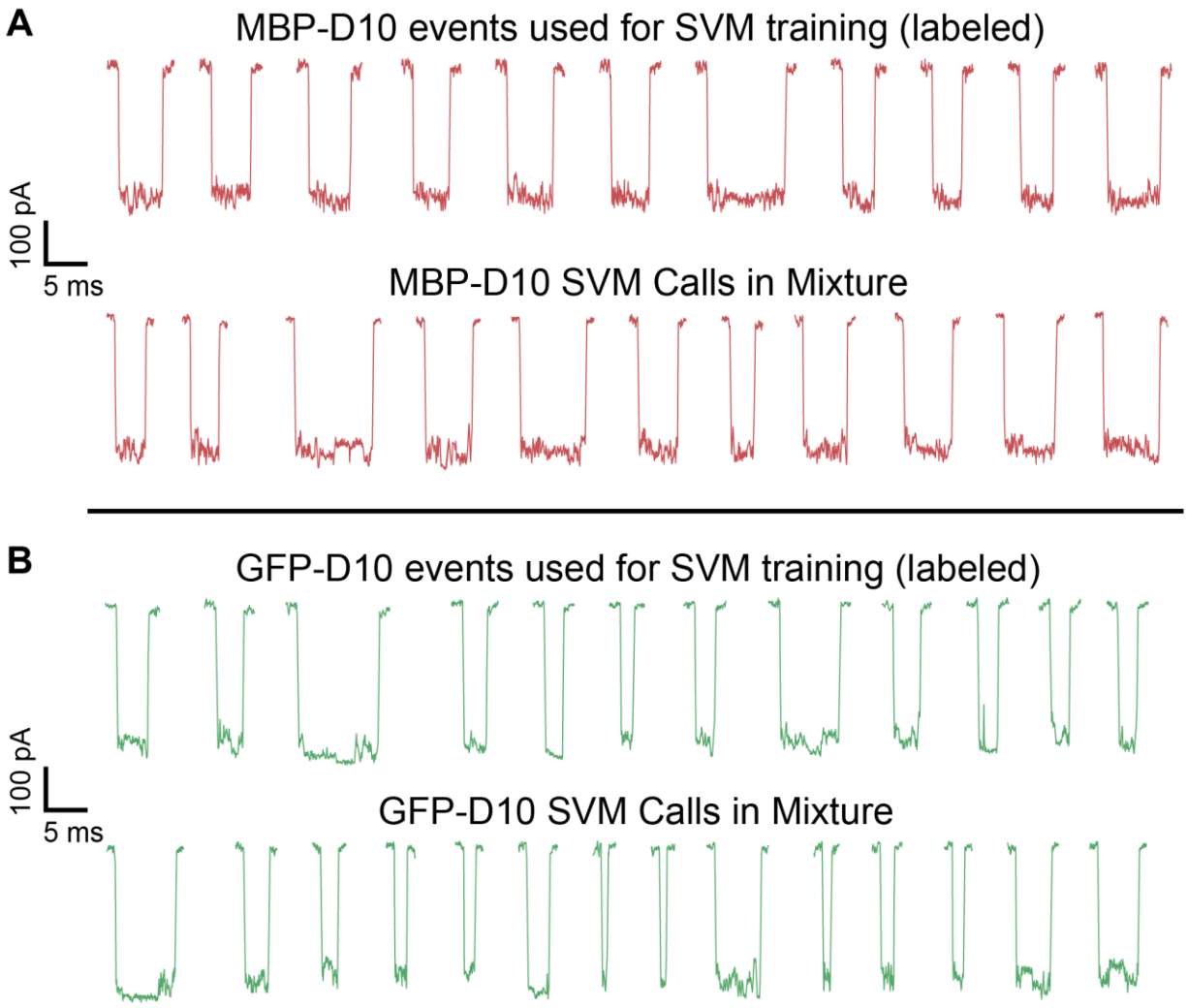


**Figure S25.** Event traces from pure MBP-D10 and GFP-D10 experiments used for training SVM model, where 51 input features from each event were used to construct SVM decision boundaries (upper). Traces of unlabeled translocation events from mixture experiment that were classified by trained SVM model (lower). 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.

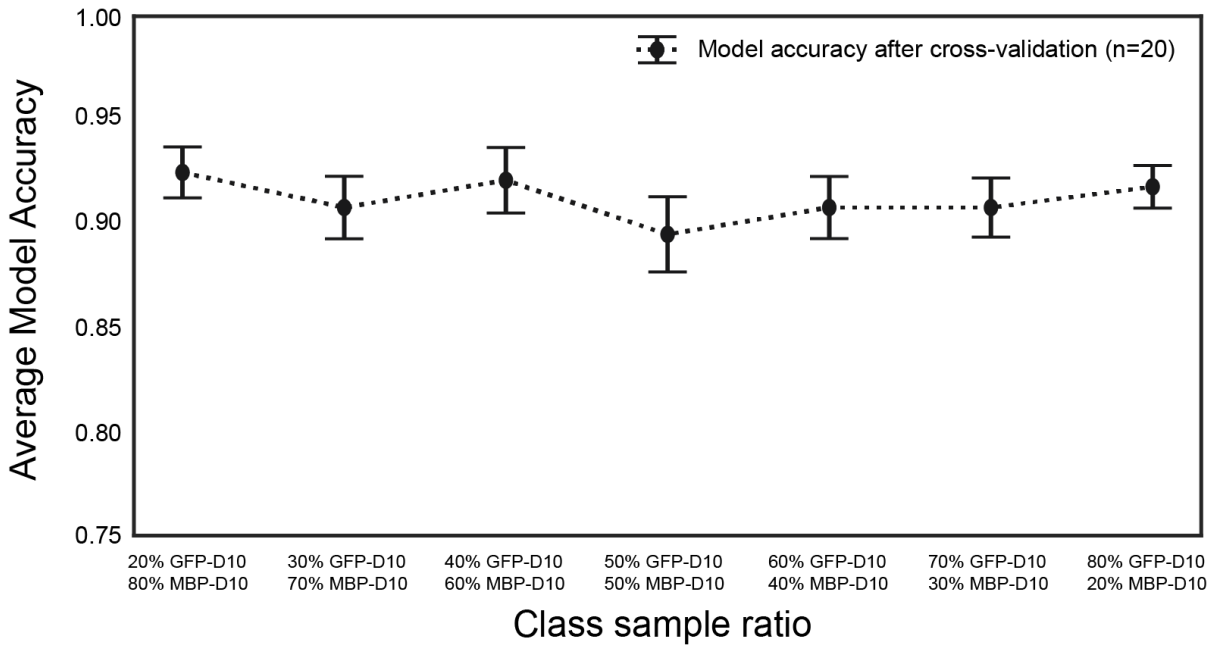


**Figure S26.** SVM classification accuracy as a function of class sample ratio. The balanced dataset used to generate the SVM for unlabeled mixture classification was made up of 381 GFP-D10 events and 381 MBP-D10. The SVM performance was further analyzed by truncating the same dataset to test the model under different ratios of MBP-D10 and GFP-D10. Prior to training and testing the model each time, different percentages of GFP-D10 and MBP-D10 events (shown on the x-axis of the figure) were randomly selected from the balanced dataset to generate a new, imbalanced sub dataset. For each imbalanced dataset, the SVM model was cross validated 20 times. The mean and standard error of the accuracy score each sub dataset is shown.

**Table S4.** The classification accuracy of the labeled events from each pure file after removing it from the dataset prior to SVM training and testing. The SVM was trained and tested on a balanced dataset that did not include any of the events from the excised file. The model was then saved and used to call the events in each pure/labeled dataset.

| File Name | Number of Events | Balanced Accuracy | | |
|---|---|---|---|---|
| | | MBP10D Calls | GFP10D Calls | Accuracy |
| MBP10D_B081920_006_trace_10kHz.bin | 367 | 332 | 35 | 90.00% |
| MBP10D_B081920_009_trace_10kHz.bin | 300 | 282 | 18 | 94% |
| MBP10D_B081920_007_trace_10kHz.bin | 101 | 89 | 12 | 88% |
| MBP10D_B081920_008_trace_10kHz.bin | 94 | 80 | 14 | 85.00% |
| | | | | |
| GFP10D_B081920_008_trace_10kHz.bin | 76 | 8 | 68 | 90% |
| GFP10D_B091520_003_trace_10kHz.bin | 48 | 3 | 45 | 93.70% |
| GFP10D_B091520_002_trace_10kHz.bin | 45 | 3 | 42 | 93% |
| GFP10D_B091520_000_trace_10kHz.bin | 43 | 2 | 41 | 95% |
| GFP10D_B091520_010_trace_10kHz.bin | 32 | 1 | 31 | 96.80% |
| GFP10D_B091520_007_trace_10kHz.bin | 27 | 0 | 27 | 100% |
| GFP10D_B091520_012_trace_10kHz.bin | 26 | 3 | 23 | 88.50% |
| GFP10D_B091520_009_trace_10kHz.bin | 25 | 3 | 22 | 88% |
| GFP10D_B091520_004_trace_10kHz.bin | 24 | 4 | 20 | 83.30% |
| GFP10D_B091520_005_trace_10kHz.bin | 19 | 1 | 18 | 94.70% |
| GFP10D_B091520_001_trace_10kHz.bin | 7 | 1 | 6 | 86% |
| GFP10D_B091520_011_trace_10kHz.bin | 6 | 0 | 6 | 100% |
| GFP10D_B091520_006_trace_10kHz.bin | 3 | 1 | 2 | 66% |

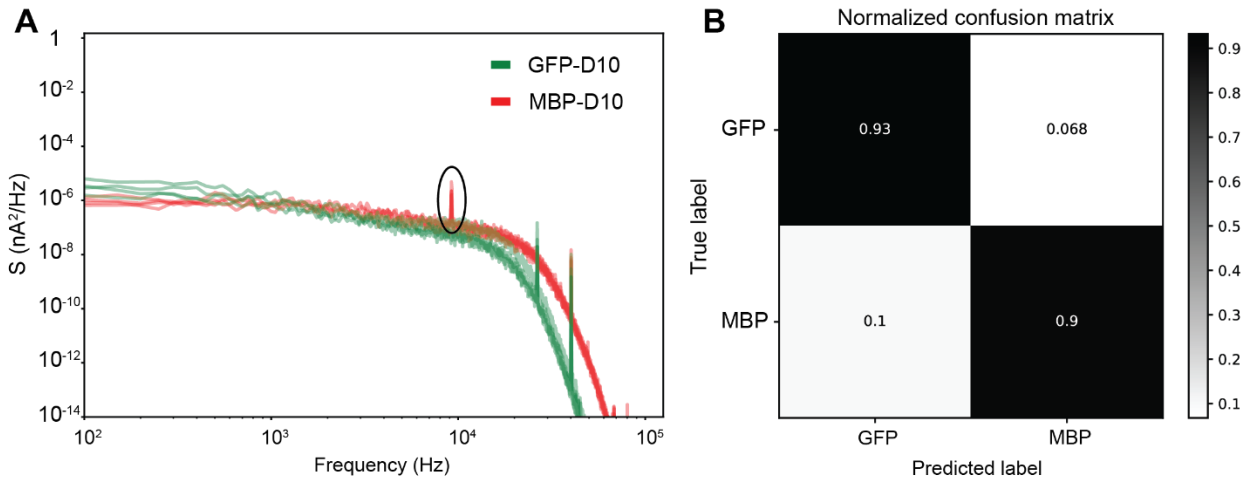


**Figure S27.** Classification Accuracy after comparing MBP-D10 and GFP-D10 baseline power spectral density. **A)** Power spectral density of the baseline for each experiment was extracted and superimposed; the circled spike located at around 9190 Hz is consistent in all 4 MBP-D10 experiments conducted in different days. **B)** The confusion matrix for the SVM model after a 2 kHz, 4 kHz, and 6 kHz low-pass Bessel filter was applied to all the events prior to computing the Fourier coefficients for subsequent training and testing.

**Table S5.** Sample size and event parameters used for $softDTW_\gamma$comparison of MBP-D10 and D10-MBP. Events with dwell times between 2 ms and 7 ms were used for this analysis and shown is the smoothing parameter $\gamma$ that was used in the $softDTW_\gamma$ function to generate the smooth average of each translocation signal.

| Type | Minimum Duration (ms) | Maximum Duration (ms) | Maximum current value (pA) | (m) Total number of selected events | $\gamma$ |
|---|---|---|---|---|---|
| MBP-D10 | 2 | 7 | 130 | 623 | 1.14 |
| D10-MBP | 2 | 7 | 130 | 111 | 1.66 |

## Supplementary Note 2: Soft-DTW Analysis

Turning to the signal content, we first analyzed the two cases where N-terminus vs. C-terminus entry of MBP were studied. The datasets used were MBP-D10 and D10-MBP (both 2 M GdmCl, 1 M KCl, 175 mV). To extract an "average shape" for MBP-D10 and D10-MBP events, barycenters (Fréchet means) were used. A barycenter or Frechet mean is a new target time-series $\vec{x}$ that produces the minimum sum of "distance metrics" to all elements $\vec{y}_{1\ldots m}$ in a time-series dataset $\boldsymbol{Y}$ evaluated one-by-one:

$$\min_{\vec{x}\in\Re^n} \sum_{i=1}^{m} \frac{1}{n\cdot m}\ softDTW_\gamma\ (\vec{x}\ , \overrightarrow{y_i}) \qquad \textbf{(S4)}$$

Where $n$ is the number of points in each time series, in this case resampled to a fixed length of 200 for both protein variants, and each $\vec{y}_i$ is an event's current array divided by the open pore current value $I_0$ to normalize for slight variations among different experiments. The function $softDTW_\gamma$ provides a smoothed dynamic time warping distance value between the two time-series inputs with the smoothing parameter $\gamma$ [1]. The result of this barycenter computation is a smooth curve, representing the centroid or the "essence" of the translocation events in the dataset. The analysis was performed using tools provided in the tslearn library (v 0.4.1) [2]. This computation was repeated for the two protein variants mentioned above. **Table S5** shows the restrictions applied to select the events, total events extracted, and the suggested smoothing parameter $\gamma$ calculated for each set using tslearn's tools.

The barycenters $\vec{x}_{MBPD10}$ and $\vec{x}_{D10MBP}$ are shown in **Figure 4A** along with 100 of their respective resampled events in the background. These curves show a smooth trend of how the events of each protein type tend to progress on average. While both samples show an overall downward trend in their barycenters, they distinctly have opposite locations for local maxima and minima. This DTW method was used to demonstrate unidirectional translocation. While the same soft-DTW metric function can be used for classification purposes, it was omitted in this study to keep the classification features used in the SVM analysis simple and computationally efficient.

We note that a similar "average shape" analysis could have been performed using signal segmentation and segment averaging across all events of each protein type. However, the time-domain variations across hundreds of events would lead to a highly flat and less informative average shape. This is mainly due to variations in translocation kinetics. The DTW analysis allowed us to minimize the effect of random time variations and rather focus on the overall trend when producing the average curve (barycenter).

## Supplementary Note 3: SVM Analysis

To explore 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 (2 M GdmCl, 1 M KCl, 175 mV) experiments and extended to unlabeled mixture experiments (1:1 ratio). Using the support vector machine (SVM) module provided by the library scikit-learn (v 0.22.2.post1) [3], an SVM classifier with a radial basis function (RBF) kernel was created and trained to discriminate MBP-D10 and GFP-D10 in a mixed sample nanopore experiment. Supervised training, cross validation, and labeled testing was done on pure MBP-D10 and GFP-D10 (2 M GdmCl, 1 M KCl, 175 mV) experiments. Translocation events with dwell times between 500us and 20ms were extracted, giving us 862 MBP-D10 events and 381 GFP-

D10 events for training and testing the model. Event extraction was done with a filter-derivative parsing method to remove event start and end borders (drops), allowing for meaningful computation of Fourier coefficients and event current maxima as SVM features. Since we were limited by the number of GFP-D10 events, we made our dataset balanced prior to training and testing by randomly selecting 381 MBP-D10 events. The model was trained on 80% of the data and was subsequently tested on the remaining 20%. The training set consisted of 609 total events, where 302 were MBP-D10 samples and 307 were GFP-D10 samples. The test set had 153 total events, where 79 were MBP-D10 samples and 74 GFP-D10 samples.

The following 51 input features were extracted from each translocation event from every pure experimental data file and were trained on to generate the SVM decision boundaries: The first 25 Fourier coefficients ($c_{1-25}$), mean and standard deviation values for event segments (where each event was segmented into 10 equally sized subsequences), minimum ion current value in the event divided by the maximum ($I_{min}/I_{max}$), mean ion current of the event divided by standard deviation ($I_{\mu}/I_{\sigma}$), maximum ion current divided by baseline current ($I_{max}/I_0$), minimum ion current divided by baseline current ($I_{min}/I_0$), minimum ion current divided by mean current ($I_{min}/I_{\mu}$), and minimum ion current divided by standard deviation ($I_{min}/I_{\sigma}$). Dwell time was not one of the 51 input features used generate the SVM to show that proteins with or without overlapping durations could be resolved. Using the preprocessing.StandardScaler.fit_transform function in scikit-learn [3], the training set from the data frame comprising the balanced data set with its 51 features was normalized prior to training and testing the SVM. The normalization parameters from the training set were used to normalize the test set (20%) and later the unlabeled mixture dataset using the sk.transform() function.

After training, the SVM model that was used on the unlabeled mixture data set had a classification accuracy of 93.4% after SVM training and testing (accuracy reported here came after testing on 20% of the pure data set), correctly calling 70/74 GFP-D10 testing samples (94.5%) and 73/79 MBP-D10 (93.2%) testing samples, which is shown in the confusion matrix **Figure S24**. Using scikit-learn's SVM probability function, the probability scores were computed to estimate the confidence of the classification for each translocation event that was used to test the SVM. The 73 correctly called MBP-D10 test samples had an average probability score of 94.3%, where 59 samples scored above 90%, 4 samples scored between 80-90%, 2 samples scored between 70-80%, and 8 scored below 70%. The 70 correctly called GFP-D10 test samples had an average probability score of 94.5%, where 58 samples scored above 90%, 4 samples scored between 80-90%, 4 samples scored between 70-80%, and 4 scored below 70%. To ensure that our model accuracy was not due to a lucky train-test split, the SVM was cross-validated by randomly shuffling the order of the events in the dataset prior to the 80/20 split and fitting the model independently 20 times. The mean classification accuracy is 92% with a standard deviation of 4%.

To further evaluate the performance of the SVM model, we conducted three different analyses. The first analysis was to look at the accuracy of the SVM as a function of class size for training and testing. We took the balanced dataset of MBP-D10 and GFP-D10 (381 and 381 events each) and randomly removed events of either class to create an imbalance sub dataset for SVM training and testing. We generated imbalanced datasets with 20% GFP-D10, 80% MBP-D10; 30% GFP-D10, 70% MBP-D10; 40% GFP-D10, 60% MBP-D10; 50% GFP-D10, 50% MBP-D10; 60% GFP-D10, 40% MBP-D10; 70% GFP-D10, 30% MBP-D10; 80% GFP-D10, 20% MBP-D10. For consistency, the SVM for each imbalanced subset had an 80/20 split. Additionally, each dataset was randomly shuffled 20 times prior to division into train and tests sets for cross-validation. The mean accuracy and standard error of each cross-validated SVM is shown in **Figure S26**. The results show that the SVM accuracy decreases as it becomes more balanced, however, all 7 subsets had a mean accuracy greater than 85%.

For the second analysis, we wanted to test the performance of the SVM model on labeled MBP-D10 and GFP-D10 events. We removed, one by one, all the events associated with each pure experimental file from the dataset and trained the SVM. Next, we took the trained model and tested it on the file that was removed. The results in **table S4** show that the SVM maintains high accuracy (85% and greater) for each experiment file, confirming that the 51 features, including the first 25 Fourier coefficients of each event, captures enough information about the profile of the analyte to reliably distinguish different proteins.

For the third analysis, we considered the possibility that experiment-specific noise during recording could arise as the dominant distinction when using the first 25 Fourier coefficients of each event to fit the SVM classifier. To demonstrate that noise leakage is not determining the hyperplane of the SVM, and instead it is fitting to signal characteristics of the protein being sensed at the pore, we generated the power spectral density (PSD) of the baseline noise for every MBP-D10 and GFP-D10 experiment used for SVM training. Shown in **Figure S27 A**, superimposing the PSD of the 4 MBP-D10 experiments with 4 GFP-D10, a distinct and consistent spike was observed at around 9190 Hz in all the MBP-D10 experiments that was not present in any of the GFP-D10 experiments. To test whether this artifact contributes to the classification accuracy of the SVM (since we are implementing the first 25 Fourier coefficients of each event as a feature), we trained and tested the model after attenuating this artifact by applying a 2 kHz, 4 kHz, and 6 kHz digital low-pass Bessel filter to each event prior to the extraction of the Fourier coefficients. All three models generated the same accuracy and confusion matrix, shown in **Figure S27 B**. Compared to the confusion matrix of the model that was trained on events that were filtered at 10 kHz during the experiments (**Figure S24**), the reduction in accuracy is minimal.

The model was saved and implemented on the unlabeled mixed class experiment. These were the results: 1,535 translocation events were parsed and 1,142 of those events were within the 500us and 20ms dwell time range used to train the SVM model; for consistency, these were events we classified for this analysis. The mixed sample experiment had a 1:1 ratio of MBP-D10 and GFP-D10, 559 of the events were classified as MBP-D10 (49%) and 583 were classified as GFP-D10 (51%). It is not intuitive that the event ratio would line up with the mixture concentration (1:1) because the capture efficiency could be different for the two proteins in the mixture. Hence, to corroborate these results we took the dwell time histograms of GFP-D10 and MBP-D10 from the pure experiments and fit each one to its own Gaussian function (**Formulas S1 and S2**). The functions were then linearly combined and weighted to fit the dwell time histogram of the mixture dataset. Based on the results of the histogram analysis, shown in **Figure S17**, 53% of the distribution is made up of GFP-D10 and 47% is made up of MBP-D10, which is consistent with the SVM results. Note that the slight difference between the event counts in the SVM analysis and histogram analysis in **Figure S17** is due to usage of a different event extractor algorithm. The MBP-D10 calls had an average probability confidence score of 85%, where 284 samples scored above 90%, 106 samples scored between 80-90%, 64 samples scored between 70%-80%, and 105 scored below 70%. The GFP-D10 calls had an average probability confidence score of 85% as well, where 313 samples scored above 90%, 73 samples scored between 80%-90%, 78 samples scored between 70-80%, and 119 scored below 70%.

## References


1 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).

2 Tavenard, R. *et al.* Tslearn, A Machine Learning Toolkit for Time Series Data. *J. Mach. Learn. Res.* **21**, 1-6 (2020).

3 Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825-2830 (2011).