

# Multi-pass, single-molecule nanopore reading of long protein strands

<https://doi.org/10.1038/s41586-024-07935-7>

Received: 28 June 2023

Accepted: 9 August 2024

Published online: 11 September 2024

Open access

 Check for updates

Keisuke Motone<sup>1,4,5</sup>, Daphne Kontogiorgos-Heintz<sup>1,5</sup>, Jasmine Wee<sup>1</sup>, Kyoko Kurihara<sup>1</sup>, Sangbeom Yang<sup>1</sup>, Gwendolin Roote<sup>1</sup>, Oren E. Fox<sup>1</sup>, Yishu Fang<sup>1</sup>, Melissa Queen<sup>1</sup>, Mattias Tolhurst<sup>2</sup>, Nicolas Cardozo<sup>2</sup>, Miten Jain<sup>3</sup> & Jeff Nivala<sup>1,2</sup>✉

The ability to sequence single protein molecules in their native, full-length form would enable a more comprehensive understanding of proteomic diversity. Current technologies, however, are limited in achieving this goal<sup>1,2</sup>. Here, we establish a method for the long-range, single-molecule reading of intact protein strands on a commercial nanopore sensor array. By using the ClpX unfoldase to ratchet proteins through a CsgG nanopore<sup>3,4</sup>, we provide single-molecule evidence that ClpX translocates substrates in two-residue steps. This mechanism achieves sensitivity to single amino acids on synthetic protein strands hundreds of amino acids in length, enabling the sequencing of combinations of single-amino-acid substitutions and the mapping of post-translational modifications, such as phosphorylation. To enhance classification accuracy further, we demonstrate the ability to reread individual protein molecules multiple times, and we explore the potential for highly accurate protein barcode sequencing. Furthermore, we develop a biophysical model that can simulate raw nanopore signals *a priori* on the basis of residue volume and charge, enhancing the interpretation of raw signal data. Finally, we apply these methods to examine full-length, folded protein domains for complete end-to-end analysis. These results provide proof of concept for a platform that has the potential to identify and characterize full-length proteoforms at single-molecule resolution.

Annotating the complexity of protein variation is important for understanding biological processes, identifying disease states and developing effective therapeutics. Proteoform diversity refers to the vast array of protein variations that can exist because of differences in transcription, translation and post-translational modifications (PTMs), which can occur through enzymatic (such as phosphorylation) and non-enzymatic (for example, spontaneous deamidation) processes<sup>5</sup>. These variations occur independently and in combination with each other on single protein molecules, creating a ‘PTM code’ that has unique and specific roles in driving biological processes<sup>6–8</sup>. The ability to sequence single protein molecules in their natural, full-length state could elucidate this proteoform diversity and its underlying code. However, current methods, including Edman degradation and mass spectrometry, have difficulty analysing full-length proteins from complex samples and face problems with detection sensitivity, dynamic range, analytical throughput and instrumentation cost<sup>1,2</sup>. To address these challenges, complementary or potentially disruptive platforms for next-generation protein analysis and sequencing have been proposed, including single-molecule fluorescence labelling and affinity-based approaches<sup>9–13</sup>. However, these emerging techniques also have limitations compared with nanopore technology<sup>14</sup>, which has the potential to achieve direct, label-free, full-length protein sequencing<sup>15</sup>.

Nanopore technology consists of a nanometre-sized pore in an insulating membrane that separates two electrolyte-filled wells<sup>16</sup>. A voltage applied across the membrane drives ionic current flow through the nanopore sensor. When individual analyte molecules pass through the pore, they can generate a detectable change in the signal. This change can provide insight into the molecular nature of the analyte. Although originally viewed, and now commercialized, as a technique for sequencing nucleic acid strands, nanopore sensing has great potential for protein analysis<sup>16,17</sup>. It has been used for the discrimination of peptides and proteins<sup>18–28</sup>, the real-time measurement of protein–protein<sup>29</sup> and protein–ligand interactions<sup>30</sup>, and aptamer-mediated protein detection<sup>30,31</sup>. Protein nanopores have also shown promise in identifying amino acids and PTMs, such as those involved in phosphorylation and glycosylation<sup>32–35</sup>, that serve as important biomarkers of cell states and diseases. Previous studies have demonstrated some ability to read DNA-conjugated peptide strands using DNA-processive molecular motors, such as a helicase or a polymerase<sup>36–38</sup>. Furthermore, rereading of peptide fragments using this strategy have made it possible to resolve among a small subset of single amino acid substitutions with high accuracy<sup>37</sup>. Despite this progress, using nanopores to obtain sequence information from intact, full-length proteins has been hindered by the difficulty of driving long protein strands through the sensor, owing to the neutrally

<sup>1</sup>Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA. <sup>2</sup>Molecular Engineering and Science Institute, University of Washington, Seattle, WA, USA. <sup>3</sup>Department of Bioengineering, Department of Physics, Northeastern University, Boston, MA, USA. <sup>4</sup>Present address: Department of Biotechnology, Graduate School of Engineering, Osaka University, Suita, Japan. <sup>5</sup>These authors contributed equally: Keisuke Motone, Daphne Kontogiorgos-Heintz. ✉e-mail: jmdn@cs.washington.edu

charged polypeptide backbone, varying charge states of amino acid side chains, and stable tertiary structures<sup>39</sup>.

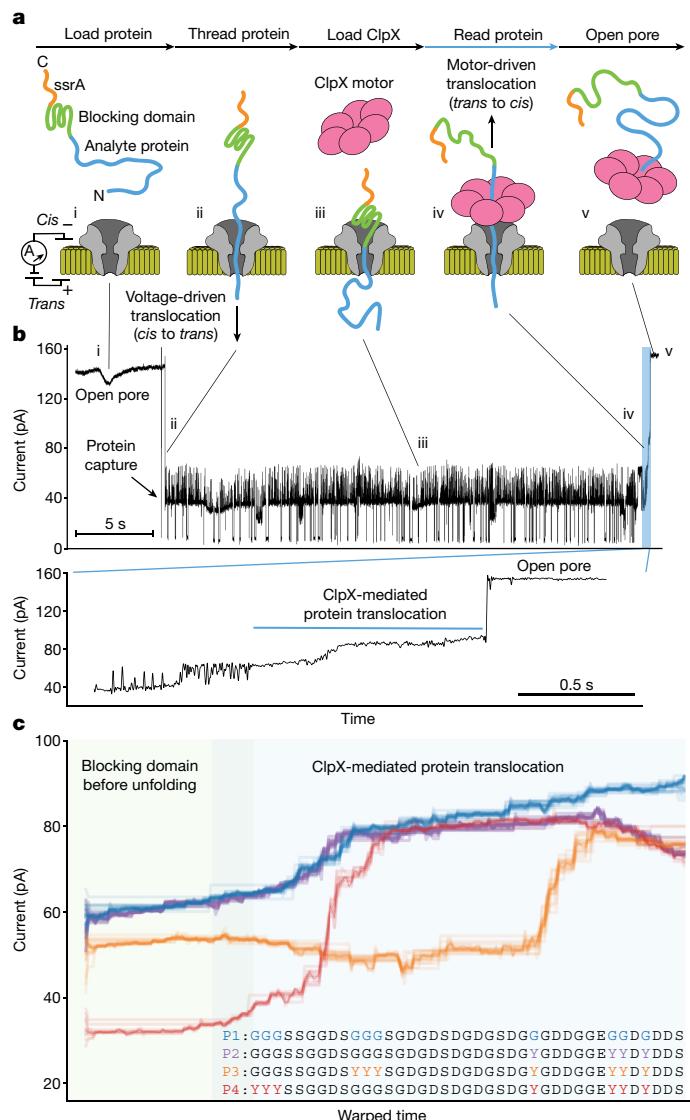
To overcome the problems of reading full-length proteins, here we have developed a technique to reversibly thread long protein strands into a CsgG pore<sup>40</sup> using electrophoresis, and then enzymatically pull them back out of the pore using the protein unfoldase and translocase activity of ClpX<sup>41</sup>. Unlike the rapid initial stage of threading the protein into the pore using electrophoretic force, the unfoldase-mediated translocation of proteins back out of the pore leads to slow, reproducible ionic current signals. This method has resulted in the processive translocation of long proteins, enabling the detection of single amino acid substitutions and PTMs across protein strands up to hundreds of amino acids in length. We have also developed an approach to rereading the same protein strand multiple times. Furthermore, this method enables the unfolding and translocation of a model folded protein domain for linear, end-to-end analysis.

## ClpX-mediated protein reading on an array

We and others have previously developed approaches to unfoldase-mediated protein translocation through nanopores<sup>42–44</sup>, but these methods required complex experimental set-ups and did not demonstrate the single amino acid sensitivity required for sequencing. In our previous methodology, the unfoldase and its cofactors were located in the *trans*-side solution, opposite to the location of the protein substrate addition<sup>42,43</sup>. This set-up rendered the method incompatible with commercial high-throughput nanopore sensor-array devices, such as MinION from Oxford Nanopore Technologies, which do not allow access to the *trans* compartment solution. To overcome the need for *trans* motor addition, we designed a more streamlined two-step process. First, the protein substrate is threaded into the nanopore by electrophoretic force (*cis*-to-*trans*). Then, ClpX is added to the *cis* solution to steadily pull the substrate protein back out of the pore (*trans*-to-*cis*) (Fig. 1a).

We first synthesized a protein to evaluate this method. It comprised an unstructured, negatively charged N-terminal sequence of 42 amino acids rich in glycine, serine and aspartic acid, to facilitate electrophoretic capture in the pore, attached to a stably folded domain (Smt3). This was followed by a short, positively charged sequence (an RGS repeat) and a ClpX-binding ssrA tag at the C-terminal end (protein P1; Supplementary Fig. 1). The RGS and the folded domain were included to inhibit complete translocation of the protein through the pore, thereby preserving the accessibility of the ssrA tag in the *cis* compartment. After introducing P1 into a MinION R9.4.1 flow cell incorporating a CsgG pore variant (Oxford Nanopore Technologies)<sup>3</sup> and applying a voltage of  $\sim$ 180 mV, we observed current blockades associated with the capture of the negatively charged protein tail in the pores. To test the ability of ClpX to extract the captured protein from the nanopore, we then introduced a buffer solution containing ClpX and ATP into the flow cell. Under these conditions, we observed deep ionic current blockades, characteristic of capture of the substrate protein in the nanopore, returning to the open channel state in a stepwise manner some time after the addition of ClpX. We also determined that these events were ATP dependent and occurred at a slower rate in the presence of ATPyS (Extended Data Fig. 1), an ATP analogue that is more difficult for ClpX to hydrolyse<sup>45</sup>. These results are consistent with our model that ClpX was binding to the ssrA tag and translocating the captured protein out of the nanopore with C-to-N-terminal directionality.

If this were true, we reasoned that mutations in the tail domain of the protein would induce alterations in the ionic current states observed during ClpX-mediated translocation of the protein through the nanopore. To test this, we synthesized three new proteins (P2, P3 and P4), each containing several tyrosine mutations at distinct positions of the polyGSD sequence (Supplementary Fig. 1). To compare the signal profiles of the four protein sequences directly, we created ensemble

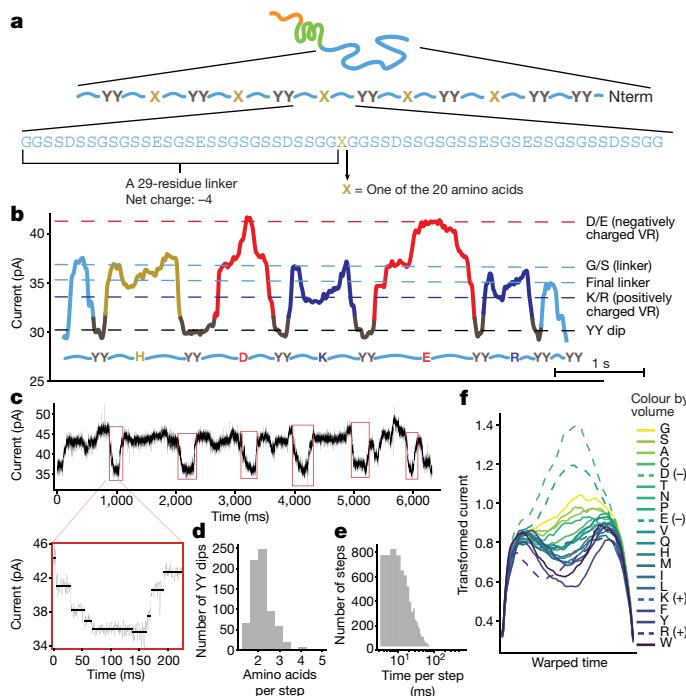


**Fig. 1 | Nanopore protein reading using an unfoldase.** **a**, Schematic of the *cis*-based unfoldase approach on the MinION platform. The roman numerals correspond to the ionic current states in **b**. **b**, Example trace of protein P1. Deep spikes in the capture state are hypothesized to be transient structural fluctuations of the Smt3 domain in the pore. State iii can be discerned from a transient drop in current when the ClpX solution is initially loaded into the flow cell. **c**, Ensemble traces for protein P1 (blue,  $n = 34$ ) and mutants P2 (purple,  $n = 17$ ), P3 (orange,  $n = 21$ ) and P4 (red,  $n = 12$ ). Protein sequences are oriented from C to N, with all mutation regions shown in colour.

ionic current traces for each of these proteins, as shown in Fig. 1c (the individual traces are in Supplementary Fig. 2). This revealed that the main differences across the translocation signals corresponded with the positions of the tyrosine mutations along the protein strands. Moreover, comparing all-versus-all signal dynamic time warping (DTW) distances revealed that the sets of translocation signals generated by each unique protein sequence formed distinct clusters, differentiating them from every other protein. This was statistically supported by permutational multivariate analysis of variance (PERMANOVA) with  $P < 1 \times 10^{-6}$  for each comparison after applying a Bonferroni correction.

## Resolving ClpX steps and substitutions

After establishing a *cis*-based ClpX approach, we next sought to investigate the sensitivity of this method to single amino acids as a first step



**Fig. 2 | Characterizing single amino acid substitutions and ClpX stepping in PASTORs.** **a**, PASTOR sequence composition. **b**, Filtered nanopore current trace of PASTOR-HDKER. Colour boundaries are defined by automated YY segmentation. **c**, Top, an example PASTOR trace. The red boxes show the manually segmented YY dips. Bottom, the black horizontal lines denote the mean of individual steps. **d**, Distribution of the mean number of residues per step in each of the YY dips;  $n = 776$  YY dips. **e**, Step dwell-time distribution. **f**, Average signal trace for the transformed VRs of each amino acid after Euclidian alignment of all the VRs equidistantly stretched to the same length. The VRs of a charged amino acid are shown as a dashed line ( $n$  of VRs and experiments are shown in Extended Data Table 1).

towards developing a long-read protein analysis method. To do this, we designed protein constructs with five repeating sequence blocks, each containing 59 amino acids. These blocks were built with a base sequence of glycine, serine, aspartic acid and glutamic acid (Fig. 2a and Supplementary Fig. 1). We introduced a single unique amino acid mutation at the central position in each block and demarcated the blocks with a double tyrosine mutation at each end. This spacing was intended to avoid entanglement of the signal contributions from single amino acid mutations by preventing them from occupying the pore at the same time. This hypothesis was grounded on prior observations indicating that around 20 amino acids can occupy the CsgG sensing region when in a stretched conformation<sup>46</sup>. We termed these strategically designed protein constructs ‘proteins for amino acid sequencing through optimized regions’ (PASTORs). We synthesized a total of eight different PASTOR variants, each containing a different sequence of mutations. The PASTOR design allowed us to analyse up to five different mutations in a single nanopore read, and the total set of eight PASTORs (C to N: HDKER, GNQST, FYWCP, AVLIM, VGDNY, TWAFH, PRMQE and KSILC) enabled us to investigate each of the 20 amino acids in two different PASTOR sequence contexts.

ClpX-mediated analysis of the PASTOR proteins manifested ionic current traces containing repetitive patterns that resulted from the seven YY mutations preceding the return to the open channel state, seen as seven repeated dips in the signal (PASTOR-HDKER is shown in Fig. 2b and the other PASTORs are shown in Supplementary Fig. 3). Between these dips, distinctive and reproducible variations in the ionic current signals were observed, corresponding to the variable amino acid mutation in each block. Using the consistent, substantial effect

of the YY mutations, we segregated the current signals into regions termed ‘YY dips’ and ‘variable regions’ (VRs) (Supplementary Fig. 4), and used these patterns to scale and segment our signals (Methods and Extended Data Fig. 2).

A close examination of the YY-dip signals revealed rapid, stepwise changes in the current level, which we reasoned must be caused by single ClpX substrate translocation steps (Fig. 2c). Previous estimates of the step size of these movements vary: single-molecule tweezer experiments suggested that ClpX translocates 5–8 amino acids per step<sup>45,46</sup>, whereas structural studies on ClpX-related protein-remodelling machines imply 2 amino acids per step<sup>47,48</sup>. To determine the step size of ClpX in our experiments, we analysed these YY-dip regions using a segmentation algorithm (examples are in Supplementary Figs. 5 and 6), filtering out dips with back-stepping or excessive noise. By dividing the number of residues contributing to the YY dips by the number of steps identified per YY dip (Methods and Extended Data Fig. 3a,b), we determined that ClpX translocates an average of approximately 1.96 residues per step (s.d. = 0.25; Fig. 2d). This was confirmed by a secondary segmentation algorithm yielding a similar mean of around 1.89 residues per step (s.d. = 0.28; Extended Data Fig. 3c–e). The dwell time of each step, capturing the duration ClpX pauses between pulling events, had a mean of 28.6 ms (s.d. = 32.3 ms; Fig. 2e). These results are in strong agreement with the step size of two amino acids hypothesized from the structural studies, and indicate that the tweezer experiments lacked the spatio-temporal resolution to resolve individual ClpX steps.

After establishing the two-residue stepping behaviour of ClpX, we shifted our focus to the VRs to explore the ionic signatures of individual amino acid mutations. Our analysis revealed that in VRs with a neutral amino acid mutation, there was a negative correlation between the ionic current levels and the volume of the amino acid (Fig. 2f and Extended Data Fig. 4a). This observation supports a volume-exclusion model in which larger amino acids block more current than their smaller counterparts do. Interestingly, the VRs containing positively charged residues (K and R) decreased the current level below the baseline sequence, whereas negatively charged residues increased it, diverging from the volume-exclusion model. This effect was greater for negatively charged residues than for positively charged ones. One possible explanation for this is that the negatively charged residues resist translocation to the negatively charged *cis* compartment, causing the protein strand to stretch and thereby decrease the total volume of protein in the pore. Conversely, a positively charged residue would be attracted to the *cis* compartment and could introduce upstream kinks in the protein strand, adding to the protein volume in the pore. The impact on signal levels could also be attributed to variations in solvation states and the mobility of ions near the charged amino acids. We hypothesize that some of the variability between VRs of the same amino acid can be attributed to non-enzymatic PTMs, such as asparagine deamidation, which is discussed in Extended Data Fig. 4b–h, Supplementary Note 1 and Supplementary Fig. 7. Collectively, these results show that this method is sensitive to single amino acid residues.

## Sequence to nanopore signal modelling

Considering the relationship between the volume and the charge of individual amino acids, and their effect on nanopore signals, we developed a biophysical model designed to simulate nanopore signals from the amino acid sequence of a protein directly. This model, which builds on previous findings<sup>46</sup>, determines a summation of the volume and charge of amino acids in a moving 20-residue window, applying a centrally positioned negative parabolic weight (Extended Data Fig. 5a–d). Extended Data Fig. 5e shows the signal generated by our model for the PASTOR-TWAFH protein sequence aligned to an actual nanopore trace of the same protein. Model signals for all proteins in this study are shown in Supplementary Fig. 8. To evaluate

our model quantitatively, we computed the DTW distance between the actual ionic current traces and the model, and compared that with the distribution of distances between the actual ionic current trace and the model of random sequences. We found that, on average, the model of the true sequence ranked in the top 0.3% of the best matches (Extended Data Fig. 5f), which confirms that the signal agreement observed in Extended Data Fig. 5e was not caused by artefacts from DTW alignment. This reinforces the assertion that our model has the capacity to simulate these current traces accurately in these sequence contexts.

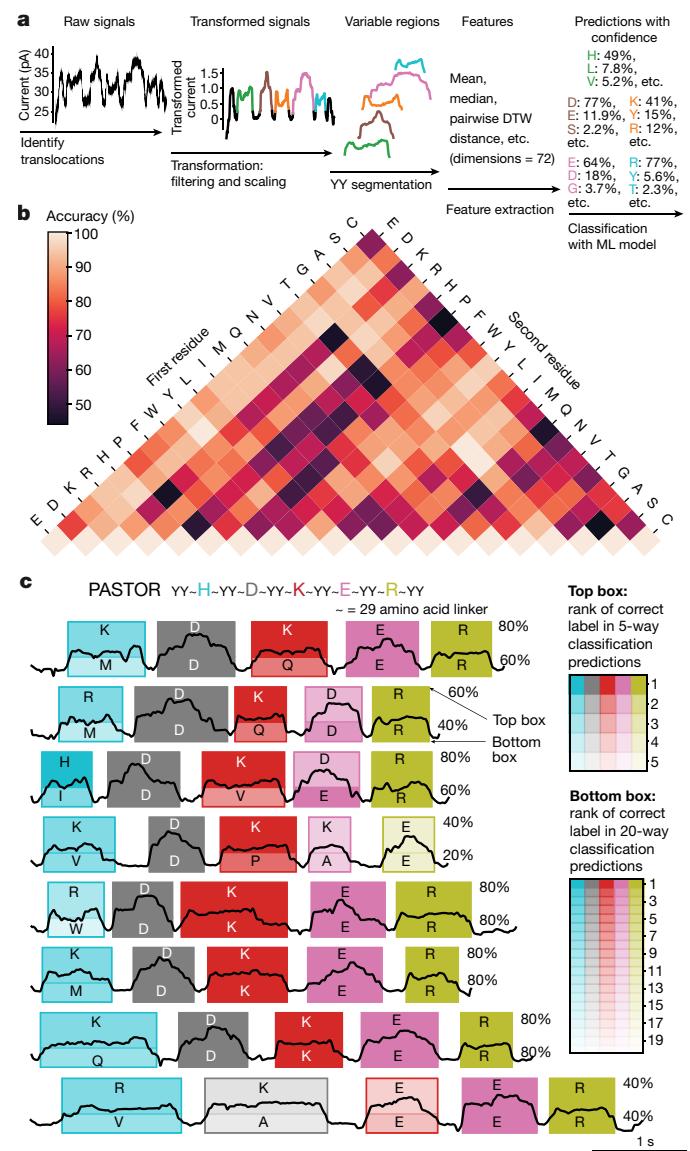
## Building a single-molecule aminocaller

Sequencing synthetic protein constructs such as PASTORs could serve diverse technological applications, including protein barcoding<sup>46,49</sup>. We addressed this by initially training machine learning models to identify the single mutation present in a VR. This process consisted of filtering and scaling each of the raw signal traces, followed by segmentation of the VR signal regions (Fig. 3a). To featurize the VR signals, we used a combination of manually curated features and DTW-distance features (Methods). We next explored several classical and deep machine learning models and found that random forests most frequently achieved the highest accuracy. All classification analyses were then done with a hyperparameter-tuned random forest evaluated on a fixed held-out test set, unless otherwise specified. We first evaluated the discrimination of all pairs of amino acids (Fig. 3b and Extended Data Fig. 6a). Pairs of amino acids with dissimilar volumes, or in which one was negatively charged, achieved the highest accuracy classifications. For example, tyrosine versus aspartate exhibited 100% discrimination accuracy. Some pairs of amino acids, such as leucine and isoleucine, proved to be more challenging, owing to their inherent physico-chemical similarities. Amino acids with high variance signals, such as cysteine, were also more difficult to distinguish from others. We then moved to training models to classify among particular sets of three amino acids (for example, G, Y and D) in which the model achieved 95% single-read accuracy. Expanding this to 5-way classifications (for example, G, V, W, R and D), the model maintained high performance, achieving an accuracy rate of 86% (Extended Data Table 2 and Extended Data Fig. 6b). In the most challenging task, a 20-way amino acid classification, our top-performing model substantially outperformed a dummy classifier, obtaining an accuracy of 28% compared with just 5.5% for the dummy. When we considered top-N accuracy measurements, our model attained 67% accuracy for top-5 and 81% for top-8 accuracy in the 20-way classification task (Extended Data Fig. 6c).

Building on these results, we integrated our classifiers downstream of the PASTOR segmenter to develop an end-to-end PASTOR ‘aminocaller’. We then amino-called a set of PASTOR reads from the classification test set (Fig. 3c and Supplementary Fig. 9). Overall sequencing accuracy per read averaged about 62% and 42% for the HDKER sequence, and roughly 51% and 21% for the AVLIM sequence, using 5-way and 20-way classification models respectively.

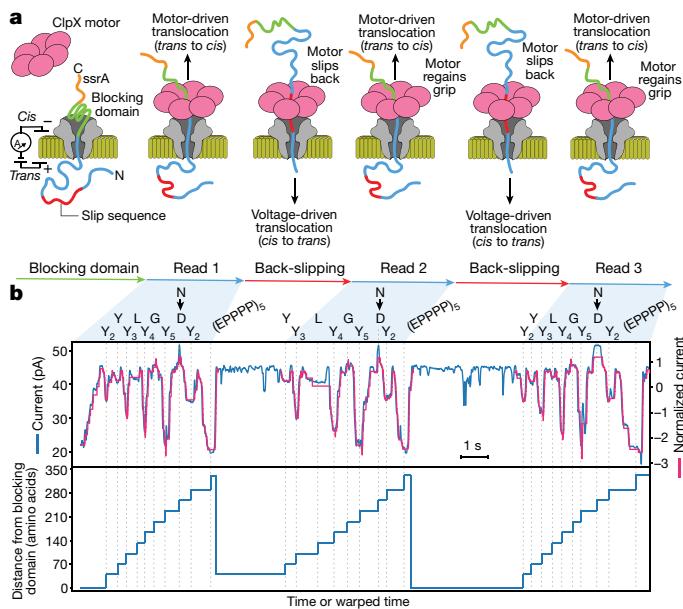
## Unfoldase slip sequence enables rereading

After developing the aminocaller, we aimed to improve the accuracy of our single-molecule sequencing approach by developing a method to reread single protein molecules. A previous study suggested that ClpX may have difficulty gripping particular polypeptide sequences, such as polyproline, on which the ClpXP complex showed slow degradation rates<sup>41</sup>. This prompted us to hypothesize that incorporating a ‘slippery’ amino acid sequence near the N terminus of a PASTOR would induce ClpX to momentarily lose its grip of the strand (Fig. 4a). Consequently, the substrate protein would be free to reread into the pore by electrophoresis. Rereading would stop enzyme-mediated translocation, and it would resume once ClpX regains its grip on the substrate. To test



**Fig. 3 | Single-molecule nanopore sequencing of single amino acid mutations.** **a**, Pipeline for PASTOR VR classification with machine learning (ML) models. **b**, Heatmap showing test accuracies in discriminating between all pairs of amino acid VR mutations, averaged over five random forests ( $n$  for VRs and experiments is shown in Extended Data Table 1). **c**, Example sequencing traces in the test set, for PASTOR-HDKER. Transformed ionic current traces are plotted with a box around the variable regions defined by the YY segmenter. The colour intensity of the boxes represents the ranking of the true class in the aminocaller’s prediction for each VR. For the 5-way classification task (top box shading), the classes are the five mutations found in that protein, whereas the 20-way classification task (bottom box shading) considers all possible amino acid classes. In each box, the letter corresponds to the model’s top prediction. Darker shades denote a more-accurate prediction, indicating that the correct label ranked high in the model’s predictions.

this strategy, we constructed a new PASTOR (PASTOR-reread) with two important sequence features: first, a proline-rich ‘slip’ sequence repeat (EPPPP)<sub>5</sub> positioned near the N terminus; and second, VRs separated by an increasing number of tyrosine residues, ranging from two to five (Supplementary Fig. 1). We reasoned that the distinct current levels that our signal-to-sequence model predicted for each of the repeats would enable us to estimate the slip distance (Supplementary Fig. 8). Indeed, nanopore signals produced by PASTOR-reread generally exhibited repeated signal patterns that closely aligned with our



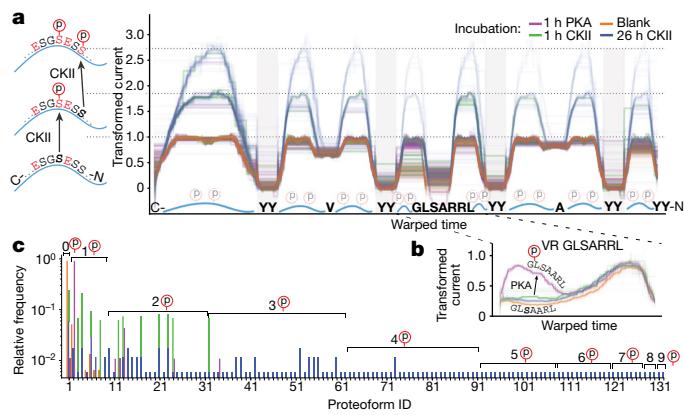
**Fig. 4 | Rereading single protein molecules multiple times with an unfoldase slip sequence.** **a**, Working model of rereading. **b**, Top box shows example trace of PASTOR-reread showing three almost-complete reread events (blue trace). Our model's predicted signal for the PASTOR-reread sequence (pink trace) was aligned to each reread. The fourth VR contains an asparagine mutation, but the corresponding signal level consistently resembles aspartate in all three instances of this PASTOR-reread trace. The modelled sequence was changed to contain an aspartate to reflect the putative PTM. Bottom box shows the approximate region of the strand that is in the nanopore over time.

model's prediction before returning to the open channel state (Fig. 4b and Supplementary Fig. 10). By using the tyrosine repeat regions as a measure of slip length, we observed that slipping distances were usually either short ranges (50–100 amino acids) or extended across the entire PASTOR unstructured region (more than 300 amino acids), accounting for roughly 40% and 30% of all rereads, respectively (Extended Data Fig. 7a). The effect of ClpX concentration on slipping distance and frequency is discussed in Supplementary Note 2 and Extended Data Fig. 7b,c.

We next investigated the potential of single-molecule rereading for increasing sequencing accuracy. Using a simulation discussed in Supplementary Note 3, the accuracy for the 20-way amino acid classification task improved from 28% to 61% (compared with a 5% random baseline) with 10 rereads (Extended Data Fig. 7d). Likewise, the accuracy for a 7-way classification task improved from 66% to 99% (against a 14% random baseline).

### Simulating protein barcode sequencing

Having determined the capacity for high-accuracy sequencing through PASTOR rereading and the ability to design PASTOR proteins with customizable VR sequences, we simulated the PASTOR VR sequence space with varying constraints, with a view to applications in protein barcoding. On the basis of the accuracy rates of our models (Extended Data Fig. 7d), we computed the number of distinct barcodes that could be generated at a given accuracy level. This calculation considered varying degrees of rereading and two different VR segment numbers per protein barcode (five and ten VRs). For example, our findings indicate that with ten VRs and ten rereads, it is feasible to generate libraries of more than 1 million or 1 billion unique PASTOR barcodes that are decodable with a single-molecule accuracy of more than 95% or more than 81%, respectively (Extended Data Fig. 7e and Supplementary Table 1).

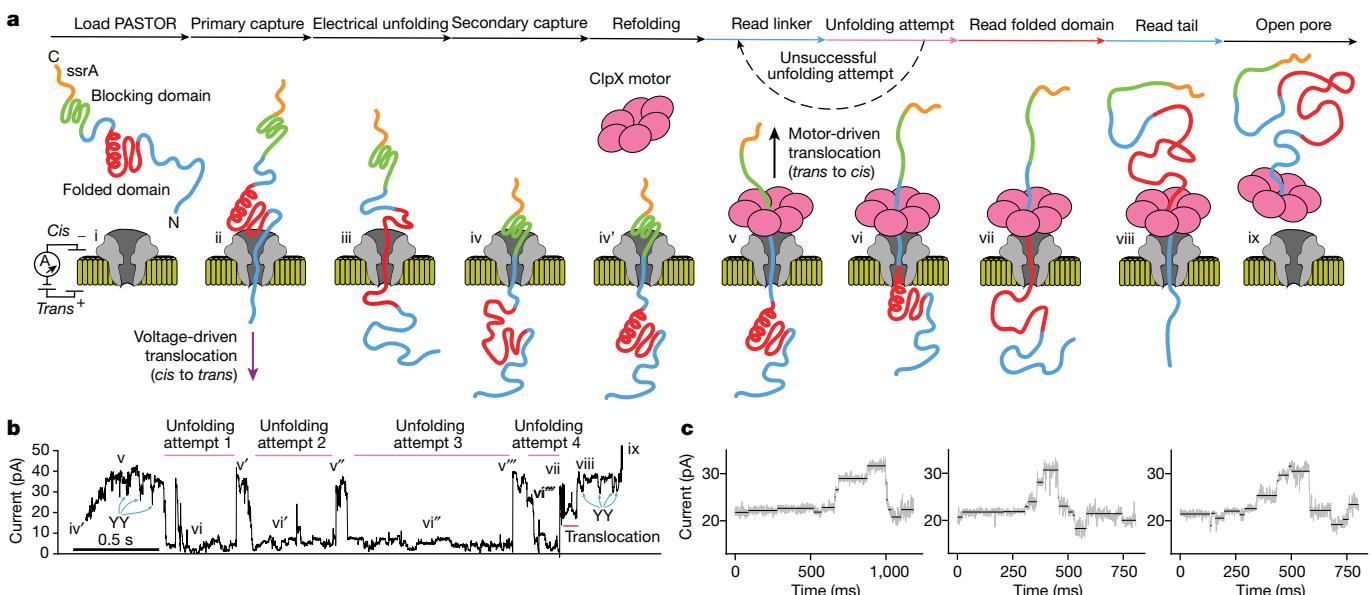


### Monitoring and mapping enzymatic PTMs

Demonstrating the ability to detect and map phosphorylations across long protein strands would be an important step towards developing a technology capable of identifying distinct full-length proteoforms. To do this, we focused on two serine/threonine protein kinases with distinct recognition motifs: protein kinase A (PKA), which recognizes the canonical motif RRXS, and casein kinase II (CKII), which targets the sequence SXXD/E<sup>50</sup>. To see if we could effectively characterize the differential activity of these two kinases using our nanopore reading approach, we designed a new substrate, PASTOR-phos (Fig. 5a and Supplementary Figs. 1 and 11). In this design, we inserted the substrate peptide LRRASLG ('kemptide') of PKA into one of the VRs, making it specific for recognizing the kinase<sup>51</sup>. To investigate CKII in PASTOR-phos, we used the original 29-amino acid linker sequences from the PASTOR, which inherently contain a CKII motif, while reducing the total number of VRs.

After incubating PASTOR-phos with PKA for 1 h, we performed nanopore analysis, which found a substantial increase in ionic current in 91 of the 92 reads (98.9%) in the kemptide VR compared with the baseline (Fig. 5b and Extended Data Fig. 8a). This increase in ionic current is consistent with expectations for the negatively charged phosphoserine, which carries a charge of -2, to enhance ionic flow. Conversely, 361 of the 368 non-kemptide VRs and linker sequences (98.1%) showed no substantial signal changes. These results are consistent with PKA activity being specific to the RRXS motif.

When the same substrate (PASTOR-phos) was treated with CKII for 1 h, we observed high read-to-read variability manifested by large increases in current levels that were found to be concentrated in the eight linker sequences containing the CKII phosphorylation motif (Fig. 5a). The maximum peak transformed current of the VRs and linker sequences incubated in CKII showed a significant increase compared with that of the PKA and the no-kinase (blank) and incubation conditions (Extended Data Fig. 8a). This indicates that the method can discern site-specific phosphorylation events, demonstrating specificity to CKII. Interestingly, a small portion of linkers had signal increases much higher than the others, indicating that they were being phosphorylated to a greater extent (Fig. 5a and Extended Data Fig. 8a). Analysis of the phosphorylated linker sequence revealed that phosphorylation at the



**Fig. 6 | Processive reading of folded protein domains.** **a**, Working model of ClpX-mediated processing of folded proteins. The roman numerals correspond to the ionic current states in **b**. **b**, Example trace of PASTOR-titin. **c**, Example

traces of titin translocation (state vii), with black horizontal lines denoting the mean of individual putative ClpX steps, found with the Bayesian segmentation algorithm (Methods).

initial motif induces the formation of a secondary CKII motif, SX<sub>n</sub>PS, which has been described previously<sup>50</sup>. We hypothesized that the linkers with much higher signal levels were phosphorylated at both serine positions. To test this, we reasoned that extending the incubation time of CKII with the substrate should increase the frequency of both single and, consequently, double phosphorylation events. Supporting this hypothesis, data from PASTOR-phos after a 26-h incubation revealed increased occurrences of both putative single and double phosphorylations in the linker sequences (Extended Data Fig. 8b,c).

Given the abundance of potential CKII phosphorylation sites in the PASTOR-phos sequence, numerous combinatorially unique proteoforms are possible (a total of 13,122). To map our reads to these various modified forms, we integrated phosphoserine into our sequence-to-signal model (Methods). This approach allowed us to align nanopore traces with the predicted sequence-to-signal profiles for each phosphorylation state across all VRs and linkers, as shown in Supplementary Fig. 12. Consequently, we identified and quantitatively assessed more than 100 distinct full-length proteoforms of PASTOR-phos, across reads obtained from the baseline, PKA and CKII experiments (Fig. 5c, Supplementary Fig. 13 and Supplementary Tables 2 and 3). For example, the 26-h CKII incubation resulted in single molecules containing as many as nine phosphorylated residues.

## **Processive reading of folded domains**

Progressing beyond synthetic, unstructured sequences, we next evaluated the effectiveness of our *cis*-based unfoldase method on protein sequences that contain a folded domain. For this purpose, we analysed a PASTOR protein with the titin I27<sub>VISPL</sub> domain, which consists of 89 amino acids arranged into 8 β-strands<sup>52</sup>, inserted into the third VR position (PASTOR-titin; Supplementary Fig. 1). Unlike unstructured proteins, nanopore traces of PASTOR-titin yielded an initial two-step electrophoretic nanopore capture state, indicating that the folded titin domain was first captured on the nanopore at state ii and then electrically unfolded at transition state iii to produce the typical PASTOR capture signal at state iv, manifested by the Smt3 domain on the pore (Fig. 6a and Extended Data Fig. 9a). After adding ClpX to the *cis* compartment, we observed a translocation signal corresponding to the leading VRs and YY regions, which are tethered to the C terminus

of the titin domain (state v). Subsequently, we observed a distinct and deep blockade state that we interpreted as ClpX attempting to unfold the titin domain (state vi), which presumably refolds in the *trans* compartment after the initial translocation. This deep blockade state often reverted back to the previous state, indicating an unsuccessful ClpX unfolding attempt and indicating that ClpX slipped back on the protein strand<sup>52,53</sup>. After a successful unfolding attempt, we observed putative translocation of the titin domain (state vii; Fig. 6b and Supplementary Fig. 14). In the titin translocation signal, individual ClpX steps could be seen (Fig. 6c). After translocation of the titin domain, we observed characteristic signal features corresponding to the downstream VRs and YY regions (state viii) before transitioning back to an open-pore state (state ix).

To confirm our understanding of the unfolding and translocation states, we performed experiments using a variant of titin I127 (PASTOR-D-titin) with a destabilized tertiary structure, introduced through double-point mutations (C47E and C63E) on two buried cysteines<sup>43,52</sup>. Comparing PASTOR-D-titin (Extended Data Fig. 9b) with PASTOR-titin allowed us to explore the effect of the tertiary structure of titin on the resulting current signals. This generated two notable differences: first, PASTOR-D-titin displayed unique signal features at the putative unfolding state vi, indicating structural disparities between the two variants; and second, states v and vi were observed only once in PASTOR-D-titin before the presumptive translocation state vii, in contrast to PASTOR-titin in which they were typically observed multiple times, leading to a substantial difference between the distribution of the unfolding times of PASTOR-titin and PASTOR-D-titin (Extended Data Fig. 9c). These differences can be attributed to the more stable, unfolding-resistant titin domain of PASTOR-titin compared with that of PASTOR-D-titin. In PASTOR-titin, repeated observations of states v and vi, which were not present for PASTOR-D-titin, support the conclusion that they result from unsuccessful unfolding attempts and ClpX back-slipping events triggered by the stable titin domain. Also, despite their dissimilar structural stabilities, PASTOR-titin and PASTOR-D-titin demonstrated similar signals during the putative translocation state vii (Extended Data Fig. 9d-f). This similarity reflects their nearly identical primary amino acid sequences. The observation of similar signals at the proposed translocation state vii between PASTOR-titin and PASTOR-D-titin,

# Article

despite their differences in structural stability, underscores the role of the primary amino acid sequence in this process. It indicates that the primary sequence is the main determinant of the translocation signal through the nanopore, whereas structural variations have greater influence on the preceding unfolding state.

We next tested PASTOR constructs with the amyloid- $\beta$  protein 1–42 (PASTOR-A $\beta$ 42) and its shorter derivative 1–15 (PASTOR-A $\beta$ 15), which have distinct amino acid sequences and lengths from the titin domain. We reasoned that A $\beta$ 42 and A $\beta$ 15 would generate brief unfolding states because they are partly but not fully structured in their monomeric forms<sup>54</sup>. As expected, after nanopore analysis they yielded ionic current traces similar to PASTOR-D-titin overall but with distinct features in unfolding state vi (Extended Data Fig. 9g and Supplementary Fig. 14). Furthermore, comparing their putative translocation states (state vii) using DTW distance, we observed that the signals generated by PASTOR-A $\beta$ 42 and PASTOR-A $\beta$ 15 share similarities to, but are distinct from, signals generated by PASTOR-titin and PASTOR-D-titin (Extended Data Fig. 9d–f), reflective of the translocation state being dependent on protein primary sequence. Overall, the dwell times of these different states also correspond well with their respective sequence lengths across all the PASTOR proteins, indicating a translocation rate of around 63 amino acids per second (an average dwell time of about 16 ms per amino acid) (Extended Data Fig. 9h). This is close to previous estimates of ClpX translocation speed and the observation that the rate of ClpX-mediated protein translocation is relatively constant regardless of protein sequence<sup>43,55</sup>.

Finally, we assessed our predictive model using these proteins. Because the model does not factor in the signal features linked with unfolding, we analysed the signal segment after the unfolding state until the completion of the translocation (states vii–viii). Using the same comparison technique as previously implemented for the PASTOR protein models, we found that the average model of the PASTORs containing folded domains ranked in the top 0.04% of the best matches (Extended Data Fig. 9i). This evaluation is evidence that our model can adequately simulate these current traces in the specified sequence contexts.

## Discussion

We have introduced a new approach for single-molecule reading of long protein strands using nanopores and an unfoldase motor protein. This method achieves single-amino-acid sensitivity and demonstrates the capability to reread and sequence amino acid substitutions in defined sequence contexts across long protein strands. This could immediately advance protein barcoding technology, as we project the ability to design libraries of synthetic peptide sequences (more than 1 billion). Moreover, we applied our method to detect and map the activities of distinct kinases, achieving site-specific detection of enzymatic PTMs along extended protein sequences and the relative quantification of more than 100 putative proteoforms of a single synthetic protein substrate. The implications and challenges of PTM detection are outlined in Supplementary Discussion 1.

We also established that ClpX translocates proteins through the nanopore in a stepwise manner, in agreement with structural studies that suggested a fundamental step size of two amino acids. We then laid the groundwork for a biophysical model capable of simulating nanopore signals that are generated when individual protein sequences are pulled through the nanopore by the unfoldase. This result could eventually enable a ‘lookup table’ approach reminiscent of mass spectrometry, facilitating full-length, single-molecule protein identification and fingerprinting (Supplementary Discussion 2). However, further work is needed to explore more native protein sequences. Furthermore, directly determining amino acid sequences from the nanopore signal (*de novo* sequencing) will be even more difficult and will require extensive datasets for training *de novo* amino acid callers.

Finally, we have demonstrated the full-length reading of a model folded protein domain, which is an important result as we move towards reading natural protein molecules. In the present system, electrophoretic protein unfolding (*cis* to *trans*; state iii in Fig. 6a) and ClpX-mediated protein unfolding (*trans* to *cis*; state vii in Fig. 6a) are key to achieving full-length folded-domain analysis. It is likely that some protein domains will exhibit greater resistance to unfolding than the substrates explored in this study. In such cases, extra strategies could be used to facilitate unfolding, such as the use of denaturants and electro-osmotic flow<sup>56</sup>. As we begin to turn our attention to natural proteins, this methodology will require synthetic N- and C-terminal sequences that can be appended using existing terminus-specific chemical-conjugation techniques<sup>39,57,58</sup> (Supplementary Discussion 3).

One factor that currently impedes data collection throughput is the two-step flow cell loading process. Looking ahead, we can imagine a system that operates continuously, with the unfoldase prebound to the protein analyte but prevented from initiating unfolding activity until the protein strand is captured by the pore. This method, which mirrors strategies developed for the nanopore sequencing of DNA<sup>59</sup>, has the potential to greatly increase throughput. There are challenges with experimental yield (Extended Data Fig. 10 and Supplementary Discussion 4) and read-out accuracy, particularly for domains with secondary structure (Supplementary Discussion 5), and these can be addressed in future work. In conclusion, this work serves as a stepping stone towards full-length protein identification, capable of achieving the highest level of proteoform resolution. Furthermore, it promises immediate advances, particularly in the context of protein barcoding and PTM-monitoring applications.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-024-07935-7>.

- Chandramouli, K. & Qian, P.-Y. Proteomics: challenges, techniques and possibilities to overcome biological sample complexity. *Hum. Genomics Proteomics* **2009**, 239204 (2009).
- Dupree, E. J. et al. A critical review of bottom-up proteomics: the good, the bad, and the future of this field. *Proteomes* **8**, 14 (2020).
- Van der Verren, S. E. et al. A dual-constriction biological nanopore resolves homonucleotide sequences with high fidelity. *Nat. Biotechnol.* **38**, 1415–1420 (2020).
- Dorey, A. & Horwka, S. Nanopore DNA sequencing technologies and their applications towards single-molecule proteomics. *Nat. Chem.* **16**, 314–334 (2024).
- Smith, L. M., Kelleher, N. L. & The Consortium for Top Down Proteomics. Proteoform: a single term describing protein complexity. *Nat. Methods* **10**, 186–187 (2013).
- Lothrop, A. P., Torres, M. P. & Fuchs, S. M. Deciphering post-translational modification codes. *FEBS Lett.* **587**, 1247–1257 (2013).
- Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41–45 (2000).
- Thomson, M. & Gunawardena, J. Unlimited multistability in multisite phosphorylation systems. *Nature* **460**, 274–277 (2009).
- Restrepo-Pérez, L., Joo, C. & Dekker, C. Paving the way to single-molecule protein sequencing. *Nat. Nanotechnol.* **13**, 786–796 (2018).
- Alfaro, J. A. et al. The emerging landscape of single-molecule protein sequencing technologies. *Nat. Methods* **18**, 604–617 (2021).
- Swaminathan, J. et al. Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. *Nat. Biotechnol.* <https://doi.org/10.1038/nbt.4278> (2018).
- Reed, B. D. et al. Real-time dynamic single-molecule protein sequencing on an integrated semiconductor device. *Science* **378**, 186–192 (2022).
- Ohayon, S. et al. Full-length single protein molecules tracking and counting in thin silicon channels. *Adv. Mater.* **36**, e2314319 (2024).
- Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nat. Biotechnol.* **34**, 518–524 (2016).
- Timp, W. & Timp, G. Beyond mass spectrometry, the next step in proteomics. *Sci. Adv.* **6**, eaax8978 (2020).
- Robertson, J. W. F., Ghimire, M. L. & Reiner, J. E. Nanopore sensing: a physical-chemical approach. *Biochim. Biophys. Acta Biomembr.* **1863**, 183644 (2021).
- Varongchayakul, N., Song, J., Meller, A. & Grinstaff, M. W. Single-molecule protein sensing in a nanopore: a tutorial. *Chem. Soc. Rev.* **47**, 8512–8524 (2018).

18. Restrepo-Pérez, L., Wong, C. H., Maglia, G., Dekker, C. & Joo, C. Label-free detection of post-translational modifications with a nanopore. *Nano Lett.* **19**, 7957–7964 (2019).
19. Chavis, A. E. et al. Single molecule nanopore spectrometry for peptide detection. *ACS Sens.* **2**, 1319–1328 (2017).
20. Restrepo-Pérez, L. et al. Resolving chemical modifications to a single amino acid within a peptide using a biological nanopore. *ACS Nano* **13**, 13668–13676 (2019).
21. Yusko, E. C. et al. Real-time shape approximation and fingerprinting of single proteins using a nanopore. *Nat. Nanotechnol.* **12**, 360–367 (2017).
22. Wei, R., Gatterdam, V., Wieneke, R., Tampé, R. & Rant, U. Stochastic sensing of proteins with receptor-modified solid-state nanopores. *Nat. Nanotechnol.* **7**, 257–263 (2012).
23. Waduge, P. et al. Nanopore-based measurements of protein size, fluctuations, and conformational changes. *ACS Nano* **11**, 5706–5716 (2017).
24. Ouldali, H. et al. Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore. *Nat. Biotechnol.* **38**, 176–181 (2020).
25. Lucas, F. L. R., Versloot, R. C. A., Yakovleva, L., Walvoort, M. T. C. & Maglia, G. Protein identification by nanopore peptide profiling. *Nat. Commun.* **12**, 5795 (2021).
26. Rodriguez-Larrea, D. Single-aminoacid discrimination in proteins with homogeneous nanopore sensors and neural networks. *Biosens. Bioelectron.* **180**, 113108 (2021).
27. Kennedy, E., Dong, Z., Tenant, C. & Timp, G. Reading the primary structure of a protein with 0.07 nm<sup>3</sup> resolution using a subnanometre-diameter pore. *Nat. Nanotechnol.* **11**, 968–976 (2016).
28. Dong, Z., Kennedy, E., Hokmabadi, M. & Timp, G. Discriminating residue substitutions in a single protein molecule using a sub-nanopore. *ACS Nano* **11**, 5440–5452 (2017).
29. Thakur, A. K. & Movileanu, L. Real-time measurement of protein-protein interactions at single-molecule resolution using a biological nanopore. *Nat. Biotechnol.* <https://doi.org/10.1038/nbt.4316> (2018).
30. Harrington, L., Cheley, S., Alexander, L. T., Knapp, S. & Bayley, H. Stochastic detection of Pim protein kinases reveals electrostatically enhanced association of a peptide substrate. *Proc. Natl Acad. Sci. USA* **110**, E4417–E4426 (2013).
31. Reynaud, L., Bouchet-Spinelli, A., Raillon, C. & Buhot, A. Sensing with nanopores and aptamers: a way forward. *Sensors* **20**, 4495 (2020).
32. Rosen, C. B., Rodriguez-Larrea, D. & Bayley, H. Single-molecule site-specific detection of protein phosphorylation with a nanopore. *Nat. Biotechnol.* **32**, 179–181 (2014).
33. Versloot, R. C. A. et al. Quantification of protein glycosylation using nanopores. *Nano Lett.* **22**, 5357–5364 (2022).
34. Martin-Baniandres, P. et al. Enzyme-less nanopore detection of post-translational modifications within long polypeptides. *Nat. Nanotechnol.* **18**, 1335–1340 (2023).
35. Nova, I. C. et al. Detection of phosphorylation post-translational modifications along single peptides with nanopores. *Nat. Biotechnol.* <https://doi.org/10.1038/s41587-023-01839-z> (2024).
36. Yan, S. et al. Single molecule ratcheting motion of peptides in a Mycobacterium smegmatis porin A (MspA) nanopore. *Nano Lett.* **21**, 6703–6710 (2021).
37. Brinkerhoff, H., Kang, A. S. W., Liu, J., Aksimentiev, A. & Dekker, C. Multiple rereads of single proteins at single-amino acid resolution using nanopores. *Science* **374**, 1509–1513 (2021).
38. Chen, Z. et al. Controlled movement of ssDNA conjugated peptide through Mycobacterium smegmatis porin A (MspA) nanopore by a helicase motor for peptide sequencing application. *Chem. Sci.* **12**, 15750–15756 (2021).
39. Motone, K., Cardozo, N. & Nivala, J. Herding cats: label-based approaches in protein translocation through nanopore sensors for single-molecule protein sequence analysis. *iScience* **24**, 103032 (2021).
40. Goyal, P. et al. Structural and mechanistic insights into the bacterial amyloid secretion channel CsgG. *Nature* **516**, 250–253 (2014).
41. Barkow, S. R., Levchenko, I., Baker, T. A. & Sauer, R. T. Polypeptide translocation by the AAA+ClpXP protease machine. *Chem. Biol.* **16**, 605–612 (2009).
42. Nivala, J., Marks, D. B. & Akeson, M. Unfoldase-mediated protein translocation through an α-hemolysin nanopore. *Nat. Biotechnol.* **31**, 247–250 (2013).
43. Nivala, J., Mulroney, L., Li, G., Schreiber, J. & Akeson, M. Discrimination among protein variants using an unfoldase-coupled nanopore. *ACS Nano* **8**, 12365–12375 (2014).
44. Zhang, S. et al. Bottom-up fabrication of a proteasome–nanopore that unravels and processes single proteins. *Nat. Chem.* **13**, 1192–1199 (2021).
45. Burton, R. E., Baker, T. A. & Sauer, R. T. Energy-dependent degradation: linkage between ClpX-catalyzed nucleotide hydrolysis and protein-substrate processing. *Protein Sci.* **12**, 893–902 (2003).
46. Cardozo, N. et al. Multiplexed direct detection of barcoded protein reporters on a nanopore array. *Nat. Biotechnol.* **40**, 42–46 (2022).
47. Fei, X. et al. Structures of the ATP-fueled ClpXP proteolytic machine bound to protein substrate. *eLife* **9**, e52774 (2020).
48. de la Peña, A. H., Goodall, E. A., Gates, S. N., Lander, G. C. & Martin, A. Substrate-engaged 26S proteasome structures reveal mechanisms for ATP-hydrolysis-driven translocation. *Science* **362**, eaav0725 (2018).
49. Wroblewska, A. et al. Protein barcodes enable high-dimensional single-cell CRISPR screens. *Cell* **175**, 1141–1155 (2018).
50. Johnson, J. L. et al. An atlas of substrate specificities for the human serine/threonine kinase. *Nature* **613**, 759–766 (2023).
51. Szmagierski, A., Guidotti, A. & Costa, E. Endogenous protein kinase inhibitors. Purification, characterization, and distribution in different tissues. *J. Biol. Chem.* **252**, 3848–3853 (1977).
52. Kenniston, J. A., Baker, T. A., Fernandez, J. M. & Sauer, R. T. Linkage between ATP consumption and mechanical unfolding during the protein processing reactions of an AAA+ degradation machine. *Cell* **114**, 511–520 (2003).
53. Maillard, R. A. et al. ClpXP generates mechanical force to unfold and translocate its protein substrates. *Cell* **145**, 459–469 (2011).
54. Ball, K. A., Phillips, A. H., Wemmer, D. E. & Head-Gordon, T. Differences in β-strand populations of monomeric Aβ40 and Aβ42. *Biophys. J.* **104**, 2714–2724 (2013).
55. van Ginkel, J. et al. Single-molecule peptide fingerprinting. *Proc. Natl Acad. Sci. USA* **115**, 3338–3343 (2018).
56. Yu, L. et al. Unidirectional single-file transport of full-length proteins through a nanopore. *Nat. Biotechnol.* <https://doi.org/10.1038/s41587-022-01598-3> (2023).
57. De Rosa, L., Di Stasi, R., Romanelli, A. & D’Andrea, L. D. Exploiting protein N-terminus for site-specific bioconjugation. *Molecules* **26**, 3521 (2021).
58. Bloom, S. et al. Decarboxylative alkylation for site-selective bioconjugation of native proteins via oxidation potentials. *Nat. Chem.* **10**, 205–211 (2018).
59. Olasagasti, F. et al. Replication of individual DNA molecules under electronic control using a protein nanopore. *Nat. Nanotechnol.* **5**, 798–806 (2010).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024

## Methods

### Expression and purification of proteins

Plasmids for analyte proteins were constructed using gBlocks (Integrated DNA Technologies) inserted into the pET-49b(+) plasmid (Novagen), with a dihydrofolate reductase domain, a polyhistidine tag and a TEV cleavage site upstream of the sequence encoding an analyte protein. The NEBuilder HiFi DNA assembly and Q5 site-directed mutagenesis kits (New England Biolabs) were used for plasmid construction. Cloning was done using NEB 5- $\alpha$ -competent *Escherichia coli* cells. Plasmid sequences were verified by Sanger sequencing through Genewiz. Protein expression was induced overnight at 30 °C with BL21 (DE3) *E. coli* cells in Overnight Express Instant TB medium (Novagen). Proteins were purified by immobilized metal affinity chromatography (IMAC) with TALON metal affinity cobalt resin and its associated buffer set (Takara), following the manufacturer's instructions. Proteins were cleaved with TEV protease (New England Biolabs) and further purified by reverse IMAC. Purified proteins were concentrated using ultracentrifugal filters with a 10 kDa cutoff (Amicon) and stored in the short term at 4 °C or in the long term at -80 °C until use.

A covalently linked hexamer of an N-terminal truncated ClpX variant (ClpX- $\Delta N_6$ )<sup>60</sup> was prepared using the BLR *E. coli* strain as described previously<sup>43</sup>. In brief, cells were grown to an optical density at 600 nm ( $OD_{600}$ ) of around 0.6 in LB medium and then incubated in the presence of 0.5 mM isopropyl  $\beta$ -D-1-thiogalactopyranoside (IPTG) at 23 °C for about 3 h to induce ClpX expression. ClpX was purified by IMAC and anion-exchange chromatography. Purified ClpX was stored at -80 °C in small aliquots until use. ClpP expression was induced at an  $OD_{600}$  of around 0.6 with 0.5 mM IPTG at 30 °C for about 3 h<sup>43</sup>. ClpP was purified by IMAC and stored at -80 °C until use.

### PTM assays

For asparagine deamidation, protein (around 1 mg ml<sup>-1</sup>) was incubated overnight in 100 mM sodium bicarbonate buffer (pH 9.6) at 25 °C to catalyse deamidation. For protein phosphorylation with kinase, protein was incubated with either 50,000 units per ml PKA (New England Biolabs) or 10,000 units per ml CKII (New England Biolabs) in a protein kinase buffer (10 mM MgCl<sub>2</sub>, 0.1 mM EDTA, 2 mM DTT, 0.01% Brij 35, 260  $\mu$ M ATP and 50 mM Tris-HCl, pH 7.5) at 30 °C. The protein solution was used for nanopore analysis immediately after the incubation without purification.

### MinION experiments

All the experiments were done on the MinION platform using R9.4.1 flow cells. Run conditions were set with a custom MinKNOW script (available from Oxford Nanopore Technologies) at a temperature of 30 °C and a constant voltage of -140 mV with a 3 kHz sampling frequency, except for initial proteins P1–P4, for which runs were performed at a constant voltage of -180 mV with a 10 kHz sampling frequency. Using the priming port, flow cells were first washed with 1 ml *cis* running buffer (200 mM KCl, 5 mM MgCl<sub>2</sub>, 10% glycerol and 25 mM HEPES-KOH, pH 7.6) and then loaded with 200  $\mu$ l protein analyte in *cis* running buffer at a final concentration of 500 nM, unless otherwise specified. Following the observation of protein captures in the pores, flow cells were washed with 1 ml *cis* running buffer to remove uncaptured proteins and subsequently loaded with 75  $\mu$ l *cis* running buffer supplemented with 4 mM ATP and 200 nM ClpX- $\Delta N_6$ , unless otherwise specified. The flow cell was washed about 4 min after analyte loading in the initial method, and around 6 min and 2 min after analyte loading at concentrations of 5 nM and 500 nM, respectively, in the optimized method (Extended Data Fig. 10a). For MinION runs in the high-salt condition (Extended Data Fig. 6b), a buffer containing 400 mM KCl, 5 mM MgCl<sub>2</sub> and 25 mM HEPES-KOH (pH 7.6) was used instead of standard *cis* running buffer to see if it would improve the signal-to-noise ratio.

### Bulk degradation assays

The time-course degradation assay of the PASTOR-HDKER protein was performed in *cis* running buffer with 6  $\mu$ M PASTOR-HDKER, 150 nM ClpX- $\Delta N_6$ , 300 nM ClpP<sub>14</sub> and an ATP-regeneration mix (4 mM ATP, 16 mM creatine phosphate and 7 units per ml creatine phosphokinase) at 30 °C. Incubation was stopped by denaturing samples in Laemmli buffer at 95 °C for 5 min. Samples were run on SDS-PAGE and stained with Coomassie blue to quantify the protein bands using the ImageJ software.

### Nanopore signal analysis

**Preprocessing.** To help identify ClpX-mediated protein translocations, we established detection thresholds using specific statistical parameters (standard deviation, median value, standard deviation of the mean of windows, and the ratio of values relative to the open pore value) indicative of translocation to ionic current blockades preceding a return to the open channel state. This analysis was used to assist the process of manually checking traces for translocations, and translocations with particularly high noise or disruptions were discarded. PASTOR proteins were auto-segmented as described below, with the exception of those containing folded domains and PASTOR-rereads, which were segmented manually. PASTOR-reread rereads with a complete Y<sub>2</sub>-Y<sub>3</sub>-Y<sub>4</sub>-Y<sub>5</sub>-Y<sub>2</sub> signal were assumed to be full-length reads with a back-slipping distance of 310 amino acids. Partial rereads missing the signal(s) of the C-terminal Y<sub>2</sub>, Y<sub>3</sub>, Y<sub>4</sub> and Y<sub>5</sub> were assigned to have back-slipping distances of 250, 188, 125 and 61 amino acids, respectively. All figures with raw traces (those shown in pA) had a low-pass Bessel filter applied using SciPy with  $N = 10$  and  $W_n = 0.025$ , except for those showing stepping analysis (Figs. 2c and 6c, Extended Data Fig. 3 and Supplementary Figs. 5 and 6), which had  $W_n = 0.7$ . Before use in data analysis, traces were smoothed by applying a low-pass Bessel filter with  $N = 10$  and  $W_n = 0.03$  with SciPy, and by applying average downsampling by a factor of 50 for proteins P1–4, 20 for the 8 PASTORs and 10 for the other proteins. Then, to scale, the segment was split into tenths, and the median of the minima of each tenth and the median of the maxima of each tenth were used as the min and max, respectively, to perform min–max scaling (Extended Data Fig. 2b). For PASTOR-phos, the signals were iteratively scaled. We first used this approach, then DTW-aligned traces to two canonical presegmented traces and selected the alignment with the lowest DTW distance. The max value of the N-terminal VR was multiplied by 1.4, and the max value of VR GLSARRL was multiplied by 1.2, and the minimal max was used as the max value for min–max scaling. This was repeated after realigning to the canonical traces and segmenting the VRs. Unless otherwise specified, 'normalized' refers to z-score normalization, as in 'normalized current' when comparing a model signal with experimental signals.

**Signal alignment.** To align signals, we used DTW<sup>61</sup> and normalized the DTW distances by dividing by the sum of the lengths of the two signals. To describe the similarity of a set of traces, we computed the DTW distance between all pairs of traces. In *t*-distributed stochastic neighbor embedding (*t*-SNE) plots, we then clustered traces on the vector of its DTW distances to all other traces. To create ensemble traces, we first identified the trace with the lowest mean DTW distance to all other traces and stretched it to create  $\mathbf{T}_{\text{medoid}} = [t_1, t_2, \dots, t_n]$ , where  $n$  is the mean length of all traces. We then DTW-aligned every other trace to  $\mathbf{T}_{\text{medoid}}$  and created  $\mathbf{T}_{\text{consensus}} = [\text{median}(\text{alignments to } t_1), \text{median}(\text{alignments to } t_2), \dots, \text{median}(\text{alignments to } t_n)]$ . Ensemble traces in Fig. 1c, Fig. 5b and Extended Data Fig. 9d show all traces aligned to the  $\mathbf{T}_{\text{consensus}}$ , but do not plot  $\mathbf{T}_{\text{consensus}}$ .

**Protein sequence-to-signal model.** To describe the amino acids, we used their volumes<sup>62</sup> and their charges at pH 7.6, at which the histidine residue is assumed to be neutral. The volume of phosphoserine

was estimated as  $126.6 \text{ cm}^3 \text{ mol}^{-1}$ , on the basis of a linear regression of molecular weight to volume of the other residues. The model signal,  $S = [S_1, S_2, \dots, S_{n-19}]$ , of amino acid sequence  $[aa_1, aa_2, \dots, aa_n]$  is calculated by computing the signal for each of the  $n-19$  windows of width 20 (Extended Data Fig. 5a–d). The vector  $\mathbf{X}_i$  describes the window starting at index  $i$  in the sequence. The  $j$ -th index in  $\mathbf{X}_i$  is  $1 + V_c \times \text{volume}(aa_{i+j}) + P_c \times \text{PositiveCharge}(aa_{i+j}) + N_c \times \text{NegativeCharge}(aa_{i+j})$ , for  $0 \leq j < 20$ , where the functions PositiveCharge and NegativeCharge take 1 if the residue has a positive or negative charge, respectively, and 0 otherwise. The constants representing weights between charge and volume,  $V_c = -3.9 \times 10^{-3}$ ,  $N_c = 4.08 \times 10^{-1}$  and  $P_c = -8.16 \times 10^{-2}$ , were determined empirically to minimize the average post-DTW distance of a training subset of protein traces to the model of their sequences. To weight the values in  $\mathbf{X}_i$ , we use a vector  $\mathbf{PW}$  (parabolic weight) of length 20 containing values representing a negative, centrally positioned parabolic curve. The  $i$ -th index in  $S$  is then finally computed as the dot product of  $\mathbf{X}_i$  and  $\mathbf{PW}$ .

**ClpX step identification.** For this analysis, the signals were not scaled or downsampled. They were filtered with a low-pass Bessel filter with  $N = 10$  and  $W_n = 0.7$ . For this analysis, YY dips were extracted manually, including portions of the signal that would otherwise be considered part of the VR in this study, to best capture the entire portion for which the double tyrosines contribute to the signal. The number of residues per YY dip was calculated as  $pw/d$ , where  $p$  is the mean proportion of the total translocation dwell time spent in these regions (0.318; Extended Data Fig. 3a),  $w$  is the total number of reading windows in the sequence (359; Extended Data Fig. 1) and  $d$  is the number of YY dips per read (6). We primarily used a Bayesian-based algorithm<sup>63</sup> to identify steps, unless otherwise noted. When applying this algorithm, a minimum length of 10 observations and a threshold of 18 was used. A total of 776 YY-dip regions were analysed, comprising 45% of all the YY dips in the dataset, omitting dips affected by potential backstepping (non-monotonic steps) or excessive noise. This selection was made by excluding YY dips that did not follow the pattern of the mean of each segmented step monotonically decreasing to the minimum and then monotonically increasing. A secondary *t*-test-based algorithm<sup>64</sup> was also used to confirm the results of the stepping rate, which was used in a different study of ClpX stepping behaviour<sup>65</sup>. When using the *t*-test-based algorithm, a minimum window length of 10 observations and a threshold *P*-value of  $5 \times 10^{-5}$  were used, and a total of 456 dips were analysed.

**YY segmentation.** To identify the YY dips and VRs, a single PASTOR trace was segmented manually into each coloured section in Fig. 2a, and the remainder of the traces were aligned to it with DTW. The corresponding regions were assigned the label from the one manually segmented trace (Supplementary Fig. 4). For PASTOR-phos, two canonical traces were segmented manually, and the rest of the traces were aligned to both, and then labels were assigned according to the canonical trace with the lowest DTW distance.

**VR classification.** We used scikit-learn to develop and test classical machine learning models and Pytorch to develop and test convolutional neural-network models. The test set was composed of all current traces from a given set of experiments to create an out-of-sample test set. The set of test experiments was selected using linear programming (Python package Pulp) to ensure at least 12 VRs with each amino acid in the test set, and minimizing the test set size. We decided to use 12 because it gave the closest to an 80–20 train–test split: 79.6% of the VRs were in the training set and 20.4% were in the testing set (full counts are shown in Extended Data Table 1a). In classification tasks for which only VRs corresponding to a subset of amino acids were used, the test set was composed of a subset of this test set. We performed hyperparameter tuning with scikit-optimize on the training set using 5-fold cross-validation. The optimal parameters were: *n\_estimators* = 250,

*min\_samples\_leaf* = 2, *max\_features* = ‘log2’, *max\_depth* = 20, *ccp\_alpha* = 0.0001, *class\_weight* = ‘balanced\_subsample’ and *criterion* = ‘gini’. All the results in Fig. 3b,c, Extended Data Fig. 6, Extended Data Table 2 and Supplementary Fig. 9 are from models evaluated on the test set. All the VRs containing an asparagine with a maximum transformed value above 1.3 had their labels changed to aspartate. In training all classical models, we upsampled minority classes, such that there was an equal representation of all classes in the training set. When training the convolutional neural network (CNN) in Extended Data Fig. 6c, we weighted the loss inversely proportional to each label’s class representation in the training set. To featurize the VRs, we performed principal component analysis on the vector of its DTW distances to all VRs in the training set to reduce the size of the vector to 64. We also used the median, max, middle, mean, dip, mean absolute value of the derivative and median absolute value of the derivative of the transformed signals, as well as the standard deviation of the raw (unfiltered, unscaled) signal. The CNN had the transformed signal as input. It was trained with a stochastic gradient descent optimizer with a learning rate of 0.01, had four convolutional layers followed by a gated recurrent unit (GRU) and then a fully connected layer, and was initialized with Kaiming initialization. Max pooling and a ReLU activation function were applied after each convolutional layer. The dummy classifier was implemented with the scikit-learn dummy classifier with default parameters.

**Reread simulation.** To collect the results shown in Extended Data Fig. 7d,e, we used a random forest without hyperparameter tuning and used 100 randomly selected 80–20 train–train splits. This was necessary to estimate the accuracy well enough with a large number of rereads, given the data limitation and the need to group samples in the test set.

**Barcode error correction.** To calculate the accuracy of barcode identification when using linear error-correcting codes, we started with our accuracy,  $p_{\text{VR}}$ , of identifying a VR given an alphabet size,  $a$ , of 2, 4, 8 or 16. For a given  $a$  and number of VRs,  $L$ , we calculated the number of bits,  $n = L \times \log_2(a)$ , that could be encoded in a protein. We simulated the accuracy with error correction,  $p'$ , when  $n-k$  of the bits were allocated to linear error-correcting codes, for all integers  $k = 1$  to  $n$ . We did this by conducting 50,000 trials of: first, encoding a random integer from 0 to  $2^k$  with a generating matrix into a message of  $n$  bits; second, randomly and independently, with probability  $p_{\text{VR}}$ , changing each of the  $n/\log_2(a)$  consecutive sets of  $\log_2(a)$  bits in the encoded message (to a different set of bits of the same length) to simulate misclassifying one VR; and third, decoding the number with syndrome decoding. We calculated  $p'$  to be the percentage of trials in which the decoded number was the same as the original random number.

**Phosphorylation detection.** Each section (C-terminal linker, VR V, VR GLSARRL, VR A and N-terminal linker) was extracted with YY segmentation. For each section, the transformed current was aligned to the model of all possible phosphorylation states, shown in Supplementary Fig. 12. We determined the number of phosphorylations in each section by the number of phosphorylations in the best-matching (lowest DTW distance) phosphorylation-state model (Supplementary Table 2) to the actual trace. When describing the signal increase in VR GLSARRL caused by PKA (Extended Data Fig. 8a), only the portion of the section up to the  $(n/3)$ -th index, where  $n$  is the length of the YY-segmented VR GLSARRL, was used because that is where PKA causes the signal to increase, as seen in Fig. 6b.

#### Null-hypothesis tests

All PERMANOVA tests were done on the DTW distance matrix of signals using scikit-bio and  $10^6$  permutations, unless we used a Bonferroni correction, in which case  $n \times 10^6$  permutations were used, where  $n$  is the number of comparisons performed. Kruskal–Wallis, *T* and

# Article

Mann–Whitney *U* tests were performed using SciPy. Reported *P* values were multiplied by *n* if we noted that we used a Bonferroni correction. All tests were two-sided unless stated otherwise, and *P* values were considered significant if *P* < 0.05.

## Materials availability

Protein expression plasmids are available at Addgene.

## Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

Data are available on GitHub at <https://github.com/uwmisl/PASTOR-sequencing> and Zenodo at <https://doi.org/10.5281/zenodo.12713840> (ref. 66). Custom MinION MinNOW run scripts can be obtained from Oxford Nanopore Technologies on request.

## Code availability

Code for the analyses is available on GitHub at <https://github.com/uwmisl/PASTOR-sequencing> and Zenodo at <https://doi.org/10.5281/zenodo.12713840> (ref. 66).

60. Martin, A., Baker, T. A. & Sauer, R. T. Rebuilt AAA + motors reveal operating principles for ATP-fuelled machines. *Nature* **437**, 1115–1120 (2005).
61. Giorgino, T. Computing and visualizing dynamic time warping alignments in R: the dtw package. *J. Stat. Softw.* **31**, 1–24 (2009).
62. Zamyatnin, A. A. Amino acid, peptide, and protein volume in solution. *Annu. Rev. Biophys. Bioeng.* **13**, 145–165 (1984).

63. Schreiber, J. & Karplus, K. Segmentation of noisy signals generated by a nanopore. Preprint at *bioRxiv* <https://doi.org/10.1101/014258> (2015).
64. Carter, N. J. & Cross, R. A. Mechanics of the kinesin step. *Nature* **435**, 308–312 (2005).
65. Sen, M. et al. The ClpXP protease unfolds substrates using a constant rate of pulling but different gears. *Cell* **155**, 636–646 (2013).
66. Kontogiorgos-Heintz, D., Yang, S., & Nivala, J. PASTOR-sequencing. Zenodo <https://doi.org/10.5281/zenodo.1271384060> (2024).

**Acknowledgements** We thank A. Martin and K. Dong (University of California, Berkeley) for discussions on ClpX stepping activity; J. Graham, K. Sabharwal, J. Wallace, R. Gutierrez and others at Oxford Nanopore Technologies for discussions and for providing the configurable MinION transcript. Funding was provided by US National Institutes of Health grant R01HG012545 (to J.N.); National Science Foundation grant EF-2021552 (J.N.); Oxford Nanopore Technologies-sponsored research agreement AM03 (J.N.); Longevity Impetus Grant from Norn Group and Hevolution Foundation (J.N.); Japan Society for the Promotion of Science overseas research fellowships (K.M.); AMED scholarship program for young researchers related to drug discovery and development (K.M.); Marilyn Fries endowed regental fellowship in computer science and engineering (D.K.H.); and a Fulbright Science and Innovation Graduate Award (M.T.).

**Author contributions** K.M., J.W., K.K., G.R., O.E.F., Y.F., M.T. and N.C. performed wet-lab experiments. D.K.-H. developed the data-processing pipelines. D.K.-H. and S.Y. developed the computational models. K.M., D.K.-H., J.W., K.K., M.J. and M.Q. performed data analysis. J.N. supervised the project. J.N. and K.M. conceived the project. K.M., D.K.-H. and J.N. wrote and edited the manuscript.

**Competing interests** The University of Washington has filed provisional patent applications covering protein rereading (K.M. and J.N.) and sequence-to-signal simulation methods (D.K.-H., M.Q., and J.N.). J.N. is a consultant to Oxford Nanopore Technologies and holds share options in the company. The other authors declare no competing interests.

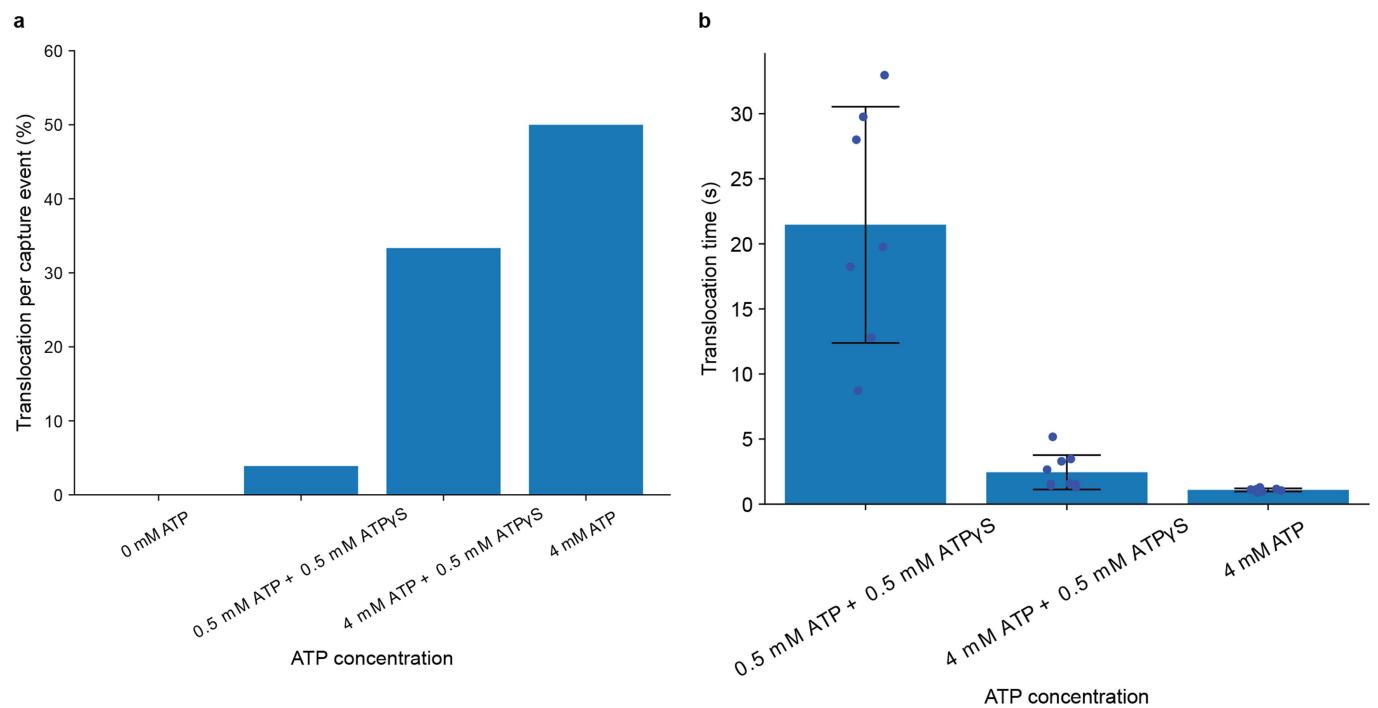
## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41586-024-07935-7>.

**Correspondence and requests for materials** should be addressed to Jeff Nivala.

**Peer review information** *Nature* thanks Aleksei Aksimentiev, David Rodriguez-Larrea and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.

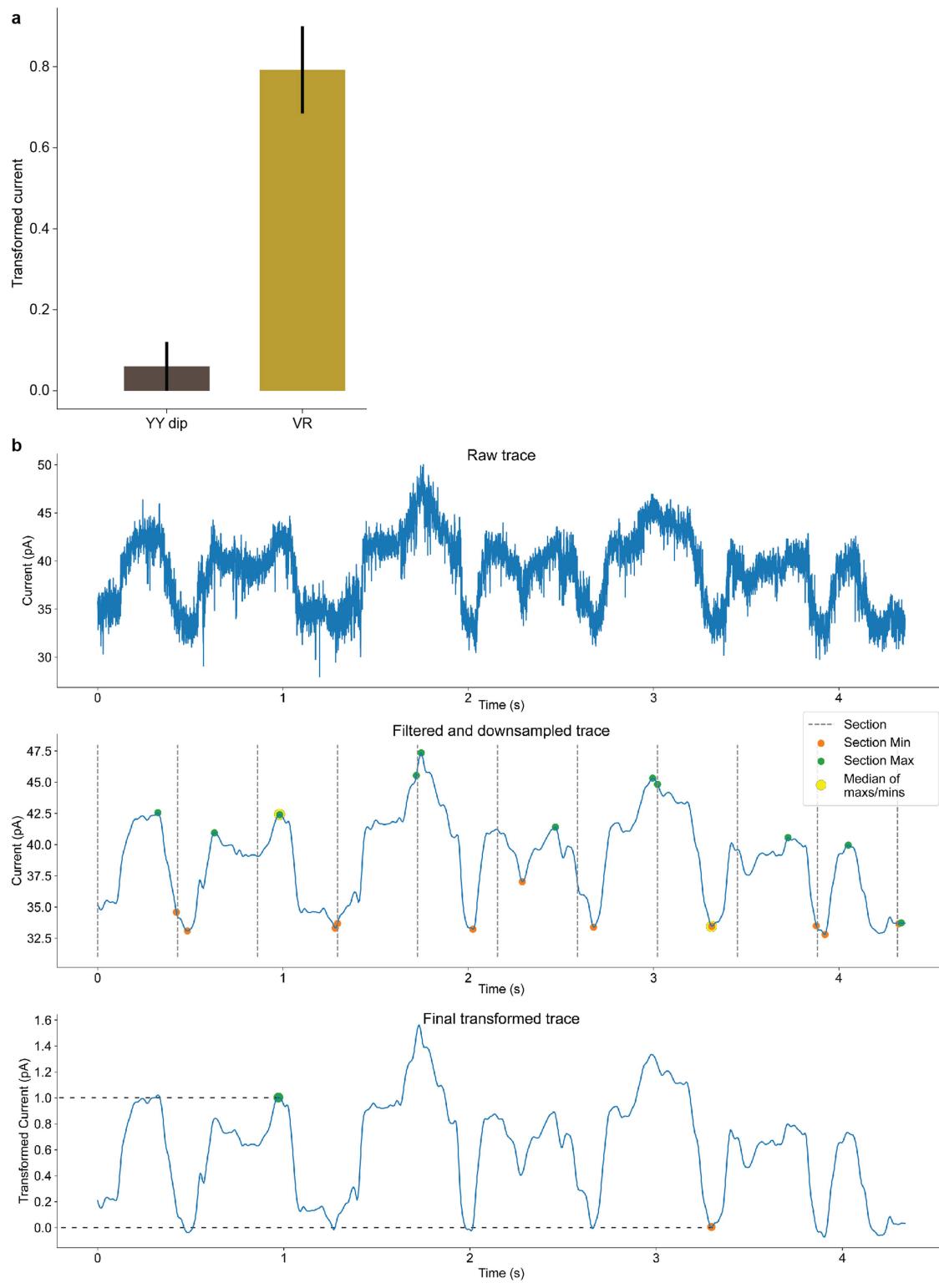
**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



**Extended Data Fig. 1 | ClpX-mediated translocation.** **a**, Fraction of ClpX-mediated translocation events observed following capture events in the presence of no ATP ( $n = 230$ ), 0.5 mM ATP + 0.5 mM ATP $\gamma$ S ( $n = 180$ ), 4 mM ATP + 0.5 mM ATP $\gamma$ S ( $n = 27$ ), or, 4 mM ATP ( $n = 16$ ). **b**, ClpX-mediated

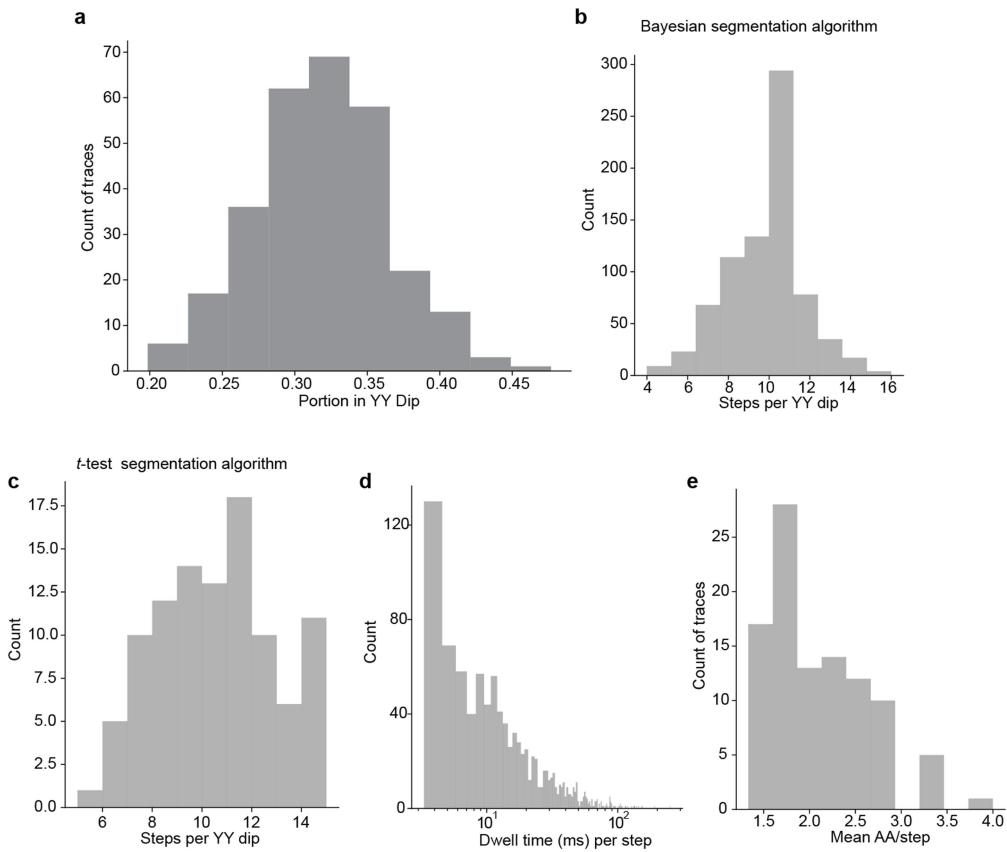
translocation time in the presence of 0.5 mM ATP + 0.5 mM ATP $\gamma$ S ( $n = 7$ ), 4 mM ATP + 0.5 mM ATP $\gamma$ S ( $n = 9$ ), or 4 mM ATP ( $n = 8$ ). Error bars denote standard deviations.

# Article



**Extended Data Fig. 2 | Consistency of YY dips and VRs in PASTORs enables scaling of ionic current traces.** **a**, Mean transformed current levels of YY and VR PASTOR segments. Error bars denote standard deviation.  $n=1828$  for YY

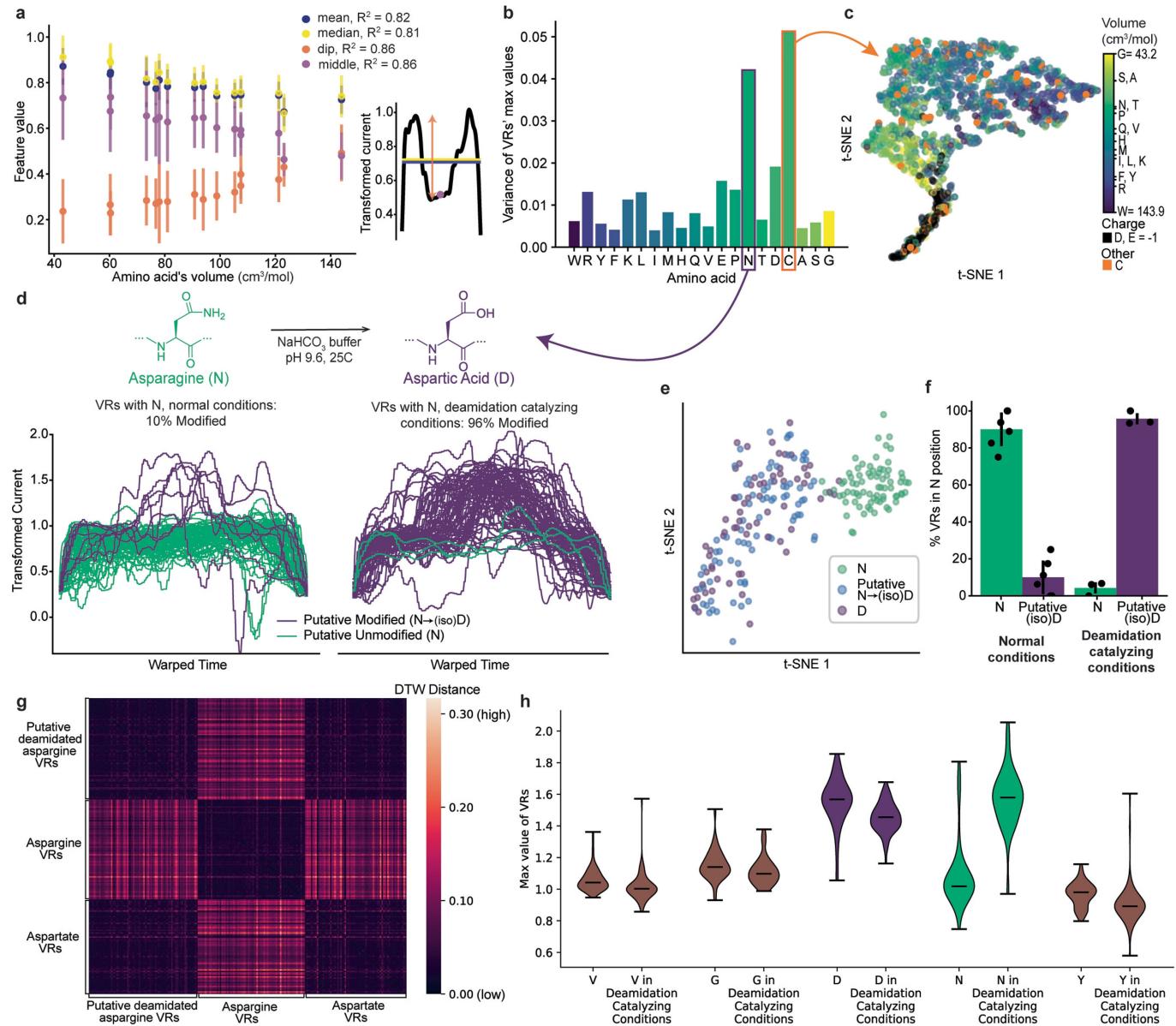
dips and 1525 for VRs. There was a total of 305 PASTOR traces analyzed. **b**, Depiction of the process of scaling signals to the “transformed” current described in Methods.



**Extended Data Fig. 3 | ClpX stepping behavior.** **a**, Distribution of the proportion of time (out of the total duration of the signal) spent within the manually segmented YY-dip regions, for  $n = 305$  traces. Mean is 0.318 and median is 0.319. This portion was used to estimate ClpX's step size (Methods). **b**, Number of steps for each of the YY dips without back steps using

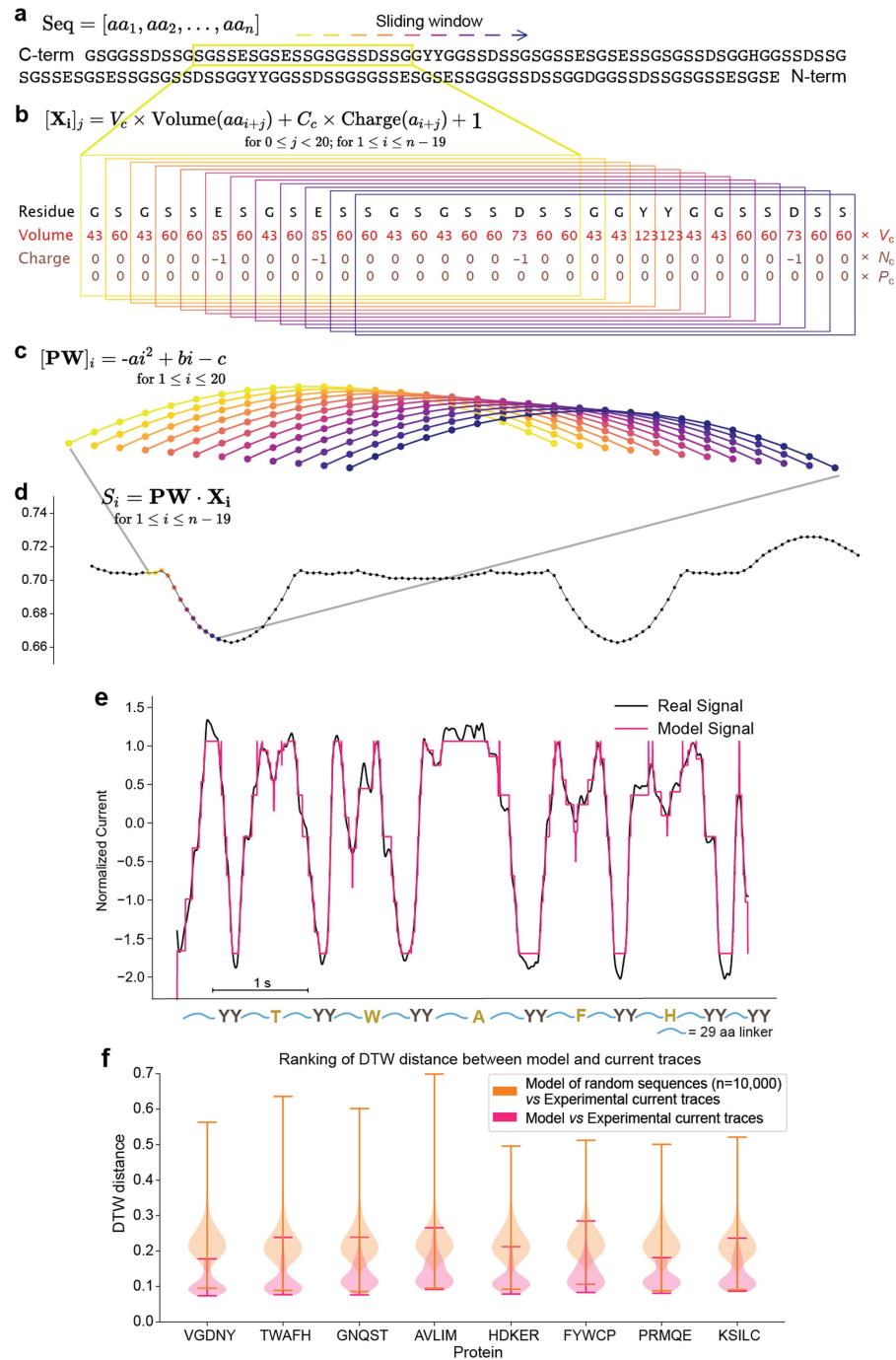
Bayesian-based YY-segmentation,  $n = 776$ . **c-e**, Stepping behavior statistics when calculated with t-test segmentation method,  $n = 456$ . Note that this  $n$  is different from **b**, because with the different segmentation algorithm, different putative backsteps were found and subsequently different dips were filtered (Methods).

# Article



**Extended Data Fig. 4 | Variable regions in PASTOR.** **a**, Scatter plot of various features of the uncharged VRs, with error bars denoting one standard deviation, center point denoting the mean, and explanation of the features to the right. **n** of VRs, traces, and experiments shown in panels **a–c** shown in Extended Data Table 1. **b**, Bar blot of the variance of the max value of the transformed VRs corresponding to each amino acid. **c**, t-SNE map showing clustering of the pairwise DTW distance between each amino acid, with all amino acids other than D, E, and C being colored by the volume, the negative amino acids colored black, and C being highlighted in orange. **d**, Plot of all the VRs corresponding to asparagine in normal conditions (left) and in conditions that catalyze the deamidation of asparagine to aspartate (right). Lines colored teal if the max value of the transformed signal  $<1.3$ , and purple otherwise.  $n = 81$  for normal conditions and 77 for deamidation conditions. **e**, t-SNE plot as in **c**, showing only asparagine and aspartate VRs. Asparagine VRs are colored blue if the max

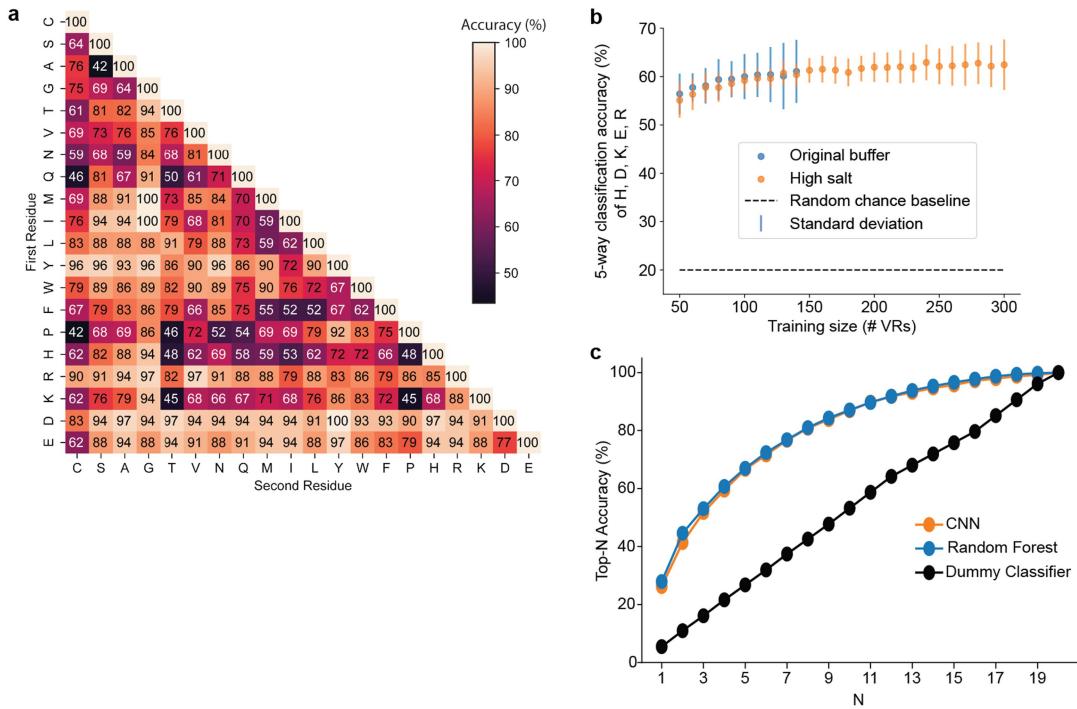
value of the transformed signal  $>=1.3$ , and green otherwise. Asparagine VRs form a distinct cluster from aspartate and putative deamidated asparagine VRs ( $p_{\text{PERMANOVA}} < 1 \times 10^{-6}$ ). Putative deamidated asparagine and aspartate are indistinguishable ( $p_{\text{PERMANOVA}} = 0.8$ ). **f**, Bar plot displaying mean percent of mutations that have been putatively deamidated or not (same threshold as in **d**, **e**) in VRs corresponding to asparagine, across technical replicates with  $n = 6, 4, 3$ , and 3 from left to right. Error bars denote standard deviations. **g**, Distance matrix of the DTW-distances between the aspartate, asparagine, and putative post-translationally modified asparagine to aspartate VRs shown in **e**. **h**, Violin plots showing distribution of the maximum height of transformed VRs, in normal and deamidation catalyzing conditions, for asparagine (N, green), and aspartate (D, purple), and the three other amino acid substitutions in PASTOR-VGDNY (valine, glycine, and tyrosine, brown). Horizontal lines denote min, median, and max.  $n = 88, 77, 81, 77, 68$ , and 77 from left to right.



**Extended Data Fig. 5 | A biophysical model for simulating nanopore ionic current traces directly from protein sequence.** **a–d**, Description of model signal generation. **a**, A protein sequence to be modeled. **b**, Calculation of the volume and charge, scaled, for all amino acids in the window of size 20. **c**, Parabolic weighting of the values within a window. **d**, Plotting the value  $S$  for each window, by computing the dot product of the parabolic weight array and the window array, to create the full model signal. **e**, Comparison between the nanopore signal of an example ionic current signal of PASTOR-TWAFH (black

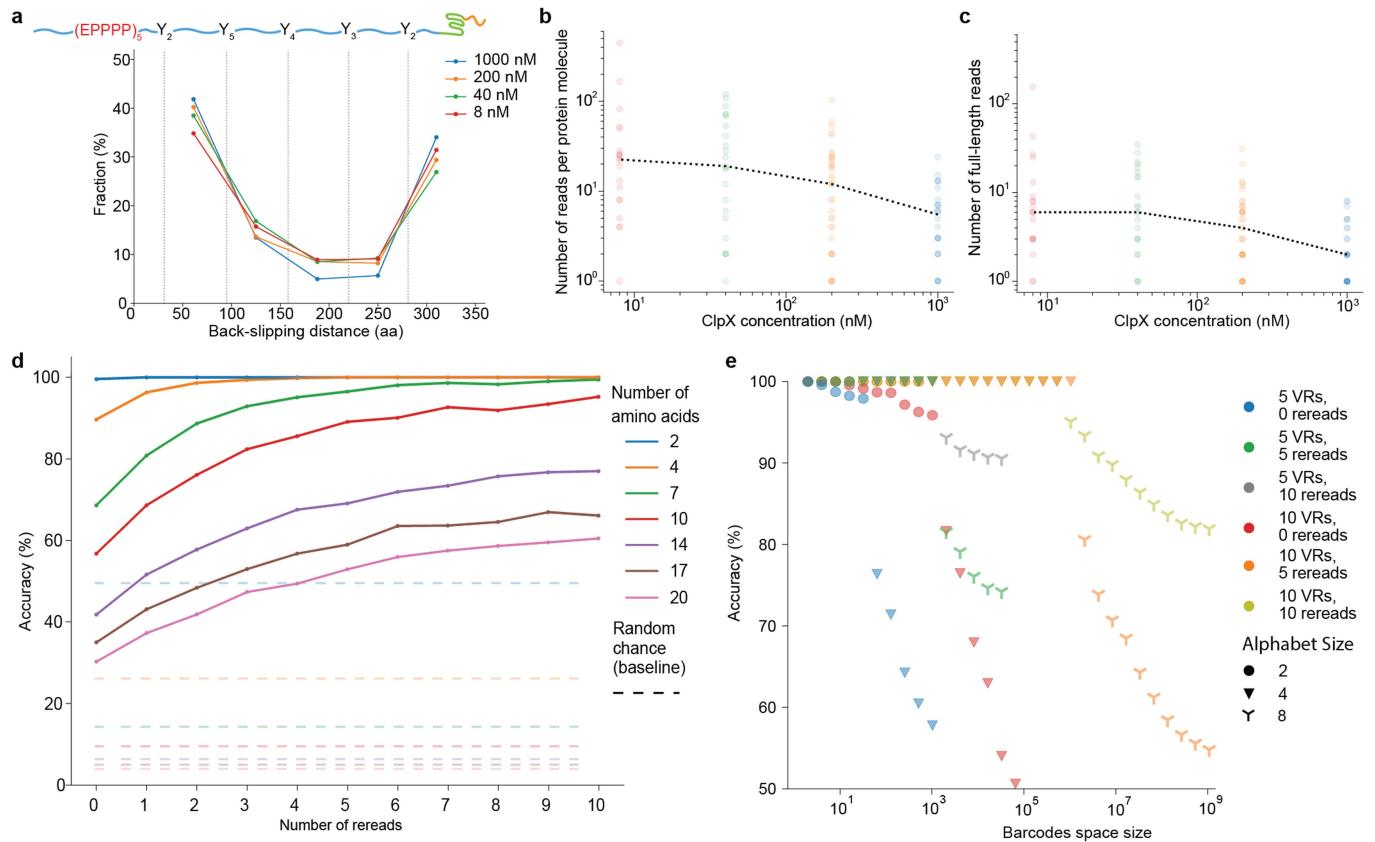
line) and the modeled signal generated for the same protein sequence (pink line). Model signal shown with the time axis aligned to the experimental trace using DTW. **f**, Distributions of the DTW distances between the real (experimental) signal traces and the model signals of the same sequence (pink), or between the real signal traces and the model signals of 10,000 random sequences derived from the same amino acid distribution as the real sequence (orange). n of experimental traces ranges from 27 to 55.

# Article



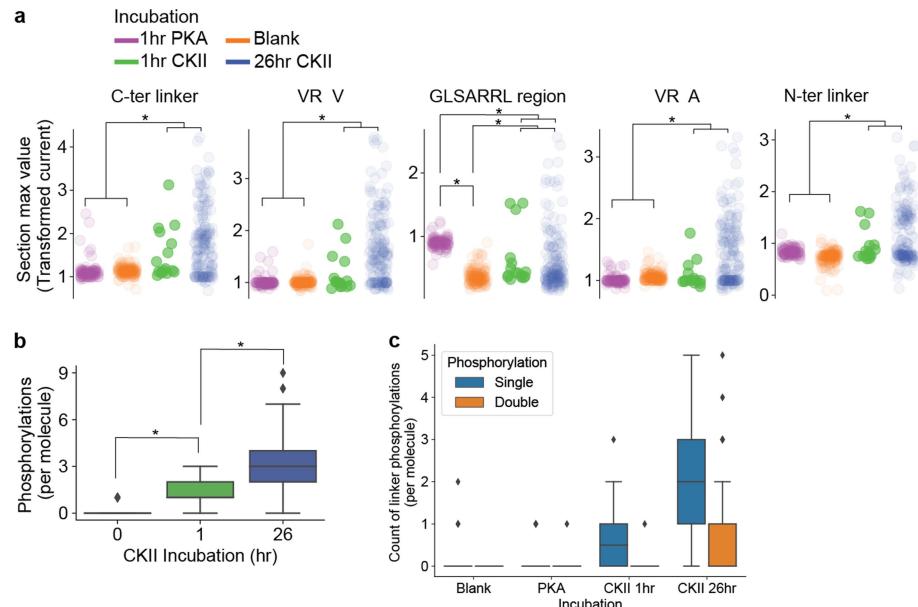
**Extended Data Fig. 6 | Classification of single-amino acid mutations with a Random Forest model.** **a**, Heatmap each pairwise classification accuracy by a Random Forest model evaluated on the fixed test set, as in main Fig. 3b, with values. **b**, The accuracy of 5-way classification of the H, D, K, E, and R VRs with various training sizes, to compare the quality of data with different buffer conditions. Each condition was trained and tested on 100 different Random Forest models, each trained on a random train-test split. The extra data was

allocated to the testing set. The original buffer data is the data used in Figs. 2, 3, with  $n$  in Extended Data Table 1. Both conditions consist of 2 independent runs. The models' performance was consistent across both standard and elevated salt conditions. **c**, Accuracy in a 20-way classification when "accuracy" is defined as the correct label being in the top-N most probable classes. The dummy classifier chooses one label at random. Results averaged over 20 models.



**Extended Data Fig. 7 | Rereading with an unfoldase slip sequence and estimating its impact on barcode sequencing accuracy.** **a**, Estimated back-slipping distance for ClpX concentrations at 1000 nM ( $n = 141$ ), 200 nM ( $n = 609$ ), 40 nM ( $n = 777$ ), and 8 nM ( $n = 999$ ). The very first full-length read (Read 1) of each analyte protein molecule was excluded from this analysis. **b** and **c**, Number of all reads and full-length reads per PASTOR-reread molecule, respectively. The dotted lines indicate medians for ClpX concentrations at 1000 nM ( $n = 26$ ), 200 nM ( $n = 37$ ), 40 nM ( $n = 23$ ), and 8 nM ( $n = 20$ ). **d**, Simulated

effect of rereading on 2 (Y, D), 4 (A, W, R, D), 7 (G, Q, W, F, R, D, E), 10 (A, G, V, N, Y, W, F, R, D, E), 14 (C, A, G, T, V, N, Q, M, Y, W, F, R, D, E), 17 (C, S, A, G, T, V, N, Q, M, I, Y, W, F, H, R, D, E), and 20-way (all 20 a.a.) classification tasks, compared to a baseline random classifier. Each value is the average over 100 train-test trials. **e**, Projected sequencing accuracy of barcode designs using the accuracies from d. Points of the same color represent different amounts of bits allocated to error correcting codes (see Methods).

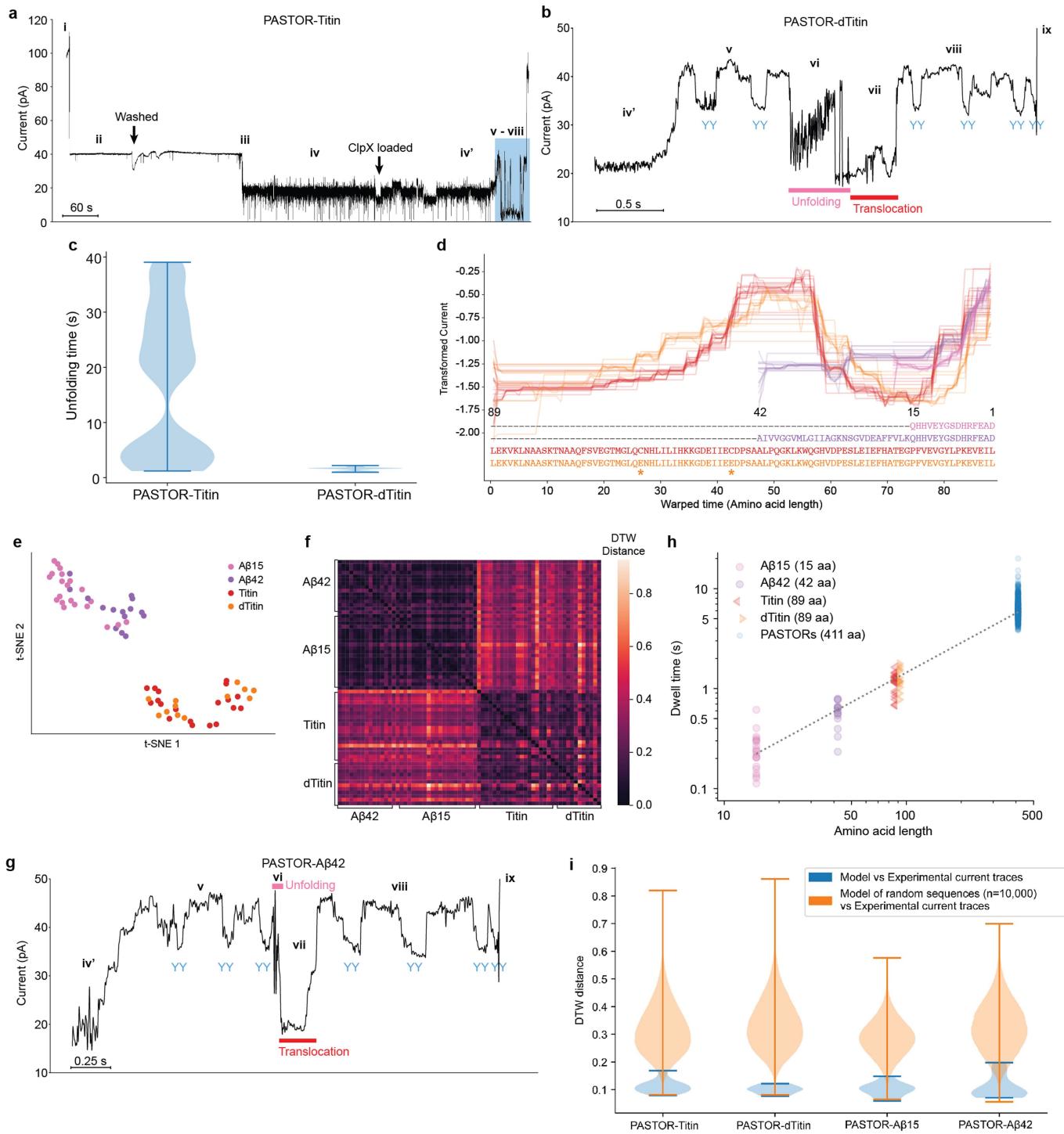


**Extended Data Fig. 8 | Reading kinase phosphorylation activity on single protein molecules.** **a**, Maximum transformed signal value for each trace. Transparency of each scatter point is proportional to the  $n$  of traces for that condition. For each of C-terminal linker, VR V, VR A, and N-terminal linker conditions, the CKII incubation conditions' maximum values were significantly higher than the blank and PKA incubation conditions ( $p_{\text{Mann-Whitney, one-sided}} < 10^{-8}$  for each, after Bonferroni correction). GLSARRL region corresponds to the C-terminal third of the VR. The GLSARRL region's maximum values are significantly higher in the PKA than the blank incubation condition ( $p_{\text{Mann-Whitney, one-way}} = 5 \times 10^{-39}$ , after Bonferroni correction) and the two CKII incubation conditions are significantly different from each of the two other conditions ( $p_{\text{Mann-Whitney}} < 10^{-5}$  for each, after Bonferroni correction).

**b**, Interquartile range (IQR) of the number of putative phosphorylations per molecule. No kinase incubation shows fewer phosphorylations than 1 hr incubation ( $p_{\text{Mann-Whitney, one-way}} = 4 \times 10^{-16}$ ), and 1 hr incubation in CKII shows fewer

phosphorylations than 26 hr incubation ( $p_{\text{Mann-Whitney, one-way}} = 5 \times 10^{-6}$ ). Center line, box, whiskers, and diamonds represent median, IQR, 1.5IQR, and outliers, respectively. **c**, Interquartile ranges of putative linker phosphorylations per molecule for each of the kinase incubation conditions, corresponding to single or double phosphorylations on a linker. Center line, box, whiskers, and diamonds represent median, interquartile range (IQR), 1.5IQR, and outliers, respectively. CKII 26 hr incubation molecules have significantly more putative single phosphorylations than the CKII 1 hr incubation condition ( $p_{\text{Mann-Whitney, one-sided, Bonferroni corrected}} = 0.002$ ) and the blank and PKA incubation conditions ( $p_{\text{Mann-Whitney, one-sided, Bonferroni corrected}} = 3 \times 10^{-28}$ ). CKII 26 hr incubation molecules also have significantly more putative double phosphorylations than the CKII 1 hr incubation condition ( $p_{\text{Mann-Whitney, one-sided, Bonferroni corrected}} = 0.01$ ) and the blank and PKA incubation conditions ( $p_{\text{Mann-Whitney, one-sided, Bonferroni corrected}} = 3 \times 10^{-58}$ ). For all panels,  $n$  of traces and experiments in Extended Data Table 3.

\* $P < 10^{-5}$ .



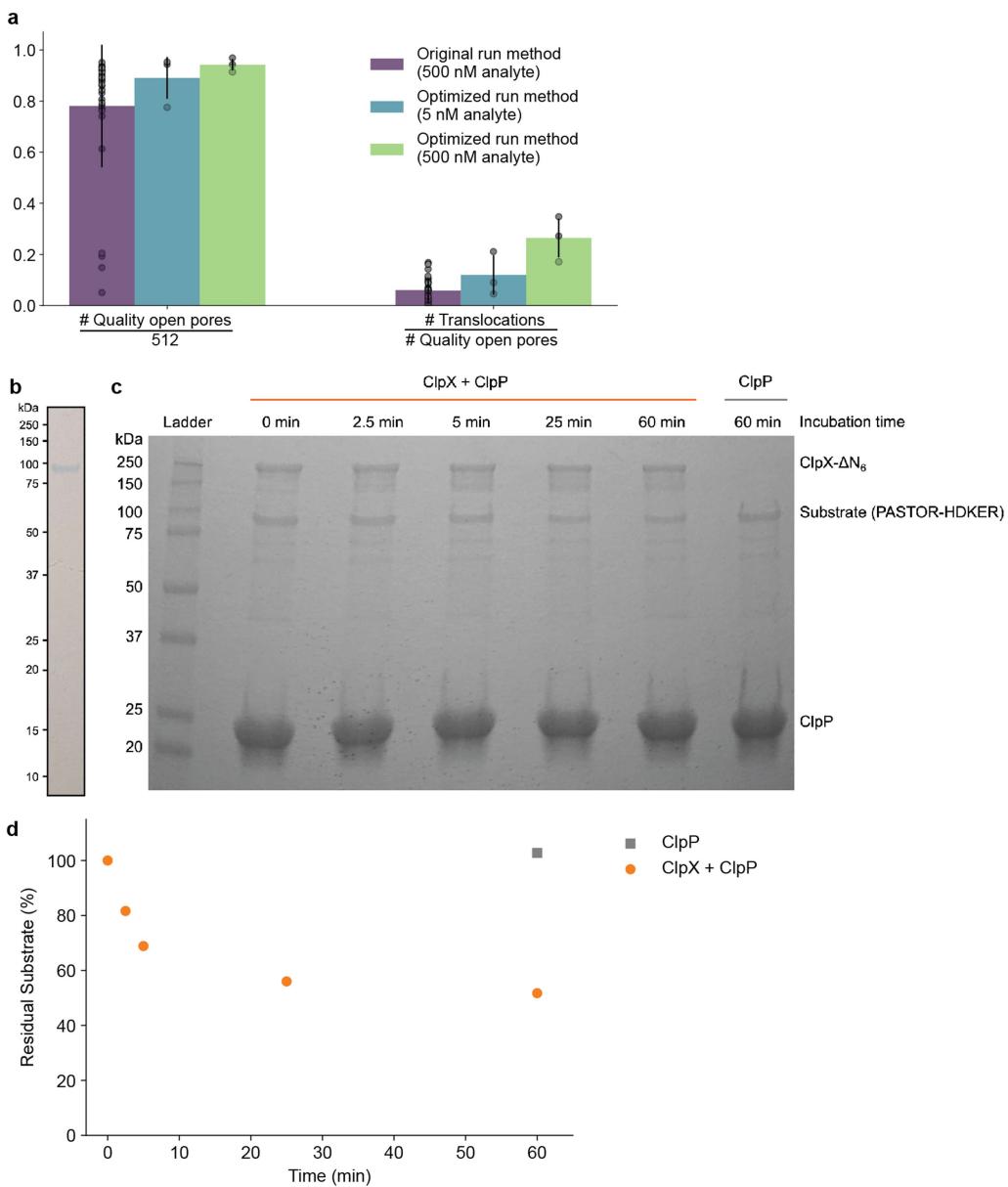
**Extended Data Fig. 9** | See next page for caption.

# Article

## Extended Data Fig. 9 | Processive reading of folded protein domains.

**a**, Example trace of PASTOR-Titin, not zoomed into translocation state (open pore to open pore). Roman numerals correspond to states described in main Fig. 6a. **b**, Example trace of PASTOR-dTitin. **c**, Distribution of total unfolding time for Titin ( $n = 21$ ) and dTitin ( $n = 14$ ). **d**, Ensemble traces of state vii of PASTOR-A $\beta$ 15 (pink,  $n = 21$ ), -A $\beta$ 42 (purple,  $n = 15$ ), -Titin (red,  $n = 20$ ), and -dTitin (orange,  $n = 12$ ). Protein sequences are shown in the C-to-N direction, and asterisks represent the C47E and C63E mutations between Titin and dTitin. **e**, t-SNE plot based on pairwise DTW distances for state vii, showing A $\beta$ 15 and A $\beta$ 42 form a distinct cluster from Titin and dTitin ( $p_{\text{PERMANOVA}} \leq 1 \times 10^{-6}$ ). A $\beta$ 15 vs A $\beta$ 42 and Titin vs dTitin states vii are indistinguishable ( $p_{\text{PERMANOVA}} = 0.99, 0.67$ , respectively). **f**, Distance matrix of the DTW-distances between the traces of A $\beta$ 15, A $\beta$ 42, Titin and dTitin, shown in e. **g**, Example trace of PASTOR-A $\beta$ 42. **h**, Relationship between protein length and translocation time. State vii dwell time is plotted for A $\beta$ 15, A $\beta$ 42, Titin, and dTitin, as well as translocation time for

the 8 PASTORs with no folded domain insert ( $n = 672$ ). The dotted line was fitted with the mean dwell times of each protein class (slope corresponds to a translocation rate of 16 ms/aa or 63 aa/sec,  $R^2 = 0.998$ ). **i**, Distributions of the DTW distances of each of the protein translocations for folded domain proteins to model signal(s). In blue, they are compared to the model signal of the protein sequence, and in orange, they are each compared to the model of 10,000 random sequences derived from the same sequence distribution. The protein translocations include the regions corresponding to the folded domain translocation (state vii) and the N-terminal half of the PASTOR YY dips and VRs (state viii). The signals corresponding to the C-terminal half of the PASTOR YY dips and VRs (state v) and the folded domain unfolding (state vi) are excluded from the analysis, because the model does not predict unfolding patterns (main Fig. 6b).  $n = 20, 12, 21$ , and 15 for PASTOR-Titin, PASTOR-dTitin, PASTOR-A $\beta$ 15, and PASTOR-A $\beta$ 42, respectively.



**Extended Data Fig. 10 | Quantification of ClpX-mediated protein translocations.** **a**, Yield of nanopore runs on the R9.4.1 flow cell. Quality open pore denotes a pore that was consistently at open pore current at the time of analyte loading. The R9.4.1 flow cell has a maximum of 512 pores available for measurement and the number of quality open pores used for measurement fluctuates depending on the flow cell condition. The initial run method with an analyte concentration of 500 nM (purple) includes PASTOR and PASTOR-phos data. The optimized run method with an analyte concentration of 5 nM (blue) includes PASTOR-HDKER data. The optimized run method with an analyte concentration of 500 nM (green) includes PASTOR-HDKER and PASTOR-phos data.  $n = 35, 3$ , and  $3$ , for the original run (500 nM analyte), optimized run (5 nM analyte), and optimized run (500 nM analyte) conditions, respectively. The number of translocations per quality open pore was significantly ( $p_{t\text{-test, one-sided}} =$

$5 \times 10^{-8}$ ) higher for the optimized run method (500 nM analyte) condition than the original run method (500 nM analyte), and the other comparisons were non-significant ( $p_{t\text{-test, original vs. 5nm optimized}} > 0.05$ ;  $p_{t\text{-test, one-sided, 5nM optimized vs 500nM optimized}} > 0.05$ ). Error bars denote standard deviations. **b**, SDS-PAGE analysis of purified PASTOR-HDKER protein. The protein band appears at a higher position on the gel than the actual molecular weight of the protein (50.2 kDa) due to its highly net negatively-charged state. **c**, Bulk ClpXP degradation assay on purified PASTOR-HDKER protein. The substrate protein was incubated with an ATP regeneration mix and ClpP in the presence or absence of ClpX. **d**, Residual PASTOR substrate was quantified based on the peak area of the PASTOR-HDKER protein band normalized by the ClpP protein band on each lane using ImageJ software. Raw gels shown in Supplementary Figs. 15 and 16.

# Article

**Extended Data Table 1 | PASTOR data**

**a**

Residue	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Total	82	63	69	79	56	81	63	89	70	89	98	81	72	91	79	82	75	88	56	62
Training set	65	51	52	62	44	65	46	72	53	72	81	65	60	75	62	66	59	71	44	50
Testing set	17	12	17	17	12	16	17	17	17	17	17	16	12	16	17	16	16	17	12	12

**b**

PASTOR	HDKER	GNQST	FYWCP	AVLIM	VGDNY	TWAFH	PRMQE	KSILC
Traces	36	48	29	55	33	27	43	34
Independent Experiments	2	3	6	3	3	4	3	5

PASTOR data relate to Figs. 2f and 3, Extended Data Figs. 4a–c, 5f, 6, 7e,f; Extended Data Table 2; and Supplementary Table 1. **a**, Counts of variable regions. **b**, Counts of traces and technical replicates for each PASTOR.

Extended Data Table 2 | Set-wise classification accuracy

Set Size	Random Forest Accuracy (%)	AAs
2	100	Y, D
3	95.0	G, Y, D
4	86.3	A, W, R, D
5	85.5	G, V, W, R, D
6	70.6	C, G, L, Y, R, D
7	65.9	G, Q, W, F, R, D, E
8	64.0	A, T, I, Y, W, R, D, E
9	59.0	G, V, N, L, Y, W, R, D, E
10	53.8	A, G, V, N, Y, W, F, R, D, E
11	50.0	C, A, G, V, N, Y, W, F, R, D, E
12	47.1	C, A, G, T, V, Q, I, Y, W, R, D, E
13	44.8	A, G, T, V, N, Q, M, Y, W, F, R, D, E
14	40.4	C, A, G, T, V, N, Q, M, Y, W, F, R, D, E
15	37.7	C, S, G, T, V, N, Q, M, Y, W, F, R, K, D, E
16	35.6	C, A, G, T, V, N, Q, M, L, Y, W, F, R, K, D, E
17	32.7	C, S, A, G, T, V, N, Q, M, I, Y, W, F, H, R, D, E
18	29.1	C, S, A, G, T, N, M, I, L, Y, W, F, P, H, R, K, D, E
19	28.9	C, S, A, G, T, V, N, Q, M, I, L, Y, W, F, P, R, K, D, E
20	28.3	all 20

Accuracy of a hyperparameter-optimized Random Forest classifying VRs when the train and test set was composed only of VRs corresponding to the mentioned amino acids.

# Article

**Extended Data Table 3 | Counts of PASTOR-phos**

Condition	Blank/No kinase (1 hr)	PKA (1 hr)	CKII (1 hr)	CKII (26 hr)
Traces	155	92	16	171
Independent Experiments	2	2	1	3

Corresponding author(s): Jeff Nivala

Last updated by author(s): July 10, 2024

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

#### Data collection

All experiments were performed on the MinION platform using R9.4.1 flow cells. Run conditions were set with a custom MinNOW script (available from Oxford Nanopore Technologies) at a temperature of 30°C and a constant voltage of -140 mV with a 3 kHz sampling frequency, except for initial proteins P1–P4, in which runs were performed at a constant voltage of -180 mV with a 10 kHz sampling frequency. Using the priming port, flow cells were first washed with 1 mL cis running buffer (200 mM KCl, 5 mM MgCl<sub>2</sub>, 10% glycerol, and 25 mM HEPES-KOH, pH 7.6) and then loaded with 200  $\mu$ L of protein analyte in cis running buffer at a final concentration of 500 nM unless otherwise specified. Following observation of protein captures in the pores, flow cells were washed with 1 mL of cis running buffer to remove uncaptured proteins and subsequently loaded with 75  $\mu$ L of cis running buffer supplemented with 4 mM ATP and 200 nM of ClpX-ΔN6 unless otherwise specified. The flow cell was washed ~4 min after the analyte loading in the initial method while the flow cell was washed ~6 min and ~2 min after the loading of the analyte at the concentration of 5 nM and 500 nM, respectively, in the optimized method (Supplementary Fig. 29). For MinION runs in the high salt condition (Supplementary Fig. 18), a buffer containing 400 mM KCl, 5 mM MgCl<sub>2</sub>, and 25 mM HEPES-KOH (pH 7.6) was used in place of standard cis running buffer.

**Data analysis**

Custom code written for this paper can be found at <https://github.com/uwmisl/PASTOR-sequencing> and <https://doi.org/10.5281/ZENODO.12713840>. The analysis pipeline for a PASTOR sequencing run begins with extracting the segments of the raw nanopore signal that contain translocations. We used a script to assist in this process, which we are including in the Github repository, but was assisted by manually checking each translocation. PASTORs are then scaled, transformed, segmented, and featurized, using code that we are including in the Github repository. The code that was used to train a Random Forest, train the CNN, to perform rereading accuracy simulations, and to perform barcode error correction are also included in the Github repository. It also contains the code for the sequence to signal model. The segmentation code to study ClpX stepping with both approaches, the normalization code used for PASTOR-phos, and the DTW medoid calculation and alignment code are available on the Github as well.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

**Data**

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

We are including the raw signals of all segments used for analysis in <https://github.com/uwmisl/PASTOR-sequencing/tree/main/data> and <https://doi.org/10.5281/ZENODO.12713840> as .json files, which can be loaded as a dataframe in python. Because of the large files size, we are including one example unprocessed .fasta file, the direct output from the MinION device, on the Github and Zenodo.

**Research involving human participants, their data, or biological material**

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

**Reporting on sex and gender**

*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used. Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*

**Reporting on race, ethnicity, or other socially relevant groupings**

*Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status). Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.) Please provide details about how you controlled for confounding variables in your analyses.*

**Population characteristics**

*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*

**Recruitment**

*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*

**Ethics oversight**

*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

**Field-specific reporting**

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences       Behavioural & social sciences       Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://nature.com/documents/nr-reporting-summary-flat.pdf)

**Life sciences study design**

All studies must disclose on these points even when the disclosure is negative.

**Sample size**

No sample size calculation was performed.

Data exclusions	Traces with substantial noise were discarded prior to analyses
Replication	Experiments were replicated as indicated, typically using technical replicates n=>3.
Randomization	Not applicable to this study; does not include subjects that require allocation into experimental groups.
Blinding	Not applicable to this study; does not include subjects that require allocation into blinded experimental groups.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems		Methods	
n/a	Involved in the study	n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies	<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines	<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology	<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Plants		