nature methods

Article

https://doi.org/10.1038/s41592-023-02021-8

Unambiguous discrimination of all 20 proteinogenic amino acids and their modifications by nanopore

Received: 26 January 2023

Accepted: 21 August 2023

Published online: 25 September 2023

Check for updates

Kefan Wang[1,2,4], Shanyu Zhang[1,2,4], Xiao Zhou[3,4], Xian Yang[1,2], Xinyue Li[1,2], Yuqin Wang[1,2], Pingping Fan[1,2], Yunqi Xiao[1,2], Wen Sun[1,2], Panke Zhang[1], Wenfei Li[3] & Shuo Huang[1,2]

Natural proteins are composed of 20 proteinogenic amino acids and their post-translational modifications (PTMs). However, due to the lack of a suitable nanopore sensor that can simultaneously discriminate between all 20 amino acids and their PTMs, direct sequencing of protein with nanopores has not yet been realized. Here, we present an engineered hetero-octameric *Mycobacterium smegmatis* porin A (MspA) nanopore containing a sole Ni^{2+} modification. It enables full discrimination of all 20 proteinogenic amino acids and 4 representative modified amino acids, N^{ω},N^{ω}-dimethyl-arginine (Me-R), *O*-acetyl-threonine (Ac-T), N^{4}-(β-*N*-acetyl-D-glucosaminyl)-asparagine (GlcNAc-N) and *O*-phosphoserine (P-S). Assisted by machine learning, an accuracy of 98.6% was achieved. Amino acid supplement tablets and peptidase-digested amino acids from peptides were also analyzed using this strategy. This capacity for simultaneous discrimination of all 20 proteinogenic amino acids and their PTMs suggests the potential to achieve protein sequencing using this nanopore-based strategy.

Proteins are important executors of life activities[1] but only a few techniques, such as Edman degradation[2] and mass spectrometry[1], have the capacity to determine the amino acid sequence of proteins. Detection limits in protein sequencing also hinder the characterization of low-abundance proteins. A single-molecule protein sequencer could provide improved sensitivity and information of post-translational modifications (PTMs). Nanopore, a versatile single-molecule sensor that has enabled remarkable progress in nucleic acid sequencing, has become a promising candidate. Although significant efforts were made to achieve nanopore translocation of proteins, no sequence information could be obtained solely from uncontrolled protein translocation[3]. Following a nanopore-induced phase-shift sequencing (NIPSS) strategy[4,5], a peptide–oligonucleotide conjugate can be scanned by a nanopore to report trace signatures containing sequence-dependent peptide information. This approach is, however, still hindered by nanopore resolution, which is insufficient for reliable protein sequence decoding due to the complexity of the sequence combination of the 20 proteinogenic amino acids[6–8].

An alternative approach is to sequence protein in a sequencing by hydrolysis approach, in which peptidase-digested amino acids are read sequentially by a nanopore, similar to that demonstrated with a proteasome nanopore[9]. This, however, requires a nanopore that can identify all proteinogenic amino acids as well as their PTMs unambiguously, and this has not yet been achieved. Previously, a Cu^{II}-phenanthroline modified α-hemolysin (α-HL) nanopore was shown to have achieved direct identification of five pairs of amino acid enantiomers[10]. It failed, however, to simultaneously discriminate between all 20 amino acids due to the insufficient resolution of α-HL. An aerolysin nanopore was

[1]State Key Laboratory of Analytical Chemistry for Life Sciences, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing, China. [2]Chemistry and Biomedicine Innovation Center (ChemBIC), Nanjing University, Nanjing, China. [3]Collaborative Innovation Center of Advanced Microstructures, National Laboratory of Solid State Microstructure, Department of Physics, Nanjing University, Nanjing, China. [4]These authors contributed equally: Kefan Wang, Shanyu Zhang, Xiao Zhou. e-mail: shuo.huang@nju.edu.cn

https://doi.org/10.1038/s41592-023-02021-8

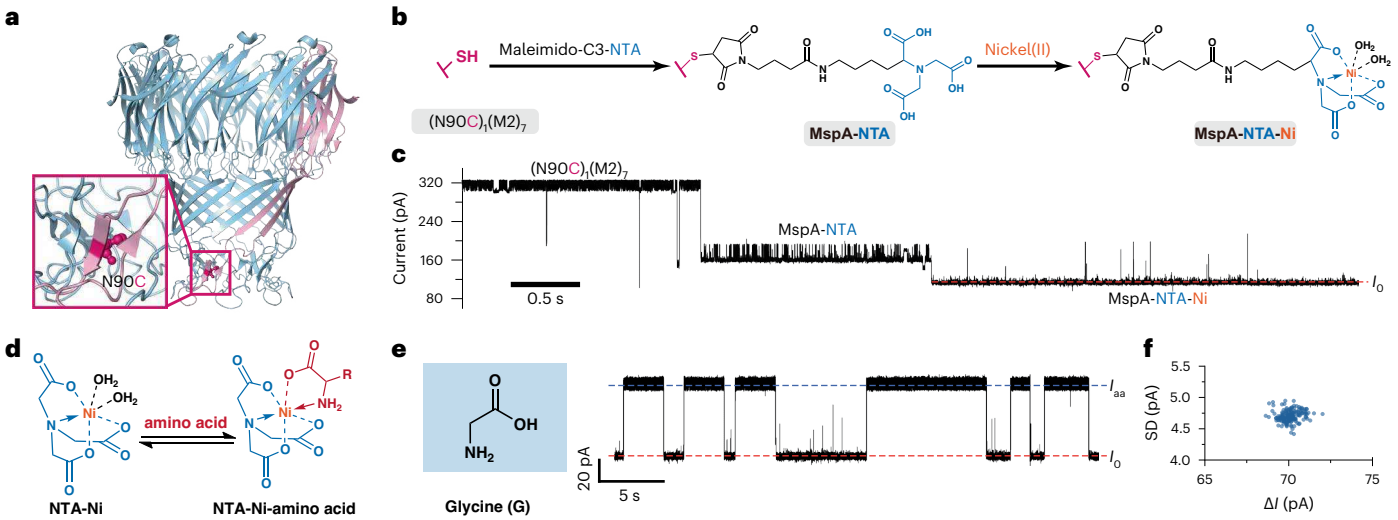


Fig. 1 | Construction of a Ni-NTA-modified nanopore for amino acid sensing. **a**, The structure of $(N90C)_1(M2)_7$. $(N90C)_1(M2)_7$ is a hetero-octameric MspA containing a sole cysteine residue (pink) at site 90 in one of its monomeric components. **b**, The construction of a Ni-NTA-modified nanopore. Maleimido-C3-NTA reacts with the cysteine residue of $(N90C)_1(M2)_7$ by a maleimide-thiol reaction to form MspA-NTA. A Ni^{2+} was subsequently chelated by MspA-NTA. For simplicity, this nickel-modified pore is referred to as MspA-NTA-Ni. **c**, Real-time characterization of Ni-NTA modification monitored by single-channel recording. The measurements were performed as described in Methods. Maleimido-C3-NTA was added to *cis* at a final concentration of 200 μM for NTA modification. Afterwards, nickel sulfate was added to *trans* with a final concentration of 50 μM to trigger nickel chelation. The success of each reaction step results in an abrupt decrease in the current amplitude and a change in the current noise. Here, I_0 stands for the open pore current of MspA-NTA-Ni. **d**, The mechanism of amino acid sensing using MspA-NTA-Ni. **e**, A representative trace of amino acid sensing by MspA-NTA-Ni. Glycine was used as the model amino acid. With a continually applied potential of +100 mV and the addition of glycine to *cis* with a final concentration of 2 mM, nanopore events appearing as reversible switching between I_{aa} and I_0 were immediately observed. For demonstration, the trace was Butterworth low-pass filtered with a cut-off frequency of 300 Hz. **f**, The scatter plot of ΔI versus SD of events acquired as described in **e**. ΔI is defined as $\Delta I = I_{aa} - I_0$. A total of 165 events were used to generate the plot ($n = 165$).

also used in the discrimination of octapeptides containing a single terminal amino acid difference[11]. The analytes of this approach are short peptides rather than standalone amino acids and it was stated that only 13 out of 20 peptides of this kind were identified[11]. Some other approaches of nanopore amino acid identification have been reported, but direct identification of all 20 proteinogenic amino acids has still not been realized[12]. Single-molecule identification of amino acids may be performed by recognition tunneling[13], but the reported event discrimination is still unsatisfactory. The consistency of a manufactured tunneling junction device and its coupling to a nanopore sensor also pose other technical challenges.

Mycobacterium smegmatis porin A (MspA)[14] is a conically shaped biological nanopore that is used widely in nanopore sequencing of nucleic acids[4,5]. An engineered MspA can also be used as a nanoreactor that can monitor single-molecule chemical reactions. Ions and small molecules, such as tetrachloroaurate(III)[15], neuron transmitters[16], anti-COVID-19 drugs[17], catecholamine enantiomers[18], monosaccharides[19], nucleoside monophosphates[20] and alditols[21], have been identified using MspA nanopores containing suitable reactive adapters. Inspired by immobilized metal-affinity chromatography (IMAC)[22], in which a nickel-nitrilotriacetic acid (Ni-NTA) affinity column is used to purify recombinant proteins containing a hexahistidine tag, a hetero-octameric MspA containing a sole Ni-NTA adapter at its pore constriction was designed and prepared for amino acid sensing. Although Ni-NTA modification applied to the whole internal lumen of solid state nanopores was previously reported in the detection of histamine[23] and His-tagged proteins[24], a biological nanopore containing a sole Ni^{2+} modification has not been reported to date.

Construction of MspA-NTA-Ni

To introduce a single NTA adapter site-specifically to the pore constriction of MspA, a hetero-octameric MspA mutant, also referred to as $(N90C)_1(M2)_7$ (Fig. 1a), was first prepared (Methods)[19–21]. $(N90C)_1(M2)_7$, which consists of one monomeric subunit containing a sole cysteine and seven monomeric subunits lacking any cysteine, was previously generated for nanopore modification of maleimide derivatives by a Michael addition reaction[19–21]. In this work, a maleimido-C3-nitrilotriacetic acid (maleimido-C3-NTA) reacts with the cysteine residue of $(N90C)_1(M2)_7$, so that a sole NTA adapter is site-specifically introduced to the pore constriction. For simplicity, this NTA-modified MspA hetero-octamer is referred to here as MspA-NTA (Fig. 1b). A nickel ion (Ni^{2+}) can then be chelated by the NTA adapter of MspA-NTA to form an MspA nanopore containing a sole Ni^{2+} located at its pore constriction. For simplicity, this Ni^{2+}-modified MspA is referred to here as MspA-NTA-Ni.

All of the reaction processes described above were monitored in real time with single-channel recording (Fig. 1c). Experimentally, the measurement was performed with a custom measurement chamber with two compartments each containing a buffer of 1.5 M KCl and 10 mM *N*-cyclohexyl-2-aminoethanesulfonic acid (CHES) at pH 9.0 (Methods). By convention, the electrically grounded compartment is defined as *cis* and the opposite compartment is defined as *trans*. A transmembrane voltage of +100 mV was applied continually. During single-channel recording, the open pore current of a single $(N90C)_1(M2)_7$ measures ~310 pA. At this stage, fluctuating noise (5.95 ± 0.19 pA) was also observed due to the existence of an unmodified cysteine at the pore constriction[25]. Maleimido-C3-NTA was added to *cis* to reach a final concentration of 200 μM. Immediately afterwards, an abrupt and irreversible current drop of ~150 pA was recorded, indicating the success of the NTA modification and the generation of MspA-NTA (Fig. 1c). The open pore current of MspA-NTA also reflects dynamic switching between two major current levels (158 ± 2 pA and 189 ± 2 pA), probably due to the existence of an unoccupied NTA adapter. After further addition of nickel sulfate to *trans* with a final concentration of 50 μM, another irreversible current drop of ~50 pA was observed, confirming the success of Ni^{2+} binding to MspA-NTA and the formation of MspA-NTA-Ni (Fig. 1c). At this stage, although some transient spike

noise was seen, dynamic current fluctuations previously observed with $(N90C)_1(M2)_7$ and MspA-NTA were no longer observable, confirming that the previously observed current switching of MspA-NTA was due to the presence of the unoccupied NTA, and that the Ni^{2+} was now tightly bound to the NTA adapter. During the time-extended measurement, no further irreversible current drop was observed, confirming that only one NTA adapter exists in the pore lumen and that the MspA-NTA-Ni contains only a sole Ni^{2+} modification. In this condition, the newly formed MspA-NTA-Ni remained unchanged in continuous measurements of ~3 h (Supplementary Fig. 1). Accordingly, the preparation of MspA-NTA-Ni is well characterized at the single-molecule level. The conductance features of $(N90C)_1(M2)_7$, MspA-NTA and MspA-NTA-Ni were also recorded for future reference (Supplementary Fig. 2 and Supplementary Table 1). Given that there is only a single modification site on the hetero-octamer $(N90C)_1(M2)_7$ and that the modified NTA can bind only a single Ni^{2+}, the nanopore conductance corresponding to the state of MspA-NTA and MspA-NTA-Ni is independent of the concentration, respectively, of the maleimido-C3-NTA and nickel sulfate used during nanopore modification.

MspA-NTA can also be prepared in batches by incubating $(N90C)_1(M2)_7$ with maleimido-C3-NTA (Methods). The MspA-NTA generated in this way can be used directly without any further treatment and its reported open pore current is identical to that previously characterized by single-channel recording (Fig. 1c). Furthermore, with a single MspA-NTA inserted, addition of nickel sulfate to *trans* to a final concentration of 50 µM immediately results in the formation of MspA-NTA-Ni. With a +100 mV applied bias, the open pore current of MspA-NTA-Ni (I_0) measures ~115 pA, which is consistent with that observed in Fig. 1c.

Amino acid sensing

Amino acids, which contain both an amino and a carboxyl group, are bidentate ligands that can react reversibly with metal ions[26]. When diffusing to the pore constriction of MspA-NTA-Ni, amino acids are expected to bond with the immobilized Ni^{2+} to form a ternary complex (Fig. 1d). Given that the binding between Ni^{2+} and an amino acid is considerably weaker than that between Ni^{2+} and NTA[27,28], it is expected that the binding and the dissociation of amino acids would fail to trigger the dissociation of Ni^{2+} from the NTA adapter. Thus, this configuration permits continuous and time-extended measurement of different amino acids. To support this, glycine, the simplest amino acid, was used as a model analyte (Fig. 1e). Experimentally, the measurement was performed with batch-prepared MspA-NTA in a 1.5 M KCl buffer (1.5 M KCl, 10 mM CHES, pH 9.0) with a continually applied transmembrane voltage of +100 mV. With a single MspA-NTA in the membrane, nickel sulfate was added to *trans* to a final concentration of 50 µM, which immediately triggers the formation of MspA-NTA-Ni. Upon the addition of glycine to *cis* with a final concentration of 2 mM, successive nanopore events appearing as current fluctuations between I_0 and the event current (I_{aa}, which is larger than I_0), were observed immediately (Fig. 1e and Supplementary Video 1). Furthermore, the rate of event appearance also increases when the final concentration of glycine added to *cis* is increased from 0.5 mM to 50 mM (Supplementary Fig. 3), confirming that the I_{aa} was generated by glycine binding. Even with 50 µM added glycine, the corresponding events were still detectable but with a much lower rate of event appearance (Supplementary Fig. 4). To describe the sensing events quantitatively, core parameters such as open pore current (I_0), event current (I_{aa}), noise amplitude (SD; the standard deviation of the event noise), dwell time (t_{off}), inter-event duration (t_{on}), mean inter-event duration (τ_{on}) and mean dwell time (τ_{off}) are defined and summarized in Supplementary Fig. 5. The reciprocal of the mean inter-event duration ($1/\tau_{on}$, $n = 3$) is proportional to the concentration of glycine, which is consistent with a bimolecular model (Supplementary Fig. 3b and Supplementary Table 2). The reciprocal of the mean dwell time ($1/\tau_{off}$, $n = 3$), however, is independent of the glycine concentration, consistent with a unimolecular model (Supplementary Fig. 3b and Supplementary Table 2). Generally, the rate of glycine event appearance increases when the buffer pH is upregulated (Supplementary Fig. 6 and Supplementary Table 3), thus a pH 9.0 buffer (1.5 M KCl, 10 mM CHES, pH 9.0) was used for all subsequent measurements, when not otherwise stated.

The blockage amplitude ΔI is defined as $\Delta I = I_{aa} - I_0$. For glycine, the ΔI is measured at ~70 pA. However, amino acid sensing events acquired with a Cu(II) modified α-HL[10] measure only 2–5 pA. By contrast, MspA (ref. 29), which has a conical lumen geometry and focuses the ionic current to the pore constriction, produces a greater event amplitude for small molecules than α-HL, which has a cylindrical lumen[30]. The event scatter plot of ΔI versus the noise amplitude (that is, SD) also shows a single and narrowly distributed population of events (Fig. 1f), indicating that both of the event features of sensing are extremely consistent between events. The NTA-Ni^{2+} adapter, which chemically restricts the conformation of amino acid analytes, plays a critical role in the production of events. The larger event amplitude and the high consistency of the event features are critical in the discrimination of different amino acids, although there are only subtle differences. This sensing capacity, however, could not be achieved when an MspA containing no NTA adapter or an MspA-NTA containing no Ni^{2+} was tested (Supplementary Fig. 7). To sum up, this shows that the MspA nanopore and the NTA-Ni^{2+} adapter are pivotal in the performance of amino acid sensing. To the best of our knowledge, however, a biological nanopore containing an NTA or a Ni^{2+} modification has not been reported previously.

Discrimination of 20 proteinogenic amino acids

To show how different proteinogenic amino acids are distinguished by MspA-NTA-Ni, identical measurements were performed with various proteinogenic amino acids (Fig. 2a). In independent measurements with different amino acids, each type of amino acid produces a unique event shape (Supplementary Figs. 8–11). This is more clearly seen in Fig. 2a, in which all representative amino acid events are shown together for comparison. Generally, all amino acid sensing events are positive, that is, $I_{aa} > I_0$. In addition, the blockage levels of amino acid events all show telegraphic switching between two levels (Supplementary Fig. 12). This telegraphic switching, which generates unique event features for different amino acids, is extremely useful in the discrimination of all 20 proteinogenic amino acids, again demonstrating the importance of the NTA-Ni^{2+} adapter.

Although most amino acids produce a single type of sensing event, histidine and proline each have two types of sensing events. In histidine, its imidazole side chain[31] may also additionally bond with Ni^{2+}, generating diverse binding configurations discriminable by MspA-NTA-Ni (Supplementary Video 2). Proline is the only cyclic amino acid of the 20 proteinogenic amino acids, and the α-amino group of proline is attached directly to its side chain. This particular chemical structure may generate configurations different from that of other proteinogenic amino acids. For simplicity, the type 1 and type 2 events of histidine and proline are referred to here as H1/H2 and P1/P2, respectively.

Based on three independent measurements with each amino acid ($n = 3$), the above-described sensing events generate highly reproducible data. The generated core event parameters summarized in Supplementary Table 4 list the quantitative details. Generally, the ΔI of different amino acid events is 38–100 pA, which is a much wider range than that previously reported for α-HL, which gives a ΔI of only 2–5 pA[10]. The low resolution of α-HL is thus able to discriminate only between five pairs of amino acid enantiomers[10] and failed to achieve simultaneous discrimination of all 20 proteinogenic amino acids, which is extremely important for nanopore protein sequencing. Here, by simultaneously considering ΔI and SD (Fig. 2b), events corresponding to the binding of 20 proteinogenic amino acids are well discriminated. Although the P2 events have some overlap with the H1 events in the two-dimensional

https://doi.org/10.1038/s41592-023-02021-8

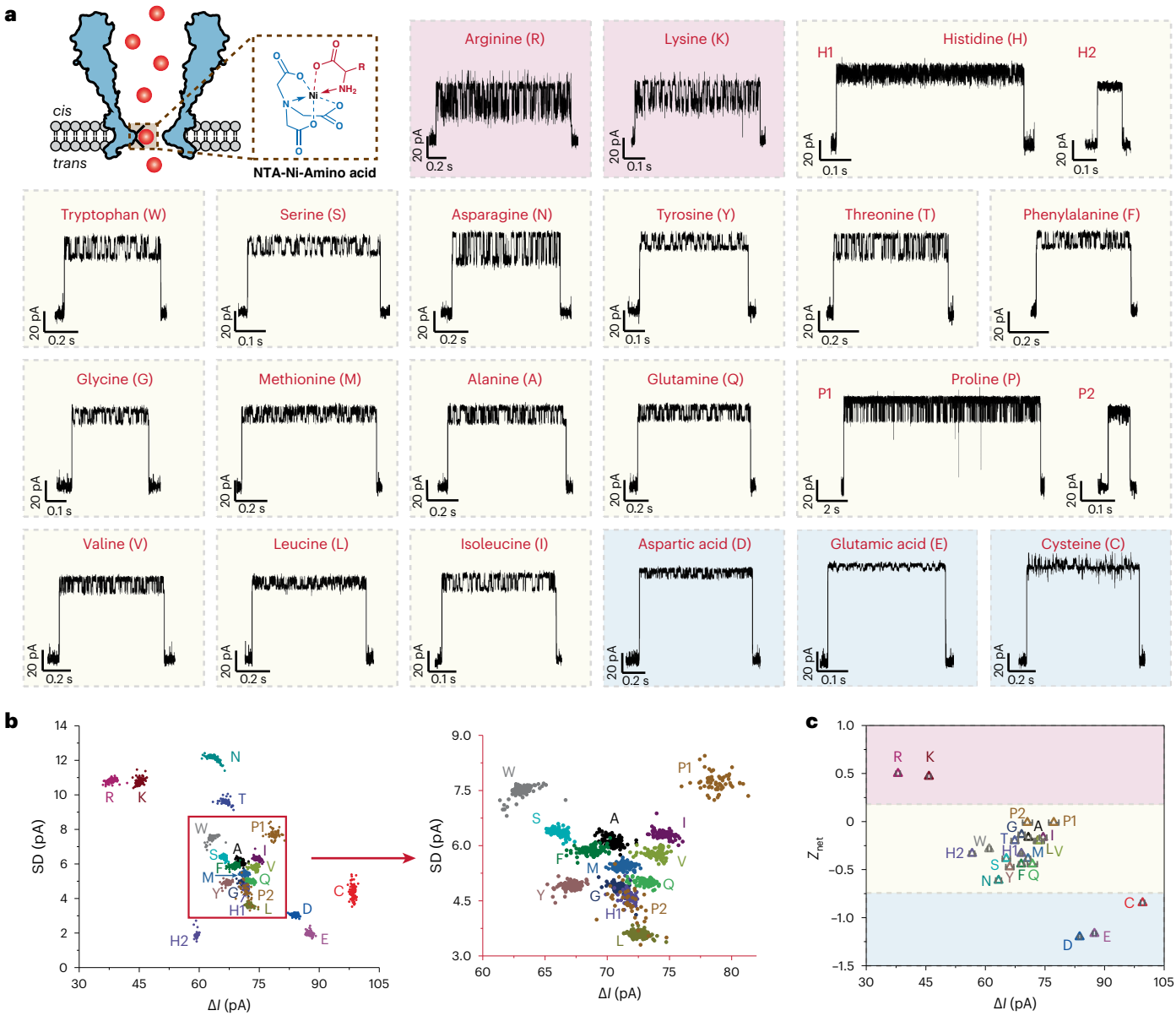


Fig. 2 | Discrimination of 20 amino acids using MspA-NTA-Ni. a, The schematics of amino acid sensing (top left) and representative events generated by different amino acids when measured with MspA-NTA-Ni. The measurements were carried out as described in Methods. A total of 20 proteinogenic amino acids were separately added to *cis* with a final concentration of 2 mM (A, C, F, G, H, K, M, N, Q, R, S, T, V, W, Y), 4 mM (D, E, I, L) or 40 mM (P) (Supplementary Figs. 8–11 and Supplementary Table 4). The final concentration of proline was set higher to compensate for its low rate of event appearance. Histidine and proline both produce two types of nanopore events, defined respectively as H1/2 and P1/2. According to their net charge (Z_{net}), all 20 amino acids were classified into three groups, in which amino acids with positive charge, weak negative charge and strong negative charge were marked with a red, yellow or blue background, respectively. **b**, The scatter plot of Δ*I* versus SD of events acquired with different amino acids. One hundred events acquired with each amino acid were used to generate the plot, according to which, most amino acid events are fully distinguishable. To clarify the detail, the events inside the red box are further zoomed in and shown on the right. Although the events corresponding to P2 and H1 appear to overlap in the plot, their event characteristics are visually different and can be discriminated when other event features such as dwell time, skewness and kurtosis are simultaneously considered. **c**, The correlation between Δ*I* and Z_{net} of amino acids. Generally, the blockage amplitude (Δ*I*) is larger when the net charge of the amino acid is more negative. The color background in the plot is consistent with that in **a**.

scatter plot, their event shapes are significantly different and can be distinguished by the different dwell times (t_{off}) (Supplementary Fig. 13). Discrimination between leucine and its isomer, isoleucine, is difficult using only mass spectrometry, but they are able to be clearly discriminated using MspA-NTA-Ni (Extended Data Fig. 1), again demonstrating the very effective resolution of MspA-NTA-Ni for amino acid sensing.

No clear correlation could be seen in the plots of Δ*I* against the volume or the molecular weight of amino acids (Supplementary Fig. 14). However, by plotting the mean Δ*I* against the net charge of different amino acids (Supplementary Table 5), it can be seen that the Δ*I* of events acquired with more negatively charged amino acids is generally larger (Fig. 2c). Also, the appearance of a negatively charged analyte at the MspA constriction, such as a carboxymethyl guanine[32], or the generation of an anionic boronate ester[16] generally indicates enhanced channel conductance. The same phenomenon was also observed in the significant reduction of channel conductance of MspA when the negatively charged aspartic acid at the pore constriction was mutated to asparagine, which is electrically neutral[14,33]. This sensitivity

https://doi.org/10.1038/s41592-023-02021-8

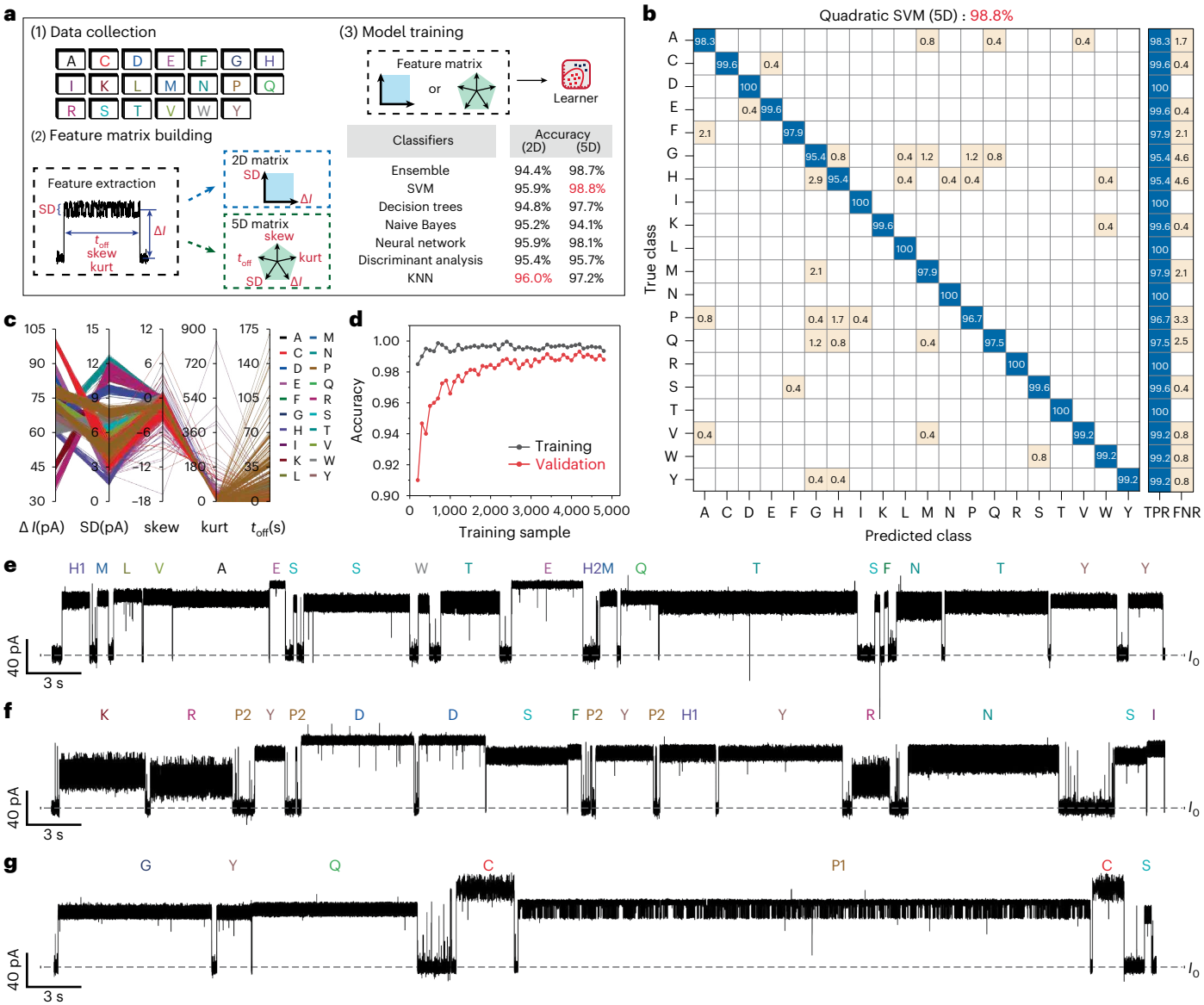


Fig. 3 | Identification of 20 amino acids by machine learning. a, The workflow of machine learning. In brief, sensing events separately acquired with 20 amino acids were collected to form a dataset. Five event features (ΔI, SD, skewness (skew), kurtosis (kurt) and t_{off}) were extracted from each event to form a feature matrix. A 2D feature matrix and a 5D feature matrix were built for machine learning. The 2D matrix contains only two features (Δ*I* and SD), similar to that in a 2D scatter plot (Fig. 2b). The 5D matrix, which contains all five features, includes more information from sensing. Machine learning was performed with the Classification Learner toolbox of MATLAB. Seven classifiers were evaluated with 10-fold cross-validation to screen the best-performing model. For the 2D matrix, the highest validation accuracy is 96.0% (Supplementary Table 6). For the 5D matrix, the highest validation accuracy reaches 98.8%, achieved by the quadratic SVM model (Supplementary Table 7). **b**, The confusion matrix of amino acid classification generated by the quadratic SVM model using the 5D feature matrix. TPR (true-positive rate) and FNR (false-positive rate) represent the correct or false classification of each true class, respectively. **c**, The parallel coordinate plots generated from the 5D feature matrix. **d**, The learning curve of the quadratic SVM model for varying sample size. **e**–**g**, Representative traces acquired during simultaneous sensing of all 20 amino acids. The measurements were performed as described in Methods. All amino acids were simultaneously added to *cis*. The final concentration of H and C was 0.1 mM. The concentration of F, M, N, T, S was 0.5 mM. The concentration of P was 20 mM. The concentration of all remaining amino acids was 1 mM. Zoomed-in views of these traces are shown in Supplementary Figs. 21–23. The events were predicted with the trained quadratic SVM model.

of MspA to charge is important in the discrimination between amino acids that are similar in mass or volume but which differ in charge, such as glutamic acid (molecular weight, 146.12; $Z_{net} = -1.28$) versus glutamine (molecular weight, 146.15; $Z_{net} = -0.66$), and arginine (volume, 188.2; $Z_{net} = +0.62$) versus phenylalanine (volume, 189.7; $Z_{net} = -0.47$).

By applying a +1 mV applied potential, which minimizes the contribution of the electrophoretic force and the electroosmotic flow, amino acid events were still clearly detectable (Supplementary Fig. 15), suggesting that the amino acids can spontaneously diffuse to the pore constriction to trigger event generation. This also explains why all 20 proteinogenic amino acids, which are differently charged, can be simultaneously detected in the same set-up (Fig. 3e–g). For the same reason, amino acid sensing can be carried out regardless of whether the amino acids were added to *cis* or *trans* (Supplementary Figs. 16–18). However, the electrophoretic force still regulates the rate of event appearance for electrically charged amino acids. With the same applied potential, the addition of electrically charged amino acids to *cis* or *trans* would produce a noticeable difference in their event detection

https://doi.org/10.1038/s41592-023-02021-8

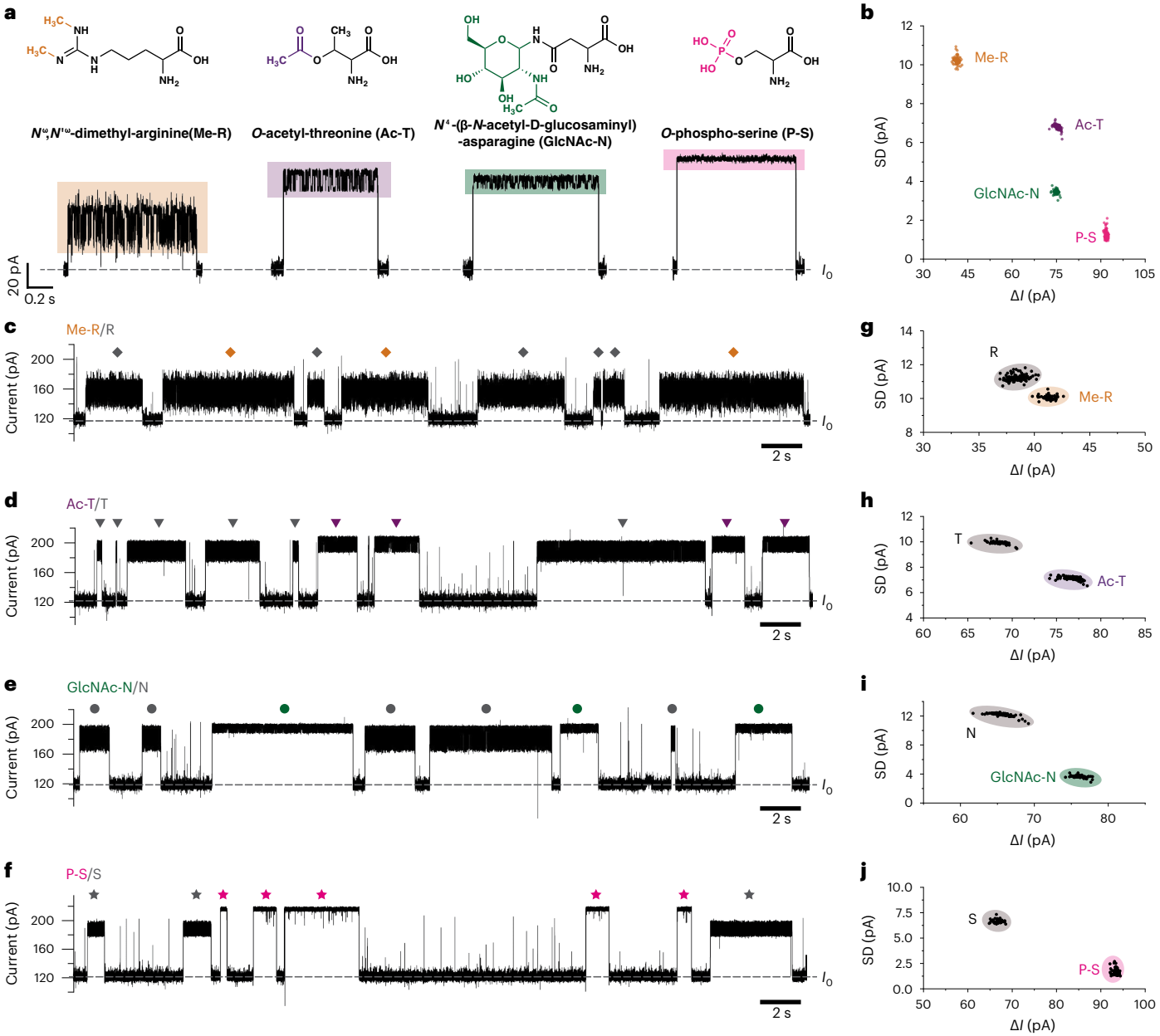


Fig. 4 | Identification of amino acids with PTMs. The measurements were performed as described in Methods. **a**, Top: amino acids with the PTMs methylation, acetylation, glycosylation and phosphorylation, as represented by N^{ω},N'^{ω}-dimethyl-arginine (Me-R), *O*-acetyl-threonine (Ac-T), N^4-(β-*N*-acetyl-D-glucosaminyl)-asparagine (GlcNAc-N) and *O*-phospho-serine (P-S), respectively. Bottom: representative nanopore events produced by the corresponding amino acids (Supplementary Fig. 25 and Supplementary Table 8). The open pore current (I_0) of MspA-NTA-Ni is marked with a dashed line. The current levels resulting from amino acid binding are marked with different color bands. **b**, The scatter plot of ΔI versus SD of events produced by the four amino acids described in **a**. The data points represent 100 events acquired for each amino acid. During the measurement, each analyte was solely added to the *cis* chamber at a final concentration of 2 mM. **c**–**f**, Representative traces acquired from simultaneous sensing of Me-R/R (**c**), Ac-T/T (**d**), GlcNAc-N/N (**e**) and P-S/S (**f**). Each amino acid was added to *cis* with a final concentration of 2 mM. Events were identified and marked with orange (Me-R), purple (Ac-T), green (GlcNAc-N), pink (P-S) or gray symbols (for the corresponding unmodified amino acids), above each event. Noticeable differences in event shapes are seen in each comparison pair. **g**–**j**, The event scatter plots of Δ*I* versus SD of events produced by Me-R/R (**g**, $n = 180$), Ac-T/T (**h**, $n = 158$), GlcNAc-N/N (**i**, $n = 180$) and P-S/S (**j**, $n = 150$). In each plot, two fully separated populations of events, generated by the unmodified and modified amino acids, respectively, are seen.

frequency (Supplementary Figs. 17 and 18). However, for electrically neutral amino acid such as glycine, the addition of amino acids to *cis* or *trans* results in a similar detection frequency (Supplementary Fig. 16).

Molecular dynamics simulations were conducted using a GROMACS package[34] for the MspA embedded in the POPC (1-palmitoyl-2-oleoylphosphatidylcholine) lipid bilayer at 300 K and 1 atm with a salt concentration of 1.5 M. An NTA adapter was established and was covalently connected to the side chain S atom at site 90 of the first monomeric subunit of the pore model according to experimental set-up. Ni^{2+} and glycine were added to the adapter, respectively, to simulate the states corresponding to NTA, NTA-Ni or NTA-Ni-Gly during a nanopore measurement. An external electric field of 0.15 V per 10 nm along the direction perpendicular to the membrane plane was applied (Methods). According to the simulation results, prior to Ni^{2+} and glycine binding, the NTA adapter tends to bend towards the *trans* side of the membrane. In this state the NTA is less conformationally confined and

https://doi.org/10.1038/s41592-023-02021-8

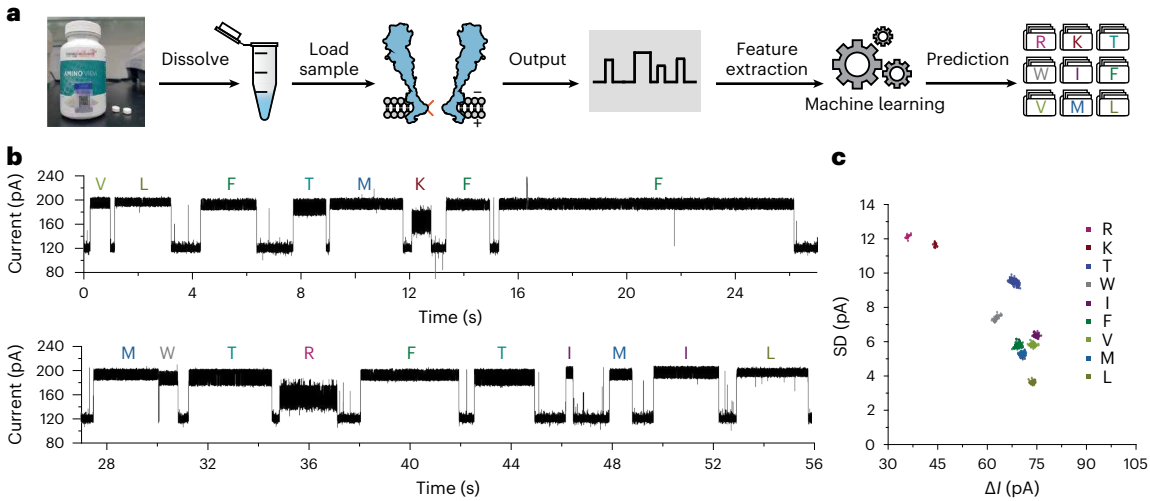


Fig. 5 | Rapid analysis of amino acid tablets using MspA-NTA-Ni. a, Schematic diagram of the workflow. The compound amino acid tablets (Kingnature) were ground into powder and dissolved in a KCl buffer (1.5 M KCl, 10 mM CHES, pH 9.0) with a final concentration of 50 mg ml^{-1}. Then, 10 µl of the solution was added to *cis*. Amino acid events were immediately observed and the event identities were predicted by machine learning. **b**, A representative trace acquired during nanopore sensing of the amino acid tablet. More details are shown in a zoomed-in view of the trace in Supplementary Fig. 28. Nanopore measurements were performed as described in Methods. All events were predicted by the previously trained quadratic SVM model. **c**, The event scatter plot of Δ*I* versus SD generated using events acquired from a 90 min continuous recording (n = 1,117) as described in **b**. Nine populations of events corresponding to R, K, T, W, I, F, V, M and L, respectively, were identified by machine learning. The identified amino acid components were consistent with that described in the product information.

may spontaneously switch between multiple conformations in the pore lumen, a phenomenon that might explain why telegraphic noise was observed at this stage during single-channel recording. However, when Ni^{2+} was bound to the NTA adapter, the Ni^{2+} induces strong interactions between amino acids in the pore lumen and the NTA adapter. This results in an extremely tightly bound configuration of the NTA-Ni adapter in the pore lumen, which explains why a low noise current level was consistently observed during single-channel recording (Fig. 1c). Also, in this state the narrow pore constriction is more occupied by the whole NTA-Ni adapter and a lower channel conductance is expected. Furthermore, upon binding with a glycine, the original strong interaction between the pore lumen and the Ni^{2+} is diminished because the Ni^{2+} is now occupied by the bound glycine. This results in a release of the whole NTA-Ni-Gly adapter from the narrowest spot of the pore constriction with a resulting increase in the channel conductance (Supplementary Figs. 19 and 20), a phenomenon that also explains why all amino acid sensing events are positive (that is, $I_{aa} > I_0$). At this stage, the NTA adapter is again loosely confined in its conformation, which might be the reason why all amino acid events generate highly fluctuating noise (Fig. 2a).

Identification of amino acids by machine learning

To automate event identification and to avoid the bias caused by human judgment, a fair and objective custom machine learning algorithm was developed for amino acid identification (Methods). The overall process of machine learning includes data collection, feature matrix building and model training (Fig. 3a). A total of 6,000 events separately acquired with different amino acids were first collected, to form a dataset. Five event features (that is, Δ*I*, SD, t_{off}, skewness and kurtosis) of the blockage level of each event were extracted using a custom MATLAB code to form a feature matrix (5D). All events in the matrix have known labels because they were separately generated from a known amino acid. The feature matrix was passed to the Classification Learner toolbox of MATLAB for training. Seven inbuilt classifiers, that is, ensemble, SVM (support vector machine), decision trees, naive Bayes, neural network, discriminant analysis and KNN (*k*-nearest neighbor) were evaluated. To avoid overfitting, the model performance was evaluated with 10-fold cross-validation. The derived quadratic SVM model, which has a 98.8% validation accuracy, was found to be the best-performing model (Fig. 3a and Supplementary Tables 6 and 7).

The previously obtained 5D feature matrix was also simplified to a matrix containing only two event parameters (2D), that is, Δ*I* and SD. The results of the 2D feature matrix were used as input for training and validation. However, the reported best validation accuracy dropped to 96.0%, indicating that the 5D feature matrix, which contains more information, clearly outperforms its 2D counterpart. Viewed in a different way, a machine learning program that simultaneously considers five event features is more accurate than the 2D scatter plot of Δ*I* versus SD (Fig. 2b).

The confusion matrix produced by the quadratic SVM model is shown in Fig. 3b, in which all amino acid events have a minimum true-positive rate of 95% (aspartic acid, isoleucine, leucine, asparagine, arginine and threonine even had a true-positive rate of 100%). Although a clear overlap between histidine and proline events was observed in the 2D scatter plot of Δ*I* versus SD (Fig. 2b), the validation accuracy of these two amino acids reached 95.4% and 96.7% respectively, by simultaneously considering the five event features and using machine learning (Fig. 3b). The parallel coordinate plot generated by the 5D feature matrix is also shown in Fig. 3c. To estimate the efficiency of model training, a learning curve was produced (Fig. 3d), which showed that a minimum of 1,500 input events is sufficient to achieve an accuracy of 98%. Furthermore, the trained quadratic SVM model was used to identify unlabeled amino acid events acquired with a mixture of all 20 proteinogenic amino acids (Supplementary Video 3). Representative traces are shown in Fig. 3e–g and all events were predicted and labeled by machine learning (Fig. 3e–g and Supplementary Figs. 21–23). As a measure of its performance, the corresponding event scatter plot of Δ*I* versus SD before and after event identification by machine learning is shown in Supplementary Fig. 24.

Identification of amino acids containing PTMs

PTMs, which are the chemical modification of proteins after translation, are critical in the modulation of a wide variety of protein functions. It is estimated that 50–90% of proteins in the human body are

https://doi.org/10.1038/s41592-023-02021-8

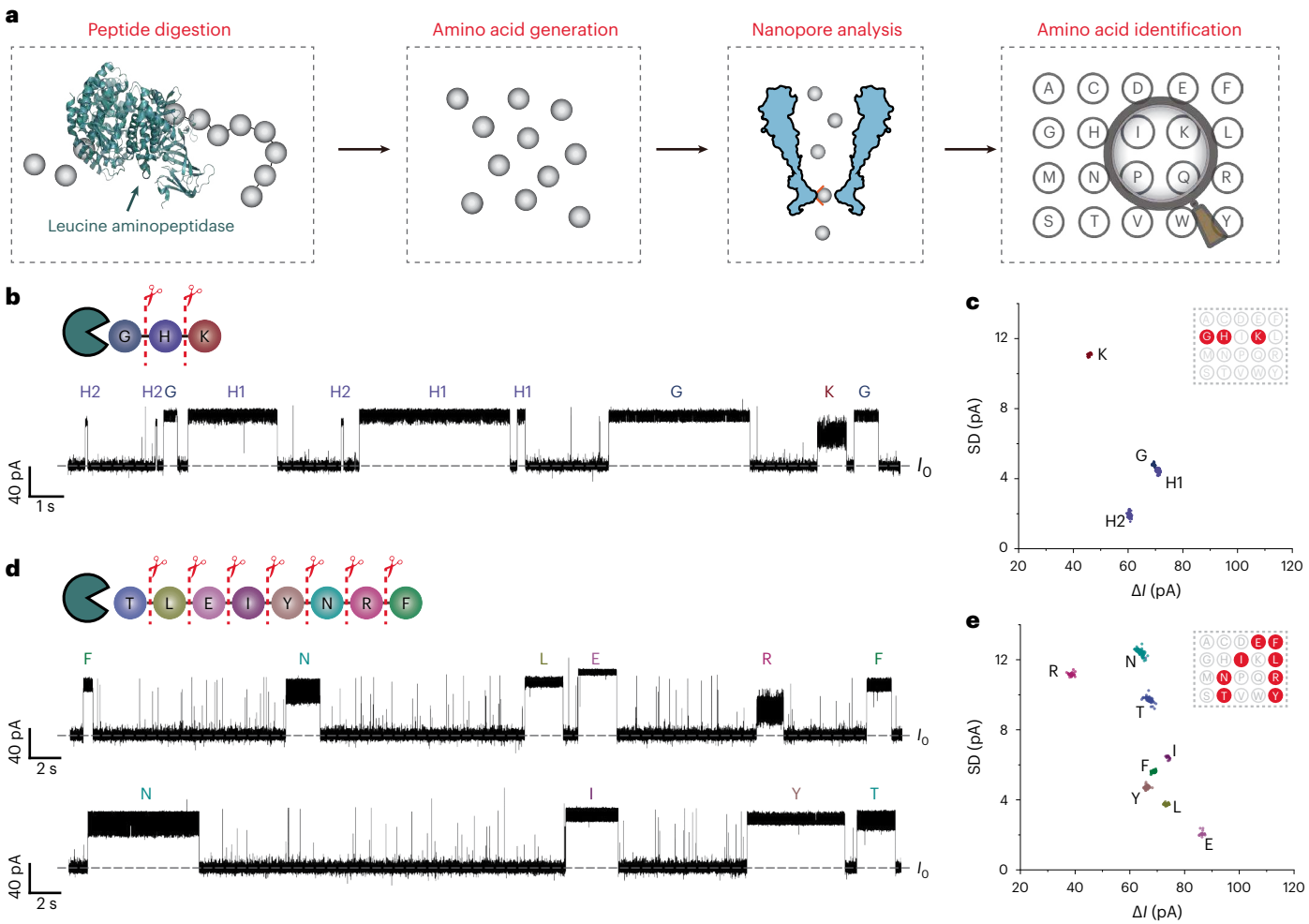


Fig. 6 | Identification of proteolytically cleaved amino acids. a, Schematic diagram of the identification of proteolytically cleaved amino acids from peptide using MspA-NTA-Ni. Leucine aminopeptidase (LAP) was used to digest the peptide and generate the amino acids. The amino acids were then identified by MspA-NTA-Ni, enabling confirmation of the amino acid components of the peptide. **b**, A representative trace of amino acids after LAP treatment of GHK peptide. **c**, The event scatter plot of Δ*I* versus SD for the events acquired as described in **b**. Data from a 90 min continuous trace are used. Four populations of events were identified by machine learning. Here, H1 and H2 represent two separate populations of events. **d**, Representative traces of amino acids after LAP treatment of octapeptide (TLEIYNRF). **e**, The event scatter plot of Δ*I* versus SD for events acquired as described in **d**. Data from a 90 min continuous trace are used. Eight populations corresponding to R, N, T, I, F, Y, L and E, respectively, were identified by machine learning. All nanopore measurements (**b**–**e**) were performed as described in Methods. A total of 40 µl filtrate of the LAP digestion product was added to *cis* prior to measurement. The event identities (**b**–**e**) were predicted by the previously trained machine learning algorithm. The open pore current (I_0) is marked with a dashed line in **b** and **d**.

post-translationally modified[35]. Accurate identification of PTMs is crucial for the understanding of cellular function as well as the related physiological and pathological processes. Although nanopore sensing of PTMs on peptides or proteins has been previously reported[36,37], a nanopore that can directly recognize individual amino acids containing PTMs has never been described, to the best of our knowledge. A recent report using the NIPSS strategy has demonstrated nanopore discrimination of peptides with individual phosphothreonine substitutions[37]. However, any substitution of phosphothreonine in the peptide will interfere with the nanopore reading of the neighboring amino acids, demonstrating an insufficient spatial resolution of that approach. It also fails to demonstrate nanopore discrimination of other PTMs[37].

Four common amino acids containing PTMs[35], that is, N^{ω},N^{ω}-dimethyl-arginine (Me-R), *O*-acetyl-threonine (Ac-T), N^4-(β-*N*-acetyl-D-glucosaminyl)-asparagine (GlcNAc-N) and *O*-phospho-serine (P-S) were used as model analytes (Fig. 4a). They demonstrate, respectively methylation, acetylation, glycosylation and phosphorylation, which are widely observed in natural proteins. When measured using MspA-NTA-Ni (Fig. 4a and Supplementary Fig. 25), representative events generated by these modified amino acids had unique event features, clearly distinguishable from events of proteinogenic amino acids (Fig. 2a). Core event parameters, as derived from three independent measurements for each condition, are also summarized in Supplementary Table 8 to show their consistency. The event scatter plot of Δ*I* versus the noise amplitude (that is, SD) of events of all modified amino acids (Fig. 4b), in which four fully separated populations of events are shown, confirms that MspA-NTA-Ni is also suitable for amino acids containing PTMs and that a high resolution of sensing is achieved.

These four modified amino acids and their unmodified precursors were also simultaneously sensed by the same nanopore (Fig. 4c–f and Supplementary Video 4). The results from each comparison pair (Supplementary Fig. 26) show two fully separated populations of events in the corresponding scatter plots (Fig. 4g–j). Although demonstrated with only four modified amino acids, this sensing strategy is in principle suitable for other types of modification such as hydroxylation, nitration or sulfation.

Furthermore, the results of the nanopore sensing of the four modified amino acids were complemented by the existing machine learning

algorithm (Extended Data Fig. 2a). The machine learning model is identical to that described in Fig. 3, but a total of 24 rather than 20 classes of amino acid data were used as input. Five event features were used for machine learning, and 10-fold cross-validation was used to evaluate the model performance (Supplementary Table 9). The quadratic SVM model had the highest validation accuracy of 98.6%, which is only 0.2% lower than that for 20 amino acids (98.8%). The corresponding confusion matrix is also shown in Extended Data Fig. 2b, in which the accuracy of all 24 amino acids is above 95.0% and the accuracy of 15 amino acids exceeds 98.0%. The event scatter plot of Δ*I* versus SD of all 24 amino acids is summarized in Extended Data Fig. 2c. Acknowledging the high resolution of MspA, the inclusion of extra data acquired with amino acids containing PTMs does not diminish the performance of the machine learning program, suggesting that the current strategy could handle even more types of amino acids in the future.

Rapid analysis of compound amino acid tablets

The high resolution of MspA-NTA-Ni and the high performance of the accompanying machine learning algorithm suggest that this sensing strategy could be used to analyze amino acid components in real biological samples. Amino acids are important for nutrition and are critical for the health and daily activities of humans. A variety of healthcare products designed to supplement nutrition, enhance immunity and renew physiological functions contain amino acids[38]. For a demonstration, a commercially available 'compound amino acid tablet' containing eight essential amino acids[39] (leucine, isoleucine, lysine, methionine, phenylalanine, threonine, tryptophan and valine) and one semi-essential amino acid (arginine) was analyzed using MspA-NTA-Ni (Fig. 5a and Supplementary Fig. 27).

The tablets were first pulverized and then dissolved in a KCl buffer (1.5 M KCl, 10 mM CHES, pH 9.0) at a concentration of 50 mg ml^{-1}. With a single MspA-NTA-Ni, the amino acid tablet solution was added to *cis* at a final concentration of 1 mg ml^{-1}. The corresponding types of amino acid events were immediately observed during single-channel recording. The identities of all events were automatically recognized by the previously trained quadratic SVM model (Fig. 5b and Supplementary Fig. 28). Events from a 90 minute continually recorded trace were used to generate the event scatter plot (Fig. 5c), in which nine clearly delineated populations of amino acid events are seen. They corresponded, respectively, to arginine, lysine, threonine, tryptophan, valine, phenylalanine, isoleucine, leucine and methionine, consistent with that described in the tablet's product manual. This confirmed that our sensing strategy is robust, consistent and can be directly applied in the quality control of nutrition products. Although a tablet normally contains other components such as starch and inorganic salts, the NTA-Ni^{2+} adapter provides sufficient selectivity to avoid interference from other components in natural samples. This suggests the feasibility of direct identification of amino acids in blood serum, urine or milk serum samples without complicated treatment, which would be useful in clinical diagnosis or nutrition analysis.

Identification of proteolytically cleaved amino acids

To evaluate whether the demonstrated sensing strategy may be used in the analysis of amino acid composition of peptides or proteins, the same principle was further applied in the identification of proteolytically cleaved amino acids (Fig. 6a). Leucine aminopeptidase (LAP) is an exopeptidase that catalyzes amino acid cleavage from the N terminus of the polypeptide chain[40] and has a broad substrate compatibility. Thus, LAP was used to cleave different target peptides into free amino acids prior to nanopore measurements.

GHK (glycyl-L-histidyl-L-lysine) is a naturally occurring tripeptide that is widely found in human serum[41]. It has a high copper affinity and has anti-inflammatory and tissue remodeling features. Experimentally, GHK was first incubated with LAP at 37 °C for 12 hours to achieve complete peptide cleavage (Methods). The product was then ultrafiltered to remove the enzyme, after which 40 µl filtrate was added to MspA-NTA-Ni, and the measurement was similarly carried out (Methods). Immediately afterwards, nanopore events corresponding to amino acids were consecutively reported (Fig. 6b). Events acquired from a 90 minute continually recorded trace were used to generate the scatter plot of Δ*I* versus SD (Fig. 6c and Supplementary Fig. 29). To remove non-clustered background noise, a DBSCAN (density-based spatial clustering of applications with noise) analysis was performed (Supplementary Fig. 29). Here, the non-clustered events may result from interfering molecules introduced from the enzymatic digestion buffer. Afterwards, four clusters of events were observed. According to the previously trained machine learning algorithm, they were identified, respectively, as K, G, H1 and H2 (Fig. 6c), fully consistent with the amino acid composition of the GHK peptide. Here, H1 and H2 are the two types of events generated by histidine, as noted above (Fig. 2a and Supplementary Video 2).

To demonstrate the generalizability of this assay to other peptides, a custom-synthesized octapeptide with a sequence of Thr-Leu-Glu-Ile-Tyr-Asn-Arg-Phe (TLEIYNRF) was identically treated and measured with MspA-NTA-Ni. A representative trace of the nanopore sensing of the TLEIYNRF digestion product is shown in Fig. 6d, in which the events corresponding to the expected amino acid identities are seen. After DBSCAN treatment followed by machine learning prediction (Fig. 6e and Supplementary Fig. 29), eight clearly delineated event populations were identified and they correspond, respectively, to R, N, T, I, F, Y, L and E, consistent with the sequence of the source peptide. To this end, this successfully demonstrates the capacity of MspA-NTA-Ni to identify proteolytically cleaved amino acids.

Discussion

A Ni^{2+}-modified MspA hetero-octamer (MspA-NTA-Ni) has been designed and used for amino acid sensing (Supplementary Video 5). It demonstrates clear discrimination of all 20 proteinogenic amino acids and 4 representative amino acids containing PTMs. This sensing configuration has remarkable stability and robustness, and can perform consistent and continuous measurement for several hours (Supplementary Fig. 30). The conical lumen geometry of MspA and that of the NTA-Ni^{2+} complex play a critical role in the generation of highly characteristic and reproducible amino acid events and, when this unique nanopore configuration of MspA-NTA-Ni is combined with a custom machine learning algorithm, it has a general accuracy of 98.6%. This capacity of amino acid sensing is also applied in the analysis of compound amino acid tablets, suggesting its potential use in clinical diagnosis and nutrition analysis. Furthermore, this principle has been extended to the identification of proteolytically cleaved amino acids, to demonstrate a nanopore-based strategy in the analysis of the amino acid composition of peptides or proteins. In the future, MspA-NTA-Ni may be conjugated with a protease, which would enable amino acids produced by hydrolysis of a target protein to be sequentially identified by the nanopore to achieve single-molecule protein sequencing[9].

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/s41592-023-02021-8.

References

1. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
2. Edman, P. Method for determination of the amino acid sequence in peptides. *Acta Chem. Scand.* **4**, 283–293 (1950).

3. Nivala, J., Marks, D. B. & Akeson, M. Unfoldase-mediated protein translocation through an alpha-hemolysin nanopore. *Nat. Biotechnol.* **31**, 247–250 (2013).
4. Yan, S. et al. Direct sequencing of 2′-deoxy-2′-fluoroarabinonucleic acid (FANA) using nanopore-induced phase-shift sequencing (NIPSS). *Chem. Sci.* **10**, 3110–3117 (2019).
5. Zhang, J. et al. Direct microRNA sequencing using nanopore-induced phase-shift sequencing. *iScience* **23**, 100916 (2020).
6. Yan, S. et al. Single molecule ratcheting motion of peptides in a *Mycobacterium smegmatis* porin A (MspA) nanopore. *Nano Lett.* **21**, 6703–6710 (2021).
7. Brinkerhoff, H., Kang, A. S., Liu, J., Aksimentiev, A. & Dekker, C. Multiple rereads of single proteins at single-amino acid resolution using nanopores. *Science* **374**, 1509–1513 (2021).
8. 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).
9. Zhang, S. et al. Bottom-up fabrication of a proteasome–nanopore that unravels and processes single proteins. *Nat. Chem.* **13**, 1192–1199 (2021).
10. Boersma, A. J. & Bayley, H. Continuous stochastic detection of amino acid enantiomers with a protein nanopore. *Angew. Chem. Int. Ed. Engl.* **51**, 9606–9609 (2012).
11. Ouldali, H. et al. Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore. *Nat. Biotechnol.* **38**, 176–181 (2020).
12. Hu, Z. L., Huo, M. Z., Ying, Y. L. & Long, Y. T. Biological nanopore approach for single-molecule protein sequencing. *Angew. Chem. Int Ed. Engl.* **60**, 14738–14749 (2021).
13. Zhao, Y. et al. Single-molecule spectroscopy of amino acids and peptides by recognition tunnelling. *Nat. Nanotechnol.* **9**, 466–473 (2014).
14. Faller, M., Niederweis, M. & Schulz, G. E. The structure of a mycobacterial outer-membrane channel. *Science* **303**, 1189–1192 (2004).
15. Cao, J. et al. Giant single molecule chemistry events observed from a tetrachloroaurate(III) embedded *Mycobacterium smegmatis* porin A nanopore. *Nat. Commun.* **10**, 5668 (2019).
16. Jia, W. et al. Programmable nano-reactors for stochastic sensing. *Nat. Commun.* **12**, 5811 (2021).
17. Jia, W. et al. A nanopore based molnupiravir sensor. *ACS Sens.* **7**, 1564–1571 (2022).
18. Jia, W. et al. Identification of single-molecule catecholamine enantiomers using a programmable nanopore. *ACS Nano* **16**, 6615–6624 (2022).
19. Zhang, S. et al. A nanopore-based saccharide sensor. *Angew. Chem. Int. Ed. Engl.* **61**, e202203769 (2022).
20. Wang, Y. et al. Identification of nucleoside monophosphates and their epigenetic modifications using an engineered nanopore. *Nat. Nanotechnol.* **17**, 976–983 (2022).
21. Liu, Y. et al. Nanopore identification of alditol epimers and their application in rapid analysis of alditol-containing drinks and healthcare products. *J. Am. Chem. Soc.* **144**, 13717–13728 (2022).
22. Hochuli, E., Döbeli, H. & Schacher, A. New metal chelate adsorbent selective for proteins and peptides containing neighbouring histidine residues. *J. Chromatogr.* **411**, 177–184 (1987).
23. Ali, M. et al. Label-free histamine detection with nanofluidic diodes through metal ion displacement mechanism. *Colloids Surf. B Biointerfaces* **150**, 201–208 (2017).
24. Wei, R., Gatterdam, V., Wieneke, R., Tampe, R. & Rant, U. Stochastic sensing of proteins with receptor-modified solid-state nanopores. *Nat. Nanotechnol.* **7**, 257–263 (2012).
25. Choi, L. S. & Bayley, H. S-nitrosothiol chemistry at the single-molecule level. *Angew. Chem. Int. Ed. Engl.* **51**, 7972–7976 (2012).
26. Shimazaki, Y., Takani, M. & Yamauchi, O. Metal complexes of amino acids and amino acid side chain groups. Structures and properties. *Dalton Trans.* **14**, 7854–7869 (2009).
27. Martell, A. E. & Smith, R. M. in *Critical Stability Constants* (eds Martell, A. E. & Smith, R. M.) 1–58 (Springer US, 1982).
28. Anderegg, G. Critical survey of stability constants of NTA complexes. *Pure Appl. Chem.* **54**, 2693–2758 (1982).
29. Zhang, J. et al. Mapping potential engineering sites of *Mycobacterium smegmatis* porin A (MspA) to form a nanoreactor. *ACS Sens.* **6**, 2449–2456 (2021).
30. Song, L. et al. Structure of staphylococcal α-hemolysin, a heptameric transmembrane pore. *Science* **274**, 1859–1865 (1996).
31. Kiseleva, I. et al. Thermodynamic study of mixed-ligand complex formation of copper(II) and nickel(II) nitrilotriacetates with amino acids in solution. I. *Polyhedron* **51**, 10–17 (2013).
32. Wang, Y. et al. Nanopore sequencing accurately identifies the mutagenic DNA lesion O^6-carboxymethyl guanine and reveals its behavior in replication. *Angew. Chem. Int. Ed. Engl.* **58**, 8432–8436 (2019).
33. Butler, T. Z., Pavlenok, M., Derrington, I. M., Niederweis, M. & Gundlach, J. H. Single-molecule DNA detection with an engineered MspA protein nanopore. *Proc. Natl Acad. Sci. USA* **105**, 20647–20652 (2008).
34. Abraham, M. J. et al. GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19–25 (2015).
35. Doyle, H. A. & Mamula, M. J. Post-translational protein modifications in antigen recognition and autoimmunity. *Trends Immunol.* **22**, 443–449 (2001).
36. 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).
37. 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 (2023).
38. Meng, W. J., Li, Y. & Zhou, Z. G. Anaphylactic shock and lethal anaphylaxis caused by compound amino acid solution, a nutritional treatment widely used in China. *Amino Acids* **42**, 2501–2505 (2012).
39. Hoffer, L. J. Human protein and amino acid requirements. *JPEN J. Parenter. Enteral Nutr.* **40**, 460–474 (2016).
40. Grembecka, J., Mucha, A., Cierpicki, T. & Kafarski, P. The most potent organophosphorus inhibitors of leucine aminopeptidase. Structure-based design, chemistry, and activity. *J. Med. Chem.* **46**, 2641–2655 (2003).
41. Dou, Y., Lee, A., Zhu, L., Morton, J. & Ladiges, W. The potential of GHK as an anti-aging peptide. *Aging Pathobiol. Ther.* **2**, 58–61 (2020).

Methods

Nanopore preparation

$(N90C)_1(M2)_7$ is a hetero-octameric protein, composed of one unit of N90C MspA-H6 and seven units of M2 MspA-D16H6 (ref. 19). $(N90C)_1(M2)_7$ contains a sole cysteine residue that is designed for site-specific chemical modification. N90C MspA-H6 and M2 MspA-D16H6 both contain a hexahistidine (H6) tail, which was introduced to assist protein purification by nickel affinity chromatography. The 16-aspartic acid (D16) placed on the monomer M2 MspA-D16H6 was designed to enhance the discrimination between different heterogeneous $(N90C)_1(M2)_7$ assemblies during gel electrophoresis. Prior to the preparation of $(N90C)_1(M2)_7$, both genes, that is, M2 MspA-D16H6 and N90C MspA-H6, were synthesized and simultaneously cloned into a pETDuet-1 plasmid by Genscript. To prepare $(N90C)_1(M2)_7$, the reconstructed plasmid was expressed with *Escherichia coli* strain BL21(DE3) plysS-competent cells and the expression products were purified by nickel affinity chromatography. Further separation of different hetero-octameric MspA was performed using polyacrylamide gel electrophoresis, during which the protein band corresponding to the $(N90C)_1(M2)_7$ assembly was identified. The corresponding band was excised from the gel. The protein was recovered from the gel band for subsequent use without any further purification.

M2 MspA is a homo-octamer. The gene coding for M2 MspA was inserted in a pET-30a (+) vector by GenScript[42]. The plasmid DNA was expressed with *E. coli* strain BL21(DE3) plysS-competent cells and the expression product was purified by nickel affinity chromatography. The prepared M2 MspA, which contains no reactive sites, was used as the reference nanopore (Supplementary Fig. 7).

Nanopore modification

To modify $(N90C)_1(M2)_7$ with a nitrilotriacetic acid (NTA), the prepared $(N90C)_1(M2)_7$ and maleimido-C3-NTA (20 mM) were mixed and co-incubated for 1 h at room temperature at a volume ratio of 1:8. The resulting product, which is referred to as MspA-NTA, was immediately used or stored at −80 °C.

The chelation of Ni^{2+} by MspA-NTA, which produces a Ni^{2+}-modified MspA nanopore, referred to as MspA-NTA-Ni, is monitored using single-channel recording.

Electrolyte buffer preparations

The KCl buffers (1.5 M KCl, 10 mM MES, pH 6.0; 1.5 M KCl, 10 mM MOPS, pH 7.0; 1.5 M KCl, 10 mM HEPES, pH 8.0; 1.5 M KCl, 10 mM CHES, pH 9.0) were prepared with Milli-Q water. The buffer was then pretreated with Chelex 100 resin for 12 h to remove polyvalent metal ions. After this, the mixture was filtered through a membrane (0.2 µm) to remove the resin. Finally, the pH of the electrolyte buffers was adjusted to the desired value.

Nanopore measurements

Nanopore measurements were carried out in a homemade Faraday cage placed on an optical table (Jiangxi Liansheng Technology). The measurement device, which consists of two chambers, was custom-made. Conventionally, the electrically grounded chamber is defined as *cis* and the opposing chamber is defined as *trans*. The two chambers are separated by a Teflon film containing a drilled aperture (~100 µm) at the center. The aperture was pretreated with 0.5% (v/v) hexadecane in pentane prior to each use. Then, each chamber was filled with 0.5 ml KCl buffer. A pair of Ag/AgCl electrodes were immersed in both chambers and electrically connected to a patch-clamp amplifier to form a closed circuit. A drop of DPhPC (diphytanoylphosphatidylcholine, 5 mg ml^{-1} in pentane) was then added to each chamber to form a lipid bilayer on the aperture. Subsequently, nanopores were added to the *cis* chamber to initiate pore insertions. Upon a single nanopore insertion, the *cis* chamber was immediately replaced with fresh KCl buffer to prevent further pore insertions.

All electrophysiological measurements were performed with an Axonpatch 200B patch-clamp amplifier paired with a Digidata 1550B digitizer at room temperature. All single-channel recordings were sampled at 25 kHz and low-pass filtered with a corner frequency of 1 kHz. Unless otherwise stated, all measurements were performed with a buffer of 1.5 M KCl, 10 mM CHES, pH 9.0 and an external voltage of +100 mV at room temperature. All analytes were added to the *cis* chamber to the desired final concentration.

The chelation of Ni^{2+} by MspA-NTA was performed during electrophysiological measurements. Prior to nanopore insertion, Ni^{2+} was added to the *trans* chamber at a final concentration of 50 µM. With a single MspA-NTA inserted, the Ni^{2+} present in *trans* will bond with the NTA on the pore to form a ternary complex termed MspA-NTA-Ni.

Data analysis

All nanopore events were detected from raw single-channel recording traces using the single-channel search function in Clampfit 10.7 (Molecular Devices). Events with a dwell time <10 ms were ignored. From each event, five event features, that is, ΔI, SD, skewness, kurtosis and t_{off}, were extracted using a custom MATLAB program. All subsequent data processing was performed with Origin 2021.

Machine learning was performed using the Classification Learner toolbox of MATLAB. A total of 300 nanopore events from each amino acid class were collected to form a labeled dataset. The label of the dataset for each event is assigned as the amino acid type used for data generation. The event features (ΔI, SD, skewness, kurtosis and t_{off}) extracted from nanopore events acquired for each known amino acid were collected to form a feature matrix. This feature matrix was then randomly split into a training set (80%) and a testing set (20%). The training set and the testing set were used as input, respectively, by the Classification Learner for model training and testing. A series of inbuilt classifiers of MATLAB, that is, Ensemble, Decision Trees, Discriminant Analysis, Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Neural Network were evaluated. To avoid overfitting, 10-fold cross-validation was performed and the corresponding validation accuracy and test accuracy were determined. The 10-fold cross-validation was performed by randomly and equally splitting the training set into 10 subsets and using each subset in turn as the validation set, with the remaining nine subsets being used to train the classifier. The cross-validation process is repeated 10 times, and the average validation accuracy is used as the evaluation criterion for the classification model. Furthermore, the best-performing model was screened according to the results of 10-fold cross-validation, and the trained model was used to predict unlabeled data. A confusion matrix was generated based on the results of the model. A learning curve with varying sample sizes was used to estimate the efficiency of model training. DBSCAN analysis was performed using Python. The epsilon was set to 0.1 and the minimum number of points was set to 10. The code for the machine learning model and the corresponding training data are provided on figshare: https://figshare.com/articles/software/Amino_acid-classifier/23995890

Calculation of the net charge (Z_{net}) of amino acids

The net charge of the amino acid could be derived as the sum of the charges of all its ionizable groups at a given pH. The charge of each ionizable group can be quantified using to the Henderson–Hasselbalch equation[43]:

$$pH = pK_a + \log([A^-]/[HA])$$

where K_a is the dissociation constant of weak acid and $pK_a = -\lg K_a$. [HA] and [A^-] represent the molarities of the weak acid and the conjugate base, respectively. Therefore, the ratio of the weak acid and the conjugate base of each ionizable group at the given pH (pH_{test}) can be determined from the pK_a values. For positive side chain groups and

the free amino groups of amino acids, the positive charge, Z_{pos}, is thus calculated by:

$$Z_{pos} = 1/\left[1 + 10^{(pH_{test} - pK_a)}\right]$$

For negatively charged side chain groups and the free carboxyl groups of amino acids, the negative charge, Z_{neg}, is thus calculated by:

$$Z_{neg} = -1/\left[1 + 10^{(pK_a - pH_{test})}\right]$$

The net charge of an amino acid can then be derived from:

$$Z_{net} = \sum Z_{pos} + \sum Z_{neg}$$

Peptide digestion

Leucine aminopeptidase (LAP) was used to catalyze the release of free amino acids from the N terminus of the peptide. A tripeptide (GHK) and an octapeptide (TLEIYNRF) were used separately as model peptides for the demonstration. The two peptides were separately dissolved in a 50 mM sodium phosphate, pH 8.0 buffer with a concentration of 10 mg ml^{-1}. LAP was prepared in a 50 mM sodium phosphate, pH 8.0 buffer with a concentration of 50 mg ml^{-1} (>7 U mg^{-1}). To initiate the hydrolysis reaction, 10 µl LAP solution, 20 µl 10 mM $MgCl_2$ and 70 µl peptide solution were mixed and incubated at 37 °C for 12 h in a dry block incubator. The reaction was stopped by heating the mixture to 80 °C for 5 min to inactive the LAP. Afterwards, another 100 µl ultrapure water was added to the mixture and loaded into an ultracentrifuge tube with a 10 kDa molecular weight cut-off. The filtration was then performed at 8,000 r.p.m. for 60 min at 4 °C. The filtrate was collected and stored at 4 °C for subsequent use.

Molecular dynamics simulations

The molecular dynamics simulations were conducted using GROMACS 2021.2[34] with the AMBER ff19SB and lipid21 force field[44,45]. Following the experimental set-up, the mutations D90N/D91N/D93N/D118R/D134R/E139K were introduced into the MspA. In addition, Asn90 of monomer A was replaced by a Cys, with its side chain S atom being covalently connected to an NTA via a linker maleimide (maleimido-C3-NTA). In addition to the above system (referred to as MspA-NTA hereafter), another two systems were also prepared: a system with a Ni^{2+} bonded to the MspA-NTA (MspA-NTA-Ni) and a system with a glycine attached to the Ni^{2+}(MspA-NTA-Ni-Gly). The force field parameters of the Cys-NTA and Gly were extracted using the packages Sobtop[46] and Multiwfn[47] following the protocol given in the literature[46]. The simulation systems were prepared using the CHARMM-GUI web server[48]. For the Ni^{2+}, the force field parameters developed by Li and Merz[49] were used. The crystal structure of the MspA (Protein Data Bank code 1UUN)[14] was used to set up the atomic coordinates of the MspA for the initial structures in the simulations. A POPC lipid bilayer with a size of 12 × 12 nm^2 was added surrounding the MspA. The system was solvated in a rectangular water box with a periodic boundary condition. K^+ and Cl^- ions corresponding to a salt concentration of 1.5 M, the same concentration as that used in the experiments, were added at random positions in the box. The smooth particle-mesh Ewald method was used for the calculations of the long-range electrostatic interactions. A cut-off distance of 1.2 nm was applied to the van der Waals interactions and the short-range part of the electrostatic interactions. For each of the three systems, at least five independent simulations with different initial conditions were carried out. In the simulations, the systems were first minimized for 1,000 steps. Then the systems were heated to 300 K and relaxed for 0.25 ns under the NVT (constant temperature, constant volume) ensemble, which was followed by another round of relaxation simulations under the NPT (constant temperature, constant pressure) ensemble at 300 K and 1 atm for 1.6 ns. The product simulations were conducted under the NPT ensemble at 300 K and 1 atm for at least 100 ns with a time step of 2 fs. During all of the above heating, relaxation and production molecular dynamics simulations, an external electric field of 0.15 V per 10 nm along the direction perpendicular to the membrane plane was applied, which gives a transmembrane voltage close to that used in the experiments. Meanwhile, a harmonic positional restraint was applied to the Cα atoms of the MspA with a spring constant of 500 kJ mol^{-1} nm^{-2} and to all of the heavy atoms of the lipid molecules with a spring constant of 1,000 kJ mol^{-1} nm^{-2}. A harmonic potential was applied to restrain the Ni^{2+} to within chelation distance of the NTA O atoms. For the system MspA-NTA-Ni-Gly, the backbone N and O atoms of Gly were restrained to the first coordination shell of Ni^{2+} by applying a harmonic potential during all of the simulation stages. A harmonic potential was applied also between the Ni^{2+} and NTA O atoms.

To analyze the structural features of the NTA in the narrow constriction region of the above three systems, the contact probabilities between NTA and the side chains of Asn90 and Asn91 in each of the monomers were calculated. The side chains of Asn90 and Asn91 are located at the inner side of the narrow constriction of the porin. Therefore, the formation of contacts between NTA and these residues tends to have a larger effect on the porin blockade. A contact between NTA and the monomer X (X represents the monomer index) was formed if the closest distance between the heavy atoms in the side chains of the residues Asn90/Asn91 of the protomer X and the N and O atoms of the NTA is less than 3.5 Å, or if the closest distance between the heavy atoms in the side chains of the residues Asn90/Asn91 of the monomer X and the Ni^{2+} bonded to NTA is less than 5.0 Å. Owing to the limitation of the accessible simulation time length, the molecular dynamics simulations here cannot capture the full coordination event of the Asn90/Asn91 side chains to the first coordination shell of the Ni^{2+} in the system MspA-NTA-Ni. The observed contacts were mainly contributed by the water-mediated coordination. Here, a relatively larger distance cut-off for the Ni^{2+} coordination distance was used to consider the water-mediated coordination. In the calculation of the contact probabilities, the snapshots from the first 50 ns of each trajectory were omitted.

Reporting summary

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

Data availability

Data supporting the findings of this study are given in the main text and the Supplementary Information. All source data are provided with this paper. All data used to train, evaluate and test the machine learning model are available on figshare. Please follow the link: https://figshare.com/articles/software/Amino_acid-classifier/23995890 for download. Source data are provided with this paper.

Code availability

The custom machine learning code is available on figshare as 'Amino acid-classifier'. Please follow the link: https://figshare.com/articles/software/Amino_acid-classifier/23995890 for download.

References

42. Wang, Y. et al. Osmosis-driven motion-type modulation of biological nanopores for parallel optical nucleic acid sensing. *ACS Appl. Mater. Interfaces* **10**, 7788–7797 (2018).
43. Moore, D. S. Amino acid and peptide net charges: a simple calculational procedure. *Biochemical Educ.* **13**, 10–11 (1985).
44. Tian, C. et al. ff19SB: amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. *J. Chem. Theory Comput.* **16**, 528–552 (2020).
45. Dickson, C. J., Walker, R. C. & Gould, I. R. Lipid21: complex lipid membrane simulations with AMBER. *J. Chem. Theory Comput.* **18**, 1726–1736 (2022).

46. Lu, T. Sobtop, version 1.0 (dev3.1), http://sobereva.com/soft/Sobtop (accessed 15 August 2022).
47. Lu, T. & Chen, F. Multiwfn: a multifunctional wavefunction analyzer. *J. Comput. Chem.* **33**, 580–592 (2012).
48. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.* **29**, 1859–1865 (2008).
49. Li, P. & Merz, K. M. Jr. Taking into account the ion-induced dipole interaction in the nonbonded model of ions. *J. Chem. Theory Comput.* **10**, 289–297 (2014).

Acknowledgements

This project was funded by the National Key R&D Program of China (grant no. 2022YFA1304602, to S.H.), National Natural Science Foundation of China (grant no. 22225405 and no. 31972917, to S.H.), the Fundamental Research Funds for the Central Universities (grant no. 020514380257 to S.H.), Programs for high-level entrepreneurial and innovative talents introduction of Jiangsu Province (individual and group program, to S.H.), Natural Science Foundation of Jiangsu Province (grant no. BK20200009, to S.H.), State Key Laboratory of Analytical Chemistry for Life Science (grant no. 5431ZZXM2204, to S.H.) and the China Postdoctoral Science Foundation (grant no. 2021M691508 and grant no. 2022T150308, to Y.W.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

S.H., K.W. and S.Z. conceived the project. S.Z. and K.W. performed the pore engineering. K.W., X.Y., X.L. and W.S. performed the measurements. X.Z. and W.L. conducted the molecular dynamics simulations. Y.W., P.F. and Y.X. designed the machine learning algorithms. K.W. and Y.W. prepared the supplementary videos. P.Z. set up the instruments. S.H. and K.W. wrote the paper. S.H. supervised the project.

Competing interests

S.H., S.Z., K.W. and Y.W. have filed patents describing the preparation of heterogeneous MspA and its applications thereof. All other authors have no competing interests.

Additional information

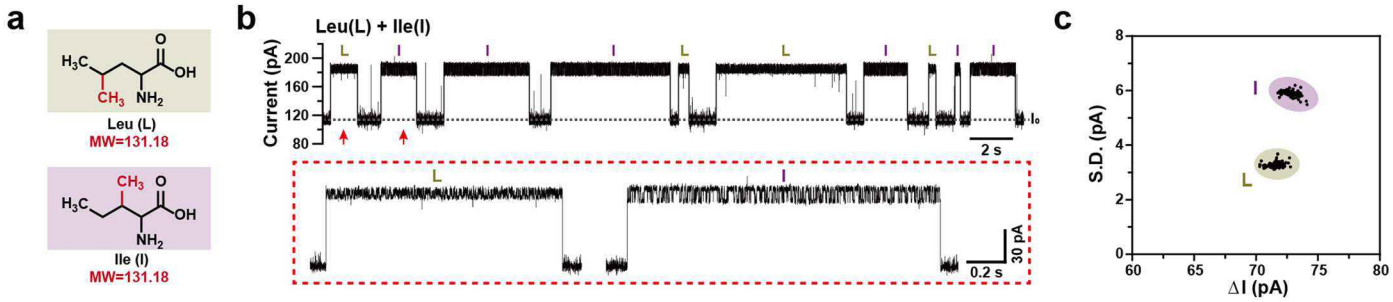
Extended data are available for this paper at https://doi.org/10.1038/s41592-023-02021-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-023-02021-8.

Correspondence and requests for materials should be addressed to Shuo Huang.

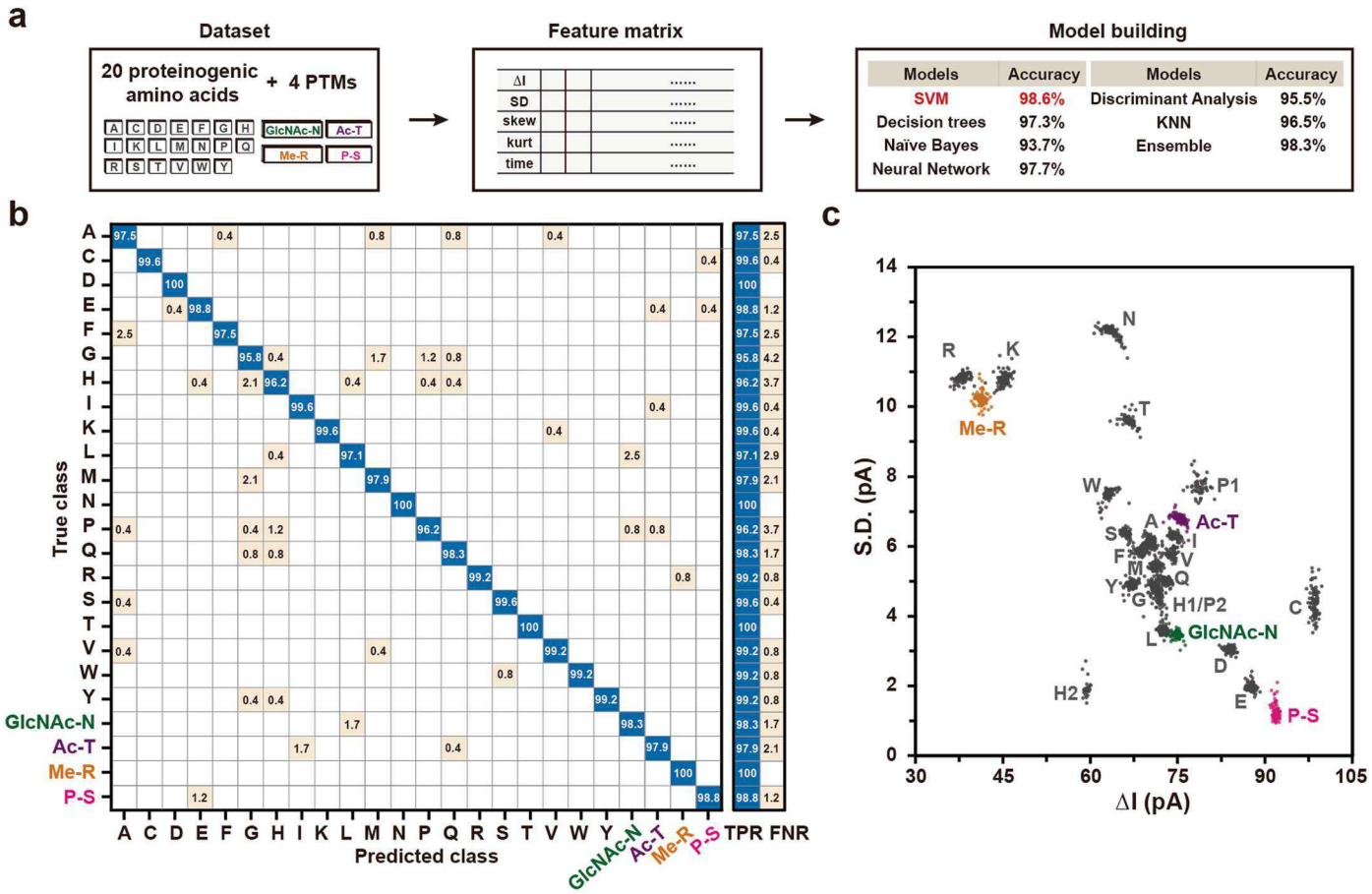
Peer review information *Nature Methods* thanks Jeff Nivala, Sukanya Punthambaker and Meni Wanunu for their contribution to the peer review of this work. Peer reviewer reports are available. Primary Handling Editor: Arunima Singh, in collaboration with the *Nature Methods* team. Peer reviewer reports are available.

https://doi.org/10.1038/s41592-023-02021-8



Extended Data Fig. 1 | Simultaneous sensing of leucine and isoleucine. The measurements were carried out as described in Methods. A 1.5 M KCl buffer (1.5 M KCl, 10 mM CHES, pH 9.0) was used. A transmembrane voltage of +100 mV was continually applied. Nickel sulfate was added to *trans* with a final concentration of 50 μM. (**a**) The chemical structures of leucine (Leu, L) and isoleucine (Ile, I). Leucine and isoleucine are isomers with identical mass. (**b**) Top: A representative trace acquired during simultaneous sensing of leucine and isoleucine. Each amino acid was added to *cis* with a final concentration of 1 mM. Bottom: Representative events of leucine and isoleucine. The events are taken from the continuous trace (top) marked with red arrows. I_0 represents the open pore current of MspA-NTA-Ni. Events caused by leucine and isoleucine are easily identifiable. (**c**) The event scatter plot of ΔI versus *S. D.* generated from results of (**b**). 274 successive events were used to generate the statistics. Though leucine and isoleucine have indistinguishable MW, they are fully discriminated by nanopore.

https://doi.org/10.1038/s41592-023-02021-8



Extended Data Fig. 2 | Machine-learning assisted identification of twenty-four amino acids. (**a**) The machine-learning workflow. Sensing events acquired with twenty proteinogenic amino acids and four modified amino acids were collected to form a database. Three-hundred events were randomly selected from each amino acid class to form a labeled dataset. Five event features including ΔI, S.D., skew, kurt and t_{off} were extracted from the events to form a feature matrix. After evaluation with ten-fold cross-validation, the quadratic SVM model was found to be the optimum model by demonstrating a validation accuracy of 98.6% (Supplementary Table 9). (**b**) The confusion matrix result of twenty-four amino acids classification performed with the trained quadratic SVM model. The row of the matrix represents the true class and the column represents the predicted class. (**c**) The scatter plot of ΔI versus *S. D.* generated by results of nanopore measurements of 20 proteinogenic amino acids (gray dots) as well as four amino acids containing PTMs (colorful dots). One hundred successive events of each amino acid were used to generate the statistics. The distribution of the four modified amino acids can be fully discriminated from that of the twenty proteinogenic amino acids.

nature portfolio

Corresponding author(s): Shuo Huang

Last updated by author(s): Jul 14, 2023

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	Single channel recordings were performed by Axon 200B and the accompanied commerically available software Axonpatch 10.7
Data analysis	Software packages used and the associated package versions Clamfit-10.7-Software used for event detection from the nanopore trace. Origin-2021-Software used for plot generation. Matlab-R2021a-Software used for event feature extraction and machine learning. Python-2.7-Program used for cluster analysis.

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

Data supporting the findings of this study are shown in the main text and the Supplementary Information. All source data are provided with this paper. All data used to train, evaluate and test the machine learning model is shared on google drive. Please follow the link: https://drive.google.com/file/d/1cNV4yCBuvJ_5_Nn6SA25b7Fek_EhuHiu/view for download.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N.A.
Population characteristics	N.A.
Recruitment	N.A.
Ethics oversight	N.A.

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

Life sciences study design

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

Sample size	Each nanopore measurement was run continuously to collect all nanopore events within the measurement period. A sample is thus defined as all nanopore events acquired at each condition. In the establishment of the machine learning model, each sample contains at least 300 nanopore events. For the measurement of concentration dependence, each sample contains all nanopore events collected within a 10 min continuous data acquisition.
Data exclusions	No data were excluded from the analysis
Replication	All conclusions were drawn based on results independently acquired from at least three nanopores (N=3). This is to exclude potential pore to pore variations.
Randomization	Randomization is not relevant to this study since no experimental groups were compared.
Blinding	Our experiment didn't involve subjective trials. Blinding is not necessary in this study.

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

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging