PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Simulation of single-protein nanopore sensing shows feasibility for whole-proteome identification

Shilo Ohayon[1☯], Arik Girsault[1☯], Maisa Nasser[1], Shai Shen-Orr[2], Amit Meller[1,3]*

1 Department of Biomedical Engineering, Technion–IIT, Haifa, Israel, **2** Rapport Faculty of Medicine, Technion–IIT, Haifa, Israel, **3** Department of Biomedical Engineering, Boston University, Boston, Massachusetts, United States of America

☯ These authors contributed equally to this work.
* ameller@technion.ac.il

Citation: Ohayon S, Girsault A, Nasser M, Shen-Orr S, Meller A (2019) Simulation of single-protein nanopore sensing shows feasibility for whole-proteome identification. PLoS Comput Biol 15(5): e1007067. https://doi.org/10.1371/journal.pcbi.1007067

Editor: Aleksei Aksimentiev, University of Illinois at Urbana-Champaign, UNITED STATES

Received: February 13, 2019

Accepted: May 2, 2019

Published: May 30, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: AM was funded by the Israel Science Foundation (www.isf.org.il), I-Core 1902/12 award. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Single-molecule techniques for protein sequencing are making headway towards single-cell proteomics and are projected to propel our understanding of cellular biology and disease. Yet, single cell proteomics presents a substantial unmet challenge due to the unavailability of protein amplification techniques, and the vast dynamic-range of protein expression in cells. Here, we describe and computationally investigate the feasibility of a novel approach for single-protein identification using tri-color fluorescence and plasmonic-nanopore devices. Comprehensive computer simulations of denatured protein translocation processes through the nanopores show that the tri-color fluorescence time-traces retain sufficient information to permit pattern-recognition algorithms to correctly identify the vast majority of proteins in the human proteome. Importantly, even when taking into account realistic experimental conditions, which restrict the spatial and temporal resolutions as well as the labeling efficiency, and add substantial noise, a deep-learning protein classifier achieves 97% whole-proteome accuracies. Applying our approach for protein datasets of clinical relevancy, such as the plasma proteome or cytokine panels, we obtain ~98% correct protein identification. This study suggests the feasibility of a method for accurate and high-throughput protein identification, which is highly versatile and applicable.

Author summary

Macromolecules identification methods are central for most biological and biomedical studies, and while the field of genomics advanced to single-molecule resolution, the proteomic field still relies on bulk and costly techniques. We describe a solution for single protein identification, based on the analysis of optical traces obtained from fluorescently-labeled proteins threaded through a nanopore and processed by a pattern recognition algorithm. To evaluate the feasibility of our method we constructed computer simulations of the system, producing and analyzing nearly 10^8 individual protein translocations from the human Swiss-Prot database. Our results suggest protein identification of >95% for the

whole human proteome, even under non-ideal conditions. These results constitute the basis for a novel whole proteome identification method, with single molecule resolution.

Introduction

Modern DNA sequencing techniques have revolutionized genomics [1], but extending these methods to routine proteome analysis, and specifically to single-cell proteomics, remains a global unmet challenge. This is attributed to the fundamental complexity of the proteome: protein expression level spans several orders of magnitude, from a single copy to tens of thousands of copies per cell; and the total number of proteins in each cell is staggering [2]. Given the lack of *in-vitro* protein amplification assays the ability to accurately quantify both abundant and rare proteins hinges on the development of single-protein identification methods that also feature extraordinary-high sensing throughput. To date, however, protein sequencing techniques, such as mass-spectrometry, have not reached single-molecule resolution, and rely on bulk averaging from hundreds of cells or more [3]. Affinity-based method can reach single protein sensitivity [4], but depend on limited repertoires of antibodies, thus severely hindering their applicability for proteome-wide analyses. Consequently, in the past few years single-molecule approaches for proteome analysis based on Edman degradation [5] or FRET [6] have been proposed. To date, however, profiling of the entire proteome of individual cells remains the ultimate challenge in proteomics [7].

Nanopores are single-molecule biosensors adapted for DNA sequencing, as well as other biosensing applications [8,9]. Recent nanopore studies extended nucleic-acid detection to proteins, demonstrating that ion current traces contain information about protein size, charge and structure [10–17]. However, to date, the challenge of deconvolving the electrical ion-current trace to determine the protein's amino-acid sequence from the time-dependent electrical signal has remained elusive. In an analogy to the field of transcriptomics, in many practical cases it is sufficient to identify and quantify each protein among the repertoire of known proteins, instead of re-sequencing it. Yao and co-workers showed theoretically that most proteins in the human proteome database can be uniquely identified by the order of appearance of just two amino-acids, lysine and cysteine (K and C, respectively) [18]. But taking into account experimental errors, for example due to false calling of an amino-acid, or an unlabeled amino-acid, sharply reduces the ID accuracy. Motivated by recent experiments suggesting the ability to translocate SDS-denatured proteins through either small nanopores (~0.5 nm) [19], or large nanopores [20] (~10 nm), and the possibility to differentiate among polypeptides based on optical sensing in nanopore [21], we here introduce a protein ID method that according to simulation remains robust against the expected experimental errors. We show that relatively low-resolution, tri-color, optical fingerprints produced during the passage of proteins through a nanopore, preserve sufficient information to allow a deep-learning classification algorithm to accurately identify the entire human proteome with >95% accuracy. Even in cases where the apparent spatial and temporal resolutions of the optical system appear to be prohibitively low, and the amino-acids labelling efficiency is incomplete, whole proteome ID efficiency remains high and robust. Particularly, the expected protein ID efficiency is of an extremely high clinical relevancy. We illustrate the broad applicability of the method by analyzing the human plasma proteome, as well as commercially-available cytokine identification panel based on antibodies, showing that our antibody-free method can readily surpass current techniques in a number of key parameters, while displaying a near perfect accuracy.

Results

Simulation of nanopore-based recognition of proteins

In our method, proteins extracted from any source (serum, tissue or cells), are denatured using urea and SDS (Fig 1A). Three amino-acids lysine (K), cysteine (C) and methionine (M)

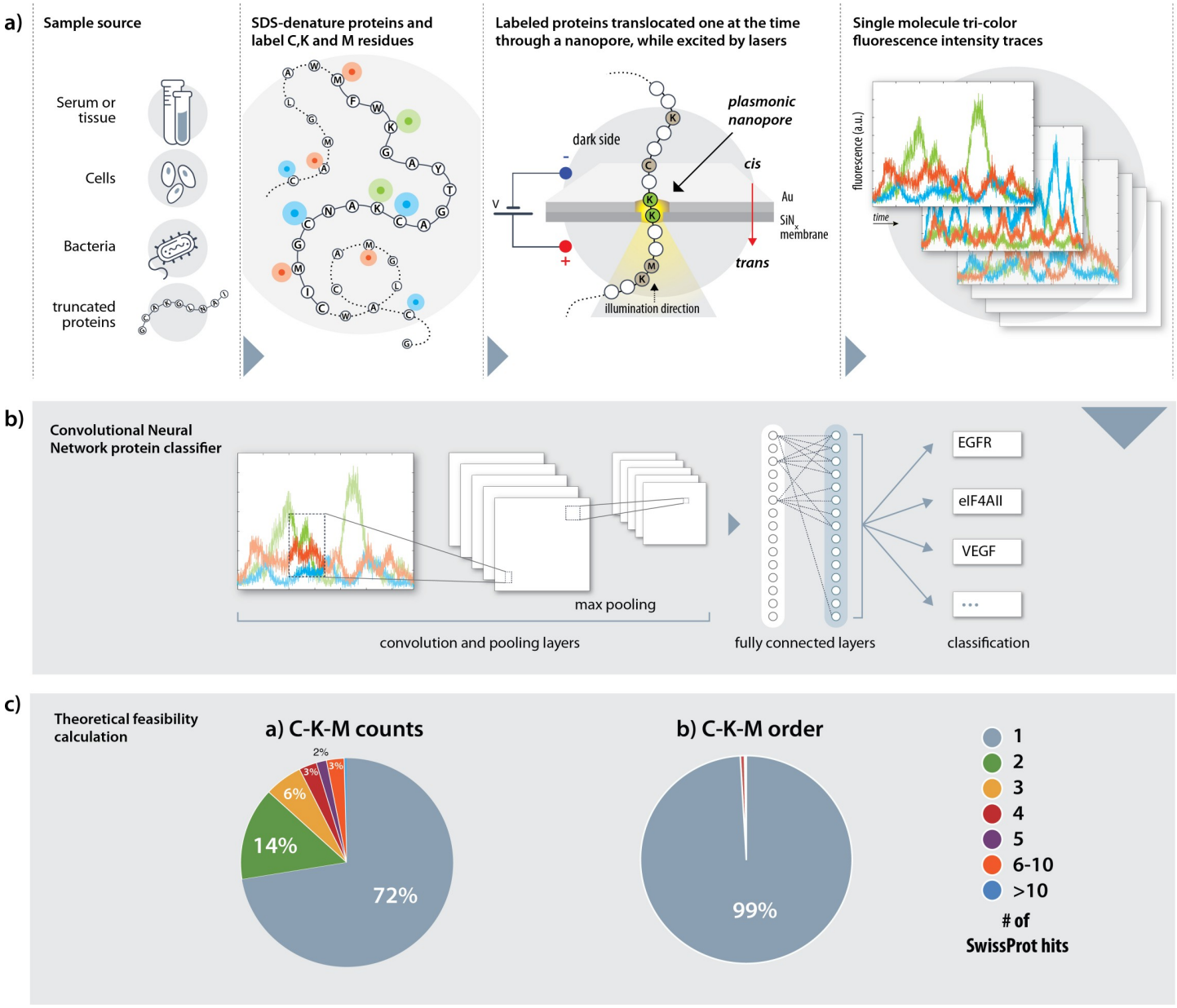


Fig 1. An overview of the Nanopore, tri-color protein identification method. (a) A tentative sample process flow. The protein sample is first denatured using SDS and cysteines (C), lysines (K) and methionines (M) are labeled with three spectrally-resolvable fluorophores. The labeled, SDS-denatured proteins are then threaded through a nanopore and excited by a laser light focused by a plasmonic architecture. The plasmonic field ensures local excitation of small portions of the denatured proteins. Finally, the photon emissions from each protein are measured in three channels, one for each fluorophore species, to create a tri-color optical trace per translocation. (b) A pre-trained convolutional neural network (CNN) classifier subsequently examines and classifies each trace, extracting its relevant features using a convolutional, an activation, a pooling and a fully connected layer, to identify the protein. (c) A theoretical evaluation of whole proteome fingerprinting based on complete labeling of C, K and M amino-acids. Counting only the number of labeled amino-acids yields unique identifications (ID) of 72% of all proteins. The remaining 28% of proteins were not uniquely identified and were either identified as one out of two (green slice) or more proteins as indicated by color. Considering also the order of the labeled amino-acids increases the unique ID fraction to 99%.

https://doi.org/10.1371/journal.pcbi.1007067.g001

are labeled with three different fluorophores using three orthogonal chemistries: the primary-amines in lysines are targeted with NHS esters; thiols in cysteines are targeted with maleimide groups, and methionines are labeled using the two-step redox-activated chemical tagging [22]. The negatively charged SDS-denatured polypeptides are electrophoretically threaded, one at a time, through a sub-5 nanometer pore fabricated in a thin insulating membrane to ensure single file threading of the SDS-coated polypeptide. The voltage, nanopore diameter and other factors, such as solution viscosity are used to regulate the protein translocations speed. The nanopore is illuminated using laser beams for multi-color excitation [23]. The excitation volume (Fig 1A, yellow highlighted region) is centered with the nanopore, and importantly, its axial depth is confined by plasmonic focusing of the incident electromagnetic field [24]. Consequently, depending on the excitation depth, either a single or multiple labeled amino-acids will be simultaneously illuminated, during the passage of the protein. Three-color fluorescence time traces ("fingerprints") are recorded for each protein passage and are classified using deep-learning (Fig 1B).

The theoretical likelihood of protein ID can be tested by calculating the percentages of unique matches of all proteins in the human Swiss-Prot database [25] based on the number and the order of appearance of three amino-acids only. Simply counting the number of K, C and M residues in each protein identifies 72% of the total proteins uniquely, and another 14% identified as either one of two proteins in which one of them is the correct match (online methods). Moreover, the percentage of uniquely identified proteins is close to 99% with the determination of the KCM order of appearance along all proteins in the human proteome database (Fig 1C). Thus, in principle, the boundaries for the expected ID accuracies fundamentally permit whole-proteome, single-protein, identification.

The theoretical analysis shown in Fig 1C may be considered as an upper limit for the accuracy of a protein ID method based on a three amino-acid labelling, which neglects inter-dye distances. However, it ignores experimental limitations, such as the sensing spatial and temporal constraints, the labelling efficiency and the photophysical properties of fluorophores. These factors are likely to impact the accuracy of the protein ID method, and hence must be considered. To this end we developed a detailed photophysical model to numerically calculate the time-dependent photon emission during the passage of each SDS-denatured protein through a solid-state nanopore. Our model consists of three layers: first, we used Finite Difference Time Domain (FDTD) computations to evaluate the expected electromagnetic field distribution for a simple plasmonic structure fabricated on top of the nanopore (Materials and Methods). Second, an amino-acid labelling simulation was applied to each protein, in order to generate partial labelling of each of the three target amino-acids. Finally, SDS-denatured proteins were allowed to slide through the plasmonic nanopore complex while illuminated at three distinct wavelengths. The expected detected photon emissions were calculated at each step of the protein translocation taking into account the photophysical properties of the fluorophores, as well as energy transfer (FRET), bleaching kinetics and collection efficiencies. This allowed us to generate detailed photon emission time traces for each and every protein translocation.

To illustrate our method, we schematically show in Fig 2A snap-shots of the system at two time points during the passage of the PSD protein. This figure is plotted in scale to illustrate the relative dimensions of the plasmonic field, the nanopore and the SDS-coated polypeptide chain (marked as orange layer around the chain). Specifically, the axial FWHM of the plasmonic field is 20 nm calculated from the FDTD field distribution, and the nanopore diameter is 3 nm. Each protein was modeled as a fully-denatured, SDS-coated, wormlike polymer [26], translocating across the nanopore at an instantaneous velocity $u_i = \langle u \rangle + \delta u_i$ where $\langle u \rangle$ is its average velocity, and the random term δu_i accounts for thermal fluctuations in its motion. Since the SDS-coated biopolymers have a Kuhn length of approximately 7 nm [26], they can

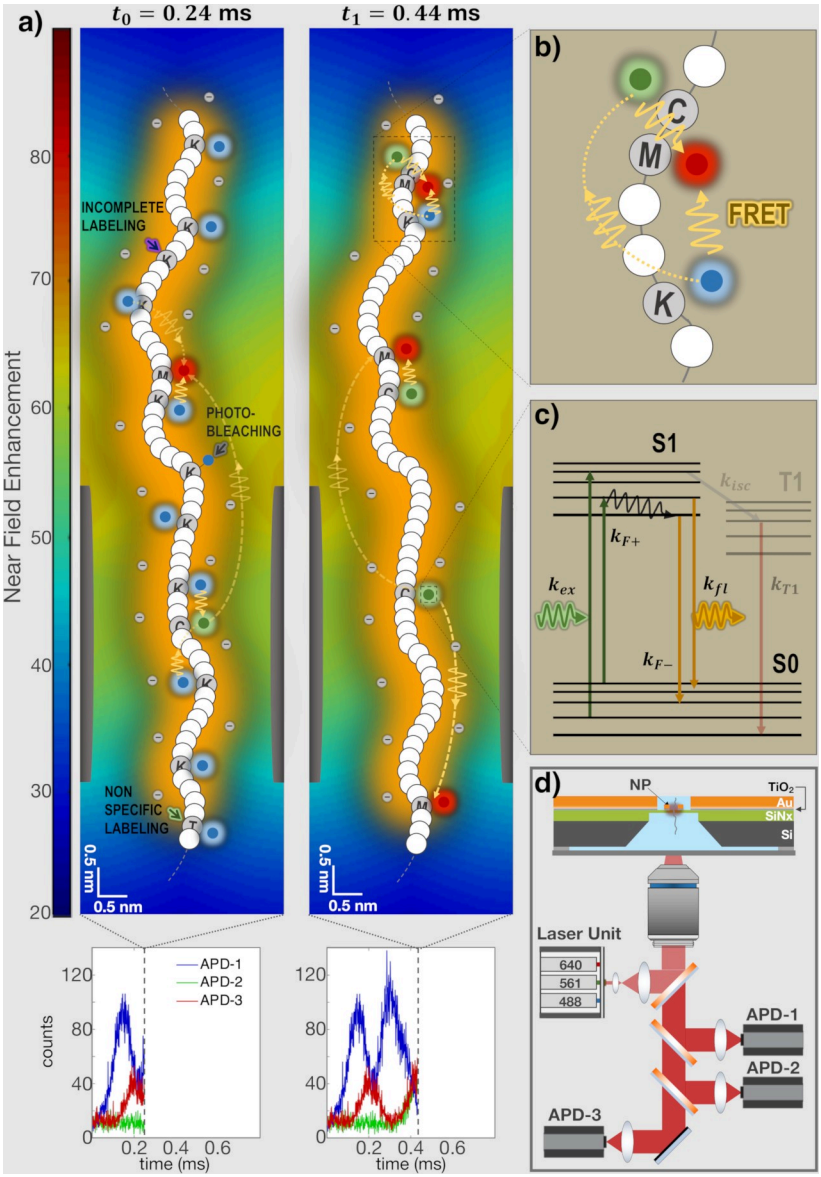


Fig 2. Simulation of the fluorescence signals generated during the translocation of the SDS-denatured PH and SEC7 domain-containing (PSD) protein. (a) The nanopore diameter and height were set to 3 and 5 nm, respectively, and the plasmonic architecture deposited on its 'top-side' produced a confined excitation profile (14–20 nm axial full-width half maximum) whose color map displayed on the left indicates the excitation near field enhancement at a wavelength of 640 nm (modeled using FDTD, see S1 Note). Two snapshots of the translocation process are shown and denoted by the timepoints t_0 and t_1 at which they were respectively taken. Energy transfer, photo-bleaching, incomplete labeling and non-specific labeling are indicated by dotted yellow lines, solid grey, purple and green arrows, respectively. (b) Zoomed in region of the polypeptide in which Forster resonance energy transfer (FRET) is shown in greater details. In this configuration, energy was transferred from lysine fluorophores to cysteine and methionine emitters, and from cysteine to methionine fluorophores. (c) The fluorescence emission rate of each labeled amino-acid was modeled as either a two-state or three-state system (see online methods for further details and in which k_{F+} and k_{F-} refer to $k_{FRET,+}$ and $k_{FRET,-}$, respectively). k_{exc} denotes the absorption rate, k_{isc} the inter-system crossing rate and k_{T1} the triplet state relaxation rate. Fluorophores are depicted in a color which denote the excitation wavelength with which they are excited or the channel to which they belong. (d) Schematics of the nanopore chip and optical system, which includes a high NA water immersion objective lens, three excitation laser lines and corresponding APDs. The nanopore chip is made of four consecutive layers: silicon (grey), silicon nitride (green) in which the nanopore is drilled, titanium oxide (grey blue) and gold (orange).

https://doi.org/10.1371/journal.pcbi.1007067.g002

be assumed to be partially-stretched (unfolded) wormlike polymers during translocation through a sub ~5 nm pore. Moreover, when threaded through a 3 nm pore, the roughly 2 nm wide SDS-coated proteins are confined laterally in a small volume in the nanopore proximity where the electromagnetic field remains nearly constant. Hence, in this study the protein translocations can be treated as one dimensional [27]. The excitation profile calculated from the FDTD simulations was approximated by a one-dimensional Gaussian function as shown in S1 Fig. The fluorescence emission rate of each labeled amino-acid while passing through the excitation zone was modeled as a two-state system (Fig 2C), as described in the Materials and Methods section. Triplet state transition rates, which may result in microsecond-long dark-states were also investigated (equations not shown) based on literature values of three specific fluorophores [28–30]. We explicitly took into consideration energy transfer rates (Fig 2B and 2C), which directly depend on the amino-acid sequence, as well as photo-bleaching rates (indicated by dotted yellow lines and solid grey arrows in Fig 2, respectively). At each time step of the simulation the emitted light from all fluorophores residing in the excitation zone were split to three spectrally-resolved, photon-counter channels as shown in Fig 2D. In addition to the collection and detection efficiency of each channel, we also considered photon statistics by incorporating shot-noise.

The labeling efficiency was modeled by randomly positioning fluorophores at the K, C and M amino-acid, such that in each protein only a fraction Γ_j of them (j represents K, C or M) was actually labelled (indicated by purple arrows in Fig 2A). In all the following computational results presented the three amino-acids, K, C and M were labelled by Atto488, Atto565 and Atto647N fluorophores, and the fluorophores properties were taken into account when simulating the photon emission rates. Additionally, we introduced cross-labelling efficiency (green arrows in Fig 2A), although this is known to be negligible [31].

In order to estimate the translocation velocity of SDS-denatured polypeptides we performed electrical translocation measurements using SDS-denatured albumin (585 amino-acids) proteins using ~4 nm-wide solid-state nanopores, as described in the Materials and Methods section. Representative translocation events measured at a bias voltage of V = 300 mV, in which a single blockage current level is observed, are shown in Fig 3A. Examining a statistical set of >900 translocation events showed a single blockade current level ($I_B = 0.7$) indicative of single-file polypeptide translocations. This experiment supports the assumption that proteins are likely to be fully denatured as they thread through the narrow nanopore, in agreement with a previous publication [20]. Fig 3B displays an overlay of the scatter plot of the fractional blockade current I_B versus the translocation dwell-time t_D, with its corresponding density map. The area delimited by the dashed red lines approximate the typical full-width-half-maximum of a Gaussian centered on the characteristic dwell-time (94.3±7.2 μs as determined by the histogram shown in the inlet panel). Accordingly, we estimate the mean translocation velocity by 0.2 cm/s. Notably, this velocity is slower than the previous report, presumably due to the fact that in our experiments a much smaller nanopore was used.

We first focus on the simulated optical signals calculated for two proteins having nearly the same length: the EGF precursor, and its receptor EGFR (1208 and 1210 amino-acids, respectively). Under near-ideal experimental conditions (100% labelling, 0.5 nm resolution, and velocity of 0.035 cm/s) their tri-color fingerprints were readily distinguishable from each other, despite similar K, C and M compositions, and followed the actual K,C,M amino-acid order in each protein (Fig 4A). We then extended our protein translocation simulations under much lower spatial resolutions, lower labelling efficiencies and higher translocation velocities. As expected, in the more realistic conditions we no longer can resolve individual fluorophore photon bursts, associated to single K, C or M residues. Instead, the resulting signals appear as continuous tri-color fingerprints of each protein translocation. Importantly, however, the

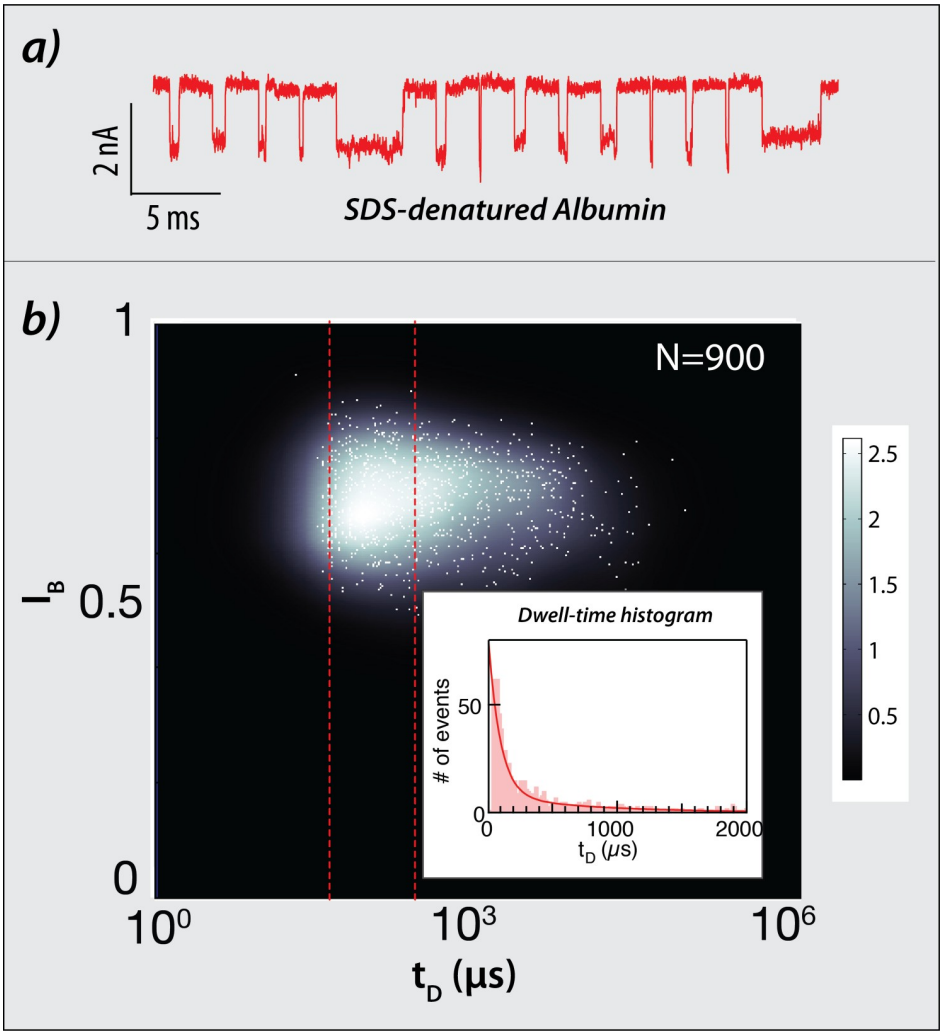


Fig 3. Measurements of SDS-denatured human serum albumin translocations through solid-state nanopores. (a) Electrical events of albumin translocating through a 4 nm-wide nanopore measured at 300 mV. (b) Scatter plot of the fractional blockade current I_B versus the translocation time t, with its corresponding density map. The number of translocations events displayed amounts to 900. The inset shows the dwell-time histogram, fitted to an exponential decay with characteristic time of 94.3±7.2 μs.

https://doi.org/10.1371/journal.pcbi.1007067.g003

fingerprints, even at the poorest resolution of 50 nm maintain an overall pattern characteristic of each protein (Fig 4B). Analyzing $>5{\cdot}10^7$ single protein translocations events, under different conditions suggest that even at 100 nm resolution some characteristic features of each protein are preserved (S2 Fig). Moreover, we expect that small variations in the nanopore size would result in different translocation velocities. To evaluate this effect, we repeated the translocation simulation experiments at mean values of 0.035, 0.2 and 2 cm/s and increasing the translocation velocity fluctuations (20%, 30% and 40% of the mean velocity). Our result presented in (S3, S4 & S8 Figs) suggest that as long as the velocity is in the order of ~ 0.2 cm/s (or below) in accordance with our experimental result (Fig 3), the identification accuracy remains sufficiently high.

We tested the similarity among repeated translocations of the same proteins, which were subject to different labeling and random velocity fluctuations, by evaluating the Pearson correlation coefficients between all pairs of 50 translocation repeats of the same protein. The results, showed in all cases high values (0.85–0.97) when considering auto-correlation (Fig 5, diagonal

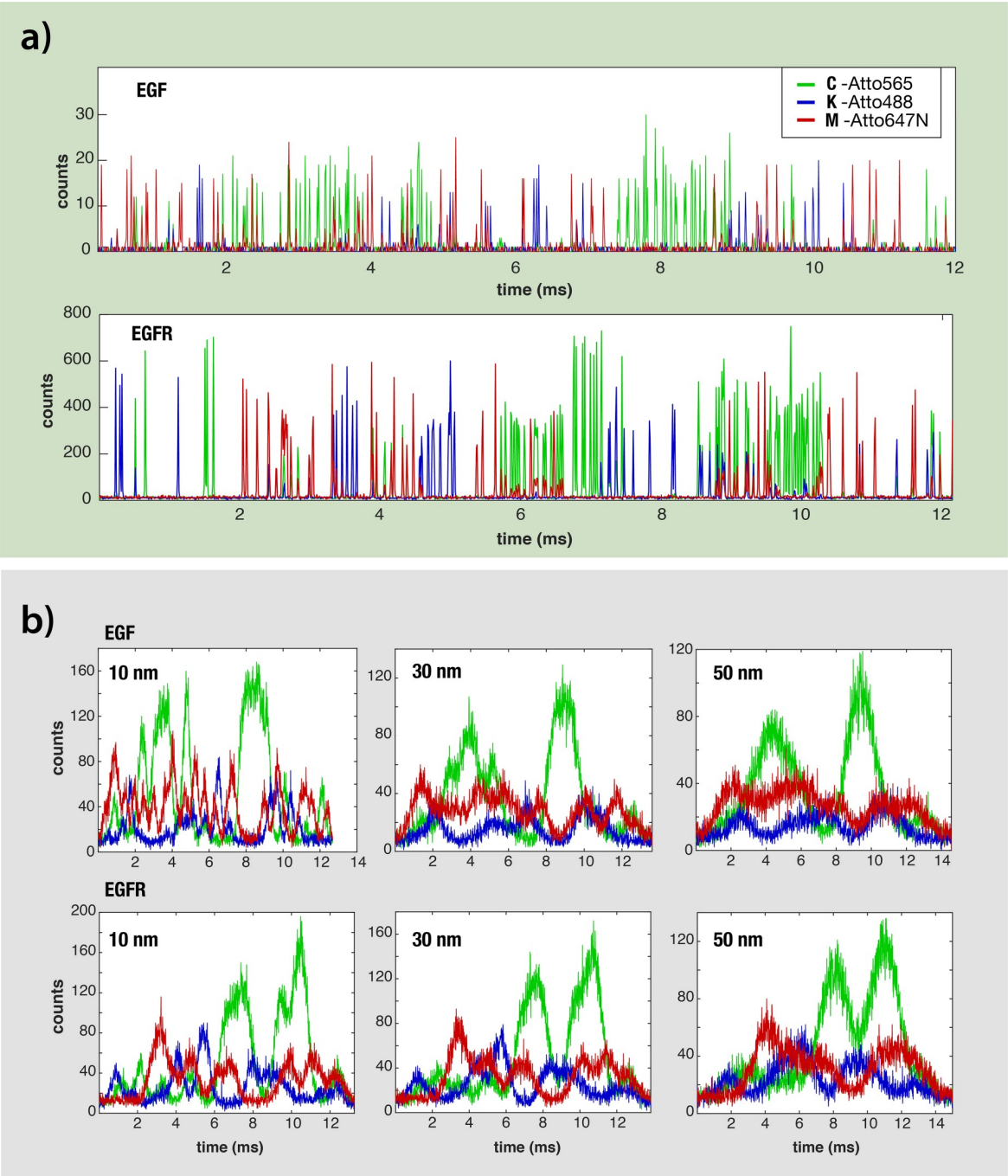


Fig 4. Simulated optical traces of epidermal growth factor (EGF) precursor protein and its receptor EGFR produced under different conditions. The C, K and M amino-acids were labeled using three different fluorophores as indicated. (a) Optical signals simulated using a spatial resolution of 0.5 nm and a labelling efficiency of 100%. (b) optical signals simulated using three distinct spatial resolutions: 10, 30 and 50 nm (from left to right).

https://doi.org/10.1371/journal.pcbi.1007067.g004

values). In contrast, attempting to cross-correlate among 5 different, randomly-chosen, proteins produced in most cases much lower Pearson coefficient values (0.03–0.35). Obviously, this is just a small fraction of all possible cross-correlations. However, even as is, this sample of data suggests that the protein translocation simulator generates highly-reproducible signals.

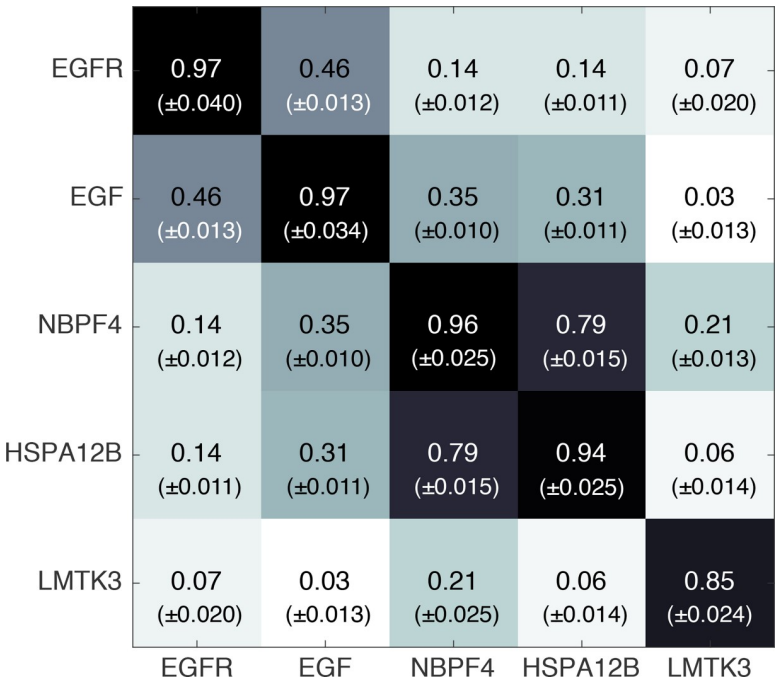
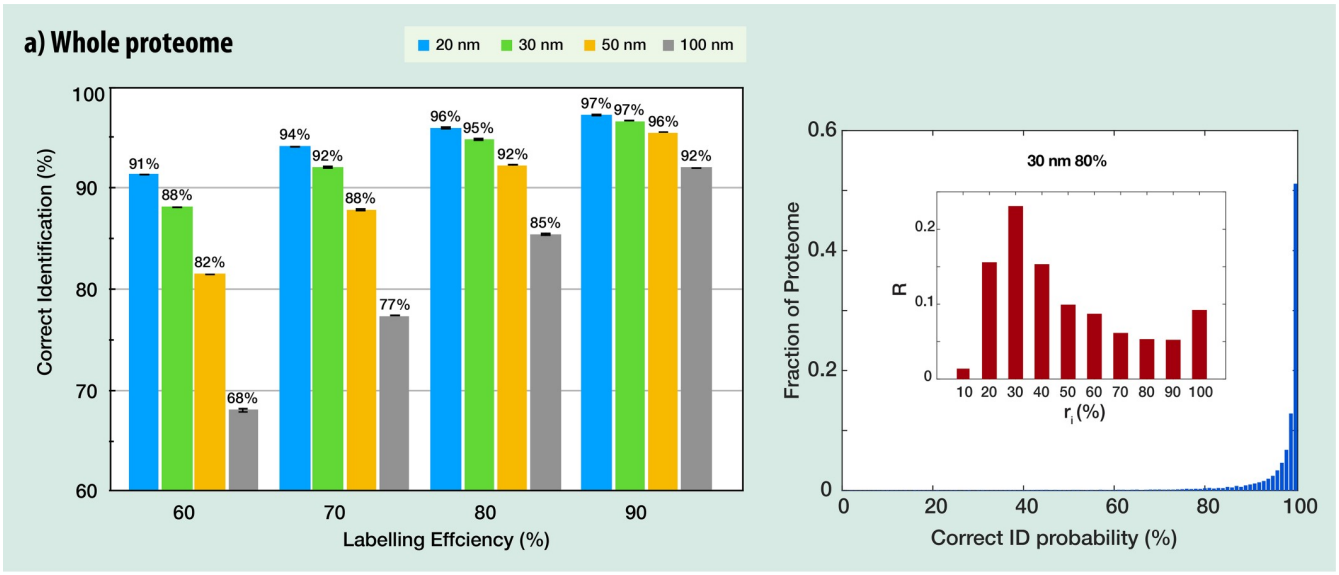


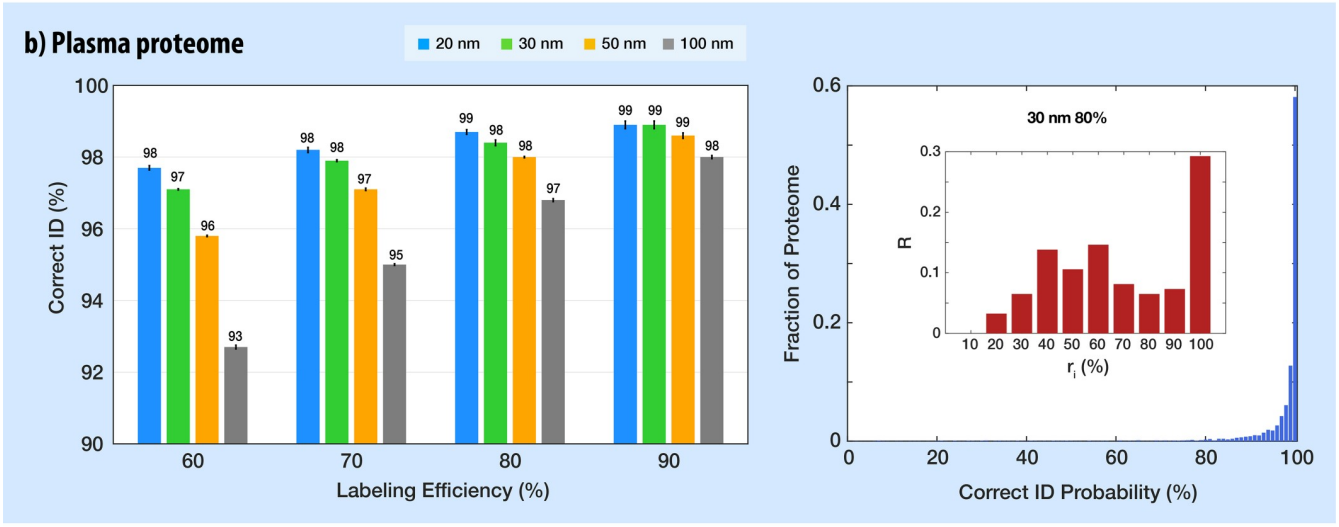
Fig 5. Pearson correlation among pairs of five simulated proteins photon traces. The elements of the correlation matrix, consisting of all Pearson correlation coefficients between all pairs of 50 translocation repeats, were first transformed to Fisher's z, subsequently averaged and finally transformed back into an "average" Pearson correlation coefficient [32]. The standard deviation is given in parentheses.

https://doi.org/10.1371/journal.pcbi.1007067.g005

Whole-proteome protein ID using deep-learning classification

Next we vastly scaled-up our simulations to include thousands of different proteins, each one repeated hundreds of times under different labeling efficiencies, translocation velocities and spatial resolutions. The accurate classification of noisy, low-resolution, time-dependent signals is often encountered in areas such as image and speech recognition and is effectively handled by Convolutional Neural Networks (CNN) approaches [33,34]. We postulated that provided sufficient training, the CNN would be able to identify most proteins based on the tri-color fingerprints. To check this hypothesis, we set up deep-learning whole-proteome analyses. First, we trained the CNN network using a large data set containing at least 80 individual nanopore passages of each protein in the Swiss-Prot database. Then the CNN was presented with new protein translocation events and queried as to the protein identity. This procedure was repeated at least 5 times for whole-proteome analysis allowing us to establish the mean ID accuracy and its standard deviation, for 16 different experimental conditions (Fig 6A). Starting with the highest labelling efficiency (90%, right-hand set) we observed that 96%-97% of all protein translocations were correctly identified, as long as the spatial resolution was $\leq$50nm. The correctly identified protein fraction dropped down to 92% using a 100 nm resolution. A similar pattern can be observed for the other labelling efficiencies with somewhat lower numbers. In the worst-case scenario considered here (100 nm resolution and only 60% labeling efficiency) the CNN nevertheless was able to correctly classify 68% of all translocation events, similar to the ideal case considered in Fig 1C, (C, K, M counts only). In other words, despite the fact that 40% of the target amino-acids were not labeled, and the resolution of the probing was about a third of the optical diffraction limit, the pattern recognition algorithm identified correctly nearly 70% of all protein translocation events. When the labelling efficiency was


a) Whole proteome
20 nm
30 nm
50 nm
100 nm
Correct Identification (%)
Labelling Effciency (%)
91%
88%
82%
68%
94%
92%
88%
77%
96%
95%
92%
85%
97%
97%
96%
92%
Fraction of Proteome
Correct ID probability (%)
30 nm 80%
R
r_i (%)


b) Plasma proteome
20 nm
30 nm
50 nm
100 nm
Correct ID (%)
Labeling Efficiency (%)
98
97
96
93
98
98
97
95
99
98
98
97
99
99
99
98
Fraction of Proteome
Correct ID Probability (%)
30 nm 80%
R
r_i (%)

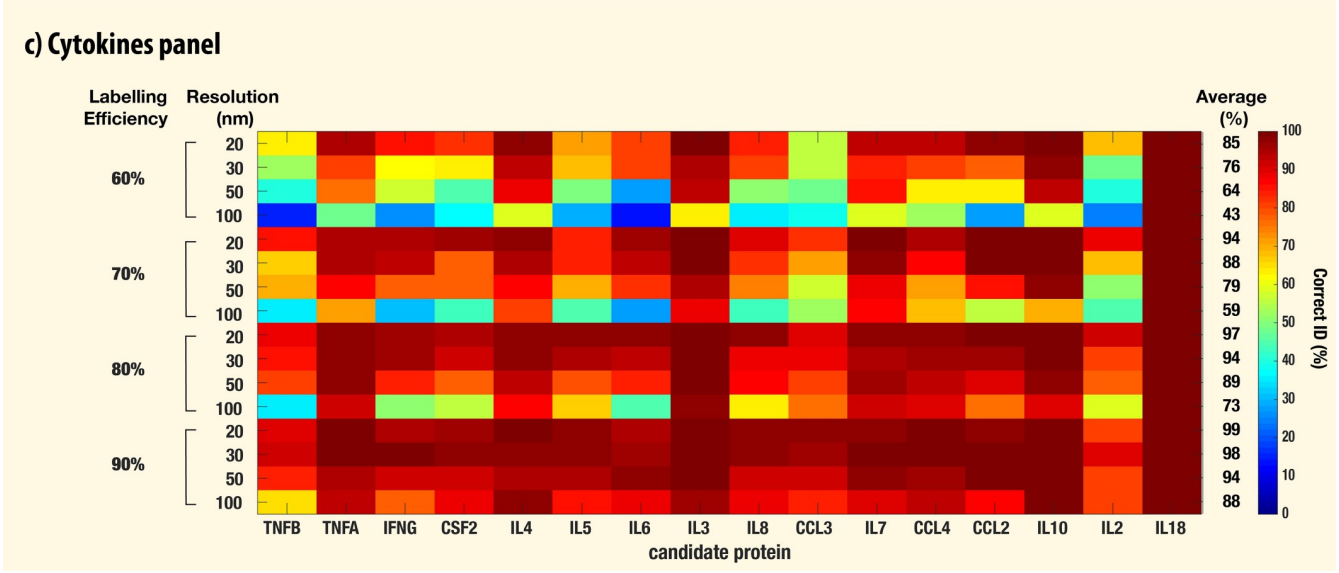

c) Cytokines panel
Labelling Efficiency
Resolution (nm)
60%
70%
80%
90%
20
30
50
100
Average (%)
85
76
64
43
94
88
79
59
97
94
89
73
99
98
94
88
Correct ID (%)
TNFB
TNFA
IFNG
CSF2
IL4
IL5
IL6
IL3
IL8
CCL3
IL7
CCL4
CCL2
IL10
IL2
IL18
candidate protein

Fig 6. CNN-based classification results of: a) whole proteome, b) plasma proteome, and c) a cytokine panel. The fractions of the correctly-identified translocation events from whole-proteome classifications repeated five times are shown in a) and b) left panels. Each classification consisted of five separate training-and-testing of a CNN using 100 translocation events per protein (a total of ~10^7 events), whose resulting correct identifications were averaged. These experiments and analyses were performed under four different spatial resolutions (20, 30, 50 and 100 nm) and labelling efficiencies (60, 70, 80 and 90%). Right-hand panels show the fraction of the proteome correctly identified with probability p when considering a spatial resolution of 30 nm for different labeling efficiencies. The bin size was set to 1%. The insets display the degree of randomness in misclassification. The bin height is given by the fraction of mis-identified proteins R (i.e. proteins that had at least 10% of their events misclassified) at different r_i (fraction of identical mismatch) intervals: $r_i = \max_j n_{ij}/N_i$ for each protein i, where n_{ij} is the number of translocation events misidentified to protein j and N_i the total number of misclassified translocation events. The bin width–r_i interval size–was set to 10%. Other experimental conditions are provided in supporting information file. c) Cytokines panel identification using the same proteins as in the ELISA set "CytokineMAP A". The heat-map represents the correct ID of each cytokine under the specified labelling efficiency and resolution. The average correct ID is provided in the right-hand column.

https://doi.org/10.1371/journal.pcbi.1007067.g006

improved to the expected standards (between 70%-90%) [22,35], and the sensing resolution assumed to be in the 20–30 nm, the correct identification of all translocation was roughly 95%. Increasing the translocation speed of proteins by nearly two orders of magnitude to 2 cm/s (an order of magnitude higher than the mean measured velocity in Fig 3), reduced the ID accuracy (S8 Fig). However, for high labeling efficiencies (80% and 90%) the ID accuracy was high (72% and 81%, respectively).

In addition to the mean accuracies, the CNN algorithm produces a "confusion matrix", which presents the number of times each and every protein x was identified as protein y (where x and y could be any of the proteins in the set). We used this information to calculate the probability density function (pdf) of correct ID for each and every classification set, namely the likelihood that a given protein is correctly identified with probability p. The pdf of correct ID calculated for the case of 30 nm resolution and 80% labelling efficiency (Fig 6A right panel) indicates that 51%, 71% and 89.2% of proteins were correctly identified with probability of 1.0, 0.98–1.0 and 0.9–1.0, respectively. The probability distributions for all other conditions are shown in SI S5 and S6 Figs.

We also analyzed the results for misclassified proteins. Specifically, we were interested to know whether a misclassified protein is likely to be deterministically or randomly misclassified. To investigate the degree of randomness in misclassification, we first selected proteins that had at least 10% misclassified events. Then, we determined the fraction of identical mismatch $r_i = \max_j n_{ij}/N_i$ for each protein i, where n_{ij} is the number of translocation events misidentified to protein j and N_i the total number of misclassified translocation events. With this a high r_i was characteristic of a deterministic misidentification, i.e. protein i is consistently mistaken with another specific protein j, and conversely a low r_i was indicative of a rather random misidentification. As shown in the right panel of Fig 6A, proteins were often confused with several others, suggesting a relatively high degree of randomness in misclassification, while only 10% were consistently mis-identified, that is with the same partner. The distributions for all other conditions are shown in SI S5 and S7 Figs.

Identification of plasma proteome and cytokines panels

We further evaluated the performance of our approach for clinically-relevant applications including whole human plasma proteome and a cytokine panel. In both studies, we kept the CNN training at the whole human proteome, rather than restricting it to the clinical sub-set. Then we presented nanopore translocation traces of the plasma/cytokines proteins and evaluated the classification accuracy as before. Interestingly for the high-spatial resolutions (20 nm and 30 nm) the correct ID of the 3852 plasma proteins was only slightly larger than the whole proteome accuracy at the different labelling efficiencies, reflecting the fact that there is a small set of proteins that are hard to be classified in both cases (Fig 6A and 6B right panel). However,

at the lower resolutions, especially for the 100 nm case in which we observed a significant drop in the ID accuracy for the whole proteome results, we still obtained very high scores for the plasma proteome. Even at the lowest labelling efficiency of 60% at 100 nm resolution the CNN classified correctly 93% of all translocations (Fig 6B). In addition, the fraction of proteins correctly identified with probability between 0.9–1.0 improved over that of the whole-proteome classification, reaching 96.8% for the case of 30nm resolution and 80% labeling efficiency. Finally, close to 30% of mis-identified proteins were consistently mistaken with another specific partner, suggesting that the accuracy of classification could be further significantly improved by relaxing the requirements of correct ID for selected proteins. These results indicate that single-molecule plasma proteome application, which holds great clinical value, does not require extremely-stringent experimental resolutions or super-efficient labelling chemistries (S9–S11 Figs).

The cytokine panel (CytokineMAP [36]) contains 16 proteins involved in inflammation, immune response and repair. We evaluated the CNN classification under 16 different experimental conditions (Fig 6C). At the lowest labelling efficiency of 60% the ID accuracy drops between 43% - 85%, and at the realistic 80% labelling we obtain correct ID in the range of 73% - 97%. However, despite the functional similarity between the candidate cytokines, and the wide range of conditions tested, each was distinguishable from all other cytokines within the commercial test panel. This indicates that our approach has the potential to meet the requirements of a broad range of clinically relevant applications–that are less demanding than whole-proteome identification–with extremely high accuracies and yet very poor experimental conditions (S12 Fig).

Discussion

Single-molecule protein ID and quantification techniques are on the verge of revolutionizing the field of proteomics by enabling researches to achieve single-cell proteomics and to identify low abundance proteins that are essential biomarkers in biomedical and clinical research [7]. Specifically, nanopore discrimination among poly-peptides based solely on two color labeling of C and K residues has recently been demonstrated [21]. Here, we have proposed and simulated the feasibility and limits of a novel method for single-molecule protein ID and quantification using tri-color amino-acid tags and a plasmonic nanopore device. Specifically, we designed a simulator that incorporates a range of physical phenomena to predict and model the behavior of our proposed device and performed a computational analysis taking into account a broad range of experimental conditions to characterize its performance. Importantly, we developed a whole-proteome single-molecule identification algorithm based on convolutional neural networks providing high accuracies (>90% overall), reaching up to 95–97% in challenging but attainable experimental conditions. To facilitate the computational efforts, in this study we approximated each protein translocation dwell-time using a Gaussian distribution function. Notably, past studies [37] successfully utilized CNN to identify signals from exponentially-distributed time-dependent signals, which may better reflect the experimental dwell-time distribution (Fig 3). However, further studies will be required to evaluate the full impact of the temporal distributions of proteins translocation dwell-time on the CNN identification accuracy.

In clinical samples lysine residues may be post-translationally modified hence reducing their labelling efficiency. To account for this effect and for the limitations in the chemical labelling yield, we evaluated the protein identification accuracy under partial labelling conditions. Our results (Fig 6) show that our tri-color protein identification method nevertheless largely circumvents this potential issue, yielding very high accuracies for up to 40% of unlabeled

residues. This is attributed to a redundancy in the tri-color labelling scheme that provides a higher degree of robustness against partial labelling.

Solid-state nanopores can process tens of individual proteins per second, and importantly because our method does not rely exclusively on measurements of the ion-current through the pore, it lends itself for parallel readout of high-density nanopore arrays fabricated on a sub mm^2 membranes, using multi-pixel single-photon sensors [38]. The versatility and robustness of convolutional neural networks tremendously simplify any calibration procedures and even potentially allow protein ID based on partial reads [39]. This ensures that the whole-proteome ID is reliable and compatible with a wide variety of systems, able to overcome real experimental challenges. Furthermore, in many cases (notably for the plasma proteome) misidentified proteins were consistently confused with another specific protein, which in a broad range of applications such as identifying disease-specific biomarkers, may not pose a significant issue as only small-subsets of the proteome are considered, or since the quantification of proteins can be cross-examined with expected counts (e.g. low, medium or high abundance). Finally, we evaluated the expected efficacy of our approach with commercially available applications, even resolving functionally similar proteins in rather poor experimental conditions.

Methods

A theoretical analysis of the proteins ID based on 2 or 3 amino-acid tags

The theoretical identification values were calculated using the human proteome Swiss-Prot database, which contains 20,328 entries. For each entry we extracted the number of the target amino-acids (C, K and M), as well as their order of appearance. For example, the p53 protein would either be characterized by its C,K,M counts (10, 20, 12, respectively) or by the sequence below: MKMMMKKCKMCKCMKMCCCMCCMMCCKKKKKKMKKKKKKKMK, in which all intervening amino-acids were deleted. Proteins having identical characteristic sequences (or C, K and M counts) are grouped together. A protein is identified when it is the sole member of a group. In the case of p53, both the C, K and M counts and the characteristic sequence gave a unique identification. The pie charts (Fig 1C) distribute the proteins according to the size of the group in which they belong to.

A protein labeler program

Each protein primary sequence was transformed into a string ($B(i)$) to which we assigned a value of 1, 2 or 3 corresponding to each of the three *aa* tags (K, C, and M), respectively; and 0 for all other *aa* in the protein sequence. To account for partial or nonspecific labelling a set of randomly selected labeled positions in the string were omitted according to a given labeling efficiency (η_L), and a set of artificial labeled positions were inserted according to a given non-specific labeling efficiency (η_{NS}). It is important to note that nonspecific labeling did not affect all *aa* equally. For instance, in generating a barcode for lysine (K) positions, nonspecific labeling could only be inserted at positions of either threonine, serine and tyrosine (amino-acids which have been shown to compete with NHS-ester-based labeling) with a probability of typically 1% [31]. The strings were generated for the entire Swiss-Prot data base, and were re-generated each time to simulate an uneven labelling of the same protein data sets, as well as whenever we used different values of η_L and η_{NS}.

Finite difference time domain calculation of plasmonic fields

The three-dimensional near field enhancement of the plasmonic structure (2D vertical cross-section shown in Fig 2A) was determined using a finite difference time domain (FDTD) [40]

method solving for Maxwell's time-dependent electromagnetic equations. The architecture over which the FDTD computations were performed comprised a 10 nm-thick silicon (Si) membrane–exhibiting a 3 nm-wide nanopore–on top of which a gold (Au) plasmonic structure was deposited (Fig 2D). An additional 2 nm-thick titanium oxide (TiO_2) adhesive layer was inserted in between the Au structures and underlying Si membrane. The plasmonic structure consisted of a gold ring (inner and outer diameter of 12 and 32 nm, respectively, and a height of 40 nm) centered at the nanopore and embedded inside a gold nanowell (diameter of 120 nm and a height of 100 nm). Water was used as the immersion media.

The excitation field was modeled as a total-field scattering-field source (TFSFS) [41] and the spatial sampling frequency was set to 5 nm^{-1} (taking 60 frequency points over the 500–800 nm wavelength range). The FDTD boundary conditions consisted of 8-layer PMLs (perfectly matched layers) symmetric in the x axis and antisymmetric in the y axis thus minimizing the reflections and the computational cost, respectively. Frequency domain power monitors only were incorporated in the simulation to determine the near field enhancement in the vicinity of the nanopore. All numerical simulations were performed using Lumerical FDTD Solutions (Lumerical, Inc).

Simulation of nanopore-based optical sensing of proteins

To simulate the translocation of the linearized protein through the nanopore, we assumed a unidirectional motion with steps of a single *aa* length ($\Delta \approx 0.35$ nm) and an average velocity u (cm/s). To account for thermal fluctuations in this process we added a random noise term δu at each step (δu can be positive or negative). Hence the simulation step time of the i-th *aa* was defined as $\tau_i = \Delta / (u + \delta u)$. The average protein velocity value was typically ~0.2 cm/s, based on experiments using SDS denatured proteins in solid-state nanopores as shown in Fig 3. Additionally, we tested faster translocations (2 cm/s). The fluorescence emission rate of each fluorophore n in our system $K_{fl,j,n}(t)$ was modeled as a two-state system:

$$K_{fl,j,n}(t) = k_{fl,j} P_{j,n}(t) \quad \text{Eq 1}$$

where $j = 1..3$ correspond to each of the three excitation/emission channels, k_{fl} the fluorescence transition rate and $P_n(t)$ the occupation probability of the excited molecular state $S1$. The fluorophores are excited by up to three laser lines corresponding to the three channels, that form sub-wavelength excitation volumes by means of a plasmonic nanostructure or total internal reflection. The axial full width at half maximum of our Gaussian excitation volume I_{ex} is defined as ξ and is allowed to vary from 5 nm to 200 nm in order to account for broad possible experimental conditions. The emitted light from the three-color channels is assumed to be acquired with given efficiencies η_j, which include both the optical transmission efficiencies and the photodetector efficiencies. The photon counts I_i^j at each channel j during each step i of the protein translocation is then determined by summing the emissions of all the fluorophores n that resides within the excitation volume. Namely:

$$I_i^j = \eta_j \sum_n K_{fl,j,n}(t_i) + k_{bg}\tau_i = \eta_j \sum_n k_{fl,j} P_{j,n}(t_i) + k_{bg}\tau_i \quad \text{Eq 2}$$

$$\begin{cases} P_{j,n}(t_i) = P_{j,n}(t_{i-1}) + \left(\dfrac{k_{ex,j}(n)}{k_j(n)} - P_{j,n}(t_{i-1}) \right) (1 - e^{-k_j(n) t_i}) & \text{Eq 3} \\ k_j(n) = k_{ex,j}(n) + k_{S1,j} = \sigma_{ex,j} \dfrac{I_{ex,j}(n) \lambda_{ex,j}}{h c_0} + \tau_{S1,j}^{-1} & \text{Eq 4} \end{cases}$$

where k_{bg} is the background emission rate, t_i the time at which step the translocation occurred

such that $t_i - t_{i-1} = \tau_i$, $k_{ex,j}(n)$ is the excitation rate of the fluorophore n of channel j, $\sigma_{ex,j}$ is its absorption coefficient, $\lambda_{ex,j}$ is the excitation wavelength and $\tau_{S1,j}$ is its excited state lifetime.

The number of cycles (S0→S1→S0) undergone by each fluorophore was capped to account for photobleaching according to a decaying exponential distribution. Specifically, the maximum number of cycles performed by each fluorophore before photobleaching was given by a random number drawn from a decaying exponential distribution with a characteristic decay of ~10^6. Finally, we applied a Poisson distribution to the photon counts I_i^j to simulate shot noise.

To include energy transfer (such as Förster Energy Transfer and homo-transfer) in our system we calculated a 2D distance matrix for each fluorophore in our system. The distances between the labelled *aa's* (or fluorophores) in each linearized protein were subsequently used to calculate the Förster energy transfers of each fluorophore from and to each of its neighboring emitters. As a proxy for the exact energy transfer, two additional transition rates accounting for energy gain and loss were incorporated in the fluorophore two-state model:

$$\begin{cases} k_{FRET+,j}(n) = \dfrac{1}{hc_0}\sum_i\sum_{m\neq n}\sigma_{ex,i}I_{ex,i}(m)E_{n\leftarrow m}\lambda_{ex,i} & \text{Eq 5} \\ k_{FRET-,j}(n) = \dfrac{\sigma_{ex,j}I_{ex,j}(n)\lambda_{ex,j}}{hc_0}\sum_i\sum_{m\neq n}E_{m\leftarrow n} & \text{Eq 6} \end{cases}$$

where $E_{m\leftarrow n} = (1+(|x_m - x_n|/R_{0,m\leftarrow n})^6)^{-1}$ is the FRET energy transfer efficiency from fluorophore n to m, x_n is the position of fluorophore n along the denatured protein and $R_{0,m\leftarrow n}$ is the Förster-radius of the (n,m) dye pair when considering an energy transfer from fluorophore n to m. The transition rates $k_{ex,j}(n)$ and $k_j(n)$ in Eq 4 were corrected to account for FRET accordingly:

$$\begin{cases} k_{ex,j}(n) \rightarrow k_{ex,j}(n) + k_{FRET+,j}(n) \\ k_j(n) \rightarrow k_j(n) + k_{FRET+,j}(n) + k_{FRET-,j}(n) \end{cases}$$

The code was implemented using MATLAB, and the optical readouts of the three channels were determined by running this procedure for each labeling string.

Protein classification and mapping of optical reads to protein IDs

For the purpose of a multi-class (the human proteome comprises more than twenty thousand proteins) classification of time-series that exhibit specific patterns, we used convolutional neural networks (CNN) that have shown great promise in the field of pattern recognition, including image classification, which similarly requires tens of thousands of classes [42,43]. Specifically, we used the python deep learning package Keras on a four GPU architecture (NVIDIA Tesla K40), which leads to a CNN whole-proteome training time of ~2 h only. The CNN model relied on four sequential layers–a convolutional layer, a normalization layer in which dropout was applied and a pooling layer–followed by a multi-layer perceptron. In brief, the convolutional layer filters (at a given step or stride size) the translocation time-series with a large set of kernels of a specific size. The resulting activation or feature map it provides is further transformed by the normalization layer such that the mean and standard deviation of the activation map approach zero and one, respectively. Next, the dropout circumvents overfitting of the CNN to the training dataset by setting a random subset of activations to zero. The last pooling layer performs a down-sampling operation on the activation map to further prevent overfitting of the training dataset and the computational load. The multi-layer perceptron

consists of a single densely-connected neural network layer, each neuron outputting the probability of belonging to the class it represents ('softmax' activation function).

The hyper-parameters were optimized according to standard procedures, that is maximizing the accuracy of the CNN trained over five to ten epochs per hyper-parameter set. Once finely adjusted, the CNN was trained using twenty epochs to yield the greatest accuracy. The protein identification accuracy as determined by the CNN was calculated as the fraction of correctly classified translocation events from the test dataset. We partitioned randomly the dataset into five pairs of training and testing sub-sets, and for which we determined the identification accuracy. The final accuracy was calculated as the average between them where a typical test set included ~400,000 translocation events.

SDS-denatured protein translocation experiments

Solid-state nanopores were fabricated using a laser drilling method in 17 nm-thick SiN_x membranes as described previously [44]. Human serum albumin (Biological Industries Inc. 30-O595-A) was first treated by TCEP (5 mM) at room temperature for 30 min to break disulfide bonds and subsequently denatured at 90˚C for 5 min in PBS with 2% sodium-dodecyl sulfate (SDS). The resulting albumin concentration was further diluted (100:1) to <1 nM in buffer (PBS/0.4M NaCl/ 0.1% SDS/ 1mM EDTA) for nanopore translocation experiments performed under a 300 mW bias. A custom-made LabVIEW interface was used to acquire and analyze each event. Scatter plots and dwell-time distributions were generated using Igor Pro (Wavemetrics).

Supporting information

S1 Note. All the numerical simulations were performed using Lumerical FDTD Solutions (Lumerical, Inc), solving for Maxwell's equations using a finite-difference time-domain method. Additional information regarding the FDTD simulations can be found in the Supporting Information.
(DOCX)

S1 Fig. Simulated optical traces of different proteins with or without a fluorophore triplet state. The spatial resolution and labeling efficiency were fixed in all cases to 30nm and 100%, respectively. Left column shows the simulated optical traces using a two-state (ground and excited) fluorophore model; right column using a three-state (ground, excited and triplet) model. Transition rates in between all states were determined according to the manufacturer (when available) and to published work (see Article).
(TIFF)

S2 Fig. Simulated optical traces of the epidermal growth factor (EGF) precursor protein and its receptor EGFR generated using two spatial resolutions: 100 and 150nm (from left to right). The labeling efficiency was set to 100% and the average translocation velocity to 0.0035 cm/s.
(TIFF)

S3 Fig. Simulated optical traces of the epidermal growth factor (EGF) precursor protein in different experimental conditions. (a) optical signals simulated using a spatial resolution of 0.5nm and a labelling efficiency of 100%. (b) optical signals simulated using three distinct spatial resolutions: 10, 30 and 50nm (top), three distinct labeling efficiencies: 90%, 80% and 70% (middle), three velocity fluctuations: 20%, 30% and 40% of the mean translocation velocity $v = 0.035$ cm/s (bottom).
(TIFF)

S4 Fig. Simulated optical traces of the B Double Prime 1 (BDP1) protein in different experimental conditions. (a) optical signals simulated using a spatial resolution of 0.5nm and a labelling efficiency of 100%. (b) optical signals simulated using three distinct spatial resolutions: 10, 30 and 50nm (top), three distinct labeling efficiencies: 90%, 80% and 70% (middle), three velocity fluctuations: 20%, 30% and 40% of the mean translocation velocity $v = 0.035\ cm/s$ (bottom).
(TIFF)

S5 Fig. Whole-proteome probability density function of correct identification and degree of randomness in misclassification at 30nm. The fraction of the whole proteome that was correctly identified with probability p (a) and the degree of randomness in misclassification (b) were determined for 30nm and four labeling efficiencies (60, 70, and 90%; the remaining 80% as well as the CNN accuracy bar plot are shown in the article). The bin size was set to 1% in all histograms. The bin height of histograms in (b) is given by the fraction of mis-identified proteins R (i.e. proteins that had at least 10% of their events misclassified) at different r_i (fraction of identical mismatch) intervals: $r_i = \max_j n_{ij}/N_i$ for each protein i, where n_{ij} is the number of translocation events misidentified to protein j and N_i the total number of mis-classified translocation events. High is characteristic of a low degree of randomness, and vice-versa low of a high degree of randomness. The bin width–r_i interval size–was set to 10%. The value in parentheses indicate the percentage of mis-identified proteins of a whole-proteome experiment.
(TIFF)

S6 Fig. Whole-proteome probability density function of correct identification for different experimental conditions. The fraction of the proteome that was correctly identified with probability p was determined for three spatial resolutions (20, 50 and 100nm; 30nm shown in article) and four labeling efficiencies (60, 70, 80 and 90%). The bin size was set to 1% in all histograms.
(TIFF)

S7 Fig. Whole-proteome degree of randomness in misclassification for different experimental conditions. The bin height is given by the fraction of mis-identified proteins R (i.e. proteins that had at least 10% of their events misclassified) at different r_i (fraction of identical mismatch) intervals: $r_i = \max_j n_{ij}/N_i$ for each protein i, where n_{ij} is the number of translocation events misidentified to protein j and N_i the total number of mis-classified translocation events. High is characteristic of a low degree of randomness, and vice-versa low of a high degree of randomness. The bin width–r_i interval size–was set to 10%. The value in parentheses indicate the percentage of mis-identified proteins of a whole-proteome experiment. The degree of randomness in misclassification was determined for three spatial resolutions (20, 50 and 100nm; 30nm shown in article) and four labeling efficiencies (60, 70, 80 and 90%).
(TIFF)

S8 Fig. Whole-proteome protein identification accuracy as a function of amino-acid dwell time and labelling efficiency. The spatial resolution was fixed to 30 nm and the dwell-time was defined as the time it took a peptide to translocate over the length of a single amino-acid. The corresponding translocation velocities are 2, 0.2 and 0.035 cm/s. The APD binning was set to 1 μs. The CNN classification was still robust to low labeling efficiency and realistic spatial and temporal resolutions, expected in real experiments.
(TIF)

S9 Fig. Plasma-proteome probability density function of correct identification and degree of randomness in misclassification at 30nm. The fraction of the plasma proteome that was correctly identified with probability p (a) and the degree of randomness in misclassification (b) were determined for 30nm and four labeling efficiencies (60, 70, and 90%; the remaining 80% as well as the CNN accuracy bar plot are shown in the article). The bin size was set to 1% in all histograms. The bin height of histograms in (b) is given by the fraction of mis-identified proteins R (i.e. proteins that had at least 10% of their events misclassified) at different r_i (fraction of identical mismatch) intervals: $r_i = \max_j n_{ij}/N_i$ for each protein i, where n_{ij} is the number of translocation events misidentified to protein j and N_i the total number of mis-classified translocation events. High is characteristic of a low degree of randomness, and vice-versa low of a high degree of randomness. The bin width–r_i interval size–was set to 10%. The value in parentheses indicate the percentage of mis-identified proteins of a plasma-proteome experiment.
(TIFF)

S10 Fig. Plasma-proteome probability density function of correct identification for different experimental conditions. The fraction of the plasma proteome that was correctly identified with probability p was determined for three spatial resolutions (20, 50 and 100nm; 30nm shown in article) and four labeling efficiencies (60, 70, 80 and 90%). The bin size was set to 1% in all histograms.
(TIFF)

S11 Fig. Plasma-proteome degree of randomness in misclassification for different experimental conditions. The bin height is given by the fraction of mis-identified proteins R (i.e. proteins that had at least 10% of their events misclassified) at different r_i (fraction of identical mismatch) intervals: $r_i = \max_j n_{ij}/N_i$ for each protein i, where n_{ij} is the number of translocation events misidentified to protein j and N_i the total number of mis-classified translocation events. High is characteristic of a low degree of randomness, and vice-versa low of a high degree of randomness. The bin width–r_i interval size–was set to 10%. The value in parentheses indicate the percentage of mis-identified proteins of a plasma-proteome experiment. The degree of randomness in misclassification was determined for three spatial resolutions (20, 50 and 100nm; 30nm shown in article) and four labeling efficiencies (60, 70, 80 and 90%).
(TIFF)

S12 Fig. Identification of proteins targeted by different commercial ELISA sets. a) Whole-proteome CNN accuracy of the CytokineMAP B kit proteins for four spatial resolutions (20, 30, 50 and 100nm) and four labeling efficiencies (60, 70, 80 and 90%). b) Whole-proteome CNN accuracy of the MetabolicMAP kit proteins for four spatial resolutions (20, 30, 50 and 100nm) and four labeling efficiencies (60, 70, 80 and 90%). c) Whole-proteome CNN accuracy of the NeuroMAP A kit proteins and misclassification distribution for four spatial resolutions (20, 30, 50 and 100nm) and four labeling efficiencies (60, 70, 80 and 90%).
(TIFF)

S13 Fig. Simulated optical traces of different proteins with or without a fluorophore triplet state. The spatial resolution and labeling efficiency were fixed in all cases to 30nm and 100%, respectively. Left column shows the simulated traces optical traces using a two-state (ground and excited) fluorophore model; right column using a three-state (ground, excited and triplet) model. Transition rates in between all states were determined according to the manufacturer (when available) and to published work (see Article).
(TIFF)

Acknowledgments

We acknowledge the personnel at the Boston University Shared Computing Cluster (SCC) for providing technical support.

Author Contributions

Conceptualization: Shilo Ohayon, Arik Girsault, Shai Shen-Orr, Amit Meller.

Funding acquisition: Amit Meller.

Methodology: Shilo Ohayon, Arik Girsault, Amit Meller.

Resources: Amit Meller.

Software: Shilo Ohayon, Arik Girsault, Maisa Nasser.

Supervision: Amit Meller.

Writing – original draft: Shilo Ohayon, Arik Girsault, Shai Shen-Orr, Amit Meller.

Writing – review & editing: Shilo Ohayon, Arik Girsault, Maisa Nasser, Shai Shen-Orr, Amit Meller.

References

1. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et al. DNA sequencing at 40: Past, present and future. Nature. 2017; 550(7676):345–53. https://doi.org/10.1038/nature24286 PMID: 29019985
2. Milo R, Jorgensen P, Moran U, Weber G, Springer M. BioNumbers The database of key numbers in molecular and cell biology. Nucleic Acids Res. 2009; 38:D750–3. https://doi.org/10.1093/nar/gkp889 PMID: 19854939
3. Bekker-Jensen DB, Kelstrup CD, Batth TS, Larsen SC, Haldrup C, Bramsen JB, et al. An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. Cell Syst. 2017 Jun; 4(6):587–599.e4. https://doi.org/10.1016/j.cels.2017.05.009 PMID: 28601559
4. Bendall SC, Nolan GP, Roederer M, Chattopadhyay PK. A deep profiler's guide to cytometry. Trends Immunol. 2012 Jul; 33(7):323–32. https://doi.org/10.1016/j.it.2012.02.010 PMID: 22476049
5. Swaminathan J, Boulgakov AA, Hernandez ET, Bardo AM, Bachman JL, Marotta J, et al. Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. Nat Biotechnol. 2018 Oct 22; 36 (11):1076–82.
6. van Ginkel J, Filius M, Szczepaniak M, Tulinski P, Meyer AS, Joo C. Single-molecule peptide fingerprinting. Proc Natl Acad Sci. 2018 Mar 27; 115(13):3338–43. https://doi.org/10.1073/pnas.1707207115 PMID: 29531063
7. Restrepo-Pérez L, Joo C, Dekker C. Paving the way to single-molecule protein sequencing. Nat Nanotechnol. 2018 Sep 6; 13(9):786–96. https://doi.org/10.1038/s41565-018-0236-6 PMID: 30190617
8. Dekker C. Solid-state nanopores. Nat Nanotechnol. 2007 Apr 4; 2(4):209–15. https://doi.org/10.1038/nnano.2007.27 PMID: 18654264
9. Bayley H. Nanopore Sequencing: From Imagination to Reality. Clin Chem. 2015 Jan 1; 61(1):25–31. https://doi.org/10.1373/clinchem.2014.223016 PMID: 25477535
10. Yusko EC, Johnson JM, Majd S, Prangkio P, Rollings RC, Li J, et al. Controlling protein translocation through nanopores with bio-inspired fluid walls. Nat Nanotechnol. 2011 Apr 20; 6(4):253–60. https://doi.org/10.1038/nnano.2011.12 PMID: 21336266
11. Plesa C, Kowalczyk SW, Zinsmeester R, Grosberg AY, Rabin Y, Dekker C. Fast Translocation of Proteins through Solid State Nanopores. Nano Lett. 2013 Feb 13; 13(2):658–63. https://doi.org/10.1021/nl3042678 PMID: 23343345
12. Nir I, Huttner D, Meller A. Direct Sensing and Discrimination among Ubiquitin and Ubiquitin Chains Using Solid-State Nanopores. Biophys J. 2015 May; 108(9):2340–9. https://doi.org/10.1016/j.bpj.2015.03.025 PMID: 25954891
13. Van Meervelt V, Soskine M, Singh S, Schuurman-Wolters GK, Wijma HJ, Poolman B, et al. Real-Time Conformational Changes and Controlled Orientation of Native Proteins Inside a Protein Nanoreactor. J Am Chem Soc. 2017 Dec 27; 139(51):18640–6. https://doi.org/10.1021/jacs.7b10106 PMID: 29206456

14. Waduge P, Hu R, Bandarkar P, Yamazaki H, Cressiot B, Zhao Q, et al. Nanopore-Based Measurements of Protein Size, Fluctuations, and Conformational Changes. ACS Nano. 2017 Jun 27; 11 (6):5706–16. https://doi.org/10.1021/acsnano.7b01212 PMID: 28471644

15. Varongchayakul N, Huttner D, Grinstaff MW, Meller A. Sensing Native Protein Solution Structures Using a Solid-state Nanopore: Unraveling the States of VEGF. Sci Rep. 2018 Dec 17; 8(1):1017. https://doi.org/10.1038/s41598-018-19332-y PMID: 29343861

16. Varongchayakul N, Song J, Meller A, Grinstaff MW. Single-molecule protein sensing in a nanopore: a tutorial. Chem Soc Rev. 2018; 47(23):8512–24. https://doi.org/10.1039/c8cs00106e PMID: 30328860

17. Chinappi M, Cecconi F. Protein sequencing via nanopore based devices: A nanofluidics perspective. Journal of Physics Condensed Matter. 2018.

18. Yao Y, Docter M, van Ginkel J, de Ridder D, Joo C. Single-molecule protein sequencing through fingerprinting: computational assessment. Phys Biol. 2015 Aug 12; 12(5):055003. https://doi.org/10.1088/1478-3975/12/5/055003 PMID: 26266455

19. Kennedy E, Dong Z, Tennant C, Timp G. Reading the primary structure of a protein with 0.07 nm3 resolution using a subnanometre-diameter pore. Nat Nanotechnol. 2016 Nov 25; 11(11):968–76. https://doi.org/10.1038/nnano.2016.120 PMID: 27454878

20. Restrepo-Pérez L, John S, Aksimentiev A, Joo C, Dekker C. SDS-assisted protein transport through solid-state nanopores. Nanoscale. 2017; 9(32):11685–93. https://doi.org/10.1039/c7nr02450a PMID: 28776058

21. Wang R, Gilboa T, Song J, Huttner D, Grinstaff MW, Meller A. Single-Molecule Discrimination of Labeled DNAs and Polypeptides Using Photoluminescent-Free TiO_2 Nanopores. ACS Nano. 2018 Nov 27; 12(11):11648–56. https://doi.org/10.1021/acsnano.8b07055 PMID: 30372037

22. Lin S, Yang X, Jia S, Weeks AM, Hornsby M, Lee PS, et al. Redox-based reagents for chemoselective methionine bioconjugation. Science (80-). 2017 Feb 10; 355(6325):597–602. https://doi.org/10.1126/science.aal3316 PMID: 28183972

23. Assad ON, Di Fiori N, Squires AH, Meller A. Two Color DNA Barcode Detection in Photoluminescence Suppressed Silicon Nitride Nanopores. Nano Lett. 2015 Jan 14; 15(1):745–52. https://doi.org/10.1021/nl504459c PMID: 25522780

24. Assad ON, Gilboa T, Spitzberg J, Juhasz M, Weinhold E, Meller A. Light-Enhancing Plasmonic-Nanopore Biosensor for Superior Single-Molecule Detection. Adv Mater. 2017 Mar; 29(9):1605442.

25. Apweiler R. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. 2004 Jan; 32 (90001):115D – 119.

26. Staikos G, Dondos A. Study of the sodium dodecyl sulphate–protein complexes: evidence of their wormlike conformation by treating them as random coil polymers. Colloid Polym Sci. 2009 Aug 10; 287 (8):1001–4.

27. Muthukumar M. Communication: Charge, diffusion, and mobility of proteins through nanopores. J Chem Phys. 2014 Aug 28; 141(8):081104. https://doi.org/10.1063/1.4894401 PMID: 25172998

28. Yeow EKL, Melnikov SM, Bell TDM, De Schryver FC, Hofkens J. Characterizing the Fluorescence Intermittency and Photobleaching Kinetics of Dye Molecules Immobilized on a Glass Surface. J Phys Chem A. 2006 Feb; 110(5):1726–34. https://doi.org/10.1021/jp055496r PMID: 16451001

29. Eggeling C, Widengren J, Brand L, Schaffer J, Felekyan S, Seidel CAM. Analysis of Photobleaching in Single-Molecule Multicolor Excitation and Förster Resonance Energy Transfer Measurements †. J Phys Chem A. 2006 Mar; 110(9):2979–95. https://doi.org/10.1021/jp054581w PMID: 16509620

30. Blom H, Chmyrov A, Hassler K, Davis LM, Widengren J. Triplet-State Investigations of Fluorescent Dyes at Dielectric Interfaces Using Total Internal Reflection Fluorescence Correlation Spectroscopy. J Phys Chem A. 2009 May 14; 113(19):5554–66. https://doi.org/10.1021/jp8110088 PMID: 19374408

31. Abello N, Kerstjens HAM, Postma DS, Bischoff R. Selective Acylation of Primary Amines in Peptides and Proteins. J Proteome Res. 2007 Dec; 6(12):4770–6. https://doi.org/10.1021/pr070154e PMID: 18001078

32. Corey DM, Dunlap WP, Burke MJ. Averaging Correlations: Expected Values and Bias in Combined Pearson r s and Fisher's z Transformations. J Gen Psychol. 1998 Jul; 125(3):245–61.

33. Zheng Y, Liu Q, Chen E, Ge Y, Zhao JL. Time Series Classification Using Multi-Channels Deep Convolutional Neural Networks. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2014. p. 298–310.

34. Acharya UR, Oh SL, Hagiwara Y, Tan JH, Adeli H. Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. Comput Biol Med. 2018 Sep; 100:270–8. https://doi.org/10.1016/j.compbiomed.2017.09.017 PMID: 28974302

35. Pretzer E, Wiktorowicz JE. Saturation fluorescence labeling of proteins for proteomic analyses. Anal Biochem. 2008 Mar; 374(2):250–62. https://doi.org/10.1016/j.ab.2007.12.014 PMID: 18191033

36. Myriadbm.
37. Misiunas K, Ermann N, Keyser UF. QuipuNet: Convolutional Neural Network for Single-Molecule Nanopore Sensing. Nano Lett. 2018 Jun 13; 18(6):4040–5. https://doi.org/10.1021/acs.nanolett.8b01709 PMID: 29845855
38. Gilboa T, Meller A. Optical sensing and analyte manipulation in solid-state nanopores. Analyst. 2015; 140(14):4733–47. https://doi.org/10.1039/c4an02388a PMID: 25684652
39. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning Deep Features for Discriminative Localization. cv-foundation.org. 2016. p. 2921–9.
40. Yee Kane. Numerical solution of initial boundary value problems involving maxwell's equations in isotropic media. IEEE Trans Antennas Propag. 1966 May; 14(3):302–7.
41. Merewether DE, Fisher R, Smith FW. On Implementing a Numeric Huygen's Source Scheme in a Finite Difference Program to Illuminate Scattering Bodies. IEEE Trans Nucl Sci. 1980; 27(6):1829–33.
42. Simonyan K, 1409.1556 AZ arXiv preprint arXiv, 2014. Very deep convolutional networks for large-scale image recognition. arxiv.org.
43. Ciresan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2012. p. 3642–9.
44. Gilboa T, Zrehen A, Girsault A, Meller A. Optically-Monitored Nanopore Fabrication Using a Focused Laser Beam. Sci Rep. 2018, 8: 9765. https://doi.org/10.1038/s41598-018-28136-z PMID: 29950607