analytical chemistry

pubs.acs.org/ac Article

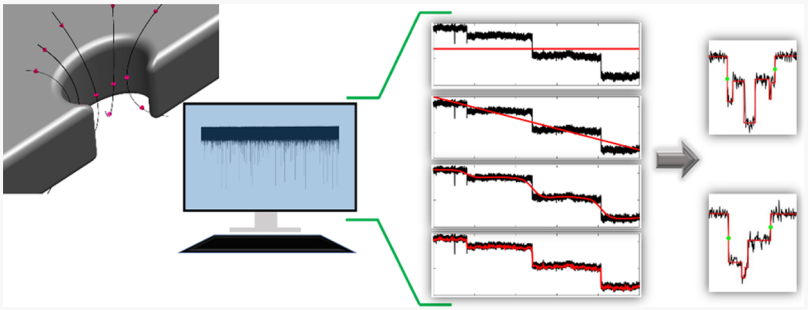
# Nanopore Data Analysis: Baseline Construction and Abrupt Change-Based Multilevel Fitting

Y. M. Nuwan D. Y. Bandara, Jugal Saharia, Buddini I. Karawdeniya, Patrick Kluth,* and Min Jun Kim*

Cite This: *Anal. Chem.* 2021, 93, 11710–11718 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

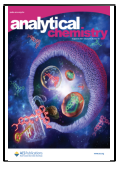




**ABSTRACT:** Solid-state nanopore technology delivers single-molecule resolution information, and the quality of the deliverables hinges on the capability of the analysis platform to extract maximum possible events and fit them appropriately. In this work, we present an analysis platform with four baseline fitting methods adaptive to a wide range of nanopore traces (including those with a step or abrupt changes where pre-existing platforms fail) to maximize extractable events (2× improvement in some cases) and multilevel event fitting capability. The baseline fitting methods, in the increasing order of robustness and computational cost, include arithmetic mean, linear fit, Gaussian smoothing, and Gaussian smoothing and regressed mixing. The performance was tested with ultra-stable to vigorously fluctuating current profiles, and the event count increased with increasing fitting robustness prominently for vigorously fluctuating profiles. Turning points of events were clustered using the *dbscan* method, followed by segmentation into preliminary levels based on abrupt changes in the signal level, which were then iteratively refined to deduce the final levels of the event. Finally, we show the utility of clustering for multilevel DNA data analysis, followed by the assessment of protein translocation profiles.

Solid-state nanopores (SSNs) offer capabilities transcending those of average ensemble tools with an expanding footprint in single-molecule science with advantages such as low cost, high throughput, scalability, and robustness. The operational principle is ostensibly simple: an analyte is added to one side and driven across the pore in response to an applied voltage perturbing the open-pore current, stamping analyte-specific information. The applications of SSNs span a host of bio-macromolecules,[1−3] bio- and synthetic particles,[4−6] and polymers.[7,8] SSNs, unlike their biological counterparts, are somewhat notorious for open-pore drifts due to pore enlargement over time,[9,10] among other reasons. Moreover, even when open-pore drift is absent, current fluctuations (*i.e.*, noisy open-pore profiles) are not uncommon in most SSNs and become more apparent at higher applied voltages, as noted in our previous work.[1,11,12] The open-pore current drift due to pore enlargement (with time) could be largely negated with a window-based analysis, where the analysis is performed on equisized trace segments rather than on the entire trace itself.[13] However, local and more pronounced open-pore fluctuations are more challenging and could easily lead to flawed event detection (*i.e.*, baseline structures being flagged as events) and/or undercounting of event populations. Although one could argue to use a smaller window size where fluctuations become insignificant, it would alienate longer duration events from the analysis and become problematic where such events are significant in the event population. Thus, baseline processing is particularly important to extract maximum possible events, especially in vigorously fluctuating current profiles. Such current fluctuations are more apparent in 2D materials such as graphene,[14] hexagonal boron nitride (h-BN),[15] and molybdenum disulfide ($MoS_2$)[16] and somewhat less in silicon nitride ($Si_xN_y$) pores fabricated using controlled dielectric breakdown (CDB) and chemically tuned CDB (CT-CDB),[11] as will be shown later. Thus, the noise level depends on the fabrication method and device architecture.[17,18] The noise and

**Received:** April 18, 2021
**Accepted:** July 28, 2021
**Published:** August 17, 2021






https://doi.org/10.1021/acs.analchem.1c01646
*Anal. Chem.* 2021, 93, 11710−11718

baseline fluctuations can be quantified through, for example, power spectral density graphs and root mean square of the open-pore current as a function of time. Analysis methods are challenged by pores that have high noise factors and fluctuations.

While a simple duration and depth analysis might be adequate for preliminary analysis, it would not yield information on finer conformations (*e.g.*, knots in the case of DNA). Thus, there has been great interest in the nanopore community to develop tools and methods for robust data analysis. More notable methods include the CUSUM-based multilevel fitting,[13,19,20] MOSAIC algorithm,[21] and hidden Markov model approaches.[22,23] The *OpenNanopore* application (CUSUM-based) developed by the Radenovic group[19] has, to some extent, paved the way for a standardized analysis approach. Like baseline profiles, the current signal within the event is not immune to fluctuations. This could lead to over/underestimation of signal levels. Here, we used a clustering-based method (density-based spatial clustering of applications with noise, *dbscan* clustering, for which technical information about the execution of the clustering method can be readily found in the *MathWorks* website) to (i) identify the boundaries of the event and for (ii) preliminary estimation of the levels within the event. These clusters (*i.e.*, points grouped based on the likelihood to be similar to each other than to those in any other group) then serve as the initial *guess* for the number of levels within the event for segmentation of the event to levels based on abrupt signal level changes. However, the performance of unsupervised clustering methods hinges on the initial parameter thresholds defined by the user. Thus, we have introduced a set of checkpoints in the proposed algorithm for the analysis to be insusceptible to improper clustering. For example, if overclustering takes place, the initial guess for the number of levels would exceed the true level count. In such cases, adjacent levels are iteratively compared using a user-defined current or standard deviation-dependent threshold (discussed later in the manuscript) to see if the levels should stay separated or be merged. Furthermore, in multilevel events (*i.e.*, events with more than one step change such as those arising from non-linear DNA translocations[24,25]), a simple mean of the data points could sometimes be insufficient to represent the current value of the level ($\Delta I_{level}$) due to capture of edge points leading to adjacent levels. This could, especially in attenuated levels, underestimate the $\Delta I_{level}$. A turning point-based method is used in the program to overcome this shortcoming. The versatility of our program was tested with both double-stranded DNA (dsDNA) and the *holo* form of the human serum transferrin (*holo*-hSTf) protein, using two different membrane types, and three fabrication methods. While dsDNA is the gold standard for nanopore experiments, proteins offer unique challenges due to their charge heterogeneity and conformational changes in solutions in response to external stimuli such as solution chemistry and applied voltage. Any analyte, irrespective of its type, must first diffuse from the bulk to the capture zone of the nanopore, after which its transport becomes drift-dominant and is eventually funneled through the pore in response to the applied electric field. The transport could be either diffusion- or barrier-limited, largely depending on the applied voltage and size of the molecule. Moreover, the capture rate provides insights into the limiting mechanism with a linear voltage response, indicating a diffusion-limited transport mechanism, and an exponential response, pointing to a barrier-limited transport phenomenon. However, we did not explore the transport phenomena in detail, as this study is more focused on the extraction of events. The protein hSTf plays an important role in iron homeostasis, its primary function being the transport of iron from the blood into the cells. Voltage-driven protein unfolding during electrophoretic translocation through nanopores has been studied previously.[1,15] The conformation of hSTf and thereby its function is pH-dependent, presenting challenges during both nanopore experiments and data extraction. We coined the name *EventPro* for the program developed in this study, which is downloadable through the research websites of the corresponding authors. Supporting Information Section S1 outlines the graphical user interface of *EventPro* and the settings associated with it.

It is worthwhile to note that much of the focus has been on the multilevel fitting of events, overshadowing the significance of baseline processing. In this paper, we first discuss baseline processing methods tailored for both fluctuating and appreciably stable pore profiles, with the former requiring an adaptive baseline fitting approach that is sensitive to local variations. Such pre-processing allows for the identification of events that would otherwise be subsumed by local baseline variations. This step can thus be identified as the first critical step to determine the proper flagging of events and by extension their final cumulative count. Second, the multilevel fitting of events is discussed. These two steps are discussed separately so that they can be adopted independently of each other depending on the requirements of the user. Since all functions are MATLAB-based, they can be integrated easily into existing workspaces.

## EXPERIMENTAL SECTION

**Nanopore Fabrication, Biomolecule and Electrolyte Preparation, and Data Acquisition.** The h-BN nanopore fabrication processes are outlined in our previous work,[15,26] and the nanopores in $Si_xN_y$ purchased from Norcada (NBPX5001Z-HR, nominally ~12 nm thick) were fabricated using CDB and[27] CT-CDB.[11] dsDNA (10787018, Fisher Scientific) was used as supplied, and the stock solutions of hSTf (T0665, Sigma-Aldrich, USA) were prepared by dissolving the as-supplied solid in >18 MΩ cm ultra-pure water (ARS-102 Aries high-purity water systems). Electrolytes (purchased from Sigma-Aldrich) were prepared by dissolving LiCl (213233) or KCl (P9333) in ultra-pure water. All electrolytes were buffered using 10 mM tris-buffer (J61036, Fisher Scientific), and the pH was adjusted by adding concentrated drops of HCl (H1758, Sigma-Aldrich) and/or KOH (306568, Sigma-Aldrich) and measured using an Orion Star pH meter. All electrical measurements were acquired using an Axopatch 200B (Molecular Devices, LLC), low-pass filtered at either 10 kHz (hSTf) or 100 kHz (dsDNA) using the inbuilt Bessel filter of the Axopatch, sampled at 200 kHz (hSTf) or 250 kHz (dsDNA), and digitized using Digidata 1440A (Molecular Devices, LLC). The basis for oversampling stems from the requirements of reconstructing the analogue signal from digitized samples with minimal distortion. For an ideal aliasing filter, this requires 2× the bandwidth. Since it is not possible experimentally, for time-domain analysis, it is recommended to sample at least 5× the bandwidth with at least 10× being considered good. However, in most cases, we have resorted to a 20× sampling rate, which is not uncommon in the nanopore community. For 100 kHz low-pass filtering, we have used the maximum sampling rate allowed by the


https://doi.org/10.1021/acs.analchem.1c01646
Anal. Chem. 2021, 93, 11710−11718

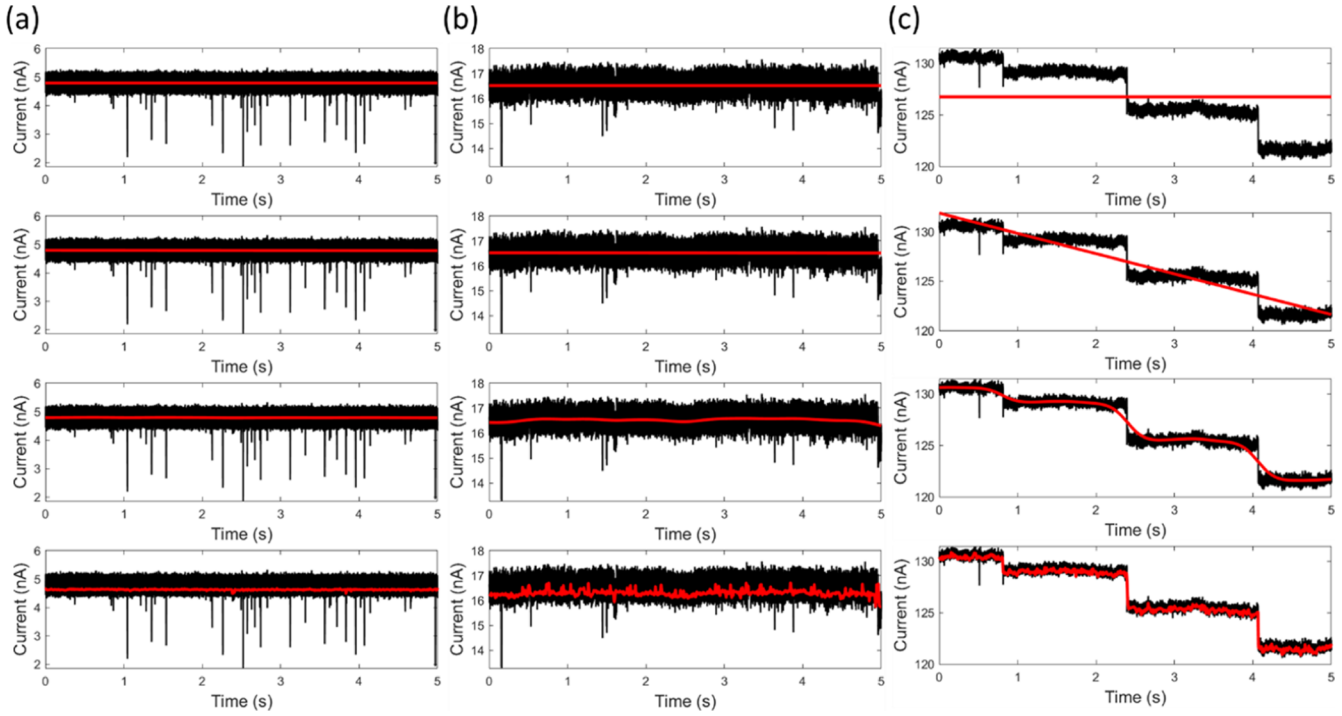



**Figure 1.** Representative 5 s traces corresponding to (a) dsDNA translocating through the $Si_xN_y$ nanopore fabricated by CT-CDB, (b) dsDNA translocating through the $Si_xN_y$ nanopore fabricated by CDB, and (c) *holo*-hSTf translocating through the h-BN nanopore fabricated by a transmission electron microscope (TEM). For h-BN, we deliberately chose an inferior trace to test the performance of each fitting approach. Each open-pore trace was fitted using the arithmetic mean (first row), linear fit (second row), Gaussian smoothing (third row), and Gaussian smoothing and regressed mixing (last row) approaches discussed in the main text. A 5 s analysis window (rather than 500 ms as outlined in the main text) was chosen to highlight the performance of each fitting approach.

digitizer (*i.e.*, 250 kHz). For hSTf, it was not possible to use the 100 kHz low-pass filter setting due to insufficient signal-to-noise ratio. It should be noted that faster translocations require operating at higher bandwidths, which comes at the cost of higher baseline noise. Although the use of a lower low-pass filter setting permits circumventing the noise issues to some extent, it comes at the expense of signal attenuation, as outlined in Supporting Information Section S4.

## RESULTS AND DISCUSSION

**Baseline Fitting.** We discuss four baseline fitting strategies along with their (i) pros and cons, and (ii) suitability for a given nanopore platform. We used baseline fitting strictly to detect events, whereas further analysis to identify levels was done using raw data rather than the baseline-corrected data because the latter could change the event structures. The open-pore current ($I_0$) trace is first segmented into windows of identical length defined by the user (typical length is about 500 ms with the choice dependent on the average duration of a resistive pulse and the open-pore stability).

*Baseline Fitting Method 1: Arithmetic Mean of $I_0$.* As the name suggests, in this method, the arithmetic mean of the open-pore current trace ($\mu_{I_0}$) in the analysis window is computed (Figure 1a–c, first row). This is the fastest method computationally and is most suitable for nanopores where the baseline does not appreciably change with time (*e.g.*, CT-CDB nanopores). However, this method could provide a false estimation of $\mu_{I_0}$ for nanopores with vigorous open-pore current fluctuations (Figure 1c, top row).

*Baseline Fitting Method 2: Linear Fit to $I_0$.* This method is more suitable for pores where the open-pore current gradually increases (or decreases) with time partly due to pore enlargement or other reasons. The open-pore current trace in the analysis window is fitted with a first-degree polynomial to construct a fit for the $I_0$ profile (Figure 1a–c, second row). The same method can also be used for nanopores with stable open-pore currents that have slight fluctuations with time for better estimation of the baseline. This method is the second-fastest method computationally (typically ∼10% slower than type 1).

*Baseline Fitting Method 3: Gaussian Smoothened $I_0$.* A fit to the open-pore current trace in the analysis window is developed by smoothing it using a Gaussian filter with the inbuilt *smoothdata* function of MATLAB ($I_{0,\mathrm{smoothdata}}$, Figure 1a–c, third row). This is the third-fastest method computationally (typically ∼5× slower than type 1) and is suited for pores that have fluctuating baselines (*i.e.*, random fluctuations within the analysis window). The method is more adaptive to baseline fluctuations, unlike the first two methods. Its use for traces with step changes should be done with utmost caution, as it could lead to poor estimation near the edges of the steps, as seen in Figure 1c (third row).

*Baseline Fitting Method 4: Gaussian Smoothing and Regressed Mixing of $I_0$.* Using the $I_{0,\mathrm{smoothdata}}$ developed under *baseline fitting method 3*, a secondary fit line is developed by gradually increasing the values of the fit line by modifying it with $n'_{\sigma,\mathrm{thresh}}\cdot\sigma_{\mathrm{trace}}$ (*i.e.*, $I_{0,\mathrm{smoothdata}}^{\mathrm{secondary,upper}} = I_{0,\mathrm{smoothdata}} + n'_{\sigma,\mathrm{thresh}}\cdot\sigma_{\mathrm{trace}}$ where $n'_{\sigma,\mathrm{thresh}}$ is a numerical variable) until ∼99.5% of the trace is $< I_{0,\mathrm{smoothdata}}^{\mathrm{secondary}}$ (the upper boundary of the current trace). The lower boundary is then defined as $I_{0,\mathrm{smoothdata}}^{\mathrm{secondary,lower}} = I_{0,\mathrm{smoothdata}} - n'_{\sigma,\mathrm{thresh}}\cdot\sigma_{\mathrm{trace}}$. Thus, the baseline can be defined as current points that are within the $I_{0,\mathrm{smoothdata}} \pm n'_{\sigma,\mathrm{thresh}}\cdot\sigma_{\mathrm{trace}}$ range. A fit


https://doi.org/10.1021/acs.analchem.1c01646
Anal. Chem. 2021, 93, 11710−11718

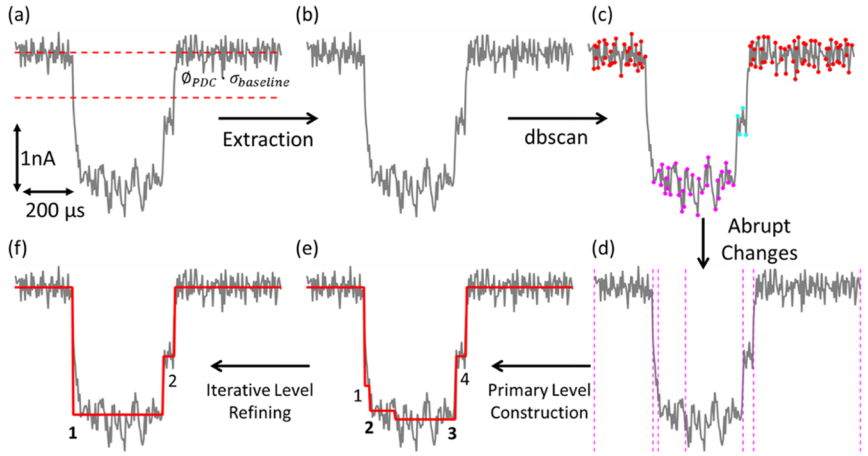



**Figure 2.** Flow diagram of the multilevel fitting approach. In brevity, (a,b) baseline fitting is first performed using one of the four methods described in Figure 1 followed by extraction of events. In this instance, events are defined as perturbations that are at least $\varphi_{PDC}\cdot\sigma_{baseline}$ deep. (c) Turning points (*i.e.*, peak and valley points) of the padded event are then used as the data points for *dbscan* clustering. (d) The clusters ($n_{clusters}$) found in panel (c) are then used as the seeding points to find abrupt changes in the signal level. This step tends to overestimate levels than what is truly present since we use $n_{clusters}$ + 2 as the seed value (reasons outlined in the main text). The vertical magenta dashed lines bracket each level. (e) A weighted mean approach (see Supporting Information Section S2) is then used to calculate the current value for each level. (f) The events are iteratively checked against their adjacent levels to see if they are sufficiently apart compared to a user-defined threshold (*i.e.*, to see if they should be merged or kept as individual levels).

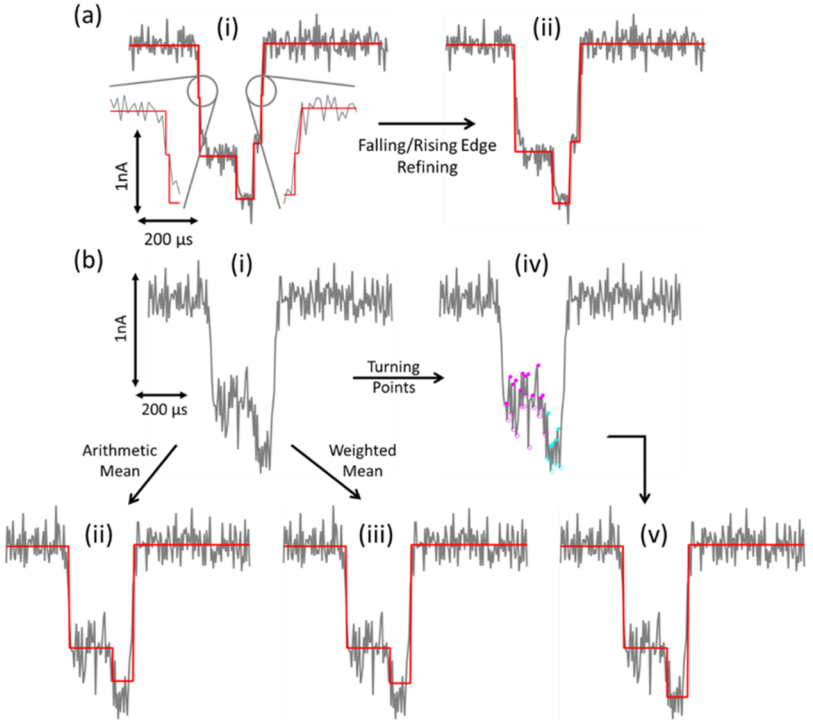



**Figure 3.** (a) Falling and rising edge refining method where (i) falsely identified levels due to subtle current fluctuations in the falling/rising edges (zoomed-in on either end of the event) are removed through (ii) Gaussian smoothing. (b) (i) The current value of each level ($\Delta I_{level}$) for a given event could be calculated either with (ii) arithmetic or (iii) weighted mean of the data points. However, since this could sometimes underestimate the $\Delta I_{level}$, we first find (iv) turning points in each level followed by (v) structure- and length-dependent mean calculation, as outlined in the main text and Supporting Information Section S2.

to this baseline is then developed using the inbuilt *msbackadj* function. This type of baseline development is computationally expensive (typically >10× slower than type 1), yet it offers capabilities to successfully develop a fit line for vigorously fluctuating baseline profiles and even for those with step changes (Figure 1c, last row). However, this should only be used if neither of the previous baseline types can be used because of the computational resources it requires.

**Multilevel Fitting and Event Characterization.** After the identification and fitting of the baseline using one of the abovementioned four methods, events are flagged as current perturbations that have deviated at least (i) $\varphi_{PDC}\cdot\sigma_{baseline}$ from the baseline fit, where $\varphi_{PDC}$ is the peak detection coefficient


https://doi.org/10.1021/acs.analchem.1c01646
Anal. Chem. 2021, 93, 11710−11718

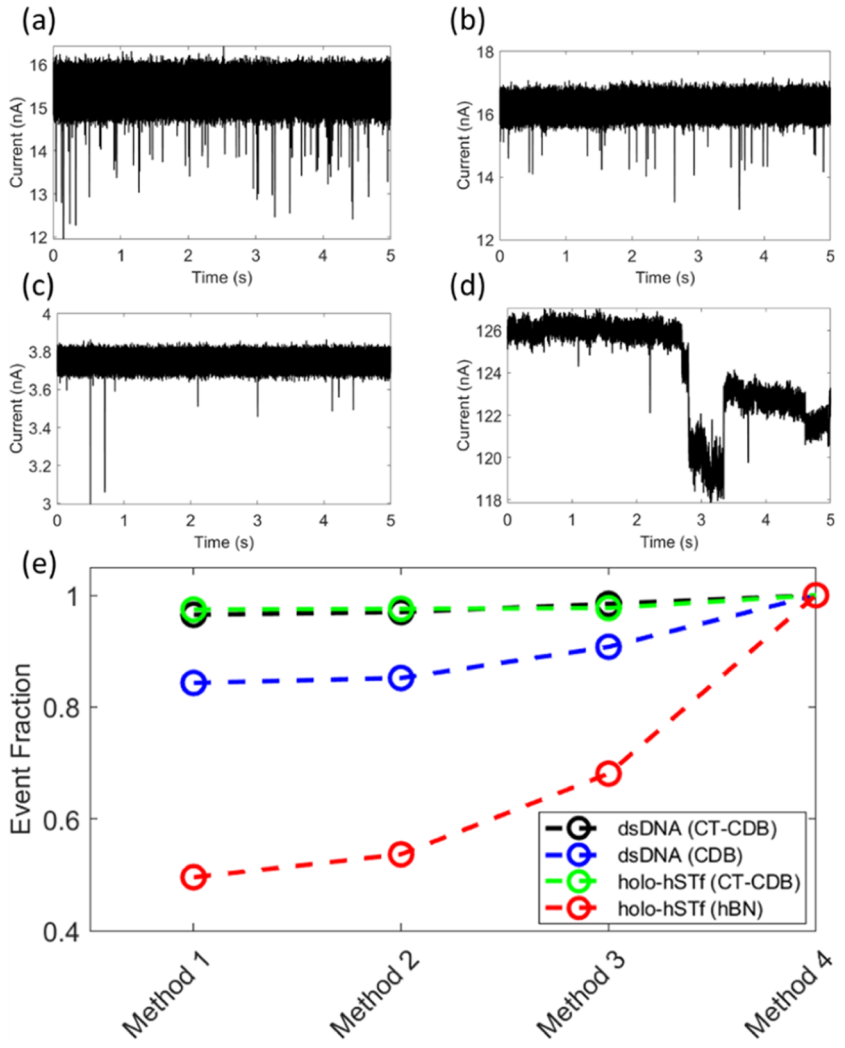



**Figure 4.** 5 s event traces corresponding to (a) dsDNA translocating through a CT-CDB-fabricated $Si_xN_y$ pore, (b) dsDNA translocating through a CDB-fabricated $Si_xN_y$ pore, (c) *holo*-hSTf translocating through a CT-CDB-fabricated $Si_xN_y$ pore, and (d) *holo*-hSTf translocating through a TEM-fabricated h-BN pore (extended traces are shown in Figures S4 and S6). (e) Event counts from each of the four baseline fitting methods normalized to the counts from the *Gaussian smoothing and regressed mixing*, $I_0$, (method 4) fitting approach. All pores used in this comparison are ∼9−10 nm in diameter.

and $\sigma_{baseline}$ is the standard deviation of the baseline in the analysis window, or (ii) $\Delta I_{PDC}$, where $\Delta I_{PDC}$ is a user-defined current threshold. Most elementary analysis (*i.e.*, simple analysis) metrics of an event include the maximum depth ($\Delta I_{max}$) and the duration ($\Delta t$). However, these metrics are highly susceptible to point variations and do not yield insights into multilevel features of an event—commonly seen with long polymeric molecules such as DNA. Steps 2 and 4 of the Multilevel Fitting and Event Characterization section highlight measures taken to safeguard against such subtle variations. The multilevel fitting flow of *EventPro* is shown in Figure 2.

*Step 1: Identification of Levels and Baseline Points.* The events are padded by adding $n_{baseline}$ points (typically 50) to either side of the event (*i.e.*, padded event). Then, the peak and valley points (*i.e.*, turning points) of the padded event are identified. These points provide a preliminary collection of selected points in each level for robust clustering (Figure 2c). Without this step, the points in the padded event would be too closely populated for adequate clustering—adjacent levels would be evaluated as a single cluster rather than separate clusters. Afterward, these points are clustered using the *dbscan* method with the threshold for a neighborhood search radius set to $\sigma_{baseline}$. Our choice for *dbscan* stems from the fact that levels could be arbitrarily shaped and the knowledge of the number of levels in a padded event is typically unknown. Abrupt changes in the padded event are then deduced using the inbuilt *findchangepts* function (Figure 2d) with the maximum number of abrupt changes defined using the cluster count ($n_{clusters}$ + 2). The "+2" is to safeguard against poor clustering of events where $n_{clusters}$ is <2, which does not permit the *findchangepts* function to identify event boundaries properly. The newly found abrupt change points allow for proper differentiation of the event region from the baseline region and, by extension, levels within the event. This newly defined event region could be different from the previously defined event region using the simple analysis because the latter is susceptible to point variation and is prone to failure with structures that have falling/rising edges that convolve gradually with the baseline region. To safeguard against oversegmentation of the padded event, the algorithm looks at the mean current level difference of the nearest neighboring levels. If they are within a user-defined threshold (either a user-


https://doi.org/10.1021/acs.analchem.1c01646
Anal. Chem. 2021, 93, 11710−11718

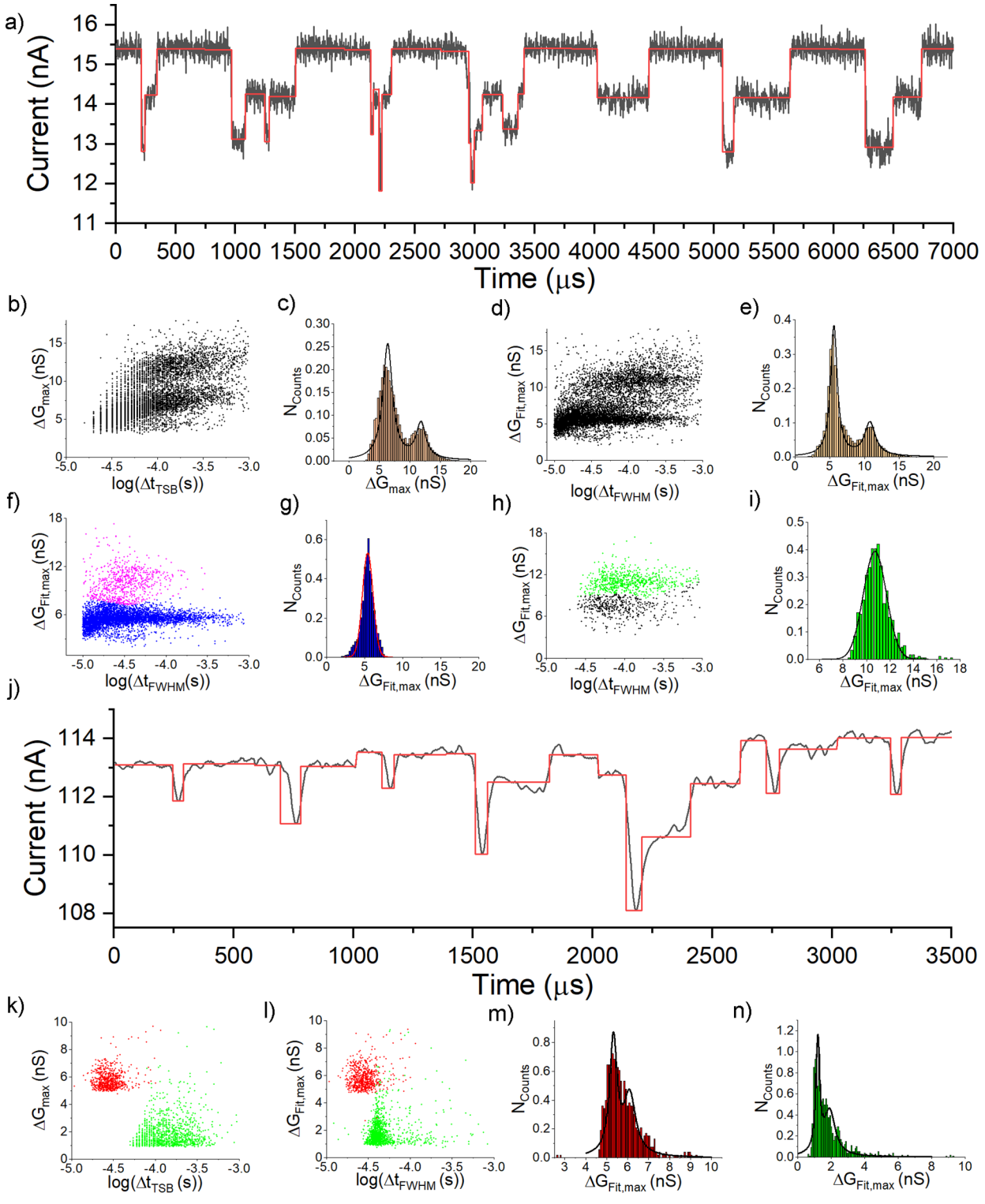



**Figure 5.** (a) Concatenated events corresponding to dsDNA translocations (through a CT-CDB-fabricated $Si_xN_y$ pore) with black and red traces corresponding to raw and fit data, respectively. (b) Scatter plot of $\Delta G_{max}$ *vs* $\log(\Delta t_{TSB})$ and (c) histogram corresponding to $\Delta G_{max}$ from simple event analysis. (d) Scatter plot of the deepest current level from the multilevel fitting approach ($\Delta G_{fit,max}$) *vs* $\log(\Delta t_{FWHM})$ and (e) histogram corresponding to $\Delta G_{fit,max}$. (f) Scatter plot of single-level events clustered into two groups using Gaussian mixture clustering with blue and magenta scattered populations corresponding to single-file and looped translocation conformations, respectively. (g) Histogram of $\Delta G_{fit,max}$ corresponding to single-file translocations from panel (f). (h) Scatter plot of dual-level events clustered into two groups using Gaussian mixture clustering with green and black scattered populations corresponding to deep and shallow dual-level translocations, respectively. (i) Histogram of $\Delta G_{fit,max}$ corresponding to deep dual-level translocations from panel (h). (j) Concatenated events corresponding to hSTf translocations (though h-BN at +800 mV) with black and red traces corresponding to raw and fit data, respectively. Scatter plots corresponding to (k) $\Delta G_{max}$ *vs* $\log(\Delta t_{TSB})$ and (l) $\Delta G_{fit,max}$ *vs* $\log(\Delta t_{FWHM})$ with red and green scattered populations corresponding to +100 and +800 mV, respectively. Histograms of $\Delta G_{fit,max}$ at (m) +100 and (n) +800 mV.



https://doi.org/10.1021/acs.analchem.1c01646
*Anal. Chem.* 2021, 93, 11710−11718

defined current, $\Delta I_{user}$, or as $n{\cdot}\sigma_{baseline}$, where $n$ is typically set to 3), the two levels are merged (Figure 2e,f). This is done iteratively until there are no levels to be merged further.

*Step 2: Refining Rising and Falling Edges.* The first and last levels of the event tend to be affected by the structure of the rising and falling edges of the event, respectively. That is, if these edges are not smooth (have point fluctuations), such fluctuations could be misidentified as separate levels (Figure 3a). This would lead to a false overestimation of the level count of the event. Thus, we used a Gaussian smoothing step for the rising and falling edges, and if any peaks or valleys are present in the smoothened edges, assigning levels to the same is permitted and eliminated otherwise (Figure 3a).

*Step 3: Calculation of the Current Value of Each Event Level.* While the more straightforward method is to use the mean of the data points in each level to calculate the representative current value of the level, this method is more suitable for long levels and for levels that do not appreciably capture the edges leading to adjacent levels. However, for short events, this could drastically underestimate the current level. Thus, to eliminate these subtleties, *EventPro* looks at the peak and valley points of each level. These define the upper and lower boundaries of a given level, and the mean of these essentially provides an estimation that is less susceptible to the presence of portions of adjacent edges ($\mu_{level,turning\ points}$), as seen in Figure 3b. The assignment of the current value for a given level ($\Delta I_{level}$) is discussed in Supporting Information Section S2 that takes level length and its structure (*i.e.*, upward, downward, or leveled) into account. Events are then qualified as acceptable or poor based on the standard deviation of the longest level and existence of points deviating greater than the threshold used to iteratively refine levels (see Supporting Information Section S3 for more details).

*Step 4: Translocation Time Calculation.* The more commonly used methods for $\Delta t$ estimation include full width at half maximum ($\Delta t_{FWHM}$), modified stop point,[28,29] and two sides of the event ($\Delta t_{TSB}$). The $\Delta t_{TSB}$ approach has been shown to drastically overestimate $\Delta t$, whereas $\Delta t_{FWHM}$ and modified stop point methods yield values in close agreement with expected $\Delta t$.[28] *EventPro* uses the $\Delta t_{FWHM}$ method. If $\Delta t < 2T_r$, where $T_r \cong 0.3321/f_c$ with $f_c$ being the cutoff frequency of the filter, the event would be attenuated due to the finite response time of the filter. For example, if $f_c$ = 10 kHz (the commonly used filter setting), $2T_r$ would be ∼66 $\mu$s.[30] To account for the finite spacing of points in the falling and rising edges of the event, 10−90% of the step height of each edge is fitted with a linear function for $\Delta t_{FWHM}$ calculation to be independent of point spacings and subtle variations in the edges. To validate this approach, we used a function generator to simulate events with a known pulse width ($\Delta t_{input}$) and compared the calculated $\Delta t_{FWHM}$ from the linear fit approach with $\Delta t_{input}$, as discussed in Supporting Information Section S4. We observed that $\Delta t_{FWHM}$ was in good agreement with $\Delta t_{input}$ for pulses longer than 40 $\mu$s, as seen in Figure S3. Thus, one must be mindful of the erroneous (over) estimation of $\Delta t$ for pulses shorter than 40 $\mu$s with the 10 kHz low-pass filter setting.

**Implementation of EventPro with Biomolecules.** In this step, we tested the performance of *EventPro* with dsDNA and protein (hSTf). Figure 4 shows representative current traces and the normalized event count with each baseline fitting type for the conditions outlined in the caption. From Figure 4, it is apparent that with deteriorating open-pore current quality, the fraction of events extracted decreases drastically from method 4 to method 1. For example, for hSTf translocating through h-BN nanopore under +800 mV, only ∼0.5× events were extracted using method 1 compared to method 4, whereas for dsDNA translocating through a CT-CDB-fabricated nanopore, the fraction obtained using method 1 improves to ∼0.95× compared to that in method 4. One should be mindful of the fact that method 4 is >10× slower than method 1, and thus, a tradeoff between the event count and the computational economy should be considered when choosing a method for baseline fitting.

*Data Analysis.* While it is possible to perform a lengthy analysis for all the analytes and pore fabrication conditions provided herein, for brevity, we restrict further analysis to the two extreme cases: dsDNA translocation through the CT-CDB nanopore (ultra-stable open-pore trace, $I_{std}$ = 130 pA, where $I_{std}$ is the standard deviation of the open-pore current computed from a ∼500 ms section) and hSTf translocation through h-BN nanopore (most vigorously fluctuating trace, $I_{std}$ = 320 pA). Further representations of noise and baseline fluctuations are shown in Supporting Information Section S6. Resistive pulses resulting from the translocation of dsDNA through a ∼10.8 nm-diameter $Si_xN_y$ nanopore were extracted by first fitting the baseline using the *arithmetic mean of* $I_0$ method and then setting $\varphi_{PDC}$ to 5. The events were then fitted using the multilevel fitting approach (Figure 5a). The scatter plot corresponding to the maximum change in conductance due to dsDNA translocation ($\Delta G_{max}$) versus log($\Delta t_{TSB}$) is shown in Figure 5b (*i.e.*, simple analysis), while the deepest current blockade of the fit levels ($\Delta G_{fit,max}$) versus log-($\Delta t_{FWHM}$) is shown in Figure 5d. The histograms corresponding to $\Delta G_{max}$ and $\Delta G_{fit,max}$ are shown in Figure 5c,e, respectively. Both the histograms were then fitted with a Lorentzian mixture model (see the *Histogram Fitting* in Supporting Information Section S7 for more details). From the histograms, it is evident that the simple analysis yields a wider distribution of $\Delta G$ compared to the multilevel fitting approach. This is not surprising since the simple analysis is susceptible to point variations and would invariably result in a wider distribution. The multilevel analysis permitted us to separate single-level (Figure 5f) and dual-level (Figure 5h) translocations from other multilevel translocations. Single-level translocations can be either single-file or looped translocations with the latter typically producing 2× deeper blockades compared to the former. Thus, to separate each of these from the scatter plot shown in Figure 5f, we performed a simple Gaussian mixture clustering (see the *Clustering of Scatter Plots* in Supporting Information Section S7), whereby the single-file translocations (blue) were separable from the looped translocations (magenta). Fitting of the histogram corresponding to single-file translocations (Figure 5g) yielded a mean value of ∼5.4 nS, which is in good agreement with that from the fit of the histogram of the total population (*i.e.*, Figure 5e, first peak). This clustering approach revealed that ∼69% of translocations were single-level type with 60% (∼87% of the single-file population) being single file and ∼9% being looped translocations. Similarly, the dual-step translocations were also clustered, as seen in Figure 5g. As seen in Figure S10, there are shallow dual-level events that have been observed previously.[19] Fitting of the histogram corresponding to deep dual-level translocations (Figure 5i) yielded a mean value of ∼10.7 nS, which is in good agreement with that from the fit of the histogram of the total population (*i.e.*, Figure 5e, second peak).


https://doi.org/10.1021/acs.analchem.1c01646
Anal. Chem. 2021, 93, 11710−11718

Furthermore, only ∼15% of translocations were dual-level with ∼11% (∼75% of the dual-level population) being deep blockades and 4% being shallow dual-level blockades. The change in open-pore conductance due to dsDNA through a ∼10.8 nm-diameter pore was estimated to be between 4.95 and 6.9 nS (see the *Modeling DNA Translocation* under Supporting Information Section S7 for more details). This was found to be in close agreement with the peak values from each of the analyses corresponding to single-file DNA translocations: ∼6.4 and ∼5.6 nS from simple and multilevel analysis, respectively.

Finally, the more vigorously fluctuating current trace—hSTf through an h-BN nanopore—was analyzed. For this, we used the *Gaussian smoothing and regressed mixing of* $I_0$ baseline fitting method (raw traces are shown in Supporting Information Figure S6) with the corresponding scatter plots and histograms shown in Figure 5k−n. While the simple analysis yields a broad scatter plot distribution (Figure 5k), the multilevel fitting approach with the $\log(\Delta t_{FWHM})$ approach yields a narrow distribution (Figure 5l). Histograms corresponding to $\Delta G_{fit,max}$ at +100 and +800 mV are shown in Figure 5m,n, respectively. It is interesting to note that with increasing applied voltage, the translocation time slows down and the change in conductance decreases. We believe that as shown in our previous work, the voltage-driven protein unfolding is the reason for these observations, where the increase in molecular length due to voltage-driven unfolding could outweigh the increasing velocity with increasing voltage with the former (i.e., unfolding) elongating the molecule (slowing down the translocation), while the latter (i.e., increasing voltage) would increase the translocation speed.[1,7] Furthermore, the voltage-driven unfolding would reduce the molecular volume and, by extension, the magnitude of the blockade depth. Each of the $\Delta G_{fit,max}$ distributions shown in Figure 5m,n was fitted with a Lorentzian−Lorentzian mixture distribution. Although the bimodal distribution is not as well separated as dsDNA, it could correspond to unfolded (lower $\Delta G$) and folded (high $\Delta G$) conformations of the protein, as noted previously.[1,7,15]

## CONCLUSIONS

We present four baseline fitting methods, *viz.*, arithmetic mean, linear fit, Gaussian smoothing, and Gaussian smoothing and regressed mixing, each with increasing robustness and computational cost, that can be implemented across a wide spectrum of open-pore current traces. The baseline fitting must be chosen depending on the behavior of the open-pore trace: ultra-stable profiles can be fitted with the arithmetic mean method (*e.g.*, CT-CDB-fabricated nanopores), while more rigorously fluctuating baselines require the more adaptive yet computationally expensive Gaussian smoothing and regressed mixing approach. Proper identification of the baseline is imperative for the effective identification of events. For example, when the open-pore current fluctuates with time (*i.e.*, noisy open-pore traces), the baseline fitting method must be adaptive to such fluctuations; otherwise, it could over/underestimate the open-pore current and, by extension, fail to recognize events. We note that identification of the maximum number of events may come at the expense of analysis efficiency, especially in the case of fluctuating open-pore profiles (more time needed to complete the analysis), yet it compiles a better representation of the translocation population. Afterward, multilevel fitting was performed, which startsed with clustering of turning points of the events using the *dbscan* method. Subsequently, abrupt changes in the signal levels were found, followed by iterative refining of the event levels based on the user-defined step height threshold. The current representation of each level was calculated using a combination of the mean of turning points and weighted mean, which takes the structure (*i.e.*, upward, downward, or connecting) and length of events into account. The rising and falling edges of the event were refined, as these were found to be susceptible to subtle current fluctuations, leading to false generation of levels. Events were then qualified as acceptable or poor based on the standard deviation of the longest level and existence of points deviating greater than the threshold used to iteratively refine levels. The performance of the program was tested using data collected from the translocation of dsDNA and hSTf through nanopores fabricated by CDB, CT-CDB, and TEM through $Si_xN_y$ and h-BN membranes. Finally, the translocation characteristics of the two extreme cases were examined: dsDNA through CT-CDB nanopores (ultra-stable open-pore current) and hSTf through h-BN nanopores (vigorously fluctuating open-pore trace). Clustering was used to separate looped events from single-file translocations and shallow dual-level events from deep dual-level events in dsDNA. In a nutshell, the proposed analysis platform carries advantages such as iterative level refining to safeguard against subtle current fluctuations that lead to false levels, calculation of the level mean based on turning points to counter shallow falling/rising edges, the full width at the half maximum approach to calculate the event duration instead of using the two sides of the blockade method, refining of falling and rising edges of the events to prevent false event classifications, especially in the case of shallow and/or noisy drops, and classification of events as acceptable or poor for the ease of the user.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.1c01646.

Graphical user interface of *EventPro* and input parameters; calculation of $\Delta I_{level}$; qualifying of events; calculation of translocation time; event profiles; noise comparison; and fitting of DNA data (PDF)

## AUTHOR INFORMATION


### Corresponding Authors

**Patrick Kluth** − *Department of Electronic Materials Engineering, Research School of Physics, Australian National University, Canberra, Australian Capital Territory 2601, Australia*; Email: patrick.kluth@anu.edu.au

**Min Jun Kim** − *Department of Mechanical Engineering, Southern Methodist University, Dallas, Texas 75275, United States*; orcid.org/0000-0002-0819-1644; Email: mjkim@lyle.smu.edu

### Authors

**Y. M. Nuwan D. Y. Bandara** − *Department of Electronic Materials Engineering, Research School of Physics, Australian National University, Canberra, Australian Capital Territory 2601, Australia*; orcid.org/0000-0003-1921-8467

**Jugal Saharia** − *Department of Mechanical Engineering, Southern Methodist University, Dallas, Texas 75275, United States*

**Buddini I. Karawdeniya** − *Department of Electronic Materials Engineering, Research School of Physics, Australian National University, Canberra, Australian Capital Territory 2601, Australia*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.1c01646

### Notes

The authors declare no competing financial interest.
Y.M.N.D.Y.B. is currently employed at the University of California, Riverside.

## ■ ACKNOWLEDGMENTS

Experimental work was supported by the National Science Foundation (CBET#2041340 and #2022374) and the National Institutes of Health (R21CA240220). B.I.K. and P.K. were supported by Our Health in Our Hands ANU Grand Challenge. P.K. acknowledges the Australian Research Council for financial support. The authors would like to thank their respective institutions for the MATLAB and Mathematica licenses provided for this work. We thank Matthew ODonohue for his diligent proofreading of this manuscript.

## ■ REFERENCES

(1) Saharia, J.; Bandara, Y. M. N. D. Y.; Goyal, G.; Lee, J. S.; Karawdeniya, B. I.; Kim, M. J. *ACS Nano* **2019**, *13*, 4246−4254.
(2) Karawdeniya, B. I.; Bandara, Y. M. N. D. Y.; Nichols, J. W.; Chevalier, R. B.; Dwyer, J. R. *Nat. Commun.* **2018**, *9*, 3278.
(3) Farimani, A. B.; Min, K.; Aluru, N. R. *ACS Nano* **2014**, *8*, 7914−7922.
(4) Darvish, A.; Lee, J. S.; Peng, B.; Saharia, J.; VenkatKalyana Sundaram, R.; Goyal, G.; Bandara, N.; Ahn, C. W.; Kim, J.; Dutta, P.; Chaiken, I.; Kim, M. J. *Electrophoresis* **2019**, *40*, 776−783.
(5) McMullen, A.; De Haan, H. W.; Tang, J. X.; Stein, D. *Nat. Commun.* **2014**, *5*, 4171.
(6) Goyal, G.; Freedman, K. J.; Kim, M. J. *Anal. Chem.* **2013**, *85*, 8180−8187.
(7) Bandara, Y. M. N. D. Y.; Tang, J.; Saharia, J.; Rogowski, L. W.; Ahn, C. W.; Kim, M. J. *Anal. Chem.* **2019**, *91*, 13665−13674.
(8) Robertson, J. W. F.; Rodrigues, C. G.; Stanford, V. M.; Rubinson, K. A.; Krasilnikov, O. V.; Kasianowicz, J. J. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 8207−8211.
(9) Charron, M.; Briggs, K.; King, S.; Waugh, M.; Tabard-Cossa, V. *Anal. Chem.* **2019**, *91*, 12228−12237.
(10) Briggs, K.; Madejski, G.; Magill, M.; Kastritis, K.; de Haan, H. W.; McGrath, J. L.; Tabard-Cossa, V. *Nano Lett.* **2018**, *18*, 660−668.
(11) Bandara, Y. M. N. D. Y.; Saharia, J.; Karawdeniya, B. I.; Hagan, J. T.; Dwyer, J. R.; Kim, M. J. *Nanotechnology* **2020**, *31*, 335707.
(12) Saharia, J.; Bandara, Y. M. N. D. Y.; Karawdeniya, B. I.; Alexandrakis, G.; Kim, M. J. *Electrophoresis* **2021**, *42*, 899−909.
(13) Plesa, C.; Dekker, C. *Nanotechnology* **2015**, *26*, 084003.
(14) Heerema, S. J.; Schneider, G. F.; Rozemuller, M.; Vicarelli, L.; Zandbergen, H. W.; Dekker, C. *Nanotechnology* **2015**, *26*, 074001.
(15) Saharia, J.; Bandara, Y. M. N. D. Y.; Lee, J. S.; Wang, Q.; Kim, M. J.; Kim, M. J. *Electrophoresis* **2020**, *41*, 630−637.
(16) Gu, C.; Yu, Z.; Li, X.; Zhu, X.; Cao, Z.; Ye, Z.; Jin, C.; Liu, Y. *Appl. Phys. Lett.* **2019**, *115*, 223702.
(17) Tabard-Cossa, V.; Trivedi, D.; Wiggin, M.; Jetha, N. N.; Marziali, A. *Nanotechnology* **2007**, *18*, 305505.
(18) Liang, S.; Xiang, F.; Tang, Z.; Nouri, R.; He, X.; Dong, M.; Guan, W. *Nanotechnol. Precis. Eng.* **2020**, *3*, 9−17.
(19) Raillon, C.; Granjon, P.; Graf, M.; Steinbock, L. J.; Radenovic, A. *Nanoscale* **2012**, *4*, 4916−4924.
(20) Liu, K.; Pan, C.; Kuhn, A.; Nievergelt, A. P.; Fantner, G. E.; Milenkovic, O.; Radenovic, A. *Nat. Commun.* **2019**, *10*, 3.
(21) Forstater, J. H.; Briggs, K.; Robertson, J. W. F.; Ettedgui, J.; Marie-Rose, O.; Vaz, C.; Kasianowicz, J. J.; Tabard-Cossa, V.; Balijepalli, A. *Anal. Chem.* **2016**, *88*, 11900−11907.
(22) Zhang, J.; Liu, X.; Ying, Y.-L.; Gu, Z.; Meng, F.-N.; Long, Y.-T. *Nanoscale* **2017**, *9*, 3458−3465.
(23) Zhang, J.-H.; Liu, X.-L.; Hu, Z.-L.; Ying, Y.-L.; Long, Y.-T. *Chem. Commun.* **2017**, *53*, 10176−10179.
(24) Sharma, R. K.; Agrawal, I.; Dai, L.; Doyle, P. S.; Garaj, S. *Nat. Commun.* **2019**, *10*, 4473.
(25) Plesa, C.; Verschueren, D.; Pud, S.; van der Torre, J.; Ruitenberg, J. W.; Witteveen, M. J.; Jonsson, M. P.; Grosberg, A. Y.; Rabin, Y.; Dekker, C. *Nat. Nanotechnol.* **2016**, *11*, 1093.
(26) Lee, J. S.; Oviedo, J. P.; Bandara, Y. M. N. D.; Peng, X.; Xia, L.; Wang, Q.; Garcia, K.; Wang, J.; Kim, M. J.; Kim, M. J. *Electrophoresis* **2021**, *42*, 991−1002.
(27) Kwok, H.; Briggs, K.; Tabard-Cossa, V. *PLoS One* **2014**, *9*, No. e92880.
(28) Pedone, D.; Firnkes, M.; Rant, U. *Anal. Chem.* **2009**, *81*, 9689−9694.
(29) Gu, Z.; Ying, Y.-L.; Cao, C.; He, P.; Long, Y.-T. *Anal. Chem.* **2015**, *87*, 907−913.
(30) Plesa, C.; Kowalczyk, S. W.; Zinsmeester, R.; Grosberg, A. Y.; Rabin, Y.; Dekker, C. *Nano Lett.* **2013**, *13*, 658−663.