Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d5dd00060b

Nano Trees: nanopore signal processing and sublevel fitting using decision trees†

Deekshant Wadhwa, [a] Philipp Mensing, [b] James Harden,[b] Paula Branco,[a] Vincent Tabard-Cossa [b] and Kyle Briggs *[b]

As the complexity of solid-state nanopore experiments increases, analysis of the resulting electrical signals to determine biomolecular details becomes a challenge. State of the art techniques for this task perform poorly when transient signal characteristics approach the bandwidth limitations of the measurement electronics. In this work, we address this challenge through an algorithm, called Nano Trees, for fitting piecewise constant functions. Nano Trees leverages machine learning algorithms to provide fits to the noisy piecewise constant data that is characteristic of nanopore ionic current signals, producing accurate fits on transients as short as twice the rise time of the measurement system. We demonstrate the performance of our algorithm on several real and synthetic datasets. These findings underscore the generalizability and accuracy of this approach in the regime of fast molecular translocations.

Received 10th February 2025
Accepted 26th May 2025

DOI: 10.1039/d5dd00060b

rsc.li/digitaldiscovery

Introduction

Nanopores are nanometer-scale holes in thin dielectric membranes similar in size to a single molecule of protein or DNA. They operate in a conductive solution by sustaining a steady ionic current under the influence of an applied voltage, which is transiently blocked when a biomolecule translocates the pore. By studying the duration and depth of the blockage, as well as the patterns in the current blockage signature, one can determine the physical properties of that molecule, such as size, shape, orientation, folding, and branching.[1–5] Nanopores are used for various molecular detection applications, including DNA sequencing,[6] detection of biomarkers of disease in clinically relevant biofluids,[7–12] and decoding of digital information stored in molecular carriers.[12,13] However, interpretation of these complex signals is a challenge.

The problem of fitting noisy piecewise constant data is ubiquitous, appearing in many scientific fields aside from nanopore analysis. For example, the same problem appears in the analysis of anomalous network traffic[14,15] and neuronal activity patterns.[16,17] Effective techniques to analyze this type of signal have the potential to be generally useful beyond just nanopore science. Nanopore analysis tasks involve the categorization of signals, for example, to recognize a rare target biomolecule signature from a complex mixture. Solid-state nanopore data often suffers from issues arising from the speed of molecular translocation, causing important signal features to be distorted by the rise time of the measurement electronics.[18] Present analysis methods often struggle to classify and identify molecules due to the fast kinetics of molecular passage compared to available measurement bandwidth and to their associated inability to accurately characterize fast transient signals.[18,19] Currently, the need for nanopore data analysis is served by a varied patchwork of techniques, many of which are specific to a single experimental context, such as basecalling[20–22] or event classification.[23] The lack of a framework that standardizes the general analysis case has led to differences in statistical treatment and makes quantitative comparisons between labs challenging. There is, therefore, a need for a method that can be readily adjusted to work effectively across multiple molecular targets and nanopore types.

While numerous methods have been proposed over the years, only a few can fit an arbitrary number of piecewise constant sublevels away from the baseline within a noisy and bandwidth-limited nanopore signal, and/or categorize events by type through recognition of patterns encoded in the sublevel structure. A commonly used class of methods involves variations on the Cumulative Sum (CUSUM) algorithm.[24–26] This algorithm[18,19,27] assumes that the signal to be analyzed is a piecewise constant signal overlaid with Gaussian-distributed noise (though it can be generalized to other noise distributions)[17] and iteratively applies a modified *t*-test to each new data point to determine the likelihood that the local mean has undergone a step change of known magnitude. Like most

[a]Department of Computer Science, University of Ottawa, Canada
[b]Department of Physics, University of Ottawa, Canada. E-mail: kbriggs@uOttawa.ca

† Electronic supplementary information (ESI) available: The Nano Trees pipeline (ESI Section S1); detailed pseudo-code algorithms for updating current estimates function (ESI Section S2), merging small current steps pass (ESI Section S3), sublevel categorization procedure (ESI Section S4), and splitting sublevels pass (ESI Section S5); glossary of all hyperparameters used by Nano Trees (ESI Section S6); steps for hyperparameter tuning (ESI Section S7); hyperparameter values used by Nano Trees for fitting all datasets presented in this research (ESI Section S8). See DOI: https://doi.org/10.1039/d5dd00060b

nanopore analysis frameworks, CUSUM performs poorly when fitting transients that are short compared to the response time of the measurement system,[19] leading to miscalculations of sublevel duration and blockage depth when transients are faster than 4 times the rise time of the system, or missing them entirely. Data clustering algorithms, like DBSCAN[24] (density-based spatial clustering of applications with noise), have also been used to get an initial guess of the number of sublevels, detect abrupt changes, and then iteratively checked against their adjacent levels to see if they are sufficiently apart compared to a user-defined threshold, otherwise merged.[25]

An alternative to probabilistic sublevel fitting is to use an approximation to the transfer function of the measurement system to extract the underlying signal from the distorted measurement. The Adaptive Time-Series Analysis (ADEPT) algorithm[18,19,28,29] fits a linear sum of exponential step functions over the event to determine the position of individual sublevels using standard nonlinear fitting techniques. This algorithm works under the assumption that the nanopore operates as a simple RC-equivalent circuit and that the rate-limiting factor that dominates the signal distortion is the response time of the measurement electronics, an assumption that breaks down when the signal is subjected to heavy filtering that can result in the time-response of the filter dominating that of the RC response.[18] While quite effective for events with a relatively small number of well-separated sublevels (typically just one or two), the use of nonlinear fitting over many parameters (3 independent parameters for each sublevel) means that the algorithm performs increasingly poorly as events get more complex and often suffers from difficult-to-debug numerical errors that result in rejection of valid events. Moreover, because an estimate of the number of sublevels needed for fitting is

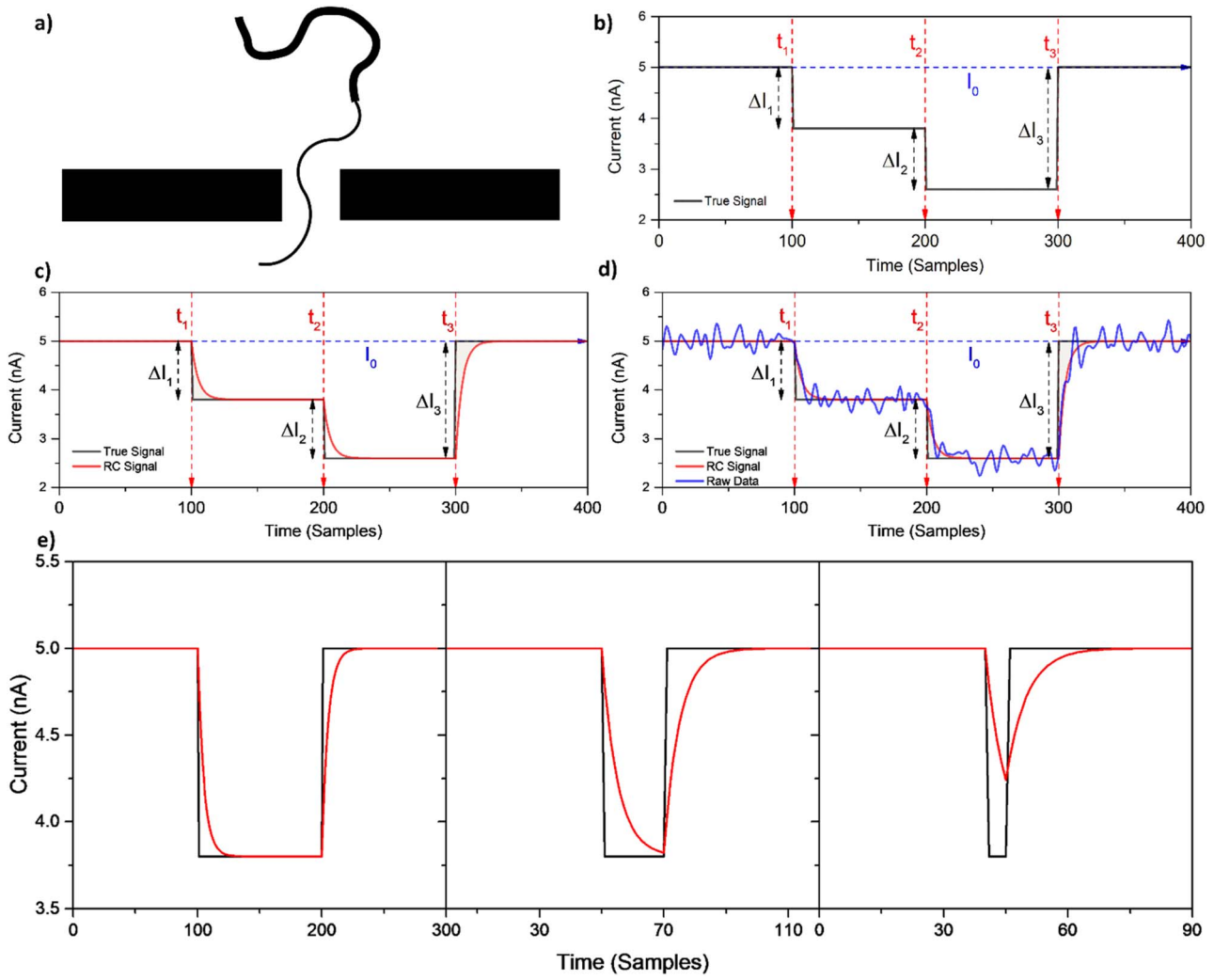

Fig. 1 (a) a schematic diagram of a mock molecule of varying thickness translocating a nanopore from top to bottom. (b) The signal produced by this molecule without noise and with infinite measurement bandwidth. (c) The signal produced by this molecule, considering finite bandwidth electronics, using the prediction from.[19] (d) The signal that would be produced by this molecule overlaid with systemic noise. (e) Examples of the level of distortion of a noiseless step function arising from bandwidth limitations as the duration of the translocation approaches the temporal resolution limits of the measurement system, with rise time set to 5 samples and event durations of 20, 4, and 1 times the rise time, respectively. Adapted with permission.[18]

a required input to the algorithm, the same challenges with respect to heavily distorted sublevels exist. This approach was recently extended using more general functional forms.[6,27,30–32] The approach suggested by Lucas *et al.*[30] is a statistical approach for characterising short-lived events, but it is not clear from published data how well it generalises to multi-step events. Another algorithm is presented by Gu *et al.*[31] uses second order differentials to extract blockage states from single level events, but generalization to multi-step events is not demonstrated. CUSUM+ is widely used as the basis for event fitting in other packages and is chosen as the main point of comparison for the algorithm developed in this work. Other approaches to fitting signals to nanopore data have been developed,[27,30,32] but all suffer from challenges when considering short transients approaching the response time of the measurement electronics.

In this work, we present a framework to improve fitting and characterization of nanopore signals that contain fast transient events, showing excellent fit accuracy down to twice the system response time. We also present a framework for optimizing fitting parameters, which we anticipate will assist with the standardization of statistical analysis of complex nanopore signals across different experimental contexts. Results of this optimized fitting can help with the task of categorizing nanopore events in mixed samples.[33] A representative example of a nanopore signal undergoing a step change as a result of the translocation of two mock molecular states is depicted in Fig. 1. The conceived molecule generating such a two-state signal is illustrated in Fig. 1a, having its diameter increase halfway along its contour length. The true underlying signal that we hope to extract is well-approximated by a piecewise constant function, as can be seen in the schematic in Fig. 1b. However, a truly instantaneous transition between states requires infinite bandwidth, and finite response time and bandwidth limitations imposed by measurement electronics, as well as any low-pass filtering applied, result in distortion of the signal. The case where the rise time is dictated by the RC response of the measurement circuit is shown in Fig. 1c. The system is also subject to several sources of electrical noise that further distort the signal, which are discussed in detail in other studies.[34] An example of the full raw signal (*i.e.*, subject to noise, low-pass filtering, and bandwidth limitations) can be seen in Fig. 1d. A more detailed description of the sources of nanopore noise is available elsewhere.[35]

This distortion becomes especially problematic when the duration of an important feature of the signal approaches the response time of the measurement system or the bandwidth of the recording device, which causes the signal to be attenuated as shown in Fig. 1e, and to vanish entirely into the noise for durations that are shorter than the system rise time.

Consequently, to accurately analyze nanopore data, the approach should be to first denoise the signal (for example, through Bessel filtering or wavelet filtering[36]), then to correct distortions arising from the finite bandwidth of the system, and finally to evaluate the physical validity of the extracted sublevel structure. With the goal of enabling such an accurate decoding of sublevels all the way down to the bandwidth- and hardware-imposed limitations of nanopore measurement, we present here a method of decoding and fitting sublevels to nanopore

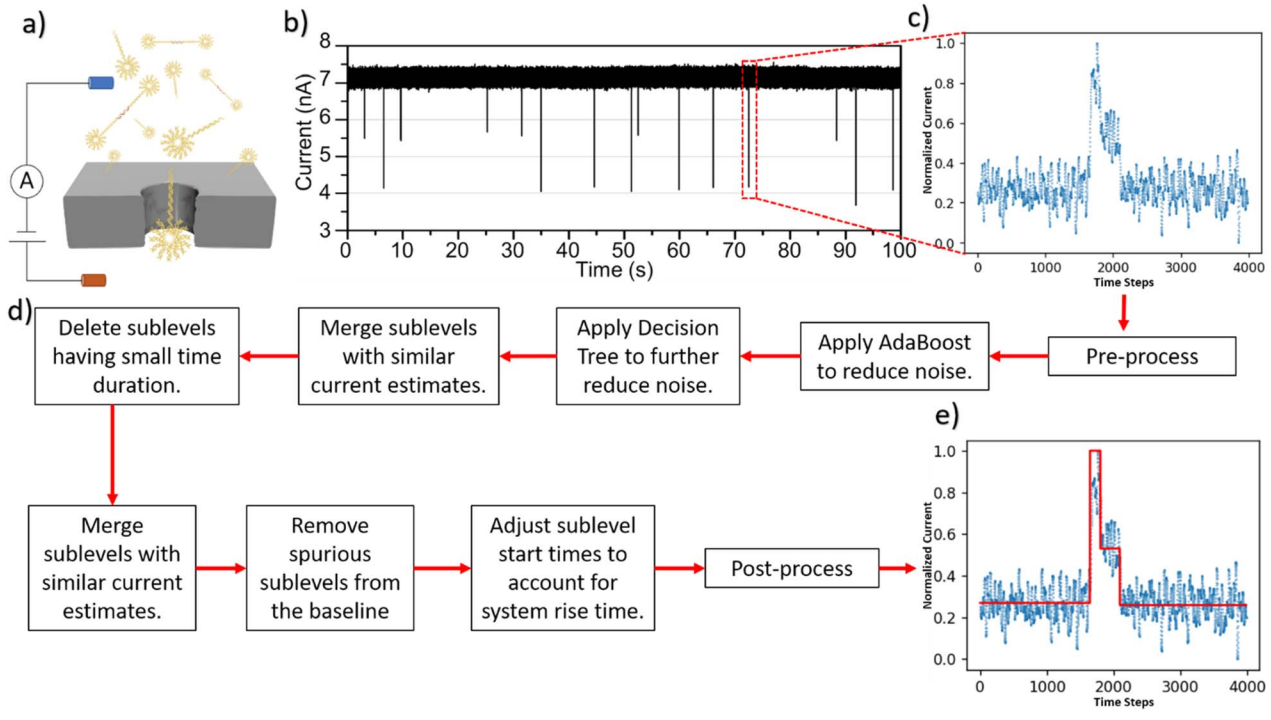

Fig. 2 (a) A representation of a nanopore and a synthetic biomolecule passing through it under the influence of an applied voltage. (b) The current trace produced by a sample molecule translocating a nanopore. (c) An event out of several in the recorded current trace. (d) Flow chart representing various passes in the Nano Trees Algorithm. (e) The sampled event fitted using Nano Trees.

data, termed Nano Trees, intended for use in decoding arbitrary piecewise constant signals overlaid with noise and systemic distortion.

Experimental

The Nano Trees pipeline consists of steps that iteratively smooth and improve the overall fit by refining the estimate of the times at which significant step changes occur in the data. It begins by using decision trees[37,38] and adaptive boosting[39–41] to denoise the data, followed by iterative refinement of the fit using modular passes over the data. At each step, the signal becomes progressively smoother as physically insignificant sublevels are removed according to criteria specified by the user through adjustable hyperparameters discussed in detail in ESI Section S1.† The current version of the code is implemented in Python. It can run on any modern desktop computer.

Each of the steps that forms the pipeline is a modular component that operates independently of the others, and can be reordered or reapplied as needed. The selection and order of passes used here were found to be effective for fitting the data discussed in this work, but is not necessarily prescribed for all nanopore data, and can easily be updated as needed. Full automation of the selection of these hyperparameters is the subject of ongoing work.

The full Nano Trees pipeline, as well as a pseudocode implementation, is described in detail in ESI Section S1† and in a related master's thesis.[42] In short, data is normalized and subsequently grouped into sublevels in a hierarchical manner, beginning with an overfit of the data and iteratively merging sublevels and improving the fit through increasingly fine-grained passes over the underlying data until it is deemed to be physically accurate according to a set of context-specific, user-specified hyperparameters. Fig. 2 provides a block diagram view of the process that employs different supervised machine learning algorithms for fitting nanopore data.

Results and discussion

The performance of this algorithm is tested on two datasets comprising both real and synthetic nanopore data. These specific datasets have been chosen to cover a broad range of event shapes and use cases, highlighting the generality of the approach. The first dataset is a synthetic dataset for which the ground truth is known, and hence, the results produced by both algorithms on this dataset can be objectively compared. The second dataset is comprised of real translocations of biomolecules through solid-state nanopores.

Synthetic data

We began the testing under controlled conditions of a synthetic dataset for which we know the true underlying signal shape, and we compared Nano Trees to the CUSUM+ algorithm, which has been widely used for nanopore data analysis in the nanopore community, being at the core of at least four separate frameworks (these being OpenNanopore, CUSUM+ itself,

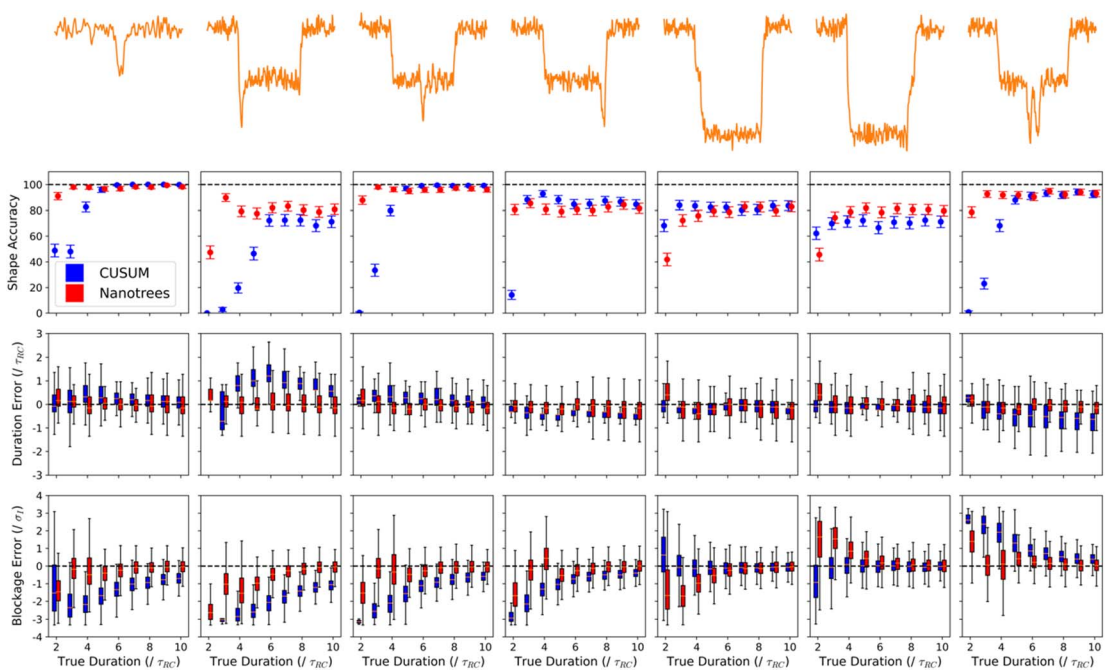

Fig. 3 Comparing Nano Trees to CUSUM+ on a synthetic dataset for 7 classes of signals. Top row: example of each signal class. Second row: metric of fit quality: shape accuracy. Defined as the fraction of events that have the correct number and ordering of sublevels in the resulting fit and blockage depths within three standard deviations of the true value. Third row: metric of fit quality: the error in the estimate of the duration of the transient. Bottom row: metric of fit quality: the error in the estimate of the blockage level for the transient. In all rows, the *x* axis is plotted as a multiple of the rise time used to simulate the events, while the *y* axis is plotted as a multiple of the simulated open pore standard deviation. Error is defined to be the signed difference between the fitted quantity and the known ground truth that was used to simulate the event.

MOSAIC, and Pyth-Ion).[18,19,27,33,43,44] This data was generated using ESI Script 1,† with event sublevels chosen to represent the most common difficult-to-fit motifs found in nanopore signals in the literature. Any nanopore event can be constructed as a linear combination of these basic subevents. These motifs are shown in the top row of Fig. 3 and are composed of the following classes:

Class 1. Event consisting of a single transient symmetric peaked sublevel.

Class 2. Event with transient symmetric peaked sublevel (asymmetric peak) at the beginning.

Class 3. Event with transient symmetric peaked sublevel (symmetric peak) in the middle.

Class 4. Event with asymmetric transient peaked sublevel at the end.

Class 5. Event with a transient sloped sublevel at the beginning.

Class 6. Event with a transient sloped sublevel at the end.

Class 7. Event containing a double peaked sublevel pair with transient separation.

For these classes of events, we varied the duration of the sublevel in class 1, the duration of the peaked sublevels in classes 2–4, the duration of the sloped sublevel in classes 5 and 6, and the gap between the two peaks in class 7. The true duration is varied in the inclusive range of 2–10 times the simulated rise time of the system.[19] For the sake of a direct comparison with real data, the rise time is simulated to be 1 μs. The signal was overlaid with uncorrelated white noise such that the SNR of the various transients is 6 times the standard deviation of the baseline current. The signal is then sampled at 5 MHz and low-pass filtered to a bandwidth of 1 MHz to show a typical nanopore measurement using state-of-the-art electronics. In this dataset, we have 500 events contributing to every data point, for 4500 events per class and 31 500 events overall.

The first metric is shape accuracy, which simply considers whether the sublevel structure of the fit is correct. To pass, a fit must have the correct number of sublevels in the correct order with respect to depth, have an error in the fitted blockage depth of the transient level not exceeding three standard deviations of the baseline noise, and have an error on the fitted duration of the transient level not exceeding 3 times the system rise time. The second row in Fig. 3 shows the percentage of events that passed as a function of transient duration. It is immediately clear that Nano Trees outperforms CUSUM+ significantly when transients are shorter than 4 times the system rise time, while matching performance for transients that achieve a steady state. Only the fits that had the correct shape (as defined above) were assessed for blockage and duration accuracy.

Of the fits that have the correct shape, we also compared the error in both the duration (third row in Fig. 3) and blockage depth (bottom row in Fig. 3, where a value of zero indicates a perfectly accurate fit). Duration error is calculated as the signed difference between the fitted duration of the transient part of the signal and the known ground truth duration used to simulate the event. Blockage error is calculated as the signed difference between the fitted blockage level and the true blockage depth used in the simulation. Both errors are expressed as multiples of the standard deviation of the baseline noise or the rise time of the system, as appropriate. A value of zero for either metric indicates a fit that coincides with the ground truth. As with the fraction of events for which the shape is accurate, Nano Trees outperforms CUSUM+ for short transients without sacrificing performance on longer ones. It is worth noting that both algorithms underestimate true blockage depth to increasing degrees as the transient duration approaches zero, which is to be expected given systemic distortion, though this effect is suppressed in Nano Trees compared to CUSUM+. Of note, the sloped sublevels (classes 5

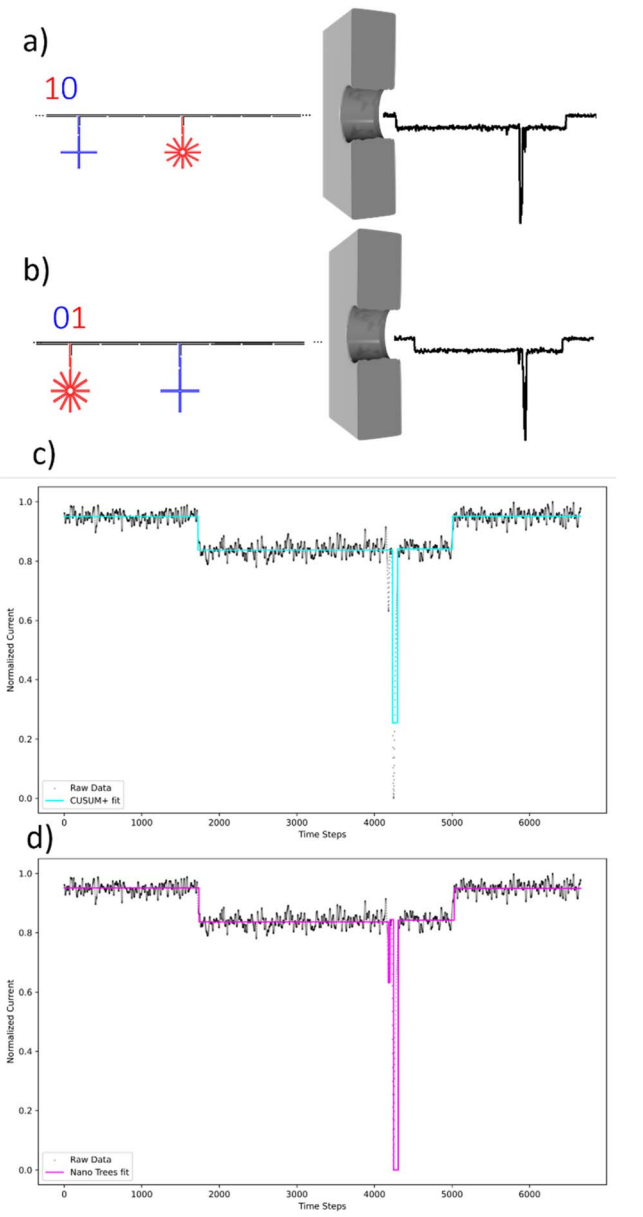

Fig. 4 (a) Schematic view of a two-bit barcode in the 10-configuration translocating a nanopore. The 4-arm star represents the 0 bit, the 12-arm star represents the 1 bit. (b) Schematic view of a two-bit barcode in the 01-configuration translocating a nanopore. (c) CUSUM+ fit on DNA Nanostructures Barcodes event. (d) Nano Trees fit on DNA Nanostructures Barcodes event.

and 6) display a clear directional bias, with Nano Trees underestimating or overestimating the blockage depth depending on whether the event is sloping upwards or downwards to a greater degree than CUSUM+. This is likely because the transients are themselves asymmetric and can be improved using more sophisticated algorithms to estimate the depth of sloped sublevels. In principle, approaches such as ADEPT[19] or the approaches suggested by Lucas *et al.*[30] or Gu *et al.*[31] could be implemented as additional modular passes in Nano Trees pipeline to correct these errors.

These signals provide clear insight into the limitations and types of errors that arise when using Nano Trees for fitting and highlight the improvements available over incumbent analysis methods in the regime of fast transients. The insights from this synthetic data are critical to inform evaluation of the quality of fits to real datasets, and to understand the strengths and limitations of the approach.

Decoding DNA nanostructure barcodes

To demonstrate the utility of this approach in the context of a real nanopore experiment, we apply Nano Trees to analyze a dataset for which our existing CUSUM+ framework failed consistently, with the intention of correcting these failures. This data arises from the translocation of a complex molecule that encodes information in the form of a double-stranded DNA (dsDNA) backbone with binding sites to which side chains can bind through DNA hybridization. These side chains carry a DNA nanostructure—a star with either 4 or 12 dsDNA arms,[7] which produce clearly distinguished blockage levels when they pass through a nanopore and are used to represent a digital 0 or 1, respectively. Events for which the 12-arm star appears closest to the end of the molecule are classified as "10 events" while events for which the 4-arm star is nearest the end are classified as "01 events". Because of the need to achieve high storage density, these side chains are quite small and result, in the ideal case, in symmetric peaked sublevels of differentiated depth and of short duration relative to the rise time of our measurement electronics. These transient levels are too fast for CUSUM+ to consistently recognize, let alone fit accurately, as shown in Fig. 4. Overall, compared to CUSUM+, Nano Trees fitting results in a much higher percentage of proper fits and accurate molecular classifications based on sublevel structure, as estimated by an expert human reviewer, shown in Table 1. As this is a real dataset, the underlying ground truth for any event is not available. The results in Table 1 are generated through manual inspection of each event and the corresponding fits from both algorithms by a human expert.

Table 1 Accuracy results on DNA Nanostructures Barcodes dataset of 292 events (147 from 10 dataset and 145 events from 01 dataset), compared to the ground truth assess through manual classification by a human operator. 10 Dataset contains events with 12 dsDNA arm star attached before 4 dsDNA arm star representing a binary 10. 01 Dataset contains events with 4 dsDNA arm star attached before 12 dsDNA arm star, representing a binary 01. Experiments were performed in 3.6 M LiCl at 150 mV bias voltage in a solid-state pore of size 14.2 nm. Current traces were low-pass Bessel filtered to 250 kHz for event detection and fitting

	Nano Trees fit accuracy	CUSUM+ fit accuracy
10 dataset	94.55%	59.18%
01 dataset	89.65%	48.96%
Combined	92.12%	54.10%

Conclusions

We developed the Nano Trees algorithm for improved sublevel fitting of nanopore data, to push the limits of temporal resolution, and to reduce instances of fitting errors and false negatives when considering fast transients in nanopore data. The code has a modular structure that allows for independent fitting steps to be introduced in any order to improve fit quality, allowing for a high degree of tunability across different experimental contexts. When the arrangement of fitting blocks used here is applied to various synthetic and real datasets, this algorithm outperforms CUSUM+. Fast transient as short as twice the rise time of the system can now be accurately fitted, though there remains room for improvement in the accuracy of fitting, particularly when considering sloped sublevels.

The algorithm is versatile and tuneable through its various hyperparameters to adjust to any type of piecewise constant data with systemic distortion and additive noise, and we are actively working on full automation of the relevant parameters. When applied to real nanopore data in particularly challenging cases, Nano Trees outperforms CUSUM+ by a considerable margin, while still falling short of the accuracy required for fully unsupervised decoding of arbitrary nanopore signals. Ongoing research and improvement will focus on reducing the false negative chance when transients approach the system's rise time, while inclusion of methods developed by others[19,30,31] as additional passes in the Nano Trees pipeline may provide a means to correct for directional bias in sublevel current estimation.

In the long term, the sublevel structure extracted by this analysis framework will form the feature set used by event classification methods in a variety of contexts, including single-molecular diagnostics, protein detection, and molecular information storage.

The definition and effect of each hyperparameter on the resulting fit is discussed in the ESI Section S6.† Users are advised to follow the hyperparameter descriptions and tuning procedures discussed in ESI Section S7† for obtaining the optimal configuration of these hyperparameters to obtain good fits. The hyperparameters used for the case studies in this work are given in ESI Section S8.†

Data availability

Part of the data supporting this article is available in the ESI.† Other datasets, including the synthetic data and translocation data for DNA barcode molecules, are available on the Federated Research Data Repository at **https://doi.org/10.20383/**

103.01212. A detailed description of the data is provided both directly in the repository and in ESI Section S9.†

Author contributions

DW wrote the first draft and implemented the Nano Trees software pipeline. DW, KB and PB designed the Nano Trees approach. PK produced experimental data. KB, JH, and VTC designed the study. All authors edited and approved the final manuscript.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We thank Dr Zachary Roelen for sharing nanopore data. All authors would like to acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [funding reference number ALLRP/586371-2023], The Ontario Centre of Innovation ("OCI"), (funding Project #: 35449), and Oxford Nanopore Technologies.

Notes and references

1 E. C. Yusko, B. R. Bruhn, O. M. Eggenberger, J. Houghtaling, R. C. Rollings, N. C. Walsh, S. Nandivada, M. Pindrus, A. R. Hall, D. Sept, J. Li, D. S. Kalonia and M. Mayer, *Nat. Nanotechnol.*, 2017, **12**, 360–367.
2 P. Waduge, R. Hu, P. Bandarkar, H. Yamazaki, B. Cressiot, Q. Zhao, P. C. Whitford and M. Wanunu, *ACS Nano*, 2017, **11**, 5706–5716.
3 V. Van Meervelt, M. Soskine, S. Singh, G. K. Schuurman-Wolters, H. J. Wijma, B. Poolman and G. Maglia, *J. Am. Chem. Soc.*, 2017, **139**, 18640–18646.
4 W. Si and A. Aksimentiev, *ACS Nano*, 2017, **11**, 7091–7100.
5 K. Briggs, H. Kwok and V. Tabard-Cossa, *Small*, 2014, **10**, 2077–2086.
6 D. Deamer, M. Akeson and D. Branton, *Nat. Biotechnol.*, 2016, **34**, 518–524.
7 L. He, D. R. Tessier, K. Briggs, M. Tsangaris, M. Charron, E. M. McConnell, D. Lomovtsev and V. Tabard-Cossa, *Nat. Commun.*, 2021, **12**, 5348.
8 S. King, K. Briggs, R. Slinger and V. Tabard-Cossa, *ACS Sens.*, 2022, **7**, 207–214.
9 S. E. Sandler, R. I. Horne, S. Rocchetti, R. Novak, N.-S. Hsu, M. Castellana Cruz, Z. Faidon Brotzakis, R. C. Gregory, S. Chia, G. J. L. Bernardes, U. F. Keyser and M. Vendruscolo, *J. Am. Chem. Soc.*, 2023, **145**, 25776–25788.
10 N. Das, B. Chakraborty and C. RoyChaudhuri, *Talanta*, 2022, **243**, 123368.
11 F. Rivas, O. K. Zahid, H. L. Reesink, B. T. Peal, A. J. Nixon, P. L. DeAngelis, A. Skardal, E. Rahbar and A. R. Hall, *Nat. Commun.*, 2018, **9**, 1037.
12 N. A. W. Bell and U. F. Keyser, *Nat. Nanotechnol.*, 2016, **11**, 645–651.
13 K. Chen, J. Zhu, F. Bošković and U. F. Keyser, *Nano Lett.*, 2020, **20**, 3754–3760.
14 C. Callegari, S. Giordano, M. Pagano and T. Pepe, *Comput Secur.*, 2012, **31**, 727–735.
15 M. Severo and J. Gama, Discovery Science. DS 2006, *Lect. Notes Comput. Sci.*, 2006, **4265**, 243–254.
16 A. Thiel, M. Greschner, C. W. Eurich, J. Ammermüller and J. Kretzberg, *J. Neurophysiol.*, 2007, **98**, 2285–2296.
17 L. Koepcke, G. Ashida and J. Kretzberg, *Front. Syst. Neurosci.*, 2016, **10**, 51.
18 K. Briggs, *Solid-State Nanopores: Fabrication, Application, and Analysis*, University of Ottawa, 2018.
19 J. H. Forstater, K. Briggs, J. W. F. Robertson, J. Ettedgui, O. Marie-Rose, C. Vaz, J. J. Kasianowicz, V. Tabard-Cossa and A. Balijepalli, *Anal. Chem.*, 2016, **88**, 11900–11907.
20 R. R. Wick, L. M. Judd and K. E. Holt, *Genome Biol.*, 2019, **20**, 129.
21 S. R. Choi and M. Lee, *Biology*, 2023, **12**, 1033.
22 M. Pagès-Gallego and J. de Ridder, *Genome Biol.*, 2023, **24**, 71.
23 K. Misiunas, N. Ermann and U. F. Keyser, *Nano Lett.*, 2018, **18**, 4040–4045.
24 E. S. Page, *Biometrika*, 1954, **41**, 100.
25 P. Granjon, *HAL Open Science*, 2014, hal-00914697.
26 R. El Sibai, Y. Chabchoub, R. Chiky, J. Demerjian and K. Barbar, in Web and Wireless Geographical Information Systems. W2GIS 2018. *Lecture Notes in Computer Science*, Springer, 2018, pp. 25–40.
27 C. Raillon, P. Granjon, M. Graf, L. J. Steinbock and A. Radenovic, *Nanoscale*, 2012, **4**, 4916.
28 A. Balijepalli, J. Ettedgui, A. T. Cornio, J. W. F. Robertson, K. P. Cheung, J. J. Kasianowicz and C. Vaz, *ACS Nano*, 2015, **9**, 12583.
29 A. Balijepalli, J. Ettedgui, A. T. Cornio, J. W. F. Robertson, K. P. Cheung, J. J. Kasianowicz and C. Vaz, *ACS Nano*, 2014, **8**, 1547–1553.
30 F. L. R. Lucas, K. Willems, M. J. Tadema, K. M. Tych, G. Maglia and C. Wloka, *ACS Omega*, 2022, **7**, 26040–26046.
31 Z. Gu, Y. L. Ying, C. Cao, P. He and Y. T. Long, *Anal. Chem.*, 2015, **87**, 907–913.
32 C. Plesa and C. Dekker, *Nanotechnology*, 2015, **26**, 084003.
33 Z. Roelen, K. Briggs and V. Tabard-Cossa, *ACS Sens.*, 2023, **8**, 2809–2823.
34 V. Tabard-Cossa, D. Trivedi, M. Wiggin, N. N. Jetha and A. Marziali, *Nanotechnology*, 2007, **18**, 305505.
35 V. Tabard-Cossa, in *Engineered Nanopores for Bioanalytical Applications*, Elsevier, 2013, pp. 59–93.
36 S. Shekar, C.-C. Chien, A. Hartel, P. Ong, O. B. Clarke, A. Marks, M. Drndic and K. L. Shepard, *Nano Lett.*, 2019, **19**, 1090–1097.
37 L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, *Classification and Regression Trees*, Wadsworth International Group, Belmont, CA, 1984.
38 T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, Springer New York, New York, NY, 2009.
39 Y. Freund and R. E. Schapire, *J. Comput. Syst. Sci.*, 1997, **55**, 119–139.

40 H. Drucker, in *ICML*, 1997, vol. 97, p. e115.
41 R. Avnimelech and N. Intrator, *Neural Comput.*, 1999, **11**, 499–520.
42 D. Wadhwa, *Nano Trees: Machine Learning Enhanced Signal Processing of Nanopore Data*, Université d'Ottawa/University of Ottawa, 2024.
43 R. Hu, Z. Zhang, L. Tian, G. Wei, Z. Wang, M. Wanunu, W. Si and Q. Zhao, *ACS Nano*, 2025, **19**, 11403–11411.
44 R. K. Lokareddy, C.-F. D. Hou, F. Forti, S. M. Iglesias, F. Li, M. Pavlenok, D. S. Horner, M. Niederweis, F. Briani and G. Cingolani, *Nat. Commun.*, 2024, **15**, 8482.