Journal Template

Molecular Dynamics of Peptide Sequencing through MoS_2 Solid-State Nanopores for Binary Encoding Applications

Andreina Urquiola Hernández[1], **Christophe Guyeux**[2], **Adrien Nicolaï**[1,*] **and Patrick Senet**[1]

[1] *Laboratoire Interdisciplinaire Carnot de Bourgogne, UMR 6303 CNRS, Université de Bourgogne, Dijon, France*
[2] *Institut FEMTO-ST, UMR 6174 CNRS, Université de Franche-Comté, Besançon, France*

This manuscript was compiled on August 29, 2024

Abstract

Biological peptides have emerged as promising candidates for data storage applications due to their versatility and programmability. Recent advances in peptide synthesis and sequencing technologies have enabled the development of peptide-based data storage systems for realizing novel information storage technologies with enhanced capacity, durability, and data access speeds. In this study, we performed peptide sequencing of 12 distinct sequences through a single-layer MoS_2 Solid-State Nanopore (SSN) using Molecular Dynamics (MD). Peptide sequences were comprised of 1 positively charged, 1 negatively charged, and 4 neutral amino acids, with the position of amino acids in the sequence being changed to generate all possible configurations. From MD, the goal was to evaluate the efficiency of these peptide sequences to represent binary information based on ionic current traces monitored during their passage through the nanopore. A classification approach using the LightGBM classifier was developed to analyze different sequence characteristics such as the influence of position of amino acids in the peptide sequence or the spacing between charged amino acids. This approach was successful to identify peptide sequence pairs relevant for encoding binary data. In addition, MD simulations allowed us to establish the nonlinear relationship between amino acid positions inside the nanopore and ionic current fluctuations to eliminate false positives and to enable effective training of machine learning algorithms. These very promising results allowed us to highlight the best approaches for peptide design as building blocks for molecular information storage using MoS_2 SSN. Particularly, criterion of the position of charged and neutral amino acids was preferred to design peptides representing binary bits. Finally, this study enhances our understanding of peptide-based data storage systems, highlighting their potential for creating efficient, scalable, and reliable molecular data storage solutions.

Keywords: *Solid-State Nanopores, MoS_2, Peptide Design, Sequencing, Data Storage, Molecular Dynamics, Ionic Current, Classification*

Corresponding author: Adrien Nicolaï *E-mail address:* adrien.nicolai@u-bourgogne.fr
DOI: https://www.doi.org/exampledoi/XXXXXXXXXX
Received: March 20, 2024 **Revised:** April 16, 2024 **Accepted:** April 20, 2024 **Published:** May 21, 2024

INTRODUCTION

Peptides, once relegated to the realm of biological molecules, are now emerging as promising candidates for data storage applications [1]. Their inherent versatility and programmability make them very attractive to encode digital information in a compact and efficient manner. Furthermore, advancements in peptide synthesis and sequencing technologies have facilitated the fabrication and readout of peptide-based data storage systems [1], [2]. Experimental techniques such as mass spectrometry enable the precise construction and interrogation of peptide libraries tailored for data storage applications [3]. This method focuses on using simple molecules arranged on an array plate, ordered both physically and by mass. It provides an alternative archival storage solution that is stable, energy-efficient, and secure. The encoding process is flexible, simple, and relies on physical manipulations rather than additional synthesis. Reading is achieved using a mass spectrometer, offering more information than traditional methods. While sensitivity varies, even small amounts of molecules can generate a spectrum. Advances in mass spectrometry technology promise further improvements, allowing for increased storage capacity per array with higher resolution spectrometers. A new approach for data storage using peptide sequences was previously reported in the literature [1], where the arrangement of amino acids encodes digital bits. Raw data were initially converted into sequences of amino acids, or peptides. To retrieve the information, peptides were sequenced and sequences were converted back into digital bits, and then subsequently decoded into raw data. To facilitate efficient synthesis and sequencing, encoded strings were divided into smaller parts with address indicators ensuring correct ordering upon retrieval. Successful synthesis, detection, and sequencing of peptides were achieved through careful selection of amino acids composing biological peptides. Moreover, assessing peptide length was crucial to increase the probability of achieving successful complete sequencing [1]. Finally, it has been demonstrated that shorter peptides offer easier synthesis and sequencing, resulting in fewer missed fragmentation. Conversely, longer peptides have the capacity to store more data per peptide, thereby reducing the overall number of peptides needed, along with the associated addresses and error correction overhead for equivalent data volumes. To strike a balance, the peptide length was standardized to 18 amino acids in [1]. Other parameters must be considered such as the selection and arrangement of amino acids within peptides. One approach involves using the distinct physico-chemical properties of the 20 natural amino acids to encode information. For example, hydrophobic and hydrophilic amino acids can represent binary values, while specific sequences or motifs may serve as markers for data retrieval and decoding. In a very recent work [4], the authors propose another approach that represents a successful integration of deep learning and structure-based modeling for precise peptide design. This method combines a Gated Recurrent Unit-based Variational Autoencoder with Rosetta FlexPepDock to generate peptide sequences and assess their binding affinity. Molecular Dynamics (MD) simulations were then performed to fine-tune the selection of peptides for experimental validation.

Due to its portability, nanopore sequencing-based technologies have garnered significant interest for DNA storage technology [5]–[14]. To characterize nanopore data storage channel, a computational simulator model was developed [5]. Theoretical signals generated by the simulator are validated by comparing them with real experimental signals, assessing sample differences and bio-molecular errors. The simulator offers the flexibility to specify sequencing coverage size, accommodating different sequencing redundancy levels in various experimental setups. This feature helps to evaluate the effectiveness of logical and sequencing/physical redundancy, guiding the design of encoding/decoding schemes and reconstruction methods. In the design of biological peptides for data storage applications, researchers aim to exploit their sequence-specific properties to represent binary information. By strategically arranging amino acids within the peptide chain, unique sequences can encode digital data in the form of bits. Moreover, peptides offer the potential for high-density storage due to their small size and the vast combinatorial possibilities of amino acid arrangements. The storage density of the peptide method, using only eight amino acids as monomers, could be 3.72 times greater than that of the DNA method, using four nucleotides as monomers [1]. Furthermore, this storage density can be enhanced even more by using 16 or more amino acids. However, a retrievable data density of 1.7×10^{10} bits/g is achieved using the peptide method, which is approximately nine orders of magnitude lower than that of the DNA method [1], [15]. This is due to the difference in how DNA and peptides can be amplified and detected. DNA can be amplified using polymerase chain reaction before sequencing, allowing a much smaller quantity of DNA to be used to retrieve data. Peptides, on the other hand, cannot be amplified in the same way, meaning that a larger quantity of peptides is required for data retrieval. This results in a lower practical data density for peptides compared to DNA. Nevertheless, there is potential for significant improvement in peptide-based data density. Advances in peptide sequencing and detection at much smaller scales (attomole, yoctomole, or even single-molecule) could bring the practical data density of peptides closer to their theoretical potential [2], [16], [17], thereby reducing the current gap between peptide and DNA data storage methods.

Effective design and analysis of peptides involves considering various factors such as sequence, amino acid composition, length, and solubility. Peptides derived from native proteins may require alterations, focusing on non-essential amino acids. Longer peptides often result in decreased purity [18], while hydrophobic amino acids can impact solubility [19]–[22]. Avoiding sequences prone to β-sheet formation [23] and ensuring a balance of charged and uncharged amino acids is also crucial [24]. By carefully assessing these factors, researchers can optimize peptide design for efficient assembly, purification, and solubility of the final product. Moreover, another study emphasizes the importance of efficiently predicting nucleotide identity [25]. Originally, they evaluated the classification performance of individual nucleotides (dAMP, dTMP, dCMP, or dGMP) using data from experiments conducted with specific pore diameters. Input variables included dwell time, the height/depth of ionic current blockade, the mean ionic blockade current, and the number of distinct ionic current jumps within a single translocation event. Based on these data, they discussed the relationship between classification schemes derived from unsupervised learning and the supervised models employed. Looking ahead, the design of peptides for data storage holds promise for realizing novel information storage technologies with enhanced capacity, durability, and data access speeds.

The exploration of peptides for data storage has highlighted the potential of integrating biological components with advanced computational models to improve the processes of encoding and retrieving information, as demonstrated by advances in nucleotide classification. This sets a precedent for future innovations in the field of data storage and retrieval. As we move towards exploring peptide design for data storage, it becomes evident that leveraging mathematical approaches for pattern recognition can significantly enhance our ability to decode complex biological signals and reveal biological insights, as demonstrated in several studies involving nanopore sensor data [26]–[29]. Moreover, the potential of machine learning algorithms to reveal biological insights inherent within nanopore sensor output data has been demonstrated in several studies [30]–[33]. From the development of SquiggleNet for real-time, direct classification of signals to the exploration of deep learning models for gene detection, computational approaches offer promising results to unravel the complexities of genetic information encoded in nanopore signals. Custom-designed informational polymers can be effectively deciphered using a specific variant of aerolysin biological nanopore (K238A) [34]. Through a bio-inspired framework, a single-bit resolution was achieved using a deep learning approach. This method allowed the accurate decoding of digital sequences containing up to 4 bits of information. The structure of aerolysin pore can potentially be fine-tuned to optimize translocation for better reading efficiency. In addition, the identity and relative concentration of polymer mixtures were effectively detected without prior knowledge. Therefore, there is a vast potential in exploring the chemical diversity of informational polymers to enhance decoding by biological nanopores. By hybridizing with DNA nucleobases, these polymers retain advantages of synthetic DNA for data storage. For example, different terminal nucleobases allow for more efficient capture and threading by the nanopore, enabling potential use of canonical DNA bases to define data structure for random access [35]. In parallel, advancements in nanopore sequencing simulations for DNA data storage applications and the development of nanopore-based DNA hard drives demonstrate innovative approaches to rewritable and secure data storage [6], [7]. Efforts to expand the molecular alphabet of DNA-based data storage systems, coupled with neural network nanopore readout processing, offer promising avenues for enhancing the capacity and efficiency of digital data storage using DNA nanostructures and solid-state nanopores, paving the way for future advancements in molecular data storage [8], [9]. In a very recent work [36], we demonstrated that single-layer MoS_2 nanopore sensors can differentiate in a distinct manner positively and negatively charged from neutral amino acids using MD and unsupervised machine learning techniques. We defined coarse grained sequences of proteins which consist of replacing the primary sequence of a protein made of the 20 amino acids to a sequence made of three types of amino acids depending on their charge: A for positive, B for negative and C for neutral amino acids.

In the present work, we performed translocation experiments of 12 different peptide sequences of amino acids through single-layer MoS_2 nanopores using MD (Fig. 1). Sequences were made of one positive (K, Lysine), one negative (E, Glutamic acid), and four neutral amino acids (A, Alanine) which were arranged in various configurations. Moreover, each sequence was chemically linked to a short polycationic charge carrier made of four Lysine, which facilitates the threading and capture of the peptide through the pore [37]. The goal of the present work was to evaluate their efficiency to represent binary information based on the ionic current traces monitored during their passage through the pore. For this purpose, we explored a supervised Machine Learning (ML) approach, *i.e.* classification approach, to study the influence of various criteria such as the position in the sequence and spacing between charged amino acids. We used the LightGBM classifier, known for its leaf-wise tree growth strategy minimizes loss, leading to faster convergence and better accuracy compared to other boosting algorithms, to identify pairs of peptide sequences potentially relevant for encoding binary data. The main advantage of this numerical approach compared to experiments is to establish the non linear relationship between the amino acid positions, which are known in MD, and the ionic current traces, as measured experimentally. This allows us to eliminate false positives, which appear as current modifications without "true" passage of the protein (in a sense of significant), and to train ML algorithms on the simulated current traces using the concept of coarse-grained sequencing of proteins proposed in a previous work [36].

MATERIALS AND METHODS

Molecular Dynamics

We performed extensive unbiased all-atom MD simulations in explicit solvent to simulate the translocation of 12 different peptide sequences through single-layer MoS_2 nanopore of diameter $D = 1.5$ nm, immersed in a 1M KCl electrolyte [36]. The simulation box, as represented in Fig. 1A, is made of around 100,000 atoms in total. MD was carried out with the GROMACS software package (version 2018.2 [38] in double precision), using AMBER99sb*-ILDN-q force-field [39]. Force-field parameters for MoS_2 are given in details in a previous work [36]. During translocation simulations, MoS_2 nanoporous membrane serves as the separation between cis and trans compartments (Fig. 1A). The peptide is initially positioned at a vertical distance of approximately 2.5 nm above the membrane, in the cis compartment. Prior to production, systems were equilibrated in the NVT (T = 300 K), first and then in the NPT ensemble (P = 1 bar) without any applied electric field. These equilibration runs lasted each for 100 ps, allowing the system to relax, first, to the desired temperature, and, second, to the desired volume. After equilibration, production run starts with random initial velocities and by applying an external uniform electric field across the membrane, corresponding to a voltage of 1 V. The duration of each production run was 400 ns in NVT ensemble, with a time step of 2 fs. In this work, 12 distinct peptide sequences made of 6 amino acids were studied, each of them connected chemically to a short polycationic charge carrier (4 Lysine, +4), as done in previous works [36], [37], [40]. Each peptide is made of 4 Ala (neutral, labeled hereafter A), 1 Lys (positive, labeled hereafter B), and 1 Glu (negative, labeled hereafter C), which are distributed at different positions in the sequence. To reduce the ensemble of sequences made of 4 A, 1 B and 1 C, a constraint of 2 consecutive A in the sequence was used, reducing the ensemble from 30 sequences to 12 sequences, as shown in Fig. 1B. In total, 50 runs of 400 ns were performed for each of the 12 sequences, resulting in a total simulation time of 240 μs.

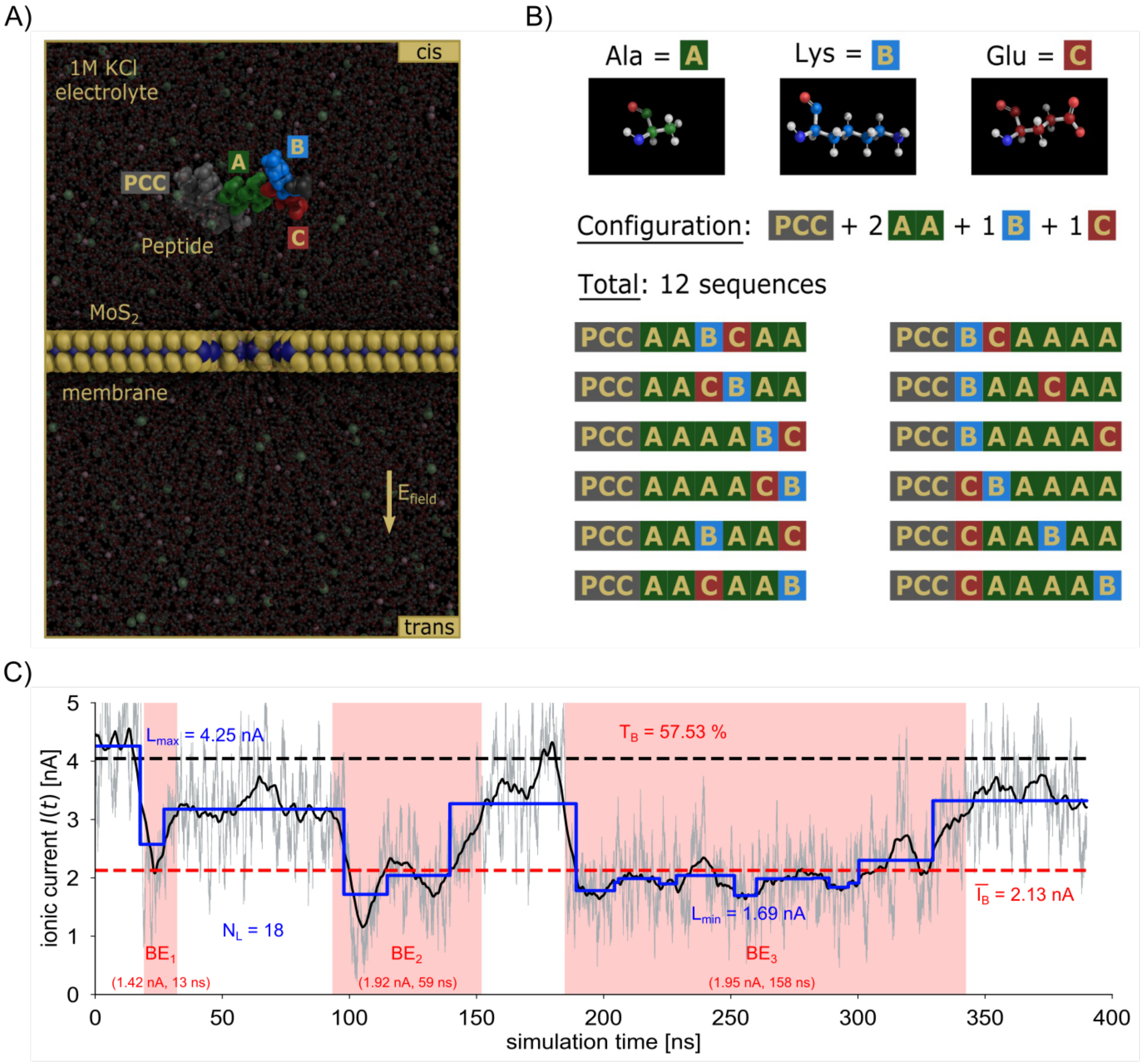

Figure 1. (A) Atomic representation of the solid-state nanopore sensor studied in the present work. The system is made of a MoS_2 nanoporous membrane (D = 1.5 nm), immersed in 1 M KCl electrolyte, plus a biological peptide. Atoms are shown with spheres: membrane (Mo: blue, S: yellow); amino acids (Alanine: green, lysine: blue and glutamic acid: red, polycationic charge carrier: gray). Ions and water molecules are represented with transparent spheres (K^+: palegreen, Cl^-: lightpink) and balls and sticks (O_w: red and H_w: white), respectively. (B) Design of peptide sequences studied in the present work. Neutral (A), positively (B) and negatively (C) charged amino acids are shown in green, blue and red, respectively. Polycationic Charge Carrier (PCC), made of 4 Lysines, is shown in gray. (C) A typical ionic current time series (in nA) as a function of time (in ns) and monitored during molecular dynamics simulations. Red areas represent blockade events extracted using the two-threshold method and their corresponding current drop ΔI_B and dwell time τ_B are also indicated. Blue lines represent structural breaks within the time series, extracting N_L = 18 ionic current levels in this example. Minimum and maximum blockade levels, L_{min} and L_{max} respectively, are also indicated, as well as the blockade duration T_B and the average of the blockade current value $\overline{I_B}$.

Ionic Current Time Series

Ionic current traces were extracted from MD production runs by tracking z-coordinates of K$^+$ and Cl$^-$ ions over time, and computed as follows:

$$I(t) = \frac{1}{\Delta t L_z} \sum_{i=1}^{N_{\text{ions}}} q_i \left[z_i(t + \Delta t) - z_i(t) \right] \tag{1}$$

where Δt represents the time interval between MD snapshots selected for calculations ($\Delta t = 1$ ns), L_z corresponds to the dimension of the simulation box in the z-direction, which aligns with the applied electric field direction, N_{ions} corresponds to the total number of ions in the simulation box, q_i is the charge of ion i, and $z_i(t)$ corresponds to the z-coordinate of ion i at time t. Ionic current was monitored every 10 ps during MD simulations, leading to time series length of 39,901 data points for each production run. Finally, traces were filtered to remove high frequency fluctuations by computing the moving mean of each trace over 1,001 samples.

Peptide Induced Blockade Events

From ionic current traces as shown in Fig. 1C, identification of peptide-induced Blockade Events (BEs) was performed using a two-threshold method. Initially, a threshold, referred to as th_1, is utilized to detect possible BEs. Threshold 1 was defined as $th_1 = \overline{I_0} - 4\sigma_0$, where $\overline{I_0}$ is the mean open pore ionic current and σ_0 is its standard deviation. For single-layer MoS_2 nanopores of diameter $D = 1.5$ nm, values of $\overline{I_0}$ and σ_0 are 4.04 and 0.23 nA, respectively. Using this threshold provides the advantage of effectively reducing the open pore ionic current fluctuations observed throughout translocation experiments. After identifying possible BEs based on th_1, we computed the corresponding probability density $P(I_B)$ of the event and a single Gaussian distribution was fitted to the data. Finally, if the mean value of the Gaussian distribution is below threshold th_2, which is defined as $th_2 = th_1 - \sigma_0 = \overline{I_0} - 5\sigma_0$, the event was definitely classified as a peptide-induced blockade event.

Moreover, each BE was defined by 2 parameters: i) its duration or dwell time τ_B and ii) its depth or current drop ΔI_B, which was computed as the difference between $\overline{I_0}$ and the mean blockade ionic current $\overline{I_B}$ of the event. From MD, the majority of the 12 peptide sequences presents between 60 and 100 BEs throughout 50 runs, except for sequences AABCAA, BAAAAC, AACAAB, AABAAC, and BAACAA, which present more than 100 BEs. BAAAAC is the sequence with the most BEs (nearly 200), which represents a significant difference (56 BEs) compared to the peptide sequence AABCAA with the second most BEs. Moreover, 2-D maps representing dwell time τ_B and current drop ΔI_B for each sequence and presented in Supplementary Materials (Fig. S1) present overall similar distributions but with some specific differences among them. In all peptide sequences, ionic current drops ΔI_B did not exceed 3.0 nA, except for a single BE observed in sequence CAABAA. Additionally, for each peptide sequence where B precedes C, named S$_{BC}$ sequence group (6 sequences), more than 97% of BEs exhibit ΔI_B<2.0 nA. In fact, the number of BEs with drops below 1.5 nA remains quite high for some sequences, such as BAAAAC (92%), BAACAA (84%), and AABAAC (82%), averaging 79% across all peptide sequences S$_{BC}$. In contrast, for peptide sequences where C precedes B, named S$_{CB}$ sequence group (6 sequences), BEs with ΔI_B<2.0 nA represent 88% of the dataset on average. However, those BEs with drops below 1.5 nA represent an average of 46%. Regarding dwell time τ_B of BEs, BAACAA peptide sequence does not present any BE with a duration τ_B greater than 100 ns, whereas BCAAAA and BAAAAC sequences present only 4% and 10% of such BEs, respectively. Overall, the majority of BEs in S$_{BC}$ group are characterized by τ_B<100 ns, averaging 90% across all these sequences. Additionally, only three peptide sequences in S$_{CB}$ present more than 30% of BEs with τ_B>100 ns. It concerns CBAAAA, AACBAA, and CAABAA peptide sequences, averaging 26% across all the sequences. On the other hand, peptide sequence BCAAAA shows the largest number of BEs with τ_B<10 ns, representing only 8% of the data.

Machine Learning Techniques

To identify significant changes in ionic current traces to uncover hidden patterns within their fluctuations, we employed first structural break detection. The identification of each level in ionic current traces for each MD run (Fig. 1C) has been performed and treated as a potential feature for the classification model of peptide sequences. Chow test, a tool for detecting structural breaks and evaluating parameter stability in regression models, was utilized for this purpose. We employed scikit-learn, an open-source Python library for machine learning, to conduct the detection. Basically, Chow test partitions the data into two subsets and examines whether the coefficients of the linear regressions remain consistent across them [41], [42]. Rejecting the null hypothesis indicates structural changes. The procedure involves fitting a regression equation to the complete set of observations, including both subsets, and calculating the residual sum of squares. Next, separate regression equations are fitted to each subset, and the residual sum of squares for these individual regressions are calculated. The ratio of the difference between the combined residual sum of squares and the sum of the residual sums of squares from the separate regressions to this latter sum follows an F-distribution under the null hypothesis, once adjusted for the corresponding degrees of freedom. This method has variations depending on whether both samples have enough observations to derive a regression equation (i.e., the observations exceed the number of estimated regression parameters) or if one sample has more observations than the estimated parameters while the other sample lacks sufficient observations.

This preliminary postprocessing of the data enables precise characterization and extraction of essential features necessary for accurate classification of ionic current observations, ultimately facilitating the recognition of sequences useful for efficient information encoding. LightGBM was selected as the algorithm for the classification problem of peptide sequences due to its advantageous features and capabilities, its efficiency in handling large datasets, and fast training speed. Based on some experiments conducted on a variety of public datasets, LightGBM has been shown to greatly speed up the training process of conventional Gradient Boosting Decision Trees (GBDT), achieving up to a 20-fold increase in speed while preserving nearly identical accuracy. Additionally, this algorithm incorporates two specific techniques: Gradient-based One-Side Sampling and Exclusive Feature Bundling. These techniques are designed to handle large datasets and a high number of features, respectively. Experimental results in [43] indicate that LightGBM significantly surpasses eXtreme Gradient Boosting (XGBoost) and Stochastic Gradient Boosting (SGB) in both computational speed and memory efficiency. Additionally, its ability to handle imbalanced datasets through class weights and its flexibility in parameter tuning further enhanced its suitability. Overall, LightGBM provided an efficient solution for the classification problem performed here.

In the supervised learning process applied hereafter, the training and testing datasets were divided in a 70 to 30% ratio. Cross-validation was employed to select hyperparameters such as the number of estimators, the maximum depth, and the learning rate. Additionally, a grid search was performed, specifying the model, parameter grid, scoring metric, and cross-validation strategy. An exhaustive feature selection method was implemented, involving a comprehensive search where all possible combinations of features are evaluated. This means conducting a brute-force evaluation of feature subsets, with the optimal subset being chosen by optimizing a specified performance metric for a given classifier. Given the small number of features extracted from the statistical analysis, computational complexity was not a problem. The evaluation of the performance of the different feature combinations in the classification task was conducted using four metrics: accuracy, which calculates the percentage of correctly predicted instances out of all predictions; recall, which measures the percentage of true positives over the sum of true positives and false negatives; precision, which calculates the percentage of true positives out of all instances predicted as positive; and F1-score, which represents the harmonic mean of precision and recall. These four metrics together offer a comprehensive view of the model performances from multiple angles, as they allow measuring the overall correctness of the model, minimizing false positives and false negatives, and maintaining a balance between the latter two. It therefore allows for informed decisions about implementation and adjustment. After a preliminary model selection process, the model with the best performance was ultimately chosen using the confusion matrix.

RESULTS

Analysis of Blockade Events Dataset and Feature Selection for Peptide Sequence Classification

First, a preliminary statistical analysis of ionic current traces dataset was conducted in order to extract information that could be potentially relevant for the identification of peptide sequences as they pass through MoS_2 nanopores. A total of six features were selected among tens of them analyzed, with three of them being extracted from the detection of BEs per simulation and the other three extracted from the full ionic current trace per simulation. This approach was conceived to incorporate more comprehensive information about the dynamics of the translocation process observed during MD. It leads to a well balanced dataset between the 12 peptide sequences since the same number of MD runs of the same duration were performed, leading to $n = 50$ observations with 6 features per peptide sequence. In details, it concerns i) the number of BEs per simulation N_B (feature F_1); ii) the number of ionic current levels per simulation N_L (feature F_2); iii) the minimum level of ionic current within a simulation L_{min} (feature F_3); iv) the maximum level of ionic current within a simulation L_{max} (feature F_4); v) the blockade duration per simulation T_B (feature F_5), which is defined as the ratio between the sum of individual BE duration within a simulation and the total simulation time (400 ns) and vi) the mean blockade ionic current per simulation $\overline{I_B}$ (feature F_6), which is defined as the average of BE ionic current values. These six features are highlighted for a given MD simulation of a given sequence in Fig. 1C. In this example, three BEs were detected using the two-threshold method in the present simulation (N_B=3). It corresponds to 22,956 values of blockade ionic current over the 39,901 values of the full time series, which leads to a blockade duration of 57.53 %. The corresponding average of the 22,956 blockade ionic current values is $\overline{I_B} = 2.13$ nA. Furthermore, from the trace presented in Fig. 1C, 18 levels of current ($N_L = 18$) were detected using structural break detection. From these 18 levels, the minimum and maximum levels of ionic current were $L_{min} = 1.69$ nA and $L_{max} = 4.25$ nA, respectively. As mentioned above, other features were tested such as the number of simulation with/without BEs, the mean and the standard deviation of blockade ionic current I_B, dwell time τ_B and ionic current drop ΔI_B per BE, the highest absolute value of ionic current per simulation, the first location of the minimum and maximum value of ionic current per simulation, the kurtosis, the median, the root mean square, the sample skewness, the standard deviation and the variation coefficient of ionic current per simulation, most of them were calculated using the Python package tsfresh (Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests).

Fig. 2 shows statistical distributions of the 6 selected features, represented as box plots for each peptide sequence. Blockade events dataset analysis revealed the presence of two distinct groups of peptide sequences, each comprised of 6 sequences (Fig. 2A).The defining characteristic among peptide sequences within each group lies in the position along the sequence of positively charged amino acid B, that is driven in the direction of the applied electric field (shown in Fig. 1A) and negatively charged amino acid C that is propelled in the opposite direction. In the first group of sequences S_{BC} (in blue), B precedes C in the 6 peptide sequences and, in the second group, named S_{CB} (in red), C precedes B in the 6 peptide sequences (Fig. 2A). For most of the features described above, we observed that peptide sequences within each group share very similar properties that clearly help us distinguishing them from the other group (Fig. 2B). First, feature F_1 (N_B) shows a notable difference in the median of the distribution for peptide sequences AABCAA, BAAAAC, and AABAAC, which are higher than those of the other peptide sequences. In addition, all peptide sequences in S_{CB} group show outliers greater than the maximum non-outlier, while these are present only in two peptide sequences of the S_{BC} group, i.e., AAAABC and AABAAC. Most of the peptide sequences in S_{CB} exhibit a distribution with lower dispersion and a lower median than most of peptide sequences in S_{BC}. It is clearly observed that feature F_3 (L_{min}) shows lower medians for peptide sequences in S_{CB} (lower than 2.3 nA), as well as greater dispersion than peptide sequences in S_{BC}, whereas peptide sequences in S_{BC} present a median greater than 2.4 nA. In the case of feature F_2 (N_L), a similar behavior is also observed between the two groups of sequences S_{BC} and S_{CB}, i.e. lower medians ($N_L < 22$ ionic current levels) and greater dispersion for peptide sequences in S_{CB}, whereas the medians for peptide sequences in S_{BC} shows $N_L \geq 22$ ionic current levels. For feature F_4 (L_{max}), probability densities behave similarly for all peptide sequences, with a median around 4.0 nA. This is because L_{max}, which represents the highest level of ionic current, is at the same level as the open pore ionic current $\overline{I_0}$. However, peptide sequence BAAAAC shows a notable dispersion for L_{max} compared to other sequences. On the other hand, for feature F_5 (T_B), distributions of peptide sequences in S_{CB} generally show greater dispersion than peptide sequences in S_{BC}, with some exceptions such as AABAAC and AABCAA sequences which are very wide compared to the others sequences in S_{BC}. Peptide sequence BCAAAA is characterized by a very short median (4 ns) compared to the others, while the sequence with the longest median is AACAAB (54 ns). Sequences in S_{CB} present, on average, a median of around 21 ns, whereas sequences in S_{BC} present, on average, a median of around 31 ns. Finally, for feature F_6 ($\overline{I_B}$), medians of probability densities of all peptide sequences are uniform. However, it shows a lower mean ($I_B < 2.5$ nA) and a greater dispersion for peptide sequences in S_{CB} than in peptide sequences in S_{BC}. BCAAAA peptide sequence presents a singular behavior compared to the other sequences with a large variability and values $\overline{I_B} < 2.5$ nA.

Probability densities P of the six features for the two groups of peptide sequences, S_{BC} and S_{CB}, were computed by applying the Gaussian Mixture Model (GMM) algorithm combined with the Bayesian Information Criterion (BIC), as presented in Fig. 2C. It involves a total of 300 data points for each feature in each group. First, $P(N_B)$ exhibits two sub-populations for the two groups S_{BC} and S_{CB} of peptide sequences, with similar means for the sub-population with the largest weight ($< N_B >= 1.18$ for S_{CB} and $< N_B >= 1.39$ for S_{BC}). The second sub-population shows larger differences, with peptide sequences in S_{BC} group presenting more events per simulation(S_{BC}: $< N_B >= 4.14$ and S_{CB}: $< N_B >= 3.07$). Second, probability densities of sensing time $P(T_B)$ exhibit two sub-populations for group S_{CB} and three for group S_{BC}. Sub-populations for group S_{BC} are centered around 5%, 30%, and 70%, whereas for group S_{CB}, they are centered around 12% and 55%, making them clearly distinguishable from each other. Third, concerning the mean blockade ionic current, $P(\overline{I_B})$ exhibits three sub-populations for peptide sequences in group S_{BC} and two for sequences in group S_{CB}. Particularly, the main sub-populations for each group are clearly distinguishable from each other which may favor the classification task (S_{BC}: $< \overline{I_B} >= 2.70$ nA vs. S_{CB}: $< \overline{I_B} >= 2.32$ nA). Fourth, $P(N_L)$ which represents the probability density of the number of current levels per simulation, exhibits two sub-populations for peptide sequences in group S_{CB}, while group S_{BC} is only characterized by one population. Moreover, main sub-populations of each sequence group are also clearly distinct from each other (S_{BC}: $< N_L >= 24.80$ vs. S_{CB}: $< N_L >= 16.63$). Similarly, $P(L_{min})$ for pep-

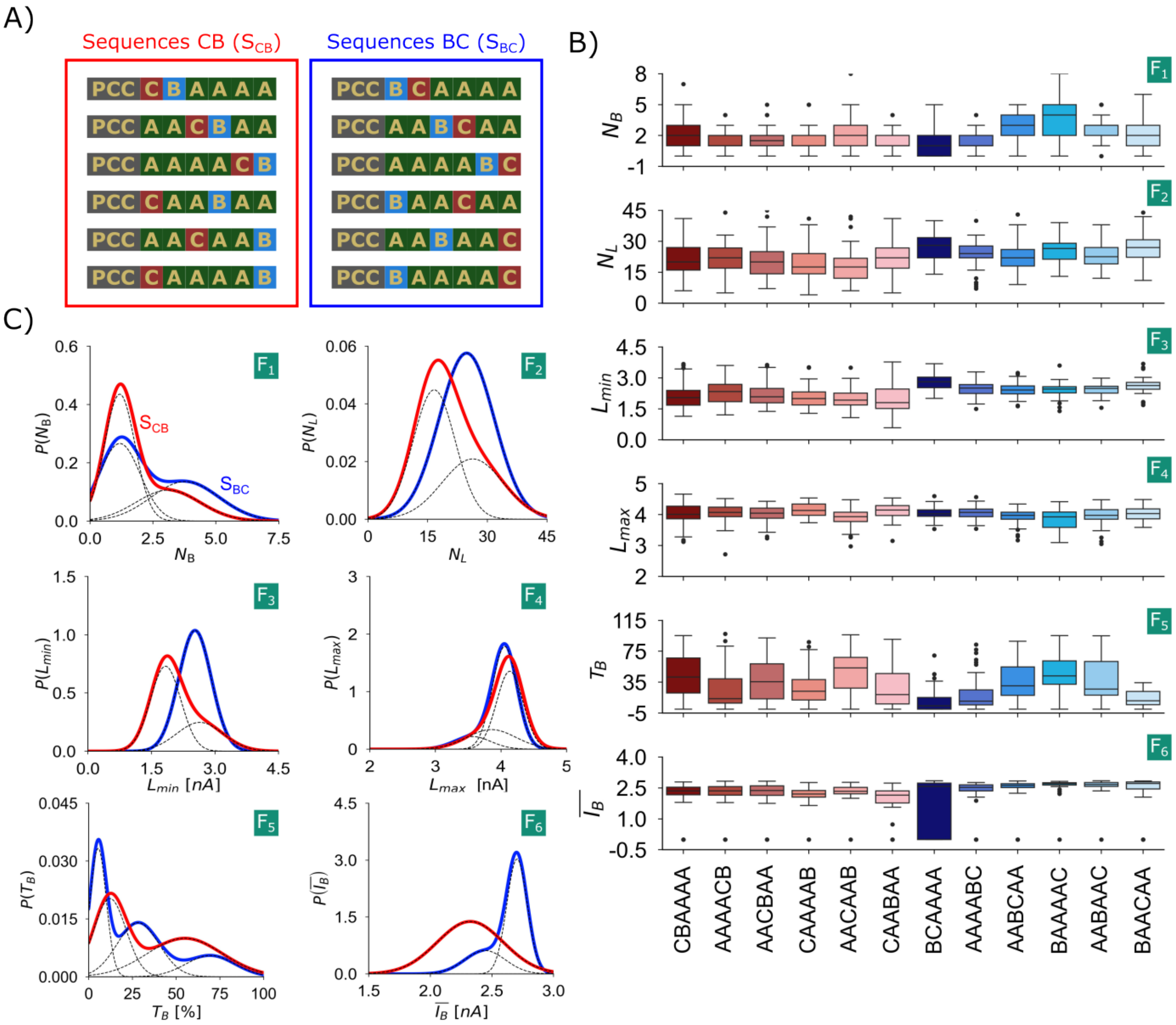

Figure 2. (A) Groups of peptide sequences S$_{BC}$ and S$_{CB}$. (B) Box plot of the six features extracted from the ionic current dataset as a function of peptide sequence: the number of BEs (N_B), the number of blockade ionic current levels (N_L), the minimum and maximum ionic current level (L_{min} and L_{max}), the total sensing time (T_B) and the mean blockade ionic current ($\overline{I_B}$). (C) Probability densities of the six features for each group of peptide sequences: S$_{BC}$ (in bluish colors) and S$_{CB}$ (in reddish colors).

tide sequences in group S$_{CB}$ exhibits two sub-populations, whereas group S$_{BC}$ presents only one population. Main sub-populations in both groups are well separated from each other, with $< L_{min} >= 2.52$ and 1.83 nA for peptide sequences in S$_{BC}$ and S$_{CB}$ groups, respectively. Finally, $P(L_{max})$ for both groups exhibit two sub-populations, with the main sub-populations being very close to each other (S$_{CB}$: $< L_{max} >= 4.05$ nA vs. S$_{BC}$: $< L_{max} >= 4.13$ nA), which was expected as they represent ionic current values corresponding to an open pore situation. However, slight differences are observed due to a "shadow" effect of the peptide above the pore, as already mentioned and described in a previous work [40]. Therefore, this feature is also included to evaluate peptide sequence classification performances. To conclude, statistical analysis of the dataset and feature selection were crucial for revealing the potential of ionic current characteristics as discriminatory features for the classification task of these two peptide sequence groups. The main sub-population for the sensing time T_B in sequences S$_{CB}$ has a mean of approximately 55%, which is significantly larger than the mean for the main sub-population in sequences S$_{BC}$ (28%). This disparity may contribute to the observation that the main sub-population for features L_{min} and $\overline{I_B}$ shows lower mean values for peptide sequences in S$_{CB}$ compared to sequences in S$_{BC}$, as longer sensing times could lead to a larger drop of ionic current. As illustrated in Fig. S5, the sensing time for motif C, which exhibits the longest duration (over 30%), is larger for peptide sequences in S$_{CB}$ group than in S$_{BC}$ group. This difference may also explain why the overall sensing time is longer for S$_{CB}$ compared to S$_{BC}$. Additionally, as previously discussed, BEs in sequences S$_{CB}$ generally exhibit a longer dwell time.

Classification of Peptides according to the Position of Charged Amino Acids in their Sequences

Once two distinct peptide sequence groups have been determined, their potential as well-distinguishable sequences for binary encoding was evaluated using a classification technique (supervised learning). First, from the six features presented in Fig. 2, a comparison between accuracy scores of models made of different combinations was performed. Results are shown in Fig. 3A with red asterisks highlighting the combinations leading to the best accuracy scores. Other metrics as precision, recall and F1-score were also evaluated (see Supplementary Materials, Fig. S2). Model selection process showed that overall the best score is obtained for the combination of features F$_{1,3,6}$ (N_B,

L_{min}, $\overline{I_B}$) with accuracy: 0.775, precision: 0.766, recall: 0.824 and F1-score: 0.780. Among the 63 possible feature combinations (some of them are shown in Fig. 3A), 76% of them achieved an accuracy larger than 0.7 and the combination of features $F_{3,4,6}$ (L_{min}, L_{max}, $\overline{I_B}$) achieves the highest accuracy, with a value of 0.777. The next four combinations with the highest accuracy are: $F_{3,6}$, $F_{1,3,6}$, $F_{1,3,4,6}$ and $F_{2,3,4,6}$. Additionally, regarding the precision, 76% of the combinations achieve a precision larger than 0.7. The combination of features with the highest precision is $F_{3,6}$ (L_{min}, $\overline{I_B}$) and $F_{1,3,6}$ (N_B, L_{min}, $\overline{I_B}$), with a value of 0.766. Regarding the recall, 41% of the 63 possible combinations exhibit a recall larger than 0.8. Feature combinations $F_{1,5,6}$ (N_B, T_B, $\overline{I_B}$) achieve the highest recall with a value of 0.837. Finally, regarding F1-score, 76% of the 63 combinations achieve scores larger than 0.7. Feature combination $F_{1,3,6}$ (N_B, L_{min}, $\overline{I_B}$) shows the highest F1-score, with a value of 0.780. From these results, feature F_6 ($\overline{I_B}$), followed by feature F_3 (L_{min}), are the most impactful for improving the classification model's performance, as they appear in all the combinations with the best metric scores. It means that the average blockade current and the minimum current level per simulation are crucial to differentiate ionic current traces of both groups of sequences, S_{BC} and S_{CB}. Then, using $F_{1,3,6}$ combination (N_B, L_{min} and $\overline{I_B}$), we computed the confusion matrix of the classification model which shows an average identification accuracy of 72% (Fig. 3B). In total, the model correctly identifies 65% of peptide sequences belonging to group S_{CB} and 79 % of peptide sequences belonging to group S_{BC}. Moreover, values of the three evaluated classification metrics, i.e. precision, recall, and F1-score are comprised between 0.65 and 0.85, with consistently better performances to classify peptide sequences in group S_{BC} compared to S_{CB}. However, false negative rate is high for class S_{CB} (35%), which suggests that the classification model sometimes

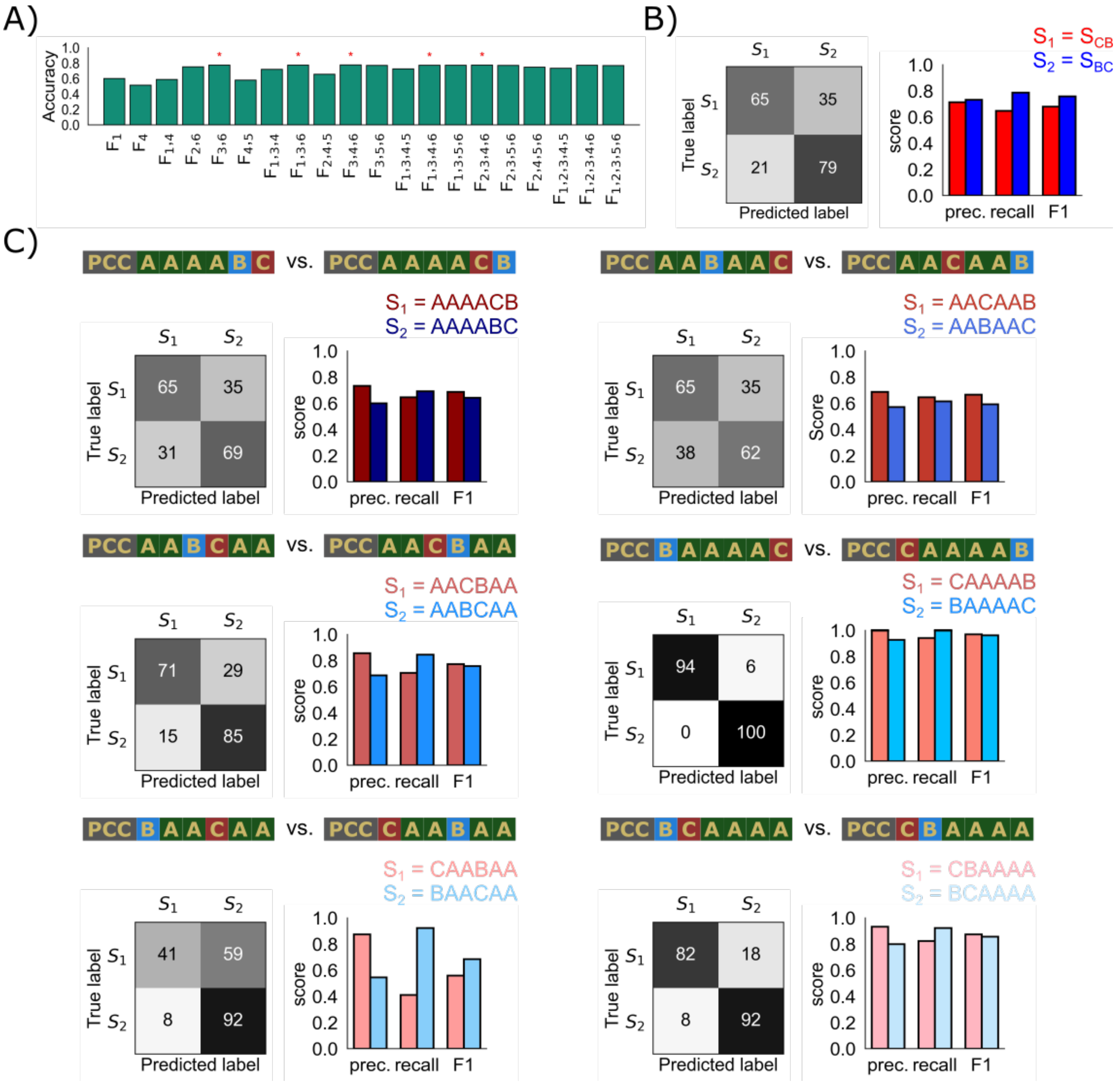

Figure 3. (A) Accuracy scores of the classification of the two groups of peptide sequences S_{BC} and S_{CB} using LightGBM classifier and evaluated for different combinations of features. Red asterisks indicate combinations of features with the highest accuracy scores. (B) Confusion matrix of the classification for peptide sequences in S_{BC} vs. S_{CB} group using a combination of features $F_{1,3,6}$. Precision, recall, and F1-score calculated for both groups S_{BC} and S_{CB} are shown as bar plots. (C) Confusion matrices and precision, recall and F1-score metrics for binary classifications between a peptide sequence in S_{CB} group versus its corresponding peptide sequence in S_{BC} group.

confuses peptide sequences in S_{CB} as S_{BC}. In summary, although the classification model developed here shows good overall performances and is very promising, particularly for class S_{BC}, there is room for improvement in identifying peptide sequences in group S_{CB}.

Then, pairwise sequence classification was performed to determine the design of peptide sequences which offers the best classification performances. Therefore, classification was performed between each pair of peptide sequences with one belonging to S_{CB} group and its corresponding sequences in S_{BC} group, resulting in six classification tasks being carried out. Results are shown in Fig. 3C. Classification tasks of peptide sequences AAAACB vs. AAAABC and AACAAB vs. AABAAC show that the best score is achieved with F_2 (N_L) as the key feature for the former, and a combination of $F_{2,3}$ features (N_L + L_{min}) for the latter. In addition, classification performances generally resemble classification of peptide sequence in S_{CB} vs. S_{BC} group, with precision, recall, and F1-score metrics around 0.7, although it shows lower values for classification of AACAAB vs. AABAAC peptide sequences. The average accuracy score decreases to 67% for AAAACB vs. AAAABC and 63% for AACAAB vs. AABAAC, consequently increasing the false negative and the false positive rates (see Fig. 3B). Based on the analyzed metrics, in both cases the classification shows similar performance for both sequences. Furthermore, we observed that classification tasks of peptide sequences CAAAAB vs. BAAAAC, CBAAAA vs. BCAAAA, and AACBAA vs. AABCAA present excellent scores, with average accuracy scores of 97%, 87%, and 78%, respectively. Combinations of features chosen for each classification task based on the best performance were $F_{1,2,4,5,6}$ for CAAAAB vs. BAAAAC, $F_{1,3,5}$ for CBAAAA vs. BCAAAA, and $F_{4,6}$ for AACBAA vs. AABCAA. In general, false negative and positive rates are quite low, with the highest false negative rate being 29.41 % and the highest false positive rate being 15%, both for the classification with the lowest performances (AACBAA vs. AABCAA). However, a null false positive rate and a false negative rate of only 5.88 % were achieved for the best classification model among all pairwise sequence comparisons, i.e. CAAAAB vs. BAAAAC. Regarding the other metrics, classification of CAAAAB vs. BAAAAC sequences achieved values larger than 0.9, classification of CBAAAA vs. BCAAAA sequences achieved values larger than 0.8, and classification of AACBAA vs. AABCAA sequences achieved values larger than 0.7 (Fig. 3C). Similarly, the three models perform better to classify peptide sequences where B (positively charged amino acid) precedes C (negatively charged amino acid). Lastly, classification of peptide sequences CAABAA vs. BAACAA shows the least efficient scores, especially for CAABAA sequence, with significant differences observed in the precision and recall metrics for both peptide sequences. Features selected for this classification task were $F_{1,3,4,6}$ (N_B, L_{min}, L_{max} and $\overline{I_B}$). False negative rate is quite high (59%), suggesting that the model frequently classifies peptide sequences CAABAA as BAACAA. However, false positive rate is extremely low (8%), which means that a very few ionic current traces of BAACAA sequence were incorrectly classified. Peptide sequence CAABAA shows good performance for the precision, with a value of approximately 0.9, but poor performance for the recall, with a value around 0.4. On the contrary, peptide sequence BAACAA shows excellent performance for the recall (around 0.9), but a lower precision, around 0.5 (Fig. 3C). This tells us that, although the classification model developed here is very effective at correctly detecting peptide sequences CAABAA (high precision), it has difficulty identifying peptide sequences CAABAA when predicting them (low recall). On the other hand, the classification model is able to detect most peptide sequences BAACAA (high recall), but the precision of these predictions is bad.

To summarize, CAAAAB and BAAAAC peptide sequences exhibit by far the best classification performances, achieving an accuracy of 0.97, which makes them the best pair of sequences among all those studied to represent bits 0 and 1 in binary information encoding. Therefore, they are the most promising candidate for accurately reading digital information encoded in a peptide sequence with single-bit resolution. This demonstrates the sensitivity of MoS_2 nanopore sensors in detecting and differentiating sequences with excellent performances, particularly when charged amino acids are separated by four neutral amino acids (the largest separation tested in the present work among all sequences). Overall, all pairwise classification tasks yielded good accuracy scores, except for classification of peptide sequences CAABAA vs. BAACAA, where there is a marked disparity between performances of both classes (Fig. 3C).

Classification of Peptides according to the Spacing between Charged Amino Acids in their Sequences

To evaluate information about peptide sequences that are selective in addition to the position of charged amino acids, to distinguish sequences within each S_{CB} and S_{BC} groups, two classification approaches were carried out separately based on two different criteria and following the same strategy as described above. First, within each group of peptide sequences S_{CB} and S_{BC}, sequences were classified using the information about the spacing between charged amino acids in the sequence. We consider two subgroups: i) sequences for which charged amino acids (B and C) are separated in the sequence by neutral amino acids (A), named $S_{CB}^{sep.}$ and comprised of AACAAB, CAABAA and CAAAAB peptide sequences; ii) sequences for which charged amino acids (B and C) are consecutive or linked together in the sequence, named $S_{CB}^{tog.}$ and comprised of AACBAA, CBAAAA and AAAACB. The same subgroups of peptide sequences can be done for S_{BC} group, resulting in two classification problems $S_{CB}^{sep.}$ vs. $S_{CB}^{tog.}$ (Fig. 4A) and $S_{BC}^{sep.}$ vs. $S_{BC}^{tog.}$ (Fig. 4C). After carrying out model selection process for classification of peptide sequences in S_{CB}^{sep} vs. S_{CB}^{tog}, the combination of the features$F_{1,2,4,6}$ was selected (N_B, N_L, L_{max} and $\overline{I_B}$). It leads an average classification accuracy of 61%, with notable false positive (32%) and false negative(45%) rates. It indicates that the model faces difficulties distinguishing peptide sequences in $S_{CB}^{tog.}$ vs. $S_{CB}^{sep.}$ (Fig. 4A). In addition, classification of peptide sequences in $S_{CB}^{sep.}$ class performs better than in $S_{CB}^{tog.}$ class in all evaluated metrics (precision, recall, and F1-score), with values larger than 0.5 in all metrics for $S_{CB}^{tog.}$ class and larger than 0.6 in all metrics for $S_{CB}^{sep.}$ class. Classification of peptide sequences in S_{BC}^{sep} vs. S_{BC}^{tog} showed better performances compared to S_{CB} group, with a feature combination comprised of $F_{3,4,5,6}$ (L_{min}, L_{max}, T_B and $\overline{I_B}$). The average classification accuracy was 70% and false positive and false negative rates were of 32% and 28%, respectively. For both sequences, precision, recall, and F1-score metrics show similar values around 0.7. These results indicate that classification models developed here present a good performance in predicting peptide sequences in the two subgroups of peptide sequences for which charged amino acids are separated or together in the sequence. However, this criterion appears to be less crucial than the relative position of charged amino acids within the sequence.

The second criterion tested here is based on the number of consecutive neutral amino acids (A) in the peptide sequence. Therefore, we separated peptide sequences into two subgroups: i) sequences with a maximum of two consecutive neutral amino acids, named S_{CB}^{2A} and S_{BC}^{2A}; ii) sequences with a maximum of four consecutive neutral amino acids, named S_{CB}^{4A} and S_{BC}^{4A} (Fig. 4B and D). The best combination of features to classify peptide sequences in S_{CB}^{4A} vs. S_{CB}^{2A} subgroups was $F_{2,3,4,5,6}$ (N_L, L_{min}, L_{max}, T_B and $\overline{I_B}$). Results show that peptide sequences in S_{CB}^{4A} are classified with precision of approximately 65 %, which indicates that the model has acceptable reliability for this subgroup of sequences. Nevertheless, peptide sequences in S_{CB}^{2A} were classified with a much lower precision around 40 %. Recall score for peptide sequences in S_{CB}^{4A} subgroup is very low (around 0.3), suggesting that the classifier struggles to correctly identify sequences in this subgroup, whereas recall scores for peptide sequences in S_{CB}^{2A} is significantly higher (around 80%, Fig. 4B). Additionally, average classification accuracy is 53% and false positive and false negative rates

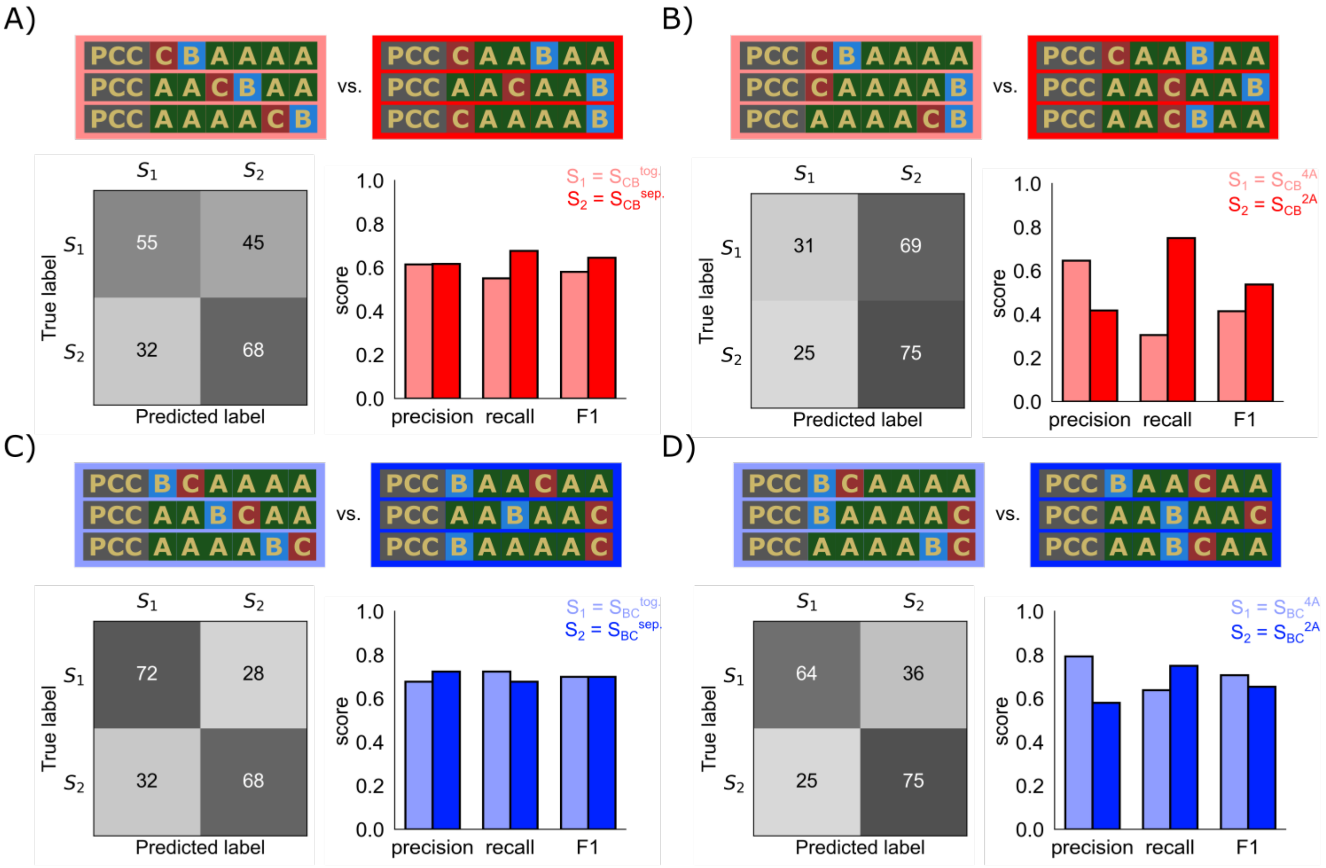

Figure 4. (A) Confusion matrices and precision, recall and F1-score metrics for binary classifications between peptide sequences in $S_{CB}^{tog.}$ vs. $S_{CB}^{sep.}$ group. (B) Confusion matrices and precision, recall and F1-score metrics for binary classifications between peptide sequences in S_{CB}^{4A} vs. S_{CB}^{2A} group. (C) Confusion matrices and precision, recall and F1-score metrics for binary classifications between peptide sequences in $S_{BC}^{tog.}$ vs. $S_{BC}^{sep.}$ group. (D) Confusion matrices and precision, recall and F1-score metrics for binary classifications between peptide sequences in S_{BC}^{4A} vs. S_{BC}^{2A} group.

are of 25% and 69%, respectively. It means that there is a significant percentage of peptide sequences in S_{CB}^{2A} that are not correctly classified and, more importantly, a large majority of peptide sequences in S_{CB}^{4A} are not correctly classified. Therefore, the classification model in this case is quite effective for peptide sequences in S_{CB}^{2A} subgroup but not for classifying peptide sequences in S_{CB}^{4A} subgroup. It leads to a very low model accuracy score (0.48). Overall, performances of the different classification models tested here show room for improvement, especially in terms of balance between S_{CB}^{2A} and S_{CB}^{4A} classes as the results indicate a significant imbalance in the performances of both classes. Therefore, optimization and tuning strategies, such as feature engineering, could be considered moving forward. On the other hand, the best combination of features to classify peptide sequences in S_{BC}^{4A} vs. S_{BC}^{2A} subgroups was F_1 (N_B) with an average classification accuracy of 69%. False positive and false negative rates were 25% and 36% respectively, suggesting that a significant percentage of traces in both classes are not correctly classified. The overall analysis indicates that peptide sequences S_{BC}^{2A} are much better classified across all metrics (precision, recall, and F1-score) compared to peptide sequences in S_{BC}^{4A}, which means that the model performs better for S_{BC}^{2A} than for S_{BC}^{4A} (Fig. 4D).

To conclude, based on the evaluation of classification scores and by taking into account the spacing between charged amino acids, we demonstrate that classifying peptide sequences in S_{BC} group based on the proximity or separation between the two charged amino acids in the sequence yields to better scores than classifying peptide sequences in S_{CB} group using the same criterion, with an accuracy of 0.70 and 0.62 respectively (Fig. 4A and C). Similarly, based on the number of consecutive neutral amino acids in peptide sequences, classifications scores show better performances to classify peptide sequences in S_{BC} group. This study indicates that when comparing both groups of peptide sequences, it is more likely to distinguish sequences in S_{BC} group according to both criteria of charged and neutral residue positions and without significant differences between the two criteria. On the opposite, peptide sequences in S_{CB} group are more sensitive to one criterion over the second, especially for sequences in S_{CB}^{4A} subgroup. According to our data, it would be more effective to use S_{BC} type sequences to identify 0 or 1 bits, as the results show very good scores for distinguishing them effectively using single-layer MoS_2 nanopores.

Classification of Peptide Sequence Motifs

Design of long peptide sequences capable of encoding binary digits continuously can be advanced through a method in which peptides are composed of specific amino acids to represent digital bits. Initially, raw data are encoded as long strings of 0s and 1s. These strings are then translated into sequences of amino acids, or peptides, according to predefined assignments. To retrieve the data, peptides are sequenced and the resulting sequences are converted back into binary digits, which are decoded to reproduce the original data. To enhance this approach, we explored the impact of amino acid motifs of different length on ionic current traces recorded during MD. Unlike the previous sections where the role of the order of each amino acid in the sequence was studied, each sequence having the same composition, this section focuses solely on the amino acid composition of sequence motifs, regardless of the order in the sequence.

Fluctuations observed in ionic current traces during peptide translocation through SSN necessitate a thorough understanding of the non-linear relationship between the amino acid presence inside the pore and the monitored ionic current. Molecular Dynamics is essential to establish this relationship since it precisely tracks Cartesian coordinates of all the atoms of the system and especially of the nanopore and of the peptide at each time step of simulation. To determine which amino acids of the peptide are inside the nanopore at a given time, a geometric criterion based on the volume of the amino acid inside the pore was employed. Each amino acid along the sequence was modelled as a sphere, centered at the center of mass of the amino acid and of radius $R_{a.a.}$, which is proportional to the volume of the amino acid. Then, if more than 30% of the volume of the amino acid is inside the pore at a given time, the amino acid is considered to be inside. This criterion enabled the extraction of information regarding the presence of a single or multiple amino acids simultaneously, called sequence motif, in single-layer MoS$_2$ nanopores. Moreover, we assigned ionic current values $I_c(t)$ to the presence of each sequence motif inside the pore at a given time t, based on the geometric criterion described above. In total, 19 different sequence motifs were identified and correspond to motifs made of: i) a single amino acid (A, B or C), ii) a pair of amino acids (AA, AB, AC or BC), iii) a triplet of amino acids (AAA, AAB, AAC or ABC), iv) a quartet of amino acids (AAAA, AAAB, AAAC or AACB), v) a quintet of amino acids (AAAAB, AAAAC or AAACB), and vi) a sextet of amino acids (AAAACB). Regarding the frequency of appearance of the different motifs in MD trajectories, single amino acid motifs were detected in all twelve sequences. As the number of amino acids per motif increases, the number of sequences where these motifs are present decreases. The most probable motifs extracted from the twelve peptide sequences were C, A, AA, and AC, with a frequency of 35%, 19%, 18%, and 13% of the total presence of all motifs, respectively (see Supplementary Material, Fig. S5). Fig. 5A depicts the distribution of ionic current associated with single amino acid motifs, which shows well-distinguished peaks despite overlapping, particularly for negatively charged amino acid C.

We performed multiclass classification tasks of peptide sequence motifs using the same strategy as before, but this time using ionic current values (see the probability densities of ionic current shown in Fig. 5A) associated with each motif as the only input variable. The aim here was to examine the influence of amino acid motifs of different lengths on ionic current traces recorded during MD at a shorter sequence length scale. This insight will be valuable for the future design of longer peptide sequences capable of encoding 0 and 1 bits within the same sequence which makes it crucial to select suitable amino acids to comprise the peptides. First, concerning motifs made of a single amino acid, we found that all three motifs A, B and C can be classified with an accuracy larger than 0.6. However, positively charged amino acid B shows issues with precision score. Furthermore, accuracy score of classification tasks degrades when identifying peptide sequence motifs consisting of two amino acids, as shown in Fig. 5B, with the motif AA being presenting the best score. This motif is, among all motifs made of two amino acids, characterized by the highest frequency during MD simulations (see Supplementary Materials, Fig. S5). Same trends persist for peptide sequence motifs made of three amino acids with accuracy scores not exceeding 0.56, which is also the accuracy score for peptide sequence motif AAB, the third most frequent motif made of two amino acids and identified from MD. Finally, by looking at all three metric scores shown in Fig. 5B (middle panel and right panel), only peptide sequence motif AA achieved good scores and peptide sequence motif AAC shows high precision score. The other motifs are characterized by low classification scores. However, by studying pairwise motif binary classification (Fig. 5C, D and E), accuracy, precision, recall and F1-scores increase significantly. For instance, Fig. 5C shows the different binary classifications performed between motifs made of a single amino acid. Classification between charged amino acids B vs. C shows the best classification scores among all three binary classifications, with an average accuracy score of 87%. In addition, precision, recall, and F1-score are quite large for negatively charged amino acid. However, positively charged amino acid B shows precision limitations, which involves a low F1-score. This is due to imbalance dataset between B and C classes (see Supplementary Materials, Table S3). On the other hand, the comparison between neutral amino acid A and negatively charged amino acid C presents a lower average classification score, with an accuracy of 75%. However, both amino acids present more balanced precision, recall, and F1-scores compared to the other two binary classifications. Finally, the comparison between neutral amino acid A and positively charged amino acid B shows an average classification accuracy of 72%, with excellent precision, recall, and F1-score for neutral amino acid A but low scores for positively charged amino acid B, especially in precision and F1-score.

For longer motifs made of two amino acids (Fig. 5D), the best average classification accuracy scores are obtained for AA vs. AB (84%), AB vs. AC (79%), and AA vs. BC (78%), with significant potential for designing longer peptide sequences capable of encoding binary digits. For these longer motifs, the precision is quite low for AA and AC due to imbalance dataset between the classes (see Supplementary Materials, Table S4). Classification of AC vs. BC peptide sequence motifs presents an average accuracy of 76%, with good recall and F1-scores for both motifs. It shows low precision for AA motif once again due to imbalanced dataset between the two classes. Classifications of peptide sequence motifs AB vs. BC and AA vs. AC show low accuracy for classes AC (48%) and AB (53%), despite dataset being relatively well-balanced between the classes and without significant overlap between ionic current distributions of both motifs. Last but not least, for peptide sequence motifs made of three amino acids (Fig. 5E), most of classification tasks trained and tested here lead to low accuracy scores for one of the classes. This may be due to a larger overlap between ionic current distributions of motifs (Fig. 5A, right panel) or imbalance in the dataset between the corresponding classes. Same observations were made for precision, recall, and F1-score metrics, for which there is a significant difference between the two classes. These results clearly indicate that it is much more difficulty for MoS$_2$ nanopores to detect sequence motifs made of three amino acids due to its sub-nm thickness and therefore representing binary data in peptide sequences using motifs made of three amino acids is not appropriate for 2D SSN.

CONCLUDING DISCUSSION

In this study, we performed MD simulations of the translocation of twelve different peptide sequences with the same composition (1 positively charged, 1 negatively charged and 4 neutral amino acids) through single-layer MoS$_2$ nanopores. By changing the configuration of the sequence, i.e. the position and spacing between charged amino acids, the goal was to explore the feasibility of differentiating between the twelve peptide sequences in order to design peptide sequences for binary encoding applications. From statistical dataset analysis and classification tasks using LightGBM classifier, we identified six promising features in ionic current time series recorded during MD, i.e. the number of BEs per simulation N_B (F$_1$), the number of ionic current levels per simulation N_L (F$_2$), the minimum level of ionic current within a simulation L_{min} (F$_3$), the maximum level of ionic current within a simulation L_{max} (F$_4$), the blockade duration per simulation T_B (F$_5$) and the mean blockade ionic current per simulation $\overline{I_B}$ (F$_6$). The corresponding feature subsets were further evaluated using four usual evaluation metrics for classifiers, i.e. accuracy, precision, recall and F1-score. First, our findings revealed the presence of two distinct groups of six sequences, determined by the relative position of the positively charged amino acid (B) compared to the negatively charged amino acid (C). This is explained by the fact that the direction of the electric field breaks the symmetry of the device with respect to the sign of charge transport. These groups of sequences were named S$_{CB}$ and S$_{BC}$ peptide sequence groups. Furthermore, as already shown in a previous work [36], this study highlights the significance of charge distribution along the peptide sequence on the discriminatory capacity for peptide sequencing through MoS$_2$

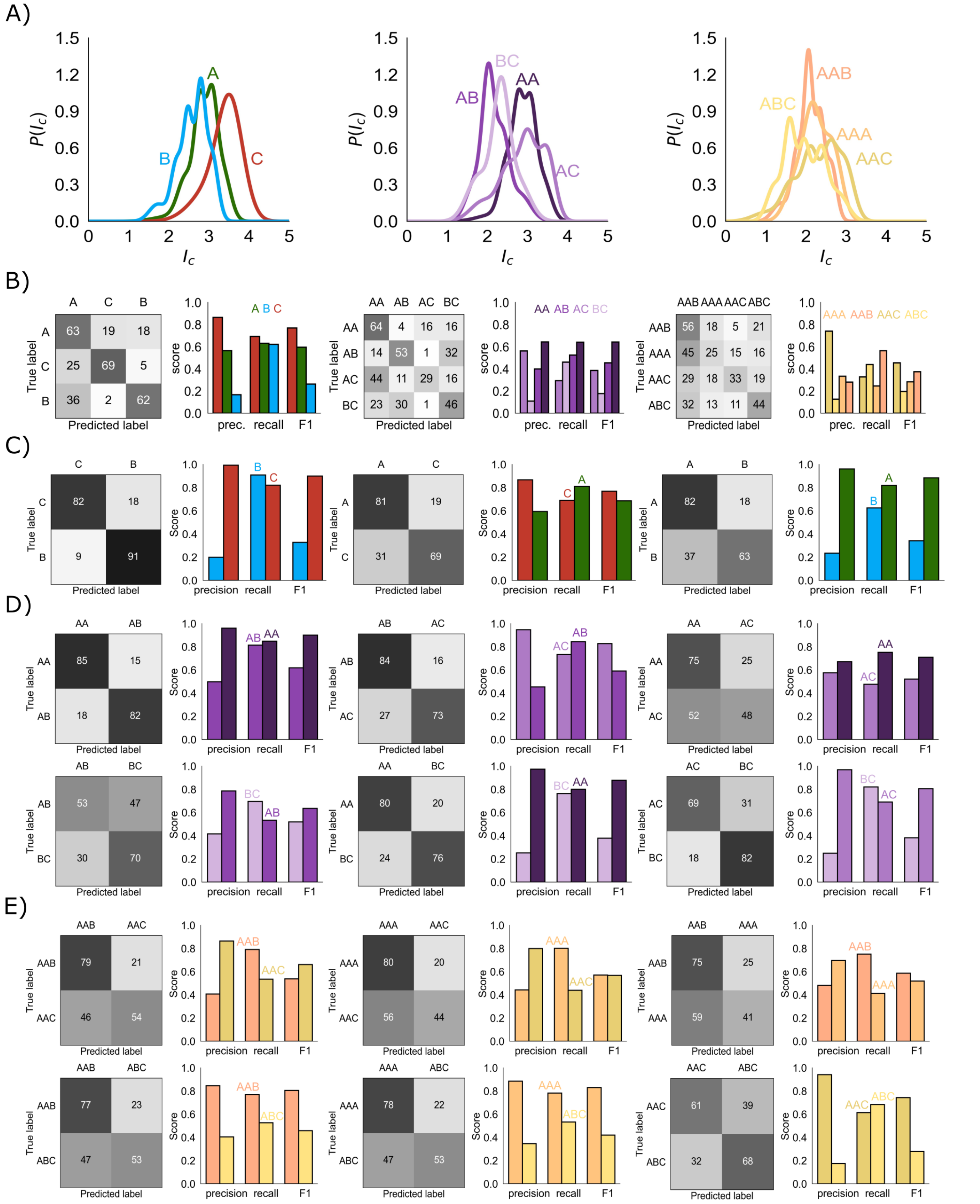

Figure 5. (A) Probability densities of ionic current $P(I_c)$ for peptide sequence motifs made of one (left panel), two (middle panel), or three amino acids (right panel), identified and extracted from MD trajectories. (B) Confusion matrices and precision, recall and F1-score metrics for tertiary and quaternary classifications between peptide sequence motifs made of one (left panel), two (middle panel) and three (right panel) amino acids. (C) Confusion matrices and precision, recall and F1-score metrics for binary classifications between peptide sequence motifs made of one amino acid (A, B, C). (D) Confusion matrices and precision, recall and F1-score metrics for binary classifications between peptide sequence motifs made of two amino acids (AA, AB, AC, BC). (E) Confusion matrices and precision, recall and F1-score metrics for binary classifications between peptide sequence motifs made of three amino acids (AAA, AAB, AAC, ABC).

nanopores.

Furthermore, we observed a strong correlation between discrimination accuracy and the separation in the sequence between charged amino acids, whether they are i) positioned apart from each other in the sequence; ii) adjacent to each other; or iii) depending on the number of adjacent neutral amino acids between them. This suggests a potential underlying mechanism influencing the detection capability of MoS$_2$ nanopore sensors for peptide sequencing. When classifying peptide sequences belonging to S$_{CB}$ group against their corresponding peptide sequences in S$_{BC}$ group (pairwise comparison), large classification accuracy scores were achieved. This is particularly true when charged amino acids are in the first position, whether the two charges are together such as CBAAAA vs. BCAAAA, or separated by four neutral charges such as CAAAAB vs. BAAAAC, as well as in the case of AACBAA vs. AABCAA. These findings highlight the critical roles of i) charged amino acid positions in the design of peptide sequences for binary data storage and ii) MoS$_2$ SSN ability to recognize the permutation of these charged amino acids within the sequence. Classification within each peptide sequence group based on the separation between charged amino acids reveals that S$_{CB}$ group presents the most challenging classification task due to its average accuracy score, while peptide sequences in S$_{BC}$ group are well classified, whether B and C amino acids are consecutive or separated by neutral amino acids in the sequence. This can be explained by the fact that, due to the direction of the electric field, forces on C and B act in opposite directions. Additionally, B is the amino acid that forms the PCC, introducing another source of asymmetry in the system, which influences both the applied force and the conformation of the peptide. Similarly, when considering the criterion of the length of separation by neutral amino acids, peptide sequences in S$_{CB}^{4A}$ group exhibit poor accuracy score. These findings indicate that classification of peptide sequences in S$_{BC}$ group generally outperforms classification of peptide sequences in S$_{CB}$ group. However, the criterion of separation between charged amino acids enhances precision within the classification of peptide sequences in S$_{CB}$ group. One of the most significant results of the present work is related to the importance of certain features extracted from ionic current traces in the classification of peptide sequences, primarily features F_3, i.e., L_{min}, and F_6, i.e., $\overline{I_B}$. These characteristics of peptide induced blockade events played a crucial role in the classifier that were ultimately selected, demonstrating that these features capture the dynamics ofblockade events in a single ionic current value through MD simulations.

The precise information provided by Molecular Dynamics is the exact position of the peptide and its amino acids as they translocate through the pore at any given time. It allows us to analyze in details peptide sequence translocations, focusing on the presence of amino acids inside the pore in order to correlate coordinates information with recorded ionic current. Therefore, we quantified the most frequent amino acid patterns within the pore, enabling more extensive extraction of ionic current data for further analysis. Peptide sequence motifs that were predominantly identified in the twelve sequences are made of one, two or three amino acids, for a total of eleven different motifs, namely A, B, C, AA, AB, AC, BC, AAA, AAB, AAC, ABC. Classification based on the length of these peptide sequence motifs showed that short motifs made of one amino acid in length exhibit much more distinct characteristics, which allow for better classification scores, whereas longer motifs may induce an increase of complexity or variability for such 2-D nanopores, leading to reduced classification performances. However, binary classification of peptide sequence motifs allowed us to determine which pairs of motifs could be differentiated. For motifs made of one amino acid, classification task shows excellent accuracy, particularly among charged amino acids, demonstrating a clear distinction between positively and negatively charged motifs. For motifs made of two or three amino acids, performances range from moderate to excellent, with some motifs of two-amino-acid length standing out, such as AA vs. BC, AA vs. AB, and AB vs. AC. Selection of pairs of shorter peptide sequence motifs that can be differentiated using 2-D SSN would enable in the future the design of longer sequences representing ’0’ and ’1’ bits . Our results suggest that sequence motifs made of one or two amino acids show great potential, particularly by comparison with motifs made of three amino acids. Sequence pairs (AA, AB) and (AB, AC) are among the best candidates for binary representations in longer peptide sequences as they show the best results in the classification tasks. Similarly, among the three binary classifications of motifs made of a single amino acid, sequence pairs (C, B) and (A, C) emerged as promising candidates.

Finally, results presented here propose various approaches for designing peptides that can be differentiated from each other, potentially serving as building blocks for data storage in biological molecules. Different criteria concerning the position of charged and neutral amino acids in the sequence as well as the spacing between charged amino acids could be used to design peptides that store 0 and 1 bits, contributing to the goal of synthesizing biological peptides made of amino acids for binary encoding applications. Exploring other structural features or modifying peptide sequences based on these findings may further enhance their potential use in molecular data storage applications since choosing classes of biological molecules that offer prolonged stability, with no energy required for storage, is one long-term objective of this area of research.

REFERENCES

[1] C. C. A. Ng, W. M. Tam, H. Yin, *et al.*, “Data storage using peptide sequences,” *Nature Communications*, vol. 12, no. 1, p. 4242, 2021.

[2] B. Sun, J. R. Kovatch, A. Badiong, and N. Merbouh, “Optimization and Modeling of Quadrupole Orbitrap Parameters for Sensitive Analysis toward Single-Cell Proteomics,” *Journal of Proteome Research*, vol. 16, no. 10, pp. 3711–3721, 2017.

[3] B. J. Cafferty, A. S. Ten, M. J. Fink, *et al.*, “Storage of information using small organic molecules,” *ACS Central Science*, vol. 5, no. 5, pp. 911–916, 2019.

[4] S. Chen, T. Lin, R. Basu, *et al.*, “Design of target specific peptide inhibitors using generative deep learning and molecular dynamics simulations,” *Nature Communications*, vol. 15, no. 1, p. 1611, 2024.

[5] S. Sankar, W. Yixin, M. Noor-A-Rahim, E. Gunawan, Y. L. Guan, and C. L. Poh, *D2sim: A computational simulator for nanopore sequencing based dna data storage*, 2024. DOI: 10.1101/2024.03.17.585393.

[6] E. G. S. Antonio, T. Heinis, L. Carteron, M. Dimopoulou, and M. Antonini, “Nanopore sequencing simulator for dna data storage,” in *2021 International Conference on Visual Communications and Image Processing (VCIP)*, 2021, pp. 1–5.

[7] K. Chen, J. Zhu, F. Bošković, and U. F. Keyser, “Nanopore-based dna hard drives for rewritable and secure data storage,” *Nano Letters*, vol. 20, no. 5, pp. 3754–3760, 2020. (visited on 05/08/2024).

[8] “Expanding the Molecular Alphabet of DNA-Based Data Storage Systems with Neural Network Nanopore Readout Processing,” *Nano Lett*, vol. 22, no. 5, pp. 1905–1914, 2022.

[9] K. Chen, J. Kong, J. Zhu, N. Ermann, P. Predki, and U. F. Keyser, “Digital data storage using dna nanostructures and solid-state nanopores,” *Nano Letters*, vol. 19, no. 2, pp. 1210–1215, 2019. (visited on 05/08/2024).

[10] R. Lopez, Y.-J. Chen, S. Dumas Ang, *et al.*, “Dna assembly for nanopore data storage readout,” *Nature Communications*, vol. 10, no. 1, p. 2933, 2019.

[11] B. Hamoum, E. Dupraz, L. Conde-Canencia, and D. Lavenier, "Channel model with memory for dna data storage with nanopore sequencing," in *2021 11th International Symposium on Topics in Coding (ISTC)*, 2021, pp. 1–5.

[12] B. Hamoum and E. Dupraz, "Channel Model and Decoder With Memory for DNA Data Storage With Nanopore Sequencing," *IEEE Access*, vol. 11, pp. 52 075–52 087, 2023.

[13] R. Hulett, S. Chandak, and M. Wootters, "On coding for an abstracted nanopore channel for dna storage," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 2465–2470.

[14] S. M. H. T. Yazdi, R. Gabrys, and O. Milenkovic, "Portable and Error-Free DNA-Based Data Storage," *Scientific Reports*, vol. 7, no. 1, p. 5011, 2017. (visited on 05/23/2024).

[15] L. Organick, Y.-J. Chen, S. Dumas Ang, *et al.*, "Probing the physical limits of reliable dna data retrieval," *Nature Communications*, vol. 11, no. 1, p. 616, 2020.

[16] S. A. Trauger, E. P. Go, Z. Shen, *et al.*, "High Sensitivity and Analyte Capture with Desorption/Ionization Mass Spectrometry on Silylated Porous Silicon," *Analytical Chemistry*, vol. 76, no. 15, pp. 4484–4489, 2004.

[17] N. Callahan, J. Tullman, Z. Kelman, and J. Marino, "Strategies for Development of a Next-Generation Protein Sequencing Platform," *Trends in Biochemical Sciences*, vol. 45, no. 1, pp. 76–89, 2020.

[18] A. Isidro-Llobet, M. N. Kenworthy, S. Mukherjee, *et al.*, "Sustainability challenges in peptide synthesis and purification: From r&d to production," *The Journal of Organic Chemistry*, vol. 84, no. 8, pp. 4615–4628, 2019.

[19] M. Murby, E. Samuelsson, T. N. Nguyen, *et al.*, "Hydrophobicity engineering to increase solubility and stability of a recombinant protein from respiratory syncytial virus," *European Journal of Biochemistry*, vol. 230, no. 1, pp. 38–44, 1995.

[20] M. Ptak-Kaczor, M. Banach, K. Stapor, P. Fabian, L. Konieczny, and I. Roterman, "Solubility and aggregation of selected proteins interpreted on the basis of hydrophobicity distribution," *International Journal of Molecular Sciences*, vol. 22, no. 9, p. 5002, 2021. (visited on 05/24/2024).

[21] L. K. Mosavi and Z. Peng, "Structure-based substitutions for increased solubility of a designed protein," *Protein Engineering, Design and Selection*, vol. 16, no. 10, pp. 739–745, 2003. (visited on 05/24/2024).

[22] Y. Kuroda, A. Suenaga, Y. Sato, S. Kosuda, and M. Taiji, "All-atom molecular dynamics analysis of multi-peptide systems reproduces peptide solubility in line with experimental observations," *Scientific Reports*, vol. 6, p. 19 479, 2016.

[23] G. A. Grant, "Synthetic peptides for production of antibodies that recognize intact proteins," *Current Protocols in Molecular Biology*, vol. 59, no. 1, pp. 11.16.1–11.16.19, 2002.

[24] J. M. Jensen, V. Xue, L. Stretz, T. Mandal, L. L. Reich, and A. E. Keating, "Peptide design by optimization on a data-parameterized protein interaction landscape," *Proceedings of the National Academy of Sciences*, vol. 115, no. 44, E10342–E10351, 2018. (visited on 05/24/2024).

[25] A. Díaz Carral, M. Ostertag, and M. Fyta, "Deep learning for nanopore ionic current blockades," *The Journal of Chemical Physics*, vol. 154, no. 4, p. 044 111, 2021.

[26] M. Sakamoto, K. Hori, and T. Yamamoto, "Machine-learning-assisted bacteria identification in ac nanopore measurement," *Sensors and Materials*, vol. 35, no. 9, pp. 3161–3171, 2023.

[27] M. Tsutsui, T. Takaai, K. Yokota, T. Kawai, and T. Washio, "Deep learning-enhanced nanopore sensing of single-nanoparticle translocation dynamics," *Small Methods*, vol. 5, no. 7, 2021.

[28] M. K. Jena and B. Pathak, "Development of an artificially intelligent nanopore for high-throughput dna sequencing with a machine-learning-aided quantum-tunneling approach," *Nano Letters*, vol. 23, no. 7, pp. 2511–2521, 2023.

[29] S. J. Greive, L. Bacri, B. Cressiot, and J. Pelta, "Identification of conformational variants for bradykinin biomarker peptides from a biofluid using a nanopore and machine learning," *ACS Nano*, vol. 18, no. 1, pp. 539–550, 2023.

[30] M. Nykrynova, V. Barton, R. Jakubicek, M. Bezdicek, M. Lengerova, and H. Skutkova, *Using deep learning for gene detection and classification in raw nanopore signals*, 2021. DOI: 10.1101/2021.12.23.473143.

[31] Y. K. Wan, C. Hendra, P. N. Pratanwanich, and J. Göke, "Beyond sequencing: Machine learning algorithms extract biology hidden in nanopore signal data," *Trends in Genetics*, vol. 38, no. 3, pp. 246–257, 2022.

[32] Y. Bao, J. Wadden, J. R. Erb-Downward, *et al.*, "Squigglenet: Real-time, direct classification of nanopore signals," *Genome Biology*, vol. 22, no. 1, p. 298, 2021.

[33] M. Nykrynova, R. Jakubicek, V. Barton, M. Bezdicek, M. Lengerova, and H. Skutkova, "Using deep learning for gene detection and classification in raw nanopore signals," *Frontiers in Microbiology*, vol. 13, 2022.

[34] C. Cao, L. F. Krapp, A. Al Ouahabi, *et al.*, "Aerolysin nanopores decode digital information stored in tailored macromolecular analytes," *Science Advances*, vol. 6, no. 50, eabc2661, 2020.

[35] L. Organick, S. D. Ang, Y.-J. Chen, *et al.*, "Random access in large-scale dna data storage," *Nature Biotechnology*, vol. 36, no. 3, pp. 242–248, 2018.

[36] A. Urquiola Hernández, P. Delarue, C. Guyeux, A. Nicolaï, and P. Senet, "Single-layer mos2 solid-state nanopores for coarse-grained sequencing of proteins," *Frontiers in Nanotechnology*, vol. 5, 2023.

[37] A. Nicolaï, M. D. Barrios Pérez, P. Delarue, V. Meunier, M. Drndić, and P. Senet, "Molecular dynamics investigation of polylysine peptide translocation through mos2 nanopores," *The Journal of Physical Chemistry B*, vol. 123, no. 10, pp. 2342–2353, 2019.

[38] M. J. Abraham, D. van der Spoel, E. Lindahl, B. Hess, and the GROMACS development team, *GROMACS User Manual version 2018.2*, 2018. [Online]. Available: www.gromacs.org.

[39] R. B. Best, D. de Sancho, and J. Mittal, "Residue-specific alpha-helix propensities from molecular simulation," *Biophysical Journal*, vol. 102, no. 6, pp. 1462–1467, 2012.

[40] A. Nicolaï, A. Rath, P. Delarue, and P. Senet, "Nanopore sensing of single-biomolecules: A new procedure to identify protein sequence motifs from molecular dynamics," *Nanoscale*, vol. 12, no. 44, pp. 22 743–22 753, 2020.

[41] G. C. Chow, "Tests of equality between sets of coefficients in two linear regressions," *Econometrica*, vol. 28, no. 3, pp. 591–605, 1960.

[42] H. Lee, "Using the chow test to analyze regression discontinuities," *Tutorials in Quantitative Methods for Psychology*, vol. 4, 2008.

[43] G. Ke, Q. Meng, T. Finley, *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017.