

## ARTICLE OPEN

Identification of amino acids with sensitive nanoporous MoS<sub>2</sub>: towards machine learning-based predictionAmir Barati Farimani<sup>1,2,3</sup>, Mohammad Heiranian<sup>1,2</sup> and Narayana R. Aluru<sup>1,2</sup>

Protein detection plays a key role in determining the single point mutations which can cause a variety of diseases. Nanopore sequencing provides a label-free, single base, fast and long reading platform, which makes it amenable for personalized medicine. A challenge facing nanopore technology is the noise in ionic current. Here, we show that a nanoporous single-layer molybdenum disulfide (MoS<sub>2</sub>) can detect individual amino acids in a polypeptide chain (16 units) with a high accuracy and distinguishability. Using extensive molecular dynamics simulations (with a total aggregate simulation time of 66 μs) and machine learning techniques, we featurize and cluster the ionic current and residence time of the 20 amino acids and identify the fingerprints of the signals. Using logistic regression, nearest neighbor, and random forest classifiers, the sensor reading is predicted with an accuracy of 72.45, 94.55, and 99.6%, respectively. In addition, using advanced ML classification techniques, we are able to theoretically predict over 2.8 million hypothetical sensor readings' amino acid types.

*npj 2D Materials and Applications* (2018)2:14; doi:10.1038/s41699-018-0060-8

## INTRODUCTION

In recent years, nanopore-sequencing technology has matured to an extent where it is now used in the industry (e.g., Oxford Nanopore, Genia) as a long-read, single base resolution, and label-free detection platform. The parallel use of thousands of pores and the ability to read a chain of DNA thousands of times empower the sequencing technology with single base resolution. Specifically, with the advances in data alignment, machine learning and data processing, the fast and long-read sequencing features obtained from the nanopore-sequencing technology have become practicable since multiple readings and statistical learning can alleviate the problem of noise in detection.<sup>1–5</sup> Ionic current and residence time (signal bandwidth) of each base in the nanopore are the primary detection signatures, which are used as features in data classification.<sup>6–8</sup> However, the fact that these signals overlap with each other for different DNA bases, has inspired scientists to propose other signal types or to functionalize the existing nanopores and discover new materials for improved signal resolution.<sup>9,10</sup> The conventional Si<sub>3</sub>N<sub>4</sub> pore and other solid state nanopores have a thickness of several nanometers<sup>2</sup> limiting the single base recognition of biological molecules.<sup>3,6</sup> The thickness of these membranes results in the encapsulation of multiple DNA/amino acid bases/residues, making it difficult to decompose the signal associated with each base.<sup>11–15</sup>

A single-layer molybdenum disulfide (MoS<sub>2</sub>) which is a 2-dimensional (2D) material with a three-atom thick lattice of Mo and S atoms has an appropriate bandgap for use as an electronic transistor.<sup>16</sup> Recent experimental<sup>17,18</sup> and computational<sup>19</sup> work has shown that a MoS<sub>2</sub> nanopore has a high signal-to-noise ratio for DNA base detection making it an attractive material for biological sensing. In addition, the stickiness of the biological molecules to the mouth of MoS<sub>2</sub> nanopore can be significantly

lowered by irradiation of sulfur atoms creating a molybdenum dominated pore.<sup>18</sup> Compared to other 2D materials (graphene and boron nitride), MoS<sub>2</sub> nanopores can operate without degradation for an extended duration of time.<sup>18,20</sup>

Many diseases including cancer, diabetes and digestive disorders are caused by malfunctioning of ribosomes (Ribosomopathies).<sup>21</sup> As a result, the amino acid chain might have a faulty transcription from DNA. Therefore, identification of an amino acid chain is necessary for diagnostic purposes and early stage of cancer/other disease detection.<sup>22</sup> In fact, the data acquired from proteomic fingerprints can be more trustworthy in defining the health status of humans than the genome; but it is important to note that acquiring and analyzing such data is more difficult.<sup>23</sup>

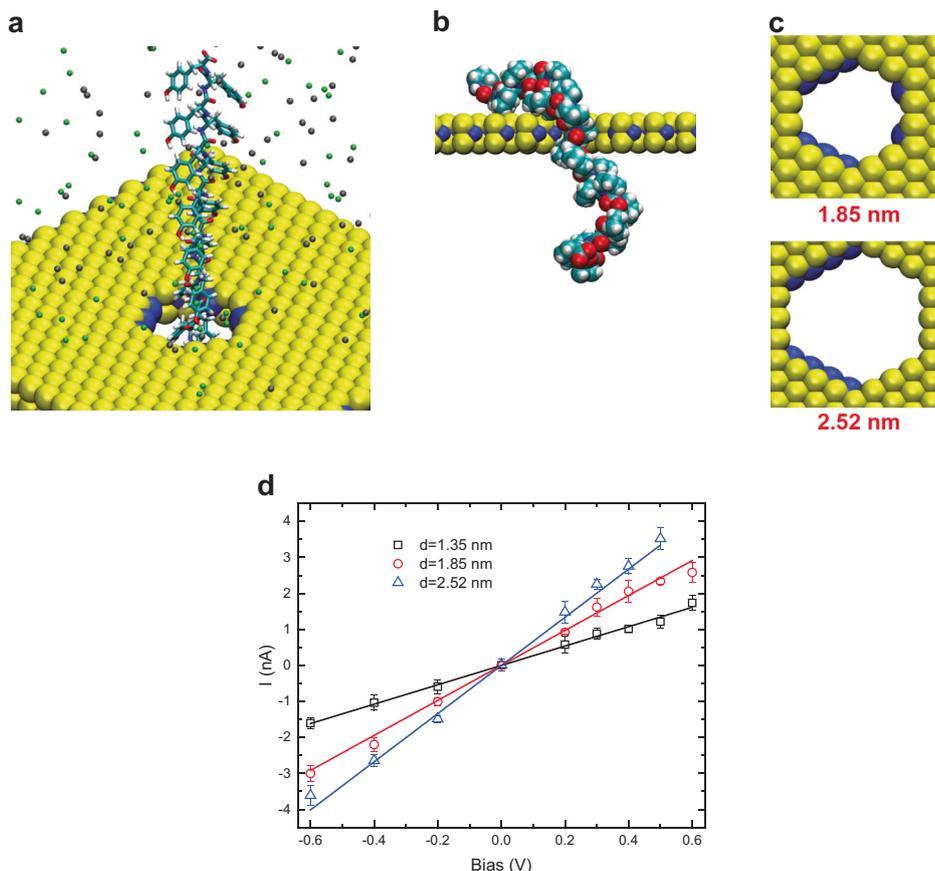
Recent work has used graphene nanopores for protein identification using molecular dynamics (MD) simulations where stepwise translocation of polypeptides has been shown. These results make interpretation of ionic current signal feasible.<sup>24</sup> Kolmogorov et al.<sup>25</sup> used machine learning (random forest and support vector machines) to process the ionic current signals obtained from solid-state nanopore sequencing of a polypeptide chain. They showed that the signals obtained using a sub-nanometer pore is sensitive enough for the detection of protein sequence.

Given the significance of amino acids identification in health diagnostics and the potential of MoS<sub>2</sub> material, we characterized the ionic current and residence time associated with the 20 standard amino acids by translocating them through a single-layer MoS<sub>2</sub> nanopore using extensive simulations. Supervised and unsupervised machine learning and classification techniques were used to classify and detect signals with a high prediction accuracy of up to 99.6%.

<sup>1</sup>Department of Mechanical Science and Engineering, Beckman Institute for Advanced Science and Technology, Urbana, IL 61801, USA; <sup>2</sup>University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA and <sup>3</sup>Department of Bioengineering, Stanford University, Stanford, CA 94305, USA  
Correspondence: Narayana R. Aluru (aluru@illinois.edu)

Received: 8 November 2017 Revised: 2 April 2018 Accepted: 24 April 2018

Published online: 24 May 2018



**Fig. 1** Schematic of the simulations and the I-V curve. **a** Simulation set up for the polypeptide chain with 16 units (here, TYR (Y)), MoS<sub>2</sub> nanopore, and ions. Water is not shown. Mo and S atoms are in blue and yellow, respectively. **b** A snapshot of Proline polypeptide translocation through the MoS<sub>2</sub> nanopore and its conformational change. **c** Comparison of the 1.85 and 2.52 nm diameter nanopores and the arrangement of Mo and S atoms at the edge of the pores. **d** I-V characteristics of MoS<sub>2</sub> nanopores with pore diameters of 1.35, 1.85, and 2.52 nm (the error bars were generated with four simulation trajectories for each data point)

## RESULTS AND DISCUSSION

By performing detailed MD simulations, we demonstrate that a single-layer nanoporous MoS<sub>2</sub> can identify individual amino acids with high accuracy. Figure 1 shows the schematic of a typical simulation box consisting of a single-layer MoS<sub>2</sub> membrane (with the pores considered), an amino acid chain (16 identical residues), water and ions. The water molecules are not shown in the figure.

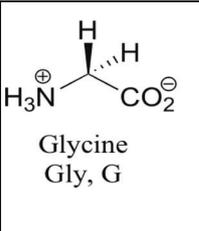
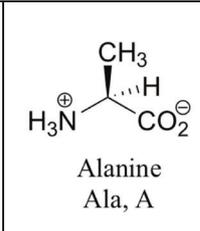
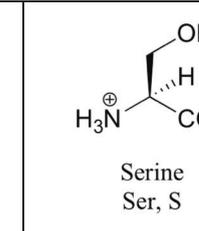
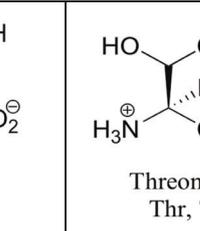
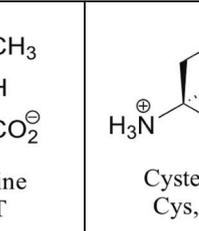
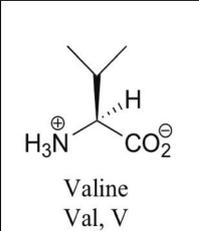
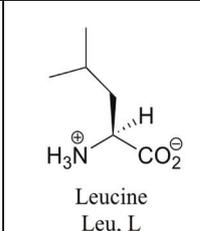
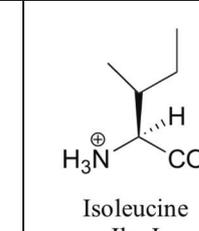
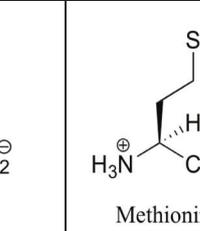
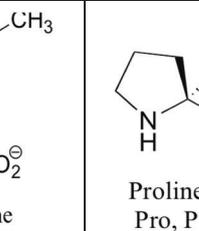
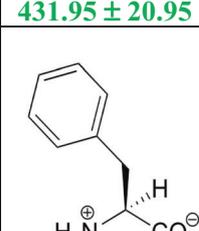
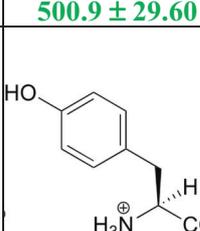
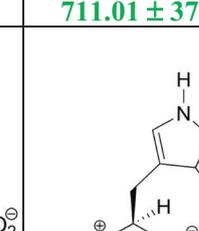
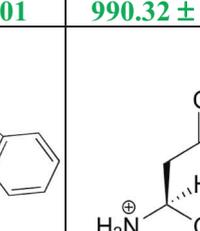
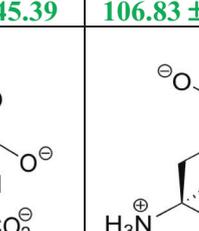
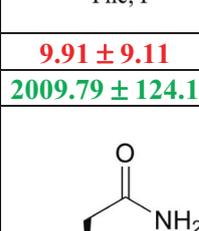
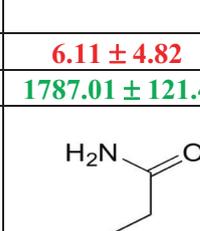
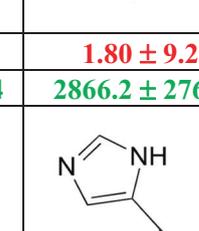
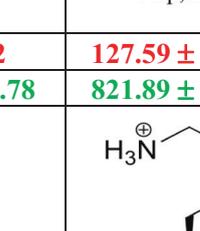
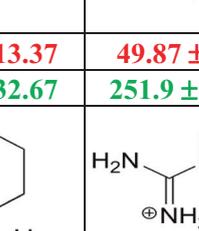
### Ionic currents and residence times

To characterize the ionic current characteristics for different pore sizes and validate the simulations against experimentally known properties (e.g., conductance), we performed MD simulations with a bare pore (with no amino acid chain in the system) applying both positive and negative biases. The I-V curves for different sizes of the nanopore in MoS<sub>2</sub>, in the absence of the polypeptide chains, show an Ohmic behavior (Fig. 1d). The conductance (G) is computed to be 2.69, 4.96, and 10.64 nS for the 1.35, 1.85, and 2.52 nm diameter pores, respectively. These conductance values are in good agreement with the experimentally measured conductance of MoS<sub>2</sub> nanopores with similar pore diameters.<sup>26</sup> In addition, the conductance values obtained from the model developed by Wannanu et al.<sup>27</sup> are comparable to the values computed in the simulations.

The amino acids are translocated through the nanopore using an external force under a constant bias of 200 mV. All the production simulations were performed for the pore with a

diameter of 1.85 nm. The average values of ionic currents and residence times for all the amino acids are tabulated in Fig. 2. The largest ionic current values (with 55–65% blockade) are observed for the amino acids (G, A, S, T, C) in the first row of Fig. 2. G, A, S, T, C are among the smallest amino acids based on the combined vdW radii.<sup>28</sup> G, A and S exhibited a very fast translocation with an average residence time per amino acid of 33.02, 64.52, and 23.10 ps, respectively. Of all the amino acids in Fig. 2, Methionine exhibits an inconsistent ionic current of  $-18.21$  pA. We investigated this anomaly by probing further into the components of the current (anion and cation contributions). We observed an intermittent anion binding to the thiol groups of Methionine resulting in a strong negative ionic current due to the fact that the Methionine is pulled by the external forces and drags the anions in the direction of the electric field (see the supporting information).

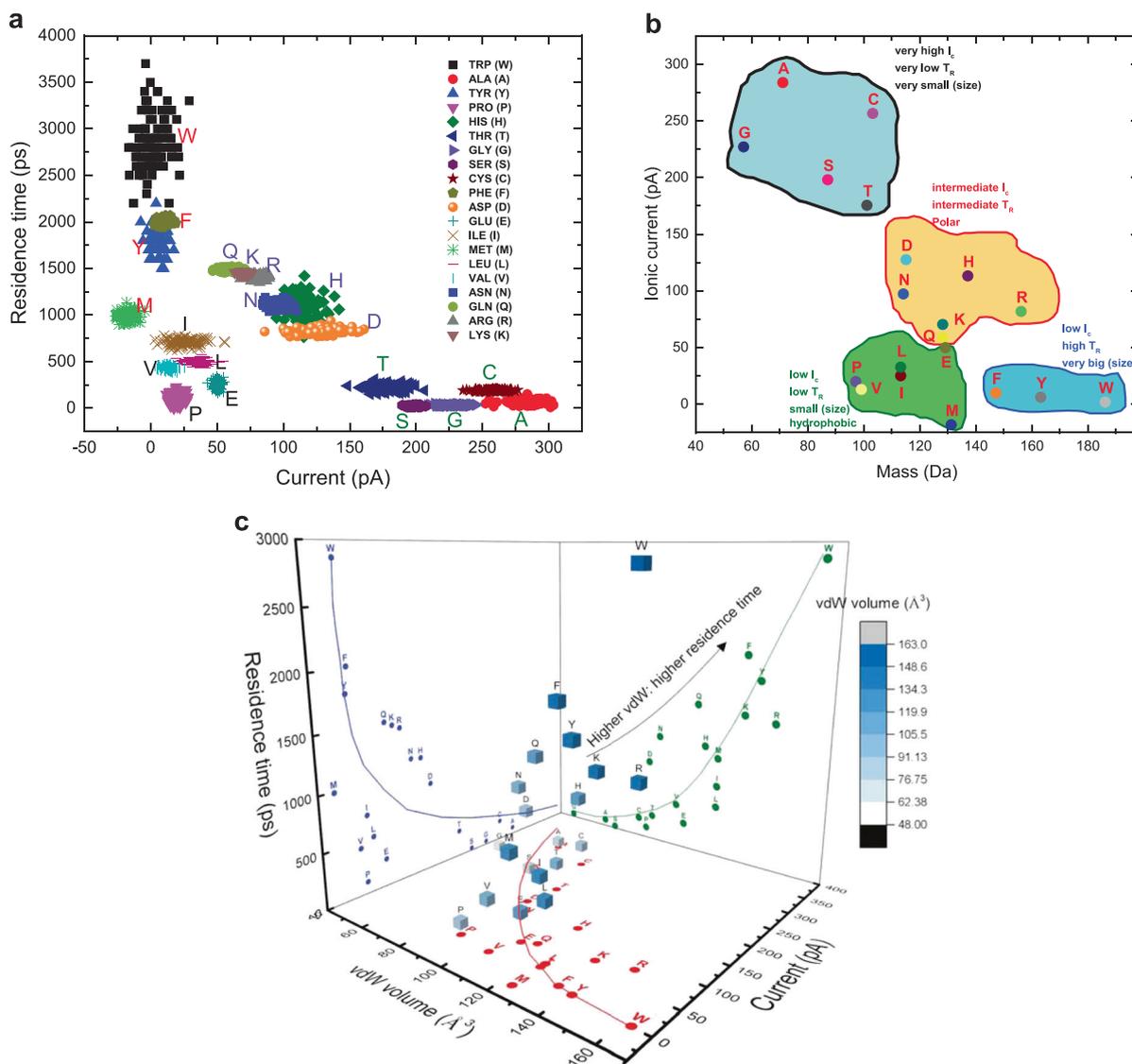
In Fig. 3a, the residence time and ionic current statistics for all the 20 amino acids are shown. L, V, E, M, P, and I amino acids show similar ionic current and residence time statistics. Their residence time is between  $\sim 100$  and  $\sim 1000$  ps and their ionic current is between  $\sim 13$  and  $\sim 50$  pA (except for the anomalous negative current of M amino acid). These amino acids have both smaller sizes and higher residence times compared to G, A, S, T, and C amino acids. The common feature among this class of amino acids is their hydrophobic nature (Fig. 3b). A 3D scattered plot of the vdW volume of each amino acid (vdW volumes are taken from ref. <sup>28</sup>) as a function of its residence time and ionic current is shown in

 Glycine Gly, G	 Alanine Ala, A	 Serine Ser, S	 Threonine Thr, T	 Cysteine Cys, C
<b>227.14 ± 9.56</b>	<b>283.94 ± 9.97</b>	<b>198.12 ± 3.76</b>	<b>175.5 ± 10.41</b>	<b>256.68 ± 8.84</b>
<b>33.02 ± 8.76</b>	<b>64.52 ± 25.66</b>	<b>23.1 ± 6.69</b>	<b>249.92 ± 31.26</b>	<b>190.52 ± 8.25</b>
 Valine Val, V	 Leucine Leu, L	 Isoleucine Ile, I	 Methionine Met, M	 Proline Pro, P
<b>13.02 ± 4.07</b>	<b>32.72 ± 6.52</b>	<b>25.11 ± 9.87</b>	<b>-18.21 ± 4.60</b>	<b>19.92 ± 3.68</b>
<b>431.95 ± 20.95</b>	<b>500.9 ± 29.60</b>	<b>711.01 ± 37.01</b>	<b>990.32 ± 45.39</b>	<b>106.83 ± 124.1</b>
 Phenylalanine Phe, F	 Tyrosine Tyr, Y	 Tryptophan Trp, W	 Aspartate Asp, D	 Glutamate Glu, E
<b>9.91 ± 9.11</b>	<b>6.11 ± 4.82</b>	<b>1.80 ± 9.22</b>	<b>127.59 ± 13.37</b>	<b>49.87 ± 2.31</b>
<b>2009.79 ± 124.1</b>	<b>1787.01 ± 121.4</b>	<b>2866.2 ± 276.78</b>	<b>821.89 ± 32.67</b>	<b>251.9 ± 37.01</b>
 Asparagine Asn, N	 Glutamine Gln, Q	 Histidine His, H	 Lysine Lys, K	 Arginine Arg, R
<b>97.28 ± 5.31</b>	<b>58.65 ± 5.41</b>	<b>113.2 ± 9.86</b>	<b>70.33 ± 2.27</b>	<b>81.86 ± 4.75</b>
<b>1105.78 ± 33.59</b>	<b>1489.81 ± 17.25</b>	<b>1091.21 ± 113.41</b>	<b>1450 ± 9.34</b>	<b>1410.91 ± 20.99</b>

**Fig. 2** Ionic current and residence time. Ionic current (in red, first number and in pA) and residence time (in green, second number and in ps per residue) associated with each amino acid. Each residence time value is an average over 100 simulations with an applied force of 0.7643 pN, and each current value is an average over the same 100 simulations and an additional set of 100 simulations with different applied forces. The pristine (bare) pore current is  $571 \pm 2.11$  pA. All these simulations were performed for the pore with a diameter of 1.85 nm

Fig. 3c (a similar plot is available in the supporting information for the mass of amino acids). The obvious trend in this data is the correlation of the volume of amino acids (or mass) with their ionic current, and residence time (higher volume (or mass) mostly leads to higher residence time and lower current). We find that the residence time ( $T_R$ , ps) and the mass ( $M$ , Da) are related by a power law relation,  $T_R = 4.66(10^{-9})M^{2.339}$  (supporting information).

Larger amino acids such as W, Y and F have larger residence times ( $\sim 1500$  ps  $< T_R < \sim 4000$  ps) and low ionic currents ( $< 10$  pA, where the bare pore current is  $571 \pm 2.1$  pA). W, Y, and F have a phenyl group that occupies most of the pore (the phenyl group alone has a vdW diameter of 7.5 Å) and the blockade is higher (see the supporting information) for these amino acids. Specifically, W contains an  $\alpha$ -amino group, an  $\alpha$ -carboxylic acid group, and a side



**Fig. 3** Physical properties of amino acids based on the ionic currents and residence times. **a** Ionic current and residence time data for 20 amino acids with their respective labels (see Fig. 2 for the errors). **b** Fingerprints and classification of amino acids categorized based on the strength of their ionic current and residence time, size (mass), and chemical properties. **c** 3D plot of vdW volume ( $\text{\AA}^3$ ) for each amino acid vs. their average residence time (ps) and ionic current (pA) (the 3D data is presented by the squares). The squares are provided with a clearer vdW volume presentation. Residence time-current data is shown in blue color, vdW-current is shown in red color and vdW-residence time data is shown in green color. Amino acids with higher volumes (or masses) exhibit higher residence time and lower current

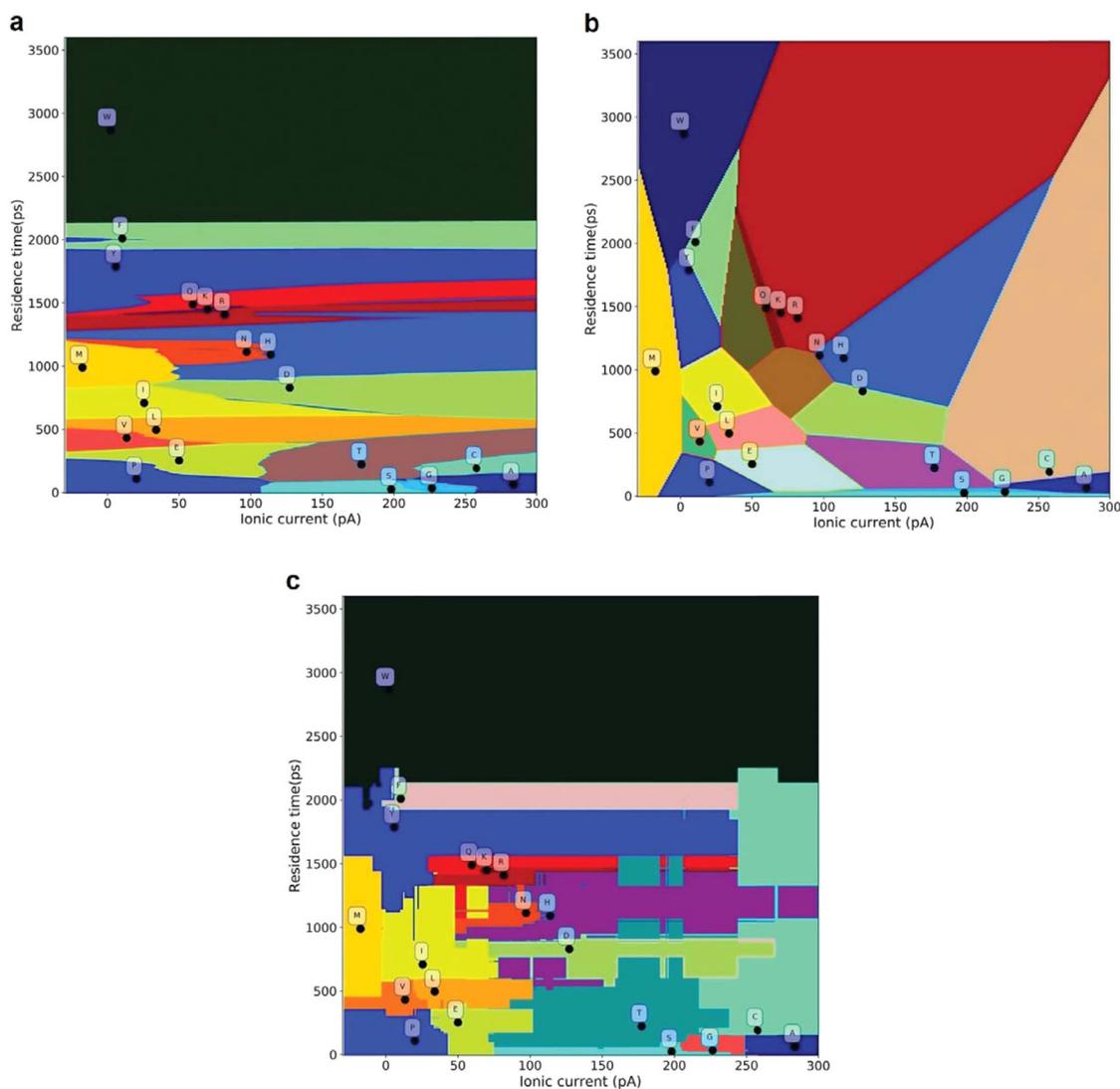
chain indole, making it the largest amino acid. W's ionic current (1.8 pA, with 99.9% pore blockade) depicts the larger size of this amino acid (Fig. 3c). In the supporting information, we compared the blockade of the  $\text{MoS}_2$  nanopore in different configurations of ALA (A) and TRP (W) inside the pore.

Glutamate (E) and Tryptophan (W) have the lowest and highest standard deviation (STD) of both ionic currents and residence times among all amino acids, respectively. Amino acids N, H, D and also Q, K and R have both intermediate ionic current and residence time ( $\sim 50 \text{ pA} < I_c < \sim 200 \text{ pA}$  and  $\sim 500 \text{ ps} < T_R < \sim 2000 \text{ ps}$ ). For these amino acids with intermediate residence time and ionic current, the common feature is their polar behavior (Fig. 3b). Based on the  $I_c$  and  $T_R$  values, we classified the amino acids into very small (G, A, S, T, C: very low  $T_R$  and High  $I_c$ ), very big with phenyl groups (F, W, Y: very high  $T_R$  and very low  $I_c$ ), small and hydrophobic (L, V, P, M, E, I: low  $T_R$  and low  $I_c$ ) and polar (N, H, D, Q, K, R: intermediate  $I_c$  and intermediate  $T_R$ ) as shown in Fig. 3b.

It is noteworthy that these simulations are for short homopeptides. In protein identification platforms with different amino acids in a sequence, the relative distance between the cluster centers can also be used to identify an amino acid. To this end, we can use ML with extrapolative properties.

#### Machine learning (ML)-based prediction

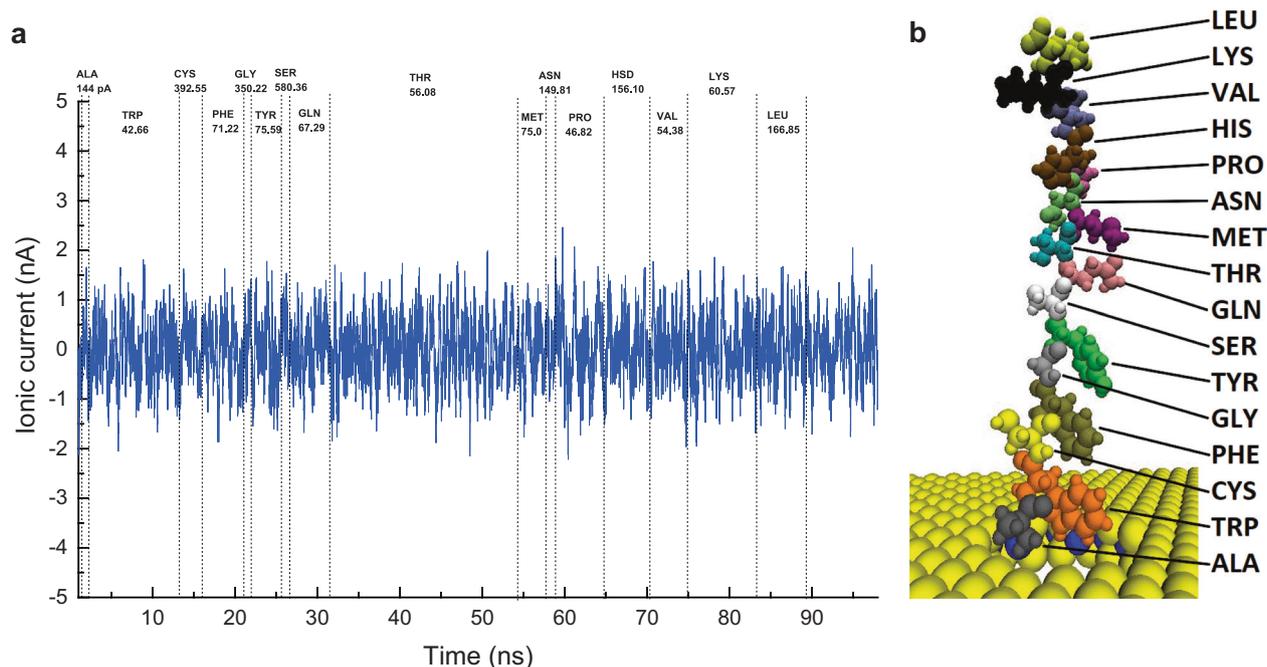
In practice, the data acquired from amino acid translocation through a  $\text{MoS}_2$  nanopore is unlabeled and a priori information is not available on the sensing ability (i.e., identifying the type of amino acid) of the nanopore. Here, we employ ML-based models to predict the amino acid type if the ionic current and residence time are known from nanopore detection. In addition, we also attempt to identify the most accurate and predictive ML model. To estimate the accuracy of amino acid class (or type) prediction of each model, the Monte Carlo Cross Validation (MCCV) algorithm is



**Fig. 4** Machine learning classification. Comparison of different machine learning models and their prediction capability in mapping the ionic current-residence time landscape. Each colored region represents an amino acid type (each letter indicates the type of amino acid) predicted by the models using the training data. Note that the colors in each plot are not correlated to the other ones. The solid dots along with the labels on the plots represent the mean values of the actual data (presented in Fig. 3a). **a** Prediction based on Nearest Neighbor Model with  $k = 3$  **b** Prediction based on Logistic Regression (large red region belongs to the class of R). **c** Prediction based on Random Forest Model with the number of estimators = 9

used. The MCCV randomly splits the dataset into test (unlabeled) and train sets (labeled). We split the dataset keeping 80% of it for training and the rest for testing the prediction by the trained models. In the training process, we used three ML models, namely, k-nearest neighbor, logistic regression and random forest, available in the Scikit-learn machine learning package in Python.<sup>29</sup> Initially, using each model, the training was performed with the labeled data from all 20 amino acids having 100 data points each for the current and residence time. The details on the three ML models and the optimization technique to determine the parameters are described in the supporting information. To identify and predict the class of a future reading (amino acid translocation data from the nanopore), we generated a grid with mesh size  $0.5 \times 0.5$  (pA, ps) with the upper and lower bounds of the current (along x-axis) and residence time (along y-axis). The upper and lower bounds of ionic current and residence time (current:  $-45$ – $330$  pA, residence time:  $0.1$ – $3800$  ps) can be obtained from Fig. 3a. To account for data at the boundaries, we extended the upper and lower limits by 2 pA and 2 ps. A total

number of 2,885,714 grid squares were generated in current/residence time space. Each grid square can be thought of as a hypothetical sensor reading. Many of these grid squares in the top right corner of the current/residence time plot are beyond the scope of the available data; therefore, the practical number of hypothetical readings is lower in actuality. The amino acid class/type of each grid square is predicted using all the three ML models. The prediction landscapes are shown in Fig. 4. The prediction is denoted by different colors for different amino acid types. The mean values of the labeled data (from Fig. 3a) are marked on each figure to compare with the prediction. The prediction of amino acid class regions with k-nearest neighbor is affected by different scales of ionic current (two orders of magnitude variation) and residence times (three orders of magnitude variation) (Fig. 4a). Since the scale of residence time is higher than that of ionic current, the prediction of the region associated with each amino acid is dominated by the residence time data (Fig. 4a). Logistic regression shows smooth and straight decision boundaries (Fig. 4b) whereas random forest shows



**Fig. 5** Heterogeneous chain of amino acids. **a** The ionic current as a function of time for one of the simulations of the mixed amino acids chain. The translocation of each amino acid, while inside the pore, is specified by two dotted lines. **b** 16 different amino acids are shown in different colors

flexible and mixed decision boundaries (Fig. 4c). The accuracy of logistic regression,  $k$ -nearest neighbor and random forest is 72.45, 94.55, and 99.6%, respectively. From these results, we can conclude that random forest and  $k$ -nearest neighbor are the best models for the prediction. Random forest exhibits the best accuracy (up to 99.6%) in the regions where training data (labeled data) is populated.

#### Heterogeneous chain of amino acids

The ML classification is used to predict the type of amino acids in a practical amino acids chain. As the actual proteins do not typically contain chains of polypeptides, a chain of mixed type amino acids has been translocated in 10 different long-time simulations (each  $\sim 100$  ns) in which a smaller force per residue of 0.4516 pN is applied. The ionic current as a function of translocation time is plotted for one of the simulations (Fig. 5a). The chain consists of 16 amino acids as shown in Fig. 5b. To alleviate the challenge of noise, the ionic current and residence time of each amino acid are averaged over the 10 simulations (Table 1). The resulting currents and residence times are then used as unlabeled data by the polypeptides trained classification models in Fig. 4. Since a smaller force per residue is applied, the residence times are much larger than that of the polypeptides. Therefore, to use the classification from the training on the polypeptides, which are based on larger forces, the currents and residence times of both polypeptides and mixed amino acids chain are normalized by the maximum values of current and residence time, respectively (Table 1). The type of each amino acid in the mixed chain is predicted and the accuracy of prediction for logistic regression,  $k$ -nearest neighbor and random forest is found to be 31.25, 43.75, and 62.50%, respectively. The lower accuracy is due to the large noise in the chains of mixed amino acids. The noise is expected to be much lower in experiments where the translocation times of each residue are much longer.

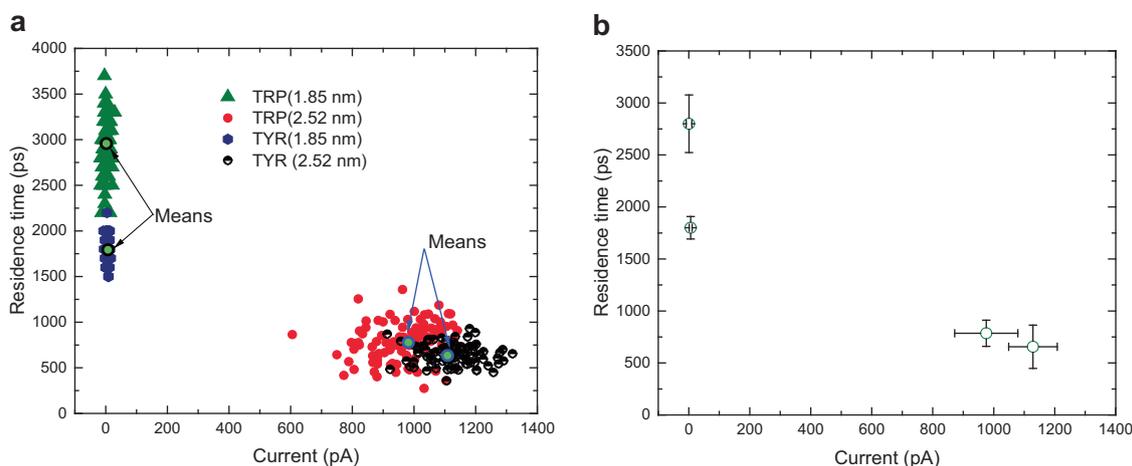
#### MoS<sub>2</sub> pore sensitivity

To characterize the sensitivity of MoS<sub>2</sub> pore with regard to its size, another set of simulations for two pore sizes of  $d = 1.85$  nm and  $d = 2.52$  nm were performed ( $d = 1.35$  nm pore is not used as it is too small to accommodate a residue of a large amino acid). We selected two amino acids (W and Y) having similar ionic currents and residence times (Fig. 3a). A force of 0.7643 pN was applied. We performed 100 simulations with an applied transmembrane bias of 200 mV for each amino acid (200 simulations in total). All other simulation parameters were kept the same except for the pore size. The ionic current and residence time data represent different patterns and values for the two pore diameters (Fig. 6). For  $d = 1.85$  nm, TYR (Y) and TRP (W) show quite distinguishable residence times with the mean values of ionic current and residence time (5.83 pA, 1787 ps) and (2.23 pA, 2866 ps), respectively (these values are for 100 samples per each pore diameter and the average values differ slightly from the data in Fig. 2). For  $d = 2.52$  nm, the ionic current and residence times clusters overlap and does not exhibit the distinguishability of the 1.85 nm pore (Fig. 6). The mean values of ionic current and residence time are (TYR: 1125.91 pA, 645.84 ps) and (TRP: 964.92 pA, 786.92 ps). The inter cluster distance (reported as a dimensionless number and computed by treating each unit of current and time as one unit) between TYR and TRP, for the pore diameters of 1.85 nm and 2.52 nm is 1079 and 51.9, respectively. Based on this data and inter cluster distances, it can be inferred that the pore with a smaller diameter ( $d = 1.85$  nm) is 20 times more sensitive compared to the 2.52 nm pore (the ratio of the distance between cluster means of 1.85 nm pore to 2.25 nm pore is  $\sim 20$ ). In general, as the diameter increases, the sensitivity decreases significantly. Our conclusion is consistent with the experimental result of Feng et al.<sup>18</sup> for DNA identification in different MoS<sub>2</sub> pore sizes. It is also notable that there is a lower limit to the nanopore size for the translocation of biomolecules. For DNA, pore sizes smaller than 2.0 nm will cause a permanent blockade of pore due to the large size of bases and the presence of deoxyribose. For amino acids, it can be as small as  $\sim 1.80$  nm.

**Table 1.** Machine learning prediction of the heterogeneous chain of amino acids

Amino acid	$I_c$ (pA)	$T_R$ (ps)	$I_c/I_{c-max}$	$T_R/T_{R-max}$	Prediction of KNN	Prediction of random forest	Prediction of logistic regression
ALA	94.48295075	807.7	0.473851	0.033599	THR	<b>ALA</b>	<b>ALA</b>
TRP	1.577286347	24039.1	0.007910	1.000000	<b>TRP</b>	<b>TRP</b>	<b>TRP</b>
CYS	198.7587907	3709.5	0.996815	0.154311	ALA	<b>CYS</b>	ALA
PHE	14.78847776	11363.9	0.074167	0.472726	TRP	TRP	<b>PHE</b>
GLY	113.0418062	2171.3	0.566927	0.090324	THR	THR	THR
TYR	47.65469162	4932.7	0.238998	0.205195	ILE	<b>TYR</b>	ILE
SER	199.3938263	2634.3	1.000000	0.109584	<b>SER</b>	ALA	ALA
GLN	37.21509764	10705.1	0.186641	0.445320	<b>GLN</b>	<b>GLN</b>	<b>GLN</b>
THR	69.70952663	14615.8	0.349607	0.608001	ARG	<b>THR</b>	ARG
MET	30.24249828	8317.9	0.151672	0.346015	GLN	HIS	ARG
ASN	73.96791658	3418.3	0.370964	0.142198	<b>ASN</b>	<b>ASN</b>	THR
PRO	13.43813196	12125.4	0.067395	0.504403	<b>PRO</b>	TYR	TYR
HIS	87.66930508	9906.5	0.439679	0.412099	<b>HIS</b>	GLN	ARG
VAL	10.73088303	9417.5	0.053818	0.391758	<b>VAL</b>	<b>VAL</b>	TYR
LYS	46.76167160	14818.6	0.234519	0.616437	TRP	<b>LYS</b>	GLN
LEU	43.47634199	10428.1	0.218043	0.433797	GLN	<b>LEU</b>	<b>LEU</b>

The averaged ionic current and residence time as well as the normalized values (by the maximum value) for each amino acid, and the prediction of each classification model are presented. The correctly predicted amino acids are highlighted in bold.



**Fig. 6** Sensitivity of MoS<sub>2</sub> nanopore. **a** Comparison of ionic current and residence time of TRP and TYR translocation through MoS<sub>2</sub> nanopores of two different diameters of 1.85 and 2.52 nm. The clusters are distinguishable for the 1.85 nm pore while they overlap for 2.52 nm pore. The bare pore current for the 1.85 and 2.52 nm pores are 571 and 1610 pA, respectively. The mean value of each cluster is also shown. **b** The mean and error bars of each cluster clearly show that the clusters are distinguishable by residence time for the 1.85 nm pore

In summary, we have proposed a single-layer MoS<sub>2</sub> nanopore for amino acid detection using ionic current, residence time and machine learning-based predictive models. We demonstrated that the acquisition of current-residence time (with at least 1000 reads per amino acid type) would yield an accuracy of up to 99.6% for amino acid detection assisted with machine learning. The 20 amino acids, each with 16 units, were categorized into different classes based on the similarity of current-residence time data. We found that the W, Y, and F residence times are significantly higher than those of A, C, G, S, and T. However, their ionic current data has an inverse relationship with the residence time (i.e., higher residence time, less current and more blockade). We also found that the polar characteristics of K, Q, R, N, H, and D give rise to intermediate current levels and residence times. The hydrophobic amino acids, in spite of their small size, exhibit very low currents

and intermediate residence times. We also employed and compared different machine learning classification techniques to predict the amino acid type based on the current-residence time training data.

## METHODS

Molecular Dynamics simulations were performed using the LAMMPS package.<sup>30</sup> Each simulation box consists of a single-layer MoS<sub>2</sub> membrane, an amino acid chain (16 identical residues), water and ions (Fig. 1.a). The MoS<sub>2</sub> membrane, amino acids, water molecules and ions were generated using visual molecular dynamics.<sup>31</sup> A nanopore is created in MoS<sub>2</sub> membrane by removing the unwanted atoms in the center of the MoS<sub>2</sub> lattice. The MoS<sub>2</sub> pore has both Mo and S atoms at its edge to replicate the most probable pore architecture in experiments. The system has dimensions of 6 × 6 × 10 nm in x, y, and z, respectively. The membrane is

fixed in the xy plane at  $z=0$  and the amino acid chain is initially placed just above the nanopore center along the of the pore (Fig. 1.a). The box containing the membrane and the polypeptide is solvated with water molecules with a padding of  $\sim 3$  nm in z. Then the resulting box is ionized with a molarity of  $\sim 1$  M (sodium and chloride). A typical simulation box contains about 32,000 atoms. A snapshot of proline chain translocating through the MoS<sub>2</sub> pore is depicted in Fig. 1b. The pore (diameter = 1.85 nm) used in the production simulations has an accessible area of 156.9 Å<sup>2</sup>. The pore diameters are computed using the method in ref. <sup>32</sup>. To characterize the size-dependent pore sensitivity of MoS<sub>2</sub> and study the I-V curves, we created two other pores with an average diameter of 1.35 and 2.52 nm (Fig. 1c).

The TIP3P water model was used and the SHAKE algorithm was employed to maintain the rigidity of the water molecules. The CHARMM27 force field<sup>33</sup> was used for water molecules, ions and polypeptides. For MoS<sub>2</sub>, the non-bonded interactions were modeled by the combination rule to obtain the Lennard-Jones (LJ) parameters.<sup>34</sup> The LJ cutoff distance was 12 Å. The long range electrostatic interactions were computed by Particle-Particle-Mesh.<sup>35</sup> Periodic boundary conditions were applied in all the three orthogonal directions.

For each simulation, the energy of the system was minimized for 30,000 steps. Next, to reach the equilibrium water density (1 g/cm<sup>3</sup>), the system was simulated in NPT ensemble for 2 ns at a pressure of 1 atm and a temperature of 300 K. Then, the system is further equilibrated in NVT ensemble for 2 ns. Temperature was maintained at 300 K by using the Nosé-Hoover thermostat with a time constant of 0.1 ps.<sup>36,37</sup> The production simulations were performed in NVT ensemble for up to 40 ns depending on the translocation speed of polypeptides. In each simulation, a constant bias of 200 mV was applied in the z-direction. Since only 4 out of 20 amino acids are charged, an external force per residue was needed to pull the polypeptide chain through the nanopore. Different forces were applied in the range of 0.4169–1.1811 pN to find the common applied force that can successfully translocate all twenty amino acids (see supporting information for translocation history vs. different applied forces). For residence times, we ran 100 simulations per amino acid type (total 2000 trajectories) with an applied force of 0.7643 pN, but we used different external forces for ionic current calculations to better mimic the experimental conditions. During all the simulations, the MoS<sub>2</sub> atoms were held fixed in space (See supporting movie for Proline translocation through a MoS<sub>2</sub> nanopore).

We monitored the time-dependent ionic current,  $I(t)$ , in the pore. We computed the ionic current through the nanopore by using the definition of current,  $I(t) = dq/dt$ , as  $I(t) = \frac{1}{L_z} \sum_i q_i \left| \frac{z_i(t+\delta t) - z_i(t)}{\delta t} \right|$ , where the sum is for all the ions,  $\delta t$  is chosen to be 5 ps, and  $z_i$  and  $q_i$  are the z-coordinate and charge of ion  $i$ , and  $n$  is the total number of ions, respectively.<sup>10</sup> To get less noisy ionic current, we skip the first and last 10% of the trajectory of the translocation. For all the trajectories used in the post-processing analyses, the presence of amino acids is checked frame by frame inside the pore. The frames, where amino acids are absent, are discarded. The residence time per residue is computed by dividing the total translocation time of the polypeptide chain translocated through the nanopore by the number of residues.

To achieve statistically meaningful data, we ran 4283 simulations in parallel (4103 translocation simulations, 80 I-V characteristic simulations, 100 pore sensitivity simulations). For each amino acid, at least 200 replica trajectories are used (out of 4103 simulations, 103 simulations were discarded because of no translocation event). This resulted in an aggregate data of  $\sim 2$  TB as well as a total combined simulation time of  $\sim 66$   $\mu$ s (see the supporting information for details on all the simulation sets.). We used Python Numpy for post-processing the data.<sup>38</sup> After collecting the ionic currents and residence times for the 20 amino acids, we created two sets of unlabeled and labeled (by the names of acid types) data for machine learning analyses. In the labeled data, we have three columns consisting of ionic currents ( $I_c$ ), residence times ( $T_R$ ) and amino acids labels (e.g., THR, TYR, ALA, ...). In the unlabeled data, we removed the third column with amino acid labels. Both the unlabeled and labeled data are available as supporting information (Amino\_acid\_IR.xlsx).

### Supporting Information Available

Machine Learning parameter selection and optimization technique, pore blockade for ALA and TRP for different conformations inside the pore, effect of mass of amino acids, translocation history of residues versus different applied forces, and Methionine thiol binding to anions.

### Data availability

The data related to the findings of this work are available from the corresponding author subject to a reasonable request.

### ACKNOWLEDGEMENTS

This work is supported by NSF under grants 1420882, 1506619, 1708852, 1720701, 1720633, and 1545907. We acknowledge the use of the parallel computing resource Blue Waters provided by the University of Illinois and the National Center for Supercomputing Applications.

### AUTHOR CONTRIBUTIONS

A.B.F. and M.H. carried out the research and wrote the paper. N.R.A. supervised the work.

### ADDITIONAL INFORMATION

**Supplementary information** accompanies the paper on the *npj 2D Materials and Applications* website (<https://doi.org/10.1038/s41699-018-0060-8>).

**Competing interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### REFERENCES

- Wei, R. S., Gatterdam, V., Wieneke, R., Tampe, R. & Rant, U. Stochastic sensing of proteins with receptor-modified solid-state nanopores. *Nat. Nanotechnol.* **7**, 257–263 (2012).
- Yusko, E. C. et al. Controlling protein translocation through nanopores with bio-inspired fluid walls. *Nat. Nanotechnol.* **6**, 253–260 (2011).
- Han, A. et al. Label-free detection of single protein molecules and protein-protein interactions using synthetic nanopores. *Anal. Chem.* **80**, 4651–4658 (2008).
- Han, A. P. et al. Sensing protein molecules using nanofabricated pores. *Appl. Phys. Lett.* <https://doi.org/10.1063/1.2180868> (2006).
- Kowalczyk, S. W. et al. Single-molecule transport across an individual biomimetic nuclear pore complex. *Nat. Nanotechnol.* **6**, 433–438 (2011).
- Majd, S. et al. Applications of biological pores in nanomedicine, sensing, and nanoelectronics. *Curr. Opin. Biotechnol.* **21**, 439–476 (2010).
- Aksimentiev, A., Heng, J. B., Timp, G. & Schulten, K. Microscopic kinetics of DNA translocation through synthetic nanopores. *Biophys. J.* **87**, 2086–2097 (2004).
- Steinbock, L. J. et al. Probing the size of proteins with glass nanopores. *Nanoscale* **6**, 14380–14387 (2014).
- Farimani, A. B., Heiraniyan, M. & Aluru, N. R. Electromechanical signatures for dna sequencing through a mechanosensitive nanopore. *J. Phys. Chem. Lett.* **6**, 650–657 (2015).
- Farimani, A. B., Dibaenia, P. & Aluru, N. R. DNA origami-graphene hybrid nanopore for dna detection. *ACS Appl. Mater. Interfaces* **9**, 92–100 (2017).
- Barati Farimani, A., Heiraniyan, M., Min, K. & Aluru, N. R. Antibody subclass detection using graphene nanopores. *J. Phys. Chem. Lett.* **8**, 1670–1676 (2017).
- Branton, D. et al. The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* **26**, 1146–1153 (2008).
- Bayley, H. & Cremer, P. S. Stochastic sensors inspired by biology. *Nature* **413**, 226–230 (2001).
- Larkin, J., Henley, R. Y., Muthukumar, M., Rosenstein, J. K. & Wanunu, M. High-bandwidth protein analysis using solid-state nanopores. *Biophys. J.* **106**, 696–704 (2014).
- Venkatesan, B. M. & Bashir, R. Nanopore sensors for nucleic acid analysis. *Nat. Nanotechnol.* **6**, 615–624 (2011).
- Radisavljevic, B., Radenovic, A., Brivio, J., Giacometti, V. & Kis, A. Single-layer MoS<sub>2</sub> transistors. *Nat. Nanotechnol.* **6**, 147–150 (2011).
- Liu, K., Feng, J. D., Kis, A. & Radenovic, A. Atomically thin molybdenum disulfide nanopores with high sensitivity for DNA translocation. *ACS Nano* **8**, 2504–2511 (2014).
- Feng, J. D. et al. Identification of single nucleotides in MoS<sub>2</sub> nanopores. *Nat. Nanotechnol.* **10**, 1070 (2015).
- Farimani, A. B., Min, K. & Aluru, N. R. DNA base detection using a single-layer MoS<sub>2</sub>. *ACS Nano* **8**, 7914–7922 (2014).
- Feng, J. et al. Electrochemical reaction in single layer MoS<sub>2</sub>: nanopores opened atom by atom. *Nano. Lett.* **15**, 3431–3438 (2015).

21. Sulima, S. O. et al. Bypass of the pre-60S ribosomal quality control as a pathway to oncogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 5640–5645 (2014).
22. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237–1251 (2013).
23. Zhao, Y. et al. Single-molecule Spectroscopy of amino acids and peptides by recognition tunnelling. *Nat. Nanotechnol.* **9**, 466–473 (2014).
24. Wilson, J., Sloman, L., He, Z. & Aksimentiev, A. Graphene nanopores for protein sequencing. *Adv. Funct. Mater.* **26**, 4830–4838 (2016).
25. Kolmogorov, M., Kennedy, E., Dong, Z., Timp, G. & Pevzner, P. A. Single-molecule protein identification by sub-nanopore sensors. *PLoS. Comput. Biol.* **13**, e1005356 (2017).
26. Feng, J. D. et al. Single-layer MoS<sub>2</sub> nanopores as nanopower generators. *Nature* **536**, 197+ (2016).
27. Wanunu, M. et al. Rapid electronic detection of probe-specific microRNAs using thin nanopore sensors. *Nat. Nanotechnol.* **5**, 807–814 (2010).
28. Darby, N. J. & Creighton, T. E. *Protein Structure*. (IRL Press at Oxford, 1993).
29. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
30. Plimpton, S. Fast parallel algorithms for short-range molecular-dynamics. *J. Comput. Phys.* **117**, 1–19 (1995).
31. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph. Model.* **14**, 33–38 (1996).
32. Heiranian, M., Farimani, A. B. & Aluru, N. R. Water desalination with a single-layer MoS<sub>2</sub> nanopore. *Nat. Commun.* **6**, 6 (2015).
33. MacKerell, A. D. & Banavali, N. K. All-atom empirical force field for nucleic acids: II. Application to molecular dynamics simulations of DNA and RNA in solution. *J. Comput. Chem.* **21**, 105–120 (2000).
34. Liang, T., Phillpot, S. R. & Sinnott, S. B. Parametrization of a reactive many-body potential for Mo-S systems. *Phys. Rev. B.* <https://doi.org/10.1103/PhysRevB.79.245110> (2009).
35. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald—an N-Log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
36. Nose, S. A unified formulation of the constant temperature molecular-dynamics methods. *J. Chem. Phys.* **81**, 511–519 (1984).
37. Hoover, W. G. Canonical dynamics—equilibrium phase-space distributions. *Phys. Rev. A.* **31**, 1695–1697 (1985).
38. Oliphant, T. E. Python for scientific computing. *Comput. Sci. Eng.* **9**, 10–20 (2007).



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018