# IDENTIFICATION OF AMINO ACIDS WITH SENSITIVE NANOPOROUS $MoS_2$: TOWARD MACHINE LEARNING-BASED PREDICTION

**Allocation:** Illinois/135 Knh
**PI:** Narayana R. Aluru[1]

[1]University of Illinois at Urbana–Champaign

## EXECUTIVE SUMMARY

Identifying a chain of amino acids can enable breakthrough advances in early diagnosis of disease and the health status of the human body. Many diseases, including cancer, diabetes, and digestive disorders, are caused by malfunctioning ribosomes leading to defective proteins. Therefore, sequencing an amino acid chain helps diagnose diseases at early stages.

In this study using petascale-based molecular simulations with a total aggregate simulation time of 66 microseconds ($\mu$s), the researcher demonstrated that a nanoporous single-layer molybdenum disulfide ($MoS_2$) can detect individual amino acids in a polypeptide chain with high accuracy and distinguishability. With the aid of machine learning techniques, the PI featurized and clustered the ionic current and residence time of the 20 amino acids and identified the fingerprints of the signals. In addition, using advanced machine learning classification techniques, the PI was able to predict the amino acid type of over 2.8 million hypothetical sensors.

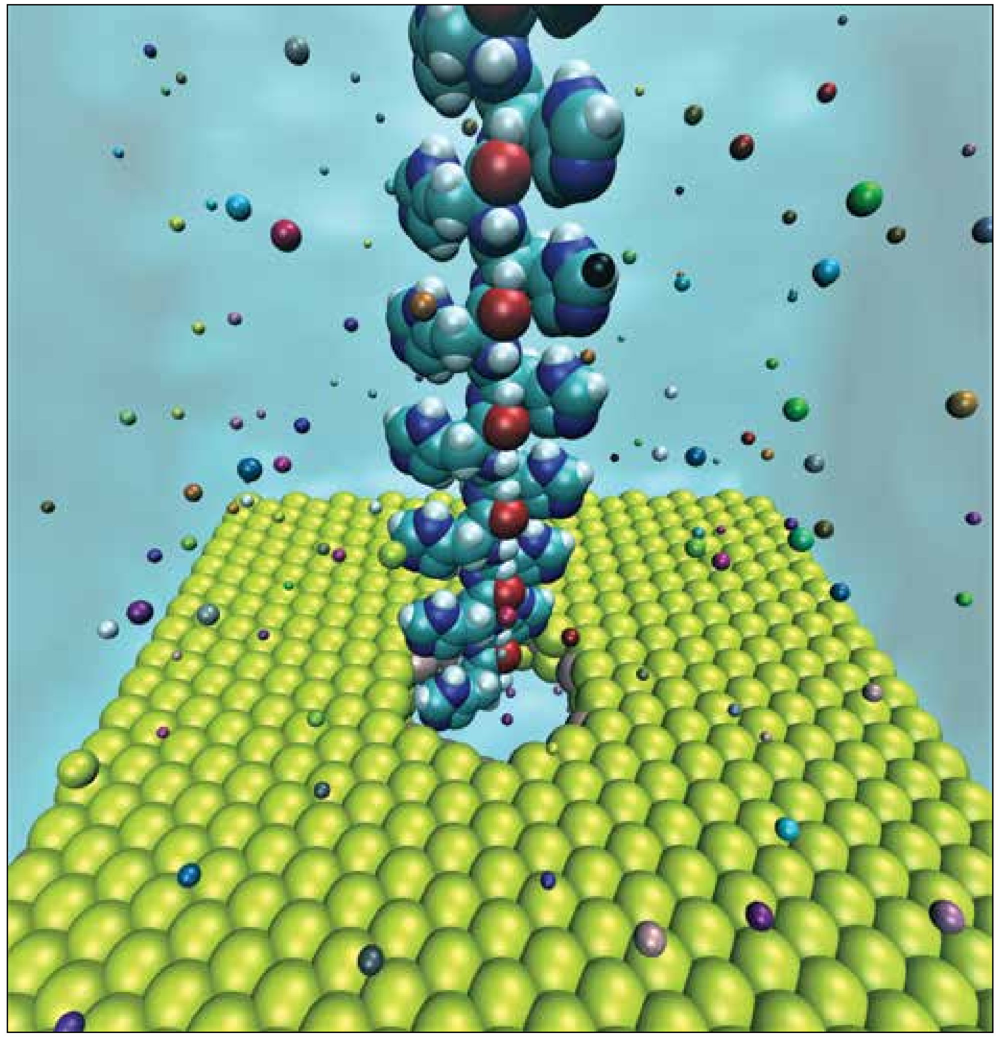

Figure 1: A snapshot of proline polypeptide translocation through the $MoS_2$ nanopore.

## RESEARCH CHALLENGE

DNA sequencing using nanopore technology has evolved significantly during the last few years. Oxford Nanopore Technologies Ltd. is currently fabricating a USB stick-sized device that can sequence DNA in several hours. In recent years, both biological and synthetic nanopores have been used for "label-free," high-resolution DNA sequencing. In addition to DNA sequencing, detection of proteins can lead to advances in improving the health status of the human body.

The challenges posed by biological molecule detection using nanopore technology are the low signal-to-noise ratio, pore degradation owing to multiple uses, the identification of single bases in real time, and the high speed of translocation [1,2]. Engineering the translocation of molecules through biological/synthetic nanopores has been defined as one of the challenging problems of biotechnology. This project has shown that an ultrathin $MoS_2$ nanopore is capable of detecting and identifying all 20 standard amino acids in proteins by using extensive molecular dynamics simulations.

## METHODS & CODES

Molecular dynamics (MD) simulations were performed using the large-scale atomic/molecular massively parallel simulator (LAMMPS). LAMMPS is an open source classical MD code for simulation of liquid, solid, and gas phases. The LAMMPS-based simulations involved three different interatomic potentials: Tersoff potential, Lennard–Jones potential, and long-range Coulombic via particle–particle particle–mesh. Each simulation box contained about 32,000 atoms consisting of a monolayer $MoS_2$, an amino acid chain, water molecules, and ions. The amino acid chain was pulled through the nanopore using an external force. Fig. 1 shows a proline chain translocating through the nanopore of $MoS_2$.

## RESULTS & IMPACT

This study has shown that a single-layer $MoS_2$ nanopore can detect individual amino acids in a polypeptide chain with high accuracy and distinguishability. Using extensive MD simulations (with a total aggregate simulation time of 66 $\mu$s) the ionic current and residence times of each residue of amino acid types was characterized and featurized. The amino acids were clustered into different groups based on their physical properties (*e.g.*, size, polarity, and hydrophobicity). In addition, the type of amino acid was classified using machine learning techniques for any future ionic current and residence time sensor readings. Logistic regression, nearest neighbor, and random forest machine learning classifiers resulted in the prediction of amino acid types with an accuracy of 72.45%, 94.55% and 99.6%, respectively.

Identifying protein chains is necessary for diagnostic purposes and early-stage detection of cancer and other diseases. In fact, the data acquired from proteomic fingerprints can be as crucial as the genome in defining the health status of humans. The proposed high-precision, single-base resolution, and fast biological molecule sequencing using nanopore technology can lead to fabrication of inexpensive personal health diagnostic devices, improving the health status of individuals. This will enable the rapidly emerging fields of predictive and personalized medicine and will mark a significant leap forward for clinical genomics and proteomics.

## WHY BLUE WATERS

This project involved 4,103 extensive MD simulations with up to 50,000 atoms and an aggregate simulation time of 66 $\mu$s. These expensive computations would not have been possible without a petascale supercomputer. Also, the LAMMPS MD package scales almost linearly with the number of cores up to 100 on Blue Waters.

## PUBLICATIONS & DATA SETS

A. B. Farimani, M. Heiranian, and N. R. Aluru, "Identification of amino acids with sensitive nanoporous $MoS_2$: Towards machine learning-based prediction," *npj 2D Mater. Appl.*, vol. 2, no. 1, May 2018.