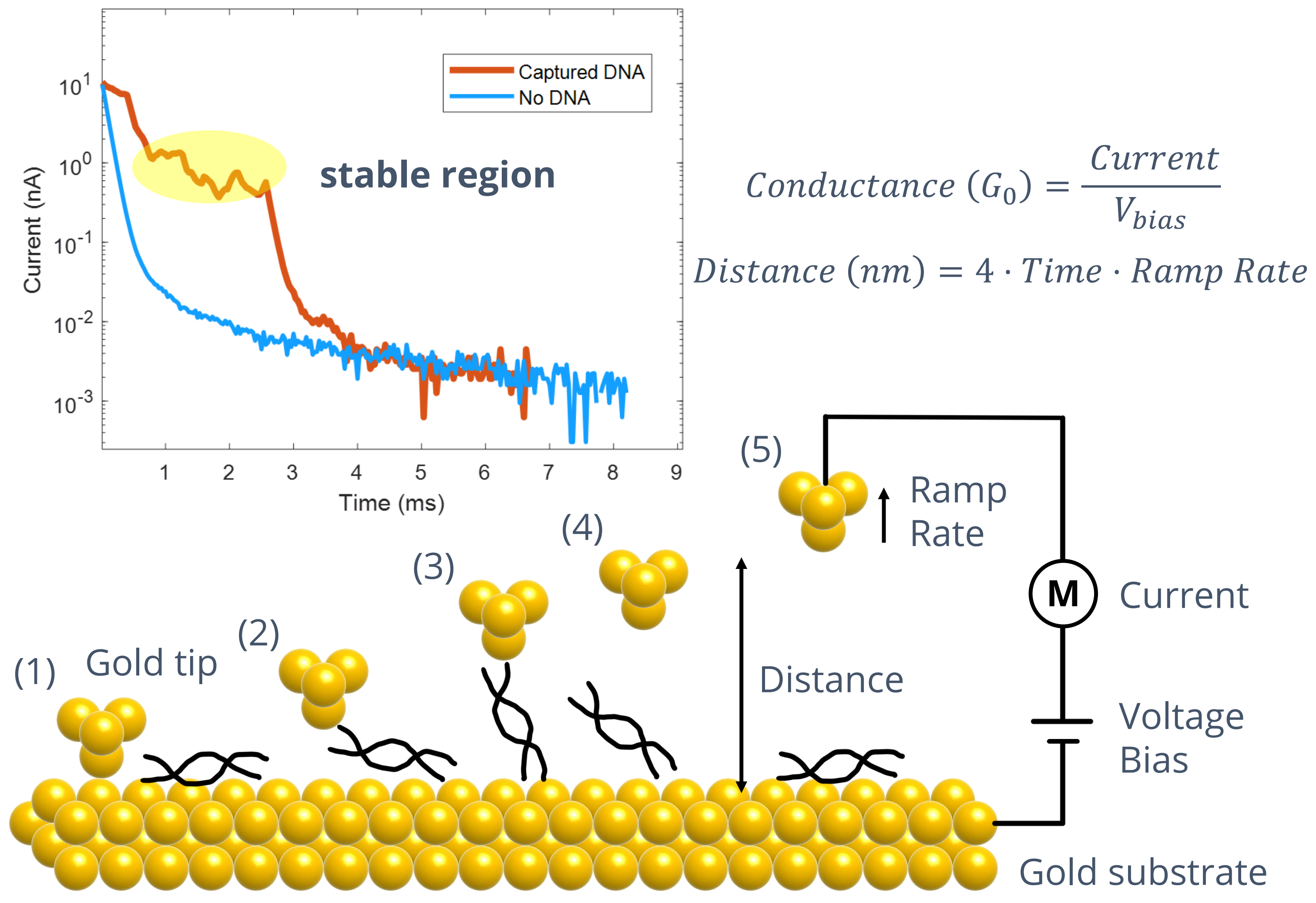
Time Series Classification
combining a Recursive Piecewise Linear Approximation Method with Machine Learning Algorithm

STUDENTS: YIREN WANG, HONGNING WANG

Motivation

- **Goal:** Identify the sequence of a DNA molecule sample from a known database, focusing on single base pair mismatches.
- **How:** Using the **time series conductance traces** recorded by the all-electronic **Single Molecule Break Junction (SMBJ)** method.
- **Hypothesis:** Molecular conductance often exhibits regions where the conductance remains relatively stable over time. These **stable regions** are the primary source of unique conductance and displacement information for different sequences. Conductance probability distributions derived from the stable regions are naturally immune to stochastic variations and should enable more accurate and robust detection.
- **Method:** For the classifier, we use an ensemble learning method **XGBoost,** with a **CNN** (convolutional neural network) model for feature extraction.
- **Key result :** We introduce a method using a recursive **Piecewise Linear Approximation (PLA)** approach to extract conductance-time segments with absolute values of slopes below a specified threshold. It shows an impressive **performance boost (approximately 20%)** for all sequences. About **20-30 experimental runs** are sufficient for over 95% accuracy and real-time DNA sequence identification.
- **Limitation:** Two sequences, differing by a single base mismatch, require a higher slope threshold with the PLA approach. Additional physics-based modeling is required to understand why these sequences are hard to classify.

SMBJ Setup and Experimental Parameters



$$Conductance\ (G_0) = \frac{Current}{V_{bias}}$$

$$Distance\ (nm) = 4 \cdot Time \cdot Ramp\ Rate$$

Process Flow

Current vs. time trace

Pre-processing

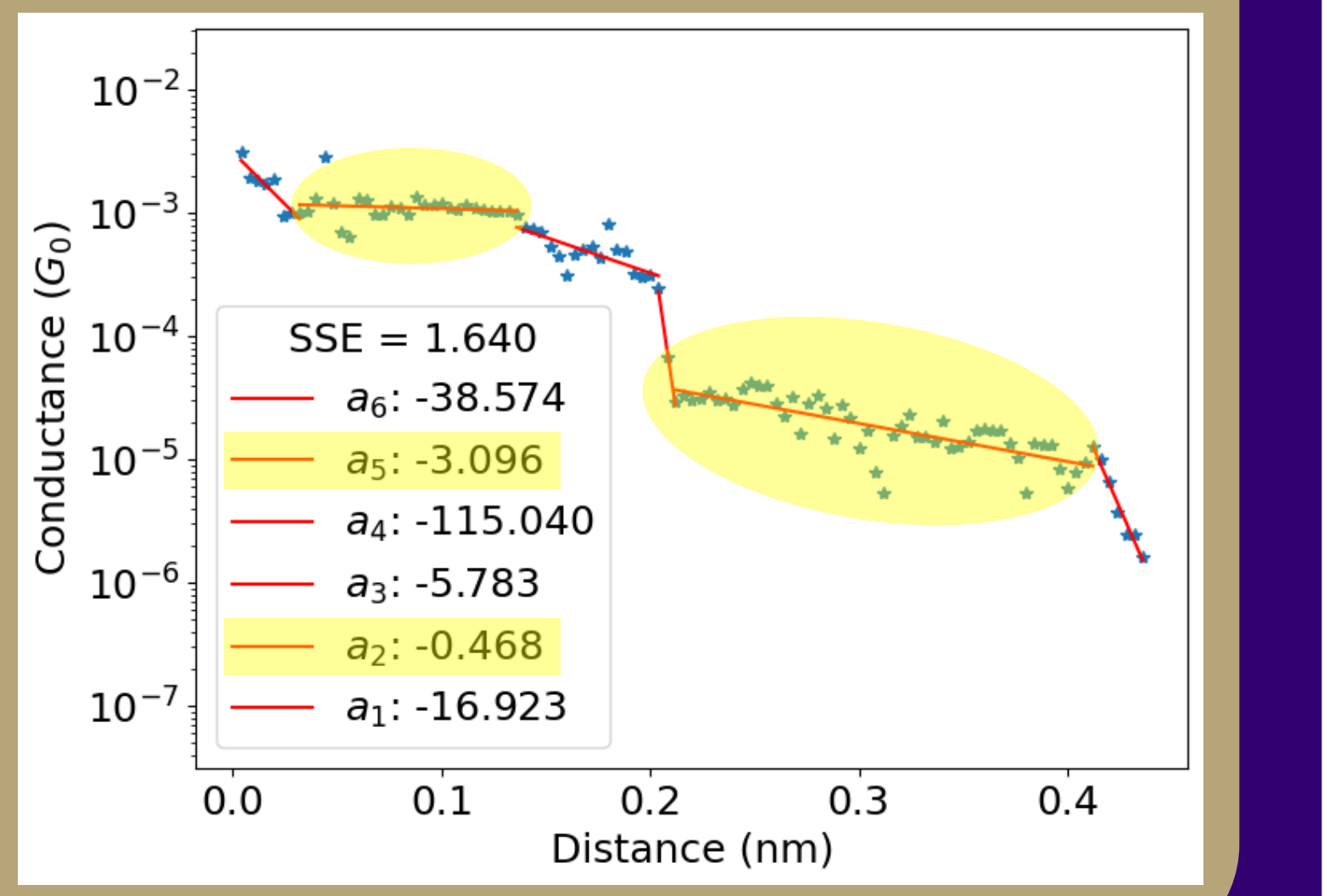
High/low clipping

Piecewise Linear Approximation

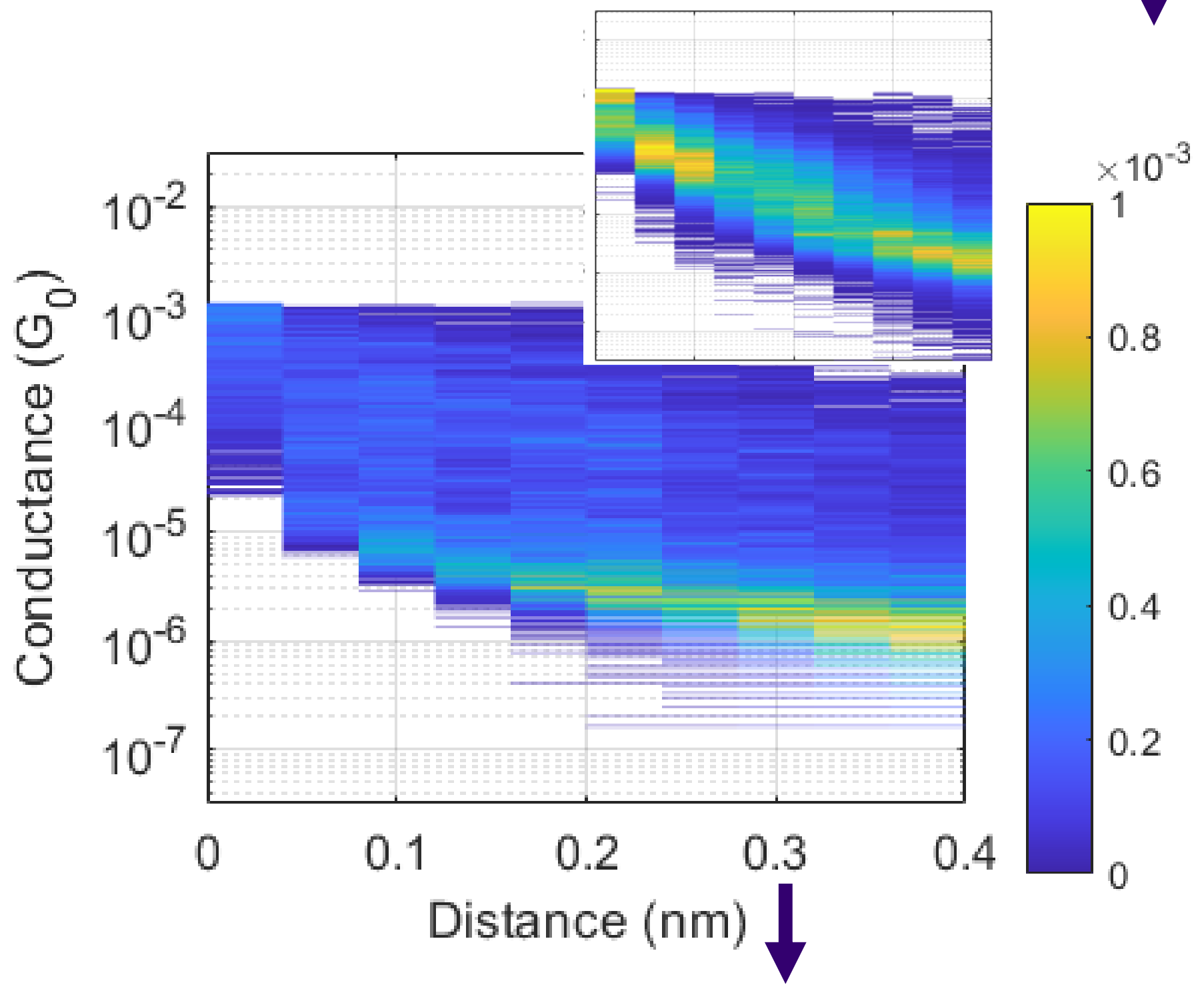
- PLA finds the best piecewise-linear approximation for a specific number of segments. It minimizes the residual for each segment recursively.

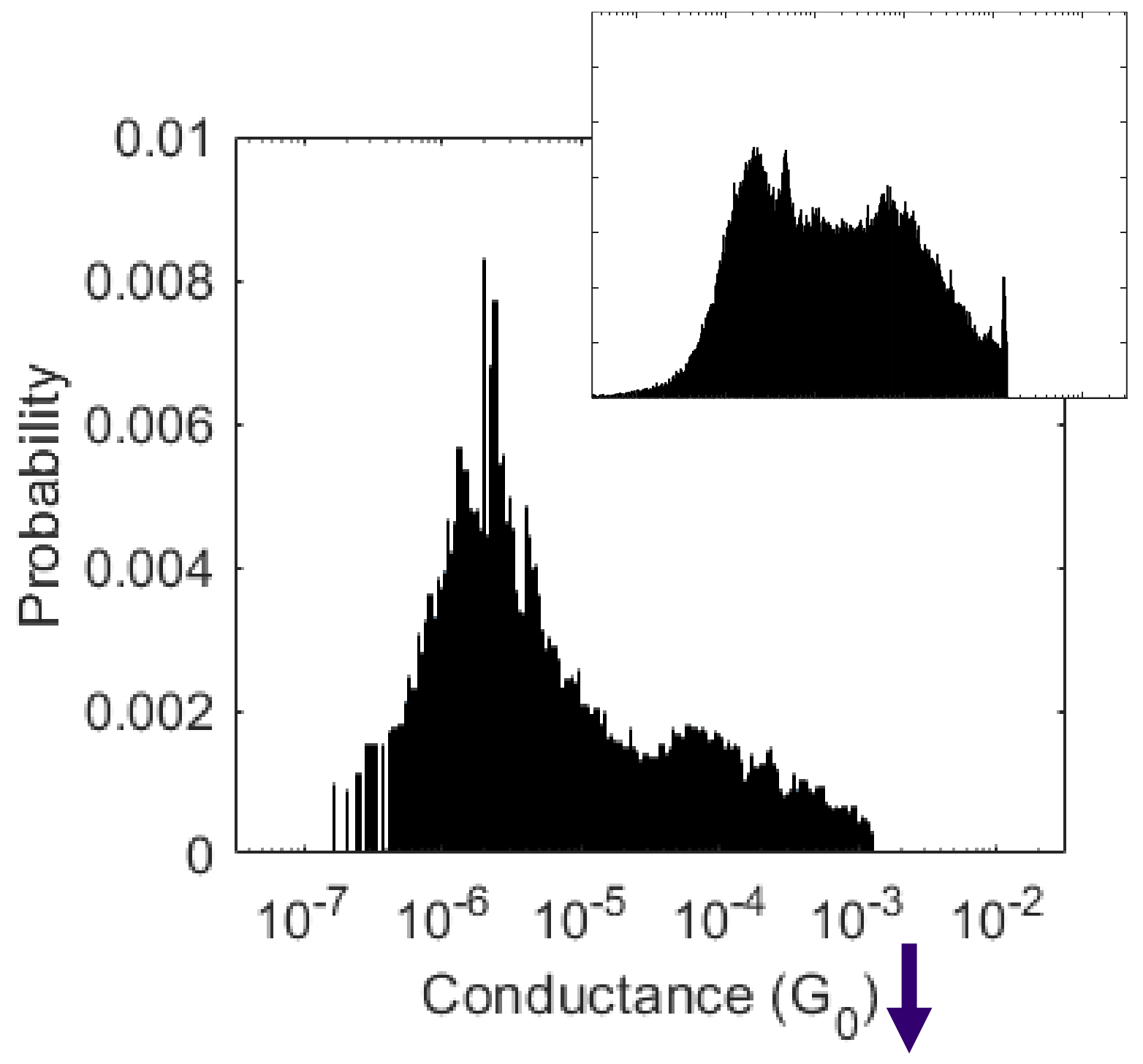
$$\min E = \sum_{s=1}^{S} \sum_{j=k(s)}^{k(s)+m(s)-1} \left(g_j - a_s d_j - b_s\right)^2$$
$$\mathrm{w.r.t}\ \{\mathrm{a_s}, \mathrm{b_s}\text{: } 1 \le s \le S\}$$

$$F(j,S) = \min_{S \le i \le j} \left(F(i,S-1) + E(i,j)\right)$$



Conductance vs. distance trace





CNN Feature Extraction → **XGBoost**

Alpha

True Class \ Predicted Class	MM1	MM2	PM
MM1	99.6%	0.1%	0.4%
MM2		99.9%	0.1%
PM	0.7%	0.1%	99.2%

Beta

True Class \ Predicted Class	MM1	MM2	MM3	PM
MM1	82.6%		17.4%	
MM2		100.0%		
MM3	16.5%		83.5%	
PM				100.0%

Delta

True Class \ Predicted Class	MM1	MM2	PM
MM1	99.6%	0.1%	0.3%
MM2	0.1%	99.6%	0.2%
PM	0.6%	0.2%	99.2%

Results

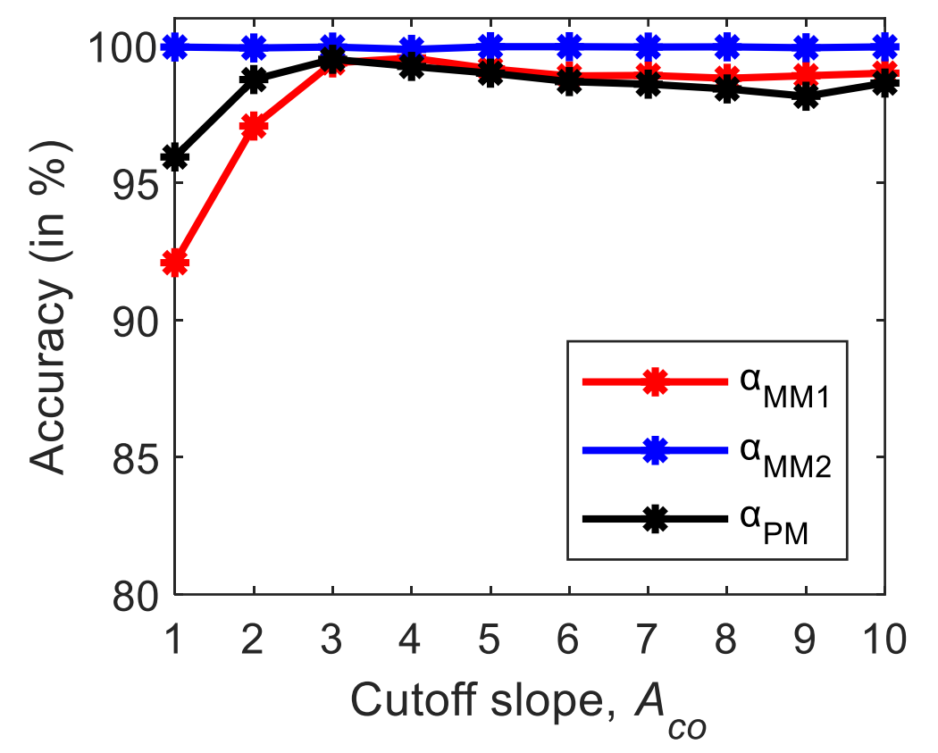
Approach	ML Algorithm	Average Accuracy sample size = 30	
		R^2 test + low pass filter	PLA with averaged distributions *
1D conductance histogram	XGBoost	83.96%	95.61%
2D conductance vs. distance histogram	CNN + XGBoost	86.40%	96.31%

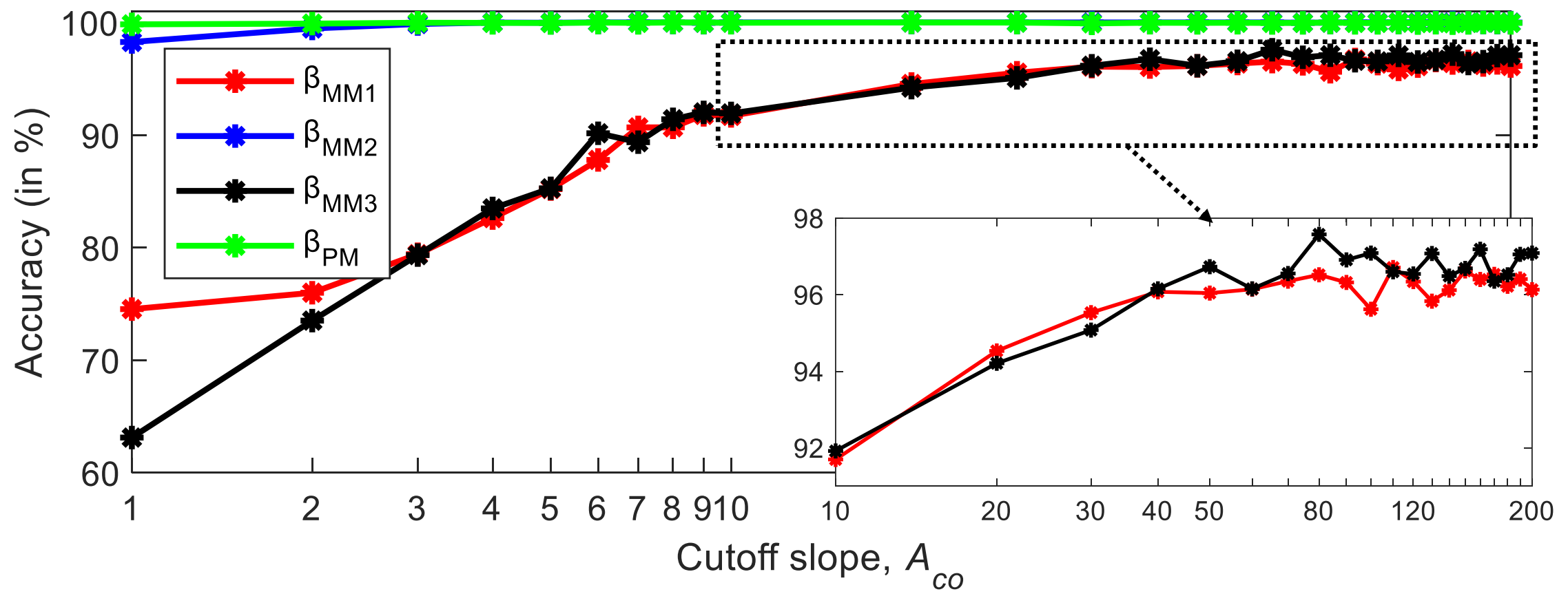
* averaging over various experimental parameters

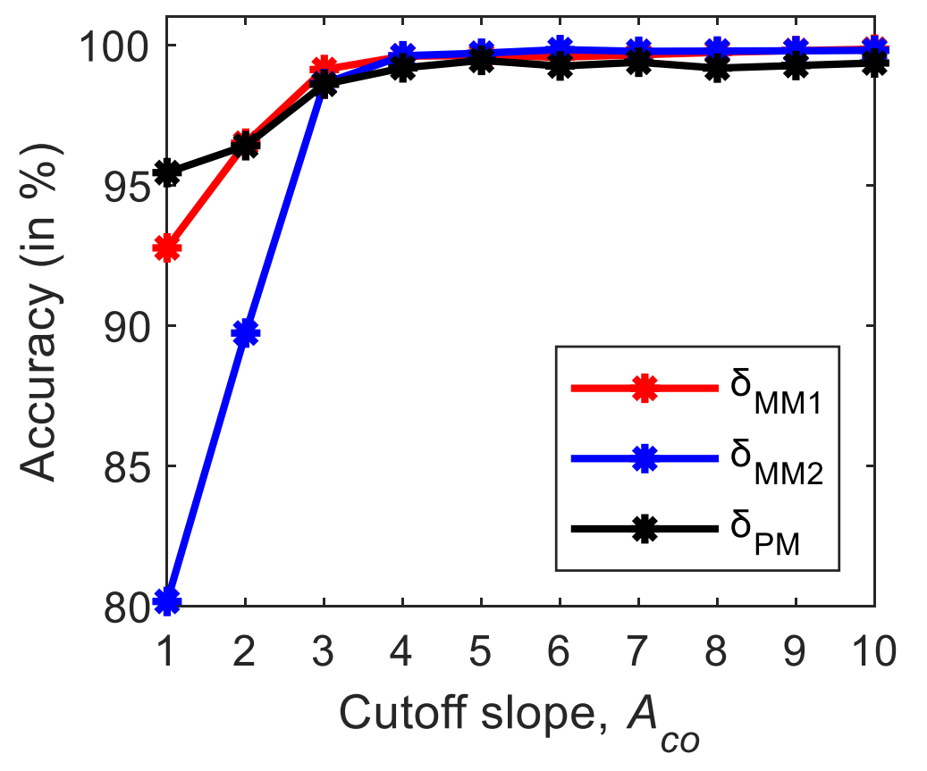
Performance Analysis w.r.t Cutoff Slope

Intuitively, increasing the **sample size** (the # of conductance traces used to compute a histogram) should improve the classification accuracy. For Beta MM1 and Beta MM3, we need a much larger sample size to reach similar accuracies as other sequences.

- With PLA method, we introduce **cutoff slope** threshold as the critical parameter.
- Beta MM1 and Beta MM3 require a cutoff slope of 80 to reach similar accuracies as other sequences processed with a cutoff slope of 4.
- **In general, a sample size of 30 is adequate to yield classifiers which exceed 95% accuracy for all ten sequences**.







References and Acknowledgments

Python package, GitHub link:
https://github.com/ethanwyr/SMBJClassifier

- Experimental data: Joshua Hihath's group, Arizona State University

- Y. Wang, H. Wang, A. K. Das, and M. P. Anantram, “Time-series Conductance Segments Extraction with Recursive Piecewise Linear Approximation Method.” preprint, 2025
- Y. Wang, H. Wang, A. K. Das, and M. P. Anantram, “A ML Framework for Genetic Sequence Identification using 2D Electrical Conductance Probability Distributions from Mixed Data Sets,” IEEE Transactions on Computational Biology and Bioinformatics, pp. 1–11, 2025, doi: 10.1109/TCBBIO.2025.3536282.
- Z. Aminiranjbar, C. Gultakti, M. Alangari, Y. Wang, B. Demir, et al., “Identifying SARS-CoV-2 Variants Using Single-Molecule Conductance Measurements,” ACS Sens, May 2024, doi: 10.1021/acssensors.3c02734.
- Y. Wang, M. Alangari, J. Hihath, A. K. Das, and M. P. Anantram, “A machine learning approach for accurate and real-time DNA sequence identification,” BMC Genomics, vol. 22, no. 1, p. 525, Dec. 2021

ELECTRICAL & COMPUTER ENGINEERING
UNIVERSITY of WASHINGTON

ADVISERS: ARINDAM DAS, M.P. ANANTRAM

SPONSOR: NATIONAL SCIENCE FOUNDATION