

Lost in Translation: Harnessing the Power of Nanopore Electrical Signals in Genomics

Computational Genomics Summer Institute (CGSI 2026)
Rancho Palos Verdes, CA, USA. July 9 2026

Can Firtina | canfirtina@gmail.com | cs.umd.edu/~firtina

Assistant Professor of Computer Science at UMD



Analyzing Genomes Reveals Key Insights



Personalized Medicine



Genome Editing

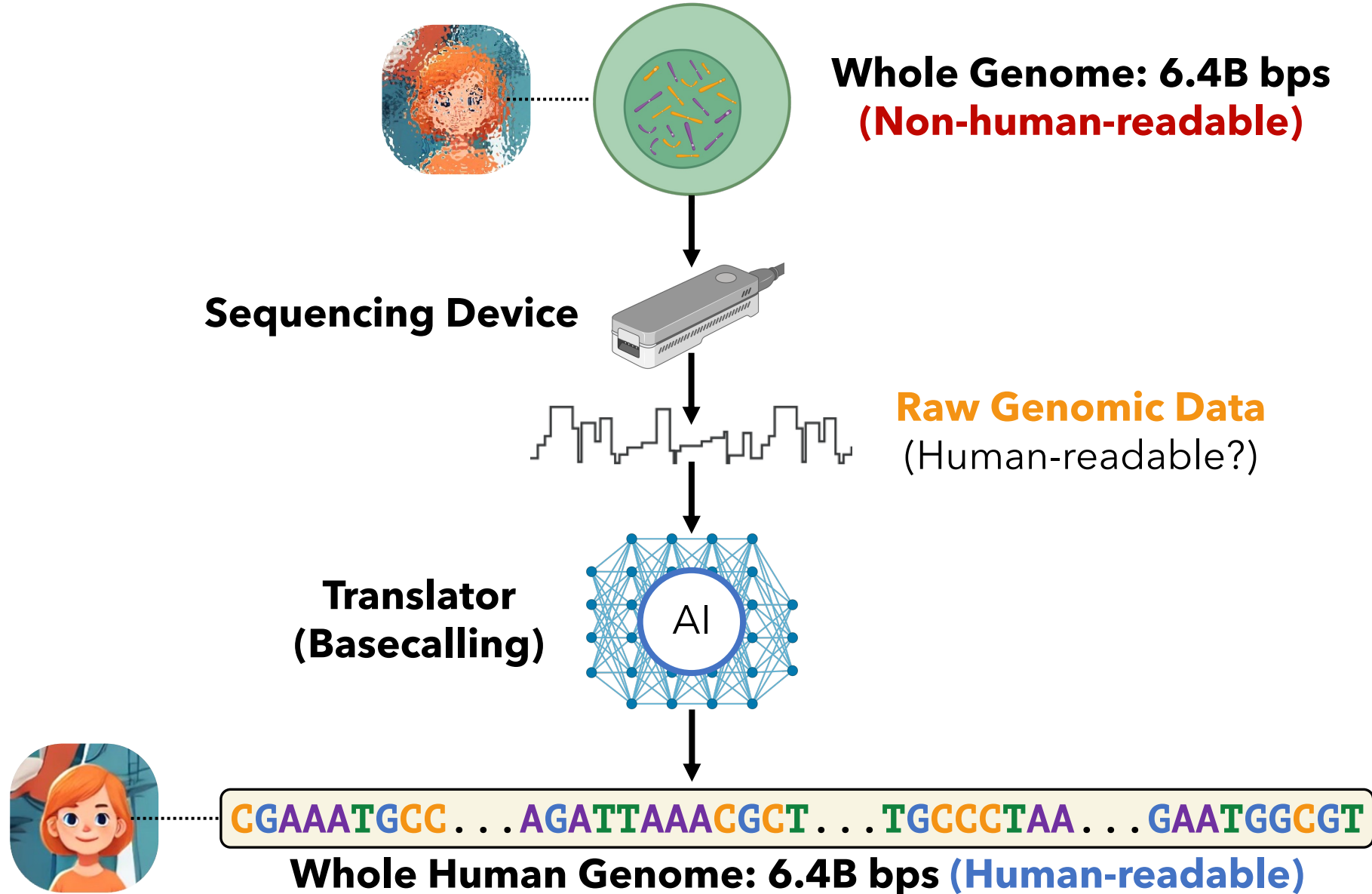


Detecting **pathogens**
in the environment

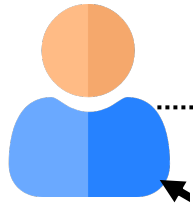


Rapid surveillance of
disease outbreaks

Generating the Entire Genomic Sequence



Genomic Data: Giant Jigsaw Puzzle to Solve



Human (Reference) Genome: 3.2 Billion bps (haploid)

CGAAATGCC...AGATTAAACGGT...TGCCCTAA...GAATGGCGT



?

CGAAATGCC...AGATTAAATGGT...TGCCCTAA...GAATGGCGT

Entire Genomic Sequence of an Organism



Small pieces (reads)



Unknown origins



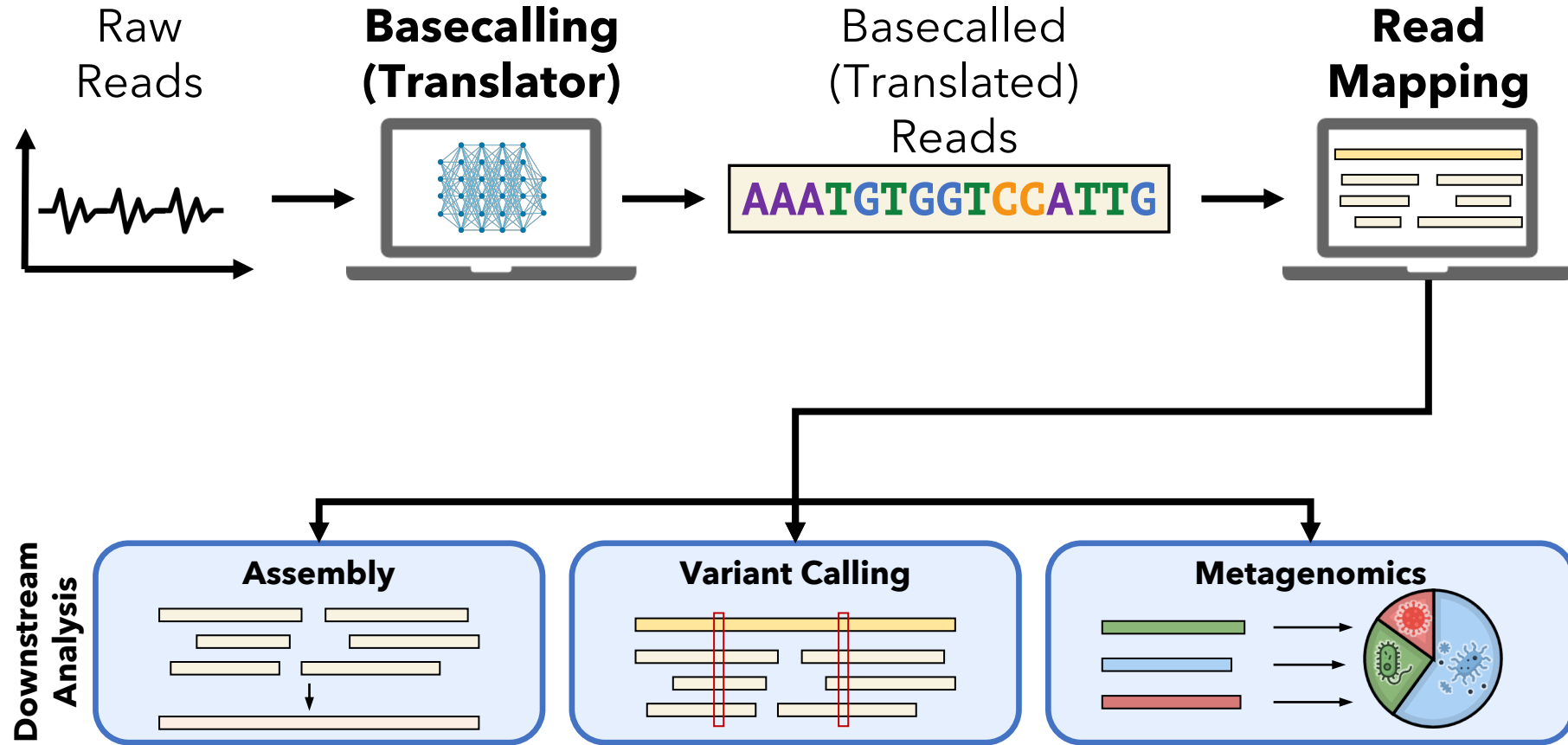
Non-identical reference genome



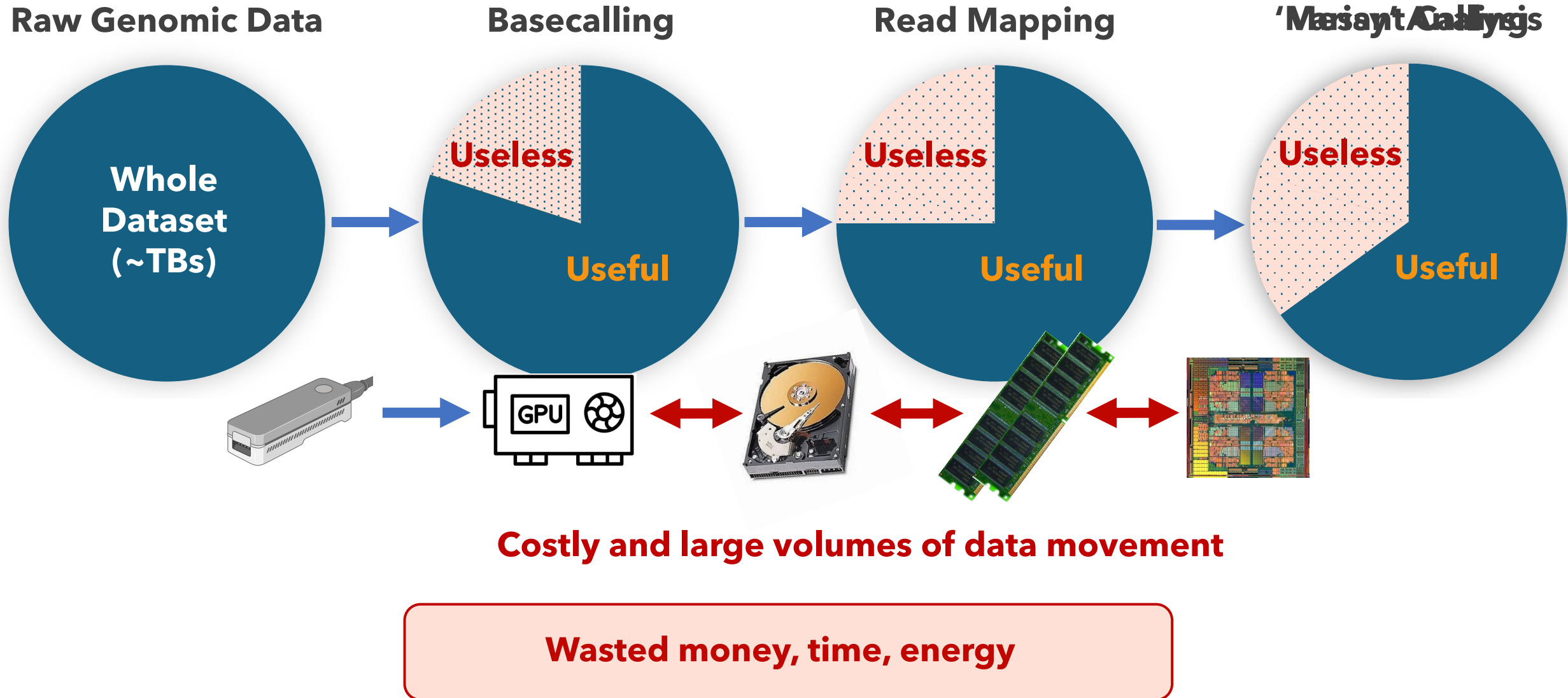
Large volume of data to analyze

**Dealing with Massive
and Noisy Datasets
(TBs to PBs)**

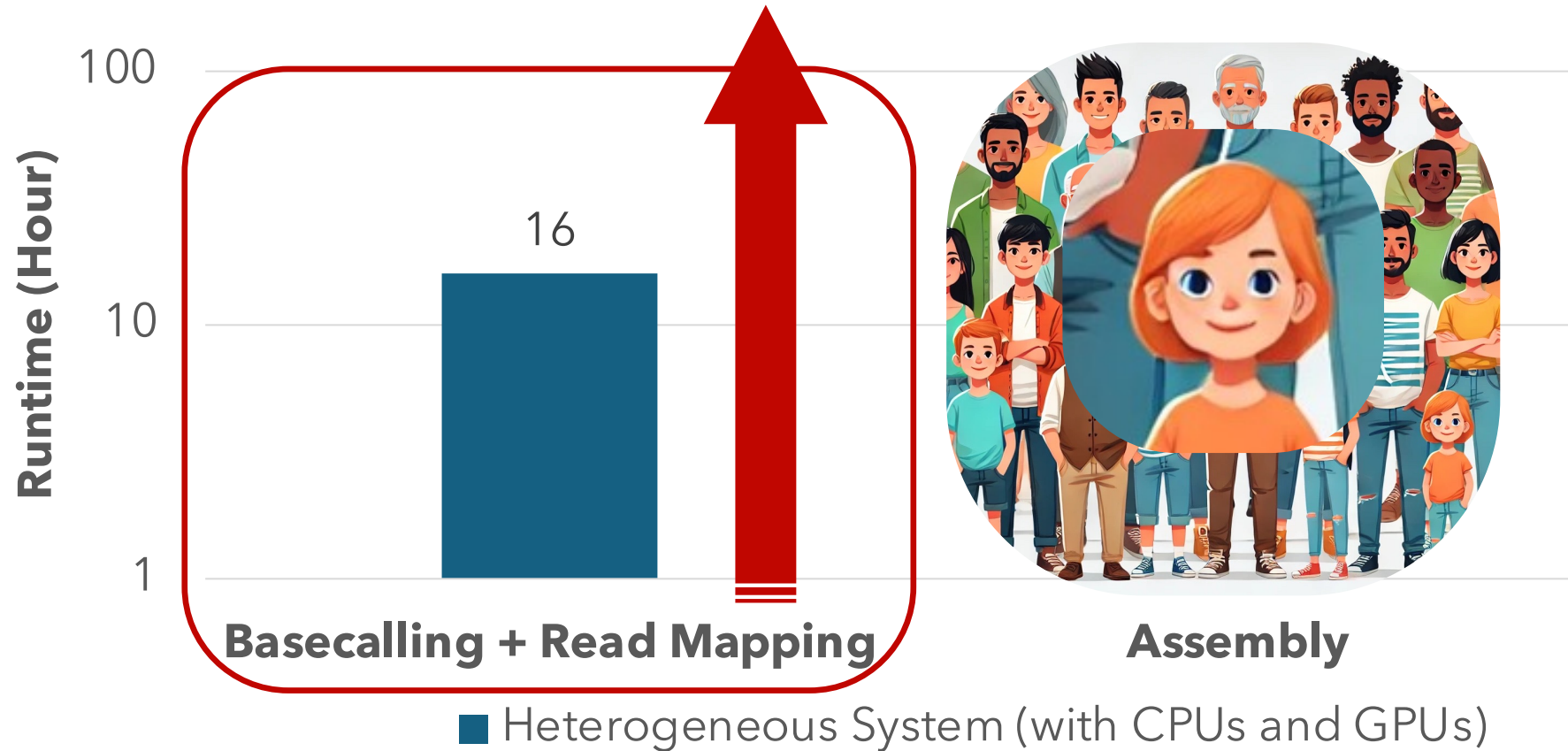
A Common Genome Analysis Pipeline



A Pipeline That Wastes Resources



A Pipeline That is Computationally Costly



Significant computational overhead

Limited Application Scope and Accuracy



**Energy-efficient analysis for
Resource-constrained devices**

**Urgent analysis
within minutes**

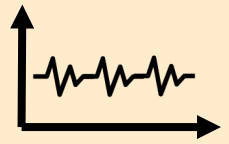


**Large-scale analysis
without sacrificing accuracy
and performance**

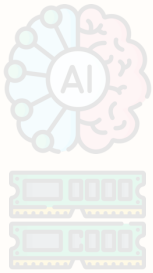


STORM Research Group

Research Directions



Real-Time and Portable
Genomic Signal Analysis



Data-centric and
AI-Driven Solutions to
Analyze Genomic Data



Accurate and Fast
Algorithms Scalable to
Large Volumes of
Genomic Data

**Today's
Talk**

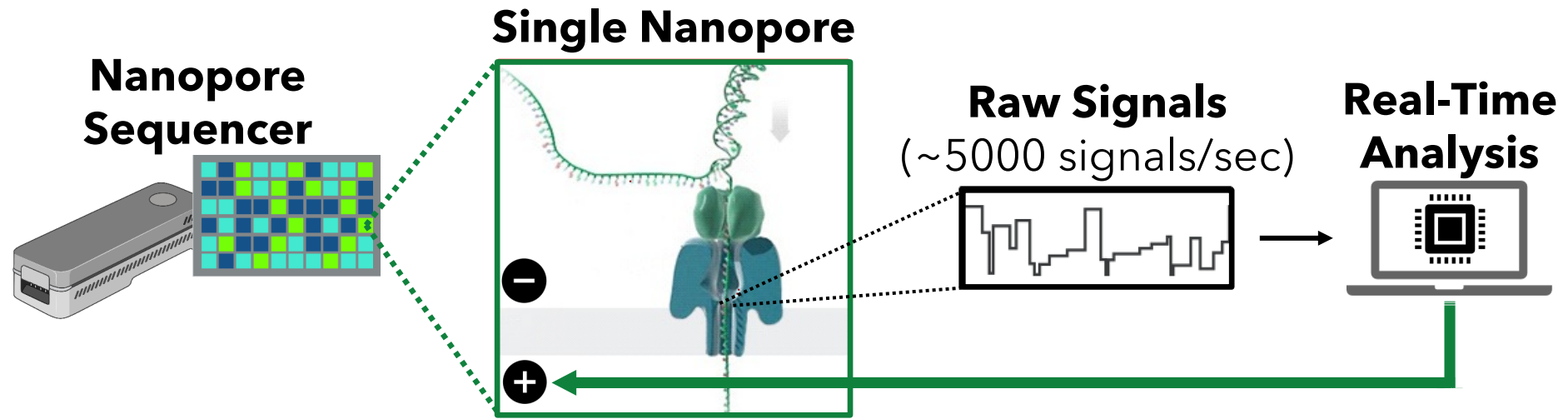


Goal

Enable fundamentally
better integration of
biological data analysis
in our **everyday lives**



Generating Electrical Signals from DNA



Democratizing Genome Analysis

Sequencing the genome of the freshwater angelfish

By Indeever Madireddy

Backed by Clive, Zach Mueller, Venky, Charlie Kinsella, Srisail, Johan Sosa, David Lang, and Cindy Wu



BioCurious San Jose, California Biology Computer Science DOI: 10.18258/27105

\$1,036

Raised of \$900 Goal

115%

Funded on 10/12/22

Successfully Funded

? How does this work?

First Ever Whole Genome Sequencing and De Novo Assembly of the Freshwater Angelfish, *Pterophyllum scalare*

Indeever Madireddy

BioCurious, Santa Clara, CA, USA

BASIS Independent Silicon Valley, San Jose, CA, USA

Correspondence to: Indeever Madireddy indeever@gmail.com

<https://experiment.com/projects/sequencing-the-genome-of-the-freshwater-angelfish>



Indeever Madireddy

@indeever_m



Wow. I didn't realize this fun project of mine would gain so much traction! I did in fact use my dead pet fish for this research. Even though my fish isn't alive, I'm glad a piece of it is still living on forever! @nanopore

Democratizing Genome Analysis



Seth Howes ✓
@SethSHowes



I sequenced my genome at home, on my kitchen table.

I wrote up exactly how I did it - the equipment, protocol, theory, and cost:



Seth Howes ✓ @SethSHowes · Apr 22



My first sequencing run produced terrible data, in part because I used DNA from cheek cells.

So this time, I'm using blood.

For nanopore sequencing, DNA input quality matters a lot. You want a tight distribution of long fragments, because that's how you maximise yield from the

[I want to sequence my genome at home](#)

From iwantosequencemygenomeathome.com

4:34 AM · Apr 19, 2026 · **1.2M** Views



107



858



4.6K



7.3K



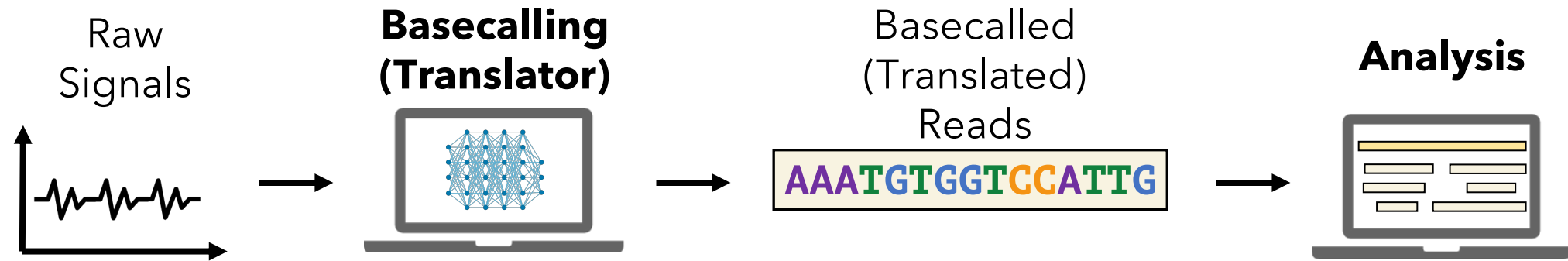
Relevant ▾

[View quotes >](#)

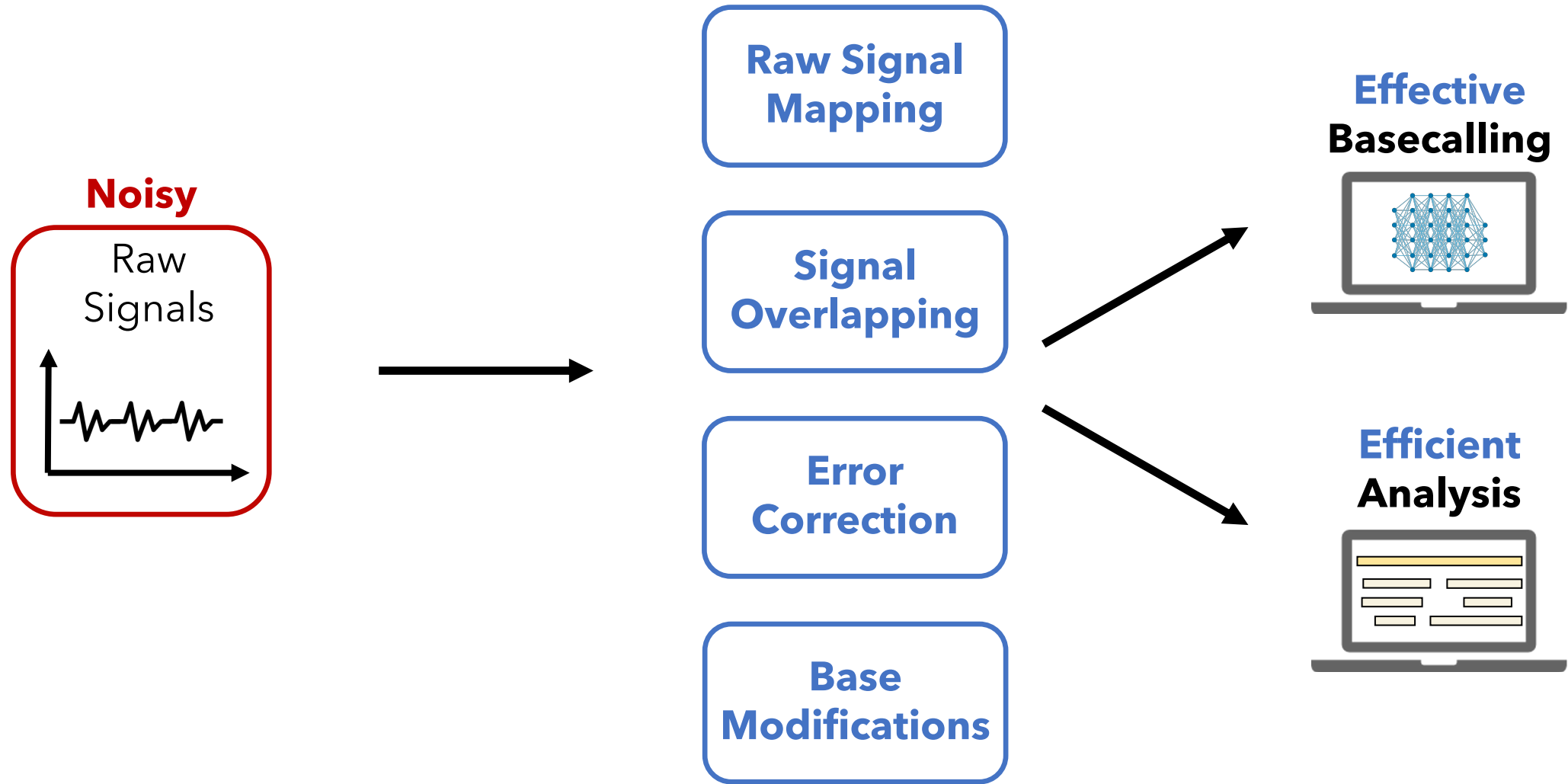
<https://x.com/SethSHowes/status/2045782975380406623>

**Future of genomic analysis
demands portable, scalable,
secure, and private solutions**

A Common Genome Analysis Pipeline

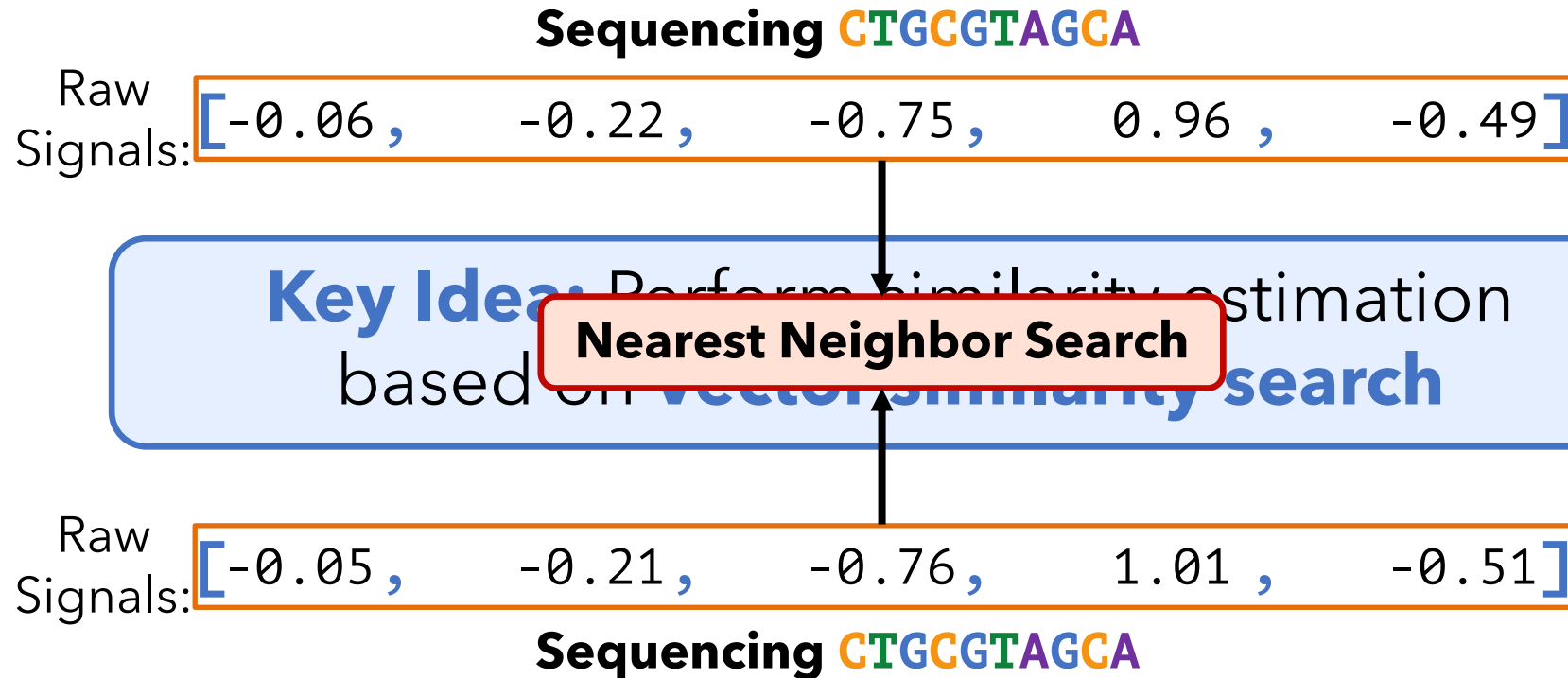


Key Perspective: Analyze Signals (Before Translation)



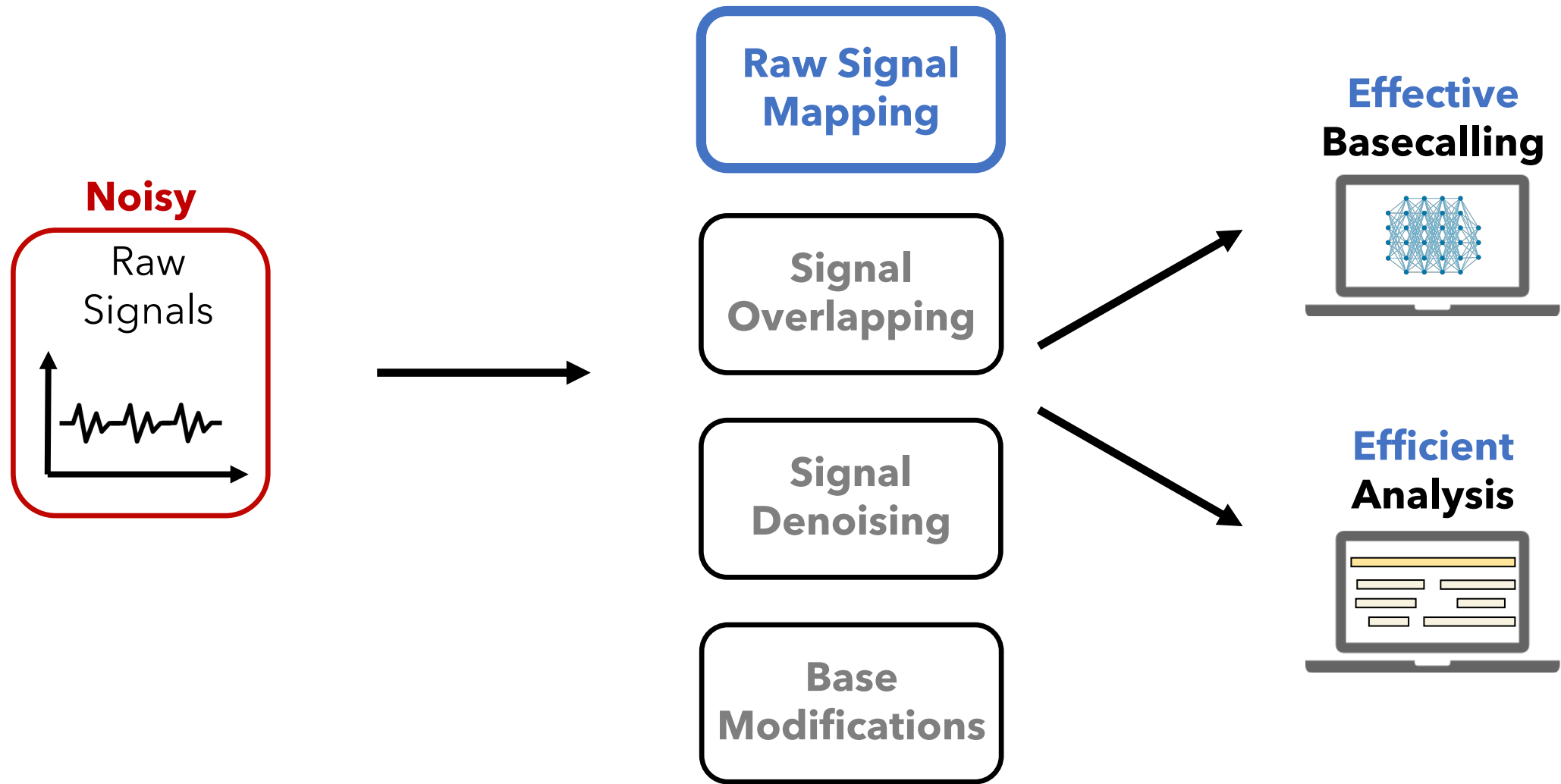
Noise in Raw Electrical Signals

Observation: Identical molecules generate similar raw signals

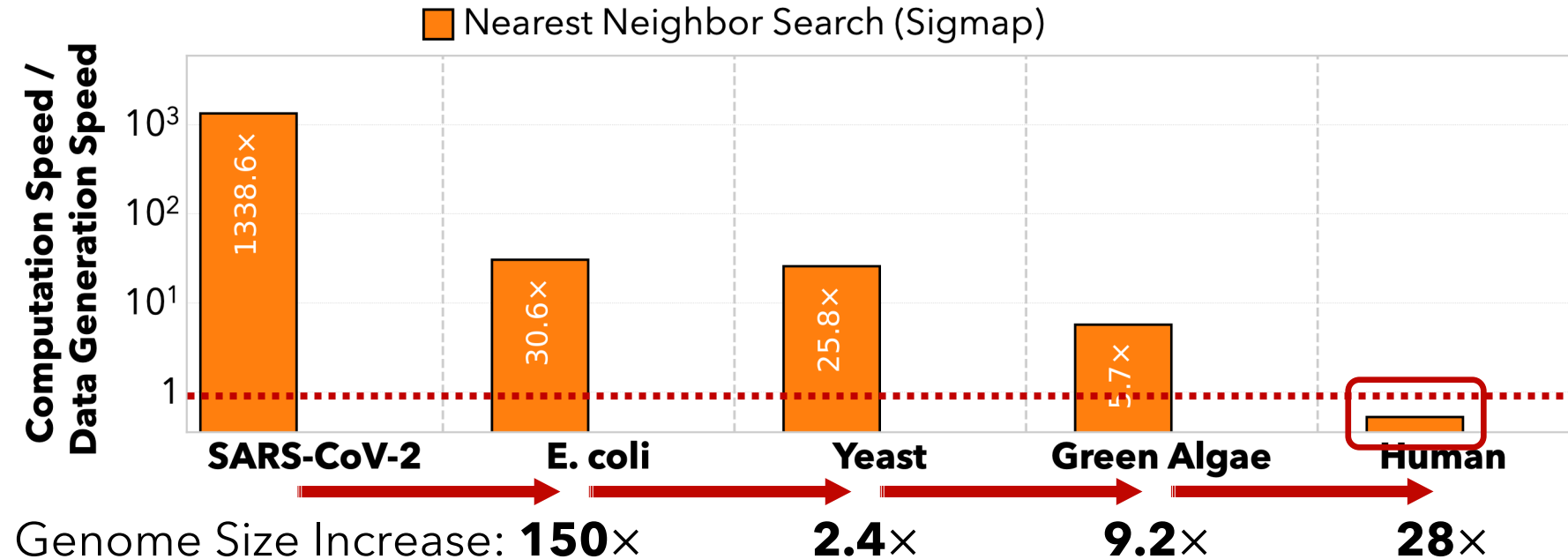


Problem: Too Costly for Large Genomes

Raw Signal Mapping



Scalability Issues Limit Real-Time Analysis



No Real-Time Analysis for Larger Genomes:
Searching mechanism does not scale well

New Frontiers: Raw Signal Analysis [ISMB '23]

- [Can Firtina](#), Nika Mansouri Ghiasi, Joel Lindegger, Gagandeep Singh, Meryem Banu Cavlak, Haiyu Mao, and Onur Mutlu,

"RawHash: enabling fast and accurate real-time analysis of raw nanopore signals for large genomes"

Proceedings of the [31st Annual Conference on Intelligent Systems for Molecular Biology \(ISMB 2023\)](#) and the [22nd European Conference on Computational Biology \(ECCB 2023\)](#), Lyon, France, July 2023.

Paper [\[online\]](#) [\[pdf\]](#) [\[code\]](#)

Conference Talk [\[video\]](#) [\[pptx\]](#) [\[pdf\]](#)

Preprint (bioRxiv) [\[online\]](#) [\[pdf\]](#)

Social Media Thread [\[Twitter \(X\)\]](#)

Bioinformatics, 2023, **39**, i297–i307
<https://doi.org/10.1093/bioinformatics/btad272>
ISMB/ECCB 2023



RawHash: enabling fast and accurate real-time analysis of raw nanopore signals for large genomes

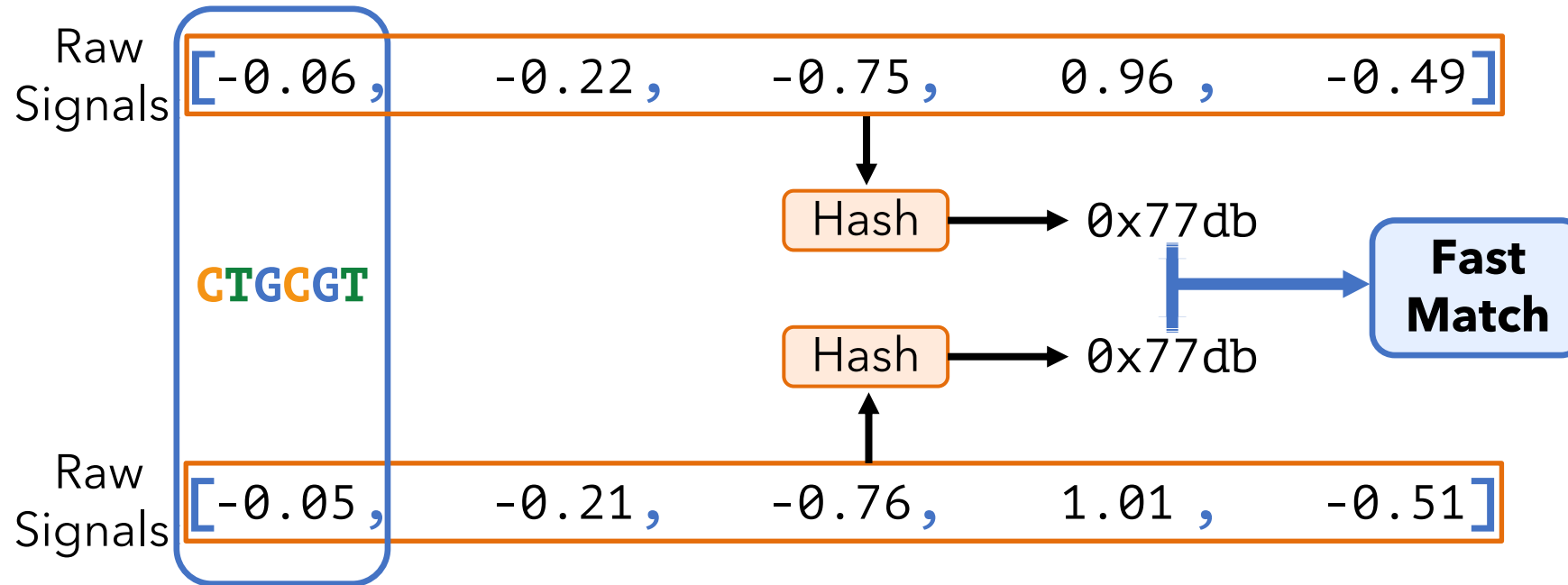
Can Firtina ^{1,*}, Nika Mansouri Ghiasi ¹, Joel Lindegger ¹, Gagandeep Singh ¹,
Meryem Banu Cavlak ¹, Haiyu Mao ¹, Onur Mutlu ^{1,*}

¹Department of Information Technology and Electrical Engineering, ETH Zurich, 8092 Zurich, Switzerland

*Corresponding author. Department of Information Technology and Electrical Engineering, ETH Zurich, Gloriastrasse 35, 8092 Zurich, Switzerland.
E-mail: firtinac@ethz.ch (C.F.), omutlu@ethz.ch (O.M.)

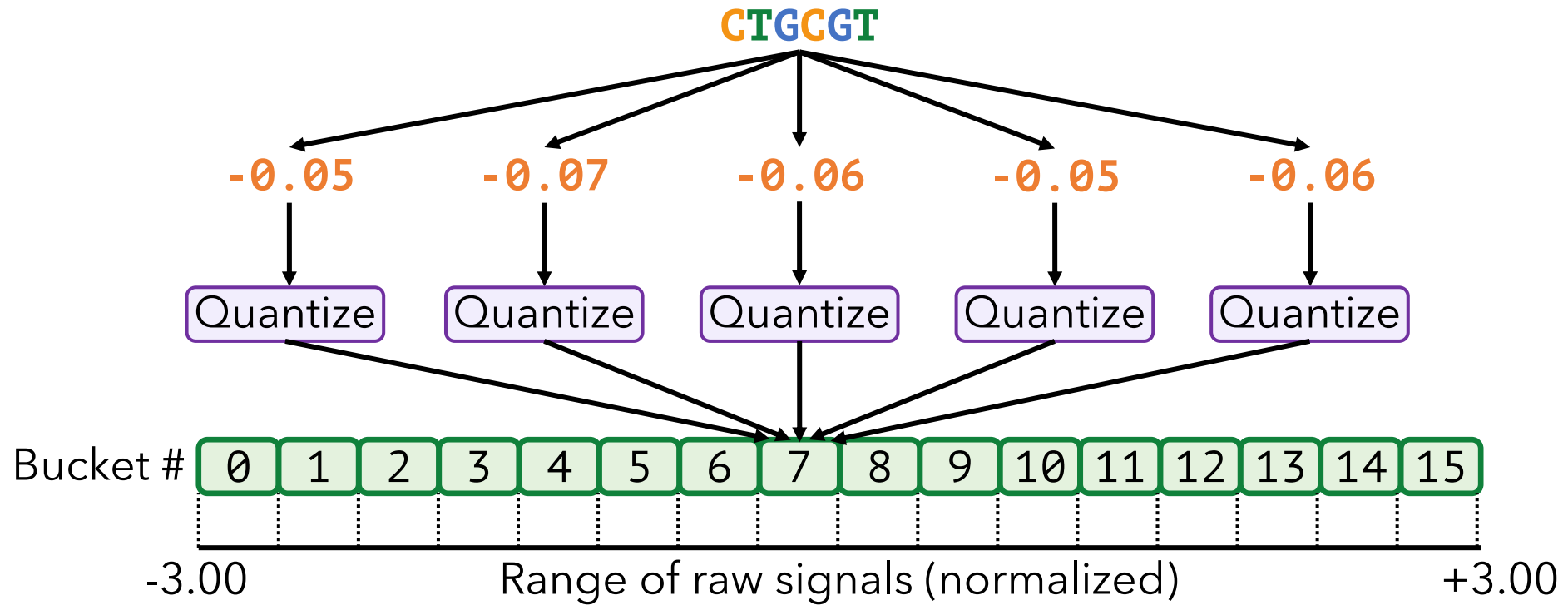
RawHash's Key Idea

Key Idea: Perform similarity estimation based on **quick exact matches**



Challenge: How to generate **the same hash value** for **similar enough signals with various noise types**?

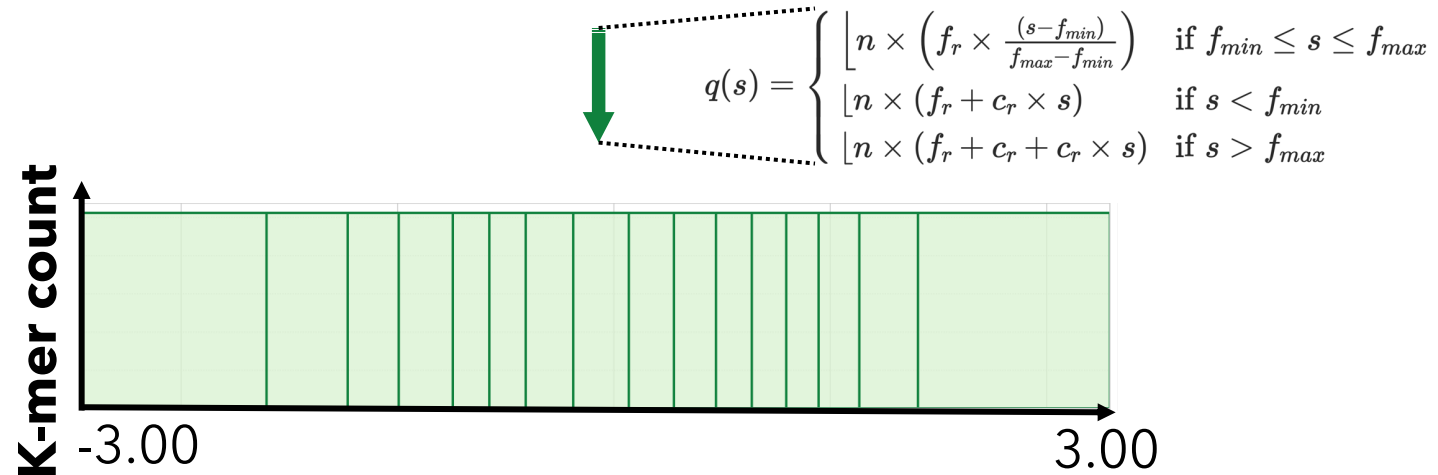
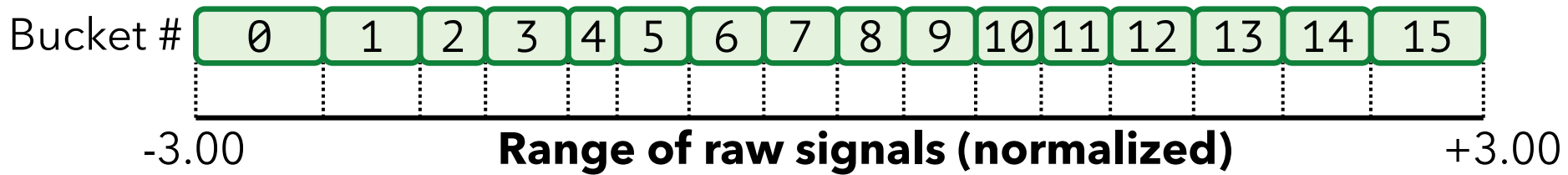
RawHash's Key Idea – Quantization



Enables matching raw signals by eliminating slight differences

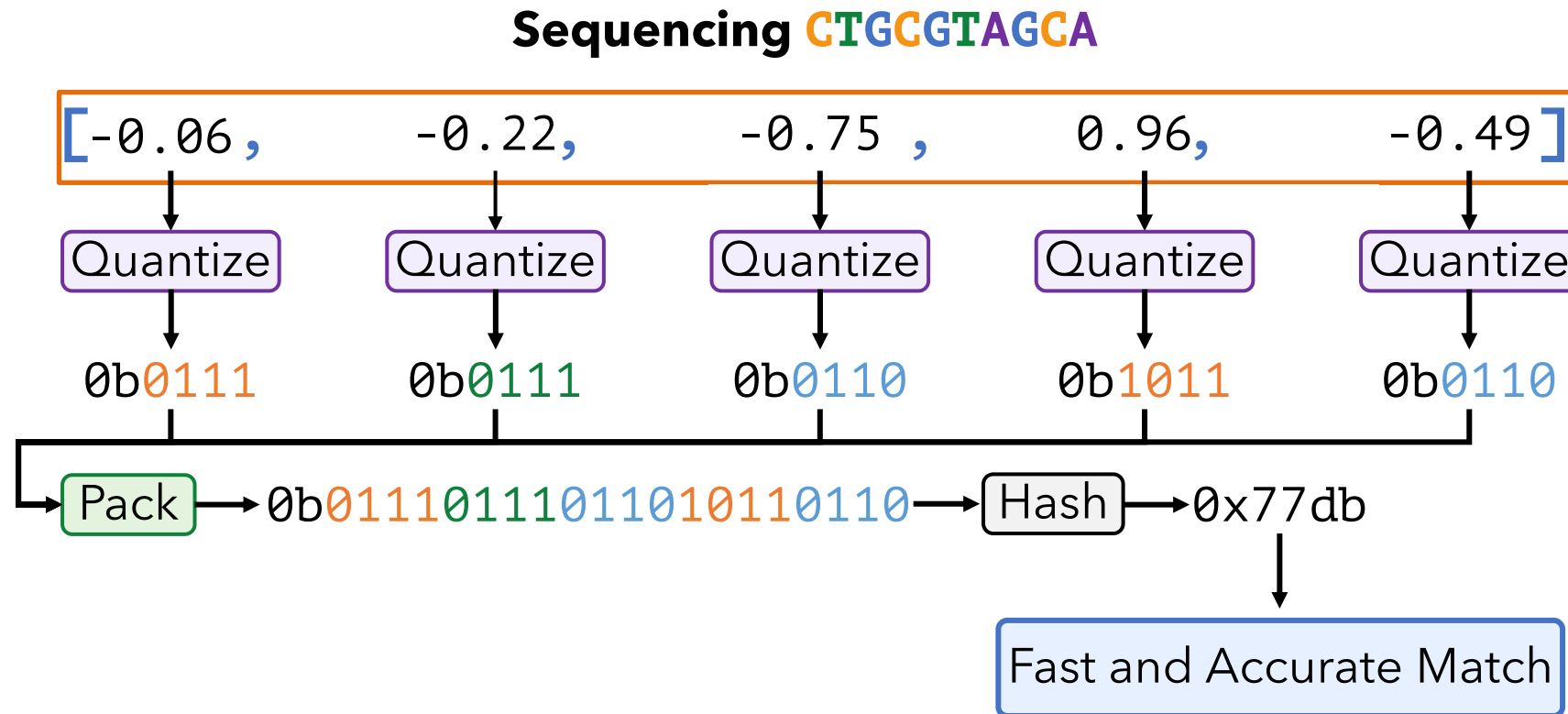
Adaptive Quantization

- **Key Idea:** Quantizing raw signals with **non-equal bucket widths** by leveraging raw signal distribution



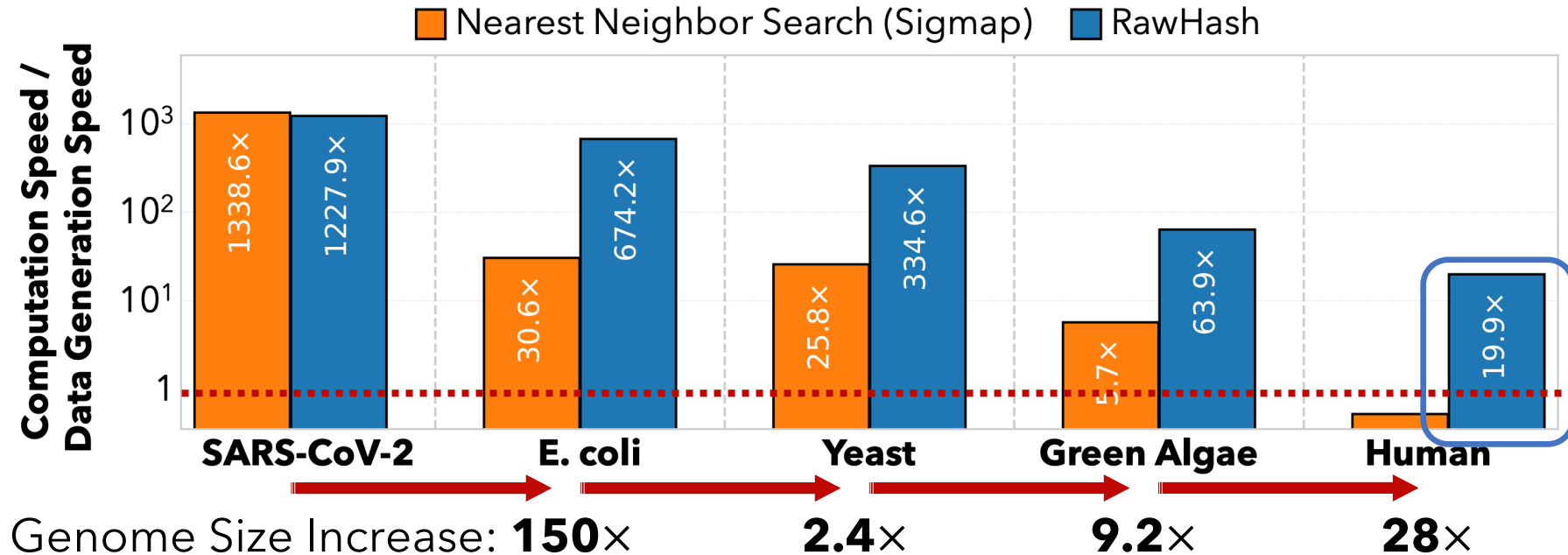
Better bucket utilization

RawHash Key Idea – Hash-based Matching



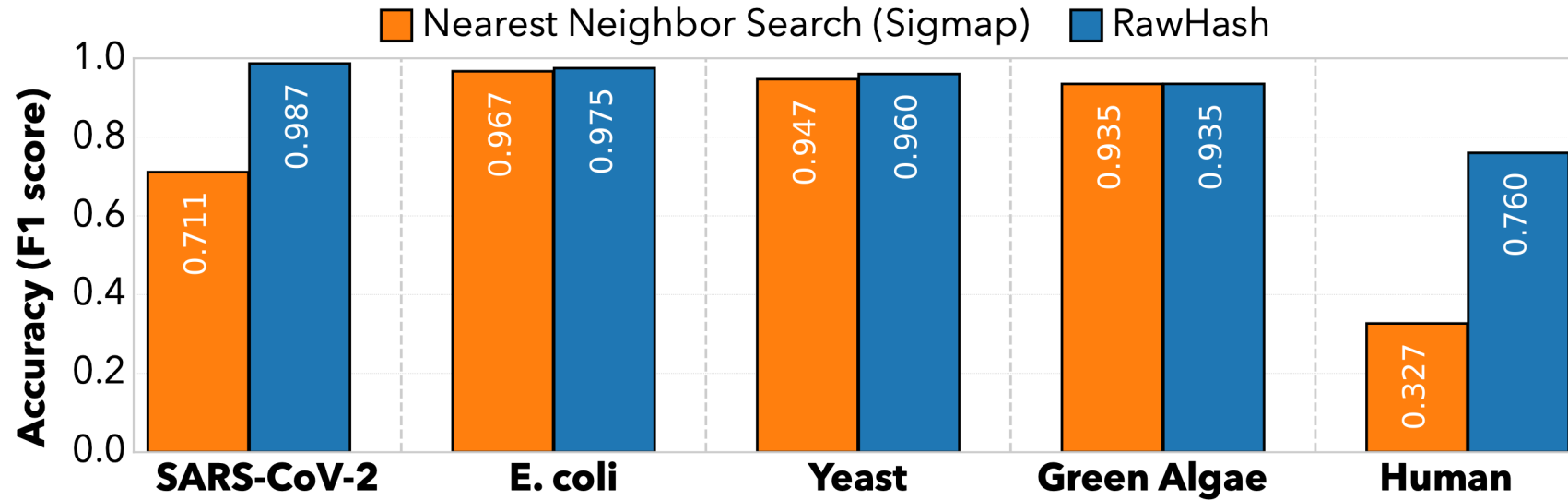
Key Contribution: **The first** hash-based search (and indexing) mechanism for raw signals

Real-Time Analysis Benefits from Fast Search



Real-time Analysis for Human Genomes:
Fast search instead of costly neighbor search

Fast Search Leads to Accurate Search



Best accuracy for all genomes:

Hash values span longer sequences than vectors ->
More unique and informative matches

(Almost) Identical accuracy to basecalled analysis:

Is the small difference because RawHash is even better?

RawHash is Open Source

 **RawHash** Public

Edit Pins

Unwatch 7

Fork 10

Starred 64

main 9 Branches 3 Tags

Go to file

Add file

Code

 **canfirtina** Rawsamble evaluation scripts cc9f2b9 · 29 days ago 49 Commits

extern	new R10 parameters	6 months ago
gitfigures	rawsamble info in README	2 years ago
src	new R10 parameters	6 months ago
test	Rawsamble evaluation scripts	29 days ago
.gitignore	Rawsamble evaluation scripts	29 days ago
.gitmodules	ZSTD sobmodule for POD5	3 years ago
LICENSE	R10 k-mer models can be parsed now as well.	3 years ago
Makefile	A minor fix in Makefile	3 years ago
README.md	new R10 parameters	6 months ago
code_of_conduct.md	Moving to multiple headers than a single one to improve ...	3 years ago

README Code of conduct GPL-3.0 license

 **RawHash**
<https://github.com/CMU-SAFARI/RawHash>
<https://github.com/STORMGroup/RawHash2>

About

RawHash can accurately and efficiently map raw nanopore signals to reference genomes of varying sizes (e.g., from viral to a human genomes) in real-time without basecalling. Described by Firtina et al. (published at https://academic.oup.com/bioinformatics/article/39/Supplement_1/i297/7210440).

academic.oup.com/bioinformatics/art...

bioinformatics

nanopore

seeding

segmentation

event-detection

genome-analysis

hash-tables

contamination

read-mapping

relative-abundances

nanopore-sequencing

nanopore-analysis-pipeline

nanopore-reads

nanopore-data

nanopore-minion

raw-signal

rawhash

raw-nanopore-signal-analysis

Readme

GPL-3.0 license

Code of conduct

Activity

Custom properties

64 stars

7 watching

STORM

26

Real-Time Raw Signal Analysis [Bioinformatics '24]

- [Can Firtina](#), Melina Soysal, Joël Lindegger, and Onur Mutlu,
"RawHash2: Mapping Raw Nanopore Signals Using Hash-Based Seeding and Adaptive Quantization"
[Bioinformatics](#), July 2024.
Paper [\[online\]](#) [\[pdf\]](#) [\[code\]](#)

Bioinformatics, 2024, **40(8)**, btae478
<https://doi.org/10.1093/bioinformatics/btae478>
Advance Access Publication Date: 30 July 2024
Applications Note



Sequence analysis

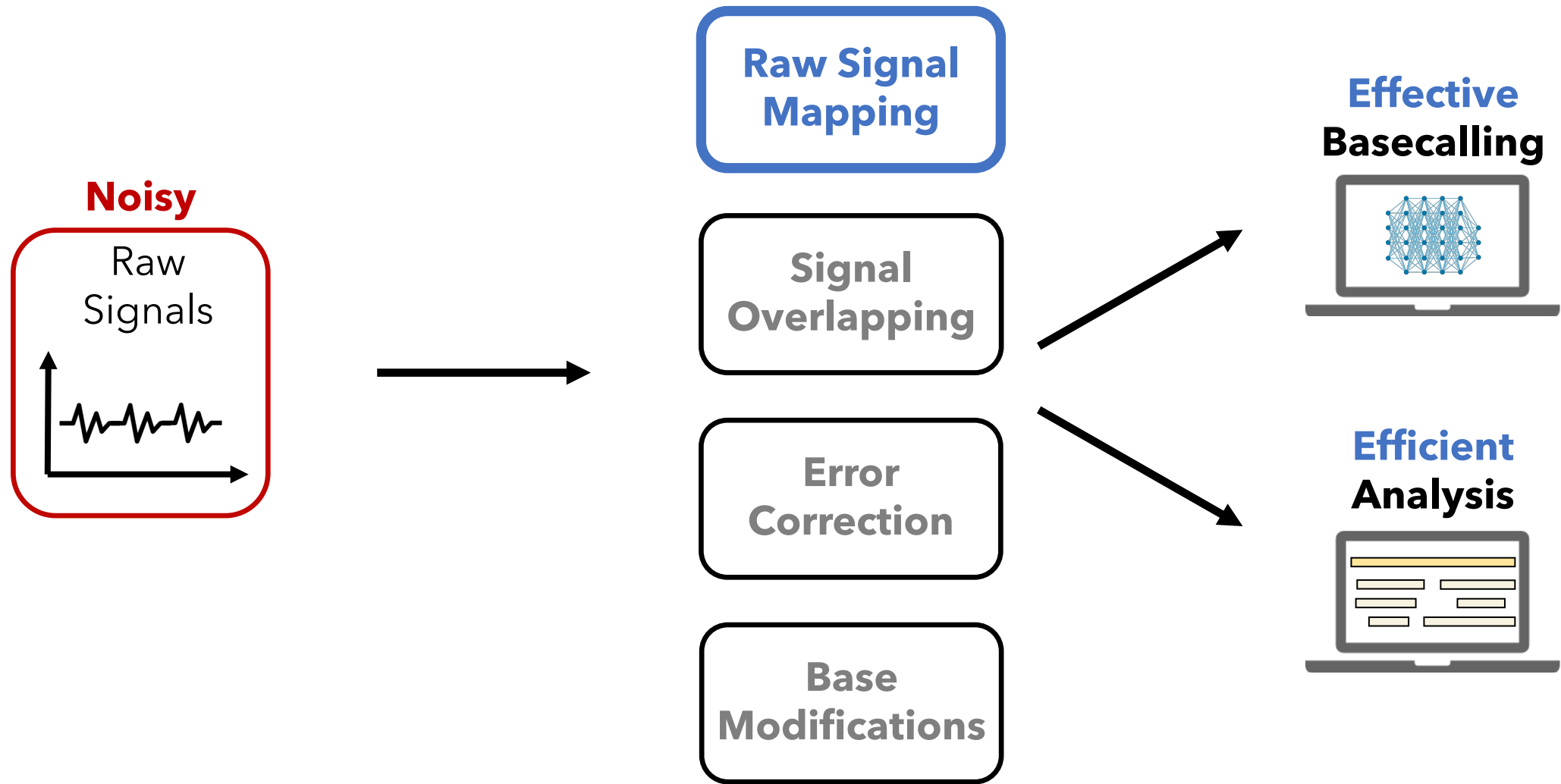
RawHash2: mapping raw nanopore signals using hash-based seeding and adaptive quantization

Can Firtina ^{1,*}, Melina Soysal ¹, Joël Lindegger ¹, Onur Mutlu ^{1,*}

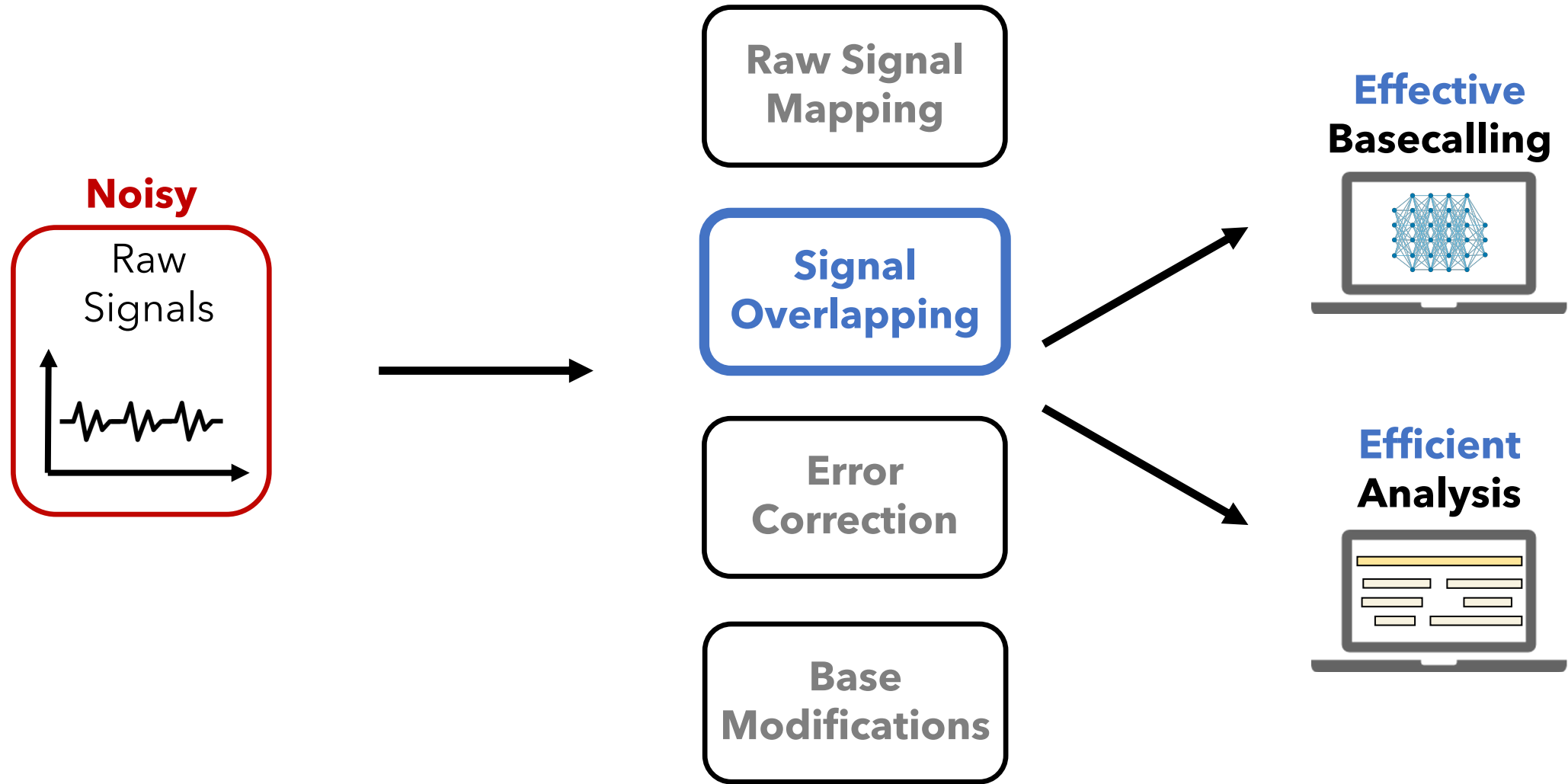
¹Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich 8092, Switzerland

*Corresponding authors. Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich 8092, Switzerland.
E-mail: firtinac@ethz.ch (C.F.); omutlu@ethz.ch (O.M.)

Raw Signal Mapping



Signal Overlapping



New Directions: Overlapping Raw Signals

- **Rawsamble: Overlapping Raw Nanopore Signals using a Hash-based Seeding Mechanism**

Can Firtina, Maximilian Mordig, Harun Mustafa, Sayan Goswami, Nika Mansouri Ghiasi, Stefano Mercoglian, Furkan Eris, Joel Lindegger, Andre Kahles, and Onur Mutlu

[Bioinformatics](#), February 2026

[\[online link\]](#) [\[pdf\]](#) [\[code\]](#)

[ISMB 2024](#) Talk [\[video\]](#) [\[pptx\]](#) [\[pdf\]](#)

[CSHL 2024 \(Biological Data Science\)](#) Talk [\[video\]](#) [\[pptx\]](#) [\[pdf\]](#)

Social Media Thread [\[Twitter \(X\)\]](#) [\[LinkedIn\]](#)

Bioinformatics, 2026, **42**(3), btag087
<https://doi.org/10.1093/bioinformatics/btag087>
Published: 26 February 2026
Original Paper | Sequence Analysis

OXFORD

Rawsamble: overlapping raw nanopore signals using a hash-based seeding mechanism

Can Firtina^{1,2,*}, Maximilian Mordig^{3,4}, Harun Mustafa^{3,5,6}, Sayan Goswami³,
Nika Mansouri Ghiasi¹, Stefano Mercoglian¹, Furkan Eris¹, Joel Lindegger¹, André Kahles^{3,5,6,*},
Onur Mutlu^{1,*}

¹Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich, 8092, Switzerland

²Department of Computer Science, University of Maryland, MD 20742, United States

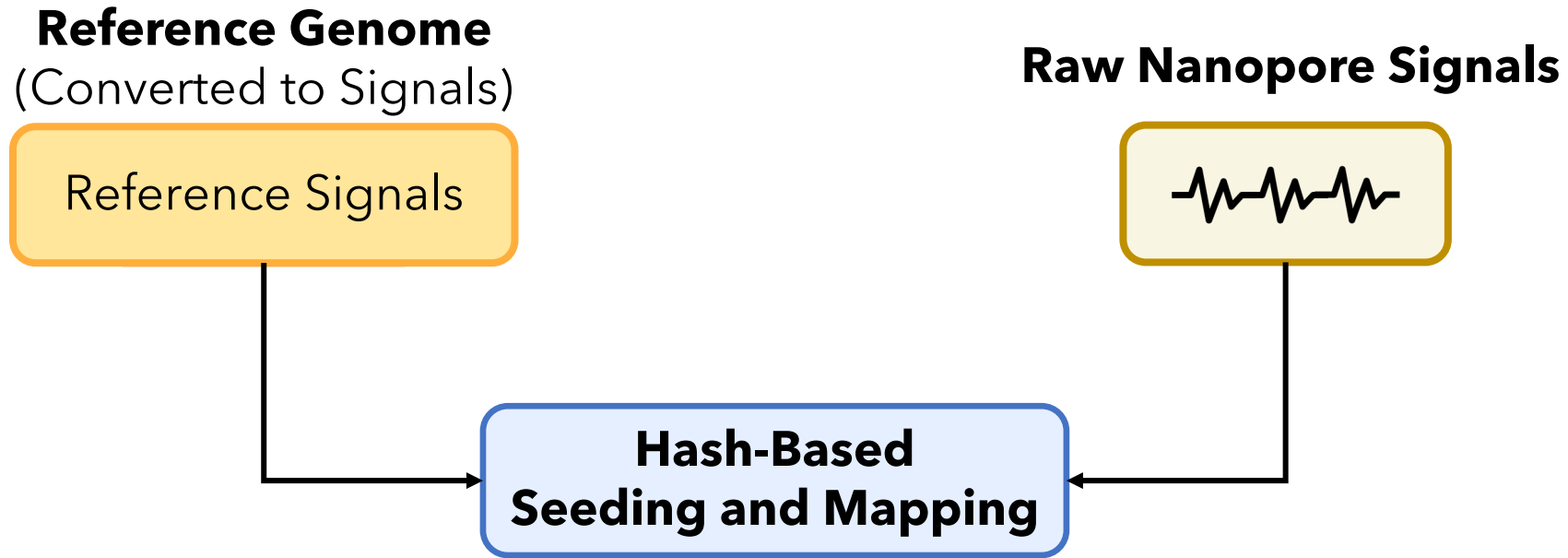
³Department of Computer Science, ETH Zurich, Zurich, 8092, Switzerland

⁴Max Planck Institute for Intelligent Systems, Tübingen, 72076, Germany

⁵University Hospital Zurich, Zurich, 8091, Switzerland

⁶Swiss Institute of Bioinformatics, Zurich, 8057, Switzerland

Beyond Reference Mapping: Overlapping

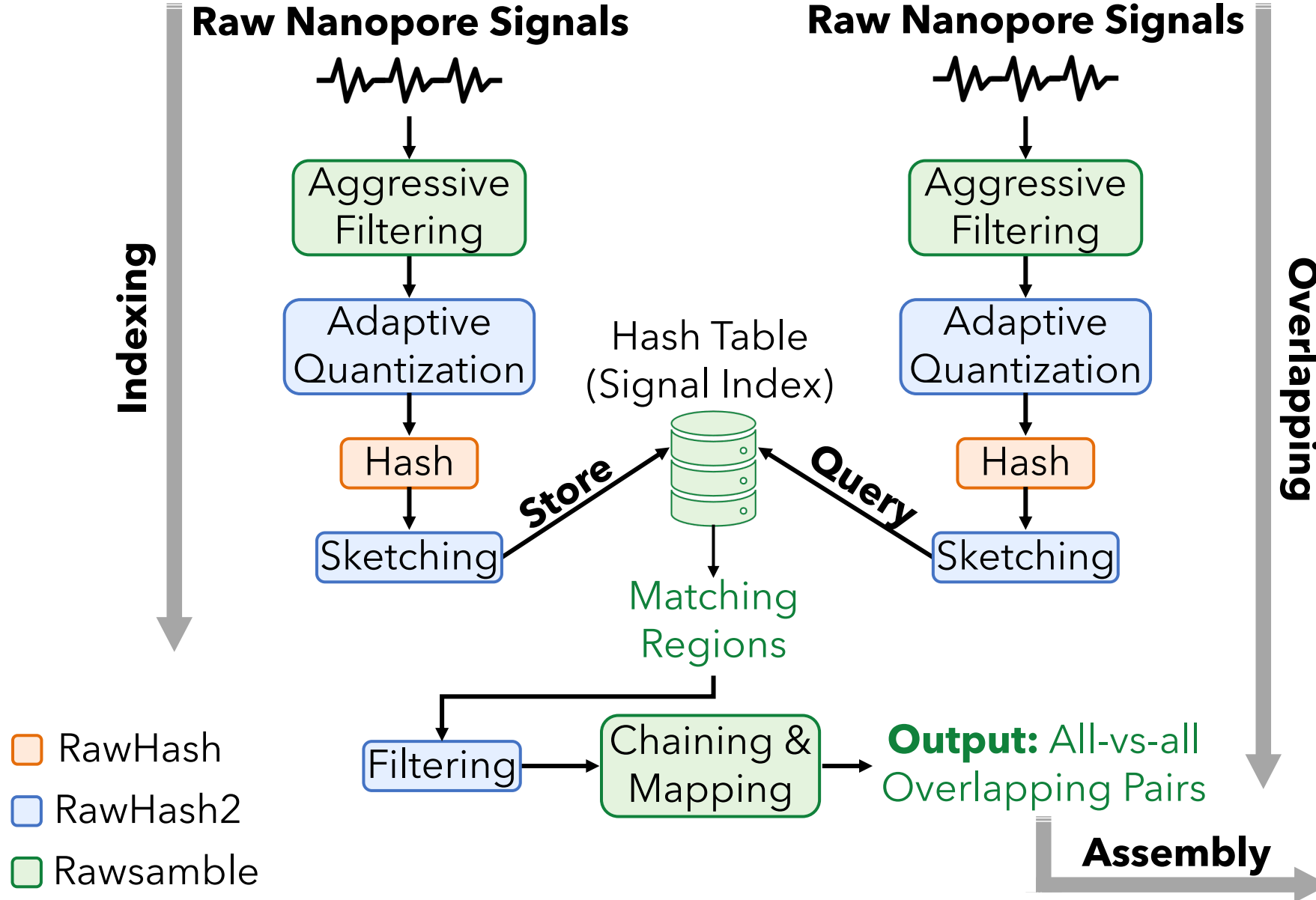


 **Challenge:** Reference genome is not always available

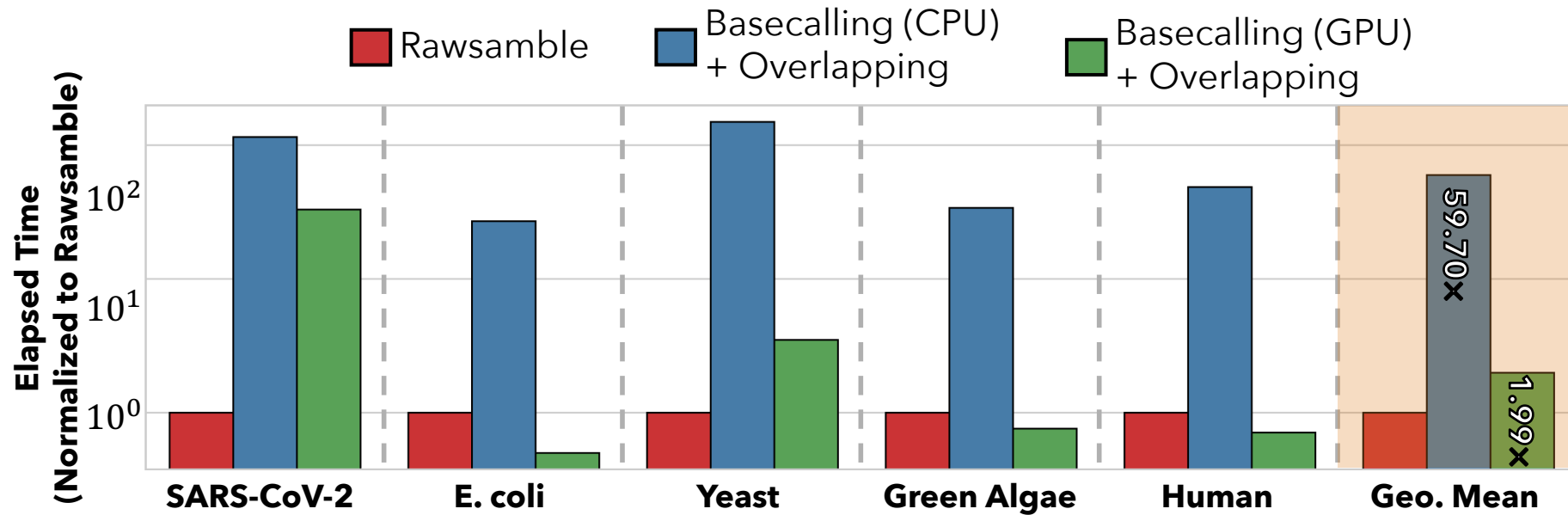
 **Assembly:** Constructing genome from **overlapping reads**

 Existing solutions cannot find overlapping reads **without basecalling**

Rawsample Overview



Eliminating Basecalling from the Pipeline

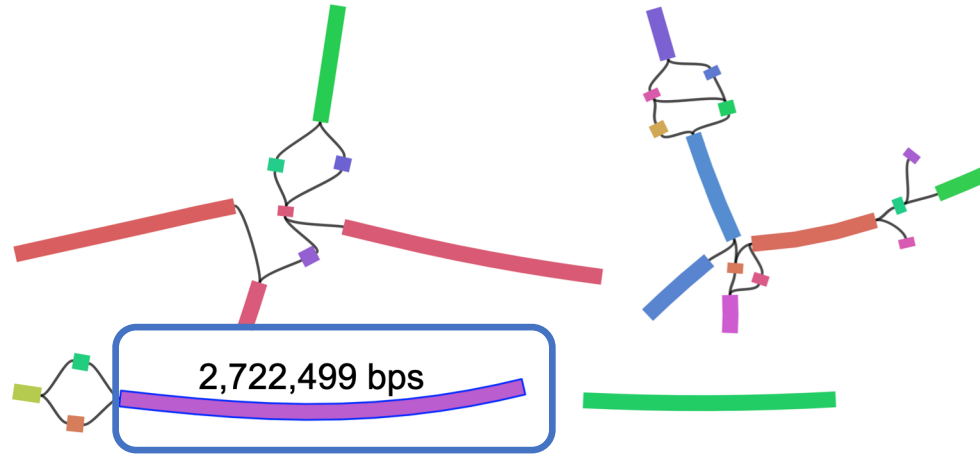


Average speedup of 60× than CPU-based basecalling:
Eliminating basecalling leaves a significant room
for performance improvements

Average speedup of 2× than the **GPU-based basecalling**:
GPU vs. GPU comparison is likely to provide even better results

Key Results – Miniasm Assemblies from Overlaps

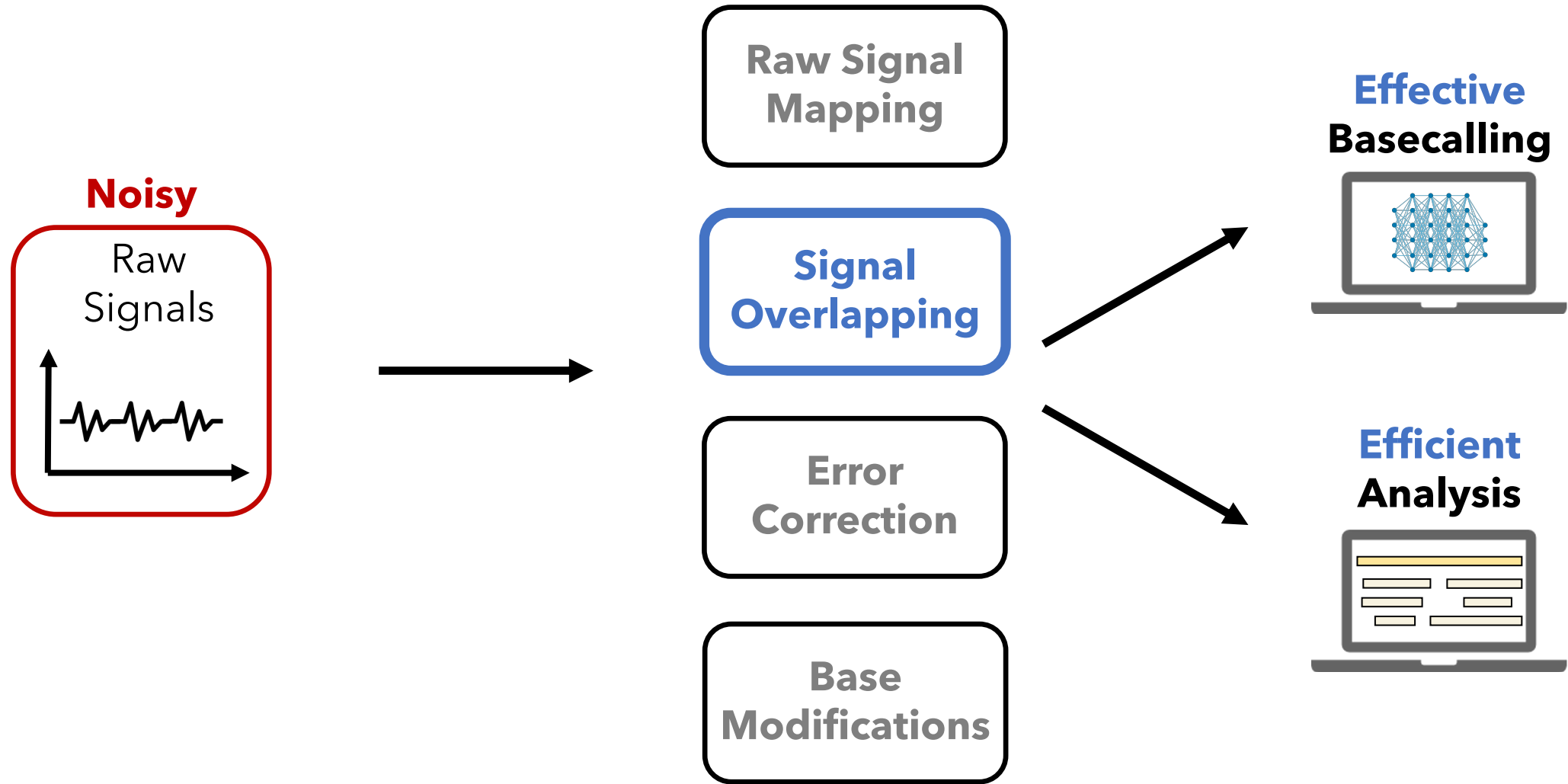
**E. coli Assembly
(From the Rawsamble Overlaps)**



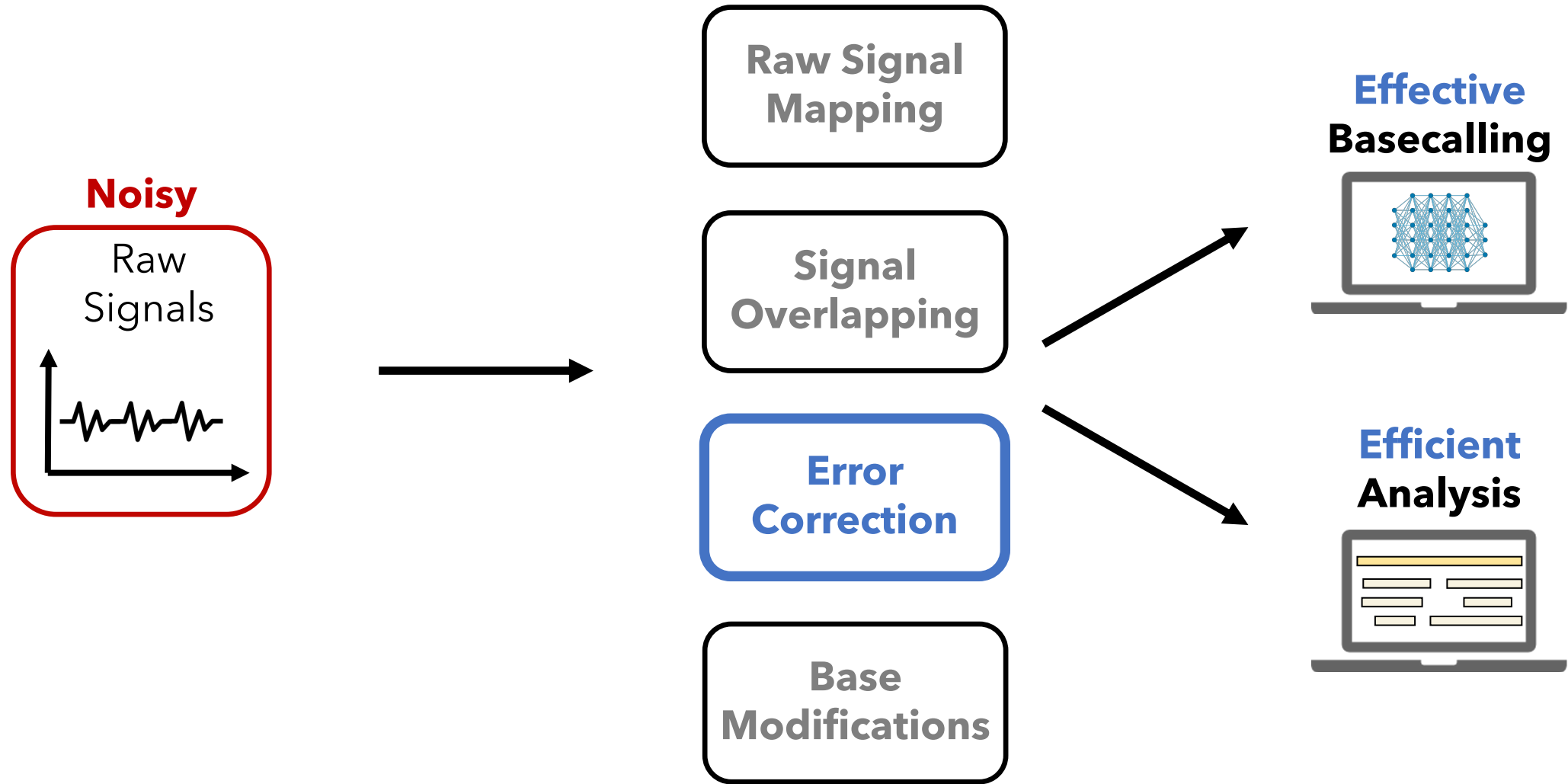
~400× longer assembled pieces
than the average read length:
Informative overlaps lead to
longer paths

**It is time to rethink if we want to translate
individual reads separately**

Signal Overlapping



Signal Denoising / Error Correction



Correcting Raw Signals [arXiv '26 & ISMB'26]

- **CRANE: Correcting Errors in Raw Nanopore Signals Using Hidden Markov Models**

Simon Ambrozak, Ulysse McConnell, Bhargav Srinivasan, Burak Ozkan, Ernest Zhang,
and Can Firtina

[arXiv](#), March 2026

[\[online link\]](#) [\[pdf\]](#) [\[code\]](#)

Talk at ISMB 2026 (HiTSeq Track) on Tuesday, July 14!

CRANE: Correcting Errors in Raw Nanopore Signals Using Hidden Markov Models

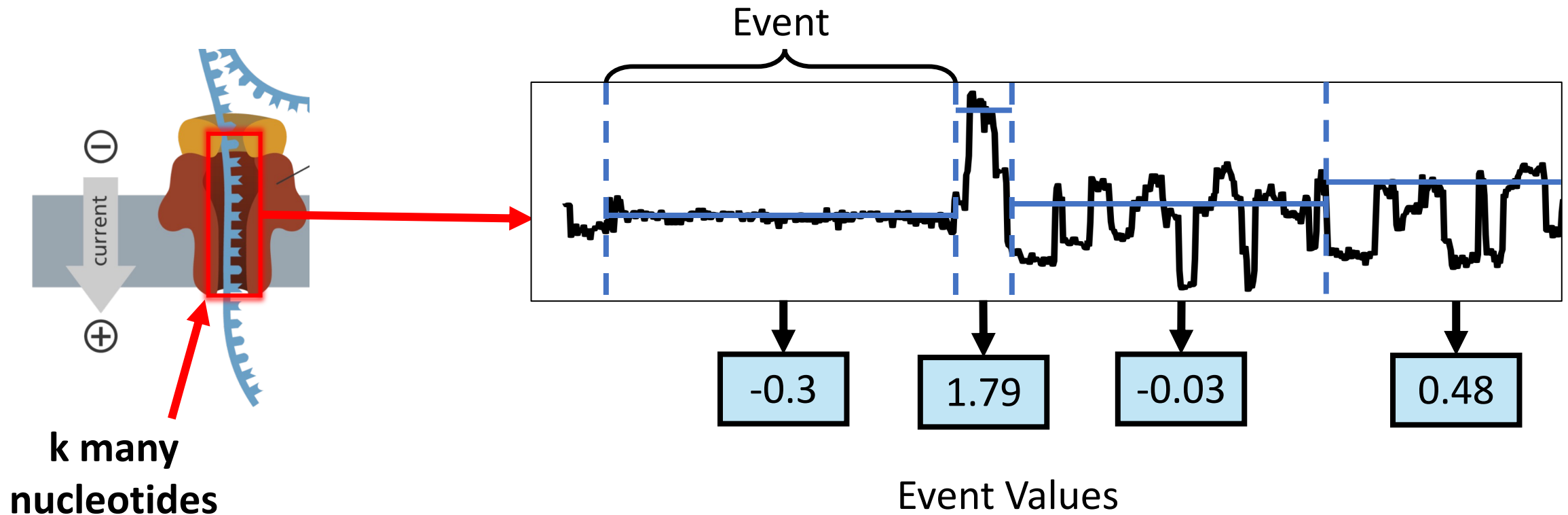
Simon Ambrozak¹ Ulysse McConnell^{1,2} Bhargav Srinivasan¹

Burak Ozkan^{1,3} Ernest Zhang¹ Can Firtina¹

¹*University of Maryland* ²*ETH Zurich* ³*Bilkent University*

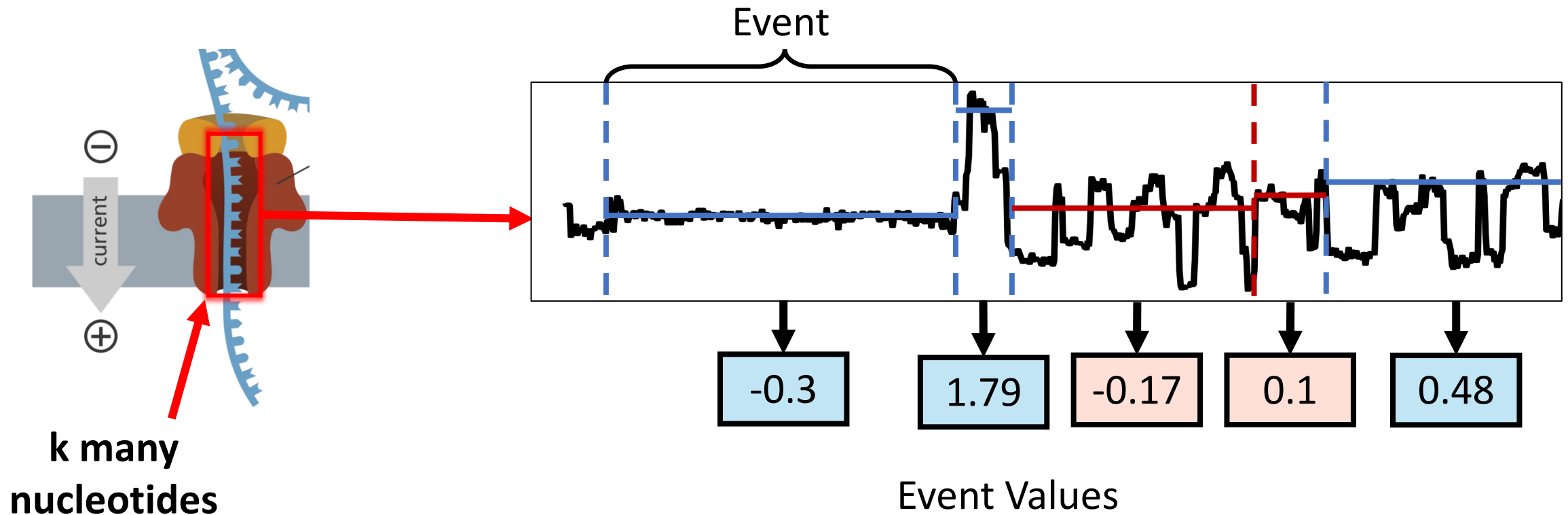
Segmenting Raw Signals Combats Noise

- Multiple measurements are taken for each k-mer
- **Segmentation:** Identify regions of signals corresponding to a particular k-mer



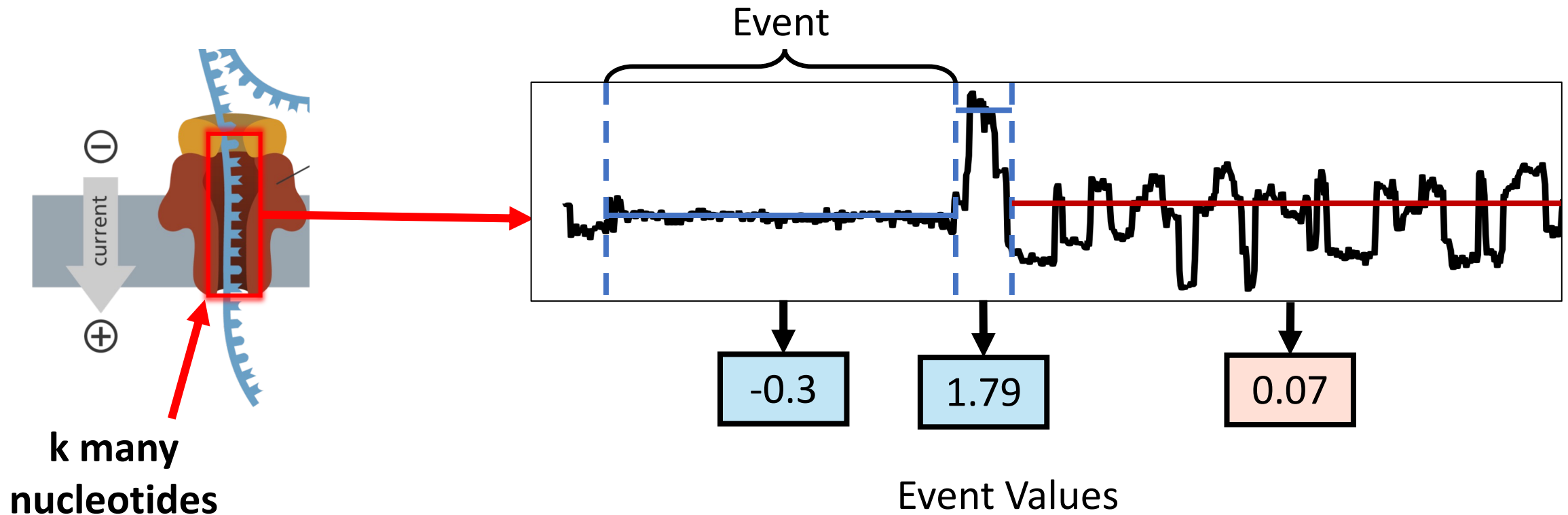
Segmenting Raw Signals Combats Noise

- Multiple measurements are taken for each k-mer
- **Segmentation:** Identify regions of signals corresponding to a particular k-mer
- **Errors:** Noise can distort events, causing segmentation errors

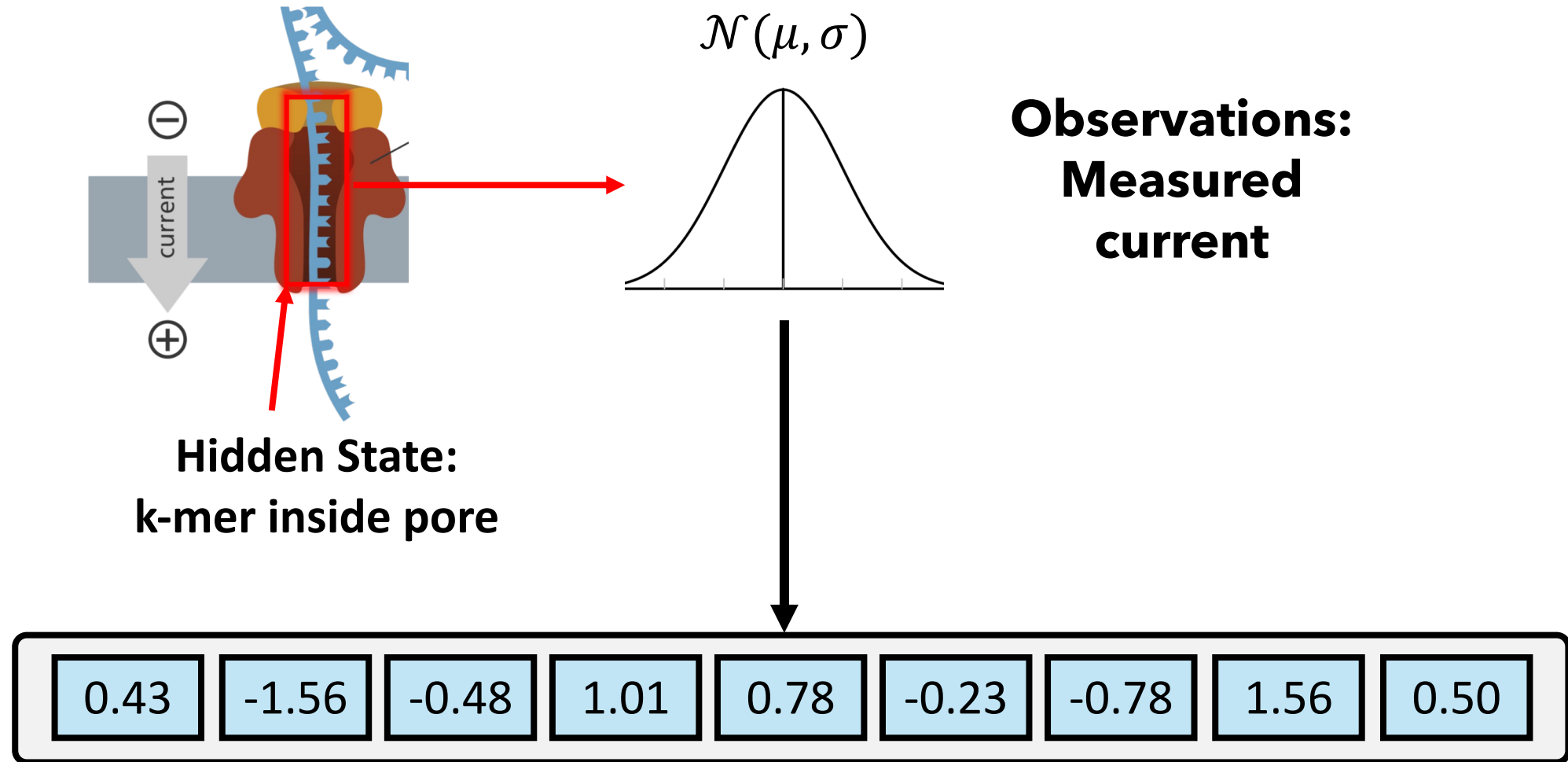


Segmenting Raw Signals Combats Noise

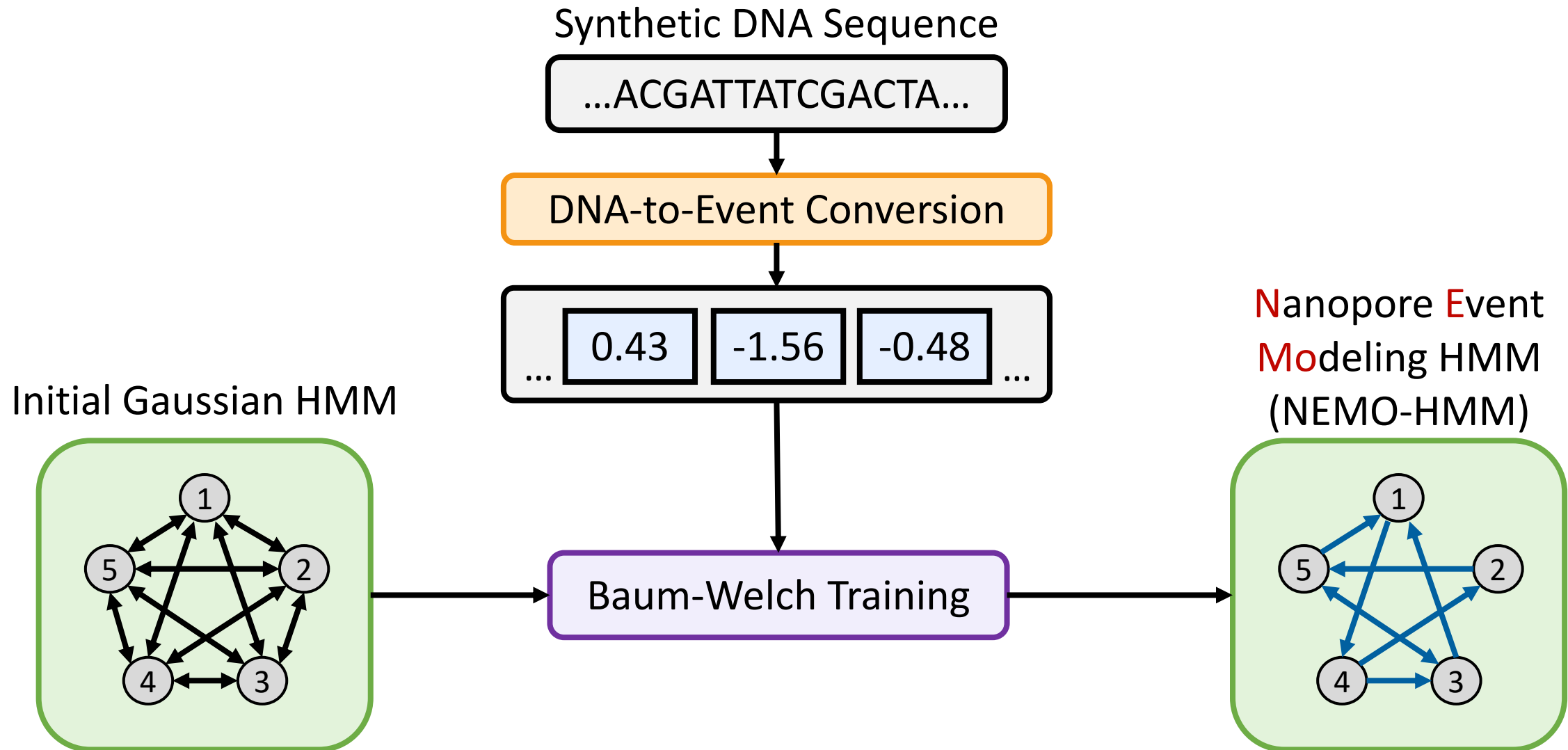
- Multiple measurements are taken for each k-mer
- **Segmentation:** Identify regions of signals corresponding to a particular k-mer
- **Errors:** Noise can distort events, causing segmentation errors



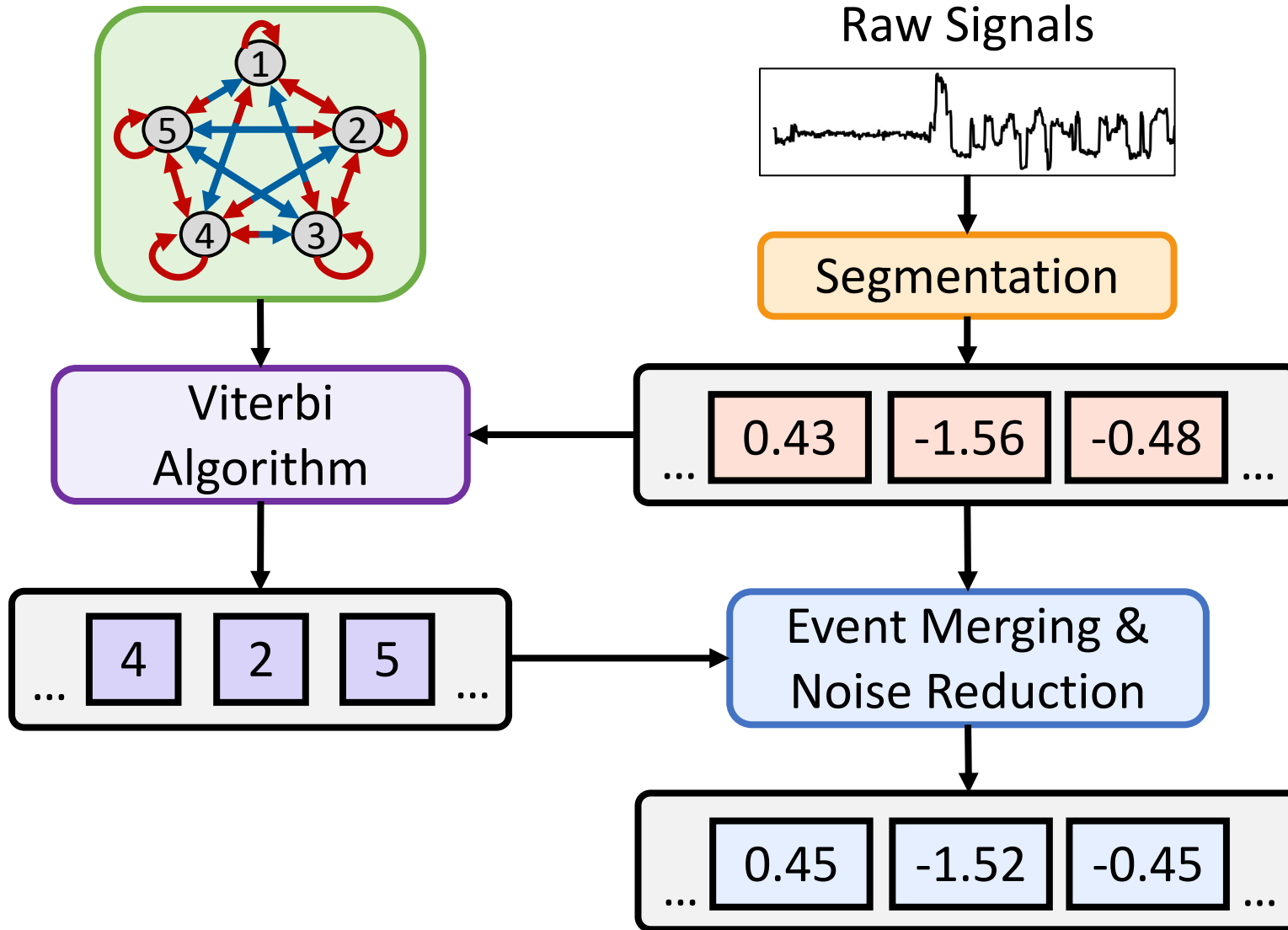
Key Observation: HMMs Can Model Nanopores



Key Idea: Train An HMM To Model A Nanopore

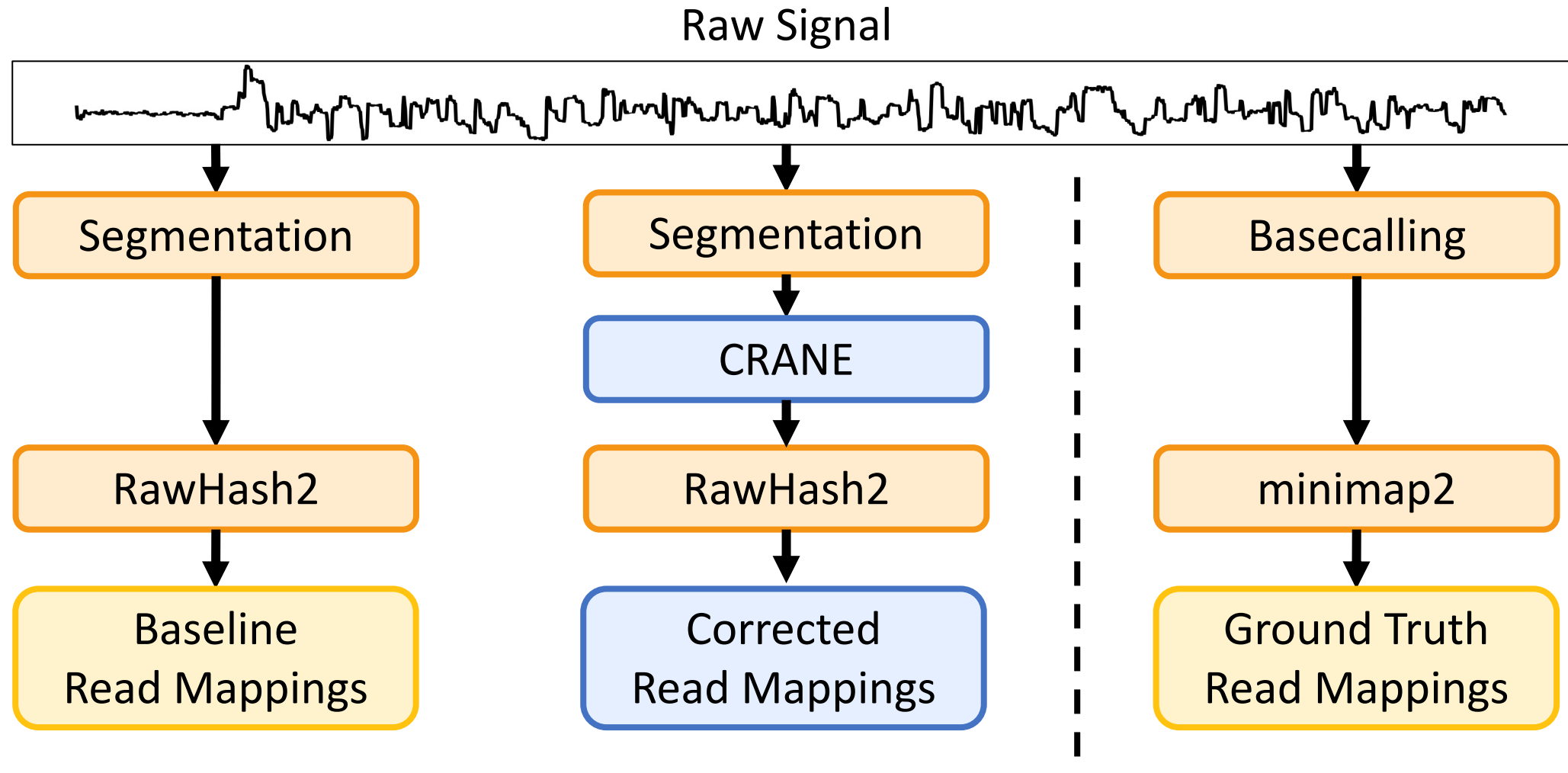


CRANE Error Correction



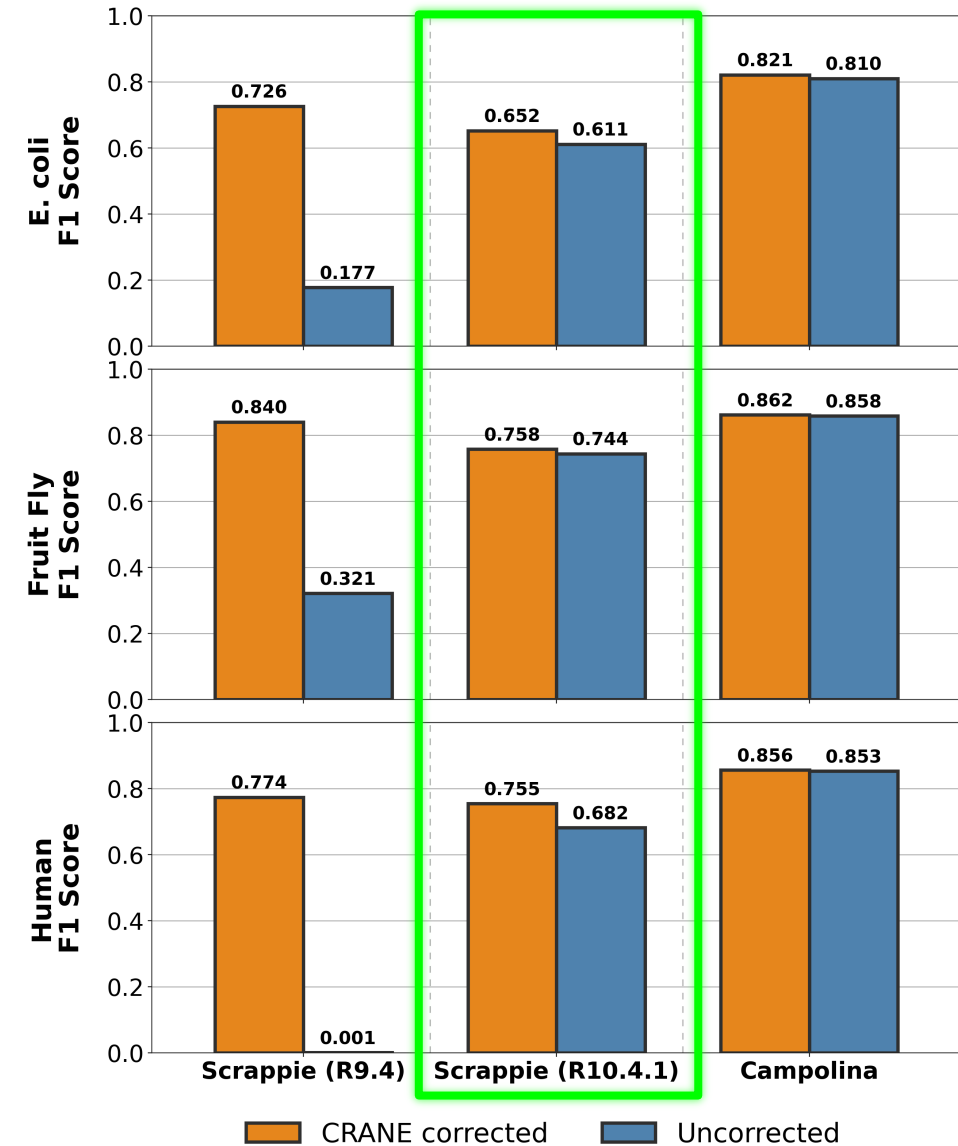
We Evaluate Our Approach Using RawHash2

- RawHash2 is the state-of-the-art **raw signal read-mapping** tool

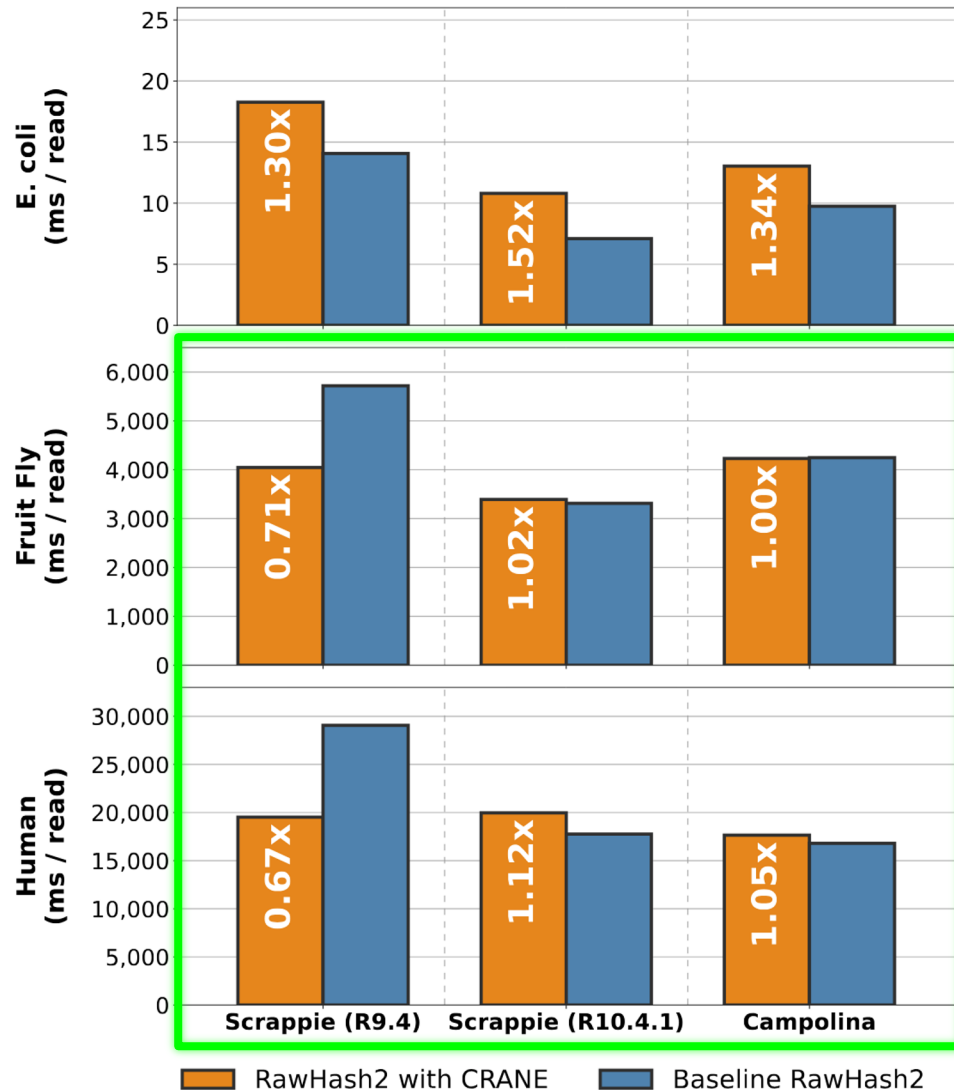


Accuracy Results - RawHash2 Read-mapping w/o HPC

CRANE improves state of the art
read mapping F1 score by
up to 11%



Runtime Results - CRANE + RawHash2 w/o HPC



CRANE accounts for **<1%** of total runtime for human genome

Source Code

- Easily reproduce results
- Train your own HMMs
- Parameter search scripts



[Code](#)

CRANEPublic

Edit PinsWatch0Fork1Starred3

main2 Branches0 Tags

Go to fileTAdd fileCode

Simon AmbrozakImproved navigation in READMEs6090f04 · 2 months ago51 Commits

extern/kmer_models	First commit	4 months ago
src	Fix some README path bugs, remove legacy code	2 months ago
test	Improved navigation in READMEs	2 months ago
train	Improved navigation in READMEs	2 months ago
.gitignore	Add segmentation module	4 months ago
.gitmodules	Migrate to cpp for segmentation	3 months ago
LICENSE	Initial commit	4 months ago
Makefile	Updated readmes, added QoL	2 months ago
README.md	Improved navigation in READMEs	2 months ago

READMEGPL-3.0 license

CRANE Overview

About

CRANE is an HMM-based tool to correct raw nanopore signal events. To do this, it 1) Trains an HMM on synthetic and experimental nanopore event data and 2) uses the Viterbi algorithm and the learned HMM to correct novel signals. Detailed information on the CRANE tool can be found here: <https://arxiv.org/abs/2603.20420>

arxiv.org/abs/2603.20420

Readme

GPL-3.0 license

Activity

Custom properties

3 stars

0 watching

1 fork

Report repository

Releases

Correcting Raw Signals [arXiv '26 & ISMB'26]

- **CRANE: Correcting Errors in Raw Nanopore Signals Using Hidden Markov Models**

Simon Ambrozak, Ulysse McConnell, Bhargav Srinivasan, Burak Ozkan, Ernest Zhang,
and Can Firtina

[arXiv](#), March 2026

[\[online link\]](#) [\[pdf\]](#) [\[code\]](#)

Talk at ISMB 2026 (HiTSeq Track) on Tuesday, July 14!

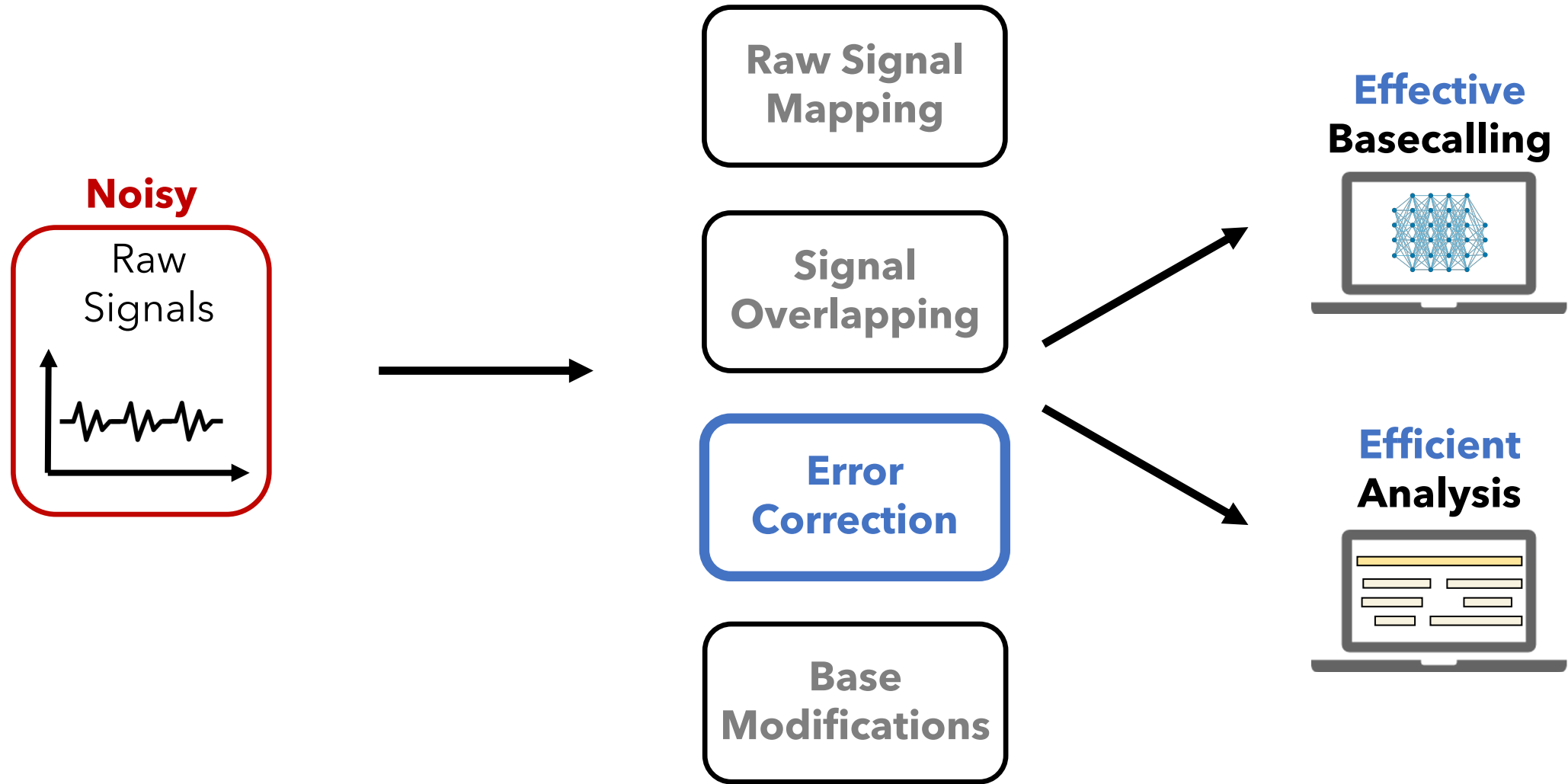
CRANE: Correcting Errors in Raw Nanopore Signals Using Hidden Markov Models

Simon Ambrozak¹ Ulysse McConnell^{1,2} Bhargav Srinivasan¹

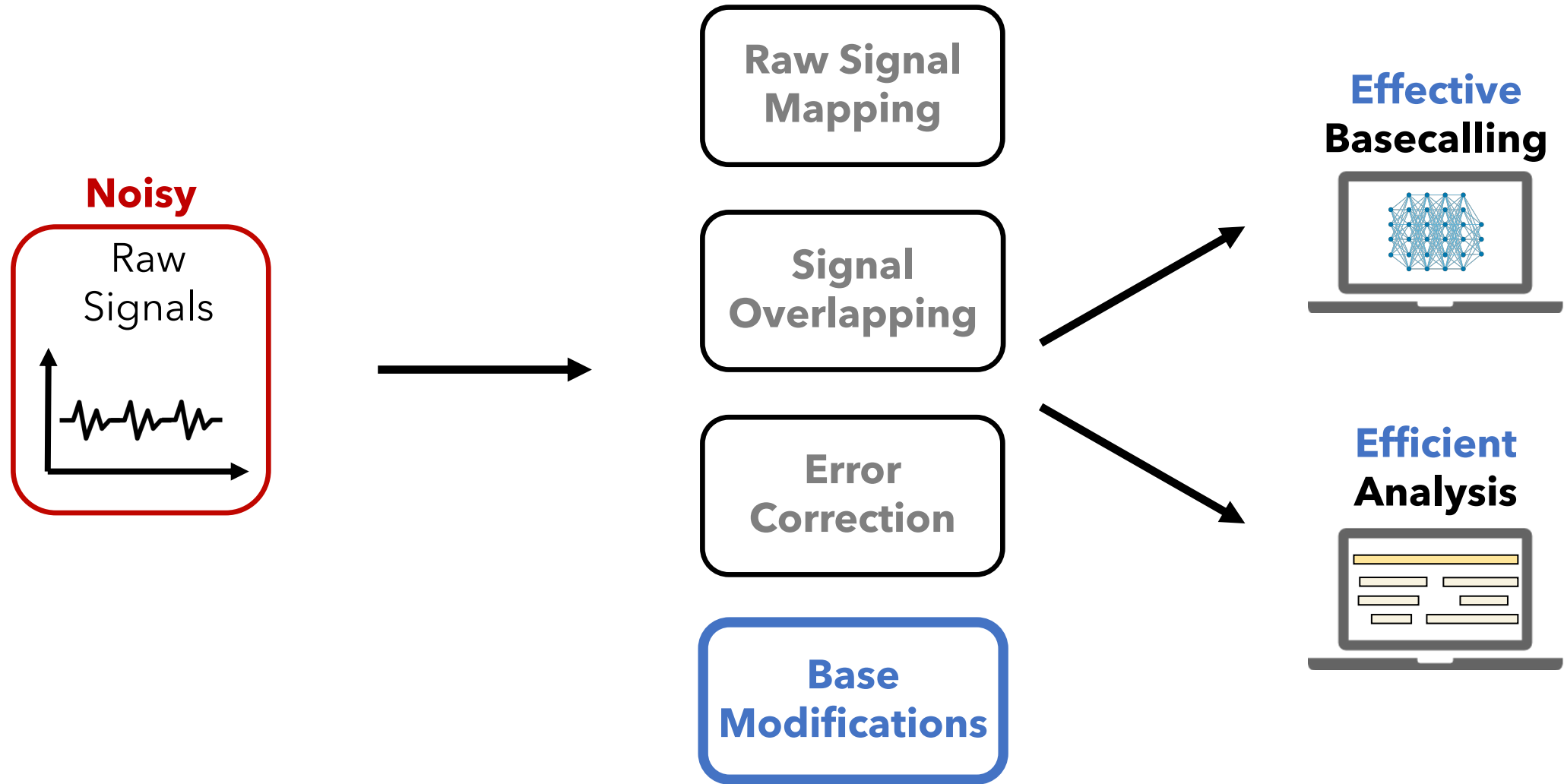
Burak Ozkan^{1,3} Ernest Zhang¹ Can Firtina¹

¹*University of Maryland* ²*ETH Zurich* ³*Bilkent University*

Signal Denoising / Error Correction

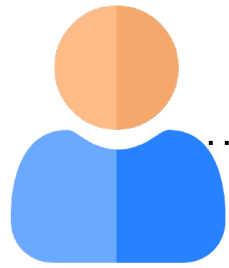


Identify Base Modifications



Motivation: Understanding Beyond DNA Letters

Human (Reference) Genome: 3.2 Billion bps (haploid)



CGAAATGCC...AGATTAAACGCT...TGCCCTAA...GAATGGCGT

5
m
C

5
m
C

5
m
C

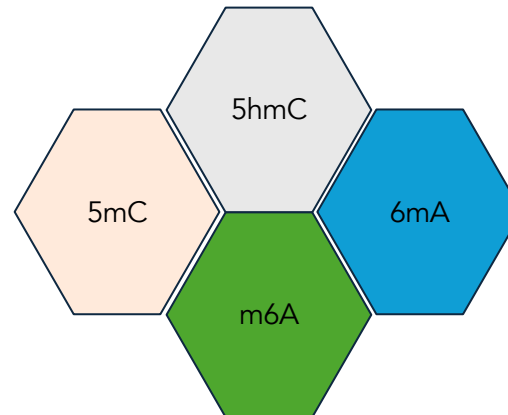
6
m
A

6
m
A

6
m
A

m
6
A

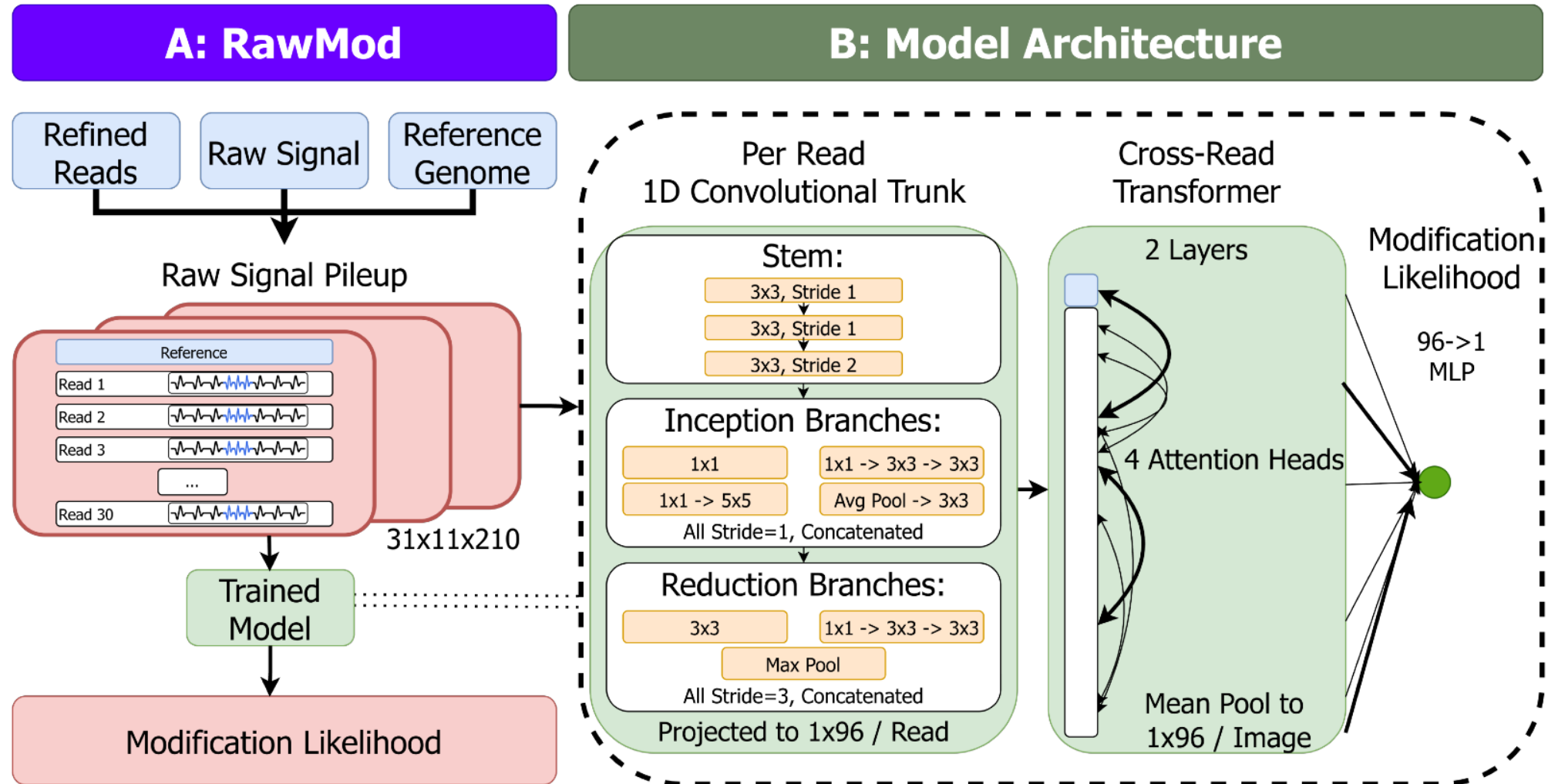
6
m
A



DNA Modification
Profiles

Our Approach: RawMod

- Pileup signals to identify if a position is modified/unmodified (binary)



Poster Presentation at ISMB on Monday (#A-153)

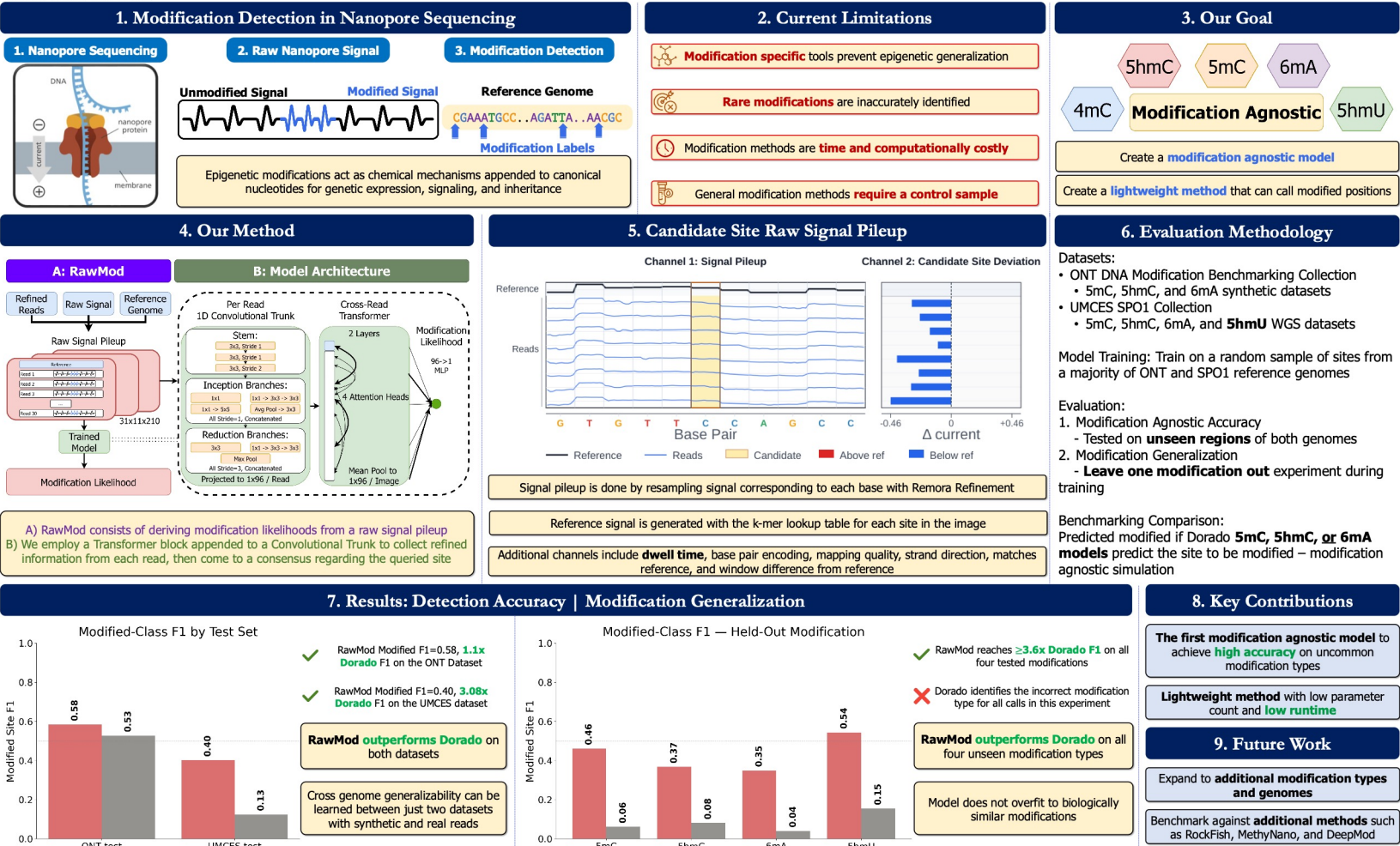
RawMod: A Deep Learning Approach to Modification Agnostic Classification in Nanopore Sequencing

Bhargav Srinivasan, Simon Ambrozak, Can Firtina
University of Maryland Department of Computer Science

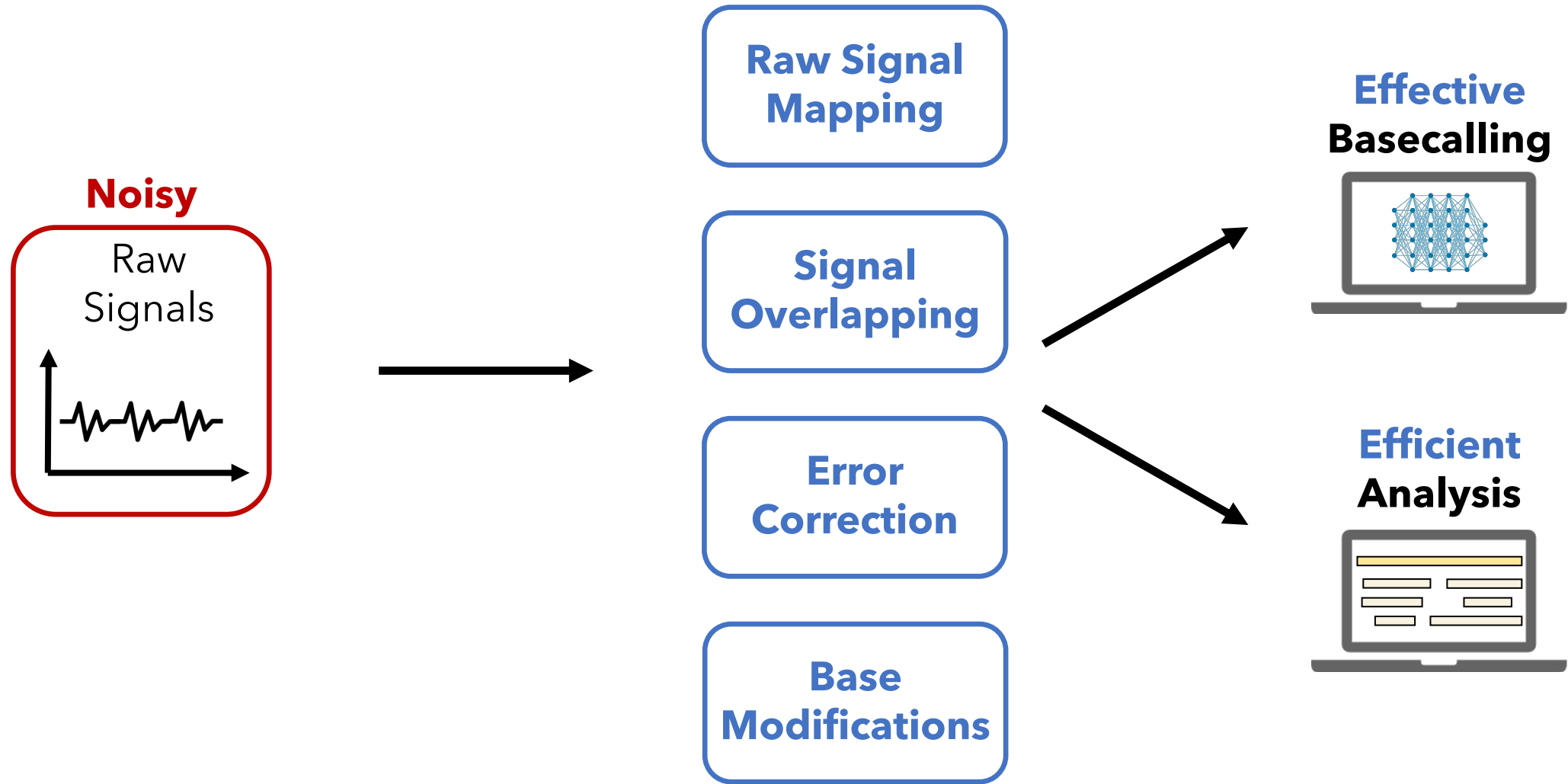


Contact: bds062@umd.edu

STORM

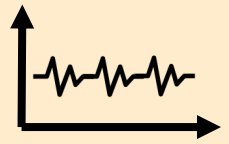


Key Perspective: Analyze Signals (Before Translation)

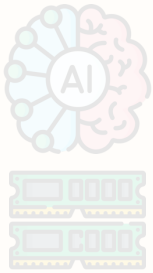


STORM Research Group

Research Directions



Real-Time and Portable
Genomic Signal Analysis



Data-centric and
AI-Driven Solutions to
Analyze Genomic Data



Accurate and Fast
Algorithms Scalable to
Large Volumes of
Genomic Data

**Today's
Talk**



Goal

Enable fundamentally
better integration of
biological data analysis
in our **everyday lives**



Lost in Translation: Harnessing the Power of Nanopore Electrical Signals in Genomics

Computational Genomics Summer Institute (CGSI 2026)
Rancho Palos Verdes, CA, USA. July 9 2026

Can Firtina | canfirtina@gmail.com | cs.umd.edu/~firtina

Assistant Professor of Computer Science at UMD



Backup Slides

Basecalling using PIM [MICRO '23]

- Taha Shahroodi, Gagandeep Singh, Mahdi Zahedi, Haiyu Mao, Joel Lindegger, Can Firtina, Stephan Wong, Onur Mutlu, and Said Hamdioui,
"Swordfish: A Framework for Evaluating Deep Neural Network-based Basecalling using Computation-In-Memory with Non-Ideal Memristors"
*Proceedings of the 56th International Symposium on Microarchitecture (**MICRO**), Toronto, ON, Canada, November 2023.*
[\[Slides \(pptx\) \(pdf\)\]](#)
[\[arXiv version\]](#)

Swordfish: A Framework for Evaluating Deep Neural Network-based Basecalling using Computation-In-Memory with Non-Ideal Memristors

Taha Shahroodi¹ Gagandeep Singh^{2,3} Mahdi Zahedi¹ Haiyu Mao³ Joel Lindegger³ Can Firtina³
Stephan Wong¹ Onur Mutlu³ Said Hamdioui¹

¹TU Delft ²AMD Research ³ETH Zürich

DNN Hardware Acceleration

DNN execution is dominated by:

Vector-Matrix
Multiplication (VMM)

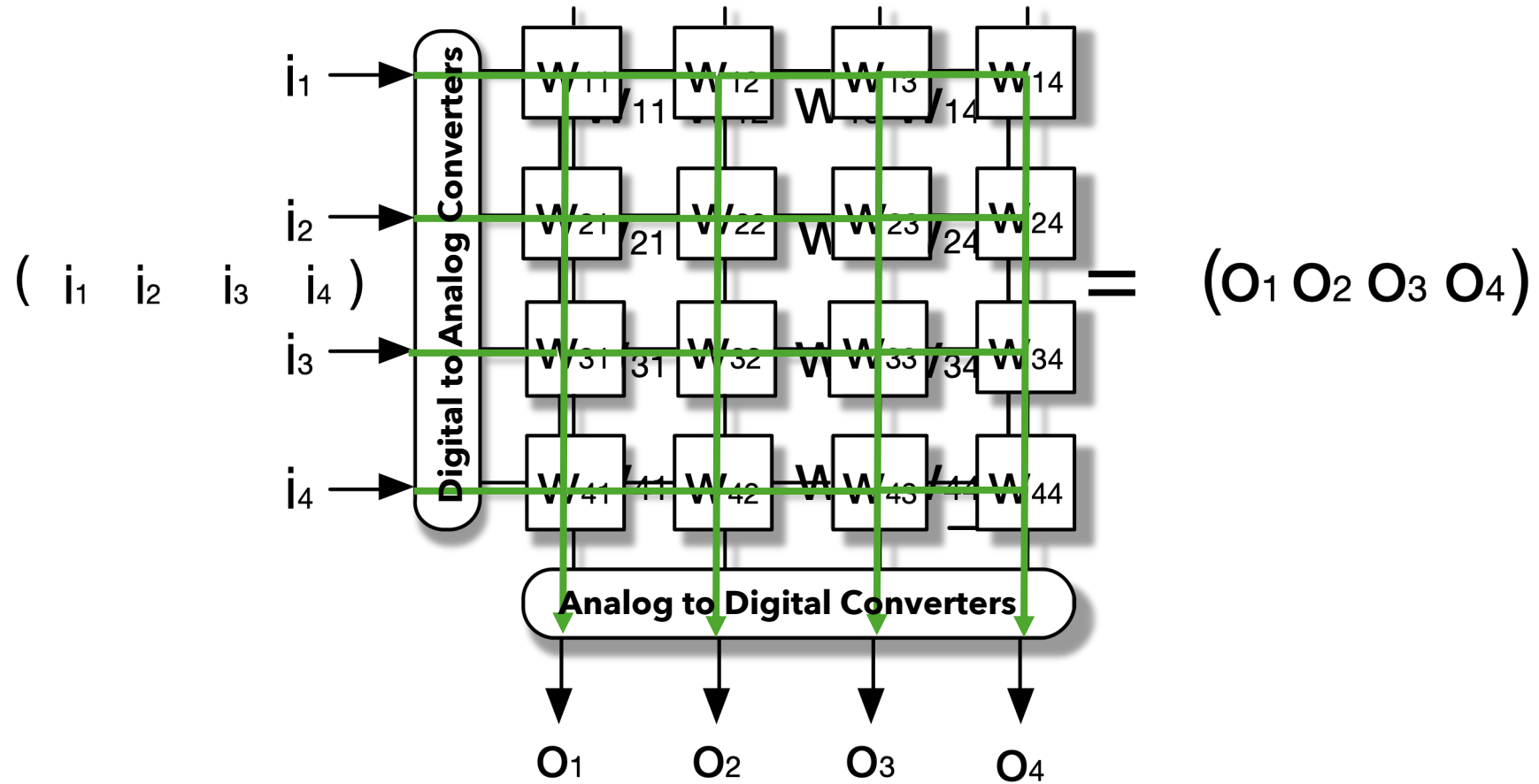
Data movement
between memory and
accelerator

Memristor-based PIM for DNN Acceleration

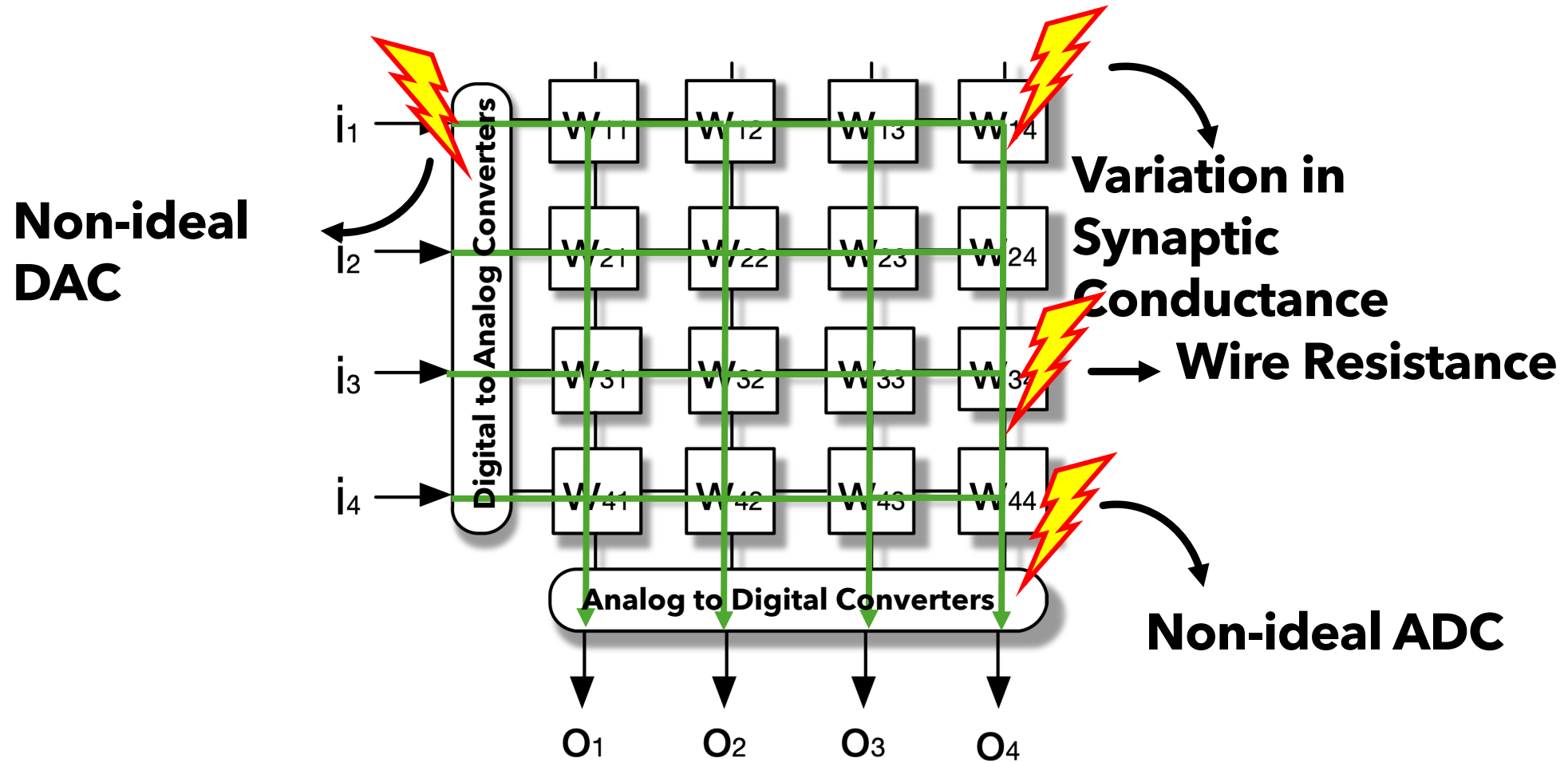
Memristor-based
crossbars support
VMM

Processing in Memory
(PIM) minimizes data
movement

VMM in Memristor-based Crossbars

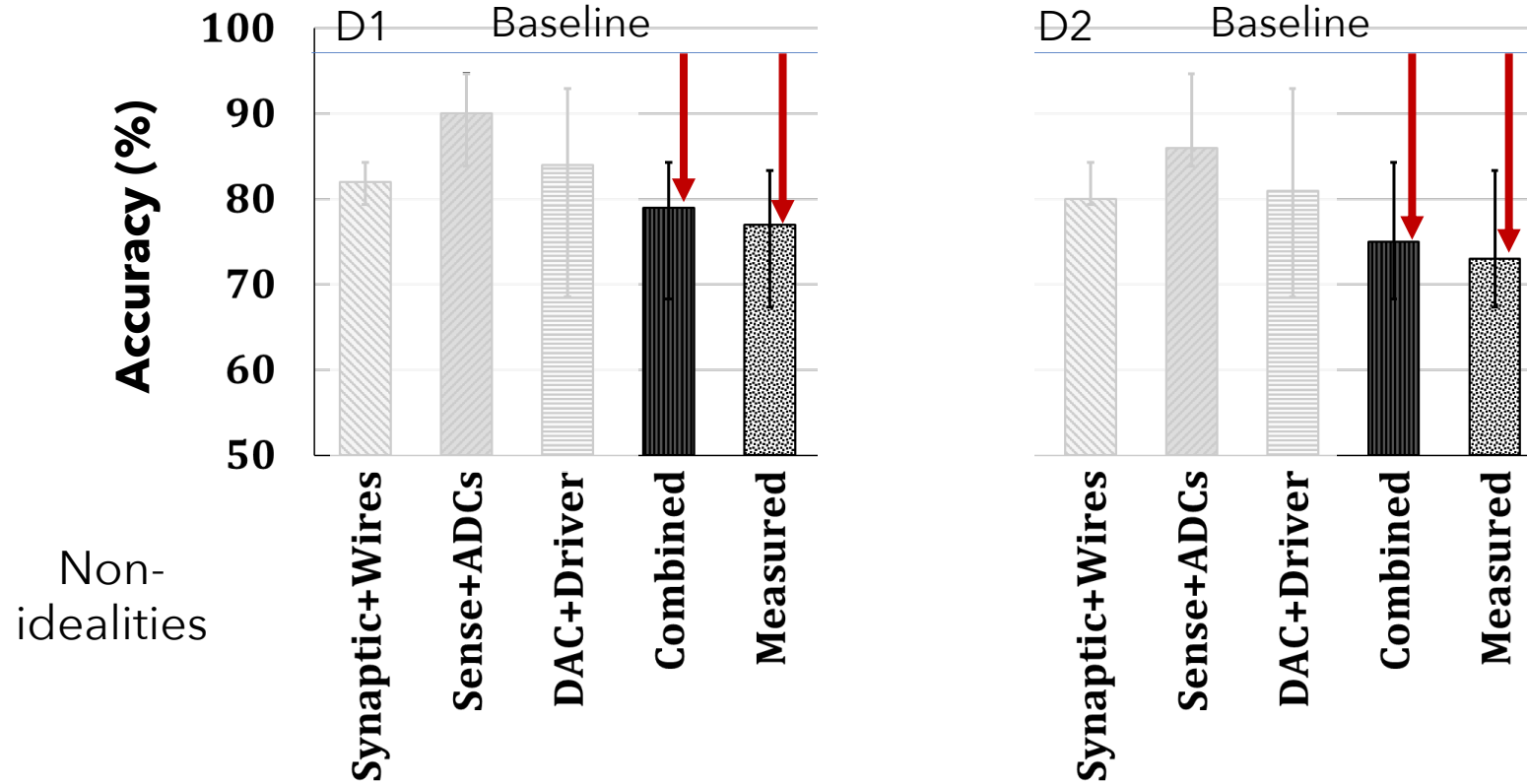


Non-idealities in Memristor-based Crossbars



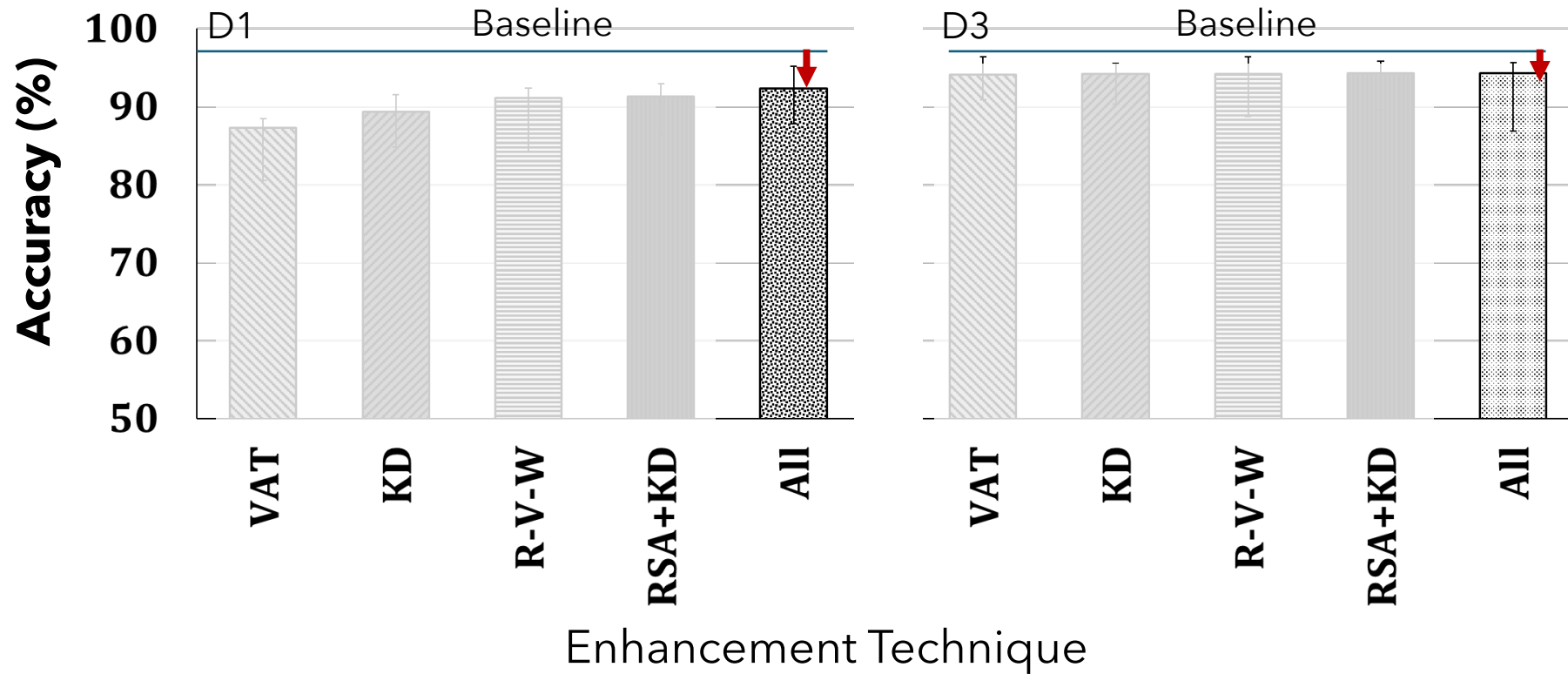
Subpar utilization of trained models
due to non-idealities

Accuracy: All Non-idealities without Mitigation



Combined non-idealities leads to significant accuracy loss (**>18%**)

Accuracy: Enhancement Techniques on All Non-idealities



Accuracy loss drops to 6% from 18% after handling all non-idealities

ML-based Genome Analysis via PIM [MICRO '22]

- Haiyu Mao, Mohammed Alser, Mohammad Sadrosadati, Can Firtina, Akanksha Baranwal, Damla Senol Cali, Aditya Manglik, Nour Almadhoun Alserr, and Onur Mutlu, **"GenPIP: In-Memory Acceleration of Genome Analysis via Tight Integration of Basecalling and Read Mapping"**

Proceedings of the 55th International Symposium on Microarchitecture (**MICRO**), Chicago, IL, USA, October 2022.

[\[Slides \(pptx\) \(pdf\)\]](#)

[\[Longer Lecture Slides \(pptx\) \(pdf\)\]](#)

[\[Lecture Video\]](#) (25 minutes)

[\[arXiv version\]](#)

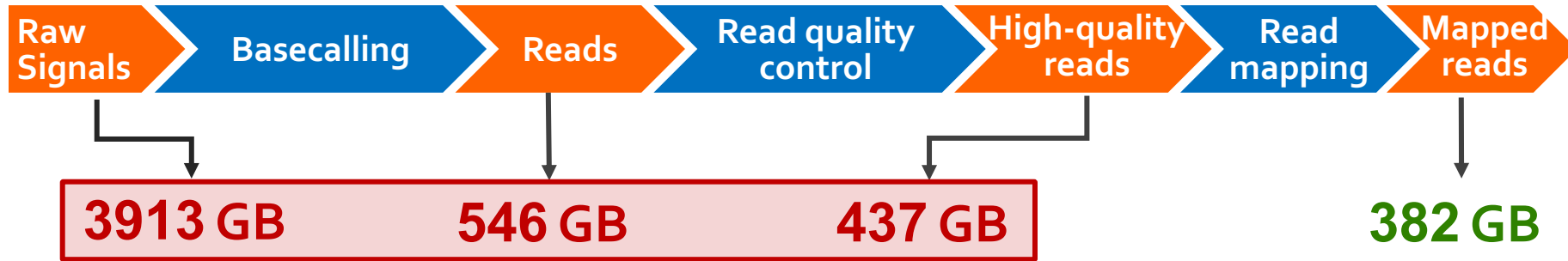
GenPIP: In-Memory Acceleration of Genome Analysis via Tight Integration of Basecalling and Read Mapping

Haiyu Mao¹ Mohammed Alser¹ Mohammad Sadrosadati¹ Can Firtina¹ Akanksha Baranwal¹
Damla Senol Cali² Aditya Manglik¹ Nour Almadhoun Alserr¹ Onur Mutlu¹

¹*ETH Zürich*

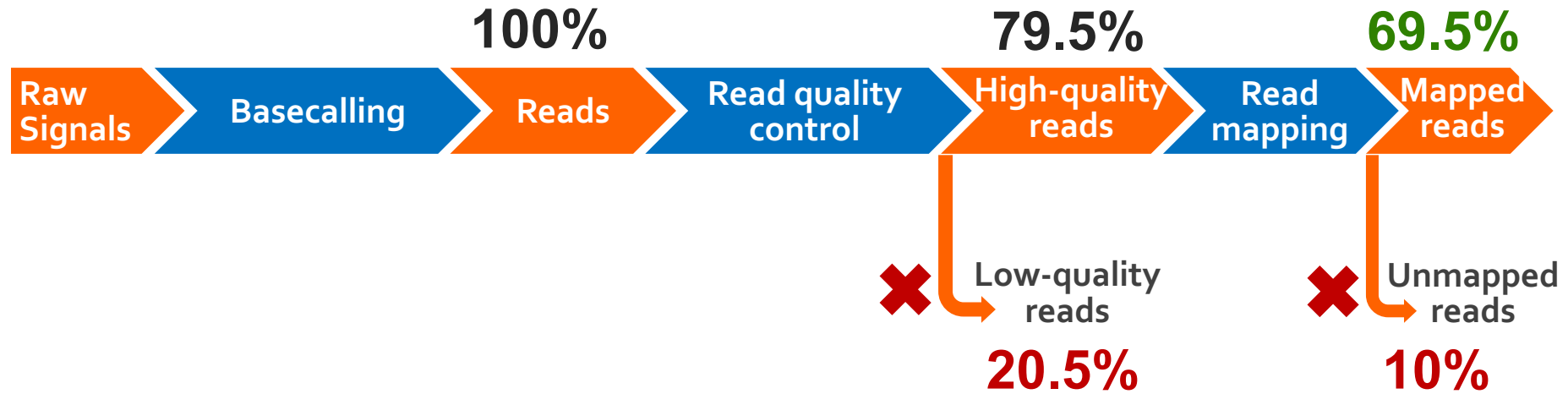
²*Bionano Genomics*

Problem 1: Large Data Movement



Large data movement between genome analysis steps

Problem 2: Wasted Computation



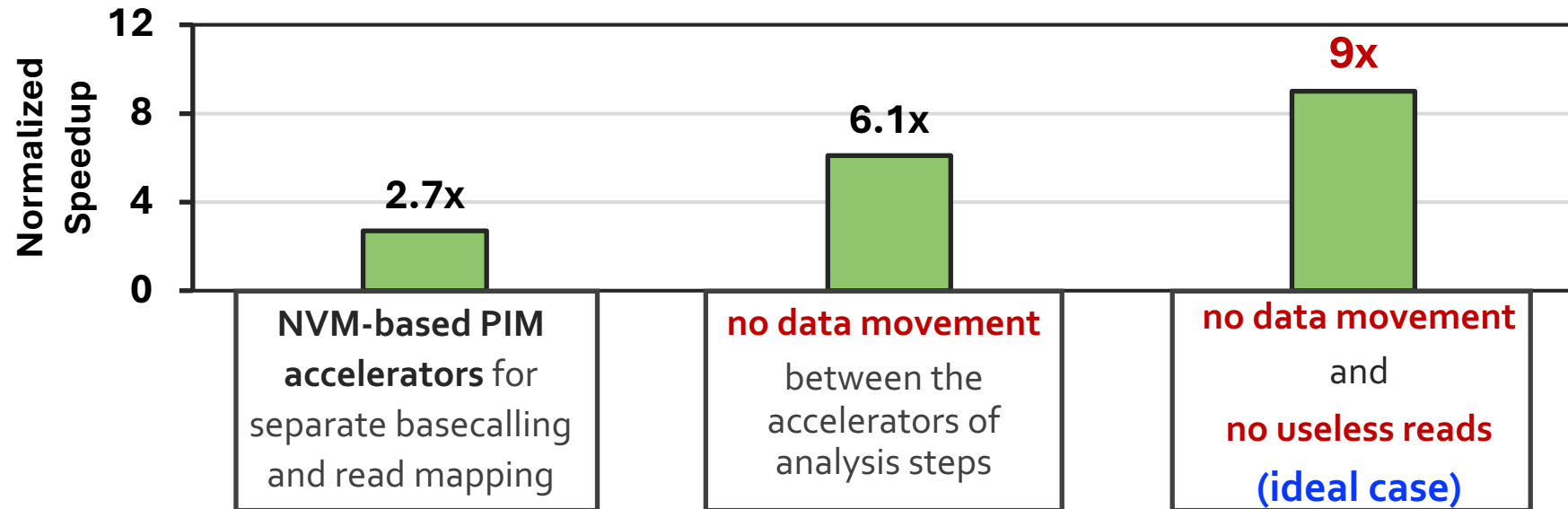
A considerable amount of computation on **useless data** due to

- Low-quality reads
- Unmapped reads

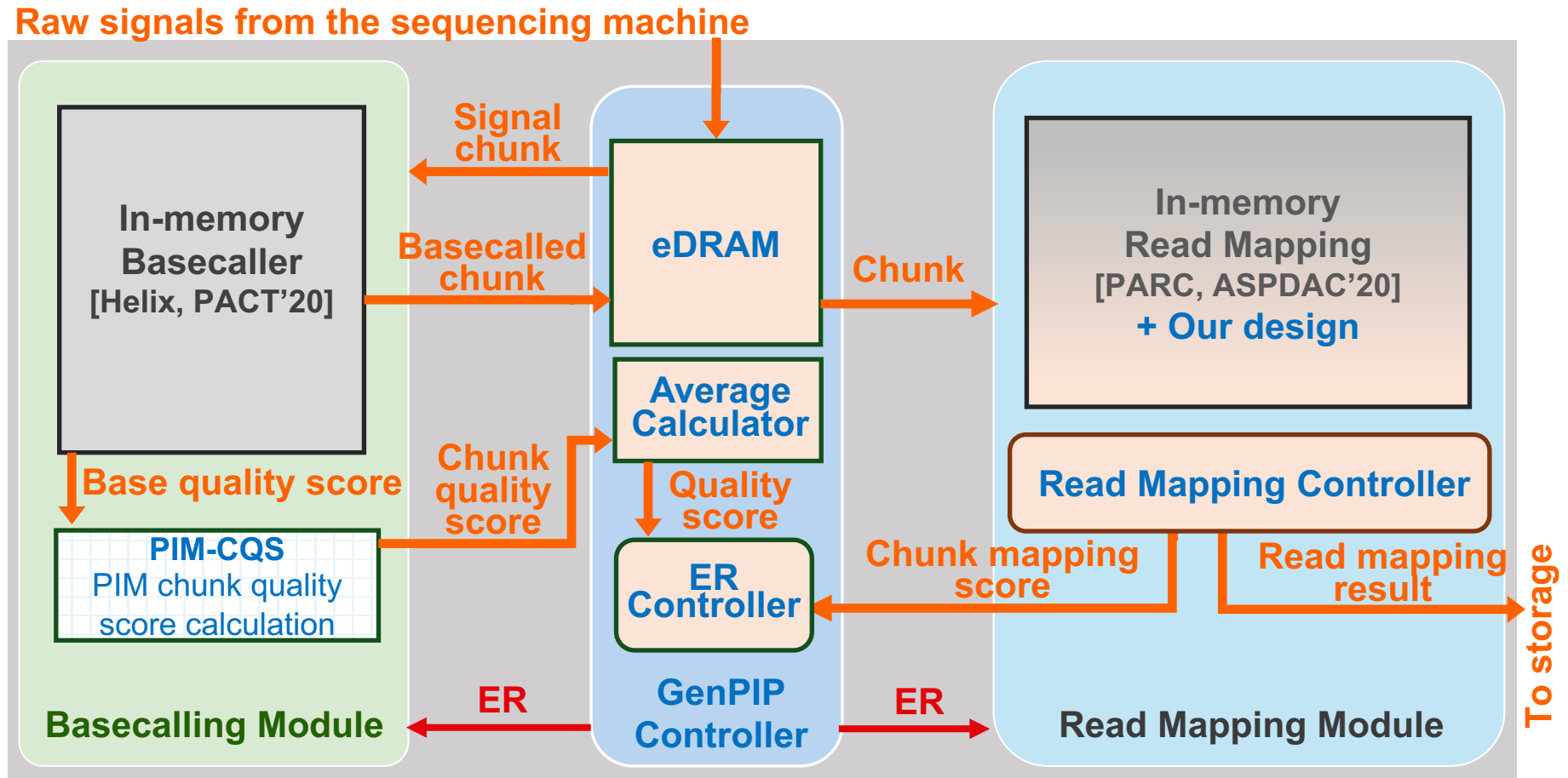
Goal and Opportunities

Goal: Efficiently accelerate the entire genome analysis pipeline while **minimizing data movement and useless computation**

We perform a study to quantify potential performance benefits



GenPIP Overview

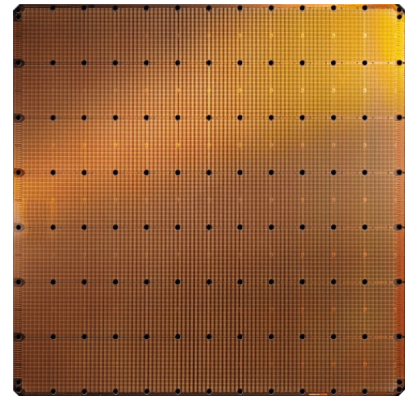


<https://arxiv.org/pdf/2209.08600.pdf>

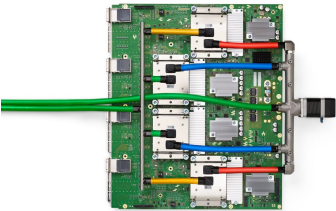
Acceleration Efforts for AI and Genomics

Cerebras WSE-3

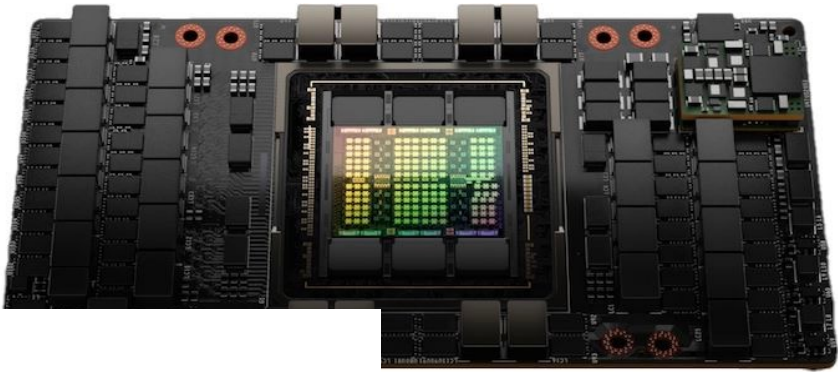
The largest ML accelerator chip



Google's TPUs:



NVIDIA H100



Scale your studies with ease

Process high-throughput data quickly with hardware acceleration

Process high-throughput data quickly with hardware acceleration. With four field-programmable gate arrays (FPGAs) onboard you have the most powerful DRAGEN analysis ever enabling you to process NovaSeq X System data easily. Perform up to four simultaneous applications per flow cell in a single run.

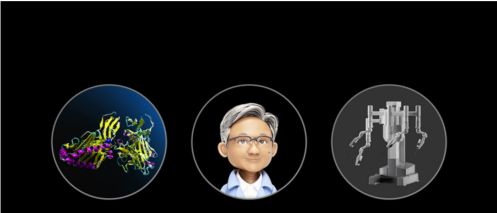
Reduce data footprint, manage and store data easily with lower costs and lower energy consumption, with built-in compression that reduces FASTQ file sizes by up to 80%.

Stream data directly to Illumina Connected Analytics or BaseSpace Sequence Hub on the cloud for scalable data management, analysis, and aggregation.

NVIDIA Partners With Industry Leaders to Advance Genomics, Drug Discovery and Healthcare

IQVIA, Illumina, Mayo Clinic and Arc Institute Harness NVIDIA AI and Accelerated Computing to Transform \$10 Trillion Healthcare and Life Sciences Industry

January 13, 2025



Absci and AMD Announce Collaboration and Strategic Investment to Accelerate the Future of AI Drug Discovery

01/08/2025

PDF Version

Quickly Aligning Raw Signals [IEEE Access '24]

- Joël Lindegger, Can Firtina, Nika Mansouri Ghiasi, Mohammad Sadrosadati, Mohammed Alser, and Onur Mutlu,

"RawAlign: Accurate, Fast, and Scalable Raw Nanopore Signal Mapping via Combining Seeding and Alignment"

[IEEE Access](#), December 2024.

Paper [\[online\]](#) [\[pdf\]](#) [\[code\]](#)

RawAlign: Accurate, Fast, and Scalable Raw Nanopore Signal Mapping via Combining Seeding and Alignment

**JOËL LINDEGGER^{ID}, CAN FIRTINA^{ID}, NIKA MANSOURI GHIASI, MOHAMMAD SADROSADATI^{ID},
MOHAMMED ALSER^{ID}, AND ONUR MUTLU^{ID}, (Fellow, IEEE)**

Department of Information Technology and Electrical Engineering, ETH Zürich, 8092 Zürich, Switzerland

Corresponding authors: Joël Lindegger (jmlindegger@gmail.com), Can Firtina (firtinac@ethz.ch), and Onur Mutlu (omutlu@ethz.ch)

This work was supported in part by the Semiconductor Research Corporation (SRC), in part by the ETH Future Computing Laboratory, in part by European Union's Horizon Programme for Research and Innovation under Grant 101047160-BioPIM, and in part by the Swiss National Science Foundation (SNSF) under Grant 200021_213084.

RawAlign Overview



Benchmarking Raw Signal Analysis [ACM-BCB '25]

- Furkan Eris, Ulysse McConnell, Can Firtina, and Onur Mutlu,
**"RawBench: A Comprehensive Benchmarking Framework
for Raw Nanopore Signal Analysis Techniques"**

Proceedings of the [16th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics \(ACM-BCB 2025\)](#), Philadelphia, PA, USA, October 2025.

Paper [\[online\]](#) [\[pdf\]](#)

Conference Talk [\[pptx\]](#) [\[pdf\]](#)

RawBench: A Comprehensive Benchmarking Framework for Raw Nanopore Signal Analysis Techniques

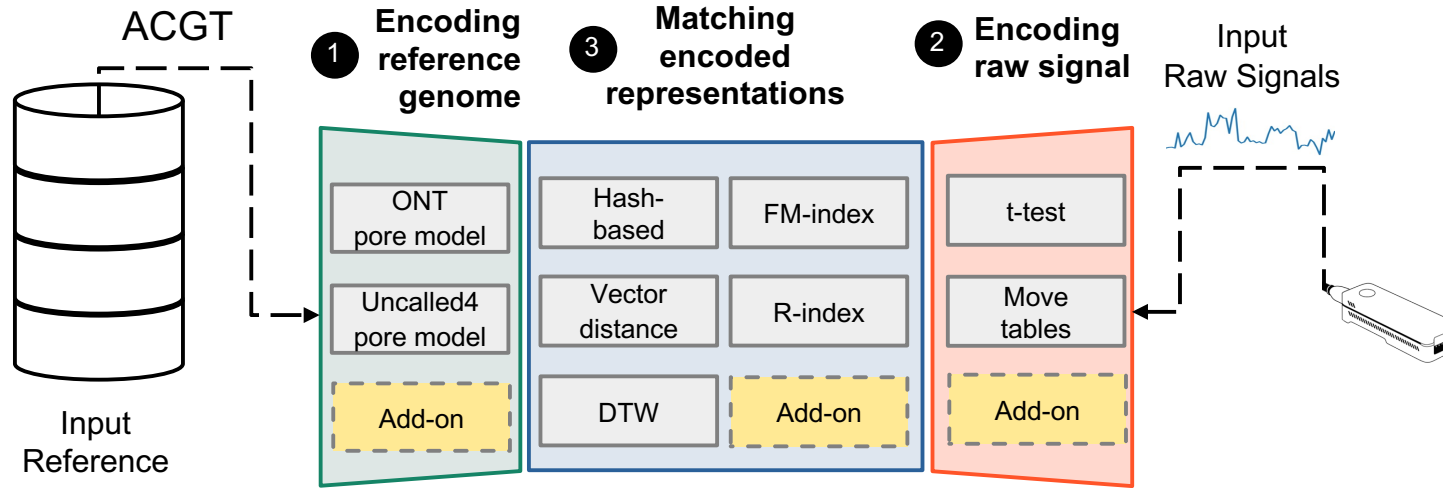
Furkan Eris
ETH Zurich
Zurich, Switzerland
furkan6olm@gmail.com

Can Firtina*
University of Maryland, College Park
MD, USA
firtina@umd.edu

Ulysse McConnell
ETH Zurich
Zurich, Switzerland
umcconnell@student.ethz.ch

Onur Mutlu*
ETH Zurich
Zurich, Switzerland
omutlu@gmail.com

RawBench Benchmarking Framework



1 Encoding Reference Genome

- Uses learned pore models by ONT and community

2 Encoding Raw Signal

- Utilizes statistical and ML-based encoding techniques

3 Matching Encoded Representations

- Compares representations that are now encoded into the same space

In-Storage Raw Signal Analysis [ICS '25]

- Melina Soysal, Konstantina Koliogeorgi, Can Firtina, Nika Mansouri Ghiasi, Rakesh Nadig, Haiyu Mao, Geraldo Francisco Oliveira Junior, Yu Liang, Klea Zambaku, Mohammad Sadrosadati, and Onur Mutlu,

"MARS: Processing-In-Memory Acceleration of Raw Signal Genome Analysis Inside the Storage Subsystem"

Proceedings of the [39th International Conference on Supercomputing \(ICS 2025\)](#), Salt Lake City, Utah, USA, June 2025.

Paper [\[online\]](#) [\[pdf\]](#)

Conference Talk [\[video\]](#) [\[pptx\]](#) [\[pdf\]](#)

MARS: Processing-In-Memory Acceleration of Raw Signal Genome Analysis Inside the Storage Subsystem

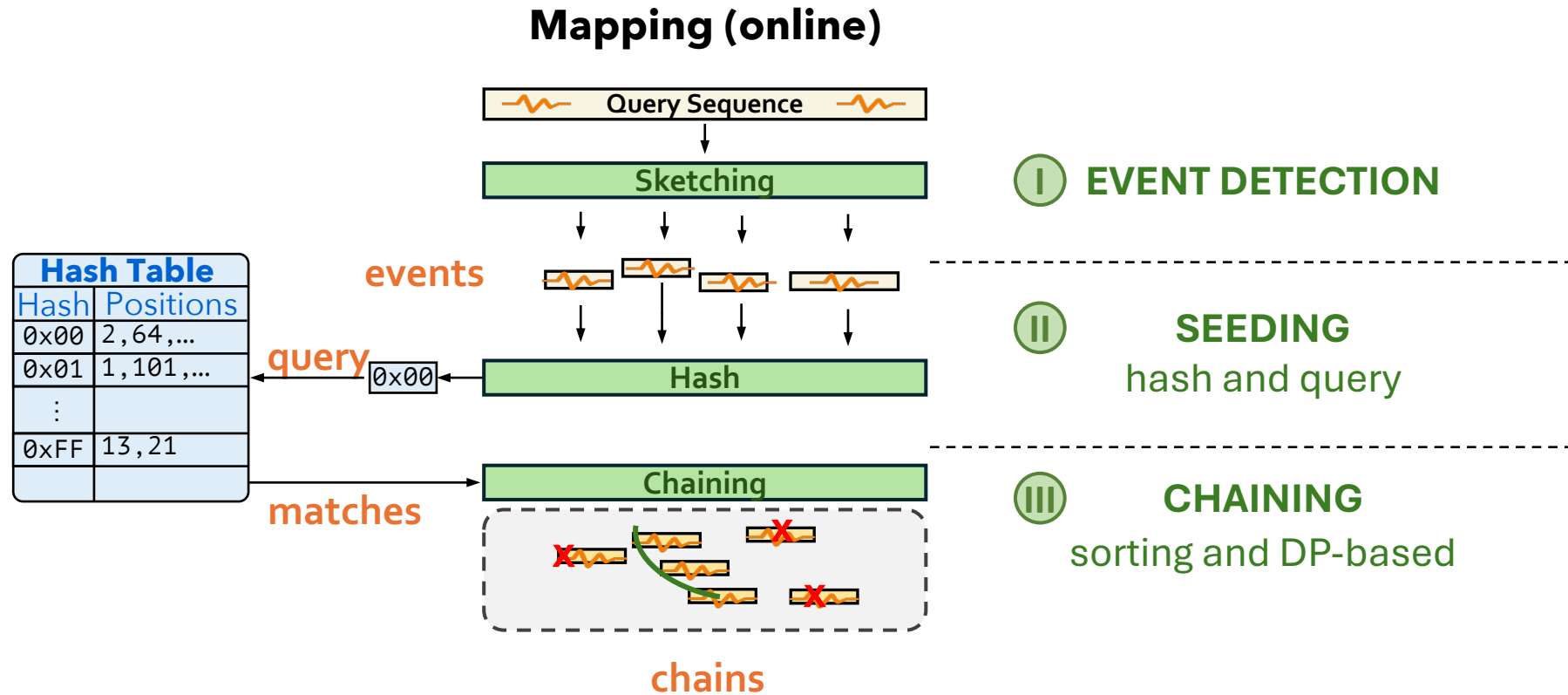
Melina Soysal[†] Konstantina Koliogeorgi[†] Can Firtina[†] Nika Mansouri Ghiasi[†]
Rakesh Nadig[†] Haiyu Mayo^{*} Geraldo F. Oliveira[†]
Yu Liang[†] Klea Zambaku[†] Mohammad Sadrosadati[†] Onur Mutlu[†]

[†] *ETH Zürich*

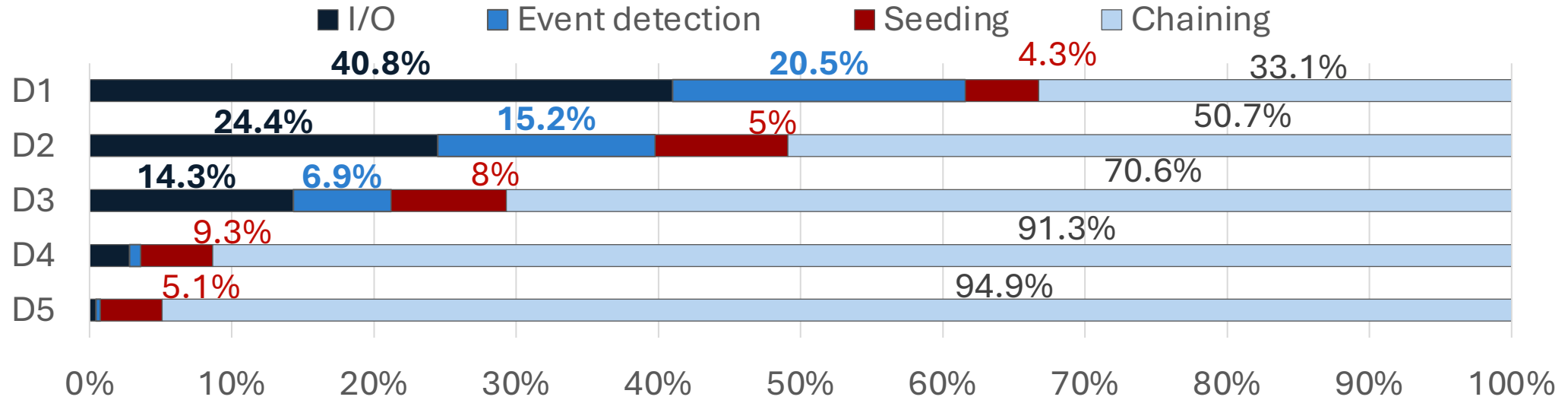
^{*} *King's College London*

Read Mapping with Raw Signals

High Level Overview of Read Mapping in Raw Signal Genome Analysis (RSGA)



Motivation

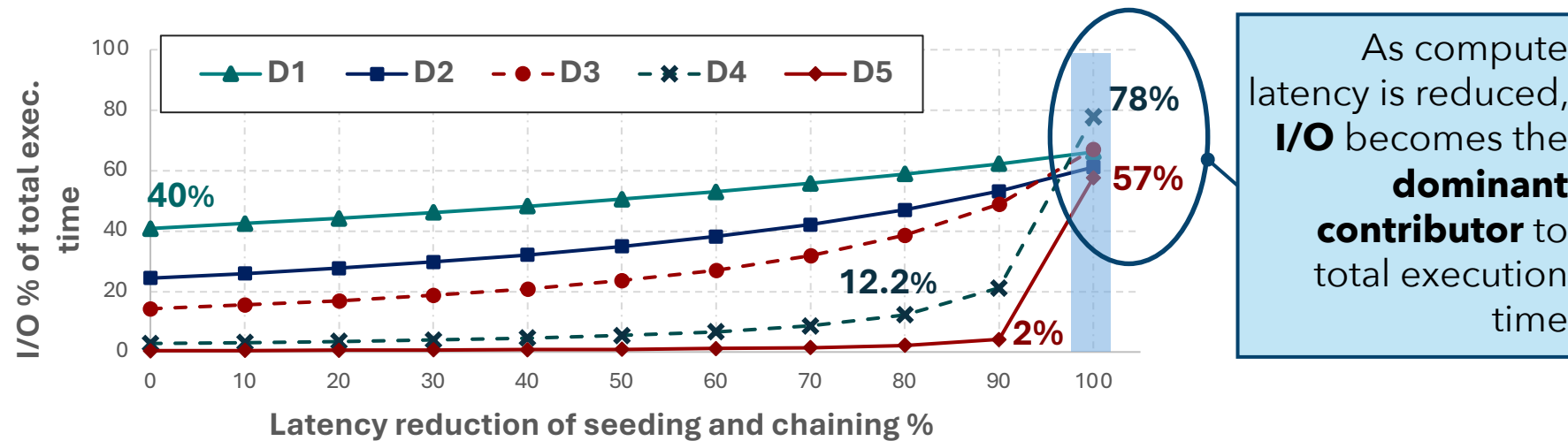


Chaining & **seeding** consistently dominate runtime

Event detection & **I/O overhead** are non-negligible for some datasets

Impact of I/O data movement

We model **I/O data movement overhead** when accelerating commonly targeted steps in the pipeline: **seeding** and **chaining**

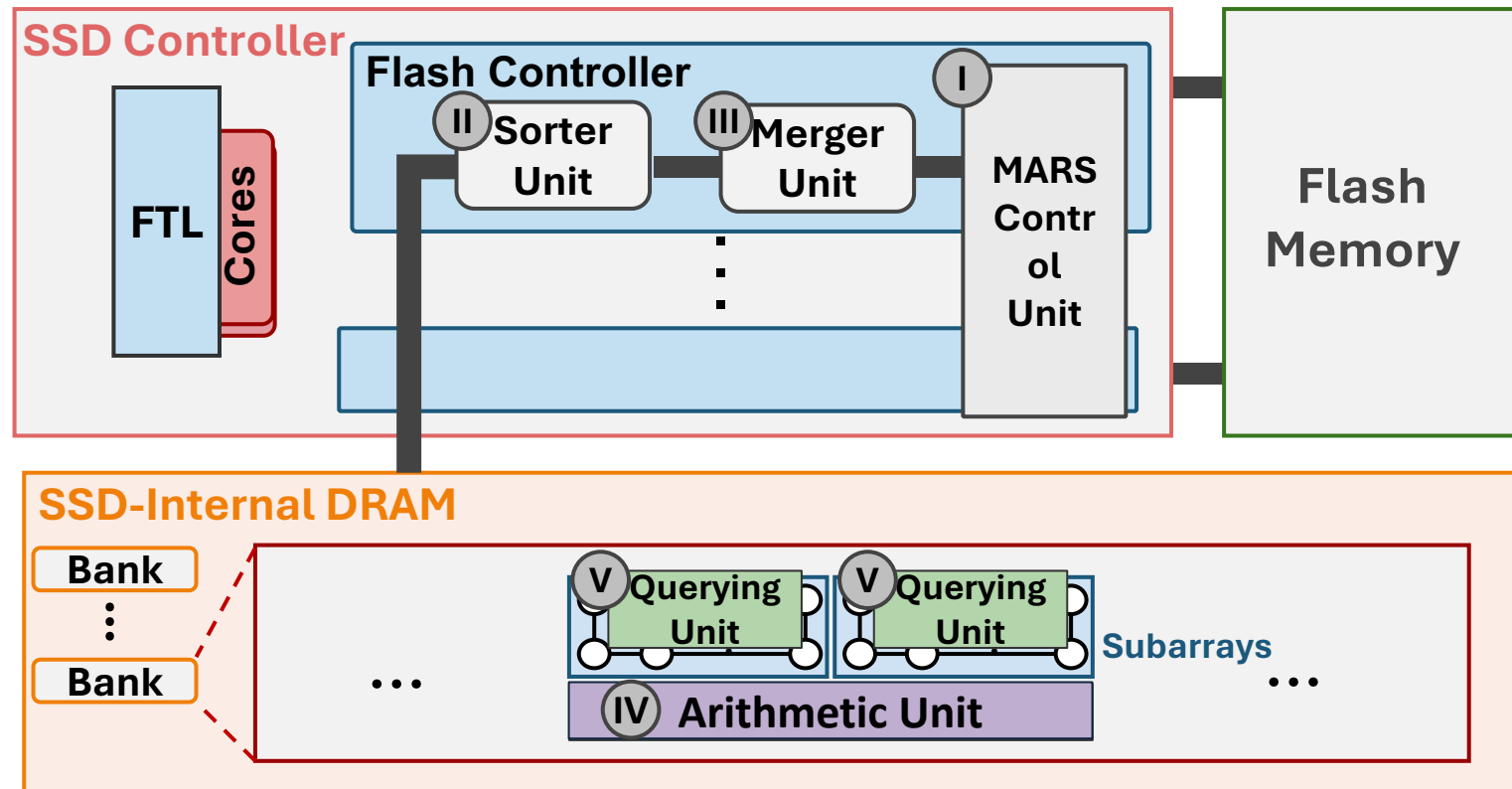


I/O data movement from SSD becomes dominant overhead as compute steps are accelerated in RSGA

MARS Architecture & System - Overview

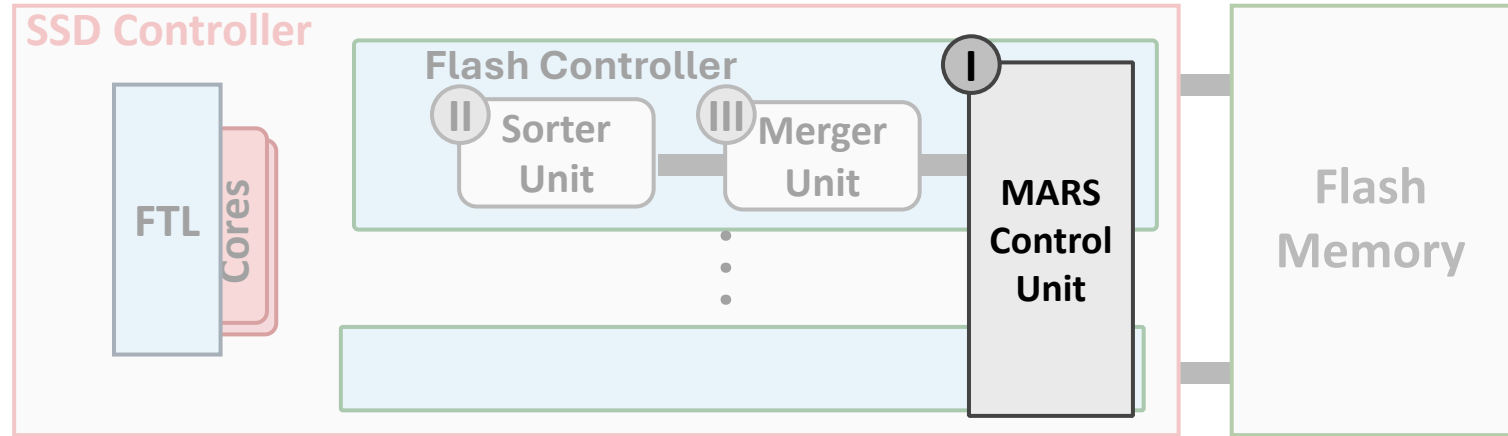
In-Storage-Processing system for *RSGA*

- **Conventional Mode:** SSD performs standard SSD storage operations
- **Accelerator Mode:** SSD executes the full RSGA pipeline without host intervention



MARS Architecture & System (I/IV)

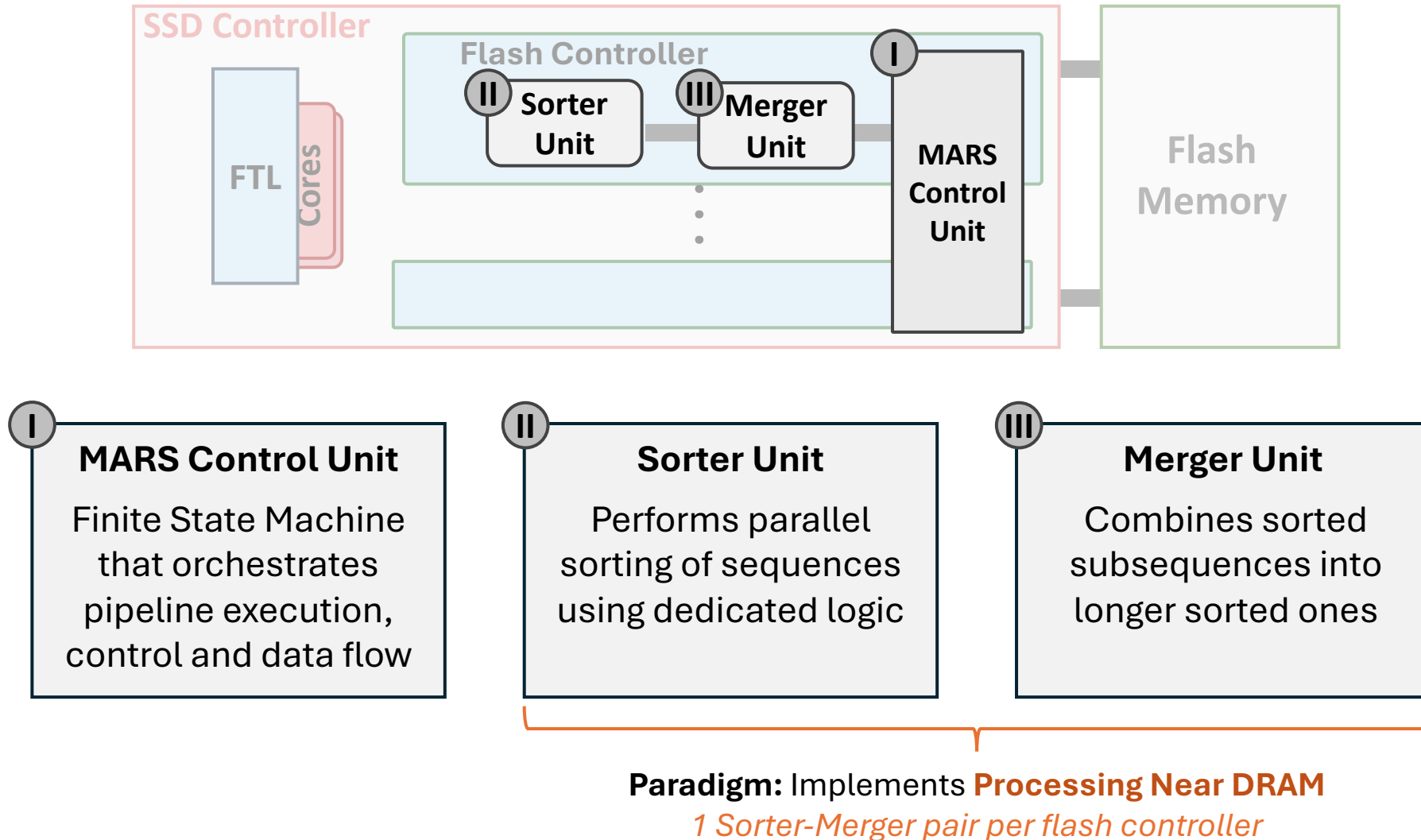
SSD Controller Components



I **MARS Control Unit**
Finite State Machine
that orchestrates
pipeline execution,
control and data flow

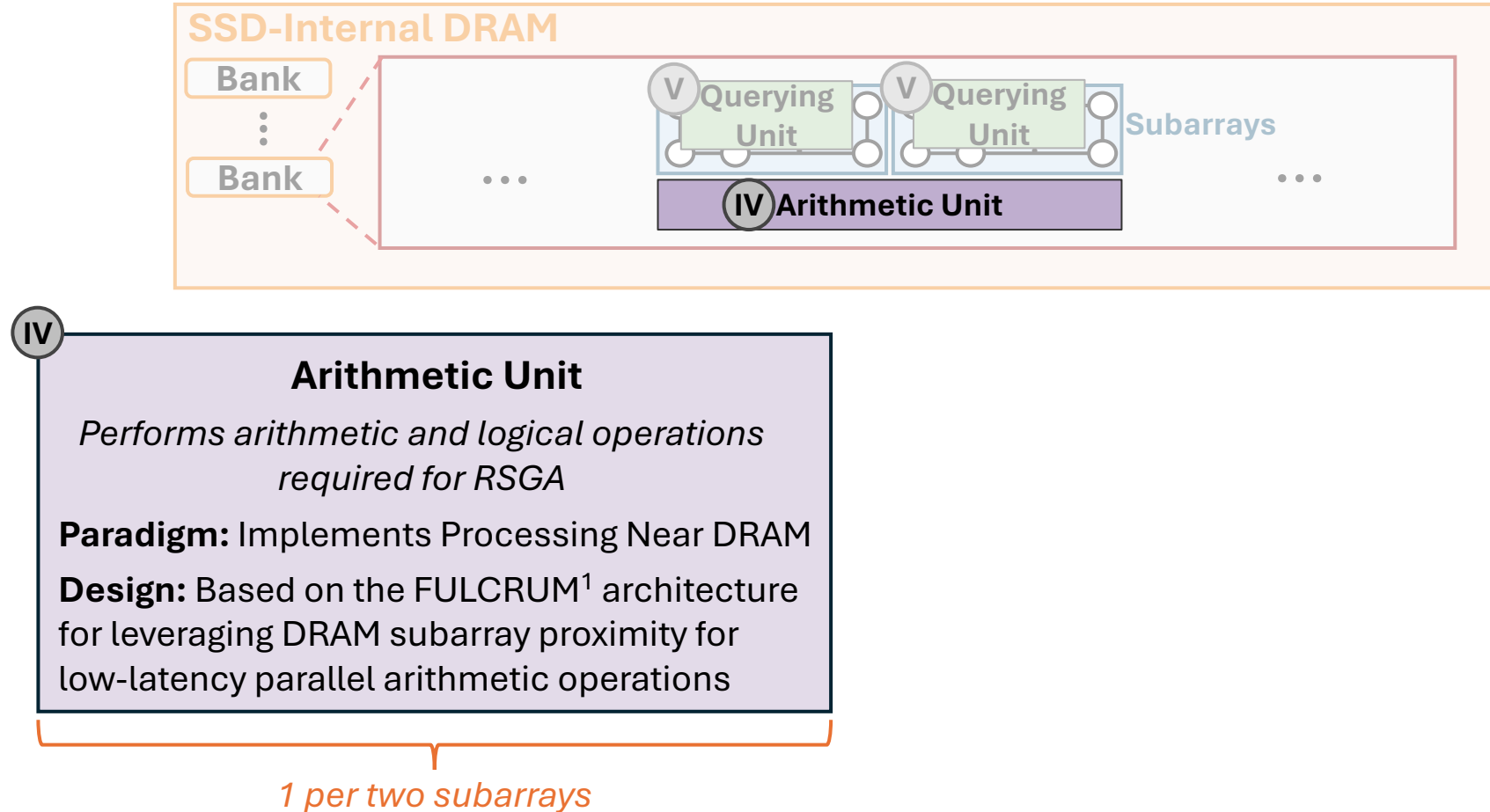
MARS Architecture & System (II/IV)

SSD Controller Components



MARS Architecture & System (III/IV)

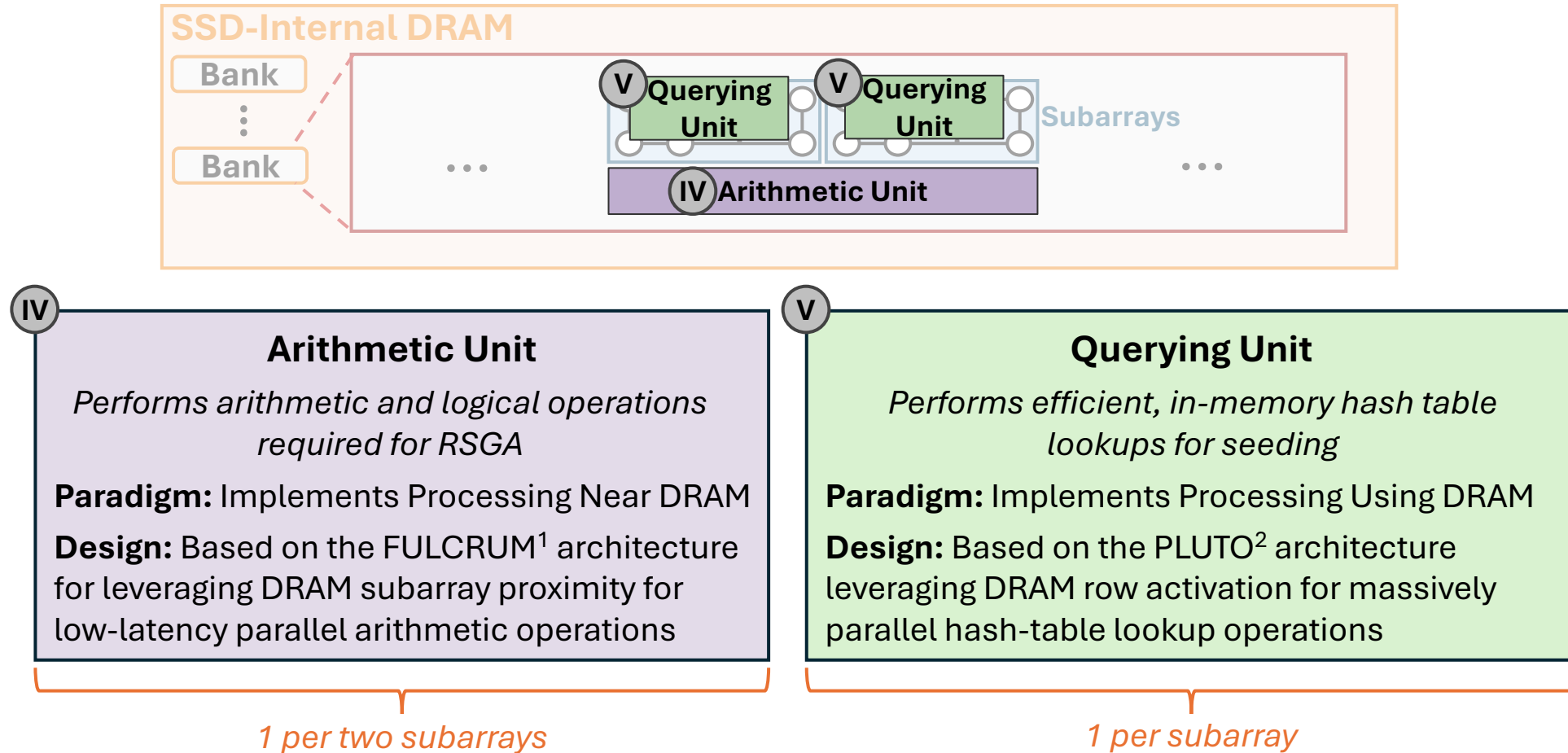
SSD-Internal DRAM Components



1: [Lenjani, HPCA'20];

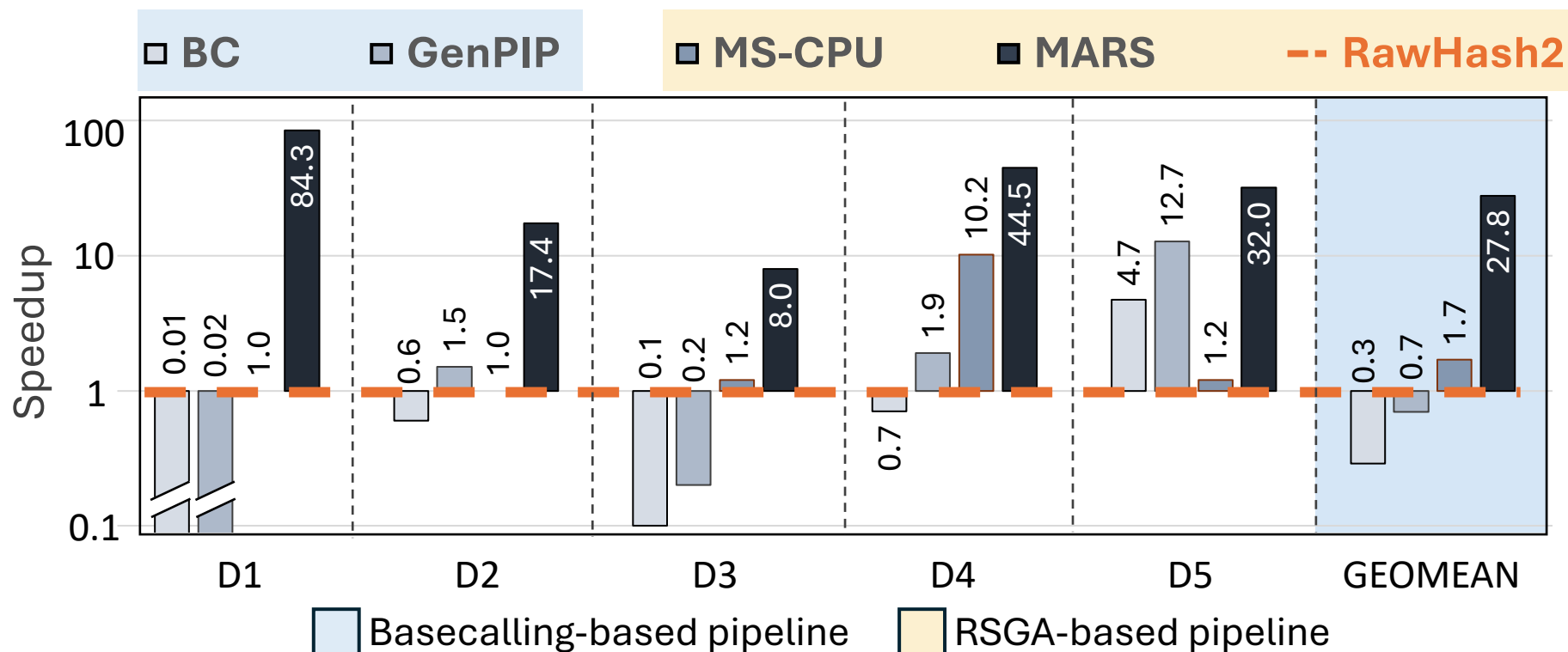
MARS Architecture & System (IV/IV)

SSD-Internal DRAM Components



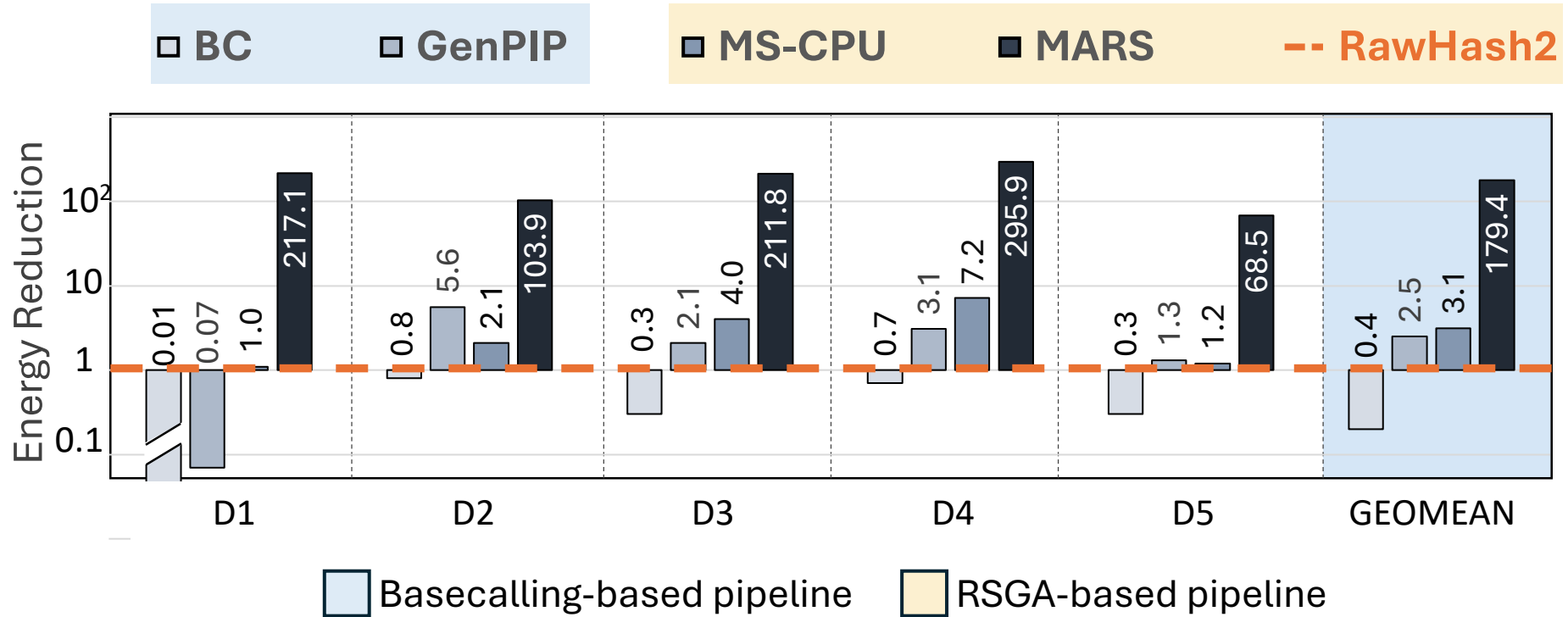
1: [Lenjani, HPCA'20]; 2: [Ferreira+, MICRO'22]

Performance Evaluation



- Key Observations
- MARS outperforms all other baselines across all datasets
 - MARS outperforms GenPIP adopting PIM outside storage
We fundamentally solve the I/O data movement overhead problem
 - MARS-CPU provide speedup of 1.7x over RH2
Our SW modifications are effective in reducing the computational load

Energy Evaluation



Key Observation

- MARS reduces energy consumption over all other baselines across all datasets
 - Improvement stems from*
 - (1) *reduced execution time*
 - (2) *fundamentally addressing the I/O data movement overhead*

Processing Data Where It Makes Sense
Provides Significant Performance
and Energy Benefits

Executive Summary

Problem:

Noise and signal processing errors reduce the accuracy of raw signal analyses

Goal:

Detect and correct errors in raw signal data

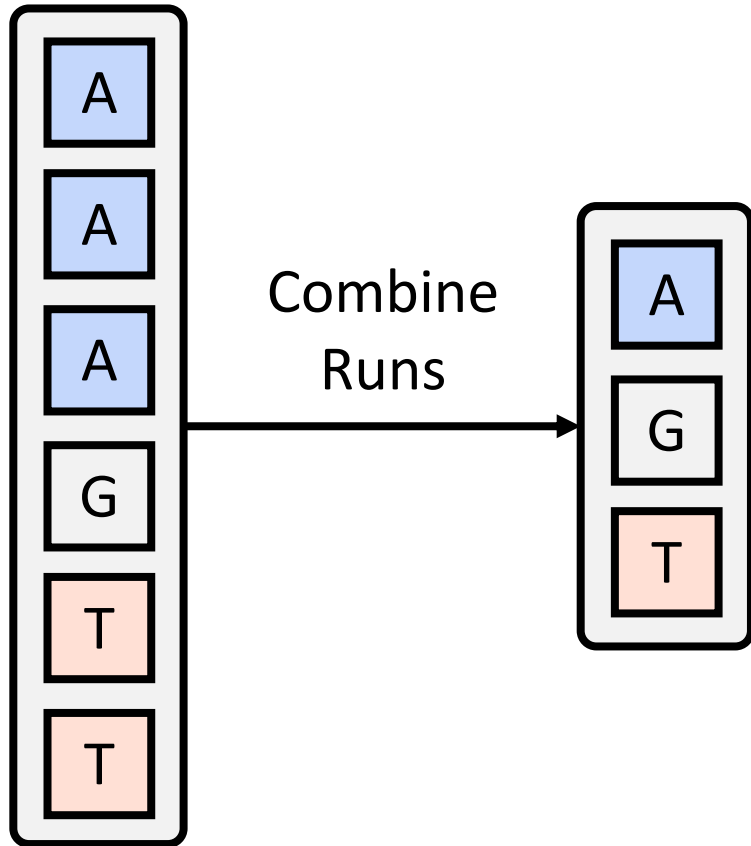
Key Contributions:

- 1) CRANE - the first sophisticated mechanism for removing errors in raw signals
- 2) Framework which can optimize CRANE for various signal analysis pipelines

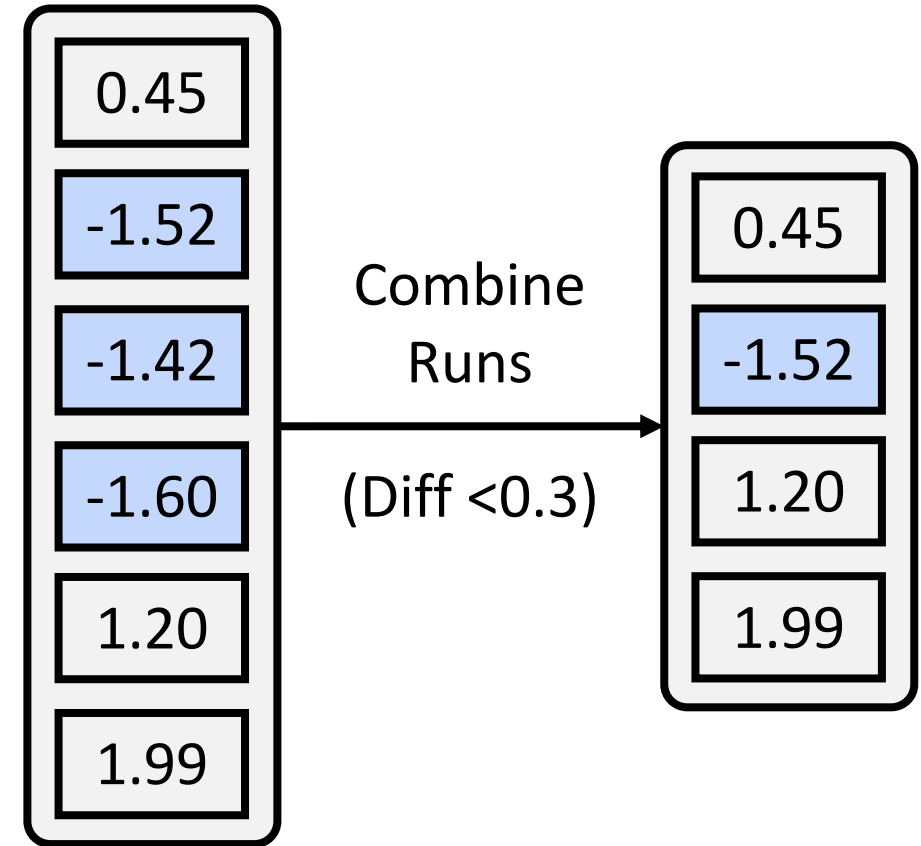
Key Results: CRANE consistently increases SOTA read mapping F1 score by up to **11%** while accounting for less than **1%** of total runtime for large genomes

Existing Error Correction Methods Are Heuristic

DNA Sequence
Homopolymer Compression



Nanopore Events
Homopolymer Compression



Existing Error Correction Methods Are Heuristic

DNA Sequence

Homopolymer Compression

Nanopore Events

Homopolymer Compression

How can we quickly and accurately correct errors
in raw nanopore events?

T

T

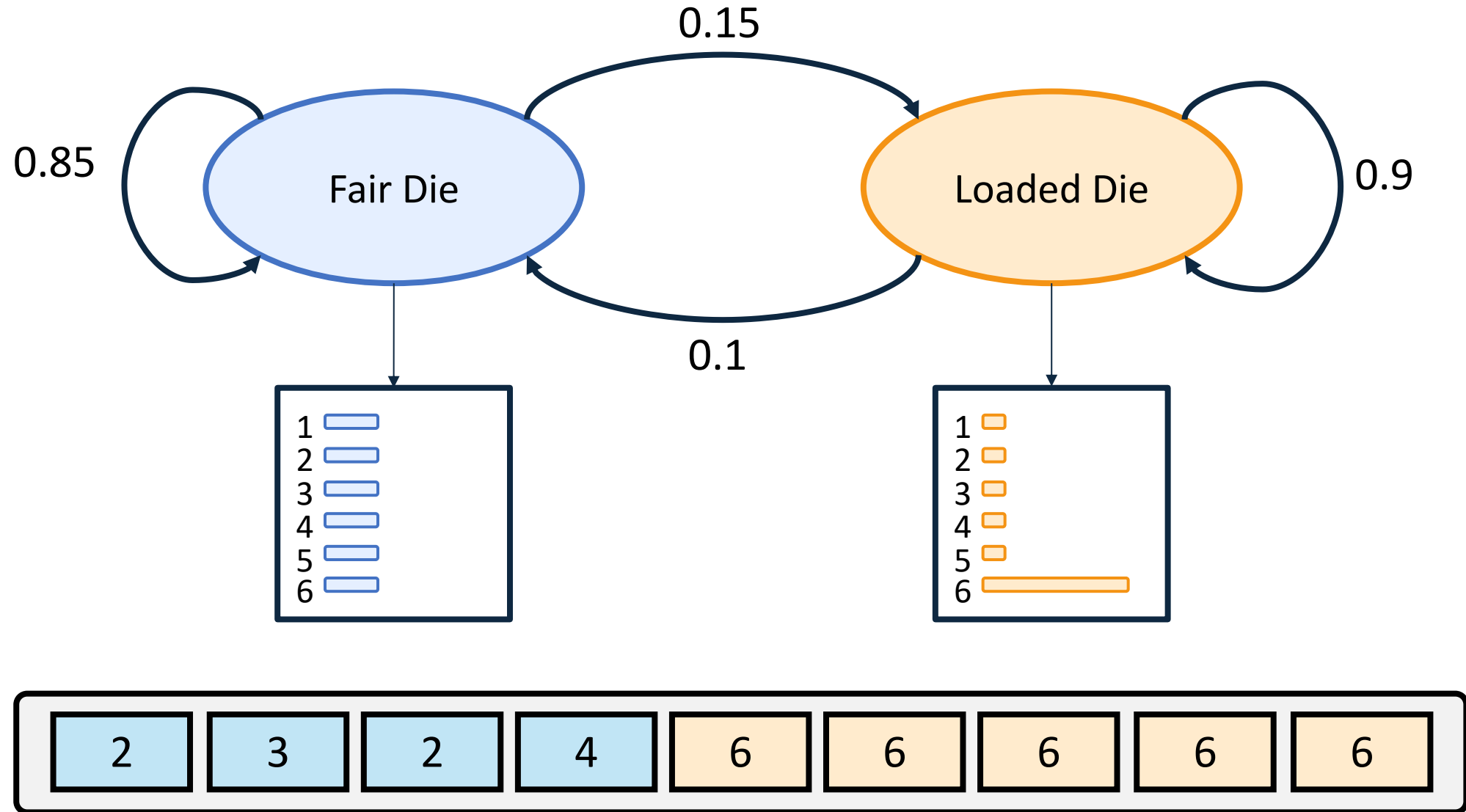
T

1.20

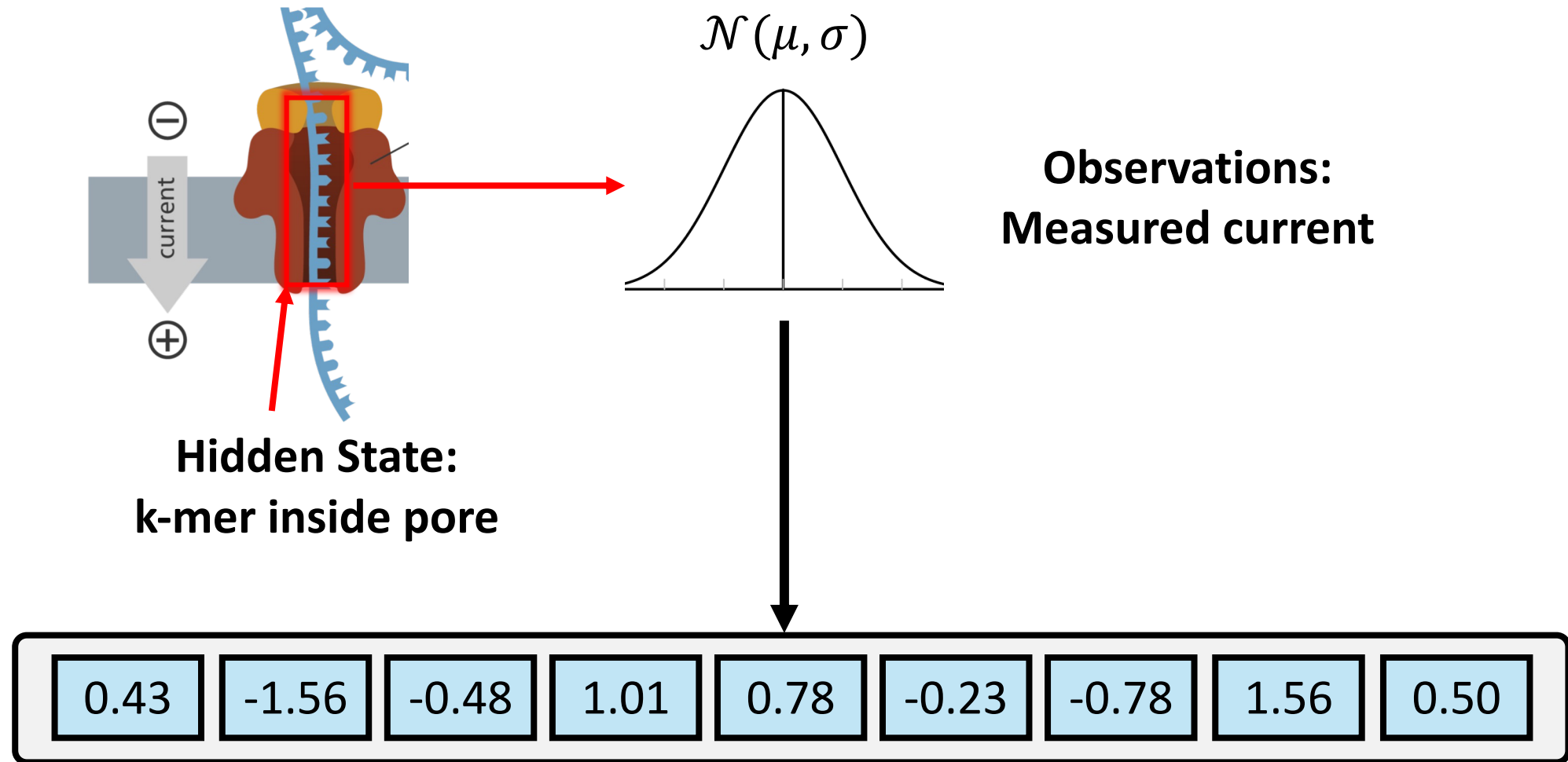
1.99

1.99

HMMs Model Unobservable Processes



HMMs Can Model Nanopores



Outline

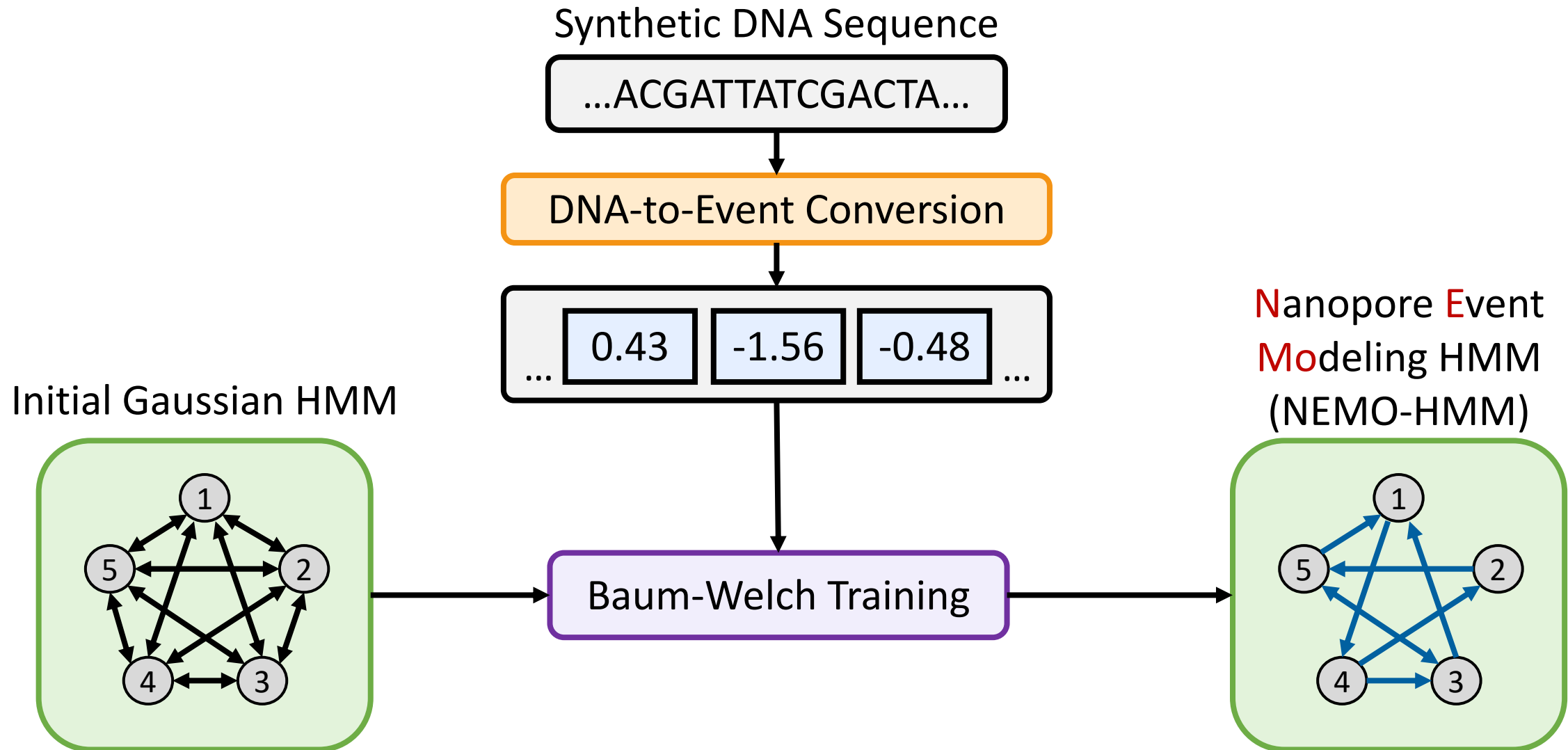
Background

CRANE

Results

Conclusion

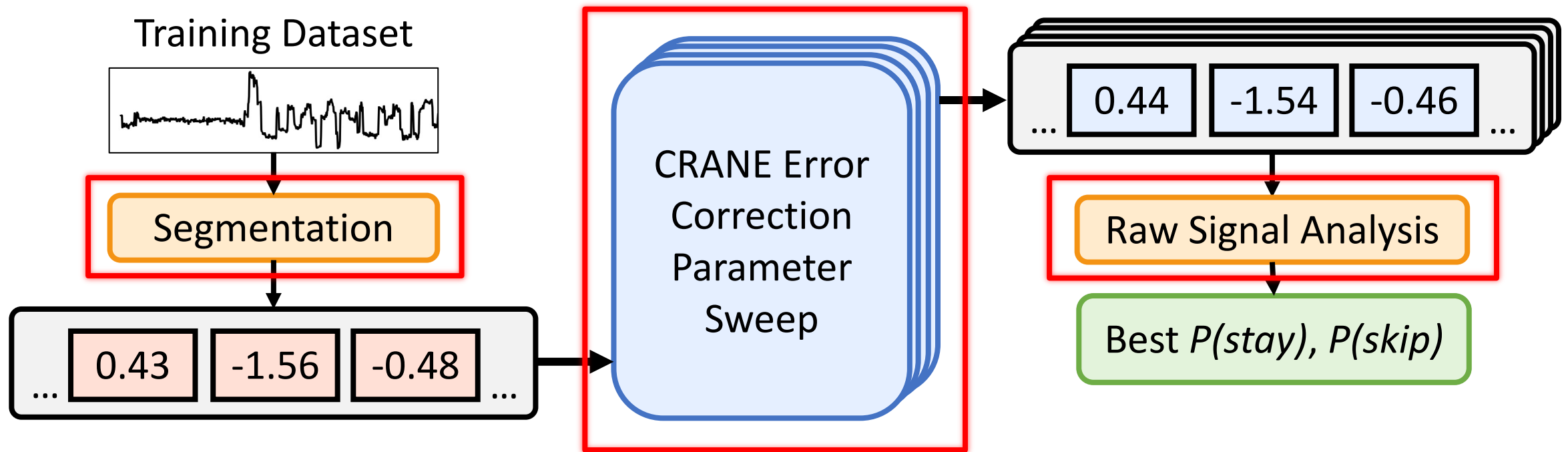
Key Idea: Train An HMM To Model A Nanopore



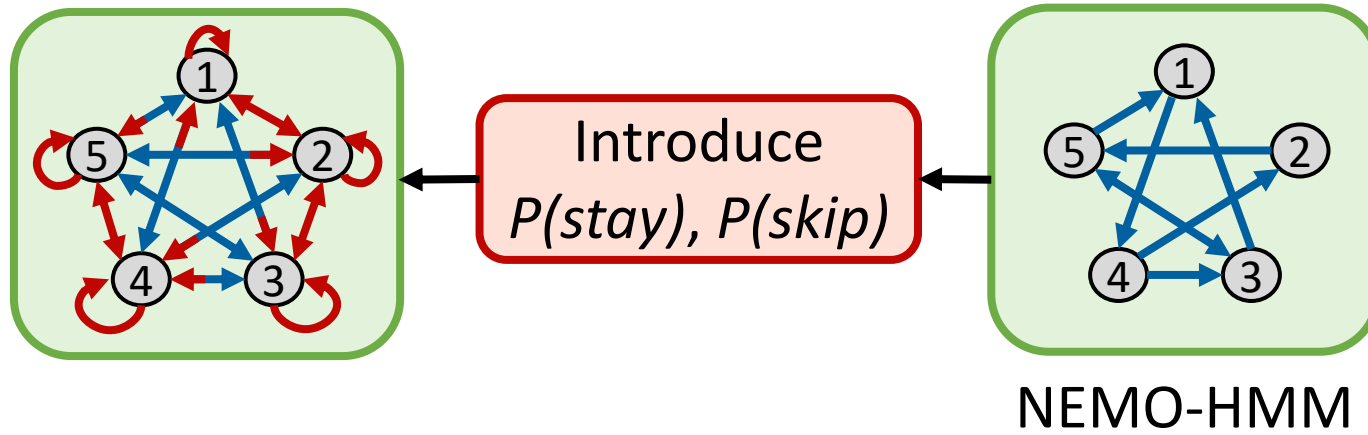
CRANE Can Be Tuned For Any Raw Signal Pipeline

A parameter search is performed
over:

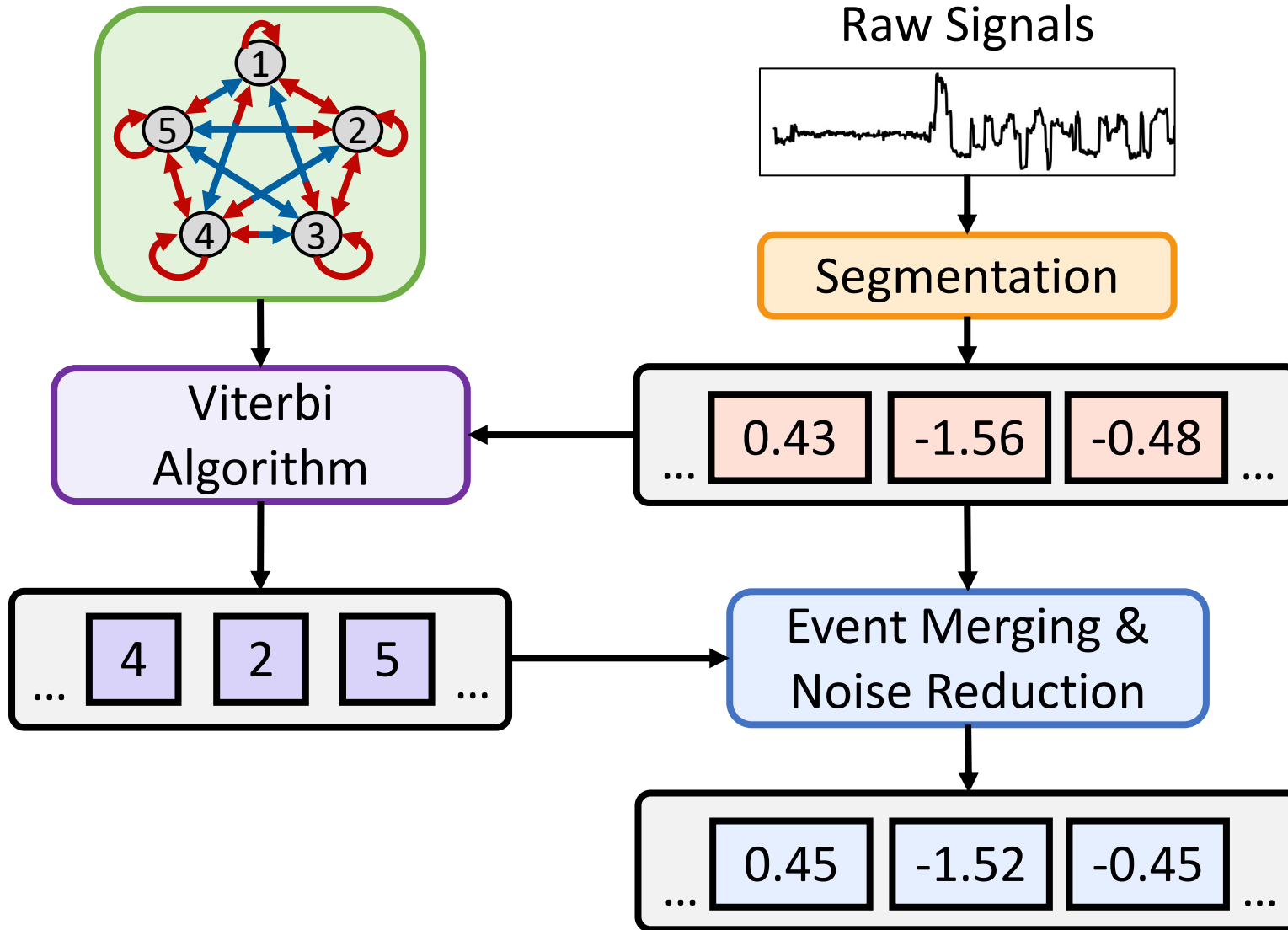
- $P(stay)$
- $P(skip)$



CRANE Error Correction

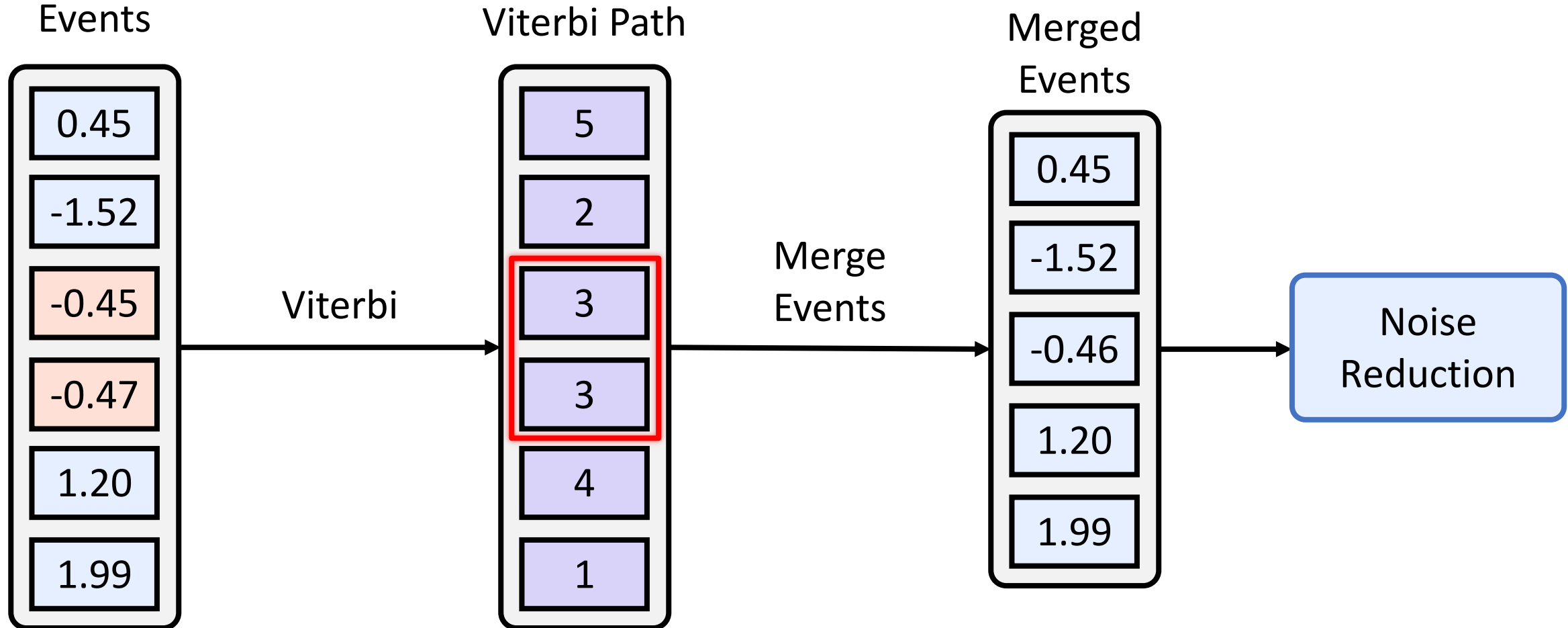


CRANE Error Correction



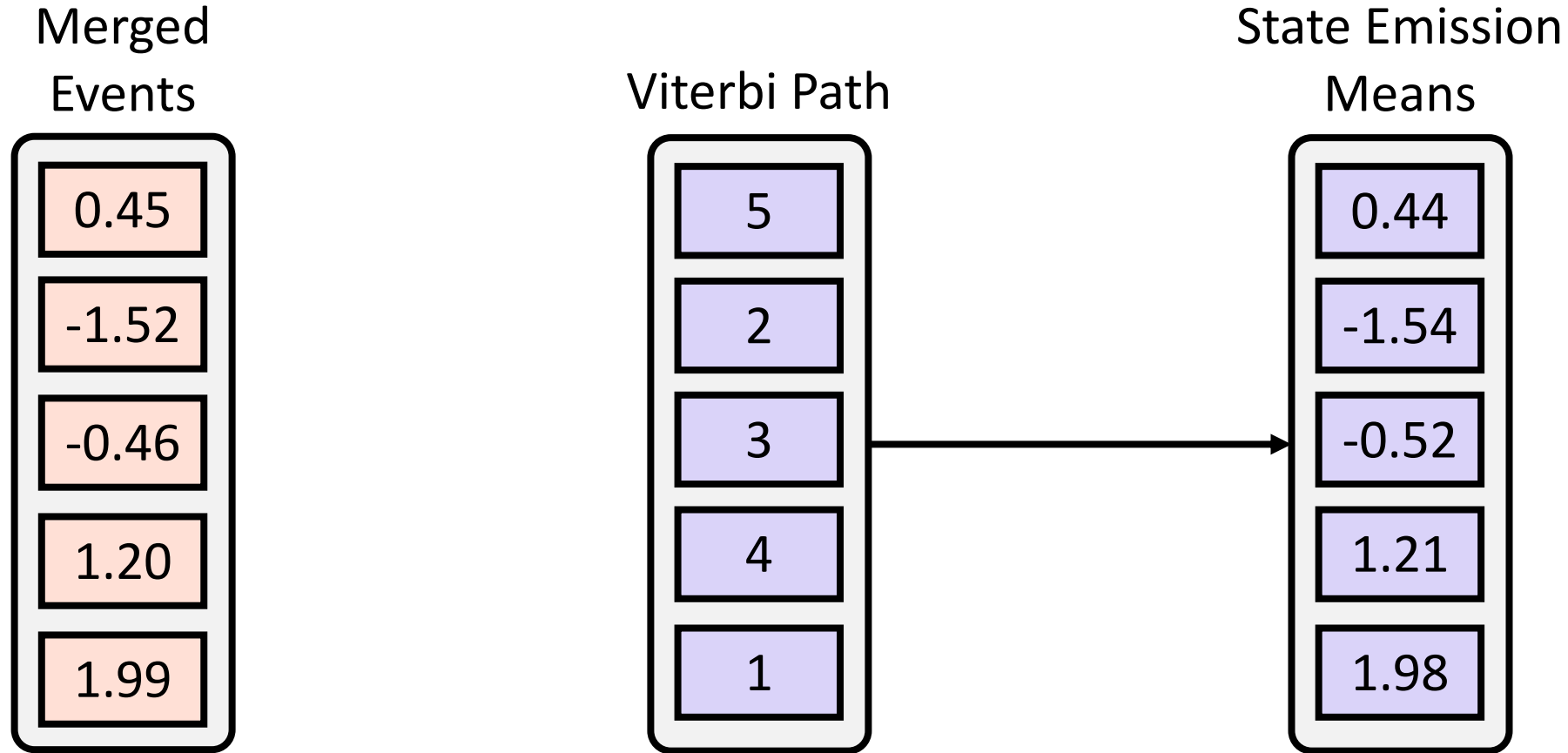
CRANE Merges Duplicate Events

- Segmentation algorithms may **oversegment** the signal



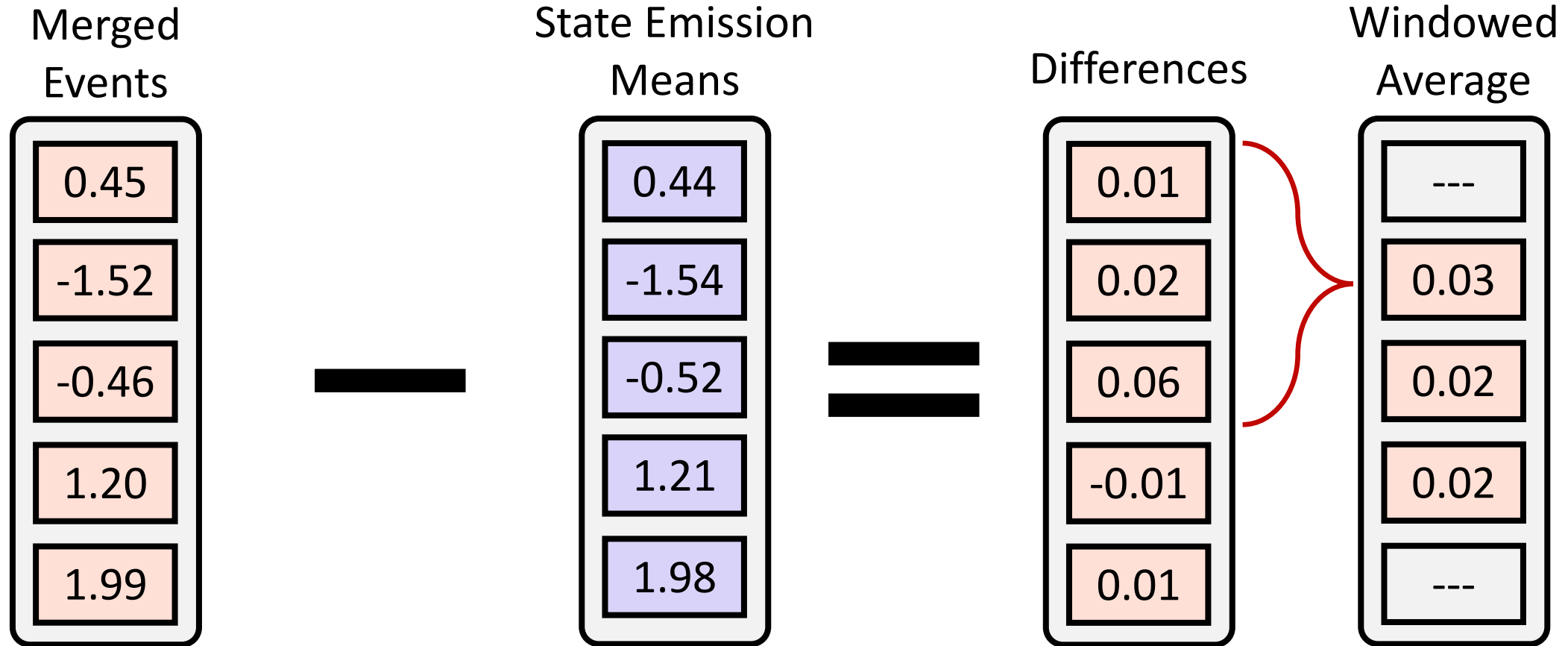
CRANE Reduces Noise In The Raw Signal

- CRANE can detect background noise which **spans multiple events**



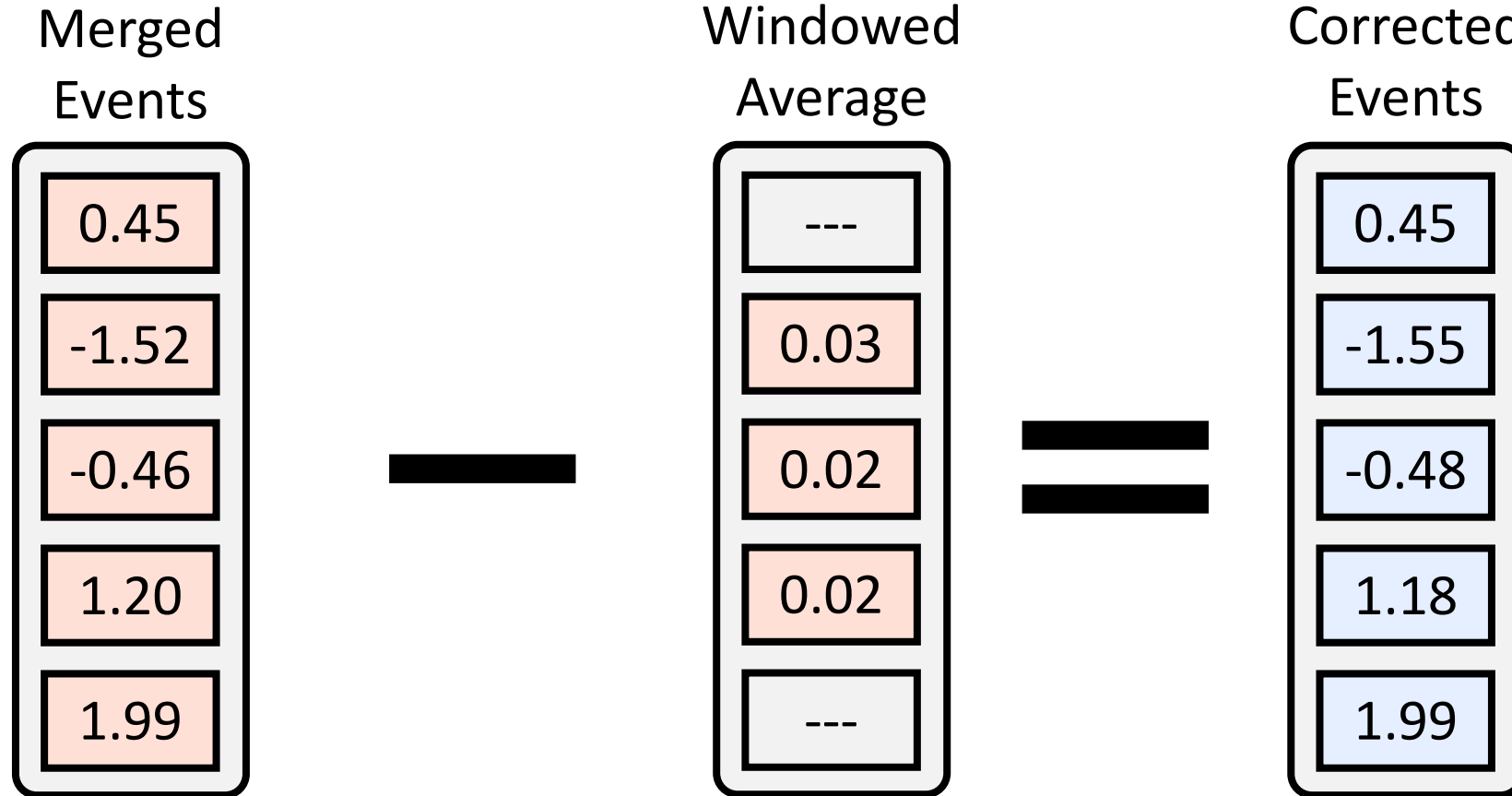
CRANE Reduces Noise In The Raw Signal

- CRANE can detect background noise which **spans multiple events**



CRANE Reduces Noise In The Raw Signal

- CRANE can detect background noise which **spans multiple events**



Outline

Background

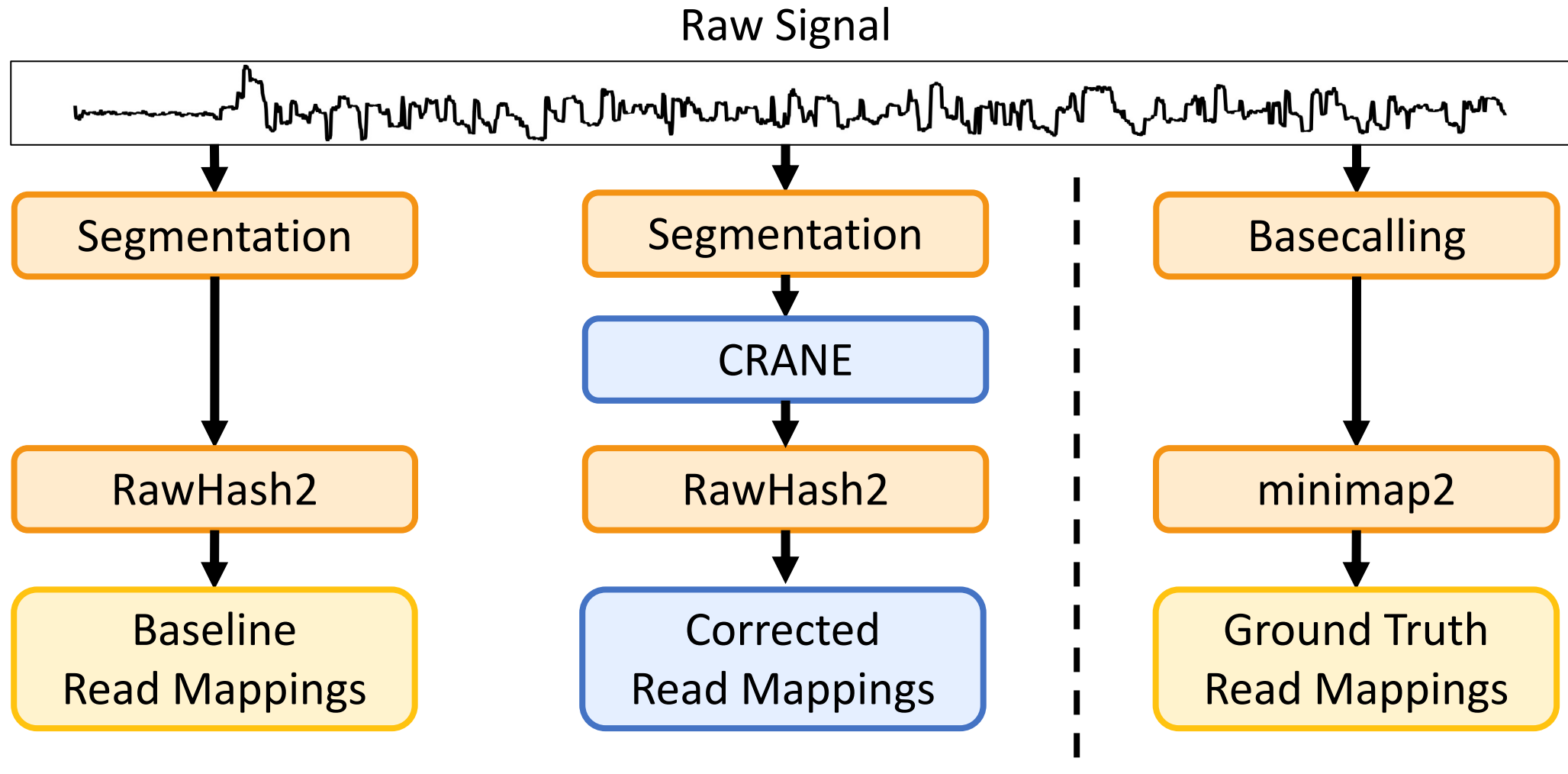
CRANE

Results

Conclusion

We Evaluate Our Approach Using RawHash2

- RawHash2 is a **raw signal read-mapping** tool



Datasets

- 4 datasets
- 12,000 *E. coli* reads for parameter searches
- 10,000 *E. coli*, Fruit Fly, and Human reads for evaluation

	Organism	Chemistry	Flow Cell	Reads	Avg. Length
D0	<i>E. coli</i> CFT073	R10.4.1 e8.2	FLO-MIN114	12,000	794
D1	<i>E. coli</i> CFT073	R10.4.1 e8.2	FLO-MIN114	10,000	764
D2	<i>D. melanogaster</i>	R10.4.1 e8.2	FLO-MIN114	10,000	5,998
D3	<i>H. sapiens</i> (HG002)	R10.4.1 e8.2	FLO-PRO114M	10,000	16,943

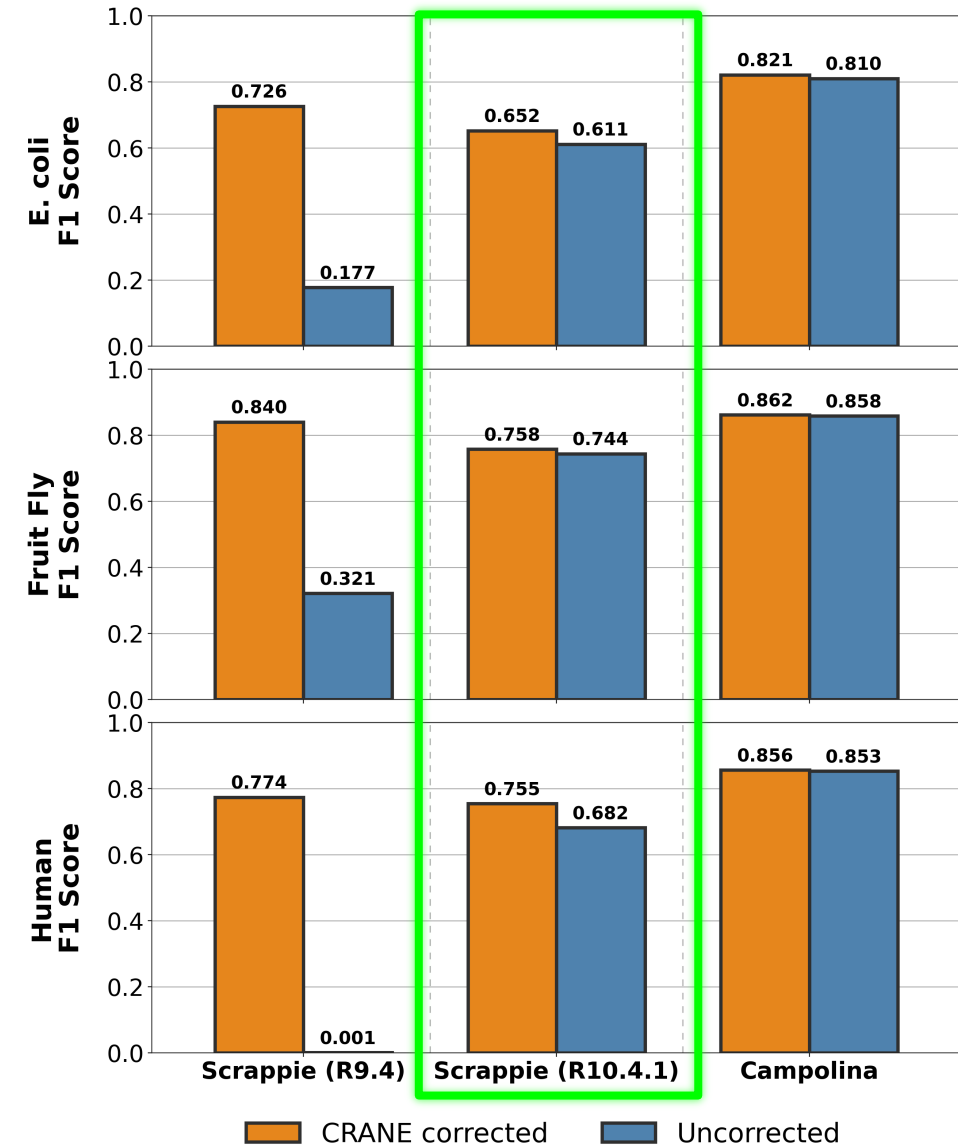
Evaluation Methodology

We evaluate CRANE's impact on **RawHash2's** read-mapping F1 score:

- Across three segmentation algorithms
 - Scrappie R9
 - Scrappie R10
 - Campolina (ML-based)
- With and without an existing error correction mechanism
 - Homopolymer Compression (HPC)

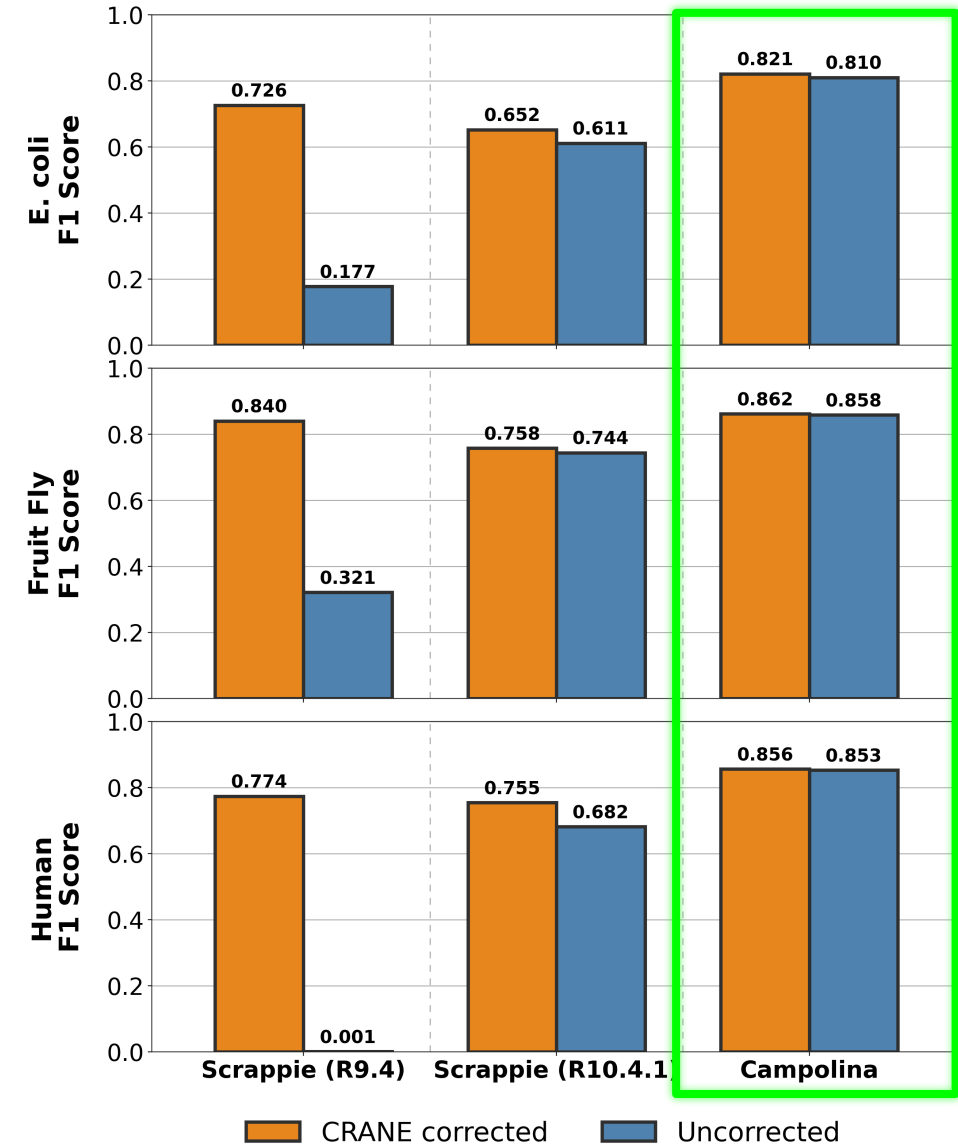
Accuracy Results - RawHash2 Read-mapping w/o HPC

CRANE improves state of the art
read mapping F1 score by
up to 11%



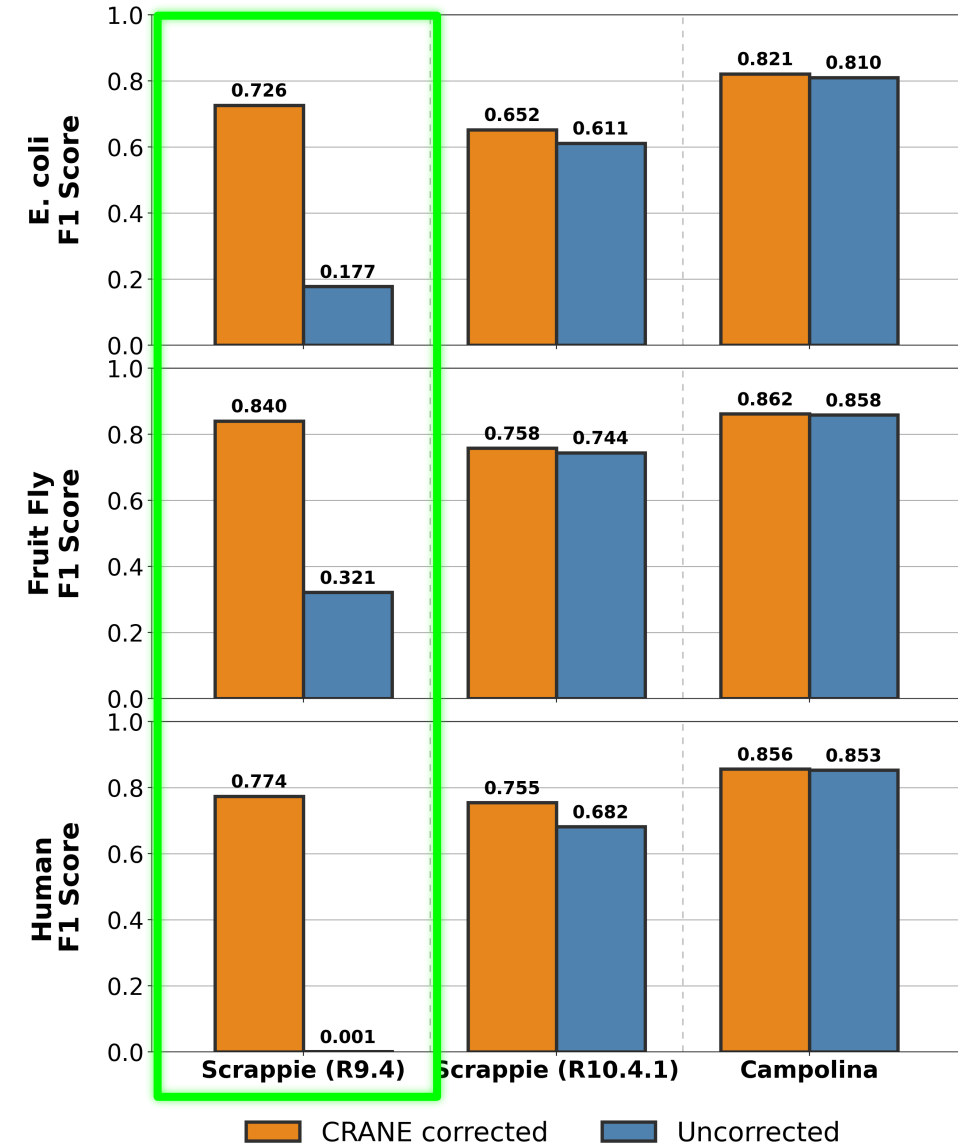
Accuracy Results - RawHash2 Read-mapping w/o HPC

CRANE improves state of the art
read mapping F1 score by
up to 11%



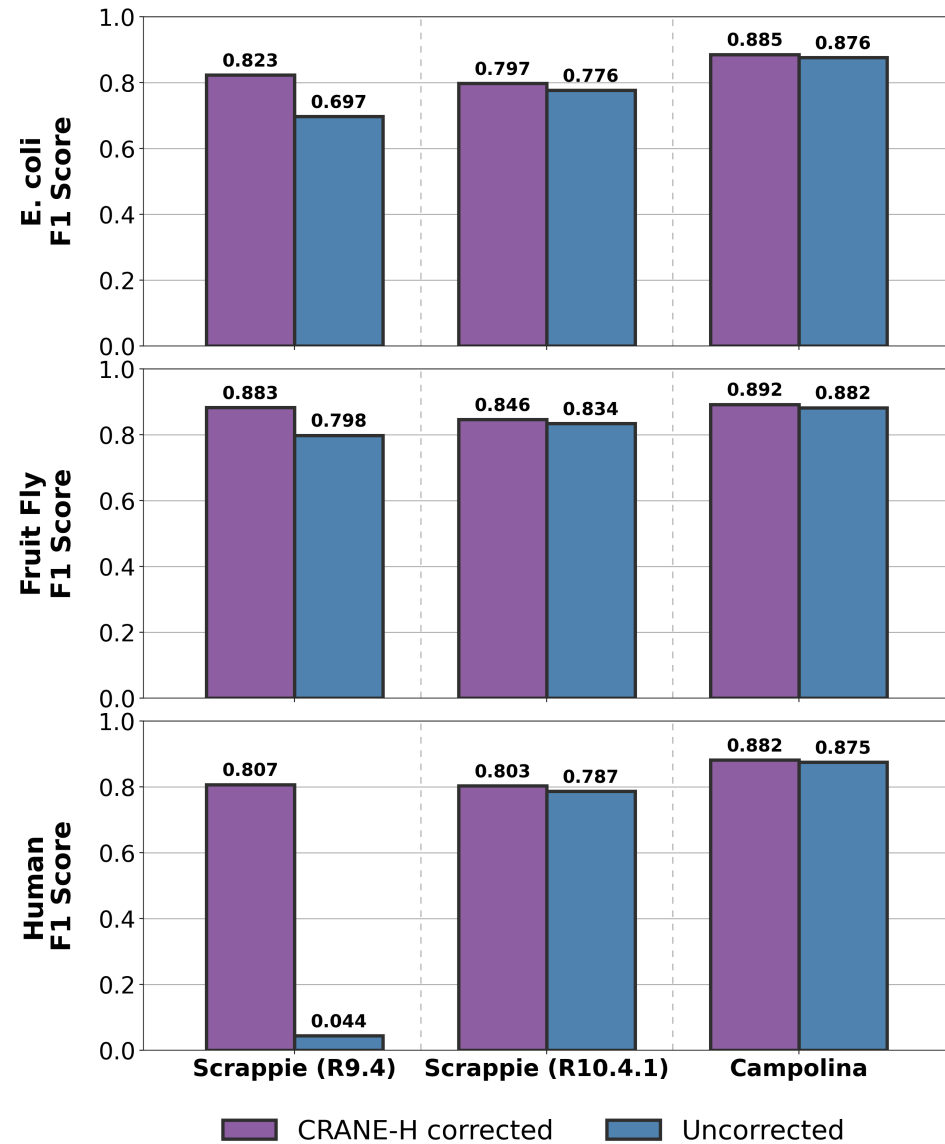
Accuracy Results - RawHash2 Read-mapping w/o HPC

CRANE improves state of the art
read mapping F1 score by
up to 11%

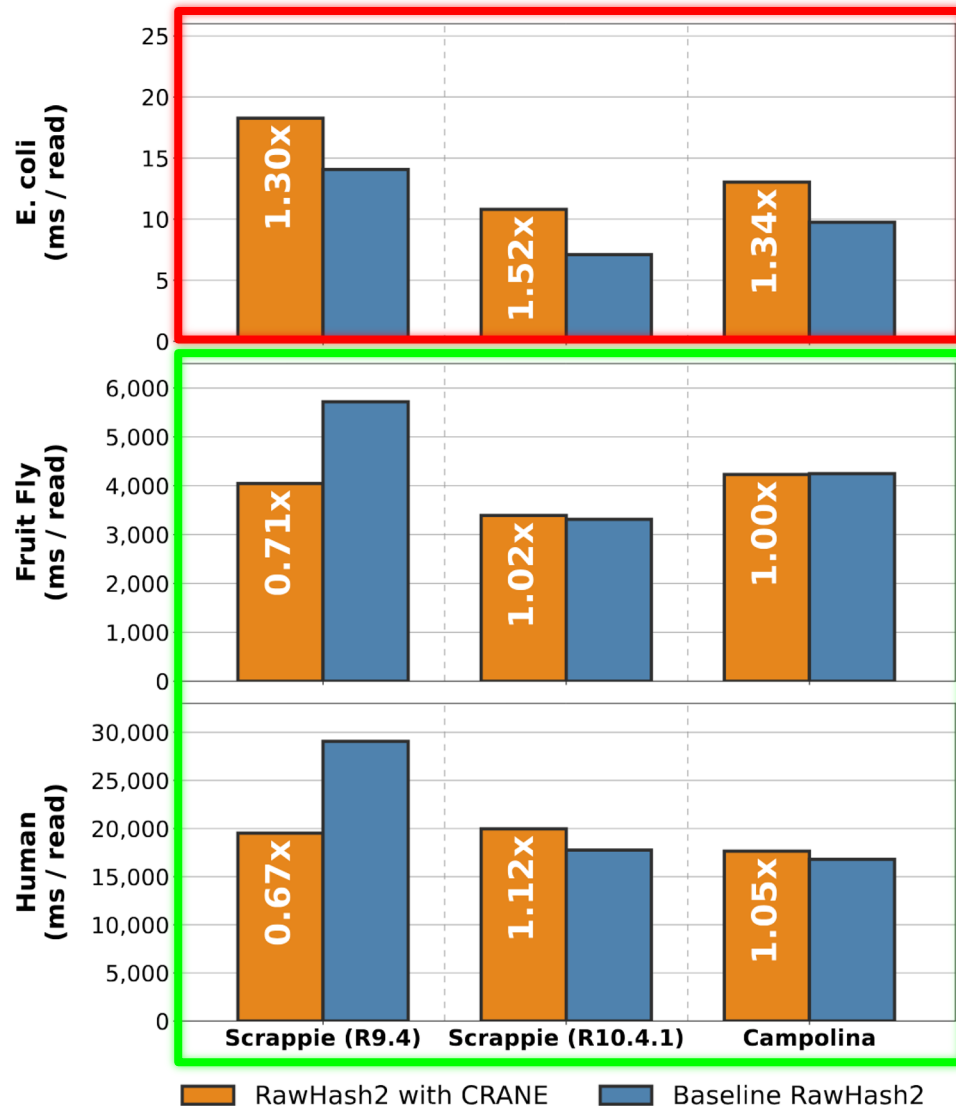


Accuracy Results - RawHash2 Read-mapping w/ HPC

CRANE consistently improves
the F1 score of RawHash2,
with and without HPC



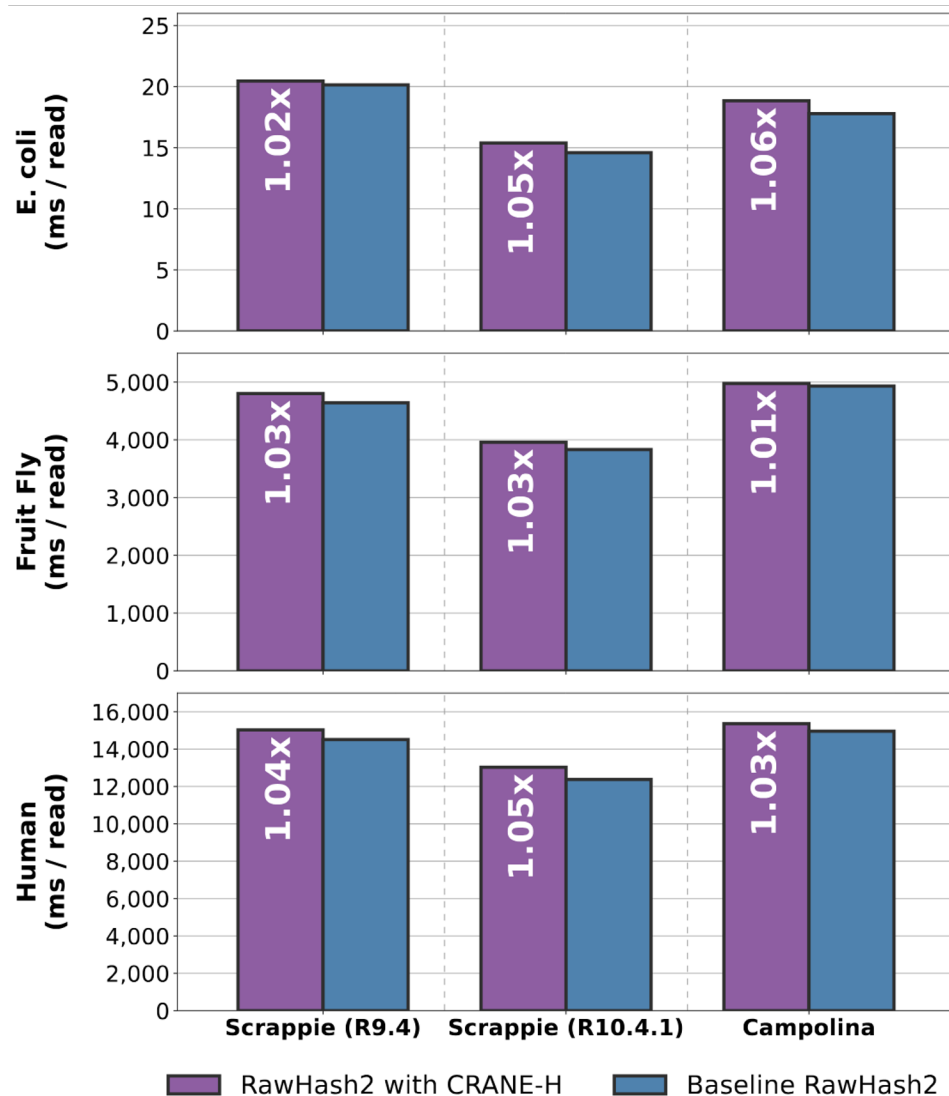
Runtime Results - CRANE + RawHash2 w/o HPC



CRANE adds a small total runtime increase (**<1.12x**) for larger genomes

CRANE accounts for **<1%** of total runtime for human genome

Runtime Results - CRANE + RawHash2 w/ HPC



With HPC, CRANE has a negligible effect on runtime (<1.06x)