Stanford

Overcoming high nanopore basecaller error rates for DNA storage via basecaller-decoder integration and convolutional codes

Shubham Chandak

Stanford University

ICASSP 2020

Team and funding


Shubham Chandak


Kedar Tatwawadi


Joachim Neu


Jay Mardia


Billy Lau

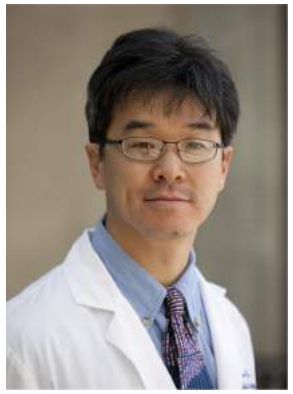

Matt Kubit


Reyna Hulett


Peter Griffin


Tsachy Weissman


Mary Wootters


Hanlee Ji

SemiSynBio: Highly scalable random access DNA data storage with nanopore-based reading

Beckman Center Innovative Technology Seed Grant

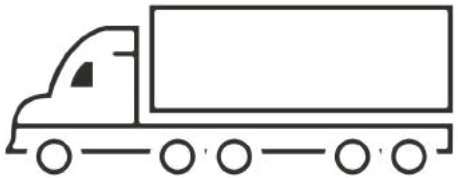
Scalable Long-Term DNA Storage with Error Correction and Random-Access Retrieval

National Institutes of Health

Motivation

200 Petabyte

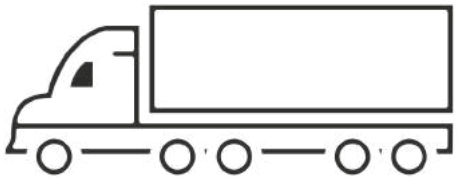
200 Petabyte



40,000 x 5 TByte HDDs
40 tons

10s of years

200 Petabyte



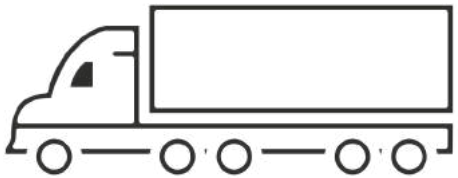
40,000 x 5 TByte HDDs
40 tons

10s of years

DNA
1 gram

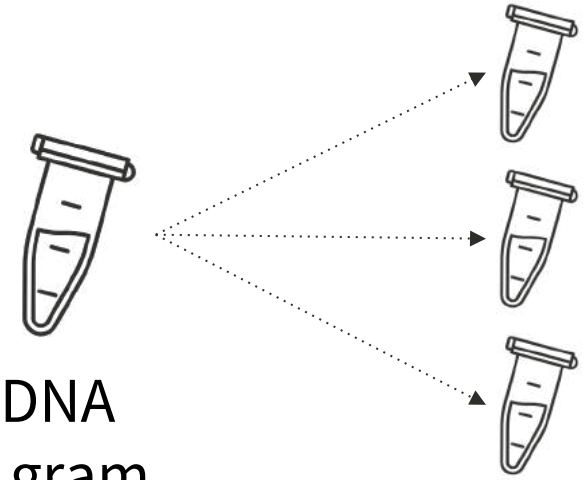
1,000s of years

200 Petabyte



40,000 x 5 TByte HDDs
40 tons

10s of years



DNA
1 gram

Easy duplication

1,000s of years

DNA storage setup

Building block: synthesis

- Ability to “**write/synthesize**” artificial DNA (sequence of {A,C,G,T})

CustomArray | GenScript Make Research Easy

Current ability: short ssDNA oligos (~150nt) at scale

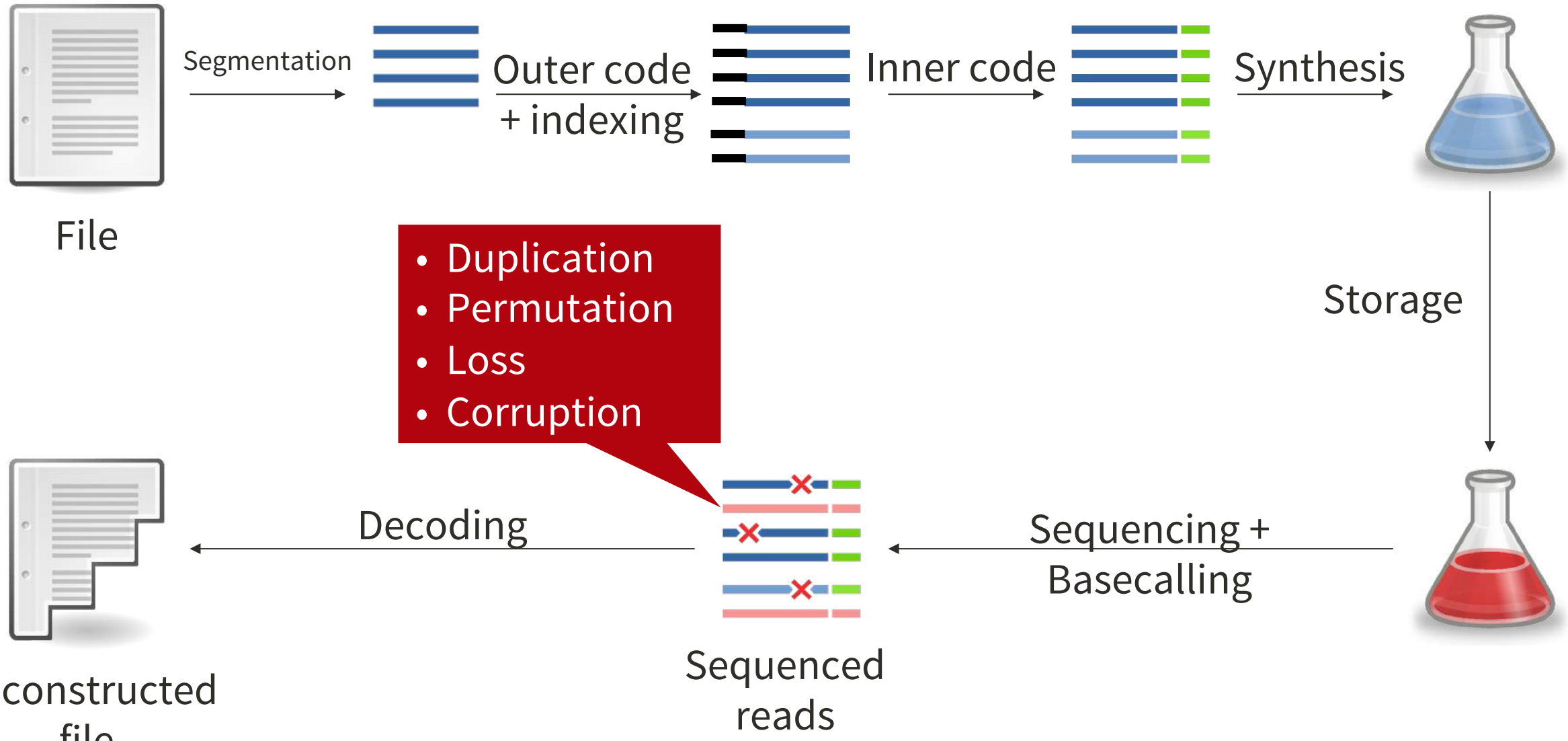
DNA Synthesis is not perfect: Usually has ~1% insertion/Deletion error

Building block: sequencing

- Nanopore sequencing: portable, real time

https://directorsblog.nih.gov/2018/02/06/sequencing-human-genome-with-pocket-sized-nanopore-device/

Typical DNA Storage System



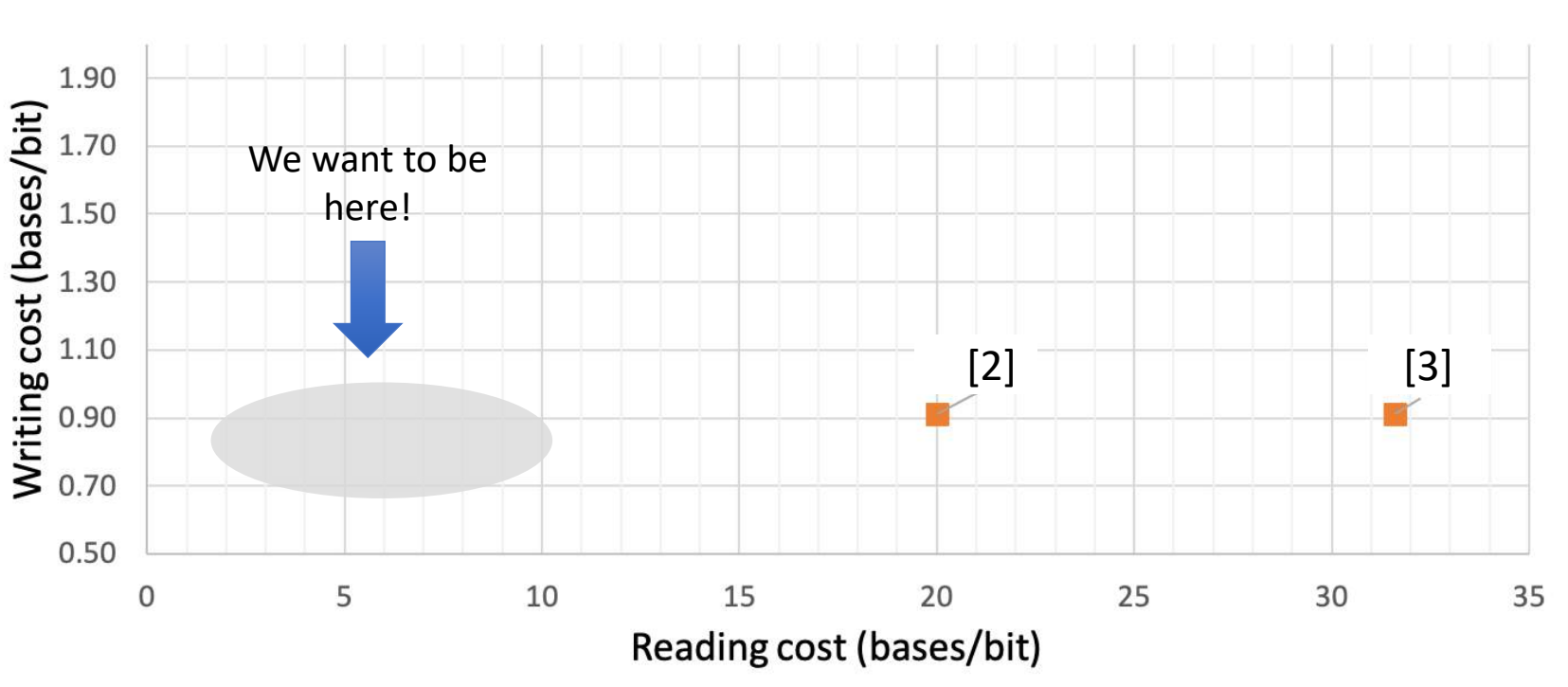
Challenges

- High basecall error rates for nanopore sequencing
  - 5-10% edit distance
  - Predominantly insertion and deletion errors
- Lack of good error correction codes for this setting

Challenges

- High basecall error rates for nanopore sequencing
  - 5-10% edit distance
  - Predominantly insertion and deletion errors
- Lack of good error correction codes for this setting
- Most previous works rely on consensus over multiple reads – **high reading cost**
  - Sequence the input lot of times (~30-40x)
  - Cluster by *index*, and perform “averaging” to reduce the error

Previous Works

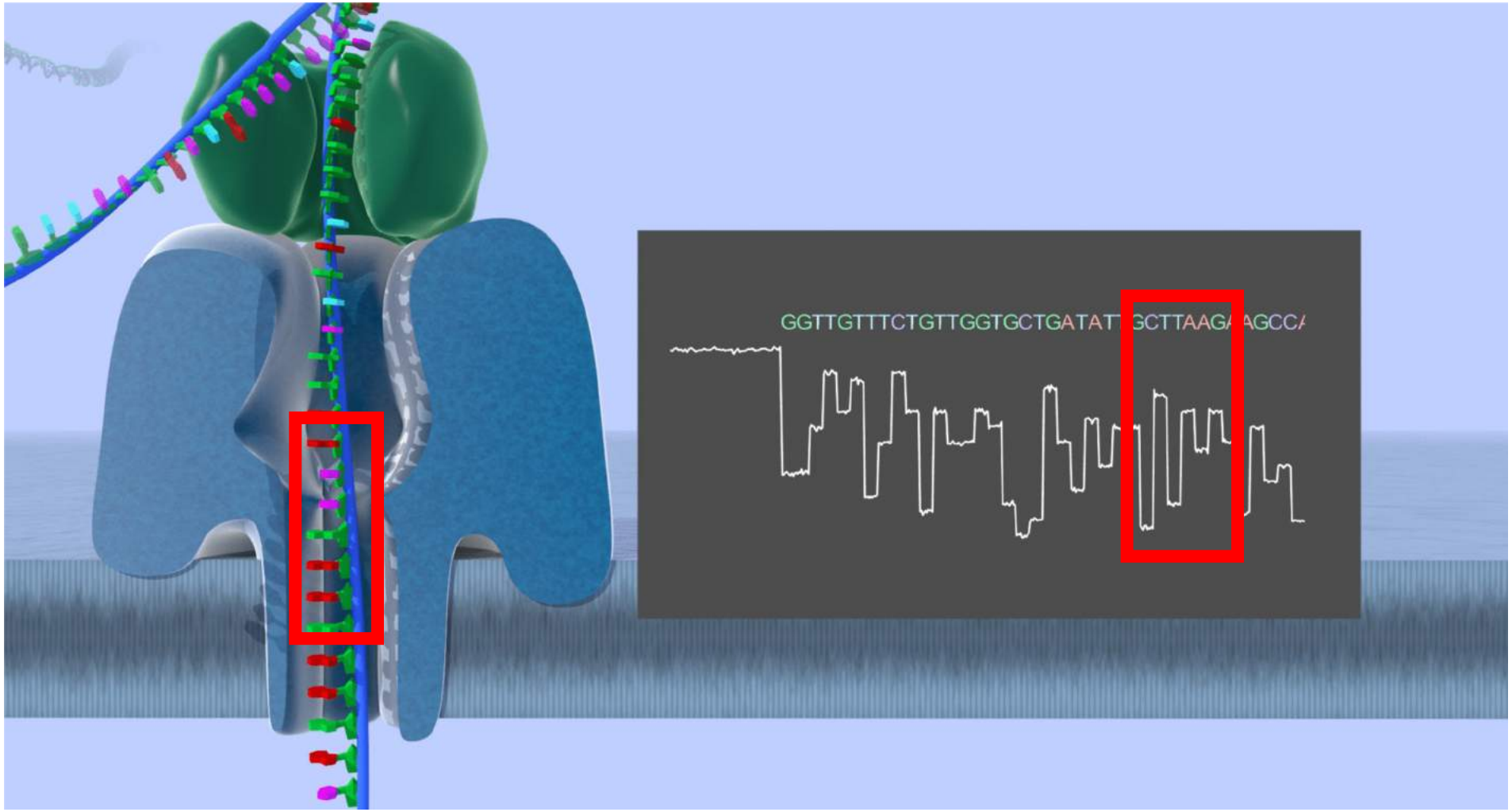


[2] L. Organick *et al.*, “Random access in large-scale DNA data storage," *Nature biotechnology*, vol. 36, no. 3, p. 242, 2018.

[3] Randolph Lopez et al., “DNA assembly for nanopore data storage readout,” Nature communications, vol. 10, no. 1, pp. 2933, 2019.

Methods

Nanopore Physics

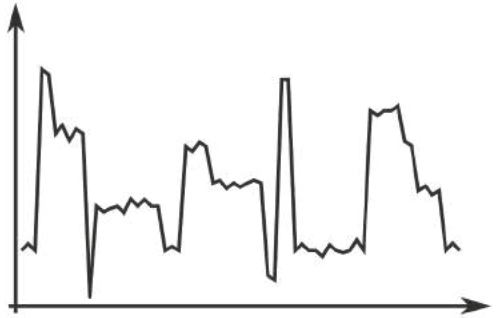


Nanopore Sequencing Model

… ACGTACGTACGT …

Nanopore sequencing channel

- Memory (inter-symbol interference)
- Base skips
- Fading
- Random symbol duration
- Noise



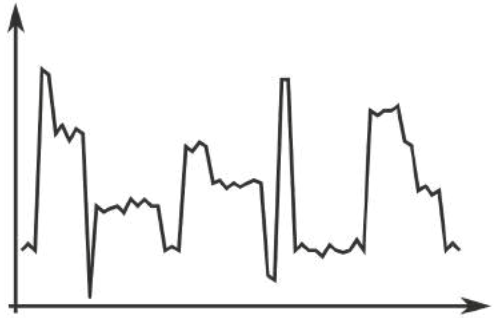
Source: "Models and Information-Theoretic Bounds for Nanopore Sequencing", Wei Mao et al., IEEE Trans. Inf. Theory 2017

Nanopore Sequencing Model

… ACGTACGTACGT …

Nanopore sequencing channel

- Memory (inter-symbol interference)
- Base skips
- Fading
- Random symbol duration
- Noise



VERY HARD TO MODEL AND ANALYZE FAITHFULLY

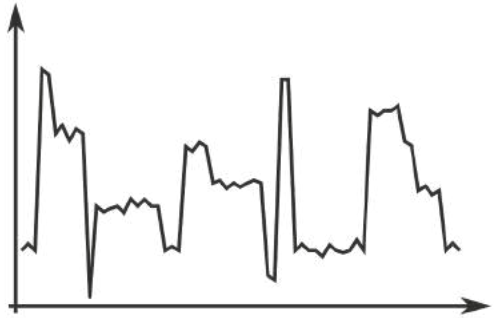
Source: "Models and Information-Theoretic Bounds for Nanopore Sequencing", Wei Mao et al., IEEE Trans. Inf. Theory 2017

Nanopore Sequencing Model

… ACGTACGTACGT …

Nanopore sequencing channel

- Memory (inter-symbol interference)
- Base skips
- Fading
- Random symbol duration
- Noise



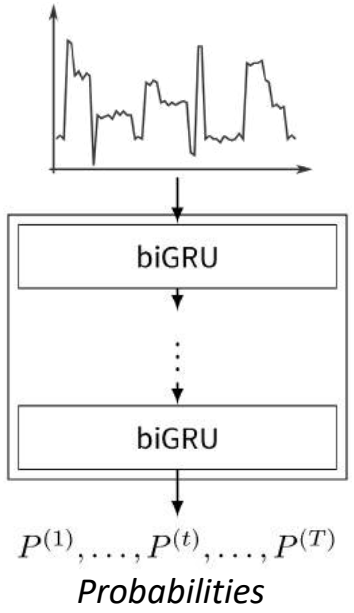
VERY HARD TO MODEL AND ANALYZE FAITHFULLY

COMBINE STRENGTHS OF MACHINE LEARNING & CODING THEORY!

Source: "Models and Information-Theoretic Bounds for Nanopore Sequencing", Wei Mao et al., IEEE Trans. Inf. Theory 2017

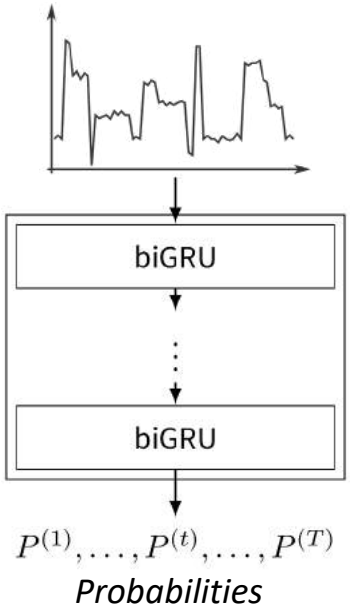
Key idea

Key idea



Using Flappie basecaller (Oxford Nanopore)

Key idea

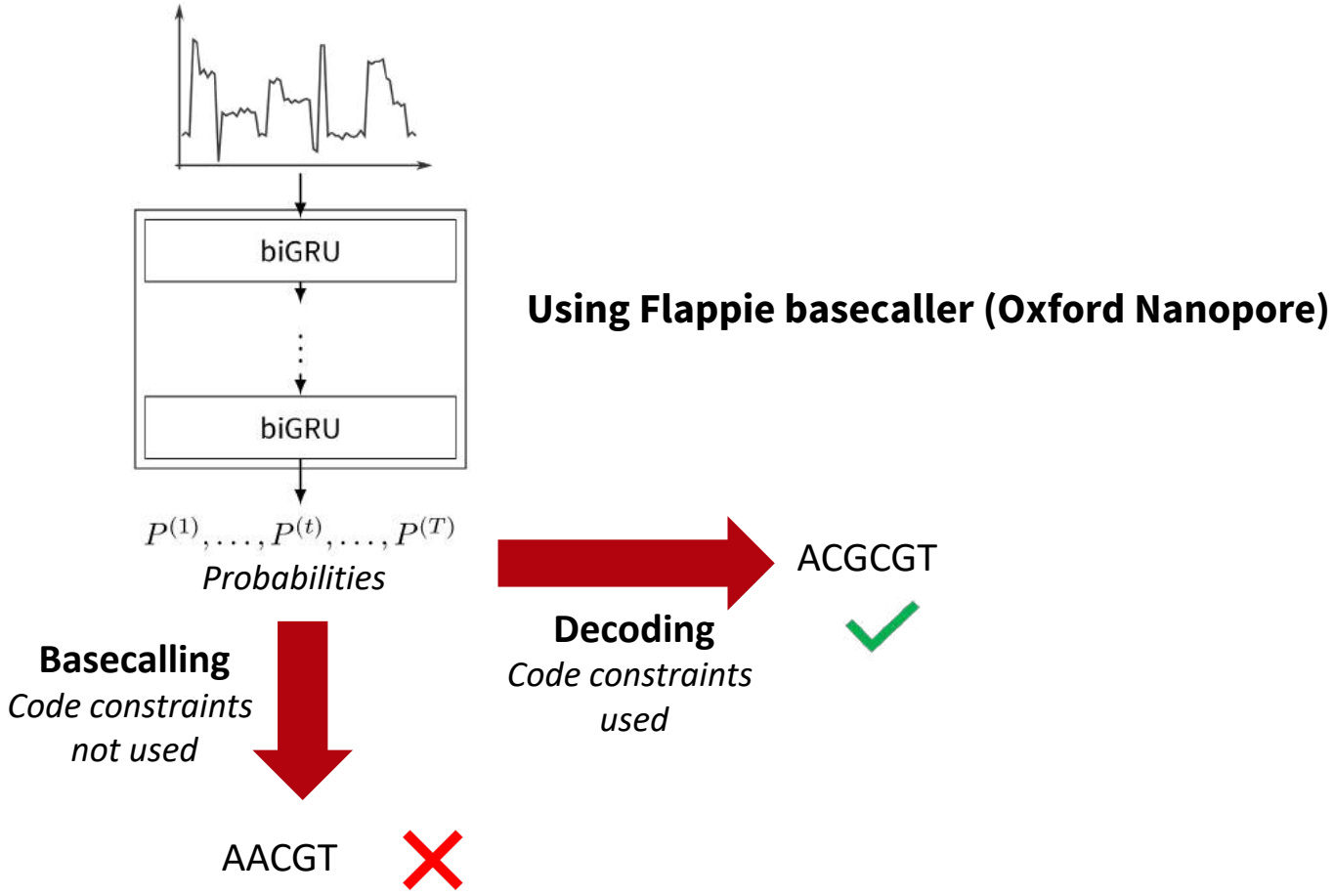


Basecalling
Code constraints not used

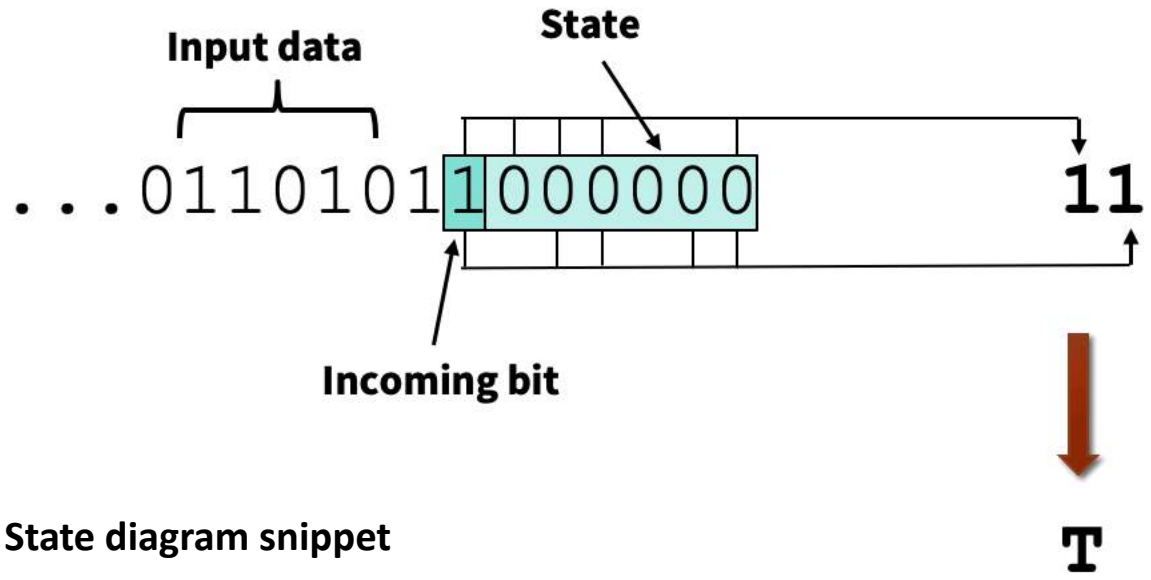
AACGT ✗

Using Flappie basecaller (Oxford Nanopore)

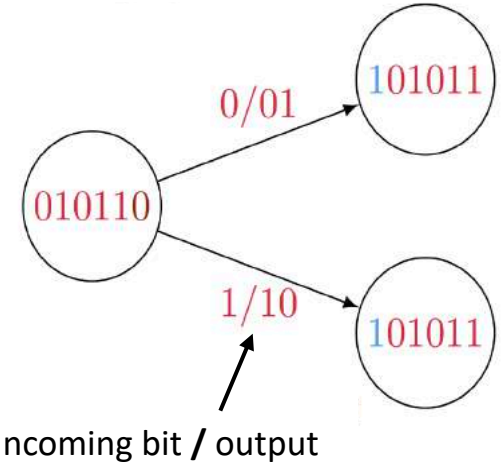
Key idea


biGRU
biGRU
$P^{(1)}, \ldots, P^{(t)}, \ldots, P^{(T)}$
Probabilities
Using Flappie basecaller (Oxford Nanopore)
Decoding
Code constraints used
ACGCGT
Basecalling
Code constraints not used
AACGT

Convolutional Codes as the Inner Code



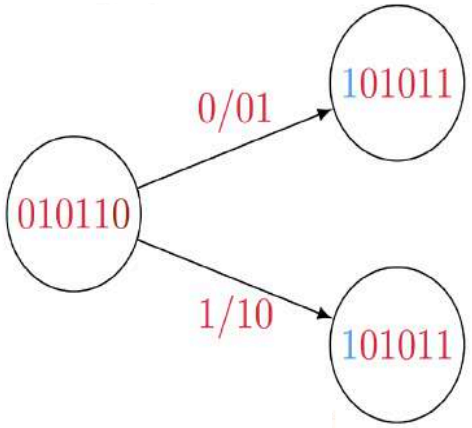
State diagram snippet



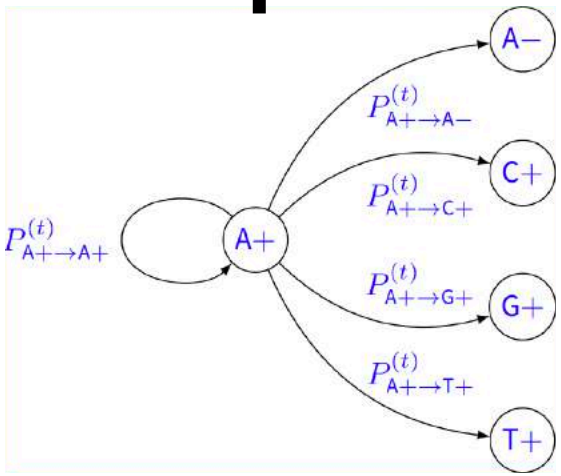
Convolution code parameters:
r = 1/2 (rate)
m = 6 (memory)

Basecaller-decoder integration

Convolutional code



+

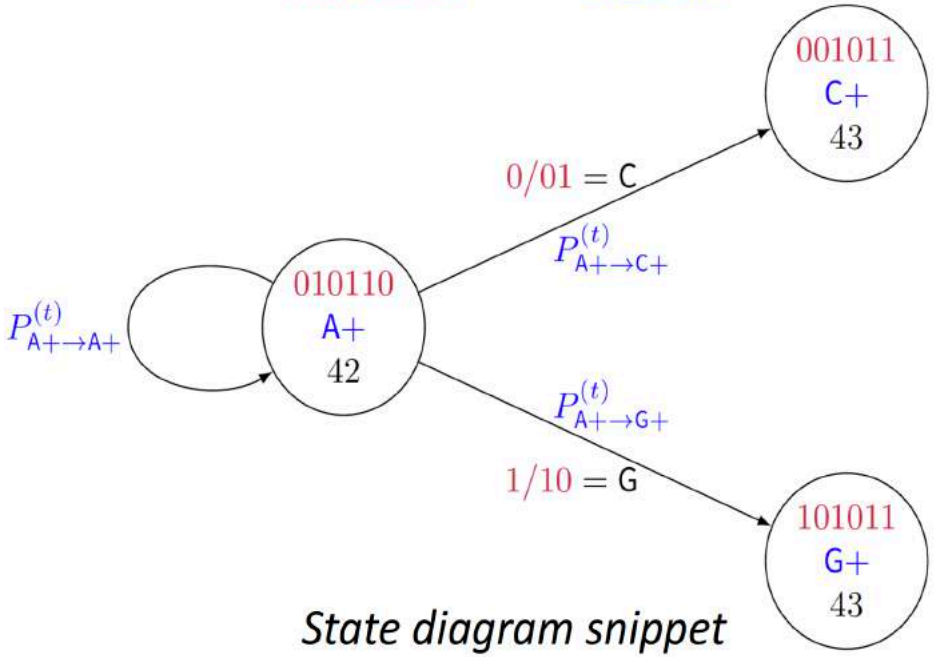


NN-modeling based transition probabilities

=

Combining NN-modeling + convolutional codes

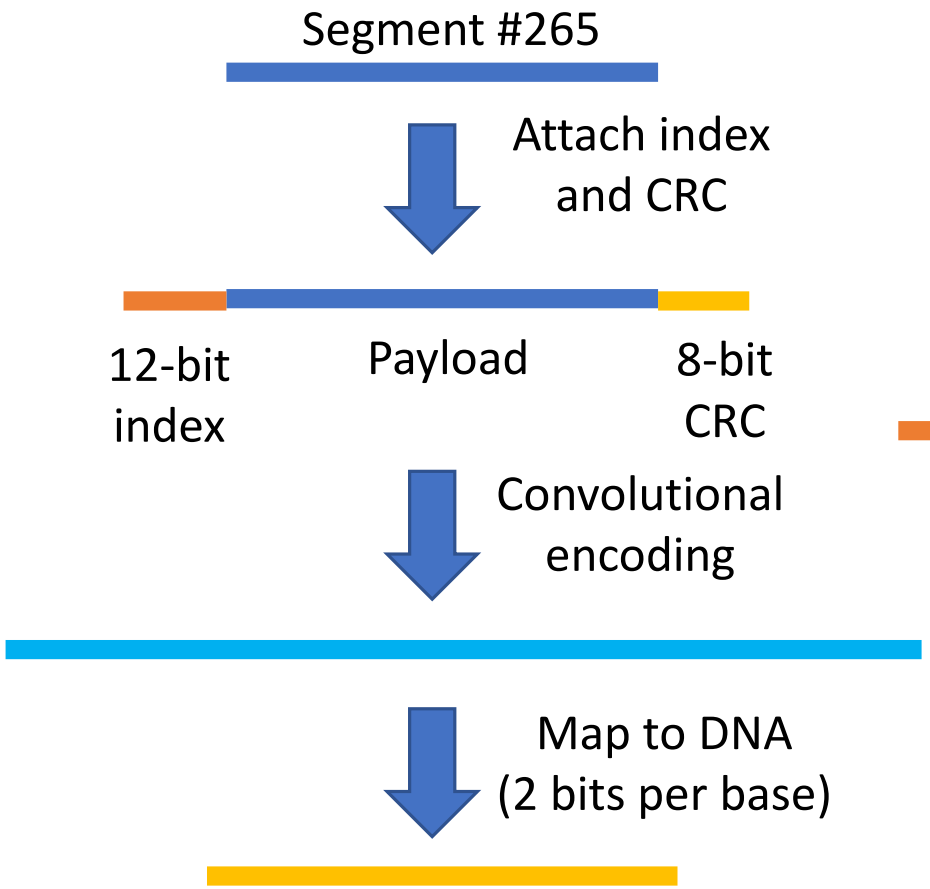
state = (state$_{convolutional}$, state$_{basecaller}$, position)



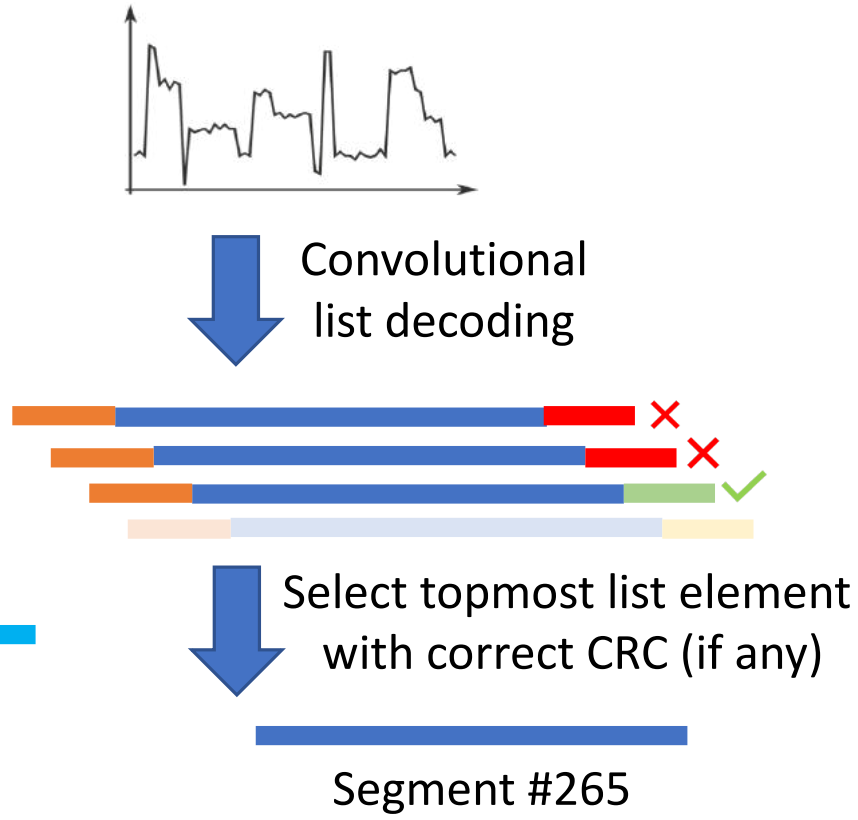
State diagram snippet

Perform Viterbi decoding using the modified state diagram

Overall Inner Code design



(a) Inner code encoding



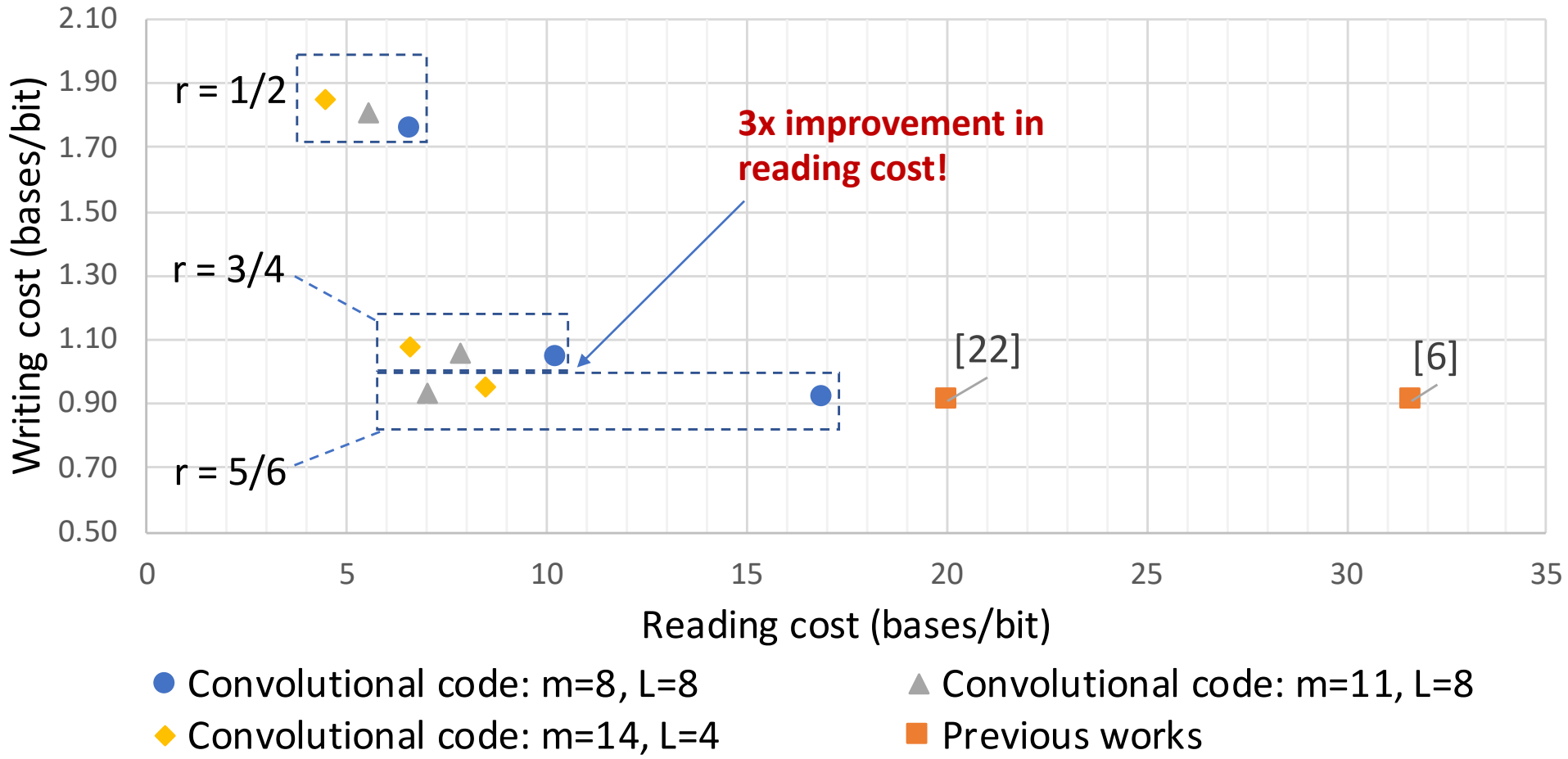
(b) Inner code decoding

Experiments and results

Experiments

- **Data**: 11KB of data: The Gettysburg Address, UN Declaration, “I have a Dream” Speech, poem collections, …
- **Final Error Correction Code Design:**
  - Reed Solomon outer code: 30% redundancy (default)
  - Pretrained Model from the ONT Flappie Basecaller
- **Synthesis:** Data Synthesized using CustomArray synthesis, into oligos of length ~165
- **Experiments:**
  - Rate of convolution code: r = 1/2, 3/4, 5/6
  - Memory: m = 8,11,14
  - List Size: 4, 8

Results



[6] L. Organick *et al.*, “Random access in large-scale DNA data storage," *Nature biotechnology*, vol. 36, no. 3, p. 242, 2018.

[22] Randolph Lopez et al., “DNA assembly for nanopore data storage readout,” Nature communications, vol. 10, no. 1, pp. 2933, 2019.

Conclusions and future work

- Novel error-correction mechanism for nanopore sequencing based DNA storage
  - Use “soft-information” from raw signal to improve decoding
  - Use neural net in basecaller to distil information from “hard-to-model” raw signal
  - Use convolutional codes that align nicely with sequential nanopore model
- Requires 3x fewer reads for decoding than previous works

Conclusions and future work

- Novel error-correction mechanism for nanopore sequencing based DNA storage
  - Use “soft-information” from raw signal to improve decoding
  - Use neural net in basecaller to distil information from “hard-to-model” raw signal
  - Use convolutional codes that align nicely with sequential nanopore model
- Requires 3x fewer reads for decoding than previous works
- Future work:
  - Optimization of convolutional code and CRC parameters
  - Finetuning of neural network model and use of improved basecallers
  - Application to other novel synthesis methodologies

Thank You!

Code and data available at

https://github.com/shubhamchandak94/nanopore_dna_storage