CMSC 829C: Algorithms, AI, and Hardware for Biological Data

Lecture 6: AI in Genomics

Fall 2026

September 17, 2026

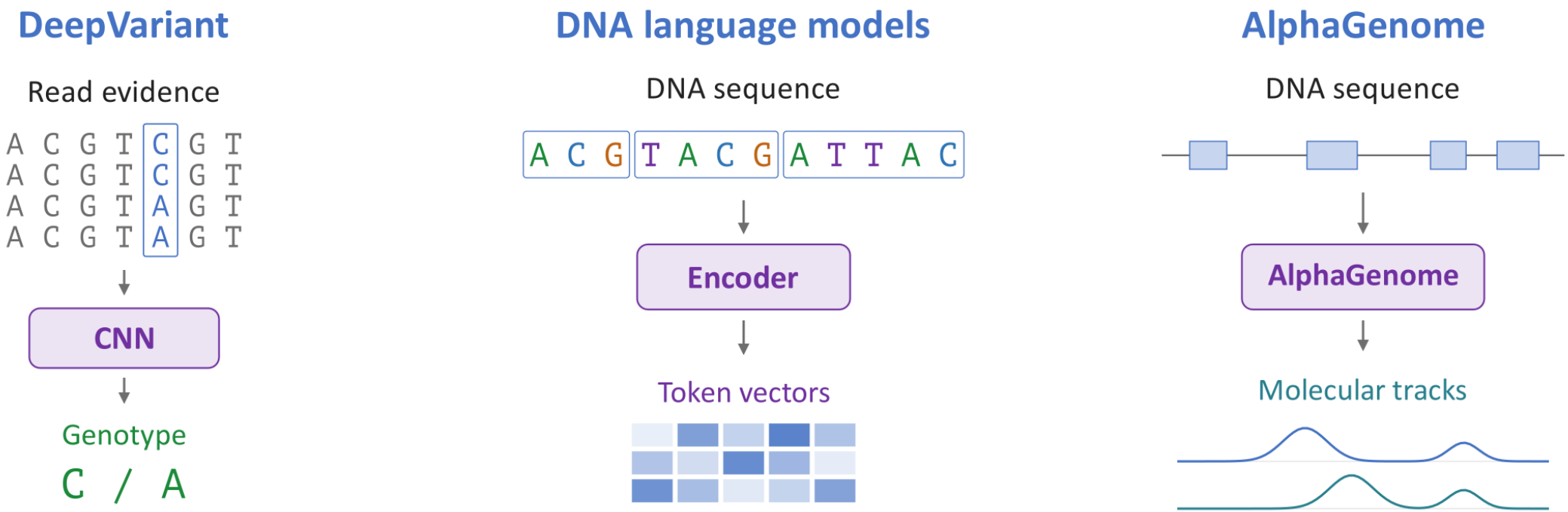
Can Firtina | canfirtina@gmail.com | cs.umd.edu/~firtina

Assistant Professor of Computer Science at University of Maryland (UMD)

UNIVERSITY OF MARYLAND

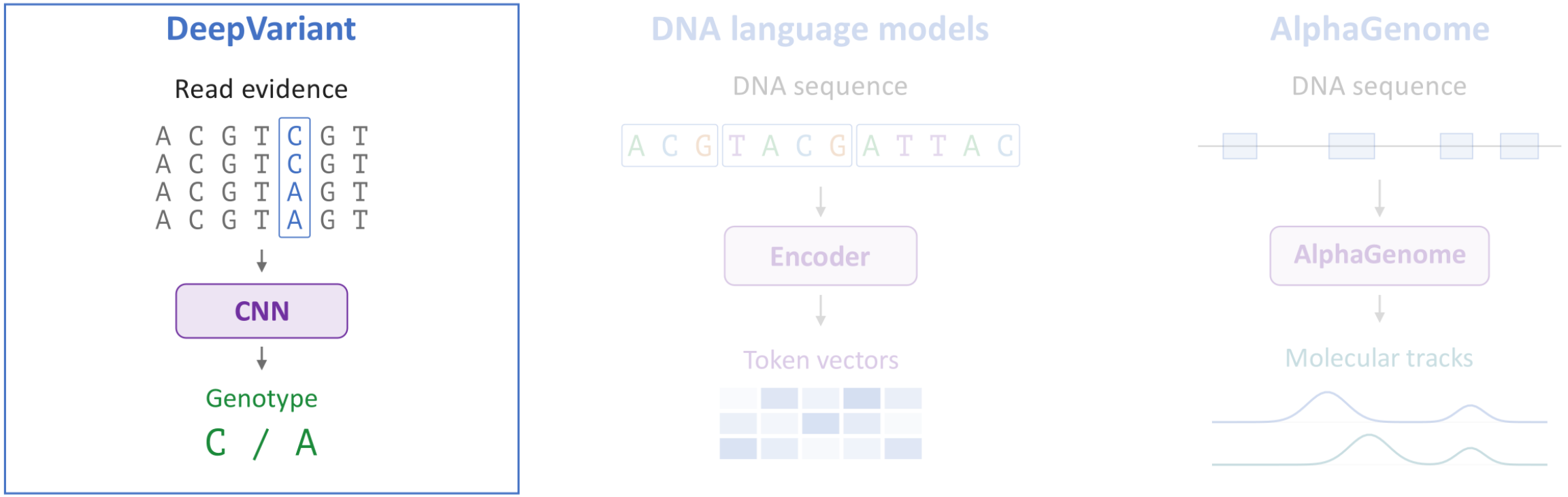
Agenda

- DeepVariant: Identify variants from read evidence
- DNA language models: Learn representations
- AlphaGenome: Predict molecular activity and variant effects



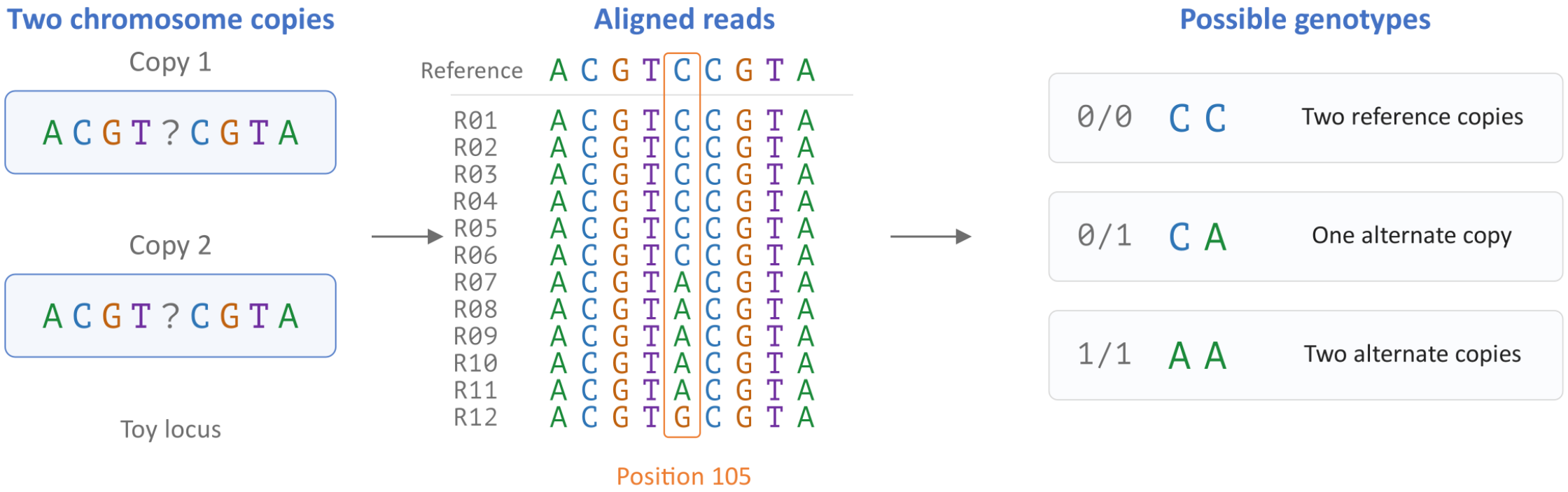
AI in Genomics

- **DeepVariant: Identify variants from read evidence**
- DNA language models: Learn reusable sequence representations
- AlphaGenome: Predict molecular activity and variant effects



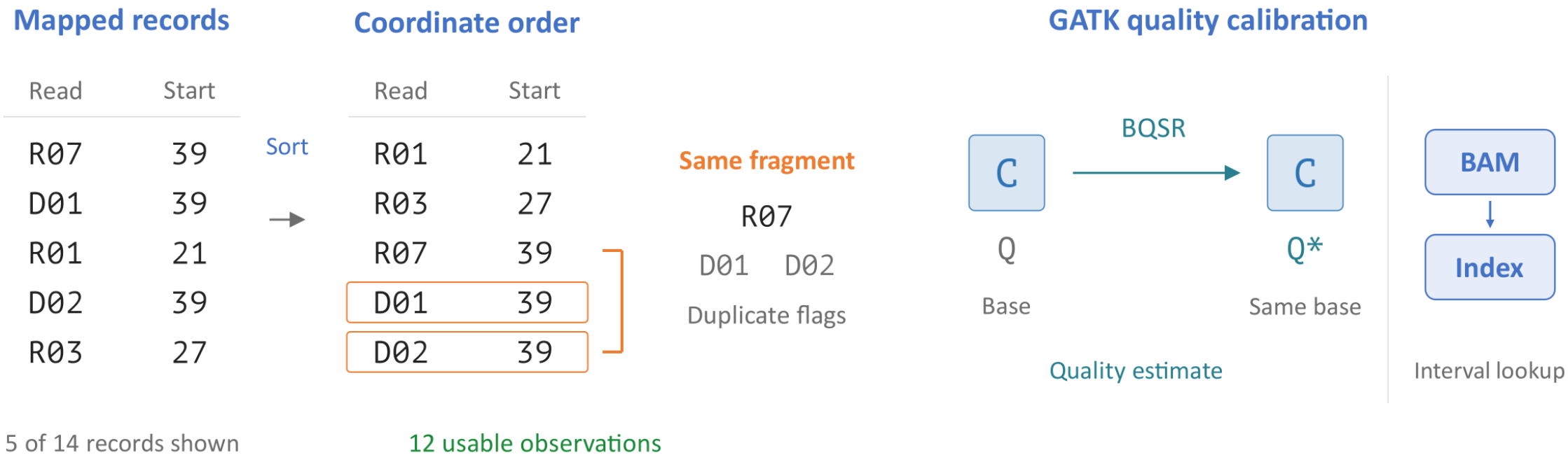
Which Differences Are Real?

- Diploid region: two chromosome copies
- Mapped reads provide noisy observations of both copies
- Genotype: the two sequence alternatives (alleles) at a site



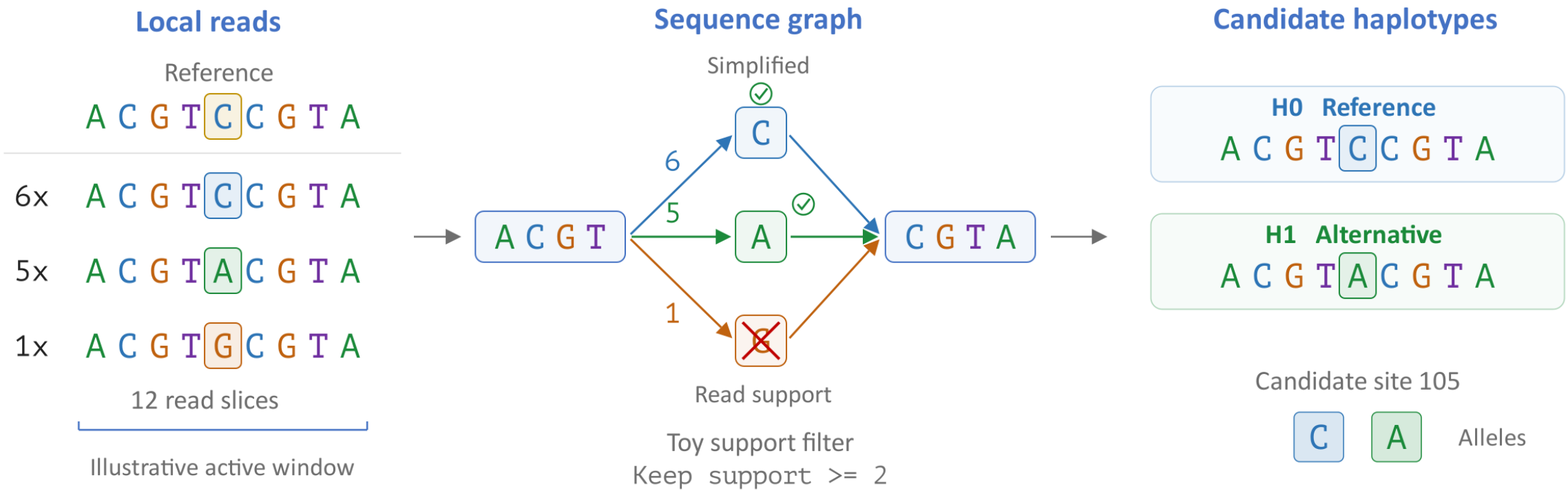
Prepare the Read Evidence

- Sort alignments so nearby reads can be processed together
- Flag likely copies of one fragment to limit repeated evidence
- Base quality score recalibration (BQSR) adjusts error estimates



HaplotypeCaller Proposes Local Sequences

- Genome Analysis Toolkit (GATK): focus on regions with variation
- Assemble candidate haplotypes: possible local DNA sequences
- Align candidates to the reference to locate possible variants



Score Reads against Each Haplotype

- Likelihood: how likely a read is under a candidate haplotype
- Pair hidden Markov model (PairHMM): sum over possible alignments
- Repeat for every read and haplotype to form a likelihood matrix

One read and candidate

R07

A C G T A C G T A

H1

A C G T A C G T A

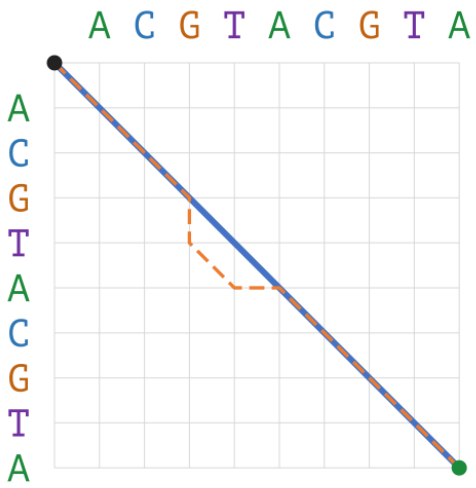
Base quality Q30

Toy error inputs

Qins 45 Qdel 45

Qgap 10

Alignment paths



M Both bases
I Read only
D Haplotype only

Sum paths

Path 1

0.09905346

Path 2

2.68e-14

Other paths

3.51e-06

0.09905697

P(Read | H1)

Log10 = -1.004

Read likelihoods

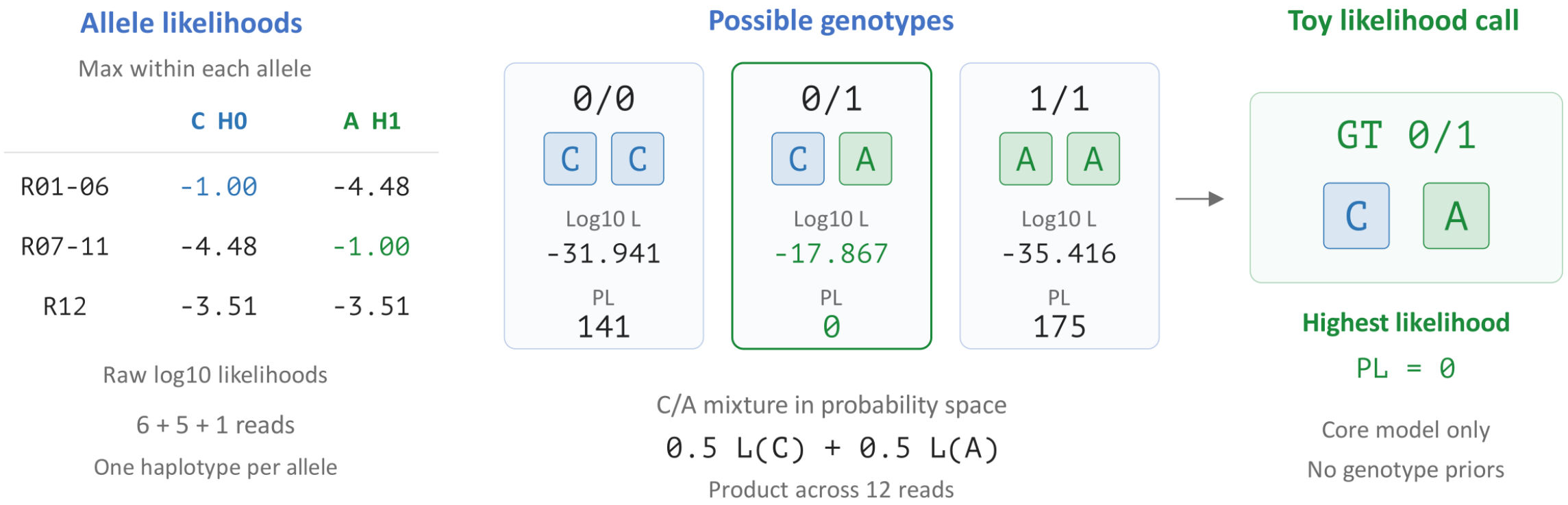
	H0 C	H1 A
R01	-1.004	-4.480
R02	-1.004	-4.480
R03	-1.004	-4.480
R04	-1.004	-4.480
R05	-1.004	-4.480
R06	-1.004	-4.480
R07	-4.481	-1.004
R08	-4.481	-1.004
R09	-4.481	-1.004
R10	-4.481	-1.004
R11	-4.481	-1.004
R12	-3.512	-3.512

Raw log10 likelihoods

Toy PairHMM core

From Read Evidence to a Genotype

- Keep each read's best haplotype likelihood for each allele
- Combine read likelihoods for C/C, C/A, and A/A
- Phred-scaled genotype likelihoods (PL): the best genotype has PL 0



Reproducibility Issues [Bioinformatics '16]

- **On genomic repeats and reproducibility**
  Can Firtina, Can Alkan
  Bioinformatics, August 2016.
  [Code] [Link]

Bioinformatics, 32(15), 2016, 2243–2247
doi: 10.1093/bioinformatics/btw139
Advance Access Publication Date: 11 March 2016
Discovery Note

OXFORD

Sequence analysis

On genomic repeats and reproducibility

Can Firtina and Can Alkan*

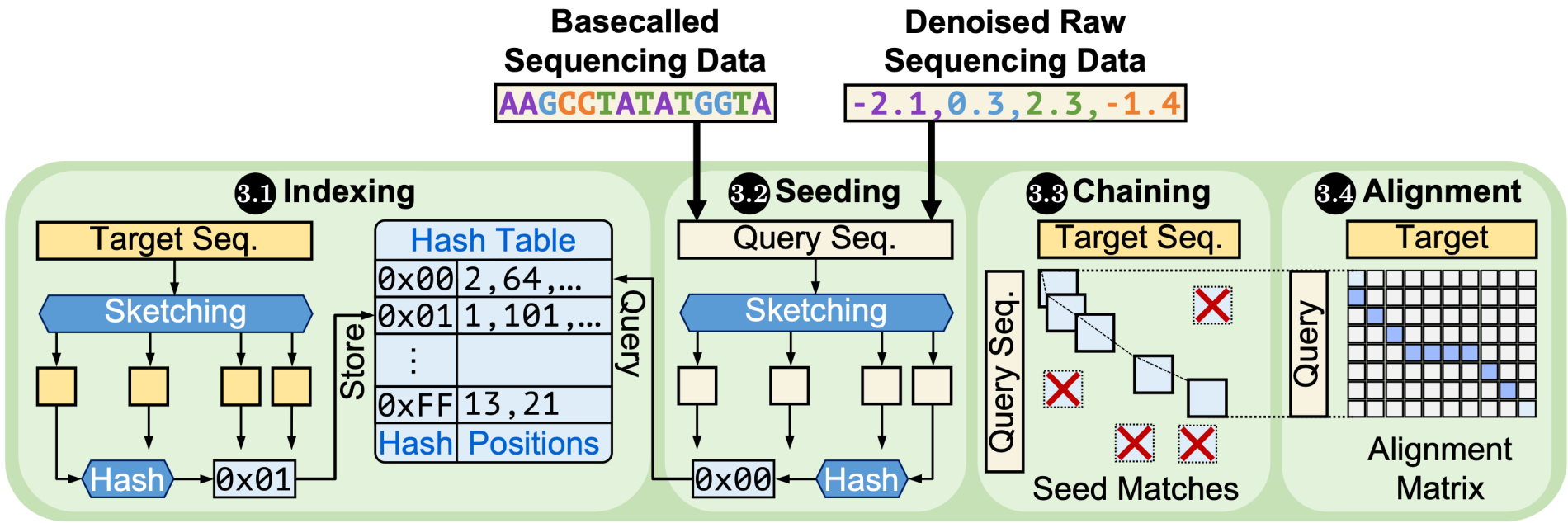
Department of Computer Engineering, Bilkent University, Ankara 06800, Turkey

*To whom correspondence should be addressed.
Associate Editor: Inanc Birol

Received on October 26, 2015; revised on January 16, 2016; accepted on March 7, 2016

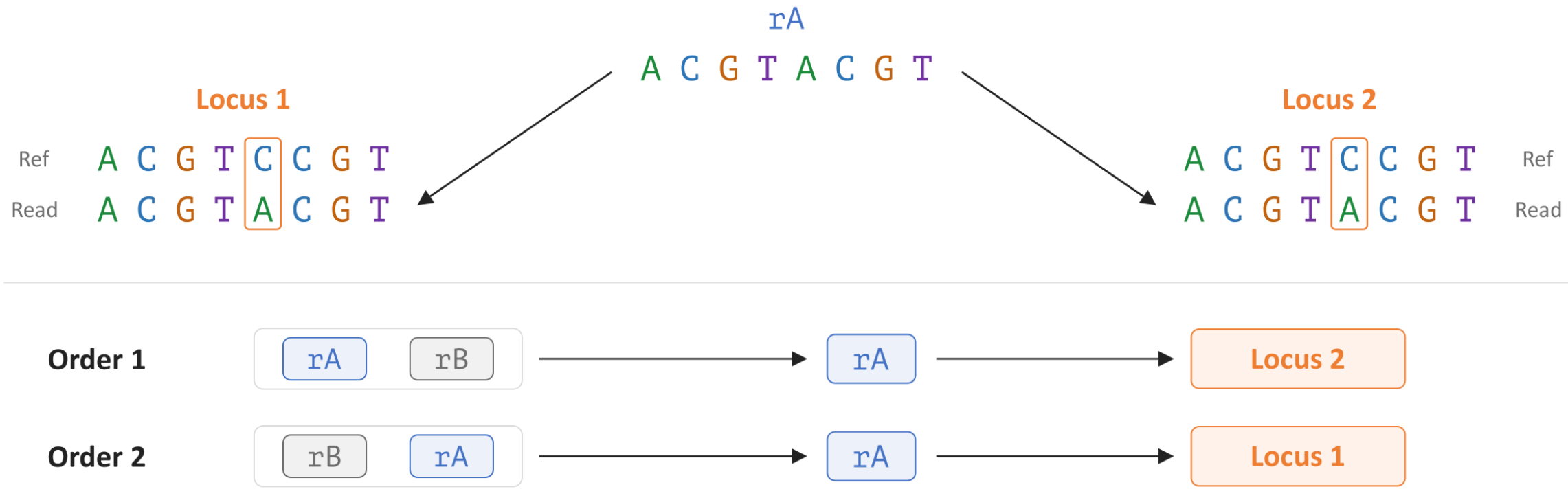
Recall: Read Mapping

- Reads are matched to positions in a reference genome
- What if several locations are equally good?



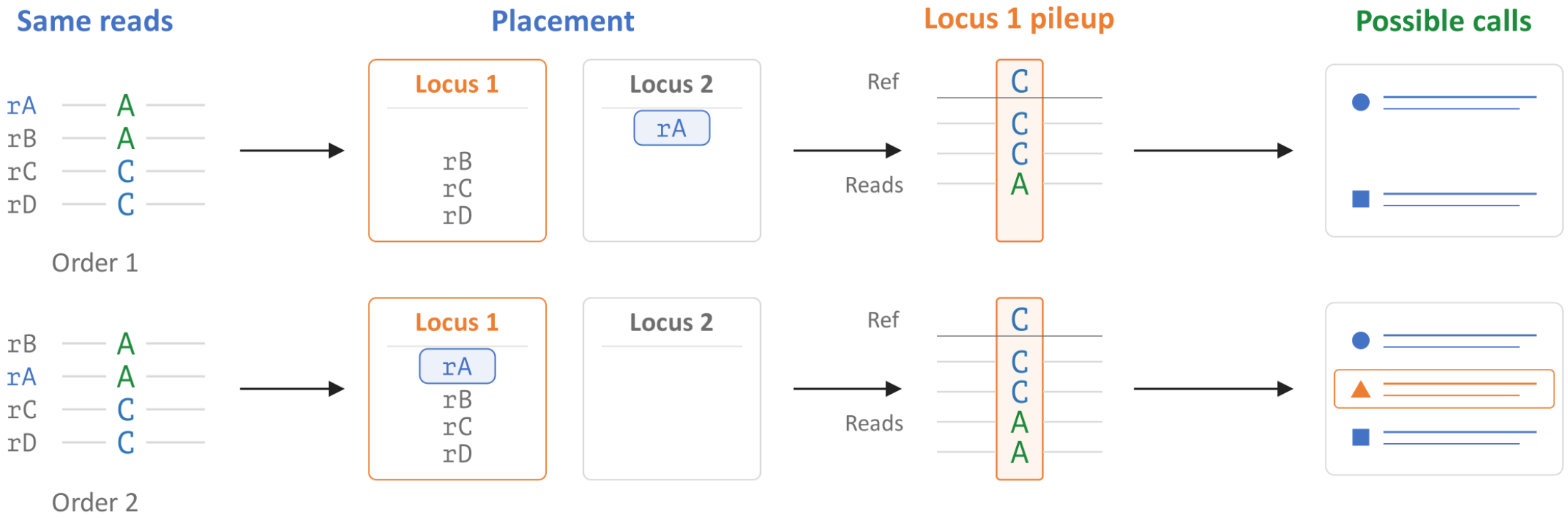
Ambiguous Mapping

- Which location should be reported for this read?
- Burrows–Wheeler Aligner with maximal exact matches (BWA-MEM)
- Pseudorandom tie-breaking depends on the read's input position



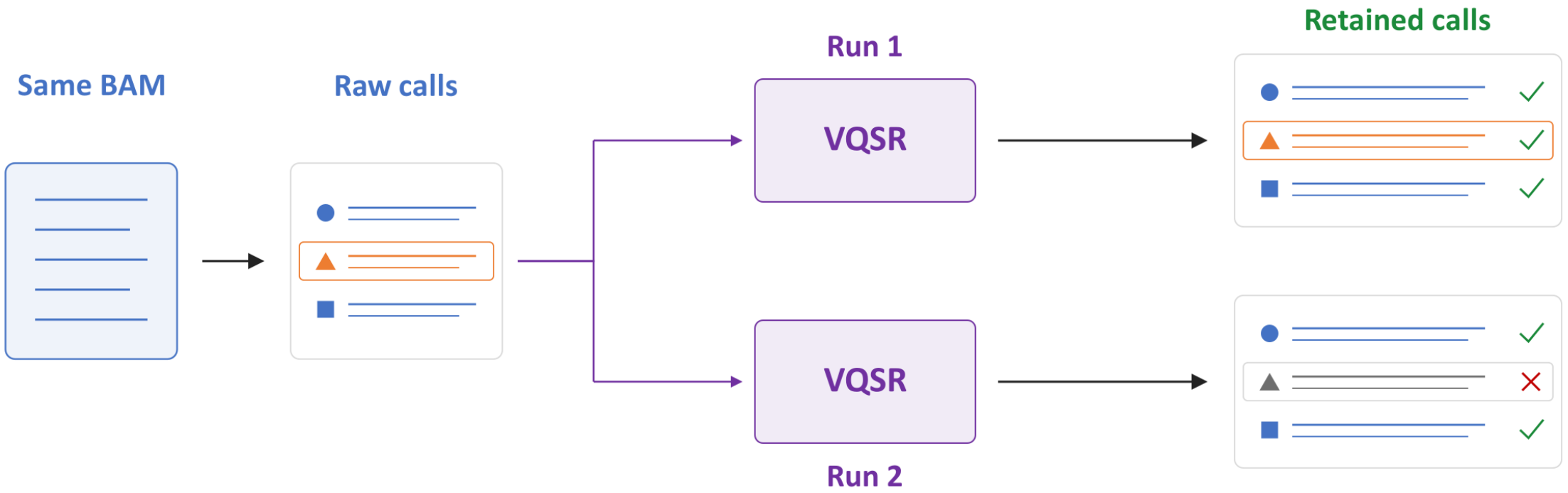
Could the Variant Calls Change?

- Would the same reads always produce the same variant calls?
- Changed placements can change the information a caller uses
- **Observation:** Different variant calling results after shuffling the same reads



Same Alignments, Same Final Calls?

- How about the **exact same Binary Alignment/Map (BAM) file**?
- 2016: variant quality score recalibration (VQSR) changed filtering
  - Different variant calling even when using the same BAM input



Reproducibility Issues [Bioinformatics '16]

- **On genomic repeats and reproducibility**
  Can Firtina, Can Alkan
  Bioinformatics, August 2016.
  [Code] [Link]

Bioinformatics, 32(15), 2016, 2243–2247
doi: 10.1093/bioinformatics/btw139
Advance Access Publication Date: 11 March 2016
Discovery Note

OXFORD

Sequence analysis

On genomic repeats and reproducibility

Can Firtina and Can Alkan*

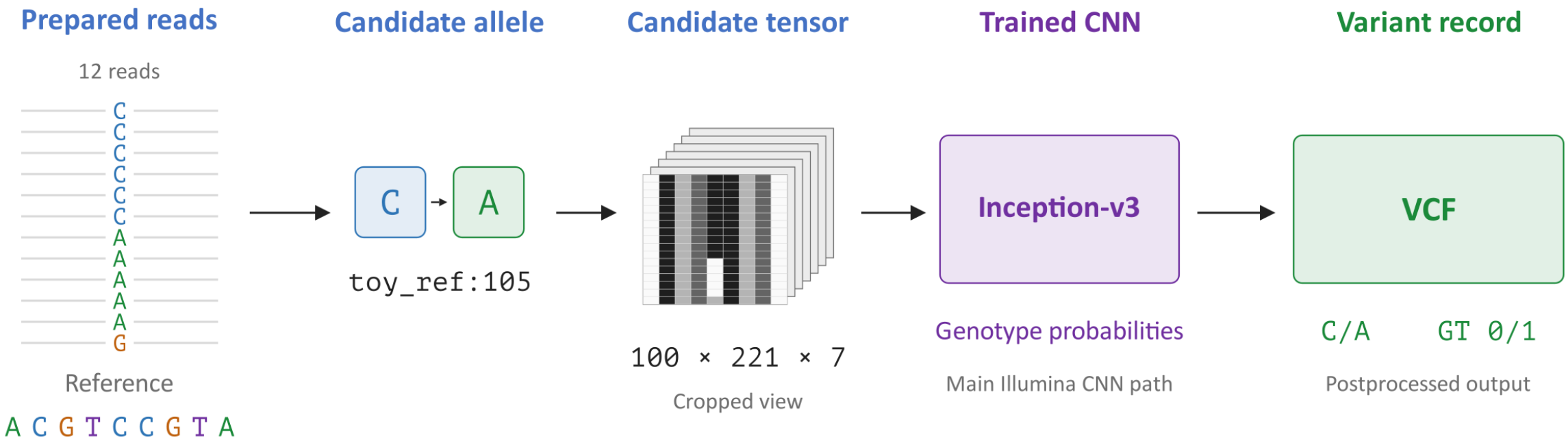
Department of Computer Engineering, Bilkent University, Ankara 06800, Turkey

*To whom correspondence should be addressed.
Associate Editor: Inanc Birol

Received on October 26, 2015; revised on January 16, 2016; accepted on March 7, 2016

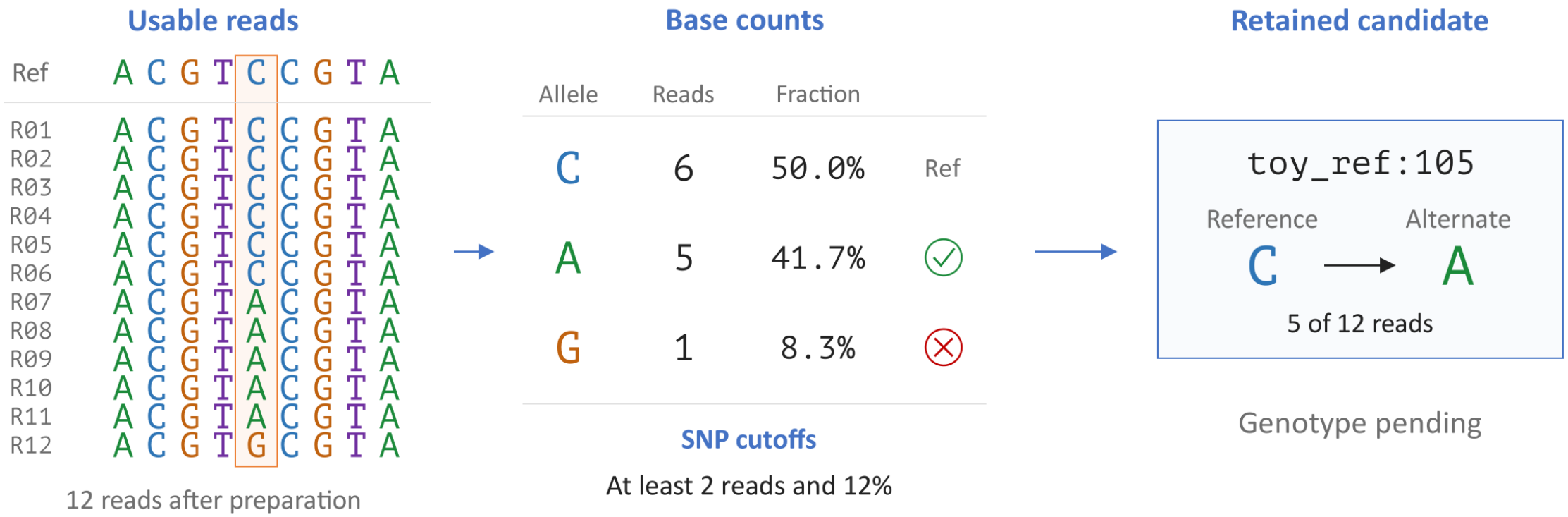
DeepVariant Learns to Interpret Read Evidence

- DeepVariant (Poplin et al., 2018): learn from read evidence
- DeepVariant 1.10 uses a convolutional neural network (CNN)
- Genotype probabilities processed into variant records



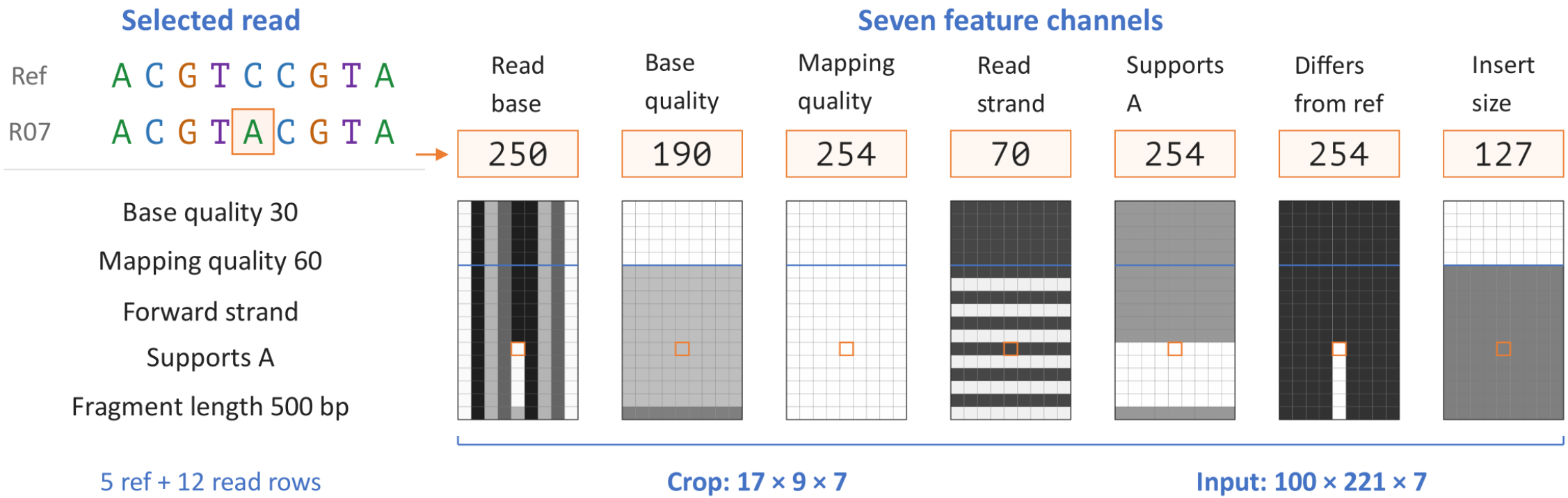
Propose Candidate Variants

- Count usable read observations at each reference position
- Single-nucleotide polymorphism (SNP) cutoffs: ≥2 reads and ≥12% support
- Passing alleles become candidates for genotype evaluation



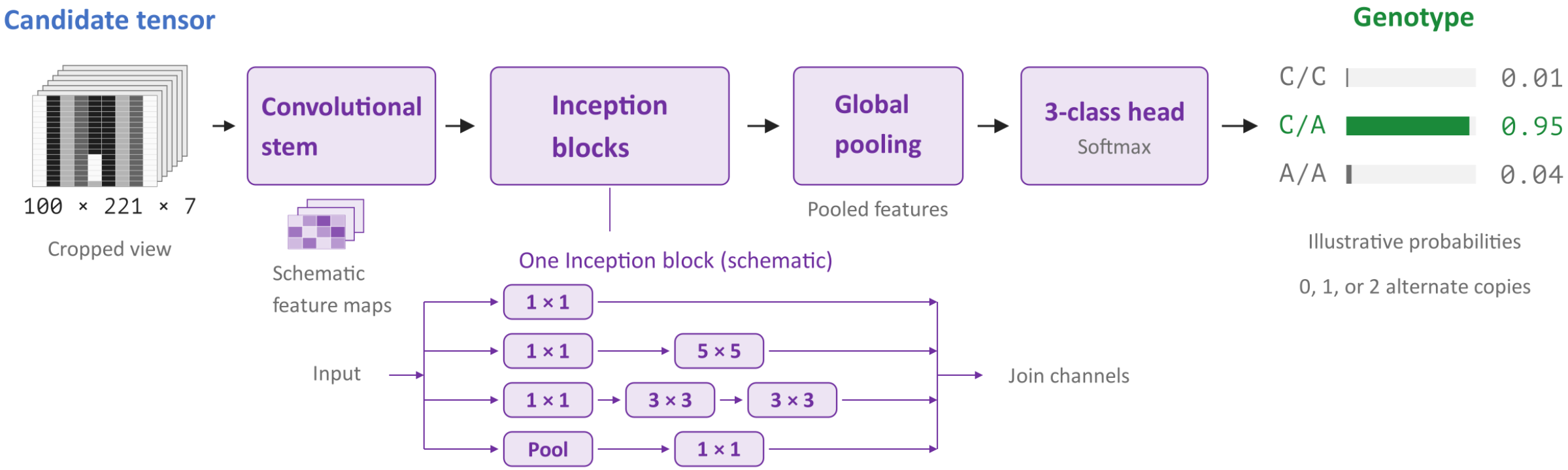
Encode the Evidence as Seven Features

- Each read position encoded by seven numerical features
- Rows aligned by genomic position, with five reference rows
- Illumina CNN input: 100 rows × 221 positions × 7 channels



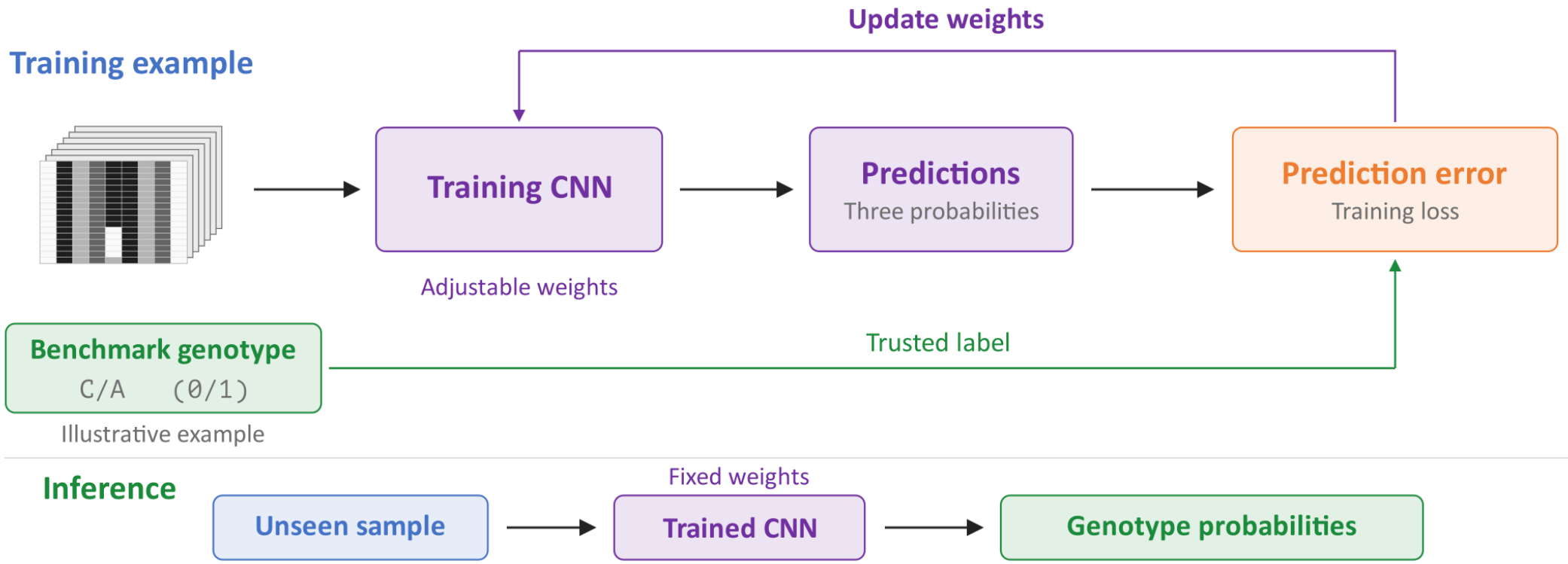
From Encoded Evidence to Genotype Probabilities

- Inception-v3 filters combine nearby values across channels
- Parallel branches merged, then summarized by average pooling
- Three probabilities for zero, one or two alternate copies



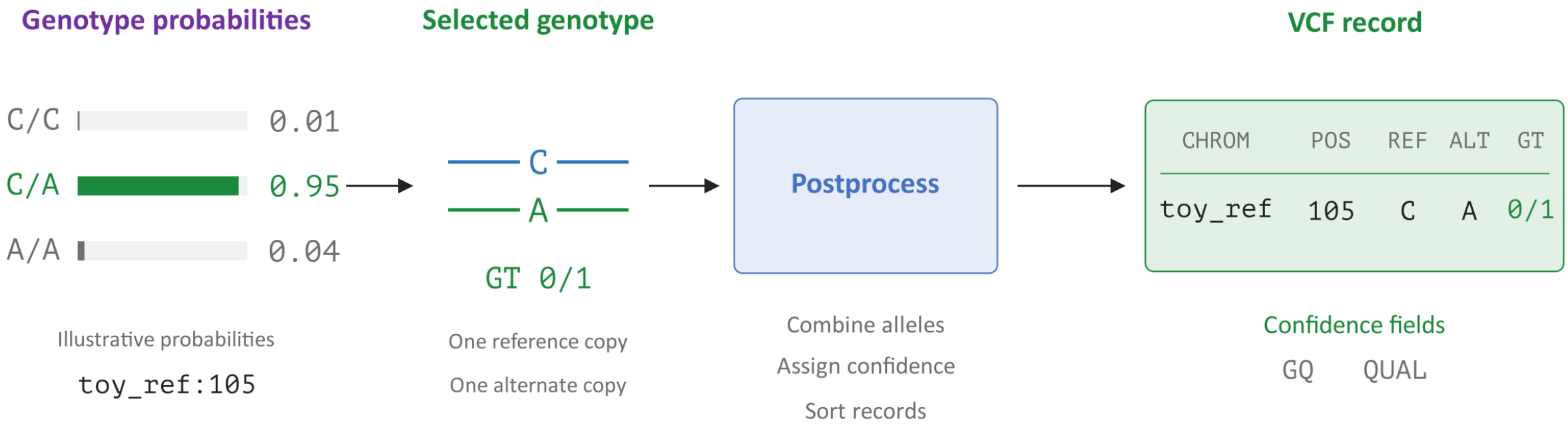
Learn from Benchmark Genotypes

- Training examples paired with benchmark genotype labels
- Prediction errors guide repeated updates to network weights
- Inference applies fixed weights to new samples



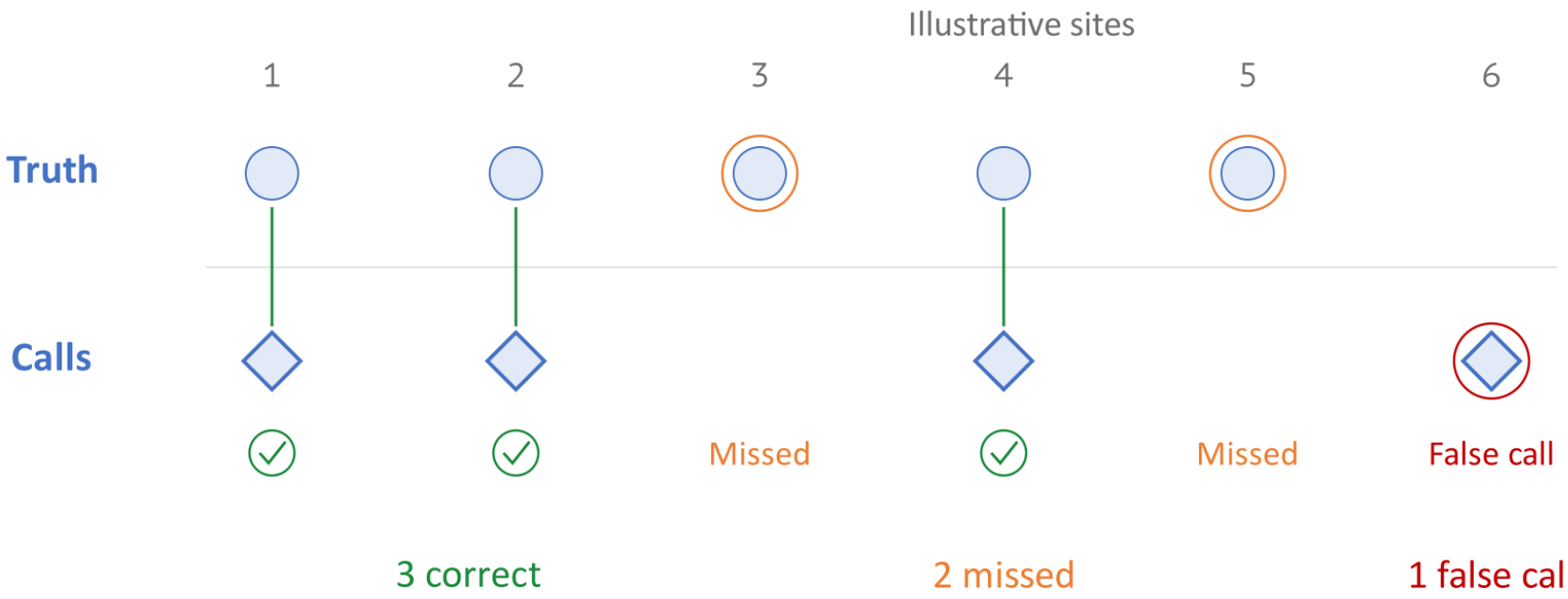
Turn Predictions into a Variant Record

- For this candidate, the largest probability selects C/A
- Postprocessing combines predictions and assigns confidence
- Variant Call Format (VCF) records: position, alleles, genotype, confidence



Measure Correct Calls and Missed Variants

- Compare calls with benchmark variants in confident regions
- Separate correct calls, missed variants and false calls
- Recall measures recovery, precision measures correctness



Recall

3 / 5 = 60%

Of the true variants

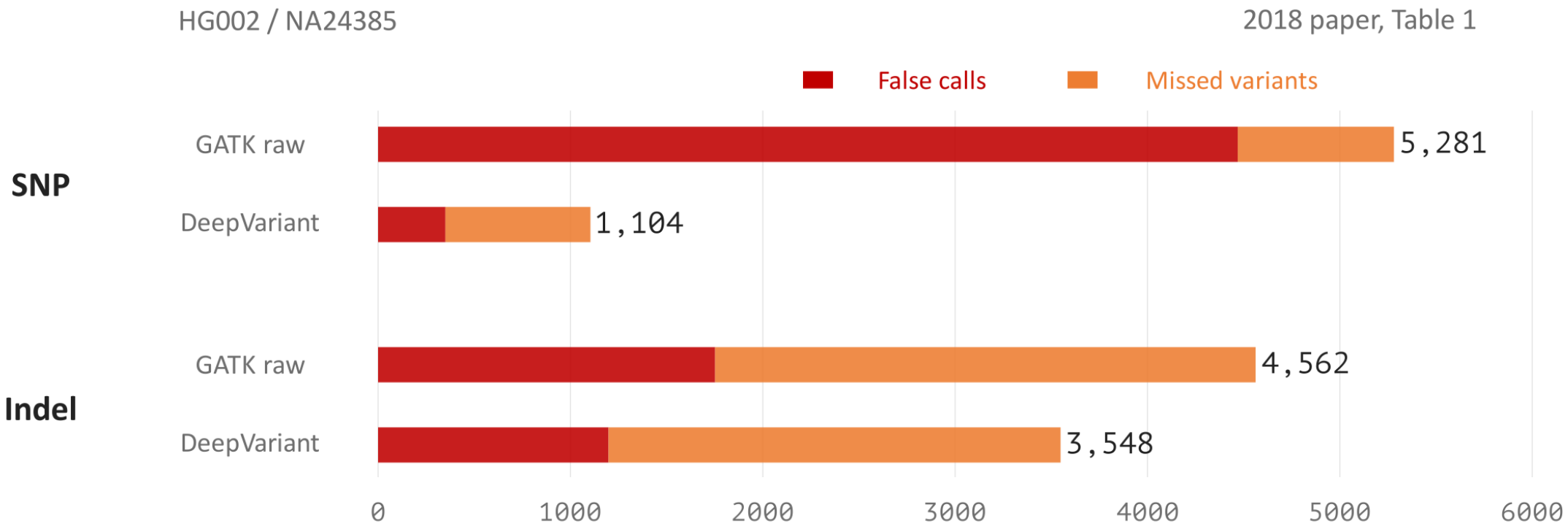
Precision

3 / 4 = 75%

Of the reported calls

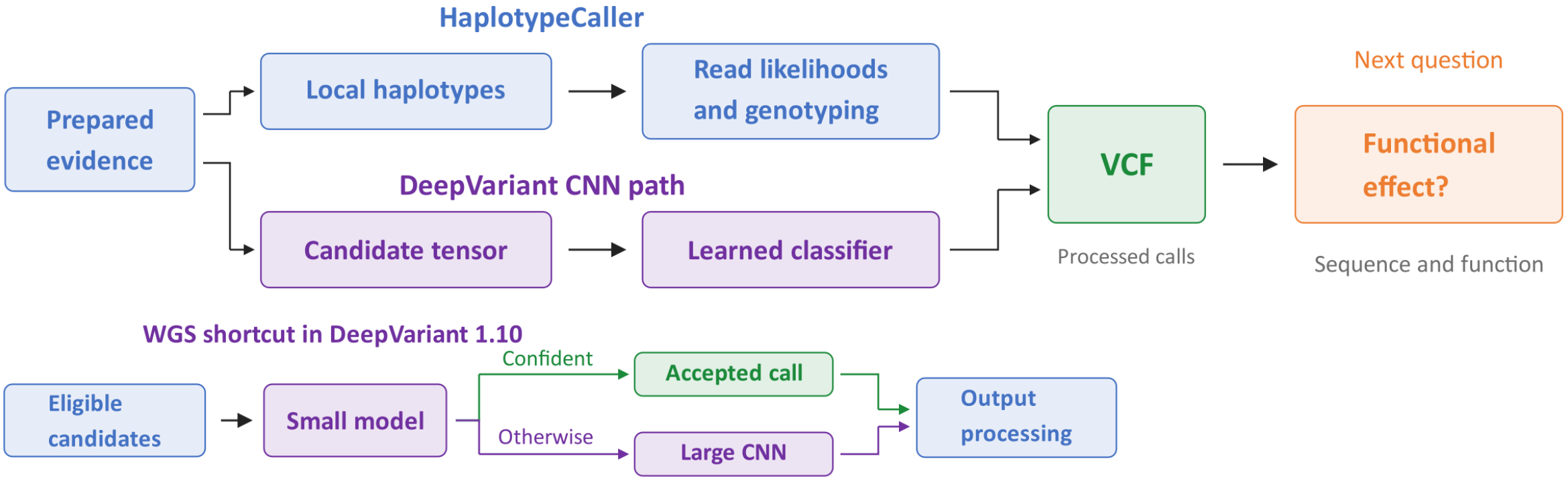
A Historical Result from the 2018 Paper

- Genome in a Bottle (GIAB) v3.2.2 benchmark, sample HG002
- SNP false calls + misses: 5,281 for GATK raw, 1,104 for DeepVariant
- Insertion/deletion (indel) errors: 4,562 for GATK raw, 3,548 for DeepVariant



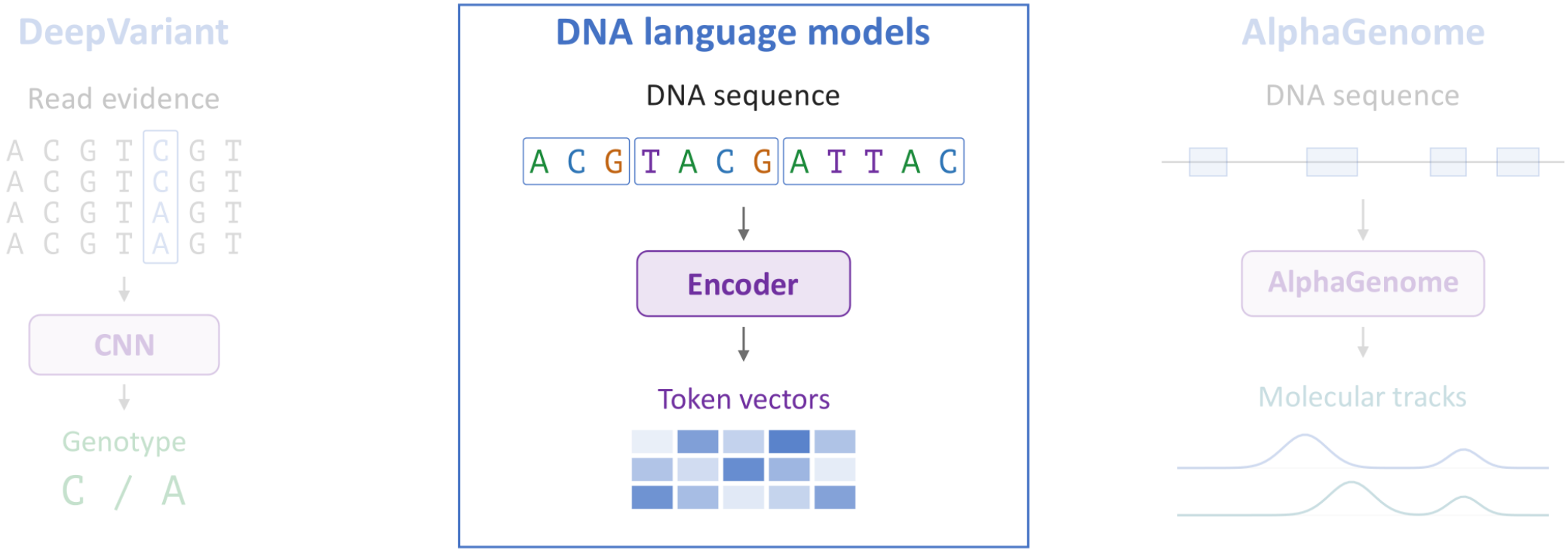
What Changes with a Learned Classifier?

- HaplotypeCaller uses likelihoods, DeepVariant learns a classifier
- DeepVariant 1.10: confident early calls for whole-genome sequencing (WGS)
- Variant calling finds DNA changes, functional models assess effects



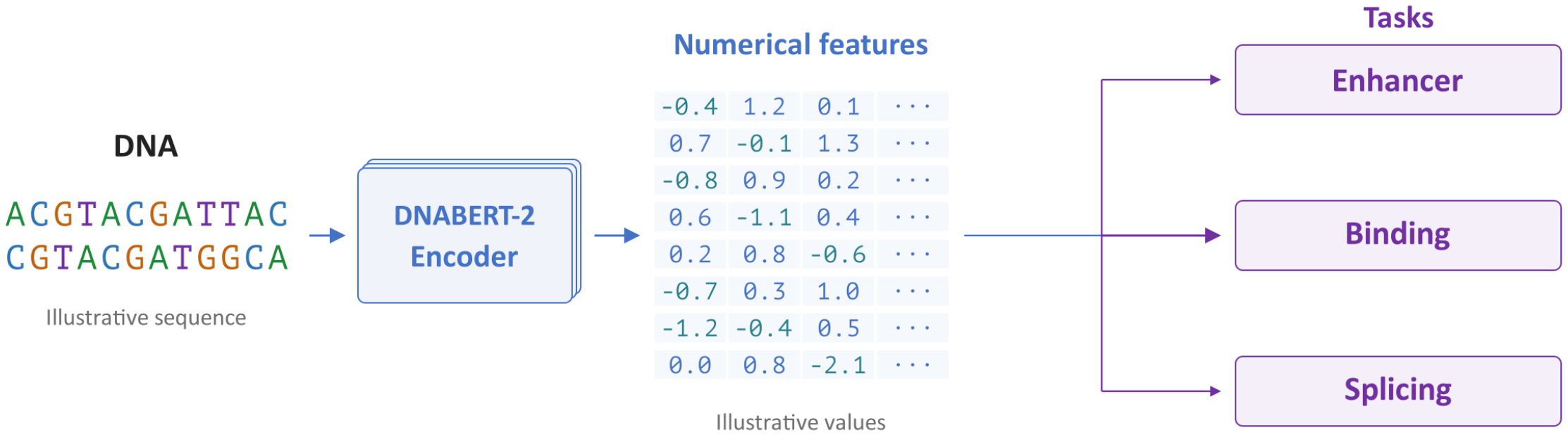
AI in Genomics

- DeepVariant: Identify variants from read evidence
- **DNA language models: Learn reusable sequence representations**
- AlphaGenome: Predict molecular activity and variant effects



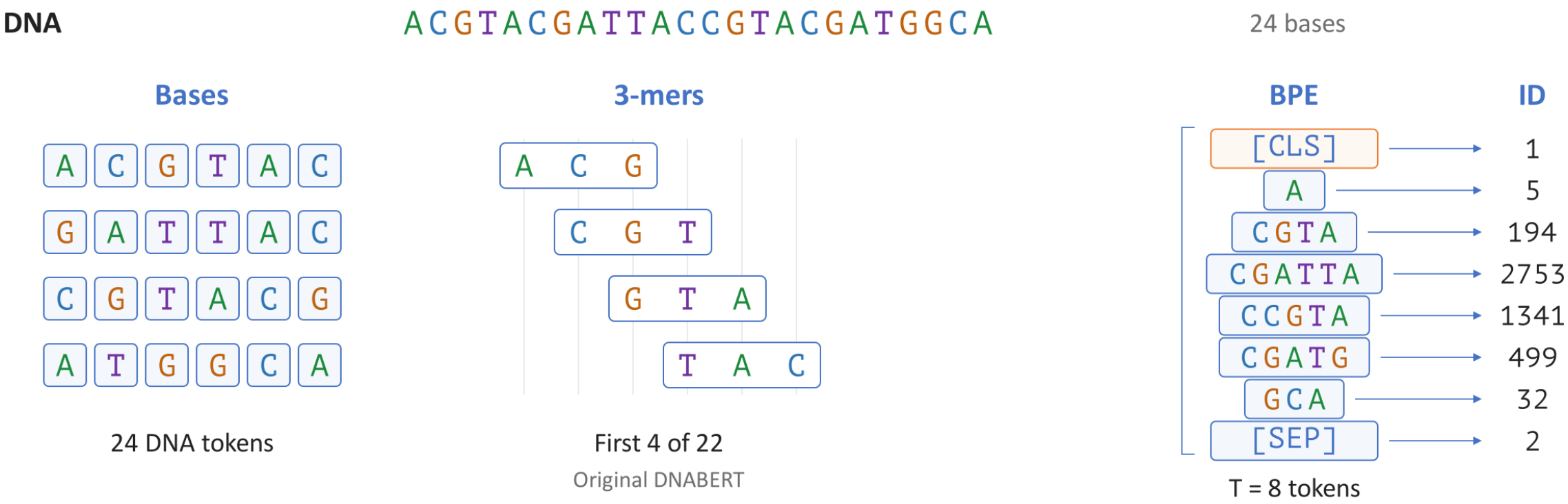
Learning DNA Representations

- DNA language models learn reusable representations of sequence
- These numerical features can support many prediction tasks



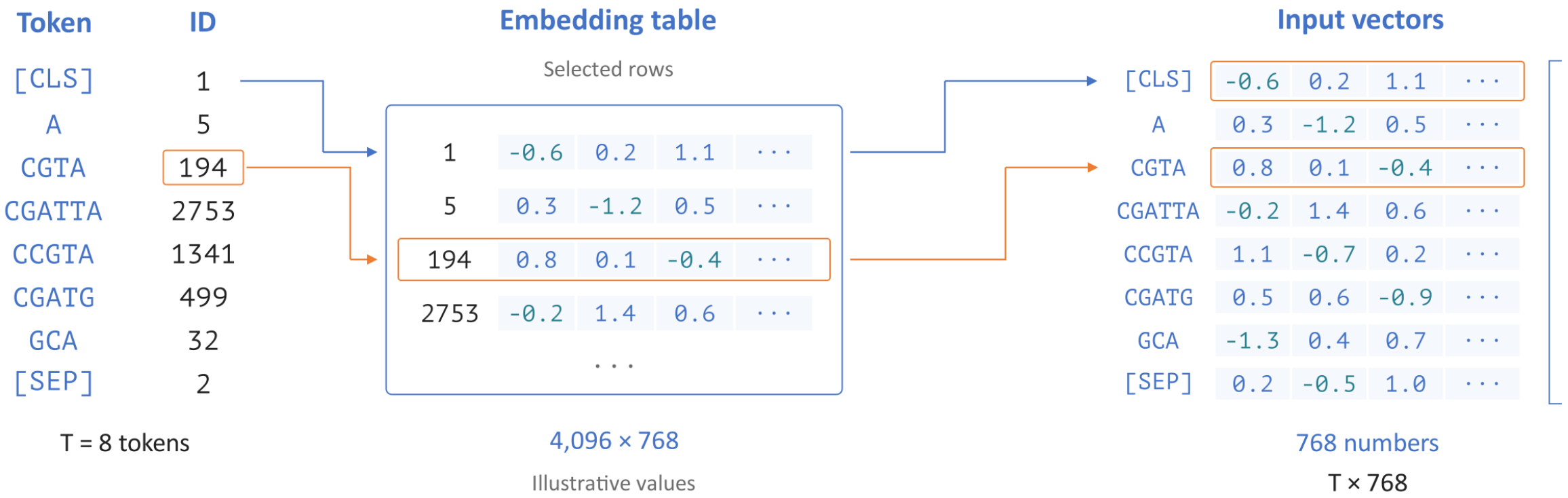
How DNA Becomes Tokens

- Tokenization divides DNA into pieces and assigns integer identifiers (IDs)
- Bases, k-mers, or byte-pair encoding (BPE) provide the tokens
- DNABERT-2 uses variable-length BPE tokens



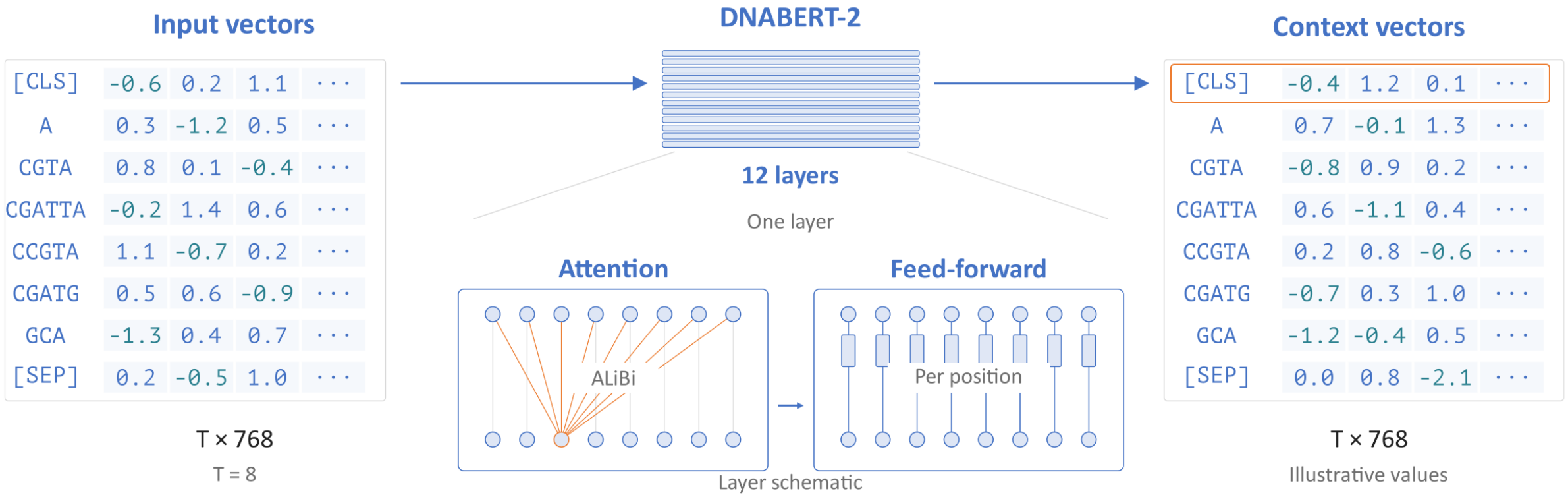
From Tokens to Vectors

- A learned embedding table maps each token ID to a vector
- DNABERT-2 uses 768 numbers per token, including special tokens
- These input vectors do not yet incorporate the surrounding sequence



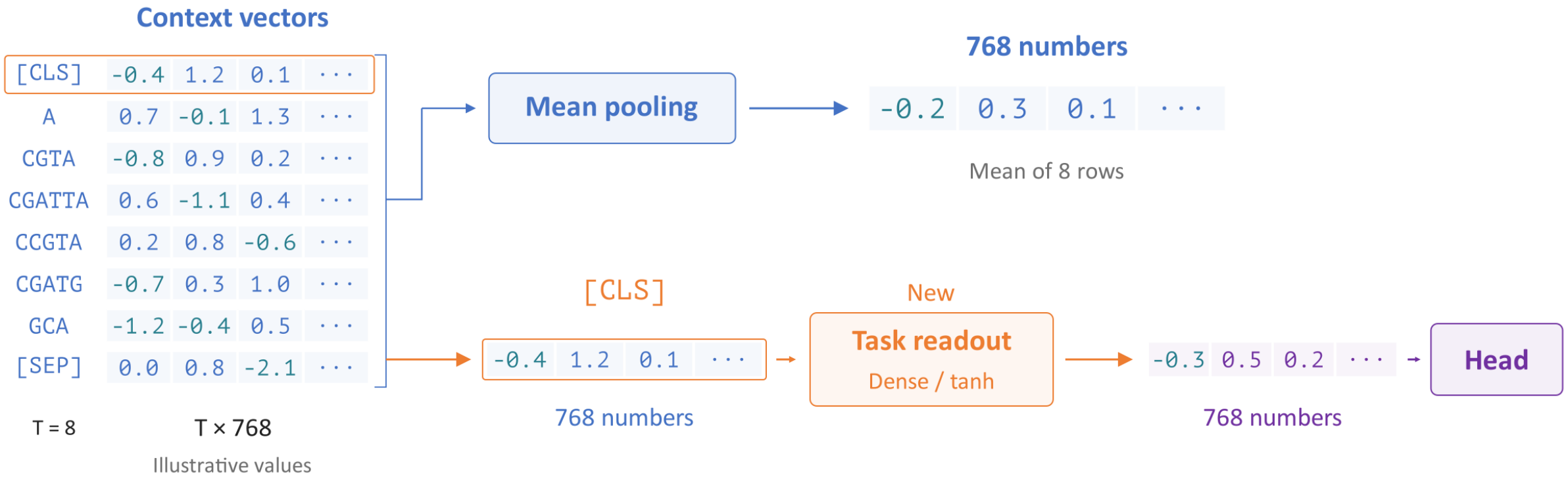
The Transformer Encoder

- Self-attention combines information from different token positions
- Each layer also transforms the features at every position
- The encoder returns one contextual vector per token



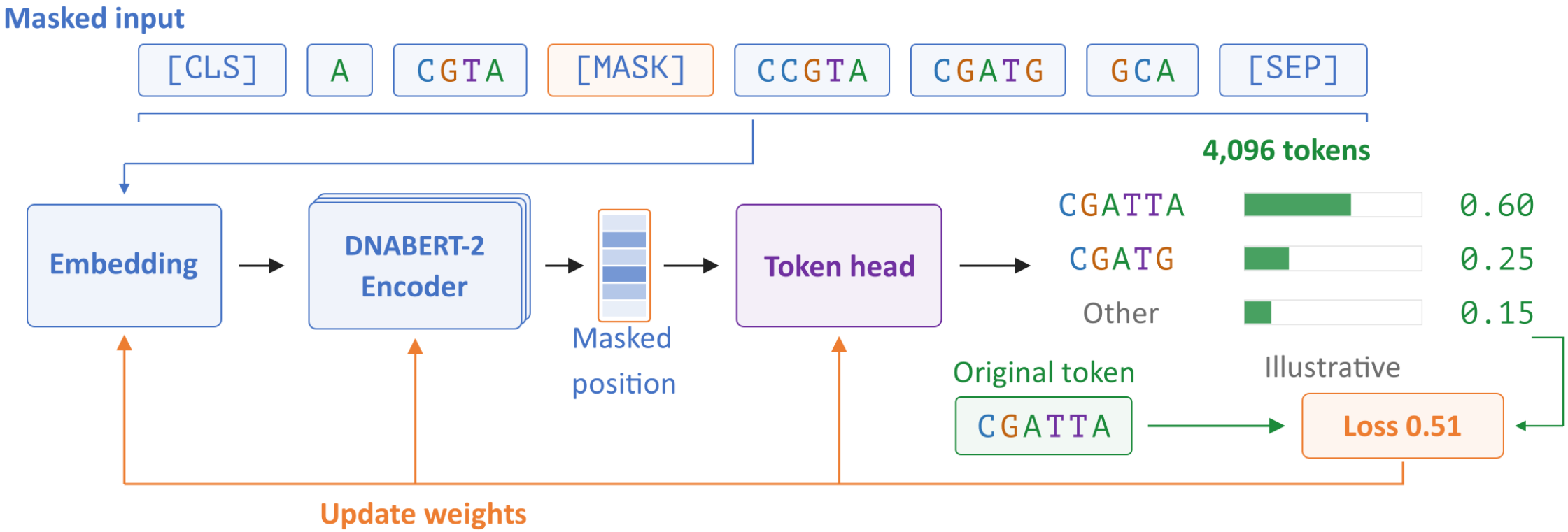
A Vector for a Sequence

- Summarize token vectors to represent the whole sequence
- Mean pooling averages each feature across the token positions
- DNABERT-2 uses the classification token ([CLS]) for its readout



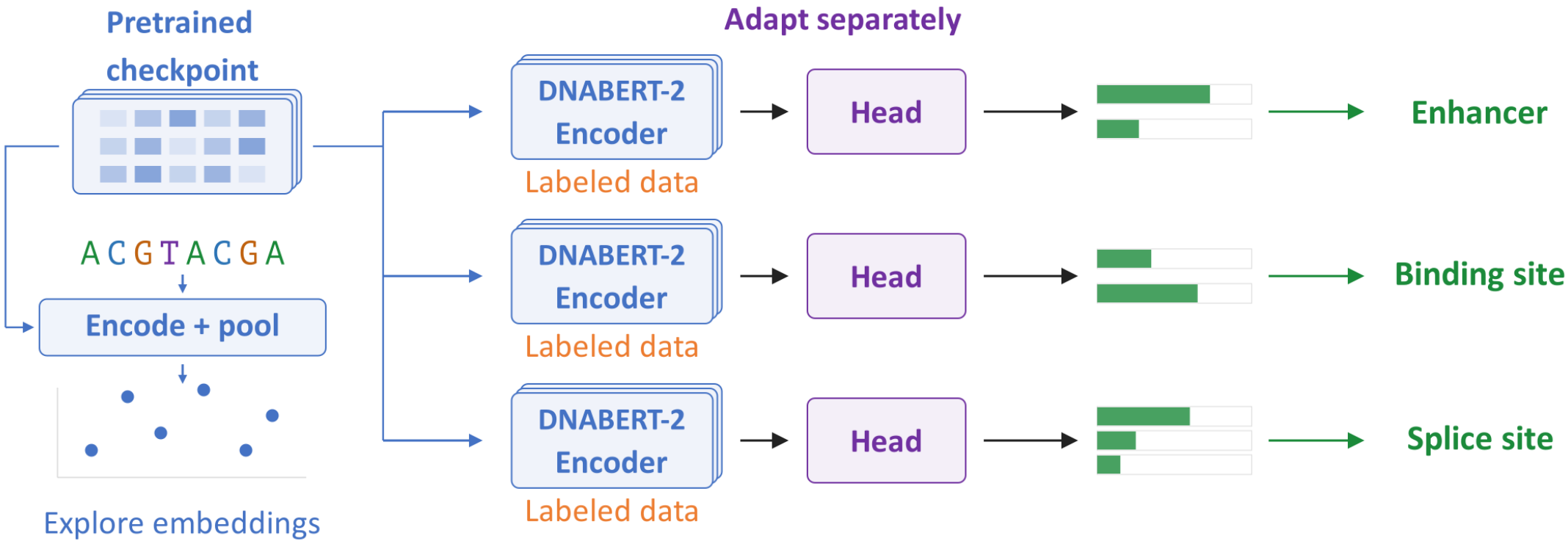
Pretraining on DNA

- Pretraining predicts masked tokens using surrounding DNA
- The target comes from the original sequence, without task labels
- Prediction errors update the embedding table, encoder, and output head



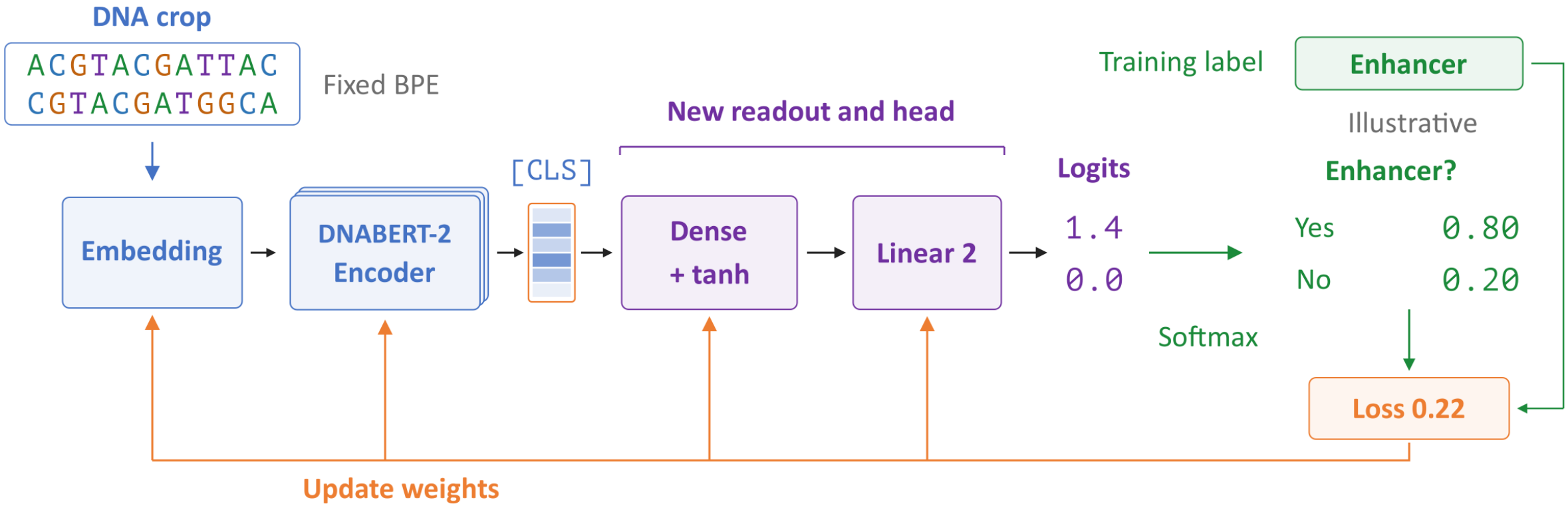
One Model, Many Tasks

- Reuse learned features to explore sequences or train prediction models
- Examples include enhancer, binding-site, and splice-site prediction
- Each prediction task needs an appropriate head and labeled examples



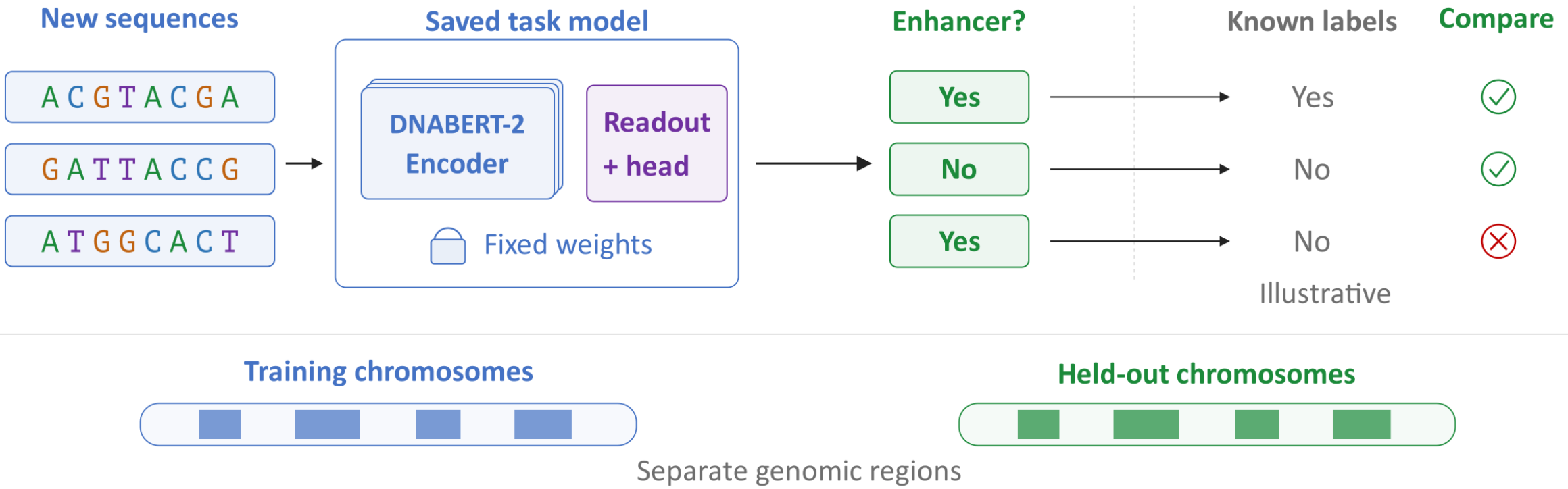
Full Fine-Tuning

- Example: classify labeled DNA regions as enhancer or non-enhancer
- Add a two-class head and compare predictions with the labels
- Full fine-tuning updates the pretrained model and the new head



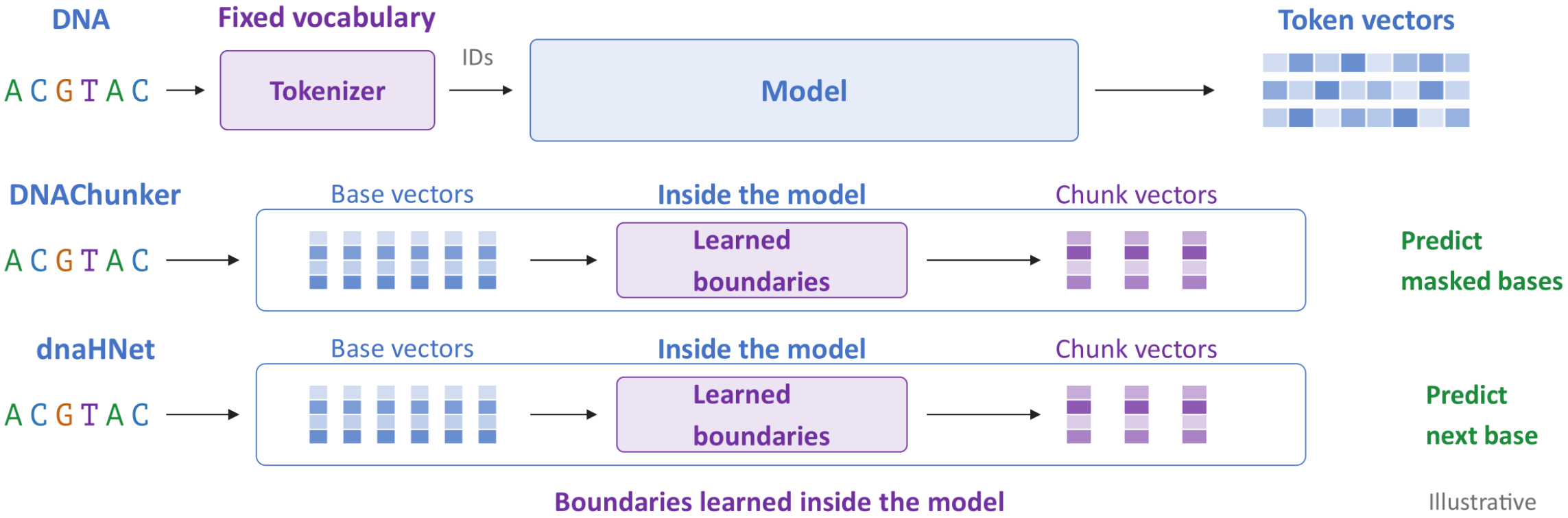
Prediction and Evaluation

- For a new sequence, run the fine-tuned model without updating it
- Evaluate on held-out sequences with independent labels
- Related or overlapping sequences can make evaluation too optimistic



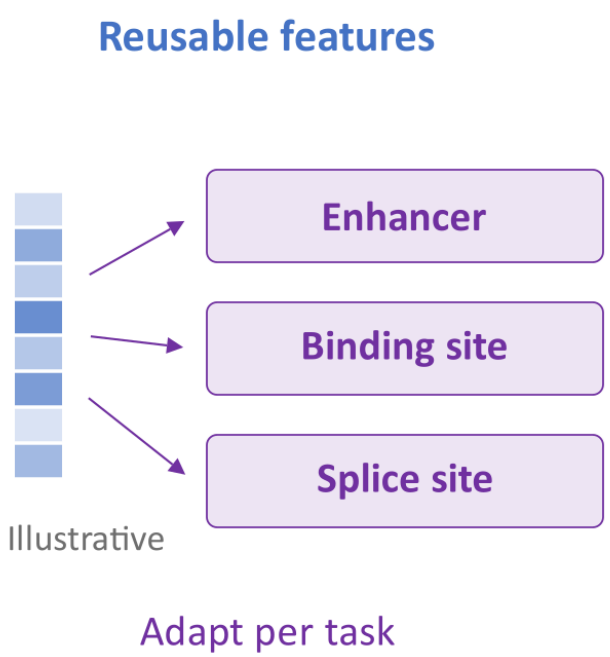
Why Tokenization Matters

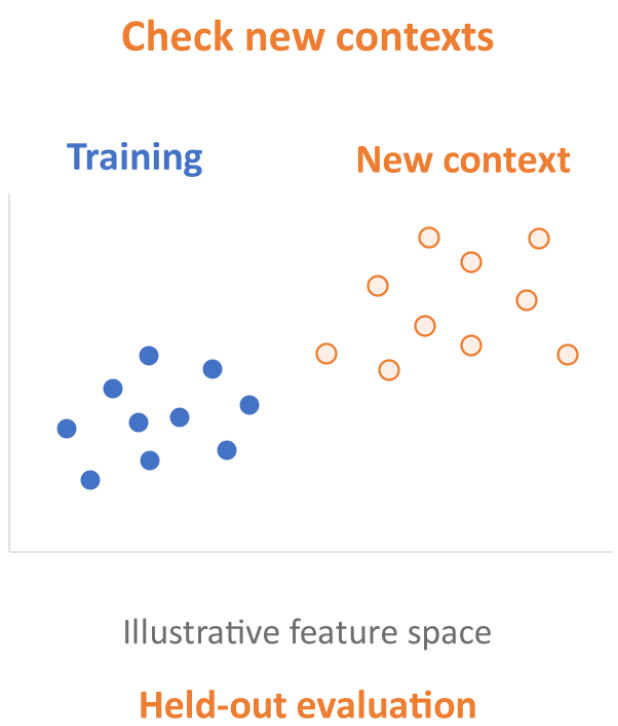
- The choice of DNA tokens changes what the model sees
- DNAChunker learns boundaries during masked-sequence pretraining
- dnaHNet learns hierarchical boundaries in an autoregressive model

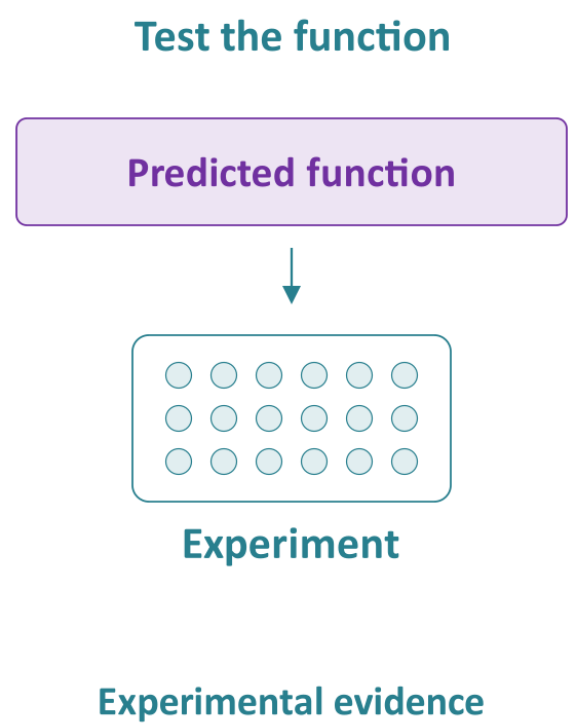


Benefits and Limitations

- Pretraining can provide useful starting features for several tasks
- Performance depends on training data, task, and biological context
- A useful embedding is not automatically a biological explanation

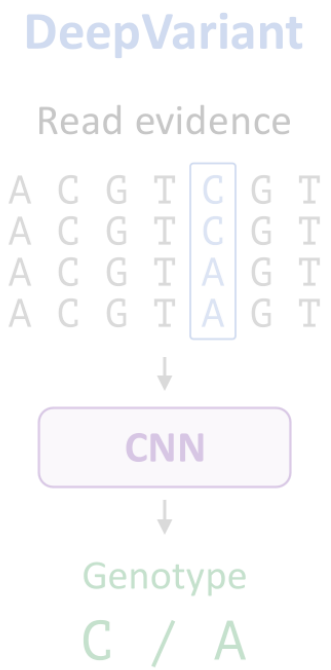


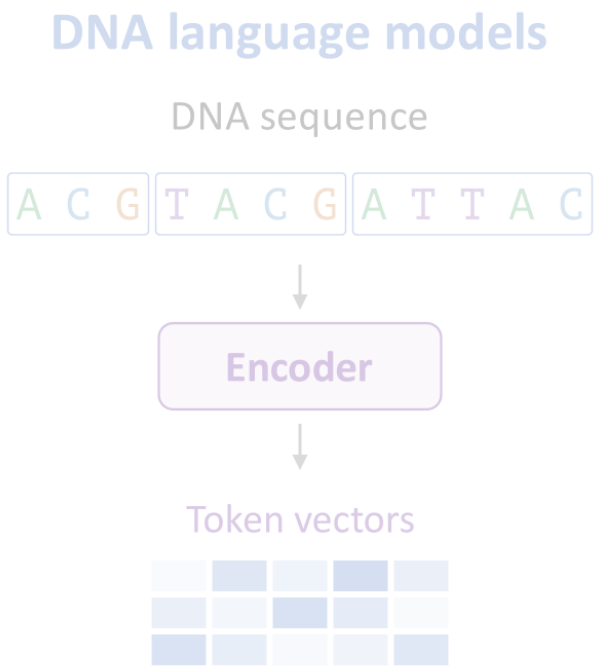


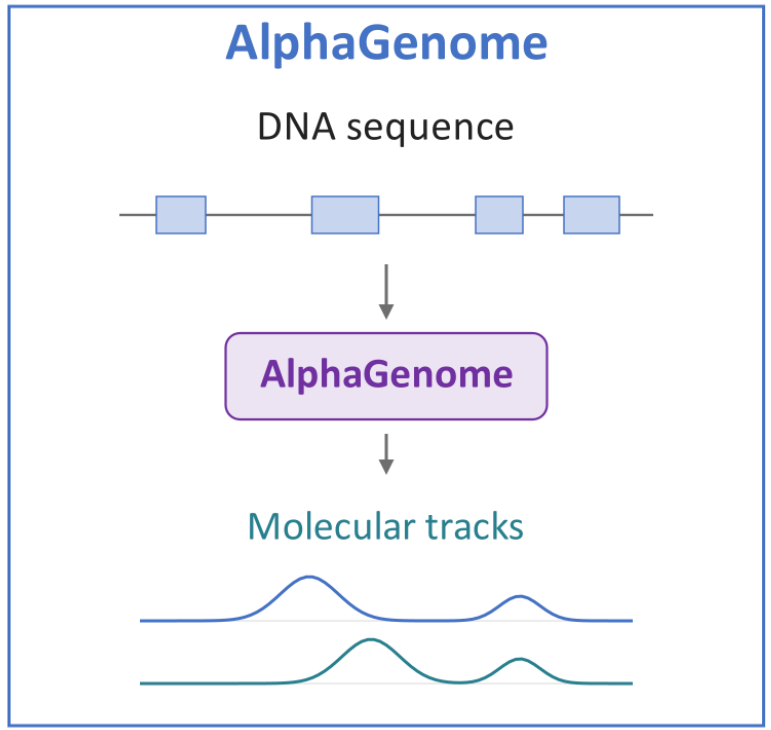


AI in Genomics

- DeepVariant: Identify variants from read evidence
- DNA language models: Learn reusable sequence representations
- **AlphaGenome: Predict molecular activity and variant effects**

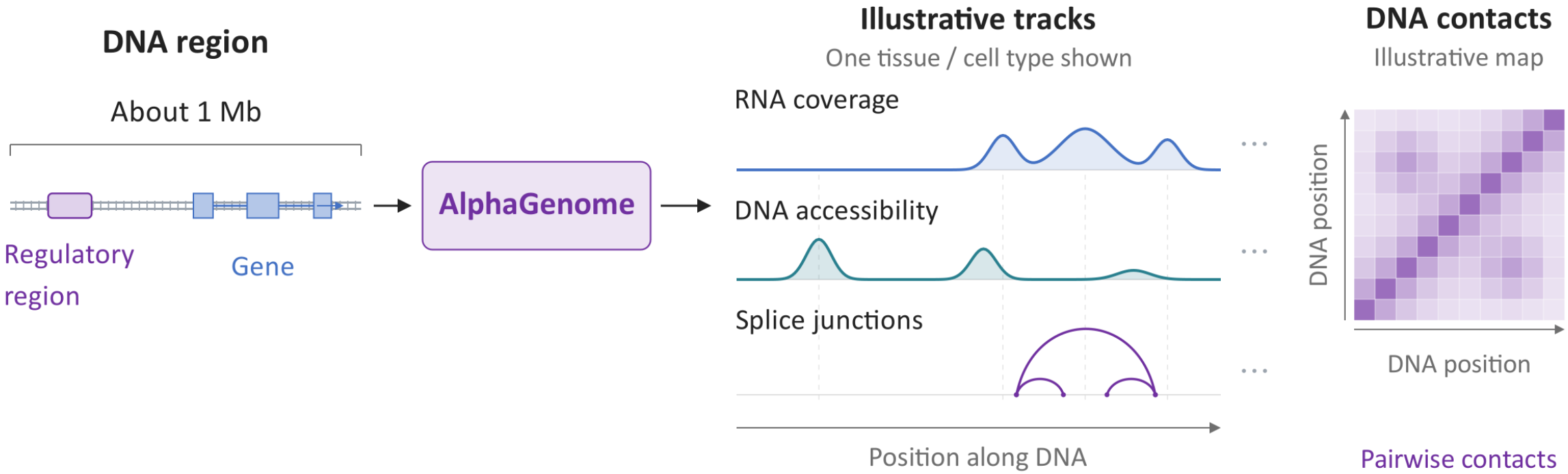






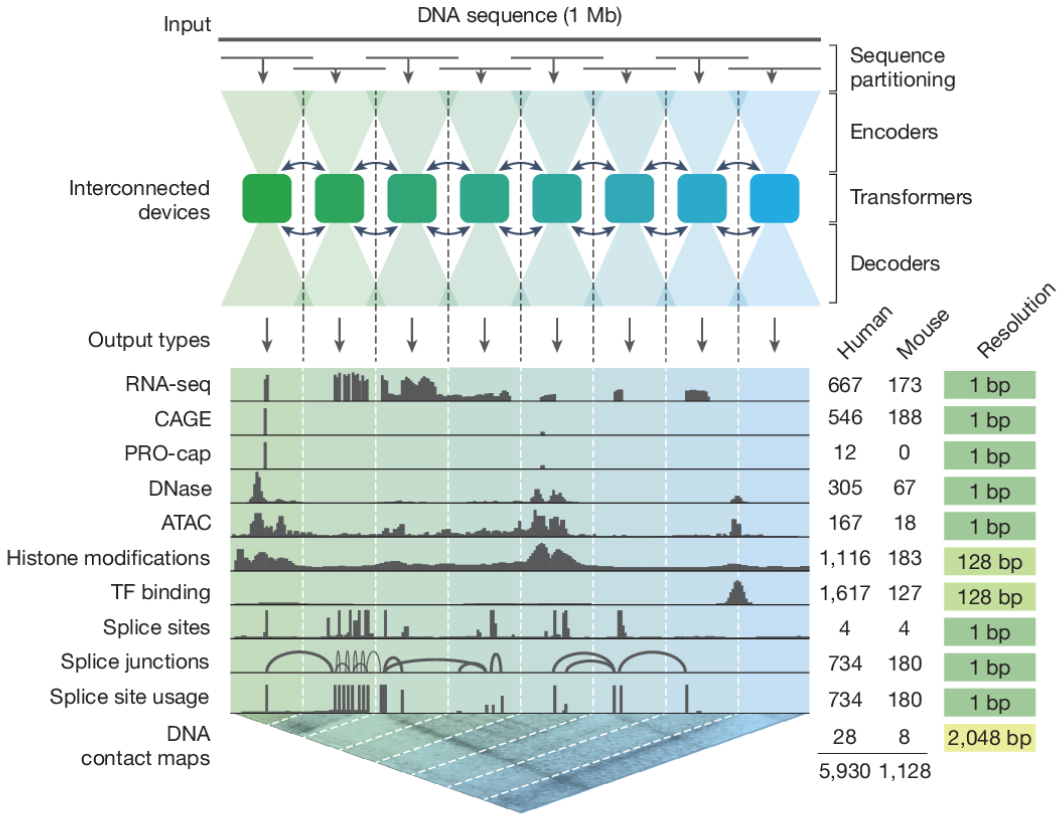
AlphaGenome: Sequence to Function

- Predict molecular measurements from about one million DNA bases
- Genome tracks describe activity along DNA in a tissue or cell type
- Outputs include expression, splicing, accessibility, and DNA contacts



AlphaGenome: Sequence to Function

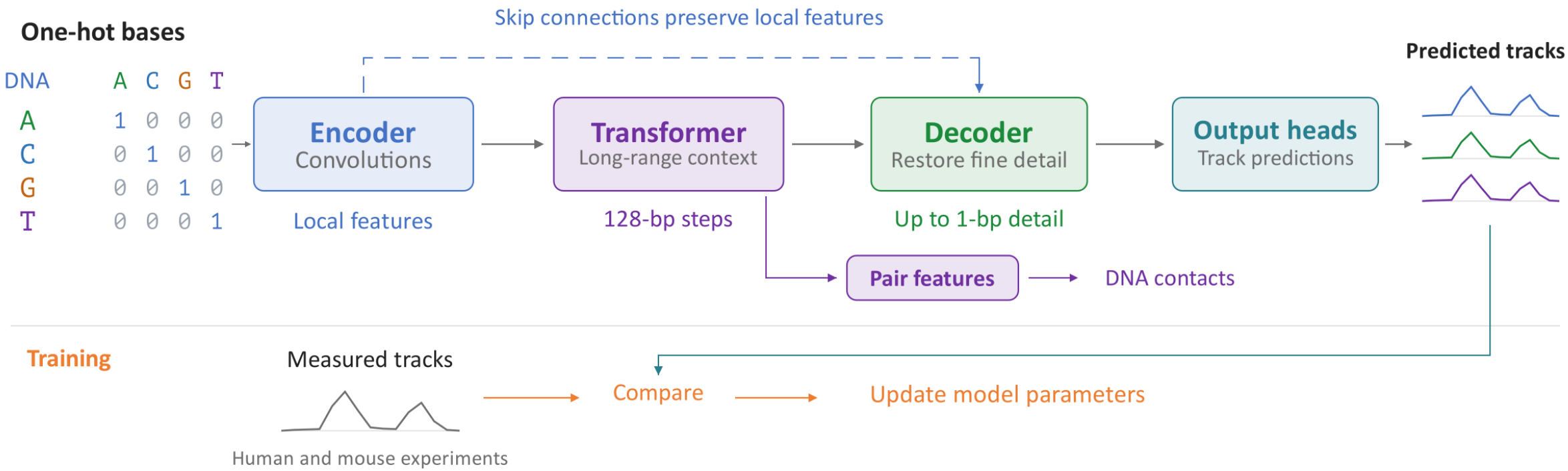
- Predict molecular measurements from about one million DNA bases
- Genome tracks describe activity along DNA in a tissue or cell type
- Outputs include expression, splicing, accessibility, and DNA contacts



https://www.nature.com/articles/s41586-025-10014-0

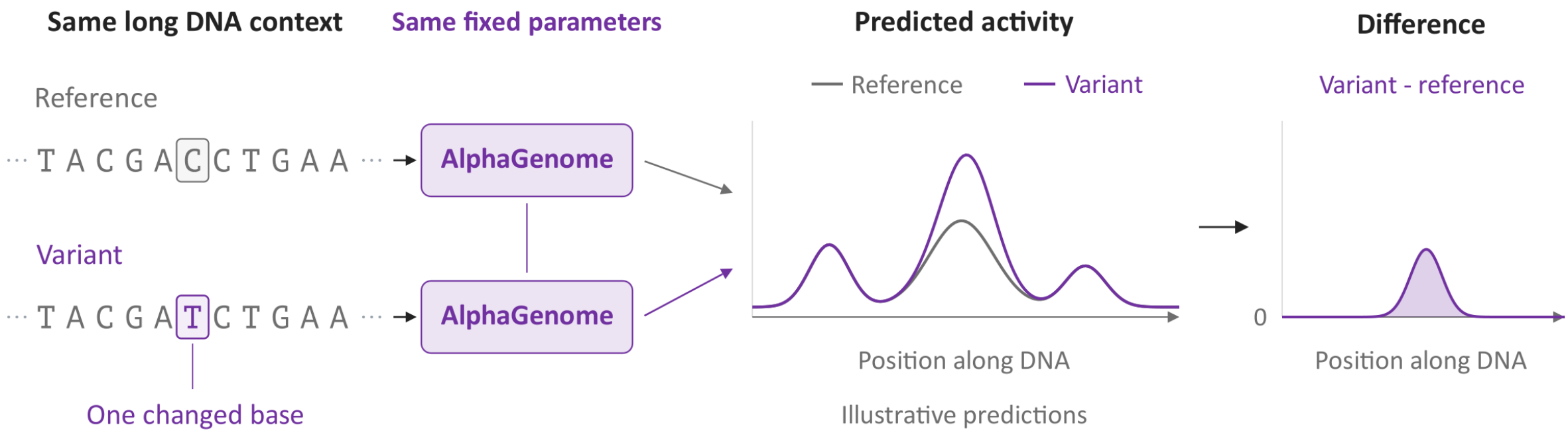
How AlphaGenome Works

- A base-by-base encoding feeds convolutional and transformer layers
- The decoder combines local detail with long-range context for prediction
- Training fits human and mouse experimental tracks



Predicting a Variant's Effects

- Change the sequence and keep the trained model fixed
- Predict the same molecular tracks for reference and variant
- Compare the predictions to estimate the molecular effect



An Insertion near TAL1

- A two-base insertion near TAL1 creates a binding motif for the MYB protein
- The model predicts increased accessibility near the insertion
- It also predicts increased TAL1 expression about 7,500 bases away

Published example

Reference

C

↓

Variant

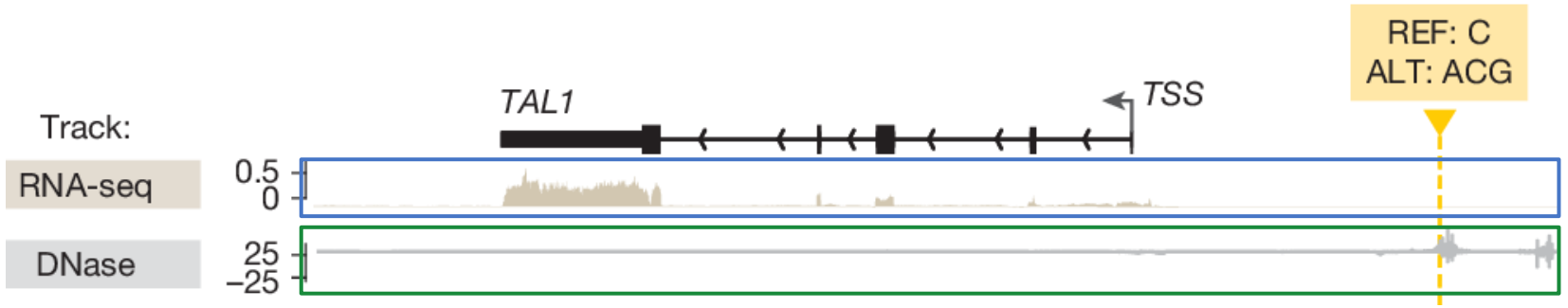
ACG

+2 bases

New MYB motif

Predicted change: variant – reference

Published tracks • Figure 6b



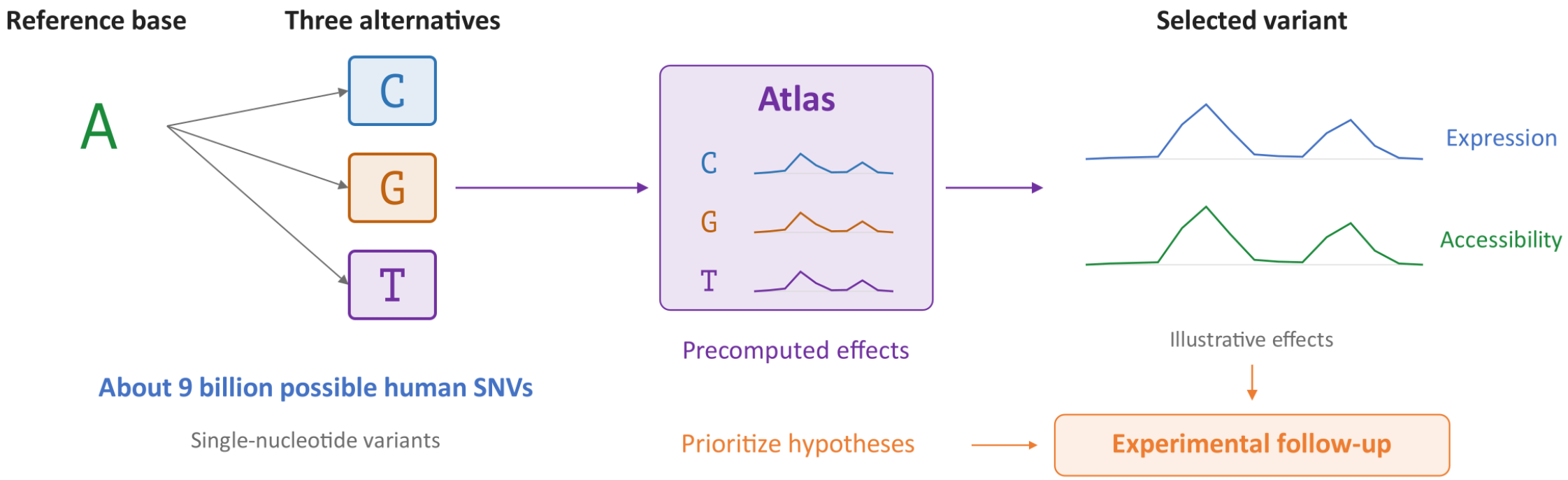
Accessibility increases near the insertion

TAL1 expression increases about 7.5 kb away

CD34+ common myeloid progenitors

Using AlphaGenome Predictions

- Atlas stores about 9 billion human single-nucleotide variant (SNV) predictions
- Look up a variant to inspect predicted effects and prioritize follow-up
- Cell-type coverage and distant regulatory effects remain limitations



Questions?

Next Week

- Hardware Accelerators for Genomics

CMSC 829C: Algorithms, AI, and Hardware for Biological Data

Lecture 6: AI in Genomics

Fall 2026

September 17, 2026

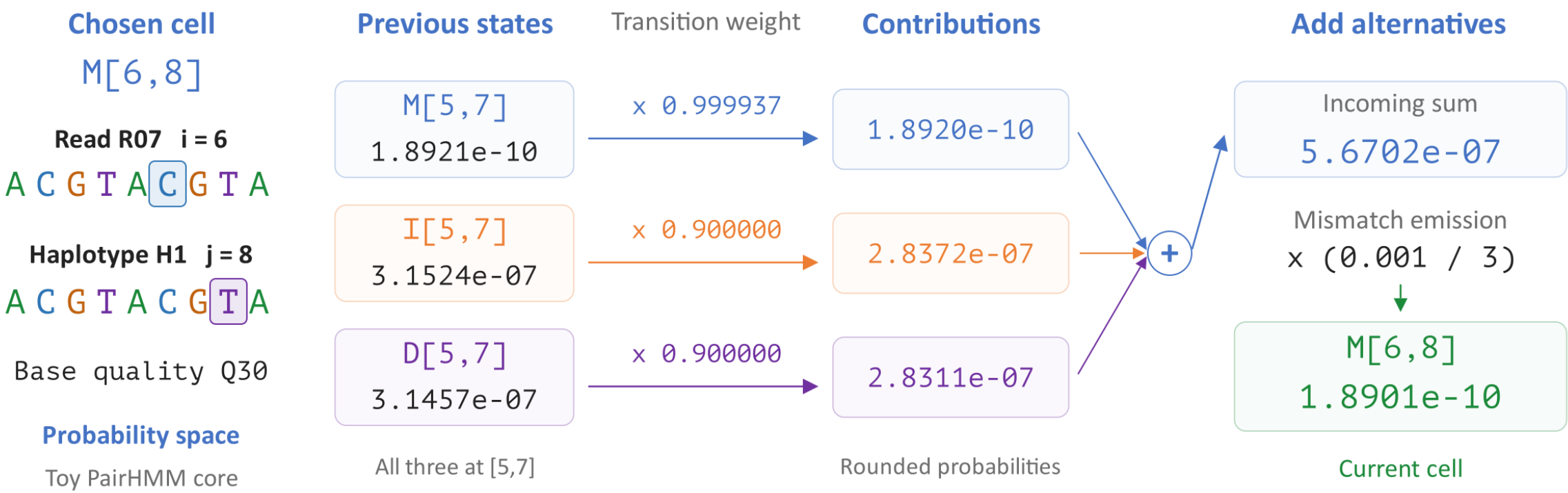
Can Firtina | canfirtina@gmail.com | cs.umd.edu/~firtina

Assistant Professor of Computer Science at University of Maryland (UMD)

UNIVERSITY OF MARYLAND

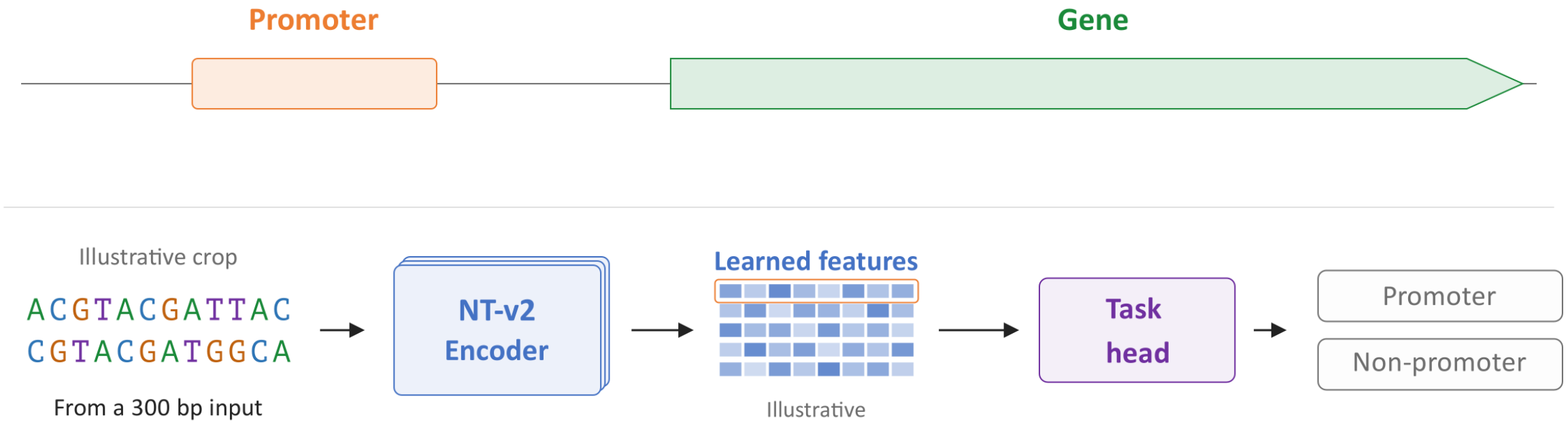
Fill One PairHMM Cell

- Each dynamic programming (DP) cell sums probabilities of incoming paths
- Weight predecessor states and add their contributions
- Apply the base emission probability to obtain the current cell



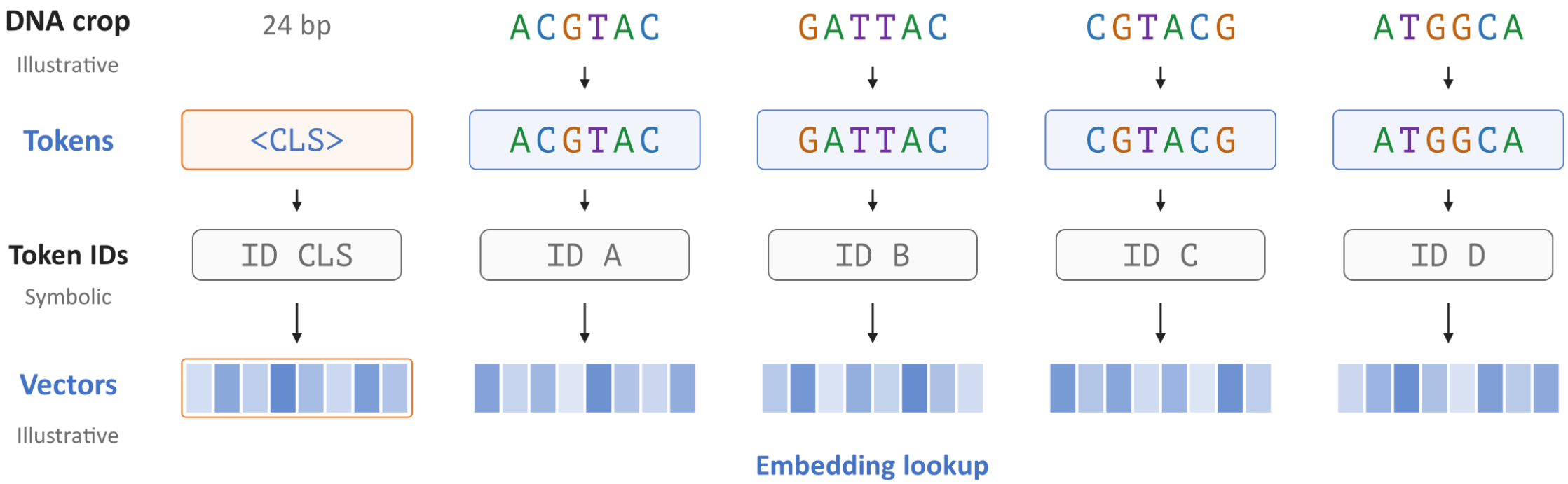
Learning from DNA

- Identify promoters, DNA regions that help start transcription
- Learn reusable features from genomic sequences
- Adapt the model to a task using labeled examples



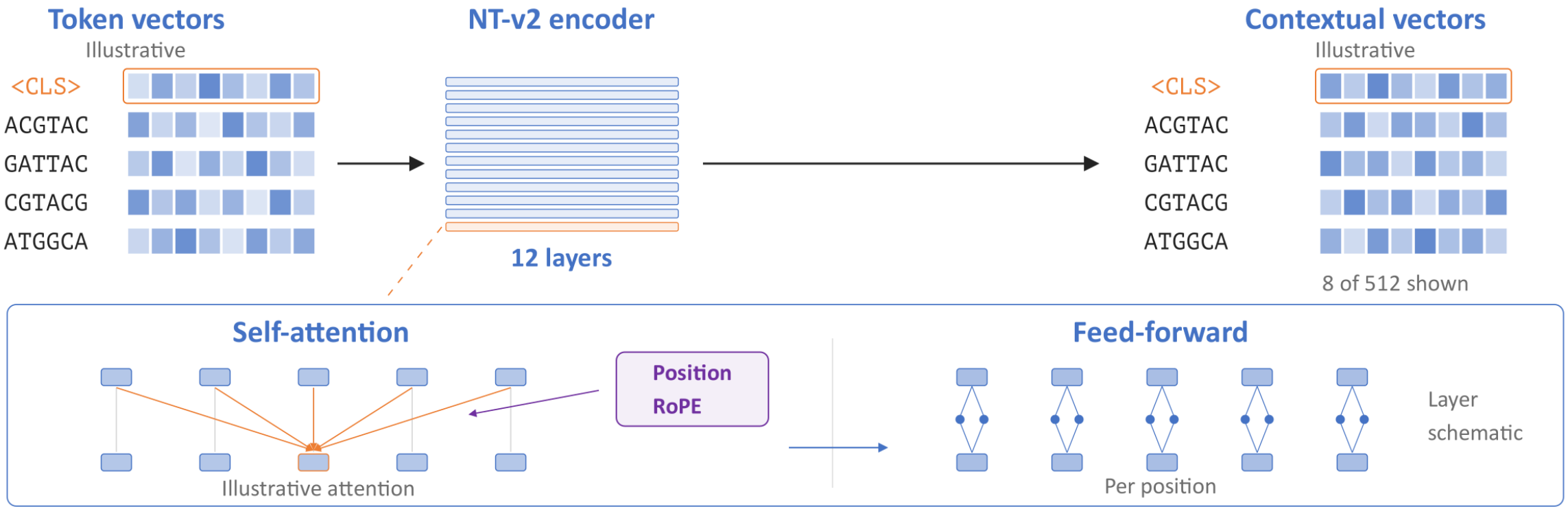
From DNA to Model Inputs

- Group DNA into nonoverlapping six-base tokens
- Prepend <CLS>, a special token for classification
- Convert token IDs into learned embedding vectors



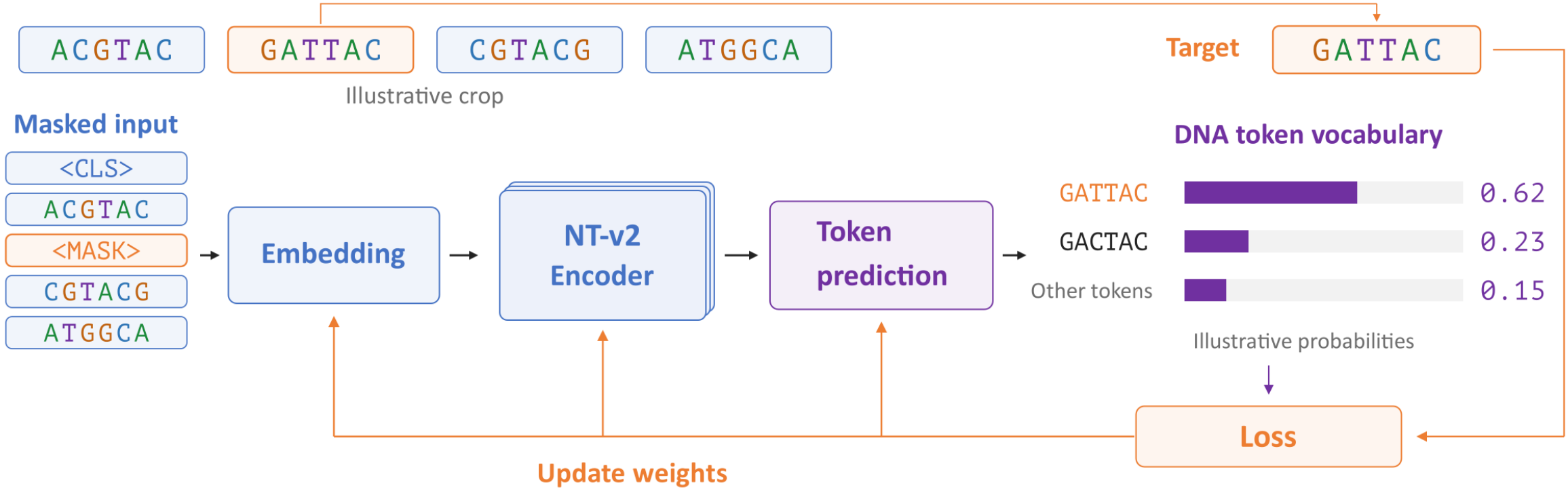
The Transformer Encoder

- Repeated Transformer layers build sequence context
- Self-attention combines information across positions
- Feed-forward layers transform each position's vector



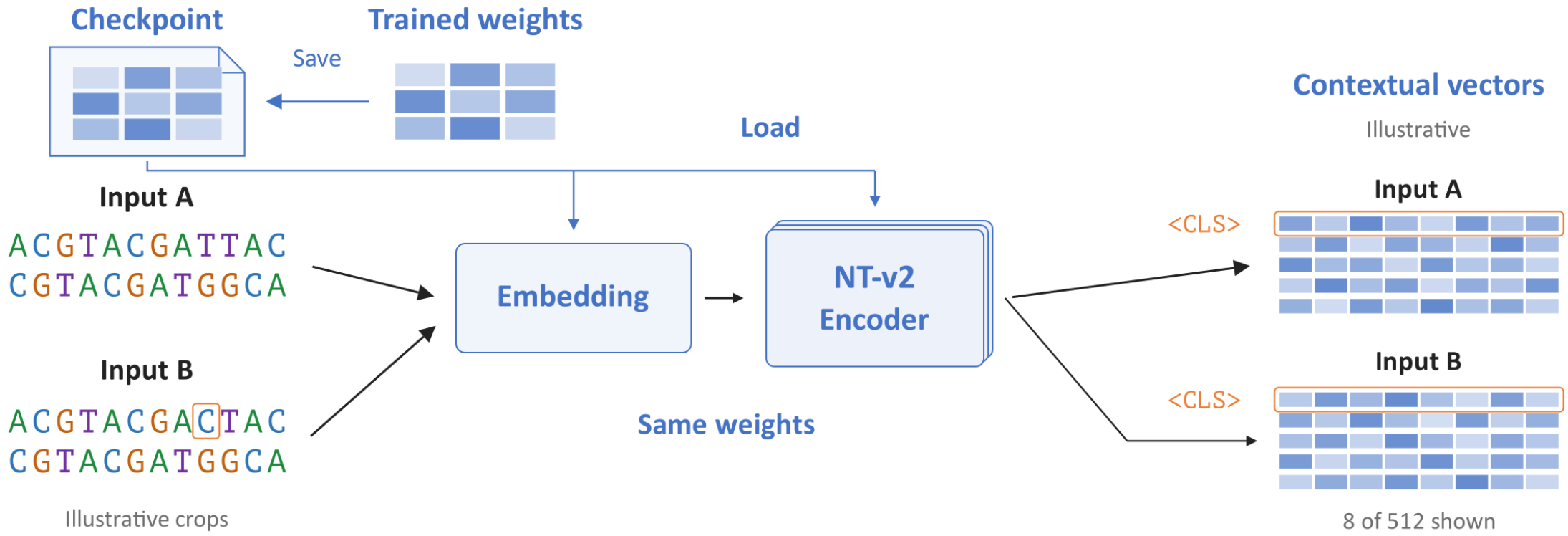
Pretraining on DNA

- The original DNA supplies the training targets
- Predict hidden tokens from the surrounding sequence
- Prediction error guides updates to model weights



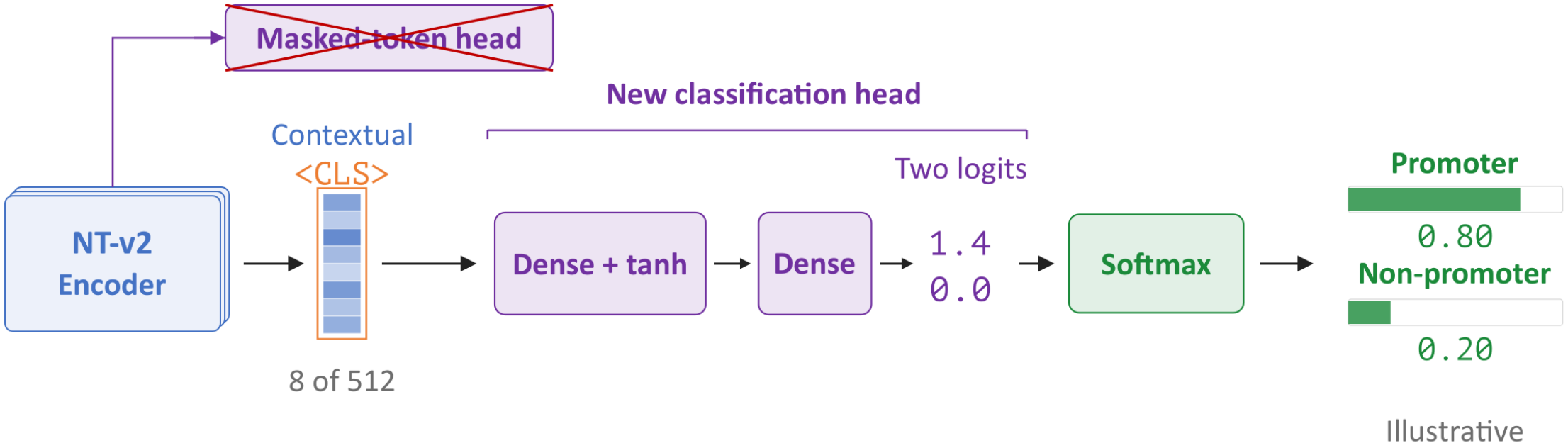
Reusing a Pretrained Model

- A checkpoint stores the learned model weights
- Load those weights to process a new DNA sequence
- Compute contextual embeddings separately for each input



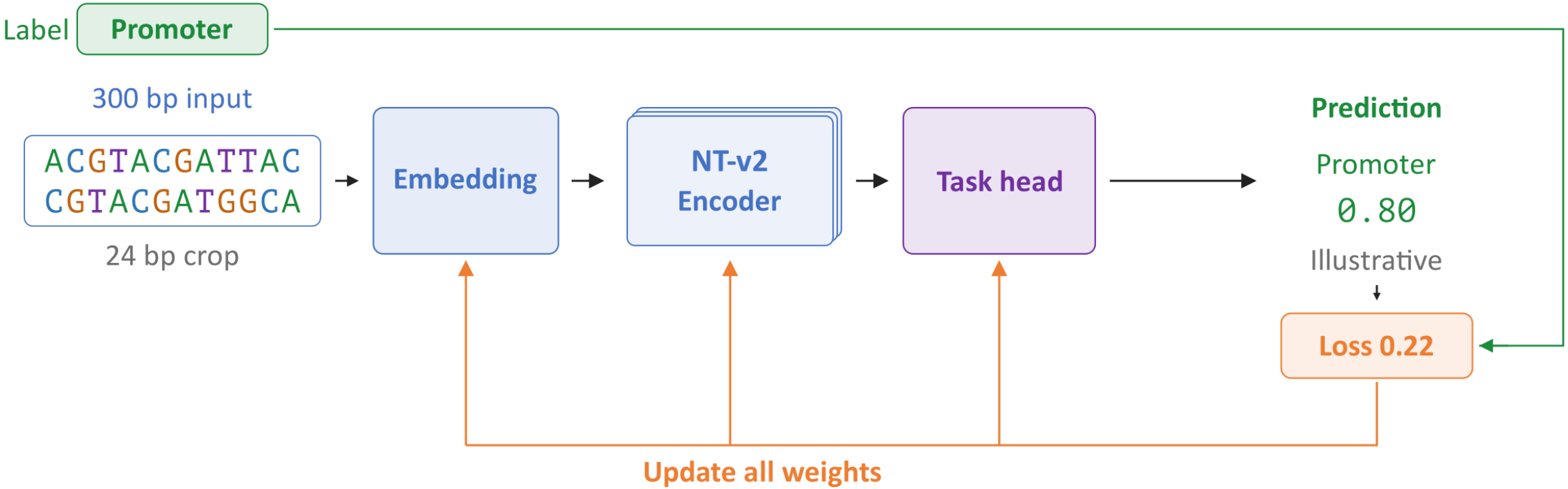
Promoter Classification

- Use the final contextual <CLS> vector
- A new classification head produces two class scores
- Softmax converts the scores into class probabilities



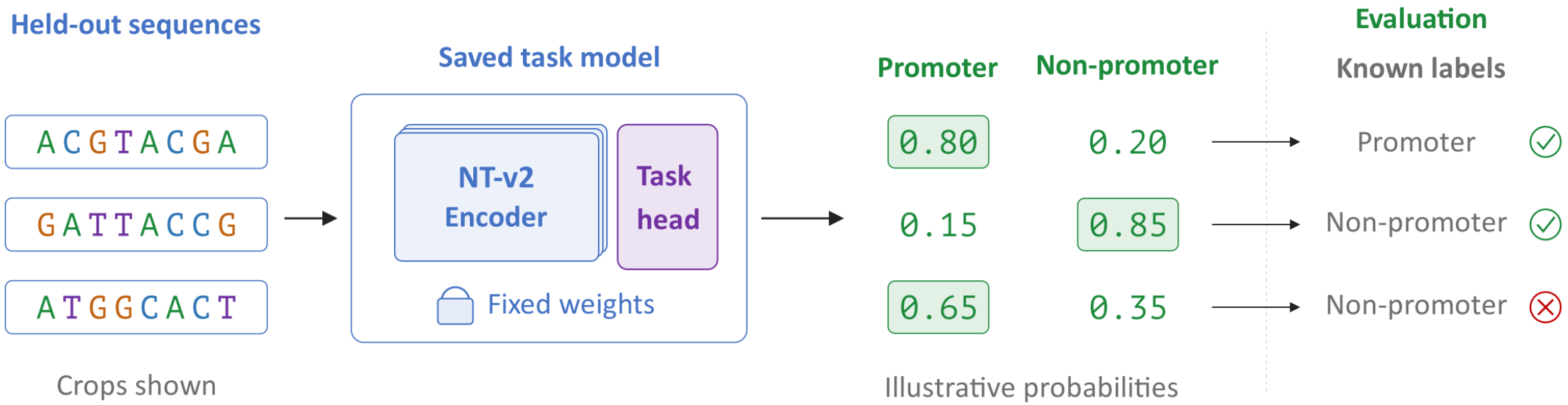
Full Fine-Tuning

- Train with sequences labeled promoter or non-promoter
- Compare each prediction with its label to compute loss
- Update the pretrained model and the new head together



Prediction and Evaluation

- Evaluate on sequences excluded from fine-tuning
- Inference uses trained weights without updating them
- Compare predictions with held-out labels



Predicting the Next Base

- NT-v2 predicts masked tokens using both sides
- Evo 2 predicts the next base from the preceding sequence
- Append a predicted base and repeat to generate DNA