

Input sequence generations

Tokenization methods

[CLS] CTCGCT GTTGTA TGCCGA A A [SEP]

[CLS] GGGCTC GGTACC GTATCA G T [SEP]

[CLS] CGGCTC CCGCAT GACTGG A T [SEP]

The non-overlapping 6-mer

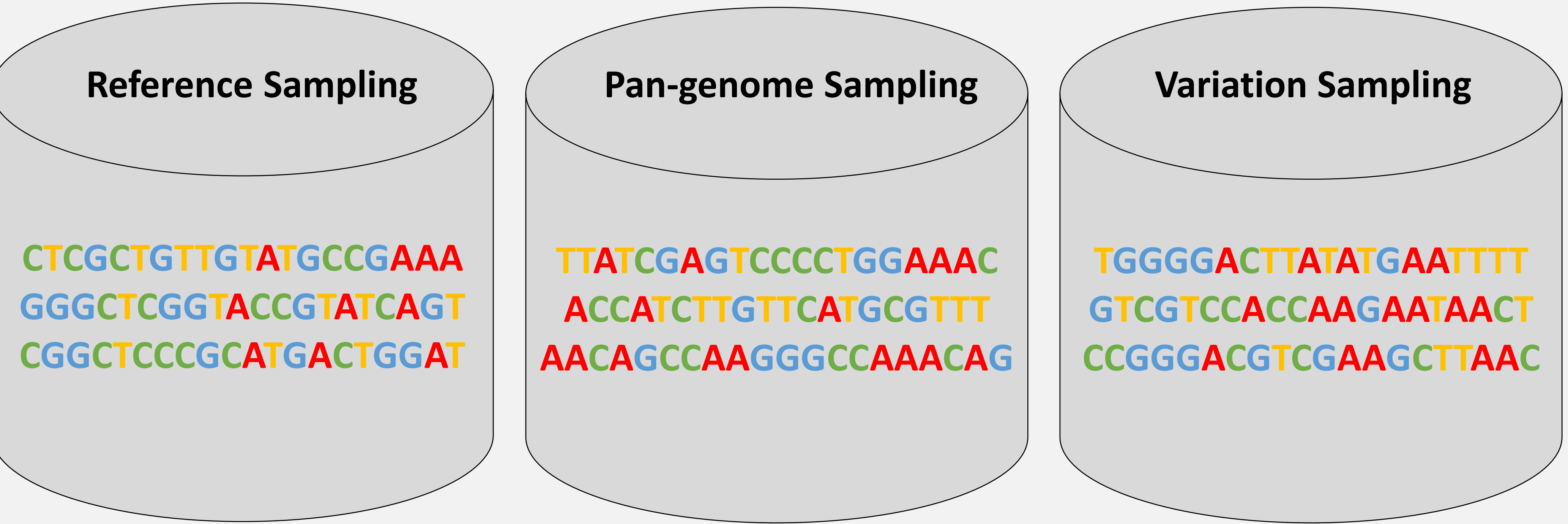
[CLS] CTCG CTGT TGTAT GCCGA [SEP] [PAD]

[CLS] GGGCT CGGTA CCGT ATCA [SEP] [PAD]

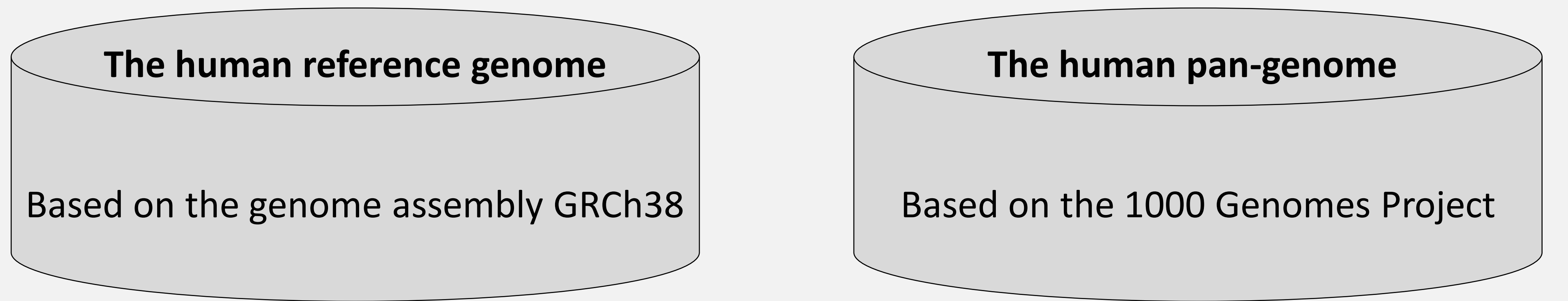
[CLS] CGG CTCCC GCAT GACT GG [SEP]

BPE

Sampling strategies



Datasets

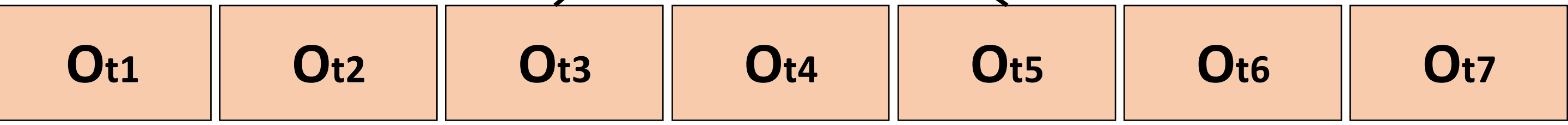


Output Sequence

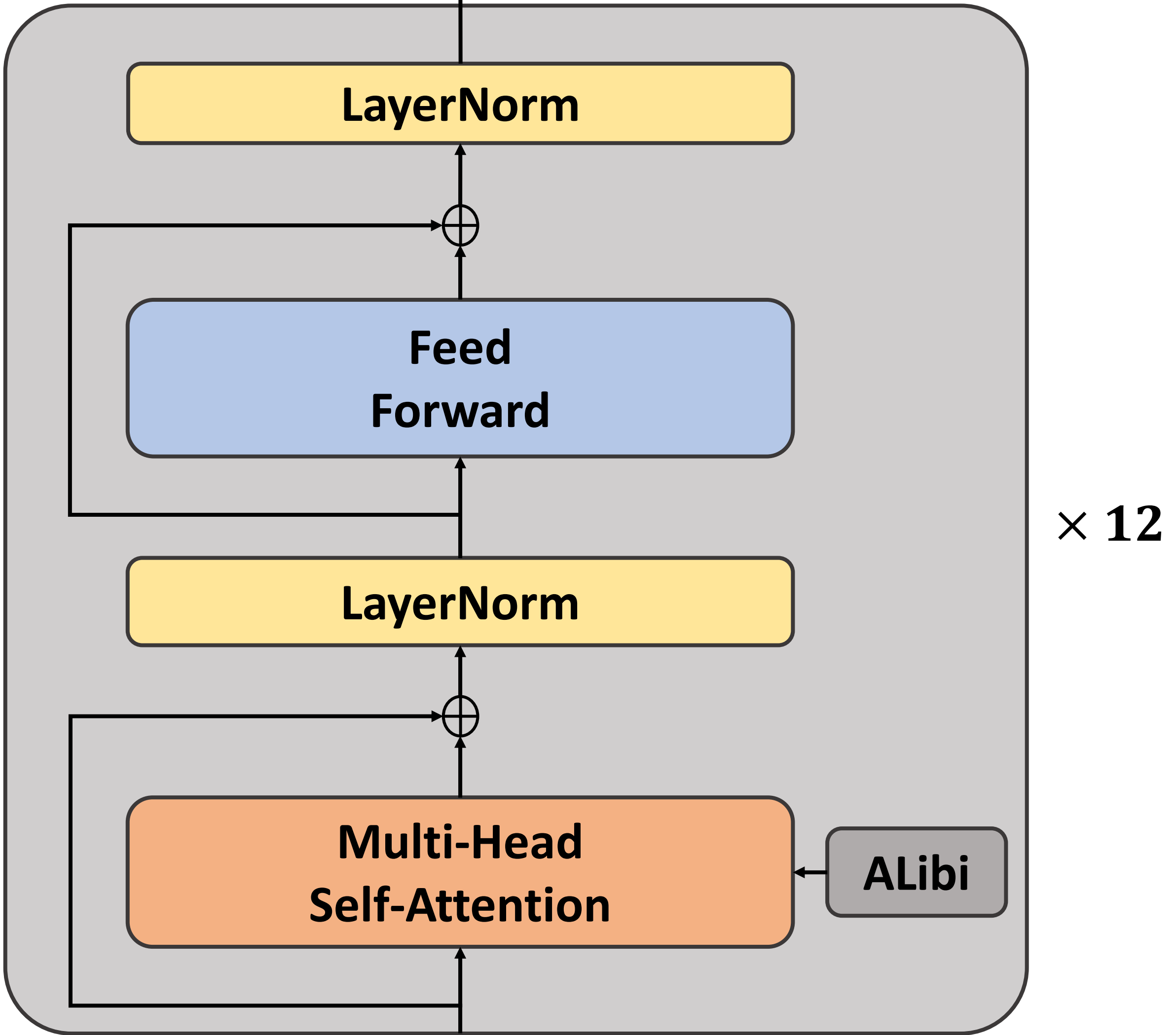
[CLS] CTCG CTGT TGTAT GCCGA [SEP] [PAD]

Classification Layer

Last Hidden State



DNABERT-2



Token Embedding



Masked Input Sequence

[CLS] CTCG [MASK] TGTAT [MASK] [SEP] [PAD]

Input Sequence

[CLS] CTCG CTGT TGTAT GCCGA [SEP] [PAD]