C.La.P.: Leveraging ATAC-seq with transformers for context-sensitive cis-regulatory modeling

P16.075.C

Panos Firbas Nisantzis✉, Carolina Gonçalves, Gonzalo G. de Polavieja

Champalimaud Foundation

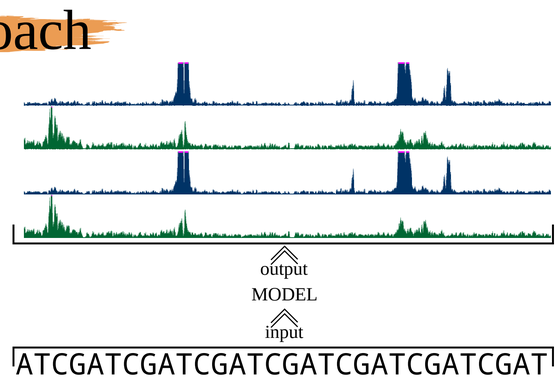
✉panos.firbas@research.fchampalimaud.org carolina.goncalves@research.fchampalimaud.org gonzalo.polavieja@neuro.fchampalimaud.org

Transformers are emerging as a promising tool in cis-regulation and chromatin modeling, but the typical design of relying solely on the reference genomic sequences as input severely limits their applicability in biological research. We introduce C.La.P. (Chromatin Language Processing), a transformer model that integrates genomic sequences with ATAC-seq signal to model individual cis-regulatory elements (CREs). The inclusion of ATAC-seq signal as input facilitates the modeling of Transcription Factor binding sites through the 'shadows' created by DNA-bound proteins. Our approach allows the model to make context-dependent predictions, unlocking its potential for real-world applications.

The typical modeling approach

Inputs large windows of reference genomic sequence to simulate large numbers of trained features.

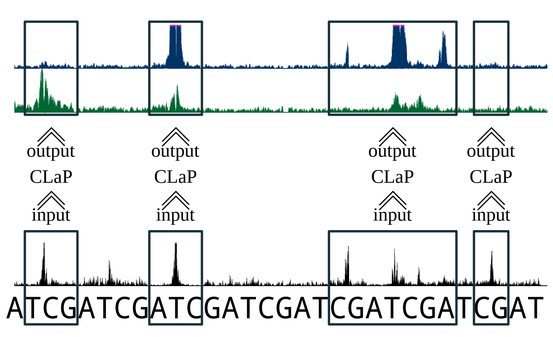
→ no biological context awareness
→ CRE regions are always included
→ non-CRE regions are also included
→ even the largest windows are too small



Our approach

Uses ATAC-seq to isolate CREs as samples and includes the ATAC-seq signal in the model.

→ ATAC-seq helps TF binding modeling
→ ATAC-seq provides context
→ inactive CRE regions are not included
→ non-CRE regions are not included
→ foundational model



The model, which consists of a convolutional tokenizer, followed by a transformer-encoder body and the task heads, was pre-trained with a masking task and fine-tuned with a nucleotide classification task for CTCF, H3k27ac, and H3k4me3. It achieves F1 scores of over 0.85 per nucleotide for all three classes.

Model Input
genomic sequence : 5 features
ATAC-seq signal: 2 features
DNA-shape: 20 features
masking: 1 feature

Datasets
all data from ENCODE project

pre-training:
75 experiments (ATAC-seq)
4.7M training samples
0.53M validation samples

fine-tuning:
33 experiments (ATAC-seq and CTCF,H3k27ac,H3k4me3 ChIP-seq)
1.5M training samples

Tokenizing/Embedding

convolutional layer 0
kernel size: 3
stride: 1
padding: none
output features: base(512), large(1024)

convolutional layer 1
kernel size: 9
stride: 1
padding: none
output features: base(768), large(1536)

convolutional layer 2
kernel size: 13
stride: 1
padding: none
output features: base(768), large(2048)

Encoder
Flash Attention
Alibi positional encoding
dropout: 0.1
#Layers: base(4), large(16)
#Heads: base(12), large(32)
Hidden size: base(768), large(2048)
Feed Fwd. size: base(3072), large(4096)

Training Hyperparameters
batch size: 5

Noam Optmizer
factor: 0.1
warmup steps: 2000
decay:0

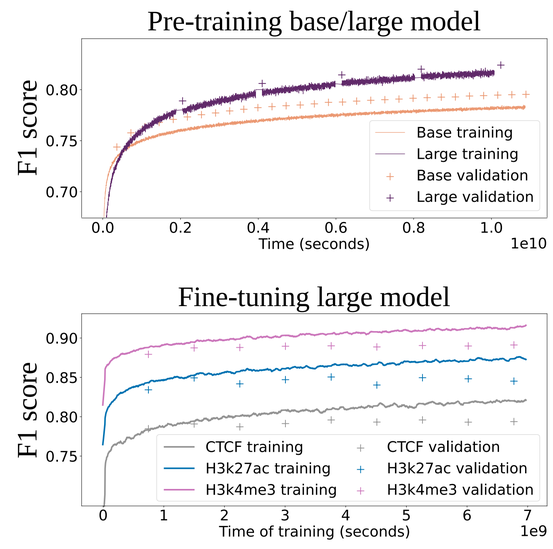
Pretraining Task Head
Single Linear Layer
4 classes output

Binary Cross Entropy loss only on masked tokens

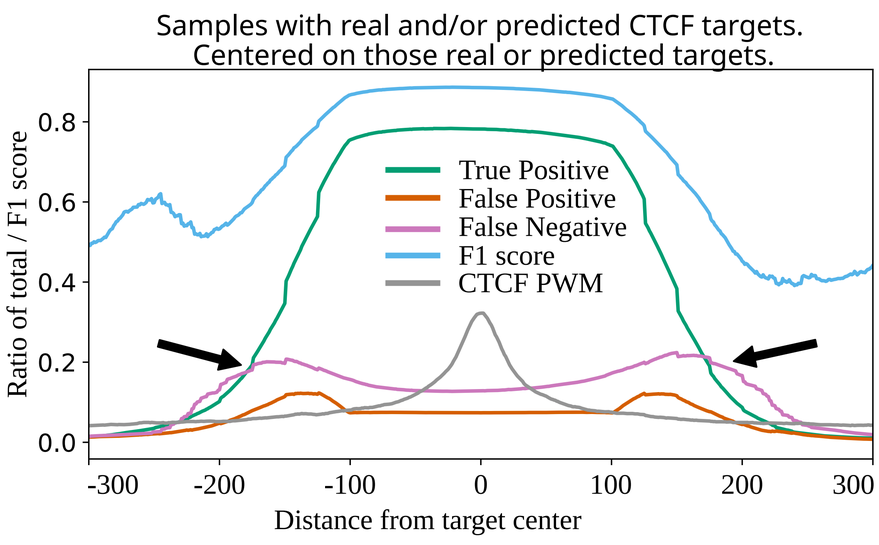
Fine-tuning Task Head

Single Linear Layer
3 classes output

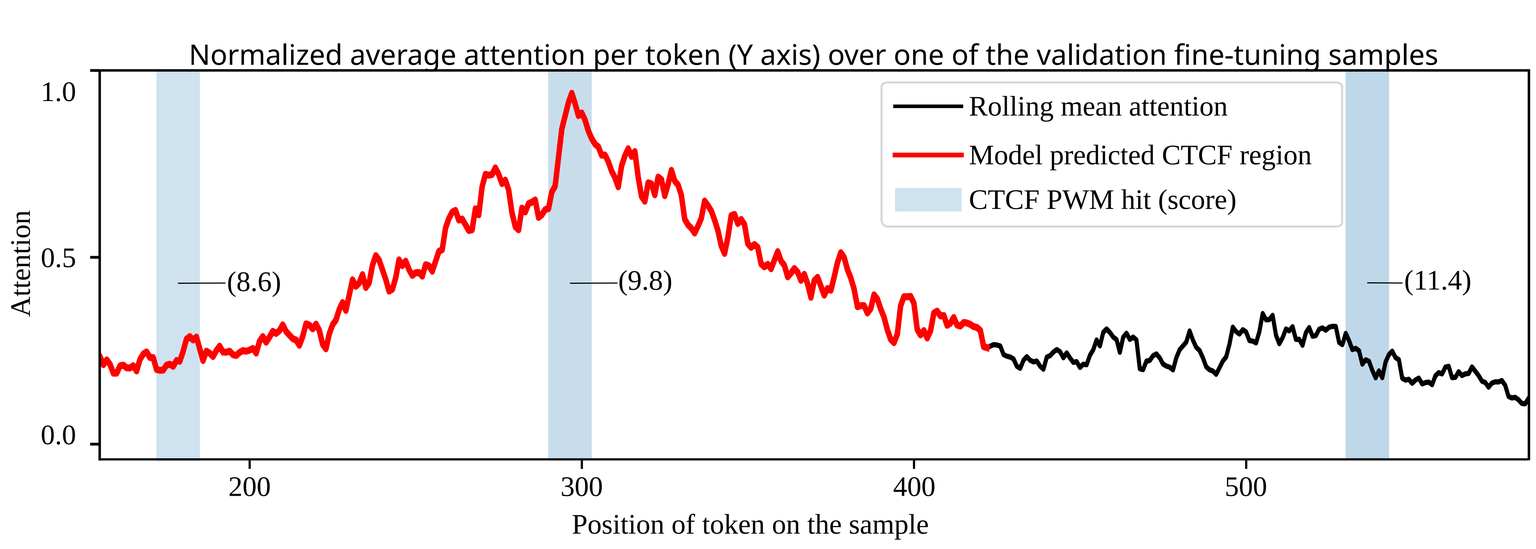
Binary Cross Entropy per token, per class.



In the case of the CTCF class, the model accurately predicts CTCF binding sites, but has a harder time in predicting the edges of the target ranges. Analysis of its attention mechanism reveals its ability to pinpoint CTCF binding sites with nucleotide-level precision. These predicted binding sites overwhelmingly overlap with CTCF PWM hits, but not always those with the highest score.

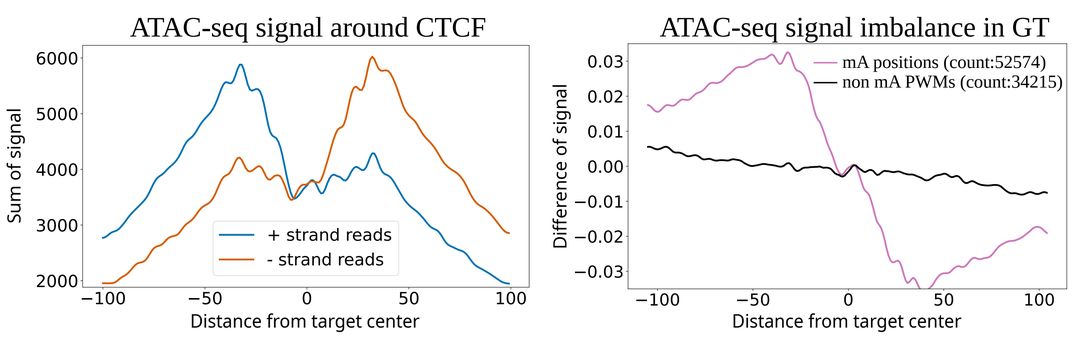


CTCF ChIP-seq signal extends for a random number of nucleotides away from the binding site at the center. C.La.P.'s False Positive and False Negative predictions are more common at the edges of real and predicted targets showing that the model captures the binding site but misses the borders.



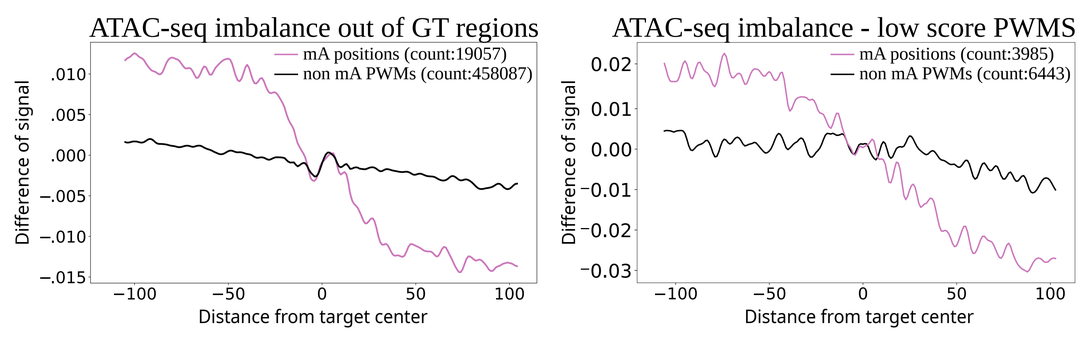
With applying attention, we produce a normalized attention score per token for any fine-tuning sample (Red and black line in the figure above). When the sample contains a ground truth or predicted CTCF target, the model's attention-peaks (mA) overwhelmingly overlap with hits for CTCF PWMs. While these mA positions typically overlap PWMs with high scores, that is not always the case as is showcased in the example above.

C.La.P. identifies real protein-DNA binding events, as evidenced by the characteristic ATAC-seq signal imbalance, even when those have not been captured by the ChIP-seq assay.



Left: By aggregating the ATAC-seq signal over multiple mA positions, we visualize the characteristic imbalance of + and - sequencing reads around events of protein binding.

Right: We can visualize the imbalance in a single line by subtracting the two signals. The imbalance around PWMs that overlap mA positions is much higher than those that don't because C.La.P. identifies real protein binding events.



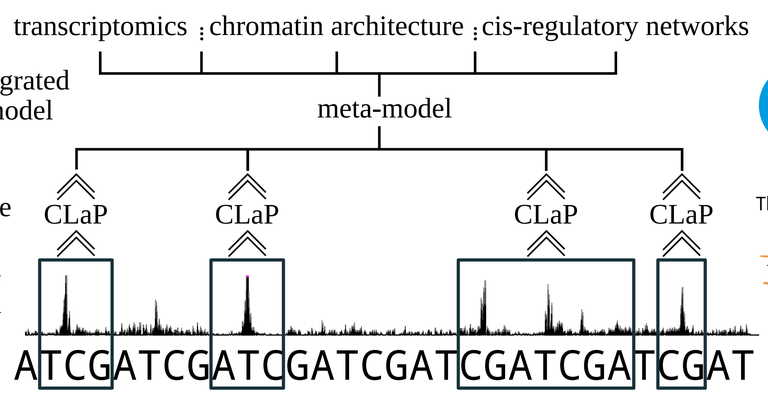
Left: C.La.P.'s mA positions at times overlap PWMs outside of ground truth CTCF regions. The ATAC-seq signal imbalance shows that the model detects protein binding events even when those have not been captured by the ChIP-seq assay.

Right: PWM hits with very low score (here with score lower than 6) exibit higher ATAC-seq signal imbalance when overlapping C.La.P.'s mA positions. The model relies on more than the genomic sequence to model protein binding.

Future directions

Multiple instances of C.La.P. can be integrated in a single meta-model to contextually model higher order biological functions.

➙ TF-binding and cis-regulation regulate those functions
➙ Many results at the price of one assay
➙ Modern professional GPUs necessary



Funding

Champalimaud Foundation
FindingPheno

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 952914

➙ Contact us for more !