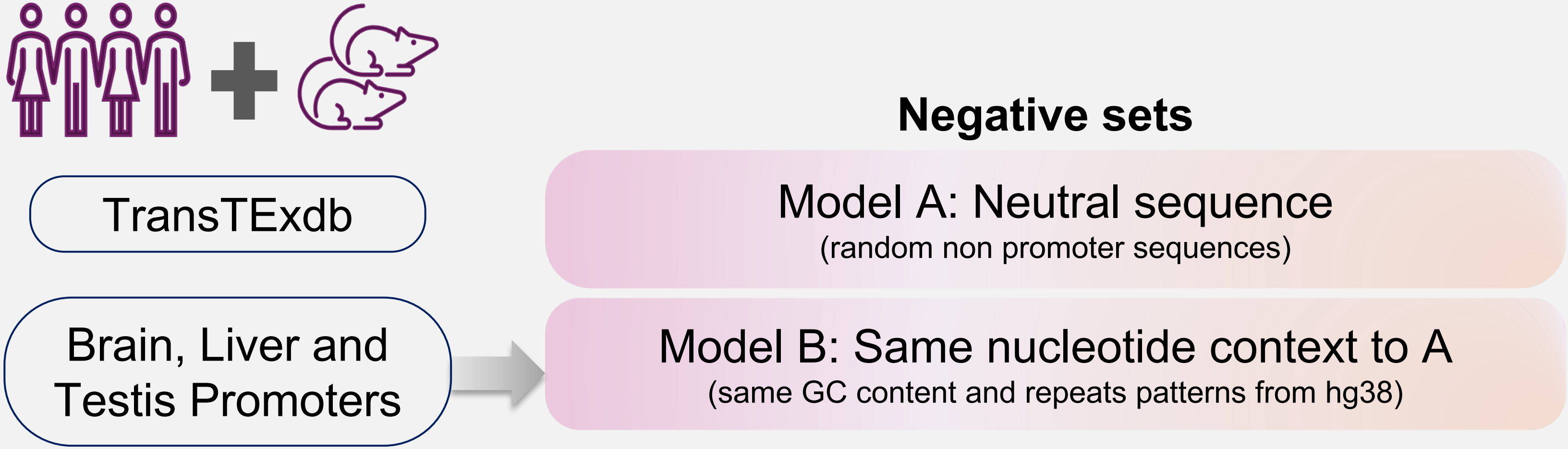
I. Data Preparation



TransTExdb

Brain, Liver and Testis Promoters

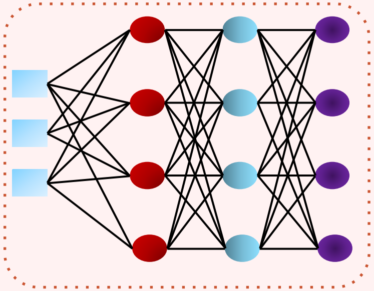
Negative sets

Model A: Neutral sequence
(random non promoter sequences)

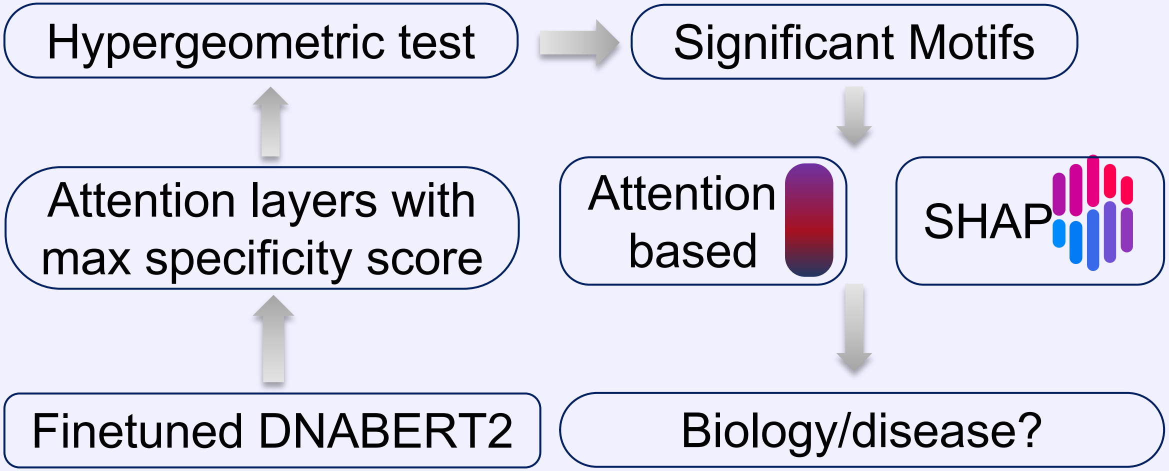
Model B: Same nucleotide context to A
(same GC content and repeats patterns from hg38)

II. Finetuning

- mmseq2
- 1:1 dataset
- 80:10:10 split



III. Motif discovery



Hypergeometric test

Significant Motifs

Attention layers with max specificity score

Attention based

SHAP

Finetuned DNABERT2

Biology/disease?