Dr. Gerstein has been extensively involved in leading roles across numerous significant genomic research consortia to better understand gene regulation and non-coding variation[73–75], including ENCODE[68], GENCODE[76], the 1000 Genomes Project[28,30,34], and the Impact of Genomic Variation on Function (IGVF) Consortium. Furthermore, the Gerstein lab has developed achieved significant advancements in the methodologies used to prioritize variants at multiple levels, including coding and non-coding variant prioritization, rare somatic and germline burden tests, and detailed allelic analysis[34,77–80]. This work has enhanced our understanding of the genomic architecture by analyzing protein-coding and non-coding regions.

The Gerstein lab recently led efforts to develop the EN-TEx resource, a multi-tissue epigenomic dataset comprising >1,600 assays mapped to the personal genomes of four individuals, and analyzed the impact of non-coding DNA variant effects towards transcription factor (TF) binding and histone modification[66]. We specifically used transformer-based language models to predict variant impact towards allele-specific behavior, such as transcription factor binding or gene expression, based on sequential contexts. Traditionally, allele-specific behavior is measured by mapping functional genomics data to personalized diploid genome sequences and using statistical tests to find changes in read depth between haplotypes at heterozygous SNVs[80]. Here, we trained a transformer model incorporating DNABERT[2] to predict which heterozygous SNVs exhibit allele-specific activity using the local sequence context (200-bp window) around the SNV (***Fig. 9A***). We used attention layers in the transformer model to calculate *attention scores* representing dependencies within the sequence, enabling the model to capture complex interactions analogous to grammars in natural languages. Our model outperformed several baselines in terms of prediction accuracy; for example, we accurately predicted allele-specific expression and binding for several transcription factors (CTCF and POL2) and histone modifications (H3K4me3 and H3K27ac) (***Fig. 9B***). Furthermore, attention scores from the transformer model showed similar patterns to enrichment of related TF motifs, and highlighted regions important for prediction, recapitulating motifs known to affect TF binding and revealing potential new motifs (***Fig. 9C***). When combined with tissue-specific epigenetic signals, the sequence-based scores contributed significantly to predict differential variant effects across tissues. These findings demonstrate that transformer models can learn dependencies between genomic sequential patterns without prior knowledge to provide novel insights into the mechanisms underlying variant effects.

The Gerstein lab's similar work on the DECODE framework leveraged sophisticated deep-learning architectures to refine genomic annotations[81] by training deep neural networks for precise enhancer prediction and localization. Recent efforts in the lab have developed an advanced framework that integrates text with multi-modal molecular representations, including 1D sequences, 2D interactions, and 3D structures[82]. Additionally, we recently fine-tuned the ESMFold LLM[10] on downstream bioinformatics tasks to predict protein phases (PPs), and demonstrated its superior performance compared to classical benchmarks such as random forest model predictions on the test set.

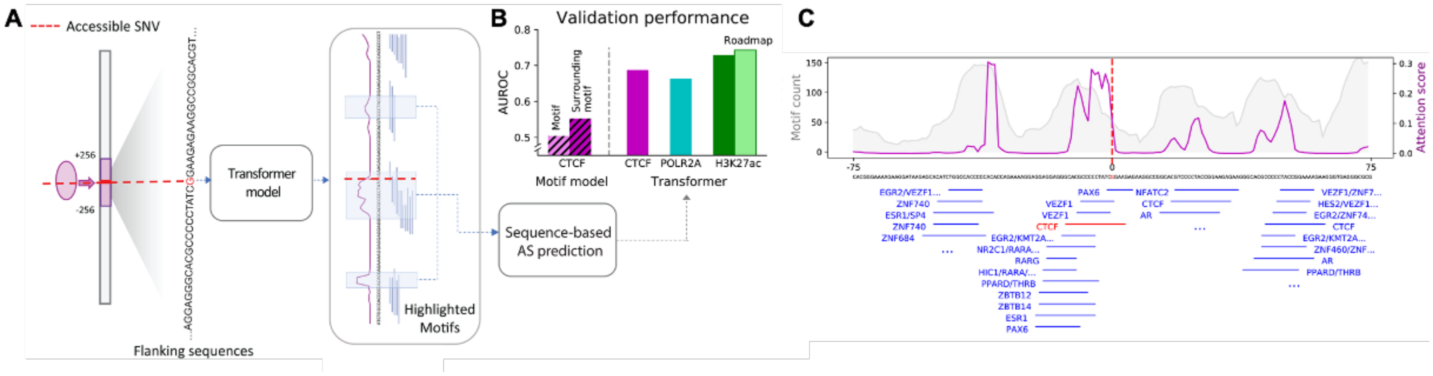

**Figure 9. Transformer model for predicting allele-specific behavior in the EN-TEx resource.** (**A**) *Overview schematic of the sequence-based predictive model.* (**B**) *Boxplots showing average performance of models to predict AS activity.* (**C**) *Attention patterns learned by the model.* Those in the flanking regions of CTCF AS SNV (magenta) show strong consistency with motif enrichment (gray).