**Gerstein lab experience in deep-learning models for genome annotation**: Dr. Gerstein has been extensively involved in leading roles across numerous significant genomic research consortia to better understand gene regulation and non-coding variation35–37, including ENCODE27, GENCODE38, the 1000 Genomes Project39–41, and IGVF42. Furthermore, the Gerstein lab has 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 analysis40,43–46. 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 modification25. We specifically used transformer-based language models to predict variant impact towards allele-specific (AS) 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 SNVs46. Here, we trained a transformer model incorporating DNABERT2 to predict which heterozygous SNVs exhibit allele-specific activity using the local sequence context (200-bp window) around the SNV (***Fig. 6A***). 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. 6B***). 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. 6C***). 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.

**Figure 6. Transformer model for predicting allele-specific (AS) 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).

**Gerstein lab experience in consortium work and ML/AI models**: Our teams have extensive experience working with collaborating across large consortia to develop resources for the genomics community. Notable, Dr. Gerstein is a member of the HGSVC, and contributed to flagship HGSVC papers and separate working-group papers to produce catalogs of functional genomics elements disrupted by SVs48,50,51. Both groups are also involved in research collaborations within the 1000 Genomes Consortium39,41,52–55. The Gerstein lab has also collaborated with computational and experimental groups across multiple consortia, including leading integrative analysis efforts for the ENCODE, modENCODE, and PsychENCODE consortia35,37,60–63. The lab is currently leading Data Coordination Centers for the PsychENCODE, dGTEx, SCORCH, and IGVF consortia, each of which entails standard dataset processing, developing uniform pipelines and ML/AI models to assess functional genomics data, and interfacing with the broader consortium for future tool development.

Additionally, our teams have broad experience with implementing standardized AI/ML models for a variety of use cases, each following best-practice guidelines for model development, benchmarking, and dissemination for community use. For example, the Gerstein lab’s DECODE framework leveraged sophisticated deep-learning architectures to refine genomic annotations64 by training deep neural networks for precise enhancer prediction and localization. Additionally, we recently fine-tuned the ESMFold LLM12 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. Furthermore, our BIOCODER project showcased the effectiveness of LLMs in managing and interpreting diverse biological data formats65. We initially developed MolLM, a pre-trained model that captures biomedical text and molecular information, enhancing performance66. We then found that MedAgents, a multi-disciplinary collaboration framework, significantly improved LLM reasoning in medicine67 , and ML-Bench demonstrated LLMs' ability to utilize open-source libraries68. Additionally, our structure-aware fine-tuning improved LLMs' capability to generate complex structured data69 , and the BioCoder benchmark illustrated our proficiency in bioinformatics coding and domain-specific challenges65. Finally, we fine-tuned an LLM to predict protein phase transitions, showing superior performance and interpretability, particularly for Alzheimer’s disease-related proteins70.

**REFERENCES**

2 Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics 2021;37:2112–20.

12 Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023. <https://doi.org/10.1126/science.ade2574>.

25 Rozowsky J, Gao J, Borsari B, Yang YT, Galeev T, Gürsoy G, et al. The EN-TEx resource of multi-tissue personal epigenomes & variant-impact models. Cell 2023;186:1493–511.e40.

27 ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 2020;583:699–710.

35 Emani PS, Liu JJ, Clarke D, Jensen M, Warrell J, Gupta C, et al. Single-cell genomics and regulatory networks for 388 human brains. Science 2024;384:eadi5199.

36 PsychENCODE Consortium, Akbarian S, Liu C, Knowles JA, Vaccarino FM, Farnham PJ, et al. The PsychENCODE project. Nat Neurosci 2015;18:1707–12.

37 Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, et al. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science 2010;330:1775–87.

38 Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 2019;47:D766–73.

39 Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* 2015;**526**:75–81.

40 Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, *et al.* Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. *Science* 2013;**342**:1235587.

41 The 1000 Genomes Project Consortium. A map of human genome variation from population scale sequencing. *Nature* 2010;**467**:1061.

42 IGVF Consortium. Deciphering the impact of genomic variation on function. Nature 2024;633:47–57.

43 Balasubramanian S, Fu Y, Pawashe M, McGillivray P, Jin M, Liu J, *et al.* Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes. *Nat Commun* 2017;**8**:382.

44 Lochovsky L, Zhang J, Fu Y, Khurana E, Gerstein M. LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. *Nucleic Acids Res* 2015;**43**:8123–34.

45 Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, *et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* 2014;**15**:480.

46 Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* 2011;**7**:522.

47 Schatz MC, Philippakis AA, Afgan E, Banks E, Carey VJ, Carroll RJ, *et al.* Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. *Cell Genom* 2022;**2.**: <https://doi.org/10.1016/j.xgen.2021.100085>.

48 Logsdon GA, Ebert P, Audano PA, Loftus M, Porubsky D, Ebler J, et al. Complex genetic variation in nearly complete human genomes. bioRxiv 2024:2024.09.24.614721. <https://doi.org/10.1101/2024.09.24.614721>.

49 Novak AM, Chung D, Hickey G, Djebali S, Yokoyama TT, Garrison E, et al. Efficient indexing and querying of annotations in a pangenome graph. bioRxiv 2024. <https://doi.org/10.1101/2024.10.12.618009>.

50 Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science 2021;372.: <https://doi.org/10.1126/science.abf7117>.

51 Li C, Bonder MJ, Syed S, Jensen M, Human Genome Structural Variation Consortium (HGSVC), HGSVC Functional Analysis Working Group, et al. An integrative TAD catalog in lymphoblastoid cell lines discloses the functional impact of deletions and insertions in human genomes. Genome Res 2024;34:2304–18.

52 A global reference for human genetic variation. Nature 2015;526:68–74.

53 An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491:56–65.

54 Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, et al. Mapping copy number variation by population-scale genome sequencing. Nature 2011;470:59–65.

55 Delaneau O, Marchini J. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nat Commun 2014;5:1–9.

60 Gerstein MB, Rozowsky J, Yan K-K, Wang D, Cheng C, Brown JB, *et al.* Comparative analysis of the transcriptome across distant species. *Nature* 2014;**512**:445–8.

61 Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, *et al.* Comparative analysis of regulatory information and circuits across distant species. *Nature* 2014;**512**:453–6.

62 Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, *et al.* Landscape of transcription in human cells. *Nature* 2012;**489**:101–8.

63 Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* 2018;**362.**: <https://doi.org/10.1126/science.aat8464>.

64 Chen Z, Zhang J, Liu J, Dai Y, Lee D, Min MR, *et al.* DECODE: a Deep-learning framework for Condensing enhancers and refining boundaries with large-scale functional assays. *Bioinformatics* 2021;**37**:i280–8.

65 Tang X, Qian B, Gao R, Chen J, Chen X, Gerstein MB. BioCoder: a benchmark for bioinformatics code generation with large language models. *Bioinformatics* 2024;**40**:i266–76.

66 Tang X, Tran A, Tan J, Gerstein MB. MolLM: A Unified Language Model for Integrating Biomedical Text with 2D and 3D Molecular Representations. *bioRxiv* 2024:2023.11.25.568656. <https://doi.org/10.1101/2023.11.25.568656>.

67 Tang X, Zou A, Zhang Z, Li Z, Zhao Y, Zhang X, *et al.* MedAgents: Large Language Models as Collaborators for Zero-Shot Medical Reasoning. Presented at the Findings of the Association for Computational Linguistics ACL 2024, Bangkok, Thailand and virtual meeting.

68 Tang X, Liu Y, Cai Z, Shao Y, Lu J, Zhang Y, *et al.* ML-bench: Evaluating Large Language Models and agents for machine learning tasks on repository-level code. *arXiv [csCL]* 2023. <https://doi.org/10.48550/ARXIV.2311.09835>.

69 Tang X, Zong Y, Phang J, Zhao Y, Zhou W, Cohan A, *et al.* Struc-Bench: Are Large Language Models Good at Generating Complex Structured Tabular Data? Presented at the Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), Mexico City, Mexico.

70 Frank M, Ni P, Jensen M, Gerstein MB. Leveraging a large language model to predict protein phase transition: A physical, multiscale, and interpretable approach. *Proc Natl Acad Sci U S A* 2024;**121**:e2320510121.