

Neuronal and microglial fine-tuned 3D DNA foundation models dissect genetic variants in Alzheimer's disease

Feixiong Cheng

chengf@ccf.org

Cleveland Clinic <https://orcid.org/0000-0002-1736-2847>

Yunxiao Ren

Cleveland Clinic

Maria Paz

Cleveland Clinic <https://orcid.org/0000-0003-0862-5673>

Xiaoyu Yang

Cleveland Clinic

Yichen Li

Cleveland Clinic <https://orcid.org/0009-0005-1355-5018>

Yuan Hou

Cleveland Clinic

Yang Li

Washington University in St. Louis <https://orcid.org/0000-0001-6997-6018>

Fulai Jin

Case Western Reserve University <https://orcid.org/0000-0003-0025-4337>

Ming Hu

Cleveland Clinic Lerner research institute <https://orcid.org/0000-0003-0987-2916>

Andrew Pieper

Case Western Reserve University

Article

Keywords:

Posted Date: February 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8920766/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Neuronal and microglial fine-tuned 3D DNA foundation models dissect genetic variants in Alzheimer's disease

Yunxiao Ren^{1,2}, Maria Rivera Paz^{1,2}, Xiaoyu Yang^{1,2}, Yichen Li^{1,2}, Yuan Hou^{1,2}, Yang E.
Li³, Fulai Jin^{4,5}, Andrew A. Pieper⁶⁻¹¹, Ming Hu^{2,12}, Feixiong Cheng^{1,2,5,12}

¹Cleveland Clinic Genome Center, Cleveland Clinic Research, Cleveland Clinic,
Cleveland, OH 44195, USA

²Department of Genomic Sciences and Systems Biology, Cleveland Clinic Research,
Cleveland Clinic, Cleveland, OH 44195, USA

³Department of Neurosurgery and Genetics, Washington University School of Medicine,
St. Louis, MO 63110, USA

⁴Department of Genetics and Genome Sciences, Case Western Reserve University,
School of Medicine, Cleveland, OH 44106, USA

⁵Case Comprehensive Cancer Center, Case Western Reserve University School of
Medicine, Cleveland, OH 44106, USA

⁶Department of Psychiatry, Neurosciences, and Pathology, Case Western Reserve
University, Cleveland, OH 44106, USA

⁷Department of Pathology, Case Western Reserve University, Cleveland, OH, 44106,
USA

⁸Department of Neurosciences, Case Western Reserve University, Cleveland, OH,
44106, USA

⁹Brain Health Medicines Center, Harrington Discovery Institute, University Hospitals
Cleveland Medical Center, Cleveland, OH, 44106, USA

23 ¹⁰Institute for Transformative Molecular Medicine, School of Medicine, Case Western
24 Reserve University, Cleveland, OH, 44106, USA

25 ¹¹Geriatric Psychiatry, GRECC, Louis Stokes Cleveland VA Medical Center; Cleveland,
26 OH 44106, USA

27 ¹²Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Case
28 Western Reserve University, Cleveland, OH 44195, USA

29

30 *Correspondence to: Feixiong Cheng, PhD

31 Cleveland Clinic Genome Center

32 Address: 9500 Euclid Ave, Cleveland, Ohio 44195, USA

33 Tel: +1-216-444-7654; Fax: +1-216-636-0009

34 Email: chengf@ccf.org

35

Abstract

Although genetic variants are among the strongest risk factors for Alzheimer's disease (AD), most associated variants reside in non-coding regions with unclear mechanisms. Here, we address that gap by fine-tuning the Akita DNA foundation model on high-resolution Hi-C data generated from neuronal progenitor cells (NPCs) and microglia to produce *iNeuroAkita*, a cell-context-aware three-dimensional (3D) genome model. *iNeuroAkita* shows improved correlation with experimental Hi-C contact maps and more accurate contact predictions at key AD loci (e.g. *APOE* and *PSEN2*), and captures CTCF-dependent chromatin loops. Using *in silico* saturation mutagenesis (ISM) across 11,856 AD-associated single-nucleotide variants (SNVs) and 120 structural variants (SVs), we quantified impacts on chromatin spatial organization and identified two SNVs predicted to strongly disrupt 3D chromatin architecture: rs636317 at the *MS4A* locus that is predicted to disrupt local chromatic interactions and microglia eQTL rs7599054 that creates gained interactions connecting its surrounding region to *CCNT2* and *TMEM163*, consistent with enhancer–promoter activation in epigenomic and single-cell transcriptomic data. We also identified AD-associated deletions at the *KANSL1* and *ARMS2* that rewire local regulatory contacts and associate with cell-type-specific expression changes, demonstrating *iNeuroAkita*'s utility for systematic characterization of noncoding AD risk variants underlying 3D genome regulation.

Introduction

Alzheimer's disease (AD) is influenced by inherited genetic variation, with genome-wide association studies (GWAS) indicating that a substantial portion of risk is attributable to common single-nucleotide variants (SNVs)^{1–4}. A growing body of evidence indicates that many AD-associated variants exert their effects by perturbing three-dimensional (3D) chromatin architecture^{5–8}, for example by altering chromatin loops, modifying long-range enhancer–promoter interactions⁹, or affecting transcription factor and CTCF binding affinity^{1,10–12}. Since 3D genome structure is critical for cell-type-specific gene regulation^{13,14}, characterizing how AD risk variants reshape chromatin folding is essential for interpreting their functional consequences.

Functional studies have further revealed that AD risk variants are enriched within regulatory elements and often exert cell-type-specific effects, particularly in microglia^{7,12,15,16}. However, these studies have primarily examined transcriptional consequences and have not systematically assessed direct effects on 3D chromatin architecture. Experimental approaches that measure variant-induced changes in chromatin folding (for example, CRISPR perturbations and high-resolution chromatin conformation capture assays) are labor-intensive, and technically demanding, making them impractical to scale to the thousands of common variants identified by large AD GWAS studies.

Structural variations (SVs) comprise a complementary and often more disruptive class of genetic variations that can reorganize chromatin architecture across kilobase- to-megabase scales by altering gene dosages, deleting *cis*-regulatory elements, or rewiring long-range chromatin interactions^{17,18}. One illustrative example is the *APP*

duplication, a rare but fully penetrant cause of early-onset AD¹⁹. Recent large-cohort whole-genome sequencing efforts have begun to implicate SVs more broadly in AD^{19,20}. For instance, Wang et al.¹⁹ reported several ultra-rare SVs associated with AD that disrupt splice acceptor or donor sites in AD risk genes, including *ABCA7*, *APP*, *PLCG2*, and *SORL1*. Analysis of ROS/MAP cohorts have also linked SVs to explain clinical, cognitive, and neuropathologic traits, identifying more than twenty SVs in linkage disequilibrium (LD) with GWAS lead variants that associate with AD-related phenotypes²⁰. Despite increasing evidence that SVs influence AD susceptibility, their effects on 3D genome organization and regulatory logic remain poorly characterized, largely because of experimental constraints.

There is thus a clear need for computational frameworks that predict how genetic variations alter 3D chromatin architecture in disease-relevant cell types. Advances in sequence-to-structure modeling, including deep learning–based DNA foundation models^{21,22}, show that primary DNA sequence encodes substantial information about chromatin folding. Models such as Akita²¹ and related architectures can predict contact maps directly from DNA sequence, enabling scalable *in silico* interrogation of variant-induced structural perturbations at genome-wide scale. However, most existing models were trained in non-neural or otherwise AD-irrelevant cell types, such as GM12878 lymphoblastoid cells, IMR-90 lung fibroblasts, HCT116 colorectal carcinoma cells, human foreskin fibroblasts (HFF), and H1 human embryonic stem cells, which limits their utility for studying for neurodegenerative disorders.

Here, we address this gap by fine-tuning a DNA foundation model — Akita on high-resolution Hi-C data from neuronal progenitor cells (NPCs) and microglia to learn

the AD-relevant contexts, producing *iNeuroAkita*. This adaptation enables systematic prediction of variant-induced changes in chromatin architecture. Applying this framework to AD-associated SNVs and SVs reveals widespread disruption of chromatin spatial organization, including CTCF-mediated boundary weakening and long-range regulatory rewiring. This work demonstrates how cell-type-aware DNA foundation models provide a scalable and mechanistic approach to interpreting AD-associated genetic variants through the 3D genome perspective.

Results

NPC and microglia fine-tuned 3D genome models enables prediction of chromatin architecture

To interpret how AD-associated variants alter chromatin architecture, we developed a cell-type-aware 3D genome model tailored to human brain-relevant cell types. Existing models such as Akita²¹ are trained in non-neural cells—including GM12878 lymphoblastoid cells, IMR-90 lung fibroblasts, HCT116 colorectal carcinoma cells, human foreskin fibroblasts (HFF), and H1 human embryonic stem cells— and therefore do not capture the chromatin landscapes of NPCs or microglia, which both play important roles in AD pathogenesis^{11,23}. We therefore fine-tuned Akita on Hi-C data generated from NPCs and healthy human primary microglia (**Fig. 1**).

We generated Hi-C data from NPCs that were differentiated from an induced pluripotent stem cell (iPSC) line derived from an AD human patient donor and processed into high-quality chromatin contact maps (**Supplementary Table S1**). This data was partitioned into training, validation, and held-out test sets to support model

optimization and unbiased evaluation. To further assess model generalizability, we incorporated publicly available human NPC Hi-C data²⁴ as an independent test resource. In parallel, Hi-C data from human primary microglia²⁵ were also processed and similarly split into training, validation, and test sets (**Fig. 1**).

All Hi-C datasets were converted into uniformly 2kb (2,048 bp) binned contact matrices and paired with corresponding hg38 reference genome to train the model on megabase-scale genomic windows (1,048,576 bp). Starting from pre-trained Akita parameters²¹, the model was fine-tuned on NPCs and microglia data independently, with key hyper-parameters optimized to maximize predictive performance using Dragonfly²⁶ (see Methods). This new cell-type-aware model, which we term *iNeuroAkita*, provides a foundation for evaluating variant-induced disruptions of 3D chromatin organization in AD-relevant cellular contexts (**Fig. 1**).

Our *iNeuroAkita* model showed overall high predictive accuracy on both the independent NPC test set (**Fig. 2a**) and an internal held-out test set (**Supplementary Fig. S1**), demonstrating robust generalization across datasets. On independent NPC test data, the model reached a global mean squared error (MSE) of 0.1144 and a Pearson correlation coefficient (PCC) of 0.6988, where lower MSE and higher correlation value means better performance, indicating robust reconstruction of experimentally derived 3D contact maps. The *iNeuroAkita* model also performed well on the microglia held-out test set, achieving a global MSE of 0.1256 and a PCC of 0.5152 on its held-out test set (**Fig. 2b**).

We next compared the *iNeuroAkita* model with the Akita using the independent NPC test dataset. Across most chromosomes, the *iNeuroAkita* model yielded

significantly lower MSE and higher correlation, confirming consistent improvements in predictive fidelity (**Fig. 2c**). Although a few chromosomes did not reach statistical significance, the overall trend favored the fine-tuned model. For instance, *iNeuroAkita* substantially outperformed Akita with respect to the AD-associated *APOE* locus on chromosome 19 (**Fig. 2d**), *PSEN2*, *CLU* and *TOMM40* (**Supplementary Fig. S2**). In the *APOE* region, *iNeuroAkita* model captured local chromatin architecture (PCC = 0.733, MSE = 0.1324) with elevated performance compared with the baseline Akita (PCC = 0.425, MSE = 0.2829) (**Fig. 2d**). These results demonstrate that *iNeuroAkita* fine-tuning significantly improved the accuracy of 3D genome prediction in AD-relevant neural contexts.

Disruption of CTCF motifs rewires local chromatin structures in NPCs

To further evaluate whether *iNeuroAkita* model accurately captures known sequence determinants of chromatin architecture, we focused on CTCF, a key architectural protein whose binding motifs anchor chromatin loops and insulate topologically associating domains (TADs)²⁷. Because CTCF plays a central role in shaping 3D genome organization, evaluating how the model responds to perturbations in CTCF motifs provides a stringent biological test of model fidelity.

We first extracted core CTCF-motif regions from JASPAR (2024)²⁸ and intersected these peaks with genomic intervals derived from the independent NPC test set and CTCF ChIP-seq data in NPC (**Supplementary Table S2**). For windows containing at least one CTCF site, the absolute CTCF coordinates were converted to positions relative to the 1-Mb sequence. Within each ~10-bp CTCF core motif, we introduced random single-nucleotide substitutions, replacing each reference base with

one of the three alternative nucleotides (A/C/G/T excluding the reference base), sampled uniformly at random, while preserving motif length and genomic coordinates. Ten independent mutated sequences were generated per CTCF-containing window, producing ten mutated (MUT) sequences for downstream in silico perturbation analyses. Across all tested CTCF loci, we computed a quantitative disruption score based on the difference between wild-type (WT) and mutated predictions (MUT) (**Supplementary Table S3**). We highlight the three loci with the strongest disruption scores in **Fig. 3a-c**. In each case, WT predictions closely matched the corresponding experimental Hi-C contact maps (PCC = 0.744, 0.559, and 0.704, MSE = 0.109, 0.130, and 0.173), whereas mutations of the CTCF motif produced localized and substantial alterations in 3D contact patterns (PCC = 0.423, 0.145, and 0.320, MSE = 0.222, 0.348, and 0.340). The weighted differential maps (DIFF = MUT – WT Prediction) further highlight focal gains (purple) and losses (green) of chromatin interactions at the disrupted motifs.

At the first locus (chr8:80.65–81.70 Mb) (**Fig. 3a**), the mutated CTCF site near 8.09 Mb (**Supplementary Table S2**) anchors a strong chromatin loop that demarcates two local domains in the WT structure (Top panel of **Fig. 3a**, and experimental contact map). This loop is faithfully reconstructed by *iNeuroAkita* model but becomes markedly attenuated following motif disruption, with the differential map showing a clear loss of interactions across the TAD corner and compensatory gains in neighboring regions. Notably, this locus lies adjacent to *FABP5*, a fatty-acid-binding protein implicated in neurogenesis and neural stem/progenitor cell function, providing a plausible NPC-relevant target within this reorganized domain²⁹.

At the second locus on chromosome 5 (chr5: 76.33-77.38 Mb; **Fig. 3b**), mutation of one of CTCF sites (7.68Mb, **Supplementary Table S2**) near the *S100Z* gene alters local chromatin interaction patterns, as seen by diffused domain structure and mixed purple–green signals across the boundary in the differential map. At the third locus on chromosome 1 (chr1: 30.45-31.50 Mb, **Fig. 3c**), CTCF disruption sites (3.08Mb and 3.11Mb, **Supplementary Table S2**) near *SDC3* and *NKAIN1* produces extensive rewiring of long-range chromatin interactions, with strong differential signals along the diagonal and at loop anchors. *SDC3* (syndecan-3) is a neuronally enriched heparan sulfate proteoglycan with high transcript expression in neural tissues, and plays important roles in synapse formation, neurite outgrowth, and neuronal migration³⁰. *NKAIN1* interacts with Na⁺/K⁺-ATPase and is expressed in neurons³¹, suggesting that changes in their 3D regulatory environment could be functionally relevant in neural lineages.

Together, these examples show that the *iNeuroAkita* model not only reproduces experimentally observed chromatin architecture but also captures the sequence-specific consequences of CTCF motif disruption at loci harboring neurodevelopmentally relevant genes.

AD-associated regulatory variants reveal focal 3D chromatin disruption at the *MS4A* locus

We next turned to inspect the 3D structural consequences of 42 experimentally determined, candidate regulatory variants across 15 AD GWAS loci and 3 progressive supranuclear palsy (PSP) loci¹² (**Fig. 4a, Supplementary Table S4**). These variants represent the *cis*-regulatory candidates emerging from prior functional genomic work¹²

but have not been systematically evaluated for their effects on chromatin architecture. We generated 1,048,576-bp WT and mutant sequences centered on each of the 42 candidate regulatory variants. We then test whether *iNeuroAkita* could predict the corresponding mutated contact maps. To quantify the magnitude of 3D structural changes across all variants, we aggregated multiple difference-based metrics into a single composite disruption score (*Z score*) (see Methods), capturing the overall strength, extent, and consistency of mutation-induced changes in predicted mutated contact maps. Across all 42 loci, rs636317 within the *MS4A* gene cluster was identified as a top candidate. This variant showed an elevated disruption score ($z = 9.4$; $p < 0.001$) (**Supplementary Table S5, Supplementary Fig. S3**), whereas all remaining variants clustered near the baseline, producing little or no detectable change in 3D genome. Notably, the rs636317 mutation produced a pronounced focal disturbance in the predicted contact map: a strong loss of chromatin interaction near the *MS4A4E* region (**Fig. 4b**).

Integration of chromatin interaction changes with cell-type-specific regulatory landscapes further reveals the functional consequences of rs636317 (**Fig. 4c**). One disrupted loop anchor lies near *MRPL16*, where both microglia and neurons show strong chromatin activity (H3K27ac, H3K4me3, and ATAC-seq peaks) (**Fig. 4c**). The predicted reduction of this contact is consistent with altered long-range chromatin interactions involving upstream regulatory elements following the mutation. Another lost interaction is observed at the *MS4A4E* locus (**Fig. 4c**). This site exhibits strong microglia-specific enhancer activity (H3K27ac + ATAC), indicating that rs636317 may disrupt a functional microglial enhancer-promoter contact. This aligns with earlier

CRISPR perturbation studies that rs636317 reduces *MS4A4E* expression specifically in microglia, suggesting *MS4A4E* as a variant-responsive target¹².

Together, these findings suggest that rs636317 may influence microglia-specific regulatory architecture within the *MS4A* locus, and may be associated with reduced *MS4A4E* expression.

3D genome modeling of microglial eQTLs reveals structurally disruptive variants

We next applied *iNeuroAkita* model to 11,814 microglia-specific fine-mapped eQTL variants (**Supplementary Table S6**), comparing predicted contact maps and mutated contact maps and summarized differences into a composite disruption score. We identified 37 variants with significant 3D perturbations ($z > 6$, adjusted $p < 0.001$, **Supplementary Table S6, Supplementary Fig. S4**), highlighting rare but detectable microglial regulatory variants may exert structural effects on chromatin folding.

Among the 37 variants, rs7599054 emerged as the top-ranked locus, showing a pronounced composite disruption ($z = 116.8$, adjusted $p < 0.001$), far exceeding all other candidates (**Fig. 5a**). Notably, rs7599054 has been identified as a microglia eQTL with *TMEM163* as its target gene. Mutation of rs7599054 produced localized but substantial rearrangement of microglia-specific chromatin interactions (**Fig. 5b**), and the differential map revealed a gain of chromatin interaction linking the region surrounding rs7599054 to the *CCNT2* and *TMEM163* genes—contacts that are absent or considerably weaker in the WT configuration.

Integrative regulatory profiles further contextualize these alterations (**Fig. 5c**). For example, rs7599054 lies within an active microglial regulatory region marked by strong H3K27 acetylation (H3K27ac), ATAC accessibility, and CTCF binding, suggesting that

the variant resides in a structurally important enhancer–boundary region. The gained interaction created by the mutation anchors in the vicinity of *CCNT2*, which displays clear H3K27ac, H3K4me3, and ATAC-seq signal in both microglia and neurons, consistent with a transcriptionally active promoter/enhancer hub. The shift in chromatin architecture may enhance physical proximity between the rs7599054 regulatory element and the *CCNT2* promoter, potentially favoring increased regulatory influence at this gene. Additional gained contacts extend toward *TMEM163*, which also harbors microglia-enriched enhancer marks.

To assess whether these structural changes correspond to transcriptional effects in human brain cell types, we examined single-cell RNA-seq data from the SEA-AD cohort³². *CCNT2* expression was selectively elevated in microglia from AD-associated dementia individuals, whereas expression was comparatively low in excitatory and inhibitory neurons (**Fig. 5d**). This microglia-specific upregulation aligns with the chromatin rewiring predicted by *iNeuroAkita*. Together, these results identify rs7599054 as a structurally disruptive microglial eQTL whose mutation strengthens enhancer–promoter connections at *CCNT2*.

Structural variants at the *KANSL1* and *ARMS2* loci rewire chromatin loops and alter gene regulation in AD

Structural variants have the potential to reshape chromatin topology more profoundly than single-nucleotide changes. To systematically assess the impact of SVs on 3D genome organization, we applied our *iNeuroAkita* to 120 AD-associated SVs curated from two recent large-scale studies^{19,20} (**Supplementary Table 8**; see Methods). For each SV, we performed *in silico* mutagenesis and quantified predicted changes in

chromatin contact maps using MSE and PCC metrics (**Supplementary Table 9**) implemented by the SupreMo tool³³.

Across the SV set, longer variants spanning multiple regulatory elements or gene bodies frequently exhibited higher global perturbation scores, reflecting widespread reorganization of local chromatin contacts. For example, a 7,837-bp deletion encompassing multiple exons of *PLCG2* produced extensive predicted interaction changes across several neighboring genes, including *PLCG2*, *SDR42E1*, *CMIP*, and *HSD17B2* (**Supplementary Fig. S5**).

It is important to note that large SVs frequently generate complex and spatially diffuse 3D effects that are difficult to attribute to specific regulatory mechanisms. To illustrate more localized and interpretable structural perturbations, we therefore focused on a smaller 314-bp intronic deletion within *KANSL1*, which, despite a more moderate global impact score, produced clear and spatially confined alteration of chromatin folding. This deletion intensified and redistributed interactions near *LRRC37A*, *ARL17A*, and *ARL17B* (**Fig. 6a**), suggesting a targeted redistribution of contacts within this densely regulated locus.

Integration with epigenomic data (**Fig. 6b**) shows that these gained interactions converge on regions enriched for H3K27ac, H3K4me3, and ATAC-seq signal in microglia and neurons, indicating that the deletion strengthens connectivity among active enhancer and promoter elements. The affected anchors lie directly adjacent to *ARL17A*, *ARL17B*, and *LRRC37A*, implicating these genes as potential regulatory targets of the altered chromatin architecture. Consistent with altered chromatin contacts, single-cell RNA-seq profiles from the SEA-AD cohort showed selective upregulation of

ARL17A and *ARL17B* in microglia and inhibitory neurons from AD patients, and elevated *LRRC37A* and *ARL17B* expression in excitatory neurons (**Fig. 6c**). These transcriptional differences align with the structural changes predicted by *iNeuroAkita*, consistent with the hypothesis that the deletion in *KANSL1* perturbs local regulatory architecture surrounding *ARL17A*, *ARL17B*, and *LRRC37A*.

We further identified a 445-bp deletion in *ARMS2* by *iNeuroAkita*, located within an enhancer-rich region upstream of *HTRA1* (**Fig. 7a, 7b**). The DEL-445 allele produced a distinct 3D architectural shift, including strengthened interactions spanning the *PLEKHA1* promoter, a well-established AD risk gene³⁴, as well as nearby regulatory regions encompassing *BTBD16* and *ARMS2* (**Fig. 7a, 7b**).

Integration with epigenomic data (**Fig. 7b**) revealed that the deleted region overlaps CTCF binding sites, as well as close multiple H3K27ac- and ATAC-marked enhancers in microglia and neurons. The altered contact pattern increases physical proximity between the deletion-adjacent region and regulatory elements positioned near *PLEKHA1* and *HTRA1*. In line with these structural changes, single-cell RNA-seq analysis reveals elevated *PLEKHA1* expression in microglia from AD patients (**Fig. 7c**). This expression pattern corresponds to the region showing enhanced contact strength in *iNeuroAkita*, suggesting that DEL-445 may influence transcriptional activity at the *PLEKHA1* locus.

Discussion

In this study, we fine-tuned the Akita model using high-resolution Hi-C data generated from NPCs and microglia, two brain cell types that are central to AD pathogenesis. This cell-type-informed training substantially improved overall prediction accuracy compared

with Akita model and enhanced performance at AD-relevant loci, enabling fine-grained prediction of how AD-associated variants remodel chromatin architecture in disease-relevant cellular contexts. We also demonstrated that *iNeuroAkita* accurately captures CTCF-dependent chromatin folding, as evidenced by consistent and localized structural disruptions following targeted CTCF motif mutagenesis. The strongest effects occurred at loci harboring neuronal or neurodevelopmentally relevant genes. Thus, *iNeuroAkita* not only learns general DNA sequence determinants of chromatin architecture but is also sensitive to regulatory landscapes of AD-relevant cell-types.

We also demonstrated that *iNeuroAkita* enables prediction of AD/ADRD variant-driven changes in chromatin architecture within AD-relevant cell types. Among the 42 high-confidence AD-associated variants analyzed, rs636317 produced one of the strongest predicted 3D structural disruptions. This variant alters only a single nucleotide within a CTCF binding motif¹², demonstrating that even subtle DNA sequence changes can reorganize local chromatin architecture. Importantly, not all variants residing in CTCF motifs generated large structural changes³⁵, such as rs6701713 and rs755951 (**Supplementary Fig. S3**), highlighting the context-dependent nature of chromatin architecture. The functional impact of a variant on 3D genomic structure is not determined by CTCF occupancy alone³⁶, but rather by a combination of factors including motif strength³⁷, flanking sequence composition³⁸, local enhancer and promoter density^{39,40}, cooperative transcription-factor binding^{41–43}, and cell-type-specific chromatin accessibility^{44–48}. These results emphasize that 3D genome disruption is governed by local regulatory architecture, and that only a subset of CTCF-overlapping variants is poised to exert biologically meaningful structural effects.

We also screened a large number of microglia eQTLs and identified 37 eQTLs with significant impact on 3D genome, opening the door to broad applications, including variant prioritization at disease loci and mechanistic interpretation of GWAS signals. Nevertheless, predictions for SNP effects remain constrained by limitations in model resolution⁴⁹, the absence of multimodal epigenomic integration⁵⁰, and the complexity of chromatin folding *in vivo*, highlighting key areas for future methodological improvement.

Structural variants can reshape chromatin architecture far more extensively than single-nucleotide changes because they alter not only local sequence motifs but also the physical arrangement of large regulatory intervals^{5,17}. Consistent with this, our *in silico* SV perturbations revealed that many AD-associated deletions, and duplications induce substantial reconfiguration of chromatin interactions. However, interpreting the regulatory consequences of these SV-driven 3D changes is challenging, particularly for long SVs that span multiple enhancers, promoters, or even entire genes. For such variants, the predicted structural effects may reflect a mixture of direct sequence loss, altered chromatin spacing, and secondary changes in domain organization, making it difficult to attribute mechanistic specificity^{17,51}. To address this complexity, we focused mechanistic interpretation on two representative SVs whose genomic context allows clearer inference: one small intronic deletion within *KANSL1*, and one deletion overlapping the enhancer-rich *ARMS2–HTRA1* locus. In both cases, the SV does not delete the entire promoter or gene body, allowing observation of how altered 3D folding alone redirects enhancer contacts and shifts local domain structure. The *KANSL1* intronic deletion strengthened interactions around *ARL17A/B* and *LRRC37A*, consistent with localized rewiring rather than broad genomic collapse. Likewise, the *ARMS2*-

overlapping deletion increased interactions linking the enhancer cluster to the *PLEKHA1* promoter, suggesting a structural basis for observed microglia-specific transcriptional upregulation. These focused analyses illustrate how SVs may modulate 3D genome architecture in ways that are consistent with cell-type-specific regulatory changes in AD^{52,53}.

We acknowledged potential limitations of our *iNeuroAkita* model. For example, variant-driven chromatin changes are often context-dependent, which were influenced by cell-type-specific chromatin accessibility, motif strength, and local regulatory density^{46–48,54}. Also, SVs pose additional interpretive difficulty due to their larger size and multi-layered consequences on chromatin folding¹⁷. Moreover, while our models can predict structural rearrangements, determining the downstream transcriptional or functional impact requires integration with additional data modalities and experimental validation.

In the future, our *iNeuroAkit model* could be extended toward broader biological and translational applications. For instance, incorporating multi-modal epigenomic signals, such as DNA methylation, ATAC-seq, histone modifications, and transcription factor binding, could enhance cell type specificity and capture regulatory states more comprehensively, similar to emerging multi-modal models (e.g., C. Origami)^{50,54}. Second, integrating phasing or haplotype-resolved genome information could improve the ability to model allele-specific and diploid architecture, which is increasingly recognized as critical in complex disease genetics^{55–57}. Third, experimental validation through approaches such as CRISPR, promoter-capture Hi-C, or single-cell multi-omic assays will be essential to test the functional consequences of model-predicted

structural disruptions. Finally, our framework can be expanded to prioritize therapeutic targets by linking variant-driven 3D disruptions to downstream transcriptional changes in AD.

In summary, this study highlights the potential of cell-type-aware 3D genome modeling as a scalable and mechanistic framework for interpreting the consequences of non-coding AD risk variants and for guiding future functional genomics and therapeutic discovery in AD and other complex diseases.

Methods

Generation of NPCs from induced pluripotent stem cells (iPSCs)

iPSC cell lines derived from a female AD patient (iPSC # IUGB269.1, age 63), provided by the National Centralized Repository for Alzheimer's Disease and Related Dementias (NCRAD) at the Indiana University, were maintained on Matrigel-coated culture dishes in mTeSR1 medium (85850, STEMCELL Technologies) with 0.5% Antibiotic-Antimycotic (Lot#43, LRI Media Lab) in a 37 °C, 5% CO₂ supplemented incubator, and were passaged using ReLeSR (100-0484, STEMCELL Technologies).

For neural progenitor cells (NPCs), the cells from iPSC were dissociated with Accutane (GicoTMStemProTM AccutaseTM Cell Dissociation Reagent, A1110501, Fisher Scientific) and plated into Matrigel-coated 6-well tissue culture dish with a density of 1–1.5×10⁵ cells/cm². Cells were cultured in neural induction medium (05839, STEMCELL Technologies) in a 37°C, 5% CO₂ incubator to form monolayer neural progenitor cells (NPCs). These NPCs were passed on day 6, day 12 and day 18.

Generation of bulk Hi-C libraries from NPCs

Harvesting cells. Cells were harvested from culture at the NPC stage on day 18, by dissociating with Accutase for 5 mins, quenched with DMEM:F12, then basket centrifugation for 5 minutes at 500 x g. Pellet was undisturbed and washed once with 1XDPBS, resuspended in 2 mL 1XDPBS, and the volume was brought to 5 mL with 1XDPBS. Cell suspension was quantified using ThermoFisher Countess II cell counter. Two aliquots were made, into 15-mL conical tubes each, with 10 million cells respectively. For each tube, volume was brought to 5 mL with 1XDPBS. **Cross-linking cells.** 714 uL of 16% formaldehyde (ThermoFisher #28906) was added for a final formaldehyde concentration of 2%. Tubes were mixed by inversion 10 times then incubated at room temperature for 10 minutes. Arima-HiC Kit, A510008, was used for subsequent steps. 497 uL of Stop Solution 1 was added (for 12.5x dilution of glycine), mixed by inversion 10 times, and incubated at room temperature for 5 minutes. Sample was transferred to ice and incubated for 15 minutes. Cells were pelleted by basket centrifugation for 5 minutes at 500 x g. Supernatant was discarded. Pellet was resuspended in 1XDPBS and quantified using cell counter. **Estimating input amount.** Using Arima HiC protocol, we then determined that for 1 million cells, the output is 2.4 ug of DNA, and we used this as input for each reaction. Cells not used were aliquoted into 15-mL conical tubes, pelleted, and stored at -80°C. **Restriction enzyme digestion followed by DNA proximal ligation.** RE digestion and ligation were achieved with enzyme cocktail found in Arima HiC A510008, and protocol followed as instructed. **HiC Library preparation.** Arima Library Prep Module (A303011) were used for sequencing library preparation. **DNA fragmentation.** To limit sequence bias, was performed to

obtain an average fragment size of 400 bp. Covaris model S2 ultrasonicator was used, and protocol as indicated by manufacturer for target fragment size was followed. **DNA size selection.** Achieved using Beckman Coulter AMPure XP beads (A63881), conditions as instructed in Arima Hi-C protocol followed as indicated. **Biotin enrichment, end repair, dA tailing, adapter ligation, and sample indexing.** Achieved by following Arima Hi-C A303011 as instructed. Arima 7 primer pairs included in kit were utilized for unique sample indexing. **Next-generation sequencing.** Sequencing libraries were submitted to Azenta Life Sciences for next-generation sequencing and sequenced on an Illumina NovaSeq platform.

Hi-C data source and processing

Hi-C datasets for model training and evaluation were obtained from both in-house and publicly available sources. First, we generated NPC Hi-C data induced from iPSCs line derived from an AD donor. To complement this dataset, we also downloaded publicly available Hi-C data of NPCs from GEO accession GSE185192²⁴, which includes control NPCs from both male and female donors. Microglial Hi-C data were obtained from the Synapse collection (syn26161051)²⁵, providing chromatin architecture maps from healthy human primary microglia.

All Hi-C datasets were processed using the distiller-nf pipeline (<https://github.com/mirnylab/distiller-nf>)⁵⁸, following the recommended steps for read alignment, filtering, and contact matrix generation. Briefly, raw Hi-C reads were aligned to the hg38 reference genome, and alignment files were parsed to extract valid Hi-C pairs. PCR duplicates and low-quality pairs were removed using the default filtering parameters. Valid interaction pairs were then aggregated into fixed-size genomic bins to

generate binned Hi-C contact matrices. For this study, we constructed matrices at 2,048 bins per 1 Mb window, consistent with the input requirements of the Akita architecture²¹. Quality control (QC) metrics were computed for all Hi-C libraries, including the number of raw reads, uniquely mapped reads with mapping quality (MAPQ) ≥ 30 , and reads remaining after PCR duplicate removal, as well total intra-chromosomal contacts and partitioned them into short-range (<20 kb) and long-range (≥ 20 kb) interaction pairs (**Supplementary Table 1**). These QC statistics ensured that all datasets used for model training and evaluation met established Hi-C quality standards.

Model training, validation and prediction

The Akita model uses a convolutional neural network in which a Basenji-derived ‘trunk’ encodes genomic sequence features, followed by a 2D prediction head that outputs chromatic contact maps^{21,59}. We adopted this architecture on our processed Hi-C datasets from in-house NPCs and microglia, enabling the network to learn sequence–structure relationships to AD related cell types.

We prepared training, validation, and test inputs for the *iNeuroAkita* model using the *akita_data.py* pipeline (<https://github.com/calico/basenji/tree/master/manuscripts/akita>)²¹. For each dataset (in-house NPC Hi-C, microglia Hi-C) were randomly partitioned into 80% training, 10% validation, and 10% test, and an additional publicly available NPC dataset (GSE185192)²⁴ was used as an independent external test set.

To construct sequence–contact-map training examples, we generated 1,048,576-bp windows centered across the genome and excluded problematic regions using the ENCODE blacklist and the hg38 assembly-gap mask (**Supplementary data**). We

492 removed 65,536 bp from each end of the sequence window (“crop”) to avoid edge
493 artifacts, and computed interaction targets at 2,048-bp resolution with
494 observed/expected normalization²¹. To increase the number of training samples and
495 improve robustness, we slide 1Mb windows across the genome using a 262,144-bp
496 stride for training and 524,288-bp stride for validation/testing. We also applied Gaussian
497 smoothing (-k 1) and aligned binned targets using `--snap 2,048`²¹.

498 These processed datasets were used to train and evaluate *iNeuroAkita* model,
499 initialized from the pretrained Akita parameters. To optimize the training, we used
500 Dragonfly, an open-source Bayesian optimization framework
501 (<https://github.com/dragonfly/dragonfly>)²⁶, to explore ranges of key hyperparameters—
502 including learning rate, momentum, batch size, and dropout. Dragonfly-guided tuning
503 yielded the final training configuration, including the selected batch size, gradient-
504 clipping settings, and the modified 2D prediction head that provided improved
505 performance.

506 After training, each model was evaluated on held-out test sequences and on an
507 independent external NPC dataset. Predicted contact vectors were converted into full
508 symmetric contact matrices and smoothed with a Gaussian kernel ($\sigma = 2$ bins) to
509 increase signal-to-noise robustness. Model accuracy was quantified using off-diagonal
510 mean squared error (MSE) and Pearson correlation coefficient (PCC) between
511 predicted and experimental contact maps, computed both (i) globally (pooled across all
512 windows) and (ii) on a per-sequence basis to obtain correlation profiles across genomic
513 tiles with 1Mb length.

CTCF motif mutagenesis

We identified genomic CTCF motif instances by scanning the hg38 reference genome with a curated CTCF position-weight matrix from JASPAR (2024)²⁸. Motif scanning was performed using FIMO (MEME suite)⁶⁰ against the hg38 reference genome with a default background model and a liberal detection threshold ($p\text{-value} \leq 1 \times 10^{-4}$). We then applied a more stringent filter on the FIMO output, retaining only CTCF motif hits with $q\text{-value} < 0.05$ for downstream analyses. To determine which CTCF-bound sites fell within the genomic windows used for evaluating *iNeuroAkita* model, we intersected the CTCF coordinates with test datasets generated previously. For each intersecting window, the corresponding 1-Mb sequence related to the test datasets named as wild-type (WT) was extracted from the hg38 reference genome using `bedtools getfasta`⁶¹.

For windows containing at least one CTCF site, the absolute CTCF coordinates were converted to positions relative to the 1-Mb sequence. Within each ~10-bp CTCF core motif, we introduced random single-nucleotide substitutions, replacing each reference base with one of the three alternative nucleotides (A/C/G/T excluding the reference base), sampled uniformly at random, while preserving motif length and genomic coordinates. Ten independent mutated sequences were generated per CTCF-containing window, producing ten mutated (MUT) sequences for downstream *in silico* perturbation analyses.

Both WT and mutated sequences were evaluated using our fine-tuned Akita models. To quantify the structural consequences of CTCF motif perturbation, WT and MUT contact maps were compared within each window by computing (i) per-pixel squared differences, (ii) per-target disruption scores $\text{mean}((\text{WT} - \text{MUT})^2)$, and (iii)

changes in overall signal intensity $\text{mean}(\text{WT}^2) - \text{mean}(\text{MUT}^2)$. The mean value was calculated for genomic bins at measured genomic regions. Disruption scores were averaged across targets to obtain a single metric per region. Regions were then ranked to identify CTCF sites whose mutation caused the strongest changes in predicted 3D chromatin structure.

To independently corroborate CTCF occupancy in NPCs, we also downloaded and processed NPC CTCF ChIP-seq data from ENCODE (accession ENCFF782OKU). These experimentally defined peaks were used to confirm that our disrupted motifs corresponded to real CTCF-binding sites.

SNV mutagenesis

We compiled a set of 42 high-confidence dementia-associated functional variants from previously published study¹² (**Supplementary Table S4**) and integrated microglia-specific fine-mapped eQTLs⁶² (**Supplementary data**). For each variant, the corresponding wild-type and alternative alleles were retrieved from the dbSNP v155 database⁶³ for hg38. Genomic sequences were then extracted and symmetrically extended around the variant position to a total length of 1,048,576 bp, matching the input requirements of our 3D genome model. Both the wild-type and variant-centered sequences were processed through the fine-tuned models to generate paired predictions of 3D chromatin contact maps for downstream comparison.

To quantify variant-induced changes in chromatin folding, we computed several complementary disruption metrics on the MUT-WT difference map: (i) L2 Frobenius norm (L2_fro), (ii) mean squared error (MSE), (iii) mean absolute deviation (L1_mean), and (iv) maximum absolute difference (MaxAbs). For visualization, we produced signal-

weighted difference maps in which MUT–WT values are modulated by the local contact intensity (max of |WT| and |MUT|), emphasizing changes in high-signal regions.

Structure variants mutagenesis

We collected high-confidence AD and ADRD structural variants from two published studies^{19,20} (**Supplementary Table S7**) and processed them for *in silico* perturbation using SuPreMo³³, a computational framework designed to perform sequence-level perturbations. SuPreMo supports only specific SV classes —deletions (DEL), duplications (DUP), inversions (INV), and breakpoints (BND), we therefor filtered the initial collection accordingly, resulting in 120 variants suitable for analysis: 100 DEL, 18 DUP, and 2 INV. All SV coordinates were converted to the hg38 assembly using UCSC liftOver to maintain consistency with the reference genome and our fine-tuned models' requirements.

In silico saturation mutagenesis (ISM) for each SV was performed using SuPreMo³³, which generates WT and SV-altered genomic sequences, constructs 1-Mb sequence windows centered on the variant, and optionally produces predicted 3D contact maps via user-specified models. SuPreMo was executed with a ± 1 bp shift to account for local sequence context and with reverse-complement augmentation enabled.

Epigenomic data integration and annotation

We downloaded ATAC-seq, H3K27ac, and H3K4me3 datasets for neurons and microglia⁶⁴, all provided in hg19 genomic coordinates. For visualization of regulatory context, we extracted the top 0.5% strongest interaction gains or losses from both SNV- and SV-based *in silico* mutagenesis analyses. To ensure coordinate compatibility with

the epigenomic datasets, CTCF peaks, SNVs, and SVs were converted to hg19 using UCSC liftOver. These annotations, together with the predicted interaction changes, were visualized using pyGenomeTracks⁶⁵, enabling joint display of chromatin accessibility, histone modification profiles, gene models, and predicted 3D structural perturbations.

Gene expression analysis in SEA-AD single-nucleus RNA-seq

Single-nucleus RNA-seq (snRNA-seq) data from the middle temporal gyrus (MTG) were obtained from the Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD)³². Data preprocessing and normalization were performed using the Scanpy Python package⁶⁶. Metadata were standardized by harmonizing cognitive-status labels into two groups (*No_dementia* and *Dementia*) and collapsing fine-grained cell subclasses into six major cell-type categories: excitatory neurons (Ext), inhibitory neurons (Inh), microglia (Mic), astrocytes (Ast), oligodendrocyte precursor cells (OPC), and oligodendrocytes (Oli)⁶⁷. Expression values correspond to log-transformed counts per 10,000 ($\log(1+CP10K)$) provided in the processed dataset.

For each gene, per-cell normalized expression values were used together with donor and condition annotations to compute differential expression. Within each cell type, donor-level median expression was first calculated for AD-associated Dementia and *No_dementia* groups, and differential expression was summarized as the $\log_2$ fold change ($\log_2FC$) between the two conditions. The standard error (SE) of $\log_2FC$ was derived from donor-level variability. Differential patterns were visualized using $\log_2FC$ plots with SE error bars, and genes were categorized as Upregulated ($\log_2FC > 0.5$), Downregulated ($\log_2FC < -0.5$), or Modest change otherwise.

Data availability

The in-house NPC Hi-C data has been deposited in the NCBI Gene Expression Omnibus (GEO) database under accession number GSE319798. Independent NPC Hi-C data used for external evaluation were obtained from GSE185192²⁴, and primary microglia Hi-C data were downloaded from syn26161051⁶². Processed Hi-C contact matrices at 2,048-bp resolution are provided in the **Supplementary data** (<https://zenodo.org/records/18662341>). The set of 42 high-confidence functional variants associated with dementia was obtained from a previously published study¹² (**Supplementary Table S5**). Microglia eQTL datasets used for variant interpretation are reported in **Supplementary Table S6**. Structural variants (SVs) analyzed in this study were collected from two published AD/ADRD genomic studies^{19,20} (**Supplementary Table S8**). CTCF motif datasets were downloaded from the JASPAR (2024)²⁸, and NPC CTCF ChIP-seq data were obtained from the ENCODE Project (accession ENCFF782OKU). The ENCODE blacklist regions and the hg38 assembly-gap mask used for model input data generation are included in the **Supplementary data**. Processed ATAC-seq and ChIP-seq (H3K27ac and H3K4me3) datasets for neurons and microglia were downloaded from the GLASS Brain Cell Types resource (https://genome.ucsc.edu/s/nottalexi/glassLab_BrainCellTypes_hg19). Single-nucleus RNA-seq data from the Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD)³² were accessed from: https://sea-ad-single-cell-profiling.s3.amazonaws.com/index.html#MTG/RNAseq/SEAAD_MTG_RNAseq_final-nuclei.2024-02-13.h5ad.

Code availability

The source code is available on GitHub at <https://github.com/ChengF-Lab/iNeuroAkita>.

References

1. Frydas, A., Wauters, E., van der Zee, J. & Van Broeckhoven, C. Uncovering the impact of noncoding variants in neurodegenerative brain diseases. *Trends in Genetics* **38**, 258–272 (2022).
2. Cuyvers, E. & Sleegers, K. Genetic variations underlying Alzheimer's disease: evidence from genome-wide association studies and beyond. *Lancet Neurol* **15**, 857–868 (2016).
3. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates A β , tau, immunity and lipid processing. *Nat Genet* **51**, 414–430 (2019).
4. Bellenguez, C. *et al.* New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat Genet* **54**, 412–436 (2022).
5. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat Rev Genet* **19**, 453–467 (2018).
6. Orozco, G., Schoenfelder, S., Walker, N., Eyre, S. & Fraser, P. 3D genome organization links non-coding disease-associated variants to genes. *Front Cell Dev Biol* **10**, 995388 (2022).
7. Nott, A. *et al.* Brain cell type-specific enhancer–promoter interactome maps and disease-risk association. *Science* **366**, 1134–1139 (2019).

8. Zeng, B. *et al.* Genetic regulation of cell type–specific chromatin accessibility shapes brain disease etiology. *Science* **384**, eadh4265 (2024).
9. Corradin, O. & Scacheri, P. C. Enhancer variants: evaluating functions in common disease. *Genome Med* **6**, 85 (2014).
10. Green, G. S. *et al.* Cellular communities reveal trajectories of brain ageing and Alzheimer’s disease. *Nature* **633**, 634–645 (2024).
11. Sun, N. *et al.* Human microglial state dynamics in Alzheimer’s disease progression. *Cell* **186**, 4386–4403.e29 (2023).
12. Cooper, Y. A. *et al.* Functional regulatory variants implicate distinct transcriptional networks in dementia. *Science* **377**, eabi8654 (2022).
13. Liu, S. *et al.* Cell type–specific 3D-genome organization and transcription regulation in the brain. *Science Advances* **11**, eadv2067 (2025).
14. Logeman, B. L., Grieco, S. F., Holmes, T. C. & Xu, X. Unfolding neural diversity: how dynamic three-dimensional genome architecture regulates brain function and disease. *Mol Psychiatry* **30**, 3781–3791 (2025).
15. Li, Y. E. *et al.* A comparative atlas of single-cell chromatin accessibility in the human brain. *Science* **382**, eadf7044 (2023).
16. Schwartzentruber, J. *et al.* Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer’s disease risk genes. *Nat Genet* **53**, 392–402 (2021).
17. Sreenivasan, V. K. A., Yumiceba, V. & Spielmann, M. Structural variants in the 3D genome as drivers of disease. *Nat Rev Genet* 1–19 (2025) doi:10.1038/s41576-025-00862-x.

- 672 18. Kaivola, K. *et al.* Genome-wide structural variant analysis identifies risk loci for
673 non-Alzheimer's dementias. *Cell Genomics* **3**, 100316 (2023).
- 674 19. Wang, H. *et al.* Structural variation detection and association analysis of whole-
675 genome-sequence data from 16,543 Alzheimer's disease sequencing project
676 subjects. *Alzheimer's & Dementia* **21**, e70277 (2025).
- 677 20. Vialle, R. A. *et al.* Structural variants linked to Alzheimer's disease and other
678 common age-related clinical and neuropathologic traits. *Genome Medicine* **17**, 20
679 (2025).
- 680 21. Fudenberg, G., Kelley, D. R. & Pollard, K. S. Predicting 3D genome folding from
681 DNA sequence with Akita. *Nat Methods* **17**, 1111–1117 (2020).
- 682 22. Schwessinger, R. *et al.* DeepC: predicting 3D genome folding using megabase-
683 scale transfer learning. *Nat Methods* **17**, 1118–1124 (2020).
- 684 23. Emani, P. S. *et al.* Single-cell genomics and regulatory networks for 388 human
685 brains. *Science* **384**, eadi5199 (2024).
- 686 24. Meharena, H. S. *et al.* Down-syndrome-induced senescence disrupts the nuclear
687 architecture of neural progenitors. *Cell Stem Cell* **29**, 116-130.e7 (2022).
- 688 25. Kosoy, R. *et al.* Genetics of the human microglia regulome refines Alzheimer's
689 disease risk loci. *Nat Genet* **54**, 1145–1154 (2022).
- 690 26. dragonfly/dragonfly. dragonfly (2025).
- 691 27. Rahman, S. & Roussos, P. The 3D Genome in Brain Development: An
692 Exploration of Molecular Mechanisms and Experimental Methods. *J Exp Neurosci* **19**,
693 26331055241293455 (2024).

- 694 28. Rauluseviciute, I. *et al.* JASPAR 2024: 20th anniversary of the open-access
695 database of transcription factor binding profiles. *Nucleic Acids Res* **52**, D174–D182
696 (2024).
- 697 29. Low, Y. L., Jin, L., Morris, E. R., Pan, Y. & Nicolazzo, J. A. Pioglitazone
698 Increases Blood-Brain Barrier Expression of Fatty Acid-Binding Protein 5 and
699 Docosahexaenoic Acid Trafficking into the Brain. *Mol Pharm* **17**, 873–884 (2020).
- 700 30. Reizes, O., Benoit, S. C. & Clegg, D. J. The role of syndecans in the regulation of
701 body weight and synaptic plasticity. *The International Journal of Biochemistry & Cell*
702 *Biology* **40**, 28–45 (2008).
- 703 31. Gorokhova, S., Bibert, S., Geering, K. & Heintz, N. A novel family of
704 transmembrane proteins interacting with β subunits of the Na,K-ATPase. *Hum Mol*
705 *Genet* **16**, 2394–2410 (2007).
- 706 32. Gabitto, M. I. *et al.* Integrated multimodal cell atlas of Alzheimer’s disease. *Nat*
707 *Neurosci* 1–18 (2024) doi:10.1038/s41593-024-01774-5.
- 708 33. Gjoni, K. & Pollard, K. S. SuPreMo: a computational tool for streamlining in silico
709 perturbation using sequence-based predictive models. *Bioinformatics* **40**, btae340
710 (2024).
- 711 34. Zhai, Y. *et al.* Identification of JAZF1, KNOP1, and PLEKHA1 as causally
712 associated genes and drug targets for Alzheimer’s disease: a summary data-based
713 Mendelian randomization study. *Inflammopharmacol* **32**, 3913–3923 (2024).
- 714 35. de Wit, E. *et al.* CTCF Binding Polarity Determines Chromatin Looping. *Molecular*
715 *Cell* **60**, 676–684 (2015).

- 716 36. Wutz, G. *et al.* Topologically associating domains and chromatin loops depend
717 on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J* **36**,
718 3573–3599 (2017).
- 719 37. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution
720 reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
- 721 38. Rohs, R. *et al.* The role of DNA shape in protein–DNA recognition. *Nature* **461**,
722 1248–1253 (2009).
- 723 39. Friedman, M. J., Wagner, T., Lee, H., Rosenfeld, M. G. & Oh, S. Enhancer–
724 promoter specificity in gene transcription: molecular mechanisms and disease
725 associations. *Exp Mol Med* **56**, 772–787 (2024).
- 726 40. Razin, S. V., Ulianov, S. V. & Iarovaia, O. V. Enhancer Function in the 3D
727 Genome. *Genes (Basel)* **14**, 1277 (2023).
- 728 41. Kim, S. & Shendure, J. Mechanisms of Interplay between Transcription Factors
729 and the 3D Genome. *Molecular Cell* **76**, 306–319 (2019).
- 730 42. Nie, Y., Shu, C. & Sun, X. Cooperative binding of transcription factors in the
731 human genome. *Genomics* **112**, 3427–3434 (2020).
- 732 43. Johnston, A. D., Simões-Pires, C. A., Thompson, T. V., Suzuki, M. & Greally, J.
733 M. Functional genetic variants can mediate their regulatory effects through alteration
734 of transcription factor binding. *Nat Commun* **10**, 3472 (2019).
- 735 44. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the
736 regulatory epigenome. *Nat Rev Genet* **20**, 207–220 (2019).
- 737 45. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat Rev*
738 *Genet* **17**, 661–678 (2016).

- 739 46. Liu, S. *et al.* Cell type–specific 3D-genome organization and transcription
740 regulation in the brain. *Science Advances* **11**, eadv2067 (2025).
- 741 47. Song, M. *et al.* Cell-type-specific 3D epigenomes in the developing human
742 cortex. *Nature* **587**, 644–649 (2020).
- 743 48. Winick-Ng, W. *et al.* Cell-type specialization is encoded by specific chromatin
744 topologies. *Nature* **599**, 684–691 (2021).
- 745 49. Li, H. *et al.* Mapping chromatin structure at base-pair resolution unveils a unified
746 model of cis-regulatory element interactions. *Cell* **0**, (2025).
- 747 50. Avsec, Ž. *et al.* AlphaGenome: advancing regulatory variant effect prediction with
748 a unified DNA sequence model.
- 749 51. Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause
750 pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
- 751 52. Bai, J. *et al.* High-throughput in silico screen uncovers key regulators of 3D
752 genome architecture. 2025.12.09.693120 Preprint at
753 <https://doi.org/10.64898/2025.12.09.693120> (2025).
- 754 53. Dekker, J. *et al.* An integrated view of the structure and function of the human 4D
755 nucleome. *Nature* 1–18 (2025) doi:10.1038/s41586-025-09890-3.
- 756 54. Tan, J. *et al.* Cell-type-specific prediction of 3D chromatin organization enables
757 high-throughput in silico genetic screening. *Nat Biotechnol* **41**, 1140–1150 (2023).
- 758 55. Accessible technology for diploid high-order 3D human genome analysis. *Nat*
759 *Struct Mol Biol* **32**, 1141–1142 (2025).
- 760 56. Tu, X. *et al.* Deep genomic models of allele-specific measurements.
761 2025.04.09.648060 Preprint at <https://doi.org/10.1101/2025.04.09.648060> (2025).

57. Atak, Z. K. *et al.* Interpretation of allele-specific chromatin accessibility using cell state-aware deep learning. *Genome Res* **31**, 1082–1096 (2021).
58. open2c/distiller-nf. Open Chromosome Collective (2025).
59. Kelley, D. R. *et al.* Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res* **28**, 739–750 (2018).
60. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
61. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
62. Kosoy, R. *et al.* Genetics of the human microglia regulome refines Alzheimer's disease risk loci. *Nat Genet* **54**, 1145–1154 (2022).
63. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308–311 (2001).
64. Nott, A. *et al.* Brain cell type-specific enhancer–promoter interactome maps and disease-risk association. *Science* **366**, 1134–1139 (2019).
65. Lopez-Delisle, L. *et al.* pyGenomeTracks: reproducible plots for multivariate genomic datasets. *Bioinformatics* **37**, 422–423 (2021).
66. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19**, 15 (2018).
67. Ren, Y. *et al.* Single-cell digital twins identify drug targets and repurposable medicine in Alzheimer's disease. 2025.01.14.633081 Preprint at <https://doi.org/10.1101/2025.01.14.633081> (2025).

Acknowledgements

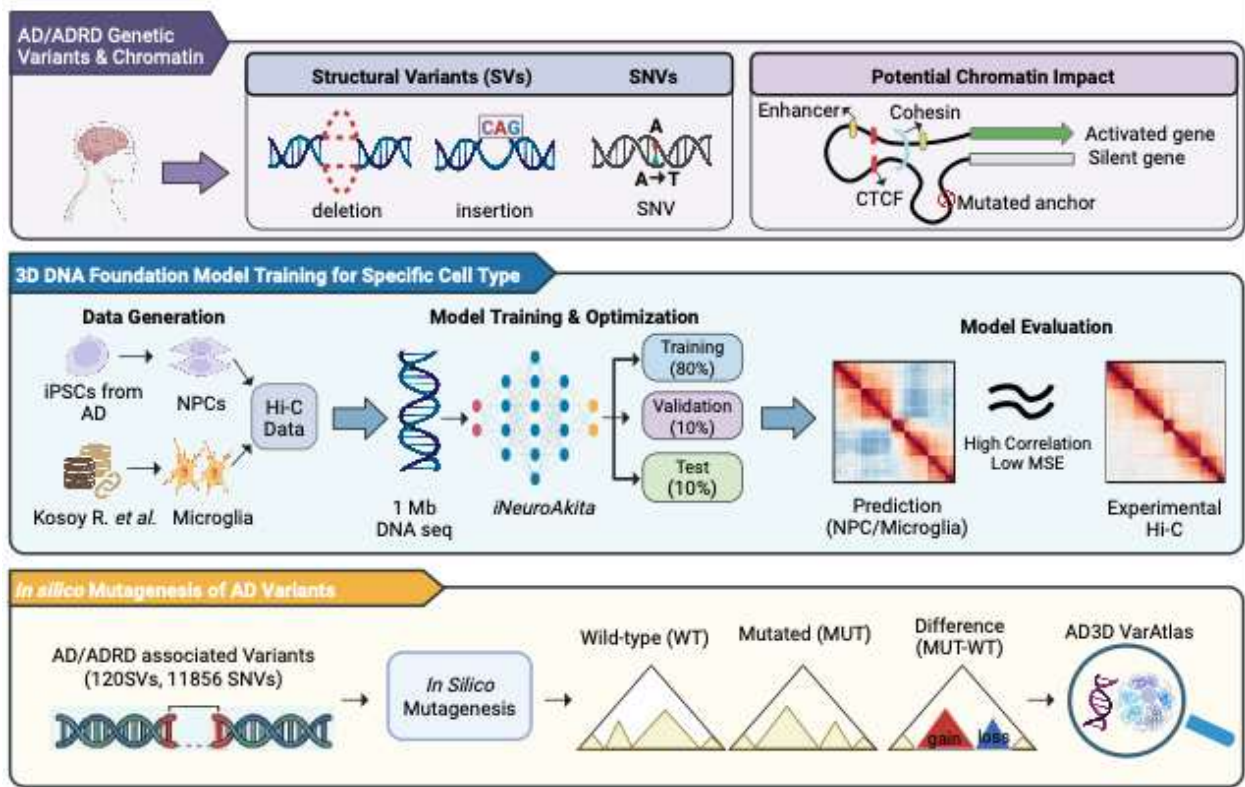
This work was supported by the National Institute on Aging (NIA) under Award Number U01AG073323, R01AG066707, R01AG084250, R01AG076448, R01AG082118, R01AG092462, R01AG092591, RF1AG082211, and R33AG083003, the National Institute of Neurological Disorders and Stroke (NINDS) under Award Number RF1NS133812, National Institutes of Health (NIH) under Award Number 1OT2OD038083-01, the Alzheimer's Association award (ALZDISCOVERY-1051936), the Alzheimer's Disease Drug Discovery Foundation (ADDF), and Dr. Keyhan and Dr. Jafar Mobasseri Endowed Chair for Innovative Research to F.C. This work was supported in part by the Rebecca E. Barchas, MD, Professorship in Translational Psychiatry, the Valour Foundation, Project 19PABH134580006-AHA/Allen Initiative in Brain Health and Cognitive Impairment, the Elizabeth Ring Mather & William Gwinn Mather Fund, S. Livingston Samuel Mather Trust, and the Louis Stokes VA Medical Center resources and facilities to A.A.P. M.H. was partially supported by the NIH grant R35HG011922.

Author's contribution

F.C. conceived the study. Y.R. designed and performed all data analyses. X.Y., M.H., Y.E.L., Y.H., and F.J. interpreted the data analysis. M.R. R., and Y. L. performed cell culture and differentiation and Hi-C sequencing library construction. Y.R. drafted the manuscript. Y.R., M.H., Y.E.L., and F.C. critically revised the manuscript. All authors gave final approval of the manuscript.

Competing interests: The authors declare no competing interests.

808 **Main Figures**



809
810 **Figure 1: Schematic overview of this study.**

811 Overview of the study design integrating Alzheimer's disease (AD)-associated genetic
812 variants with fine-tuned 3D genome modeling. Top: AD/ADRD risk variants, including
813 structural variants and SNPs, were analyzed for their potential to alter chromatin
814 architecture. Middle: A 3D DNA foundation model was fine-tuned using Hi-C data from
815 neuronal progenitor cells (NPCs) and microglia. The model was trained and validated
816 on 1 Mb genomic sequences to predict 3D contact maps that accurately recapitulate
817 experimental Hi-C measurements. Bottom: In silico mutagenesis (ISM) was performed
818 to simulate the impact of AD-associated variants on chromatin folding. Predicted wild-
819 type and mutant 3D maps were compared to quantify structural disruptions and
820 compiled into the AD related 3D atlas altered by variants (AD3D VarAtlas), a reference
821 for variant-induced 3D genome alterations in Alzheimer's disease. Figure created with
822 BioRender.com.

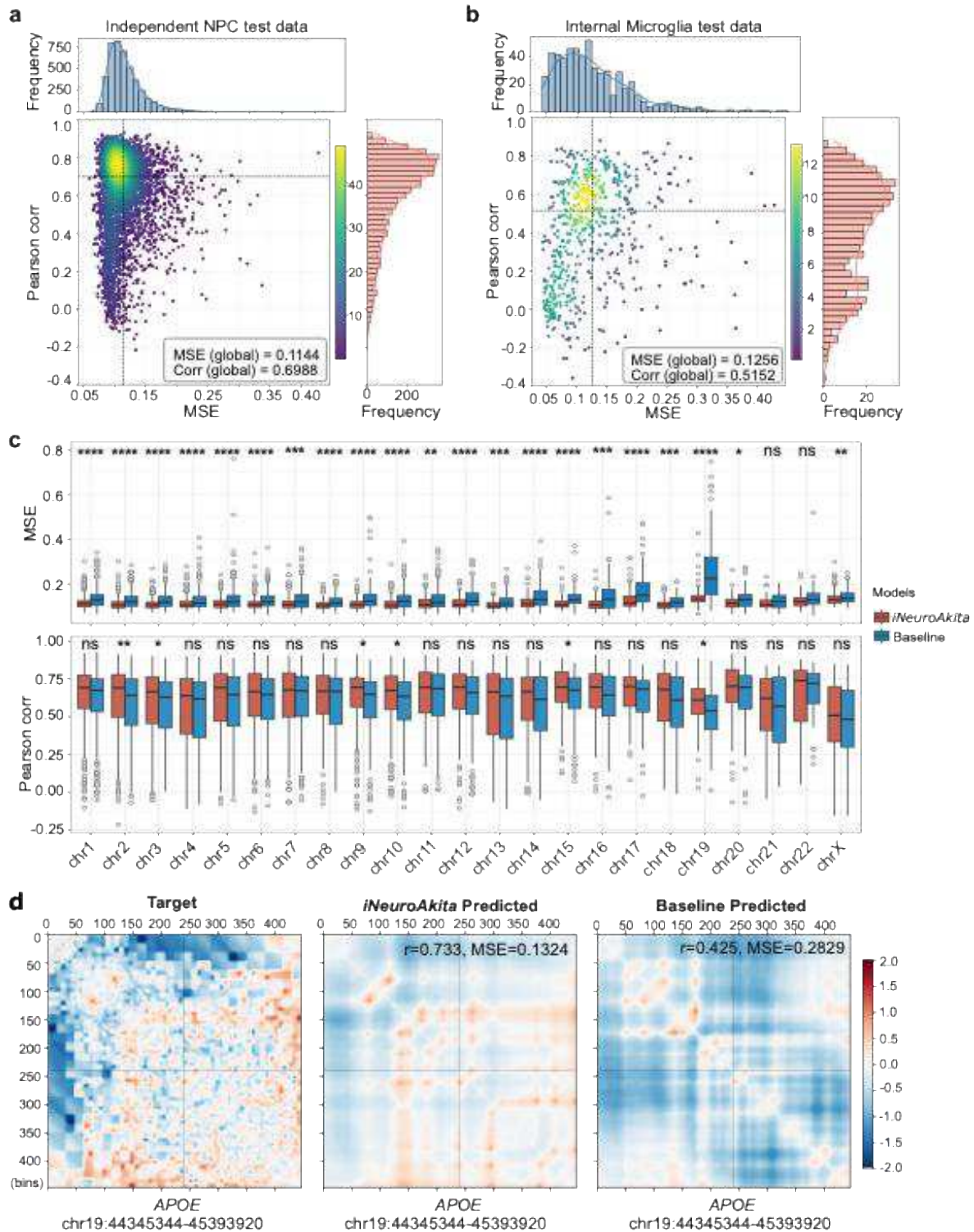


Figure 2: Overall performance evaluation of *iNeuroAkita* model on neuronal progenitor cells (NPCs) and microglia cells.

826 **(a)** Evaluation of the *iNeuroAkita* model on an independent NPC test dataset, showing
827 prediction accuracy related to MSE and Pearson correlation across genomic regions.
828 **(b)** Performance of the *iNeuroAkita* model on held-out microglia test data. **(c)**
829 Comparison between the *iNeuroAkita* model and the baseline Akita model on the
830 independent NPC dataset related to MSE (up) and Pearson correlation (down). **(d)**
831 Example of predicted by *iNeuroAkita* model, and baseline model versus experimental
832 contact maps of one region from the independent NPC test data.

833

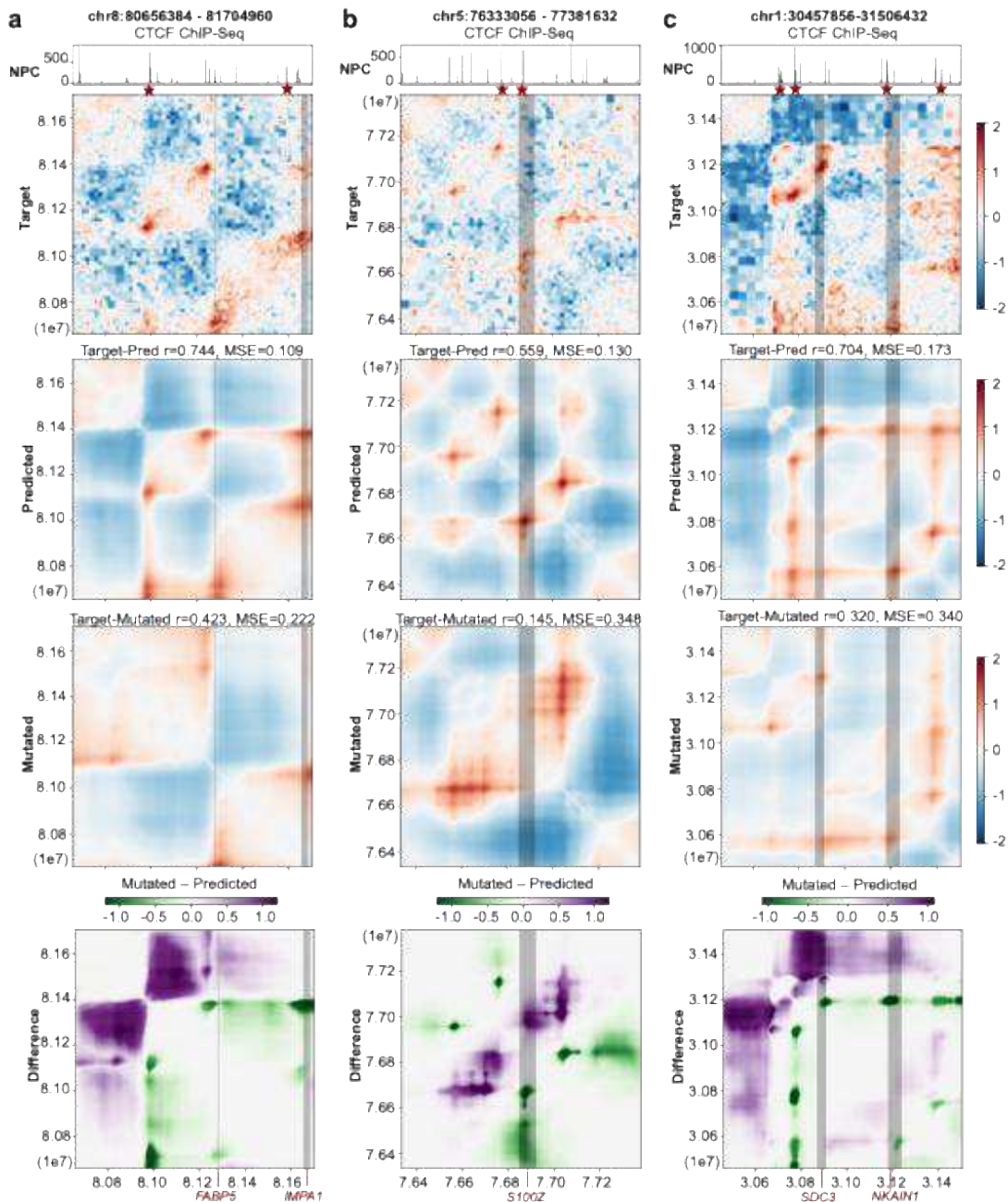


Figure 3: CTCF motif disruption alters 3D chromatin organization as predicted by the *iNeuroAkita* model

(a–c) Representative loci showing the impact of *in silico* CTCF motif mutation on chromatin contact patterns. Each panel includes, from top to bottom: NPC CTCF ChIP-seq signal, experimental Hi-C contact map, *iNeuroAkita* model prediction, and predicted contact map after *in silico* CTCF mutation, and the weighted differential map (DIFF = Mutated – Predicted). Red stars mark the positions of the mutated CTCF motif(s). Gray shaded vertical regions highlight the genomic intervals encompassing nearby genes of interest within each locus.

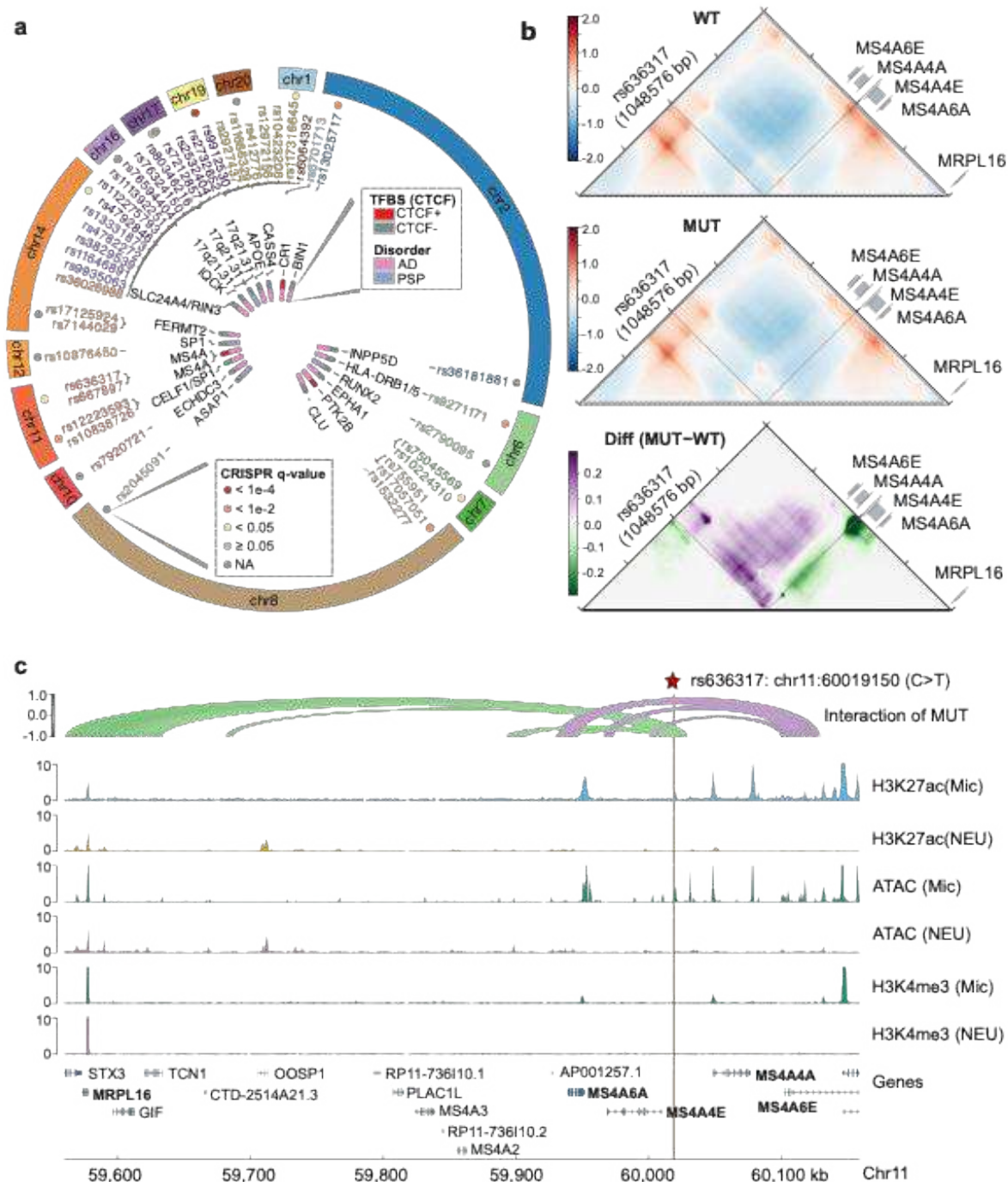


Figure4: *iNeuroAkita* model identifies chromatin interaction disruption caused by dementia-associated SNVs.

(a) CRISPR-validated dementia-associated SNPs used for 3D genome structure prediction. Forty-two functional variants identified in previous CRISPR screening studies

849 were analyzed using *iNeuroAkita* model. Variants are annotated by chromosome, CTCF
850 binding and associated disorder category (e.g., AD, PSP). **(b)** Predicted 3D contact
851 maps in 1M length for the wild type (WT) and mutated (MUT) alleles of the SNP
852 (rs636317, chr11:60,019,150 (hg19), C>T). The bottom panel displays the weighted
853 differential map ($\text{DIFF} = \text{MUT} - \text{WT}$), summarizing mutation-induced contact changes
854 using a signal-weighted difference metric that emphasizes structurally meaningful gains
855 (red) and losses (blue) of interactions. **(c)** Integrative epigenomic view of the disrupted
856 region. The upper track displays predicted gains (red arcs) and losses (blue arcs) of
857 chromatin interactions after mutation, followed by ChIP-seq (H3K27ac, H3K4me3) and
858 ATAC-seq tracks from microglia (Mic) and neurons (NEU). Gene annotations at the
859 bottom highlight candidate mutation-responsive target genes—including *MS4A4A*,
860 *MS4A4E*, and *MS4A6A*—that reside within altered 3D contacts.

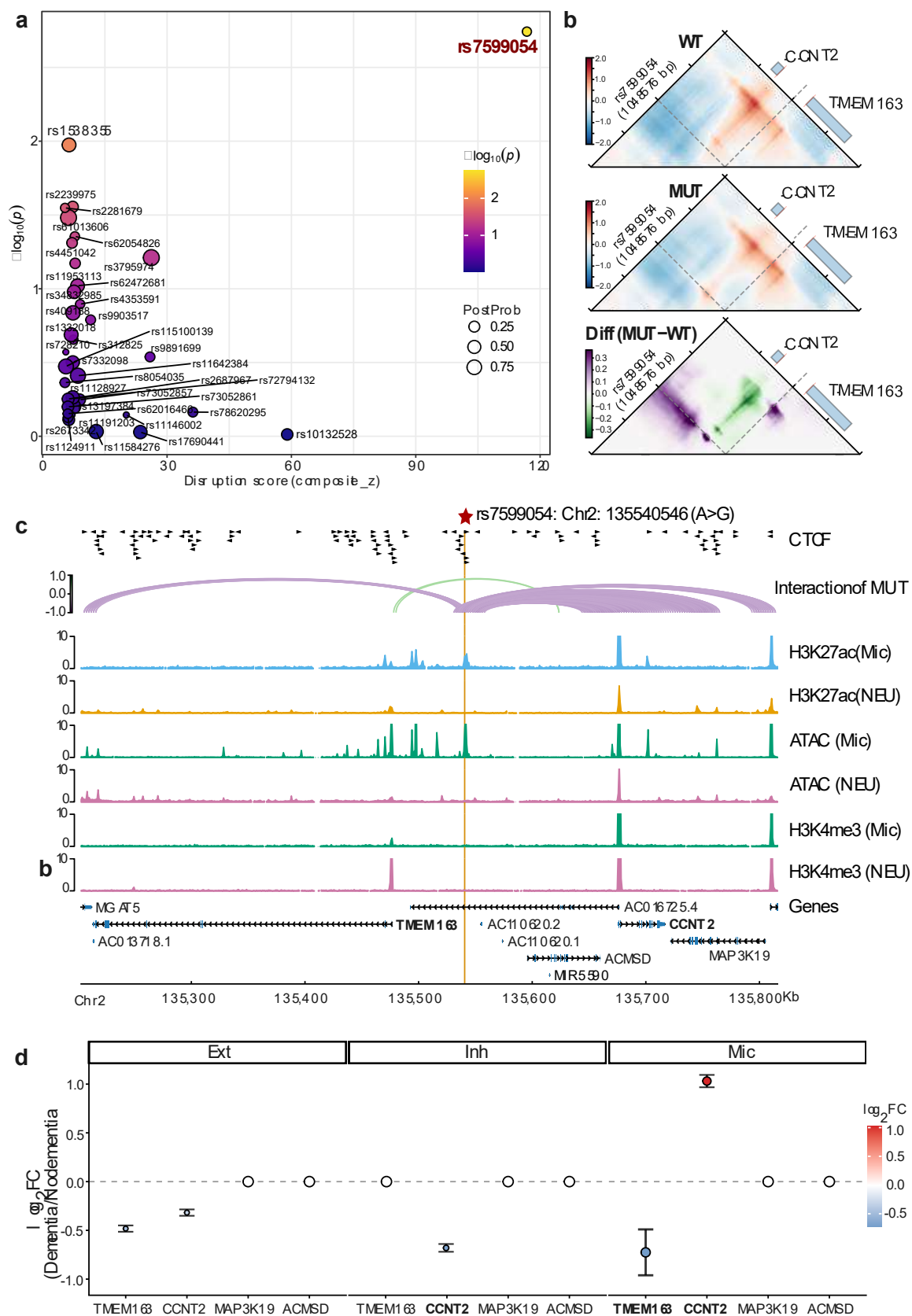


Figure 5: Microglia fine-mapped eQTL variant rs7599054 exhibits strong 3D chromatin disruption and altered gene regulation

(a) Top-disruptive variants from microglia fine-mapped eQTLs. Scatter plot showing the relationship between 3D chromatin disruption score (x-axis) and GWAS $-\log_{10}(p)$ value (y-axis) for variants derived from microglia fine-mapped eQTL data. Variant rs7599054 shows the highest composite disruption score and significant disease association. **(b)** Predicted 3D chromatin contact maps for the rs7599054 locus. Wild type (WT) and mutated (MUT) allele predictions from the *iNeuroAkita* model in 1Mb region. The bottom panel displays the weighted differential interaction map (DIFF = MUT – WT), which highlights structurally meaningful contact gains (red) and losses (blue) by incorporating signal magnitude into the difference calculation. **(c)** Integrative epigenomic view of the disrupted region. The upper panel shows predicted gain (red) and loss (blue) chromatin interaction arcs following mutation with CTCF binding information. Below are ChIP-seq and ATAC-seq tracks for microglia (Mic) and neurons (NEU), including H3K27ac and H3K4me3 profiles, illustrating active regulatory elements overlapping the disrupted loops and candidate target genes. **(d)** Variant-associated gene expression changes in dementia vs. non-dementia samples. Single-cell RNA-seq data from SEA-AD show differential expression of nearby genes (*TMEM163*, *CCNT2*, *MAP3K19*, *ACMSD*) across excitatory (Ext), inhibitory (Inh), and microglial (Mic) populations, highlighting cell-type-specific transcriptional effects linked to the disrupted locus.

Figure 6: A structural variant at the *KANSL1* locus disrupts 3D chromatin organization and alters neuronal gene regulation.

(a) Predicted 3D chromatin contact maps for the *KANSL1* structural variant (DEL-314). Wild type (WT) and mutated (MUT) allele predictions from the *iNeuroAkita* model show a large local reorganization of chromatin topology at chr17:44,223,907–44,224,221(hg19) (314 bp deletion). **(b)** Integrative epigenomic view of the disrupted *MAPT* region. Predicted gain (red) and loss (blue) chromatin interactions after the deletion are overlaid with CTCF binding, H3K27ac, H3K4me3, and ATAC-seq profiles from microglia (Mic) and neurons (NEU). The affected region overlaps regulatory elements and genes including *ARL17A*, *ARL17B*, and *LRRC37A*. **(c)** Cell-type-specific expression changes in dementia versus non-dementia samples. Single-cell RNA-seq data from the SEA-AD cohort show altered expression of nearby genes (*ARL17A*, *ARL17B*, *LRRC37A*, *MAPT*, *KANSL1*) across excitatory (Ext), inhibitory (Inh), and microglial (Mic) populations, highlighting possible downstream transcriptional effects of the *MAPT*-linked structural variant.

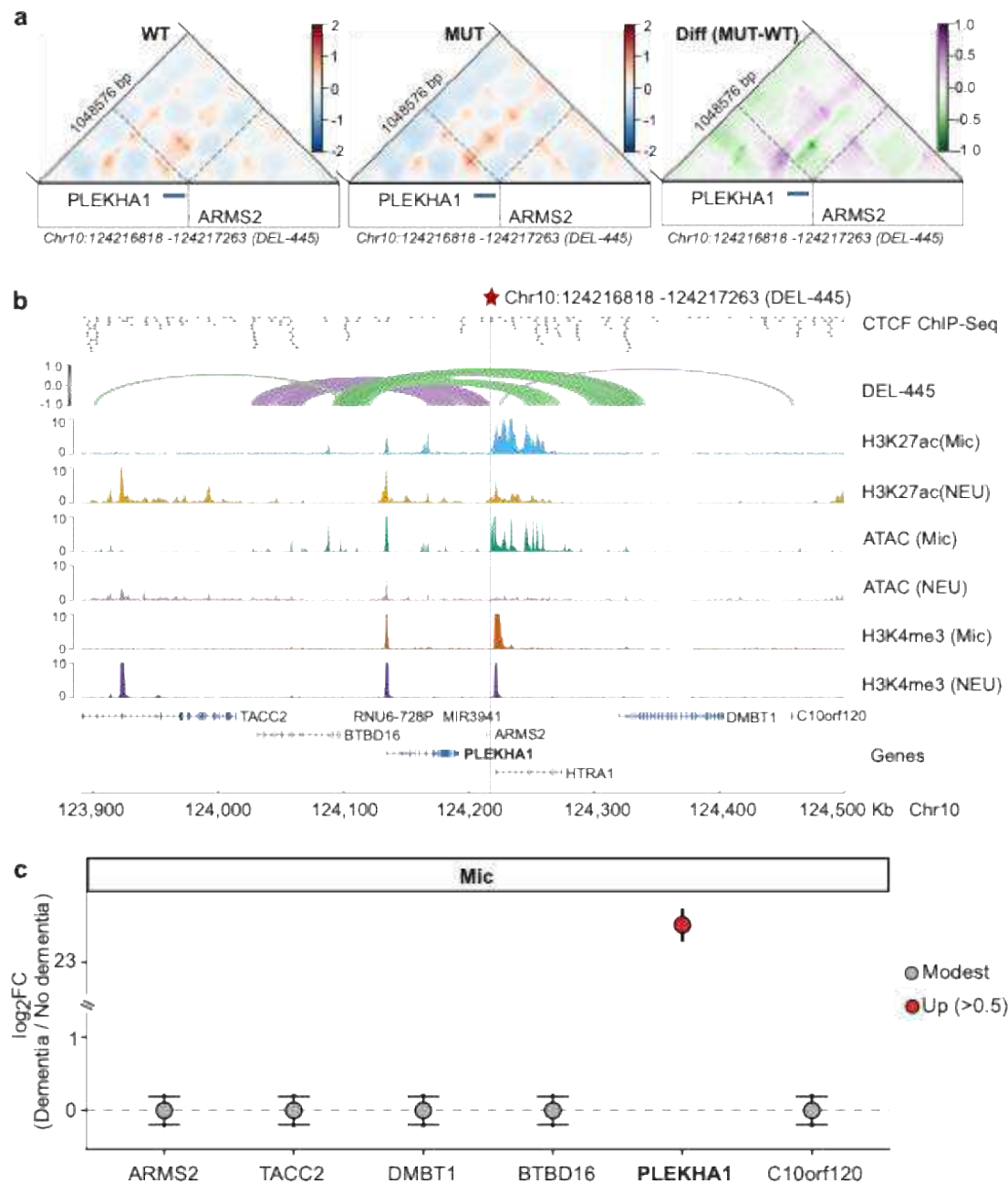


Figure 7. A structural variant near the *ARMS2* locus disrupts 3D chromatin architecture and alters gene expression.

(a) Predicted 3D chromatin contact maps for the *ARMS2* structural variant (DEL-445). Reference (REF) and mutated (MUT) allele predictions from the fine-tuned 3D genome model show local reorganization of chromatin loops at chr10:124,216,818–124,217,263 (hg19) (445 bp deletion). **(b)** Integrative epigenomic landscape of the disrupted region.

Predicted gain (red) and loss (blue) chromatin interactions following the deletion are displayed together with CTCF binding and epigenomic profiles. ChIP-seq (H3K27ac, H3K4me3) and ATAC-seq data from microglia (Mic) and neurons (NEU) highlight active enhancers overlapping the disrupted loop anchors and regulatory elements near *ARMS2*, *HTRA1*, and *PLEKHA1*. **(c)** Cell-type–specific transcriptional effects in dementia. Single-cell RNA-seq data from SEA-AD show elevated expression of *ARMS2* and *PLEKHA1* in microglia (Mic) from dementia samples compared to non-dementia controls, suggesting potential regulatory activation associated with the 3D structural disruption.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryTableS1.zip](#)
- [SupplementaryTableS2.zip](#)
- [SupplementaryFigS1.pdf](#)
- [SupplementaryTableS8.zip](#)
- [SupplementaryTableS9.zip](#)
- [SupplementaryTableS3.zip](#)
- [SupplementaryTableS4.zip](#)
- [SupplementaryTableS7.zip](#)
- [SupplementaryFigS3.pdf](#)
- [SupplementaryTableS5.zip](#)
- [SupplementaryFigS2.pdf](#)
- [SupplementaryTableS6.zip](#)
- [SupplementaryFigS5.pdf](#)
- [SupplementaryFigS4.pdf](#)
- [SupplementaryTableLegends.pdf](#)
- [SupplementaryFigureLegends.pdf](#)