Deep Learning-Based Prediction of Variant Effects on Chromatin Accessibility During Dynamic Neuronal Activation

Zicheng Wang & Xin He
Department of Human Genetics
The University of Chicago
Chicago, IL 60637
{zichengwang,xinhe}@uchicago.edu

Abstract

Uncovering the genetic basis of neuropsychiatric diseases (NPDs) is challenging, as most risk variants are noncoding and operate via dynamic, context-dependent mechanisms—many of which remain hidden under static baseline conditions and only emerge during neuronal stimulation. In this work, we evaluate the capacity of ChromBPNet, a sequence-to-function deep learning framework, to predict variant effects across dynamic cellular states. Using iPSC-derived neurons at 0hr (unstimulated), 1hr, and 6hr of KCl stimulation for training, we validated model performance using empirical chromatin accessibility QTLs (caQTLs). ChromBPNet achieves high performance in prioritizing functional variants (average precision 0.51–0.60; 5.6–$6.6\times$ baseline enrichment) and accurately predicts allelic log fold-changes (Pearson $r = 0.61$–0.69; 82–86% directional concordance). Crucially, we observed poor cross-lineage generalization, as models trained on distinct cell types (e.g., microglia) failed to predict neuronal variant effects (Pearson $r \approx 0.2$). These results demonstrate that sequence-based models capture robust regulatory features within a lineage but require context-matched training data to generalize across distinct chromatin landscapes.

1 Introduction

Genome-wide association studies (GWAS) have mapped hundreds of risk loci to neuropsychiatric disorders (NPDs), yet their functional interpretation remains a major challenge (Yao et al., 2021; Smeland et al., 2025). The vast majority of these variants are located in non-coding regions and exert their effects through highly context-specific mechanisms. Crucially, the regulatory effects of many NPD risk variants only emerge during neuronal stimulation (Boulting et al., 2021; Ma et al., 2024; Roussos et al., 2016). By relying solely on 'resting state' baseline data, current models overlook these dynamic regulatory landscapes, leaving a significant portion of disease heritability unexplained.

While experimental approaches such as chromatin accessibility quantitative trait loci (caQTL) mapping can reveal context-specific regulatory effects, they remain resource-intensive and constrained by sample availability and genetic diversity. Recent sequence-to-function deep learning models, including AlphaGenome (Avsec et al., 2026), Borzoi (Linder et al., 2025), and ChromBPNet (Pampari et al., 2025), offer a scalable alternative. They predict functional genomic readouts directly from DNA sequence and enable in silico mutagenesis. These models are largely trained on bulk epigenomic and transcriptomic data collected under baseline conditions across tissues and cell types, allowing broad variant effect prediction without large-scale genetic sampling. However, because their training data primarily capture steady-state regulatory landscapes, an open question remains: do these models capture the regulatory grammar governing stimulus-dependent responses, or mainly reflect static accessibility patterns?

In this work, we evaluated the reliability of sequence-based variant effect scores by comparing them against observed caQTLs in a dynamic cellular environment. We utilized a comprehensive iPSC-derived neuron dataset, including GABAergic and glutamatergic lineages profiled at an unstimulated baseline (0hr) and following 1hr and 6hr of KCl stimulation (Liang et al., 2025). With three lineages and three stimulation timepoints, the dataset yields nine unique cell-state combinations, enabling mapping of dynamic caQTLs characterized by stimulation-dependent changes in genetic effect sizes.

Our results demonstrate that while ChromBPNet achieves high accuracy in prioritizing functional variants and predicting allelic effect sizes within neuronal lineages, there is a significant performance drop when models are applied across distinct cell types. This underscores a critical requirement for context-matched training data to accurately capture the lineage-specific regulatory grammar that sequence-based models otherwise fail to generalize from unrelated chromatin landscapes.

2 Methodology

2.1 Dataset and Preprocessing

We utilized a single-nucleus ATAC-seq (snATAC-seq) dataset from Liang et al. (2025) to investigate variant-level regulatory effects in resting and stimulated human neurons. The dataset contains three distinct cell lineages, GABAergic neurons (GABA), $NEFM^-$ glutamatergic neurons (nmglut), and $NEFM^+$ glutamatergic neurons (npglut), each profiled across three temporal states: unstimulated (0hr) and KCl-stimulated (1hr and 6hr). We utilized caQTLs from this study, which the authors mapped across a cohort of 95 cell lines. For ChromBPNet training, we obtained snATAC-seq fragments, consensus peak sets, and cell type annotations from a subset of 18 cell lines from Batch 024.

2.2 ChromBPNet Training Pipeline

To mitigate coverage-related biases, we downsampled snATAC-seq reads to ensure uniform library sizes across all cell types and stimulation states. We observed that neuronal stimulation induces high concentrations of transcription factors, which can introduce motif artifacts in background regions; therefore, the bias model was trained exclusively on unstimulated (0hr) neurons and subsequently applied to the corresponding 1hr and 6hr states.

The ChromBPNet models were trained on pseudobulk data for each of the nine contexts, with bias factorization. We then computed variant effect scores (predicted log fold-change) for 501,417 variants located within peaks using the bias-corrected models. To optimize computational efficiency, variant scores were calculated using the forward strand only. P-values were estimated using 1 million shuffled background sequences.

2.3 Benchmarking and Evaluation

We focused our evaluation exclusively on variants located within open chromatin regions (OCRs). We implemented a lead-variant selection strategy: for each peak with at least one caQTL (cPeak), we retained only the caQTL with the lowest empirical P-value. If multiple variants had the same lowest P-value, we selected the one closest to the peak summit, ensuring a direct comparison between the strongest empirical signal and the model-predicted score. We assessed model performance using two primary metrics:

1. **Precision-Recall Analysis:** We measured the model's ability to distinguish causal variants from non-regulatory background noise. Each lead caQTL was matched with 10 negative control variants (empirical $P > 0.1$), establishing a baseline precision of 1/11 (0.09). We calculated the average precision (AP) using the absolute predicted variant scores ($|\log \mathrm{FC}|$) as the ranking metric.
2. **Quantitative Agreement and Directionality:** To evaluate the accuracy of the predictions, we computed the Pearson correlation (r) between the model-predicted variant effects (predicted $\log \mathrm{FC}$) and the empirically observed caQTL effect sizes. Furthermore, we assessed directional concordance, defined as the percentage of significant lead caQTLs where the sign of the predicted effect matched the sign of the observed experimental effect.

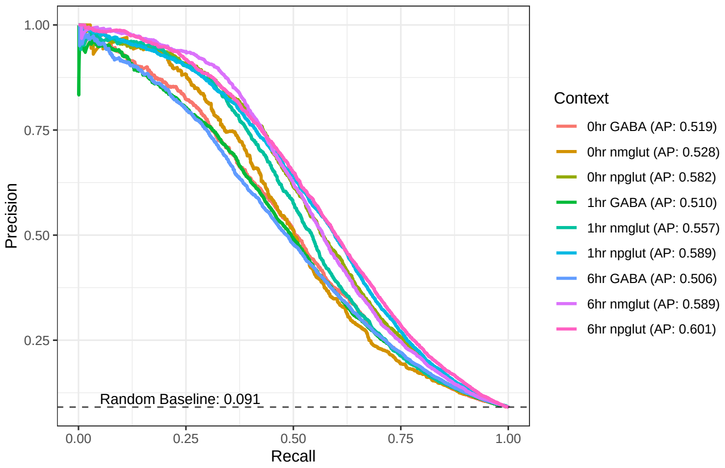


Figure 1: Evaluation of variant prioritization across cell types and stimulation states. Each curve contrasts lead caQTLs against matched negative control variants within accessibility peaks, using $|\log \mathrm{FC}|$ as the ranking metric. Average precision (AP) values for each context are reported.

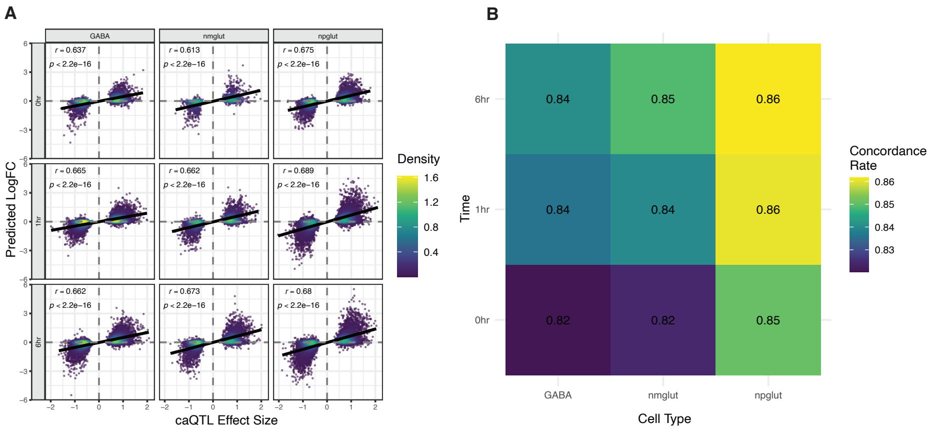


Figure 2: Evaluation of ChromBPNet predicted effect sizes and directional concordance against lead caQTLs. (a) Correlation between ChromBPNet-predicted variant scores and empirical caQTL effect sizes. (b) Directional concordance between predicted variant scores and observed caQTL effects.

3 RESULTS

We first assessed ChromBPNet's ability to distinguish functional from non-regulatory variants using Precision-Recall analysis (Figure 1). Across nine contexts, AP values were consistently high, ranging from 0.51 in 6hr GABA to 0.60 in 6hr npglut, indicating strong performance in identifying variants that influence chromatin state. These values correspond to a 5.6-6.6 fold enrichment over a baseline of random SNPs located within accessibility peaks, demonstrating the model's robust capacity to identify variants that modulate chromatin states.

When restricted to significant lead caQTLs, ChromBPNet's logFC predictions showed moderate to strong correlations with observed caQTL effect sizes ranging from 0.61 to 0.69 (Figure 2a). Crucially, the model displayed high directional concordance, with the sign of the predicted effect matching the caQTL effect in 82–86% of lead caQTLs (Figure 2b). Notably, this performance remained stable across stimulus time points (1hr and 6hr), suggesting that the model successfully captures regulatory logic that persists during cellular activation.

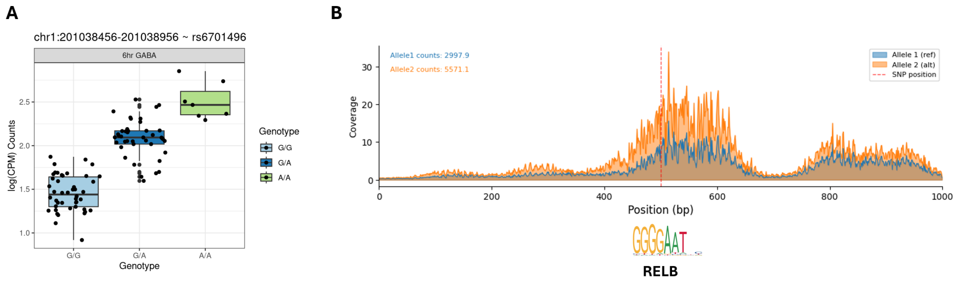


Figure 3: Mechanistic nomination of a SCZ risk variant (rs6701496) in 6hr GABAergic neurons. (a) Box plot of normalized total accessibility of the region nearby the variant, grouped by genotype. (b) Predicted in silico accessibility profiles for the reference and alternative alleles across a 1 Kb window centered on the variant.

To assess the limits of cross-lineage generalization, we evaluated the performance of ChromBPNet models trained on unrelated lineages, specifically microglia (Supplementary Figure 1a) and smooth muscle cells (SMC; Supplementary Figure 1b), against our neuronal caQTLs. When applied to neuron caQTLs, these models showed a significant drop in performance, with Pearson correlations falling to approximately 0.2. This underscores the necessity of context-specific training data for accurate variant effect prediction in complex, lineage-specific landscapes.

Our framework successfully nominated variants with potential roles in neuropsychiatric disease. For instance, rs6701496, a schizophrenia (SCZ) risk variant, was identified as a strong caQTL in 6hr stimulated GABAergic neurons (Figure 3a). ChromBPNet predicted increased chromatin accessibility for the alternative allele, consistent with the altered affinity of a RELB transcription factor motif (Figure 3b). RELB is a non-canonical NF-κB subunit regulating immune-related transcriptional programs (Gupta et al., 2019), and NF-κB dysregulation has been reported in schizophrenia and inflammatory states (Murphy et al., 2022). These results suggest that the rs6701496 variant modifies the affinity of a RELB motif, potentially altering activation-linked neuroimmune regulatory programs that contribute to schizophrenia risk.

4 Conclusion and Future Directions

In this study, we extended the application of ChromBPNet into the dynamic landscape of stimulated neurons, training models to decode the sequence basis of chromatin accessibility across a temporal activation gradient (0hr, 1hr, and 6hr). ChromBPNet reliably prioritized functional variants within accessible chromatin, achieving average precision of 0.51–0.60 (approximately 5.6–6.6 $\times$ enrichment over the locus-matched baseline). For lead caQTLs, predicted log fold-changes were quantitatively aligned with measured effect sizes (Pearson $r = 0.61$–0.69), and the sign of the predicted effect agreed with the empirical slope in 82–86% of cases. Importantly, this accuracy was stable at 1hr and 6hr, indicating that the learned regulatory grammar generalizes across stimulation states.

While ChromBPNet accurately predicts variant effects in neuronal contexts, our results reveal a key limitation in cross-lineage generalization. Models trained on unrelated cell types, such as microglia or smooth muscle, showed a marked performance drop on neuronal caQTLs, with Pearson correlations falling to $r \approx 0.2$. These findings indicate that sequence-to-function models depend on context-matched training data to capture lineage-specific chromatin landscapes.

To bridge the gap between deep-learning predictions and population-level QTL measurements, three advancements are crucial. First, integrating formal fine-mapping (Wang et al., 2020) will better isolate causal variants from those in linkage disequilibrium, sharpening performance benchmarks. Second, adopting personalized genomic frameworks like SAGE-net (Spiro et al., 2025) will move beyond reference sequences to capture how individual variation influences caQTL effect sizes. Finally, pairing snATAC-seq with snRNA-seq will be critical to determine how sequence-driven chromatin changes translate into functional, stimulus-induced transcription. Together, these steps will enable sequence-to-function models to move beyond variant prioritization toward mechanistic prediction of disease-associated regulatory processes.

REFERENCES

Žiga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R Taylor, Tom Ward, Clare Bycroft, Lauren Nicolaisen, Eirini Arvaniti, Joshua Pan, et al. Advancing regulatory variant effect prediction with alphagenome. *Nature*, 649(8099):1206–1218, 2026.

Gabriella L Boulting, Ershela Durresi, Bulent Ataman, Maxwell A Sherman, Kevin Mei, David A Harmin, Ava C Carter, Daniel R Hochbaum, Adam J Granger, Jesse M Engreitz, et al. Activity-dependent regulome of human gabaergic neurons reveals new patterns of gene regulation and neurological disease heritability. *Nature Neuroscience*, 24(3):437–448, 2021.

Angela S Gupta, Debolina D Biswas, La Shardai N Brown, Karli Mockenhaupt, Michael Marone, Andrew Hoskins, Ulrich Siebenlist, and Tomasz Kordula. A detrimental role of relb in mature oligodendrocytes during experimental acute encephalomyelitis. *Journal of neuroinflammation*, 16(1):161, 2019.

Lifan Liang, Siwei Zhang, Zicheng Wang, Hanwen Zhang, Chuxuan Li, Alexandra C Duhe, Xiaotong Sun, Xiaoyuan Zhong, Alena Kozlova, Brendan Jamison, et al. Single-cell multiomics of neuronal activation reveals context-dependent genetic control of brain disorders. *bioRxiv*, 2025.

Johannes Linder, Divyanshi Srivastava, Han Yuan, Vikram Agarwal, and David R Kelley. Predicting rna-seq coverage from dna sequence as a unifying model of gene regulation. *Nature Genetics*, 57 (4):949–961, 2025.

Yixuan Ma, Jaroslav Bendl, Brigham J Hartley, John F Fullard, Rawan Abdelaal, Seok-Man Ho, Roman Kosoy, Peter Gochman, Judith Rapoport, Gabriel E Hoffman, et al. Activity-dependent transcriptional program in ngn2+ neurons enriched for genetic risk for brain-related disorders. *Biological psychiatry*, 95(2):187–198, 2024.

Caitlin E Murphy, Adam K Walker, Maryanne O'Donnell, Cherrie Galletly, Andrew R Lloyd, Dennis Liu, Cynthia Shannon Weickert, and Thomas W Weickert. Peripheral nf-κb dysregulation in people with schizophrenia drives inflammation: putative anti-inflammatory functions of nf-κb kinases. *Translational Psychiatry*, 12(1):21, 2022.

Anusri Pampari, Anna Shcherbina, Evgeny Z Kvon, Michael Kosicki, Surag Nair, Soumya Kundu, Arwa S Kathiria, Viviana I Risca, Kristiina Kuningas, Kaur Alasoo, et al. Chrombpnet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants. *bioRxiv*, 2025.

Panos Roussos, Boris Guennewig, Dominik C Kaczorowski, Guy Barry, and Kristen J Brennand. Activity-dependent changes in gene expression in schizophrenia human-induced pluripotent stem cell neurons. *JAMA psychiatry*, 73(11):1180–1188, 2016.

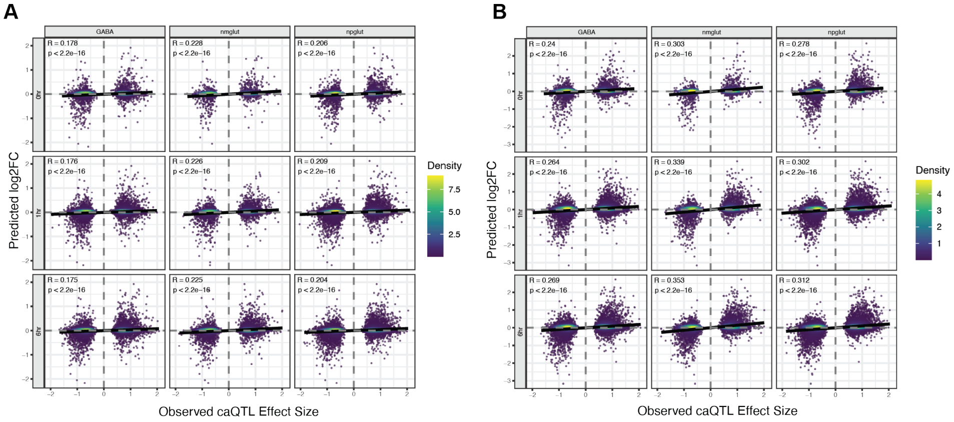
Olav B Smeland, Gleda Kutrolli, Shahram Bahrami, Vera Fominykh, Nadine Parker, Julian Fuhrer, Guy FL Hindley, Linn Rødevand, Piotr Jaholkowski, Markos Tesfaye, et al. A genome-wide analysis of the shared genetic risk architecture of complex neurological and psychiatric disorders. *Nature Neuroscience*, pp. 1–12, 2025.

Anna E Spiro, Xinming Tu, Yilun Sheng, Alexander Sasse, Rezwan Hosseini, Maria Chikina, and Sara Mostafavi. A scalable approach to investigating sequence-to-expression prediction from personal genomes. *bioRxiv*, pp. 2025–02, 2025.

Gao Wang, Abhishek Sarkar, Peter Carbonetto, and Matthew Stephens. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(5):1273–1300, 2020.

Xueming Yao, Joseph T Glessner, Junyi Li, Xiaohui Qi, Xiaoyuan Hou, Chonggui Zhu, Xiaoge Li, Michael E March, Liu Yang, Frank D Mentch, et al. Integrative analysis of genome-wide association studies identifies novel loci associated with neuropsychiatric disorders. *Translational psychiatry*, 11(1):69, 2021.

A Appendix



Supplementary Figure 1: Limited cross-lineage generalization of variant effect predictions. Scatter plots showing correlations between neuronal caQTL effects and ChromBPNet models trained on **(a)** Microglia and **(b)** Smooth Muscle Cells (SMC). The significant drop in performance ($r \approx 0.2$) highlights the requirement for context-matched training data.