



Deep learning predicts DNA methylation regulatory variants in the human brain and elucidates the genetics of psychiatric disorders

Jiyun Zhou^{a,b}, Qiang Chen^a, Patricia R. Braun^b, Kira A. Perzel Mandell^{a,c}, Andrew E. Jaffe^{a,b,c,d,e,f}, Hao Yang Tan^{a,b}, Thomas M. Hyde^{a,b,g}, Joel E. Kleinman^{a,b}, James B. Potash^b, Gen Shinozaki^h, Daniel R. Weinberger^{a,b,c,d,g}, and Shizhong Han^{a,b,c,1}

Edited by Joseph Ecker, Salk Institute for Biological Studies, La Jolla, CA; received April 11, 2022; accepted July 18, 2022

There is growing evidence for the role of DNA methylation (DNAm) quantitative trait loci (mQTLs) in the genetics of complex traits, including psychiatric disorders. However, due to extensive linkage disequilibrium (LD) of the genome, it is challenging to identify causal genetic variations that drive DNAm levels by population-based genetic association studies. This limits the utility of mQTLs for fine-mapping risk loci underlying psychiatric disorders identified by genome-wide association studies (GWAS). Here we present INTERACT, a deep learning model that integrates convolutional neural networks with transformer, to predict effects of genetic variations on DNAm levels at CpG sites in the human brain. We show that INTERACT-derived DNAm regulatory variants are not confounded by LD, are concentrated in regulatory genomic regions in the human brain, and are convergent with mQTL evidence from genetic association analysis. We further demonstrate that predicted DNAm regulatory variants are enriched for heritability of brain-related traits and improve polygenic risk prediction for schizophrenia across diverse ancestry samples. Finally, we applied predicted DNAm regulatory variants for fine-mapping schizophrenia GWAS risk loci to identify potential novel risk genes. Our study shows the power of a deep learning approach to identify functional regulatory variants that may elucidate the genetic basis of complex traits.

DNA methylation quantitative trait loci (mQTL) | convolutional neural network (CNN) | transformer | regulatory variants | GWAS

Genome-wide association studies (GWAS) have achieved remarkable success in identifying genetic associations with complex traits, including many hundreds of risk loci uncovered for psychiatric disorders and brain-related traits (1–3). However, it has been challenging to identify causal variants within GWAS risk loci and their target genes, largely due to extensive linkage disequilibrium (LD) across the genome (4) and our incomplete knowledge of the noncoding genome where most GWAS hits reside (5). Nonetheless, accumulating evidence indicates that noncoding risk variants are enriched in regulatory DNA regions in tissues and cell types related to disease, suggesting that noncoding risk variants may act through regulation of gene expression in disease-relevant tissues and cell types (6).

DNA methylation (DNAm) plays an important role in transcriptional regulation, brain development, and function (7, 8). Genetic control of DNAm levels may be one functional mechanism of noncoding risk variants underlying psychiatric disorders (9). Indeed, population-based genetic association studies have identified numerous DNAm quantitative trait loci (mQTLs) associated with DNAm levels at CpG sites in different tissues, including in the human brain (10–15). There has been considerable interest in utilizing mQTLs to interpret the functional consequences of noncoding risk variants within GWAS risk loci, including via colocalization (16), Mendelian randomization (17), and the imputation-driven methylome-wide association study (18). However, as with other types of molecular QTLs, due to extensive LD across the genome, it is challenging to identify causal genetic variations that drive DNAm levels of CpG sites by genetic association studies. This limits the utility of mQTLs for pinpointing causal variants within GWAS risk loci and identifying target genes.

As a complementary approach to QTL discovery, deep learning techniques have been employed to predict effects of genetic variations on molecular traits, such as gene expression and chromatin marks, in bulk tissues and cell lines (19–21). The general idea is to first build prediction models for molecular traits from local DNA sequences, and then estimate the impacts of genetic variations on molecular traits by the difference of predicted levels of molecular traits between the two DNA sequences of different alleles. In contrast to traditional molecular QTLs that are confounded by LD, regulatory variants

Significance

Genetic association studies have identified widespread DNA methylation (DNAm) quantitative trait loci (mQTLs) that may illuminate causal genetic variations underlying human complex traits. However, due to extensive linkage disequilibrium (LD) of the genome, it is challenging to identify causal genetic variations that drive DNAm levels. Here we present a deep learning model to predict effects of genetic variations on DNAm levels in the human brain. We demonstrate that deep learning-derived DNAm regulatory variants are not confounded by LD, are convergent with mQTL evidence from genetic association analysis, and underlie the genetic basis of brain-related traits. Our study shows the power of a deep learning approach to identify functional regulatory variants that may elucidate the genetic basis of complex traits.

Author contributions: T.M.H., J.E.K., J.B.P., G.S., D.R.W., and S.H. designed research; J.Z., Q.C., A.E.J., H.Y.T., and S.H. performed research; J.Z., Q.C., P.R.B., K.A.P.M., A.E.J., H.Y.T., and S.H. analyzed data; and J.Z., P.R.B., J.B.P., D.R.W., and S.H. wrote the paper.

Competing interest statement: A.E.J. is a current employee and shareholder of Neumora Therapeutics.

This article is a PNAS Direct Submission.

Copyright © 2022 the Author(s). Published by PNAS. This article is distributed under [Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 \(CC BY-NC-ND\)](https://creativecommons.org/licenses/by-nc-nd/4.0/).

¹To whom correspondence may be addressed. Email: Shizhong.Han@libd.org.

This article contains supporting information online at <http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2206069119/-DCSupplemental>.

Published August 15, 2022.

predicted by deep learning techniques in theory do not suffer from the confounding effect of LD, because deep learning predicts regulatory variants by estimating their impacts on DNA motifs associated with molecular traits.

However, most previous deep learning-driven studies were based on a convolutional neural network (CNN) that is usually unable to capture long-range dependencies. This may result in suboptimal prediction for molecular traits, because long-range interactions of DNA elements are not unusual in gene regulation, such as co-occurring DNA motifs that act cooperatively (22). The recurrent neural network approach was developed to address long-range dependencies, but it suffers from memory loss for distant features and computational inefficiency (23). Recently, the transformer model, characterized by the self-attention mechanism, is becoming a popular approach to capture long-range dependencies (24). The transformer has become the model of choice for natural language processing, and has been successfully applied to learning representations for DNA and protein sequences (25, 26). It is also notable that the self-attention mechanism of transformer is a key feature of AlphaFold, a deep learning architecture that has led to a breakthrough in predicting protein structure from sequence (27).

The recent advances in deep learning techniques, along with the availability of high-throughput DNAm quantification in the human brain, have opened up new opportunities to predict functional variants regulating DNAm levels in the human brain, and potentially, to further our understanding of the genetic basis of psychiatric disorders. However, we are not aware of any study that has taken advantage of these opportunities. Two studies applied deep learning to predict DNAm levels or states from local DNA sequences in cell lines and single cells (28, 29), but both studies were based on CNN models without the context of the human brain. Moreover, only one study applied the model to predict genetic impacts on DNAm states (29), but no further work was conducted to relate predicted regulatory variants with genetics of complex traits.

Here, we present a deep learning model, the integrated CNN and transformer (INTERACT), to predict effects of genetic variations on DNAm levels in the human brain. We show the superior performance of INTERACT in predicting DNAm levels from local DNA sequences compared to several other CNN-based models. We demonstrate that DNAm regulatory variants predicted by the INTERACT model are not confounded by LD, are convergent with independent mQTL evidence, and underlie the genetic basis of brain-related traits. Finally, we apply predicted DNAm regulatory variants for fine-mapping schizophrenia GWAS risk loci, and show that the approach identifies potential novel risk genes.

Results

INTERACT Model for Predicting Tissue-Specific DNAm Levels.

We designed INTERACT, which integrates CNN and transformer to predict DNAm levels of CpG sites from local DNA sequences (Fig. 1A). The CNN module enables the capture of local DNA sequence features, whereas the transformer module aims to detect distant features that may act jointly. To evaluate whether INTERACT is able to learn tissue-specific information that determines tissue-specific DNAm levels, we trained four tissue-specific INTERACT models to predict DNAm levels of ~800,000 CpG sites measured by the EPIC Methylation array in samples of four different tissues (brain, blood, saliva, and buccal) collected from the same subjects in our previous study (30). Given the complexity of the INTERACT model and the

limited number of CpG sites on the EPIC array as training samples, we pretrained INTERACT by predicting DNAm levels of ~26 million CpG sites in the hippocampus measured by whole-genome bisulfite sequencing (WGBS) in our previous study (31). Pretrained INTERACT was then used to fine-tune each tissue-specific INTERACT model by predicting DNAm levels of CpG sites on the EPIC array for samples of each tissue.

We evaluated the performance of each fine-tuned tissue-specific INTERACT model (hereafter referred to as tissue-specific INTERACT unless otherwise specified) in predicting DNAm levels of independent CpG sites not used for model training. Overall, we observed strong correlations between the predicted and observed DNAm levels of independent CpG sites across four tissue-specific models, with the average Spearman correlation ranging from 0.78 in blood to 0.82 in brain samples (Fig. 1B). As an example, Fig. 1C shows the scatter plot for the observed and predicted DNAm levels by the brain-specific INTERACT in one brain sample (Spearman correlation $r = 0.82$, $P < 2.2 \times 10^{-16}$). We compared the performance of each tissue-specific INTERACT with several alternative models in predicting DNAm levels, including INTERACT without pretraining, standard CNN, and MRCNN, another CNN-based model published previously (28) (Fig. 1B and Dataset S1). The fine-tuned INTERACT achieved the best performance, as measured by both Spearman correlation and mean squared error across four tissue-specific models, followed by the INTERACT without pretraining, and the CNN model. The MRCNN showed the worst performance across all four tissue-specific models.

We further examined model prediction performance for CpG sites of different content in brain samples. Specifically, we grouped CpG sites by specific regulatory regions located in the human brain (i.e., promoter vs. enhancer) or by their DNAm level (low: 0 to 0.1; intermediate: 0.1 to 0.9; high: 0.9 to 1). We observed that: 1) INTERACT has better prediction performance for CpG sites within enhancers than within promoters, and 2) INTERACT has better prediction performance for CpG sites with intermediate DNAm levels than with either high or low DNAm levels (SI Appendix, Table S1). However, the correlations between measured and predicted DNAm levels for each subgroup of CpG sites were lower than the global correlation (average Spearman correlation = 0.82) across all independent CpG sites, which could be explained by the global correlation being dominated by CpG sites of very low and very high DNAm levels (Fig. 1C). Nonetheless, INTERACT consistently outperformed CNN and MRCNN in predicting DNAm levels for each subgroup of CpG sites (SI Appendix, Table S1).

We evaluated whether INTERACT captures tissue-specific information in multiple ways. First, we observed that the top two principal components (PCs) of predicted DNAm levels of independent CpG sites can clearly separate samples of four different tissues (Fig. 1D), suggesting that our tissue-specific models learned tissue-specific information in predicting DNAm levels. Second, we computed the correlation of differential DNAm effect between each tissue pair at independent CpG sites based on observed and predicted DNAm levels. INTERACT showed a higher correlation on average (0.47) across tissue pairs than INTERACT without pretraining (0.36), CNN (0.21), and MRCNN (0.13) (SI Appendix, Table S2). Third, we investigated how well INTERACT can recover tissue-specific differentially methylated regions (tDMRs) based on observed DNAm levels for each tissue pair. INTERACT can recover an average of ~30% tDMRs across tissue pairs, which were higher than INTERACT without pretraining (24%), CNN (15%), and MRCNN (5%) (SI Appendix, Table S3). These data further

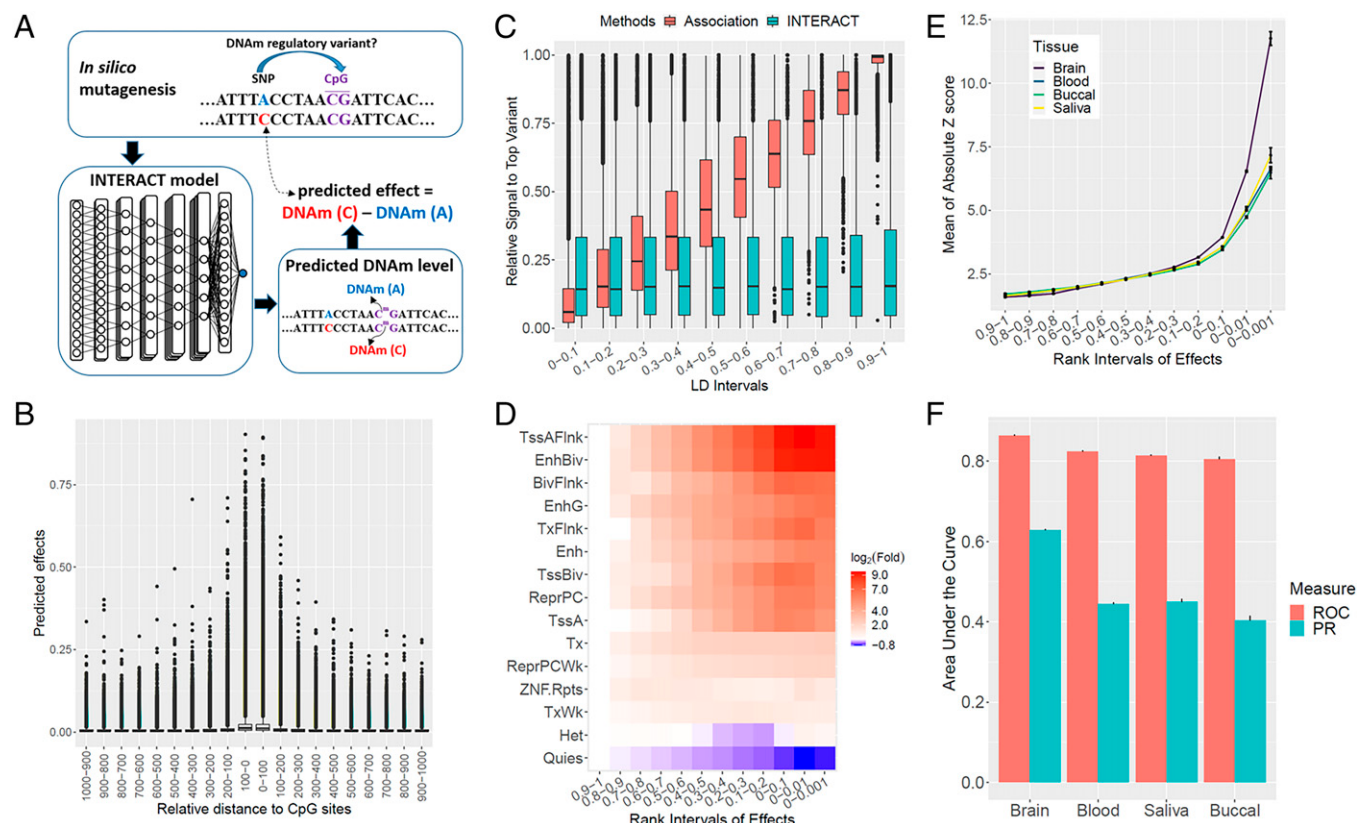


Fig. 2. In silico discovery, characterization, and validation of DNAm regulatory variants. (A) A schematic view of in silico discovery of DNAm regulatory variants. (B) Comparison of predicted effects of variants on DNAm levels vs. their relative distance to CpG sites. Variants are grouped by their relative distance to CpG sites on the x axis. A box plot represents the distribution of predicted effects of variants on DNAm levels by the brain-specific model in one brain sample. (C) Comparison of relative signal of each variant to the top variant of the strongest signal (y axis) derived from either association analysis or INTERACT vs. the LD strength between the variant and the top variant (x axis) of each CpG site. Relative signal from association analysis were measured by the $[-\log_{10}(P \text{ value})]$ of each variant divided by the maximum value of top variant for each CpG site. Relative signal from INTERACT were measured by the absolute value of predicted effect of each variant divided by the maximum value of top variant for each CpG site. (D) Enrichment of 15-core chromatin states in the DLPCF from the Epigenome Roadmap project among variants ranked at different intervals by their predicted effects on DNAm levels in one brain sample. Rank interval “0–0.001” represents variants of large effect and ranked in the top 0.1%. The color gradient represents $\log_2$ (enrichment fold) of variants in each rank interval for their enrichment of each chromatin state compared to variants ranked in the bottom (0.9 to 1). (E) Comparing SNP effects on DNAm levels of SNP–CpG pairs (x axis) predicted by each tissue-specific INTERACT vs. average mQTLs signals of the same SNP–CpG pairs (y axis) from association analysis. Each tissue-specific model predicts SNP effects on DNAm levels of SNP–CpG pairs in each sample of the corresponding tissue used for training the model. SNP–CpG pairs are then ranked by their predicted effects in each training sample, and average mQTLs signals are calculated for SNP–CpG pairs in each rank interval. Rank interval “0–0.001” represents SNP–CpG pairs of large predicted effect and ranked in the top 0.1%. Each point and its error bar on the curve represent mean and SD of average mQTLs signals of SNP–CpG pairs in each rank interval across training samples of the same tissue. (F) Comparison of tissue-specific INTERACT models for their performance in predicting causal mQTLs in the gold-standard dataset. Each tissue-specific INTERACT predicts SNP effects on DNAm levels of SNP–CpG pairs of gold-standard dataset in each sample of the corresponding tissue used for training the model. AUC-ROC and AUC-PR are calculated for each tissue-specific INTERACT based on predicted SNP effects on DNAm levels of SNP–CpG pairs by the model in each training sample. The height of each bar and its error bar represents the mean and SD of AUC-ROC (AUC-PR) across training samples of the same tissue.

levels by our tissue-specific INTERACT models. Specifically, we computed mQTL association statistics of all single-nucleotide polymorphism (SNP)–CpG pairs for SNPs within a 2-kb window of CpG sites on a 450K array, based on a dataset from our previous study of mQTLs in the prefrontal cortex (15). We then compared mQTL signals for SNP–CpG pairs grouped by their effects on DNAm levels predicted by each tissue-specific INTERACT. We reasoned that if predicted effects were reliable, SNP–CpG pairs with greater effects would show stronger association evidence. This was indeed the case, as SNP–CpG pairs of greater predicted effects had a higher absolute value of the z-statistic on average, and the pattern was stronger for effects predicted by the brain-specific model than by nonbrain-specific models (Fig. 2E), reinforcing that context matters in predicting regulatory variants acting in the brain. Moreover, we observed that SNP–CpG pairs of greater predicted effects had a higher rate of consistency in direction of effect with mQTLs, with the strongest pattern observed for effects predicted by the brain-specific model (SI Appendix, Fig. S1), providing reassurance about model quality.

To further quantify the performance of tissue-specific INTERACT models in predicting causal mQTLs, we built a “gold-standard” dataset comprised of SNP–CpG pairs with high-confidence causal mQTLs and null SNPs (no effects on DNAm). Specifically, gold standards were built from statistical fine-mapping using the SuSiE (41) based on the same mQTL dataset from the prefrontal cortex described above (15). We collected 541 SNP–CpG pairs with high-confidence causal mQTLs (posterior inclusion probability [pip] = 1, Bonferroni-corrected association $P < 0.05$) and 5,024 SNP–CpG pairs of no mQTLs (pip = 0, association $P > 0.99$). We then evaluated the ability of our models to separate the two groups of SNP–CpG pairs by predicting SNP effects on DNAm levels for each SNP–CpG pair. The brain-specific INTERACT achieved the best discrimination performance measured by the area under the receiver operating characteristic curve (AUC-ROC = 0.86), followed by blood- (AUC-ROC = 0.82), saliva- (AUC-ROC = 0.81), and buccal-specific INTERACT (AUC-ROC = 0.80). The brain-specific INTERACT achieved more notable

gain measured by the area under the precision recall curve (AUC-PR = 0.63) compared to the other three models (blood, 0.44; saliva, 0.45; buccal, 0.40) (Fig. 2*F*), reinforcing that context matters in predicting regulatory variants in the human brain.

DNAm Regulatory Variants Predicted by the Brain-Specific INTERACT Were Enriched for Heritability of Brain-Related Traits. We performed stratified LD-score regression (S-LDSC) to examine whether DNAm regulatory variants predicted by the brain-specific INTERACT were enriched for heritability of 13 brain-related traits, and 1 nonbrain trait, human height. As a comparison, the same analyses were conducted for mQTLs (false-discovery rate [FDR] < 0.01) and fine-mapped mQTLs (fmQTLs, $\text{pip} > 0.1$) computed from the same mQTL dataset from the prefrontal cortex described above (15). Overall, we observed that variants of larger predicted effects were more strongly enriched for heritability of a number of brain-related traits, but not for human height (Fig. 3*A* and Dataset S3). The strongest enrichment was observed for schizophrenia by variants ranked in the top 10%, which represented 14% of annotation SNPs, but explained 35% of schizophrenia heritability (enrichment fold = 2.5, $P = 1.5 \times 10^{-4}$, FDR = 0.008). This level of enrichment was stronger than fmQTLs that had a similar proportion of annotation SNPs (13%), but explained only 20% of schizophrenia heritability (enrichment fold = 1.6, $P = 1.8 \times 10^{-4}$, FDR = 0.01). On the other hand, mQTLs were enriched for heritability of only three traits (bipolar disorder, Parkinson disease, and human height), but fmQTLs showed stronger enrichment for the same three traits, plus enrichment for a number of new traits (schizophrenia, depression, education years, cigarettes per day, and drinks per week) (Dataset S4), reinforcing the role of mQTLs in brain-related traits, but also a need to identify causal mQTLs. Additionally, it was

notable that both mQTLs and fmQTLs showed strong enrichment for heritability of human height, but top-ranked variants predicted by the brain-specific INTERACT did not, suggesting that INTERACT tends to predict DNAm regulatory variants that are unique to brain-related traits.

To evaluate the tissue context for variants implicated in brain-related traits, we performed S-LDSC for variants predicted by the blood-specific INTERACT (SI Appendix, Fig. S2 and Dataset S3). Consistent with the brain-specific INTERACT, the top 10% ranked variants from the blood-specific INTERACT showed the highest enrichment for heritability of a number of brain-related traits, but their strength was generally less than what we observed from the brain-specific INTERACT. For example, the enrichment was only nominally significant for schizophrenia (fold = 2.1, $P = 0.032$), neuroticism (fold = 2, $P = 0.030$), and education years (fold = 1.8, $P = 0.034$), whereas the top 10% ranked variants from the brain-specific INTERACT showed stronger evidence for the same traits (schizophrenia, fold = 2.5, $P = 1.5 \times 10^{-4}$; neuroticism, fold = 2.5, $P = 3.6 \times 10^{-4}$; education years, fold = 2.3, $P = 2.3 \times 10^{-4}$). However, we did observe that the top of 10% variants from the blood-specific INTERACT showed stronger enrichment for heritability of Alzheimer's disease (fold = 4.5, $P = 0.033$) compared to the brain-specific INTERACT (fold = 0.8, $P = 0.93$), possibly reflecting the more immune-related association of this disease than other brain-related traits. Overall, this analysis highlights that tissue context matters in identifying regulatory variants implicated in brain-related traits.

DNAm Regulatory Variants Predicted by the Brain-Specific INTERACT Improve Polygenic Risk Prediction for Schizophrenia. Given the strong evidence that top-ranked variants predicted by the brain-specific INTERACT were enriched for

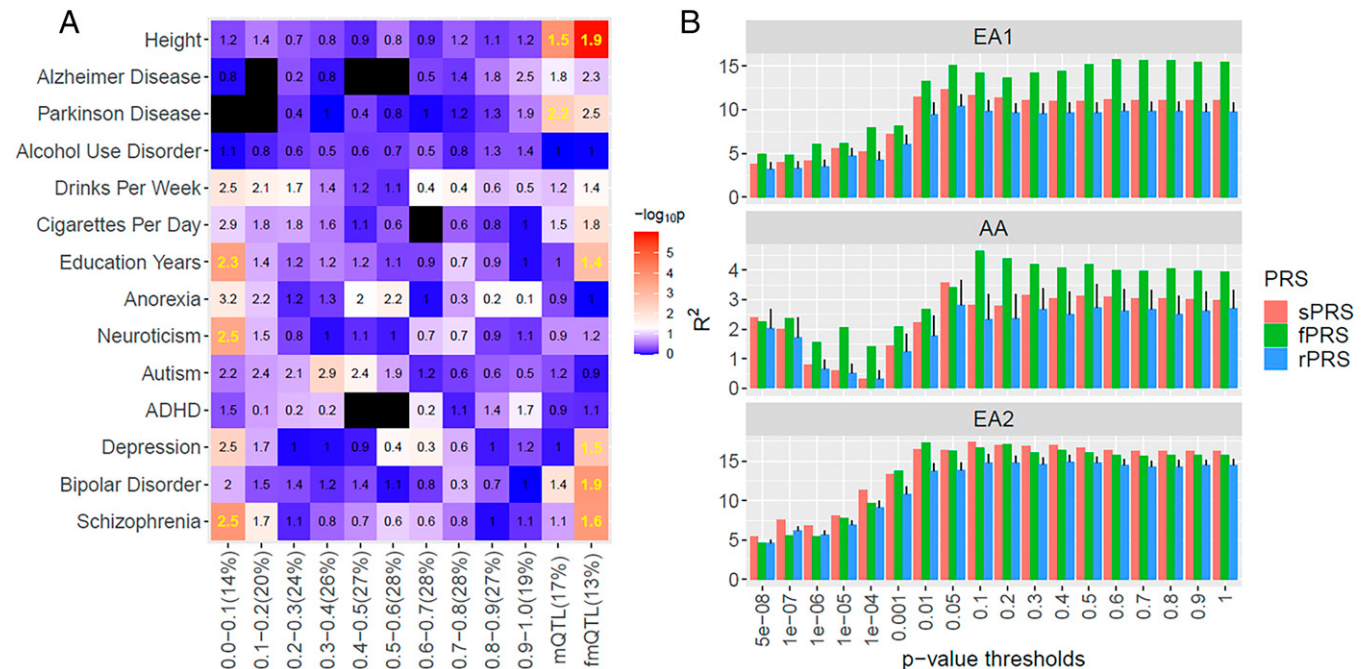


Fig. 3. DNAm regulatory variants predicted by the brain-specific INTERACT underlie the genetic basis of brain-related traits. (A) Heritability enrichment analysis for variants predicted by the brain-specific INTERACT. The x axis represents variants ranked at different intervals by their predicted effects. Interval "0-0.1" represents variants of high effect and ranked in the top 10%. As a comparison, mQTL and fmQTLs in the DLPFC are also included. The percent in brackets represents proportion of annotation SNPs included for S-LDSC. The color gradient represents significance levels for enriched heritability. The black color represents negative heritability estimates from S-LDSC. The numbers within each square are heritability enrichment fold and numbers in bold indicate FDR significant (FDR < 0.05) after multiple testing correction. (B) Comparison of prediction performance of three types of PRS for schizophrenia case-control status. fPRS: functional PRS computed from predicted DNAm regulatory variants; sPRS: standard PRS; rPRS: random PRS computed from random SNPs matched for the number of SNPs with fPRS. Error bar above rPRS represents SD of R^2 across 100 random iterations of rPRS.

schizophrenia heritability, we further examined whether the top ranked variants could improve polygenic risk prediction for schizophrenia case-control status in three independent samples, including two samples of European ancestry (European-American [EA] 1, 178 cases, 343 controls; EA2, 660 cases, 685 controls) and one sample of African ancestry [(African American [AA], 130 cases, 276 controls). We computed polygenic risk scores (PRS) by the standard approach of LD pruning and P -value thresholding (P+T), but used only the top 20% of ranked variants that were enriched for schizophrenia heritability as shown by S-LDSC analysis. We called the PRS computed this way as “functional PRS” (fPRS) since it was solely based on potential DNAm regulatory variants. We compared the predictive performance of fPRS with standard PRS (sPRS) computed by the standard P -value thresholding approach. Because the number of SNPs pruned for fPRS was much less than for sPRS, to make the fPRS and sPRS more comparable, we also computed a random PRS (rPRS) that was matched for the same number of SNPs used for fPRS at each P -value threshold.

We first investigated the predictive performance of fPRS using the top 20% of ranked variants predicted across all brain

samples used for training the brain-specific INTERACT. Despite the smaller number of pruned SNPs, fPRS had a higher prediction value (R^2) for schizophrenia case-control status than did sPRS in EA1 and AA (Fig. 3B and Dataset S5). The relative increase in the maximum R^2 across all P -value thresholds was 28% (fPRS: 15.7% vs. sPRS: 12.3%) and 31% (fPRS: 4.7% vs. sPRS: 3.1%) in EA1 and AA, respectively. A further increase of R^2 was observed when fPRS was compared to rPRS at each P -value threshold, ranging from 30 to 87% in EA1 and from 10% to 4.5-fold in AA. In EA2, the maximum R^2 of fPRS (17.3%) was smaller than that, but close to the maximum R^2 of sPRS (17.4%). When compared to rPRS, fPRS still achieved better predictive performance in EA2 at P -value thresholds $\geq 1 \times 10^{-5}$, with the increase of R^2 ranging from 5 to 27%.

Fine-Mapping Schizophrenia GWAS Risk Loci through DNAm Regulatory Variants Predicted by the Brain-Specific INTERACT. We utilized DNAm regulatory variants predicted by the brain-specific INTERACT to identify putative causal variants and causal genes within schizophrenia GWAS risk loci. Fig. 4A shows the workflow of our fine-mapping strategy. We first

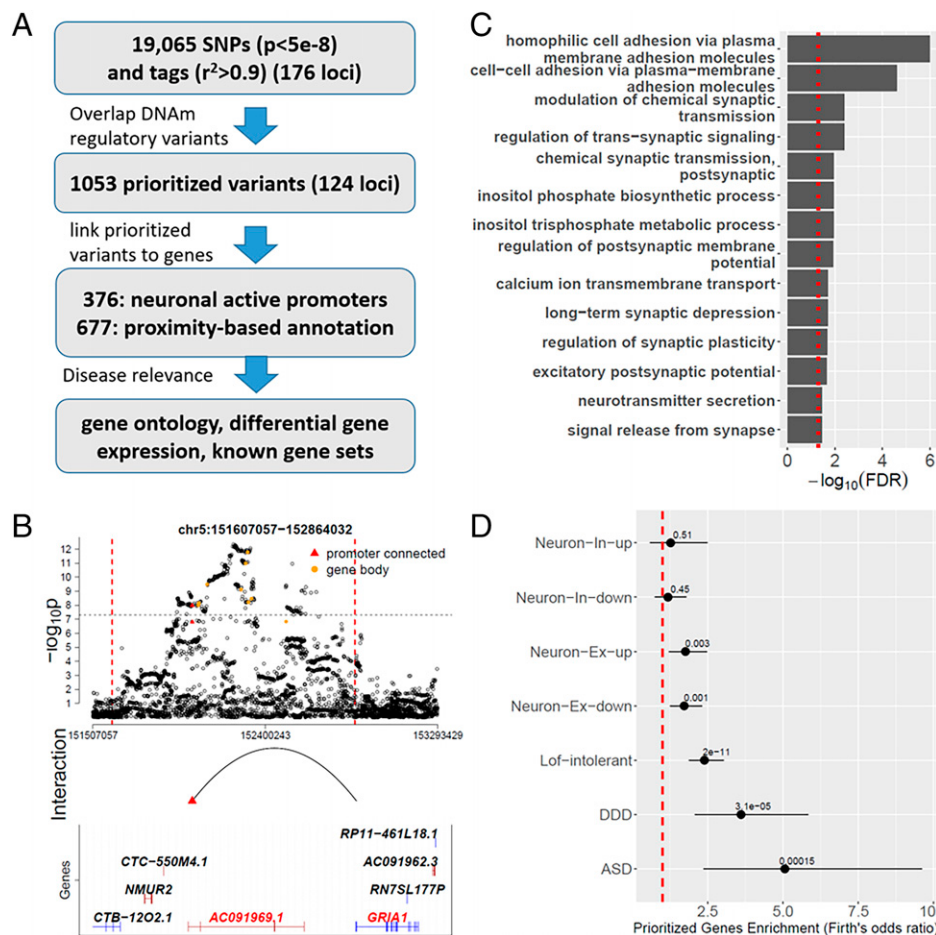


Fig. 4. Fine-mapping schizophrenia GWAS risk loci. (A) Overview of fine-mapping strategy. (B) Fine-mapping result for one risk locus. (Top) Regional plot for GWAS association signals. The two vertical dotted red lines indicate risk locus interval. The colored points indicate prioritized risk variants and their annotations (red triangle, variants connected to active promoters in neuron; orange circle, variants within gene bodies). (Middle) Distal regulation of prioritized risk variants with target gene *GRIA1*. (Bottom) All genes within risk locus. Gene names in red indicate prioritized risk genes. Gene frames are drawn based on the longest transcript from ENSEMBLE annotation (hg19). (C) Gene ontology enrichment analysis for prioritized risk genes. The dotted vertical red line indicates significant threshold after FDR correction ($FDR < 0.05$). (D) Gene set enrichment analysis for prioritized risk genes. The y axis represents gene sets: ASD, autism risk genes identified by integrated de novo mutations and rare variants analysis; DDD, genes enriched for de novo mutations in developmental disorder cases; Lof-intolerant, loss-of-function intolerant genes; Neuron-Ex-down, decreased gene expression in excitatory neurons of schizophrenia cases; Neuron-Ex-up, increased gene expression in excitatory neurons of schizophrenia cases; Neuron-In-down, decreased gene expression in inhibitory neurons of schizophrenia cases; Neuron-In-up, increased gene expression in inhibitory neurons of schizophrenia cases. Each horizontal line represents the odds-ratio (OR) and 95% confidence interval of the association between prioritized risk genes and genes of each gene set. Association was computed by logistic regression using Firth's bias reduction method adjusting for gene size, with all protein coding genes as background. Numbers above each line are P values. Dotted red line indicates no enrichment.

collected 14,444 genome-wide significant SNPs ($P < 5 \times 10^{-8}$) and 4,621 tag SNPs that had strong LD ($r^2 > 0.9$) with those genome-wide significant SNPs, resulting in 19,065 candidate SNPs across 176 autosomal risk loci. We then overlapped these candidate SNPs with top ranked variants that are potentially DNAm regulatory variants (FDR < 0.05), determined empirically by our assembled gold-standard dataset in each training brain sample (Dataset S6). After overlapping, 1,053 variants were retained as putative causal SNPs across 124 risk loci (Dataset S7). To link candidate causal SNPs to target genes they might regulate, we first assigned 376 SNPs to active promoters they overlapped or distally contacted in the neuronal cell type of the human brain, motivated by abundant prior evidence that the neuron is the most relevant cell type where schizophrenia risk variants may exert their effects (42). The remaining 677 SNPs were sequentially assigned to gene bodies (554 SNPs), nearest genes (97 SNPs, within 100 kb), or intergenic regions (26 SNPs, nearest gene > 100 kb) by proximity-based annotations. After linking candidate causal SNPs to their putative gene targets, 44 loci had only one candidate gene, 35 loci had two candidate genes, 9 loci had three candidate genes, and 35 loci contained more than three candidate genes.

It was notable that, among the 124 risk loci mapped by prioritized variants, 58 loci contained risk variants that were distally linked to active promoters in neurons, including 28 where risk variants were linked to active promoters even beyond risk loci (SI Appendix, Fig. S3). The link of risk variants with distal promoters altered our interpretation of GWAS risk loci in multiple ways. First, the distal regulation revealed risk genes that could otherwise be missed if we only examined genes within risk loci. One example was a risk locus on chromosome 5 (chr5:151,607,057 – 152,864,032) in which no obvious candidate genes could be identified within the locus, but two risk variants were linked to *GRN1*, a well-known candidate gene for schizophrenia (43) (Fig. 4B). Second, the distal regulation indicated alternative causal genes within risk loci. One example is the *DRD2* risk locus where prioritized risk genes not only included *DRD2*, but also *TTC12*, which was contacted by a prioritized variant. Third, the distal regulation revealed potential functional mechanisms of noncoding risk variants for risk loci where prioritized risk variants were linked to promoters of high confidence risk genes (e.g., *CACNA1C*, *GRIN2A*, and *CNTN4*) (44–46).

We evaluated the relevance of prioritized genes with schizophrenia in multiple ways. First, gene ontology analysis of prioritized genes revealed pathways strongly implicated in schizophrenia, particularly synaptic transmission (Fig. 4C). Second, we examined whether prioritized genes tended to be differentially expressed genes ($P < 0.05$) between schizophrenia cases and controls in neuronal cell types utilizing an existing dataset (47). Prioritized genes were enriched for both up-regulated and down-regulated genes in excitatory neurons, but not in inhibitory neurons (Fig. 4D). Third, we observed that prioritized genes were enriched for multiple gene sets that have been implicated in schizophrenia, including loss-of-function intolerant genes and genes involved in autism and developmental disorders (Fig. 4D). The convergence of prioritized genes with our previous knowledge about schizophrenia genetics supports our fine-mapping strategy.

Discussion

Here, we have described a deep learning model, INTERACT, to identify DNAm regulatory variants in the human brain, and

further investigated the contribution of predicted DNAm regulatory variants to the genetic basis of brain-related traits. Our study reveals the inherent power of the DNA sequence to encode tissue-specific DNAm patterns and the ability of our tissue-specific INTERACT to learn DNA sequence features that shape tissue-specific DNAm levels. We show that genetic effects on DNAm levels predicted by the brain-specific INTERACT were not confounded by LD, were convergent with mQTL evidence from association analysis, and had discriminative power in separating causal mQTLs from null SNPs. We demonstrate that DNAm regulatory variants predicted by our brain-specific model were enriched for heritability of brain-related traits and improved polygenic risk prediction for schizophrenia across ancestry samples. Finally, we applied predicted DNAm regulatory variants for prioritizing risk genes within schizophrenia GWAS risk loci, which led to the altered interpretation of GWAS risk loci and the identification of potentially novel risk genes.

Compared to existing deep learning models, our designed INTERACT model has three key advantages in predicting functional regulatory variants. First, due to the self-attention mechanism of the transformer module, INTERACT has the advantage of detecting distant interacting features that could otherwise be missed by a standard CNN. Second, the pretraining and fine-tuning design enables knowledge transfer from WGBS datasets to EPIC array datasets, leveraging strengths and limitations of both datasets. To wit, WGBS measures a much larger number of CpG sites, but is subject to measurement bias due to variability of sequencing depth, while the EPIC array measures a lesser number of CpG sites, but is more accurate. Third, we trained the INTERACT within the biological context of brain-related traits, the human brain. As a result, compared to the nonbrain-specific model, the brain-specific model predicts DNAm regulatory variants that were more strongly converged with mQTL evidence in human brain and explained more heritability of brain-related traits.

This work should be viewed in light of several limitations. First, our model takes input DNA sequences of only 2 kb, and hence it can only predict DNAm regulatory variants located within 1 kb of each CpG site. Future work can be developed to design models that take longer DNA sequences and hence detect regulatory variants beyond 1 kb of each CpG site. Second, our model was trained to predict DNAm levels in bulk brain tissues, and the *in silico* mutagenesis we employed may not identify DNAm regulatory variants that are brain cell-type-specific. Future studies could be extended to train models that predict brain cell-type-specific DNAm levels, followed by *in silico* discovery of brain cell-type-specific DNAm regulatory variants. Third, variants derived from our model can only help uncover noncoding risk variants that act through regulation of DNAm levels, but it is not able to reveal risk variants acting through mechanisms that are independent of DNAm regulation. Our model can be extended to identify functional regulatory variants for other types of molecular traits, such as gene expression and histone modification, in the human brain or in specific brain cell types.

Methods

Training Datasets. Training datasets for the INTERACT model were based on DNAm data we collected in our previous study for comparing DNAm between brain and peripheral tissues from the same living subjects (30). Details about tissue acquisition, processing, and experimental and bioinformatics procedures related to the DNAm data were described in our prior study. Dataset S8 shows details about brain and nonbrain samples used in this study. Briefly, we

collected DNA samples from resected brain tissues of five different brain regions (frontal cortex, temporal cortex, occipital cortex, temporal pole, and hippocampus) from 21 subjects. We also collected DNA from blood, saliva, and buccal samples from the same 21 subjects. DNAm was measured by the Infinium HumanMethylationEPIC array for ~800,000 CpG sites. DNAm data were processed using the R packages Minfi (48) and RnBeads (49).

Our pretraining dataset for INTERACT was based on the WGBS dataset we generated in our previous study for two adult brain regions (the hippocampus and the dorsolateral prefrontal cortex, DLPFC); details about study samples, data generation, and data processing have been described in our prior report (31). In brief, we performed WGBS targeting 30× coverage on ~300 samples across two brain regions consisting of patients with schizophrenia and unaffected controls. The raw WGBS data were processed and aligned to the GRCh38.p12 genome. DNAm was called by Bismark (50) and smoothed by the R package bsseq (v1.18) (51). After data processing and quality control (QC), 165 DLPFC samples and 179 hippocampal samples were assessed for 29,401,795 CpG sites across the epigenome, with an average postprocessing coverage of 17.3 reads per CpG site. INTERACT was pretrained by predicting the smoothed DNAm levels across 26,416,185 autosomal CpG sites of hippocampal samples.

For both training and pretraining datasets, we split CpG sites into three independent sets by chromosomes for model training and evaluation. We used CpG sites on chromosomes 1 to 20 as the training set and CpG sites on chromosome 21 as the validation set for model tuning. CpG sites on chromosome 22 were used as the testing set for prediction performance evaluation.

INTERACT Model. INTERACT contains three main modules: CNN, transformer, and fully connected network. The model takes a one-hot encoded DNA sequence of 2 kb as input, and the DNAm level of the CpG site centered in the DNA sequence as output. We explored prediction performance for the brain-specific INTERACT model taking different lengths of DNA input (1, 2, 3, and 4 kb), and the 2-kb length generated the best prediction performance (Dataset S9), supporting our choice of 2-kb input DNA length. Details of our model architecture are provided in [SI Appendix, Table S4](#) and codes are publicly available in our GitHub. Below are descriptions of each module.

CNN module. The CNN module contains three convolution layers with 400 filters for each layer. The features learned by convolution layers are fed into a max pooling layer. The feature matrix output by the maximum pooling layer is fed into a dropout layer to avoid overfitting, with a dropout rate of 0.5.

Transformer module. Output by the CNN module is fed into the transformer module. The transformer module contains a stack of eight identical layers. Each layer includes two sublayers: the first is a multihead self-attention layer and the second is a simple, position-wise fully connected feed-forward network. In addition, there is residual connect around each sublayer and each sublayer is followed by layer normalization.

Fully connected network. The features learned by the transformer module are fed into a fully connected network for predicting DNAm levels at CpG sites in samples of the same tissue. Our fully connected network contains one hidden layer and an output layer. For pretraining, the output layer contains 179 units corresponding prediction tasks for the 179 hippocampal samples in the WGBS dataset. For fine-tuning, the output layer contains 21 units corresponding prediction tasks for the 21 samples of each tissue in the EPIC array dataset. The output layer predicts DNAm levels and scales prediction values into the 0 to 1 range using the sigmoid function.

Pretraining and Fine-Tuning. Given the model complexity of INTERACT and the limited number of CpG sites on the EPIC array as training samples, we pretrained INTERACT by predicting DNAm levels of ~26 million autosomal CpG sites in 179 hippocampal samples measured by WGBS in our previous study (31). Pretrained INTERACT was then used to fine-tune each tissue-specific INTERACT model by predicting DNAm levels of CpG sites on the EPIC array for 21 samples of each tissue. The reason we chose a WGBS dataset for pretraining instead of for final training was because WGBS suffers measurement bias due to variability of sequencing coverage across CpG sites. However, we reasoned that the large number of CpG sites measured by WGBS still contain informative features underlying DNAm levels, which can be learned by the pretrained model and further transferred to each tissue-specific model, and can hence improve prediction performance of the final model.

Alternative Models to Compare with INTERACT for Predicting DNAm Levels. We compared our fine-tuned INTERACT model with three alternative models for their performance in predicting DNAm levels. 1) INTERACT without pretraining: This was the INTERACT model started with random weights. It was trained by CpG sites on the EPIC array and no WGBS data were involved. 2) CNN model: This had the same structure of CNN module, followed by the same fully connected network module as described in the INTERACT model. 3) MRCNN: This was a CNN-based model published previously (28). It includes four convolution layers, one max-pooling layer following the second convolution layer, one dropout layer, and one fully connected network. The input of MRCNN was a one-hot encoded DNA sequence of length 400; the second convolutional layer reshapes the sequence into a two-dimensional array.

tDMRs. Following a previous approach (52), we identified tDMRs for each tissue pair based on observed or predicted DNAm levels of independent CpG sites not used for model training. First, we applied a linear mixed model to identify tDMPs that were significant after Bonferroni correction ($P < 0.05/17,421$ independent CpG sites) and had an effect size of $\geq 20\%$ difference in DNAm between two tissues. The linear mixed model was applied to each CpG, with a fixed effect for tissue type and a random effect for each individual to correct for any interindividual variation. ComBat was applied to correct DNAm levels by removing batch effects that may arise from different EPIC arrays, with sex, age, and tissue type included as covariates for batch effect estimation (53). Second, we used the tDMPs status to identify tDMRs in which at least three tDMPs were detected with an inter-CpG distance of ≤ 1 kb while allowing ≤ 3 nontDMPs within the tDMRs. Finally, we estimated the proportion of tDMRs from observed DNAm levels that can be recovered by tDMRs from predicted DNAm levels. A tDMR from observed DNAm levels was declared as recovered if its $\geq 50\%$ region overlapped with a tDMR identified by predicted DNAm levels.

DNA Motif Analysis. We examined filters in the first convolutional layer of each tissue-specific INTERACT model to identify DNA motifs associated with DNAm levels in each tissue. Following the method described previously (29), DNA motifs were discovered for each filter by a subset of sequences where the filter produces an activation value higher than half of the maximum activation value across all subsets of sequences the filter has scanned. Selected subsets of sequences are then aligned to generate a position weight matrix, which was further matched to annotated TF binding DNA motifs in the *Homo sapiens* CIS-BP database using the Tomtom v4.10.1 motif comparison tool from the MEME suite (33). We consider matches at FDR < 0.05 significant for each filter.

In Silico Mutagenesis. We performed in silico mutagenesis to identify functional genetic variations that regulate DNAm levels. Briefly, we first introduced a variant allele in a given DNA sequence of 1 kb flanking a CpG site. Each sequence (with or without the introduced variant allele) was then passed to the trained INTERACT model to get the predicted DNAm level of the CpG site centered within each sequence. The impact of the introduced variant on DNAm level of the CpG site was estimated by the difference of the predicted DNAm levels of the CpG site within the two sequences. We did this in silico mutagenesis for 9,042,066 SNPs (minor allele frequency [MAF] > 0.005) observed in the European ancestry samples of 1000 Genomes and located within a 1-kb window of 25,564,506 autosomal CpG sites, resulting in estimated effects on DNAm levels for 182,211,784 SNP-CpG pairs in each sample used for training the tissue-specific INTERACT model. SNPs were ranked in descending order by their maximum absolute values of predicted effects for all of their paired CpG sites in each sample of the same tissue used for training the tissue-specific model.

mQTLs Analysis. We obtained mQTL evidence from a dataset generated by our previous mQTL study of the prefrontal cortex from the BrainSeq consortium (15). Information on tissue acquisition, processing, curation, dissection, and experimental and bioinformatics procedures related to the methylation data, and genotype data processing was described in prior reports. Briefly, DNAm was measured using the Illumina HumanMethylation450 (450k) array for ~485,000 CpG sites. DNAm data were processed using the R package Minfi (48). GWAS data were imputed into 1000 Genomes Phase 3 variants using SHAPEIT2 (54) and IMPUTE2 (55). We computed mQTL association statistics for SNPs within 1 kb of each CpG site in a subset of 238 samples of European ancestry, adjusting

for age, sex, schizophrenia diagnosis, the top five PCs from genotype data, and the top 10 PCs from DNAm levels.

Fine-Mapping mQTLs. We performed statistical fine-mapping using SuSiE to identify causal mQTLs that drive DNAm levels of CpG sites, based on the same mQTL dataset described above (15). SuSiE (41) is an approach based on Bayesian variable selection in regression and it showed superior performance in fine-mapping compared to other approaches including CAVIAR (56), FINEMAP (57), and DAP-G (58). Specifically, we performed fine-mapping for all SNPs within 20 kb of each autosomal CpG site as captured by the 450K array, adjusting for the same set of covariates described above for mQTL analysis, in 1-kb window analysis. We set $L = 10$ for the largest number of causal mQTLs and let SuSiE estimate prior effect size from data.

S-LDSC Regression. We performed S-LDSC regression (59) to evaluate the enrichment of heritability of brain-related traits for variants ranked at different intervals by their impact on DNAm levels predicted by the INTERACT model. As a comparison, we also conducted S-LDSC analysis for mQTLs derived from association analysis, including 1,065,054 mQTLs without fine-mapping ($FDR < 0.01$) and 858,174 fine-mapped mQTLs ($pip > 0.1$) as described above. We selected 13 brain-related traits with GWAS sample sizes $> 40,000$, including a wide range of behavioral traits, and psychiatric and neurological disorders. We also included one nonbrain trait, human height, as a control to examine whether our findings are specific to brain-related traits. We downloaded GWAS summary statistics of each trait from the sources listed in [Dataset S10](#). Following recommendations from the LDSC resource website (<https://alkesgroup.broadinstitute.org/LDSCORE>), S-LDSC was run for each list of variants with the baseline LD model v2.2 that included 97 annotations to control for the LD between variants with other functional annotations in the genome. We used HapMap Project Phase 3 SNPs as regression SNPs, and 1000 Genomes SNPs of European ancestry samples as reference SNPs, which were all downloaded from the LDSC resource website.

PRS for Schizophrenia. We calculated schizophrenia PRS in three independent samples, including two samples of European ancestry (EA1, 178 cases, 343 controls; EA2, 660 cases, 685 controls) and one sample of African ancestry (AA, 130 cases, 276 controls). The samples of EA1 and AA were postmortem brain specimens donated through the Offices of the Chief Medical Examiners of the District of Columbia and of the Commonwealth of Virginia, Northern District to the National Institute of Mental Health Brain Tissue Collection at the NIH in Bethesda, Maryland, according to NIH Institutional Review Board guidelines (Protocol #90-M-0142). Autotaped informed consent was obtained from legal next of kin on every case (as these donations occurred after death, and thus the donors themselves could not consent). Details of the donation process have been described previously (60, 61). The sample of EA2 included participants that were all Caucasians of European ancestry and were selected from the Clinical Brain Disorders Branch Sibling Study of schizophrenia at the National Institute of Mental Health (D.R.W., Principal Investigator).

GWAS data for each sample were collected by different types of Illumina microarray over the years. We imputed GWAS data into 1000 Genomes Phase 3 variants based on a common set of SNPs ($\sim 300,000$) across all chips in each study sample, following the best-practice guidelines of IMPUTE2. In brief, pre-phasing was first performed with SHAPEIT to infer haplotypes for samples based on autosomal SNPs with an MAF greater than 0.01. Imputation was carried out on prephased haplotypes using IMPUTE2 against reference data from the 1000 Genomes Phase 3 haplotype reference. After postimputation QC (SNP missing rate < 0.05 , MAF > 0.01 , imputation quality score > 0.9 , and Hardy-Weinberg equilibrium $> 10^{-5}$), 6,229,206, 5,947,269, and 4,993,796 autosomal variants were retained for EA1, EA2, and AA samples, respectively. To make PRS more comparable across the three samples, we used the same set of 4,834,021 SNPs that passed QC across all three samples. Imputed dosage data of SNPs were used for calculating the PRS in each sample, as described below.

We computed PRS by the standard approach of LD pruning and P -value thresholding, but used three types of SNPs resulting in three types of PRS: fPRS, sPRS, and rPRS. 1) fPRS was computed from the top 20% of variants ranked by their predicted effects on DNAm levels by the brain-specific INTERACT model; 2) sPRS was computed from all of the SNPs that passed QC; 3) for rPRS, to make fPRS and sPRS more comparable, we randomly selected a subset pruned SNPs for sPRS to match the number of pruned SNPs for fPRS at each P -value

threshold. We repeated the process 100 times and the average predictive performance of rPRS across 100 iterations was compared with fPRS. SNPs were pruned using P -value informed clumping in PLINK (62), with a cutoff of $r^2 = 0.1$ within a 500-kb window. We calculated 18 levels of PRS using pruned SNPs under 18 different thresholds of P values: $5e-08$, $1e-07$, $1e-06$, $1e-05$, $1e-04$, 0.001 , 0.01 , 0.05 , 0.1 , 0.2 , 0.3 , 0.4 , 0.5 , 0.6 , 0.7 , 0.8 , 0.9 and 1 . Weights used for PRS calculation were the natural log (odds ratio) of SNPs estimated from a metaanalysis of PGC3 GWAS of European ancestry samples (63), from which we removed samples used here for PRS calculation. We tested associations between each type of PRS and schizophrenia case-control status adjusting for population stratification using the top five PCs from genotype data as covariates. The portion of the variance in schizophrenia case-control status explained by the PRS was assessed by the Nagelkerke pseudo R^2 .

Fine-Mapping Schizophrenia GWAS Risk Loci.

Define GWAS risk loci. We clumped significant SNPs from schizophrenia GWAS summary statistics [PGC3, European ancestry samples (64)] into 176 independent risk loci on autosomes using PLINK. Specifically, we first obtained 299 index SNPs ($P < 5 \times 10^{-8}$) that were LD-independent and had $r^2 < 0.1$ within a 3-Mb window. Risk loci for each index SNP were defined as being 50-kb upstream of the leftmost and 50-kb downstream of the rightmost SNPs that were within a 3-Mb window of an index SNP and had $r^2 > 0.2$ with the index SNP. Risk loci were merged if they were within 50 kb, resulting in 176 independent risk loci.

Prioritize candidate causal SNPs. We prioritized putative causal SNPs within risk loci by overlapping significant SNPs ($P < 5 \times 10^{-8}$) and their tagged SNPs ($r^2 > 0.9$) with DNAm regulatory variants predicted by the brain-specific model across all brain samples used for training the model. DNAm regulatory variants were top ranked variants with predicted effect sizes larger than a predefined cutoff that was determined for each training brain sample separately using our assembled gold-standard dataset (SNP-CpG pairs with high-confident causal mQTLs or null SNPs). Specifically, we predicted SNP effect on DNAm level for each SNP-CpG pair in each training brain sample. We then ranked all SNP-CpG pairs based on the absolute value of their predicted effect sizes. We calculated the FDR at each data point of the ranked list and the effect cutoff was set as the smallest effect at which $FDR \leq 0.05$ could be achieved. The effect cutoffs were ~ 0.05 across 21 brain samples used for training the brain-specific INTERACT model ([Dataset S6](#)).

Assign candidate causal SNPs to target genes. Utilizing an epigenomics dataset of neuronal cell types in the human brain (65), we first assigned prioritized risk variants to putative causal genes if the genes had active promoters in the neuronal cell type that overlapped with prioritized risk variants. Putative causal genes were also assigned if they had active promoters that could be connected to prioritized risk variants through distal regulation. Details for identification of active promoters and their interactions with distal regulatory regions in the neuronal cell types were described in the prior report (65). Briefly, active promoters were identified by H3K4me3 peaks that overlap H3K27ac within 2,000 bp of a nearest transcription start site. Interactions between active promoters and distal regulatory regions were detected by proximity ligation-assisted chromatin immunoprecipitation sequencing (PLAC-seq). To link risk variants to distal promoters, each risk variant was extended by 2,500 bp in both directions and then intersected with PLAC-seq bins that had significant interactions in neurons. If one of the PLAC-seq bins overlapped a risk variant-region and another bin overlapped an active promoter, the active promoter was then distally linked with the risk variant. If the risk variant could not be assigned to overlapping or distal active promoters in neurons, we then assigned them to genes according to proximity-based annotations from the Ensembl Variant Effect Predictor (66).

Gene Set Enrichment Analysis. We used clusterProfiler for gene ontology enrichment analysis (67). We also tested the enrichment of prioritized schizophrenia risk genes for a number of gene sets: 1) differentially expressed genes ($P < 0.05$) between schizophrenia cases and controls in excitatory and inhibitory neuronal cell types (47); 2) 3,123 loss-of-function intolerant genes ($pLI > 0.9$) (68); 3) 102 significant ($FDR < 10\%$) autism-spectrum disorder genes (69); 4) 293 genes enriched for de novo mutation in developmental disorder cases (70). Enrichment was tested by logistic regression with Firth's bias reduction using the R package "logistf." All analyses were adjusted for gene size.

Data, Materials, and Software Availability. The codes in this paper have been deposited in GitHub, <https://github.com/LieberInstitute/INTERACT> (71). All other data are available in the main text and supporting information.

ACKNOWLEDGMENTS. We are grateful for the vision and generosity of the Lieber and Maltz families, who made this work possible; the families who donated to this research; two anonymous reviewers for their thoughtful criticisms, comments, and suggestions on early versions of the manuscript; Dr. Olga Troyanskaya and Dr. Christopher Park for their review of the manuscript and helpful suggestions; and the Psychiatric Genomics Consortium, UK BioBank, GSCAN (the GWAS & Sequencing Consortium of Alcohol and Nicotine), the International Parkinson's Disease Genomics Consortium, and the CTGlab (the Complex Trait Genetics laboratory at the VU University Amsterdam the Amsterdam University Medical Centre) for making their GWAS results publicly available, Computing support from The Joint High Performance Computing Exchange (J H P C E) facility in the Department of Biostatistics at the Johns Hopkins Bloomberg School of Public

Health and Maryland Advanced Research Computing Center (MARCC). This study was supported by NIH Grants R01MH121394, R01MH112751, and R01AA024486 (to S.H.), R01MH119165 (to G.S.). Part of the human brain data from this study was generated from the Lieber Institute for Brain Development and supported by Institute funds.

Author affiliations: ^aLieber Institute for Brain Development, The Johns Hopkins Medical Campus, Baltimore, MD 21287; ^bDepartment of Psychiatry and Behavioral Sciences, The Johns Hopkins University School of Medicine, Baltimore, MD 21287; ^cDepartment of Genetic Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD 21205; ^dDepartment of Neuroscience, The Johns Hopkins University School of Medicine, Baltimore, MD 21205; ^eDepartment of Mental Health, The Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205; ^fDepartment of Biostatistics, The Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205; ^gDepartment of Neurology, The Johns Hopkins University School of Medicine, Baltimore, MD 21205; and ^hDepartment of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Palo Alto, CA 94305

- Schizophrenia Working Group of the Psychiatric Genomics Consortium, Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
- N. R. Wray *et al.*; eQTLGen; 23andMe; Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).
- E. A. Stahl *et al.*; eQTLGen Consortium; BIOS Consortium; Bipolar Disorder Working Group of the Psychiatric Genomics Consortium, Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat. Genet.* **51**, 793–803 (2019).
- D. J. Schaid, W. Chen, N. B. Larson, From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
- F. Zhang, J. R. Lupski, Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24** (R1), R102–R110 (2015).
- M. T. Maurano *et al.*, Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
- C. Colantuoni *et al.*, Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature* **478**, 519–523 (2011).
- S. Numata *et al.*, DNA methylation signatures in development and aging of the human prefrontal cortex. *Am. J. Hum. Genet.* **90**, 260–272 (2012).
- L. Ma *et al.*; BrainSeq Consortium, Schizophrenia risk variants influence multiple classes of transcripts of sorting nexin 19 (SNX19). *Mol. Psychiatry* **25**, 831–843 (2020).
- N. E. Banovich *et al.*, Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet.* **10**, e1004663 (2014).
- X. Zhang *et al.*, Linking the genetic architecture of cytosine modifications with human complex traits. *Hum. Mol. Genet.* **23**, 5893–5905 (2014).
- T. R. Gaunt *et al.*, Systematic identification of genetic influences on methylation across the human life course. *Genome Biol.* **17**, 61 (2016).
- J. L. McClay *et al.*; Swedish Schizophrenia Consortium, High density methylation QTL analysis in human blood via next-generation sequencing of the methylated genomic DNA fraction. *Genome Biol.* **16**, 291 (2015).
- H. Schulz *et al.*, Genome-wide mapping of genetic determinants influencing DNA methylation and gene expression in human hippocampus. *Nat. Commun.* **8**, 1511 (2017).
- A. E. Jaffe *et al.*, Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nat. Neurosci.* **19**, 40–47 (2016).
- C. Giambartolomei *et al.*; CommonMind Consortium, A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* **34**, 2538–2545 (2018).
- E. Hannon *et al.*, Leveraging DNA-methylation quantitative-trait loci to characterize the relationship between methylomic variation, gene expression, and complex traits. *Am. J. Hum. Genet.* **103**, 654–665 (2018).
- S. Han *et al.*, Integrating brain methylome with GWAS for psychiatric risk gene discovery. *BioRxiv* [Preprint] (2018). <https://www.biorxiv.org/content/10.1101/440206v1> (Accessed 11 January 2022).
- J. Zhou *et al.*, Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).
- G. E. Hoffman, J. Bendl, K. Girdhar, E. E. Schadt, P. Roussos, Functional interpretation of genetic variants using deep learning predicts impact on chromatin accessibility and histone modification. *Nucleic Acids Res.* **47**, 10597–10611 (2019).
- D. R. Kelley *et al.*, Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* **28**, 739–750 (2018).
- K. Van Roey, N. E. Davey, Motif co-regulation and co-operativity are common mechanisms in transcriptional, post-transcriptional and post-translational regulation. *Cell Commun. Signal.* **13**, 45 (2015).
- S. Hochreiter, J. Schmidhuber, Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
- A. Vaswani *et al.*, Attention is all you need. *arXiv [Preprint]* (2018) <https://arxiv.org/abs/1706.03762> (Accessed 11 January 2022).
- Y. Ji, Z. Zhou, H. Liu, R. V. Davuluri, DNABERT: Pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* **13**, btob083 (2021).
- N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, M. Linal, ProteinBERT: A universal deep-learning model of protein sequence and function. *Bioinformatics* **38**, btac020 (2022).
- J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
- Q. Tian *et al.*, MRCNN: A deep learning model for regression of genome-wide DNA methylation. *BMC Genomics* **20** (suppl. 2), 192 (2019).
- C. Angermueller, H. J. Lee, W. Reik, O. Stegle, DeepCpG: Accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* **18**, 67 (2017).
- P. R. Braun *et al.*, Genome-wide DNA methylation comparison between live human brain and peripheral tissues within individuals. *Transl. Psychiatry* **9**, 47 (2019).
- K. A. Perzel Mandell *et al.*, Genome-wide sequencing-based identification of methylation quantitative trait loci and their role in schizophrenia risk. *Nat. Commun.* **12**, 5251 (2021).
- C. A. Markunas *et al.*, Genome-wide DNA methylation differences in nucleus accumbens of smokers vs. nonsmokers. *Neuropsychopharmacology* **46**, 554–560 (2021).
- S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, W. S. Noble, Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).
- J. Aruga, T. Inoue, J. Hoshino, K. Mikoshiba, Zic2 controls cerebellar development in cooperation with Zic1. *J. Neurosci.* **22**, 218–225 (2002).
- K. Zhang *et al.*, Imbalance of excitatory/inhibitory neuron differentiation in neurodevelopmental disorders with an NR2F1 point mutation. *Cell Rep.* **31**, 107521 (2020).
- M. T. Heneka, G. E. Landreth, PPARs in the brain. *Biochim. Biophys. Acta* **1771**, 1031–1045 (2007).
- M. Y. Chiang *et al.*, An essential role for retinoid receptors RARbeta and RXRgamma in long-term potentiation and depression. *Neuron* **21**, 1353–1361 (1998).
- M. Nomoto *et al.*, Dysfunction of the RAR/RXR signaling pathway in the forebrain impairs hippocampal memory and synaptic plasticity. *Mol. Brain* **5**, 8 (2012).
- L. D. Moore, T. Le, G. Fan, DNA methylation and its basic function. *Neuropsychopharmacology* **38**, 23–38 (2013).
- Z. Sun *et al.*, EGR1 recruits TET1 to shape the brain methylome during development and upon neuronal activity. *Nat. Commun.* **10**, 3892 (2019).
- G. Wang, A. Sarkar, P. Carbonetto, M. Stephens, A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. B* **82**, 1273–1300 (2020).
- N. G. Skene *et al.*; Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).
- C. Barkus *et al.*, What causes aberrant salience in schizophrenia? A role for impaired short-term habituation and the GRIA1 (GluA1) AMPA receptor subunit. *Mol. Psychiatry* **19**, 1060–1070 (2014).
- A. L. Moon, N. Haan, L. S. Wilkinson, K. L. Thomas, J. Hall, CACNA1C: Association with psychiatric disorders, behavior, and neurogenesis. *Schizophr. Bull.* **44**, 958–965 (2018).
- H. Yuan, C. M. Low, O. A. Moody, A. Jenkins, S. F. Traynelis, Ionotropic GABA and glutamate receptor mutations and human neurologic diseases. *Mol. Pharmacol.* **88**, 203–217 (2015).
- A. Oguro-Ando *et al.*, Cntn4, a risk gene for neuropsychiatric disorders, modulates hippocampal synaptic plasticity and behavior. *Transl. Psychiatry* **11**, 106 (2021).
- W. B. Ruzicka *et al.*, Single-cell dissection of schizophrenia reveals neurodevelopmental-synaptic axis and transcriptional resilience. *medRxiv [Preprint]* (2020). <https://www.medrxiv.org/content/10.1101/2020.11.06.20225342v1> (Accessed 11 January 2022).
- M. J. Aryee *et al.*, Minfi: A flexible and comprehensive Bioconductor package for the analysis of Illumina DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
- Y. Assenov *et al.*, Comprehensive analysis of DNA methylation data with RnBeads. *Nat. Methods* **11**, 1138–1140 (2014).
- F. Krueger, S. R. Andrews, Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
- K. D. Hansen, B. Langmead, R. A. Irizarry, BSsmooth: From whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.* **13**, R83 (2012).
- R. C. Sliker *et al.*, Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array. *Epigenetics Chromatin* **6**, 26 (2013).
- W. E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
- O. Delaneau, J. Marchini, J. F. Zagury, A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).
- B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, G. R. Abecasis, Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
- F. Hormozdizari, E. Kostem, E. Y. Kang, B. Pasaniuc, E. Eskin, Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).
- C. Benner *et al.*, FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
- X. Wen, Y. Lee, F. Luca, R. Pique-Regi, Efficient integrative multi-SNP association analysis via deterministic approximation of posteriors. *Am. J. Hum. Genet.* **98**, 1114–1129 (2016).
- H. K. Finucane *et al.*; ReproGen Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; RACI Consortium, Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

60. A. Deep-Soboslay *et al.*, Reliability of psychiatric diagnosis in postmortem research. *Biol. Psychiatry* **57**, 96–101 (2005).
61. B. K. Lipska *et al.*, Critical factors in gene expression in postmortem human brain: Focus on studies in schizophrenia. *Biol. Psychiatry* **60**, 650–658 (2006).
62. S. Purcell *et al.*, PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
63. V. Trubetskoy *et al.*, Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* . 2022 Apr;604(7906):502–508. doi: 10.1038/s41586-022-04434-5. Epub 2022 Apr 8.
64. V. Trubetskoy *et al.*; Indonesia Schizophrenia Consortium; PsychENCODE; Psychosis Endophenotypes International Consortium; SynGO Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium, Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).
65. A. Nott *et al.*, Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. *Science* **366**, 1134–1139 (2019).
66. W. McLaren *et al.*, The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
67. G. Yu, L. G. Wang, Y. Han, Q. Y. He, clusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
68. M. Lek *et al.*; Exome Aggregation Consortium, Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
69. F. K. Satterstrom *et al.*; Autism Sequencing Consortium; iPSYCH-Broad Consortium, Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* **180**, 568–584.e23 (2020).
70. J. Kaplanis *et al.*; Deciphering Developmental Disorders Study, Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).
71. J. Zhou *et al.*, Deep learning predicts DNA methylation regulatory variants in the human brain and elucidates the genetics of psychiatric disorders. GitHub. <https://github.com/LieberInstitute/INTERACT>. Deposited 12 April 2022.