PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

DeepGenePrior: A deep learning model for prioritizing genes affected by copy number variants

Zahra Rahaie[1], Hamid R. Rabiee[1]*, Hamid Alinejad-Rokny[2]*

1 BCB Group, DML, Department of Computer Engineering, Sharif University of Technology, Tehran, Iran, **2** UNSW Biomedical Machine Learning Lab (BML), the Graduate School of Biomedical Engineering, UNSW Sydney, Sydney, Australia

* rabiee@sharif.edu (HRR); h.alinejad@unsw.edu.au (HA-R)

Citation: Rahaie Z, Rabiee HR, Alinejad-Rokny H (2023) DeepGenePrior: A deep learning model for prioritizing genes affected by copy number variants. PLoS Comput Biol 19(7): e1011249. https://doi.org/10.1371/journal.pcbi.1011249

Editor: William Stafford Noble, University of Washington, UNITED STATES

Received: December 7, 2022

Accepted: June 6, 2023

Published: July 24, 2023

Data Availability Statement: The code and the data are available at: http://git.dml.ir/z_rahaie/DGP.

Funding: HRR was partially supported by IR National Science Foundation (INSF), Grant No. 96006077. This work was also supported by the UNSW Scientia Program Fellowship and the Australian Research Council Discovery Early Career Researcher Award (DECRA) under Grant No. DE220101210 to HAR. This study makes use of data generated by the DECIPHER community. A complete list of centers that contributed to the generation of the data is available from https://

Abstract

The genetic etiology of brain disorders is highly heterogeneous, characterized by abnormalities in the development of the central nervous system that lead to diminished physical or intellectual capabilities. The process of determining which gene drives disease, known as "gene prioritization," is not entirely understood. Genome-wide searches for gene-disease associations are still underdeveloped due to reliance on previous discoveries and evidence sources with false positive or negative relations. This paper introduces DeepGenePrior, a model based on deep neural networks that prioritizes candidate genes in genetic diseases. Using the well-studied Variational AutoEncoder (VAE), we developed a score to measure the impact of genes on target diseases. Unlike other methods that use prior data to select candidate genes, based on the "guilt by association" principle and auxiliary data sources like protein networks, our study exclusively employs copy number variants (CNVs) for gene prioritization. By analyzing CNVs from 74,811 individuals with autism, schizophrenia, and developmental delay, we identified genes that best distinguish cases from controls. Our findings indicate a 12% increase in fold enrichment in brain-expressed genes compared to previous studies and a 15% increase in genes associated with mouse nervous system phenotypes. Furthermore, we identified common deletions in *ZDHHC8*, *DGCR5*, and *CATG00000022283* among the top genes related to all three disorders, suggesting a common etiology among these clinically distinct conditions. DeepGenePrior is publicly available online at http://git.dml.ir/z_rahaie/DGP to address obstacles in existing gene prioritization studies identifying candidate genes.

Author summary

DeepGenePrior is a deep learning-based method for prioritizing genes in genetic diseases. Conventional tools utilize the guilt by association principle, which relies on prior knowledge to identify novel genes. In contrast, our method does not use any prior information. Furthermore, other tools rely on auxiliary data, including false positive or negative relations, which may lead to erroneous associations. Another group of methods relies on

deciphergenomics.org/about/stats and via email from contact@deciphergenomics.org. Funding for the DECIPHER project was provided by Welcome, Grant No. WT223718/Z/21/Z. Those who carried out the original analysis and collection of the data in DECIPHER project bear no responsibility for the analysis or interpretation of the analyses provided in this study. Analysis was made possible with computational resources provided by the UNSW BioMedical Machine Learning Laboratory (BML) Servers with funding from the UNSW Scientia Program Fellowship. The funders had no role in study design, data collection and analysis, the decision to publish, or the preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

hypothesis testing, and fundamental issues regarding this group have been widely discussed in different papers.

We compared the results of DeepGenePrior with both statistical and machine learning studies against biological and classification benchmarks. Our method's results outperformed current works in three brain disorders: autism, schizophrenia, and developmental delay.

Introduction

Brain Disorders (BD) [1] are a group of disorders that affect the development of the nervous system, leading to dysfunctional brain functions that can influence memory, emotion, and learning ability. Well-studied loci associated with autism (a type of BD) include deletions in 16p11.2 [2–4] and duplications in 15q3 [5,6]. Genetic factors related to autism include TBX1 (involved in the regulation of development and associated with the 22q11.2 deletion syndrome), SHANK3 (a synaptic scaffolding gene), NLGN4 (a neuroligin gene), PCDH10 (a protocadherin gene), and NHE9 [7,8]. Other genes such as NRXN1, SHANK2, CNTN4, CNTNAP2, DPYD, DPP6, RFWD2, NLGN1, ASTN2, SYNGAP1, and DLGAP2, as well as DDX53-PTCHD1, are candidate genes for autism.

Schizophrenia (SCZ) is another disorder under the umbrella of brain disorders. CNVs disrupt several genes associated with SCZ, including TBX1 (also associated with autism), ERBB4 (encodes a receptor for NDF/heregulin), SLC1A3 (a glutamate transporter), RAPGEF4 (a nucleotide exchange factor), and CIT (a neuronal Rho-target gene) [7,8]. 7q11.2 and 15q13.3 have been reported as associated with SCZ [9]. In SCZ, a large (3 Mb) deletion on chromosome 22q11.21 is a significant risk factor [10], and other loci, including deletions at 1q21.1, deletions at 3q29, duplications of 16p11.2, deletions at 15q13.3, exonic deletions at 2p16.3, and duplications at 7q36.3, have also been reported [10]. Deletions in 1q24 (including the FMO group of genes and DNM3), 2q33.1 (SATB2), and 2p16.1 (NRXN1) are well-known variations associated with developmental delay (DD) [11].

Research on the genetics of diseases has implications for diagnosing, treating, and developing drugs for these disorders. Understanding the genetic etiology of brain disorders can provide valuable insights into effective prevention and treatment methods. Gene prioritization, the process of identifying genes that most likely contribute to a disease or phenotype, can be used in BDs. This work uses case and control copy number variants as input to prioritize causal genes associated with BDs.

The prioritization of genes relies on various types of evidence. According to [12], gene-disease associations are grouped into five categories, namely functional, cross-species, same-compartment, mutation, and textual. The first category examines molecule interactions [13], while the second category discusses homolog genes that cause similar phenotypes in other organisms [14]. Same-compartment evidence is based on the fact that the gene is involved in known disease-associated pathways or compartments, such as the cell membrane or nucleus [15]. Mutation evidence is based on Single Nucleotide Polymorphism (SNP) and structural variants, which is also the focus of this study [16]. Text evidence can be obtained from online collections like PubMed [17].

Several gene prioritization methods have been reviewed in [18–21], and from a methodological point of view, they can be classified into statistical and machine learning methods. The first group primarily employs hypothesis testing, such as exact tests like Fisher's or permutation tests, to determine whether a gene is associated or not. However, several studies have reported p-value fallacies, such as distributional assumptions, limitations in data collection,

and misleading results [22]. In addition, power loss and dependent values are discussed in detail as other criticisms of marginal p-values in [3]. Other issues can arise with these types of analyses, such as not considering all the heterogeneous features of genes.

Machine learning (ML) methods often rely on the 'guilt by association' (GBA) principle [23–25]. This principle suggests that the new genes associated with a disease interact with the most recently discovered genes in a network that encodes similarities between genes. Inference of different types of networks can then lead to the discovery of new genes. In other words, ML methods require seed data (in this case, genes that implicitly characterize the disorder) [18] and a similarity metric to determine which candidate genes are similar or associated with the seed(s). However, issues arise with this approach, as discussed in [23,24]. For instance, it is impossible to discover a novel gene association that does not relate to the previous ones. Additionally, genes of a novel genetic disease with unknown roots cannot be found [26,27] due to the dependency of these methods on prior information.

The issues discussed above hinder an ideal gene prioritization solution. To overcome these issues, we propose the DeepGenePrior method, which falls in the fourth category suggested in [12], as a deep learning architecture for gene prioritization. DeepGenePrior uses the well-studied autoencoder architecture with a variational learning framework. The Variational AutoEncoder [28,29] (VAE) is the stochastic variant of the autoencoder. Our method uses Copy Number Variants (CNV) data for gene prioritization. We train the network of neurons with all CNVs of cases and controls for all three diseases, followed by fine-tuning with the CNVs of the target disease. Controls and cases have zero and one CNV labels for the supervised learning phase. Finally, we build a score for every gene using the network weights and prioritize them. Fig 1 summarizes the method.

Our proposed method addresses gaps in previous studies and offers several advantages. First, it does not rely on theoretical assumptions like those in the hypothesis testing. Second, it does not require seed data, which is needed for methods based on guilt by association. Third, it does not rely on networks with false relations, like protein-protein networks.

We used CNVs from brain disorders to evaluate our method and compared them against major tools. We identified significantly mutated genes and found that our method detects genes that are 12% more enriched in brain expression than other tools. Furthermore, we compared the detected genes to those that cause nervous system phenotypes in mice and found our results to be 15% more enriched than other methods.

In addition, we examined genes that were exclusively overrepresented in one gender and analyzed the relationships between the detected genes and various phenotypes in the DECIPHER data source and the gene ontology of the putative genes. We found three genes common among the top genes associated with all three diseases: *ZDHHC8*, *DGCR5*, and *CATG00000022283*. According to the literature [30], defects found in *ZDHHC8* can be linked to susceptibility to schizophrenia. Also, we found that deletions in *CYFIP1*, *PRODH*, *XXBAC*, *B444P24*, *LINC00896*, *ZDHHC8*, *AC006547*, *NIPA2*, *RTN4R*, *NIPA1*, and *TUBGCP5* are associated with schizophrenia and developmental delay.

The following section describes our algorithm, the data we used, and the experiments we conducted. We then discuss our results before presenting our conclusions and future work in the final section.

Results

Prioritization of the genes in BDs

A deep learning model was utilized to identify the genes associated with BDs. The model was trained using copy number variants (CNVs) from all cases and controls, and the resulting

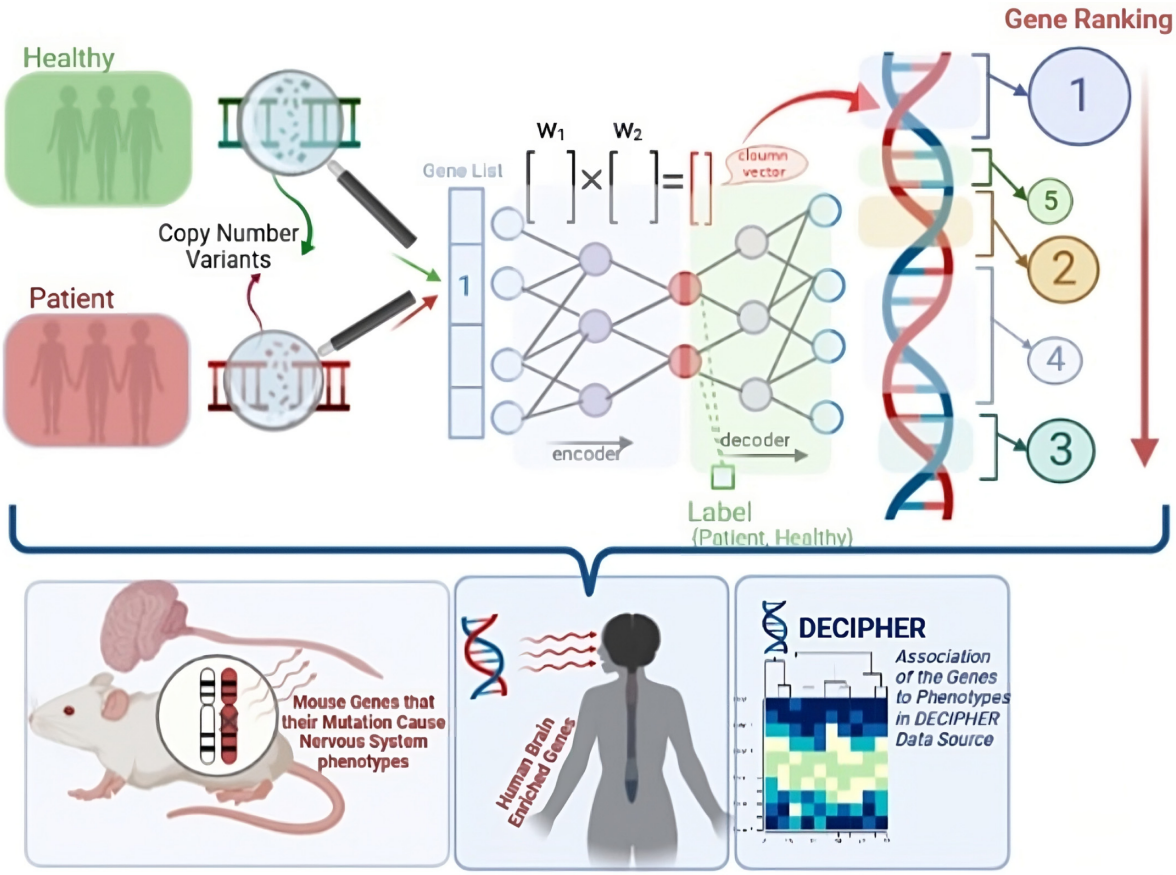

Fig 1. Summary of the Method and Analyses of the Results (Created with BioRender.com). A deep learning model learns the distinctions between cases and controls; then, the learned weights are used to prioritize the genes. After training, the results are evaluated using mutant mouse genes, human brain-enriched genes, DECIPHER data, and gene ontology analyses.

https://doi.org/10.1371/journal.pcbi.1011249.g001

model weights were employed to determine scores. The UCSC Lift Genome Annotations [31] tool was employed to convert all CNVs to the hg19 genome, and the locations of all CNVs were confirmed using NCBI remap tools [32]. CNVs smaller than one kilobase pair were excluded from the analysis.

The study showcases Tables 1, 2, and 3, displaying the top 40 genes for each disorder, accompanied by their respective p-values. Table 4 illustrates the methodology employed for Fisher's exact test. Specifically, CaseOV represents the overlaps between cases and the genes, while ControlOV represents the overlaps between controls and the genes.

The study presents Tables 1, 2, and 3, which exhibit the top 40 genes for each disorder along with their corresponding p-values. Table 4 shows the formulation of Fisher's exact test. CaseOV represents the number of overlaps between cases and the genes, while controlOV represents the number of overlaps between controls and the genes.

Furthermore, we examined the genes that are associated with all of the three disorders and those linked with only two of them. *COMT* deletion is common between ASD and SCZ, while deletions in *CYFIP1*, *PRODH*, *XXBAC-B444P24*, *LINC00896*, *ZDHHC8*, *AC006547*, *NIPA2*, *RTN4R*, *NIPA1*, and *TUBGCP5* are common between SCZ and DD. Next, common genes between ASD and DD are deletions in *FAM57B*, *SHANK3*, and *BDH1*, and the shared genes between the three disorders were deletions in *DGCR5* and *ZDHHC8*.

Table 1. Top 40 genes associated with developmental delay are presented. The model's top findings on the developmental delay (DD) data source are reported herein. Each row provides information on gene names, overlapping cases and controls, P-value, and the type of genetic variation.

Gene name	Status	P-Value	CaseOV	ControlOV	Gene name	Status	P-Value	CaseOV	ControlOV
TDRP	dup	5.19E-159	293	18	*NANOG*	del	6.13E-122	221	12
DGCR5	del	2.17E-159	288	15	*SLC2A14*	del	6.13E-122	221	12
PRODH	del	1.46E-154	282	16	*CYFIP1*	del	5.85E-118	230	21
LCE3E	del	1.51E-153	257	5	*NIPA2*	del	5.85E-118	230	21
ERICH1	dup	5.32E-140	275	26	*TUBGCP5*	del	1.31E-117	231	22
CATG0101427	dup	1.97E-143	260	14	*NIPA1*	del	5.85E-118	230	21
ERICH1-AS1	dup	2.72E-147	254	8	*CATG0022283*	del	5.85E-118	230	21
RP11-462G22	dup	3.88E-150	249	4	*CATG0022286*	del	5.85E-118	230	21
CATG0074892	dup	1.64E-149	248	4	*CATG0024378*	del	5.85E-118	230	21
CATG0074890	dup	6.38E-148	248	5	*SLC2A3*	del	8.47E-121	221	13
CATG0074891	dup	6.38E-148	248	5	*CATG0007863*	del	1.02E-120	219	12
RP11-462G22	dup	6.38E-148	248	5	*CATG0011162*	del	2.79E-118	215	12
CATG0074887	dup	2.69E-147	247	5	*CATG0024374*	del	2.85E-112	222	22
CATG0101432	dup	6.15E-138	251	14	*LCE3D*	del	1.40E-105	256	57
NPHP1	dup	5.77E-133	260	24	*RTN4R*	del	3.99E-116	183	0
MALL	dup	1.29E-132	261	25	*ZDHHC8*	del	3.99E-116	183	0
RP11-378A12	dup	6.89E-135	246	14	*AC006547*	del	3.99E-116	183	0
RP11-134O21	dup	9.55E-125	243	22	*LINC00896*	del	3.99E-116	183	0
CATG0117958	dup	2.42E-124	234	17	*XXBAC-B444P24*	del	3.99E-116	183	0
TMEM72-AS1	dup	1.68E-123	236	19	*OTUD7A*	dup	3.29E-97	228	46

https://doi.org/10.1371/journal.pcbi.1011249.t001

Table 2. Top 40 genes associated with schizophrenia are identified and presented. The model was trained using the schizophrenia data source, and the top results are reported herein. Each gene entry includes information on case and control overlaps, type of genetic variation, and corresponding P-value.

Gene Name	Status	P-value	CaseOV	ControlOV	Gene Name	Status	P-value	CaseOV	ControlOV
DGCR6	del	1.79E-06	129	60	*MED15*	del	1.79E-16	59	1
PRODH	del	7.19E-06	130	64	*DGCR8*	del	1.17E-17	58	0
DGCR5	del	7.19E-06	130	64	*CATG0058213*	del	1.17E-17	58	0
AC009133	dup	2.11E-14	66	5	*TMEM219*	dup	6.54E-14	61	4
MVP	dup	1.03E-14	64	4	*PTPRT*	del	2.77E-05	110	53
CDIPT	dup	1.03E-14	64	4	*CATG0058203*	del	2.31E-17	57	0
SEZ6L2	dup	1.03E-14	64	4	*CATG0058206*	del	2.31E-17	57	0
CATG0027072	dup	1.03E-14	64	4	*CATG0058209*	del	2.31E-17	57	0
CDIPT-AS1	dup	1.03E-14	64	4	*CLTCL1*	del	2.31E-17	57	0
ASPHD1	dup	1.90E-14	63	4	*COMT*	del	3.47E-16	58	1
TRMT2A	del	5.97E-18	59	0	*NIPA2*	del	0.00012	97	48
RANBP1	del	5.97E-18	59	0	*NIPA1*	del	0.00012	97	48
ZDHHC8	del	5.97E-18	59	0	*CATG0022283*	del	0.00012	97	48
AC006547	del	5.97E-18	59	0	*CATG0022286*	del	0.00012	97	48
LINC00896	del	5.97E-18	59	0	*CYFIP1*	del	0.000179	97	49
XXBAC-B444P24	del	5.97E-18	59	0	*TUBGCP5*	del	0.000179	97	49
QPRT	dup	1.98E-12	65	7	*CATG0024378*	del	0.00016	96	48
KCTD13	dup	3.53E-14	62	4	*BOLA2B*	dup	2.95E-11	51	4
PAGR1	dup	3.53E-14	62	4	*CATG0022287*	del	0.000191	94	47
RTN4R	del	1.17E-17	58	0	*AC023490*	del	4.04E-14	46	0

https://doi.org/10.1371/journal.pcbi.1011249.t002

Table 3. The top 40 genes associated with Autism Spectrum Disorder (ASD) are presented in this study. These genes are identified as having the highest likelihood of causing ASD based on their variations.

Gene Name	Status	P-Value	CaseOV	ControlOV	Gene Name	Status	P-Value	CaseOV	ControlOV
DGCR2	del	2.38E-44	420	3	*GABRA5*	dup	9.82E-28	235	0
ARVCF	del	3.77E-44	418	3	*OCA2*	dup	8.85E-28	234	0
GNB1L	del	3.77E-44	418	3	*CATG00000022283*	del	5.01E-08	263	32
CATG00000058206	del	6.34E-44	417	3	*CATG00000022351*	dup	2.34E-27	231	0
COMT	del	6.34E-44	417	3	*NRXN1*	del	2.08E-12	249	19
ZDHHC8	del	6.34E-44	417	3	*LINC00624*	del	3.24E-15	241	13
HIRA	del	6.34E-44	417	3	*XXBAC-B135H6*	del	1.66E-21	229	4
TBX1	del	6.34E-44	417	3	*BCL9*	del	1.01E-16	234	10
CDIPT	del	4.15E-28	387	15	*CHD1L*	del	1.01E-16	234	10
SEZ6L2	del	4.15E-28	387	15	*FMO5*	del	1.01E-16	234	10
ASPHD1	del	4.15E-28	387	15	*ACP6*	del	1.01E-16	234	10
KCTD13	del	4.15E-28	387	15	*CATG00000092640*	del	5.84E-17	141	0
CATG00000027072	del	4.15E-28	387	15	*CATG00000058020*	del	1.96E-15	142	1
CDIPT-AS1	del	4.15E-28	387	15	*RFC2*	del	1.94E-15	141	1
ALDOA	del	4.04E-28	386	15	*WBSCR22*	del	9.53E-17	140	0
FAM57B	del	4.04E-28	386	15	*GTF2I*	del	9.53E-17	140	0
CHRNA7	del	1.75E-26	239	1	*STX1A*	del	9.53E-17	140	0
DGCR5	del	3.40E-21	244	6	*EIF4H*	del	1.94E-15	141	1
GABRB3	dup	9.82E-28	235	0	*DNAJC30*	del	9.53E-17	140	0
OTUD7A	del	2.64E-26	236	1	*VPS37D*	del	9.53E-17	140	0

https://doi.org/10.1371/journal.pcbi.1011249.t003

In the subsequent sections, a comparison was made with machine learning methods, followed by a search for genes displaying brain-enriched expression. Notably, it was observed that many genes associated with brain disorders possess brain-enriched functions [33]. We compared our results with similar studies, demonstrating that our research successfully identifies more brain-enriched genes than previous investigations.

Furthermore, we compare our findings to genes that cause nervous system phenotypes in mice, which were obtained from the MGI repository [34]. Our study demonstrates a higher fold enrichment than similar studies. The next step is identifying genotype-phenotype relationships using the DECIPHER data source [35], focusing on phenotypes exhibiting high enrichment levels.

In addition, we used WebGestalt [36] to perform gene ontology analyses of coding genes, with a focus on examining Gene Ontology (GO), Human Phenotype Ontology (HPO), and associated disease terms.

Comparison with machine learning methods

Next, we compare our method with machine learning methods for the gene prioritization problem. The selected algorithms were guided backpropagation (GBP) [37], deepLIFT [38],

Table 4. A contingency table was constructed to apply Fisher's exact test. This table will be utilized in the analysis to calculate the p-value for the genes and DNA segments under investigation.

# of case samples overlapped with the gene	# of case samples not overlapped with the gene
# of control samples overlapped with the gene	# of control samples not overlapped with the gene

https://doi.org/10.1371/journal.pcbi.1011249.t004

Table 5. A comparison was conducted with other machine learning methods to assess performance. The accuracies and ROC AUCs of various machine learning techniques were reported for three datasets. It was found that DeepGenePrior outperformed the other methods, demonstrating higher accuracy and ROC AUC values.

	SCZ		ASD		DD	
	Accuracy	ROC AUC	Accuracy	ROC AUC	Accuracy	ROC AUC
DeepGenePrior	80	.86	82	.87	83	.89
DeepGenePrior without Pretraining	74	.81	73	.80	71	.79
DeepLIFT [77]	60	.65	62	.66	63	.68
Guided Backpropagation [78]	65	.72	69	.76	64	.69

https://doi.org/10.1371/journal.pcbi.1011249.t005

and DeepGenePrior (without pre-training). The third choice is to show the effect of pre-training on the performance of the whole model (an ablation study). DeepLIFT [38] is a reference-based global feature importance algorithm that uses a correlation score to measure the input's effect on the model's output. Guided backpropagation is a global feature importance that is gradient-based.

The performance benchmarks are computed as follows. The model is trained comprehensively, and important genes are selected based on their respective weights. Subsequently, the model is retrained using the identified important genes as inputs and the disease status as the output. The performance evaluation is then reported based on the test set. Global methods mainly suffer from many computations and estimates (making the model inaccurate). DeepLIFT needs a reference for calculation; the reference is very influential in the final results of the model and may cause the model to choose the wrong inputs.

Guided backpropagation needs gradients, and it has been proven that the gradients can sometimes be noisy, resulting in the selection of irrelevant features. Other methods need several simple local surrogate models to interpolate the manifold in high-dimensional models (like LIME [39]); these surrogates impose massive calculations and imprecise the model.

Some advantages of our proposed method are that it does not need the reference, does not rely on noisy data, and is not local, and there is a way to inject unlabeled as well as labeled data in the model.

The Python torch Captum [40] implementations of these algorithms were used for the comparison.

The results were reported in Table 5 regarding the accuracy and ROC AUC. Our DeepGenePrior algorithm Performs higher than the others. Besides, ROC curves are shown in Fig 2.

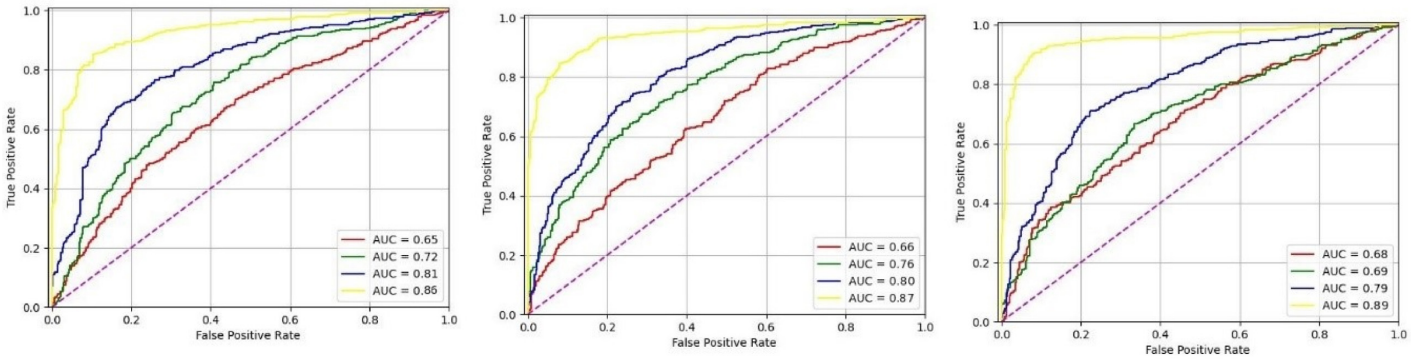

Fig 2. The Area Under the Curve (AUC) values for different machine learning methods. The yellow curve represents DeepGenePrior, the blue curve represents DeepGenePrior without Pretraining, the green curve represents Guided Backpropagation [37], and the Red is for DeepLIFT [38].

https://doi.org/10.1371/journal.pcbi.1011249.g002

Overrepresentation of tissue-specific genes

Several studies (such as [41] and [42]) claim that brain-enriched genes play an important role in BDs. To determine whether the detected genes are overrepresented in the brain tissue, we compute the fraction of coding and non-coding genes that have been enriched (background expectation) and compare it with the percentage of genes that have overlapped with deleted or duplicated CNVs.

The authors of [41] provide a list of brain-enriched genes. To obtain this list, they used the FANTOM5 CAGE-associated transcriptome [43] to identify coding and long non-coding RNA genes in the regions and examined their expression patterns across sample types.

In addition to alternative methods, we incorporated two gene prioritization tools, GeneFriends [44] and ToppGene [45], both accessible online. GeneFriends applies the guilt by association approach, while ToppGene identifies candidate genes based on functional similarity to the training gene list. However, these tools possess certain limitations. Notably, they have a restricted capacity for accommodating large datasets, require seed data for achieving results (following the guilt by association principle), and rely on parameter tuning by the user, such as setting a Pearson correlation threshold and an FDR threshold. For this analysis, the default parameter values were employed.

Fig 3 presents the results of brain-enriched coding genes fold enrichment, and Fig 4 illustrates brain-enriched lncRNA genes fold enrichment.

The results of our study are compared with those of Coe et al. [11] and Cooper et al. [46], two important studies of developmental delay. They were also compared with PLINK [47] and SNATCNV [41], publicly available tools with state-of-the-art performance.

In the list of brain-enriched genes related to ASD and SCZ, *DGCR2* specifies a protein proposed to be important in neural crest cell migration [30]. The *ZDHHC8* gene, strongly associated with ASD and SCZ [30], is another gene to note.

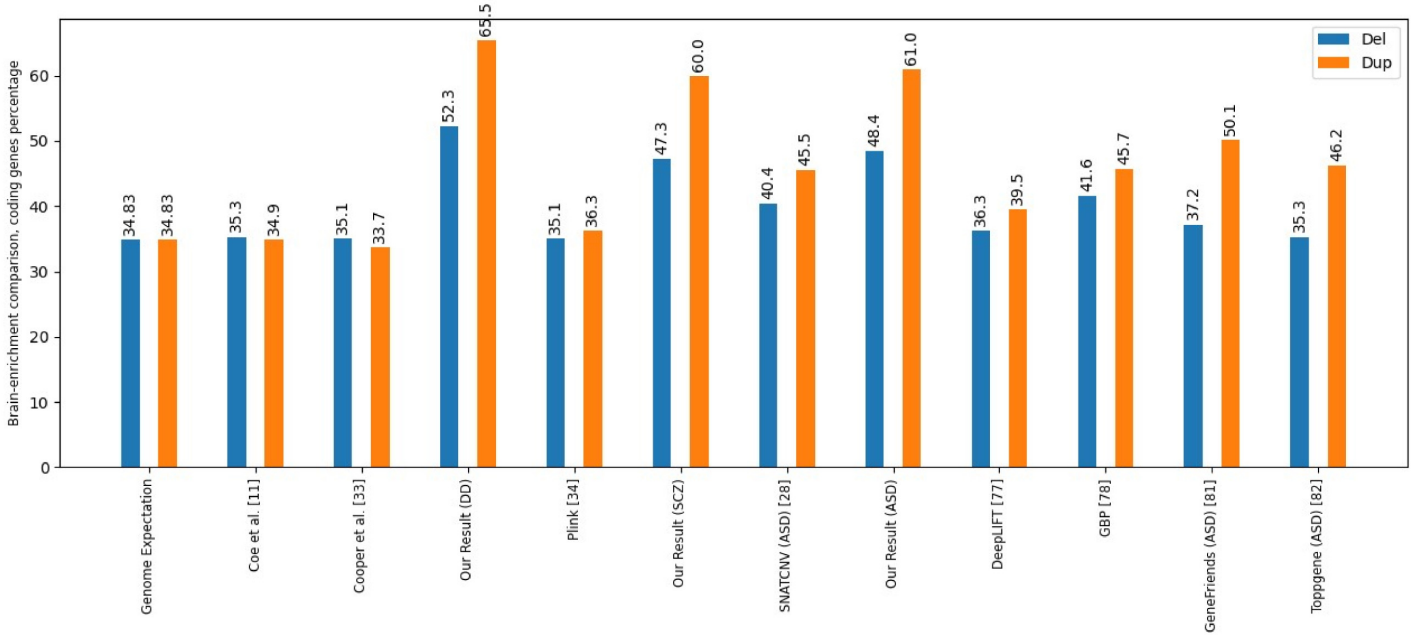

Fig 3. Brain-enrichment comparison, coding genes. This figure compares brain-enriched coding genes for different tools and methods. The percentage of brain-enriched coding genes was evaluated for two variation types, namely deletion, and duplication.

https://doi.org/10.1371/journal.pcbi.1011249.g003

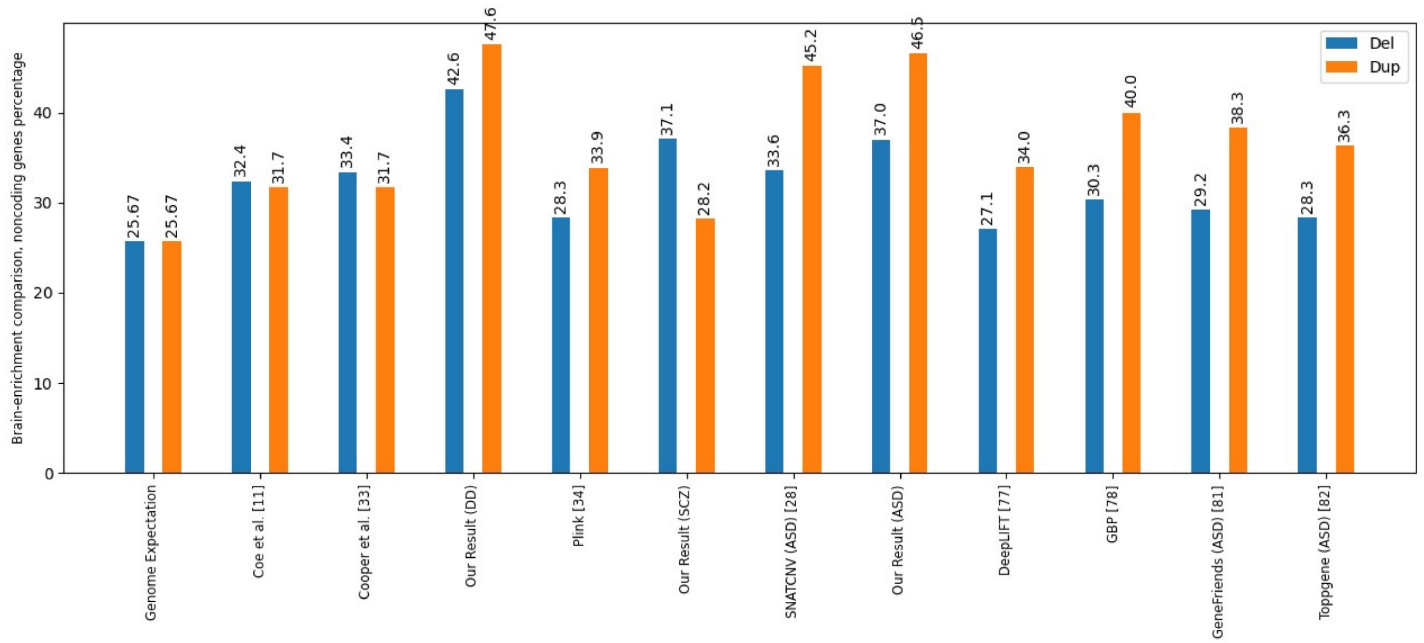

Fig 4. Brain-enrichment comparison, noncoding genes. This figure illustrates the comparison of different tools and papers based on the percentage of brain-enriched noncoding genes.

https://doi.org/10.1371/journal.pcbi.1011249.g004

Next, we have some brain-enriched genes associated with SCZ; *RTN4R* is a gene in which adult central nervous systems are likely to be affected by its role in regulating axonal regeneration and plasticity. *CATG00000058203* and *Septin5* and *CATG00000057131* are some brain-enriched genes associated with ASD and SCZ, previously mentioned in [41].

As for the developmental delay, the *DGCR5*, *PRODH*, *NIPA1*, *TUBGCP5*, *RTN4R*, *ZDHHC8*, *CRKL*, and *SERPIND1* genes are also brain-enriched and associated with the disease. Most of them are from the 22nd and 15th chromosomes (22q11.21).

Gene segregation analysis of male and female patients

Long-standing research shows that females are more tolerant of mutations than males, which explains why males are more prone to brain disorders such as autism. New studies also confirm the validity of previous findings [48–50] that male cases show more significant enrichment than female cases when comparing the ratios of cases to controls. In this research, we pointed out that some genes are more biased towards males, for example, deletion in *PHF2 (ENSG00000197724)*, duplication in *NRXN1 (ENSG00000179915)*, and deletions in *WDFY3 (ENSG00000163625)*, *PHF3 (ENSG00000118482)*, *MED13L (ENSG00000123066)*, and *WAC (ENSG00000095787)*, are more frequently seen in males than females for the developmental delay.

Besides, we performed the same analysis with ASD CNVs. We found that the *PTCHD1 (ENSG00000165186)* gene deletion occurred more in male than female patients. (Table 6 provides the details for these claims, and the chi-square test confirms the results).

DECIPHER analysis

We used DECIPHER [35], the genotype-phenotype data source for almost 12,600 patients with CNVs, to analyze phenotypes associated with candidate genes.

Table 6. Gender Bias Analysis of the Brain Disorders. This table presents the Gender Bias Analysis of Brain Disorders, highlighting genes associated with one gender.

Gene Name	% Male Cases	% Male Controls	%Cases/ %Controls (Male)	% Female Cases	% Female Controls	%Cases/ %Controls (Female)	Log2 (Male/ Female Enrichment)
PTCHD1(Del) ♂	.29	.08	3.7	.11	.07	1.6	1.21
PHF2(Del) ♂	.05	.08	.62	.01	.07	.21	1.59
NRXN1(Dup) ♂	.07	.08	.87	.04	.14	.31	1.49
WDFY3(Del) ♂	.08	.08	.99	.03	.07	.41	1.27
PHF3 (Del) ♂	.98	.15	6.53	.45	.27	1.66	1.97
MED13L(Del) ♂	.73	.08	9.125	.17	.06	2.83	2.06
WAC(Del) ♂	1.2	.4	3	.21	.31	.677	2.15

https://doi.org/10.1371/journal.pcbi.1011249.t006

To investigate the relationship between genes and phenotypes, we calculate the ratio of overlapped samples with the specific phenotype to the number of overlapped samples for a putative gene. Figs 5 (DD), 6 (SCZ), and 7 (ASD) depict the respective heatmaps for each target disorder.

Some of the highlighted phenotypes related to the target diseases are obesity (HP:0001513), autism (HP: 0000717), behavioral abnormality (HP: 0000708), irregularity of the face (HP: 00000271), and seizures (HP:0001250).

Children with autism are more likely to suffer from medical comorbidities. For example, we found macrocephaly (HP:00000256), hydrocephalus (HP:00000238), cerebral palsy (HP:0100021), migraine (HP:0002076), sleep disturbance (HP:0002360), and failure to thrive (HP:0001508) which was also mentioned in [51] as the phenotypes that co-occur in the autism. For schizophrenia, DECIPHER analysis revealed phenotypes such as obsessive-compulsive behavior (HP:0000722), anxiety (HP:0000739), and depression (HP:0000716), as well explained in [52]. *MVP* duplication, overrepresented in SCZ, is associated with depression (HP:0000716).

Regarding the developmental delay, secondary conditions such as microcephaly (HP:0000252) and anxiety (HP:0000739) can be proposed, which was also suggested in [53]; This disorder has received less research. *BCL9*, *FMO5*, and *GPR89B* deletions related to microcephaly are also overrepresented in DD. *NIPA1* duplication, associated with anxiety, is among the top genes of DD. In [53], microdeletion of the *NF1* gene is found to be associated with microcephaly and DD.

Our model deduces a set of genes for a target genetic disease. We investigated the set of phenotypes related to the genes; the specific relationship between genes and phenotypes shows that there can be diversity in the etiology of the disease, which implies that the occurrence of a phenotype in a target disease is influenced by what candidate genes are mutated in the patient.

Analysis of biological processes and phenotypic ontologies of candidate genes

As part of our analysis, we used WebGestalt [36] to investigate the associations between identified genes and specific gene ontologies (GOs), human phenotype ontologies (HPOs), and disease terms [54,55].

Some examples of the discovered disease ontology terms were intellectual disability, language development disorders, poor school performance (for the developmental delay), autistic

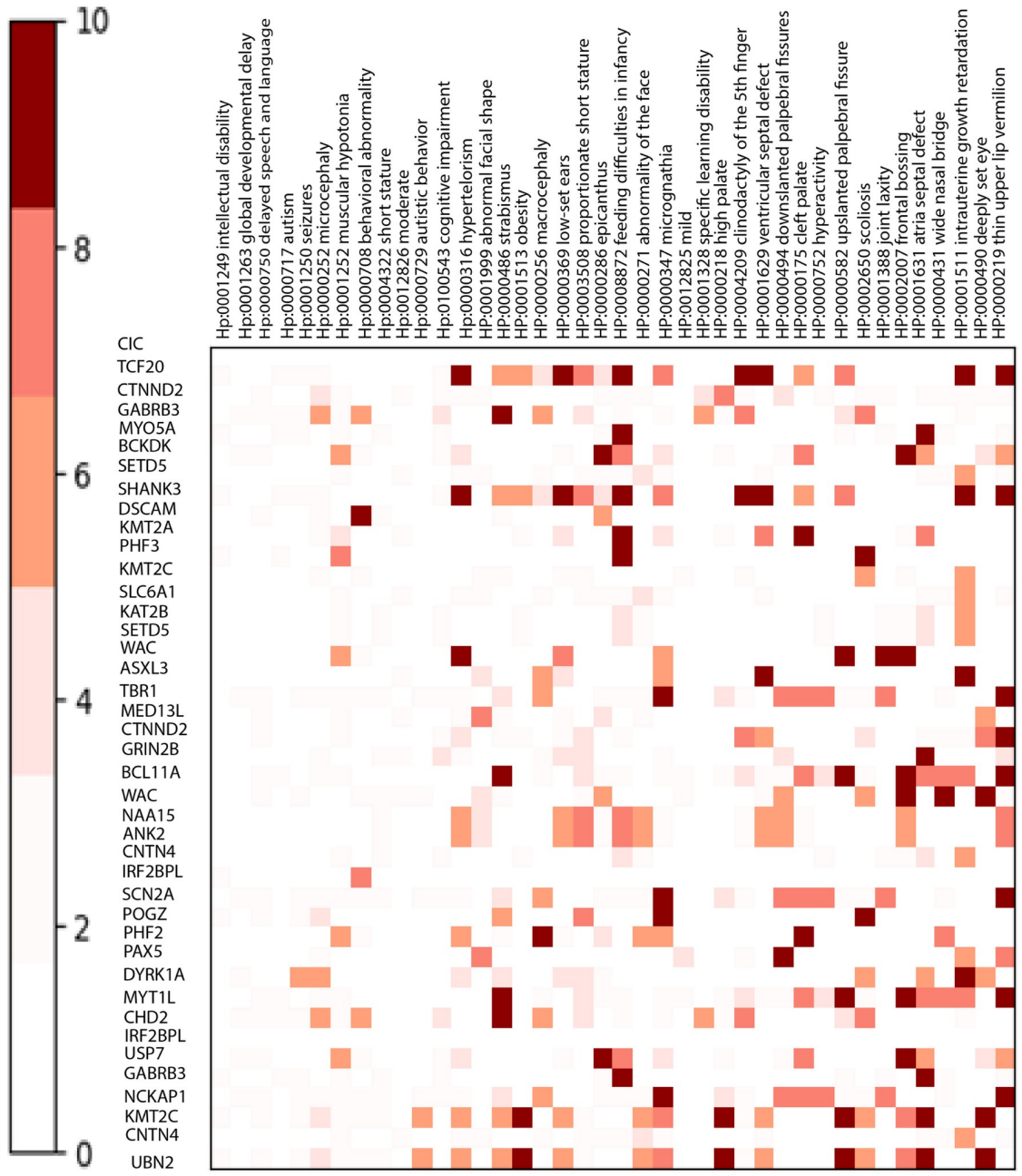

Fig 5. Heatmap for developmental delay. This figure showcases a Heatmap for Developmental Delay, providing insights into the relationship between candidate genes and DECIPHER phenotypes. The heatmap demonstrates a strong correlation between genes and phenotypes, depicted by the prominent dark red color.

https://doi.org/10.1371/journal.pcbi.1011249.g005

disorder, and language development disorders. Language development disorders are discussed in [56] as a comorbidity of BDs.

In the associated HPO terms, some examples were autistic behavior, delayed speech and language development, intellectual disability, severe global developmental delay, abnormal social behavior, impaired social interactions, and abnormally aggressive, impulsive, or violent

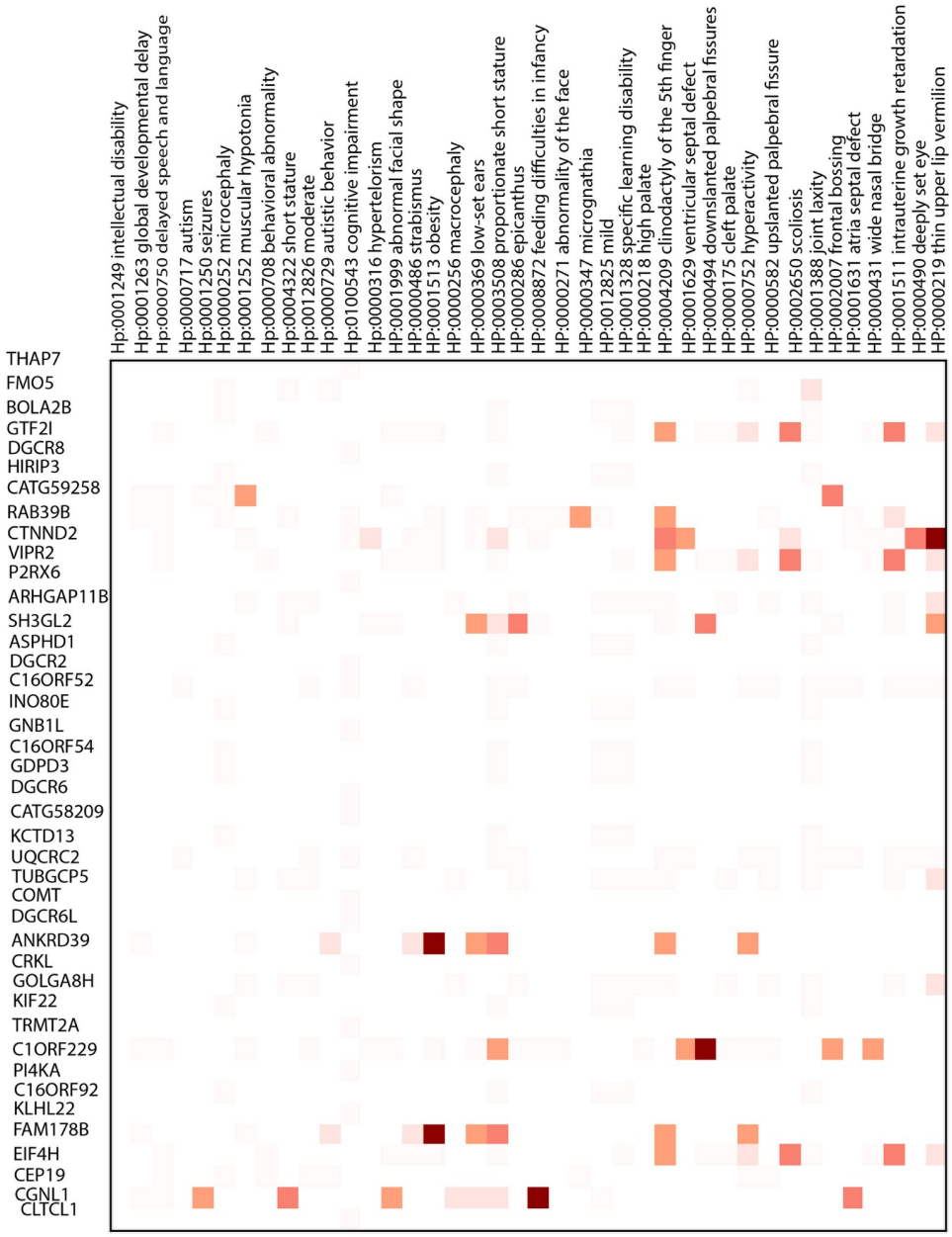

Fig 6. Heatmap for schizophrenia. This figure presents a Heatmap for Schizophrenia, where the horizontal labels represent genes associated with SCZ, and the vertical labels represent DECIPHER phenotypes. Detailed explanations of the results can be found in the accompanying text.

https://doi.org/10.1371/journal.pcbi.1011249.g006

behavior. Abnormal behavior is mentioned in [57], and impaired social interaction is discussed in [56] as phenotypes related to BDs.

The highlighted Gene Ontology terms include cognition, dendrite development, and synapse organization. In [58], dendrite development is pointed out to be associated with BDs, and the relationship between synapse organization and BDs is addressed in [59]. Tables 7, 8, and 9 summarize the results.

Fig 7. Heatmap for autism. The color legend is similar to the last heatmap.

https://doi.org/10.1371/journal.pcbi.1011249.g007

Overrepresentation of homologs of coding genes causative of nervous system phenotypes in the mutated mouse

Studying animal genetic mutations provides insight into disease mechanisms and treatments for brain disorders. Several animal models have been developed to uncover the disorder's process [60]. Mutant mice with specific defects in the nervous system are among them. Models based on mutant mice replicate key symptoms of brain disorders.

We investigate what percentage of the causative genes have homologs in mouse genes whose mutation causes nervous system phenotypes. For this purpose, we used the Mouse

Table 7. WebGestalt Analysis for Developmental Delay. Three types of analyses were conducted for genes associated with developmental delay (DD). The table presents the p-value, false discovery rate (FDR), and the number of genes for each analyzed trait.

Description	p-value	FDR	#Genes	Type of Analysis
Autism spectrum disorders	4.02E-07	6.84E-04	7	Disease Ontology Terms
Intellectual disability	2.41E-06	2.05E-03	14	Disease Ontology Terms
Language development disorders	2.81E-05	1.60E-02	3	Disease Ontology Terms
Epileptic encephalopathy	7.90E-05	2.61E-02	4	Disease Ontology Terms
Mental retardation	1.38E-04	2.61E-02	11	Disease Ontology Terms
Low intelligence	1.38E-04	2.61E-02	11	Disease Ontology Terms
Mental deficiency	1.38E-04	2.61E-02	11	Disease Ontology Terms
Poor school performance	1.38E-04	2.61E-02	11	Disease Ontology Terms
Dull intelligence	1.38E-04	2.61E-02	11	Disease Ontology Terms
Cognition	1.03E-05	8.29E-03	8	Gene Ontology
Intraspecies interaction between organisms	8.84E-05	2.36E-02	4	Gene Ontology
Covalent chromatin modification	9.89E-05	2.36E-02	7	Gene Ontology
Chemical synaptic transmission, postsynaptic	1.44E-04	2.36E-02	5	Gene Ontology
Regulation of membrane potential	1.72E-04	2.36E-02	8	Gene Ontology
Multi-organism behavior	1.76E-04	2.36E-02	4	Gene Ontology
Neuron projection guidance	2.34E-04	2.44E-02	6	Gene Ontology
Dendrite development	2.43E-04	2.44E-02	6	Gene Ontology
Neuron projection organization	5.50E-04	4.91E-02	4	Gene Ontology

https://doi.org/10.1371/journal.pcbi.1011249.t007

Genome Informatics (MGI) database to identify genes related to the mouse nervous system and their human homologs.

Fig 8 presents an analysis of the proportions of homologs among the identified genes exhibiting nervous system phenotypes in mice. The findings reveal that the coding genes identified by our method show a higher percentage of homologs in mutant mouse models displaying nervous system phenotypes compared to the results obtained from other methods. We also evaluated two gene prioritization tools, GeneFriends [44] and ToppGene [45]. For example, some genes that have orthologs in mice with nervous system phenotypes are *SEPTIN5*, *RTN4R*, and *ZDHHC8*. These genes are common among the three disorders.

Statistical analysis

Subsequently, an independent statistical analysis is conducted to compare the outcomes of DeepGenePrior with hypothesis testing performed in similar studies. Sample results are depicted in Fig 9. To assess whether the observed associations are statistically significant or occur by chance, 100,000 random permutations of case and control labels were performed. The corresponding results are illustrated in the respective diagrams.

Discussion

In this paper, we presented a deep learning approach that uses a variational autoencoder to analyze CNVs and prioritize genes within them systematically. Our deep learning model learns how features are distributed over samples, which enables us to predict the likelihood that a gene variation will cause a specific disease. We applied our method to three disorders under the umbrella term of brain disorders. We examined the results for overrepresentation of enriched brain coding, long non-coding RNA genes, and mouse orthologs with nervous system

Table 8. WebGestalt Analysis for Schizophrenia. The table displays the results of the WebGestalt analysis conducted for schizophrenia. Several traits have demonstrated significant correlations with brain disorders, as determined by various types of analysis.

Description	P-Value	FDR	#Genes	Type of Analysis
Blepharophimosis	4.04E-07	4.29E-04	5	Disease Ontology Terms
Hernia, Inguinal	7.50E-07	4.29E-04	5	Disease Ontology Terms
Chronic otitis media	1.01E-06	4.29E-04	4	Disease Ontology Terms
ear infection chronic	1.01E-06	4.29E-04	4	Disease Ontology Terms
Proteinuria	8.52E-06	9.68E-04	5	Disease Ontology Terms
Redundant skin	9.10E-06	9.68E-04	3	Disease Ontology Terms
Bunion	9.10E-06	9.68E-04	3	Disease Ontology Terms
Hallux Valgus	9.10E-06	9.68E-04	3	Disease Ontology Terms
Sloping shoulders	9.10E-06	9.68E-04	3	Disease Ontology Terms
Congenital anomaly of neck	9.10E-06	9.68E-04	3	Disease Ontology Terms
Deformity of neck	9.10E-06	9.68E-04	3	Disease Ontology Terms
Malformation of the neck	9.10E-06	9.68E-04	3	Disease Ontology Terms
Hypoplastic toenails	9.10E-06	9.68E-04	3	Disease Ontology Terms
Phonophobia	9.10E-06	9.68E-04	3	Disease Ontology Terms
Colon diverticulum anatomic structure	9.10E-06	9.68E-04	3	Disease Ontology Terms
Diverticular disease of the colon	9.10E-06	9.68E-04	3	Disease Ontology Terms
Pointed chin	1.05E-05	1.05E-03	4	Disease Ontology Terms
Sacral dimples	1.81E-05	1.62E-03	3	Disease Ontology Terms
Pulmonary Stenosis	2.30E-05	1.62E-03	4	Disease Ontology Terms
Prominent lower lip	2.89E-05	1.62E-03	4	Disease Ontology Terms
Posterior embryotoxic	4.02E-08	3.95E-05	6	Human Phenotype Ontology
Abnormality of the line of Schwalbe	4.02E-08	3.95E-05	6	Human Phenotype Ontology
Retinal vascular tortuosity	4.02E-08	3.95E-05	6	Human Phenotype Ontology
Abnormal systemic arterial morphology	6.88E-08	5.06E-05	10	Human Phenotype Ontology
Retinal arteriolar tortuosity	4.40E-07	2.59E-04	4	Human Phenotype Ontology
Abnormal aortic morphology	5.50E-07	2.70E-04	8	Human Phenotype Ontology
Abnormal concentration of calcium in the blood	6.93E-07	2.92E-04	6	Human Phenotype Ontology
Patellar dislocation	5.94E-06	1.82E-03	4	Human Phenotype Ontology
Multiple renal cysts	5.94E-06	1.82E-03	4	Human Phenotype Ontology
Tetralogy of Fallot	7.50E-06	1.82E-03	6	Human Phenotype Ontology
Abnormality of calcium homeostasis	7.50E-06	1.82E-03	6	Human Phenotype Ontology
Blepharophimosis	8.79E-06	1.82E-03	6	Human Phenotype Ontology
Conotruncal defect	8.79E-06	1.82E-03	6	Human Phenotype Ontology
Inguinal hernia	8.82E-06	1.82E-03	7	Human Phenotype Ontology
Myocardial infarction	9.25E-06	1.82E-03	4	Human Phenotype Ontology
Atrophy/Degeneration involving the corticospinal tracts	1.37E-05	2.25E-03	4	Human Phenotype Ontology
Abnormality of divalent inorganic cation homeostasis	1.38E-05	2.25E-03	6	Human Phenotype Ontology
Abnormal ventriculoarterial connection	1.38E-05	2.25E-03	6	Human Phenotype Ontology
Abnormal connection of the cardiac segments	1.59E-05	2.46E-03	6	Human Phenotype Ontology
Cholelithiasis	1.97E-05	2.63E-03	4	Human Phenotype Ontology

https://doi.org/10.1371/journal.pcbi.1011249.t008

phenotypes. Additionally, we used the DECIPHER data to investigate how variations in the identified genes influence other traits. Furthermore, we conducted gene ontology analyses.

We analyzed 118,968 case CNVs from 48,748 patients and 76,528 control CNVs from 26,063 healthy individuals for gene associations with brain disorders. Among the top 40 genes associated with developmental delay, *DGCR6*, *PRODH*, *DGCR5*, and *ZDHHC8* are

Table 9. WebGestalt analysis for autism spectrum disorder.

Description	P-value	FDR	#Genes	Type of Analysis
Autism Spectrum Disorders	3.73E-08	6.34E-05	7	Disease Ontology Terms
Autistic Disorder	1.36E-06	1.16E-03	9	Disease Ontology Terms
Language Development Disorders	1.07E-05	6.10E-03	3	Disease Ontology Terms
Intellectual Disability	9.44E-05	3.89E-02	10	Disease Ontology Terms
Autistic behavior	1.14E-04	3.89E-02	3	Disease Ontology Terms
Intraspecies Interaction Between Organisms	8.79E-09	7.07E-06	6	Gene Ontology Terms
Multi-Organism Behavior	2.62E-08	1.05E-05	6	Gene Ontology Terms
Cognition	4.24E-07	1.14E-04	8	Gene Ontology Terms
Chemical Synaptic Transmission, Postsynaptic	9.33E-07	1.87E-04	6	Gene Ontology Terms
Dendrite Development	2.54E-05	4.08E-03	6	Gene Ontology Terms
Adult Behavior	3.05E-05	4.09E-03	5	Gene Ontology Terms
Membrane Biogenesis	6.32E-05	7.22E-03	3	Gene Ontology Terms
Synapse Organization	7.18E-05	7.22E-03	7	Gene Ontology Terms
Regulation Of Membrane Potential	8.37E-05	7.48E-03	7	Gene Ontology Terms
Respiratory Gaseous Exchange	2.21E-04	1.71E-02	3	Gene Ontology Terms
Peptidyl-Lysine Modification	2.33E-04	1.71E-02	5	Gene Ontology Terms
Neuron Projection Guidance	3.09E-04	2.07E-02	3	Gene Ontology Terms
Regulation Of Neurological System Process	4.82E-04	2.98E-02	4	Gene Ontology Terms
Covalent Chromatin Modification	8.51E-04	4.89E-02	5	Gene Ontology Terms

https://doi.org/10.1371/journal.pcbi.1011249.t009

potential candidates for involvement in DiGeorge syndrome pathology and schizophrenia. Additionally, the expression of *MVP* may serve as a prognostic marker for several types of cancer. For schizophrenia, *DGCR6* and *PRODH* are well-known candidate genes, and *DGCR5* is a long non-coding RNA gene with a high score for causing schizophrenia.

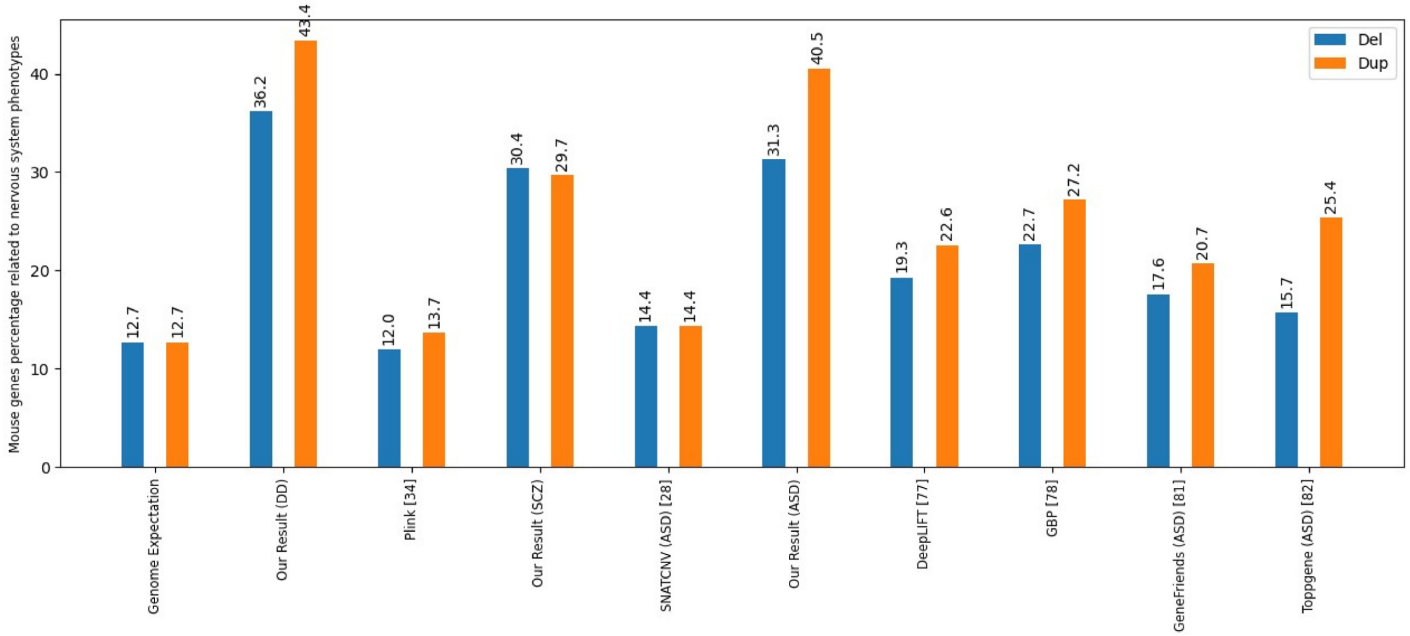

Fig 8. Percentage of Mouse Genes Associated with Nervous System Phenotypes. The figure compares results obtained from various tools and methods, categorized according to types of variation. The analysis focuses on the proportion of mice genes contributing to nervous system phenotypes.

https://doi.org/10.1371/journal.pcbi.1011249.g008

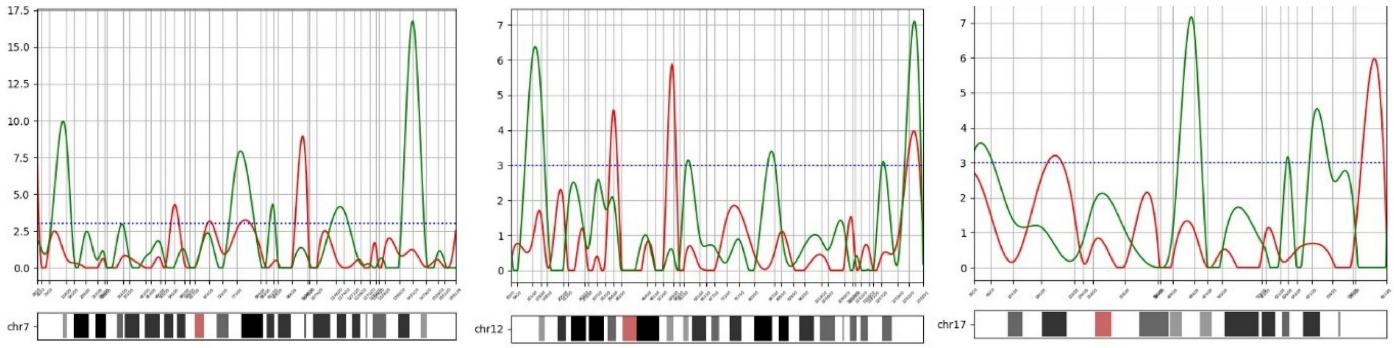

Fig 9. Distribution of -log10(p-value) across Chromosomes with Deletion and Duplication Samples. The figure displays the distribution of -log10(p-value) using Fisher's exact test for deletion (green curves) and duplication (red curves). The horizontal axis represents the chromosome loci, while the vertical axis represents the -log10(p-value).

https://doi.org/10.1371/journal.pcbi.1011249.g009

Furthermore, *SEZ6L2*, *CDIPTOSP*, *ASPHD1*, and *RANBP1* are potential candidate genes for schizophrenia. For autism spectrum disorder, *DGCR2*, *ARVCF*, *GNB1L*, *COMT*, *ZDHHC8*, *CHRNA7*, and *NRXN1* are candidate genes with various associated developmental defects.

Schizophrenia is a complex and debilitating mental disorder associated with genetic factors. One of the well-known candidate genes for schizophrenia is *DGCR6*, which codes for a protein. In addition, mutations in *PRODH* (Proline Dehydrogenase 1), located on 22q11.21, have been linked with susceptibility to schizophrenia (*SCZD4*). Another potential genetic factor for schizophrenia is *DGCR5*, which is a long non-coding RNA (lncRNA) with a high score for causing schizophrenia. [30]

SEZ6L2 (Seizure Related 6 Homolog Like 2), located in 16p11.2, is another gene implicated in mental disorders. This region is thought to hold candidate genes for autism spectrum disorder. *CDIPTOSP* (CDIP Transferase Opposite Strand, Pseudogene) is a lncRNA gene associated with Central Nervous System Germ Cell Tumor disease. *ASPHD1* (Aspartate Beta-Hydroxylase Domain Containing 1) is another gene linked with schizophrenia (specifically, Schizophrenia 3). Lastly, *RANBP1* (RAN Binding Protein 1) is a protein-coding gene linked with Digeorge Syndrome.

Autism spectrum disorder (ASD) is a complex developmental disorder linked to genetic factors. One such factor is the deletion of *DGCR2*, which has been associated with a wide range of developmental defects. These defects are collectively called *CATCH 22*, which stands for DiGeorge syndrome, velocardiofacial syndrome, conotruncal anomaly face syndrome, and isolated conotruncal cardiac defects. Additionally, the *ARVCF* gene is responsible for autosomal dominant Velo-Cardio-Facial syndrome (VCFS), which is characterized by cleft palate, conotruncal heart defects, and facial dysmorphology. *GNB1L* is another gene that is deleted in DiGeorge syndrome. [61,62]

Schizophrenia and panic disorder are two other mental disorders that have been linked to genetic factors. One such factor is the *COMT* (Catechol-O-Methyltransferase) gene, which codes for a protein and has been associated with both schizophrenia and Panic Disorder 1. Another gene linked to schizophrenia is *ZDHHC8* (Zinc Finger DHHC-Type Palmitoyltransferase 8), which is located on chromosome 6q24-q25.

We also investigated the gender distribution of CNVs in BDs. We found that duplication in *NRXN1* and deletion in *PTCHD1* are more frequently observed in males than females for some of the BDs.

We observed that some brain-enriched coding genes were significantly expressed in all three disorders. Examples include *DGCR2*, *SEPTIN5*, and *ARVCF*, which are located on chromosome 22 and have a deletion associated with these disorders. These three genes were among the top ten coding brain-enriched genes related to the three disorders. We also found that *DGCR5*, a noncoding brain-enriched gene previously known as a biomarker for Huntington's disease, is highly associated with DD. *AC000068* is a noncoding brain-enriched gene associated with SCZ and ASD. *SEPTIN5* has been previously shown to be associated with ASD and SCZ, while *DGCR2* is mainly known to be associated with SCZ. *AC004471* is a noncoding brain gene among the top 10 genes related to SCZ, ASD, and DD.

Among the top genes with significant brain expression, some have orthologs in mice that showed nervous system phenotypes. *SEPTIN5*, *ZDHHC8*, *RTN4R*, and *KCTD13* are the top genes for ASD and SCZ, while *RTN4R* and *ZDHHC8* also rank highly in DD. *SEZ6L2* is top in ASD but has a lower rank in SCZ. *ZDHHC8* and *RTN4R* are genes with nervous system morphological and physiological phenotypes, while *SEPTIN5* shows only nervous and physiological phenotypes in mice.

In the next step, we used DECIPHER [35] to examine the relationship between the detected genes and other phenotypes. We found that delayed speech, language, and autism were associated with the set of genes. According to our findings, seizures were associated with SCZ; this relationship was previously discussed in [63].

Microcephaly [64] and macrocephaly [65] are two reverse phenotypes associated with ASD and SCZ. Additionally, 'abnormal facial shape' is associated with all three disorders [66], which has also been studied in [67]. *CACNA1H* is one of the genes related to some overrepresented phenotypes [68], and *TCF20*, discussed in [69], is another gene highlighted in the heatmap of developmental delay.

We performed gene ontology analysis for the detected genes using the WebGestalt tool. This tool allowed us to perform gene ontology analysis, human phenotype ontology (HPO) analysis, and disease ontology analysis separately. For the disease ontology, some terms were "Language Development Disorders," "Autistic behavior," and "Congenital neck anomaly." Overrepresented HPO terms included "Severe global developmental delay," "abnormal social behavior," "Delayed speech and language development," and "Intellectual disability." Some of the most common gene ontology terms were "dendrite development," "cognition," and "Regulation of Neurological System Process." In summary, these findings support the biological relevance of the method-identified genes to genetic factors that contribute to brain disorders.

Although the application of our model focused on three specific brain disorders, it is important to note that our method is not limited to these disorders alone. The versatility of our approach allows for its application in any case-control study involving copy number variants associated with different target disorders. Consequently, the method inherently generates a list of candidate genes specific to the target disorder.

In future research, we plan to explore integrating network analysis techniques and combining mutational data with other auxiliary information, such as proteins or other modalities. This integration will enable the utilization of various modeling tools, such as graphs, to uncover additional patterns within the mutational data.

Materials and methods

Data and preprocessing

In this study, we analyzed three case-control datasets comprising individuals with brain disorders, namely autism spectrum disorder, schizophrenia, and developmental delay. After preprocessing and quality control, the autism spectrum disorder dataset consisted of 47,119 cases and

24,858 control copy number variants (CNVs), as documented in the AUTDB database [41]. The schizophrenia dataset comprised 42,046 cases and 40,414 control CNVs [70]. The developmental delay dataset included 29,803 cases and 11,256 control CNVs. These datasets were selected based on their relevance to the genetic etiology of brain disorders and the availability of reliable and well-curated CNV data.

The final data source for developmental delay comprised two independent datasets with two different data types: NSTD 54 [46] and NSTD 100 [11]. We utilized the NSTD 100 dataset, which includes gender data. All CNVs in this dataset are rare, with a frequency of less than 1% of the population. Further details regarding these CNVs are reported in Table 10.

We used two supplementary data sources in our study. The first is the FANTOM5 (Functional Annotation of the Mammalian Genome 5) Atlas [71], which includes 21,069 protein-coding and 27,920 non-coding genes.

The second data source we used is the Database of Chromosomal Imbalance and Phenotype in Humans Using ENSEMBL Resources (DECIPHER, February 1st, 2017) [35]. This dataset contains information on patients, CNVs, and phenotypes such as ASD, DD, and SCZ. We investigate DECIPHER website to analyze the relationship between genes and other phenotypes and to augment and pretrain our system. Table 11 shows the statistics of the dataset.

In DECIPHER, there were 1,548 patients with ASD-related phenotypes, including 'HP:0000717' (autism), 'HP:0000729' (autistic behavior), and 'HP:0000753' (autism with high cognitive abilities). The dataset also contained 2,144 patients with DD-related phenotypes, including 'HP:0001263' (global developmental delay), 'HP:0011342' (mild global developmental delay), 'HP:0011344' (severe global developmental delay), 'HP:0011343' (moderate global developmental delay), and 'HP:0012758' (neurodevelopmental delay).

This paper also analyzed tissue-enriched genes with a high brain expression level compared to other tissues. We utilized the list of brain-enriched genes provided in [41]. While [42] highlights the impact of brain-enriched genes on autism spectrum disorder, our study focuses on their effect on schizophrenia and developmental delay.

Table 10. Statistics of different datasets. The number of case and control individuals, along with the number of CNVs, were reported in the table.

Dataset	of case CNVs	of control CNVs	Sum	Ratio	of Patients	of Healthy	Sum	Ratio
Autism spectrum disorder	47,119	24,858	71,977	~1.89	19,663	6,479	26,142	~3.03
Schizophrenia	42,046	40,414	82,460	~1.05	28,684*	28,893*	57,577	~0.99
Developmental delay (NSTD 100)	29,803	11,256	41,059	~2.64	29,085	19,584	48,669	~1.52

* 13k Affy and 15k Illumina for cases, 14K Affy, and 14k Illumina for controls.

https://doi.org/10.1371/journal.pcbi.1011249.t010

Table 11. DECIPHER statistics [35]. DECIPHER is a genotype-phenotype data source that can be used to investigate the associations between genes and traits.

Num of Patients	~12,600 Patients	
Num of CNVs	~16,600 CNVs	
Num of Phenotypes	~2,615 Phenotypes	
Num of Autism Patients	~ 1,548 Patients	Related Phenotypes: Autism, Autistic behavior, Autism with high cognitive abilities
Num of Developmental Delay Patients	~ 2,144 Patients	Related Phenotypes: Global developmental delay, Mild global developmental delay, severe global developmental delay, Neurodevelopmental delay, Moderate global developmental delay
Num of Schizophrenia Patients	~ 12 Patients	Related Phenotypes: Schizophrenia, Schizencephaly

https://doi.org/10.1371/journal.pcbi.1011249.t011

Additionally, we used MGI (Mouse Genome Informatics) data [34] to determine if candidate genes related to disease cause a nervous system phenotype in mice, following a similar approach as [41]. HTML was parsed from pages covering nervous system phenotype (MP:0003631) [72], abnormal nervous system morphology (MP:0003632) [73], and abnormal nervous system physiology (MP:0003633) [74]. The mapping was performed using [75]. The data preprocessing involved CNV filtering, conversion, and supplementary data cleansing (DECIPHER data analysis, FANTOM5 data, etc.).

For CNV filtering and conversion, we filtered out CNVs smaller than one kbps (similar to other studies such as [11,41,46]). The CNV studies also had different coordinates (hg17, hg18, and hg19). Therefore, we unified all CNVs and converted them to hg19 using the UCSC Lift Genome Annotations tools [31]. Moreover, we removed Y chromosome CNVs due to insufficient data, eliminating all CNVs with missing values.

We removed patients without phenotypes during supplementary data cleansing while using the DECIPHER data. Preprocessing was unnecessary for Fantom5, MGI, and brain genes since all gene coordinates were already in the hg19 format and ready for processing.

Furthermore, we removed some genes that were not the result of the model, such as genes that overlapped more with controls than cases or genes that did not overlap with CNVs.

A formal overview of a gene prioritization system

In the context of gene prioritization, the process can be conceptualized as a system where the input consists of a target disease and a comprehensive gene list. Depending on the methodology employed for gene processing, various additional datasets may also be incorporated as auxiliary input. These datasets could include protein networks, pathway data, or reliable candidate genes associated with the target disease, thereby leveraging the "guilt by association" principle. The desired output is a list of candidate genes, which can be sorted or unsorted, representing the outcome of prioritization or classification. Furthermore, a scoring system may be implemented to indicate the likelihood of a gene's association with a particular phenotype or disease. The discriminatory algorithm aims to infer the role of each gene in the development of the target disease.

This section aims to provide a formal definition of our work. Consider a case-control study about a specific target disease. This study comprises copy number variants observed in both patient and healthy control groups. The CNVs can be defined as quadruples, characterized as follows:

$$CNV_{set} = \{(ch; dosage; strt; stp) | ch\ in\ \{1..24\}; dosage\ in\ \{del; dup\}; strt < stp\}. \quad (1)$$

where ch is the chromosome number, the dosage is the type of CNV, either deletion or duplication, and strt and stp determine the region of the chromosome where the variation occurs. The CNVs are for people (specified by an identifier) whose features (like gender, other phenotypes, etc.) may or may not be available.

This CNV is available in two sets: one for cases and one for controls.

$$D_{input} = D_{case} \cup D_{control} \quad (2)$$

$$D_{case|control=\{(p_{id};CNV_{set})|p_{id}\ is\ the\ id\ of\ an\ individual;CNV_{set}\ is\ the\ rare\ CNVs\ for\ him\}} \quad (3)$$

Each rare CNV is related to an individual (characterized by p_id), either healthy or patient. Additionally, the dataset may optionally include auxiliary data for each individual, such as gender information. This supplementary information enables us to explore the discriminatory

role of genes for each gender. Our objective is to address the gene prioritization problem utilizing a set of rare copy number variants.

The method overview

Compared to conventional machine learning methods, deep learning approaches offer the advantage of constructing a feature hierarchy and reducing data dimensions. This facilitates the identification of hidden patterns within the data more effectively than alternative approaches. An example of deep learning is the autoencoder, which plays a crucial role in dimensionality reduction and generating a concise, high-level representation of the data through a hierarchical arrangement of features [6]. The autoencoder consists of an encoder network (inference network) that progressively transforms the input into a low-dimensional latent representation and a decoder network (generative network) that strives to reconstruct the output to closely resemble the original input. Autoencoders have been widely employed in various bioinformatics problems [76–78].

Combining autoencoders with the variational learning framework results in the development of Variational Autoencoders VAE [28,79]. VAEs further enhance the capabilities of autoencoders. Fig 10 presents an overview of the VAE, illustrating its schematic representation.

The primary distinction between autoencoders and their variational counterpart lies in their inherent nature. Autoencoders operate deterministically, while variational autoencoders (VAEs) adopt a probabilistic approach. VAEs, in particular, employ regularization techniques to prevent overfitting during training. VAEs are founded upon the Bayesian theorem and inference principles, incorporating a regularization constraint. This framework assumes that the latent representation follows a multivariate Gaussian distribution, denoted as $N(\mu, \sigma)$.

Numerous studies have demonstrated that VAE exhibits enhanced stability during training and produces less ambiguous outputs than other generative models. This improved performance can be attributed to VAE's optimization of precise objective functions rooted in likelihood principles [81]. The posterior distribution in VAE is modeled as a Gaussian distribution, characterized by its mean and variance. It has been theoretically proven that this Gaussian distribution can approximate any function effectively. The primary objective of the VAE model is to encode the input data into a Gaussian distribution, estimating its mean and covariance.

VAE, a deep generative model that utilizes variational inference, is designed to discover a low-dimensional latent representation, denoted as z, for high-dimensional input data X, following the probability distribution P(X). To capture the intrinsic information of the input dataset, P(z|X), the estimation of the posterior distribution becomes necessary, which is typically intractable. By employing variational inference, a distribution family Q(z|X) (referred to as the variational distribution) is introduced to approximate the P(z|X) distribution. The objective is to minimize the Kullback-Leibler (KL) divergence (D) between these two distributions, serving as a dissimilarity measure.

$$D[Q(z|X)||P(z|X)] = E_{z\sim Q}[\log Q(z|X) - \log P(z|X)]. \quad (4)$$

After some calculations, we have the following objective function, which is the variational

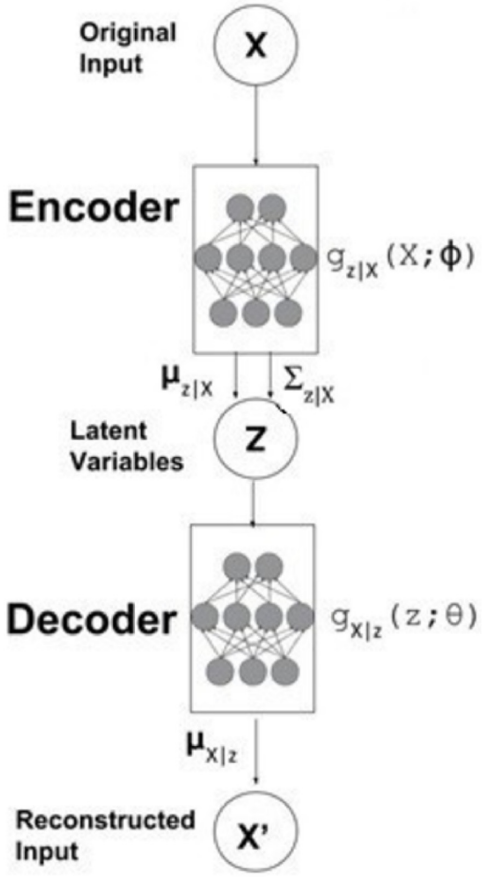

Fig 10. Visualization of a Two-Step Semisupervised Variational Autoencoder (VAE) Process. The figure illustrates the two steps involved in training the VAE. Initially, the VAE is trained in an unsupervised manner. In the second step, one part of the VAE is utilized for training with labels, introducing supervised learning. [80].

https://doi.org/10.1371/journal.pcbi.1011249.g010

lower bound on log-likelihood:

$$\begin{aligned} &\log p_\theta(x) \geq \ell_{VAE} \\ &= \log P(X) - D[Q(z|X)||P(z|X)] \\ &= E_{z\sim Q}[\log P(X|z)] - D(Q(z|X)||P(z)] \end{aligned} \tag{5}$$

The first term is the expectation over the approximate posterior distribution (named as reconstruction error), and the second term (KL distance) is the regularizer (P (z) is standard Gaussian Distribution, N(0, I)). Q(z|X) is the encoding distribution, and P (X|z) is the decoding distribution.

Utilizing these equations transforms the minimization task into a maximization task. The encoder, denoted as Q(z|X), and the decoder, denoted as P(X|z), play crucial roles in this process. This goal can be achieved using deep neural networks coupled with stochastic gradient variational Bayes. In the VAE framework, the encoder component is employed to generate the parameters of the variational distribution. To mitigate overfitting, the dropout technique can be applied. The recognition model Q(z|X) takes the form of a multi-dimensional Gaussian distribution, where the network generates the mean and covariance of this Gaussian distribution.

As for the latent space, a standard Gaussian distribution N(0, I) is employed as the prior distribution.

The loss function in VAE comprises two terms: the reconstruction loss, which facilitates efficient encoding and decoding similar to an autoencoder, and the regularization term, also known as the latent loss, which imposes constraints on the latent space. The regularization term approximates the latent space to follow a standard Gaussian distribution. To incorporate the regularization, the VAE loss function incorporates the Kullback-Leibler divergence, which encourages the covariance matrix to be close to the identity matrix and the mean to be zero.

The training process of the deep learning models consists of two phases: pretraining and fine-tuning. During the pretraining phase, the autoencoder is trained to learn high-level features from all the CNVs associated with the disorders. In the subsequent fine-tuning step, the decoder is set aside, and only the dedicated CNVs specific to the target disease are utilized for training.

The method details

In this section, we explain our method for prioritizing genes. An overview of the method is provided in Fig 11.

A deep learning model is proposed for this task. According to the dataset for each disease, we have a copy number of variants for patients and healthy individuals. Each set of copy

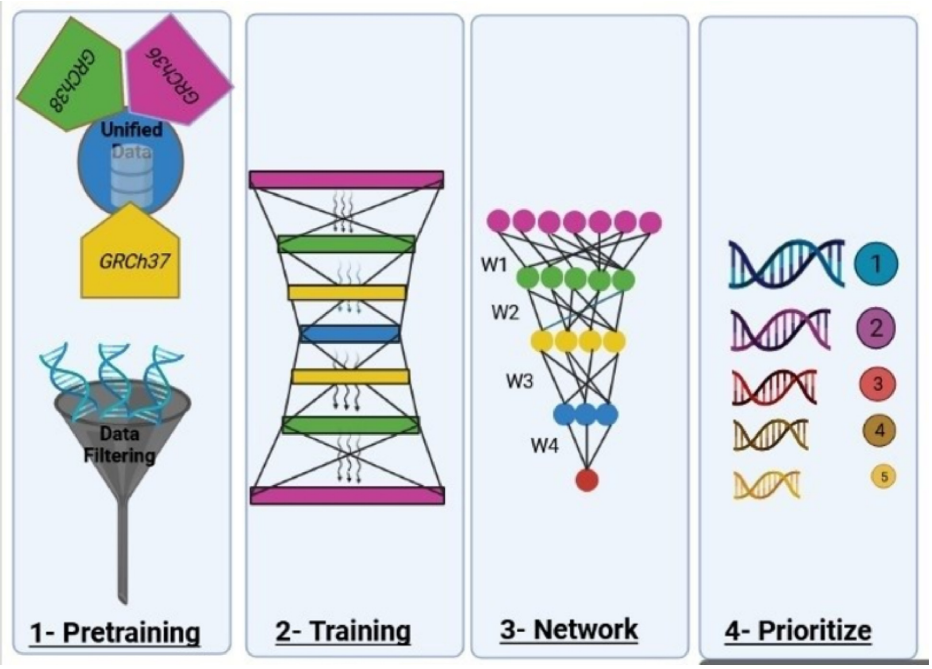

Fig 11. Overview of the Proposed Method. This figure presents a schematic representation of the entire process involved in the proposed method. The process consists of several sequential steps, as illustrated from left to right. The first step involves data preprocessing for data preparation. Initially, the data is obtained in various formats such as hg18, hg19, etc. To establish uniformity, the data is converted into a unified format, specifically hg19. Additionally, this step eliminates redundant, useless, and incomplete features from the data. In the second step, a model is constructed using the cleaned data. This model takes the form of an autoencoder. Subsequently, the weights of the network are adjusted using the corresponding labels. These labels assign values of zero or one to distinguish between healthy and patient individuals. In the fourth step, the autoencoder's coefficients are utilized to prioritize the genes. The importance of each gene is represented by the size of its corresponding icon in the figure. Larger icons correspond to more significant genes, whereas smaller icons indicate less important genes.

https://doi.org/10.1371/journal.pcbi.1011249.g011

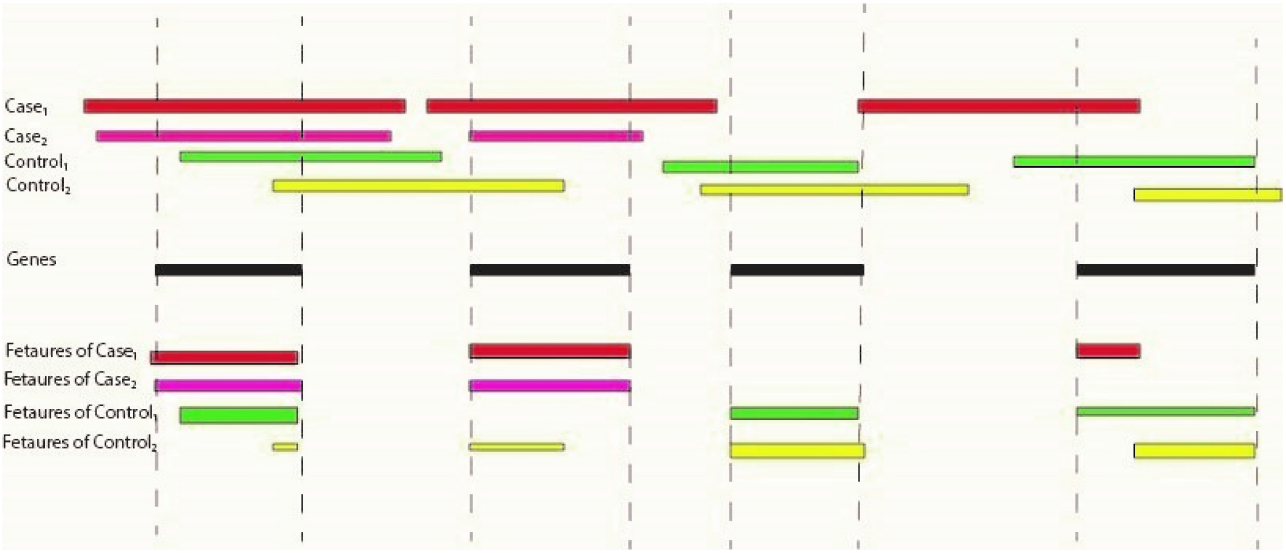

Fig 12. Features Generated from Cases and Controls. This figure presents the features derived from a group of cases and controls. Specifically, the figure depicts two cases and two controls, along with four genes of interest. The copy number variants (CNVs) observed in patients and healthy individuals are visually represented as rectangles in the top section of the figure. Furthermore, the overlaps, which signify the values of features for each case and control, are illustrated in the bottom section. These overlaps provide insight into the shared characteristics between the cases and controls.

https://doi.org/10.1371/journal.pcbi.1011249.g012

number variants for an individual has some overlaps with genes, which are features that feed into our deep learning. This is shown in Fig 12.

We have a list of genes that we want to determine whether their expression will affect disease incidence; besides, we have a list of cases and controls with CNVs for a target disease. We want to convert them to a supervised learning algorithm.

We need to convert CNVs to genes for each healthy and patient individual. Computing overlaps can do this. For the set of genes preprocessed, as discussed before, we measure the length of overlap (in kbps) with the CNVs of an individual. The label of the training set is whether the person is healthy or patient (zero or one).

In the pretraining phase of the model, we used all the CNVs of the brain disorders (autism + schizophrenia + developmental delay). In the next stage, fine-tuning, the CNV of a specific disease is used. Thus, here we have used semi-supervised learning.

After our VAE has been fully trained, we just use the encoder part directly for the next step:

1. Train a VAE using all our data points and transform our data (X) into the latent space (Z variables) (We use all data in this step).
2. Solve a standard supervised learning problem with (Z, Y) pairs (Y is the label set).

The learning algorithm for the whole process is shown in Fig 13. In this algorithm, p is the true posterior, q is the approximate posterior distribution, z is the latent variable, θ is the decoder (z|x) parameters (generative model), and φ is the encoder (x|z) parameters (inference model).

Let's suppose that the encoder weights are represented by $W^m_{i\times j}$, where m is the layer number, i is the output size in the last layer, and j is the input size in the current layer (no connection is determined by zero). As we know, the final layer that will be attached to the encoder is the label; and its size is one (whether the individual is patient (= one) or healthy (= zero)).

Algorithm Learning in model

while generativeTraining() **do**
$\mathcal{D} \leftarrow$ getRandomMiniBatch()
$\mathbf{z}_i \sim q_\phi(\mathbf{z}_i|\mathbf{x}_i) \quad \forall \mathbf{x}_i \in \mathcal{D}$
$\mathcal{J} \leftarrow \sum_n \mathcal{J}(\mathbf{x}_i)$
$(\mathbf{g}_\theta, \mathbf{g}_\phi) \leftarrow (\frac{\partial \mathcal{J}}{\partial \theta}, \frac{\partial \mathcal{J}}{\partial \phi})$
$(\boldsymbol{\theta}, \boldsymbol{\phi}) \leftarrow (\boldsymbol{\theta}, \boldsymbol{\phi}) + \boldsymbol{\Gamma}(\mathbf{g}_\theta, \mathbf{g}_\phi)$
end while
while discriminativeTraining() **do**
$\mathcal{D} \leftarrow$ getLabeledRandomMiniBatch()
$\mathbf{z}_i \sim q_\phi(\mathbf{z}_i|\mathbf{x}_i) \quad \forall \{\mathbf{x}_i, y_i\} \in \mathcal{D}$
trainClassifier($\{\mathbf{z}_i, y_i\}$)
end while

Fig 13. Variational autoencoder(VAE), the learning process algorithm [82].

https://doi.org/10.1371/journal.pcbi.1011249.g013

If we multiply all weights matrices together, the result has the size input size × 1 (the matrices are multiplicable since the output of the last layer equals the input of the next layer). The resulting matrix (precisely column vector) can rank genes according to the label (the label is the status of the disease), and this is the same thing we want to model. The formulation is as follows:

$$W^{final}_{I\times 1} = W^1_{I\times .} \times W^2 \times \ldots \times W^M_{.\times 1} \tag{6}$$

The specification of the deep learning model is such that a binary classification task is accomplished. The size of each layer is the square root of the number of nodes of the last layer. The final layer has a binary outcome, the last activation function is sigmoid, and loss function is binary cross-entropy, and the optimization algorithm is Adam.

Additionally, we investigate the novelty of the top ten genes in three disorders by conducting a comprehensive literature search (searching the 'gene name' + 'disorder name,' the gene will be labeled as known if a meaningful result is obtained. Table 12 shows the results of this investigation.

The detail of the implementation

The deep learning model has a training phase, which needs a training set including cases and controls. We developed the system with Python and PyTorch [83]. We used cross-validation and grid search to tune the parameters (like the number of neurons in each layer).

The activation functions are empirically selected Rectified Linear Units, and the weights were optimized by an adaptive optimization algorithm (Adam) [84] to minimize reconstruction error and loss. The decoder has a symmetrical structure to the encoder. The learning rate, decay rate, and epoch were set to 0.001 and 1, and at most 10,000, respectively. Also, we restrict connections to some extent for a reduction in parameters. The train/test ratio is set to 80/20. The number of layers is at most three.

Table 12. Top ten genes suggested to be candidates for brain disorders and their status in last publications.

	Gene Name	Lastly Discovered
Developmental Delay	*TDRP*	Novel
	DGCR5	Novel
	PRODH	Novel
	LCE3E	Novel
	ERICH1	Known
	CATG0101427	Novel
	ERICH1-AS1	Novel
	RP11-462G22	Novel
	CATG0074892	Novel
	CATG0074890	Novel
Schizophrenia	*CATG0074891*	Novel
	DGCR6	Known
	PRODH	Known
	DGCR5	Known
	AC009133	Novel
	MVP	Novel
	CDIPT	Known
	SEZ6L2	Known
	CATG0027072	Novel
	CDIPT-AS1	Novel
Autism Spectrum Disorder	*DGCR2*	Known
	ARVCF	Known
	GNB1L	Known
	CATG00000058206	Novel
	COMT	Known
	ZDHHC8	Known
	HIRA	Novel
	TBX1	Known
	CDIPT	Known
	SEZ6L2	Known

https://doi.org/10.1371/journal.pcbi.1011249.t012

Supporting information

S1 Table. Details of the results for Autism Spectrum Disorder.
(XLSX)

S2 Table. Details of the results for Schizophrenia.
(XLSX)

S3 Table. Details of the results for Developmental Delay.
(XLSX)

S1 Fig. The common genes between disorders, 'del' is short for deletion.
(EPS)

S2 Fig. Distribution of CNV length in different chromosomes for SCZ disease; y – Axis is the $\times 10^5$. The numbers on top of the plot show the number of cases and controls. The red color (left) represents cases, and the blue (right) represents controls.
(EPS)

S3 Fig. Distribution of CNV length in different chromosomes for ASD disease. Y-Axis is the $\times 10^6$. The numbers on top of the plot show the number of cases and controls. The red color (left) represents cases, and the blue (right) represents controls.
(EPS)

S4 Fig. Distribution of CNV length in different chromosomes for DD disease. Y-Axis is the $\times 10^6$. The numbers on top of the plot show the number of cases and controls. The red color (left) represents cases, and the blue (right) represents controls.
(EPS)

S5 Fig. Demographic Distribution of DD and ASD datasets.
(EPS)

S6 Fig. Decipher Phenotypes Frequency.
(EPS)

S7 Fig. Details of the setup of the method. Since the technique is semisupervised, the first step is to use the data without labels to pretrain the network. The next step is to use the target data to fine-tune it. Next, we use the coefficients of the network to derive a score for each of the features of the input network, i.e., genes. The values of the scores are then sorted so that the relative usefulness of the genes can be evaluated.
(EPS)

Author Contributions

Conceptualization: Zahra Rahaie, Hamid R. Rabiee, Hamid Alinejad-Rokny.

Data curation: Zahra Rahaie.

Formal analysis: Zahra Rahaie, Hamid R. Rabiee, Hamid Alinejad-Rokny.

Investigation: Zahra Rahaie.

Methodology: Zahra Rahaie, Hamid R. Rabiee, Hamid Alinejad-Rokny.

Project administration: Hamid R. Rabiee, Hamid Alinejad-Rokny.

Software: Zahra Rahaie.

Supervision: Hamid R. Rabiee, Hamid Alinejad-Rokny.

Validation: Zahra Rahaie, Hamid R. Rabiee, Hamid Alinejad-Rokny.

Visualization: Zahra Rahaie.

Writing – original draft: Zahra Rahaie.

Writing – review & editing: Hamid R. Rabiee, Hamid Alinejad-Rokny.

References

1. Raj MR, Sreeja A. Analysis of computational gene prioritization approaches. Procedia computer science. 2018 Jan 1; 143:395–410. https://doi.org/10.1016/j.procs.2018.10.411
2. Lan W, Wang J, Li M, Peng W, Wu F. Computational approaches for prioritizing candidate disease genes based on PPI networks. Tsinghua Science and Technology. 2015 Oct 13; 20(5):500–12. https://doi.org/10.1109/TST.2015.7297749
3. Kumar AA, Van Laer L, Alaerts M, Ardeshirdavani A, Moreau Y, Laukens K, et al. pBRIT: gene prioritization by correlating functional and phenotypic annotations through integrative data fusion. Bioinformatics. 2018 Jul 1; 34(13):2254–62. https://doi.org/10.1093/bioinformatics/bty079 PMID: 29452392

4. Nitsch D, Gonçalves JP, Ojeda F, De Moor B, Moreau Y. Candidate gene prioritization by network analysis of differential expression using machine learning approaches. BMC bioinformatics. 2010 Dec; 11 (1):1–6. https://doi.org/10.1186/1471-2105-11-460 PMID: 20840752
5. Glaab E, Bacardit J, Garibaldi JM, Krasnogor N. Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data. PloS one. 2012 Jul 11; 7(7):e39932. https://doi.org/10.1371/journal.pone.0039932 PMID: 22808075
6. Baldi P. Autoencoders, unsupervised learning, and deep architectures. InProceedings of ICML workshop on unsupervised and transfer learning 2012 Jun 27 (pp. 37–49). JMLR Workshop and Conference Proceedings.
7. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS. SUSPECTS: enabling fast and effective prioritization of positional candidates. Bioinformatics. 2006 Mar 15; 22(6):773–4. https://doi.org/10.1093/bioinformatics/btk031 PMID: 16423925
8. Hutz JE, Kraja AT, McLeod HL, Province MA. CANDID: a flexible method for prioritizing candidate genes for complex human traits. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society. 2008 Dec; 32(8):779–90. https://doi.org/10.1002/gepi.20346 PMID: 18613097
9. Cheng MC, Chien WH, Huang YS, Fang TH, Chen CH. Translational Study of Copy Number Variations in Schizophrenia. International Journal of Molecular Sciences. 2021 Dec 31; 23(1):457. https://doi.org/10.3390/ijms23010457 PMID: 35008879
10. Malhotra D, Sebat J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. Cell. 2012 Mar 16; 148(6):1223–41. https://doi.org/10.1016/j.cell.2012.02.039 PMID: 22424231
11. Coe Bradley P., et al. "Refining analyses of copy number variation identifies specific genes associated with developmental delay." Nature genetics 46.10 (2014): 1063–1071. https://doi.org/10.1038/ng.3092 PMID: 25217958
12. Chapter Bromberg Y. 15: disease gene prioritization. PLoS computational biology. 2013 Apr 25; 9(4): e1002902. https://doi.org/10.1371/journal.pcbi.1002902 PMID: 23633938
13. Tranchevent LC, Ardeshirdavani A, ElShal S, Alcaide D, Aerts J, Auboeuf D, et al. Candidate gene prioritization with Endeavour. Nucleic acids research. 2016 Jul 8; 44(W1):W117–21. https://doi.org/10.1093/nar/gkw365 PMID: 27131783
14. Stäubert C, Tarnow P, Brumm H, Pitra C, Gudermann T, Gruters A, et al. Evolutionary aspects in evaluating mutations in the melanocortin 4 receptor. Endocrinology. 2007 Oct 1; 148(10):4642–8. https://doi.org/10.1210/en.2007-0138 PMID: 17628007
15. Jiang BB, Wang JG, Wang Y, Xiao J. Gene prioritization for type 2 diabetes in tissue-specific protein interaction networks. Systems Biology. 2009 Sep; 10801131:319–28.
16. Mefford HC, Muhle H, Ostertag P, von Spiczak S, Buysse K, Baker C, et al. Genome-wide copy number variation in epilepsy: novel susceptibility loci in idiopathic generalized and focal epilepsies. PLoS genetics. 2010 May 20; 6(5):e1000962. https://doi.org/10.1371/journal.pgen.1000962 PMID: 20502679
17. Altman RB, Bergman CM, Blake J, Blaschke C, Cohen A, Gannon F, et al. Text mining for biology-the way forward: opinions from leading scientists. Genome biology. 2008 Sep; 9(2):1–5. https://doi.org/10.1186/gb-2008-9-s2-s7 PMID: 18834498
18. Zolotareva O, Kleine M. A survey of gene prioritization tools for Mendelian and complex human diseases. Journal of integrative bioinformatics. 2019 Dec 1; 16(4). https://doi.org/10.1515/jib-2018-0069 PMID: 31494632
19. Moreau Y, Tranchevent LC. Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nature Reviews Genetics. 2012 Aug; 13(8):523–36. https://doi.org/10.1038/nrg3253 PMID: 22751426
20. Börnigen D, Tranchevent LC, Bonachela-Capdevila F, Devriendt K, De Moor B, De Causmaecker P, et al. An unbiased evaluation of gene prioritization tools. Bioinformatics. 2012 Dec 1; 28(23):3081–8. https://doi.org/10.1093/bioinformatics/bts581 PMID: 23047555
21. Seyyedrazzagi E, Navimipour NJ. Disease genes prioritizing mechanisms: a comprehensive and systematic literature review. Network Modeling Analysis in Health Informatics and Bioinformatics. 2017 Dec; 6(1):1–5. https://doi.org/10.1007/s13721-017-0154-9
22. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Annals of internal medicine. 1999 Jun 15; 130(12):995–1004. https://doi.org/10.7326/0003-4819-130-12-199906150-00008 PMID: 10383371
23. Gillis Jesse, and Pavlidis Paul. "Guilt by association" is the exception rather than the rule in gene networks." PLoS computational biology 8.3 (2012): e1002444.
24. Gunning Margot, and Pavlidis Paul. "Guilt by association" is not competitive with genetic association for identifying autism risk genes." *Scientific Reports* 11.1 (2021): 15950.

25. Fisher Aaron, Rudin Cynthia, and Dominici Francesca. "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously." *J. Mach. Learn. Res.* 20.177 (2019): 1–81. PMID: 34335110
26. Boudellioua I, Kulmanov M, Schofield PN, Gkoutos GV, Hoehndorf R. DeepPVP: phenotype-based prioritization of causative variants using deep learning. BMC bioinformatics. 2019 Dec; 20(1):1–8. https://doi.org/10.1101/311621
27. Zakeri P, Simm J, Arany A, ElShal S, Moreau Y. Gene prioritization using Bayesian matrix factorization with genomic and phenotypic side information. Bioinformatics. 2018 Jul 1; 34(13):i447–56. https://doi.org/10.1093/bioinformatics/bty289 PMID: 29949967
28. Kingma DP, Welling M. Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114. 2013 Dec 20.
29. Kingma D, Welling M. Efficient gradient-based inference through transformations between Bayes nets and neural nets. In International Conference on Machine Learning 2014 Jun 18 (pp. 1782–1790). PMLR.
30. Molinard-Chenu A, Dayer A. The candidate schizophrenia risk gene DGCR2 regulates early steps of corticogenesis. Biological Psychiatry. 2018 Apr 15; 83(8):692–706. https://doi.org/10.1016/j.biopsych.2017.11.015 PMID: 29305086
31. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome research. 2002 Jun 1; 12(6):996–1006. https://doi.org/10.1101/gr.229102 PMID: 12045153
32. The Remap Tool. https://www.ncbi.nlm.nih.gov/genome/tools/remap.
33. Cardoso AR, Lopes-Marques M, Silva RM, Serrano C, Amorim A, et al. Essential genetic findings in neurodevelopmental disorders. Human genomics. 2019 Dec; 13(1):1–7.
34. Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA, Mouse Genome Database Group. The Mouse Genome Database (MGD): mouse biology and model systems. Nucleic acids research. 2008 Jan; 36 (suppl_1): D724–8. https://doi.org/10.1093/nar/gkm961 PMID: 18158299
35. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, et al. DECIPHER: database of chromosomal imbalance and phenotype in humans using ensemble resources. The American Journal of Human Genetics. 2009 Apr 10; 84(4):524–33. https://doi.org/10.1016/j.ajhg.2009.03.010 PMID: 19344873
36. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic acids research. 2019 Jul 2; 47(W1): W199–205. https://doi.org/10.1093/nar/gkz401 PMID: 31114916
37. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806. 2014 Dec 21.
38. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. International conference on machine learning 2017 Jul 17 (pp. 3145–3153). PMLR.
39. Ribeiro MT, Singh S, Guestrin C. " Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 1135–1144).
40. Kokhlikyan N, Miglani V, Martin M, Wang E, Alsallakh B, Reynolds J, et al. Captum: A unified and generic model interpretability library for PyTorch. arXiv preprint arXiv:2009.07896. 2020 Sep 16.
41. Alinejad-Rokny H, Heng JI, Forrest AR. Brain-enriched coding and long non-coding RNA genes are overrepresented in recurrent neurodevelopmental disorder CNVs. Cell Reports. 2020 Oct 27; 33 (4):108307. https://doi.org/10.1016/j.celrep.2020.108307 PMID: 33113368
42. Pinto D, Delaby E, Merico D, Barbosa M, Merikangas A, Klei L, et al. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. The American Journal of Human Genetics. 2014 May 1; 94(5):677–94. https://doi.org/10.1016/j.ajhg.2014.03.018 PMID: 24768552
43. Hon CC, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJ, Gough J, et al. An atlas of human long non-coding RNAs with accurate 5′ ends. Nature. 2017 Mar; 543(7644):199–204. https://doi.org/10.1038/nature21374 PMID: 28241135
44. Raina Priyanka, Guinea Rodrigo, Chatsirisupachai Kasit, Lopes Inês, Farooq Zoya, Guinea Cristina, et al, GeneFriends: gene co-expression databases and tools for humans and model organisms, Nucleic Acids Research, 2022; gkac1031, https://doi.org/10.1093/nar/gkac1031 PMID: 36454018
45. Chen J, Bardes EE, Aronow BJ, Jegga AG 2009. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Research. https://doi.org/10.1093/nar/gkp427 PMID: 19465376
46. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, et al. A copy number variation morbidity map of developmental delay. Nature genetics. 2011 Sep; 43(9):838–46. https://doi.org/10.1038/ng.909 PMID: 21841781

47. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics. 2007 Sep 1; 81(3):559–75. https://doi.org/10.1086/519795 PMID: 17701901

48. May T, Adesina I, McGillivray J, Rinehart NJ. Sex differences in neurodevelopmental disorders. Current opinion in neurology. 2019 Aug 1; 32(4):622–6. https://doi.org/10.1097/WCO.0000000000000714 PMID: 31135460

49. Rinehart NJ, Cornish KM, Tonge BJ. Gender differences in neurodevelopmental disorders: Autism and fragile x syndrome. Biological basis of sex differences in psychopharmacology. 2010:209–29. https://doi.org/10.1007/7854_2010_96 PMID: 21769728

50. Brentani H. Gender, Genetic, And Environmental Factors In The Neurodevelopmental Disorders. European Neuropsychopharmacology. 2019 Jan 1; 29:S745–6. https://doi.org/10.1016/j.euroneuro.2017.06.083

51. Al-Beltagi M. Autism medical comorbidities. World journal of clinical pediatrics. 2021 May 9; 10(3):15. https://doi.org/10.5409/wjcp.v10.i3.15 PMID: 33972922

52. Buckley PF, Miller BJ, Lehrer DS, Castle DJ. Psychiatric comorbidities and schizophrenia. Schizophrenia bulletin. 2009 Mar 1; 35(2):383–402. https://doi.org/10.1093/schbul/sbn135 PMID: 19011234

53. Xie B, Fan X, Lei Y, Chen R, Wang J, Fu C, et al. A novel de novo microdeletion at 17q11. 2 adjacent to NF1 gene associated with developmental delay, short stature, microcephaly and dysmorphic features. Molecular cytogenetics. 2016 Dec; 9(1):1–5. https://doi.org/10.1186/s13039-016-0251-y PMID: 27247625

54. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nature genetics. 2000 May; 25(1):25–9. https://doi.org/10.1038/75556 PMID: 10802651

55. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic acids research. 2016 Oct 19:gkw943. https://doi.org/10.1093/nar/gkw943 PMID: 27924018

56. Schilbach L. Autism and other disorders of social interaction: where we are and where to go from here. European Archives of Psychiatry and Clinical Neuroscience. 2022 Feb 9:1–3. https://doi.org/10.1007/s00406-022-01391-y PMID: 35141808

57. Hisaoka T, Komori T, Kitamura T, Morikawa Y. Abnormal behaviors relevant to neurodevelopmental disorders in Kirrel3-knockout mice. Scientific reports. 2018 Jan 23; 8(1):1–2. https://doi.org/10.1038/s41598-018-19844-7 PMID: 29362445

58. Martínez-Cerdeño V. Dendrite and spine modifications in autism and related neurodevelopmental disorders in patients and animal models. Developmental neurobiology. 2017 Apr; 77(4):393–404. https://doi.org/10.1002/dneu.22417 PMID: 27390186

59. Zieger HL, Choquet D. Nanoscale synapse organization and dysfunction in neurodevelopmental disorders. Neurobiology of Disease. 2021 Oct 1; 158:105453. https://doi.org/10.1016/j.nbd.2021.105453 PMID: 34314857

60. Fallah MS, Eubanks JH. Seizures in mouse models of rare neurodevelopmental disorders. Neuroscience. 2020 Oct 1; 445:50–68. https://doi.org/10.1016/j.neuroscience.2020.01.041 PMID: 32059984

61. Xue Chen, et al. Progress and assessment of lncRNA DGCR5 in malignant phenotype and immune infiltration of human cancers. American Journal of Cancer Research 11.1 (2021): 1. PMID: 33520356

62. Suzuki G, Harper KM, Hiramoto T, Sawamura T, Lee M, Kang G, Tanigaki K, et al. Sept5 deficiency exerts pleiotropic influence on affective behaviors and cognitive functions in mice. Human molecular genetics. 2009 May 1; 18(9):1652–60. https://doi.org/10.1093/hmg/ddp086 PMID: 19240081

63. Hyde TM, Weinberger DR. Seizures and schizophrenia. Schizophrenia bulletin. 1997 Jan 1; 23(4):611–22. https://doi.org/10.1093/schbul/23.4.611 PMID: 9365998

64. Kunugi H, Takei N, Murray RM, Saito K, Nanko S. Small head circumference at birth in schizophrenia. Schizophrenia research. 1996 May 1; 20(1–2):165–70. https://doi.org/10.1016/0920-9964(96)00007-2 PMID: 8794505

65. Klein S, Sharifi-Hannauer P, Martinez-Agosto JA. Macrocephaly as a clinical indicator of genetic subtypes in autism. Autism Research. 2013 Feb; 6(1):51–6. https://doi.org/10.1002/aur.1266 PMID: 23361946

66. Tripi G, Roux S, Matranga D, Maniscalco L, Glorioso P, Bonnet-Brilhault F, et al. Cranio-facial characteristics in children with autism spectrum disorders (ASD). Journal of Clinical Medicine. 2019 May 9; 8(5):641. https://doi.org/10.3390/jcm8050641 PMID: 31075935

67. Hosseini MP, Beary M, Hadsell A, Messersmith R, Soltanian-Zadeh H. Deep Learning for Autism Diagnosis and Facial Analysis in Children. Frontiers in Computational Neuroscience. 2021; 15. https://doi.org/10.3389/fncom.2021.789998 PMID: 35126078

68. Chourasia N, Ossó-Rivera H, Ghosh A, Von Allmen G, Koenig MK. Expanding the phenotypic spectrum of CACNA1H mutations. Pediatric Neurology. 2019 Apr 1; 93:50–5. https://doi.org/10.1016/j.pediatrneurol.2018.11.017 PMID: 30686625

69. Torti E, Keren B, Palmer EE, Zhu Z, Afenjar A, Anderson I, et al. Variants in TCF20 in neurodevelopmental disability: description of 27 new patients and review of literature. Genetics in Medicine. 2019 Sep; 21(9):2036–42. https://doi.org/10.1038/s41436-019-0454-9 PMID: 30739909

70. Marshall CR, Howrigan DP, Merico D, Thiruvahindrapuram B, Wu W, Greer DS, et al. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. Nature genetics. 2017 Jan; 49(1):27–35. https://doi.org/10.1038/ng.3725 PMID: 27869829

71. DGT RP, FANTOM Consortium. A promoter-level mammalian expression atlas. Nature. 2014 Mar 27; 507(7493):462–70. https://doi.org/10.1038/nature13182 PMID: 24670764

72. Mouse 0003631. MGI 6.22, Mammalian Phenotype Ontology Annotations, Last Updated 07/04/2023, http://www.informatics.jax.org/mp/annotations/MP:0003631.

73. Mouse 0003632. MGI 6.22, Mammalian Phenotype Ontology Annotations, Last Updated 07/04/2023, http://www.informatics.jax.org/mp/annotations/MP:0003632.

74. Mouse 0003633. MGI 6.22, Mammalian Phenotype Ontology Annotations, Last Updated 07/04/2023, http://www.informatics.jax.org/mp/annotations/MP:0003633.

75. Mouse Homologene. MGI 6.22, Mammalian Phenotype Ontology Annotations, Last Updated 07/10/2023, http://www.informatics.jax.org/downloads/reports/HGNC_AllianceHomology.rpt.

76. Chicco D, Sadowski P, Baldi P. Deep autoencoder neural networks for gene ontology annotation predictions. In Proceedings of the 5th ACM conference on bioinformatics, computational biology, and health informatics 2014 Sep 20 (pp. 533–540).

77. Chen L, Cai C, Chen V, Lu X. Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model. InBMC bioinformatics 2016 Dec (Vol. 17, No. 1, pp. 97–107). BioMed Central. https://doi.org/10.1186/s12859-015-0852-1 PMID: 26818848

78. Svensson V, Gayoso A, Yosef N, Pachter L. Interpretable factor models of single-cell RNA-seq via variational autoencoders. Bioinformatics. 2020 Jun 1; 36(11):3418–21. https://doi.org/10.1093/bioinformatics/btaa169 PMID: 32176273

79. Doersch C. Tutorial on variational autoencoders. arXiv preprint arXiv:1606.05908. 2016 Jun 19.

80. Keng B. Semi-supervised learning with variational autoencoders. Self-published via Github. io. 2017 Sep.

81. Genevay A, Peyré G, Cuturi M. GAN and VAE from an optimal transport point of view. arXiv preprint arXiv:1706.01807. 2017 Jun 6.

82. Kingma DP, Mohamed S, Jimenez Rezende D, Welling M. Semi-supervised learning with deep generative models. Advances in neural information processing systems. 2014; 27.

83. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Advances in Neural Information Processing Systems 32. Curran Associates, Inc.; 2019. p. 8024–35.

84. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014 Dec 22.