medRxiv preprint doi: https://doi.org/10.1101/2023.01.30.23285175; this version posted February 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

An Expert-guided Hierarchical Graph Attention Network for Post-traumatic Stress Disorder Highly-associative Genetic Biomarkers Identification

Qi Zhang[1, #], Yang Han[1, #], Jacqueline CK Lam[1, *, #], Ruiqiao Bai[1], Illana Gozes[2], Victor OK Li[1,*, #]

[1]Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam, Hong Kong

[2]Department of Human Molecular Genetics and Biochemistry, Sackler Faculty of Medicine, Adams Super Center for Brain Studies and Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel

[#]Equal contributions

[*]Corresponding and Senior authors (jcklam@eee.hku.hk, vli@eee.hku.hk)

Abstract

Post-traumatic Stress Disorder (PTSD) is a common debilitating mental disorder, that occurs in some individuals following extremely traumatic events. Traditional identification of Genetic Markers (GM) for PTSD is mainly based on a statistical clinical approach by comparing PTSD patients with normal controls. However, these statistical studies present limitations, often generating inconsistent results. Few studies have yet examined thoroughly the role of somatic mutations, PTSD disease pathways and their relationships. Capitalizing on deep learning techniques, we have developed a novel hierarchical graph attention network to identify highly correlational GM (HGMs) of PTSD. The network presents the following novelties: First, both a hierarchical graph structure and a graph attention mechanism have been integrated into a model to develop a graph attention network (GAtN) model. Second, domain-specific knowledge, including somatic mutations, genes, PTSD pathways and their correlations have been incorporated into the graph structures. Third, 12 somatic mutations having high or moderate impacts on proteins or genes have been identified as the potential HGMs for PTSD. Fourth, our study is carefully guided by prominent PTSD literature or clinical experts of the field; any high saliency HGMs generated from our model are further verified by existing PTSD-related authoritative medical journals. Our study illustrates the utility and significance of a hybrid approach, integrating both AI and expert-guided/domain-specific knowledge for thorough identification of biomarkers of PTSD, while building on the nature of convergence and divergence of PTSD pathways. Our expert-guided AI-driven methodology can be extended to other pathological-based HGM identification studies; it will transform the methodology of biomarker identification for different life-threatening diseases to speed up the complex lengthy procedures of new biomarkers identification.

Introduction

Post-traumatic Stress Disorder (PTSD) is a common and debilitating mental disorder that occurs in some individuals following exposure to extremely traumatic events, such as life-threatening accidents or natural disasters [1,2]. It leads to symptoms such as re-experiencing (e.g. having trauma-related memories that intrude into what is currently happening), avoidance of stimuli associated with the trauma, negative changes in cognition and mood, and hyperarousal [2-7]. These symptoms could cause serious and long-lasting problems, including unemployment, marital instability, physical illness, and early mortalities [3,8-14]. Family, twin and

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.01.30.23285175; this version posted February 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

molecular genetic studies have suggested that genetic factors contribute to the development of PTSD [13-23]. However, despite more than a decade of research efforts, robust Highly-associative Genetic Markers (HGMs) of PTSD remain largely unknown [4].

Traditional methods identifying genetic markers related to PTSD are mainly based on statistical analysis and comparisons among PTSD patients and normal controls [2,4,11,14,17,20-22,24-50]. Two major approaches are candidate gene association studies and Genome-wide Association Studies (GWAS) [2,4,11,14,17,20-22,24-54].

In candidate gene association studies, only a few selected genetic markers are involved in the analysis, and the selection is mainly based on existing biological knowledge obtained from prior research on PTSD-related neurobiological processes [17,20,24-26]. Much research efforts have been made in candidate gene association studies [27-43]. For instance, the FKBP5 gene, which is an important regulator of the stress system, has been suggested to have single-nucleotide polymorphisms associated with PTSD through interactions with child abuse [3,27,31]. However, the majority of PTSD HGMs could hardly be identified by candidate gene association studies, since such studies are usually limited to genetic regions where there are prior hypotheses about their roles in the development or maintenance of PTSD, whereas people's prior understanding on the pathophysiology of PTSD is incomprehensive or even incorrect [14,17,21].

In GWAS, the frequencies of hundreds to millions of genetic variants across the entire genome are compared simultaneously between those with and without PTSD [17,20,21,24,25,44,45]. It is a hypothesis-free approach avoiding the need for prior knowledge of specific candidate genes/variants, and thus capable of identifying unknown mechanisms [20,24,45]. It has gained momentum in recent years for PTSD-related genetic marker identification [2,4,11,14,22,24,46-54]. For example, a GWAS conducted on veterans and their intimate partners reported a genome-wide significant association between PTSD and a single-nucleotide polymorphism (rs8042149) located in the RORA gene [46]. However, given the large number of genetic markers simultaneously involved in the statistical analysis, large sample sizes are required for GWAS to have statistical power [44,55]. Results of GWAS are also best combined with known or putative PTSD-related biological knowledge, in order to avoid spurious findings caused by sampling bias [20].

There are other problems with traditional genetic studies of PTSD. A major issue is the inconsistency among research results when different samples are investigated [4,21,25,26,41,43,56-59]. Besides, previous PTSD-related studies have mainly focused on germline mutations, with the underlying hypothesis that PTSD-related genetic factors are heritable, while less attention has been paid to somatic mutations [2,4,11,14,22,27,30-33,37-40,42,43,47-54,56,59]. In 2021, Sragovich et al. have extracted somatic mutations from RNA-seq data, and utilized STRING analysis, which is a method based on protein interaction information, to identify crucial PTSD-related genes with somatic mutations[60]. Eight genes have been identified in the study, including TSC1, FMR1, GSK3B, EZR, TNF, IL1R2, CASP1 and CASP4 [60]. However, studies in this field are still in the beginning stage.

In recent years, Artificial Intelligent (AI) techniques have been utilized in finding genetic markers associated with diseases, though not on PTSD. For instance, to classify whether a gene is associated with Parkinson's Disease (PD), a neural network-based ensemble (n-semble) method based on protein features has been put forward, reaching 88.9%, 90.9% and 89.8% for the precision, recall and F score in a five-fold validation, respectively [61]. Another PD-related gene prediction model named N2A-SVM has also been proposed based on protein interaction information and techniques including Node2vec, the autoencoder and the support vector machine [62]. Its area under the receiver operating characteristic (ROC) curve reaches 0.7289 for classifying whether a gene is associated with PD in a ten-fold validation [62]. A similar methodology has also been applied to the disease Multiple Sclerosis (MS), and achieved 70.11% accuracy for classifying whether a gene is associated with MS in a five-fold validation [63]. Besides, Chang et al. [64] has proposed a deep learning method based on a sparse auto-encoder to identify cancer-related genes, by extracting features from protein expression profiles and protein interaction information, and achieves ROC value over 0.8 in predicting cancer-related genes. There are also studies applying AI methods for identifying Alzheimer's Disease (AD)-related genes, using techniques including autoencoders, stepwise artificial neural networks, convolutional neural networks and conditional generative adversarial networks [65-68]. Given the lack of ground truth, most of those papers have not reported their exact accuracy for identifying AD-

medRxiv preprint doi: https://doi.org/10.1101/2023.01.30.23285175; this version posted February 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

related genes, while some of them have claimed that previous studies/analyses (related to AD/neurodegenerative diseases or gene functions) could support some genes they identified [65-68]. Besides, in 2021, Li et al. have published a plan on designing an AI-driven causal graph model to identify the HGMs for AD in the future [69]. Moreover, utilizing data on 83 diseases, a feed-forward neural network has been designed for disease diagnosis based on information of gene expression and disease pathways, and sensitivity analysis has been performed to identify associations between diseases and genes [70]. There are literature supports for 70% of the top 10 disease-gene associations identified in the study [70]. Thus, AI techniques seem promising in identifying genetic markers for diseases.

In this paper, we have developed a graph-based deep learning diagnosis model to identify probable HGMs for PTSD, utilizing hierarchical graph structures and graph attention mechanisms [71,72]. Compared to previous studies, our novelties are listed as follows:

- We have constructed a hierarchical biomedical graph representing different layers of one's biological system, ranging from somatic mutations, genes, to pathways, while incorporating a variety of domain-specific knowledge during the graph construction process, including the impact level of somatic mutations to proteins (i.e. CADD scores), lengths of genes, and the number of PTSD-related pathways on which each gene locates.
- We have utilized graph attention mechanisms to calculate the weights of complex gene-gene interconnections to complement the domain-specific pathway-based information. We have also used attention mechanisms to capture crucial long genes and genes located in more PTSD-related pathways.
- We have proposed a novel saliency score that calculates PTSD-related risk for somatic mutations utilizing the novel hierarchical graph attention network (H-GAtN). A list of HGMs have been identified based on the saliency score and on the literature.

Results

Experimental setting and evaluation

The model was implemented in PyTorch [73]. The training epoch was 30 and the optimizer was Adam (initial learning rate = 10^{-6}, reduced by one-tenth after every 10 epochs). L2 regularization has been applied (weight decay = 10^{-7}). The hidden dimensions of the embedding layer, the mutation graph, the gene graph, and the final fully connected layer were 8, 16, 32, and 16, respectively. 80% of the patients and 80% of the normal control subjects were randomly selected as the training set, and the rest were used as the testing set. The percentage of misclassified subjects (the subject would be classified as PTSD patients when the predicted PTSD probability is larger than 0.5, otherwise classified as normal controls) in the testing set was used to evaluate the model's classification accuracy, and the final error rate could be lowered to 11.8%. The area under the ROC curve was 0.90.

Top HGMs identification

There were 13566 high/moderate-impact somatic mutations detected in the subjects, which were taken as input features of the model. The studied mutations have high or moderate impacts on the proteins of 4561 genes. After model training, saliency scores of all mutations were calculated. Table 1 shows detailed information about the mutations with the top 20 saliency scores and the gene affected by each mutation. The type of mutations includes insertion-deletion (INDEL), and single-nucleotide variant (SNV).

Ranking (based on Saliency Score)	Chromosome	Position	Type	Reference allele	Alternative allele	Affected gene
1	8	30180593	INDEL	CAG	C	DCTN6
2	18	51176872	SNV	C	T	MEX3C
3	3	111606778	SNV	G	A	CD96
4	11	118349865	INDEL	G	GA	CD3G
5	6	41198325	SNV	C	A	TREML2

medRxiv preprint doi: https://doi.org/10.1101/2023.01.30.23285175; this version posted February 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

6	14	102050157	SNV	C	T	DYNC1H1
7	6	33008108	SNV	C	T	HLA-DOA
8	7	100373707	INDEL	AT	A	PILRA
9	9	137039846	SNV	G	A	NPDC1
10	1	158292315	SNV	G	A	CD1C
11	12	57741942	SNV	T	C	AGAP2
12	19	51415990	SNV	G	C	SIGLEC10
13	19	51414971	SNV	C	G	SIGLEC10
14	19	54574980	SNV	A	G	LILRA2
15	7	100399296	SNV	C	T	PILRA
16	6	32937309	SNV	G	T	HLA-DMB
17	14	61549756	SNV	T	G	PRKCH
18	12	10315316	SNV	A	G	KLRD1
19	17	27306800	SNV	T	G	WSB1
20	3	10301346	SNV	C	T	SEC13

Table 1. Somatic mutations with the top 20 saliency scores

Among these top 20 mutations, we have identified 12 of them as HGMs based on the related literature. A detailed description of the related literature for these top 20 mutations could be found in Table 3. We selected the HGMs comprehensively considering the closeness of their relationship to PTSD, the number of related research articles, the citations, and the impact factor of the journals. These 12 identified HGMs were further divided into three tiers given the strength of literature support. The identified HGMs and their tiers are listed in Table 2.

Saliency score ranking	**Chromosome**	**Position**	**Type**	**Reference allele**	**Alternative allele**	**Gene affected**	**Tier**
6	14	102050157	SNV	C	T	DYNC1H1	1
7	6	33008108	SNV	C	T	HLA-DOA	2
17	14	61549756	SNV	T	G	PRKCH	
19	17	27306800	SNV	T	G	WSB1	
20	3	10301346	SNV	C	T	SEC13	
1	8	30180593	INDEL	CAG	C	DCTN6	3
2	18	51176872	SNV	C	T	MEX3C	
5	6	41198325	SNV	C	A	TREML2	
8	7	100373707	INDEL	AT	A	PILRA	
9	9	137039846	SNV	G	A	NPDC1	
11	12	57741942	SNV	T	C	AGAP2	
15	7	100399296	SNV	C	T	PILRA	

Definition:
Tier 1: The affected gene is taken as closely related to PTSD by more than x number of authoritative/representative journal articles.
Tier 2: The affected gene is taken as related to PTSD, but the number of related articles is limited.
Tier 3: The affected gene is taken as related to more than one kind of neural/mental disease, by more than x number of authoritative/representative journal articles.

Table 2. List of Identified HGMs

HGMs validation

As mentioned in the introduction section, in Sragovich et al.[60], eight genes with somatic mutations have been suggested to be potentially related to PTSD, including TSC1, FMR1, GSK3B, EZR, TNF, IL1R2, CASP1, and CASP4. Specifically, in the dataset used in our study, there are 12 somatic mutations with high

medRxiv preprint doi: https://doi.org/10.1101/2023.01.30.23285175; this version posted February 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

or moderate impacts on those eight genes. We have checked the saliency score ranking of those 12 somatic mutations for reference. Figure 1 shows the saliency scores of all somatic mutations considered in our model, and mutations on the left have higher saliency scores. Vertical red lines correspond to the ranking of the 12 reference somatic mutations. It could be seen that they are distributed across the left-hand side of the figure, which implies the identified risky mutations in Sragovich et al.[60] also have relatively higher saliency scores calculated by our model. Hence, to a certain extent, our saliency score supports the potential importance of genes identified in Sragovich et al.[60]. The 12 reference mutations are not the ones with top saliency scores in our study. One possible reason is that Sragovich et al.[60] concentrated on the high-impact mutations while our study focused on both high-impact and moderate-impact mutations.

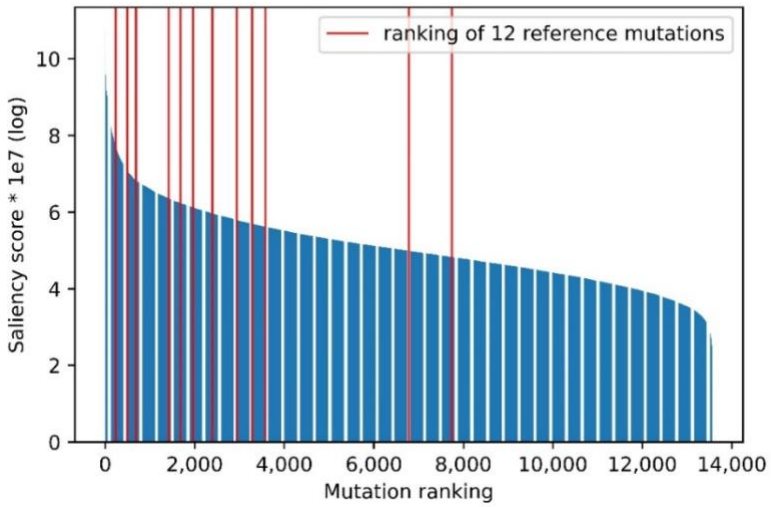

Figure 1. Ranking based on mutation saliency score

To further validate the model's capability of identifying potential HGMs, we investigated the genes affected by the top 20 HGMs identified by our model, specifically their presence in research related to PTSD or other mental/neural diseases. The supportive literature is summarized in Table 3. It could be seen that most of the genes affected by the top-ranking mutations have also been identified as biomarkers for mental diseases or neural diseases by other researchers. Among them, the gene DYNC1H1 has been recognized by many researchers as not only a potential biomarker but also a key target in the treatment of PTSD. The supportive literature further validates our model's capability of identifying potential HGMs.

Ranking based on saliency score	Affected gene	Related disease	Representative literature
1	DCTN6	Neural diseases	It was found that DCTN6 deficiency enhances aging in mouse brains[74]. DCTN6 was also identified to have close relation to protein PQBP1, which has been linked to intellectual disability disorders and progressive neurodegenerative diseases[75].
2	MEX3C	Mental diseases and neural diseases	MEX3C was identified as one of the risk genes contributing to Neurodegenerative brain diseases[76]. The location of the gene 18q21.2 has been suggested highly correlated with some mental disorders including schizophrenia[77] and depression[78].
3	CD96	Other diseases	CD96 was related to immune system[79].
4	CD3G	Mental diseases	CD3G was related to immune system. It was identified to be associated with prenatal depressive symptoms[80].

medRxiv preprint doi: https://doi.org/10.1101/2023.01.30.23285175; this version posted February 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

5	TREML2	Neural diseases	TREML2 was identified to be Alzheimer's disease risk genes[81]. Missense variant in TREML2 was found to be protective against AD[82].
6	DYNC1H1	PTSD	DYNC1H1 was not only identified as biomarker for diagnosis of PTSD[83], but also a key target in the treatment of PTSD[84].
7	HLA-DOA	PTSD	HLA-DOA was related to immune system. It was identified to be related to PTSD in an article using Transcriptome-wide association studies (TWAS)[85].
8, 15	PILRA	Neural diseases	PILRA was identified to be correlated to Alzheimer's disease by many articles[86,87].
9	NPDC1	Mental diseases and neural diseases	NPDC1 was identified to be correlated to Alzheimer's disease[88]. It was also identified to be associated with development and prognosis of schizophrenia[89].
10	CD1C	Mental diseases	CD1C was related to immune system. It was suggested to have a good diagnostic performance in major depressive disorder[90].
11	AGAP2	Mental diseases	AGAP2 was identified as risk gene related to autism in many research articles[91]. It was also identified as differentially methylated gene related to Alzheimer's disease[92].
12, 13	SIGLEC10	Other diseases	SIGLEC10 was related to immune system[93].
14	LILRA2	Mental diseases	LILRA2 was identified as differentially expressed transcripts and genes in a study on loneliness[94].
16	HLA-DMB	Other diseases	HLA-DMB was mainly related to immune system. It was identified as a gene associated to schizophrenia[95].
17	PRKCH	PTSD	PRKCH was identified as a gene significantly upregulated in PTSD cases compared to controls[96]. It was also identified as a gene closely associated with stroke[97].
18	KLRD1	Mental diseases	KLRD1 was identified as possible therapeutic targets of stress-related disorders[98].
19	WSB1	PTSD	WSB1 was identified as a gene significantly upregulated in PTSD cases compared to controls[96]. It was also suggested to be protective towards Parkinson's disease[99].
20	SEC13	PTSD	SEC13 is part of the complex mTORC1, which is identified as a key to the formation and also the treatment target of PTSD[100,101] and Alzheimer's disease[102].

Table 3. A description of the top 20 somatic mutations from representative literature

Network-learned edges

The graph attention convolutional network in our model could learn a set of edge parameters from the training data, which represents the network's belief in the connection strength between gene pairs. In Fig. 2, we visualized the connectivity of gene node DYNC1H1, which is affected by the top identified HGM. The first-order neighbors are the gene nodes connected to DYNC1H1 with the largest edge parameters. The orange-colored nodes are the overlap of these first-order neighbours and the genes affected by mutations of the top 20 saliency scores. Since the gene DYNC1H1 has been implied to be correlated to PTSD by a sufficient number of research articles, these network-learned neighboring genes could be suggestive candidates for potential PTSD-related pathway studies.

medRxiv preprint doi: https://doi.org/10.1101/2023.01.30.23285175; this version posted February 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

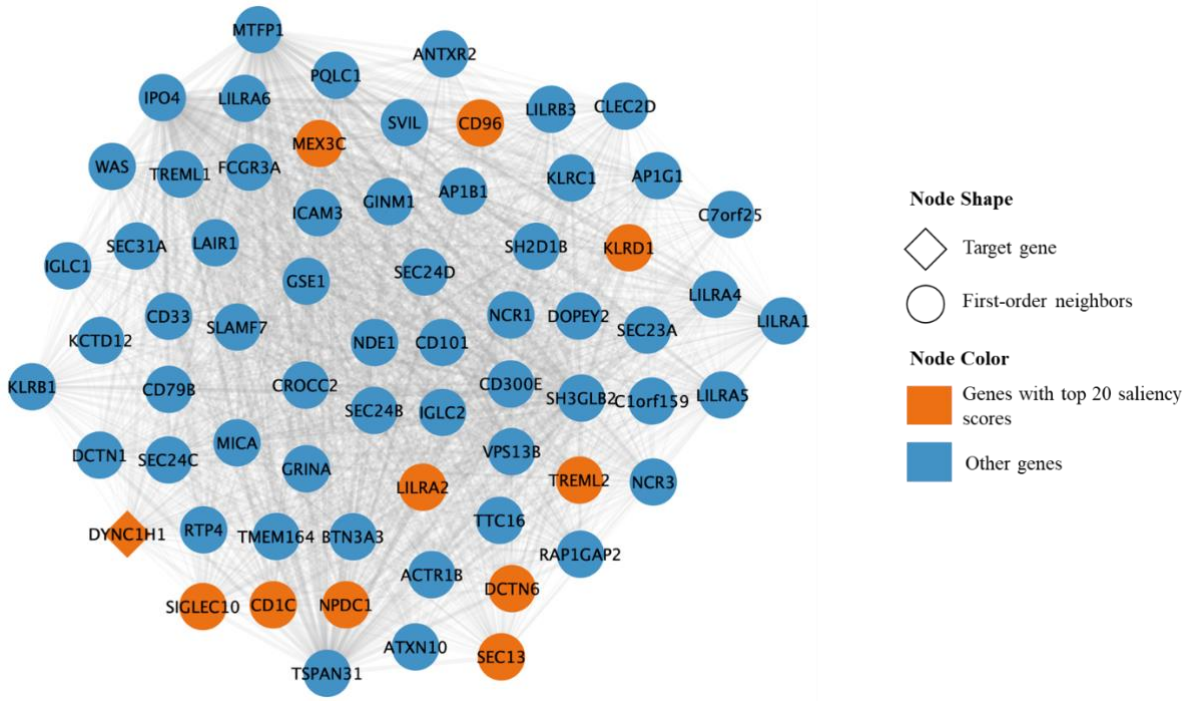

Figure 2. Visualization o network-learned graph edges for gene DYNC1H1

Discussion

A novel graph-based deep learning diagnosis model has been developed to identify probable HGMs for PTSD. Hierarchical graph structures and graph attention mechanisms have been utilized in the model to incorporate a variety of domain knowledge on somatic mutations, genes, pathways, and their correlations [71,72].

Our model has identified 12 mutations as potential HGMs for PTSD, and there are 11 genes whose corresponding proteins are affected by these 12 mutations at high or moderate impact levels. Among the identified high-risk genes, DYNC1H1 was also identified in previous research as not only a biomarker but also a key treatment target for PTSD. For other identified genes, literature support could also be found proving their correlation to PTSD or mental/neural diseases. The learned edge parameters of our graph attention network also provide suggestive candidates for PTSD-related pathway discovery.

It is worth noting that 5 out of the 11 high-risk genes (TREML2, PILRA, NPDC1, AGAP2, SEC13) identified to be related to PTSD by our model were suggested to be closely related to Alzheimer's Disease in previous research. This implies that there may be underlying connections between these two diseases. Besides, two of our identified genes were suggested to be correlated to depressive symptoms (CD3G, CD1C) and two of them were suggested to be correlated to schizophrenia (NPDC1, HLA-DMB). These findings may provide new insights for future research on not only biomarker identification, but also potential treatment studies like drug repurposing.

Last but not the least, a notable proportion of the top 20 mutations affect genes that are mainly related to the immune system (CD96, CD3G, CD1C, HLA-DOA, HLA-DMB, SIGLEC10). As shown in Table 3, HLA-DMB and HLA-DOA were identified to be associated with schizophrenia[95]. In that study, the authors suggested that their regression analysis supported disease mechanisms that involve the activity of immunity-related pathways in the brain. The similar findings in this study could further demonstrate the importance of the role that immune system plays in the PTSD disease mechanism.

In the future, research efforts are encouraged on improving the HGM identification procedure. A limitation of the current study is the small number of subjects, which might lead to overfitting in the model training. In the future, larger PTSD somatic mutation datasets are in need. In addition, few-shot learning techniques that are tailor-made for HGM identification are suggested to be developed, so as to

medRxiv preprint doi: https://doi.org/10.1101/2023.01.30.23285175; this version posted February 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

fundamentally reduce the burden in data collection. Moreover, future studies might also consider the incorporation of quantitative PTSD phenotype measurement, additional demographic information and environmental factors into the model, such as the severity and duration of subjects' trauma exposure [50,103,104].

Method

Data Collection and Pre-processing

Two types of data have been collected in this study, including sample data and knowledge-based data. The sample data contained genetic information and demographic information of the studied samples, and were used as the network input. The knowledge-based data contained the biomedical information of mankind, including pathway data and CADD scores, and were used in the construction of the neural network. Figure 3 illustrates our data extraction procedure. Details are specified in the following sections:

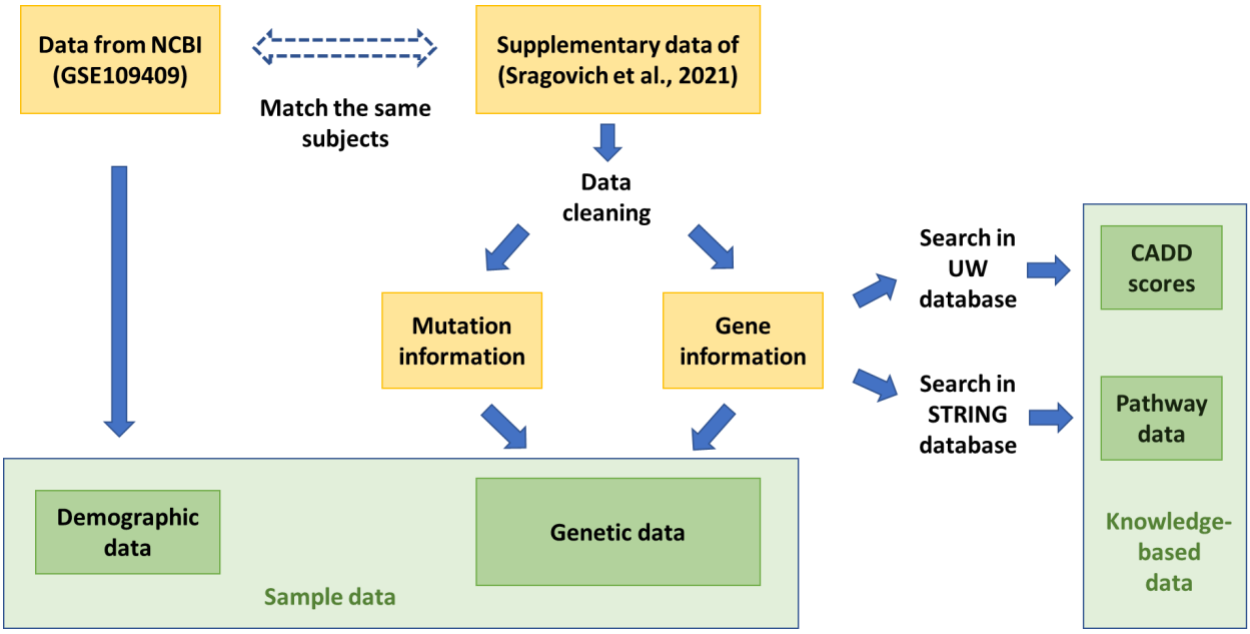

Figure 3. Data extraction procedures: National center for biotechnology information (NCBI), Gene expression omnibus series (GSE), Search tool for retrieval of interacting genes/proteins (STRING), On-line CADD scoring system of Washington University (UW) [105-108]

Genetic data

The genetic data were obtained from the supplementary table of a study conducted by Sragovich et al., which extracted somatic mutations from blood samples of 85 Canadian infantry soldiers, including 27 PTSD patients and 58 normal controls [60]. Specifically, the dataset contains information on somatic Single-nucleotide Variants (SNVs) and Insertions/Deletions (INDELs) with high/moderate impacts on proteins, and information on genes of the proteins they influence. The SNVs and INDELs were mapped to the GRCh38 human reference genome [60]. Some mismatched columns in the table have been detected and corrected in the data cleaning. Synonyms referring to the same gene in the dataset have been replaced by the latest gene symbol among them using the Ensembl website and MyGene[109,110]. Sragovich et al. [60] was based on RNA-seq data of subjects, and RNA-seq read frequencies of SNVs/INDELs have also been extracted from its supplementary table. The SNVs/INDELs without read frequency information in partial subjects (taking up around 1% of all SNVs and INDELs in all subjects) were taken as non-existing mutations in corresponding subjects. Moreover, to account for the fact that somatic mutations are more likely to be implicated in long genes, the lengths of genes were obtained from the Ensembl website and MyGene[109,110].

medRxiv preprint doi: https://doi.org/10.1101/2023.01.30.23285175; this version posted February 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Pathway data

The pathway information was obtained from the STRING database by setting all genes with somatic mutations in the dataset of Sragovich et al.[60] as inputs, and the STRING database returned pathways that the input genes are enriched in [105,111]. Specifically, pathways or biological processes from three databases, including the Biological Process (Gene Ontology) database, the KEGG pathway database and the Reactome pathway database, have been extracted [112-114]. Then, PTSD-related pathways were selected from them by a domain expert to generate the final PTSD-related pathway dataset [115].

CADD scores

The CADD (Combined Annotation Dependent Depletion) score tool[116] is a logistic regression model for the calculation of variant impact, and it has been utilized to identify genetic markers in previous studies[117]. The CADD tool requires the following variant information: CHROM, POS, REF, and ALT. For the somatic mutations with high/moderate impacts on proteins, we obtained their Phred-scaled CADD scores from the online calculation system provided by Washington University[108].

Demographic data

For each subject, we obtained the corresponding demographic information, including age group (six age groups, including 18-24, 25-30, 31-36, 37-42, 43-50, 50-61 years old) and gender, from the original dataset used in Sragovich, et al [60] (i.e. NCBI accession number: GSE109409 [107,118]).

Ethics approval and consent to participate

Datasets of [60,107,118] used in this study were obtained following the research protocol accepted by the Human Research Ethics Committee (HREC) of Defense Research and Development Canada (DRDC) - Protocol 2017-019, and informed consent was obtained from all participants [60,107,118].

Methodology

This study proposed a novel hierarchical graph attention network (H-GAtN) to identify probable HGMs for PTSD. The proposed methodology consists of four steps. Firstly, a hierarchical biomedical graph was constructed utilizing the genetic and pathway data we collected, with nodes corresponding to genes/mutations and edges corresponding to their biomedical connection. Secondly, we trained a hierarchical graph attention network to learn a high-level graph representation of the constructed biomedical graph. Thirdly, the learned high-level feature from the proposed graph neural network was combined with other demographic features, including age and gender. The concatenated features were fed into a deep feedforward neural network for the final prediction of PTSD probability. Finally, we calculated the saliency score for each somatic mutation and obtained the top 20 probable candidates for HGMs.

Knowledge-based hierarchical graph construction: mutations, genes, and pathways

A graph as a non-linear data structure can represent the interaction of an arbitrary number of nodes with arbitrary connectivity status, and is thus widely used to model complex real-life scenarios including social networks, traffic forecasting, etc. Previously, graphical convolutional neural networks have also been proven successful in dealing with molecular interaction [119] and mutation-related disease prediction [120,121]. Therefore, we adopted a graph data structure in our PTSD diagnosis scenario, and a biomedical graph was constructed to model the PTSD-related genes and mutations, making it possible for a deep neural network to learn the correlations and interactions of these biomedical concepts.

Specifically, we constructed a hierarchical biomedical graph capturing (1) the interactions of somatic mutations, (2) the interactions of genes, and (3) the hierarchical linkages between somatic mutations and genes (see Figure 4). The first hierarchy of the biomedical graph was a subgraph with each node representing a mutation associated with PTSD. The edges in this subgraph represented the mutation-mutation interactions, and all somatic mutations located on the same gene were linked to each other by undirected edges. The second hierarchy of the graph was a subgraph with each node representing a gene associated with PTSD. The edges in this subgraph represented the gene-gene interactions, and were constructed according to the PTSD-related pathways. All genes involved in the same pathway were linked

medRxiv preprint doi: https://doi.org/10.1101/2023.01.30.23285175; this version posted February 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

to each other by undirected edges. The two subgraphs were connected by mutation-gene edges. Each mutation-gene edge connected a somatic mutation to the gene on which the mutation has a high/moderate impact. The weights for the mutation-mutation edges and the gene-gene edges in the subgraph were uniformly set to 1. The weights for the mutation-gene edges were assigned with the corresponding CADD score, and were normalized to the range from 0 to 1.

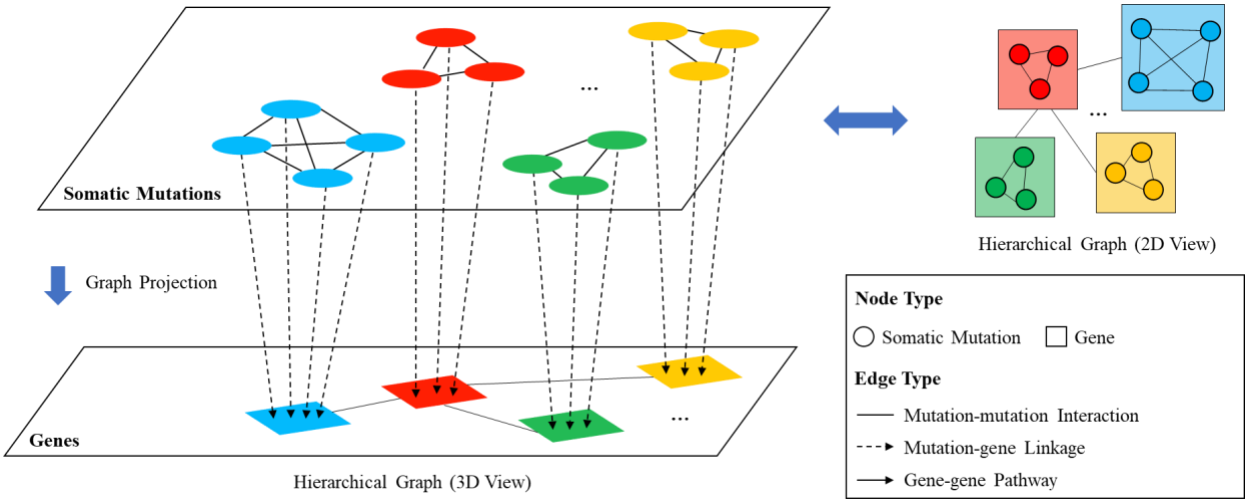

Figure 4. Hierarchical graph construction and projection

GNN model structure

After constructing the biomedical graph using domain-specific knowledge, we trained a hierarchical graph attention network to learn a high-level graph representation of the constructed biomedical graph (see Figure 5). The input of the network was the mutation frequency of one subject, and the output was a real number between 0 and 1 indicating the probability of the subject having PTSD.

Each somatic mutation was represented by a one-hot feature vector multiplied by its frequency. The one-hot representations were first embedded into a low-dimensional dense space by an embedding layer. The embedded mutation feature vectors were then fed to the first graph convolutional layer, which conducted graph convolution based on the mutation subgraph. The resulting mutation node features were fed to a graph projection layer, which adjusted the node number and dimension of the node features. Then the projected node features were fed into the second graph convolutional layer, which conducted graph convolution based on the gene subgraph. To better address the complex gene-gene interconnections that are yet to be captured by domain knowledge, a graph attention convolutional layer was incorporated as the third graph convolutional layer. After the hierarchical graph learning, an attention-based fusion layer was incorporated on top of the graph neural network to combine the gene node features and generate a supreme node with its feature vector reflecting the overall information for the subject.

The feature vector learned by our H-GAtN model was then concatenated with the demographic features of the subject. The concatenated vector was fed into a feedforward neural network which generated the final prediction of a real-number probability.

medRxiv preprint doi: https://doi.org/10.1101/2023.01.30.23285175; this version posted February 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

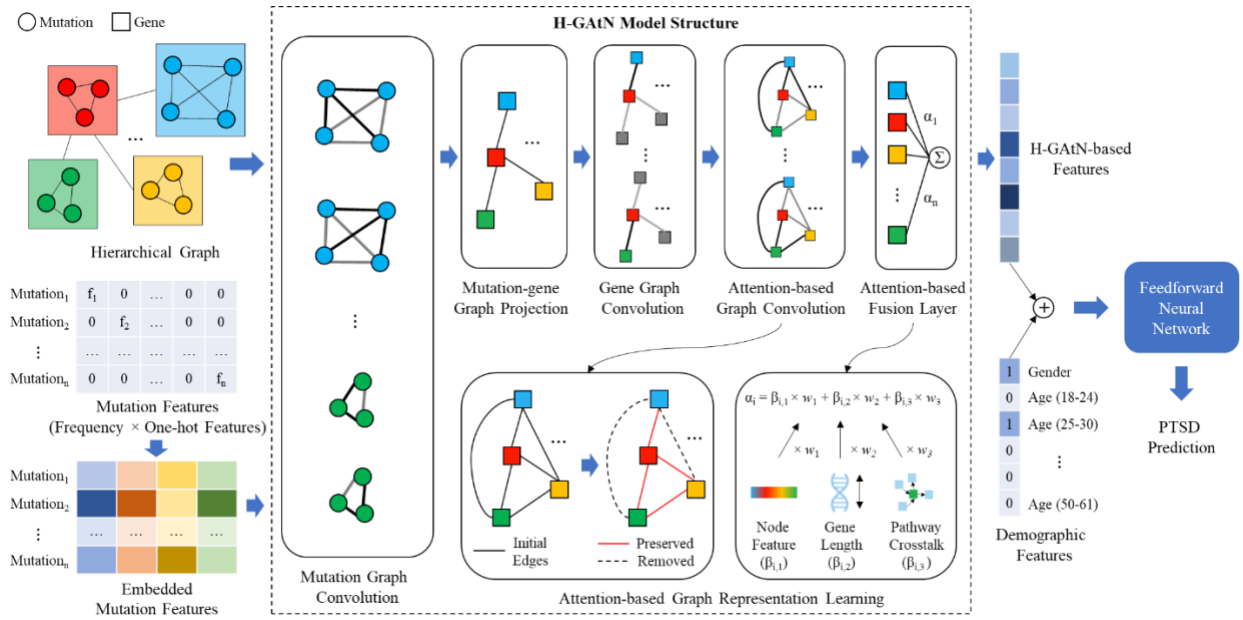

Figure 5. H-GAtN model structure

Graph projection layer

The graph projection layer was an important structure in our H-GAtN model, designed as a bridge chaining the two graph convolutional layers applied on two subgraphs. The input of the layer was the learned mutation node features $\boldsymbol{U}^{(n\times j)}$, and the layer projected the input into gene node features $\boldsymbol{V}^{(m\times i)}$. n and j are the number of mutation nodes and the dimension of node features in the mutation subgraph. m and i are the number of gene nodes and the dimension of node features in the gene subgraph. Equation (1) illustrates the projection operation. $\boldsymbol{E}^{(m\times n)}$ was a matrix composed of the mutation-gene edges, and left-multiplying $\boldsymbol{U}$ by $\boldsymbol{E}$ was equivalent to calculating a weighted sum for mutation nodes connected to the same gene. The mutations with higher CADD scores would have higher weights during the process. $W^{(j\times i)}$ was a set of learnable network parameters that transferred the dimension of the resulting node features.

$$\boldsymbol{V}^{(m\times i)} = \boldsymbol{E}^{(m\times n)}\boldsymbol{U}^{(n\times j)}\boldsymbol{W}^{(j\times i)} \tag{1}$$

Graph attention layer and attention-based fusion layer

Different types of graph convolutional layers were used at different hierarchies of the graph, adapting to the specific characteristics of the data. In the first hierarchy, the vanilla version of graph convolutional layer was applied to process the mutation nodes with known connections. In the second hierarchy, an attention-based graph convolutional layer was also incorporated in addition to the vanilla version of graph convolutional layer. Adopting the GAT layer proposed in Veličković et al. [72], the structure of the graph attention convolutional layer is shown in Figure 5. The layer did not require prior knowledge of the pathway information, and all the edges were uniformly initialized to 1. During the training process, the model iteratively optimized its parameters including the edge weights. After training the model on the training data for some epochs, the model preserved the edges that were useful, and removed those useless edges by assigning them a small weight. The graph attention convolutional layer was capable of discovering the pathways that are currently unknown. Stacking it with the vanilla version of the graph convolutional layer, we obtained a network that combined the advantages of the domain-knowledge-based approach and the data-driven approach.

The attention-based fusion layer utilized domain-specific knowledge to merge all the gene nodes into one single supreme node. Our attention mechanism included feature-based attention, gene-length-based attention, and pathway-based attention. Different components of the attention score captured the importance

medRxiv preprint doi: https://doi.org/10.1101/2023.01.30.23285175; this version posted February 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

of each gene in different aspects. Equations (2-3) illustrate how the final attention score for each gene node was calculated. α_{n1} was the feature-based attention score, which measured the similarity of each node feature to a set of learned parameters. α_{n2} was the gene-length-based attention score for the n_{th} gene, and was determined by the total number of nucleotides within that gene. α_{n3} was the pathway-based attention score of the n_{th} gene, and was determined by the number of pathways which included that gene.

$$\alpha_i = \beta_{i,1} \times w_1 + \beta_{i,2} \times w_2 + \beta_{i,3} \times w_3 \tag{2}$$

$$\beta_{i,1} = u_i \cdot v \tag{3}$$

where

α_i is the final attention score of the i_{th} gene,
α_{i1}, α_{i2}, and α_{i3} are three attention scores of the i_{th} gene calculated in different ways,
w_1, w_2, and w_3 are trainable scalar variables,
u_i is the node vector of the n_{th} gene,
v is a trainable vector variable.

Demographic features

The learned high-level features from the proposed H-GAtN model were combined with other demographic features, including age and gender. The categorical demographic data were first processed into dummy variables, resulting in a 0/1 vector with 6 components (one for gender information and five for age group information). The demographic features were first rescaled to the same magnitude as the learned H-GAtN-based features, and then concatenated to the learned features. The concatenated vector was fed into a deep feedforward neural network for the final prediction. Ablation study proved that adding the demographic information improved the model accuracy of identifying PTSD patients, reducing the error rate from 17.6% to 11.8%.

Saliency analysis

We conducted a saliency analysis to better understand how each mutation could affect the prediction of the final output, thus revealing the importance of each mutation in causing the disease. Previous work in computer vision [122] has shown that the gradients with respect to the input values could reflect how much each input feature contributes to the output value. The predicted output (in our case whether the subject has PTSD) of a single subject could be approximated by a linear expression, shown in Equation (4). The magnitude of each dimension of the gradient indicated the relative sensitiveness of that particular input feature. After averaging the gradient on the whole dataset, we defined the saliency score using Equation (5), where D is the whole dataset, including training and testing data. After training the model, we calculated the saliency score for each input somatic mutation, and obtained the relative importance of each mutation.

$$\hat{y}(\mathbf{x}) \approx \boldsymbol{w}(\boldsymbol{x})^T \mathbf{x} + b \tag{4}$$

$$\mathbf{s} = \sum_{\mathbf{x},y \in D} \frac{\boldsymbol{w}(\mathbf{x})}{|D|} \tag{5}$$

Data Availability

Partial genetic data supporting this study is available at the supplementary data of Sragovich et al. [60]. The demographic data used in this article is available at GEO with the accession number: GSE109409. Other datasets generated in this study will be made available upon request to the corresponding authors.

Code Availability

medRxiv preprint doi: https://doi.org/10.1101/2023.01.30.23285175; this version posted February 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

The code for this study will be made available upon request to the corresponding authors.

References

1 Yehuda, R. Post-traumatic stress disorder. *New England Journal of Medicine* **346**, 108-114 (2002).

2 Duncan, L. E. *et al.* Largest GWAS of PTSD (N= 20070) yields genetic overlap with schizophrenia and sex differences in heritability. *Molecular Psychiatry* **23**, 666-673 (2018).

3 Lu, L. *Machine learning approaches for biomarker identification and subgroup discovery for post-traumatic stress disorder*, The University of Memphis, (2020).

4 Nievergelt, C. M. *et al.* Genomic approaches to posttraumatic stress disorder: The psychiatric genomic consortium initiative. *Biological Psychiatry* **83**, 831-839 (2018).

5 APA. *Diagnostic and Statistical Manual of Mental Disorders*. 5th edn, Vol. 21 (American Psychiatric Association (APA), 2013).

6 Tull, M. *Types of re-experiences in PTSD*, <https://www.verywellmind.com/re-experiencing-2797325> (2020).

7 Brewin, C. R. Re-experiencing traumatic events in PTSD: New avenues in research on intrusive memories and flashbacks. *European Journal of Psychotraumatology* **6**, 27180 (2015).

8 Del Gaizo, A. L., Elhai, J. D. & Weaver, T. L. Posttraumatic stress disorder, poor physical health and substance use behaviors in a national trauma-exposed sample. *Psychiatry Research* **188**, 390-395 (2011).

9 Flood, A. M. *et al.* Prospective study of externalizing and internalizing subtypes of posttraumatic stress disorder and their relationship to mortality among Vietnam veterans. *Comprehensive Psychiatry* **51**, 236-242 (2010).

10 Greenberg, P. E. *et al.* The economic burden of anxiety disorders in the 1990s. *Journal of Clinical Psychiatry* **60**, 427-435 (1999).

11 Ashley-Koch, A. E. *et al.* Genome-wide association study of posttraumatic stress disorder in a cohort of Iraq–Afghanistan era veterans. *Journal of Affective Disorders* **184**, 225-234 (2015).

12 Kessler, R. C. Posttraumatic stress disorder: The burden to the individual and to society. *Journal of Clinical Psychiatry* **61**, 4-14 (2000).

13 Koenen, K. C. Genetics of posttraumatic stress disorder: Review and recommendations for future studies. *Journal of traumatic stress* **20**, 737-750 (2007).

14 Xie, P. *et al.* Genome-wide association study identifies new susceptibility loci for posttraumatic stress disorder. *Biological Psychiatry* **74**, 656-663 (2013).

15 Breslau, N., Davis, G. C., Andreski, P. & Peterson, E. Traumatic events and posttraumatic stress disorder in an urban population of young adults. *Archives of General Psychiatry* **48**, 216-222 (1991).

16 Radant, A., Tsuang, D., Peskind, E. R., McFall, M. & Raskind, W. Biological markers and diagnostic accuracy in the genetics of posttraumatic stress disorder. *Psychiatry Research* **102**, 203-215 (2001).

17 Cornelis, M. C., Nugent, N. R., Amstadter, A. B. & Koenen, K. C. Genetics of post-traumatic stress disorder: Review and recommendations for genome-wide association studies. *Current Psychiatry Reports* **12**, 313-326 (2010).

18 Goldberg, J., True, W. R., Eisen, S. A. & Henderson, W. G. A twin study of the effects of the Vietnam war on posttraumatic stress disorder. *Jama* **263**, 1227-1232 (1990).

medRxiv preprint doi: https://doi.org/10.1101/2023.01.30.23285175; this version posted February 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

19 Xian, H. *et al.* Genetic and environmental influences on posttraumatic stress disorder, alcohol and drug dependence in twin pairs. *Drug and Alcohol Dependence* **61**, 95-102 (2000).

20 Yehuda, R., Koenen, K. C., Galea, S. & Flory, J. D. The role of genes in defining a molecular biology of PTSD. *Disease Markers* **30**, 67-76 (2011).

21 Smoller, J. W. The genetics of stress-related disorders: PTSD, depression, and anxiety disorders. *Neuropsychopharmacology* **41**, 297-319 (2016).

22 Almli, L. M. *et al.* A genome-wide identified risk variant for PTSD is a methylation quantitative trait locus and confers decreased cortical activation to fearful faces. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **168**, 327-336 (2015).

23 Stein, M. B., Jang, K. L., Taylor, S., Vernon, P. A. & Livesley, W. J. Genetic and environmental influences on trauma exposure and posttraumatic stress disorder symptoms: A twin study. *American Journal of Psychiatry* **159**, 1675-1681 (2002).

24 Sheerin, C. M., Lind, M. J., Bountress, K. E., Nugent, N. R. & Amstadter, A. B. The genetics and epigenetics of PTSD: Overview, recent advances, and future directions. *Current Opinion in Psychology* **14**, 5-11 (2017).

25 Banerjee, S. B., Morrison, F. G. & Ressler, K. J. Genetic approaches for the study of PTSD: Advances and challenges. *Neuroscience Letters* **649**, 139-146 (2017).

26 Polimanti, R. & Wendt, F. R. Posttraumatic stress disorder: From gene discovery to disease biology. *Psychological Medicine*, 1-11 (2021).

27 Binder, E. B. *et al.* Association of FKBP5 polymorphisms and childhood abuse with risk of posttraumatic stress disorder symptoms in adults. *Jama* **299**, 1291-1305 (2008).

28 Gressier, F. *et al.* The 5-HTTLPR polymorphism and Posttraumatic Stress Disorder: A meta-analysis. *Journal of traumatic stress* **26**, 645-653 (2013).

29 Segman, R. *et al.* Association between the dopamine transporter gene and posttraumatic stress disorder. *Molecular Psychiatry* **7**, 903-907 (2002).

30 Lee, H. J. *et al.* Influence of the serotonin transporter promoter gene polymorphism on susceptibility to posttraumatic stress disorder. *Depression and Anxiety* **21**, 135-139 (2005).

31 Xie, P. *et al.* Interaction of FKBP5 with childhood adversity on risk for post-traumatic stress disorder. *Neuropsychopharmacology* **35**, 1684-1692 (2010).

32 Kolassa, I.-T., Kolassa, S., Ertl, V., Papassotiropoulos, A. & Dominique, J.-F. The risk of posttraumatic stress disorder after trauma depends on traumatic load and the catechol-o-methyltransferase val158met polymorphism. *Biological Psychiatry* **67**, 304-308 (2010).

33 Amstadter, A. B. *et al.* Variant in RGS2 moderates posttraumatic stress symptoms following potentially traumatic event exposure. *Journal of anxiety disorders* **23**, 369-373 (2009).

34 Comings, D. E. *et al.* The dopamine D2 receptor locus as a modifying gene in neuropsychiatric disorders. *Jama* **266**, 1793-1800 (1991).

35 Comings, D., Muhleman, D. & Gysin, R. Dopamine D2 receptor (DRD2) gene and susceptibility to posttraumatic stress disorder: A study and replication. *Biological Psychiatry* **40**, 368-372 (1996).

36 Kilpatrick, D. G. *et al.* The serotonin transporter genotype and social support and moderation of posttraumatic stress disorder and depression in hurricane-exposed adults. *American Journal of Psychiatry* **164**, 1693-1699 (2007).

37 Zhang, L. *et al.* Genetic association of FKBP5 with PTSD in US service members deployed to Iraq and Afghanistan. *Journal of Psychiatric Research* **122**, 48-53 (2020).

medRxiv preprint doi: https://doi.org/10.1101/2023.01.30.23285175; this version posted February 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

38 Ressler, K. J. *et al.* Post-traumatic stress disorder is associated with PACAP and the PAC1 receptor. *Nature* **470**, 492-497 (2011).

39 Zhang, H. *et al.* Brain derived neurotrophic factor (BDNF) gene variants and Alzheimer's disease, affective disorders, posttraumatic stress disorder, schizophrenia, and substance dependence. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **141**, 387-393 (2006).

40 Wolf, E. J. *et al.* Corticotropin releasing hormone receptor 2 (CRHR-2) gene is associated with decreased risk and severity of posttraumatic stress disorder in women. *Depression and Anxiety* **30**, 1161-1169 (2013).

41 Gelernter, J. *et al.* No association between D2 dopamine receptor (DRD2)"A" system alleles, or DRD2 haplotypes, and posttraumatic stress disorder. *Biological Psychiatry* **45**, 620-625 (1999).

42 Wilker, S. *et al.* The role of memory-related gene WWC1 (KIBRA) in lifetime posttraumatic stress disorder: Evidence from two independent samples from African conflict regions. *Biological Psychiatry* **74**, 664-671 (2013).

43 Guffanti, G. *et al.* No association between RORA polymorphisms and PTSD in two independent samples. *Molecular Psychiatry* **19**, 1056-1057 (2014).

44 Daskalakis, N. P., Rijal, C. M., King, C., Huckins, L. M. & Ressler, K. J. Recent genetics and epigenetics approaches to PTSD. *Current Psychiatry Reports* **20**, 1-12 (2018).

45 Skelton, K., Ressler, K. J., Norrholm, S. D., Jovanovic, T. & Bradley-Davino, B. PTSD and gene variants: New pathways and new thinking. *Neuropharmacology* **62**, 628-637 (2012).

46 Logue, M. W. *et al.* A genome-wide association study of post-traumatic stress disorder identifies the retinoid-related orphan receptor alpha (RORA) gene as a significant risk locus. *Molecular Psychiatry* **18**, 937-942 (2013).

47 Guffanti, G. *et al.* Genome-wide association study implicates a novel RNA gene, the lincRNA AC068718. 1, as a risk factor for post-traumatic stress disorder in women. *Psychoneuroendocrinology* **38**, 3029-3038 (2013).

48 Stein, M. B. *et al.* Genome-wide association studies of posttraumatic stress disorder in 2 cohorts of US army soldiers. *JAMA Psychiatry* **73**, 695-704 (2016).

49 Nievergelt, C. M. *et al.* Genomic predictors of combat stress vulnerability and resilience in US marines: A Genome-wide association study across multiple ancestries implicates PRTFDC1 as a potential PTSD gene. *Psychoneuroendocrinology* **51**, 459-471 (2015).

50 Maihofer, A. X. *et al.* Enhancing discovery of genetic variants for PTSD through integration of quantitative phenotypes and trauma exposure information. *Biological Psychiatry* (2021).

51 Almli, L. M. *et al.* Follow-up and extension of a prior genome-wide association study of posttraumatic stress disorder: gene× environment associations and structural magnetic resonance imaging in a highly traumatized African-American civilian population. *Biological Psychiatry* **76**, e3-e4 (2014).

52 Wolf, E. J. *et al.* A genome-wide association study of clinical symptoms of dissociation in a trauma-exposed sample. *Depression and Anxiety* **31**, 352-360 (2014).

53 Kilaru, V. *et al.* Genome-wide gene-based analysis suggests an association between Neuroligin 1 (NLGN1) and post-traumatic stress disorder. *Translational Psychiatry* **6**, e820-e820 (2016).

medRxiv preprint doi: https://doi.org/10.1101/2023.01.30.23285175; this version posted February 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

54 Melroy-Greif, W. E., Wilhelmsen, K. C., Yehuda, R. & Ehlers, C. L. Genome-wide association study of post-traumatic stress disorder in two high-risk populations. *Twin Research and Human Genetics* **20**, 197-207 (2017).

55 Voisey, J., Young, R. M., Lawford, B. R. & Morris, C. P. Progress towards understanding the genetics of posttraumatic stress disorder. *Journal of Anxiety Disorders* **28**, 873-883 (2014).

56 Zhang, K. *et al.* An overview of posttraumatic stress disorder genetic studies by analyzing and integrating genetic data into genetic database PTSDgene. *Neuroscience & Biobehavioral Reviews* **83**, 647-656 (2017).

57 Sheerin, C. M. *et al.* Meta-analysis of associations between hypothalamic-pituitary-adrenal axis genes and risk of posttraumatic stress disorder. *Journal of Traumatic Stress* **33**, 688-698 (2020).

58 Sullivan, P. F. Spurious genetic associations. *Biological Psychiatry* **61**, 1121-1126 (2007).

59 Morey, R. A. *et al.* Genome-wide association study of subcortical brain volume in PTSD cases and trauma-exposed controls. *Translational Psychiatry* **7**, 1-10 (2017).

60 Sragovich, S., Gershovits, M., Lam, J. C., Li, V. O. & Gozes, I. Putative blood somatic mutations in post-traumatic stress disorder-symptomatic soldiers: High impact of cytoskeletal and inflammatory proteins. *Journal of Alzheimer's Disease*, 1-12 (2021).

61 Arora, P., Mishra, A. & Malhi, A. N-semble-based method for identifying Parkinson's disease genes. *Neural Computing and Applications*, 1-11 (2021).

62 Peng, J., Guan, J. & Shang, X. Predicting Parkinson's disease genes based on Node2vec and autoencoder. *Frontiers in genetics* **10**, 226 (2019).

63 Liu, H. *et al.* Predicting the disease genes of multiple sclerosis based on network representation learning. *Frontiers in Genetics* **11** (2020).

64 Chang, J.-W. *et al.* A deep learning model based on sparse auto-encoder for prioritizing cancer-related genes and drug target combinations. *Carcinogenesis* **40**, 624-632 (2019).

65 Lee, T. & Lee, H. Prediction of Alzheimer's disease using blood gene expression data. *Scientific Reports* **10**, 1-13 (2020).

66 Athilakshmi, R., Jacob, S. G. & Rajavel, R. in *Advances in Big Data and Cloud Computing* 547-554 (Springer, 2019).

67 Zafeiris, D., Rutella, S. & Ball, G. R. An artificial neural network integrated pipeline for biomarker discovery using Alzheimer's disease as a case study. *Computational and Structural Biotechnology Journal* **16**, 77-87 (2018).

68 Liu, Y., Li, Z., Ge, Q., Lin, N. & Xiong, M. Deep feature selection and causal analysis of Alzheimer's disease. *Frontiers in Neuroscience* **13**, 1198 (2019).

69 Li, V. O. *et al.* Designing a protocol adopting an artificial intelligence (AI) – driven approach for early diagnosis of late-onset Alzheimer's disease. *Journal of Molecular Neuroscience*, 1-9 (2021).

70 Gaudelet, T. *et al.* Unveiling new disease, pathway, and gene associations via multi-scale neural network. *PLOS ONE* **15**, e0231059 (2020).

71 Ying, R. *et al.* Hierarchical graph representation learning with differentiable pooling. *arXiv preprint arXiv:1806.08804* (2018).

72 Veličković, P. *et al.* Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

73 Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32**, 8026-8037 (2019).

medRxiv preprint doi: https://doi.org/10.1101/2023.01.30.23285175; this version posted February 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

74 Sharoar, M. G., Zhou, J., Benoit, M., He, W. & Yan, R. Dynactin 6 deficiency enhances aging-associated dystrophic neurite formation in mouse brains. *Neurobiology of Aging* **107**, 21-29 (2021).

75 Kunde, S. *et al.* The X-chromosome-linked intellectual disability protein PQBP1 is a component of neuronal RNA granules and regulates the appearance of stress granules. *Human molecular genetics* **20**, 4916-4931 (2011).

76 Perrone, F., Cacace, R., van der Zee, J. & Van Broeckhoven, C. Emerging genetic complexity and rare genetic variants in neurodegenerative brain diseases. *Genome medicine* **13**, 1-13 (2021).

77 Stefansson, H. *et al.* Common variants conferring risk of schizophrenia. *Nature* **460**, 744-747 (2009).

78 Zubenko, G. S., Hughes, H. B., Stiffler, J. S., Zubenko, W. N. & Kaplan, B. B. Genome survey for susceptibility loci for recurrent, early-onset major depression: results at 10cM resolution. *American Journal of Medical Genetics* **114**, 413-422 (2002).

79 Liu, F. *et al.* CD96, a new immune checkpoint, correlates with immune profile and clinical outcome of glioma. *Scientific reports* **10**, 1-10 (2020).

80 Kallak, T. K. *et al.* Maternal prenatal depressive symptoms and toddler behavior: an umbilical cord blood epigenome-wide association study. *Translational psychiatry* **12**, 1-11 (2022).

81 Hodges, A. K., Piers, T. M., Collier, D., Cousins, O. & Pocock, J. M. Pathways linking Alzheimer's disease risk genes expressed highly in microglia. *Neuroimmunology and Neuroinflammation* **8**, 245 (2021).

82 Benitez, B. A. *et al.* Missense variant in TREML2 protects against Alzheimer's disease. *Neurobiology of aging* **35**, 1510. e1519-1510. e1526 (2014).

83 Montalvo-Ortiz, J. L. *et al.* Epigenome-wide association study of posttraumatic stress disorder identifies novel loci in US military veterans. *Translational psychiatry* **12**, 1-9 (2022).

84 Su, A. *et al.* Integrated transcriptomic and metabolomic analysis of rat serum to investigate potential target of puerarin in the treatment post-traumatic stress disorder. *Annals of Translational Medicine* **9** (2021).

85 Logue, M. W. *et al.* Gene expression in the dorsolateral and ventromedial prefrontal cortices implicates immune-related gene networks in PTSD. *Neurobiology of stress* **15**, 100398 (2021).

86 Miller, J., McKinnon, L., Murcia, J. D. G., Kauwe, J. & Ridge, P. G. Synonymous variant rs2405442 in PILRA is associated with Alzheimer's disease and affects RNA expression by destroying a ramp sequence: Basic science and pathogenesis: Genetics and omics of AD. *Alzheimer's & Dementia* **16**, e045988 (2020).

87 Patel, T. *et al.* Whole-exome sequencing of the BDR cohort: evidence to support the role of the PILRA gene in Alzheimer's disease. *Neuropathology and applied neurobiology* **44**, 506-521 (2018).

88 Rathore, N. *et al.* Paired Immunoglobulin-like Type 2 Receptor Alpha G78R variant alters ligand binding and confers protection to Alzheimer's disease. *PLoS genetics* **14**, e1007427 (2018).

89 Liu, J. *et al.* The association of DNA methylation and brain volume in healthy individuals and schizophrenia patients. *Schizophrenia research* **169**, 447-452 (2015).

medRxiv preprint doi: https://doi.org/10.1101/2023.01.30.23285175; this version posted February 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

90 Ning, L. *et al.* A novel 4 immune-related genes as diagnostic markers and correlated with immune infiltrates in major depressive disorder. *BMC immunology* **23**, 1-10 (2022).

91 C Yuen, R. K. *et al.* Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nature neuroscience* **20**, 602-611 (2017).

92 Zhang, L. *et al.* Epigenome-wide meta-analysis of DNA methylation differences in prefrontal cortex implicates the immune processes in Alzheimer's disease. *Nature communications* **11**, 1-13 (2020).

93 Chen, G.-Y., Tang, J., Zheng, P. & Liu, Y. CD24 and Siglec-10 selectively repress tissue damage–induced immune responses. *Science* **323**, 1722-1725 (2009).

94 Canli, T. *et al.* Differential transcriptome expression in human nucleus accumbens as a function of loneliness. *Molecular psychiatry* **22**, 1069-1078 (2017).

95 Håvik, B. *et al.* The complement control-related genes CSMD1 and CSMD2 associate to schizophrenia. *Biological psychiatry* **70**, 35-42 (2011).

96 Guardado, P. *et al.* Altered gene expression of the innate immune, neuroendocrine, and nuclear factor-kappa B (NF-κB) systems is associated with posttraumatic stress disorder in military personnel. *Journal of anxiety disorders* **38**, 9-20 (2016).

97 Stephenson, J. Genetic Stroke Risk. *JAMA* **297**, 686-686 (2007).

98 Breen, M. S. *et al.* Acute psychological stress induces short-term variable immune response. *Brain, Behavior, and Immunity* **53**, 172-182 (2016).

99 Nucifora, F. C. *et al.* Ubiqutination via K27 and K29 chains signals aggregation and neuronal protection of LRRK2 by WSB1. *Nature communications* **7**, 1-11 (2016).

100 Girgenti, M. J., Ghosal, S., LoPresto, D., Taylor, J. R. & Duman, R. S. Ketamine accelerates fear extinction via mTORC1 signaling. *Neurobiology of Disease* **100**, 1-8 (2017).

101 Oh, J.-Y. *et al.* Acupuncture modulates stress response by the mTOR signaling pathway in a rat post-traumatic stress disorder model. *Scientific Reports* **8**, 1-17 (2018).

102 Maiese, K. Taking aim at Alzheimer's disease through the mammalian target of rapamycin. *Annals of medicine* **46**, 587-596 (2014).

103 Almli, L. M., Fani, N., Smith, A. K. & Ressler, K. J. Genetic approaches to understanding post-traumatic stress disorder. *International Journal of Neuropsychopharmacology* **17**, 355-370 (2014).

104 Gillespie, C. F., Phifer, J., Bradley, B. & Ressler, K. J. Risk and resilience: Genetic and environmental influences on development of the stress response. *Depression and Anxiety* **26**, 984-992 (2009).

105 STRING. *STRING*, <https://string-db.org/> (2021).

106 NCBI. *Welcome to NCBI*, <https://www.ncbi.nlm.nih.gov/> (2021).

107 NCBI. *Series GSE109409*, <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE109409> (2021).

108 *CADD Score*, <https://cadd.gs.washington.edu/download> (2022).

109 Ensembl. *e!Ensembl*, <https://asia.ensembl.org/index.html> (2021).

110 MyGene. *Welcome to MyGene.py's documentation!*, <https://docs.mygene.info/projects/mygene-py/en/latest/> (2021).

111 Szklarczyk, D. *et al.* STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* **47**, D607-D613 (2019).

112 Geneontology. *The gene ontology resource*, <http://geneontology.org/> (2021).

medRxiv preprint doi: https://doi.org/10.1101/2023.01.30.23285175; this version posted February 1, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

113 KEGG. *KEGG PATHWAY Database*, <https://www.genome.jp/kegg/pathway.html> (2021).
114 Reactome. *Reactome*, <https://reactome.org/> (2021).
115 TAU. *Prof. Illana Gozes*, <https://english.tau.ac.il/profile/igozes> (2021).
116 Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research* **47**, D886-D894 (2019).
117 Park, J. S. *et al.* Brain somatic mutations observed in Alzheimer's disease associated with aging and dysregulation of tau phosphorylation. *Nature communications* **10**, 1-12 (2019).
118 Boscarino, C. *et al.* Using next-generation sequencing transcriptomics to determine markers of post-traumatic symptoms: Preliminary findings from a post-deployment cohort of soldiers. *G3: Genes, Genomes, Genetics* **9**, 463-471 (2019).
119 Klicpera, J., Groß, J. & Günnemann, S. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123* (2020).
120 Zhang, Y., Chen, Y. & Hu, T. PANDA: Prioritization of autism-genes using network-based deep-learning approach. *Genetic Epidemiology* **44**, 382-394 (2020).
121 Chereda, H. *et al.* Explaining decisions of graph convolutional neural networks: patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer. *Genome Medicine* **13**, 1-16 (2021).
122 Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

Acknowledgements
This research is supported in part by the US National Academy of Medicine Healthy Longevity Catalyst Award (Hong Kong), 2021 and 2022.

Author contributions
J.L. and V.L. put forward the research question, the initial methodology, contributed to the design and modification of the model architecture, revised the manuscript, and acquired research funding. Q.Z., Y.H., and R.B. further developed the neural network model for biomarker identification, based on the detailed proposal put forward by V.L. and J.L. R.B. and Y.H. collected and processed the input data. Q.Z. implemented the model and R.B. revised an intermediate version of the codes. Q.Z., R.B., and H.Y. wrote the first draft., I.G. provided the PTSD dataset and information about the PTSD-related pathways. All authors reviewed the manuscript.

Competing interests
The authors declare no competing interests.