mFusion: A multiscale fusion method bridging neuroimags to genes through neurotransmissions in mental health disorders

Luolong Cao[1,+], Zhenyi Wang[2,3,+], Zhiyuan Yuan[1,*], Qiang Luo[1,4,5,*]

[1] National Clinical Research Center for Aging and Medicine at Huashan Hospital, Institute of Science and Technology for Brain-Inspired Intelligence, Research Institute of Intelligent Complex Systems, Fudan University, Shanghai 200433, China

[2]Shanghai Institute of Hematology, State Key Laboratory of Medical Genomics, National Research Center for Translational Medicine (shanghai), Ruijin Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, 200025 Shanghai, China.

[3]MOE Key Laboratory of Bioinformatics; Bioinformatics Division and Center for Synthetic & Systems Biology, BNRist; Department of Automation, Tsinghua University, Beijing 100084, China.

[4] State Key Laboratory of Medical Neurobiology and MOE Frontiers Center for Brain Science, Institutes of Brain Science, Fudan University, Shanghai 200032, China.

[5] Shanghai Research Center of Acupuncture & Meridian, Shanghai 200433, China.

[+] These authors contributed equally

* These authors jointly supervised this work

Supplement Methods on pre-process of AHBA gene expression data

We used microarray data of brain tissues from 6 neurotypical donors provided by the Allen human brain atlas (AHBA) (http://human.brain-map.org). In this database, gene expressions were normalized across all samples and genes using BF normalization methods. We followed the AHBA pre-processing pipeline suggested by Arnatkevičiūtė et al.[1] and implemented in our previous publication2, including the following steps:

1) Probe-to-gene re-annotation: This reannotation was done by using the reference genome assembly

GRCh38.p12 (released in 2017/12). Following the research of Shen et al[2], 45,461 (77.5%) probes were annotated to unique genes. This re-annotated set of 45,461 probes corresponded to 19,951 unique genes.

2) Data filtering: The AHBA provided a binary indicator for those expressions did not exceed the background in at least 50% of all cortical and subcortical samples across all subjects. Excluding these probes gave a set of 31,342 probes and 15,409 unique genes.

3) Probe selection: As recommended by Arnatkevičiūtė et al, expressions were averaged among probes assigned to the same gene. As reported in our previous publication[2], the mean approach and the max approach gave highly correlated gene expressions (r=0.88).

4) Using the Harvard-Oxford atlas, the samples were separated into the cortical and the subcortical areas. Only the samples in the left hemisphere were used in the following analyses, since the samples in the right hemisphere were collected from only 2 of the 6 donors.

5) Normalization: The expression data were first normalized within sample and across-gene, and then normalized across samples. Given the systematic differences in gene expressions between the cortical and the subcortical areas, the normalizations were conducted separately for the samples from the subcortical and the cortical tissues. One gene failed the normalization and therefore was deleted, resulting 15,408 unique genes.

Supplement Methods of Partial Least Squares (PLS) regression process

Partial Least Squares (PLS) regression employs an iterative approach to compute latent variables via Singular Value Decomposition (SVD). Each iteration yields a set of orthogonal latent variables for both the predictor matrix X and the response matrix Y, along with their respective regression coefficients. The

procedure can be delineated as follows:

1). Standardization: Initially, the predictor matrix $\boldsymbol{X}$ and the response matrix $\boldsymbol{Y}$ are standardized, as usually we used Z-scores. This step ensures that each variable has a mean of zero and a standard deviation of one. We defined the covariance matrix as $\boldsymbol{R_1 = XY}$.

2). Covariance Decomposition: Subsequently, the covariance matrix $\boldsymbol{R_1}$ is decomposed using SVD, yielding the matrices $\boldsymbol{U}$, $\boldsymbol{S}$, and $\boldsymbol{V}$, such that $\boldsymbol{R_1 = USV^T}$. The first pair of singular vectors (i.e., the first columns of $\boldsymbol{U}$ and $\boldsymbol{V}$) are denoted as $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$ and the first singular value (i.e., the first diagonal entry of $\boldsymbol{S}$) is denoted $\boldsymbol{s}_1$. The first singular value represents the maximum covariance between the singular vectors.

3). Selection of Weights and Formation of Latent Variables: The $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$, are extracted as the weights for $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. And we derive the first pair of latent variables (i.e. components), $\boldsymbol{t_X}$ and $\boldsymbol{t_Y}$, by $\boldsymbol{t_X = X \times u_1}$ and $\boldsymbol{t_Y = Y \times v_1}$. Concurrently, regression models are constructed: $\boldsymbol{X}$ is regressed onto $\boldsymbol{t_X}$ to yield $\boldsymbol{\widetilde{X}_1}$, and $\boldsymbol{Y}$ is regressed onto $\boldsymbol{t_Y}$ to yield $\boldsymbol{\widetilde{Y}_1}$, where $\boldsymbol{\widetilde{X}_1 = t_X \times t_X^T \times X}$ and $\boldsymbol{\widetilde{Y}_1 = t_Y \times v_1^T}$.

4). Residual Calculation and Subsequent Component Extraction: A new covariance matrix $\boldsymbol{R}_2$ is constructed utilizing the residual information from $\boldsymbol{X}$ after accounting for $\boldsymbol{t_X}$ ($\boldsymbol{X_1 = X - \widetilde{X}_1}$) and $\boldsymbol{Y}$ after accounting for $\boldsymbol{t_Y}$ ($\boldsymbol{Y_1 = Y - \widetilde{Y}_1}$). This matrix serves as the input for the subsequent iteration of component extraction ($\boldsymbol{R}_2 = \boldsymbol{X}_1 \times \boldsymbol{Y}_1$).

5). Iteration: Steps 2) to 4) are reiterated until the model's predictive accuracy reaches an acceptable threshold or until the predefined number of components l is attained, with the upper limit of l being the rank of matrix $\boldsymbol{X}$.

The statistical significance of these latent variables that guarantees the generality of results was determined by the percentage of variance explained by the regression model through 1,000 permutations, while the weight of each predictor that index signal reliability was assessed and normalizes as Z-score

using 1,000 bootstraps[3].

Supplementary References

1 Arnatkeviciute, A., Fulcher, B. D. & Fornito, A. A practical guide to linking brain-wide gene expression and neuroimaging data. *Neuroimage* **189**, 353-367, doi:10.1016/j.neuroimage.2019.01.011 (2019).

2 Shen, C. *et al.* What Is the Link Between Attention-Deficit/Hyperactivity Disorder and Sleep Disturbance? A Multimodal Examination of Longitudinal Relationships and Brain Structure Using Large-Scale Population-Based Cohorts. *Biol Psychiatry* **88**, 459-469, doi:10.1016/j.biopsych.2020.03.010 (2020).

3 Krishnan, A., Williams, L. J., McIntosh, A. R. & Abdi, H. Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review. *NeuroImage* **56**, 455-475, doi:https://doi.org/10.1016/j.neuroimage.2010.07.034 (2011).

Content of Supplementary Figures

Content of Supplementary tables in Supplementary Data 1

Table S1: Descriptions on 45 neuroimaging PET maps.

Table S2: List of SCZ risk genes from four standard databases.

Table S3: List of ASD risk genes from four standard databases.

Table S4: Score of genes from mFusion related to SCZ.

Table S5: Score of genes from mFusion related to ASD.

Table S6: Enrichment analysis results of top 1541 genes related to SCZ.

Table S7: Enrichment analysis results of top 1541 genes related to ASD.

Table S8: Morphological correlation matrix of eight disorders.

Table S9: Expressional correlation matrix of eight disorders.

Table S10: Genetic correlation matrix of eight disorders.

Table S11: Enrichment analysis results of three Gene sets corresponding Figure 7E.

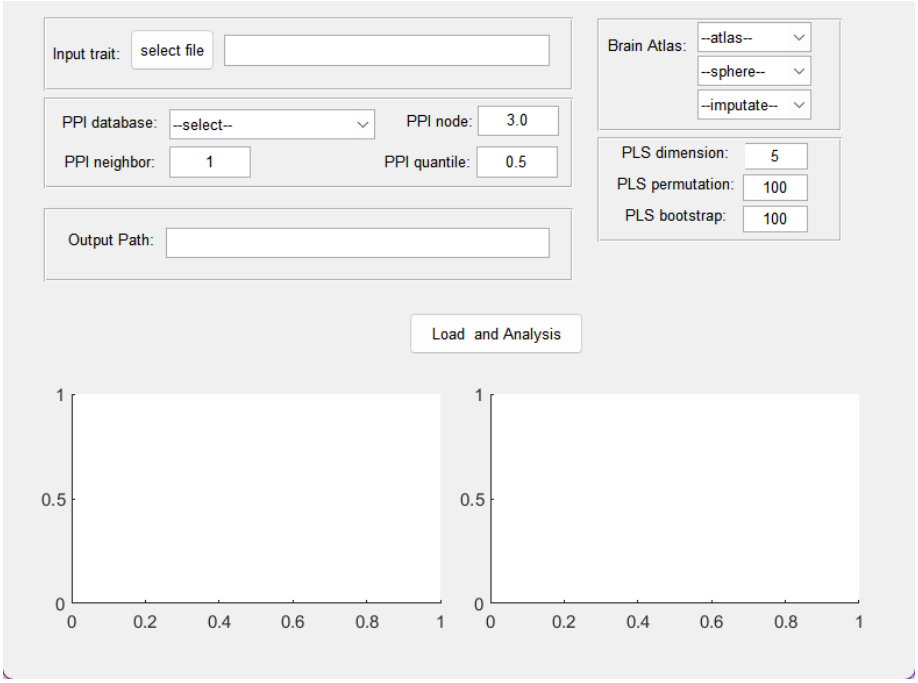
Table S12: Gene-neurotransmission-trait pathway scores of eight disorders.

Table S13: Cortical thickness differences of eight mental disorders collected by ENIGMA consortium.

Table S14: GWAS Statistics of eight disorders.

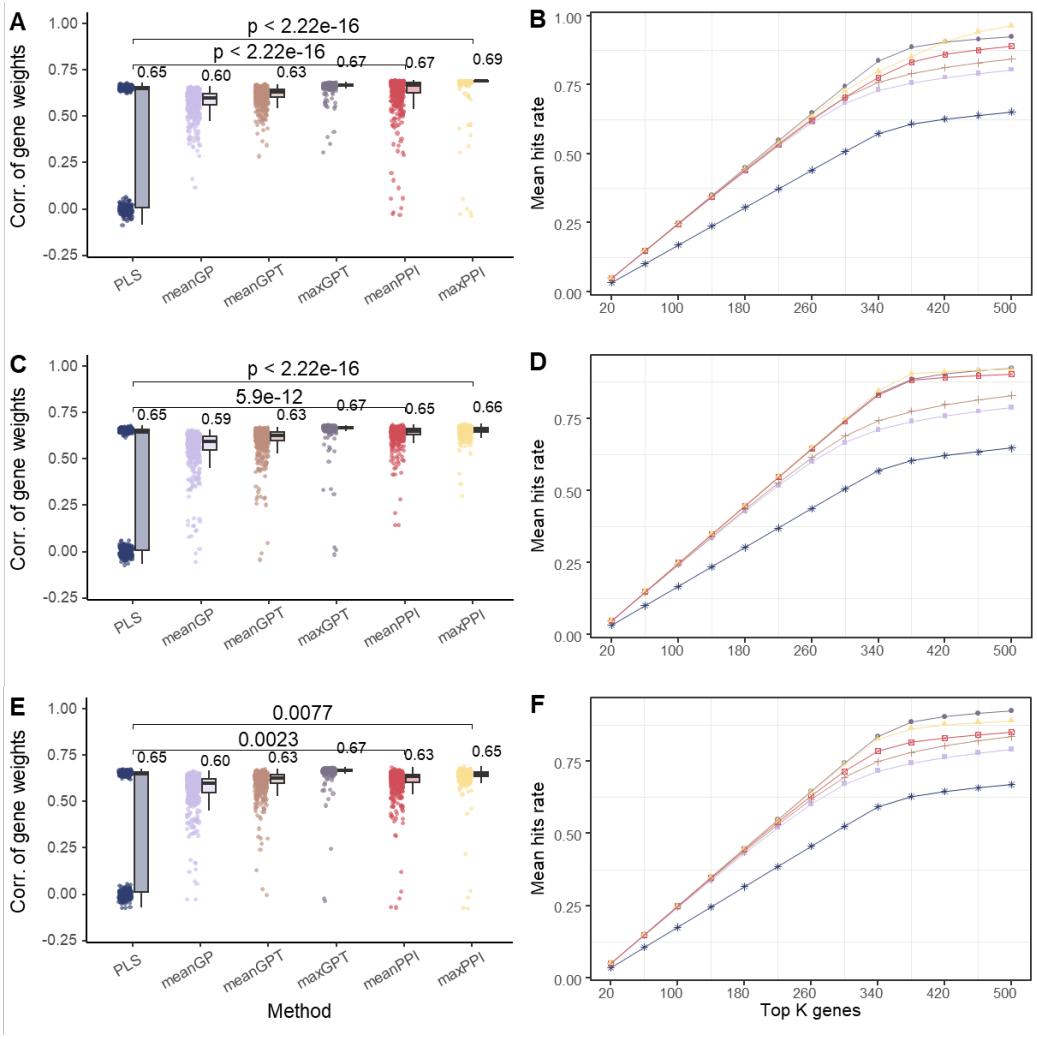
Supplementary Figures

Figure S1. The working interface of mFusion toolkit.



First, the user selects input files containing brain regions and their corresponding traits. Then, they choose the brain atlas, which serves as the standard coordinate system for input traits, AHBA gene expressions, and biomolecular PET maps. Next, users can configure the parameters of the PPI network. The "PPI database" contains protein full links or only physical links, the "PPI neighbor" (default = 1) defines the neighbor steps for a hub gene in the PPI network, the "PPI node" (default = 3) defines the which PET maps to consider in the PPI network, while the "PPI quantile" (default = 0.5) sets the edges' interaction confidence or strength. Finally, users can initiate the analysis by clicking the "Load and Analysis" button to load AHBA gene expression data and biomolecular data as well as PPI network, and to run analysis to generate scores for all genes.

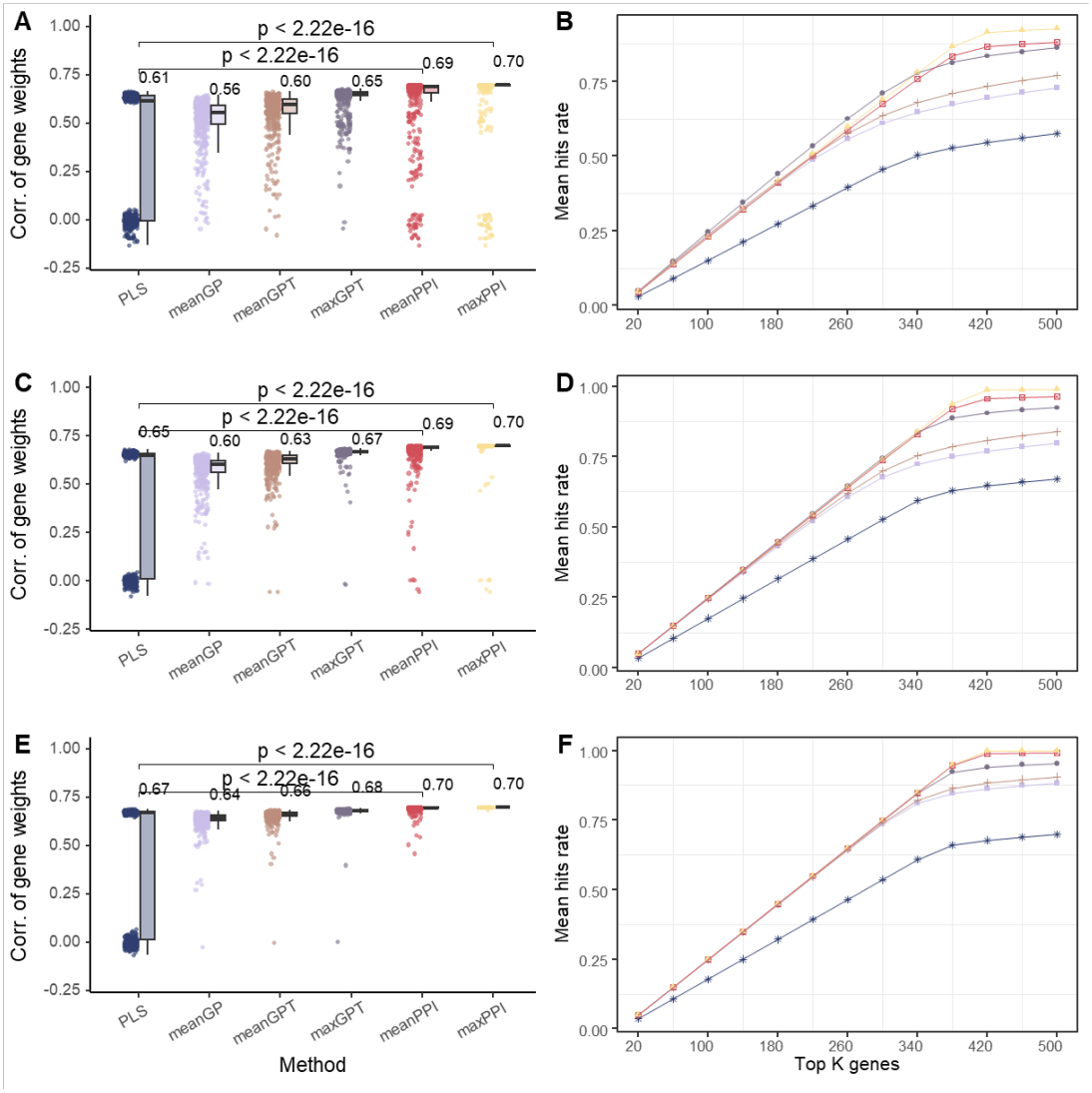
Figure S2. Evaluation of fusion methods from simulated datasets with three kinds of PPI perturbation.



A: At PPI perturbation 1 (i.e., randomly shuffled 30% of the elements within the adjacency matrix), the correlation between real gene weights and fusion weights measured by different fusion methods of 500 simulated experiments. The lower whisker extends from the first quartile (Q1) to the smallest data point that is within 1.5 * interquartile range (IQR) below Q1. The upper whisker extends from the third quartile (Q3) to the largest data point that is within 1.5 * IQR above Q3. The number next to bar represents the median of the population (using unpaired Wilcoxon test). **B:** At PPI perturbation 1, average hit rates of genes in all 500 simulations. The hit rate was measured by the rate of really active genes in top K genes ranked by specific fusion method. **C:** correlation bar plot of 500 simulations at

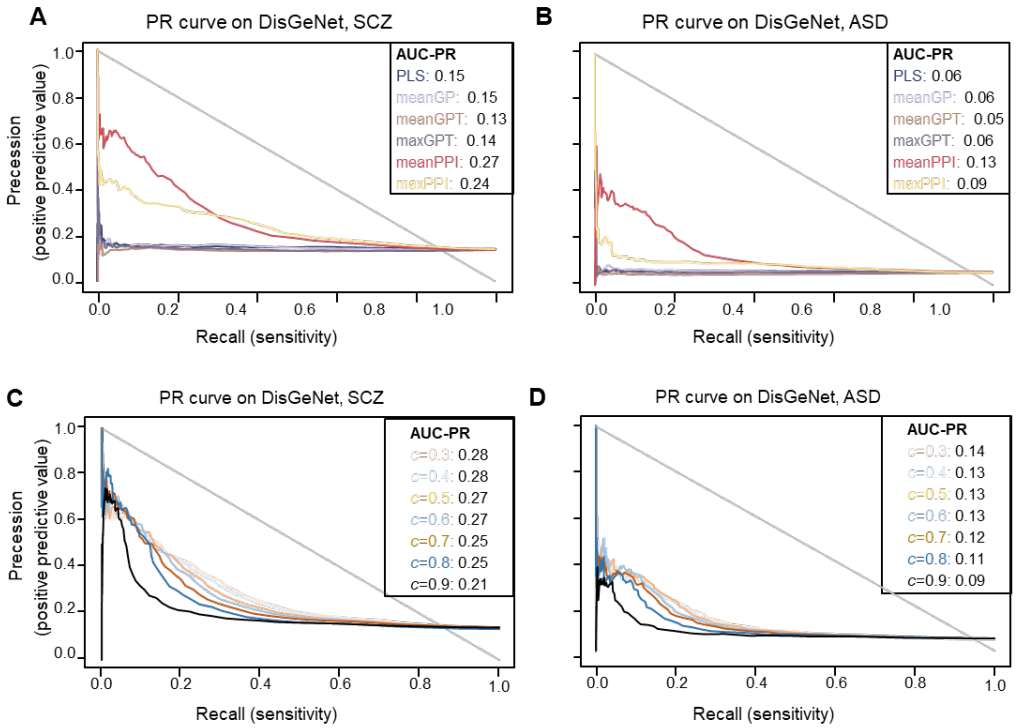
PPI perturbation 2 (i.e., set the minimum 30% of the elements in the adjacency matrix to be zero). **D:** average hit rates of genes at PPI perturbation 2. **E:** correlation bar plot of 500 simulations at PPI perturbation 3 (i.e., randomly shuffled 30% of the elements, and then set the minimum 30% of the elements in the adjacency matrix to be zero). **F:** average hit rates of genes at PPI perturbation 3.

Figure S3. Evaluation of fusion methods from simulated datasets with different spatial resolution of brain maps.



A: When the number of brain regions *n* is 100, the correlation between real gene weights and fusion weights measured by different fusion methods of 500 simulated experiments. **B:** When *n* is 100, average hit rates of genes in all 500 simulations. The hit rate was measured by the rate of really active genes in top K genes ranked by specific fusion method. **C:** When *n* is 200 (the same as in Figure 2a), correlation bar plot of 500 simulations. **D:** When *n* is 200 (the same as in Figure 2b), average hit rates of genes. **E:** When *n* is 500, correlation bar plot of 500 simulations. **F:** When *n* is 500, average hit rates of genes.

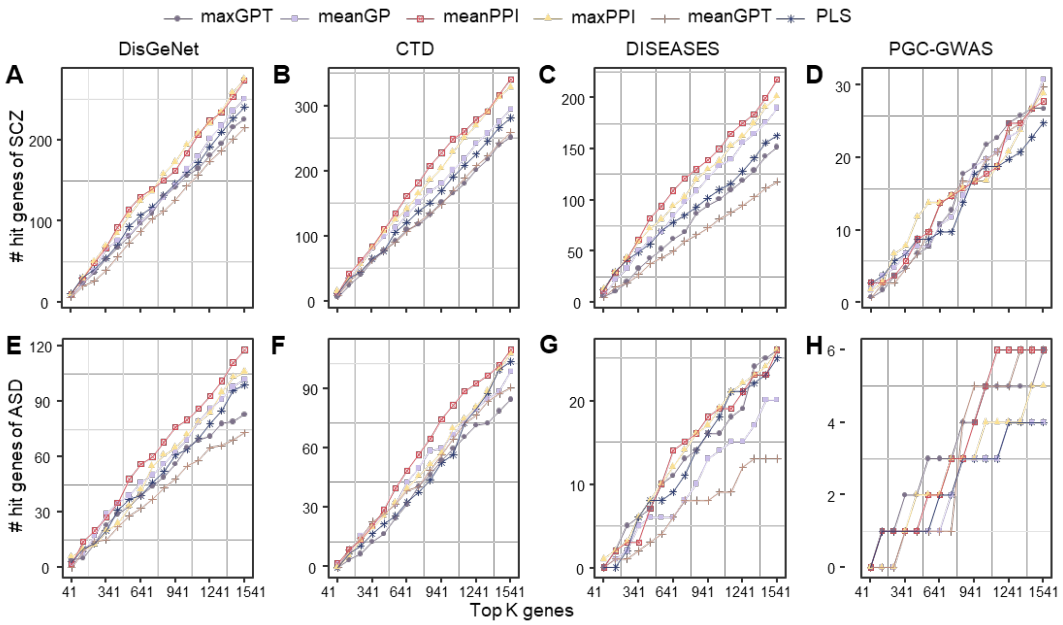
Figure S4. Evaluation of fusion methods on DisGeNet database for two disease datasets when PPI depth was set as 2.



A: PR (precision-recall) curve of different fusion methods for SCZ. **B:** PR (precision-recall) curve of different fusion methods for ASD. **C:** PR (precision-recall) curve of meanPPI fusion method for SCZ at

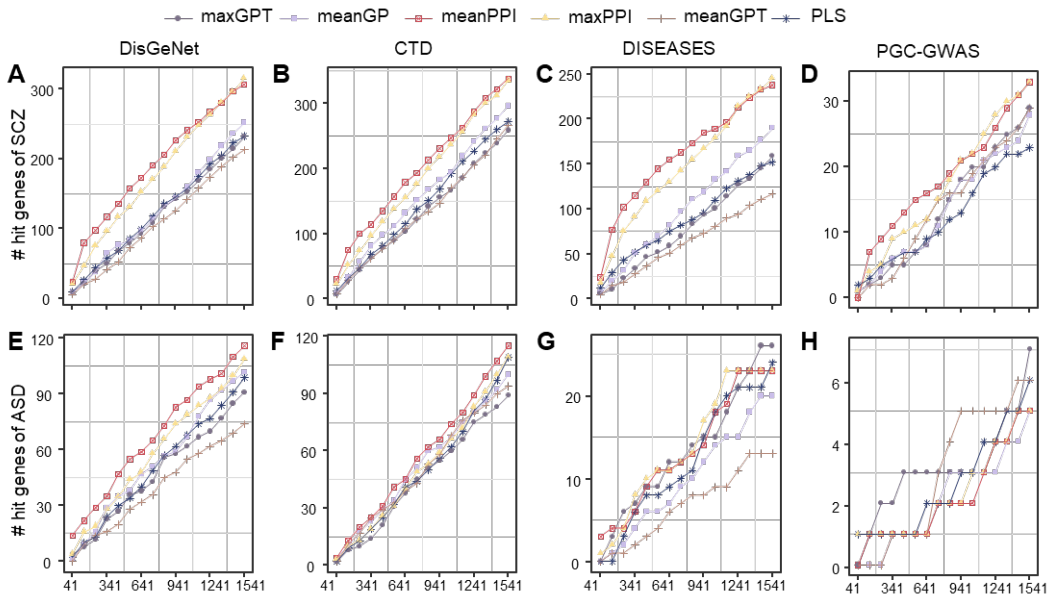
different PPI confidence when setting PPI depth as 1. **D:** PR (precision-recall) curve of meanPPI fusion method for ASD at different PPI confidence when setting PPI depth as 1.

Figure S5. Evaluation of fusion methods for two disease datasets when using the full PPI when setting the depth as 2.



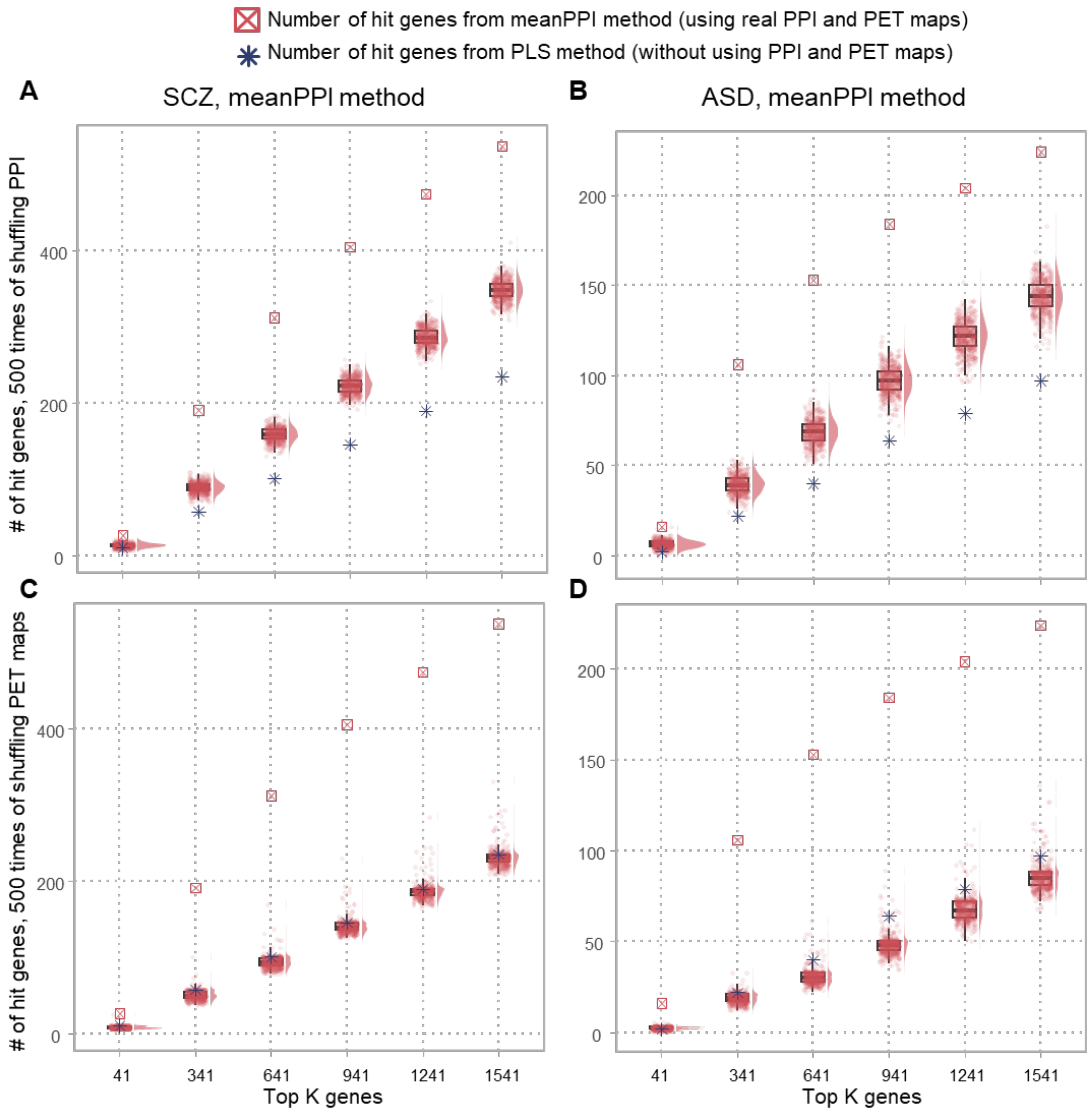
A~H: Number of overlapped genes for SCZ (A~D) and ASD (E~H) in different standard datebases: DisGeNet, CTD, DISEASES, and PGC-GWAS datasets.

Figure S6. Evaluation of fusion methods for two disease datasets when using the physical subnetwork of PPI when setting the depth as 1.



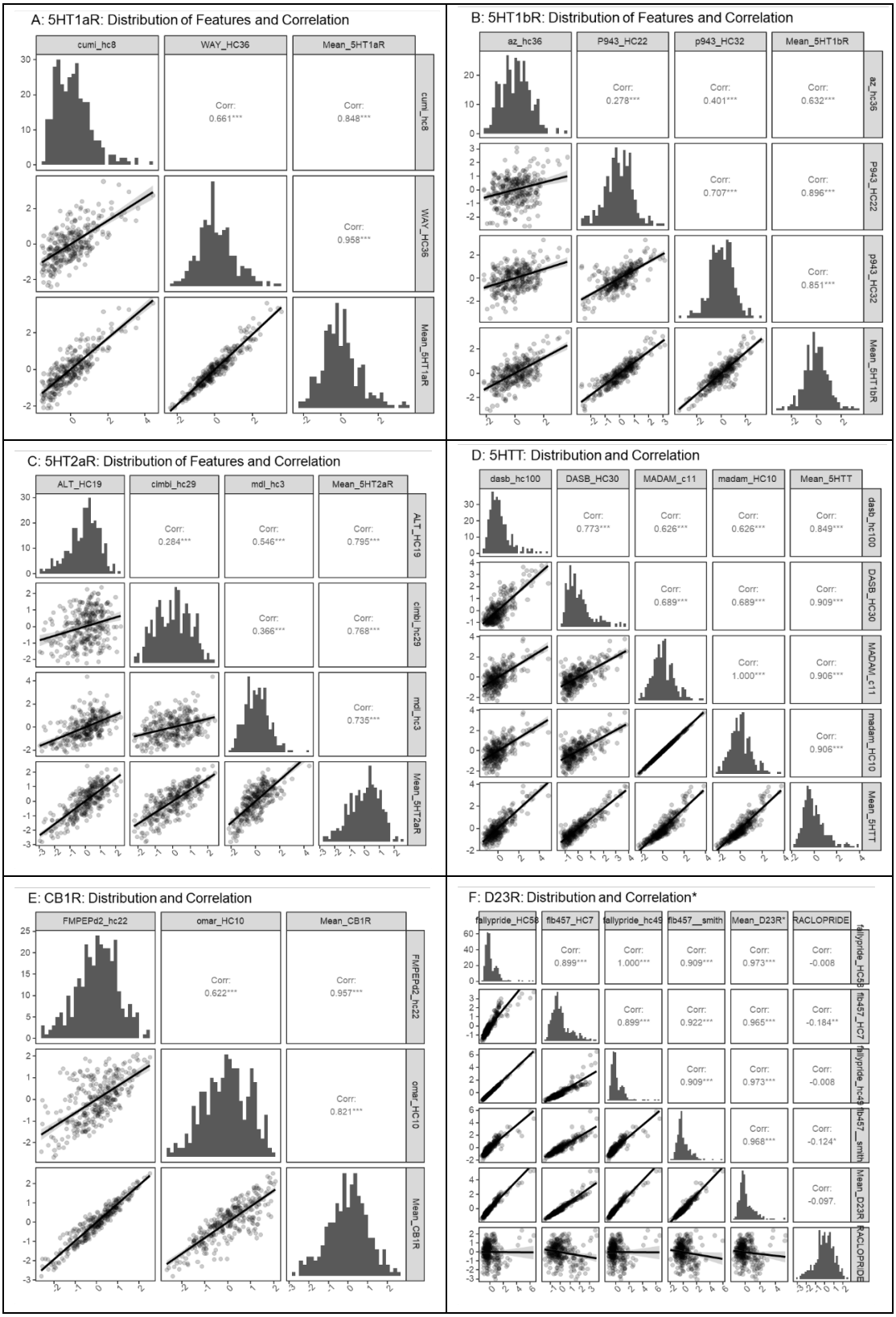
A~H: Number of overlapped genes for SCZ (A~D) and ASD (E~H) in different standard datebases: DisGeNet, CTD, DISEASES, and PGC-GWAS datasets.

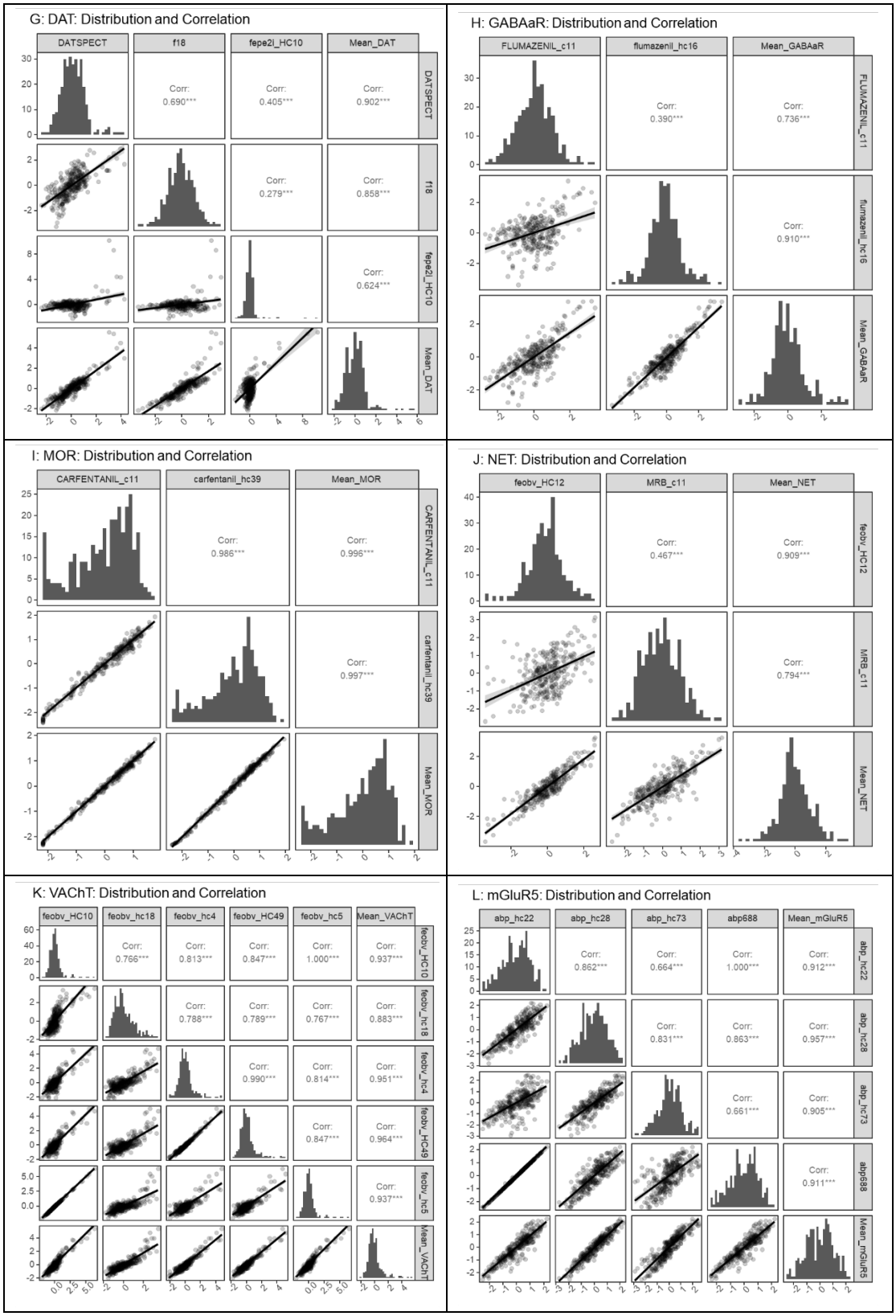
Figure S7. The distribution of hit gene numbers of SCZ and ASD disease on DisGeNet database.



A-B: The meanPPI method uses real or shuffled PPIs to obtain the number of hits in the top K gene. **C-D:** The meanPPI method uses real or shuffled PET maps to obtain the number of hits in the top K gene. The lower whisker extends from the first quartile (Q1) to the smallest data point that is within 1.5 * interquartile range (IQR) below Q1. The upper whisker extends from the third quartile (Q3) to the largest data point that is within 1.5 * IQR above Q3.

Figure S8. Distribution and correlation plot for PET maps of 12 kinds of overlapped proteins.


A: 5HT1aR: Distribution of Features and Correlation
cumi_hc8
WAY_HC36
Mean_5HT1aR
Corr: 0.661***
Corr: 0.848***
Corr: 0.958***
B: 5HT1bR: Distribution of Features and Correlation
az_hc36
P943_HC22
p943_HC32
Mean_5HT1bR
Corr: 0.278***
Corr: 0.401***
Corr: 0.632***
Corr: 0.707***
Corr: 0.896***
Corr: 0.851***
C: 5HT2aR: Distribution of Features and Correlation
ALT_HC19
cimbi_hc29
mdl_hc3
Mean_5HT2aR
Corr: 0.284***
Corr: 0.546***
Corr: 0.795***
Corr: 0.366***
Corr: 0.768***
Corr: 0.735***
D: 5HTT: Distribution and Correlation
dasb_hc100
DASB_HC30
MADAM_c11
madam_HC10
Mean_5HTT
Corr: 0.773***
Corr: 0.626***
Corr: 0.626***
Corr: 0.849***
Corr: 0.689***
Corr: 0.689***
Corr: 0.909***
Corr: 1.000***
Corr: 0.906***
Corr: 0.906***
E: CB1R: Distribution and Correlation
FMPEPd2_hc22
omar_HC10
Mean_CB1R
Corr: 0.622***
Corr: 0.957***
Corr: 0.821***
F: D23R: Distribution and Correlation*
fallypride_HC58
flb457_HC7
fallypride_hc49
flb457__smith
Mean_D23R*
RACLOPRIDE
Corr: 0.899***
Corr: 1.000***
Corr: 0.909***
Corr: 0.973***
Corr: -0.008
Corr: 0.922***
Corr: 0.965***
Corr: -0.184**
Corr: 0.909***
Corr: 0.973***
Corr: -0.008
Corr: 0.968***
Corr: -0.124*
Corr: -0.097.



*: Because the “RACLOPRIDE” PET of protein “D2R” is negatively associated with the other 4 maps, the “Mean_D23R”is calculated using the average of the other 4 maps except the “RACLOPRIDE”.

Figure S9. Performance when using 20 non-repetitive maps to different fusion methods at DisGeNet database for SCZ (A: AUC-ROC curve, B: AUC-PR curve) and ASD (D: AUC-ROC, E: AUC-PR). C and F: Comparison the gene scores of meanPPI method when using 45 PET maps or 20 PET maps for SCZ and ASD, separately.

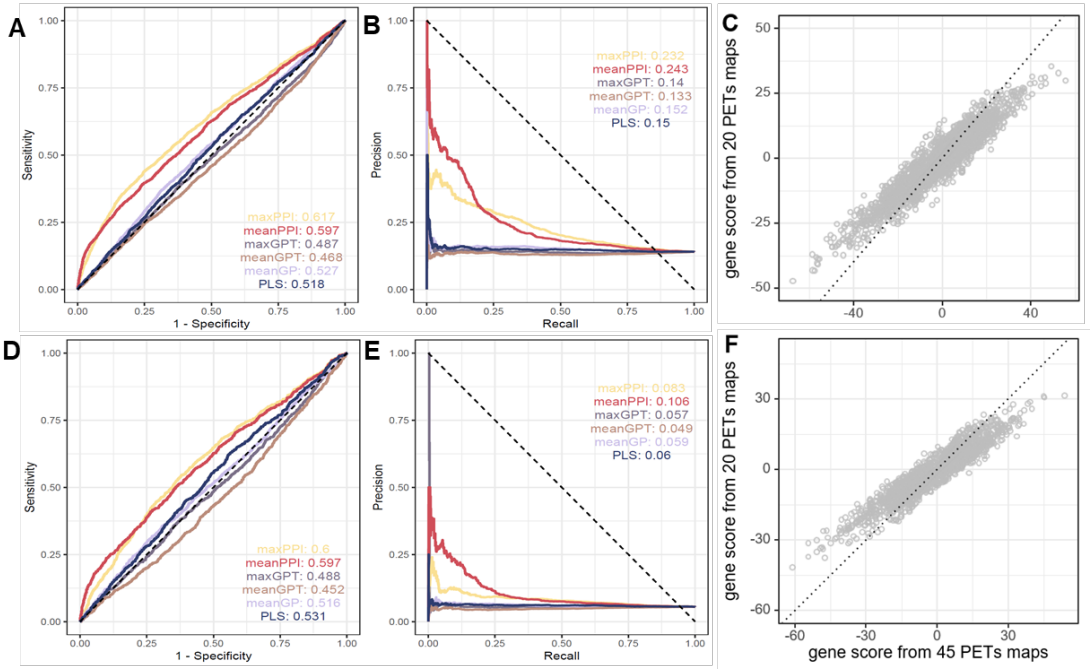


Figure S10. Number of overlapped genes for SCZ (A~D) and ASD (E~H) disease in different standard datebases when using 20 non-repetitive PET maps only at different top K genes.

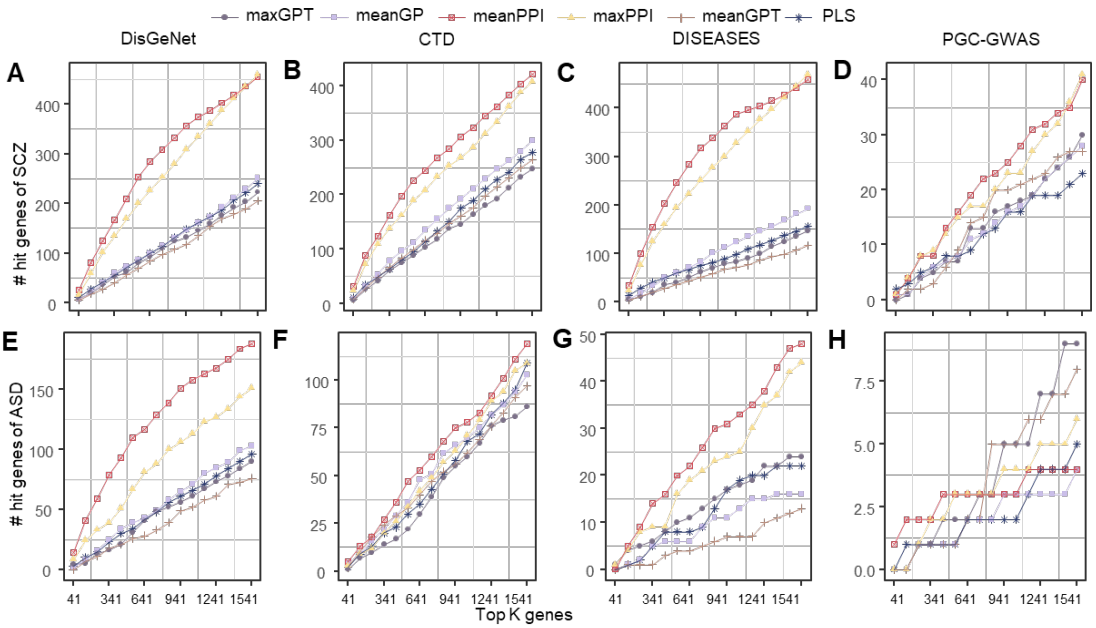


Figure S11. A: PPI network for gene KCNC1 with 20 neurochemical architectures (measured by PET maps) at STRING database with default parameters. B: Average scores of gene-neurotransmission -trait pathways across different neurotransmission types for eight disorders.

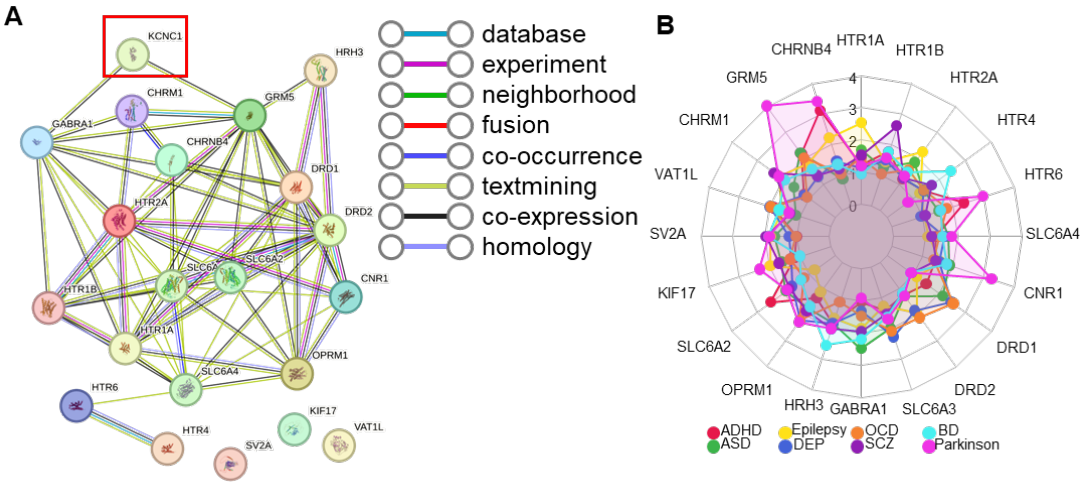
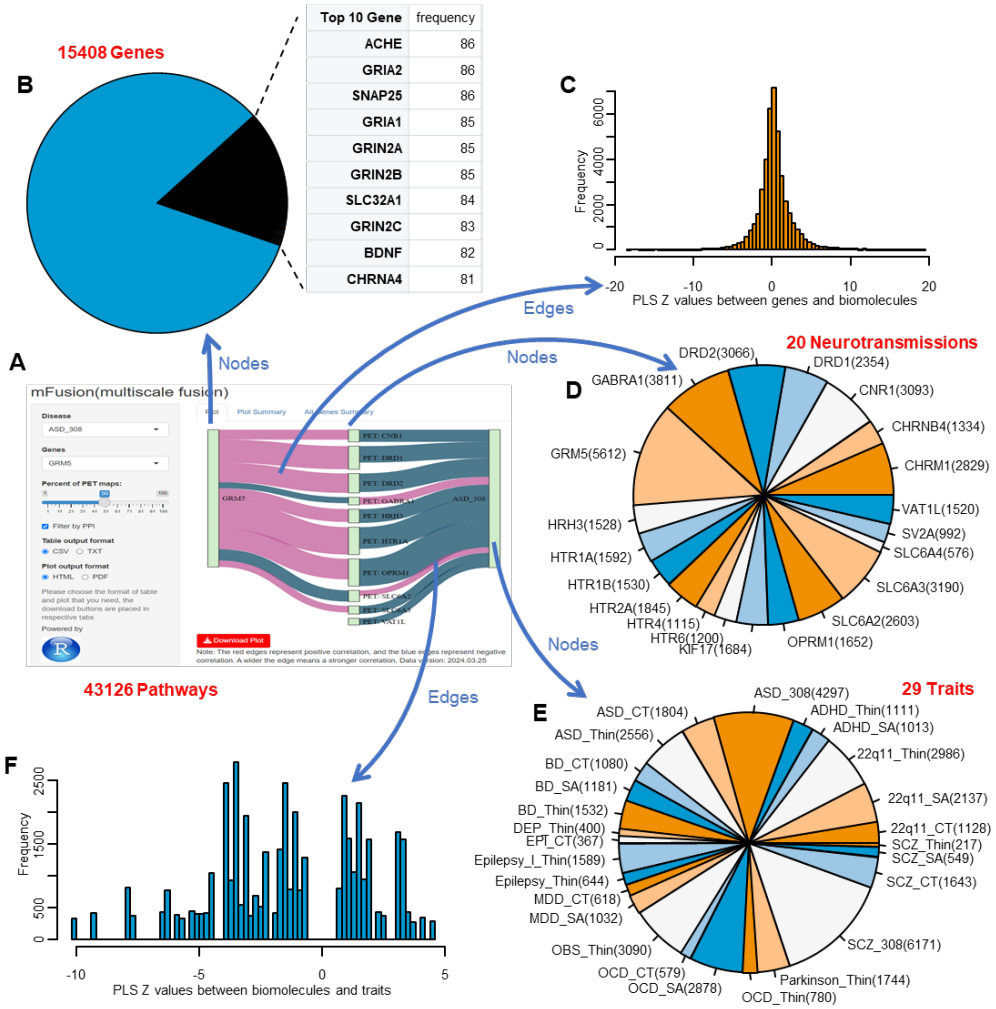


Figure S12. Gene-neurotransmission- trait pathways database generated by mFusion method.



A: Sanky plot of Gene-neurotransmission- trait pathway database, accessible var searchable web pages using the shinyAPPs platform (https://xomicsbio.shinyapps.io/mfusion_shiny/). Red edges indicate positive correlations, blue edges represent negative correlations. A wider the edge indicates a stronger correlation. B: This database contains 15,408 genes totally. The table on the right shows the top 10 frequently occurring genes. C: Distribution curve of the strength of PLS associations (measured by Z values) between genes and neurotransmissions. D: This database contains a total of 20 neurotransmissions (and correspond frequency). E: This database contains a total of 29 disease traits (and correspond frequency). Thin: thinning index generated from Cohen's *d* value of cortical thickness; CT: cortical thickness; SA: surface area. F: Strength distribution curve of PLS associations (Z values) between neurotransmissions and traits.