nature portfolio

Peer Review File

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

Corresponding Author: Professor Qiang Luo

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
In “mFusion: A multiscale fusion method for bridging neuroimaging traits to genes through neurotransmitter systems in mental health disorders”, Cao et al develop a new method to weigh genes in terms of their relevance towards a specific neuroimaging trait (e.g. case versus control cortical thickness difference in a disorder). This method combines the weights from different PLS analyses, getting to the question of how a gene may be influencing a trait, how a protein (e.g. neurotransmitter receptor) may be influencing a trait, and whether the gene influences the trait through the protein by way of protein-protein interaction. The authors test the method via simulations and empirically and find that the weighing method that involves genetic influence, protein influence, and crucially protein-protein interaction, outperforms other methods in finding previously established (and possibly new) genes that are related to a disorder. Ultimately this work makes an important contribution by demonstrating that consideration of the multiscale mechanisms in the brain - especially how they interact with one another - helps in identifying potential pathological processes. I have several questions and suggestions which I hope will improve the manuscript prior to publication.

[1] I did not fully understand the mFusion method and hope my points of confusion will help clarify the manuscript. Much of the confusion likely stems from unclear terminology. I am personally familiar with “weight” being the elements in the matrix U and V after the decomposition of the covariance matrix into USV (U are weights of the genes, S are singular values, and V are weights of the trait/neurotransmitter receptor maps). “Score” refers to the product of the original data matrix (e.g. $G_{n, u}$, where n is number of regions and u is number of genes) with the gene weights ($U_{u, l}$, where l is number of latent variable). “Latent variable” is the set of gene weights, singular value, and trait/receptor weight(s). It would be helpful if the authors could define the terminology and avoid using terms like “weight”, “score”, and “latent variable” if they do not mean the standard PLS definition of these terms. At least starting the PLS methodology with a definition of how PLS works and the terms the authors are using throughout would be very helpful in understanding mFusion.

[2] In Line 401, the authors define weight of the ith gene as the z-score of bootstrapped coefficients. Why not use the weight of the ith gene as defined by the PLS decomposition in the first matrix U?

[3] Line 409: I was confused by the definition of $Z_{g, p}(i, j)$ - what is the ith gene weight for the jth receptor map? Given my understanding of PLS (detailed in point 1), the ith gene weight would be U[i, 0] (for the first latent variable) and the jth neurotransmission weight would be V[0, j] so where is Z(i, j) coming from?

[4] I’m also confused by the terminology “neurotransmission-related neurotransmissions” (e.g. lines 407, 410).

[5] I recommend explicitly reporting where data came from, for example the name of the toolbox (e.g. neuromaps, enigmatoolbox) and link. I also strongly recommend citing the original study that collected this data. I noticed the authors provide links to papers for each receptor image but do not seem to cite these papers despite using this data.

[6] Where is the PD enigma data coming from? It is not included in the enigmatoolbox.

[7] In line 345 the authors mention “cerebral metabolic maps” - what are these? And why are they being referred to as

neurotransmission images? Likewise, the list of neurotransmission images includes UCBJ, a tracer for SV2A which is not a neurotransmitter receptor or transporter.

[8] Quality of gene expression and PET receptor data may make a big difference to the results. It is likely difficult to address how data quality impacts results but I at least recommend discussing this point in the limitations. For example, the authors include a lot of neurotransmitter receptor/transporter maps, including many that overlap in terms of protein that they bind to and even tracer. How reliable are these maps? Why not curate a more specialized list based on the quality of the image? Or at least compare images that are mapping the same protein to assess reliability of these images. For the gene expression data, the authors could filter out genes that have low reliability across donors (i.e. low differential stability).

[9] I thought the inclusion of the PPI networks was interesting and was happy to see the superior performance of PPI mFusion methods. I wonder if it's possible to benchmark their performance against a null, to be convinced that it is the biological relationship between gene and protein, as well as their specific topographic expression patterns across the brain, that result in superior performance. For example, are the number of empirical hits significantly greater when using the empirical PPI network than when using a random PPI network with equal density?

[10] To continue on the point above: if the PLS is run with a randomly shuffled (or rotated, to preserve spatial autocorrelation) gene or receptor matrix, do the number of gene hits significantly decrease?

[11] Please list the seven morphological parameters referenced in line 373.

[12] In regards to the following sentence (line 399): "We selected the first PLS component if it had a p-value < 0.05 by permutation; otherwise, the component that explained the most variance of Y." For the gene-trait and neurotransmission-trait PLSs, there is only one latent variable, which makes this sentence confusing. Perhaps stating this sentence with respect to the gene-neurotransmission PLS only? How common are non-significant latent variables in the empirical analyses?

[13] I recommend reporting the median correlation between real and fusion gene weights in each bin such that readers know by how much the different fusion methods improve the median correlation. (Fig 2a)

[14] In the section starting at line 207, the authors compare how many gene hits are found using the meanPPI method vs regular PLS. Could the authors also compare the average weight of these gene hits? I could imagine a possibility where the PLS discovers fewer genes but these genes are all at the very top of the list, whereas with meanPPI it may be more difficult to determine which genes are relevant and which are not, if no a priori information on relevant genes is known.

[15] There were some parts of the Results I felt were better left for the Discussion, including the text starting at line 221 and the entire section at line 241.

[16] Could the authors discuss how the very different spatial resolutions of microarray-derived gene expression and PET-derived receptor densities may impact the mFusion method? This could also be tested in the simulation by making the simulated receptor data more smooth.

[17] I recommend explaining how you are clustering disorders early in the section starting at line 253.

[18] Although this method is being presented as a tool for finding gene hits in mental disorders, I see potential for this method to be applied to any brain map. I recommend expanding on the different applications of this method in the Discussion.

[19] Very minor but there is something wrong with the grammatical structure of the sentence beginning at line 150 ("On the other hand, the number of hits in different top K genes…"). Please clarify.

Reviewer #2

(Remarks to the Author)
Revealing the molecular pathways underling mental health disorders is critical for its diagnosis and treatment. In this work, Cao et al. implemented a strategy for this task by integrating information from spatial gene expression, neuroimaging-based traits and neurotransmitter systems. Although approaches have been proposed to link gene expression and traits, this proposed method demonstrates its advance in bridging the gap from gene expression to disorders using PET images of neurotransmitter receptors/transporters. The authors also showed that the usage of PPI effectively improve the accuracy of their method for identifying disease-related genes.

This work is important, and the data analyses are comprehensive. I recommend publication if the following concerns are properly addressed:

Major concerns:
1. In both numeric simulation and real-world applications, the meanGPT and maxGPT approach demonstrated similar, if not worse, performance as PLS in several tasks (e.g. Figure 2a and Figure 3). Does it mean the inclusion of PET image alone is not helpful? On the other hand, approaches using PPI showed robust outperformance. Could the authors provide the rationale/evidence why PPI is important and how it benefits the approach?

2. Several points in Figure 2 need to be clarified.
1) The visibility of key elements in the boxplot of Figure 2a needs to be improved. The median value is hardly visible in any boxes. Additionally, why are the ranges of correlations so narrow for mFusion method but wide for PLS?
2) In Figure 2c, the AUC-ROC of maxPPI is 1, why is this possible? Could the authors provide details for generating the ROC curve?
3) In the simulation analyses, the authors used $\tilde{w}_X \times \tilde{w}_M^T$ to represent the PPI network. How much does this reflect the PPI in real world?

Minor concerns:
1. In Figure 2, 3, 5 and 6, the authors used PLS as the target for comparison. How was PLS performed? More details need to be provided.

2. The title for Figure 3, “Performance with different threshold for pruning the PPI network”, does not match its figures.

3. The authors claimed that Table S6-7 showed the outperformance of meanPPI. But these tables only provide the mFusion/PLS scores.

4. P-values and statistical tests need to be provided whenever necessary, for example:
“Among the ASD related genes in the DisGeNet database, the number of hit genes in the top K gene sets identified by the meanPPI method was also significantly greater than that identified by other five methods”.

5. What’s the y-axis of Figure 5 a-b?

6. In Figure 6 e-f, does the color shade represents Z-scores? If yes, why are the z-scores ranging from <(-15) to >15? What’s the meaning of shapes in front of gene names?

7. Website needs detailed user instructions.

8. The following sentences are broken:
1) “Complex interactions across multiple scales from genes, through neurotransmitters, to neural networks.”
2) “On the other hand, the number of hits in different top K genes (ranging from 41 to 1,541 genes; 1,541 is 10% of the total 15,408 genes) referring to knowledge databases (Figure 3 c-j; Table 1).”

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have appropriately addressed all my comments and I endorse the publication of this manuscript.

Reviewer #2

(Remarks to the Author)
The authors have adequately addressed the concerns, and I agree to its publication.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.
In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.
The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

Dear Editors and Reviewers:

Thank you for giving us this opportunity to submit a revised version of our manuscript **(ID: COMMSBIO-24-4155)**. Those comments are all valuable and very helpful for revising and improving this manuscript. We have studied comments carefully and have made correction which we hope meet with approval. Revised portion are marked with different colors in the paper. The main corrections in the paper and the responds to the reviewer's comments one by one are as flowing:

Reviewer 1:

We thank the Reviewer for endorsing our approach of considering the multiscale mechanisms in the brain to help in identifying potential pathological processes. We also thank the Reviewer for the insightful comments. By following these comments, we have now significantly improved the manuscript. Our reply to each comment and the corresponding revisions are provided in detail below.

[1] I did not fully understand the mFusion method and hope my points of confusion will help clarify the manuscript. Much of the confusion likely stems from unclear terminology. I am personally familiar with "weight" being the elements in the matrix U and V after the decomposition of the covariance matrix into USV (U are weights of the genes, S are singular values, and V are weights of the trait/neurotransmitter receptor maps). "Score" refers to the product of the original data matrix (e.g. G_{n, u}, where n is number of regions and u is number of genes) with the gene weights (U_{u, l}, where l is number of latent variable). "Latent variable" is the set of gene weights, singular value, and trait/receptor weight(s). It would be helpful if the authors could define the terminology and avoid using terms like "weight", "score", and "latent variable" if they do not mean the standard PLS definition of these terms. At least starting the PLS methodology with a definition of how PLS works and the terms the authors are using throughout would be very helpful in understanding mFusion.

--Thanks for this comment. Sorry we haven't made it clear. We have now clarified the PLS regression method and the definitions in the Methods and unified the terminology used throughout the manuscript.

Methods

"We examined three types of associations, including gene-trait, PET-trait, and gene-PET associations. Each type of association was separately estimated by partial least squares (PLS) regression, a widely employed method for evaluating the association between a 1-dimention response variable ($\boldsymbol{Y}_{n\times1}$) and a multidimensional predictor

($\boldsymbol{X}_{n\times p}$), where n is the sample size and p is the dimension of predictors[1-3]. Let L be the rank of matrix $\boldsymbol{X}$, the PLS regression iteratively computes L latent variables (*i.e.*, components) for $\boldsymbol{X}$ by singular value decomposition (SVD). Suppose the data matrices have already been normalized, if not, we normalize (*i.e.*, Z-score) them before entering the algorithm. At the initial step, the covariance matrix was decomposed by SVD as $\boldsymbol{X}^{\mathrm{T}}_{n\times p}\boldsymbol{Y}_{n\times 1} = \boldsymbol{U}_{p\times 1} \times s$, where $\boldsymbol{U}$ is an orthogonal vector and s is the corresponding singular value. The first component of was $\boldsymbol{X}$ formed as $\boldsymbol{t}_1 = \boldsymbol{X}_{n\times p} \times \boldsymbol{U}_{p\times 1}$. Therefore, $\boldsymbol{U}$ specifies the weight of each predictor for the first component (note as u_i, $i = 1,2,\cdots,n$). Next, data $\boldsymbol{X}$ is regressed on the first component $\boldsymbol{t}_1$ and the residuals $\widetilde{\boldsymbol{X}}_{n\times p}$ are used as the data for the next iteration. The second component is given by applying the SVD to $\widetilde{\boldsymbol{X}}^{\mathrm{T}}_{n\times p}\boldsymbol{Y}_{n\times 1}$. The iteration stops when L components are established as $(\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_L)$.

For each component ($\boldsymbol{t}_l, l = \mathbf{1}, \mathbf{2}, \cdots, L$) established from $\boldsymbol{X}$, the effect size of its association with $\boldsymbol{Y}$ can be estimated by the variance explained in the linear regression of $\boldsymbol{Y}$ on this component. The significance of this association was assessed by 1,000 permutations, *i.e.*, randomly shuffling the elements of $\boldsymbol{Y}$ to re-conduct the PLS regression and re-calculate the variance explained. Following the literature[1], the estimation error in each weight u_i is established empirically by 1,000 bootstraps as the standard error in the bootstrapped estimations (σ_i) and is adjusted for by u_i/σ_i. The error-adjusted weights entered the following analyses. For brevity, the error-adjusted weight was referred to as the Z-score in the following texts." **(Line 456~480)**

[2] In Line 401, the authors define weight of the ith gene as the z-score of bootstrapped coefficients. Why not use the weight of the ith gene as defined by the PLS decomposition in the first matrix U?

--Thank you for the comment. By following the literature (Vértes *et.al.* 2016), the estimation error (σ_i) in its weight u_i for the ith gene is empirically established by bootstraps, so as to account for the potentially sampling bias. After adjusting for the estimation error, the genes can then be ranked by the error-adjusted weight (u_i/σ_i). We apologize for not making this clear. To clarify this point, we have now revised the Methods section as below:

Methods

"Following the literature[1], the estimation error in each weight u_i is established empirically by 1,000 bootstraps as the standard error in the bootstrapped estimations

(σ_i) and is adjusted for by u_i/σ_i. The error-adjusted weights entered the following analyses." **(Line 476~480)**

Reference

Vértes, P. E. *et al.* Gene transcription profiles associated with inter-modular hubs and connection distance in human functional magnetic resonance imaging networks. *Philos Trans R Soc Lond B Biol Sci* **371**, doi:10.1098/rstb.2015.0362 (2016).

[3] Line 409: I was confused by the definition of Z_{g, p}(i, j) - what is the ith gene weight for the jth receptor map? Given my understanding of PLS (detailed in point 1), the ith gene weight would be U[i, 0] (for the first latent variable) and the jth neurotransmission weight would be V[0, j] so where is Z(i, j) coming from?

--We apologize for not making this clear. We used the PLS regression instead of the PLS correlation. We have now described this method carefully as detailed in our reply to your first comment. To make it clear, we have replaced the expression of 'correlation' with 'association'. For example, we changed 'multiscale correlations' into 'multiscale associations', and also used terms of 'gene-trait association', 'gene-PET association', 'PET-trait association'.

Particularly, for the gene-PET association, instead of calculating the PLS correlation between 45 PET maps and 15,408 genes, we conducted 45 different PLS regressions for each of the PET maps separately. Therefore, for the jth PET map, a PLS regression was used to establish the PLS-weights for the associations of this PET map with all genes, noted by $Z_{g,p}(i,j)$ ($i = 1,2,\cdots,15408$). When we completed all 45 PLS regressions, we established this PLS-weight matrix for the gene-PET associations as $Z_{g,p}(i,j)$ ($i = 1,2,\cdots,15408$; $j = 1,2,\cdots,45$). We have now clarified this definition in the Methods.

Methods

"The gene-PET associations were assessed by 45 PLS regressions separately for each of the 45 PET maps, and the regional gene expression matrix $\boldsymbol{X}_{n\times p}$ was the predictors. Therefore, for the j-th PET map ($j = 1,2,\cdots,v$), the normalized Z-score $Z_{g,p}(i,j)$ was used to measure the association of the i-th gene to this PET map ($i = 1,2,\cdots,n$)." **(Line 491~494)**

[4] I'm also confused by the terminology "neurotransmission-related neurotransmissions" (e.g. lines 407, 410).

--Thank you for your careful review. To make it clear, we have changed all these terms

to “Neurotransmission-related PET maps”.

[5] I recommend explicitly reporting where data came from, for example the name of the toolbox (e.g. neuromaps, enigmatoolbox) and link. I also strongly recommend citing the original study that collected this data. I noticed the authors provide links to papers for each receptor image but do not seem to cite these papers despite using this data.

--Thank you for this comment. We have now provided both the links and the citations for the toolboxes from which we had extracted the neuroimaging trait data, including the neurompas, enigma toolbox, and JuSpace. Now, we have included a new table (Table 2) in the main text to list all the PET maps used in the current study. In this table, we have listed the targeted protein, corresponding neurotransmitter, tracer, sample demographics, and citation. In Table S1, we have also provided the links to the data sets we used in this study.

Methods

“In this study, we curated a comprehensive database comprising 45 neurotransmission-related PET maps for 9 neurotransmitter systems and synaptic density. Among them, 36 maps were provided in the neuromaps toolbox (https://netneurolab.github.io/neuromaps/index.html)[4], 6 were available through the JuSpace toolbox (https://github.com/juryxy/JuSpace)[5], and 3 were available at the PET imaging database provided by Hansen et. al.[6] (https://github.com/netneurolab/hansen_receptors/tree/main/data/PET_nifti_images).” **(Line 402~408)**

“The ENIGMA consortium and ENIGMA toolbox (https://enigma-toolbox.readthedocs.io/en/latest/index.html#)[7] have provided the structural case-control differences for eight mental disorders,…” **(Line 426~428)**

“Table 2. Neurotransmission-related PET maps included in analyses. BP_{ND}: parametric and regional non-displaceable binding potential.; B_{max}: density (pmol ml^{-1}) converted from binding potential (5-HT) or distributional volume (GABA) using autoradiography-derived densities. V_T: tracer distribution volume. SUVR: standardized uptake value ratio. Values in parentheses (under n) indicate the number of females. The Protein column indicate the protein names in the STRING database. Supplementary Table S1 also includes more extensive methodological details, such as Excitatory/Inhibitory, Ionotropic/Metabotropic, and Source toolkit.” **(Line 993~1001)**

[6] Where is the PD enigma data coming from? It is not included in the enigmatoolbox.

--This data set has been published by the ENIGMA-Parkinson's Study (Lannsma, *et al*. Mov Disord. 2021). We have now clarified this point and cited the corresponding reference.

Methods

"The ENIGMA consortium and ENIGMA toolbox (https://enigma-toolbox.readthedocs.io/en/latest/index.html#)[7] have provided the structural case-control differences for eight mental disorders, …, Parkinson's disease (PD)[8], …" **(Line426~428)**

Reference:

Laansma, M. A. *et al.* International Multicenter Analysis of Brain Structure Across Clinical Stages of Parkinson's Disease. *Mov Disord* **36**, 2583-2594, doi:10.1002/mds.28706 (2021).

[7] In line 345 the authors mention "cerebral metabolic maps" - what are these? And why are they being referred to as neurotransmission images? Likewise, the list of neurotransmission images includes UCBJ, a tracer for SV2A which is not a neurotransmitter receptor or transporter.

--We apologize for this overlooking, that the PET map for cerebral metabolism was not used in the final analyses as they are not directly related to neurotransmission. We have now removed this mentioning.

Regarding the SV2A, it is a widely distributed synaptic vesicle membrane glycoprotein in the brain, is associated with the regulation of a variety of neurotransmitter systems (Finnema, *et.al*. Sci Transl Med. 2016). This including inhibitory neurotransmitter such as the release of GABA (gamma-aminobutyric acid) (Tokudome, *et. al*. Scientific Report. 2016), excitatory neurotransmitters such as glutamate (Rossi, *et. al*. Front Neurosci. 2022). Moreover, SV2s is associated with several kinds of nervous system diseases such as epilepsy and AD, also shows that it is very important to the normal function of the nervous system[9]. Therefore, SV2A is taken into consideration in this paper. We have now emphasized this point in Table 2.

Table 2

"*: The synaptic vesicle glycoprotein 2A(SV2A) is targeted by PET imaging to quantify synaptic density in human brains[10]." **(Line 1000~1001)**

Reference:

Finnema, S. J. *et al.* Imaging synaptic density in the living human brain. *Sci Transl Med* **8**, 348ra396, doi:10.1126/scitranslmed.aaf6667 (2016).

Tokudome, K. *et al.* Synaptic vesicle glycoprotein 2A (SV2A) regulates kindling epileptogenesis via GABAergic neurotransmission. *Scientific Reports* **6**, 27420,

doi:10.1038/srep27420 (2016).

Rossi, R., Arjmand, S., Bærentzen, S. L., Gjedde, A. & Landau, A. M. Synaptic Vesicle Glycoprotein 2A: Features and Functions. *Front Neurosci* **16**, 864514, doi:10.3389/fnins.2022.864514 (2022).

[8] Quality of gene expression and PET receptor data may make a big difference to the results. It is likely difficult to address how data quality impacts results but I at least recommend discussing this point in the limitations. For example, the authors include a lot of neurotransmitter receptor/transporter maps, including many that overlap in terms of protein that they bind to and even tracer. How reliable are these maps? Why not curate a more specialized list based on the quality of the image? Or at least compare images that are mapping the same protein to assess reliability of these images. For the gene expression data, the authors could filter out genes that have low reliability across donors (i.e. low differential stability).

--Thank you for your recommendation, the quality of gene expression data obtained from the AHBA is an important factor. We have described in detail the processing pipeline and the algorithm parameters in the supplementary Methods. To emphasize this point, we have now included one more limitation.

Limitations

"Second, the choice of processing parameters may influence the AHBA gene expression data[11]. To mitigate this challenge, we had normalized the expression values and focused only on analyses related to the relative rank of genes as opposed to the absolute values."
(Line 367~370)

--As we do not have the raw data for these PET images, it is difficult for us to evaluate their quality. However, as the maps have already been published in the literatures, their quality control procedures have been verified in these previous publications. Here, to enrich the amount of data, we used all the PET maps we could collected.

As you recommended, for PET receptor data that overlap in terms of protein, we calculated there spatial Pearson correlation at DK308 parcellation atlas (showed in **Supplementary Figure S12**), we found a significant positive association ($p<0.01$) between all overlapped maps except for one PET map of Dopamine Receptor. The results were consistent with a previous report (Hansen *et al*. 2022). These correlations have now been provided in the **Supplementary Figure S12**.

Reference

Hansen, J. Y. *et al.* Mapping neurotransmitter systems to the structural and functional organization of the human neocortex. *Nature Neuroscience* **25**, 1569-1581, doi:10.1038/s41593-022-01186-3 (2022).

We also averaged the PET maps for the same molecule to have 20 distinct maps, and reconducted the main analysis. We found that the gene scores given by the mFusion method were significantly correlated with those reported in the main text (spearman correlation for gene scores was r=0.97, p<2e-16, and r=0.98, p<2e-16, respectively). The verification analysis has now been reported in the **Methods, Results, and Figure S8-S9.**

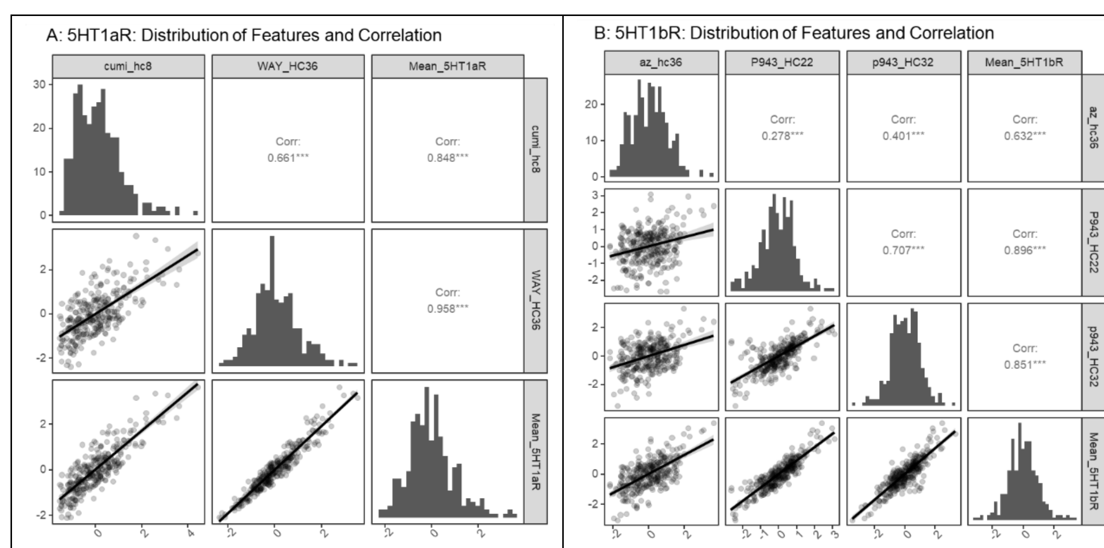
Methods

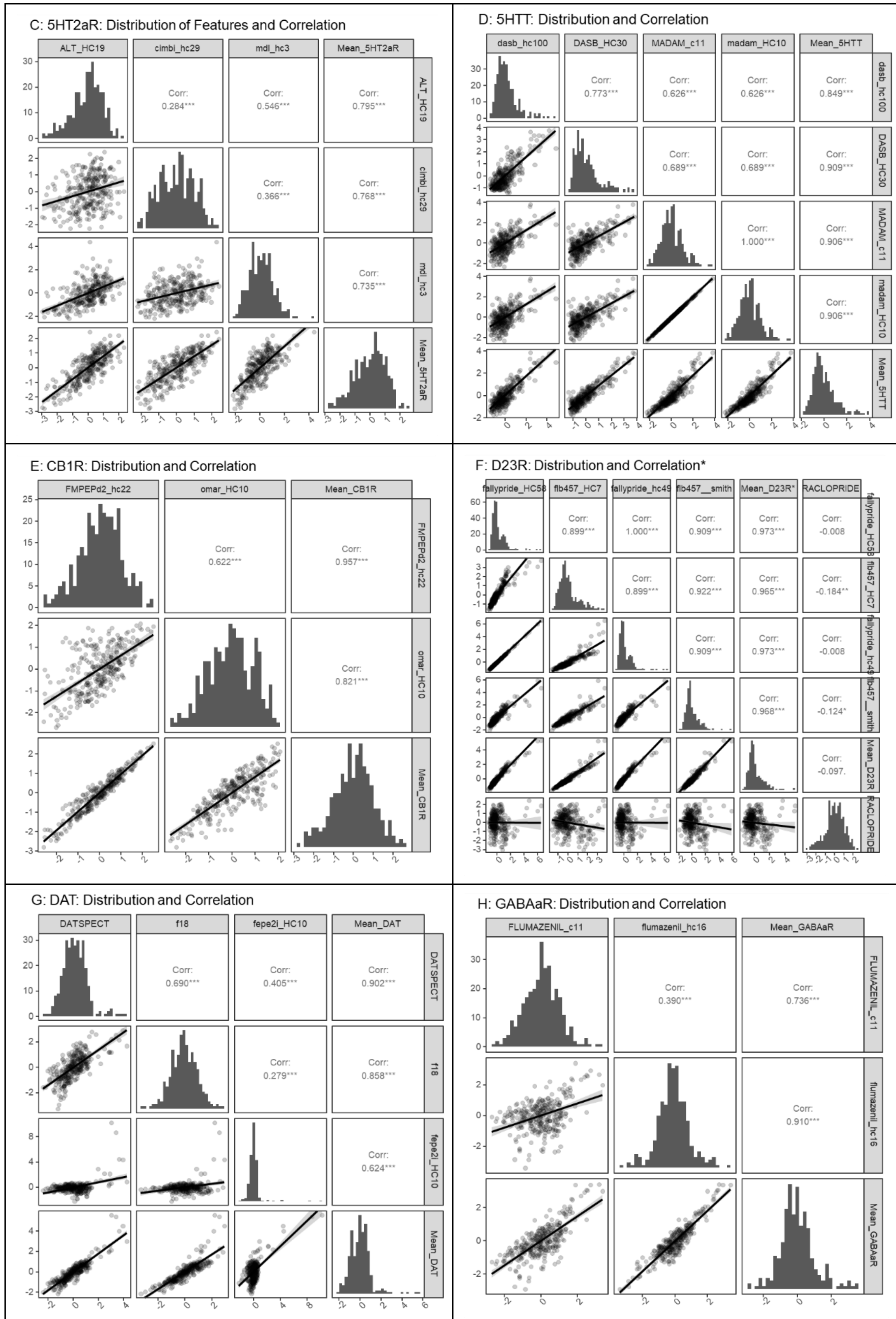
"Furthermore, we averaged the PET maps measuring the same molecule and got 20 PET maps each of which measured a different molecule (Figure S18). To test the potential bias introduced by the redundant PET maps for some molecules, we compared the gene scores given the mFusion using these 20 unique PET maps with the gene scores given by using all the 45 PET maps. If these scores are significantly correlated with each other, there is little evidence for the bias." **(Line 630~635)**

Results

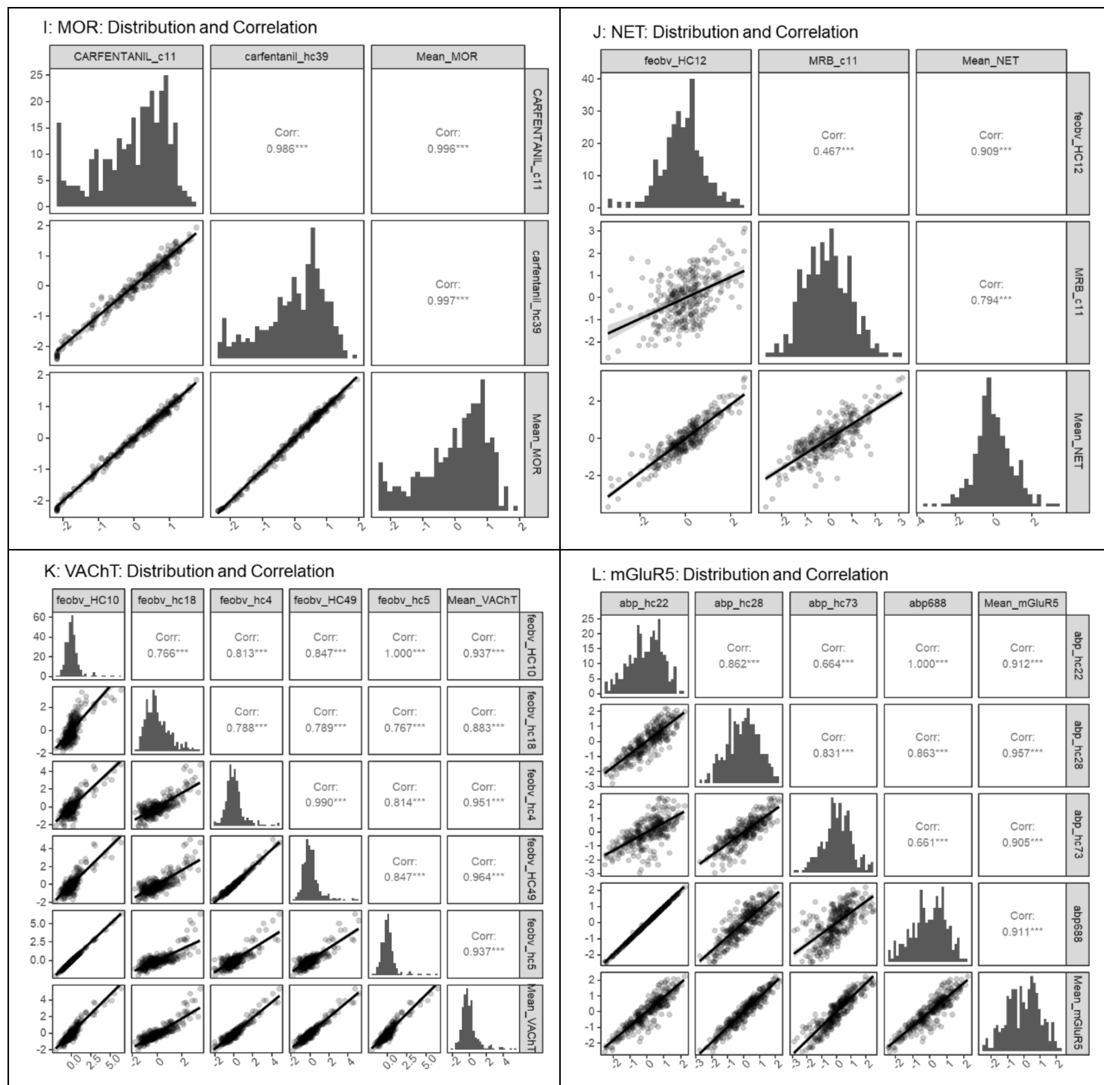
"To assess the effect of the quality of PET maps on the results, the 45 redundant maps were synthesized and averaged into 20 unique maps (Figure S8). Subsequently, the characteristics of SCZ and ASD were reanalyzed (Figures S9-S10). The meanPPI method demonstrated remarkable consistency with the primary findings regarding the identification of disease risk genes, exhibiting a spearman correlation for gene scores of r = 0.97 (p < 2e-16) and r = 0.98 (p < 2e-16), respectively (Figure S9). Furthermore, both the meanPPI and maxPPI methods emerged as the most effective approaches (Figure S10)." **(Line 213~220)**

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

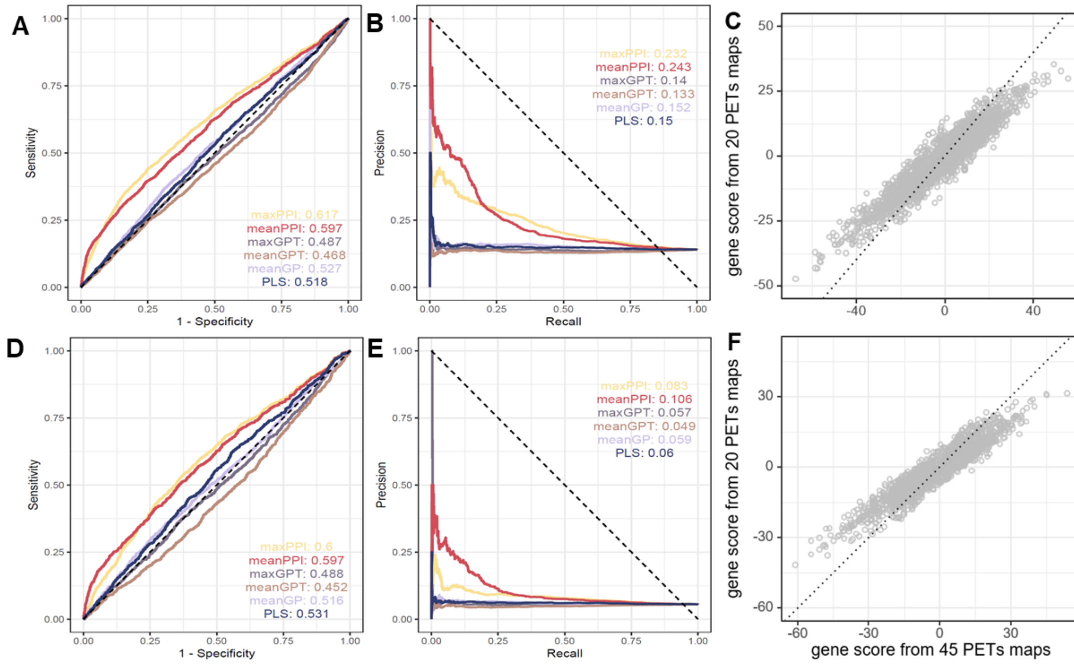




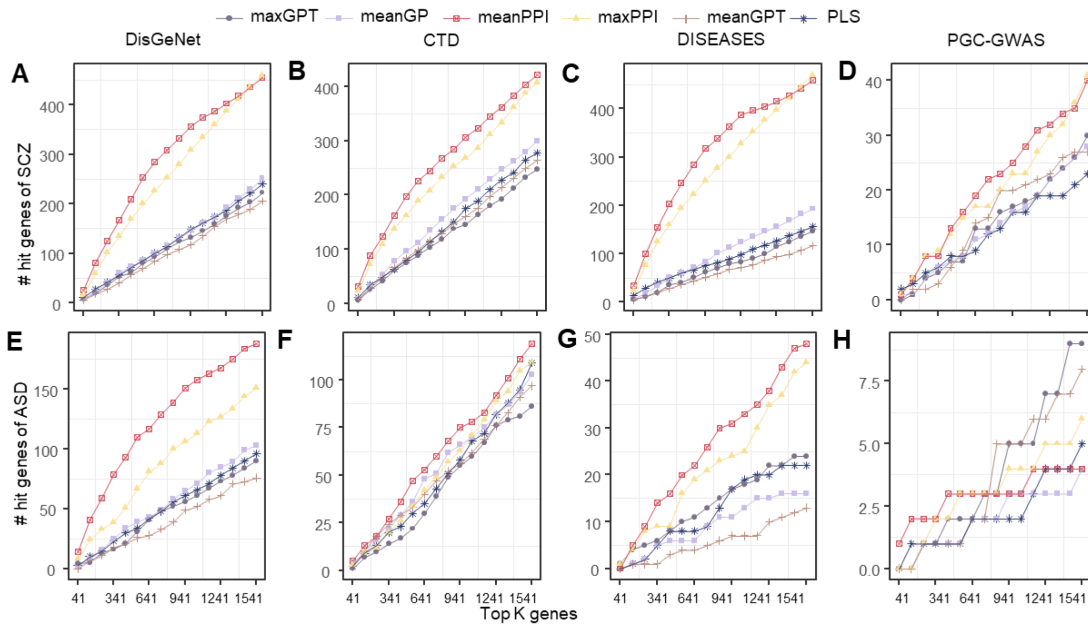
C: 5HT2aR: Distribution of Features and Correlation
ALT_HC19
cimbi_hc29
mdl_hc3
Mean_5HT2aR
Corr: 0.284***
Corr: 0.546***
Corr: 0.795***
Corr: 0.366***
Corr: 0.768***
Corr: 0.735***
D: 5HTT: Distribution and Correlation
dasb_hc100
DASB_HC30
MADAM_c11
madam_HC10
Mean_5HTT
Corr: 0.773***
Corr: 0.626***
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: 0.899***
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.
G: DAT: Distribution and Correlation
DATSPECT
f18
fepe2l_HC10
Mean_DAT
Corr: 0.690***
Corr: 0.405***
Corr: 0.902***
Corr: 0.279***
Corr: 0.858***
Corr: 0.624***
H: GABAaR: Distribution and Correlation
FLUMAZENIL_c11
flumazenil_hc16
Mean_GABAaR
Corr: 0.390***
Corr: 0.736***
Corr: 0.910***



Supplementary 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.



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



[9] I thought the inclusion of the PPI networks was interesting and was happy to see the superior performance of PPI mFusion methods. I wonder if it's possible to benchmark their performance against a null, to be convinced that it is the biological relationship between gene and protein, as well as their specific topographic expression patterns across the brain, that result in superior performance. For example, are the number of empirical hits significantly greater when using the empirical PPI network

than when using a random PPI network with equal density?

--Thank you for your valuable advice. We curated a null distribution of PPI by permuting the nodes of STRING (v11.5) PPI 500 times, and re-analyzed SCZ and ASD disease by mFusion using these PPIs. We found that the true PPI information significantly enhanced the ability of the mFusion-meanPPI method to find disease-associated genes. We have now included these details in the Methods, Results, and Supplementary Figure S7 A-B.

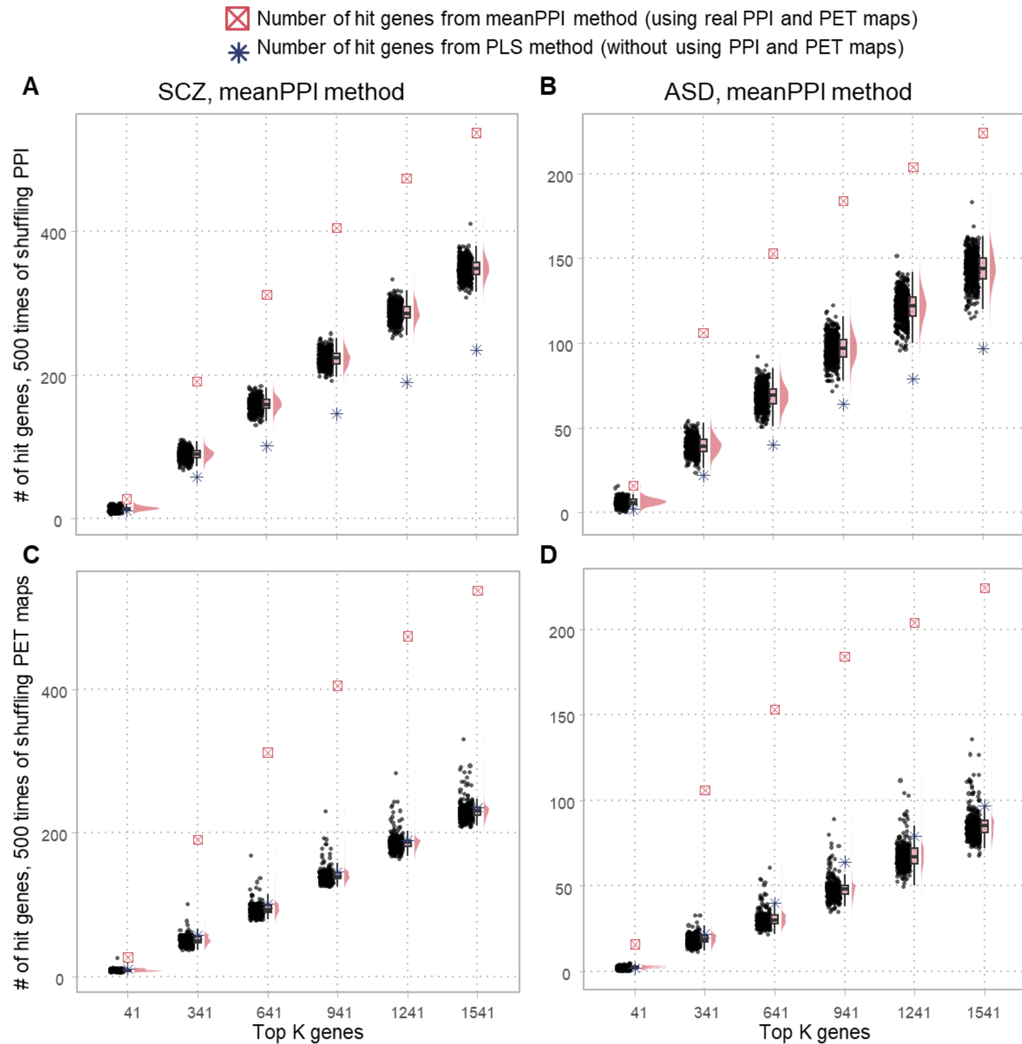
Methods

"To assess the influence of PPI information on the mFusion method, 500 times of shuffling were conducted on the nodes within the STRING PPI networks. This type of shift preserved the network properties of the original PPI and the characteristics of the node itself, while modifying the neighbor relationship of each node. The null distribution of results in random PPI networks was obtained and compared with results from the real STRING PPI networks." **(Line 619~624)**

Results

"In order to evaluate the importance of PPIs in the context of the mFusion-meanPPI methodology, a comparative analysis was conducted on SCZ and ASD phenotypes separately. The analysis comprised a computational evaluation of 500 randomly generated PPIs for each disease, with the resulting null distribution of the number of hit genes presented in Figure S7. The results demonstrated that the application of the meanPPI method using real PPI data markedly augmented the capacity to identify hit genes compared to the use of random PPI." **(line 201~207)**

Supplementary 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. ."



[10] To continue on the point above: if the PLS is run with a randomly shuffled (or rotated, to preserve spatial autocorrelation) gene or receptor matrix, do the number of gene hits significantly decrease?

--Per your suggestion, we have now randomly shuffled each of the PET maps for 500 times, and indeed found that the number of gene hits significantly decreased. We found that these results significantly strengthened our findings. We have now updated the Methods, Results and included new Supplementary Figure S7 C-D.

Methods

"Similarly, we permutated the 45 PET maps 500 times (by shuffling the order of brain regions across 45 maps concurrently at each time) to evaluate their contributions to the mFusion methodology. If the number of hit genes significantly reduced, the molecular distributions given by the real PET maps contributed significantly to mFusion." **(Line 624~628)**

Results

"In addition, a similar permutation was made for the 45 PET maps (see Methods) and reapplied to the analysis of the SCZ and ASD disease. The results in Figure S7 C-D revealed a marked reduction in the ability of the meanPPI method in pinpointing disease-associated genes, thereby indicating that real PET maps are pivotal in the meanPPI method." **(Line 207~211)**

[11] Please list the seven morphological parameters referenced in line 373.

--We have now listed these parameters as the following:

Methods

"In our investigation, we incorporated a brain map depicting case-control differences in morphological similarity, specifically the correlation of seven morphological parameters (*i.e.*, the gray matter volume, surface area, cortical thickness, Gaussian curvature, mean curvature, fractional anisotropy, and mean diffusivity) derived from MRI and diffusion-weighted imaging data, concerning schizophrenia." **(Line 437~441)**

[12] In regards to the following sentence (line 399): "We selected the first PLS component if it had a p-value < 0.05 by permutation; otherwise, the component that explained the most variance of Y." For the gene-trait and neurotransmission-trait PLSs, there is only one latent variable, which makes this sentence confusing. Perhaps stating this sentence with respect to the gene-neurotransmission PLS only? How common are non-significant latent variables in the empirical analyses?

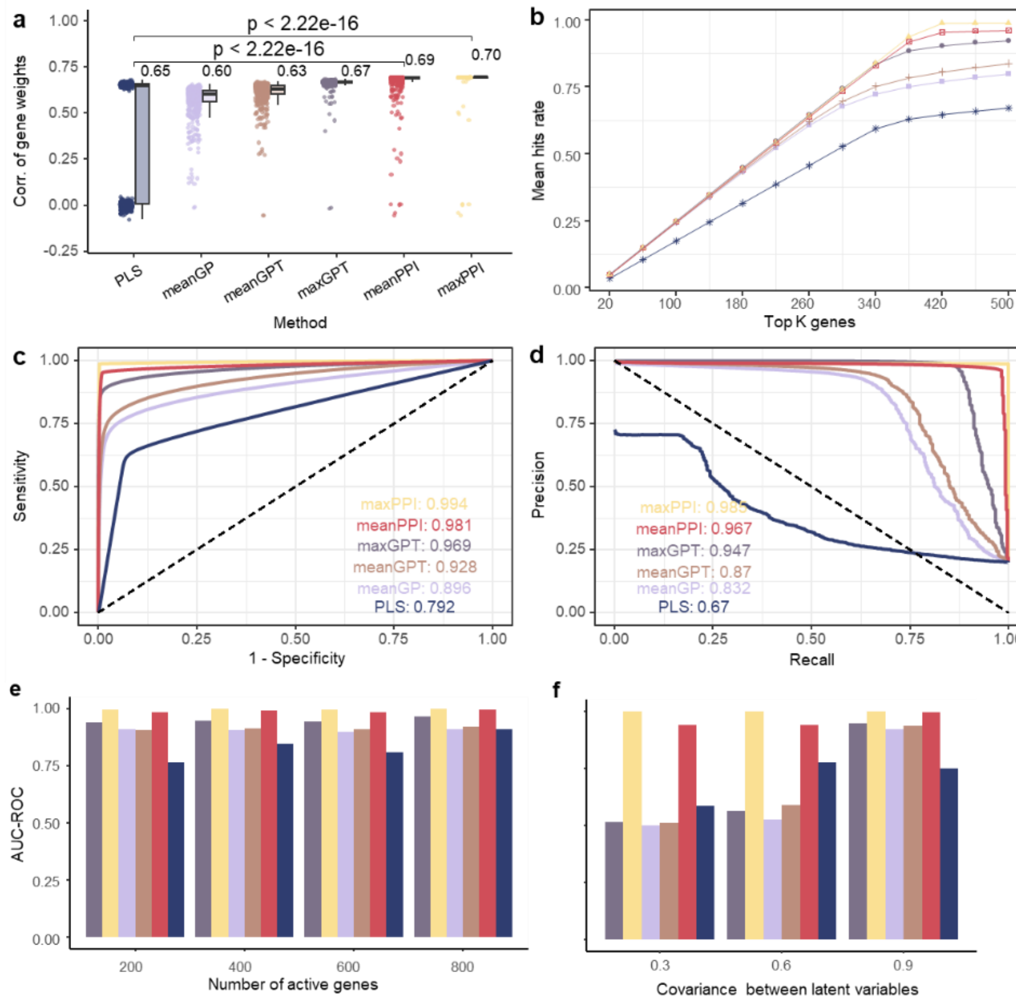
--We apologize for not making this clear. In our PLS regressions, the outcome was always 1-dim, while the predictors had multiple dimensions. Therefore, the PLS regressions could give many components of predictors. We needed to test the significance of these correlations between these components and the outcome variable by permutations. We have now clarified the PLS regression method as detailed in our reply to your first comment.

Empirically, in the current study, the PLS regression could not identify significant components for three (*i.e.*, epilepsy, OCD, and ASD) out of eight diseases. Therefore, in these cases, we used the components explaining the most variance in the outcome variables.

[13] I recommend reporting the median correlation between real and fusion gene weights in each bin such that readers know by how much the different fusion methods improve the median correlation. (Fig 2a)

--Thank you for your advice. We have now reported these medians in the **Figure 2a.**

Line 1018: "**Figure 2. Evaluation of fusion methods from simulated datasets.** a: The correlation between real gene weights and fusion weights measured by different fusion methods of 500 simulated experiments. The number next to bar represents the median of the population (using unpaired Wilcoxon test)."



[14] In the section starting at line 207, the authors compare how many gene hits are found using the meanPPI method vs regular PLS. Could the authors also compare the average weight of these gene hits? I could imagine a possibility where the PLS discovers fewer genes but these genes are all at the very top of the list, whereas with meanPPI it may be more difficult to determine which genes are relevant and which are not, if no a priori information on relevant genes is known.

--Yes, the genes ranked first by the regular PLS regression had higher PLS-regression weights as shown in Figure 6 a-b. And we newly clarified this meaning at manuscript: "**a-b:** Gene scores from meanPPI method and PLS method." **(Line 1061)**

However, the higher PLS-regression weight alone identified many genes that could not be found in the reference gene sets for SCZ and ASD. In another words, there were

many false positives, since the PLS-regression is essentially a multivariate approach which is prone to overfitting. Using the meanPPI, we had successfully reduced the false positives, as the hit rate was higher in genes with high meanPPI scores than that in genes with high PLS-regression weights.

Results

"By comparing gene scores with disease-related genes listed in the DisGeNet database, we observed that higher meanPPI fusion scores were associated with higher hit rates. Since the PLS-regression is essentially a multivariate approach, which is prone to overfitting, we found more false positives in the genes with high PLS-regression weights. In contrast, we demonstrated that the mFusion-meanPPI approach reduced the false positive rate by combining the information from multiscale." **(Line 252~258)**

[15] There were some parts of the Results I felt were better left for the Discussion, including the text starting at line 221 and the entire section at line 241.

--Thank you for your suggestion. We have now moved these parts to the discussion section between **line331~360.**

[16] Could the authors discuss how the very different spatial resolutions of microarray-derived gene expression and PET-derived receptor densities may impact the mFusion method? This could also be tested in the simulation by making the simulated receptor data smoother.

--Thank you for this insightful comment. We have now conducted the simulation experiment per your suggestion. When we changed the number of brain regions ***n*** to100, 200, or 500, the simulation results showed that the higher the spatial resolution was, the better the mFusion performed. We have now reported these simulations as follows:

Methods

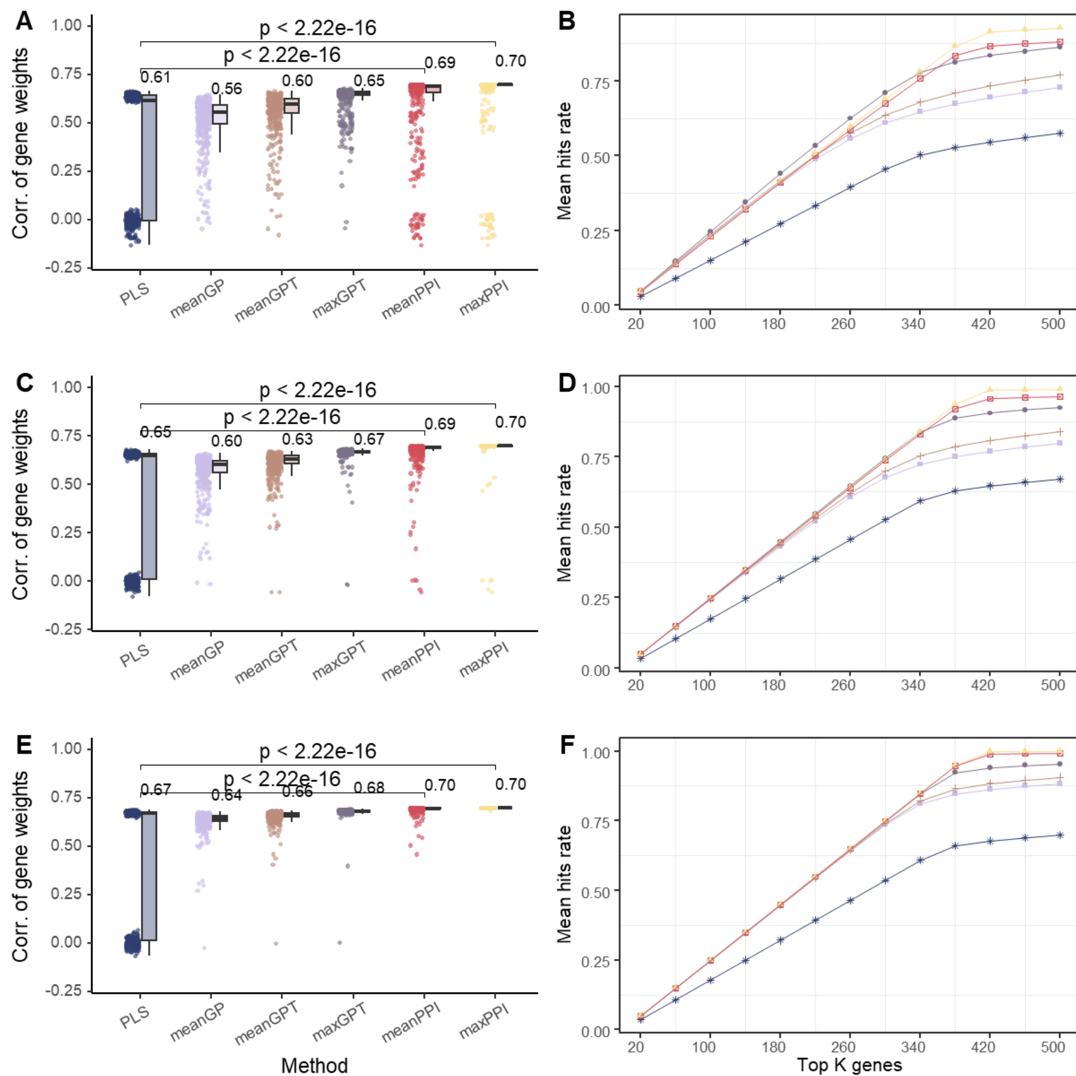
"Thirdly, we changed the dimension of ***X, Y,*** and ***Z*** matrix (i.e., the brain regions *n* ranges in 100, 200, or 500) to investigate the influence of varying spatial resolutions of microarray-derived gene expression and PET-derived receptor maps to the mFusion method." **(Line 573~576)**

Results

"Thirdly, we conducted a simulation of brain maps at three distinct spatial resolutions. Specifically, the number of brain regions (*n*) was varied between 100, 200, and 500 (see Methods for further details), as delineated in Figure S3. The results of this simulation demonstrated a positive correlation between the spatial resolution of the ***X, Y***, and ***Z*** matrices and the efficacy of the methods in identifying activated genes. Notably, the meanPPI and maxPPI methodologies consistently exhibited superior performance compared to other methods, exhibiting a level of stability that highlights their

robustness in high-resolution brain mapping analyses." **(Line 152~159)**

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.



[17] I recommend explaining how you are clustering disorders early in the section starting at line 253.

--Thanks for your recommendation, and we added more explanation on the clustering

process at this section.

Results

"We applied the mFusion-meanPPI algorithm to neuroimaging traits of eight disorder cohorts separately (Figure 7a, see Methods section), and prioritized top 10% genes based on their Z-scores. Spearman correlation analysis of these genes was performed to assess the similarity between each pair of disorders. Following this, hierarchical clustering was applied to the spearman correlation coefficients among these diseases, resulting in the identification of three distinct clusters. These clusters reflected the expressional association among these diseases, as inferred from the gene Z-scores." **(Line 280~286)**

[18] Although this method is being presented as a tool for finding gene hits in mental disorders, I see potential for this method to be applied to any brain map. I recommend expanding on the different applications of this method in the Discussion.

--Thanks for you this suggestion. To emphasize this point, we have now revised the Discussion as the following:

Discussion

"In the current study, we demonstrated the performance of the proposed mFusion as a tool for finding gene hits in mental disorders using the PET maps, it is worth noting that the method could be applied to any brain maps, such as the functional MRI or magnetoencephalography, single-photon emission computed tomography, *etc*." **(Line 325~329)**

[19] Very minor but there is something wrong with the grammatical structure of the sentence beginning at line 150 ("On the other hand, the number of hits in different top K genes…"). Please clarify.

--Thank you for your careful review. It has now been corrected.

Results

"On the other hand, we compared the number of hits in the top K genes given by various methods. When we varied the parameter K from 41 to 1541, where 1541 was 10% of the total of 15,408 genes, we found that the proposed methods had consistently more hits as compared with the other algorithms (Figure 3 c-j)." **(Line 167~170)**

Reviewer 2:

Thank you for endorsing our work as both important and comprehensive. By following your comments, we have significantly improved the manuscripts. Our reply to each of the comment is detailed below.

1. In both numeric simulation and real-world applications, the meanGPT and maxGPT approach demonstrated similar, if not worse, performance as PLS in several tasks (e.g. Figure 2a and Figure 3). Does it mean the inclusion of PET image alone is not helpful?On the other hand, approaches using PPI showed robust outperformance. Could the authors provide the rationale/evidence why PPI is important and how it benefits the approach?

--Thank you for the comment. The reason might be that the use of PPI in the meanPPI method works as a filter of noise. In the GPT approach, all PET-molecules were considered. However, the PLS regression, a multivariate approach, is prone to over-fitting. Therefore, significant noises were introduced by those over-fitted PET-gene associations. In contrast, by restricting our considerations within those PET-molecules that had PPI-connectivity with the genes, the PPI approaches significantly reduced such noises. To demonstrate this point, we have now permuted the PPI network for 500 times and found that the PPI indeed significantly improved the algorithm performances.

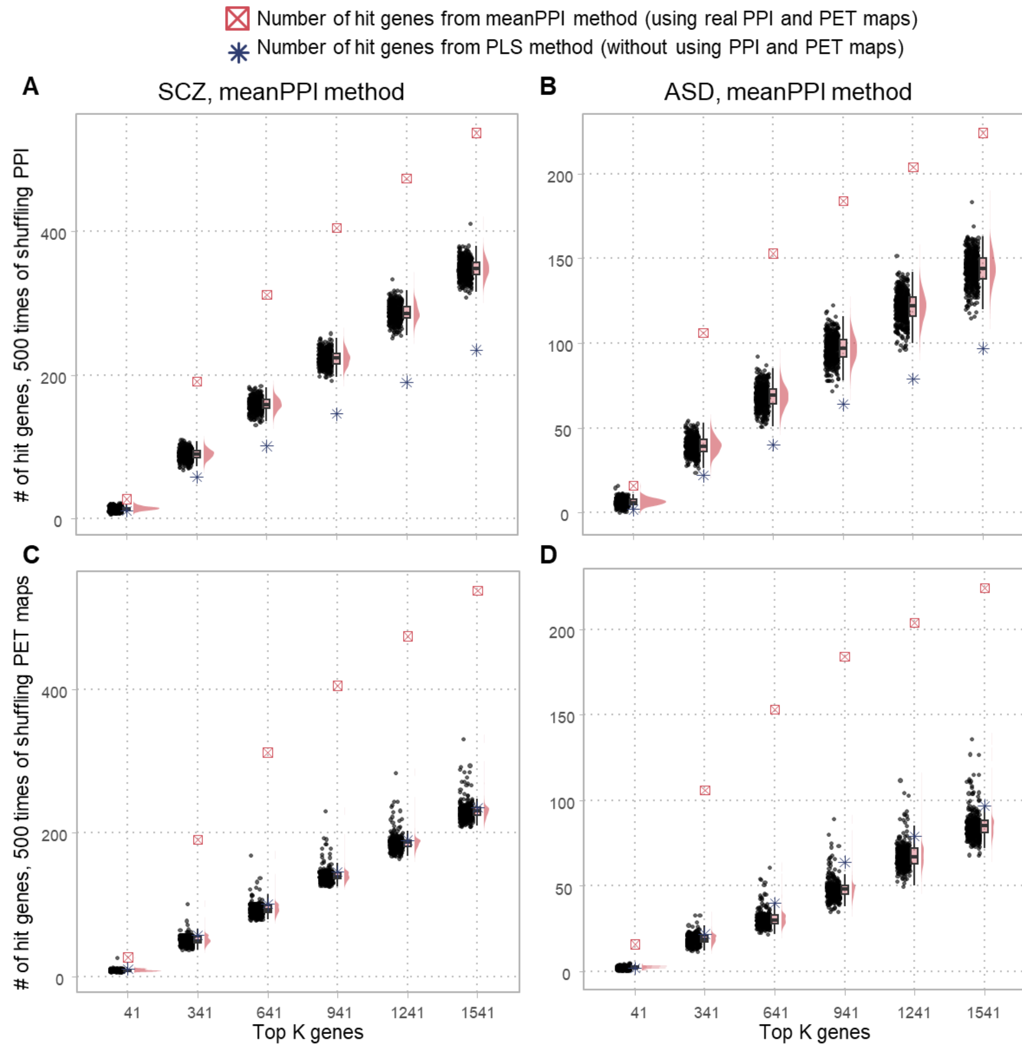
Methods

"To assess the influence of PPI information on the mFusion method, 500 times of shuffling were conducted on the nodes within the STRING PPI networks. This type of shift preserved the network properties of the original PPI and the characteristics of the node itself, while modifying the neighbor relationship of each node. The null distribution of results in random PPI networks was obtained and compared with results from the real STRING PPI networks." **(Line 619~624)**

Results

"In order to evaluate the importance of PPIs in the context of the mFusion-meanPPI methodology, a comparative analysis was conducted on SCZ and ASD phenotypes separately. The analysis comprised a computational evaluation of 500 randomly generated PPIs for each disease, with the resulting null distribution of the number of hit genes presented in Figure S7. The results demonstrated that the application of the meanPPI method using real PPI data markedly augmented the capacity to identify hit genes compared to the use of random PPI." **(Line 201~207)**

Supplementary 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 genes."



2. Several points in Figure 2 need to be clarified.
1) The visibility of key elements in the boxplot of Figure 2a needs to be improved. The median value is hardly visible in any boxes. Additionally, why are the ranges of correlations so narrow for mFusion method but wide for PLS?

--Thanks for your advice. We have annotated the median in Figure2a, and added captions for this.

The correlation bar range of PLS is much larger than other methods, the reason is that the PLS method simply considers the correlation between variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, and is prone to overfitting and giving data-sensitive results. However, other methods have stable performance (i.e., the ranges of correlations are narrow) in all simulation experiments, because the influence of variable $\boldsymbol{M}$ on $\boldsymbol{Y}$ has also been taken into consideration in the mFusion framework.

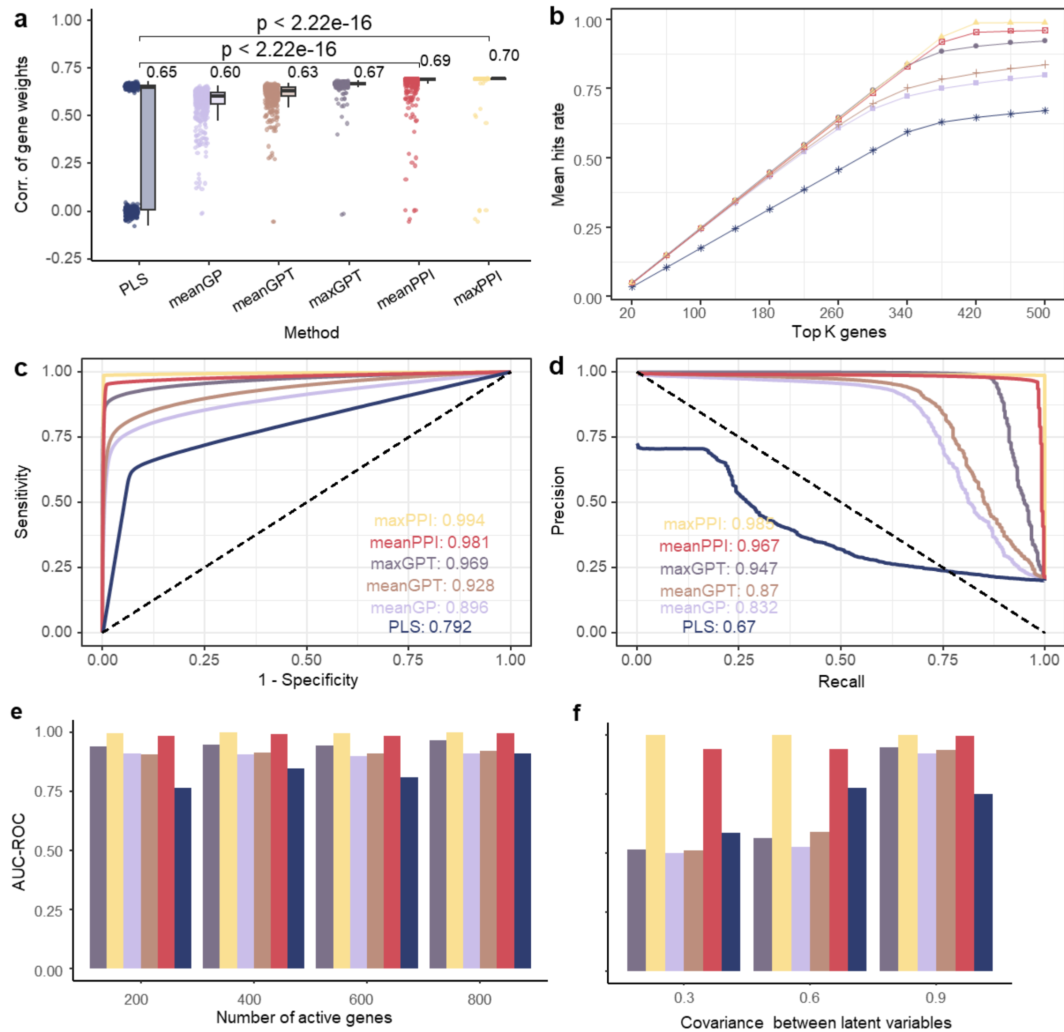


Figure 2. Evaluation of fusion methods from simulated datasets. a: The correlation between real gene weights and fusion weights measured by different fusion methods of 500 simulated experiments. The number next to bar represents the median of the population (using unpaired Wilcoxon test). **b:** Average 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**: ROC (Receiver Operating Characteristic) curve of different fusion methods on simulation data. **d**: PR (precision-recall) curve of different fusion methods on simulation data. **e**: AUC-ROC value of different fusion method when number of active genes changed. **f**: AUC-ROC value of different fusion method when covariance between latent variables changed. **(Line 1015~1026)**

2) In Figure 2c, the AUC-ROC of maxPPI is 1, why is this possible? Could the authors provide details for generating the ROC curve?

--Thanks for your comment on this point. In simulation experiments, $\tilde{\boldsymbol{w}}_{\boldsymbol{X}} \times \tilde{\boldsymbol{w}}_{\boldsymbol{M}}^{\boldsymbol{T}}$ is a completely accurate connection matrix, providing the perfect PPI information that greatly improves the performances of maxPPI and meanPPI methods, so the AUC-ROC

of maxPPI is 1. We add this description into the caption of Figure 2c. **(Line 1021~1023)**

About the generation of the ROC curve: At each simulation, the first ***K*** genes with the highest score (the ***K*** was predefined) were predicted as active genes while others were deemed as inactive. So, the predictive question was seemed as a binary classification task, and the true positive rate and false positive rate was computed at each test. In this way, we generated the mean AUC-ROC curve of total 500 simulation results. We newly clarified this process at the manuscript:

Methods:
At each experiment, the first ***K*** genes with the highest score (the ***K*** was predefined) were predicted as active genes while others were deemed as inactive. A larger AUC-ROC signifies a more robust classification ability of the model. We used R package "multiROC" to compute the mean value of true positive and false positive genes at all simulated experiment. **(Line 585~589)**

3) In the simulation analyses, the authors used w˜_X×w˜_M^T to represent the PPI network. How much does this reflect the PPI in real world?

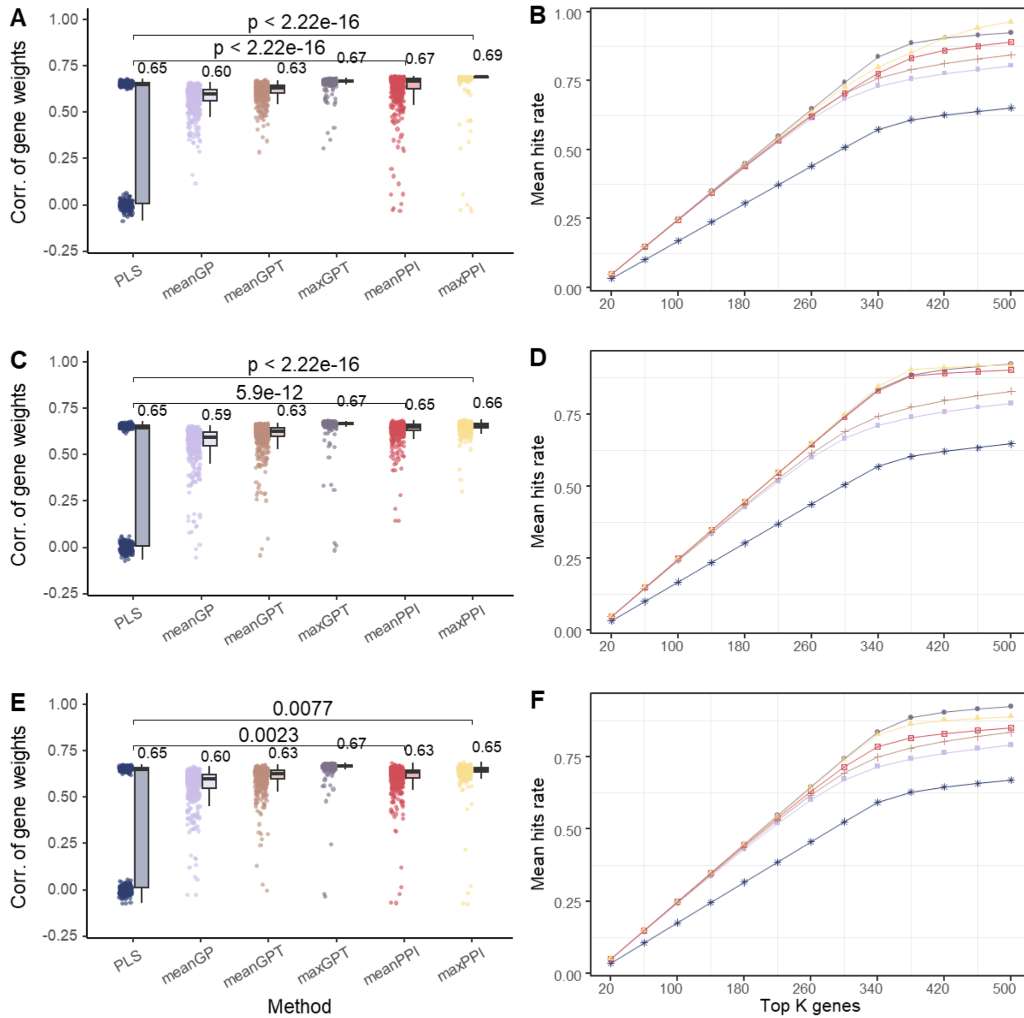
--Thank you for this comment. In the current framework, the PPI is mainly used as a filter for noise. Therefore, in the simulation we did not seek for preserving the network properties of the real PPI. However, inspired by your comment, we realized that there was a difference between the real PPI and the simulated PPI, as the simulated PPI is completed recorded but the real PPI is only partially observed. Therefore, we conducted an additional simulation experiment by using different proportions of the true PPI information. We confirmed that the maxPPI and meanPPI methods performed consistently better than the other algorithms in these conditions. We have now reported these results as follows:

Methods
"Furthermore, to evaluate the impact of Protein-Protein Interaction (PPI) networks on the mFusion method, we conducted a series of perturbations on the PPI network. Firstly, 30% of the elements within the adjacency matrix $\tilde{\boldsymbol{w}}_{\boldsymbol{X}} \times \tilde{\boldsymbol{w}}_{\boldsymbol{M}}^{\boldsymbol{T}}$ were randomly shuffled, thereby introducing a degree of artificiality to the PPI, akin to the inherent noise present in real-world PPI networks. Secondly, we set a minimum 30% of the elements in the adjacency matrix to zero, thereby simulating the incompleteness of PPI that is often encountered in biological systems. Thirdly, we changed the dimension of ***X, Y,*** and ***Z*** matrix (i.e., the brain regions n ranges in 100, 200, or 500) to investigate the influence of varying spatial resolutions of microarray-derived gene expression and PET maps to the mFusion method. In conclusion, we executed 500 repetitions of the mFusion method on these perturbed PPI matrices to evaluate the robustness and sensitivity to such perturbations." **(Line 567~578)**

Results

"And then, three kinds of perturbations were performed on the PPI networks to illustrated the influence of PPI information on the mFusion method for 500 repetitions: (1) randomly shuffle 30% of the elements within the adjacency matrix $\widetilde{\boldsymbol{w}}_{\boldsymbol{X}} \times \widetilde{\boldsymbol{w}}_{\boldsymbol{M}}^{\boldsymbol{T}}$; (2) set the minimum 30% of the elements in the adjacency matrix to be zero; (3) randomly shuffle 30% of the elements, and then set the minimum 30% of the elements in the adjacency matrix to be zero. We found that the meanPPI and maxPPI methods consistently outperformed their counterparts in all three conditions (Figure S2)." **(Line 144~150)**

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 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.



Minor concerns:

1. In Figure 2, 3, 5 and 6, the authors used PLS as the target for comparison. How was PLS performed? More details need to be provided.

--Thanks for your advice. We extended the explanation of PLS regression analysis in the **Methods**.

Methods

"We examined three types of associations, including gene-trait, PET-trait, and gene-PET associations. Each type of association was separately estimated by partial least squares (PLS) regression, a widely employed method for evaluating the association between a 1-dimention response variable ($\boldsymbol{Y}_{n\times 1}$) and a multidimensional predictor ($\boldsymbol{X}_{n\times p}$), where n is the sample size and p is the dimension of predictors[1-3]. Let L be the rank of matrix $\boldsymbol{X}$, the PLS regression iteratively computes L latent variables (*i.e.*, components) for $\boldsymbol{X}$ by singular value decomposition (SVD). Suppose the data matrices

have already been normalized, if not, we normalize (*i.e.*, Z-score) them before entering the algorithm. At the initial step, the covariance matrix was decomposed by SVD as $\boldsymbol{X}_{n\times p}^{\mathrm{T}}\boldsymbol{Y}_{n\times 1} = \boldsymbol{U}_{p\times 1} \times s$, where $\boldsymbol{U}$ is an orthogonal vector and s is the corresponding singular value. The first component of was $\boldsymbol{X}$ formed as $\boldsymbol{t}_1 = \boldsymbol{X}_{n\times p} \times \boldsymbol{U}_{p\times 1}$. Therefore, $\boldsymbol{U}$ specifies the weight of each predictor for the first component (note as u_i, $i = 1,2,\cdots,n$). Next, data $\boldsymbol{X}$ is regressed on the first component $\boldsymbol{t}_1$ and the residuals $\widetilde{\boldsymbol{X}}_{n\times p}$ are used as the data for the next iteration. The second component is given by applying the SVD to $\widetilde{\boldsymbol{X}}_{n\times p}^{\mathrm{T}}\boldsymbol{Y}_{n\times 1}$. The iteration stops when L components are established as $(\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_L)$.

For each component $(\boldsymbol{t}_l, l = \mathbf{1}, \mathbf{2}, \cdots, L)$ established from $\boldsymbol{X}$, the effect size of its association with $\boldsymbol{Y}$ can be estimated by the variance explained in the linear regression of $\boldsymbol{Y}$ on this component. The significance of this association was assessed by 1,000 permutations, *i.e.*, randomly shuffling the elements of $\boldsymbol{Y}$ to re-conduct the PLS regression and re-calculate the variance explained. Following the literature[1], the estimation error in each weight u_i is established empirically by 1,000 bootstraps as the standard error in the bootstrapped estimations (σ_i) and is adjusted for by u_i/σ_i. The error-adjusted weights entered the following analyses. For brevity, the error-adjusted weight was referred to as the Z-score in the following texts." **(Line 456~480)**

2. The title for Figure 3, "Performance with different threshold for pruning the PPI network", does not match its figures.

--Thanks for your careful review, we apologized for this typo and have corrected the title of Figure 3 as: "Figure 3. Performance on SCZ and ASD disease of fusion methods under different disease databases." **(Line 1030)**

3. The authors claimed that Table S6-7 showed the outperformance of meanPPI. But these tables only provide the mFusion/PLS scores.

--Thanks for reminding. We apologized for not making this clear. The meaning here is the Figure 3c-j showed the outperformance of meanPPI, TableS6 is the dependent data of Figure 3c-g, TableS7 is the dependent data of Figure 3g-j. We have clarified this meaning in manuscript.

"On the other hand, we compared the number of hits in the top K genes given by various methods. When we varied the parameter K from 41 to 1541, where 1541 was 10% of the total of 15,408 genes, we found that the proposed methods had consistently more hits as compared with the other algorithms (Figure 3 c-j)." **(Line 167~170)**

4. P-values and statistical tests need to be provided whenever necessary, for example: "Among the ASD related genes in the DisGeNet database, the number of hit genes in the top K gene sets identified by the meanPPI method was also significantly greater than that identified by other five methods".

--Thank you for your advice. We have checked and clarified P-values and statistical tests in this manuscript.

Line 131: "Figure 2a, unpaired Wilcoxon test, 500 times of simulations"

Line 172: "Figure 3c; *p-value* < 0.001, paired Wilcoxon test for meanPPI and PLS method."

Line 176: "Figure 3g; *p-value* < 0.001, paired Wilcoxon test for meanPPI and PLS method."

5. What's the y-axis of Figure 5 a-b?

--Sorry we haven't made it clear. We append caption of Figure 5: "The Y-axis lists disease with categories in alphabetical order." **(Line 1048)**

6. In Figure 6 e-f, does the color shade represents Z-scores? If yes, why are the z-scores ranging from <(-15) to >15? What's the meaning of shapes in front of gene names?

--Thank you for your comments. The values and colors here represent the Z-scores of PLS performed by gene-PET pairs. And the z-scores range from -18.39~19.04. For the convenience of visualization. The legend has been modified into -20.0~20.0.

Point shapes of genes in (e)-(f) means the same as in (a)-(b). We added texts to make it clear **at line 1069:** "Point shapes of genes in (e)-(f) means the same as in (a)-(b)."

7. Website needs detailed user instructions.

--Thank you for your recommendations. To make the mFusion method convenient to users, we've uploaded the toolkit source code on GitHub, and written the key pieces for using this toolkit. The GitHub website also contains the "Use instructions and parameter selection", "Explanation of the output files", and "Posthoc enrichment analysis with R code".

8. The following sentences are broken:

1) "Complex interactions across multiple scales from genes, through neurotransmitters, to neural networks."

--Thank you for your review. We have changed this sentence into: "There exists many complex interactions across multiple scales from genes, through neurotransmitters, to neural networks." **(Line 56)**

2) "On the other hand, the number of hits in different top K genes (ranging from 41 to 1,541 genes; 1,541 is 10% of the total 15,408 genes) referring to knowledge databases (Figure 3 c-j; Table 1)."

--Thank you for your review. We have changed this sentence into: "On the other hand, we compared the hit number of genes of various methods when taking different top K genes (K ranged from 41 to 1541 genes; 1541 is 10% of the total of 15,408 genes). Figure 3 c-j (corresponding to different reference knowledge bases in Table 1) shows that the two methods have higher hit numbers." **(Line 167~170)**

Reference used in manuscript:

1 Vértes, P. E. *et al.* Gene transcription profiles associated with inter-modular hubs and connection distance in human functional magnetic resonance imaging networks. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **371**, doi:10.1098/rstb.2015.0362 (2016).

2 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).

3 Abdi, H. Partial least squares regression and projection on latent structure regression (PLS Regression). *WIREs Computational Statistics* **2**, 97-106, doi:https://doi.org/10.1002/wics.51 (2010).

4 Markello, R. D. *et al.* neuromaps: structural and functional interpretation of brain maps. *Nature Methods* **19**, 1472-1479, doi:10.1038/s41592-022-01625-w (2022).

5 Dukart, J. *et al.* JuSpace: A tool for spatial correlation analyses of magnetic resonance imaging data with nuclear imaging derived neurotransmitter maps. *Human Brain Mapping* **42**, 555-566, doi:https://doi.org/10.1002/hbm.25244 (2021).

6 Hansen, J. Y. *et al.* Mapping neurotransmitter systems to the structural and functional organization of the human neocortex. *Nature Neuroscience* **25**, 1569-1581, doi:10.1038/s41593-022-01186-3 (2022).

7 Larivière, S. *et al.* The ENIGMA Toolbox: multiscale neural contextualization of multisite neuroimaging datasets. *Nature Methods* **18**, 698-700, doi:10.1038/s41592-021-01186-4 (2021).

8 Laansma, M. A. *et al.* International Multicenter Analysis of Brain Structure Across Clinical Stages of Parkinson's Disease. *Mov Disord* **36**, 2583-2594, doi:10.1002/mds.28706 (2021).

9 Stout, K. A., Dunn, A. R., Hoffman, C. & Miller, G. W. The Synaptic Vesicle Glycoprotein 2: Structure, Function, and Disease Relevance. *ACS Chem Neurosci*

10, 3927-3938, doi:10.1021/acschemneuro.9b00351 (2019).
10 Finnema, S. J. *et al.* Imaging synaptic density in the living human brain. *Sci Transl Med* **8**, 348ra396, doi:10.1126/scitranslmed.aaf6667 (2016).
11 Markello, R. D. *et al.* Standardizing workflows in imaging transcriptomics with the abagen toolbox. *eLife* **10**, e72129, doi:10.7554/eLife.72129 (2021).

Decision on manuscript:

Your manuscript entitled "mFusion: A multiscale fusion method bridging neuroimages to genes through neurotransmissions in mental health disorders" has now been seen again by our referees, whose comments appear below. In light of their advice I am delighted to say that we are happy, in principle, to publish a suitably revised version in Communications Biology.

REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

The authors have appropriately addressed all my comments and I endorse the publication of this manuscript.

-- We thank you for your thoughtful assessment of our manuscript and for your suggestions to improve it.

Reviewer #2 (Remarks to the Author):

The authors have adequately addressed the concerns, and I agree to its publication.

-- We thank you for taking time to assess our manuscript, and for your positive feedback.