nature portfolio

Peer Review File

Transformer-Based Deep Learning Enhances Discovery in Migraine GWAS

Corresponding Author: Dr Xiao Chang

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)
Dr. Meng and co-authors investigated the genetic architecture of migraine using both cross-trait GWAS analysis and deep learning techniques using GWAS summary statistics from previously published studies. To do so, they used GWAS datasets of major depressive disorder (MDD), a complex disorder often associated with migraine and presenting a genetic overlap with migraine. The authors claimed to uncover 293 novel genetic loci (previously unreported for migraine) and new signaling pathways related to nitrogen compound metabolism and cation binding. The manuscript is well-written, and the results could clearly add to the literature of migraine genetics. The methods used seem appropriate and the GWAS summary statistics used are from large meta-analyses. Although previous studies have already applied machine learning techniques to enhance gene discovery for migraine (PMIDs: 39144710; 40326299), the novelty of the manuscript resides in the development of the Transformer-based Deep learning model (InsightGWAS) using MDD and in the identification of potential novel genetic loci and biological pathways underlying migraine susceptibility. Specific comments below:
Major comments
1. My first important concern is regarding the study design. Why only MDD was selected for the cross-trait GWAS and deep learning analyses? As based on the genetic correlation results (Table 1), other psychiatric disorders, including ADHD and ASD, shared a significant genetic background with migraine.
2. My other main concern is about the validation of the potential novel migraine loci identified by InsightGWAS. I recommend replicating those migraine loci in independent cohorts.
3. Results, page 5, line 98: “The most significant genetic correlation was observed between migraine and MDD (rg = -0.29, P = 1.78 × 10-33), indicating a substantial overlap in their genetic architectures”. However, in Table 1, it is reported “rg = 0.29”. Please confirm if a negative (or positive) genetic correlation was observed between MDD and migraine. Further, if the genetic correlation between migraine and MDD was negative (rg=-0.29), how would this affect the interpretation of overlapping genetic architectures.
4. If MDD and migraine have unique genetic determinants and distinct clinical hallmarks, is it valid to use transfer learning to identify genetic loci for migraine using MDD summary statistics?
5. How are “previously unreported” or novel loci defined? Based on distance with previously reported loci ? If it is the case, I would also recommend including an appropriate, strong, filter for LD as well e.g. r2 < 0.05 and/or an approach like GCTA COJO, to evaluate which loci are independent of other lead SNPs previously reported.
6. Introduction: Recent large GWAS/genetic investigations of migraine (Choquet H, et al. Commun Biol. 2021. PMID: 34294844; Bjornsdottir G, et al. Nat Genet. 2023.PMID: 37884687) are not cited. The overlap in terms of findings with those previous works and the current study should be discussed.
7. It would be helpful to conduct some simulations comparing the performance of InsightGWAS to other current and widely-used GWAS techniques. I am also interested to see QQ plots showing the different methods for testing the loci-trait associations under null simulations, examples of power of the different methods for GWAS analysis in simulations, and a comparison of which loci/genes are identified for each GWAS method (overlapping and different identifications)
8. I would suggest conducting a comparison of InsightGWAS against other transfer learning GWAS techniques (e.g., transferGWAS)? Why is InsightGWAS preferable to other established techniques? Please provide the potential advantages and disadvantages.
9. Results, lines 109-124: It would be helpful to model and compare the statistical power and inflation factors of InsightGWAS, MTAG, and condFDR.
10. The Transfer learning technique identified most of the migraine-associated loci previously reported in the Hautakangas’s paper, and also identified an “eightfold increase in SNP coverage through transfer learning”. Are the new loci identified

valid? How can we confirm that the new loci are still migraine-specific if MDD was used to extend SNP coverage? As above-mentioned, replication using independent cohorts is needed to validate the findings.
11. It may be helpful to identify overlapping loci between migraine, MDD, and other brain-related traits. This would allow identification of neurological-related biological pathways vs. migraine- or MDD- specific pathways/processes.

Minor comments
1. In the abstract (and Intro), please add the relevance of leveraging GWAS datasets of (only) MDD to better understand the genetic architecture of migraine.
2. Introduction: missing references for recent large GWAS/genetic investigations of migraine (Choquet H, et al. Commun Biol. 2021. PMID: 34294844; Bjornsdottir G, et al. Nat Genet. 2023.PMID: 37884687) and post-GWAS studies (Meyers TJ, et al. HGG Adv. 2023. PMID: 37415806; Gui J, et al. J Headache Pain. 2024.PMID: 38840241. Ghaffar A, et al. Hum Genet. 2023. PMID: 37245199; Xiong Z, et al. J Headache Pain. 2024. PMID: 39261750; Sun X, et al. J Headache Pain. 2024. PMID: 39039470)
3. Page 4, line 91: “GWAS summary statistics for migraine and five psychiatric conditions”, however, six psychiatric conditions are reported in Table 1 (including ASD).
4. Italicize gene name NUDT12 on line 117.
5. Need citations for “condFDR” and “MTAG” on line 112
6. Define “cdfr12” threshold on line 120
7. Page 9, line 223-224: “Furthermore, a recent large-scale sequencing study on migraine identified 13 novel loci.” Please provide the reference.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
Thank you for the opportunity to review “Transformer-Based Deep Learning Enhances Discovery in Migraine GWAS.”
The manuscript tackles an important challenge—rescuing sub-threshold GWAS hits by integrating functional annotations via a Transformer and transfer-learning from a well-powered psychiatric trait. The idea is timely and, if rigorously validated, could advance locus discovery.
However, several critical gaps must be addressed before this work can be considered for publication.

1) Statistical Validation of Performance Gains

ROC/AUC Significance: Figure 2 shows modest AUC improvements over DeepGWAS, XGBoost, but no confidence intervals or DeLong tests are reported. Without bootstrap-derived 95% CIs or pairwise significance tests, it is impossible to know whether these gains are meaningful or within noise.

2) Ablation of Transfer Learning

The core premise is that MDD pre-training boosts migraine discovery—but no “no-transfer” baseline (training the Transformer from scratch on migraine only) is reported.

3) Sensitivity to Genetic Correlation and Trait Selection

Why MDD? Table 1 shows other correlated traits (ADHD, ASD). It remains unclear how transfer performance varies with genetic correlation strength.

4) Feature-Importance and Nonlinear Interaction Evidence

Although the Transformer’s self-attention is used for modeling nonlinear interactions (also mentioned as advantage over DeepGWAS), no empirical demonstration is provided. Similarly, the relative contribution of p-value versus annotations is not quantified.

5) p-Value Scaling Bias in Transfer

Mixing extreme p-values from MDD (e.g. 10^{-12}) with more modest migraine p-values (10^{-8}) can train the model to focus on p-value scale rather than leveraging annotation context.

6) Redundant Pleiotropy Analysis

The manuscript devotes a section to pleiotropy (MTAG/condFDR) between migraine and MDD, yet those pleiotropic SNPs are never used for training, labeling, or loss-weighting—nor do they inform any downstream analysis beyond the initial correlation. As a result, the section adds neither methodological value nor practical insight.

7) Lack of Method-Level Benchmarking Across Traits and Conditions

Although the manuscript is presented as a generalizable method, it is only demonstrated on a single target trait (migraine) with one pre-training partner (MDD). There is no systematic exploration of the method's behavior across varying GWAS characteristics—such as differing sample sizes, maximum p-value strengths, or levels of genetic correlation—nor any "boundary condition" analyses that would guide users on when transfer learning is likely to help or hurt.

Recommendation: Major revision.

(Remarks on code availability)
The provided PyTorch scripts correctly implement the Transformer architecture, pre-training, fine-tuning and inference pipelines as described.
However, they lack flexible configuration (all hyperparameters hard-coded), reproducibility controls (no random-seed setting), structured logging, and modularization.

Reviewer #4

(Remarks to the Author)
This article presents a novel framework that combines cross-trait GWAS and Transformer-based deep learning to enhance genetic discovery in migraine. By applying transfer learning from MDD GWAS data, the authors introduce InsightGWAS, which improves the identification of both known and novel migraine-associated loci. The approach reveals shared genetic architecture with psychiatric disorders and uncovers biologically relevant pathways. While the methodology is innovative and technically fine, some improvements in clarity and rigor are needed to enhance reproducibility.

Comments:
1. The model's predictions were evaluated only on withheld migraine data from the same dataset used in training. There is no external dataset used for true validation. This raises concerns about potential overfitting, despite high performance metrics.
2. Given the complexity of the Transformer architecture, it would be helpful to include sensitivity analyses to show how robust the model's predictions are to different input features, hyperparameter choices, or label definitions.
3. While other psychiatric traits were analyzed via LDSC, the cross-trait GWAS and transfer learning focused exclusively on MDD. The manuscript would benefit from clarifying whether cross-trait meta-analysis was also attempted with other disorders (e.g., bipolar disorder or anxiety), and if so, how those results compared. Providing a stronger rationale for selecting MDD—beyond its high genetic correlation with migraine—would help justify this focus.
4. While the manuscript positions InsightGWAS as an advancement over DeepGWAS, the comparison lacks sufficient methodological depth. A clearer explanation of how InsightGWAS differs architecturally and conceptually from previous models—and why these changes matter—would make the contribution more distinct.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have done an excellent job addressing my comments.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
1) Performance Comparison with Other Methods

AUCs for Transformer (0.9994) and DeepGWAS (0.9990) are nearly identical, statistically significant with 10k bootstraps, but practically trivial.

For a methods paper, what matters is whether this tiny AUC gain translates into more robust novel discoveries, lower FDR, or better replication rates.

Right now, the performance differences are too small to justify the claim that the Transformer substantially improves over DeepGWAS.

The comparative advantage of InsightGWAS over existing methods is overstated; its performance is extremely close to DeepGWAS, and the added complexity may not be justified without clearer gains in real discovery.

2) Transfer Learning vs. Baseline Transformer

Training accuracy is not the right metric; both models already exceed 98–99%.

What matters is generalization: validation/test AUC, calibration, FDR, and especially the number of novel loci discovered.

The observed instability in the baseline curve may simply reflect hyperparameter choice, not an inherent weakness.

Overall, it does not convincingly show that transfer learning meaningfully improves migraine discovery over a well-tuned baseline Transformer.

3) Attention & Feature Importance

Attention maps do not prove nonlinear interaction modeling; they only show weight correlations. Rigorous interaction tests (e.g. Shapley interaction indices, synthetic data benchmarks) are needed.

Reported metric is accuracy, which is misleading under class imbalance. AUC or PR-AUC would be more convincing.

Little biological interpretation is provided (e.g. do the high-attention features align with migraine-relevant biology?).

While annotations add some predictive value, the evidence that Transformers uniquely capture nonlinear interactions (vs. DeepGWAS or simpler models) is weak. The model still seems heavily p-value driven.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
1. The inclusion of the MVP migraine GWAS as an independent dataset meaningfully strengthens the validation of the model. The reported replication rate is reasonable and provides supportive evidence of robustness.

2. The ablation experiments, label definition assessments, and hyperparameter tests are clear and informative. These results collectively support the robustness of the model under various conditions.

3. The additional cross-trait analyses and quantitative comparisons convincingly explain the choice of MDD as the primary trait for cross-trait GWAS and transfer learning. The rationale is well presented.

4.The clarification of architectural and conceptual differences improves the manuscript's positioning. The distinctions from DeepGWAS are now sufficiently outlined for the reader.

(Remarks on code availability)

Reply to Reviewers' Comments

We sincerely thank the editor and reviewers for their thoughtful and constructive feedback. We have carefully revised the manuscript in response to each comment and believe the changes have substantially strengthened the study. A point-by-point response follows below.

Reviewer #1 (Remarks to the Author):

Dr. Meng and co-authors investigated the genetic architecture of migraine using both crosstrait GWAS analysis and deep learning techniques using GWAS summary statistics from previously published studies. To do so, they used GWAS datasets of major depressive disorder (MDD), a complex disorder often associated with migraine and presenting a genetic overlap with migraine. The authors claimed to uncover 293 novel genetic loci (previously unreported for migraine) and new signaling pathways related to nitrogen compound metabolism and cation binding. The manuscript is well-written, and the results could clearly add to the literature of migraine genetics. The methods used seem appropriate and the GWAS summary statistics used are from large meta-analyses. Although previous studies have already applied machine learning techniques to enhance gene discovery for migraine (PMIDs: 39144710; 40326299), the novelty of the manuscript resides in the development of the Transformer-based Deep learning model (InsightGWAS) using MDD and in the identification of potential novel genetic loci and biological pathways underlying migraine susceptibility. Specific comments below:
Major comments
1. My first important concern is regarding the study design. Why only MDD was selected for the cross-trait GWAS and deep learning analyses? As based on the genetic correlation results (Table 1), other psychiatric disorders, including ADHD and ASD, shared a significant genetic background with migraine.

Reply: We thank the reviewer for raising this important question regarding our study design, particularly our decision to focus on major depressive disorder (MDD) for cross-trait GWAS and deep learning analyses.

First, as shown in Table 1, MDD exhibits the strongest genetic correlation with migraine (r_g = 0.29, p = 1.78×10^{-33}) among all psychiatric traits evaluated. While ADHD (r_g = 0.198, p = 5.33×10^{-11}) and ASD (r_g = 0.115, p = 0.0029) also share polygenic overlap with migraine, their genetic correlations are

notably lower, suggesting a less robust shared genetic architecture and thus a weaker foundation for transfer learning in this context.

Second, the MDD GWAS includes 170,756 cases and 329,443 controls, which is substantially larger than the GWAS datasets for ADHD (38,691 cases; 186,843 controls) and ASD (18,381 cases; 27,969 controls). Notably, both ADHD and ASD sample sizes are smaller than that of the migraine GWAS itself (53,109 cases; 316,078 controls), which undermines the fundamental principle of transfer learning—that the source dataset should be larger and more informative than the target. Using lower-powered source traits would limit the robustness of model pretraining and reduce the effectiveness of knowledge transfer.

Third, in terms of variant-level signal, the MDD GWAS identified 102 genome-wide significant loci, which yielded approximately 4,600 positive SNPs for model training. In comparison, ADHD and ASD yielded only 1,428 and 93 positive variants, respectively—both of which are fewer than the 1,834 positive variants available from the migraine GWAS itself. The limited number of positive labels in ADHD and ASD results in severe class imbalance and restricts the model's ability to learn and generalize meaningful patterns. By contrast, MDD offers a richer and more balanced training set that supports stable optimization and more accurate prediction in a deep learning framework.

To further evaluate this design choice, we conducted additional cross-trait GWAS using MTAG and condFDR using ADHD, ASD, ANX, BIP and SCZ. However, none of these analyses identified additional genome-wide significant loci beyond those already present in the original migraine GWAS. These negative findings further support our conclusion that MDD is the most suitable source trait for transfer learning in this study. We have updated the main text accordingly to incorporate these

findings and clarify the rationale for selecting MDD (Page 6, Lines 128-135; Page 8, Lines 190-204; Tables S4-S13).

2. My other main concern is about the validation of the potential novel migraine loci identified by InsightGWAS. I recommend replicating those migraine loci in independent cohorts.

Reply: We appreciate the reviewer's comment regarding the validation of the novel migraine loci identified by InsightGWAS in independent cohorts.

As most published migraine GWAS to date have been conducted by the International Headache Genetics Consortium (IHGC), there is substantial overlap across existing studies, limiting opportunities for fully independent replication. However, we were fortunate to identify a newly released and independent large-scale GWAS from the VA Million Veteran Program (MVP), recently published in Science (Diversity and scale: Genetic architecture of 2068 traits in the VA Million Veteran Program) (Verma A et al.,2024). This resource provided publicly available summary statistics for migraine based on two phenotype ascertainment approaches: (i) GCST90475546: Migraine headaches, based on self-reported symptoms (28,635 cases; 315,668 controls), and (ii) GCST90475837: Migraine (PheCode 340), based on ICD-coded clinical diagnosis (31,836 cases; 437,667 controls). For each locus, we examined the lead variant and applied a replication threshold of $P < 0.05$. Although the phenotype definitions and case ascertainment methods in the MVP cohort are not fully aligned with those used in IHGC-based GWAS, 117 loci replicated in at least one of the two MVP migraine datasets, and 23 loci replicated in both. These findings provide further support for the validity of a substantial subset of the novel loci prioritized by InsightGWAS.

We have added these results to the revised manuscript (Pages 11-12, Lines 299-321; Table S15).

Reference

Verma A, Huffman JE, Rodriguez A, et al. Diversity and scale: Genetic architecture of 2068 traits in the VA Million Veteran Program. Science. 2024 Jul 19;385(6706):eadj1182. doi: 10.1126/science.adj1182. Epub 2024 Jul 19. PMID: 39024449.

3. Results, page 5, line 98: "The most significant genetic correlation was observed between migraine and MDD (rg = -0.29, P = 1.78 × 10-33), indicating a substantial overlap in their genetic architectures". However, in Table 1, it is reported "rg = 0.29". Please confirm if a negative (or positive) genetic correlation was observed between MDD and migraine. Further, if the genetic correlation between migraine and MDD was negative (rg=-0.29), how would this affect the interpretation of overlapping genetic architectures.

Reply: We sincerely thank the reviewer for the careful reading and for pointing out the inconsistency in the reported genetic correlation. The correct genetic correlation between migraine and MDD is positive (r_g = 0.29, P = 1.78 × 10^{-33}), as shown in Table 1. The value in the Results section ("r_g = –0.29") was a typographical error. We have corrected this in the revised manuscript and carefully reviewed all reported genetic correlations to ensure consistency and accuracy throughout the text (Page 5, Line 104).

4. If MDD and migraine have unique genetic determinants and distinct clinical hallmarks, is it valid to use transfer learning to identify genetic loci for migraine using MDD summary statistics?

Reply: We appreciate the reviewer's insightful question. While it is true that migraine and MDD have distinct clinical characteristics and partially non-overlapping genetic architectures, transfer learning does not require phenotypic equivalence. Rather, it is designed to leverage informative patterns learned from a related trait to enhance discovery in a data-limited target context. In brief, the model was first pretrained on MDD GWAS summary statistics to capture generalizable patterns of genetic association. It was then fine-tuned using migraine GWAS data to adapt the model to migraine-specific

genetic architecture. Finally, the inference was carried out solely on migraine data, ensuring that all findings were directly relevant to migraine genetics.

A more detailed explanation is as follows:

MDD was selected as the source trait not due to clinical similarity, but because of its large GWAS sample size (170,756 cases and 329,443 controls) and high signal density (102 genome-wide significant loci, ~4,600 positive variants), which together enable effective training of deep learning models. In contrast, ADHD and ASD offer fewer than 1,500 and 100 positive variants, respectively—less than even the migraine GWAS (1,834 positive variants), making them less suitable for this purpose.

Critically, our InsightGWAS framework follows a two-stage transfer learning design:

Pretraining on MDD allows the model to learn general associations between variant-level features (e.g., statistical metrics, annotations) and likelihood of disease relevance.

Fine-tuning on migraine enables the model to adapt to migraine-specific signal patterns. During this step, model parameters are updated using only migraine GWAS data, allowing us to retain transferable insights while ensuring all final discoveries are specific to the migraine context.

This structure reflects a central principle of transfer learning: shared structure can be exploited to improve prediction while accommodating differences between source and target tasks. **Importantly, all loci prioritized by InsightGWAS in this study result from this fine-tuned model applied to migraine data only (not from MDD).** Additionally, many of the novel loci identified were replicated in an independent migraine cohort (MVP), reinforcing the validity of this approach.

We have updated the manuscript to clarify this point and to underscore that all biological interpretations and discoveries are based exclusively on migraine datasets (Page 9, Lines 228-241).

P.S.:

To address another reviewer's comment regarding the transfer learning step of our framework (Reviewer #3, comment 2), we conducted a direct comparison between the transfer learning and non-transfer (baseline) versions of InsightGWAS. As expected, the transfer learning model demonstrated improved performance in the migraine setting, where a well-powered and genetically correlated source trait (MDD) was available. However, recognizing that transfer learning may not be suitable for all trait pairs—particularly when no appropriate source phenotype with sufficient genetic overlap exists—we have made both versions of the model publicly available in our GitHub repository (https://github.com/ziangmeng/MA-MDD-Transformer-based-model-/tree/main). This provides users the flexibility to select or benchmark either approach according to the characteristics of their target GWAS.

5. How are "previously unreported" or novel loci defined? Based on distance with previously reported loci? If it is the case, I would also recommend including an appropriate, strong, filter for LD as well e.g. r2 < 0.05 and/or an approach like GCTA COJO, to evaluate which loci are independent of other lead SNPs previously reported.

Reply: We appreciate the reviewer's comment. In this study, the definition of 'novel' loci was based on linkage disequilibrium (LD) (r^2). Specifically, loci with an LD r^2 less than 0.05 with previously reported loci associated with migraine were considered novel. Additional clarification on this aspect has been provided in the revised manuscript (Page 11, Lines 279-280).

6. Introduction: Recent large GWAS/genetic investigations of migraine (Choquet H, et al. Commun Biol. 2021. PMID: 34294844; Bjornsdottir G, et al. Nat Genet. 2023.PMID: 37884687) are not cited. The overlap in terms of findings with those previous works and the current study should be discussed.

Reply: We sincerely appreciate the reviewer's suggestion to cite recent large-scale GWAS studies on migraine. We have now incorporated these important references in the Introduction and Results section of our revised manuscript (Page 3, Line 44; Page 11, Lines 287-297; Page 12, Lines 309-321).

7. It would be helpful to conduct some simulations comparing the performance of InsightGWAS to other current and widely-used GWAS techniques. I am also interested to see QQ plots showing the different methods for testing the loci-trait associations under null simulations, examples of power of the different methods for GWAS analysis in simulations, and a comparison of which loci/genes are identified for each GWAS method (overlapping and different identifications)

Reply: We thank the reviewer for this thoughtful suggestion. However, we would like to clarify that InsightGWAS is not a traditional GWAS method that performs statistical association testing or generates p-values. Rather, it is a post-GWAS variant prioritization framework that leverages learned patterns from a pretrained model and SNP-level functional features to re-rank variants from existing GWAS summary statistics. As such, InsightGWAS is not directly applicable in simulation settings based on null hypothesis testing, and does not generate association statistics suitable for QQ plot comparison or power analysis in the conventional sense.

Instead, in this paper, we demonstrate the utility of InsightGWAS through real-data comparisons. For example, using GWAS summary statistics from a migraine study excluding 23andMe data (~49K cases), InsightGWAS identified 367 independent signals, including all 41 genome-wide significant loci previously reported. Notably, it also prioritized 53 additional loci that were only discovered in a much larger GWAS with ~102K cases. Furthermore, in line with the reviewer's previous suggestion to assess the validity of the prioritized loci, we tested them in an independent migraine GWAS from the VA Million Veteran Program. Among the 293 novel loci predicted by InsightGWAS, 117 were replicated ($P < 0.05$) in at least one MVP dataset, and 23 loci were supported by both, providing additional

evidence for the biological relevance and generalizability of many of the signals identified by our model.

8. I would suggest conducting a comparison of InsightGWAS against other transfer learning GWAS techniques (e.g., transferGWAS)? Why is InsightGWAS preferable to other established techniques? Please provide the potential advantages and disadvantages.

Reply: We thank the reviewer for the thoughtful suggestion. While both InsightGWAS and transferGWAS utilize transfer learning principles, they are fundamentally different in scope, input data, and analytical goals, and thus not directly comparable.

transferGWAS is designed for GWAS of complex image-derived phenotypes. It first uses a convolutional neural network (CNN) pretrained on external image datasets to extract low-dimensional image embeddings, which are then used as quantitative traits in downstream GWAS using linear mixed models. The transfer occurs at the phenotype level, leveraging pretrained image representations to define new phenotypes for association testing.

In contrast, InsightGWAS is a post-GWAS prioritization framework. It takes as input existing GWAS summary statistics (not raw phenotype or genotype data) and integrates functional annotations of SNPs. The model is pretrained on one trait (e.g., MDD) and fine-tuned and applied to a second trait (e.g., migraine) to improve variant prioritization. The transfer occurs at the model level, not the phenotype level. It does not perform de novo association testing or produce p-values, and is instead designed to enhance discovery by re-ranking variants based on functional and statistical patterns.

9. Results, lines 109-124: It would be helpful to model and compare the statistical power and inflation factors of InsightGWAS, MTAG, and condFDR.

Reply: We thank the reviewer for this valuable suggestion. However, we would like to clarify that InsightGWAS is not a statistical testing framework and does not generate p-values, so conventional measures such as genomic inflation factors (λ_{GC}) or statistical power curves cannot be directly applied.

In contrast to MTAG and condFDR, which generate association statistics and can be assessed using metrics like genomic control and QQ plots, InsightGWAS functions as a post-GWAS prioritization tool. Its goal is to improve the ranking of variants based on integrated functional and statistical features, especially in underpowered GWAS contexts. Therefore, comparisons in terms of λ or type I error are not meaningful or appropriate for InsightGWAS.

As an alternative, we assessed the effectiveness of InsightGWAS by comparing its prioritized loci with those identified in larger, more powered GWAS (e.g., the full Hautakangas migraine GWAS including 23andMe), as well as testing replication in an independent cohort (MVP). These analyses demonstrate that InsightGWAS can successfully recover loci that are missed in smaller GWAS but later confirmed in larger studies or independent datasets, highlighting its practical utility in signal enrichment.

We have added clarification in the revised Results to reflect these methodological differences and our evaluation strategy (Page 7, Lines 157-165).

10. The Transfer learning technique identified most of the migraine-associated loci previously reported in the Hautakangas's paper, and also identified an "eightfold increase in SNP coverage through transfer learning". Are the new loci identified valid? How can we confirm that the new loci are still migraine-specific if MDD was used to extend SNP coverage? As above-mentioned, replication using independent cohorts is needed to validate the findings.

Reply: We thank the reviewer for raising this important question. While the pretraining stage of InsightGWAS uses MDD GWAS data to learn transferable patterns of genetic architecture, the inference stage is conducted entirely on migraine GWAS summary statistics. This means that all

variant prioritization and signal enhancement are specific to the migraine trait and based solely on migraine data.

To evaluate whether the additional loci identified by InsightGWAS are truly relevant to migraine rather than reflecting patterns from the pretraining trait (MDD), we performed replication analysis using an independent migraine GWAS from the VA Million Veteran Program (MVP). Among the 293 novel loci identified by InsightGWAS, 117 were replicated ($P < 0.05$) in at least one MVP migraine dataset, and 23 loci were replicated in both phenotype definitions, despite differences in cohort composition and migraine phenotype ascertainment. These findings provide further support for the validity and migraine-specificity of the signals prioritized by InsightGWAS.

Moreover, InsightGWAS is designed to identify loci with sub-threshold association evidence that may be missed by conventional GWAS due to limited power. The model does not transfer p-values or associations from MDD, but rather uses MDD-pretrained representations to improve the ranking of SNPs in the migraine context.

We have clarified this distinction and included the replication results in the revised manuscript to address this concern (Page 9-10, Lines 228-257; Pages 11-12, Lines 299-321; Table S15).

11. It may be helpful to identify overlapping loci between migraine, MDD, and other brain-related traits. This would allow identification of neurological-related biological pathways vs. migraine or MDD- specific pathways/processes.

Reply: We thank the reviewer for this insightful comment. To assess whether the transfer learning model introduced bias toward MDD or other neuropsychiatric traits, we systematically examined the overlap between the 367 independent loci prioritized by InsightGWAS for migraine and previously reported loci for MDD, ADHD, and ASD.

Using LD-based matching ($r^2 > 0.05$) to known GWAS hits, we found that only 7 loci overlapped with MDD-associated variants, 2 with ADHD, and none with ASD (see Table R1). This suggests that the vast majority of InsightGWAS-prioritized loci are migraine-specific and are not driven by signals from the pretraining trait.

These findings support that while the model leveraged generalizable genetic patterns learned from MDD GWAS during the pretraining phase, its inference and prediction were made entirely on migraine GWAS data, and we found no evidence of bias toward unrelated traits. Thus, the transfer learning component serves to enhance signal detection while preserving trait specificity.

Table R1: Overlap between InsightGWAS-prioritized migraine loci and neuropsychiatric GWAS loci.

Migraine SNP (InsightGWAS)	**Matched SNP (other neuropsychiatric-traits GWAS)**	**Trait**	**LD (r^2)**
rs7618883	rs13084037	MDD	0.1386
rs144070260	rs13084037	MDD	0.0559
rs11547261	rs13084037	MDD	0.0572
rs3883861	rs6783233	MDD	0.0640
rs3013736	rs7030813	MDD	0.4476
rs55707505	rs1045430	MDD	0.1261
rs9607782	rs5995992	MDD	0.7984
rs895219	rs1438898	ADHD	0.0799
rs4842676	rs704061	ADHD	0.1008
—	—	ASD	—

Note: Overlap was defined as $r^2 > 0.05$ based on 1000 Genomes EUR LD.

Minor comments

1. In the abstract (and Intro), please add the relevance of leveraging GWAS datasets of (only) MDD to better understand the genetic architecture of migraine.

Reply: We thank the reviewer for this thoughtful suggestion. We fully agree that explaining the choice of MDD is important for understanding the study design. However, the abstract is limited to 150 words under *Nature Communications* guidelines, and we were unable to include detailed justification within this space. We respectfully hope the reviewer understands this limitation. To address the concern, we have revised the Introduction and Main text to clearly describe the rationale for using MDD, based on its genetic correlation with migraine and suitability for transfer learning (Page 4, Lines 83-86).

2. Introduction: missing references for recent large GWAS/genetic investigations of migraine (Choquet H, et al. Commun Biol. 2021. PMID: 34294844; Bjornsdottir G, et al. Nat Genet. 2023.PMID: 37884687) and post-GWAS studies (Meyers TJ, et al. HGG Adv. 2023. PMID: 37415806; Gui J, et al. J Headache Pain. 2024.PMID: 38840241. Ghaffar A, et al. Hum Genet. 2023. PMID: 37245199; Xiong Z, et al. J Headache Pain. 2024. PMID: 39261750; Sun X, et al. J Headache Pain. 2024. PMID: 39039470)

Reply: We thank the reviewer for highlighting the omission of recent key references related to migraine genetics. We have now incorporated all the suggested citations for recent migraine GWAS studies and post-GWAS investigations in the Introduction section of our revised manuscript (Page 3, Lines 44-48).

3. Page 4, line 91: "GWAS summary statistics for migraine and five psychiatric conditions", however, six psychiatric conditions are reported in Table 1 (including ASD).

Reply: We thank the reviewer for pointing out this inconsistency. The text has been corrected to accurately reflect that six psychiatric conditions (including ASD) were analyzed (Page 5, Line 97).

4. Italicize gene name NUDT12 on line 117.

Reply: Thank you for catching this. We have italicized NUDT12 as required (Page 5, Line 123).

5. Need citations for "condFDR" and "MTAG" on line 112

Reply: We thank the reviewer for pointing out this omission. We have now added the appropriate citations for both methods (Page 5, Line 117).

6. Define "cdfr12" threshold on line 120

Reply: We appreciate the reviewer's attention to methodological clarity. We have now explicitly defined "cdfr12" as conditional FDR < 0.01 in the revised manuscript (Page 6, Line 127).

7. Page 9, line 223-224: "Furthermore, a recent large-scale sequencing study on migraine identified 13 novel loci." Please provide the reference.

Reply: Thank you for pointing this out, we have now added the appropriate reference at the indicated location (Page 11, Line 288).

Reviewer #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reply: We appreciate the reviewer's contribution to the evaluation of our manuscript and acknowledge the collaborative peer review process under the Nature Communications initiative.

Reviewer #3 (Remarks to the Author):

Thank you for the opportunity to review "Transformer-Based Deep Learning Enhances Discovery in Migraine GWAS."The manuscript tackles an important challenge—rescuing sub-threshold GWAS hits by integrating functional annotations via a Transformer and transfer-learning from a well-powered psychiatric trait. The idea is timely and, if rigorously validated, could advance locus discovery. However, several critical gaps must be addressed before this work can be considered for publication.

1) Statistical Validation of Performance Gains

ROC/AUC Significance: Figure 2 shows modest AUC improvements over DeepGWAS, XGBoost, but no confidence intervals or DeLong tests are reported. Without bootstrap-derived 95% CIs or pairwise significance tests, it is impossible to know whether these gains are meaningful or within noise.

Reply: We thank the reviewer for the thoughtful suggestion regarding statistical validation of model performance. To address this concern, we conducted extensive bootstrap analysis with 10,000 replicates to evaluate the robustness and significance of performance gains across methods. As shown in Figure R1, the bootstrap-derived AUC distributions demonstrate that our Transformer-based InsightGWAS model consistently outperformed DeepGWAS, XGBoost, Logistic Regression, and Ridge Regression.

These intervals confirm that the observed improvements are robust and unlikely to arise from random variation. We have added these results and visualizations to the revised manuscript (Page 9, Lines 219-223 and Figure S2).

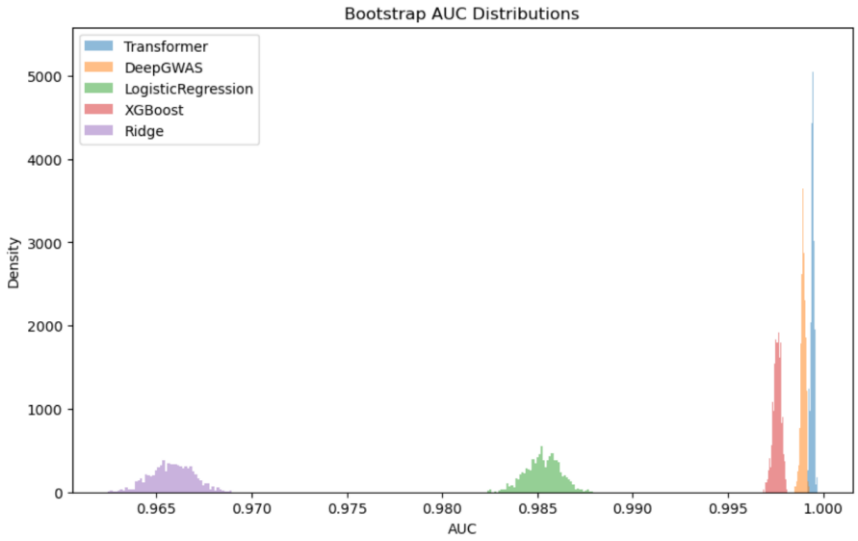

Figure R1. Bootstrap AUC distributions across different models.

Distribution of AUC values computed from 10,000 bootstrap replicates for each model: Transformer-based InsightGWAS (blue, mean AUC = 0.9994, 95% CI = 0.9992–0.9996), DeepGWAS (orange,

0.9990, 0.9987–0.9992), Logistic Regression (green, 0.9853, 0.9834–0.9869), XGBoost (red, 0.9976, 0.9971–0.9980), and Ridge Regression (purple, 0.9658, 0.9635–0.9680).

2) Ablation of Transfer Learning

The core premise is that MDD pre-training boosts migraine discovery—but no "no-transfer" baseline (training the Transformer from scratch on migraine only) is reported.

Reply: We thank the reviewer for highlighting the importance of evaluating the contribution of transfer learning. To evaluate the effectiveness of MDD-based pretraining, we conducted a controlled experiment where both the baseline Transformer (trained from scratch on migraine data only) and the transfer model (initialized from the pretrained MDD model) were trained under identical settings: 20 epochs, a learning rate of 0.01, and the same training set.

As shown in Training Accuracy Comparison Between Transfer and Baseline Models (Figure R2), the transfer model consistently achieves higher accuracy than the baseline model at nearly every training epoch. While the baseline model exhibits greater variance and multiple sharp drops in accuracy during training (e.g., falling below 95% at several epochs), the transfer model maintains stable and high performance, with accuracy consistently above 98% and peaking at 99.37%. This suggests that the pretrained weights from the MDD dataset provide a strong initialization that accelerates convergence and improves robustness against fluctuations in model updates.

Additionally, recognizing that the effectiveness of transfer learning depends on the availability of suitable source traits with sufficient genetic correlation, we have released the full source code—including both the transfer learning and baseline (non-transfer) versions—on our GitHub repository (https://github.com/ziangmeng/MA-MDD-Transformer-based-model-/tree/main). This enables users to flexibly apply or adapt the method according to the characteristics of their target phenotype,

including scenarios where no appropriate source trait is available for transfer.

We have revised the manuscript accordingly (Pages 9-10, Lines 243-257; Pages 19, Lines 540-541).

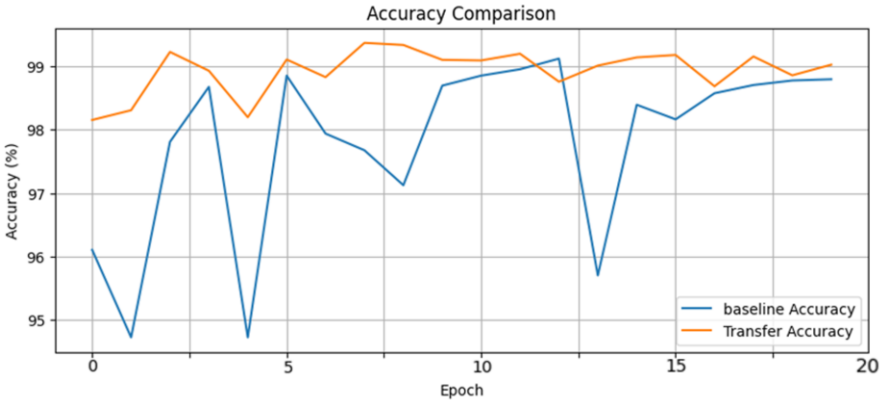

Figure R2. Comparison of classification accuracy between transfer learning and baseline models across training epochs.

Training accuracy (%) over 20 epochs for the Transformer-based InsightGWAS model with (orange) and without (blue) transfer learning.

3) Sensitivity to Genetic Correlation and Trait Selection Why MDD? Table 1 shows other correlated traits (ADHD, ASD). It remains unclear how transfer performance varies with genetic correlation strength.

Reply: We sincerely appreciate the reviewer's insightful question regarding trait selection and the relationship between genetic correlation strength and transfer learning performance. This important consideration is indeed closely related to the first concern raised by Reviewer #1, and we are grateful for the opportunity to further elaborate on our methodological rationale.

In our study, we selected MDD as the source trait for transfer learning based on a combination of biological relevance, statistical justification, and empirical feasibility. While ADHD and ASD also exhibit positive genetic correlations with migraine (r_g = 0.198 and 0.115, respectively), MDD shows the strongest correlation (r_g = 0.29, P = 1.78×10^{-33}), suggesting a higher degree of polygenic overlap.

More critically, the GWAS for MDD offers considerably greater statistical power and richer training information. The MDD dataset comprises 170,756 cases and 329,443 controls, yielding 102 genome-wide significant loci and approximately 4,600 positive SNPs for model training. In contrast, ADHD and ASD GWAS report only 27 and 5 genome-wide loci, corresponding to 1,428 and 93 positive variants, respectively—both fewer than the 1,834 positive variants identified in the migraine GWAS itself. From this perspective, ADHD and ASD do not offer any advantage in data quantity or signal strength over migraine. Consequently, using them as source traits would undermine the core rationale of transfer learning and could even lead to inferior performance compared to a baseline model trained solely on migraine data.

We have clarified these points in the revised manuscript (Page 8, Lines 190-204).

4) Feature-Importance and Nonlinear Interaction Evidence

Although the Transformer's self-attention is used for modeling nonlinear interactions (also mentioned as advantage over DeepGWAS), no empirical demonstration is provided. Similarly, the relative contribution of p-value versus annotations is not quantified.

Reply: We appreciate the reviewer's insightful comment on the need for empirical evidence supporting the Transformer model's ability to capture nonlinear feature interactions and quantify the relative contribution of different input modalities. To address this point, we conducted two complementary analyses.

First, we performed an attention-weight analysis using the final layer of the Transformer encoder trained on the migraine dataset. As shown in Figure R3, the resulting self-attention heatmap highlights clear cross-feature dependencies—most notably between GWAS statistics (e.g., p-values) and regulatory annotations such as eQTLs and chromatin accessibility features. These interactions are not confined to within-group signals but span distinct biological feature types, suggesting that the

model effectively captures nonlinear relationships between statistical significance and functional relevance.

Second, we conducted an ablation study by grouping all input features into four functional categories: (1) position-related features (chromosome, base-pair position, LD score, allele frequency); (2) GWAS statistical features (including p-value, effect size, standard error, and sample size); (3) molecular QTL annotations (e.g., sQTL, eQTL); and (4) regulatory and OCR-based features (such as ENCODE, transcription factor footprints, and previously reported GWAS signals). Five separate models were trained using the same architecture and protocol: one using all features and four additional models each excluding a single feature group. Removing GWAS statistics led to the largest drop in accuracy (from 0.9840 to 0.9634), underscoring their strong predictive power. However, excluding molecular QTLs (accuracy: 0.9723) and regulatory/OCR annotations (accuracy: 0.9827) also resulted in measurable performance reductions. These results demonstrate that while statistical features are important, functional annotations contribute substantially to performance, confirming that the Transformer model integrates diverse biological information rather than relying solely on p-value scale. We have revised the manuscript accordingly (Pages 7-8, Lines 167-188; Figure S1).

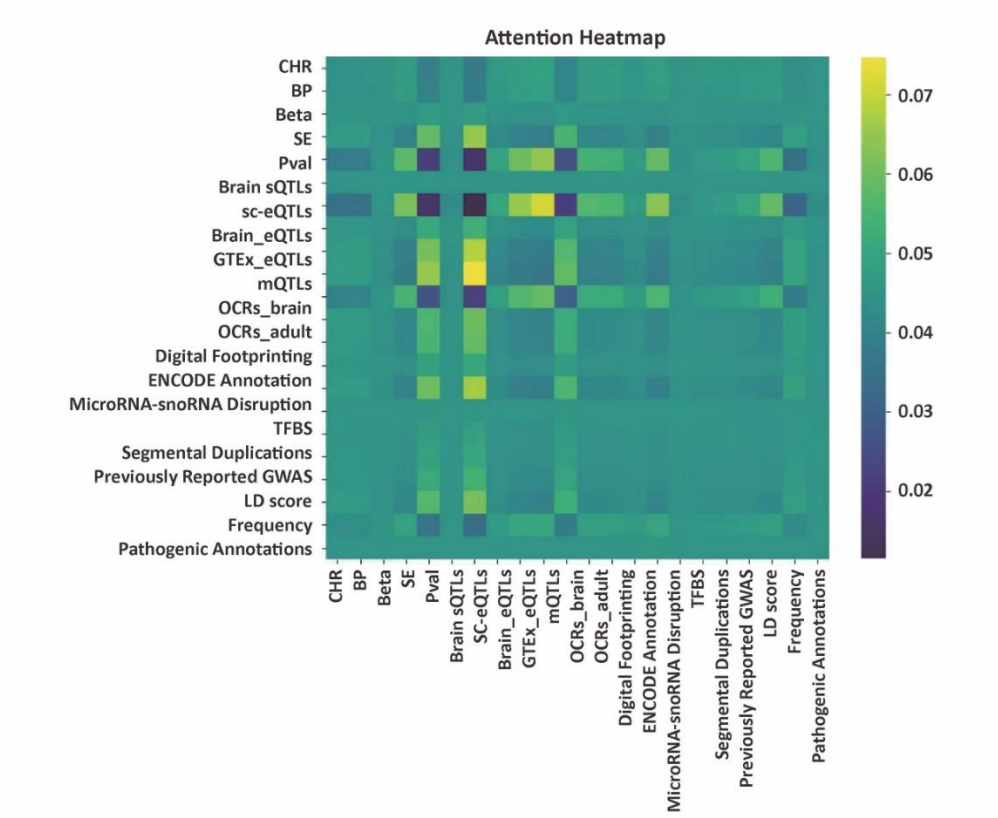

Figure R3. Feature-wise self-attention heatmap in the Transformer model.

This heatmap visualizes the self-attention scores among variant-level features learned by the final layer of the Transformer encoder during fine-tuning. The x- and y-axes represent input feature dimensions, grouped into four functional categories: (1) position-related features (chromosome, base-pair position, LD score, and allele frequency); (2) GWAS statistical features (p-value, effect size, standard error, and sample size); (3) molecular QTL annotations (eQTL, sQTL, and other transcript-level QTLs); and (4) regulatory and OCR-based features (ENCODE signals, transcription factor footprints, and previously reported GWAS hits). The color intensity reflects the magnitude of self-attention assigned by the model from one feature to another, with warmer colors indicating stronger cross-feature dependencies. The pattern highlights how the Transformer integrates statistical and functional inputs in a context-aware and nonlinear manner.

5) p-Value Scaling Bias in Transfer

Mixing extreme p-values from MDD (e.g. 10^{-12}) with more modest migraine p-values (10^{-8}) can train the model to focus on p-value scale rather than leveraging annotation context.

Reply: We appreciate the reviewer's concern regarding potential overreliance on p-values,

particularly the risk that mixing highly significant MDD p-values (e.g., 10^{-12}) with more modest

migraine p-values (e.g., 10^{-8}) could bias the model toward p-value scale rather than biological context. We would like to clarify that in our framework, no such mixing occurs during supervised training. While MDD data are used for pretraining in the transfer learning setting, the final training, fine-tuning, and inference are conducted exclusively on migraine data—using migraine-specific summary statistics and annotations. Thus, all prediction is based solely on the migraine p-value distribution.

To further evaluate the model's reliance on different input modalities, we performed an ablation study by removing feature groups and measuring accuracy drops. Removing the GWAS statistics group (which includes p-value, beta, SE, and N) led to the largest accuracy decline (from 0.9840 to 0.9634), confirming its importance. However, removing functional annotations such as molecular QTLs (accuracy: 0.9723) and open chromatin/regulatory features (accuracy: 0.9827) also resulted in measurable drops, indicating that the model does not depend on p-values alone. Instead, it integrates diverse biological signals to inform variant prioritization.

These results support that although p-value features are influential, the model does not simply rank variants by magnitude. Rather, it captures complementary annotation-driven patterns relevant to migraine biology. We have added this clarification and the results of the ablation analysis to the revised manuscript (Pages 7-8, Lines 167-188; Figure S1).

6) Redundant Pleiotropy Analysis

The manuscript devotes a section to pleiotropy (MTAG/condFDR) between migraine and MDD, yet those pleiotropic SNPs are never used for training, labeling, or loss-weighting—nor do they inform any downstream analysis beyond the initial correlation. As a result, the section adds neither methodological value nor practical insight.

Reply: We thank the reviewer for this important point. Our goal is to advance the identification of migraine-associated loci beyond the current discoveries from GWAS performed by the International Headache Genetics Consortium (IHGC). As MTAG and condFDR are valuable extensions of single-trait GWAS, we initially applied these methods to identify additional migraine loci. However, in our analysis, they yielded only a few new signals, suggesting the limited marginal gain that conventional methods can offer beyond existing GWAS results.

These limitations led us to explore deep learning methods, resulting in the development of InsightGWAS, which leverages functional annotations and a Transformer model for variant prioritization. Migraine is particularly suitable for this approach due to its genetic characteristics and the availability of two large-scale, high-quality datasets.

While InsightGWAS performed well in the migraine setting, we acknowledge that the method is still evolving and will require further validation in other traits. We have clarified this methodological progression and its current limitations in the revised manuscript (Page 16, Lines 446-452).

7) Lack of Method-Level Benchmarking Across Traits and Conditions

Although the manuscript is presented as a generalizable method, it is only demonstrated on a single target trait (migraine) with one pre-training partner (MDD). There is no systematic exploration of the method's behavior across varying GWAS characteristics—such as differing sample sizes, maximum p-value strengths, or levels of genetic correlation—nor any "boundary condition" analyses that would guide users on when transfer learning is likely to help or hurt.

Reply: We thank the reviewer for this insightful comment. As noted in our response to the previous point, InsightGWAS is not yet intended as a fully generalizable framework across all trait pairs. In this study, we focus on advancing migraine GWAS discovery, and selected migraine as the target trait due to its well-characterized genetic architecture and the availability of two large, high-quality GWAS

datasets. MDD was chosen as the source trait because of its large sample size and moderate genetic correlation with migraine, which made it a biologically and statistically compatible candidate for transfer learning.

We acknowledge that a systematic evaluation across traits with varying GWAS characteristics—such as different sample sizes, p-value distributions, or degrees of polygenicity—would be valuable. However, our current implementation includes trait-specific feature sets, which limits its immediate applicability to other conditions without retraining or reconfiguration. Moreover, the success of transfer learning depends on the availability of well-matched source traits, which may not always be present.

We are actively extending InsightGWAS to support additional disease contexts and to explore its performance under broader conditions. These directions are part of ongoing work, and the current limitations have been explicitly discussed in the revised Discussion section (Page 16, Lines 446-452).

Recommendation: Major revision.

Reviewer #3 (Remarks on code availability):

The provided PyTorch scripts correctly implement the Transformer architecture, pre-training, fine-tuning and inference pipelines as described.

However, they lack flexible configuration (all hyperparameters hard-coded), reproducibility controls (no random-seed setting), structured logging, and modularization.

Reply: We thank the reviewer for the insightful comments regarding the provided PyTorch scripts.

To address the issue of flexible configuration, we have updated all scripts to accept command-line arguments for key hyperparameters, including the number of attention heads, Transformer layers, hidden dimensions, batch size, learning rate, number of training epochs, and classification threshold. This enables users to conveniently adjust experimental settings without modifying the source code directly. The accompanying README has been expanded to clearly document each configurable option, with default values and example commands for training, transfer learning, inference, and baseline comparisons.

Regarding reproducibility, structured logging, and modularization, we have uploaded the exact pretrained models used in our study to the GitHub repository (https://github.com/ziangmeng/MA-MDD-Transformer-based-model-/tree/main) to ensure full reproducibility of our results, independent of implementation-level variability. Specifically:

model/transformer_model_mdd.pth: Pretrained on MDD GWAS data

model/transformer_model_finetuned.pth: Fine-tuned on migraine GWAS data

These checkpoints correspond to the final models reported in the manuscript and are sufficient to reproduce all key findings, even though the example scripts have been intentionally simplified for clarity and educational purposes. We have clarified this in the “Code availability” section (Page 22, Lines 628-634).

Reviewer #4 (Remarks to the Author):

This article presents a novel framework that combines cross-trait GWAS and Transformerbased deep learning to enhance genetic discovery in migraine. By applying transfer learning from MDD GWAS data, the authors introduce InsightGWAS, which improves the identification of both known and novel migraine-associated loci. The approach reveals shared genetic

architecture with psychiatric disorders and uncovers biologically relevant pathways. While the methodology is innovative and technically fine, some improvements in clarity and rigor are needed to enhance reproducibility.

Comments:

1. The model's predictions were evaluated only on withheld migraine data from the same dataset used in training. There is no external dataset used for true validation. This raises concerns about potential overfitting, despite high performance metrics.

Reply: We thank the reviewer for raising this important point.

In the revised manuscript, we addressed this concern by conducting independent validation using newly released migraine GWAS summary statistics from the VA Million Veteran Program (MVP). These data—published in Science (Verma A et al.,2024) and absent from prior migraine meta-analyses—represent an entirely independent cohort that, despite employing distinct case definitions and ascertainment strategies, provides a valuable opportunity for external replication.

Specifically, we examined whether the novel loci identified by InsightGWAS replicated in migraine under two different clinical definitions: GCST90475546 (self-reported migraine; 28,635 cases, 315,668 controls) and GCST90475837 (ICD-coded migraine diagnosis, PheCode 340; 31,836 cases, 437,667 controls).

Applying a nominal significance threshold ($P < 0.05$), 117 of the 293 novel loci (40%) replicated in at least one of the two MVP datasets, and 23 loci were replicated in both. This level of replication—despite differences in phenotype definition and ascertainment method—supports the robustness and potential biological relevance of the loci prioritized by InsightGWAS.

We have included this validation analysis and its results in the revised manuscript (Pages 11-12, Lines 299-321; Table S15).

Reference

Verma A, Huffman JE, Rodriguez A, et al. Diversity and scale: Genetic architecture of 2068 traits in the VA Million Veteran Program. Science. 2024 Jul 19;385(6706):eadj1182. doi: 10.1126/science.adj1182. Epub 2024 Jul 19. PMID: 39024449.

2. Given the complexity of the Transformer architecture, it would be helpful to include sensitivity analyses to show how robust the model's predictions are to different input features, hyperparameter choices, or label definitions.

Reply: We thank the reviewer for raising this important point. To evaluate the robustness of our model given the complexity of the Transformer architecture, we conducted a series of sensitivity analyses.

First, we assessed model sensitivity to input features through ablation experiments. The full model achieved an accuracy of 0.9840. Removing GWAS summary statistics (P-value, beta, standard error, sample size) caused the largest drop in accuracy (0.9634), followed by molecular QTLs (0.9723), positional features (0.9765), and OCR/regulatory annotations (0.9827). These findings indicate that model performance relies on the integration of multiple biologically relevant feature groups rather than any single input type (Pages 7-8, Lines 167-188; Figure S1).

Second, regarding label definition, we applied widely accepted GWAS thresholds, designating SNPs with $P < 5\times10^{-8}$ as positive. This strict criterion ensures well-separated, high-confidence labels for training. Notably, the enhanced performance of the transfer learning model reflects the value of pretraining on a well-powered trait (MDD), which includes a larger number of genome-wide significant loci and provides stronger supervisory signal than migraine. This facilitates faster convergence and improved generalization after fine-tuning. To further evaluate model design, we

compared transfer learning (pretrained on MDD) with a baseline model (trained solely on migraine), confirming that pretraining consistently improved performance as shown in Figure R2（Figure S3 in the revised manuscript）.

Finally, we examined hyperparameter sensitivity by varying the number of training epochs. As shown in Figure R2, the transfer model consistently outperformed the baseline across epochs, suggesting stable performance under reasonable training configurations.

These results collectively demonstrate that InsightGWAS is robust to input feature composition, label definitions, and training parameters. Relevant details and results have been incorporated into the revised manuscript and Supplementary Figures (Pages 7-8, Lines 167-188; Figure S1；Pages 9-10, Lines 243-257; Figure S3).

3. While other psychiatric traits were analyzed via LDSC, the cross-trait GWAS and transfer learning focused exclusively on MDD. The manuscript would benefit from clarifying whether cross-trait meta-analysis was also attempted with other disorders (e.g., bipolar disorder or anxiety), and if so, how those results compared. Providing a stronger rationale for selecting MDD—beyond its high genetic correlation with migraine—would help justify this focus.

Reply: We appreciate the reviewer's insightful comment regarding our focus on MDD for cross-trait analyses. To directly address this point, we conducted additional cross-trait GWAS using MTAG and condFDR with five psychiatric traits beyond MDD—namely ADHD, ASD, ANX, BIP, and SCZ. However, none of these analyses yielded additional genome-wide significant loci beyond those already identified in the original migraine GWAS. In contrast, cross-trait analyses with MDD produced a few novel loci, highlighting its uniquely informative role in this context.

Furthermore, we emphasize that the rationale for prioritizing MDD goes beyond genetic correlation alone. MDD has the largest available GWAS sample size among the psychiatric traits analyzed

(170,756 cases and 329,443 controls), and consequently provides a much richer set of positive variants (approximately 4,600) for modeling purposes. In contrast, ADHD and ASD yielded only 1,428 and 93 positive variants, respectively—both fewer than the 1,834 variants from the migraine GWAS itself. This discrepancy limits their utility in downstream transfer learning and prioritization.

Taken together, these observations support our focus on MDD as the most statistically powered and biologically informative trait for cross-trait GWAS and transfer learning in this study. We have updated the manuscript accordingly to clarify this rationale and empirical support (Page 8, Lines 190-204).

4. While the manuscript positions InsightGWAS as an advancement over DeepGWAS, the comparison lacks sufficient methodological depth. A clearer explanation of how InsightGWAS differs architecturally and conceptually from previous models—and why these changes matter—would make the contribution more distinct.

Reply: We thank the reviewer for the comment. We have clarified these differences and their significance in the revised manuscript (Pages 7-8, Lines 167-188).

Reply to Reviewers' Comments

We sincerely thank the editor and reviewers for their thoughtful and constructive feedback. We have carefully revised the manuscript in response to each comment and believe the changes have substantially strengthened the study. A point-by-point response follows below.

Reviewer #1 (Remarks to the Author):

The authors have done an excellent job addressing my comments.

Response: We are glad that our revisions have addressed your concerns, and we greatly appreciate your support of our work.

Reviewer #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Response: Thank you for your helpful comments and for participating in the Early Career Reviewer program. We appreciate your time and thoughtful feedback.

Reviewer #3 (Remarks to the Author):

1) Performance Comparison with Other Methods

AUCs for Transformer (0.9994) and DeepGWAS (0.9990) are nearly identical, statistically significant with 10k bootstraps, but practically trivial.

For a methods paper, what matters is whether this tiny AUC gain translates into more robust novel discoveries, lower FDR, or better replication rates.

Right now, the performance differences are too small to justify the claim that the Transformer substantially improves over DeepGWAS.

The comparative advantage of InsightGWAS over existing methods is overstated; its performance is extremely close to DeepGWAS, and the added complexity may not be justified without clearer gains in real discovery.

Response: We thank the reviewer for this important observation. We fully agree that the incremental improvement in AUC compared to DeepGWAS is numerically small, and on its own would not justify a claim of major methodological advancement. The central aim of our work, however, is not to introduce a purely methods-focused paper, but rather to leverage an exploratory framework in order to advance migraine locus discovery. Within this context, the most substantial progress relative to the original DeepGWAS study stems from our careful curation and expansion of functional features, which substantially improved the overall quality of variant prioritization. The reliability of these findings was further reinforced by successful replication in the independent migraine GWAS from MVP. Indeed, comparison with the original DeepGWAS results highlights that our feature choices account for much of the gain in discovery power (https://pubmed.ncbi.nlm.nih.gov/36824788/, Figure S3). Once these features were optimized, we acknowledge that the additional benefit of adopting a Transformer architecture over deep neural networks or other classifiers is limited. We have addressed this point directly in the revised Discussion to provide a more balanced perspective on the methodological contribution and to emphasize that the primary value of this work lies in the biological insights gained for migraine (Page 17, Lines 457-472).

2) Transfer Learning vs. Baseline Transformer

Training accuracy is not the right metric; both models already exceed 98–99%.

What matters is generalization: validation/test AUC, calibration, FDR, and especially the number of

novel loci discovered.

The observed instability in the baseline curve may simply reflect hyperparameter choice, not an inherent weakness.

Overall, it does not convincingly show that transfer learning meaningfully improves migraine discovery over a well-tuned baseline Transformer.

Response: We thank the reviewer for this insightful comment and agree with the concerns raised. Our use of transfer learning here should be viewed as an exploratory effort specifically in the context of migraine, where there is both a strong genetic correlation with MDD and the advantage of leveraging the much larger MDD GWAS dataset to pre-train the model. These characteristics made migraine a reasonable setting in which to test the potential value of transfer learning. Nonetheless, we acknowledge that in the present study the performance gains from transfer learning over a baseline Transformer were modest.

We also acknowledge that if extended to other diseases, the benefit of transfer learning may vary considerably depending on whether an appropriate genetically correlated trait exists and whether sufficiently large GWAS data are available for that trait. In some settings transfer learning may provide advantages, while in others a baseline Transformer may perform just as well or even better. For this reason, we provide both models and have made the baseline implementation available alongside InsightGWAS. We have also emphasized this point in the revised Discussion to clarify that the added value of transfer learning should not be assumed universally, but rather depends on disease context, data availability, and trait relationships (Page 17, Lines 457-472).

3) Attention & Feature Importance

Attention maps do not prove nonlinear interaction modeling; they only show weight correlations.

Rigorous interaction tests (e.g. Shapley interaction indices, synthetic data benchmarks) are needed.

Response: We agree with the reviewer's comment that attention maps do not constitute rigorous evidence of nonlinear interaction modeling. In response, we removed the original Figure S1. Feature-wise self-attention heatmap in the Transformer model and replaced it with new figures based on a permutation-based pairwise interaction test. The principle of this analysis is as follows: we first evaluated the model's baseline performance on the validation set, then measured how much predictive accuracy (AUC and AP) dropped when shuffling the values of individual features, and finally repeated the procedure while shuffling two features simultaneously. To quantify interaction, we compared the paired drop with the sum of the two individual drops. If the joint perturbation caused a greater decrease than expected from the sum of the individual effects, we interpreted this as evidence of a nonlinear, non-additive interaction. This approach provides a direct and computationally tractable test of interaction effects, and serves as a practical alternative to Shapley interaction indices or synthetic data benchmarks, which, while theoretically rigorous, are prohibitively slow to compute on genome-scale data. The results, now shown in Supplementary Figures S1, reveal that the strongest interactions occur among GWAS statistical features (β , SE, p-value, and LD score), while regulatory annotations contribute mainly marginal effects. This confirms the reviewer's observation: although the Transformer is capable of capturing some nonlinear feature interactions, the dominant predictive signal remains concentrated in GWAS-derived features. We also revised the manuscript accordingly (Pages 7-8, Lines 176-192).

Reported metric is accuracy, which is misleading under class imbalance. AUC or PR-AUC would be

more convincing.

Response: We thank the reviewer for raising this important point regarding performance metrics under class imbalance. In accordance with the reviewer’s suggestion, we have now incorporated AUC results alongside accuracy for all ablation experiments:

All Features: AUC = 0.9990 | Accuracy = 0.9840

No Group 1 (Position): AUC = 0.9968 | Accuracy = 0.9765

No Group 2 (GWAS Stats): AUC = 0.9869 | Accuracy = 0.9634

No Group 3 (QTL): AUC = 0.9987 | Accuracy = 0.9723

No Group 4 (OCR & Regulatory): AUC = 0.9986 | Accuracy = 0.9827

We have updated the manuscript accordingly (Page 8, Lines 192-199).

Little biological interpretation is provided (e.g. do the high-attention features align with migraine-relevant biology?).

Response: We thank the reviewer for this helpful comment. In our feature set, the variables that most directly reflect migraine biology are the GWAS summary statistics themselves—namely the odds ratios (OR) and P-values—since these are derived directly from migraine GWAS and represent the statistical evidence of association with the disease. It is therefore expected that these features receive high attention weights, as they anchor the model to migraine-specific signals. Other annotations, such as eQTLs and sc-eQTLs, play an auxiliary role by adding biological context. They help the model discriminate between variants with borderline statistical significance that may arise by chance and those more likely to have true functional relevance. This is particularly valuable when regulatory

annotations are observed in brain tissues or neuronal cell types, where they provide additional support for biological plausibility.

We have clarified this point in the revised manuscript (Pages 7-8, Lines 176-199, Pages 17, Lines 457-465), emphasizing that while the model is necessarily p-value driven, the integration of functional annotations refines prioritization by filtering out random signals and highlighting variants that are both statistically supported and functionally plausible.

While annotations add some predictive value, the evidence that Transformers uniquely capture nonlinear interactions (vs. DeepGWAS or simpler models) is weak. The model still seems heavily p-value driven.

Response: We agree with the reviewer that the model appears to be strongly influenced by p-values. This outcome is expected and appropriate, given that our framework is fundamentally built on GWAS summary statistics. The purpose of InsightGWAS is precisely to detect true associations that fall below the conventional genome-wide significance threshold. Such variants, while not reaching full significance, necessarily exhibit at least some degree of p-value enrichment. In this context, p-values provide the primary statistical signal, while functional annotations and other features serve a complementary role—helping to distinguish between variants with borderline significance arising by chance and those more likely to be biologically meaningful. We have clarified this point in the revised manuscript to better align expectations about the relative contributions of p-values and annotations in our model (Pages 7-8, Lines 176-192).

Reviewer #4 (Remarks to the Author):

1. The inclusion of the MVP migraine GWAS as an independent dataset meaningfully strengthens the validation of the model. The reported replication rate is reasonable and provides supportive evidence

of robustness.

Response: We appreciate your positive evaluation of the replication using the MVP migraine GWAS dataset.

2. The ablation experiments, label definition assessments, and hyperparameter tests are clear and informative. These results collectively support the robustness of the model under various conditions.

Response: Thank you for acknowledging the clarity and informativeness of our model robustness assessments.

3. The additional cross-trait analyses and quantitative comparisons convincingly explain the choice of MDD as the primary trait for cross-trait GWAS and transfer learning. The rationale is well presented.

Response: We are grateful for your recognition of the rationale behind the choice of MDD for cross-trait analysis.

4.The clarification of architectural and conceptual differences improves the manuscript's positioning. The distinctions from DeepGWAS are now sufficiently outlined for the reader.

Response: We thank you for noting the improved clarity regarding the distinctions from DeepGWAS.