EFGPP: Exploratory framework for genotype-phenotype prediction

Muhammad Muneeb[1,2] and David B. Ascher[1,2,*]

[1]School of Chemistry and Molecular Biology, The University of Queensland, Brisbane, 4067, Australia
[2]Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, 3004, Australia
[*]Correspondence: d.ascher@uq.edu.au

SUMMARY

Genotype-phenotype prediction requires diverse data, including genetic variants, functional annotations (FA), linkage disequilibrium, polygenic risk scores (PRS), covariates, and genome-wide association studies (GWAS). We present a framework to optimize data integration for phenotype prediction. Using UK Biobank genotypes (733 samples), 2 GWAS for migraine, 3 for depression, publicly available FA, and 4 PRS tools, we tested various combinations of datasets for migraine prediction. The best individual dataset performance achieved a test AUC of 0.64 (±0.14). When different combinations were formed—configuration 1 (migraine-related data sources) and configuration 2 (migraine and depression-related data sources)—the combination of covariates, PRS-PLINK, and weighted-annotated genotype data in configuration 1 achieved a test AUC of 0.69 (±0.13). In configuration 2, combining unweighted-annotated genotype data and PRS-LDAK achieved a test AUC of 0.66 (±0.04). We observed that the inclusion of PRS, covariates, PRS from AnnoPred and LDAK, and annotated genotype data improves prediction performance.

KEYWORDS

genotype-phenotype prediction, genetics, gwas, machine learning, polygenic risk scores

INTRODUCTION

Genotype-phenotype prediction involves predicting traits and diseases based on genetic datasets[1,2]. It has diverse applications, including understanding disease mechanisms[3], personalized medicine[4,5], case-control classification, disease treatment responses[6], and exploring the interplay between genetic diseases[7,8].

Data from various biological, environmental[9], and phenotypic dimensions are gathered and incorporated into a coherent framework for robust predictive modeling[10]. These data sources include, but are not limited to, genotype data obtained by selecting top single nucleotide polymorphisms (SNPs) using p-value thresholding[11] from genome-wide association studies (GWAS). Covariates encompass medical conditions, metabolite levels, sex, and other demographic or clinical variables that may influence phenotypic expression[12]. Principal component analysis (PCA) is applied to the genotype data to account for population stratification and reduce dimensionality[13]. Functional annotations (FA) provide insights into the biological relevance of genetic variants, assisting in prioritizing SNPs for prediction models[14]. GWAS offers summary statistics highlighting associations between genetic variants and traits, serving as a foundation for identifying significant SNPs[15]. Polygenic risk scores (PRS)[16,17], derived from GWAS data, aggregate the effects of multiple genetic variants to estimate an individual's genetic predisposition to specific traits or diseases. These data sources are combined to create a feature set, and machine learning (ML)

or deep learning (DL) algorithms are employed for phenotype prediction[10,18]. Researchers have further enhanced prediction performance by integrating multiple GWAS[19,20] for one phenotype, leveraging GWAS from related phenotypes[21] and populations[22], and combining multiple PRS[23]. For instance, combining GWAS data from migraine and depression enables the development of multi-trait PRS, capturing shared genetic architectures between related conditions.

To address the wide variety of datasets and methodologies for integration, we propose an exploratory framework that (1) generates individual datasets for covariates, genotype data (weighted by GWAS effect sizes or unweighted, with or without functional annotations), PRS (using PLINK[24], PRSice-2[25], AnnoPred[26], LDAK[27]), and PCA; (2) selects the most effective individual datasets based on training/validation performance and stability; and (3) systematically combines datasets across various categories to construct a composite dataset, identifying combinations that enhance predictive power for case-control classification. The framework supports multiple GWAS from single or related phenotypes/populations, integrates FA, and includes PRS from various tools, allowing genetic and non-genetic factors inclusion to improve the performance of genotype-phenotype predictions. Figure 1, shows the overall workflow of EEGPP.

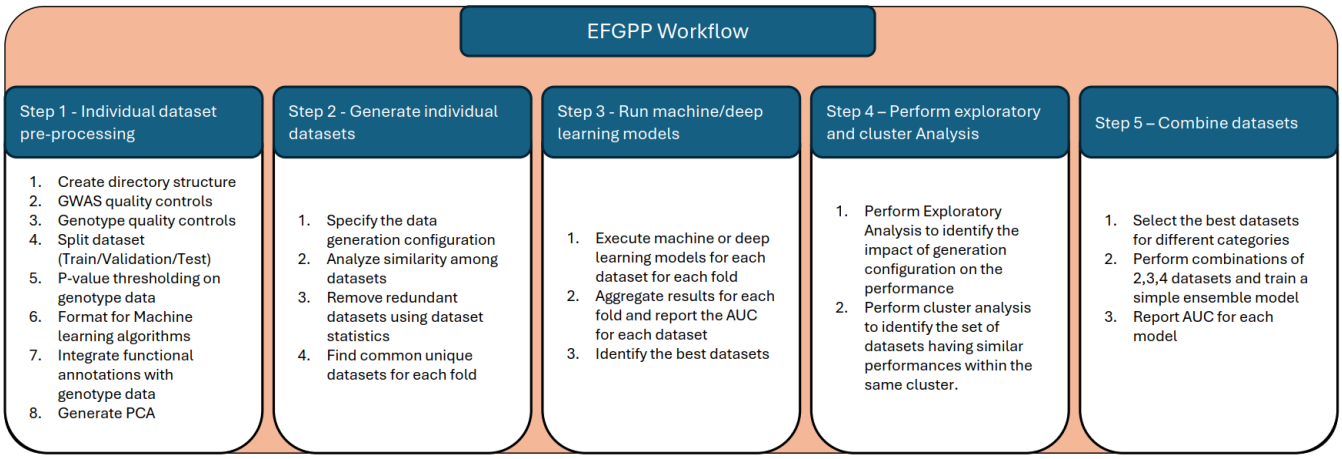

Figure 1: **An overview of EFGPP workflow.** The analysis pipeline comprises several steps, which can be categorized into five steps. **Step 1 - Individual Dataset Pre-processing** involves creating a directory structure, applying GWAS quality controls, conducting genotype quality controls, splitting the dataset into training, validation, and testing sets, applying p-value thresholding on genotype data, formatting the data for ML algorithms, integrating FA with genotype data, and generating PCA. **Step 2 - Generate Individual Datasets** focuses on defining the data generation configuration, analyzing similarities among datasets, eliminating redundant datasets, and identifying common unique datasets for each fold. **Step 3 - Run Machine/Deep Learning Models** includes executing ML/DL models on each dataset across every fold, aggregating results for each fold, and reporting the AUC for each dataset. **Step 4 - Perform Exploratory and Cluster Analysis** involves conducting exploratory analysis to assess the influence of the generation configuration on performance and executing cluster analysis to identify the set of datasets exhibiting similar performance within the same cluster. Finally, **Step 5 - Combine Datasets** includes selecting the best datasets for different categories, generating combinations of 2, 3, and 4 datasets to train a simple ensemble model, and reporting the AUC for each model.

Material and Methods

Dataset

We considered six base genetic datasets for this study: GWAS, Genotype, Covariates, PRS, LD, and FA (Figure 2). The GWAS data was obtained from the GWAS Catalog (`https://www.ebi.ac.uk/gwas/`). Genotype data and covariates were sourced from the UK Biobank. PRS were generated using PRS tools. LD data were downloaded from LDAK (`https://dougspeed.com/reference-panel/`). FAs were retrieved from AnnoPred (`https://github.com/yiminghu/AnnoPred`).

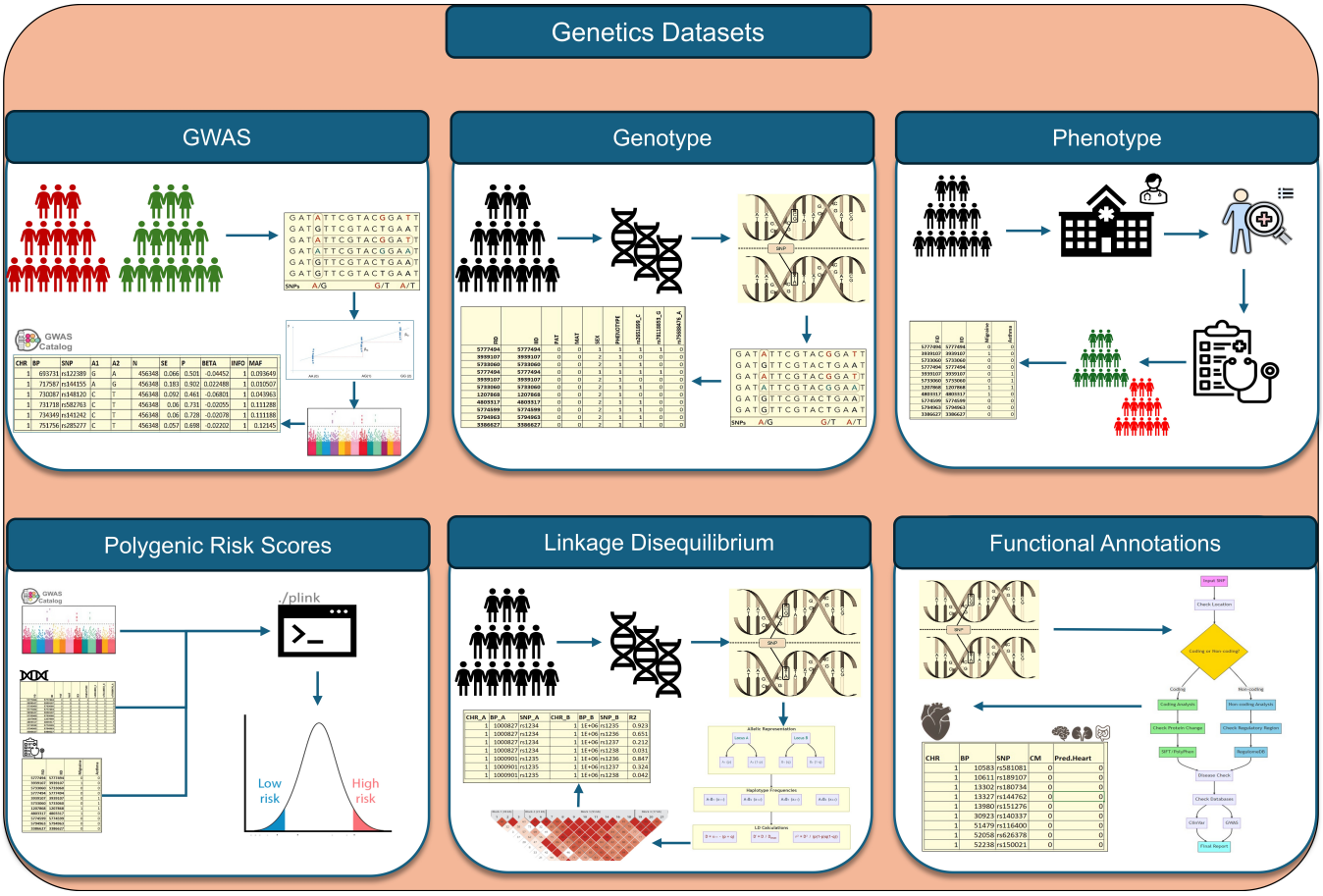

Figure 2: **An overview of genetic datasets considered in this study.** (1) GWAS analyzes genetic variations across large populations to identify SNPs associated with specific traits or diseases. After quality control measures, statistical analysis (e.g., linear regression for quantitative traits) identifies significant associations, often visualized in Manhattan plots. Information for all SNPs is stored in a format containing Chromosomes, Base pairs, SNP ID, P-values, Effect size/odds ratio, Reference allele, Alternative allele, and Minor allele frequency. (2) DNA is sequenced using platforms like Illumina or Oxford Nanopore, and variants are identified and saved for each sample in a tabular format for future analysis. (3) Phenotype data—such as clinical traits—are collected through standardized medical assessments and integrated with genetic data. (4) PRS aggregates these genetic associations to estimate an individual's risk for complex diseases. (5) LD analysis explores the non-random association of genetic variants, aiding in the identification of causal variants. (6) Functional annotation links genetic variants to biological processes to assess the impact of variants on gene function.

Modeling

We retrieved GWAS files for migraine (2) and depression (3) from the GWAS catalog using GWASPokerforPRS (`https://github.com/MuhammadMuneeb007/GWASPokerforPRS`). The corresponding accession IDs and download links can be found on GitHub (GWASFiles). GWAS files were subjected to quality controls, retaining SNPs with Minor Allele Frequency (MAF) $>$ 0.01 and Imputation Information Score (INFO) $>$ 0.8, while removing ambiguous and missing SNPs[28,29]. Genotype data were extracted from the UK Biobank, which included 733 participants reporting comorbid conditions (hypertension, asthma, depression, osteoarthritis, high cholesterol, irritable bowel syndrome, hypothyroidism, hay fever, migraine, and gastric reflux) and 135 Nuclear Magnetic Resonance (NMR) metabolomic biomarkers. Comorbid conditions and metabolic biomarkers level were utilized as covariates after excluding migraine and depression. Phenotype data and corresponding data sources (Covariates and genotype data) were split using stratified folds (5) to maintain the same ratio and order of cases and controls across the train (80%), validation (10%), and test (10%) sets for both phenotypes. Quality control was performed on all training sets for each GWAS and genotype combination, with thresholds for MAF of 0.01, Hardy-Weinberg Equilibrium (HWE) of 1×10^{-6}, Genotype Missingness of 0.1, Individual Missingness of 0.1, and a Relative Cutoff of 0.125[30,31]. For all GWAS files, we processed genotype data separately. For example, the data were processed for various combinations, including "migraine" (genotype data) with "migraine.gz" (GWAS file) and "depression" (genotype data) with files like "depression_11.gz," "depression_4.gz," and others (GWAS files). Top (50, 100, 500, and 1000) SNPs were selected using p-value thresholds[32] on genotype data, and annotations from AnnoPred were appended to the genotype data, creating two genotype matrices—one with annotations and one without[11]. Using PLINK, PCA (first 10 components) was calculated using genotype data. For PRS calculations, tools like PLINK, PRSice-2, AnnoPred, and LDAK were used (`https://muhammadmuneeb007.github.io/PRSTools/Introduction.html`). Depending on the tool input requirements, the PRS model was fitted using genotype data, GWAS, covariates, LD, FA, and PCA. PRS scores were generated using log distributed p-values ranging from a minimum p-value of 1×10^{-10} to 1, with 200 intervals, resulting in 200 risk scores for each sample. Some PRS tools (AnnoPred, LDAK) produced new betas/effects for each SNP, and SNPs in the genotype matrix were multiplied by these betas, resulting in four genotype matrices for each set of top SNPs. Data from all sources for one GWAS file is categorized as 1) Genotype data, 2) Genotype data with functional annotations, 3) Genotype data weighted by the GWAS file betas, 4) Genotype data weighted by the GWAS file betas and functional annotations, 5) Covariates, 6) PCA results, and 7) PRS for each tool. These categories represent the possible inputs that can be used for training models in genotype-phenotype prediction (Figure 3).

The number of base datasets generated is calculated using Equation 1.

$$D_{\text{total}} = \sum_{r \in C} \binom{|P|}{r} \prod_{i=1}^{r} (2|S| + |S||N|(|W_{\text{files}}| + 1) + |M|) \tag{1}$$

where:

$C =$ Number of combinations
$r =$ Number of datasets in each combination
$|P| =$ Number of Phenotype-GWAS pairs
$|S| =$ Number of scaling options (True/False)
$|N| =$ Number of top SNPs and annotated SNPs genotype matrix
$|W_{\text{files}}| =$ Number of PRS models $\times$ Number of Phenotype-GWAS pairs
$|M| =$ Number of PRS models

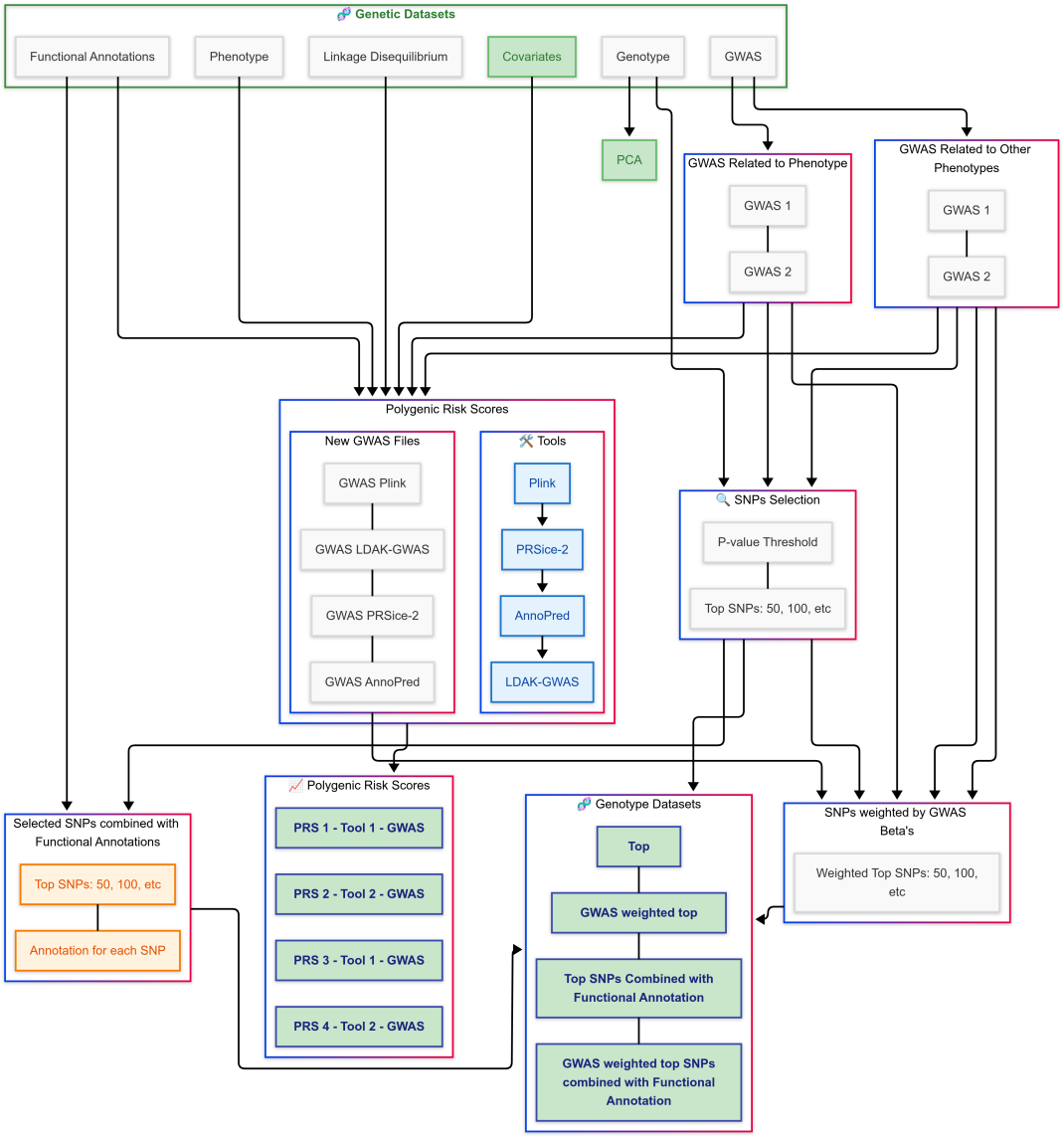

Figure 3: **This diagram illustrates a comprehensive flowchart for generating base datasets for predicting one phenotype**. The process begins with various genetic datasets, including genotype data, GWAS files, phenotype information, FA, LD, and covariates. Significant SNPs are selected based on p-value thresholds for each GWAS and Phenotype pair. PCA is performed on each genotype dataset, which, in our case, remains consistent across all datasets. PRS for each pair are generated using tools such as AnnoPred, PLINK, PRSice-2, and LDAK. Subsequently, weighted, unweighted, annotated, and unannotated genotype data are generated for each pair. The selected datasets can be combined into different configurations, including two-component (e.g., PRS + Covariates, PCA + SNPs) and three-component (e.g., PRS + PCA + Covariates) combinations, as well as a complete model that integrates PRS, PCA, SNPs, and Covariates. The workflow highlights systematically exploring genetic and non-genetic factors for phenotype prediction.

Parameters for Configuration 1: This configuration involves using only migraine GWAS files.

$$
\begin{aligned}
|P| &= 2 \text{ (migraine with 2 GWAS files)} \\
|S| &= 1 \text{ (False only)} \\
|N| &= 9 \text{ (all SNP options)} \\
|W_{\text{files}}| &= 4 \times 2 = 8 \text{ (PRS models} \times \text{Phenotype-GWAS pairs)} \\
|M| &= 4 \text{ (PLINK, PRSice-2, AnnoPred, LDAK-GWAS)}
\end{aligned}
$$

Calculation for $r = 1$:

$$
\begin{aligned}
D_{\text{total}} &= \sum_{r \in \{1\}} \binom{2}{r} \prod_{i=1}^{r} (2 \cdot 1 + 1 \cdot 9 \cdot (8 + 1) + 4) \\
&= \binom{2}{1} \cdot 87 \\
&= 2 \cdot 87 \\
&= 174
\end{aligned}
$$

Parameters for Configuration 2: For the second configuration with both migraine and depression GWAS files.

$$
\begin{aligned}
|P| &= 5 \text{ (2 migraine + 3 depression pairs)} \\
|S| &= 1 \text{ (False only)} \\
|N| &= 9 \text{ (all SNP options)} \\
|W_{\text{files}}| &= 4 \times 5 = 20 \text{ (PRS models} \times \text{Phenotype-GWAS pairs)} \\
|M| &= 4 \text{ (PLINK, PRSice-2, AnnoPred, LDAK-GWAS)}
\end{aligned}
$$

Calculation for $r = 1$:

$$
\begin{aligned}
D_{\text{total}} &= \sum_{r \in \{1\}} \binom{5}{r} \prod_{i=1}^{r} (2 \cdot 1 + 1 \cdot 9 \cdot (20 + 1) + 4) \\
&= \binom{5}{1} \cdot 195 \\
&= 5 \cdot 195 \\
&= 975
\end{aligned}
$$

For Configurations 1 and 2, 175 and 975 individual datasets were generated, respectively. Since executing ML/DL models for 2^{175} and 2^{975} combinations of datasets is not feasible, we reduced the dataset count by removing duplicates using the Kolmogorov-Smirnov statistic[33] (i.e., `similarity = stats. ks_2samp(data1.flatten(), data2.flatten()).statistic`). We retained the first dataset if the similarity score was equal to 1. As a result, 62 unique datasets were retained for Configuration 1, and 489 unique datasets for Configuration 2.

For each dataset, 12 different classification ML algorithms were used, ranging from basic models (Decision Trees, Logistic Regression) to more advanced ensembles (Random Forest, XGBoost, LightGBM, CatBoost) and neural network architectures. All models were optimized via hyperparameter tuning and configured to handle class imbalance, and results were aggregated per fold (details are provided on GitHub (CoreML.py)). Additionally, one can use deep learning architectures ranging from simple feed-forward networks (FNN) and LSTM to 1D convolutional networks (1D-CNN), transformer-based models, graph neural networks, tabular data models

(TabNet, DeepFM), and hybrid architectures. These deep learning models, optimized for binary classification tasks with techniques like batch normalization, dropout, and early stopping, are available on GitHub (CoreDL.py). Further analysis was performed using ML models.

Results for each dataset were aggregated over each fold, and datasets with validation performance exceeding 0.6 were retained (though a different threshold can be chosen to reduce the number of datasets for further analysis). Using ML models, 49 and 355 datasets passed in Configurations 1 and 2, respectively.

We systematically classified the retained datasets into distinct categories to minimize combinations. The categories were as follows: (1) Covariates, (2) PCA, (3) Genotype_WeightStatus_AnnotationStatus _GWASFile, and (4) PRS_GWAS File, resulting in 1, 1, 2×2×2, and 1×2 datasets (Total = 12), respectively. Using the category, PRS_PRSTOOL_GWASFile would result in 1×4×2 = 8 datasets (Total = 18). Depending on the categories to explore further, one can create different categories. The datasets from each category were ranked using a composite score 2, and the top datasets were retained for further analysis. The composite score is calculated as:

$$\begin{aligned}\text{Composite\_Score} = {} & 0.25 \cdot \text{Validation AUC (Normalized)} + \\ & 0.25 \cdot \text{Train-Validation Gap (Normalized)} + \\ & 0.25 \cdot \text{Train Stability (Normalized)} + \\ & 0.25 \cdot \text{Validation Stability (Normalized)}\end{aligned} \tag{2}$$

Based on these composite scores, the top datasets from each category were selected, and finally, we trained and tested all combinations of 2, 3, 4, and 5 datasets. Each dataset combination was trained using a simple deep-learning model within a stacked ensemble learning framework. At the first level, simple neural networks with two dense layers and dropout for regularization were created for each dataset. These base models were trained independently, utilizing early stopping and class weight balancing to prevent overfitting. At the second level, a meta-model was used to combine the predictions from all base models to generate final predictions. The results were then merged for each fold. We applied selection criteria for each dataset combination as specified in equation 2 to identify the best combinations.

Genotype-phenotype prediction is modeled using the following framework (Equation 3):

$$Y = f\left(\alpha_{\text{PRS}} \cdot \text{PRS} + \alpha_{\text{PCA}} \cdot \text{PCA} + \alpha_{\text{SNPs}} \cdot \text{SNPs} + \alpha_{\text{Cov}} \cdot \text{Covariates}\right) \tag{3}$$

In this model, Y represents the phenotype, which could be a disease status or a quantitative trait. The function f (ML or DL model) defines the relationship between the predictors and the outcome, where the predictors include multiple PRS, PCA, individual SNPs, and covariates. The coefficients α represent whether a specific dataset is being selected or not. This framework allows for including PRS from multiple tools and modified genotype data, enabling a more comprehensive prediction by integrating genetic and non-genetic factors.

RESULTS

We generated datasets for configuration 1 (175 datasets) and configuration 2 (975 datasets). After removing duplicates, 62 unique datasets remained for configuration 1 and 489 for configuration 2. The dataset parameters and a list of unique datasets (UniqueDatasets.csv) are available on GitHub (Configuration1/Datasets.csv, Configuration2/Datasets.csv). For each dataset, we used 12 ML models, and the results for each unique dataset are available at (Configuration1/ResultsIndividualDataset.csv) for Configuration 1 and (Configuration2/ ResultsIndividualDataset.csv) for Configuration 2. For each dataset, 12 models were used. The results were

summed for each fold, and among the 12 models, the one with the highest composite score was selected. This process yields one optimal model for each dataset.

The optimal test performance for Configuration 1 and 2 was using weighted genotype type (432 features), the migraine.gz GWAS file with 50 annotated SNPs (snps_annotated_50), and the Logistic Regression model. The model demonstrated a training AUC of 0.655758 (±0.02147), validation AUC of 0.637858 (±0.086888), and test AUC of 0.643899 (±0.143235). The best performance in terms of stability for Configuration 1 was using PRS (125 features), the migraine.gz GWAS file with LDAK-GWAS, and the Naive Bayes ML model. The model achieved a training AUC of 0.670031 (±0.027755), validation AUC of 0.638314 (±0.161106), and test AUC of 0.575402 (±0.028343). For Configuration 2, multiple datasets generated stable performance. The model using PRS (125 features), the migraine.gz GWAS file with LDAK-GWAS, and Naive Bayes achieved a training AUC of 0.670031 (±0.027755), validation AUC of 0.638314 (±0.161106), and test AUC of 0.575402 (±0.028343). The model using genotype data with the depression_17.gz GWAS file with 5000 unweighted SNPs demonstrated a training AUC of 0.733086 (±0.019131), validation AUC of 0.592498 (±0.121865), and test AUC of 0.541196 (±0.070488).

We conducted a cluster analysis using the training dataset to identify similar datasets, analyze their distribution, and uncover hidden patterns. For each dataset, we calculated the following features: mean (μ), standard deviation (σ), skewness, kurtosis, quartiles (Q25, Q50, Q75), interquartile range (IQR), and PCA variance ratio based on the first two principal components. The data was subsequently processed through dimensionality reduction techniques, including PCA, t-distributed Stochastic Neighbor Embedding (t-SNE)[34], and Uniform Manifold Approximation and Projection (UMAP)[35]. Afterward, K-means clustering with five clusters (k=5) was applied to the training data for fold 1 in Configuration 1 (Figure 4), although this approach can be extended to other folds and Configuration 2.

Each clustering method revealed distribution differences among the datasets; however, no clear pattern linking specific clusters with high test AUC values emerged. While clustering helps focus subsequent analyses on individual clusters, we performed further evaluations using all datasets.

For each Configuration, exploratory data analysis was performed to reveal hidden patterns and assess the impact of data generation parameters on the performance of the individual datasets. For Configurations 1 and 2, the results are shown in Figures 5 and 6.

For Configuration 1, datasets were divided into different categories. We used two GWAS files for Configuration 1: GWAS1 (migraine.gz) and GWAS2 (migraine_5.gz). Among the datasets using migraine.gz, the best-performing one was selected using a composite score selection criterion. GWAS1 (migraine.gz) achieved better AUC metrics (training: 0.95, validation: 0.75, test: 0.63) compared to GWAS2 (migraine_5.gz). The dataset performance assessment showed distinct AUC score distributions across different data modalities. Performance was constant for the covariates and PCA datasets (given that only one dataset was available for each category). In contrast, datasets based on genotype data and PRS showed variability, indicating their potential for enhanced performance. While the use of weight files on genotype data did not substantially impact performance overall, the weighted genotype datasets did improve prediction accuracy for some cases. The training and test performances were comparable for genotype datasets with 50 SNPs and 50 annotated SNPs, suggesting these datasets provide a generalized performance. For all other datasets, the test performance remained nearly identical. Finally, among the evaluated models—except for PRSice-2—the test performance was generally low. Of the machine learning models used, logistic regression demonstrated the highest stability for the training and validation datasets. Collectively, this analysis indicates that for Configuration 1, the combination of GWAS1 (migraine.gz), genotype-based data with 50 SNPs and 50 annotated SNPs, the presence of weight files, the PRSice-2 model, and the Logistic Regression model yield a more

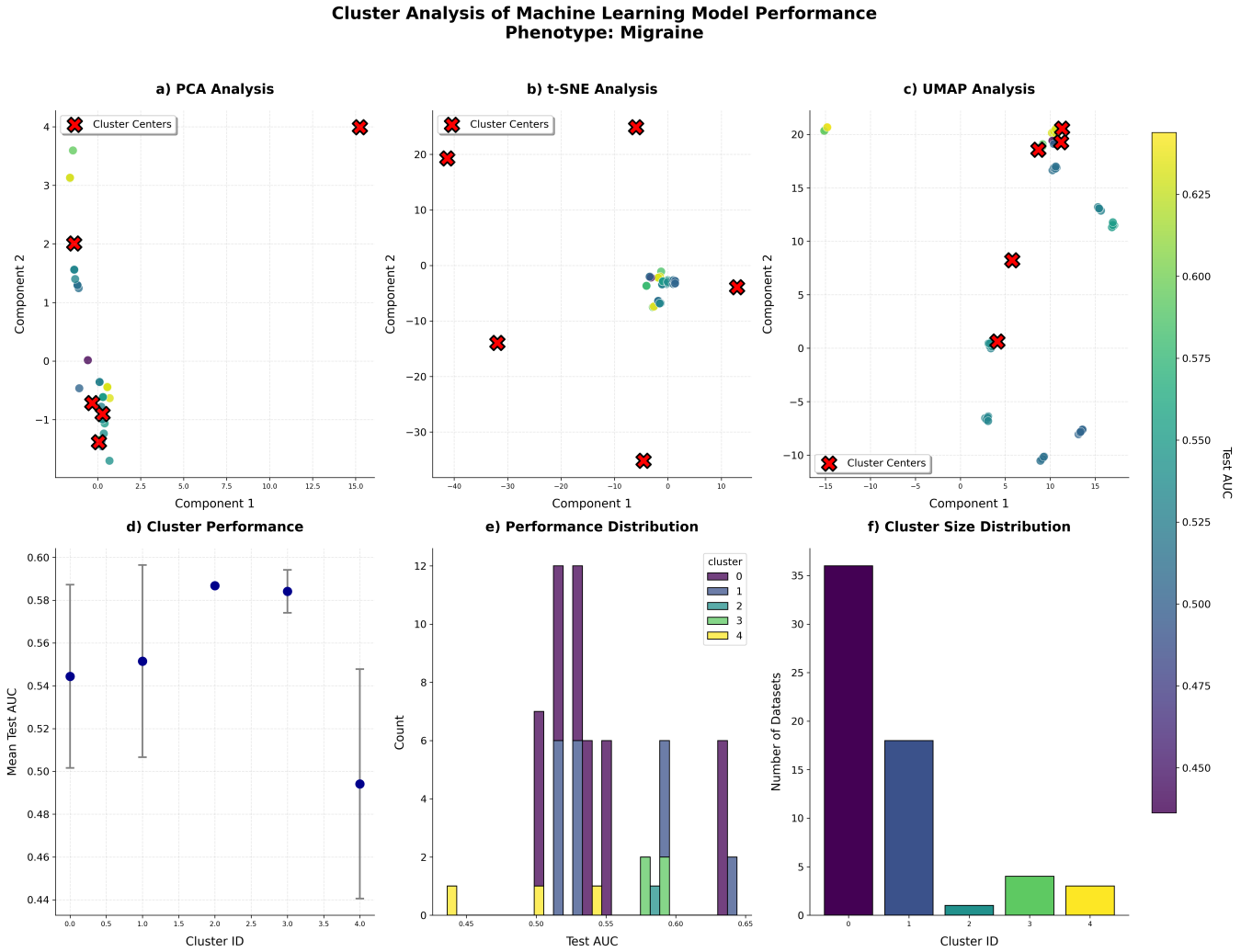

Figure 4: **Overview of Cluster Analysis for Configuration 1.** K-means clustering analysis with 5 clusters (k=5) was carried out on the training data for fold 1 in configuration 1, but it can also be applied to other folds and configuration 2. The first three panels show the PCA, t-SNE, and UMAP analyses, with cluster centers indicated by red crosses in the two-dimensional space. These plots demonstrate distinct groupings within the datasets and varying performance levels, ranging from 0.450 to 0.625. Quantitative assessments indicate that Cluster 3 achieved the highest mean test AUC of approximately 0.58, while Cluster 5 exhibited lower performance, around 0.49. Lastly, the distribution of cluster sizes indicated a noticeable imbalance, as Cluster 1 contained the majority of datasets (around 35, mostly genotype datasets), in contrast to Clusters 3 (covariates), 4, and 5, which encompassed only 2–4 datasets each. These observations highlight the heterogeneity of the datasets in configuration 1.

generalized performance (Figure 5).

For Configuration 2, all datasets were divided into different categories based on the GWAS files used. The following GWAS files were used: GWAS1 (migraine.gz), GWAS2 (migraine_5.gz), GWAS3 (depression_11.gz), GWAS4 (depression_17.gz), and GWAS5 (depression_4.gz). Among the datasets incorporating PRS, genotype data, and PCA, the best-performing dataset was selected using a composite score selection criterion. GWAS1 (migraine.gz) achieved the best AUC metrics (training: 0.95, validation: 0.75, test: 0.63) compared to all other GWAS files, which showed similar train, validation, and test performances. The dataset performance assessment demonstrated distinct AUC score distributions across different data modalities. Performance was

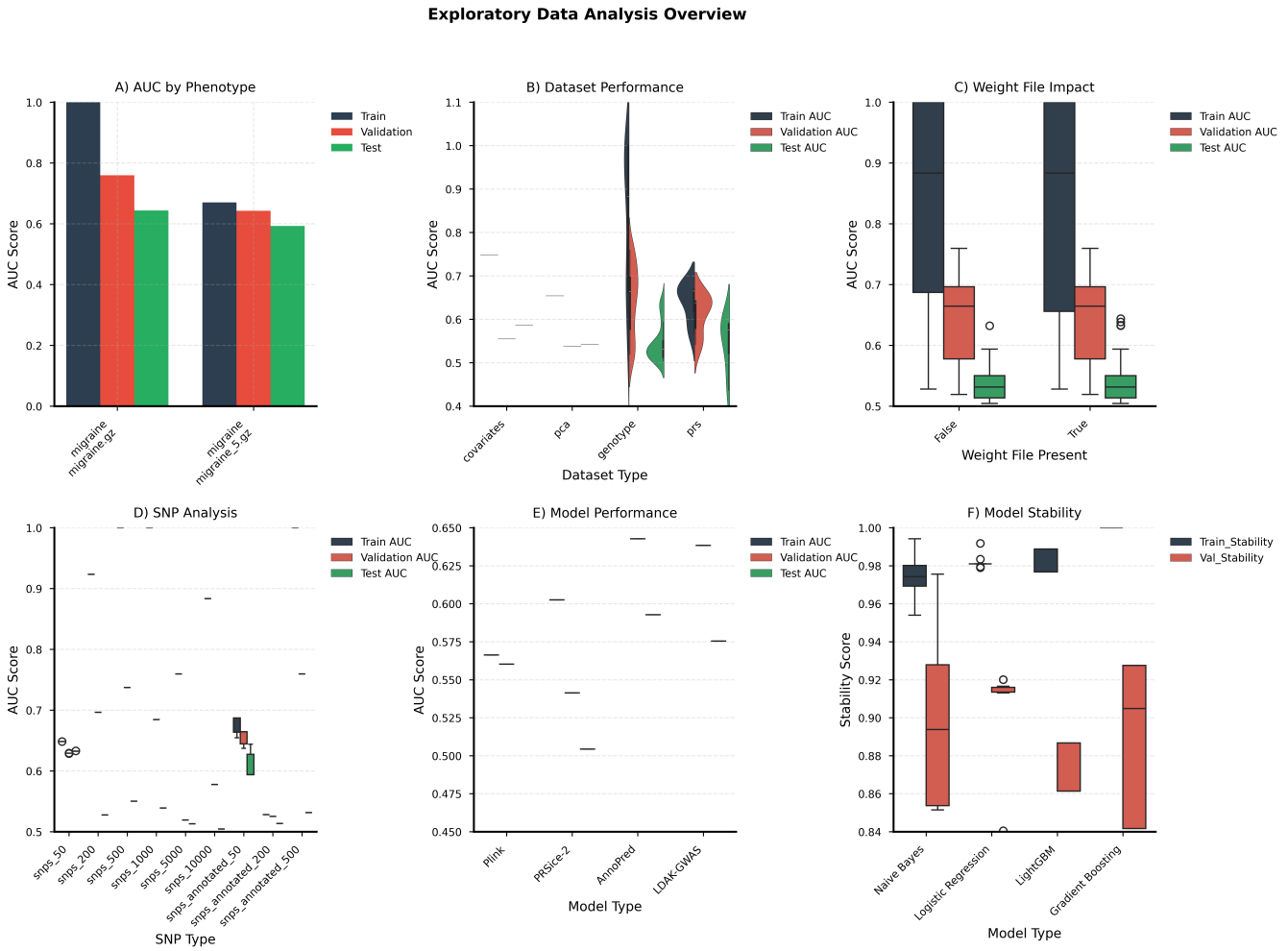

Figure 5: **An overview of the impact of data generation parameters on training, validation, and test performance for Configuration 1.** A) Shows the dataset's best performance for each GWAS. B) Comparative analysis of performance across different dataset types, featuring violin plots that illustrate the distribution of AUC scores for each dataset category. C) Impact assessment of weight file incorporation on model performance, showing how the presence of a weight file affects prediction AUC. D) Performance analysis based on the number of SNPs, highlighting the relationship between various SNPs (genotype features) and performance. E) Comparison of PRS models' performance, illustrating the distribution of AUC scores across different PRS models. F) Stability analysis of machine learning models across all datasets. This analysis was conducted on the top models for each dataset, which were chosen based on the highest composite score.

constant for covariates (with only one dataset available in this category). In contrast, datasets based on genotype data, PCA, and PRS displayed variability. Although PCA performance was expected to be consistent, differences in clumping and pruning operations when using different GWAS files may lead to variations in PCA computations. Weighted genotype datasets improved AUC for training, validation, and test datasets. Moreover, genotype datasets with 50 SNPs and 50 annotated SNPs showed comparable training and test performances, suggesting these configurations provide generalized performance. It is important to note that when using genotype data for depression, the performance for annotated 50 SNPs decreased, as inferred from the comparative plot (D) in Figure 5. The PRS models exhibited interesting behavior: when using PRS derived from depression for migraine prediction, performance increased, suggesting that

depression PRS may be helpful for migraine prediction[36]. The neural network demonstrated the highest and most consistent stability between training and validation datasets among the machine learning models evaluated. These analyses indicate that for Configuration 2, the combination of GWAS1 (migraine.gz), genotype-based data with 50 SNPs and 50 annotated SNPs, the use of weight files, PRS models derived from both depression and migraine, and the neural network model yields better performance (Figure 6).

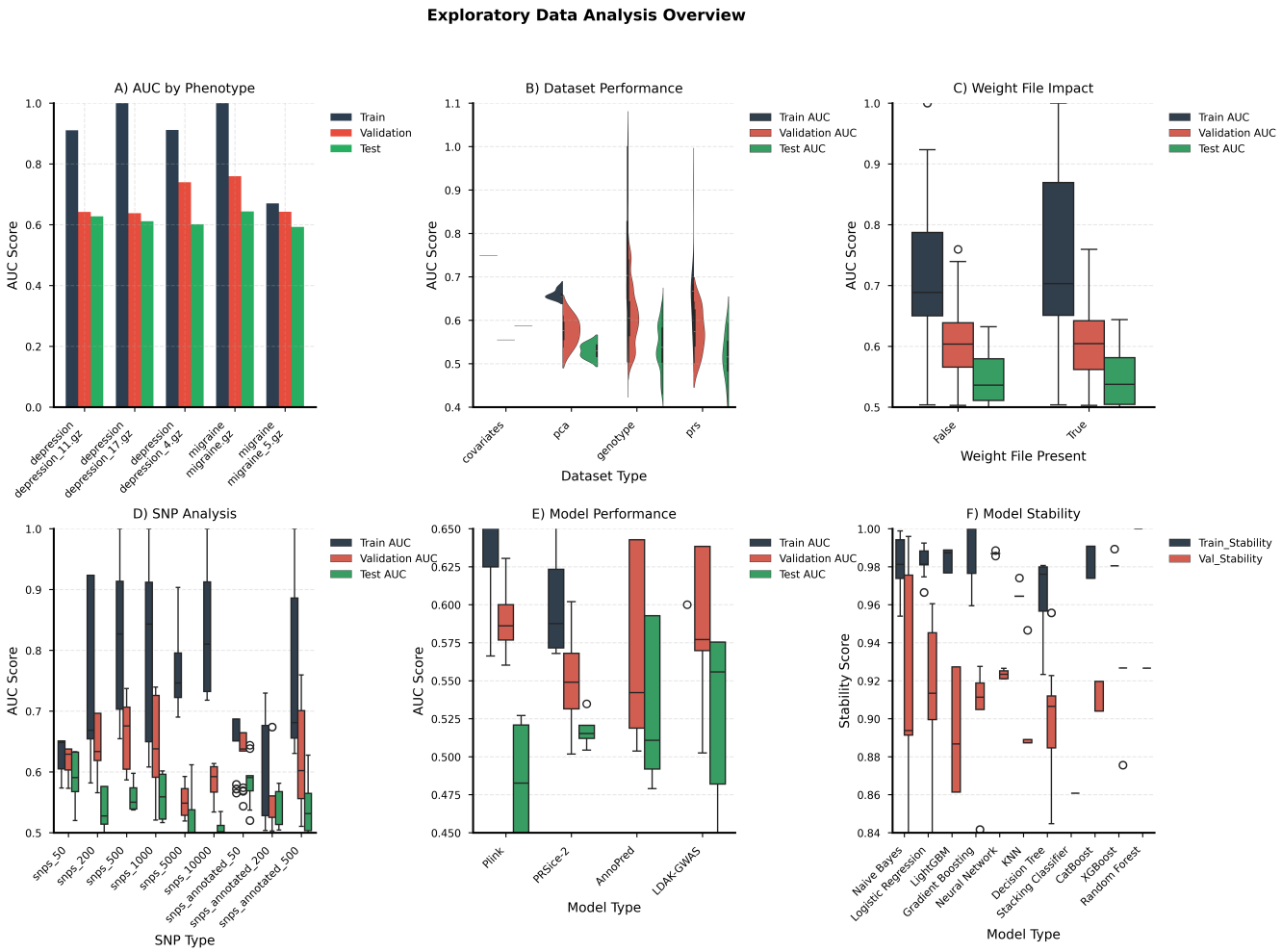

Figure 6: **An overview of the impact of data generation parameters on training, validation, and test performance for configuration1.** A) Displays the best performance of all GWAS. B) Comparative analysis of performance across different dataset types, featuring violin plots that illustrate the distribution of AUC scores for each dataset category. C) Impact assessment of weight file incorporation on model performance, showing how the presence of a weight file affects prediction AUC. D) Performance analysis based on the number of SNPs, highlighting the relationship between various SNPs (genotype features) and performance. E) Comparison of PRS models' performance, illustrating the distribution of AUC scores across different PRS models. F) Stability analysis of machine learning models across all datasets. All analyses were conducted on the best models for each dataset, which were selected based on the best composite score.

We further reduced the number of datasets by applying a validation AUC threshold of > 0.6 and ranked the remaining datasets based on the composite score. The remaining datasets, which comprise approximately 49 and 355 datasets for Configurations 1 and 2, respectively, are available on GitHub (Configuration1/ResultsTop10.csv and Configuration2/ResultsTop10.csv).

Finally, we selected datasets based on different categories (including covariates, PCA, genotype data, and PRS) to form the combinations of the datasets. These categories were defined based on dataset generation parameters. For Configuration 1, we selected 11 datasets, while for Configuration 2, we selected 16. For both configurations, dataset selection was guided by composite score ranking and a validation threshold of > 0.60. The selected categories included covariates, genotype data (both weighted and unweighted), annotated and unannotated datasets (irrespective of the GWAS file), PCA, and PRS (PRS calculated using GWAS files and PRS tools). The selected datasets for Configuration 1 and Configuration 2 are shown in Table 1 and on GitHub (Configuration1/best_datasets.csv and Configuration2/best_datasets.csv).

Datasets	**Configuration 1 categories**	**Configuration 2 categories**
Covariates	Covariates	Covariates
Genotype	Genotype_UW_annotated_migraine	Genotype_UW_annotated
Genotype	Genotype_UW_not_annotated_migraine	Genotype_UW_not_annotated
Genotype	Genotype_W_annotated_migraine	Genotype_W_annotated
Genotype	Genotype_W_not_annotated_migraine	Genotype_W_not_annotated
PCA	PCA	PCA
PRS	PRS_migraine_5_AnnoPred	PRS_depression_4_AnnoPred
PRS	PRS_migraine_5_LDAK-GWAS	PRS_migraine_AnnoPred
PRS	PRS_migraine_AnnoPred	PRS_migraine_5_AnnoPred
PRS	PRS_migraine_LDAK-GWAS	PRS_depression_11_LDAK-GWAS
PRS	PRS_migraine_PLINK	PRS_migraine_LDAK-GWAS
PRS	–	PRS_migraine_5_LDAK-GWAS
PRS	–	PRS_depression_11_PRSice-2
PRS	–	PRS_depression_11_PLINK
PRS	–	PRS_depression_4_PLINK
PRS	–	PRS_migraine_PLINK

Table 1: **Table showing the final selected datasets for both configurations**. Covariates and PCA appeared in both categories. *UW means genotype data was not weighted by the GWAS file. For genotype data, we used the format `Genotype_{Weighted/Unweighted}_{Annotated/NotAnnotated}_{GWASFile}` to form categories. For Configuration 1, PRS datasets followed the format `PRS_{GWASFile}_{PRStool}`. Configuration 2 included more GWAS files, resulting in five additional datasets.

We created dataset combinations and trained a simple neural network on each. For Configuration 1 (11 datasets), we created combinations of 2–5 datasets ($\binom{11}{2}$–$\binom{11}{5}$), yielding 1012 unique model configurations. For Configuration 2 (16 datasets), similar combinations resulted in 6868 configurations. Datasets with validation AUC < 0.50 were discarded. Results for each combination are available on GitHub (Configuration1/ResultsML1.csv, Configuration2/ResultsML2.csv).

We analyzed the best-performing combinations based on the test performance. Tables 2 and 3 show the results for Configuration 1 and Configuration 2. Each configuration involves different combinations of selected datasets used to train a simple neural network.

For Configuration 1, the combination of `Covariates, PRS_migraine_PLINK, Genotype_UW_not_annotated_migraine` achieved a training AUC of 0.952524 (±0.038708), validation AUC of 0.739561 (±0.150103), and test AUC of 0.685461 (±0.127186). This dataset combination demonstrated high performance compared to others due to its balanced integration of three complementary data types: patient data (covariates), genetic risk scores (PRS_migraine_PLINK), and raw genetic data (Genotype_UW_not_annotated_migraine). An interesting pattern to observe is that when weighted and annotated SNPs are used along with AnnoPred, a better generalization performance is achieved. The fourth result indicates that raw genotype data can be skipped from combination modeling, and strong performance can still be attained using only `Covariates, PCA, PRS_migraine_LDAK-GWAS, PRS_migraine_AnnoPred`, with a training AUC of 0.910635 (±0.081524), validation AUC of 0.688332 (±0.088731), and test AUC of 0.668391

Combination Size	Datasets	Train AUC	Val AUC	Test AUC
3	Covariates, PCA, PRS_migraine_PLINK	0.90	0.66	0.69
3	Covariates, PRS_migraine_PLINK, Genotype_UW_not_annotated_migraine	0.95	0.74	0.69
5	Covariates, PCA, PRS_migraine_PLINK, Genotype_W_not_annotated_migraine, PRS_migraine_LDAK-GWAS	0.78	0.67	0.68
4	Covariates, PCA, PRS_migraine_LDAK-GWAS, PRS_migraine_AnnoPred	0.91	0.69	0.67
2	Genotype_W_annotated_migraine, PRS_migraine_AnnoPred	0.73	0.70	0.66

Table 2: **Results for Configuration 1.** The first, second, third, fourth, and fifth columns show the number of dataset combinations, the datasets included in the combinations, and the training, validation, and test AUC.

(±0.117305). The inclusion of `Covariates` and `PRS_migraine_PLINK` in top-performing combinations highlights their crucial role in genotype-phenotype prediction (Table 2).

Combination Size	Datasets	Train AUC	Val AUC	Test AUC
2	Genotype_UW_annotated, PRS_LDAK-GWAS_migraine_5	0.80	0.71	0.66
2	Genotype_UW_annotated, PRS_AnnoPred_depression_4	0.87	0.74	0.66
4	PRS_Plink_depression_11, PRS_Plink_depression_4, Genotype_UW_not_annotated, Genotype_UW_annotated	0.90	0.72	0.63
5	PRS_Plink_depression_11, PRS_LDAK-GWAS_migraine, PRS_AnnoPred_depression_4, Genotype_W_not_annotated, Genotype_UW_annotated	0.98	0.80	0.62
3	PRS_Plink_depression_11, Genotype_UW_annotated, Genotype_W_not_annotated	0.85	0.78	0.62

Table 3: **Results for Configuration 2.** The first, second, third, fourth, and fifth columns show the number of dataset combinations, the datasets included in the combinations, and the training, validation, and test AUC.

For Configuration 2, the combination of `Genotype_UW_annotated, PRS_LDAK-GWAS_migraine_5` achieved a training AUC of 0.798282 (±0.263385), validation AUC of 0.71399 (±0.122399), and test AUC of 0.663471 (±0.048293). Another two-dataset combination using `Genotype_UW_annotated, PRS_AnnoPred_depression_4` showed higher training performance with a training AUC of 0.870221 (±0.166343), validation AUC of 0.735277 (±0.109561), and test AUC of 0.65844 (±0.063303). These two-dataset combinations demonstrate strong generalization performance. The five-dataset combination (`PRS_Plink_depression_11, PRS_LDAK-GWAS_migraine, PRS_AnnoPred_depression_4, Genotype_W_not_annotated, Genotype_UW_annotated`) achieved a training AUC of 0.980247 (±0.011184), validation AUC of 0.80284 (±0.148088), and test AUC of 0.622748 (±0.057671). This suggests potential overfitting due to incorporating multiple datasets.

A notable pattern is that combinations including both depression-related PRS scores (PRS_Plink_depression and PRS_AnnoPred_depression) tend to achieve higher training and validation scores but show slightly reduced test performance. The consistent presence of Genotype_UW_annotated in all top-performing combinations highlights its crucial role in genotype-phenotype prediction, while the inclusion of both depression and migraine PRS scores in various combinations suggests the potential benefit of leveraging cross-disorder genetic information (Table 3).

DISCUSSION

Incorporating diverse genetic and non-genetic data sources is essential for robust genotype-phenotype prediction. Researchers have employed multiple methodologies to enhance predictive performance, including integrating various PRS, FA, and genetic data across different diseases and populations. However, leveraging such a broad range of datasets requires a systematic evaluation and integration framework that is adaptable to additional diseases, populations, and risk scores, as proposed in this study.

The framework can be used to (1) evaluate whether risk scores from related diseases can predict a primary phenotype, (2) identify the optimal GWAS files among those available in public repositories and the GWAS Catalog, (3) determine the optimal number of SNPs for phenotype prediction, (4) select the most suitable PRS tools, (5) assess the performance impact of incorporating weighted and non-weighted genotype data, (6) explore transfer learning, where GWAS data from one population aid in predicting phenotypes in a secondary population, (7) identify the best machine learning and deep learning models for a specific phenotype, (8) extend research by incorporating additional data sources, (9) facilitate the evaluation of cross-disorder genetic relationships through systematic testing of how variants from one disorder predict outcomes in another, (10) contribute to the development of personalized medicine approaches by identifying the most predictive combinations of genetic and non-genetic factors, and (11) optimize biomarker panels by determining which biomarkers contribute most to combined prediction models.

We considered migraine as a primary phenotype and depression as a secondary phenotype, incorporating 2 migraine and 3 depression GWAS, 4 PRS tools, functional annotations, as well as weighted and unweighted as well as annotated and unannotated genotype data to test the framework. We observed that combining multiple datasets or selective subsets of datasets enhances predictive performance. Risk scores and annotated genotype data derived from phenotypes related to the primary phenotype can also be used for prediction, as demonstrated by the contribution of the depression risk score in our analysis. Our findings indicate that the integration of PCA, covariates, PRS from AnnoPred and LDAK, as well as annotated and weighted genotype data, can be combined to form a more robust and generic predictor.

There are a few limitations to this study. First, forming categories and selecting the data requires strong exploratory data analysis, and managing a large number of datasets can be quite complex, as it also demands substantial storage. Although we restricted the analysis to SNPs below 10000, the generation process can take one to two days for larger datasets. Testing all combinations of the datasets is computationally expensive and requires a significant amount of computational resources. Moreover, the genetic datasets are quite large, necessitating the reduction of individual dataset dimensions, as implemented in the manuscript. The distribution of the different genetic datasets shows considerable diversity

RESOURCE AVAILABILITY

The genotype data can be downloaded from the UK Biobank. The annotation information is available from AnnoPred https://github.com/yiminghu/AnnoPred, and to obtain the processed version, use the following link: https://drive.google.com/drive/folders/19PMporIqzUj9IY3FNXbnzBI-aBy8KmT?usp=sharing. The GWAS files can be found at the GWAS Catalog https://www.ebi.ac.uk/gwas/. The code is accessible in the GitHub repository: https://github.com/MuhammadMuneeb007/EFGPP.

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, David B. Ascher (d.ascher@uq.edu.au).

Data and code availability

- The genotype data can be downloaded from the UK Biobank.
- The code is accessible in the GitHub repository: https://github.com/MuhammadMuneeb007/EFGPP.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

Not applicable.

AUTHOR CONTRIBUTIONS

Conceptualization, M.M. and D.A.; methodology, M.M. and D.A.; investigation, M.M.; writing–original draft, M.M.; writing-–review & editing, M.M. and D.A.; funding acquisition, D.A.; resources, M.M. and D.A.; supervision, D.A.

DECLARATION OF INTERESTS

The authors declare no competing interests

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES

Claude and GitHub Copilot were utilized to review the code and make a flowchart. After utilizing these tools, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

1. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 years of gwas discovery: Biology, function, and translation. The American Journal of Human Genetics *101*, 5–22. URL: http://dx.doi.org/10.1016/j.ajhg.2017.06.005. doi: 10.1016/j.ajhg.2017.06.005.

2. Medvedev, A., Mishra Sharma, S., Tsatsorin, E., Nabieva, E., and Yarotsky, D. (2022). Human genotype-to-phenotype predictions: Boosting accuracy with nonlinear models. PLOS ONE *17*, e0273293. URL: http://dx.doi.org/10.1371/journal.pone.0273293. doi: 10.1371/journal.pone.0273293.

3. Chalmer, M.A., Esserlind, A.L., Olesen, J., and Hansen, T.F. (2018). Polygenic risk score: use in migraine research. J Headache Pain *19*, 29–10. doi: 10.1186/s10194-018-0856-0.

4. LISTGARTEN, J., STEGLE, O., MORRIS, Q., BRENNER, S.E., and PARTS, L. (2013). Personalized medicine: From genotypes and molecular phenotypes towards therapy- session introduction. In Biocomputing 2014. WORLD SCIENTIFIC pp. 224–228. URL: http://dx.doi.org/10.1142/9789814583220_0022. doi: 10.1142/9789814583220_0022.

5. Dong, X., Xiao, T., Chen, B., Lu, Y., and Zhou, W. (2022). Precision medicine via the integration of phenotype-genotype information in neonatal genome project. Fundamental Research *2*, 873–884. URL: http://dx.doi.org/10.1016/j.fmre.2022.07.003. doi: 10.1016/j.fmre.2022.07.003.

6. Kogelman, L.J.A., Esserlind, A.L., Francke Christensen, A., Awasthi, S., Ripke, S., Ingason, A., Davidsson, O.B., Erikstrup, C., Hjalgrim, H., Ullum, H., Olesen, J., and Folkmann Hansen, T. (2019). Migraine polygenic risk score associates with efficacy of migraine-specific drugs. Neurol Genet *5*, e364–e364. doi: 10.1212/NXG.0000000000000364. Go to Neurology.org/NG for full disclosures. Funding information is provided at the end of the article. The DBDS Genomic Consortium and the International Headache Genetics Consortium coinvestgators are listed in appendices 2 and 3 at the end of the article. The Article Processing Charge was funded by the authors.

7. Ligthart, L., Hottenga, J.J., Lewis, C.M., Farmer, A.E., Craig, I.W., Breen, G., Willemsen, G., Vink, J.M., Middeldorp, C.M., Byrne, E.M., Heath, A.C., Madden, P.A.F., Pergadia, M.L., Montgomery, G.W., Martin, N.G., Penninx, B.W.J.H., McGuffin, P., Boomsma, D.I., and Nyholt (2014). Genetic risk score analysis indicates migraine with and without comorbid depression are genetically different disorders. Hum Genet *133*, 173–186. doi: 10.1007/s00439-013-1370-8. These authors contributed equally to this work.

8. Muneeb, M., Feng, S., and Henschel, A. (2022). Transfer learning for genotype–phenotype prediction using deep learning models. BMC Bioinformatics *23*. URL: http://dx.doi.org/10.1186/s12859-022-05036-8. doi: 10.1186/s12859-022-05036-8.

9. Hunter, D.J. (2005). Gene–environment interactions in human diseases. Nature Reviews Genetics *6*, 287–298. URL: http://dx.doi.org/10.1038/nrg1578. doi: 10.1038/nrg1578.

10. Guo, T., and Li, X. (2023). Machine learning for predicting phenotype from genotype and environment. Current Opinion in Biotechnology *79*, 102853. URL: http://dx.doi.org/10.1016/j.copbio.2022.102853. doi: 10.1016/j.copbio.2022.102853.

11. Fadista, J., Manning, A.K., Florez, J.C., and Groop, L. (2016). The (in)famous gwas p-value threshold revisited and updated for low-frequency variants. European Journal of Human Genetics *24*, 1202–1205. URL: http://dx.doi.org/10.1038/ejhg.2015.269. doi: 10.1038/ejhg.2015.269.

12. McCaw, Z.R., Colthurst, T., Yun, T., Furlotte, N.A., Carroll, A., Alipanahi, B., McLean, C.Y., and Hormozdiari, F. (2022). Deepnull models non-linear covariate effects to improve phenotypic prediction and association power. Nature Communications *13*. URL: http://dx.doi.org/10.1038/s41467-021-27930-0. doi: 10.1038/s41467-021-27930-0.

13. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics *38*, 904–909. URL: http://dx.doi.org/10.1038/ng1847. doi: 10.1038/ng1847.

14. Torkamani, A., Scott-Van Zeeland, A.A., Topol, E.J., and Schork, N.J. (2011). Annotating individual human genomes. Genomics *98*, 233–241. URL: http://dx.doi.org/10.1016/j.ygeno.2011.07.006. doi: 10.1016/j.ygeno.2011.07.006.

15. Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nature Reviews Genetics *20*, 467–484. URL: http://dx.doi.org/10.1038/s41576-019-0127-1. doi: 10.1038/s41576-019-0127-1.

16. Torkamani, A., Wineinger, N.E., and Topol, E.J. (2018). The personal and clinical utility of polygenic risk scores. Nature Reviews Genetics *19*, 581–590. URL: http://dx.doi.org/10.1038/s41576-018-0018-x. doi: 10.1038/s41576-018-0018-x.

17. Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature Genetics *50*, 1219–1224. URL: http://dx.doi.org/10.1038/s41588-018-0183-z. doi: 10.1038/s41588-018-0183-z.

18. Sehrawat, S., Najafian, K., and Jin, L. (2023). Predicting phenotypes from novel genomic markers using deep learning. Bioinformatics Advances *3*. URL: http://dx.doi.org/10.1093/bioadv/vbad028. doi: 10.1093/bioadv/vbad028.

19. Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A., and Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. Nature Reviews Genetics *16*, 85–97. URL: http://dx.doi.org/10.1038/nrg3868. doi: 10.1038/nrg3868.

20. Garreta, L., Cerón-Souza, I., Palacio, M.R., and Reyes-Herrera, P.H. (2021). Multigwas: An integrative tool for genome wide association studies in tetraploid organisms. Ecology and Evolution *11*, 7411–7426. URL: http://dx.doi.org/10.1002/ece3.7572. doi: 10.1002/ece3.7572.

21. Turley, P., Walters, R.K., Maghzian, O., Okbay, A., Lee, J.J., Fontana, M.A., Nguyen-Viet, T.A., Wedow, R., Zacher, M., Furlotte, N.A., Magnusson, P., Oskarsson, S., Johannesson, M., Visscher, P.M., Laibson, D., Cesarini, D., Neale, B.M., and Benjamin, D.J. (2018). Multi-trait analysis of genome-wide association summary statistics using mtag. Nature Genetics *50*, 229–237. URL: http://dx.doi.org/10.1038/s41588-017-0009-4. doi: 10.1038/s41588-017-0009-4.

22. Ishigaki, K., Sakaue, S., Terao, C., Luo, Y., Sonehara, K., Yamaguchi, K., Amariuta, T., Too, C.L., Laufer, V.A., Scott, I.C., Viatte, S., Takahashi, M., Ohmura, K., Murasawa, A., Hashimoto, M., Ito, H., Hammoudeh, M., Emadi, S.A., Masri, B.K., Halabi, H., Badsha, H., Uthman, I.W., Wu, X., Lin, L., Li, T., Plant, D., Barton, A., Orozco, G., Verstappen, S.M.M., Bowes, J., MacGregor, A.J., Honda, S., Koido, M., Tomizuka, K., Kamatani, Y., Tanaka, H., Tanaka, E., Suzuki, A., Maeda, Y., Yamamoto, K., Miyawaki, S., Xie, G., Zhang, J., Amos, C.I., Keystone, E., Wolbink, G., van der Horst-Bruinsma, I., Cui, J., Liao, K.P., Carroll, R.J., Lee, H.S., Bang, S.Y., Siminovitch, K.A., de Vries, N., Alfredsson, L., Rantapää-Dahlqvist, S., Karlson, E.W., Bae, S.C., Kimberly, R.P., Edberg, J.C., Mariette, X., Huizinga, T., Dieudé, P., Schneider, M., Kerick, M., Denny, J.C., Matsuda, K., Matsuo, K., Mimori, T., Matsuda, F., Fujio, K., Tanaka, Y., Kumanogoh, A., Traylor, M., Lewis, C.M., Eyre, S., Xu, H., Saxena, R., Arayssi, T., Kochi, Y., Ikari, K., Harigai, M., Gregersen, P.K., Yamamoto, K., Louis Bridges, S., Padyukov, L., Martin, J., Klareskog, L., Okada, Y., and Raychaudhuri, S. (2022). Multi-ancestry genome-wide association analyses identify novel genetic mechanisms in rheumatoid arthritis. Nature Genetics *54*, 1640–1651. URL: http://dx.doi.org/10.1038/s41588-022-01213-w. doi: 10.1038/s41588-022-01213-w.

23. Truong, B., Hull, L.E., Ruan, Y., Huang, Q.Q., Hornsby, W., Martin, H., van Heel, D.A., Wang, Y., Martin, A.R., Lee, S.H., and Natarajan, P. (2024). Integrative polygenic risk score improves the prediction accuracy of complex traits and diseases. Cell Genomics *4*, 100523. URL: http://dx.doi.org/10.1016/j.xgen.2024.100523. doi: 10.1016/j.xgen.2024.100523.

24. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics *81*, 559â€“575. URL: http://dx.doi.org/10.1086/519795. doi: 10.1086/519795.

25. Choi, S.W., and Oâ€™Reilly, P.F. (2019). Prsice-2: Polygenic risk score software for biobank-scale data. GigaScience *8*. URL: http://dx.doi.org/10.1093/gigascience/giz082. doi: 10.1093/gigascience/giz082.

26. Zheng, Z., Liu, S., Sidorenko, J., Wang, Y., Lin, T., Yengo, L., Turley, P., Ani, A., Wang, R., Nolte, I.M., Snieder, H., Aguirre-Gamboa, R., Deelen, P., Franke, L., Kuivenhoven, J.A., Lopera Maya, E.A., Sanna, S., Swertz, M.A., Vonk, J.M., Wijmenga, C., Yang, J., Wray, N.R., Goddard, M.E., Visscher, P.M., and Zeng, J. (2024). Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. Nature Genetics *56*, 767â€“777. URL: http://dx.doi.org/10.1038/s41588-024-01704-y. doi: 10.1038/s41588-024-01704-y.

27. Zhang, Q., Privé, F., Vilhjálmsson, B., and Speed, D. (2021). Improved genetic prediction of complex traits from individual-level data or summary statistics. Nature Communications *12*. URL: http://dx.doi.org/10.1038/s41467-021-24485-y. doi: 10.1038/s41467-021-24485-y.

28. Turner, S., Armstrong, L.L., Bradford, Y., Carlson, C.S., Crawford, D.C., Crenshaw, A.T., de Andrade, M., Doheny, K.F., Haines, J.L., Hayes, G., Jarvik, G., Jiang, L., Kullo, I.J., Li, R., Ling, H., Manolio, T.A., Matsumoto, M., McCarty, C.A., McDavid, A.N., Mirel, D.B., Paschall, J.E., Pugh, E.W., Rasmussen, L.V., Wilke, R.A., Zuvich, R.L., and Ritchie, M.D.

(2011). Quality control procedures for genome-wide association studies. Current Protocols in Human Genetics *68*. URL: http://dx.doi.org/10.1002/0471142905.hg0119s68. doi: 10.1002/0471142905.hg0119s68.

29. Truong, V.Q., Woerner, J.A., Cherlin, T.A., Bradford, Y., Lucas, A.M., Okeh, C.C., Shivakumar, M.K., Hui, D.H., Kumar, R., Pividori, M., Jones, S.C., Bossa, A.C., Turner, S.D., Ritchie, M.D., and Verma, S.S. (2022). Quality control procedures for genome-wide association studies. Current Protocols *2*. URL: http://dx.doi.org/10.1002/cpz1.603. doi: 10.1002/cpz1.603.

30. Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P., and Zondervan, K.T. (2010). Data quality control in genetic case-control association studies. Nature Protocols *5*, 1564–1573. URL: http://dx.doi.org/10.1038/nprot.2010.116. doi: 10.1038/nprot.2010.116.

31. Pavan, S., Delvento, C., Ricciardi, L., Lotti, C., Ciani, E., and D'Agostino, N. (2020). Recommendations for choosing the genotyping method and best practices for quality control in crop genome-wide association studies. Frontiers in Genetics *11*. URL: http://dx.doi.org/10.3389/fgene.2020.00447. doi: 10.3389/fgene.2020.00447.

32. Zeng, J., Zheng, Z., Liu, S., Sidorenko, J., Yengo, L., Turley, P., Ani, A., Wang, R., Nolte, I., Snieder, H., Yang, J., Wray, N., Goddard, M., and Visscher, P. (2023). Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. European Neuropsychopharmacology *75*, S29–S30. URL: http://dx.doi.org/10.1016/j.euroneuro.2023.08.063. doi: 10.1016/j.euroneuro.2023.08.063.

33. Massey, F.J. (1951). The kolmogorov-smirnov test for goodness of fit. Journal of the American Statistical Association *46*, 68–78. URL: http://dx.doi.org/10.1080/01621459.1951.10500769. doi: 10.1080/01621459.1951.10500769.

34. Hinton, G.E., and Roweis, S. (2002). Stochastic neighbor embedding. In S. Becker, S. Thrun, and K. Obermayer, eds. Advances in Neural Information Processing Systems vol. 15. MIT Press. URL: https://proceedings.neurips.cc/paper_files/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf.

35. McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv. URL: https://arxiv.org/abs/1802.03426. doi: 10.48550/ARXIV.1802.03426.

36. Chalmer, M.A., Esserlind, A.L., Olesen, J., and Hansen, T.F. (2018). Polygenic risk score: use in migraine research. The Journal of Headache and Pain *19*. URL: http://dx.doi.org/10.1186/s10194-018-0856-0. doi: 10.1186/s10194-018-0856-0.