Unsupervised Ensemble Learning for Efficient Integration of Pre-trained Polygenic Risk Scores

Chenyin Gao[1], Justin D. Tubbs[2, 3, 4], Yi Han[5], Min Guo[1], Sijia Li[1], Erica Ma[1], Dailin Luo[1], Jordan W. Smoller[2, 6], Phil H. Lee[2, 3, 4], and Rui Duan[1,6]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[2]Center for Genomic Medicine, Mass General Hospital, Boston, MA, USA
[3]Department of Psychiatry, Harvard Medical School, Boston, MA, USA
[4]Stanley Center for Psychiatry, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[5]Department of Statistics, Columbia University, New York, NY, USA
[6]Center for Precision Psychiatry, Mass General Hospital, Boston, MA, USA
Email: *rduan@hsph.harvard.edu*

Abstract

The growing availability of pre-trained polygenic risk score (PRS) models has enabled their integration into real-world applications, reducing the need for extensive data labeling, training, and calibration. However, selecting the most suitable PRS model for a specific target population remains challenging, due to issues such as limited transferability, data heterogeneity, and the scarcity of observed phenotype in real-world settings. Ensemble learning offers a promising avenue to enhance the predictive accuracy of genetic risk assessments, but most existing methods often rely on observed phenotype data or additional genome-wide association studies (GWAS) from the target population to optimize ensemble weights, limiting their utility in real-time implementation. Here, we present the UNsupervised enSemble PRS (*UNSemblePRS*), an unsupervised ensemble learning framework, that combines pre-trained PRS models without requiring phenotype data or summaries from the target population. Unlike traditional supervised approaches, UNSemblePRS aggregates models based on prediction concordance across a curated subset of candidate PRS models. We evaluated UNSemblePRS using both continuous and binary traits in the All of Us database, demonstrating its scalability and robust performance across diverse populations. These results underscore UNSemblePRS as an accessible tool for integrating PRS models into real-world contexts, offering broad applicability as the availability of PRS models continues to expand.

1 Introduction

Polygenic risk score (PRS) has emerged as a powerful tool to predict an individual's genetic predisposition to specific phenotypes. Typically, PRSs aggregate hundreds to millions of genetic variants identified in genome-wide association studies (GWAS) through a weighted sum of allele doses, each multiplied by its respective effect size (Khera et al., 2018; Kullo et al., 2022). Over the past decades, numerous PRS methods, such as LDpred (Vilhjálmsson et al., 2015), LDpred2 (Privé et al., 2020), lassosum (Mak et al., 2017), and PRS-CS (Ge et al., 2019), have been developed to more accurately quantify genetic predisposition for a variety of clinically relevant measures (e.g., body mass index (BMI) (Khera et al., 2019)) and diseases (e.g.,

breast cancer (BC) (Mavaddat et al., 2019)). For many clinical use cases, PRSs are becoming increasingly important for risk stratification and early identification of high-risk individuals.

With the growing emphasis on open science, researchers are progressively making their PRS models publicly accessible. Efforts to support this movement include the development of open-source repositories like the PGS Catalog (`https://www.pgscatalog.org`), which archives and documents publicly available PRS models. In just several years, over 5000 models spanning more than 650 traits have been shared, with this number continuing to grow rapidly. For widely studied traits such as height, Body Mass Index (BMI), and Breast Cancer (BC), more than 100 pre-trained models are available. These models, developed by various research groups, utilize diverse statistical and machine learning techniques applied to a range of datasets. The availability of these models promotes transparency, reproducibility, and collaboration in genetic research, creating new opportunities that leverages model-sharing and the integration of diverse data sources (Lambert et al., 2024b).

Despite recent advancements, from an implementation perspective, practitioners aiming to apply PRS models for risk stratification or other tasks requiring the quantification of genetic risk for specific traits often encounter uncertainty regarding best practices for leveraging pre-trained models in their own practice. This uncertainty arises from several factors. First, while a wide range of methods for training PRS models is available, there is no strong evidence that any one method consistently outperforms others (Ma and Zhou, 2021). Furthermore, the performance of PRS models is known to vary significantly across different biobanks and populations (Wojcik et al., 2019; Duncan et al., 2019; Fatumo et al., 2022), due to potential differences in contextual factors such as age, sex, and health status. Another significant challenge is that most GWAS studies predominantly include individuals of European (EUR) ancestry, leading to PRS models derived from these studies often performing poorly in non-EUR populations (Ding et al., 2022; Hou et al., 2024).

These uncertainties make it challenging to directly select the most suitable pre-trained PRS models for a given target population in a real-world implementation. To address this limitation, increasing efforts have been made to develop new PRS models by leveraging ancestry-specific GWAS summary statistics or utilizing multi-population GWAS meta-analyses that incorporate cross-population data to develop more robust PRS models with improved transferability across diverse populations (Ruan et al., 2022; Xiao et al., 2022). While retraining or calibrating PRS models for target populations can improve predictive accuracy, this approach is often not feasible due to the lack of complete or available GWAS data and linkage disequilibrium (LD) matrices for many non-EUR populations (Wray et al., 2007; Martin et al., 2017). Additionally, developing new PRS models requires substantial computational resources, stringent quality control processes, and multidisciplinary expertise in data pre-processing. These demands significantly hinder the practicality of widespread PRS model development and application.

As pre-trained models become more widely accessible, the development of effective methods for combining them has gained increasing importance, enabling researchers to leverage the strengths of individual models for improved overall performance. Various ensemble techniques have been designed to aggregate multiple PRS models tailored to target populations. Examples include multi-PRS (Albiñana et al., 2023), PRSmix (Truong et al., 2024), and PROSPER (Zhang et al., 2024), which employ ensemble learning strategies such as lasso, ridge, and SuperLearner to combine PRS models and optimize predictive power for specific populations or phenotypes (Márquez-Luna et al., 2017; Lloyd-Jones et al., 2019; Zhao et al., 2024).

However, training these ensemble models often requires phenotype data or additional GWAS from the target population, which may not always be readily available. For instance, evaluating genetic risk scores for a particular disease in a relatively young population can be challenging, as most individuals may not yet have experienced disease onset. Additionally, accurately defining the phenotype of interest for supervised ensemble training from large datasets, such as electronic health records, can be both costly and labor-intensive, often necessitating specialized clinical expertise (Lobach and Detmer, 2007). These challenges present significant barriers to the widespread application of supervised PRS ensemble methods.

In this paper, we introduce an easy-to-implement UNsupervised enSemble PRS (*UNSemblePRS*) framework that leverages pre-trained PRS models without requiring phenotype data from the target population. Unlike traditional supervised ensemble learning approaches, which derive weights by optimizing predictive power for the target trait, our unsupervised method aggregates models by leveraging prediction concordance across multiple PRS models. This approach mitigates unwanted prediction variability across models, resulting in improved prediction accuracy. UNSemblePRS enables users to specify a collection of candidate models tailored to their target population of interest and generates an aggregated PRS score that delivers stable performance. It is robust in scenarios where some candidate models are of lower quality or perform poorly in the target population. Furthermore, UNSemblePRS is highly scalable and can be flexibly adjusted to accommodate real-time population shifts, making it an effective and adaptable solution for diverse research applications. Integrated with open-source resources such as the PGS Catalog, UNSemblePRS provides a valuable tool for the research community, promoting accessibility, scalability, and reproducibility in genetic risk prediction. Using data from the All of Us (AoU) Research Hub, we evaluate the performance of UNSemblePRS, benchmarking it against the best-performing single PRS predictions across multiple phenotypes. Our results demonstrate its ability to provide high-quality genetic predictions across diverse traits. In addition, we assess the transferability and interpretability of UNSemblePRS, highlighting its strengths, limitations, and potential for future applications as more PRS models become publicly available.

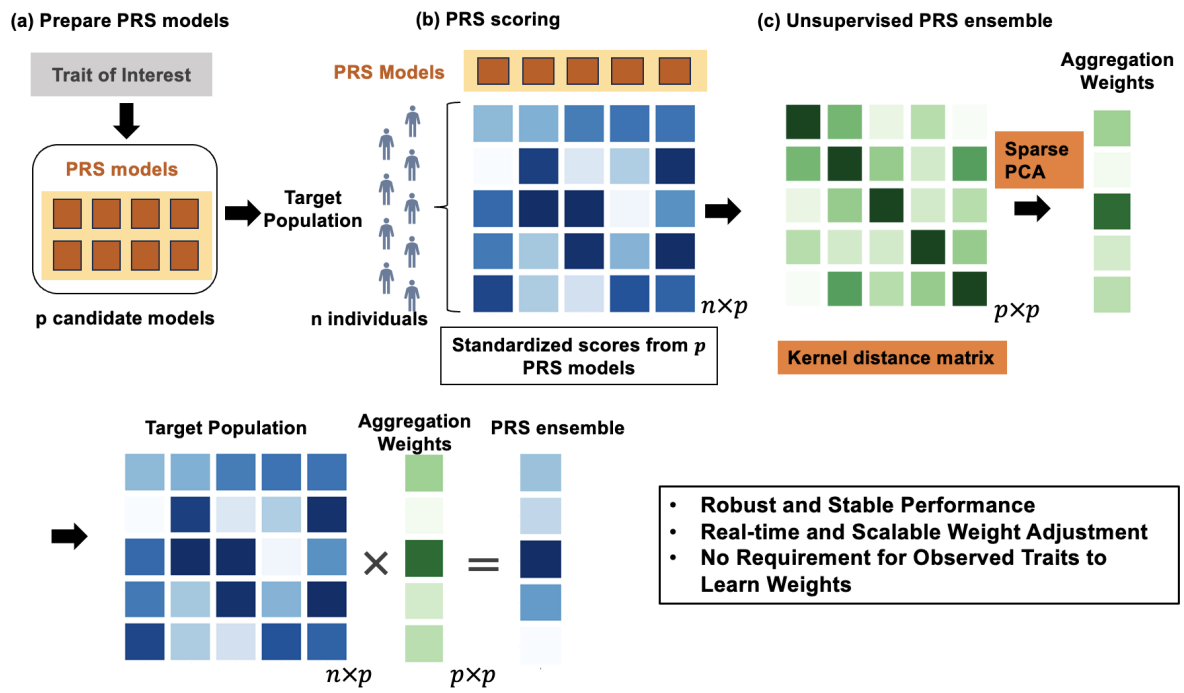

Figure 1: Workflow of UnsemblePRS, including 1) preparing PRS models, 2) applying the PRS models on the target population 3) performing UnsemblePRS to obtain the final aggregated score.

2 Results

2.1 Overview of Methods and Study Design

The UNSemblePRS method: The UNSemblePRS method integrates multiple pre-trained PRS models to generate aggregated risk scores for a target cohort of n individuals with available genotypic data. For a specific trait of interest, the process begins by selecting a set of candidate PRS models, which may have been trained on diverse populations, datasets, or using different PRS methodologies. In this work, we utilize models from the PGS catalog; however, users have the flexibility to curate candidate models based on their available resources and preferences. For each candidate model, PRS scores are computed for all individuals in the target cohort. With p candidate models, this step produces an $n \times p$ matrix, where the entry at position (i, j) corresponds to the PRS score for the i-th individual derived from the j-th model. This matrix serves as the basis for subsequent aggregation steps, enabling the generation of robust and comprehensive risk predictions.

The core of UNSemblePRS lies in assigning appropriate weights to enhance predictive power and ensure robustness across different environments. From an implementation standpoint, we assume that the trait of interest is not observed in the target cohort, which requires assigning weights in an unsupervised manner.

UNSemblePRS builds upon prior work on unsupervised model aggregation. Specifically, Parisi et al. (2014) introduced a meta-classifier that integrates a set of binary classifiers under the assumption of conditional independence between classifiers. Ma et al. (2023) extended this idea to integrate multiple dimension reduction algorithms, assuming the output of each algorithm can be expressed as a signal plus homoskedastic Gaussian noise. Coombes et al. (2020) proposed using principal component analysis (PCA) to refine tuning parameter choices in PRS models. All the aforementioned methods share a common strategy: applying spectral decomposition to a concordance matrix derived from the outputs of all classifiers or algorithms and obtaining a consensus output through a weighted average of the models, with the weights determined by the eigenvectors.

UNSemblePRS employs sparse kernel PCA on the standardized PRS matrix of dimension $n \times p$ to capture relationships among the PRS models (Figure 1). The standardized PRS matrix is constructed by centering and normalizing the PRS scores from each candidate model to have a mean of zero and a variance of one. The kernel function, which defines the similarity between PRS scores, maps the PRS matrix into a higher-dimensional feature space. In our work, we choose the kernel function as a polynomial function of degree three, where its explicit form can be found in the Online Method; results based on different kernel functions are included in the Supplementary Material for sensitivity analysis. This mapping enables the capture of more flexible and nonlinear relationships among the PRS scores. Within this feature space, sparse PCA is applied to identify a set of sparse eigenvectors that can reconstruct the features as accurately as possible, where the non-zero entries in the eigenvector indicate the selection of PRS models. Finally, the aggregation weights for combining the selected PRS scores into the final aggregated PRS are derived from the eigenvector generated by non-sparse kernel PCA separately for each ancestry population in the target cohort.

The rationale behind this approach is that mapping the scores into a non-linear space using the kernel product allows the leading eigenvector, which captures the largest portion of variation among the models, to effectively reflect the relative importance of the PRS models. In addition, the sparsity assumption helps mitigate issues arising from models with redundant information, low-quality training data, or errors introduced during uploading or downloading processes. These considerations are particularly relevant for open-source models available in the PGS catalog. First, many models are trained on the same GWAS data sources, leading to redundant information. Second, some models are constructed using only significant single-nucleotide polymorphisms (SNPs) from underpowered studies, resulting in scores that may be nearly non-predictive. Third, as the PGS catalog allows users to upload models, human errors, such as documenting the wrong trait or misaligning PGS weights, can occur. As demonstrated in our empirical study, applying a sparsity threshold to exclude less important models improves predictive performance. In UNSemblePRS, we set a default threshold of 50%, excluding PRS models assigned weights in the lower half. However, this thresh-

old is adjustable based on users' prior knowledge of redundancy or potential adversarial models among the candidates. A detailed mathematical formulation and step-by-step procedures are provided in the Online Method.

Evaluation of UNSemblePRS using All of Us Data The performance of UNSemblePRS is evaluated using data from the AoU Research Hub (All of Us Research Program, 2019), a large and diverse U.S.-based cohort study supporting biomedical research to improve health outcomes. We assess the performance of UNSemblePRS on five continuous traits and five binary traits. The continuous traits include height, BMI, high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, and diastolic blood pressure (DBP). The binary traits include BC, inflammatory bowel disease (IBD), chronic kidney disease (CKD), atrial fibrillation (AF), and coronary artery disease (CAD). The Experimental Factor Ontology (EFO) IDs of these traits can be found in the Supplementary Table A1. We downloaded all PRS models available in the PGS catalog as of May 2024 for each trait, corresponding to each trait based on their EFO IDs. The PRS models that either do not contain chromosome and position information or have no overlapping SNPs with individuals from the AoU database are excluded from our analysis. A total of 85 models are scored for BMI, 93 for height, 90 for HDL cholesterol, 104 for LDL cholesterol, and 51 for DBP. For the binary traits, we scored 135 models for BC, 20 for IBD, 20 for CKD, 37 for AF, and 67 for CAD.

To evaluate the proposed method, a total of 285,269 individuals aged 21 and older from the AoU database with available genotypic data were extracted. The target cohort for evaluation was randomly selected from a subset of participants with observed phenotypes for the trait of interest. Detailed sample size information for each phenotype is provided in Supplementary Table A1.The target cohort size is set to 10000, representing the scale of a moderately sized study in real-world practice where PRS applications are needed for various downstream tasks, such as developing a risk prediction model or accounting for genetic risk in comparative effectiveness studies. In addition to the genotype and phenotype data, we also extracted a set of covariates including age, sex, and calculated the 30 genetic principal components (PCs) for the target cohort (Price et al., 2006). Leveraging the diversity of the AoU data, we are able to assess the performance of UNSemblePRS in both the overall population and in different subpopulations defined by ancestral groups and other demographic factors. More details regarding the quality control of the genotypic data, ancestry annotation, and extraction of each trait can be found in the Online Method.

We compare UNSemblePRS to several alternative methods for unsupervised model aggregation. These include (1) the average of all standardized PRS models, which assigns equal importance to each candidate model (denoted as "Avg"), and (2) the average of PRSs weighted by the size of the relevant GWAS studies (denoted as "SSAvg"). Additionally, we consider using (3) the PRS derived from the largest GWAS study

(denoted as "MaxPRS"), motivated by emerging evidence that the size of the GWAS study can significantly impact the predictive power of PRS models (Márquez-Luna et al., 2017; Grinde et al., 2019; Weissbrod et al., 2022; Lehmann et al., 2023). We also consider (4) the approach involves deriving weights using PCA applied to the correlation matrix of the candidate PRS models (denoted as "PCA"), as proposed by Coombes et al. (2020) for combining PRS models with different tuning parameters using the clumping and thresholding method. Beyond these unsupervised aggregation methods, we also benchmark UNSemblePRS against (5) the best-performing candidate model selected based on observed traits (denoted as "best"). However, it is important to note that in real-world scenarios where the traits are unobserved, identifying the best model is not feasible.

For the predictive performance assessment with continuous traits, we report the increase in explained variance on the linear scale, measured as the incremental R^2. This is estimated by the difference between the R^2 of the full linear model, which includes PRS and covariates, and the reduced linear model with only covariates. For binary traits, we provide the adjusted area under the receiver operating characteristic curve (AUC), accounting for the effects of covariates. In addition to assessing the predictive performance of UNSemblePRS, we examined its transferability across the context-specific populations, the impact of sparsity thresholds on model performance, and the interpretability of its weights compared to relevant benchmark methods. Further details of the implementation are provided in the Online Method section.

2.2 Key findings

Unsupervised PRS ensemble demonstrates comparable performance to the best PRS across multiple traits and diverse ancestry populations Figure 2 illustrates the predictive performance of the compared methods for both quantitative and binary traits. Meanwhile, we also calculate the difference in predictive accuracy for each unsupervised method relative to the best single PRS model, averaged across all continuous or binary traits (see the heatmaps on the left of Figure 2). These evaluations are conducted on the entire target cohort (ALL) and stratified by the ancestry annotation; additional evaluation on the performance stratified by the age and sex are provided in the Supplementary Figures A1 and A2.

For continuous traits, UNSemblePRS achieves predictive accuracy comparable to the best single PRS model across all ancestry populations. On average, its performance difference is around 5% relative to the best model in the overall population, less than 3% in EUR and American (AMR) populations, 7% in African-ancestry population (AFR) and 14% in East Asian population (EAS). Notably, UNSemblePRS significantly outperforms other unsupervised methods for aggregating models.On average across continuous traits, UNSemblePRS exhibits the smallest relative decrease in performance compared to the best single

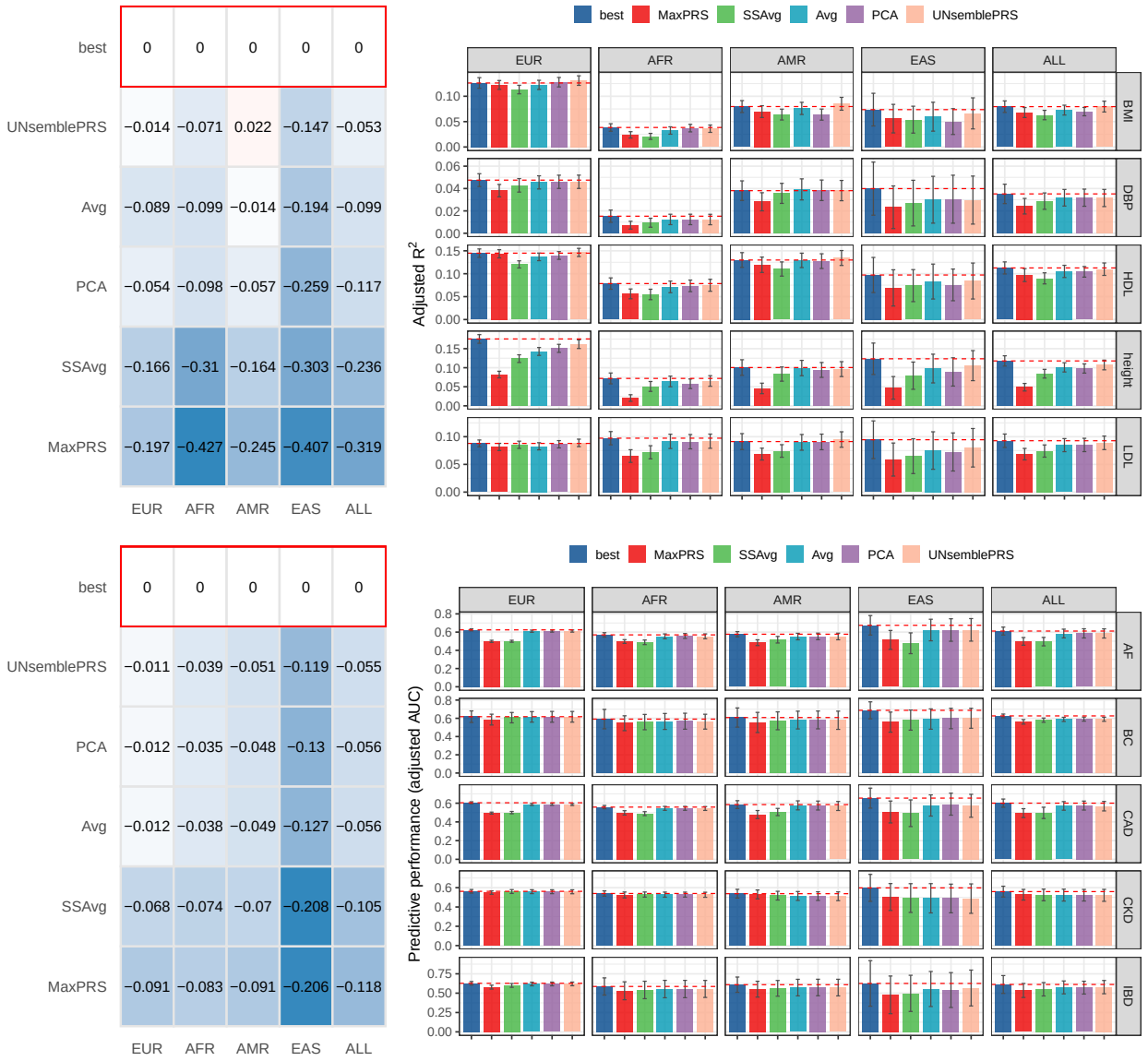

Figure 2: **(Left)** Relative differences in predictive performance compared to the best single PRS model, evaluated across 100 random samples (N = 10,000) and averaged over continuous and binary traits in the overall target cohort and stratified by genetic ancestry. **(Right)** Predictive performance for continuous and binary traits in the overall target cohort and across ancestry groups. Error bars represent ±1 standard deviation computed from the random samples. Dashed lines denote the performance of the best single PRS model.

model, being 46.9% smaller than the unweighted average and 55.1% smaller than PCA. Methods that aggregate PRS models using GWAS sample sizes or selected based on the largest GWAS sample size perform worse than other approaches, suggesting that larger GWAS sample sizes do not directly translate into better predictive performance for PRS models. In addition, UNSemblePRS remains the most effective unsupervised method for predicting binary traits. However, the improvement is less obvious compared to the performance for continuous traits. Predicting binary traits poses challenges, as many candidate models from the PGS catalog exhibit relatively low predictive power in the AoU database. Consequently, this limited model performance leads to similar weights being assigned to each candidate model across multiple ancestry populations, causing the performance more similar to the average method.

UNSemblePRS assigns higher weights to models with better predictive performance. We show that the weights learned by UNSemblePRS demonstrate stronger alignment with the true predictive performance of each candidate model. For each population, we assess the predictive performance of each model using the observed traits and regress it against the weights assigned by PCA and UNSemblePRS after proper scaling.

Figure 3 presents scatter plots and fitted curves for all the candidate PRS models across the considered traits. Both weights and true predictive performance are normalized to have zero mean and unit variance for each set of trait-specific candidate PRS models. The fitted R^2 values are used to evaluate alignment or concordance between the actual model performance and the assigned weights. For continuous traits, the concordance can reach as high as $R^2 = 0.79$, indicating that models with higher predictive performance are consistently assigned larger weights. In contrast, for the PCA method, the concordance can be as low as 0.01, suggesting that higher weights may be assigned to models with relatively low performance. For binary traits, the weights assigned by UNSemblePRS show weaker concordance with true predictive performance, especially in non-EUR populations, although they still outperform PCA weights in alignment. This issue likely arises because most PRS models for binary traits in the PGS catalog are trained on EUR GWAS samples, resulting in poor and indistinguishable performance in non-EUR populations. These findings highlight UNSemblePRS's ability to effectively estimate the relative performance of each candidate model and rank them accordingly for both continuous and binary traits.

UNSemblePRS shows better transferability than the best model The best PRS model, identified using the observed traits, often varies across populations. This variation can lead to transferability challenges, as the best model for one dataset may perform poorly in another population. To illustrate this, we utilize word clouds (Figure 4) to show the frequency at which each trait-specific PRS model is identified as the best

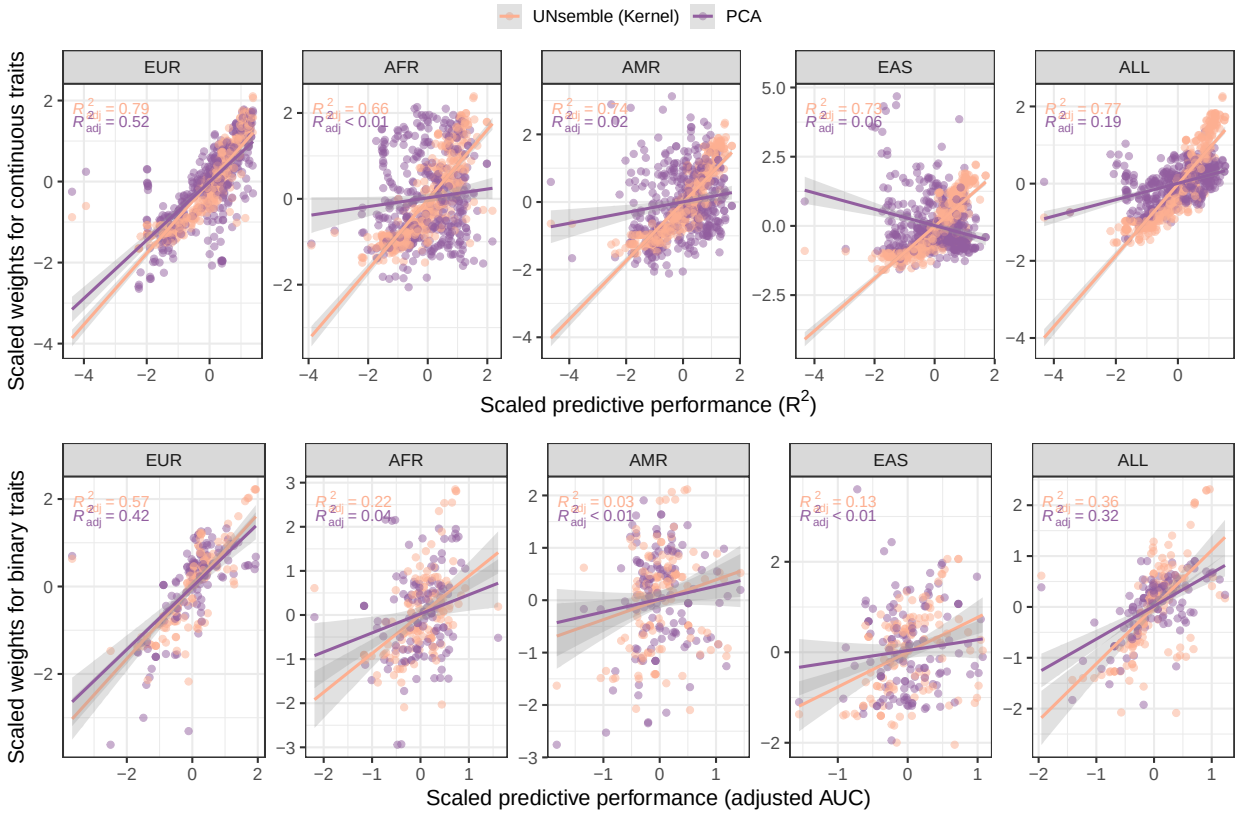

Figure 3: The scaled aggregation weights assigned to each candidate PRS model by the kernel and standard PCA methods, compared to the corresponding scaled predictive performance (R^2 or adjusted AUC) across multiple ancestry populations and the overall cohort. Each point represents a PRS model from the PGS Catalog.

model across 100 independent target cohorts within each ancestry population. The size of each PRS model's name in the word cloud is proportional to the number of times it is selected as the best model, with models fewer than twice excluded for clarity.

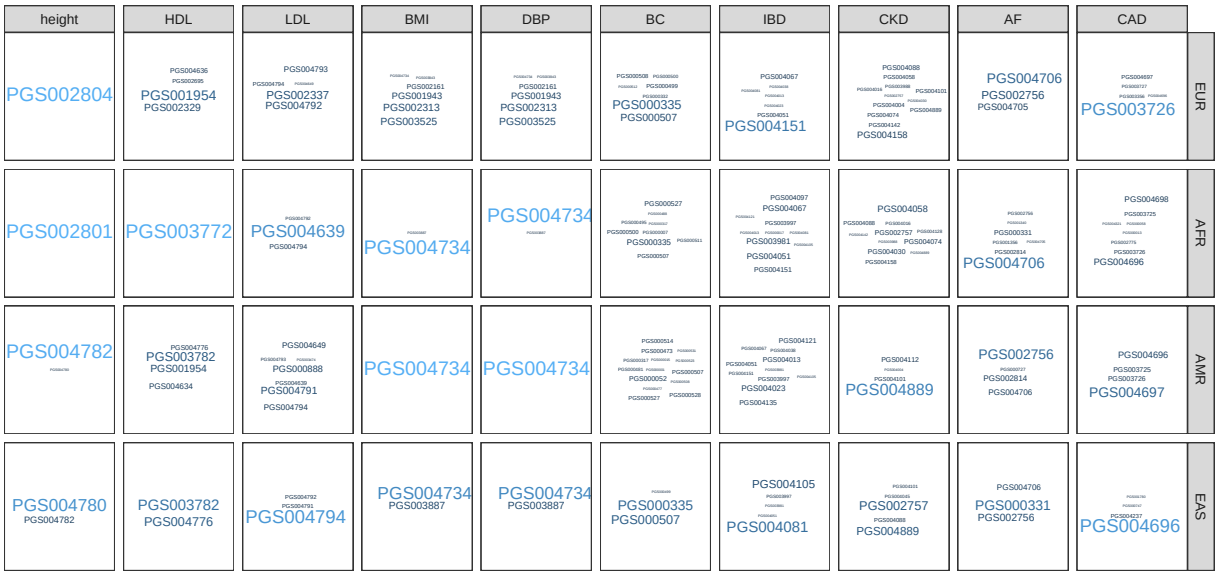

Figure 4: Word cloud visualization of the most frequently selected best PRS model across 100 repeated random samples (N = 10,000) for different traits and ancestry populations. The text size is proportional to the frequency with which each PRS model was identified as the best across the 100 draws.

Figure 4 demonstrates that for traits, such as height, the best PRS model is relatively consistent within each ancestral group, but varies across ancestry populations. The metadata of these PRS models reveals they are predominantly trained using GWAS studies based on the corresponding ancestry populations. In contrast for other traits, like LDL and BC, the best PRS model varies widely within and across ancestral groups, highlighting the poor generalizability and transferability of these trait-specific PRS models (Liu et al., 2021; Wang et al., 2022).

Although the weights in UNSemblePRS can be easily recalibrated for a new population, we also evaluate their transferability by applying the pre-learned UNSemblePRS weights directly to a new population, which has different context features, such as sex, age, and genetic ancestry, from the target cohort used for deriving the weights. Specifically, we train the best model and determine the weights for UNSemblePRS using a dataset of 10000 samples randomly selected from the AoU data. We then evaluate the model's performance on an independent population drawn from the remaining AoU database, which exhibits a different distribution of sex, age or genetic ancestry. Transferability is assessed by comparing the relative performance changes between the independent samples and the target cohort.

Figure 5 presents the relative changes in predictive performance (measured by R^2 for continuous traits

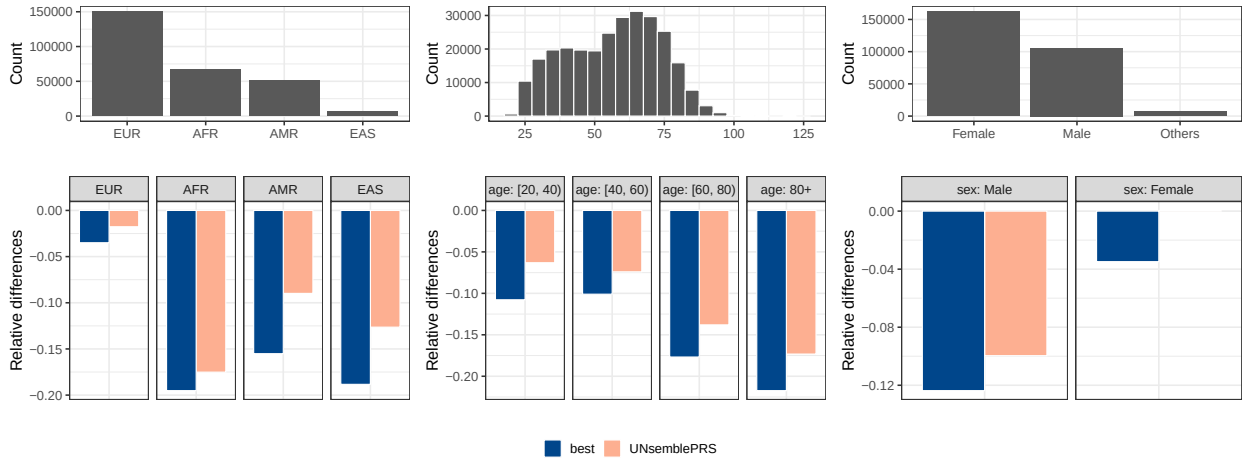

Figure 5: Relative changes in predictive performance of the best single PRS model and UNSemblePRS in independent samples compared to the target cohort, averaged over continuous and binary traits across 100 random draws (N = 10,000) for each context-specific population. The distributions of context-specific features for the target cohort are shown in the upper panel.

and AUC for binary traits) averaged over all the considered traits across each context-specific population. Regarding the impact of genetic ancestry on the PRS prediction, performance changes are modest ($\leq 5\%$) on average in EUR population for both methods since the original target cohort used to learn the weights and select the best model is dominated by the EUR population. Notably, the predictive performance in UNSemblePRS tends to decrease less than that of the best model, with a decline of 1.8% for UNSemblePRS compared to 3.5% for the best model. The result indicates that UNSemblePRS is more generalizable to individuals who share the same ancestry origin but are not included in the target cohort (i.e., the independent samples from the EUR population). When the independent samples exclusively consist of the AFR, AMR, or EAS populations, the predictive performances show more pronounced declines. Specifically, the performance drop of the best model when transporting from the original cohort to the non-EUR populations ranges from $15.0\% - 20.0\%$, while UNSemblePRS shows a smaller decline of 12.2%, with 15.9% for the AFR population, 8.5% for the AMR population, and 12.2% for the EAS population, respectively.

The performance drops when transporting from the original cohort to individuals aged between 20 and 60 have a decrease in PRS accuracy of 6.1% for UNSemblePRS compared to 10.4% for the best model. For those over 80, the decrease is 17.7% for UNSemblePRS compared to 21.7% for the best model. This disparity might be attributed to the large proportion of individuals between ages 20 and 60 ($\geq 50\%$) and the small proportion over 80 ($\leq 7\%$) in the target cohort, which leads to poor transferability for the senior population. Furthermore, UNSemblePRS shows improved transferability compared to the best model across sex-specific populations, with an average decline of 5% for UNSemblePRS compared to 8% for the best model. Overall, UNSemblePRS demonstrates smaller relative changes across both sex, age, and genetic

ancestry when averaging over all traits, compared to the best model. This indicates that UNSemblePRS is less sensitive to shifts in contextual features in new independent samples.

Sparsity in UNSemblePRS helps to increase the predictive performance The sparsity threshold in UNSemblePRS mitigates the impact of low-quality or corrupted models on the final aggregated PRS, using a data-driven approach. The sparsity level determines the proportion of models excluded by assigning their weights to 0 after the initial weights are derived from kernel PCA. The choice of the sparsity level can thus influence the performance of UNSemblePRS, and an optimal level may depend on prior knowledge about the candidate PRS models.

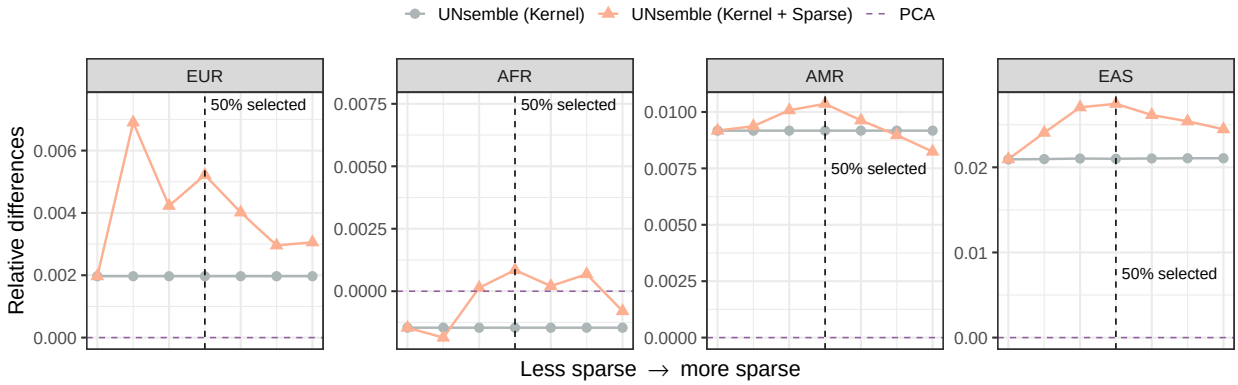

Figure 6: Changes in the relative performance of UNSemblePRS with and without sparsity, compared to PCA, stratified by ancestry populations and averaged across all traits, as the level of sparsity varies.

We compare the proposed UNSemblePRS methods with different sparsity levels in Figure 6, which illustrates the predictive performance across various levels of sparsity and ancestry populations, averaged over all the considered traits. Since UNSemblePRS utilizes kernel mapping for the relationship between PRS models, compared to PCA, it yields superior predictive performance for all the sparsity levels considered, except for the AFR population where low sparsity yields slightly worse performance than PCA. As the sparsity level increases, we see increased performance across all ancestry populations, as shown in Figure 6. This demonstrates that promoting sparsity is beneficial for selecting useful PRS models and removing unwanted redundant information. However, the sparsity level should not be set too high, as it risks excluding informative models and reducing performance. When using the PGS catalog as the source of pre-trained PRS models, we recommend a default sparsity level of 50%.

Scalability A key advantage of UNSemblePRS is its scalability. Leveraging sparse kernel PCA, weights can be computed in less than two seconds for a dataset with 10000 subjects and approximately 100 candidate models, enabling real-time implementation. For larger target sample sizes, UNSemblePRS supports

distributed computing. This approach, partitions samples into random batches, computes kernel matrices within each batch, and combines the results. This strategy achieves runtime comparable to processing a single batch. Additionally, distributed computing facilitates the integration of data across multiple biobanks or institutions. This is particularly valuable when working with underrepresented populations, where combining data from multiple institutions can improve weight estimation. The distributed strategy allows for sharing only a $p \times p$ kernel matrix across sites, enabling joint learning of UNSemblePRS weights while maintaining data privacy. Details for the distributed algorithm are provided in the Online Method.

3 Discussion

In response to the growing emphasis on open science and open-source machine learning and the increasing availability of pre-trained PRS models, we introduce UNSemblePRS, an efficient method for aggregating pre-trained PRS models to deliver robust and reliable performance in target cohorts in real time. By leveraging the numerous PRS models available in databases such as the PGS Catalog, UNSemblePRS avoids the complexity associated with calibrating or retraining models for target populations, which can often be computationally intensive and may be infeasible for some underrepresented populations. Using data from the AoU Research Hub, we demonstrate that UNSemblePRS achieves performance comparable to the best PRS model across diverse ancestry populations and traits. Additionally, it shows strong consistency between the weights assigned to each candidate model and their predictive performance. When applied to a new population, UNSemblePRS exhibits better transferability than the best model selected based on observed traits. Recalibrating the weights further enhances its predictive performance.

Despite the robustness of UNSemblePRS, performance gaps persist across ancestral groups, with predictive performance in non-EUR populations generally lower than in EUR populations. Even when comparing the best-performing model for each ancestral population, significant gaps remain for most traits. These disparities are rooted in the underrepresentation of non-EUR populations in genetic studies (Ju et al., 2022). In the PGS catalog, the majority of models are still trained on EUR GWAS or EUR-dominated GWAS due to the lack of genetic data from populations of non-European ancestry, leading to lower-quality models for non-EUR populations (Ding et al., 2022; Hou et al., 2024). As UNSemblePRS relies on the diversity and quality of its candidate set of pre-trained models, its performance in non-EUR populations is currently suboptimal. However, we anticipate that as recognition of these disparities grows and more diverse samples are collected, researchers will develop higher-quality models for underrepresented populations in the near future. UNSemblePRS will then play a critical role in leveraging the strengths of individual models to achieve improved performance in real-world applications.

Moving forward, the integration of pre-trained PRS models could benefit from incorporating external information about the reliability or similarity between the training data of the PRS models and the target population. For instance, the PGS catalog provides details on the ancestral composition of GWAS training and validation samples, as well as the methods and datasets used to develop the PRS models (Lambert et al., 2021, 2024a). While this information is not completely reported for all models, it could be utilized to guide the learning of optimal weights. However, striking the right balance between leveraging prior information and relying on purely data-driven weights will be critical. Users may need to carefully calibrate this balance, as the optimal approach could vary across different traits and populations.

Acknowledgement

We gratefully acknowledge All of Us participants for their contributions, without whom this research would not have been possible. We also thank the National Institutes of Health's All of Us Research Program for making available the participant data examined in this study. This work was supported by the National Institutes of Health (R01MH137218 and R01GM148494).

Data and Code Availability

The code to implement UNSemblePRS is available on GitHub, `https://github.com/biostat-duan-lab/UNSemblePRS`. This study used data from the All of Us Research Program, accessed through the All of Us Researcher Workbench `https://researchallofus.org/`. The dataset used in this study is from the Controlled Tier Dataset v7 (2022Q4R9) release. Data are available to approved researchers through the All of Us Researcher Workbench following completion of the required registration and training. Due to privacy and confidentiality policies, individual-level data cannot be shared directly. Researchers interested in accessing the All of Us dataset may apply for access through the All of Us Research Program.

References

Albiñana, C., Zhu, Z., Schork, A. J., Ingason, A., Aschard, H., Brikell, I., Bulik, C. M., Petersen, L. V., Agerbo, E., Grove, J., et al. (2023). Multi-pgs enhances polygenic prediction by combining 937 polygenic scores. *Nature Communications*, 14(1):4702.

All of Us Research Program (2019). The "all of us" research program. *New England Journal of Medicine*, 381(7):668–676.

All of Us Research Program (2024). Genomic data in the all of us research program. *Nature*, 627(8003):340–346.

Cao, Y., Zhang, A., and Li, H. (2020). Multisample estimation of bacterial composition matrices in metagenomics data. *Biometrika*, 107(1):75–92.

Coombes, B. J., Ploner, A., Bergen, S. E., and Biernacka, J. M. (2020). A principal component approach to improve association testing with polygenic risk scores. *Genetic Epidemiology*, 44(7):676–686.

Ding, Y., Hou, K., Burch, K. S., Lapinska, S., Privé, F., Vilhjálmsson, B., Sankararaman, S., and Pasaniuc, B. (2022). Large uncertainty in individual polygenic risk score estimation impacts prs-based risk stratification. *Nature Genetics*, 54(1):30–39.

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R., and Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications*, 10(1):3328.

Erichson, N. B., Zheng, P., Manohar, K., Brunton, S. L., Kutz, J. N., and Aravkin, A. Y. (2020). Sparse principal component analysis via variable projection. *SIAM Journal on Applied Mathematics*, 80(2):977–1002.

Fatumo, S., Chikowore, T., Choudhury, A., Ayub, M., Martin, A. R., and Kuchenbaecker, K. (2022). A roadmap to increase diversity in genomic studies. *Nature Medicine*, 28(2):243–250.

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., and Smoller, J. W. (2019). Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nature Communications*, 10(1):1776.

Grinde, K. E., Qi, Q., Thornton, T. A., Liu, S., Shadyab, A. H., Chan, K. H. K., Reiner, A. P., and Sofer, T. (2019). Generalizing polygenic risk scores from europeans to hispanics/latinos. *Genetic Epidemiology*, 43(1):50–62.

Hou, K., Xu, Z., Ding, Y., Mandla, R., Shi, Z., Boulier, K., Harpak, A., and Pasaniuc, B. (2024). Calibrated prediction intervals for polygenic scores across diverse contexts. *Nature Genetics*, pages 1–11.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327.

Ju, D., Hui, D., Hammond, D. A., Wonkam, A., and Tishkoff, S. A. (2022). Importance of including non-european populations in large human genetic studies to enhance precision medicine. *Annual Review of Biomedical Data Science*, 5(1):321–339.

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., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9):1219–1224.

Khera, A. V., Chaffin, M., Wade, K. H., Zahid, S., Brancale, J., Xia, R., Distefano, M., Senol-Cosar, O., Haas, M. E., Bick, A., et al. (2019). Polygenic prediction of weight and obesity trajectories from birth to adulthood. *Cell*, 177(3):587–596.

Kullo, I. J., Lewis, C. M., Inouye, M., Martin, A. R., Ripatti, S., and Chatterjee, N. (2022). Polygenic scores in biomedical research. *Nature Reviews Genetics*, 23(9):524–532.

Lambert, S. A., Gil, L., Jupp, S., Ritchie, S. C., Xu, Y., Buniello, A., McMahon, A., Abraham, G., Chapman, M., Parkinson, H., et al. (2021). The polygenic score catalog as an open database for reproducibility and systematic evaluation. *Nature Genetics*, 53(4):420–425.

Lambert, S. A., Wingfield, B., Gibson, J. T., Gil, L., Ramachandran, S., Yvon, F., Saverimuttu, S., Tinsley, E., Lewis, E., Ritchie, S. C., et al. (2024a). Enhancing the polygenic score catalog with tools for score calculation and ancestry normalization. *Nature Genetics*, pages 1–6.

Lambert, S. A., Wingfield, B., Gibson, J. T., Gil, L., Ramachandran, S., Yvon, F., Saverimuttu, S., Tinsley, E., Lewis, E., Ritchie, S. C., et al. (2024b). The polygenic score catalog: new functionality and tools to enable fair research. *medRxiv*, pages 2024–05.

Lehmann, B., Mackintosh, M., McVean, G., and Holmes, C. (2023). Optimal strategies for learning multi-ancestry polygenic scores vary across traits. *Nature Communications*, 14(1):4023.

Liu, C., Zeinomar, N., Chung, W. K., Kiryluk, K., Gharavi, A. G., Hripcsak, G., Crew, K. D., Shang, N., Khan, A., Fasel, D., et al. (2021). Generalizability of polygenic risk scores for breast cancer among women with european, african, and latinx ancestry. *JAMA Network Open*, 4(8):e2119084–e2119084.

Lloyd-Jones, L. R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K. E., Wang, H., Zheng, Z., Magi, R., Esko, T., et al. (2019). Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nature Communications*, 10(1):5086.

Lobach, D. F. and Detmer, D. E. (2007). Research challenges for electronic health records. *American Journal of Preventive Medicine*, 32(5):S104–S111.

Ma, R., Sun, E. D., and Zou, J. (2023). A spectral method for assessing and combining multiple data visualizations. *Nature Communications*, 14(1):780.

Ma, Y. and Zhou, X. (2021). Genetic prediction of complex traits with polygenic scores: a statistical review. *Trends in Genetics*, 37(11):995–1011.

Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X., and Sham, P. C. (2017). Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology*, 41(6):469–480.

Márquez-Luna, C., Loh, P.-R., Consortium, S. A. T. . D. S., Consortium, S. T. . D., and Price, A. L. (2017). Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genetic Epidemiology*, 41(8):811–823.

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., Daly, M. J., Bustamante, C. D., and Kenny, E. E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, 100(4):635–649.

Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., Tyrer, J. P., Chen, T.-H., Wang, Q., Bolla, M. K., et al. (2019). Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *The American Journal of Human Genetics*, 104(1):21–34.

Parisi, F., Strino, F., Nadler, B., and Kluger, Y. (2014). Ranking and combining multiple predictors without labeled data. *Proceedings of the National Academy of Sciences*, 111(4):1253–1258.

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(8):904–909.

Privé, F., Arbel, J., and Vilhjálmsson, B. J. (2020). Ldpred2: better, faster, stronger. *Bioinformatics*, 36(22-23):5424–5431.

Ruan, Y., Lin, Y.-F., Feng, Y.-C. A., Chen, C.-Y., Lam, M., Guo, Z., He, L., Sawa, A., Martin, A. R., et al. (2022). Improving polygenic prediction in ancestrally diverse populations. *Nature Genetics*, 54(5):573–580.

Salmon, J., Harmany, Z., Deledalle, C.-A., and Willett, R. (2014). Poisson noise reduction with non-local pca. *Journal of Mathematical Imaging and Vision*, 48:279–294.

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., et al. (2024). Integrative polygenic risk score improves the prediction accuracy of complex traits and diseases. *Cell Genomics*, 4(4).

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, 97(4):576–592.

Wang, Y., Tsuo, K., Kanai, M., Neale, B. M., and Martin, A. R. (2022). Challenges and opportunities for developing more generalizable polygenic risk scores. *Annual Review of Biomedical Data Science*, 5(1):293–320.

Weissbrod, O., Kanai, M., Shi, H., Gazal, S., Peyrot, W. J., Khera, A. V., Okada, Y., Martin, A. R., Finucane, H. K., et al. (2022). Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nature Genetics*, 54(4):450–458.

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., Highland, H. M., Patel, Y. M., Sorokin, E. P., Avery, C. L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, 570(7762):514–518.

Wray, N. R., Goddard, M. E., and Visscher, P. M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research*, 17(10):1520–1528.

Xiao, J., Cai, M., Hu, X., Wan, X., Chen, G., and Yang, C. (2022). Xpxp: improving polygenic prediction by cross-population and cross-phenotype analysis. *Bioinformatics*, 38(7):1947–1955.

Zhang, H., Zhan, J., Jin, J., Zhang, J., Lu, W., Zhao, R., Ahearn, T. U., Yu, Z., O'Connell, J., Jiang, Y., et al. (2023). A new method for multiancestry polygenic prediction improves performance across diverse populations. *Nature Genetics*, 55(10):1757–1768.

Zhang, J., Zhan, J., Jin, J., Ma, C., Zhao, R., O'Connell, J., Jiang, Y., 23andMe Research Team, Koelsch, B. L., Zhang, H., et al. (2024). An ensemble penalized regression method for multi-ancestry polygenic risk prediction. *Nature Communications*, 15(1):3238.

Zhao, Z., Dorn, S., Wu, Y., Yang, X., Jin, J., and Lu, Q. (2024). One score to rule them all: regularized ensemble polygenic risk prediction with gwas summary statistics. *bioRxiv*, pages 2024–11.

A Online Supplementary Material

Online method for UNSemblePRS: Our unsupervised Polygenic Risk Score (PRS) ensemble method is grounded in the spiked covariance model (Johnstone, 2001), which has been extensively utilized across a range of applications (Lobach and Detmer, 2007; Salmon et al., 2014; Cao et al., 2020). In this model, let $X_{n\times p}$ represent the standardized matrix consisting of p PRS scores for n individuals with respect to a single phenotype, where each column has a mean of zero and a unit length. Standardizing multiple PRSs helps to mitigate variations in model scales and ensures comparability across models. We assume that the covariance matrix of the p PRS models, denoted by $\Sigma_{p\times p}$ follows $\Sigma_{p\times p} = \Omega_{p\times p} + \sigma^2 I_{p\times p}$ where $\Omega_{p\times p} = vv^\intercal$ and v is a unit vector representing relatively each model's performance, and $\sigma^2 I_{p\times p}$ represents the noise induced by sampling and other prediction errors. The unit vector v is referred to as the spike, which indicates the strength of the predictive signal for each PRS model, and therefore reflects each model's importance.

Our goal is to recover the vector v and use it as the aggregation weight for the unsupervised ensemble learning. The estimation of v is equivalent to estimating the eigenvector of the sample covariance matrix $\Sigma_{p\times p}$. However, the sample covariance $\Sigma_{p\times p}$ only captures the linear relationships among these PRS scores. To generalize, we apply a kernel function Φ to map the score matrix $X_{n\times p}$ into a non-linear space $F_{p\times p}$, where nonlinear relationships among the PRS scores can be explored. In UNSemblePRS, the default kernel function Φ is chosen to be the polynomial kernel with a degree of 3. Specifically, for PRS scores $X_{\bullet,k}$ and $X_{\bullet,k'}$ predicted by the k-th and k'-th PRS models, we measure their relationships by computing $(X_{\bullet,k}^\intercal X_{\bullet,k'} + 1)^3$ for the (k, k')-th entry of the feature space $F_{p\times p}$; other kernel functions are available in the R package `kernlab`.

Next, we employ the sparse PCA algorithm from Erichson et al. (2020) to remove redundant signals from the kernel space. More specifically, the rank-one sparse PCA algorithm is applied to the mapped kernel spaces, minimizing the following objective function:

$$\widehat{v} = \arg\min_{v} \left\{ \frac{1}{2} \|F_{p\times p} - F_{p\times p} vv^\intercal\|_F^2 + \lambda \|v\|_1 \right\}, \tag{1}$$

where λ is the tuning parameter, and $\|v\|_1$ is the ℓ_1 norm for the vector v to promote sparsity. Finally, the final aggregated PRS score is derived by combining the standardized PRS matrix $X_{n\times p}$ according to the aggregation weights $\widehat{v}$; see Algorithm 1 for the detailed procedures.

UNSemblePRS for distributed learning The UNSemblePRS is particularly suited for a distributed learning framework. By computing the correlation matrix $\Sigma^{(s)}$ for each site-specific PRS score $X_{n_s\times p}^{(s)}$ as $\Sigma^{(s)} = (X_{n_s\times p}^{(s)})^\intercal X_{n_s\times p}^{(s)}$, we can obtain the distributed correlation matrix $\bar{\Sigma} = \sum_{s=1}^{S} w_s \Sigma^{(s)}$ for the central

Algorithm 1 UNSemblePRS Method for PRS Ensemble Learning

Require: Matrix $X_{n\times p}$: standardized PRS scores for n individuals across p PRS models.
Ensure: Final aggregated PRS score for the n target samples.

1: Obtain the feature space $F_{p\times p}$ by applying polynomial kernel with degree of 3 to all pairwise standardized PRS scores.
2: Apply sparse PCA to the transformed feature space $F_{p\times p}$ and obtain the weights:

$$\widehat{v} = \arg\min_{v} \frac{1}{2}\|F_{p\times p} - F_{p\times p} v v^{\intercal}\|_F^2 + \lambda\|v\|_1.$$

3: Compute the final aggregated PRS score $y^{\text{UNSemblePRS}} = X\widehat{v}$.

server, where w_s are user-specified site weights that reflect prior information about the data quality at each site. Next, we extract the distributed feature space $\bar{F}_{p\times p}$ by applying the polynomial kernel to $\bar{\Sigma}$. Finally, the distributed aggregation weights are obtained by applying the rank-one sparse PCA algorithm to $\bar{F}_{p\times p}$. Distributed learning for other kernel functions should be discussed on a case-by-case basis and is omitted here. This process refines the ensemble learning approach while preserving privacy concerns and offers a more accurate polygenic prediction for a more complex target cohort; see Algorithm 2 for the complete implementation.

Algorithm 2 Distributed UNSemblePRS Method for PRS Ensemble Learning

Require: Site-specific matrix $X^{(s)}_{n_s\times p}$: standardized PRS scores for n_s individuals across p PRS models at site $s = 1, \cdots, S$.
Ensure: Final aggregated PRS scores for n_s individuals models at site $s = 1, \cdots, S$.

1: Each site shares the correlation matrix $\Sigma^{(s)} = (X^{(s)}_{n_s\times p})^{\intercal} X^{(s)}_{n_s\times p}$ with the central server.
2: At the central server, obtain the federated feature space $\bar{F}_{p\times p}$ by applying the polynomial kernel to $\bar{\Sigma} = \sum_{s=1}^{S} w_s \Sigma^{(s)}$.
3: Obtain the federated aggregation weights:

$$\widehat{v} = \arg\min_{v} \frac{1}{2}\|\bar{F}_{p\times p} - \bar{F}_{p\times p} v v^{\intercal}\|_F^2 + \lambda\|v\|_1.$$

4: The central server broadcast the weight $\widehat{v}$.
5: Compute the final federated aggregated PRS score $y^{\text{UNSemblePRS},s} = X^{(s)}\widehat{v}$ at site $s = 1, \cdots, S$.

AoU Data extraction and curation The All of Us (AoU) Research Program is a population-based, longitudinal initiative designed to recruit at least one million participants from across the United States to advance precision medicine (All of Us Research Program, 2019). Ongoing enrollment captures extensive health-related data—ranging from genomic data and electronic health records to lifestyle factors and environmental exposures—to elucidate how these elements collectively influence health and disease. Our analyses leverage the latest All of Us Controlled Tier Dataset v7 (2022Q4R9). The program operates under a centralized institutional review board, and all participants provide informed consent. The present study is

a secondary analysis of de-identified AoU participants, approved as nonhuman participants research by the AoU Research Program Resource Access Board, with a waiver of the informed consent from the participants.

Genotyping and quality control: Genotyping and primary quality control (QC) are performed by the centralized data processing team of the AoU Research Program, summarized previously (All of Us Research Program, 2024). The AoU microarray genotyping dataset consists of 1.8 million single nucleotide variants (SNVs) genotyped for 312925 participants, mapped to the hg38/GRCh38 reference genome. QC ensures that samples are sex-concordant (i.e., genotype-based self-reported sex match at birth), have a missing call rate below 2%, and contamination rates under 3%. Genetic ancestry and relatedness are determined for the participants with genome-wide short-read sequencing data.

Polygenic scoring: For PRS generation, we use genotyped SNPs in the 1000 Genomes Project Phase 3 because this reference panel offers comprehensive global representation, compatibility with the AoU dataset's genetic architecture, and reliable imputation accuracy, making it well-suited for ancestry-aware polygenic score analyses.

Ancestry annotation: Ancestry groups are defined based on the Human Genome Diversity Project and 1000 Genomes reference panels, including African/African American (AFR), Admixed/ Latino/ Native American (AMR), East Asian (EAS), European (EUR), Middle Eastern (MID), South Asian (SAS), and Other/Multi (OTH, representing individuals not belonging to a single ancestry or with balanced admixture). For the remaining 285269 individuals with genome-wide genotyping data, ancestry and relatedness are assigned using a K-means clustering algorithm ($k = 5$) applied to 16 genetic principal components (PCs), as described by the AoU Genomics Initiative. Within each predicted ancestry population, PCA is performed using PC-AiR in the R `GENESIS` package to account for genetic sub-structuring while controlling for relatedness. Phasing and imputation are conducted with ShapeIT (v5.0) and minimac (v4-1.0.2), using the 1000 Genomes Project Phase 3 as a reference panel. Post-QC genotype data includes bi-allelic SNPs with a missing rate below 2%, Hardy-Weinberg equilibrium p-value less than 1×10^{-5}, minor allele frequency above 0.1%, and imputation quality scores of at least 0.8 for each population.

Extraction of traits and covariates: The considered traits for the AoU database are summarized in Table A1.Continuous traits are curated as the average of the most recent five measurements, or fewer if less than five measurements are available. Phenotype labels for the considered binary traits are determined as cases when individuals have been diagnosed based on ICD and SNOMED criteria, or a documented disease history for the target trait. Age for each individual is defined as the difference between the current year, 2023, and the individual's birth year.

Table A1: Summary of considered traits, number of models, and sample sizes of the AoU data

	Phenotype	EFO	# PRS	# Participants
Continuous				
	Body Mass Index	EFO_0004340	93	268834
	Height	EFO_0004339	85	269570
	High-Density Lipoprotein	EFO_0004612	90	113180
	Low-Density Lipoprotein	EFO_0004611	104	85890
	Diastolic Blood Pressure	EFO_0006336	51	222367
Binary				
	Breast Cancer	EFO_0000305	135	223106
	Inflammatory Bowel Disease	EFO_0003767	20	223106
	Chronic Kidney Disease	EFO_0003884	20	223106
	Atrial Fibrillation	EFO_0000275	37	71437
	Coronary Artery Disease	EFO_0001645	67	71437

Model evaluation for quantitative and binary traits We randomly draw 10000 individuals from the AoU data who have both genotypic data and observed trait of interest as the target cohort. The number of participants with observed phenotypes for each trait is presented in Table A1. A typical study involving PRS aims to test the association between PRS and a target trait. This association is evaluated using standard metrics for association testing or goodness-of-fit assessment. For continuous traits, once covariates such as age, sex, and genetic principal components are accounted for in the model, we report the incremental phenotypic variance explained, denoted by R^2. This measure represents the increase in R^2 due to the inclusion of the PRS in the regression model, thereby isolating the unique explanatory power of the PRS itself. When dealing with binary traits, a common practice is to assess the model's discriminative ability through the area under the Receiver Operating Characteristic curve (AUC), adjusted for covariates to ensure the contribution of PRS is accurately reflected (Zhang et al., 2023).

Each PRS method is applied to the target cohort stratified by sex, age, and ancestry annotation. The prediction accuracy is assessed across 100 repeated draws of target cohorts (Figure 2, Figures A1 and A2). The trait BC is excluded from the analysis for the stratified evaluation based on sex. The best model is selected from the candidate PRS scores with the largest association with the target trait in the given stratified population. The average of PRSs, weighted by the size of the relevant GWAS studies, and the PRS derived from the largest GWAS study are based on a subset of the candidate PRS scores with metadata available at the PGS Catalog. The aggregation weights for the PCA and UNSemblePRS methods are obtained based on the stratified population. The sparsity level for the UNSemblePRS is set to 50% by default using candidate PRS scores from the PGS Catalog.

Furthermore, for each trait considered, both predictive metrics and aggregation weights of candidate PRS models are standardized to have zero mean and unit variance. The scaled predictive metrics, measured by

R^2 for continuous traits and adjusted AUC for binary traits, are then compared with the scaled aggregation weights derived from UNSemblePRS and standard PCA analysis in scatter plots (Figure 3). Each point represents a PRS model, and the curves are fitted by linear regression models for these two methods, respectively. The concordance between the actual model performance and the assigned weights is computed as the R^2 of the fitted linear regression models.

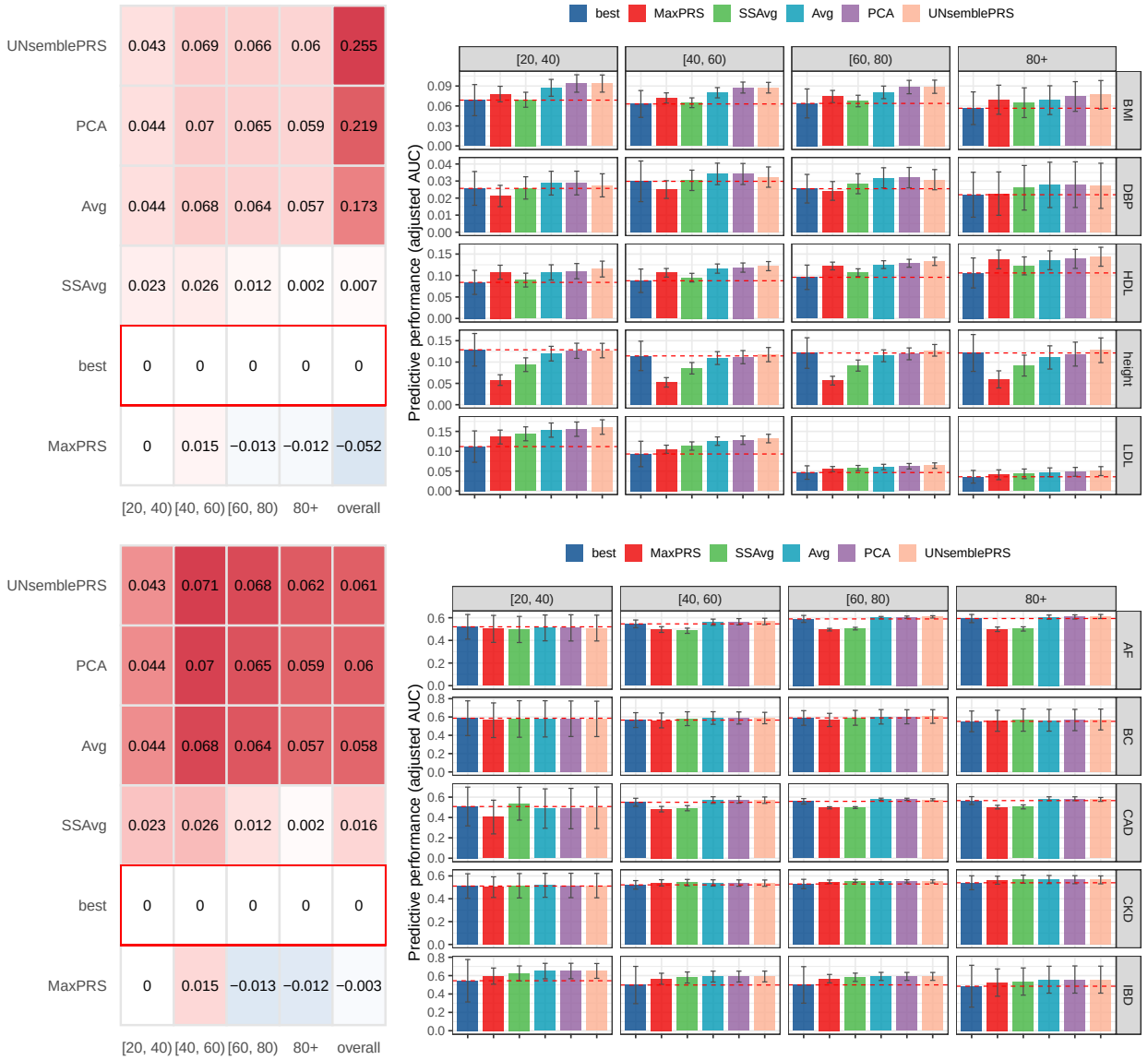

Figure A1: **(Left)** Relative differences in predictive performance compared to the best single PRS model, evaluated across 100 random samples (N = 10,000) and averaged over continuous and binary traits in the overall target cohort and stratified by age groups. **(Right)** Predictive performance for continuous and binary traits in the overall target cohort and across age groups. Error bars represent ±1 standard deviation computed from the random samples. Dashed lines denote the performance of the best single PRS model.

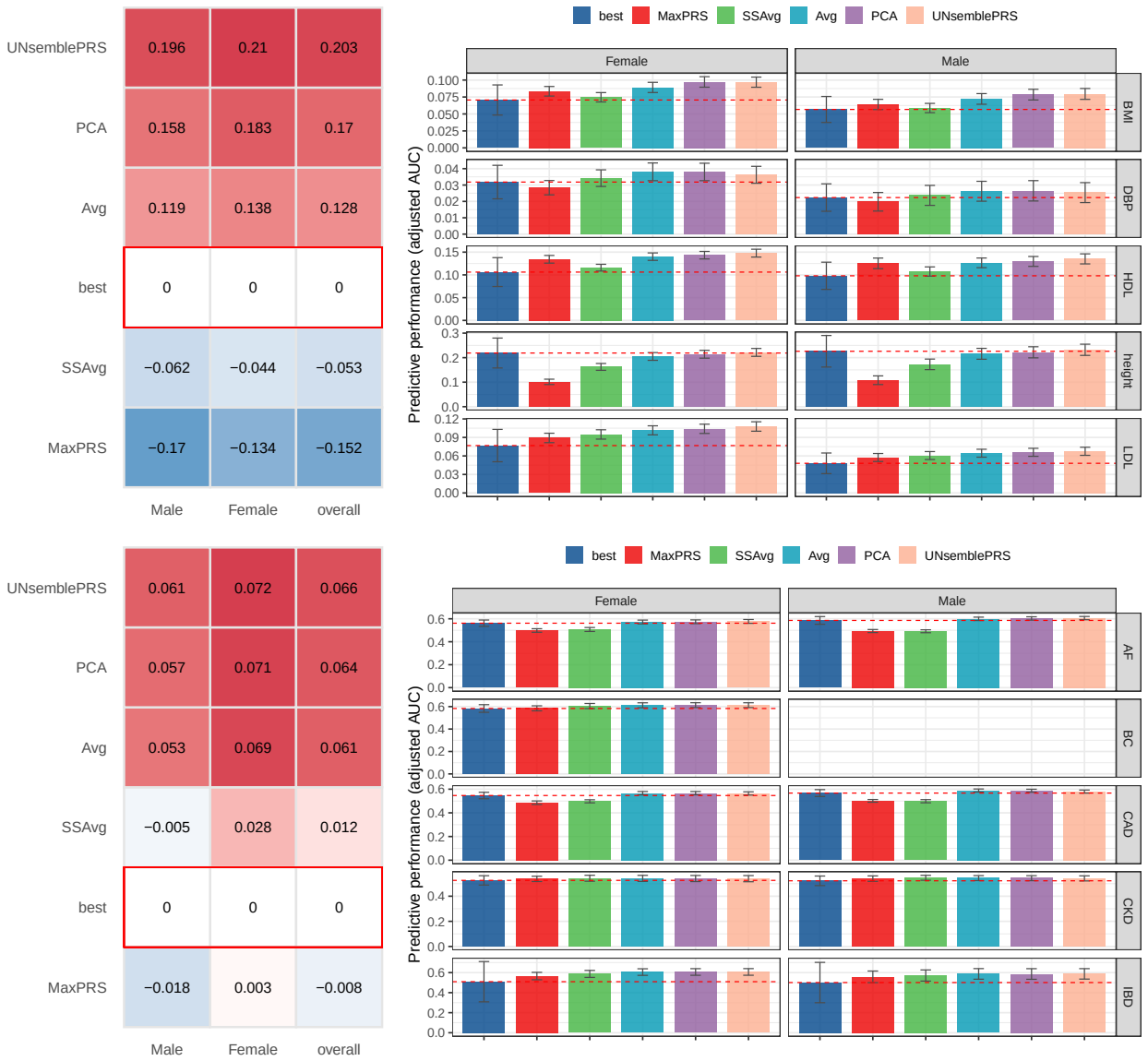

Figure A2: **(Left)** Relative differences in predictive performance compared to the best single PRS model, evaluated across 100 random samples (N = 10,000) and averaged over continuous and binary traits in the overall target cohort and stratified by sex groups. **(Right)** Predictive performance for continuous and binary traits in the overall target cohort and across sex groups. Error bars represent ±1 standard deviation computed from the random samples. Dashed lines denote the performance of the best single PRS model.

Assessing model transferability To assess the transferability of the UNSemblePRS, we apply the pre-learned UNSemblePRS weights from the target cohort to independent samples that consist exclusively of one ancestry population. Specifically, for each randomly sampled target cohort, we randomly draw independent samples from the remaining individuals in the AoU database, ensuring the independent samples have the same sample size as the target cohort and consist exclusively of a single ancestry population (i.e., EUR, AFR, AMR, or EAS) or include individuals from a specific age ranges (i.e., $[20, 40)$, $[40, 60)$, $[60, 80)$, and 80+) or sex groups (male or female). The changes in predictive performance relative to the target cohorts are computed for the UNSemblePRS and the best single PRS models, as presented in Figure 5.

Sensitivity analysis on the impact of varying sparsity threshold To examine the impact of sparsity level on the performance of UNSemblePRS, we assess the model's predictive performance by applying the UNSemblePRS to the target cohorts under varying levels of the tuning parameter λ. Specifically, the proportion of selected candidate PRS models is ranging from 100% to 10%. The changes in predictive performance relative to the PCA analysis are computed for the UNSemblePRS with and without sparsity, as well as for the best single PRS model, as presented in Figure 6. This evaluation allows us to understand the sensitivity of the UNSemblePRS to different sparsity levels and how these levels impact the performance of ensemble learning for the considered traits.

Sensitivity analysis on the impact of kernel function To examine the impact of various kernel functions on the performance of UNSemblePRS, we evaluate the model's predictive performance using three distinct kernel functions: 1) a polynomial function of degree 3, $(X_{\bullet,k}^{\intercal} X_{\bullet,k'} + 1)^3$; 2) a polynomial function of degree 4, $(X_{\bullet,k}^{\intercal} X_{\bullet,k'} + 1)^4$; and 3) a Gaussian radial basis kernel, $\exp(-\|X_{\bullet,k} - X_{\bullet,k'}\|^2)$. Each UNSemblePRS model is assessed on the replicated ancestry-specific target cohorts. The results are presented in Figure A3. It can be observed that UNSemblePRS achieves prediction accuracy comparable to the best single PRS model across different ancestry populations, regardless of the chosen kernel function. On average, the predictive performance for the continuous traits is 0.083 for the polynomial kernel and 0.080 for the Gaussian radial basis kernel, while the predictive performance for the binary traits is 0.568 for the polynomial kernel and 0.561 for the Gaussian radial basis kernel. This suggests that the model can be robust to kernel choice, provided that some level of flexible nonlinear relationships is captured. In practice, the selection of a specific kernel function should be made on a case-by-case basis, guided by domain knowledge.

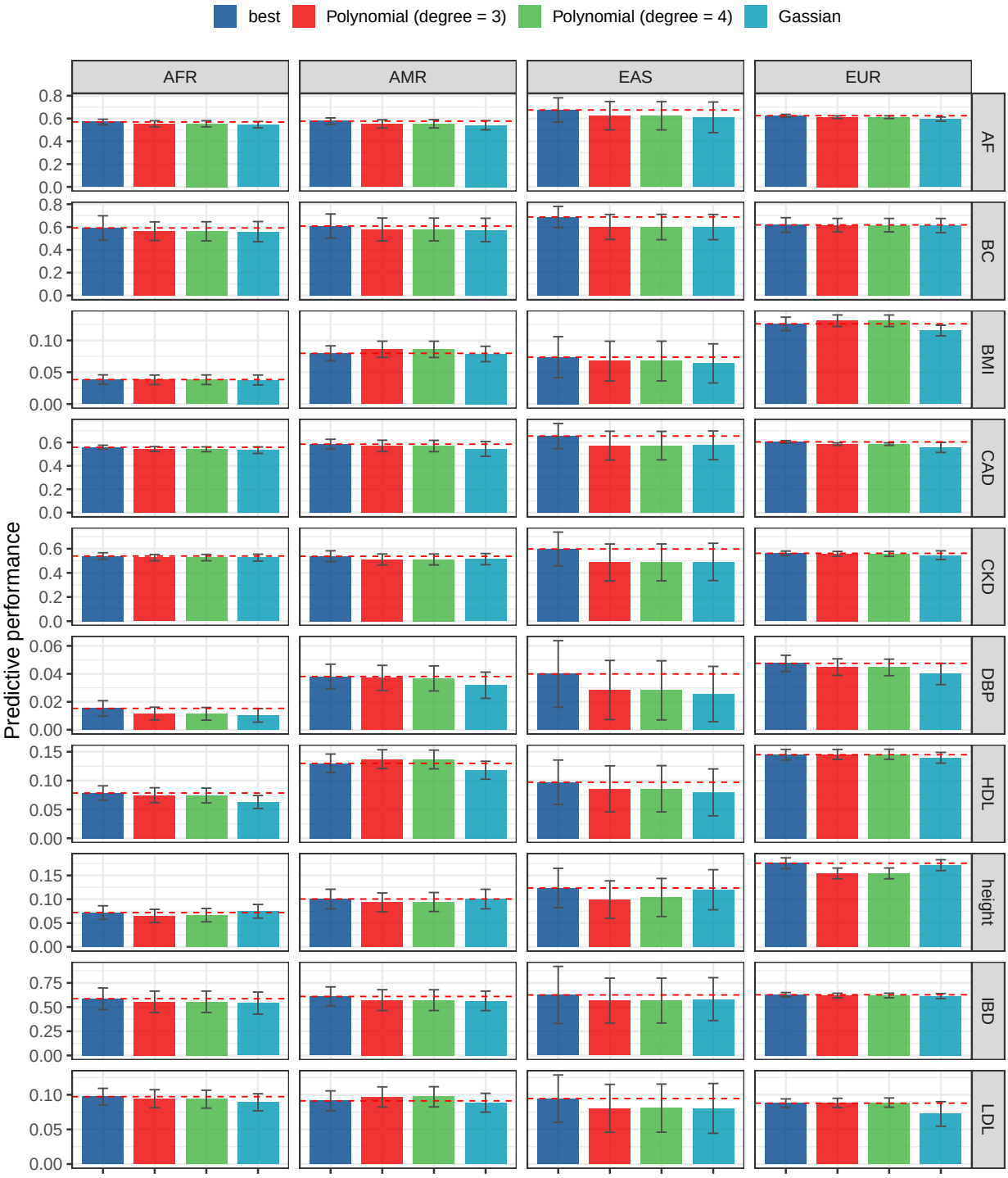

Figure A3: The predictive performances of the UNSemblePRS with different kernel functions for continuous and binary traits on the overall target cohort and separately on each ancestry population.