nature communications

Article | Article in Press

https://doi.org/10.1038/s41467-026-77128-5

A pre-train and fine-tune framework for adaptive boosting of pre-trained polygenic risk scores

Received: 13 April 2025

Accepted: 26 June 2026

Published online: 29 August 2026

Check for updates

Jie Hu[1,14], Raelynn Chen[2,14], Maxwell Salvatore[1], Olivia Wu[2], Okan Bilge Ozdemir[2], Yiwen Lu[1], Shawn N. Murphy[3], Elizabeth W. Karlson[4], Atlas Khan[5], Krzysztof Kiryluk[6], Iftikhar J. Kullo[7], Johanna L. Smith[8], Eimear E. Kenny[9], Yuan Luo[10], Zaldy S. Tan[11], William G. La Cava[12,13], Marylyn D. Ritchie[1], Yong Chen[1] ✉ & Ruowang Li[2] ✉

Polygenic risk scores are widely used for predicting genetic risk across complex diseases and traits, and several pre-trained models have been developed. Few approaches leverage these pre-trained polygenic risk scores to further refine predictive performance. Here, we present Adaptive Boosting of pre-trained Polygenic Risk Socres, a fine-tuning framework that refines pre-trained polygenic risk score models through adaptive variable selection and model boosting to identify additional predictive signals that may not be fully captured by the original models. Simulations show that our framework can identify signals orthogonal to pre-trained polygenic risk scores while controlling false discovery rates. Using UK Biobank data, we fine-tune pre-trained polygenic risk scores for binary diseases and continuous traits, and validate the results across three independent datasets: All of Us, eMERGE, and Penn Medicine Biobank. Real data analyses show that Adaptive Boosting of pre-trained Polygenic Risk Scores achieves statistically significant improvements in several scenarios while maintaining competitive performance in others.

Genome-wide association studies (GWAS) have identified a large number of genetic loci associated with complex traits and diseases, providing the foundation for polygenic risk prediction[1–3]. While GWAS have primarily focused on identifying significant genetic associations, polygenic risk score (PRS) methods leverage GWAS findings to combine associations across SNPs into a single number aimed at improving disease risk prediction and stratification[4–14]. Notably, public repositories, such as the PGS Catalog, store and make available thousands of published pre-trained PRS developed using various methodologies across hundreds of traits, facilitating the independent application and evaluation of these models[15].

Despite the wide availability of pre-trained PRS, several unsolved challenges can limit their optimal application. First, the majority of GWAS have assumed a linear relationship between a SNP's effect allele

[1]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, USA. [2]Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [3]Mass General Brigham, Boston, MA, USA. [4]Division of Rheumatology, Immunity and Inflammation, Brigham and Women's Hospital, Boston, MA, USA. [5]Columbia University Medical Center (CUMC), Columbia University, New York, NY, USA. [6]Division of Nephrology, Department of Medicine, Vagelos College of Physicians & Surgeons, Columbia University, New York, NY, USA. [7]Division of Cardiovascular Diseases, Mayo Clinic, Rochester, MN, USA. [8]Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN, USA. [9]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [10]Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA. [11]Department of Neurology & Medicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [12]Department of Pediatrics, Havard Medical School, Boston, MA, USA. [13]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA. [14]These authors contributed equally: Jie Hu, Raelynn Chen. ✉e-mail: ychen123@pennmedicine.upenn.edu; Ruowang.Li@cshs.org

and disease risk (called additive encoding), which restricts downstream analyses like PRS to only capture additive genetic effects[16–18]. While alternative models have explored capturing non-additive effects of individual SNPs, they often overlook the information already contained in pre-trained PRS[19,20]. Second, differences between PRS methodologies and the datasets they are applied to can result in suboptimal performance. Even within the same method, differences in parameter choice can cause potential predictive SNPs being omitted[21,22]. These inconsistencies can hinder the transferability of the models to different datasets. Third, due to differences in the populations used to develop pre-trained PRS and the intended target population, estimation biases and efficiency losses may occur[23]. As recent studies have shown, no single PRS methodology consistently outperforms others[21]. Consequently, each pre-trained PRS may omit different critical predictive signals, leaving potentially substantial room for improvement in polygenic risk prediction.

To address these challenges, we propose a pre-training and fine-tuning framework, Adaptive Boosting of pre-trained Polygenic Risk Scores (AB-PRS), as illustrated in Fig. 1. AB-PRS adaptively incorporates additional uncaptured genetic signals robustly selected from multiple datasets, with the goal of improving the predictive performance of PRS models. By utilizing a boosting algorithm[24], we treat pre-trained PRS as the strong learner to capture robust additive genetic effects estimated from large-scale training datasets, while integrating additional fine-tuned genetic signals not captured by the pre-trained model as weak learners to refine the model. Unlike traditional PRS methods, which rely solely on a training dataset, AB-PRS uses both training and validation datasets for the robust selection of fine-tuned predictors. Our

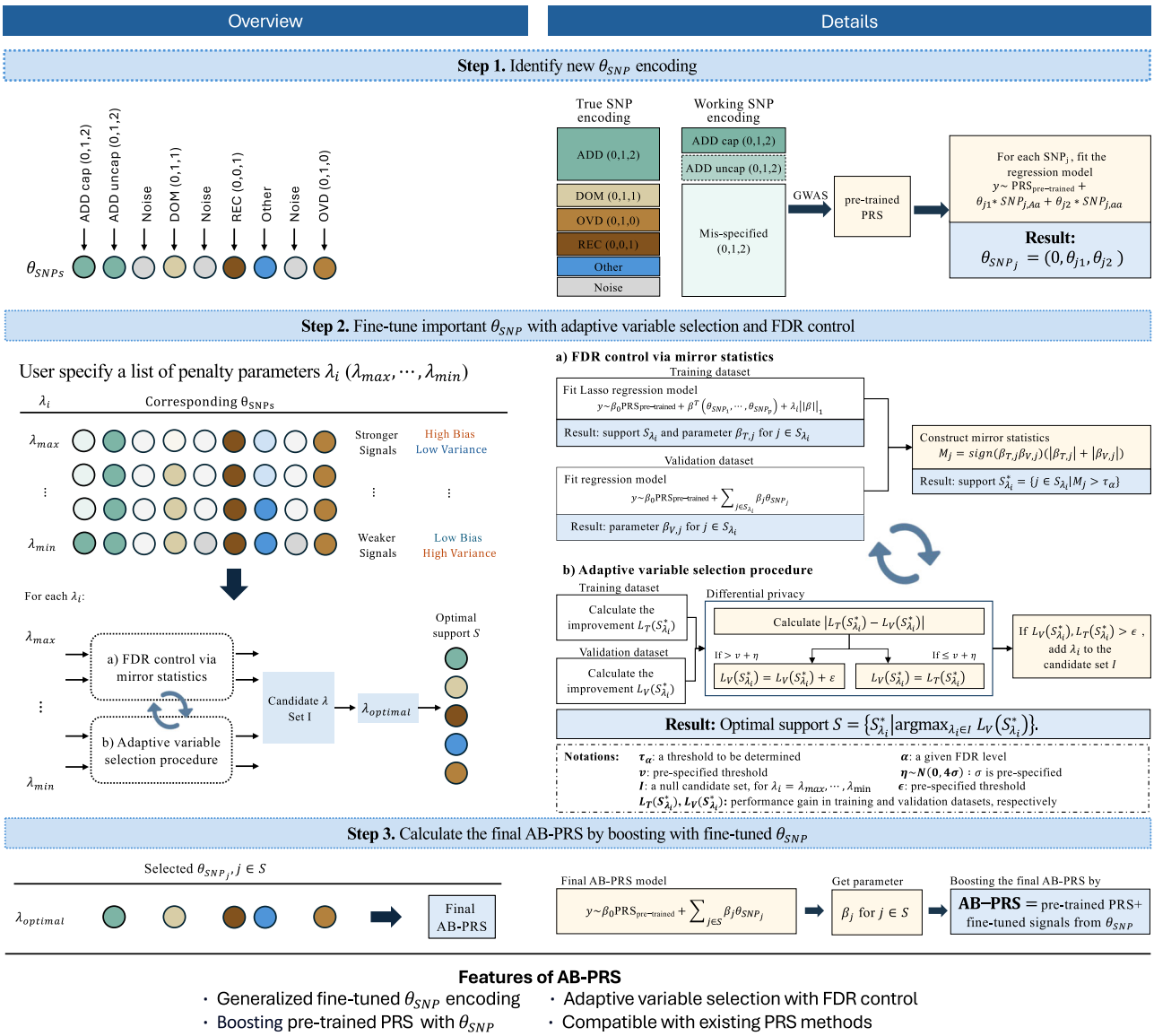


Fig. 1 | Schematic representation of Adaptively Boosting of Pre-trained PRS (AB-PRS) framework. The AB-PRS method consists of three key steps to fine-tune existing PRSs. In Step 1, a data-driven SNP transformation, θ_{SNP}, is constructed for each SNP to encode signals orthogonal to the pre-trained PRS. θ_{SNP} is designed to capture signals not accounted for by the pre-trained model, including both additive and non-additive effects. In Step 2, significant θ_{SNP} are selected for fine-tuning using an adaptive variable selection process with FDR control. This step involves two substeps: Step 2a: Each θ_{SNP} is evaluated for robustness across multiple datasets to ensure generalizability and reduce the likelihood of false positives. Step 2b: An adaptive selection procedure is applied to optimize model parameters while minimizing overfitting risks from reusing the training dataset. In Step 3, the robust θ_{SNP} identified in Step 2 are integrated with the original pre-trained PRS through model boosting, resulting in the final fine-tuned AB-PRS model. The left panel provides a flow-process diagram of the method, while the right panel offers a detailed explanation of the method. ADD: additive; DOM: dominant; REC: recessive; OVD: over-dominance; ADD cap: captured additive; ADD uncap: uncaptured additive.

method employs a two-step process with false discovery rate (FDR) control to ensure the robustness and generalizability of selected signals. First, we use LASSO selection on the training dataset to identify potentially important fine-tuned predictors. To refine this selection, we then apply mirror statistics[25–28] on the validation dataset to filter out noise variables while controlling the FDR. In addition, the selection process includes an adaptive component that permits the repeated use of the validation dataset for iterative tuning of the LASSO regularization parameter, ensuring robust signal selection without overfitting[29].

Using simulation studies, we demonstrated AB-PRS's effectiveness to integrate uncaptured genetic signals to improve predictive performance and risk stratification compared to pre-trained PRS. We further evaluated AB-PRS on a range of complex diseases and traits, including Alzheimer's Disease (AD), Breast Cancer, Hypertension, Type 2 Diabetes (T2D), BMI, Cholesterol, High-Density Lipoprotein (HDL) and Low-Density Lipoprotein (LDL), using the UK Biobank (UKBB) dataset[30]. The models were validated across three large-scale independent datasets, All of Us (AoU)[31,32], the Penn Medicine Biobank (PMBB)[33], and eMERGE[34,35]. In all cases, AB-PRS performed comparably to or better than pre-trained PRS, demonstrating its potential for broad applicability in improving polygenic risk prediction.

Results

Demonstrating advantages and interpretability of AB-PRS by simulation

We conducted simulation analyses on continuous outcomes under mixed genetic effects to evaluate the ability of θ_{SNP} to capture signals omitted by the pre-trained PRS. We benchmarked the proposed fine-tuned AB-PRS against other comparable variable selection strategies, using pre-trained PRSs computed with PRS (basic), PRS-CS, LDpred and MegaPRS. Since the prediction performance exhibited similar patterns across these four baseline methods, we present the results based on PRS-CS in the main paper and provide the results for PRS (basic), LDpred and MegaPRS in Supplementary Section 1. The prediction performance of the methods was evaluated using R^2 that was calculated on testing data that is independent of the training and validation datasets. In addition, to evaluate the performance of θ_{SNPs} selection across different methods, we assessed the proportion of correctly identified θ_{SNPs} under various types of genetic effects in the low-dimensional setting, where Bayesian variable selection methods were also included.

The results in Fig. 2a, using baseline pre-trained PRS calculated from PRS-CS, indicated that the proposed AB-PRS method generally achieved superior prediction performance compared to other methods, as measured by R^2 across all scenarios. For all methods, prediction accuracy decreased as the number of causal variants increased or heritability decreased, because with more causal SNPs in LD or lower heritability, it became increasingly difficult to distinguish true signals from noise. Focusing on the LASSO method, we observed that its performance was even worse than the pre-trained PRS obtained from PRS-CS when heritability was 0.1, particularly with a relatively large number of causal SNPs or a small sample size. This highlights the importance of FDR control when selecting the θ_{SNPs}. Comparing AB-PRS and AB-PRS (single val), we found that when the number of causal SNPs was small, they showed comparable performance. However, AB-PRS outperformed AB-PRS (single val) when the number of causal SNPs was large, suggesting that reusing the validation set improved prediction performance under more realistic scenarios. These patterns are further supported by replicate-level paired difference tests, which confirm that AB-PRS achieves statistically significant improvements over the baseline pre-trained PRS across all simulated scenarios, whereas alternative methods exhibit either comparable or significantly inferior performance in certain settings. The simulation results with pre-trained PRS calculated using PRS (basic), LDpred and MegaPRS were presented in Supplementary Figs. 1–3, which showed similar patterns to Fig. 2a; therefore, we omitted further discussion here.

Regarding the θ_{SNPs} selection performance (right panel of Supplementary Fig. 4), the LASSO method selected nearly all important predictors for uncaptured ADD, DOM, REC, and OVD genetic effects, showing at least an 80% selection proportion across all scenarios. However, it also included many captured ADD and noise predictors, with a noise selection proportion of at least 50% across all scenarios, which negatively affected prediction performance. In contrast, the Bayesian method selected very few captured ADD and noise predictors, with a selection proportion of no more than 10% across all scenarios. However, it also selected substantially fewer important predictors; for example, it selected less than 25% of non-additive genetic effects shown in the first row when H = 0.1, resulting in decreased predictive performance. The AB-PRS method effectively controlled the FDR by selecting very few captured ADD and noise predictors (less than 28%), while also selecting many important predictors, over 50% of uncaptured additive, DOM, and COM SNPs across all scenarios. This approach combined the advantages of both the LASSO and Bayesian methods while avoiding their shortcomings, thereby resulting in improved prediction performance across all scenarios.

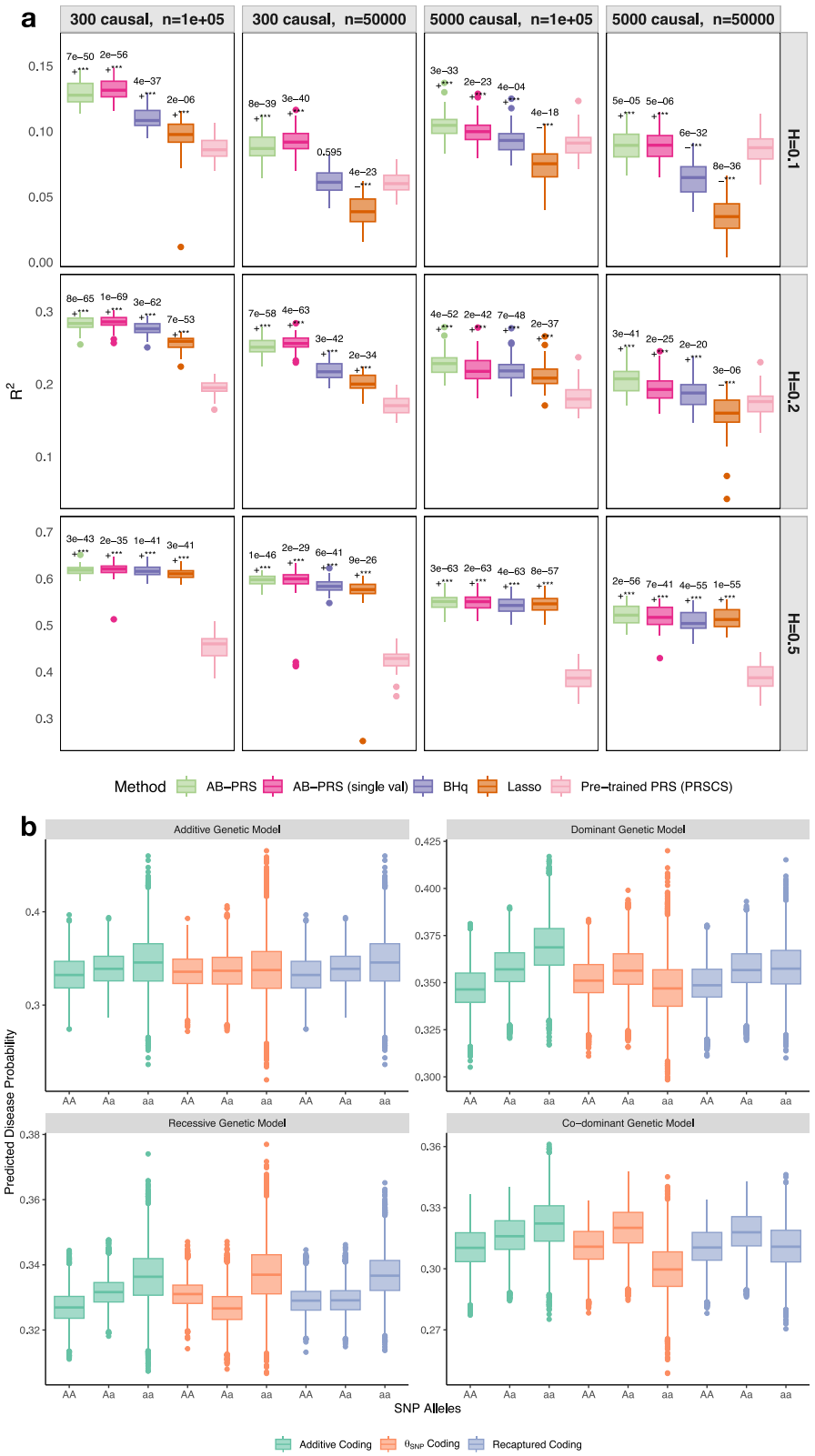
We then evaluated whether combining the fine-tuned θ_{SNP} with signals from an assumed additive model could recapture the true simulated genetic effects. Fig. 2b illustrates that when the true genetic effects follow an additive model, the θ_{SNP} encoding does not capture additional signals, as reflected by the similar predicted disease probabilities across the three allelic combinations. This effectively serves as a negative control. However, when SNPs exhibit non-additive signals (shown in the orange box-plots), the θ_{SNP} encoding produces distinctly different patterns. These variations highlight the discrepancy between the true underlying genetic model and the incorrectly specified additive model. For instance, the over-dominance coding (0, 1, 0) deviates more from the additive coding (0, 1, 2) than the dominant coding (0, 1, 1) leading θ_{SNP} to capture more signals under the over-dominance model than the dominant model for allelic combination aa. By augmenting the additive PRS with orthogonal signals from θ_{SNP}, the true genetic effects can be recovered, as demonstrated by the recaptured coding (blue box-plots).

Consistent improvement of AB-PRS in binary and continuous traits in UKBB

From the analysis of four binary traits (top row of Fig. 3a), AB-PRS generally yields AUC values that are comparable to or higher than those of the pre-trained PRS across all data sources. The most notable improvement is observed in the PGS-based models, where Alzheimer's disease (AD) achieves an 8.43% increase in AUC. For the consortium-based models, the largest gain appears for breast cancer, with an 11.12% improvement. In the FinnGen-based models, AD again shows the strongest enhancement, with a 2.31% increase in AUC. In several of these scenarios, the differences are statistically significant, while the remaining are competitive. These results indicate that AB-PRS yields measurable and robust improvements for binary phenotypes, though the degree of benefit varies across data sources.

For the continuous traits (bottom row of Fig. 3a), AB-PRS also improves the proportion of explained variance in several settings. The PGS-based model for cholesterol shows the most striking gain, improving 3.28 × over the pre-trained PRS. The PGS-based model for LDL also exhibits a notable enhancement of 31.46%, BMI of 27.59%, and HDL of 25.93%. Overall, AB-PRS improves predictive performance for both binary and continuous traits, with the degree of improvement depending on the trait and pre-trained model. For several traits, performance differences between AB-PRS and the pre-trained PRS are modest, reflecting the strength of the baseline models.

https://doi.org/10.1038/s41467-026-77128-5



To evaluate whether the AB-PRS method captures additional useful signals that are not identified by the pre-trained PRS, we investigated the θ_{SNPs} selected by the AB-PRS method and categorized them as ADD, DOM, REC, and OVD SNPs, respectively. The results for binary outcomes (top of Fig. 4) and continuous outcomes (bottom of Fig. 4) showed that AB-PRS effectively selected SNPs across different genetic models, capturing both uncaptured and mis-specified signals. The ratio of the sum of these three non-additive SNPs to all the selected θ_{SNPs} is over 50% for binary outcomes. However, for continuous traits, the missed additive signals are more dominant in the selected SNP set. The Manhattan plot (Fig. 5) also shows that many SNPs exhibit smaller p-values under non-additive models. For instance, numerous REC SNPs (brown triangles) are detected across all traits, showing the smallest p-values among the four genetic models. These results suggest that, in

Fig. 2 | Simulation results. (**a**) Simulation results for mixed types of genetic effects. Prediction accuracy was quantified by R^2 between the observed and predicted traits in the testing set. The four columns correspond to genetic architectures with 300 and 5000 causal SNPs simulated under a point-normal model, with sample sizes n = 50, 000 and n = 100, 000, respectively. The three rows correspond to heritability levels of 0.1, 0.2, and 0.5. For each scenario, simulations were independently replicated 50 times, and the baseline pre-trained PRS was constructed using PRSCS. Statistical significance was assessed using two-sided paired t-tests based on replicate-level differences in R^2 between each method and the baseline pre-trained PRS. No multiple testing correction was applied. Symbols indicate statistically significant differences: "+" denotes higher R^2 than the baseline and "-" denotes lower R^2. Exact p-values are shown above each box, corresponding to paired comparisons with the baseline method. Statistical significance is indicated by * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$). No symbol is shown when the performance difference is not statistically significant. (**b**) Disease probability box plots for different SNP encoding. Data are simulated for binary phenotype under the same setup as (a). The simulation was independently repeated 100 times, with 100 effective SNPs simulated per replicate; therefore, each boxplot pools n = 10, 000 effective SNPs. For each SNP, three encoding schemes are considered: regular additive coding, θ_{SNP} coding, and recaptured coding. θ_{SNP} coding captures the measurements of the association bias for alleles Aa and aa (described in section 4.3). Recaptured coding combines the previous two encoding schemes and reconstructs the genetic signals (described in section 4.6.4). Boxes represent the interquartile range (IQR; 25th–75th percentiles) with the median indicated by the center line. Whiskers extend to the minimum and maximum values within 1.5 × IQR, and values beyond this range are plotted as outliers. Source data are provided as a Source Data file.

addition to additive models (green square), the AB-PRS method detects numerous SNPs with larger associations under the other three non-additive models across all traits. Since limiting the analysis to four pre-defined genetic models may underestimate the diverse genetic effects present in real-world data, we performed a clustering analysis of the selected fine-tuned SNPs to evaluate the types of models from the real data. From Fig. 6, we observed that some clusters display ADD signals, such as cluster 3 in Breast Cancer, Hypertension, T2D, and Cholesterol, while some other clusters show DOM signals, like cluster 2 in LDL. Some clusters exhibit REC signals, such as cluster 4 in AD and cluster 6 in Breast Cancer, and clusters like cluster 1 in Cholesterol show OVD signals. Additionally, the remaining clusters exhibited a great diversity of genetic effects in the real data. In summary, the results in Fig. 6 demonstrated that the AB-PRS method detects a diverse range of genetic signals across all traits in real data.

In addition to reporting the AUC for binary outcomes, we also computed the odds ratio (OR) (Fig. 7a) between the top 10% and bottom 10% risk quantiles derived from the pre-trained PRS and AB-PRS models, respectively. The most substantial improvement is observed for Alzheimer's disease (AD), where AB-PRS achieves a 72.44% higher OR overall, with a particularly strong gain among females (84.10%). Hypertension also shows clear improvement, with a 15.51% increase overall and 23.09% among females. For type 2 diabetes (T2D), the overall OR rises by 12.71% in females. Overall, these results confirm that AB-PRS strengthens risk stratification for most binary phenotypes in UKBB, with especially pronounced gains for AD and hypertension.

For the continuous traits in UKBB, the top recovery rate (TRR) (Fig. 7b) was used as an analog to the odds ratio to assess stratification performance. Across all four traits, AB-PRS consistently outperforms the pre-trained PRS, demonstrating an improved ability to identify individuals with extreme observed values. The most substantial improvement is observed for cholesterol, with a 21.77% increase overall and a stronger gain among males (30.62%). HDL also shows a clear enhancement, with TRR increasing by 12.56% for females and 9.10% overall. For LDL and BMI, the improvements are moderate but consistent, with overall gains of 12.34% and 10.94%, respectively. These results indicate that AB-PRS provides better alignment between predicted and observed trait extremes, suggesting improved predictive calibration for continuous phenotypes.

Demonstrating generalizability of AB-PRS through application to external testing Data

For evaluation on AoU, on the binary traits (top row of Fig. 3b), AB-PRS achieves AUCs that are broadly comparable to those of pre-trained PRS, with improvements observed in several settings across model sources. Among consortium-based models, breast cancer achieves the largest improvement with a 9.94% increase. For PGS-based models, one notable gain occurs for AD, showing a 4.04% AUC increase. For the continuous traits (bottom row), LDL in the PGS-based model shows the most pronounced improvement, with R^2 increasing by 59.5% over the pre-trained PRS, BMI improving by 28.62%, and HDL improving by 20.55%. Improvements are more modest for other traits and model sources.

Across the four binary traits of eMERGE (top row of Fig. 3c), AB-PRS demonstrates AUC performance similar to pre-trained PRS, and in multiple scenarios, exhibits superior discrimination. The PGS-based model for AD achieves the largest AUC improvement at 11.61%, followed by the consortium-based model for AD with a 4.27% increase. For the continuous traits (bottom row), the strongest R^2 gain is again seen in cholesterol under the PGS-based model, increasing 2.06 × , while BMI improves by 19.88%, and HDL by 17.78%. Overall, AB-PRS demonstrates effective external generalization in eMERGE, with particularly strong results for cholesterol and HDL.

In the PMBB dataset (top row of Fig. 3d), AB-PRS again maintains competitive AUC performance relative to pre-trained PRS, with gains observed in several configurations. The PGS-based model for AD yields the largest AUC gain of 8.74%, while the consortium-based model for breast cancer follows with a 3.19% improvement. The FinnGen-based models show smaller yet consistent enhancements across traits. For the continuous traits (bottom row), AB-PRS produces some of the most substantial variance gains observed externally. The PGS-based model for cholesterol improves 2.84 × , and HDL increases 29.91% in the same model, BMI by 26.03%, and LDL by 25.74%. For several traits, performance differences are modest, reflecting the strength of the baseline models and the inherent difficulty of further improvement.

The risk stratification performance for the four binary traits was compared between pre-trained PRS and AB-PRS across the three external datasets (Fig. 7a). Alzheimer's disease (AD) consistently shows the largest improvements, with the most substantial gain observed in PMBB (2.08 × among females), followed by AoU (58.45% among female) and eMERGE (68.42% among females). Type 2 diabetes (T2D) also exhibits notable improvements, particularly in AoU (4.46%) and PMBB (1.82%). Hypertension shows smaller but consistent increases, with the largest gain observed in AoU (7.11% among females). In contrast, breast cancer results are mixed, with small improvements in AoU (2.74%) and eMERGE (6.37%).

For the continuous traits in the external datasets (Fig. 7b), AB-PRS consistently improves the top recovery rate (TRR) compared with the pre-trained PRS. In the AoU dataset, the largest improvement is observed for HDL among males, with an 11.74% increase, followed by LDL among females (9.82%) and BMI among females (3.20%). In the eMERGE dataset, cholesterol shows the most substantial enhancement, with a 35.86% increase among males and a 19.10% gain among females. LDL also demonstrates strong improvement, with an 18.52% increase among females. In the PMBB dataset, the largest TRR increase is found for cholesterol among females (26.63%), followed by HDL among females (19.53%) and BMI among all individuals (5.20%). Overall, AB-PRS demonstrates robust generalization for continuous traits across external datasets, with particularly strong recovery of top-

https://doi.org/10.1038/s41467-026-77128-5

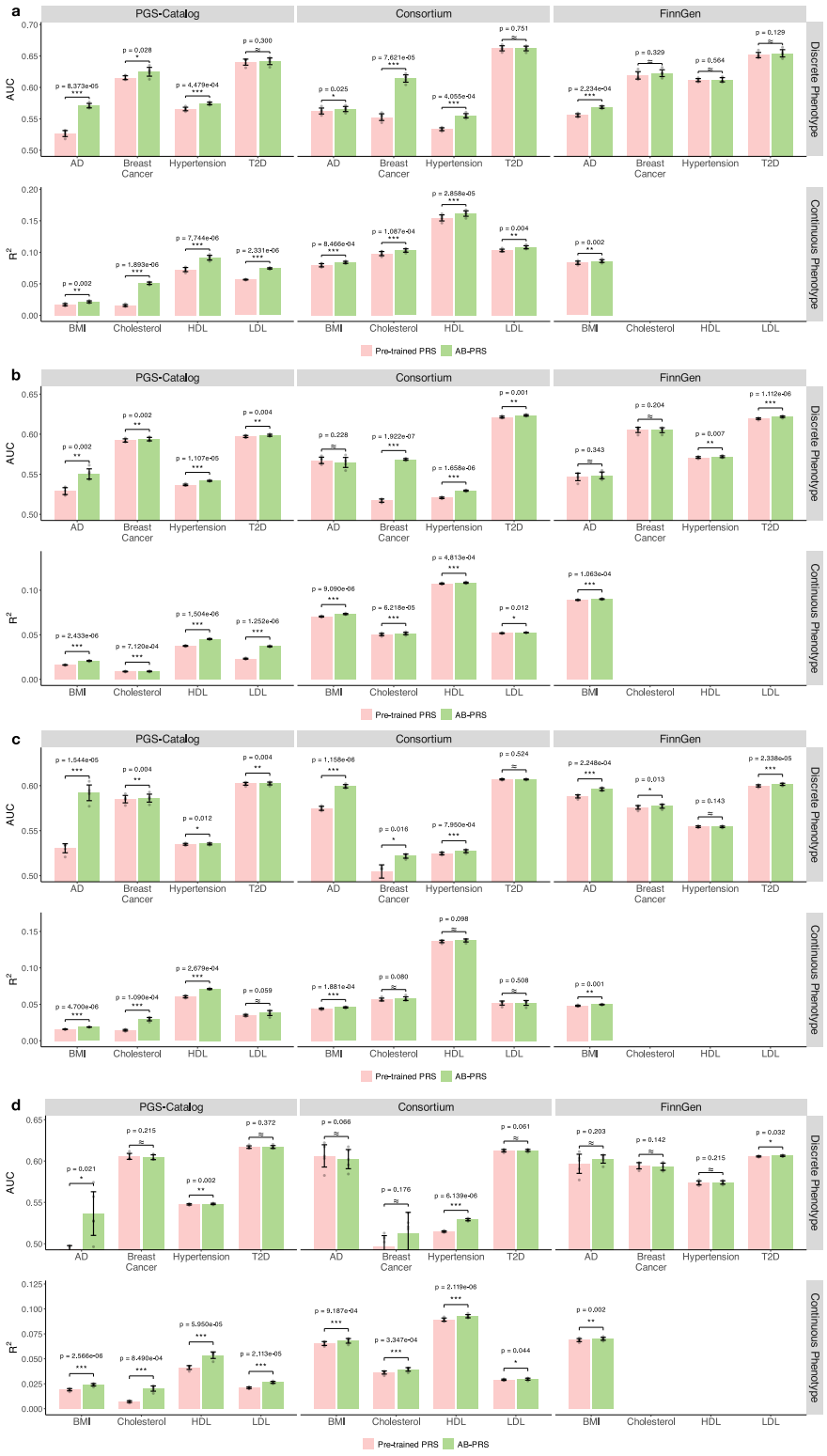


Fig. 3 | The testing data AUC for binary traits and the test data R^2 for continuous outcomes. Each subfigure corresponded to one evaluation cohort: **a** UK Biobank (internal testing data), **b** All of Us (external testing data), **c** eMERGE (external testing data), and **d** PMBB (external testing data). Within each subfigure, the first row showed the AUC of four binary traits (AD, breast cancer, hypertension, and T2D), and the second row showed the R^2 of four continuous traits (BMI, cholesterol, HDL, and LDL). For each row, three models were compared: PGS Catalog, consortium-based, and FinnGen-based models. For internal testing (UKBB), the confidence intervals were evaluated using five-fold cross-validation (CV), with each bar based on $n = 5$ CV folds, and the best-performing CV model was subsequently applied to the external datasets. For each external dataset, 20% of the data was used for fine-tuning in five non-overlapping splits, with each bar based on $n = 5$ fine-tuning/test evaluations. Bars show the mean metric value, and error bars represent the 95% confidence interval across replicates. Symbols indicate the results of two-sided paired *t*tests comparing Pre-trained PRS and AB-PRS within each phenotype; no multiple-comparison adjustment was applied. Statistical significance is denoted by * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$), while " ≈ " denotes a non-significant difference. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-026-77128-5

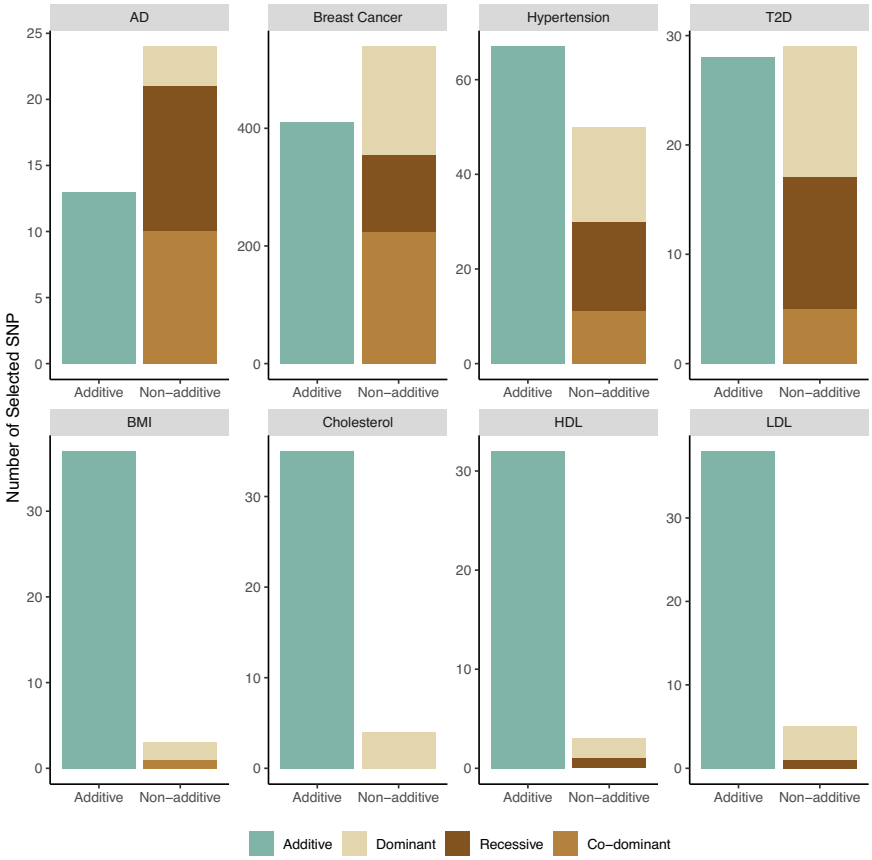


Fig. 4 | The selected number of different θ_{SNP} types for binary traits (top) and continuous traits (bottom). The selected θ_{SNPs} were categorized into ADD, DOM, REC, and OVD groups based on the *P*-values calculated under the specified SNP encoding. Specifically, for each selected θ_{SNP}, its encoding type was reassigned to ADD, DOM, REC, and OVD, and the regression models $Y \sim SNP_{ADD}$, $Y \sim SNP_{DOM}$, $Y \sim SNP_{REC}$, and $Y \sim SNP_{OVD}$ were fitted to obtain the corresponding *p*-values. Each SNP was then categorized into ADD, DOM, REC, or OVD if the corresponding *P*-value was the smallest and less than 0.01. If all *p*-values were greater than 0.01, the SNP was classified into the other encoding type and excluded from the figure. Association testing was performed in R using two-sided tests from logistic regression for binary phenotypes and linear regression for continuous phenotypes; no multiple-comparison adjustment was applied across the four SNP-encoding tests. Source data are provided as a Source Data file.

performing individuals for lipid- and BMI-related phenotypes, and consistent improvements across both sexes.

For the consortium-derived models, we analyzed four quantitative phenotypes across four external evaluation cohorts, each further divided by sex (male, female, and all), resulting in 48 TRR evaluations. Among these, 33 (68.8%) showed improved performance for AB-PRS compared with the pre-trained PRS. For FinnGen, which included one quantitative phenotype and three sex groups across four evaluation cohorts, a total of 12 TRR evaluations were performed, and 8 (66.7%) of them showed improvement with AB-PRS (Supplementary Figs. 5b and 6b).

For binary traits, we applied the same phenotype, cohort, and sex design; however, the male subgroup was not available for breast cancer, yielding 44 total evaluations. In these, 33 out of 44 (75.0%) consortium-based and 27 out of 44 (61.4%) FinnGen-based comparisons showed higher OR values for AB-PRS (Supplementary Figs. 5a and 6a).

Discussion

Numerous PRS have been developed for a variety of human diseases and traits. Open databases, such as the PGS Catalog, curate publicly available or user-uploaded pre-trained PRS developed using diverse data sources, methodologies, and phenotypes of interest. The availability of pre-trained PRS enables rapid application of these scores for various diseases. However, due to the diversity in PRS methods and training data sources, pre-trained PRS do not always capture the full spectrum of true genetic signals predictive of diseases.

In this study, we developed a method, AB-PRS, to improve the performance of pre-trained PRS by fine-tuning additional signals orthogonal to those identified by the original PRS. Notably, AB-PRS is applicable to scores generated by any polygenic risk score methodology, as the algorithm leverages existing scores to identify additional genetic signals not captured by the pre-trained model and incorporates these signals as components of the model. AB-PRS provides a framework for fine-tuning pre-trained polygenic risk scores. Existing methods that aim to improve the identification of additional genetic signals typically do not utilize the information available in pre-trained models. Instead, these methods often start from a null model, omitting the advancements achieved in existing PRS methodology development[19,20]. Furthermore, most current methods partition genetic signals into additive and non-additive effects, focusing primarily on the non-additive components that have yet to be captured[36]. In contrast, AB-PRS adopts a data-driven approach to capture any signals orthogonal to the pre-trained PRS, including uncaptured additive effects and genetic effects that extend beyond traditional discrete classifications (e.g., dominant or recessive). This flexibility allows AB-PRS to efficiently capture a broader spectrum of genetic signals.

Through simulation studies, we demonstrated that AB-PRS can fine-tune signals orthogonal to the pre-trained PRS and improve predictive performance by incorporating these additional signals,

https://doi.org/10.1038/s41467-026-77128-5

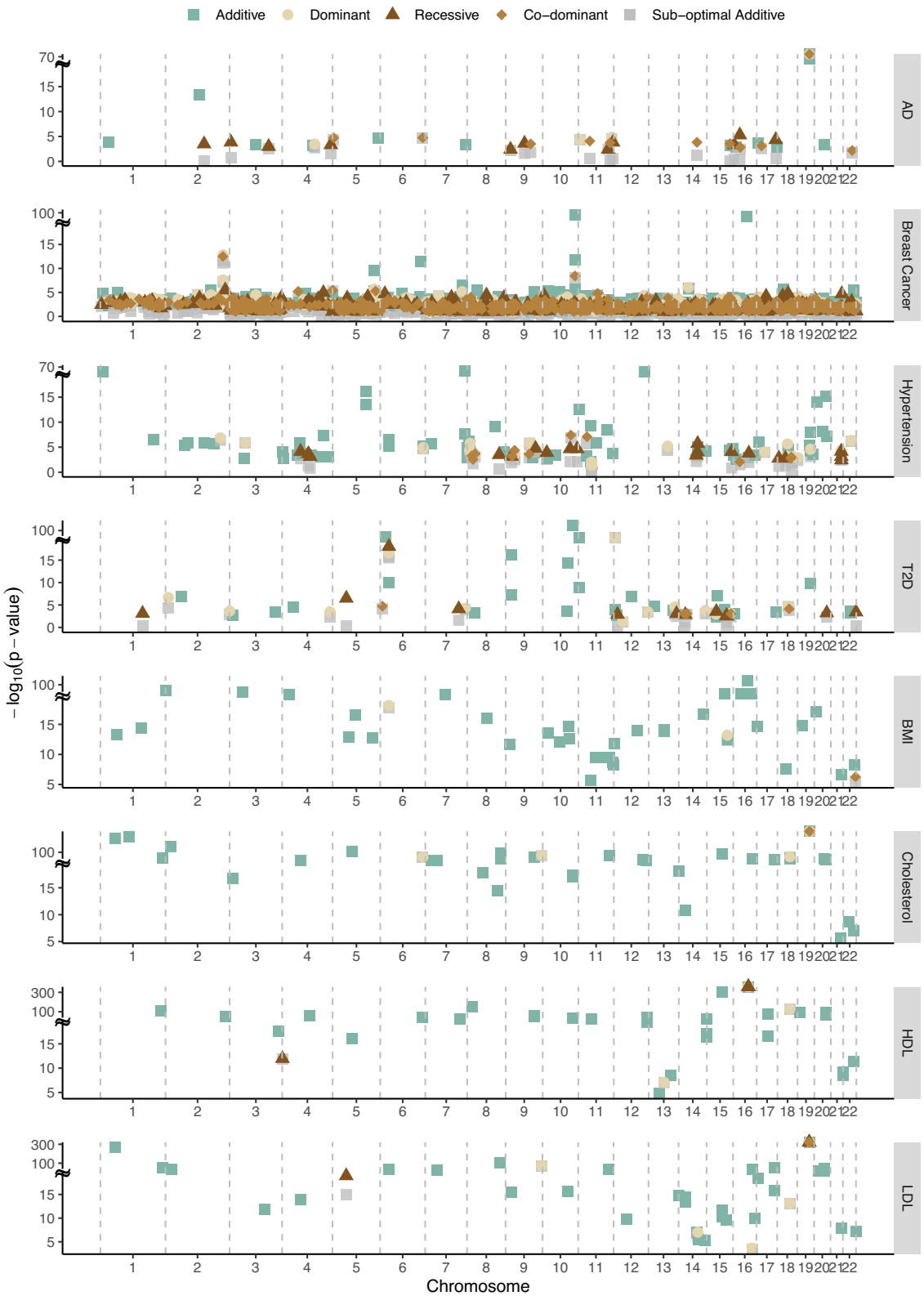


Fig. 5 | Manhattan plots for AB-PRS selected SNPs on UKBB internal testing data. Run GWAS with different genetic models for all selected SNPs on UKBB internal testing data. Identify the genetic model of each SNP by picking the most significant genetic signal based on *p*-value. For SNPs identified as non-additive, plot the sub-optimal additive signal as well. Association testing was performed in R using two-sided tests from logistic regression for binary phenotypes and linear regression for continuous phenotypes; no multiple-comparison adjustment was applied across the four SNP-encoding tests. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-026-77128-5

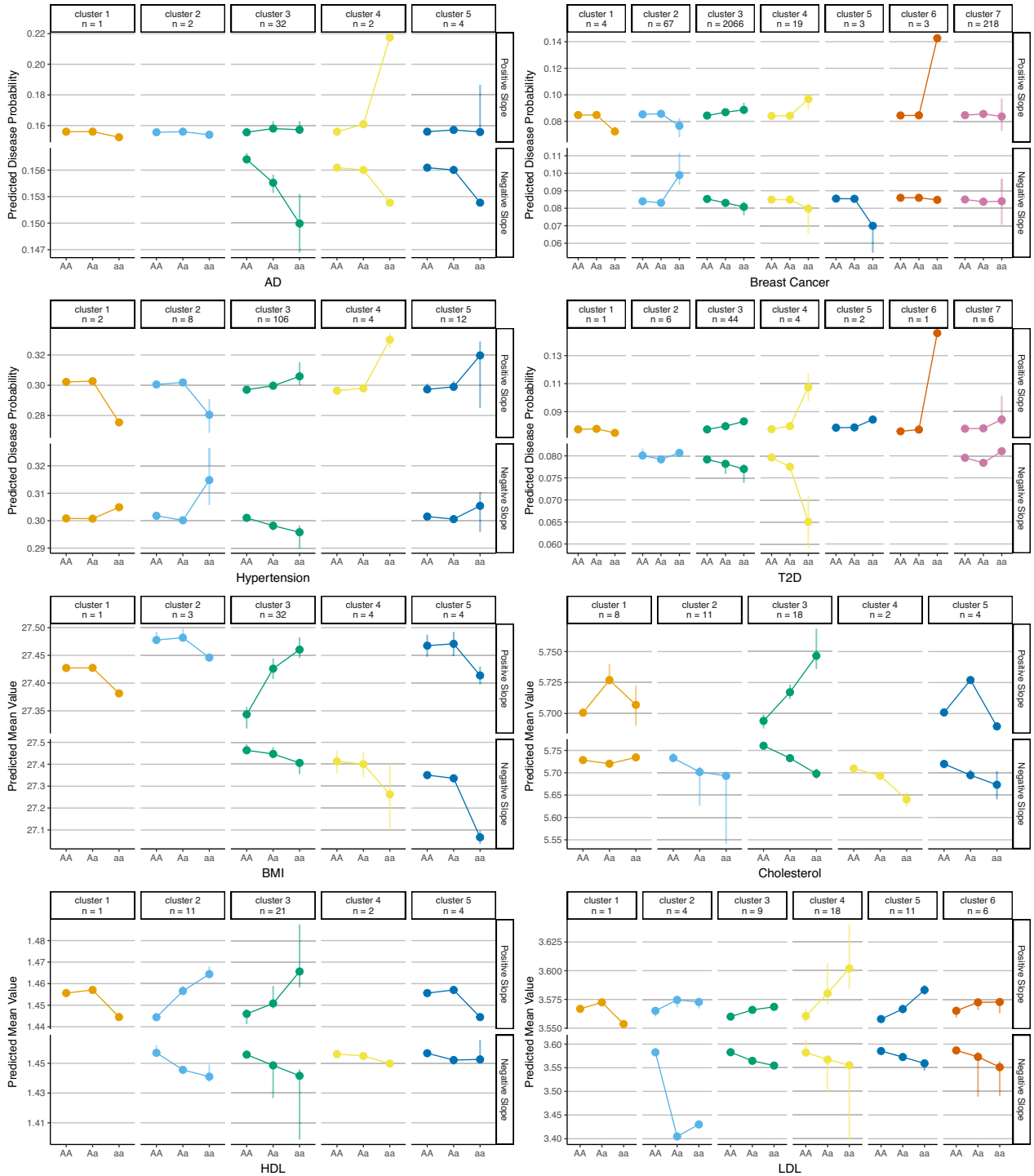


Fig. 6 | Recaptured coding of AB-PRS selected SNPs on internal testing data. K-means clustering was applied to all selected SNPs, with the number of clusters determined via the elbow method. For each cluster, SNP encoding patterns were recaptured and associated with predicted disease probabilities (for binary traits) or predicted mean values (for continuous traits) on the internal testing set. Subgroups were ordered by cluster centers, with the final group representing outliers beyond the 5–95th percentile slope ratio range. Positive and negative prediction components were separated to avoid signal cancellation. Points show the median predicted value, and vertical error bars represent the interquartile range, from the 25th to 75th percentile. *n* denotes the number of SNPs included in each cluster. Source data are provided as a Source Data file.

effectively boosting the pre-trained PRS. As illustrated in the right panel of Supplementary Fig. 4, when the true genetic effects consist of a mix of captured additive and non-additive effects, AB-PRS effectively identifies SNPs with non-additive effects and substantially increases predictive performance for the continuous traits. We also evaluated alternative variable selection approaches, including BHq, standard LASSO regression, and Bayesian type of methods. Bayesian generally has inferior power to identify nonadditive SNPs, achieving only 25% and 50% power to identify DOM and OVD SNPs. In comparison, LASSO regression exhibits nearly 80% power to select non-additive SNPs but

https://doi.org/10.1038/s41467-026-77128-5

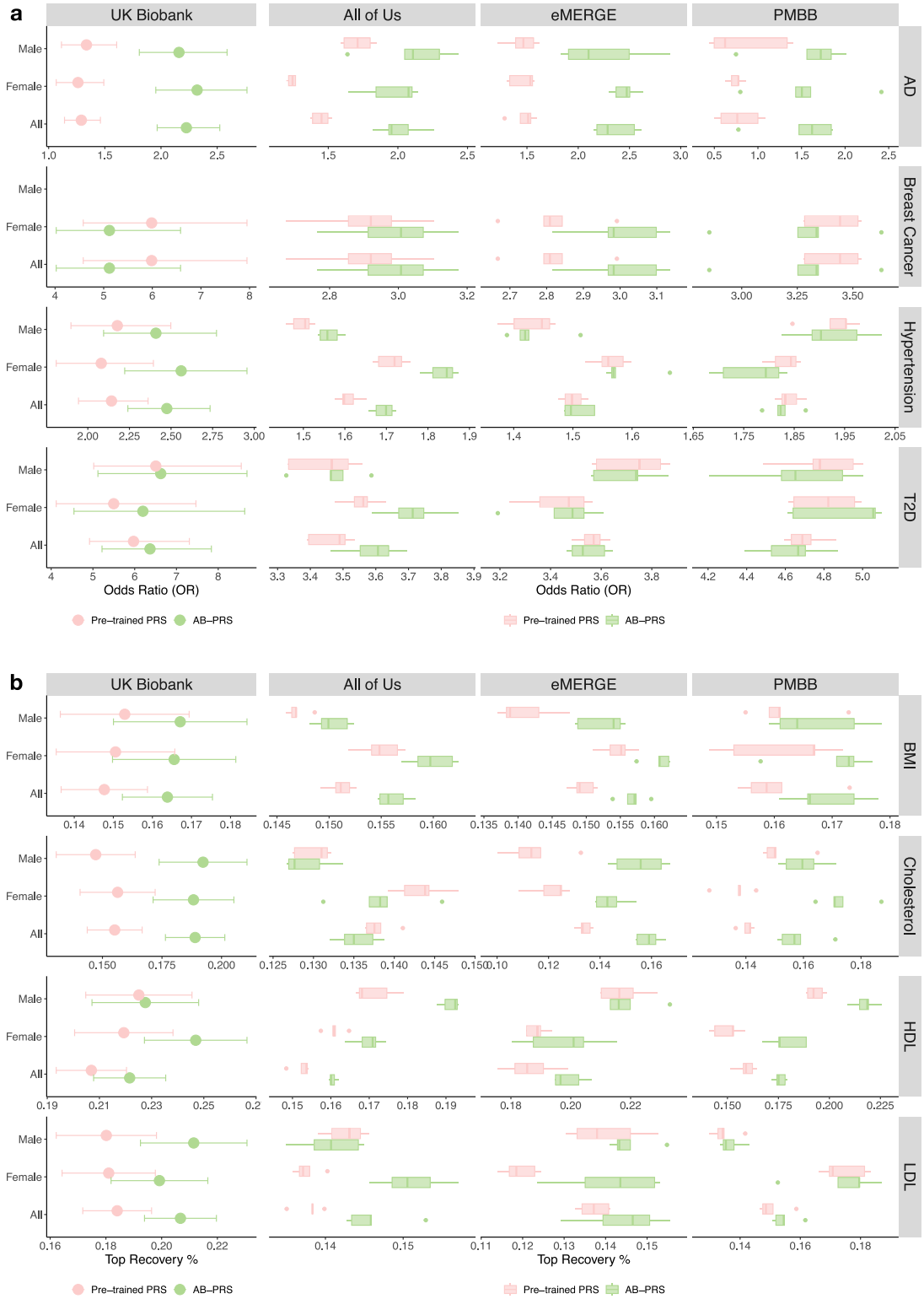


Fig. 7 | Comparison of odds ratio and top recovery rate between pre-trained PRS and AB-PRS. a shows the odds ratios (ORs) for binary traits, and subfigure **b** presents the top recovery rates (TRRs) for continuous traits, both comparing the stratification performance of AB-PRS and pre-trained PRS models trained using PGS-catalog data. In (**a**), ORs are presented for UK Biobank internal testing data as the mean OR (round dot) for the top 10% risk quantile versus the bottom 10%, with 95% confidence intervals (whiskers). For external datasets (AoU, eMERGE, and PMBB), ORs are calculated across five fine-tuning splits and shown using box plots. Breast cancer data were collected only for females, and male results are therefore omitted. In (**b**), the TRRs serves as an analog to the OR for continuous traits, quantifying the proportion of truly top 10% individuals (by observed phenotype) correctly identified within the top 10% of model-predicted values. UK Biobank results are displayed as means with 95% confidence intervals, and external dataset results as box plots. Case/control and sample size details for each phenotype are provided in Supplementary Data 1–4. Boxes represent the interquartile range (IQR; 25–75th percentiles) with the median indicated by the center line. Whiskers extend to the minimum and maximum values within 1.5 × IQR, and values beyond this range are plotted as outliers. Source data are provided as a Source Data file.

at the cost of largely increased false positives. LASSO selected between 50-60% of noise SNPs and 30-50% of additive SNPs. The substantial increase in the false positive rate diminishes its almost perfect detection power. Our approach demonstrates a balance between FDR and power, combining the strengths of both the LASSO and BHq methods while mitigating their limitations, therefore resulting in better prediction performance across all scenarios. Moreover, as shown in Supplementary Table 1, AB-PRS and Lasso exhibit comparable scalability in both runtime and memory usage. Although AB-PRS incurs slightly higher runtime than Lasso due to its additional FDR step, the memory usage remains similar. Therefore, if a computation is feasible for Lasso, it will also be feasible for AB-PRS. Figure 2b shows that when the true genetic effects are additive, the SNP coding, θ_{SNP}, does not capture additional signals, as indicated by the similar predicted disease probabilities for the three allelic combinations. In contrast, θ_{SNP} exhibits drastically different patterns when the simulated non-additive SNPs are modeled under the additive encoding (green boxplots). These differential patterns reflect the distances between the true coding and the mis-specified additive coding. For example, the over-dominance coding (0, 1, 0) is more dissimilar to the additive coding (0, 1, 2) compared to the dominant coding (0, 1, 1). Consequently, the θ_{SNP} coding captures more signals under the over-dominance model compared to the dominant model. By boosting the additive PRS with orthogonal signals from θ_{SNP}, the true simulated effects can be recaptured.

In addition, comparison results from the UKBB testing data (Fig. 3a), together with those from the three external datasets (AoU, eMERGE, and PMBB), indicate that AB-PRS generally matches and in several cases improves upon the performance of the pre-trained PRS by incorporating selected important θ_{SNPs}. Importantly, in each setting, the pre-trained PRS was chosen from among multiple state-of-the-art approaches (e.g., PRS-CS, LDpred2, and MegaPRS), thereby representing a near-optimized baseline for the target population. Under such a strong starting point, substantial further gains are inherently difficult to achieve. In this context, the ability of AB-PRS to preserve performance while achieving statistically significant improvements in multiple scenarios highlights its robustness and generalizability across different pre-trained PRS models. While we demonstrated that AB-PRS can capture omitted additive and additional non-linear genetic effects, including dominant, recessive, and over-dominance effects, in both simulation studies (Fig. 2a) and real data analyses (Figs. 4, 5), realistic scenarios present a more complex picture. The captured effects are not confined to these three types of genetic models (Fig. 6). The clustering results in real data analyses provide support for why the AB-PRS method is compatible with or outperforms pre-trained PRS, as it has the ability to identify numerous additional signals for fine-tuning that pre-trained PRS methods either miss or fail to capture. As shown in equation (1), the θ_{SNP} encoding can capture any signals orthogonal to the pre-trained PRS. Thus, if pre-trained PRS omit SNPs with true additive signals or SNPs with non-linear effects that do not strictly follow the three listed models, AB-PRS can capture complementary predictive components. We acknowledge that some of the observed non-additive signals may, in certain settings, reflect incomplete modeling of the underlying genetic architecture, for example, due to LD structure within genomic regions, rather than necessarily reflecting genuine non-linear genetic effects. When causal variants are not directly observed or adequately tagged, correlated SNPs may exhibit apparent non-linear associations due to model misspecification. This phenomenon may be more likely to arise for binary traits analyzed under logistic regression, because marginalizing over an unobserved correlated variant generally produces a mixture of logistic curves, which does not necessarily preserve a single logit-linear relationship in the observed SNP. As a result, deviations from additivity may occur even when the underlying causal effect operates additively on the log-odds scale. To illustrate this possibility, we provide a simple proof-of-concept simulation in Section S.1.1 of the Supplementary Section 1.1. The example represents a deliberately simplified and stylized configuration intended to clarify the mechanism, whereas real genomic architectures are typically more complex and heterogeneous. It serves as a conceptual illustration of how omission of a correlated unobserved variant under logistic regression can induce apparent departures from additivity at the observed SNP. In line with this possibility, we observed a higher frequency of detected non-additive signals for binary traits compared to continuous traits (Fig. 5). However, we do not interpret this pattern as definitive evidence of model misspecification, as true biological dominance or interaction effects may also contribute. Furthermore, phenotype misclassification or heterogeneity in case definitions may introduce additional distortions that manifest as apparent non-additive patterns, particularly in complex diseases with variable diagnostic criteria. Importantly, even when such deviations arise from incomplete tagging rather than biological dominance per se, modeling these effects can still improve predictive performance, which is the primary objective of AB-PRS. In addition, the risk stratification analysis of odds ratios between high-risk and low-risk groups (Fig. 7a) shows that AB-PRS generally yields comparable or higher OR estimates across most datasets and binary traits, with the exception of breast cancer. These findings suggest that AB-PRS maintains effective risk stratification performance while incorporating complementary predictive signals. We also examine potential gender differences in Fig. 7a. The results indicate that the variation between genders is minimal, suggesting that the OR estimates are consistent across both genders.

There are several key considerations and limitations in this study that warrant future research. First, many biobank-linked EHR datasets, such as the UK Biobank, are widely used for developing PRS. Consequently, one of the pre-trained PRSs (Hypertension from PGS catalog) utilized in this study was developed using the same datasets, leading to sample overlap in the training dataset. However, it is important to note that this overlap was limited to only the training phase and did not affect the independent validation cohorts. Second, substantial variations exist among cohorts in terms of participant and phenotype characteristics. For instance, eMERGE and PMBB cohorts exhibit nearly double the breast cancer rate and 1.5 times the T2D rate compared to AoU. Meanwhile, the UKBB has a higher prevalence of AD cases due to the use of a proxy phenotype based on parental diagnosis and age. These differences underscore the challenges of developing generalizable models for diverse populations with varying geographic locations and demographic profiles. Furthermore, there are variations in the methodologies used to develop the pre-trained PRSs, ranging from genome-wide significant SNP selection to more advanced regression-based and Bayesian approaches. These factors collectively emphasize the importance of model fine-tuning, as pre-trained models are unlikely to fully account for cohort- and model- specific variations. Third, while the fine-tuning strategy improved the prediction and risk stratification for most pre-trained PRSs, breast cancer was an exception. Although the prediction AUC for the breast cancer AB-PRS was consistently equal to or better than the pre-trained PRS, the risk-stratification analysis yielded mixed results. Similar or positive outcomes were observed in eMERGE and AoU, whereas PMBB showed negative results. This could be attributed to several factors.

Breast cancer, unlike other traits, is a more complex disease for digital phenotyping. The PheCode algorithm does not currently account for the heterogeneity of breast cancer subtypes, which are likely to vary substantially across datasets. Additionally, the pre-trained PRS for breast cancer was already highly predictive compared to those for other traits, leaving less room for improvement through fine-tuning. Furthermore, inherent randomness in the thresholds used for risk stratification analysis could influence the results. While AUC measures overall prediction performance across all individuals, risk stratification analysis relies on predefined thresholds to classify

individuals into high- and low-risk groups. Exploration of alternative thresholds has demonstrated improved performance for breast cancer risk stratification. However, the results presented here were selected to emphasize the discrepancy between AUC and risk stratification analyses. In contrast, the fine-tuned PRS for AD demonstrated the greatest improvement among all binary diseases. This is largely because the pre-trained PRS excluded APOE SNPs, the most impactful genetic variants associated with AD. Excluding APOE SNPs is a common practice to assess the contribution of weaker SNPs to AD risk. In this study, the AD PRS served as a positive control to evaluate the ability of the AB-PRS method to identify and incorporate APOE SNPs through fine-tuning. Post-hoc analysis confirmed that APOE SNPs were successfully selected by the method. Together, these results illustrate a key property of the pre-train-fine-tune framework: when the base model is suboptimal, fine-tuning can substantially enhance performance and vice versa. In addition, the AB-PRS method requires local, individual-level genetic data for fine-tuning, which necessitates a higher level of access to datasets. Despite this limitation, the fine-tuning approach demonstrates large potential, as it enables continuous updates to PRSs with emerging evidence, ultimately improving the clinical applicability of genetic risk scores. Finally, a further limitation is that we did not evaluate cross-ancestry fine-tuning. Extending a large European-derived PRS to non-European populations is an important application, but it requires additional methodological development to handle cross-ancestry genetic architecture differences, as well as sufficiently powered non-European training and validation datasets. We identify ancestry-aware fine-tuning as a key direction for future work once larger multi-ancestry cohorts become available.

Methods

Ethics

All analyses were conducted in accordance with relevant guidelines and regulations for research involving human participants and in accordance with the principles of the Declaration of Helsinki. UK Biobank data were accessed under Application Number 86494, which has received ethical approval from the UK Biobank Research Ethics Committee. The information of individuals from All of US included in our analyses has been collected according to the All of Us Research Program Operational Protocol (https://allofus.nih.gov/article/all-us-research-program-protocol). The detailed consent process of All of Us is described on https://allofus.nih.gov/about/protocol/all-us-consent-process. The eMERGE Network data were accessed through controlled-access procedures with appropriate institutional approvals. Data from the Penn Medicine Biobank (PMBB) were used under IRB protocol #813913 at the University of Pennsylvania. All data used in this study were de-identified prior to analysis.

Overview of AB-PRS

AB-PRS is designed to improve the performance of pre-trained PRS by using a fine-tuning framework to incorporate genetic signals that may have been missed or biased in the original model. Examples of these fine-tuned signals include non-additive SNP effects overlooked in standard additive GWAS PRS, uncaptured signals due to variations in genetic architecture, the choice of PRS methodologies, parameter selection, and underlying model assumptions, as well as population differences and data heterogeneities between source and target datasets.

To capture these additional signals, we developed a fine-tuning framework that begins with a generalized SNP transformation, θ_{SNP} encoding (step 1 of Fig. 1), which identifies orthogonal genetic signals not accounted for by the existing pre-trained PRS. Next, a two-step procedure selects the predictive θ_{SNP} with FDR control (step 2 of Fig. 1). In this process, LASSO selection identifies potential predictors; the validation dataset then controls FDR and avoids overfitting with an adaptive component for iterative tuning of regularized parameters. By incorporating these selected fine-tuned signals through a boosting algorithm, AB-PRS aims to improve the performance of pre-trained PRS models (step 3 of Fig. 1).

Generalized θ_{SNP} encoding

As noted above, pre-trained PRS models can perform suboptimally in real-data applications due to bias or their inability to flexibly capture signals. Therefore, we aim to fine-tune the pre-trained PRS by incorporating additional signals that are orthogonal to the existing pre-trained PRS to improve the performance. These orthogonal signals are captured using the developed generalized θ_{SNP} encoding. To illustrate the main idea, we use the mis-specified additive coding as an example. Traditional methods, for instance, often rely on the GWAS approach to obtain pre-trained PRS. Specifically, if β_j is the association strength between the phenotype and SNP_j, then the pre-trained PRS is given by $\mathrm{PRS} = \sum_{j=1}^{p} \beta_j SNP_j$. However, existing GWAS typically evaluate SNPs using an additive model (ADD) and encode them as (0, 1, 2), regardless of their true encoding. In reality, the true SNP encoding is not solely additive, but can be other types, for example, dominant (DOM), encoded as (0, 1, 1); recessive (REC), encoded as (0, 0, 1); and over-dominance (OVD), encoded as (0, 1, 0). Mis-specifying those non-additive SNP encodings as only additive encodings would result in incorrect associations being captured by pre-trained PRS using GWAS. In other words, there may still be underlying associations that are either not fully captured or are captured with bias in pre-trained PRS. Additionally, truly additive SNPs might be omitted due to differences in PRS methodologies, the statistical power of source datasets to detect true associations, and other influencing factors.

In this work, we aim to develop a generalized θ_{SNP} encoding to measure the uncaptured or not fully captured associations for each SNP that are orthogonal to the pre-trained PRS. Specifically, for SNP_j we fit the regression model

$$Y \sim \beta_0 \cdot \mathrm{PRS}_{\mathrm{pre-trained}} + \theta_{j1} SNP_{j,Aa} + \theta_{j2} SNP_{j,aa}, \tag{1}$$

to obtain the coefficients θ_{j1} and θ_{j2} for $SNP_{j,Aa}$ and $SNP_{j,aa}$, where Y is the phenotype of interest. Here, $SNP_{j,Aa}$ and $SNP_{j,aa}$ are indicator variables for the heterozygote and homozygous risk-allele genotypes (AA as reference). Then, a generalized linear model can be fitted according/corresponding to the form of Y. The coefficients θ_{j1} and θ_{j2} can be interpreted as measurements of the associations for alleles Aa and aa of SNP_j, respectively, which are orthogonal to the pre-trained PRS. Then, each SNP is encoded using the θ_{SNP} coding, where $\theta_{SNP_j} = (0, \theta_{j1}, \theta_{j2})$. Essentially, θ_{SNP} provides an encoding for each SNP, capturing the signals that were not accounted for in the pre-trained PRS

Adaptive variable selection with FDR control and boosting selected θ_{SNPs}

Since only a subset of SNPs is expected to be informative for predicting disease risk, not all θ_{SNP} would contribute to fine-tuning. For instance, additive SNPs that are already well captured by existing pre-trained PRS are unlikely to have their corresponding θ_{SNP} values included. Therefore, variable selection techniques, like LASSO regression, can be employed to identify important θ_{SNPs}. It is desirable to control the False Discovery Rate (FDR)[37] in variable selection procedures. We propose the following procedure to control the FDR in LASSO models, which can otherwise capture a considerable amount of noise. Let S_0 denote the true signal support, and let S be the selected signal support, the FDR is denoted as

$$\mathrm{FDR} = \mathbb{E}(\mathrm{FDP}), \quad \text{where} \quad \mathrm{FDP} = \frac{|S \cap S_0|}{|S| \vee 1}, \tag{2}$$

where $a \vee b = \max\{a, b\}$, $|\cdot|$ denote the cardinality of a set, and FDP stands for "false discovery proportion".

Then, we will develop a selection procedure to identify the important θ_{SNPs} while simultaneously controlling the FDR. Let $\mathbf{Y}$ denote the $n \times 1$ vector of phenotype, $\boldsymbol{\theta}$ denote the $n \times p$ θ_{SNP} encoding matrix, and $\boldsymbol{\beta}$ denote the $p \times 1$ vector of θ_{SNP} effects. As illustrated in step 2a of Fig. 1, we split the data into training and validation datasets of size n_T and n_V, respectively, where $n = n_T + n_V$. Firstly, in the training dataset, for a given tuning parameter λ_i in a pre-specified tuning parameter set $\{\lambda_{max}, \ldots, \lambda_{min}\}$, we employ a regression model with LASSO penalty in the form

$$Y \sim \beta_0 \cdot \mathrm{PRS}_{\mathrm{pre-trained}} + \sum_{j=1}^{p} \beta_{T,j}\theta_{SNP_j} + \lambda_i \parallel \boldsymbol{\beta} \parallel_1, \tag{3}$$

to obtain the coefficient estimates $\beta_{T,j}$, and denote the selected support using LASSO under λ_i by $S_{\lambda_i} = \{j \in \{1, \ldots, p\} | \beta_{T,j} \neq 0\}$. Secondly, in the validation dataset, based on the support S_{λ_i} selected in the first step, we fit the following regression model

$$Y \sim \beta_0 \cdot \mathrm{PRS}_{\mathrm{pre-trained}} + \sum_{j \in S_{\lambda_i}} \beta_{V,j}\theta_{SNP_j}, \tag{4}$$

to obtain the coefficient estimates $\beta_{V,j}$ for $j \in S_{\lambda_i}$. Then, we consider using the method of mirror statistics proposed by Dai et al.[26] for variable selection while controlling the FDR. The mirror statistic for feature j from training and validation estimates is defined as

$$M_j = \mathrm{sign}(\beta_{T,j}\beta_{V,j}) f(|\beta_{T,j}|, |\beta_{V,j}|), \tag{5}$$

where $\mathrm{sign}(\cdot)$ is the sign function, $f(x, y)$ is a nonnegative, exchangeable and monotonically increasing function defined on $\mathbb{R}^+ \times \mathbb{R}^+$. We adopt the recommendation of Dai et al.[25] to employ $f(x, y) = x + y$, so that the mirror statistic M_j uses the sign of $\beta_{T,j}\beta_{V,j}$ and a magnitude that combines the contributions from both training and validation estimates. In other words, the magnitude reflects the sum of the absolute values of the estimates, taking into account whether they are aligned or opposite in sign. This choice has been shown to be optimal in a simplified setting, as illustrated by Dai et al.[25]. By definition, the mirror statistic M_j in equation (5) satisfies the following two properties:

(i) A θ_{SNP} is more likely to be considered relevant if its mirror statistic is larger.

(ii) The distribution of the mirror statistic M_j under the null is (asymptotically) symmetric about zero.Based on these two properties, given a FDR level α, it is reasonable to select the features by $S^*_{\lambda_i} = \{j : M_j > \tau_\alpha\}$, with the threshold τ_α determined by the minimal non-negative τ for which the estimated FDP is less than alpha:

$$\tau_\alpha = \min\left\{ \tau > 0 : \widehat{\mathrm{FDP}}(\tau) = \frac{\#\left\{j : M_j < -\tau\right\}}{\#\left\{j : M_j > \tau\right\} \vee 1} \leq \alpha \right\}. \tag{6}$$

This controls the FDR at the prespecified level α.

In step 2a of Fig. 1, we use both the training and validation datasets to select robust θ_{SNPs} for fine-tuning under a specific λ_i. Note that the tuning parameter λ_i controls the support size at the first LASSO selection step. A large tuning parameter may lead to the loss of important θ_{SNPs} and consequently a loss of power, while a small tuning parameter may result in a large number of noises being included, affecting the FDR control step. In step 2b of Fig. 1, we aim to identify the optimal tuning parameter by repeatedly using the training and validation datasets. Additionally, we employ an adaptive component proposed by Dwork et al.[29], which borrows ideas from differential privacy, to address overfitting issues arising from the reuse of both training and validation data. Consider a pre-specified tuning parameter set $\{\lambda_{max}, \ldots, \lambda_{min}\}$, and let I be a null candidate index set. For $\lambda_i \in \{\lambda_{max}, \ldots, \lambda_{min}\}$, we can obtain a corresponding support $S^*_{\lambda_i}$ through the first LASSO selection and second FDR control steps. Then we can calculate the corresponding improvements $L_T(S^*_{\lambda_i})$ and $L_V(S^*_{\lambda_i})$ in the training dataset and validation dataset, respectively. Here, the improvement $L_T(S^*_{\lambda_i})$ denotes the improvement of the loss obtained by the model $Y \sim \mathrm{PRS} + \sum_{j \in S^*_{\lambda_i}} \beta_{T,j}\theta_{SNP_j}$ over the loss obtained by the model $Y \sim \mathrm{PRS}$. For example, in the logistic regression, $L_T(S^*_{\lambda_i})$ can represent the increase in AUC achieved by the model $Y \sim \mathrm{PRS} + \sum_{j \in S^*_{\lambda_i}} \beta_{T,j}\theta_{SNP_j}$ over the AUC obtained by the model $Y \sim \mathrm{PRS}$. $L_V(S^*_{\lambda_i})$ is similarly calculated as $L_T(S^*_{\lambda_i})$, but it is based on the validation dataset. Specifically, we define

$$L_V(S^*_{\lambda_i}) = \begin{cases} L_V(S^*_{\lambda_i}) + \delta & 0.5em \text{ if } |\mathrm{L_V}(S^*_{\lambda_i}) - \mathrm{L_T}(S^*_{\lambda_i})| > \upsilon + \eta \\ L_T(S^*_{\lambda_i}) & 0.5em \text{ Otherwise}, \end{cases} \tag{7}$$

where $\delta, \eta \sim \mathcal{N}(0, \sigma^2)$ and υ is a pre-specified threshold. Given threshold $\epsilon > 0$, when both $L_T(S^*_{\lambda_i}) > \epsilon$ and $L_V(S^*_{\lambda_i}) > \epsilon$, add the λ_i into the candidate set I. In our analysis, we set $\sigma^2 = 0.025$, $\upsilon = 0.01$ and $\epsilon = 10^{-6}$. Finally, we can obtain the optimal support by $S = \{S^*_{\hat{\lambda}_i} | \arg\max_{\lambda_i \in I} L_V(S^*_{\lambda_i})\}$. Once the important θ_{SNP} set S is determined, the final AB-PRS is computed as

$$AB - PRS = \beta_0 \cdot \mathrm{PRS}_{\mathrm{pre-trained}} + \sum_{j \in S} \beta_j \theta_{SNP_j}, \tag{8}$$

where the coefficients β_0 and $\{\beta_j; j \in S\}$ are estimated from the regression model

$$Y \sim \beta_0 \cdot \mathrm{PRS}_{\mathrm{pre-trained}} + \sum_{j \in S} \beta_j \theta_{SNP_j}. \tag{9}$$

By doing this, the pre-trained PRS acts as the baseline predictor, while the selected θ_{SNP_j} terms provide additional refinements. The phrase "boosting" is used here in a conceptual sense to indicate the enhancement of the pre-trained PRS through fine-tuned SNP-level effects, rather than referring to iterative boosting algorithms, such as AdaBoost or gradient boosting[38].

Pre-trained PRS from public resources

The PGS Catalog[15,39] is an open database of published polygenic scores (PGS) for various traits and diseases, providing detailed information on their construction, validation, and performance. The pre-trained PGS models (hereafter referred to as PGS-based models) available in the catalog are widely used for evaluation and application due to their rigorous curation and standardization by the PGS Catalog team.

In addition to the PGS Catalog, we also incorporated pre-trained PRS from the FinnGen project[40], which combines extensive genotype and phenotype data from Finnish biobanks and national health registries. FinnGen provides large-scale GWAS summary statistics for numerous diseases and quantitative traits, enabling the derivation of high-quality PRS that capture population-specific genetic architecture.

Furthermore, we included PRS derived from several large consortium-based GWAS studies to ensure comprehensive coverage across key complex traits. Specifically, summary statistics were obtained for Alzheimer's disease from Jung et al.[41], for hypertension from Wojcik et al.[42], for breast cancer from BCAC Zhang et al.[43], for type 2 diabetes from DIAGRAM of Glucose et al.[44], for HDL, LDL, and total cholesterol from the GLGC study Graham et al.[45], and for BMI from the GIANT consortium Locke et al.[46]. These consortium-level resources provide well-powered and harmonized GWAS results across diverse populations, facilitating robust cross-cohort PRS evaluation.

Simulation studies

Simulating outcomes under mixed types of genetic effects. In the simulation settings, outcomes were generated based on a mixture of SNPs exhibiting additive, uncaptured additive, and non-additive genetic effects, where the non-additive effects comprised dominant, recessive, and over-dominance SNPs. In particular, the uncaptured ADD SNPs contribute to outcome generation but are not captured by pre-trained PRS. We simulated outcomes under the linear model:

$$Y = \beta_0 + \mathbf{X}^\top \boldsymbol{\beta} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{10}$$

where Y is a continuous phenotype, $\mathbf{X}$ is a $p \times 1$ matrix of mixed type of SNPs. Specifically, we first independently generated 50 uncaptured ADD, 50 DOM, 50 REC, and 50 OVD SNPs with a minor allele frequency of 0.05 under Hardy-Weinberg equilibrium. The remaining SNPs were taken directly from real genetic data in the UK Biobank to mimic realistic LD structure and genetic architecture observed in real-world analyses, with around 300,000 total SNPs. The effect size β_j was drawn from $\mathcal{N}(0, H/M)$ if SNP_j was causal, and set to 0 otherwise, where H is the heritability explained by genome-wide genetic markers (known as SNP heritability), and M is the total number of causal SNPs. We considered three heritability levels $H \in \{0.1, 0.2, 0.5\}$, corresponding to 10, 20, and 50% of the total phenotypic variance explained by genetic factors, respectively. Among causal SNPs, the first 200 non-additive and uncaptured additive SNPs were always set as causal, while the number of additive causal SNPs was varied between 100 and 4800. This resulted in total causal SNP counts of 300 and 5000, representing sparse and dense genetic architectures, respectively. For each combination of the genetic architecture and the total sample size $n \in (50,000, 100,000)$, the simulation was repeated 50 times.

Calculating pre-trained PRS. In all simulated datasets, we considered three approaches to derive pre-trained PRS, each following standard practices in the literature:

(i) PRS (Basic). A GWAS was performed under the additive genetic model, and the resulting effect size estimates were used to construct the pre-trained PRS.

(ii) PRS-CS. We also applied PRS-CS[11], a Bayesian regression framework that leverages continuous shrinkage (CS) priors to model SNP effect sizes. This method adaptively shrinks SNP effects according to their posterior inclusion probabilities while accounting for linkage disequilibrium (LD), often improving prediction accuracy over traditional GWAS-based approaches.

(iii) LDpred. We also included LDpred[47], a widely used method that adjusts GWAS summary statistics by explicitly modeling LD among SNPs. LDpred assumes a prior on the fraction of causal variants and employs a Gibbs sampling algorithm to generate posterior mean effect size estimates, which are then used to construct PRS. The implementation was obtained from the official GitHub repository (https://github.com/bvilhjal/ldpred).

(iv) MegaPRS. We further incorporated MegaPRS[48], a method designed to efficiently construct polygenic scores from large-scale biobank data by combining penalized regression, summary-statistic-based modeling, and cross-validation for tuning parameter selection. MegaPRS integrates multiple regularization frameworks (lasso, ridge, elastic net) and accounts for LD through a reference panel, enabling robust and scalable PRS estimation in large datasets.

For all four methods, uncaptured additive SNPs were excluded from the calculation of the pre-trained PRS.

Methods comparisons for variable selection under various simulation settings. We compared AB-PRS performance alongside a variable selection method without FDR control (classical LASSO[49]) and a method with FDR control (the Benjamini-Hochberg (BHq) procedure[37]). The BHq method relies on p-values and ensures exact FDR control when all p-values are independent. Benjamini and Yekutieli[50] later extended the BHq procedure to handle dependent p-values. We additionally considered Bayesian variable selection methods in this study. Specifically, we implemented the variational Bayesian variable selection approach using the `varbvs` R package[51]. Due to the substantial computational burden of Bayesian approaches, we restricted their inclusion to the low-dimensional setting ($p = 500$). For high-dimensional scenarios, we excluded Bayesian methods because of their prohibitive runtime. To assess whether adaptive reuse of the validation set improved prediction performance, we also considered a benchmark version of the AB-PRS method that used the validation data only once, which we referred to as AB-PRS (Single Val).

For simulations with mixed types of genetic effects, each dataset was divided into three non-overlapping partitions with 50% training, 25% validation, and 25% testing splits. AB-PRS was applied to the training and validation data to select θ_{SNP} with an FDR level less than 0.1, i.e., $\alpha = 0.1$, a threshold commonly used in the literature[25,26]. We also set $\alpha = 0.1$ in our real data applications. The predictive performance of all methods was evaluated on the testing data using R^2.

Recaptured coding by aggregating additive and θ_{SNP} coding. To assess whether θ_{SNP} can capture orthogonal signals not accounted for by traditional PRS, we simulated outcomes under four genetic models: additive, dominant, recessive, and overdominance. Each simulated outcome was associated exclusively with one type of genetic effect. Regardless of the true underlying effects, we modeled SNP associations using the additive model, following the common practice in GWAS. For each SNP_j, we calculated the corresponding θ_{SNP_j}. To evaluate whether combining the assumed additive model with the θ_{SNP} model could reconstruct the true genetic model, we summed the predictive probabilities for each allele from both models to form the recaptured coding. The predicted response value under recaptured coding is defined as

$$P_{recap}(x) = \frac{P_{add}(x) + P_{\theta}(x)}{2} \quad \text{for} \quad x = (\text{AA}, \text{Aa}, \text{aa}), \tag{11}$$

where $P_{add}(x)$ and $P_{\theta}(x)$ represent predictive probabilities for ADD SNP and θ_{SNP} under allele x, respectively. To minimize random variability, we repeated the entire process 100 times.

Application of AB-PRS to UKBB

UK Biobank data. The UK Biobank is a population-based cohort study with linked genetics and phenotype data, encompassing approximately 500,000 individuals representing diverse regions across the United Kingdom[30]. Genotyping was conducted utilizing two types of genotype arrays, namely the UK BiLEVE Axiom Array or the UK Biobank Axiom Array, organized into 106 batches and imputed using the merged UK10K and 1000 Genomes phase 3 reference panels. We performed a series of quality control (QC) measures to ensure the quality of the analyzed dataset. For sample-level QC, we retained individuals that passed the following criteria: (1). Only unrelated individuals were retained. Among related individuals, one individual from each pair was systematically removed to prevent undue influence from familial genetic connections. The threshold for relatedness was set at the level of second-degree relatives, as indicated by an identity-by-descent $\hat{\pi}$ value equal to or greater than 0.25. (2) Only individuals with self-reported White British ancestry were selected. This ensures compatibility with the population used to generate pre-trained PRS. (3) Individuals with matched self-reported and genetically inferred sexes. (4). Individuals with heterozygosity within three standard deviations from the mean. For SNP-level QC, we excluded SNPs that: (1) have a missing rate of more than 5%, (2) minor allele frequency less than 1%,

(3) have an imputation quality info score less than 0.8, (4) are duplicated or ambiguous (5) have a Hardy-Weinberg equilibrium p-value less than 10^{-10}. After quality control, 397,356 participants and 1,997,903 SNPs remained for analysis.

Phenotype and covariate extraction. For breast cancer, hypertension, and T2D, case and control statuses were determined using the widely adopted PheCode algorithm, which relies on the inclusion and exclusion of disease-specific ICD-10 codes[52]. For AD, we employed an AD-by-proxy algorithm that incorporates an individual's diagnoses, parental diagnoses, and age to determine case and control statuses[53]. The continuous traits were extracted from the UK Biobank (UKBB) data fields: field 21001 for BMI, field 30690 for cholesterol, field 30760 for HDL, and field 30780 for LDL. Only measurements from the first visit were used. In addition, covariates were extracted to be adjusted as confounders in the model, including principal components (field 22009), sex (field 22001 & 31), age (field 21022), and genotype measurement batch (field 22000).

Variants filtering. When applying the proposed AB-PRS method to the UKBB, we performed variant filtering to reduce computational time and prevent potential false-positive signals. For each trait, we selected candidate SNPs associated with the trait using a lenient p-value threshold of less than 0.05 based on published GWAS summary statistics[54]. After converting all candidate SNPs into generalized θ_{SNP} encoding, the next step was to select predictive θ_{SNPs} using adaptive variable selection and FDR control techniques. However, including all θ_{SNPs} for selection would be computationally and memory-intensive given the ultra-high dimensionality (e.g., 250,000–400,000 dimensions for real traits). Considering that the pre-trained PRS already captures the majority of strong signals, we assume that the remaining signals in the θ_{SNPs} are very sparse. Therefore, to reduce high dimensionality, we first used threshold methods to screen out the majority of θ_{SNPs} that are unlikely to be related to the outcomes. Then, we selected important θ_{SNPs} from the remaining subset using the methods described in Step 2 of the algorithm (Fig. 1). In this paper, we introduce two threshold methods for screening. One method relies on the p-value threshold, approaching it from the perspective of statistical significance. Specifically, we compute the p-value for each θ_{SNP} and select those with a p-value lower than a specified threshold (set to 0.001 in our analysis). Alternatively, we also selected SNPs based on improvements in AUC for binary outcomes or R^2 for continuous outcomes, approaching it from the prediction perspective. To make the results comparable between these two approaches, we select the top p_0 θ_{SNPs} with the highest AUC or R^2 improvement, where p_0 is the number of θ_{SNPs} determined by the p-value thresholding method.

Model training and evaluation. For the selected samples that passed quality control, we randomly split them into 80% for model training, 10% for model validation, and 10% for held-out testing sets. This choice was guided by a simulation study evaluating model performance across different training data proportions while keeping 10% of the data fixed for testing. The results indicated that performance was optimal when the training set comprised approximately 70–80% of the data (See Supplementary Fig. 1). We selected 80% to align with the 5-fold cross-validation (CV) setup. For the breast cancer phenotype, the same data split proportions were used; however, only female individuals were included. AB-PRS was applied to the training data to generate generalized θ_{SNPs}. The validation data was then used to select θ_{SNPs} that demonstrated robust and improved predictive performance. The model parameters were tuned using both the training and validation data through the adaptive process. Finally, we evaluated the performance of the fine-tuned AB-PRS in predicting phenotypic outcomes using the testing data. For binary and continuous phenotypes, the models' performance was assessed using the Area Under the Curve (AUC) and the R-squared (R^2) metric, respectively. To formally compare predictive performance across methods, we conducted paired difference tests. In the simulation studies, differences in R^2 between each method and the baseline pre-trained PRS were evaluated using two-sided paired t-tests across independent simulation replicates. For real data analyses, performance differences in AUC (for binary traits) and R^2 (for continuous traits) were assessed using paired tests across the five cross-validation folds under identical data splitting. To assess the ability of the AB-PRS model to identify individuals with high genetic risk of disease, we also calculated the odds ratio for the top 10% risk quantile versus the bottom 10%, with 95% confidence intervals. For continuous traits, odds ratios are not directly defined since the outcomes are not binary. To provide an analogous measure of stratification ability, we define a metric termed the top recovery rate (TRR). The TRR quantifies how effectively a model identifies individuals with extreme (top-ranked) observed trait values. Specifically, for each model, individuals are ranked according to their predicted values based on either the pre-trained PRS or AB-PRS. Let $\mathcal{T}$ true denote the set of individuals within the top p% of the observed (true) trait distribution, and $\mathcal{T}$ pred denote the set of individuals within the top p% of the model-predicted values. The top recovery rate is then defined as

$$\mathrm{TRR} = \frac{|\mathcal{T}\ \mathrm{true} \cap \mathcal{T}\ \mathrm{pred}|}{|\mathcal{T}_{\mathrm{true}}|}. \tag{12}$$

This metric measures the proportion of truly top-performing individuals correctly captured by the model-predicted top group. In this study, we use 10% as an illustrative threshold, analogous to the top-10% vs. bottom-10% comparison employed in the odds ratio analysis for binary outcomes.

Application of AB-PRS to external testing datasets

Introduction to datasets. The All of Us (AoU) program[31], a U.S. National Institutes of Health EHR-linked biobank, includes approximately 413,450 participants with a target enrollment of one million diverse individuals. Phenotypic data collection involves physical measurements, laboratory results, EHR information, and survey responses. Genetic information is obtained through short-read and long-read whole-genome sequencing, as well as microarray genotyping. AoU intentionally oversamples groups historically underrepresented in biomedical research, making it valuable for examining health outcomes across various demographics. To ensure participant privacy, AoU employs robust data security and de-identification protocols.

The Electronic Medical Records and Genomics Network (eMERGE)[35] constitutes a network of EHR-linked biobanks within several U.S. healthcare systems. As of November 2024, eMERGE comprises over 136,000 participants across its cohorts. The collected phenotypic data include a variety of health conditions and traits derived from EHRs, such as physical measurements, lab results, and clinical diagnoses. eMERGE was established to facilitate research on gene-disease associations and predictive genomic medicine by integrating clinical and genetic data. The network adheres to strict ethical guidelines and data-sharing agreements to protect participant confidentiality.

The Penn Medicine Biobank (PMBB)[33], based at the University of Pennsylvania, is an EHR-linked biobank with enrollment exceeding 260,000 participants as of November 2024. The dataset offers whole exome and genome-wide SNP arrays, identifying over 2.4 million rare single-nucleotide variants and insertion-deletion events, along with more than 39 million common variants. Genetic data are connected to EHR data, encompassing a wide range of health conditions and traits, detailed physical measurements, laboratory results, and prescribed medications. The biobank implements comprehensive data governance policies to ensure ethical use and privacy protection for participants.

Quality control for external testing data. The quality control process for participants in the external testing datasets involved several key steps to ensure data integrity and consistency. First, cohorts were restricted to participants of genetically-inferred recent European ancestry and whose genetically-inferred sex matched their self-reported sex. Related participants were excluded to reduce confounding due to shared genetics among relatives. Lastly, we removed any participants with missing phenotype data to maintain a complete dataset for subsequent analyses.

For continuous phenotypes, an additional range-based filtering was applied using ranges derived from the UKBB data. Specifically, trait values were required to fall within the following ranges: BMI (12.1212–74.6837 kg/m^2), cholesterol (10.818–278.28 mg/dL), HDL (3.942–79.218 mg/dL), and LDL (4.788–176.346 mg/dL). This filtering was applied to external testing data to ensure consistency in outcome ranges between the testing and training datasets.

Evaluation after local fine-tuning. To account for the differences between the external testing dataset and UKBB, and to mitigate potential overfitting, when applying the AB-PRS to the external testing dataset, we performed an additional fine-tuning step using 20% of the external testing data and evaluated the final fine-tuned model using the remaining 80% of the dataset. First,

$$y^{UKBB} \sim \alpha^{UKBB} + \beta_0^{UKBB} \cdot PRS^{UKBB} + \sum_{j\in S} \beta_j^{UKBB} \cdot \theta_{SNP_j}^{UKBB},$$

where PRS^{UKBB} and $\theta_{SNP_j}^{UKBB}$ are external PRS and AB-PRS selected θ_{SNP} on UKBB. Define $m_1 = \beta_0^{UKBB} \cdot PRS^{external}$ and $m_2 = \sum_{j\in S} \beta_j^{UKBB} \cdot \theta_{SNP_j}^{external}$, then run the following regression model on 20% of external testing data:

$$y^{external} \sim \alpha^{external} + \beta^{external} \cdot m_1 + \beta_\theta^{external} \cdot m_2$$

to get the estimates of $\beta^{external}$ and $\beta_\theta^{external}$. Then calculate the AB-PRS on the remaining 80% of the external testing dataset:

$$AB-PRS = \beta^{external} \cdot \left[\beta_{PRS}^{UKBB} \cdot PRS^{external}\right] + \beta_\theta^{external} \cdot \left[\sum_{j=1}^{p} \beta_j^{UKBB} \cdot \theta_{SNP_j}^{external}\right].$$

We divided the external dataset into five equal parts, each containing 20% of the individuals. For each part, we used it as the fine-tuning subset and evaluated AUC or R^2 on the remaining 80%. This procedure produced five evaluation values, from which the average model performance and the 95% confidence intervals were obtained.

Analysis of the fine-tuned genetic signals

Inferring the optimal genetic models of selected SNPs. For each selected SNP in each phenotype, we run separate outcome regression models under the ADD, DOM, REC, and OVD genetic models to obtain the corresponding p-values: p_{ADD}, p_{DOM}, p_{REC}, and p_{OVD}, respectively. The optimal genetic model for each SNP was determined by selecting the model with the minimum p-value among the four. A significance threshold of 0.01 was applied to ensure that the selected SNPs were representative of their respective genetic models; SNPs that did not meet this threshold were excluded from further analysis

Data-driven clustering of genetic models. Clustering analysis was performed using the selected SNPs to investigate the types of genetic coding observed in real data. The input of each selected SNP is obtained by first constructing the recaptured coding and then calculating the slope ratio $\frac{k_{aa,Aa}}{k_{Aa,AA}}$, where $k_{aa,Aa}$ is the slope of of predicted disease prevalence between allelic combination *aa* and *Aa* in recaptured coding, i.e. $k_{aa,Aa} = \frac{P_{recap}(aa) - P_{recap}(Aa)}{2-1}$, and similarly, $k_{Aa,AA} = \frac{P_{recap}(Aa) - P_{recap}(AA)}{1-0}$. If the slope ratio of falls below the 5% quantile or above the 95% quantile, we classify the corresponding θ_{SNP} as an outlier. For slope ratios within this central range, we applied the K-means algorithm to cluster them into different groups, with the optimal number of groups determined using the Elbow method[55].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The individual-level genotype and phenotype data analyzed in this study are available through the UK Biobank data access process at http://www.ukbiobank.ac.uk/register-apply/. The analyzed phenotypes, genotype data, trait definitions, and study cohorts are described in Section 4.7.1. GWAS summary statistics used for constructing the baseline PRSs were obtained from the Neale Lab UK Biobank GWAS release at http://www.nealelab.is/uk-biobank/. We used the imputed_v3 GWAS summary statistics from both sexes, except for breast cancer, for which the female-specific GWAS summary statistics were used. The GWAS labels used in this study were 21001 (body mass index), 30690 (cholesterol), 30760 (HDL), 30780 (LDL), AD (Alzheimer's disease), C50 (breast cancer), E4_DM2 (type 2 diabetes), and I10 (hypertension). The eMERGE dataset is available through the eMERGE Network and requires data access approval. Detailed information on accessing eMERGE data can be found at https://emerge-network.org/. The PMBB dataset is available through the Penn Medicine Biobank at the University of Pennsylvania. Access to PMBB requires an application and institutional review board (IRB) approval. More information about PMBB data access can be found at https://www.med.upenn.edu/biobank/. The All of Us Research Program stores data in a secure cloud-based research environment, and approved researchers can access participant-level data and analytical tools through the All of Us Researcher Workbench. Application details are available at https://www.researchallofus.org/register/. Further details regarding the AoU, eMERGE, and PMBB datasets, including cohort composition, available genotype and phenotype data, and study-specific sample selection procedures, are provided in Section 4.8.1. PGS Catalog is publicly available at https://www.pgscatalog.org/. The specific PGS identifiers used in the analyses are provided in Supplementary Data 5. Source data are provided with this paper.

Code availability

Source code for the AB-PRS method is freely available from the GitHub repository at https://github.com/Cedars-CIG/ABPRS, where the getting-started tutorial is also available at https://cedars-cig.github.io/ABPRS/. The code is released under the MIT License, and the same license is provided in the GitHub repository in a LICENSE file. The code is publicly available without access restrictions, subject to the terms of the license. A citable version of the code corresponding to this study has been archived on Figshare at https://doi.org/10.6084/m9.figshare.32145724[56].

References

1. Abdellaoui, A., Yengo, L., Verweij, K. J. H. & Visscher, P. M. 15 years of GWAS discovery: realizing the promise. *Am. J. Hum. Genet.* **110**, 179–194 (2023).
2. Loos, R. J. F. 15 years of genome-wide association studies and no signs of slowing down. *Nat. Commun.* **11**, 5900 (2020).
3. Hindorff, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).

4. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
5. Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Hum. Mol. Genet.* **28**, R133–R142 (2019).
6. Choi, S. W., Mak, T. S. H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).
7. Li, R., Chen, Y., Ritchie, M. D. & Moore, J. H. Electronic health records and polygenic risk scores for predicting disease risk. *Nat. Rev. Genet.* **21**, 493–502 (2020).
8. Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
9. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics* **36**, 5424–5431 (2020).
10. Choi, S. W. & O'Reilly, P. F. PRSice-2: polygenic risk score software for biobank-scale data. *GigaScience* **8**, giz082 (2019).
11. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y. C. A. & Smoller, J. W. Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).
12. Hoggart, C. J. et al. BridgePRS leverages shared genetic effects across ancestries to increase polygenic risk score portability. *Nat. Genet.* **56**, 180–186 (2023).
13. Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X. & Sham, P. C. Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.* **41**, 469–480 (2017).
14. Lloyd-Jones, L. R. et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* **10**, 5086 (2019).
15. Lambert, S. A. et al. Enhancing the polygenic score catalog with tools for score calculation and ancestry normalization. *Nat. Genet.* **56**, 1989–1994 (2024).
16. Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.* **4**, e1000008 (2008).
17. Hill, W. G. & Mackay, T. F. C. Ds Falconer and introduction to quantitative genetics. *Genetics* **167**, 1529–1536 (2004).
18. Manolio, T. A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
19. Ohta, R., Tanigawa, Y., Suzuki, Y., Kellis, M. & Morishita, S. A polygenic score method boosted by non-additive models. *Nat. Commun.* **15**, 4433 (2024).
20. Hall, M. A. et al. Novel EDGE encoding method enhances ability to identify genetic interactions. *PLoS Genet.* **17**, e1009534 (2021).
21. Monti, R. et al. Evaluation of polygenic scoring methods in five biobanks shows larger variation between biobanks than methods and finds benefits of ensemble learning. *Am. J. Hum. Genet.* **111**, 1431–1447 (2024).
22. Ni, G. et al. A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts. *Biol. Psychiatry* **90**, 611–620 (2021).
23. Kachuri, L. et al. Principles and methods for transferring polygenic risk scores across global populations. *Nat. Rev. Genet.* **25**, 8–25 (2024).
24. Breiman, L. Arcing classifier (with discussion and a rejoinder by the author). *Ann. Stat.* **26**, 801–849 (1998).
25. Dai, C., Lin, B., Xing, X. & Liu, J. S. False discovery rate control via data splitting. *J. Am. Stat. Assoc.* **118**, 2503–2520 (2023).
26. Dai, C., Lin, B., Xing, X. & Liu, J. S. A scale-free approach for false discovery rate control in generalized linear models. *J. Am. Stat. Assoc.* **118**, 1551–1565 (2023).
27. Xing, X., Zhao, Z. & Liu, J. S. Controlling false discovery rate using Gaussian mirrors. *J. Am. Stat. Assoc.* **118**, 222–241 (2023).
28. Hu, J. et al. Federated feature selection with false discovery rate control. *J. R. Stat. Soc. Ser. B* 88, 978–997 (2025).
29. Dwork, C. et al. The reusable holdout: preserving validity in adaptive data analysis. *Science* **349**, 636–638 (2015).
30. Bycroft, C. et al. The uk biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
31. All of Us Research Program Investigators 2019. The All of Us research program. *N. Engl. J. Med.* **381**, 676–668 (2019).
32. Ramirez, A. H. et al. The All of Us research program: Data quality, utility, and diversity. *Patterns* **3**, 100570 (2022).
33. Verma, A. et al. The Penn Medicine Biobank: towards a genomics-enabled learning healthcare system to accelerate precision medicine in a diverse population. *J. Pers. Med.* **12**, 1974 (2022).
34. Kho, A. N. et al. Electronic medical records for genetic research: results of the emerge consortium. *Sci. Transl. Med.* **3**, 79re1 (2011).
35. McCarty, C. A. et al. The Emerge Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genomics* **4**, 13 (2011).
36. Palmer, D. S. et al. Analysis of genetic dominance in the UK Biobank. *Science* **379**, 1341–1348 (2023).
37. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
38. Bühlmann, P. & Hothorn, T. Boosting algorithms: regularization, prediction and model fitting. *Stat. Sci.* **22**, 477–505 (2007).
39. Lambert, S. A. et al. The polygenic score catalog as an open database for reproducibility and systematic evaluation. *Nat. Genet.* **53**, 420–425 (2021).
40. Kurki, M. I. et al. Finngen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
41. Jung, S.-H. et al. Transferability of alzheimer disease polygenic risk score across populations and its association with alzheimer disease-related phenotypes. *JAMA Netw. Open* **5**, e2247162 (2022).
42. Wojcik, G. L. et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
43. Zhang, H. et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nat. Genet.* **52**, 572–581 (2020).
44. of Glucose, M.-A. et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
45. Graham, S. E. et al. The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021).
46. Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
47. Vilhjálmsson, B. J. et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
48. Zhang, Q., Privé, F., Vilhjálmsson, B. & Speed, D. Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nat. Commun.* **12**, 4192 (2021).
49. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288 (1996).
50. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).
51. Carbonetto, P. & Stephens, M. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal.* **7**, 73–108 (2012).
52. Wu, P. et al. Mapping icd-10 and icd-10-cm codes to phecodes: workflow development and initial evaluation. *JMIR Med. Inform.* **7**, e14325 (2019).
53. Jansen, I. E. et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).

54. Lab, N. Round 2 results of GWAS in the UK Biobank (2018). http://www.nealelab.is/uk-biobank/. Released 1st August 2018.
55. Thorndike, R. L. Who belongs in the family? *Psychometrika* **18**, 267–276 (1953).
56. Li, R. et al. Code for: a pre-train and fine-tune framework for adaptively boosting polygenic risk prediction (ab-prs) https://doi.org/10.6084/m9.figshare.32145724.(2026).

Acknowledgements

We acknowledge the UK Biobank. This research has been conducted using the UK Biobank Resource under Application Number 86494. We acknowledge the Penn Medicine BioBank (PMBB) for providing data and thank the patient participants of Penn Medicine who consented to participate in this research program. We would also like to thank the Penn Medicine BioBank team and Regeneron Genetics Center for providing genetic variant data for analysis. The PMBB is approved under IRB protocol# 813913 and supported by Perelman School of Medicine at the University of Pennsylvania, a gift from the Smilow family, and the National Center for Advancing Translational Sciences of the National Institutes of Health under CTSA award number UL1TR001878. We gratefully acknowledge all of the 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 [and/or samples and/or cohort] examined in this study. We thank the patient-participants of Penn Medicine who consented to participate in this research program.

Author contributions

J.H., R.C., Y.C., and R.L. conceived the study and developed the methodology. J.H. and R.C. implemented the methods, developed the code, performed simulations, and conducted real data analyses using UK Biobank and All of Us datasets. M.S. performed analyses using the eMERGE and PMBB datasets. O.B.O. assisted data processing in the All of Us dataset. O.W. developed the R package and tutorial. Y.L. and O.W. designed and created the overview figure. J.H., R.C., and R.L. wrote the manuscript. S.N.M., B.K., A.K., K.K., I.J.K., J.L.S., E.E.K., Y.L., and M.D.R. contributed to the curation and processing of the eMERGE dataset and provided expertise related to its use. Z.S.T. provided expertise in interpreting the results. W.G.L. contributed to the development of the methodology. All authors reviewed the manuscript, provided critical feedback, and approved the final manuscript.

Funding

J.H. and Y.C.'s efforts were supported in part by National Institutes of Health (U01TR003709, RF1AG077820, R01AG073435, R01DK128237). R.L. is supported by the National Institutes of Health grants P30 AG094848 and U01 AG066833 and in part by Cedars Sinai Department of Neurology/Jona Goldrich Center for Alzheimer's & Memory Disorders. R.C. and O.B.O. are supported by the Department of Computational Biomedicine, Cedars Sinai.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-77128-5.

Correspondence and requests for materials should be addressed to Yong Chen or Ruowang Li.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.