nature communications

Article

https://doi.org/10.1038/s41467-026-76154-7

Modeling nonlinear and interaction effects of spatiotemporal and nongenetic factors improves prediction for complex traits

Received: 31 October 2025

Accepted: 22 July 2026

Published online: 29 August 2026

Check for updates

Ross DeVito [1] & Melissa Gymrek [1,2,3]

Adjusting for nongenetic factors improves genetic association testing and polygenic scores, yet most studies rely on linear adjustments for limited covariate sets. Location and time covariates can proxy environmental exposures but are rarely included, and linear adjustments cannot capture their nonlinear effects or interactions. We adopt a null model approach where an auxiliary nonlinear model predicts phenotypes from covariates alone. This prediction is then included as an additional covariate in downstream analysis. Using 16 phenotypes in the UK Biobank, we show gradient boosted decision tree nulls including spatiotemporal features improve covariate modeling. Incorporating these nonlinear spatiotemporal covariate predictions improves polygenic prediction for all phenotypes (median 7.3% R^2 gain with BASIL). Model interpretation reveals covariate interactions including sex-specific age effects, seasonal patterns, and complex geospatial dependencies. Together, these results demonstrate that a small addition to existing polygenic score workflows can improve predictions and provide insights into environmental effects for complex traits.

Statistical methods for understanding the genetic basis of complex traits, such as genome-wide association studies (GWAS), polygenic scores (PGS), and fine-mapping, rely on identifying associations between genetic variants and phenotypes. Adjusting for other sources of phenotypic variation with covariates in these models is critical to improve statistical power and reduce confounding[1–3]. Common covariates include age and sex, which explain a substantial portion of variability for many phenotypes, as well as genotype principal components (PCs), which help control for confounding due to population structure or shared ancestry[1,3–6]. These covariates may also serve as proxies for unmeasured sources of variation[6], such as cultural or lifestyle differences.

Standard approaches for modeling covariates face multiple limitations. First, covariates are typically added as additional terms to linear models, which scale efficiently to handle large-scale data. However, this approach assumes that covariates have linear effects and do not interact with each other or with genetic variants included in the model. Second, although these models are simple and straightforward to interpret, violation of the linear model assumption can make model interpretation unreliable[7]. Third, most analyses rely on a limited set of covariates (e.g., age, sex, PCs), ignoring other features such as spatiotemporal information with potentially large impacts on phenotypes.

DeepNull[5] addressed the first of these limitations by modeling nonlinear and interaction effects among covariates. In their framework, a deep learning model predicts a phenotype from covariates (age, sex, genotyping array) alone. The prediction from this auxiliary model, which they refer to as a null model, is then used as a single covariate in downstream linear models. This two-stage process effectively removes complex effects of covariates not captured by standard models while remaining compatible with linear association testing implemented by existing tools. Their study showed this approach improved statistical power in GWAS simulations with covariate interactions or nonlinear effects without losing power when there are only linear effects, increased the number of significant signals in real

[1]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA. [2]Department of Pediatrics, University of California San Diego, La Jolla, CA, USA. [3]Department of Medicine, University of California San Diego, La Jolla, CA, USA. e-mail: mgymrek@ucsd.edu

https://doi.org/10.1038/s41467-026-76154-7

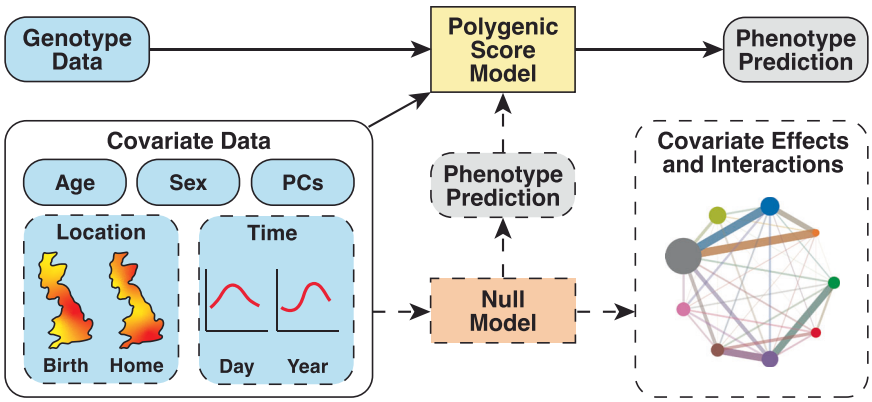


Fig. 1 | Overview of using a spatiotemporal covariate null model for polygenic scores. Standard elements of a polygenic score (PGS) model are denoted with solid outlines and arrows, while additional elements considered here (spatiotemporal covariates and nonlinear null models) are indicated by dashed outlines and arrows. A standard PGS is trained to predict a phenotype based on genotype data and covariates (blue boxes). In the null model framework, we extended this by adding an auxiliary nonlinear null model (orange box), which predicts the phenotype based on the covariates alone. In addition to standard covariates of age, sex, and genotype PCs, we incorporate location and time covariates to capture spatiotemporal effects and model interpretation techniques to identify important interactions between covariates. The null model prediction (gray box with dashed border) is then used as an additional covariate input to the PGS model of choice, allowing it to capture nonlinear covariate effects and interactions.

phenotypes, and improved PGS performance. While the DeepNull approach can in theory also address the second challenge of interpretation by overcoming model misspecification imposed by linear models, the model types used (deep neural networks) can be difficult to reliably interpret[8–11]. Whereas DeepNull focused solely on modeling covariate effects, alternative strategies have been developed that perform similar modeling of nonlinear effects and interactions across both genotype and covariate features[12–14]. However, those approaches necessitate aggressive feature selection to reduce the number of genetic variants considered in order to make training feasible, making it possible that truly causal variants will be excluded from the final model. As a result, we focus on covariate modeling here, though the null model approach could be integrated into the feature selection process of methods that jointly consider both genotypes and covariates in nonlinear models.

The third challenge is which additional covariates to include. The set of covariates included is critical regardless of the modeling approach used. DeepNull, like most standard GWAS, relied on a small number of commonly used covariates (e.g., age and sex). While some environmental covariates are well-known to influence certain traits (e.g., sun exposure for vitamin D[15] or physical activity for heart disease[16]), these features are often not readily available, and it can be cumbersome to manually identify relevant covariates for each trait. Instead, we sought to identify a set of general-purpose covariates capable of acting as proxies for a wide array of environmental factors. This led us to focus on spatiotemporal features: birth and home locations, and time of both year and day of assessment. These covariates are typically available in large cohorts and are broadly relevant across many traits, as they reflect spatial and temporal variation in environmental factors such as climate, pollution, and behavior. Importantly, their effects are unlikely to be well captured by linear models. For example, urban and rural areas present distinct exposures[17], while many biological and behavioral phenomena follow cyclical daily[18,19] or seasonal patterns[19,20]. Further, some biomarkers have been demonstrated to show interaction effects between sex and other covariates[21,22]. Nonlinear modeling is therefore particularly useful for capturing these complex relationships and improving the accuracy of covariate adjustment.

Here, we evaluated the impact of modeling nonlinear and interaction effects of spatiotemporal covariates on PGS prediction accuracy, focusing on 16 continuous and three case-control traits available for up to 316,867 White British individuals from the UK Biobank[23]. To model covariates, we adapted the null model framework from DeepNull, as it provides a modular and computationally efficient method to evaluate and integrate the contribution of nongenetic covariates. We evaluated three null model classes (linear regression, neural networks, and gradient-boosted decision trees) on different subsets of available covariates. We then integrated null models into polygenic scores using BASIL[24] and PRS-CS[25]. Finally, we applied a model interpretation method to identify specific covariate interactions that explain a large fraction of phenotypic variation.

Results

Overview of null model evaluation framework

We began by evaluating the performance of null models, which predict a phenotype of interest based on covariates alone (Fig. 1). We considered three variations for each of three classes of models: linear regression, neural networks, and gradient-boosted decision trees. The linear models served as a baseline for assessing the performance improvement of the two nonlinear model classes. For this class, we included a standard linear regression, as well as LASSO and ridge-penalized regressions.

For the neural network class, the first variation used the neural network architecture and training procedure used in the DeepNull study[5]. A potential weakness of the training procedure used there is that it is trained for a set number of iterations with a fixed learning rate and thus could be prone to overfitting. For the second and third neural network model variations, we therefore considered alternative training procedures that mitigate this risk by adding early stopping, weight decay, and a learning rate reduction on plateau (denoted as ESWP, see Methods). The second variation used the original DeepNull architecture, and the third used a smaller architecture with three hidden layers instead of four.

Finally, we considered gradient-boosted decision trees, which are generally easier and cheaper to train[26,27] and which the DeepNull study reported to show similar performance to their neural networks. For this, we used XGBoost[28] with early stopping. The three variations consisted of the default XGBoost parameters (a maximum of 100 estimators and a learning rate of 0.3), as well as two larger versions with up to 1000 or 2500 estimators, both using a learning rate reduced to 0.25 and a maximum depth of 4 to mitigate overfitting.

For each of the nine null model variations (three variations each for each class), we tested eight sets of covariate inputs. In addition to a baseline set with 1) just age and sex, which were included in all models,

https://doi.org/10.1038/s41467-026-76154-7

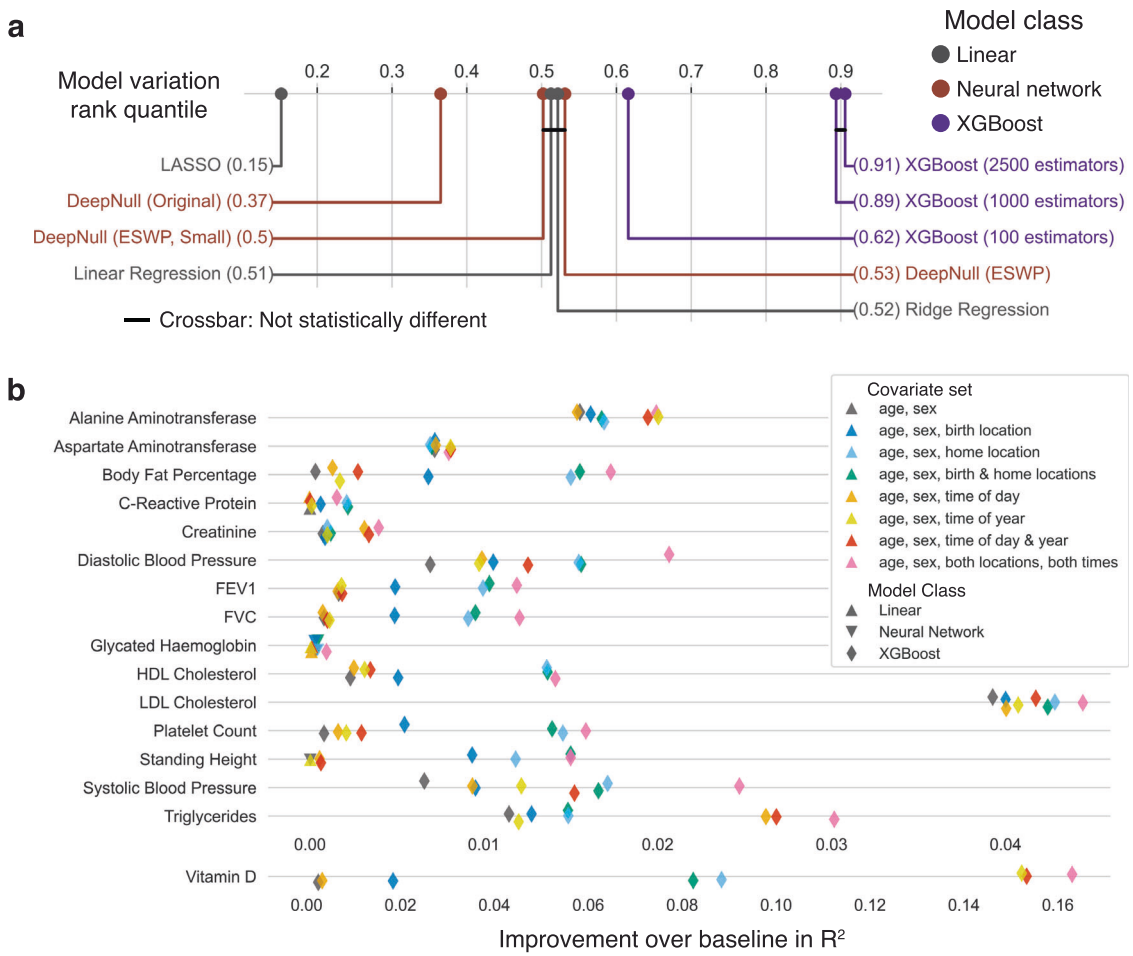


Fig. 2 | Null model evaluation results. a Model variation critical difference plot. The x-axis shows the quantile of a model variation's R^2 rank across all pairs, where a higher quantile indicates better predictive power on average. The horizontal crossbar connects model variants that were not statistically different with a Nemenyi post-hoc test α of 0.05 (following a significant Friedman test). Color indicates model class (gray=linear; red=neural network; purple=XGBoost). See Methods for details of model training, including use of early stopping, AdamW, and learning rate reduction on plateau (ESWP). **b** Best null model improvement for each covariate set over the baseline approach. The x-axis shows the improvement in R^2 of the best-performing model variation for each covariate set we evaluated compared to a baseline linear model with age and sex alone. Marker shape indicates the class of the model variation. Color indicates the covariate set. The best performer for each covariate set with or without PCs included is shown. The results in both panels are from a five-fold cross-validation using the training set. Scores underlying all components of this figure can be found in Supplementary Data 2 and the numbers of samples is the training set size for each trait in Supplementary Data 1.

the additional covariate sets added 2) home location, 3) birth location, 4) time of day of sample collection, 5) time of year of sample collection, 6) home and birth location, 7) time of day and year, and 8) all the time and location covariates. We tested each of these eight covariate sets with and without the first 20 genotype principal components (PCs) included as additional covariates.

Evaluating null model performance on continuous phenotypes

Each null model variation and covariate set was evaluated using 16 continuous phenotypes available for between 277,224 and 316,867 unrelated White British participants in the UK Biobank[23]. These included anthropometric measurements, markers of lung, kidney, and liver function, blood pressure and lipid levels, and other blood-assay-based biomarkers (Supplementary Data 1). We randomly divided participants into 8:1:1 train, validation, and test sets. For each phenotype, the subset of these sets for which phenotype data was available was used (Supplementary Data 1). We considered a total of 2304 models (phenotype/covariate/model variation combinations). To preserve the validation and test sets for the downstream PGS task (described below) while still providing predictions for all individuals, just the training set was used to train and evaluate the null models on their own using five-fold cross-validation.

We first evaluated the performance of the different model class variations (Supplementary Data 2). We performed two types of evaluations. First, we considered overall performance across the different covariate sets and phenotypes. For this, we used a nonparametric ANOVA Friedman test followed by a Nemenyi post-hoc test to determine pairwise significant differences between each model class variation (reviewed in [29]). This method is based on ranking each phenotype/covariate/model combination by performance, blocked by phenotype-covariate set pair (such that model variations were ranked separately for each of the 256 phenotype-covariate set pairs). We primarily used R^2 as the performance metric for each model, though the results when using mean absolute error (MAE) are similar (Supplementary Fig. 1).

We observed significant differences in performance across model variations (Friedman $\chi^2(8) = 1234$, p-value = 4.4e-261, Kendall's $W = 0.60$). Overall, the two largest XGBoost models (maximum either 2500 or 1000 estimators) showed significantly better performance than any other model variation (at Nemenyi $\alpha = 0.05$ for all pairwise comparisons; Fig. 2a). Across all phenotype/covariate combinations, the median improvement for the largest XGBoost model compared to a linear regression null was a 4.3% increase in R^2 and a 0.55% decrease in MAE. By average rank, the original DeepNull neural network was the second-worst performer, only outperforming the LASSO regression

and significantly worse than all other model variations. The neural network models with the additional modifications to prevent overfitting performed better, with results similar to the linear and ridge regressions. The smallest XGBoost model variation tested, which used default parameters (100 estimators) and early stopping, was the third-best performer; it was significantly better than all non-XGBoost models but significantly worse than the two larger variations.

Second, we identified the top-performing model variation for each phenotype or phenotype/covariate set pair. Overall, we observed that the XGBoost null models achieved the best performance by far the most often. The best performing model in terms of R^2 for all phenotypes was one of the larger XGBoost variants (10 phenotypes for 1000 estimators and 6 with 2500). In terms of MAE, 12/16 phenotypes performed best with an XGBoost model (8 with 2500 estimators and 4 with 1000), while the remaining four phenotypes (ALT, AST, C-reactive protein, and glycated hemoglobin) performed best with the original DeepNull neural network. When considering the best model variation for each phenotype across all covariate sets, the XGBoost variations were again the best performers most often. Linear null model variations performed best in terms of R^2 for 28/256 pairs and best in terms of MAE for just 2/256 pairs. The neural network null models were best for 6/256 and 29/256 pairs by R^2 and MAE, respectively.

We next compared the relative performance of different covariate sets for each phenotype, in this case blocking results by phenotype-model variation pairs to evaluate aggregate differences across all phenotypes (Friedman $\chi^2(15) = 718$, p-value = 3.2e-143, Kendall's $W = 0.33$). The covariate sets that included all additional location and time features, with and without PCs, showed top overall performance in terms of average R^2 and MAE ranks, though just adding the location features was not significantly lower (Nemenyi $\alpha \geq 0.05$). On the other hand, the baseline covariate set of age and sex was the worst performer by average rank. Adding PCs, time of day, or time of year alone did not significantly outperform just age and sex. Amongst the covariate sets that were significantly better than the baseline age and sex, median percent improvement in terms of R^2 across all phenotypes and model variations ranged from 0.3 to 2.5% (median 1.6% without PCs and 1.2% with PCs).

We characterized the top-performing covariate set by phenotype, in each case considering the model variation/covariate-set pair with the best performance for that phenotype (Fig. 2b). All best-performing models included additional time and/or location features. Using the additional time and location features without PCs was best for 12/16 phenotypes in terms of R^2 and MAE. A covariate set with different additional covariates performed best for the remaining four phenotypes: time of year for ALT, both time of day and year for AST, all locations for C-reactive protein, and time, location, and PCs for standing height. Due to the variation in covariate set performance by phenotype, we continued to evaluate different phenotype set options in the following section, where we integrated the null models with the additional time and location features into PGSs.

A possible alternative approach to using a nonlinear model is to encode covariates with combinations of basis functions. For example, the temporal features could be encoded with combinations of cyclic basis functions that match the cyclic nature of the underlying factor. We evaluated using cyclic basis function representations of the temporal covariates (time of day, month of year, and day of year) for vitamin D, triglycerides, and systolic blood pressure (Supplementary Data 3; Methods). We trained the nine model variations using age, sex, and all additional spatiotemporal covariates with and without the additional cyclic basis function representations. R^2 increased a median of 7.7% for the linear models across all variations and phenotypes and 3.0% for the neural network null models when the cyclic basis functions were included. For XGBoost null models, there was a median 0.1% decrease in performance. Even with the improvements from the cyclic basis function representations of temporal features, the linear and neural network models still performed worse than the XGBoost nulls without the basis functions. The neural network nulls were on average 5.9% worse for vitamin D, 10.2% worse for systolic blood pressure, and 22.3% worse for triglycerides. The linear nulls averaged 11.3%, 13.3%, and 41.1% worse, respectively. Overall, these results suggest the XGBoost models capture cyclical or other non-linear patterns without the need for knowledge of these patterns a priori and without the need to explicitly build them into the feature representations.

We tracked the runtime and cost of training the null model for the HDL cholesterol phenotype (Supplementary Data 4). The XGBoost and linear nulls had the lowest costs and runtimes, with both having averages of £0.08 and 5 min to train a single model on Amazon Web Services instances through the UK Biobank Research Analysis Platform (UKB-RAP). The cost of training a neural network null model was about 23 times higher, with an average cost of £1.85 and a runtime of an hour and 38 min. The original DeepNull training scheme was the most expensive: £3.31 with a time of 2 h and 55 min a model. Our modifications to the neural network training procedure resulted in average costs of £1.12 and £1.11, with average training times both rounding to 59 min.

Finally, we evaluated how our null models generalize to participants of non-White British self-reported ethnicity (Supplementary Data 5 and Supplementary Fig. 2). We compared the largest XGBoost null model using age, sex, and all spatiotemporal covariates to a linear ridge regression null with the same inputs across four self-reported ethnicities (Methods): Black or Black British (median n per phenotype 1940), Asian or Asian British (including Chinese; median 866), White and Black (median 674), and White and Asian (median 522). Performance generally declined among non-White British participants, varying substantially across groups. When compared against the White British XGBoost performance, the median decrease in R^2 for the XGBoost null was 77.2% for Black or Black British, 38.2% for Asian or Asian British, 16.4% for White and Black, and −4.8% for White and Asian participants (where a negative decrease indicates improvement). For the linear model, the corresponding decreases compared to the XGBoost null in White British were 114.7, 47.5, 22.0, and 9.2%. In all four cohorts, XGBoost outperformed the linear model, indicating that the nonlinear modeling advantage is maintained in non-White British populations. For several phenotypes, particularly FEV1, FVC, and vitamin D, R^2 was negative for some groups, indicating that the model predictions trained on White British participants were worse than predicting the group means. These results should be interpreted cautiously, given the small sample sizes and wide confidence intervals, but they suggest that while the null model approach generalizes, training on ancestry-matched data or data including the corresponding ethnicities would likely improve performance for underrepresented groups.

Overall, our results demonstrate that XGBoost achieved the best performance for the task of computing null models from covariates. Additionally, the XGBoost models showed more consistent performance across the covariate set options tested (Fig. 2a), are less expensive (Supplementary Data 4) and easier to train[27,30,31], and allow application of interpretability methods that are efficient with desirable properties (e.g., guarantees of local accuracy and consistency, generalization to any-order interactions)[7,32]. Therefore, the following sections focus on the XGBoost model with up to 2500 estimators as the null model, but continue evaluating different covariate set options across phenotypes.

Nonlinear null models of covariates improve PGS performance

Previous work found that including the null model output as a covariate input to GWAS should improve statistical power and thus improve prediction accuracy of PGSs derived from GWAS results[5]. We tested the impact of using our improved nonlinear null models for PGS construction using two separate methods. The first, PRS-CS[25], is a

summary statistics-based method in which the summary statistics of a single GWAS are adjusted according to a Bayesian model using an external linkage disequilibrium (LD) panel. The second, BASIL[24], is an iterative method in which GWAS is used to select candidate variants as input to a LASSO model. This process is repeated iteratively, with the following rounds fitting to the LASSO residuals from the previous iteration. Both methods are based on linear models. For the association testing steps for each method, we incorporated the null model prediction when the overall model included a null as well as individual terms for the additional covariates described above. Both model types used imputed common variants (MAF ≥ 1%).

We evaluated these methods' performance on the held-out test set with and without null models using the 16 continuous phenotypes considered in the previous section. We evaluated them using the standard covariates (age, sex, and genotype PCs), as well as location, time, and both location and time features, as covariates for the PGS model. For each of these covariate options, we tested a version with and without the prediction of the null model as an additional covariate. As in DeepNull[5], covariates were included as linear terms to the PGS model even when they are already modeled by the null model, since that model only captures the relationships between the covariates and the phenotype, not the covariates and the genotypes. To evaluate the relative benefits of including or excluding the PCs as input to the null model, we also tested a version using all additional covariates as well as the PCs as input. Performance for these variations for the BASIL models is shown in Fig. 3a and for PRS-CS in Supplementary Fig. 3. Full results for each covariate set and model choice are provided in Supplementary Data 6.

For all phenotypes for both BASIL and PRS-CS, the best performance was achieved when a nonlinear null model was used with the additional time and/or location features, with the best covariate set varying across traits. For most settings tested, BASIL models outperformed those from PRS-CS, but overall trends across covariate sets and models were similar for both. For 11/16 phenotypes for BASIL and 12/16 phenotypes for PRS-CS, performance measured in R^2 with a nonlinear null and additional spatiotemporal features was greater than the upper boundary of the 95% confidence interval (bootstrap $n = 100{,}000$) of the standard approach (age, sex, and PCs as covariates without a nonlinear null model). In contrast, just adding additional covariates without a null model resulted in performance outside this confidence interval for just 2/16 traits with BASIL (vitamin D and triglycerides) and 3/16 with PRS-CS (vitamin D, triglycerides, and systolic blood pressure). There were five cases with BASIL and 4 with PRS-CS where a nonlinear null model with just age and sex as inputs led to performance improvements outside the 95% CI of the standard approach. Of all 128 possible cases for which we tested the same set of covariates with and without the nonlinear null model, the null model only showed worse performance in two cases, both of which had R^2 decreases of less than 0.0005. GWAS that incorporated the null model with the additional spatiotemporal features resulted in an increased number of significant independent loci for 11/16 phenotypes (Supplementary Data 7), with a particularly dramatic improvement for vitamin D (increase from 2 to 42 loci). A two-sided Fisher-Pitman test[33,34] between the standard approach and an alternative per PGS model type resulted in a statistically significant (Bonferroni corrected p-value < 0.01) result where the alternative approach had a greater mean for 71/128 comparisons for PRS-CS and 87/128 for BASIL (Supplementary Data 8). There was only one case where there was a significant difference for which the alternative approach performed worse (BASIL using only additional time covariates for LDL-C). Overall, these results show that incorporating a null model with spatiotemporal features in almost all cases improves or does not harm performance. Further, the combination of spatiotemporal covariates with nonlinear null models provided greater improvement over incorporating either these additional covariates or null models alone.

The covariate set that achieved the best R^2 for the majority of phenotypes was the nonlinear null model prediction with the additional location and time covariates (best for 11/16 phenotypes with BASIL and 12/16 phenotypes with PRS-CS). For BASIL and PRS-CS respectively, the median improvement in R^2 for this approach was 7.3% and 15.5% compared to the standard approach of just including age, sex, and PCs as covariates to the PGS model, 7.5 and 11.4% compared to additionally including the location and time covariates as linear covariates to the PGS model, and 4.7 and 6.2% compared to using age and sex as covariates with a nonlinear null model. The optimal covariate set varied across phenotypes. For example, for BASIL, there were two phenotypes (C-reactive protein and standing height) for which the best performance was achieved when location but not time features were included, and one (aspartate aminotransferase) for which the best performance was when time but not location was included. BASIL PGS models with nonlinear nulls achieved R^2 outside the 95% CI of the standard approach for 11 of 16 continuous phenotypes each when using either additional location, location and time, or location, time, and PC features. For PRS-CS, this was for 12, 11, and 11 phenotypes, respectively. Using a null model with all additional spatiotemporal covariates never resulted in an R^2 or MAE lower than the 95% CI of the best performing variation when principal components were not included in the null model. When PCs were included in the null model's input, the only case of either metric being lower than the best model's confidence interval was R^2 for C-reactive protein.

We further investigated the relative contribution of our null models (XGBoost or linear) compared to the PGSs for prediction performance (Supplementary Fig. 4a). For all but three continuous phenotypes (body fat percentage, vitamin D, and systolic blood pressure), the standard BASIL PGS outperformed the null models (covariates) alone, with the best performance achieved for all cases when the PGS was combined with the XGBoost null. The median percent improvement for the standard BASIL PGS over the linear null using age and sex was 77.2%, while the median improvement over the XGBoost null with the additional spatiotemporal covariates was 20.1%. The median improvements for the BASIL PGS using the additional spatiotemporal covariates and an XGBoost null that incorporated those additional covariates were 99.0% over the linear null and 48.5% over the XGBoost spatiotemporal null. Our results suggest the null models and PGSs contain complementary information that, when combined, result in the best overall prediction.

To expand our comparison of XGBoost and neural network null models to their effects on downstream PGS, we trained additional BASIL and PRS-CS models for the four phenotypes where a neural network was the best-performing null model in terms of MAE. These PGSs used the prediction of the specific neural network variation and covariate set that performed best in the null evaluation. In all four cases, the original DeepNull[5] neural network was the best performer. For alanine aminotransferase, all spatiotemporal features were included; for aspartate aminotransferase, time of year was included; for C-reactive protein, time of day and the genetic principal components were included; and for glycated hemoglobin, all location features were included as input to the neural network. The BASIL and PRS-CS models had age, sex, PCs, and the spatiotemporal covariates as input in addition to the prediction of either the largest XGBoost null with age, sex, PCs, and the spatiotemporal covariates as input or the corresponding best neural network null. The PGSs using these neural network nulls performed worse than those using XGBoost nulls in all cases (Supplementary Data 9). The difference in performance was significant for all phenotypes except for aspartate aminotransferase (Bonferroni corrected two-sided Fisher-Pitman test[33,34] p-values < 0.001). Even for cases where neural network-based nulls achieve the best covariate-only prediction performance, the XGBoost nulls appear to be slightly more robust for adjusting for covariate effects in the full PGS models.

https://doi.org/10.1038/s41467-026-76154-7

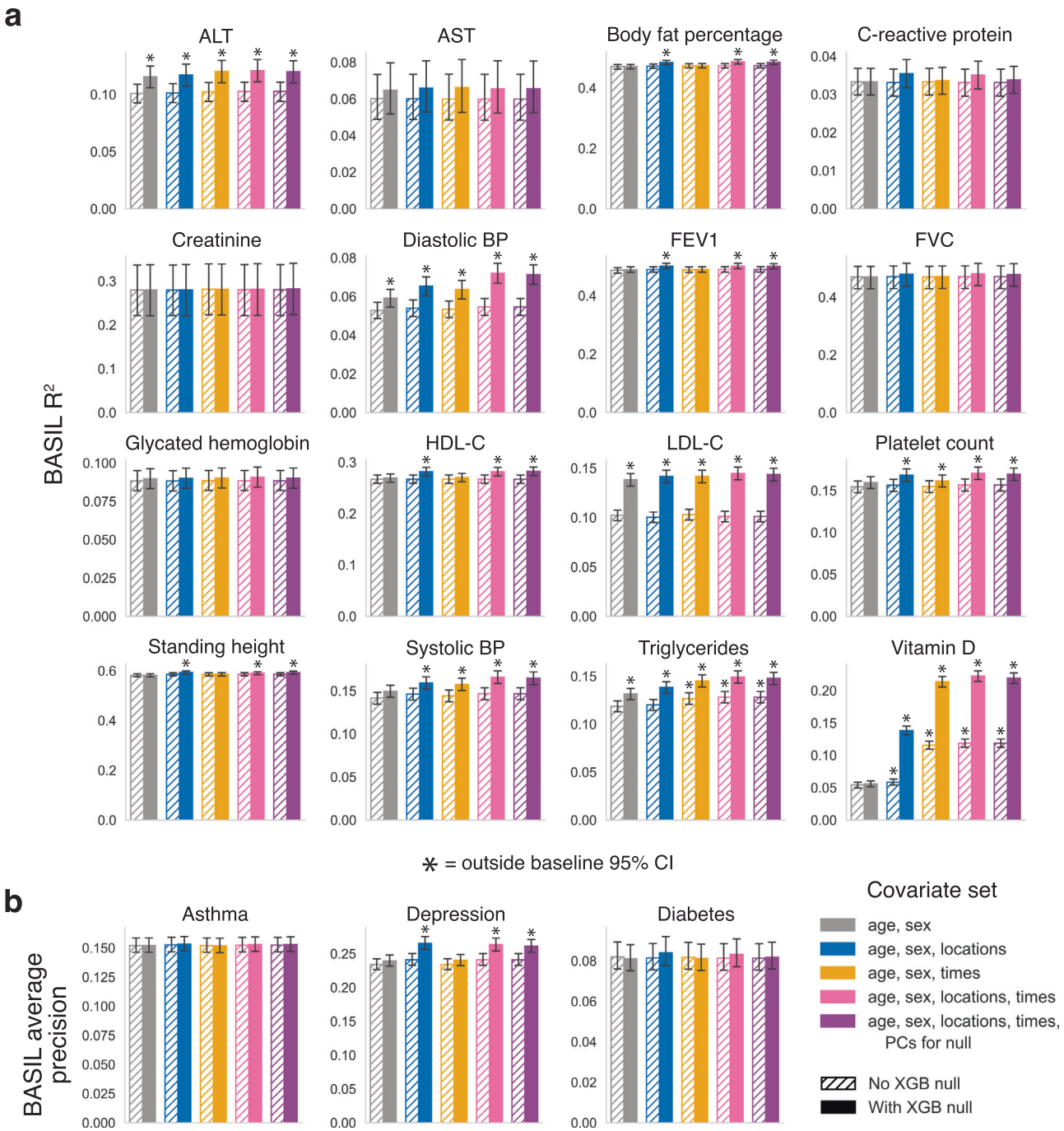


Fig. 3 | BASIL PGS model performance across model and covariate set options. **a** Performance for continuous phenotypes (data provided in Supplementary Data 6). Bars show test set performance of BASIL PGS models, measured in R^2, for the 16 continuous phenotypes when the different covariate options were evaluated with and without a nonlinear null. Error bars show 95% bootstrapped confidence intervals ($n = 100{,}000$) around the test set R^2. For each covariate set, the hatched bar shows the R^2 when no nonlinear null was used and filled bars show performance when the nonlinear XGBoost null (XGB) was included. Asterisks above bars indicate that their value is greater than the 95% confidence interval (CI) of the baseline age, sex, and no null model approach. The pink and purple hashed bars are the same, as the only difference between the "age, sex, locations, times" and "age, sex, locations, times, PCs for null" covariate sets is whether the PCs are included as input to the null in addition to the other covariates. **b** Performance for case-control phenotypes (data provided in Supplementary Data 6). Bars show the test set performance of the BASIL PGS models for the three case-control traits. Average precision was used to evaluate performance. As for **a**, the error bars show 95% bootstrapped confidence intervals ($n = 100{,}000$) around the test set R^2 and the hashed pink and purple bars are the same. Train, validation, and test set sizes for all phenotypes are provided in Supplementary Data 1.

We also evaluated how the nonlinear null models and additional spatiotemporal covariates impacted the performance of BASIL for three case-control traits from UK Biobank: asthma, depression, and diabetes (Fig. 3b). For these phenotypes, we used 317,511 labeled samples with case control ratios of 1:6.05, 1:4.59, and 1:18.2, respectively (Supplementary Data 1). For all three phenotypes, the best results occurred when a null model was used with the addition of location features, though the degree of impact varied. Depression was the only phenotype for which there was improvement in average precision and area under the receiver operating curve (AUC) outside of

the 95% confidence interval of the standard approach. The average precision improved 13.5% over the standard approach and 10.1% compared to including the location features but only modeling them linearly. The XGBoost spatiotemporal null outperformed the standard BASIL PGS for depression (Supplementary Fig. 4b). The improvements in AUC were 6.1% and 5.4%, respectively. Using a null model with locations, locations and times, or locations, times, and PCs for depression were the only instances for case-control phenotypes where an alternative model performed significantly differently from the baseline (Bonferroni corrected Fisher-Pitman p-value < 0.01; Supplementary Data 8). For asthma and diabetes, the improvements in average precision and AUC were minor and never outside of the 95% confidence interval of the standard approach. Both the standard BASIL PGS and the version incorporating the largest XGBoost null with additional spatiotemporal covariates were outperformed by that XGBoost spatiotemporal null alone (Supplementary Fig. 4b).

Ultimately, for continuous traits, we found that adding a nonlinear null and additional spatiotemporal features to PGS models were both advantageous, but using both together performed best. Using both resulted in performance outside the 95% confidence interval of the baseline for more than twice as many phenotypes compared to applying either of those strategies individually. The improvements in terms of R^2 achieved by null models with spatiotemporal features over baseline models is highly correlated with the improvement in R^2 for BASIL PGSs when the nonlinear null and spatiotemporal features are added (Pearson correlation = 0.9985, p-value 5.4e-19; Supplementary Fig. 5). For the three case-control phenotypes we did not see a clear trend of improvement, but the one example for which a model exceeded the 95% confidence interval of the baseline was when location features and a nonlinear null were used in conjunction.

Identifying covariate interactions in null models

Unlike linear models, nonlinear models can be used to understand interactions between their input features in addition to the inputs' separate additive effects. Given the notable improvements in null model and PGS performance using models that account for nonlinear effects and covariate interactions, we next sought to identify specific interactions within our extended set of covariates that are important for explaining phenotypic variation. Here, we focused on XGBoost, which we used above in PGS models. These models are a good fit for understanding both effects of individual covariates and interactions, since interpretability methods exist that can efficiently compute effects of individual factors, first-order, and higher-order interactions by taking advantage of the underlying tree-based structure of the model[7,32]

We used TreeSHAP-IQ[32] to compute first-order Shapley Interaction Index (1-SII) values[35] to identify covariate interactions. One complication when using Shapley value-based interpretability methods is that they provide local (i.e., for one individual) measures of impact or importance when we are mostly interested in understanding the global importance, which reflects the aggregate importance across the population. To identify covariate interactions that were globally important, we looked at the mean magnitude of the local first-order SII values for the held out test set (Supplementary Data 10). We then visualized these global relationships using network plots similar to those used by TreeSHAP-IQ for visualizing local feature attribution in addition to examining 1-SII values for individual features and pairs of features (Fig. 4 and Supplementary Fig. 6).

We identified several influential categories of interactions contributing to the traits studied. First, we found multiple traits (ALT, AST, LDL-C, triglycerides, vitamin D) where the age-sex interaction had a greater mean 1-SII magnitude than age and/or sex alone (Fig. 4a, b; Supplementary Fig. 6a, b, n, r, s). For some phenotypes (ALT, AST, LDL-C, and triglycerides), the trajectories with age are largely in opposite directions for males vs. females. These interactions are similar to those shown by studies in other cohorts for LDL-C and triglyceride levels[21], ALT[36], and systolic blood pressure[37]. LDL-C (Fig. 4a and Supplementary Fig. 6n) most clearly demonstrates the importance of modeling age and sex covariates in a way that accounts for age-by-sex interactions. If sex alone were considered, females would be deemed to have higher LDL-C levels. On the other hand, when considering age, it is clear that females' levels are only higher on average in and after the second age quintile (ages 49–55), likely as a result of menopause[21]. If age alone were considered, the different trajectories for males and females with age would be missed. The importance of this interaction is reflected in the mean 1-SII magnitudes for LDL-C, where the value for the age-sex interaction was greater than age or sex alone (Supplementary Fig. 6n). Aspartate Aminotransferase (AST) provides a similar example: considering age alone would miss the opposite trend observed for males compared to females (Fig. 4b). The mean 1-SII magnitude for the age-sex interaction is less than that of sex but greater than age (Supplementary Fig. 6b).

Second, we identified a strong interaction between sex and time of year for vitamin D levels (Fig. 4c). The sex-day of year interaction had the sixth-highest mean 1-SII magnitude (Supplementary Fig. 6s). Vitamin D levels show a sinusoidal pattern, with peak values in August and lowest levels in February. The means for males and females are similar (less than a tenth of a nanomole per liter difference), but the amplitude of the sinusoidal pattern for males is larger. This is supported by other studies that report greater vitamin D variability over the course of the year for men than women, possibly explained by lower levels of adipose tissue[15]. This time-dependent variability may also explain why other studies looking at sex differences in vitamin D levels without accounting for time of year are inconsistent[15].

Finally, we identified multiple strong interactions involving location. For some traits, we observed interactions between home and birth location coordinates. This indicated there may be some relationship between where someone lives compared to where they were born and the phenotype. For example, for C-reactive protein (CRP) the birth and home east coordinate interaction ranked fifth highest in mean 1-SII magnitude and the birth and home north coordinate interaction ranked seventh (Supplementary Fig. 6e). Looking into this further, we found the birth-to-home distance is negatively correlated with CRP levels (Spearman r = -0.077; $P < 10^{-100}$). This trend is seen for both males and females when visualizing CRP distributions by birth-to-home location quintile (0.7–3.8 km, 3.8–8.7 km, 8.7–29.5 km, 29.5–147.5 km, and 147.5–1035.0 km from their birth location; Fig. 4d).

In another example, for current or prior depression, the self-reported rate increased with increased birth-to-home location distance (Fig. 4e). This result matched the findings of a 2023 study on internal migration in the United States, which found that the rate of depression among internal migrants was 1.259 times higher than for nonmigrants[38]. We found significant positive correlations between depression and birth-to-home location distance (point-biserial r = 0.076, rank-biserial r = 0.15, both p-values < 10^{-100}). The birth and home north coordinate interaction had the third highest mean 1-SII magnitude following only age and sex on their own (Supplementary Fig. 6g). The birth and home east interaction was fifth highest, one behind the home north coordinate. Importantly, our results do not distinguish if the cause is due to differences in attributes between those who do or do not migrate or properties of the locations involved. Still, our results help explain why incorporating interactions between features can improve PGS prediction accuracy.

Discussion

In this study, we demonstrate that improved modeling of the non-genetic component of traits can substantially enhance the performance of PGSs. Our results highlight that the combination of spatiotemporal features and nonlinear null models leads to more significant improvements in PGS accuracy than either modification on its

https://doi.org/10.1038/s41467-026-76154-7

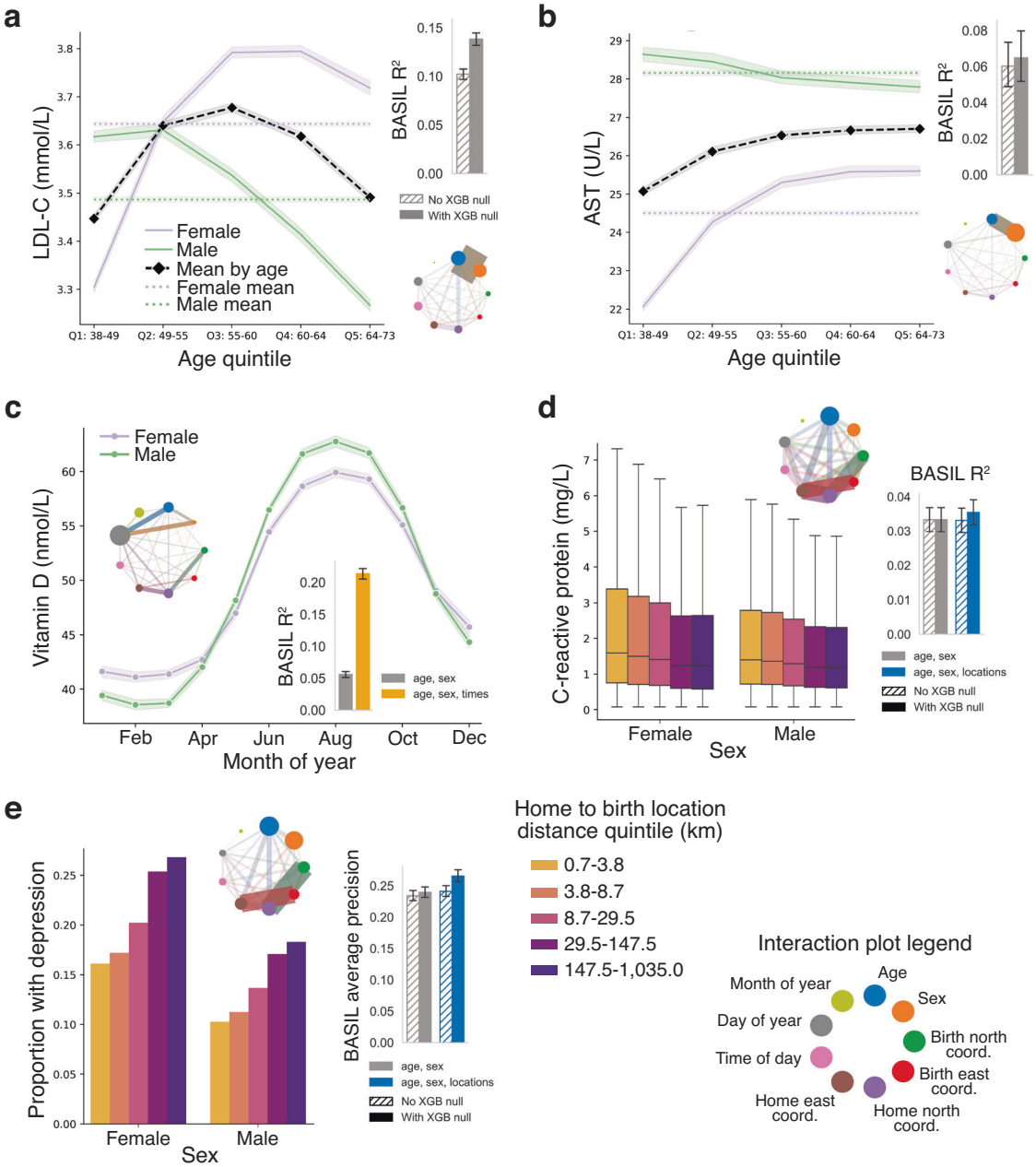


Fig. 4 | Covariate interactions identified by null model interpretation. a Age-sex interactions for LDL cholesterol (LDL-C). The main plot shows mean LDL-C levels by sex across age quintiles, plus overall sex and age group means in all participants. Shading indicates 99% confidence intervals (bootstrap $n = 1000$). The bar plot (top right) shows BASIL PGS model improvement on the test set when a nonlinear null enables modeling these interactions. The bottom right shows a covariate interaction network plot defined below. **b** Age-sex interactions for aspartate aminotransferase (AST). Panels are the same as **a**. **c** Sex-time of year interaction for vitamin D. The main plot shows the average vitamin D levels for males and females over the year, binned by month rather than day of year, which is collinear but identified by TreeSHAP-IQ as having greater impact (upper left interaction plot). Shading shows 99% confidence intervals (bootstrap $n = 1000$). The bar plot shows PGS improvement on the test set when time covariates are included with a nonlinear null, which can capture this cyclical behavior. **d** Birth-to-home distance interaction for C-reactive protein (CRP). The box plot shows CRP levels across five birth-to-home distance quintiles for males and females. The center line indicates the median, the box limits represent the 25th and 75th percentiles (Q1 and Q3), and the whiskers extend to the farthest datapoint within Q1-1.5*IQR and Q3 + 1.5*IQR where IQR gives the interquartile range (Q3-Q1). The legend below **d** shows quintile colors used in (**d**) and (**e**). The interaction network plot is shown above. The bar plot showing improvement when location features are modeled with and without a nonlinear null is to the right. **e** Birth-to-home distance interaction for depression. The bar plot shows the proportion per sex and distance quintile reporting past or current depression. The interaction network plot is above. The BASIL polygenic score performance bar plot shows the largest improvement when location features are included and modeled nonlinearly. For all panels, the SHAP interaction plots and BASIL performance bar plots use the held out test set, while phenotype trend plots use all available data. For all performance bar plots, error bars indicate 95% bootstrapped confidence intervals ($n = 100{,}000$) and XGB stands for XGBoost. Hatched and solid bars denote model performance on the test set without and with nulls, respectively (data in Supplementary Data 6). In interaction plots, circle sizes indicate mean SHAP value magnitudes for covariates (data in Supplementary Data 11), edge widths indicate the mean 1-SII value magnitudes for pairwise interactions (data in Supplementary Data 10; Supplementary Fig. 6), and sample sizes are the test set sizes in Supplementary Data 1. All overall, train, validation, and test set sizes can be found in Supplementary Data 1.

own. We found this to be particularly true for the continuous phenotypes tested here, where the combination of these two improvements resulted in the best performance for all phenotypes and R^2 outside of the 95% confidence interval of the baseline for 11/16 traits with BASIL and 12/16 with PRS-CS. For the three case-control phenotypes we evaluated we saw modest differences in performance, except for depression where the combination of a nonlinear null model and location features resulted in performance outside the baseline 95% CI.

The utility of specific spatiotemporal covariates was phenotype dependent. For phenotypes known to have diurnal or seasonal rhythms, like vitamin D or blood pressure, the time-based covariates were highly informative. Conversely, for standing height, which is highly heritable and for which environmental exposures likely have an effect before adulthood[39], as expected, adding factors, such as time of assessment, has no impact on performance. Similarly, for the case-control phenotypes, the time of assessment was not relevant to whether a person would have previously developed any of the conditions, and as expected, they had either no effect or a detrimental effect on performance. In other case-control applications like longitudinal studies, time features could be more relevant for understanding case-control traits. In practice, researchers should consider the underlying dynamics of a trait when selecting covariates or compare the performance of null models, which are relatively cheap and easy to train, to evaluate different covariate sets. Nonetheless, for the majority of continuous traits we tested, a model incorporating the full set of spatiotemporal covariates performed best.

A strength of the null model approach is that, beyond improving prediction performance, it provides insight into how nongenetic factors and their interplay impact different traits. By modeling nonlinear and interaction effects among covariates, these models reveal dependencies that linear adjustments fail to capture. In this study, we detected multiple covariate interactions, including examples where phenotype means exhibit opposite age-related trends between sexes (ALT, AST, LDL-C, and triglycerides), cyclical temporal patterns for vitamin D reflecting seasonal variation, and complex spatial dependencies between birth and home locations. These location differences can capture effects of distance or direction moved, as well as non-additive interactions between the effects of where one grew up and now lives. Together, these findings illustrate how nonlinear null models can uncover interpretable relationships that are obscured under standard linear frameworks.

Our study has several limitations. First, the training and PGS evaluation were conducted exclusively on individuals of White British ancestry in the UK Biobank. While our evaluation of null model performance on non-White British participants suggests that the nonlinear modeling advantage is maintained across ethnic groups, we observed substantial performance degradation for some groups, particularly Black or Black British participants. The specific environmental patterns learned by our models are likely population- and location-specific, and retraining on ancestry-matched data would likely improve performance. Our analysis of non-White British groups was also limited by small sample sizes, which resulted in wide confidence intervals and should be interpreted cautiously. We focused our non-White British evaluation on the null models alone, as this highlighted the differences resulting from the null modeling as opposed to differences in modeling genetic factors. Second, spatiotemporal variables are, in most cases, proxies for true environmental exposures. Our models and methods for interpreting them cannot determine the specific causal factor. For example, if location is determined to be important, we still would not know if that were due to the weather, pollution, or lifestyle differences in an area. Third, while our method improved predictions for continuous traits, the benefit for case-control traits was modest. This could be due to a smaller environmental contribution for the selected phenotypes, lower statistical power, or challenges in phenotype definition (e.g., self-reports of past instances of depression).

In our study, we focused on integrating null models into commonly used classes of PGSs based on linear models. In this approach, the null models are trained separately on covariates, and the resulting output is used as an additional linear term in association testing or PGS construction. As discussed in the introduction, an alternative approach is to perform feature selection on the genotypes and then use a nonlinear model to predict the phenotype based on a model that jointly considers the subset of genotypes and covariates[12–14]. In this setup, the feature selection is a key step that could be similarly improved with the integration of a null model of covariates, which is an important topic of future work.

Overall, we demonstrate that the incorporation of nonlinear and interaction effects of spatiotemporal covariates into association testing models can, in some cases, result in large performance gains with only a small additional computational overhead beyond standard approaches. Importantly, our framework, which builds on DeepNull[5], outputs a single term summarizing these covariate contributions that can be easily added as a covariate to any downstream methods for complex trait analysis, such as genome-wide association studies, PGS, and fine-mapping. Further, location and time-related covariates are already collected by most large biobanks, and the gradient boosted decision tree null models are relatively cheap and easy to train and interpret. This makes our approach easy to implement for future studies. Ultimately, this work demonstrates the importance of modeling the environmental component of complex traits and provides a practical framework for integrating these features into large-scale genetic studies to improve prediction and discovery.

Methods

Ethics

This research complied with all relevant ethical regulations and was conducted in accordance with the principles of the Declaration of Helsinki. Ethics approval for this study is covered by the 2021 North West Multi-center Research Ethics Committee approval renewal (REC reference: 21/NW/0157), which covers UK Biobank[23] researchers without the need for further approval. All participants provided written informed consent prior to participation and were informed of their right to withdraw at any time. Access to the UK Biobank was provided through application number 46122.

Participant data

All analyses were performed using data from the UK Biobank (UKB)[23] on the UKB Research Analyses Platform. Our sample and genotype quality control procedure was primarily based on published pipelines[40,41]. For all parts of this study excluding the null model evaluation on non-White British participants, we excluded participants who were not identified as White British ancestry by field 22006 (which combines self-reported ethnicity with genetic principal components analysis), did not meet the quality control (QC) criteria for inclusion in the UKB principal components computation (which included subsetting to individuals unrelated to the 3rd degree, field 22020), were labeled as outliers for heterozygosity or missing rate (field 22027), had sex chromosome aneuploidy (field 22019), or exhibited sex discordance (fields 31 and 22001). This resulted in 337,081 participants. These individuals were used for the genotype QC, during which additional participants were also removed based on their genotype missingness rate (see below). To avoid having missing covariate values, we further subsetted to the 317,511 participants for whom home and birth location data were available for all analyses. These remaining participants were randomly split into 8:1:1 training, validation, and test sets. The final sample counts per phenotype are provided in Supplementary Data 1.

For the null model evaluation of participants with non-White British ancestry, participants were assigned to an ethnic group based on their self-reported ethnic background (field 21000). The Black or

Black British group contains participants who selected "Black or Black British" as their ethnic group, Asian or Asian British contains participants who selected "Asian or Asian British" or "Chinese" as their ethnic group, White and Black contains participants who selected "Mixed" for their ethnic group and then "White and Black Caribbean" or "White and Black African" for their ethnic background, and White and Asian contains participants who selected "Mixed" for their ethnic group and then "White and Asian" for their background. To be included in the non-White British groups for evaluation, participants must not have been part of the White British sample defined by field 22006 and must have had values for all covariates used as null model inputs.

Genotype data

We used the version 3 UKB imputed genotypes. The following genotype QC steps, based on those used in previous pipelines for UKB analysis[40,41] were performed with plink2 v2.00a2.3LM[42,43] using the sample set described above: We began by removing variants with a minor allele frequency less than 0.01, an INFO score of less than 0.3, or that were multiallelic. A coarse missingness filter was then applied, removing participants and then variants with a missingness rate greater than 20%. Finally, variants with a missingness rate greater than 2% were removed, followed by participants with a missingness rate greater than 2%[40]. Hardy-Weinberg equilibrium (HWE) filtering was applied upstream to genotyping batches by UK Biobank as part of their QC pipeline and HWE is an assumption of the imputation method (IMPUTE4[44]), so we did not perform additional HWE filtering[23,41].

Covariate data

Covariate information was derived from UKB fields described below. Following the sample QC, genotype QC, and removal of samples with missing location data, there were no missing values. The age covariate was obtained from the age at recruitment (field 21022), which reflects a participant's age at their initial assessment center visit. Sex was coded as a binary variable (field 31). We used the first 20 genetic principal components, as provided by UKB in field 22009.

For location covariates, we considered participants' home and birth locations. These features are encoded as easting-northing coordinate pairs using the Ordnance Survey National Grid (OSGB) reference system (fields 130 and 129 for birth, and 22702 and 22704 for home).

For time covariates, we included both the time of day and the time of year based on the time the blood sample was collected during the initial assessment center visit (field 3166). Time of day was represented continuously in terms of minutes past midnight. The time of year was represented with integers for both the month and the day of the year.

Phenotype data

Sixteen continuous phenotypes were used, each derived from the initial assessment visit for a single UK Biobank field: alanine aminotransferase (field 30620), aspartate aminotransferase (field 30650), body fat percentage (field 23099), C-reactive protein (field 30710), creatinine (field 30700), diastolic blood pressure (field 4079), forced expiratory volume in 1-second (FEV1, field 3063), forced vital capacity (FVC, field 3062), glycated hemoglobin (field 30750), HDL cholesterol (field 30760), LDL direct (LDL-C, field 30780), platelet count (field 30080), standing height (field 50), systolic blood pressure (field 4080), triglycerides (field 30870), and vitamin D (field 30890).

Three case-control phenotypes were used. Asthma cases were defined using the version 2.0 algorithmically defined outcomes (ADOs) made available by UKB. Participants were classified as cases if they had a value for "source of asthma report" (field 42015), which included self-reported only, hospital primary, death primary, hospital secondary, or death contributory. Depression cases were defined based on responses to the online mental well-being questionnaire. Participants were classified as cases if they had a value for "duration of worst depression" (field 20438). Values in this field indicated participants had answered "yes" to either "Have you ever had a time in your life when you felt sad, blue, or depressed for two weeks or more in a row?" (field 20446) or "Have you ever had a time in your life lasting two weeks or more when you lost interest in most things like hobbies, work, or activities that usually give you pleasure?" (field 20441). Diabetes cases were identified using self-reported diagnosis of diabetes by a doctor (field 2443, "yes" for any instance).

Null models

The null models, which predict a phenotype based solely on covariates, were trained using the training set of participants with 5-fold cross-validation, as in DeepNull[5]. Predictions for training set participants were from the fold in which they were held out. Predictions for members of the validation and test sets were computed as the average of the predicted values across the five folds. When evaluating null model performance on its own, we used only the held-out predictions for the training set to avoid biasing results when integrating the null model into PGSs (below).

We evaluated three classes of null models: linear, XGBoost, and neural network. The linear models were implemented using scikit-learn[45]. The linear regression used the default parameters. The LASSO and ridge regression models used the default parameters except for α, which was set to 0.5 and 0.1, respectively.

XGBoost[28] is an implementation of gradient-boosted decision trees (GBDTs). We used the Python API default parameters except where otherwise noted. For all model variations, 10% of the participants in the folds used for training were held out as a validation set for early stopping. Root mean square error (RMSE) was used as the early stopping evaluation metric. The smallest of the three variations we evaluated used up to 100 boosting rounds (i.e., estimators) and stopped early if there was no improvement in 25 rounds. The two larger versions used up to 1000 or 2500 boosting rounds and waited 50 rounds before early stopping. To further avoid overfitting, the two larger versions used a maximum tree depth of 4 and reduced the learning rate from 0.3 to 0.25.

The neural network models were implemented using PyTorch[46]. One variation was a reimplementation of the original DeepNull[5] network architecture and training procedure. The architecture has two components: a linear residual connection and a multi-layer perceptron (MLP) with four hidden layers of sizes 64, 64, 32, and 16 with Rectified Linear Unit (ReLU) activations between them. Both components take the input variables as input, and their outputs are summed to get the output. The Adam optimizer[47] was used with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. Training was run for 1000 epochs with a batch size of 1024 and a learning rate of 10^{-4}. Mean squared error (MSE) loss was used for regressions and binary cross-entropy for classification.

The other two variations modified the training procedure to help prevent overfitting. Both used early stopping, in which 10% of the participants in the folds used for training were held out as a validation set. These variations were trained for up to 5000 epochs but stopped early and reverted to their best-performing weights if the loss did not improve for 75 epochs. The AdamW optimization algorithm[48] was used with the default PyTorch parameters due to its weight decay regularization. Training was run with a batch size of 1024 and an initial learning rate of 10^{-4}. The learning rate was reduced by a factor of 0.1 if training plateaued (the loss did not improve by more than 10^{-4} for more than 25 epochs). These versions are indicated as ESWP in Fig. 2, referring to early stopping, AdamW, and plateau. One variation (ESWP, Small) used the same architecture as the original DeepNull model, while the other instead used three hidden layers of sizes 64, 32, and 16.

Cyclic basis functions to represent time features

To evaluate null model performance using cyclic basis functions for temporal features (time of day, month of year, day of year), each was

additionally encoded with two basis functions: sin(θ) and cos(θ) where θ is 2π(feature value / feature range). Feature range was set to 1440 minutes for time of day, 12 for month of year, and 365.25 for day of year. Full results using these representations are in Supplementary Data 3.

Polygenic score models

The PGS models, which used either BASIL (no version information available)[24] or PRS-CS v1.1.0[25], predicted a phenotype based on genotypes, genetic principal components, nongenetic covariates, and, in some cases, the prediction of a nonlinear null model. Evaluation was performed using the test set. Before these models were fit, the location coordinates, time of day, day of year, and null model predictions were standardized to have a mean of zero and a standard deviation of one. For PRS-CS, the month of the year was treated as a categorical variable, while for BASIL, it was left as an integer. This was done as PRS-CS can accept categorical input variables.

BASIL models were trained on the training set of participants, while the validation set was used for early stopping and selecting the λ hyperparameter used in BASIL. Training was performed for up to 25 iterations, with early stopping if the validation metric did not improve for more than two iterations. An elastic-net mixing parameter of $\alpha = 1.0$ was used, corresponding to the LASSO penalty. The default parameters of the authors' implementation were otherwise used.

PRS-CS is a summary statistics-based method, so we began by running an association test with plink2 v2.00a6LM[42,43] using the generalized linear model (--glm flag) on the training set. The only departure from the default parameters was increasing the maximum variance inflation factor to 10^6. This change allowed us to run the association test despite the correlation between covariates, especially the null model prediction and the other covariates it is based on, being high. DeepNull[5] recommends incorporating the underlying covariates into a PGS in addition to the null prediction, as the null does not account for associations between the covariates and genotypes. We evaluated the effect of removing covariates that are highly correlated with the null prediction on a BASIL PGS, which includes multiple rounds of plink2 association tests. The highest magnitude correlation between the null prediction and age was for glycated hemoglobin ($r = 0.96$) and for sex creatinine ($r = 0.98$). When we removed age but included all other covariates and the null prediction for glycated hemoglobin, R^2 increased just 0.2% over using all covariates and the null, which was not significant (two-sided Fisher-Pitman test[33,34] p-value = 0.29). Removing sex as a covariate for creatinine resulted in a small but significant decrease in PGS R^2 (0.282 to 0.280, p-value = 2e-5). The covariate with the next largest magnitude correlation with a null prediction was day of year for vitamin D ($r = 0.56$).

For the PRS-CS PGS, the genotypes and covariates, optionally including the null model prediction, were used as inputs to the association test, whereas the following PRS-CS step used only the summary statistics for the genotypes. For PRS-CS, the 1000 Genomes European LD reference panel provided on the PRS-CS website was used. This was selected over the UKB LD reference panel since it was orthogonal to the samples in our analysis. Participants were then scored using the posterior effect sizes from PRS-CS. To reincorporate the covariates and rescale the score to the actual phenotype, a linear wrapper model was then fit using the validation set. This linear model used either a scikit-learn[45] linear regression for continuous traits or a logistic regression for case-control traits with the default parameters. The input variables were the score from summing the PRS-CS effects, the covariates, and the optional null model prediction. The covariates were scaled as described above, except that months were instead one-hot encoded.

The association test run as part of the PRS-CS PGS was used to evaluate the impact of the null on GWAS summary statistics. Supplementary Data 7 provides the counts of significant (association p-value $< 5 \times 10^{-8}$) variants and independent loci using either the standard approach or with the XGBoost nonlinear null model using all the additional spatiotemporal features, as well as the number that were common or unique to each approach. Independent loci were identified as in Karczewski et al. (2025)[49], using the plink2[42,43] --clump command with a clumping radius of 500 kb, a clump r^2 of 0.1, and an LD panel consisting of a subset of 5000 training set participants. If clumps from the two sets of summary statistics (i.e., standard and with null) had lead variants within 500 kb of each other, the clumps were greedily matched to the nearest clump from the other set and the paired clumps were considered to be a single locus.

Null model interpretation

We used TreeSHAP-IQ[32] to interpret the trained XGBoost models to identify covariate effects and interactions. This was done using the test set of participants and all five of the XGBoost models (one per training fold) that make up one full null model. For each participant and model, this resulted in SHAP values for each covariate (Supplementary Data 11) and a matrix of first-order Shapley Interaction Index (1-SII) values. This matrix provides the 1-SII value for each covariate on its own as well as every pairwise interaction. Both the SHAP and 1-SII values are local, meaning they reflect the impact of a covariate for one participant. To make these values global (i.e., reflect their impact across the entire cohort), we used the average magnitude of the 1-SII values across all models and test set participants when identifying important interactions. Mean 1-SII values for all traits are provided in Supplementary Fig. 6 and Supplementary Data 10.

Reporting summary

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

Data availability

We used genetic, covariate, and phenotype datasets provided by UK Biobank under application number 46122 (see https://www.ukbiobank.ac.uk/use-our-data/). Individual level data for UK Biobank participants are available through the UK Biobank Research Analysis Platform. Source data are provided with this paper in the Supplementary Data files.

Code availability

Code and workflows used for this study are available at https://github.com/gymrek-lab/DeeperNull[50], along with documentation on how to train null models, compute Shapley and 1-SII values, and the functions of the UK Biobank workflows. The code is available under the GNU General Public License v3.0.

References

1. 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).
2. Dor, E. et al. *Selecting Covariates Genome-Wide Assoc. Stud.* https://doi.org/10.1101/2023.02.07.527425 (2023).
3. Uffelmann, E. et al. Genome-wide association studies. *Nat. Rev. Methods Prim.* **1**, 59 (2021).
4. Sinnott-Armstrong, N. et al. Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat. Genet.* **53**, 185–194 (2021).
5. McCaw, Z. R. et al. DeepNull models non-linear covariate effects to improve phenotypic prediction and association power. *Nat. Commun.* **13**, 241 (2022).
6. Aschard, H. et al. Covariate selection for association screening in multiphenotype genetic studies. *Nat. Genet.* **49**, 1789–1795 (2017).
7. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).

8. Wang, Y., Zhang, T., Guo, X. & Shen, Z. Gradient-based feature attribution in explainable AI: a technical review. Preprint at https://doi.org/10.48550/arXiv.2403.10415 (2024).
9. Zhang, J. et al. Revisiting the trustworthiness of saliency methods in radiology AI. *Radiol. Artif. Intell.* **6**, e220221 (2024).
10. Ghorbani, A., Abid, A. & Zou, J. Interpretation of neural networks is fragile. *Proc. AAAI Conf. Artif. Intell.* **33**, 3681–3688 (2019).
11. Majdandzic, A., Rajesh, C. & Koo, P. K. Correcting gradient-based interpretations of deep neural networks for genomics. *Genome Biol.* **24**, 109 (2023).
12. Medvedev, A., Mishra Sharma, S., Tsatsorin, E., Nabieva, E. & Yarotsky, D. Human genotype-to-phenotype predictions: boosting accuracy with nonlinear models. *PLoS ONE* **17**, e0273293 (2022).
13. Sigurdsson, A. I. et al. Complex genetic effects linked to plasma protein abundance in the UK Biobank. *Nat. Commun.* **17**, 533 (2025).
14. Meléndez, A. et al. Assessing Tree-Based Phenotype Prediction on the UK Biobank. In *Proc. IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 3804–3810 https://doi.org/10.1109/BIBM58861.2023.10385960.(IEEE, 2023).
15. Wierzbicka, A. & Oczkowicz, M. Sex differences in vitamin D metabolism, serum levels and action. *Br. J. Nutr.* **128**, 2115–2130 (2022).
16. Nystoriak, M. A. & Bhatnagar, A. Cardiovascular effects and benefits of exercise. *Front. Cardiovasc. Med.* **5**, 135 (2018).
17. Davidson, T., Boardman, J. D. & Hunter, L. M. Exploring rural-urban differences in polygenic associations for health among older adults in the United States. *J. Rural Soc. Sci.* **37**, 4 (2022).
18. Tartour, K. & Padmanabhan, K. The clock takes shape-24 h dynamics in genome topology. *Front. Cell Dev. Biol.* **9**, 799971 (2021).
19. Wucher, V., Sodaei, R., Amador, R., Irimia, M. & Guigó, R. Day-night and seasonal variation of human gene expression across tissues. *BioRxiv Prepr. Serv. Biol*. 2021.02.28.433266 https://doi.org/10.1101/2021.02.28.433266 (2022).
20. Dopico, X. C. et al. Widespread seasonal gene expression reveals annual differences in human immunity and physiology. *Nat. Commun.* **6**, 7000 (2015).
21. Holven, K. B. & Roeters Van Lennep, J. Sex differences in lipids: a life course approach. *Atherosclerosis* **384**, 117270 (2023).
22. McCartney, D. L. et al. An epigenome-wide association study of sex-specific chronological aging. *Genome Med* **12**, 1 (2020).
23. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
24. Qian, J. et al. A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLOS Genet.* **16**, e1009141 (2020).
25. 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).
26. Shwartz-Ziv, R. & Armon, A. Tabular data: deep learning is not all you need. *Inf. Fusion* **81**, 84–90 (2022).
27. Grinsztajn, L., Oyallon, E. & Varoquaux, G. Why do tree-based models still outperform deep learning on tabular data? *Advances in neural information processing systems* **35**, 507–520 (2022).
28. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016). https://doi.org/10.1145/2939672.2939785.
29. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **7**, 1–30 (2006).
30. McElfresh, D. et al. When do neural nets outperform boosted trees on tabular data? *Adv. Neural Inf. Process. Syst.* **36**, 76336–76369 (2023).
31. Yang, B., Gül, M. & Chen, Y. Comparative analysis of deep learning and tree-based models in power demand prediction: accuracy, interpretability, and computational efficiency. *J. Build. Phys.* **49**, 127–169 (2025).
32. Muschalik, M., Fumagalli, F., Hammer, B. & Hüllermeier, E. Beyond treeSHAP: efficient computation of any-order shapley interactions for tree ensembles. *Proc. AAAI Conf. Artif. Intell.* **38**, 14388–14396 (2024).
33. Berry, K. J., Mielke, P. W. & Mielke, H. W. The Fisher-Pitman permutation test: an attractive alternative to the *F* test. *Psychol. Rep.* **90**, 495–502 (2002).
34. Bestgen, Y. Please, don't forget the difference and the confidence interval when seeking for the state-of-the-art status. *Proceedings of the Language Resources and Evaluation Conference* 5956–5962 https://doi.org/10.63317/2gavpzzwqkcx (2022).
35. Grabisch, M. & Roubens, M. An axiomatic approach to the concept of interaction among players in cooperative games. *Int. J. Game Theory* **28**, 547–565 (1999).
36. Petroff, D. et al. Age dependence of liver enzymes: an analysis of over 1,300,000 consecutive blood samples. *Clin. Gastroenterol. Hepatol.* **20**, 641–650 (2022).
37. Wong, D., Gerry, S., Shamout, F., Clifton, D. A. & Pimentel, M. A. F. Cross-sectional centiles of blood pressure by age and sex: a four-hospital database retrospective observational analysis. *BMJ Open* **10**, e033618 (2020).
38. Liao, Z.-X., Tan, X.-M., Zhao, Y.-Y., Sun, X.-C. & Yi, F.-L. Association between internal migration experience and depressive symptoms: analysis of PSID data. *BMC Public Health* **23**, 1145 (2023).
39. Jelenkovic, A. et al. Genetic and environmental influences on human height from infancy through adulthood at different levels of parental education. *Sci. Rep.* **10**, 7974 (2020).
40. Marees, A. T. et al. A tutorial on conducting genome-wide association studies: quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.* **27**, e1608 (2018).
41. Collister, J. A., Liu, X. & Clifton, L. Calculating polygenic risk scores (PRS) in UK Biobank: a practical guide for epidemiologists. *Front. Genet.* **13**, 818574 (2022).
42. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, s13742-015-0047–8 (2015).
43. Purcell, S. & Chang, C. PLINK 2.0. www.cog-genomics.org/plink/2.0/ (2024).
44. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
45. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
46. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. *Proceedings of the 33rd International Conference on Neural Information Processing Systems* 8026–8037 (2019).
47. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *Proc. 3rd International Conference on Learning Representations* (ICLR, 2015).
48. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *Proc. 7th International Conference on Learning Representations* (ICLR, 2019).
49. Karczewski, K. J. et al. Pan-UK Biobank genome-wide association analyses enhance discovery and resolution of ancestry-enriched effects. *Nat. Genet.* **57**, 2408–2417 (2025).
50. Ross DeVito. gymrek-lab/DeeperNull: v1.0.0. https://doi.org/10.5281/ZENODO.20131325 (2026).

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 46122. This work uses data provided by patients and collected by the NHS as part of their care and support. We gratefully acknowledge UK Biobank participants for their contributions, without whom this research would not have been possible. We thank Tara Mirmira, Michael Lamkin, and Arya Massarat for helpful comments on the manuscript.

Author contributions

R.D. and M.G. conceptualized the project. R.D. performed all analyses and experiments, prepared the figures and wrote the manuscript. M.G. supervised analyses and helped with writing the manuscript. Both authors reviewed and approved the final manuscript.

Funding

M.G. and R.D. disclose support for the research of this work from NIH grant RM1HG011558.

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/s41467-026-76154-7.

Correspondence and requests for materials should be addressed to Melissa Gymrek.

Peer review information *Nature Communications* thanks Zachary McCaw, Md. Moksedul Momin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

© The Author(s) 2026