







DeepNull models non-linear covariate effects to improve phenotypic prediction and association power

Zachary R. McCaw ^{1,3}, Thomas Colthurst^{2,3}, Taedong Yun ², Nicholas A. Furlotte¹, Andrew Carroll ¹, Babak Alipanahi ¹, Cory Y. McLean ^{2,4}✉ & Farhad Hormozdiari ^{2,4}✉

Genome-wide association studies (GWASs) examine the association between genotype and phenotype while adjusting for a set of covariates. Although the covariates may have non-linear or interactive effects, due to the challenge of specifying the model, GWAS often neglect such terms. Here we introduce DeepNull, a method that identifies and adjusts for non-linear and interactive covariate effects using a deep neural network. In analyses of simulated and real data, we demonstrate that DeepNull maintains tight control of the type I error while increasing statistical power by up to 20% in the presence of non-linear and interactive effects. Moreover, in the absence of such effects, DeepNull incurs no loss of power. When applied to 10 phenotypes from the UK Biobank ($n = 370K$), DeepNull discovered more hits (+6%) and loci (+7%), on average, than conventional association analyses, many of which are biologically plausible or have previously been reported. Finally, DeepNull improves upon linear modeling for phenotypic prediction (+23% on average).

¹Google Health, Palo Alto, CA, USA. ²Google Health, Cambridge, MA, USA. ³These authors contributed equally: Zachary R. McCaw, Thomas Colthurst. ⁴These authors jointly supervised this work: Cory Y. McLean, Farhad Hormozdiari. ✉email: cym@google.com; fhormoz@google.com

Genome-wide association studies (GWASs) aim to detect genetic variants or single-nucleotide polymorphisms (SNPs) that are associated with complex traits and diseases. Over the past decade, GWASs have successfully identified thousands of variants associated with various and diverse phenotypes^{1–6}. These associations have expanded our knowledge of biological mechanisms⁷ and improved our ability to predict phenotypic risk⁸.

In most GWAS, the association strength between genotype and phenotype is assessed while adjusting for a set of covariates, such as age, sex, and principal components (PCs) of the genetic relatedness matrix. Covariates are included in GWAS for two main reasons: to increase precision and to reduce confounding. In the linear model setting, adjustment for a covariate will improve precision if the distribution of the phenotype differs across levels of the covariate. For example, when performing GWAS on height, males and females have different means. Adjusting for sex reduces residual variation, and thereby increases power to detect an association between height and the candidate SNPs. Note, however, that omitting sex from the association test is entirely valid. In contrast, omitting a confounder will result in a biased test of association. By definition, a confounder is a common cause of the exposure (i.e. genotype) and the outcome (i.e. phenotype)⁹. In GWAS, a potential confounder is genetic ancestry: two ancestral groups may differ with respect to minor allele frequency (MAF) at common SNPs and, for unrelated reasons, in their phenotypic means. Failure to adjust for ancestry will lead to spurious associations between the phenotype and the SNPs whose MAFs differ across ancestries, inflating the type I error of the association test. To reduce confounding due to population substructure, or the presence of genetically related subgroups within the cohort, multiple genetic PCs are commonly included as covariates during association testing^{10,11}.

The simplest form of covariate adjustment is to include a linear term for the covariate in the association model. If the phenotypic mean changes non-linearly with the covariate, the residual variation may be further reduced by including higher order adjustments, such as quadratic or interaction terms, as in the following recent examples^{12–14}. Shrine et al.¹² included age^2 as a covariate when studying chronic obstructive pulmonary disease; Chen et al.¹³ included squared body mass index (BMI^2) when studying obstructive sleep apnea; and Kosmicki et al.¹⁴ included an age by sex interaction ($\text{age} \times \text{sex}$) when studying COVID-19 disease outcomes. Although these recent works have recognized the potential importance of modeling non-linear covariate effects, no systematic approach has been described for detecting the appropriate non-linear functions to adjust for in GWAS. The difficulty stems from the exponential number of possible interactions that can arise from a finite set of covariates (e.g. $\text{age} \times \text{sex}$, $\text{age}^2 \times \text{sex}$, \dots), and the infinite number of possible transformations of any given continuous covariate (square, logarithm, exponentiation, etc.). Lastly, the optimal number of covariate interactions is not known a priori and requires evaluating different possibilities (Supplementary Table 1).

In this work, we address the issue of model misspecification in GWAS; specifically, misspecification of the relationship between the phenotype and covariates. DeepNull uses a flexible deep neural network (DNN) to learn this potentially complex and non-linear relationship, then adjusts for the network's expectation of the phenotype (based on covariates only) during association testing. Although simpler models (e.g. a second-order interaction model) may suffice in particular cases, the DNN architecture is sufficiently expressive to capture the broad range of phenotype-covariate relationships that researchers might encounter in practice. Moreover, no loss of power is observed when the relationship between the phenotype and covariates is in fact linear.

Using simulated data, we show that DeepNull markedly improves association power and phenotypic prediction in the presence of non-linear covariate effects, and retains equivalent performance in the absence of non-linear effects. We then demonstrate improvements in association power and phenotype prediction across 10 phenotypes from the UK Biobank (UKB)¹⁵, indicating DeepNull's potential for broad utility in biobank-scale GWAS. We provide DeepNull as freely available open-source software (Code Availability) for straightforward integration into existing GWAS association platforms.

Results

DeepNull overview. DeepNull trains a DNN to predict a phenotype of interest from covariates not directly derived from genotypic data (hereafter “non-genetic covariates”). Due to its ability to approximate any continuous mapping^{16,17}, the DNN can capture complex non-linear relationships between the phenotype and covariates. When performing genetic association testing, the DNN's prediction of the phenotype for each individual is included as a single additional covariate within the association model. Adjusting for the DNN's prediction in the association model is equivalent to regressing it out from both phenotype and genotype. By flexibly modeling the association between phenotype and non-genetic covariates, DeepNull reduces the residual variation, and thereby increases the statistical power (Supplementary Fig. 1, Supplementary Note).

Consider a quantitative phenotype ascertained for a sample of n individuals genotyped at m SNPs. Let $Y = (y_i)_{i=1}^n$ denote the $n \times 1$ phenotype vector, where y_i is the phenotypic value of the i th individual; let $\mathbf{G} = [g_{ij}]$ denote the $n \times m$ sample by SNP genotype matrix, where g_{ij} is the minor allele count for the i th individual at the j th variant. Let $\bar{\mathbf{G}} = [\bar{g}_{ij}] \in \mathbb{R}^{n \times m}$ denote the standardized version of \mathbf{G} , in which columns have been centered and scaled to have mean zero and unit variance. Furthermore, let h be a (possibly non-linear) function that predicts the phenotype from non-genetic covariates; we learn h using a DNN trained with cross-validation on the sample. The DeepNull association model is as follows:

$$Y = \bar{\mathbf{G}}_j \beta_j + \tilde{\mathbf{X}} \gamma + H(\mathbf{X}) \gamma_h + \varepsilon. \quad (1)$$

Here β_j is the effect sizes for the j th variant on the phenotype; $\tilde{\mathbf{X}} = [x_{ik}]$ is the $n \times (p + g)$ covariate matrix that includes p non-genetic covariates (e.g. age and sex) and g adjustments for genetic confounding (e.g. genetic PCs); γ is the $(p + g) \times 1$ vector of association coefficients for all covariates. Compared with the standard GWAS association model, the DeepNull association model differs only by the inclusion of a single additional term $H(\mathbf{X}) \gamma_h$; \mathbf{X} is the $n \times p$ subset of $\tilde{\mathbf{X}}$ consisting of non-genetic covariates (see “Methods”); $H: \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^n$ is the function that applies h row-wise to \mathbf{X} ; and γ_h is the scalar association coefficient for the DNN's prediction of the phenotype based on non-genetic covariates.

DeepNull and Baseline perform similarly under linear effects.

We simulated phenotypes based on genotypes and covariates from the UK Biobank¹⁵. Standardized age, sex, and `genotyping_array` served as true covariates for 10,000 randomly sampled individuals (“Methods”). First, we considered a linear effect for covariates on phenotypes ($f(x) = \gamma x$). We simulated 100 phenotypes for each of six different genetic architectures with varying amounts of phenotypic variance explained by the genetic data (σ_g^2) and by covariates (σ_x^2): (i) $\sigma_g^2 = 0.2$ and $\sigma_x^2 = 0.1$; (ii) $\sigma_g^2 = 0.2$ and $\sigma_x^2 = 0.2$; (iii) $\sigma_g^2 = 0.4$ and $\sigma_x^2 = 0.1$; (iv) $\sigma_g^2 = 0.4$ and $\sigma_x^2 = 0.2$; (v) $\sigma_g^2 = 0.4$ and $\sigma_x^2 = 0.4$; and (vi) $\sigma_g^2 = 0.6$ and

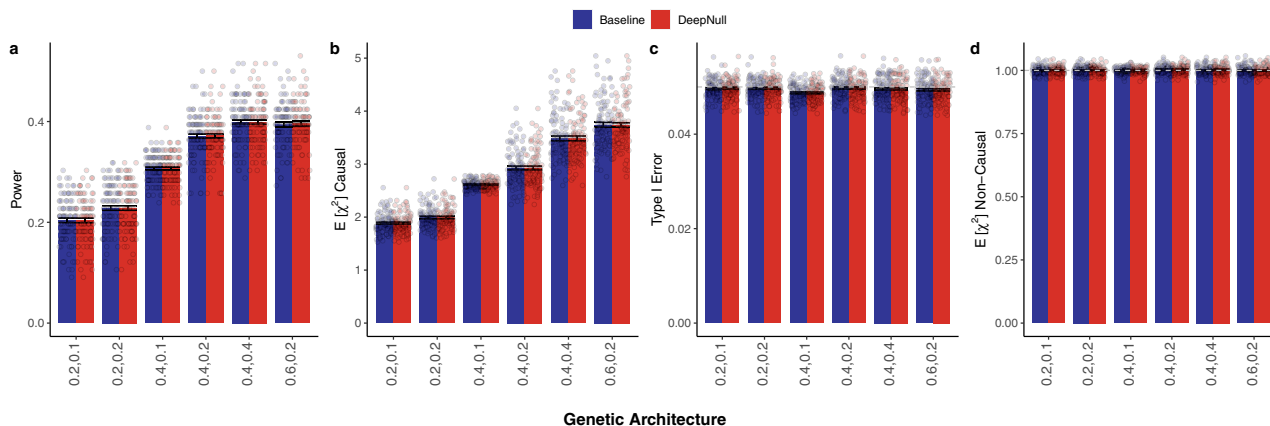


Fig. 1 DeepNull and baseline model achieve similar results under simulated linear covariate effects. **a** Statistical power, and **(b)** expected χ^2 statistics for variants in the causal chromosome (chr22); **(c)** type I error, and **(d)** expected χ^2 statistics for variants on the non-causal chromosomes (chr1 and chr2). In the case of power and the expected χ^2 statistics in the causal chromosome, higher is better. Methods should have a type I error of 0.05 (gray dashed horizontal line). The expected χ^2 statistics for the non-causal chromosomes should be 1 (gray dashed horizontal line). X-axis values indicate the proportion of phenotypic variance explained by genotypes and covariates, respectively. Error bars are the standard error of the mean for each estimate and each bar plot summarizes results from $n=100$ independent simulation replicates. None of the quantities shown is significantly different between Baseline and DeepNull (Wilcoxon signed-rank one-sided test). Source data are provided as a Source Data file.

$\sigma_x^2 = 0.2$. Causal variants were randomly embedded within chr22 and non-causal variants within chr1 and chr2. We compared the DeepNull GWAS with standard GWAS (hereafter referred to as “Baseline”), each of which was performed using BOLT-LMM¹⁸ (“Methods”). Statistical power and expected χ^2 statistics for the causal chromosome (chr22) were similar for DeepNull and Baseline (Fig. 1a, b, Supplementary Table 2). Statistical power for both DeepNull and Baseline increased as genetic heritability σ_g^2 increased, which is expected since the non-centrality parameter of the χ^2 test increases with the heritability. Additionally, the type I error was maintained at the nominal level, and the expected χ^2 statistics for non-causal variants are similar for both methods (Fig. 1c, d). Thus, DeepNull and Baseline produce similar GWAS results when the effect of the covariates on the phenotype is linear. Lastly, DeepNull and Baseline perform similarly both when excluding non-confounding covariates (i.e., hidden non-confounding covariates, Supplementary Table 3) and when including irrelevant covariates (Supplementary Table 4).

DeepNull increases power when covariates interact. We simulated phenotypes using a similar process as described above and used standardized age, sex, genotyping_array, age², age \times sex, and age \times genotyping_array as true covariates. However, both DeepNull and Baseline are only given age, sex, genotyping_array as known covariates. This simulation setting explores the case where the true covariates are known but their possible interactions are not. DeepNull had higher statistical power (2–13% relative improvement) than baseline, and higher expected χ^2 statistics at causal variants (2–20% relative improvement) across all genetic architectures (Fig. 2a, b, Supplementary Table 5). Importantly, both DeepNull and Baseline control the type I error and generate similar expected χ^2 statistics for non-causal variants (Fig. 2c, d).

DeepNull increases power under non-linear models. We simulated phenotypes using a similar process as described above and again used age, sex, genotyping_array, age², age \times sex, and age \times genotyping_array as true covariates. However, here we fix the genetic architecture ($\sigma_g^2 = 0.4$ and $\sigma_x^2 = 0.4$) and consider non-linear effects of the covariates on the phenotype by using different non-linear functions for $f(\cdot)$ in Eq.

(9): $\sin(x)$, $\exp(x)$, $\log(|x|)$, and $\text{sigmoid}(x)$. Again, both DeepNull and Baseline are only given age, sex, and genotyping_array as known covariates. In all cases, DeepNull outperforms Baseline both in terms of statistical power (3%–9% relative improvement) and expected χ^2 statistics (13%–22% relative improvement), while both methods control the type I error (Supplementary Table 6).

DeepNull is computationally efficient (Supplementary Notes) and its power increases as the sample size increases (Supplementary Notes; Supplementary Fig. 2, Supplementary Table 7). Finally, DeepNull’s results are not affected by random seed initialization (Supplementary Notes; Supplementary Fig. 3).

DeepNull detects more hits than Baseline GWAS on real data.

To explore whether applying DeepNull is beneficial in non-simulated data, we performed GWAS for ten phenotypes from the UK Biobank, using both Baseline and DeepNull. These were: alkaline phosphatase (ALP), alanine aminotransferase (ALT), aspartate aminotransferase (AST), apolipoprotein B (ApoB), calcium, glaucoma referral probability (GRP), LDL cholesterol (LDL), phosphate, sex hormone-binding globulin (SHBG), and triglycerides (TG), each of which has evidence of potentially non-linear relationships between covariates and the phenotype (Supplementary Figs. 4–13). All phenotypes except GRP were extracted directly from the UK Biobank. age, sex, and genotyping_array were considered as input covariates for DeepNull’s DNN (Supplementary Table 8). We performed GWAS for these phenotypes using age, sex, genotyping_array, and the top 15 genetic PCs as covariates.

GRP differs from the other phenotypes considered in that it was derived from color retinal fundus images, using the model presented in Alipanahi et al.¹⁹. As in that study, we are interested in biological signals for glaucoma that are not driven by the vertical cup-to-disc ratio (VCDR). Thus, for GRP only, several additional covariates were included in the association model: VCDRvisit, refractive-error, and image-gradability. To train DeepNull’s DNN, we used VCDRvisit, age, sex, and genotyping_array to predict GRP. We then performed GWAS for GRP using age, sex, genotyping_array, the top 15 PCs, VCDRvisit, refractive-error, and image-gradability as covariates.

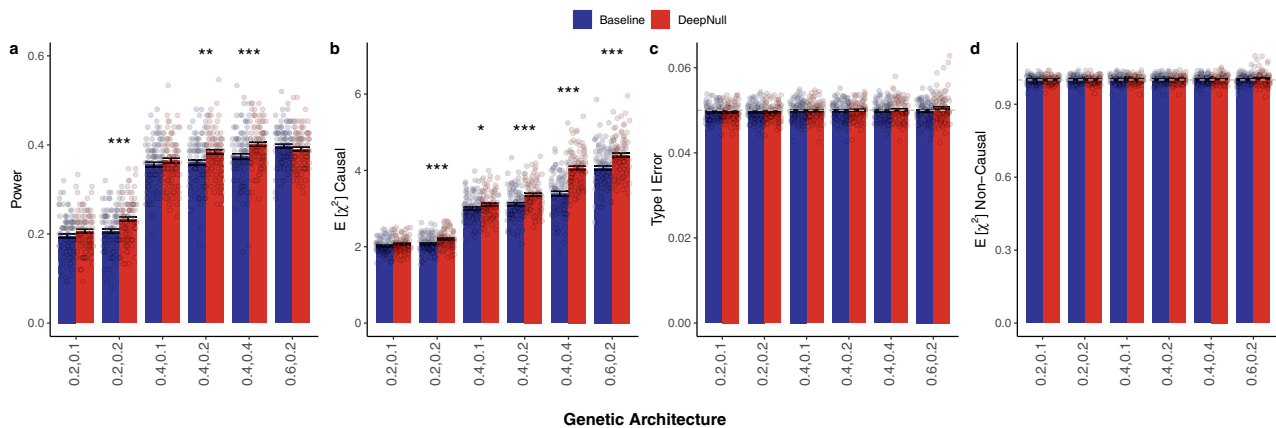


Fig. 2 DeepNull increases power in the presence of covariate interactions. **a** Statistical power, and **(b)** expected χ^2 statistics for variants in the causal chromosome (*chr22*); **(c)** type I Error, and **(d)** expected χ^2 statistics for variants in the non-causal chromosomes (*chr1* and *chr2*). In the case of power and expected χ^2 statistics for the causal chromosome, higher is better. Methods should maintain a type I error of no more than 0.05, which is shown by the dashed gray horizontal line. For the non-causal chromosomes, the expected χ^2 statistics should be 1, which is also shown in dashed gray horizontal line. X-axis values indicate the proportion of phenotypic variance explained by genotypes and covariates, respectively. Error bars are the standard error of the mean for each estimate and each bar plot summarizes results from $n = 100$ independent simulation replicates. The numerical results are shown in Supplementary Table 5. Indicators for P value (Wilcoxon signed-rank one-sided test) ranges: $P \leq 0.05$, $**P \leq 0.01$, $***P \leq 0.001$. Source data are provided as a Source Data file.

Table 1 DeepNull improves association results relative to the Baseline model on ten phenotypes from the UK Biobank.

Pheno	<i>n</i>	#Hits		%Improve	#Loci		%Improve
		Baseline	DeepNull		Baseline	DeepNull	
ALP	416,232	1697	1759	3.65%	336	350	4.17%
ALT	416,057	371	379	2.16%	173	174	0.58%
AST	414,743	337	351	4.15%	137	145	5.84%
ApoB	414,639	1172	1219	4.01%	200	217	8.50%
Calcium	381,934	726	739	1.79%	272	281	3.31%
GRP	65,896	28	38	35.71%	26	38	46.15%
LDL	415,892	950	993	4.53%	193	202	4.66%
Phosphate	381,362	658	664	0.91%	224	229	2.23%
SHBG	378,459	1084	1120	3.32%	319	323	1.25%
TG	416,295	1221	1254	2.70%	261	266	1.92%
Avg.	370,151	824.4	851.6	6.29%	214.1	222.5	7.86%

n is the sample size, hits refers to the number of independent genome-wide significant associations detected, and loci is the number of independent regions after merging hits within 250 kb. Bold values in the table indicate the best results.

Phenotypic abbreviations: ALP alkaline phosphatase, ALT alanine aminotransferase, AST aspartate aminotransferase, ApoB Apolipoprotein B, GRP glaucoma referral probability, LDL low-density lipoprotein, SHBG sex hormone-binding globulin, TG triglycerides.

For all GWAS, we first verified that the DeepNull prediction was consistent across all five data folds (Supplementary Table 9). After running GWAS across the entire dataset, we computed the stratified LD score regression (S-LDSC) intercept^{20,21} to determine whether there was evidence of inflation due to confounding. In no case did the S-LDSC intercept differ significantly from 1, providing no evidence of inflation due to confounding in our analysis (Supplementary Table 10). In addition, the SNP-heritability of all phenotypes was estimated from both the DeepNull and Baseline summary statistics. For all phenotypes except GRP, the heritability was nominally, though not significantly, greater with DeepNull (Supplementary Table 10).

DeepNull detects more genome-wide significant hits (i.e. independent lead variants) and loci (independent regions after merging hits within 250 kbp together; see Methods) than Baseline for all phenotypes examined (Table 1). For example, we found 46% more significant loci (38 vs. 26) for GRP using DeepNull compared to the Baseline model. Similarly, in the case of LDL, we detected 202 significant loci using DeepNull compared to the 193 significant loci detected with Baseline (4.5% more hits and

4.7% more loci). In addition, 99 of the DeepNull loci were replicated in the GWAS catalog compared with 96 loci for Baseline (Supplementary Fig. 14). For ApoB, DeepNull detected 1219 hits compared to 1172 hits detected by Baseline (4.0% improvement) and DeepNull detected 217 significant loci compared to 200 significant loci obtained from Baseline (8.5% improvement; see Table 1). In addition, 166 of the DeepNull loci were replicated in the GWAS catalog compared with 165 loci for Baseline (Supplementary Fig. 15). For these three phenotypes, we further investigated the biological significance of the detected associations using FUMA²² (Supplementary Table 11). For GRP, 42 gene sets, predominantly related to pigmentation, were enriched among DeepNull's results, whereas none were enriched among the Baseline results. For LDL, DeepNull detected more gene sets overall (955 Baseline vs. 1000 DeepNull), although the gene sets detected by Baseline scored higher in terms of the average $-\log_{10}(p\text{-value})$ (8.60 Baseline vs. 8.38 DeepNull). However, when focusing on the subset related to lipid metabolism, DeepNull detected more gene sets (65 Baseline vs. 72 DeepNull) and did so at a higher level of significance (average

$-\log_{10}(p\text{-value})$: 13.88 Baseline vs. 14.34 DeepNull). For ApoB, DeepNull detected fewer gene sets overall (983 Baseline vs. 946 DeepNull), but at a higher level of significance (average $-\log_{10}(p\text{-value})$: 7.65 Baseline vs. 7.81 DeepNull). The gene sets detected by DeepNull related to lipid metabolism and neurological conditions, including Alzheimer's disease.

Overall, the average percentage improvement with DeepNull, taken across phenotypes, was 6.29% for significant hits and 7.86% for loci (Table 1). The average number of hits increased by 3.29%, from 824.4 for Baseline to 851.6 for DeepNull, and the average number of loci increased by 3.93%, from 214.1 to 222.5. In addition, the median number of hits and loci increased by 3.48% and 3.74%, respectively. Lastly, DeepNull tends to have a higher level of significance for variants compared to Baseline (Supplementary Figs. 16–25).

To further understand the source of the DeepNull improvements, we evaluated three additional Baseline models of increasing complexity and a gradient boosted decision tree (GBDT) non-linear model. The first model, which we call "Baseline+ReLU", featurizes age into five additional covariates by applying the ReLU function at different thresholds (and solely for GRP, also featurizes VCDRvisit in the same way). We observed that while Baseline+ReLU generally identified more significant hits and loci than Baseline, DeepNull consistently outperformed both baseline methods (Supplementary Table 12). The second model, which we call "Second-order Baseline", extends the Baseline model to include all second-order interactions between age, sex, and genotyping_array: age^2 , $\text{age} \times \text{sex}$, $\text{age} \times \text{genotyping_array}$, and $\text{sex} \times \text{genotyping_array}$. Although the additional second-order interaction covariates consistently improve over the Baseline model results, DeepNull detects as many or more significant loci than Second-order Baseline for nine of the 10 phenotypes (Supplementary Table 13). For AST, LDL, phosphate, and TG, Second-order Baseline and DeepNull detected similar numbers of hits and loci (Supplementary Tables 14 and 15), providing evidence that the hits and loci not found by the Baseline model, which does not include interactions, were in fact true signals. The utility of DeepNull arises because the optimal order of covariate interactions is unknown a priori (Supplementary Table 1), exhaustively

enumerating higher order interactions in impractical, and attempting to do so will likely introduce collinearity. Next, we compared the number of hits and loci of DeepNull with an extended Baseline model that performs sex-specific spline fitting (Methods) and observed that DeepNull outperforms this Baseline extension as well (compare Supplementary Tables 14, 16 for hits and Supplementary Tables 15, 17 for loci). Finally, we compared the number of hits and loci of DeepNull with a non-linear GBDT model (Methods) and observed similar numbers of hits and loci (Supplementary Tables 16, 17).

DeepNull improves phenotype prediction for UKB phenotypes.

An important feature of DeepNull is that it provides additional signal for phenotype prediction. Typically, phenotype prediction models are created using a linear combination of common covariates (such as age and sex) and a polygenic risk score (PRS) defined using GWAS association results. Covariate interactions or higher order terms are occasionally included, but typically in an ad hoc fashion. DeepNull provides a way to easily include potential covariate interactions or higher order terms. The Baseline model includes a PRS computed using PLINK ($\text{PRS}_{\text{baseline}}$) and linear covariate effects ($\text{PRS}_{\text{baseline}} + \text{Linear covariates}$). The DeepNull-Baseline model includes a PRS computed in the same way except using association results from DeepNull ($\text{PRS}_{\text{DeepNull}} + \text{Linear covariates}$), and DeepNull is a model that includes both the DeepNull-based PRS and the DeepNull prediction (non-linear covariate effects).

When compared to the Baseline model, the DeepNull model performs significantly better in terms of the Pearson R^2 (Fig. 3). We calculated R^2 following previous works^{23,24}. We observed that in the case of GRP, LDL, calcium, and ApoB, DeepNull improves phenotype prediction by 83.42%, 40.33%, 23.90%, and 21.61%, respectively. Overall, DeepNull improves phenotype prediction (average improvement = 23.72%, median improvement = 16.08%) across the ten phenotypes analyzed (average $n = 370\text{K}$; Supplementary Table 18). In addition, DeepNull has an average R^2 of 0.1940 compared to Baseline average R^2 of 0.1315 (33.65% improvement; Supplementary Table 18). To determine whether the improved predictive power stems from more accurate GWAS

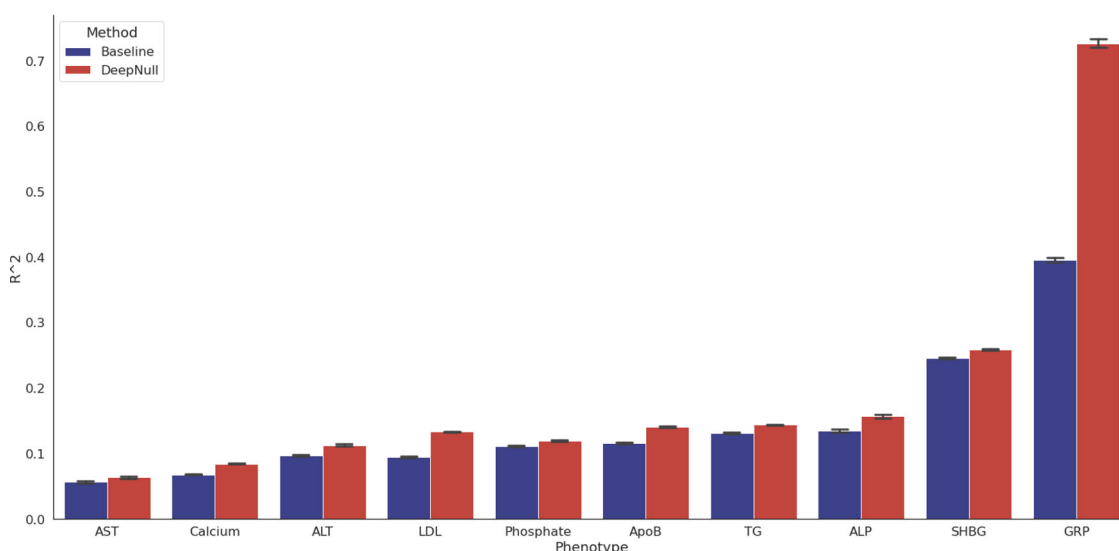


Fig. 3 DeepNull improves phenotype prediction compared to Baseline. The X-axis provides the phenotype names and the Y-axis is the R^2 where R is Pearson's correlation between true and predicted value of phenotypes. Center of each bar indicates the computed R^2 over all samples and the error bars indicate the standard error. Standard errors are computed by performing bootstrapping for each phenotype ($n = 1000$ bootstrapping trials). Phenotypic abbreviations: alkaline phosphatase (ALP), alanine aminotransferase (ALT), aspartate aminotransferase (AST), apolipoprotein B (ApoB), glaucoma referral probability (GRP), LDL cholesterol (LDL), sex hormone-binding globulin (SHBG), and triglycerides (TG).

effect size estimates or inclusion of the DeepNull DNN prediction, we examined predictive performance of a model that uses age, sex, and $\text{PRS}_{\text{DeepNull}}$ ("DeepNull-Baseline"). This model produces slightly higher R^2 compared to Baseline for seven of the ten phenotypes, though the difference is not statistically significant for any phenotype (Supplementary Table 18), indicating that most of the improved predictive power arises due to better modeling the effects of non-genetic factors. Next, we compared phenotype prediction of DeepNull to an extended Baseline model that incorporates second-order interactions (additional covariates such as age^2 , $\text{age} \times \text{sex}$, $\text{age} \times \text{genotyping_array}$). The second-order Baseline model produces similar R^2 to DeepNull for many of the phenotypes, but DeepNull increases phenotype prediction of GRP by 11.81% (compare Supplementary Tables 13, 18). Third, we compared phenotype prediction of DeepNull to an extended Baseline model that performs sex-specific spline fitting (Methods) and observed that DeepNull outperforms this Baseline extension as well (compare Supplementary Tables 18, 19). Finally, we compared phenotypic prediction of DeepNull to a non-linear GBDT model ("Methods") and observed similar performance (Supplementary Tables 20, 21).

DeepNull's covariates should remain in the association model.

When performing genetic association analysis via the model shown in Eq. (1), the covariates X input row-wise to the DNN prediction function h are also included as components of the linear term $\bar{X}y$. This secondary adjustment for X is necessary because h captures the association between the covariates X_i and the phenotype y_i , but does not capture any association between the covariates X_i and genotype \bar{g}_{ij} . Failure to include X_i in the final association model is comparable to projecting X_i out of y_i but not g_{ij} . To empirically demonstrate the necessity of adjusting X_i in the final association model, we generated phenotypes via

$$y_i = \bar{g}_i\beta + x_i\gamma_1 + x_i^2\gamma_2 + \epsilon_i.$$

For this simulation only, \bar{g}_i was generated as a continuous random variable, allowing for fine control of the correlation between \bar{g}_i and x_i , and the model h for predicting y_i from x_i was the oracle model

$$y_i = x_i\gamma_1 + x_i^2\gamma_2 + \epsilon_i.$$

We compare two methods for estimating the genetic effect β . The unadjusted model incorporates the prediction $h(x_i)$ of y_i based on x_i but omits x_i from the association model, emulating the exclusion of covariates provided to DeepNull from the association model as shown in Eq. (1),

$$y_i = \bar{g}_i\beta + h(x_i)\gamma_h + \epsilon_i. \quad (2)$$

The adjusted model includes both $h(x_i)$ and a linear correction for x_i , emulating the application of (1) in practice where the functional form linking y_i and x_i is unknown,

$$y_i = \bar{g}_i\beta + x_i\gamma_1 + h(x_i)\gamma_h + \epsilon_i. \quad (3)$$

Figure 4 presents the relative bias of the unadjusted and linearly adjusted models for estimating the association parameter β . The relative bias for estimating β from the generative model, which represents the best possible performance, is also provided. For these simulations $\gamma_1 = 2$, $\gamma_2 = -1$, and $\beta \in \{\pm 1, \pm 2, \pm 3\}$; the correlation between \bar{g}_i and x_i was 0.5. The unadjusted estimate is generally biased. The magnitude and direction of the bias depend on the coefficients of the generative model. For the unadjusted estimator to be unbiased, \bar{g}_i and x_i must be independent. Since the dependence of \bar{g}_i and x_i is seldom clear, and the linearly adjusted model is unbiased in either case, we adopted the linearly adjusted model for all other analyses. Moreover, the linearly adjusted

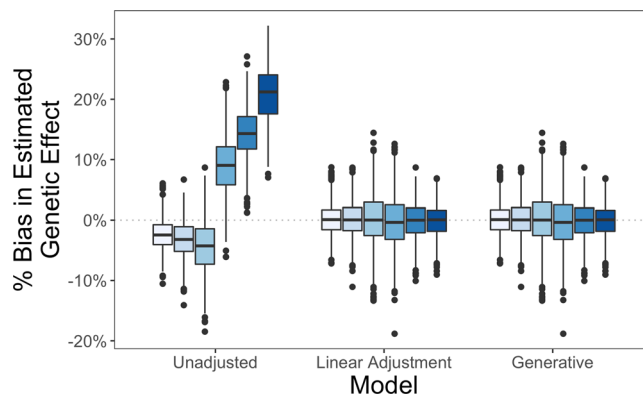


Fig. 4 Adjusting for covariates provided to DeepNull during association testing is necessary to avoid bias. The unadjusted model regresses y_i on \bar{g}_i and $h(x_i)$, the prediction of y_i based on x_i , omitting x_i from the association model. This approach results in biased estimation of the genetic effect. The linear adjustment model regresses y_i on \bar{g}_i , x_i , and $h(x_i)$. This approach is unbiased. The generative model regresses y_i on \bar{g}_i , x_i , and x_i^2 . This represents the best possible performance. Each box plot summarizes results from $n = 10^3$ independent simulation replicates. The box demarcates, from top to bottom, the 75th, 50th, and 25th percentiles of the corresponding distribution. The whiskers extend between the largest and smallest values within 1.5 times the interquartile range. Any values outside the whiskers are marked by points. Source data are provided as a Source Data file.

estimator remained unbiased in the presence of lower- and higher order covariate effects (Supplementary Figs. 26, 27).

Discussion

A typical GWAS examines the association between genotypes and the phenotype of interest while adjusting for a set of covariates. While covariates potentially have non-linear effects on the phenotype in many real world settings, due to the challenge of specifying the model, GWAS seldom include non-linear terms. Although it is theoretically possible to model the non-linear effects by considering all possible covariate interactions in a linear model, this approach has multiple limitations. First, the optimal order of covariate interactions is unknown a priori (Supplementary Table 1) as it depends on the particular phenotype and set of covariates. Second, adding higher order covariate interactions requires careful analysis to avoid overfitting and collinearity. We proposed a new framework, DeepNull, that can model the non-linear effect of covariates on phenotypes when such non-linearity exists. We show that DeepNull can substantially improve phenotype prediction. In addition, we show that DeepNull achieves results similar to a standard GWAS when the effect of covariate on the phenotype is linear, and can significantly outperform a standard GWAS when the covariate effects are non-linear. DeepNull reduces residual variation, thereby increasing statistical power (Supplementary Fig. 1).

Increasing the statistical power of GWAS is an area of active research that aims to uncover the many variants, each with individually small effect sizes, that collectively explain substantial variation in complex traits and diseases. Multiple complementary approaches have been proposed for increasing statistical power. The most fundamental is to increase the sample size²⁵. However, when resources are limited, the sample size cannot be increased indefinitely, and power can be improved through the use of more refined statistical analyses. Linear mixed models (LMMs) were introduced to perform GWASs that include related individuals, who are not statistically independent^{18,26–33}. An orthogonal modeling-based approach is to remap or transform the

phenotype to make the distribution of phenotypic residuals more nearly normal^{34–38}. While normality of the phenotypic residuals is not necessary for valid association testing, standard association tests are most powerful when the residuals are in fact normally distributed. The final class of methods increases power by leveraging external data on the prior biological plausibility of the variants under study. Highly conserved variants, variants in exons, and protein-coding variants all have higher prior probability of being causal than variants in intergenic regions. A series of methods have been developed that incorporate functional data to detect biologically important variants and up-weight their association statistics or reduce their significance thresholds^{39–44}. By focusing on capturing non-linear covariate effects, DeepNull constitutes a distinct approach to improving statistical power of GWAS, one which can be used in combination with any or all of the approaches discussed above.

We note several limitations of our work. First, while training the DeepNull model, we assume individuals (e.g. samples) are independent. Although this is a general assumption among machine learning methods and optimization frameworks, this is not necessarily true in the presence of related individuals. Thus, we believe that an ML optimizer that can incorporate sample relatedness may improve the prediction accuracy of DeepNull’s DNN. Importantly, although DeepNull makes the independence assumption during training, this does not mean that type I error is not controlled. Our analyses used BOLT-LMM to perform the association testing, which does correctly account for the relatedness between individuals. Second, DeepNull does not attempt to model possible genotype-covariate ($G \times X$) or genotype-genotype ($G \times G$) interactions. This limitation is shared by standard GWAS and can only be overcome by employing different statistical models that explicitly capture these interactions during association testing. Third, DeepNull’s DNN is not easily interpretable compared to less expressive models such as the Baseline model. For improving GWAS power, this is not a major limitation as the parameter of interest is the coefficient describing the relationship between genotype and phenotype. By estimating this coefficient within a linear model that incorporates DeepNull’s prediction of phenotype, we obtain a more precise estimate of the genetic effect. For more interpretable phenotypic prediction, possibly at the expense of some prediction accuracy, it may be beneficial to use an alternative non-linear model such as spline regression, generalized additive models⁴⁵, symbolic regression⁴⁶, or neural additive model⁴⁷. Alternatively, the trained DeepNull model can be interrogated with a variety of methods^{48–51}, although we note that DNN interpretability is still an active and evolving area of research. Lastly, DeepNull is a proof of concept. For some phenotypes, a simpler model such as the Second-order Baseline model may suffice to capture the phenotype-covariate relationship. For others, an alternative non-linear model such as a GBDT may perform similarly to DeepNull’s DNN; for the 10 example UKB phenotypes presented here, a GBDT implemented in XGBoost provided similar performance. Although XGBoost and DNN performed similarly for these phenotypes, the added flexibility of DNNs may prove advantageous for other phenotypes or sets of covariates. For example, DNNs can handle complex inputs such as image and text that XGBoost typically cannot. Importantly, we observed in all cases that DeepNull performed as well or better than current standard practice, and the underlying DNN is sufficiently expressive to capture many of the phenotype-covariate relationships likely to be encountered in practice.

By accurately modeling the non-linear interactions between covariates and the phenotype of interest, DeepNull improved phenotype prediction and association power, both in simulations and on 10 UKB phenotypes. Software for performing end-to-end cross-validated training and prediction is freely available (Code

Availability). The resulting phenotypic predictions can readily be included among the input data to commonly-used GWAS models, including PLINK and BOLT-LMM. The improved performance of DeepNull, combined with its ease of use, suggest that it or similar approaches to modeling non-linear covariate effects should become a standard component of performing phenotypic prediction and association testing.

Methods

Notation: We use bold capital letters to indicate matrices, non-bold capital letters to indicate vectors, and non-bold lowercase letters to indicate scalars.

Standard GWAS. We consider GWAS of a quantitative trait for a sample of n individuals genotyped at m SNPs. Let $Y = (y_i)_{i=1}^n$ denote the $n \times 1$ phenotype vector, where y_i is the phenotypic value of the i th individual, and $G = [g_{ij}]$ the $n \times m$ sample by SNP genotypes matrix, where g_{ij} is the minor allele count for the i th individual at the j th variant. Since human genomes are diploid, each variant has 3 possible minor allele counts: $g_{ij} \in \{0, 1, 2\}$. $G_{\cdot j} = (g_{ij})_{i=1}^n$ is a vector of minor allele counts for all individuals at the j th SNP. For simplicity, assume the phenotypes and genotypes are standardized to have zero mean and unit variance. Let $\bar{G} = [\bar{g}_{ij}] \in \mathbb{R}^{n \times m}$ be the standardized version of G , i.e. the empirical mean and variance of $\bar{G}_{\cdot j}$ are zero and one, respectively: $\frac{1}{n} \sum_i \bar{g}_{ij} = 0$ and $\frac{1}{n} \sum_i \bar{g}_{ij}^2 = 1$ for each j th SNP.

A typical GWAS assumes the effect of each variant on the phenotype is linear and additive. Thus, we have the following generative model:

$$Y = \bar{G}\beta + X\gamma + \epsilon \tag{4}$$

where β is the $m \times 1$ vector of effect sizes for each variant on the phenotype, $X = [x_{ik}]$ is the $n \times q$ covariate matrix, including covariates such as age and sex, and γ is the $q \times 1$ vector of association coefficients for the covariates. Let X indicate covariates not directly derived from genotypic data (“non-genetic covariates”). For genotypes $g_{ij} \in \{0, 1, 2\}$ the assumptions of linearity and additivity are not restrictive. On the other hand, a typical GWAS also assumes that the covariates are linearly associated with the phenotype. This is a far more restrictive assumption if any of the covariates are continuous. $\epsilon = (\epsilon_i)_{i=1}^n$ is an $n \times 1$ residual vector that models the environmental effects and measurement noise.

To perform a GWAS, each variant is individually tested for association with the phenotype. For example, the j th variant is tested for association using the following model:

$$Y = \bar{G}_{\cdot j}\beta_j + \tilde{X}\tilde{\gamma} + \epsilon \tag{5}$$

Here \tilde{X} contains the known set of covariates (e.g. age and sex), in addition to adjustments for confounding that become necessary when the genotypes at SNPs $\tilde{j} \neq j$ are omitted from the model shown in Eq. (4). Confounding due to the presence of genetically related subgroups within the sample, for example subgroups of individuals with common ancestry, is referred to as population structure, and is commonly accounted for by including the top several genetic PCs in \tilde{X} ^{10,11,52}.

The model in Eq. (5) can be simplified by projecting away the covariates^{18,53}. Define $P = I - \tilde{X}(\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T$, which is the projection onto the orthogonal complement of the linear subspace spanned by \tilde{X} . Multiplying Eq. (5) through by P on the left yields:

$$PY = P\bar{G}_{\cdot j}\beta_j + \epsilon^* \tag{6}$$

The projected phenotype PY is the residual from regression of Y on \tilde{X} . Likewise, $P\bar{G}_{\cdot j}$ is the residual from regression of $\bar{G}_{\cdot j}$ on \tilde{X} . Importantly, if $\bar{G}_{\cdot j}$ and \tilde{X} are dependent, which is necessarily true if \tilde{X} contains confounders of the genotype-phenotype relationship, then $P\bar{G}_{\cdot j}$ will differ from $\bar{G}_{\cdot j}$. Consequently, an analysis that residualizes only Y with respect to \tilde{X} will be misspecified. Instead, to remove dependence on \tilde{X} , both Y and $\bar{G}_{\cdot j}$ should be residualized in Eq. (5).

Though including genotypic PCs can control for population structure, it fails to correct for cryptic or family relatedness between individuals^{26,27,54,55}. LMMs were introduced to GWAS to overcome these limitations^{18,26–33}. LMMs account for random variation in the phenotypic mean that is correlated with the genetic relatedness of the individuals under study, and have proven effective for increasing power even when the kinship among subjects is more distant^{18,32,33}. We use BOLT-LMM^{18,33} to perform our analyses and we refer to it as the Baseline method.

DeepNull model. In this work, we consider a model in which the covariates have potentially non-linear effect on the phenotypes. The corresponding generative model for an individual i can be written as

$$y_i = \bar{G}_i\beta + f(X_i)\gamma + \epsilon_i$$

where all variables are defined identically as in formula (4), $f: \mathbb{R}^q \rightarrow \mathbb{R}$ is any

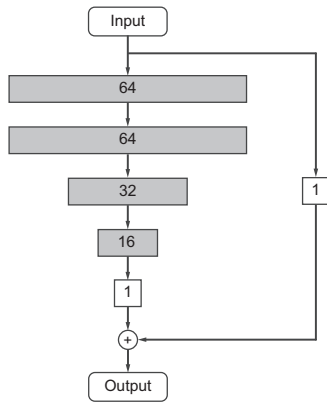


Fig. 5 DeepNull DNN model architecture. Each rectangle represents one layer and all layers are fully connected. Shaded layers use the ReLU activation and the non-shaded layers do not use an activation function (i.e. linear connection). The input is the set of known covariates and the output is the predicted phenotype.

(potentially non-linear) function, $\bar{G}_i = (\bar{g}_{ij})_{j=1}^m$, and $X_i = (x_{ik})_{k=1}^q$. In vector form,

$$Y = \bar{G}\beta + F(X)\gamma_f + \epsilon$$

where $F: \mathbb{R}^{n \times q} \rightarrow \mathbb{R}^n$ is the function that applies f to each row of X .

We convert the estimation of $u_i = f(X_i)$ into a learning problem, where we predict u_i using y_i and X_i as targets and input features, respectively. In other words, we train a model h using the covariates X_i and the phenotype y_i by minimizing

$$\|y_i - h(X_i)\|^2. \tag{7}$$

We designed a DNN architecture for modeling the function h (Fig. 5). We explored the model proposed previously to detect interpretable statistical interactions⁵⁶ but found that a simpler model with an explicit linear effect performed equally well on four UKB phenotypes tested (data not shown). The resulting model is inspired by residual networks⁵⁷ and consists of two components. One component (the shorter path from input to output in Fig. 5) is linear, to directly represent the linear effect of the covariates on the phenotype. The other component (the longer path in Fig. 5) is a multi-layer perceptron (MLP), to model a potentially non-linear effect of the covariates. The MLP component has 4 hidden layers, all of which use the Rectified Linear Unit (ReLU) activation.

In an equation form, the DeepNull model can be written as

$$h(X_i) = H^{(5)} + H^{(6)},$$

where

$$\begin{aligned} H^{(1)} &= \phi(W_{64 \times q}^{(1)} X_i + B_{64 \times 1}^{(1)}) \\ H^{(2)} &= \phi(W_{64 \times 64}^{(2)} H^{(1)} + B_{64 \times 1}^{(2)}) \\ H^{(3)} &= \phi(W_{32 \times 64}^{(3)} H^{(2)} + B_{32 \times 1}^{(3)}) \\ H^{(4)} &= \phi(W_{16 \times 32}^{(4)} H^{(3)} + B_{16 \times 1}^{(4)}) \\ H^{(5)} &= W_{1 \times 16}^{(5)} H^{(4)} + B_{1 \times 1}^{(5)} \\ H^{(6)} &= W_{1 \times q}^{(6)} X_i + B_{1 \times 1}^{(6)} \end{aligned}$$

and ϕ is the coordinate-wise ReLU function, i.e.

$$\phi\left(x_p\right)_p = \left(\max(0, x_p)\right)_{p=1}^p.$$

DeepNull learns

$$W = \{W_{64 \times q}^{(1)}, W_{64 \times 64}^{(2)}, W_{32 \times 64}^{(3)}, W_{16 \times 32}^{(4)}, W_{16 \times 32}^{(4)}, W_{1 \times 16}^{(5)}, W_{1 \times q}^{(6)}\}$$

and

$$B = \{B_{64 \times 1}^{(1)}, B_{64 \times 1}^{(2)}, B_{32 \times 1}^{(3)}, B_{16 \times 1}^{(4)}, B_{16 \times 1}^{(4)}, B_{1 \times 1}^{(5)}, B_{1 \times 1}^{(6)}\}$$

by minimizing the mean squared error in (7) using the Adam optimizer⁵⁸ implemented in Keras for TensorFlow 2. Adam is run with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. We also used a batch_size of 1024 and a learning_rate of 10^{-4} . We train DeepNull for 1,000 epochs (running DeepNull with more epochs can improve the results with the cost of increasing the training time), without early stopping, batch normalization, or dropout. Kernel initializers were set to default (glorot_uniform) and bias initializers were set to default (zeros).

Performing GWAS using DeepNull. After training DeepNull, we use the following model to test for association between the j th variant and the phenotype:

$$y_i = \bar{g}_{ij}\beta_j + h(X_i)\gamma_h + \bar{X}_i\gamma + \epsilon.$$

The vectorized form of the above association test is

$$Y = \bar{G}_j\beta_j + H(X)\gamma_h + \bar{X}\gamma + \epsilon. \tag{8}$$

Where $H: \mathbb{R}^{n \times q} \rightarrow \mathbb{R}^n$ is the function that applies h to each row of X . Compared to the standard GWAS association model in Eq. (5), the DeepNull association model differs only by the inclusion of an extra term $H(X)\gamma_h$, where $h(X_i)$ is the DNN’s prediction of the phenotype, based on non-genetic covariates only, and γ_h is a scalar association coefficient. As in the model shown in Eq. (5), \bar{X} includes both non-genetic covariates (e.g. age and sex) and adjustments for confounding (e.g. genetic PCs) while X excludes PCs. PCs are excluded because the aim of DeepNull is to predict phenotypes without utilizing genetic data, whereas the PCs are computed from genotypes. In addition, higher order interactions of PCs may capture true biological signals that it is not desirable to remove (e.g. conditional associations) in GWAS.

To avoid overfitting, DeepNull should be trained and run on distinct sets of individuals. However, to maximize the GWAS’s statistical power, all individuals in the cohort should receive DeepNull predictions. To satisfy both of these criteria, we split the cohort by individual into k partitions. For each selected partition, we train a DeepNull model using data from $k - 2$ of the other partitions and use the remaining partition for validation and model selection. The model that performs best on the validation partition is then used to predict all individuals in the selected partition. The partitioning scheme ensures that each partition is used as the validation/selection partition exactly once.

Simulation framework. We simulate data using the model

$$Y = \bar{G}\beta + \sum_{k=1}^q f(X_k)\gamma_k + \epsilon \tag{9}$$

where X_k is the value of the k -th covariate for all individuals, γ_k is the effect size, and $f(\cdot)$ is an arbitrary function from \mathbb{R} to \mathbb{R} , such as the identity $f(x) = x$ or exponential function $f(x) = \exp(x)$. For $j = 1, \dots, m$, the variant effect sizes β_j are drawn independently from a normal distribution with mean zero and variance equal to $\frac{\sigma_g^2}{m}$ where $\sigma_g^2 \in [0, 1]$ is the proportion of phenotypic variance explained by genotype (i.e., the heritability) and m is the number of causal variants:

$\beta_j \stackrel{iid}{\sim} \mathcal{N}(0, \frac{\sigma_g^2}{m})$. Similarly, the covariate effects are drawn independently from a normal distribution with mean zero and variance equal to $\frac{\sigma_x^2}{q}$ such that σ_x^2 is the proportion of phenotypic variance explained by the covariates: $\gamma_k \stackrel{iid}{\sim} \mathcal{N}(0, \frac{\sigma_x^2}{q})$.

Lastly, ϵ is drawn from another independent normal distribution with mean 0 and variance $1 - (\sigma_g^2 + \sigma_x^2)$: $\epsilon \sim \mathcal{N}(0, 1 - \sigma_g^2 - \sigma_x^2)$. Under this model, $\mathbb{E}(Y) = 0$ and $\mathbb{V}(Y) = \mathbb{E}(Y^2) = 1$. In the case $f(\cdot)$ is the identity function $f(x) = x$, our simulation framework is similar to previous works^{18,32}.

Phenotypes were simulated based on genotypes and covariates from the UKB. Age, sex, and genotype_array were included as covariates. Causal variants were selected uniformly at random from chr22 such that 1% variants (i.e., 127 variants) were causal. Association testing was performed using BOLT-LMM³³ applied to chromosomes chr1, chr2, and chr22. BOLT-LMM is a linear mixed model that incorporates a Bayesian spike-and-slab prior for the random effects attributed to variants other than that being tested. The prior allows for a non-infinitesimal genetic architecture, in which a mixture of both small and large effect variants influence the phenotype. Specifically, the BOLT-LMM association statistic arises from Eq. (8) with the inclusion of an additional random effect $\bar{G}^{(-j)}\delta$. Here $\bar{G}^{(-j)}$ denotes genotype at all variants not on the same chromosome as variant j , and the components of δ follow the spike-and-slab prior¹⁸.

In our setting, chr1 and chr2 are utilized to compute the type I error of the association test, which is the proportion of non-causal variants erroneously associated with the phenotype at a given significance threshold α (e.g. $\alpha = 0.05$). For null SNPs, the expected χ^2 statistic is 1. Methods that effectively control type I error are compared with respect to their power for correctly rejecting the null hypothesis⁵⁹⁻⁶¹, and their expected χ^2 statistics^{18,32,33}. Power is defined as the probability of correctly detecting that a variant with a non-zero effect size is causal⁵⁹⁻⁶¹. Additionally, the expected χ^2 statistic of an association method is a proxy for its prediction accuracy^{18,32,33}.

UKB GWAS evaluation. All GWASs were performed in a subset of UKB individuals of European genetic ancestry, identified as in Alipanahi et al.¹⁹. Briefly, the medioid of the top 15 genetic PC values of all individuals with self-reported “British” ancestry was computed, then the distance from each individual in UKB to the British medioid was computed and all individuals within a distance of 40 were retained. The threshold of 40 was selected based on the 99th percentile of distances of individuals who self-identify as British or Irish.

Association testing was performed via BOLT-LMM^{18,33} (Code Availability) with covariates specific to each experiment. GWAS “hits” were defined as genome-

wide significant (i.e. $P \leq 5 \times 10^{-8}$) lead variants that are independent at an R^2 threshold of 0.1. Hits were identified using the `--clump` command in PLINK (Code Availability). The linkage disequilibrium (LD) calculation was based on a reference panel of 10,000 randomly sampled unrelated subjects of European ancestry from the UKB. The span of each hit was defined based on the set of reference panel variants in LD with the hit at $R^2 \geq 0.1$. GWAS "loci" were defined by merging hits within 250 Kbp.

Comparison of two GWAS results G_1 and G_2 for shared and unique hits was performed by examining overlap of the hit spans; a given hit H_1 from G_1 is classified as shared if the span of any hit from G_2 overlaps it, otherwise it is classified as unique.

Comparison of our GWAS with the GWAS catalog (Code Availability) was performed analogously to comparing two GWAS. We used `gwas_catalog_v1.0.2-associations_e100_r2021-04-05` and converted coordinates from GRCh38 to GRCh37 using UCSC LiftOver (Code Availability) with default parameters. All catalog variants whose "DISEASE/TRAIT" column matched the phenotype of interest and were genome-wide significant were converted into loci by merging variants within 250 Kbp.

Learning phenotype-covariates relationship via spline regression. We can learn the non-linear relationship between the phenotype and covariates by fitting sex-specific spline regression models to predict the desired phenotype using a set of covariates. For each sex, we learn an independent spline regression model based on the other non-genetic covariates. We utilized the python scikit-learn package (Code Availability) to perform spline fitting.

Learning phenotype-covariates relationship via XGBoost. We can also learn the non-linear relationship between the phenotype and covariates by fitting gradient boosted decision trees. XGBoost (Code Availability) is one existing implementation of gradient boosted decision trees. We utilized XGBoost to learn the non-linear relationship. The optimal XGBoost hyperparameters were selected by performing black-box hyperparameter optimization with Google Vizier⁶². The optimization objective was to minimize root mean squared error for the `totalprotein` phenotype in UKB. The dataset was randomly split into train (80%) and test (20%) folds. The optimal parameters identified, and used for all 10 UKB phenotypes, were the following: `max_depth = 3`, `eta = 0.3190`, `alpha = 0.6577`, and `lambda = 2`.

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

Data availability

This work used genotyped and phenotypes from the UK Biobank study (<https://www.ukbiobank.ac.uk>) and our accessed was approved under application 65275. We utilized GWAS Catalog (<https://www.ebi.ac.uk/gwas/>) for replication analysis. Source data are provided with this paper.

Code availability

DeepNull software is available for download from GitHub (<https://github.com/google-health/genomics-research/tree/main/nonlinear-covariate-gwas>) or installation via PyPI (<https://pypi.org/project/deepnull/>). We used the following tools: BOLT-LMM (<https://data.broadinstitute.org/alkesgroup/bolt-lmm>), S-LDSC (<https://data.broadinstitute.org/alkesgroup/ldscore>), PLINK, scikit-learn (<https://scikit-learn.org/stable/>), TensorFlow (<https://www.tensorflow.org>), UCSC LiftOver (<https://genome.ucsc.edu/cgi-bin/hgLiftOver>), and XGBoost (<https://xgboost.readthedocs.io/en/latest/>).

Received: 18 May 2021; Accepted: 9 December 2021;

Published online: 11 January 2022

References

- Hakonarson, H. et al. A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature* **448**, 591–594 (2007).
- Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).
- International Multiple Sclerosis Genetics Consortium (IMSGC) et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.* **45**, 1353–1360 (2013).
- Ripke, S. et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* **45**, 1150–1159 (2013).
- Köttgen, A. et al. Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nat. Genet.* **45**, 145–154 (2013).
- Buniello, A. et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
- Claussnitzer, M. et al. FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* **373**, 895–907 (2015).
- 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).
- Imbens, G. W. & Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (Cambridge University Press, 2015) ISBN 0521885884.
- Marchini, J., Cardon, L. R., Phillips, M. S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nat. Genet.* **36**, 512–517 (2004).
- Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
- Shrine, N. et al. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat. Genet.* **51**, 481–493 (2019).
- Chen, H. et al. Multiethnic Meta-Analysis identifies RAI1 as a possible obstructive sleep apnea-related quantitative trait locus in men. *Am. J. Respir. Cell Mol. Biol.* **58**, 391–401 (2018).
- Kosmicki, J. A. et al. Genetic association analysis of SARS-CoV-2 infection in 455,838 UK biobank participants. *medRxiv* <https://doi.org/10.1101/2020.10.28.20221804> (2020).
- Bycroft, C. et al. The UK biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
- Leshno, M., Ya. Lin, V., Pinkus, A. & Schocken, S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Netw.* **6**, 861–867 (1993).
- Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **4**, 251–257 (1991).
- Loh, P.-R. et al. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
- Alipanahi, B. et al. Large-scale machine-learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology. *Am. J. Hum. Genet.* **108**, 1217–1230 (2021).
- Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
- Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
- Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
- Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
- Lehmann, B. C. L., Mackintosh, M., McVean, G. & Holmes, C. C. High trait variability in optimal polygenic prediction strategy within multiple-ancestry cohorts. *bioRxiv* <https://doi.org/10.1101/2021.01.15.426781> (2021).
- Visscher, P. M. et al. 10 years of gwas discovery: Biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
- Min Kang, H. et al. Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).
- Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
- Zhang, Z. et al. Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).
- Yang, J., Hong Lee, S., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
- Lippert, C. et al. FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8**, 833–835 (2011).
- Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
- Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014).
- Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).
- Scuteri, A. et al. Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet.* **3**, e115 (2007).
- Fusi, N., Lippert, C., Lawrence, N. D. & Stegle, O. Warped linear mixed models for the genetic analysis of transformed phenotypes. *Nat. Commun.* **5**, 4890 (2014).
- GTEX Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
- McCaw, Z. R., Lane, J. M., Saxena, R., Redline, S., & Lin, X. Operating characteristics of the rankbased inverse normal transformation for quantitative trait analysis in genomewide association studies. *Biometrics* **76**, 1262–1272 (2020).

38. GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
39. Eskin, E. Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. *Genome Res.* **18**, 653–660 (2008).
40. Darnell, G., Duong, D., Han, B. & Eskin, E. Incorporating prior information into association studies. *Bioinformatics* **28**, i147–i153 (2012).
41. Duong, D. et al. Using genomic annotations increases statistical power to detect egenes. *Bioinformatics* **32**, i156–i163 (2016).
42. Wen, X., Lee, Y., Luca, F. & Pique-Regi, R. Efficient integrative Multi-SNP association analysis via deterministic approximation of posteriors. *Am. J. Hum. Genet.* **98**, 1114–1129 (2016).
43. Wen, X. Molecular QTL discovery incorporating genomic annotations using Bayesian false discovery rate control. *Ann. Appl. Statistics* **10**, 1619–1638 (2016).
44. Kichaev, G. et al. Leveraging polygenic functional enrichment to improve GWAS power. *Am. J. Hum. Genet.* **104**, 65–75 (2019).
45. Hastie, T. J. & Tibshirani, R. J. *Generalized Additive Models* (Routledge, 1990).
46. Koza, J. R. *Genetic Programming: on the Programming of Computers by Means of Natural Selection* Vol. 1 (MIT Press, 1992).
47. Agarwal, R., Frosst, N., Zhang, X., Caruana, R., & Hinton, G. E. Neural additive models: interpretable machine learning with neural nets. In *NeurIPS 2021 proceedings* <https://proceedings.neurips.cc/paper/2021/file/251bd0442dfcc53b5a761e050f8022b8-Paper.pdf> (2021).
48. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inform. Process. Syst.* **31**, 4768–4777 (2017).
49. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning* (eds Precup, D. & Teh, Y. W.) 3145–3153 **70**, (PMLR, 2017).
50. Alaa, A. M. & van der Schaar, M. Demystifying black-box models with symbolic metamodelling. *Adv. Neural Inform. Process. Syst.* **32**, 11304–11314 (2019).
51. Crabbe, J., Zhang, Y., Zame, W. & van der Schaar, M. Learning outside the black-box: the pursuit of interpretable models. *Adv. Neural Inform. Process. Syst.* **33**, 17838–17849 (2020).
52. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
53. McCulloch, C. E., & Searle, S. R. *Generalized, Linear, and Mixed Models* (Wiley, 2000).
54. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463 (2010).
55. Sul, J. H. & Eskin, E. Mixed models can correct for population structure for genomic regions under selection. *Nat. Rev. Genet.* **14**, 300–300 (2013).
56. Tsang, M., Cheng, D., & Liu, Y. Detecting statistical interactions from neural network weights. In *International Conference on Learning Representations* (2018).
57. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778 (IEEE, 2016).
58. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In *International Conference for Learning Representations (ICLR) proceedings 2015* <https://arxiv.org/abs/1412.6980> (2015).
59. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9**, e1003348 (2013).
60. Sham, P. C. & Purcell, S. M. Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.* **15**, 335–346 (2014).
61. Eskin, E. Discovering genes involved in disease and the mystery of missing heritability. *Commun. ACM* **58**, 80–87 (2015).
62. Golovin, D. et al. Google vizier: a service for black-box optimization. In *Proc. 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2017). <https://doi.org/10.1145/3097983.3098043>.

Acknowledgements

This research has been conducted using the UK Biobank Resource application 65275. We are grateful to Alkes L. Price for helpful comments on the manuscript. We are extremely thankful for Babak Behsaz's contribution to our in-house GWAS pipeline and Justin Cosentino for insightful comments and discussion regarding neural network interpretability.

Author contributions

C.Y.M. and F.H. conceived the study. Z.R.M., B.A., C.Y.M., and F.H. designed the study. Z.R.M., T.C., T.Y., N.F., C.Y.M., and F.H. performed experiments. Z.R.M., T.C., T.Y., N.F., A.C., B.A., C.Y.M., and F.H. analyzed results. Z.R.M., C.Y.M., and F.H. wrote the manuscript. All authors contributed to the final version of the manuscript.

Competing interests

All authors are employees of Google LLC. This study was funded by Google LLC.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-27930-0>.

Correspondence and requests for materials should be addressed to Cory Y. McLean or Farhad Hormozdiari.

Peer review information Nature Communications thanks Nick Shrine and the other anonymous reviewer(s) for their contribution to the peer review this work. Peer reviewer reports are available.

Reprints and permission 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 license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license 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 license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022