bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

DeepNull: Modeling non-linear covariate effects improves phenotype prediction and association power

Zachary R. McCaw[1], Thomas Colthurst[2], Taedong Yun[2], Nicholas A. Furlotte[1], Andrew Carroll[1], Babak Alipanahi[1], Cory Y. McLean[2,*], Farhad Hormozdiari[2,*]

1 Google Health, Palo Alto, CA, USA

2 Google Health, Cambridge, MA, USA

These authors contributed equally: Zachary R. McCaw, Thomas Colthurst

These authors jointly supervised this work: Cory Y. McLean, Farhad Hormozdiari

* Corresponding author: cym@google.com, fhormoz@google.com

Abstract

Genome-wide association studies (GWAS) 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).

Introduction

GWAS aim to detect genetic variants or single-nucleotide polymorphisms (SNPs) that are associated with complex traits and diseases. Over the past decade, GWAS have successfully identified

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

thousands of variants associated with various and diverse phenotypes [1–6]. These associations have expanded our knowledge of biological mechanisms [7] and improved our ability to predict phenotypic risk [8].

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 $\texttt{age}^2$ as a covariate when studying chronic obstructive pulmonary disease; Chen *et al.* [13] included squared body mass index ($\texttt{BMI}^2$) when studying obstructive sleep apnea; and Kosmicki *et al.* [14] included an age by sex interaction ($\texttt{age} \times \texttt{sex}$) 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

can arise from a finite set of covariates (e.g. $\texttt{age} \times \texttt{sex}, \texttt{age}^2 \times \texttt{sex}, \cdots$), 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 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 (see URLs) 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

statistical power (Supplementary Figure 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 ith individual; let $\boldsymbol{G} = [g_{ij}]$ 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 $\bar{\boldsymbol{G}} = [\bar{g}_{ij}] \in \mathbb{R}^{n \times m}$ denote the standardized version of $\boldsymbol{G}$, 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:

$$Y = \bar{G}_{\cdot j}\beta_j + \tilde{\boldsymbol{X}}\gamma + H(\boldsymbol{X})\gamma_h + \varepsilon. \tag{1}$$

Here β_j is the effect sizes for the jth variant on the phenotype; $\tilde{\boldsymbol{X}} = [x_{ik}]$ 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 $H(\boldsymbol{X})\gamma_h$: $\boldsymbol{X}$ is the $n \times p$ subset of $\tilde{\boldsymbol{X}}$ consisting of non-genetic covariates (see Methods); $H : \mathbb{R}^{n \times p} \to \mathbb{R}^n$ is the function that applies h row-wise to $\boldsymbol{X}$; 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 Biobank [15]. 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 $\sigma_x^2 = 0.2$. Causal variants were randomly embedded within `chr22` and non-causal variants within `chr1` and `chr2`. We compared the DeepNull

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

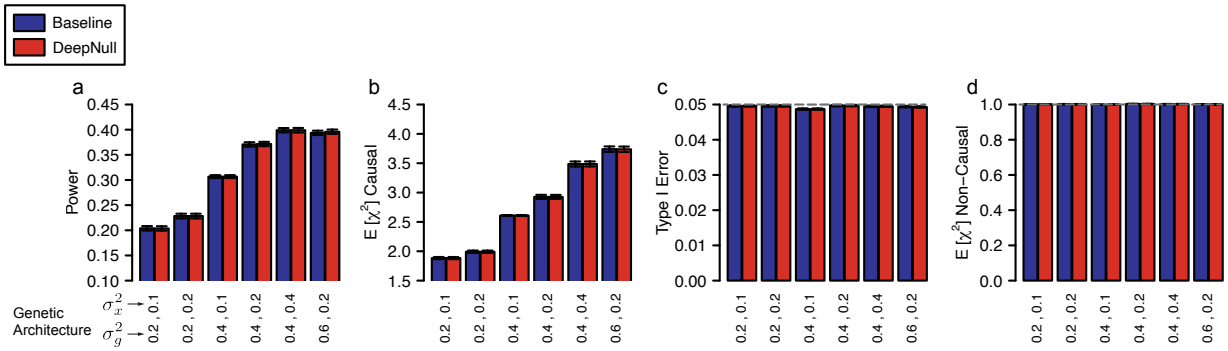


Figure 1: **DeepNull and baseline model achieve similar results under simulated linear covariate effects.** (a) Statistical power, (b) expected χ^2 statistics for variants in the causal chromosome (`chr22`), (c) type I error, and (d) expected χ^2 statistics for variants on the non-causal chromosomes (`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 (grey dashed horizontal line). The expected χ^2 statistics for the non-causal chromosomes should be 1 (grey 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. None of the quantities shown is significantly different between Baseline and DeepNull (Wilcoxon signed-rank test).

GWAS with standard GWAS (hereafter referred to as "Baseline"), each of which was performed using BOLT-LMM [18] (Methods). Statistical power and expected χ^2 statistics for the causal chromosome (`chr22`) were similar for DeepNull and Baseline (Figure 1a,b, Supplementary Table 2 and Supplementary Data 1). 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 (Figure 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`, `age`^{2}, `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 interac-

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

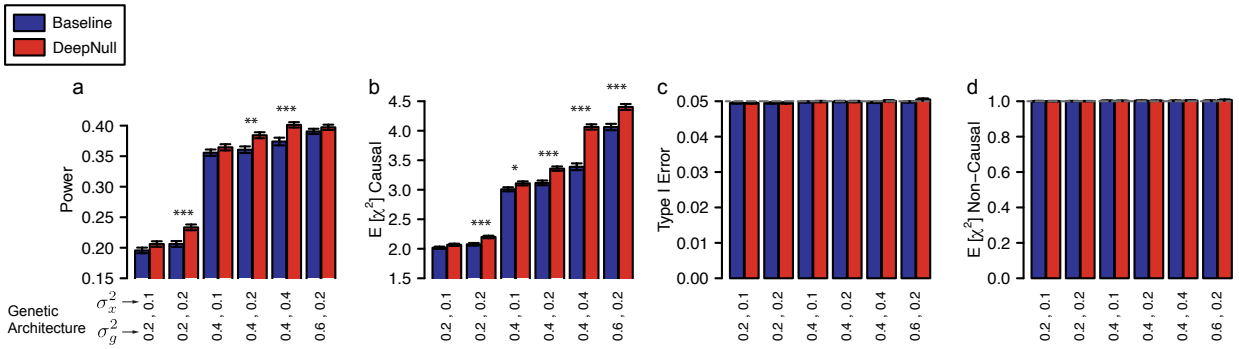


Figure 2: **DeepNull increases power in the presence of covariate interactions.** (a) Statistical power, (b) expected χ^2 statistics for variants in the causal chromosome (`chr22`), (c) type I Error, and (d) expected χ^2 statistics for variants in the non-causal chromosomes (`chr1` and `chr2`.). In the 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 grey horizontal line. For the non-causal chromosomes, the expected χ^2 statistics should be 1, which is also shown in dashed grey 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. The numerical results are shown in Supplementary Table 5. Indicators for P value (Wilcoxon signed-rank test) ranges: $^{*}P \leq 0.05$, $^{**}P \leq 0.01$, $^{***}P \leq 0.001$.

tions 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 (Figure 2a,b, Supplementary Table 5, and Supplementary Data 2). Importantly, both DeepNull and Baseline control the type I error and generate similar expected χ^2 statistics for non-causal variants (Figure 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`, `age`^{2}, `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 Equation (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 Figure 2, Supplementary Table 7).

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Finally, DeepNull's results are not affected by random seed initialization (Supplementary Notes; Supplementary Figure 3).

Pheno	n	**#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%

Table 1: **DeepNull improves association results relative to the Baseline model on ten phenotypes from the UK Biobank.** 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. 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), and TG (Triglycerides).

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 Figures 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: `VCDR visit`,

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

`refractive error`, and `image gradability`. To train DeepNull's DNN, we used `VCDR visit`, `age`, `sex`, and `genotyping_array` to predict GRP. We then performed GWAS for GRP using `age`, `sex`, `genotyping_array`, the top 15 PCs, `VCDR visit`, `refractive error`, and `image gradability` as covariates.

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 Figure 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; Table 1). In addition, 166 of the DeepNull loci were replicated in the GWAS catalog compared with 165 loci for Baseline (Supplementary Figure 15). For these three phenotypes, we further investigated the biological significance of the detected associations using FUMA [22] (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-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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(average $-\log_{10}$(p-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-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 Figures 16–25).

To further understand the source of the DeepNull improvements, we evaluated three additional Baseline models of increasing complexity. 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 `VCDR visit` 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}, `age` $\times$ `sex`, `age` $\times$ `genotyping_array`, and `sex` $\times$ `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. Lastly, 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 and 16 for hits and Supplementary Tables 15 and 17 for loci).

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 ($\texttt{PRS}_{\texttt{baseline}}$) and linear covariate effects ($\texttt{PRS}_{\texttt{baseline}}$+ Linear covariates). The "DeepNull-Baseline" model includes a PRS computed in the same way except using association results from DeepNull ($\texttt{PRS}_{\texttt{DeepNull}}$+ 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 (Figure 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=370K; 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 effect size estimates or inclusion of the DeepNull DNN prediction, we examined predictive performance of a model that uses `age`, `sex`, and $\texttt{PRS}_{\texttt{DeepNull}}$ ("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 $\texttt{age}^2$, $\texttt{age} \times \texttt{sex}$, $\texttt{age} \times \texttt{genotyping\_array}$). 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 and 18). Lastly, we compared phenotype prediction of DeepNull to an extended Baseline model

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

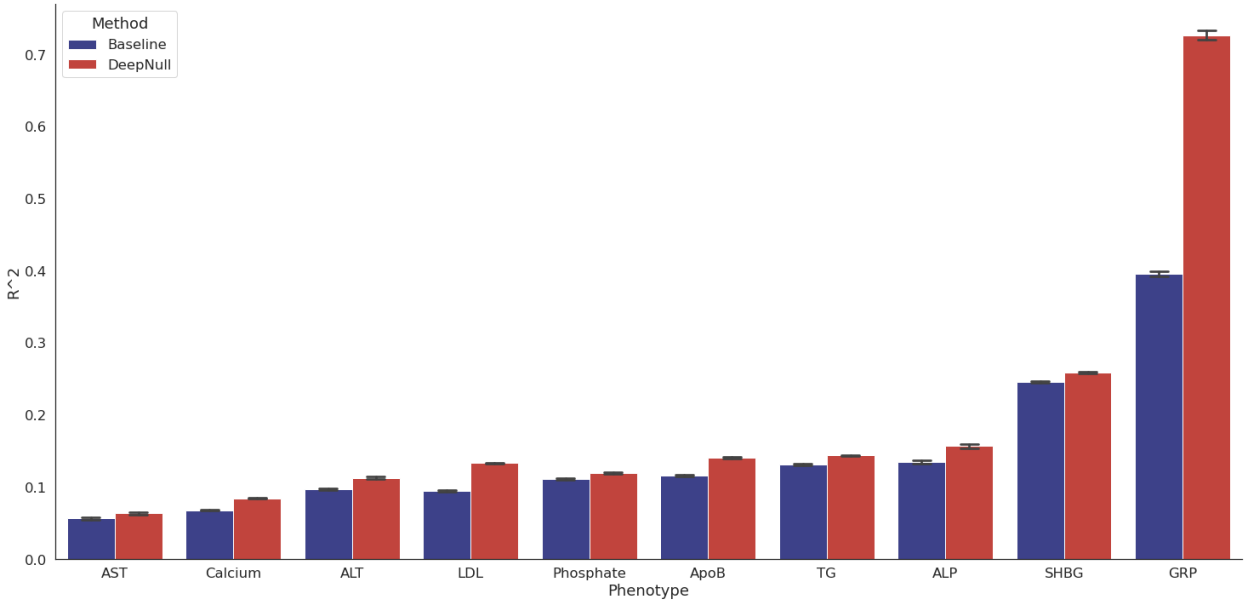


Figure 3: **DeepNull improves phenotype prediction compared to Baseline.** The X-axis is the phenotype names and the Y-axis is the R^2 where R is the Pearson's correlation between true and predicted value of phenotypes. The error bars indicate the standard error.

that performs sex-specific spline fitting (Methods) and observed that DeepNull outperforms this Baseline extension as well (compare Supplementary Tables 18 and 19).

DeepNull's covariates should remain in the association model.

When performing genetic association analysis via the model shown in Equation (1), the covariates X input row-wise to the DNN prediction function h are also included as components of the linear term $\tilde{X}\gamma$. 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

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

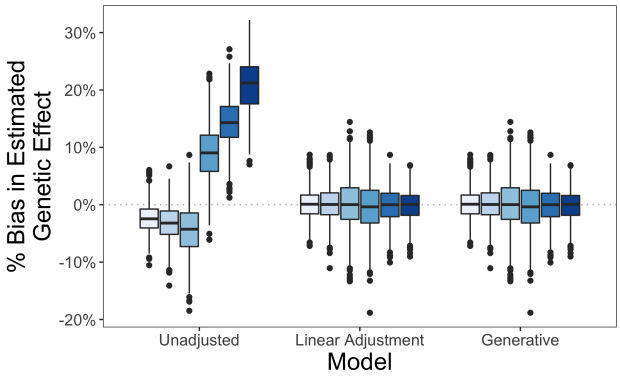


Figure 4: **Adjusting for covariates provided to DeepNull during association testing is necessary to avoid bias.** The unadjusted model regresses y_i on $\bar{g}_i$ and $h(x_i)$, the prediction of y_i based on x_i, omitting x_i from the association model. This approach results in biased estimation of 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.

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 Equation (1),

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

The adjusted model includes both $h(x_i)$ 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,

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 estimator remained unbiased in the presence of lower- and higher-order covariate effects (Supplementary Figures 26 and 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 covariate effect 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 Figure 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 size [25]. 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 GWAS including related individuals, who are not statistically independent

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[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, and thus 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 necessary 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 to 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, using a non-linear model such as spline regression or generalized additive model [45], symbolic regression [46], or neural additive model [47] may be beneficial. Alternatively, the trained DeepNull model can be interrogated with a variety of methods [48–51], though we note that DNN inter-

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

pretability is 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 boosted trees may equal or possibly outperform DeepNull's DNN. For example, we observed that XGBoost obtained similar GWAS hits, loci, and phenotypic predictions for the 10 example UKB phenotypes (Supplementary Tables 16, 17, and 20). 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 (see URLs). 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 $\boldsymbol{G} = [g_{ij}]$ 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 $\bar{\boldsymbol{G}} = [\bar{g}_{ij}] \in \mathbb{R}^{n \times m}$ be the standardized version of $\boldsymbol{G}$, 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:

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

where β is the $m \times 1$ vector of effect sizes for each variant on the phenotype, $\boldsymbol{X} = [x_{ik}]$ 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 $\boldsymbol{X}$ 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. $\varepsilon = (\varepsilon_i)_{i=1}^{n}$ 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:

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

Here $\tilde{\boldsymbol{X}}$ 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 Equation (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 $\tilde{\boldsymbol{X}}$ [10, 11, 52].

The model in Equation (5) can be simplified by projecting away the covariates [18, 53]. Define $\boldsymbol{P} = \boldsymbol{I} - \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T$, which is the projection onto the orthogonal complement of the linear

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

subspace spanned by $\boldsymbol{X}$. Multiplying Equation (5) through by $\boldsymbol{P}$ on the left yields:

$$\boldsymbol{P}Y = \boldsymbol{P}\bar{G}_{\cdot j}\beta_j + \varepsilon^*. \tag{6}$$

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

$$y_i = \bar{G}_{i\cdot}\beta + f(X_{i\cdot})\gamma_f + \varepsilon_i$$

where all variables are defined identically as in formula (4), $f : \mathbb{R}^q \to \mathbb{R}$ is any (potentially non-linear) function, $\bar{G}_{i\cdot} = (\bar{g}_{ij})_{j=1}^m$, and $X_{i\cdot} = (x_{ik})_{k=1}^q$. In vector form,

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

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

We convert the estimation of $u_i = f(X_{i\cdot})$ into a learning problem, where we predict u_i using y_i and $X_{i\cdot}$ as targets and input features, respectively. In other words, we train a model h using the

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

covariates $X_{i\cdot}$ and the phenotype y_i by minimizing

$$\|y_i - h(X_{i\cdot})\|^2. \tag{7}$$

We designed a DNN architecture for modeling the function h (Figure 5). We explored the model proposed previously to detect interpretable statistical interactions [56] 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 networks [57] and consists of two components. One component (the shorter path from input to output in Figure 5) is linear, to directly represent the linear effect of the covariates on the phenotype. The other component (the longer path in Figure 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 h can be written as

$$h(X_{i\cdot}) = H^{(5)} + H^{(6)},$$

where

$$\begin{aligned}
H^{(1)} &= \phi(\boldsymbol{W}^{(1)}_{64\times q} X_{i\cdot} + B^{(1)}_{64\times 1}) \\
H^{(2)} &= \phi(\boldsymbol{W}^{(2)}_{64\times 64} H^{(1)} + B^{(2)}_{64\times 1}) \\
H^{(3)} &= \phi(\boldsymbol{W}^{(3)}_{32\times 64} H^{(2)} + B^{(3)}_{32\times 1}) \\
H^{(4)} &= \phi(\boldsymbol{W}^{(4)}_{16\times 32} H^{(3)} + B^{(4)}_{16\times 1}) \\
H^{(5)} &= \boldsymbol{W}^{(5)}_{1\times 16} H^{(4)} + B^{(5)}_{1\times 1} \\
H^{(6)} &= \boldsymbol{W}^{(6)}_{1\times q} X_{i\cdot} + B^{(6)}_{1\times 1}
\end{aligned}$$

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

$$\phi\Big((x_p)_{p=1}^{P}\Big) = \Big(\max(0, x_p)\Big)_{p=1}^{P}.$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

DeepNull learns

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

and

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

by minimizing the mean squared error in (7) using the `Adam` optimizer [58] 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`).

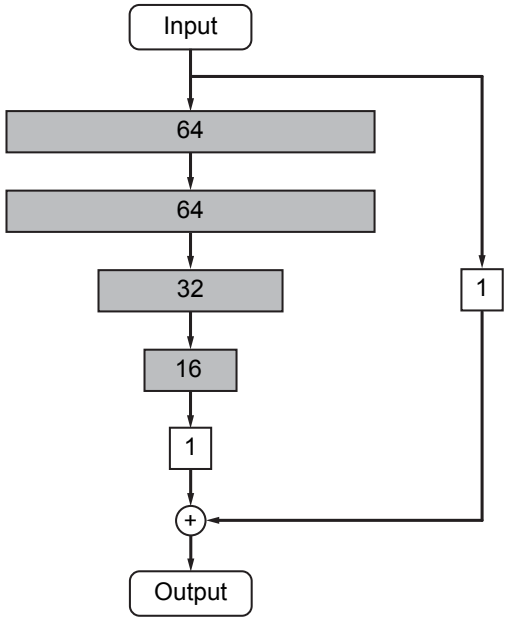


Figure 5: **DeepNull DNN model architecture.** Each rectangle represents one layer and all layers are fully connected. Shaded layers use the ReLU activation and the non-shaded layers do not use an activation function (i.e. linear connection). The input is the set of known covariates and the output is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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Performing GWAS using DeepNull.

After training DeepNull, we use the following model to test for association between the jth variant and the phenotype:

$$y_i = \bar{g}_{ij}\beta_j + h(X_{i.})\gamma_h + \tilde{X}_{i.}\gamma + \varepsilon.$$

The vectorized form of the above association test is

$$Y = \bar{G}_{.j}\beta_j + H(\boldsymbol{X})\gamma_h + \tilde{\boldsymbol{X}}\gamma + \varepsilon. \tag{8}$$

where $H : \mathbb{R}^{n\times q} \to \mathbb{R}^n$ is the function that applies h to each row of $\boldsymbol{X}$. Compared to the standard GWAS association model in Equation (5), the DeepNull association model differs only by the inclusion of an extra term $H(\boldsymbol{X})\gamma_h$, 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 Equation (5), $\tilde{\boldsymbol{X}}$ includes both non-genetic covariates (e.g. age and sex) and adjustments for confounding (e.g. genetic PCs) while $\boldsymbol{X}$ 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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Simulation framework.

We simulate data using the model

$$Y = \bar{\boldsymbol{G}}\beta + \sum_{k=1}^{q} f(X_{\cdot k})\gamma_k + \varepsilon \tag{9}$$

where $X_{\cdot k}$ 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 $j = 1, \cdots, m$, 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: $\beta_j \overset{iid}{\sim} \mathcal{N}(0, \frac{\sigma_g^2}{m})$. 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: $\gamma_k \overset{iid}{\sim} \mathcal{N}(0, \frac{\sigma_x^2}{q})$. Lastly, ε is drawn from another independent normal distribution with mean 0 and variance $1 - (\sigma_g^2 + \sigma_x^2)$: $\varepsilon \sim \mathcal{N}(0, 1 - \sigma_g^2 - \sigma_x^2)$. 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-LMM [33] 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 Equation (8) with the inclusion of an additional random effect $\bar{\boldsymbol{G}}^{(-j)}\delta$. Here $\bar{\boldsymbol{G}}^{(-j)}$ denotes genotype at all variants not on the same chromosome as variant j, and the components of δ follow the spike-and-slab prior [18].

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. α=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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

null hypothesis [59–61], 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 causal [59–61]. Additionally, the expected χ^2 statistic of an association method is a proxy for its prediction accuracy [18, 32, 33].

UKB GWAS evaluation.

All GWAS 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] (see URLs) 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 (see URLs). 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 (see URLs) 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 (see URLs) 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,

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

we learn an independent spline regression model based on the other non-genetic covariates. We utilized the python scikit-learn package (URLs) 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 (URLs) 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 Vizier [62]. The optimization objective was to minimize root mean squared error for the `total protein` 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`.

URLs

BOLT-LMM software: `https://data.broadinstitute.org/alkesgroup/bolt-lmm`

BaselineLD annotations: `https://data.broadinstitute.org/alkesgroup/ldscore`

DeepNull software: `https://github.com/google-health/genomics-research/tree/main/nonlinear-covariate-gwas`

GWAS Catalog: `https://www.ebi.ac.uk/gwas/`

PLINK software: `https://www.cog-genomics.org/plink1.9`

scikit-learn: `https://scikit-learn.org/stable/`

TensorFlow: `https://www.tensorflow.org`

UCSC LiftOver: `https://genome.ucsc.edu/cgi-bin/hgLiftOver`

UK Biobank study: `https://www.ukbiobank.ac.uk`

XGBoost: `https://xgboost.readthedocs.io/en/latest/`

Data Availability

This work used genotyped and phenotypes from the UK Biobank study (see URLs).

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Code Availability

DeepNull software is available for download from GitHub (see URLs) or installation via PyPI (`https://pypi.org/project/deepnull/`).

References

[1] Hakon Hakonarson, Struan F A Grant, Jonathan P Bradfield, Luc Marchand, Cecilia E Kim, Joseph T Glessner, Rosemarie Grabs, Tracy Casalunovo, Shayne P Taback, Edward C Frackelton, Margaret L Lawson, Luke J Robinson, Robert Skraban, Yang Lu, Rosetta M Chiavacci, Charles A Stanley, Susan E Kirsch, Eric F Rappaport, Jordan S Orange, Dimitri S Monos, Marcella Devoto, Hui-Qi Qu, and Constantin Polychronakos. A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature*, 448(7153):591–594, August 2007.

[2] D Altshuler, M J Daly, and E S Lander. Genetic mapping in human disease. *Science*, 322 (5903):881–888, 2008.

[3] International Multiple Sclerosis Genetics Consortium (IMSGC), Ashley H Beecham, Nikolaos A Patsopoulos, Dionysia K Xifara, Mary F Davis, Anu Kemppinen, Chris Cotsapas, Tejas S Shah, Chris Spencer, David Booth, An Goris, Annette Oturai, Janna Saarela, Bertrand Fontaine, Bernhard Hemmer, Claes Martin, Frauke Zipp, Sandra D'Alfonso, Filippo Martinelli-Boneschi, Bruce Taylor, Hanne F Harbo, Ingrid Kockum, Jan Hillert, Tomas Olsson, Maria Ban, Jorge R Oksenberg, Rogier Hintzen, Lisa F Barcellos, Wellcome Trust Case Control Consortium 2 (WTCCC2), International IBD Genetics Consortium (IIBDGC), Cristina Agliardi, Lars Alfredsson, Mehdi Alizadeh, Carl Anderson, Robert Andrews, Helle Bach Søndergaard, Amie Baker, Gavin Band, Sergio E Baranzini, Nadia Barizzone, Jeffrey Barrett, Céline Bellenguez, Laura Bergamaschi, Luisa Bernardinelli, Achim Berthele, Viola Biberacher, Thomas M C Binder, Hannah Blackburn, Izaura L Bomfim, Paola Brambilla, Simon Broadley, Bruno Brochet, Lou Brundin, Dorothea Buck, Helmut Butzkueven, Stacy J Caillier, William Camu, Wassila Carpentier, Paola Cavalla, Elisabeth G Celius, Irène Coman, Giancarlo Comi, Lucia Corrado, Leentje Cosemans, Isabelle Cournu-Rebeix, Bruce A C Cree, Daniele Cusi, Vincent Damotte, Gilles Defer, Silvia R Delgado, Panos Deloukas, Alessia di Sapio, Alexander T

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Dilthey, Peter Donnelly, Bénédicte Dubois, Martin Duddy, Sarah Edkins, Irina Elovaara, Federica Esposito, Nikos Evangelou, Barnaby Fiddes, Judith Field, Andre Franke, Colin Freeman, Irene Y Frohlich, Daniela Galimberti, Christian Gieger, Pierre-Antoine Gourraud, Christiane Graetz, Andrew Graham, Verena Grummel, Clara Guaschino, Athena Hadjixenofontos, Hakon Hakonarson, Christopher Halfpenny, Gillian Hall, Per Hall, Anders Hamsten, James Harley, Timothy Harrower, Clive Hawkins, Garrett Hellenthal, Charles Hillier, Jeremy Hobart, Muni Hoshi, Sarah E Hunt, Maja Jagodic, Ilijas Jelčić, Angela Jochim, Brian Kendall, Allan Kermode, Trevor Kilpatrick, Keijo Koivisto, Ioanna Konidari, Thomas Korn, Helena Kronsbein, Cordelia Langford, Malin Larsson, Mark Lathrop, Christine Lebrun-Frenay, Jeannette Lechner-Scott, Michelle H Lee, Maurizio A Leone, Virpi Leppä, Giuseppe Liberatore, Benedicte A Lie, Christina M Lill, Magdalena Lindén, Jenny Link, Felix Luessi, Jan Lycke, Fabio Macciardi, Satu Männistö, Clara P Manrique, Roland Martin, Vittorio Martinelli, Deborah Mason, Gordon Mazibrada, Cristin McCabe, Inger-Lise Mero, Julia Mescheriakova, Loukas Moutsianas, Kjell-Morten Myhr, Guy Nagels, Richard Nicholas, Petra Nilsson, Fredrik Piehl, Matti Pirinen, Siân E Price, Hong Quach, Mauri Reunanen, Wim Robberecht, Neil P Robertson, Mariaemma Rodegher, David Rog, Marco Salvetti, Nathalie C Schnetz-Boutaud, Finn Sellebjerg, Rebecca C Selter, Catherine Schaefer, Sandip Shaunak, Ling Shen, Simon Shields, Volker Siffrin, Mark Slee, Per Soelberg Sorensen, Melissa Sorosina, Mireia Sospedra, Anne Spurkland, Amy Strange, Emilie Sundqvist, Vincent Thijs, John Thorpe, Anna Ticca, Pentti Tienari, Cornelia van Duijn, Elizabeth M Visser, Steve Vucic, Helga Westerlind, James S Wiley, Alastair Wilkins, James F Wilson, Juliane Winkelmann, John Zajicek, Eva Zindler, Jonathan L Haines, Margaret A Pericak-Vance, Adrian J Ivinson, Graeme Stewart, David Hafler, Stephen L Hauser, Alastair Compston, Gil McVean, Philip De Jager, Stephen J Sawcer, and Jacob L McCauley. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.*, 45(11):1353–1360, November 2013.

[4] Stephan Ripke, Colm O'Dushlaine, Kimberly Chambert, Jennifer L Moran, Anna K Kähler, Susanne Akterin, Sarah E Bergen, Ann L Collins, James J Crowley, Menachem Fromer, Yunjung Kim, Sang Hong Lee, Patrik K E Magnusson, Nick Sanchez, Eli A Stahl, Stephanie Williams, Naomi R Wray, Kai Xia, Francesco Bettella, Anders D Borglum, Brendan K Bulik-

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Sullivan, Paul Cormican, Nick Craddock, Christiaan de Leeuw, Naser Durmishi, Michael Gill, Vera Golimbet, Marian L Hamshere, Peter Holmans, David M Hougaard, Kenneth S Kendler, Kuang Lin, Derek W Morris, Ole Mors, Preben B Mortensen, Benjamin M Neale, Francis A O'Neill, Michael J Owen, Milica Pejovic Milovancevic, Danielle Posthuma, John Powell, Alexander L Richards, Brien P Riley, Douglas Ruderfer, Dan Rujescu, Engilbert Sigurdsson, Teimuraz Silagadze, August B Smit, Hreinn Stefansson, Stacy Steinberg, Jaana Suvisaari, Sarah Tosato, Matthijs Verhage, James T Walters, Multicenter Genetic Studies of Schizophrenia Consortium, Douglas F Levinson, Pablo V Gejman, Kenneth S Kendler, Claudine Laurent, Bryan J Mowry, Michael C O'Donovan, Michael J Owen, Ann E Pulver, Brien P Riley, Sibylle G Schwab, Dieter B Wildenauer, Frank Dudbridge, Peter Holmans, Jianxin Shi, Margot Albus, Madeline Alexander, Dominique Campion, David Cohen, Dimitris Dikeos, Jubao Duan, Peter Eichhammer, Stephanie Godard, Mark Hansen, F Bernard Lerer, Kung-Yee Liang, Wolfgang Maier, Jacques Mallet, Deborah A Nertney, Gerald Nestadt, Nadine Norton, Francis A O'Neill, George N Papadimitriou, Robert Ribble, Alan R Sanders, Jeremy M Silverman, Dermot Walsh, Nigel M Williams, Brandon Wormley, Psychosis Endophenotypes International Consortium, Maria J Arranz, Steven Bakker, Stephan Bender, Elvira Bramon, David Collier, Benedicto Crespo-Facorro, Jeremy Hall, Conrad Iyegbe, Assen Jablensky, Rene S Kahn, Luba Kalaydjieva, Stephen Lawrie, Cathryn M Lewis, Kuang Lin, Don H Linszen, Ignacio Mata, Andrew McIntosh, Robin M Murray, Roel A Ophoff, John Powell, Dan Rujescu, Jim Van Os, Muriel Walshe, Matthias Weisbrod, Durk Wiersma, Wellcome Trust Case Control Consortium 2, Peter Donnelly, Ines Barroso, Jenefer M Blackwell, Elvira Bramon, Matthew A Brown, Juan P Casas, Aiden P Corvin, Panos Deloukas, Audrey Duncanson, Janusz Jankowski, Hugh S Markus, Christopher G Mathew, Colin N A Palmer, Robert Plomin, Anna Rautanen, Stephen J Sawcer, Richard C Trembath, Ananth C Viswanathan, Nicholas W Wood, Chris C A Spencer, Gavin Band, Céline Bellenguez, Colin Freeman, Garrett Hellenthal, Eleni Giannoulatou, Matti Pirinen, Richard D Pearson, Amy Strange, Zhan Su, Damjan Vukcevic, Peter Donnelly, Cordelia Langford, Sarah E Hunt, Sarah Edkins, Rhian Gwilliam, Hannah Blackburn, Suzannah J Bumpstead, Serge Dronov, Matthew Gillman, Emma Gray, Naomi Hammond, Alagurevathi Jayakumar, Owen T McCann, Jennifer Liddle, Simon C Potter, Radhi Ravindrarajah, Michelle Ricketts, Avazeh Tashakkori-Ghanbaria, Matthew J Waller, Paul Weston,

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Sara Widaa, Pamela Whittaker, Ines Barroso, Panos Deloukas, Christopher G Mathew, Jenefer M Blackwell, Matthew A Brown, Aiden P Corvin, Mark I McCarthy, Chris C A Spencer, Elvira Bramon, Aiden P Corvin, Michael C O'Donovan, Kari Stefansson, Edward Scolnick, Shaun Purcell, Steven A McCarroll, Pamela Sklar, Christina M Hultman, and Patrick F Sullivan. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.*, 45(10):1150–1159, October 2013.

[5] Anna Köttgen, Eva Albrecht, Alexander Teumer, Veronique Vitart, Jan Krumsiek, Claudia Hundertmark, Giorgio Pistis, Daniela Ruggiero, Conall M O'Seaghdha, Toomas Haller, Qiong Yang, Toshiko Tanaka, Andrew D Johnson, Zoltán Kutalik, Albert V Smith, Julia Shi, Maksim Struchalin, Rita P S Middelberg, Morris J Brown, Angelo L Gaffo, Nicola Pirastu, Guo Li, Caroline Hayward, Tatijana Zemunik, Jennifer Huffman, Loic Yengo, Jing Hua Zhao, Ayse Demirkan, Mary F Feitosa, Xuan Liu, Giovanni Malerba, Lorna M Lopez, Pim van der Harst, Xinzhong Li, Marcus E Kleber, Andrew A Hicks, Ilja M Nolte, Asa Johansson, Federico Murgia, Sarah H Wild, Stephan J L Bakker, John F Peden, Abbas Dehghan, Maristella Steri, Albert Tenesa, Vasiliki Lagou, Perttu Salo, Massimo Mangino, Lynda M Rose, Terho Lehtimäki, Owen M Woodward, Yukinori Okada, Adrienne Tin, Christian Müller, Christopher Oldmeadow, Margus Putku, Darina Czamara, Peter Kraft, Laura Frogheri, Gian Andri Thun, Anne Grotevendt, Gauti Kjartan Gislason, Tamara B Harris, Lenore J Launer, Patrick McArdle, Alan R Shuldiner, Eric Boerwinkle, Josef Coresh, Helena Schmidt, Michael Schallert, Nicholas G Martin, Grant W Montgomery, Michiaki Kubo, Yusuke Nakamura, Toshihiro Tanaka, Patricia B Munroe, Nilesh J Samani, David R Jacobs, Jr, Kiang Liu, Pio D'Adamo, Sheila Ulivi, Jerome I Rotter, Bruce M Psaty, Peter Vollenweider, Gerard Waeber, Susan Campbell, Olivier Devuyst, Pau Navarro, Ivana Kolcic, Nicholas Hastie, Beverley Balkau, Philippe Froguel, Tõnu Esko, Andres Salumets, Kay Tee Khaw, Claudia Langenberg, Nicholas J Wareham, Aaron Isaacs, Aldi Kraja, Qunyuan Zhang, Philipp S Wild, Rodney J Scott, Elizabeth G Holliday, Elin Org, Margus Viigimaa, Stefania Bandinelli, Jeffrey E Metter, Antonio Lupo, Elisabetta Trabetti, Rossella Sorice, Angela Döring, Eva Lattka, Konstantin Strauch, Fabian Theis, Melanie Waldenberger, H-Erich Wichmann, Gail Davies, Alan J Gow, Marcel Bruinenberg, LifeLines Cohort Study, Ronald P Stolk, Jaspal S Kooner,

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Weihua Zhang, Bernhard R Winkelmann, Bernhard O Boehm, Susanne Lucae, Brenda W Penninx, Johannes H Smit, Gary Curhan, Poorva Mudgal, Robert M Plenge, Laura Portas, Ivana Persico, Mirna Kirin, James F Wilson, Irene Mateo Leach, Wiek H van Gilst, Anuj Goel, Halit Ongen, Albert Hofman, Fernando Rivadeneira, Andre G Uitterlinden, Medea Imboden, Arnold von Eckardstein, Francesco Cucca, Ramaiah Nagaraja, Maria Grazia Piras, Matthias Nauck, Claudia Schurmann, Kathrin Budde, Florian Ernst, Susan M Farrington, Evropi Theodoratou, Inga Prokopenko, Michael Stumvoll, Antti Jula, Markus Perola, Veikko Salomaa, So-Youn Shin, Tim D Spector, Cinzia Sala, Paul M Ridker, Mika Kähönen, Jorma Viikari, Christian Hengstenberg, Christopher P Nelson, CARDIoGRAM Consortium, DIAGRAM Consortium, ICBP Consortium, MAGIC Consortium, James F Meschia, Michael A Nalls, Pankaj Sharma, Andrew B Singleton, Naoyuki Kamatani, Tanja Zeller, Michel Burnier, John Attia, Maris Laan, Norman Klopp, Hans L Hillege, Stefan Kloiber, Hyon Choi, Mario Pirastu, Silvia Tore, Nicole M Probst-Hensch, Henry Völzke, Vilmundur Gudnason, Afshin Parsa, Reinhold Schmidt, John B Whitfield, Myriam Fornage, Paolo Gasparini, David S Siscovick, Ozren Polašek, Harry Campbell, Igor Rudan, Nabila Bouatia-Naji, Andres Metspalu, Ruth J F Loos, Cornelia M van Duijn, Ingrid B Borecki, Luigi Ferrucci, Giovanni Gambaro, Ian J Deary, Bruce H R Wolffenbuttel, John C Chambers, Winfried März, Peter P Pramstaller, Harold Snieder, Ulf Gyllensten, Alan F Wright, Gerjan Navis, Hugh Watkins, Jacqueline C M Witteman, Serena Sanna, Sabine Schipf, Malcolm G Dunlop, Anke Tönjes, Samuli Ripatti, Nicole Soranzo, Daniela Toniolo, Daniel I Chasman, Olli Raitakari, W H Linda Kao, Marina Ciullo, Caroline S Fox, Mark Caulfield, Murielle Bochud, and Christian Gieger. Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nat. Genet.*, 45(2):145–154, February 2013.

[6] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, Daniel Suveges, Olga Vrousgou, Patricia L Whetzel, Ridwan Amode, Jose A Guillen, Harpreet S Riat, Stephen J Trevanion, Peggy Hall, Heather Junkins, Paul Flicek, Tony Burdett, Lucia A Hindorff, Fiona Cunningham, and Helen Parkinson. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

2019. *Nucleic Acids Res.*, 47(D1):D1005–D1012, January 2019.

[7] Melina Claussnitzer, Simon N Dankel, Kyoung-Han Kim, Gerald Quon, Wouter Meuleman, Christine Haugen, Viktoria Glunk, Isabel S Sousa, Jacqueline L Beaudry, Vijitha Puviindran, Nezar A Abdennur, Jannel Liu, Per-Arne Svensson, Yi-Hsiang Hsu, Daniel J Drucker, Gunnar Mellgren, Chi-Chung Hui, Hans Hauner, and Manolis Kellis. FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.*, 373(10):895–907, September 2015.

[8] Amit V Khera, Mark Chaffin, Krishna G Aragam, Mary E Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S Lander, Steven A Lubitz, Patrick T Ellinor, and Sekar Kathiresan. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.*, 50(9):1219–1224, September 2018.

[9] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, USA, 2015. ISBN 0521885884.

[10] Jonathan Marchini, Lon R Cardon, Michael S Phillips, and Peter Donnelly. The effects of human population structure on large genetic association studies. *Nat. Genet.*, 36(5):512–517, May 2004.

[11] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, 38(8):904–909, August 2006.

[12] Nick Shrine, Anna L Guyatt, A Mesut Erzurumluoglu, Victoria E Jackson, Brian D Hobbs, Carl A Melbourne, Chiara Batini, Katherine A Fawcett, Kijoung Song, Phuwanat Sakornsakolpat, Xingnan Li, Ruth Boxall, Nicola F Reeve, Ma'en Obeidat, Jing Hua Zhao, Matthias Wielscher, Stefan Weiss, Katherine A Kentistou, James P Cook, Benjamin B Sun, Jian Zhou, Jennie Hui, Stefan Karrasch, Medea Imboden, Sarah E Harris, Jonathan Marten, Stefan Enroth, Shona M Kerr, Ida Surakka, Veronique Vitart, Terho Lehtimäki, Richard J Allen, Per S Bakke, Terri H Beaty, Eugene R Bleecker, Yohan Bossé, Corry-Anke Brandsma, Zhengming Chen, James D Crapo, John Danesh, Dawn L DeMeo, Frank Dudbridge, Ralf Ewert, Christian Gieger, Amund Gulsvik, Anna L Hansell, Ke Hao, Joshua D Hoffman, John E Hokanson, Georg Homuth, Peter K Joshi, Philippe Joubert, Claudia Langenberg, Xuan Li, Liming

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Li, Kuang Lin, Lars Lind, Nicholas Locantore, Jian'an Luan, Anubha Mahajan, Joseph C Maranville, Alison Murray, David C Nickle, Richard Packer, Margaret M Parker, Megan L Paynton, David J Porteous, Dmitry Prokopenko, Dandi Qiao, Rajesh Rawal, Heiko Runz, Ian Sayers, Don D Sin, Blair H Smith, María Soler Artigas, David Sparrow, Ruth Tal-Singer, Paul R H J Timmers, Maarten Van den Berge, John C Whittaker, Prescott G Woodruff, Laura M Yerges-Armstrong, Olga G Troyanskaya, Olli T Raitakari, Mika Kähönen, Ozren Polašek, Ulf Gyllensten, Igor Rudan, Ian J Deary, Nicole M Probst-Hensch, Holger Schulz, Alan L James, James F Wilson, Beate Stubbe, Eleftheria Zeggini, Marjo-Riitta Jarvelin, Nick Wareham, Edwin K Silverman, Caroline Hayward, Andrew P Morris, Adam S Butterworth, Robert A Scott, Robin G Walters, Deborah A Meyers, Michael H Cho, David P Strachan, Ian P Hall, Martin D Tobin, Louise V Wain, and Understanding Society Scientific Group. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat. Genet.*, 51(3):481–493, March 2019.

[13] Han Chen, Brian E Cade, Kevin J Gleason, Andrew C Bjonnes, Adrienne M Stilp, Tamar Sofer, Matthew P Conomos, Sonia Ancoli-Israel, Raanan Arens, Ali Azarbarzin, Graeme I Bell, Jennifer E Below, Sung Chun, Daniel S Evans, Ralf Ewert, Alexis C Frazier-Wood, Sina A Gharib, José Haba-Rubio, Erika W Hagen, Raphael Heinzer, David R Hillman, W Craig Johnson, Zoltan Kutalik, Jacqueline M Lane, Emma K Larkin, Seung Ku Lee, Jingjing Liang, Jose S Loredo, Sutapa Mukherjee, Lyle J Palmer, George J Papanicolaou, Thomas Penzel, Paul E Peppard, Wendy S Post, Alberto R Ramos, Ken Rice, Jerome I Rotter, Scott A Sands, Neomi A Shah, Chol Shin, Katie L Stone, Beate Stubbe, Jae Hoon Sul, Mehdi Tafti, Kent D Taylor, Alexander Teumer, Timothy A Thornton, Gregory J Tranah, Chaolong Wang, Heming Wang, Simon C Warby, D Andrew Wellman, Phyllis C Zee, Craig L Hanis, Cathy C Laurie, Daniel J Gottlieb, Sanjay R Patel, Xiaofeng Zhu, Shamil R Sunyaev, Richa Saxena, Xihong Lin, and Susan Redline. Multiethnic Meta-Analysis identifies RAI1 as a possible obstructive sleep apnea-related quantitative trait locus in men. *Am. J. Respir. Cell Mol. Biol.*, 58(3): 391–401, March 2018.

[14] J A Kosmicki, J E Horowitz, N Banerjee, R Lanche, A Marcketta, E Maxwell, Xiaodong Bai, D Sun, J Backman, D Sharma, C O'Dushlaine, A Yadav, A J Mansfield, A Li, J Mbatchou,

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

K Watanabe, L Gurski, S McCarthy, A Locke, S Khalid, O Chazara, Y Huang, E Kvikstad, A Nadkar, A O'Neill, P Nioi, M M Parker, S Petrovski, H Runz, J D Szustakowski, Q Wang, Regeneron Genetics Center, UKB Exome Sequencing Consortium, M Jones, S Balasubramanian, W Salerno, A Shuldiner, J Marchini, J Overton, L Habegger, M N Cantor, J Reid, A Baras, G R Abecasis, and M A Ferreira. Genetic association analysis of SARS-CoV-2 infection in 455,838 UK biobank participants. November 2020.

[15] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, October 2018.

[16] Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, January 1993. doi: 10.1016/s0893-6080(05)80131-5. URL `https://doi.org/10.1016/s0893-6080(05)80131-5`.

[17] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991. doi: 10.1016/0893-6080(91)90009-t. URL `https://doi.org/10.1016/0893-6080(91)90009-t`.

[18] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjálmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, Nick Patterson, and Alkes L Price. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.*, 47(3):284–290, March 2015.

[19] Babak Alipanahi, Farhad Hormozdiari, Babak Behsaz, Justin Cosentino, Zachary R. McCaw, Emanuel Schorsch, D. Sculley, Elizabeth H. Dorfman, Paul J. Foster, Lily H. Peng, Sonia Phene, Naama Hammel, Andrew Carroll, Anthony P. Khawaja, and Cory Y. McLean. Large-scale machine-learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology. *The American Journal of Human Genetics*, 108(7):1217–1230, July

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

2021. doi: 10.1016/j.ajhg.2021.05.004. URL https://doi.org/10.1016/j.ajhg.2021.05.004.

[20] Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Laramie Duncan, John R B Perry, Nick Patterson, Elise B Robinson, Mark J Daly, Alkes L Price, and Benjamin M Neale. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.*, 47(11): 1236–1241, November 2015.

[21] Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, Stephan Ripke, Felix R Day, ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, RACI Consortium, Shaun Purcell, Eli Stahl, Sara Lindstrom, John R B Perry, Yukinori Okada, Soumya Raychaudhuri, Mark J Daly, Nick Patterson, Benjamin M Neale, and Alkes L Price. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.*, 47(11):1228–1235, November 2015.

[22] K Watanabe, E Taskesen, A van Bochoven, and D Posthuma. Functional mapping and annotation of genetic associations with fuma. 8(1):1826, 11 2017. doi: 10.1038/s41467-017-01261-5.

[23] Alicia R. Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51(4):584–591, March 2019. doi: 10.1038/s41588-019-0379-x. URL https://doi.org/10.1038/s41588-019-0379-x.

[24] B.C.L. Lehmann, M. Mackintosh, G. McVean, and C.C. Holmes. High trait variability in optimal polygenic prediction strategy within multiple-ancestry cohorts. January 2021. doi: 10.1101/2021.01.15.426781. URL https://doi.org/10.1101/2021.01.15.426781.

[25] PM Visscher, NR Wray, Q Zhang, P Sklar, MI McCarthy, MA Brown, and J Yang. 10 years of gwas discovery: Biology, function, and translation. *Am J Hum Genet*, 101(1):5–22, 2017.

[26] Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, 2008.

[27] Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-Yee Kong, Nelson B Freimer, Chiara Sabatti, and Eleazar Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354, 2010.

[28] Zhiwu Zhang, Elhan Ersoz, Chao-Qiang Lai, Rory J Todhunter, Hemant K Tiwari, Michael A Gore, Peter J Bradbury, Jianming Yu, Donna K Arnett, Jose M Ordovas, and Edward S Buckler. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, 42(4):355–360, 2010.

[29] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, 88(1):76–82, January 2011.

[30] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835, 2011.

[31] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7):821–824, 2012.

[32] Jian Yang, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.*, 46 (2):100–106, February 2014.

[33] Po-Ru Loh, Gleb Kichaev, Steven Gazal, Armin P Schoech, and Alkes L Price. Mixed-model association for biobank-scale datasets. *Nat. Genet.*, 50(7):906–908, July 2018.

[34] Angelo Scuteri, Serena Sanna, Wei-Min Chen, Manuela Uda, Giuseppe Albai, James Strait, Samer Najjar, Ramaiah Nagaraja, Marco Orrú, Gianluca Usala, Mariano Dei, Sandra Lai, Andrea Maschio, Fabio Busonero, Antonella Mulas, Georg B Ehret, Ashley A Fink, Alan B Weder, Richard S Cooper, Pilar Galan, Aravinda Chakravarti, David Schlessinger, Antonio Cao, Edward Lakatta, and Gonçalo R Abecasis. Genome-wide association scan shows genetic

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

variants in the FTO gene are associated with obesity-related traits. *PLoS Genet.*, 3(7):e115, July 2007.

[35] Nicolo Fusi, Christoph Lippert, Neil D Lawrence, and Oliver Stegle. Warped linear mixed models for the genetic analysis of transformed phenotypes. *Nature Communications*, 5(1), 2014.

[36] GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*, 550 (7675):204–213, 2017.

[37] Zachary R McCaw, Jacqueline M Lane, Richa Saxena, Susan Redline, and Xihong Lin. Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics*, 76(4):1262–1272, 2020.

[38] GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, September 2020.

[39] Eleazar Eskin. Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. *Genome Res.*, 18(4):653–660, April 2008.

[40] G Darnell, D Duong, B Han, and E Eskin. Incorporating prior information into association studies. *Bioinformatics*, 28(12):i147–i153, 2012.

[41] Dat Duong, Jennifer Zou, Farhad Hormozdiari, Jae Hoon Sul, Jason Ernst, Buhm Han, and Eleazar Eskin. Using genomic annotations increases statistical power to detect egenes. *Bioinformatics*, 32(12):i156–i163, 2016.

[42] Xiaoquan Wen, Yeji Lee, Francesca Luca, and Roger Pique-Regi. Efficient integrative Multi-SNP association analysis via deterministic approximation of posteriors. *The American Journal of Human Genetics*, 98(6):1114–1129, 2016.

[43] Xiaoquan Wen. Molecular QTL discovery incorporating genomic annotations using bayesian false discovery rate control. *The Annals of Applied Statistics*, 10(3):1619–1638, 2016.

[44] Gleb Kichaev, Gaurav Bhatia, Po-Ru Loh, Steven Gazal, Kathryn Burch, Malika K Freund, Armin Schoech, Bogdan Pasaniuc, and Alkes L Price. Leveraging polygenic functional en-

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

richment to improve GWAS power. *The American Journal of Human Genetics*, 104(1):65–75, 2019.

[45] Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*. Routledge, 1990.

[46] John R Koza. *Genetic programming: on the programming of computers by means of natural selection*, volume 1. MIT press, 1992.

[47] Rishabh Agarwal, Nicholas Frosst, Xuezhou Zhang, Rich Caruana, and Geoffrey E Hinton. Neural additive models: Interpretable machine learning with neural nets. *arXiv*, 2020.

[48] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 31:4768–4777, 2017.

[49] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017.

[50] Ahmed M Alaa and Mihaela van der Schaar. Demystifying black-box models with symbolic metamodels. *Advances in Neural Information Processing Systems*, 32:11304–11314, 2019.

[51] Jonathan Crabbe, Yao Zhang, William Zame, and Mihaela van der Schaar. Learning outside the black-box: The pursuit of interpretable models. *Advances in Neural Information Processing Systems*, 33:17838–17849, 2020.

[52] Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS Genet.*, 2(12):e190, December 2006.

[53] Charles E. McCulloch and Shayle R. Searle. *Generalized, Linear, and Mixed Models*. Wiley, December 2000. doi: 10.1002/0471722073. URL `https://doi.org/10.1002/0471722073`.

[54] Alkes L. Price, Noah A. Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, June 2010. doi: 10.1038/nrg2813. URL `https://doi.org/10.1038/nrg2813`.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[55] Jae Hoon Sul and Eleazar Eskin. Mixed models can correct for population structure for genomic regions under selection. *Nature Reviews Genetics*, 14(4):300–300, February 2013. doi: 10.1038/nrg2813-c1. URL `https://doi.org/10.1038/nrg2813-c1`.

[56] Michael Tsang, Dehua Cheng, and Yan Liu. Detecting statistical interactions from neural network weights. In *International Conference on Learning Representations*, 2018.

[57] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[58] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, December 2014.

[59] Frank Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.*, 9(3): e1003348, March 2013.

[60] Pak C Sham and Shaun M Purcell. Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.*, 15(5):335–346, May 2014.

[61] Eleazar Eskin. Discovering genes involved in disease and the mystery of missing heritability. *Commun. ACM*, 58(10):80–87, September 2015.

[62] Daniel Golovin, Benjamin Solnik, Subhodeep Moitra, Greg Kochanski, John Karro, and D. Sculley. Google vizier: A service for black-box optimization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, August 2017. doi: 10.1145/3097983.3098043. URL `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.

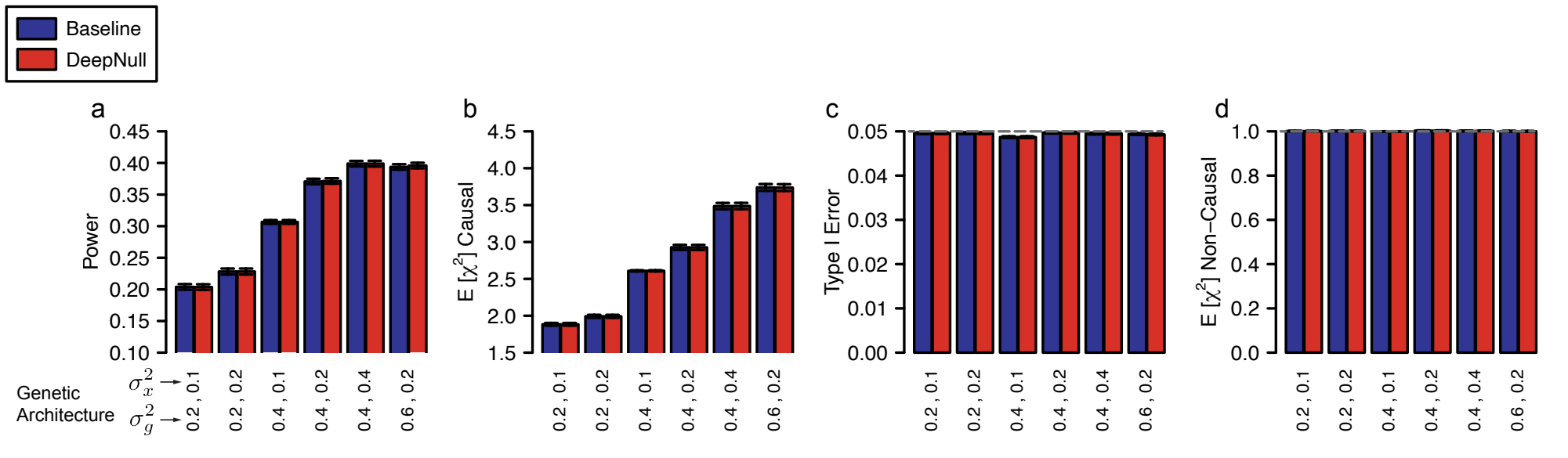
bioRxiv preprint doi: https://doi.org/10.1101/2021.05.26.445783; this version posted September 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

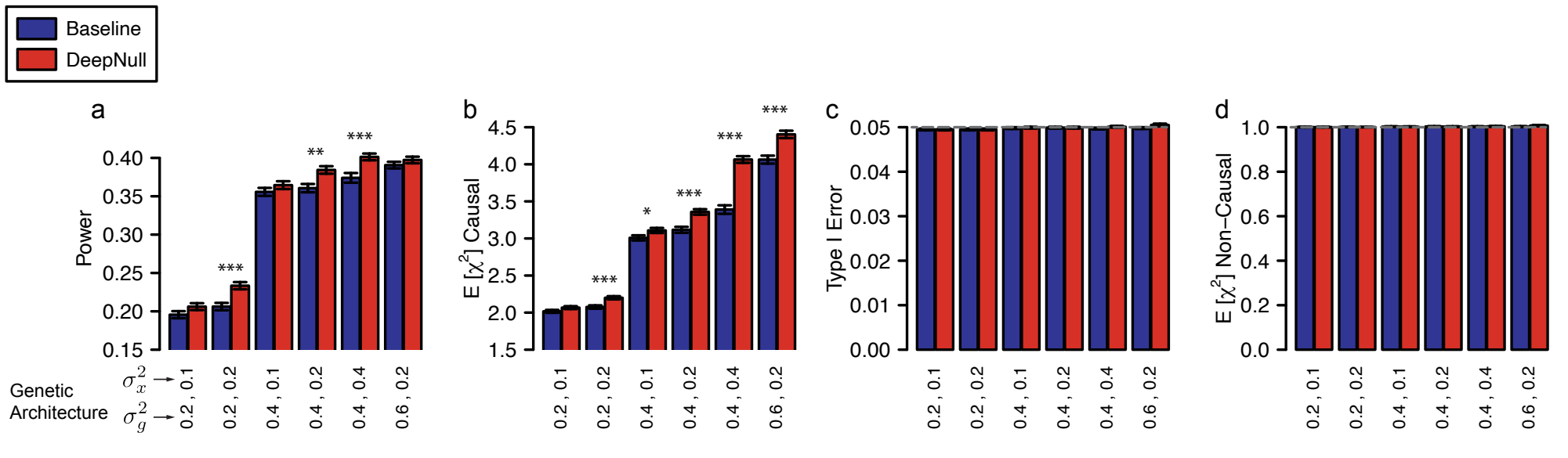
Author Contributions

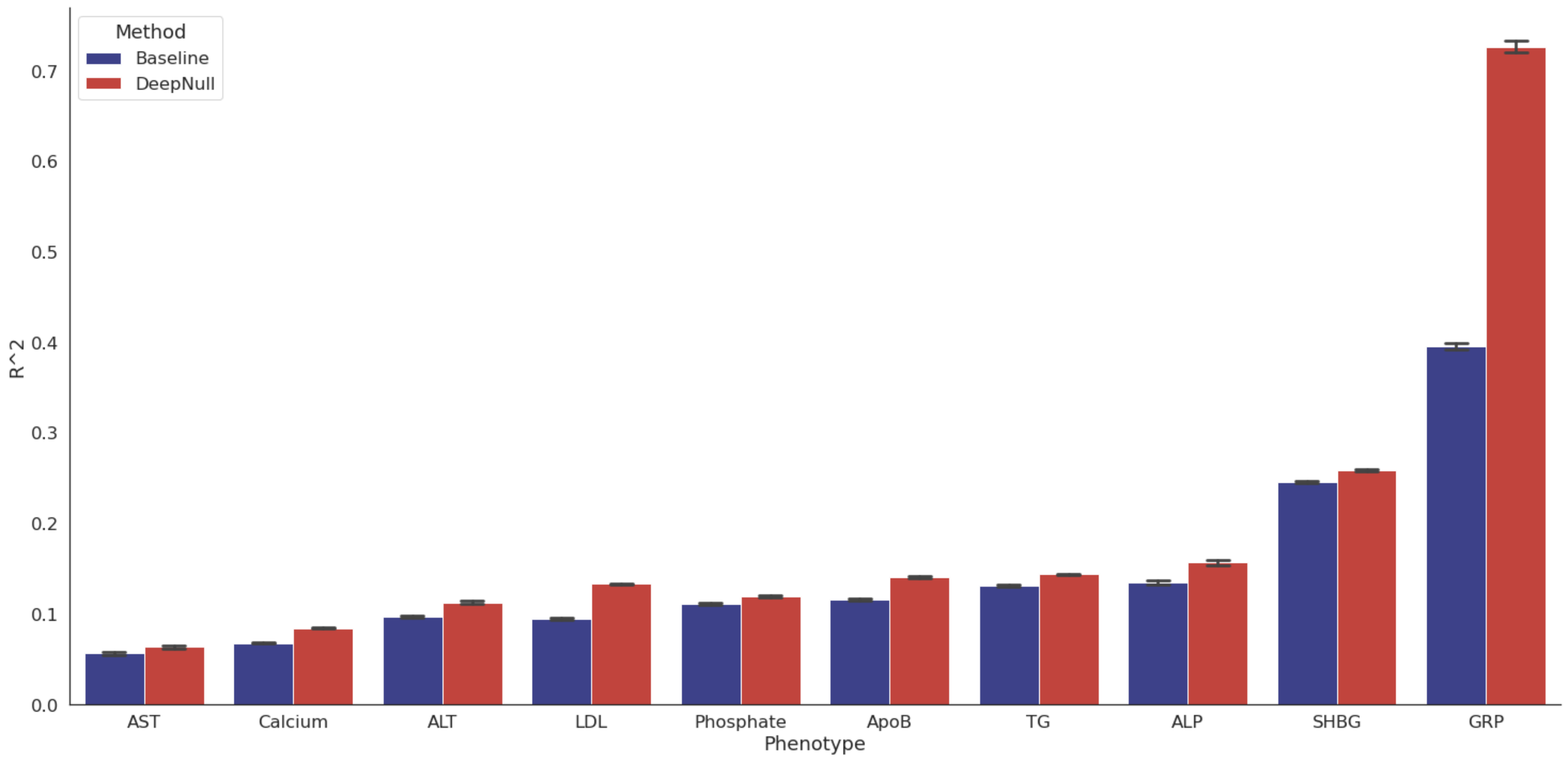
CYM and FH conceived the study. ZRM, BA, CYM, and FH designed the study. ZRM, TC, TY, NF, CYM, and FH performed experiments. ZRM, TC, TY, NF, AC, BA, CYM, and FH analyzed results. ZRM, CYM, and FH wrote the manuscript. All authors contributed to the final version of the manuscript.

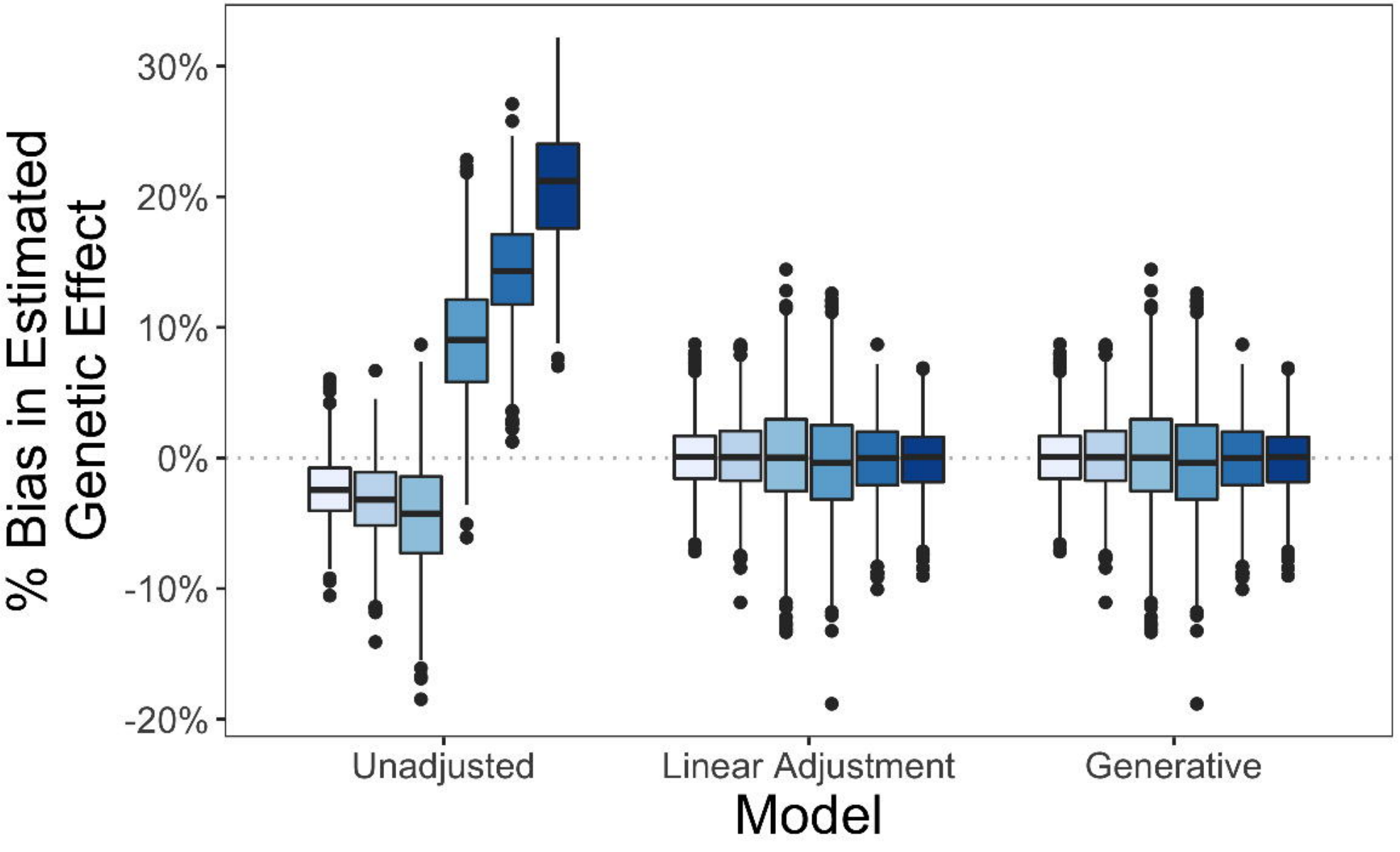
Competing Interests

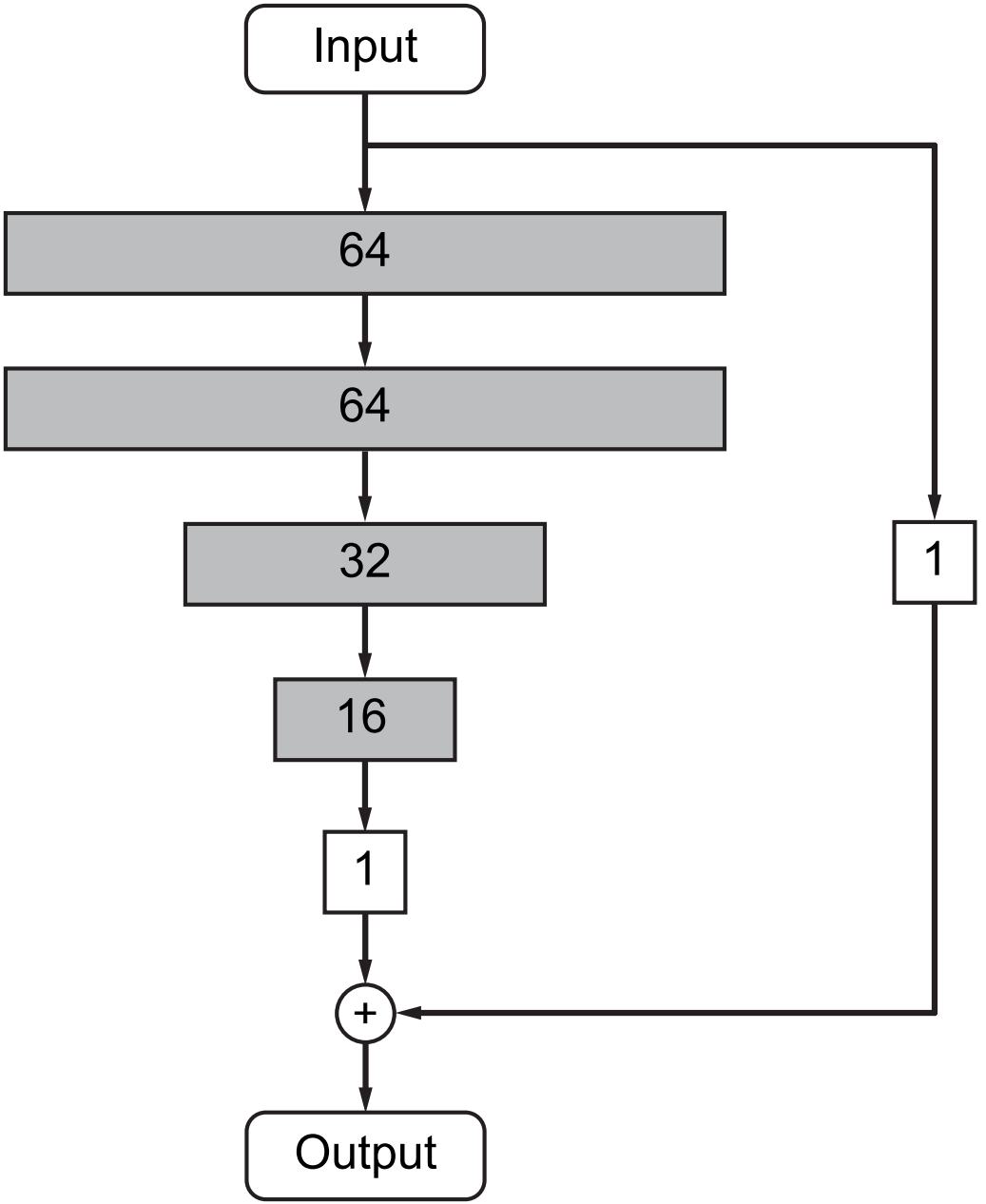
All authors are employees of Google LLC. This study was funded by Google LLC.


Baseline
DeepNull
a
Power
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
b
E [χ^2] Causal
4.5
4.0
3.5
3.0
2.5
2.0
1.5
c
Type I Error
0.05
0.04
0.03
0.02
0.01
0.00
d
E [χ^2] Non–Causal
1.0
0.8
0.6
0.4
0.2
0.0
Genetic Architecture
$\sigma_x^2 \rightarrow$
$\sigma_g^2 \rightarrow$
0.2 , 0.1
0.2 , 0.2
0.4 , 0.1
0.4 , 0.2
0.4 , 0.4
0.6 , 0.2


Baseline
DeepNull
a
Power
0.40
0.35
0.30
0.25
0.20
0.15
***
**
***
b
E [χ^2] Causal
4.5
4.0
3.5
3.0
2.5
2.0
1.5
***
*
***
***
***
c
Type I Error
0.05
0.04
0.03
0.02
0.01
0.00
d
E [χ^2] Non−Causal
1.0
0.8
0.6
0.4
0.2
0.0
Genetic Architecture
$\sigma_x^2 \rightarrow$
$\sigma_g^2 \rightarrow$
0.2 , 0.1
0.2 , 0.2
0.4 , 0.1
0.4 , 0.2
0.4 , 0.4
0.6 , 0.2


Method
Baseline
DeepNull
R^2
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
AST
Calcium
ALT
LDL
Phosphate
ApoB
TG
ALP
SHBG
GRP
Phenotype


% Bias in Estimated Genetic Effect
30%
20%
10%
0%
-10%
-20%
Unadjusted
Linear Adjustment
Generative
Model


Input
64
64
32
16
1
1
+
Output