

Deep Learning for Polygenic Risk Prediction

Costa Georgantas^{1,2}, Zoltán Kutalik², and Jonas Richiardi^{1,2}

¹Lausanne University Hospital, Lausanne, CH

²University of Lausanne, Lausanne, CH

Polygenic risk scores (PRS) are relative measures of an individual's genetic propensity to a particular trait or disease. Most PRS methods assume that mutation effects scale linearly with the number of alleles and are constant across individuals. While these assumptions simplify computation, they increase error, particularly for less-represented racial groups. We developed and provide Delphi (deep learning for phenotype inference), a deep-learning method that relaxes these assumptions to produce more predictive PRS. In contrast to other methods, Delphi can integrate up to hundreds of thousands of SNPs as input. We compare our results to a standard, linear PRS model, lasso regression, and a gradient-boosted trees-based method. We show that deep learning can be an effective approach to genetic risk prediction. We report a relative increase in the percentage variance explained compared to the state-of-the-art by 11.4% for body mass index, 18.9% for systolic blood pressure, 7.5% for LDL, 35% for C-reactive protein, 16.2% for height, 29.6 % for pulse rate; in addition, Delphi provides 2% absolute explained variance for blood glucose while other tested methods were non-predictive. Furthermore, we show that Delphi tends to increase the weight of high-effect mutations. This work demonstrates an effective deep learning method for modeling genetic risk that also showed to generalize well when evaluated on individuals from non-European ancestries.

Correspondence: costa.georgantas@chuv.ch

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

21 Introduction

22 The total genetic component of common traits and diseases is attributable, at least in part, to a combination of small
23 effects from a large number of mutations on the entire genome (1). Genome-wide association studies (GWAS) can
24 identify univariate relationships between common single nucleotide polymorphisms (SNPs) and a given trait. A GWAS
25 output comprises an estimated effect size coupled with a P-value of association for each tested SNP. A single scalar
26 indicating relative genetic risk can be obtained by summing up the number of alleles weighted by the estimated
27 effect size of SNPs, with or without non-genetic risk factors (2). These so-called polygenic risk scores (PRS) are
28 commonly used to quantify an individual's genetic propensity for a particular trait or disease and have potential clinical
29 applications in prevention, diagnosis, and treatment (3; 4; 5).

30 Methods for PRS estimation have evolved considerably over the past decade. It was first found that including mutations
31 below GWAS statistical significance would increase predictive power (6; 7). Taking linkage disequilibrium (LD) into
32 account by either clumping and thresholding (C+T) or by using a shrinkage method also improved performance (8; 9).
33 More recent work has included advancements in statistical learning and an improved understanding of biology to
34 increase the predictive performance of PRS. For instance, Bayesian approaches can also consider minor allele
35 frequency (MAF) (10) or incorporate functional priors (11) to modify the effect size estimates. These different methods
36 generally offer marginal improvements over one another and suffer from similar limitations: effects are constant and
37 scale linearly with the number of alleles.

38 PRS typically become significantly less predictive when applied to other less represented ancestries (12). This
39 performance drop can be partly attributed to allele frequency differences between cohorts and other genetic and
40 environmental factors. These limitations hinder the application of PRS in medical settings (13), and this performance
41 gap can only be bridged with additional data collection from under-represented ancestries. Multiple approaches have
42 been proposed to increase the generalizability of PRS, for instance, by aggregating results from multiple GWAS
43 studies (14; 15) or prioritizing functional variants (16). Recently, increased prediction performance was observed
44 through the use of a gradient-boosted model taking a standard PRS and a selection of high-impact SNPs as inputs (17).
45 Very recently, GenoBoost (18), another gradient-boosted approach, showed improved performance by modeling
46 non-additive mutation effects.

47 Deep learning (DL) offers the ability to learn complex patterns directly from large labeled datasets with minimal
48 assumptions. In genetics, DL has been applied for many problems such as variant calling (19), motif discovery (20),
49 and image-derived phenotyping for GWAS (21; 22). Explainable DL approaches (23) could provide additional insight
50 into the genetic factors influencing the disease. Abe et al. recently constructed a knowledge graph (24) to generate
51 text-based explanations for individual variants. Using deep learning for genetic risk prediction could provide unique
52 advantages, as overparametrization has recently been shown to improve generalization (25), which is important for
53 PRS to be applicable in under-represented populations.

54 Using DL for PRS estimation has been attempted before, although the proposed approaches consisted in using shallow
55 networks (max. 4 fully connected or convolutional layers) on a small set of SNPs (max. 5K) (26; 27; 28; 29; 30; 31). In
56 those examples, DL was shown only marginally to improve results, if at all. For instance, Badré et al. (29) found that
57 including 5273 high-impact SNPs in a deep neural network slightly improved the predictive performance of PRS for
58 breast cancer over logistic regression, and including more SNPs did not improve performance. Zhou et al. (31) showed
59 that a small neural network with three fully connected layers improved Alzheimer's disease genetic risk prediction in a
60 small ($N \approx 10K$) cohort.

61 In this work, we propose Delphi (deep learning for phenotype inference), a deep learning method that alleviates
62 some of the issues of PRS mentioned above by tuning risk score estimates in a data-driven and hypothesis-free
63 manner. In contrast to previous methods, we use a transformer architecture to capture non-linear interactions. Unlike
64 other approaches, we modify effect sizes before the summation, allowing allele effects to depend on sex, , and other
65 mutations. Our method can fine-tune effects from any classical PRS method such as LDpred (8) and Lassosum (9).
66 We report state-of-the-art results for 5 phenotypes from the UK Biobank dataset (UKBB) (32), and show that Delphi
67 tends to increase the estimated effect of high-impact mutations. We also validate our predictions on individuals from
68 under-represented ancestries and show that Delphi generalizes better than other tested approaches.

69 Results

70 The Delphi Method

71 At a high level, Delphi (Figure 1) uses genotyping and covariate information to learn perturbations of mutation effect
 72 estimates. Our approach contained two main steps. (1) the dataset was split into training, validation, and test sets
 73 before pre-processing. Mutation effect sizes were estimated with standard PRS techniques, and genotyping data
 74 was converted into a format enabling fast loading during training. (2) In the training step, a covariate model based
 75 on gradient-boosted trees (33) estimated the phenotype from age, sex, and genetic principal components, and a
 76 deep neural network learned to perturb individual effect sizes for all mutations included in the PRS summation. The
 77 modified effect sizes were then summed up to form a personalized PRS. The covariate model outputs and the PRS
 78 summation were finally linearly combined to form the final prediction.

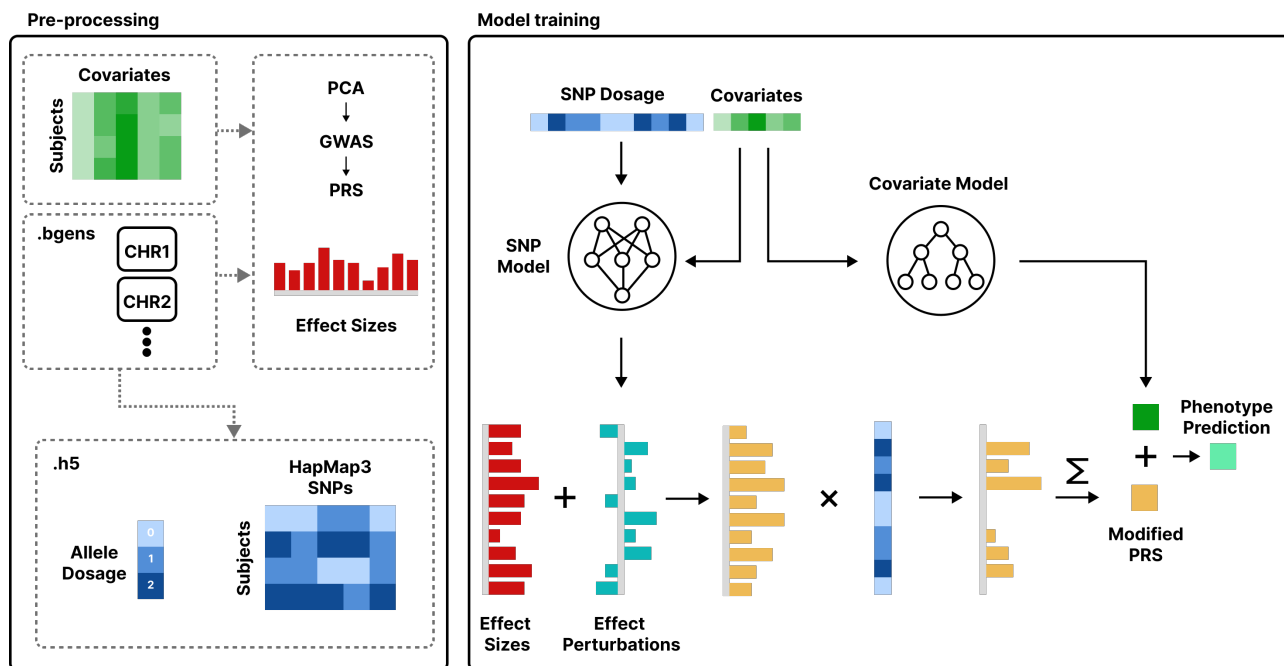


Figure 1. Overview of the Delphi method. The data is split into training, validation, and test sets before pre-processing. A GWAS is conducted for the phenotype in question on the training set, followed by a PRS method. A transformer neural network learns to modify the effect size estimates during training depending on other SNP dosages and covariates on the training data set. Model selection for the neural network and the PRS methods is done using the validation set. Modified effect sizes are summed up and aggregated with the predictions of a boosted trees covariate model to form a new PRS. Prediction results are all evaluated on the held-out test set.

79 GWAS and PRS

80 485'231 UKBB subjects were randomly split into different sets. The training set was used for principal component
 81 analysis (PCA), GWAS, PRS computation, and deep neural network training. The validation set was used for PRS
 82 validation and model selection after training. The held-out test set remained unseen until the final evaluation. We only
 83 considered 1.3M SNPs from the HapMap3 set (34) with an INFO score > 0.8 and MAF > 0.01 . PCA on the genotype
 84 matrix was used to capture population structure.

85 GWAS for all phenotypes only included subjects within the training set from British-white ancestry (UKBB field 22006)
 86 to reduce spurious associations, and any subjects further than three standard deviations away from the first six principal
 87 components were removed. Sex, age, and the first 20 principal components were used as covariates. Classical
 88 PRS methods use LD, MAF, and other measures to reweight the effect estimates. We found some performance
 89 improvement by using these re-weighted effect estimates as a baseline instead of the GWAS summary statistics.
 90 PRS were obtained with three different methods: C+T, Lassosum (9), and LDpred (8). The pre-processing step was
 91 implemented in R, using the bigsnpr (35) library.

92 Learning perturbations of mutation effects

93 The second step consists of learning individualized effect perturbations. As in GWAS, covariates were age, sex, and
 94 the first 20 PC loading. Before training, an XGBoost model was fitted on covariate data and is referred to as the

95 covariate model. Separately, the genotype data was converted from .bgen files to a hierarchical format that allows for
96 fast data retrieval of all HapMap3 SNPs of a small number of subjects. We trained the neural network on the residuals
97 of this model, which made convergence easier when some covariates had a high impact on the phenotype. The neural
98 network's architecture was a standard 8-layer transformer with variable sequence length depending on the number of
99 input SNPs. SNPs were aggregated into fixed-size groups and linearly mapped to form a sequence of embeddings of
100 size 512. In addition, covariates were included as the first embedding in the sequence, and zero padding was used
101 when necessary. The transformer's output was then mapped back into a vector the size of the number of input SNPs.
102 This vector represents individualized variations in the SNP effect. As in traditional PRS methods, these modified
103 effects were then summed up and linearly mapped in combination with the output of the covariate model to form a
104 final prediction. A graphical overview of the method is presented in figure 1.

105 Baseline PRS results

106 Three PRS methods (C+T, Lassosum2, and LDpred2) were compared to provide baselines. We compared the
107 proportion of explained variance (EVR) for all phenotypes. Predictions were made with a linear or logistic model, using
108 age, sex, and the first 20 genetic PCs as covariates and the estimated score. Our results are displayed in Figure 2.
109 LDpred2 outperformed the other two tested methods on all tested phenotypes. Thus, we chose the effect estimates
110 from LDpred2 as the baseline for our method in all further analyses. We also compared the performance of three
111 variants of LDpred2: LDpred2-grid and LDpred2-auto. We found that LDpred2-auto was superior to LDpred2-grid for
112 all tested phenotypes and used this variant.

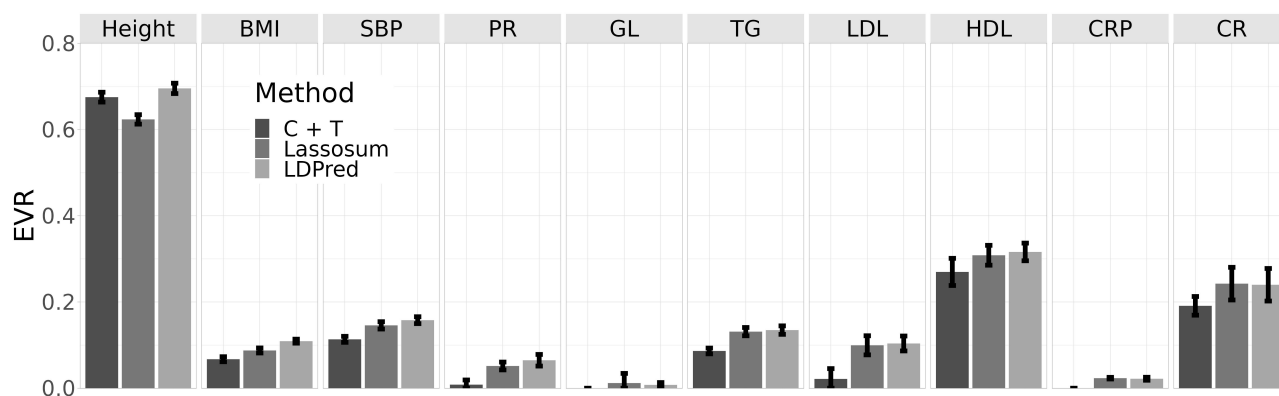


Figure 2. C+T, Lassosum2 and LDpred2 linear PRS results on the validation set. We show the best-performing model of three independent data splits. Validation sets were used to determine the optimal parameters for each method. Error bars indicate the standard deviation between splits. EVR: explained variance, BMI: Body mass index, CRP: C-Reactive protein, GL: glucose, LDL: low-density lipoproteins, SBP: systolic blood pressure.

113 Trait Prediction

114 We evaluated the performance of Delphi on ten continuous phenotypes, using three different train/test splits, and
115 used explained variance as performance metrics. We also compared Delphi with linear and Lasso regressions and an
116 approach using XGBoost to modify effect sizes (17) with the base weights from LDpred. Hyperparameters for all three
117 methods were tuned with three-way cross-validation on the same validation set. Results showed that Delphi resulted
118 in lower error than other approaches on all phenotypes Figure 3 provides detailed results. It should be noted that the
119 explained variance from benchmarked PRS methods is lower than the ones shown on the validation set in Figure2, as
120 the validation set was used for parameter tuning on this set for all 3 methods.

121 We then compared Delphi prediction to the next best method, XGboost, in terms of the distribution of errors for ten
122 phenotypes. Delphi generally tended to have fewer large prediction errors, as shown by the ratio of quartile difference
123 between the two methods (Figure 4). This is especially visible for height, for which ratios had to be bounded between
124 0.9 and 1.1 for visibility. We suppose the sharp gain in performance for height is due to the fact that this phenotype is
125 known to be highly polygenic and to exhibit SNP-sex and SNP-PC interactions (36), which makes this phenotype
126 particularly suitable for our approach. This difference in prediction distribution is also visible from histograms of
127 distances between predicted and ground truth decile values, as shown in figure 5.

128 Predictions for Delphi showed, in general, lower absolute error than XGboost prediction (supplementary section 1).
129 The relative increase in the percentage variance explained compared to the state-of-the-art was 11% for body mass
130 index, 19% for systolic blood pressure, and 35% for C-reactive protein; in addition, Delphi provided 2% absolute
131 explained variance for blood glucose while other tested methods were non-predictive.

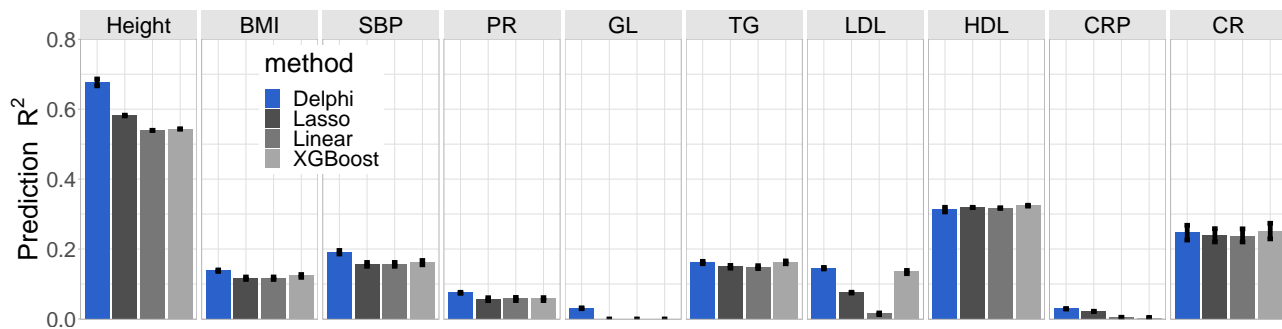


Figure 3. Accuracy of polygenic predictions for ten phenotypes in the UK Biobank. We report results for a linear PRS model, lasso regression, and XGBoost models, including the dosage of multiple high-impact SNPs as input, and our method. See Figure 2 for acronyms.

132 Performance comparison on non-white British for multiple phenotypes

133 We also compared performance on the subset of the test set with non-British white ancestries ($N \approx 13K$, depending on
 134 split and phenotype). Ancestry was determined according to Field 22006, which indicates subjects who self-identified
 135 as 'White British' according to Field 21000 and have very similar genetic ancestry based on a principal components
 136 analysis of the genotypes. Results are shown in Figure 6. The relative increase in the percentage variance explained
 137 compared to the state-of-the-art on the set with non-British white ancestries was 18% for body mass index, 25%
 138 for systolic blood pressure, 2% for LDL, and 15% for C-reactive protein; in addition, Delphi provided 2% absolute
 139 explained variance for blood glucose while other tested methods were non-predictive. Notably, the EVR is higher for
 140 non-British white individuals for some phenotypes (BMI and LDL). This predictive gain is due to the reduction of the
 141 total variance and is not reflected in the mean absolute error (see supplementary section 1). Results on the British
 142 white set were otherwise very similar to the total set shown in figure 3 as these individuals from approximately 95% of
 143 the held-out test set.

144 We also compared performance on subsets of the test set that self-reported as either African ($N \approx 560$), Chinese
 145 ($N \approx 290$), or Indian ($N \approx 920$). Results are shown in figure 7. Despite the low number of subjects in each group, Delphi
 146 outperforms other tested approaches on most phenotypes.

147 Observed Trends in Effect modulation

148 We observed interesting patterns when inspecting the average effect modulations before the summation. As shown in
 149 Figure 8, Delphi tends to down-weight the absolute effect of SNPs with low absolute effect. Interestingly, we do not
 150 observe the same trend when grouping SNPs by minor allele frequency decile. Other SNP-heritability estimation
 151 methods such as LDAK (37) include MAF, LD estimates and functional annotations to refine predictions. As LDPred
 152 modifies the effect estimates before any modification by the deep neural network, we expect the LD structure to
 153 be included in the effect estimates. This observation might indicate that the absolute effect may be an additional
 154 parameter of interest for future Bayesian methods.

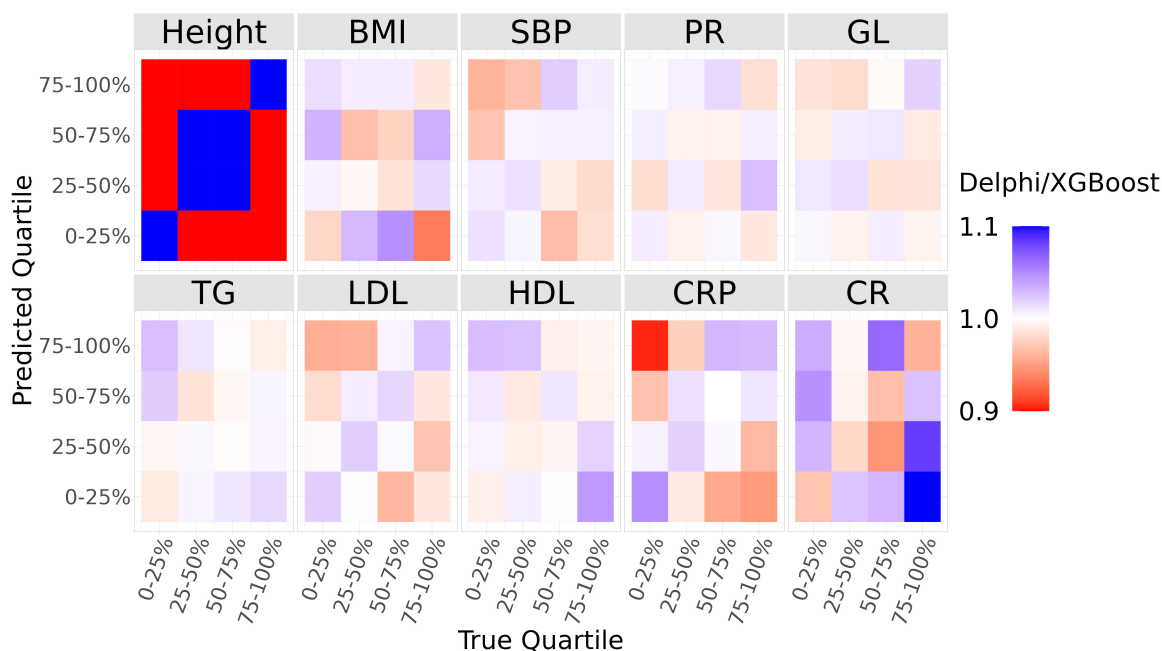


Figure 4. Ratio of quartile distributions of predictions between Delphi and XGBoost, on the test set for five phenotypes. Although the proportion of correctly binned subjects (same predicted and ground-truth quartiles) is similar for both methods, Delphi tends to avoid extreme differences between prediction and ground-truth. Values for height were bounded between 0.9 and 1.1 for visibility; original values are in the range of 0.2 and 1.3.

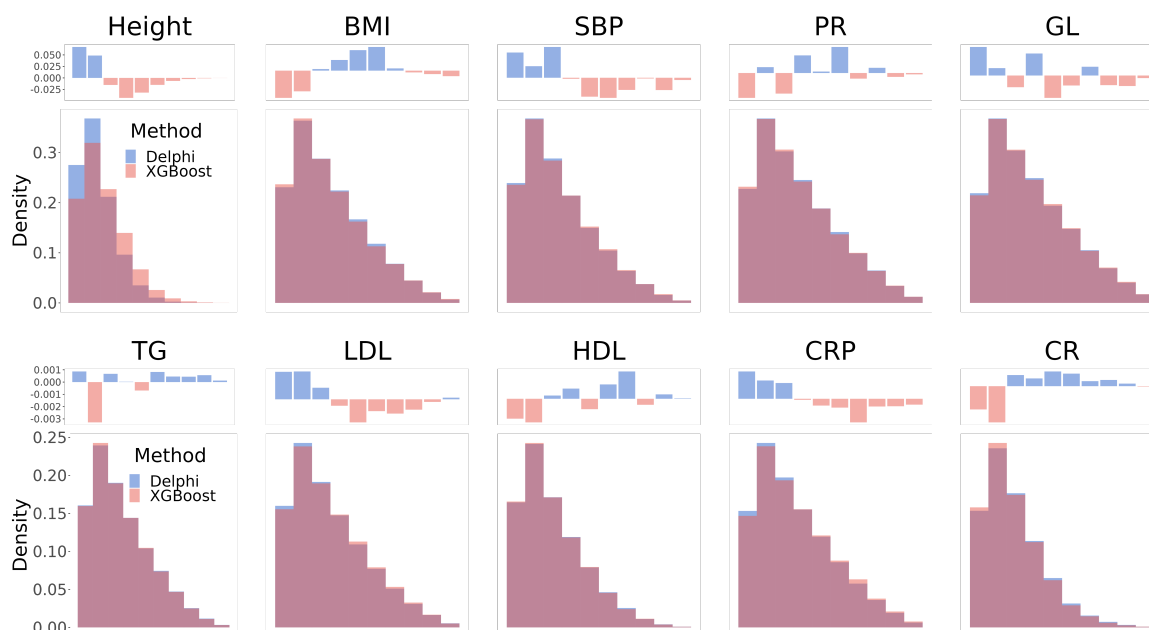


Figure 5. Histogram of the absolute difference between predicted and true deciles on the test set for five phenotypes. Delphi consistently bins subjects more adequately than XGBoost for most phenotypes.

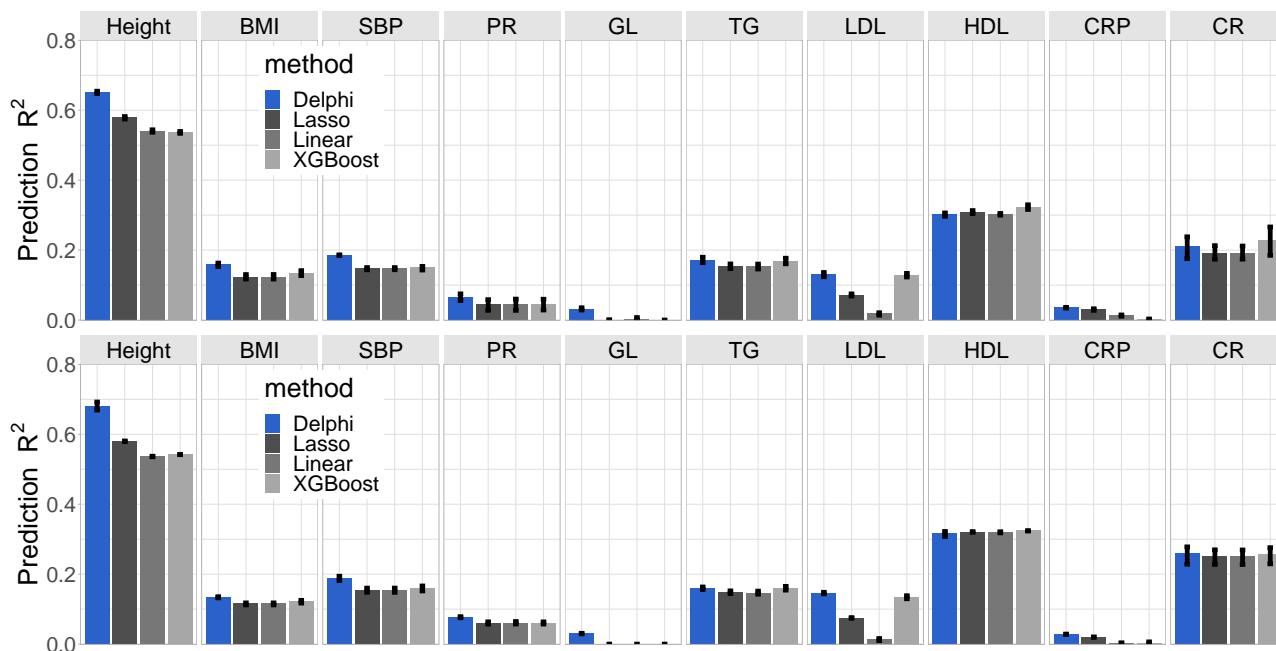


Figure 6. Accuracy of polygenic predictions for ten phenotypes in the UK Biobank. Top: predictions for ten phenotypes in the UK Biobank on individuals with non-British white ancestry. Bottom: Prediction results for individuals with British white ancestry. We report results for a linear PRS model, lasso regression, and XGBoost, including the dosage of multiple high-impact SNPs as input, and our method.

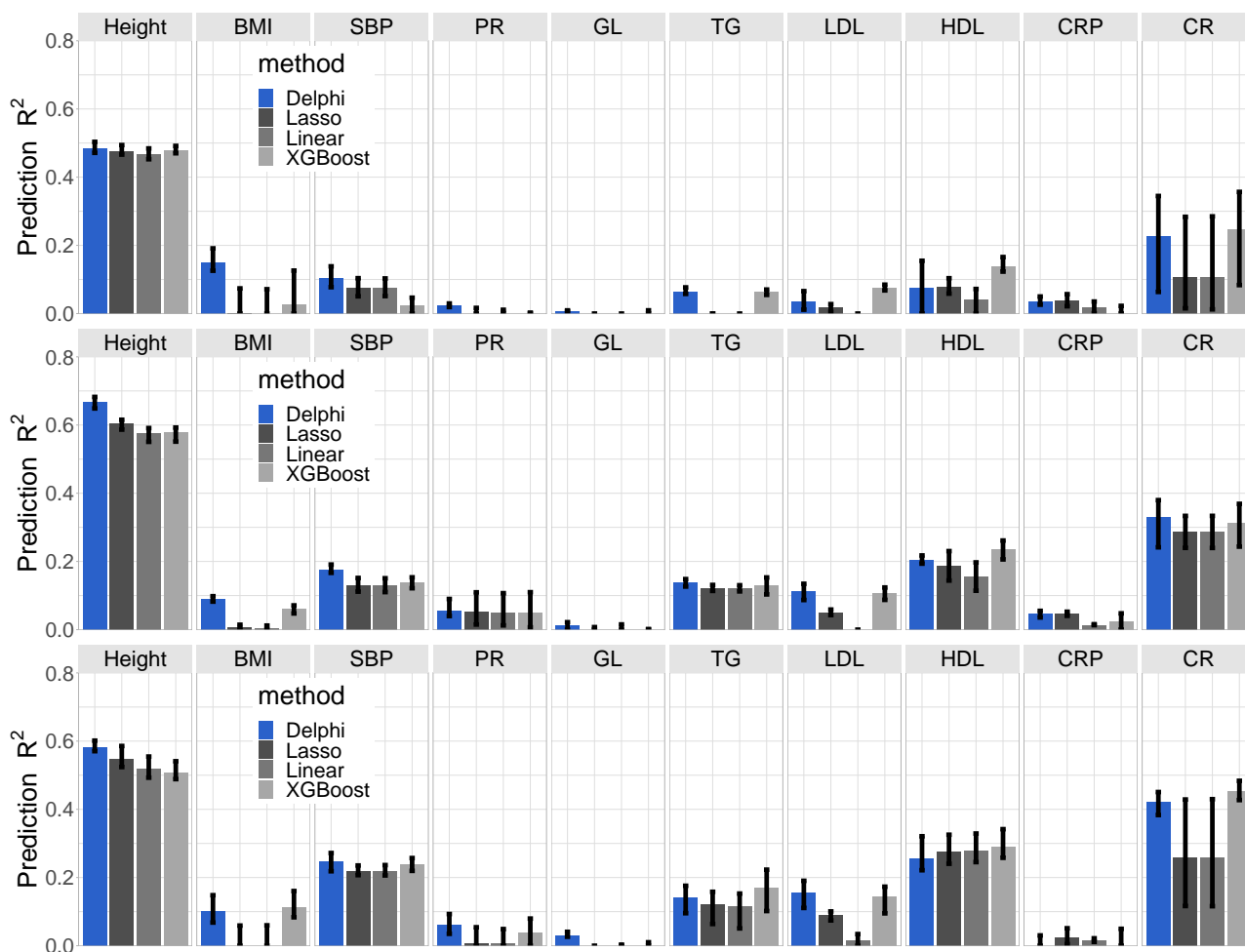


Figure 7. Accuracy of polygenic predictions for ten continuous phenotypes in the UK Biobank for three self-reported ethnic background. Top: African, middle: Indian, bottom: Chinese

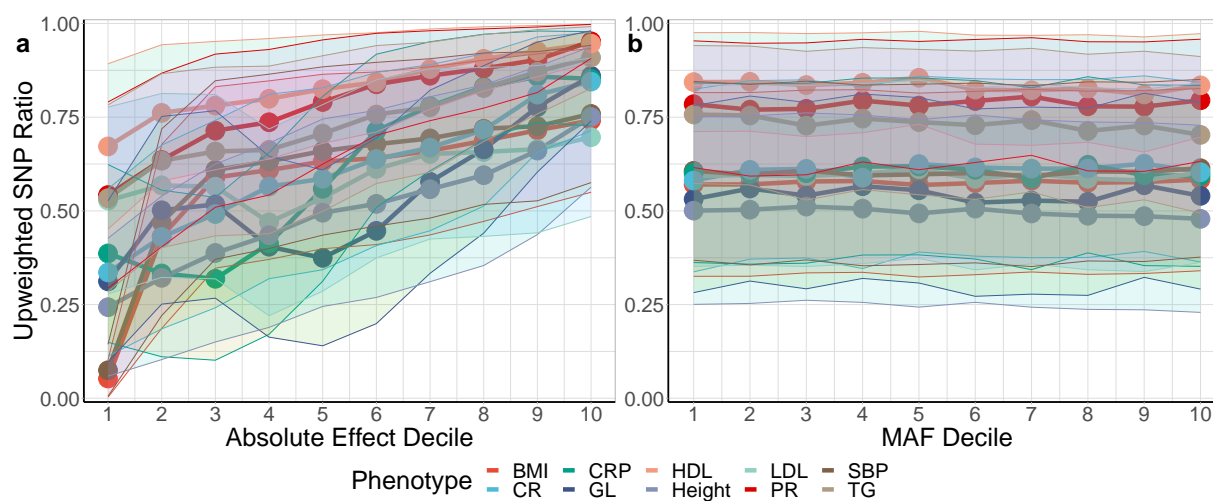


Figure 8. a) The ratio of up-weighted SNPs in Delphi by absolute effect size decile was estimated with LDpred2 for ten phenotypes. This ratio was computed by dividing the number of SNPs with higher estimated effect on average on the test set with the total number of SNPs for search decile. b) Ratio of up-weighted SNPs in Delphi by minor allele frequency.

155 Discussion

156 We introduced Delphi, a deep-learning-based method for trait prediction from genetic data. We demonstrated that deep
157 learning enhances the predictive power of polygenic risk scores. Treating genetic risk estimation as a deep prediction
158 problem allowed us to relax the usual assumptions of traditional PRS methods, yielding significant performance
159 improvements on multiple phenotypes over previous PRS computation methods.

160 Several studies have tested deep learning approaches for phenotype inference from genetic data. These approaches
161 all have a similar structure: use GWAS summary statistics to select a subset of SNPs, then use these as input for the
162 neural network. Uppu et al. (26) used a 3-layer feed-forward network applied to breast cancer data. To our knowledge,
163 this study contains the first use of a neural network for genetic risk prediction. In contrast, Bellot et al. (27) did not
164 find any performance gain when comparing convolutional and fully connected neural networks to traditional methods.
165 Recently, Huang et al. proposed DL-PRS (28), a method that also uses a shallow network to predict COPD, achieving
166 marginal performance gains over traditional methods on UKBiobank data. A similar approach has been used by Badre
167 et al. (29) with a 4-layer FC neural network on breast cancer data. Very recently, Zhou et al. (31) used a graph neural
168 network for Alzheimer's prediction by constructing a graph from a few correlated loci. Elgart et al. (17) showed the
169 strongest evidence for the superiority of non-linear methods for PRS by using gradient-boosted trees. This publication
170 obtained robust results across multiple traits, which motivated our study.

171 All previously mentioned approaches only consider a small subset of SNPs (typically less than 1000) as input and
172 become less predictive when including more small-effect SNPs. Training becomes difficult as smaller effects add
173 noise to the input due to their minimal individual impact on the phenotype and a lack of clearly exploitable patterns.
174 This is particularly a problem for PRS, which can include tens of thousands of SNPs. To guide the neural network
175 towards meaningful predictions, we chose to perturb the estimated effect sizes rather than predicting the phenotype
176 directly. As a result, we can effectively integrate up to a hundred thousand SNPs as input, which would not be feasible
177 with other methods.

178 We have also shown that our method generalizes well when evaluated on individuals from non-European ancestries,
179 although our training set is composed of 95 % European. This is an essential point for the success of PRS in any
180 clinical setting, or their application can potentially reinforce racial bias (38). Our approach could be combined with
181 other methods for the standardization of PRS, for instance, by combining summary statistics from multiple GWAS
182 studies (39) or through some debiasing measure (40). The performance and fairness of PRS is an ongoing problem
183 and requires more data acquisition from non-European cohorts. To reduce these disparities, it is necessary to assess
184 and maintain prediction performance for all populations thoroughly.

185 Our study presents several limitations. The high dimensionality of the data, combined with the sizeable but still limited
186 number of samples, means some trade-offs had to be taken to maintain consistent prediction performance across all
187 traits. Similarly to other PRS methods, the most crucial hyperparameter to tune is the minimum threshold probability
188 for SNP inclusion. This threshold also affects the number of SNPs we batch in a single embedding vector, and
189 training can diverge when including too many non-significant SNPs. A threshold of $> 1\%$ in MAF also limits our ability
190 to generalize to other ancestries but is required for estimating effect sizes. Similarly, we removed individuals from
191 non-European ancestry for GWAS to avoid spurious associations but kept them during training. Our approach also
192 takes significantly more computational power and time than the other compared methods.

193 Delphi tends to increase the effect of SNPs with high effect estimates and down-weights low effect SNPs. Similar
194 heuristics have been shown to improve heritability estimates by tuning effects based on minor allele frequency (MAF)
195 (10). Although MAF is correlated with effect size, we found no such association by inspecting variations modulation
196 and MAF quantiles. Unfortunately, we also found that the effect estimates of individual SNPs would vary drastically
197 between different data splits, making the interpretation of SNP effect modulation challenging, as the variations of the
198 neural network were much smaller than the differences from data randomization. This limitation might be alleviated by
199 including summary statistics from another cohort.

200 **Methods**

201 **UK Biobank**

202 The UK Biobank (UKBB) (32) is a large-scale ongoing prospective study including over half a million individuals
203 from across the United Kingdom. Participants were first recruited between 2006 and 2010 and underwent extensive
204 testing, including blood biomarkers, health and lifestyle questionnaires, and genotyping. Longitudinal hospitalization
205 data for any disease represented by an ICD-10 code is also provided between the recruitment date and the present
206 time. UKBB contains genotypes for 488,377 individuals at the time of download (March 2023), 409,519 of which are
207 from 'white British' ancestry. Ancestry was inferred using the data field 22006, which uses self-reports and principal
208 component analysis of the genotypes. Variant quality control included the removal of SNPs with imputation info
209 score < 0.8 and retaining SNPs with hard-call genotypes of > 0.9 probability and $MAF > 0.01$. To reduce the initial
210 dimensionality of the data, we only considered 1,054,330 HapMap3 (HM3) (41) SNPs as they have shown to be a
211 sufficient set for traditional PRS methods (42) and are the standard set for polygenic risk score evaluation.

212 **Data Splits and Phenotypes**

213 We evaluated the performance on all ten traits of our method using three independent train/validation/test splits. For
214 the quality control of our samples, we only considered 407,008 subjects used in the principal components analysis of
215 the UKB dataset (field 22020). These subjects are unrelated, did not withdraw consent from the study, and passed
216 some genotyping quality control tests. Subjects were not selected based on ancestry at this stage. We used 80%
217 (325,606) of the dataset for training, 5000 subjects for validation, and the rest (76,402 subjects) for testing. Some
218 individuals were further removed depending on missing data for each phenotype. We kept this exact split for the
219 preprocessing and the training of the neural network. The same training set was used to compute the GWAS and
220 train the neural network. The validation set was used to select the best hyperparameters for polygenic risk scores and
221 benchmark algorithms and to stop the deep neural network training. We assessed the performance of our method on
222 ten continuous phenotypes. BMI, height, SBP, LDL, and C reactive protein values were taken directly from the UK
223 Biobank first time point measurements.

224 The LD reference panel used for LDpred was previously computed (42) with some individuals from the test set. The
225 choice of LD reference panel was shown to have a limited impact on performance (42), and the same weights from
226 LDpred were used for all benchmarked methods. Finally, this panel did not contain individuals from non-British white
227 ancestry, ensuring that performance results on non-British white (see section [Performance comparison on non-white](#)
228 [British for multiple phenotypes](#)) are unbiased.

229 **PCA and GWAS**

230 PCA eigenvectors were obtained from HM3 SNPs using only genotype information from the training set for each
231 data split. As recommended (43), we used a truncated PCA method with initial pruning that iteratively removes
232 long-range LD regions. For GWAS computation, the training set was pruned by removing individuals with no British
233 white ancestry (field 22006) and who were beyond two standard deviations of the Mahalanobis distance of the first 6
234 PCs. This additional subject selection was only applied for the GWAS to prevent spurious relationships that can arise
235 with heterogeneous cohorts. Covariates for the regression included age, sex, the first 20 principal components, age^2 ,
236 $age \cdot sex$ and $age^2 \cdot sex$. PCA and GWAS were computed using the bigsnpr (35) R package (version 1.9.10).

237 **PRS Computation**

238 Polygenic risk scores were computed for each phenotype and data split using clumping and thresholding (C+T),
239 lassosum2, and LDpred2. In C+T, correlated variants are first clumped together, leaving only the ones with the lowest
240 P-values while others are removed. We used 50 P-value thresholds combined with stacking to learn an optimal linear
241 combination of C+T scores in a 10-fold cross-validation on the train set. The remaining variants are then pruned by
242 discarding the ones with a P-value larger than a chosen significance level. Lassosum uses L1 and L2 regularization on
243 the effect sizes and a linkage disequilibrium (LD) correlation matrix to penalize correlated and low-effect variants. The
244 regularization coefficients were chosen by measuring model performance on the validation set. LDpred is a Bayesian
245 method that uses a prior on effect sizes and an LD correlation matrix to re-weight effect estimates. LDpred2-auto
246 (44) is a variant of LDpred in which two key model parameters, the SNP heritability and polygenicity, are estimated
247 from the data. The LD correlation matrix was obtained from a reference panel (42). We used the validation set to
248 identify optimal hyper-parameters such as P-value cutoffs and regularization coefficients for each method. We used
249 the C+T, Lassosum2, and LDpred2 implementations of the bigsnpr (35) R package (version 1.9.10), using the default
250 hyperparameters ranges for each PRS method.

251 Network Architecture

252 We designed a deep learning neural network (DNN) to predict phenotypes from genetic data, illustrated in Figure 9.
 253 During training, SNPs are loaded in memory as a matrix of size $B \times S$, where B represents the batch size and S is the
 254 number of SNPs. To reduce the dimensionality of the data, SNPs were filtered by a tunable p-value threshold T . The
 255 neural network's architecture is an 8-layer pre-norm transformer (45) with two attention heads and GELU activation
 256 function (46). Similarly to vision transformers (45), we batched SNPs into arbitrary patches of length L to form a
 257 sequence of embeddings, using zero-padding to complete the last embedding. Inputs were then linearly mapped to
 258 match the input size of the transformer (512 in all experiments), and a vector containing covariate information was
 259 added as the first embedding of the sequence.

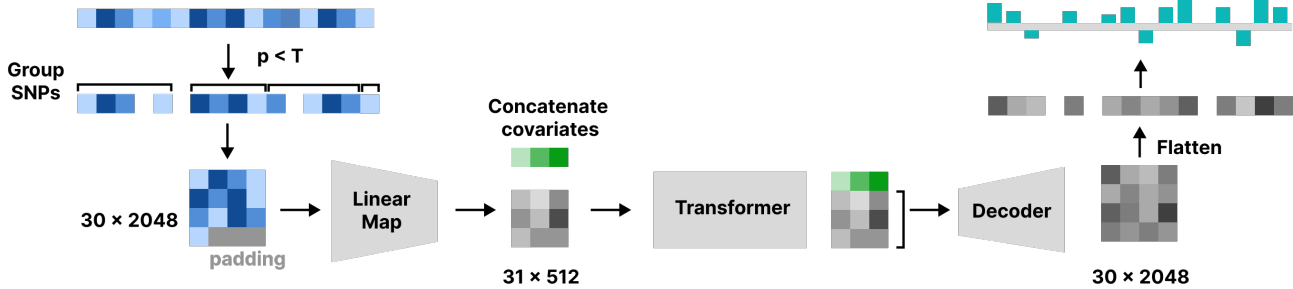


Figure 9. Overview of the architecture of the neural network.

260 We found that training directly on the phenotype would result in divergent training due to the large dimensionality
 261 of the data, the low impact of individual SNPs, and the fact that phenotypes are not fully described by their inputs.
 262 To remedy this problem, we used the effect sizes β_i from a classical PRS method to guide the neural network's
 263 predictions. We decoded the output of the transformer using a linear layer to match the original input size and predict
 264 variations of each effect using a \tanh activation function:

$$\beta'_i = \left(1 + \frac{1}{2} \tanh(f_{\theta}^i(x))\right) \beta_i, \quad (1)$$

265 where β_i is the effect size from the PRS, β'_i is the effect modified by the neural network f_{θ}^i , and x is the input to the
 266 neural network (covariates + 100K SNPs).

267 Each output β'_i represents an individual modification of the estimated effect that depends on covariates and the
 268 presence of other SNPs. Modified effect estimates were then summed up to form the modified PRS prediction of the
 269 DNN, $y_{DNN} = \sum \beta'_i$. We designed the DNN such that, were it to output only zeros, we would recover the unmodified
 270 PRS score. We found that using the effect estimates as a guide during training to be the only way for the neural
 271 network to output predictive results.

272 We used the effect sizes from LDpred2 to train the neural network in all our experiments. We used a batch size
 273 of $B = 512$, transformer input size of 512, feed-forward dimension of 512, and 0.3 dropout during training. The
 274 covariate vector included the same covariates from the GWAS for each trait. The patch size ($L \in \{128, 2048\}$), P-value
 275 thresholds (range $0.01 \cdot 10^{-6}$), and learning rate (range $0.05 \cdot 5 \cdot 10^{-4}$) were individually tuned for each trait and were
 276 the only parameters that varied between traits. For a specific trait, we used the same patch size and p-value threshold
 277 for each data split. We used a linear decay for the learning rate with 300 warmup steps. Models were trained on a
 278 single NVIDIA GeForce RTX 3090 (24 GB) and were composed of approximately 14M parameters. We used the
 279 AdamW optimizer (47) with $\epsilon = 10 \cdot 10^{-8}$, $\beta = (0.9, 0.999)$. Training averaged between 8 to 12 hours for each trait.

280 For binary traits, we used the ROC AUC as the evaluation metric.

281 For all phenotypes, we used explained variance (Equation 2) as the evaluation metric:

$$\text{EVR} = 1 - \frac{\text{var}(y - \hat{y})}{\text{var}(y)}, \quad (2)$$

282 where y is the ground truth and \hat{y} is the prediction.

283 We used these metrics to select the best-performing model on the validation set and for the final evaluation of the
 284 held-out test set. We used smooth L1 loss (48) during training.

285 Interestingly, we observed different convergence patterns for each phenotype. Some, like BMI and SBP, tended to
 286 converge after one epoch despite a very low learning rate and even overfit after this point. On the other hand, height
 287 required a much larger learning rate and converged after around 20 epochs. Binary traits included fewer SNPs due to
 288 differences in the distribution of GWAS p-values. Consequently, binary traits required a smaller patch size. We tuned
 289 the patch size such that the sequence length lay between 20 and 70, keeping patch sizes multiples of 2 between 256
 290 and 2048.

291 Covariate Model

292 As some covariates can greatly impact the phenotype (e.g., sex and height), we found that directly using the phenotype
293 as a ground truth would make the neural network diverge during training for some phenotypes. To solve this problem,
294 we used another model that only used the covariates as input to predict the phenotype and trained the deep neural
295 network on the residuals. We chose XGBoost, a gradient-boosted trees method similar to the method we used to
296 benchmark, without the additional high-impact SNPs. To be consistent, we used the same hyperparameters across
297 data splits and phenotypes with 3-fold cross-validation for optimal model selection. For the XGBoost hyperparameters,
298 we used a maximum depth of 5, $\alpha = 0, \gamma = 0, \eta = 0.01$, a subsample of 80%, and a minimum child weight of 10 in all
299 our experiments. The weighted sum of effect estimates with P-values lower than 0.05 but higher than the P-value
300 threshold of the deep neural network was then added back to the output of the covariate model. Finally, The DNN
301 predictions y_{DNN} were linearly combined with the covariate model to form the final prediction.

302 Data Loading During Training

303 Gene sequence variations formats such as bgen and pgen are compressed and optimized to query a single variant at
304 a time to enable fast GWAS analysis. For our purposes, we needed a format that could allow us to efficiently load
305 in memory all HM3 variants for a small number of subjects. We chose to convert the HM3 SNPs in bgen format to
306 a Hierarchical Data Format (HDF5) with a Python script using the h5py (49) and bgen_reader (50) libraries. When
307 loading the data, we implemented an efficient dataloader that merges genotype and phenotype information. This data
308 format allowed us to load a batch of 512 samples containing 1.1M SNPs in memory in less than 10 ms, which was
309 acceptable for training. Bgens were converted to a single HDF file with a Python script, which only needed to be done
310 once. We encoded allele dosage as 0, 1, and 2 for homozygous reference, heterozygous, and homozygous allele.
311 The samples were ordered as in the sample file from the .bgen of the first chromosome.

312 Adding in low effects as constants

313 To keep the inputs' dimensionality relatively low and avoid including extremely small effects, we summed up the effects
314 that were lower than 0.05% of the maximum and only included the others in the input of the neural network. The sum
315 of the smaller effects was then added to the y_{DNN} output. Assuming that the deep neural network output only zeros,
316 the network's architecture is such that output would be the same as the LDpred weighted sum.

317 Model performance evaluation and comparison to existing methods

318 We compared the performance of our approach to three other state-of-the-art methods. We used the polygenic risk
319 score predictions from LDpred2 for each method and included the same covariates as our approach. It was recently
320 found (17) that including high-impact SNPs in a non-linear model can increase the quality of genetic prediction. We
321 modified this existing method to be computationally feasible while enabling fair comparison. To be precise, instead
322 of filtering SNPs with LASSO regression before XGboost, which we found to be computationally expensive due to
323 the size of UKB, we filtered them by P-value thresholding. We considered for inclusion in the model all SNPs with
324 a p value $< 10^{-4}$ using our GWAS summary statistics for the corresponding trait, keeping the same data splits as
325 previously described. We then used eight relative thresholds α values between 0 and 1 and kept SNPs with a P-value
326 in the top T percentile.

327 We fitted XGBoost and LASSO models by including covariates (sex, age, first 20 PCs), selected SNPs, and the
328 LDpred2 risk score prediction. We selected the model that minimized the MSE for each phenotype. For XGBoost, we
329 always used a learning rate of 0.01, maximum depth of 5, minimum child weight of 10, and subsample of 80%. Each
330 model was fit using 3-fold cross-validation on the training set, allowing up to 2000 boosted trees with early stopping
331 after 20 rounds. We repeated this process for all 10 traits and 3 data splits. Analysis was conducted using Python 3
332 and the scikit-learn and xgboost packages.

333 Data Analysis with R

334 Data analysis was performed with publicly available packages: tidyverse v1.3.1 (51), and dplyr v1.0.8 (52).

335 Data Analysis with Python

336 The covariate model was implemented using the xgboost python library (53). The deep learning model was
337 implemented in Pytorch (54).

338 **Code Availability**

339 Code used for processing genetic data, GWAS analyses, and training of the neural network for this manuscript is
340 provided on a dedicated GitLab repository <https://gitlab.com/cgeo/delphi>.

341 **Data availability**

342 No data were generated in the present study. UK Biobank data are publicly available by application (<https://www.ukbiobank.ac.uk/enable-your-research/register>).
343

344 **Acknowledgements**

345 This research has been conducted using the UK Biobank resource under application number 80108, with funding
346 from the Swiss National Science Foundation (Sinergia CRSII5_202276/1).

Bibliography

1. Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorf, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, Aravinda Chakravarti, Judy H. Cho, Alan E. Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N. Rotimi, Montgomery Slatkin, David Valle, Alice S. Whittemore, Michael Boehnke, Andrew G. Clark, Evan E. Eichler, Greg Gibson, Jonathan L. Haines, Trudy F. C. Mackay, Steven A. McCarroll, and Peter M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, October 2009.
2. Michael D. Osterman, Tyler G. Kinzy, and Jessica N. Cooke Bailey. Polygenic Risk Scores. *Current Protocols*, 1(5):e126, May 2021.
3. Ali Torkamani, Nathan E. Wineinger, and Eric J. Topol. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19(9):581–590, September 2018.
4. Cathryn M. Lewis and Evangelos Vassos. Polygenic risk scores: from research tools to clinical instruments. *Genome Medicine*, 12(1):44, May 2020.
5. Jack W. O'Sullivan, Anna Shcherbina, Johanne M. Justesen, Mintu Turakhia, Marco Perez, Hannah Wand, Catherine Tcheandjieu, Shoa L. Clarke, Manuel A. Rivas, and Euan A. Ashley. Combining Clinical and Polygenic Risk Improves Stroke Prediction Among Individuals With Atrial Fibrillation. *Circulation. Genomic and Precision Medicine*, 14(3):e003168, June 2021.
6. David M. Evans, Peter M. Visscher, and Naomi R. Wray. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Human Molecular Genetics*, 18(18):3525–3531, September 2009.
7. Shaun M. Purcell, Naomi R. Wray, Jennifer L. Stone, Peter M. Visscher, Michael C. O'Donovan, Patrick F. Sullivan, Pamela Sklar, Shaun M. Purcell (Leader), Jennifer L. Stone, Patrick F. Sullivan, Douglas M. Ruderfer, Andrew McQuillin, Derek W. Morris, Colm T. O'Dushlaine, Aiden Corvin, Peter A. Holmans, Michael C. O'Donovan, Pamela Sklar, Naomi R. Wray, Stuart Macgregor, Pamela Sklar, Patrick F. Sullivan, Michael C. O'Donovan, Peter M. Visscher, Hugh Gurling, Douglas H. R. Blackwood, Aiden Corvin, Nick J. Craddock, Michael Gill, Christina M. Hultman, George K. Kirov, Paul Lichtenstein, Andrew McQuillin, Walter J. Muir, Michael C. O'Donovan, Michael J. Owen, Carlos N. Pato, Shaun M. Purcell, Edward M. Scolnick, David St Clair, Jennifer L. Stone, Patrick F. Sullivan, Pamela Sklar (Leader), Michael C. O'Donovan, George K. Kirov, Nick J. Craddock, Peter A. Holmans, Nigel M. Williams, Lyudmila Georgieva, Ivan Nikolov, N. Norton, H. Williams, Draga Toncheva, Vihra Milanova, Michael J. Owen, Christina M. Hultman, Paul Lichtenstein, Emma F. Thelander, Patrick Sullivan, Derek W. Morris, Colm T. O'Dushlaine, Elaine Kenny, Emma M. Quinn, Michael Gill, Aiden Corvin, Andrew McQuillin, Khalid Choudhury, Susmita Datta, Jonathan Pimm, Srinivasa Thirumalai, Vinay Puri, Robert Krasucki, Jacob Lawrence, Digby Quedest, Nicholas Bass, Hugh Gurling, Caroline Crombie, Gillian Fraser, Soh Leh Kuan, Nicholas Walker, David St Clair, Douglas H. R. Blackwood, Walter J. Muir, Kevin A. McGhee, Ben Pickard, Pat Malloy, Alan W. Maclean, Margaret Van Beck, Naomi R. Wray, Stuart Macgregor, Peter M. Visscher, Michele T. Pato, Helena Medeiros, Frank Middleton, Celia Carvalho, Christopher Morley, Ayman Fanous, David Conti, James A. Knowles, Carlos Paz Ferreira, Antonio Macedo, M. Helena Azevedo, Carlos N. Pato, Jennifer L. Stone, Douglas M. Ruderfer, Andrew N. Kirby, Manuel A. R. Ferreira, Mark J. Daly, Shaun M. Purcell, Pamela Sklar, Shaun M. Purcell, Jennifer L. Stone, Kimberly Chambert, Douglas M. Ruderfer, Finny Kuruvilla, Stacey B. Gabriel, Kristin Ardlie, Jennifer L. Moran, Mark J. Daly, Edward M. Scolnick, Pamela Sklar, The International Schizophrenia Consortium, Manuscript preparation, Data analysis, GWAS analysis subgroup, Polygene analyses subgroup, Management committee, Cardiff University, Karolinska Institutet/University of North Carolina at Chapel Hill, Trinity College Dublin, University College London, University of Aberdeen, University of Edinburgh, Queensland Institute of Medical Research, University of Southern California, Massachusetts General Hospital, and Stanley Center for Psychiatric Research and Broad Institute of MIT and Harvard. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–752, August 2009.
8. Bjarni J. Vilhjálmsson, Jian Yang, Hilary K. Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, Po-Ru Loh, Gaurav Bhatia, Ron Do, Tristan Hayeck, Hong-Hee Won, Sekar Kathiresan, Michele Pato, Carlos Pato, Rulla Tamimi, Eli Stahl, Noah Zaitlen, Bogdan Pasaniuc, Gillian Belbin, Eimear E. Kenny, Mikkel H. Schierup, Philip De Jager, Nikolaos A. Patsopoulos, Steve McCarroll, Mark Daly, Shaun Purcell, Daniel Chasman, Benjamin Neale, Michael Goddard, Peter M. Visscher, Peter Kraft, Nick Patterson, and Alkes L. Price. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *American Journal of Human Genetics*, 97(4):576–592, October 2015.
9. Timothy Shin Heng Mak, Robert Milan Porsch, Shing Wan Choi, Xueya Zhou, and Pak Chung Sham. Polygenic scores via penalized regression on summary statistics: MAK et al. *Genetic Epidemiology*, 41(6):469–480, September 2017.
10. Doug Speed, Na Cai, UCLEB Consortium, Michael R. Johnson, Sergey Nejentsev, and David J. Balding. Reevaluation of SNP heritability in complex human traits. *Nature Genetics*, 49(7):986–992, July 2017.
11. Carla Márquez-Luna, Steven Gazal, Po-Ru Loh, Samuel S. Kim, Nicholas Furlotte, Adam Auton, and Alkes L. Price. Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *Nature Communications*, 12(1):6052, October 2021.
12. L. Duncan, H. Shen, B. Gelaye, J. Meijssen, K. Ressler, M. Feldman, R. Peterson, and B. Domingue. Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications*, 10(1):3328, July 2019.
13. 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, April 2019.
14. Carla Márquez-Luna, Po-Ru Loh, South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA Type 2 Diabetes Consortium, and Alkes L. Price. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genetic Epidemiology*, 41(8):811–823, December 2017.
15. Yunfeng Ruan, Yen-Feng Lin, Yen-Chen Anne Feng, Chia-Yen Chen, Max Lam, Zhenglin Guo, Lin He, Akira Sawa, Alicia R. Martin, Shengying Qin, Hailiang Huang, and Tian Ge. Improving polygenic prediction in ancestrally diverse populations. *Nature Genetics*, 54(5):573–580, May 2022.
16. Tiffany Amariuta, Kazuyoshi Ishigaki, Hiroki Sugishita, Tazro Ohta, Masaru Koido, Kushal K. Dey, Koichi Matsuda, Yoshinori Murakami, Alkes L. Price, Eiryo Kawakami, Chikashi Terao, and Soumya Raychaudhuri. Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. *Nature Genetics*,

- 52(12):1346–1354, December 2020.
- 417 17. Michael Elgart, Genevieve Lyons, Santiago Romero-Brufau, Nuzulul Kurniansyah, Jennifer A. Brody, Xiuqing Guo, Henry J.
418 Lin, Laura Raffield, Yan Gao, Han Chen, Paul de Vries, Donald M. Lloyd-Jones, Leslie A. Lange, Gina M. Peloso, Myriam
419 Fornage, Jerome I. Rotter, Stephen S. Rich, Alanna C. Morrison, Bruce M. Psaty, Daniel Levy, Susan Redline, Paul de Vries,
420 and Tamar Sofer. Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse
421 human populations. *Communications Biology*, 5(1):1–12, August 2022.
 - 422 18. Rikifumi Ohta, Yosuke Tanigawa, Yuta Suzuki, Manolis Kellis, and Shinichi Morishita. A polygenic score method boosted by
423 non-additive models. *Nature Communications*, 15(1):4433, May 2024.
 - 424 19. Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo
425 Dijamco, Nam Nguyen, Pegah T. Afshar, Sam S. Gross, Lizzie Dorfman, Cory Y. McLean, and Mark A. DePristo. A universal
426 SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36(10):983–987, November 2018.
 - 427 20. Avanti Shrikumar, Katherine Tian, Žiga Avsec, Anna Shcherbina, Abhimanyu Banerjee, Mahfuza Sharmin, Surag Nair, and
428 Anshul Kundaje. Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version
429 0.5.6.5, April 2020. arXiv:1811.00416 [cs, q-bio, stat].
 - 430 21. Matthias Kirchler, Stefan Konigorski, Matthias Norden, Christian Meltendorf, Marius Kloft, Claudia Schurmann, and Christoph
431 Lippert. transferGWAS: GWAS of images using deep transfer learning. *Bioinformatics (Oxford, England)*, 38(14):3621–3628,
432 July 2022.
 - 433 22. Upamanyu Ghose, William Sproviero, Laura Winchester, Marco Fernandes, Danielle Newby, Brittany Ulm, Liu Shi, Qiang Liu,
434 Cassandra Adams, Ashwag Albukhari, Majid Almansouri, Hani Choudhry, Cornelia van Duijn, and Alejo Nevado-Holgado.
435 Genome wide association neural networks (GWANN) identify novel genes linked to family history of Alzheimer's disease in
436 the UK BioBank, June 2022.
 - 437 23. Arno van Hilten, Steven A. Kushner, Manfred Kayser, M. Arfan Ikram, Hieab H. H. Adams, Caroline C. W. Klaver, Wiro J.
438 Niessen, and Gennady V. Roshchupkin. GenNet framework: interpretable deep learning for predicting phenotypes from
439 genetic data. *Communications Biology*, 4(1):1–9, September 2021.
 - 440 24. Shuya Abe, Shinichiro Tago, Kazuaki Yokoyama, Miho Ogawa, Tomomi Takei, Seiya Imoto, and Masaru Fuji. Explainable AI
441 for Estimating Pathogenicity of Genetic Variants Using Large-Scale Knowledge Graphs. *Cancers*, 15(4):1118, January 2023.
 - 442 25. Nilesh Tripuraneni, Ben Adlam, and Jeffrey Pennington. Overparameterization Improves Robustness to Covariate Shift in High
443 Dimensions. In *Advances in Neural Information Processing Systems*, volume 34, pages 13883–13897. Curran Associates,
444 Inc., 2021.
 - 445 26. Suneetha Uppu, Aneesh Krishna, and Raj Gopalan. TOWARDS DEEP LEARNING IN GENOME-WIDE ASSOCIATION
446 INTERACTION STUDIES. *PACIS 2016 Proceedings*, June 2016.
 - 447 27. Pau Bellot, Gustavo de Los Campos, and Miguel Pérez-Enciso. Can Deep Learning Improve Genomic Prediction of Complex
448 Human Traits? *Genetics*, 210(3):809–819, November 2018.
 - 449 28. Sijia Huang, Xiao Ji, Michael Cho, Jaehyun Joo, and Jason Moore. DL-PRS: a novel deep learning approach to polygenic risk
450 score. Technical report, 2021. Type: article.
 - 451 29. Adrien Badré, Li Zhang, Wellington Muchero, Justin C. Reynolds, and Chongle Pan. Deep neural network improves the
452 estimation of polygenic risk scores for breast cancer. *Journal of Human Genetics*, 66(4):359–369, April 2021.
 - 453 30. Nimrod Ashkenazy, Martin Feder, Ofer M. Shir, and Sarel Hübner. GWANN: Implementing deep learning in genome wide
454 association studies, June 2022.
 - 455 31. Xiaopu Zhou, Yu Chen, Fanny C. F. Ip, Yuanbing Jiang, Han Cao, Ge Lv, Huan Zhong, Jiahang Chen, Tao Ye, Yuewen Chen,
456 Yulin Zhang, Shuangshuang Ma, Ronnie M. N. Lo, Estella P. S. Tong, Vincent C. T. Mok, Timothy C. Y. Kwok, Qihao Guo, Kin Y.
457 Mok, Maryam Shoai, John Hardy, Lei Chen, Amy K. Y. Fu, and Nancy Y. Ip. Deep learning-based polygenic risk analysis for
458 Alzheimer's disease prediction. *Communications Medicine*, 3(1):1–20, April 2023.
 - 459 32. Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green,
460 Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory
461 Collins. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle
462 and Old Age. *PLOS Medicine*, 12(3):e1001779, March 2015.
 - 463 33. Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD*
464 *International Conference on Knowledge Discovery and Data Mining*, pages 785–794, San Francisco California USA, August
465 2016. ACM.
 - 466 34. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, October 2007.
 - 467 35. Analysis of Massive SNP Arrays.
 - 468 36. Carrie Zhu, Matthew J. Ming, Jared M. Cole, Michael D. Edge, Mark Kirkpatrick, and Arbel Harpak. Amplification is the
469 primary mode of gene-by-sex interaction in complex human traits. *Cell Genomics*, 3(5):100297, May 2023.
 - 470 37. Doug Speed, John Holmes, and David J. Balding. Evaluating and improving heritability models using summary statistics.
471 *Nature Genetics*, 52(4):458–462, April 2020.
 - 472 38. Adebawale Adeyemo, Mary K. Balaconis, Deanna R. Darnes, Segun Fatumo, Palmira Granados Moreno, Chani J. Hodonsky,
473 Michael Inouye, Masahiro Kanai, Kazuto Kato, Bartha M. Knoppers, Anna C. F. Lewis, Alicia R. Martin, Mark I. McCarthy,
474 Michelle N. Meyer, Yukinori Okada, J. Brent Richards, Lucas Richter, Samuli Ripatti, Charles N. Rotimi, Saskia C. Sanderson,
475 Amy C. Sturm, Ricardo A. Verdugo, Elisabeth Widen, Cristen J. Willer, Genevieve L. Wojcik, Alicia Zhou, and Polygenic Risk
476 Score Task Force of the International Common Disease Alliance. Responsible use of polygenic risk scores in the clinic:
477 potential benefits, risks and gaps. *Nature Medicine*, 27(11):1876–1884, November 2021.
 - 478 39. Omer Weissbrod, Masahiro Kanai, Huwenbo Shi, Steven Gazal, Wouter J. Peyrot, Amit V. Khera, Yukinori Okada, Alicia R.
479 Martin, Hilary K. Finucane, and Alkes L. Price. Leveraging fine-mapping and multipopulation training data to improve
480 cross-population polygenic risk scores. *Nature Genetics*, 54(4):450–458, April 2022.
 - 481 40. Diego Machado Reyes, Aritra Bose, Ehud Karavani, and Laxmi Parida. FairPRS: adjusting for admixed populations in polygenic
482 risk scores using invariant risk minimization. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*,
483 28:198–208, 2023.
 - 484 41. David M. Altshuler, Richard A. Gibbs, Leena Peltonen, David M. Altshuler, Richard A. Gibbs, Leena Peltonen, Emmanouil
485 Dermitzakis, Stephen F. Schaffner, Fuli Yu, Leena Peltonen, Emmanouil Dermitzakis, Penelope E. Bonnen, David M.

- 486 Altshuler, Richard A. Gibbs, Paul I. W. de Bakker, Panos Deloukas, Stacey B. Gabriel, Rhian Gwilliam, Sarah Hunt, Michael
487 Inouye, Xiaoming Jia, Aarno Palotie, Melissa Parkin, Pamela Whittaker, Fuli Yu, Kyle Chang, Alicia Hawes, Lora R. Lewis,
488 Yanru Ren, David Wheeler, Richard A. Gibbs, Donna Marie Muzny, Chris Barnes, Katayoon Darvishi, Matthew Hurles,
489 Joshua M. Korn, Kati Kristiansson, Charles Lee, Steven A. McCarroll, James Nemesh, Emmanouil Dermitzakis, Alon Keinan,
490 Stephen B. Montgomery, Samuela Pollack, Alkes L. Price, Nicole Soranzo, Penelope E. Bonnen, Richard A. Gibbs, Claudia
491 Gonzaga-Jauregui, Alon Keinan, Alkes L. Price, Fuli Yu, Verner Anttila, Wendy Brodeur, Mark J. Daly, Stephen Leslie, Gil
492 McVean, Loukas Moutsianas, Huy Nguyen, Stephen F. Schaffner, Qingrun Zhang, Mohammed J. R. Ghorri, Ralph McGinnis,
493 William McLaren, Samuela Pollack, Alkes L. Price, Stephen F. Schaffner, Fumihiko Takeuchi, Sharon R. Grossman, Ilya
494 Shlyakhter, Elizabeth B. Hostetter, Pardis C. Sabeti, Clement A. Adebamowo, Morris W. Foster, Deborah R. Gordon, Julio
495 Licinio, Maria Cristina Manca, Patricia A. Marshall, Ichiro Matsuda, Duncan Ngare, Vivian Ota Wang, Deepa Reddy, Charles N.
496 Rotimi, Charmaine D. Royal, Richard R. Sharp, Changqing Zeng, Lisa D. Brooks, Jean E. McEwen, The International HapMap
497 3 Consortium, Principal investigators, Project coordination leaders, Manuscript writing group, Genotyping and QC, ENCODE
498 3 sequencing and SNP discovery, Copy number variation typing and analysis, Population analysis, Low frequency variation
499 analysis, Linkage disequilibrium and haplotype sharing analysis, Imputation, Natural selection, Community engagement
500 and sample collection groups, and Scientific management. Integrating common and rare genetic variation in diverse human
501 populations. *Nature*, 467(7311):52–58, September 2010.
- 502 42. Florian Privé, Julyan Arbel, Hugues Aschard, and Bjarni J. Vilhjálmsson. Identifying and correcting for misspecifications in
503 GWAS summary statistics and polygenic scores. *HGG advances*, 3(4):100136, October 2022.
- 504 43. Abdel Abdellaoui, Jouke-Jan Hottenga, Peter de Knijff, Michel G. Nivard, Xiangjun Xiao, Paul Scheet, Andrew Brooks, Erik A.
505 Ehli, Yueshan Hu, Gareth E. Davies, James J. Hudziak, Patrick F. Sullivan, Toos van Beijsterveldt, Gonneke Willemssen, Eco J.
506 de Geus, Brenda W. J. H. Penninx, and Dorret I. Boomsma. Population structure, migration, and diversifying selection in the
507 Netherlands. *European Journal of Human Genetics*, 21(11):1277–1285, November 2013.
- 508 44. Florian Privé, Julyan Arbel, and Bjarni J. Vilhjálmsson. LDpred2: better, faster, stronger. *Bioinformatics (Oxford, England)*,
509 36(22-23):5424–5431, April 2021.
- 510 45. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa
511 Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16
512 Words: Transformers for Image Recognition at Scale, June 2021. arXiv:2010.11929 [cs].
- 513 46. Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and
514 Tie-Yan Liu. On Layer Normalization in the Transformer Architecture, June 2020. arXiv:2002.04745 [cs, stat].
- 515 47. Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. arXiv:1412.6980 [cs].
- 516 48. Ross Girshick. Fast R-CNN, September 2015. arXiv:1504.08083 [cs].
- 517 49. HDF5 for Python.
- 518 50. Bgen-reader's documentation — bgen-reader 4.0.8 documentation.
- 519 51. Hadley Wickham and RStudio. tidyverse: Easily Install and Load the 'Tidyverse', February 2023.
- 520 52. Hadley Wickham, Romain François, Lionel Henry, Kirill Müller, Davis Vaughan, Posit Software, and PBC. dplyr: A Grammar of
521 Data Manipulation, November 2023.
- 522 53. xgboost: XGBoost Python Package.
- 523 54. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin,
524 Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan
525 Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style,
526 High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran
527 Associates, Inc., 2019.