Improving Neural Networks for Genotype-Phenotype Prediction Using Published Summary Statistics

Tianyu Cui¹, Khaoula El Mekkaoui¹, Aki Havulinna^{2,3}, Pekka Marttinen^{1,2,5}, and Samuel Kaski^{1,4,5}

¹ Department of Computer Science, Aalto University, Finland, firstname.lastname@aalto.fi

² Finnish Institute for Health and Welfare (THL), Helsinki, Finland, aki.havulinna@thl.fi

³ Institute for Molecular Medicine Finland, FIMM-HiLIFE, Helsinki, Finland

⁴ Department of Computer Science, University of Manchester, UK

⁵ Equal contribution

Abstract. Phenotype prediction is a necessity in numerous applications in genetics. However, when the size of the individual-level data of the cohort of interest is small, statistical learning algorithms, from linear regression to neural networks, usually fail due to insufficient data. Fortunately, summary statistics from genome-wide association studies (GWAS) on other large cohorts are often publicly available. In this work, we propose a new regularization method, namely, *main effect prior (MEP)*, for making use of GWAS summary statistics from external datasets. The main effect prior is generally applicable for machine learning algorithms, such as neural networks and linear regression. With simulation and real-world experiments, we show empirically that MEP improves the prediction performance on both homogeneous and heterogeneous datasets. Moreover, deep neural networks with MEP outperform standard baselines even when the training set is small.

Keywords: Neural network · Summary statistics · Regularization · Phenotype prediction.

Fig. 1. Overview. Prediction accuracy of a model on the target cohort, with few individual-level training data, is improved by incorporating the GWAS summary statistics from an external cohort through our main effect prior (MEP). This does not require accessing the individual-level data of the external cohort.

1 Introduction

Genotype-based phenotype prediction is central in various genetics applications. For example, accurate prediction of human complex traits and diseases from genotype has shown great promise in precision medicine and healthcare [1,7,47]. In genomic selection, predicting the grain yield with genotype and environment factors is revolutionizing plant breeding and increasing food production [34,38,39]. In bacterial genomics, predicting phenotypes such as antibiotic resistance has become a fundamental task to understand bacteria [27,40].

Existing approaches for phenotype prediction can be divided into two categories, depending on whether the modeler has access to individual-level data. Classical statistical genetics approaches, such as BLUP [3,17,33] and its extensions [16,21,44], demonstrate impressive predictive performance when individual-level data are available. Moreover, standard machine learning methods are becoming appealing alternatives because fewer assumptions are made about the genetic mechanism underlying the traits of interest, ranging from linear models, e.g., Lasso [12,28], to complex models, e.g., neural networks [2,18,38,39,41,55], that capture nonlinear main effects and feature interactions. However, due to the few assumptions made by deep neural networks, suitable inductive biases [37,52], such as neural network architectures and regularizations, are needed to learn the optimal function, especially for small data. On the other hand, accessing individual-level genomic data is cumbersome due to the enormous volume of the data and often also restricted due to privacy considerations such as the GDPR [50], which recently have made the summary statistics approach attractive: Models are learned using summary statistics from genome-wide association studies (GWAS) without access to individual-level data [15,31,57]. These methods build on summary statistics, including the estimated main effect and standard error of each single-nucleotide polymorphism (SNP) as well as the estimated linkage disequilibrium (LD) matrix, which are generally easier to share than individual-level datasets.

The accuracy of predictive models trained with individual-level data or summary statistics data for several traits has already been promising [31,32] with large biobank datasets of target cohorts. However, a more realistic situation is that users only have a *small* number of individual-level data for the cohort of interest, in which case neural networks and even linear models may fail due to insufficient data. Fortunately, the GWAS summary statistics of the same traits may be publicly available for a different cohort with a large amount of data. Therefore, the question of how to leverage the external publicly available summary statistics to improve phenotype prediction with complex models on small datasets is an engaging research question with various applications.

In this paper, we focus on predicting continuous human phenotypes, such as metabolites, with genotype data on a small target cohort but with access to the GWAS summary statistics from an external dataset, as shown in Figure 1. We propose a novel regularization term that incorporates prior knowledge about the main effect of each SNP from the external GWAS summary statistics, and the regularization can be widely applied to any optimization-based machine learning method, such as linear models and neural networks. We

demonstrate that the *main effect prior (MEP)* consistently improves all models, and neural networks with MEP outperform existing approaches even when the training dataset is small.

2 Background

2.1 Neural networks and regularization

Recent advances in neural networks have demonstrated strong predictive performance with large training datasets in various fields [13,25]. However, neural networks are prone to overfitting when encountered with small and noisy datasets, such as genotype-phenotype datasets. A number of regularization techniques, e.g., weight decay [26] and Lasso neural network [29], have been applied to constrain the flexibility of neural networks; They are all equivalent to zero-mean priors from a Bayesian perspective. More sophisticated regularization techniques have been proposed to incorporate extra domain knowledge into the model beyond sparsity [43]. For example, when the importance score of each feature is known *a priori*, attribution priors could be used to regularize the feature importance in the model, and methods such as DeepSHAP [48] and Contextual Decomposition [42] agree with the corresponding prior. While this way of regularizing the model has improved performance in specific problem domains, interpretation of the results requires *local interpretations* [51], i.e., the feature importance specific to each data point, assigned by humans, instead of *global interpretations*, such as main effects. A further complication is that feature importance scores in the deep learning community have a very different meaning from the main effects that are typically considered in a GWAS. Thus, new ways of regularizing the model are needed for neural networks to incorporate external GWAS results.

2.2 Bayesian approaches to incorporate domain knowledge

Priors on the proportion of variance explained (PVE) have been proposed for linear regression [56] and neural networks [11] to improve feature selection and prediction in genetics. Building of informative priors for neural networks has also been studied in the function space. For example, Gaussian processes [14,45] have been proposed as a way of defining functional priors to encode rich functional structures with specific kernels. Bayesian approaches have the capability of marginalizing over all possible weights and capturing the predictive uncertainty [52], with the cost of increased computational complexity, because approximate inference methods are usually needed for neural networks. In this paper, we focus on the regularization-based approaches and leave the Bayesian extensions for future work.

3 Methods

3.1 Neural networks with residual connections

Let $\mathcal{D} = \{(X, Y) | X \in \{0, 1, 2\}^{N \times p}, Y \in \mathbb{R}^N\}$ denote a training dataset with N individuals, p SNPs, and one phenotype variable, where $(\mathbf{x}_i, y_i) \in \mathcal{D}$ represents the SNPs coded as 0,1 or 2 and phenotype of the i th individual. A neural network (NN) is a function mapping inputs to outputs by aggregating predictions of simple but non-linear functions. Selecting a suitable architecture for the NN is critical because it reflects the inductive biases of the model. For example, convolutional layers learn translation-invariant functions, which are especially helpful in computer vision tasks [19].

In this paper we use a multi-layer perceptron (MLP) with a residual connection as the model architecture,

$$f(\mathbf{x}_i; \mathbf{w}) = \mathbf{x}_i \mathbf{w}_{\text{LR}} + f_{\text{MLP}}(\mathbf{x}_i; \mathbf{w}_{\text{MLP}}), \quad (1)$$

where the width and depth of $f_{\text{MLP}}(\cdot)$ can be fixed beforehand or cross-validated. The residual network is essentially combining a linear regression with an MLP, which makes the neural network easier to train [22]. Moreover, it has been shown that residual networks are usually better than models without residual connections in genetics applications [8,29,30], and especially suitable for genotype-phenotype prediction, where linear main effects of the genotype often explain most of the phenotype variance.

3.2 Main effect prior (MEP)

We introduce the main effect prior as a novel regularization method for training neural networks. In the standard training procedure of neural networks, we find the optimal model parameters $\hat{\mathbf{w}}$ by minimizing a loss function, which consists of the average prediction error, $\mathcal{L}(f(\mathbf{x}_i; \mathbf{w}), y_i)$, on the training set and a regularization function $\Omega(\mathbf{w})$ on the model parameters $\mathbf{w}$ to avoid overfitting,

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^N \mathcal{L}(f(\mathbf{x}_i; \mathbf{w}), y_i) + \lambda \Omega(\mathbf{w}). \quad (2)$$

Here a cross-validated hyper-parameter λ on the training set controls the regularization strength. In the explanation prior framework [42,43,51], another regularization function, $\Omega'(\phi(\mathbf{w}), \phi^*)$, measuring the difference between the feature importance of the model and prior feature importance, is added:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^N \mathcal{L}(f(\mathbf{x}_i; \mathbf{w}), y_i) + \lambda \Omega(\mathbf{w}) + \lambda' \Omega'(\phi(\mathbf{w}), \phi^*). \quad (3)$$

Here $\phi(\mathbf{w})$ represents the feature importance of the neural network $f(\cdot; \mathbf{w})$, and ϕ^* is the prior feature importance provided by the user.

The published GWAS summary statistics are essentially the linear regression coefficients of SNPs when predicting the phenotype variable using each SNP independently. In order to derive priors from GWAS summary statistics, we calculate the multivariate linear regression which best approximates the neural network, and regularize its coefficients according to the explanation prior framework (Equation 3) towards the joint effects of SNPs, which can be derived from the published GWAS summary statistics (see Section 3.3). Specifically, we define the feature importance vector $\phi(\mathbf{w})$ as the coefficient vector of a multiple linear regression that predicts $f(X; \mathbf{w})$, i.e., the projection of neural network predictions on the X :

$$\phi(\mathbf{w}) = (X^T X)^{-1} X^T f(X; \mathbf{w}). \quad (4)$$

Assuming we have the prior joint main effects and corresponding standard deviations of SNPs, i.e., $(\hat{\psi}, \hat{\sigma}_{\psi})$, derived from the GWAS summary statistics of an external dataset, we define the main effect prior (MEP) as

$$\Omega'(\phi(\mathbf{w}), \phi^*) = \left\| \frac{\phi(\mathbf{w}) - \hat{\psi}}{\hat{\sigma}_{\psi}} \right\|_1, \quad (5)$$

and the regularization coefficient λ' is cross-validated in the same way as λ . Note that Equation 5 defines an *implicit* prior over neural network functions, instead of an *explicit* prior over the weights, in a fashion similar to the common weight decay regularization, which corresponds to the Gaussian prior. In our case, this prior encodes the inductive bias that the main effects implied by the neural network model should comply with the published summary statistics from the external data. In essence, our goal is to approximate the neural network with linear regression to provide importance weights matching earlier summary statistics, a priori.

In this work, we focus on continuous phenotype prediction problems, so the prediction error $\mathcal{L}(f(\mathbf{x}_i; \mathbf{w}), y_i)$ is measured with the squared error. We use L_1 regularization to avoid overfitting, analogously to the LassoNet [29], i.e., $\Omega(\mathbf{w}) = \|\mathbf{w}\|_1$, because the genotype matrix is potentially high-dimensional.

3.3 Constructing the prior

In this section, we provide details of constructing the prior joint main effects, i.e., $(\hat{\psi}, \hat{\sigma}_{\psi})$ in Equation 5, from existing GWAS summary statistics.

Estimating the prior joint effects from marginal effects In GWAS, the reported summary statistics are usually marginal effects of SNPs that are tested one by one with single-SNP models, which do not take the correlation between SNPs (i.e., the linkage disequilibrium) into account. In order to use the MEP defined in Equation 5, we first convert the marginal effects from GWAS summary statistics into joint effects.

We apply the earlier-studied approach [54] to estimate the joint effects of multiple SNPs from corresponding marginal effects without accessing the individual-level data. If the individual-level genotype-phenotype data is available, we could estimate the joint effects $(\hat{\mathbf{b}}, \hat{\sigma}_{\mathbf{b}})$ and marginal effects $(\hat{\beta}, \hat{\sigma}_{\beta})$ of multiple SNPs with the least-squares approach

$$\begin{aligned}\hat{\mathbf{b}} &= (X_e^T X_e)^{-1} X_e^T Y_e, \quad \hat{\sigma}_{\mathbf{b}}^2 = \sigma_J^2 (X_e^T X_e)^{-1}; \\ \hat{\beta} &= D_e^{-1} X_e^T Y_e, \quad \hat{\sigma}_{\beta}^2 = \sigma_M^2 D_e^{-1},\end{aligned}\tag{6}$$

where (X_e, Y_e) represents the external genotype-phenotype data with size N_e where the summary statistics have been estimated, D_e is the diagonal matrix with entries from the diagonal of $X_e^T X_e$, and σ_J^2 and σ_M^2 are the residual variances of the multi-SNP linear model and multiple single-SNP linear models. We assume that the mean of each SNP has been removed for simplicity.

From Equation 6, we can write the relations between $\hat{\mathbf{b}}$ and $\hat{\beta}$ immediately [54]:

$$\hat{\mathbf{b}} = (X_e^T X_e)^{-1} D_e \hat{\beta},\tag{7}$$

where $(X_e^T X_e)^{-1}$ and D_e are unavailable to us because individual-level data have not been published. However, assuming Hardy-Weinberg equilibrium (HWE) and knowing the allele frequency f_j of SNP j , the j th element of $D_{e(j)}$ be estimated by $D_{e(j)} = 2f_j(1 - f_j)N_e$. Moreover, knowing the LD matrix Σ_e of the external dataset, we can estimate $X_e^T X_e$ with

$$X_e^T X_e = D_e^{1/2} \Sigma_e D_e^{1/2}.\tag{8}$$

When Σ_e is unavailable, the LD matrix from a reference sample (e.g., the training data) can be used instead.

Following the definition of residual variances of single and multiple linear regressions, we have the following relations between σ_J^2 and σ_M^2 :

$$\hat{\sigma}_J^2 = \frac{Y_e^T Y_e - \hat{\mathbf{b}}^T D_e \hat{\beta}}{N - p}, \quad \hat{\sigma}_{M(j)}^2 = \frac{Y_e^T Y_e - D_e(j) \hat{\beta}_j}{N - 1}.\tag{9}$$

Hence, we can compute $\hat{\sigma}_J^2$ by first estimating $Y_e^T Y_e = \hat{\sigma}_{M(j)}^2 (N - 1) + D_e(j) \hat{\beta}_j$. Then the standard deviation $\hat{\sigma}_{\mathbf{b}}^2$ can be estimated with Equation 6 [54]. There are sophisticated approaches based on these ideas, such as LDpred [49] and PRS-CS [15], which additionally accommodate different effect size distributions within the Bayesian framework, and which can be applied to reconstruct the prior joint effects from GWAS summary statistics. In this work, we adopt the straightforward idea presented above and leave the exploration of the possible further enhancements for future work.

Calibrating the prior joint effects on the training set After obtaining the prior joint effects $(\hat{\mathbf{b}}, \hat{\sigma}_{\mathbf{b}})$ from external GWAS summary statistics, we propose to calibrate the prior joint effects on a target cohort. This is needed for example because the phenotype might have been scaled differently in different cohorts due to different measurement standards [24,35,36]. Moreover, it is necessary to normalize the phenotype of the training data to improve the neural network training. For these reasons, the raw prior joint effects usually reflect the actual influences on the target phenotype only up to a scaling factor.

We use a simple linear regression on the training set to calibrate the prior joint effect, by scaling the phenotype predicted with summary statistics $\mathbf{x}_i \hat{\mathbf{b}}$ according to the target phenotype,

$$\hat{y}_i = \alpha + \hat{\alpha}_1 \mathbf{x}_i \hat{\mathbf{b}}.\tag{10}$$

Thus, the calibrated prior joint effects that will be used in Equation 5 are given by $(\hat{\psi}, \hat{\sigma}_{\psi}) = (\hat{\alpha}_1 \hat{\mathbf{b}}, \hat{\alpha}_1 \hat{\sigma}_{\mathbf{b}})$. In Section 4.2, we show that this calibration is necessary especially when the external summary statistics are available from a different cohort than the target cohort.

4 Experiments

In this section, we first analyze multiple simulated data sets with different configurations to demonstrate the benefits of MEP. Then we apply the approach to two real-world metabolite data sets. Lastly, we conduct several ablation studies and a sensitivity analysis on the proposed method.

4.1 Simulation study

Setting: To construct the genotype dataset X for the simulation study, we select all SNPs (after basic quality control) from different numbers p_g of genes associated with several cholesterol phenotypes from the UK Biobank [5], resulting in p SNPs in total. We simulate the phenotype Y by first simulating the “gene expression level” G_g of each gene g according to a linear function of all SNPs $X_{(g)}$ in that gene, and then generate phenotypes Y with main and interaction effects of G_g :

$$\begin{aligned} G_g &= X_{(g)}\alpha_{(g)}, \forall g \in \{1, \dots, p_g\}, \\ Y &= \sum_{g=1}^{p_g} w_g G_g + \sum_{1 \leq j \neq k \leq p_g} f_{jk}(G_j, G_k) + \epsilon. \end{aligned} \quad (11)$$

In the simulator, $f_{jk}(\cdot, \cdot)$ represents the interaction function between genes j and k , chosen from three types: multiplicative, $G_j \times G_k$, maximum, $\max\{G_j, G_k\}$, and squared difference, $(G_j - G_k)^2$. We set the parameters related to the main effect, i.e., $\alpha_{(g)}$ and w_g , such that the main effects explain $\sim 90\%$ of the signal variance, and we tune the variance of noise ϵ to vary the signal-to-noise ratio (S/N) of the data. For each simulation, we randomly select 100,000 individuals to construct the external summary statistics. From the remaining individuals, we select N in the training set and 20,000 in the test set; hence, this corresponds to the simple setup where the summary statistics are from the same cohort as the target data. We study the performance of the MEP for different numbers of training data N , numbers of SNPs p , and signal-to-noise ratios (S/N).

The two-stage generative process in Equation 11 reflects the idealized domain knowledge that SNPs within a gene affect how the gene is expressed, and the combined expression of multiple genes affects the phenotype. Furthermore, it reduces the complexity of simulating pairwise interactions, from $p(p-1)/2$ to $p_g(p_g-1)/2$ interactions. However, unlike [10], we do not assume this hierarchical structure of the simulator in the neural network that is used for prediction.

Results: We show the results of the simulation study with different settings in Figure 2. In general, we observe that neural networks with MEP, NN+MEP (external), that incorporate the summary statistics from the external set outperform neural networks without MEP (NN) and linear regression (LR), consistently in all settings. Moreover, we see in the middle panel that as more individual-level training data are available, the performance of NN+MEP(external) and NN becomes indistinguishable (e.g., when $N=80k$). In the last panel of Figure 2, we study how the number of SNPs affects the results. We notice that although the performance of all methods decreases as the number of SNPs increases (while keeping the total S/N fixed), the gap between NN+MEP(external) and NN increases. This is because models without MEP fail to learn the effects of high-dimensional SNPs from insufficient training data. Since the main effects of SNPs have already been captured by GWAS summary statistics on the large external dataset in MEP(external), using MEP is especially beneficial for high-dimensional data.

The figure also plots results of the “golden standard” baseline constructed by training NNs on both the training set and the external set (NN+external data). This assumes that we can access the individual-level data of the large external cohort instead of only the GWAS summary statistics, and is hence a non-realistic upper bound of results. As expected, NN+external data results are better than all the realistic methods including NN+MEP(external) because the external dataset contains more information than its summary statistics. However, the performance gaps are relatively small when S/N is small (e.g., S/N=0.1 in the left panel), where the linear effects captured by the GWAS summary statistics are likely to explain most of the signal.

4.2 Real-world datasets

In this section, we consider the prediction of six different NMR metabolite phenotypes determined at Nightingale Health [53]: cholesterol and triglycerides in HDL, LDL, and VLDL, available in two real-world datasets UK Biobank [5] and FINRISK [4]. We use these two data sets to construct two kinds of realistic test cases: 1) the *homogeneous* case, where the summary statistics and the target dataset are both from different subsets of the same cohort, and 2) the *heterogeneous* case where the summary statistics are from an external cohort

Fig. 2. Results of the simulation study. Each bar shows the averaged test PVE (over 50 repeated experiments), and the error bar is the corresponding 95% CI. An unrealistic upper bound (NN+external data) is given by assuming that the individual-level external dataset is available, instead of only the GWAS summary statistics. With MEP that incorporates the GWAS summary statistics from the external dataset, NN+MEP(external) outperforms models trained on the target dataset without MEP in all settings.

that is different from the cohort of the target data. In both cases, we use a large subset of the UK Biobank as the external data from which the summary statistics are extracted, and either a small subset of the UK Biobank or the FINRISK as the target data set in the homogeneous and heterogeneous cases, respectively.

Setting: To construct the homogeneous case study, we randomly split the UK Biobank dataset, which consists of 117,981 individuals with fully measured NMR metabolomics, into two parts: an external dataset with 110,000 individuals for which we estimate the GWAS summary statistics with PLINK1.9 [6], and a target dataset with 7,981 individuals, to train and test the prediction models. The heterogeneous case is exactly the same, except that we use the whole FINRISK (DILGOM07 more specifically) dataset, with 4,620 individuals, as the target data set (See Figure 1).

As suggested in [2,18], we reduce the dimensionality of the genotype in the target dataset and thus reduce the training complexity of the neural network, by applying LD clumping, which removes strongly correlated SNPs based on the GWAS summary statistics in the external dataset. The numbers of remaining SNPs for each phenotype are shown in Table 2 to 5. Moreover, we further test that our conclusions are not sensitive to the p-value threshold in LD clumping that determines the number of selected SNPs (See Section 4.4). For each experiment, we randomly split the target data into training (80%) and test (20%) sets. The genotypes are mean-centered and phenotypes standardized as part of data preprocessing. The hyper-parameters of each method are chosen via 5-fold cross-validation on the training set. We repeat each experiment with 10 different splits to evaluate the significance, and we measure the performance with the proportion of variance explained (PVE), i.e., R^2 , on the test set.

Baselines: We consider 9 different approaches: 1. linear regression (LR), specifically BLUP [3]; 2. ridge regression (Ridge), a LR with L_2 regularization; 3. lasso regression (Lasso), a LR with L_1 regularization [28]; 4. elastic net, a LR with a combination of L_1 and L_2 regularization [20]; 5. gradient boosting tree (XGBoost), implemented with XGBoost [9]; 6. neural networks (NN), a 2-hidden-layer MLP with a residual connection discussed in Section 3.1; 7. the joint effect model derived from external GWAS (Section 3.3); 8. linear regression with MEP (LR+MEP(external)), i.e., replacing the neural network $f(\mathbf{x}_i; y_i)$ in Equation 3 with a linear regression; 9. neural networks with MEP (NN+MEP(external)).

Results: We first study whether the calibration of the main effects proposed in Equation 10 is necessary. In Table 1, we show the predictive performance, in terms of the test PVE, of the joint MEP before and after the calibration on the training set. We see that even in the simple homogeneous case, where the summary statistics and the target data are both from the UK Biobank, the calibration step improves the prediction in 3 out of 6 phenotype variables. Moreover, in the heterogeneous case, where the target data set is instead from the FINRISK, the calibration significantly improves the prediction for all phenotypes considered.

Table 1. Results of the calibration experiment. Averaged test PVE (over 10 random splits) of the joint effect prior before and after the calibration step on each phenotype in the homogeneous (UK Biobank, UKB) and heterogeneous (FINRISK, FIN) cases. The best result in each experiment is boldfaced, and the corresponding statistical significance is marked as stars. We observe that the calibration step is especially beneficial in the FINRISK dataset (heterogeneous cases).

Priors	HDL.C		LDL.C		VLDL.C	
	UKB	FIN	UKB	FIN	UKB	FIN
Joint effects	0.078	0.052	0.060	0.048	0.064	0.004
Calibrated joint effects	0.078	0.055***	0.064***	0.050*	0.065	0.026***
	HDL.TG		LDL.TG		VLDL.TG	
	UKB	FIN	UKB	FIN	UKB	FIN
Joint effects	0.074	0.023	0.084	0.032	0.063	0.034
Calibrated joint effects	0.075*	0.039***	0.086*	0.040***	0.063	0.036***

¹ *, **, *** represent the p-value of corresponding test < 0.05, 0.01, 0.001 respectively.

Table 2. Comparison of the methods in the homogeneous case. Averaged test PVE (over 10 random splits) of all baseline methods on the target dataset (a subset of the UK Biobank). The dimension (p) and size of the training set (N) are shown for each phenotype. The best result in each experiment is boldfaced. We show the statistical significance of whether NN+MEP(external) is better than the second-best, with stars on the superscript, and better than without the MEP (NN), with stars on the subscript.

Phenotype (N, p)	HDL.C (5,758, 267)	LDL.C (5,758, 407)	VLDL.C (5,758, 128)	HDL.TG (5,758, 472)	LDL.TG (5,758, 493)	VLDL.TG (5,758, 274)
LR	0.049	0.034	0.056	0.006	0.02	0.026
Ridge	0.060	0.045	0.060	0.052	0.071	0.049
Lasso	0.059	0.052	0.058	0.060	0.078	0.056
Elastic Net	0.060	0.052	0.060	0.061	0.081	0.056
XGBoost	0.035	0.049	0.058	0.044	0.056	0.039
NN	0.063	0.052	0.061	0.060	0.083	0.057
MEP(external)	0.078	0.064	0.064	0.075	0.086	0.063
LR+MEP(external)	0.079	0.064	0.066	0.075	0.086	0.063
NN+MEP(external)	0.082***	0.066***	0.069**	0.076***	0.088***	0.063*

¹ The cyan parts of methods use GWAS summary statistics from the external cohort.

² *, **, *** represent the p-value of corresponding test < 0.05, 0.01, 0.001 respectively.

Next, we compare the predictive performance of the different methods in the homogeneous and heterogeneous cases in Tables 2 and 3, respectively. First, we see that our proposed NN+MEP(external) has the (shared) highest accuracy for all 6 phenotypes in both cases. To check whether the improvement is statistically significant, we conduct two one-sided t-tests on the differences between test PVEs over 10 random splits to confirm: 1. whether the best model (NN+MEP(external)) is significantly better than the second-best model, and 2. whether NN+MEP(external) is significantly better than NN, i.e., the same model but without the MEP from the external dataset. The results of these two tests are shown in Tables 2 and 3 using stars in the superscript or subscript, respectively.

Specifically to the homogeneous case (Table 2), we see that using the MEP (external) as such (i.e., just using the model trained with the external summary statistics but without any adjustment with the target data) is already better than all baselines trained on the training data without the MEP. This is because the size of the external dataset is much larger than the training data and the effect sizes are homogeneous. However, even in this homogeneous case, information from the training set is helpful, as LR+MEP(external) and NN+MEP(external) improve the prediction performance further. Furthermore, NN+MEP(external) is significantly better than LR+MEP(external) in 5 out of 6 phenotypes, which shows the effect of incorporating nonlinear main effects and interactions in the prediction model.

In the more challenging heterogeneous case (Table 3), we notice that, different from the homogeneous case, MEP(external) as such does not always perform better than models trained on the small training set, e.g., on phenotypes VLDL.C, HDL.TG, and LDL.TG, even when the main effects are calibrated. This is an instance of negative transfer: if the external cohort is very different, findings from it may not generalize to

Table 3. Comparison of the methods in the heterogeneous case. Averaged test PVE (over 10 random splits) of all baseline methods on the target dataset (FINRISK). The dimension (p) and size of the training set (N) are shown for each phenotype. The best result in each experiment is boldfaced. We show the statistical significance of whether NN+MEP(external) is better than the second-best, with stars on the superscript, and better than without the MEP (NN), with stars on the subscript.

Phenotype (N, p)	HDL.C (3,696, 205)	LDL.C (3,696, 322)	VLDL.C (3,696, 100)	HDL.TG (3,696, 384)	LDL.TG (3,696, 392)	VLDL.TG (3,696, 212)
LR	0.005	0.024	0.035	-0.038	0.040	-0.014
Ridge	0.040	0.028	0.044	0.033	0.034	0.028
Lasso	0.040	0.043	0.046	0.052	0.050	0.034
Elastic Net	0.041	0.044	0.043	0.052	0.048	0.033
XGBoost	0.024	0.045	0.022	0.040	0.038	0.012
NN	0.044	0.044	0.047	0.046	0.043	0.032
MEP(external)	0.055	0.041	0.027	0.039	0.040	0.036
LR+MEP(external)	0.055	0.056	0.046	0.050	0.054	0.041
NN+MEP(external)	0.058[*]_{***}	0.057[*]_{***}	0.048[*]_*	0.057^{**}_{**}	0.056^{***}	0.043^{***}

¹ The cyan parts of methods use GWAS summary statistics from the external cohort.

² *, **, *** represent the p-value of corresponding test < 0.05, 0.01, 0.001 respectively.

the target cohort. This could be handled with hierarchical modeling, but doing so accurately would require access to the individual-level data instead of mere publicly available summary statistics. Consequently, when both training data from the target cohort and summary statistics from the external cohort are available, NN+MEP(external) achieves the significantly best prediction performance in 4 out of 6 phenotypes, and is always significantly better than not including the external MEP.

4.3 Ablation studies

In this section, we analyze the MEP in detail with several ablation studies.

We first study if using the joint effects of summary statistics (in Section 3.3) is significantly better than using the marginal effects, i.e., effects estimated separately for each SNP, as the prior to regularize the model (NN+MEP(marginal)). Instead of defining the feature importance in Equation 4 using the joint effects, we use the marginal effects of neural network predictions, such that

$$\phi(\mathbf{w}) = D^{-1} X^T f(X; \mathbf{w}), \quad (12)$$

where D is the diagonal matrix with entries from the diagonal of $X^T X$. Moreover, we use the raw marginal effects ($\hat{\beta}, \hat{\sigma}_{\beta}$) from GWAS to construct the prior instead of the derived joint prior in the MEP:

$$\Omega'(\phi(\mathbf{w}), \phi^*) = \left\| \frac{\phi(\mathbf{w}) - \hat{\beta}}{\hat{\sigma}_{\beta}} \right\|_1. \quad (13)$$

We show the performance of NN+MEP(marginal) in Table 4. We see that although using marginal effects as the prior improves prediction, for instance, in the phenotypes LDL.TG and VLDL.TG, the improvement is rather moderate compared with using the joint effects. This is because the correlation structure of SNPs is ignored in marginal effects, and they are especially important in phenotype prediction tasks [54].

We then study the effect of removing the standard deviation $\hat{\sigma}_{\psi}$ denominator from the MEP, i.e., using

$$\Omega'(\phi(\mathbf{w}), \phi^*) = \left\| \phi(\mathbf{w}) - \hat{\phi} \right\|_1, \quad (14)$$

because the standard deviation term is often not provided in the more sophisticated constructions of the joint effects, such as LDpred [49] and PRS-CS [15]. From Table 4, we observe that NN+MEP-STD is still much better than NN without any MEP. Although NN+MEP-STD is worse than NN+MEP(external), the difference is rather small in some phenotypes such as LDL.TG and VLDL.TG. Therefore, MEP is still applicable even the standard deviation term is not available.

Table 4. Results of the ablation studies. The studies are based on the heterogeneous setup (i.e. FINRISK as the target datasets). The performance is measured by the average test PVE over 10 random splits. The dimension (p) and size of the training set (N) are shown for each phenotype. We observe that using the marginal effects as prior (NN+MEP(marginal)) only has moderate improvement compared with NN+MEP(external). Moreover, removing the standard deviation term from MEP (i.e., NN+MEP-STD) does not affect the results significantly.

Phenotype (N, p)	HDL.C (3,696, 205)	LDL.C (3,696, 322)	VLDL.C (3,696, 100)	HDL.TG (3,696, 384)	LDL.TG (3,696, 392)	VLDL.TG (3,696, 212)
NN	0.044	0.044	0.047	0.046	0.043	0.032
NN+MEP(marginal)	0.044	0.045	0.048	0.050	0.055	0.037
NN+MEP-STD	0.059	0.051	0.046	0.051	0.059	0.042
NN+MEP(empirical)	0.045	0.044	0.046	0.046	0.044	0.032
NN+MEP(external)	0.061	0.056	0.048	0.057	0.060	0.043

¹ The cyan parts of methods use GWAS summary statistics from the external cohort.

Table 5. Results of the sensitivity study w.r.t. the p-value threshold (per column) in LD clumping. The study is based on the heterogeneous setup (the HDL.C phenotype in FINRISK). The performance is measured by the average test PVE over 10 random splits. The dimension (p) and size of the training set (N) are shown for each p-value threshold. We show the statistical significance of whether NN+MEP(external) is better than the second-best, with stars on the superscript, and better than without the MEP (NN), with stars on the subscript.

p-value threshold (N, p)	5e-6 (3,696, 676)	5e-7 (3,696, 342)	5e-8 (3,696, 205)	5e-9 (3,696, 130)	5e-10 (3,696, 108)
LR	-0.040	-0.026	0.005	0.024	0.030
Ridge	0.035	0.036	0.040	0.043	0.043
Lasso	0.036	0.037	0.040	0.043	0.044
Elastic Net	0.036	0.037	0.041	0.043	0.043
XGBoost	0.023	0.023	0.024	0.024	0.024
NN	0.037	0.042	0.044	0.044	0.044
MEP(external)	0.063	0.061	0.055	0.051	0.048
LR+MEP(external)	0.063	0.060	0.055	0.052	0.052
NN+MEP(external)	0.065 [*] _{***}	0.062 _{***}	0.058 [*] _{***}	0.054 [*] _{***}	0.052 _{***}

¹ The cyan parts of methods use GWAS summary statistics from the external cohort.

² *, **, *** represent the p-value of corresponding test < 0.05, 0.01, 0.001 respectively.

Lastly, we study a heuristic empirical approach, where the joint effects are estimated using linear regression in the training dataset in the target cohort NN+MEP(empirical), instead of on the external dataset. This corresponds to the situation where not even the GWAS summary statistics from another dataset are available. As expected, the test PVE of NN+MEP(empirical) is close to that of the NN, because no additional information is incorporated.

4.4 Sensitivity analysis

In Table 5, we analyze how sensitive the performance of each method is w.r.t. the p-value threshold in the LD clumping step, which affects the number of selected SNPs, on the HDL.C phenotype in FINRISK (the heterogeneous setting).

We observe that as we increase the p-value threshold, i.e., select more SNPs, the performance of the models without the MEP decreases (even achieve negative PVEs due to the overfitting), because they fail to learn the information in the additional SNPs from insufficient training data. However, the models trained with MEP achieve better performance with more SNPs, because the information from additional SNPs has already been captured by the GWAS summary statistics of the large cohort with MEP(external). Moreover, the neural network with the MEP, NN+MEP(external), consistently outperforms others.

5 Conclusion and discussion

In this work, we focused on genotype-phenotype prediction where a small individual-level target dataset and the GWAS summary statistics from a large external dataset were available. Due to the limited amount of data, even linear regression trained on the target dataset alone often failed to generalize well, not to mention complex models such as neural networks. We introduced a main effect prior for leveraging the GWAS summary statistics from an external dataset, which are commonly available, unlike additional individual-level data. Calculating the main effect prior only requires model predictions and input features; Thus, it is widely applicable with any machine learning model, although we focused on neural networks in this work. In both simulation studies and real-world datasets, we observed that neural networks could achieve significantly better prediction accuracy with the proposed main effect prior, compared with standard baselines, even on small training datasets.

Neural networks have many attractive properties in phenotype prediction, such as their ability to model arbitrary nonlinear main effects and feature interactions, but some challenges remain for future work. First, neural networks require massive individual-level training data, which are often unavailable due to privacy considerations. Here, we proposed a tractable solution by accessing the summary statistics from another large study. However, this may not be optimal because nonlinear effects and interactions in the external set, which could potentially benefit neural networks a lot, are lost when forming standard GWAS summary statistics. Recent advances in simulating and sharing individual-level data with differential privacy [23,46] could be a possible solution. Second, it would be interesting to devise model architectures that jointly incorporate the whole genome, instead of a subset of SNPs, to predict the phenotype. Unlike linear regression or summary statistics approaches, the number of parameters required to learn in a neural network is much larger than its number of input features. Moreover, we found that the relatively simple neural network architecture used here, i.e., a fully connected MLP with residual connections, was computationally heavy for whole-genome prediction. We resolved this with the two-stage approach whereby the neural network was built on genes selected previously by other methods. New architectures, including convolutional layers [2], that automatically achieve the required sparsity and parameter sharing would thus be needed for whole-genome studies. Moreover, Bayesian approaches, e.g., PRS-CS [15], designed specifically for estimating the joint genome-wide SNP effects from their summary statistics, would be particularly beneficial for constructing the MEP. Lastly, the proposed main effect prior could be generalized to classification settings, which cover some of the attractive phenotypes, e.g., different diseases, and we leave this as another interesting future direction.

Acknowledgement

This work was supported by the Academy of Finland (Flagship programme: Finnish Center for Artificial Intelligence, FCAI, grants 319264, 292334, 286607, 294015, 336033, 315896, 341763), and EU Horizon 2020 (INTERVENE, grant no. 101016775). We also acknowledge the computational resources provided by the Aalto Science-IT Project from Computer Science IT.

References

1. Aronson, S.J., Rehm, H.L.: Building the foundation for genomics in precision medicine. *Nature* **526**(7573), 336–342 (2015) [2](#)
2. Badré, A., Zhang, L., Muchero, W., Reynolds, J.C., Pan, C.: Deep neural network improves the estimation of polygenic risk scores for breast cancer. *Journal of Human Genetics* **66**(4), 359–369 (2021) [2](#), [7](#), [11](#)
3. Bloom, J.S., Ehrenreich, I.M., Loo, W.T., Lite, T.L.V., Kruglyak, L.: Finding the sources of missing heritability in a yeast cross. *Nature* **494**(7436), 234–237 (2013) [2](#), [7](#)
4. Borodulin, K., Tolonen, H., Jousilahti, P., Jula, A., Juolevi, A., Koskinen, S., Kuulasmaa, K., Laatikainen, T., Männistö, S., Peltonen, M., Perola, M., Puska, P., Salomaa, V., Sundvall, J., Virtanen, S.M., Vartiainen, E.: Cohort Profile: The National FINRISK Study. *International Journal of Epidemiology* **47**(3), 696–696i (11 2017) [6](#)
5. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O’Connell, J., et al.: The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**(7726), 203–209 (2018) [6](#)

6. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J.: Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience* **4**(1), s13742–015 (2015) [7](#)
7. Chatterjee, N., Shi, J., García-Closas, M.: Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics* **17**(7), 392–406 (2016) [2](#)
8. Chen, M.L., Dodd, A., Royer, J., Freschi, L., Schito, M., Ezewudo, M., Kohane, I.S., Beam, A., Farhat, M.: Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in mycobacterium tuberculosis resistance prediction. *EBioMedicine* **43**, 356–369 (2019) [3](#)
9. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., et al.: Xgboost: extreme gradient boosting. R package version 0.4-2 **1**(4), 1–4 (2015) [7](#)
10. Cui, T., El Mekkaoui, K., Reinval, J., Havulinna, A.S., Marttinen, P., Kaski, S.: Gene-gene interaction detection with deep learning. *bioRxiv* (2021) [6](#)
11. Cui, T., Havulinna, A., Marttinen, P., Kaski, S.: Informative Bayesian Neural Network Priors for Weak Signals. *Bayesian Analysis* pp. 1 – 31 (2021) [3](#)
12. De Los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., Cotes, J.M.: Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* **182**(1), 375–385 (2009) [2](#)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018) [3](#)
14. Flam-Shepherd, D., Requeima, J., Duvenaud, D.: Mapping Gaussian process priors to Bayesian neural networks. In: *NIPS Bayesian deep learning workshop* (2017) [3](#)
15. Ge, T., Chen, C.Y., Ni, Y., Feng, Y.C.A., Smoller, J.W.: Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nature communications* **10**(1), 1–10 (2019) [2](#), [5](#), [9](#), [11](#)
16. Gianola, D., Fernando, R.L., Stella, A.: Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* **173**(3), 1761–1776 (2006) [2](#)
17. Goddard, M.E., Wray, N.R., Verbyla, K., Visscher, P.M.: Estimating effects and making predictions from genome-wide marker data. *Statistical Science* **24**(4), 517–529 (2009) [2](#)
18. Gola, D., Erdmann, J., Müller-Myhsok, B., Schunkert, H., König, I.R.: Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status. *Genetic epidemiology* **44**(2), 125–138 (2020) [2](#), [7](#)
19. Goodfellow, I., Bengio, Y., Courville, A.: *Deep Learning*. MIT Press (2016), <http://www.deeplearningbook.org> [3](#)
20. Grinberg, N.F., Orhobor, O.I., King, R.D.: An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat. *Machine Learning* **109**(2), 251–277 (2020) [7](#)
21. Habier, D., Fernando, R.L., Kizilkaya, K., Garrick, D.J.: Extension of the bayesian alphabet for genomic selection. *BMC bioinformatics* **12**(1), 1–12 (2011) [2](#)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016) [3](#)
23. Jälkö, J., Lagerspetz, E., Haukka, J., Tarkoma, S., Honkela, A., Kaski, S.: Privacy-preserving data sharing via probabilistic modeling. *Patterns* **2**(7), 100271 (2021) [11](#)
24. Kerminen, S., Martin, A.R., Koskela, J., Ruotsalainen, S.E., Havulinna, A.S., Surakka, I., Palotie, A., Perola, M., Salomaa, V., Daly, M.J., et al.: Geographic variation and bias in the polygenic scores of complex diseases and traits in finland. *The American Journal of Human Genetics* **104**(6), 1169–1181 (2019) [5](#)
25. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25**, 1097–1105 (2012) [3](#)
26. Krogh, A., Hertz, J.A.: A simple weight decay can improve generalization. In: *Advances in neural information processing systems*. pp. 950–957 (1992) [3](#)
27. Lees, J.A., Mai, T.T., Galardini, M., Wheeler, N.E., Horsfield, S.T., Parkhill, J., Corander, J.: Improved prediction of bacterial genotype-phenotype associations using interpretable pangenome-spanning regressions. *MBio* **11**(4), e01344–20 (2020) [2](#)
28. Lello, L., Avery, S.G., Tellier, L., Vazquez, A.I., de Los Campos, G., Hsu, S.D.: Accurate genomic prediction of human height. *Genetics* **210**(2), 477–497 (2018) [2](#), [7](#)
29. Lemhadri, I., Ruan, F., Abraham, L., Tibshirani, R.: Lassonet: A neural network with feature sparsity. *Journal of Machine Learning Research* **22**(127), 1–29 (2021) [3](#), [4](#)
30. Liu, Y., Wang, D., He, F., Wang, J., Joshi, T., Xu, D.: Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. *Frontiers in genetics* **10**, 1091 (2019) [3](#)
31. Lloyd-Jones, L.R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K.E., Wang, H., Zheng, Z., Magi, R., Esko, T., et al.: Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nature communications* **10**(1), 1–11 (2019) [2](#)
32. Loh, P.R., Kichaev, G., Gazal, S., Schoech, A.P., Price, A.L.: Mixed-model association for biobank-scale datasets. *Nature genetics* **50**(7), 906–908 (2018) [2](#)
33. de Los Campos, G., Vazquez, A.I., Fernando, R., Klimentidis, Y.C., Sorensen, D.: Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS genetics* **9**(7), e1003608 (2013) [2](#)

34. Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J., Ma, C.: A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta* **248**(5), 1307–1318 (2018) [2](#)
35. Marrugat, J., d'Agostino, R., Sullivan, L., Elosua, R., Wilson, P., Ordovas, J., Solanas, P., Cordon, F., Ramos, R., Sala, J., et al.: An adaptation of the framingham coronary heart disease risk function to european mediterranean areas. *Journal of Epidemiology & Community Health* **57**(8), 634–638 (2003) [5](#)
36. Mishra, A., McClelland, R.L., Inoue, L.Y., Kerr, K.F.: Recalibration methods for improved clinical utility of risk scores. *Medical Decision Making* p. 0272989X2111044697 (2021) [5](#)
37. Mitchell, T.: *Machine learning* (1997) [2](#)
38. Montesinos-López, A., Montesinos-López, O.A., Gianola, D., Crossa, J., Hernández-Suárez, C.M.: Multi-environment genomic prediction of plant traits using deep learners with dense architecture. *G3: Genes, Genomes, Genetics* **8**(12), 3813–3828 (2018) [2](#)
39. Montesinos-López, O.A., Martín-Vallejo, J., Crossa, J., Gianola, D., Hernández-Suárez, C.M., Montesinos-López, A., Juliana, P., Singh, R.: New deep learning genomic-based prediction model for multiple traits with binary, ordinal, and continuous phenotypes. *G3: Genes, genomes, genetics* **9**(5), 1545–1556 (2019) [2](#)
40. Moradigaravand, D., Palm, M., Farewell, A., Mustonen, V., Warringer, J., Parts, L.: Prediction of antibiotic resistance in escherichia coli from large-scale pan-genome data. *PLoS computational biology* **14**(12), e1006258 (2018) [2](#)
41. Muneeb, M., Henschel, A.: Eye-color and type-2 diabetes phenotype prediction from genotype data using deep learning methods. *BMC bioinformatics* **22**(1), 1–26 (2021) [2](#)
42. Rieger, L., Singh, C., Murdoch, W., Yu, B.: Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In: *International Conference on Machine Learning*. pp. 8116–8126. PMLR (2020) [3](#), [4](#)
43. Ross, A.S., Hughes, M.C., Doshi-Velez, F.: Right for the right reasons: training differentiable models by constraining their explanations. In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. pp. 2662–2670 (2017) [3](#), [4](#)
44. Speed, D., Balding, D.J.: Multiblup: improved snp-based prediction for complex traits. *Genome research* **24**(9), 1550–1557 (2014) [2](#)
45. Sun, S., Zhang, G., Shi, J., Grosse, R.: Functional variational Bayesian neural networks. *arXiv preprint arXiv:1903.05779* (2019) [3](#)
46. Tajeddine, R., Jätkö, J., Kaski, S., Honkela, A.: Privacy-preserving data sharing on vertically partitioned data (2020) [11](#)
47. Torkamani, A., Wineinger, N.E., Topol, E.J.: The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* **19**(9), 581–590 (2018) [2](#)
48. Tseng, A., Shrikumar, A., Kundaje, A.: Fourier-transform-based attribution priors improve the interpretability and stability of deep learning models for genomics. *Advances in Neural Information Processing Systems* **33** (2020) [3](#)
49. Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.R., Bhatia, G., Do, R., et al.: Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The american journal of human genetics* **97**(4), 576–592 (2015) [5](#), [9](#)
50. Voigt, P., Von dem Bussche, A.: *The eu general data protection regulation (gdpr). A Practical Guide*, 1st Ed., Cham: Springer International Publishing **10**, 3152676 (2017) [2](#)
51. Weinberger, E., Janizek, J., Lee, S.I.: Learning deep attribution priors based on prior knowledge. *arXiv preprint arXiv:1912.10065* (2019) [3](#), [4](#)
52. Wilson, A.G., Izmailov, P.: Bayesian deep learning and a probabilistic perspective of generalization. *arXiv preprint arXiv:2002.08791* (2020) [2](#), [3](#)
53. Würtz, P., Kangas, A.J., Soininen, P., Lawlor, D.A., Davey Smith, G., Ala-Korpela, M.: Quantitative serum nuclear magnetic resonance metabolomics in large-scale epidemiology: a primer on-omic technologies. *American journal of epidemiology* **186**(9), 1084–1096 (2017) [6](#)
54. Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Weedon, M.N., Loos, R.J., et al.: Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nature genetics* **44**(4), 369–375 (2012) [5](#), [9](#)
55. Zeng, S., Mao, Z., Ren, Y., Wang, D., Xu, D., Joshi, T.: G2pdeep: a web-based deep-learning framework for quantitative phenotype prediction and discovery of genomic markers. *Nucleic Acids Research* (2021) [2](#)
56. Zhang, Y., Bondell, H.D.: Variable selection via penalized credible regions with dirichlet-laplace global-local shrinkage priors. *Bayesian Analysis* **13**(3), 823–844 (2018) [3](#)
57. Zhu, X., Stephens, M.: Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *The annals of applied statistics* **11**(3), 1561 (2017) [2](#)