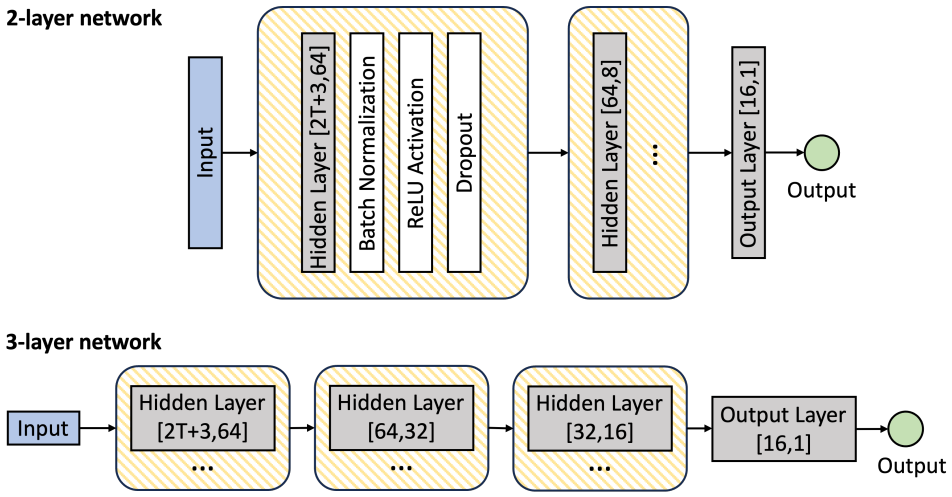

Figure S1: **Architectures of the neural network models used for shrinkage.** Input consists of observed effect $\hat{\beta}_i$, standard error se_i, LD score ℓ_i of the focal variant i, as well as observed effects $\hat{\beta}_j$ and LD values r_{ij} of the top T variants that are in highest LD (r_{ij}) with the focal variant i. Output is the regularized effect $\tilde{\beta}_i$, aiming at approximate the unknown true effect βi. Input and output dimensions of each layer are included in brackets. The dropout layer is not included in the last hidden layer in each model.

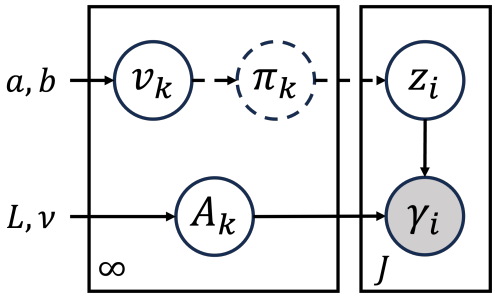

Figure S2: **A graphical representation of the zero-mean infinite-mixture (Eq. 4) using plate notation.** Nodes are random variables and edges represent the dependence between them. The observed variable is shaded. Uncircled letters are hyperparameters. Dashed lines indicate that z is dependent on v through π, but π is not explicitly included as a variational variable in the inference since it can be deterministically represented by v.

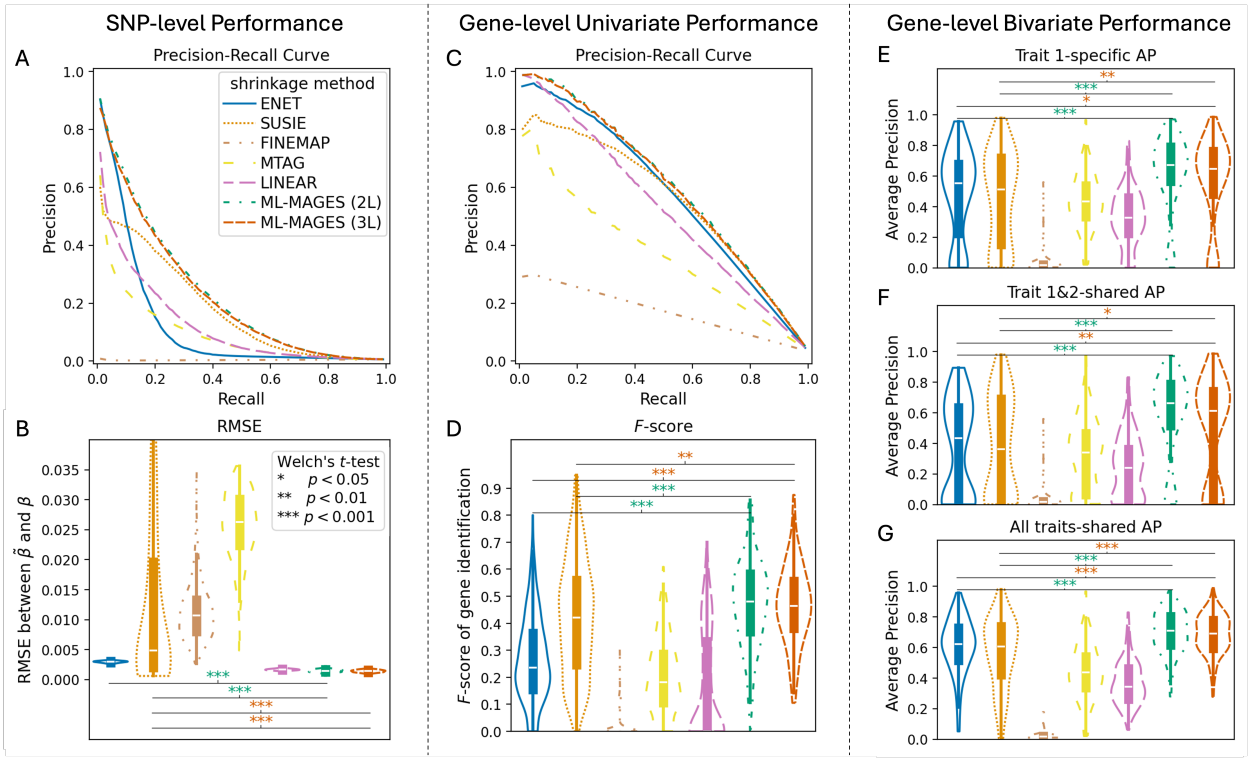

Figure S3: **Our NN methods *ML-MAGES 2L* and *3L* outperform the others in shrinking inflated GWA effect sizes and subsequently identifying associated genes in simulations with three traits.** The figure design follows that of Fig. 2 in the main text. Legends shown in panel A apply to all panels; each violin plot ordered from left to right as *ENET*, *SUSIE*, *FINEMAP*, *MTAG*, *LINEAR*, *ML-MAGES 2L* and *3L*. The significance of the comparisons using Welch's t-test are indicated on the violin plots. **Left** (SNP-level performance): comparing the regularized effects and the true effects of each simulation. **Center** (gene-level univariate performance): comparing univariate enrichment test with the simulated ground truth. **Right** (gene-level bivariate performance): comparing aggregated effect at gene-level based on bivariate association clustering output with the simulated ground truth (i.e., trait-specific vs. shared). **A**: Precision-recall curve (PRC) averaged across all 300 simulations (by interpolation), where the positives are the true non-zero effects and the precision-recall pairs are obtained by thresholding $|\tilde{\beta}|$. **B**: RMSE between β and $\tilde{\beta}$. **C**: PRC averaged across all 300 simulations (by interpolation), where the true positives are the truely associated genes and the precision-recall pairs are obtained by thresholding negative log of p-values from enrichment tests. **D**: F-score of identifying associated genes, where genes with an FDR-adjusted $p < 0.05$ from the enrichment test is identified as associated. **E**: Trait-specific average precision (AP) for identifying genes with trait-specific association to simulated trait 1, when ranking genes by the sum of absolute effects of variants in trait-specific clusters and comparing against genes being truly trait-specific. **F**: Average precision (AP) for identifying genes associated to both traits 1 and 2, when ranking the genes by the sum of absolute effects of variants in clusters of shared association between these two traits and comparing against the ground-truth. **G**: Average precision (AP) for identifying genes associated to all three traits 1, 2, and 3, when ranking the genes by the sum of absolute effects of variants in clusters of shared association among all traits and comparing against the ground-truth.

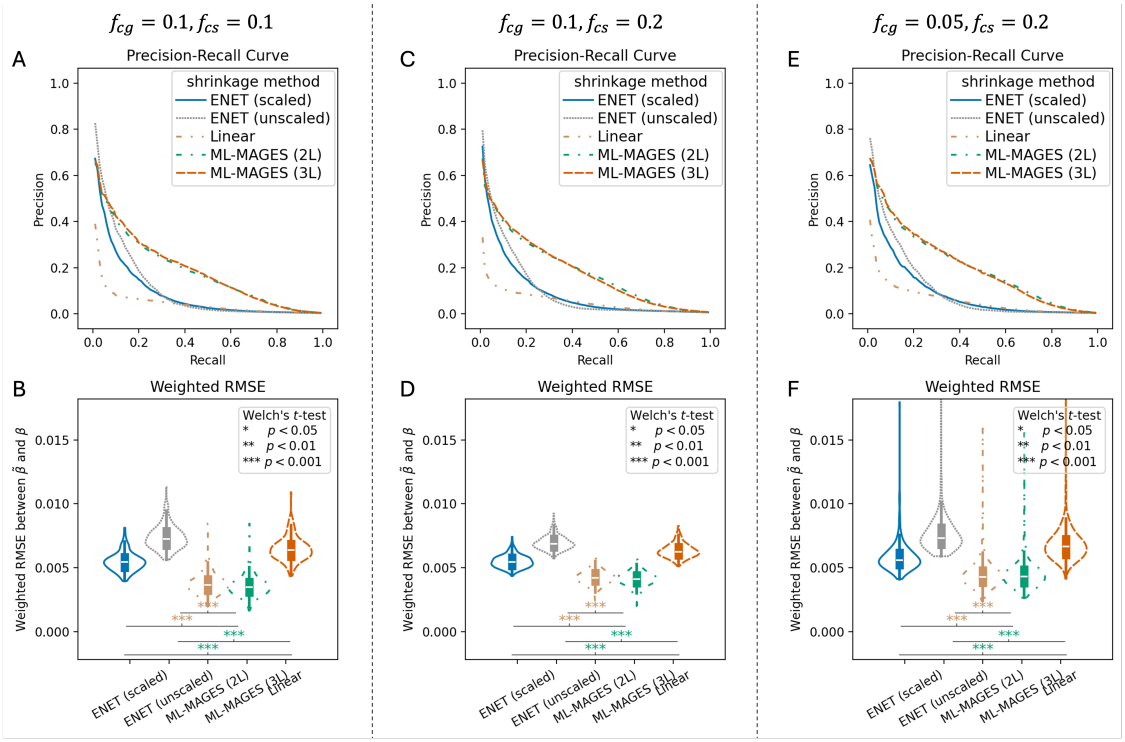

Figure S4: ***ML-MAGES 2L and 3L*** **outperform elastic net and linear neural network model on shrinking inflated GWA effect sizes when models are trained on synthetic data simulated from imputed UKB genotypes.** Outputs of 10 separately trained models for each NN architecture are ensembled to provide the final shrinkage results. The three simulation settings, **A-B**, **C-D**, and **E-F**, differ in the fraction of causal genes (f_{cg}) and the fraction of causal variants in each causal gene (f_{cs}), labeled on top of each column). The fraction of causal variants not in any causal genes is fixed to be 0.1. The two NN architectures, *ML-MAGES* (2L) and *ML-MAGES* (3L), and the single-layer neural network for comparison, labeled as *LINEAR*, are each averaged across 10 models trained independently. The performances of *ML-MAGES* (2L) and *ML-MAGES* (3L) are compared to that of *LINEAR*, as well as that of elastic net using both untransformed and transformed synthetic effects, labeled as "scaled" and "unscaled". The transformed synthetic effects are scaled to match the distribution of the summary statistics of mean corpuscular volume (MCV) from UKB and are used to construct NN inputs. The unscaled effects closely reflect the simulation that Cheng et al. (2020) used for evaluating gene-e. The figure style follows that of Fig. 2 panels A and B, where on top shows the precision-recall curves and on bottom shows the weighted RMSE, with the weights applied inversely proportional to the fraction of variants with true non-zero and zero effects. The significance of the comparisons using Welch's t-test are indicated on the bottom of the violin plots.

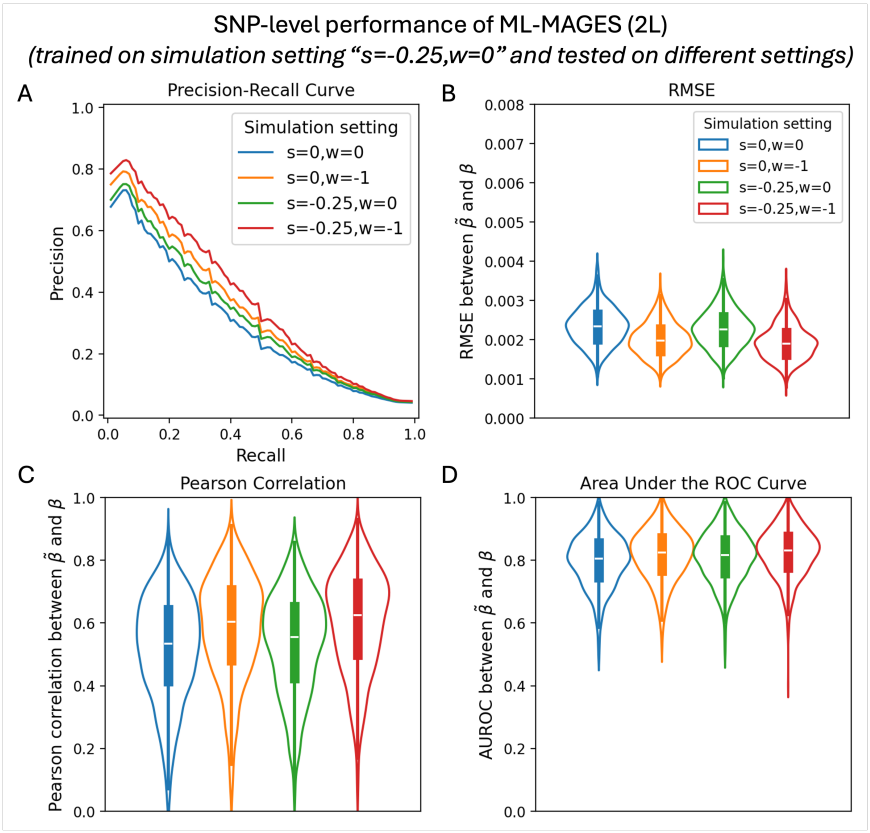

Figure S5: ***ML-MAGES* models trained on a simulation setting show robust performance on data generated under different simulation settings.** The ten ML-MAGES (2L) models are trained on data simulated with $s = -0.25$ and $w = 0$, where s controls the influence of allele frequency on effect size variance through heterozygosity, and w controls the dependence on LD score (Eq. 10), with a value of zero indicating no influence or dependence. The models are evaluated on data under the same simulation setting, as well as three other simulation settings: $s = 0, w = 0$, $s = 0, w = -1$, and $s = -0.25, w = -1$. Each setting has 400 simulations. **A**: Precision-recall curve (PRC) averaged across all 400 simulations (by interpolation), where the positives are the true non-zero effects and the precision-recall pairs are obtained by thresholding $|\tilde{\beta}|$. **B**: RMSE between β and $\tilde{\beta}$. **C**: Pearson correlation between β and $\tilde{\beta}$. **D**: Aread under the ROC curve (AUROC) for β and $\tilde{\beta}$. Model performance remains consistent across different simulation settings, indicating that our NN models are relatively robust to variation in model assumptions.

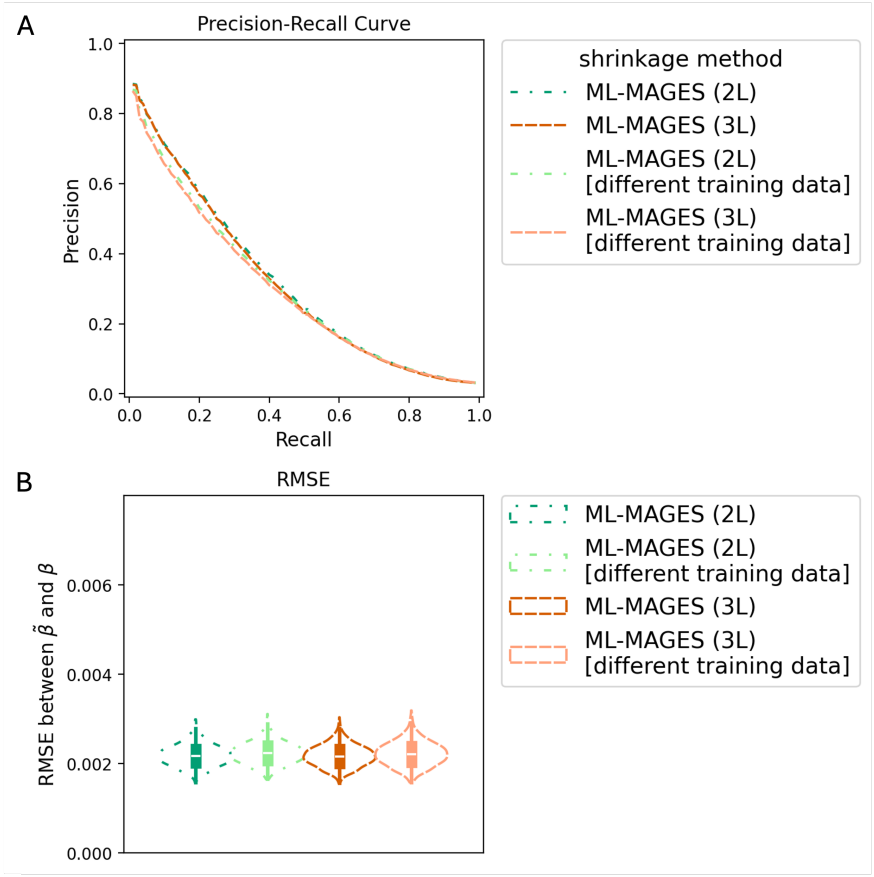

Figure S6: ***ML-MAGES* models, trained on simulations from a different set of chromosomes, demonstrate consistent performance on synthetic data.** Models marked with "different training data" were trained on simulations generated from Chromosomes 13, 14, 16, and 17, with Chromosome 15 left for validation. Models without the mark are the same ones shown in Fig. 2, which were trained on simulations generated from Chromosomes 18, 19, 21 and 22, with Chromosome 20 left for validation. Performances are evaluated on 100 simulations of effects generated on 1,000-variant segments from Chromosome 15. **A**: Precision-recall curve (PRC) averaged across all simulations, where the positives are the true non-zero effects and the precision-recall pairs are obtained by thresholding $|\tilde{\beta}|$. **B**: RMSE between β and $\tilde{\beta}$. All performances are similar across models.

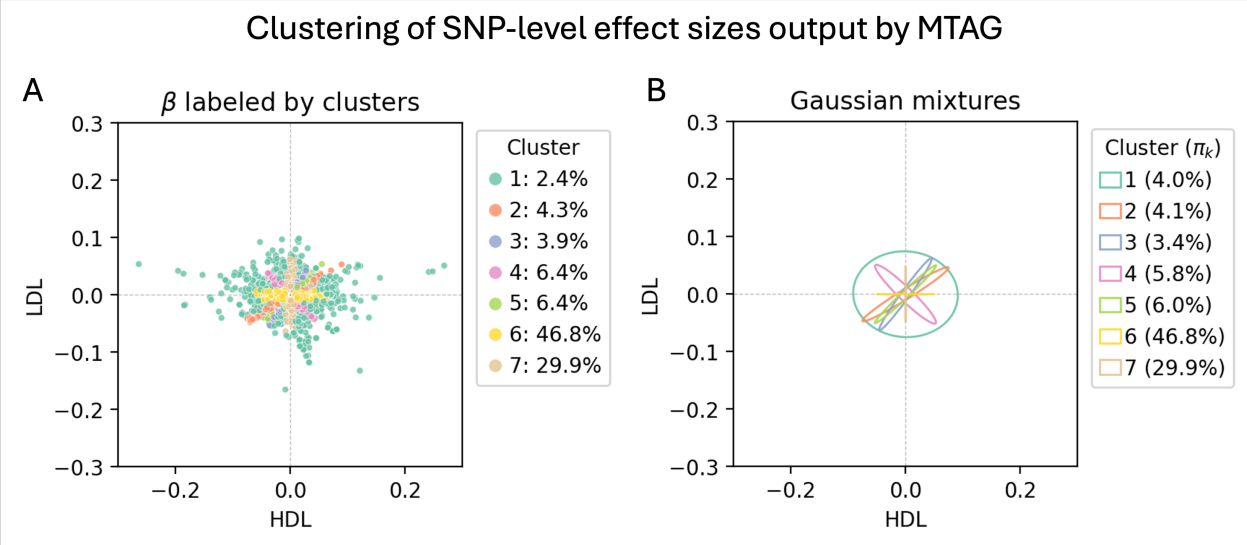

Figure S7: **Trait-specific effect sizes output by *MTAG* (Turley et al., 2018) when two traits, HDL and LDL, are analyzed jointly by the method.** **A**: Scatter plot of trait-specific effects for HDL versus those for LDL for UK Biobank genotyped variants, colored by clusters detected by the association clustering. **B**: Covariance ellipses of Gaussian mixtures representing clusters inferred from association clustering. *MTAG* does not shrink effect sizes, and its output effects are not all strictly "trait-specific", as large shared association clusters (Cls. 1–5) remain evident in the clustering results.