GEMCONT: Genetics-based Multimodal Contrastive Learning Enhances Phenotypic embeddings and Boosts Genetic Discovery

Daniel Sens[1,2,3] DANIEL.SENS@HELMHOLTZ-MUNICH.DE
Liubov Shilova[1,2,3,4] LIUBOV.SHILOVA@HELMHOLTZ-MUNICH.DE
Adrian V. Dalca[5,6] ADALCA@MIT.EDU
Julia A. Schnabel[3,7,8,*] JULIA.SCHNABEL@HELMHOLTZ-MUNICH.DE
Francesco Paolo Casale[1,2,3,*] FRANCESCOPAOLO.CASALE@HELMHOLTZ-MUNICH.DE

[1] *Institute of AI for Health, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany*
[2] *Helmholtz Pioneer Campus, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany*
[3] *School of Computation, Information and Technology, Technical University of Munich, Garching, Germany*
[4] *Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany*
[5] *A.A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA*
[6] *Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA*
[7] *Institute of Machine Learning in Biomedical Imaging, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany*
[8] *School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK*

Editors: Accepted for publication at MIDL 2026

Abstract

Genetic variation provides stable, time-invariant markers of disease risk and can therefore reveal upstream mechanisms underlying complex traits. Genome-wide association studies (GWAS) have identified thousands of loci associated with disease, yet most remain difficult to interpret because the intermediate phenotypes linking genotype to disease are unknown. Here, we address the question whether disease-associated genetic loci can be directly used to extract such risk-related features from quantitative phenotypes, including functional tests and medical imaging. We introduce **GEMCONT** (GEnetics-based Multimodal CONTrastive Learning), a multimodal contrastive learning framework that aligns genotype and phenotype representations in a shared latent space. Unlike task-agnostic multimodal pretraining, GEMCONT is disease-conditioned: GWAS-informed variant panels act as targeted supervision to learn risk-relevant imaging embeddings. To reflect the weak, additive nature of genetic effects, it employs a linear genetic encoder alongside a deep phenotypic encoder. We validate GEMCONT in controlled simulations and apply it to two real-world settings: spirometry curves for asthma and retinal fundus images for glaucoma. In both, GEMCONT improves disease risk prediction and enhances recovery of genetic associations compared with standard unsupervised or polygenic risk–based models.

* Corresponding authors.

Altogether, our results demonstrate that incorporating stable genetic supervision into multimodal representation learning enables the extraction of genetically informed risk traits, refining disease phenotypes and improving the interpretability of association studies.

Keywords: Multimodal Contrastive Learning, Imaging Genetics, Genome-Wide Association Studies, Machine Learning–Derived Phenotypes, Medical Imaging

1. Introduction

Genome-wide association studies (GWAS) have identified thousands of genetic loci associated with human diseases and complex traits (Visscher et al., 2017; Manolio et al., 2009). Because germline variation is fixed and precedes disease onset, genetic associations provide upstream information about biological mechanisms. Yet for most loci, the functional link between genotype and phenotype remains unknown (Tam et al., 2019). Recent work in imaging genetics has begun to address this challenge by using high-dimensional biomedical data—such as medical images or physiological recordings—to derive quantitative phenotypes that better reflect underlying biology (Wright and Herzberg, 2021; Tracy, 2008; Robinson, 2012). Early approaches relied on manually defined regions or handcrafted measurements, while more recent studies leverage machine learning to learn compact phenotypic representations directly from raw data (Zech et al., 2018). These representations have proven valuable for association studies: for instance, supervised networks trained on clinical outcomes can uncover novel genetic loci (Rakowski et al., 2024; Kirchler et al., 2022), and unsupervised embeddings can reveal heritable structure (Yun et al., 2024; Xie et al., 2024). However, unsupervised representation learning tends to capture the dominant axes of variation in the data and may overlook disease-related effects when these correspond to more subtle or less frequent patterns (Shilova et al., 2025).

To address these challenges, we introduce **GEMCONT** (GEnetics-based Multimodal CONTrastive learning), a multimodal contrastive learning framework for imaging-genetics analysis (Fig. 1). GEMCONT aligns medical imaging data with disease-associated genetic variants in a shared latent space for a given disease. Through this alignment, GEMCONT learns disease-specific imaging embeddings that capture risk-relevant variation and are predictive of future disease. By focusing on disease-specific, risk-aligned embeddings, GEMCONT differs from prior multimodal frameworks such as ContIG (Taleb et al., 2022) and MRM (Yang et al., 2023), which use genetic or molecular modalities for task-agnostic multimodal pretraining followed by downstream fine-tuning. The contributions of this work are threefold:

1. **Genetics-informed contrastive learning.** We adapt multimodal contrastive learning to disease-focused imaging-genetics analysis through GEMCONT, which (i) selects disease-associated variants from external GWAS summary statistics to define targeted genetic supervision, and (ii) employs a linear genetic projector for efficient and interpretable variant contributions (Fig. 1).

2. **Benchmarking across genetic architectures.** We benchmark GEMCONT using controlled simulations with known disease-associated latent traits and causal variants, evaluating performance across sample sizes and genetic architectures to determine when contrastive alignment improves latent trait recovery.

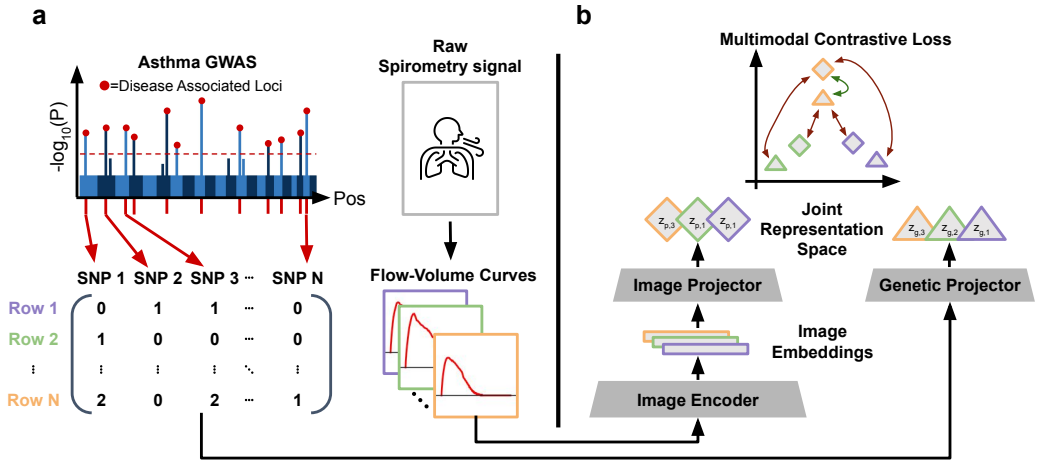

Figure 1: **Application of GEMCONT to asthma variants and spirometry images.** **(a)** Genetic variants significantly associated with asthma are extracted from the imputed UK Biobank data. **(b)** Corresponding raw spirometry signals are converted into Flow-Volume Curve images, and an asymmetric dual-encoder is trained to align genotype and image embeddings in a joint representation space through the multimodal contrastive loss.

3. **Applications to population imaging.** We apply GEMCONT to two use cases in the UK Biobank (Sudlow et al., 2015). First, we recover asthma-related spirometry embeddings by integrating asthma-associated variants with flow–volume curve data. Second, we recover a glaucoma-related latent trait from retinal fundus images using glaucoma-associated variants and imaging data. In both settings, genetics-guided contrastive learning improves disease risk prediction and strengthens genetic association analyses compared to standard self-supervised approaches.

Together, these results establish contrastive learning with genetic supervision as a principled approach for constructing disease-specific, GWAS-aware phenotypes from high-dimensional medical imaging data.

2. Methods

The approach implemented in GEMCONT can be formulated as two-step process: (i) learning a joint representation space where genetic and imaging-derived features are co-embedded, and (ii) validating the extracted imaging embeddings in statistical association analyses and disease classification. Below, we describe the co-embedding pipeline and the statistical validation procedures.

2.1. Contrastive Learning for Genetics-Image Alignment

Contrastive learning for genetics-imaging alignment builds on the CLIP (Radford et al., 2021) framework, introducing key adaptations to address the unique challenges of genetic data, which are characterized by sparse and weak additive effects on phenotypes. Formally, given a dataset of paired genotype and phenotype samples $\mathcal{D} = \{(x_{g,i}, x_{p,i})\}_{i=1}^{N}$, each genotype sample $x_g \in \{0,1,2\}^S$ represents an allele count vector of S disease-associated variants, while x_p denotes a high-content phenotype (e.g., medical images). Following the principle implemented in the CLIP model, GEMCONT learns joint embeddings of genetic and imaging data by maximizing agreement between modalities from the same individual while encouraging separation between individuals. This objective enables the co-embedding of genetic and phenotypic features into a shared latent space, facilitating the discovery of biologically meaningful genotype-phenotype relationships.

Multimodal Contrastive Learning Objective. GEMCONT processes genotype and phenotype data through modality-specific encoders (Fig. 1). The phenotype encoder f_{θ_p} maps x_p to an intermediate embedding e_p, which is then projected onto a unit-norm latent space as z_p. The genotype data are processed through a linear genotype projector, mapping x_g to a unit-norm latent embedding z_g. During training, we sample a mini-batch $\mathcal{B} \subset \mathcal{D}$ of paired genotype–phenotype samples and optimize the phenotype encoder, phenotype projector, and genotype projector to align genetic and phenotypic projections by maximizing similarity within individuals while minimizing it across individuals. This is achieved using the multimodal contrastive loss (Radford et al., 2021):

$$\mathcal{L} = \frac{1}{2}\left(\mathcal{L}_{g\to p} + \mathcal{L}_{g\leftarrow p}\right), \tag{1}$$

where

$$\mathcal{L}_{g\to p} = -\sum_{j\in\mathcal{B}} \log \frac{\exp(z_{g,j}^T z_{p,j}/\tau)}{\sum_{k\in\mathcal{B},\, k\neq j} \exp(z_{g,j}^T z_{p,k}/\tau)}, \tag{2}$$

and $\mathcal{L}_{g\leftarrow p}$ is defined analogously, swapping g and p. Similar to CLIP, $\tau > 0$ is a learnable temperature parameter.

Adaptations for genetic data. To address the sparsity and additive nature of genetic effects, GEMCONT introduces two key adaptations:

1. **Selection of informative variants.** We extract relevant genetic features from genome-wide association study (GWAS) summary statistics, which quantify associations (e.g., p-values, effect sizes) between millions of variants and a disease of interest. From these statistics, we select independent genome-wide significant loci through a clumping procedure (Purcell et al., 2007), iteratively retaining the most significant variant while removing correlated neighbors within a 5 Megabase (Mb) window.

2. **Linear projection of genotypes.** Genetic effects are predominantly additive with limited evidence for interactions between variants (Hill et al., 2005), making a linear projection sufficient for mapping selected variants into the latent space. This reduces model complexity while preserving key genetic signals.

2.2. Genetic Association Analysis of Learned Embeddings

Multi-trait GWAS for embedding analysis. To assess whether GEMCONT-derived embeddings capture meaningful genetic signals, we perform a multi-trait genome-wide association study (GWAS) on a held-out set of samples not used for training. We adapt the single-variant model from (Lippert et al., 2014; Casale et al., 2015), modeling each embedding dimension as a quantitative trait influenced by genetic variation. Let $\boldsymbol{E} \in \mathbb{R}^{N\times D}$ be the embedding matrix, $\boldsymbol{g} \in \{0,1,2\}^{N\times 1}$ a genotype vector, and $\boldsymbol{F} \in \mathbb{R}^{N\times K}$ a matrix of covariates. The model is:

$$\boldsymbol{E} = \boldsymbol{F}\boldsymbol{A} + \boldsymbol{g}\boldsymbol{b}^T + \boldsymbol{\Psi}, \tag{3}$$

where $\boldsymbol{A} \in \mathbb{R}^{D\times K}$ and $\boldsymbol{b} \in \mathbb{R}^{D\times 1}$ capture covariate and genetic effects, respectively, and $\boldsymbol{\Psi} \sim \mathcal{N}(0, \boldsymbol{C})$ models residual noise with a learnable covariance matrix $\boldsymbol{C} \in \mathbb{R}^{D\times D}$. Following (Lippert et al., 2014; Casale et al., 2015), $\boldsymbol{C}$ is estimated under the null model, while a single scaling factor per variant is optimized under the alternative model to control false positives efficiently (Korte et al., 2012). We test whether $\boldsymbol{b} \neq 0$, obtaining p-values via a likelihood ratio test with D degrees of freedom. For simulations (Sec. 3.1), we do not adjust for covariates. In both real-world applications (Sec. 3.2, Sec. 3.3), all embedding-GWAS tests adjust for genotyping array, assessment center, sex, age, age^2, sex-by-age, sex-by-age^2, height, height^2, BMI, and the top 20 genetic principal components. To address feature correlation and non-Gaussian distributions, embeddings are projected onto their top D principal components and rank-normalized before association testing. Independent genome-wide significant loci ($p < 5 \times 10^{-8}$) are identified using PLINK's clumping procedure (Purcell et al., 2007), which retains only approximately independent genetic associations by removing variants in linkage disequilibrium ($r^2 < 0.05$) within a 5 Mb window.

Assessing overlap with disease GWAS. To evaluate whether GEMCONT-derived embeddings capture known disease-associated genetic signals, we compare the genomic loci identified in our embedding-based GWAS to those from a standard disease GWAS. Specifically, we measure the fraction of independent genome-wide significant loci ($p < 5 \times 10^{-8}$) identified in the disease GWAS that are also recovered at genome-wide significance in the embedding GWAS. This assessment is performed on a held-out test set, distinct from the training data used for learning embeddings.

External disease GWAS and meta-analysis. To define the disease-specific variant panels used by GEMCONT, we rely on large external genome-wide association studies (GWAS) for asthma and glaucoma from the Million Veteran Program (Verma et al., 2023) and FinnGen (Kurki et al., 2023). For each disease, we harmonize summary statistics across cohorts and combine them using a fixed-effect inverse-variance meta-analysis in METAL (Willer et al., 2010). From the resulting meta-analytic GWAS, we selected variants with association $p < 10^{-5}$ and applied LD clumping in PLINK (5 Mb window, $r^2 < 0.05$) (Purcell et al., 2007), yielding an approximately independent set of disease-enriched SNPs. The 10^{-5} threshold is an intermediate cut-off that has been used when selecting variants from GWAS loci for Mendelian randomization analyses (Davey Smith and Hemani, 2014; Jin et al., 2024) and lies within the range of p-value thresholds typically explored in clumping-and-thresholding polygenic risk score methods (Choi et al., 2020). This choice provides a panel of variants that is strongly enriched for disease-associated signal while remaining sufficiently large to supervise the phenotype encoder.

3. Experiments and Results

We evaluate GEMCONT's ability to (i) enhance genetic signal for phenotype- or disease-associated variants and (ii) recover the underlying latent phenotype in simulations, and via disease risk prediction as a proxy in real data applications. We conduct three experiments: a controlled simulation study to assess performance under varying genetic architectures and two real-world applications to UK Biobank (Sudlow et al., 2015) data. In the first application we analyze flow-volume curves - used in asthma diagnosis (Jayasooriya et al., 2023) - and integrate genetic variants associated with asthma. In the second we apply our framework to retina fundus images, which are used in glaucoma diagnosis (Saha et al., 2023), and co-embed them with variants associated with glaucoma. We additionally report robustness analyses comparing GEMCONT with linear versus nonlinear genetic projectors and evaluating sensitivity to the GWAS variant-selection threshold (Sec. 3.4; Tab. 1).

Compared methods. In the simulation and spirometry experiments, we compare GEMCONT against two established self-supervised embedding methods: a variational autoencoder (VAE), which learns latent representations by optimizing a reconstruction objective under a latent prior (Kingma and Welling, 2014), and SimCLR (Chen et al., 2020), a contrastive learning approach that maximizes agreement between augmented views of the same input. For the fundus application, we leverage the retina foundation model RetFound (Zhou et al., 2023, 2025), a large-scale self-supervised model pretrained on nearly one million fundus images, which provides a strong unsupervised reference without the need to retrain SimCLR or VAE baselines. In the real-data applications, we additionally include baselines tailored to disease prediction. First, we use a simple multimodal model in which the genetic branch is reduced to a single polygenic risk score (PRS) for the target disease, computed as the sum of GEMCONT's input variants weighted by their GWAS effect sizes and fed as a univariate input to the genetic projector. We refer to this as the PRS baseline. Second, to assess whether genetics-driven phenotype embeddings provide added value over conventional clinical markers, we benchmark all spirometry models against the $\mathrm{FEV}_1/\mathrm{FVC}$ ratio and all fundus models against the cup-to-disc ratio, both widely used functional (Lambert et al., 2015) and imaging-derived (Gordon et al., 2002; Foster et al., 2002) biomarkers in their respective diagnostic domains. Finally, in the fundus experiment we leverage a strong retinal foundation model (a DINOv2-pretrained ViT backbone) and consider three configurations on top of it: a frozen RetFound (Zhou et al., 2023, 2025) baseline, GEMCONT, and a supervised upper-bound model (Sec. 3.3).

3.1. Benchmarking in Simulated Data

Setup. To evaluate GEMCONT in a controlled setting, we design a simulation framework where a latent genetic trait influences imaging features, mimicking real-world genotype-phenotype relationships. We use EMNIST (Cohen et al., 2017), a dataset of 814,255 grayscale handwritten characters across 62 classes, and define the latent phenotype as the rotation angle of each character. We systematically vary key factors: training set size (N_{train}), genetic variance explained (h_g), and phenotype transformation strength ($\alpha_{\max}$). A fixed 100K test set is used across all experiments, with five random splits for training, where N_{train} is varied (default: 100K). After training, we extract image embeddings and evaluate:

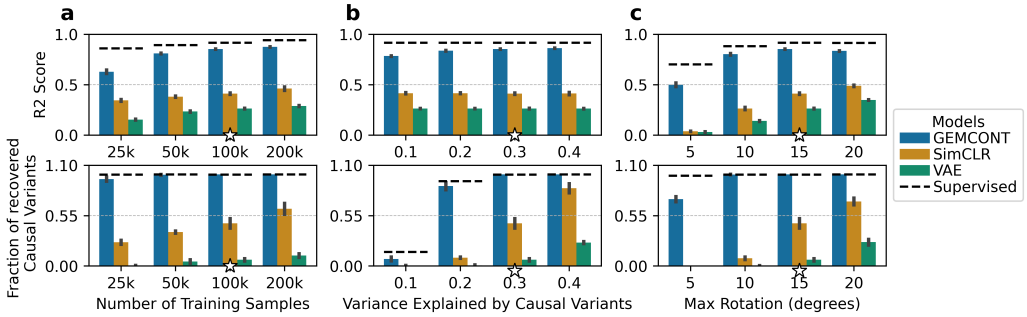

Figure 2: **Evaluation of phenotype and genetic signal recovery.** Performance of GEMCONT, SimCLR, VAE, and a supervised model is assessed under varying training set sizes (a), genetic variance explained (b), and maximum rotation (c). The first row shows the R^2 score for predicting the latent phenotype z from the learned embeddings, while the second row presents the fraction of causal variants identified at genome-wide significance ($p < 5 \times 10^{-8}$). Each bar represents the mean ± standard deviation across five random splits. Stars denote standard values held constant while other parameters were varied.

1. **Phenotype recovery:** Predicting z from embeddings using ridge regression, measured by R^2 on the test set.

2. **Genetic recovery:** Identifying genome-wide significant variants ($p < 5 \times 10^{-8}$) via multi-trait GWAS (Sec. 2.2).

Simulation strategy. We first subsample N images stratified across character labels. Next, we simulate a genotype matrix $\boldsymbol{G} \in \{0, 1, 2\}^{N \times S}$ for S variants, where each entry represents allele counts drawn from a binomial distribution: $\boldsymbol{G}_{i,j} \sim \text{Binomial}(2, f_j)$ with minor allele frequency f_j. The latent phenotype $\boldsymbol{z}$ is then generated as a weighted combination of genetic effects and environmental noise, controlling the proportion of variance explained by genetics (h_g):

$$\boldsymbol{z} = \sqrt{h_g} \cdot \widetilde{\boldsymbol{G}\boldsymbol{\beta}} + \sqrt{1 - h_g} \cdot \widetilde{\boldsymbol{z}_n}, \tag{4}$$

where $\widetilde{\boldsymbol{x}} = \frac{\boldsymbol{x} - \mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})}$ denotes the standardized version of a vector $\boldsymbol{x}$, with $\mu(\cdot)$ and $\sigma(\cdot)$ denoting the mean and standard deviation of its elements, respectively. Here, each variant effect size $\boldsymbol{\beta}$ is sampled from $\{-1, 1\}$ with equal probability, and $\boldsymbol{z}_n \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_N)$ models environmental noise. We then define rotation angles as $\boldsymbol{\alpha} = \alpha_{\max} \cdot \tanh(c \cdot \boldsymbol{z})$, where c is chosen to prevent saturation of tanh across samples. Each image is rotated by its corresponding α_j, creating a dataset of genetic-image pairs directly linked through rotation.

Results. Figure 2 summarizes the results. GEMCONT outperforms SimCLR and VAE across all settings. Performance saturates beyond 100K training samples, though GEMCONT maintains a significant advantage (Fig. 2a). Genetic variance (h_g) has minimal impact on phenotype recovery but strongly affects variant detection, with GEMCONT consistently identifying more causal variants (Fig. 2b). Finally, lower rotation angles ($\alpha_{\max}$) degrade baseline performance more than GEMCONT, which remains robust across conditions

(Fig. 2c). As expected, a supervised model serves as an upper bound for both phenotype and variant recovery.

3.2. Application to Spirometry and Asthma

Experimental setup. We generate flow–volume curves following (Yun et al., 2024) and compute the FEV_1/FVC ratio, a key biomarker for asthma diagnosis (Lambert et al., 2015). Similar to recent work that applies CNNs directly to images of spirometry flow–volume curves for quality control (Martins et al., 2025; Wang et al., 2022), we rasterize each trajectory into a standardized 256×256 grayscale image at 200 dpi and use these images as input to the encoder. Genetic variants associated with asthma are selected from external GWAS summary statistics using clumping (Sec. 2.2), yielding 551 approximately independent variants. To ensure that spirometry curves reflect baseline lung function, we exclude participants who reported using a chest inhaler or smoking a cigarette within the last hour before testing, in line with clinical spirometry preparation guidelines (Paraskeva et al., 2011). After matching imaging and genetic data and restricting to individuals of European ancestry, we retain 227,332 participants. To obtain stable estimates and quantify variability, we perform five random 50/50 train/validation splits and evaluate (i) disease recovery and (ii) genetic signal enrichment in the embeddings. Disease recovery is assessed using L2-regularized logistic regression to predict asthma from pre-spirometry diagnosis and diagnosis within five years post-assessment, reporting ROC AUC. Genetic signal enrichment is quantified by performing multi-trait GWAS on the embeddings and computing the fraction of independent asthma-associated variants that remain significant after Bonferroni correction (Sec. 2.2). Figure 3 summarizes the results.

Results. GEMCONT achieves the highest recall of asthma-associated loci, though all models recover only a small fraction (Fig. 3a), consistent with expectations given our smaller sample size relative to the effective sample size of the GWAS meta-analysis. For asthma prediction, GEMCONT significantly ($p < 0.05$) outperforms the compared models at baseline and approaches the supervised model's upper bound for future diagnoses (Fig. 3b). Finally, Fig. 3c displays violin plots of the first principal component of the image embeddings (PC1) and the FEV_1/FVC ratio, stratified by asthma status; both measures show modest distributional shifts between cases and controls, indicating that PC1 captures asthma-related variation that is comparable in magnitude to the classical biomarker but derived directly from the flow–volume curves.

3.3. Application to Fundus Images and Glaucoma

Experimental setup. We analyzed color fundus images from the first imaging visit of UK Biobank participants (Sudlow et al., 2015) and filtered images using the MCF-Net model (Fu et al., 2019), excluding images with a rejection probability above 80%. We further excluded fundus images from participants who reported prior surgery or laser treatment for glaucoma, as this affects biomarkers for glaucoma risk and can impact fundus morphology (Lesk et al., 1999; Raghu et al., 2012; Pillunat et al., 2023). Genetic variants associated with glaucoma were selected from external GWAS summary statistics using clumping (Sec. 2.2), yielding 1,535 approximately independent variants. After merging with genetic data for individuals of European ancestry, we retained 36,349 participants with at least one usable fundus image.

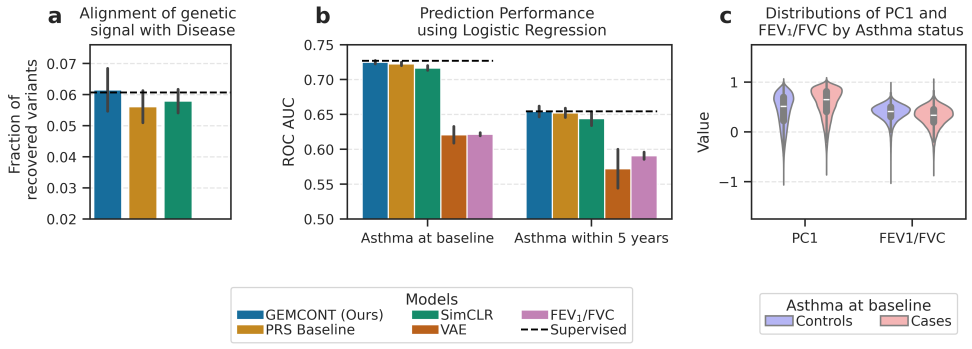

Figure 3: **Asthma prediction and genetic signal enrichment.** Comparison of GEMCONT, PRS baseline, SimCLR, VAE, and a supervised model trained on the latent phenotype. (a) Fraction of independent genome-wide significant asthma-associated variants recovered in method-specific GWAS after multiple testing correction. (b) ROC AUC for asthma classification at baseline and within 5 years post-assessment. (c) Distributions of the first principal component (PC1) of GEMCONT image embeddings and of FEV_1/FVC, stratified by asthma status. Violin plots show normalized values for controls (blue) and cases (red), with black bars indicating median and inter-quartile range. Results are mean ± standard deviation across 5 random 50/50 splits.

For individuals with two images, we randomly sampled left or right eye with equal probability during training whenever the individual was drawn into a batch. During validation, if both eyes were available, we extracted image embeddings for each eye and used their mean as the final embedding. We build on a Vision Transformer (ViT) base encoder pretrained using DINOv2 on retinal images (Zhou et al., 2023, 2025), which we keep frozen and use as an online feature extractor during training (Kolesnikov et al., 2020; Vo et al., 2025). Standard image augmentations are applied before feeding inputs through the frozen encoder (Sec. 3.5). On top of this backbone, we consider three image-based configurations. First, RetFound uses the frozen ViT features with simple mean pooling over patch tokens; no additional representation learning is performed, and the resulting embeddings are used directly in downstream GWAS and logistic regression. Second, GEMCONT adds an attention-pooling layer (Ilse et al., 2018) and a lightweight two-layer MLP to map pooled features to image embeddings, which are then aligned with genetic embeddings using the multimodal contrastive objective. Third, a supervised model shares the same architecture as GEMCONT but is optimized directly for glaucoma classification, providing an approximate upper bound. As in the spirometry experiment (Sec. 3.2), we perform five random 50/50 train/validation splits and evaluate (i) disease recovery and (ii) genetic signal enrichment in the embeddings. Disease recovery is assessed using L2-regularized logistic regression to predict glaucoma from diagnosis by the time of image acquisition and diagnosis within five years post-assessment, reporting ROC AUC. Genetic signal enrichment is quantified via multi-trait GWAS on the embeddings, measuring the fraction of independent glaucoma-associated variants that remain significant after Bonferroni correction (Sec. 2.2). Figure 4 summarizes the results.

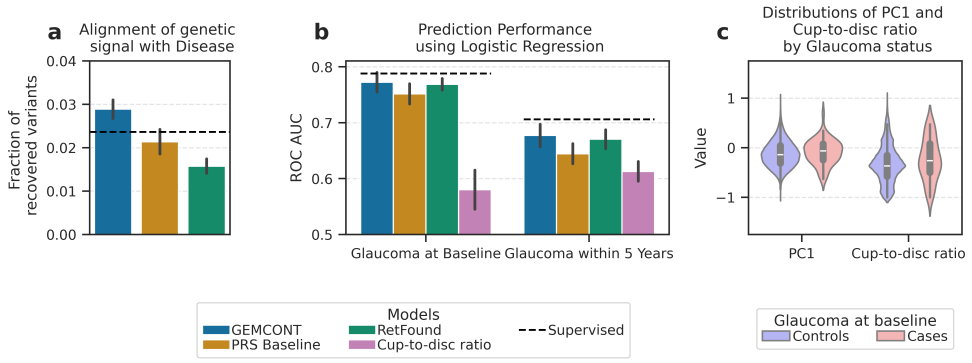

Figure 4: **Glaucoma prediction and genetic signal enrichment.** Comparison of GEMCONT, a PRS-only baseline, RetFound embeddings, a supervised upper-bound model (dashed line), and the cup-to-disc ratio clinical biomarker. (a) Fraction of independent genome-wide significant glaucoma-associated variants recovered after multiple testing correction. (b) ROC AUC for glaucoma classification at baseline and within 5 years post-assessment. (c) Distributions of the first principal component (PC1) of GEMCONT image embeddings and of cup-to-disc ratio, stratified by glaucoma status. Violin plots show normalized values for controls (blue) and cases (red), with black bars indicating median and inter-quartile range. Results are mean ± standard deviation across 5 random 50/50 splits.

Results. GEMCONT recovers the largest fraction of independent glaucoma-associated variants, outperforming the PRS and RetFound baselines as well as the supervised upper bound in terms of alignment between embeddings and disease loci (Fig. 4a). For disease prediction, GEMCONT is competitive with the supervised model and consistently outperforms both the PRS baseline and the cup-to-disc ratio for glaucoma at image acquisition and for incident glaucoma within five years (Fig. 4b). Finally, both the first principal component of the GEMCONT embeddings and the cup-to-disc ratio show case–control shifts, with PC1 exhibiting slightly stronger separation, indicating that the learned representation captures glaucoma-related variation that is at least comparable to this established imaging biomarker (Fig. 4c).

3.4. Robustness and Sensitivity Analyses

We conducted robustness analyses to evaluate two key design choices: (i) the use of a linear genetic projector and (ii) the GWAS variant-selection threshold. Replacing the linear genetic projector with a two-layer MLP (batch normalization + ReLU) did not produce consistent gains in disease-prediction AUC or genetic-signal recovery across asthma and glaucoma, supporting the linear inductive bias under weak and sparse genetic effects (Tab. 1). Likewise, applying a stricter GWAS-significant supervision panel ($p < 5 \times 10^{-8}$, LD-clumped) resulted in comparable or slightly lower AUC and no systematic improvement in recovery relative to the default $p < 10^{-5}$ panel (Tab. 1).

Table 1: **GEMCONT ablation and sensitivity analyses on asthma (spirometry) and glaucoma (fundus).** ROC AUC is reported for baseline diagnosis and diagnosis within 5 years. Genetic recovery is the fraction of independent loci from the corresponding external disease GWAS recovered as significant in embedding-GWAS at $p < 5 \times 10^{-8}$ (held-out data). Values are mean ± std across 5 random splits.

Experiment	Method (setting)	ROC AUC		Genetic recovery
		baseline	within 5 years	
Asthma (spirometry)	GEMCONT	0.725 ± 0.002	0.654 ± 0.008	0.062 ± 0.007
	GEMCONT (MLP)	0.723 ± 0.003	0.654 ± 0.008	0.060 ± 0.005
	GEMCONT (strict SNP panel)	0.723 ± 0.002	0.652 ± 0.008	0.057 ± 0.005
Glaucoma (fundus)	GEMCONT	0.772 ± 0.017	0.677 ± 0.020	0.029 ± 0.002
	GEMCONT (MLP)	0.768 ± 0.010	0.670 ± 0.020	0.033 ± 0.004
	GEMCONT (strict SNP panel)	0.763 ± 0.011	0.667 ± 0.021	0.029 ± 0.003

3.5. Implementation Details

All models were implemented in PyTorch (Paszke et al., 2019) and trained for 150 epochs with a batch size of 1024 using AdamW (Loshchilov and Hutter, 2017) (base learning rate 3×10^{-4}, weight decay 1×10^{-4}) and a cosine-annealing schedule with a 10-epoch warm-up. For all models and experiments, we set the embedding dimension to $D = 256$ and used a single linear layer as the phenotype projector (output dimension $\mathbb{R}^{D/2}$). Embeddings are ℓ_2-normalized before computing similarities. The temperature τ in the multimodal contrastive loss is implemented as a learnable scalar (initialized to $\tau_0 = 0.07$). For the supervised baselines we additionally applied early stopping on the validation loss with a patience of 50 epochs. To control for population structure, we regress out the top 20 genetic PCs from both single-variant dosages and PRS features before training, following standard UK Biobank practice (Bycroft et al., 2018; Canela-Xandri et al., 2018). For the genetic branch, weights connected to each input variant were initialized to the corresponding effect size from the external meta-analytic GWAS, and genetic inputs were augmented using SCARF (Bahri et al., 2021) with corruption probability $p = 0.1$. Image augmentations were adapted to each experiment: random erasing for the EMNIST simulation to preserve the rotation signal, random resized crops with Gaussian blur for spirometry, and random resized crops, color jitter, and Gaussian blur for fundus images. Training was performed on a single NVIDIA H100 (80GB) GPU; a typical GEMCONT run (150 epochs, batch size 1024) takes ∼10 hours wall-clock, with runtime dominated by the image branch (particularly in fundus).

4. Conclusion and Future Work

We introduced GEMCONT, a genetics-based multimodal contrastive learning framework that aligns genotype and imaging embeddings to emphasize disease-relevant variation. By leveraging disease-associated genetic variants as supervision, GEMCONT learns imaging representations predictive of future disease risk, positioning genetics as a biologically grounded

supervisory signal for medical imaging. The framework contributes directly to the medical imaging domain by producing disease-predictive embeddings under genetic supervision.

Our findings reinforce a central goal of imaging genetics: identifying intermediate imaging biomarkers that mediate the relationship between genetic variation and disease (Elliott et al., 2018; Meyer et al., 2020). In this context, GEMCONT operationalizes this principle by coupling disease-associated variants with phenotypic representations, guiding the learned imaging features toward mechanistic axes of disease risk. This extends our earlier work on genetics-supervised biomarker discovery (Sens et al., 2024). While as expected only a modest subset of the disease-associated variants is mediated through the imaging modality under study, each recovered locus represents a testable hypothesis linking genetic variation to an interpretable phenotypic feature.

Empirically, GEMCONT was evaluated across datasets of increasing complexity—from a multimodal MNIST benchmark to spirometry-derived flow–volume curves for asthma and retinal fundus photographs for glaucoma—demonstrating robust predictive performance across modalities. In particular, the glaucoma experiment highlights the practical relevance of GEMCONT within a canonical medical-imaging setting and its ability to refine foundation-model embeddings through disease-specific, genetics-based fine-tuning. Methodologically, we confirmed two core design hypotheses: (i) a linear genetic encoder effectively captures additive genotype–phenotype relationships without measurable gains from additional non-linear modeling; and (ii) performance remains stable under a stricter genome-wide significance threshold, indicating robustness to variant inclusion criteria.

Despite these strengths, several limitations remain. First, although ancestry-related biases were mitigated by regressing out genetic principal components and restricting analyses to unrelated individuals of homogeneous ancestry, residual population or site effects may persist despite covariate adjustment, and disentangling these confounders remains an open challenge for multimodal contrastive frameworks. Second, while the comparison between GEMCONT and the PRS baseline used identical variant panels to isolate the effect of individual variants versus aggregate modeling, future work will extend benchmarking to broader variant sets and polygenic scores from genetically correlated traits. Third, uncertainty in GWAS summary statistics was not explicitly modeled, and incorporating uncertainty-weighted variant selection represents a principled avenue for future development. Finally, although the fixed embedding dimensionality of $D = 256$ yielded stable results across all experiments, a systematic exploration of the trade-off between latent dimensionality and model performance across modalities will be valuable for further optimization.

Looking ahead, integrating generative decoders with GEMCONT represents a promising direction to enhance interpretability and facilitate imaging biomarker discovery. Building on emerging frameworks (Chaudhary et al., 2026, 2025; Shilova et al., 2025), such extensions could enable direct visualization of variant-driven imaging changes by decoding along genetic directions in the latent space. Beyond interpretability, future work will extend GEMCONT to additional imaging modalities (e.g., brain MRI) and integrate it within genetic causal inference frameworks (Davey Smith and Hemani, 2014; Sens et al., 2024). Through these developments, GEMCONT will advance the broader goal of genetics-informed imaging by linking genetic variation to intermediate phenotypes that mediate disease risk, thereby supporting biomarker discovery and patient stratification.

Acknowledgments

This research has been conducted using the UK Biobank Resource (Application Number 87065). FPC and DS were funded by the Free State of Bavaria's Hightech Agenda through the Institute of AI for Health (AIH). FPC acknowledges support from the Chan Zuckerberg Initiative (CZI) through the AI Residency Program. DS and LS acknowledge the support of the research school Munich School for Data Science (MUDS). AVD acknowledges support from NIH grant R01EB033773. JAS acknowledges funding from the German Federal Ministry of Education and Research and the Bavarian State Ministry for Science and the Arts under the Munich Centre for Machine Learning (MCML), and from the German Academic Exchange Service (DAAD) under the Konrad Zuse School of Excellence for Reliable AI (RelAI).

References

Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. SCARF: Self-supervised contrastive learning using random feature corruption. arXiv [cs.LG], 29 June 2021. URL https://openreview.net/pdf?id=CuV_qYkmKb3.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK biobank resource with deep phenotyping and genomic data. Nature, 562(7726):203–209, October 2018. ISSN 0028-0836,1476-4687. doi: 10.1038/s41586-018-0579-z. URL https://www.nature.com/articles/s41586-018-0579-z.

Oriol Canela-Xandri, Konrad Rawlik, and Albert Tenesa. An atlas of genetic associations in UK biobank. Nature Genetics, 50(11):1593–1599, November 2018. ISSN 1061-4036,1546-1718. doi: 10.1038/s41588-018-0248-z. URL http://dx.doi.org/10.1038/s41588-018-0248-z.

Francesco Paolo Casale, Barbara Rakitsch, Christoph Lippert, and Oliver Stegle. Efficient set tests for the genetic analysis of correlated traits. Nature methods, 12(8):755–758, August 2015. ISSN 1548-7091,1548-7105. doi: 10.1038/nmeth.3439. URL http://dx.doi.org/10.1038/nmeth.3439.

Shubham Chaudhary, Almut Voigts, Sergey Vilov, Matthias Heinig, and Francesco Paolo Casale. AI-based histopathology phenotyping reveals germline loci shaping breast cancer morphology. pages 199–212, 31 December 2025. URL https://raw.githubusercontent.com/mlresearch/v311/main/assets/chaudhary25a/chaudhary25a.pdf.

Shubham Chaudhary, Almut Voigts, Michael Bereket, Matthew L Albert, Kristina Schwamborn, Eleftheria Zeggini, and Francesco Paolo Casale. HistoGWAS: an AI-enabled framework for automated genetic analysis of tissue phenotypes in histology cohorts. Genome Biology, 27(1), 31 March 2026. ISSN 1474-7596,1474-760X. doi: 10.1186/s13059-026-04031-z. URL http://dx.doi.org/10.1186/s13059-026-04031-z.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. arXiv [cs.LG], 13 February 2020. URL http://proceedings.mlr.press/v119/chen20j/chen20j.pdf.

Shing Wan Choi, Timothy Shin-Heng Mak, and Paul F O'Reilly. Tutorial: a guide to performing polygenic risk score analyses. Nature protocols, 15(9):2759–2772, 24 September 2020. ISSN 1754-2189,1750-2799. doi: 10.1038/s41596-020-0353-1. URL https://www.nature.com/articles/s41596-020-0353-1.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: an extension of MNIST to handwritten letters. arXiv [cs.CV], 17 February 2017. URL http://arxiv.org/abs/1702.05373.

George Davey Smith and Gibran Hemani. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Human molecular genetics, 23(R1):R89–98, 15 September 2014. ISSN 0964-6906,1460-2083. doi: 10.1093/hmg/ddu328. URL http://dx.doi.org/10.1093/hmg/ddu328.

Lloyd T Elliott, Kevin Sharp, Fidel Alfaro-Almagro, Sinan Shi, Karla L Miller, Gwenaëlle Douaud, Jonathan Marchini, and Stephen M Smith. Genome-wide association studies of brain imaging phenotypes in UK biobank. Nature, 562(7726):210–216, October 2018. ISSN 0028-0836,1476-4687. doi: 10.1038/s41586-018-0571-7. URL https://www.nature.com/articles/s41586-018-0571-7.

Paul J Foster, Ralf Buhrmann, Harry A Quigley, and Gordon J Johnson. The definition and classification of glaucoma in prevalence surveys. The British journal of ophthalmology, 86 (2):238–242, February 2002. ISSN 0007-1161,1468-2079. doi: 10.1136/bjo.86.2.238. URL https://bjo.bmj.com/content/86/2/238.short.

Huazhu Fu, Boyang Wang, Jianbing Shen, Shanshan Cui, Yanwu Xu, Jiang Liu, and Ling Shao. Evaluation of retinal image quality assessment networks in different color-spaces. In Lecture Notes in Computer Science, Lecture notes in computer science, pages 48–56. Springer International Publishing, Cham, 2019. ISBN 9783030322380,9783030322397. doi: 10.1007/978-3-030-32239-7_6. URL http://dx.doi.org/10.1007/978-3-030-32239-7_6.

Mae O Gordon, Julia A Beiser, James D Brandt, Dale K Heuer, Eve J Higginbotham, Chris A Johnson, John L Keltner, J Philip Miller, Richard K Parrish, 2nd, M Roy Wilson, and Michael A Kass. The ocular hypertension treatment study: baseline factors that predict the onset of primary open-angle glaucoma. Archives of ophthalmology, 120(6):714–20; discussion 829–30, June 2002. ISSN 0003-9950,1538-3601. doi: 10.1001/archopht.120.6.714. URL http://dx.doi.org/10.1001/archopht.120.6.714.

William (bill) W G Hill, Michael E Goddard, and Peter M Visscher. Data and theory point to mainly additive genetic variance for complex traits. PLoS genetics, preprint (2008):e8, 2005. ISSN 1553-7390,1553-7404. doi: 10.1371/journal.pgen.1000008.eor. URL http://dx.doi.org/10.1371/journal.pgen.1000008.eor.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In Jennifer Dy and Andreas Krause, editors, Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 2127–2136. PMLR, 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

S Jayasooriya, M Stolbrink, E M Khoo, I T Sunte, J I Awuru, M Cohen, D C Lam, A Spanevello, D Visca, R Centis, G B Migliori, A C Ayuk, J A Buendia, B I Awokola, B E Del-Rio-Navarro, S Muteti-Fana, M Lao-Araya, P Chiarella, H Badellino, S W Somwe, M P Anand, J R Garcí-Corzo, A Bekele, M E Soto-Martinez, B H M Ngahane, M Florin, K Voyi, K Tabbah, B Bakki, A Alexander, B L Garba, E M Salvador, G B Fischer, A G Falade, Zorica ŽivkoviĆ, S J Romero-Tapia, G E Erhabor, H Zar, B Gemicioglu, H V Brandão, X Kurhasani, N El-Sharif, V Singh, J C Ranasinghe, S T Kudagammana, M R Masjedi, J N Velásquez, A Jain, I Cherrez-Ojeda, L F M Valdeavellano, R M Gómez, E Mesonjesi, B M Morfin-Maciel, A E Ndikum, G B Mukiibi, B K Reddy, O Yusuf, S Tarright-Mahi, J V Mérida-Palacio, S K Kabra, E Nkhama, N R Filho, V B Zhjegi, K Mortimer, S Rylance, and R R Masekela. Clinical standards for the diagnosis and management of asthma in low- and middle-income countries. The international journal of tuberculosis and lung disease: the official journal of the International Union against Tuberculosis and Lung Disease, 27 (9):658–667, 1 September 2023. ISSN 1027-3719,1815-7920. doi: 10.5588/ijtld.23.0203. URL http://dx.doi.org/10.5588/ijtld.23.0203.

Yongxiu Jin, Chenxi Han, Dongliang Yang, and Shanlin Gao. Association between gut microbiota and diabetic nephropathy: a mendelian randomization study. Frontiers in microbiology, 15:1309871, 27 March 2024. ISSN 1664-302X. doi: 10.3389/fmicb.2024.1309871. URL http://dx.doi.org/10.3389/fmicb.2024.1309871.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014. URL http://arxiv.org/abs/1312.6114.

Matthias Kirchler, Stefan Konigorski, Matthias Norden, Christian Meltendorf, Marius Kloft, Claudia Schurmann, and Christoph Lippert. transferGWAS: GWAS of images using deep transfer learning. Bioinformatics (Oxford, England), 38(14):3621–3628, 11 July 2022. ISSN 1367-4803,1367-4811. doi: 10.1093/bioinformatics/btac369. URL http://dx.doi.org/10.1093/bioinformatics/btac369.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (BiT): General visual representation learning. In Computer Vision – ECCV 2020, Lecture notes in computer science, pages 491–507. Springer International Publishing, Cham, 2020. ISBN 9783030585570,9783030585587. doi: 10.1007/978-3-030-58558-7_29. URL http://dx.doi.org/10.1007/978-3-030-58558-7_29.

Arthur Korte, Bjarni J Vilhjálmsson, Vincent Segura, Alexander Platt, Quan Long, and Magnus Nordborg. A mixed-model approach for genome-wide association studies of

correlated traits in structured populations. Nature genetics, 44(9):1066–1071, September 2012. ISSN 1061-4036,1546-1718. doi: 10.1038/ng.2376. URL http://dx.doi.org/10.1038/ng.2376.

Mitja I Kurki, Juha Karjalainen, Priit Palta, Timo P Sipilä, Kati Kristiansson, Kati M Donner, Mary P Reeve, Hannele Laivuori, Mervi Aavikko, Mari A Kaunisto, Anu Loukola, Elisa Lahtela, Hannele Mattsson, Päivi Laiho, Pietro Della Briotta Parolo, Arto A Lehisto, Masahiro Kanai, Nina Mars, Joel Rämö, Tuomo Kiiskinen, Henrike O Heyne, Kumar Veerapen, Sina Rüeger, Susanna Lemmelä, Wei Zhou, Sanni Ruotsalainen, Kalle Pärn, Tero Hiekkalinna, Sami Koskelainen, Teemu Paajanen, Vincent Llorens, Javier Gracia-Tabuenca, Harri Siirtola, Kadri Reis, Abdelrahman G Elnahas, Benjamin Sun, Christopher N Foley, Katriina Aalto-Setälä, Kaur Alasoo, Mikko Arvas, Kirsi Auro, Shameek Biswas, Argyro Bizaki-Vallaskangas, Olli Carpen, Chia-Yen Chen, Oluwaseun A Dada, Zhihao Ding, Margaret G Ehm, Kari Eklund, Martti Färkkilä, Hilary Finucane, Andrea Ganna, Awaisa Ghazal, Robert R Graham, Eric M Green, Antti Hakanen, Marco Hautalahti, Åsa K Hedman, Mikko Hiltunen, Reetta Hinttala, Iiris Hovatta, Xinli Hu, Adriana Huertas-Vazquez, Laura Huilaja, Julie Hunkapiller, Howard Jacob, Jan-Nygaard Jensen, Heikki Joensuu, Sally John, Valtteri Julkunen, Marc Jung, Juhani Junttila, Kai Kaarniranta, Mika Kähönen, Risto Kajanne, Lila Kallio, Reetta Kälviäinen, Jaakko Kaprio, FinnGen, Nurlan Kerimov, Johannes Kettunen, Elina Kilpeläinen, Terhi Kilpi, Katherine Klinger, Veli-Matti Kosma, Teijo Kuopio, Venla Kurra, Triin Laisk, Jari Laukkanen, Nathan Lawless, Aoxing Liu, Simonne Longerich, Reedik Mägi, Johanna Mäkelä, Antti Mäkitie, Anders Malarstig, Arto Mannermaa, Joseph Maranville, Athena Matakidou, Tuomo Meretoja, Sahar V Mozaffari, Mari E K Niemi, Marianna Niemi, Teemu Niiranen, Christopher J O Donnell, Ma En Obeidat, George Okafo, Hanna M Ollila, Antti Palomäki, Tuula Palotie, Jukka Partanen, Dirk S Paul, Margit Pelkonen, Rion K Pendergrass, Slavé Petrovski, Anne Pitkäranta, Adam Platt, David Pulford, Eero Punkka, Pirkko Pussinen, Neha Raghavan, Fedik Rahimov, Deepak Rajpal, Nicole A Renaud, Bridget Riley-Gillis, Rodosthenis Rodosthenous, Elmo Saarentaus, Aino Salminen, Eveliina Salminen, Veikko Salomaa, Johanna Schleutker, Raisa Serpi, Huei-Yi Shen, Richard Siegel, Kaisa Silander, Sanna Siltanen, Sirpa Soini, Hilkka Soininen, Jae Hoon Sul, Ioanna Tachmazidou, Kaisa Tasanen, Pentti Tienari, Sanna Toppila-Salmi, Taru Tukiainen, Tiinamaija Tuomi, Joni A Turunen, Jacob C Ulirsch, Felix Vaura, Petri Virolainen, Jeffrey Waring, Dawn Waterworth, Robert Yang, Mari Nelis, Anu Reigo, Andres Metspalu, Lili Milani, Tõnu Esko, Caroline Fox, Aki S Havulinna, Markus Perola, Samuli Ripatti, Anu Jalanko, Tarja Laitinen, Tomi P Mäkelä, Robert Plenge, Mark McCarthy, Heiko Runz, Mark J Daly, and Aarno Palotie. FinnGen provides genetic insights from a well-phenotyped isolated population. Nature, 613(7944): 508–518, January 2023. ISSN 0028-0836,1476-4687. doi: 10.1038/s41586-022-05473-8. URL http://dx.doi.org/10.1038/s41586-022-05473-8.

Allison Lambert, M Bradley Drummond, Christine Wei, Charles Irvin, David Kaminsky, Meredith McCormack, and Robert Wise. Diagnostic accuracy of FEV1/forced vital capacity ratio z scores in asthmatic patients. The journal of allergy and clinical immunology, 136 (3):649–653.e4, September 2015. ISSN 0091-6749,1097-6825. doi: 10.1016/j.jaci.2015.02.027. URL http://dx.doi.org/10.1016/j.jaci.2015.02.027.

Mark R Lesk, George L Spaeth, Augusto Azuara-Blanco, Silvana V Araujo, L Jay Katz, Annette K Terebuli, Richard P Wilson, Marlene R Moster, and Courtland M Schmidt. Reversal of optic disc cupping after glaucoma surgery, analysed with a scanning laser tomograph. Journal of glaucoma, 8(Supplement 1):S11, February 1999. ISSN 1057-0829,1536-481X. doi: 10.1097/00061198-199902001-00022. URL http://dx.doi.org/10.1097/00061198-199902001-00022.

Christoph Lippert, Franceso Paolo Casale, Barbara Rakitsch, and Oliver Stegle. LIMIX: genetic analysis of multiple traits. bioRxiv, 21 May 2014. doi: 10.1101/003905. URL http://dx.doi.org/10.1101/003905.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv [cs.LG], 14 November 2017. URL http://arxiv.org/abs/1711.05101.

Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, 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, 8 October 2009. ISSN 0028-0836,1476-4687. doi: 10.1038/nature08494. URL http://dx.doi.org/10.1038/nature08494.

Carla Martins, Henrique Barros, and André Moreira. Transfer learning in spirometry: CNN models for automated flow-volume curve quality control in paediatric populations. Computers in biology and medicine, 184(109341):109341, January 2025. ISSN 0010-4825,1879-0534. doi: 10.1016/j.compbiomed.2024.109341. URL http://dx.doi.org/10.1016/j.compbiomed.2024.109341.

Hannah V Meyer, Timothy J W Dawes, Marta Serrani, Wenjia Bai, Paweł Tokarczuk, Jiashen Cai, Antonio de Marvao, Albert Henry, R Thomas Lumbers, Jakob Gierten, Thomas Thumberger, Joachim Wittbrodt, James S Ware, Daniel Rueckert, Paul M Matthews, Sanjay K Prasad, Maria L Costantino, Stuart A Cook, Ewan Birney, and Declan P O'Regan. Genetic and functional insights into the fractal structure of the heart. Nature, 584(7822): 589–594, August 2020. ISSN 0028-0836,1476-4687. doi: 10.1038/s41586-020-2635-8. URL http://dx.doi.org/10.1038/s41586-020-2635-8.

Miranda A Paraskeva, Brigitte M Borg, and Matthew T Naughton. Spirometry. Australian family physician, 40(4):216, 2011. ISSN 0300-8495. URL https://www.racgp.org.au/getattachment/b2aef6c3-a6fb-46bf-9acf-f0215484f04c/Spirometry.aspx.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 3 December 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Karin R Pillunat, Florian T A Kretz, Stefan Koinzer, Christoph Ehlken, Lutz E Pillunat, and Karsten Klabe. Effectiveness and safety of VISULAS® green selective laser trabeculoplasty: a prospective, interventional multicenter clinical investigation. International ophthalmology, 43(7):2215–2224, July 2023. ISSN 0165-5701,1573-2630. doi: 10.1007/s10792-022-02617-7. URL http://dx.doi.org/10.1007/s10792-022-02617-7.

Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, Mark J Daly, and Pak C Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics, 81(3):559–575, September 2007. ISSN 0002-9297. doi: 10.1086/519795. URL http://dx.doi.org/10.1086/519795.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. arXiv [cs.CV], 26 February 2021. URL https://proceedings.mlr.press/v139/radford21a/radford21a.pdf.

N Raghu, S S Pandav, S Kaushik, P Ichhpujani, and A Gupta. Effect of trabeculectomy on RNFL thickness and optic disc parameters using optical coherence tomography. Eye, 26 (8):1131–1137, August 2012. ISSN 0950-222X,1476-5454. doi: 10.1038/eye.2012.115. URL http://dx.doi.org/10.1038/eye.2012.115.

Alexander Rakowski, Remo Monti, and Christoph Lippert. TransferGWAS of T1-weighted brain MRI data from UK biobank. PLoS genetics, 20(12):e1011332, December 2024. ISSN 1553-7390,1553-7404. doi: 10.1371/journal.pgen.1011332. URL http://dx.doi.org/10.1371/journal.pgen.1011332.

Peter N Robinson. Deep phenotyping for precision medicine. Human mutation, 33(5): 777–780, May 2012. ISSN 1059-7794,1098-1004. doi: 10.1002/humu.22080. URL http://dx.doi.org/10.1002/humu.22080.

Sajib Saha, Janardhan Vignarajan, and Shaun Frost. A fast and fully automated system for glaucoma detection using color fundus photographs. Scientific reports, 13(1):18408, 27 October 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-44473-0. URL http://dx.doi.org/10.1038/s41598-023-44473-0.

Daniel Sens, Liubov Shilova, Ludwig Gräf, Maria Grebenshchikova, Bjoern M Eskofier, and Francesco Paolo Casale. Genetics-driven risk predictions leveraging the mendelian randomization framework. Genome Research, 34(9):1276–1285, 11 October 2024. ISSN 1088-9051,1549-5469. doi: 10.1101/gr.279252.124. URL http://dx.doi.org/10.1101/gr.279252.124.

Liubov Shilova, Daniel Sens, Ayshan Aliyeva, Shubham Chaudhary, Qiaohan Xu, Emmanuelle Salin, Johannes Schiefelbein, Ben Asani, Oana Veronica Amarie, Elida Schneltzer, Ayellet V Segrè, Julia A Schnabel, Na Cai, Bjoern M Eskofier, and Francesco Paolo Casale. REECAP: Contrastive learning of retinal aging reveals genetic loci linking morphology to eye disease. medRxiv: the preprint server for health sciences, 27 November 2025. doi: 10.1101/2025.11.19.25340555. URL http://dx.doi.org/10.1101/2025.11.19.25340555.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine, 12(3):e1001779, March 2015. ISSN 1549-1277,1549-1676. doi: 10.1371/journal.pmed.1001779. URL http://dx.doi.org/10.1371/journal.pmed.1001779.

Aiham Taleb, Matthias Kirchler, Remo Monti, and Christoph Lippert. ContIG: Self-supervised multimodal contrastive learning for medical imaging with genetics. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, June 2022. doi: 10.1109/cvpr52688.2022.02024. URL https://openaccess.thecvf.com/content/CVPR2022/papers/Taleb_ContIG_Self-Supervised_Multimodal_Contrastive_Learning_for_Medical_Imaging_With_Genetics_CVPR_2022_paper.pdf.

Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. Benefits and limitations of genome-wide association studies. Nature reviews. Genetics, 20 (8):467–484, August 2019. ISSN 1471-0056,1471-0064. doi: 10.1038/s41576-019-0127-1. URL http://dx.doi.org/10.1038/s41576-019-0127-1.

Russell P Tracy. 'deep phenotyping': characterizing populations in the era of genomics and systems biology. Current opinion in lipidology, 19(2):151–157, April 2008. ISSN 0957-9672,1473-6535. doi: 10.1097/MOL.0b013e3282f73893. URL https://journals.lww.com/co-lipidology/fulltext/2008/04000/_deep_phenotyping___characterizing_populations_in.9.aspx.

Anurag Verma, Jennifer E Huffman, Alex Rodriguez, Mitchell Conery, Molei Liu, Yuk-Lam Ho, Youngdae Kim, David A Heise, Lindsay Guare, Vidul Ayakulangara Panickan, Helene Garcon, Franciel Linares, Lauren Costa, Ian Goethert, Ryan Tipton, Jacqueline Honerlaw, Laura Davies, Stacey Whitbourne, Jeremy Cohen, Daniel C Posner, Rahul Sangar, Michael Murray, Xuan Wang, Daniel R Dochtermann, Poornima Devineni, Yunling Shi, Tarak Nath Nandi, Themistocles L Assimes, Charles A Brunette, Robert J Carroll, Royce Clifford, Scott Duvall, Joel Gelernter, Adriana Hung, Sudha K Iyengar, Jacob Joseph, Rachel Kember, Henry Kranzler, Daniel Levey, Shiuh-Wen Luoh, Victoria C Merritt, Cassie Overstreet, Joseph D Deak, Struan F A Grant, Renato Polimanti, Panos Roussos, Yan V Sun, Sanan Venkatesh, Georgios Voloudakis, Amy Justice, Edmon Begoli, Rachel Ramoni, Georgia Tourassi, Saiju Pyarajan, Philip S Tsao, Christopher J O'Donnell, Sumitra Muralidhar, Jennifer Moser, Juan P Casas, Alexander G Bick, Wei Zhou, Tianxi Cai, Benjamin F Voight, Kelly Cho, Michael J Gaziano, Ravi K Madduri, Scott M Damrauer, and Katherine P Liao. Diversity and scale: Genetic architecture of 2,068 traits in the VA million veteran program. medRxiv: the preprint server for health sciences, 29 June 2023. doi: 10.1101/2023.06.28.23291975. URL http://dx.doi.org/10.1101/2023.06.28.23291975.

Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of GWAS discovery: Biology, function, and translation.

The American Journal of Human Genetics, 101(1):5–22, July 2017. ISSN 0002-9297,1537-6605. doi: 10.1016/j.ajhg.2017.06.005. URL http://dx.doi.org/10.1016/j.ajhg.2017.06.005.

Hung Q Vo, Lin Wang, Kelvin K Wong, Chika F Ezeana, Xiaohui Yu, Wei Yang, Jenny Chang, Hien V Nguyen, and Stephen T C Wong. Frozen large-scale pretrained vision-language models are the effective foundational backbone for multimodal breast cancer prediction. IEEE journal of biomedical and health informatics, 29(5):3234–3246, May 2025. ISSN 2168-2194,2168-2208. doi: 10.1109/JBHI.2024.3507638. URL http://dx.doi.org/10.1109/JBHI.2024.3507638.

Yimin Wang, Yicong Li, Wenya Chen, Changzheng Zhang, Lijuan Liang, Ruibo Huang, Jianling Liang, Dandan Tu, Yi Gao, Jinping Zheng, and Nanshan Zhong. Deep learning for spirometry quality assurance with spirometric indices and curves. Respiratory research, 23(1):98, 21 April 2022. ISSN 1465-9921,1465-993X. doi: 10.1186/s12931-022-02014-9. URL http://dx.doi.org/10.1186/s12931-022-02014-9.

Cristen J Willer, Yun Li, and Gonçalo R Abecasis. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics (Oxford, England), 26(17):2190–2191, 1 September 2010. ISSN 1367-4803,1367-4811. doi: 10.1093/bioinformatics/btq340. URL https://dx.doi.org/10.1093/bioinformatics/btq340.

J T Wright and M C Herzberg. Science for the next century: Deep phenotyping. Journal of dental research, 100(8):785–789, July 2021. ISSN 0022-0345,1544-0591. doi: 10.1177/00220345211001850. URL http://dx.doi.org/10.1177/00220345211001850.

Ziqian Xie, Tao Zhang, Sangbae Kim, Jiaxiong Lu, Wanheng Zhang, Cheng-Hui Lin, Man-Ru Wu, Alexander Davis, Roomasa Channa, Luca Giancardo, Han Chen, Sui Wang, Rui Chen, and Degui Zhi. iGWAS: Image-based genome-wide association of self-supervised deep phenotyping of retina fundus images. PLoS genetics, 20(5):e1011273, May 2024. ISSN 1553-7390,1553-7404. doi: 10.1371/journal.pgen.1011273. URL http://dx.doi.org/10.1371/journal.pgen.1011273.

Qiushi Yang, Wuyang Li, Baopu Li, and Yixuan Yuan. MRM: Masked relation modeling for medical image pre-training with genetics. In 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pages 21395–21405. IEEE, 1 October 2023. doi: 10.1109/iccv51070.2023.01961. URL http://dx.doi.org/10.1109/iccv51070.2023.01961.

Taedong Yun, Justin Cosentino, Babak Behsaz, Zachary R McCaw, Davin Hill, Robert Luben, Dongbing Lai, John Bates, Howard Yang, Tae-Hwi Schwantes-An, Yuchen Zhou, Anthony P Khawaja, Andrew Carroll, Brian D Hobbs, Michael H Cho, Cory Y McLean, and Farhad Hormozdiari. Unsupervised representation learning on high-dimensional clinical data improves genomic discovery and prediction. Nature genetics, 56(8):1604–1613, 8 August 2024. ISSN 1061-4036,1546-1718. doi: 10.1038/s41588-024-01831-6. URL https://www.nature.com/articles/s41588-024-01831-6.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric K Oermann. Confounding variables can degrade generalization performance of

radiological deep learning models. arXiv [cs.CV], 1 July 2018. URL http://arxiv.org/abs/1807.00431.

Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, Yuka Kihara, UK Biobank Eye & Vision Consortium, Andre Altmann, Aaron Y Lee, Eric J Topol, Alastair K Denniston, Daniel C Alexander, and Pearse A Keane. A foundation model for generalizable disease detection from retinal images. Nature, 622(7981):156–163, October 2023. ISSN 0028-0836,1476-4687. doi: 10.1038/s41586-023-06555-x. URL http://dx.doi.org/10.1038/s41586-023-06555-x.

Yukun Zhou, Zheyuan Wang, Yilan Wu, Ariel Yuhan Ong, Siegfried Wagner, Eden Ruffell, Mark Chia, Zhouyu Guan, Lie Ju, Justin Engelmann, David Merle, Tingyao Li, Jia Shu, Paul Nderitu, Ke Zou, Jocelyn Hui Lin Goh, Qingshan Hou, Xiaoxuan Liu, Yaxing Wang, Yih Chung Tham, Andre Altmann, Carol Cheung, Daniel Alexander, Eric Topol, Alastair Denniston, Tien Yin Wong, Bin Sheng, and Pearse A Keane. Revealing the impact of pre-training data on medical foundation models. Research Square, 3 April 2025. doi: 10.21203/rs.3.rs-6080254/v1. URL http://dx.doi.org/10.21203/rs.3.rs-6080254/v1.