EmbedGEM: A framework to evaluate the utility of embeddings for genetic discovery

Sumit Mukherjee,[1,*] Zachary R McCaw,[1], Jingwen Pei,[1] Anna Merkoulovitch,[1] Raghav Tandon,[2]

Tom Soare,[1] David Amar,[1] Hari Somineni,[1] Christoph Klein,[1] Santhosh Satapati,[1] David Lloyd,[1]

Christopher Probert,[1] insitro Research Team,[1] Daphne Koller,[1] Colm O'Dushlaine,[1] Theofanis Karaletsos[1]

Abstract

Machine learning derived embeddings are a compressed representation of high content data modalities obtained through deep learning models[1]. Embeddings have been hypothesized to capture detailed information about disease states and have been qualitatively shown to be useful in genetic discovery. Despite their promise, embeddings have some drawbacks: i) they are often confounded by covariates, and ii) their disease relevance is hard to ascertain. In this work we describe a framework to systematically evaluate the utility of embeddings in genetic discovery called EmbedGEM (**Embed**ding **G**enetic **E**valuation **M**ethods). Although, motivated by applications to embeddings, EmbedGEM is equally applicable for other multivariate traits as well.

EmbedGEM focuses on comparing embeddings along two axes: i) heritability of the embeddings, and ii) ability to identify 'disease relevant' variants. We use the number of genome-wide significant signals and mean/median chi-square statistic as a proxy for the heritability of multivariate traits. To evaluate disease relevance, we compute polygenic risk scores for each orthogonalized component of the embedding (or multivariate comparators) and evaluate their association with a held-out set of patients with high-confidence disease traits. While we introduce some relatively straightforward ways to evaluate heritability and disease relevance, we foresee that our framework can be easily extended by adding more metrics.

We demonstrate the utility of EmbedGEM by using it to evaluate embedding and non-embedding traits in two separate datasets: i) a synthetic dataset simulated to demonstrate the ability of the framework to correctly rank traits based on their heritability and disease relevance, ii) data from the UK Biobank focused on NAFLD relevant traits. EmbedGEM is implemented in the form of an easy to use Python-based workflow (https://github.com/insitro/EmbedGEM) .

Keywords: Representation learning, Genome-Wide Association Study, Polygenic Risk Scores.

[1] Insitro, South San Francisco, California, USA.

[2] Center for Machine Learning, Georgia Institute of Technology, Georgia, USA.

1 Introduction

Representation learning is a crucial aspect of modern day machine learning, aiming to discover meaningful and informative representations of data. Machine learned representations are often simply called *embeddings*. Deep neural networks have emerged as a powerful tool for representation learning, demonstrating remarkable success across various domains. Representation learning was first introduced in the form of unsupervised pre-training [2], where a deep neural network was trained on unlabeled data to initialize subsequent supervised learning tasks. Autoencoders were the next generation of representation learning algorithms, which are neural networks trained to reconstruct input, enabling the learning of compact and informative representations [3]. More recently, self-supervised learning has gained attention as a promising approach for representation learning, where the model learns representations by maximizing agreement between differently augmented views of the same data [4]. These advancements have significantly improved tasks like image classification, natural language processing, and recommendation systems [5]. Representation learning has also been successfully used in several healthcare applications. For instance, it has been employed in analysis of electronic health records in [6] to enable personalized disease progression prediction. More recently, self-supervised learning of electronic health records was employed to facilitate downstream tasks such as patient similarity and disease prediction [7].

In recent years, embeddings have become increasingly used in genetic discovery [8, 9, 10, 11, 12] helping to link new seemingly relevant variants to disease indications of interest. While some studies have evaluated the utility of using embeddings for genetic discovery heuristically, there is currently little systematic work evaluating the value of embeddings in genetic discovery. In [13], the authors generate Polygenic Risc Scores (PRSs) from their embeddings in addition to performing univariate Genome-wide Association Studies (GWAS) for each component. While this was not developed as an explicit way to evaluate disease relevance of embeddings, it served as a motivation for our approach to evaluate disease relevance. In [9] the authors conducted univariate GWAS for each PC of embeddings and analyzed extreme samples to visually interpret the PCs and assessed clinical relevance using PheWAS [14]. Similarly, [12, 11] performed GWAS on each dimension of the embeddings and conducted genetic correlation analysis with other relevant traits to evaluate the relevance of the different embedding dimensions.

Several other studies have utilized more indirect approaches to evaluate the relevance of embeddings. Dadousis and colleagues [10] conducted a GWAS using latent variables related to milk protein composition and cheesemaking traits in dairy cattle. They then tested nominally significant genes for gene set enrichment to qualitatively evaluate the results derived from embeddings. In another study [15, 16], the authors present an approach to learn a low dimensional representation from paired GTEx data. They validated the methodology by assessing tissue-specific variation in latent factors and investigating associations

between factors and visual features through imaging QTL. Mukherjee and colleagues generated unsupervised 2D representations learned from bulk RNA-Seq data and used a metric based on patient distance from the prototypical baseline as a quantitative phenotype in GWAS [8]. The authors validate the metric against other known clinical metrics to quantify disease severity.

Unlike prior works which were focused on simply performing genetic discovery on embeddings, here we propose a coherent formal framework to evaluate the utility of embeddings in genetic discovery, EmbedGEM (Figure 1). EmbedGEM evaluates embeddings along two common dimensions: i) heritability, ii) disease relevance. We demonstrate the utility of EmbedGEM in comparing the utility of embeddings as well as other relevant traits for both a simulated and real dataset. Finally, we share a software implementation of EmbedGEM along with primer on how to use it.

2 Background

2.1 Mathematical formulation of representation learning

Let X denote the input data space and Y denote the learned representation space. Given a dataset of the form $\mathbb{D} = (x_1, l_1), (x_2, l_2), ..., (x_n, l_n)$, where $x_i \in X$ is an input sample and l_i is its corresponding optional label, the goal of representation learning is to find a mapping function $f : X \to Y$ that captures the underlying structure and meaningful features of the data. This is often formulated as a supervised learning problem, where the objective is to minimize a loss function $L(g(f(x_i)), l_i)$ that measures the discrepancy between the predicted labels using the learned representations $g(f(x_i))$ and the true label l_i. In the absence of meaningful labels, representation learning can also be formulated as an unsupervised or self-supervised learning problem where the loss function is often of the form $L(g(f(x_i)), x_i)$. Here, g represents some function that maps $f(x)$ to the domain of l or x. In both cases, the primary output of the representation learning algorithm for a given data sample x_i is $y_i = f(x_i)$, which is often a lower dimension representation of the input (commonly referred to as embeddings).

2.2 Genome-wide association study (GWAS)

GWAS involves analyzing a large number of single nucleotide polymorphisms (SNPs) across the genome to identify associations between specific genetic variants and phenotypic traits. The mathematical formulation of GWAS can be represented as a statistical model, where the phenotype is regressed on the genotype (encoded by G), which represents the number of risk alleles an individual carries at the genetic variant. This

can be expressed at a high level as:

$$Y = \beta_0 + \beta G + \sum_{j}^{m} \gamma_j X_j + \epsilon$$

where Y is the phenotype, $(X_1, X_2, ..., X_m)$ are the covariates, β is the regression coefficient for the variant encoded by G, $(\gamma_1, \gamma_2, ..., \gamma_m)$ are the regression coefficients for the covariates and ϵ is the error term. This formulation allows for the identification of variants that are significantly associated with the phenotype of interest. The associations are run separately for each variant because of the high dimensionality of the variant set. A detailed review on GWAS methodology can be found in [17, 18].

2.3 LD-based clumping Clumping

Clumping based on Linkage Disequilibrium (LD) is a process used in genetic studies to identify and retain only the most significantly associated variants within a region of linkage disequilibrium (LD) [19, 20]. This is done to reduce redundancy due to the correlation between variants in close proximity to each other. The mathematical formulation of clumping can be represented as a selection process, where for each LD region, only the variant with the lowest p-value is retained. This can be expressed at a high level as:

$$S = \{s_i : p(s_i) = \min_{s_j \in LD(i)} p(s_j)\}$$

where S is the set of selected variants, s_i is a variant, $p(s_i)$ is the p-value of variant s_i, and $LD(i)$ is the set of variants in the same LD region as s_i. An LD region is defined as a genomic region where the variants exhibit high pairwise correlation, indicating that they are inherited together more frequently than expected by chance. LD regions are typically determined using measures such as the squared correlation coefficient (r^2) or the D' statistic. This formulation allows for the reduction of redundancy in the variant set, while retaining the most significantly associated variants within each LD region. It is often the case that clumping is performed after a GWAS to refine the set of associated variants.

2.4 Polygenic Risk Scores (PRS)

Polygenic risk scores (PRS) have emerged as a powerful tool in genetic epidemiology for predicting an individual's risk of developing complex traits and diseases. A PRS is calculated by summing the weighted contributions of multiple genetic variants across the genome. The mathematical formulation of PRS can be

expressed as:

$$\text{PRS}_i = \sum_{j=1}^{J} \beta_i G_{ij}$$

where PRS_i is the polygenic risk score for subject i, G_{ij} represents the number of risk alleles subject i carries at genetic variant j, and β_j represents the effect size of each variant estimated from large-scale GWAS. By aggregating the effects of multiple genetic variants, PRS provides a quantitative measure of genetic predisposition to a particular trait or disease. A comprehensive tutorial on PRS can be found in [21, 22, 23].

3 Methods and Datasets

In this section we detail the EmbedGEM workflow, explaining the two aspects of the workflow: i) evaluating heritability of embeddings, ii) assessing the *disease relevance* of the associations. As seen in Figure 1, the heritability evaluation comprises of deriving multivariate GWAS summary statistics from univariate GWAS summary statistics of orthogonalized components of the embedding (section 3.1.1). This is then used to compute metrics that quantify the heritability of the embeddings (section 3.1.2). For evaluating disease relevance, we assess the performance of polygenic risk scores from the orthogonalized dimensions of embeddings in predicting a disease trait of interest (section 3.1.3).

Aside from introducing the different components of the workflow, this section also introduces the datasets used to evaluate EmbedGEM. First we detail a simulated dataset that we use to demonstrate the ability of EmbedGEM to correctly order different embeddings (section 3.2). Then we introduce a real world example, along with methods used to process the data and generate embeddings (section 3.3).

3.1 Evaluation methods

3.1.1 GWAS methodology for multivariate traits

Below we describe the approaches to compare multivariate traits along the two aforementioned axes, we first describe a simple method to obtain multivariate GWAS summary statistics using univariate summary statistics from orthogonal traits.

Given K traits $(Y_1, \ldots, Y_K)$, we perform Principal Component Analysis (PCA) to obtain k orthogonal principal components (PCs), denoted as $\text{PC}_1, \ldots, \text{PC}_K$. Subsequently, we conduct single-trait GWAS for the first $m \leq K$ PCs (where m was either user-selected or determined through an automated procedure)

using the following model:

$$\mathrm{PC}_k = \beta_k G + \sum_{j=1}^{J} \gamma_{ij} X_j + e$$

Here pc_i is the trait PC, G is genotype at the variant of interest, and $(X_1, \ldots, X_J)$ are a set of covariates. The effect of genotype on the ith PC is captured by β_i. Because the principal components are orthogonal by construction, the per-component Wald statistics $Z_i = \hat{\beta}_i / \mathbb{SE}(\hat{\beta}_i)$ are independent and multivariate normal asymptotically. Under the null hypothesis of no association, the combined test statistic $T = \sum_{i=1}^{m} Z_i^2$ follows a central χ^2 distribution with m degrees of freedom. We utilize $T \sim \chi^2_m(0)$ to evaluate the hypothesis:

$$H_0 : G \text{ is not associated with any of the } m \text{ PCs}$$

$$H_a : G \text{ is associated with at least one of the } m \text{ PCs}$$

Note: PCA is just one way to generate orthogonal components of embeddings and was used to generate the results in this paper. However, the EmbedGEM workflow doesn't dependent on any particular way of orthogonalizing traits.

3.1.2 Evaluating heritability

The goal of evaluating heritability is to compare the different multivariate traits (e.g. different embeddings) in terms of their ability to manifest genetic associations. In our approach, which is commonly used in the field, we examine the following metrics calculated from the GWAS summary statistics of the multivariate trait of interest:

- Number of independent (clumped) genome-wide significant (GWS) variants (p-value $\leq 5 \times 10^{-8}$).
- Mean/Median χ^2 statistic of the independent (clumped) genome-wide significant hits for the multivariate trait.

The mean/median χ^2 statistic at independent genome-wide significant (GWS) variants is an assessment of signal strength. An approach that provides a higher mean/median χ^2 provides better power, potentially allowing for genetic discoveries with smaller sample sizes. The number GWS variants is a more qualitative proxy for the same. The primary drawback of these metrics is that we have no way of knowing which of the identified GWS variants are false positives. If a gold-standard set of variants (e.g. variants known to be associated with NAFLD) is available, the multivariate traits may also be compared with respect to the average χ^2 at these variants.

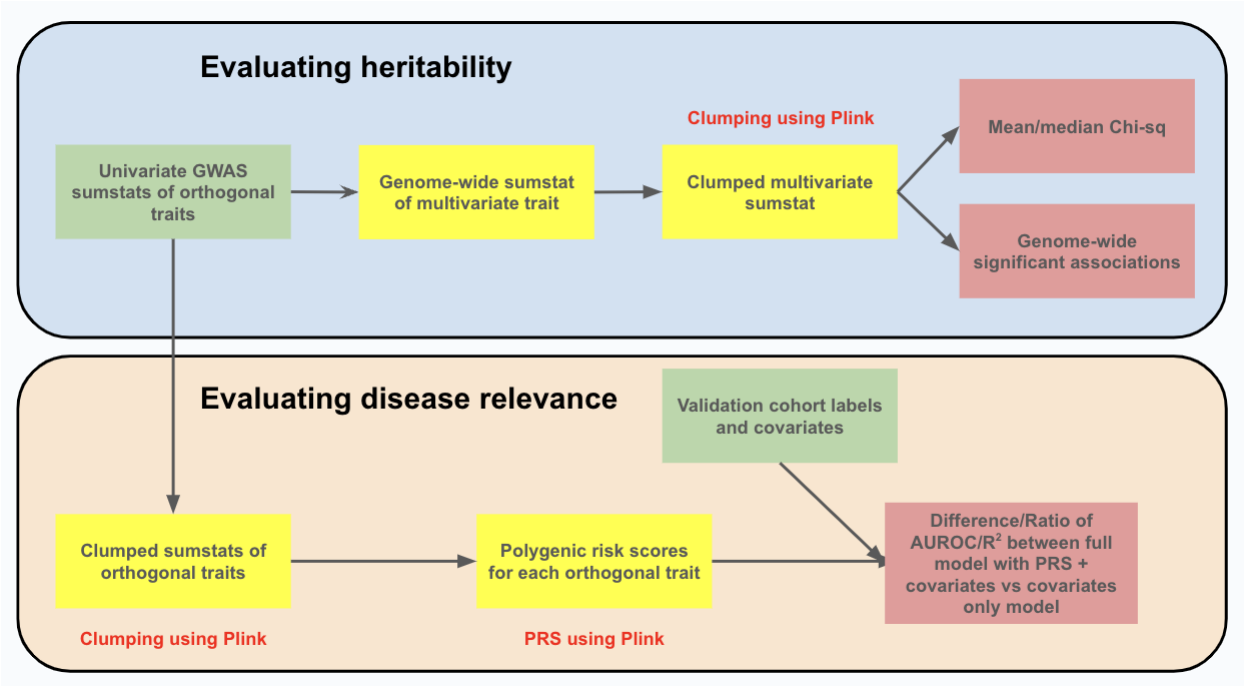

Figure 1: **Overview of the genetic validation workflow of EmbedGEM.** Green boxes indicate inputs, yellow boxes indicate intermediate outputs and red boxes indicate output metrics. Parts of the workflow that utilize `plink` commands are highlighted using red text.

3.1.3 Evaluating disease relevance

The purpose of the disease relevance evaluation is to compare embeddings with respect to their ability to identify disease-relevant variants. More specifically, we evaluate the disease relevance using the PRSs for all orthogonalized traits (e.g. embedding PCs) by comparing a full disease prediction model, which includes the PRSs, to a reduced model which excludes them. For the ith subject, let D_i denote the disease or evaluation trait, (PRS_{ik}) that subject's polygenic score with respect to the kth embedding, and (X_{ij}) a collection of covariates. We fit the following full and reduced generalized linear models (GLM) using the observed data:

$$\text{Full}: \quad g\{\mathbb{E}(D_i|\text{PRS}_{ik}, X_{ij})\} = \sum_{k=1}^{K} \beta_k \text{PRS}_{ik} + \sum_{j=1}^{J} \gamma_j X_{ij},$$

$$\text{Reduced}: \quad g\{\mathbb{E}(D_i|X_{ij})\} = \sum_{j=1}^{J} \gamma_j X_{ij}.$$

For a binary trait, logistic models are fit, while for a continuous trait, linear models are fit. From a fitted GLM, a prediction $\hat{D}_i$ of the disease trait is obtained, which can be compared with the observed D_i using a variety of metrics. For binary traits, we calculate the area under the receiver operating characteristic (AUROC) and the area under the precision-recall curve (AUPRC). For continuous traits, we calculate the square correlation and the mean absolute error (MAE). The full and reduced models are compared with

respect to a contrast Δ (e.g. difference or ratio) of the evaluation metrics. To assess whether addition of the PRSs significantly improves disease relevance, we estimate the distribution of Δ under the null by first permuting the PRSs, rendering them non-informative, then taking B bootstrap resamples of the data (with replacement). For each resample b, the full and reduced models are fit, and metric contrast Δ_b calculated. Letting Δ_{obs} denote the observed contrast (on the unpermuted data), the final p-value is:

$$p = \frac{1}{1+B}\left\{1 + \sum_{b=1}^{B} \mathbb{I}(|\Delta_b| \geq |\Delta_{\text{obs}}|)\right\}.$$

3.2 Simulation of toy dataset

To demonstrate the validity of the results generated by our workflow, we generate an illustrative toy example. Starting from `plink` genotypes for the EUR 1KG cohorts [24] ($n = 522$), LD pruning at $R^2 \leq 0.1$ was applied to obtain a set of variants in linkage equilibrium, then 100 variants were sampled at random from each chromosome (2,200 variants in total). Two orthogonal univariate traits, Y_1 and Y_2, were simulated to represent embeddings. Each Z_k was generated from an infinitesimal model:

$$\beta_k \sim N\left(0, \frac{h_k^2}{n_k} I\right), \qquad Y_k = G_k \beta_k + \epsilon_k,$$

where G_k is the set of standardized genetic variants affecting Y_k, β_k is the effect size vector, $\epsilon_k \sim N(0, 1-h_k^2)$ is a residual, h_k^2 is the heritability of Y_k, and n_k is the number of causal variants. Each simulated embedding depended on $n_k = 1000$ variants, with 500 variants affecting both embeddings and 500 variants unique to each embedding.

The evaluation trait Y was generated via a linear combination of the embedding traits and a covariate representing sex:

$$D = Y_1\gamma_1 + Y_2\gamma_2 + X\gamma_X + \varepsilon, \qquad \varepsilon \sim N(0, 1).$$

Three scenarios were considered:

i. **High heritability and high disease relevance**: Here the embeddings were simulated to have 99% heritability, and γ_1 and γ_2 where set to large values, relative to γ_X and ε, such that D was highly dependent on the embeddings.

ii. **High heritability but low disease relevance**: Here the embeddings remained unchanged but the evaluation trait was permuted such that variants associated with Y_1 and Y_2 were unassociated with D.

iii. **Low heritability and low disease relevance**: Here both the embeddings (Y_1, Y_2) and the evalua-

tion traits D were permuted such that the embeddings were not heritable and D did not depend on the embeddings.

3.3 Data processing of real-world dataset

3.3.1 Cohort description: Non-Alcoholic Fatty Liver Disease (NAFLD) in UK Biobank

Non-Alcoholic Fatty Liver Disease (NAFLD) is a prevalent and complex chronic liver condition, characterized by the accumulation of excess fat in the liver of individuals who consume little to no alcohol [25]. It is closely associated with metabolic syndrome and is considered as the hepatic manifestation of the syndrome. NAFLD encompasses a wide spectrum of liver damage, ranging from simple steatosis to non-alcoholic steatohepatitis (NASH), advanced fibrosis, and cirrhosis. The global prevalence of NAFLD is estimated to be approximately 25%, making it a significant public health concern [25].

The UK Biobank is a large-scale biomedical database and research resource, containing in-depth genetic and health information from half a million UK individuals aged 40-69 years [26, 27]. The resource includes data on a wide range of health-related outcomes, which are being further enriched via linkage to diverse medical, social, and environmental records. Neck-to-knee MRIs from UK Biobank have previously been used to extract various adiposity or organ traits [28, 29]. He we utilized the method outlined in [29] to impute adiposity traits for approximately thirty six thousand patients from their neck-to-knee MRIs. These deep imputation models were also our source for supervised embeddings.

We identified 203 fields that have a known or speculative association with the disease of interest (included in supplementary materials). The traits were grouped into: i) NMR metabolomics, ii) abdominal composition variables, iii) blood biochemistry markers. We also identified 48 covariates that might affect traits of interest independently from the disease processes of interest. The list of covariates were grouped into: i) body size, ii) addiction information, iii) medical status information, iv) medication usage information, v) baseline characteristics.

To curate a cohort for evaluating the disease relevance, we identified 1774 NAFLD cases and 2209 NAFLD controls using ICD-10 codes found in UK Biobank. These individuals were removed from the discovery cohort (individuals with neck-to-knee MRIs).

3.3.2 Processing of neck-to-knee MRI images

We use the processing pipeline outlined in [29] to process the neck-to-knee MRI images. Six MRI stations for each participant were first merged into a unified voxel grid using trilinear interpolation. This resulted in a single volume of $(224 \times 174 \times 370)$ voxels for each signal type (fat and water). Subsequently, these volumes are transformed into a two-dimensional representations by aggregating all values along two viewing axes. This produces a mean intensity projection for both the coronal and sagittal views, which were then tiled side by side. This procedure was separately executed for both the water and fat signals. The resultant images were individually normalized and downsampled to create two color channels of a single image with dimensions of $(256 \times 256 \times 2)$ pixels. A third channel was generated by taking the average of the first two channels.

3.3.3 Learning supervised embeddings

Supervised embeddings were learned using the process described in [29] for each adiposity trait separately (see list of traits in appendix) The modeling procedure involved replacing the last layer of an ImageNet pre-trained ResNet-50 [30] model with a linear layer (to obtain regression predictions). The model was then finetuned end-to-end using 80% of the labeled data for training and 20% of the data for testing. An adaptive learning rate schedule was used for the Adam optimizer, as stated in [29]. 2048 dimensional embeddings were extracted from the flattened output of the last convolutional layer.

4 Software implementation details

Our genetic validation pipeline is implemented by combining several commonly used `plink` commands through a *redun* [31] based workflow. Redun is a powerful workflow engine developed by *insitro*. It is designed to facilitate the execution of computational workflows, particularly in the field of bioinformatics. Redun provides a robust framework for defining, running, and managing workflows, ensuring reproducibility and traceability of results. In particular, redun enables us to seamlessly scale up from local compute to compute backends such as AWS Batch for larger datasets. As seen in Figure 1 and evident from previous sections, our workflow comprises of two main portions: i) evaluation of heritability, ii) evaluation of disease relevance. Our workflow is implemented to run end-to-end based using a single python script which can be executed from the command line as follows:

```
redun run generate_dataset.py main --config config.yaml
```

Here, `config.yaml` is the config file which contains links to the inputs needed for the workflow execution as well as parameters used by various sub-modules. An overview of the different config parameters can be found in the Supplement as well as in the Readme document of our github repository.

5 Experiment and results

In this section, we present the results of our experiments using EmbedGEM. We first demonstrate its utility in comparing three archetypal simulated scenarios. We then move on to real-world data, where we compare the embeddings with other multivariate traits in the context of NAFLD. Finally, we explore the utility of different types of embeddings, demonstrating that not all embeddings are equally useful in genetic discovery.

5.1 EmbedGEM orders archetypal toy scenarios correctly based on heritability and disease relevance

Traits	Disease relevance				Heritability	
	r^2 ratio	r^2 p-value	MAE ratio	MAE p-value	No. of hits	mean χ^2
High heritability, high disease relevance	35.547	0.227	0.863	0.001	6	16.572
Low heritability, low disease relevance	N/A	N/A	N/A	N/A	0	0
High heritability, low disease relevance	2.447	0.955	0.995	0.834	6	16.572

Table 1: **Comparison of archetypal toy scenarios.** The table shows the results of our EmbedGEM for three scenarios: high heritability and high disease relevance, low heritability and low disease relevance, and high heritability and low disease relevance. The framework correctly orders the three scenarios in terms of heritability and disease relevance. The high heritability traits have a high mean χ^2/number of genome-wide significant hits, while the low heritability phenotype has a mean χ^2 and number of genome-wide significant hits equal to 0. The disease relevance is also correctly identified, with only the high heritability phenotype with high disease relevance exhibiting a low MAE ratio and p-value.

As described in Section 3.2, we generated three archetypal scenarios to demonstrate the utility of EmbedGEM namely, i) high heritability and high disease relevance, ii) low heritability and low disease relevance, iii) high heritability and low disease relevance. In all scenarios, we treated the two simulated traits as one multivariate trait. In table 1 we can see that EmbedGEM correctly ranks the traits in terms of both heritability as well as disease relevance. For instance, we can see that both high heritability multivariate traits have high mean χ^2 and/or number of genome-wide significant hits, while the low heritability multivariate trait has Mean χ^2 and number of genome-wide equal to 0. We also see the disease relevance is also correctly identified with only the high heritability phenotype with high disease relevance exhibiting a low MAE ratio and p-value.

5.2 Comparing embeddings with other multivariate traits highlights the importance of evaluating disease relevance

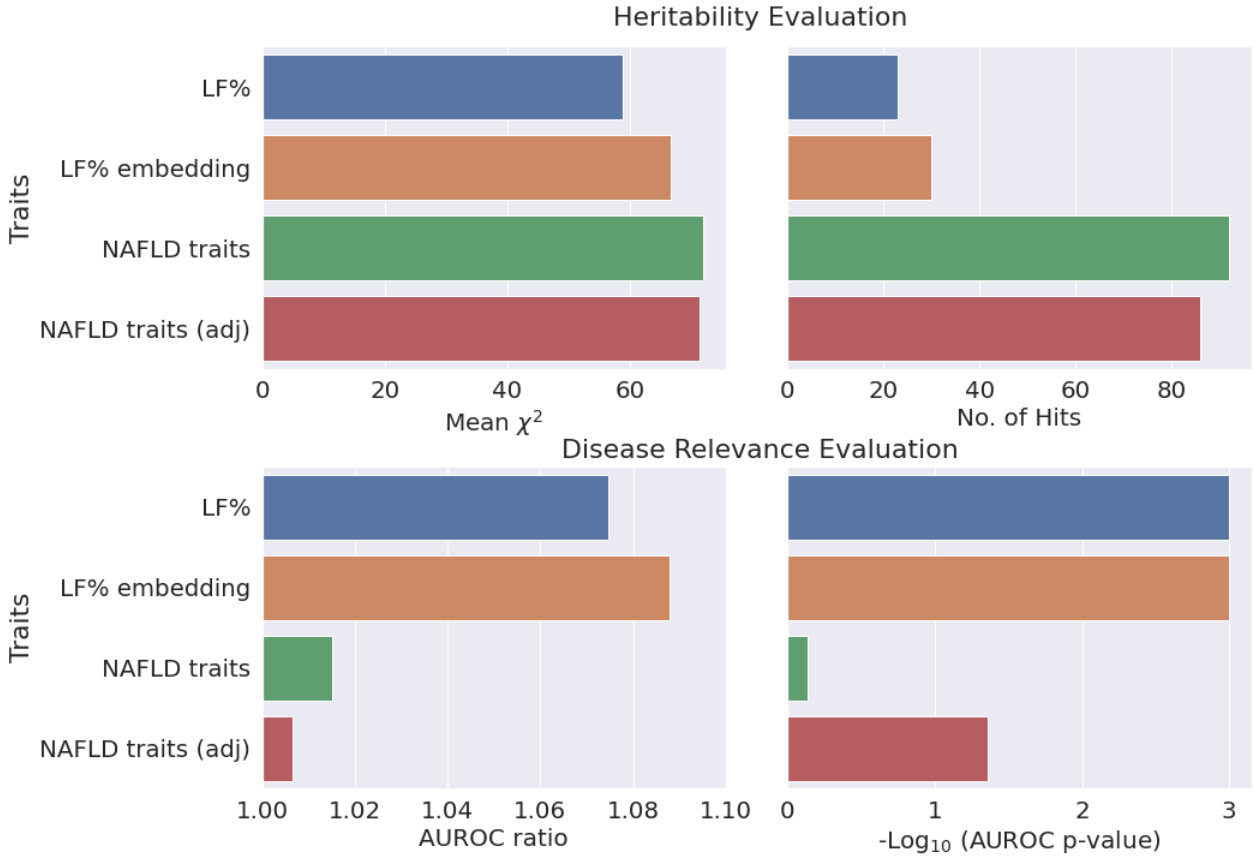

Figure 2: **Comparison of heritability and disease relevance between LF embeddings and other traits.** The figure illustrates the heritability and disease relevance of liver fat (LF) embeddings, LF percentage predictions, and 203 NAFLD relevant traits. LF embeddings, despite having lower heritability than the 203 traits, show higher disease relevance and heritability than the univariate LF predictions, suggesting that strong supervised embeddings might be more powerful than strong univariate biomarkers.

To evaluate the utility of embedding derived traits with respect to non-embedding derived univariate and multi-variate traits in a real human cohort, we selected the following comparators in the UK Biobank NAFLD cohort:

- Liver fat percentage predictions from a supervised machine learning model trained on neck-to-knee MRIs.
- Raw embeddings extracted from the supervised model trained to predict liver fat percentage from neck-to-knee MRIs.

- 203 NAFLD relevant traits with and without adjustment for covariates, treated as a multivariate trait.

Liver fat percentage is known to be a very strong biomarker for NAFLD and MRI imputed LF% has been used for genetic discovery for NAFLD previously[29], hence it serves as a strong univariate baseline. The 203 NAFLD relevant traits were selected as a non-embedding but multivariate baseline since they should both be heritable as well as somewhat disease relevant. To ensure a fair comparison between traits, we only included individuals who had no missing values for any of the traits, thereby ensuring that all GWAS analyses were conducted on the same sample size. For each multivariate trait, we took the first 5 PCs as the input to our workflow.

As seen in Figure 2, we can see that while LF% embeddings show lower mean χ^2 and fewer genome-wide significant hits than the 203 traits, the associations derived from the embeddings are much more disease relevant than them. Notably we see that LF% embedding has a higher disease relevance and higher heritability than the univariate LF predictions. This suggests that strong supervised embeddings have more statistical power than strong univariate biomarkers. Furthermore, it importantly highlights a flaw in simply using proxies of heritability to ascertain the utility of embeddings in genetic discovery [9, 8].

5.3 Not all embeddings are equally useful

Traits	Disease relevance				Heritability	
	AUC ratio	AUC p-value	AUPRC ratio	AUPRC p-value	No. of hits	mean χ^2
LF% embedding	1.088	0.001	1.162	0.001	30	66.615
ATFMVL embedding	1.017	0.51	1.046	0.29	20	33.556
ResNet embedding	N/A	N/A	N/A	N/A	7	32.440

Table 2: **Comparison of heritability and disease relevance between different embeddings.** The table illustrates the heritability and disease relevance of embeddings derived from different machine learning models. We show that depending on how the embeddings are trained, they may have dramatically different utility in genetic discovery. While, a LF% embedding has both high heritability and disease relevance, embeddings of a weaker disease proxy (ATFMVL) has much lower disease relevance and embeddings from ResNet pre-trained on ImageNet show no disease relevance at all.

While the previous sub-section demonstrated that compared to strong univariate biomarkers and non-embedding multivariate traits, embeddings may be both more heritable and more disease relevant, we next wanted to see if these benefits extend embeddings trained on less disease specific tasks as well. To study this, we compared LF% embeddings with embeddings derived from two other models: i) a supervised model

trained to predict anterior thigh fat-free muscle volume of the left side (ATFMVL) from whole-body MRIs, which has a low correlation with LF% ($r^2 = 0.21$), and ii) ImageNet pre-trained ResNet model [30].
As seen in Table 2, embeddings derived from LF% are much more heritable and more disease relevant than the two other embeddings. While both ATFMVL and Imagenet pre-trained ResNet embeddings show lower heritability and much lower disease relevance than LF% embeddings. The PCs of the Imagenet pre-trained ResNet embeddings don't produce any independent genome-wide signalsand hence fails to produce any disease relevance results. We speculate that this is because the representations learned by the Imagenet pre-trained ResNet model are not specific to human biology and probably capture unrelated features such as low level imaging features like shapes. While these are often quite useful in predictive tasks like types of animals, they may be neither heritable nor disease relevant.

6 Conclusion

Here we described the first framework to evaluate the utility of embeddings for genetic discovery. We designed EmbedGEM to be re-usable in any setting where multivariate traits need to be compared for their utility in disease relevant genetic discovery. The genetic validation pipeline is implemented using a combination of several commonly used plink commands through a redun-based workflow. The workflow comprises two main portions: evaluation of heritability and evaluation of disease relevance. To evaluate heritability, the sumstats of orthogonalized traits are combined to get multivariate GWAS sumstats, which are then clumped using plink. Metrics for evaluating heritability are then computed and reported. To evaluate disease relevance, the sumstats of the orthogonalized traits are clumped using plink, followed by generating a PRS for each trait. These PRSs are then used for association testing and generating metrics. The entire workflow is implemented to run end-to-end using a single python script, ensuring reproducibility and traceability of results.

We show that EmbedGEM successfully ranked the simulated traits in terms of both heritability and disease relevance. We then demonstrated the utility of EmbedGEM on real data from the UK Biobank. We showed that in a comparison between embedding derived traits with other multivariate traits, the embeddings showed lower heritability compared to the 203 non-embedding traits. However, the associations derived from the embeddings were found to be much more disease relevant. Notably, the LF% embedding had higher disease relevance and heritability than the univariate liver fat predictions, suggesting that strong supervised embeddings may have higher statistical power than even very strong univariate biomarkers (e.g. LF%, which is also used to diagnose the disease of interest NAFLD).

Furthermore, we used EmbedGEM to compare embeddings derived from different machine learning models. Our results clearly showed that not all embeddings are valuable for genetic discovery and the value of the embedding is tied to training process used to generate them. For instance, embeddings derived from Ima-

geNet pre-trained ResNet showed low heritability and no disease relevance to NAFLD.

Building on this, the software implementation of EmbedGEM has been meticulously crafted to offer easy extensibility, allowing the computation of additional metrics using the intermediate outputs produced by a highly standardized workflow. For example, users seeking to expand EmbedGEM with novel methods for assessing disease relevance, such as survival or multiple traits, can achieve this by capitalizing on the PRSs for each PC of the multivariate trait. Similarly, if users wish to modify the PRS computation process and employ their own custom tool, they can effortlessly do so using the clumped summary statistics of each PC. Furthermore, the introduction of new methods for evaluating heritability is feasible by leveraging the multivariate summary statistics generated by EmbedGEM.

We envision EmbedGEM as a key benchmarking tool for the systematic evaluation of different embeddings and other multivariate traits in their utility for genetic discovery. By providing a standardized platform for comparison, EmbedGEM provides researchers with a standardized approach to make informed decisions about the most effective embeddings for their specific genetic discovery tasks. This will not only streamline the process of genetic discovery but also foster a more collaborative and transparent research environment.

Acknowledgements

The authors would like thank the participants of the UK Biobank, whose data were used with permission. This research was conducted using the UK Biobank Resource under approved Application Number 51766.

insitro Research Team Banner and Contribution Statements

All contributors are listed in alphabetical order.

Downloading, preprocessing, and curation of UK Biobank data

- Francesco Paolo Casale,[1,2,3,4] Eilon Sharon,[1] Thomas W. Soare,[1] James Warren[1].

Software engineering

- **Statistical genetics pipeline**: Magdalena Borecka,[1] Francesco Paolo Casale,[1,2,3,4] Anna Merkoulovitch,[1] Colm O'Dushlaine,[1] Thomas W. Soare,[1] Paul Sud,[1] Baris Ungun[1].
- **redun development**: Robin Betz,[1] Edward Chee,[1] Patrick R. Conrad,[1] Kevin Ford,[1] Christoph Klein,[1] Donald Naegely,[1] Matthew Rasmussen[1].

Affiliations:

[1] insitro, South San Francisco, CA, USA.

[2] Institute of AI for Health, Helmholtz Munich, Neuherberg, Germany.

[3] Helmholtz Pioneer Campus, Helmholtz, Munich, Neuherberg, Germany.

[4] School of Computation, Information and Technology, Technical University of Munich, Munich, Germany.

Declaration of Interests

SM, ZRM, JP, AM, TS, DA, HS, CK, SS, DL, CP, members of the insitro Research Team, DK, CO, and TK are current or former employees and shareholders of Insitro.

References

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[2] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[3] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[5] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[6] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.

[7] Wei-Hung Weng and Peter Szolovits. Representation learning for electronic health records. *arXiv preprint arXiv:1909.09248*, 2019.

[8] Sumit Mukherjee, Laura Heath, Christoph Preuss, Suman Jayadev, Gwenn A Garden, Anna K Greenwood, Solveig K Sieberts, Philip L De Jager, Nilüfer Ertekin-Taner, Gregory W Carter, et al. Molecular estimation of neurodegeneration pseudotime in older brains. *Nature communications*, 11(1):5781, 2020.

[9] Matthias Kirchler, Stefan Konigorski, Matthias Norden, Christian Meltendorf, Marius Kloft, Claudia Schurmann, and Christoph Lippert. transfergwas: Gwas of images using deep transfer learning. *Bioinformatics*, 38(14):3621–3628, 2022.

[10] Christos Dadousis, Sara Pegolo, Guilherme JM Rosa, Giovanni Bittante, and Alessio Cecchinato. Genome-wide association and pathway-based analysis using latent variables related to milk protein composition and cheesemaking traits in dairy cattle. *Journal of Dairy Science*, 100(11):9085–9102, 2017.

[11] Ziqian Xie, Tao Zhang, Sangbae Kim, Jiaxiong Lu, Wanheng Zhang, Cheng-Hui Lin, Man-Ru Wu, Alexander Davis, Roomasa Channa, Luca Giancardo, et al. igwas: image-based genome-wide association of self-supervised deep phenotyping of human medical images. *medRxiv*, pages 2022–05, 2022.

[12] Khush Patel, Ziqian Xie, Hao Yuan, Sheikh Muhammad Saiful Islam, Wanheng Zhang, Assaf Gottlieb, Han Chen, Luca Giancardo, Alexander Knaack, Evan Fletcher, et al. New phenotype discovery method by unsupervised deep representation learning empowers genetic association studies of brain imaging. *medRxiv*, pages 2022–12, 2022.

[13] Taedong Yun, Justin Cosentino, Babak Behsaz, Zachary R McCaw, Davin Hill, Robert Luben, Dongbing Lai, John Bates, Howard Yang, Tae-Hwi Schwantes-An, et al. Unsupervised representation learning improves genomic discovery and risk prediction for respiratory and circulatory functions and diseases. *medRxiv*, 2023.

[14] Joshua C Denny, Marylyn D Ritchie, Melissa A Basford, Jill M Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R Masys, Dan M Roden, and Dana C Crawford. Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, 26(9):1205–1210, 2010.

[15] Gregory Gundersen, Bianca Dumitrascu, Jordan T Ash, and Barbara E Engelhardt. End-to-end training of deep probabilistic cca on paired biomedical observations. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, 2020.

[16] Jordan T Ash, Gregory Darnell, Daniel Munro, and Barbara E Engelhardt. Joint analysis of expression levels and histological images identifies genes associated with tissue morphology. *Nature communications*, 12(1):1609, 2021.

[17] Andries T Marees, Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis, Cynthia Marie-Claire, and Eske M Derks. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International journal of methods in psychiatric research*, 27(2):e1608, 2018.

[18] Mark I McCarthy, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John PA Ioannidis, and Joel N Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews genetics*, 9(5):356–369, 2008.

[19] Florian Privé, Bjarni J Vilhjálmsson, Hugues Aschard, and Michael GB Blum. Making the most of clumping and thresholding for polygenic scores. *The American Journal of Human Genetics*, 105(6):1213–1221, 2019.

[20] Yagoub Adam, Chaimae Samtal, Jean-tristan Brandenburg, Oluwadamilare Falola, and Ezekiel Adebiyi. Performing post-genome-wide association study analysis: overview, challenges and recommendations. *F1000Research*, 10, 2021.

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

[22] 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, 2020.

[23] Cathryn M Lewis and Evangelos Vassos. Polygenic risk scores: from research tools to clinical instruments. *Genome medicine*, 12(1):1–11, 2020.

[24] 1000 Genomes Project Consortium Corresponding authors Auton Adam adam. auton@ gmail. com 1 b Abecasis Gonçalo R. goncalo@ umich. edu 2 c, Production group Baylor College of Medicine Gibbs Richard A.(Principal Investigator) 14 Boerwinkle Eric 14 Doddapaneni Harsha 14 Han Yi 14 Korchina Viktoriya 14 Kovar Christie 14 Lee Sandra 14 Muzny Donna 14 Reid Jeffrey G. 14 Zhu Yiming 14, Broad Institute of MIT, Harvard Lander Eric S.(Principal Investigator) 13 Altshuler David M. 3 Gabriel Stacey B.(Co-Chair) 13 Gupta Namrata 13, Coriell Institute for Medical Research Gharani Neda 31 Toji Lorraine H. 31 Gerry Norman P. 31 Resch Alissa M. 31, Illumina Bentley David R.(Principal Investigator) 5 Grocock Russell 5 Humphray Sean 5 James Terena 5 Kingsbury Zoya 5, McDonnell Genome Institute at Washington University Mardis Elaine R.(Co-Principal Investigator)(Co-Chair) 22 Wilson Richard K.(Co-Principal Investigator) 22 Fulton Lucinda 22 Fulton Robert 22, et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[25] Zobair M Younossi, Aaron B Koenig, Dinan Abdelatif, Yousef Fazel, Linda Henry, and Mark Wymer. Global epidemiology of nonalcoholic fatty liver disease—meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology*, 64(1):73–84, 2016.

[26] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. 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, 2015.

[27] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[28] Yi Liu, Nicolas Basty, Brandon Whitcher, Jimmy D Bell, Elena P Sorokin, Nick van Bruggen, E Louise Thomas, and Madeleine Cule. Genetic architecture of 11 organ traits derived from abdominal mri using deep learning. *Elife*, 10:e65554, 2021.

[29] Taro Langner, Fredrik K Gustafsson, Benny Avelin, Robin Strand, Håkan Ahlström, and Joel Kullberg. Uncertainty-aware body composition analysis with deep regression ensembles on uk biobank mri. *Computerized Medical Imaging and Graphics*, 93:101994, 2021.

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

[31] insitro. Redun: A python package for managing computational workflows. `https://github.com/insitro/redun`, 2021.