Predicting GWAS peaks from autoencoder embeddings of genome wide experimental assays

Anna Banaszak, Niels Ringler, Dominika Młynarczyk, Alex Fastner

Supervisor: Felix Brechtmann

Abstract

Genome Wide Association Studies (GWAS) is a powerful but complex bioinformatics analysis. It is an analysis over many genomes to identify associations between genomic variants and certain traits or phenotypes. The problem we try to address in this project is that an association of a gene to a trait does not necessitate that it is causal. Thus we aim to incorporate more data from specific sources to try and find truly causal genes.
Our goal in this project is to create embeddings from GTEx and CRISPR data to use as input to a model predicting GWAS peaks. We do this using a Variational Autoencoder (VAE) to create these embeddings as well as to reduce the input dimensionality. We train the VAE model on both datasets individually and take the Latent spaces to use as embeddings later. To evaluate our results we compare our embeddings GTEx/CRISPR performance to various PCAs. PCAs were both provided by our supervisor but we also made our own. The embeddings we create are run through a type of ensemble machine learning method Extreme Gradient Boosting regression (XGBOOST) and we evaluate them based on R-squared values.
First we check the impact of covariates alone with the PCA of the GTEx and CRISPR data. The covariates surprisingly provide a lot of information to the prediction. Then we try both our generated CRISPR, and GTEx embeddings from our Variational Autoencoder individually. Next we try to improve on performance by using them both together. Then finally we test our best models on our test set. Our embeddings showed some improvement on some of the traits, but overall PCA was still better for 14 of 20 traits. The task is very noise prone and it proved difficult to achieve a high R^2 value.

Introduction

Motivation

Genome Wide Association Studies (GWAS) is an analysis of many genomes to identify associations between genomic variants and certain traits or phenotypes. These traits can be diseases like covid or cancer, but also biomarker traits such as red blood cell count or the amount of vitamin D. When comparing the Single Nucleotide Polymorphisms (SNPs) of individuals with a given trait to those without, one can find differences. If this difference is too large to have likely occurred by chance, one can say that this SNP is associated with the phenotype (trait) that one is examining. By using a large dataset and statistical power one hopes to link these genomic variants to a specific trait.
With few exceptions, most diseases are complex and are caused by many mutations or variants. Finding an association between certain SNPs with a disease/trait is helpful in narrowing the scope of what is involved. This is a primary analysis which can inform further investigation into specific genes/pathways and more specific biochemistry to better understand the biological truth of the trait/disease. One downside to GWAS is that many of the SNPs detected correlate with a given phenotype but are not causal. Filtering the causal SNPs from the correlated ones is a difficult task.

Goal

The goal of this project is to try to leverage information from other datasets, like gene expression, and protein-protein interaction, to further investigate possible gene-trait causality. To do that we create lower dimensional embeddings of these datasets and use them as an input for a regression task to predict GWAS peaks. Based on the obtained results one is able to trace back the contribution of a particular dataset, chromosomes or even genes into the prediction. We hope that this information could provide additional links between genes and GWAS peaks to better understand their interaction.
The main motivation to use a lower dimensional representation of the datasets is to reduce the number of features in the prediction task. If we were using raw datasets the regression would be ill-defined as the number of features would be greater than the number of samples. However, we also hope that dimensionality reduction can help in preserving the most important information in the particular datasets and reduce information redundancy.

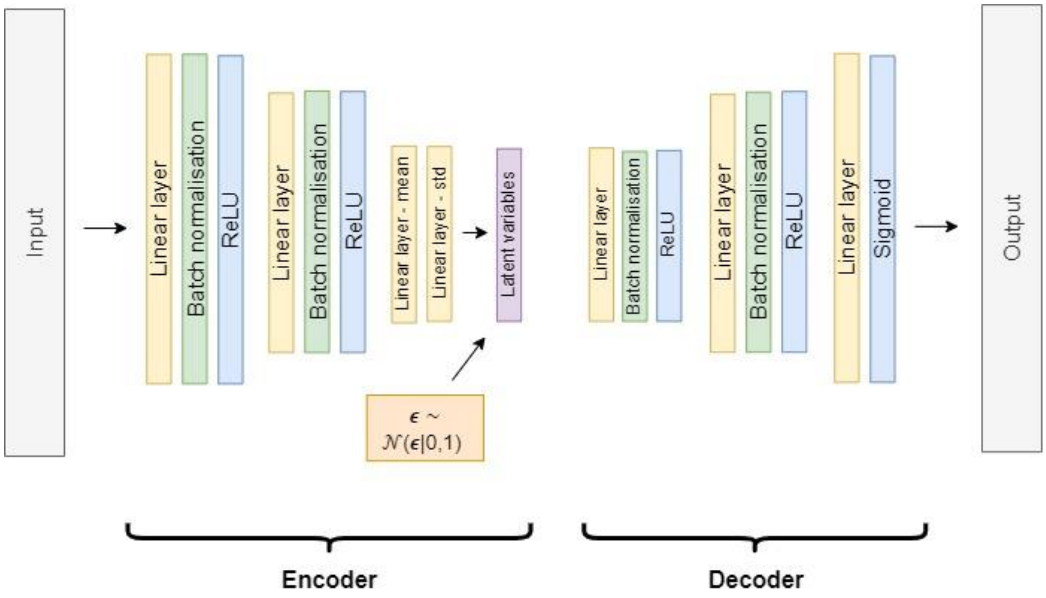
Datasets

In the project we used several datasets: GWAS, GTEx, CRISPR, STRING, and precomputed embedding derived from protein sequences (Elnaggar et al.).
The GWAS dataset we use comes from Pan UK-Biobank, a large database of paired genetic and phenotypic data. Gene aggregated GWAS data that was pre-processed with MAGMA (de Leeuw, C. A., (2015)) was also provided to us. This preprocessing takes given p-values and computes a Z-stat score which is the GWAS prediction.
For our analysis we focused on two datasets: GTEx and CRISPR. In particular, we worked with the most current release (GTEx Analysis V8) of the GTEx Portal.

The GTEx Project studies tissue-specific gene expression and regulation. This dataset includes samples from 54 non-diseased tissue sites across 1000 individuals, which results in 18527 genes each with 17384 features.
The CRISPR dataset comes from Project Achilles. We worked with the DepMap Public 18Q3 release. By using CRISPR-Cas-9 to select and knock out individual genes, it is possible to identify the ones crucial to a cell lines survival. We use this data as one of our inputs to our VAE with the hope of gaining information about the cruciality of certain genes and possible associations. The CRISPR dataset contains 17591 genes from 482 tissues.
We also use precomputed embeddings from STRING protein-protein interactions and precomputed embedding derived from protein sequences (Elnaggar et al.). to test our performance later.

Variational Autoencoder

Variational Autoencoder (VAE) (Kingma et al) is a neural network architecture, composed of an encoder and a decoder. It is trained to minimize the error between the input and the encoded and then reconstructed data. The input is mapped to a distribution over the latent space. Latent variables are then sampled from the obtained distribution and fed into the decoder. The loss function consists of the reconstruction term and the regularization term, which ensures that the distribution is close to normal and the latent space is better organized. VAE can be used for dimensionality reduction. The size of the latent variables layer is smaller than that of the input and causes the model to learn how to describe given data using a reduced number of features.



PCA

Principal component Analysis (PCA) is a statistical process of finding the components that explain the variance in your data the best. This is a great way to quickly find a low dimensional representation that still holds most of the information of the data. Principal components are the linear combination that explains the most variance, in descending order. By computing the principal components and using them to perform a change of basis ignoring some lower principal components it is possible to reduce dimensionality.
Later on we compare our embeddings generated with VAE with just using PCA provided by our supervisor and a PCA we generated ourselves.

Evaluation metric

R-squared (R^2) is a value between 0 and 1, and denotes the proportion of the variation in the dependent variable that can be explained by the independent one. It can be interpreted as a goodness-of-fit of the regression model.

Methodology

Overview

Although our contribution mainly focuses on creating well performing embedding for tabular datasets, it is crucial to present the architecture of the prediction method (provided by Felix Brechtmann). This is at the same time the evaluation method for assessing the feasibility of created embeddings. Input to the regression task is the multimodal data and output is the prediction of the aggregated Z-stat value. Performance evaluation of the regression is done using aggregated R^2 values among all the genes.

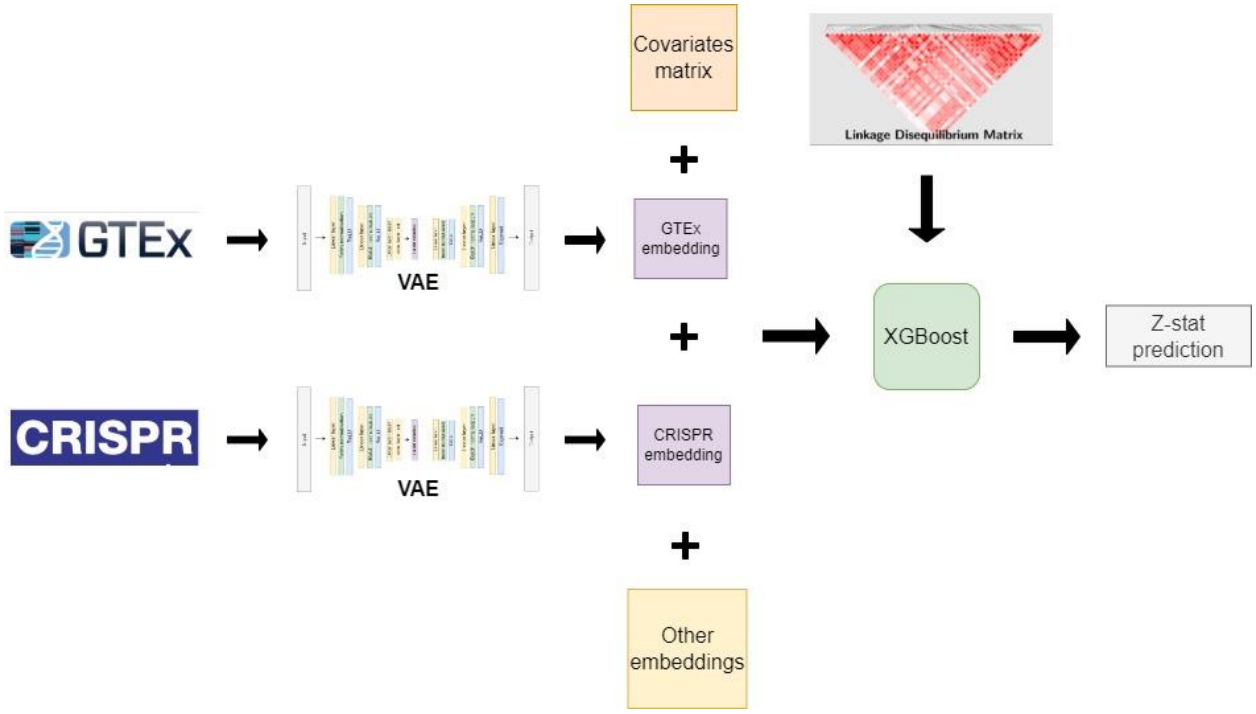
Evaluation framework

As GWAS peaks are a continuously valued data used machine learning technique for prediction is the Extreme Gradient Boosting regression (XGBoost) Chen, T. (2016, August) - a type of ensemble machine learning method. The loss function is Mean Squared Error.

The first step of the evaluation is preprocessing data obtained from MAGMA software. It is later used to create the covariates matrix and Linkage-disequilibrium matrix. LD matrix is the non-random association between alleles at given loci. When 2 loci have an association of genes that exceeds a threshold provided by random chance, they are said to be in Linkage-disequilibrium. The Covariates matrix contains additional data about the genes such as genesize and gene density which are additional independent factors.

The architecture of the evaluation framework is presented on the Figure below. The LD matrix is used to project the Z-stat values into LY. Embeddings are concatenated to covariates matrix and later fed into the XGBoost regressor along with LY as the regression objective. As the final step

the backprojection using Linkage-disequilibrium (LD) matrix is performed on the regression output to obtain the actual results.



In order to assess the quality of created embeddings we performed a comparative evaluation based on aggregated R^2 value with the use of a modified cross validation method. We used leave-one-chromosome-out strategy, which meant that in every fold genes coming from the same chromosome were left out as a validation set. The main objective was to outperform the provided PCA embeddings for particular data sets separately, and with concatenated embeddings as a final test. We also investigated the contribution of each dataset to the predictions.

Tabular data preprocessing

In order to reduce technical biases caused by irrelevant genes filtering out all non protein coding genes was performed as a common preprocessing step among both datasets. Afterwards, we applied min-max normalization to rescale data to range in [0,1] to assure stability of dimensionality reduction with use of VAE model.

GTEx

We applied the log10 transformation to the GTEx data in order to reduce the magnitude of the absolute values and the skewness of the distribution. This preprocessing step improved the results of the VAE model. Moreover, we investigated the usefulness of mean-centering the data by gene, but did not observe an improvement in the results.

CRISPR

The idea of using the CRISPR dataset comes from the assumption that the degree of correlation of the gene knockout profiles reflects their functional relationship. However, there are technical confoundings that we took into account as described in (Boyle et al., (2018)). In particular, it is known that olfactory receptors often expose highly correlated profiles, which leads to technical biases in the end. Also in order to statistically detect coessential interactions, we followed an approach based on generalized least squares (GLS) as described in (Wainberg et al., (2019)), resulting in a symmetric gene by gene matrix of p-values. We did not only focus on the p-values, but also examined the residuals gained by the GLS. However, it turned out that none of the preprocessing steps improved the performance of the final embedding. We conclude that the VAE is capable of outperforming any of our attempts of preprocessing by training on the raw and unprocessed data.

Dimensionality reduction using VAE

In order to find the best model, we performed a wide model search in terms of VAE architecture and model hyperparameters. The best performing model had 3 linear layers with ReLU activation function. The loss function was a sum of Mean Squared Error and KL-divergence between reconstruction and input. We used an Adam optimizer with the learning rate and learning rate scheduler tuned for each dataset separately. Sizes of linear layers were adjusted to the size of input data accordingly. In order to find the optimal number of training epochs, the data was split into training and validation sets with ratio 3:1.
After finding the most suitable VAE model, the whole dataset was compressed into lower dimensional embedding using the trained encoder.

GTEx

In the best performing model on the GTEx data, the encoder and the decoder consisted of three fully connected layers with the output sizes 1024, 512 and 100. Batch normalization was not used. The starting learning rate was 0.0003. The loss was calculated as a sum of MSE and KL-divergence with ratio 3:2.

CRISPR

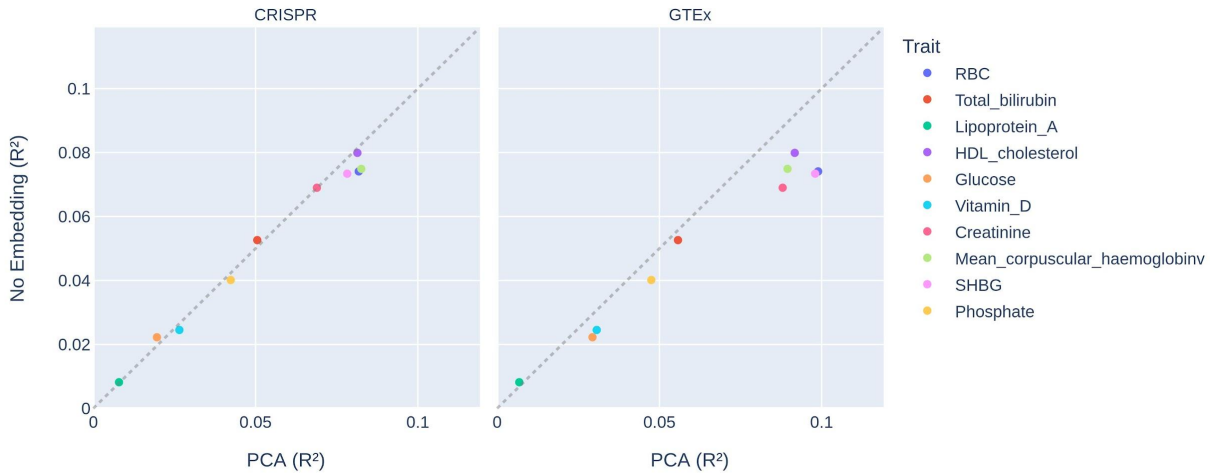
Also for the dataset of CRISPR screens we used encoder and decoder with three fully connected layers, with batch normalization. Output sizes of the best performing model were 320, 256 and 100. We started with a learning rate of 0.001 and decreased it after 100 epochs of learning. The loss was calculated as a sum of MSE and KL-divergence as with the GTEx dataset. We achieved similar performance on the validation error of the VAE with many different settings of hyper-parameters. Nonetheless the GWAS predictions of the embeddings were noticeably different.

Results

In the following experiments, PCA embeddings of CRISPR and GTEx datasets were treated as a baseline over which we tried to improve the regression performance. Results of our project consist of several comparative evaluations. To avoid information leakage we chose several traits as the validation set (RBC, Total bilirubin, Lipoprotein A, HDL cholesterol, Glucose, Vitamin D, Creatine, Mean corpuscular hemoglobin v, SHBG, Phosphate) and others as the test set after choosing the best embeddings. All of the following experiments, except for the final one (Experiment 5), were performed on the validation set.

Experiment 1 - datasets contribution

First, we compared the performance of the regression on the covariates matrix alone and with PCA embedding of GTEx or CRISPR datasets. The idea behind this test was to see the contribution of each dataset to the explained variance.



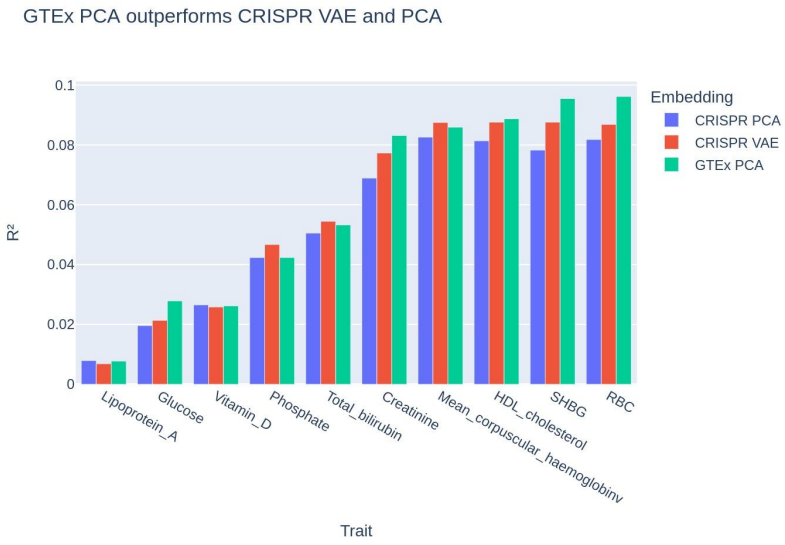
Traits which tend to have higher R^2 values, like RBC, SHBC, HDL cholesterol, benefit more from providing embeddings as additional data to the regressor input. We can observe similar behavior in both datasets. The contribution of the GTEx dataset is relatively high compared to the contribution of CRISPR.

Experiment 2 - CRISPR embeddings study

A comparative study between different embeddings of CRISPR dataset.

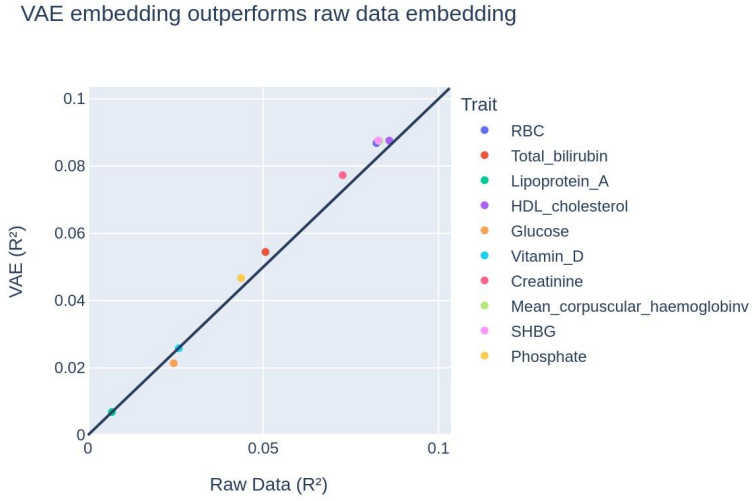
First of all we noticed that the PCA with 256 dimensions that we used as baseline model did not predict GWAS signals better than the raw data used as embedding. It showed only slight improvement on traits that were explained with a comparably small R^2, in particular Lipoprotein A and Vitamin D.

Reducing the number of principal components obviously also reduced the prediction power of the PCA embedding. We noticed that a truncation of the principal components down to 50 explained 47% percent of the variance of the dataset, but did not contribute anything to the GWAS prediction anymore. It performed similarly to a random vector used as embedding or using no embedding at all and the covariates only.



In contrast to that, the PCA of the GTEx dataset predicted GWAS signals more successfully. We verified this by comparing the R^2 values of both PCAs on GTEx and on CRISPR screens datasets.

Although these findings suggested that the CRISPR screens dataset might not contain enough useful information for a solid GWAS prediction, we tried the VAE approach and thus extracting a structured low dimensional representation that would reflect the coessentiality of the genes. In the end, our best VAE embedding was able to outperform the CRISPR PCA embedding. It even outperformed the raw data as embedding in all of the traits except Phosphate.



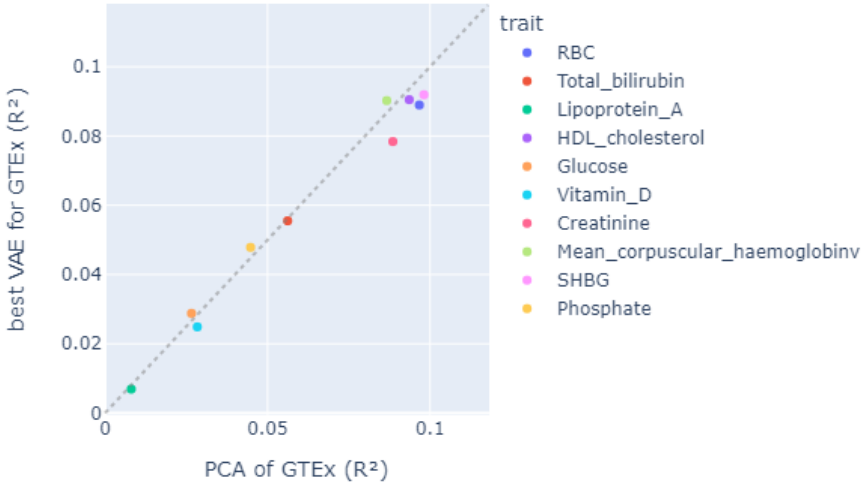
Experiment 3 - GTEx embeddings study

A comparative study between different embeddings of GTEx dataset.

Best VAE embedding vs precomputed PCA embedding

The VAE model achieving the best results on GTEx data had three linear layers, no batch normalization and a latent space of size 100. The dataset was preprocessed using the log10 transformation.

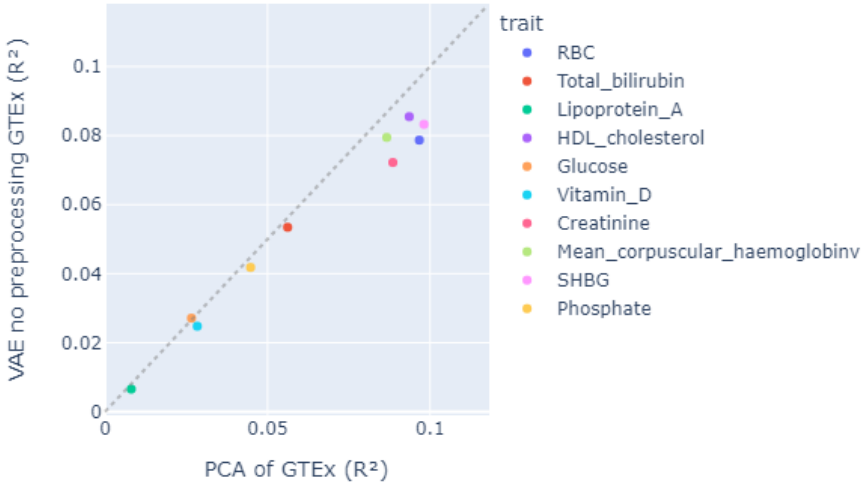
The best performing embedding of the GTEx data obtained using the VAE model did not manage to improve over the precomputed PCA in most of the validation traits. The results for traits: Glucose, Mean corpuscular hemoglobin and Phosphate were slightly better than the ones obtained by the PCA embedding.



VAE with no preprocessing vs precomputed PCA

A VAE model, which performed the best on raw GTEx data had three linear layers, batch normalization and latent space of size 256.

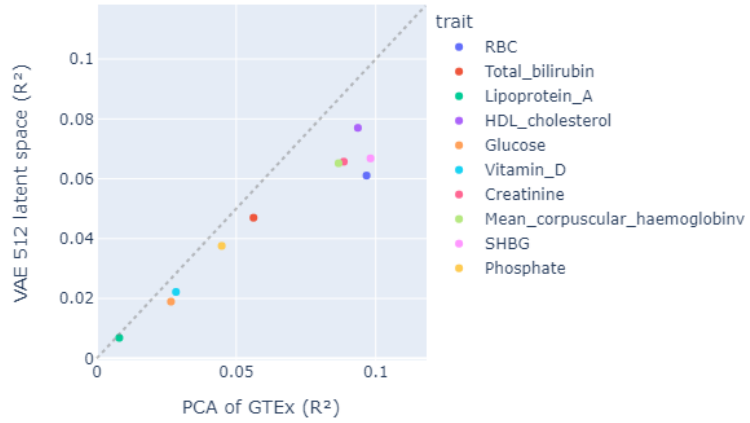
Without applying any preprocessing to the GTEx data the embedding achieves worse results than the PCA embedding in all of the traits except for Glucose.



VAE embedding with 512 dimensions vs precomputed PCA

A VAE model with three linear layers, no batch normalization and a latent space of size 512 applied to data preprocessed using log10 transformation.

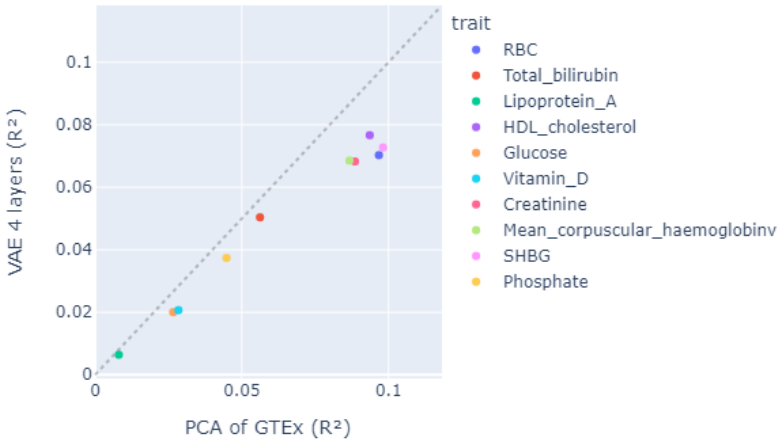
VAE embedding with more dimensions worsens the results achieved for all traits. However, we can observe that the deterioration is bigger for traits that tend to have higher R^2 values, like RBC, SHBC, HDL cholesterol.



VAE with bigger hidden layers vs precomputed PCA

A VAE model with four linear layers with increased sizes (8192, 2048, 512, 100), batch normalization and the latent space of size 100 applied to data preprocessed using log10 transformation.

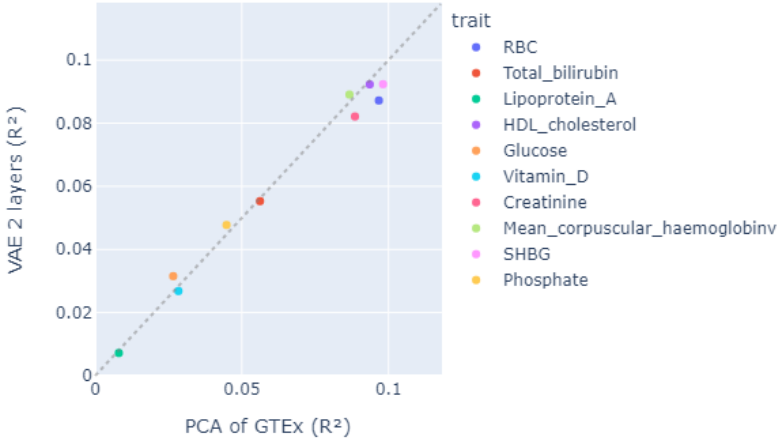
Using more complex VAE architecture does not help the model to better describe the input data and does not result in any improvements.



VAE with two hidden layers vs precomputed PCA

A VAE model with two linear layers (1024, 100), batch normalization and the latent space of size 100 applied to data preprocessed using log10 transformation.

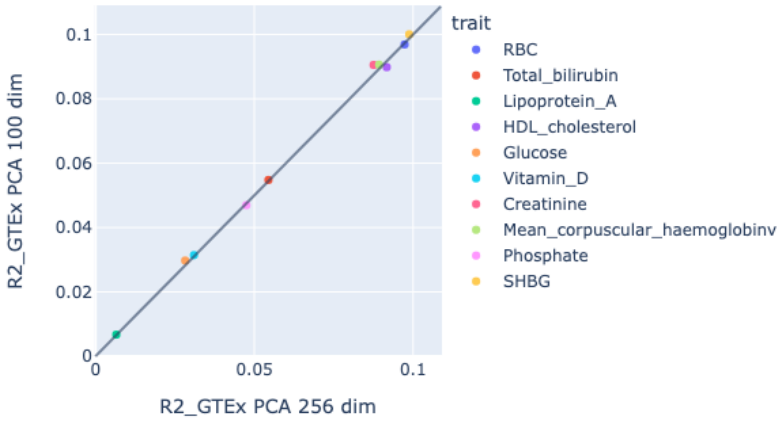
In this case the results were comparable to those of the best VAE embedding shown before. The results for traits: Glucose, Mean corpuscular hemoglobin and Phosphate were slightly better compared to PCA.



PCA embedding with 256 dimensions vs 100 dimensions

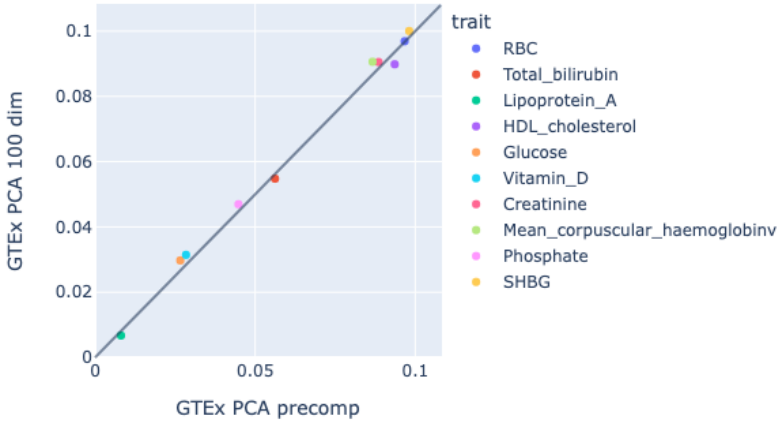
Since we did not manage to create a VAE embedding on GTEx data which would outperform provided precomputed PCA embedding we decided to create our PCA embedding on normalized GTEx data. We aimed to find a more accurate representation by reducing the number of dimensions. This Pca is performed on normalized GTEx data.

Reducing the number of dimensions from 256 to 100 did not significantly affect regression performance although smaller embedding has slightly better results.



Our PCA embedding vs precomputed PCA embedding

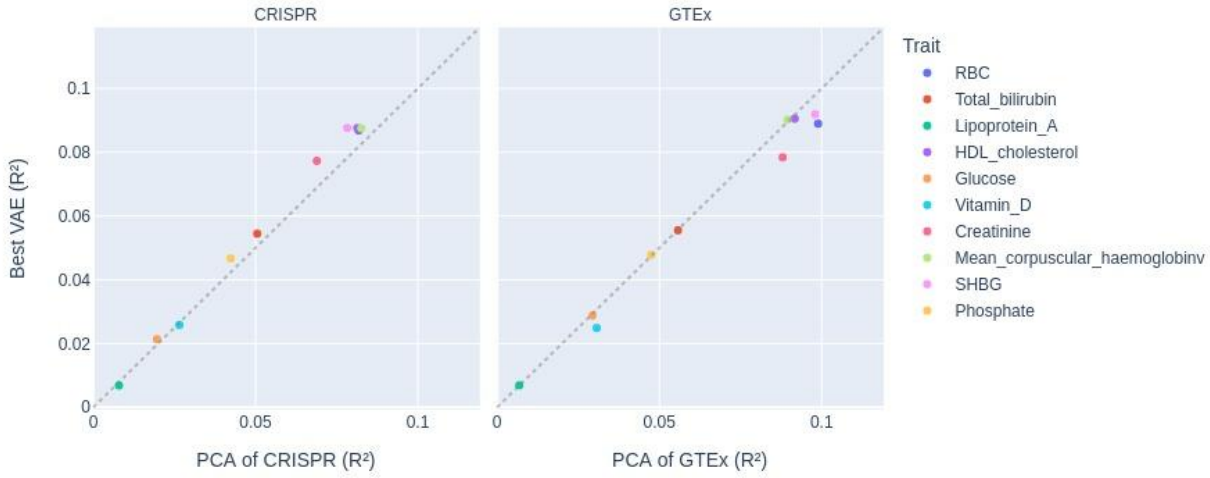
We managed to outperform the provided 256-dimensional PCA embedding with our own 100-dimensional PCA embedding on normalized data.



Experiment 4 - VAE embeddings for CRISPR and GTEx

Comparative study between best CRISPR and GTEx embeddings. The aim of this experiment is to assess which dataset is performing better in terms of aggregated R^2 value.

VAE embedding vs PCA embedding for CRISPR and GTEx

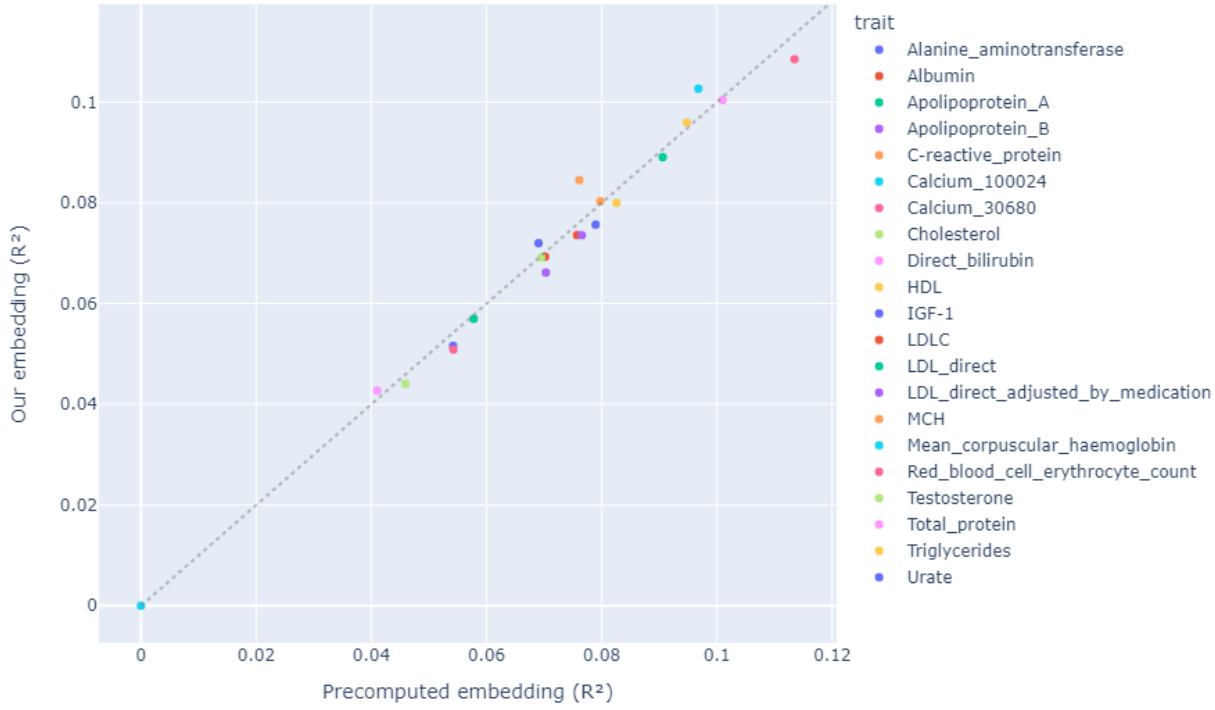


The best performing embedding of the GTEx data obtained using the VAE model did not manage to improve over the PCA algorithm in most of the validation traits. The results for traits: Glucose, Mean corpuscular hemoglobin and Phosphate were slightly better.
However we managed to outperform the PCA of the CRISPR screen dataset with a VAE embedding of the CRISPR screen dataset. Since the latter dataset is comparatively much smaller, we conclude that the VAE is able to support less informational input data better.
Still the best performing VAE embedding on the CRISPR screen dataset could also not outperform the PCA of the GTEx dataset.

Experiment 5 - final tests of created embeddings

Combined best VAE embeddings vs provided precomputed one

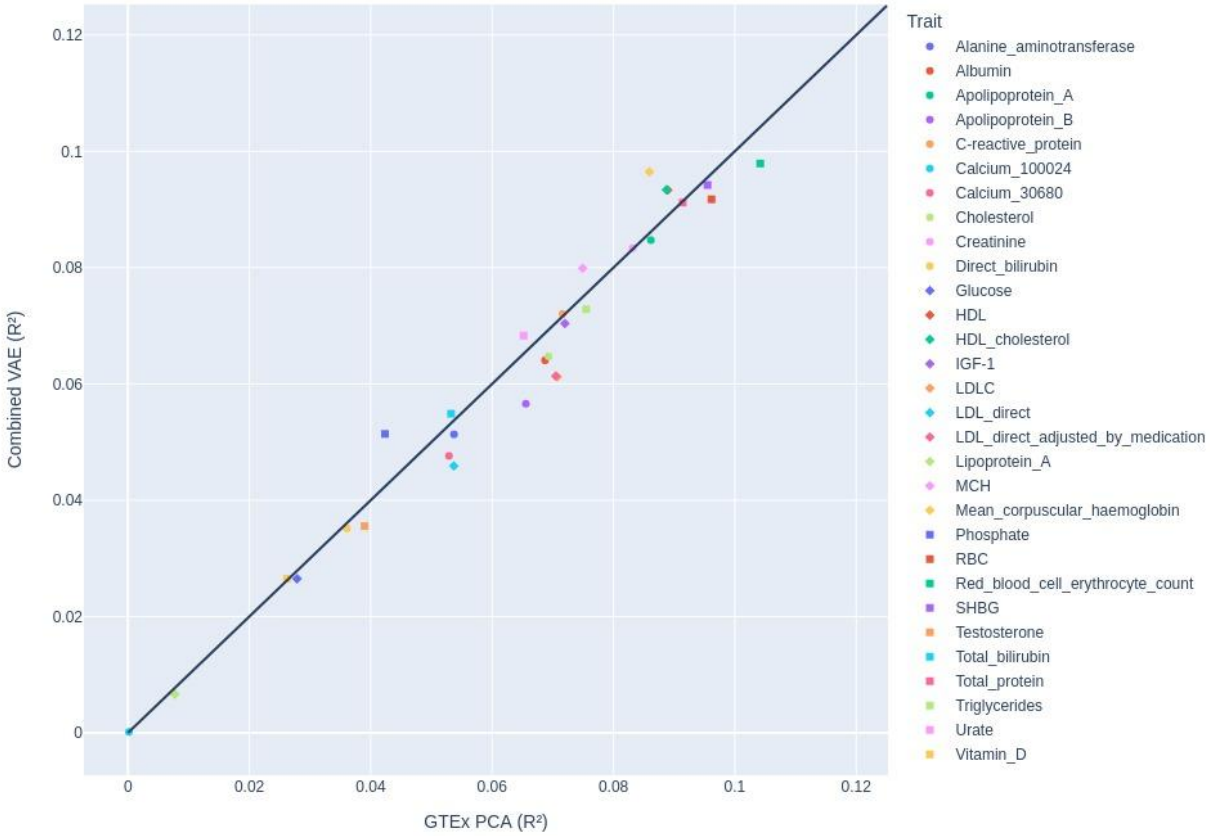
Test on the concatenated embeddings (best VAE embedding of GTEx, best VAE embedding of CRISPR, precomputed embedding from STRING protein-protein interactions, precomputed embedding derived from protein sequences) vs provided embeddings on traits not used in validation above tasks above.



The embedding created using our best results from VAE models achieves better results for 6 out of 20 traits from the test set. The precomputed embedding explains the majority of traits better.

Combined VAE embedding for CRISPR and GTEx vs PCA embedding for GTEx

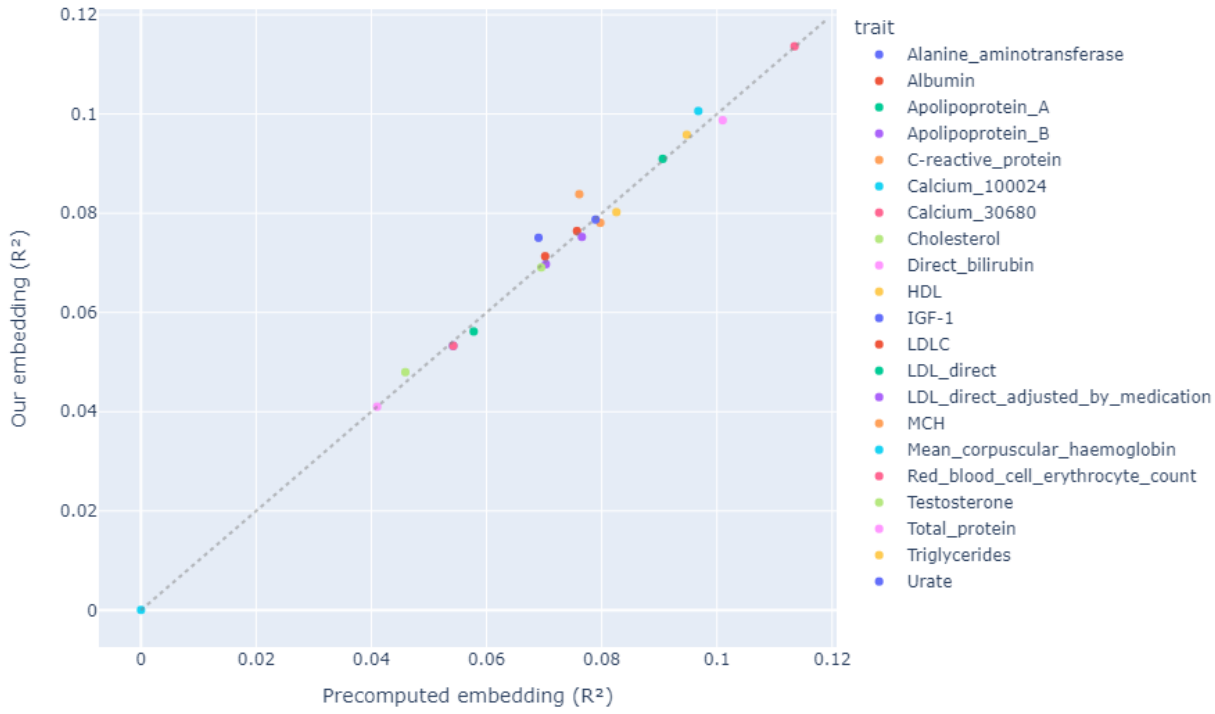
Trait	GTEx PCA (R^2)	Combined VAE (R^2)	R^2 increase (%)
Phosphate	0.042	0.051	21.4
Mean_corpuscular_haemoglobin	0.086	0.096	12.3
MCH	0.075	0.080	6.6
HDL_cholesterol	0.089	0.093	5.3
HDL	0.089	0.093	5.0



The concatenated CRISPR and GTEx embeddings showed good results in certain traits: They had an increase in the R^2 value of more than 20 % on phosphate compared to the PCA of GTEx. Mean corpuscular hemoglobin, MCH and HDL increased in R^2 value by 5 to 10 %. However the PCA of GTEx could explain the majority, that is 19 from 30 traits better.

Combined best created embeddings vs precomputed one

Test on the concatenated embeddings (best PCA embedding of GTEx, best VAE embedding of CRISPR, precomputed embedding from STRING protein-protein interactions, precomputed embedding derived from protein sequences) vs provided embeddings on traits not used in validation above tasks above.



In 11 out of 21 traits we see a performance improvement of created embeddings over the provided embeddings. In some of the traits, like Apolioprotein_B or IGF-1, the improvement is significant. This concludes that using VAE embedding for CRISPR dataset and 100-dimensional PCA embedding on normalized GTEx data results in higher R^2 values not only separately but also as a combined input to the regression task.

Discussion

One of the most important observations is the high contribution of covariates matrix to the R^2 values in our evaluation. Adding embeddings to the regression input did not result in significant improvements in aggregated R^2 values. Obtained R^2 values are generally low, which might be blamed on the nature of the task and low signal noise ratio in the GWAS data.
Both PCA and VAE embeddings performed better when having 100 dimensions compared to 256. Simpler models worked better for preprocessed GTEx data; increasing the number of hidden layers or expanding the sizes in order to reduce the number of dimensions of the input data was not helpful.
Moreover it takes time to get familiar with the datasets and understand how they relate to each other. Lack of previous domain knowledge might be a significant overhead in projects of this kind.
Dimensionality reduction using a VAE was useful for the CRISPR screen dataset but not for the GTEx dataset. Hence we assume that the VAE approach can be helpful, especially for smaller dimensional input. Also, from a practical point of view, tuning the VAE model on a large dataset is an expensive task in terms of computational resources, which makes it hard to use it on the

GTEx data. VAE embedding might not result in a significant performance improvement and has a much higher computational overhead.
Since most of the preprocessing that we tried on the CRISPR screen dataset did not improve the performance of the final embedding, we conclude that the VAE is capable of learning many of the necessary preprocessing steps on its own.

Conclusion

We were able to find a better performing embedding compared to the existing one although we had to leverage both of the dimensionality reduction methods - PCA and VAE. Choosing the right dimensionality reduction method depends on the data itself. Another decision factor might be the computational complexity and execution or development time constraints.

References

1. Weeks et al. 2020; Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases
2. Trofimov et al. 2020; Factorized embeddings learns rich and biologically meaningful embedding spaces using factorized tensor decomposition
3. GTEx Portal. (n.d.). Retrieved July 22, 2022, from https://gtexportal.org/home/
4. DepMap: The Cancer Dependency Map Project at Broad Institute. (n.d.). Retrieved July 22, 2022, from https://depmap.org/portal/
5. STRING: functional protein association networks. (n.d.). Retrieved July 22, 2022, from https://string-db.org/
6. Pan UKBB | Pan UKBB. (n.d.). Retrieved July 22, 2022, from https://pan.ukbb.broadinstitute.org/
7. Index of /public/lecture/ml4rg/gene_embedding_projects. (n.d.). Retrieved July 22, 2022, from https://www.cmm.in.tum.de/public/lecture/ml4rg/gene_embedding_projects/
8. de Leeuw, C. A., Mooij, J. M., Heskes, T., & Posthuma, D. (2015). MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLoS Computational Biology, 11(4), e1004219. https://doi.org/10.1371/journal.pcbi.1004219
9. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system
10. Kingma, Diederik & Welling, Max. (2014). Auto-Encoding Variational Bayes.
11. Boyle et al. 2018; High-resolution mapping of cancer cell networks using co-functional interactions
12. Wainberg et al. 2019; A genome-wide almanac of co-essential modules assigns function to uncharacterized genes