Use of Phenotypes in Deep Learning-Based Dimensionality Reduction

UPPSALA UNIVERSITET

Filip Thor filip.thor@it.uu.se   Carl Nettelblad carl.nettelblad@it.uu.se

Division of Scientific Computing, Department of Information Technology   Science for Life Laboratory (SciLifeLab)

Code and Additional Figures:

Introduction

Dimensionality reduction is usually a first step for researchers wanting to study the population structure of a genetic dataset. We develop SNP-based dimensionality reduction models using deep learning. In this work, we leverage phenotypes both as a means of evaluating embeddings and as extra information when training the model to incorporate more information in our embeddings.

Use of PCA in Population Genomics

PCA is a linear method that computes a lower dimensional representation with the aim of preserving as much variance as possible. Often in population genomics, PCA is applied to a dataset to visualize the data, showing that samples from the same population have a similar genetic makeup. It is used as a de facto standard technique, but the accuracy of the mapping is seldom discussed, and newer methods like t-SNE can resolve relations at a finer scale.

Why Use Deep Learning?

With neural networks (NN) we can use more information in the data, such as the ordering and position of the SNPs.

We find an explicit mapping $f : \mathbb{R}^D \rightarrow \mathbb{R}^d$, and thus projecting new samples onto a learned embedding is easy, in contrast to non-parametric models. The framework is versatile, and a good higher-dimensional feature extraction should be a good foundation for further work on phenotype prediction.

Contrastive Learning for Dimensionality Reduction

We utilize **Contrastive Learning** (Figure 1), which minimizes distances in the embedding space between similar samples and maximizes distances to dissimilar ones, while using no information on population labels.

This is done by assigning each sample positive and negative samples. The network is trained by penalizing large embedding distances for similar samples. The positives are defined as being masked views of the original, and the negatives are taken as some random sample in the batch. The implementation is built on top of the deep learning framework GCAE [1].

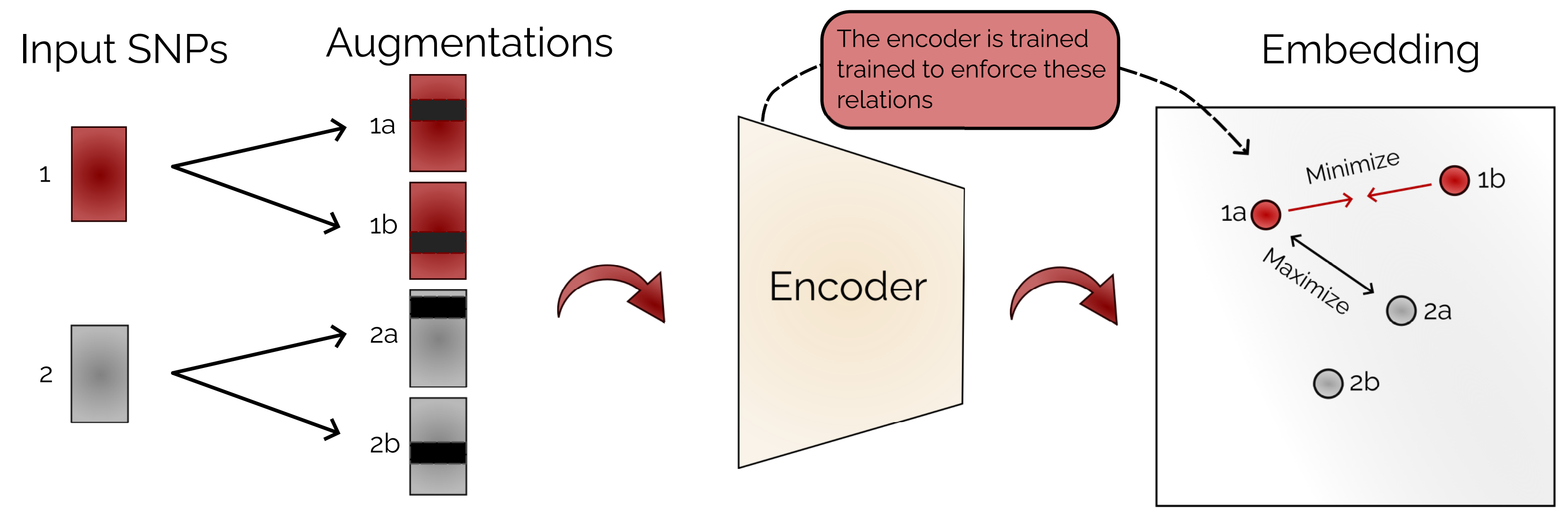

Figure 1. Contrastive learning concept. An encoder neural network model is trained to embed two slightly altered versions of the same sample close to each other.

Quantitative Evaluation of Embeddings

Population classification A good embedding should map samples of similar origin or ancestry close. It has been observed that t-SNE captures local structure well but has worse global performance, while the opposite is true for PCA. As a measure of performance of how well the method groups samples, we use a KNN classifier for the population label, and report the F1-score.

The predicted class y_i for sample i is computed as $y_i = \text{argmax}\left(\sum_{j \in \mathcal{N}_i^k} \tilde{y}_j\right)$, where $\tilde{y}$ is a one-hot encoding of the data label, and $\mathcal{N}_i^k$ is the set of k closest samples in the embedding space. By varying the value of k, we can see how well the methods perform on different scales.

Phenotype prediction The other way we propose to evaluate an embedding is to look at its predictive power for phenotypes. A good embedding should retain the vital information needed for a phenotypic regression model to perform well. Given a projection method $\mathcal{P} : \mathbb{R}^D \rightarrow \mathbb{R}^d$, a dataset $\{\mathcal{X}_i, y_i\}_{i=1}^N$, and a regression model $\mathcal{R}$, we evaluate the Pearson correlation coefficient p for the predicted $\hat{y}$ and true phenotypes y, where $\hat{y} = \mathcal{R}(\mathcal{P}(\mathcal{X}))$.

Embeddings of Potato Data

Setting the embedding dimension to $d = 2$, we can produce PCA-like visualizations. Figure 2 shows an example using a potato dataset (669 samples, 8 classes, 7k SNPs) [2]. The samples are part of a breeding program, where T1 is the first generation, and T2/T3 are generations that have undergone selection. Here we use PCA, t-SNE, and two versions of contrastive learning. They all yield some clustering of the data, we now turn to attempting to quantify their differences.

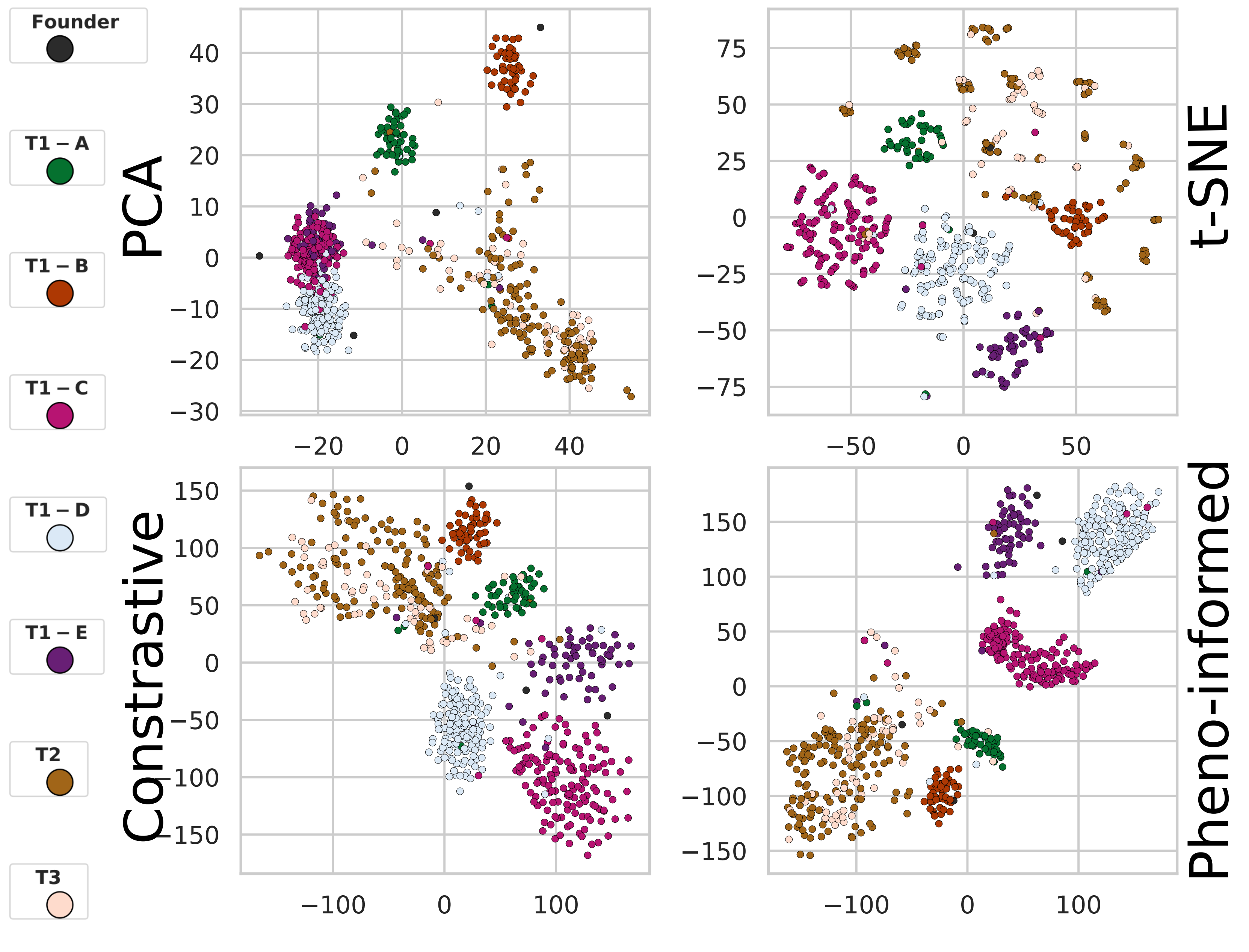

Figure 2. Dimensionality reduction using PCA, t-SNE, and our deep learning methods on a potato dataset.

Incorporating Phenotypes in Training

The deep learning framework is versatile, where model architecture and training strategy can be tailored to the application. As is the definition of the similarity of samples in the contrastive learning setting, which enables the following:

Idea: Choose negatives as the samples with the largest phenotypic differences.

We evaluate our deep learning model both with and without the use of phenotypes and compare them with contemporary methods, all using embeddings of **dimension 2**. All presented results show means quantities over 10 different train/test splits.

Figure 3 shows the population classification score, using a KNN classifier with varying k, which shows the trends in model performance on different scales. The score is computed over all samples. Table 1 shows the phenotype prediction performance using the full genotypes, and the embeddings, and Figure 4 shows the KNN phenotype prediction score for varying k.

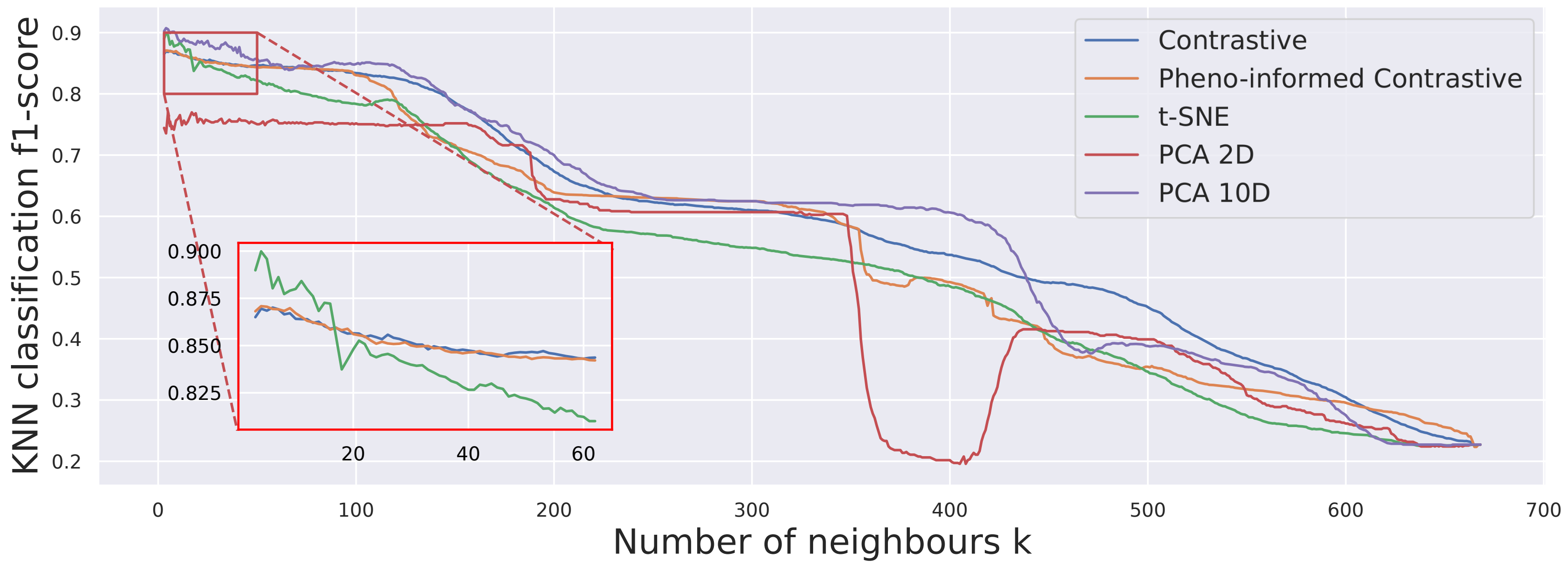

Figure 3. Population classification F1 score using a KNN classifier, with increasing k.

Method/input	Full data	2D PCA	2D t-SNE	2D Cont.	2D Cont . pheno
GBLUP	0.7016	0.6735	0.5635	0.5798	0.6167
KNN	0.6951	0.6958	0.6912	0.7013	**0.7082**
Random Forest	**0.7199**	0.6685	0.6747	0.6945	0.6966

Table 1. Phenotype prediction (tuber number) test performance in terms of Pearson correlation coefficient for an 80/20 train/test split for three different regression models using both the full set of SNPs and 2D embeddings.

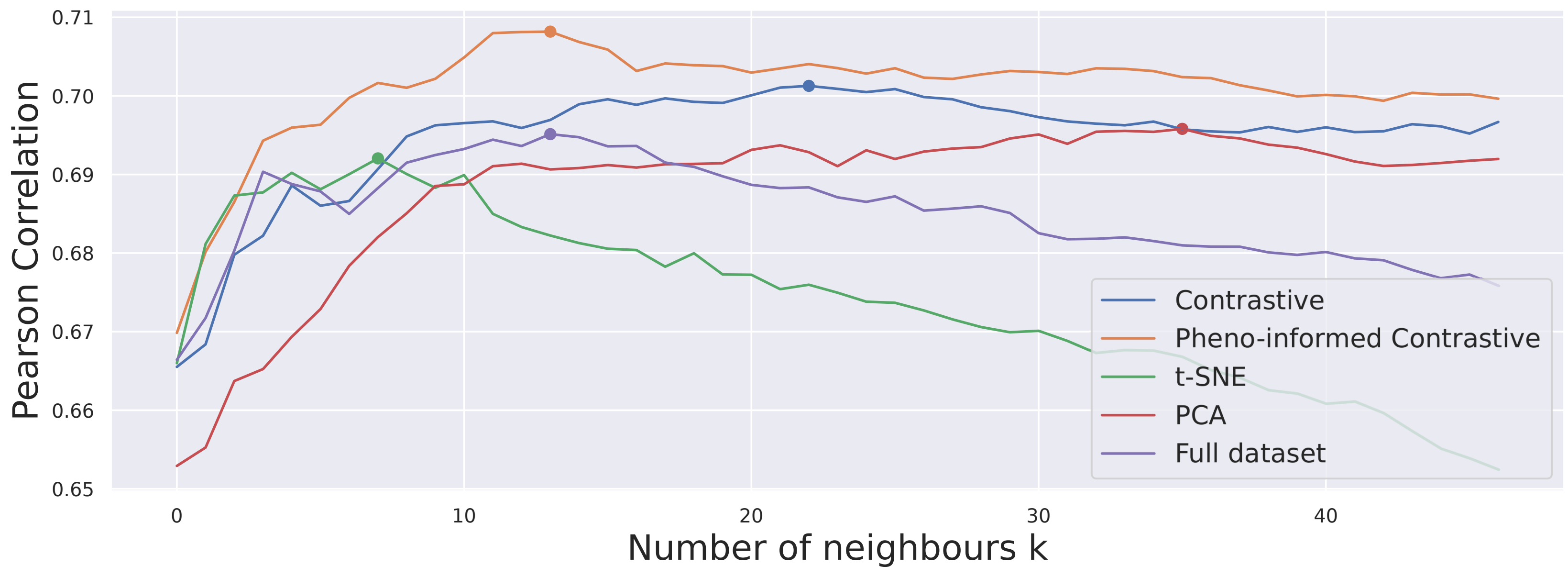

Figure 4. Phenotype prediction score using different values of k in KNN regression.

Conclusions and Future Work

The contrastive learning implementation performs slightly worse than t-SNE in peak F1-score on population classification. Still, it retains global structure better, as seen in the t-SNE embedding F1-score dropping quickly as k increases.

The produced 2D embeddings can be used in phenotype prediction, with close to the same performance as on the full 7K SNP dataset. Incorporating phenotypes in training improves phenotype prediction, without drastically lowering the population classification score.

For both PCA and t-SNE, the embeddings have used the genotypes of all samples before fitting the regression model, which is not the case for the NN models. This is a strength of a parameterized model which (unlike t-SNE) has a straightforward way of running inference.

Future work should focus on developing methods for causal genotype-phenotype relations, as this is currently missing in the contrastive learning approach.

Acknowledgements and References

Compute resources provided by SNIC through National Supercomputer Centre NSC) at Linköping University under Project Berzelius 2022-17, 2022-165, and 2023-258. Project funded by Formas, The Swedish government research council for sustainable development, Grant no 2020-00712, Deep Learning for Analyzing Population Structure and Genotype-Phenotype Mapping.

[1] K. Ausmees and C. Nettelblad. A deep learning framework for characterization of genotype data. *G3 Genes|Genomes|Genetics*, 12(3), 01 2022. ISSN 2160-1836. doi: 10.1093/g3journal/jkac020. jkac020.

[2] C. Selga, F. Reslow, P. Pérez-Rodríguez, and R. Ortiz. The power of genomic estimated breeding values for selection when using a finite population size in genetic improvement of tetraploid potato. *G3 Genes|Genomes|Genetics*, 12(1):jkab362, 10 2021. ISSN 2160-1836. doi: 10.1093/g3journal/jkab362. URL https://doi.org/10.1093/g3journal/jkab362.