On two dimensional mappings of SNP marker data and CNNs: overcoming the limitations of existing methods using Fermat distance.

J. Elenter[1], G. Etchebarne[1], I. Hounie[1], M.I. Fariello[1], F. Lecumberry[1]

1. Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay

Introduction and motivation

In recent years, **Convolutional Neural Networks** have attracted great attention, establishing state-of-the-art results in many fields, most notably, in Computer Vision. Not only do CNNs drastically reduce the number of parameters in comparison to fully-connected networks, but they also inherit adequate inductive biases for dealing with images, such as translational invariance. In an attempt to leverage their success and ubiquity, approaches mapping non-euclidian data into two dimensional image-like feature maps, which are used as inputs to CNN architectures, have been proposed.

Convolutional models exploit local structure in signals. In the case of genomic sequences, this local structure corresponds to **linkage disequilibrium** and **cross-marker** interaction phenomena [1]. Although similar SNPs are often neighboring markers, they can also be located in distant regions of the genome. Thus, simply sliding a convolutional kernel through the genomic sequence (1D Convolution) may prevent the model from capturing interactions between distant SNPs.

This suggests that it may be beneficial to find a representation of the genome in which **similar SNPs** are close and dissimilar SNPs are further apart. Such alternative representation of the genome could catalyze the extraction of structural information via convolutions. In this sense, an image-like (2-dimensional) representation seems like a coherent choice since it enables the use of popular image processing tools and two-dimensional CNNs.

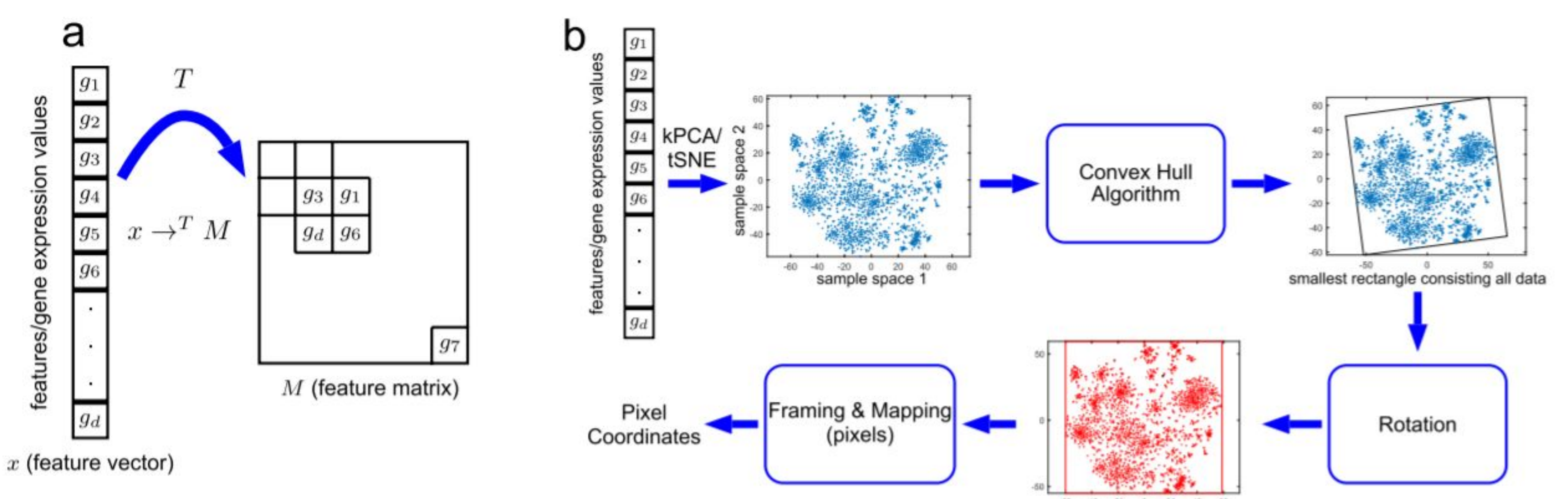
Data description

The yeast dataset [2] consists of **1008 samples** of **yeast** which were obtained as the cross of a laboratory and a wine strain. Both parent strains correspond to the species *Saccharomyces cerevisiae*.

The sequences of raw data were post-processed into 30.594 high confidence SNPs which were uniformly sampled to obtain the final **11.623 binary markers**. These were coded as follows: 1 if the sequence variation came from the wine strain and 0 if it came from the laboratory strain. The phenotype in question is yeast population growth and it was measured in 4 different environments: Lactate, Lactose, Sorbitol and Xylose.

Genome to image

In [3], Sharma et al. propose a technique called *DeepInsight* (DI) for transforming a set of numerical sequences into a set of image-like matrices. The key step in DI is mapping each feature (in this case SNPs) to the cartesian plane, which is then discretized and framed using Convex Hull. This mapping is done through t-SNE or kPCA.



(a) An illustration of transformation from feature vector to feature matrix.
(b) An illustration of the DeepInsight methodology to transform a feature vector to image pixels, taken from [3].

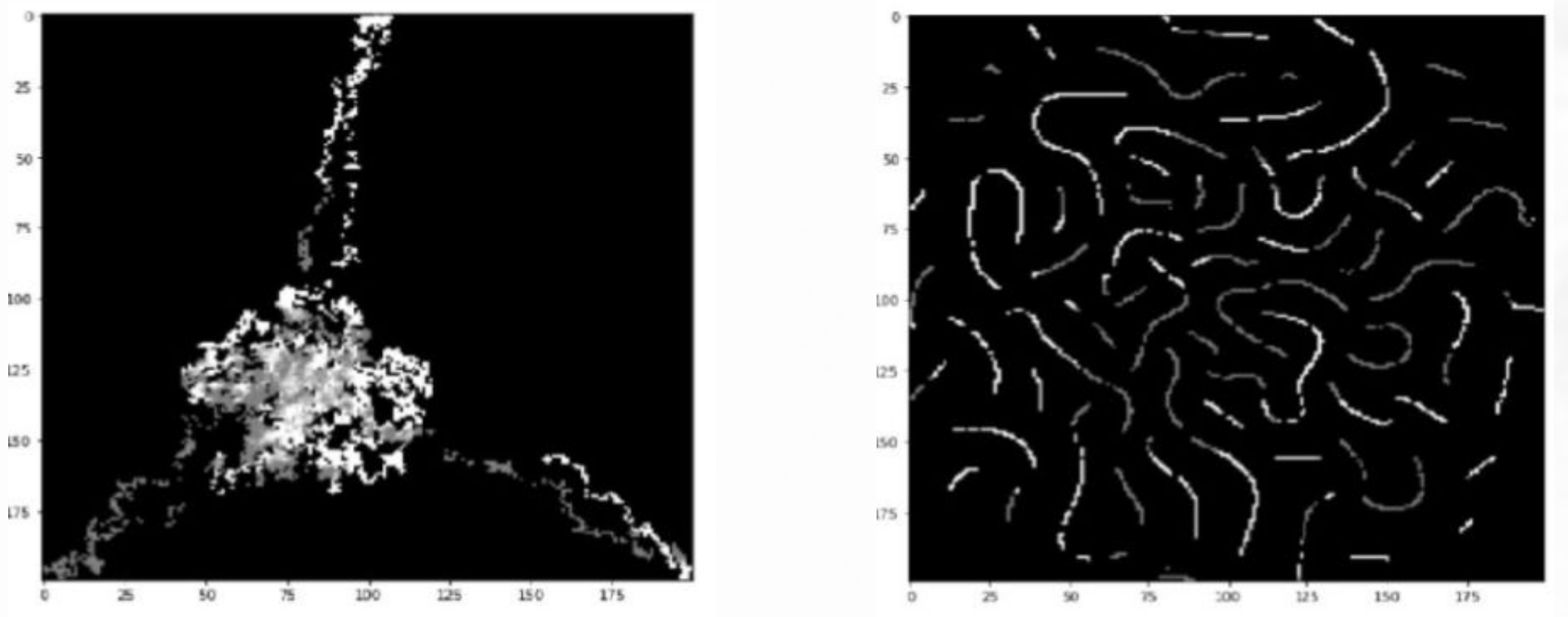


Image representation of yeast using k-PCA (left) and t-SNE (right).

Acknowledgements

This work was partially funded by project ANII FSDA 1-2018-1-154364. The experiments presented in this work were carried out using ClusterUy (site: https://cluster.uy).

References

[1] Abdollahi-Arpanahi, R., Gianola, D., & Peñagaricano, F. (2020). Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genetics Selection Evolution*

[2] Bloom, Joshua S and Ehrenreich (2013). Finding the sources of missing heritability in a yeast cross. Nature Publishing Group.

[3] Sharma, Alok, et al. "DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture." *Scientific reports* 9.1 (2019): 1-7.

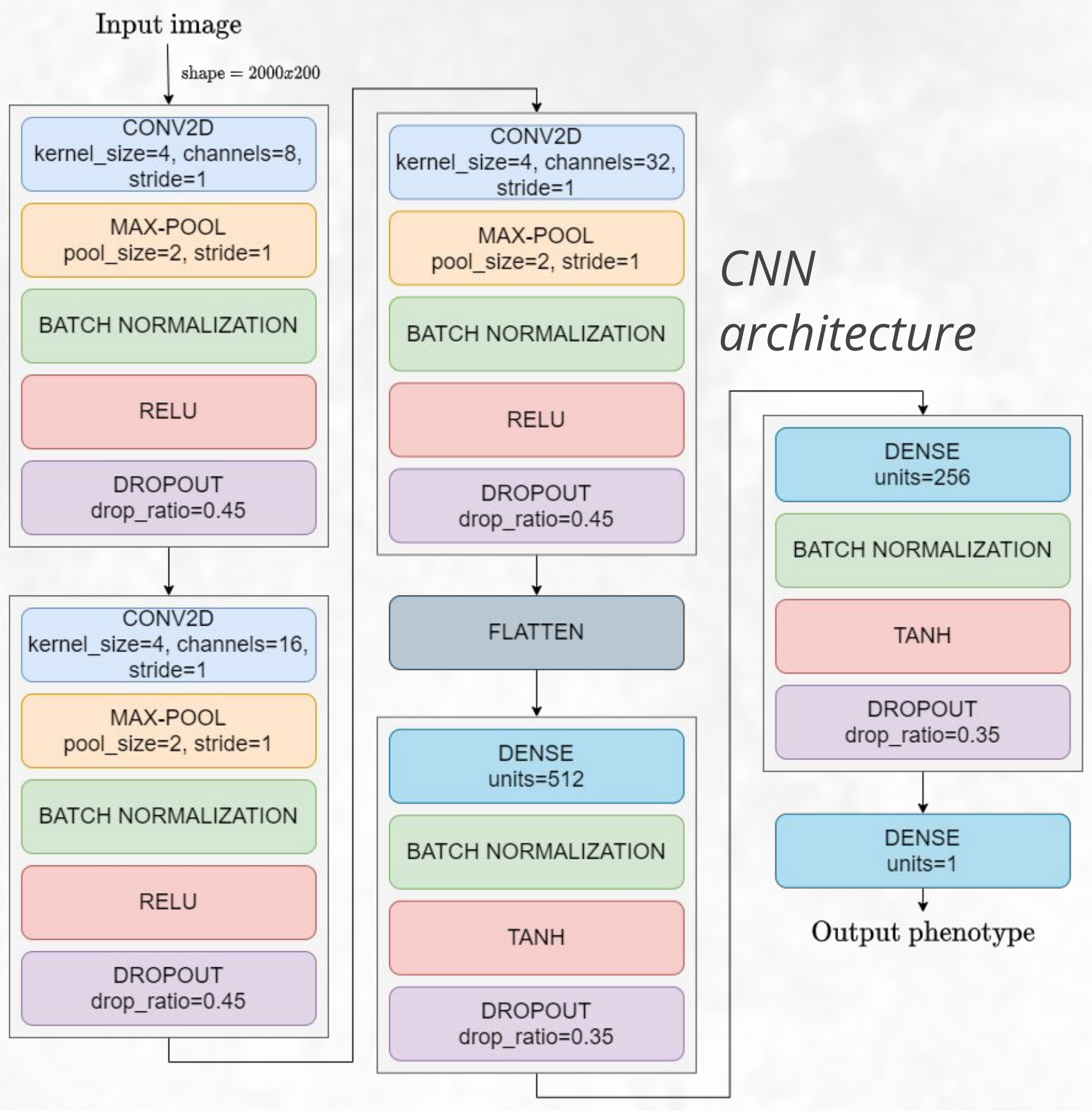
[4] Grinberg, Nastasiya F., Oghenejokpeme I. Orhobor, and Ross D. King. "An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat." *Machine learning* 109.2 (2020): 251-277.

[5] Sapienza, F., Groisman, P., & Jonckheere, M. (2018). Weighted Geodesic Distance Following Fermat's Principle.

Phenotype prediction and random mapping

In order to evaluate the extent to which t-SNE and k-PCA capture the high-dimensional structure of the yeast genome, these mappings are compared to a random mapping.

A 2-D CNN is trained and fine-tuned on the three resulting image datasets. Predictive accuracy is measured in terms of R^2, which is a standard metric in genome-enabled prediction. Results are also compared to those reported by Grinberg in [4].



CNN architecture

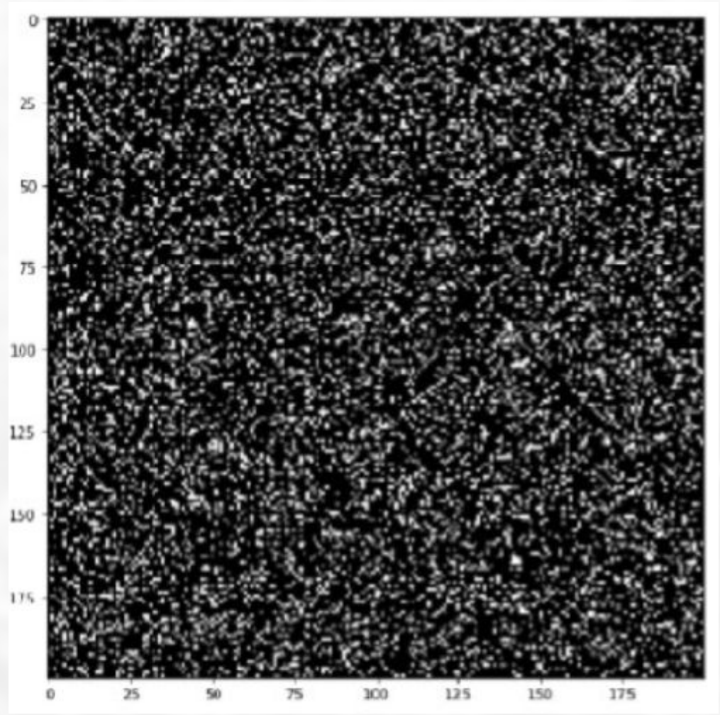
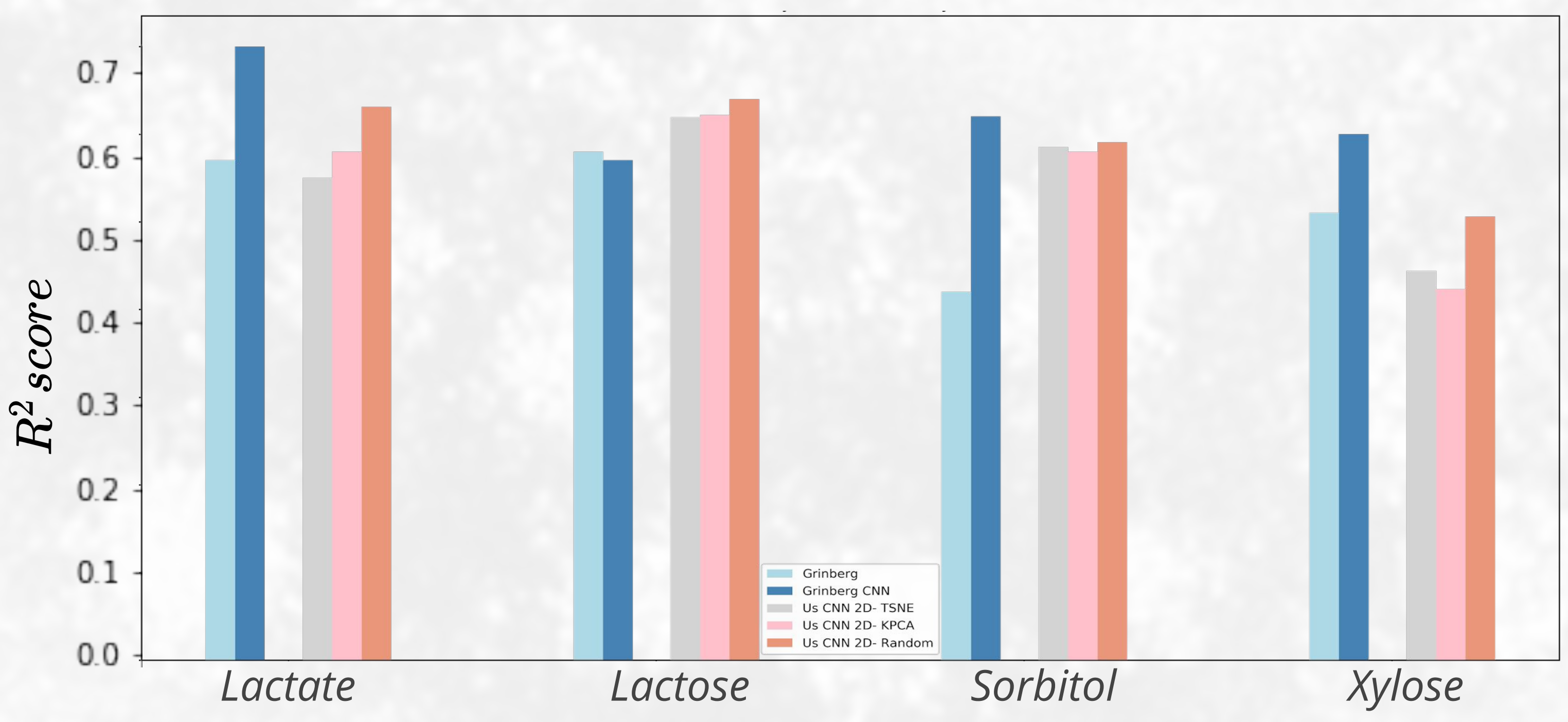


Image representation of yeast using random mapping.

Predictive accuracy by environment and by model.



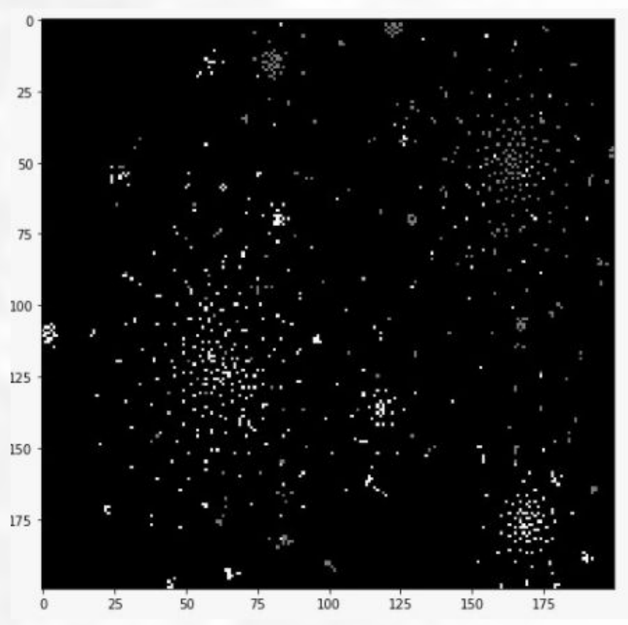
A CNN trained on random mapping of genomes performs better in 3 out of 4 environments. Although comparable to the results in the literature, the DI pipeline does not seem to recover and exploit local structure in the yeast genome.

An Alternative: t-SNE+Fermat distance

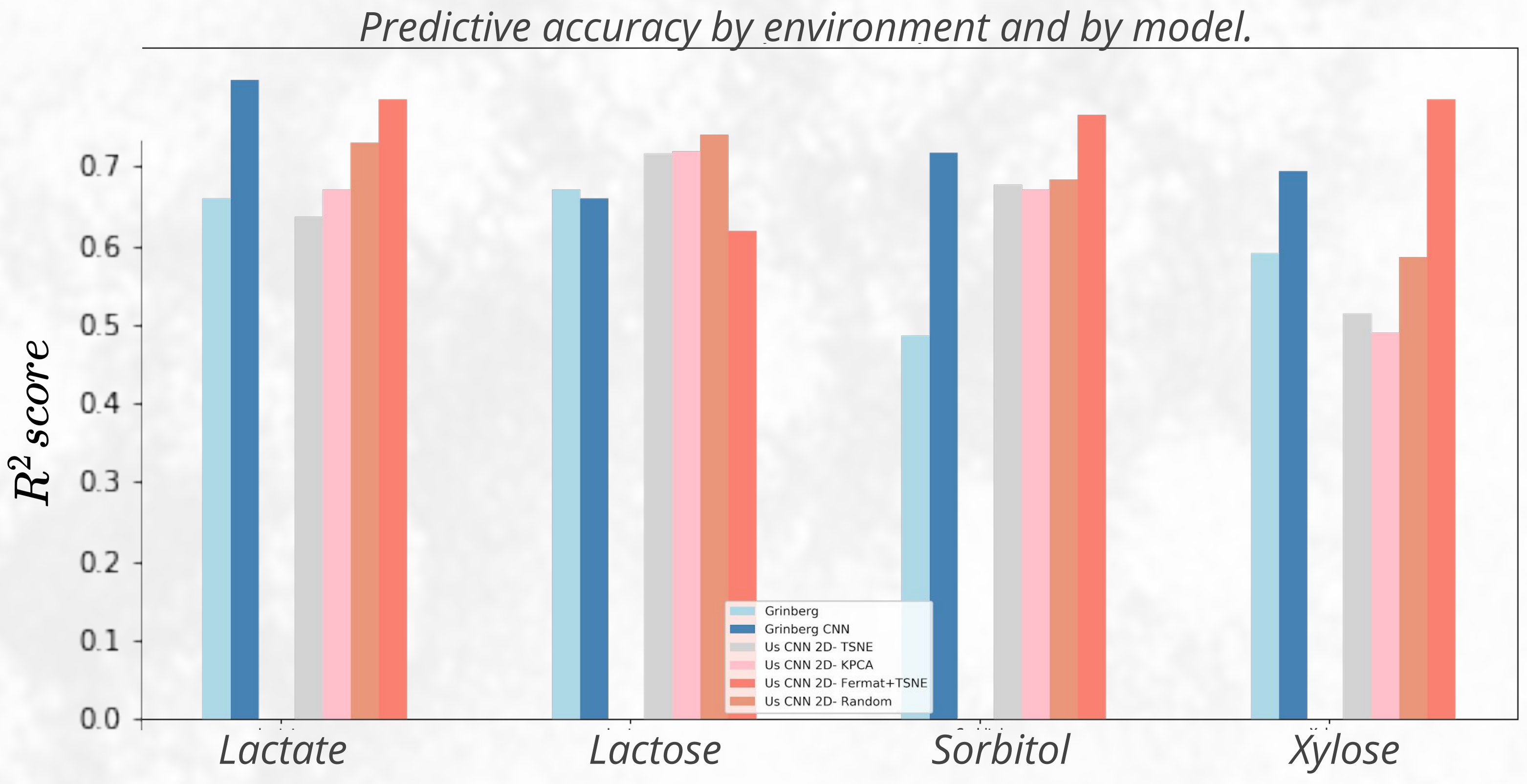
Fermat distance, introduced in [5] for clustering, takes into account both the underlying manifold and the density from which the data is sampled.

$$\mathscr{D}_{\mathbb{X}_n}(x,y) = \min_{\substack{(q_1,\ldots,q_K)\in\mathbb{X}_n^K \\ q_1=x,\, q_K=y,\, K\geq 2}} \sum_{i=1}^{K-1} \ell(q_{i+1},q_i)^{\alpha}.$$

where ℓ is a distance (e.g: euclidean) and $\alpha > 1$.



Predictive accuracy by environment and by model.



Conclusions and perspectives

Experiments on random mappings suggest that the pipeline outlined in [3] to create image-like embeddings of genomes does not effectively capture the high-dimensional structure of the data.

Nonetheless, the CNN trained on images built with t-SNE and Fermat distance slightly outperforms other mappings (in all environments) and results found in the literature (in two out of four environments). Thus, it presents a promising alternative which may potentiate the use of 2D-CNNs on SNP markers and other types of genetic data.