Latent Space Phenotyping: Automatic Image-Based Phenotyping for Treatment Studies

Jordan Ubbens[1], Mikolaj Cieslak[2], Przemyslaw Prusinkiewicz[2], and Ian Stavness[1]

[1]*Department of Computer Science, University of Saskatchewan*
[2]*Department of Computer Science, University of Calgary*

Abstract

Association mapping studies have enabled researchers to identify candidate loci for many important environmental resistance factors, including agronomically relevant resistance traits in plants. However, traditional genome-by-environment studies such as these require a phenotyping pipeline which is capable of accurately and consistently measuring stress responses, typically in an automated high-throughput context using image processing. In this work, we present Latent Space Phenotyping (LSP), a novel phenotyping method which is able to automatically detect and quantify response to treatment directly from images. Using two synthetically generated image datasets, we first show that LSP is able to successfully recover the simulated QTL in both simple and complex synthetic imagery. We then demonstrate an example application of an interspecific cross of the model C_4 grass *Setaria*. We propose LSP as an alternative to traditional image analysis methods for phenotyping, enabling association mapping studies without the need for engineering complex image processing pipelines.

1 Introduction

The combined use of association mapping techniques, such as genome-wide association studies (GWAS), and high-throughput phenotyping has yielded many candidate loci for agronomically important quantitative traits in plants [4]. For food crops, genome-wide analysis of susceptibility or resistance to abiotic stress factors such as drought or nitrogen deficiency [8, 24], salinity [5], or other factors leads to the discovery of genetic differences underlying these agronomically important characteristics. These genotype-by-environment ($G \times E$) studies are capable of identifying resistance alleles which could result in a tolerance to a wider variety of environmental conditions if, for example, introgressed into agricultural lines.

Part of the recent success of such studies, especially in plants, has been due to high-throughput imaging technologies, giving researchers access to large, high-resolution image datasets for phenotypic analysis. Using these datasets in conjunction with tools to measure phenotypes from images, traits of interest can be automatically quantified to compile large collections of phenotype data for use in association mapping or breeding [7]. Utilizing high-throughput phenotyping data, susceptibility or resistance to a treatment is quantified by identifying and extracting macroscopic traits known to be correlated with changes caused

by the treatment. For example, measurements of the number of pixels representing vegetation as a proxy for biomass can be used to quantify drought stress, since biomass is known to be affected by drought stress, allowing researchers to find quantitative trait loci (QTL) correlating with drought resistance or water use efficiency [8]. However, image-based phenotyping techniques fall behind genotyping techniques, resulting in challenges for researchers [10]. Although there is a wide selection of software tools available for extracting phenotype information from images, the design and implementation of specific phenotyping pipelines is often required for individual studies due to inconsistencies between datasets. This is true of both traditional image analysis where thresholds and parameters need to be adjusted, as well as more recent machine learning techniques which require the time-consuming manual annotation of training data. In addition, some phenotypes are difficult to measure from images, and ad-hoc solutions tailored to a particular imaging modality or dataset are often required in place of more general ones.

To overcome the many challenges associated with image-based phenotyping, we present Latent Space Phenotyping (LSP), a novel automatic image analysis technique for quantifying response to a treatment. The key characteristic of LSP in comparison to existing image-based phenotyping methods is that the phenotype estimated using an image analysis pipeline is replaced with an abstract learned concept of the response to treatment, inferred automatically from the image data using deep learning. In this way, any visually consistent response can be detected and differentiated, whether that response is a difference in size, shape, colour, or complex morphological changes such as wilting. The proposed technique has two fundamental advantages when compared to traditional methods. LSP removes the burden of image-based phenotyping from researchers, essentially eliminating the phenotyping bottleneck for treatment studies. This is a substantial advantage, as the challenges presented by phenotyping significantly impede progress on important biological and agronomic problems. Further, by abstracting the visual response to the treatment, LSP is able to detect and quantify complex morphological changes and combined changes of multiple phenotypes which would not only be extremely difficult to quantify using an image processing pipeline, but may not even be apparent to a researcher as correlating with the treatment. LSP leverages the considerable discriminative power of deep learning techniques to obviate the practice of selecting an image feature to use as a proxy for measuring differences in the response to treatment.

LSP receives its name from a family of techniques broadly known as latent variable models. These models include many popular techniques such as Generative Adversarial Networks (GANs) [11] as well as Variational AutoEncoders (VAEs) [14]. These techniques map input data, such as images, to a lower-dimensional latent representation. Information is lost during the learning of this representation, compressing the samples, which can then be converted back into an approximate reconstruction of the original input. While generative models such as GANs and VAEs sample from this latent space in order to produce new samples from the input distribution, LSP instead maps paths through the latent space to quantify the temporal progression of the effect of the treatment. This work is related to previously described methods which are capable of automatically quantifying differences in morphology between individuals, notably the persistent-homology (PH) method [18]. While PH is focused on automatic shape description, LSP instead learns temporal models of stress which can be dependent on, or independent from, shape.

In this paper, we describe the proposed method in detail and provide results on both synthetically generated and real-world datasets. Finally, we address potential limitations and avenues for future work in latent space techniques in the biological sciences.

2 Methods

LSP consists of a three-stage process. Individual images are embedded as vectors in an n-dimensional *latent space*. Then, a decoder performs the reverse process and projects these points in latent space back into the image space. A geodesic distance in the latent space is calculated by mapping a path through this decoder function. The output for an individual corresponds to the distance travelled through the latent space from the initial to the final timepoints. These outputs can be used as trait values with any existing genome-wide association software tools, or interpreted as objective response ratings to inform breeding decisions. A diagram showing the complete process is shown in Figure S1. A complete implementation of LSP, called `LSP-Lab`, is provided at `https://github.com/p2irc/lsplab`.

2.1 Data

Performing an LSP analysis requires an image dataset, comprised of images taken at an arbitrary number (U) of time points during cultivation for each individual in each of the treatment and control conditions. The initial timepoint should ideally be zero days after stress (DAS), in order to establish this as the baseline for determination of the effect of the treatment. Controlled imaging (using imaging booths, stages, or incubators) is recommended in order to maintain consistency in image characteristics such as distance from the camera and the position of the specimen in the frame. However, the method is robust to noise in the images (such as variations in lighting) as long as the noise is consistent between both treatment and control samples.

2.2 Learning an Embedding

To learn the visual features correlating with treatment, LSP utilizes a learned projection of images from the population into an n-dimensional *latent space*, a process known as *embedding*. The embedding is shaped by a supervised learning task, which trains a convolutional neural network (CNN) to extract visual features relevant to the discrimination of treatment and control samples. Performing the embedding allows the method to learn the latent structure of the response, and gives the method the ability to overlook any morphological or temporal characteristics that may be different between accessions, but do not correspond to response to the treatment.

The process of learning the embedding requires only treatment/control labels. The input to the training process is a sequence of images taken for each individual in the treatment and control conditions. The dataset is evenly divided by individual into five folds and training is conducted on each fold in sequence. For each fold, that fold is held out as the *embedding set* and the four remaining folds function as the *training set*. Images in the training set are standardized by subtracting the mean pixel value and dividing by the standard deviation, and then used as input to a CNN. For each time point image, the activations of the last fully connected layer in this CNN are used as input to a Long-Short Term Memory (LSTM) network (Figure 2.1).

We describe both CNNs and LSTMs briefly here, but refer to the literature for more detailed summaries of deep learning in general and these network variants in particular [16, 29, 31]. A CNN can be used to learn local feature extractors from image data. The capability of a CNN to learn a complex representation of the data in this way allows the technique to perform well in many complicated image analysis tasks, such as image classification,

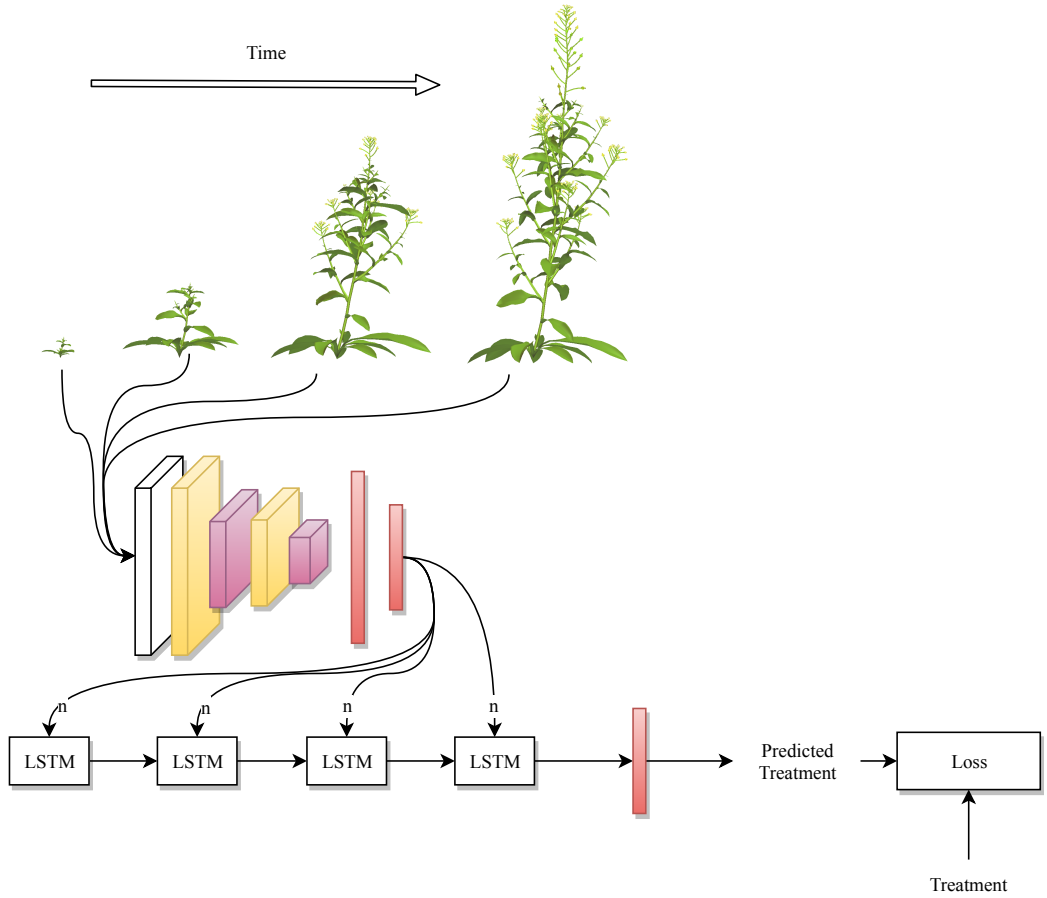

Figure 2.1: The deep network used in learning an embedding. A CNN takes a sequence of images at various timepoints and feeds outputs to an LSTM, which in turn is used to predict the treatment. The LSTM is removed and the CNN is retained to embed new samples.

object detection, semantic and instance segmentation, and many other application areas [16]. CNNs have been used extensively in the recent literature on image-based plant phenotyping, showing promise in several areas, including disease detection and organ counting [13, 21, 22, 25, 31]. For the process of learning an embedding, we implement a simple six-layer convolutional neural network as described in Table S1. Larger architectures were tested and found to show no difference in the experiments discussed here. The only important consideration is the number of units in the last fully connected layer, which will determine the dimensionality of the latent space (n). It was empirically observed that larger values of n facilitated convergence in the decoder step (Section 2.3). However, a smaller value of n will reduce overfitting, which can be a problem when performing the embedding learning step with small datasets. The value $n = 16$ was chosen as the default value and used for all experiments presented here.

Recurrent neural networks (RNNs) are an extension to neural networks which allows

for the use of sequential data. RNNs are a popular tool for time series, video, and natural language problems, for which sequence is an important factor. Briefly, RNNs maintain an internal state which is updated through the sequence, allowing them to incorporate information about the past into the current time point. LSTMs are an extension to RNNs which incorporate a more complicated internal state which is capable of selectively retaining information about the past. LSTMs have also appeared in the plant phenotyping literature, demonstrating that they are able to successfully learn a model of temporal growth dynamics in an accession classification task [29]. LSTMs have also been used as a model of spatial attention in leaf segmentation [27].

The final time point of the LSTM feeds into a fully connected layer, the output of which uses the treatment/control labels of the training images as classification targets, using a standard sigmoid cross-entropy loss for training. After the system has finished training for each fold, the images of the embedding set are then projected into latent space. This embedding is given by the activations of the second fully connected layer in the CNN. In this way, each of the images in each of the treatment and control sequences can be encoded as n-dimensional vectors. The latent space for the first fold is considered to be the *canonical space* and the points embedded into it from the first fold as described above are referred to as the *canonical samples*. For each subsequent fold, the process is repeated as above to determine embeddings of the points in the test fold into a new space. Afterwards, these points are projected into the canonical space via a linear transformation. This transformation is learned by fitting a linear regression model between the canonical samples in the n-dimensional canonical space, and the same images projected into the present n-dimensional space. The goodness of fit is assessed by measuring the coefficient of determination between these two sets of vectors. Finally, this linear transformation is used to project the embedding set into the canonical space. Although this linear transformation can lose the non-linear details of the learned space, it is used instead of a learned non-linear mapping in order to prevent overfitting.

The final result of the embedding step is all images k projected into the same n-dimensional space, which can be visualized as an approximated *ordination plot* using a dimensionality reduction technique (here we use PCA). To obtain the time-relative embedding k_m for the ordination plot, we subtract the vector for the first time point in the image sequence from the vector for the final time point. We then subtract the (now time-normalized) vector for the control sample from the vector for the treated sample to arrive at the final vector for the accession k'_m for the ordination plot.

$$k'_m = k_{m_t} - k_{m_c} = (k_{m_{c_U}} - k_{m_{c_0}}) - (k_{m_{t_U}} - k_{m_{t_0}}) \tag{1}$$

The ordination plot is used only for visualization purposes, since distances on the ordination plot do not correspond to semantic distance between samples, an issue discussed in Section 2.4. Creating an ordination plot with exact distances between accessions would require calculating m^2 pair-wise paths in the latent space, which is intractable. However, generating the ordination plot using euclidean distances between embeddings often illustrates stratification of samples in the latent space, albeit with approximate accuracy.

2.3 Training a Decoder

The second phase of the method involves training a decoder which performs the same function as the embedding process described in Section 2.2, but in reverse. The purpose of the decoder is to define the mapping from latent vectors to image space, discovering the latent

structure in the image space, and allowing us to calculate geodesic paths in the latent space during the subsequent phase (Section 2.4).

The structure of the decoder network consists of a series of convolutional layers followed by learned upsampling layers, which increase the spatial resolution of the input (Table S2). Images from samples in the training set of the final fold are projected by the finalized CNN into the latent space, and then a separate CNN projects these latent space vectors back into the input space (Figure 2.2) where the loss function quantifies the difference between the original image and its reconstruction provided by the decoder in terms of mean squared error (MSE). Note that this process is different than an autoencoder, since the encoder and decoder are not trained end to end.

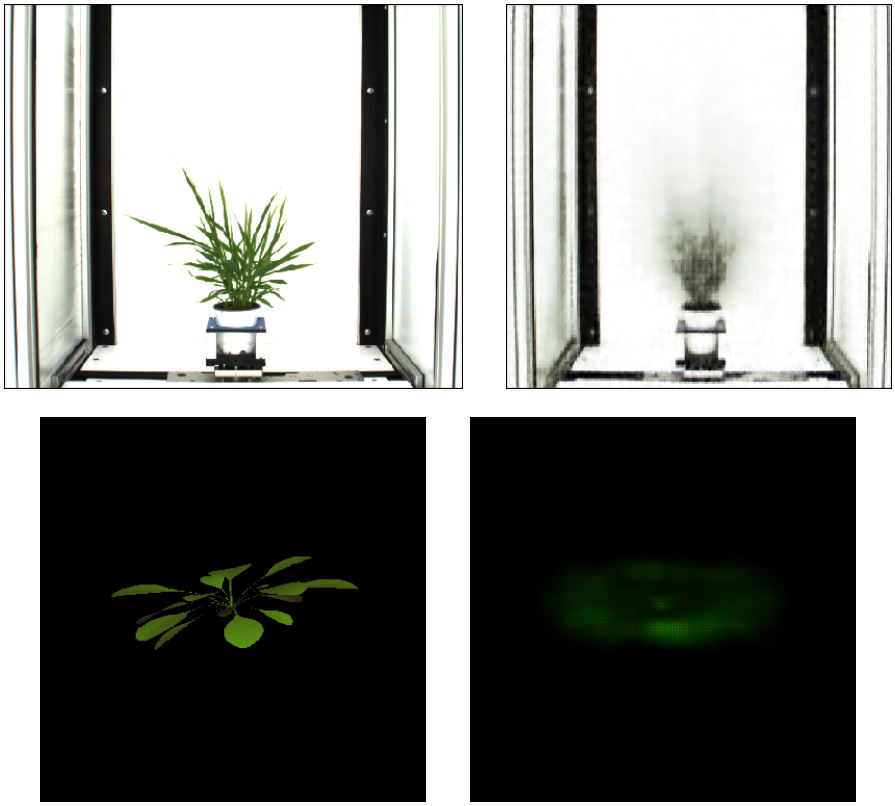

Figure 2.2: Real images from the Setaria and synthetic Arabidopsis datasets (left), and the same images predicted from their latent space encodings by a decoder network (right).

Since the embeddings are derived from the supervised classification task, the only features which are encoded in the latent representation are those which are correlated with stress. For example, in Figure 2.2 (bottom) the induced angle of the synthetic rosette (the plant leans slightly to the left) is not reflected in the decoder's prediction, since plant angle is not encoded in the latent space due to it not being correlated with the simulated response to treatment. The leaf elevation angle, however, does match between the real and the predicted images. The use of MSE creates decoded images which appear blurry - this is an expected result, and helps produce smooth interpolations in the image space when calculating paths

as described in Section 2.4

2.4 Calculating Geodesics in Latent Space

In the final part of the process, we seek to describe the distance traveled in the latent space over time for each of the treated and control sequences, with respect to their embeddings. In particular, we are interested in characterizing the *semantic distance* between images at the initial and final timepoints - that is, the distance between these images in terms of stress propagation. This characterization of semantic distance needs to be considered in terms of *geodesic distance* in the latent space, rather than euclidean distance in the latent space or the in the image space. Figure S2 illustrates the difference between the euclidean distance and the geodesic distance in the latent space for a toy example.

In Section 2.3 we defined a decoder (or a *generator function*) $g : \mathcal{Z} \mapsto \mathcal{X}$ where $\mathcal{Z}$ is the latent space and $\mathcal{X}$ is the input space of the CNN (Figure 2.2). Since g is trivially non-linear, this implies that Z is not a euclidean space, but a (pseudo-) Riemann manifold [2]. The geodesic distance of a path γ on a latent space Riemann manifold in the continuous case is given by

$$Length(\gamma_t) = \int_0^1 ||J_{\gamma_t} \frac{d\gamma_t}{dt}||\ dt, \quad J_{\gamma_t} = \left.\frac{\delta g}{\delta z}\right|_{z=\gamma_t}, \tag{2}$$

mapping the path through the generator function g via the Jacobian J_{γ_t} [2]. Minimizing this path in the discrete case can be accomplished by optimizing the squared pair-wise distance between a series of intermediate path vertices, minimizing

$$\underset{z}{\text{argmin}} \sum_{i=1}^{j} m(g(z_i),\, g(z_{i-1}))^2 \tag{3}$$

where m is a difference function [15]. Performing this optimization on the latent spaces generated by LSP is possible using a standard choice for m such as L_2 distance in the pixel space, since distance in pixels is a sufficiently principled choice for our application.

Using the embeddings of the images for the initial and final timepoints provides the start and end points for a path through the latent space. The embeddings of the intermediate timepoints are also computed, and these are used as stationary vertices on the path. Since more vertices means a more accurate discrete approximation of the geodesic path, we interpolate additional intermediate vertices between the stationary vertices. These vertices are calculated by optimizing Equation 3. For all experiments we use as close to, but not more than, 30 vertices for the path, with an equal number of intermediate vertices between each pair of stationary vertices. Instead of performing progressive subdivision as in [15], we start from a linear interpolation between stationary vertices. This allows us to perform the optimization all at once, instead of dividing the task into multiple successive optimizations which is potentially more expensive. Calculating the total path length as in the sum in Equation 3 describes the individual in a single unitless scalar value, indicating the difference in semantic distance travelled over the course of the trial.

3 Results

In order to validate the proposed method, we describe three experiments where LSP is used. The first two experiments demonstrate results on synthetic datasets. Synthetic images of

rosettes [30] and roots [20] have been used previously to train neural networks for phenotyping tasks. Here we use synthetically generated image data as it allows us to introduce specific variance in the imagery, and then recover that variance on the other end by running a GWAS on the simulated population. First, we demonstrate a dataset consisting of simple synthetic imagery of circles with varying growth rates and a single simulated QTL corresponding to resistance to a hypothetical treatment, allowing us to demonstrate how the method operates under ideal circumstances with consistent effects and visually simple imagery. Next, we use synthetic imagery generated using a model of the *Arabidopsis thaliana* rosette. Using a dynamic architectural plant model allows us to simulate a more complex response to treatment, in this case, a difference in leaf elevation angle. Finally, we demonstrate an application to plants using a recombinant inbred line of *Setaria* treated with drought stress [8].

The latent distance measurements provided by LSP can be used in conjunction with any standard association mapping software package. For the results with synthetic data, we use FaST-LMM [19] to perform this analysis and generate the Manhattan plot. For the Setaria RIL experiment, we use the biparental linkage mapping pipeline provided by the authors of the dataset.

3.1 Synthetic circles

The first experiment we present is intended to show how the LSP method performs under optimal conditions. For this purpose we use a simple model of the *Arabidopsis thaliana* rosette which depicts individuals as white circles on black backgrounds, with a hypothetical treatment causing a decreased growth rate of the circle over time in this simple model.

To generate the synthetic circles dataset, we begin from a real *Arabidopsis thaliana* genotype database known as the *Arabidopsis thaliana polymorphism database* [12]. This dataset includes 214,051 SNPs for 1,307 different accessions of Arabidopsis. A single causal SNP was chosen at random, and we let that SNP represent a polymorphism which confers resistance to a hypothetical treatment that affects the growth rate of the circle. For each of the control and treatment conditions, a sequence of six time points is generated, with images representing a circle growing from an initial diameter (sampled from a normal distribution) to a final diameter. The growth rate of the diameter is drawn from a normal distribution, parameterized according to condition (treatment or control) and resistance (major or minor allele at the simulated causal locus).

Performing an LSP analysis of this dataset allows us to forgo phenotyping and use the synthetic image data as input directly. The ordination plot representing the learned embedding of the image data as well as the manhattan plot are shown in Figure 3.1. LSP is able to recover the simulated causal locus with no false positives in this simple case. The causal locus explains 66.8% of the variance in latent space.

The ordination plot (Figure 3.1, left) shows that samples from this simple model stratify in the latent space into discrete clusters based on the presence of the simulated minor allele. Relating LSP to the established method of using image processing to extract the growth rate phenotype, we examine the correlation between pair-wise distances in the latent space and differences in measured phenotype between the same accessions. There is significant correlation between calculated geodesic distances in the latent space and the relative growth in the number of white pixels in the synthetic circles dataset ($R = 0.93$, $p < 0.01$).

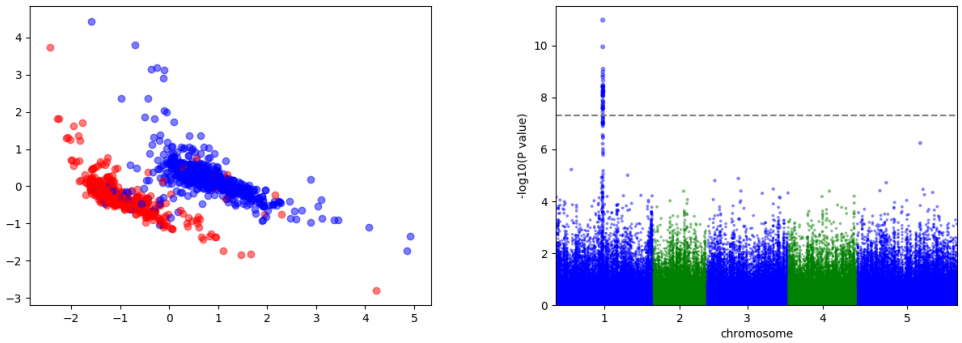

Figure 3.1: Result of LSP on the synthetic circles dataset. The ordination plot (left) describes each accession in latent space. Accessions with the presence of the minor allele at the simulated causal locus are colored red. Manhattan plot (right).

3.2 *Arabidopsis thaliana* model

In this experiment, we demonstrate the extension of LSP from the simple synthetic imagery used in Section 3.1 to more complex synthetic imagery and a more visually complex response to the simulated treatment. We also verify the claim that LSP is able to learn arbitrary and complex responses to the treatment beyond simple changes in growth rate.

For this purpose, we used an existing L-system-based model of an Arabidopsis thaliana rosette [30]. It is based on observations and measurements of the development of real Arabidopsis rosettes [23]. The model was run in the lpfg simulation program [1], which simulated the development of the plant over time, and rendered the resulting images. We selected seven of these images corresponding to different time points of the simulation for the LSP analysis. The same Arabidopsis genotype database was used as with the previous synthetic dataset, as well as the same simulated causal SNP which confers resistance to the simulated treatment.

In this case, the elevation angle of the plant's leaves is sampled from a normal distribution which is parameterized according to whether the sample is untreated, treated-and-resistant, or treated-and-not-resistant. Figure S3 shows the effect of the simulated treatment where the angle of the leaves on the treated plant is increased relative to the untreated sample. Other parameters in the model, such as growth rate, are normally sampled for each accession. It should be noted that, unlike in the synthetic circles dataset, the growth rate of the simulated Arabidopsis plant is completely uncorrelated from the treatment, as are multiple other model parameters. This means that, although the effect of the treatment is still visually apparent, the CNN must learn a more complex visual concept and cannot rely on measuring plant pixels to discriminate between treated and untreated samples. Since the leaf elevation is modulated as a function of plant maturity, the effect of the treatment is not visible in plants with a low growth rate, adding considerable noise and further increasing the complexity of the task. Also note that performing phenotyping on this image dataset would be challenging, since estimating leaf angle from images is a nontrivial image processing task, especially in the absence of depth information [3, 6]. Therefore we omit results using a traditional phenotyping pipeline.

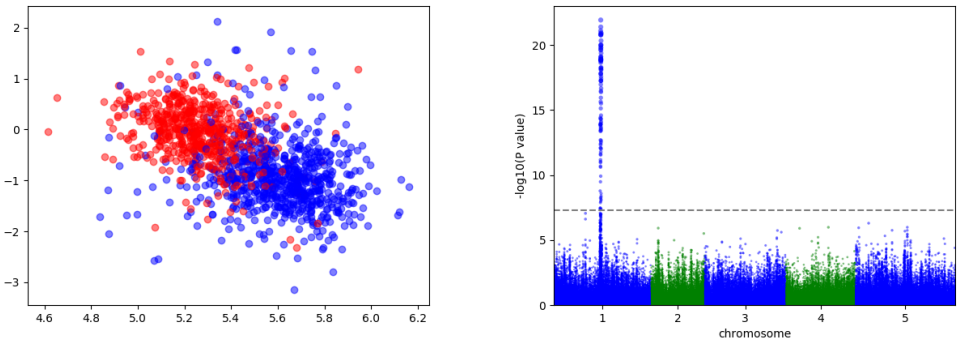

Figure 3.2: Results for LSP on the synthetic Arabidopsis model: ordination plot (left) and manhattan plot (right). Accessions with the presence of the minor allele at the simulated causal locus are colored red in the ordination plot.

The result of LSP on the synthetic Arabidopsis dataset shown in Figure 3.2 demonstrates that the phenotypic noise and complex imagery create an ordination plot with significantly less stratification than with the simple phenotype in the first experiment with synthetic imagery (Figure 3.1, left). Despite this added complexity, the method is able to successfully determine the simulated causal locus on chromosome one with no false positives. The causal SNP explains 57.5% of the variance in latent space.

3.3 Setaria RIL (*S. italica* x *S. viridis*)

In our final experiment with natural imagery, we use a published dataset of a recombinant inbred line (RIL) population of the C_4 model grass *Setaria* [9, 8]. The dataset includes drought and well-watered conditions and has been used to detect QTL relevant to water use efficiency and drought tolerance [26, 8].

The dataset was used as provided by the authors of the original study [9] with a few modifications. The image data was downsampled to 411 by 490 pixels, to allow for a more practical input size for the CNN. The February 1st time point was selected as the initial time point, since many of the earlier time points were taken before emergence. In total, 1,138 individuals and six time points were used. The SNP calls were used as provided by the authors, resulting in a collection of 1,595 SNPs for this experiment. The latent distance values generated by the proposed method were used as trait values for the multiple QTL biparental linkage mapping pipeline provided by the authors of the dataset, in order to replicate the methodology used in the published results.

A histogram of latent distance values for individuals in each of the water-limited and well-watered conditions are shown in Figure 3.3 (left). These generated latent distance traits identified one QTL associated with drought resistance in the Setaria RIL population (7@51.9, Figure 3.3, right). This detected locus is one of the QTL identified by Feldman et al. as correlating to plant size, water use, and model fit in a water use model. Although we were only successful in replicating one of the genotype by environment QTL in the published study, these reported QTL correspond to a water use model incorporating evapotranspiration, not a single trait derived directly from the images such as vegetation

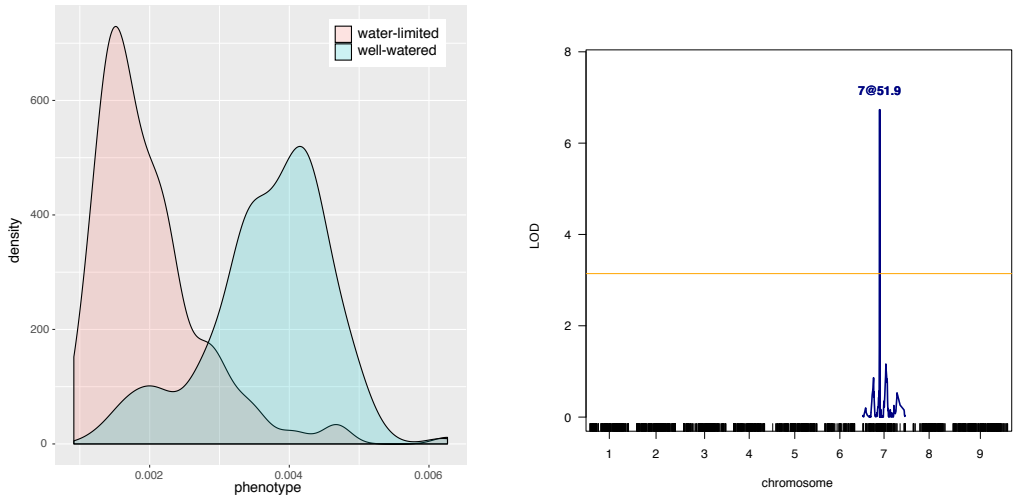

Figure 3.3: Histogram of generated latent distance values for plants in the water-limtied and well-watered conditions for the Setaria RIL dataset (left) and LOD plot (right), showing QTL for comparison between water-limited and well-watered conditions.

area. Additionally, the referenced study only agreed with a previous study finding drought stress QTL in this interspecific cross on one loci (albeit with a different set of markers) [26].

4 Discussion

The method as described has some limitations, including increased computational requirements compared to the majority of image-based phenotyping techniques. Since the method involves multiple deep neural networks, the use of GPUs is advisable to perform these optimizations in a tractable amount of time. The experiments presented here were performed on two GPUs and the time required per experiment ranged from two to eight hours depending on the number of accessions and the number of timepoints in the dataset. In addition, since the method is non-deterministic because of randomized initial weights and random mini-batching (as with all deep learning methods), repeating the same experiment may output different results.

Beyond computational requirements, another limitation of the method is a substantial difference in interpretability compared to GWAS using standard image-based phenotyping techniques. The traits measured with these standard techniques have a direct and interpretable relationship with the response to the treatment – for example, it has been shown that the number of plant pixels in an image can be used as a proxy for biomass [17]. Therefore, the measured phenotype can be directly interpreted as the biomass of the sample and QTL can be found which correlate with the effect of the treatment on biomass. In the case of LSP, the individual's response to the treatment is abstracted and quantified only relative to other individuals in the dataset. Interpretability techniques such as saliency maps [28] (Figure S7) can help to elucidate relevant regions in the images, but the measurements still lack a direct biological interpretation in the same way as measurements of biomass. There-

fore, candidate loci obtained through LSP must be interpreted differently, and biological explanations must be inferred from the function of the detected loci.

The results presented in Section 3 demonstrate the capability of LSP to automatically form accurate learned concepts of response to treatment from images and recover QTL with a very low false positive rate. As an automated system, the proposed method is exempt from the considerable challenges which arise in developing and deploying image analysis pipelines to first measure phenotype from images. It is also free from assumptions about which visually evident features are caused by the treatment, automatically detecting area, leaf angle, and drought stress in three different experiments. Replicating more candidate loci from existing studies will help continue to validate the technique and encourage further study on latent methods in the biological sciences.

Data Availability

Full datasets and utility scripts needed for reproducing the results and figures presented in Section 3 can be found at `https://figshare.com/s/f710381c04c01e2ba319`. The data for the Setaria RIL experiment are available from the authors [8].

Acknowledgements

We would like to gratefully acknowledge the Baxter group at the Donald Danforth Plant Science Center for the use of the publicly available Setaria RIL dataset. This research was funded by a Canada First Research Excellence Fund grant from the Natural Sciences and Engineering Research Council of Canada.

References

[1] Virtual laboratory. `http://www.algorithmicbotany.org/virtual_laboratory/`. Accessed: 2017-08-01.

[2] Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent Space Oddity: on the Curvature of Deep Generative Models. (2):1–15, 2017.

[3] Bernhard Biskup, Hanno Scharr, Ulrich Schurr, and Uwe Rascher. A stereo imaging system for measuring structural parameters of plant canopies. *Plant, Cell & Environment*, 30(10):1299–1308, 2007.

[4] Benjamin Brachi, Geoffrey P. Morris, and Justin O. Borevitz. Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biology*, 12(10):232, Oct 2011.

[5] Malachy T. Campbell, Avi C. Knecht, Bettina Berger, Chris J. Brien, Dong Wang, and Harkamal Walia. Integrating Image-Based Phenomics and Association Analysis to Dissect the Genetic Architecture of Temporal Salinity Responses in Rice. *Plant Physiology*, 168(4):1476–1489, 2015.

[6] Tino Dornbusch, Séverine Lorrain, Dmitry Kuznetsov, Arnaud Fortier, Robin Liechti, Ioannis Xenarios, and Christian Fankhauser. Measuring the diurnal pattern of leaf hyponasty and growth in arabidopsis–a novel phenotyping approach using laser scanning. *Functional Plant Biology*, 39(11):860–869, 2012.

[7] Noah Fahlgren, Malia A. Gehan, and Ivan Baxter. Lights, camera, action: High-throughput plant phenotyping is ready for a close-up. *Current Opinion in Plant Biology*, 24:93–99, 2015.

[8] Max J. Feldman, Patrick Z. Ellsworth, Noah Fahlgren, Malia A. Gehan, Asaph B. Cousins, and Ivan Baxter. Components of Water Use Efficiency Have Unique Genetic Signatures in the Model C4 Grass Setaria. *Plant Physiology*, 178(2):699–715, 2018.

[9] Max J. Feldman, Rachel E. Paul, Darshi Banan, Jennifer F. Barrett, Jose Sebastian, Muh Ching Yee, Hui Jiang, Alexander E. Lipka, Thomas P. Brutnell, José R. Dinneny, Andrew D.B. Leakey, and Ivan Baxter. *Time dependent genetic analysis links field and controlled environment phenotypes in the model C4 grass Setaria*, volume 13. 2017.

[10] Robert T. Furbank and Mark Tester. Phenomics - technologies to relieve the phenotyping bottleneck. *Trends in Plant Science*, 16(12):635–644, 2011.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[12] Matthew W. Horton, Angela M. Hancock, Yu S. Huang, Christopher Toomajian, Susanna Atwell, Adam Auton, N. Wayan Muliyati, Alexander Platt, F. Gianluca Sperone, Bjarni J. Vilhjálmsson, Magnus Nordborg, Justin O. Borevitz, and Joy Bergelson. Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nature Genetics*, 44(2):212–216, 2012.

[13] Andreas Kamilaris and Francesc X. Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147(February):70–90, 2018.

[14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[15] Samuli Laine. Feature-Based Metrics for Exploring the Latent Space of Generative Models. *ICLR Workshop*, 7:1–4, 2018.

[16] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[17] Dario Leister, Claudio Varotto, Paolo Pesaresi, Alexandra Niwergall, and Francesco Salamini. Large-scale evaluation of plant growth in Arabidopsis thaliana by non-invasive image analysis, 1999.

[18] Mao Li, Margaret H. Frank, Viktoriya Coneva, Washington Mio, Daniel H. Chitwood, and Christopher N. Topp. The persistent homology mathematical framework provides enhanced genotype-to-phenotype associations for plant morphology. *Plant Physiology*, 177(4):1382–1395, 2018.

[19] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M. Kadie, Robert I. Davidson, and David Heckerman. FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835, 2011.

[20] Guillaume Lobet, Iko T. Koevoets, Manuel Noll, Patrick E. Meyer, Pierre Tocquin, Loïc Pagès, and Claire Périlleux. Using a Structural Root System Model to Evaluate and Improve the Accuracy of Root Image Analysis Pipelines. *Frontiers in Plant Science*, 8(April):1–11, 2017.

[21] Hao Lu, Zhiguo Cao, Yang Xiao, Bohan Zhuang, and Chunhua Shen. TasselNet : Counting maize tassels in the wild via local counts regression network. *Plant Methods*, pages 1–14, 2017.

[22] Sharada Prasanna Mohanty, David P. Hughes, and Marcel Salathé. Using Deep Learning for Image-Based Plant Disease Detection. *Frontiers in Plant Science*, 7(September):1–7, 2016.

[23] Lars Mündermann, Yvette Erasmus, Brendan Lane, Enrico Coen, and Przemyslaw Prusinkiewicz. Quantitative modeling of Arabidopsis development. *Plant Physiology*, 139(2):960–968, 2005.

[24] E. H. Neilson, A. M. Edwards, C. K. Blomstedt, B. Berger, B. Lindberg Møller, and R. M. Gleadow. Utilization of a high-throughput shoot imaging system to examine the dynamic phenotypic responses of a c4 cereal crop plant to nitrogen and water deficiency over time. *Journal of Experimental Botany*, 66(7):1817–1832, 2015.

[25] Michael P. Pound, Jonathan A. Atkinson, Alexandra J. Townsend, Michael H. Wilson, Marcus Griffiths, Aaron S. Jackson, Adrian Bulat, Georgios Tzimiropoulos, Darren M. Wells, Erik H. Murchie, Tony P. Pridmore, and Andrew P. French. Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. *GigaScience*, 6(10):1–10, 2017.

[26] Lufeng Qie, Guanqing Jia, Wenying Zhang, James Schnable, Zhonglin Shang, Wei Li, Binhui Liu, Mingzhe Li, Yang Chai, Hui Zhi, and Xianmin Diao. Mapping of Quantitative Trait Locus (QTLs) that contribute to germination and early seedling drought tolerance in the interspecific cross Setaria italica x Setaria viridis. *PLoS ONE*, 9(7):3–10, 2014.

[27] Bernardino Romera-Paredes and Philip Hilaire Sean Torr. Recurrent instance segmentation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, pages 312–329. Springer International Publishing, 2016.

[28] J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015.

[29] Sarah Taghavi Namin, Mohammad Esmaeilzadeh, Mohammad Najafi, Tim B. Brown, and Justin O. Borevitz. Deep Phenotyping: Deep Learning For Temporal Phenotype/Genotype Classification. *bioRxiv*, pages 1–29, 2017.

[30] Jordan R. Ubbens, Mikolaj Cieslak, Przemyslaw Prusinkiewicz, and Ian Stavness. The use of plant models in deep learning: an application to leaf counting in rosette plants. *Plant Methods*, 14(1):6, 2018.

[31] Jordan R. Ubbens and Ian Stavness. Deep Plant Phenomics: A Deep Learning Platform for Complex Plant Phenotyping Tasks. *Frontiers in Plant Science*, 8(July), 2017.

Supplementary Information

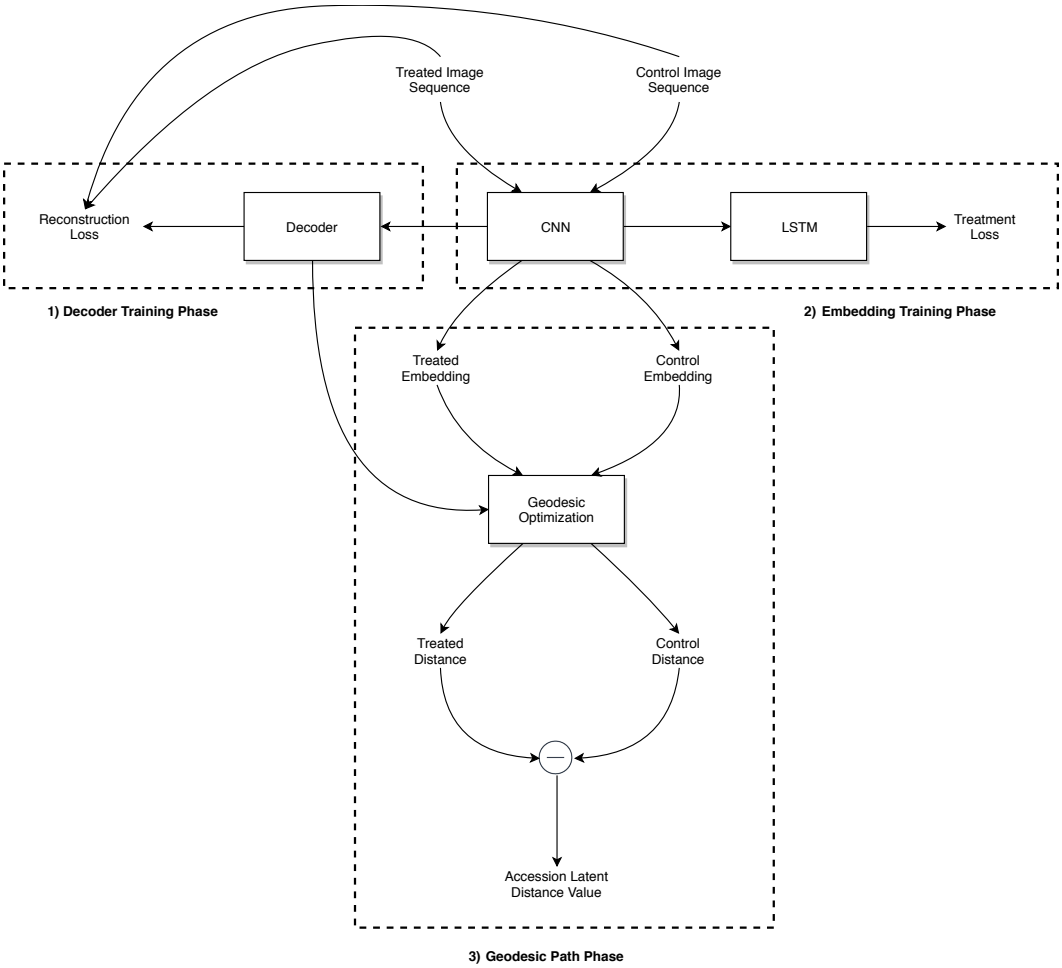

Figure S1: Overview of the proposed technique. The process consists of three phases, which take place in sequence. First, an embedding process projects images into latent space. Second, a decoder is trained to convert these embeddings back into the input space. Lastly, the decoder is used to calculate a geodesic path between the embeddings for the initial and final timepoints.

	Size	Depth	Stride
conv + relu	3×3	1	1
max pool	3×3		3
conv + relu	3×3	16	1
max pool	3×3		3
conv + relu	3×3	32	1
max pool	3×3		3
conv + relu	3×3	32	1
max pool	3×3		2
fully connected + relu	64		
fully connected + relu	16		

Table S1: Architecture details for the convolutional neural network used in the embedding.

	Size	Depth	Stride
conv + relu	3×3	16	1
upsample	3×3		2
conv + relu	3×3	32	1
conv + relu	3×3	32	1
conv + relu	3×3	32	1
upsample	3×3		2
conv + relu	3×3	32	1
conv + relu	3×3	32	1
conv + relu	3×3	32	1
upsample	3×3		2
conv + relu	3×3	64	1
conv + relu	3×3	64	1
conv + relu	3×3	64	1
upsample	3×3		2
conv + relu	3×3	64	1
conv + relu	3×3	64	1
conv + relu	3×3	64	1
upsample	3×3		2
conv + relu	3×3	32	1
conv + relu	3×3	32	1
conv + relu	3×3	16	1
upsample	3×3		2
conv + relu	3×3	16	1
upsample	3×3		2
conv + relu	1×1	16	1

Table S2: Architecture details for the decoder network.

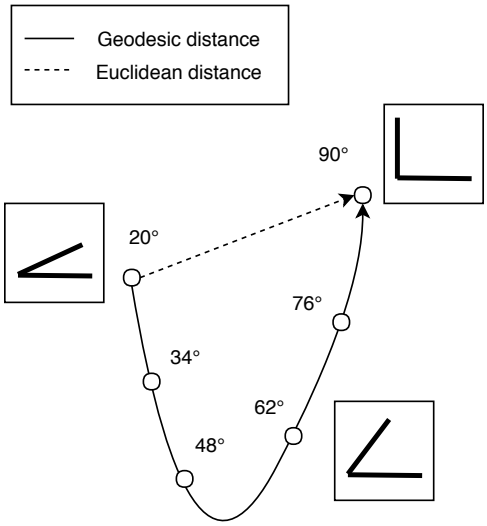

Figure S2: Distance between embeddings of images of lines at different angles in a hypothetical latent space. Here, the *semantic distance* is the difference in the interior angle. The euclidean distance between images in the image space is constant, and the euclidean distance (dotted arrow) between their encodings in the latent space is evidently not representative of the semantic distance. However, the geodesic path (solid arrow) between images represents the semantic distance well.

Figure S3: Untreated (left) and treated non-resistant (right) synthetic Arabidopsis plants at the final timepoint, showing differences in leaf elevation angle.

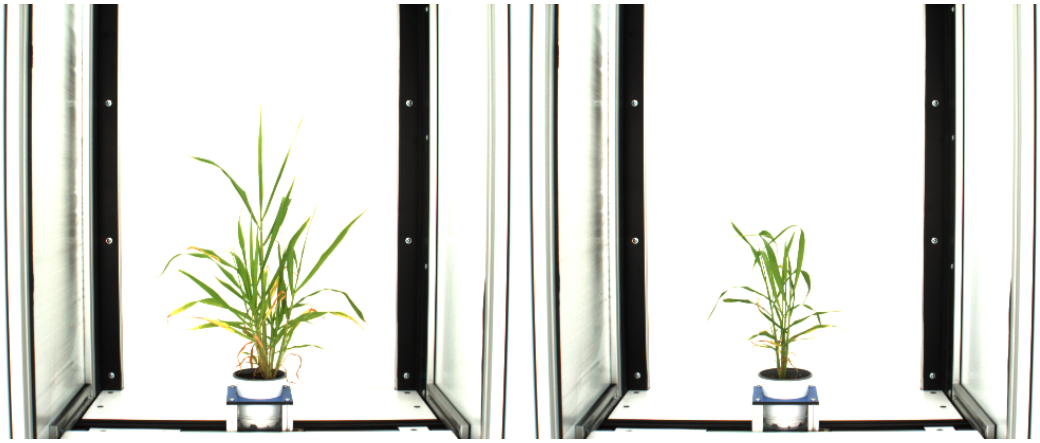

Figure S4: Well-watered (left) and water-limited (right) examples of a particular line from the Setaria RIL population [9]

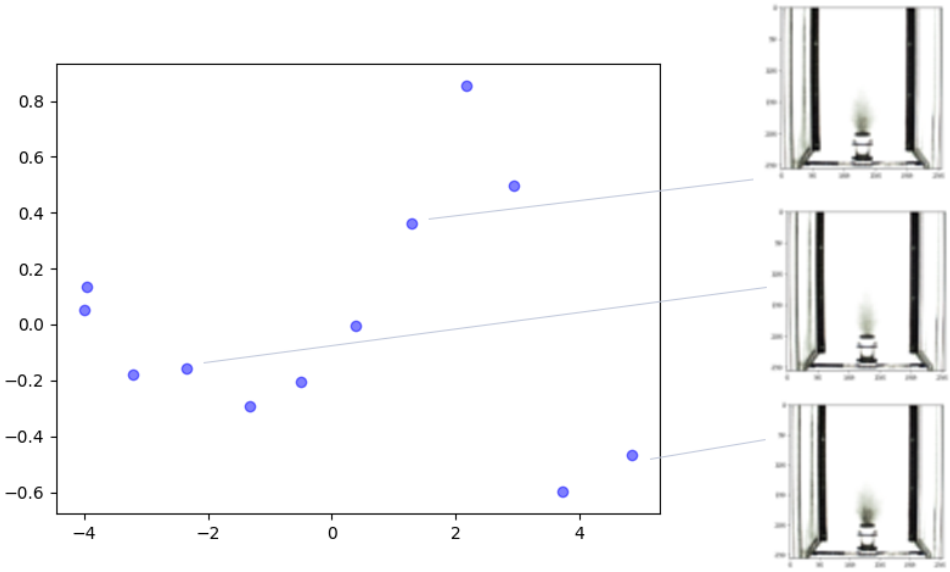

Figure S5: Shortest geodesic path between initial and final timepoints for the Setaria dataset in the 16-dimensional latent space, cast to two dimensions using PCA.

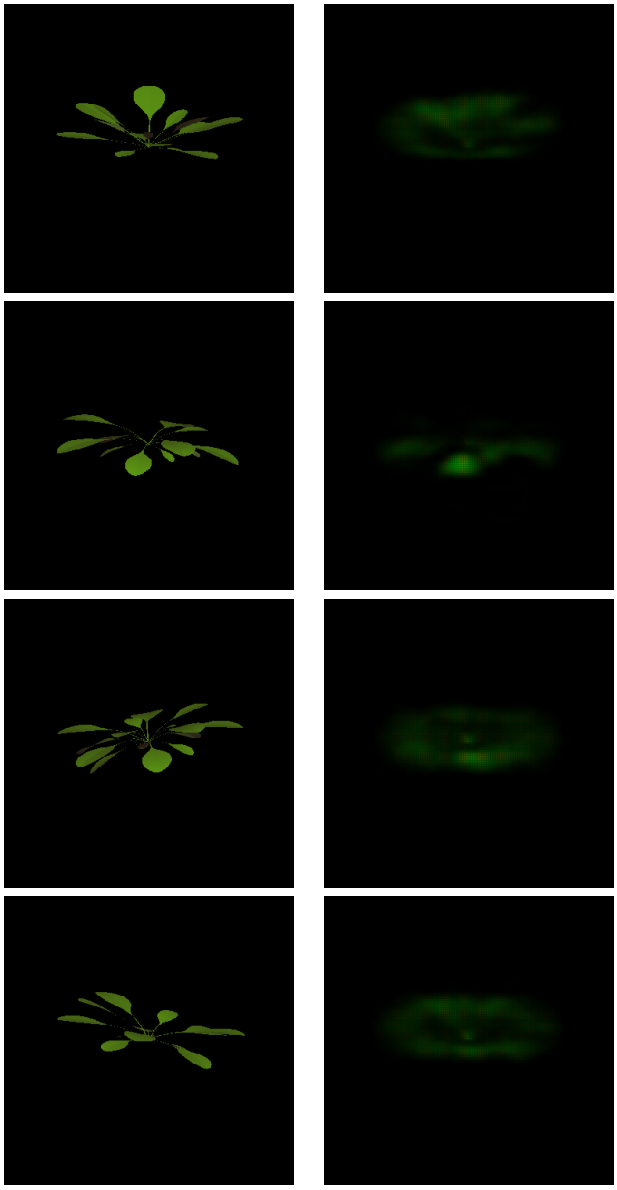

Figure S6: Additional examples of synthetic Arabidopsis rosettes (left) decoded from their latent space vectors (right).

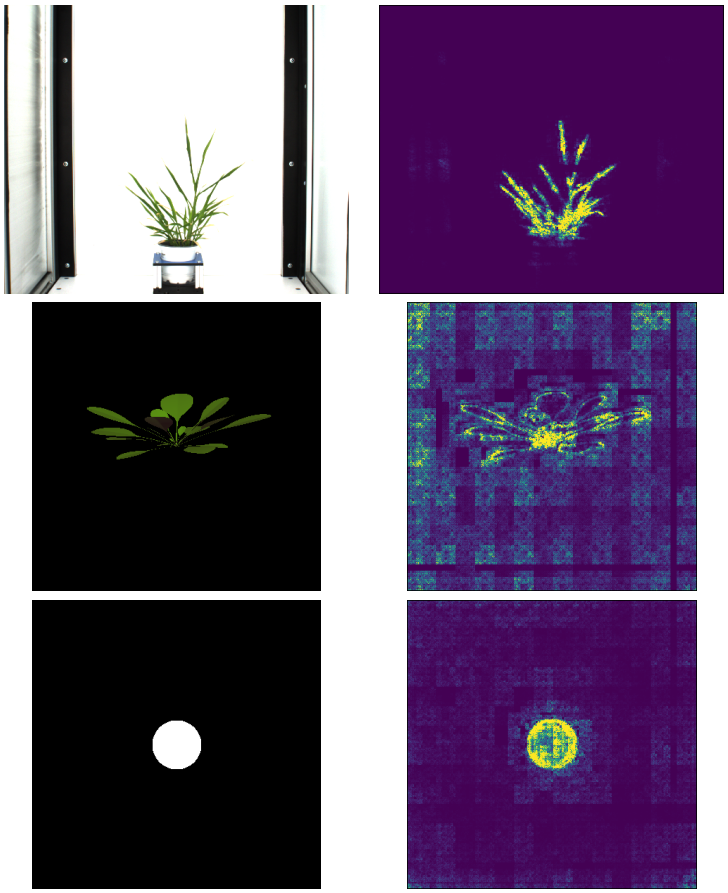

Figure S7: Examples images from the Setaria, synthetic Arabidopsis, and synthetic circles datasets (left) and corresponding saliency maps genetated using guided backpropagation (right). Intensity is higher for the pixels which have high saliency with respect to the L_2 norm of the latent space embedding of the image.