araDEEPopsis: From images to phenotypic traits using deep transfer learning

SHORT TITLE:
Transfer learning in plant phenotyping

AUTHORS:
Patrick Hüther[1,*,‡], Niklas Schandry[1,2,*], Katharina Jandrasits[1], Ilja Bezrukov[3], Claude Becker[1,2,‡]

AFFILIATION:
[1] Gregor Mendel Institute of Molecular Plant Biology (GMI), Austrian Academy of Sciences, Vienna BioCenter (VBC), Dr. Bohr-Gasse 3, 1030 Vienna, Austria
[2] Genetics, Faculty of Biology, Ludwig-Maximilians-University München, 82152 Martinsried, Germany
[3] Department of Molecular Biology, Max Planck Institute of Developmental Biology, 72076 Tübingen, Germany

‡To whom correspondence should be addressed: claude.becker@gmi.oeaw.ac.at; patrick.huether@gmi.oeaw.ac.at

*These authors contributed equally.

Abstract

Linking plant phenotype to genotype, i.e., identifying genetic determinants of phenotypic traits, is the goal of plant breeders and geneticists alike. While the ever-growing genomic resources and the rapid decrease of sequencing costs have greatly facilitated obtaining the critical amount of genomic data, collecting phenotypic data for large numbers of plants remains a bottleneck. Many phenotyping strategies rely on recording images of plants, which makes it necessary to extract phenotypic measurements from these images in a fast and robust way. Common image segmentation tools for plant phenotyping mostly rely on color information, which is error-prone when background or plant color are variable in the experiment or deviate from the underlying expectations. We have developed araDEEPopsis, a versatile, fully open-source pipeline to extract phenotypic measurements from plant images in an unsupervised manner. araDEEPopsis was built around the deep-learning model DeepLabV3+ and re-trained for segmentation of *Arabidopsis thaliana* rosettes. It uses semantic segmentation to classify leaf tissue into up to three categories: healthy, anthocyanin-rich, and senescent. This makes araDEEPopsis particularly powerful at quantitative phenotyping from early to late developmental stages, of mutants with aberrant leaf color and/or phenotype, and of plants growing in stressful conditions, where leaf-color may deviate from green. Using araDEEPopsis on a panel of 210 natural Arabidopsis accessions, we were able to not only accurately segment phenotypically diverse genotypes but also to map known loci related to anthocyanin production and early necrosis using the araDEEPopsis output in genome-wide association analyses. Our pipeline is able to handle images of diverse origins, image quality, and background composition, and could even accurately segment images of a distantly related Brassicaceae. Because it can be deployed on virtually any common operating system and is compatible with several high-performance computing environments, araDEEPopsis can be used independently of bioinformatics expertise and computing resources.

Introduction

The phenotyping bottleneck in plant -omics studies

Over the last decades, molecular techniques have steadily increased in throughput, while they have kept decreasing in cost. A prime example for this development is nucleic acid sequencing, which has followed a trend analogous to Moore's law in computer science. However, phenotyping methods, i.e., methods aimed at determining the physical shape of an organism and at measuring morphological parameters, have not kept up with this pace, which leads to a "phenotyping bottleneck" [1] in the design and execution of scientific studies. Such a phenotyping bottleneck constitutes one of the major challenges also in plant biology, where there are two major underlying causes. The first one is data acquisition, which in most plant phenotyping scenarios is equivalent to acquiring standardized images of plants growing in a controlled environment. Plant phenotyping requires space and dedicated infrastructure that can accommodate the developmental and growth transitions that occur over a plant's life. Moreover, plant development needs to be phenotyped over relatively long periods of time. For example, in the case of the model plant *Arabidopsis thaliana* (referred to as simply Arabidopsis for the remainder of the text), a relatively small and

fast-growing species, a phenotyping experiment typically runs for several weeks or months, depending on the phenotype of interest. In their vegetative phase, i.e., before they produce shoots, flowers, and seeds, Arabidopsis plants grow in relatively small, flat rosettes and can therefore be considered two-dimensional. Because the whole plant area is visible from the top during this phase, top-view phenotyping of the plant to determine size, growth rate, leaf development, etc. is straight-forward. However, while high-throughput image acquisition has become almost trivial, robustly and faithfully extracting meaningful data from these images has not.

Overcoming the data acquisition challenge, for example by means of a dedicated plant phenotyping facility, does not mitigate the second cause of the phenotyping bottleneck, which is data analysis. In the context of image-based phenotyping data, this includes automated image processing and object detection, object segmentation, and extraction of quantitative phenotypic trait data.

Challenges in automated image analysis

The first key step on the way from image to quantitative phenotype data is also the most difficult one: defining the area in the image that depicts the object of interest, in this case the plant. On a small set of images, this segmentation can be done manually by delineating the plant object using e.g. *ImageJ* [2,3] or similar software. On large image datasets with hundreds or thousands of individual images, such as they are typical for experiments from phenotyping platforms, this task needs to be automated, both to speed up the process and to neutralize user bias. Commonly used software for segmenting plant objects from digital images relies on color information. In simple terms, color information of digital images is stored in tables, with information on each pixel stored in a corresponding cell. While grayscale images are stored in a single table, each cell containing the grayscale value of the corresponding pixel, color image information is stored in several such tables. For example, for images using the very common additive color model RGB, three separate tables store information on red (R), green (G) and blue (B) color channel intensities for each pixel. A very simple approach to differentiate plants from background is to assume that, since plants are green, one can assign all pixels that pass a certain threshold in the green channel to the object ‘plant’ and ignore all other pixels. This approach is called binary thresholding and works well if, and only if, the assumption is correct that all plants in the experiment are green and that the remaining area of the image is not. In reality, this assumption is often violated, e.g. because plants produce high quantities of anthocyanins under certain circumstances, by which they might turn a red or purple hue, or because they develop chlorosis or become senescent, in which case they turn yellow or brown and, ultimately, white. Moreover, when growing larger over the course of the experiment, plants from a neighbouring pot often protrude their leaves into the image area of the monitored plant, thus creating additional green areas that should not be taken into account. Even the background color might fluctuate, either because of the constraints of the experimental setup or because of other organisms such as algae growing on the soil surface. Therefore, color-based segmentation often has to be verified by laborious and time-consuming visual inspection of the individual images to identify and correct false segmentation and artifacts. We therefore argue that in plant phenotyping, color-based image segmentation, such as it is currently applied in many common segmentation tools, depends on many image parameters,

some of which are difficult to control in the experiment. The method becomes error-prone and unreliable when the underlying assumptions are violated. We therefore propose that robust plant phenotype measurement should derive from alternative approaches.

Machine learning methods in image analysis

Alternative approaches to object segmentation that employ machine learning methods such as Gaussian mixture models (GMM) [4] or Naive Bayes classifiers [5] to solve the aforementioned problems of color-based segmentation are available. While these offer high flexibility, their implementation and application to new datasets requires substantial programming knowledge.
In recent years, successful alternative approaches to segment particular classes of objects from images have come from the deep learning field. Convolutional Neural Networks (CNNs) have proven invaluable for image classification and segmentation tasks [6–9]. They perform well on data that has a local structure, such as it is inherently found in image data where values of neighbouring pixels tend to be highly correlated and contain recurring structures such as corners and edges.
These models are supervised, which means that when provided with a set of manually annotated images containing the “ground truth” for each pixel, the model will attempt to derive rules for the classification of pixels based on the training dataset. This training process is iterative and usually done over many thousands of iterations, along which the model attempts to map input images to their corresponding ground truth annotations. A loss function is calculated to estimate the error between input and output, and the model subsequently tries to minimize this error via back-propagation of error [10]. Weight matrices throughout the network layers are updated along a gradient, following the chain rule to calculate partial derivatives of the loss function with regard to the layer weights. Due to the non-convex nature of the loss function, error minimization typically reaches only local minima, requiring careful selection of model parameters and a large and diverse set of training data to avoid overfitting. The latter is usually scarce because ground truth data for supervised learning have to be generated by meticulous manual annotation. Depending on the nature of the desired feature to be extracted, such a task is typically labour-intensive and time-consuming because of the large number of data points required to train a well-generalising deep learning model *de novo.*
In a process referred to as transfer learning, trained model architectures that do well at pixel classification can be re-trained on new data that contain new classes while retaining already trained weights in layers that extract low-level features such as shapes and edges. In consequence, to re-train a model on plant rosettes, the dataset on which the model was originally trained does not need to contain plant rosettes. Transfer learning is mainly about updating the last layer of the model, which is much faster and requires considerably less training data than designing and training a completely new model [11]

Here, we introduce araDEEPopsis, a pipeline centered around the deep-learning model DeepLabV3+, retrained for segmentation of Arabidopsis rosettes. Initially designed to simply identify image areas containing plants (rosettes), we further developed the tool to discriminate three different health states of leaves. We show how araDEEPopsis can be applied to reliably segment Arabidopsis rosettes independent of their shape, age, and health state, which was not

possible with color-based segmentation algorithms. By non-invasively extracting color index data from the segmented leaf area, araDEEPopsis delivered highly resolved quantitative data that we successfully applied in genome-wide association (GWA) analysis of anthocyanin content. To the best of our knowledge, araDEEPopsis is the first tool to return quantitative data on plant senescence, by which we were able to identify genetic variants that drive premature senescence. Because Arabidopsis is the most widely used model species in molecular plant biology and plant genetics, we centred our pipeline on this species and show that araDEEPopsis can accurately segment other species, too.

araDEEPopsis is a versatile and robust tool that can extract various biologically relevant plant phenotypes with unprecedented accuracy. Because the pipeline is written in Nextflow, researchers with little computational background can install and run it on their personal computer or in a high-performance computing environment.

Results

Color-guided segmentation can be misguided by image and object parameters

We wanted to explore performance and accuracy of classical color-based segmentation and of a self-learning algorithm in segmenting Arabidopsis rosettes from a large automated phenotyping experiment. We monitored growth of 210 Arabidopsis accessions from the 1001genomes panel [12] in natural soil, under climate controlled conditions, with six replicates per genotype, from seed to flowering stage. Using an automated phenotyping system (https://www.viennabiocenter.org/facilities/plant-sciences/phenotron/), we recorded top-view images twice per day, cropped them to frames containing single pots, then subjected them to rosette area measurements. We first used *LemnaGrid* (Lemnatec), which relies on color channel information to segment plants in top-view images and is commonly used for Arabidopsis phenotyping [13]. This resulted in accurate segmentation of young and healthy plants composed mainly of green tissue (Fig 1). However, because the plants in our dataset were grown for long periods in natural soil, many accumulated high levels of anthocyanins in their leaves, either because of genetic determinants or as a general stress response to the natural soil and its abiotic or biotic composition; others showed onset of senescence at later stages of the experiment. On images of such plants, segmentation often failed completely or resulted in only partial or inaccurate segmentation of the plant (Fig 1), showing that color-based segmentation is sensitive to the deviations from the expected color composition of the object to be detected.

Training of a proof-of-principle model

Assuming that self-learning algorithms would achieve higher accuracy while at the same time being more robust towards shifts in color patterns, we developed araDEEPopsis (*Ara*bidopsis *deep*learning based *op*timal *s*emantic *i*mage *s*egmentation), which is built around the deep-learning model DeepLabV3+. We hypothesized that such a method would faithfully extract

Arabidopsis rosettes from top-view images, independent of developmental stage or phenotype. To test this, we generated a 'ground truth' training set on which the model could be trained. We initially generated a small dataset of 300 images, in which we manually annotated the rosette area, deliberately excluding senescent leaves. From these 300 annotations, we randomly selected 80% for training the model and kept 20% separate for evaluating its accuracy.
After training, this model, which we will refer to as 'model A' from here on, performed well on plants at various developmental stages (Fig 1), which encouraged us to go forward and generate a much larger and more fine-grained ground truth dataset.

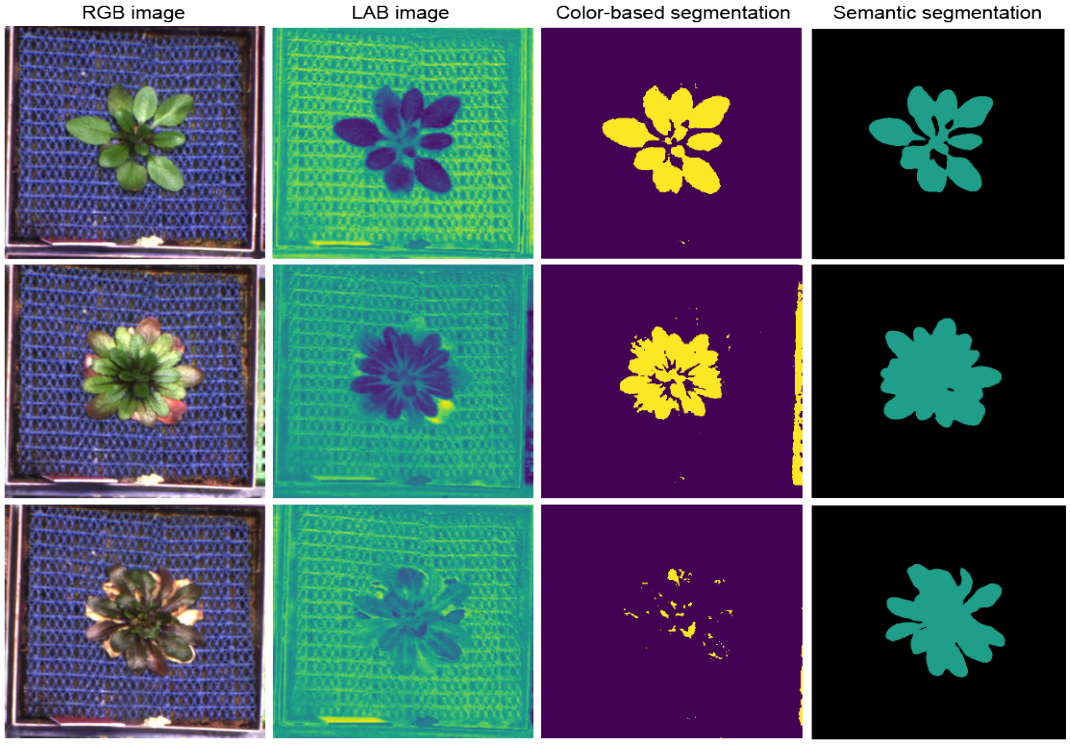

Fig 1. Performance of color-based vs. semantic segmentation. The leftmost column shows the original RGB images of a representative Arabidopsis plant from our phenotyping experiment at three different developmental stages. The second column shows the same images transformed into the LAB colorspace, which enhances contrast between green and other colors. Binary thresholding (third column), based on LAB input and used for color-based segmentation, results in difficulties to correctly segment older or anthocyanin-containing plants. Semantic segmentation by araDEEPopsis is insensitive to color and contrast of the plant (rightmost column).

Advanced models for differentiated plant area classification to classify individual leaves

From the example images shown in Fig 1, it is apparent that the health state changes not only in the context of the whole rosette but that there are substantial differences between individual leaves or parts of leaves, with some accumulating high anthocyanin levels and others entering senescence as part of the plant's life cycle or in response to stress. We therefore asked if it would be possible to train a model that would be able to semantically extract such features from leaves and assign them to different classes, depending on the phenotypic appearance. Not only would

this allow finer resolution when assessing the overall state of the plant but would also enable discerning differently colored areas.

Reasoning that it should be possible to segment senescent leaves separately, we generated a second, larger and more finely annotated training dataset compared to the initial one. This second dataset consisted of 1,375 manually annotated top-view images of *A. thaliana* plants from the phenotyping experiment described above, from which we again kept 20% separate for validation purposes. The images were selected semi-randomly such that they covered all various developmental stages and included healthy-looking as well as stressed plants that exhibited altered color profiles due to anthocyanin accumulation or senescence. Instead of manually annotating the whole rosette, as we had done for the initial training set, we annotated all leaves of each rosette individually and manually assigned each leaf to one of three different classes, depending on its appearance:

- green
- anthocyanin- rich
- senescent or dead

From these annotations, we generated image masks for two additional models complementing the initial one-class model A. For the two-class-model B, we classified senescent (class_senesc) vs. non-senescent leaves (class_norm), whereas the three-class-model C was trained to differentiate between senescent (class_senesc), anthocyanin-rich (class_antho), and green (class_norm) areas (Fig 2A). We then used these annotated three sets of masks for transfer learning of a publicly available *xception65 [7]* based DeepLabV3+ checkpoint that has been pre-trained on the ImageNet dataset (see Methods) [14,15].

After training of all three models had completed, we assessed their performance according to the mean intersection over union (mIoU) for all pixel-level classifications, which is defined as the mean fraction of true positives divided by the sum of true positives, false positives and false negatives over all annotated classes. On the validation dataset, model A, B, and C reached an mIoU of 96.8%, 84.0%, 83.1%, respectively.

Next, we compared how each model performed at segmenting all 148,792 rosette images of the dataset. First, we asked how much the different classes contributed to the area segmented as 'plant'. Averaged across all images, model A, which ignores senescent leaves, classified 12.5% of image area as plant tissue (Fig 2B). Models B and C both resulted in approximately 14% of image area classified as plant area (including senescent areas ignored by model A) (Fig 2B). The fraction of class_senescent segmentation was almost identical in both models. To further test whether our most complex model, model C, performed as expected, we analysed the ratio of the three classes over the course of the experiment. As expected, we saw a sharp increase in anthocyanin-rich area from 30 days after sowing onwards, followed by an increase in senescent segmentations 10 days later (Fig 2C). Relative classification of green tissue decreased accordingly. This showed that model C was overall able to capture senescence becoming more and more frequent the older the plants became, accurately reflecting plant development.

Next, we wanted to compare the three models, to see if they behaved differently in segmenting the combined rosette area. Using the total pixel area classified as leaves in each image, we found that models B (two-class) and C (three-class) correlated most strongly ($R^2 = 0.998$; Fig 3C). Model A, our one-class model trained on annotations of whole rosettes and not of individual leaves, correlated less well with models B and C ($R^2 = 0.985$ and 0.982, respectively) and had a tendency

to segment larger areas than either of the other two (Fig 2C). Visual inspection of overlays of the segmentations produced by model A with the original image revealed that model A segmentations frequently included areas between leaves, which were ignored by models B and C, explaining the larger rosette sizes measured by model A (see segmentations in Fig 2C). We believe that these differences are related to the different annotation strategies (whole rosettes in model A and more refined per-leaf annotations for the larger datasets in models B and C). When generating the ground truth for models B and C, we noticed that it is non-trivial to make consistent decisions as to when a leaf should be annotated as senescent or rich in anthocyanins, as the transitions are gradual, and classification can become subjective. This might also explain the decrease in mIoU scores that we observed with the two- and three-class models. We tried to mitigate user bias by having different individuals generate annotations, thus ensuring that the model would faithfully learn the features of interest.

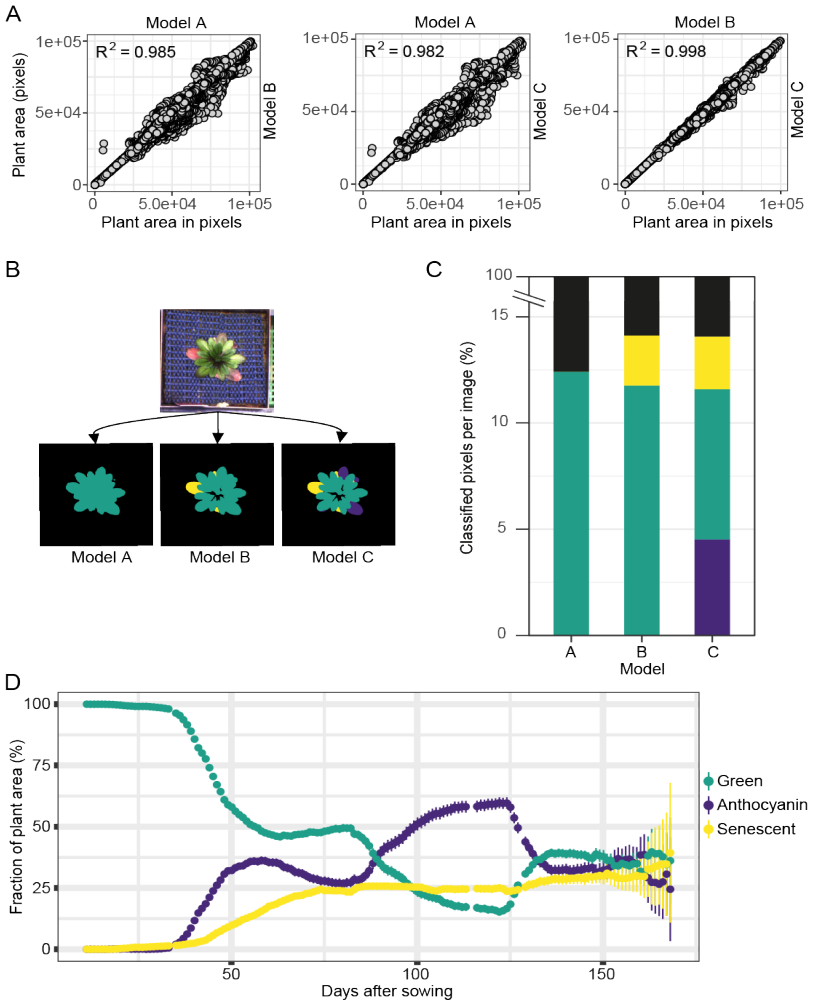

Fig 2. The three models available through araDEEPopsis. (A) Scatterplots showing correlations between measurements of plant area (includes all leaf states except senescence) between the one-class model A, the two-class model B, and the three-class model C. (B) Example of segmentation results from different models for a single Arabidopsis individual. (C) Overall comparison for percentage of all classified pixels across all 148,792 images between the three available models. (D) Relative number of pixels classified as green, anthocyanin rich or senescent by model C over time. The mean percentage of pixels assigned to each class is shown per day. Error bars indicate 95% confidence intervals.

araDEEPopsis pipeline

The araDEEPopsis pipeline is written in Nextflow [16]. The pipeline is fully open-source, licenced under GPLv3, and is presented in detail in the following sections. Briefly, the pipeline takes a folder of images as an input, splits the total image set into batches of equal size, and performs semantic segmentation on the images using a model of choice (see above *"Advanced models..."*). The semantic segmentation output is then used to extract morphometric and color index traits from each individual image (see below *"Extraction of morphometric and color index traits"*) . Quality control and exploratory analyses can be carried out after the pipeline has finished by launching a bundled *Shiny* [17] application. In addition to offering some straightforward visualisation of the results, the *Shiny* application also provides an interface to merge metadata and araDEEPopsis output for downstream analysis (Fig 3).

araDEEPopsis is designed to segment plants from images that contain only a single plant and we do not recommend usage on images containing multiple plants. In case their respective phenotyping system records whole or partial trays of plant pots, as is common for many automated phenotyping platforms, users will be required to pre-process these images and divide them into sub-images, each only containing a single individual plant.

Extraction of morphometric and color index traits from segmented images

Next, we used the segmentation masks obtained from araDEEPopsis to extract morphometric and color index traits of the plant (Fig 4), using the python library scikit-image [18,19]). For example, major and minor axes of an object-surrounding ellipse are a measure of the aspect ratio of the plant and inform on whether the rosette is rather round or rather elongated along one axis. The area of a convex hull surrounding the object is a measure of plant size. The solidity of the plant object is calculated by dividing the area of the plant by the area of the convex hull. Finally, dividing the plant area by the area of its bounding box, i.e., a minimal box enclosing the object, is representative of the overall extent of the plant (Fig 4). Depending on the type of downstream analyses, these indirect object parameters can be used as proxies for overall or specific growth.
The segmentation masks are also used to extract color index information from the plant area. These traits are based on the intensity of each RGB color channel per pixel. Simple color index traits are, for example, the average intensity of the green channel over all pixels classified as 'plant'. We have implemented the color index traits described by [18] in our pipeline (Fig 4C). These traits are calculated for each class individually and for a compound class termed “plant_region”, which contains both class_norm and class_antho (only for model C). Details on the color index traits are shown in Fig 4C and provided in the Methods section.

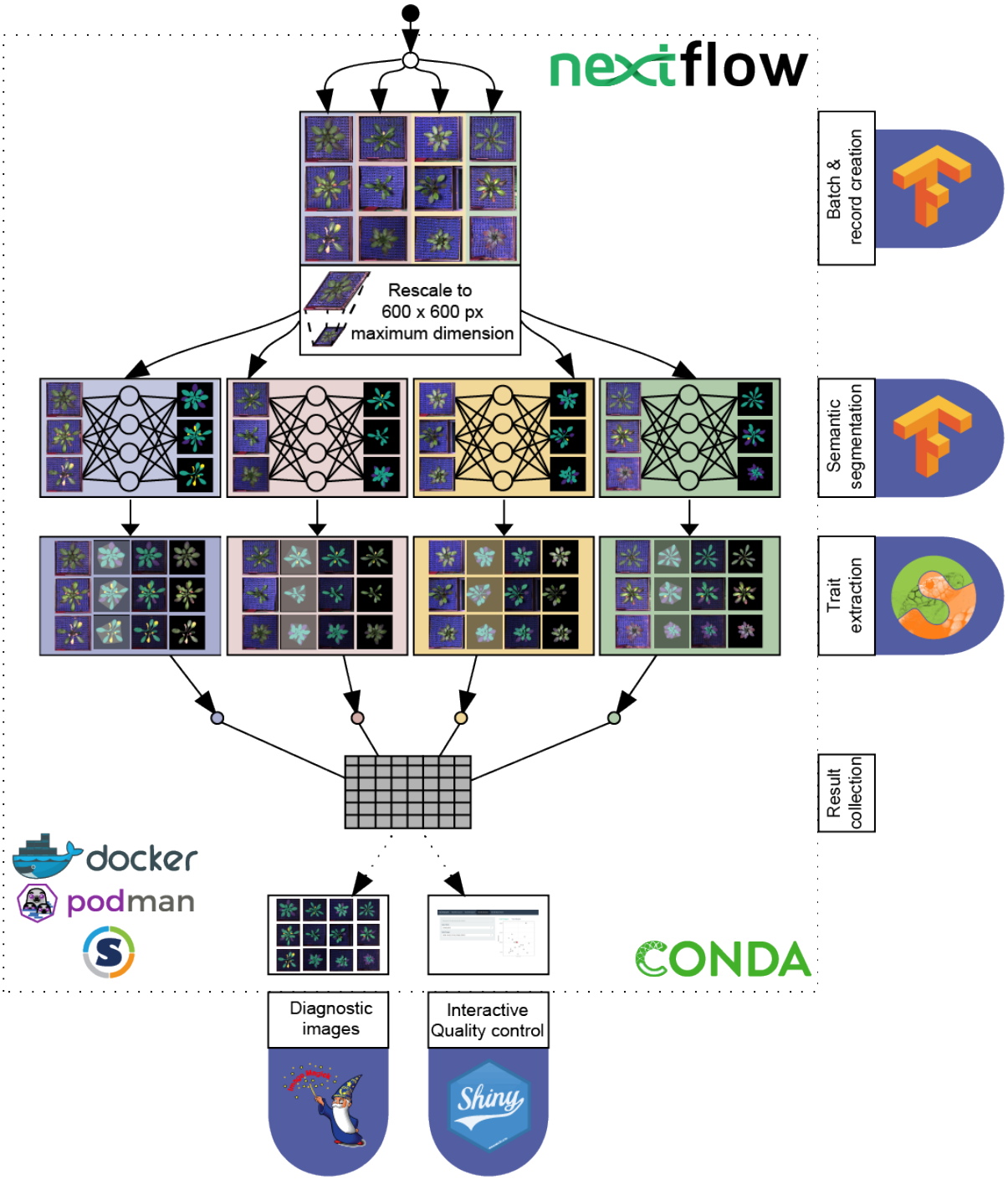

Fig 3. The araDEEPopsis pipeline. A folder containing image files is passed to the pipeline. The total number of images is split into batches of a fixed size (indicated by different background colors). Batches are processed in parallel: first, the segmentation is performed on each batch, and then traits are extracted from the segmented images. The output from all batches is collected in one final table, the results table. In addition, the pipeline produces diagnostic images, which show the segmentation results overlayed on the original image, color coded segmentation masks, and background-subtracted plant rosettes. These diagnostic images can be explored in a *Shiny* app [17], which is launched at the end of an araDEEPopsis run, such that users can visually inspect the segmentations and verify their accuracy.

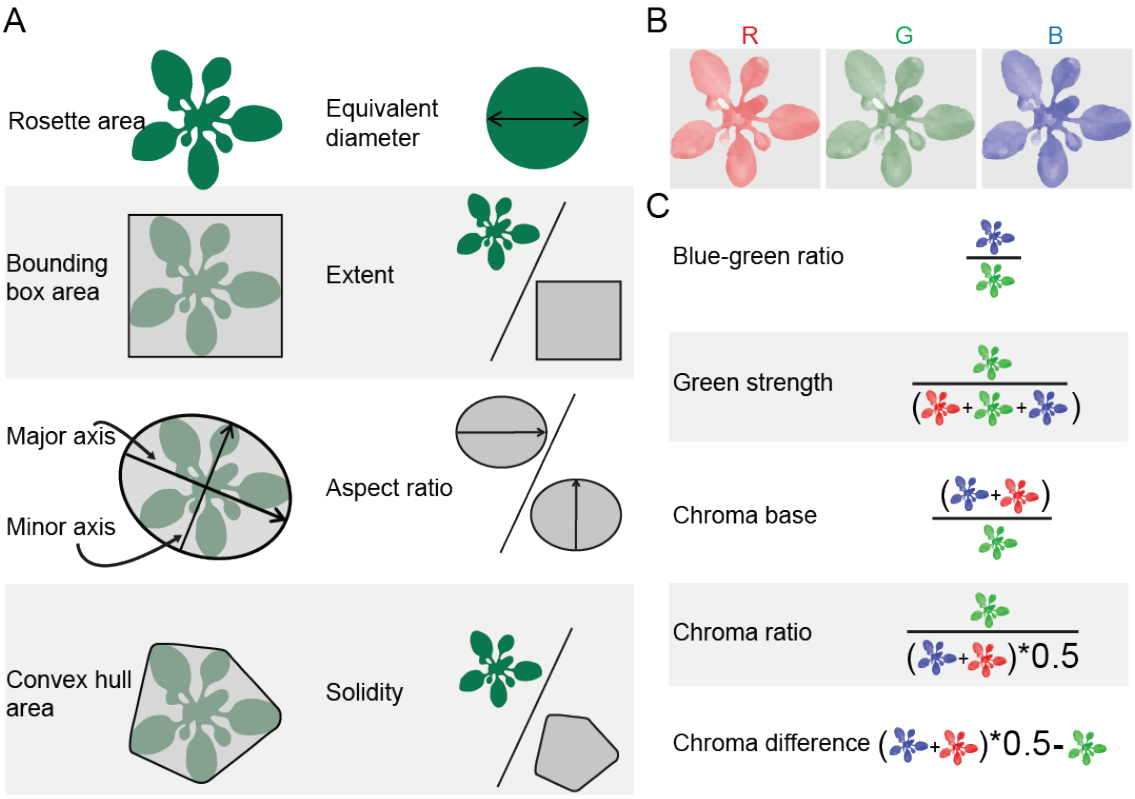

Fig 4: Morphometric and color index measurements extracted from segmentation masks. (A) Morphometric traits extracted using the python library scikit-image [18,19]). (B) Separation of the segmented plant image into red (R), green (G), and blue (B) color channels. (C) color-channel indices calculated as described in [18]. The differently colored plants represent mean values of the different color channels; see (B).

Validation against public datasets

To validate the accuracy of our pipeline, we reanalysed publicly available images and compared the output of araDEEPopsis to the published analyses. This served two purposes: first, we wanted to verify that rosette features extracted based on araDEEPopsis segmentation were accurately reproducing published data. Second, we wanted to test whether our pipeline would remain robust when using data generated on different phenotyping platforms, with different image recording systems, pot sizes and shapes, and background composition. We used images from three published studies [20–22], which were generated on the RAPA platform [23], and re-analyzed them using araDEEPopsis.

Correlation analysis of published rosette area and our own measurements resulted in R^2 values of 0.996 [21] and 0.979 [20], respectively (Fig 5). Despite this overall very strong correlation, we noticed individual images in which our segmentation disagreed with the published one. When inspecting the respective images and the segmentation masks from both analyses, we could confirm that the respective rosette segmentations in the original analysis were inaccurate or incomplete, and that araDEEPopsis segmentation was in strong agreement with the actual rosette in the image, even when plants had strong aberrant phenotypes or grew on low-contrast background (Fig 5).

Re-analysis of a third dataset [22], generated on yet another phenotyping platform (https://www.psb.ugent.be/phenotyping/pippa) [24]), also showed high correlation between our and published measurements (Fig S1). Unfortunately, original segmentations were not available for these images, but when inspecting some of the outliers, we again noticed highly accurate segmentation by araDEEPopsis relative to the actual image.

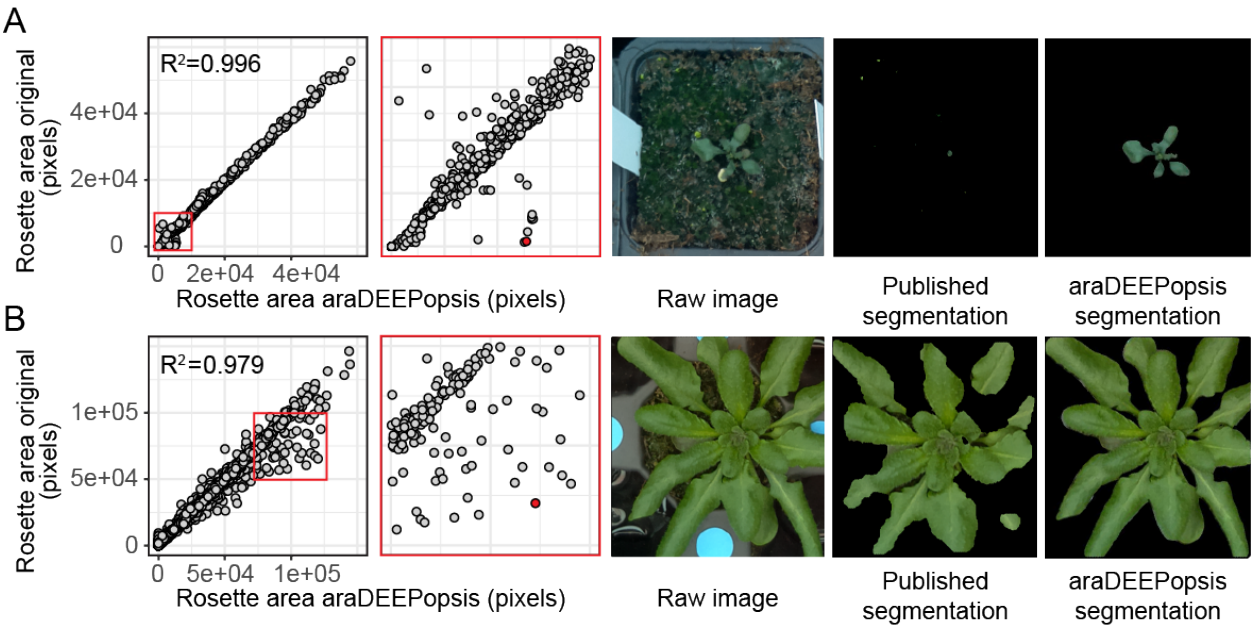

Fig 5. Validation of araDEEPopsis output against published data. (A) Validation against data from [21]. The leftmost panel shows the correlation between values produced by araDEEPopsis against published data. The second panel is a magnification of the area boxed-in in red in the first panel, highlighting disagreeing measurements (red dot). The third panel shows the original image to the highlighted data point panel two. Panel four shows the original segmentation, panel five the segmentation by araDEEPopsis. (B) Validation against [20] with the same order as in (A).

Validation by genome wide association studies

Validation of color index traits

The ultimate purpose of phenotype measurements often is to relate them to genetic data in order to understand genetic determinants that control phenotypic traits. We therefore wanted to test whether araDEEPopsis provided accurate and meaningful measurements for Arabidopsis that could be used to search for genetic associations in genome-wide association (GWA) studies, relying on the 1001genomes variant information [12]. Instead of morphometric traits such as size and growth rate, which are usually highly polygenic and therefore difficult to test, we wondered if we could make use of additional layers of image information provided by araDEEPopsis as a proxy for physiological parameters. We hypothesized that RGB color information of the plant, i.e., the respective pixel intensities in the three color channels, would provide information on anthocyanin content and hence on the physiological state of the plant; it was previously shown that tissue color highly correlates with anthocyanin content [18]. Using the araDEEPopsis segmentations, which accurately reflect the area covered by the rosette, we extracted color

information by collecting data from the different RGB channels from the segmented object area, as explained above (Fig 4B,C). We then used *limix* [25] to perform GWA analysis on the "chroma ratio" (Fig 4C) of the rosette area 37 d after sowing. The "chroma ratio" is calculated by dividing the mean intensities in the green channel, divided by half of the sum of the intensities of blue and red channels. It is therefore inversely correlated with the amount of anthocyanin accumulation, and increases when anthocyanin content is low. We found strong associations with regions on chromosomes 1, 2, 4 and 5 (Fig 6A). When ranking by -log10(p), the second-highest ranking SNP is on chromosome 4 and the closest annotated gene from this SNP is *ANTHOCYANINLESS2* (*ANL2;* AT4G00730). *ANL2* encodes a homeobox transcription factor and has been implicated in controlling anthocyanin accumulation in sub-epidermal tissues [26]. Mutants in *ANL2* accumulate less anthocyanin in sub-epidermal leaf tissue. In our data, accessions carrying the alternative allele near *ANL2* displayed a higher chroma ratio, indicative of lower anthocyanin accumulation (Fig 6B). This shows that araDEEPopsis is able to non-invasively extract quantitative color index data of the whole rosette that can be applied to genetic association analysis.

Validation of pixel-level classifications

Having validated that color profiles of the whole plant area can be used to extract informative traits, we wondered if we would be able to identify genetic variants significantly associated with the relative amount of senescent or dead tissue during early development. We conducted a GWA analysis in which we used the relative amount of senescent tissue at 25 d after sowing as the phenotypic trait. This is early in the vegetative phase of plant development, and the relative senescent plant area is small compared to later stages (see Fig 3C). We found several genomic regions that displayed significant associations with this phenotype, the most striking of which was located on chromosome 4. The highest-ranking SNP in that locus was located within the coding region of ACCELERATED CELL DEATH 6 (*ACD6;* AT4G14400). The *ACD6* locus has been extensively studied in Arabidopsis and was identified as associated with vegetative growth, microbial resistance, and necrosis, in particular in F1 hybrids [27–29]. Our plants were grown on natural, non-sterilized soil from a field site near Zurich in Switzerland [30]; it is therefore fair to assume that the microbial load and potential biotic stress levels were higher than in standard potting soil. We observed a ~8-fold difference in median relative senescent tissue per plant (Fig 6C,D) from approximately 0.5% of total segmented pixels in accessions carrying the reference allele to around 4% in those carrying the alternative allele. This is in line with published studies, which showed that certain alleles of *ACD6* cause auto-immunity in Arabidopsis, which phenotypically manifests as premature senescence or necrosis [27].

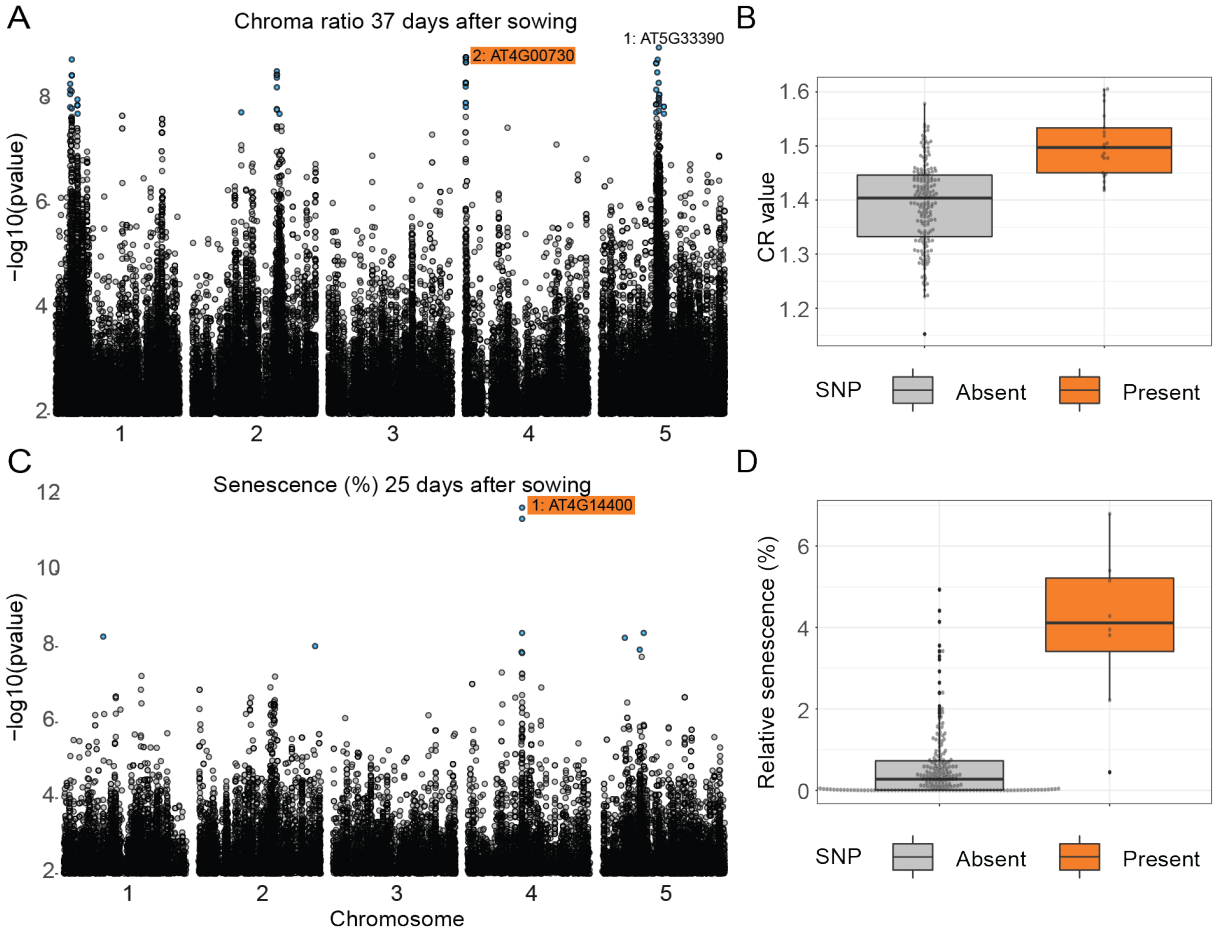

Fig 6. Genome-wide association (GWA) analyses based on araDEEPopsis output. (A) Results of GWA on the trait "chroma ratio" 37 days after sowing. The log10-transformed p-value for each SNP is plotted (only SNPs with minor allele frequency >5% are shown). SNPs that are significant after Bonferroni correction are shown in bluel. For the two highest-ranking SNPs, the closest gene is indicated. (B) Chroma ratio of plants 37 days after sowing, split by accessions carrying the reference and alternative alleles of the SNP close to *AT2G00730* (*ANL2*), highlighted in (A). (C) Same as (A) for the trait "relative senescence" 25 days after sowing. AT4G14400 corresponds to *ACD6* (see text). (D) Relative senescence 25 days after sowing, split by accessions carrying the reference and alternative alleles of the SNP close to *ACD6*, highlighted in (C).

Analysis of other plant species

While our specific goal was to be able to faithfully segment Arabidopsis rosettes, we wanted to see if our method would generalize to other plant species with similar overall morphology. We therefore tested araDEEPopsis on top-view images of the Brassicaceae *Thlaspi arvense*, which has a similar leaf shape and phyllotaxis than those of Arabidopsis but does not for a flat rosette on the ground. Without training a model on images of *T. arvense*, i.e., using the model trained on Arabidopsis, segmentation masks matched the original plant highly accurately (Fig 7), showing that araDEEPopsis is robust to variations in plant morphology, as long as these do not deviate from the general architecture that was used in the training set.

Fig 7. Segmentation of *T. arvense*. The leftmost panel shows the original image of a *T. arvense* individual. The middle panel shows the segmented mask and convex hull area. The rightmost panel shows the plant accurately cropped from the original image using the mask generated by araDEEPopsis.

Discussion

Here, we have presented araDEEPopsis, a versatile tool to extract phenotypic measurements from small or large sets of plant images in an unsupervised manner. araDEEPopsis is able to faithfully identify, segment and measure plants from top-view images of rosettes such as those of Arabidopsis, using deep learning methodology. Our tool is easy to use, runs on personal computer and high-performance computing environments, and is robust against variations in overall plant appearance, image quality, and background composition, making it superior to common color-segmentation based tools.

A fast and easy-to-use segmentation tool

Our models were trained on sets of 240 and >1000 images, respectively, with manually annotated 'ground truth' segmentations, which required a non-negligible amount of manual labor. We believe that these investments were warranted, because araDEEPopsis produced highly accurate segmentations and morphometric measurements that highly correlated with published data; occasional deviations from the originally reported data could in most cases be related to more accurate segmentation by araDEEPopsis. Besides being accurate and requiring limited labour, the method is also fast: the fully automated analysis of 100 images took 23 minutes on a personal computer (8GB ram, Intel i5 2GHz). Depending on available resources such as memory and number of cores, the implementation in nextflow enables straight-forward parallelization, allowing for a significant increase in speed when deployed in high-performance computing environments. At the same time, nextflow is agnostic to the computational infrastructure being used, making it straightforward to deploy araDEEPopsis to any type of computer. While training of the model greatly benefitted from the availability of GPUs, image predictions can also be carried out using CPUs in a time-efficient manner.

araDEEPopsis is robust to image characteristics and background composition

Many of the images in our training dataset had blue plastic meshes as background. This could raise the concern that the model might have learned to classify pixels belonging to the background

rather than the plant of interest, i.e., to segment the plant as "non-background", which would render the tool unreliable when using it on images with a different background composition. However, by testing araDEEPopsis on images acquired in other phenotyping environments using different physical backgrounds, and on images of potted plants with substantial algal growth surrounding the Arabidopsis plant of interest, we observed that segmentation of leaves remained accurate and was agnostic to the background (see Fig 4a). Ideally, araDEEPopsis should be used with images that are homogeneous within one set regarding light intensity, exposure time, etc. Segmentation still works for images with various light intensities (Fig S2), and plant size and geometry can still be analyzed, but quantification of color intensities based on the original image is no longer comparable in that case.

Accurate determination of plant health state

araDEEPopsis not only segments rosettes and extracts morphometric parameters, it also allows to extract color channel information, which can be used to make assessments of the plant's anthocyanin content and overall health status. We have shown that these results can be used, for example, for quantitative genetics approaches (Fig. 6).
Besides a simple one-class model (model A) that is able to segment non-senescent leaves,, araDEEPopsis includes two models that allow segmentation of anthocyanin-rich and/or senescent areas, depending on their color composition. To the best of our knowledge, this makes araDEEPopsis the first tool capable of automatically and reliably classifying senescent leaves and of distinguishing healthy, green leaves from stressed, anthocyanin-loaded ones. Our models could also be extended to additional classes, depending on the specific needs of researchers and the phenotypes of interest.

Outlook and Perspective

Our study shows that transfer learning is a valuable approach to overcome the phenotyping bottleneck. Our trained models have restrictions with regard to the angle at which the images are recorded and currently perform best for top-view images. The model was designed for the analysis of *Arabidopsis* rosettes but was able to also segment images of other species. We believe that the approach could be extended to a broader range of species and also to other types of plant images, e.g. of side-view angles. These adjustments would require expert annotation of the corresponding images but can be performed by an experienced researcher and trained helpers in a reasonable amount of time. Alternatively, it would be desirable to collect possibly pre-existing annotations from different research groups and various species into a centralized repository that could serve as a powerful resource for phenomics in plant science and breeding.
ImageNet, the dataset that was used to pretrain the baseline model we built upon, contains images of dogs, airplanes, bicycles, etc., but no detailed annotations of plant species. It is therefore remarkable that ImageNet pretrained models enable such high accuracy when retraining the last layer with a relatively small set of 300 manually annotated images. Ultimately, this highlights the potential such publicly available databases and models hold for research and suggests that we should exploit such resources more frequently.

While our current implementation focusses on large-scale image data from HT phenotyping platforms installed in controlled greenhouse environments, we envision the approach to also be beneficial for field phenotyping purposes. For example, choosing a smaller network backbone such as *MobileNetV3* [31] could enable field researchers to measure plant morphometry on the go using their smartphone and could thus facilitate data collection and interpretation in the field.

Methods

Plant growth

Plants were grown in long-day conditions (16h light (21°C), 8 h dark (16°C), 140µE/m²s) with 60% relative humidity and watered twice per week. Sowing was done on 20th September 2018 and plants were imaged two times per day. Plants were monitored daily for emerging inflorescences and flowering accessions were removed from the growth chamber. Plants that had not flowered before 11th December 2018 were subjected to vernalization: the chamber was cooled to 4°C (ramped). During watering, the temperature was raised to 5°C. This program ended on 21st January 2019.

Training & Validation

The *Computer Vision Annotation Tool* [32] was used for manual image annotation; custom scripts were used to produce annotation masks from the XML output. The publicly available *DeepLabV3+* [14,15] code was modified to enable model training on our own annotated training sets. The code used for training as well as download links for our annotated training datasets is available here: https://github.com/phue/models/tree/aradeepopsis_manuscript/research/deeplab.
For model evaluation, we split the annotated sets 80:20: 80% of the images were used to train the model and 20% for its evaluation.
A transfer learning strategy was employed by using a model checkpoint based on the *xception65* architecture [7] that has been pretrained on the *ImageNet* dataset [33]. Starting from that, training was implemented in an asynchronous manner using between-graph replication on our in-house slurm cluster, allowing to use a total of 16 Nvidia Tesla V100 GPUs across 4 compute nodes. The training was performed according to the protocol described in literature [14,15] with the following changes: To account for the number of GPUs, the training batch size was set to 128 and a base learning rate of 0.1 was applied and decayed according to a polynomial function after a burn-in period of 2000 training steps. Images were randomly cropped to 321x321 pixels, and training was stopped after 75,000 iterations.

Implementation

Based on the trained models, an image analysis pipeline was implemented in *Nextflow [16]*. The workflow is outlined in Fig 1. *Nextflow* allows external pipeline dependencies to be packaged in a *Docker* container which we provide at https://hub.docker.com/r/beckerlab/aradeepopsis/. The

container can be run using *Docker* [34], *podman* [35] , or *Singularity* [36]. Alternatively, dependencies can be automatically installed into a *Conda* [37] environment.
The pipeline follows a scatter-gather strategy to scatter the input data into equally sized batches of arbitrary size allowing for parallel processing, after which the results are again collected into a single output table. After the initial splitting, all images in one batch are first converted to TFrecord files using *TensorFlow* [38]. The records are then served to the trained *TensorFlow* model of choice in order to generate segmentation masks, containing pixelwise classification.
In the next step, morphometric traits are extracted from the using the regionprops function of the python library *scikit-image* [19] on a per-class basis. In addition, color channel information is extracted in the form of mean pixel intensities for each of the channels in the original RGB image. This happens within the region determined by the segmentation mask and is also done for each class. Based on the color channel means, chroma indices are calculated as follows [18]

$$Chroma_{base} = \frac{N_{blue} + N_{red}}{N_{green}}$$

$$Chroma_{difference} = \frac{N_{blue} + N_{red}}{2} - N_{green}$$

$$Chroma_{ratio} = \frac{N_{green}}{\frac{N_{blue} + N_{red}}{2}}$$

$$BG_{Ratio} = \frac{N_{blue}}{N_{green}}$$

$$S_{green} = \frac{N_{green}}{N_{red} + N_{green} + N_{blue}}$$

The results from these measurements are then collected from all processed batches, ultimately resulting in a single table containing a total of 78 traits. Based on the segmentation masks in combination with their corresponding input images, the pipeline also produces diagnostic images showing a color-coded mask, the cropped plant region, the convex hull and an overlay image between original and mask.
These single-plant diagnostics are then, in an optional step, merged to produce summary diagnostics using *ImageMagick [39]* or can be viewed in an interactive *Shiny* application [17], allowing for fine-grained inspection of segmentation quality, pixel classification, correlations between traits as well as time-resolved data visualization if appropriate metadata is provided.

Genome-wide association studies

We performed GWA analysis on the traits produced by araDEEPopsis using limix [25]. We used the average of each trait per accession per day, and performed GWA analysis by fitting a linear mixed model using *limix* [25]. To associate phenotype and single nucleotide polymorphisms, we used the 1,135 genotype SNP matrix and the corresponding kinship matrix, subset to those accessions where we had trait information. We screened the results for interesting trait-date combinations and followed these up using *Arabidopsis* specific tools developed in-house

(https://github.com/Gregor-Mendel-Institute/gwaR). The analysis is detailed in supplementary document 1.

Acknowledgements

We thank James M. Watson for critical reading of the manuscript and valuable comments. Karina Weiser Lobão helped in the manual image annotation. We thank Dario Galanti and Oliver Bossdorf for providing images of *T. arvense*. Klaus Schlaeppi and Selma Cadot provided the field soil. We thank Núria Serra Serra and Jorge Isaac Rodriguez for testing the analysis pipeline. Erich Birngruber gave advice on distributed model training on a slurm cluster.
Arabidopsis phenotyping was performed with support of the Plant Sciences Facility at Vienna BioCenter Core Facilities GmbH (VBCF), member of the Vienna BioCenter (VBC), Austria. This work was supported by a grant of the European Research Council (ERC) to C.B. (grant ID 716823, "FEAR-SAP"), by the Austrian Academy of Sciences (ÖAW), and the Max Planck Society.

References

1. Furbank RT, Tester M. Phenomics--technologies to relieve the phenotyping bottleneck. Trends Plant Sci. 2011;16: 635–644.

2. Schneider C a., Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods. 2012;9: 671–675.

3. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, et al. Fiji: an open-source platform for biological-image analysis. Nat Methods. 2012;9: 676–682.

4. Rother C, Kolmogorov V, Blake A. "GrabCut": interactive foreground extraction using iterated graph cuts. ACM SIGGRAPH 2004 Papers. New York, NY, USA: Association for Computing Machinery; 2004. pp. 309–314.

5. Gehan MA, Fahlgren N, Abbasi A, Berry JC, Callen ST, Chavez L, et al. PlantCV v2: Image analysis software for high-throughput plant phenotyping. PeerJ. 2017;5: e4088.

6. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. openaccess.thecvf.com; 2016. pp. 770–778.

7. Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions. arXiv [cs.CV]. 2016. Available: http://arxiv.org/abs/1610.02357

8. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in Neural Information Processing Systems 25. Curran Associates, Inc.; 2012. pp. 1097–1105.

9. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86: 2278–2324.

10. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323: 533–536.

11. Sharif Razavian A, Azizpour H, Sullivan J, Carlsson S. CNN features off-the-shelf: an astounding baseline for recognition. Proceedings of the IEEE conference on computer vision and pattern recognition workshops. cv-foundation.org; 2014. pp. 806–813.

12. 1001 Genomes Consortium. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell. 2016;166: 481–491.

13. Arvidsson S, Pérez-Rodríguez P, Mueller-Roeber B. A growth phenotyping pipeline for Arabidopsis thaliana integrating image analysis and rosette area modeling for robust quantification of genotype effects. New Phytol. 2011;191: 895–907.

14. Chen L-C, Papandreou G, Schroff F, Adam H. Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv [cs.CV]. 2017. Available: http://arxiv.org/abs/1706.05587

15. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. arXiv [cs.CV]. 2018. Available: http://arxiv.org/abs/1802.02611

16. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35: 316–319.

17. RStudio I. Shiny: Easy Web Applications in R. 2014.

18. Del Valle JC, Gallardo-López A, Buide ML, Whittall JB, Narbona E. Digital photography provides a fast, reliable, and noninvasive method to estimate anthocyanin pigment concentration in reproductive and vegetative plant tissues. Ecol Evol. 2018;8: 3064–3076.

19. van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. scikit-image: image processing in Python. PeerJ. 2014;2: e453.

20. Barragan CA, Wu R, Kim S-T, Xi W, Habring A, Hagmann J, et al. RPW8/HR repeats control NLR activation in Arabidopsis thaliana. PLoS Genet. 2019;15: e1008313.

21. Capovilla G, Delhomme N, Collani S, Shutava I, Bezrukov I, Symeonidi E, et al. PORCUPINE regulates development in response to temperature through alternative splicing. Nat Plants. 2018;4: 534–539.

22. Baute J, Polyn S, De Block J, Blomme J, Van Lijsebettens M, Inzé D. F-Box Protein FBX92 Affects Leaf Size in Arabidopsis thaliana. Plant Cell Physiol. 2017;58: 962–975.

23. Vasseur F, Bresson J, Wang G, Schwab R, Weigel D. Image-based methods for phenotyping growth dynamics and fitness components in Arabidopsis thaliana. Plant Methods. 2018;14: 63.

24. Coppens F, Wuyts N, Inzé D, Dhondt S. Unlocking the potential of plant phenotyping data through integration and data-driven approaches. Current Opinion in Systems Biology. 2017;4: 58–63.

25. Lippert C, Casale FP, Rakitsch B, Stegle O. LIMIX: genetic analysis of multiple traits. bioRxiv. 2014. p. 003905. doi:10.1101/003905

26. Kubo H, Peeters AJ, Aarts MG, Pereira A, Koornneef M. ANTHOCYANINLESS2, a homeobox gene affecting anthocyanin distribution and root development in Arabidopsis. Plant Cell. 1999;11: 1217–1226.

27. Todesco M, Balasubramanian S, Hu TT, Traw MB, Horton M, Epple P, et al. Natural allelic variation underlying a major fitness trade-off in Arabidopsis thaliana. Nature. 2010;465: 632–636.

28. Świadek M, Proost S, Sieh D, Yu J, Todesco M, Jorzig C, et al. Novel allelic variants in ACD6 cause hybrid necrosis in local collection of Arabidopsis thaliana. New Phytol. 2017;213: 900–915.

29. Zhu W, Zaidem M, Van de Weyer A-L, Gutaker RM, Chae E, Kim S-T, et al. Modulation of ACD6 dependent hyperimmunity by natural alleles of an Arabidopsis thaliana NLR resistance gene. PLoS Genet. 2018;14: e1007628.

30. Hu L, Robert CAM, Cadot S, Zhang X, Ye M, Li B, et al. Root exudate metabolites drive plant-soil feedbacks on growth and defense by shaping the rhizosphere microbiota. Nat Commun. 2018;9: 2738.

31. Howard A, Sandler M, Chu G, Chen L-C, Chen B, Tan M, et al. Searching for MobileNetV3. arXiv [cs.CV]. 2019. Available: http://arxiv.org/abs/1905.02244

32. Sekachev B. Computer Vision Annotation Tool: A Universal Approach to Data Annotation. In: Intel [Internet]. 1 Mar 2019 [cited 26 Feb 2020]. Available: https://software.intel.com/en-us/articles/computer-vision-annotation-tool-a-universal-approach-to-data-annotation

33. Deng J, Dong W, Socher R, Li L-J, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2009. pp. 248–255.

34. Docker. Available: https://www.docker.com/

35. Podman. Available: https://podman.io/

36. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLoS One. 2017;12: e0177459.

37. Conda. Available: https://conda.io/en/latest/

38. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: A System for Large-Scale Machine Learning. Available: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf

39. ImageMagick Studio LLC. ImageMagick. Available: https://imagemagick.org/