

Unlocking the power of AI for phenotyping fruit morphology in *Arabidopsis*

Kieran Atkins¹, Gina A. Garzón-Martínez², Andrew Lloyd¹, John H. Doonan^{1,*}, and Chuan Lu^{3,*}

¹National Plant Phenomics Centre, IBERS, Aberystwyth University, Aberystwyth SY23 3EE, UK

²Centro de Investigación Tibaitatá, Corporación Colombiana de Investigación Agropecuaria (Agrosavia), Mosquera, Cundinamarca, 250047, Colombia

³Computer Science Department, Aberystwyth University, Aberystwyth SY23 3DB, UK

*Correspondence addresses. John H. Doonan and Chuan Lu, Gogerddan, IBERS, Aberystwyth University, Plas Gogerddan, Aberystwyth, Ceredigion, SY23 3EE, UK.

E-mail: cul@aber.ac.uk; jhd2@aber.ac.uk

Abstract

Deep learning can revolutionise high-throughput image-based phenotyping by automating the measurement of complex traits, a task that is often labour-intensive, time-consuming, and prone to human error. However, its precision and adaptability in accurately phenotyping organ-level traits, such as fruit morphology, remain to be fully evaluated. Establishing the links between phenotypic and genotypic variation is essential for uncovering the genetic basis of traits and can also provide an orthologous test of pipeline effectiveness. In this study, we assess the efficacy of deep learning for measuring variation in fruit morphology in *Arabidopsis* using images from a multiparent advanced generation intercross (MAGIC) mapping family. We trained an instance segmentation model and developed a pipeline to phenotype *Arabidopsis* fruit morphology, based on the model outputs. Our model achieved strong performance with an average precision of 88.0% for detection and 55.9% for segmentation. Quantitative trait locus analysis of the derived phenotypic metrics of the MAGIC population identified significant loci associated with fruit morphology. This analysis, based on automated phenotyping of 332,194 individual fruits, underscores the capability of deep learning as a robust tool for phenotyping large populations. Our pipeline for quantifying pod morphological traits is scalable and provides high-quality phenotype data, facilitating genetic analysis and gene discovery, as well as advancing crop breeding research.

Keywords: plant phenotyping, *Arabidopsis*, fruit morphology, instance segmentation, deep learning; QTL analysis; MAGIC population

Introduction

Phenotyping fruit morphology

Complex traits, such as organ size, are closely related to plant productivity in both crop breeding and natural settings. These traits influence ecological processes like plant–plant competition and reproductive success. Seed production is an important component of reproductive success, intimately associated with specialised organs called fruits that develop from a specialised organ within the flower called the gynoecium. Fruits are the main harvested product from row crops such as wheat, rice, and oats as well as from fruit crops such as strawberries, tomatoes, bush fruits, and apples. Quantity, quality, and other fruit-related traits are therefore of immense agronomic importance, underpinning the economics of modern agriculture. Climate change, increased pest diversity, and an ever-growing population threaten reliable food production, and therefore it is becoming increasingly crucial to identify and phenotype these traits at scale, with a view to integrating knowledge gained into crop production [1].

The factors that influence fruit size and shape are many and can interact in a complex manner (reviewed by Cucinotta et al. [2]). Intrinsic factors include plant signalling molecules such as auxin and gibberellin, whose levels depend on genetic, environmental, and developmental variables. The potential ultimate fruit size is partly set before pollination by, for example, the number of ovules present in the gynoecium, but the realisation of fruit size often depends on the success of pollination. For example, the dry simple fruits found in the Brassica or cabbage family usually

will develop only if pollination has been successful [3]. Within a given species, fruit size and particularly fruit length typically reflect the number of seeds present [4, 5]. The Brassica family also has sophisticated genetic resources available, both in crops such as rapeseed/canola [6] and in undomesticated research models such as *Arabidopsis* [7]. These resources and associated mapping techniques provide an orthologous means of testing the biological significance of any novel phenotyping system.

Arabidopsis thaliana produces many fruits, termed *siliques*, whose small size and fragile structure make it difficult to collect morphological phenotype data by hand consistently and reliably. Unlike many domesticated crops that retain their seeds, *Arabidopsis* is a wild plant in which the seed pods shatter soon after ripening. With often more than 100 *siliques* per plant and experimental designs delivering large populations (thousands to tens of thousands) of plants for genetic and ecological studies, new phenotyping approaches are essential to capture these data in the short time window available. RGB imaging provides a low-cost solution that minimises handling of the individual fruits, provided the computer vision tools are capable of extracting phenotypic information, again at low cost and in a timely manner.

Image-based plant phenotyping offers a nondestructive, high-throughput approach to data collection. While 3-dimensional (3D) imaging, hyperspectral imaging, and multimodal techniques provide rich information, 2-dimensional (2D) RGB imaging remains a practical choice due to its simplicity and efficiency [8, 9]. However, accurate object detection and measurement within plant images

Received: October 21, 2024. Revised: December 14, 2024. Accepted: January 6, 2025

© The Author(s) 2025. Published by Oxford University Press GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

can be challenging due to issues such as organ overlap and the projection of 3D structures into 2D.

Classical computer vision (CV) uses hand-crafted algorithms to analyse images. Classical CV methods can successfully measure fruit morphology [10, 11] and have been used to generate phenotypes for genetic analysis [12, 13]. However, classical CV is typically only suitable for collecting traits from simple scenes (i.e., isolated plants and organs; highly colour-differentiated organs) and is usually unable to robustly handle complex scenes, such as overlapping samples or the presence of other plant biomass. Large-scale phenotyping, especially for fragile or hard-to-harvest fruits, is more effective while the fruits remain attached to the stem. Deep learning offers an alternative approach to classical CV when analysing complex scenes. Deep learning models are trained to replicate the relationship between inputted images and hand-collected training data to then infer outputs on novel, unseen samples. This also allows deep learning to be trained to handle detection and segmentation with specimens still in their biological contexts. Deep learning models are inherently scalable and can be trained or retrained to cope with subjects beyond their initial training and have been used successfully in high-throughput phenotyping [14, 15].

Deep learning-based approaches

Convolutional neural networks (CNNs) are widely used in image-based deep learning. CNNs consist of layers of trainable image filters that progressively learn to extract features for analysis. Lower-level features, such as edges and corners, are captured in earlier layers, while higher-level features, such as object parts and semantic concepts, are developed in the later layers. These higher-level features are then used for higher-level tasks, such as image classification. For example, Atila et al. [16] developed an EfficientNet CNN to classify the presence of different diseases on the leaves of multiple species, achieving a final accuracy of 99.91% in their most successful model. Similarly, Morshed et al. [17] used DenseNet to assess the quality of harvested fruit, classifying each image as having good, bad, or mixed-quality fruit across 6 species, with an accuracy of 99.67%.

However, extracting quantitative data on individual organs that are still attached to the plant requires a different model architecture. For example, Hamidinekoo et al. [18] created DeepPod, a CNN pipeline for counting *Arabidopsis* siliques in images of flattened fruit-bearing branches. Using DenseNet, overlapping grid patches within an image were classified as background, base, body, or tip of a silique, and the individual siliques were reconstructed by joining the segmented pixels in a postprocessing step. While valuable phenotype data can be extracted from organ detection and localisation, they cannot be used to capture detailed morphometric data on the detected objects. To more effectively phenotype the morphology of the fruit, instance segmentation models may offer a better solution.

Instance segmentation combines both object detection with semantic segmentation, detecting objects and segmenting each instance of an object from the background. Instance segmentation models are also designed to be agnostic to the specific shapes of objects being detected. Two predominant methods employed in instance segmentation models are single-stage and 2-stage. Single-stage methods, like YOLACT, an extension of YOLO, predict object bounding boxes and segmentation masks in a single pass through the network [19, 20]. Variants of these architectures have been applied in plant phenotyping for rapid inference, such as identifying tomato fruit and cherry branches for robotic fruit

picking and real-time detection and counting of objects like tea buds [21]. However, these methods often prioritize real-time application over achieving the highest accuracy.

Two-stage methods, in contrast, have an initial object proposal stage followed by a classification stage, which tends to result in more precise detections [22] while still generating outputs in negligible time. Segmentation is then performed within predicted bounding boxes or cluster pixel projections into instances. Mask R-CNN [23], one of the most prominent 2-stage architectures, has been used for phenotyping tasks such as detecting leaf disease [24] and barley seed segmentation from dense scenes [25]. Two-stage methods can be improved by integrating attention mechanisms, like transformer architectures [26, 27], to enable spatially aware feature learning and potentially more accurate segmentation by eliminating the region proposal phase. Although transformers have been used in 3D segmentation of point clouds, such as extraction of rapeseed pod traits [28], their higher computational demands and greater need for training data can make them less practical than R-CNN models in certain situations [27]. Alternatively, multistage methods like Cascade Mask R-CNN can enhance accuracy by progressively refining detection [27, 29], allowing the model to better handle variations in scale and occlusion commonly found in agricultural imagery. In this study, we choose Cascade Mask R-CNN for instance segmentation.

Deep learning has been successfully applied in phenotyping and the results effectively linked to the genome. For example, Wang et al. [30] used a CNN to identify the developmental stages of wheat in the field, evaluating morphology without precise measurement and subsequently used these results for quantitative trait locus (QTL) analysis, detecting novel epistatic interactions affecting flowering time. SegFormer [31], a transformer-based semantic segmentation model, was used to segment rice panicles from shoots. The segmented images enabled extraction of plant-level traits that were then used to identify QTLs associated with variation in yield. However, the model only segmented panicle from background within the images, limiting its resolution for extraction of individual organ traits such as grain length and width, which were instead obtained through manual phenotyping [32]. Thus, it remains uncertain whether deep learning can generate high-quality phenotype data on complex morphological traits, such as fruit size and shape, that can be reliably associated with genetic variation and environmental conditions. For a truly high-throughput phenotyping system, the data should be collected with minimal human intervention and have sufficient resolution to detect small differences between genotypes as well as treatment-induced variation.

To formally assess whether deep learning can be used for precise morphological quantification, we developed an image-based fruit phenotyping pipeline leveraging an instance segmentation deep learning model to extract morphological trait variability and associate this with the genome. We utilised images of fruit-bearing stems of *A. thaliana*, harvested from a multiparent advanced generation intercross (MAGIC) population [33] where samples had been grown as either single plants in isolation or groups of 4 under competitive conditions. This dataset enabled us to assess the ability of a deep learning phenotyping system to accurately capture organ-level morphometric traits in an experimental setting and at a scale useful for the identification of causative genetic variation. To reduce manual effort, we developed a data annotation tool based on the image manipulation software GIMP, speeding up the data annotation process. We trained a Cascade Mask R-CNN model to detect and accurately segment *Arabidopsis* siliques. The phenotypic data derived from these segmentations

were used for QTL analysis, showing that deep learning-derived phenotype data can be confidently associated with known QTL. Furthermore, potential novel genomic regions affecting trait variation were identified, demonstrating the potential of deep learning for gene discovery.

Materials and Methods

Image data

In this study, we utilised images of stem material from an experiment investigating the genetic basis of sibling-sibling competition in *A. thaliana*, using 485 lines from the MAGIC population by Biernaskie et al. [34]. The population was a MAGIC consisting of a core set of 527 recombinant inbred lines (RILs), derived from 19 intermated accessions [33]. Plants were grown in pots containing either single plants (isolation) or in groups of 4 of the same line (groups). A total of 2,801 pots represented 485 genotypes across 2 treatments, and 3 experimental batches, with 3 replicates per batch, were initially collected. Of these, 2,099 were ultimately used for analysis (details in the subsection on phenotypic data extraction).

Mature inflorescences and stems from a single pot, containing either 1 plant or 4 plants depending on the treatment, constituted a single sample for imaging. Thus, all the stems from a single pot were flattened and scanned at 300 dpi using a flatbed scanner. In total, 7,132 images were taken and saved as PNG files. The full experimental design behind the experiment is described by Biernaskie et al. [34]. For our study, we trained a model on a subset of these images, performed inference on the full dataset, and created a phenotyping pipeline to extract morphometric traits from the outputs. This workflow is shown in Fig. 1.

Image annotation using the GIMP image annotator

Effective training of instance segmentation networks often requires that objects are well annotated within each image. Typically, image annotation involves drawing a polygon along the boundaries of each object. However, our image collection tends to feature clear boundaries between the plant biomass and the background, which can be exploited using classical computer vision algorithms.

We evaluated various existing segmentation annotation tools. Many popular tools, such as VGG Image Annotator (VIA) [35], rely on manual polygon annotation without offering any computer vision assistance. While tools like CVAT [36] provide integration with deep learning models like the segment anything model (SAM [37]) for automatic segmentation based on keypoints or bounding boxes, as well as some basic CV functionalities via OpenCV, we found that these approaches yielded poor segmentation results with imprecise borders for our specific dataset. Therefore, we explored the tools available in the GNU Image Manipulation Program (GIMP) and found that its suite of selection tools could quickly create high-quality segmentations. To further streamline this process, we developed a custom plug-in named the GIMP Image Annotator (GIÅ), which allows users to save selected regions as mask images, along with metadata such as object classes, in a designated folder.

GIÅ features an intuitive graphical user interface (GUI) that integrates seamlessly with GIMP, making it a robust training data collection tool. This interactive tool applies existing computer vision algorithms that rely on colour similarity to segment specific structures, such as siliques, within images (known as Fuzzy

Select within the software package). GIÅ facilitates the annotation of well-separated, unoccluded siliques. For partially occluded siliques, typically only the ambiguous, occluded region needs hand-drawn boundaries to accurately define the object. This tool significantly reduces the effort required for precise annotation of each object within an image. Using GIÅ, we produced a dataset of 55 images with 2,940 annotated siliques.

We employed an amodal instance segmentation approach, which involves segmenting both visible and occluded pixels of an object [38]. By saving masks separately for each instance, the approach allows overlapping regions to be attributed to multiple instances. This approach is particularly useful for scenarios where objects overlap and partially occlude each other, providing a more comprehensive representation of the scene.

Model development

Dataset train and test split

To evaluate the model's outputs on unseen samples, we randomly split the annotated dataset into an 80% training set and a 20% testing set, resulting in 44 training images with 2,282 individual silique annotations and 11 testing images with 652 silique annotations.

Cascade Mask R-CNN model training

To capture morphometric data, we used an instance segmentation deep learning model, Cascade Mask R-CNN [23, 29], an improvement over Mask R-CNN by utilising a cascading architecture. This architecture introduces a series of cascaded region of interest (RoI) heads, each trained on progressively harder examples, by using larger IoU thresholds for bounding box detection, leading to improved performance, especially for difficult objects.

The model is capable of both detecting each object (silique) and outputting a mask that denotes the pixels belonging to that specific object instance. These outputted masks form the basis for our phenotype data. Underlying Cascade Mask R-CNN, we used a ResNetX 1.6GF backbone [39] with a feature pyramid network. We initialised the network with weights pretrained on COCO (common objects in context) [40]. COCO is a commonly used, high-quality dataset, and models trained on COCO have already learned strong lower-level object detection features that form a strong model initialisation point. We initialise the weights without any parameter freezing to allow the model to adapt to the largely different domain of our dataset.

The model was then trained on the training set for 36 epochs, using stochastic gradient descent with a learning rate of 0.01. We found that these parameters allowed the model to converge well on our dataset without overfitting. We included an initial linear learning rate warm-up for the first 100 iterations, starting at 0.00001 and ending at the learning rate of 0.01. This was chosen to stop the sharp initial updates from the initial COCO weights, which can lead to instability [41]. The learning rate was set to decay by a factor of 0.1 at epochs 24 and 33 to allow for fine-tuning later in the training. Stochastic gradient descent also included a momentum value of 0.9 and a weight decay value of 0.0001. The model was trained for 36 epochs with 2x NVIDIA A100 48 GB GPUs and batch size 1, resulting 792 training iterations, taking 1 hour and 21 minutes.

The images within our dataset were too large to fit on the GPU at native resolution during training, so to reduce memory usage while preserving silique detail, images were scaled to 4,590 × 3,240, 90% of the original size. As the size of the siliques within these images already very small, we found downsampling the size

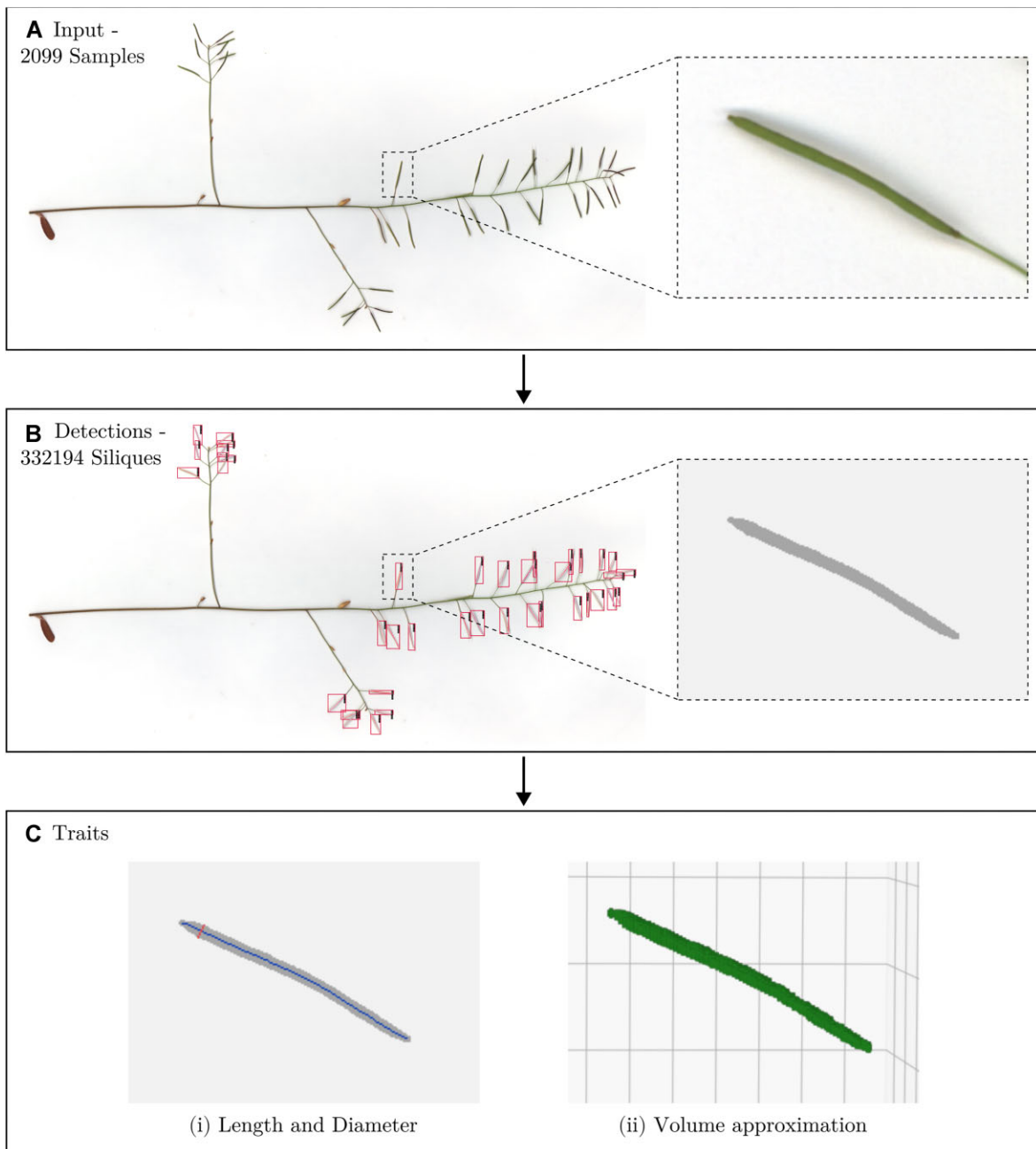


Figure 1: Workflow. (A) An example input image alongside a zoomed cutout showing an example siliqua. (B) The output from the network on example input, alongside an example generated mask. (C) Derived traits from the generated mask; left shows the length (blue, spinal line) and the maximum diameter (red, lateral line) overlaid on the mask, and right shows a 3D visualisation of the siliqua, based on the diameter values along the length of the siliqua. In total, 2,099 samples representing 362 lines were processed by the pipeline, outputting 332,194 independently measured siliques.

of the images further resulted in poor segmentation results in the higher precision $AP_{0.75}$ metric. We inferred that the higher resolution was therefore necessary to achieve the quality of segmentation and fine detail required in the ultimate phenotype data for relation to the genome. Moreover, while alternative solutions, such as dividing images into smaller patches, could preserve resolution and detail for siliques, they introduce complexity in splitting and reassembling, particularly for future studies focusing on positional branch-level traits. Thus, we chose a higher resolution (compared to the popular setting of less than $1,000 \times 1,000$) that balances image quality with the capabilities of our available GPU resources.

Model evaluation for instance segmentation

We evaluated the model's success in 2 independent ways: first, by comparing its outputs on previously unseen samples with manually annotated ground truth data and, second, by evaluating the model's phenotypic data using QTL analysis to identify phenotype–genotype correlations.

Evaluating the model on the unseen test data requires direct comparison of the model's predicted outputs against the hand-collected annotation data. Predicted objects (represented by a bounding box or mask) are considered correctly identified if the IoU between the prediction and the ground truth exceeds a certain threshold. Model detection is categorised as either true

Table 1: Individual silique traits—names, abbreviations, and descriptions of each of the traits derived from a single silique

Trait	Abbr.	Description
Silique area	SA	The raw binary mask output from the network, calculated as the number of segmented pixels, converted into mm^2 . Illustrated as the grey region in Fig. 1B.
Silique length	SL	Skeletonisation (using Zhang and Suen [42] thinning algorithm) of the binary mask to define a median line representing the primary structure. This algorithm erodes edges of a binary mask, until only a single, spinal line of pixels remains. The Euclidean distance of the line of pixels approximates the length of the silique, converted in mm . Illustrated as a blue spinal line of the pod in Fig. 1C(i).
Silique diameter	SD	Calculation of the Euclidean distance transform of the object, finding the largest value along the skeletonised length line and taking double that value to be an approximation of diameter (or the largest distance to the nearest edge among all points along the length), converted into mm . Illustrated as a red lateral line in Fig. 1C(i).
Silique volume	SV	Approximation of the silique volume, based on calculating the volume of a series of 1-voxel-high cylinders/disks at each point along the spinal line, with the sum approximating the full silique volume, converted into mm^3 . A 3D illustration in Fig. 1C(ii).

positives (*TP*) or false positives (*FP*) depending on whether they align with ground-truth annotations for a given IoU threshold. Any annotations the model failed to detect are considered false negatives (*FN*). We calculate precision (proportion of correctly identified objects *TP* against all detected objects $TP + FP$) and recall (proportion of correctly identified objects *TP* to the total number of objects within the sample $TP + FN$) across all detected siliques for both object detection and instance segmentation. We report average precision (*AP*) and average recall (*AR*), representing the average precision or recall values across IoU thresholds of 0.5, 0.75, and the range from 0.5 to 0.95 at 0.05 increments. These metrics are standard in evaluating object detection models and allowed us to check if the model detected all objects within the image, without littering the image with spurious detections.

Phenotypic data extraction

Silique morphological traits

The outputted masks from the instance segmentation model are used to calculate several metrics representing silique morphology. The calculated traits are outlined in Table 1 and visualised in Fig. 1. Silique length represents the spinal line, between the pedicel–silique junction to the tip of the silique (remains of the stigma), while the silique diameter represents the maximum width. The projected area (referred to hereafter as “area”) is the area of the mask outputted by the model.

The silique volume is based on the Euclidean distance transform values computed from the silique mask. For each point along the central length line (spinal line) of the silique, the distance transform is calculated as the distance to the nearest border of the mask along a line orthogonal to the length line. These values serve as the radius, r , of a series of 1-voxel-high cylinders, with volume $\pi r^2 h$, where $h = 1$, centred on each point of the spinal line. Summing the volumes of all cylinders along the spinal line approximates the total silique volume, assuming a cylinder-like morphology where each cross section is circular with a radius equal to the distance transform.

Phenotypic metrics at the sample level

Prior to QTL analysis, we employed a 2-level filtering process to ensure that the extracted phenotypic data were biologically relevant. The first level was applied to the pots (i.e., samples), excluding aberrant samples unrelated to the quality of the model predictions. The second level was applied to the siliques, remov-

ing model-detected siliques based on the quality of the generated masks to improve the accuracy of the extracted traits.

The pot-level filtering was necessary because, in the original experiment that produced the image data, some individual plants exhibited unusually large numbers of aborted fruits. The cause was unclear, as no specific genotype or growing location followed this pattern. Therefore, adhering to the original experimental protocol [34], any individual plant with more than 2 branches showing a high number of aborting fruits had all replicate samples of that same genotype removed. In total, 702 samples were therefore removed from the QTL analysis, resulting in a dataset of 2,099 samples representing 2 treatments of 362 genetically distinct MAGIC lines.

Next, we programmatically examined the structure of each silique mask and filtered out masks with multiple connected components, ensuring each mask represented a single connected shape. For length calculations, we verified that the generated skeletons were continuous lines without significant spurs (i.e., exceeding 5 pixels in length). This ensured that the masks conformed to the expected cylindrical morphology and filtered out aberrant masks, for example, masks covering more than a single detected silique.

We extracted silique morphology features from the filtered silique masks, including length, diameter, volume, and area (see Fig. 1). For each sample, we calculated various aggregate silique morphological trait statistics, including mean, percentiles, and relative standard deviation (*RSD*, the ratio of standard deviation to the mean). *RSD* provides a measure of the overall spread of each silique morphology trait within a sample, a potentially desirable trait for breeding. We explored different percentiles (from 95th to 5th) of fruit morphology traits, representing the best and worst fruits produced.

QTL analysis

To assess the impact of competition on silique morphology traits at the population level, we conducted a paired-sample t-test on sample-level data, disregarding genotypic variation. This analysis allowed us to evaluate the utility of the phenotype data before examining genetic differences.

For the QTL analysis, we utilised the sample-level silique metrics (including mean, percentiles, and *RSD*) across all samples of each treatment and genotype. We calculated heritability estimates of the aggregate metrics and found the P_{95} phenotypes had the highest heritability estimate in all the silique traits for both

Table 2: Detection and segmentation AP and AR results—IoU threshold values denoted in subscripts

Metric	AP _{0.5:0.95}	AP _{0.5}	AP _{0.75}	AR _{0.5:0.95}
Detection	88.0	95.8	94.5	90.7
Segmentation	55.9	94.2	66.0	61.5

treatments. The 95th percentile phenotype also represents the best a plant or group of plants can produce, which is a phenotype of interest in breeding applications. We therefore focused primarily on this phenotype for the QTL analysis. QTL mappings were performed using linear mixed models, with the kinship matrix calculated using the leave-one-chromosome-out (LOCO) method, unless otherwise specified. We estimated the parental effects of each of the founder alleles at putative QTL by fitting a single-QTL model at the QTL position and outputting the estimated regression coefficients of each of the founder alleles. The genetic map used was constructed by Kover et al. [33], consisting of 1,250 markers across the 5 chromosomes. There were 274, 210, 248, 228, and 290 markers for each chromosome, respectively, with an average spacing between markers of 0.4 cM and a maximum gap of 4.5 cM. To account for potential batch effects in the original experiment, we included batch as a covariate in the analysis. The Bayes credible interval with a 0.95 probability was calculated to estimate the probable region containing each QTL. Confidence thresholds were calculated using permutation tests where number of permutations = 1,000 and $p < 0.05$.

To identify genes related to silique development, we utilised the GOSLIM ontology system [43] to collect Gene Ontology terms related to fruit development, organ development, and organ morphogenesis (GO:001015, GO:009402, GO:1905392—Set 1) and seed development (GO:0048316—Set 2), including their child terms. With these ontological terms, we searched the TAIR9 database [7] for genes with known effects on fruit phenotypes. The specific ontologies used are listed in [Supplementary Table S3](#). Genes within the Bayes credible interval were cross-referenced with our detected QTL, and the gene nearest to the QTL peak was identified.

Software

Models were created and trained using Python library `MMDetection` v3.3 [44], built upon `PyTorch` v2.2.1 [45]. Model phenotyping pipeline was built upon Python libraries `scikit-image` v0.22.0 [46] and `opencv` v4.9.0 [47]. QTL analysis was performed using `R/qt12` v0.36 [48], and genomic data were provided by `R/atMAGIC` v0.1.0 [49], which in turn was retrieved from R. Mott's and P. Kover's server (<http://mtweb.cs.ucl.ac.uk/mus/www/magic/>).

Results

Model performance on silique detection and segmentation

We initially evaluate the direct outputs of the model, which consist of predicted bounding boxes and associated masks for the detected siliques for given thresholds, reporting precision and recall metrics for both silique detection and segmentation, as outlined in the method section on model evaluation.

Table 2 shows the detection and segmentation AP and AR results from the test split of the annotated dataset. For detection, the model achieves an average precision of over 95% for siliques when an IoU threshold of 0.75 is used on the detected boxes. We observe similar performance for the average of IoU thresholds

ranging from 0.50 to 0.95, achieving a precision of 88% and recall of 91%. For segmentation, the model achieves an average precision of 66% for siliques when an IoU threshold of 0.75 is used on the segmentation mask. The model also achieves a segmentation IoU score of over 0.50 in 94% of detected siliques. The model successfully detects and creates quality segmentations of most siliques within the unseen test set. We show example qualitative results in Fig. 2, where we see consistent, accurate segmentation of each silique, even in situations of partial occlusion.

We also note the amodal instance segmentation performance of our model. We found that in situations of occlusion, our model was typically capable of segmenting both siliques amodally, even more so if the siliques were perpendicular to each other. We believe this is due to the visual clues of tips and bases of the siliques informing the detector and the fact that perpendicular occlusion fares better is due to the intersection area of the 2 objects being reduced, giving more information to the model to consider them as 2 distinct instances. As long as neighbouring bounding boxes don't largely overlap, so as not to get suppressed by non-max suppression, and there are enough visual clues in the image, the model is capable of handling amodal segmentation in situations of occlusion.

Treatment effects on the distributions of phenotype data

Given the model's strong performance in detection and segmentation, we then evaluated the plant-level phenotype data extracted from the model predictions to assess their ability to detect treatment effects, specifically the impact of competition on silique morphology within the MAGIC population.

Figure 3 shows the distribution of the 95th percentile phenotypes of the 4 silique traits across the population of 362 lines for each treatment. For the length, volume, and area phenotypes, we see a normal-like distribution for both treatments, with a reduction in the overall values for plants grown in competition. This indicates that for the population, the group growth treatment reduces silique length and volume. Diameter has much less spread than the other phenotypes, but this is, at least partially, the result of the imaging process. The diameter values are extremely small (approximately 5–10 pixels across), so any minor variation will be binned into the available pixels, and therefore metrics such as P_{95} that take a value at a specific location in the distribution will result in selecting the same value for many samples, resulting in a metric with less resolution. However, the P_{95} metric still best represents the maximum potential of each plant; therefore, we chose to move forward using it for our analysis.

We performed a paired-sample t-test using the sample-level summary statistics, specifically the 95th percentile of silique morphology traits, for each line and treatment. For the phenotypic metrics (95th percentile, mean, and RSD) of length, diameter, and volume, we found a statistically significant difference between the 2 treatments at $p < 0.05$ across the population, except for the 95th percentile of diameter. The p values for each phenotypic metric are provided in [Supplementary Table S2](#). This indicates a statistically significant difference in most of these phenotypic metrics between the 2 treatment distributions at the population level.

QTL mapping, parental effects, and alignment of identified QTL to known genes

While the response of the entire population to competition was clear and significant, there was extensive phenotypic variation among the genotypes for all traits. We therefore used QTL



Figure 2: Example qualitative results on unseen images from the full *Arabidopsis thaliana* dataset. Red boxes represent detections of siliques. The red mask within each box detection represents the detected mask for that detected silique. Percent values tied to each detection represent the confidence score. Upper row shows full stem material; lower row shows zoomed regions.

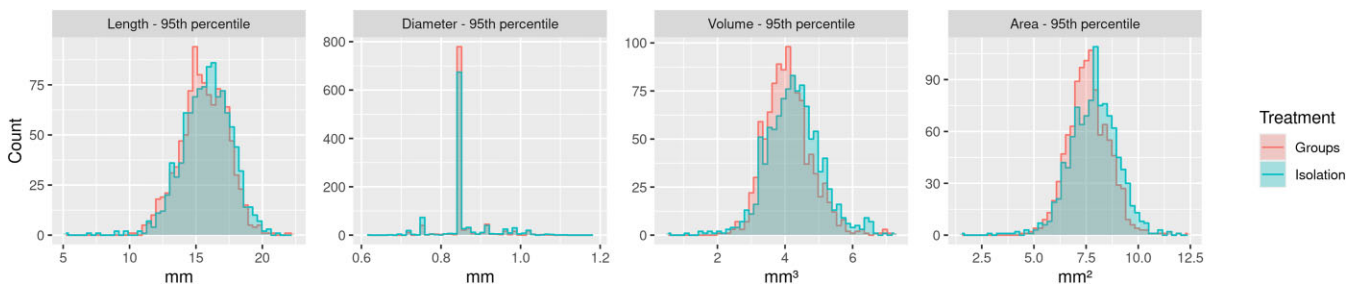


Figure 3: The 95th percentile (P_{95}) phenotype distributions—histograms of the distribution of the P_{95} for length, diameter, volume, and area, with each treatment overlaid. Group treatment is represented in red, and isolation treatment is represented in blue. Data from the filtered population of the 362 lines. Note: All figures have free axis scales.

analysis to test whether the deep learning–derived phenotypes were suitable for genetic analysis.

Figure 4 shows the QTL mapping of length, diameter, volume, and area P_{95} phenotypes under both treatments, with the calculated significance thresholds represented as a horizontal dashed

line. Genes with relevant ontologies related to fruit development, organ development, and organ morphogenesis are shown in grey (complete list is available in supplementary data). A complete list of detected QTL for the 95th percentile phenotypes is provided in Table 3 alongside 95% Bayes credible intervals.

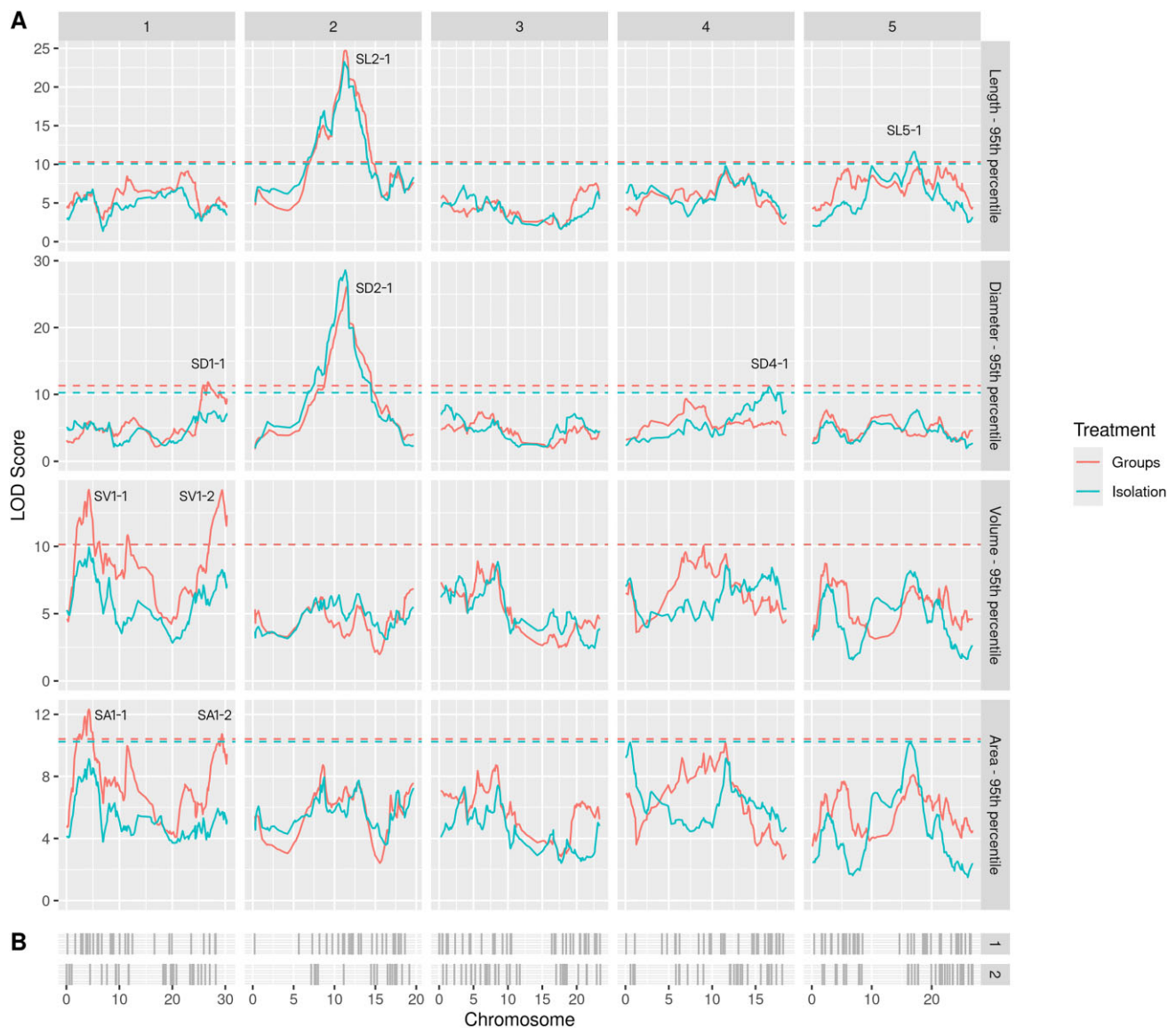


Figure 4: (A) QTL mapping for 95th percentile (P_{95}) phenotypes for length, diameter, volume, and area. Horizontal dashed lines represent the significance threshold calculated using permutation tests (number of permutations = 1,000, $p < 0.05$). Detected QTL designations are written and positioned next to associated peaks. Group treatment is represented in red and isolation in blue. (B) Genes with relevant ontologies related to fruit development, organ development, and organ morphogenesis (set 1) and seed development (set 2) shown as vertical lines indicating the starting position of the gene. Set number is shown in facet title on the right side. Data are from the filtered population of 362 lines.

Table 3: List of all significant 95th percentile (P_{95}) phenotype QTL for all traits. Bayes credible interval calculated with probability = 0.95. Nearest gene is calculated by the closest gene to peak within the Bayes credible interval. Data from the filtered population of the 362 lines.

Trait	QTL	Treatment	Chromosome	LOD	Interval (Mb)	Peak (Mb)	Nearest Gene
Silique length	SL2-1	Isolation	2	23.25	11.143–11.588	11.200	ER
	SL2-1	Groups	2	24.73	11.143–11.541	11.324	ER
	SL5-1	Isolation	5	11.65	16.186–17.676	17.169	SCP2
Silique diameter	SD2-1	Isolation	2	28.60	10.793–11.480	11.324	ER
	SD2-1	Groups	2	26.06	11.324–11.588	11.541	PGY1
	SD1-1	Groups	1	11.85	25.735–28.976	26.682	SERK1
	SD4-1	Isolation	4	11.20	15.787–17.804	16.575	ADC2
Silique volume	SV1-1	Groups	1	14.22	03.398–04.569	04.249	ESR1
	SV1-2	Groups	1	14.19	28.405–29.814	29.389	×
Silique area	SA1-1	Groups	1	12.33	02.548–04.927	04.249	ESR1
	SA1-2	Groups	1	10.73	27.852–29.893	29.389	BZR1

For the silique length phenotype, strong QTL signals were detected on chromosome 2 with significant peaks, ranging from 11.143 to 11.588 Mb for both treatments, suggesting a treatment-invariant QTL (designated as SL2-1). The closest gene with relevant ontology to the QTL peak is *ER*, or *erecta*. The *erecta* mutation, at 11.208 to 11.215 Mb, produces short fat club-like siliques [50]—a decreasing effect on length and an increasing effect on diameter. To determine the parental source of the variation at this locus for silique length and silique width, we fitted a single-QTL model using a linear mixed model and calculated the parental effect of each of the founder alleles. Figure 5A shows the effect of the founder alleles on the length P_{95} phenotype. The LER (Landsberg Erecta) founder allele provides the main decreasing effect for SL2-1 with a smaller contribution from Can-0 about the QTL peak. The *ER* gene (8.4 kbs away from the peak) is shown as a dashed black line. Figure 5B illustrates the opposite, with an increasing effect for diameter (SD2-1) for the same founder allele.

Additionally, there are 2 significant peaks on chromosome 1 for both the volume and area P_{95} phenotypes within the region of 3.101 – 3.398 Mb and peaks at 4.249 Mb. These loci show a significant association for the group treatment only, suggesting the P_{95} phenotypes of these traits are treatment-sensitive. The first locus related to volume, SV1-1 with a peak at 4.249 Mb, had the closest gene *ESR1*. The second locus, SV1-2 with a peak at 29.389 Mb and a Bayes credible interval of 28.579 – 29.814 Mb, had no genes in our ontology list near it, suggesting a novel locus. It is worth noting, however, that the highly related area phenotype also detected a QTL at the same peak as SV1-2 but with a wider credible interval and found the nearby gene *BZR1* from our ontology list. Figure 6 shows the significant QTL associated with the volume P_{95} phenotype.

Discussion

This report describes the successful application of an instance segmentation model based on Cascade Mask R-CNN, along with a pipeline for quantitative phenotyping of *A. thaliana* siliques at scale. Our deep learning approach enabled the extraction of morphological traits, which, when combined with genetic data, led to the identification of high-confidence QTL. As far as we know, this is the first time that fruit morphological data, generated by a deep learning system, have been successfully linked to genomic variation. Many previous studies have identified QTL influencing fruit morphology, in a wide variety of crops [12, 13, 51–54]. However, these tend to involve manual or simple classical CV-based measurements of isolated fruits. For example, Gegas et al. [51] used a high-throughput imaging approach with harvested grain (botanically, a single-seeded fruit) to dissect the genetic architecture of kernel variation in bread wheat and its close relatives.

Where fruits are easily harvested and handled (e.g., tomatoes, squashes, and bananas, amongst others), it is possible to phenotype manually and at a sufficient scale to undertake genetic analyses. This is even possible with minor crops such as Chinese bayberry, *Myrica rubra* [55]. However, many types of fruit, particularly from undomesticated species such as the model experimental species *Arabidopsis*, are inherently difficult to harvest due to the inherent fragility of the ripe fruits. Therefore, whole stems carrying the siliques were harvested to minimise shattering, and the samples were then pressed before imaging to maximise the accuracy of trait measurements. Alternative methods, such as multiview RGB imaging of whole plants followed by structure-from-motion 3-D modelling [56, 57], were considered too slow to enable data capture from the entire population. Likewise, computed to-

mography scanning of plants with long thin stems that needed careful immobilisation was not feasible at scale.

While it is technically possible to extract morphometric data manually from physical samples, it is difficult to complete this sort of operation at scale and within a time frame where such samples typically retain their structural coherence. Thus, a high-resolution image dataset of experimental samples was created and stored within days of harvest, thereby immortalising the information for later extraction and analysis. An alternative to imaging would be to employ a relatively large number of trained staff, but this is often not financially feasible even for crops. One person, however, can achieve imaging of large sample numbers at minimal cost as it requires ubiquitous low-cost hardware. Phenotype data such as fruit position, angle, and colour exist within the images used within this study but would need pipelines developed for high-throughput extraction. This potential of further data extraction is only possible due to the imaging process and tools that computer vision provides. As computer vision tools progress, more phenotype data will be able to be extracted from these populations without the cost of setting up new experiments.

The simplicity and speed of 2D imaging and flat-bed scanning allowed us to critically evaluate the use of deep learning phenotyping on a large population that was designed to address fundamental biological questions [34]. However, due to the inherent complexity of the material being analysed and the relatively small training dataset, the deep learning model generated unwanted artefacts. To address this, we employed basic filtering techniques. We exclude all masks that return more than 1 connected component, which typically occurs when a bounding box has enclosed more than 1 silique or parts of 2 or more neighbouring siliques. In such cases, the mask generation branch of the network attempts to include those pixels in the detection. We also excluded masks that have more than 1 skeletal line (after pruning smaller skeletal branches), as this, like multiple connected components, only typically occurs when regions of neighbouring siliques have been included in the mask output. Overall, these artefacts occur in 7.3% of detections on (median) average per image. Appropriate filtering ensured that the final dataset was suitable for downstream genetic analysis.

Collecting high-resolution silique morphological data at this scale is not typically possible using manual measurements. Typical manual QTL approaches tend to estimate genetic potential from a relatively small subset of hand-collected fruits [5, 58–60]. Our approach allowed us to analyse the entire silique output from each sample, generating various statistical measures, including mean, percentiles, and relative standard deviation (RSD). Notably, the 95th percentile of each silique trait gave the strongest overall heritability (ranging from 0.27–0.43). This was confirmed by QTL mapping of the mean versus P_{95} metrics (Fig. 4 and Supplementary Fig. S3): while the overall landscape was similar, the P_{95} phenotype gave higher confidence scores. As our system generates morphological data on very large numbers of individually detected fruit with defined criteria for including or excluding defined components of data, it removes subjective bias from the analysis, allowing full exploration of different facets of fruit quality.

Uniformity of fruit size within a cultivar is another desirable trait depending on the crop. Thus, millers prefer uniform grain size in many small-grain cereals for processing reasons [61], whereas “standard” scales for uniformity are defined in breeding many fruits. Notably in tomato, uniformity has been reported as highest in small-fruited varieties [62], but particular alleles of quantitative trait locus *qMULTIPLE INFLORESCENCE BRANCH 2* (*qMIB2*), a *SPATULA*-like gene, can contribute to fruit size stability

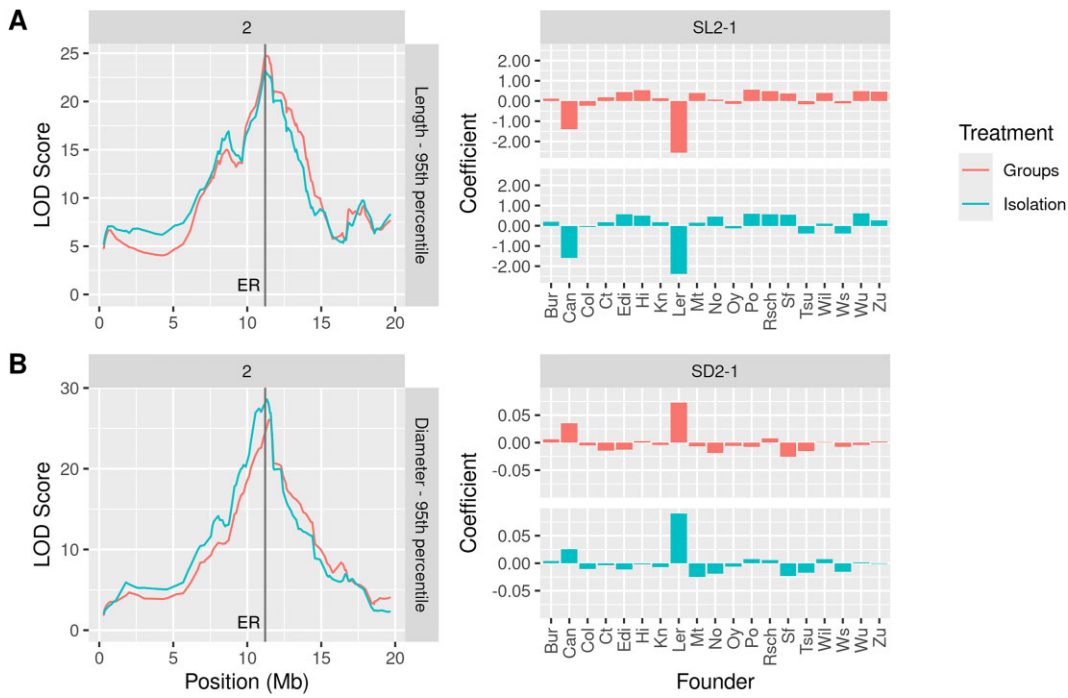


Figure 5: QTL mapping outputs and parental effect coefficients for each of the founder alleles for (A) length P₉₅ and (B) diameter P₉₅ phenotypes comparing group and isolation treatments. QTL analysis outputs (left) display the peak positions of QTL on chromosome 2, while the corresponding QTL names are shown above the parental effect coefficient bar plots (right). The position of the gene ER is indicated by a solid vertical line. Data are from the filtered population of 362 lines.

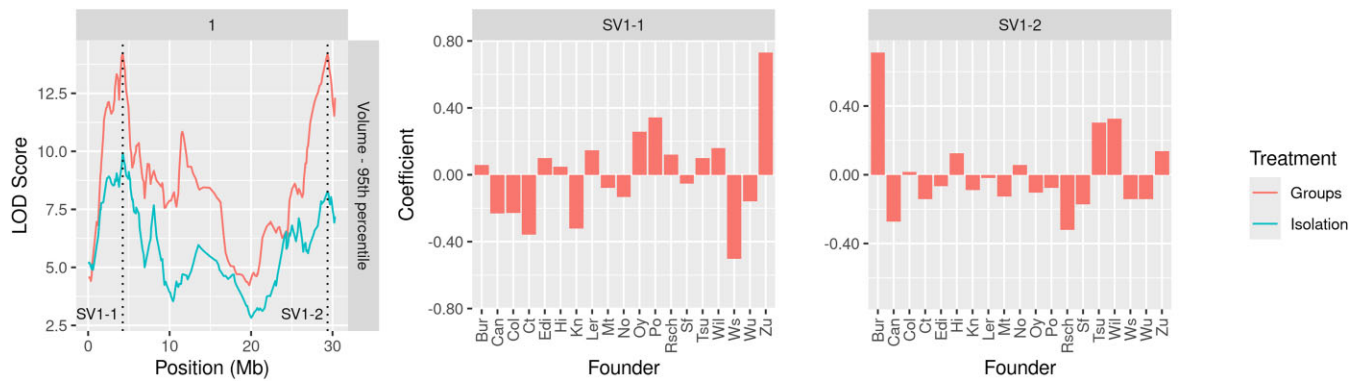


Figure 6: QTL mapping outputs and parental effect coefficients for the significant QTL associated with the volume P₉₅ phenotype, comparing group and isolation treatments. The peak positions of detected QTL on chromosome 1 are indicated by vertical dashed lines. Parental effect coefficient bar plots for the 2 significant QTLs in the group treatment are provided. No significant QTLs were found for the isolation treatment at these positions. Data are from the filtered population of 362 lines.

in larger fruited strains [63]. To assess uniformity, we used RSD, but this had low heritability and did not reveal any significant QTL, despite the presence of a significant treatment effect that affected the QTL mapping of RSD, as reflected in LOD (logarithm of the odds) scores in certain genomic regions (Supplementary Fig. S4). Variation in silique morphology is perhaps influenced more by development than genetics or treatment as the inflorescence stem is determinate, with later flowers producing shorter siliques [64].

Conventional wisdom suggests that small training datasets, like ours, typically lead to overfitting in deep learning models. However, despite being trained on only 44 images and around 2,000 annotated siliques, our model generalised well to the remaining dataset. This is likely due to the uniform background and the high number of single-class objects per image. The abundance of objects per image also makes the loss value at each train-

ing step more informative, which is crucial given the variability in siliques' shape, size, openness, and fertility within the same plant. Our image and data acquisition strategy enables effective fine-tuning of a COCO pretrained model for our specific task. We chose to use a smaller backbone for the model, ultimately deciding to use RegNetX 1.6GF, a model with approximately half the parameter count of ResNet-50. The RegNet family of CNNs is designed to have more efficient-sized depth and width to reduce redundancy in parameters. We found that using a smaller backbone resulted in equal or improved detection and segmentation results compared to its larger parameter-count siblings. This implies the larger models overfit easily on this constrained dataset, and the regularisation effect of reducing the model size benefits this task. Reducing model size also allowed more memory to fit the large images during training.

Conclusion

This article has presented a deep learning pipeline capable of phenotyping morphometric traits of siliques of *A. thaliana*. It demonstrates an innovative application of deep learning models as a phenotyping tool, focusing on the capability of relating phenotypic variation to the genome. In particular, our work expands the knowledge of what is possible with automatically collected phenotype data through deep learning, revealing that high-confidence QTL can be identified on the basis of morphology phenotypes from a deep learning model. The systematic verification results demonstrate that our deep learning phenotyping data have the quality and accuracy required for genetic analysis of continuous morphometric characteristics of attached plant organs, in the context of high-precision quantitative phenotyping.

The phenotypic and genotypic datasets, as well as the code and models developed in this study, are fully available at the open data repository. We ensure reproducibility by adhering to FAIR principles and providing step-by-step documentation, allowing for ease of replication by the scientific community. In keeping with the principles of open science, our pipeline not only provides immediate insights into silique morphology but also acts as a foundational tool for future open-source development. Researchers working in different phenomics environments can modify our pipeline, allowing for extended applications beyond *A. thaliana*. The scalability of our pipeline underscores its potential for application in other large-scale phenotyping studies, such as those in crop breeding programs or ecological field studies. As deep learning continues to evolve, models like ours are well positioned to be further reinforced to handle the increasing data complexity and volume inherent to high-throughput plant phenotyping.

Our phenotyping approach has direct implications for precision agriculture and breeding, allowing for the high-resolution quantification of key agronomic traits. Adaptation to other plant species with similar architectures, especially Brassica seed crops, should be relatively straightforward. While our model generalizes well on this relatively focused dataset, future research could explore more diverse datasets or augment existing data to improve robustness. Additionally, these image collections contain a wealth of additional interesting information, so expanding the pipeline to detect other plant structures (such as petioles, stems, and flowers) and features (e.g., branch angles, numbers, and organ positions) could enhance its applicability to broader plant phenomics studies. We hope this extensive image dataset and associated genetic tools will motivate others to reuse these data.

Availability of Source Code

MorphPod: Deep learning phenotyping of Arabidopsis fruit morphology

- <https://github.com/kieranatkins/silique-detector/> [65]
- Operating system: Platform independent
- Programming language: Python, R
- Other requirements: see public environment file released under GNU GPL v3
- bio.tools: [biotools:morphpod](https://bio.tools/morphpod)
- RRID: [MorphPod \(RRID:SCR_026174\)](https://www.ebi.ac.uk/rrid/SCR_026174)
- This code has also been archived in Software Heritage [66].

GIMP Image Annotator (GIÅ): a lightweight GIMP plug-in for computer vision-assisted image annotation

- <https://github.com/kieranatkins/gimp-image-annotator/> [67]
- Operating system: Platform independent

- Programming language: Python
- Other requirements: GIMP
- released under GNU GPL v3
- bio.tools: [biotools:gimp_image_annotator](https://bio.tools/gimp_image_annotator)
- RRID: [gimp_image_annotator \(RRID:SCR_026175\)](https://www.ebi.ac.uk/rrid/SCR_026175)
- Workflow hub: [10.48546/workflowhub.workflow.1229.1](https://www.workflowhub.eu/workflows/10.48546/workflowhub.workflow.1229.1) [68]

Additional Files

Supplementary Fig. S1. Data Collection GUI: a screenshot of the GUI running alongside GIMP to manage the collection of training data.

Supplementary Fig. S2. Heritability estimations: estimations of heritability of each of the explored phenotypes, for each silique trait. Calculated via a linear mixed model using an overall kinship matrix, with batch introduced as a covariate. Each silique trait is shown with each phenotype metric derived from the data, including the mean (μ), P_{95} , and RSD. Data are from the filtered population of the 362 lines.

Supplementary Fig. S3. (A) QTL mapping for mean phenotypes for length, diameter, volume, and area. Horizontal dashed lines represent the significance threshold calculated using permutation tests (number of permutations = 1,000, $p < 0.05$). (B) Genes with relevant ontologies related to fruit development, organ development, and organ morphogenesis (set 1) and seed development (set 2) are shown as vertical lines indicating the starting positions of the gene. Group treatment is represented in red and isolation in blue. Data are from the filtered population of the 362 lines.

Supplementary Fig. S4. (A) QTL mapping for relative standard deviation (RSD) phenotypes for length, diameter, volume, and area. Horizontal dashed lines represent the significance threshold calculated using permutation tests (number of permutations = 1,000, $p < 0.05$). (B) Genes with relevant ontologies related to fruit development, organ development, and organ morphogenesis (set 1) and seed development (set 2) are shown as vertical lines indicating the starting position of the gene. Group treatment is represented in red and isolation in blue. Data are from the filtered population of the 362 lines.

Supplementary Table S1. Phenotypes: descriptions of calculated phenotypes for each genotype.

Supplementary Table S2. Paired-sample t-test results for the population-level analysis of the effect of competition for the P_{95} phenotypes, comparing the distributions of both treatments. Data from the filtered population of the 362 lines.

Supplementary Table S3. Gene ontologies used to subset the list of genes.

Abbreviations

AP: average precision; AR: average recall; CNN: convolutional neural network; COCO: common objects in context; IoU: intersection-over-union (also Jaccard Index); LOD: logarithm of the odds; MAGIC: multiparent advanced generation intercross; QTL: quantitative trait locus; R-CNN: region-based convolutional neural network; RIL: recombinant inbred line; RSD: relative standard deviation (also coefficient of variation).

Acknowledgements

We thank Dr. Jay Biernaskie from University of Oxford for allowing us to use the experimental dataset prior to publication. Thanks to Ralph Lidner and Al-Fida M. Ahmed for their assistance with data collection.

Author Contributions

K.A. annotated the dataset, trained the model, created the phenotyping pipeline, and performed the analyses. K.A. wrote the manuscript. G.G.M. designed and performed the original experiment and collected the image dataset. A.L. oversaw the QTL and gene ontology analyses. J.H.D. and C.L. obtained the funding and supervised the project. All authors proofread and agreed on the manuscript.

Funding

We acknowledge the support of the EPPN2020 grant (No. 731013), the National Capability in Plant Phenotyping (BBS/E/W/0012844A), the Miscanthus AI project (EPSRC EP/Y005430/1), the Supercomputing Wales project, and the AccelerateAI project (partly funded by ERDF through the Welsh Government). K.A. was supported by a PhD studentship from the UKRI BBSRC FoodBioSystems Doctoral Training Partnership (DTP), grant number BB/T008776/1; G.G.M. was supported by the AberDoc scholarship from Aberystwyth University; A.L. is funded through a UK Research and Innovation Future Leaders Fellowship (MR/T043253/1).

Data Availability

The dataset for model development and validation with QTL analysis is available in Zenodo, under a CC0 license [69]. DOME-ML annotations can be found in the DOME Registry [70].

Competing Interests

The authors declare that they have no competing interests.

References

1. Marsh JI, Hu H, Gill M, et al. Crop breeding for a changing climate: integrating phenomics and genomics with bioinformatics. *Theor Appl Genet* 2021;134(6):1677–90. <https://doi.org/10.1007/s00122-021-03820-3>.
2. Cucinotta M, Di Marzo M, Guazzotti A, et al. Gynoecium size and ovule number are interconnected traits that impact seed yield. *J Exp Bot* 2020;71(9):2479–89. <https://doi.org/10.1093/jxb/eraa050>.
3. Vivian-Smith A, Luo M, Chaudhury A, et al. Fruit development is actively restricted in the absence of fertilization in Arabidopsis. *Development* 2001;128(12):2321–31. <https://doi.org/10.1242/dev.128.12.2321>.
4. Zhang L, Yang G, Liu P, et al. Genetic and correlation analysis of silique-traits in Brassica napus L. by quantitative trait locus mapping. *Theor Appl Genet* 2011;122(1):21–31. <https://doi.org/10.1007/s00122-010-1419-1>.
5. Siles L, Hassall KL, Sanchis Gritsch C, et al. Uncovering trait associations resulting in maximal seed yield in winter and spring oilseed rape. *Front Plant Sci* 2021;12:697576. <https://doi.org/10.3389/fpls.2021.697576>.
6. Ton LB, Neik TX, Batley J. The use of genetic and gene technologies in shaping modern rapeseed cultivars (Brassica napus L.). *Genes* 2020;11(10):1161. <https://doi.org/10.3390/genes1101161>.
7. Berardini TZ, Reiser L, Li D, et al. The Arabidopsis information resource: making and mining the “gold standard” annotated reference plant genome. *Genesis* 2015;53(8):474–85. <https://doi.org/10.1002/dvg.22877>.
8. Pieruschka R, Schurr U. Plant phenotyping: past, present, and future. *Plant Phenomics* 2019;2019:7507131. <https://doi.org/10.34133/2019/7507131>.
9. Yang W, Feng H, Zhang X, et al. Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. *Mol Plant* 2020;13(2):187–214. <https://doi.org/10.1016/j.molp.2020.01.008>.
10. Berry JC, Fahlgren N, Pokorny AA, et al. An automated, high-throughput method for standardizing image color profiles to improve image-based plant phenotyping. *PeerJ* 2018;6:e5727. <https://doi.org/10.7717/peerj.5727>.
11. He H, Ma X, Guan H. A calculation method of phenotypic traits of soybean pods based on image processing technology. *Ecol Inform* 2022;69:101676. <https://doi.org/10.1016/j.ecoinf.2022.101676>.
12. Diaz-Garcia L, Covarrubias-Pazarán G, Schlautman B, et al. Image-based phenotyping for identification of QTL determining fruit shape and size in American cranberry (*Vaccinium macrocarpon* L.). *PeerJ* 2018;6:e5461. <https://doi.org/10.7717/peerj.5461>.
13. Zingaretti L, Monfort A, Pérez-Enciso M. Automatic fruit morphology phenome and genetic analysis: an application in the Octoploid Strawberry. *Plant Phenomics* 2021;2021:1–14. <https://doi.org/10.34133/2021/9812910>.
14. Pound MP, Atkinson JA, Townsend AJ, et al. Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. *Gigascience* 2017;6(10):gix083. <https://doi.org/10.1093/gigascience/gix083>.
15. Gill T, Gill S, Chopra Y, et al. A comprehensive review of high throughput phenotyping and machine learning for plant stress phenotyping. *Plant Phenomics* 2022;2:3. <https://doi.org/10.1007/s43657-022-00048-z>.
16. Atila Ü, Uçar M, Akyol K, et al. Plant leaf disease classification using EfficientNet deep learning model. *Ecol Inform* 2021;61:101182. <https://doi.org/10.1016/j.ecoinf.2020.101182>.
17. Morshed MS, Ahmed S, Ahmed T, et al. Fruit quality assessment with densely connected convolutional neural network. In: 2022 12th International Conference on Electrical and Computer Engineering (ICECE). New York City, NY, United States: IEEE, 2022, 1–4. <https://doi.org/10.1109/ICECE57408.2022.10088873>.
18. Hamidinekoo A, Garzón-Martínez GA, Ghahremani M, et al. DeepPod: a convolutional neural network based quantification of fruit number in Arabidopsis. *Gigascience* 2020;9(3):giaa012. <https://doi.org/10.1093/gigascience/giaa012>.
19. Bolya D, Zhou C, Xiao F, et al. YOLACT: real-time instance segmentation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019, 9156–65. <https://doi.org/10.1109/ICCV.2019.00925>.
20. Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016, 779–88. <https://doi.org/10.1109/CVPR.2016.91>.
21. Li Y, Feng Q, Liu C, et al. MTA-YOLACT: multitask-aware network on fruit bunch identification for cherry tomato robotic harvesting. *Eur J Agron* 2023;146:126812. <https://doi.org/10.1016/j.eja.2023.126812>.
22. Carranza-García M, Torres-Mateo J, Lara-Benítez P, et al. On the performance of one-stage and two-stage object detectors in autonomous vehicles using camera data. *Remote Sensing* 2021;13(1):89. <https://doi.org/10.3390/rs13010089>.
23. He K, Gkioxari G, Dollár P, et al. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017.

24. Afzaal U, Bhattarai B, Pandeya YR, et al. An instance segmentation model for strawberry diseases based on mask R-CNN. *Sensors* 2021;21(19):6565. <https://doi.org/10.3390/s21196565>.
25. Toda Y, Okura F, Ito J, et al. Training instance segmentation neural network with synthetic datasets for crop seed phenotyping. *Commun Biol* 2020;3:173. <https://doi.org/10.1038/s42003-020-0905-5>.
26. Cheng B, Misra I, Schwing AG, et al. Masked-attention mask transformer for universal image segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, 2022, 1290–99.
27. Gu W, Bai S, Kong L. A review on 2D instance segmentation based on deep neural networks. *Image Vision Comput* 2022;120:104401. <https://doi.org/10.1016/j.imavis.2022.104401>.
28. Du R, Ma Z, Xie P, et al. PST: plant segmentation transformer for 3D point clouds of rapeseed plants at the podding stage. *ISPRS J Photogramm Remote Sens* 2023;195:380–92. <https://doi.org/10.1016/j.isprsjprs.2022.11.022>.
29. Cai Z, Vasconcelos N. Cascade R-CNN: high quality object detection and instance segmentation. *IEEE Trans Pattern Anal Mach Intell* 2021;43(5):1483–98. <https://doi.org/10.1109/TPAMI.2019.2956516>.
30. Wang X, Xuan H, Evers B, et al. High-throughput phenotyping with deep learning gives insight into the genetic architecture of flowering time in wheat. *Gigascience* 2019;8(11):giz120. <https://doi.org/10.1093/gigascience/giz120>.
31. Xie E, Wang W, Yu Z, et al. SegFormer: simple and efficient design for semantic segmentation with transformers. In: *Proceedings of the 35th International Conference on Neural Information Processing Systems*. Vancouver, Canada: Curran Associates Inc., 2021.
32. Geng Z, Lu Y, Duan L, et al. High-throughput phenotyping and deep learning to analyze dynamic panicle growth and dissect the genetic architecture of yield formation. *Crop Environ* 2024;3(1):1–11. <https://doi.org/10.1016/j.crope.2023.10.005>.
33. Kover PX, Valdar W, Trakalo J, et al. A multiparent advanced generation inter-cross to fine-map quantitative traits in Arabidopsis thaliana. *PLoS Genet* 2009;5(7):1–15. <https://doi.org/10.1371/journal.pgen.1000551>.
34. Biernaskie JM, Garzón-Martínez GA, Corke FMK, et al. Under revision, uncovering the hidden genetic basis of plant competitiveness and group productivity. *Proc R Soc B Biol Sci* 2025. in press.
35. Dutta A, Zisserman A. The VIA annotation software for images, audio and video. In: *Proceedings of the 27th ACM International Conference on Multimedia*. MM '19. New York: ACM, 2019. <https://doi.org/10.1145/3343031.3350535>.
36. Corporation CVAT.ai. Computer Vision Annotation Tool (CVAT). 2023. <https://github.com/cvat-ai/cvat>. Accessed 14 December 2024.
37. Kirillov A, Mintun E, Ravi N, et al. Segment Anything. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE, 2023, 3992–4003.
38. Li K, Malik J. Amodal instance segmentation. In: *Computer Vision – ECCV 2016*. ECCV 2016. Lecture Notes in Computer Science. Amsterdam, The Netherlands: Springer nature, 2016.
39. Radosavovic I, Kosaraju RP, Girshick R, et al. Designing network design spaces. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, 2020.
40. Lin TY, Maire M, Belongie S et al., Microsoft COCO: common objects in context. In: Fleet D, Pajdla T, Schiele B, et al. eds. *Computer Vision—ECCV 2014*. Cham: Springer International Publishing; 2014:740–55.
41. Gilmer J, Ghorbani B, Garg A, et al. A loss curvature perspective on training instability in deep learning. In: *The Tenth International Conference on Learning Representations ICLR 2022*. Virtual Event: OpenReview.net, 2021.
42. Zhang TY, Suen CY. A fast parallel algorithm for thinning digital patterns. *Commun ACM* 1984;27(3):236–39. <https://doi.org/10.1145/357994.358023>.
43. Berardini TZ, Mundodi S, Reiser L, et al. Functional annotation of the Arabidopsis genome using controlled vocabularies. *Plant Physiol* 2004;135(2):745–55. <https://doi.org/10.1104/pp.104.040071>.
44. Chen K, Wang J, Pang J et al., MMDetection: open MMLab detection toolbox and benchmark. Technical Report. 2019. <https://arxiv.org/abs/1906.07155>. Accessed 14 December 2024.
45. Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, vol. 33, Red Hook, NY, USA: Curran Associates Inc., 2019:721.
46. van der Walt S, Schönberger JL, Nunez-Iglesias J, et al. scikit-image: image processing in Python. *PeerJ* 2014;2:e453. <https://doi.org/10.7717/peerj.453>.
47. Bradski G. The OpenCV Library. *Dr Dobb's Journal of Software Tools* 2000.
48. Broman KW, Gatti DM, Simecek P, et al. R/qlt2: software for mapping quantitative trait loci with high-dimensional data and multi-parent populations. *Genetics* 2019;211(2):495–502. <https://doi.org/10.1534/genetics.118.301595>.
49. Tavares H. atMAGIC: Data package with R/qlt2 objects for the Arabidopsis MAGIC lines. 2024. R package version 0.1.0, commit 0b8e985975a0d0277cdc0bb4cc7a0dc39393070b. <https://github.com/tavareshugo/atMAGIC>. Accessed 14 December 2024.
50. Torii KU, Mitsukawa N, Oosumi T, et al. The Arabidopsis ERECTA gene encodes a putative receptor protein kinase with extracellular leucine-rich repeats. *Plant Cell* 1996;8(4):735–46. <https://doi.org/10.1105/tpc.8.4.735>.
51. Gegas VC, Nazari A, Griffiths S, et al. A genetic framework for grain size and shape variation in wheat. *Plant Cell* 2010;22(4):1046–56. <https://doi.org/10.1105/tpc.110.074153>.
52. Pan Y, Wang Y, McGregor C, et al. Genetic architecture of fruit size and shape variation in cucurbits: a comparative perspective. *Theor Appl Genet* 2020;133:1–21. <https://doi.org/10.1007/s00122-019-03481-3>.
53. Hussain Q, Zhan J, Liang H, et al. Key genes and mechanisms underlying natural variation of silique length in oilseed rape (*Brassica napus* L.) germplasm. *Crop J* 2022;10(3):617–26. <https://doi.org/10.1016/j.cj.2021.08.010>.
54. Hong Y, Zhang M, Zhu J, et al. Genome-wide association studies reveal novel loci for grain size in two-rowed barley (*Hordeum vulgare* L.). *Theor Appl Genet* 2024;137(3):58. <https://doi.org/10.1007/s00122-024-04562-8>.
55. Zhang W, Chen K, Zhang B, et al. Postharvest responses of Chinese bayberry fruit. *Postharvest Biol Technol* 2005;37(3):241–51. <https://doi.org/10.1016/j.postharvbio.2005.05.005>.
56. Ghahremani M, Williams K, Corke F, et al. Direct and accurate feature extraction from 3D point clouds of plants using RANSAC. *Comput Electron Agric* 2021;187:106240. <https://doi.org/10.1016/j.compag.2021.106240>.
57. Ghahremani M, Williams K, Corke FMK, et al. Deep segmentation of point clouds of wheat. *Front Plant Sci* 2021;12:608732. <https://doi.org/10.3389/fpls.2021.608732>.

58. Zhao X, Yu K, Pang C, et al. QTL analysis of five silique-related traits in *brassica napus* L. across multiple environments. *Front Plant Sci* 2021;12:766271. <https://doi.org/10.3389/fpls.2021.766271>.
59. Jiao Y, Zhang K, Cai G, et al. Fine mapping and candidate gene analysis of a major locus controlling ovule abortion and seed number per silique in *Brassica napus* L. *Theor Appl Genet* 2021;134(8):2517–30. <https://doi.org/10.1007/s00122-021-03839-6>.
60. Yu P, Gao Z, Hua Z. Contrasting impacts of ubiquitin over-expression on *arabidopsis* growth and development. *Plants* 2024;13(11):1485. <https://doi.org/10.3390/plants13111485>.
61. Esvelt Klos K, Yimer BA, Howarth CJ, et al. The genetic architecture of milling quality in spring oat lines of the collaborative oat research enterprise. *Foods* 2021;10(10):2479. <https://doi.org/10.3390/foods10102479>.
62. Macua JI, Lahoz I, Bozal JM, et al. Industrial quality of cherry tomato varieties in Navarre. *Acta Hort.* 2007;758:181–84. <https://doi.org/10.17660/ActaHortic.2007.758.20>.
63. Sun S, Liu Z, Wang X, et al. Genetic control of thermomorphogenesis in tomato inflorescences. *Nat Commun* 2024;15(1):1472. <https://doi.org/10.1038/s41467-024-45722-0>.
64. González-Suárez P, Walker CH, Lock T, et al. FLOWERING LOCUS T-mediated thermal signalling regulates age-dependent inflorescence development in *Arabidopsis thaliana*. *J Exp Bot* 2024;75(14):4400–14. <https://doi.org/10.1093/jxb/erae094>.
65. MorphPod. MorphPod (version 1.0). 2024. <https://github.com/kieranatkins/silique-detector/releases/tag/v1.0.0>. Accessed 14 December 2024.
66. Atkins K, Garzón Martínez GA, Lloyd A et al. Unlocking the power of AI for phenotyping fruit morphology in *Arabidopsis* (Version 1), [Computer software]. Software Heritage. 2024. <https://archive.softwareheritage.org/swh:1:snp:ae0de35e2b49271b5a8f6dc682a957ac882c2cf2;origin=https://github.com/kieranatkins/gimp-image-annotator>. Accessed 14 December 2024.
67. GIÀ. GIMP Image Annotator (GIÀ) (version 1.0). GitHub. 2024. <https://github.com/kieranatkins/gimp-image-annotator/releases/tag/v1.0.0>. Accessed 14 December 2024.
68. Atkins K. GIMP Image Annotator. WorkflowHub. 2024. <https://doi.org/10.48546/WORKFLOWHUB.WORKFLOW.1229.1>. Accessed 14 December 2024.
69. Atkins K, Garzón Martínez GA, Lloyd A, et al. Data for: unlocking the Power of AI for fruit phenotyping: a genetic validation study in *Arabidopsis* [data set]. Zenodo. 2024. <https://doi.org/10.20391/283ce324-6a96-4cc8-8168-51f48354f7cf>. Accessed 14 December 2024.
70. Atkins K, Garzón Martínez GA, Lloyd A, et al. Unlocking the power of AI for phenotyping fruit morphology in *Arabidopsis* [DOME-ML Annotations]. DOME-ML Registry. 2024. <https://registry.dome-ml.org/review/a8q3rb7qrv>. Accessed 14 December 2024.