StomataBench: Measuring Taxonomic Generalization in Stomatal Detection

Sainath Reddy Gummi
South Dakota State University

Hankui K. Zhang
South Dakota State University

Chulwoo Pack
South Dakota State University

Shyam Solanki
South Dakota State University

Young Chang
South Dakota State University
Manitoba Agriculture

Abstract

Stomatal density is widely measured in plant biology because it links cellular anatomy to photosynthesis, transpiration, water-use efficiency, and global change responses across Plantae. Recent work has addressed scalable stomatal phenotyping by training object detectors to localize and count stomata in microscopy images. Unfortunately, most existing stomatal datasets evaluate detectors within a single acquisition pipeline, leaving open whether their predictions remain reliable under out-of-distribution (OOD) deployment. This matters because detection counts are not the final scientific output. The resulting density measurement is the quantity used for biological interpretation.

In contrast, StomataBench evaluates stomatal detection as a scientific measurement problem under structured OOD shift. StomataBench pairs a single training set, OD-Train, with four held-out test sets that separate taxonomic shift, imaging-system shift, annotator shift, and combined multi-source shift. OD-Train contains 21 woody species, 20 eudicots and one magnoliid, and no monocot, gymnosperm, or fern. To make OOD evaluation biologically meaningful, held-out species are arranged on a six-level OOD ladder grounded in taxonomic distance. Beyond stomatal localization, the evaluation tests whether detector outputs preserve calibrated species-level variation in stomatal density. StomataBench provides a benchmark for testing whether stomatal detectors remain useful as measurement tools under deployment-relevant biological and imaging shifts.

1 Introduction

Stomatal density, defined as the number of microscopic gas-exchange pores per unit area of the plant surface, is a measurable plant trait that links cellular anatomy to photosynthesis, transpiration, water-use efficiency, and global change biology [1]. Estimating this trait at scale has therefore become an important target for high-throughput phenotyping pipelines based on object detectors [2, 3]. In this setting, the detector is not the final scientific output. The resulting density measurement is the quantity used for biological interpretation. A model may produce visually accurate bounding boxes while still yielding biased counts, and those biases can lead directly to biased biological conclusions. This happens because a count includes every box above one confidence threshold, so extra or duplicate boxes and missed stomata shift it even when the matched boxes fit well.

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Track on Evaluations and Datasets.

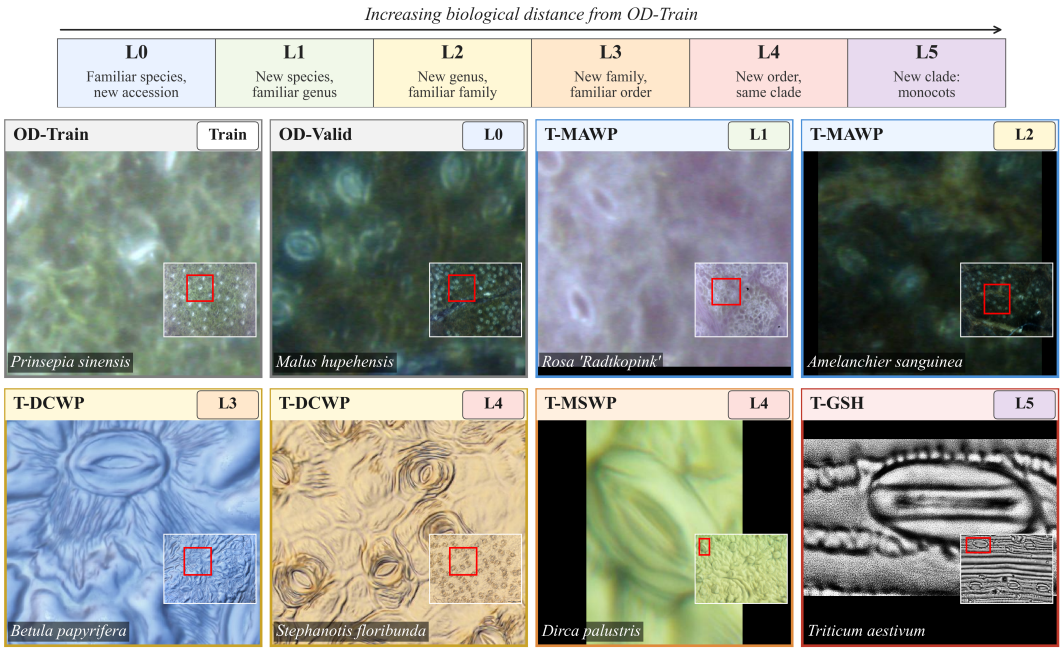


Figure 1: StomataBench at a glance. Representative micrographs from the training set (OD-Train), the OD-Valid validation set, and the four test sets, arranged along the OOD ladder (L0–L5) (Table 2).

Stomatal detection is therefore a measurement task framed as a detection task. The central question is not only whether a detector produces accurate boxes, but whether its predicted counts remain accurate and preserve biologically meaningful density variation when applied to new species, and across microscope and instrument variability, differing illumination conditions, different annotation protocols, and data from different laboratories [2, 3, 4]. Existing stomatal datasets largely evaluate performance within a single acquisition pipeline and report COCO average precision (AP), which can hide biologically important failure modes in downstream density estimation [4, 5].

StomataBench is a collection of datasets for stomatal detection and density estimation under structured distribution shift. StomataBench pairs a single training set, OD-Train, with four held-out test sets designed to investigate distinct deployment failure modes (Table 1, Figure 1). OD-Train contains 3,154 microscopy images and 134,802 annotated stomata across 21 woody-plant species, all flowering plants: 20 eudicots and one magnoliid. It has no monocots (such as grasses and cereals), gymnosperms (such as conifers), or ferns. T-MAWP tests taxonomic generalization under the same microscope and annotation protocol. T-DCWP tests generalization across multiple woody-plant species with a different microscope [2]. T-MSWP targets imaging-system and annotation shift within woody plants [3]. T-GSH aggregates published multi-source datasets of nine taxa, from eudicots such as soybean and poplar to monocot crops such as wheat, barley, rice, and maize. Its calibrated wheat subset serves as a monocot extrapolation stress test [4].

StomataBench evaluation is designed around two principles. First, out-of-distribution (OOD) evaluation is grounded in taxonomic distance. Species in the test sets are arranged on a six-level OOD ladder (L0–L5) ordered by clade divergence from the training distribution [2, 6]. Standard OOD benchmarks often define shift using synthetic corruptions or dataset-specific domain splits [7, 8]. These settings are useful, but they do not measure how biologically far the test organisms are from the training organisms. Taxonomic distance provides a meaningful shift axis for stomatal phenotyping because it orders test species by how closely they are related to the training species, and stomatal complex type and epidermal morphology carry taxonomic information. Measuring generalization along this OOD ladder shows, species by species, where detector performance changes, but it does not replace validation in the intended deployment domain. Second, we evaluate detectors both as detectors and as measurement tools: with standard COCO localization metrics, and with count error, density error, and species-level density R^2 at a fixed confidence threshold. The COCO metrics

measure whether the predicted boxes land on the stomata [9]. The count and density metrics ask a different question: whether those boxes lead to counts that preserve the biologically meaningful variation in stomatal density, which is the measurement used for downstream trait discovery [10].

A broad empirical study of 15 fine-tuned detectors spanning two-stage CNNs, one-stage CNNs, YOLO-family models, real-time deformable transformers, and foundation-model-derived detectors shows that detector transfer is not a single-axis problem [11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21]. The best detector changes across shift types, and strong localization performance does not necessarily imply reliable downstream density estimation.

Contributions.

- **Dataset.** StomataBench provides a custom OD-Train training set, a custom T-MAWP held-out test set, and aggregated external test sets from published sources.
- **Evaluation protocol.** StomataBench introduces a taxonomically grounded OOD ladder (L0–L5) and evaluates detectors with standard COCO localization metrics and with count error, density error, and calibrated species-level stomatal-density R^2, connecting detector performance to downstream phenotyping reliability.
- **Empirical study.** Across 15 fine-tuned detectors spanning multiple architecture families, StomataBench shows that the best detector is shift-dependent and that detection AP can disagree with density-based trait recovery.

Data availability. Datasets, annotations, metadata, and per-image prediction files are released at `https://huggingface.co/datasets/Sainath001/StomataBench`, fine-tuned model checkpoints at `https://huggingface.co/Sainath001/stomatabench_neurips_weights`, and evaluation code and result tables at `https://github.com/gummi001/stomatabench_neurips_eval`. All images and annotations are released under CC-BY-NC-4.0 with Croissant metadata, and the datasheet is Section S1.

Table 1: StomataBench dataset composition and evaluation subsets (scene-primary: one image per scene). Shift from OD-Train: T taxonomic, I imaging system, A annotator.

Dataset	Shift	Taxa	OOD levels	Released all images		Scene-primary AP, F1, count MAE		Calibrated density	
				Images	Stomata	Images	Stomata	Images	Stomata
Training and validation									
OD-Train	–	21	–	3,154	134,802	3,139	134,614	3,139	134,614
OD-Valid	–	5	L0	138	8,792	138	8,792	138	8,792
Held-out test sets									
T-MAWP	T	17	L1 L2 L3 L4	456	17,164	456	17,164	456	17,164
T-DCWP	T+I	39	L0 L1 L2 L3 L4	119	2,858	119	2,858	119	2,858
T-MSWP	T+I+A	18	L1 L3 L4	10,715	514,568	10,284	503,898	9,479	421,290
T-GSH	T+I+A	9	L4 L5	8,229	319,161	6,943	288,543	1,930	40,993

2 Related Work

Accurate stomatal phenotyping under OOD shift. Traditional phenotyping workflows rely on microscopy followed by manual counting or morphometric measurement in tools such as ImageJ. These workflows are accurate but labor-intensive, limiting their use in large-scale breeding, ecological, genotypic and phenotypic studies. Recent work has addressed this bottleneck by training deep learning models to localize, count, and segment stomata in microscopy images [4]. Systems such as StomataCounter [2], LabelStoma [22], SSD-based detectors for wheat [23] and oil palm [24], and StomataScorer for maize [25] show that automated stomatal phenotyping can work well within a single acquisition pipeline. However, these evaluations leave open whether detector predictions remain reliable on images from new species, microscopes, laboratories, or annotation protocols.

OOD evaluation asks whether models remain reliable when test data differ from training data. In plant phenotyping, this matters because deployment often involves new species, new clades, differ-

ent imaging systems, and different annotation practices [5]. StomataBench follows this motivation by separating taxonomic shift, imaging-system shift, annotator shift, and combined multi-source shift into distinct held-out test sets.

Object-detection benchmarks usually evaluate localization quality using metrics such as COCO AP and recall [9]. These metrics are necessary, but they are not always sufficient when detections are used as scientific measurements. In stomatal phenotyping, the downstream quantity is density, not the bounding box itself. A detector may localize many stomata correctly while still producing biased counts because of false positives, false negatives, confidence thresholds, or calibration differences. Related microscopy and phenotyping studies therefore often report count error or correlation in addition to detection scores [10, 25]. StomataBench adopts this measurement view: it tests whether detector outputs preserve species-level density variation.

Table 2: The six-level OOD ladder (illustrated in Figure 1): the biological holdout and generalization regime of each level, with the held-out test species, scene-primary images and calibrated images at that level.

Level	Biological holdout	Generalization regime	Species	Images	Calibrated
L0[†]	Familiar species, new accession	No taxonomic shift (control)	1	3	3
L1	New species, familiar genus	Within-genus shift	24	6,455	5,650
L2	New genus, familiar family	Within-family shift	6	60	60
L3	New family, familiar order	Within-order shift	21	783	783
L4	New order, same clade	Within-clade shift	24	4,868	3,671
L5	New clade: monocots	Cross-clade shift	6	5,633	1,817

[†] L0 also holds the OD-Valid validation set (5 species, 138 images, 138 calibrated). It is not a held-out test and is not counted above.

Object detection model families. Modern object detection has evolved through several architectural paradigms. Two-stage detectors such as Faster R-CNN [11, 13, 14] generate class-agnostic region proposals with a Region Proposal Network and then refine these candidates using a second-stage classifier and box regressor. This proposal-refinement design remains a strong baseline for small-object detection because candidate regions are explicitly revisited before final prediction. In contrast, one-stage dense detectors such as RetinaNet [12, 13, 14] remove the proposal stage and directly predict boxes and class probabilities over multi-scale feature maps, using focal loss to address foreground–background imbalance.

The YOLO family represents a deployment-oriented branch of one-stage detection, emphasizing fast single-pass inference, efficient feature aggregation, and strong training-time augmentation [15, 16]. YOLOv26-X further reflects the recent shift toward end-to-end real-time detection without non-maximum suppression (NMS). Transformer-based detectors provide a different formulation by casting detection as set prediction. RT-DETRv2 and RT-DETRv4 [18, 19] extend this paradigm toward real-time inference through efficient multi-scale encoding and a deformable-attention decoder. RT-DETRv4 further incorporates training-time vision-foundation-model distillation, using DINOv3 as a teacher through Deep Semantic Injection and Gradient-guided Adaptive Modulation, while preserving inference-time efficiency [19, 21].

Finally, foundation-model-derived detectors provide a complementary route to stomatal detection by recasting category-specific detection as language-conditioned or concept-prompt-conditioned set prediction. Grounding DINO with a Swin-Base backbone (GDINO-SB) [17] represents language-aligned open-set detection. It combines a visual backbone with a BERT-based text encoder, fuses image and text features through cross-modal attention, and uses language-guided query selection followed by a cross-modality decoder to produce bounding boxes grounded to the input phrase.

SAM 3 [20] represents a newer promptable-concept vision foundation model. Although the SAM family is primarily known for promptable segmentation, SAM 3 differs from earlier SAM models by including an image-level detector that can localize all instances matching a concept prompt. Architecturally, SAM 3 combines a vision encoder, text encoder, geometry encoder, DETR-style encoder–decoder, learned object queries, and a mask decoder, allowing the same model to produce boxes, masks, and confidence scores for concept-matching instances. Here, SAM3-OD denotes our detection-only use of the official SAM 3 model rather than a modified architecture.

Including GDINO-SB and SAM3-OD lets StomataBench compare foundation-model-derived detectors with closed-set detectors under the same shifts, without isolating which of their components explains a difference.

3 Datasets

StomataBench comprises one training set (OD-Train), one validation set (OD-Valid), and four held-out test sets (T-MAWP, T-DCWP, T-MSWP, T-GSH). Table 1 and Figure 1 summarize the splits and taxonomic distance.

OD-Train. The training set consists of 3,154 microscopy images and 134,802 annotated stomatal complexes across 21 species in 10 botanical families (Table S1, Figure S1).

Coverage is concentrated in Rosales, with secondary representation from Ericales, Cornales, Ranunculales, Vitales, Fagales, Sapindales, Dipsacales, and Laurales. Each stomatal complex is annotated as a single-class (*stomata*) axis-aligned bounding box.

OD-Valid. The validation set contains 138 images and 8,792 stomata from five species (Table S2a, Figure S2a). All five species are in OD-Train, but the images come from different accessions (L0).

T-MAWP. The first test set contains 456 images and 17,164 stomata across 17 species (Table S2b, Figure S2b). It uses the same microscope, working distance, magnification, calibration, and annotation protocol as OD-Train and OD-Valid, but includes species-level taxonomic holdouts spanning the OOD ladder used for density evaluation (Table 2).

Imaging and annotation protocol for OD-Train, OD-Valid, and T-MAWP. Images from OD-Train, OD-Valid, and T-MAWP were sampled from living accessions at the Arnold Arboretum of Harvard University, Boston. All images were captured by the authors using a ProScope HR handheld digital microscope (USB 2.0, 400× magnification) connected to a tripod-mounted laptop running ProScope HR v1.2.1. Image resolution is 1280×1024 px, with calibration scale $s = 0.5988\,\mu\mathrm{m\,px}^{-1}$ (167 px per 0.1 mm) measured using a stage micrometer. Imaging was performed on the abaxial (lower) leaf surface. The clear contact dome was gently pressed against the leaf to stabilize the field of view, and the integrated LED ring provided coaxial fill illumination. Field sessions were conducted under natural daylight, and images with strong specular highlights from direct solar reflections were excluded during annotation. For hardwood species, the lowest accessible branch was sampled. Visible stomata were annotated as bounding boxes in RectLabel Pro under the supervision of a botanist. No image augmentation was applied before annotation.

T-DCWP. The second test set contains 119 images and 2,858 stomata across 39 species (Table S3, Figure S3). Source images are publicly available and were captured in Washington, DC, using a high-resolution Olympus BX-63 microscope under a non-destructive imaging protocol [2]. A curated subset was annotated by the authors using the same single-class bounding-box protocol as OD-Train and T-MAWP. Images with calibrated scale bars are used for species-level density evaluation (Table 2).

T-MSWP. The third test set, T-MSWP, contains 10,715 images and 514,568 annotated stomata across 18 taxa (Table S4, Figure S4). It primarily evaluates imaging-system and annotation shift because images were acquired using a different protocol from OD-Train, and the annotations were made by the original authors and obtained from their public release, in YOLO TXT format with two classes, *stomata* and *stomata pore*. We retained only the *stomata* class and converted the annotations to COCO format. Source images are publicly available and were captured at Mississippi State University [3].

T-MSWP contains both order-overlapping and order-unseen eudicot taxa relative to OD-Train. The Level-4 OOD subset consists of taxa from orders absent from OD-Train: *Ilex opaca* in Aquifoliales, *Fraxinus caroliniana* in Lamiales, *Populus* spp. and *Salix* spp. in Malpighiales, and *Dirca palustris* in Malvales. Density uses the per-image calibration supplied by the source authors. Images at X100 magnification were used for detection and counting only (Table S6).

T-GSH. The fourth test set contains 8,229 images and 319,161 stomata aggregated from Stomata Hub source datasets (Table S5, Figure S5) [2, 4, 22, 23, 24, 25, 26, 27, 28]. Original annotations were provided in XML and converted to COCO format for evaluation. They were made by the authors of the source datasets, then checked by botanists at Stomata Hub, who applied Contrast

Limited Adaptive Histogram Equalization (CLAHE) to reduce inter-source color bias and host the images on their website [4]. T-GSH holds all L5 species in the detection benchmark, the monocots. For density evaluation, only the calibrated images are used (Table 1), and the calibrated Gibbs wheat subset (*Triticum aestivum*, Table S5) serves as the L5 monocot stress test (Table 2) [26].

4 Models

Fifteen open-source object detectors were evaluated, spanning five architecture families: two-stage convolutional detectors, one-stage convolutional detectors, YOLO-family detectors, real-time deformable transformer detectors, and foundation-model-derived detectors. The benchmark includes Faster R-CNN variants with ResNet and ResNeXt backbones, RetinaNet, YOLOv8/11/12/26, RT-DETRv2, RT-DETRv4, GDINO-SB, and SAM3-OD [11, 12, 13, 14, 15, 16, 17, 18, 19, 20]. Model sizes range from 31.4M to 839.4M parameters. All detectors were fine-tuned once on OD-Train using a single *stomata* class and evaluated unchanged on every held-out test set. No model was given taxonomic labels, dataset identity, or test-set-specific tuning. Architectures, parameter counts, learning rates, training schedules, and batch sizes are summarized in Table S7.

Pre-processing and training. Each detector used its native framework pre-processing pipeline to avoid introducing model-specific custom data handling. Detectron2 models were trained from COCO-format annotations and used the standard shortest-side resize of 800 px with maximum size 1333 px and ImageNet BGR normalization [14]. Ultralytics YOLO models used YOLO-format annotations, letterboxed inputs to 1024×1024, and retained the default Ultralytics training augmentations [15]. RT-DETRv2 used the HuggingFace auto-processor with images passed at native 1280×1024 resolution and ImageNet RGB normalization [18]. RT-DETRv4 used its official training configuration with 640×640 RGB inputs [19, 21]. GDINO-SB used the official HuggingFace processor with the text prompt "a stomata ." [17, 29]. SAM3-OD used the ODinW text-only configuration with 1008×1008 inputs and the prompt "stomata" during fine-tuning and inference [20]. All models were optimized with AdamW. All pretrained checkpoints were fine-tuned end-to-end on OD-Train.

Table 3: COCO box AP (%) on OD-Valid and the four held-out test sets. Best **bold**, second-best underlined.

Model	OD-Valid		T-MAWP		T-DCWP		T-MSWP		T-GSH	
	$AP_{50:95}$	AP_{50}	$AP_{50:95}$	AP_{50}	$AP_{50:95}$	AP_{50}	$AP_{50:95}$	AP_{50}	$AP_{50:95}$	AP_{50}
Two-stage convolutional detectors										
FRCNN-R50-1×	42.59	86.66	33.92	77.44	20.99	51.62	19.67	41.24	8.29	24.14
FRCNN-R50-3×	42.85	84.80	32.29	73.44	16.69	41.49	17.66	38.44	7.50	21.10
FRCNN-R101-1×	41.73	84.68	33.21	75.46	6.55	19.76	20.52	44.83	3.80	10.69
FRCNN-X101	40.87	82.94	32.47	75.66	12.01	32.71	19.01	38.84	3.10	9.66
One-stage convolutional detectors										
Retina-R50-1×	42.39	86.05	34.40	74.77	12.99	37.69	20.13	41.95	8.85	24.52
Retina-R50-3×	42.05	85.55	32.89	71.73	15.02	40.49	16.74	37.83	10.54	29.40
Retina-R101-3×	41.86	87.27	33.57	76.58	6.66	23.57	13.85	31.78	9.06	23.48
YOLO-family detectors										
YOLOv8	47.13	88.26	36.14	77.06	9.10	24.23	23.69	43.77	4.68	12.23
YOLOv11	47.78	88.94	<u>37.76</u>	<u>80.02</u>	27.98	63.57	<u>44.57</u>	**77.89**	13.56	37.26
YOLOv12	<u>48.30</u>	89.76	37.38	79.70	18.52	48.56	29.10	65.43	13.46	35.27
YOLOv26	47.13	88.22	37.45	78.76	27.74	65.08	31.81	62.93	14.96	43.09
Transformer detectors										
RT-DETRv2	46.05	<u>90.67</u>	35.72	78.44	25.90	58.27	24.40	52.20	15.07	39.84
RT-DETRv4	**49.15**	**92.94**	**39.08**	**82.74**	25.98	65.84	26.87	58.66	19.40	50.18
Foundation-model-derived detectors										
GDINO-SB	43.40	86.74	33.64	75.11	<u>43.26</u>	<u>81.59</u>	**45.28**	<u>76.27</u>	<u>20.96</u>	<u>52.57</u>
SAM3-OD	43.74	86.54	35.40	78.82	**46.46**	**85.58**	21.90	46.53	**24.93**	**63.18**

5 Evaluation

StomataBench evaluation is based on two complementary axes: (i) detector localization under held-out distribution shift, and (ii) calibrated stomatal-density recovery for biological trait discovery. The first evaluation treats each test set as an object-detection benchmark and reports standard COCO bounding-box metrics [9]. The second treats detections as measurements of a biological phenotype and asks whether predicted counts preserve image-level stomatal-density variation within species.

Detection benchmark. For standard object-detection evaluation, each model outputs a ranked list of bounding boxes and confidence scores for the single class *stomata*. Predictions are exported at a low score threshold of 0.001 (or a model's higher score floor, Table S8) to preserve the ranked list used by COCO AP integration. We report COCO bounding-box AP averaged over intersection-over-union (IoU) thresholds $t \in \{0.50, 0.55, \ldots, 0.95\}$, denoted $\mathrm{AP}_{50:95}$, together with AP_{50} and AR@100, the average recall with at most 100 detections per image. AP and AR are percentages computed with `pycocotools`. True positives (TP), false positives (FP), false negatives (FN) and F1 use the boxes above the confidence threshold τ, matched one-to-one to annotated stomata at IoU 0.50. Count mean absolute error (MAE), in stomata per image, needs no box matching. Full results at $\tau = 0.25$ are in Section S3, the other thresholds in Section S4, and paired-image bootstrap intervals in Sections S5 and S6.

Stomatal density for trait discovery. Stomatal density (stomata mm^{-2}) is the phenotype of biological interest. For each calibrated image, ground-truth and predicted stomatal counts are converted to density using the physical field-of-view area. For image i, let (W_i, H_i) denote image width and height in pixels and let s_i denote the calibration scale in $\mu\mathrm{m\,px}^{-1}$. The field-of-view area is $A_i = W_i H_i s_i^2 / 10^6$ mm^2. With c_i the ground-truth stomata count and $\hat{c}_i^m(\tau)$ the number of boxes from model m with confidence at least τ, the ground-truth and predicted densities are $\rho_i = c_i / A_i$ and $\hat{\rho}_i^m(\tau) = \hat{c}_i^m(\tau) / A_i$ (stomata mm^{-2}). OD-Train, OD-Valid, and T-MAWP use the ProScope calibration $s = 0.5988\,\mu\mathrm{m\,px}^{-1}$. External sources use their source-specific calibration values. Species are grouped into OOD levels by the first taxonomic rank at which they differ from the species in OD-Train (Table 2), with families, orders, and clades following APG IV [30]. The levels are ordinal, not a continuous phylogenetic distance, and detection difficulty is not assumed to rise monotonically with every level. A Lauraceae species introduces a new family within a represented order and is L3. *Magnolia virginiana* introduces order Magnoliales within the represented magnoliid clade and is L4. Per-species values are in Section S7.

For each species s, let $\mathcal{I}_s = \{i : \mathrm{species}(i) = s\}$ denote its test images, $n_s = |\mathcal{I}_s|$, and $\bar{\rho}_s$ denote the mean ground-truth density for that species. For model m at confidence threshold τ, species-level density agreement is measured as

$$R^2_{s,m}(\tau) = 1 - \frac{\sum_{i \in \mathcal{I}_s} \left(\rho_i - \hat{\rho}_i^m(\tau)\right)^2}{\sum_{i \in \mathcal{I}_s} \left(\rho_i - \bar{\rho}_s\right)^2}.$$

Results are reported at $\tau = 0.25$, one permissive threshold for every detector and test set, so that counts rather than high-confidence boxes drive density and no model or dataset is tuned at test time. We also report the shared and model-specific validation thresholds (Table S8), which are selected on OD-Valid and frozen before testing, and fixed 0.50 and 0.75 (Figure 2). The coefficient satisfies $R^2_{s,m} \in (-\infty, 1]$. Negative values indicate that the model's predicted densities are worse than a constant reference that predicts the species mean density $\bar{\rho}_s$ for every image. Species with fewer than three calibrated images are excluded from species-level R^2 because the estimate is not meaningful with only two image-level observations. In T-DCWP, this excludes *Crataegus viridis* and *Stewartia malacodendron* ($n_s = 2$ each), which Figure S20 still shows.

6 Results

6.1 Detection benchmark across held-out test sets

Table 3 summarizes COCO bounding-box performance across the four held-out test sets. No single detector dominates across all shifts. RT-DETRv4 leads on T-MAWP, where the microscope, calibration, and annotation protocol match OD-Train. GDINO-SB leads narrowly on T-MSWP, where the dominant shift is imaging system and annotation style. SAM3-OD leads on T-DCWP and T-GSH

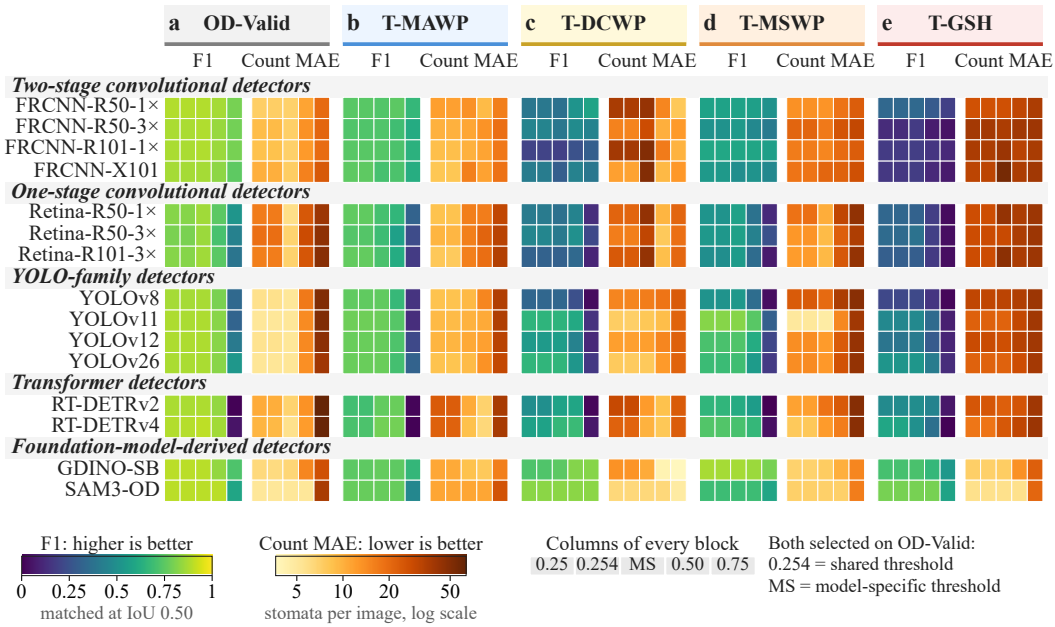


Figure 2: F1 (IoU 0.50) and count MAE of every detector at five confidence settings: fixed 0.25, 0.50 and 0.75, the shared threshold and each detector's model-specific threshold (MS), both selected on OD-Valid (Table S8). Each metric has one colour scale for all five sets, so colours can be compared across panels. Count MAE is in stomata per image on a log scale. Values at 0.25 are in Table S10 and at the other settings in Section S4.

with GDINO-SB second on both, so the two foundation-model-derived detectors transfer best to these sets, an empirical pattern whose cause the benchmark does not isolate. The T-MAWP leader RT-DETRv4 does not transfer uniformly, and the leading detector depends on the type of shift being tested. These results support evaluating stomatal detectors across biologically and visually distinct shifts rather than relying on a single held-out test set. Example predictions of every detector are in Section S10.

6.2 Stomatal density across the OOD ladder

Figure 3 summarizes within-species density R^2 at $\tau = 0.25$ for every species with at least three calibrated images. Every value is in Section S7, the other per-species metrics in Section S8, and the other thresholds in Section S9. The strongest contrast is the L5 wheat subset: on the 348 calibrated Gibbs wheat images, every detector produces negative R^2, meaning predicted densities are worse than using the wheat mean ground-truth density as a constant prediction. Localization fails there too: at IoU 0.50 only GDINO-SB and SAM3-OD recover more than 1% of wheat stomata, and at IoU 0.30 these two match most of them, so their wheat boxes are loose rather than absent (Table S18). This occurs even though SAM3-OD leads detection AP on the full T-GSH set. Detection transfer and trait recovery are therefore related but not equivalent. Wheat failure is not one shared bias direction: nine detectors overcount and six undercount (Table S18), and wheat R^2 stays negative at every threshold (Section S9).

Within T-MAWP, several species show strong positive density recovery, especially at L1. These cases show that detector-derived counts can recover biologically meaningful stomatal-density variation for some woody-plant species. However, the best-performing detector changes across species. Between fixed 0.25 and 0.75, the leading detector by density R^2 changes for 15 of 17 T-MAWP species and by count MAE for 14 (Section S9). High species-level R^2 values are distributed across Faster R-CNN, RetinaNet, YOLO, and GDINO-SB rather than being concentrated in one model family. On T-DCWP, where most species have three images, these values are descriptive. This species-level view exposes a failure mode that dataset-level AP hides: a detector can have high set-

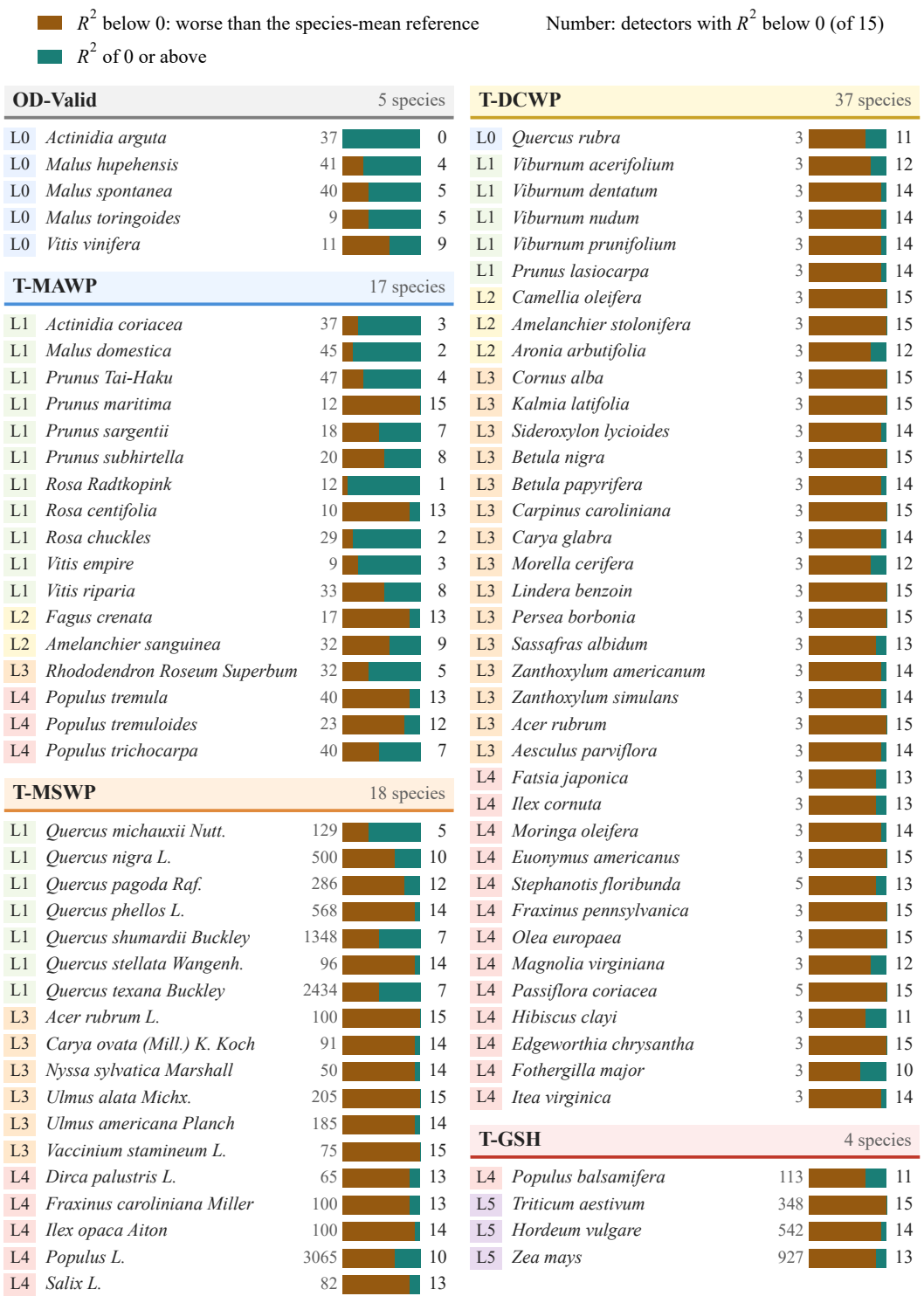


Figure 3: Within-species density R^2 at $\tau = 0.25$ for the 81 taxon–dataset groups with at least three calibrated images (number beside each name). A taxon in two test sets counts in each (*Acer rubrum*), and *Populus* and *Salix* are genus-level labels. Each bar splits the 15 detectors into R^2 below 0 (brown, worse than the species-mean reference) and 0 or above (teal). The label counts those below 0. T-DCWP values (most species have three images) are descriptive. The values are in Figures S19 and S20.

level AP while failing to recover density variation for particular species (matched examples in Section S6). AP evaluates ranked boxes over several IoUs, while count MAE and density use only the boxes kept at one confidence value, where false positives and false negatives can cancel (Figure 2). On T-MAWP at $\tau = 0.25$, RT-DETRv4 has the highest AP and recall but also many background false positives (Table S12) and one of the highest count MAEs (Table S10), while at fixed 0.50 it has the lowest count MAE (Figure 2). A species' AP leader is also its density R^2 leader for only one of the 17 T-MAWP species. AP therefore remains the standard check of localization, but alone it does not show whether a detector is a reliable density tool.

7 Limitations

Several limitations should be noted. Density estimation is constrained by the physical field of view. Larger calibrated leaf areas could provide more stable density estimates, but a wider field of view lowers the optical resolution needed to detect individual stomata [25]. At the same time, OD-Train, OD-Valid, and T-MAWP were collected from intact leaves under field conditions, making the benchmark relevant to scalable tree breeding and field phenotyping workflows.

OD-Train is taxonomically broad within woody eudicots (plus one magnoliid) but does not include monocots, gymnosperms, ferns, or basal angiosperms (the ANA grade). These are priorities for future expansion. The L5 results, from the T-GSH monocots (wheat, barley, rice, maize, and oil palm), should therefore be interpreted as a stress test for eudicot-to-monocot extrapolation, not as a complete estimate of monocot generalization. Species names follow each source dataset, so a few are older synonyms. T-MSWP and T-GSH are large heterogeneous test sets. Their per-species AP is in Section S8 (Figure S23). AP keeps the official COCO limit of 100 detections per image, which can understate detectors on images with more than 100 stomata (34% of T-MSWP and 37% of T-GSH stomata), so future analyses should address it.

Several T-MAWP and T-DCWP species have small image-level sample sizes or low within-species density variance. Species-level R^2 is sensitive to both factors and can be negative when count errors are systematic [10]. R^2 is read with caution where the ground-truth density varies little. Count MAE and bias are reported for every species (Section S8). Low-replication species are retained because they reflect a realistic biological-data regime rather than a perfectly balanced machine learning benchmark. Species-level density R^2 is reported at $\tau = 0.25$ to allow uniform comparison across detectors and datasets, but this threshold is not intended to be deployable. Duplicate detections, two boxes on one stoma, are counted in each detector's final output, without changing its NMS settings.

StomataBench currently uses single-class bounding-box annotations for stomatal complexes. It does not distinguish guard cells from pavement cells, annotate pores, estimate aperture state, or measure pore-level geometry. Extending the benchmark to keypoint detection, instance segmentation, and pore-level labels would support a broader range of stomatal phenotyping tasks [25]. Images were also collected without controlled CO_2, humidity, water status, light history, developmental stage, or time-of-day metadata [28]. StomataBench is therefore a multidisciplinary benchmark for detection generalization and density measurement under distribution shift, not a controlled physiological experiment.

8 Conclusion

We introduced StomataBench, a benchmark for testing whether stomatal detectors remain reliable under taxonomic, imaging-system, annotator, and multi-source shift. Across 15 fine-tuned detectors, the detector with the highest COCO AP changes with the type of shift, so a single held-out test set is not enough to judge transfer. High set-level AP does not guarantee reliable measurement: on the 348 calibrated Gibbs wheat images every detector has negative density R^2, including those that lead AP on the full T-GSH set. At the same time, several woody species close to the training species, mostly at L1, show strong density recovery, so detector-derived counts can support field phenotyping when the test species is near the training set and the confidence threshold is checked on validation data. Biased density estimates can mis-rank genotypes or distort drought-screening decisions. Benchmark performance is not a deployment guarantee. Target-domain calibration, confidence selection on independent validation data, and human review remain necessary.

Acknowledgments and Disclosure of Funding

This work was supported by the Hatch-Multistate project (SD00R730-23) through the SDAES and in part by SDSU's HPC resources. We acknowledge the Arnold Arboretum of Harvard University for providing access to living collections and financial support through a Sinnott Award.

The authors declare no competing interests.

References

[1] Patricia L. M. Lang, Joel M. Erberich, Lua Lopez, Clemens L. Weiß, Gabriel Amador, Hannah F. Fung, Sergio M. Latorre, Jesse R. Lasky, Hernán A. Burbano, Moisés Expósito-Alonso, and Dominique C. Bergmann. Century-long timelines of herbarium genomes predict plant stomatal response to climate change. *Nature Ecology & Evolution*, 8(9):1641–1653, 2024. doi: 10.1038/s41559-024-02481-x.

[2] Karl C. Fetter, Sven Eberhardt, Rich S. Barclay, Scott Wing, and Stephen R. Keller. StomataCounter: a neural network for automatic stomata identification and counting. *New Phytologist*, 223(3):1671–1681, 2019. doi: 10.1111/nph.15892.

[3] Jiaxin Wang, Heidi J. Renninger, and Qin Ma. Labeled temperate hardwood tree stomatal image datasets from seven taxa of *Populus* and 17 hardwood species. *Scientific Data*, 11:1, 2024. doi: 10.1038/s41597-023-02657-3.

[4] Jonathon A. Gibbs and Alexandra J. Burgess. Application of deep learning for the analysis of stomata: a review of current methods and future directions. *Journal of Experimental Botany*, 75(21):6704–6718, 2024. doi: 10.1093/jxb/erae207.

[5] Jordan Ubbens, Ian Stavness, Michael P. Pound, and Wei Guo. Deep learning in plant phenotyping: the first ten years. *Plant Phenomics*, 7(4):100062, 2025. doi: 10.1016/j.plaphe.2025.100062.

[6] Susana Magallón, Sandra Gómez-Acevedo, Luna L. Sánchez-Reyes, and Tania Hernández-Hernández. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytologist*, 207(2):437–453, 2015. doi: 10.1111/nph.13264.

[7] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[8] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 2021.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

[10] C. P. James Chen, Robin R. White, and Ryan Wright. Common pitfalls in evaluating model performance and strategies for avoidance in agricultural studies. *Computers and Electronics in Agriculture*, 234:110126, 2025. doi: 10.1016/j.compag.2025.110126.

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[12] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. doi: 10.1109/ICCV.2017.324.

[13] Yibo Sun, Zhe Sun, and Weitong Chen. The evolution of object detection methods. *Engineering Applications of Artificial Intelligence*, 133:108458, 2024. doi: 10.1016/j.engappai.2024.108458.

[14] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[15] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8. `https://github.com/ultralytics/ultralytics`, 2023. Software, version 8.0.0, AGPL-3.0 license.

[16] Yunjie Tian, Qixiang Ye, and David Doermann. YOLOv12: Attention-centric real-time object detectors. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 78433–78457. Curran Associates, Inc., 2025. doi: 10.52202/085713-2627.

[17] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, volume 15105 of *Lecture Notes in Computer Science*, pages 38–55. Springer, 2024. doi: 10.1007/978-3-031-72970-6_3.

[18] Wenyu Lv, Yian Zhao, Qinyao Chang, Kui Huang, Guanzhong Wang, and Yi Liu. RT-DETRv2: Improved baseline with bag-of-freebies for real-time detection transformer. *arXiv preprint arXiv:2407.17140*, 2024.

[19] Zijun Liao, Yian Zhao, Xin Shan, Yu Yan, Chang Liu, Lei Lu, Xiangyang Ji, and Jie Chen. RT-DETRv4: Painlessly furthering real-time object detection with vision foundation models. In *European Conference on Computer Vision*, volume 17074 of *Lecture Notes in Computer Science*, pages 95–113. Springer, 2026. doi: 10.1007/978-3-032-37029-7_7.

[20] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026.

[21] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

[22] Angela Casado-García, Arantza del-Canto, Alvaro Sanz-Saez, Usue Pérez-López, Amaia Bilbao-Kareaga, Felix B. Fritschi, Jon Miranda-Apodaca, Alberto Muñoz-Rueda, Anna Sillero-Martínez, Ander Yoldi-Achalandabaso, Maite Lacuesta, and Jónathan Heras. LabelStoma: A tool for stomata detection based on the YOLO algorithm. *Computers and Electronics in Agriculture*, 178:105751, 2020. ISSN 0168-1699. doi: 10.1016/j.compag.2020.105751. URL `https://www.sciencedirect.com/science/article/pii/S0168169920312977`.

[23] Yosuke Toda, Toshiaki Tameshige, Masakazu Tomiyama, Toshinori Kinoshita, and Kentaro K. Shimizu. An affordable image-analysis platform to accelerate stomatal phenotyping during microscopic observation. *Frontiers in Plant Science*, 12:715309, 2021. doi: 10.3389/fpls.2021.715309.

[24] Qi Bin Kwong, Yick Ching Wong, Phei Ling Lee, Muhammad Syafiq Sahaini, Yee Thung Kon, Harikrishna Kulaveerasingam, and David Ross Appleton. Automated stomata detection in oil palm with convolutional neural network. *Scientific Reports*, 11:15210, 2021. doi: 10.1038/s41598-021-94705-4.

[25] Xiuying Liang, Xichen Xu, Zhiwei Wang, Lei He, Kaiqi Zhang, Bo Liang, Junli Ye, Jiawei Shi, Xi Wu, Mingqiu Dai, and Wanneng Yang. StomataScorer: a portable and high-throughput leaf stomata trait scorer combined with deep learning and an improved CV model. *Plant Biotechnology Journal*, 20(3):577–591, 2022. doi: 10.1111/pbi.13741.

[26] Jonathon A. Gibbs, Lorna McAusland, Carlos A. Robles-Zazueta, Erik H. Murchie, and Alexandra J. Burgess. A deep learning method for fully automatic stomatal morphometry and maximal conductance estimation. *Frontiers in Plant Science*, 12:780180, 2021. doi: 10.3389/fpls.2021.780180.

[27] Sophie B. Cowling, Hamidreza Soltani, Sean Mayes, and Erik H. Murchie. Stomata Detector: High-throughput automation of stomata counting in a population of African rice (*Oryza glaberrima*) using transfer learning. *bioRxiv*, 2021. doi: 10.1101/2021.12.01.469618.

[28] Zhuangzhuang Sun, Xiao Wang, Yunlin Song, Qing Li, Jin Song, Jian Cai, Qin Zhou, Yingxin Zhong, Shichao Jin, and Dong Jiang. StomataTracker: Revealing circadian rhythms of wheat stomata with in-situ video and deep learning. *Computers and Electronics in Agriculture*, 212: 108120, 2023. doi: 10.1016/j.compag.2023.108120.

[29] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, 2020. Association for Computational Linguistics. URL `https://aclanthology.org/2020.emnlp-demos.6/`.

[30] The Angiosperm Phylogeny Group, Mark W. Chase, Maarten J. M. Christenhusz, Michael F. Fay, James W. Byng, Walter S. Judd, Douglas E. Soltis, David J. Mabberley, Alexander N. Sennikov, Pamela S. Soltis, and Peter F. Stevens. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society*, 181(1):1–20, 2016. doi: 10.1111/boj.12385.

Supplementary Information

S1 Datasheet for StomataBench

This section is the datasheet for StomataBench. It describes each set species by species and shows the taxonomic distance behind every OOD level (Section 3, Table 1 and Figure 1). Tables S1–S5 list the species of OD-Train and of each held-out test set with their family, order, images and stomata, and Figures S1–S5 show representative micrographs of each set. Density metrics use only calibrated images, and Table S6 lists the images left out and why. All images are released under CC-BY-NC-4.0.

Table S1: OD-Train species composition (21 species across 10 families). Mean density per species ranges from 30.92 to 277.27 stomata mm^{-2}.

Accession ID(s)	Scientific name	Family	Order	Images (N)	Stomata (N)	GT density (mm^{-2})
403-97*C & 905-85*A	*Actinidia arguta*	Actinidiaceae	Ericales	161	4,662	61.61
257-2006*B	*Actinidia macrosperma*	Actinidiaceae	Ericales	145	5,110	74.99
107-2004*B	*Calycanthus* 'Venus'	Calycanthaceae	Laurales	55	3,165	122.44
1690-77*B	*Clematis terniflora*	Ranunculaceae	Ranunculales	308	6,900	47.67
276*A	*Cotinus coggygria*	Anacardiaceae	Sapindales	56	5,285	200.81
14714-1*A	*Hydrangea paniculata* 'Praecox'	Hydrangeaceae	Cornales	434	6,306	30.92
427-2016*A	*Malus* 'Blanche Ames'	Rosaceae	Rosales	134	10,609	168.46
1951-80*A, 1951-80*B, 210-2001*A, 608-2016*A & 12198*B	*Malus hupehensis*	Rosaceae	Rosales	240	12,063	106.95
28-2013*A	*Malus ioensis*	Rosaceae	Rosales	49	4,966	215.64
79-2009*B, 51-2009*A, 78-2009*B, 48-2009*A, 49-2009*A, 79-2009*A, 75-2009*A & 74-2009*B	*Malus sieversii*	Rosaceae	Rosales	204	12,591	131.33
49-98*A, 46-98*A, 228-2013*B & 228-2013*C	*Malus spontanea*	Rosaceae	Rosales	128	6,724	111.77
80-2017C, 47-98*B & 50-2011B	*Malus toringoides*	Rosaceae	Rosales	62	4,685	160.78
510-94*A	*Malus* × *zumi*	Rosaceae	Rosales	172	7,846	97.06
596-86*A	*Prinsepia sinensis*	Rosaceae	Rosales	226	6,552	61.69
119-96*A	*Prunus* 'Washi-no-o'	Rosaceae	Rosales	75	3,712	105.31
13-2016*A	*Quercus rubra*	Fagaceae	Fagales	116	15,116	277.27
80-2012*A	*Rosa* 'Queen Elizabeth'	Rosaceae	Rosales	63	1,533	51.78
713-2010*MASS	*Rosa* 'Seafoam'	Rosaceae	Rosales	162	7,513	98.68
531-34*C	*Stewartia gemmata*	Theaceae	Ericales	54	1,308	51.54
158-2004*A	*Viburnum sargentii* 'Onondaga'	Adoxaceae	Dipsacales	55	1,029	39.81
332-84*A	*Vitis vinifera* × *amurensis* 'Beihong'	Vitaceae	Vitales	255	7,127	59.47
Total (21 species, 10 families)				**3,154**	**134,802**	

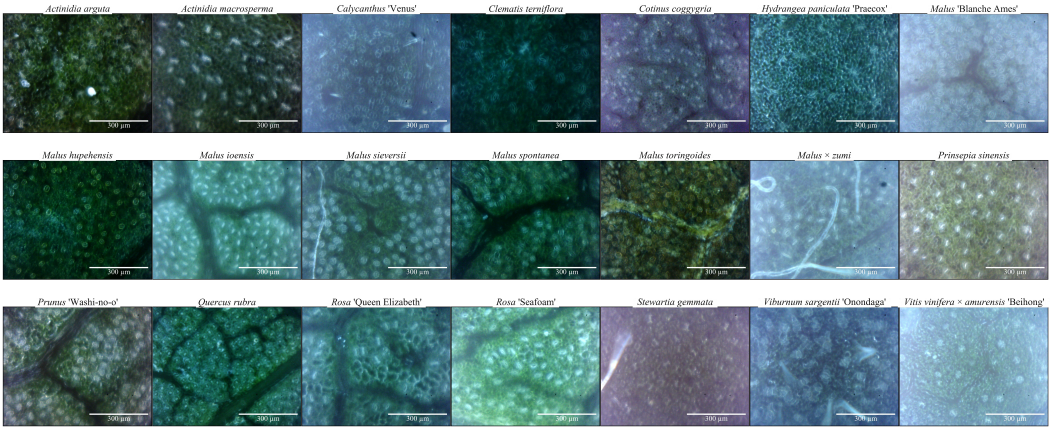


Figure S1: Representative stomata micrographs of diverse woody plants from OD-Train (300 μm scale bar).

Table S2: Species-level composition of (a) OD-Valid and (b) T-MAWP, grouped by OOD level relative to OD-Train (Table 2). (a) OD-Valid, the validation set: 138 images and 8,792 stomata from 5 OD-Train species on different accessions (L0). (b) T-MAWP: 456 images and 17,164 stomata across 17 held-out species at L1–L4.

OOD	Accession(s)	Scientific name	Family	Order	Images	Stomata	Mean density (mm^{-2})
(a) OD-Valid (validation set)							
L0	593-87*A	*Actinidia arguta*	Actinidiaceae	Ericales	37	1,434	82.5
L0	7241*A	*Malus hupehensis*	Rosaceae	Rosales	41	3,276	170.0
L0	103-2013*A, 216-2013*A	*Malus spontanea*	Rosaceae	Rosales	40	3,069	163.3
L0	202-2001*A	*Malus toringoides*	Rosaceae	Rosales	9	671	158.6
L0	331-84*A	*Vitis vinifera* × *amurensis*	Vitaceae	Vitales	11	342	66.2
(b) T-MAWP (held-out test set)							
L1	248-86*D	*Actinidia coriacea*	Actinidiaceae	Ericales	37	1,034	59.5
L1	130-2018*A, 130-2018*B	*Malus domestica*	Rosaceae	Rosales	45	1,532	72.4
L1	457-2010*A	*Prunus Tai-Haku*	Rosaceae	Rosales	47	1,675	75.8
L1	439-2008*A	*Prunus maritima*	Rosaceae	Rosales	12	1,177	208.7
L1	931-51*A	*Prunus sargentii*	Rosaceae	Rosales	18	907	107.2
L1	319-90*F	*Prunus subhirtella*	Rosaceae	Rosales	20	1,749	186.1
L1	198-2010*A	*Rosa Radtkopink*	Rosaceae	Rosales	12	222	39.4
L1	327-84*C	*Rosa centifolia*	Rosaceae	Rosales	10	285	60.6
L1	920-83*A	*Rosa chuckles*	Rosaceae	Rosales	29	913	67.0
L1	130-2019*B	*Vitis empire*	Vitaceae	Vitales	9	231	54.6
L1	1739-81*A, 302-89*C	*Vitis riparia*	Vitaceae	Vitales	33	1,242	80.1
L2	7058*B	*Amelanchier sanguinea*	Rosaceae	Rosales	32	1,307	86.9
L2	24-98*B	*Fagus crenata*	Fagaceae	Fagales	17	1,330	166.5
L3	866-50*B	*Rhododendron Roseum Superbum*	Ericaceae	Ericales	32	1,238	82.3
L4	547-59*MASS	*Populus tremula*	Salicaceae	Malpighiales	40	706	37.6
L4	170-2017*A	*Populus tremuloides*	Salicaceae	Malpighiales	23	569	52.6
L4	168-2015*E	*Populus trichocarpa*	Salicaceae	Malpighiales	40	1,047	55.7

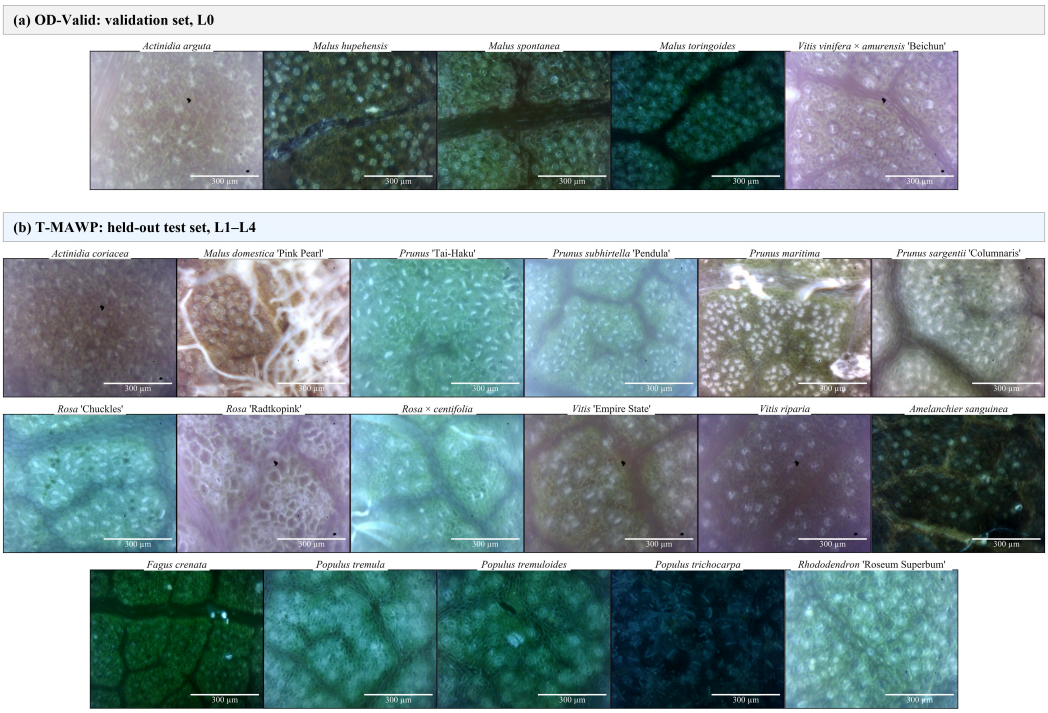


Figure S2: Representative stomata micrographs, one per species, of (a) OD-Valid, the validation set, and (b) T-MAWP, the held-out test set (300 μm scale bar).

Table S3: Species-level composition of T-DCWP by OOD level relative to OD-Train (Table 2): images, stomata and per-image mean GT density per species. Most of T-DCWP sits at L3/L4 (74% of images, 82% of stomata, in novel families and orders). Totals: 119 images and 2,858 stomata across 39 species.

Scientific name	Family	Order	Images (N)	Stomata (N)	GT density (mm^{-2})
L0: Familiar species, new accession					
Quercus rubra	Fagaceae	Fagales	3	94	337.84
		L0 subtotal (1 species)	3	94	—
L1: New species, familiar genus					
Viburnum acerifolium	Adoxaceae	Dipsacales	3	43	154.54
Viburnum dentatum	Adoxaceae	Dipsacales	3	54	194.08
Viburnum nudum	Adoxaceae	Dipsacales	3	26	93.45
Viburnum prunifolium	Adoxaceae	Dipsacales	3	48	172.51
Stewartia malacodendron	Theaceae	Ericales	2	12	64.69
Prunus lasiocarpa	Rosaceae	Rosales	3	43	154.54
		L1 subtotal (6 species)	17	226	—
L2: New genus, familiar family					
Camellia oleifera	Theaceae	Ericales	3	30	107.82
Amelanchier stolonifera	Rosaceae	Rosales	3	31	111.42
Aronia arbutifolia	Rosaceae	Rosales	3	93	334.25
Crataegus viridis	Rosaceae	Rosales	2	39	210.25
		L2 subtotal (4 species)	11	193	—
L3: New family, familiar order					
Cornus alba	Cornaceae	Cornales	3	50	179.70
Kalmia latifolia	Ericaceae	Ericales	3	43	154.54
Sideroxylon lycioides	Sapotaceae	Ericales	3	88	316.28
Betula nigra	Betulaceae	Fagales	3	36	129.39
Betula papyrifera	Betulaceae	Fagales	3	30	107.82
Carpinus caroliniana	Betulaceae	Fagales	3	118	424.10
Carya glabra	Juglandaceae	Fagales	3	86	309.09
Morella cerifera	Myricaceae	Fagales	3	108	388.16
Lindera benzoin	Lauraceae	Laurales	3	69	247.99
Persea borbonia	Lauraceae	Laurales	3	153	549.89
Sassafras albidum	Lauraceae	Laurales	3	118	424.10
Zanthoxylum americanum	Rutaceae	Sapindales	3	49	176.11
Zanthoxylum simulans	Rutaceae	Sapindales	3	22	79.07
Acer rubrum	Sapindaceae	Sapindales	3	117	420.50
Aesculus parviflora	Sapindaceae	Sapindales	3	151	542.70
		L3 subtotal (15 species)	45	1,238	—
L4: New order, same clade					
Fatsia japonica	Araliaceae	Apiales	3	48	172.51
Ilex cornuta	Aquifoliaceae	Aquifoliales	3	92	330.65
Moringa oleifera	Moringaceae	Brassicales	3	52	186.89
Euonymus americanus	Celastraceae	Celastrales	3	63	226.42
Stephanotis floribunda	Apocynaceae	Gentianales	5	195	200.47
Fraxinus pennsylvanica	Oleaceae	Lamiales	3	42	150.95
Olea europaea	Oleaceae	Lamiales	3	83	298.31
Magnolia virginiana	Magnoliaceae	Magnoliales	3	148	531.92
Passiflora coriacea	Passifloraceae	Malpighiales	5	200	198.30
Hibiscus clayi	Malvaceae	Malvales	3	35	125.79
Edgeworthia chrysantha	Thymelaeaceae	Malvales	3	37	132.98
Fothergilla major	Hamamelidaceae	Saxifragales	3	37	132.98
Itea virginica	Iteaceae	Saxifragales	3	75	269.55
		L4 subtotal (13 species)	43	1,107	—

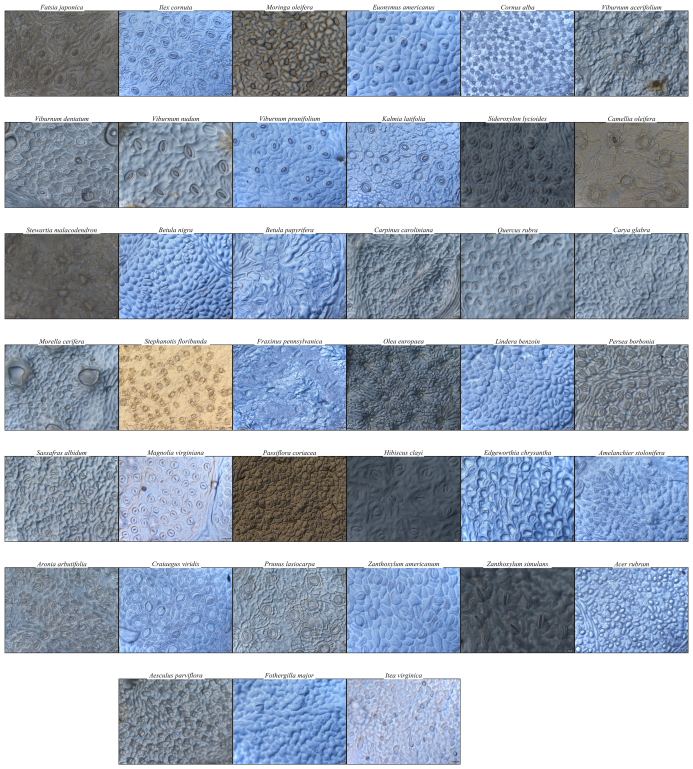


Figure S3: Representative stomata micrographs of diverse woody plants from T-DCWP.

Table S4: Species-level composition of T-MSWP by OOD level relative to OD-Train (Table 2). Calibration is resolved per image from acquisition magnification and sensor dimensions (four distinct scales among calibrated images, 239–930 px per 0.1 mm). *Quercus texana*, *Q. shumardii*, and *Populus* spp. hold ∼80% of annotated stomata. Totals: 10,715 images and 514,568 stomata across 18 taxa. Calibrated (N): released images whose physical scale passed calibration QC (others in Table S6). Density evaluation uses the 9,479 scene-primary images among them.

OOD	Scientific name	Family	Order	Images (N)	Stomata (N)	Calibrated (N)
L1	*Quercus michauxii*	Fagaceae	Fagales	129	4,735	129
L1	*Quercus nigra*	Fagaceae	Fagales	653	16,783	653
L1	*Quercus pagoda*	Fagaceae	Fagales	286	14,704	286
L1	*Quercus phellos*	Fagaceae	Fagales	568	34,027	568
L1	*Quercus shumardii*	Fagaceae	Fagales	1,612	115,138	1,348
L1	*Quercus stellata*	Fagaceae	Fagales	196	7,868	196
L1	*Quercus texana*	Fagaceae	Fagales	2,975	198,757	2,434
L3	*Vaccinium stamineum*	Ericaceae	Ericales	85	1,074	85
L3	*Carya ovata*	Juglandaceae	Fagales	91	2,097	91
L3	*Nyssa sylvatica*	Nyssaceae	Cornales	50	500	50
L3	*Acer rubrum*	Sapindaceae	Sapindales	100	1,759	100
L3	*Ulmus alata*	Ulmaceae	Rosales	205	7,374	205
L3	*Ulmus americana*	Ulmaceae	Rosales	185	4,054	185
L4	*Ilex opaca*	Aquifoliaceae	Aquifoliales	100	3,752	100
L4	*Fraxinus caroliniana*	Oleaceae	Lamiales	200	3,282	200
L4	*Populus* spp.	Salicaceae	Malpighiales	3,068	94,666	3,068
L4	*Salix* spp.	Salicaceae	Malpighiales	82	1,206	82
L4	*Dirca palustris*	Thymelaeaceae	Malvales	130	2,792	130
			Total (18 taxa)	**10,715**	**514,568**	**9,910**

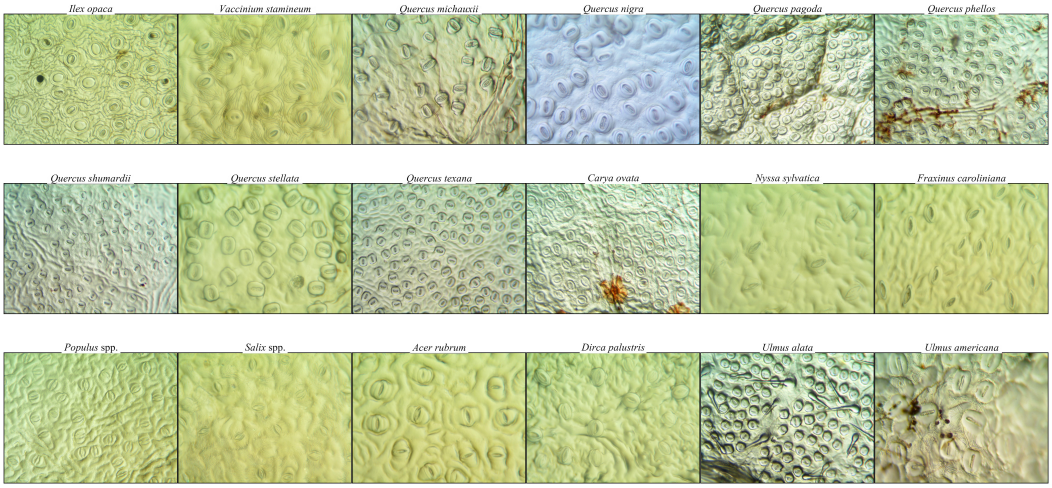


Figure S4: Representative stomata micrographs of diverse woody plants from T-MSWP.

Table S5: Source-level composition of T-GSH by OOD level relative to OD-Train (Table 2): eight identified species plus one unresolved *Triticum* group, from published datasets. T-GSH holds all L5 species, a new clade: ∼84% of images and ∼82% of stomata are monocots (Poaceae, Arecaceae), with the three LabelStoma species, including *Hordeum vulgare*, counted individually. Totals: 8,229 images and 319,161 stomata. Calibrated (N): released images whose physical scale passed calibration QC (others in Table S6). Density evaluation uses the 1,930 scene-primary images among them.

OOD	Scientific name	Family	Order	Source	Images (N)	Stomata (N)	Calibrated (N)
L4	*Glycine max*	Fabaceae	Fabales	[22]	600	33,054	0
L4	*Phaseolus vulgaris*	Fabaceae	Fabales	[22]	600	19,132	0
L4	*Populus balsamifera*	Salicaceae	Malpighiales	[2, 26]	113	4,225	113
L5	*Elaeis guineensis*	Arecaceae	Arecales	[24]	321	11,787	0
L5	*Hordeum vulgare*	Poaceae	Poales	[22]	600	15,915	600
L5	*Oryza glaberrima*	Poaceae	Poales	[27]	253	11,931	0
L5	*Triticum aestivum*	Poaceae	Poales	[26]	349	1,889	349
L5	*Triticum aestivum*	Poaceae	Poales	[28]	2,161	42,984	0
L5	*Triticum* spp.	Poaceae	Poales	[23]	732	56,556	0
L5	*Zea mays*	Poaceae	Poales	[25]	2,500	121,688	2,000
				Total	**8,229**	**319,161**	**3,062**

Table S6: Images excluded from species-level density evaluation. Detection and count evaluation use the 17,940 scene-primary images of OD-Valid and the four test sets, but density evaluation uses only the 12,122 of them that are calibrated (67.6%). The rows count all released images, redundant views included: 5,972 images, of which 5,818 are scene-primary. Exclusions are of two kinds: no calibration exists, or a nominal calibration failed a consistency check. The X100 oak images hold about as many stomata per image as the X200 images of the same species, although at half the magnification they would cover four times the leaf area, so the scale label, not a change in the leaves, is the likely cause. Ground-truth annotations are unaffected in both cases.

Species	Source subset	Set	Images	Stomata	Reason
No calibration available (4,667 images, 175,444 stomata)					
Triticum aestivum	StomataTracker [28]	T-GSH	2,161	42,984	No scale bar in image, no published scale
Triticum spp.	Toda et al. [23]	T-GSH	732	56,556	No scale bar in image, no published scale
Phaseolus vulgaris	LabelStoma	T-GSH	600	19,132	No scale bar in image, no published scale
Glycine max	LabelStoma	T-GSH	600	33,054	No scale bar in image, no published scale
Elaeis guineensis	Kwong et al. [24]	T-GSH	321	11,787	No scale bar in image, no published scale
Oryza glaberrima	Cowling et al. [27]	T-GSH	253	11,931	No scale bar in image, no published scale
Calibration failed consistency check (1,305 images, 161,696 stomata)					
Quercus texana	X100 group	T-MSWP	541	54,069	Density below the same species at X200/X400
Quercus shumardii	X100 group	T-MSWP	264	28,539	Density below the same species at X200/X400
Zea mays	TipScope [25]	T-GSH	500	79,088	Density from the published field of view disagrees with ProScope images of the same species
		Total excluded	**5,972**	**337,140**	

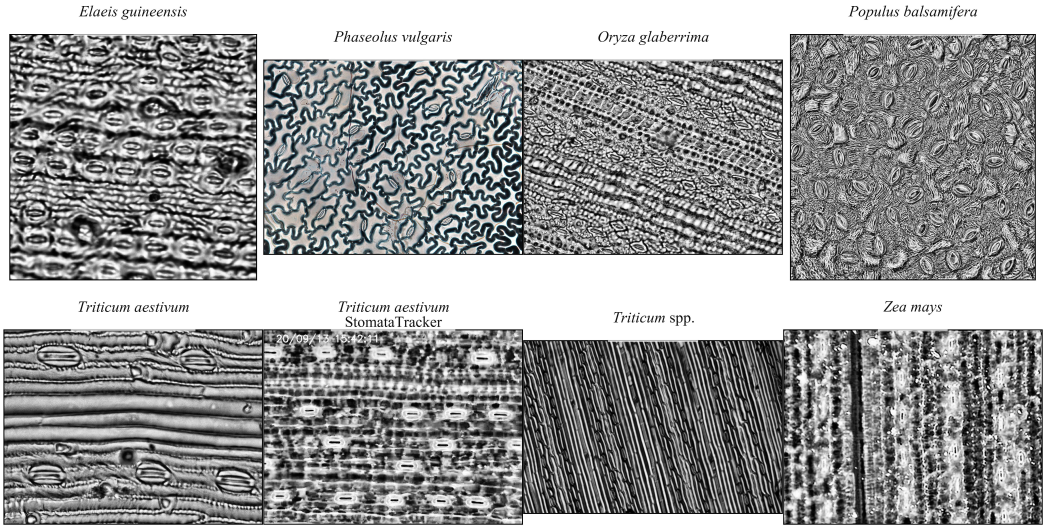


Figure S5: Representative stomata micrographs from T-GSH.

S2 Detectors and operating thresholds

Table S7 lists the detectors and their training settings (Section 4). Table S8 gives the confidence thresholds selected on OD-Valid: the shared and model-specific settings of Figure 2.

Thresholds are chosen from 950 values, 0.001 to 0.950 in steps of 0.001. For a detector and a threshold, the count error of an image is the difference between the number of boxes scoring at least the threshold and the annotated count, taken as a positive number. These errors are averaged within each of the 5 OD-Valid species and then over the species, so each species counts the same (species-balanced count MAE). A detector's threshold is the value with the lowest MAE among those within its score range, and a tie goes to the lower threshold. The shared threshold minimizes the mean of the 15 detectors' MAE over all 950 values. The 95% interval repeats the selection on 2,000 bootstrap resamples, each drawing images with replacement within every species so that the species keep equal weight, and takes the 2.5th and 97.5th percentiles. The shared threshold has no interval. The 1% plateau runs from the lowest to the highest threshold whose MAE is within 1% of the minimum. For 8 detectors some thresholds between those two ends have an MAE above that 1% limit.

Table S7: Training settings of the detectors, all fine-tuned on OD-Train with a single *stomata* class. Params: total parameters in millions. LR: peak learning rate. Schedule: total training budget. BS: batch size per step. Every GPU has 80 GB.

Model ID	Architecture	Params (M)	LR	Schedule	BS	Train GPU
Two-stage convolutional detectors [11, 14]						
FRCNN-R50-1×	Faster R-CNN, ResNet-50-FPN, 1×	41.4	5×10^{-5}	25,000 iters	4	1×A100
FRCNN-R50-3×	Faster R-CNN, ResNet-50-FPN, 3×	41.4	5×10^{-5}	25,000 iters	4	1×A100
FRCNN-R101-1×	Faster R-CNN, ResNet-101-FPN, 1×	60.4	5×10^{-5}	25,000 iters	4	1×A100
FRCNN-X101	Faster R-CNN, ResNeXt-101-FPN, 3×	104.8	5×10^{-5}	25,000 iters	4	1×A100
One-stage convolutional detectors [12, 14]						
Retina-R50-1×	RetinaNet, ResNet-50-FPN, 1×	38.0	5×10^{-5}	25,000 iters	4	1×A100
Retina-R50-3×	RetinaNet, ResNet-50-FPN, 3×	38.0	5×10^{-5}	25,000 iters	4	1×A100
Retina-R101-3×	RetinaNet, ResNet-101-FPN, 3×	57.1	5×10^{-5}	25,000 iters	4	1×A100
YOLO-family detectors [15, 16]						
YOLOv8	YOLOv8-X	68.2	8.7×10^{-6}	400 epochs	14	2×A100
YOLOv11	YOLOv11-X	56.9	8.7×10^{-6}	400 epochs	14	2×A100
YOLOv12	YOLOv12-X	59.1	8.7×10^{-6}	400 epochs	14	2×A100
YOLOv26	YOLOv26-X	58.8	8.7×10^{-6}	400 epochs	14	2×A100
Transformer detectors: RT-DETRv2 [18], RT-DETRv4 [19, 21]						
RT-DETRv2	RT-DETRv2, ResNet-50	42.9	5×10^{-5}	50 epochs	8	1×A100
RT-DETRv4	RT-DETRv4-L, HGNetV2-B4, DINOv3 teacher	31.4	2.5×10^{-4}	30 epochs	16	1×H100
Foundation-model-derived detectors: GDINO-SB [17], SAM3-OD [20]						
GDINO-SB	Grounding-DINO, Swin-Base	232.8	2×10^{-5}	50 epochs	4	1×A100
SAM3-OD	SAM3, ODinW text-only configuration	839.4	8×10^{-6}	60 epochs	32	2×H100

Table S8: Model-specific confidence thresholds, selected on OD-Valid by species-balanced count MAE (the mean of the per-species MAEs, so each species counts the same) and fixed before testing. The shared threshold, one value for all detectors, is 0.254. Section S2 gives the grid, the tie rule, the bootstrap interval and the 1% plateau. Score floor: the lowest confidence score a detector predicts. The MAE columns are species-balanced too, so they differ from the image-weighted MAE of Table S10.

Model	Score floor	Selected τ	95% interval	1% plateau	**OD-Valid count MAE**	
					at selected τ	at $\tau = 0.25$
Two-stage convolutional detectors						
FRCNN-R50-1×	0.001	0.199	[0.114, 0.371]	[0.178, 0.209]	8.28	8.41
FRCNN-R50-3×	0.001	0.105	[0.057, 0.167]	[0.098, 0.165]	7.82	9.07
FRCNN-R101-1×	0.001	0.129	[0.056, 0.166]	[0.112, 0.130]	7.44	8.67
FRCNN-X101	0.001	0.003	[0.002, 0.216]	[0.003, 0.003]	8.38	9.84
One-stage convolutional detectors						
Retina-R50-1×	0.001	0.090	[0.088, 0.113]	[0.088, 0.094]	6.84	16.19
Retina-R50-3×	0.001	0.111	[0.096, 0.116]	[0.110, 0.112]	7.47	18.14
Retina-R101-3×	0.001	0.127	[0.117, 0.134]	[0.126, 0.128]	8.30	15.52
YOLO-family detectors						
YOLOv8	0.001	0.189	[0.175, 0.231]	[0.175, 0.215]	5.32	6.08
YOLOv11	0.001	0.227	[0.192, 0.285]	[0.226, 0.239]	5.76	5.86
YOLOv12	0.001	0.199	[0.189, 0.273]	[0.195, 0.201]	5.80	6.35
YOLOv26	0.001	0.219	[0.197, 0.322]	[0.213, 0.255]	5.98	6.01
Transformer detectors						
RT-DETRv2	0.001	0.407	[0.365, 0.438]	[0.404, 0.411]	7.09	15.20
RT-DETRv4	0.001	0.388	[0.366, 0.427]	[0.385, 0.397]	5.94	13.04
Foundation-model-derived detectors						
GDINO-SB	0.250	0.299	[0.268, 0.333]	[0.297, 0.300]	6.03	7.40
SAM3-OD	0.500	0.500	[0.500, 0.518]	[0.500, 0.503]	6.10	6.10

S3 Detection and counting at the fixed threshold

This section gives the full results at the fixed threshold $\tau = 0.25$. Table S9 completes Table 3, and Table S10 gives the values behind Figure 2. Tables S11–S15 split detection into true positives, false positives, false negatives and false-positive types. Table S16 gives count bias and density error. Tables S17 and S18 relax box matching to IoU 0.30, and Table S18 adds Gibbs wheat density R^2.

Table S9: COCO box AP_{75} and AR@100 (%) on OD-Valid and the four held-out test sets, completing Table 3. Best **bold**, second-best underlined.

Model	OD-Valid		T-MAWP		T-DCWP		T-MSWP		T-GSH	
	AP_{75}	AR@100	AP_{75}	AR@100	AP_{75}	AR@100	AP_{75}	AR@100	AP_{75}	AR@100
Two-stage convolutional detectors										
FRCNN-R50-1×	33.42	49.89	20.91	44.61	13.38	36.34	18.52	31.95	3.25	18.20
FRCNN-R50-3×	34.97	49.61	20.33	42.71	11.43	34.57	15.29	29.98	2.86	14.41
FRCNN-R101-1×	32.12	49.10	20.89	44.03	2.64	20.86	17.07	31.52	1.57	13.53
FRCNN-X101	32.60	47.55	18.36	42.77	7.03	28.65	17.86	30.84	1.26	10.53
One-stage convolutional detectors										
Retina-R50-1×	35.45	48.91	23.65	43.96	5.97	36.64	18.19	36.08	4.39	18.76
Retina-R50-3×	35.44	48.59	22.83	42.23	8.63	36.10	13.73	35.44	4.62	20.22
Retina-R101-3×	32.58	49.08	20.48	44.33	1.40	28.33	10.77	32.94	5.01	19.65
YOLO-family detectors										
YOLOv8	42.96	55.97	26.69	48.95	5.07	27.97	24.94	36.24	2.90	14.00
YOLOv11	45.22	56.50	28.46	50.27	19.09	49.05	48.74	**55.69**	7.67	23.56
YOLOv12	**45.97**	**57.04**	27.55	**50.44**	10.99	36.40	21.96	39.20	7.84	21.98
YOLOv26	43.79	56.21	28.61	49.96	18.92	47.26	30.34	44.13	6.60	26.78
Transformer detectors										
RT-DETRv2	38.83	55.81	24.15	49.36	19.31	44.94	19.92	37.85	6.82	23.85
RT-DETRv4	43.83	56.91	**29.16**	50.11	16.70	43.23	21.02	40.07	9.29	29.39
Foundation-model-derived detectors										
GDINO-SB	35.09	51.12	22.69	43.26	44.49	55.21	**51.80**	52.79	12.00	30.22
SAM3-OD	35.25	50.77	24.09	45.48	**48.22**	**58.01**	20.07	35.37	**14.36**	**35.61**

Table S10: F1 (IoU 0.50) and image-weighted count MAE (the mean over images) at confidence threshold $\tau = 0.25$, the fixed-threshold values of Figure 2. Best **bold**, second-best underlined.

Model	OD-Valid		T-MAWP		T-DCWP		T-MSWP		T-GSH	
	F1↑	MAE↓	F1↑	MAE↓	F1↑	MAE↓	F1↑	MAE↓	F1↑	MAE↓
Two-stage convolutional detectors										
FRCNN-R50-1×	0.885	7.32	0.739	11.86	0.397	36.84	0.577	13.67	0.336	26.60
FRCNN-R50-3×	0.875	8.98	0.724	10.02	0.449	15.61	0.511	21.08	0.123	38.27
FRCNN-R101-1×	0.874	8.54	0.739	8.72	0.200	38.50	0.595	16.64	0.172	36.40
FRCNN-X101	0.866	10.18	0.751	8.06	0.416	11.72	0.538	17.88	0.163	31.68
One-stage convolutional detectors										
Retina-R50-1×	0.818	16.72	0.756	8.49	0.412	21.39	0.516	18.59	0.329	27.57
Retina-R50-3×	0.800	18.80	0.731	9.62	0.457	16.57	0.532	12.83	0.347	28.03
Retina-R101-3×	0.818	16.14	0.760	**7.88**	0.311	24.25	0.491	12.07	0.298	28.10
YOLO-family detectors										
YOLOv8	0.864	5.70	0.757	8.48	0.327	15.27	0.514	23.61	0.218	30.86
YOLOv11	0.875	5.25	**0.778**	9.11	0.651	7.36	0.818	**4.98**	0.465	22.31
YOLOv12	0.876	5.58	0.773	9.71	0.541	11.78	0.713	14.03	0.390	28.62
YOLOv26	0.872	5.28	0.771	8.32	0.653	7.65	0.683	14.12	0.523	22.07
Transformer detectors										
RT-DETRv2	0.873	10.60	0.697	23.57	0.492	29.44	0.654	11.62	0.475	25.13
RT-DETRv4	0.886	9.71	0.713	22.39	0.572	21.97	0.727	7.79	0.590	17.67
Foundation-model-derived detectors										
GDINO-SB	**0.898**	6.36	0.750	10.91	0.721	13.34	**0.867**	6.88	0.705	7.20
SAM3-OD	0.895	**5.22**	0.766	11.11	**0.826**	**6.48**	0.686	7.19	**0.795**	**5.71**

Table S11: Detection reliability on OD-Valid at $\tau = 0.25$ and IoU 0.50 (138 scene-primary images, 8,792 stomata). Predictions with score ≥ 0.25 (Pred.) are matched one-to-one, in descending score order, to the unmatched stoma with the highest IoU. IoU ≥ 0.50 gives a true positive (TP), an unmatched prediction a false positive (FP), an unmatched stoma a false negative (FN). P = TP/(TP+FP), R = TP/(TP+FN), F1 their harmonic mean. FP types: background (FP_{bg}, IoU < 0.10 with every stoma), localization (FP_{loc}, best IoU 0.10 to below 0.50), duplicate (FP_{dup}, IoU ≥ 0.50 only with already-matched stomata). Best P, R, F1 **bold**, second-best distinct <u>underlined</u>.

Model	OD-Valid									
	Pred.	TP	FP	FN	P↑	R↑	F1↑	FP_{bg}	FP_{loc}	FP_{dup}
Two-stage convolutional detectors										
FRCNN-R50-1×	8,240	7,536	704	1,256	0.915	0.857	0.885	610	94	0
FRCNN-R50-3×	7,755	7,237	518	1,555	0.933	0.823	0.875	417	101	0
FRCNN-R101-1×	7,744	7,230	514	1,562	0.934	0.822	0.874	393	121	0
FRCNN-X101	7,467	7,044	423	1,748	0.943	0.801	0.866	320	103	0
One-stage convolutional detectors										
Retina-R50-1×	6,579	6,289	290	2,503	<u>0.956</u>	0.715	0.818	214	76	0
Retina-R50-3×	6,291	6,030	261	2,762	**0.959**	0.686	0.800	193	68	0
Retina-R101-3×	6,659	6,320	339	2,472	0.949	0.719	0.818	234	105	0
YOLO-family detectors										
YOLOv8	8,363	7,410	953	1,382	0.886	0.843	0.864	874	75	4
YOLOv11	8,678	7,646	1,032	1,146	0.881	0.870	0.875	958	71	3
YOLOv12	8,526	7,582	944	1,210	0.889	0.862	0.876	873	68	3
YOLOv26	8,662	7,614	1,048	1,178	0.879	0.866	0.872	946	80	22
Transformer detectors										
RT-DETRv2	10,007	8,206	1,801	586	0.820	<u>0.933</u>	0.873	1,567	126	108
RT-DETRv4	9,970	8,308	1,662	484	0.833	**0.945**	0.886	1,399	120	143
Foundation-model-derived detectors										
GDINO-SB	8,594	7,803	791	989	0.908	0.888	**0.898**	658	85	48
SAM3-OD	8,619	7,788	831	1,004	0.904	0.886	<u>0.895</u>	757	70	4

Table S12: Detection reliability on T-MAWP at $\tau = 0.25$ and IoU 0.50 (456 scene-primary images, 17,164 stomata). Columns as in Table S11.

Model	T-MAWP									
	Pred.	TP	FP	FN	P↑	R↑	F1↑	FP_{bg}	FP_{loc}	FP_{dup}
Two-stage convolutional detectors										
FRCNN-R50-1×	20,701	13,988	6,713	3,176	0.676	0.815	0.739	6,030	675	8
FRCNN-R50-3×	15,860	11,951	3,909	5,213	0.754	0.696	0.724	3,357	537	15
FRCNN-R101-1×	16,547	12,464	4,083	4,700	0.753	0.726	0.739	3,611	463	9
FRCNN-X101	15,677	12,339	3,338	4,825	**0.787**	0.719	0.751	2,825	490	23
One-stage convolutional detectors										
Retina-R50-1×	16,310	12,651	3,659	4,513	<u>0.776</u>	0.737	0.756	3,260	397	2
Retina-R50-3×	15,759	12,033	3,726	5,131	0.764	0.701	0.731	3,269	453	4
Retina-R101-3×	17,503	13,168	4,335	3,996	0.752	0.767	0.760	3,890	437	8
YOLO-family detectors										
YOLOv8	16,622	12,786	3,836	4,378	0.769	0.745	0.757	3,472	364	0
YOLOv11	19,724	14,352	5,372	2,812	0.728	0.836	**0.778**	4,846	518	8
YOLOv12	18,966	13,962	5,004	3,202	0.736	0.813	<u>0.773</u>	4,458	540	6
YOLOv26	17,970	13,538	4,432	3,626	0.753	0.789	0.771	3,962	441	29
Transformer detectors										
RT-DETRv2	27,890	15,703	12,187	1,461	0.563	<u>0.915</u>	0.697	11,043	829	315
RT-DETRv4	27,366	15,880	11,486	1,284	0.580	**0.925**	0.713	10,397	784	305
Foundation-model-derived detectors										
GDINO-SB	21,351	14,442	6,909	2,722	0.676	0.841	0.750	5,793	1,011	105
SAM3-OD	21,981	14,995	6,986	2,169	0.682	0.874	0.766	6,162	814	10

Table S13: Detection reliability on T-DCWP at $\tau = 0.25$ and IoU 0.50 (119 scene-primary images, 2,858 stomata). Columns as in Table S11.

Model	T-DCWP									
	Pred.	TP	FP	FN	P↑	R↑	F1↑	FP_{bg}	FP_{loc}	FP_{dup}
Two-stage convolutional detectors										
FRCNN-R50-1×	7,192	1,994	5,198	864	0.277	0.698	0.397	4,648	545	5
FRCNN-R50-3×	4,037	1,549	2,488	1,309	0.384	0.542	0.449	2,087	401	0
FRCNN-R101-1×	7,193	1,004	6,189	1,854	0.140	0.351	0.200	5,547	641	1
FRCNN-X101	3,637	1,350	2,287	1,508	0.371	0.472	0.416	1,776	510	1
One-stage convolutional detectors										
Retina-R50-1×	5,257	1,671	3,586	1,187	0.318	0.585	0.412	2,856	712	18
Retina-R50-3×	4,726	1,734	2,992	1,124	0.367	0.607	0.457	2,263	690	39
Retina-R101-3×	5,724	1,333	4,391	1,525	0.233	0.466	0.311	3,006	1,358	27
YOLO-family detectors										
YOLOv8	2,233	833	1,400	2,025	0.373	0.291	0.327	1,155	233	12
YOLOv11	3,290	2,000	1,290	858	0.608	0.700	0.651	1,008	224	58
YOLOv12	2,330	1,403	927	1,455	0.602	0.491	0.541	709	214	4
YOLOv26	2,628	1,792	836	1,066	<u>0.682</u>	0.627	0.653	652	160	24
Transformer detectors										
RT-DETRv2	6,295	2,252	4,043	606	0.358	0.788	0.492	3,564	408	71
RT-DETRv4	5,426	2,368	3,058	490	0.436	0.829	0.572	2,513	413	132
Foundation-model-derived detectors										
GDINO-SB	4,384	2,609	1,775	249	0.595	<u>0.913</u>	<u>0.721</u>	1,465	230	80
SAM3-OD	3,615	2,674	941	184	**0.740**	**0.936**	**0.826**	701	192	48

Table S14: Detection reliability on T-MSWP at $\tau = 0.25$ and IoU 0.50 (10,284 scene-primary images, 503,898 stomata). Columns as in Table S11.

Model	T-MSWP									
	Pred.	TP	FP	FN	P↑	R↑	F1↑	FP_{bg}	FP_{loc}	FP_{dup}
Two-stage convolutional detectors										
FRCNN-R50-1×	374,416	253,461	120,955	250,437	0.677	0.503	0.577	28,962	91,931	62
FRCNN-R50-3×	287,417	202,228	85,189	301,670	0.704	0.401	0.511	8,204	76,956	29
FRCNN-R101-1×	338,425	250,579	87,846	253,319	0.740	0.497	0.595	28,613	59,225	8
FRCNN-X101	322,895	222,459	100,436	281,439	0.689	0.441	0.538	25,393	75,008	35
One-stage convolutional detectors										
Retina-R50-1×	314,204	211,274	102,930	292,624	0.672	0.419	0.516	16,596	85,994	340
Retina-R50-3×	375,953	234,201	141,752	269,697	0.623	0.465	0.532	21,599	119,682	471
Retina-R101-3×	395,977	220,954	175,023	282,944	0.558	0.438	0.491	24,984	148,543	1,496
YOLO-family detectors										
YOLOv8	262,148	196,697	65,451	307,201	0.750	0.390	0.514	18,048	39,735	7,668
YOLOv11	470,680	398,794	71,886	105,104	0.847	<u>0.791</u>	<u>0.818</u>	18,723	43,310	9,853
YOLOv12	360,185	307,893	52,292	196,005	<u>0.855</u>	0.611	0.713	8,161	43,670	461
YOLOv26	360,138	295,084	65,054	208,814	0.819	0.586	0.683	11,300	51,133	2,621
Transformer detectors										
RT-DETRv2	392,283	293,194	99,089	210,704	0.747	0.582	0.654	24,811	67,422	6,856
RT-DETRv4	435,979	341,470	94,509	162,428	0.783	0.678	0.727	25,351	61,415	7,743
Foundation-model-derived detectors										
GDINO-SB	440,835	409,689	31,146	94,209	**0.929**	**0.813**	**0.867**	11,987	14,871	4,288
SAM3-OD	440,069	323,601	116,468	180,297	0.735	0.642	0.686	11,347	103,751	1,370

Table S15: Detection reliability on T-GSH at $\tau = 0.25$ and IoU 0.50 (6,943 scene-primary images, 288,543 stomata). Columns as in Table S11.

Model	T-GSH									
	Pred.	TP	FP	FN	P↑	R↑	F1↑	FP_{bg}	FP_{loc}	FP_{dup}
Two-stage convolutional detectors										
FRCNN-R50-1×	151,019	73,742	77,277	214,801	0.488	0.256	0.336	63,733	13,515	29
FRCNN-R50-3×	22,866	19,098	3,768	269,445	**0.835**	0.066	0.123	2,435	1,319	14
FRCNN-R101-1×	175,493	40,023	135,470	248,520	0.228	0.139	0.172	126,903	8,495	72
FRCNN-X101	78,669	29,940	48,729	258,603	0.381	0.104	0.163	40,051	8,564	114
One-stage convolutional detectors										
Retina-R50-1×	159,599	73,780	85,819	214,763	0.462	0.256	0.329	75,039	10,743	37
Retina-R50-3×	113,612	69,728	43,884	218,815	0.614	0.242	0.347	33,489	10,322	73
Retina-R101-3×	146,773	64,896	81,877	223,647	0.442	0.225	0.298	71,414	10,334	129
YOLO-family detectors										
YOLOv8	167,202	49,634	117,568	238,909	0.297	0.172	0.218	92,584	24,944	40
YOLOv11	158,275	103,988	54,287	184,555	0.657	0.360	0.465	32,867	21,267	153
YOLOv12	108,850	77,569	31,281	210,974	0.713	0.269	0.390	20,945	10,283	53
YOLOv26	206,644	129,443	77,201	159,100	0.626	0.449	0.523	57,713	17,084	2,404
Transformer detectors										
RT-DETRv2	273,488	133,597	139,891	154,946	0.488	0.463	0.475	103,711	26,379	9,801
RT-DETRv4	205,437	145,761	59,676	142,782	0.710	0.505	0.590	46,359	9,553	3,764
Foundation-model-derived detectors										
GDINO-SB	283,028	201,546	81,482	86,997	0.712	<u>0.698</u>	<u>0.705</u>	37,150	40,212	4,120
SAM3-OD	277,134	224,720	52,414	63,823	<u>0.811</u>	**0.779**	**0.795**	24,046	28,192	176

Table S16: Signed count bias and calibrated stomatal-density error at confidence threshold $\tau = 0.25$. Bias is predicted minus annotated stomata per image, so positive means overcounting. Density MAE (stomata mm^{-2}) uses the calibrated images of Table 1. Best Density MAE↓ values are **bold** and second-best distinct values are <u>underlined</u>.

Model	OD-Valid		T-MAWP		T-DCWP		T-MSWP		T-GSH	
	Bias	Density MAE↓	Bias	Density MAE↓	Bias	Density MAE↓	Bias	Density MAE↓	Bias	Density MAE↓
Two-stage convolutional detectors										
FRCNN-R50-1×	−4.00	15.57	7.76	25.24	36.42	392.94	−12.59	178.80	−19.81	37.87
FRCNN-R50-3×	−7.51	19.10	−2.86	21.31	9.91	162.99	−21.05	264.49	−38.27	43.86
FRCNN-R101-1×	−7.59	18.16	−1.35	18.55	36.43	412.06	−16.09	237.14	−16.28	60.82
FRCNN-X101	−9.60	21.66	−3.26	<u>17.16</u>	6.55	125.26	−17.60	216.87	−30.23	38.45
One-stage convolutional detectors										
Retina-R50-1×	−16.04	35.57	−1.87	18.06	20.16	229.46	−18.45	236.58	−18.57	37.01
Retina-R50-3×	−18.12	40.01	−3.08	20.47	15.70	176.69	−12.44	148.48	−25.20	30.51
Retina-R101-3×	−15.46	34.34	0.74	**16.77**	24.08	260.94	−10.49	144.13	−20.42	25.89
YOLO-family detectors										
YOLOv8	−3.11	12.13	−1.19	18.05	−5.25	155.60	−23.51	334.88	−17.48	89.43
YOLOv11	−0.83	<u>11.16</u>	5.61	19.38	3.63	<u>74.64</u>	−3.23	**72.34**	−18.76	34.98
YOLOv12	−1.93	11.87	3.95	20.67	−4.44	123.61	−13.97	212.03	−25.88	36.74
YOLOv26	−0.94	11.22	1.77	17.71	−1.93	77.25	−13.98	214.57	−11.80	73.67
Transformer detectors										
RT-DETRv2	8.80	22.56	23.52	50.14	28.88	306.79	−10.85	128.26	−2.17	33.80
RT-DETRv4	8.54	20.66	22.37	47.63	21.58	215.91	−6.60	102.06	−11.97	29.70
Foundation-model-derived detectors										
GDINO-SB	−1.43	13.54	9.18	23.22	12.82	141.65	−6.13	<u>88.58</u>	−0.79	<u>22.99</u>
SAM3-OD	−1.25	**11.12**	10.56	23.63	6.36	**68.71**	−6.21	89.27	−1.64	**17.02**

Table S17: Recall and F1 at $\tau = 0.25$ when one-to-one matching is relaxed from IoU 0.50 to 0.30. $R_{0.50}$ is the recall of Tables S11–S15, and $F1_{0.50}$ is in Table S10. Best **bold**, second-best distinct underlined.

Model	OD-Valid			T-MAWP			T-DCWP			T-MSWP			T-GSH		
	$R_{0.50}$	$R_{0.30}$	$F1_{0.30}$	$R_{0.50}$	$R_{0.30}$	$F1_{0.30}$	$R_{0.50}$	$R_{0.30}$	$F1_{0.30}$	$R_{0.50}$	$R_{0.30}$	$F1_{0.30}$	$R_{0.50}$	$R_{0.30}$	$F1_{0.30}$
Two-stage convolutional detectors															
FRCNN-R50-1×	0.857	0.866	0.895	0.815	0.848	0.769	0.698	0.835	0.475	0.503	0.622	0.713	0.256	0.290	0.380
FRCNN-R50-3×	0.823	0.834	0.886	0.696	0.723	0.751	0.542	0.655	0.543	0.401	0.485	0.617	0.066	0.070	0.130
FRCNN-R101-1×	0.822	0.835	0.888	0.726	0.748	0.762	0.351	0.499	0.284	0.497	0.552	0.660	0.139	0.153	0.190
FRCNN-X101	0.801	0.812	0.878	0.719	0.744	0.778	0.472	0.603	0.531	0.441	0.524	0.639	0.104	0.125	0.197
One-stage convolutional detectors															
Retina-R50-1×	0.715	0.723	0.828	0.737	0.757	0.776	0.585	0.750	0.528	0.419	0.496	0.611	0.256	0.286	0.368
Retina-R50-3×	0.686	0.693	0.808	0.701	0.724	0.755	0.607	0.779	0.587	0.465	0.570	0.652	0.242	0.271	0.388
Retina-R101-3×	0.719	0.730	0.831	0.767	0.789	0.781	0.466	0.744	0.495	0.438	0.556	0.622	0.225	0.250	0.332
YOLO-family detectors															
YOLOv8	0.843	0.850	0.871	0.745	0.764	0.777	0.291	0.355	0.399	0.390	0.440	0.579	0.172	0.242	0.306
YOLOv11	0.870	0.877	0.882	0.836	0.864	**0.804**	0.700	0.765	0.711	0.791	**0.853**	**0.882**	0.360	0.432	0.558
YOLOv12	0.862	0.869	0.882	0.813	0.843	0.801	0.491	0.551	0.607	0.611	0.684	0.798	0.269	0.303	0.440
YOLOv26	0.866	0.874	0.880	0.789	0.813	0.794	0.627	0.679	0.708	0.586	0.668	0.779	0.449	0.504	0.587
Transformer detectors															
RT-DETRv2	0.933	0.942	0.881	0.915	0.950	0.724	0.788	0.879	0.549	0.582	0.667	0.750	0.463	0.519	0.532
RT-DETRv4	**0.945**	**0.953**	0.894	**0.925**	**0.961**	0.741	0.829	0.924	0.638	0.678	0.777	0.833	0.505	0.533	0.622
Foundation-model-derived detectors															
GDINO-SB	0.888	0.894	**0.904**	0.841	0.891	0.794	0.913	0.964	0.761	**0.813**	0.827	**0.882**	0.698	0.815	0.822
SAM3-OD	0.886	0.892	0.901	0.874	0.917	**0.804**	**0.936**	**0.988**	**0.872**	0.642	0.757	0.809	**0.779**	**0.871**	**0.889**

Table S18: Gibbs wheat (T-GSH, 348 calibrated scene-primary images, 1,884 annotated stomata): predicted boxes at $\tau = 0.25$, true and false positives (TP + FP = Pred.) at IoU 0.30 and 0.50, and wheat's within-species density R^2 at the fixed threshold 0.25, the shared threshold 0.254, the model-specific thresholds (MS) of Table S8 and the fixed threshold 0.50. Best **bold**, second-best distinct underlined.

Model	Pred.	IoU 0.30				IoU 0.50				Density R^2 at τ			
		TP	FP	R	P	TP	FP	R	P	0.25	0.254	MS	0.50
Two-stage convolutional detectors													
FRCNN-R50-1×	3,349	0	3,349	0.000	0.000	0	3,349	0.000	0.000	−25.96	−24.89	−48.04	−4.47
FRCNN-R50-3×	90	0	90	0.000	0.000	0	90	0.000	0.000	−4.17	−4.18	−3.61	−4.41
FRCNN-R101-1×	6,258	0	6,258	0.000	0.000	0	6,258	0.000	0.000	−88.98	−84.90	−297.93	−5.11
FRCNN-X101	351	0	351	0.000	0.000	0	351	0.000	0.000	**−3.54**	**−3.55**	−234.21	−4.05
One-stage convolutional detectors													
Retina-R50-1×	2,830	0	2,830	0.000	0.000	0	2,830	0.000	0.000	−13.25	−12.01	−582.44	−4.10
Retina-R50-3×	877	0	877	0.000	0.000	0	877	0.000	0.000	−3.59	−3.56	−88.02	−4.34
Retina-R101-3×	1,736	0	1,736	0.000	0.000	0	1,736	0.000	0.000	−5.03	−4.65	−205.95	−4.26
YOLO-family detectors													
YOLOv8	9,217	0	9,217	0.000	0.000	0	9,217	0.000	0.000	−279.26	−262.56	−734.35	−5.94
YOLOv11	1,063	1	1,062	0.001	0.001	1	1,062	0.001	0.001	−6.01	−5.78	−7.19	−4.10
YOLOv12	160	0	160	0.000	0.000	0	160	0.000	0.000	−4.08	−4.10	−3.80	−4.42
YOLOv26	8,195	2	8,193	0.001	0.000	1	8,194	0.001	0.000	−286.35	−272.29	−411.50	−12.46
Transformer detectors													
RT-DETRv2	2,675	1	2,674	0.001	0.000	0	2,675	0.000	0.000	−16.08	−14.68	**−2.94**	−3.35
RT-DETRv4	2,118	19	2,099	0.010	0.009	12	2,106	0.006	0.006	−18.97	−17.00	−3.40	−3.89
Foundation-model-derived detectors													
GDINO-SB	3,860	1,414	2,446	0.751	0.366	403	3,457	**0.214**	**0.104**	−7.49	−7.34	−5.87	**−2.78**
SAM3-OD	3,436	1,489	1,947	**0.790**	**0.433**	252	3,184	0.134	0.073	−8.15	−8.15	−8.15	−8.15

S4 Detection, count, and density reliability at other confidence thresholds

Counts and densities change with the confidence threshold, so this section repeats the set-level results at the four other settings of Figure 2. Two are selected on OD-Valid (Table S8): the shared threshold, one value for all detectors, and the model-specific threshold of each detector. The other two are the fixed thresholds $\tau = 0.50$ and $\tau = 0.75$. Figures S6–S11 print the values in the layout of Figure 2, one metric per figure: F1, count MAE, false positives, false negatives, signed count bias and density MAE. F1, false positives and false negatives use IoU 0.50, and the values at $\tau = 0.25$ are in Section S3.

F1 (IoU 0.50) — higher is better (0 – 0.5 – 1)

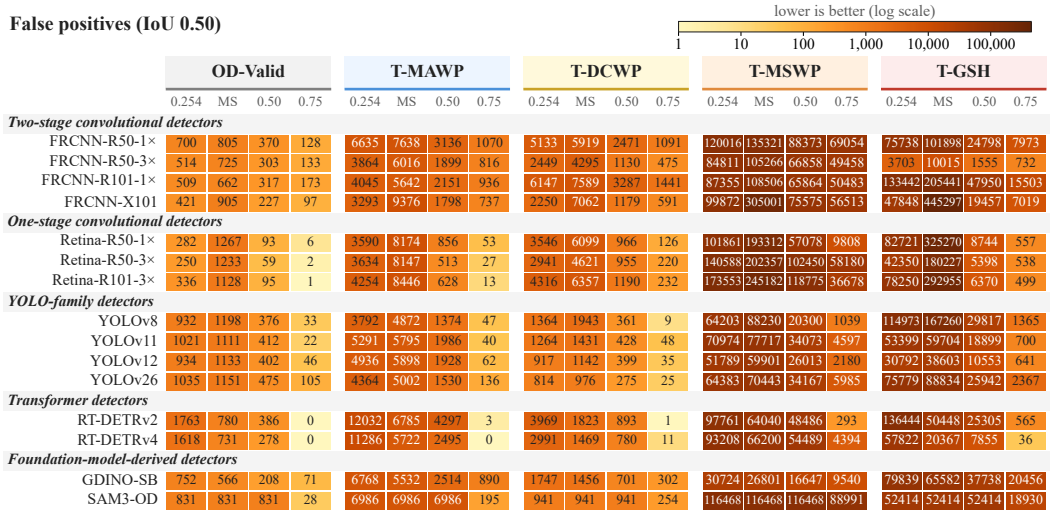
	OD-Valid				T-MAWP				T-DCWP				T-MSWP				T-GSH			
	0.254	MS	0.50	0.75	0.254	MS	0.50	0.75	0.254	MS	0.50	0.75	0.254	MS	0.50	0.75	0.254	MS	0.50	0.75
Two-stage convolutional detectors																				
FRCNN-R50-1×	0.885	0.885	0.866	0.795	0.740	0.730	0.750	0.665	0.399	0.375	0.518	0.579	0.577	0.575	0.561	0.510	0.335	0.337	0.285	0.190
FRCNN-R50-3×	0.874	0.882	0.846	0.786	0.724	0.723	0.695	0.611	0.451	0.395	0.478	0.446	0.510	0.537	0.469	0.410	0.122	0.172	0.080	0.047
FRCNN-R101-1×	0.874	0.879	0.851	0.796	0.739	0.735	0.715	0.639	0.200	0.188	0.252	0.299	0.595	0.593	0.587	0.561	0.172	0.182	0.154	0.114
FRCNN-X101	0.866	0.866	0.830	0.751	0.752	0.705	0.715	0.621	0.418	0.296	0.440	0.398	0.538	0.490	0.515	0.467	0.162	0.179	0.117	0.071
One-stage convolutional detectors																				
Retina-R50-1×	0.817	0.858	0.748	0.523	0.756	0.707	0.660	0.315	0.413	0.368	0.446	0.131	0.515	0.569	0.377	0.117	0.328	0.312	0.183	0.029
Retina-R50-3×	0.798	0.849	0.708	0.436	0.731	0.687	0.564	0.229	0.459	0.408	0.500	0.221	0.531	0.563	0.435	0.247	0.344	0.377	0.191	0.036
Retina-R101-3×	0.817	0.856	0.743	0.389	0.760	0.712	0.614	0.174	0.312	0.308	0.339	0.104	0.489	0.545	0.350	0.071	0.297	0.296	0.170	0.036
YOLO-family detectors																				
YOLOv8	0.864	0.869	0.799	0.338	0.756	0.757	0.668	0.149	0.328	0.329	0.244	0.030	0.512	0.535	0.355	0.037	0.217	0.221	0.152	0.019
YOLOv11	0.875	0.877	0.824	0.320	0.779	0.776	0.736	0.133	0.651	0.645	0.602	0.148	0.818	0.820	0.742	0.304	0.464	0.472	0.346	0.089
YOLOv12	0.875	0.878	0.821	0.409	0.773	0.769	0.721	0.219	0.541	0.553	0.438	0.115	0.712	0.725	0.614	0.211	0.389	0.411	0.284	0.103
YOLOv26	0.872	0.873	0.822	0.528	0.771	0.770	0.702	0.316	0.654	0.654	0.527	0.165	0.682	0.695	0.543	0.173	0.522	0.527	0.417	0.128
Transformer detectors																				
RT-DETRv2	0.874	0.876	0.839	0.005	0.699	0.761	0.771	0.012	0.494	0.577	0.587	0.019	0.654	0.611	0.549	0.002	0.477	0.484	0.448	0.062
RT-DETRv4	0.887	0.899	0.862	0.052	0.716	0.787	0.785	0.004	0.576	0.668	0.679	0.063	0.727	0.709	0.667	0.040	0.589	0.496	0.363	0.005
Foundation-model-derived detectors																				
GDINO-SB	0.898	0.894	0.839	0.719	0.752	0.764	0.754	0.661	0.723	0.749	0.813	0.819	0.867	0.865	0.848	0.796	0.706	0.711	0.666	0.534
SAM3-OD	0.895	0.895	0.895	0.588	0.766	0.766	0.766	0.464	0.826	0.826	0.826	0.827	0.686	0.686	0.686	0.586	0.795	0.795	0.795	0.581

Figure S6: F1 at IoU 0.50 for every detector and set at the four settings other than $\tau = 0.25$: the shared threshold $\tau = 0.254$ and each detector's model-specific threshold (MS, Table S8), both selected on OD-Valid, and the fixed $\tau = 0.50$ and 0.75. Layout as Figure 2, values printed. The values at $\tau = 0.25$ are in Section S3. SAM3-OD releases no scores below 0.50, so its shared, MS and $\tau = 0.50$ values equal those at $\tau = 0.25$.

Count MAE (stomata per image) — lower is better (log scale) (5 – 10 – 20 – 50)

	OD-Valid				T-MAWP				T-DCWP				T-MSWP				T-GSH			
	0.254	MS	0.50	0.75	0.254	MS	0.50	0.75	0.254	MS	0.50	0.75	0.254	MS	0.50	0.75	0.254	MS	0.50	0.75
Two-stage convolutional detectors																				
FRCNN-R50-1×	7.33	7.06	11.19	20.22	11.73	13.54	9.06	16.08	36.31	43.04	14.99	7.88	13.72	13.12	18.02	23.22	26.60	27.06	30.82	35.96
FRCNN-R50-3×	9.06	7.20	13.66	20.97	10.03	11.08	13.15	18.81	15.43	27.98	10.10	12.53	21.17	17.13	25.52	30.33	38.30	36.31	39.60	40.44
FRCNN-R101-1×	8.57	7.12	12.57	19.49	8.73	9.97	11.30	17.23	38.22	49.76	16.87	9.80	16.69	14.95	19.68	23.07	36.26	41.23	32.87	36.77
FRCNN-X101	10.25	7.39	15.79	24.24	8.05	15.99	11.55	18.49	11.46	49.83	9.39	12.66	17.96	20.92	22.16	26.93	31.78	55.02	36.11	38.99
One-stage convolutional detectors																				
Retina-R50-1×	16.89	5.91	24.60	41.07	8.43	16.14	16.63	30.48	21.12	44.24	9.68	21.20	18.83	10.50	30.80	44.93	27.55	43.57	36.01	40.86
Retina-R50-3×	19.07	6.83	28.16	45.94	9.53	16.22	21.37	32.70	16.11	30.96	7.82	18.95	13.02	9.06	22.68	35.63	28.20	28.90	36.31	40.72
Retina-R101-3×	16.26	6.99	24.99	48.33	7.74	15.99	19.05	34.02	23.56	43.51	8.76	20.64	12.24	9.24	24.61	43.49	28.08	46.03	36.69	40.72
YOLO-family detectors																				
YOLOv8	5.80	5.16	17.07	50.48	8.43	9.32	15.00	34.50	15.34	16.61	17.83	23.58	23.84	20.00	36.01	47.97	30.87	31.83	34.87	40.96
YOLOv11	5.25	5.18	14.38	51.38	9.01	9.86	10.75	34.87	7.24	7.98	8.88	21.66	5.05	4.69	14.84	39.69	22.42	21.66	30.14	39.51
YOLOv12	5.63	5.24	14.67	46.91	9.62	10.76	11.71	32.86	11.75	11.82	14.92	22.24	14.17	12.16	23.63	42.98	28.71	27.38	33.54	39.21
YOLOv26	5.28	5.28	13.78	39.84	8.29	9.01	13.40	30.22	7.59	7.85	12.55	21.63	14.30	12.62	26.17	43.71	22.07	22.14	27.53	38.39
Transformer detectors																				
RT-DETRv2	10.36	6.82	13.36	63.54	23.20	10.60	6.99	37.40	28.81	12.18	8.52	23.77	11.80	18.58	23.98	48.91	24.89	22.62	25.67	40.15
RT-DETRv4	9.40	5.71	12.06	62.00	21.92	8.62	6.49	37.56	21.37	9.89	7.96	23.13	7.91	12.24	16.56	47.57	17.83	24.72	31.06	41.45
Foundation-model-derived detectors																				
GDINO-SB	6.20	6.59	15.09	27.16	10.67	8.54	7.84	16.21	13.11	10.55	4.11	3.64	6.93	7.50	10.22	15.10	7.16	7.41	13.65	22.67
SAM3-OD	5.22	5.22	5.22	36.88	11.11	11.11	11.11	25.73	6.48	6.48	6.48	4.71	7.19	7.19	7.19	16.51	5.71	5.71	5.71	20.98

Figure S7: Count MAE (stomata per image) at the settings of Figure S6. Count MAE needs no box matching.



	OD-Valid				T-MAWP				T-DCWP				T-MSWP				T-GSH			
	0.254	MS	0.50	0.75	0.254	MS	0.50	0.75	0.254	MS	0.50	0.75	0.254	MS	0.50	0.75	0.254	MS	0.50	0.75
Two-stage convolutional detectors																				
FRCNN-R50-1×	700	805	370	128	6635	7638	3136	1070	5133	5919	2471	1091	120016	135321	88373	69054	75738	101898	24798	7973
FRCNN-R50-3×	514	725	303	133	3864	6016	1899	816	2449	4295	1130	475	84811	105266	66858	49458	3703	10015	1555	732
FRCNN-R101-1×	509	662	317	173	4045	5642	2151	936	6147	7589	3287	1441	87355	108506	65864	50483	133442	205441	47950	15503
FRCNN-X101	421	905	227	97	3293	9376	1798	737	2250	7062	1179	591	99872	305001	75575	56513	47848	445297	19457	7019
One-stage convolutional detectors																				
Retina-R50-1×	282	1267	93	6	3590	8174	856	53	3546	6099	966	126	101861	193312	57078	9808	82721	325270	8744	557
Retina-R50-3×	250	1233	59	2	3634	8147	513	27	2941	4621	955	220	140588	202357	102450	58180	42350	180227	5398	538
Retina-R101-3×	336	1128	95	1	4254	8446	628	13	4316	6357	1190	232	173553	245182	118775	36678	78250	292955	6370	499
YOLO-family detectors																				
YOLOv8	932	1198	376	33	3792	4872	1374	47	1364	1943	361	9	64203	88230	20300	1039	114973	167260	29817	1365
YOLOv11	1021	1111	412	22	5291	5795	1986	40	1264	1431	428	48	70974	77717	34073	4597	53399	59704	18899	700
YOLOv12	934	1133	402	46	4936	5898	1928	62	917	1142	399	35	51789	59901	26013	2180	30792	38603	10553	641
YOLOv26	1035	1151	475	105	4364	5002	1530	136	814	976	275	25	64383	70443	34167	5985	75779	88834	25942	2367
Transformer detectors																				
RT-DETRv2	1763	780	386	0	12032	6785	4297	3	3969	1823	893	1	97761	64040	48486	293	136444	50448	25305	565
RT-DETRv4	1618	731	278	0	11286	5722	2495	0	2991	1469	780	11	93208	66200	54489	4394	57822	20367	7855	36
Foundation-model-derived detectors																				
GDINO-SB	752	566	208	71	6768	5532	2514	890	1747	1456	701	302	30724	26801	16647	9540	79839	65582	37738	20456
SAM3-OD	831	831	831	28	6986	6986	6986	195	941	941	941	254	116468	116468	116468	88991	52414	52414	52414	18930

Figure S8: False positives at IoU 0.50 at the settings of Figure S6. FP and FN are diagnostic counts, defined as in Table S11.



	OD-Valid				T-MAWP				T-DCWP				T-MSWP				T-GSH			
	0.254	MS	0.50	0.75	0.254	MS	0.50	0.75	0.254	MS	0.50	0.75	0.254	MS	0.50	0.75	0.254	MS	0.50	0.75
Two-stage convolutional detectors																				
FRCNN-R50-1×	1259	1179	1800	2906	3199	2920	4990	8084	864	833	995	1247	250816	245859	273272	307738	215181	209547	236440	257458
FRCNN-R50-3×	1564	1277	2124	3013	5241	4048	7022	9251	1311	1099	1606	1902	302161	280302	329187	361361	269594	260465	276478	281519
FRCNN-R101-1×	1568	1378	2050	2863	4736	3900	6412	8674	1855	1772	1972	2103	253549	245772	267282	287577	248778	239224	260451	270218
FRCNN-X101	1755	1387	2394	3442	4844	2709	6611	9100	1510	1132	1719	2002	281794	241183	303030	333379	258830	216479	269445	277723
One-stage convolutional detectors																				
Retina-R50-1×	2523	1236	3488	5674	4543	3317	8296	13950	1193	835	1760	2649	294080	226684	373798	471895	215645	175225	258660	284265
Retina-R50-3×	2794	1403	3945	6342	5172	3926	10223	14938	1132	941	1588	2475	270845	227226	335584	424610	219714	179567	257451	283254
Retina-R101-3×	2490	1370	3543	6670	4025	3010	9288	15528	1532	1181	2032	2688	284560	223217	371830	483883	224518	187546	261085	283192
YOLO-family detectors																				
YOLOv8	1397	1124	2694	6999	4419	3746	7867	15777	2029	1914	2411	2815	308437	287916	390623	494333	239352	232056	262275	285780
YOLOv11	1162	1060	2346	7113	2839	2615	6008	15942	871	815	1443	2626	106077	99542	186480	412786	185204	180825	224319	275007
YOLOv12	1227	1020	2386	6520	3238	2753	6400	15044	1459	1331	1944	2681	196989	183302	269028	444148	211493	203837	239070	272846
YOLOv26	1196	1086	2320	5603	3668	3298	7058	13915	1075	996	1738	2599	210132	197988	303218	455527	159803	153558	205595	268722
Transformer detectors																				
RT-DETRv2	594	1339	2158	8769	1475	2453	3689	17059	619	961	1299	2830	211669	253909	294774	503270	155470	180179	197878	279312
RT-DETRv4	493	1019	1922	8556	1297	2321	4451	17126	494	690	989	2764	163146	190828	224569	493554	144073	186534	222724	287821
Foundation-model-derived detectors																				
GDINO-SB	1009	1233	2282	3819	2747	3142	5267	8245	249	274	422	665	94646	99330	120855	164494	87446	93056	125602	175887
SAM3-OD	1004	1004	1004	5117	2169	2169	2169	11925	184	184	184	666	180297	180297	180297	258191	63823	63823	63823	162774

Figure S9: False negatives at IoU 0.50 at the settings of Figure S6.

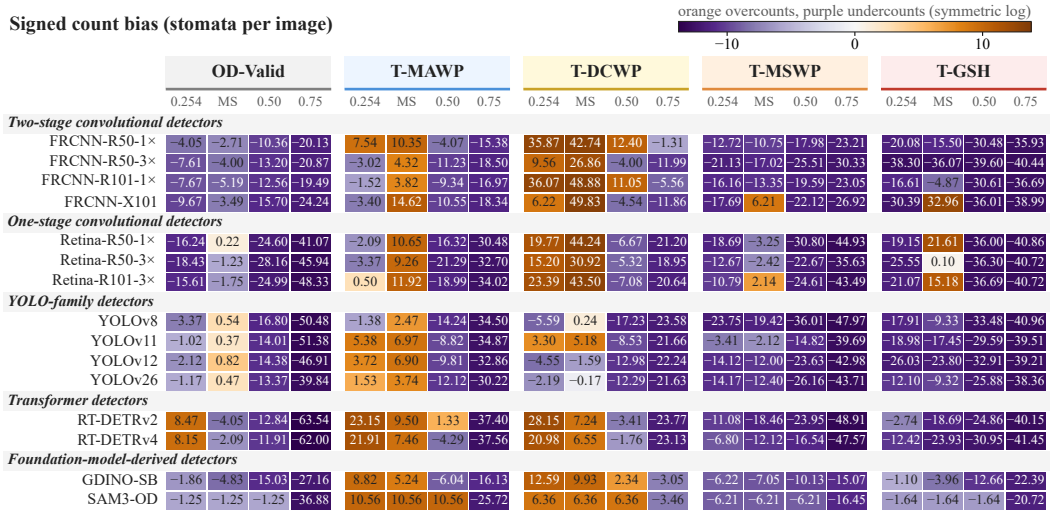


Figure S10: Signed count bias at the settings of Figure S6: predicted minus annotated stomata per image, so positive means overcounting.

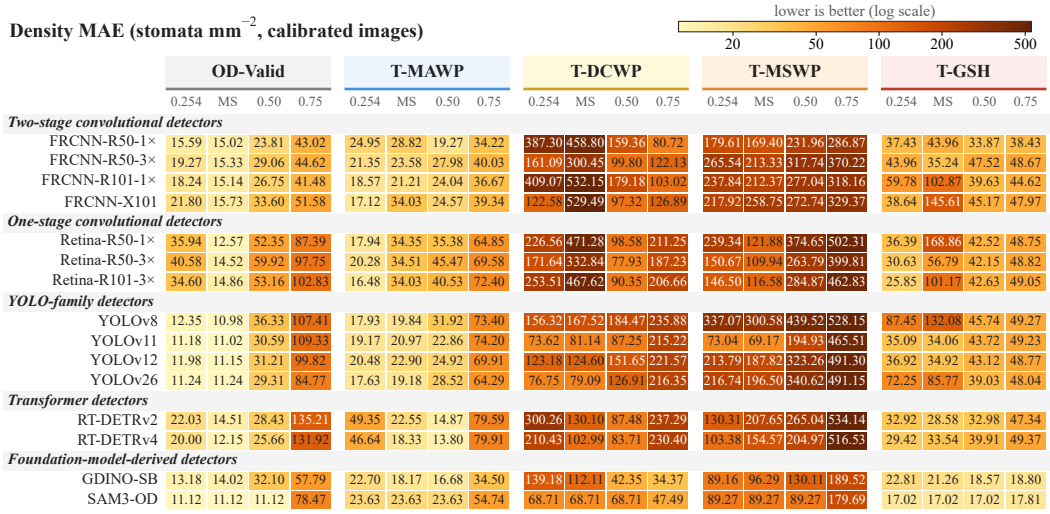


Figure S11: Density MAE (stomata mm^{-2}) at the settings of Figure S6, on the calibrated images of Table 1.

S5 Uncertainty of the set-level results

Figures S12 and S13 show count MAE and density MAE at $\tau = 0.25$ for every detector and set, each with a 95% paired-image bootstrap interval that shows how much the value changes when the test images are resampled. The resamples are paired across detectors, so all detectors are scored on the same resampled images. Overlapping intervals are not a test of whether two detectors differ, and we use the intervals only to show how stable each value is. The resamples draw images, not plants. In T-MAWP 15 of the 17 species, and in OD-Valid 4 of the 5, were imaged from a single accession, one plant or planting of the living collection, so the intervals leave out the variation between plants of a species. The exact intervals are in Section S6.

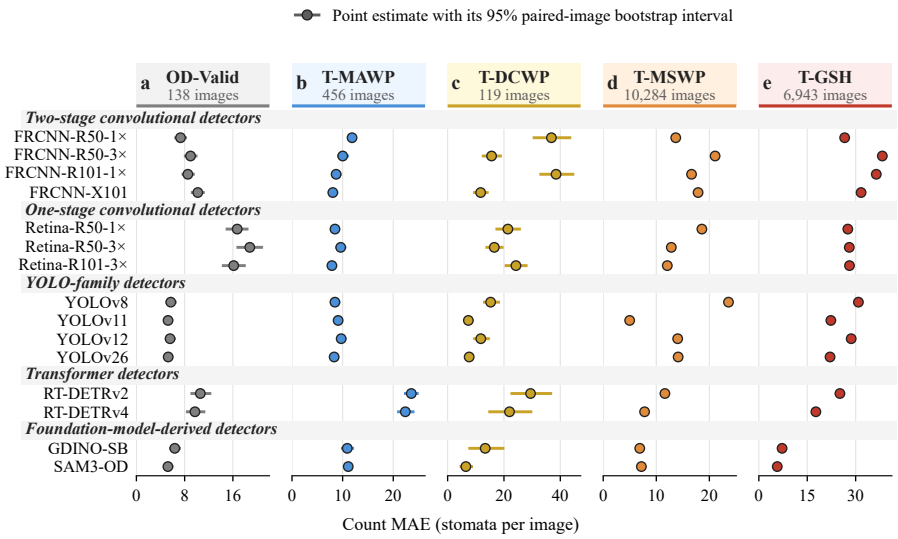


Figure S12: Count MAE (stomata per image) at $\tau = 0.25$ for every detector and set: point estimates with 95% paired-image bootstrap intervals (percentile intervals from 2,000 resamples of scene-primary images, paired across detectors). Exact values are in Table S19.

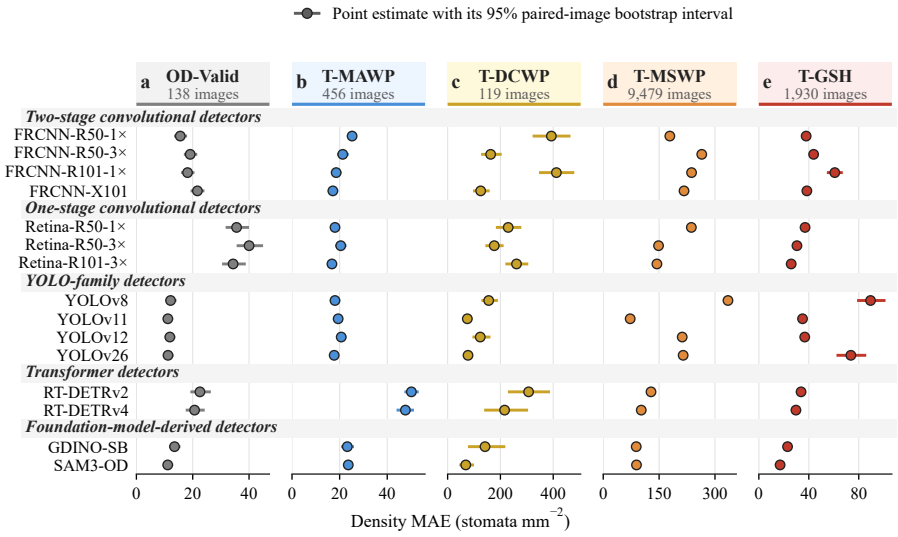


Figure S13: Density MAE (stomata mm^{-2}) at $\tau = 0.25$ for every detector and set: point estimates with 95% paired-image bootstrap intervals (percentile intervals from 2,000 resamples of scene-primary images, paired across detectors). Density uses the calibrated images (Table 1). Exact values are in Table S20.

S6 Bootstrap intervals at the fixed threshold

This section gives the 95% paired-image bootstrap intervals at $\tau = 0.25$, for the set-level results and for the species with enough images to resample. Tables S19 and S20 hold the set-level count MAE and density MAE plotted in Figures S12 and S13. Table S21 compares detection and density metrics for three example species. On the same 45 *Malus domestica* images, YOLOv12 and RT-DETRv4 have nearly the same AP (44.59 and 43.57), but at $\tau = 0.25$ their density R^2 is 0.83 and -0.67 (Table S21 and Figure S19). FRCNN-R50-3× and RT-DETRv2 are a second such pair (AP 37.87 and 38.00, density R^2 0.62 and -0.52). In both pairs the RT-DETR model over-counts all or all but one of the images, mostly with background boxes scoring below 0.50, while its counts still rise and fall with the annotated density. AP does not depend on τ, but the count does. At $\tau = 0.50$ the order reverses in both pairs, with density R^2 of 0.17 for YOLOv12 and 0.70 for RT-DETRv4, and 0.28 for FRCNN-R50-3× and 0.76 for RT-DETRv2 (Figure S47). At each detector's model-specific threshold, selected on OD-Valid (Table S8), the *Malus domestica* density R^2 is 0.79 for YOLOv12 and 0.73 for RT-DETRv4, and 0.50 for FRCNN-R50-3× and 0.62 for RT-DETRv2 (Figure S45). Close AP can therefore come with opposite density R^2 at one fixed threshold, while a threshold suited to each detector narrows the gap and removes the sign difference. On the calibrated Gibbs wheat images, by contrast, localization and density fail together, and density R^2 is below 0 for every detector at all four thresholds of Table S18. A sweep of all 999 thresholds from 0.001 to 0.999, in steps of 0.001, finds no threshold at which any detector reaches a positive wheat density R^2. The highest value is -1.13, for SAM3-OD at $\tau = 0.806$. Figures S14–S18 plot per-species count MAE, density MAE and density R^2 with intervals for every T-MAWP species, and the density metrics for the calibrated Gibbs wheat images, the L5 stress test.

Table S19: Count MAE (stomata per image) at $\tau = 0.25$ with 95% paired-image bootstrap intervals (percentile intervals from 2,000 resamples of scene-primary images, paired across detectors). Best point estimate **bold**, second-best distinct <u>underlined</u>.

Model	OD-Valid	T-MAWP	T-DCWP	T-MSWP	T-GSH
	Count MAE	Count MAE	Count MAE	Count MAE	Count MAE
Two-stage convolutional detectors					
FRCNN-R50-1×	7.32 [6.28, 8.38]	11.86 [11.04, 12.78]	36.84 [30.19, 43.89]	13.67 [13.48, 13.86]	26.60 [25.90, 27.29]
FRCNN-R50-3×	8.98 [7.87, 10.13]	10.02 [8.92, 11.24]	15.61 [12.14, 19.27]	21.08 [20.88, 21.29]	38.27 [37.31, 39.26]
FRCNN-R101-1×	8.54 [7.49, 9.68]	8.72 [7.80, 9.78]	38.50 [32.58, 44.97]	16.64 [16.48, 16.82]	36.40 [35.54, 37.26]
FRCNN-X101	10.18 [9.04, 11.34]	<u>8.06 [7.09, 9.11]</u>	11.72 [9.03, 14.56]	17.88 [17.64, 18.11]	31.68 [30.81, 32.60]
One-stage convolutional detectors					
Retina-R50-1×	16.72 [14.80, 18.59]	8.49 [7.62, 9.40]	21.39 [16.97, 26.01]	18.59 [18.32, 18.86]	27.57 [26.74, 28.41]
Retina-R50-3×	18.80 [16.59, 21.01]	9.62 [8.58, 10.73]	16.57 [13.43, 19.85]	12.83 [12.61, 13.05]	28.03 [27.17, 28.94]
Retina-R101-3×	16.14 [14.17, 18.14]	**7.88 [7.18, 8.62]**	24.25 [20.24, 28.39]	12.07 [11.84, 12.32]	28.10 [27.16, 29.03]
YOLO-family detectors					
YOLOv8	5.70 [4.84, 6.68]	8.48 [7.50, 9.52]	15.27 [12.57, 18.58]	23.61 [23.38, 23.87]	30.86 [29.77, 31.94]
YOLOv11	<u>5.25 [4.48, 6.10]</u>	9.11 [8.29, 9.97]	<u>7.36 [5.70, 9.14]</u>	**4.98 [4.89, 5.08]**	22.31 [21.61, 23.02]
YOLOv12	5.58 [4.78, 6.52]	9.71 [8.73, 10.74]	11.78 [9.10, 14.99]	14.03 [13.88, 14.18]	28.62 [27.67, 29.56]
YOLOv26	5.28 [4.46, 6.11]	8.32 [7.51, 9.24]	7.65 [6.29, 9.20]	14.12 [13.92, 14.33]	22.07 [21.26, 22.90]
Transformer detectors					
RT-DETRv2	10.60 [8.96, 12.43]	23.57 [22.13, 25.02]	29.44 [22.32, 37.11]	11.62 [11.47, 11.78]	25.13 [24.38, 25.88]
RT-DETRv4	9.71 [8.22, 11.44]	22.39 [20.76, 24.20]	21.97 [14.43, 30.06]	7.79 [7.66, 7.92]	17.67 [17.08, 18.28]
Foundation-model-derived detectors					
GDINO-SB	6.36 [5.44, 7.37]	10.91 [9.67, 12.26]	13.34 [7.34, 20.14]	<u>6.88 [6.79, 6.98]</u>	<u>7.20 [6.93, 7.48]</u>
SAM3-OD	**5.22 [4.41, 6.14]**	11.11 [10.33, 11.96]	**6.48 [4.26, 9.01]**	7.19 [7.05, 7.32]	**5.71 [5.48, 5.94]**

Table S20: Density MAE (stomata mm^{-2}) at $\tau = 0.25$ with 95% paired-image bootstrap intervals (percentile intervals from 2,000 resamples of scene-primary images, paired across detectors). Density uses the calibrated images (Table 1). Best point estimate **bold**, second-best distinct underlined.

Model	OD-Valid	T-MAWP	T-DCWP	T-MSWP	T-GSH
	Density MAE	Density MAE	Density MAE	Density MAE	Density MAE
Two-stage convolutional detectors					
FRCNN-R50-1×	15.57 [13.38, 18.01]	25.24 [23.53, 27.06]	392.94 [322.01, 465.50]	178.80 [175.77, 181.87]	37.87 [34.61, 41.35]
FRCNN-R50-3×	19.10 [16.85, 21.60]	21.31 [18.92, 23.83]	162.99 [127.02, 204.73]	264.49 [261.00, 268.01]	43.86 [41.79, 46.02]
FRCNN-R101-1×	18.16 [15.88, 20.62]	18.55 [16.56, 20.74]	412.06 [346.23, 479.87]	237.14 [233.01, 241.32]	60.82 [54.47, 67.22]
FRCNN-X101	21.66 [19.24, 24.18]	17.16 [15.08, 19.39]	125.26 [96.34, 159.19]	216.87 [213.61, 220.24]	38.45 [35.80, 41.28]
One-stage convolutional detectors					
Retina-R50-1×	35.57 [31.68, 39.98]	18.06 [16.14, 20.15]	229.46 [182.84, 279.16]	236.58 [232.86, 240.40]	37.01 [34.53, 39.69]
Retina-R50-3×	40.01 [35.60, 44.98]	20.47 [18.14, 22.87]	176.69 [142.99, 212.09]	148.48 [146.05, 150.91]	30.51 [29.13, 31.92]
Retina-R101-3×	34.34 [30.44, 38.84]	**16.77 [15.19, 18.42]**	260.94 [218.88, 305.63]	144.13 [141.52, 146.73]	25.89 [24.37, 27.56]
YOLO-family detectors					
YOLOv8	12.13 [10.33, 14.25]	18.05 [16.01, 20.21]	155.60 [128.16, 190.55]	334.88 [329.88, 340.06]	89.43 [78.67, 101.42]
YOLOv11	11.16 [9.47, 13.03]	19.38 [17.75, 21.20]	74.64 [58.15, 94.06]	**72.34 [70.45, 74.21]**	34.98 [32.33, 37.76]
YOLOv12	11.87 [10.05, 13.80]	20.67 [18.73, 22.84]	123.61 [93.99, 162.21]	212.03 [208.41, 215.87]	36.74 [33.95, 39.64]
YOLOv26	11.22 [9.59, 13.15]	17.71 [16.00, 19.73]	77.25 [62.79, 93.43]	214.57 [210.23, 218.80]	73.67 [62.15, 85.99]
Transformer detectors					
RT-DETRv2	22.56 [19.16, 26.43]	50.14 [47.20, 53.21]	306.79 [228.38, 388.11]	128.26 [126.41, 129.99]	33.80 [30.89, 36.69]
RT-DETRv4	20.66 [17.51, 24.25]	47.63 [43.91, 51.22]	215.91 [138.41, 304.85]	102.06 [100.08, 104.05]	29.70 [26.96, 33.10]
Foundation-model-derived detectors					
GDINO-SB	13.54 [11.59, 15.65]	23.22 [20.63, 26.06]	141.65 [77.18, 218.50]	88.58 [86.99, 90.08]	22.99 [21.16, 24.80]
SAM3-OD	**11.12 [9.48, 13.06]**	23.63 [22.00, 25.44]	**68.71 [44.27, 100.13]**	89.27 [87.55, 90.94]	**17.02 [15.11, 19.14]**

Table S21: Detection and density metrics for three example species at $\tau = 0.25$, one each from T-MAWP, T-MSWP and T-GSH. AP, F1, within-species density R^2 and density MAE are computed on the same images of each species. AP uses official score-ranked COCO evaluation, and F1 uses IoU 0.50. Brackets are paired image-bootstrap 95% percentile intervals (1,000 replicates for AP and 2,000 for the density metrics).

Set	Species	N	Model	AP [95% CI]	F1	Density R^2 [95% CI]	Density MAE [95% CI]
T-MAWP	*Malus domestica*	45	YOLOv12	44.59 [42.53, 46.47]	0.801	0.83 [0.70, 0.90]	10.2 [7.7, 12.6]
T-MAWP	*Malus domestica*	45	YOLOv11	45.23 [43.49, 47.11]	0.805	0.80 [0.67, 0.87]	12.4 [10.4, 14.8]
T-MAWP	*Malus domestica*	45	FRCNN-R50-3×	37.87 [35.30, 40.45]	0.759	0.62 [0.34, 0.80]	13.4 [9.4, 17.8]
T-MAWP	*Malus domestica*	45	RT-DETRv2	38.00 [35.25, 40.80]	0.719	-0.52 [-1.54, -0.01]	35.0 [29.6, 40.9]
T-MAWP	*Malus domestica*	45	RT-DETRv4	43.57 [41.61, 45.63]	0.733	-0.67 [-1.71, -0.14]	37.6 [32.6, 43.2]
T-MSWP	*Ulmus americana Planch*	185	SAM3-OD	42.24 [38.86, 45.38]	0.740	0.53 [0.34, 0.66]	51.1 [45.3, 57.5]
T-MSWP	*Ulmus americana Planch*	185	GDINO-SB	29.57 [27.31, 32.19]	0.666	-0.37 [-0.73, -0.08]	95.4 [86.4, 104.5]
T-MSWP	*Ulmus americana Planch*	185	YOLOv11	21.92 [18.80, 24.70]	0.501	-5.90 [-7.33, -4.81]	220.0 [201.0, 238.9]
T-MSWP	*Ulmus americana Planch*	185	YOLOv8	2.84 [2.17, 3.68]	0.060	-8.68 [-10.67, -7.20]	288.6 [274.6, 302.9]
T-MSWP	*Ulmus americana Planch*	185	RT-DETRv2	13.16 [10.66, 15.64]	0.452	-0.28 [-0.82, 0.12]	81.6 [71.5, 92.4]
T-GSH	*Triticum aestivum*	348	GDINO-SB	0.31 [0.25, 0.40]	0.140	-7.49 [-11.14, -5.15]	89.0 [82.7, 95.6]
T-GSH	*Triticum aestivum*	348	SAM3-OD	0.18 [0.13, 0.25]	0.095	-8.15 [-12.13, -5.64]	78.0 [69.9, 86.7]
T-GSH	*Triticum aestivum*	348	FRCNN-X101	0.00 [0.00, 0.00]	0.000	-3.54 [-4.67, -2.80]	66.9 [62.4, 71.2]
T-GSH	*Triticum aestivum*	348	Retina-R101-3×	0.00 [0.00, 0.01]	0.000	-5.03 [-7.26, -3.60]	68.6 [62.3, 74.8]
T-GSH	*Triticum aestivum*	348	YOLOv26	0.00 [0.00, 0.00]	0.000	-286.35 [-459.66, -176.79]	324.0 [270.7, 378.0]
T-GSH	*Triticum aestivum*	348	RT-DETRv2	0.00 [0.00, 0.00]	0.000	-16.08 [-24.67, -10.88]	101.8 [90.2, 113.8]

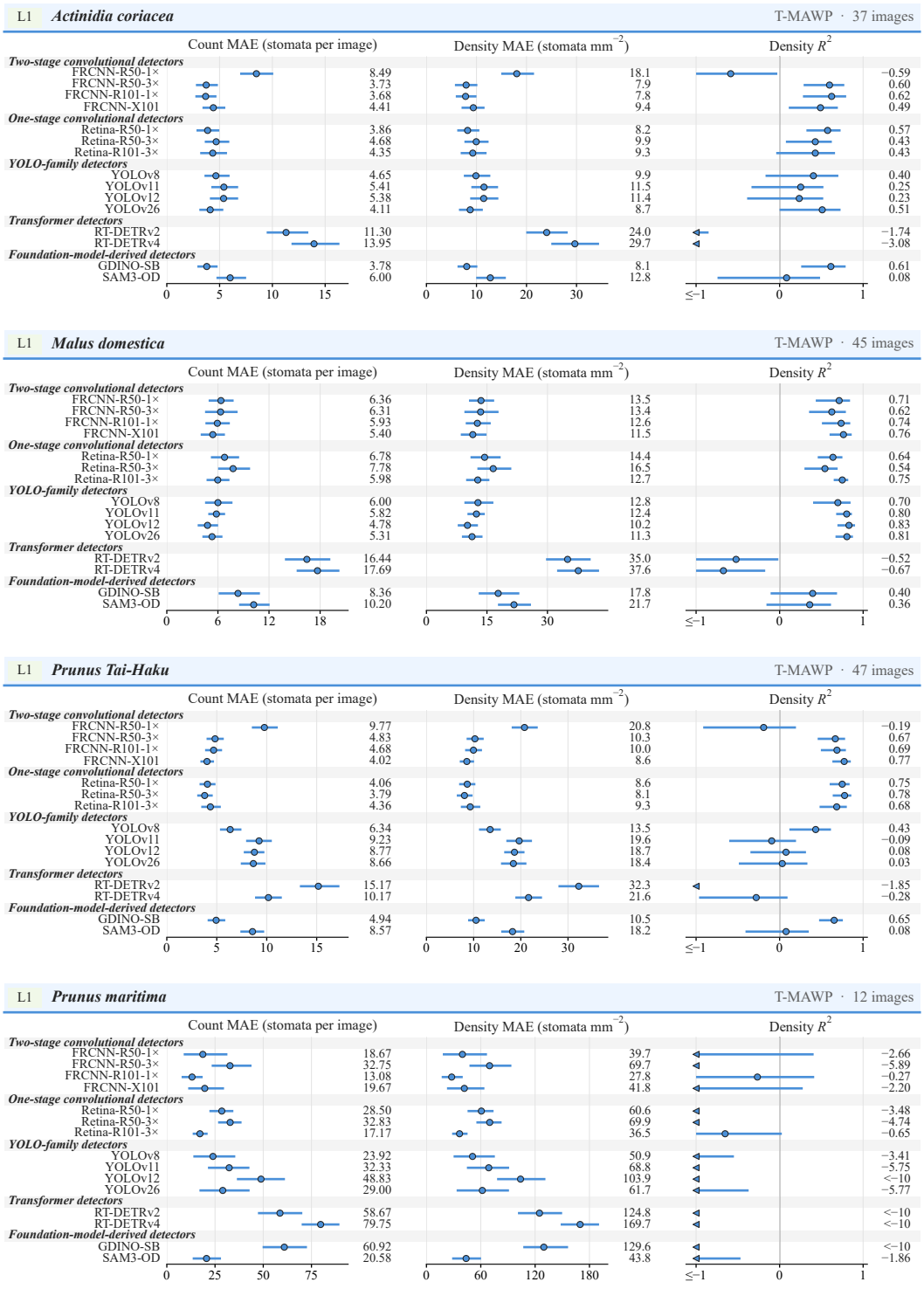


Figure S14: Per-species count MAE, density MAE and density R^2 at $\tau = 0.25$ for every detector: point estimates, printed at the right, with 95% paired-image bootstrap intervals (2,000 resamples of each species' scene-primary images). Every detector gets the same mark, and each panel has its own scale. Density R^2 below -1 is drawn at -1 with a left-pointing marker, and values below -10 print as < -10.

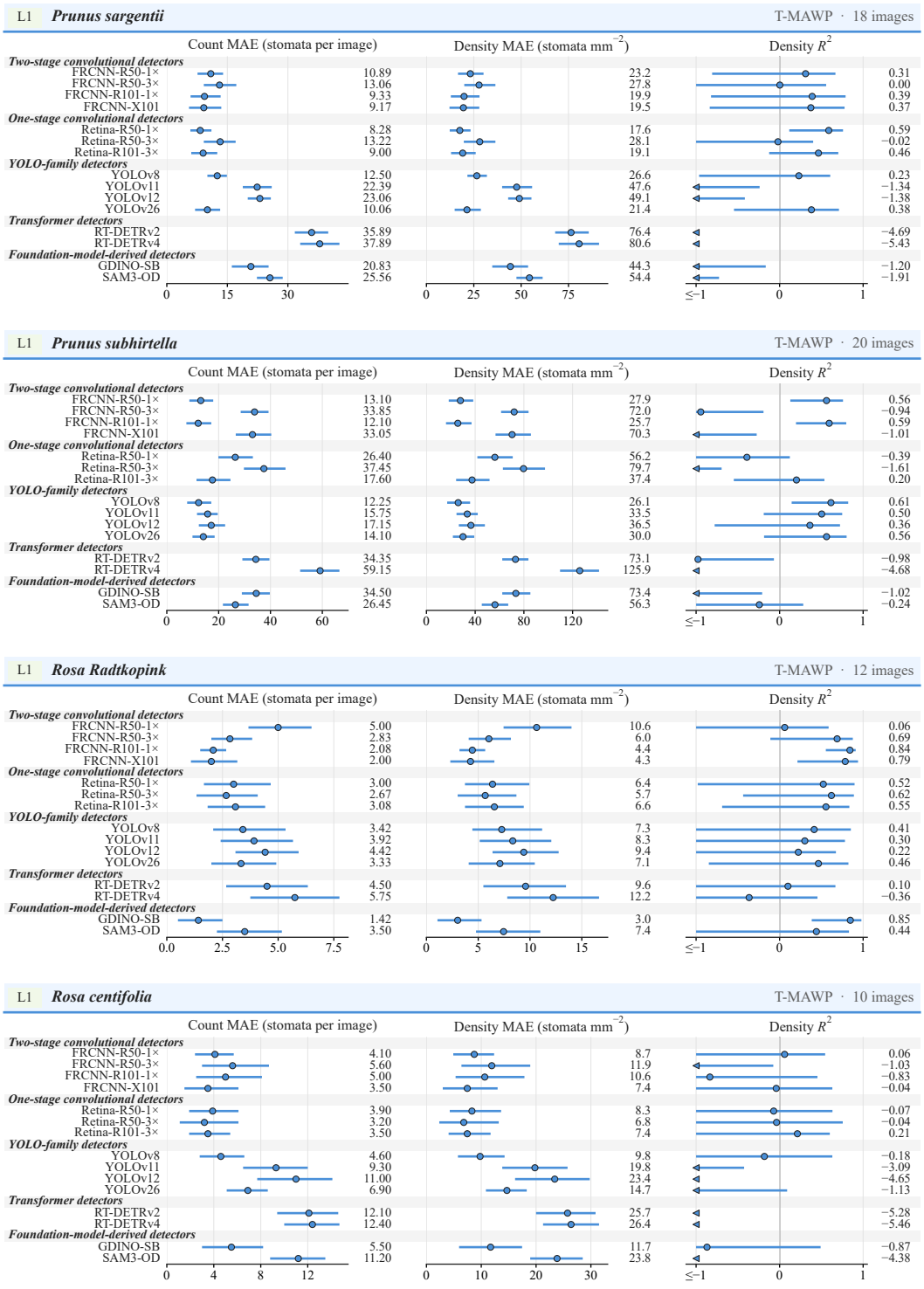


Figure S15: As Figure S14, for the next 4 T-MAWP species.

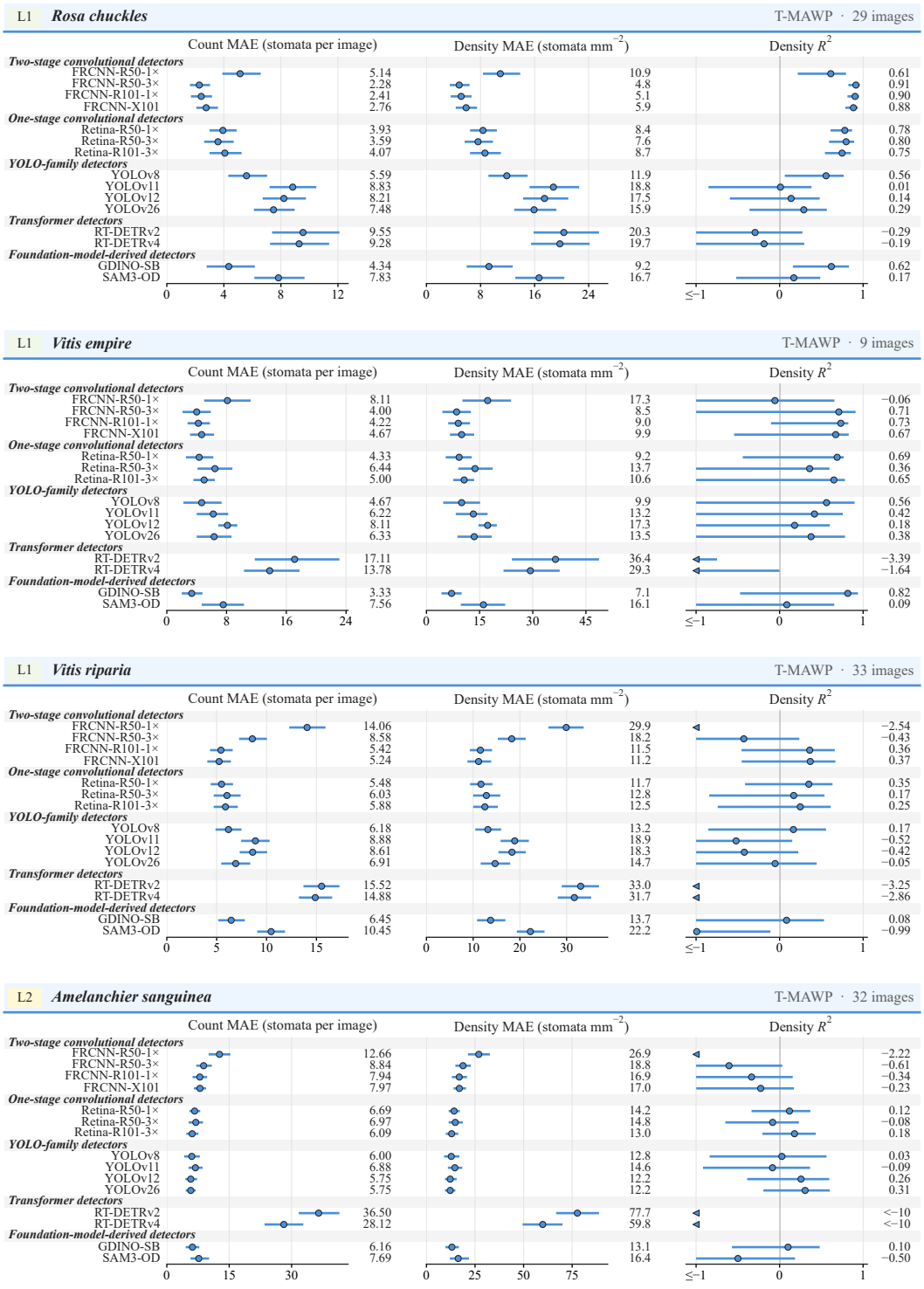


Figure S16: As Figure S14, for the next 4 T-MAWP species.

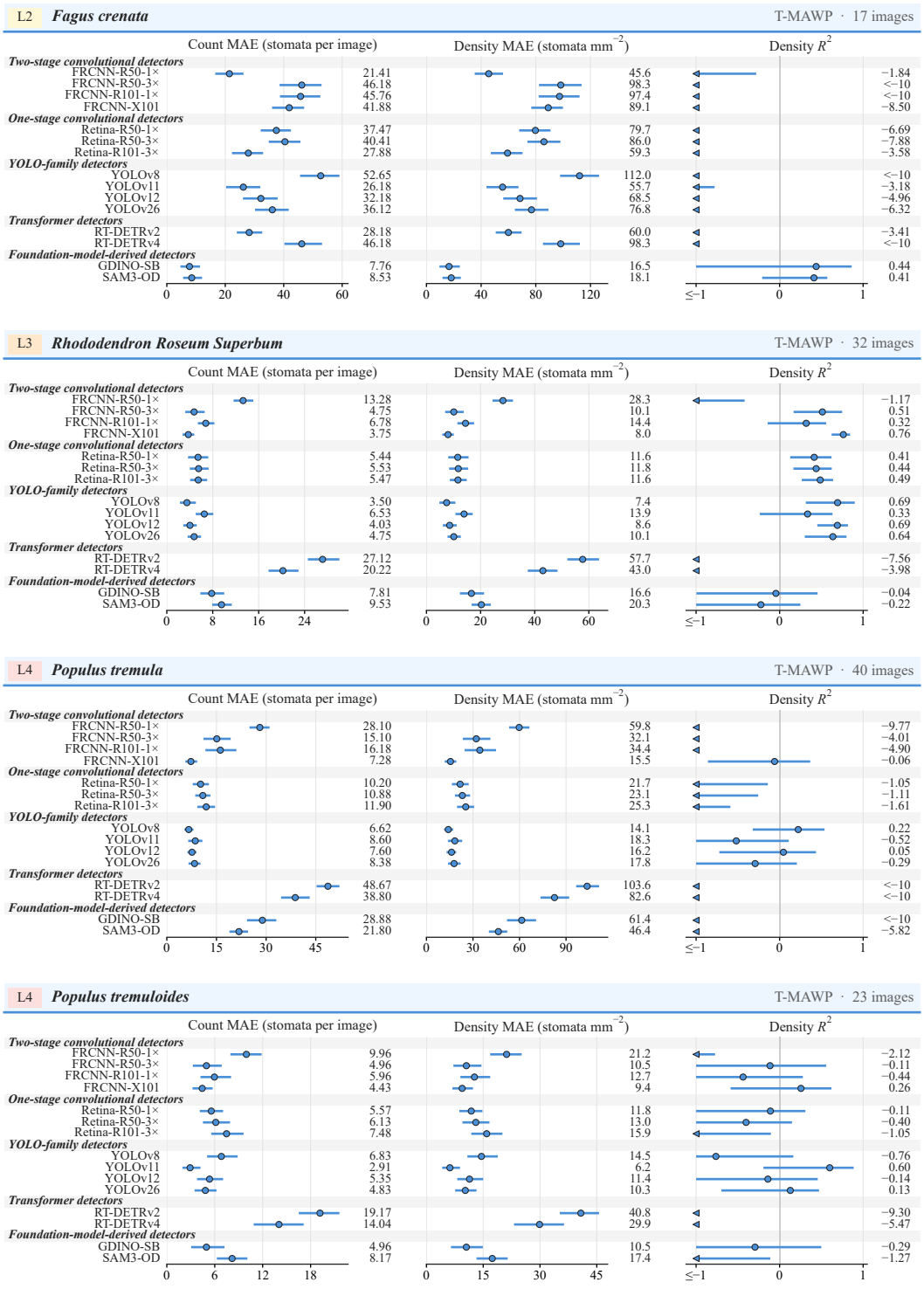


Figure S17: As Figure S14, for the next 4 T-MAWP species.

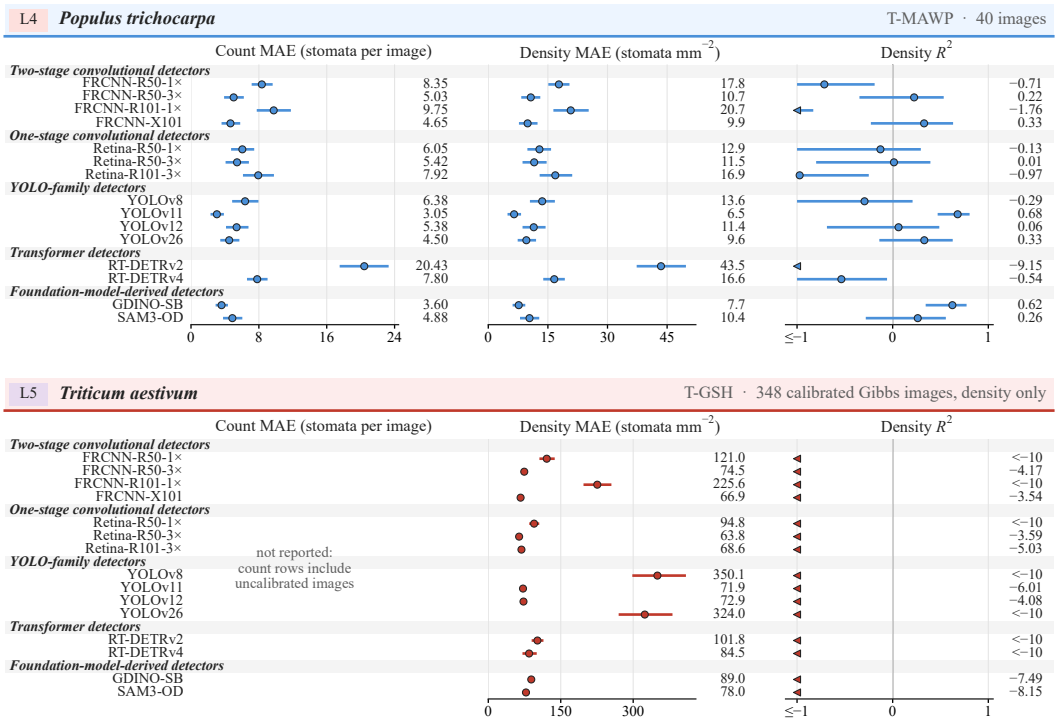


Figure S18: As Figure S14, for the last T-MAWP species and for the 348 calibrated Gibbs images of *Triticum aestivum*, the L5 stress test. Each set keeps its table colour: T-MAWP blue, T-GSH red. Wheat has density only, because its count rows also include the uncalibrated StomataTracker images.

S7 Per-species density R^2 at the fixed threshold

Figure 3 in Section 6.2 counts how often a detector's within-species density R^2 falls below zero. Figures S19 and S20 give every value, one cell per detector and calibrated species, in the colours of Figure 3. Most T-DCWP species have very few calibrated images, so their values are descriptive. Density MAE and all other per-species metrics are in Section S8.

Confidence: $\tau = 0.25$

		N	$\bar{\rho}_{GT}$	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
OD-Valid · 5 calibrated species																		
L0	*Actinidia arguta*	37	82.5	0.87	0.78	0.72	0.38	0.39	0.19	0.47	0.85	0.84	0.87	0.85	0.44	0.47	0.74	0.84
L0	*Malus hupehensis*	41	170.0	0.32	0.05	0.19	−0.06	−1.83	−2.33	−1.66	0.59	0.64	0.61	0.67	0.79	0.84	0.69	0.72
L0	*Malus spontanea*	40	163.3	0.55	0.20	0.31	0.04	−2.05	−3.34	−2.21	0.54	0.72	0.68	0.70	−0.43	−0.65	0.54	0.64
L0	*Malus toringoides*	9	158.6	0.51	0.42	0.23	0.32	−1.18	−1.67	−1.08	0.77	0.72	0.65	0.68	−3.20	−2.05	0.15	0.69
L0	*Vitis vinifera*	11	66.2	−3.20	−0.36	0.53	0.20	0.25	0.18	0.40	−0.12	−0.23	−0.56	−0.26	−4.34	−2.82	0.41	−0.81
T-MAWP · 17 calibrated species																		
L1	*Actinidia coriacea*	37	59.5	−0.59	0.60	0.62	0.49	0.57	0.43	0.43	0.40	0.25	0.23	0.51	−1.74	−3.08	0.61	0.08
L1	*Malus domestica*	45	72.4	0.71	0.62	0.74	0.76	0.64	0.54	0.75	0.70	0.80	0.83	0.81	−0.52	−0.67	0.40	0.36
L1	*Prunus Tai-Haku*	47	75.8	−0.19	0.67	0.69	0.77	0.75	0.78	0.68	0.43	−0.09	0.08	0.03	−1.85	−0.28	0.65	0.08
L1	*Prunus maritima*	12	208.7	−2.66	−5.89	−0.27	−2.20	−3.48	−4.74	−0.65	−3.41	−5.75	<−10	−5.77	<−10	<−10	<−10	−1.86
L1	*Prunus sargentii*	18	107.2	0.31	0.00	0.39	0.37	0.59	−0.02	0.46	0.23	−1.34	−1.38	0.38	−4.69	−5.43	−1.20	−1.91
L1	*Prunus subhirtella*	20	186.1	0.56	−0.94	0.59	−1.01	−0.39	−1.61	0.20	0.61	0.50	0.36	0.56	−0.98	−4.68	−1.02	−0.24
L1	*Rosa Radtkopink*	12	39.4	0.06	0.69	0.84	0.79	0.52	0.62	0.55	0.41	0.30	0.22	0.46	0.10	−0.36	0.85	0.44
L1	*Rosa centifolia*	10	60.6	0.06	−1.03	−0.83	−0.04	−0.07	−0.04	0.21	−0.18	−3.09	−4.65	−1.13	−5.28	−5.46	−0.87	−4.38
L1	*Rosa chuckles*	29	67.0	0.61	0.91	0.90	0.88	0.78	0.80	0.75	0.56	0.01	0.14	0.29	−0.29	−0.19	0.62	0.17
L1	*Vitis empire*	9	54.6	−0.06	0.71	0.73	0.67	0.69	0.36	0.65	0.56	0.42	0.18	0.38	−3.39	−1.64	0.82	0.09
L1	*Vitis riparia*	33	80.1	−2.54	−0.43	0.36	0.37	0.35	0.17	0.25	0.17	−0.52	−0.42	−0.05	−3.25	−2.86	0.08	−0.99
L2	*Fagus crenata*	17	166.5	−1.84	<−10	<−10	−8.50	−6.69	−7.88	−3.58	<−10	−3.18	−4.96	−6.32	−3.41	<−10	0.44	0.41
L2	*Amelanchier sanguinea*	32	86.9	−2.22	−0.61	−0.34	−0.23	0.12	−0.08	0.18	0.03	−0.09	0.26	0.31	<−10	<−10	0.10	−0.50
L3	*Rhododendron Roseum Superbum*	32	82.3	−1.17	0.51	0.32	0.76	0.41	0.44	0.49	0.69	0.33	0.69	0.64	−7.56	−3.98	−0.04	−0.22
L4	*Populus tremula*	40	37.6	−9.77	−4.01	−4.90	−0.06	−1.05	−1.11	−1.61	0.22	−0.52	0.05	−0.29	<−10	<−10	<−10	−5.82
L4	*Populus tremuloides*	23	52.6	−2.12	−0.11	−0.44	0.26	−0.11	−0.40	−1.05	−0.76	0.60	−0.14	0.13	−9.30	−5.47	−0.29	−1.27
L4	*Populus trichocarpa*	40	55.7	−0.71	0.22	−1.76	0.33	−0.13	0.01	−0.97	−0.29	0.68	0.06	0.33	−9.15	−0.54	0.62	0.26
T-MSWP · 18 calibrated species																		
L1	*Quercus michauxii Nutt.*	129	827.6	0.09	−0.48	−0.22	−0.02	−0.14	0.43	0.33	−2.49	0.78	0.20	0.20	0.32	0.60	0.71	0.68
L1	*Quercus nigra L.*	500	639.5	−1.22	−3.09	−3.56	−2.07	−2.56	−0.17	−0.22	−5.27	0.46	−1.77	−1.40	0.19	0.38	0.50	0.46
L1	*Quercus pagoda Raf.*	286	669.9	−2.56	−4.02	−2.91	−2.36	−3.06	−0.79	−1.02	−5.44	−0.38	−3.27	−3.62	0.49	−0.17	0.53	0.64
L1	*Quercus phellos L.*	568	710.6	−7.14	<−10	<−10	−8.29	<−10	−3.16	−3.22	<−10	0.41	<−10	<−10	−2.48	−1.03	−0.72	−0.97
L1	*Quercus shumardii Buckley*	1,348	786.8	0.08	−0.60	−0.83	−0.57	−0.96	0.06	0.04	−2.67	0.84	−0.08	−1.01	0.59	0.60	0.65	0.62
L1	*Quercus stellata Wangenh.*	96	1104.6	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	0.22	<−10	<−10	<−10	−9.36	−2.79	−2.54
L1	*Quercus texana Buckley*	2,434	724.3	0.21	−0.27	−0.54	−0.20	−0.62	0.39	0.37	−2.54	0.85	−0.33	−0.48	0.62	0.63	0.73	0.65
L3	*Acer rubrum L.*	100	483.6	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	−8.87	<−10	<−10	−7.33	−0.40	−1.99	−1.12
L3	*Carya ovata (Mill.) K. Koch*	91	407.5	−7.25	−9.61	−6.96	−4.72	−7.12	−4.27	−4.14	−9.48	−2.06	−7.51	−6.74	−1.77	−5.71	−0.38	0.15
L3	*Nyssa sylvatica Marshall*	50	274.9	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	−8.63	−3.68	−1.75	0.15
L3	*Ulmus alata Michx.*	205	594.5	−5.13	−7.77	−4.77	−2.87	−5.95	−2.88	−2.32	<−10	−3.16	−6.42	−5.81	−0.11	−1.26	−0.49	−0.18
L3	*Ulmus americana Planch*	185	309.1	−3.86	−5.58	−4.56	−2.48	−4.84	−2.98	−1.37	−8.68	−5.90	−6.75	−7.02	−0.28	−3.16	−0.37	0.53
L3	*Vaccinium stamineum L.*	75	346.4	<−10	<−10	<−10	<−10	<−10	−5.81	−7.39	<−10	−1.85	<−10	<−10	−5.95	−0.48	−0.69	−0.06
L4	*Dirca palustris L.*	65	590.5	<−10	<−10	<−10	−9.21	<−10	−7.89	−4.65	<−10	−0.08	<−10	<−10	−7.31	0.01	−0.40	0.24
L4	*Fraxinus caroliniana Miller*	100	451.2	−4.70	<−10	<−10	−9.15	<−10	−3.96	−1.32	<−10	0.23	−1.41	−8.51	−5.91	−0.39	−1.29	0.22
L4	*Ilex opaca Aiton*	100	270.2	<−10	<−10	<−10	<−10	<−10	−7.05	−2.52	<−10	0.20	<−10	<−10	<−10	−6.76	−5.50	−8.48
L4	*Populus L.*	3,065	222.3	−0.27	−6.77	−1.72	−2.24	−0.50	−0.04	0.30	−6.04	0.54	−3.27	−0.77	−2.21	0.53	0.64	0.81
L4	*Salix L.*	82	404.4	−5.11	<−10	−9.52	−4.42	−4.19	−1.18	−0.41	<−10	0.38	<−10	−2.26	−4.07	−0.17	−0.58	0.42
T-GSH · 4 calibrated species																		
L4	*Populus balsamifera*	113	269.7	0.37	−7.03	−2.12	<−10	−1.15	−0.36	0.19	<−10	−8.99	<−10	−7.19	−2.32	−0.90	0.67	0.94
L5	*Triticum aestivum*	348	78.0	<−10	−4.17	<−10	−3.54	<−10	−3.59	−5.03	<−10	−6.01	−4.08	<−10	<−10	<−10	−7.49	−8.15
L5	*Hordeum vulgare*	542	17.0	−5.69	−6.68	−5.84	−4.81	−2.73	−4.39	−2.16	−6.67	−3.06	−6.04	−1.14	−8.48	−2.22	−1.28	0.50
L5	*Zea mays*	927	30.8	<−10	<−10	<−10	<−10	<−10	<−10	<−10	−9.96	−6.77	−3.52	0.10	−5.59	−7.42	−1.03	0.49

Within-species density R^2: brown below the species-mean reference (0), teal above

≤−1 0 1

Figure S19: Within-species density R^2 at $\tau = 0.25$ for every detector and every calibrated species of OD-Valid, T-MAWP, T-MSWP and T-GSH. Values are printed, colour is clipped at -1, and values below -10 print as < -10. Teal: better than the species' own mean density (reference 0). Brown: worse. N: calibrated images. $\bar{\rho}_{GT}$: mean annotated density (stomata mm^{-2}). The 5 T-GSH species without calibrated images are left out. Density MAE for the same species is in Section S8.

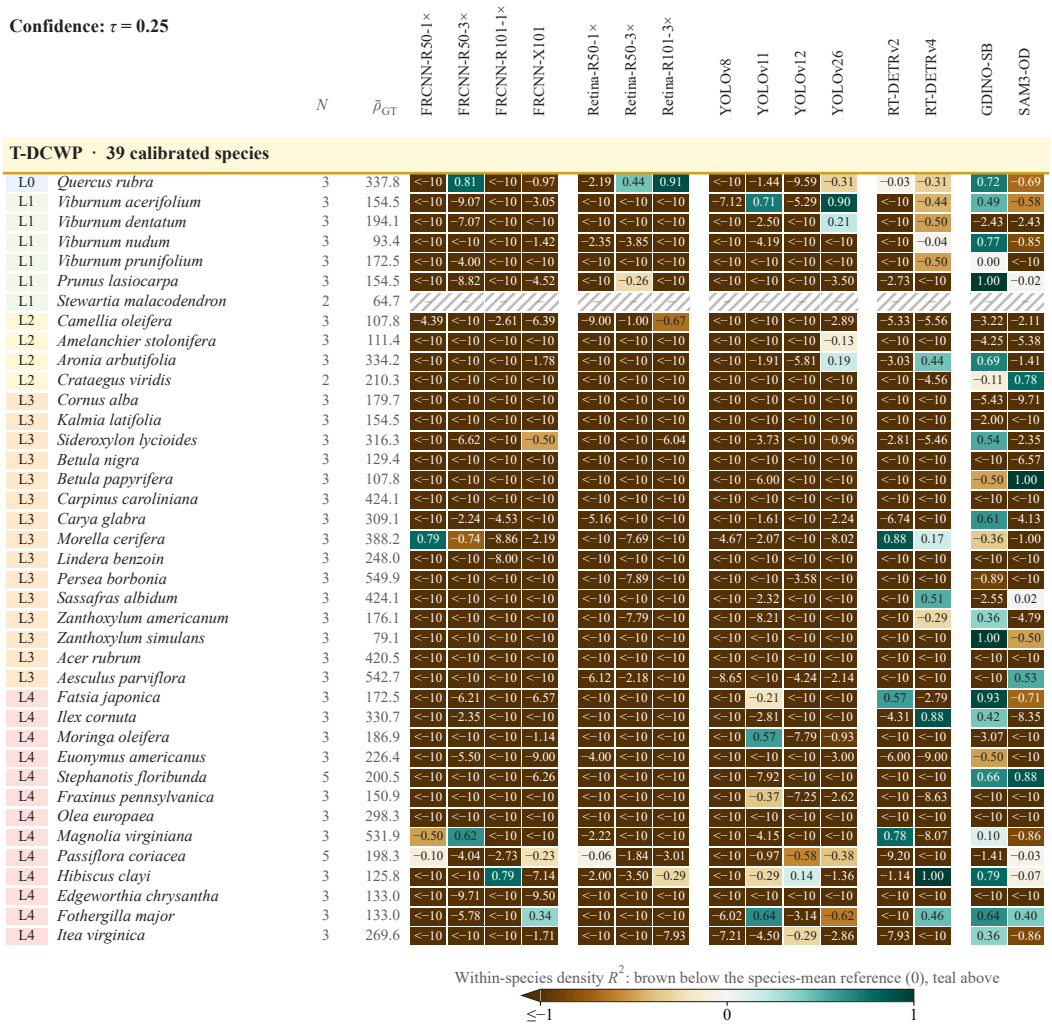


Figure S20: As Figure S19, for all 39 T-DCWP species. With at most 5 calibrated images per species, these values are descriptive. *Crataegus viridis* and *Stewartia malacodendron* have 2 images each, below the three-image rule of Section 5, so they are not in Figure 3. The two *Stewartia malacodendron* images have the same annotated density, so its R^2 is undefined (–).

S8 Per-species results at the fixed threshold

Set-level numbers average over species that a detector may handle very differently. This section reports every metric per species at $\tau = 0.25$, except COCO AP, which scores all released predictions, without an additional confidence cutoff. Each metric has two figures, one for OD-Valid, T-MAWP and T-DCWP and, a page or two later, one for T-MSWP and T-GSH. Figures S21–S42 give, in this order, COCO $AP_{50:95}$, precision, recall, F1, true positives, false positives, false negatives, count MAE, count R^2, density MAE and signed count bias. For the other metrics, matching uses IoU 0.50, and the density metrics use the calibrated images only. OD-Valid and T-MAWP images all cover the same leaf area, 0.47 mm^2, so on these two sets density is count divided by that area. There, density R^2 equals count R^2 and density MAE is 2.13 times count MAE, the same result in two units. Each row gives the number of images per species (N). T-DCWP has two to five images per species, most with three, so its per-species values are descriptive only.

Within-species density R^2 is in Figures S19 and S20 (Section S7), next to each species' mean density. These are the values counted in Figure 3.

COCO $AP_{50:95}$ (%), all released scores

lighter is higher

0 50 100

		N	$\bar{c}_{GT}$	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
OD-Valid · 5 species																		
L0	*Actinidia arguta*	37	38.8	40.7	35.7	30.7	32.6	37.8	37.1	33.5	44.1	41.5	43.4	39.5	43.9	45.7	40.9	42.4
L0	*Malus hupehensis*	41	79.9	45.2	46.1	45.9	43.8	45.0	44.7	45.2	49.7	50.6	50.6	50.3	51.1	52.7	47.3	46.7
L0	*Malus spontanea*	40	76.7	41.5	43.1	42.3	41.1	41.7	41.6	42.0	47.2	48.6	49.2	48.1	46.9	48.0	43.2	42.6
L0	*Malus toringoides*	9	74.6	34.3	34.4	35.3	33.2	33.5	32.3	33.1	37.3	39.8	41.3	39.8	37.0	39.9	35.6	32.6
L0	*Vitis vinifera*	11	31.1	45.1	44.2	45.1	46.0	48.7	47.2	47.4	45.5	47.2	46.4	46.9	41.7	47.9	41.8	43.9
T-MAWP · 17 species																		
L1	*Actinidia coriacea*	37	27.9	38.7	35.6	33.4	36.3	39.0	36.9	33.1	44.1	46.0	45.9	44.0	45.3	45.3	39.9	40.1
L1	*Malus domestica*	45	34.0	38.4	37.9	38.1	36.3	38.3	37.0	38.4	42.3	45.2	44.6	43.9	38.0	43.6	37.5	38.1
L1	*Prunus Tai-Haku*	47	35.6	47.0	46.4	46.9	43.4	51.2	50.6	41.4	52.2	52.2	50.4	49.1	50.8	52.2	51.1	47.5
L1	*Prunus maritima*	12	98.1	7.2	7.8	13.4	11.0	9.9	7.3	12.2	14.1	12.2	12.0	8.6	10.6	9.8	6.9	6.5
L1	*Prunus sargentii*	18	50.4	19.2	14.6	18.7	17.9	19.1	15.1	19.0	17.2	12.9	13.4	14.0	19.3	20.6	13.4	12.1
L1	*Prunus subhirtella*	20	87.5	32.1	29.6	36.9	27.5	29.7	22.9	34.8	36.7	37.8	40.2	38.6	35.6	37.9	33.1	35.3
L1	*Rosa Radtkopink*	12	18.5	54.0	48.5	49.6	52.4	55.4	54.6	52.6	51.0	53.6	50.6	56.5	48.6	56.2	51.9	51.8
L1	*Rosa centifolia*	10	28.5	35.4	36.9	35.9	34.9	37.2	36.9	38.3	32.4	33.6	36.9	36.6	33.9	41.4	36.2	40.0
L1	*Rosa chuckles*	29	31.5	48.7	46.2	46.8	47.6	48.4	48.4	50.6	38.6	37.9	40.1	43.5	43.6	48.4	47.7	52.6
L1	*Vitis empire*	9	25.7	35.2	36.0	39.7	38.3	40.4	40.1	42.4	38.7	42.1	42.4	42.2	35.8	43.1	37.2	39.1
L1	*Vitis riparia*	33	37.6	43.9	45.7	47.2	46.6	48.2	47.7	47.6	47.3	49.7	51.0	50.4	45.2	51.3	45.6	48.8
L2	*Fagus crenata*	17	78.2	18.8	14.6	14.2	15.9	15.3	12.9	15.6	22.1	27.1	29.7	26.1	23.2	25.0	12.3	23.6
L2	*Amelanchier sanguinea*	32	40.8	37.0	37.4	39.3	35.3	35.1	35.4	39.3	37.9	42.7	41.5	42.3	39.6	38.7	34.4	36.6
L3	*Rhododendron Roseum Superbum*	32	38.7	38.6	35.7	36.1	36.7	40.7	36.9	39.3	40.8	44.8	46.0	44.6	42.0	45.8	42.2	42.5
L4	*Populus tremula*	40	17.6	13.8	10.4	7.8	14.6	14.9	13.8	15.2	17.2	20.6	20.0	19.7	16.7	19.4	15.8	18.4
L4	*Populus tremuloides*	23	24.7	32.9	24.3	25.1	31.8	31.6	35.3	31.6	41.7	44.9	43.1	42.3	42.4	41.8	34.7	37.4
L4	*Populus trichocarpa*	40	26.2	37.8	39.4	30.1	34.7	36.1	37.1	31.8	42.3	45.4	43.9	44.5	44.5	46.8	34.8	40.6
T-DCWP · 39 species																		
L0	*Quercus rubra*	3	31.3	1.7	1.5	0.5	0.7	3.4	0.6	1.0	3.2	7.0	0.2	2.9	1.0	2.8	3.8	1.0
L1	*Viburnum acerifolium*	3	14.3	5.4	1.2	0.1	0.4	2.7	1.6	0.4	1.9	10.3	5.8	13.1	10.5	9.6	56.1	57.8
L1	*Viburnum dentatum*	3	18.0	10.6	5.7	0.1	1.2	8.5	5.9	1.4	1.5	16.8	30.6	33.5	30.0	34.4	52.4	57.7
L1	*Viburnum nudum*	3	8.7	0.4	0.1	0.0	1.0	0.4	0.1	0.1	0.1	28.2	0.2	3.0	9.5	18.2	74.2	73.4
L1	*Viburnum prunifolium*	3	16.0	35.5	14.6	4.1	25.3	8.7	8.1	8.0	3.3	53.0	20.2	40.9	51.5	37.3	29.6	41.4
L1	*Prunus lasiocarpa*	3	14.3	24.1	4.0	0.0	1.0	2.3	5.4	0.9	0.7	21.7	0.2	51.2	21.5	35.4	14.3	11.4
L1	*Stewartia malacodendron*	2	6.0	16.7	12.1	2.4	0.2	15.7	14.7	7.9	1.1	20.3	7.8	15.3	40.9	29.2	39.4	66.4
L2	*Camellia oleifera*	3	10.0	8.1	0.2	0.0	0.1	0.3	0.0	0.0	0.0	28.5	6.2	36.5	11.5	34.8	53.9	40.1
L2	*Amelanchier stolonifera*	3	10.3	0.4	3.2	0.0	0.0	3.9	2.9	0.2	0.4	2.8	0.3	32.6	19.3	21.7	33.9	48.9
L2	*Aronia arbutifolia*	3	31.0	20.8	12.6	2.4	4.1	15.2	14.7	5.3	9.3	29.5	38.8	49.6	43.6	33.3	64.0	66.3
L2	*Crataegus viridis*	2	19.5	5.6	0.5	0.0	0.0	0.2	2.5	0.7	0.2	6.2	7.0	41.3	18.7	37.2	70.8	73.5
L3	*Cornus alba*	3	16.7	1.5	3.1	0.0	0.7	3.2	7.0	1.2	0.0	9.1	31.4	26.9	46.3	25.6	58.1	70.0
L3	*Kalmia latifolia*	3	14.3	52.0	7.5	0.5	9.9	9.3	24.0	8.0	0.8	29.8	1.5	27.2	20.9	50.6	58.0	66.6
L3	*Sideroxylon lycioides*	3	29.3	59.8	51.8	37.6	43.3	50.1	51.6	34.4	50.4	62.4	44.7	61.5	62.2	62.4	64.5	69.2
L3	*Betula nigra*	3	12.0	1.0	2.0	5.0	2.3	1.1	3.1	1.3	0.4	1.0	1.3	1.0	3.2	0.3	19.3	39.7
L3	*Betula papyrifera*	3	10.0	0.7	0.1	0.0	0.1	0.6	0.3	0.4	0.0	6.9	1.1	22.2	16.6	18.2	59.9	65.8
L3	*Carpinus caroliniana*	3	39.3	11.8	15.6	8.8	16.9	19.8	20.7	12.8	23.7	37.0	20.8	21.2	30.4	13.6	68.3	22.2
L3	*Carya glabra*	3	28.7	46.7	50.5	9.6	33.9	23.7	28.1	5.9	23.7	57.2	44.3	52.0	50.4	52.1	63.6	68.3
L3	*Morella cerifera*	3	36.0	31.8	27.1	10.5	19.6	22.4	26.9	16.1	20.0	34.6	42.3	37.6	37.2	26.0	35.5	39.3
L3	*Lindera benzoin*	3	23.0	8.7	9.7	2.6	5.7	4.2	4.7	3.2	5.0	13.0	3.6	6.2	4.3	6.1	51.7	49.2
L3	*Persea borbonia*	3	51.0	12.4	23.4	4.0	5.2	17.3	23.4	6.6	11.3	34.6	35.2	38.9	14.5	38.4	50.0	54.8
L3	*Sassafras albidum*	3	39.3	2.2	10.7	0.1	5.7	6.7	8.1	1.7	1.0	14.3	7.9	13.9	9.7	25.1	52.2	47.3
L3	*Zanthoxylum americanum*	3	16.3	27.6	3.1	0.1	7.3	1.2	0.9	1.4	2.5	41.8	4.4	38.1	28.0	43.7	63.1	69.6
L3	*Zanthoxylum simulans*	3	7.3	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.5	0.0	0.1	4.0	12.9	57.8	62.9
L3	*Acer rubrum*	3	39.0	5.5	15.0	6.7	4.7	7.5	8.4	4.0	4.6	17.3	5.5	16.6	3.7	4.9	3.3	37.7
L3	*Aesculus parviflora*	3	50.3	9.1	12.2	3.2	9.1	12.7	15.9	6.1	11.6	25.8	14.3	31.7	20.6	20.9	34.4	47.8
L4	*Fatsia japonica*	3	16.0	56.3	17.5	0.1	8.2	34.7	30.5	17.6	0.6	58.8	30.6	58.7	53.6	67.5	63.3	69.3
L4	*Ilex cornuta*	3	30.7	55.2	28.9	2.0	43.5	46.6	46.7	34.7	4.3	35.8	44.8	47.7	56.6	62.7	25.1	22.3
L4	*Moringa oleifera*	3	17.3	39.1	49.1	38.8	43.9	45.7	40.8	29.0	18.1	26.5	43.4	26.3	48.8	55.7	49.4	60.1
L4	*Euonymus americanus*	3	21.0	37.5	22.1	16.0	28.4	13.1	9.4	9.2	23.2	53.4	23.2	48.3	41.5	30.3	63.5	42.6
L4	*Stephanotis floribunda*	5	39.0	33.3	21.8	19.6	9.6	27.5	24.5	16.6	6.5	32.7	20.2	30.0	39.8	23.2	44.5	42.8
L4	*Fraxinus pennsylvanica*	3	14.0	17.1	12.0	0.2	9.0	18.3	15.9	12.4	3.0	19.3	30.2	49.6	55.3	46.9	58.0	64.0
L4	*Olea europaea*	3	27.7	32.5	38.0	11.1	34.3	38.3	44.3	19.7	17.4	40.5	48.6	47.2	43.1	47.0	56.0	65.6
L4	*Magnolia virginiana*	3	49.3	64.2	58.0	24.5	55.4	31.6	48.2	18.9	39.1	39.3	55.0	44.1	54.6	54.3	66.5	73.3
L4	*Passiflora coriacea*	5	40.0	24.1	16.5	17.5	13.6	13.1	12.1	11.3	7.9	22.7	17.8	14.1	23.9	15.0	26.7	38.7
L4	*Hibiscus clayi*	3	11.7	67.6	57.6	61.1	58.2	44.2	55.7	42.3	24.4	69.7	66.9	65.7	66.2	72.4	52.9	67.4
L4	*Edgeworthia chrysantha*	3	12.3	0.5	0.5	0.1	0.8	1.2	1.6	1.6	0.8	18.2	0.2	6.9	2.8	14.6	25.1	20.1
L4	*Fothergilla major*	3	12.3	39.0	32.2	7.1	40.1	24.5	18.6	13.2	34.5	46.1	44.0	43.4	47.9	59.8	53.1	64.8
L4	*Itea virginica*	3	25.0	50.4	51.2	34.0	52.8	43.0	50.6	27.1	57.4	61.2	29.6	60.2	58.8	55.8	64.4	69.4

Figure S21: COCO $AP_{50:95}$ (%), computed from all released predictions, without an additional confidence cutoff, for every detector and every species of OD-Valid, T-MAWP and T-DCWP. Each cell prints the value. N: images. $\bar{c}_{GT}$: mean annotated stomata per image.

Precision at $\tau = 0.25$

lighter is higher

0 — 0.5 — 1

		N	$\bar{c}_{GT}$	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
OD-Valid · 5 species																		
L0	*Actinidia arguta*	37	38.8	0.90	0.91	0.90	0.92	0.94	0.95	0.89	0.89	0.86	0.88	0.87	0.84	0.86	0.96	0.91
L0	*Malus hupehensis*	41	79.9	0.97	0.98	0.98	0.98	0.99	0.98	0.99	0.90	0.91	0.91	0.90	0.91	0.92	0.96	0.94
L0	*Malus spontanea*	40	76.7	0.92	0.94	0.94	0.95	0.96	0.97	0.97	0.90	0.89	0.90	0.89	0.81	0.80	0.89	0.90
L0	*Malus toringoides*	9	74.6	0.82	0.84	0.81	0.85	0.91	0.90	0.90	0.80	0.83	0.84	0.82	0.60	0.66	0.74	0.80
L0	*Vitis vinifera*	11	31.1	0.72	0.79	0.86	0.82	0.83	0.82	0.84	0.76	0.79	0.76	0.78	0.69	0.72	0.82	0.77
T-MAWP · 17 species																		
L1	*Actinidia coriacea*	37	27.9	0.70	0.79	0.81	0.85	0.81	0.81	0.77	0.81	0.78	0.78	0.81	0.68	0.64	0.84	0.77
L1	*Malus domestica*	45	34.0	0.77	0.78	0.79	0.82	0.84	0.84	0.80	0.80	0.79	0.82	0.79	0.60	0.61	0.71	0.69
L1	*Prunus Tai-Haku*	47	35.6	0.77	0.86	0.86	0.87	0.87	0.87	0.86	0.83	0.79	0.79	0.79	0.69	0.77	0.85	0.79
L1	*Prunus maritima*	12	98.1	0.52	0.57	0.62	0.60	0.66	0.61	0.65	0.56	0.52	0.47	0.53	0.45	0.39	0.39	0.44
L1	*Prunus sargentii*	18	50.4	0.63	0.68	0.72	0.73	0.73	0.74	0.73	0.61	0.51	0.51	0.58	0.51	0.51	0.53	0.49
L1	*Prunus subhirtella*	20	87.5	0.83	0.91	0.87	0.89	0.90	0.90	0.87	0.85	0.76	0.77	0.78	0.67	0.58	0.68	0.70
L1	*Rosa Radtkopink*	12	18.5	0.77	0.86	0.92	0.91	0.85	0.85	0.83	0.82	0.81	0.79	0.82	0.79	0.76	0.88	0.84
L1	*Rosa centifolia*	10	28.5	0.77	0.83	0.85	0.84	0.80	0.85	0.81	0.76	0.70	0.67	0.75	0.65	0.67	0.75	0.69
L1	*Rosa chuckles*	29	31.5	0.82	0.87	0.89	0.88	0.85	0.85	0.84	0.80	0.76	0.76	0.77	0.75	0.76	0.85	0.79
L1	*Vitis empire*	9	25.7	0.71	0.78	0.82	0.81	0.82	0.77	0.80	0.80	0.77	0.75	0.78	0.58	0.63	0.82	0.74
L1	*Vitis riparia*	33	37.6	0.70	0.78	0.82	0.82	0.82	0.79	0.81	0.79	0.76	0.77	0.78	0.69	0.70	0.80	0.76
L2	*Fagus crenata*	17	78.2	0.76	0.79	0.78	0.79	0.85	0.83	0.80	0.89	0.84	0.89	0.88	0.58	0.53	0.63	0.72
L2	*Amelanchier sanguinea*	32	40.8	0.70	0.73	0.76	0.77	0.80	0.80	0.80	0.78	0.76	0.82	0.79	0.51	0.56	0.77	0.77
L3	*Rhododendron Roseum Superbum*	32	38.7	0.69	0.82	0.84	0.80	0.81	0.75	0.78	0.79	0.78	0.84	0.82	0.58	0.65	0.78	0.77
L4	*Populus tremula*	40	17.6	0.27	0.30	0.25	0.44	0.36	0.35	0.37	0.55	0.48	0.51	0.47	0.23	0.26	0.31	0.37
L4	*Populus tremuloides*	23	24.7	0.56	0.60	0.59	0.72	0.66	0.66	0.58	0.84	0.79	0.84	0.82	0.51	0.57	0.72	0.65
L4	*Populus trichocarpa*	40	26.2	0.66	0.81	0.81	0.79	0.70	0.71	0.65	0.84	0.81	0.87	0.85	0.54	0.71	0.80	0.77
T-DCWP · 39 species																		
L0	*Quercus rubra*	3	31.3	0.18	0.25	0.11	0.07	0.08	0.07	0.06	0.17	0.21	0.06	0.26	0.13	0.23	0.43	0.12
L1	*Viburnum acerifolium*	3	14.3	0.14	0.11	0.02	0.09	0.11	0.17	0.02	0.40	0.51	0.69	0.65	0.29	0.51	0.83	0.75
L1	*Viburnum dentatum*	3	18.0	0.10	0.26	0.01	0.13	0.15	0.14	0.02	0.17	0.47	0.92	0.81	0.36	0.82	0.80	0.80
L1	*Viburnum nudum*	3	8.7	0.06	0.00	0.00	0.14	0.00	0.00	0.01	0.00	0.73	0.00	0.50	0.22	0.76	0.93	0.87
L1	*Viburnum prunifolium*	3	16.0	0.25	0.70	0.21	0.79	0.30	0.37	0.34	0.20	0.80	0.88	0.92	0.71	0.86	0.90	0.63
L1	*Prunus lasiocarpa*	3	14.3	0.34	0.23	0.00	0.10	0.06	0.25	0.04	0.00	0.62	–	0.89	0.50	0.48	0.72	0.56
L1	*Stewartia malacodendron*	2	6.0	0.05	0.10	0.01	0.05	0.18	0.14	0.11	0.10	0.44	0.25	0.75	0.28	0.50	0.56	0.67
L2	*Camellia oleifera*	3	10.0	0.46	0.00	0.00	0.00	0.00	0.00	0.00	–	0.50	–	0.88	0.44	0.57	0.66	0.67
L2	*Amelanchier stolonifera*	3	10.3	0.01	0.04	0.00	0.01	0.05	0.04	0.00	0.05	0.18	0.20	0.75	0.11	0.55	0.83	0.86
L2	*Aronia arbutifolia*	3	31.0	0.31	0.45	0.07	0.28	0.56	0.56	0.24	0.58	0.77	0.93	0.90	0.82	0.91	0.95	0.86
L2	*Crataegus viridis*	2	19.5	0.13	0.02	0.00	0.00	0.03	0.09	0.08	0.00	0.13	–	1.00	0.38	0.97	0.93	0.97
L3	*Cornus alba*	3	16.7	0.03	0.08	0.01	0.03	0.05	0.09	0.03	0.01	0.28	0.71	0.43	0.14	0.63	0.85	0.86
L3	*Kalmia latifolia*	3	14.3	0.15	0.12	0.02	0.14	0.17	0.27	0.16	0.05	0.44	0.11	0.55	0.15	0.18	0.96	0.90
L3	*Sideroxylon lycioides*	3	29.3	0.64	0.84	0.49	0.90	0.73	0.79	0.72	0.88	0.87	0.98	0.93	0.88	0.85	0.97	0.93
L3	*Betula nigra*	3	12.0	0.03	0.06	0.06	0.05	0.07	0.08	0.07	0.04	0.05	0.12	0.05	0.04	0.02	0.05	0.63
L3	*Betula papyrifera*	3	10.0	0.02	0.01	0.00	0.00	0.03	0.02	0.03	0.00	0.27	–	0.86	0.17	0.16	1.00	1.00
L3	*Carpinus caroliniana*	3	39.3	0.47	0.60	0.43	0.59	0.42	0.51	0.19	0.68	0.71	0.75	0.78	0.74	0.57	0.91	0.77
L3	*Carya glabra*	3	28.7	0.70	0.88	0.20	0.84	0.62	0.60	0.29	0.64	0.88	0.93	0.89	0.74	0.76	0.94	0.88
L3	*Morella cerifera*	3	36.0	0.79	0.72	0.47	0.67	0.58	0.66	0.37	0.56	0.77	0.93	0.83	0.87	0.73	0.85	0.83
L3	*Lindera benzoin*	3	23.0	0.23	0.38	0.15	0.29	0.19	0.24	0.17	0.27	0.39	0.35	0.30	0.10	0.18	0.31	0.82
L3	*Persea borbonia*	3	51.0	0.30	0.55	0.14	0.21	0.41	0.53	0.25	0.44	0.68	0.71	0.79	0.33	0.57	0.91	0.84
L3	*Sassafras albidum*	3	39.3	0.18	0.42	0.02	0.37	0.28	0.34	0.07	0.02	0.55	0.58	0.72	0.34	0.71	0.81	0.86
L3	*Zanthoxylum americanum*	3	16.3	0.38	0.14	0.00	0.47	0.08	0.11	0.10	0.00	0.67	–	0.87	0.48	0.90	0.92	0.84
L3	*Zanthoxylum simulans*	3	7.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.21	0.71	1.00	0.96
L3	*Acer rubrum*	3	39.0	0.20	0.26	0.22	0.23	0.27	0.25	0.21	0.23	0.31	0.16	0.38	0.13	0.12	0.16	0.31
L3	*Aesculus parviflora*	3	50.3	0.27	0.32	0.18	0.32	0.44	0.47	0.22	0.42	0.47	0.47	0.69	0.30	0.30	0.51	0.93
L4	*Fatsia japonica*	3	16.0	0.31	0.37	0.01	0.19	0.46	0.44	0.31	0.00	0.88	0.88	0.94	0.78	0.84	0.96	0.86
L4	*Ilex cornuta*	3	30.7	0.70	0.52	0.06	0.73	0.59	0.73	0.62	0.25	0.86	0.94	0.91	0.86	0.98	0.93	0.79
L4	*Moringa oleifera*	3	17.3	0.67	0.54	0.52	0.84	0.74	0.51	0.56	0.75	0.77	0.91	0.86	0.76	0.57	0.98	0.81
L4	*Euonymus americanus*	3	21.0	0.86	0.83	0.72	0.98	0.54	0.51	0.44	0.78	0.73	0.86	0.88	0.87	0.88	0.95	0.76
L4	*Stephanotis floribunda*	5	39.0	0.55	0.79	0.35	0.57	0.74	0.60	0.52	0.35	0.68	0.86	0.66	0.57	0.53	0.94	0.94
L4	*Fraxinus pennsylvanica*	3	14.0	0.10	0.11	0.01	0.22	0.13	0.15	0.17	0.22	0.46	0.57	0.76	0.29	0.72	0.71	0.64
L4	*Olea europaea*	3	27.7	0.58	0.84	0.22	0.68	0.55	0.64	0.39	0.62	0.82	0.91	0.93	0.70	0.83	0.86	0.90
L4	*Magnolia virginiana*	3	49.3	0.86	0.88	0.39	0.89	0.66	0.68	0.47	0.92	0.84	0.91	0.80	0.88	0.79	0.95	0.91
L4	*Passiflora coriacea*	5	40.0	0.72	0.74	0.63	0.60	0.62	0.59	0.54	0.51	0.79	0.78	0.66	0.56	0.39	0.67	0.87
L4	*Hibiscus clayi*	3	11.7	0.64	0.51	0.92	0.76	0.80	0.83	0.87	1.00	0.90	0.97	0.88	0.90	1.00	0.94	0.92
L4	*Edgeworthia chrysantha*	3	12.3	0.02	0.03	0.01	0.12	0.02	0.04	0.07	0.00	0.54	0.00	0.29	0.06	0.36	0.20	0.19
L4	*Fothergilla major*	3	12.3	0.18	0.57	0.07	0.75	0.40	0.31	0.37	0.90	0.90	0.89	0.80	0.49	0.88	0.95	0.90
L4	*Itea virginica*	3	25.0	0.43	0.54	0.39	0.87	0.65	0.76	0.69	0.76	0.84	0.82	0.86	0.79	0.65	0.96	0.93

Figure S22: Precision at $\tau = 0.25$, IoU 0.50, for every detector and every species of OD-Valid, T-MAWP and T-DCWP. Each cell prints the value. N: images. $\bar{c}_{GT}$: mean annotated stomata per image. A dash (–) means the detector predicts no box for that species at $\tau = 0.25$, so precision is undefined.

COCO $AP_{50:95}$ (%), all released scores

lighter is higher

0 50 100

		N	$\bar{c}_{GT}$	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
T-MSWP · 18 species																		
L1	*Quercus michauxii Nutt.*	129	36.7	17.8	22.7	22.6	18.5	17.3	18.8	10.3	22.3	40.6	28.2	37.4	30.7	33.2	42.9	40.8
L1	*Quercus nigra L.*	500	26.8	1.2	1.0	0.9	1.0	2.4	1.1	1.1	4.2	13.5	6.1	4.3	4.7	4.9	16.4	2.1
L1	*Quercus pagoda Raf.*	286	51.4	13.9	17.0	17.6	12.6	15.2	17.1	9.8	13.4	34.3	25.3	24.6	17.2	27.9	43.9	21.3
L1	*Quercus phellos L.*	568	59.9	21.7	18.7	24.4	24.0	19.8	16.8	13.5	34.8	54.4	33.7	44.1	30.3	29.8	52.0	41.0
L1	*Quercus shumardii Buckley*	1,612	71.4	21.5	20.1	23.3	25.0	22.5	19.1	14.2	31.6	48.6	29.8	33.6	33.1	26.3	50.6	24.3
L1	*Quercus stellata Wangenh.*	96	40.2	0.5	0.1	4.7	0.3	2.5	0.7	3.3	4.9	63.6	20.0	20.7	0.1	25.2	53.0	0.0
L1	*Quercus texana Buckley*	2,975	66.8	15.8	16.0	17.5	16.0	16.2	12.8	10.5	23.9	39.5	26.4	23.9	22.4	20.3	45.5	8.9
L3	*Acer rubrum L.*	100	17.6	1.0	0.4	0.3	2.8	1.1	0.3	1.5	6.2	38.9	12.4	11.8	2.1	38.8	32.6	31.1
L3	*Carya ovata (Mill.) K. Koch*	91	23.0	9.0	10.9	9.4	7.6	11.3	14.5	6.4	12.8	33.8	27.9	29.6	19.6	32.3	36.3	52.4
L3	*Nyssa sylvatica Marshall*	50	10.0	1.8	1.8	0.4	2.5	2.7	0.7	4.7	3.7	27.9	6.3	8.7	4.2	37.7	29.1	44.4
L3	*Ulmus alata Michx.*	205	36.0	4.8	6.8	6.7	4.4	6.6	7.2	3.5	5.8	19.0	15.2	14.0	8.5	12.8	21.8	4.1
L3	*Ulmus americana Planch*	185	21.9	3.5	5.6	3.2	1.9	6.1	8.1	4.1	2.8	21.9	19.4	18.6	13.2	15.4	29.6	42.2
L3	*Vaccinium stamineum L.*	75	12.6	0.1	0.0	0.0	0.1	0.7	0.0	0.3	0.1	25.7	2.1	9.5	0.4	16.1	38.6	48.1
L4	*Dirca palustris L.*	65	21.5	6.9	5.2	6.0	7.2	5.0	4.0	6.7	5.8	53.8	12.1	36.8	11.1	45.2	50.3	62.1
L4	*Fraxinus caroliniana Miller*	100	16.4	6.6	4.4	5.0	10.3	5.4	1.6	12.1	14.1	38.7	31.4	26.6	3.7	43.8	32.4	37.0
L4	*Ilex opaca Aiton*	100	37.5	0.0	0.0	0.1	0.0	0.4	0.2	0.1	0.3	36.0	1.1	7.2	0.1	1.3	43.5	40.5
L4	*Populus L.*	3,065	30.9	39.3	33.1	34.8	30.1	41.0	42.1	42.7	19.8	55.8	45.3	51.4	30.5	49.6	46.2	53.1
L4	*Salix L.*	82	14.7	42.3	39.1	43.9	42.1	47.3	42.7	47.6	18.5	59.0	42.5	52.5	33.6	52.1	50.3	57.5
T-GSH · 9 species																		
L4	*Glycine max*	600	55.1	20.9	17.3	11.5	16.0	19.9	23.6	22.9	9.3	21.2	28.7	24.8	19.2	21.1	13.8	30.0
L4	*Phaseolus vulgaris*	597	32.0	31.2	21.7	19.4	22.1	29.8	26.9	30.8	3.8	21.3	27.5	23.7	14.3	19.9	12.5	12.1
L4	*Populus balsamifera*	113	37.4	19.9	5.6	0.7	0.3	13.5	11.8	10.7	0.0	1.8	3.5	9.8	0.7	31.9	56.1	62.8
L5	*Triticum aestivum*	2,443	17.9	4.9	12.9	0.2	0.1	6.4	12.6	3.5	18.4	40.1	38.7	36.0	39.2	38.6	44.7	40.7
L5	*Elaeis guineensis*	317	36.6	5.3	10.7	4.5	5.0	13.0	16.0	25.1	1.3	12.8	13.4	13.1	13.3	15.6	20.2	27.3
L5	*Hordeum vulgare*	542	27.9	1.2	3.0	0.7	3.0	6.6	6.5	8.8	0.1	6.9	4.6	11.7	8.8	8.6	18.7	27.5
L5	*Oryza glaberrima*	253	47.2	2.2	5.2	0.6	0.8	5.1	7.2	6.0	0.0	0.3	0.2	0.5	5.5	4.5	20.0	25.3
L5	*Triticum spp.*	651	78.4	3.8	1.5	3.2	6.0	6.8	6.1	6.0	5.5	3.6	2.8	9.0	2.9	15.1	2.8	7.6
L5	*Zea mays*	1,427	69.3	8.2	5.4	5.8	0.7	3.7	4.8	3.1	1.6	6.6	5.9	10.0	12.1	17.4	27.7	31.3

Figure S23: COCO $AP_{50:95}$ as in Figure S21, for the species of T-MSWP and T-GSH.

Precision at $\tau = 0.25$

lighter is higher

0 0.5 1

		N	$\bar{c}_{GT}$	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
T-MSWP · 18 species																		
L1	*Quercus michauxii Nutt.*	129	36.7	0.70	0.79	0.76	0.71	0.67	0.67	0.51	0.80	0.82	0.87	0.89	0.73	0.78	0.87	0.82
L1	*Quercus nigra L.*	500	26.8	0.21	0.20	0.20	0.20	0.16	0.12	0.12	0.39	0.60	0.45	0.37	0.40	0.44	0.71	0.28
L1	*Quercus pagoda Raf.*	286	51.4	0.56	0.71	0.66	0.56	0.58	0.59	0.45	0.68	0.76	0.81	0.79	0.58	0.67	0.79	0.64
L1	*Quercus phellos L.*	568	59.9	0.71	0.74	0.80	0.76	0.70	0.66	0.52	0.88	0.92	0.90	0.90	0.82	0.83	0.96	0.82
L1	*Quercus shumardii Buckley*	1,612	71.4	0.73	0.74	0.80	0.76	0.70	0.68	0.58	0.81	0.88	0.88	0.85	0.83	0.79	0.96	0.77
L1	*Quercus stellata Wangenh.*	96	40.2	0.12	0.06	0.11	0.09	0.01	0.01	0.05	0.11	0.90	0.88	0.74	0.03	0.83	0.95	0.04
L1	*Quercus texana Buckley*	2,975	66.8	0.65	0.66	0.72	0.65	0.61	0.56	0.49	0.74	0.81	0.83	0.76	0.76	0.76	0.93	0.63
L3	*Acer rubrum L.*	100	17.6	0.19	0.03	0.09	0.38	0.05	0.02	0.22	0.47	0.92	0.77	0.64	0.24	0.79	0.88	0.79
L3	*Carya ovata (Mill.) K. Koch*	91	23.0	0.49	0.63	0.48	0.41	0.53	0.54	0.39	0.63	0.77	0.84	0.81	0.59	0.64	0.83	0.86
L3	*Nyssa sylvatica Marshall*	50	10.0	0.49	0.50	0.22	0.37	0.37	0.11	0.51	0.45	0.83	0.69	0.63	0.25	0.62	0.81	0.90
L3	*Ulmus alata Michx.*	205	36.0	0.42	0.56	0.48	0.37	0.48	0.45	0.28	0.49	0.68	0.74	0.65	0.49	0.59	0.70	0.46
L3	*Ulmus americana Planch*	185	21.9	0.24	0.43	0.18	0.18	0.43	0.44	0.29	0.42	0.79	0.85	0.80	0.48	0.46	0.79	0.78
L3	*Vaccinium stamineum L.*	75	12.6	0.06	0.02	0.04	0.07	0.00	0.00	0.04	0.05	0.82	0.45	0.59	0.08	0.76	0.93	0.96
L4	*Dirca palustris L.*	65	21.5	0.54	0.38	0.46	0.39	0.35	0.26	0.46	0.45	0.89	0.85	0.85	0.50	0.84	0.89	0.89
L4	*Fraxinus caroliniana Miller*	100	16.4	0.64	0.45	0.59	0.58	0.46	0.23	0.64	0.61	0.92	0.93	0.88	0.36	0.85	0.88	0.87
L4	*Ilex opaca Aiton*	100	37.5	0.06	0.00	0.03	0.02	0.00	0.02	0.01	0.11	0.85	0.32	0.55	0.02	0.30	0.99	0.69
L4	*Populus L.*	3,065	30.9	0.77	0.93	0.82	0.80	0.85	0.82	0.78	0.70	0.93	0.96	0.95	0.80	0.92	0.95	0.97
L4	*Salix L.*	82	14.7	0.93	0.96	0.95	0.84	0.95	0.82	0.90	0.77	0.94	0.94	0.96	0.67	0.88	0.95	0.97
T-GSH · 9 species																		
L4	*Glycine max*	600	55.1	0.73	0.94	0.41	0.81	0.56	0.73	0.63	0.43	0.75	0.93	0.77	0.42	0.56	0.59	0.84
L4	*Phaseolus vulgaris*	597	32.0	0.57	0.90	0.27	0.86	0.41	0.52	0.45	0.20	0.67	0.87	0.57	0.48	0.51	0.59	0.62
L4	*Populus balsamifera*	113	37.4	0.66	0.52	0.06	0.14	0.41	0.46	0.48	0.01	0.23	0.64	0.65	0.10	0.91	0.95	0.98
L5	*Triticum aestivum*	2,443	17.9	0.17	0.85	0.02	0.02	0.27	0.53	0.20	0.44	0.74	0.83	0.61	0.58	0.87	0.88	0.81
L5	*Elaeis guineensis*	317	36.6	0.54	0.88	0.29	0.69	0.77	0.88	0.86	0.19	0.81	0.85	0.72	0.36	0.78	0.74	0.81
L5	*Hordeum vulgare*	542	27.9	0.34	0.79	0.11	0.48	0.56	0.66	0.73	0.10	0.57	0.57	0.64	0.37	0.70	0.56	0.76
L5	*Oryza glaberrima*	253	47.2	0.15	0.79	0.06	0.21	0.32	0.52	0.37	0.01	0.06	0.03	0.04	0.19	0.18	0.62	0.77
L5	*Triticum spp.*	651	78.4	0.64	0.61	0.60	0.73	0.70	0.71	0.77	0.77	0.69	0.77	0.78	0.43	0.68	0.33	0.61
L5	*Zea mays*	1,427	69.3	0.72	0.77	0.58	0.29	0.61	0.65	0.39	0.30	0.66	0.67	0.77	0.67	0.88	0.90	0.95

Figure S24: Precision as in Figure S22, for the species of T-MSWP and T-GSH.

Recall at $\tau = 0.25$

lighter is higher

0 — 0.5 — 1

		N	$\bar{c}_{GT}$	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
OD-Valid · 5 species																		
L0	*Actinidia arguta*	37	38.8	0.93	0.85	0.81	0.78	0.79	0.77	0.77	0.86	0.90	0.91	0.89	0.95	0.97	0.88	0.96
L0	*Malus hupehensis*	41	79.9	0.84	0.82	0.83	0.81	0.70	0.67	0.71	0.85	0.86	0.86	0.87	0.92	0.93	0.88	0.88
L0	*Malus spontanea*	40	76.7	0.85	0.83	0.83	0.82	0.70	0.66	0.71	0.85	0.88	0.87	0.88	0.94	0.95	0.90	0.88
L0	*Malus toringoides*	9	74.6	0.77	0.72	0.73	0.70	0.58	0.55	0.60	0.73	0.77	0.77	0.77	0.91	0.93	0.89	0.78
L0	*Vitis vinifera*	11	31.1	0.96	0.92	0.87	0.88	0.92	0.92	0.92	0.87	0.91	0.90	0.88	0.95	0.95	0.87	0.93
T-MAWP · 17 species																		
L1	*Actinidia coriacea*	37	27.9	0.89	0.81	0.80	0.80	0.84	0.80	0.84	0.82	0.89	0.91	0.88	0.95	0.96	0.85	0.92
L1	*Malus domestica*	45	34.0	0.82	0.74	0.78	0.75	0.73	0.70	0.78	0.75	0.82	0.78	0.81	0.89	0.92	0.85	0.88
L1	*Prunus Tai-Haku*	47	35.6	0.99	0.97	0.97	0.96	0.96	0.94	0.94	0.96	0.99	0.99	0.99	0.99	0.99	0.97	0.99
L1	*Prunus maritima*	12	98.1	0.45	0.38	0.60	0.50	0.47	0.41	0.55	0.65	0.66	0.71	0.45	0.71	0.71	0.64	0.53
L1	*Prunus sargentii*	18	50.4	0.76	0.52	0.63	0.61	0.65	0.56	0.64	0.74	0.73	0.74	0.68	0.87	0.89	0.75	0.74
L1	*Prunus subhirtella*	20	87.5	0.74	0.56	0.76	0.55	0.63	0.52	0.72	0.75	0.90	0.92	0.89	0.93	0.97	0.94	0.91
L1	*Rosa Radtkopink*	12	18.5	0.97	0.92	0.89	0.92	0.96	0.96	0.97	0.95	0.98	0.98	0.96	0.98	0.99	0.94	1.00
L1	*Rosa centifolia*	10	28.5	0.83	0.68	0.72	0.74	0.80	0.77	0.79	0.81	0.91	0.93	0.88	0.92	0.96	0.87	0.96
L1	*Rosa chuckles*	29	31.5	0.96	0.93	0.91	0.91	0.94	0.94	0.92	0.93	0.97	0.96	0.96	0.98	0.99	0.94	0.99
L1	*Vitis empire*	9	25.7	0.93	0.88	0.85	0.84	0.87	0.85	0.87	0.83	0.90	0.89	0.87	0.97	0.97	0.89	0.93
L1	*Vitis riparia*	33	37.6	0.96	0.94	0.91	0.88	0.89	0.90	0.92	0.90	0.94	0.94	0.91	0.97	0.98	0.92	0.96
L2	*Fagus crenata*	17	78.2	0.55	0.32	0.32	0.37	0.44	0.40	0.51	0.29	0.56	0.53	0.47	0.79	0.85	0.60	0.76
L2	*Amelanchier sanguinea*	32	40.8	0.90	0.86	0.87	0.86	0.77	0.73	0.81	0.80	0.87	0.81	0.83	0.96	0.94	0.85	0.87
L3	*Rhododendron Roseum Superbum*	32	38.7	0.93	0.74	0.72	0.83	0.76	0.74	0.79	0.81	0.90	0.83	0.84	0.98	0.99	0.93	0.95
L4	*Populus tremula*	40	17.6	0.70	0.46	0.47	0.50	0.55	0.55	0.58	0.44	0.63	0.57	0.61	0.88	0.83	0.81	0.83
L4	*Populus tremuloides*	23	24.7	0.78	0.57	0.60	0.66	0.72	0.75	0.73	0.63	0.80	0.69	0.71	0.90	0.89	0.75	0.86
L4	*Populus trichocarpa*	40	26.2	0.86	0.70	0.51	0.70	0.83	0.83	0.84	0.67	0.83	0.71	0.74	0.95	0.91	0.80	0.91
T-DCWP · 39 species																		
L0	*Quercus rubra*	3	31.3	0.21	0.26	0.14	0.06	0.09	0.07	0.06	0.13	0.22	0.05	0.24	0.14	0.24	0.44	0.13
L1	*Viburnum acerifolium*	3	14.3	0.35	0.19	0.07	0.14	0.28	0.33	0.07	0.09	0.49	0.21	0.60	0.58	0.67	0.93	0.98
L1	*Viburnum dentatum*	3	18.0	0.65	0.33	0.09	0.20	0.52	0.46	0.11	0.07	0.37	0.43	0.80	0.93	0.93	0.98	0.98
L1	*Viburnum nudum*	3	8.7	0.12	0.00	0.00	0.12	0.00	0.00	0.04	0.00	0.42	0.00	0.04	0.42	0.85	1.00	1.00
L1	*Viburnum prunifolium*	3	16.0	0.83	0.67	0.29	0.62	0.50	0.48	0.44	0.04	0.98	0.29	0.50	0.96	0.88	0.90	0.98
L1	*Prunus lasiocarpa*	3	14.3	0.65	0.12	0.00	0.09	0.12	0.30	0.14	0.00	0.12	0.00	0.58	0.58	0.74	0.72	0.65
L1	*Stewartia malacodendron*	2	6.0	0.58	0.50	0.17	0.08	0.58	0.42	0.50	0.08	0.33	0.08	0.25	1.00	0.92	0.83	1.00
L2	*Camellia oleifera*	3	10.0	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.47	0.23	0.87	0.97	0.93
L2	*Amelanchier stolonifera*	3	10.3	0.10	0.26	0.00	0.03	0.42	0.26	0.03	0.03	0.13	0.03	0.77	0.94	0.90	0.94	1.00
L2	*Aronia arbutifolia*	3	31.0	0.86	0.61	0.25	0.30	0.76	0.77	0.39	0.15	0.89	0.69	0.96	0.99	0.97	0.99	1.00
L2	*Crataegus viridis*	2	19.5	0.54	0.03	0.00	0.00	0.13	0.31	0.21	0.00	0.08	0.00	0.72	0.77	0.79	1.00	1.00
L3	*Cornus alba*	3	16.7	0.34	0.46	0.06	0.16	0.40	0.50	0.26	0.04	0.54	0.78	0.90	0.98	0.82	0.92	1.00
L3	*Kalmia latifolia*	3	14.3	1.00	0.49	0.07	0.42	0.70	0.91	0.63	0.05	0.93	0.09	0.74	0.86	1.00	1.00	1.00
L3	*Sideroxylon lycioides*	3	29.3	0.95	0.92	0.78	0.85	0.97	0.94	0.83	0.67	0.95	0.60	0.88	0.97	0.97	0.99	1.00
L3	*Betula nigra*	3	12.0	0.22	0.17	0.25	0.25	0.36	0.36	0.36	0.11	0.19	0.19	0.17	0.44	0.36	0.69	0.92
L3	*Betula papyrifera*	3	10.0	0.13	0.03	0.00	0.00	0.17	0.07	0.13	0.00	0.20	0.00	0.40	0.60	0.73	0.90	1.00
L3	*Carpinus caroliniana*	3	39.3	0.58	0.58	0.47	0.61	0.50	0.60	0.24	0.46	0.80	0.59	0.69	0.90	0.74	1.00	0.87
L3	*Carya glabra*	3	28.7	0.88	0.77	0.19	0.49	0.62	0.72	0.44	0.21	0.98	0.62	0.78	0.86	0.98	0.98	0.99
L3	*Morella cerifera*	3	36.0	0.80	0.65	0.32	0.55	0.82	0.85	0.51	0.43	0.89	0.61	0.64	0.88	0.79	0.89	0.94
L3	*Lindera benzoin*	3	23.0	0.65	0.39	0.14	0.43	0.42	0.46	0.41	0.26	0.43	0.26	0.49	0.43	0.64	0.94	0.96
L3	*Persea borbonia*	3	51.0	0.56	0.53	0.29	0.24	0.60	0.61	0.41	0.22	0.80	0.77	0.68	0.72	0.90	0.96	1.00
L3	*Sassafras albidum*	3	39.3	0.29	0.31	0.03	0.22	0.41	0.46	0.12	0.01	0.47	0.28	0.32	0.53	0.69	0.94	0.93
L3	*Zanthoxylum americanum*	3	16.3	0.94	0.10	0.00	0.31	0.10	0.12	0.18	0.00	0.82	0.00	0.27	0.90	0.94	0.98	1.00
L3	*Zanthoxylum simulans*	3	7.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.27	0.68	1.00	1.00
L3	*Acer rubrum*	3	39.0	0.75	0.72	0.60	0.55	0.56	0.48	0.46	0.36	0.59	0.63	0.39	0.74	0.73	0.86	0.97
L3	*Aesculus parviflora*	3	50.3	0.62	0.64	0.39	0.56	0.58	0.57	0.38	0.56	0.71	0.60	0.79	0.81	0.77	0.92	0.93
L4	*Fatsia japonica*	3	16.0	0.94	0.31	0.06	0.23	0.75	0.79	0.83	0.00	0.79	0.15	0.33	0.81	1.00	0.98	1.00
L4	*Ilex cornuta*	3	30.7	1.00	0.47	0.10	0.51	0.89	0.95	0.86	0.01	0.95	0.52	0.75	0.90	0.99	0.96	0.92
L4	*Moringa oleifera*	3	17.3	0.90	0.92	0.90	0.90	0.92	0.90	0.88	0.46	0.77	0.79	0.81	0.96	0.98	0.85	1.00
L4	*Euonymus americanus*	3	21.0	0.97	0.76	0.57	0.89	0.57	0.59	0.49	0.49	0.98	0.59	0.94	0.92	0.97	0.97	1.00
L4	*Stephanotis floribunda*	5	39.0	0.90	0.62	0.67	0.57	0.82	0.79	0.70	0.44	0.83	0.58	0.73	0.90	0.93	0.90	0.94
L4	*Fraxinus pennsylvanica*	3	14.0	0.71	0.45	0.10	0.40	0.71	0.69	0.67	0.05	0.43	0.38	0.60	0.98	0.98	1.00	1.00
L4	*Olea europaea*	3	27.7	0.80	0.77	0.60	0.76	0.93	0.86	0.71	0.35	0.89	0.70	0.89	0.95	0.89	0.96	1.00
L4	*Magnolia virginiana*	3	49.3	0.93	0.91	0.26	0.66	0.64	0.89	0.61	0.61	0.99	0.58	0.41	0.85	0.97	1.00	1.00
L4	*Passiflora coriacea*	5	40.0	0.81	0.50	0.71	0.65	0.60	0.56	0.60	0.36	0.66	0.68	0.56	0.82	0.73	0.79	0.96
L4	*Hibiscus clayi*	3	11.7	1.00	0.97	0.94	0.97	0.94	1.00	0.97	0.20	1.00	0.91	1.00	1.00	1.00	0.97	1.00
L4	*Edgeworthia chrysantha*	3	12.3	0.05	0.03	0.03	0.11	0.05	0.08	0.22	0.00	0.35	0.00	0.11	0.32	0.46	0.78	0.86
L4	*Fothergilla major*	3	12.3	0.84	0.70	0.32	0.81	0.70	0.62	0.73	0.49	0.95	0.68	0.95	0.92	0.95	0.95	1.00
L4	*Itea virginica*	3	25.0	0.96	0.97	0.79	0.99	0.96	0.99	0.87	0.93	0.99	0.84	0.97	0.99	0.99	1.00	1.00

Figure S25: Recall at $\tau = 0.25$, IoU 0.50, for every detector and every species of OD-Valid, T-MAWP and T-DCWP. Each cell prints the value. N: images. $\bar{c}_{GT}$: mean annotated stomata per image.

F1 at $\tau = 0.25$

lighter is higher

0 — 0.5 — 1

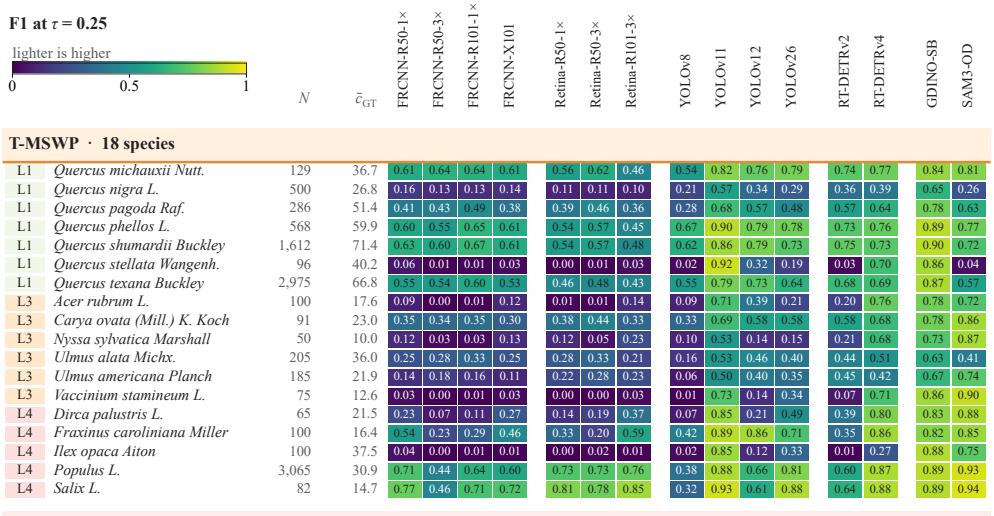
		N	$\bar{c}_{GT}$	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
OD-Valid · 5 species																		
L0	*Actinidia arguta*	37	38.8	0.91	0.88	0.85	0.84	0.86	0.85	0.83	0.87	0.88	0.89	0.88	0.90	0.91	0.92	0.93
L0	*Malus hupehensis*	41	79.9	0.90	0.89	0.90	0.89	0.82	0.80	0.82	0.88	0.88	0.88	0.88	0.91	0.92	0.92	0.91
L0	*Malus spontanea*	40	76.7	0.89	0.88	0.88	0.88	0.81	0.79	0.82	0.87	0.88	0.88	0.88	0.87	0.87	0.89	0.89
L0	*Malus toringoides*	9	74.6	0.79	0.78	0.77	0.77	0.71	0.68	0.72	0.77	0.80	0.80	0.79	0.72	0.77	0.81	0.79
L0	*Vitis vinifera*	11	31.1	0.82	0.85	0.87	0.85	0.87	0.87	0.88	0.81	0.84	0.82	0.83	0.80	0.82	0.84	0.84
T-MAWP · 17 species																		
L1	*Actinidia coriacea*	37	27.9	0.79	0.80	0.80	0.82	0.82	0.81	0.80	0.82	0.83	0.84	0.84	0.79	0.77	0.84	0.84
L1	*Malus domestica*	45	34.0	0.79	0.76	0.78	0.79	0.78	0.76	0.79	0.78	0.81	0.80	0.80	0.72	0.73	0.78	0.77
L1	*Prunus Tai-Haku*	47	35.6	0.87	0.91	0.91	0.91	0.91	0.90	0.90	0.89	0.88	0.88	0.88	0.82	0.87	0.91	0.88
L1	*Prunus maritima*	12	98.1	0.49	0.46	0.61	0.55	0.55	0.49	0.60	0.60	0.58	0.57	0.49	0.55	0.51	0.49	0.48
L1	*Prunus sargentii*	18	50.4	0.69	0.59	0.67	0.66	0.69	0.64	0.68	0.67	0.60	0.60	0.62	0.64	0.64	0.62	0.59
L1	*Prunus subhirtella*	20	87.5	0.78	0.69	0.81	0.68	0.74	0.66	0.79	0.80	0.83	0.84	0.83	0.78	0.72	0.79	0.79
L1	*Rosa Radtkopink*	12	18.5	0.86	0.89	0.91	0.92	0.90	0.90	0.90	0.88	0.88	0.87	0.88	0.87	0.86	0.91	0.91
L1	*Rosa centifolia*	10	28.5	0.80	0.75	0.78	0.79	0.80	0.81	0.80	0.78	0.79	0.78	0.81	0.76	0.79	0.81	0.80
L1	*Rosa chuckles*	29	31.5	0.89	0.90	0.90	0.89	0.89	0.89	0.88	0.86	0.85	0.85	0.85	0.85	0.86	0.89	0.88
L1	*Vitis empire*	9	25.7	0.80	0.83	0.83	0.83	0.84	0.81	0.83	0.81	0.83	0.81	0.82	0.73	0.77	0.85	0.82
L1	*Vitis riparia*	33	37.6	0.81	0.85	0.86	0.85	0.85	0.84	0.86	0.84	0.84	0.85	0.84	0.80	0.81	0.85	0.85
L2	*Fagus crenata*	17	78.2	0.64	0.46	0.46	0.50	0.58	0.54	0.62	0.44	0.67	0.66	0.61	0.67	0.65	0.61	0.74
L2	*Amelanchier sanguinea*	32	40.8	0.78	0.79	0.81	0.81	0.78	0.76	0.80	0.79	0.81	0.82	0.81	0.67	0.70	0.81	0.82
L3	*Rhododendron Roseum Superbum*	32	38.7	0.80	0.78	0.77	0.82	0.79	0.74	0.79	0.80	0.83	0.83	0.83	0.73	0.78	0.85	0.85
L4	*Populus tremula*	40	17.6	0.39	0.36	0.32	0.47	0.44	0.43	0.45	0.49	0.55	0.54	0.53	0.37	0.40	0.45	0.51
L4	*Populus tremuloides*	23	24.7	0.65	0.58	0.60	0.69	0.69	0.70	0.65	0.72	0.79	0.76	0.76	0.65	0.70	0.73	0.74
L4	*Populus trichocarpa*	40	26.2	0.75	0.75	0.63	0.74	0.76	0.77	0.73	0.75	0.82	0.78	0.79	0.69	0.80	0.80	0.84
T-DCWP · 39 species																		
L0	*Quercus rubra*	3	31.3	0.19	0.25	0.12	0.06	0.08	0.07	0.06	0.15	0.22	0.06	0.25	0.14	0.24	0.43	0.12
L1	*Viburnum acerifolium*	3	14.3	0.20	0.14	0.03	0.11	0.16	0.22	0.03	0.15	0.50	0.32	0.63	0.39	0.58	0.88	0.85
L1	*Viburnum dentatum*	3	18.0	0.17	0.29	0.02	0.16	0.23	0.22	0.04	0.10	0.41	0.58	0.80	0.52	0.87	0.88	0.88
L1	*Viburnum nudum*	3	8.7	0.08	0.00	0.00	0.13	0.00	0.00	0.02	0.00	0.54	0.00	0.07	0.29	0.80	0.96	0.93
L1	*Viburnum prunifolium*	3	16.0	0.39	0.68	0.24	0.70	0.38	0.41	0.39	0.07	0.88	0.44	0.65	0.81	0.87	0.90	0.76
L1	*Prunus lasiocarpa*	3	14.3	0.44	0.15	0.00	0.10	0.08	0.28	0.06	0.00	0.20	0.00	0.70	0.54	0.58	0.72	0.60
L1	*Stewartia malacodendron*	2	6.0	0.10	0.16	0.02	0.06	0.28	0.21	0.18	0.09	0.38	0.12	0.38	0.44	0.65	0.67	0.80
L2	*Camellia oleifera*	3	10.0	0.28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.61	0.30	0.68	0.78	0.78
L2	*Amelanchier stolonifera*	3	10.3	0.01	0.07	0.00	0.01	0.09	0.06	0.01	0.04	0.15	0.06	0.76	0.19	0.68	0.88	0.93
L2	*Aronia arbutifolia*	3	31.0	0.45	0.52	0.10	0.29	0.65	0.65	0.30	0.24	0.83	0.79	0.93	0.90	0.94	0.97	0.93
L2	*Crataegus viridis*	2	19.5	0.21	0.02	0.00	0.00	0.05	0.14	0.11	0.00	0.10	0.00	0.84	0.51	0.87	0.96	0.99
L3	*Cornus alba*	3	16.7	0.06	0.14	0.01	0.05	0.08	0.16	0.06	0.01	0.36	0.74	0.58	0.25	0.71	0.88	0.93
L3	*Kalmia latifolia*	3	14.3	0.27	0.19	0.03	0.21	0.27	0.41	0.26	0.05	0.60	0.10	0.63	0.26	0.30	0.98	0.95
L3	*Sideroxylon lycioides*	3	29.3	0.77	0.88	0.60	0.88	0.83	0.86	0.77	0.76	0.91	0.75	0.90	0.92	0.90	0.98	0.96
L3	*Betula nigra*	3	12.0	0.05	0.09	0.09	0.09	0.12	0.13	0.11	0.06	0.08	0.15	0.08	0.07	0.04	0.09	0.75
L3	*Betula papyrifera*	3	10.0	0.04	0.02	0.00	0.00	0.06	0.04	0.05	0.00	0.23	0.00	0.55	0.26	0.27	0.95	1.00
L3	*Carpinus caroliniana*	3	39.3	0.52	0.59	0.45	0.60	0.46	0.55	0.21	0.55	0.75	0.66	0.74	0.81	0.64	0.96	0.82
L3	*Carya glabra*	3	28.7	0.78	0.82	0.19	0.62	0.62	0.66	0.35	0.32	0.93	0.74	0.83	0.80	0.86	0.96	0.93
L3	*Morella cerifera*	3	36.0	0.79	0.68	0.38	0.60	0.68	0.74	0.43	0.48	0.82	0.74	0.72	0.88	0.76	0.87	0.88
L3	*Lindera benzoin*	3	23.0	0.34	0.38	0.15	0.34	0.27	0.31	0.24	0.26	0.41	0.30	0.37	0.17	0.28	0.46	0.89
L3	*Persea borbonia*	3	51.0	0.39	0.54	0.19	0.22	0.49	0.57	0.31	0.29	0.73	0.74	0.73	0.45	0.69	0.94	0.91
L3	*Sassafras albidum*	3	39.3	0.22	0.35	0.02	0.28	0.33	0.39	0.08	0.01	0.51	0.38	0.44	0.42	0.70	0.87	0.89
L3	*Zanthoxylum americanum*	3	16.3	0.54	0.12	0.00	0.37	0.09	0.11	0.13	0.00	0.73	0.00	0.41	0.63	0.92	0.95	0.92
L3	*Zanthoxylum simulans*	3	7.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.24	0.70	1.00	0.98
L3	*Acer rubrum*	3	39.0	0.32	0.38	0.32	0.33	0.36	0.33	0.29	0.28	0.41	0.26	0.38	0.21	0.21	0.27	0.47
L3	*Aesculus parviflora*	3	50.3	0.38	0.42	0.25	0.40	0.50	0.51	0.28	0.48	0.56	0.53	0.74	0.44	0.43	0.66	0.93
L4	*Fatsia japonica*	3	16.0	0.46	0.34	0.02	0.21	0.57	0.57	0.45	0.00	0.84	0.25	0.49	0.80	0.91	0.97	0.92
L4	*Ilex cornuta*	3	30.7	0.83	0.49	0.08	0.60	0.71	0.82	0.72	0.02	0.90	0.67	0.82	0.88	0.98	0.94	0.85
L4	*Moringa oleifera*	3	17.3	0.77	0.68	0.66	0.87	0.82	0.65	0.69	0.57	0.77	0.85	0.83	0.85	0.72	0.91	0.90
L4	*Euonymus americanus*	3	21.0	0.91	0.79	0.64	0.93	0.55	0.55	0.46	0.60	0.84	0.70	0.91	0.89	0.92	0.96	0.86
L4	*Stephanotis floribunda*	5	39.0	0.68	0.69	0.46	0.57	0.77	0.68	0.60	0.39	0.74	0.69	0.69	0.70	0.68	0.92	0.94
L4	*Fraxinus pennsylvanica*	3	14.0	0.17	0.17	0.02	0.29	0.22	0.24	0.27	0.08	0.44	0.46	0.67	0.45	0.83	0.83	0.78
L4	*Olea europaea*	3	27.7	0.67	0.81	0.32	0.72	0.69	0.73	0.50	0.45	0.86	0.79	0.91	0.81	0.86	0.91	0.95
L4	*Magnolia virginiana*	3	49.3	0.89	0.89	0.32	0.76	0.65	0.77	0.53	0.73	0.91	0.71	0.54	0.86	0.87	0.97	0.95
L4	*Passiflora coriacea*	5	40.0	0.76	0.60	0.67	0.62	0.61	0.57	0.57	0.42	0.72	0.73	0.61	0.66	0.51	0.72	0.91
L4	*Hibiscus clayi*	3	11.7	0.78	0.67	0.93	0.85	0.87	0.91	0.92	0.33	0.95	0.94	0.93	0.95	1.00	0.96	0.96
L4	*Edgeworthia chrysantha*	3	12.3	0.03	0.03	0.01	0.12	0.03	0.05	0.10	0.00	0.43	0.00	0.16	0.10	0.40	0.32	0.31
L4	*Fothergilla major*	3	12.3	0.30	0.63	0.11	0.78	0.51	0.41	0.49	0.63	0.92	0.77	0.86	0.64	0.91	0.95	0.95
L4	*Itea virginica*	3	25.0	0.60	0.70	0.52	0.93	0.78	0.86	0.77	0.84	0.91	0.83	0.91	0.88	0.78	0.98	0.96

Figure S26: F1 at $\tau = 0.25$, IoU 0.50, for every detector and every species of OD-Valid, T-MAWP and T-DCWP. Each cell prints the value. N: images. $\bar{c}_{GT}$: mean annotated stomata per image.



		N	$\bar{c}_{GT}$	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
T-MSWP · 18 species																		
L1	*Quercus michauxii Nutt.*	129	36.7	0.54	0.53	0.55	0.53	0.48	0.57	0.41	0.41	0.81	0.67	0.71	0.75	0.77	0.81	0.81
L1	*Quercus nigra L.*	500	26.8	0.13	0.10	0.10	0.11	0.08	0.09	0.09	0.15	0.54	0.28	0.24	0.32	0.36	0.60	0.24
L1	*Quercus pagoda Raf.*	286	51.4	0.32	0.31	0.39	0.28	0.30	0.38	0.30	0.18	0.62	0.44	0.34	0.56	0.61	0.77	0.62
L1	*Quercus phellos L.*	568	59.9	0.51	0.43	0.55	0.52	0.44	0.50	0.40	0.54	0.88	0.70	0.68	0.66	0.70	0.83	0.74
L1	*Quercus shumardii Buckley*	1,612	71.4	0.56	0.51	0.58	0.51	0.43	0.49	0.41	0.50	0.84	0.72	0.64	0.69	0.69	0.84	0.67
L1	*Quercus stellata Wangenh.*	96	40.2	0.04	0.01	0.01	0.02	0.00	0.01	0.03	0.01	0.94	0.19	0.11	0.02	0.60	0.79	0.04
L1	*Quercus texana Buckley*	2,975	66.8	0.48	0.45	0.51	0.45	0.38	0.43	0.39	0.44	0.78	0.66	0.56	0.61	0.64	0.82	0.52
L3	*Acer rubrum L.*	100	17.6	0.06	0.00	0.01	0.07	0.01	0.01	0.10	0.05	0.57	0.26	0.12	0.17	0.74	0.71	0.67
L3	*Carya ovata (Mill.) K. Koch*	91	23.0	0.27	0.23	0.28	0.23	0.29	0.36	0.29	0.22	0.63	0.44	0.45	0.57	0.72	0.75	0.86
L3	*Nyssa sylvatica Marshall*	50	10.0	0.07	0.02	0.02	0.08	0.07	0.03	0.15	0.06	0.39	0.08	0.09	0.18	0.74	0.67	0.84
L3	*Ulmus alata Michx.*	205	36.0	0.18	0.19	0.25	0.19	0.20	0.26	0.17	0.09	0.44	0.33	0.29	0.39	0.45	0.57	0.37
L3	*Ulmus americana Planch*	185	21.9	0.10	0.11	0.14	0.08	0.15	0.21	0.19	0.03	0.37	0.27	0.23	0.43	0.38	0.58	0.70
L3	*Vaccinium stamineum L.*	75	12.6	0.02	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.65	0.08	0.24	0.06	0.67	0.79	0.86
L4	*Dirca palustris L.*	65	21.5	0.15	0.04	0.07	0.21	0.09	0.16	0.31	0.04	0.81	0.12	0.35	0.31	0.77	0.77	0.87
L4	*Fraxinus caroliniana Miller*	100	16.4	0.47	0.15	0.19	0.39	0.26	0.18	0.55	0.33	0.87	0.80	0.60	0.34	0.87	0.76	0.82
L4	*Ilex opaca Aiton*	100	37.5	0.04	0.00	0.01	0.01	0.00	0.02	0.01	0.01	0.85	0.08	0.23	0.01	0.24	0.80	0.81
L4	*Populus L.*	3,065	30.9	0.66	0.29	0.52	0.47	0.63	0.66	0.74	0.26	0.84	0.51	0.71	0.48	0.83	0.84	0.90
L4	*Salix L.*	82	14.7	0.66	0.30	0.57	0.63	0.70	0.75	0.82	0.20	0.92	0.45	0.81	0.61	0.88	0.83	0.92
T-GSH · 9 species																		
L4	*Glycine max*	600	55.1	0.39	0.10	0.30	0.29	0.52	0.52	0.56	0.25	0.38	0.31	0.44	0.59	0.54	0.58	0.86
L4	*Phaseolus vulgaris*	597	32.0	0.64	0.31	0.55	0.30	0.70	0.65	0.69	0.15	0.47	0.46	0.49	0.54	0.59	0.58	0.64
L4	*Populus balsamifera*	113	37.4	0.66	0.15	0.08	0.01	0.50	0.51	0.48	0.00	0.05	0.04	0.18	0.08	0.62	0.86	0.94
L5	*Triticum aestivum*	2,443	17.9	0.17	0.12	0.02	0.02	0.28	0.33	0.20	0.46	0.78	0.73	0.83	0.88	0.73	0.86	0.89
L5	*Elaeis guineensis*	317	36.6	0.07	0.14	0.08	0.07	0.44	0.34	0.47	0.01	0.37	0.20	0.49	0.64	0.48	0.73	0.82
L5	*Hordeum vulgare*	542	27.9	0.02	0.01	0.01	0.06	0.16	0.08	0.18	0.00	0.11	0.04	0.31	0.40	0.19	0.77	0.86
L5	*Oryza glaberrima*	253	47.2	0.13	0.01	0.05	0.04	0.17	0.18	0.20	0.04	0.06	0.02	0.07	0.34	0.18	0.58	0.77
L5	*Triticum spp.*	651	78.4	0.10	0.03	0.08	0.13	0.19	0.15	0.10	0.10	0.05	0.03	0.15	0.15	0.44	0.29	0.52
L5	*Zea mays*	1,427	69.3	0.31	0.01	0.13	0.05	0.10	0.09	0.07	0.13	0.39	0.22	0.50	0.40	0.50	0.90	0.84

Figure S27: Recall as in Figure S25, for the species of T-MSWP and T-GSH.



		N	$\bar{c}_{GT}$	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
T-MSWP · 18 species																		
L1	*Quercus michauxii Nutt.*	129	36.7	0.61	0.64	0.64	0.61	0.56	0.62	0.46	0.54	0.82	0.76	0.79	0.74	0.77	0.84	0.81
L1	*Quercus nigra L.*	500	26.8	0.16	0.13	0.13	0.14	0.11	0.11	0.10	0.21	0.57	0.34	0.29	0.36	0.39	0.65	0.26
L1	*Quercus pagoda Raf.*	286	51.4	0.41	0.43	0.49	0.38	0.39	0.46	0.36	0.28	0.68	0.57	0.48	0.57	0.64	0.78	0.63
L1	*Quercus phellos L.*	568	59.9	0.60	0.55	0.65	0.61	0.54	0.57	0.45	0.67	0.90	0.79	0.78	0.73	0.76	0.89	0.77
L1	*Quercus shumardii Buckley*	1,612	71.4	0.63	0.60	0.67	0.61	0.54	0.57	0.48	0.62	0.86	0.79	0.73	0.75	0.73	0.90	0.72
L1	*Quercus stellata Wangenh.*	96	40.2	0.06	0.01	0.01	0.03	0.00	0.01	0.03	0.02	0.92	0.32	0.19	0.03	0.70	0.86	0.04
L1	*Quercus texana Buckley*	2,975	66.8	0.55	0.54	0.60	0.53	0.46	0.48	0.43	0.55	0.79	0.73	0.64	0.68	0.69	0.87	0.57
L3	*Acer rubrum L.*	100	17.6	0.09	0.00	0.01	0.12	0.01	0.01	0.14	0.09	0.71	0.39	0.21	0.20	0.76	0.78	0.72
L3	*Carya ovata (Mill.) K. Koch*	91	23.0	0.35	0.34	0.35	0.30	0.38	0.44	0.33	0.33	0.69	0.58	0.58	0.58	0.68	0.78	0.86
L3	*Nyssa sylvatica Marshall*	50	10.0	0.12	0.03	0.03	0.13	0.12	0.05	0.23	0.10	0.53	0.14	0.15	0.21	0.68	0.73	0.87
L3	*Ulmus alata Michx.*	205	36.0	0.25	0.28	0.33	0.25	0.28	0.33	0.21	0.16	0.53	0.46	0.40	0.44	0.51	0.63	0.41
L3	*Ulmus americana Planch*	185	21.9	0.14	0.18	0.16	0.11	0.22	0.28	0.23	0.06	0.50	0.40	0.35	0.45	0.42	0.67	0.74
L3	*Vaccinium stamineum L.*	75	12.6	0.03	0.00	0.01	0.03	0.00	0.00	0.03	0.01	0.73	0.14	0.34	0.07	0.71	0.86	0.90
L4	*Dirca palustris L.*	65	21.5	0.23	0.07	0.11	0.27	0.14	0.19	0.37	0.07	0.85	0.21	0.49	0.39	0.80	0.83	0.88
L4	*Fraxinus caroliniana Miller*	100	16.4	0.54	0.23	0.29	0.46	0.33	0.20	0.59	0.42	0.89	0.86	0.71	0.35	0.86	0.82	0.85
L4	*Ilex opaca Aiton*	100	37.5	0.04	0.00	0.01	0.01	0.00	0.02	0.01	0.02	0.85	0.12	0.33	0.01	0.27	0.88	0.75
L4	*Populus L.*	3,065	30.9	0.71	0.44	0.64	0.60	0.73	0.73	0.76	0.38	0.88	0.66	0.81	0.60	0.87	0.89	0.93
L4	*Salix L.*	82	14.7	0.77	0.46	0.71	0.72	0.81	0.78	0.85	0.32	0.93	0.61	0.88	0.64	0.88	0.89	0.94
T-GSH · 9 species																		
L4	*Glycine max*	600	55.1	0.51	0.18	0.35	0.43	0.54	0.61	0.60	0.32	0.51	0.47	0.56	0.49	0.55	0.58	0.85
L4	*Phaseolus vulgaris*	597	32.0	0.61	0.47	0.36	0.45	0.52	0.58	0.55	0.17	0.55	0.60	0.53	0.51	0.55	0.59	0.63
L4	*Populus balsamifera*	113	37.4	0.66	0.23	0.07	0.02	0.45	0.48	0.48	0.00	0.08	0.08	0.29	0.09	0.74	0.90	0.96
L5	*Triticum aestivum*	2,443	17.9	0.17	0.21	0.02	0.02	0.27	0.40	0.20	0.45	0.76	0.78	0.70	0.70	0.79	0.87	0.85
L5	*Elaeis guineensis*	317	36.6	0.13	0.24	0.12	0.12	0.56	0.49	0.61	0.01	0.51	0.32	0.58	0.46	0.60	0.73	0.82
L5	*Hordeum vulgare*	542	27.9	0.04	0.01	0.02	0.11	0.25	0.14	0.28	0.00	0.19	0.07	0.42	0.38	0.29	0.65	0.81
L5	*Oryza glaberrima*	253	47.2	0.14	0.02	0.06	0.07	0.22	0.27	0.26	0.02	0.06	0.02	0.05	0.25	0.18	0.60	0.77
L5	*Triticum spp.*	651	78.4	0.17	0.05	0.14	0.22	0.29	0.24	0.18	0.18	0.10	0.06	0.25	0.22	0.54	0.31	0.56
L5	*Zea mays*	1,427	69.3	0.43	0.01	0.21	0.09	0.17	0.16	0.11	0.18	0.49	0.33	0.61	0.50	0.63	0.90	0.89

Figure S28: F1 as in Figure S26, for the species of T-MSWP and T-GSH.

True positives at $\tau = 0.25$

log scale: lighter is more

≤1 10 100 10^3 10^4 10^5

		N	GT	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
OD-Valid · 5 species																		
L0	*Actinidia arguta*	37	1,434	1329	1223	1167	1115	1137	1107	1111	1231	1292	1299	1270	1368	1394	1260	1373
L0	*Malus hupehensis*	41	3,276	2750	2673	2714	2645	2287	2206	2321	2790	2822	2806	2839	3003	3042	2896	2868
L0	*Malus spontanea*	40	3,069	2612	2543	2560	2512	2160	2038	2172	2600	2704	2656	2686	2897	2923	2753	2705
L0	*Malus toringoides*	9	671	517	483	490	471	391	366	401	491	516	514	517	613	625	598	524
L0	*Vitis vinifera*	11	342	328	315	299	301	314	313	315	298	312	307	302	325	324	296	318
T-MAWP · 17 species																		
L1	*Actinidia coriacea*	37	1,034	925	839	823	825	867	826	865	848	923	946	906	981	997	877	952
L1	*Malus domestica*	45	1,532	1256	1133	1195	1150	1125	1065	1193	1151	1262	1200	1236	1361	1414	1300	1351
L1	*Prunus Tai-Haku*	47	1,675	1653	1623	1619	1609	1601	1575	1574	1615	1656	1651	1650	1659	1665	1626	1652
L1	*Prunus maritima*	12	1,177	535	448	712	588	556	481	653	767	772	831	533	838	841	749	623
L1	*Prunus sargentii*	18	907	692	468	573	556	589	508	580	670	663	668	614	789	804	677	669
L1	*Prunus subhirtella*	20	1,749	1286	975	1322	963	1101	903	1252	1315	1577	1606	1565	1623	1696	1651	1589
L1	*Rosa Radtkopink*	12	222	216	205	198	205	214	213	216	211	217	217	214	217	220	208	221
L1	*Rosa centifolia*	10	285	236	194	205	211	229	219	224	231	260	264	252	263	273	249	274
L1	*Rosa chuckles*	29	913	874	845	832	830	862	856	843	847	882	880	873	892	903	859	901
L1	*Vitis empire*	9	231	214	203	196	195	200	197	201	192	209	206	201	224	225	205	214
L1	*Vitis riparia*	33	1,242	1198	1169	1126	1098	1103	1113	1139	1118	1165	1170	1129	1204	1211	1141	1197
L2	*Fagus crenata*	17	1,330	737	429	431	490	591	534	682	387	747	700	628	1045	1127	795	1011
L2	*Amelanchier sanguinea*	32	1,307	1170	1124	1133	1120	1003	949	1057	1049	1139	1061	1090	1260	1232	1114	1138
L3	*Rhododendron Roseum Superbum*	32	1,238	1154	914	889	1032	945	912	984	1006	1110	1022	1036	1216	1220	1156	1180
L4	*Populus tremula*	40	706	494	322	330	356	388	388	409	313	446	402	432	620	588	569	585
L4	*Populus tremuloides*	23	569	443	327	342	377	408	424	417	360	454	394	406	514	509	428	488
L4	*Populus trichocarpa*	40	1,047	905	733	538	734	869	870	879	706	870	744	773	997	955	838	950
T-DCWP · 39 species																		
L0	*Quercus rubra*	3	94	20	24	13	6	8	7	6	12	21	5	23	13	23	41	12
L1	*Viburnum acerifolium*	3	43	15	8	3	6	12	14	3	4	21	9	26	25	29	40	42
L1	*Viburnum dentatum*	3	54	35	18	5	11	28	25	6	4	20	23	43	50	50	53	53
L1	*Viburnum nudum*	3	26	3	0	0	3	0	0	1	0	11	0	1	11	22	26	26
L1	*Viburnum prunifolium*	3	48	40	32	14	30	24	23	21	2	47	14	24	46	42	43	47
L1	*Prunus lasiocarpa*	3	43	28	5	0	4	5	13	6	0	5	0	25	25	32	31	28
L1	*Stewartia malacodendron*	2	12	7	6	2	1	7	5	6	1	4	1	3	12	11	10	12
L2	*Camellia oleifera*	3	30	6	0	0	0	0	0	0	0	2	0	14	7	26	29	28
L2	*Amelanchier stolonifera*	3	31	3	8	0	1	13	8	1	1	4	1	24	29	28	29	31
L2	*Aronia arbutifolia*	3	93	80	57	23	28	71	72	36	14	83	64	89	92	90	92	93
L2	*Crataegus viridis*	2	39	21	1	0	0	5	12	8	0	3	0	28	30	31	39	39
L3	*Cornus alba*	3	50	17	23	3	8	20	25	13	2	27	39	45	49	41	46	50
L3	*Kalmia latifolia*	3	43	43	21	3	18	30	39	27	2	40	4	32	37	43	43	43
L3	*Sideroxylon lycioides*	3	88	84	81	69	75	85	83	73	59	84	53	77	85	85	87	88
L3	*Betula nigra*	3	36	8	6	9	9	13	13	13	4	7	7	6	16	13	25	33
L3	*Betula papyrifera*	3	30	4	1	0	0	5	2	4	0	6	0	12	18	22	27	30
L3	*Carpinus caroliniana*	3	118	68	68	56	72	59	71	28	54	94	70	82	106	87	118	103
L3	*Carya glabra*	3	86	76	66	16	42	53	62	38	18	84	53	67	74	84	84	85
L3	*Morella cerifera*	3	108	86	70	35	59	89	92	55	46	96	66	69	95	85	96	101
L3	*Lindera benzoin*	3	69	45	27	10	30	29	32	28	18	30	18	34	30	44	65	66
L3	*Persea borbonia*	3	153	86	81	44	36	92	94	62	34	122	118	104	110	137	147	153
L3	*Sassafras albidum*	3	118	34	36	4	26	48	54	14	1	56	33	38	63	82	111	110
L3	*Zanthoxylum americanum*	3	49	46	5	0	15	5	6	9	0	40	0	13	44	46	48	49
L3	*Zanthoxylum simulans*	3	22	0	0	0	0	0	0	0	0	0	0	0	6	15	22	22
L3	*Acer rubrum*	3	117	88	84	70	64	66	56	54	42	69	74	46	86	85	101	113
L3	*Aesculus parviflora*	3	151	94	96	59	84	88	86	57	85	107	90	119	122	117	139	141
L4	*Fatsia japonica*	3	48	45	15	3	11	36	38	40	0	38	7	16	39	48	47	48
L4	*Ilex cornuta*	3	92	92	43	9	47	82	87	79	1	87	48	69	83	91	88	85
L4	*Moringa oleifera*	3	52	47	48	47	47	48	47	46	24	40	41	42	50	51	44	52
L4	*Euonymus americanus*	3	63	61	48	36	56	36	37	31	31	62	37	59	58	61	61	63
L4	*Stephanotis floribunda*	5	195	175	120	131	111	159	154	137	86	162	113	143	176	182	176	183
L4	*Fraxinus pennsylvanica*	3	42	30	19	4	17	30	29	28	2	18	16	25	41	41	42	42
L4	*Olea europaea*	3	83	66	64	50	63	77	71	59	29	74	58	74	79	74	80	83
L4	*Magnolia virginiana*	3	148	138	134	39	98	95	131	90	90	147	86	60	126	143	148	148
L4	*Passiflora coriacea*	5	200	163	100	142	130	120	111	120	72	132	135	113	164	146	158	193
L4	*Hibiscus clayi*	3	35	35	34	33	34	33	35	34	7	35	32	35	35	35	34	35
L4	*Edgeworthia chrysantha*	3	37	2	1	1	4	2	3	8	0	13	0	4	12	17	29	32
L4	*Fothergilla major*	3	37	31	26	12	30	26	23	27	18	35	25	35	34	35	35	37
L4	*Itea virginica*	3	75	72	73	59	74	72	74	65	70	74	63	73	74	74	75	75

Figure S29: True positives at $\tau = 0.25$, IoU 0.50, for every detector and every species of OD-Valid, T-MAWP and T-DCWP. Each cell prints the value. N: images. GT: annotated stomata, so TP = GT − FN.

False positives at $\tau = 0.25$

log scale: lighter is fewer

≤1 10 100 10^3 10^4 10^5

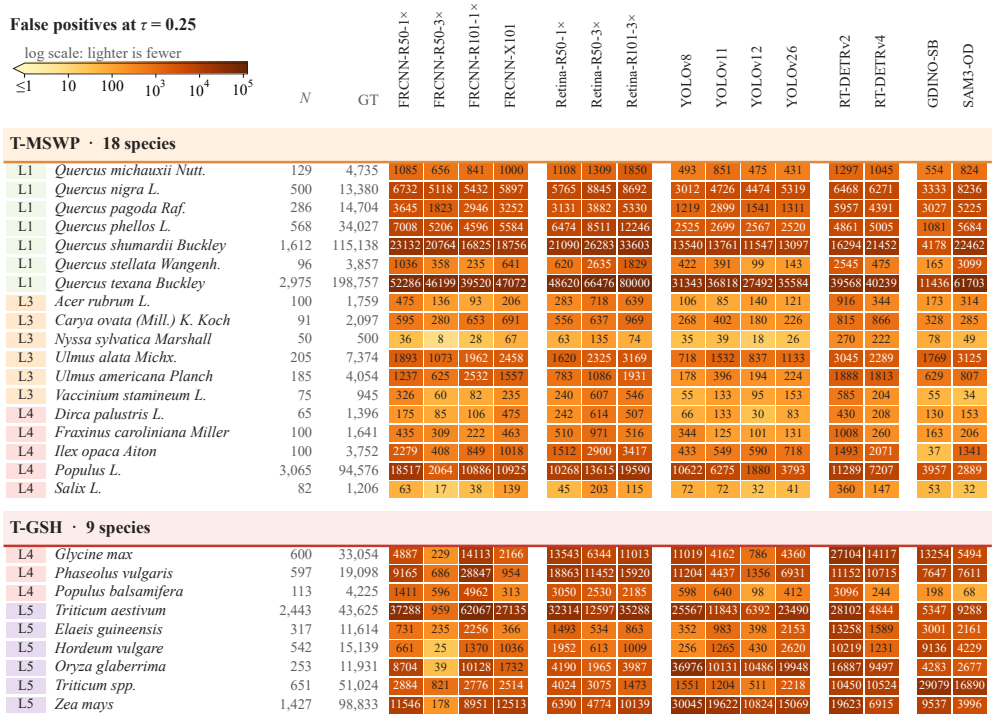
		N	GT	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
OD-Valid · 5 species																		
L0	*Actinidia arguta*	37	1,434	149	126	129	92	70	60	131	158	202	171	182	251	226	57	137
L0	*Malus hupehensis*	41	3,276	90	57	56	45	27	39	34	307	293	272	325	310	270	129	178
L0	*Malus spontanea*	40	3,069	218	158	163	140	86	55	70	275	343	304	340	680	712	330	292
L0	*Malus toringoides*	9	671	117	92	117	80	41	39	43	121	109	101	117	412	328	212	128
L0	*Vitis vinifera*	11	342	130	85	49	66	66	68	61	92	85	96	84	148	126	63	96
T-MAWP · 17 species																		
L1	*Actinidia coriacea*	37	1,034	397	225	195	146	206	189	262	194	255	271	208	471	553	171	278
L1	*Malus domestica*	45	1,532	378	321	320	245	212	205	306	287	340	265	321	893	914	522	620
L1	*Prunus Tai-Haku*	47	1,675	481	275	270	247	233	236	252	342	453	436	432	729	488	279	426
L1	*Prunus maritima*	12	1,177	492	336	430	389	285	302	354	613	709	932	464	1043	1293	1159	801
L1	*Prunus sargentii*	18	907	399	222	224	210	223	181	215	426	647	654	452	764	785	605	698
L1	*Prunus subhirtella*	20	1,749	257	97	189	125	120	97	187	241	485	484	444	813	1236	784	667
L1	*Rosa Radtkopink*	12	222	66	33	17	21	38	37	43	46	52	58	48	59	71	29	43
L1	*Rosa centifolia*	10	285	72	39	36	41	57	38	52	74	114	131	86	143	136	81	123
L1	*Rosa chuckles*	29	913	188	122	105	113	157	155	156	212	285	271	257	298	279	156	239
L1	*Vitis empire*	9	231	88	56	43	46	44	58	51	49	62	70	57	161	130	46	77
L1	*Vitis riparia*	33	1,242	504	332	249	247	250	288	267	290	364	346	327	550	522	294	382
L2	*Fagus crenata*	17	1,330	229	116	121	128	102	109	174	48	138	83	88	764	988	465	394
L2	*Amelanchier sanguinea*	32	1,307	512	406	356	336	250	233	267	288	352	234	297	1215	975	334	339
L3	*Rhododendron Roseum Superbum*	32	1,238	509	198	170	252	215	307	277	264	319	191	234	890	665	324	361
L4	*Populus tremula*	40	706	1336	740	997	445	684	721	707	256	476	388	489	2033	1670	1278	993
L4	*Populus tremuloides*	23	569	343	222	236	150	211	214	296	68	120	76	88	496	383	169	267
L4	*Populus trichocarpa*	40	1,047	462	169	125	197	372	356	469	138	201	114	140	865	398	213	278
T-DCWP · 39 species																		
L0	*Quercus rubra*	3	94	94	72	102	85	90	89	87	57	79	72	67	84	75	54	88
L1	*Viburnum acerifolium*	3	43	92	63	174	58	93	70	147	6	20	4	14	60	28	8	14
L1	*Viburnum dentatum*	3	54	319	51	387	75	161	150	241	20	23	2	10	90	11	13	13
L1	*Viburnum nudum*	3	26	45	20	104	18	23	26	71	2	4	1	1	38	7	2	4
L1	*Viburnum prunifolium*	3	48	117	14	54	8	56	40	40	8	12	2	2	19	7	5	28
L1	*Prunus lasiocarpa*	3	43	55	17	168	35	77	38	147	3	3	0	3	25	35	12	22
L1	*Stewartia malacodendron*	2	12	123	56	152	20	31	31	48	9	5	3	1	31	11	8	6
L2	*Camellia oleifera*	3	30	7	6	19	11	8	20	22	0	2	0	2	9	20	15	14
L2	*Amelanchier stolonifera*	3	31	407	181	344	111	240	214	207	19	18	4	8	239	23	6	5
L2	*Aronia arbutifolia*	3	93	181	70	330	72	55	56	111	10	25	5	10	20	9	5	15
L2	*Crataegus viridis*	2	39	141	60	196	46	153	118	98	4	20	0	0	48	1	3	1
L3	*Cornus alba*	3	50	480	264	434	251	404	240	379	349	71	16	60	290	24	8	8
L3	*Kalmia latifolia*	3	43	237	157	168	108	151	108	138	36	50	31	26	209	201	2	5
L3	*Sideroxylon lycioides*	3	88	47	15	73	8	32	22	28	8	13	1	6	12	15	3	7
L3	*Betula nigra*	3	36	254	88	145	156	172	149	178	86	136	52	106	440	533	486	19
L3	*Betula papyrifera*	3	30	176	83	217	58	146	80	132	2	16	0	2	90	114	0	0
L3	*Carpinus caroliniana*	3	118	76	46	73	50	80	68	120	25	39	23	23	38	65	11	31
L3	*Carya glabra*	3	86	33	9	66	8	33	41	94	10	11	4	8	26	26	5	12
L3	*Morella cerifera*	3	108	23	27	39	29	65	47	95	36	29	5	14	14	32	17	21
L3	*Lindera benzoin*	3	69	147	45	55	75	120	103	139	49	47	34	79	260	205	146	14
L3	*Persea borbonia*	3	153	203	67	270	139	132	85	183	44	57	49	28	224	105	14	30
L3	*Sassafras albidum*	3	118	159	50	208	45	124	107	200	46	45	24	15	120	33	26	18
L3	*Zanthoxylum americanum*	3	49	74	32	107	17	55	50	79	1	20	0	2	47	5	4	9
L3	*Zanthoxylum simulans*	3	22	20	22	54	5	18	11	36	1	5	2	1	22	6	0	1
L3	*Acer rubrum*	3	117	346	245	247	209	180	165	199	143	154	377	76	599	601	521	247
L3	*Aesculus parviflora*	3	151	249	208	268	181	111	97	203	117	121	100	53	279	277	134	11
L4	*Fatsia japonica*	3	48	102	26	290	47	42	48	91	3	5	1	1	11	9	2	8
L4	*Ilex cornuta*	3	92	39	40	135	17	57	32	49	3	14	3	7	13	2	7	22
L4	*Moringa oleifera*	3	52	23	41	43	9	17	46	36	8	12	4	7	16	39	1	12
L4	*Euonymus americanus*	3	63	10	10	14	1	31	35	40	9	23	6	8	9	8	3	20
L4	*Stephanotis floribunda*	5	195	141	32	244	85	57	104	127	157	78	18	74	133	162	12	11
L4	*Fraxinus pennsylvanica*	3	42	280	158	335	60	198	168	138	7	21	12	8	100	16	17	24
L4	*Olea europaea*	3	83	47	12	175	30	64	40	93	18	16	6	6	34	15	13	9
L4	*Magnolia virginiana*	3	148	23	19	60	12	49	63	101	8	27	8	15	18	38	8	15
L4	*Passiflora coriacea*	5	200	64	35	82	88	73	77	101	68	36	37	57	130	225	79	29
L4	*Hibiscus clayi*	3	35	20	33	3	11	8	7	5	0	4	1	5	4	0	2	3
L4	*Edgeworthia chrysantha*	3	37	111	32	100	28	103	73	113	4	11	3	10	187	30	118	135
L4	*Fothergilla major*	3	37	138	20	162	10	39	51	46	2	4	3	9	35	5	2	4
L4	*Itea virginica*	3	75	95	62	92	11	38	23	29	22	14	14	12	20	40	3	6

Figure S30: False positives at $\tau = 0.25$, IoU 0.50, for every detector and every species of OD-Valid, T-MAWP and T-DCWP. Each cell prints the value. N: images. GT: annotated stomata.



		N	GT	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
T-MSWP · 18 species																		
L1	*Quercus michauxii Nutt.*	129	4,735	2554	2531	2611	2504	2286	2690	1944	1925	3855	3186	3384	3535	3647	3827	3816
L1	*Quercus nigra L.*	500	13,380	1801	1311	1339	1479	1062	1239	1150	1964	7191	3683	3185	4339	4830	8007	3219
L1	*Quercus pagoda Raf.*	286	14,704	4660	4511	5684	4175	4346	5610	4353	2616	9148	6530	5036	8304	8949	11380	9125
L1	*Quercus phellos L.*	568	34,027	17424	14760	18651	17562	15048	16890	13523	18328	29919	23784	23271	22445	23854	28374	25115
L1	*Quercus shumardii Buckley*	1,612	115,138	64148	58325	66487	59054	50051	55961	46933	57892	96891	83364	73933	79127	79350	96966	76805
L1	*Quercus stellata Wangenh.*	96	3,857	146	22	30	62	8	26	97	53	3612	745	408	83	2317	3049	143
L1	*Quercus texana Buckley*	2,975	198,757	95482	90117	102276	88736	74681	84564	77481	87192	154441	131361	111275	122055	126584	162158	103565
L3	*Acer rubrum L.*	100	1,759	108	4	9	125	15	17	178	93	1008	462	218	292	1299	1242	1177
L3	*Carya ovata (Mill.) K. Koch*	91	2,097	564	484	593	488	615	760	609	464	1324	926	949	1193	1518	1563	1793
L3	*Nyssa sylvatica Marshall*	50	500	35	8	8	39	37	16	76	29	194	40	44	91	369	335	419
L3	*Ulmus alata Michx.*	205	7,374	1348	1387	1819	1431	1471	1930	1243	694	3228	2434	2137	2908	3344	4223	2711
L3	*Ulmus americana Planch*	185	4,054	395	462	558	340	596	845	771	130	1486	1075	923	1733	1542	2338	2854
L3	*Vaccinium stamineum L.*	75	945	22	1	3	17	1	2	22	3	616	77	223	54	634	748	809
L4	*Dirca palustris L.*	65	1,396	205	52	91	298	128	217	430	55	1129	164	486	437	1079	1081	1208
L4	*Fraxinus caroliniana Miller*	100	1,641	766	251	316	636	432	291	900	535	1420	1313	981	563	1420	1251	1353
L4	*Ilex opaca Aiton*	100	3,752	133	1	23	24	7	65	40	52	3175	284	879	28	900	2992	3047
L4	*Populus L.*	3,065	94,576	62871	27639	49398	44732	59641	62176	70221	24431	79052	47919	66779	45274	78772	79153	85337
L4	*Salix L.*	82	1,206	799	362	683	757	849	902	983	241	1105	546	973	733	1062	1002	1105
T-GSH · 9 species																		
L4	*Glycine max*	600	33,054	13022	3298	9861	9527	17336	17249	18669	8297	12667	10391	14669	19428	17827	19007	28539
L4	*Phaseolus vulgaris*	597	19,098	12272	6008	10493	5774	13359	12504	13233	2775	8990	8806	9359	10388	11185	11133	12251
L4	*Populus balsamifera*	113	4,225	2770	641	329	53	2132	2143	2034	4	192	175	781	354	2606	3635	3952
L5	*Triticum aestivum*	2,443	43,625	7597	5366	1059	688	12016	14251	8882	19891	33918	31796	36329	38385	31857	37685	38719
L5	*Elaeis guineensis*	317	11,614	845	1651	901	797	5084	3891	5510	85	4303	2307	5665	7488	5594	8488	9479
L5	*Hordeum vulgare*	542	15,139	334	93	170	964	2445	1199	2679	30	1673	581	4690	6036	2821	11614	13055
L5	*Oryza glaberrima*	253	11,931	1504	146	633	459	1972	2132	2383	533	681	280	857	4049	2127	6937	9142
L5	*Triticum spp.*	651	51,024	5153	1289	4125	6662	9443	7413	5050	5082	2693	1669	7693	7750	22648	14564	26515
L5	*Zea mays*	1,427	98,833	30245	606	12452	5016	9993	8946	6456	12937	38871	21564	49400	39719	49096	88483	83068

Figure S31: True positives as in Figure S29, for the species of T-MSWP and T-GSH.



		N	GT	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
T-MSWP · 18 species																		
L1	*Quercus michauxii Nutt.*	129	4,735	1085	656	841	1000	1108	1309	1850	493	851	475	431	1297	1045	554	824
L1	*Quercus nigra L.*	500	13,380	6732	5118	5432	5897	5765	8845	8692	3012	4726	4474	5319	6468	6271	3333	8236
L1	*Quercus pagoda Raf.*	286	14,704	3645	1823	2946	3252	3131	3882	5330	1219	2899	1541	1311	5957	4391	3027	5225
L1	*Quercus phellos L.*	568	34,027	7008	5206	4596	5584	6474	8511	12246	2525	2699	2567	2520	4861	5005	1081	5684
L1	*Quercus shumardii Buckley*	1,612	115,138	23132	20764	16825	18756	21090	26283	33603	13540	13761	11547	13097	16294	21452	4178	22462
L1	*Quercus stellata Wangenh.*	96	3,857	1036	358	235	641	620	2635	1829	422	391	99	143	2545	475	165	3099
L1	*Quercus texana Buckley*	2,975	198,757	52286	46199	39520	47072	48620	66476	80000	31343	36818	27492	35584	39568	40239	11436	61703
L3	*Acer rubrum L.*	100	1,759	475	136	93	206	283	718	639	106	85	140	121	916	344	173	314
L3	*Carya ovata (Mill.) K. Koch*	91	2,097	595	280	653	691	556	637	969	268	402	180	226	815	866	328	285
L3	*Nyssa sylvatica Marshall*	50	500	36	8	28	67	63	135	74	35	39	18	26	270	222	78	49
L3	*Ulmus alata Michx.*	205	7,374	1893	1073	1962	2458	1620	2325	3169	718	1532	837	1133	3045	2289	1769	3125
L3	*Ulmus americana Planch*	185	4,054	1237	625	2532	1557	783	1086	1931	178	396	194	224	1888	1813	629	807
L3	*Vaccinium stamineum L.*	75	945	326	60	82	235	240	607	546	55	133	95	153	585	204	55	34
L4	*Dirca palustris L.*	65	1,396	175	85	106	475	242	614	507	66	133	30	83	430	208	130	153
L4	*Fraxinus caroliniana Miller*	100	1,641	435	309	222	463	510	971	516	344	125	101	131	1008	260	163	206
L4	*Ilex opaca Aiton*	100	3,752	2279	408	849	1018	1512	2900	3417	433	549	590	718	1493	2071	37	1341
L4	*Populus L.*	3,065	94,576	18517	2064	10886	10925	10268	13615	19590	10622	6275	1880	3793	11289	7207	3957	2889
L4	*Salix L.*	82	1,206	63	17	38	139	45	203	115	72	72	32	41	360	147	53	32
T-GSH · 9 species																		
L4	*Glycine max*	600	33,054	4887	229	14113	2166	13543	6344	11013	11019	4162	786	4360	27104	14117	13254	5494
L4	*Phaseolus vulgaris*	597	19,098	9165	686	28847	954	18863	11452	15920	11204	4437	1356	6931	11152	10715	7647	7611
L4	*Populus balsamifera*	113	4,225	1411	596	4962	313	3050	2530	2185	598	640	98	412	3096	244	198	68
L5	*Triticum aestivum*	2,443	43,625	37288	959	62067	27135	32314	12597	35288	25567	11843	6392	23490	28102	4844	5347	9288
L5	*Elaeis guineensis*	317	11,614	731	235	2256	366	1493	534	863	352	983	398	2153	13258	1589	3001	2161
L5	*Hordeum vulgare*	542	15,139	661	25	1370	1036	1952	613	1009	256	1265	430	2620	10219	1231	9136	4229
L5	*Oryza glaberrima*	253	11,931	8704	39	10128	1732	4190	1965	3987	36976	10131	10486	19948	16887	9497	4283	2677
L5	*Triticum spp.*	651	51,024	2884	821	2776	2514	4024	3075	1473	1551	1204	511	2218	10450	10524	29079	16890
L5	*Zea mays*	1,427	98,833	11546	178	8951	12513	6390	4774	10139	30045	19622	10824	15069	19623	6915	9537	3996

Figure S32: False positives as in Figure S30, for the species of T-MSWP and T-GSH.

False negatives at $\tau = 0.25$

log scale: lighter is fewer

≤1 10 100 10^3 10^4 10^5

		N	GT	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
OD-Valid · 5 species																		
L0	*Actinidia arguta*	37	1,434	105	211	267	319	297	327	323	203	142	135	164	66	40	174	61
L0	*Malus hupehensis*	41	3,276	526	603	562	631	989	1070	955	486	454	470	437	273	234	380	408
L0	*Malus spontanea*	40	3,069	457	526	509	557	909	1031	897	469	365	413	383	172	146	316	364
L0	*Malus toringoides*	9	671	154	188	181	200	280	305	270	180	155	157	154	58	46	73	147
L0	*Vitis vinifera*	11	342	14	27	43	41	28	29	27	44	30	35	40	17	18	46	24
T-MAWP · 17 species																		
L1	*Actinidia coriacea*	37	1,034	109	195	211	209	167	208	169	186	111	88	128	53	37	157	82
L1	*Malus domestica*	45	1,532	276	399	337	382	407	467	339	381	270	332	296	171	118	232	181
L1	*Prunus Tai-Haku*	47	1,675	22	52	56	66	74	100	101	60	19	24	25	16	10	49	23
L1	*Prunus maritima*	12	1,177	642	729	465	589	621	696	524	410	405	346	644	339	336	428	554
L1	*Prunus sargentii*	18	907	215	439	334	351	318	399	327	237	244	239	293	118	103	230	238
L1	*Prunus subhirtella*	20	1,749	463	774	427	786	648	846	497	434	172	143	184	126	53	98	160
L1	*Rosa Radtkopink*	12	222	6	17	24	17	8	9	6	11	5	5	8	5	2	14	1
L1	*Rosa centifolia*	10	285	49	91	80	74	56	66	61	54	25	21	33	22	12	36	11
L1	*Rosa chuckles*	29	913	39	68	81	83	51	57	70	66	31	33	40	21	10	54	12
L1	*Vitis empire*	9	231	17	28	35	36	31	34	30	39	22	25	30	7	6	26	17
L1	*Vitis riparia*	33	1,242	44	73	116	144	139	129	103	124	77	72	113	38	31	101	45
L2	*Fagus crenata*	17	1,330	593	901	899	840	739	796	648	943	583	630	702	285	203	535	319
L2	*Amelanchier sanguinea*	32	1,307	137	183	174	187	304	358	250	258	168	246	217	47	75	193	169
L3	*Rhododendron Roseum Superbum*	32	1,238	84	324	349	206	293	326	254	232	128	216	202	22	18	82	58
L4	*Populus tremula*	40	706	212	384	376	350	318	318	297	393	260	304	274	86	118	137	121
L4	*Populus tremuloides*	23	569	126	242	227	192	161	145	152	209	115	175	163	55	60	141	81
L4	*Populus trichocarpa*	40	1,047	142	314	509	313	178	177	168	341	177	303	274	50	92	209	97
T-DCWP · 39 species																		
L0	*Quercus rubra*	3	94	74	70	81	88	86	87	88	82	73	89	71	81	71	53	82
L1	*Viburnum acerifolium*	3	43	28	35	40	37	31	29	40	39	22	34	17	18	14	3	1
L1	*Viburnum dentatum*	3	54	19	36	49	43	26	29	48	50	34	31	11	4	4	1	1
L1	*Viburnum nudum*	3	26	23	26	26	23	26	26	25	26	15	26	25	15	4	0	0
L1	*Viburnum prunifolium*	3	48	8	16	34	18	24	25	27	46	1	34	24	2	6	5	1
L1	*Prunus lasiocarpa*	3	43	15	38	43	39	38	30	37	43	38	43	18	18	11	12	15
L1	*Stewartia malacodendron*	2	12	5	6	10	11	5	7	6	11	8	11	9	0	1	2	0
L2	*Camellia oleifera*	3	30	24	30	30	30	30	30	30	30	28	30	16	23	4	1	2
L2	*Amelanchier stolonifera*	3	31	28	23	31	30	18	23	30	30	27	30	7	2	3	2	0
L2	*Aronia arbutifolia*	3	93	13	36	70	65	22	21	57	79	10	29	4	1	3	1	0
L2	*Crataegus viridis*	2	39	18	38	39	39	34	27	31	39	36	39	11	9	8	0	0
L3	*Cornus alba*	3	50	33	27	47	42	30	25	37	48	23	11	5	1	9	4	0
L3	*Kalmia latifolia*	3	43	0	22	40	25	13	4	16	41	3	39	11	6	0	0	0
L3	*Sideroxylon lycioides*	3	88	4	7	19	13	3	5	15	29	4	35	11	3	3	1	0
L3	*Betula nigra*	3	36	28	30	27	27	23	23	23	32	29	29	30	20	23	11	3
L3	*Betula papyrifera*	3	30	26	29	30	30	25	28	26	30	24	30	18	12	8	3	0
L3	*Carpinus caroliniana*	3	118	50	50	62	46	59	47	90	64	24	48	36	12	31	0	15
L3	*Carya glabra*	3	86	10	20	70	44	33	24	48	68	2	33	19	12	2	2	1
L3	*Morella cerifera*	3	108	22	38	73	49	19	16	53	62	12	42	39	13	23	12	7
L3	*Lindera benzoin*	3	69	24	42	59	39	40	37	41	51	39	51	35	39	25	4	3
L3	*Persea borbonia*	3	153	67	72	109	117	61	59	91	119	31	35	49	43	16	6	0
L3	*Sassafras albidum*	3	118	84	82	114	92	70	64	104	117	62	85	80	55	36	7	8
L3	*Zanthoxylum americanum*	3	49	3	44	49	34	44	43	40	49	9	49	36	5	3	1	0
L3	*Zanthoxylum simulans*	3	22	22	22	22	22	22	22	22	22	22	22	22	16	7	0	0
L3	*Acer rubrum*	3	117	29	33	47	53	51	61	63	75	48	43	71	31	32	16	4
L3	*Aesculus parviflora*	3	151	57	55	92	67	63	65	94	66	44	61	32	29	34	12	10
L4	*Fatsia japonica*	3	48	3	33	45	37	12	10	8	48	10	41	32	9	0	1	0
L4	*Ilex cornuta*	3	92	0	49	83	45	10	5	13	91	5	44	23	9	1	4	7
L4	*Moringa oleifera*	3	52	5	4	5	5	4	5	6	28	12	11	10	2	1	8	0
L4	*Euonymus americanus*	3	63	2	15	27	7	27	26	32	32	1	26	4	5	2	2	0
L4	*Stephanotis floribunda*	5	195	20	75	64	84	36	41	58	109	33	82	52	19	13	19	12
L4	*Fraxinus pennsylvanica*	3	42	12	23	38	25	12	13	14	40	24	26	17	1	1	0	0
L4	*Olea europaea*	3	83	17	19	33	20	6	12	24	54	9	25	9	4	9	3	0
L4	*Magnolia virginiana*	3	148	10	14	109	50	53	17	58	58	1	62	88	22	5	0	0
L4	*Passiflora coriacea*	5	200	37	100	58	70	80	89	80	128	68	65	87	36	54	42	7
L4	*Hibiscus clayi*	3	35	0	1	2	1	2	0	1	28	0	3	0	0	0	1	0
L4	*Edgeworthia chrysantha*	3	37	35	36	36	33	35	34	29	37	24	37	33	25	20	8	5
L4	*Fothergilla major*	3	37	6	11	25	7	11	14	10	19	2	12	2	3	2	2	0
L4	*Itea virginica*	3	75	3	2	16	1	3	1	10	5	1	12	2	1	1	0	0

Figure S33: False negatives at $\tau = 0.25$, IoU 0.50, for every detector and every species of OD-Valid, T-MAWP and T-DCWP. Each cell prints the value. N: images. GT: annotated stomata, so TP = GT − FN.

Count MAE at $\tau = 0.25$

log scale: lighter is lower error

≤1 — 10 — 100

		N	$\bar{c}_{GT}$	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
OD-Valid · 5 species																		
L0	*Actinidia arguta*	37	38.8	2.5	3.7	4.1	6.2	6.5	7.5	5.8	2.8	3.0	2.8	2.9	5.7	5.4	3.9	2.8
L0	*Malus hupehensis*	41	79.9	10.7	13.3	12.3	14.3	23.5	25.1	22.5	7.2	7.1	7.7	6.7	5.6	4.5	6.9	6.5
L0	*Malus spontanea*	40	76.7	6.4	9.3	8.8	10.4	20.6	24.4	20.7	6.2	4.4	4.8	4.8	13.4	14.2	6.8	4.8
L0	*Malus toringoides*	9	74.6	11.9	13.3	14.9	14.2	26.6	29.6	25.2	9.2	9.6	10.9	10.1	39.3	31.3	16.1	9.9
L0	*Vitis vinifera*	11	31.1	10.5	5.6	3.3	4.1	3.8	4.1	3.5	4.9	5.2	5.5	5.5	11.9	9.8	3.4	6.5
T-MAWP · 17 species																		
L1	*Actinidia coriacea*	37	27.9	8.5	3.7	3.7	4.4	3.9	4.7	4.4	4.6	5.4	5.4	4.1	11.3	13.9	3.8	6.0
L1	*Malus domestica*	45	34.0	6.4	6.3	5.9	5.4	6.8	7.8	6.0	6.0	5.8	4.8	5.3	16.4	17.7	8.4	10.2
L1	*Prunus Tai-Haku*	47	35.6	9.8	4.8	4.7	4.0	4.1	3.8	4.4	6.3	9.2	8.8	8.7	15.2	10.2	4.9	8.6
L1	*Prunus maritima*	12	98.1	18.7	32.8	13.1	19.7	28.5	32.8	17.2	23.9	32.3	48.8	29.0	58.7	79.8	60.9	20.6
L1	*Prunus sargentii*	18	50.4	10.9	13.1	9.3	9.2	8.3	13.2	9.0	12.5	22.4	23.1	10.1	35.9	37.9	20.8	25.6
L1	*Prunus subhirtella*	20	87.5	13.1	33.9	12.1	33.0	26.4	37.5	17.6	12.2	15.8	17.1	14.1	34.4	59.1	34.5	26.4
L1	*Rosa Radtkopink*	12	18.5	5.0	2.8	2.1	2.0	3.0	2.7	3.1	3.4	3.9	4.4	3.3	4.5	5.8	1.4	3.5
L1	*Rosa centifolia*	10	28.5	4.1	5.6	5.0	3.5	3.9	3.2	3.5	4.6	9.3	11.0	6.9	12.1	12.4	5.5	11.2
L1	*Rosa chuckles*	29	31.5	5.1	2.3	2.4	2.8	3.9	3.6	4.1	5.6	8.8	8.2	7.5	9.6	9.3	4.3	7.8
L1	*Vitis empire*	9	25.7	8.1	4.0	4.2	4.7	4.3	6.4	5.0	4.7	6.2	8.1	6.3	17.1	13.8	3.3	7.6
L1	*Vitis riparia*	33	37.6	14.1	8.6	5.4	5.2	5.5	6.0	5.9	6.2	8.9	8.6	6.9	15.5	14.9	6.5	10.5
L2	*Fagus crenata*	17	78.2	21.4	46.2	45.8	41.9	37.5	40.4	27.9	52.6	26.2	32.2	36.1	28.2	46.2	7.8	8.5
L2	*Amelanchier sanguinea*	32	40.8	12.7	8.8	7.9	8.0	6.7	7.0	6.1	6.0	6.9	5.8	5.8	36.5	28.1	6.2	7.7
L3	*Rhododendron Roseum Superbum*	32	38.7	13.3	4.8	6.8	3.8	5.4	5.5	5.5	3.5	6.5	4.0	4.8	27.1	20.2	7.8	9.5
L4	*Populus tremula*	40	17.6	28.1	15.1	16.2	7.3	10.2	10.9	11.9	6.6	8.6	7.6	8.4	48.7	38.8	28.9	21.8
L4	*Populus tremuloides*	23	24.7	10.0	5.0	6.0	4.4	5.6	6.1	7.5	6.8	2.9	5.3	4.8	19.2	14.0	5.0	8.2
L4	*Populus trichocarpa*	40	26.2	8.3	5.0	9.8	4.7	6.0	5.4	7.9	6.4	3.0	5.4	4.5	20.4	7.8	3.6	4.9
T-DCWP · 39 species																		
L0	*Quercus rubra*	3	31.3	6.7	0.7	8.3	2.3	3.3	1.3	0.3	8.3	2.7	5.7	2.0	1.7	2.0	1.0	2.0
L1	*Viburnum acerifolium*	3	14.3	21.3	9.3	44.7	7.0	20.7	13.7	35.7	11.0	2.0	10.0	1.0	14.0	4.7	1.7	4.3
L1	*Viburnum dentatum*	3	18.0	100.0	5.0	112.7	12.7	45.0	40.3	64.3	10.0	3.7	9.7	1.7	28.7	2.3	4.0	4.0
L1	*Viburnum nudum*	3	8.7	7.3	7.3	26.0	2.3	2.3	3.3	15.3	8.0	3.7	8.3	8.0	9.0	1.7	0.7	1.3
L1	*Viburnum prunifolium*	3	16.0	36.3	1.3	6.7	3.3	10.7	5.0	4.3	12.7	3.7	10.7	7.3	5.7	1.0	0.7	9.0
L1	*Prunus lasiocarpa*	3	14.3	13.3	7.0	41.7	6.0	13.0	2.7	36.7	13.3	11.7	14.3	5.0	5.0	8.0	0.0	2.3
L1	*Stewartia malacodendron*	2	6.0	59.0	25.0	71.0	5.5	13.0	12.0	21.0	3.0	1.5	4.0	4.0	15.5	5.0	3.0	3.0
L2	*Camellia oleifera*	3	10.0	5.7	8.0	3.7	6.3	7.3	3.3	2.7	10.0	8.7	10.0	4.7	5.3	5.3	4.7	4.0
L2	*Amelanchier stolonifera*	3	10.3	126.3	52.7	104.3	27.0	74.0	63.7	59.0	4.3	3.0	8.7	1.0	79.0	6.7	2.0	1.7
L2	*Aronia arbutifolia*	3	31.0	56.0	11.3	86.7	5.0	11.0	11.7	18.0	23.0	5.0	8.0	2.7	6.3	2.0	1.3	5.0
L2	*Crataegus viridis*	2	19.5	61.5	11.0	78.5	7.5	59.5	45.5	33.5	17.5	8.0	19.5	5.5	19.5	3.5	1.5	0.5
L3	*Cornus alba*	3	16.7	149.0	79.0	129.0	69.7	124.7	71.7	114.0	100.3	16.0	3.7	18.3	96.3	5.0	2.7	2.7
L3	*Kalmia latifolia*	3	14.3	79.0	45.0	42.7	27.7	46.0	34.7	40.7	1.7	15.7	6.7	5.0	67.7	67.0	0.7	1.7
L3	*Sideroxylon lycioides*	3	29.3	14.3	3.3	18.0	1.7	9.7	5.7	4.3	7.0	3.0	11.3	1.7	3.0	4.0	0.7	2.3
L3	*Betula nigra*	3	12.0	75.3	19.3	39.3	43.0	49.7	42.0	51.7	18.0	35.7	7.7	25.3	140.0	170.0	158.3	5.3
L3	*Betula papyrifera*	3	10.0	50.0	20.7	62.3	9.3	40.3	17.3	35.3	9.3	3.3	10.0	5.3	26.0	35.3	1.0	0.0
L3	*Carpinus caroliniana*	3	39.3	8.7	2.0	3.7	1.3	7.0	7.0	10.0	13.0	7.0	8.3	4.3	8.7	11.3	3.7	5.3
L3	*Carya glabra*	3	28.7	7.7	3.7	4.7	12.0	4.7	5.7	15.3	19.3	3.0	9.7	3.7	4.7	8.0	1.0	3.7
L3	*Morella cerifera*	3	36.0	1.7	3.7	11.3	6.7	15.3	10.3	14.0	8.7	5.7	12.3	8.3	1.0	3.0	4.3	4.7
L3	*Lindera benzoin*	3	23.0	41.0	7.0	2.0	12.0	26.7	22.0	32.7	4.0	2.7	7.7	14.7	73.7	60.0	47.3	3.7
L3	*Persea borbonia*	3	51.0	45.3	16.3	53.7	13.3	23.7	8.7	30.7	25.0	8.7	5.3	9.7	60.3	29.7	4.0	10.0
L3	*Sassafras albidum*	3	39.3	25.0	10.7	31.3	15.7	18.0	14.3	32.0	23.7	5.7	20.3	21.7	21.7	1.7	6.3	3.3
L3	*Zanthoxylum americanum*	3	16.3	23.7	4.7	19.3	5.7	5.7	3.0	13.0	16.0	3.7	16.3	11.3	14.0	1.3	1.0	3.0
L3	*Zanthoxylum simulans*	3	7.3	3.3	8.0	10.7	5.7	1.3	3.7	4.7	7.0	5.7	6.7	7.0	2.7	2.3	0.0	0.3
L3	*Acer rubrum*	3	39.0	105.7	70.7	66.7	52.0	43.0	34.7	45.3	22.7	35.3	111.3	20.3	189.3	189.7	168.3	81.0
L3	*Aesculus parviflora*	3	50.3	64.0	51.0	58.7	38.0	16.0	10.7	36.3	17.0	25.7	13.0	9.7	83.3	81.0	40.7	3.7
L4	*Fatsia japonica*	3	16.0	33.0	5.0	81.7	5.3	10.0	12.7	27.7	15.0	1.7	13.3	10.3	1.3	3.0	0.3	2.7
L4	*Ilex cornuta*	3	30.7	13.0	3.0	17.3	9.3	15.7	9.0	12.0	29.3	3.0	13.7	5.3	3.3	0.3	1.0	5.0
L4	*Moringa oleifera*	3	17.3	6.0	12.3	12.7	1.3	4.3	13.7	10.0	6.7	0.7	3.0	1.0	4.7	12.7	2.3	4.0
L4	*Euonymus americanus*	3	21.0	2.7	1.7	4.3	2.0	1.3	3.0	2.7	7.7	7.3	6.7	1.3	2.0	2.0	1.0	6.7
L4	*Stephanotis floribunda*	5	39.0	24.2	8.6	36.0	4.2	5.4	12.6	14.2	29.2	9.0	12.8	14.4	22.8	29.8	1.4	0.6
L4	*Fraxinus pennsylvanica*	3	14.0	89.3	45.0	99.0	11.7	62.0	51.7	41.3	11.0	1.7	4.7	3.0	33.0	5.0	5.7	8.0
L4	*Olea europaea*	3	27.7	10.0	2.3	47.3	3.3	19.3	9.3	23.0	12.0	3.0	6.3	1.7	10.0	2.0	3.3	3.0
L4	*Magnolia virginiana*	3	49.3	4.3	1.7	16.3	12.7	6.7	15.3	14.3	16.7	8.7	18.0	24.3	1.3	11.0	2.7	5.0
L4	*Passiflora coriacea*	5	40.0	5.4	13.0	4.8	3.6	4.2	7.2	4.2	12.0	7.6	6.8	6.8	18.8	34.6	7.4	4.4
L4	*Hibiscus clayi*	3	11.7	6.7	10.7	0.3	3.3	2.0	2.3	1.3	9.3	1.3	0.7	1.7	1.3	0.0	0.3	1.0
L4	*Edgeworthia chrysantha*	3	12.3	25.3	4.0	21.3	3.7	22.7	13.0	28.0	11.0	4.3	11.3	7.7	54.0	4.7	36.7	43.3
L4	*Fothergilla major*	3	12.3	44.0	5.7	45.7	1.7	9.3	12.3	12.0	5.7	1.3	3.7	2.3	10.7	1.7	1.3	1.3
L4	*Itea virginica*	3	25.0	30.7	20.0	25.3	3.3	11.7	7.3	6.3	5.7	4.3	2.0	3.3	6.3	13.0	1.0	2.0

Figure S34: Count MAE (stomata per image) at $\tau = 0.25$, for every detector and every species of OD-Valid, T-MAWP and T-DCWP. Each cell prints the value. N: images. $\bar{c}_{GT}$: mean annotated stomata per image.

False negatives at $\tau = 0.25$

log scale: lighter is fewer

≤1 · 10 · 100 · 10^3 · 10^4 · 10^5

		N	GT	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
T-MSWP · 18 species																		
L1	*Quercus michauxii Nutt.*	129	4,735	2181	2204	2124	2231	2449	2045	2791	2810	880	1549	1351	1200	1088	908	919
L1	*Quercus nigra L.*	500	13,380	11579	12069	12041	11901	12318	12141	12230	11416	6189	9697	10195	9041	8550	5373	10161
L1	*Quercus pagoda Raf.*	286	14,704	10044	10193	9020	10529	10358	9094	10351	12088	5556	8174	9668	6400	5755	3324	5579
L1	*Quercus phellos L.*	568	34,027	16603	19267	15376	16465	18979	17137	20504	15699	4108	10243	10756	11582	10173	5653	8912
L1	*Quercus shumardii Buckley*	1,612	115,138	50990	56813	48651	56084	65087	59177	68205	57246	18247	31774	41205	36011	35788	18172	38333
L1	*Quercus stellata Wangenh.*	96	3,857	3711	3835	3827	3795	3849	3831	3760	3804	245	3112	3449	3774	1540	808	3714
L1	*Quercus texana Buckley*	2,975	198,757	103275	108640	96481	110021	124076	114193	121276	111565	44316	67396	87482	76702	72173	36599	95192
L3	*Acer rubrum L.*	100	1,759	1651	1755	1750	1634	1744	1742	1581	1666	751	1297	1541	1467	460	517	582
L3	*Carya ovata (Mill.) K. Koch*	91	2,097	1533	1613	1504	1609	1482	1337	1488	1633	773	1171	1148	904	579	534	304
L3	*Nyssa sylvatica Marshall*	50	500	465	492	492	461	463	484	424	471	306	460	456	409	131	165	81
L3	*Ulmus alata Michx.*	205	7,374	6026	5987	5555	5943	5903	5444	6131	6680	4146	4940	5237	4466	4030	3151	4663
L3	*Ulmus americana Planch*	185	4,054	3659	3592	3496	3714	3458	3209	3283	3924	2568	2979	3131	2321	2512	1716	1200
L3	*Vaccinium stamineum L.*	75	945	923	944	942	928	944	943	923	942	329	868	722	891	311	197	136
L4	*Dirca palustris L.*	65	1,396	1191	1344	1305	1098	1268	1179	966	1341	267	1232	910	959	317	315	188
L4	*Fraxinus caroliniana Miller*	100	1,641	875	1390	1325	1005	1209	1350	741	1106	221	328	660	1078	221	390	288
L4	*Ilex opaca Aiton*	100	3,752	3619	3751	3729	3728	3745	3687	3712	3700	577	3468	2873	3724	2852	760	705
L4	*Populus L.*	3,065	94,576	31705	66937	45178	49844	34935	32400	24355	70145	15524	46657	27797	49302	15804	15423	9239
L4	*Salix L.*	82	1,206	407	844	523	449	357	304	223	965	101	660	233	473	144	204	101
T-GSH · 9 species																		
L4	*Glycine max*	600	33,054	20032	29756	23193	23527	15718	15805	14385	24757	20387	22663	18385	13626	15227	14047	4515
L4	*Phaseolus vulgaris*	597	19,098	6826	13090	8605	13324	5739	6594	5865	16323	10108	10292	9739	8710	7913	7965	6847
L4	*Populus balsamifera*	113	4,225	1455	3584	3896	4172	2093	2082	2191	4221	4033	4050	3444	3871	1619	590	273
L5	*Triticum aestivum*	2,443	43,625	36028	38259	42566	42937	31609	29374	34743	23734	9707	11829	7296	5240	11768	5940	4906
L5	*Elaeis guineensis*	317	11,614	10769	9963	10713	10817	6530	7723	6104	11529	7311	9307	5949	4126	6020	3126	2135
L5	*Hordeum vulgare*	542	15,139	14805	15046	14969	14175	12694	13940	12460	15109	13466	14558	10449	9103	12318	3525	2084
L5	*Oryza glaberrima*	253	11,931	10427	11785	11298	11472	9959	9799	9548	11398	11250	11651	11074	7882	9804	4994	2789
L5	*Triticum spp.*	651	51,024	45871	49735	46899	44362	41581	43611	45974	45942	48331	49355	43331	43274	28376	36460	24509
L5	*Zea mays*	1,427	98,833	68588	98227	86381	93817	88840	89887	92377	85896	59962	77269	49433	59114	49737	10350	15765

Figure S35: False negatives as in Figure S33, for the species of T-MSWP and T-GSH.

Count MAE at $\tau = 0.25$

log scale: lighter is lower error

≤1 · 10 · 100

		N	$\bar{c}_{GT}$	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
T-MSWP · 18 species																		
L1	*Quercus michauxii Nutt.*	129	36.7	9.2	12.2	10.7	10.3	11.3	8.1	9.2	18.0	3.8	8.6	7.5	7.1	5.5	4.7	5.0
L1	*Quercus nigra L.*	500	26.8	10.1	13.9	14.4	12.1	13.2	7.3	7.5	16.9	4.0	10.5	9.8	5.8	5.0	4.5	4.7
L1	*Quercus pagoda Raf.*	286	51.4	23.6	29.9	22.0	25.7	25.9	18.9	18.7	38.0	11.3	23.6	29.2	9.8	12.5	8.0	8.0
L1	*Quercus phellos L.*	568	59.9	16.9	24.8	19.0	19.2	22.0	15.2	14.8	23.2	3.5	13.5	14.7	11.9	9.2	8.1	8.0
L1	*Quercus shumardii Buckley*	1,612	71.4	17.4	22.4	20.0	23.3	27.3	20.5	21.5	27.2	5.4	12.6	17.5	12.6	10.2	9.4	10.4
L1	*Quercus stellata Wangenh.*	96	40.2	27.9	36.2	37.4	32.9	33.6	12.6	20.1	35.2	2.5	31.4	34.4	12.8	11.1	6.7	6.6
L1	*Quercus texana Buckley*	2,975	66.8	17.4	21.0	19.6	21.7	25.4	16.2	15.0	27.1	5.2	13.5	17.5	12.7	10.9	9.3	11.5
L3	*Acer rubrum L.*	100	17.6	11.8	16.2	16.6	14.3	14.6	10.2	9.4	15.6	6.7	11.6	14.2	5.9	2.2	3.6	2.9
L3	*Carya ovata (Mill.) K. Koch*	91	23.0	12.6	15.0	11.9	11.0	11.3	9.0	8.3	15.0	6.1	11.0	10.4	5.6	6.1	4.2	2.8
L3	*Nyssa sylvatica Marshall*	50	10.0	8.6	9.7	9.3	7.9	8.0	7.0	7.0	8.7	5.3	8.8	8.6	4.1	2.7	2.0	1.1
L3	*Ulmus alata Michx.*	205	36.0	20.3	24.0	17.8	17.1	20.9	15.3	14.7	29.1	12.9	20.0	20.0	7.9	9.5	7.9	7.8
L3	*Ulmus americana Planch*	185	21.9	13.4	16.0	11.5	12.8	14.5	11.9	8.6	20.2	11.8	15.1	15.7	5.5	8.4	6.2	3.3
L3	*Vaccinium stamineum L.*	75	12.6	8.0	11.8	11.5	9.2	9.4	4.5	5.1	11.8	2.7	10.3	7.6	4.3	2.0	2.1	1.6
L4	*Dirca palustris L.*	65	21.5	15.6	19.4	18.4	9.7	15.8	8.9	7.2	19.6	2.6	18.5	12.7	8.4	2.6	3.4	2.4
L4	*Fraxinus caroliniana Miller*	100	16.4	4.4	10.8	11.0	5.8	7.0	3.8	2.5	7.6	1.3	2.4	5.3	3.0	1.9	2.5	1.4
L4	*Ilex opaca Aiton*	100	37.5	13.4	33.4	28.8	27.1	22.3	7.9	4.4	32.7	2.1	28.8	21.6	22.3	7.8	7.2	7.5
L4	*Populus L.*	3,065	30.9	7.3	21.2	11.7	12.9	8.4	6.8	5.2	19.6	4.0	14.7	8.1	12.9	4.1	4.1	2.7
L4	*Salix L.*	82	14.7	4.4	10.1	5.9	3.9	3.9	2.4	2.0	10.9	1.1	7.7	2.7	3.7	1.6	2.1	1.1
T-GSH · 9 species																		
L4	*Glycine max*	600	55.1	26.8	49.2	35.0	35.7	20.9	17.9	16.5	25.5	27.3	36.5	24.3	31.6	18.9	11.4	5.2
L4	*Phaseolus vulgaris*	597	32.0	12.2	20.8	38.3	20.8	26.9	16.2	22.7	23.0	11.2	15.2	13.5	14.9	14.2	7.3	2.7
L4	*Populus balsamifera*	113	37.4	5.9	26.4	13.6	34.2	11.1	9.1	6.9	33.0	30.0	35.0	26.9	14.8	12.2	4.5	2.0
L5	*Triticum aestivum*	2,443	17.9	10.0	15.3	20.1	10.3	6.8	7.6	7.2	9.2	4.8	5.0	8.0	10.2	4.2	2.6	3.1
L5	*Elaeis guineensis*	317	36.6	32.1	30.7	30.1	33.0	16.3	22.7	16.6	35.3	20.1	28.1	12.9	29.4	14.2	3.5	2.3
L5	*Hordeum vulgare*	542	27.9	26.1	27.7	25.8	24.2	20.1	24.6	21.5	27.4	22.7	26.1	15.1	17.7	20.9	11.2	4.3
L5	*Oryza glaberrima*	253	47.2	45.6	46.4	46.1	39.2	34.9	33.8	35.2	128.4	40.8	51.2	71.3	55.4	48.8	14.8	6.3
L5	*Triticum spp.*	651	78.4	66.4	75.1	68.1	64.3	57.7	62.3	68.4	68.2	72.4	75.0	63.6	52.0	30.8	19.7	13.2
L5	*Zea mays*	1,427	69.3	40.1	68.7	55.1	57.2	57.8	59.6	58.2	39.3	28.4	46.8	25.4	37.4	30.2	5.8	9.7

Figure S36: Count MAE as in Figure S34, for the species of T-MSWP and T-GSH.

Count R^2 at $\tau = 0.25$

teal above the species mean (0), brown below

≤−1 — 0 — 1

		N	$\bar{c}_{GT}$	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
OD-Valid · 5 species																		
L0	*Actinidia arguta*	37	38.8	0.87	0.78	0.72	0.38	0.39	0.19	0.47	0.85	0.84	0.87	0.85	0.44	0.47	0.74	0.84
L0	*Malus hupehensis*	41	79.9	0.32	0.05	0.19	−0.06	−1.83	−2.33	−1.66	0.59	0.64	0.61	0.67	0.79	0.84	0.69	0.72
L0	*Malus spontanea*	40	76.7	0.55	0.20	0.31	0.04	−2.05	−3.34	−2.21	0.54	0.72	0.68	0.70	−0.43	−0.65	0.54	0.64
L0	*Malus toringoides*	9	74.6	0.51	0.42	0.23	0.32	−1.18	−1.67	−1.08	0.77	0.72	0.65	0.68	−3.20	−2.05	0.15	0.69
L0	*Vitis vinifera*	11	31.1	−3.20	−0.36	0.53	0.20	0.25	0.18	0.40	−0.12	−0.23	−0.56	−0.26	−4.34	−2.82	0.41	−0.81
T-MAWP · 17 species																		
L1	*Actinidia coriacea*	37	27.9	−0.59	0.60	0.62	0.49	0.57	0.43	0.43	0.40	0.25	0.23	0.51	−1.74	−3.08	0.61	0.08
L1	*Malus domestica*	45	34.0	0.71	0.62	0.74	0.76	0.64	0.54	0.75	0.70	0.80	0.83	0.81	−0.52	−0.67	0.40	0.36
L1	*Prunus Tai-Haku*	47	35.6	−0.19	0.67	0.69	0.77	0.75	0.78	0.68	0.43	−0.09	0.08	0.03	−1.85	−0.28	0.65	0.08
L1	*Prunus maritima*	12	98.1	−2.66	−5.89	−0.27	−2.20	−3.48	−4.74	−0.65	−3.41	−5.75	<−10	−5.77	<−10	<−10	<−10	−1.86
L1	*Prunus sargentii*	18	50.4	0.31	0.00	0.39	0.37	0.59	−0.02	0.46	0.23	−1.34	−1.38	0.38	−4.69	−5.43	−1.20	−1.91
L1	*Prunus subhirtella*	20	87.5	0.56	−0.94	0.59	−1.01	−0.39	−1.61	0.20	0.61	0.50	0.36	0.56	−0.98	−4.68	−1.02	−0.24
L1	*Rosa Radtkopink*	12	18.5	0.06	0.69	0.84	0.79	0.52	0.62	0.55	0.41	0.30	0.22	0.46	0.10	−0.36	0.85	0.44
L1	*Rosa centifolia*	10	28.5	0.06	−1.03	−0.83	−0.04	−0.07	−0.04	0.21	−0.18	−3.09	−4.65	−1.13	−5.28	−5.46	−0.87	−4.38
L1	*Rosa chuckles*	29	31.5	0.61	0.91	0.90	0.88	0.78	0.80	0.75	0.56	0.01	0.14	0.29	−0.29	−0.19	0.62	0.17
L1	*Vitis empire*	9	25.7	−0.06	0.71	0.73	0.67	0.69	0.36	0.65	0.56	0.42	0.18	0.38	−3.39	−1.64	0.82	0.09
L1	*Vitis riparia*	33	37.6	−2.54	−0.43	0.36	0.37	0.35	0.17	0.25	0.17	−0.52	−0.42	−0.05	−3.25	−2.86	0.08	−0.99
L2	*Fagus crenata*	17	78.2	−1.84	<−10	<−10	−8.50	−6.69	−7.88	−3.58	<−10	−3.18	−4.96	−6.32	−3.41	<−10	0.44	0.41
L2	*Amelanchier sanguinea*	32	40.8	−2.22	−0.61	−0.34	−0.23	0.12	−0.08	0.18	0.03	−0.09	0.26	0.31	<−10	<−10	0.10	−0.50
L3	*Rhododendron Roseum Superbum*	32	38.7	−1.17	0.51	0.32	0.76	0.41	0.44	0.49	0.69	0.33	0.69	0.64	−7.56	−3.98	−0.04	−0.22
L4	*Populus tremula*	40	17.6	−9.77	−4.01	−4.90	−0.06	−1.05	−1.11	−1.61	0.22	−0.52	0.05	−0.29	<−10	<−10	<−10	−5.82
L4	*Populus tremuloides*	23	24.7	−2.12	−0.11	−0.44	0.26	−0.11	−0.40	−1.05	−0.76	0.60	−0.14	0.13	−9.30	−5.47	−0.29	−1.27
L4	*Populus trichocarpa*	40	26.2	−0.71	0.22	−1.76	0.33	−0.13	0.01	−0.97	−0.29	0.68	0.06	0.33	−9.15	−0.54	0.62	0.26
T-DCWP · 39 species																		
L0	*Quercus rubra*	3	31.3	<−10	0.81	<−10	−0.97	−2.19	0.44	0.91	<−10	−1.44	−9.59	−0.31	−0.03	−0.31	0.72	−0.69
L1	*Viburnum acerifolium*	3	14.3	<−10	−9.07	<−10	−3.05	<−10	<−10	<−10	−7.12	0.71	−5.29	0.90	<−10	−0.44	0.49	−0.58
L1	*Viburnum dentatum*	3	18.0	<−10	−7.07	<−10	<−10	<−10	<−10	<−10	<−10	−2.50	<−10	0.21	<−10	−0.50	−2.43	−2.43
L1	*Viburnum nudum*	3	8.7	<−10	<−10	<−10	−1.42	−2.35	−3.85	<−10	<−10	−4.19	<−10	<−10	<−10	−0.04	0.77	−0.85
L1	*Viburnum prunifolium*	3	16.0	<−10	−4.00	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	−0.50	0.00	<−10
L1	*Prunus lasiocarpa*	3	14.3	<−10	−8.82	<−10	−4.52	<−10	−0.26	<−10	<−10	<−10	<−10	−3.50	−2.73	<−10	1.00	−0.02
L1	*Stewartia malacodendron*	2	6.0															
L2	*Camellia oleifera*	3	10.0	−4.39	<−10	−2.61	−6.39	−9.00	−1.00	−0.67	<−10	<−10	<−10	−2.89	−5.33	−5.56	−3.22	−2.11
L2	*Amelanchier stolonifera*	3	10.3	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	−0.12	<−10	<−10	−4.25	−5.37
L2	*Aronia arbutifolia*	3	31.0	<−10	<−10	<−10	−1.78	<−10	<−10	<−10	<−10	−1.91	−5.81	0.19	−3.03	0.44	0.69	−1.41
L2	*Crataegus viridis*	2	19.5	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	−4.56	−0.11	0.78
L3	*Cornus alba*	3	16.7	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	−5.43	−9.71
L3	*Kalmia latifolia*	3	14.3	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	−2.00	<−10
L3	*Sideroxylon lycioides*	3	29.3	<−10	−6.62	<−10	−0.50	<−10	<−10	−6.04	<−10	−3.73	<−10	−0.96	−2.81	−5.46	0.54	−2.35
L3	*Betula nigra*	3	12.0	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	−6.57
L3	*Betula papyrifera*	3	10.0	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	−6.00	<−10	<−10	<−10	<−10	−0.50	1.00
L3	*Carpinus caroliniana*	3	39.3	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10
L3	*Carya glabra*	3	28.7	<−10	−2.24	−4.53	<−10	−5.16	<−10	<−10	<−10	−1.61	<−10	−2.24	−6.74	<−10	0.61	−4.13
L3	*Morella cerifera*	3	36.0	0.79	−0.74	−8.86	−2.19	<−10	−7.69	<−10	−4.67	−2.07	<−10	−8.02	0.88	0.17	−0.36	−1.00
L3	*Lindera benzoin*	3	23.0	<−10	<−10	−8.00	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10
L3	*Persea borbonia*	3	51.0	<−10	<−10	<−10	<−10	<−10	−7.89	<−10	<−10	<−10	−3.58	<−10	<−10	<−10	−0.89	<−10
L3	*Sassafras albidum*	3	39.3	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	−2.32	<−10	<−10	<−10	0.51	−2.55	0.02
L3	*Zanthoxylum americanum*	3	16.3	<−10	<−10	<−10	<−10	<−10	−7.79	<−10	<−10	−8.21	<−10	<−10	<−10	−0.29	0.36	−4.79
L3	*Zanthoxylum simulans*	3	7.3	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	1.00	−0.50
L3	*Acer rubrum*	3	39.0	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10
L3	*Aesculus parviflora*	3	50.3	<−10	<−10	<−10	<−10	−6.12	−2.18	<−10	−8.65	<−10	−4.24	−2.14	<−10	<−10	<−10	0.53
L4	*Fatsia japonica*	3	16.0	<−10	−6.21	<−10	−6.57	<−10	<−10	<−10	<−10	−0.21	<−10	<−10	0.57	−2.79	0.93	−0.71
L4	*Ilex cornuta*	3	30.7	<−10	−2.35	<−10	<−10	<−10	<−10	<−10	<−10	−2.81	<−10	<−10	−4.31	0.88	0.42	−8.35
L4	*Moringa oleifera*	3	17.3	<−10	<−10	<−10	−1.14	<−10	<−10	<−10	<−10	0.57	−7.79	−0.93	<−10	<−10	−3.07	<−10
L4	*Euonymus americanus*	3	21.0	<−10	−5.50	<−10	−9.00	−4.00	<−10	<−10	<−10	<−10	<−10	−3.00	−6.00	−9.00	−0.50	<−10
L4	*Stephanotis floribunda*	5	39.0	−0.05	0.87	−1.70	0.97	0.90	0.61	0.47	−0.91	0.70	0.71	0.45	−0.23	−2.40	0.99	1.00
L4	*Fraxinus pennsylvanica*	3	14.0	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	−0.38	−7.25	−2.62	<−10	−8.62	<−10	<−10
L4	*Olea europaea*	3	27.7	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10
L4	*Magnolia virginiana*	3	49.3	−0.50	0.62	<−10	<−10	−2.22	<−10	<−10	<−10	−4.15	<−10	<−10	0.78	−8.07	0.10	−0.86
L4	*Passiflora coriacea*	5	40.0	0.92	0.48	0.94	0.96	0.97	0.90	0.97	0.74	0.79	0.84	0.87	0.04	−3.13	0.87	0.96
L4	*Hibiscus clayi*	3	11.7	<−10	<−10	0.79	−7.14	−2.00	−3.50	−0.29	<−10	−0.29	0.14	−1.36	−1.14	1.00	0.79	−0.07
L4	*Edgeworthia chrysantha*	3	12.3	<−10	−9.71	<−10	−9.50	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10	<−10
L4	*Fothergilla major*	3	12.3	<−10	−5.78	<−10	0.34	<−10	<−10	<−10	−6.02	0.64	−3.14	−0.62	<−10	0.46	0.64	0.40
L4	*Itea virginica*	3	25.0	<−10	<−10	<−10	−1.71	<−10	<−10	−7.93	−7.21	−4.50	−0.29	−2.86	−7.93	<−10	0.36	−0.86

Figure S37: Count R^2 at $\tau = 0.25$, for every detector and every species of OD-Valid, T-MAWP and T-DCWP. Each cell prints the value. N: images. $\bar{c}_{GT}$: mean annotated stomata per image. Values below -10 print as < -10. The images of *Stewartia malacodendron* have the same annotated count, so its R^2 is undefined (–).

Density MAE at $\tau = 0.25$

log scale: lighter is lower error

≤1 · 10 · 100 · 10^3

		N	$\bar{\rho}_{GT}$	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
OD-Valid · 5 calibrated species																		
L0	*Actinidia arguta*	37	82.5	5.3	7.9	8.6	13.2	13.9	15.9	12.3	6.0	6.4	6.0	6.2	12.1	11.5	8.2	5.9
L0	*Malus hupehensis*	41	170.0	22.7	28.3	26.3	30.4	49.9	53.5	47.8	15.3	15.2	16.3	14.3	11.9	9.5	14.6	13.9
L0	*Malus spontanea*	40	163.3	13.7	19.9	18.7	22.2	43.8	51.9	44.0	13.3	9.4	10.3	10.3	28.6	30.1	14.5	10.1
L0	*Malus toringoides*	9	158.6	25.3	28.4	31.7	30.3	56.5	62.9	53.7	19.6	20.3	23.2	21.5	83.7	66.7	34.3	21.0
L0	*Vitis vinifera*	11	66.2	22.4	12.0	7.0	8.7	8.1	8.7	7.4	10.4	11.0	11.8	11.6	25.3	20.9	7.2	13.9
T-MAWP · 17 calibrated species																		
L1	*Actinidia coriacea*	37	59.5	18.1	7.9	7.8	9.4	8.2	9.9	9.3	9.9	11.5	11.4	8.7	24.0	29.7	8.1	12.8
L1	*Malus domestica*	45	72.4	13.5	13.4	12.6	11.5	14.4	16.5	12.7	12.8	12.4	10.2	11.3	35.0	37.6	17.8	21.7
L1	*Prunus Tai-Haku*	47	75.8	20.8	10.3	10.0	8.6	8.6	8.1	9.3	13.5	19.6	18.7	18.4	32.3	21.6	10.5	18.2
L1	*Prunus maritima*	12	208.7	39.7	69.7	27.8	41.8	60.6	69.9	36.5	50.9	68.8	103.9	61.7	124.8	169.7	129.6	43.8
L1	*Prunus sargentii*	18	107.2	23.2	27.8	19.9	19.5	17.6	28.1	19.1	26.6	47.6	49.1	21.4	76.4	80.6	44.3	54.4
L1	*Prunus subhirtella*	20	186.1	27.9	72.0	25.7	70.3	56.2	79.7	37.4	26.1	33.5	36.5	30.0	73.1	125.9	73.4	56.3
L1	*Rosa Radtkopink*	12	39.4	10.6	6.0	4.4	4.3	6.4	5.7	6.6	7.3	8.3	9.4	7.1	9.6	12.2	3.0	7.4
L1	*Rosa centifolia*	10	60.6	8.7	11.9	10.6	7.4	8.3	6.8	7.4	9.8	19.8	23.4	14.7	25.7	26.4	11.7	23.8
L1	*Rosa chuckles*	29	67.0	10.9	4.8	5.1	5.9	8.4	7.6	8.7	11.9	18.8	17.5	15.9	20.3	19.7	9.2	16.7
L1	*Vitis empire*	9	54.6	17.3	8.5	9.0	9.9	9.2	13.7	10.6	9.9	13.2	17.3	13.5	36.4	29.3	7.1	16.1
L1	*Vitis riparia*	33	80.1	29.9	18.2	11.5	11.2	11.7	12.8	12.5	13.2	18.9	18.3	14.7	33.0	31.7	13.7	22.2
L2	*Fagus crenata*	17	166.5	45.6	98.3	97.4	89.1	79.7	86.0	59.3	112.0	55.7	68.5	76.8	60.0	98.3	16.5	18.1
L2	*Amelanchier sanguinea*	32	86.9	26.9	18.8	16.9	17.0	14.2	14.8	13.0	12.8	14.6	12.2	12.2	77.7	59.8	13.1	16.4
L3	*Rhododendron Roseum Superbum*	32	82.3	28.3	10.1	14.4	8.0	11.6	11.8	11.6	7.4	13.9	8.6	10.1	57.7	43.0	16.6	20.3
L4	*Populus tremula*	40	37.6	59.8	32.1	34.4	15.5	21.7	23.1	25.3	14.1	18.3	16.2	17.8	103.6	82.6	61.4	46.4
L4	*Populus tremuloides*	23	52.6	21.2	10.5	12.7	9.4	11.8	13.0	15.9	14.5	6.2	11.4	10.3	40.8	29.9	10.5	17.4
L4	*Populus trichocarpa*	40	55.7	17.8	10.7	20.7	9.9	12.9	11.5	16.9	13.6	6.5	11.4	9.6	43.5	16.6	7.7	10.4
T-DCWP · 39 calibrated species																		
L0	*Quercus rubra*	3	337.8	71.9	7.2	89.9	25.2	35.9	14.4	3.6	89.9	28.8	61.1	21.6	18.0	21.6	10.8	21.6
L1	*Viburnum acerifolium*	3	154.5	230.0	100.6	481.6	75.5	222.8	147.4	384.6	118.6	21.6	107.8	10.8	150.9	50.3	18.0	46.7
L1	*Viburnum dentatum*	3	194.1	1078.2	53.9	1214.8	136.6	485.2	434.9	693.7	107.8	39.5	104.2	18.0	309.1	25.2	43.1	43.1
L1	*Viburnum nudum*	3	93.4	79.1	79.1	280.3	25.2	25.2	35.9	165.3	86.3	39.5	89.9	86.3	97.0	18.0	7.2	14.4
L1	*Viburnum prunifolium*	3	172.5	391.8	14.4	71.9	35.9	115.0	53.9	46.7	136.6	39.5	115.0	79.1	61.1	10.8	7.2	97.0
L1	*Prunus lasiocarpa*	3	154.5	143.8	75.5	449.3	64.7	140.2	28.8	395.3	143.8	125.8	154.5	53.9	53.9	86.3	0.0	25.2
L1	*Stewartia malacodendron*	2	64.7	636.1	269.6	765.5	59.3	140.2	129.4	226.4	32.3	16.2	43.1	43.1	167.1	53.9	32.3	32.3
L2	*Camellia oleifera*	3	107.8	61.1	86.3	39.5	68.3	79.1	35.9	28.8	107.8	93.4	107.8	50.3	57.5	57.5	50.3	43.1
L2	*Amelanchier stolonifera*	3	111.4	1362.1	567.9	1124.9	291.1	797.9	686.5	636.1	46.7	32.3	93.4	10.8	851.8	71.9	21.6	18.0
L2	*Aronia arbutifolia*	3	334.2	603.8	122.2	934.5	53.9	118.6	125.8	194.1	248.0	53.9	86.3	28.8	68.3	21.6	14.4	53.9
L2	*Crataegus viridis*	2	210.3	663.1	118.6	846.4	80.9	641.5	490.6	361.2	188.7	86.3	210.3	59.3	210.3	37.7	16.2	5.4
L3	*Cornus alba*	3	179.7	1606.5	851.8	1390.9	751.2	1344.2	772.7	1229.2	1081.8	172.5	39.5	197.7	1038.7	53.9	28.8	28.8
L3	*Kalmia latifolia*	3	154.5	851.8	485.2	460.0	298.3	496.0	373.8	438.5	18.0	168.9	71.9	53.9	729.6	722.4	7.2	18.0
L3	*Sideroxylon lycioides*	3	316.3	154.5	35.9	194.1	18.0	104.2	61.1	46.7	75.5	32.3	122.2	18.0	32.3	43.1	7.2	25.2
L3	*Betula nigra*	3	129.4	812.3	208.5	424.1	463.6	535.5	452.8	557.1	194.1	384.6	82.7	273.1	1509.5	1833.0	1707.2	57.5
L3	*Betula papyrifera*	3	107.8	539.1	222.8	672.1	100.6	434.9	186.9	381.0	100.6	35.9	107.8	57.5	280.3	381.0	10.8	0.0
L3	*Carpinus caroliniana*	3	424.1	93.4	21.6	39.5	14.4	75.5	75.5	107.8	140.2	75.5	89.9	46.7	93.4	122.2	39.5	57.5
L3	*Carya glabra*	3	309.1	82.7	39.5	50.3	129.4	50.3	61.1	165.3	208.5	32.3	104.2	39.5	50.3	86.3	10.8	39.5
L3	*Morella cerifera*	3	388.2	18.0	39.5	122.2	71.9	165.3	111.4	150.9	93.4	61.1	133.0	89.9	10.8	32.3	46.7	50.3
L3	*Lindera benzoin*	3	248.0	442.1	75.5	21.6	129.4	287.5	237.2	352.2	43.1	28.8	82.7	158.1	794.3	646.9	510.4	39.5
L3	*Persea borbonia*	3	549.9	488.8	176.1	578.6	143.8	255.2	93.4	330.7	269.6	93.4	57.5	104.2	650.5	319.9	43.1	107.8
L3	*Sassafras albidum*	3	424.1	269.6	115.0	337.8	168.9	194.1	154.5	345.0	255.2	61.1	219.2	233.6	233.6	18.0	68.3	35.9
L3	*Zanthoxylum americanum*	3	176.1	255.2	50.3	208.5	61.1	61.1	32.3	140.2	172.5	39.5	176.1	122.2	150.9	14.4	10.8	32.3
L3	*Zanthoxylum simulans*	3	79.1	35.9	86.3	115.0	61.1	14.4	39.5	50.3	75.5	61.1	71.9	75.5	28.8	25.2	0.0	3.6
L3	*Acer rubrum*	3	420.5	1139.3	761.9	718.8	560.7	463.6	373.8	488.8	244.4	381.0	1200.4	219.2	2041.4	2045.0	1815.0	873.4
L3	*Aesculus parviflora*	3	542.7	690.1	549.9	632.6	409.7	172.5	115.0	391.8	183.3	276.7	140.2	104.2	898.5	873.4	438.5	39.5
L4	*Fatsia japonica*	3	172.5	355.8	53.9	880.5	57.5	107.8	136.6	298.3	161.7	18.0	143.8	111.4	14.4	32.3	3.6	28.8
L4	*Ilex cornuta*	3	330.7	140.2	32.3	186.9	100.6	168.9	97.0	129.4	316.3	32.3	147.4	57.5	35.9	3.6	10.8	53.9
L4	*Moringa oleifera*	3	186.9	64.7	133.0	136.6	14.4	46.7	147.4	107.8	71.9	7.2	32.3	10.8	50.3	136.6	25.2	43.1
L4	*Euonymus americanus*	3	226.4	28.8	18.0	46.7	21.6	14.4	32.3	28.8	82.7	79.1	71.9	14.4	21.6	21.6	10.8	71.9
L4	*Stephanotis floribunda*	5	200.5	196.1	64.1	335.8	35.9	53.8	127.1	149.4	159.2	40.7	106.5	79.9	125.5	98.9	7.1	3.3
L4	*Fraxinus pennsylvanica*	3	150.9	963.2	485.2	1067.4	125.8	668.5	557.1	445.7	118.6	18.0	50.3	32.3	355.8	53.9	61.1	86.3
L4	*Olea europaea*	3	298.3	107.8	25.2	510.4	35.9	208.5	100.6	248.0	129.4	32.3	68.3	18.0	107.8	21.6	35.9	32.3
L4	*Magnolia virginiana*	3	531.9	46.7	18.0	176.1	136.6	71.9	165.3	154.5	179.7	93.4	194.1	262.4	14.4	118.6	28.8	53.9
L4	*Passiflora coriacea*	5	198.3	21.2	43.5	32.6	21.2	22.8	39.1	35.9	70.1	25.5	23.4	25.0	70.6	97.2	34.8	23.4
L4	*Hibiscus clayi*	3	125.8	71.9	115.0	3.6	35.9	21.6	25.2	14.4	100.6	14.4	7.2	18.0	14.4	0.0	3.6	10.8
L4	*Edgeworthia chrysantha*	3	133.0	273.1	43.1	230.0	39.5	244.4	140.2	301.9	118.6	46.7	122.2	82.7	582.2	50.3	395.3	467.2
L4	*Fothergilla major*	3	133.0	474.4	61.1	492.4	18.0	100.6	133.0	129.4	61.1	14.4	39.5	25.2	115.0	18.0	14.4	14.4
L4	*Itea virginica*	3	269.6	330.7	215.6	273.1	35.9	125.8	79.1	68.3	61.1	46.7	21.6	35.9	68.3	140.2	10.8	21.6

Figure S38: Density MAE (stomata mm^{-2}) at $\tau = 0.25$, calibrated images only, for every detector and every species of OD-Valid, T-MAWP and T-DCWP. Each cell prints the value. N: calibrated images. $\bar{\rho}_{GT}$: mean annotated density (stomata mm^{-2}). Species without calibrated images are left out.

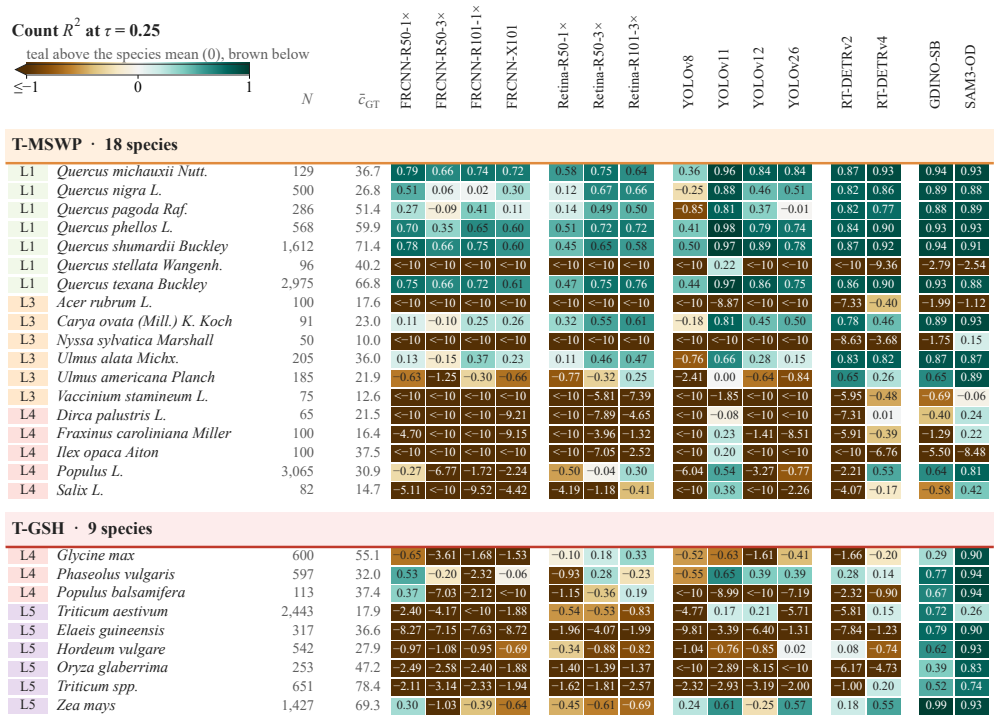


Figure S39: Count R^2 as in Figure S37, for the species of T-MSWP and T-GSH.

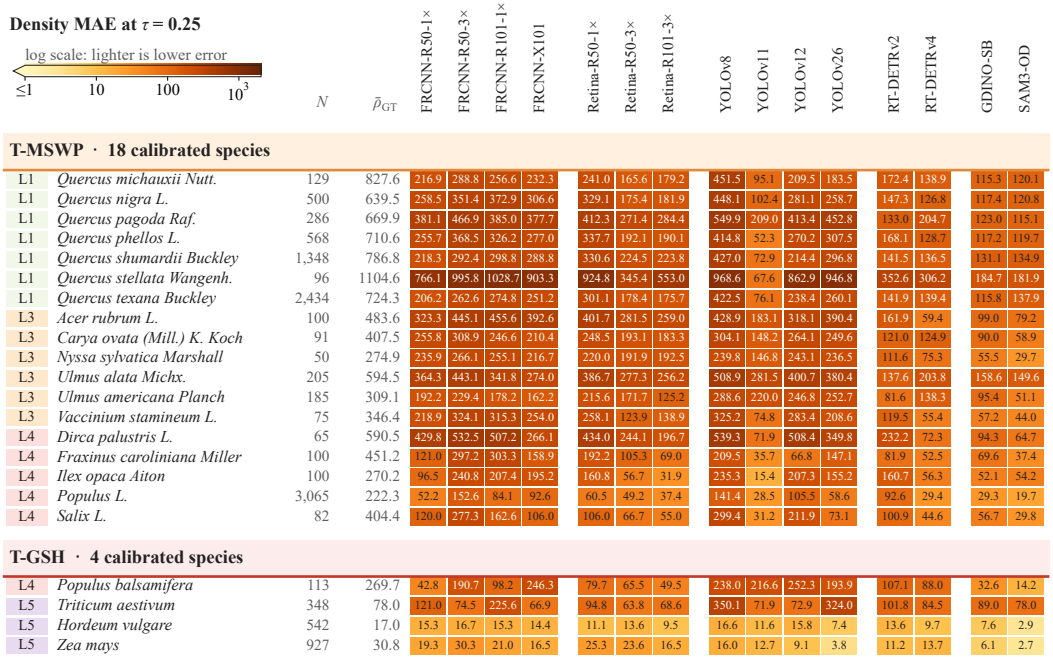


Figure S40: Density MAE as in Figure S38, for the calibrated species of T-MSWP and T-GSH.

Signed count bias at $\tau = 0.25$

orange overcounts, purple undercounts (symmetric log)

−100 0 100

		N	$\bar{c}_{GT}$	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
OD-Valid · 5 species																		
L0	*Actinidia arguta*	37	38.8	1.2	−2.3	−3.7	−6.1	−6.1	−7.2	−5.2	−1.2	1.6	1.0	0.5	5.0	5.0	−3.2	2.1
L0	*Malus hupehensis*	41	79.9	−10.6	−13.3	−12.3	−14.3	−23.5	−25.1	−22.5	−4.4	−3.9	−4.8	−2.7	0.9	0.9	−6.1	−5.6
L0	*Malus spontanea*	40	76.7	−6.0	−9.2	−8.7	−10.4	−20.6	−24.4	−20.7	−4.8	−0.6	−2.7	−1.1	12.7	14.2	0.3	−1.8
L0	*Malus toringoides*	9	74.6	−4.1	−10.7	−7.1	−13.3	−26.6	−29.6	−25.2	−6.6	−5.1	−6.2	−4.1	39.3	31.3	15.4	−2.1
L0	*Vitis vinifera*	11	31.1	10.5	5.3	0.5	2.3	3.5	3.5	3.1	4.4	5.0	5.5	4.0	11.9	9.8	1.5	6.5
T-MAWP · 17 species																		
L1	*Actinidia coriacea*	37	27.9	7.8	0.8	−0.4	−1.7	1.1	−0.5	2.5	0.2	3.9	4.9	2.2	11.3	13.9	0.4	5.3
L1	*Malus domestica*	45	34.0	2.3	−1.7	−0.4	−3.0	−4.3	−5.8	−0.7	−2.1	1.6	−1.5	0.6	16.0	17.7	6.4	9.8
L1	*Prunus Tai-Haku*	47	35.6	9.8	4.7	4.6	3.9	3.4	2.9	3.2	6.0	9.2	8.8	8.7	15.2	10.2	4.9	8.6
L1	*Prunus maritima*	12	98.1	−12.5	−32.8	−2.9	−16.7	−28.0	−32.8	−14.2	16.9	25.3	48.8	−15.0	58.7	79.8	60.9	20.6
L1	*Prunus sargentii*	18	50.4	10.2	−12.1	−6.1	−7.8	−5.3	−12.1	−6.2	10.5	22.4	23.1	8.8	35.9	37.9	20.8	25.6
L1	*Prunus subhirtella*	20	87.5	−10.3	−33.9	−11.9	−33.0	−26.4	−37.5	−15.5	−9.7	15.7	17.1	13.0	34.4	59.1	34.3	25.4
L1	*Rosa Radtkopink*	12	18.5	5.0	1.3	−0.6	0.3	2.5	2.3	3.1	2.9	3.9	4.4	3.3	4.5	5.8	1.2	3.5
L1	*Rosa centifolia*	10	28.5	2.3	−5.2	−4.4	−3.3	0.1	−2.8	−0.9	2.0	8.9	11.0	5.3	12.1	12.4	4.5	11.2
L1	*Rosa chuckles*	29	31.5	5.1	1.9	0.8	1.0	3.7	3.4	3.0	5.0	8.8	8.2	7.5	9.6	9.3	3.5	7.8
L1	*Vitis empire*	9	25.7	7.9	3.1	0.9	1.1	1.4	2.7	2.3	1.1	4.4	5.0	3.0	17.1	13.8	2.2	6.7
L1	*Vitis riparia*	33	37.6	13.9	7.8	4.0	3.1	3.4	4.8	5.0	5.0	8.7	8.3	6.5	15.5	14.9	5.8	10.2
L2	*Fagus crenata*	17	78.2	−21.4	−46.2	−45.8	−41.9	−37.5	−40.4	−27.9	−52.6	−26.2	−32.2	−36.1	28.2	46.2	−4.1	4.4
L2	*Amelanchier sanguinea*	32	40.8	11.7	7.0	5.7	4.7	−1.7	−3.9	0.5	0.9	5.8	−0.4	2.5	36.5	28.1	4.4	5.3
L3	*Rhododendron Roseum Superbum*	32	38.7	13.3	−3.9	−5.6	1.4	−2.4	−0.6	0.7	1.0	6.0	−0.8	1.0	27.1	20.2	7.6	9.5
L4	*Populus tremula*	40	17.6	28.1	8.9	15.5	2.4	9.2	10.1	10.2	−3.4	5.4	2.1	5.4	48.7	38.8	28.5	21.8
L4	*Populus tremuloides*	23	24.7	9.4	−0.9	0.4	−1.8	2.2	3.0	6.3	−6.1	0.2	−4.3	−3.3	19.2	14.0	1.2	8.1
L4	*Populus trichocarpa*	40	26.2	8.0	−3.6	−9.6	−2.9	4.8	4.5	7.5	−5.1	0.6	−4.7	−3.4	20.4	7.7	0.1	4.5
T-DCWP · 39 species																		
L0	*Quercus rubra*	3	31.3	6.7	0.7	7.0	−1.0	1.3	0.7	−0.3	−8.3	2.0	−5.7	−1.3	1.0	1.3	0.3	2.0
L1	*Viburnum acerifolium*	3	14.3	21.3	9.3	44.7	7.0	20.7	13.7	35.7	−11.0	−0.7	−10.0	−1.0	14.0	4.7	1.7	4.3
L1	*Viburnum dentatum*	3	18.0	100.0	5.0	112.7	10.7	45.0	40.3	64.3	−10.0	−3.7	−9.7	−0.3	28.7	2.3	4.0	4.0
L1	*Viburnum nudum*	3	8.7	7.3	−2.0	26.0	−1.7	−1.0	0.0	15.3	−8.0	−3.7	−8.3	−8.0	7.7	1.0	0.7	1.3
L1	*Viburnum prunifolium*	3	16.0	36.3	−0.7	6.7	−3.3	10.7	5.0	4.3	−12.7	3.7	−10.7	−7.3	5.7	0.3	0.0	9.0
L1	*Prunus lasiocarpa*	3	14.3	13.3	−7.0	41.7	−1.3	13.0	2.7	36.7	−13.3	−11.7	−14.3	−5.0	2.3	8.0	0.0	2.3
L1	*Stewartia malacodendron*	2	6.0	59.0	25.0	71.0	4.5	13.0	12.0	21.0	−1.0	−1.5	−4.0	−4.0	15.5	5.0	3.0	3.0
L2	*Camellia oleifera*	3	10.0	−5.7	−8.0	−3.7	−6.3	−7.3	−3.3	−2.7	−10.0	−8.7	−10.0	−4.7	−4.7	5.3	4.7	4.0
L2	*Amelanchier stolonifera*	3	10.3	126.3	52.7	104.3	27.0	74.0	63.7	59.0	−3.7	−3.0	−8.7	0.3	79.0	6.7	1.3	1.7
L2	*Aronia arbutifolia*	3	31.0	56.0	11.3	86.7	2.3	11.0	11.7	18.0	−23.0	5.0	−8.0	2.0	6.3	2.0	1.3	5.0
L2	*Crataegus viridis*	2	19.5	61.5	11.0	78.5	3.5	59.5	45.5	33.5	−17.5	−8.0	−19.5	−5.5	19.5	−3.5	1.5	0.5
L3	*Cornus alba*	3	16.7	149.0	79.0	129.0	69.7	124.7	71.7	114.0	100.3	16.0	1.7	18.3	96.3	5.0	1.3	2.7
L3	*Kalmia latifolia*	3	14.3	79.0	45.0	42.7	27.7	46.0	34.7	40.7	−1.7	15.7	−2.7	5.0	67.7	67.0	0.7	1.7
L3	*Sideroxylon lycioides*	3	29.3	14.3	2.7	18.0	−1.7	9.7	5.7	4.3	−7.0	3.0	−11.3	−1.7	3.0	4.0	0.7	2.3
L3	*Betula nigra*	3	12.0	75.3	19.3	39.3	43.0	49.7	42.0	51.7	18.0	35.7	7.7	25.3	140.0	170.0	158.3	5.3
L3	*Betula papyrifera*	3	10.0	50.0	18.0	62.3	9.3	40.3	17.3	35.3	−9.3	−2.7	−10.0	−5.3	26.0	35.3	−1.0	0.0
L3	*Carpinus caroliniana*	3	39.3	8.7	−1.3	3.7	1.3	7.0	7.0	10.0	−13.0	5.0	−8.3	−4.3	8.7	11.3	3.7	5.3
L3	*Carya glabra*	3	28.7	7.7	−3.7	−1.3	−12.0	0.0	5.7	15.3	−19.3	3.0	−9.7	−3.7	4.7	8.0	1.0	3.7
L3	*Morella cerifera*	3	36.0	0.3	−3.7	−11.3	−6.7	15.3	10.3	14.0	−8.7	5.7	−12.3	−8.3	0.3	3.0	1.7	4.7
L3	*Lindera benzoin*	3	23.0	41.0	1.0	−1.3	12.0	26.7	22.0	32.7	−0.7	2.7	−5.7	14.7	73.7	60.0	47.3	3.7
L3	*Persea borbonia*	3	51.0	45.3	−1.7	53.7	7.3	23.7	8.7	30.7	−25.0	8.7	4.7	−7.0	60.3	29.7	2.7	10.0
L3	*Sassafras albidum*	3	39.3	25.0	−10.7	31.3	−15.7	18.0	14.3	32.0	−23.7	−5.7	−20.3	−21.7	21.7	−1.0	6.3	3.3
L3	*Zanthoxylum americanum*	3	16.3	23.7	−4.0	19.3	−5.7	3.7	2.3	13.0	−16.0	3.7	−16.3	−11.3	14.0	0.7	1.0	3.0
L3	*Zanthoxylum simulans*	3	7.3	−0.7	0.0	10.7	−5.7	−1.3	−3.7	4.7	−7.0	−5.7	−6.7	−7.0	2.0	−0.3	0.0	0.3
L3	*Acer rubrum*	3	39.0	105.7	70.7	66.7	52.0	43.0	34.7	45.3	22.7	35.3	111.3	1.7	189.3	189.7	168.3	81.0
L3	*Aesculus parviflora*	3	50.3	64.0	51.0	58.7	38.0	16.0	10.7	36.3	17.0	25.7	13.0	7.0	83.3	81.0	40.7	0.3
L4	*Fatsia japonica*	3	16.0	33.0	−2.3	81.7	3.3	10.0	12.7	27.7	−15.0	−1.7	−13.3	−10.3	0.7	3.0	0.3	2.7
L4	*Ilex cornuta*	3	30.7	13.0	−3.0	17.3	−9.3	15.7	9.0	12.0	−29.3	3.0	−13.7	−5.3	1.3	0.3	1.0	5.0
L4	*Moringa oleifera*	3	17.3	6.0	12.3	12.7	1.3	4.3	13.7	10.0	−6.7	0.0	−2.3	−1.0	4.7	12.7	−2.3	4.0
L4	*Euonymus americanus*	3	21.0	2.7	−1.7	−4.3	−2.0	1.3	3.0	2.7	−7.7	7.3	−6.7	1.3	1.3	2.0	0.3	6.7
L4	*Stephanotis floribunda*	5	39.0	24.2	−8.6	36.0	0.2	4.2	12.6	13.8	9.6	9.0	−12.8	4.4	22.8	29.8	−1.4	−0.2
L4	*Fraxinus pennsylvanica*	3	14.0	89.3	45.0	99.0	11.7	62.0	51.7	41.3	−11.0	−1.0	−4.7	−3.0	33.0	5.0	5.7	8.0
L4	*Olea europaea*	3	27.7	10.0	−2.3	47.3	3.3	19.3	9.3	23.0	−12.0	2.3	−6.3	−1.0	10.0	2.0	3.3	3.0
L4	*Magnolia virginiana*	3	49.3	4.3	1.7	−16.3	−12.7	−1.3	15.3	14.3	−16.7	8.7	−18.0	−24.3	−1.3	11.0	2.7	5.0
L4	*Passiflora coriacea*	5	40.0	5.4	−13.0	4.8	3.6	−1.4	−2.4	4.2	−12.0	−6.4	−5.6	−6.0	18.8	34.2	7.4	4.4
L4	*Hibiscus clayi*	3	11.7	6.7	10.7	0.3	3.3	2.0	2.3	1.3	−9.3	1.3	−0.7	1.7	1.3	0.0	0.3	1.0
L4	*Edgeworthia chrysantha*	3	12.3	25.3	−1.3	21.3	−1.7	22.7	13.0	28.0	−11.0	−4.3	−11.3	−7.7	54.0	3.3	36.7	43.3
L4	*Fothergilla major*	3	12.3	44.0	3.0	45.7	1.0	9.3	12.3	12.0	−5.7	0.7	−3.0	2.3	10.7	1.0	0.0	1.3
L4	*Itea virginica*	3	25.0	30.7	20.0	25.3	3.3	11.7	7.3	6.3	5.7	4.3	0.7	3.3	6.3	13.0	1.0	2.0

Figure S41: Signed count bias (predicted minus annotated stomata per image, so positive means overcounting) at $\tau = 0.25$, for every detector and every species of OD-Valid, T-MAWP and T-DCWP. Each cell prints the value. N: images. $\bar{c}_{GT}$: mean annotated stomata per image.



Figure S42: Signed count bias as in Figure S41, for the species of T-MSWP and T-GSH.

S9 Per-species results at the other confidence thresholds

Figure 2 shows that the confidence threshold changes set-level F1 and count MAE. This section shows the effect per species, for density R^2 and count MAE, at the shared threshold, at the model-specific threshold of each detector (both selected on OD-Valid, Table S8), and at the fixed thresholds $\tau = 0.50$ and $\tau = 0.75$. Figures S43–S50 give within-species density R^2, laid out as Figures S19 and S20 at $\tau = 0.25$. Figures S51–S58 give count MAE for every species in the same layout. Count MAE needs no box matching, so no IoU applies. At $\tau = 0.25$, density R^2 is in Section S7 and count MAE in Section S8.

Between $\tau = 0.25$ and $\tau = 0.75$ the T-MAWP leader by density R^2 changes for 15 of the 17 species, and the leader by count MAE for 14. Most of these changes are not near-ties. Scored at $\tau = 0.75$, the detector that led at 0.25 trails the new leader in density R^2 by a median of 2.89 over the 15 species (largest 29.55), and by less than 0.25 for only 3 of them. In count MAE it trails by a median of 12.4 stomata per image over the 14 species. These gaps use the test labels, so they describe the results after the fact and are not a way to choose a detector.

S9.1 Within-species density R^2

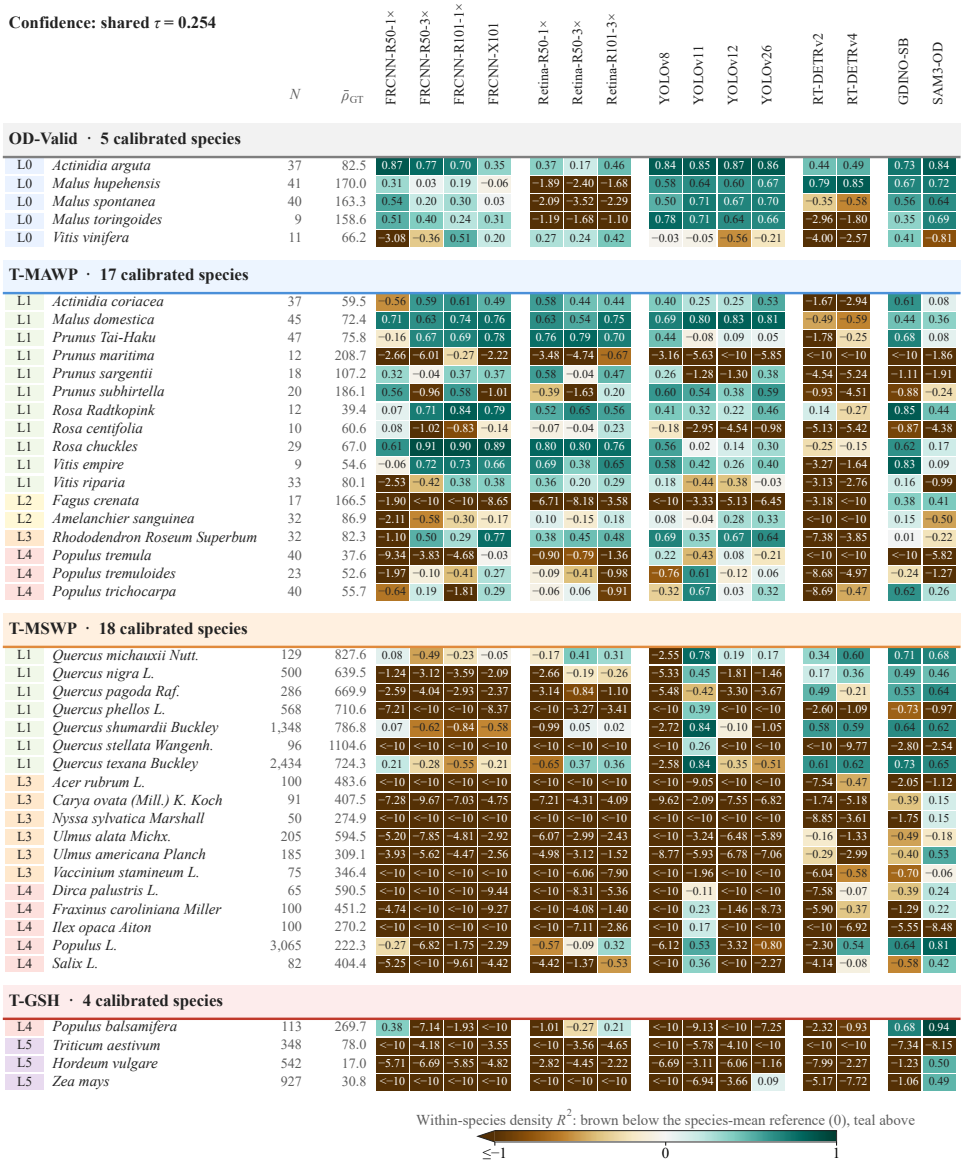


Figure S43: Within-species density R^2 at the shared threshold $\tau = 0.254$ for every detector and every calibrated species of OD-Valid, T-MAWP, T-MSWP and T-GSH, laid out as Figure S19 ($\tau = 0.25$).

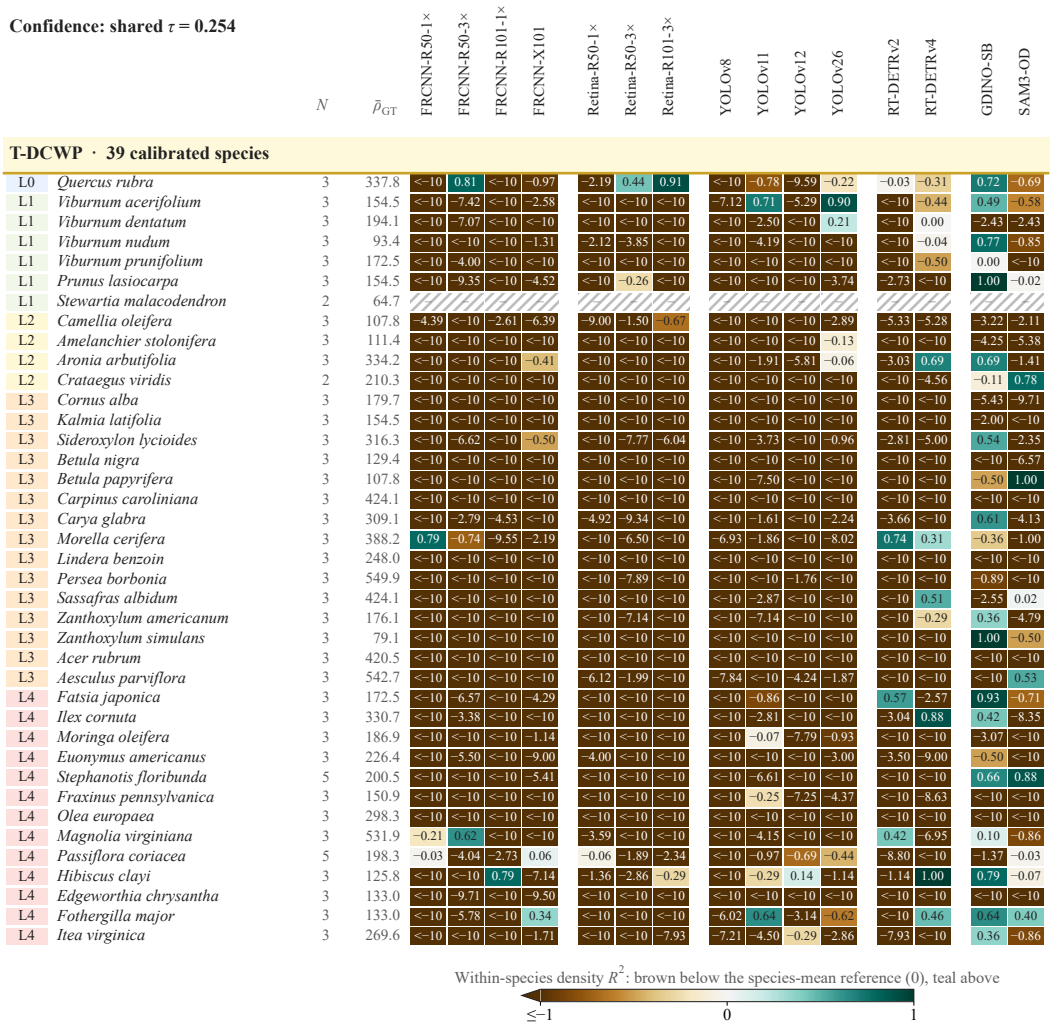


Figure S44: As Figure S43, for all 39 T-DCWP species (laid out as Figure S20).

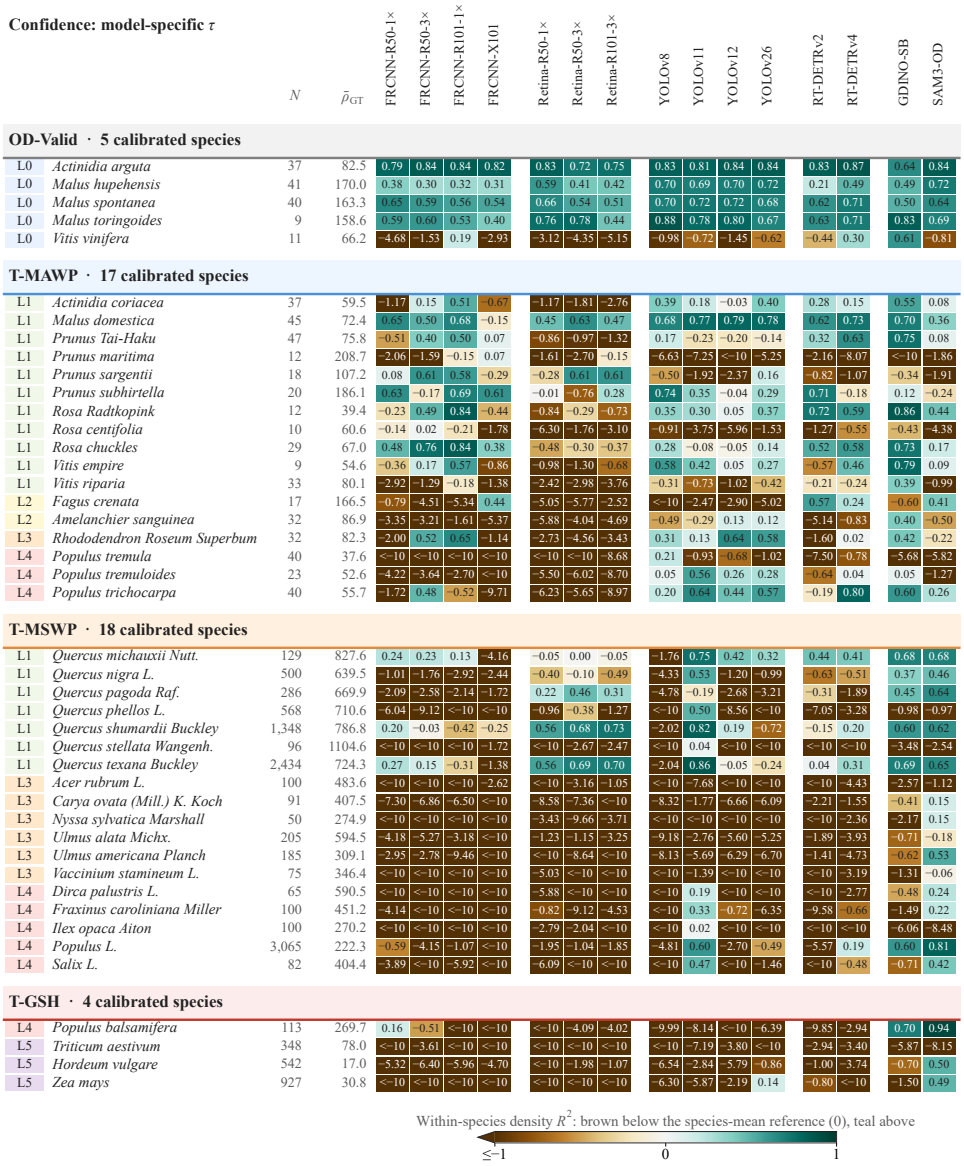


Figure S45: Within-species density R^2 at each detector's model-specific threshold (Table S8) for every detector and every calibrated species of OD-Valid, T-MAWP, T-MSWP and T-GSH, laid out as Figure S19 ($\tau = 0.25$).

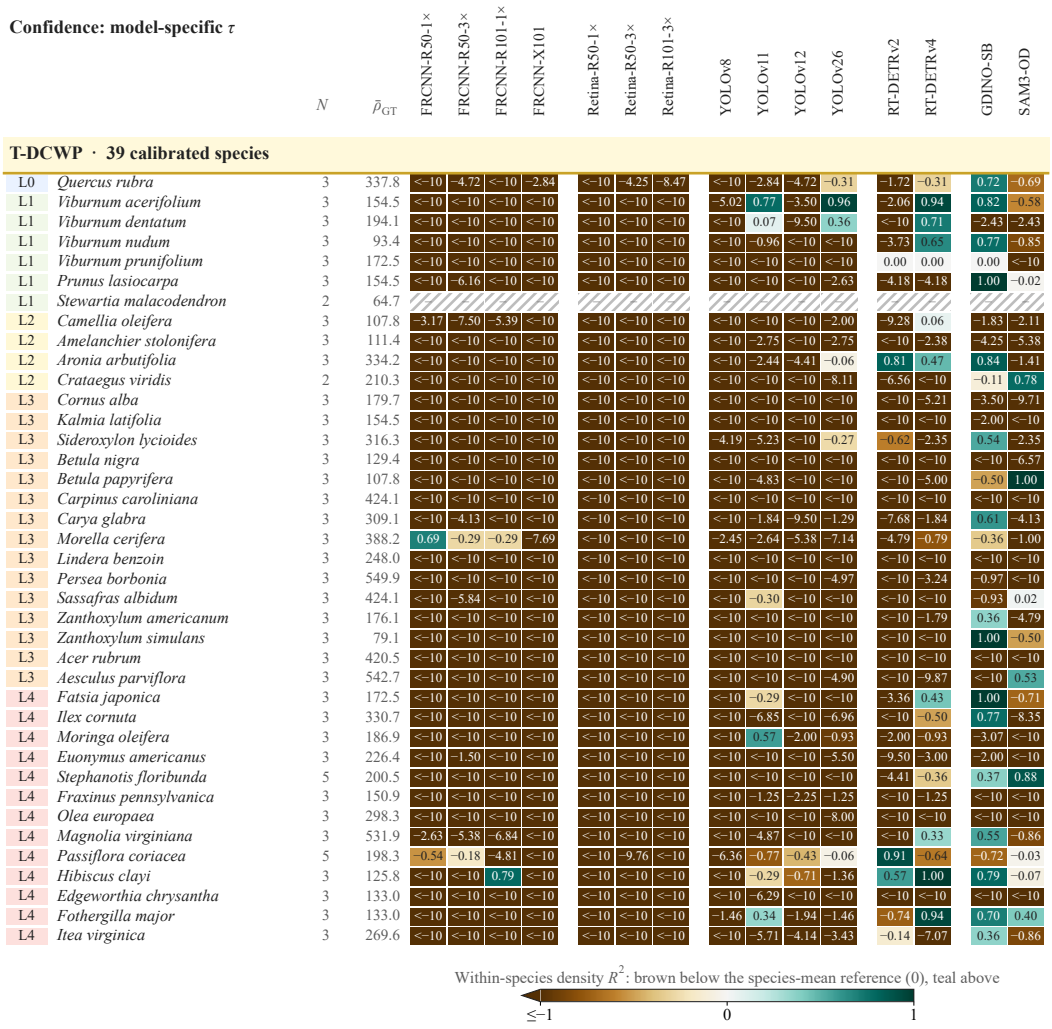


Figure S46: As Figure S45, for all 39 T-DCWP species (laid out as Figure S20).

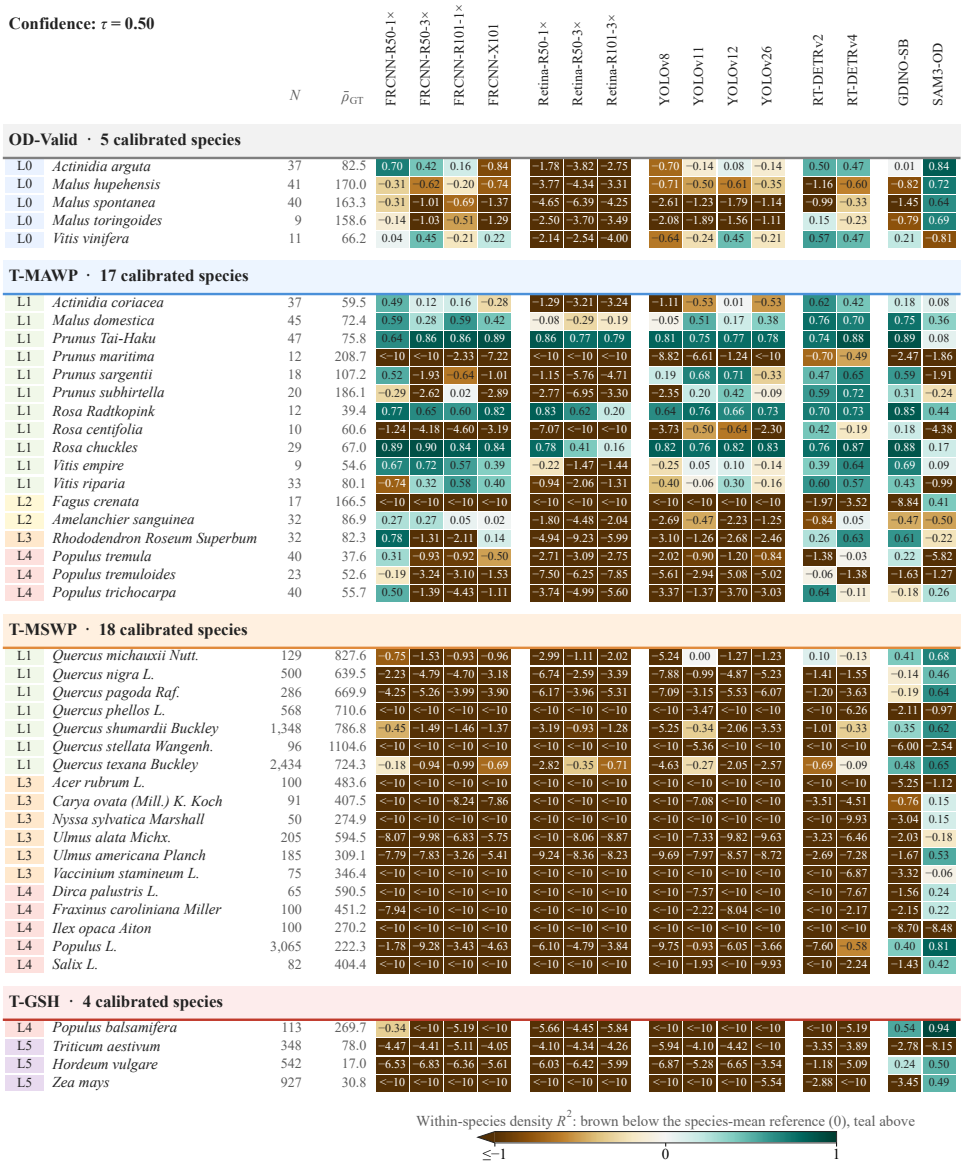


Figure S47: Within-species density R^2 at $\tau = 0.50$ for every detector and every calibrated species of OD-Valid, T-MAWP, T-MSWP and T-GSH, laid out as Figure S19 ($\tau = 0.25$).

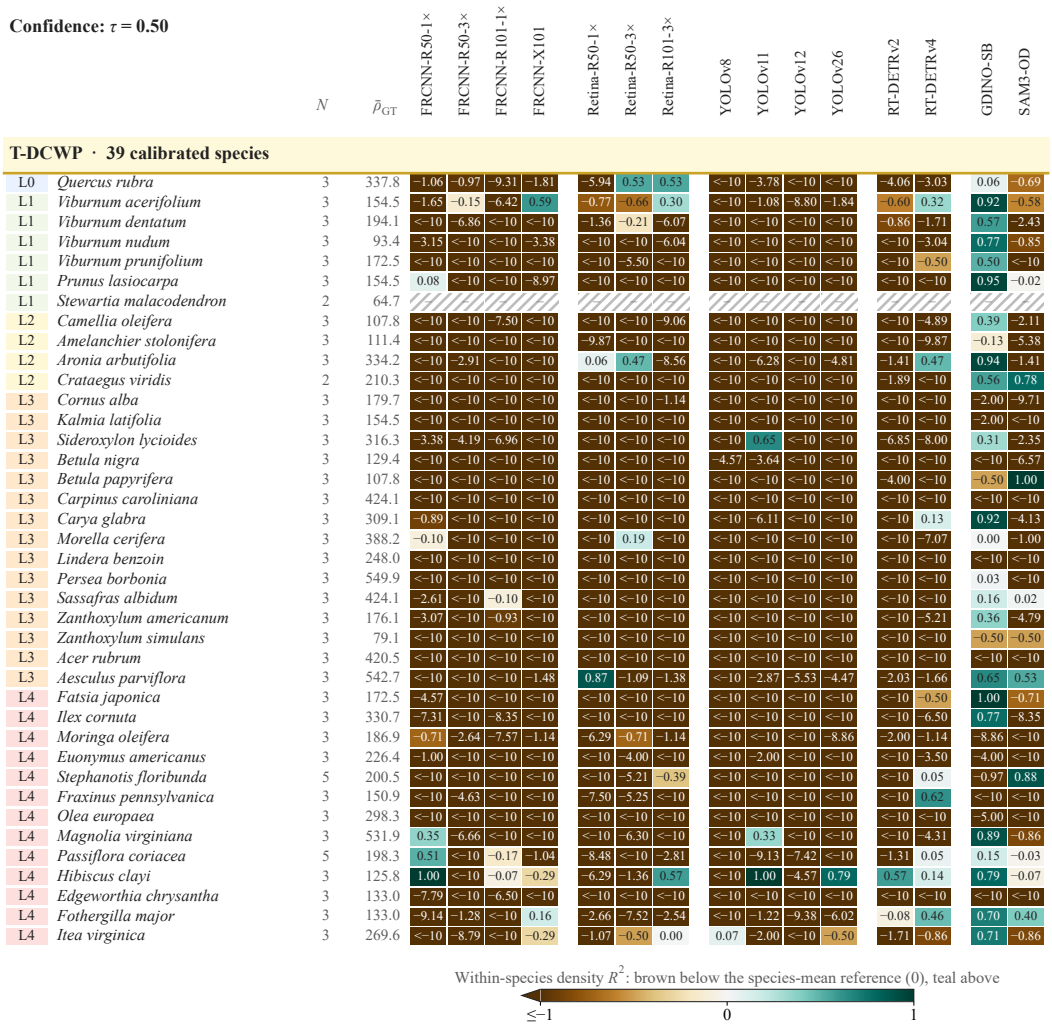


Figure S48: As Figure S47, for all 39 T-DCWP species (laid out as Figure S20).

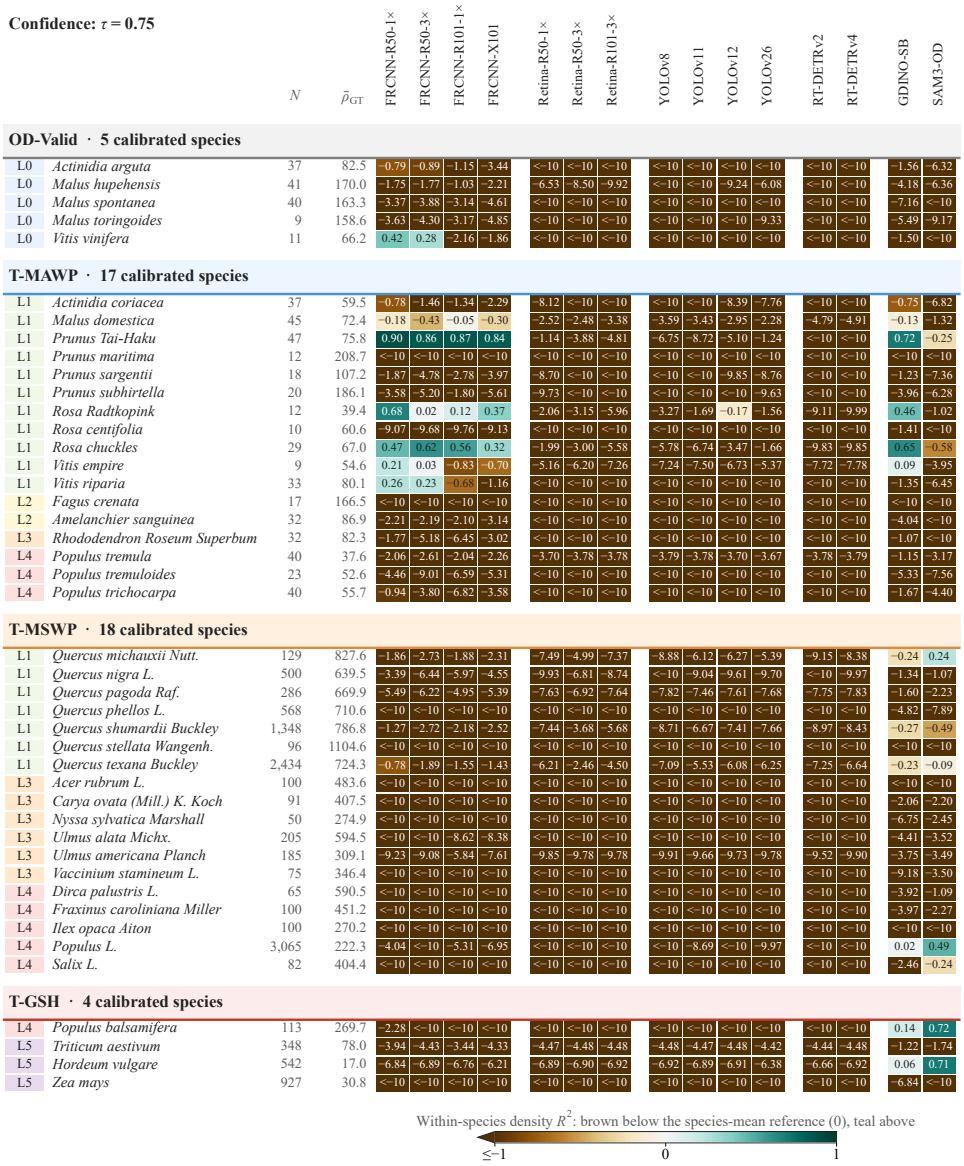


Figure S49: Within-species density R^2 at $\tau = 0.75$ for every detector and every calibrated species of OD-Valid, T-MAWP, T-MSWP and T-GSH, laid out as Figure S19 ($\tau = 0.25$).

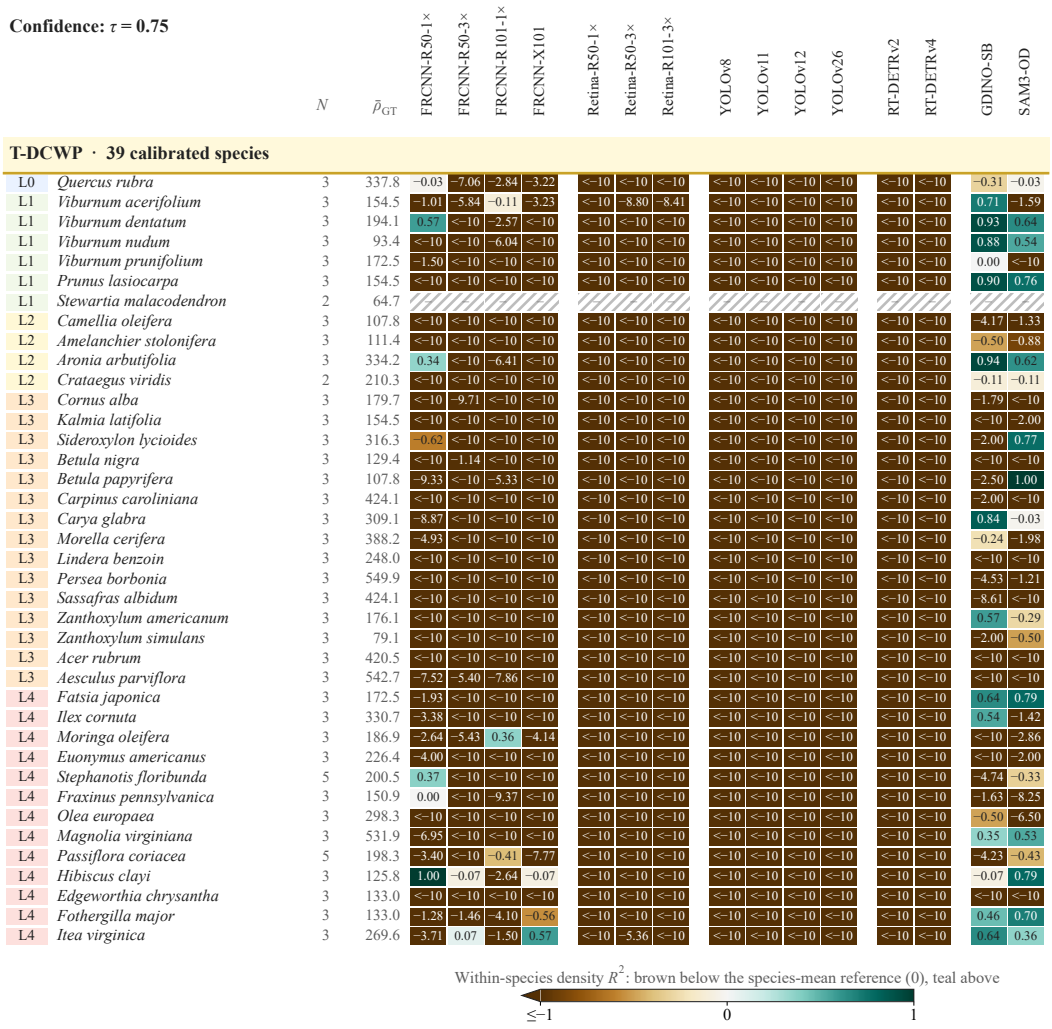


Figure S50: As Figure S49, for all 39 T-DCWP species (laid out as Figure S20).

S9.2 Count MAE

Confidence: shared $\tau = 0.254$

		N	$\bar{c}_{GT}$	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
OD-Valid · 5 species																		
L0	*Actinidia arguta*	37	38.8	2.5	3.8	4.1	6.3	6.6	7.6	5.8	2.9	3.0	2.8	2.9	5.6	5.3	3.9	2.8
L0	*Malus hupehensis*	41	79.9	10.7	13.4	12.3	14.3	23.8	25.5	22.6	7.4	7.1	7.7	6.7	5.7	4.4	7.1	6.5
L0	*Malus spontanea*	40	76.7	6.5	9.4	8.8	10.5	20.8	24.9	20.9	6.5	4.5	4.9	4.9	12.9	13.7	6.5	4.8
L0	*Malus toringoides*	9	74.6	11.9	13.7	14.9	14.3	26.7	29.7	25.4	9.0	9.9	11.2	10.3	38.2	30.0	13.7	9.9
L0	*Vitis vinifera*	11	31.1	10.4	5.6	3.4	4.2	3.7	3.9	3.4	4.7	4.8	5.6	5.3	11.5	9.5	3.4	6.5
T-MAWP · 17 species																		
L1	*Actinidia coriacea*	37	27.9	8.4	3.8	3.8	4.4	3.9	4.6	4.3	4.6	5.4	5.3	4.0	11.2	13.7	3.8	6.0
L1	*Malus domestica*	45	34.0	6.3	6.2	5.9	5.4	6.9	7.8	5.9	6.0	5.8	4.8	5.2	16.3	17.2	8.1	10.2
L1	*Prunus Tai-Haku*	47	35.6	9.6	4.8	4.7	4.0	4.0	3.6	4.3	6.3	9.2	8.7	8.6	14.9	10.0	4.7	8.6
L1	*Prunus maritima*	12	98.1	18.6	33.2	13.2	19.8	28.5	32.8	17.2	23.0	32.0	47.7	29.2	57.9	78.8	59.7	20.6
L1	*Prunus sargentii*	18	50.4	10.8	13.3	9.6	9.2	8.3	13.3	8.9	12.2	22.1	22.6	10.0	35.4	37.3	20.3	25.6
L1	*Prunus subhirtella*	20	87.5	13.2	34.0	12.2	33.1	26.4	37.6	17.6	12.4	15.1	16.9	13.8	33.7	58.2	33.4	26.4
L1	*Rosa Radtkopink*	12	18.5	4.9	2.8	2.1	2.0	3.0	2.6	3.0	3.4	3.8	4.4	3.3	4.4	5.5	1.4	3.5
L1	*Rosa centifolia*	10	28.5	4.0	5.6	5.0	3.6	3.9	3.2	3.4	4.6	9.1	10.9	6.7	11.9	12.3	5.6	11.2
L1	*Rosa chuckles*	29	31.5	5.1	2.3	2.4	2.7	3.8	3.6	3.9	5.6	8.8	8.2	7.4	9.4	9.1	4.3	7.8
L1	*Vitis empire*	9	25.7	8.1	3.9	4.2	4.7	4.3	6.3	5.0	4.6	6.2	7.7	6.2	16.9	13.8	3.2	7.6
L1	*Vitis riparia*	33	37.6	14.0	8.5	5.3	5.1	5.5	5.8	5.6	6.1	8.5	8.5	6.8	15.3	14.6	6.2	10.5
L2	*Fagus crenata*	17	78.2	21.8	46.5	46.1	42.2	37.5	41.0	27.9	52.6	26.8	32.6	36.5	27.4	45.2	8.3	8.5
L2	*Amelanchier sanguinea*	32	40.8	12.5	8.7	7.8	7.8	6.8	7.1	6.2	5.7	6.8	5.5	5.6	35.8	27.3	5.9	7.7
L3	*Rhododendron Roseum Superbum*	32	38.7	13.1	4.8	6.9	3.7	5.6	5.4	5.5	3.6	6.5	4.2	4.9	26.8	19.9	7.5	9.5
L4	*Populus tremula*	40	17.6	27.4	14.9	15.8	7.2	9.6	10.0	11.2	6.6	8.4	7.4	8.2	48.2	37.7	28.2	21.8
L4	*Populus tremuloides*	23	24.7	9.7	4.9	6.0	4.4	5.4	6.1	7.3	6.8	2.8	5.2	5.0	18.5	13.5	4.9	8.2
L4	*Populus trichocarpa*	40	26.2	8.2	5.2	9.8	4.7	5.9	5.3	7.7	6.5	3.1	5.5	4.6	20.0	7.6	3.6	4.9
T-MSWP · 18 species																		
L1	*Quercus michauxii Nutt.*	129	36.7	9.3	12.3	10.8	10.4	11.5	8.3	9.3	18.3	3.7	8.7	7.6	7.0	5.5	4.8	5.0
L1	*Quercus nigra L.*	500	26.8	10.1	14.0	14.5	12.1	13.4	7.3	7.6	17.0	4.0	10.6	9.9	5.9	5.1	4.6	4.7
L1	*Quercus pagoda Raf.*	286	51.4	23.7	30.0	22.1	25.8	26.2	19.1	19.0	38.3	11.5	23.8	29.5	9.8	12.5	7.9	8.0
L1	*Quercus phellos L.*	568	59.9	17.0	24.9	19.0	19.3	22.2	15.4	15.1	23.5	3.6	13.7	14.8	12.1	9.4	8.1	8.0
L1	*Quercus shumardii Buckley*	1,612	71.4	17.5	22.5	20.0	23.4	27.6	20.7	21.8	27.5	5.5	12.8	17.8	12.8	10.3	9.5	10.4
L1	*Quercus stellata Wangenh.*	96	40.2	28.0	36.3	37.5	32.9	33.9	12.8	20.6	35.3	2.4	31.6	34.6	13.1	11.4	6.7	6.6
L1	*Quercus texana Buckley*	2,975	66.8	17.5	21.1	19.7	21.7	25.6	16.4	15.2	27.4	5.2	13.6	17.8	12.9	11.1	9.4	11.5
L3	*Acer rubrum L.*	100	17.6	11.8	16.2	16.6	14.3	14.6	10.4	9.6	15.6	6.7	11.7	14.3	6.0	2.2	3.6	2.9
L3	*Carya ovata (Mill.) K. Koch*	91	23.0	12.6	15.0	11.9	11.0	11.4	9.1	8.3	15.2	6.0	11.0	10.5	5.6	5.9	4.2	2.8
L3	*Nyssa sylvatica Marshall*	50	10.0	8.6	9.7	9.3	7.9	8.0	7.0	7.0	8.7	5.4	8.9	8.6	4.1	2.7	2.0	1.1
L3	*Ulmus alata Michx.*	205	36.0	20.4	24.1	17.9	17.2	21.1	15.5	15.0	29.2	13.1	20.1	20.2	8.1	9.7	7.9	7.8
L3	*Ulmus americana Planch*	185	21.9	13.6	16.1	11.6	12.9	14.7	12.1	8.9	20.3	11.9	15.2	15.8	5.5	8.5	6.3	3.3
L3	*Vaccinium stamineum L.*	75	12.6	8.0	11.8	11.5	9.3	9.4	4.6	5.2	11.8	2.8	10.4	7.6	4.3	2.1	2.1	1.6
L4	*Dirca palustris L.*	65	21.5	15.7	19.4	18.5	9.8	15.9	9.2	7.7	19.6	2.7	18.6	12.8	8.6	2.8	3.4	2.4
L4	*Fraxinus caroliniana Miller*	100	16.4	4.4	10.8	11.1	5.8	7.1	3.9	2.6	7.7	1.3	2.5	5.4	3.1	1.9	2.5	1.4
L4	*Ilex opaca Aiton*	100	37.5	13.6	33.5	29.0	27.2	22.7	7.9	4.7	32.8	2.1	28.9	21.9	22.6	7.9	7.3	7.5
L4	*Populus L.*	3,065	30.9	7.3	21.3	11.7	13.0	8.6	7.0	5.2	19.8	4.0	14.7	8.2	13.1	4.1	4.1	2.7
L4	*Salix L.*	82	14.7	4.4	10.1	5.9	3.9	4.0	2.5	2.0	10.9	1.2	7.8	2.7	3.7	1.6	2.1	1.1
T-GSH · 9 species																		
L4	*Glycine max*	600	55.1	26.9	49.3	34.9	35.9	20.5	18.1	16.3	25.9	27.6	36.6	24.5	30.8	18.9	11.3	5.2
L4	*Phaseolus vulgaris*	597	32.0	12.1	20.9	37.7	20.8	26.0	15.6	21.6	22.9	11.4	15.3	13.6	14.8	13.9	7.2	2.7
L4	*Populus balsamifera*	113	37.4	5.9	26.6	13.2	34.2	10.6	8.8	6.8	33.2	30.2	35.1	27.0	14.8	12.3	4.5	2.0
L5	*Triticum aestivum*	2,443	17.9	9.8	15.3	19.8	10.4	6.7	7.8	7.1	9.1	4.8	5.0	7.8	9.8	4.2	2.6	3.1
L5	*Elaeis guineensis*	317	36.6	32.2	30.7	30.1	33.0	16.5	22.8	16.9	35.3	20.1	28.2	13.0	28.5	14.5	3.5	2.3
L5	*Hordeum vulgare*	542	27.9	26.1	27.7	25.8	24.3	20.3	24.7	21.6	27.4	22.8	26.1	15.2	17.5	21.0	11.0	4.3
L5	*Oryza glaberrima*	253	47.2	45.1	46.4	45.6	39.4	35.0	33.9	35.1	125.8	40.5	51.0	70.3	54.3	48.2	14.8	6.3
L5	*Triticum spp.*	651	78.4	66.4	75.2	68.2	64.4	58.0	62.5	68.6	68.3	72.5	75.1	63.7	52.4	31.4	19.8	13.2
L5	*Zea mays*	1,427	69.3	40.4	68.7	55.3	57.5	58.2	60.1	58.6	39.8	28.8	47.1	25.7	37.4	30.8	5.7	9.7

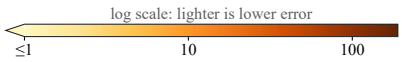


Figure S51: Per-species count MAE (stomata per image) at the shared threshold $\tau = 0.254$ for every detector and every species of OD-Valid, T-MAWP, T-MSWP and T-GSH. Each cell prints the value. Colour is on a log scale, lighter for lower error. N: images. $\bar{c}_{GT}$: mean annotated stomata per image. Count MAE needs no box matching.

Confidence: shared $\tau = 0.254$

		N	$\bar{c}_{GT}$	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
T-DCWP · 39 species																		
L0	*Quercus rubra*	3	31.3	6.7	0.7	8.3	2.3	3.3	1.3	0.3	9.0	2.3	5.7	1.7	1.7	2.0	1.0	2.0
L1	*Viburnum acerifolium*	3	14.3	21.0	8.7	44.0	6.7	20.0	13.0	35.3	11.0	2.0	10.0	1.0	14.0	4.7	1.7	4.3
L1	*Viburnum dentatum*	3	18.0	98.3	5.0	112.3	11.3	45.0	38.3	62.0	10.0	3.7	9.7	1.7	28.0	2.0	4.0	4.0
L1	*Viburnum nudum*	3	8.7	7.3	7.0	24.7	2.0	2.3	3.3	13.3	8.0	3.7	8.3	8.0	8.7	1.7	0.7	1.3
L1	*Viburnum prunifolium*	3	16.0	35.3	1.3	6.3	3.3	9.7	5.0	4.3	12.7	3.7	11.0	7.7	5.7	1.0	0.7	9.0
L1	*Prunus lasiocarpa*	3	14.3	12.3	7.3	41.3	6.0	12.3	2.7	33.7	13.3	12.0	14.3	5.3	5.0	7.3	0.0	2.3
L1	*Stewartia malacodendron*	2	6.0	57.5	24.5	70.0	5.5	13.0	11.0	20.0	2.5	2.0	4.0	4.0	15.0	5.0	3.0	3.0
L2	*Camellia oleifera*	3	10.0	5.7	8.0	3.7	6.3	7.3	3.7	2.7	10.0	8.7	10.0	4.7	5.3	5.0	4.7	4.0
L2	*Amelanchier stolonifera*	3	10.3	125.3	49.3	103.7	26.3	71.7	60.3	53.7	4.7	3.3	8.7	1.0	77.7	6.0	2.0	1.7
L2	*Aronia arbutifolia*	3	31.0	55.0	11.3	86.3	3.7	11.0	11.7	18.0	24.3	5.0	8.0	3.3	6.3	1.3	1.3	5.0
L2	*Crataegus viridis*	2	19.5	60.5	9.5	78.0	7.5	56.5	44.5	31.5	17.5	8.0	19.5	5.5	19.5	3.5	1.5	0.5
L3	*Cornus alba*	3	16.7	149.0	77.3	128.3	68.7	124.7	71.7	114.0	97.7	15.7	3.3	17.0	93.7	5.0	2.7	2.7
L3	*Kalmia latifolia*	3	14.3	78.0	45.0	42.7	26.7	46.0	33.7	40.7	2.3	15.0	6.7	5.3	66.0	65.7	0.7	1.7
L3	*Sideroxylon lycioides*	3	29.3	14.0	3.3	18.0	1.7	9.7	4.7	4.3	7.0	3.0	11.7	1.7	3.0	3.3	0.7	2.3
L3	*Betula nigra*	3	12.0	74.7	19.3	38.7	42.7	49.7	41.3	51.7	17.7	33.3	7.7	23.3	138.7	168.7	155.7	5.3
L3	*Betula papyrifera*	3	10.0	49.7	20.3	61.7	8.3	39.0	16.0	33.3	9.3	3.7	10.0	5.3	25.0	35.0	1.0	0.0
L3	*Carpinus caroliniana*	3	39.3	8.3	2.3	3.7	1.3	7.0	6.7	10.0	13.3	6.7	8.3	4.3	7.3	11.3	3.7	5.3
L3	*Carya glabra*	3	28.7	7.3	4.0	4.7	12.0	4.3	5.0	15.3	19.3	3.0	9.7	3.7	3.7	8.0	1.0	3.7
L3	*Morella cerifera*	3	36.0	1.7	3.7	11.7	6.7	15.3	9.7	13.3	10.3	5.3	12.3	8.3	1.7	2.3	4.3	4.7
L3	*Lindera benzoin*	3	23.0	40.3	7.0	2.7	12.0	26.3	21.7	32.0	4.3	2.7	7.7	14.7	72.3	58.3	46.0	3.7
L3	*Persea borbonia*	3	51.0	44.7	16.3	53.7	13.7	23.7	8.7	30.7	25.0	9.0	4.3	9.0	59.3	28.0	4.0	10.0
L3	*Sassafras albidum*	3	39.3	24.3	11.0	31.3	15.7	18.0	14.0	31.7	24.7	6.0	20.7	22.3	19.3	1.7	6.3	3.3
L3	*Zanthoxylum americanum*	3	16.3	23.0	5.0	18.7	5.7	5.3	2.7	12.3	16.0	3.3	16.3	11.3	12.0	1.3	1.0	3.0
L3	*Zanthoxylum simulans*	3	7.3	3.7	8.0	10.7	5.7	1.7	3.7	4.3	7.0	5.7	6.7	7.0	2.7	2.3	0.0	0.3
L3	*Acer rubrum*	3	39.0	105.7	70.7	66.7	51.3	43.0	34.3	45.0	22.0	34.7	109.7	20.0	187.7	186.3	166.7	81.0
L3	*Aesculus parviflora*	3	50.3	64.0	50.7	58.7	38.0	16.0	10.3	36.3	16.3	25.0	13.0	9.3	81.7	78.0	37.7	3.7
L4	*Fatsia japonica*	3	16.0	32.0	5.3	81.0	4.7	9.7	12.0	26.7	15.0	2.0	13.3	10.7	1.3	2.7	0.3	2.7
L4	*Ilex cornuta*	3	30.7	12.7	3.3	16.7	9.3	15.7	8.7	11.7	29.3	3.0	13.7	5.3	3.0	0.3	1.0	5.0
L4	*Moringa oleifera*	3	17.3	6.0	12.3	12.7	1.3	4.3	13.0	10.0	6.7	1.0	3.0	1.0	4.7	12.3	2.3	4.0
L4	*Euonymus americanus*	3	21.0	2.3	1.7	4.3	2.0	1.3	3.0	2.3	7.7	7.3	6.7	1.3	1.7	2.0	1.0	6.7
L4	*Stephanotis floribunda*	5	39.0	24.2	8.6	35.8	4.0	5.0	12.6	14.2	28.8	8.0	12.8	13.8	22.4	28.0	1.4	0.6
L4	*Fraxinus pennsylvanica*	3	14.0	88.7	44.3	97.7	11.0	61.3	51.0	39.3	11.0	1.3	4.7	3.7	32.0	5.0	5.7	8.0
L4	*Olea europaea*	3	27.7	9.7	2.7	47.0	3.3	19.3	9.3	23.0	12.7	2.7	6.3	1.7	9.3	2.0	3.3	3.0
L4	*Magnolia virginiana*	3	49.3	4.0	1.7	17.0	12.7	7.7	15.3	14.0	16.7	8.7	18.3	24.7	2.0	10.3	2.7	5.0
L4	*Passiflora coriacea*	5	40.0	5.2	13.0	4.8	3.0	4.2	7.4	4.0	12.4	7.6	7.0	7.0	18.4	33.4	7.2	4.4
L4	*Hibiscus clayi*	3	11.7	6.0	10.0	0.3	3.3	1.7	2.0	1.3	9.3	1.3	0.7	1.3	1.3	0.0	0.3	1.0
L4	*Edgeworthia chrysantha*	3	12.3	24.3	4.0	20.7	3.7	22.3	12.3	26.0	11.3	5.0	11.3	7.7	53.3	4.3	36.3	43.3
L4	*Fothergilla major*	3	12.3	41.7	5.7	44.7	1.7	8.7	11.3	11.0	5.7	1.3	3.7	2.3	10.3	1.7	1.3	1.3
L4	*Itea virginica*	3	25.0	30.0	19.7	25.3	3.3	11.7	7.3	6.3	5.7	4.3	2.0	3.3	6.3	12.7	1.0	2.0

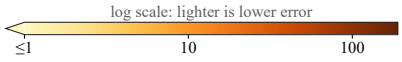


Figure S52: As Figure S51, for all 39 T-DCWP species.



Figure S53: Per-species count MAE (stomata per image) at each detector's model-specific threshold (Table S8) for every detector and every species of OD-Valid, T-MAWP, T-MSWP and T-GSH. Each cell prints the value. Colour is on a log scale, lighter for lower error. N: images. $\bar{c}_{\mathrm{GT}}$: mean annotated stomata per image. Count MAE needs no box matching.

Confidence: model-specific τ

		N	$\bar{c}_{GT}$	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
T-DCWP · 39 species																		
L0	*Quercus rubra*	3	31.3	9.7	4.3	10.3	3.0	6.7	4.0	5.0	5.7	3.0	4.3	2.0	2.3	2.0	1.0	2.0
L1	*Viburnum acerifolium*	3	14.3	28.7	29.7	65.3	55.7	37.7	24.3	40.7	9.7	1.7	8.3	0.7	5.0	1.0	1.0	4.3
L1	*Viburnum dentatum*	3	18.0	114.0	24.0	124.3	80.7	55.7	46.0	79.0	9.7	1.7	7.0	1.7	8.7	0.7	4.0	4.0
L1	*Viburnum nudum*	3	8.7	11.3	8.0	39.7	46.3	38.0	28.0	46.0	7.7	2.3	8.0	7.7	3.0	1.0	0.7	1.3
L1	*Viburnum prunifolium*	3	16.0	48.3	2.7	22.3	28.3	33.7	15.0	12.7	10.7	4.3	10.3	7.0	0.7	0.7	0.7	9.0
L1	*Prunus lasiocarpa*	3	14.3	22.0	4.7	60.7	59.3	53.0	26.3	61.3	12.7	11.0	14.3	4.3	5.0	5.7	0.0	2.3
L1	*Stewartia malacodendron*	2	6.0	76.5	58.5	101.0	94.0	65.0	57.0	75.0	4.5	2.0	3.0	3.5	7.0	1.0	2.5	3.0
L2	*Camellia oleifera*	3	10.0	5.0	5.0	5.7	20.3	13.3	13.7	39.7	9.7	8.7	10.0	4.0	7.0	1.7	3.7	4.0
L2	*Amelanchier stolonifera*	3	10.3	140.7	109.0	128.0	89.7	101.0	78.7	97.7	6.0	1.3	8.0	1.3	27.7	1.7	2.0	1.7
L2	*Aronia arbutifolia*	3	31.0	67.3	41.0	96.7	64.3	13.0	12.3	18.7	19.7	5.3	7.0	2.7	1.3	2.3	1.0	5.0
L2	*Crataegus viridis*	2	19.5	77.5	47.5	96.0	80.5	102.5	60.0	75.0	16.0	7.5	19.5	4.5	4.0	6.5	1.5	0.5
L3	*Cornus alba*	3	16.7	152.7	139.0	147.7	83.3	126.0	71.7	114.0	141.7	20.0	4.7	23.7	34.7	3.0	2.3	2.7
L3	*Kalmia latifolia*	3	14.3	85.3	76.7	58.0	84.7	84.0	66.7	53.0	6.0	18.3	9.0	7.3	22.0	24.7	0.7	1.7
L3	*Sideroxylon lycioides*	3	29.3	17.7	10.7	27.7	20.7	21.7	13.3	11.3	3.7	3.3	9.3	1.7	2.0	3.0	0.7	2.3
L3	*Betula nigra*	3	12.0	79.3	34.7	53.7	86.0	90.0	66.0	75.3	26.7	38.0	15.0	33.3	71.0	96.0	125.3	5.3
L3	*Betula papyrifera*	3	10.0	64.3	43.7	85.0	84.0	73.3	54.3	81.3	8.7	3.0	10.0	4.7	7.3	3.3	1.0	0.0
L3	*Carpinus caroliniana*	3	39.3	10.3	7.7	5.7	6.7	30.7	20.3	29.7	7.0	6.7	3.0	3.7	3.0	1.3	2.7	5.3
L3	*Carya glabra*	3	28.7	13.0	3.0	14.0	26.7	18.0	16.0	26.0	15.7	3.3	6.3	3.0	6.0	3.3	1.0	3.7
L3	*Morella cerifera*	3	36.0	1.7	3.3	4.0	11.0	30.3	23.7	36.3	6.3	6.3	8.7	8.0	9.0	4.3	4.3	4.7
L3	*Lindera benzoin*	3	23.0	47.0	12.7	9.3	51.7	67.7	47.7	70.0	7.7	5.7	8.3	19.0	29.0	17.7	32.0	3.7
L3	*Persea borbonia*	3	51.0	51.7	24.3	61.0	49.0	49.3	19.3	44.0	17.3	11.7	12.0	7.7	14.0	7.0	4.3	10.0
L3	*Sassafras albidum*	3	39.3	35.0	8.3	47.7	50.0	43.3	21.0	44.0	15.7	3.0	15.7	18.3	12.7	15.3	4.3	3.3
L3	*Zanthoxylum americanum*	3	16.3	31.3	8.7	34.0	54.0	57.3	32.7	49.0	16.0	5.0	16.0	9.7	5.0	1.7	1.0	3.0
L3	*Zanthoxylum simulans*	3	7.3	4.3	11.7	20.3	63.3	52.3	19.0	74.7	7.0	5.3	6.7	7.0	4.3	3.7	0.0	0.3
L3	*Acer rubrum*	3	39.0	111.0	99.3	78.7	61.0	60.0	56.3	53.7	43.7	40.7	129.0	24.3	78.7	122.0	144.0	81.0
L3	*Aesculus parviflora*	3	50.3	67.7	73.7	69.0	49.7	36.0	24.0	38.0	36.3	30.3	24.0	12.3	28.3	19.0	23.7	3.7
L4	*Fatsia japonica*	3	16.0	42.0	10.3	91.7	79.0	53.3	47.3	52.3	13.3	2.0	12.7	8.7	4.3	1.3	0.0	2.7
L4	*Ilex cornuta*	3	30.7	16.7	12.0	33.3	46.0	29.0	15.0	19.7	29.3	4.7	11.7	4.3	11.0	1.7	0.7	5.0
L4	*Moringa oleifera*	3	17.3	8.0	25.0	21.3	24.3	25.7	33.3	23.7	6.0	0.7	2.0	1.0	2.0	1.7	2.3	4.0
L4	*Euonymus americanus*	3	21.0	3.0	1.0	3.0	14.3	12.3	8.3	11.7	6.3	7.7	4.3	1.7	2.3	1.3	1.3	6.7
L4	*Stephanotis floribunda*	5	39.0	28.6	3.2	46.8	57.6	26.4	16.2	17.2	40.2	12.4	11.0	16.8	5.2	4.4	1.8	0.6
L4	*Fraxinus pennsylvanica*	3	14.0	100.7	101.3	128.0	86.0	77.7	57.0	66.3	9.7	2.0	2.7	2.0	12.0	2.0	5.0	8.0
L4	*Olea europaea*	3	27.7	13.3	7.3	62.7	42.3	27.3	15.3	27.0	10.0	2.7	5.3	1.3	1.7	4.0	2.7	3.0
L4	*Magnolia virginiana*	3	49.3	6.7	9.7	8.0	28.7	16.7	16.7	46.7	12.7	9.3	13.0	19.7	16.3	2.7	2.0	5.0
L4	*Passiflora coriacea*	5	40.0	6.8	4.8	7.8	21.2	17.4	6.2	11.4	7.2	5.8	6.2	5.6	0.8	7.4	5.4	4.4
L4	*Hibiscus clayi*	3	11.7	13.7	20.3	0.3	64.3	20.3	18.7	23.0	6.7	1.3	1.3	1.7	0.7	0.0	0.3	1.0
L4	*Edgeworthia chrysantha*	3	12.3	36.0	6.0	40.0	57.3	53.7	30.0	60.3	9.7	3.3	9.7	6.3	24.3	5.7	26.3	43.3
L4	*Fothergilla major*	3	12.3	53.7	13.3	63.0	42.0	34.7	37.0	34.7	3.0	1.7	3.7	3.0	2.3	0.3	1.0	1.3
L4	*Itea virginica*	3	25.0	32.7	34.0	31.3	15.3	27.3	33.7	12.0	9.0	4.7	4.0	4.0	1.3	5.0	1.0	2.0

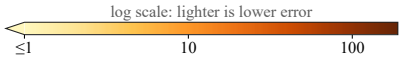


Figure S54: As Figure S53, for all 39 T-DCWP species.

Confidence: $\tau = 0.50$

		N	$\bar{c}_{GT}$	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
OD-Valid · 5 species																		
L0	*Actinidia arguta*	37	38.8	4.4	6.1	7.6	11.5	14.4	19.4	17.1	11.2	8.3	7.7	8.8	5.3	5.9	8.4	2.8
L0	*Malus hupehensis*	41	79.9	15.9	17.8	15.2	18.4	31.6	33.4	30.0	17.1	16.1	16.6	14.9	20.7	17.2	19.2	6.5
L0	*Malus spontanea*	40	76.7	12.8	16.4	14.7	17.6	28.6	32.8	27.3	21.7	16.5	18.9	16.4	15.8	13.1	17.9	4.8
L0	*Malus toringoides*	9	74.6	18.9	26.7	21.3	28.2	36.1	42.4	41.1	34.0	33.2	30.2	28.1	14.7	20.0	25.1	9.9
L0	*Vitis vinifera*	11	31.1	4.4	3.1	4.6	3.7	8.7	9.5	11.3	5.8	5.3	3.1	5.1	3.1	3.2	3.8	6.5
T-MAWP · 17 species																		
L1	*Actinidia coriacea*	37	27.9	4.2	6.1	6.1	7.3	10.2	14.2	14.4	8.6	6.5	5.6	7.1	3.5	4.8	5.4	6.0
L1	*Malus domestica*	45	34.0	7.0	9.8	7.3	9.2	13.6	14.7	14.9	13.1	9.2	12.6	10.4	5.8	6.3	6.1	10.2
L1	*Prunus Tai-Haku*	47	35.6	5.1	3.0	2.9	2.4	2.9	3.8	3.6	3.6	4.1	3.8	3.7	3.9	2.8	2.4	8.6
L1	*Prunus maritima*	12	98.1	46.6	56.4	22.8	37.3	53.2	82.1	59.7	42.0	33.8	17.6	67.4	16.8	15.3	22.6	20.6
L1	*Prunus sargentii*	18	50.4	7.9	25.0	17.6	19.9	21.1	38.9	35.8	11.4	6.0	6.2	15.0	10.0	6.4	8.0	25.6
L1	*Prunus subhirtella*	20	87.5	26.2	46.8	22.4	47.5	47.6	69.6	51.0	44.1	20.9	17.2	24.1	12.2	10.3	18.8	26.4
L1	*Rosa Radtkopink*	12	18.5	2.1	2.7	2.5	2.0	2.2	2.9	4.3	2.7	2.0	2.7	2.5	2.3	2.0	1.9	3.5
L1	*Rosa centifolia*	10	28.5	5.9	10.7	10.4	9.1	13.8	17.7	19.9	9.5	4.4	4.8	6.4	2.8	4.3	4.0	11.2
L1	*Rosa chuckles*	29	31.5	2.7	2.4	3.0	3.0	3.7	5.8	7.6	3.4	4.0	3.5	3.4	3.4	2.3	2.9	7.8
L1	*Vitis empire*	9	25.7	4.9	4.0	3.7	4.2	8.7	12.6	12.2	8.1	6.8	6.1	8.0	6.0	4.0	4.1	7.6
L1	*Vitis riparia*	33	37.6	9.6	5.2	3.9	4.6	9.0	11.8	10.0	7.5	6.4	5.2	6.9	4.2	4.0	4.5	10.5
L2	*Fagus crenata*	17	78.2	49.9	60.6	62.1	56.8	67.8	74.1	66.4	68.4	55.3	58.8	60.6	22.7	27.9	43.2	8.5
L2	*Amelanchier sanguinea*	32	40.8	5.6	5.6	6.2	5.8	12.2	17.5	12.8	13.9	8.1	12.6	10.1	9.3	6.1	7.6	7.7
L3	*Rhododendron Roseum Superbum*	32	38.7	3.4	13.4	15.7	7.7	22.7	30.0	24.5	18.7	12.9	17.2	16.8	6.6	4.8	5.0	9.5
L4	*Populus tremula*	40	17.6	6.3	10.7	10.6	9.2	15.2	16.1	15.6	14.1	10.6	11.7	10.6	12.2	7.5	6.4	21.8
L4	*Populus tremuloides*	23	24.7	5.5	11.5	11.5	8.8	17.5	16.0	17.7	15.0	11.3	14.6	14.3	5.2	7.9	9.0	8.2
L4	*Populus trichocarpa*	40	26.2	4.0	9.5	14.8	8.8	13.9	15.7	16.8	12.8	9.3	13.9	12.8	3.5	6.4	6.5	4.9
T-MSWP · 18 species																		
L1	*Quercus michauxii Nutt.*	129	36.7	13.6	16.8	13.9	14.9	22.0	16.3	19.6	27.1	10.1	15.8	15.2	10.4	11.6	7.8	5.0
L1	*Quercus nigra L.*	500	26.8	12.6	17.0	16.5	14.5	20.2	13.7	15.3	21.4	9.5	16.6	17.2	11.2	11.3	7.4	4.7
L1	*Quercus pagoda Raf.*	286	51.4	32.6	36.6	27.6	33.7	41.9	34.2	40.3	47.5	27.8	37.4	42.2	21.4	30.6	14.5	8.0
L1	*Quercus phellos L.*	568	59.9	22.0	31.1	22.6	24.6	34.3	24.7	30.3	38.7	11.7	22.8	25.3	24.7	18.3	12.2	8.0
L1	*Quercus shumardii Buckley*	1,612	71.4	21.7	28.4	22.8	28.0	40.6	30.1	33.7	46.8	18.7	27.8	35.2	29.7	22.6	14.2	10.4
L1	*Quercus stellata Wangenh.*	96	40.2	33.1	38.8	38.6	36.5	39.6	29.7	36.0	38.6	8.4	38.8	39.3	25.6	27.7	9.5	6.6
L1	*Quercus texana Buckley*	2,975	66.8	21.5	25.9	22.1	25.3	40.2	25.4	29.3	47.3	19.7	26.9	35.5	28.5	21.7	13.8	11.5
L3	*Acer rubrum L.*	100	17.6	13.8	16.8	17.0	16.1	16.8	14.8	15.6	17.1	11.8	15.3	16.8	12.0	8.5	5.5	2.9
L3	*Carya ovata (Mill.) K. Koch*	91	23.0	15.9	17.9	14.0	15.1	18.4	16.5	18.1	20.5	12.1	15.9	15.9	9.4	10.1	5.5	2.8
L3	*Nyssa sylvatica Marshall*	50	10.0	9.1	9.9	9.5	8.6	9.8	9.5	9.0	9.6	7.7	9.5	9.6	8.2	4.7	2.7	1.1
L3	*Ulmus alata Michx.*	205	36.0	25.9	28.0	21.7	23.0	30.2	24.7	28.3	34.1	23.1	27.9	28.5	17.8	22.7	13.2	7.8
L3	*Ulmus americana Planch*	185	21.9	19.0	19.2	13.7	17.1	20.7	19.4	19.5	21.6	17.0	18.9	19.4	11.4	17.7	9.4	3.3
L3	*Vaccinium stamineum L.*	75	12.6	9.6	12.2	12.0	10.9	11.8	9.6	11.2	12.4	6.8	12.0	11.0	9.7	5.1	3.6	1.6
L4	*Dirca palustris L.*	65	21.5	18.4	20.6	19.9	15.0	21.1	19.9	19.6	21.0	9.0	20.3	18.5	17.1	9.4	4.8	2.4
L4	*Fraxinus caroliniana Miller*	100	16.4	5.6	12.5	12.3	7.8	11.6	9.1	7.0	11.5	2.9	5.3	10.4	8.3	3.0	3.1	1.4
L4	*Ilex opaca Aiton*	100	37.5	21.0	36.4	33.5	32.5	35.6	19.8	31.1	36.6	10.0	35.0	33.4	34.7	16.2	9.0	7.5
L4	*Populus L.*	3,065	30.9	11.8	24.7	15.5	17.5	20.2	18.2	16.2	25.3	9.6	19.8	15.1	22.7	8.9	5.5	2.7
L4	*Salix L.*	82	14.7	6.6	12.3	7.5	6.2	10.3	8.8	8.2	12.9	2.7	10.8	5.8	9.2	3.0	2.5	1.1
T-GSH · 9 species																		
L4	*Glycine max*	600	55.1	37.5	51.8	36.0	42.7	36.8	36.6	34.1	44.0	41.9	45.1	39.1	29.4	33.2	21.0	5.2
L4	*Phaseolus vulgaris*	597	32.0	11.8	24.5	13.5	25.4	14.4	14.7	14.8	27.8	22.3	21.0	20.8	17.9	16.7	9.7	2.7
L4	*Populus balsamifera*	113	37.4	10.1	33.6	22.9	36.2	23.9	21.7	24.2	36.8	36.5	36.6	33.5	35.0	23.2	6.0	2.0
L5	*Triticum aestivum*	2,443	17.9	11.7	16.3	11.1	13.4	15.3	15.2	16.5	10.3	5.3	6.4	4.7	2.5	9.4	2.8	3.1
L5	*Elaeis guineensis*	317	36.6	34.5	32.6	33.0	34.4	26.9	30.7	28.0	36.5	26.0	32.9	22.4	8.2	24.8	7.0	2.3
L5	*Hordeum vulgare*	542	27.9	27.5	27.9	27.1	25.8	26.5	27.4	27.0	27.8	26.1	27.4	21.9	18.1	25.9	6.9	4.3
L5	*Oryza glaberrima*	253	47.2	36.4	46.9	35.7	44.6	44.9	44.5	44.7	48.0	40.9	44.7	43.7	30.3	42.2	18.9	6.3
L5	*Triticum spp.*	651	78.4	69.0	76.2	70.5	68.2	72.7	72.8	76.0	73.6	76.0	76.9	69.7	72.8	59.8	41.0	13.2
L5	*Zea mays*	1,427	69.3	52.5	69.1	62.2	64.8	68.5	69.0	68.7	58.2	50.1	60.9	45.3	50.7	62.1	21.9	9.7

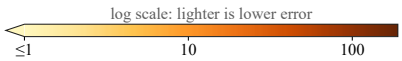


Figure S55: Per-species count MAE (stomata per image) at $\tau = 0.50$ for every detector and every species of OD-Valid, T-MAWP, T-MSWP and T-GSH. Each cell prints the value. Colour is on a log scale, lighter for lower error. N: images. $\bar{c}_{GT}$: mean annotated stomata per image. Count MAE needs no box matching.

Confidence: $\tau = 0.50$

		N	$\bar{c}_{GT}$	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
T-DCWP · 39 species																		
L0	*Quercus rubra*	3	31.3	2.7	2.3	6.0	2.7	4.7	1.0	1.0	19.7	3.7	12.7	10.7	3.3	3.7	1.3	2.0
L1	*Viburnum acerifolium*	3	14.3	5.0	4.0	10.3	2.0	4.7	5.0	2.7	13.7	5.7	12.3	6.7	4.7	3.0	0.7	4.3
L1	*Viburnum dentatum*	3	18.0	26.7	6.0	55.7	6.7	3.0	2.3	5.0	17.0	11.7	13.3	7.3	2.7	2.7	1.3	4.0
L1	*Viburnum nudum*	3	8.7	3.3	7.7	5.3	3.3	8.0	7.3	4.3	8.3	6.3	8.3	8.3	6.0	3.0	0.7	1.3
L1	*Viburnum prunifolium*	3	16.0	5.7	3.3	8.7	6.3	4.0	1.7	3.3	15.0	5.3	12.7	12.3	3.0	1.0	0.3	9.0
L1	*Prunus lasiocarpa*	3	14.3	2.3	11.0	10.7	8.0	11.0	12.0	11.0	14.3	13.7	14.3	12.0	9.0	9.0	0.3	2.3
L1	*Stewartia malacodendron*	2	6.0	13.0	0.5	24.5	4.5	1.5	4.0	2.5	5.0	5.5	6.0	4.5	4.0	1.5	2.5	3.0
L2	*Camellia oleifera*	3	10.0	8.3	8.3	7.0	8.3	9.0	8.7	7.7	10.0	9.7	10.0	9.0	8.0	5.3	1.7	4.0
L2	*Amelanchier stolonifera*	3	10.3	34.0	6.3	34.0	3.7	3.0	4.3	6.3	9.3	8.3	10.3	4.3	10.3	2.3	1.0	1.7
L2	*Aronia arbutifolia*	3	31.0	12.3	5.7	30.3	15.3	2.7	1.7	9.3	30.0	8.3	15.0	7.3	5.0	2.3	0.7	5.0
L2	*Crataegus viridis*	2	19.5	12.0	8.0	20.5	11.5	14.5	12.5	17.5	19.0	19.0	19.5	14.0	2.5	11.0	1.0	0.5
L3	*Cornus alba*	3	16.7	93.7	21.7	60.0	18.7	5.3	4.0	1.3	11.3	6.7	6.0	4.0	15.7	3.7	2.0	2.7
L3	*Kalmia latifolia*	3	14.3	35.7	11.7	16.0	9.3	1.7	5.7	3.7	12.0	7.7	12.7	6.3	4.3	7.7	0.7	1.7
L3	*Sideroxylon lycioides*	3	29.3	2.7	3.7	3.7	8.0	6.3	6.7	9.0	15.7	1.0	18.3	10.7	4.7	4.7	1.3	2.3
L3	*Betula nigra*	3	12.0	50.0	10.0	22.7	23.7	23.7	14.7	12.0	4.7	4.3	7.7	7.0	29.3	44.3	24.0	5.3
L3	*Betula papyrifera*	3	10.0	11.0	4.3	17.0	5.7	6.7	9.0	7.3	10.0	9.3	10.0	8.3	2.7	6.0	1.0	0.0
L3	*Carpinus caroliniana*	3	39.3	4.7	10.7	2.0	4.0	8.7	3.7	2.7	32.0	13.7	20.0	16.0	8.0	7.0	1.7	5.3
L3	*Carya glabra*	3	28.7	2.7	7.0	15.0	16.3	20.3	10.3	15.7	27.0	5.3	18.7	15.7	10.7	1.7	0.3	3.7
L3	*Morella cerifera*	3	36.0	3.3	14.0	19.3	14.0	14.3	3.3	13.0	26.7	10.7	20.7	17.3	13.7	10.3	3.3	4.7
L3	*Lindera benzoin*	3	23.0	20.7	7.0	8.7	4.7	5.3	6.0	6.0	19.3	15.7	18.3	12.7	4.3	8.0	3.0	3.7
L3	*Persea borbonia*	3	51.0	18.7	20.3	21.7	19.7	21.0	21.0	27.7	44.0	15.3	20.0	32.7	17.7	19.0	2.3	10.0
L3	*Sassafras albidum*	3	39.3	6.3	23.3	3.3	26.7	26.7	27.7	28.3	35.7	29.0	34.3	32.7	24.7	28.0	2.3	3.3
L3	*Zanthoxylum americanum*	3	16.3	2.3	11.0	1.0	10.3	14.7	13.7	7.7	16.0	10.7	16.3	15.3	9.3	3.0	1.0	3.0
L3	*Zanthoxylum simulans*	3	7.3	5.7	5.0	3.3	6.7	7.0	7.0	6.7	7.3	6.7	7.3	7.0	5.0	5.7	0.3	0.3
L3	*Acer rubrum*	3	39.0	84.3	36.7	43.3	25.0	24.3	17.3	14.3	21.3	8.3	38.0	29.7	12.0	65.7	72.3	81.0
L3	*Aesculus parviflora*	3	50.3	51.3	16.0	46.3	9.0	2.0	8.3	8.3	28.3	11.7	15.3	14.0	8.3	9.7	3.3	3.7
L4	*Fatsia japonica*	3	16.0	4.7	13.3	46.3	9.0	12.0	9.3	8.3	15.3	8.0	15.3	15.0	7.7	2.3	0.0	2.7
L4	*Ilex cornuta*	3	30.7	4.0	16.3	4.3	16.7	10.7	9.0	14.7	30.7	10.3	24.3	20.3	15.3	3.7	0.7	5.0
L4	*Moringa oleifera*	3	17.3	1.3	2.3	2.7	1.3	3.3	1.3	1.3	13.3	5.0	3.7	3.3	2.0	1.3	3.3	4.0
L4	*Euonymus americanus*	3	21.0	0.7	3.0	7.3	4.0	4.0	1.3	3.0	14.0	1.3	13.0	6.0	4.0	1.0	1.3	6.7
L4	*Stephanotis floribunda*	5	39.0	9.6	18.4	16.4	11.0	8.4	3.2	3.2	17.2	13.2	22.8	14.4	8.2	2.0	3.0	0.6
L4	*Fraxinus pennsylvanica*	3	14.0	26.0	3.7	15.7	5.7	3.3	4.0	8.3	13.7	7.7	9.3	10.3	5.3	1.0	4.7	8.0
L4	*Olea europaea*	3	27.7	1.7	6.0	14.3	2.7	2.0	6.3	4.3	21.0	4.7	11.3	6.3	7.0	6.7	0.7	3.0
L4	*Magnolia virginiana*	3	49.3	2.3	10.0	30.0	20.0	39.3	10.0	40.3	30.0	2.7	30.3	40.3	25.7	7.7	1.0	5.0
L4	*Passiflora coriacea*	5	40.0	3.8	19.6	3.4	8.0	16.8	19.6	11.0	25.4	18.6	17.0	20.6	9.0	5.8	4.8	4.4
L4	*Hibiscus clayi*	3	11.7	0.0	3.3	1.0	1.3	3.3	1.7	0.7	11.0	0.0	2.7	0.3	0.7	0.7	0.3	1.0
L4	*Edgeworthia chrysantha*	3	12.3	3.7	7.0	3.0	7.7	8.0	7.0	8.7	12.3	9.0	12.3	10.3	9.7	9.7	7.0	43.3
L4	*Fothergilla major*	3	12.3	6.3	3.3	19.0	2.0	4.3	6.7	4.3	8.7	2.3	6.3	5.0	2.0	1.7	1.0	1.3
L4	*Itea virginica*	3	25.0	11.7	6.3	11.0	2.0	3.0	2.3	2.0	1.7	3.3	7.0	2.3	3.3	2.0	0.7	2.0

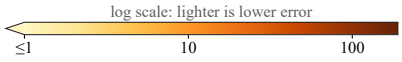


Figure S56: As Figure S55, for all 39 T-DCWP species.

Confidence: $\tau = 0.75$

		N	$\bar{c}_{GT}$	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
OD-Valid · 5 species																		
L0	*Actinidia arguta*	37	38.8	11.5	11.6	12.8	18.4	32.7	35.5	36.8	37.9	37.8	36.5	32.1	38.5	38.8	13.9	24.0
L0	*Malus hupehensis*	41	79.9	23.5	23.6	20.0	25.1	39.7	44.4	48.3	50.7	52.1	45.4	36.8	79.9	74.6	32.8	38.5
L0	*Malus spontanea*	40	76.7	24.9	26.1	23.7	28.1	49.5	57.7	59.5	64.1	65.7	59.4	49.9	76.7	76.3	34.2	46.6
L0	*Malus toringoides*	9	74.6	41.4	45.2	38.8	47.0	65.0	69.9	71.9	71.1	71.7	69.3	62.8	74.6	74.6	48.8	62.4
L0	*Vitis vinifera*	11	31.1	3.1	4.0	9.1	8.3	24.0	24.5	27.0	25.9	25.9	23.6	21.8	30.0	30.8	7.5	17.7
T-MAWP · 17 species																		
L1	*Actinidia coriacea*	37	27.9	9.0	10.4	10.3	12.2	22.0	24.8	26.7	24.7	25.2	22.1	21.2	27.7	27.9	8.7	20.3
L1	*Malus domestica*	45	34.0	14.2	15.7	13.4	15.2	26.2	25.8	29.2	29.7	29.3	28.1	25.6	33.0	33.8	14.4	21.3
L1	*Prunus Tai-Haku*	47	35.6	2.6	2.9	2.9	3.3	13.1	20.6	22.7	26.5	29.3	23.1	13.7	35.5	35.6	4.2	9.5
L1	*Prunus maritima*	12	98.1	79.7	78.6	50.1	67.8	95.6	98.0	97.8	98.1	98.1	96.8	97.7	98.1	98.1	74.6	92.4
L1	*Prunus sargentii*	18	50.4	25.1	35.9	28.8	33.3	46.6	49.9	49.9	50.1	49.6	49.0	46.6	50.3	50.4	21.2	42.9
L1	*Prunus subhirtella*	20	87.5	52.9	61.4	41.0	62.9	81.2	87.0	86.5	87.5	87.1	85.6	81.2	87.2	87.5	53.6	66.6
L1	*Rosa Radtkopink*	12	18.5	2.8	4.5	4.2	3.8	9.2	11.0	14.3	11.0	8.8	5.6	8.2	17.2	18.5	3.4	7.5
L1	*Rosa centifolia*	10	28.5	15.0	16.1	15.7	15.2	25.1	26.5	28.4	28.0	28.2	27.0	25.3	28.5	28.5	6.9	23.9
L1	*Rosa chuckles*	29	31.5	5.4	5.0	5.3	6.3	15.8	18.1	24.3	24.5	26.3	19.2	14.8	31.4	31.5	4.9	11.0
L1	*Vitis empire*	9	25.7	4.8	6.9	9.9	9.2	20.9	23.0	24.6	24.7	25.1	23.9	21.7	25.4	25.7	6.8	18.3
L1	*Vitis riparia*	33	37.6	4.6	5.1	8.6	9.8	28.6	31.0	32.7	33.2	32.9	32.8	27.5	37.5	37.5	10.1	20.2
L2	*Fagus crenata*	17	78.2	67.4	70.9	71.2	68.4	77.6	78.2	78.2	78.2	78.1	77.8	77.5	78.2	78.2	64.6	71.5
L2	*Amelanchier sanguinea*	32	40.8	13.1	13.2	12.6	15.2	32.1	35.3	34.7	38.7	38.5	36.0	32.5	40.8	40.8	16.8	27.4
L3	*Rhododendron Roseum Superbum*	32	38.7	15.0	22.9	25.2	18.1	37.0	38.3	38.1	38.2	38.4	38.3	37.3	38.7	38.7	12.4	35.2
L4	*Populus tremula*	40	17.6	13.7	14.7	13.1	14.3	17.5	17.6	17.6	17.6	17.6	17.5	17.4	17.6	17.6	10.9	16.4
L4	*Populus tremuloides*	23	24.7	13.6	18.7	16.5	14.7	23.6	23.4	23.9	22.9	23.7	23.2	22.7	24.6	24.7	14.5	17.3
L4	*Populus trichocarpa*	40	26.2	8.4	14.2	18.5	13.8	23.9	24.3	24.3	23.6	23.9	23.4	22.9	25.9	25.6	10.7	15.2
T-MSWP · 18 species																		
L1	*Quercus michauxii Nutt.*	129	36.7	18.1	21.3	17.3	20.1	33.6	28.0	33.4	36.0	30.7	31.0	29.0	36.4	35.3	11.9	9.5
L1	*Quercus nigra L.*	500	26.8	15.1	19.8	18.5	16.9	25.1	20.6	23.6	26.1	24.0	24.2	24.7	26.5	25.8	11.0	10.1
L1	*Quercus pagoda Raf.*	286	51.4	39.6	41.6	33.3	40.4	50.3	47.3	50.8	51.3	49.5	50.1	50.6	51.2	51.4	24.6	30.9
L1	*Quercus phellos L.*	568	59.9	28.2	38.0	27.1	31.4	54.3	42.7	55.9	58.5	41.9	48.5	47.7	59.9	58.2	17.9	18.0
L1	*Quercus shumardii Buckley*	1,612	71.4	28.3	36.2	26.6	34.7	64.2	48.4	62.8	69.7	54.5	62.3	63.9	71.4	70.1	21.7	26.3
L1	*Quercus stellata Wangenh.*	96	40.2	36.6	39.7	39.3	38.7	40.2	39.0	40.1	40.1	38.7	40.2	40.2	40.1	40.2	13.1	14.0
L1	*Quercus texana Buckley*	2,975	66.8	27.5	32.7	25.7	30.7	60.9	41.9	56.9	65.7	55.6	58.0	61.3	66.8	65.5	21.0	25.0
L3	*Acer rubrum L.*	100	17.6	15.4	17.3	17.3	17.1	17.6	17.1	17.5	17.6	17.4	17.5	17.6	17.5	17.6	7.6	10.9
L3	*Carya ovata (Mill.) K. Koch*	91	23.0	18.9	20.2	16.4	18.7	22.7	22.3	22.8	22.8	21.5	21.7	21.9	22.6	23.0	7.6	6.4
L3	*Nyssa sylvatica Marshall*	50	10.0	9.5	9.9	9.7	9.3	10.0	10.0	9.9	10.0	9.7	10.0	10.0	10.0	10.0	3.9	2.4
L3	*Ulmus alata Michx.*	205	36.0	29.9	31.1	25.2	27.6	35.3	32.8	35.0	35.9	35.1	35.3	35.3	35.5	35.9	18.9	20.2
L3	*Ulmus americana Planch*	185	21.9	20.8	20.7	17.4	19.5	21.8	21.6	21.8	21.9	21.3	21.6	21.7	21.6	21.9	12.3	10.7
L3	*Vaccinium stamineum L.*	75	12.6	11.1	12.5	12.3	11.9	12.5	12.3	12.6	12.6	12.4	12.6	12.5	12.6	12.6	5.7	3.5
L4	*Dirca palustris L.*	65	21.5	20.0	21.2	20.7	18.4	21.5	21.4	21.5	21.5	20.7	21.4	21.3	21.4	21.4	6.9	4.4
L4	*Fraxinus caroliniana Miller*	100	16.4	7.4	14.1	13.4	10.7	15.2	14.1	13.5	15.6	13.6	13.4	15.4	16.3	15.0	4.1	3.1
L4	*Ilex opaca Aiton*	100	37.5	27.4	37.3	36.0	35.5	37.5	34.9	37.4	37.5	36.5	37.4	37.5	37.5	37.4	11.5	20.4
L4	*Populus L.*	3,065	30.9	16.5	27.2	18.9	21.2	28.4	27.7	28.2	29.8	24.1	27.1	25.6	30.8	28.7	7.1	5.0
L4	*Salix L.*	82	14.7	9.0	13.5	9.2	8.6	14.1	13.6	13.7	14.6	11.0	14.1	12.0	14.6	14.1	3.2	1.8
T-GSH · 9 species																		
L4	*Glycine max*	600	55.1	45.5	53.3	45.0	47.9	52.9	52.3	50.7	54.4	53.9	53.5	52.4	55.1	54.7	30.5	20.8
L4	*Phaseolus vulgaris*	597	32.0	18.5	27.2	16.0	28.6	29.0	27.6	28.3	31.9	31.4	28.7	29.9	31.6	31.6	14.5	7.4
L4	*Populus balsamifera*	113	37.4	16.7	35.9	33.7	37.1	35.8	36.0	36.5	37.4	37.4	37.3	37.1	37.4	37.3	8.3	4.6
L5	*Triticum aestivum*	2,443	17.9	15.3	17.1	14.8	16.2	17.7	17.6	17.7	16.6	12.9	12.8	12.2	15.5	17.8	3.8	4.7
L5	*Elaeis guineensis*	317	36.6	35.7	34.3	34.9	35.3	34.5	35.6	35.4	36.6	34.9	36.2	33.9	35.2	36.5	12.6	7.2
L5	*Hordeum vulgare*	542	27.9	27.8	27.9	27.7	26.8	27.9	27.9	27.9	27.9	27.9	27.9	27.2	27.6	27.9	12.1	6.4
L5	*Oryza glaberrima*	253	47.2	43.6	47.1	42.6	46.4	47.0	47.1	47.1	46.8	46.9	47.0	44.5	46.2	47.2	24.9	19.3
L5	*Triticum spp.*	651	78.4	71.8	77.0	72.6	71.8	77.8	77.7	78.1	78.2	78.2	78.2	77.1	78.4	78.3	59.7	45.1
L5	*Zea mays*	1,427	69.3	61.5	69.2	66.3	68.0	69.3	69.3	69.2	68.9	69.1	68.7	67.5	67.3	69.3	45.2	53.8

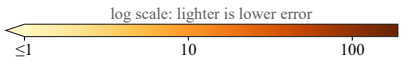


Figure S57: Per-species count MAE (stomata per image) at $\tau = 0.75$ for every detector and every species of OD-Valid, T-MAWP, T-MSWP and T-GSH. Each cell prints the value. Colour is on a log scale, lighter for lower error. N: images. $\bar{c}_{GT}$: mean annotated stomata per image. Count MAE needs no box matching.

Confidence: $\tau = 0.75$

		N	$\bar{c}_{GT}$	FRCNN-R50-1×	FRCNN-R50-3×	FRCNN-R101-1×	FRCNN-X101	Retina-R50-1×	Retina-R50-3×	Retina-R101-3×	YOLOv8	YOLOv11	YOLOv12	YOLOv26	RT-DETRv2	RT-DETRv4	GDINO-SB	SAM3-OD
T-DCWP · 39 species																		
L0	*Quercus rubra*	3	31.3	1.7	5.3	3.0	3.7	21.0	9.0	13.3	30.7	18.7	27.3	29.7	31.3	27.0	2.0	1.7
L1	*Viburnum acerifolium*	3	14.3	5.3	10.3	4.0	8.0	13.3	12.3	12.0	14.3	13.7	14.3	13.7	14.3	14.3	2.0	5.3
L1	*Viburnum dentatum*	3	18.0	1.3	10.7	2.7	13.0	15.3	15.0	15.0	18.0	17.7	17.3	16.3	17.3	18.0	0.3	1.0
L1	*Viburnum nudum*	3	8.7	6.0	8.3	4.3	6.3	8.7	8.7	8.7	8.7	8.7	8.7	8.7	8.7	8.7	0.3	0.7
L1	*Viburnum prunifolium*	3	16.0	1.0	7.0	12.3	8.0	11.3	5.7	8.3	16.0	16.0	16.0	16.0	15.3	10.7	0.7	4.3
L1	*Prunus lasiocarpa*	3	14.3	8.7	13.3	7.3	11.3	14.3	14.3	14.3	14.3	14.3	14.3	14.3	14.0	14.3	0.7	1.0
L1	*Stewartia malacodendron*	2	6.0	1.0	5.0	5.0	5.5	6.0	6.0	6.0	6.0	6.0	6.0	6.0	5.5	6.0	1.5	2.0
L2	*Camellia oleifera*	3	10.0	9.0	9.3	9.3	9.0	9.7	9.7	9.7	10.0	9.7	10.0	10.0	10.0	10.0	5.0	3.3
L2	*Amelanchier stolonifera*	3	10.3	3.3	4.3	4.0	6.7	9.7	9.3	10.3	10.3	10.3	10.3	10.0	10.0	9.7	0.7	1.0
L2	*Aronia arbutifolia*	3	31.0	2.3	19.3	7.7	23.7	20.7	27.7	29.7	31.0	28.0	26.3	25.3	31.0	27.7	0.7	2.0
L2	*Crataegus viridis*	2	19.5	8.0	15.5	10.5	18.5	19.5	19.5	19.5	19.5	19.5	19.5	19.0	19.5	19.5	1.5	1.5
L3	*Cornus alba*	3	16.7	23.0	4.0	16.0	4.3	14.3	10.3	15.7	16.3	16.3	13.0	8.0	14.7	16.3	1.7	3.7
L3	*Kalmia latifolia*	3	14.3	10.7	5.0	5.7	3.3	14.0	12.0	12.7	14.3	14.3	14.3	13.3	14.3	13.7	1.0	0.7
L3	*Sideroxylon lycioides*	3	29.3	2.0	7.0	10.7	12.3	27.3	21.7	26.7	27.3	16.3	27.7	24.7	29.3	29.0	2.7	0.7
L3	*Betula nigra*	3	12.0	21.0	2.7	9.3	10.3	9.7	9.7	11.0	12.0	12.0	12.0	12.0	12.0	12.0	7.3	11.7
L3	*Betula papyrifera*	3	10.0	4.0	8.0	2.7	8.3	10.0	10.0	10.0	10.0	10.0	10.0	10.0	9.7	10.0	2.3	0.0
L3	*Carpinus caroliniana*	3	39.3	7.7	19.0	7.0	11.0	33.3	24.3	23.0	39.3	37.7	38.3	36.7	39.3	39.0	0.7	18.0
L3	*Carya glabra*	3	28.7	6.3	12.0	23.0	20.7	28.0	27.0	28.0	28.7	24.7	27.7	27.3	28.7	28.3	0.7	1.7
L3	*Morella cerifera*	3	36.0	9.0	24.0	26.7	22.3	35.3	26.7	30.3	36.0	34.0	34.0	33.3	36.0	36.0	3.3	4.3
L3	*Lindera benzoin*	3	23.0	4.0	13.7	15.7	6.3	23.0	20.7	22.0	22.7	23.0	22.3	23.0	23.0	23.0	5.7	15.7
L3	*Persea borbonia*	3	51.0	14.0	34.0	12.0	39.3	49.3	49.0	51.0	50.7	47.7	45.7	50.0	51.0	51.0	8.0	4.7
L3	*Sassafras albidum*	3	39.3	26.0	31.0	21.7	34.0	38.3	38.7	39.3	39.0	39.0	39.0	39.0	39.3	39.3	9.7	12.7
L3	*Zanthoxylum americanum*	3	16.3	4.0	15.3	10.0	14.0	16.3	16.3	16.3	16.3	16.3	16.3	16.3	16.3	16.3	0.7	1.3
L3	*Zanthoxylum simulans*	3	7.3	7.0	6.3	3.7	7.0	7.3	7.3	7.3	7.3	7.3	7.3	7.0	7.3	7.3	0.7	0.3
L3	*Acer rubrum*	3	39.0	44.3	10.7	15.3	13.0	31.7	31.7	33.7	39.0	37.7	35.7	39.0	39.0	38.7	6.7	14.3
L3	*Aesculus parviflora*	3	50.3	16.0	13.3	15.0	20.7	42.7	44.0	48.0	49.3	46.0	46.7	43.0	50.3	50.3	23.7	28.7
L4	*Fatsia japonica*	3	16.0	3.0	15.3	7.0	13.3	15.7	14.7	16.0	16.0	16.0	16.0	16.0	16.0	15.3	1.0	1.0
L4	*Ilex cornuta*	3	30.7	2.7	25.0	20.0	22.7	28.7	26.7	30.7	30.7	30.7	30.7	30.0	30.7	28.3	0.7	2.3
L4	*Moringa oleifera*	3	17.3	2.3	2.7	1.0	2.7	11.0	7.3	9.7	17.0	15.3	11.7	13.3	17.3	16.7	4.0	2.0
L4	*Euonymus americanus*	3	21.0	1.3	5.7	11.3	5.0	17.0	6.3	9.0	21.0	18.3	21.0	18.0	21.0	16.3	2.3	1.3
L4	*Stephanotis floribunda*	5	39.0	1.6	26.8	5.6	21.6	34.6	33.0	29.6	37.8	36.4	37.4	34.4	38.8	37.0	6.0	4.0
L4	*Fraxinus pennsylvanica*	3	14.0	1.3	8.0	5.0	10.3	11.3	13.0	14.0	14.0	13.3	13.0	14.0	12.0	13.7	2.3	4.7
L4	*Olea europaea*	3	27.7	8.0	9.3	6.7	9.3	15.7	22.3	23.0	27.0	17.7	20.3	16.0	27.7	26.3	0.3	1.0
L4	*Magnolia virginiana*	3	49.3	10.3	19.0	39.3	27.7	49.3	38.0	49.3	47.0	42.3	45.7	49.0	49.3	49.3	3.0	2.3
L4	*Passiflora coriacea*	5	40.0	12.2	26.0	6.0	17.2	35.8	34.0	34.4	39.0	37.8	34.4	35.4	40.0	39.4	12.8	6.4
L4	*Hibiscus clayi*	3	11.7	0.0	1.0	2.3	1.0	11.0	6.3	7.7	11.3	5.7	7.7	9.0	11.3	10.3	1.0	0.3
L4	*Edgeworthia chrysantha*	3	12.3	9.7	10.7	10.0	9.0	11.7	11.3	12.3	12.3	12.3	12.3	12.0	12.3	12.3	6.7	11.3
L4	*Fothergilla major*	3	12.3	3.3	3.0	4.3	2.7	11.0	11.0	11.0	12.0	11.3	10.7	10.7	10.0	11.7	1.7	1.0
L4	*Itea virginica*	3	25.0	4.0	1.7	3.0	1.3	19.7	5.0	16.0	20.3	16.3	21.7	10.3	25.0	22.3	1.0	1.0

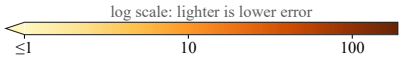


Figure S58: As Figure S57, for all 39 T-DCWP species.

S10 Qualitative visualizations

Figures S59–S61 show what the detection measures look like on single images: the predictions of every detector at the fixed threshold $\tau = 0.25$, in the order of the main tables, five per figure. Each set contributes one image, chosen by a fixed rule, not picked by hand: its scene-primary image whose annotated count is closest to the set's median, and for T-GSH the calibrated Gibbs wheat image of the L5 stress test chosen the same way. Predictions are matched to the annotations as in the evaluation. The examples follow the patterns of the tables: most stomata are matched on OD-Valid, the held-out sets mix missed stomata with extra boxes, and on the Gibbs wheat image only the two foundation-model detectors match any stoma. Single images illustrate these patterns, and the measured values are in Sections S3–S7.

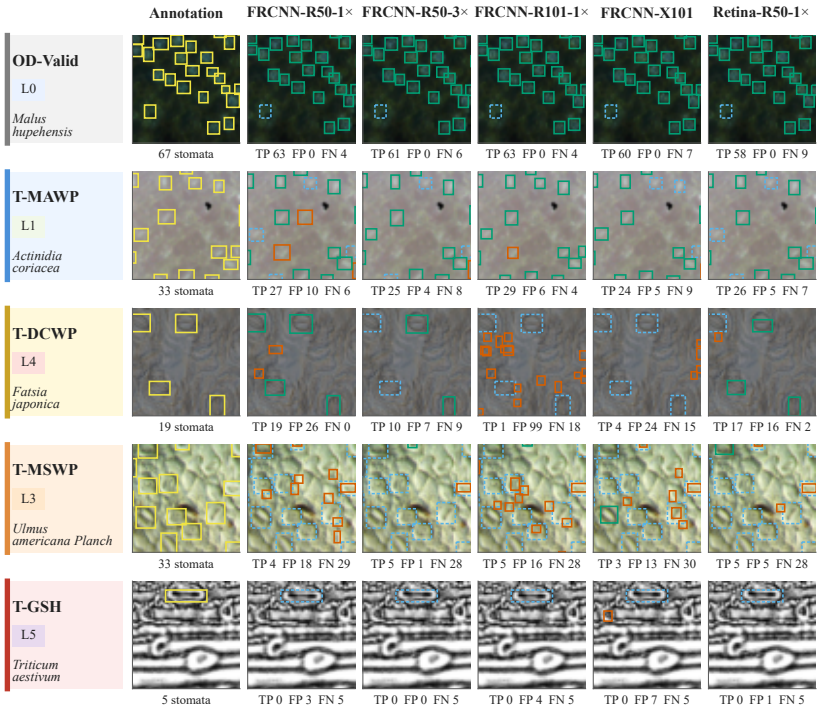


Figure S59: Predictions of FRCNN-R50-1×, FRCNN-R50-3×, FRCNN-R101-1×, FRCNN-X101, Retina-R50-1× at $\tau = 0.25$ on one image per set, next to its annotated stomata (yellow), chosen by the rule of Section S10. Matched one-to-one at IoU 0.50: true positives green, false positives orange, missed stomata blue (dashed). Each tile is a square centred on the annotated stomata, and counts are for the whole image.

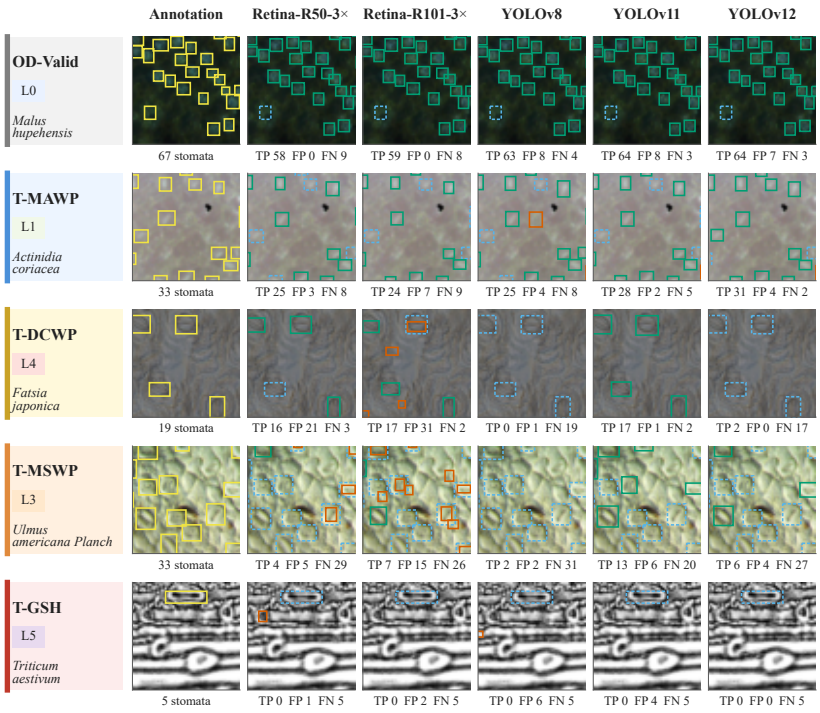


Figure S60: Predictions of Retina-R50-3×, Retina-R101-3×, YOLOv8, YOLOv11, YOLOv12 at $\tau = 0.25$ on one image per set, next to its annotated stomata (yellow), chosen by the rule of Section S10. Matched one-to-one at IoU 0.50: true positives green, false positives orange, missed stomata blue (dashed). Each tile is a square centred on the annotated stomata, and counts are for the whole image.

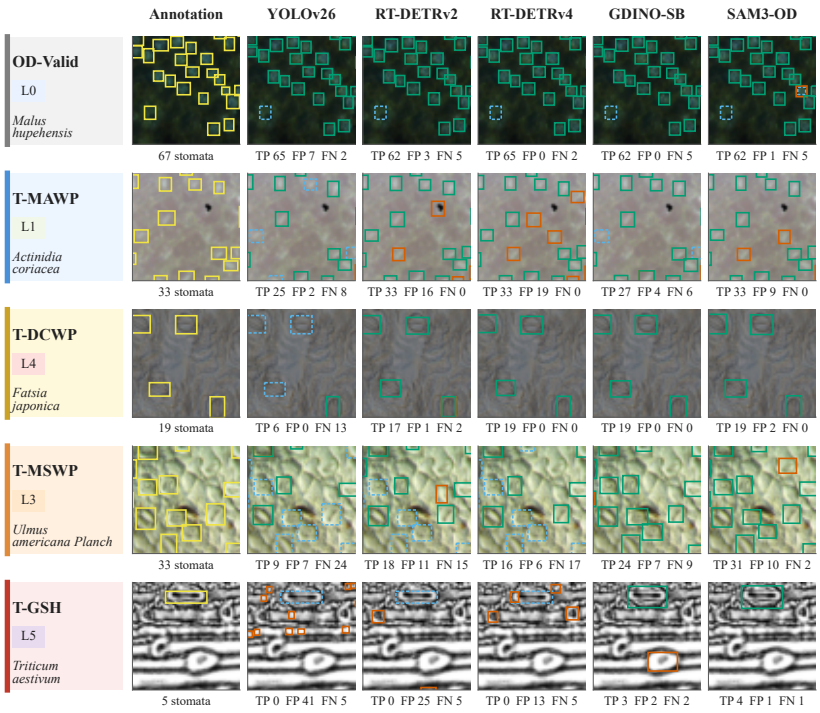


Figure S61: Predictions of YOLOv26, RT-DETRv2, RT-DETRv4, GDINO-SB, SAM3-OD at $\tau = 0.25$ on one image per set, next to its annotated stomata (yellow), chosen by the rule of Section S10. Matched one-to-one at IoU 0.50: true positives green, false positives orange, missed stomata blue (dashed). Each tile is a square centred on the annotated stomata, and counts are for the whole image.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state the contributions: the datasets, the OOD-ladder evaluation protocol, and the study of 15 fine-tuned detectors. Sections 3 to 6 support them, and Section 7 states their scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 7 covers the physical field of view, the taxonomic scope of OD-Train, low replication in some species, the fixed threshold that is not meant for deployment, duplicate detections, single-class boxes, and uncontrolled imaging conditions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper presents no theorem or formal theoretical result requiring proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 3 describes the sets, Section 4 the detectors with their pre-processing and training, and Section 5 the metrics, box matching, calibration and thresholds. Table S7 gives the training settings and Table S8 the thresholds selected on OD-Valid. The released ground-truth and prediction files and the evaluation code recompute the reported metrics without retraining. Training code is not released.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The Data availability paragraph (Section 1) links the images, annotations, metadata and per-image predictions on Hugging Face, the fine-tuned checkpoints on Hugging Face, and the evaluation code and commands on GitHub. Training code is not released.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The Datasets, Models, and Evaluation sections and Section S2 specify the splits, OD-Valid use, model configurations, confidence protocols, matching rules, calibration, aggregation, and uncertainty.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Set-level count and density MAE, and the per-species results for T-MAWP and the calibrated Gibbs wheat images, have 95% paired-image percentile bootstrap intervals from 2,000 resamples with seed 20260802 (Sections S5 and S6). The AP intervals of Table S21 use 1,000 resamples, with one seed per species (20260811 to 20260813). The intervals show the variation from sampling test images, not the variation across training runs. Other per-species values are point estimates, except the example species of Table S21.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Table S7 (Section S2) gives the type and number of GPUs used to train each detector, the GPU memory, and the training schedule in iterations or epochs. Training time and the compute of preliminary or discarded runs are not reported.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. External images are redistributed with the permission of their creators, and the data contain no people or personal information.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 8 states that detector-derived counts can support field phenotyping, that biased density estimates can mis-rank genotypes or distort drought-screening decisions, and that deployment needs target-domain calibration, independent threshold selection and human review. Section 7 states what the benchmark does not cover.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The data are plant leaf micrographs with no people or personal information, and the detectors only localize stomata. Neither poses a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Sections 3 and 4 cite the creators of every dataset and model. The authors of T-DCWP, T-MSWP and T-GSH granted permission to redistribute their images, which are released under CC-BY-NC-4.0. The license of each fine-tuned checkpoint is listed on the model card of the weights repository.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new assets are OD-Train, OD-Valid, T-MAWP, the T-DCWP annotations, the fine-tuned checkpoints and the evaluation code. Section S1 documents the species composition, accessions and calibration of each set, and Section 3 gives the imaging and annotation protocol. The Hugging Face dataset card and Croissant metadata accompany the data. Images and annotations are released under CC-BY-NC-4.0, the checkpoints under the licenses of their base models, and the code under the MIT license.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

Justification: The paper involves no crowdsourcing or research with human subjects. The first author annotated OD-Train, OD-Valid, T-MAWP and T-DCWP, and no annotators were paid.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human participants or identifiable human data are involved, so IRB review is not applicable. The study is grounded in the plant kingdom.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were used for editing, data-processing assistance, and visualization support, but not as a core model, evaluator, annotation source, or scientific decision maker. All computations were rerun from frozen source data and audited with deterministic scripts. The core methodology, scientific rigor, and originality of the research do not involve the use of LLMs.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.