bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Petal to the metal: The slow road to automating large-scale phenology labeling for herbarium specimens

Erin L. Grady[1], Raphael LaFrance[1], Daijiang Li[2], Russell Dinnage[3,4], Ellen G. Denny[5], John Deck[6] and Robert P. Guralnick[1]

1 Florida Museum of Natural History, University of Florida, Gainesville, Florida, USA

2 Department of Botany, University of Wisconsin-Madison, Madison, WI, USA

3 Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada

4 Alberta Machine Intelligence Institute, Edmonton, AB, Canada

5 USA National Phenology Network, School of Natural Resources and the Environment, University of Arizona, Tucson, Arizona, USA

6 Berkeley Natural History Museums, University of California, Berkeley, California, USA

ABSTRACT

Herbarium specimens represent critical historical records of plant phenology, yet automating annotation of reproductive structures remains challenging given the diversity of floral morphologies, specimen age and quality, and image quality. Here, we present a machine learning pipeline that uses an ensemble modeling approach to detect flowers on herbarium specimens and deliver these data to the phenology research community. After testing multiple strategies for generating training data, we found in-house expert-curated annotations were essential for producing reliable results. Expert validation found relatively strong accuracy for detecting present floral structures, but still had moderately high false negative rates. Applying the ensemble to our filtered final image dataset of 22 million records resulted in 11.1 million records labeled with flowers present. However, only 2.9 million of these contained complete metadata necessary for downstream phenology research, highlighting the need for full label digitization efforts. Still, this dataset represents a large compilation of historical herbarium-

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

derived phenology records available as a resource for the phenology community. We end by demonstrating how integrating these machine-labeled records into Phenobase, a publicly-available phenology database, expands taxonomic and temporal coverage for large-scale phenological analyses, and discuss remaining challenges and next steps.

Key Words: computer vision; data integration; herbarium specimens; machine learning; plant phenology; plant flowering

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

INTRODUCTION

Natural life cycle events, such as plant flowering, fruiting, and leaf-out, occur annually on seasonally-defined schedules shaped by environmental cues. The timing and sequence of these events, which vary depending on species and environment, underpin many biological and human processes, including species interactions (Kudo et al., 2004), carbon cycling (Richardson et al., 2013), resource acquisition (Deacy et al., 2017), agriculture (Yadav et al., 2023), tourism (Nagai et al., 2019), and allergy management (Manangan et al., 2023). Because these seasonal changes are critical and fundamentally linked with many socio-ecological processes, a whole branch of inquiry, phenology, is devoted to understanding their drivers and consequences.

A fundamental question in phenology is how different species and ecosystems have responded and will continue to respond to accelerating human-driven environmental change. The rate and direction of these phenological shifts are of particular interest as they are often considered one of the first and best indicators of broader ecological disruption (Menzel et al., 2006). For example, such shifts can interfere with key interactions that require phenological synchrony, such as pollination by interacting insects, and can have cascading consequences for species' population health, ecosystems, and people (Visser et al., 2019). Thus, access to data capturing the timing of key life stages across time, space, and taxa is increasingly of interest and importance across many scientific and applied disciplines.

Concerted phenology monitoring has been a key means to generate data needed to document phenological change. However, many regions across the globe don't have phenology monitoring programs, and in areas that do, the temporal extent is often limited. For example, the USA National Phenology Network began coalescing observational monitoring in 2007 (Crimmins et al., 2022), but national-scale phenology data before this date are limited and often of short duration (with some exceptions, see for example, Hough's efforts from 1851-1859; Guralnick et al., 2025). Beyond direct phenology monitoring, there are new resources, such as

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

iNaturalist, producing digital vouchers where phenological information can be secondarily derived (Dinnage et al., 2025). Such contemporary contributory science platforms, like iNaturalist, harness community members across the globe to generate large quantities of observations linked to images and now include means to annotate phenological states. While these resources are helping to further fill spatial and taxonomic data gaps and providing new opportunities for phenology research at scale, they are even more limited in temporal extent compared to monitoring data.

A critical component for defining and predicting phenological change is understanding phenology and its trends over decadal to century-level timescales. Herbaria, libraries that house collections of pressed plant vouchers from the field, contain specimens dating back sometimes hundreds of years and have the potential to offer valuable historical phenology data, complementing contemporary monitoring initiatives (Davis et al., 2015; Willis et al., 2017). Recent efforts to image and digitize herbarium specimens have allowed for broader use of these plant collections (Soltis 2017), with information on over 108 million angiosperm specimens (~40 million with associated images) available on the Global Biodiversity Information Facility (GBIF) as of October 2025. Ideally, each of these herbarium specimens contains valuable information about the location, collection date, species identity, and occasionally the phenological phase of that plant individual. For the vast majority of plants lacking phenology annotations, it requires significant effort to classify which, if any, phenological phases are visibly present on specimens.

Automating phenology annotation for digitized herbarium specimens is a clear next step, one that would unlock vast amounts of historical data and enable phenological change research across diverse taxa and environments. Recent machine learning efforts to classify phenological phases in photos of live plants have been successful (Dinnage et al., 2025); however, herbarium specimens present a unique set of challenges. Dried and pressed plant specimens often lose color and structure over time, reproductive structures can be lost during processing, some taxa have structures that are undetectable from digitized images alone or without

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

dissection, and specimens contain non-plant material such as labels and herbarium stamps that may interfere with machine detection (Pearson et al., 2020). Additionally, accuracy of machine phenology annotation is highly dependent on the annotator who creates training data used to calibrate models, and curating a reliable and taxonomically-diverse training dataset requires extensive effort and careful consideration.

Several works have applied machine learning to herbarium specimens, with greater success in species identification than phenological recognition tasks (Hussein et al., 2022). Phenological efforts have often focused on small, curated species and image sets (Goëau et al., 2020; Younis et al., 2020) and many have reported challenges annotating specific phenological structures such as flowers (Lorieul et al., 2019; Younis et al., 2020). More recent efforts have successfully addressed specific phenology research questions using models trained by compositing many datasets from multiple different sources (Williamson et al., 2025). Our aim here, however, was not to produce a standalone research product or answer a particular question, but to develop a trustable, usable, and integratable data resource for the broader research community. Doing so required careful curation of our own training data, which allowed us to leverage information gathered during training to reduce noise in the final data product, as we discuss belowt.

Here, we present an automated phenology annotation pipeline that accurately detects flowers present on digitized herbarium specimens. With the goal of maximizing geographic and taxonomic coverage across the angiosperm tree of life, this effort unlocks vast quantities of previously limited historical phenology information. We outline the challenges, successes, complexities and strategic decisions involved in developing the pipeline, emphasizing the need for high-quality training data generated by botanically-knowledgeable individuals and rigorous filtering on multiple dimensions both pre- and post-modeling. We discuss the trade off between strict data filtering for improved model accuracy and data retention to maximize the volume of downstream historical records annotated with phenology information. We highlight a surprising

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

and disappointing discovery that the majority of specimens lack metadata sufficient for phenological annotation, which we discovered after all the rest of the effort. Finally, we demonstrate the vast scale of previously-unavailable historical phenology data that are now available to stakeholders, and discuss the integration of these data into Phenobase (https://phenobase.netlify.app/), an online database of millions of research-ready, community-sourced phenology data from field images and from structured in situ phenology monitoring programs.

METHODS

Gathering specimen image and metadata

We opted to gather data for this work from GBIF, which provides access to billions of species occurrence records and, in many cases, associated images. We searched GBIF in mid-2023 for all “Magnoliopsida” and “Liliopsida”, using filters to select the basis of record as “preserved specimen” and media type as “images”. This produced 33,691,939 occurrence records with 35,314,854 corresponding multimedia records. We did not filter out records missing coordinates or location information so that when metadata of a record is updated in the future, its associated phenological information will have already been extracted. We downloaded this dataset and associated images in October 2024 and, using a nearest pixel filter, programmatically resized each image to 1024 pixels along the shorter edge to reduce storage and satisfy GPU memory limitations. Because services are often down and media links can change, the number of images downloaded is a subset of the total number of records in the original GBIF dataset. In sum, after contending with link issues, removing small or low quality images, and images associated with taxa that we decided to filter out (described below), we collected 22,447,563 images for downstream labeling and model training.

Labeling training data

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

A challenging aspect of developing a reliable automated pipeline for phenophase annotation on herbarium specimens across the angiosperm tree of life was generating enough high-quality training data. Our goal was to produce a trusted historical phenology data resource for broad use, and we therefore deliberately chose to manually generate our own training data rather than utilizing data from multiple external sources. As mentioned above, manually scoring pressed plant specimens for phenophases is time consuming and difficult, and a key challenge is balancing the need for botanical expertise with the capacity of expert annotators. Other initiatives have employed volunteers, undergraduate students, and paid non-experts with relative success (Brenskelle et al., 2020; Zhou et al., 2018), however these focused on a few key taxa with easily-distinguishable structures. Limiting annotations to a few taxa at the genus or family level allows for more in-depth, targeted training and familiarity with target taxa which likely lead to better results. We were unsure whether employing non-experts would be successful when scoring thousands of diverse taxa.

We first tested whether a paid undergraduate student could produce accurate phenophase annotations across all flowering plants. We employed a student to score a random selection of specimens for presence or absence of open flowers. If the student was uncertain or unable to make an informed decision about the digitized image, they had the option to mark “unknown”. We then also had a botanically-knowledgeable individual validate these annotations to assess their accuracy. As reported in Results, we found the undergraduate’s open flower annotations were not at a level usable for training a model. Based on these results and other work that suggests that one of the most important variables in annotation accuracy is simply the annotator themselves (Brenskelle et al., 2020), we made the conscious choice to employ one in-house botanically-knowledgeable individual to generate all training data. This decision greatly reduced our capacity to generate high quantities of training data, but ensured high-quality and consistency, which was our priority.

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

In addition to the first set of specimens annotated to validate the undergraduate annotations, we annotated a second additional set of specimens for open flowers to increase training data. While compiling the second set, we randomly selected specimens in families underrepresented in the first set in an effort to promote taxonomic evenness. Additional records were added to the training data iteratively through repeated bouts of image validation throughout model training.

Filtering training data

To ensure accurate and reliable results, we opted to filter out taxa pre-model-training that were challenging to manually annotate. First, we removed genera that had more than 5 records in the training dataset and an “unknown” annotation rate greater than or equal to 25%. After removing these genera, we removed families using the same criteria. Many of these removed taxa were those with small reproductive structures, such as Chenopodiaceae, or those with flowers and fruits that are difficult to distinguish from each other in images alone, like Poaceae. These taxa were removed permanently from our pipeline, both from training and downstream data. We made this choice strategically to balance data quality and quantity. Throughout this process, we found it critical to acknowledge and respond to limitations, both in human expertise, machine capabilities, and overall capacity. Final lists of removed genera and families are available in Supporting Information.

Model training and post-training thresholding

Model Ensembling Approach

One challenge with training models on all herbarium data is the vast and complex floral diversity across the whole of the group, challenging models to properly abstract features that can apply broadly and that are not easily confusable with other objects. We realized single models often performed modestly well, but that ensembled models might perform better than any single one

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

given this variability. We tried several different neural networks, architectures, and model parameters and ultimately determined that an ensemble modeling approach, combining three models, was most effective based on downstream validation statistics. Note that all three models were fine tuned versions of models that were previously pretrained on ImageNet data (Deng et al., 2009).

Despite differences across the three models (Effnet, ViT_{MSE} and VIT_{CE}, summarized in Table 1), all training followed the same general steps. First, expert-labeled records were split into 60% training, 20% testing, and 20% validation sets. After low-quality images were filtered out, the images were resized to match the size requirements of the specific model (Table 1). Images in the training dataset were augmented using random horizontal and vertical flips and "auto augment" (Cubuk et al., 2018), which uses multiple techniques including random rotation, color jitter, sharpness adjustment, and image shear. Next, all models were color adjusted to use the ImageNet color mean and standard deviation. During training, each model was saved as to five checkpoints The final best model was chosen as the checkpoint with the best F1 score on the holdout validation dataset. The F1 score is the harmonic mean of precision (how many positive model predictions are actually positive?) and recall (how many actual positives did the model detect?). We chose the F1 score because it is particularly useful for assessing predictive performance for the positive class ("flowers present"), which is the focus of this effort. This overall approach is standard for model selection and ensures comparable scoring across model approaches (e.g. EffNet CNN versus ViT) and that models are tuned, selected, and ensembled based on their best validation performance.

We ran models using a single A100 GPU on the University of Florida's HiPerGator. Each model was trained for at least 200 epochs, and we saved five checkpoints during training. At the end of every epoch, we evaluated model performance on a held-out validation set that was never used for backpropagation, which we used to assure model generalization and select the best checkpoint. We used several metrics for evaluating the model, including F1 score,

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

accuracy, validation set loss, precision, and recall. After training, each checkpoint from all three models was post-processed to determine the probability thresholds that maximized accuracy. We divided the range [0.0–1.0] into three regions defined by two thresholds. Images with a score above the high threshold were considered a positive indication of the trait, images with scores below the low threshold were considered a negative indication of the trait. Images with scores between the high and low thresholds were classified as “equivocal”, meaning the model could not decide if the trait was present or absent. We adjusted those two thresholds until the images in the test (holdout) dataset accuracy was at the maximum. We also guarded against equivocal records being more than 30% of the test dataset. We used a batch size of 32, the initial learning rate was 5e-5 with a weight decay parameter of 0.01, and an AdamW optimizer.

For each image, each model generated a final outcome of positive, negative, or equivocal. To ensemble our three models, we used a two-thirds “majority rule” approach, focusing in particular on predictions of “flowers present” (e.g., where at least two models of the three predicted present with high certainty). Since herbarium specimens often contain only a portion of a plant individual and absence on the whole plant cannot be inferred, we focused on the accuracy of “present” labels and only retained images machine-labeled as “present” in downstream data. All other cases were scored “flowers not present” or “equivocal”. Note that the ensemble itself can be undecided (equivocal), for instance with [positive, negative, equivocal] or two or more equivocal votes from the three model outputs.

Model validation

We validated ensemble results both using held-out test data and through additional expert validation. For the expert validation, we machine-labeled a random subset of images and validated 600 specimens machine-labeled as “present” for flowers and 300 specimens machine-labeled as “absent” for flowers. Consistent with the Plant Phenology Ontology (Stucky et al., 2018), “flowers” here refers to any floral structure, including both flower buds and open flowers.

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Although the models were originally trained on “open flowers”, we opted to conduct validation at the broader “flower” level, as these tasks are difficult and we felt “flower” better reflected the features detected by the ensemble models. To further understand instances where the model consensus was incorrect, we anecdotally looked at false positives and false negatives to understand general patterns in the mistakes (see Results).

Machine labeling and integration with other sources

After model validation, the remaining unlabeled downloaded images were passed through the ensemble model pipeline. We retained only images labeled as “present” by at least two out of three models and data fields were standardized for integration in Phenobase, which has a defined set of required and optional fields. As we discuss below, we dropped records that we labelled but were missing these key fields that are required for phenology research. Key fields here include: scientific name, event date, and decimal latitude and longitude. The remaining data were made publicly available through the Phenobase web portal (https://phenobase.netlify.app/).

Evaluating specimens missing critical fields

A key challenge and major limitation during this process was that a large number of specimens did not have date or location information included in their metadata. As these two pieces of information are critical for meaningful phenology data, this resulted in the loss of substantial data. To further investigate this issue, we randomly selected 100 specimens labeled as “flowers present” that were missing both date and location information. We manually annotated these specimens for whether or not they included specific (day, month and year) date information and georeferencable locality information on the specimen label.

RESULTS

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

268 **Labeling training data**

269 We worked with an undergraduate student to annotate a total of 3,525 specimens for open
270 flowers, providing them with resources for the task and asking them to use online resources
271 when needed. When validating these annotations, the undergraduate and one of our team
272 members (lead author EG), who served as a botanical expert, were in agreement 69% of the
273 time. Given this result, we chose to not include the undergraduate annotations in our final
274 training dataset. EG then generated a total of 9,004 annotations from individual specimens for
275 open flowers. After removing 49 genera and an additional seven families (Appendix S1; see
276 Supporting Information with this article), 1660 genera and 261 families remained. Our final
277 training dataset contained 7,205 specimen records, 4319 annotated as open flowers present
278 and 2886 annotated as open flowers absent. Once expert-labeled records were split, 4324
279 images were included in the training set, 1442 in the testing set, and 1439 in the validation set.
280
281 **Model fitting and validation**

282 We utilized three models (Table 1) and ensembled their results to predict the presence of
283 flowers. Our training model accuracy scores across all three models varied between 88.3% and
284 92.5%. In general, the vision transformers models (ViT) performed better (92.2% and 92.5%
285 accuracy) than the EffNet model (88.3% success). Overall, from the full ensemble, validation on
286 the held-out test data found 96.3% accuracy for detecting the presence of flowers and 86.8%
287 accuracy for detecting the absence of flowers (Appendix S2). However, expert validation on
288 records not in the initial training, testing, or validation processes found the ensemble models to
289 be 90.5% accurate for detecting the presence of flowers and 52.3% accurate for the absence of
290 flowers (Table 2). The low accuracy for detecting absences was expected and likely due both to
291 expert-validated data being more diverse than training data and that we were less concerned
292 with training for absences in our overall process. Further investigation of the false positives from
293 the expert validation found that the majority of false positives were fruits mistaken for flowers,

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

294 although some sheets had no reproductive structures present and one sheet had a fruit
295 illustration (Figure 1). Investigation into false negatives from the expert validation found a variety
296 of reasons for the mistakes, many explainable, including situations where flowers weren't
297 attached to the main plant, flowers that may have been confused as fruits, and flowers that were
298 very challenging to see (Figure 2).
299
300 **Machine labeling and integration with other sources**
301 After downloading and filtering out problematic images (images that could not be downloaded,
302 images that cannot be read, or images that were either less than 10 KB or greater than 32 MB)
303 and images of hard-to-annotate genera and families (Appendix S1), we were left with
304 22,447,563 images of digitized herbarium specimens. These images were fed through the
305 ensemble modeling pipeline and annotated for presence or absence of flowers. Of these
306 records, 11,131,820 were labeled as flowers present (meaning at least ⅔ models scored as
307 present), but only 2,913,353 of these contained valid latitude, longitude, and date information
308 (Figure 3). 5,748,788 images were scored as flowers absent (meaning at least ⅔ models scored
309 as absent), and 5,566,955 were equivocal (meaning there was no consensus among the
310 models). Only records scored as flowers present and with the needed metadata were retained
311 and integrated into Phenobase (https://phenobase.netlify.app/).
312
313 **Data coverage and evaluating specimens missing critical information**
314 The 2,913,35 records annotated as flowers present represented 383 plant families and 8,883
315 genera (Figure 4). They represent 320 years of collecting, with the earliest record dated in 1501
316 (Figure 5). We evaluated a small subset of the 8,218,467 records that were missing locality and
317 date information to determine exactly what fields were missing and if those issues could be
318 resolved with more complete digitization of labels. Out of the 100 random specimens that were
319 missing date and latitude/longitude information, 65 had a date (specific day and year) on the

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

specimen label while 35 did not; 85 had a locality that is georeferenceable and localizable to a unit lower than country, while 15 did not. A total of 57 out of 100 had both a georeferenceable specific locality and date information on the label. Only 44 of the 100 records had images retrievable from the original URL, while for 56 records the image URL failed.

DISCUSSION

While machine-automated annotation is a clear next step in addressing bottlenecks and delivering large quantities of valuable data to the research community, some annotation tasks are harder than others. Further, tasks that are relatively straightforward in one context, such as annotating flowers in photographs of live plants, can be much more challenging in other contexts, such as attempting to do the same for dried, pressed, and aged plants on herbarium sheets. Other groups have produced herbarium data using machine labeling approaches to answer their own research questions about machine learning, certain taxa, or regions, and many showcase reasonably high precision and recall (Goëau et al., 2020; Lorieul et al., 2019; Williamson et al., 2025; Younis et al., 2020). Here, however, our focus was on creating a reliable resource at the broadest possible scale and for the phenology research community as a whole; we were not focused on answering specific research questions. Our first priority was that stakeholders could trust, understand, and easily use these data. A key challenge for us was to develop a set of solutions that we felt met a reasonable bar for trustworthiness, which lead to rigorous data filtering, and to clearly communicate where these solutions may fall short. Here we are not wholly concerned with the size of the data product or any other metric of success. Rather, creating trustworthy data in this context meant carefully vetting the quality of training data, evaluating validation statistics, and being thoughtful about the data structure and documentation made available to stakeholders so they can best use these data and understand their value and limitations.

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

We also caution that the most appealing solutions for annotating flowers may be difficult at the taxonomic, spatial and temporal scale we attempted here. Our initial hope when generating training data was that we could produce flower annotations utilizing a team of people, likely undergraduate students, with limited botanical training. While this approach might work when focusing on a small subset of plants, such as a genus or family, when scaling to all angiosperms we found it untenable. A vast, complicated diversity of inflorescence morphologies interacts with herbarium sheet taphonomic processes to make it especially challenging to detect flower presence, even for an expert botanist. We argue that the ideal solution for creating high quality annotations is delegating annotation tasks to botanists working on taxonomic groups where they have expertise, but this is also impractical given capacity and logistical constraints. Our solution here was using one expert with ample practice scoring sheets, which has both benefits and drawbacks. One major benefit is that there is likely more consistency and higher quality with this method than a crowdsourced solution. This is especially important in dealing with the many cases where flower presence was uncertain, which was critical for removing certain families and genera from downstream annotation labeling. However, a major drawback in using only one annotator to generate training data is that this method is much less scalable, generating significantly fewer annotated records to use for training. Finally, more focused expertise in certain taxonomic groups may have produced less “unknowns” and higher quality training data for those groups that could lead to inclusion in downstream modeling.

A key step in our process was making thoughtful, informed decisions about removing taxa that we felt were too challenging to score correctly. While annotating training data, the annotator could choose “present” or “absent” options to inform the models, but critically they also had an “unknown” option which they were encouraged to select whenever they could not make a confident determination. Ultimately, the proportion of “unknown” annotations, for individual genera and also families, was what we used to determine which taxa were simply too challenging to score consistently for flower presence. There is likely no perfect cutoff, but we

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

371 decided that taxa where more than 25% of records were too difficult to label were likely to be
372 problematic for modeling. This decision represents a tradeoff between data quantity and quality.
373 While it is always disappointing to exclude data, we prioritized maintaining stakeholder trust by
374 favoring reliability over volume. We maintain that when attempting to scale challenging tasks
375 such as phenological annotation of herbarium specimens, it is critical to acknowledge both
376 human and technological limits and adjust methods accordingly.

377 Even with substantial taxonomic filtering, model performance was middling on both held-
378 out test data and expert validated data. Given the challenges with fitting a single model, we
379 opted for an ensemble approach, which gave us more power to interrogate multiple models and
380 arrive at a consensus. Ensembling is particularly useful because it reduces idiosyncratic errors
381 from any single model. We required a two-thirds majority to score presence from the 3 separate
382 models. This strategy required more up-front effort, but ultimately produced a more rigorous
383 labeling workflow. As in other approaches we have published (Dinnage et al., 2025), our interest
384 is in phenophase presence, and the ensembled models are fairly good at avoiding false
385 positives, but we note that the model still produced a substantial number of false absences.
386 Since we explicitly tuned the models to minimize false positives at the cost of increasing false
387 negatives, it means we are likely throwing out many images that have flowers.

388 The filtering process didn't end after we completed machine labeling. Out of 11 million
389 records labeled as flowers present, we found that more than 70% lacked a reported
390 latitude/longitude and/or a date. This is consistent with the overall proportion of specimen
391 records on GBIF that have missing coordinates or spatial issues. When we queried GBIF on
392 December 16, 2025, for example, there are 52,254,706 records with associated specimen
393 images for Magnoliopsida and Liliopsida. However, only about 32% of them (16,786,460) have
394 latitude and longitude coordinates and without geospatial issues. This data incompleteness
395 percent is disappointingly high and effectively limits the utility of these data for research, as date
396 and location information is generally required for these data to be research ready (sensu Soltis

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

et al., 2018). The root cause of this issue appears to be due to the publication of “stub records” to GBIF. These stub records may have taxon names but often lack dates and/or georeferences and they generally are from very large collections. For example, the Museum National d'Histoire Naturelle de Paris published 5,031,777 angiosperm records with images and this provider was one where records were often found to be lacking date and geospatial data. Further, we should note that many of the images we annotated are no longer available on GBIF. Out of the 100 we checked, more than half no longer had resolvable images online; most of those also came from publishers who often supply a large corpus of images. These significant issues bring home the need for efforts to expedite full label digitization to avoid these issues of missing metadata since many labels do contain dates and can be georeferenced. Such efforts are needed to support the best use of primary and secondary data (sensu Soberón and Peterson, 2004), such as phenology annotations generated from images. Creating more persistence for images tied to specimens will also better ensure reproducibility and stability when relating specimen images, core specimen metadata and secondary data derived from those images.

Despite setbacks, our ensemble modeling flower annotation pipeline unlocks vast new global phenology data at unprecedented temporal scales. Although in situ phenology data collection networks (Crimmins et al., 2022) and other initiatives to unlock phenology information from sources like iNaturalist (Dinnage et al., 2025) are providing large quantities of contemporary phenology data, data predating these initiatives are less common. They are often in the form of tabular reports that were monitored at one site over many years, although some larger citizen science projects covered multiple sites and years such as the New York Phenology Project (Fuccillo Battle et al., 2022) and Eastern USA Phenology Project (Guralnick et al., 2024) in the middle and late 1800’s. The almost 3 million labeled records that we present here represent a large dataset of historical (pre-2000) phenology now accessible for the research community. However, we note that significant gaps remain in phenology data

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

coverage across space and time, and that continued digitization and efforts to add missing metadata to existing digitized records will be essential for continuing to close these gaps.

All records labeled as flowers present from this effort are available through Phenobase, an open-access datastore of plant phenology information that integrates in situ data from phenology monitoring programs, machine-labeled data from iNaturalist images, and now machine-labeled data from herbarium specimens. Our pipeline produces phenology labels and fields consistent with the Phenobase data structure and Plant Phenology Ontology (Stucky et al., 2018), and uses a controlled vocabulary for other fields, ensuring that data users can search for phenology records at any level of granularity and get consistent results. Future work may include efforts to label herbarium specimens for fruits, however our early explorations found this task to be even more challenging than labeling for flowers, likely requiring greater human effort, and yet more rigorous filtering steps. In sum, despite the challenges, providing information on historical flowering dates across the globe and integrating with other data sources enables new analyses and insights spanning regions and taxa, enabling yet broader and more synthetic questions to be asked across temporal, taxonomic and spatial scales.

Acknowledgements

The authors would like to thank the many individuals who have contributed to collecting, processing, digitizing, and maintaining herbarium specimens, as this effort would not be possible without their dedication. Phenobase is a team effort and the authors acknowledge team members Carrie Seltzer, Ramona Walls, and Taran Lichtenberger. UFIT Research Computing provided computational resources and support. This work was funded by the National Science Foundation (DBI2223512 to Robert Guralnick and DBI2223508 to Daijiang Li).

Author Contributions

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Robert Guralnick, Daijiang Li and Raphael LaFrance initially conceived of this effort. Erin Grady operationalized most aspects of gathering training data, in consultation with Raphael LaFrance and Robert Guralnick. Raphael LaFrance developed the modeling framework for this effort with help from Russell Dinnage. Raphael LaFrance produced final models and descriptive statistics and Erin Grady led expert model validation. Ellen Denny and John Deck helped ensure that data in the right format could be ingested into Phenobase. Erin Grady and Robert Guralnick archived data. Erin Grady and Robert Guralnick wrote the manuscript with help from Raphael LaFrance, Russell Dinnage and Daijiang Li. All authors contributed to drafts and gave final approval for publication.

Data Availability Statement

The ensemble data models and a corresponding JSON file with model metadata data are housed on Zenodo (https://doi.org/10.5281/zenodo.17079402). Images used in training, validation, and testing are located here: https://zenodo.org/records/17675089. Code used for this project can be found on github (https://github.com/rafelafrance/phenobase/tree/v1.0.0). Training data and final ensemble output can be found on Zenodo (https://doi.org/10.5281/zenodo.17675089).

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Appendix S1. List of difficult-to-annotate genera and families removed from training and downstream data.

Appendix S2. Table S1. Validation results for held-out test data.

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

REFERENCES

Brenskelle, L., Guralnick, R.P., Denslow, M., and Stucky, B.J. 2020. Maximizing human effort for analyzing scientific images: A case study using digitized herbarium sheets. 2020. *Applications in Plant Sciences* 8(6): e11370. https://doi.org/10.1002/aps3.11370

Crimmins, T., Denny, E., Posthumus, E., Rosemartin, A., Croll, R., Montano, M., and Panci, H. 2022. Science and Management Advancements Made Possible by the USA National Phenology Network's Nature's Notebook Platform. *BioScience* 72(9): 908–920. https://doi.org/10.1093/biosci/biac061

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. 2018. AutoAugment: Learning augmentation policies from data. In arXiv [cs.CV]. arXiv. https://doi.org/10.48550/arXiv.1805.09501

Davis, C.C., Willis, C.G., Connolly, B., Kelly, C. and Ellison, A.M. 2015. Herbarium records are reliable sources of phenological change driven by climate and provide novel insights into species' phenological cueing mechanisms. *American Journal of Botany* 102: 1599-1609. https://doi.org/10.3732/ajb.1500237

Deacy, W.W., Armstrong, J.B., Leacock, W.B., Robbins, C.T., Gustine, D.D., Ward, E.J., Erlenbach, J.A. and Stanford, J.A. 2017. Phenological synchronization disrupts trophic interactions between Kodiak brown bears and salmon. *Proceedings of the National Academy of Sciences of the United States of America* 114 (39): 10432-10437. https://doi.org/10.1073/pnas.1705248114

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

496 Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. 2009. ImageNet: A large-scale
497 hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern*
498 *Recognition*: 248–255.
499
500 Dinnage, R., Grady, E., Neal, N., Deck, J., Denny, E., Walls, R., Seltzer, C. et al. 2025.
501 PhenoVision: A framework for automating and delivering research-ready plant
502 phenology data from field images. Methods in Ecology and Evolution 16(8): 1763-1780.
503
504 Fuccillo Battle, K., Duhon, A., Vispo, C. R., Crimmins, T. M., Rosenstiel, T. N., Armstrong-
505 Davies, L. L., and de Rivera, C. E. 2022. Citizen science across two centuries reveals
506 phenological change among plant species and functional groups in the Northeastern US.
507 *The Journal of Ecology* 110(8): 1757–1774. https://doi.org/10.1111/1365-2745.13926
508
509 Goëau, H., A. Mora-Fallas, J. Champ, N. L. R. Love, S. J. Mazer, E. Mata-Montero, A. Joly, and
510 P. Bonnet. 2020. A new fine-grained method for automated visual analysis of herbarium
511 specimens: A case study for phenological data extraction. *Applications in Plant Sciences*
512 8(6): e11368. https://doi.org/10.1002/aps3.11368
513
514 Guralnick, R., Crimmins, T., Grady, E., and Campbell, L. Phenological response to climatic
515 change depends on spring warming velocity. 2024. *Communications Earth &*
516 *Environment* 5, 634. https://doi.org/10.1038/s43247-024-01807-8
517
518 Hussein, B.R., Malik, O.A., Ong, W.H., and Slik, J.W.F. 2022. Applications of computer vision
519 and machine learning techniques for digitized herbarium specimens: A systematic
520 literature review. *Ecological Informatics* 69, 101641.
521 https://doi.org/10.1016/j.ecoinf.2022.101641

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Kudo, G., Nishikawa, Y., Kasagi, T. and Kosuge, S. 2004. Does seed production of spring ephemerals decrease when spring comes early? *Ecological Research* 19, 255–259. https://doi.org/10.1111/j.1440-1703.2003.00630.x

Lorieul, T., Pearson, K. D., Ellwood, E. R., Goëau, H., Molino, J.-F., Sweeney, P. W., Yost, J. M., et al. 2019. Toward a large-scale and deep phenological stage annotation of herbarium specimens: Case studies from temperate, tropical, and equatorial floras. *Applications in Plant Sciences* 7(3): e1233. https://doi.org/10.1002/aps3.1233

Manangan, A., Brown, C., Saha, S., Bell, J., Hess, J., Uejio, C., Fineman, S., and Schramm, P. 2021. Long-term pollen trends and associations between pollen phenology and seasonal climate in Atlanta, Georgia (1992-2018). *Annals of allergy, asthma & immunology : official publication of the American College of Allergy, Asthma, & Immunology*, *127*(4), 471–480.e4. https://doi.org/10.1016/j.anai.2021.07.012

Menzel, A., Sparks, T.H., Estrella, N., Koch, E., Aasa, A., Ahas, R., Alm-Kubler et al. 2006. European phenological response to climate change matches the warming pattern. *Global Change Biology,* 12: 1969-1976. https://doi.org/10.1111/j.1365-2486.2006.01193.x

Nagai, S., Saitoh, T.M. and Yoshitake, S. 2019. Cultural ecosystem services provided by flowering of cherry trees under climate change: a case study of the relationship between the periods of flowering and festivals. *International Journal of Biometeorol*ogy 63, 1051–1058. https://doi.org/10.1007/s00484-019-01719-9

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

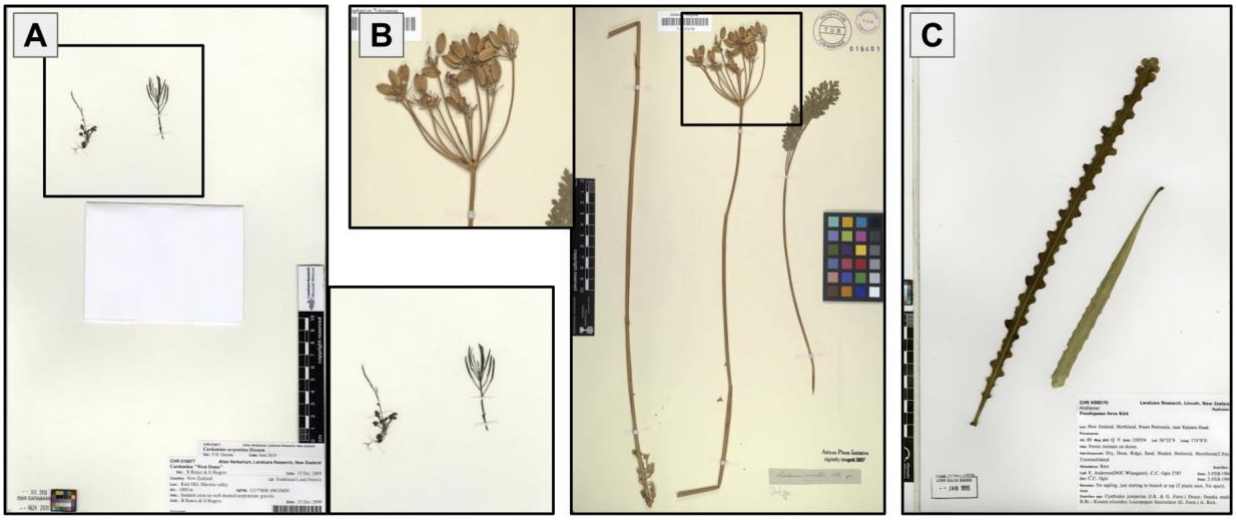
548 Pearson, K.D., Nelson, G., Aronson, M.F.J., Bonnet, P., Brenskelle, L., Davis, C.C., Denny,
549 E.G. et al. 2020. Machine Learning Using Digitized Herbarium Specimens to Advance
550 Phenological Research, *BioScience* 70(7), 610–620.
551 https://doi.org/10.1093/biosci/biaa044
552
553 Richardson, A.D., Keenan, T.F., Migliavacca, M., Ryu, Y., Sonnentag, O., Toomey, M. 2013.
554 Climate change, phenology, and phenological control of vegetation feedbacks to the
555 climate system. *Agricultural and Forest Meteorology*, 169, 156–173.
556 https://doi.org/10.1016/j.agrformet.2012.09.012
557
558 Soberón, J. and Peterson, A.T. 2004. Biodiversity informatics: managing and applying primary
559 biodiversity data. *Philosophical Transactions of the Royal Society of London*, Series B
560 Biological Sciences 359(1444):689-698. https://doi.org/10.1098/rstb.2003.1439.
561
562 Soltis, P.S. 2017. Digitization of herbaria enables novel research. *American Journal of Botany*
563 104: 1281-1284. https://doi.org/10.3732/ajb.1700281
564
565 Soltis, P. S., G. Nelson, G., and S.A.James. 2018. Green digitization: Online botanical
566 collections data answering real-world questions. *Applications in Plant Sciences 6*(2),
567 e1028. https://doi.org/10.1002/aps3.1028
568
569 Stucky, B. J., Guralnick, R., Deck, J., Denny, E. G., Bolmgren, K., and Walls, R. 2018. The plant
570 phenology ontology: A new informatics resource for large-scale integration of plant
571 phenology data. Frontiers in Plant Science 9: 517.
572 https://doi.org/10.3389/fpls.2018.00517.
573

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

574 Visser, M.E. and Gienapp, P. 2019. Evolutionary and demographic consequences of
575 phenological mismatches. *Nature Ecology & Evolution* 3, 879–885.
576 https://doi.org/10.1038/s41559-019-0880-8
577
578 Williamson, D.R., Prestø, T., Westergaard, K.B., Trascau, B.M., Vange, V., Hassel, K., Koch, W.
579 and Speed, J.D.M. 2025. Long-term trends in global flowering phenology. *New*
580 *Phytologist*. https://doi.org/10.1111/nph.70139
581
582 Willis, C.G., Ellwood, E.R., Primack, R.B., Davis, C.C., Pearson, K.D., Gallinat, A.S., Yost, J.M.
583 et al. 2017. Old Plants, New Tricks: Phenological Research Using Herbarium
584 Specimens. *Trends in Ecology & Evolution* 32(7):531-546. 10.1016/j.tree.2017.03.015
585
586 Yadav, S., Korat, J. R., Yadav, S., Mondal, K., Kumar, A., Homeshvari, and Kumar, S. 2023.
587 Impacts of Climate Change on Fruit Crops: A Comprehensive Review of Physiological,
588 Phenological, and Pest-Related Responses. *International Journal of Environment and*
589 *Climate Change* 13 (11):363–371. https://doi.org/10.9734/ijecc/2023/v13i113179.
590
591 Younis, S., Schmidt, M., Weiland, C., Dressler, S., Seeger, B., Hickler, T. 2020. Detection and
592 annotation of plant organs from digitised herbarium scans using deep learning.
593 *Biodiversity Data Journal* 8: e57090. https://doi.org/10.3897/BDJ.8.e57090
594
595 Zhou, N., Siegel, Z.D., Zarecor, S., Lee, N., Campbell, D.A., Andorf, C.M., Nettleton, D. et al.
596 2018. Crowdsourcing image analysis for plant phenomics to generate ground truth data
597 for machine learning. PLoS Computational Biology 14(7):e1006337.
598 https://doi.org/10.1371/journal.pcbi.1006337
599

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

600 **Figures**



601

602 **Figure 1.** Examples of false positives, situations where the ensemble model consensus scored
603 the image as “flowers present”, but there are no flowers present in the image. A. and B. are
604 examples where fruits were confused for flowers. This was the most common cause of false
605 positives, accounting for over half of cases. C. is an example where leaves were confused as
606 flowers. Around 30% of false positives were situations where vegetative (non-reproductive)
607 structures were confused as flowers.

608

609

610

611

612

613

614

615

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

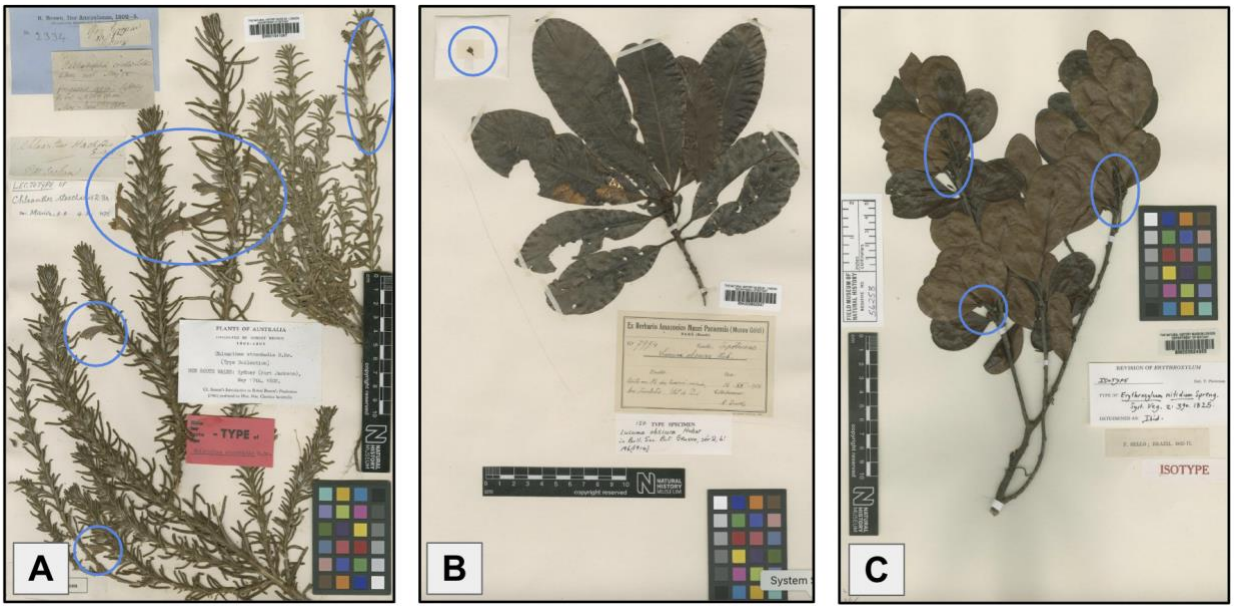


Figure 2. Examples of false negatives, situations where the ensemble model consensus scored the image as “flowers absent”, but there are flowers present in the image. A. is an example of a situation where there are obvious flowers on the sheet and the machine simply missed them. B. is an example of a situation where the floral structure is not connected to the plant. This often caused the machine to miss the flowers in the image. C. is an example where flowers are present, but just really challenging to see.

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

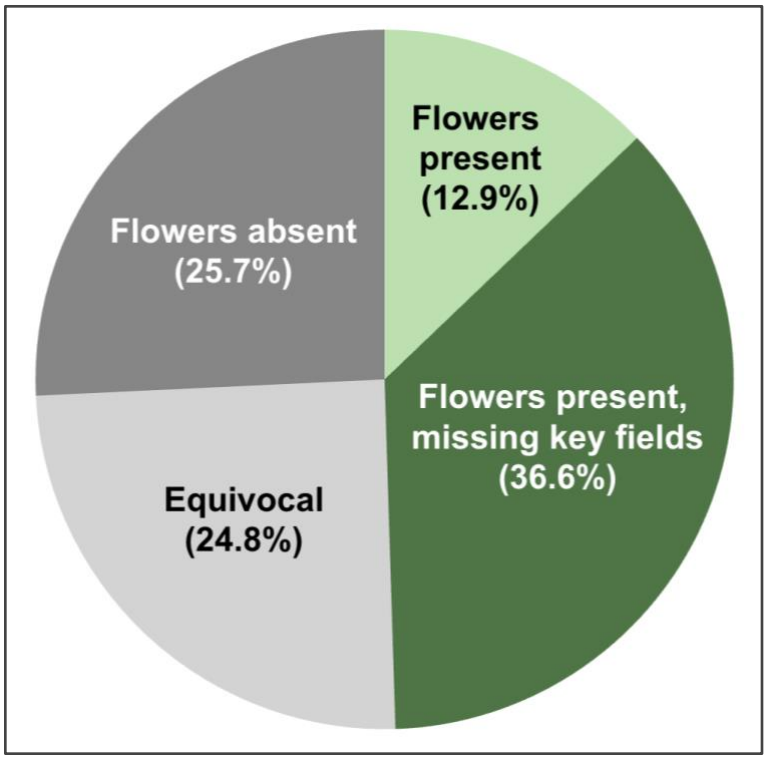
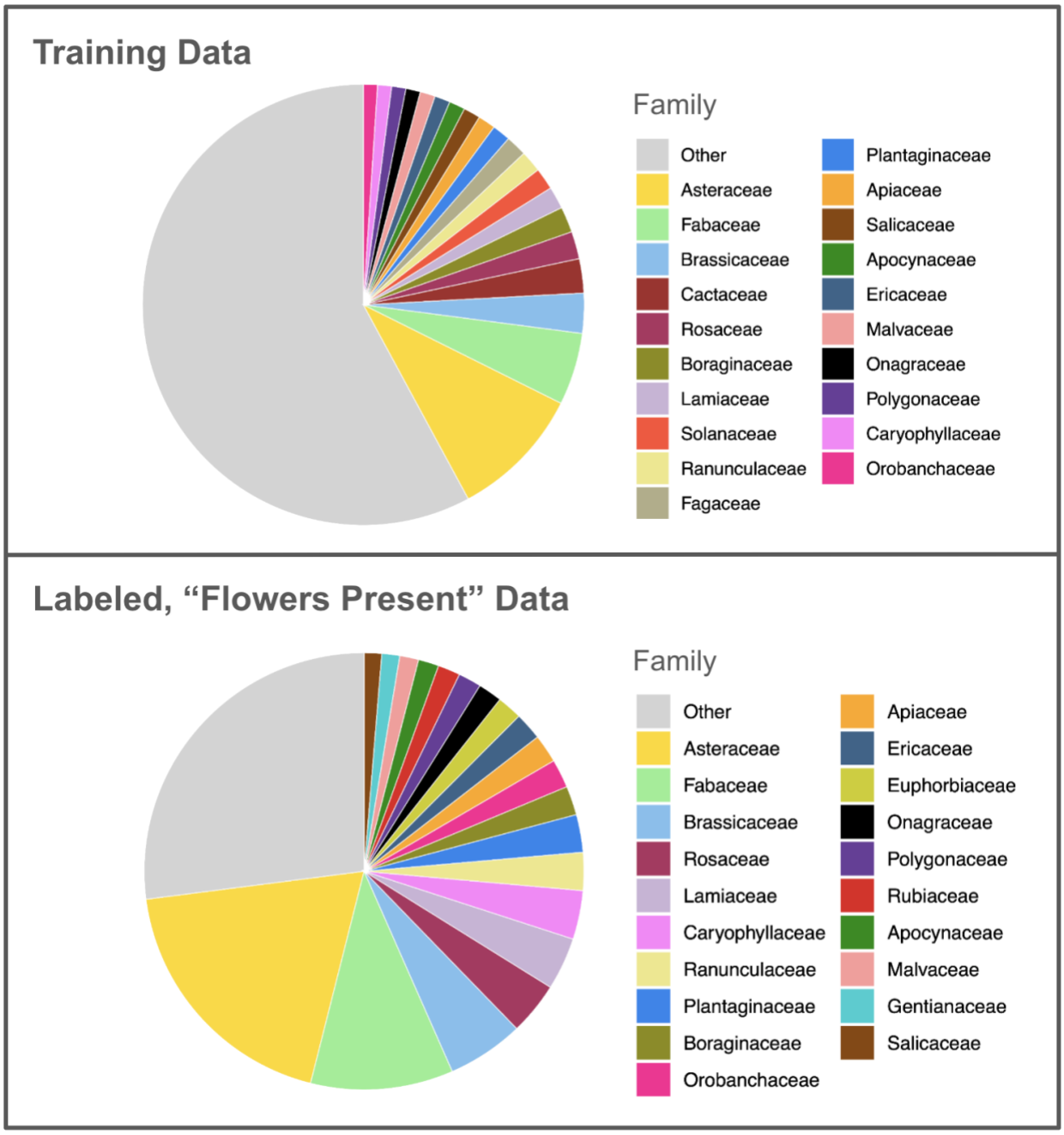


Figure 3. Percent of all records pushed through the ensemble pipeline labeled as “flowers present”, “flowers absent”, and “equivocal”, highlighting that the majority of records labeled as “flowers present” were unusable because they were missing critical date and/or location information. Only records in the “Flowers present” category that had all critical fields were retained, representing 12.9% of all records.

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.



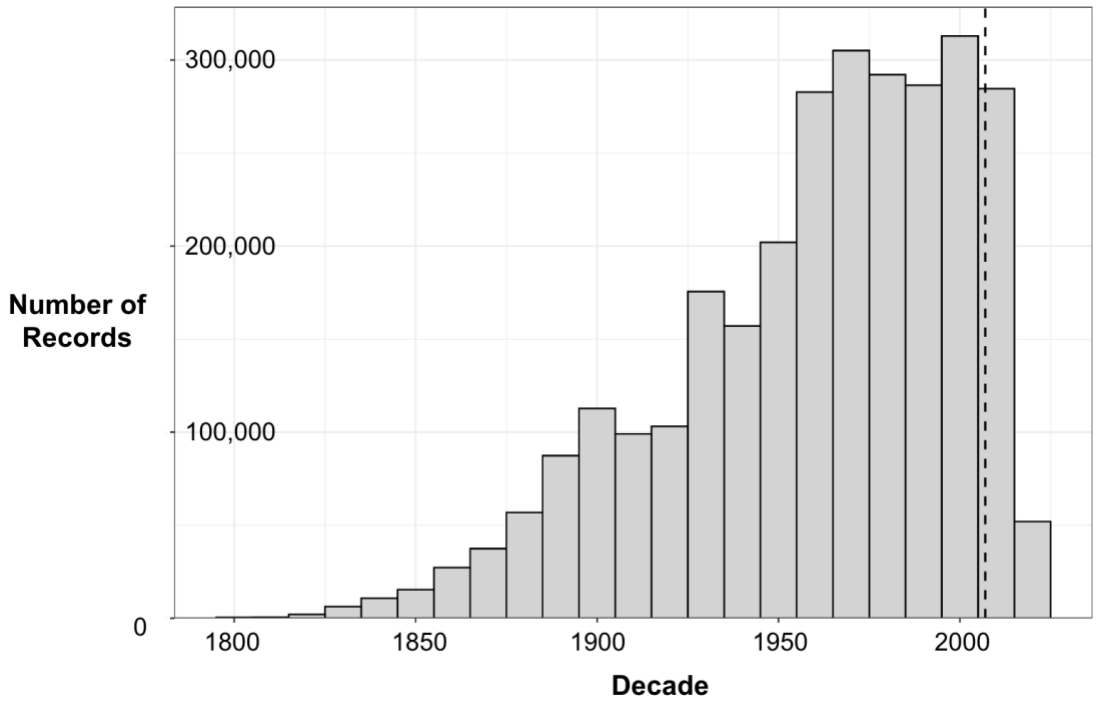
627

628 **Figure 4.** Pie charts showing the top 20 families by number of records in the training dataset

629 (includes training, testing, and validation sets) and in the labeled, present data.

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

630



631

632 **Figure 5.** Number of “flowers present” labeled records per decade, beginning at 1800. The

633 dashed line indicates 2007, the year the USA National Phenology Network was founded and the

634 year before iNaturalist began.

635

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

636 **Tables**

637 **Table 1.** Description of the three models included in the ensemble modeling approach.
638 Threshold values are cut-offs applied to each model's continuous output probabilities. The
639 difference between model 2 and model 3 is the loss function used. Mean square error loss
640 functions are more akin to confidence scores and are typically Gaussian, while cross-entropy is
641 better suited for classification tasks and tends to skew continuous probabilities more strongly to
642 0 or 1.

Model type	Image size requirements	Threshold low	Threshold high	Loss function
EffNet CNN	528 x 528 pixels	0.05	0.95	Mean squared error
Visual transformer network (ViT)	384 x 384	0.5	0.7	Mean squared error
Visual transformer network (ViT)	384 x 384	0.5	0.7	Cross entropy

643

644

bioRxiv preprint doi: https://doi.org/10.64898/2025.12.19.695552; this version posted December 22, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license.

Table 2. Expert validation of machine-labels for flowers present or absent. Bolded values represent situations where the human expert and model consensus were in agreement. Grey cell represents the accuracy relevant to the final, downstream data.

Expert Annotation	**Model consensus: Present**	**Model consensus: Absent**
Present	**90.5% (543)**	47.7% (143)
Absent	9.9% (57)	**52.3% (157)**