1. Improvements over our previous submission

This proposal argues for an innovative approach for extracting high density morphometric data from biological images which will be developed in the context of three morphologically diverse model organisms. We previously submitted a similar proposal to the CDI program in 2009, which received favorable reviews, and we thank the panel for their constructive critique. In this resubmission, we have specifically addressed the comments received from the previous review, and we believe the result is a much improved proposal. First, we have increased the diversity of structure that we will study in detail by including *Arabidopsis* (§5.2) (and adding co-PI Palanivelu). Second, we have clarified how the method makes use of, and is different from, standard image processing (early in §3). In particular, many bottom-up image analysis issues are **replaced** by Bayesian inference on a carefully constructed forward imaging function. Third, we have taken further care to explain how the typically high computational requirements of sampling methods can be mitigated (§3.4.1, §3.4.2, §3.4.3) and that our experience with *Alternaria* ([130, 131]) and learning furniture models [129] strongly suggest that our approach should be tractable (§3.5). Finally, we have clarified how we will evaluate the system (§4).

2. Introduction — a transformative approach to biological structure

We propose a new paradigm for automatically quantifying biological structure from image data. This original effort will have significant impact because algorithmic limitations in automated extraction of quantified form—numbers from structure—are currently the biggest impediment to rapid discovery in many high throughput experiments. Because biological form is tightly coupled to key issues in organism function, identity, and evolution, there is much to learn by linking morphological characteristics to other measurements spanning genetic, molecular, environmental, and fitness data. However, this enterprise requires that morphology be quantified accurately on a large scale to capture the potentially subtle variation within and across groups. Thus, there is great opportunity for accelerated discovery using high-throughput experiments incorporating image capture coupled with morphology extraction **provided that** it can be significantly automated.

Many researchers have argued that better methods to automatically quantify form are desperately needed (e.g., [44, 58, 62, 88, 95, 157]). Thus, it is not surprising that much effort has gone into quantifying form or identifying specific structures such as spores [111], leaves [40], or neurites [16-19, 86, 97-99, 102] using image data. However, so far, quantifying form in this manner has typically been very *task-specific*, and much research has been driven by specific analyses to explore particular hypotheses. This leads computational approaches being significantly influenced by what is observed in images and how they address the specific task. The challenge then becomes processing images to compute target quantities, and leads to a very *image-centric* analysis strategy.

We propose a **completely different way of thinking** applicable to many organisms and organs (Fig. 1). The key to success is to focus on the representation of the organism in 3D—an *organism-centric* view*.* For this we propose stochastic geometric models for organism parts and their assemblage that are **learned from image data** given a broad schema that is easily specified by domain scientists (§3.1). Appearance characteristics (e.g., color, texture, shape, size) are attached to organism parts. The entire representation is linked to images through an imaging model (§3.3) which is independent of the structure model. This dissociation of the

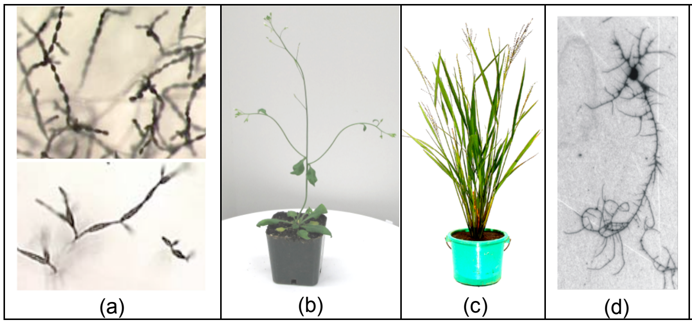

Fig 1. Examples of diverse biological form that can be addressed by the methods proposed here: (a) Microscopic fungi from the genus *Alternaria;* (b) *Arabidopsis thaliana*; (c); a variant of wild rice, *Oryza glaberrima*; (d) Drosophila brain neurons grown in culture. *Alternaria*, *Arabidopsis*, and *Oryza* will be addressed in detail. Other plants will be used to test rapid expansion to more species. Neurons are beyond project scope.

structure model from the imaging process has significant advantages over approaches that analyze morphology using image-oriented representations.

Our approach is inspired by a large body of work demonstrating that compact, grammar based models such as L-systems (e.g. [20, 83, 84, 101, 115, 152]) can model a range of plant and other forms that exhibit complex levels of phenotypic variation. However, the focus of previous efforts has been on either providing good models for computer graphics applications, or developing a representation of a particular organism for the purpose of linking a model of growth to a model of form. Quantitative use has been limited to summary statistics [114], or building classifiers [127]. Fitting such models to image data, or learning their parameters and meta-parameters from images, as we propose in detail below (§3) has seen little attention. However, we see significant untapped power in fitting such models to image data for quantitative applications. From a biological perspective, there is a significant difference between producing a believable image of a generic object, and being able to extract complex metrics of the structure from a particular object in the laboratory or the field.

Thus we propose a strategy **embodied in the following pipeline** (Fig. 2). First, a tool that leverages knowledge about structure enables domain scientists to easily generate a rough 3D model for a newly considered organism category (e.g., genus). Second, this scaffolding, together with images of multiple individuals from one or more groups (e.g., species), is used to automatically learn a stochastic geometric model for the category and groups. Collaterally, we determine the geometric structure for all the examples. Third, the learned model can then be fit to images of new examples to determine their geometric structure, as well as their phenotype and corresponding identity hypothesis (e.g., species). Finally, we can easily compute numerous morphological quantities, including many unanticipated ones.

This approach is grounded in principled statistical modeling and inference methodology. The model schema builder will be optimized for effective user input, but behind the scenes the schema will be translated to parametric Bayesian statistical models (§3.1). The schema for a category (e.g., genus) will have parameters that potentially vary over sub-categories (e.g. species), the details of this variation being exactly what we want to learn from data. Note that category parameters are typically distributional (e.g., means and variances for counts of branches on a particular plant species). Hypotheses for category parameter values for each sub-category define the distributions over structure instances for different sets of examples. The structures for individual examples are characterized through instance parameters (e.g., particular angles at which branches emerge) (§3.2). The imaging model (§3.3) links the hypothesized structure to the image data to provide a principled computation of the statistical evidence for structure parameters in the observed images.

This forward model—going from the model to data–is reversed using Bayes rule to give a posterior over the parameters (§3.4). Here, the parameters are a combination of category level parameters shared by sub-categories, instance parameters, and camera parameters for each image. To estimate parameters we optimize the posterior function for the entire collection of

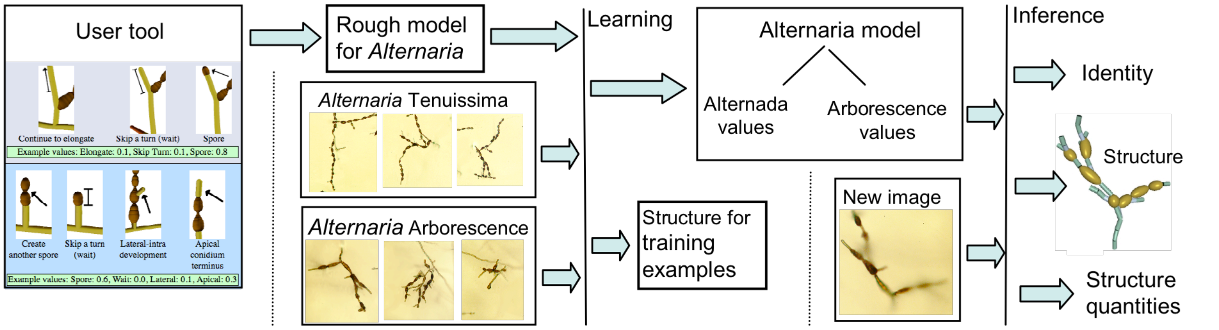

Fig 2. The general paradigm using the genus *Alternaria* as an example. A user tool enables defining the rough form of Alternaria which is automatically translated into a Bayesian model. The model is then fit to images of multiple examples of one or more species (two shown). The learned model is then applied to new images to determine identity, structure, and quantified morphology.

images. Since the posterior is very complex, the most promising methods are sampling-based approaches which are feasible due to the relatively strong model and carefully constructed image likelihoods.

Having learned a category model, we can simultaneously classify and extract numerous structural metrics for each new example. More specifically, the values of category level distributional parameters define a phenotypic group, and classification proceeds by testing each of these phenotypic models and choosing the best one. Very importantly, because we are using a generative statistical model, we have estimates for classification confidence and fit quality. Thus anomalies, perhaps suggestive of a new phenotype, can be automatically flagged. More generally, the underlying approach can support a variety of workflows, including learning additional levels of hierarchy and defining subtle phenotypes. In summary, the core capability that we propose is to determine phenotype relative to previously seen data, and quantify form relative to a representation as an assemblage of biologically relevant parts (e.g., hyphae and spores for fungi, branches and leaves for plants) and the mathematical representations for those parts.

Project goals and evaluation. We will evaluate the modeling approach with respect to the following key **aims**: **1)** We can learn models that can be fit to new examples from a group (e.g., species or strain), as demonstrated by the fits to image data being consistent with image data; **2)** That learned models are sufficient biologically relevant to support phenotype classification; **3)** That the relatively arbitrary structure quantities easily computed from models fit to images supports scientific discovery; **4)** That the process is scalable so that new organism groups become successively easier to handle using our model schema translation approach; and **5)** Putting this capability on-line is effective for education and outreach (§7).

Phenotype structure quantification in practice. Demonstrating biological relevance (goals **2** and **3**) requires tight collaborative research with domain experts on model organisms where we have libraries of carefully catalogued varieties, ongoing cultivation, and molecular data. Thus we will fully develop the methodology in three specific cases: 1) filamentous fungi in the genus *Alternaria* [116], a major environmental saprobe that also impacts human health directly (Fig. 1a); 2) well characterized and readily available strains and close relatives of the most widely researched model plant, *Arabidopsis thaliana* (a dicot) that displays a wide-range of morphological differences (Fig. 1b); and 3) wild and domestic varieties of rice (a monocot) from carefully bred lines that express specific phenotypes when grown under controlled environmental conditions (Fig. 1c).

As a key test of **aim 2**, our system should be able to quickly differentiate morphologically different fungal strains/plant varieties, which in some case vary by only extremely subtle morphological differences. We posit that this should be effective even when differentiating them requires expert knowledge, thereby demonstrating usefulness for automated identification when experts are not available. Further testing will be implicit in screens for morphologically different phenotypes among groups defined by identity, genetics, gene expression, experimental condition, or fitness. Here phenotype differences would be discovered in parallel with evaluating our ability to distinguish them in an iterative process.

Aim 3 is to **demonstrate the potential for discovery** by linking structure quantities from these organisms to environmental conditions, molecular data, and breeding strategies. This will demonstrate the power of opening up bioinformatics to morphology, i.e., phenomics. This allows multiple linkages, supporting goals such as understanding the molecular basis of vegetative growth, or finding promising genes for crop improvement. Specific experimental directions for the three domains are developed below (§5).

Transformative aspects. The proposed work pursues a fundamentally new approach to research in biological structure that is especially suited for automatically quantifying critical phenotypic characteristics across numerous other organisms. This has great transformative potential because it can, for the first time, open up morphology to bioinformatics inquiry by putting complex morphometrics from structure alongside other data. **Features of the approach that embody conceptual breakthroughs include:**

- Focusing on the representation of semantic biological structure as opposed to using image features to inform specific tasks. Thus, little additional programming is needed to implement unanticipated morphometric measurements.
- Decoupling the representation of structure from imaging processes. Imaging is considered an independent module. These two factors can be separated during inference because imaging effects and alternative structure hypotheses rarely masquerade as each other.
- Translating schema sketches based on L-systems and user-oriented concepts into Bayesian models that support learning of model details. As the body of modeled structure expands, the human supervisory effort to model new varieties rapidly decreases.
- Learning the statistical parameters for a given group from image data (e.g., the distribution of branching angles for a particular plant species) using Bayesian inference. These quantities are essential to biological understanding but are difficult to acquire by other means.
- Hierarchically arranging models so that structural knowledge across related groups (e.g., species, strains, mutant alleles) is shared.
- Fitting high-level component-based models representative of biological semantics to image data for automated quantification of form. This contrasts with the conventional low level approaches of fitting particular parameterized shapes of already known components.

Benefits. In addition to the overall scientific benefit and **intellectual merit** of understanding the origin and role of biological structure variation in many domains, this research direction also has far-reaching practical benefits. Examples of these **broader scientific impacts** include: 1) linking phenotypic diversity within disparate taxonomic groups to wealth of biological, environmental, and genetic variables; 2) automated species identification which has applications in biodiversity studies and the detection of invasive, toxic, or sensitive species (e.g., CITES [9] listed species); 3) guiding crop breeding efforts by providing larger scale and more accurate data on morphological quantities that affect yield. **Broader educational impacts** will be enhanced by the convergence of trainees from diverse disciplines and by explicit efforts to engage undergraduates and high school teachers in computation for the biosciences.

3. A computational approach for quantifying biological structure

The forward, generative model shown in Figure 3 modularizes the processes going from abstract high level representation down to pixel observations. Consider an organism category (e.g. genus) that represents the range that can be covered by a shared schema for assemblage. As described shortly (§3.1), we will develop a system for rapid rough specification of such schema, and automated translation of the user oriented representation into Bayesian models. Such schema establish the hierarchical structure of statistical parameters. For example, our proposed Alternaria fungal model might have a Poisson distribution for the length of hyphal filament below a branch, along with a meta-distribution over the Poisson mean. When this is smaller, the species growth has a more compact appearance with greater branching density. Since this is a distribution over distributions, we also model an individual being a particularly compact or non-compact example within what is typical for the species. This kind of distributional parameter implicitly affects structure topology and thus the number of parameters needed to describe an individual. Additional distributional parameters characterize parts (e.g., means and variances for leaf appearance for plants) to complete the idealized model for an individual. Parameters for environmental effects such as leaves wilting or folding or insect damage can be included here also. Finally, parameters may be a function of age or stage of development (§3.2) if the organism is to be considered at different life stages.

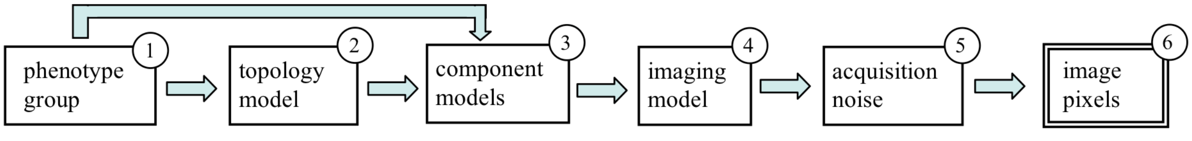

Fig 3. The forward model breaking up the modeling process into modules. The group (1) implies a distribution over topologies (2), and components (3). Component models (3) provides shape, color and texture. The camera (4) is independent of morphology, as is image acquisition (5).

Image formation(§3.3). We consider the 3D representation of structure and appearance as distinctly separate from the imaging process that maps structure to image data. While we use computer graphics to display models (e.g. Fig. 9 far below), it is important to realize that we use the term “model” to refer specifically to the underlying mathematical representation. This representation can be combined with an image formation process to characterize how biological structure give rise to real image data under a particular imaging method.

Statistical inference (§3.4). Given this forward process, the missing piece is to determine the parameters—and hence structure and possibly group characteristics—from image data. This is achieved using Bayesian inference. We distinguish between two inference processes:

1) **Fitting** (Fig. 4). Given a model for a group, fit that model to an image, thereby determining its structure. Fitting can further be used for automated **identification** by fitting multiple alternative models, and considering the relative likelihood of each fit.
2) **Learning** (Fig. 5). Determine the more general meta-parameters (e.g. distributions over branch counts) for a group by fitting a model with both group and individual parameters to the combined image data for a group simultaneously. Here, the meta-parameters are shared across the group, and individual fits are produced collaterally.

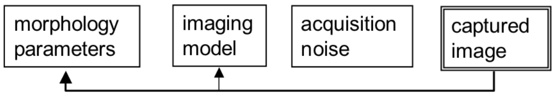

Fig 4. Inference for **fitting**. Parameters are sought that give high posterior probability to the captured image. Imaging model parameters are determined as by-products.

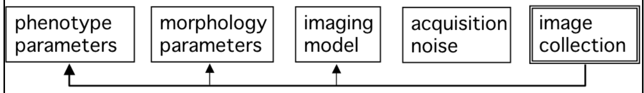

Fig 5. Inference for **learning**. Here the key parameters are for entire phenotype group. However, the morphology of individuals and the imaging model parameters are also determined.

Relation to traditional image processing. Our approach uses image processing to extract simplified low level representations (e.g., edges, textures), and to speed up inference with data driven sampling (§3.4.2, [94, 151]). Traditional image processing extracts targeted cues and assembles information “bottom up”, typically heuristically. Instead, the proposed approach evaluates hypothesis that integrate the bottom up, low level information with “top down” high level global information. These two information sources help each other immensely. All information sources are expressed statistically, and are combined using Bayesian methodology.

Unifying principles. The range of image data that our approach can potentially address is enormous. One reason is that the complexity is made modular. Topological complexity is handled independently of component complexity, and strategies for each can be shared across many organisms. Further, factors that modulate the statistics are handled hierarchically—there is variation within an individual, conditioned on sub-groups, conditioned on groups, and so on. In addition, process that potentially can be very complex, such as insect damage to leaves can be parameterized so that they apply to many plants. Finally, the effect of imaging can be separated during learning because model hypotheses rarely masquerade as viable image system parameter hypotheses and vice versa. Thus the organized approach developed here can have much more impact than research focused on isolated modeling or image analysis for specific tasks. The whole is clearly greater than the sum of the parts. However, to execute this vision, much innovation and careful research is needed. Some details on how we will proceed follow.

3.1. Sketching structure

Our approach is inspired by a large body of work demonstrating that compact, grammar based models such as L-systems (e.g. [20, 83, 84, 101, 115, 152]) can model a range of plant and other forms that exhibit complex levels of phenotypic variation. In its pure form, an L-system models organism growth using (typically) a context free grammar where productions are applied in parallel (Fig. 6,7, next page). In preliminary work, we have built a simple web interface [141] that allows domain scientists, and even middle school students [144], to model *Alternaria*, *Oryza*, *Arabidopsis*, saguaro, and creosote. Even though our interface is currently crude, users find it surprisingly efficient for creating variations of these organisms. This is in concordance with the success of other software for plant modeling based on L-systems [4, 5, 12].

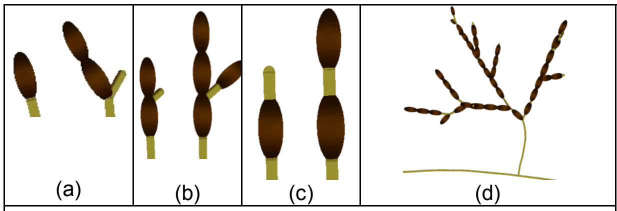

Fig 6. Building *Alternaria* sporulation structures from simple components (spores, hyphae) and simple rules that are applied stochastically. The pairs (a,b,c) illustrate possible growth rules where the structure on the left is replaced by the structure on the right. The application of rule (a) creates a branch; rules (b,c) add spores; (d) the result of applying the rules stochastically 60 times.

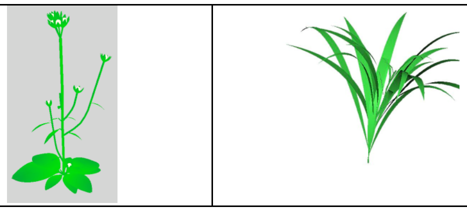

Fig 7. Rough models that we have sketched up for *Arabidopsis* and rice (leaves only). We expect that we can make the sketching process efficient and accurate. A key proposed innovation is to translate user sketches to models suitable for inference, which can then be refined using image data through Bayesian update. Here the user models provide an initial prior on the structure.

Despite the success of the L-system approach, our preliminary study suggests that pushing the growth metaphor blindly does not necessarily lead to good models for inference. For example, producing simple structures with a replacement model for tip-growth is an inefficient way to model a simple rice leaf. On the other hand, a grammar model with terminals representing entire components such as rice leaves are clearly more parsimonious for inference where the growth history is not relevant. Of course, these are both grammar models, and grammar models are in general attractive for inference [96, 106, 163, 165].

Hence we propose developing a method for automatically translating effective user oriented representations to an internal representation effective for inference. Preliminary work suggests that the independent sources of complexity discussed above will likely be exposed to the user. In particular, the user will be given tools to specify topology based on grammar rules, and be able to specify components from a pallet of possibilities. For example, there are only a limited number of basic leaf shapes that can be combined with distributions over transformations to model most leaves [67, 68]. What is less clear is when it is advantageous to represent growth and how to do so. We will resolve this as part of this project.

Again, the key enabling idea is to automatically translate the user representation into one that supports inference. Recall that the user sketch is used to provide the architecture of an organism group, with the particular statistics to be learned from image data. There are many organism groups. Hence it is critical that the sketching process is efficient, and the translated models support effective learning and inference.

3.2. Modeling and meta-modeling

More formally, the model for a group within a category is represented by the group distributional parameters, Θ^{group} that are meaningful within the corresponding category schema provided by model sketch. We can represent structure of one or more individuals as a collection of parameters in the variable length row vector, Θ^{model} given by

$$\Theta^{\text{model}} = \left[\Theta^{\text{group}} \quad \Theta_1^{\text{model}} \quad \Theta_2^{\text{model}} \quad \ldots \quad \Theta_n^{\text{model}} \right] \quad , \tag{1}$$

where each sub-vector of Θ^{model} has many elements and the number of elements in the individual part of the model, Θ_i^{model}, is not known a priori.

Priors and constraints. For a given set of parameters, Θ^{model}, our approach defines a process to compute the prior probability density, $p(\Theta^{\text{model}})$, perhaps conditioned on the sub-category (e.g., species) hypothesis, s, by

$$p\left(\Theta^{\text{model}} \middle| s\right) = p\left(\Theta^{\text{group}} \middle| s\right) \prod_i p\left(\Theta_i^{\text{model}} \middle| \Theta^{\text{group}}\right) \quad . \tag{2}$$

This encodes our understanding of the structure possibilities before consulting the image. This computation of $p(\Theta^{\text{model}})$ also embodies any constraints that the form might have. For example, the structure encoded in a particular instance hypothesis, Θ_i^{model}, might self-intersect, and we normally impose the constraint that this cannot happen, and thus $p(\Theta^{\text{model}})$ would be zero.

Sub-part characterization. In the compositional model, the identity of sub-parts (e.g., spore, leaf type) is either implicit or specified as a group or category level parameter. Each kind of sub-part requires a parameterized model for its geometry, color and texture. For example, our *Alternaria* model uses ellipsoids for spores and cylinders for hyphae. Such parameter sets are associated with a probability distribution that accounts for large scale variation of the sub-parts.

Sub-part finer scale variation and environmental effects. Beyond the variation described so far, we expect small scale variation in the sub-parts (e.g., spores are not perfect ellipsoids). This can be modeled as variance from the base model. One can conveniently absorb such variation into the imaging process, but for 3D structure, this is less statistically faithful, and hence we will integrate it at this stage. Additional environmental effects, such as loss of leaf tissue due to consumption by insects, are also most faithfully considered at this point. As noted above, models for such processes, while challenging to get right, can be parameterized so that they can be shared over many organism categories. Hence what is learned with one organism can be propagated to many others.

Temporal modeling. The parameters above can easily be conditioned on a hidden variable for organism age or stage of life. This modularizes complexities due to growth, and allows us to infer the organism age. Since the user supplied schema (§3.1) typically has a growth model, priors for the temporal aspects for the model will be readily available from our proposed approach to translate the user representations to Bayesian graphical models.

3.3. Imaging processes and data likelihood

Imaging Process. A given data set has an image formation process, $I\left(\Theta_i^{\text{model}}, \Theta^{\text{shared}}, \Theta_j^{\text{image}}\right)$, that takes a model hypothesis, Θ_i^{model}, and maps it to one or more “ideal” images based on the shared imaging parameters, Θ^{shared}, and parameters specific to each images, indexed by j, Θ_j^{image}. Thus in an experiment with multiple images, of multiple specimens, there can be imaging parameters that hierarchically apply to the experiment, the specimen, or each image, each of which may be known (thus constant), or fit simultaneously with the structure.

For example, consider an experiment where one takes multiple images—recommended to deal with occlusion—of an Arabidopsis plant that is rotated on a turntable to provide different angles of view (Fig. 8). Here some camera parameters are shared by all images but each image is associated with a different rotation angle. If we calibrate both the camera and the turntable, then the camera parameters are known. However, in the general case, some camera parameters will need to be inferred simultaneously with structure, which is quite feasible as we have shown in our work on learning models for simple furniture objects (Fig. 10 below, [129]). Regardless, for each image, the camera parameter hypothesis $(\Theta^{\text{shared}}, \Theta_j^{\text{image}})$ implies a projective transformation of the 3D model into the 2D image plane. Notice that the main challenge in combining images from multiple views—image feature correspondence—largely disappears when the multiple views are used to evaluate model hypotheses. A similar observation has previously been used in the case of surface patch models [50, 51].

For *Alternaria* (Fig. 1a, 9 below; §5.1), and other microscopic specimens, the imaging process, $I(\bullet)$, is quite different. Here we collect image Z-stacks, where each image is associated with a known image plane position. These images have significant blur where the specimen is out of focus, and to deal with this we fit parameters for the microscope point spread function simultaneously with the model [130, 131].

Measurement process. The imaging process, $I(\bullet)$, is then followed by a measurement process, with parameters, $\Theta^{\text{measurement}}$, that provides the likelihood of seeing what is actually detected in the image, based on the ideal image. This maps feature detection to probabilities, and also accounts for missing observations and imaging noise.

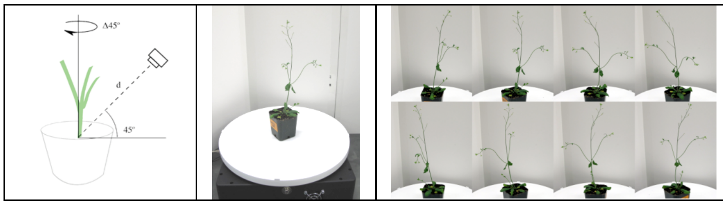

Fig 8. A convenient way to capture images of potted plants that we use for *Arabidopsis*. The turntable shown in the middle image is computer controlled. The right hand figure shows eight images captured at different rotations.

Data likelihood. The imaging and measurement processes together give the likelihood of the observed image data given the model. Focusing on a particular image, *j*, and for simplicity, a specific feature type (e.g. edge strength), *f*, then we have a product over pixels indexed by *k*:

$$p\left(D_j \mid \Theta\right)=\prod\left\{p\left(f_k \mid \Theta\right)\right\} \ , \qquad (3)$$

where $p\left(f_k \mid \Theta\right)=p\left(f_k \mid I_j\left(\Theta_i^{\text{model}}, \Theta^{\text{shared}}, \Theta_j^{\text{image}}\right), \Theta^{\text{measurement}}\right)$ accounts for imaging noise on top of contributions to the observed f_k due to both the model, and non-model sources of image pixel intensity such as spurious items in view (clutter). Note that $p(f_k \mid \Theta)$ implies a penalty for missing evidence that is expected by the model hypothesis. For example, if the model implies that an edge feature at a pixel is likely, then an observation corresponding to edge absence should give a relatively small value of $p(f_k \mid \Theta)$. Such negative evidence is important because without proper counting of missing observations, the fitting process becomes biased towards fitting more structure to account for the positive observations alone. Here spurious structure leads to predicted edges that are not observed and thus the evidence for the model is diminished.

3.4. Inferring biological structure by sampling.

The set of all relevant parameters, Θ , contains model, imaging, and measurement parameters, with model parameters being divided into ones for the group, and ones for the individual.

Posterior distribution for fitting. If we already have learned a model for the group, then Θ^{group} is known, and we seek the specific model for an instance, Θ^{model}. Collaterally, we also find the capture parameters, Θ_j^{image} , for each image, D_j of the instance (e.g., different slices of a stack, or different views). Assuming the shared camera and measurement parameters are also known, the posterior distribution for Θ^{model} (and $\left\{\Theta_j^{\text{image}}\right\}$) is

$$p\left(\Theta^{\text{model}},\left\{\Theta_j^{\text{image}}\right\} \mid \left\{D_j\right\}\right)=p\left(\Theta^{\text{model}} \mid \Theta^{\text{group}}\right) \prod_j p\left(D_j \mid \Theta_j^{\text{image}},\Theta^{\text{model}};\Theta^{\text{shared}},\Theta^{\text{measurement}}\right)\left(1/p(\left\{D_j\right\})\right) \ (4)$$

using Bayes' rule and assuming conditional independence of $\left\{D_j\right\}$, given Θ^{model}.

Posterior distribution for learning. For learning we seek Θ^{group} which requires multiple instances for the group. Here, the models for the individual instances, Θ_i^{model}, as well as image specific parameters for every image of every instance, are determined collaterally:

$$p\left(\Theta^{\text{group}} \mid \left\{D_{ij}\right\}\right)=p\left(\Theta^{\text{group}}\right)\prod_i\prod_j p\left(D_{ij} \mid \Theta_{ij}^{\text{image}} , \Theta_i^{\text{model}} , \Theta^{\text{shared}} ; \Theta^{\text{measurement}}\right)\left(1/p(\left\{D_{ij}\right\})\right) \ . \qquad (5)$$

Inference by sampling. Both (4) and (5) have the form $p\left(\Theta \mid D\right)=p\left(D \mid \Theta\right) \ p\left(\Theta\right)\left(1/p(D)\right)$ where the last factor, $1/p(D)$, is constant given the observed data, D, and need not be considered further. In both cases we seek good values of the parameters, Θ, given the observed image data, D. One form of inference is to find the parameters, Θ, that give the maximum value of the posterior distribution, *p(Θ|D)*, which is equivalent to finding

$$\max_{\Theta} \ \left\{ \ p\left(D \mid \Theta\right)p\left(\Theta\right) \ \right\} \ . \qquad (6)$$

Thus the task reduces to optimization to find the maximum a posteriori (MAP) estimator. In the case of inferring complex biological form, there are a number of factors that make doing so difficult: 1) the number of parameters is a priori unknown; 2) the number of parameters is in principle unbounded, and typically quite large (easily tens of thousands); 3) the parameter set is a mix of continuous, discrete, and categorical; and 4) the parameters have complex constraints (e.g. we must impose that structure cannot self-intersect). Markov chain Monte Carlo (MCMC) methods [30, 90, 103] provide a powerful, unified framework for performing such optimizations.

MCMC optimization simulates a Markov chain whose (unique) stationary distribution is the posterior distribution in (6). A variety of methods can do this, e.g., Langevin dynamics [103], hyperdynamics [136, 155, 156], importance sampling methods [90], adaptive cross-entropy variants [48, 122-124], and the standard Metropolis-Hastings (MH) method [90]. Under easily arranged conditions, after some time the samples come from the target distribution, and the sampler gravitates to high-probability areas thereby aiding the search for a good solution.

Abstract problem structure. Our problem has the following important, but *understudied* structure. We have topologies expressed using an indeterminate number of discrete and categorical parameters (e.g., branch counts and leaf types). Further, there is much approximate hierarchical conditional independence (leaves on one branch only minimally interact with leaves on other branches). Finally, the topology determines the number of continuous parameters for quantities such as branch angles and leave size.

Baseline sampler. Our work so far suggests the following effective sampling approach [129] as a baseline. Here we alternate between: 1) sampling with the reversible jump version [63, 64] of Metropolis-Hastings (MH) to sample the discrete and categorical parameters allowing changes in the number of them (jump moves [164]) and 2) hybrid Monte Carlo to sample the values of continuous parameters (diffusion moves [164]).

Sampling challenges. The main challenge for sampling remains computational cost. To continue to address this challenge as we scale up our method for high-throughput experiments, we will improve our sampling approaches algorithmically and computationally. For the former we note that most work in sampling focuses on producing samples suitable for integration which is *more restrictive* than finding the MAP estimate. To find the MAP, we only need to visit the modes of the distribution which is potentially easier. Sometimes this is exploited using simulated annealing [29, 30, 57, 82, 154], but this does not apply well to the search over topologies. Details on sampling challenges and strategies we will explore for dealing with them follow.

3.4.1. Challenge 1: Reducing random walk behavior in discrete/categorical sampling.

Most MCMC strategies use *reversible Markov chains*, which lead to expensive "random-walk" behavior. A Markov chain is reversible if it satisfies the *detailed balance condition,* $p(\Theta|D)\alpha(\tilde{\Theta}|\Theta) = p(\tilde{\Theta}|D)\alpha(\Theta|\tilde{\Theta})$. This means that the probability of a transition from state Θ to state $\tilde{\Theta}$ is the same as the probability of the reverse transition. Hence a move by a reversible Markov sampler can easily be undone by subsequent steps. In our problem, the state space of the sampler has a tree-like character, reflecting branching rules in the grammar. For Markov samplers to go between two topologically distinct structures S_1 and S_2 via purely local moves, e.g., adding and deleting branches, it is necessary to first delete branches from the first structure, then generate the new structure. While global transitions like splitting and joining substructures can partially alleviate this problem, reversible chain will still waste much computational time deleting and adding the same branches in this process.

This common issue motivates recently proposed non-reversible Markov chain algorithms [46, 52, 74, 78, 104, 153]. Intuitively, we would like to endow Markov samplers with the ability to remember some of the past history. A particularly simple idea that achieves this was proposed by Neal [104] (see also [153]). It is easy to demonstrate that this algorithm is extremely effective on very simple graphs, and we are investigating this approach on graphs with "bottlenecks," akin to replacing the parent node of a sub-tree with a different sub-tree.

We propose to develop and analyze non-reversible sampling approaches for topology that takes advantage of the graph structure of our problem. We already take advantage of this structure when computing the likelihood, but existing methods cannot do so for the topology sampler. Achieving this has potential for significant impact on our problem and many others.

3.4.2. Challenge 2: Avoiding becoming trapped in metastable states.

Our posterior distributions, like the potential energy surfaces of real molecules, form highly complex landscapes with multiple local maxima. Many samplers (e.g., hybrid Monte Carlo) can search for optimal parameters within this landscape as they execute a random walk through parameter space that is strongly biased towards increasing posterior probability. This has the consequence that samplers can get trapped near local maxima for many iterations before escaping, making it hard for the sampler to find a global maximum in reasonable time. We have experience with several methods that can alleviate this, include simulated annealing [29, 30, 57, 82, 154], hyperdynamics [136, 139, 155, 156], and covariance scaled sampling [135, 137, 138]. However, a more critical issue is the discrete sampling search for topology.

Data driven sampling. A related problem is that the sampler can become trapped because proposed moves are repeatedly rejected. To improve acceptance rates to reasonable levels, an important strategy we will use extensively is *data-driven heuristics* [164] which use standard image analysis to provide a probability distribution over the location of useful sub-parts such as *Alternaria* spores. We have implemented data-driven methods with a multiple image resolution strategy to make reversible jump Metropolis-Hastings MCMC relatively efficient and robust on our *Alternaria* image stacks [130, 131]. For this project, where an important goal is to automatically translate user sketches into inference engines, we propose automating the integration of data driven methods into these automatically generated engines.

3.4.3. Challenge 3: Fast evaluation of posterior distributions and their derivatives.

Our sampling strategy entails evaluating the posterior distribution, and possibly its derivatives, for every proposed change in parameters. Strategies to make this efficient follow.

Approximate conditional independence. We typically have much local structure, which amounts to significant, approximate conditional independence in the likelihood, and enables modest changes to be proposed with very little computation. For example, re-computing the likelihood in response to change in branch angle on one side of a plant, can be approximated by ignoring parts of the plant that are far away.

Analytic approximations. Because our posterior functions are so complex, initial implementations rely heavily on numerical approximations. However, there is much to be gained by deriving analytic approximations of parts of the function, thereby significantly increasing computation speed (often referred to as Rao-Blackwellisation [43, 79]). This is easier for microscope stacks than standard camera images because the later creates non-differentiable regions in parameter space due to parts being occluded.

Parallel computation. This local structure of our inference problem means that there are many opportunities for parallelization to be exploited with the rapidly expanding availability of many-core hardware. For example, a new inference approach that we will explore is to have multiple samplers running in parallel from different starting points that share information, similar to using data driven mechanisms to improve proposal quality.

Hardware acceleration We will speed up specific computations such as projecting 3D data onto 2D images using Graphics Processing Unit (GPU) hardware, which we already do for the furniture sampler [129]. Other functions such as fast Fourier transforms needed for the microscope stack imaging model are now relatively available for GPUs [59-61].

3.5. Preliminary results inferring structure using sampling

Preliminary results on modest size problems suggest that with care, fitting and learning structure by sampling is tractable. We have published results for fitting Alternaria to brightfield images stacks [130, 131] (Fig. 9) and learning models for simple furniture categories such as tables and chairs from standard images [129] (Fig 10). For furniture categories, a single view per instance yielded reasonable results. For much more complex biological form that is much more stochastic and exhibits significant self occlusion, using multiple views (as in Fig. 8) is more appropriate, especially for learning.

4. Evaluation of model learning and model fitting

Evaluation is critical, and PI Barnard has substantive experience evaluating computational methods and developing methodologies for doing so [32, 34, 37, 56, 132]. In this section we first review the relevant aspects of the computational system, and then discuss its evaluation. Demonstrating that the paradigm can be transformative for biology is considered later (§5).

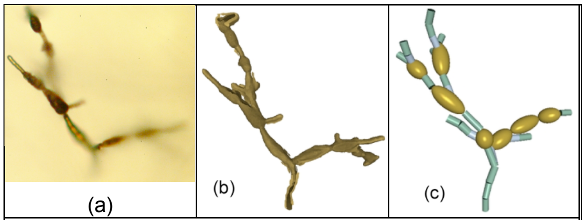

Fig 9. An example of a model instance fit to *Alternaria* using sampling. (a) One image of a stack of 80 taken at stepped focal planes. (b) A rendering of surface elements detected from the data. The structure in this representation is only in the viewer's mind. Conversely, the model fit to the data (c), explicitly captures the structure so that it can be used for quantitative analysis.

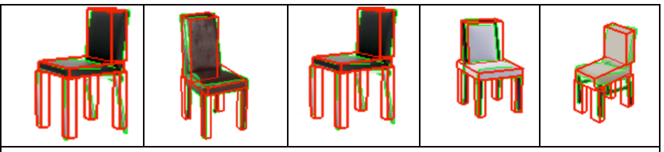

Fig 10. A example from [129] where we learn models for furniture object categories from multiple examples. Here, a chair topology model is learned using single views of eight chairs (five shown). The same model schema (assemblage of blocks) applies to tables, sofas, etc., and the structural consistency of the "species" means that a stochastic geometric model for the category (Θ^{group}) can be learned. Collaterally, specific fits and camera views for the examples are found (shown projected into the images).

Workflow summary. Referring back to Figure 2, model learning will proceed in conjunction with human input through web interfaces that provide rough structural organization (§3.1). This input will be automatically mapped to models (§3.2,§3.3) suitable for inference. Statistical inference (§3.4) over a group of individuals (learning) yields distributional parameters for the group (e.g. probability of branching) that improve upon the rough ones from the initial stage. Having learned some models (say for two species), we can fit them to new images, thereby both determining the class, and fitting the structure which can easily yield numbers that quantify the structure. The quality of the fit is automatically computed by $p(D|\Theta)$. Hence it is easy to clearly flag unexpected new structure for human inspection for possible model correction.

Evaluation of the computational approach will include: **1)** verifying that learned models can be fit to new examples of the same group, and evaluating how the quality of those fits, as measured by $p(D|\Theta)$, compare with visual determination of how well fits projected into images match the image data. This will require some measurements of those quantities by other means; **2**) quantifying the accuracy of structure numbers (e.g., branch angles) as it relates to the fit quality; and **3**) measuring how well models for multiple groups (e.g., a number of *Alternaria* species) support automatic classification through inference.

5. Biological relevance and scientific impact.

A key goal of this project is to demonstrate that we can build pipelines that extract numbers from structure that are biologically relevant. This in turn will demonstrate that the proposed paradigm will open up scientific investigation not previously possible. Validating this vision requires collaborative research on important **model** organisms for which we have libraries of carefully catalogued varieties, ongoing cultivation, molecular data, and available expertise. Thus we will fully develop the methodology using three specific model genera where we have the infrastructure in place: *Alternaria, Arabidopsis,* and *Oryza*. Further, these three domains each push different aspects of the method. *Alternaria* requires a microscope image function, the other two use (multiple) standard images. The structure and sub-structures of *Arabidopsis* and *Oryza* are quite different and are representatives, respectively of dicot and monocot plant groups, a major classification of flowering plants based upon overall growth form (Fig. 1b,1c).

5.1. Quantifying filamentous fungi in the genus *Alternaria*

Fungi in the genus *Alternaria* [116] (Fig. 1a) are found world-wide and are common saprobes important in the degradation of all types of biomass [119]. Moreover, they are important plant pathogens that impact agricultural productivity through crop yield loss, contamination with mycotoxins, as well as through the post-harvest degradation of food [53,100]. These fungi also impact human health directly through invasive mycosis and the production of spore-borne aerial allergens that contribute to the rising global incidence of allergy and asthma [49, 65].

Assessments of *Alternaria* diversity for identification purposes or basic studies in fungal biology and ecology are hampered by a paucity of easily accessible morphological features. The primary characters used for this assessment are variations in the reproductive structure which includes the filamentous spore-bearing hyphae and the reproductive propagules (spores) [133]. Both features have simple structure, yet they possess a number of discernable characters that have slight degrees of variation. Thus, current assessments of diversity are painstakingly derived by close examination and documentation of subtle morphometric differences, often across a population of samples. Additionally, most characters exhibit notable phenotypic plasticity that is impacted by environmental conditions of temperature, light, and substrate, which further complicates accurate identification [134]. Most often these differences are not discernible to the general mycological community and critical assessments are possible only by experts in the field [134]. However, the broad availability of such data, as well as the addition of data currently not extractable, would greatly facilitate most aspects of contemporary fungal research as well expand the utility of fungal morphometrics through linkage to other complex data sets including genomic and metabolomic data.

Modeling and quantifying *Alternaria* structure. We have developed preliminary models for four species grown under standard culture conditions (*A. alternata*, *A. tenuissima*, *A. arborescens*, and *A. gaisen*) and similar modeling efforts will be extended to 8 additional species, all known to be significant plant pathogens and/or produce notable allergenic spores. All species are very closely related with only minor variation at the molecular level, yet exhibit distinguishable morphological differences and morphological plasticity in response to the environment. Co-PI Pryor has assembled an *Alternaria* library of over 3000 accessions, mostly genetically fingerprinted, as a testbed for exploring phenotypic variation along species and ecotypes. Image stacks for each group will be recorded using a brightfield microscope using stepped focal planes producing images with blurring that is handled already discussed (§3.3).

Linking structure to loci. Following model development, extracted morphometrics will be mapped onto DNA-based phylogenies developed from 6 loci commonly used for systematic reconstruction to determine which morphological features are of phylogenetic utility and which are homoplasious, as well as which features have the most robust diagnostic value. Image stacks of species grown under varied culture parameters will be re-fitted to standard models to see how specific morphological features vary in response to specific environmental parameters and to determine if a specific response may also be of phylogenetic utility.

Translational studies. Similarly, spore dispersal potential and pattern will be assessed for each modeled species grown at varied culture conditions using a laminar flow air chamber specifically designed to examine aerial dispersion of fungal spores. Data on lateral spore movement and scatter will then be correlated with extracted model metrics and culture conditions to establish the hierarchy of phenotypic parameters that most significantly impact dispersion potential. This knowledge will be directly applicable to studies of fungal ecology and fitness, as well as well as epidemiological studies related to plant pathology and the incidence of human allergy and asthma. In future studies, we envision development of species-specific microarrays to directly explore gene expression as it relates to morphometrics, phenotypic response, and dispersion potential and these types of studies will be maximally productive only through linkage with our proposed morphometric models.

5.2. Automated characterization of morphological variations in *Arabidopsis* strains and other Brassicaceae species

Arabidopsis thaliana, a member of the plant family Brassicaceae, is a key research model plant that is replete with molecular and genetic resources and the phenotypic variations that exist in these plants are well documented [21, 85]. Quantifying this natural genetic variation could identify the adaptive changes that are important for their survival in natural environment. Many of these traits are controlled by quantitative alleles in the genome that results in subtle variations and they remain uncharacterized due to the lack of approaches to record and quantify them.

Modeling and Quantifying *Arabidopsis* structure. To comprehensively chronicle morphological variations during plant development, we will use 96 Arabidopsis strains [105] and six members of Brassicaceae-*A. lyrata*, *A. aeronosa*, *A. pumilla*, Capsella *rubella*, *Sisimbrium irio* and *Thellungiella halophila* [54]. We will capture the morphology of Arabidopsis and related species by taking images of potted plants growing on a computer controlled turntable as described above (§3.3, Fig 8). Specifically, we will use models fit to *Arabidopsis* images to study in detail two different aspects of development: 1) leaf morphology: venation patterns, leaf number and shape, trichome variation (shape and number), direction of the blade surface and epinasty of the blade; and 2) flower and fruit morphology: petal number and shapes, fruit length, shapes, and elongation rates. Many of the plant behaviors are influenced by external factors such as light. We will monitor the above noted traits at different light conditions such as long (16hours day and 8 hours night) and short day (8 hours day and 16 hours night) conditions. Many of these morphological traits are also influenced by the circadian rhythm; therefore, we will image the plants over a period of time (days) to identify the effect of circadian clock on the manifestation of these traits. It should be noted that these traits have never been comprehensively analyzed and are virtually impossible to quantify using existing methods.

Linking structure to loci. To directly demonstrate the power of morphology data quantification (**Aim 3**), we will use at least one each of the variable traits in leaf and flower/fruit morphology to leverage the enormous genetic resources (e.g. [7, 8]) to identify the genetic loci that control these traits. Our goal will be to answer two specific questions: Which are the genes that affect variation in a specific trait? And, what is the nature of the allelic differences? Co-PI Palanivelu has experience in positional cloning single genes that control phenotypic traits [107, 150]. If these traits are controlled by quantitative trait loci (QTL), we will employ previously developed approaches to map and clone the genetic loci that controls quantitative traits [22].

Translational studies. The imaging data on leaf and flower/fruit development that will be generated as part of this study will be an unprecedented documentation of plant development and will form the foundation for the Arabidopsis community to identify the genetic loci that control these traits. This data will also provide an opportunity to model plant growth and link form and function. Additionally, careful chronicling of plant development will also form the reference using which mutants defective in these processes could be identified in the future.

5.3. Automated extraction of *Oryza* morphology

Rice (*Oryza sativa L.*) is the most important food crop in the world and feeds half the population. It is expected that over the next 25-30 years that the rice dependent population will more than double. Basic and applied plant rice scientists are now tasked with the creation of new and improved rice varieties that yield twice as much rice grains but can grow with less water, fertilizer and need less pesticides and fungicides. In order to achieve the goal of creating a "green super rice" plant molecular biologists need to design a "new plant type" by genetic modification using conventional breeding methods, wide hybridization with wild *Oryza* species, as well as genetic engineering. A highly successful example of such ideotype breeding was the introduction of semidwarf genes – "green revolution genes" – into wheat, rice and sorghum in the 1960s that resulted in short-statured plants with twice the yield potential. Scientists at International Rice Research Institute (IRRI) are working on further modifications of plant architecture to enhance yield potential such as: 1) low tillering; 2) no unproductive tillers; 3) 200-250 grains per panicle; 4) dark green and erect leaves; and 5) vigorous and deep root systems.

The wild relatives of rice (23 *Oryza* species, 10 genome types) offer a virtually untapped reservoir of genes that could be used to modify plant architecture to improve the yield potential of cultivated rice (Fig. 1c). Over the past 5 years, Co-PI Wing and colleagues have developed an unprecedented within-genus comparative genomics platform for 13 wild *Oryza* species under the rubric the *Oryza* Map Alignment Project (OMAP) with an additional goal of generating reference genomes sequences for all major *Oryza* genomes within 2 years. Thus time is ripe to investigate relationships between *Oryza* architecture and crop improvement.

Modeling and quantifying *Oryza* structure. We will image *Oryza* plant growth and development, from seedling to mature panicle stage, for 13 *Oryza* MAP species. Imaging will be similar to that *Arabidopsis*, although imaging rice plants requires using a lager turntable. Greenhouses at UA are equipped to grow different varieties and species of rice year round and under different environment conditions. This work will provide baseline real time data on how these plants grow under differing light, water, and salt stress regimes. We will use this data to update the initial model using Bayesian inference thereby creating a grounded statistical model for morphology that covers the 13 species.

Linking structure to loci. Next we will use these learned models to measure detailed plant growth and development in 2 BC1F2 populations (~ 384 BC1F2 plants / cross) of wild AA genome species (*O. rufipogon* and *O. nivara*) crossed with elite US cultivars Lagrue and M202. Quantitative data from these studies will be used to link plant architecture phenotypes with genomic locations. Such data will permit the positional cloning of genes affecting plant architecture and will be used in crop improvement. More broadly, having such genes in hand will also lead to a comprehensive understanding of cereal plant growth and development.

Translational studies. The data from *Oryza* will be novel and we will leverage co-PI Wing's extensive contacts to disseminate the approach to rice researchers and breeders all over the world. Doing so has the potential to impact rice production relevant to nearly 2 billion people. For example, a recent study identified *submergence rice 1* gene that control plant height [55] and was subsequently used by IRRI to breed tall varieties of rice that can escape submergence from heavy floods in Bangladesh and eastern India, saving millions of pounds in crop losses [1].

6. Rapid scaling to many organisms and on-line dissemination

Our approach has potential for even broader transformative impact because it is designed so that expanding to additional organisms is relatively easy and increasing efficient. We will demonstrate this by using the system to create models for several other organisms and verify that increasingly less support from the system developers is needed as the system matures. Additional organisms will include the filamentous fungus *Aspergillus*, and the crops tomato and maize. We will also consider images from the field for southwest flora including creosote and saguaro. Field images lead to a similar image models as potted laboratory plants, but camera parameter inference does not have the benefit of the strong prior available with turntable based imaging (Fig. 8, §3.3), and image background clutter can be much more significant.

Dissemination. We will develop an on-line interface that makes it easy for everyone to experiment with our approach, download models, and contribute their own models. The integration with education and outreach (§7) will help us ensure that the system is very usable and accessible. We will also disseminate the system as robust software products. The PIs have a solid record of making data, software, and modeling tool interfaces (e.g., for *Alternaria*) available [31,36,116,141,143,145,161]. We have budgeted for a full time scientific programmer to enable good software development, both to support research activities, and to leverage that effort to provide software worthy of extended use, and additional novel use, by the community.

7. Integrating education, outreach, and training

Research students at all levels will be integral to the project, and the PI's have excellent track records in mentoring students. This project will provide unique opportunities for students across multiple-disciplines to interact. The educational goals will be further enhanced by the uniquely supportive environment for interdisciplinary study at UA. The PIs have memberships in multiple UA Graduate Interdisciplinary Programs (GIDPs), including Statistics, Applied Math, and Cognitive Science, the BIO5 institute [3], and the iPlant collaborative [15].

Advanced graduate training. The computational PIs (Barnard, Lin) are among four leads of the Uncertainty Quantification Group (UQG) [11], which provides a context for engaging math and CS graduate students in modern approaches to stochastic systems through regular informal meetings, more formal colloquia from external speakers, and workshops (e.g., [14]).

Training high school teachers. This project will recruit and train high school teachers through the AZ-START (Arizona Science Teacher Advancement and Research Training) program based at UA [2]. During the summer teachers work with life science research mentors full-time, gaining hands-on research experience with which to develop lab exercises for their high school biology classes. Our specific goal is inspire and facilitate the incorporation of quantitative image-analysis methods in those lab exercises. The teachers will perform experiments and acquire images, and work with computational researchers to analyze them.

Increasing undergraduate participation in research. The PIs have an excellent track record of mentoring undergraduate research students. Collectively, since 2001, we have worked with 45 undergraduate research students leading to many publications, abstracts, and on-line demos [35, 38, 39, 72, 86, 110, 131, 140-142, 147, 149, 162]. This project has many excellent opportunities for undergraduates to join an exciting new interdisciplinary direction. To facilitate, we will apply for REU funding, and collaborate with the existing UA Undergraduate Biology Research Program (UBRP) [13] and Minority Access to Research Careers (MARC) [10].

On-line modeling. The on-line access to the modeling system described above (§0) will provide broad access to biological structure modeling in the novel context of automated image interpretation. We envision educational users and citizen scientists taking pictures, creating models to match them, and experimenting with the inferring the structure of their examples. We will integrate explanatory text and web links to explain the underlying computational methods.

Synergy with other activities. This project will enhance our related outreach programs including our Integration of Science and Computing (ISC) Summer Camp for middle school students in targeted groups [144], EOT activities with iPlant [15], and the UA KEYS [6] program.

8. Results From Prior N.S.F. Support (last 5 years)

PI Barnard is PI on a recently awarded CAREER grant (Learning Models for Object Structure, #0747511, $450K) which is complementary to this project. Here we developed a method for learning the category topology of furniture objects from single view images of multiple examples (Fig. 10, [129], 2nd paper under review). This award also supported 4 instances of our Integration of Science and Computing Summer Camp for middle school students [144]. PI Barnard serves as a faculty advisor for the iPlant Collaborative [15]. Previously, PI Barnard worked on a subcontract to NSF funding for the Large Synoptic Survey Telescope project (LSST) (#0551161; PI William Smith) on tracking asteroids [33, 87, 113].

Co-PI Palanivelu and **PI Barnard** collaborate on work supported by IOS# 0723421 (Characterization of Pollen Tube Repulsion in *Arabidopsis thaliana*). Using map-based cloning and TAIL-PCR procedures, we identified the gene (LORELEI, AT4g26466) that was defective in a mutant we previously isolated in a screen [150]. LORELEI encodes a putative GPI-anchor containing protein with unknown function. We demonstrated that LORELEI has a role in pollen tube repulsion, pollen tube reception, double fertilization and early seed development. Barnard's role is the development of statistical models for tube tracking and pollen tube ovule interaction (one paper under review [41], a second manuscript [42]).

Co-PI Pryor has been supported on the DEB-supported project "Collaborative Research: Integrating Morphological and Molecular Systematics of *Alternaria* and Related Fungal Genera" **[**NSF # 0416283 ($264,654) 8/1/04-7/31/08;], and is currently supported on the DEB project "Estimating speciation/reticulation boundaries in asexual *Alternaria*: a genomics approach" [NSF #0918758 ($468,625) 7/1/09-6/30/12] . Significant advancements in *Alternaria* systematics have resulted from this effort by primary project participants and project collaborators. These findings are summarized in 12 publications [26, 28, 45, 69, 70, 72, 110, 117, 118, 126, 130, 131], five manuscripts in review [71, 73, 108, 109, 147], two MS thesis [27, 125], 16 abstracts presented at national and international meetings, and on-line dissemination [116, 141]. **Co-PI Pryor** is also co-PI on a Microbial Observatory project, MO: Kartchner Caverns: Habitat Scale Community Structure and Function in Carbonate Caves (MCB #0604300, $1,374,786, 10/1/06-9/30/11), resulting in two manuscript [75] and 7 abstracts [89].

Co-PI Wing: 1) The *Oryza* BAC Library Project: PI Wing, Co-PIs Soderlund, Tomkins, Jackson. Award #0208329 (9/02-9/04: $600k). The *Oryza* Map Alignment Project: PI Wing, Co-PIs Jackson, Stein, Ware. Award #0321678 (10/03-09/08: $9.7M). Objective: To establish a model system to study cereal genome evolution at the genus level. Primary objectives and outcomes describe above. **23 Pubs (+ 3 submitted)** [23-25, 47, 66, 76, 77, 80, 81, 91-93, 112, 120, 121, 128, 146, 148, 158-160, 166, 167]. **2)** Sequencing Rice Chr3 and Chr10: PI Wing, Co-PIs McCombie, Wilson, Soderlund, Jiang. NSF Award #9982594 (9/99-9/2004: $3.1M); USDA-CSREES Award #99-3517-8505 (9/99-9/2004: $3.1M). **11 Pubs**. **3)** The Maize Genome Sequencing Project: PI Wilson, Co-PIs Wing, Ware, McCombie, Schnable. Award #0527192 (11/05-10/2010: $29M [$3M to Wing]). Objective: To sequence the maize genome using a BAC-by-BAC strategy. Specific objectives: select a MTP of BAC clones (~16,000 MTP BACs & ~1,000 gap-filling BACs) from the maize physical map (AGI); construct shotgun clone libraries from all MTP BACs; sequence two 384-well plates (~6x coverage) from each BAC; prefinish each BAC; manually finish all "unique" regions of each BAC; and annotate the maize genome (CSHL). **1 Pub**. **4)** Sequencing of Chr3 Short Arms from the AA, BB, CC, BBCC Genomes of Wild Relatives of Rice for Comparative Functional and Evolutionary Genomics: PI Wing, Co-PIs: Rounsley, Jackson, Stein. Award #0638541 (10/06-9/09: $2.7M). Objective: To produce a high-quality community sequence resource for Chr3S of the AA, BB, CC and BBCC genomes of *Oryza* for comparative genomics analysis using MTP tiles of BAC clones derived from the OMAP project on a conventional Sanger platform. Progress: 1) All five Chr3S arms have been sequenced and loaded into the Gramene database; 2) Sequence annotation and analysis is currently underway.

Co-PI Lin is sole PI on a recently awarded grant, "Computational Analysis of Large Dynamical Systems," (NSF DMS-0907927). He was previously awarded an NSF Postdoctoral Fellowship (DMS-0303489).

Coordination Plan

1. Participants

All participants are at the University of Arizona. All PIs/co-PIs will work together to optimize interdisciplinary research and further the educational and outreach goals. All PIs have worked with at least one other of the PIs for several years.

Kobus Barnard, Computer Science (PI).

This project represents a key research direction for the PI's career for the next five years. His project role will include overall project steering, management and integration. His technical role will focus on computational and interdisciplinary issues. He will lead the computational research including the development of the model schema sketching capability, the translation to internal model representation, the components of that internal model, the imaging model, and, in and tight collaboration with co-PI Lin, the inference engine. He will supervise two students, one programmer, and co-supervise other students and collaborating postdocs as needed.

Kevin Lin, Math (co-PI).

Co-PI Lin will focus on primarily on mathematical issues, and secondarily on computational ones. He will help recruit math / applied math students into the project and (co)supervise them and CS students on mathematical issues. Lin will lead work on theoretical issues in sampling for structure inference such as the use of non-reversible moves.

Barry Pryor, Plant Sciences (co-PI).

Co-PI Pryor will provide image data from his extensive collection of *Alternaria* species, strains, and ecotypes, collaborate on the development of *Alternaria* models, and investigate linking quantified structure with molecular and environmental data. He will supervise one student studying imaging and morphogenesis in fungi.

Ravi Palanivelu, Plant Sciences (co-PI)

Co-PI Palanivelu will provide access to diverse strains and close relatives of *Arabidopsis* and help link the quantified structure of the organisms to molecular and environmental data. Palanivelu will co-supervise (with Wing) a PhD student focused on the application of quantified phenotype to plant biological problems.

Rod Wing, Plant Sciences (co-PI)

Co-PI Wing will provide diverse domestic and wild rice plants from the genus *Oryza* for imaging and analysis, work tightly with the computational PIs and students on developing *Oryza* models, and will lead the demonstration of biological relevance in the case of *Oryza*. The focus will be on linking genetic data to investigate how the methods developed in this project can impact crop breeding. Co-PI Wing will also help arrange imaging of several other crop plants grown under laboratory conditions. He will co-supervise one PhD student focused on the application of quantified phenotype to plant biological problems, and in specific, on imaging *Oryza* and linking *Oryza* architecture to genome loci.

Research Associate

We will recruit a research associate who will have key duties in assisting project execution, overall software system design, and robust software development. This person will also contribute to algorithmic development in collaboration with the computational grad students.

2. Management across departments and cultures

All PIs have a strong record of collaborating with each other, and other researchers in multidisciplinary projects. PI Barnard has long term, ongoing, collaboration with co-PI Pryor (7 years), and co-PI Palanivelu (5 years). PI/Co-PIs Barnard and Lin have been collaborating on the establishment and running of the University of Arizona Uncertainty Quantification Group, including running a workshop in spring 2008. Co-PI Wing has actively collaborated with Barnard on the formation of the iPlant collaborative and on the Integration of Science and Computing summer camp [144] (as has co-PI Pryor and co-PI Palanivelu). In addition, Wing and co-PI Pryor have on-going collaboration (2-3 years) on the genetic characterization of microbial communities in their Kartchner Cavern Microbial Observatory project. Thus, communication among key project personnel has already been established on a number of levels, which will ensure rapid initiation of the CDI project.

Project management meetings. The PIs and the research associate will meet every month on project management issues such as mid and longer term goals and progress towards them.

Multidisciplinary group meetings. We will have multi-disciplinary weekly group meetings for all project participants. These meetings will integrate important themes: 1) multi-disciplinary brain-storming, discussion, and presentation, with significant participation by students, with the goal of breaking down inter-disciplinary communication barriers and; 2) working towards computational solutions that apply broadly, and the development of tools that are easy to use and make sense to domain scientists and outreach participants.

Software group meetings. The computational participants will meet weekly to ensure efficient development of software components.

Campus wide workshops. Twice in the first academic year and once a year thereafter we will have broadly accessible campus wide workshops. This will be social, immersive, and reflective experience where we will disseminate project ideas and receive valuable feedback and opportunities for further collaboration and outreach. We expect excellent turnout as this is a very important topic for our highly interdisciplinary and strong biological research campus. The workshops will feature presentations from the PIs that lay out the challenges and point out common themes and more specific research presentations by students and postdocs. These presentations would be followed by breakout sessions of smaller groups to exchange ideas, answer questions, and further cultivate a sense of community.

3. Research work schedule

Education, outreach, and training activities will be ongoing and integrated with the following activities. Part of each year's software development activities will include putting tools on-line for educational purposes. Teachers will be trained each summer. Students will be involved at all levels and will contribute greatly to the project goals.

Y1	Integrative algorithm work	• Study the automated translation of grammar models. • Theoretical analysis of the sampling problem for identifying structure that might be exploited by non-reversible sampling and other proposed methods.
	Integrative software development	• Design and develop the software framework to support both modeling and inference. Software developed in preliminary work will provide initial use cases, but it is critical to establish good design at the onset • Implement a reversible jump / stochastic dynamics sampler for both imaging modalities to provide a baseline for comparison. • Release first version of general web based system that will only effectively support *Alternaria*. This will be of interest to the Alternaria research community and provide us with valuable feedback.
	Alternaria	• Setup system for collecting large quantities of Alternaria images • Develop models for more Alternaria species. • Verify the system on dozens of images of multiple species using identification and visual inspection. • Design morphometric experiments for year 2 and beyond.
	Arabidopsis	• Acquire large quantities of images of Arabidopsis • Develop models for more Arabidopsis plant growth and development
	Oryza	• Initial user modeling tool for Oryza. • Prototype translation from user representation to internal representation for inference (a use case of this capability). • Testing of inference for this image model using synthesized *Oryza* data. • Initiate *Oryza* data collection.

Y2	Integrative algorithm work	• Plant sub-part modeling including modeling damage from insects or physical folding or wilting. • Data driven proposal mechanisms • Analytic methods for the core samplers focusing first on the microscope stack image model. • Theoretical study of non-reversible sampling on graphs relevant to this domain. Implementation of a prototype on simplified data.
	Integrative software development	• Complete core modeling user interface with respect to stems and leaves including palates for many leaf shapes leaves and support of stem/branch curve modeling. • Implement the first version of automatic translation grammar models. • The next release of the on-line system will contain capability for several diverse organisms including Alternaria and Oryza.
	Alternaria	• Prototype smaller scale morphometric experiments. • Testing of structure understanding pipeline for a laboratory setting in the case of Alternaria images captured using a motorized microscope and automatically saved to the appropriate file systems.
	Arabidopsis	• Acquire large quantities images of other species related to Arabidopsis. • Develop plant development models for these species. • Characterization of morphological variations among Arabidopsis strains
	Oryza	• Initial fitting of a simple rice species to data. • Design morphometric experiments for year 3 and 4.

Y3	Integrative algorithm work	• Integration of non-reversible sampling into the inference engine. • Integration of time (or life stage) as a variable into the system. • Data driven proposal mechanisms (continued—now these need to be automatically inherited from the model). • Analytic methods for the core samplers (continued—now focusing on the 2D projection imaging model).
	Integrative software development	• Fully parallelize and optimize the inference engine • Explore gains possible by GPU programming and alternative architectures, and provide library components for doing so. • The third release of the on-line system will support a wide range of organisms, including ones quite unlike our test cases.
	Alternaria	• Linking extracted morphometrics to DNA based phylogenies. • Demonstrate accurate species identification.
	Arabidopsis	• Clone loci that control morphological trait variation in Arabidopsis
	Oryza	• Scale up Oryza modeling and fitting to more species and specimens. • Initial Oryza quantification experiments.
	Unified modeler	• Validate the unified modeling approach by adding models for tomato, creosote, maize, and saguaro.

Y4	Algorithm work	• Theoretical analysis of non-reversible jump moves. • Add flowering models. • Further integration of age into models.
	Software	• Final dissemination of efficient, robust, and usable software.
	Alternaria	• High throughput experiments on spore dispersal.
	Arabidopsis	• Confirm identification of genes that control morphological trait by performing complementation experiments
	Oryza	• Ongoing modeling and fitting. • Second phase of Oryza quantification experiments.
	Unified modeler	• Validate that on-line users can rapidly prototype novel plants with substructures that are variants of ones already available.