MINE: a new way to design genetics experiments for discovery

Torres[1], I., S. Zhang[1], A. Bouffier[1], M. Skaro[1], Yue Wu[2], L. Stupp[3], J. Arnold[1], Y. A. Chung[4], & H.-B Schuttler[5]

Target Journal: *Briefings in Bioinformatics*

[1]Institute of Bioinformatics, University of Georgia
[2]Department of Genetics, Stanford University
[3]Genetics Department, University of Georgia
[4]Plant Biology, University of Georgia
[5]Physics and Astronomy, University of Georgia

Corresponding author: Jonathan Arnold, arnold@uga.edu

Abstract
The Maximally Informative Next Experiment or MINE is a new experimental design approach for experiments, such as those in omics, in which the number of effects or parameters p greatly exceeds the number of samples n ($p > n$). Classical experimental design presumes $n > p$ for inference about parameters and its application to $p > n$ can lead to over-fitting. To overcome $p > n$, MINE is an ensemble method, which makes predictions about future experiments from an existing ensemble of models consistent with available data in order to select the most informative next experiment. Its advantages are in exploration of the data for new relationships with $n < p$ and being able to integrate smaller and more tractable experiments to replace adaptively one large classic experiment as discoveries are made. Thus, using MINE is model-guided and adaptive over time in a large omics study. Here, MINE is illustrated on two distinct multi-year experiments, one involving genetic networks in *Neurospora crassa* and a second one involving a Genome Wide Association Study or GWAS in *Sorghum bicolor* as a comparison to classic experimental design in an agricultural setting.

Key Points

- New model-guided and adaptive experimental design for omics experiments called the Maximally Informative Next Experiment or MINE is reviewed for genetics.
- Ensemble methods are reviewed to fit models when the number of parameters p greatly exceeds the number of data points n ($p >> n$).
- MINE uses a fitted model ensemble for model guided discovery when $p >> n$ to select the next MINE, thereby better distinguishing models within a fitted model ensemble.
- MINE is illustrated in nonlinear models of the clock in *Neurospora crassa* in multi-year transcriptomics experiments.
- MINE is also illustrated in linear models for a Genome Wide Association Study or GWAS in *Sorghum bicolor* for multi-year experiments to find the genes underlying biomass.

I. Introduction

The classic approach to experimental design was developed by R. A. Fisher for linear models in 1935 and had a profound effect on all of science[1,2]. Growing out of his work at the Rothamsted Experiment Station, he introduced widely the notion of precision of an experiment, randomization, ways of controlling heterogeneity through blocking and by the use of covariates, and the vast subject of experimental design in the context of linear models[3].

The focus of all of these efforts was not on discovery per se. Rather, the end goal was the precision of estimates and the power to test effects in a controlled experiment with the proper randomization and blocking practices in place. The number of replicates was such that the number of observations (n) was typically much greater than the number of effects (p) being estimated in the model. Unfortunately this is no longer the typical situation of an omics experiment[4], such as a Genome Wide Association Study (or GWAS) as an example. Instead, there may be only n=1943 samples of sorghum accessions but over p=400,000 potential effects of single nucleotide polymorphisms (or SNPs) on the complex trait of interest for an agronomic crop[5]. The methods of classic experimental design are limited to the situation of $n > p$ for inference about model parameters and are not designed for data exploration for hypothesis generation.

While the original goal was the precision of estimates[2], the new goal is discovery. The reason that precision is less important is because the goal of such studies has shifted to the discovery of relationships in the data. The focus on precision of effects can only be addressed in follow up studies when the relevant variables in the experiment have been identified and related. We desire to discover the appropriate nonlinear kinetic models that underly the biological clock at the molecular level[6] as we carry out very expensive sequence of transcriptomic experiments. We wish to discover the relation of plant functional traits to SNPs in the nuclear genome or the assemblage of fungal symbionts in the microbiome most beneficial for plant growth[7-9] using models drawn from systems and population ecology[10]. How can the classic linear models[11] and newer models of systems biology[6] guide a discovery process, in which some GWAS studies have millions of SNPs, such as those on human height[12]?

A new approach to the design of large genomics experiments is introduced, one in which model guided discovery is used adaptively[13] over time to find the variables that matter, considering a system in which the number of potential effects in the system (p) far exceeds the number of observations (n) on the system. The methodological approach utilizes ensemble methods[14] drawn from statistical physics[15] and, ultimately, Boltzmann's 19nth Century work[16]. The particular ensemble method explored here for model-guided discovery is called MINE, which stands for Maximally Informative Next Experiment[17].

II. Ensemble Methods

Ensemble methods were developed in the 19nth century by Boltzmann to describe the motion of an ideal gas[16]. In this situation there is an Avogadro number (A) of particles in a one liter box, but only 3 measurements are made: temperature, pressure, and volume. How is the motion of A particles described with 6A degrees of freedom each described with only three measurements?

With so little data, the data did not strongly support just one model. Boltzmann's solution was to give up on identifying one best model, but rather to make predictions from an ensemble of models[15]. Omics experiments face exactly the same problem[14], but the paucity of data with respect to the complexity of the model is not as severe as in the problem Boltzmann faced. The interest may be in identifying the dynamics of genes and their products in carbon metabolism[14] or the biological clock[18], but the genetics dictates that only a limited number of samples at different time points can be made to identify the system, while there are many parameters required to describe the system. For example, the number of measurements at different time points on the biological clock may be on the order of 60,000 measurements (n), but there are over 90,000 rate constants and initial conditions (p) in the model that must be estimated[19,20]. Much as for an ideal gas, averaging over the ensemble allows for detailed predictions about complex biological systems, such as the clock.

A simple example is used to illustrate the approach of ensemble methods. The first step is writing down the model specification for the measurements. There are n measurements $Y = (y_1,..., y_n)$ drawn from an unknown distribution parameterized by p parameters in $\theta = (\theta_1,...,\theta_p)$, and some of these parameters are ancillary parameters, such as α below, in which there is less interest. In addition, the variables $U = (u_1,...,u_n)$ describe the experimental conditions. For example, the list U might specify the SNPs used in a GWAS field trial. Then the model specification would take the form:

$$P(Y \mid \theta,U) = C\exp\,(-H(Y \mid \theta,U)),$$

where C is a normalization constant chosen to make the integral over the data Y to equal 1. The quantity $H(Y \mid \theta,U)$ is known as the Hamiltonian. Ideally the distribution would be observed directly, but in practice what is available are sample moments of the data Y.

Since the goal is to identify a model θ supported by the data Y, a change of viewpoint is needed. As in the method of maximum likelihood[21], the model specification $P(Y \mid \theta,U)$ is viewed as a function of the model parameters θ, and the data Y and experimental conditions U are taken as fixed:

[0] $$Q(\theta|Y,U) = \Omega^{-1}P(Y \mid \theta,U),$$

where Ω is a normalization constant chosen to make the integral over all parameters θ in the parameter space equal to 1. This normalization constant is only a function of the data Y. The magnitude of the ensemble $Q(\theta \mid Y,U)$, or $Q(\theta)$ for short, is larger when the model θ is more supported by the data Y. It may be useful to think of the ensemble $Q(\theta|Y,U)$ as a posterior distribution to the model specification $P(Y \mid \theta,U)$ with the two functions, $P(Y \mid \theta,U)$ and $Q(\theta|Y,U)$, connected by Bayes Theorem[22].

The ensemble $Q(\theta|Y,U)$, or $Q(\theta)$ for short, is the collection of models θ consistent with the available data Y. Model-averaging with respect to the model ensemble $Q(\theta)$ allows predictions about the system's behavior. Instead of identifying one model θ, a distribution of models $Q(\theta)$ is identified. With the number of parameters p being vastly greater than the number of data points n,

predictions can still be made and tested with respect to averages computed from the ensemble $Q(\theta)$.

Monte Carlo Methods are used to identify the ensemble $Q(\theta)$[15,23] because the model specifications are complicated[18,24]. A simple example will illustrate how this is done. Take the Hamiltonian viewed as a function of θ as having the following simple form:

$$H(\theta) = \beta[-\theta^2 + \alpha\theta^4)].$$

The model parameter θ is the one we are truly interested in, and the remaining parameters α and β are ancillary. A graph of the ensemble $Q(\theta) = \exp(-H(\theta))$ is shown in Figure 1d. There are two maxima in the ensemble or equivalently, two minima in the Hamiltonian. The goal is to reconstruct the ensemble $Q(\theta)$ by Monte Carlo for prediction.

In this example, we are in the perfect world in which the ensemble, or equivalently the Hamiltonian, is observed from 10,000 values after each move in the Monte Carlo experiment. To reconstruct the ensemble $Q(\theta)$ by Monte Carlo at each move a new model parameter θ' is drawn from the ensemble $Q(\theta)$ when the current proposal is the model parameter θ. The goal is to move into a region of the parameter space which is well supported by the ensemble $Q(\theta)$ in the equilibration phase. Once equilibrated many 1,000s or 10,000s of models are accumulated that are well supported to reconstruct the ensemble from the sample histogram of these θ-values[18]. The question remains how to choose the well-supported θ-values.

One greedy approach to moving in the parameter space is to draw a model parameter θ' and proceed up hill using some procedure like steepest ascent to climb the hill(s) in the ensemble. As shown in Figure 1, this might lead to a local maximum. In fact in Figure 1 there are 2 such maxima. To avoid local maxima, a model parameter θ' is drawn randomly from the ensemble $Q(\theta)$, being greedy when there is an improvement in the ensemble probability, *i.e.*, $Q(\theta') > Q(\theta)$ or equivalently $H(\theta') < H(\theta)$, but occasionally when $Q(\theta') < Q(\theta)$, move downhill anyway. The occasional downhill move may allow escape from a local maximum. In practice in systems biology it may be more appropriate to think of the ensemble surface as gently rolling hills as on a golf course because the data are limited ($n < p$).

Metropolis and colleagues[25] developed a stochastic search procedure in statistical physics for this and now many other optimization problems[23]. The probability of a move is:

$$p = \min\left(1, \frac{Q(\theta')}{Q(\theta)}\right).$$

The probability p of a move from $\theta \rightarrow \theta'$ occurs with probability 1 if the proposed move takes us up hill, but if the proposed move takes us downhill, then the probability of a move downhill decreases with the amount of drop from $Q(\theta) \rightarrow Q(\theta')$. The sequence of moves are made 10,000 or more times to move into a region of the parameter space well supported by the data during the equilibration phase.

In the equilibration phase the inferred ensemble converges to the true ensemble known as the target distribution (Figure 1). The Monte Carlo search for this simple model is successful in the

reconstruction in less than a 1,000 moves and is surprisingly quick, equilibrating in less than 20 moves. A video displays the reconstruction process (video S1). Once equilibration is achieved, another sequence of Monte Carlo moves called the accumulation phase is used to build the target distribution. In this simple example only a 1,000 moves are needed to carry the ensemble identification into the accumulation phase.

In practice a sweep is introduced to describe the number of moves taken to visit each model parameter on average once. A standard equilibration run and an accumulation run is 40,000 sweeps, which will vary in practice with the complexity of the model[18,19].

III. MINE

Once an ensemble method produces a collection of models supported by the data, then it is possible to make predictions from the ensemble distribution about the next experiment. By averaging some variable of interest over the models in the ensemble distribution $Q(\theta \mid Y,U)$, a prediction can be made, given the current data Y and the experimental conditions U. For example, Y might be plant biomasses measured in year 1 of a 5 year GWAS experiment to identify SNPs to predict biomass in sorghum with a certain collection of SNPs from the Bioenergy Accession Panel (BAP)[26]. The question is what accessions are to be used in year 2 to specify design U. In year 1, 79 accessions are measured in the GWAS and can be used to inform the SNP choice in year 2.

One way to make this choice is to select experimental conditions permitting us to distinguish the models in the ensemble $Q(\theta \mid Y,U)$ identified from Year 1. The best way to distinguish experimentally two models randomly chosen from the model ensemble is if the predictions $F(\theta,Y)$ of each model ($\theta = \theta_1 \; or \; \theta_2$) are orthogonal as shown in Figure 2. For experiment 1 on the left, the predictions of the two selected models θ_1 and θ_2 are correlated and are harder to distinguish under experimental conditions U_1. The same two models under experimental conditions U_2 are easier to distinguish - model θ_1 is easily tested again model θ_2. The goal of a MINE criterion is then to support making "the angle" between the two predictions of a random pair in the ensemble as large as possible on average in year 2 as a function of the experimental conditions U and current ensemble $Q(\theta \mid Y,U)$ identified from the data in year 1.

There are two standard ways to measure the associations between the predictions[6]. One is by the covariances between the predictions of the data Y (MINE by Covariance Ellipsoid Volume); the other is by the correlations between the predictions of the data Y (MINE by Correlation Ellipsoid Volume). There are a variety of reasons for advocating the use of MINE by Correlation Ellipsoid Volume[6]. One of the main reasons is that when there are a large number (p >>n) of almost linearly dependent observations as found in practice, it would be highly desirable to emphasize the new directions in the data Y as done by Correlation Ellipsoid Volume. The new directions in Y in the next year depend on the choice of design matrix X. Denote by E the correlation matrix between the components of Y. The MINE Correlation Ellipsoid Volume is then a determinant (det):

$$V(U) = \det\,(E(U)).$$

When the predictions are on average highly correlated (Figure 2A), the determinant is nearly zero. When the predictions are nearly orthogonal (going in new directions) (Figure 2B), the determinant is nearly 1.

A microscope analogy[11] provides insights on how MINE works (Figure 3). MINE is highly analogous to a microscope and its optics. The object in the microscope field described by the data Y is the observed system. MINE, like the optics of the microscope, picks up each component of Y through the prediction $F(\theta,Y)$ about the system. For example, $F(\theta,Y)$ could be the list of predictions of plant biomasses in a GWAS study. The optics ($F(\theta,Y)$) and likewise, MINE, then magnify the predictions to create the image or model of the system (Figure 3).

The microscope has a field of view of the object, which we refer to as the Uncertainty Volume of the new experiment Y. The uncertainty in the observations on the field of view comes from our uncertainty about the optics controlled by θ and in the measurements Y on the object. The optics (predictions) then translate the Uncertainty Volume V(U) in the sample space into an image, the Uncertainty Volume in the parameter space. The result is that an Uncertainty Volume in the sample space (object) is mapped by the optics $F(\theta,Y)$ to the Uncertainty Volume in the parameter space (image).

The magnification applied to the object is adjusted to reduce the Uncertainty Volume in the parameter space (image). Another interpretation of the image quality is given by the determinant det(E(U)). The determinant is a measure of the volume of a parallelepiped defined by the Uncertainty Volume in the Sample Space[27]. The determinant is also a measure of the Uncertainty volume in the parameter space (inside the ensemble). As the magnification knob is twiddled, the clarity of the image (model parameters) is increased and uncertainty, reduced (Figure 3). If the parallelepiped is squashed in the parameter space, less detail from the observations Y in the sample space are being retained in imaging (*i.e.*, model fitting). MINE is doing the focusing and representing the object in higher clarity in the image constructed by the observer using MINE.

*A simple model for predicting hyphal extension colonization by arbuscular mycorrhizal fungi (AMF) in plant roots to illustrate MINE***.** Mixture experiments are used in population genetics[28] and science and engineering in general[29]. Mixture experiments are examples of linear models that are the focus of experimental design[2]. In these designs there is a mixture of treatments in different proportions affecting some dependent variable of interest. Mixture experiments can be used to study how different AMF affect the health of the plant through colonization of the root system. The assembly of the AMF biome in plant roots is a product of choices imposed by the plant genotype[30,31], competition between AMF, ecological drift[9], historical contingency[32], abiotic factors such as Phosphorous (P) and Nitrogen (N) in the soil[33,34], and other factors. Consider three AMF species, S_1, S_2 and S_3, competing for colonization area in the plant roots of sorghum[9] of ONE plant genotype. These AMF are potential partners with the plant in one of the oldest symbioses on the planet[35]. Potentially the plant provides carbon, and in return potentially the AMF hyphal network provide P and N, like an extended root system. The success of this partnership is measured in part by AMF hyphal extension in the roots and the resulting biomass of the plant host[36]. To study this symbiosis the experimenter inoculates sorghum with a mixed population of at least 10% of the coenocyte cells being S_1, at least 15% are S_2 and at least 5% of the coenocyte cells being S_3. The mixed coenocyte inoculum is a co-culture in the plant root cells. Denoting respective spore percentages by u_1, u_2 and u_3, respectively, u_1, u_2 and u_3, are thus constrained by lower limits,

[1] $u_1 \geq u_1^{(lo)} = 0.10, \quad u_2 \geq u_2^{(lo)} = 0.15, \quad u_3 \geq u_3^{(lo)} = 0.05,$

and by the normalization condition

[2] $u_1 + u_2 + u_3 = 1.$

Given [2], only two of the three species fraction values can be freely chosen. In the following, we will use proportions u_1 and u_2 as those two free variables, with u_3 then being determined *via* [2]. Furthermore, the proportions u_1 and u_2 are then subject to upper and lower bounds, resulting from [1] and [2]. When referring, below, to experimenters freely choosing (u_1, u_2 , u_3), it should be understood that these choices must be within the constraints imposed by conditions [1] and [2].

Assume that, by setting appropriate experimental conditions, the experimenter can construct an inoculum with a constant total spore population size, N_c, and constant species fractions, u_1, u_2 and u_3. Assume also that, subject to the foregoing constraints [1] and [2], the experimenter can precisely set the values of u_1, u_2 , u_3 and N_c.

Each of the three AMF taxa can increase its rate of occupancy of the root space in the plant, denoted by θ_1, θ_2 and θ_3, for species, S_1, S_2 and S_3, respectively. The experimenter wishes to determine, or at least impose constraints on, the values of these rates in percent area increase, θ_1, θ_2 and θ_3, by performing a sequence of time series experiments wherein the linear filament extension in a root image, denoted by $y(t)$, is measured as a function of time, t, at certain time points,

$$t_1, \; t_2, \; \ldots, \; t_K.$$

Here, K is the total number of experimental observation time points. Each experiment thus produces a series of observed filament extension amounts, $y(t_k)$ for k=1,2,…, K, denoted by

$$y_1, \; y_2, \; \ldots, \; y_K .$$

That is, y_k is value of $y(t)$ observed at time t_k, with k=1,2,…, K labeling the different observation time points. Each of these experiments is to be performed on a spore population begun with a different combination, (u_1, u_2, u_3), of AMF inoculation fractions. For simplicity assume, however, that the values of the rates of hyphal extension, θ_1, θ_2 and θ_3, remain the same throughout all these experiments, *i.e.*, assume that the hyphal extension rates, θ_1, θ_2 and θ_3, do not change when the experimenter changes the population composition (u_1, u_2, u_3) from one experiment to the next as in a race tube experiment[6,37]. For simplicity we will refer to θ_1, θ_2 and θ_3 as the rates of colonization success.

The extraction of any information about the success rate in root colonization, θ_1, θ_2 and θ_3, from the experimental time series data, y_k, requires, a *mathematical model* which treats the rates θ_1, θ_2 and θ_3, as well as the known *experimental control parameters*, u_1, u_2 and u_3, as input parameters. The model must then use these input parameters to provide a *predicted value* for each experimental observation, y_k, the hyphal extension colonized in a plant root. For a given experimental data point, y_k, we denote the corresponding value predicted by the model by f_k. Obviously, whatever the model predicts depends on the model input parameters, θ_1, θ_2 , θ_3, u_1, u_2, and u_3, that were used to make the prediction. We will therefore often write f_k as a *function* of these input parameters, *i.e.*, as

$$f_k(\theta_1, \; \theta_2 \; , \theta_3, u_1, u_2, u_3) \; ,$$

to make it explicit that f_k is dependent on the assumed values of the rate parameters θ_1, θ_2 and θ_3, and on the given values of the control parameters, u_1, u_2, and u_3, set by the experimenter.

For the scenario assumed here, *i.e.*, for a mixed population of spores from three AMF species jointly producing percent root colonization, X, at constant rates per spore cell type, a simple mathematical model for f_k is easy to construct. Assume that the mixed cell population is

established, and starts producing colonization, at time $t = 0$, with no initial colonization length X being present at that time. Then the total percent colonization X produced by the entire AMF population in the roots, by observation time t_k, is given by:

[3] $$X = f_k(\theta_1,\ \theta_2\ , \theta_3, u_{1,}\ u_2, u_3) \ = \ N_c\, u_1\, \theta_1 t_k \ + N_c\, u_2\, \theta_2 t_k \ + N_c\, u_3 \theta_3\, t_k \ .$$

The percent colonization X can be measured in roots by bright field microscopy[38-41].
To understand this linear model, which is linear in the model parameters θ, recall here that N_c is the total number of AMF in the inoculum, and hence $N_c u_1$ is the number S_1-cells in the inoculum. Hence, $N_c u_1\, \theta_1$ is the rate of increase by all S_1-cells combined producing percentage root area X-contribution. Each spore produces a hyphopodium by which to colonize the root cortex. Recall now that

[4] (Rate of increase in colonization length) × (Time) = (total length colonized).

Hence, the length colonized, by all AMF S_1-cells combined, by time t_k, is $N_c u_1\, \theta_1\, t_k$. Likewise, the length colonized of the roots produced by all AMF S_2-cells and by all S_3-cells, by time t_k, are $N_c u_2\, \theta_2\, t_k$ and $N_c u_3\, \theta_3\, t_k$, respectively. We then obtain f_k , *i.e.*, the predicted total amount of hyphal extension colonized X produced by all cells until time t_k, by simply adding up the foregoing three X-contributions from all three AMF species. The result is eq. [3].

Suppose we have performed multiple experiments, to be labeled by an "experiment index" $\ell = 1,2, ...L$, where L is the total number of experiments. In each experiment, a different AMF species composition ($u_{1,}\ u_2,\ u_3$) was used. To distinguish these $u_{1,}\ u_2$ and u_3, used in the different experiments, we therefore have to label *them* with the additional index ℓ, as $u_1^{(\ell)}, u_2^{(\ell)}$ and $u_3^{(\ell)}$, for $\ell = 1,2, ...L$. Consequently, a different time series of X-data, $y_1,\ y_2,\ \ldots,\ y_K$, was observed in each experiment, and we therefore also have to label the observed data, $y_1,\ y_2,\ \ldots,\ y_K$ with the additional index ℓ, as $y_1^{(\ell)}, y_2^{(\ell)}, \ldots, y_K^{(\ell)}$, for $\ell = 1,2, ...L$. Also assume that each data point, $y_k^{(\ell)}$, has been measured with some experimental uncertainty, quantified by an experimental standard deviation $\sigma_k{}^{(\ell)}$. The χ^2-function (or by another name, the Hamiltonian) is then given by:

[5] $$\chi^2(\theta, U\,) \ = \ \sum_{\ell=1}^{L} \sum_{k=1}^{K} \frac{1}{(\sigma_k{}^{(\ell)})^2} \left[y_k{}^{(\ell)} - f_k(\theta, u^{(\ell)}) \right]^2 .$$

To simplify and compactify the notation, we have introduced here the following abbreviations:

[6] $$\theta := (\theta_1, \theta_2\ , \theta_3)\ ;$$

[7] $$u^{(\ell)}\ := \left(\ u_1^{(\ell)}, u_2^{(\ell)}\ , u_3^{(\ell)}\right) \quad \text{for}\ \ \ell = 1,2, ...L\ ;$$

[8] $$U := (u^{(1)}, u^{(2)}, ...u^{(L)}) = \left(\ u_1^{(1)}, u_2^{(1)}\ , u_3^{(1)},\ \ u_1^{(2)}, u_2^{(2)}\ , u_3^{(2)},\ \ \ldots\ u_1^{(L)}, u_2^{(L)}\ , u_3^{(L)}\right)$$

That is, θ (without subscript) is shorthand for a vector which comprises the rates of colonization θ_1 , θ_2 and θ_3. The $u^{(\ell)}$ (without subscript) denotes the vector of the three AMF species inoculation fractions used in experiment number ℓ, and U is the vector comprising the species fractions from *all* experiments combined. Note that θ does not have an (ℓ)-superscript here because θ_1, θ_2 and θ_3 are assumed to have the same values in all experiments.

Note that $f_k(\theta, u^{(\ell)})$ is the model prediction of hyphal extension, from eq. [3], for $y_k{}^{(\ell)}$, *i.e.*, for the k^{th} time series data point for percent root area colonized observed in the ℓ^{th} experiment. The square of the so-called *residual*, on the right-hand side of [5],

[9] $$r_k^{(\ell)}(\theta,u^{(\ell)}) := y_k{}^{(\ell)} - f_k(\theta,u^{(\ell)}) ,$$

thus measures the deviation of the model prediction $f_k(\theta,u^{(\ell)})$ from the experimental observation of hyphal extension $y_k{}^{(\ell)}$: The larger $\left(r_k^{(\ell)}\right)^2$, the worse, *i.e.*, greater, is the deviation of the model prediction, $f_k(\theta,u^{(\ell)})$, from the observed data point, $y_k{}^{(\ell)}$. By taking the sum of all squared residuals, the χ^2- function in [5] thus provides a composite measure of the overall deviation of the model predictions from the data, for *all* data points on hyphal length colonized combined. In the least-squares fitting approach the "best possible" choice of model parameters is then obtained by finding a parameter combination, $(\theta_1, \theta_2 , \theta_3)$, which minimizes this deviation, *i.e.,* by minimizing $\chi^2(\theta, U)$ with respect to θ_1, θ_2 and θ_3. In the following, let $\theta^{(\mathrm{b})} \equiv \left(\theta_1^{(\mathrm{b})}, \theta_2^{(\mathrm{b})} , \theta_3^{(\mathrm{b})}\right)$ denote that best possible parameter combination which minimizes $\chi^2(\theta, U)$.

Note, in passing, that the squared residuals entering into $\chi^2(\theta, U)$ in [5] are weighted by the reciprocals of the variances, $(\sigma_k{}^{(\ell)})^2$. This means that experimental data points with larger experimental uncertainties carry less weight and have less of an effect on the choice of the optimal, "best match" parameter combination, $\theta^{(\mathrm{b})}$, than data points with smaller experimental uncertainties. In that sense, $\theta^{(\mathrm{b})}$ can be regarded as a 'weighted compromise" between all data points, $y_k{}^{(\ell)}$.

While there are, in principle, many different ways to define an ensemble probability distribution function having these general characteristics, an obvious, simple choice for [0], supported by statistical theory[11], is given by:

[10] $$Q(\theta \mid U) = \frac{1}{\Omega}\, e^{-\chi^2(\theta, U)/2} .$$

The $1/\Omega$ -factor in eq. [10] is a normalization factor, chosen to ensure that the ensemble PDF integrates to a probability of 1. That is, for our model for a mixture experiment with $\theta \equiv (\theta_1, \theta_2 , \theta_3)$, the Ω is chosen such that

[11] $$\int_{\theta_{\mathrm{Lo}}}^{\theta_{\mathrm{Hi}}} \int_{\theta_{\mathrm{Lo}}}^{\theta_{\mathrm{Hi}}} \int_{\theta_{\mathrm{Lo}}}^{\theta_{\mathrm{Hi}}} Q(\theta_1, \theta_2 , \theta_3 \mid U)\; d\theta_1\; \mathrm{d}\theta_2\; \mathrm{d}\theta_3 = 1 .$$

Here, θ_{Lo} and θ_{Hi} denote, respectively, a reasonable lower and upper limit imposed on θ_1, θ_2 and θ_3. Eq. [10] is then to be understood to hold only when θ_1, θ_2 and θ_3 each falls within the interval between θ_{Lo} and θ_{Hi}; if θ_1 or θ_2 or $\theta_3$lies outside of this interval we set $Q(\theta \mid U) = 0$.

Notice that $Q(\theta \mid U)$ in eq. [10] has the desired general characteristics: For very large values of $\chi^2(\theta, U)$, the exponential function $e^{-\chi^2(\theta, U)/2}$, and hence $Q(\theta \mid U)$, becomes very small; for smaller values of $\chi^2(\theta, U)$, $Q(\theta \mid U)$ becomes larger. Hence, θ-choices whose model predictions agree poorly with the experimental data, will have a low probability of being drawn from $Q(\theta \mid U)$; θ-choices whose model predictions agree well with the experimental data, will have a higher probability of being drawn from $Q(\theta \mid U)$.

Given $Q(\theta \mid U)$, we can now calculate, for example, expectation values, variances and histograms of any observable quantity, $A(\theta)$, which the model allows us to predict as a function of θ. Specifically for the expectation value, E[...], and variance, σ^2[...], of such an "observable" $A(\theta)$, we need to calculate:

[12] $\mathrm{E}[A(.)] = \int_{\theta_{\mathrm{Lo}}}^{\theta_{\mathrm{Hi}}} \int_{\theta_{\mathrm{Lo}}}^{\theta_{\mathrm{Hi}}} \int_{\theta_{\mathrm{Lo}}}^{\theta_{\mathrm{Hi}}} A(\theta)\ Q(\theta \mid U)\ d^3\theta$,

with $\theta \equiv (\theta_1, \theta_2, \theta_3)$ and $d^3\theta \equiv \mathrm{d}\theta_1\, \mathrm{d}\theta_2\, \mathrm{d}\theta_3$ for short, and then

[13] $\sigma^2[A(.)] = \mathrm{E}\left[(A(.))^2\right] - (\mathrm{E}[A(.)])^2$.

Here, $\mathrm{E}\left[(A(.))^2\right]$ is obtained, analogous to $\mathrm{E}[A(.)]$, with $A(\theta)$ in eq.[12] replaced by $(A(\theta))^2$.

Within the ensemble approach, $\mathrm{E}[A(.)]$ can serve as a prediction of a representative value of $A(\theta)$, given the experimental control parameters U and prior experimental data, ${y_k}^{(\ell)}$ for all ℓ and all k. However, the ensemble approach also allows us to evaluate the *uncertainty* of that prediction, by way of $\sigma[A(.)]$. Furthermore, with similar expectation value calculations, we can also analyze in more detail the random distribution of $A(\theta)$ by way of histograms of all possible A-values. This would tell us, for example, if the values of $A(\theta)$ have a uni- or a multi-modal distribution, for random θs drawn from the ensemble $Q(\theta)$.

These are just a few examples of what kinds of data analyses and model predictions the ensemble approach itself allows us to implement. In the context of the MINE approach of experiment design, we will have to evaluate certain correlations between pairs of observables, $A(\theta)$ and $B(\theta)$, say. This will require the calculation of expectation values of the general form $\mathrm{E}[A(.)\ B(.)]$, with $A(\theta)$ in eq.[12] replaced by the product $A(\theta)B(\theta)$.

The evaluation of all the foregoing expectation values usually requires numerical techniques to carry out the θ-integrations as in eq. [12]. In general, the θ-space is very high-dimensional, far greater than the θ-dimension of $\dim(\theta)$=3 in our simple model here. Markov chain Monte Carlo methods are then the only approach available to perform the required expectation value calculations efficiently for omics experiments and field studies[15] (see Section II). In Figure 4A is a simulation of the mixture experiment with the application of the ensemble method to the simulated data. The ensemble method converges quite well to the true colonization rates θ.

Assume L prior experiments have already been performed, with experimental control parameter vectors $u^{(\ell)}$, as defined in [7], and observed values ${y_k}^{(\ell)}$, for $\ell = 1,2,...L$ and $k = 1,2, ...K$. The experimental data points, ${y_k}^{(\ell)}$, combined with the corresponding model predictions, $f_k(\theta, u^{(\ell)})$, from eq.[3], define an ensemble PDF, $Q(\theta|\ U)$, *via* eqs. [5] and [12]. The $Q(\theta|\ U)$, in turn, will determine $V(U)$, the predicted uncertain volume of the observables, to be measured in the new experiment(s), as follows:

As a simplest case, assume that we want to design just *one* new experiment, with a new experimental control parameter vector $u \equiv (u_1, u_2, u_3)$. The MINE objective is then to choose to input inoculation proportions u so as to maximize the information content of the new experiment about the rates of production of colonization by hyphal extension (X), by maximizing the predicted uncertainty volume of the observables to be measured. In our simple mixture experiment example, there are K such observables in any experiment: the hyphal extension X-amounts to be measured at times t_k, for $k = 1,2, ...K$. The predicted values for these observed hyphal extensions X are then $f_k(\theta, u)$, as defined by the model eq. [3], for given θ and u. These predicted values for these K observations can be thought of as the components of a vector in a K-

dimensional space, the so-called *observation space* (Figure 3). For a given θ and u, this vector of *predicted observations*, is in the following denoted by $f(\theta, u)$, and given by

[14] $$f(\theta, u) := (f_1(\theta, u), f_2(\theta, u), \ldots f_K(\theta, u)).$$

We can now use the ensemble PDF, $Q(\theta, U)$, to define, in some way, a volume of likely θs in θ-space. If we let θ sweep over that finite volume then, by eq. [14], $f(\theta, u)$ will sweep over some corresponding finite volume (or hyper-surface) in the observation space: the *uncertainty volume* of the predicted observation vector, to be denoted by $V(u)$, for given u, and illustrated in Figure 3.

There is of course no precise prescription of how to define a volume of likely-θ in θ-space, or a corresponding uncertainty volume, $V(u)$, in observation space. That definition is not unique: it requires some arbitrary, but reasonable choices to be made. In the following, two specific possible choices for $V(u)$ will be discussed.

MINE by Covariance Ellipsoid Volume. In the covariance matrix approach, we define $V(u)$ in terms of the uncertainty ellipsoid, constructed from the covariances of the K observable predictions, $f_1(\theta, u), f_2(\theta, u), \ldots f_K(\theta, u)$, subject to the ensemble PDF $Q(\theta, U)$. Let $D_{kj}(u)$ denote those covariance matrix elements, *i.e.*, for $k,j = 1,2, \ldots K$, let

[15] $$D_{kj}(u) := \mathrm{E}\left[f_k(., u)\, f_j(., u)\right] - \mathrm{E}[f_k(., u)]\ \mathrm{E}\left[f_j(., u)\right],$$

with expectation values E[…] defined as in eq.[12]. On general mathematical grounds, the corresponding $K \times K$ covariance matrix, $D(u)$, is symmetric and positive semi-definite. Therefore, D has K real, non-negative eigenvalues, λ_ν; and it has an orthonormal basis of corresponding K-dimensional eigenvectors, $e^{(\nu)}$, with $\nu = 1,2, \ldots K$. That is, for $k,j,\nu,\mu = 1,2, \ldots K$, we have:

[16] $$\sum_{j=1}^{K} D_{kj}(u)\, e_j^{(\nu)} = \lambda_\nu\, e_k^{(\nu)}$$

[17] $$\lambda_\nu \geq 0$$

[18] $$\sum_{j=1}^{K} e_j^{(\nu)}\, e_j^{(\mu)} = \delta_{\nu,\mu}$$

[19] $$\sum_{\nu=1}^{K} e_k^{(\nu)}\, e_j^{(\nu)} = \delta_{k,j}\ .$$

The eigenvalues and eigen vectors are of course dependent upon u, but for notational simplicity we have suppressed that functional dependence, *i.e.*, $\lambda_\nu(u)$ and $e_j^{(\nu)}(u)$, in eqs. [16]-[19].

By eq. [18], the eigenvectors, $e^{(\nu)}$, are orthogonal, *i.e.*, pairwise perpendicular to each other. The eigenvalues, λ_ν, are the variances of the predicted observation vector, $f(\theta, u)$, along the corresponding eigenvector directions. That is, if we take the projection of the vector $f(\theta, u)$ onto $e^{(\nu)}(u)$, *i.e.*, let $p^{(\nu)}(\theta,u)$ denote that projection, with

[20] $$p^{(\nu)}(\theta,u) := e^{(\nu)}(u) \cdot f(\theta,u) = \sum_{k=1}^{K} e_k^{(\nu)}(u)\, f_k(\theta,u),$$

then λ_ν is the variance of that projected $f(\theta, u)$–vector:

[21] $$\sigma^2\left[p^{(\nu)}(.,u)\right] = \lambda_\nu,$$

where $\sigma^2[...]$ is defined as in eq.[14].

The eigenvalues and eigenvectors of $D(u)$ define the so-called "covariance ellipsoid" or "error ellipsoid" of the predicted observation vector, $f(\theta, u)$, in the K-dimensional observation space:

The eigenvectors, $e^{(v)}$, can be thought of as the orientations of the principal axes of the ellipsoid; the standard deviations of the projections $p^{(v)}(\theta,u)$, *i.e.*, $\sigma\left[p^{(v)}(.,u)\right] = \sqrt{\lambda_v}$, are the lengths of the principal semi-axes along the $e^{(v)}$- direction. This ellipsoid serves as our uncertainty volume, and $V(u)$ is given by the product of the semi-axis lengths,

[22] $$V(u) = C_K \sqrt{\lambda_1(u)\ \lambda_2(u) \ldots \lambda_K(u)}\,,$$

where C_K is an unimportant geometrical prefactor,

[23] $$C_K = \frac{2\pi^{n/2}}{n\ \Gamma(\frac{n}{2})}\ ,$$

with $\Gamma(x)$ denoting Euler's gamma function. Eq. [22] can also be written in terms of the determinant of the *D*-matrix:

[24] $$V(u) = C_K \sqrt{\det\left(D(u)\right)}\ .$$

MINE by Correlation Ellipsoid Volume. In the correlation matrix approach, we define $V(u)$ in terms of an uncertainty ellipsoid constructed from the Pearson correlations of the *K* observable predictions, $f_1(\theta,u), f_2(\theta,u), \ldots f_K(\theta,u)$, subject to the ensemble PDF $Q(\theta,U)$. The Pearson correlation matrix elements, denoted by $E_{kj}(u)$, are related to the covariance matrix elements, $D_{kj}(u)$, from eq.[16], by

[25] $$E_{kj}(u) := \frac{D_{kj}(u)}{\sqrt{D_{kk}(u)\ D_{jj}(u)}}\,.$$

Note that $E_{kj}(u)$ can also be written as the covariance matrix of the predicted observations, $f_k(\theta,u)$, normalized by their standard deviations:

[26] $$E_{kj}(u) := \mathrm{E}\left[g_k(.,u)\ g_j(.,u)\ \right] - \mathrm{E}[g_k(.,u)\ ]\ \mathrm{E}\left[g_j(.,u)\ \right],$$

where

[27] $$g_k(\theta,u) := \frac{1}{\sigma[f_k(.,u)]}\ f_k(\theta,u)\ .$$

Therefore, the correlation matrix *E* has the same mathematical properties of symmetry and semi-positivity as the covariance matrix *D*. Analogous, to the covariance ellipsoid constructed from *D*, we can therefore construct a correlation ellipsoid from the eigenvalues and orthonormal eigenvectors of the correlation matrix *E*. Using the volume of the correlation ellipsoid as the uncertainty volume of the predicted observables, we then have, analogous to eq.[22],

[28] $$V(u) = \ C_K \sqrt{\kappa_1(u)\ \kappa_2(u) \ldots \kappa_K(u)}\,,$$

where $\kappa_1(u),\ \kappa_2(u),\ldots \kappa_K(u)$ are the eigenvalues of the correlation matrix *E*. Analogous to eq.[24] we can also write this as

[29] $$V(u) = C_K \sqrt{\det\left(E(u)\right)}\ .$$

The advantages of the volume for the correlation ellipsoid are several. One, V(u) is a measure of linear dependence of the observables, and the greatest gain in information from an experiment is likely to come from structuring an experiment to increase this linear independence of the observables. In fact, V(u) is 1 when the observables are linearly

independent and is 0 when there is some linear dependence in the observables. Two, V(u) has well known statistical properties when the ensemble takes the form of [10][42].
The surface of the MINE criterion in a contour plot is shown for the next mixture experiment (Figure 4B). The MINE experiment involves using ~.25 of AMF1 in the inoculum and ~0.40 of AMF2 in the inoculum to characterize the rates of hyphal extension in the next experiment. As a final note some theorems about properties of MINE have been established for the class of linear models, such as the mixture experiments[11]. Code for the ensemble methods are available on GitHub[43,44].

IV. Application of MINE to RNA profiling experiments to discover the mechanism of biological clocks

MINE was developed specifically for this kind of transcriptomics problem and used to close the loop in the computing life cycle (Figure 5) proposed by Hood and Abersold[45]. Here the application was to transcriptomic experiments to discover the mechanism underlying circadian rhythms, one of the central problems of systems biology[46]. Transcriptomic experiments have a limited number of time points n, but have many 1,000s of genes (and hence parameters (p)) to be identified in the process[6]. While both MINE criteria using the Covariance by Ellipsoid Volume and Correlation by Ellipsoid Volume were developed, only the Correlation by Ellipsoid Volume was utilized in the end in designing the experiments[6].

Transcriptomics was to be used to explore the mechanism of the biological clock in one of the most well studied model systems[47], the filamentous fungus, *Neurospora crassa*. Three major components of the clock mechanism were: (1) *frequency* (*frq)*, the gene encoding the oscillator of the system and a negative regulator[48]; (2) *white-collar-1* (*wc-1*), the gene encoding the light response element and a positive transcriptional activator for the system[49]; (3) and *white-collar - 2*(*wc-2*), a second positive transcriptional activator for the system. Together *wc-1* and *wc-2* encode WC-1 and WC-2 proteins that act as positive elements in the clock through the dimeric complex WCC=WC-1/WC-2, while *frq* encodes a protein FRQ, which acts as the negative regulator for the system. The FRQ protein provides negative feedback to *wc-1* and *wc-2*. All three of these elements appear in single copy in the *N. crassa* genome, but they have homologs in fly and mammalian systems[49,50].

Now consider a series of RNA profiling experiments which were conducted, guided by MINE to choose an informative sequence of experiments. The last in a series of three adaptive experiments guided by MINE involved a choice of whether or not to do a knockdown or overexpression experiment on: (1) *frq*; (2) *wc-1*; (3) or *wc-2*. The conventional wisdom was to mutate the oscillator gene *frq*[51].

The first step in the MINE application is to make predictions for various mutations in the clock mechanism genes using an available ensemble. The RNA profiles of all 11,000 genes were measured at each of 14 timepoints. A subset of these RNA measurements included a total of 14 timepoints on each of the three clock genes so that the prediction vector F had 42 components. Unlike previous models so far considered, the clock model is a nonlinear model (in the parameters describing the model), which specifies a genetic network of nonlinear ordinary differential equations describing the time course of the genes, their cognate RNAs and proteins[18]. The model

ensemble for this clock network was used to predict RNA profiles of *frq*, *wc-1*, and *wc-2* and their correlations under different possible experiments as shown in Figure 2.

With the correlation matrix in hand for the predictions, the MINE criterion based on the correlation volume ellipsoid in [29] was calculated as a function of the degree of knockdown of the three clock genes (Figure 5). The result was surprising. A knockdown of *wc-1* was selected as the MINE experiment and used to identify 2,323 genes responding to the knockdown[6]. The second surprise from this model-guided discovery process was that ribosome biogenesis was under clock control. This was later confirmed in mammalian systems[52].

A better way to do this MINE calculation would have been to use the transcriptomic data on all 11,000 *N. crassa* genes instead of just the three clock genes driving the system. The ensemble methods now exist for the whole genome-scale network with all of its 1,000s of genes and ensembles now exist for the entire clock network[19,20,53]. Using Graphical Processing Units (GPUs) ensemble methods, such as MINE, can now be implemented on a genomic scale with unknown network structure[19,53].

V. Application of MINE to GWAS field studies for AMF/Sorghum project

Consider a GWAS study under way to understand the genetic basis of biomass and AMF colonization in *Sorghum bicolor* using the BAP Panel[26] of 343 plant accessions of varying biomass. There are 232,303 SNPs to characterize each member of the panel after filtering for minor allele frequencies[54]. The focus is on dry weight as a measure of biomass, and AMF colonization is measured using a convolution neural network from imaging AMF in roots[41,55]. At time 0 dry weight data were available on each accession in the panel to construct an ensemble[26]. The GWAS study is running for 5 years[43,44], and in each year MINE is being used to select 79 BAP plant accessions for study. The 79 plant seedlings will be assigned randomly to 79 rows with each row consisting of 9 seedlings of identical genotype. The 79 x 9 block was replicated three times in the field. The goal is to discover a relation between plant biomass as a function of SNPs.

In the field study there are actually a number of plant features that are being measured during the sequence of MINE experiments[44]. These include plant genotype, plant eQTLs, the microbiome, tissue total Phosphorous (P) Nitrogen (N), time of harvest, and other variables relevant to plant health as measured by biomass (Figure 6). These variables used to predict biomass (as well as AMF colonization and AMF community composition) are held together in a causal diagram representing a structural equation model[56]. The standard model for GWAS experiments is the mixed linear model which is a special case of the structural equation model in which some of the independent variables are random with mean zero[57]. For purposes of illustration a mixed linear model is presented below for an adaptive GWAS experiment underway at Wellbrook Farm, Athens, GA.

There are two sets of measured inputs to the GWAS experiment in year 1 at Wellbrook Farm, Athens, GA: 1) fixed variables in the design matrix X, such as block number, harvest time of each plant, N level, P level; 2) random intercepts and slope effects Z for the genotype of each BAP accession on the other hand. The additive genetic variation in plant genotype as it effects

the dependent variable of log dry weight (*e.g.*, biomass) was captured by binning the number of alleles in a given genomic region different from the reference genome using the sum method[58]. The number of SNPs in each chromosomal region was adjusted to ensure that at least each genomic region was 50 kb in size. Since linkage disequilibrium is reported between markers separated by 3.5-35 kb[59], the 50 kb size was chosen to reduce linkage disequilibrium between bins. The result was that the number of bins was 2748 in the 750 Mb sorghum genome with each bin typically having 10-12 genes in sorghum. The number of such alleles in a bin is treated as a continuous random variable with mean 0 and variance component σ_i^2 for the ith accession, and summed over regions to obtain a random effect for an accession. The fact that each random effect is the sum of 2748 small random effects of chromosomal regions makes it plausible that the random effect of accession is normally distributed. The fixed effects of block number and BAP genotype, for example are denoted by the vector β and the random effects, by u.

These fixed and random effects are used to predict some measure of biomass, log dry weight. In the experiment there are a total of n ~ 606 plants (2-3 plants x 3 blocks x 79 rows) in the field to infer the fixed and random effects, which is less than p =2748 fixed effects + 79 variance components. The measurements to be predicted are summarized in the n x 1 vector Y. The last component of the model is the error in biomass, denoted by ϵ_i for the ith plant in the experiment. The biomass measurements are summarized in an n x 1 column vector Y.

In year 1 of the adaptive GWAS experiment there was no evidence of block effects, and N and P applied were not varied. A total of 79 BAP accessions were planted in a randomized complete block design with 3 blocks, 79 plots in each block, 1 genotype randomly assigned per plot (*i.e.*, a row), and 12 replicates per plot. The mixed linear model for this experiment reduced to:

[30] $$Y = X\beta + Zu + \epsilon,$$

where Y is a n x 1 vector of observations on biomass. The n x p matrix X is the values of the p independent variables (describing each bin) with n observations on each of the chromosomal bins. The p x 1 vector β are the fixed effects of each chromosomal region. The n x r matrix Z is the r=2748 normalized number of alleles in each accession on n plants in the field. The r x 1 vector u is a vector of random effects of each accession on biomass. The errors in the dependent variable, biomass Y, are captured in the n x 1 vector ϵ. Three of the assumptions of the model are that: 1) the random effects u are independent of the biomass errors ϵ ; 2) the errors ϵ are normally distributed with mean 0 and variance σ^2; 3) the random effects u are normally distributed with mean 0 and variance $\sigma_{j(i)}^2$ plant I with accession j(i). That is, the assumptions are that the random effects u and errors ϵ are independent and normally distributed with mean 0 and variance-covariance matrix $I\sigma_i^2$and $I\sigma^2$, respectively. In the application of [30] taking Z = X implies that both intercepts and slopes are random as in[41].

Under this model the prediction of biomass is:

$$E(Y) = X\beta.$$

The variance components and heritability are used to calculate the variance-covariance matrix V of the biomass measurements Y:

$$V \equiv VAR(Y) = \sum_{i=1}^{n} Z_i Z_i' \sigma^2_{j(i)} + \sigma^2 I = \sum_{i=1}^{n} X_i X_i' \sigma^2_{j(i)} + \sigma^2 I,$$

where $Z = [Z_1,\ldots,Z_n]' = [X_1,\ldots,X_n]'$. That is, Z_i and X_i are the ith row vectors of Z and X, respectively. Each observation Y_i has such a 1 x n row vector X_i to describe the genetics of its accession. Each term $X_i X_i' \sigma_i^2$ is an n x n block. The variance-covariance matrix is diagonal with p blocks each with the same diagonal elements $\sigma^2_{j(i)}$. The index j(i) is a lookup that returns the variance component of the ith observation as determined from accession j. Plant I has an assigned accession j.

With the assumptions above for the mixed linear model the ensemble Q can be written down as multivariate normal with the θ-vector consisting of the fixed effects β and the variance components:

$$Q(\theta) = \frac{e^{-\frac{1}{2}(Y-XB)' V^{-1} (Y-XB)}}{(2\pi)^{n/2} |V|^{1/2}}$$

In the first use of MINE in a field experiment no fertilizers were applied to the field in 2021 at Wellbrook Farm, Athens, GA[44]. The model reduced to a fixed effects model with the number of alleles in a bin as the set of independent variables using the sum method[58].

The ensemble method was used to fit the models to the published log dry weight data from 3 years from 2013-2015 averaged in Florence, South Carolina[26] to make predictions in the use of MINE[44]. Typically in an omics experiment there is prior published data available, and this should be used when available[6] to initialize the MINE sequence. A total of 1000 equilibration sweeps were done, and then 1000 sets of model parameters were accumulated, each model parameter set separated by 100 decorrelation sweeps. The chi-squared per data point was 6.12 with n = 606 dry weight measurements. As a control the ensemble run was repeated with the only change being 1000 decorrelation steps. The fitted ensemble from all the data simultaneously provides a unified framework for feature selection to avoid overfitting[43,44].

MINE was then applied to the fitted ensemble from Florence, South Carolina to select 80 accessions for use in 2022 for planting at Wellbrook Farm[44]. The MINE method used was the covariance ellipsoid in [24]. The MINE criterion V(u) conceptually involved evaluating det(D) over all possible $\binom{343}{80}$ samples from the BAP panel, which is computationally intractable. Instead as an approximation, the MINE criterion was optimized by evaluating det(D) on all $\binom{343}{3}$ triples drawn from 343 BAP accessions deposited at USDA GRIN in Griffin, GA. The result is shown graphically (**Figure 7**). Details of calculating det(D) in section III begins with calculating the covariance matrix of the predictions for next year's experiment with [15]. The eigenvalues are calculated for the covariance matrix using [16]-[19]. The eigenvalues in turn determine det(D) in [22]. Code for the ensemble methods and the MINE calculations are available on GitHub[44].

A detailed comparative analysis of MINE in GWAS with classic design approaches is available[43,44]. The use of MINE in a series of smaller 3 yearly adaptive experiments is compared directly with a large classic design on the BAP for log dry weight[26]. In this example the MINE experiments made the GWAS feasible with about 8 project participants per year in the field experiment where a larger classic design using all BAP accessions in one year involved over 20 project participants.

As a final note, in the first pass to developing the ensemble methods, such as MINE, for GWAS the random effects were assigned to accessions rather than individual genes to track the replicates across blocks in the field design each year and to keep the model simpler by limiting the number of variance components to the number of accessions rather than the number of bins. The hypothesis to be tested was simply to ascertain whether or not BAP accessions had an effect on AMF colonization. Future work will include more variance components for each chromosomal bin to fit these more complicated ensembles of models[43].

VI. Conclusion

There are other potential domains of application of MINE. One example is computer vision models for large image data sets, such as the 15 Terabyte (Tb) data set of plant root images with AMF structures in the root cells recently reported[60]. As such data sets grow, there are choices of structural categories to annotate for "balance" to build a good classification and segmentation model. Then the ratios of training, validation, and testing sets need to be selected as design parameters, the usual ratios being 8:1:1. Once a criterion for information is decided upon, such as accuracy of classification or F1 or precision as well as for segmentation (*i.e.*, Intersection over Union (IoU))[41], then MINE could be applied. Classification problems are potentially approachable by MINE in the same way as MINE has been applied to phylogenetic trees in taxon sampling[61,62]. The goal would be to select promising categories for further annotation to improve classification and segmentation adaptively as the data resource is expanded.

Other kinds of large Tb data sets are encountered in natural language processing of genomic sequences arising from protein and DNA sequence languages in microbial genomes and microbiomes[63]. The goal is to use natural language models to help predict the functional categories of sequences, most of which are unannotated. Again the same kind design questions arise. What are the most informative functional categories to explore with annotation? How do we go about the training, validating, and testing natural language models for genomic sequences? MINE would presumably help to select functional categories for discovery potential.

MINE is a discovery tool designed specifically for very large genetic data sets and is illustrated in a variety of problems within genetics here. MINE is built upon other ensemble methods[18] that have been developed for fitting models with $n << p$. These ensemble methods of model identification coupled with MINE complete a discovery cycle (Figure 6) for exploring problems in genetics. This discovery cycle has been called computing life[45]. MINE as a discovery tool completes the cycle in both analyzing and designing future costly omics experiments arising in genetics and allows an adaptive approach to solving problems in genetics.

Data Availability

The data underlying this article will be shared on reasonable request to the corresponding author.

Figure 1. Convergence of an ensemble to the target distribution: (a) ensemble after 20 moves; (b) ensemble after 100 moves; (c) ensemble after 1000 moves; (d) true ensemble or "target distribution". In panel (d) is shown the target distribution as the ensemble converges to the true target distribution under a Monte Carlo Experiment (a-c). The starting guess at to the parameter θ was 3. After each of 20, 100, and 1000 moves, 10,000 samples of θ from the resulting distribution were drawn to characterize the ensemble. The ancillary parameters were $\alpha = 0.1$ and $\beta = 1.0$. The plots were created in MATLAB_R2018B (https://www.mathworks.com/products/matlab.html).

Figure 2. Two models can be better distinguished by their predictions in the next experiment if their predictions are less correlated. The predictions of model θ_1 and θ_2 under experimental condition U_1 are the expectations $f_1 = E(\theta_1|Y,U_1)$ and $f_2 = E(\theta_2|Y,U_1)$, respectively. If the two models θ_1 and θ_2 are chosen independently from the model ensemble $Q(\theta \mid Y,U)$, the expectations are calculated with respect to the product density $Q(\theta_1|Y,U)Q(\theta_2|Y,U)$, where U = U_1 or U_2.

Figure 3. MINE is analogous in function to the optics on a microscope. The data Y are the objects in the field of view. The models θ are in the image. The MINE criterion with the predictions $F(\theta,U)$ is the optics. The Uncertainty Volume in the image is the magnification measured by the MINE criterion, V(u) = det(E(U)). From[44].

Figure 4. An ensemble method to identify the mixture experiment's hyphal extension rates θ is carried out on simulated data from the mixture experiment, and MINE is used to choose the next mixture experiment with inoculation proportions u_1 and u_2. (A) Panel A illustrates the ensemble method on simulated data for a mixture experiment of AMF colonizers. The orange lines are the true colonization rates θ. In the Monte Carlo experiment the estimated rates are plotted as a function of sweep, a visit on average once to each of the three rates θ. In the first 3000 sweeps the Monte Carlo experiment is equilibrated to get in the neighborhood of parameters θ that fit the simulated data. In the accumulation phase (last 1000 sweeps) the estimates of θ are accumulated to form the ensemble estimate. (B) Panel B presents the next MINE mixture experiment recommended. The contour plot is of the MINE criterion det(E) as a function of the mixture of spore inoculum proportions u_1 and u_2.

Figure 5. The MINE experiment is a 90% knockdown of the *wc-1* gene. The MINE criterion displayed is the correlation ellipsoid volume det(E(U)), which is graphed as a function of the remaining activity of the three clock mechanism genes. The predictions F are of the log base 10 concentrations over time of *frq*, *wc-1*, and *wc-2* mRNAs over time from the RNA profiling experiments. The mRNA levels were measured at 14 time points over an 8 hour window. The drawing is taken from[6].

Figure 6. A sequence of MINE experiments are being used in a 5 year GWAS experiment to examine the relation between biomass and SNPs in *Sorghum bicolor* using the BAP accessions[26]. MINE is used to select the BAP accessions to be used in each year in order to map AMF colonization and biomass to the sorghum genetic map in a GWAS study. Multi-scale structural equation model (SEM) for the project (center boxes and arrows). Lotka-Volterra community models are nested within the SEM and predict associations that affect biomass. The dependent variable is biomass, and the arrows in the diagram denote causal relationships between independent variables in the SEM. In this model, sorghum genotype is the primary independent variable that correlates with the remaining variables. This conceptual model will evolve continuously using the model guided discovery process of maximally informative next experiment (MINE; outer ring)[6].

Figure 7. The MINE criterion log(det(E)) was used to select 80 accessions for use in a GWAS experiment at Wellbrook Farm, GA in 2022. The top 200 selected triples of accessions are ranked by det(D). From these top 200 triples 80 distinct accessions were selected.

Acknowledgements
This work was supported by DOE DE-SC0021386 and NSF MCB-2041546.

References

1 Fisher, R. A. *The design of experiments*. (Oliver and Boyd, 1935).
2 Fisher, R. A. *The design of experiments*. 6th edn, (Oliver and Boyd, 1951).
3 John, P. W. M. *Statistical design and analysis of experiments*. (Macmillan, 1971).
4 Lasky, J. R. *et al.* Genome-environment associations in sorghum landraces predict adaptive traits. *Science Advances* **1**, e1400218 (2015). https://doi.org:10.1126/sciadv.1400218
5 Hu, Z., Olatoye, M. O., Marla, S. & Morris, G. P. An integrated genotyping-by-sequencing polymorphism map for over 10,000 sorghum genotypes. *The Plant Genome* **12**, 180044 (2019).
6 Dong, W. *et al.* Systems biology of the clock in *Neurospora crassa*. *PloS one* **3**, e3105 (2008).
7 Johnson, N. C., Wilson, G. W. T., Wilson, J. A., Miller, R. M. & Bowker, M. A. Mycorrhizal phenotypes and the Law of the Minimum. *New Phytologist* **205**, 1473-1484 (2015). https://doi.org:10.1111/nph.13172
8 Johnson, N. C., Wilson, G. W. T., Bowker, M. A., Wilson, J. A. & Miller, R. M. Resource limitation is a driver of local adaptation in mycorrhizal symbioses. *Proceedings of the National Academy of Sciences* **107**, 2093 (2010). https://doi.org:10.1073/pnas.0906710107
9 Gao, C. *et al.* Strong succession in arbuscular mycorrhizal fungal communities. *The ISME Journal* **13**, 214-226 (2019). https://doi.org:10.1038/s41396-018-0264-0
10 Johnson, N. C. *et al.* From Lilliput to Brobdingnag: Extending Models of Mycorrhizal Function across Scales. *BioScience* **56**, 889-900 (2006). https://doi.org:10.1641/0006-3568(2006)56[889:FLTBEM]2.0.CO;2
11 Bouffier, A. M., Arnold, J. & Schüttler, H. B. A MINE Alternative to D-Optimal Designs for the Linear Model. *PloS one* **9**, e110234 (2014). https://doi.org:10.1371/journal.pone.0110234
12 Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610**, 704-712 (2022). https://doi.org:10.1038/s41586-022-05275-y
13 Gouveia, G. J. *et al.* Long-term metabolomics reference material. *Analytical chemistry* **93**, 9193-9199 (2021).
14 Battogtokh, D., Asch, D. K., Case, M. E., Arnold, J. & Schuttler, H. B. An ensemble method for identifying regulatory circuits with special reference to the *qa* gene cluster of *Neurospora crassa*. *Proc Natl Acad Sci U S A* **99**, 16904-16909 (2002).
15 Landau, D. P., Binder, K., Landau, D. P. & Binder, K. Monte Carlo simulations at the periphery of physics and beyond. *A Guide to Monte Carlo Simulations in Statistical Physics; Cambridge University Press: New York, NY, USA*, 13-22 (2014).
16 Guggenheim, E. A. *Boltzmann's distribution law*. (North-Holland Pub. Co.; Interscience Publishers, 1955).

17 McGee, R. L. & Buzzard, G. T. Maximally informative next experiments for nonlinear models. *Mathematical Biosciences* **302**, 1-8 (2018). https://doi.org:https://doi.org/10.1016/j.mbs.2018.04.007

18 Yu, Y. *et al.* A genetic network for the clock of *Neurospora crassa*. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 2809-2814 (2007).

19 Al-Omari, A. *et al.* Discovering Regulators in Post-Transcriptional Control of the Biological Clock of *Neurospora crassa* Using Variable Topology Ensemble Methods on GPUs. *IEEE Access* **6**, 54582-54594 (2018). https://doi.org:10.1109/ACCESS.2018.2871876

20 Al-Omari, A. M. *et al.* Ensemble Methods for Identifying RNA Operons and Regulons in the Clock Network of Neurospora Crassa. *IEEE Access* **10**, 32510-32524 (2022). https://doi.org:10.1109/ACCESS.2022.3160481

21 Fisher, R. A. & Russell, E. J. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **222**, 309-368 (1922). https://doi.org:10.1098/rsta.1922.0009

22 Savage, L. J. *The foundations of statistics*. 2d rev. edn, (Dover Publications, 1972).

23 Landau, D. P. & Binder, K. A Guide to Monte Carlo Simulations in Statistical Physics. *Cambridge University Press* (2009).

24 Antoninka, A. J., Ritchie, M. E. & Johnson, N. C. The hidden Serengeti—Mycorrhizal fungi respond to environmental gradients. *Pedobiologia* **58**, 165-176 (2015). https://doi.org:https://doi.org/10.1016/j.pedobi.2015.08.001

25 Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* **21**, 1087-1092 (1953). https://doi.org:10.1063/1.1699114

26 Brenton, Z. W. *et al.* A Genomic Resource for the Development, Improvement, and Exploitation of Sorghum for Bioenergy. *Genetics* **204**, 21-33 (2016). https://doi.org:10.1534/genetics.115.183947

27 Mostow, G. D. & Sampson, J. H. *Linear algebra*. (McGraw-Hill, 1969).

28 Anderson, W. W., Arnold, J., Sammons, S. A. & Yardley, D. G. Frequency-dependent viabilities of Drosophila pseudoobscura karyotypes. *Heredity* **56**, 7-17 (1986). https://doi.org:10.1038/hdy.1986.2

29 Cornell, J. A. Experiments with Mixtures: A Review. *Technometrics* **15**, 437-455 (1973). https://doi.org:10.1080/00401706.1973.10489071

30 Cobb, A. B. *et al.* The role of arbuscular mycorrhizal fungi in grain production and nutrition of sorghum genotypes: Enhancing sustainability through plant-microbial partnership. *Agriculture, Ecosystems & Environment* **233**, 432-440 (2016). https://doi.org:https://doi.org/10.1016/j.agee.2016.09.024

31 Watts-Williams, S. J. *et al.* Diverse Sorghum bicolor accessions show marked variation in growth and transcriptional responses to arbuscular mycorrhizal fungi. *Plant, Cell & Environment* **42**, 1758-1774 (2019). https://doi.org:https://doi.org/10.1111/pce.13509

32 Chappell, C. R. & Fukami, T. Nectar yeasts: a natural microcosm for ecology. *Yeast* **35**, 417-423 (2018). https://doi.org:https://doi.org/10.1002/yea.3311

33 Liu, Y. *et al.* Phylogenetic structure of arbuscular mycorrhizal community shifts in response to increasing soil fertility. *Soil Biology and Biochemistry* **89**, 196-205 (2015). https://doi.org:https://doi.org/10.1016/j.soilbio.2015.07.007

34 Jiang, S. *et al.* Dynamics of arbuscular mycorrhizal fungal community structure and functioning along a nitrogen enrichment gradient in an alpine meadow ecosystem. *New Phytologist* **220**, 1222-1235 (2018). https://doi.org:https://doi.org/10.1111/nph.15112

35 Revillini, D., Gehring, C. A. & Johnson, N. C. The role of locally adapted mycorrhizas and rhizobacteria in plant–soil feedback systems. *Functional Ecology* **30**, 1086-1098 (2016). https://doi.org:https://doi.org/10.1111/1365-2435.12668

36 Oyarte Galvez, L. *et al.* A travelling-wave strategy for plant–fungal trade. *Nature*, 1-9 (2025).

37 Dharmananda, S. *Studies of the circadian clock of Neurospora crassa: light-induced phase shifting*. (University of California, 1980).

38 McGonigle, T. P., Miller, M. H., Evans, D. G., Fairchild, G. L. & Swan, J. A. A new method which gives an objective measure of colonization of roots by vesicular—arbuscular mycorrhizal fungi. *New Phytologist* **115**, 495-501 (1990). https://doi.org:https://doi.org/10.1111/j.1469-8137.1990.tb00476.x

39 Plouznikoff, K., Asins, M. J., de Boulois, H. D., Carbonell, E. A. & Declerck, S. Genetic analysis of tomato root colonization by arbuscular mycorrhizal fungi. *Annals of Botany* **124**, 933-946 (2019). https://doi.org:10.1093/aob/mcy240

40 De Vita, P. *et al.* Genetic markers associated to arbuscular mycorrhizal colonization in durum wheat. *Scientific Reports* **8**, 10612 (2018). https://doi.org:10.1038/s41598-018-29020-6

41 Zhang, S. *et al.* Computer vision models enable mixed linear modeling to predict arbuscular mycorrhizal fungal colonization using fungal morphology. *Scientific Reports* **14**, 10866 (2024). https://doi.org:10.1038/s41598-024-61181-5

42 Muirhead, R. J. *Aspects of multivariate statistical theory*. (John Wiley & Sons, 2009).

43 Torres, I. MINE: Maximally Informative Next Experiment - Genetics Application and Novel Computational Methodology. *PhD Dissertation in Bioinformatics* **University of Georgia** (2024).

44 Torres, I. Z., S.; Bouffier, A.; Schuttler, H. B.; Arnold, J. MINE: MaximallyInformative Nextr Experiment - Towards a new GWAS experiemntal design and methodology. *G3 Genes|Genomes|Genetics* **in revision** (2025).

45 Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929-934 (2001).

46 Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662-1664 (2002).

47 Dunlap, J. C. Molecular bases for circadian clocks. *Cell* **96**, 271-290 (1999).

48 Aronson, B. D., Johnson, K. A., Loros, J. J. & Dunlap, J. C. Negative feedback defining a circadian clock: autoregulation of the clock gene frequency. *Science* **263**, 1578-1584 (1994).

49 Crosthwaite, S. K., Dunlap, J. C. & Loros, J. J. Neurospora wc-1 and wc-2: transcription, photoresponses, and the origins of circadian rhythmicity. *Science* **276**, 763-769 (1997).

50 McClung, C. R., Fox, B. A. & Dunlap, J. C. The Neurospora clock gene frequency shares a sequence element with the Drosophila clock gene period. *Nature* **339**, 558-562 (1989).

51 McDonald, M. J. & Rosbash, M. Microarray analysis and organization of circadian gene expression in Drosophila. *Cell* **107**, 567-578 (2001).
52 Jouffe, C. l. *et al.* The circadian clock coordinates ribosome biogenesis. *Plos Biology* **11**, e1001455 (2013).
53 Al-Omari, A. *et al.* Discovering regulatory network topologies using ensemble methods on GPGPUs with special reference to the biological clock of *Neurospora crassa*. *IEEE Access* **3**, 27-42 (2015).
54 Brenton, Z. W. *et al.* Species-Specific Duplication Event Associated with Elevated Levels of Nonstructural Carbohydrates in Sorghum bicolor. *G3 Genes|Genomes|Genetics* **10**, 1511-1520 (2020). https://doi.org:10.1534/g3.119.400921
55 Evangelisti, E. *et al.* Deep learning-based quantification of arbuscular mycorrhizal fungi in plant roots. *New Phytologist* **232**, 2207-2219 (2021). https://doi.org:https://doi.org/10.1111/nph.17697
56 Kendall, M. G., Stuart, A., Ord, J. K. & O'Hagan, A. *Kendall's advanced theory of statistics*. 6th edn, (Edward Arnold ;
Halsted Press, 1994).
57 Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* **42**, 355-360 (2010). https://doi.org:10.1038/ng.546
58 Romagnoni, A. *et al.* Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data. *Scientific Reports* **9**, 10351 (2019). https://doi.org:10.1038/s41598-019-46649-z
59 Wang, Y.-H. *et al.* Genetic Structure and Linkage Disequilibrium in a Diverse, Representative Collection of the C4 Model Plant, Sorghum bicolor. *G3 Genes|Genomes|Genetics* **3**, 783-793 (2013). https://doi.org:10.1534/g3.112.004861
60 Zhang, S., Bourlai;, T. & Arnold, J. MycorrhisEE: a high-resolution image dataset for deep learning based quantification of Arbuscular Mycorrhizal Fungi. *Proceedings of IEEE "Big Data"* **P285** (2024).
61 Townsend, J. P. & Leuenberger, C. Taxon Sampling and the Optimal Rates of Evolution for Phylogenetic Inference. *Systematic Biology* **60**, 358-365 (2011). https://doi.org:10.1093/sysbio/syq097
62 Townsend, J. P. & Lopez-Giraldez, F. Optimal Selection of Gene and Ingroup Taxon Sampling for Resolving Phylogenetic Relationships. *Systematic Biology* **59**, 446-457 (2010). https://doi.org:10.1093/sysbio/syq025
63 Miller, D., Stern, A. & Burstein, D. Deciphering microbial gene function using natural language processing. *Nature Communications* **13**, 5731 (2022). https://doi.org:10.1038/s41467-022-33397-4

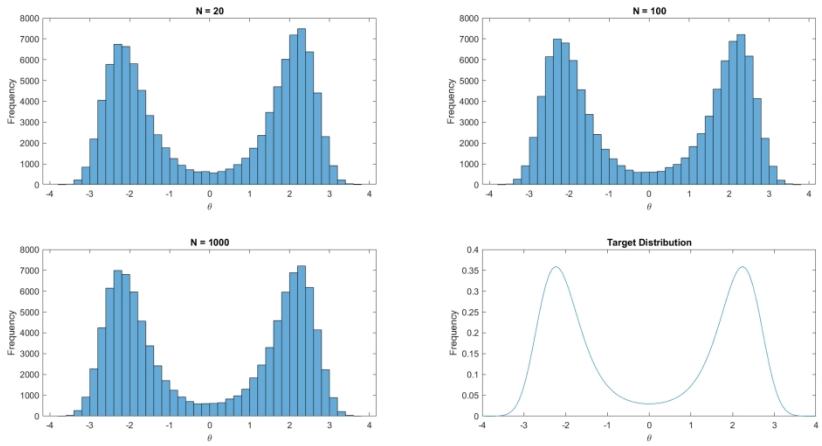

Figure 1. Convergence of an ensemble to the target distribution: (a) ensemble after 20 moves; (b) ensemble after 100 moves; (c) ensemble after 1000 moves; (d) true ensemble or "target distribution". In panel (d) is shown the target distribution as the ensemble converges to the true target distribution under a Monte Carlo Experiment (a-c). The starting guess at to the parameter θ was 3. After each of 20, 100, and 1000 moves, 10,000 samples of θ from the resulting distribution were drawn to characterize the ensemble. The ancillary parameters were α=0.1 and β=1.0. The plots were created in MATLAB_R2018B (https://www.mathworks.com/products/matlab.html).

406x198mm (300 x 300 DPI)

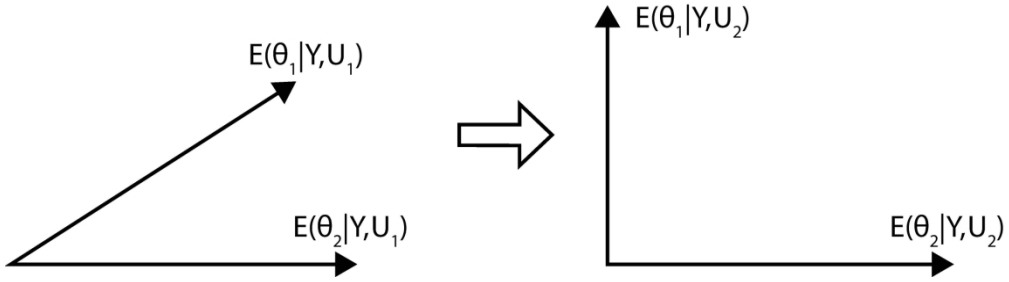

Figure 2. Two models can be better distinguished by their predictions in the next experiment if their predictions are less correlated. The predictions of model θ_1 and θ_2 under experimental condition U1 are the expectations f_1=E(θ_1 |Y,U_1) and f_2=E(θ_2 |Y,U_1), respectively. If the two models θ_1 and θ_2 are chosen independently from the model ensemble Q(θ|Y,U), the expectations are calculated with respect to the product density Q(θ_1 |Y,U)Q(θ_2 |Y,U), where U = U1 or U2.

208x57mm (300 x 300 DPI)

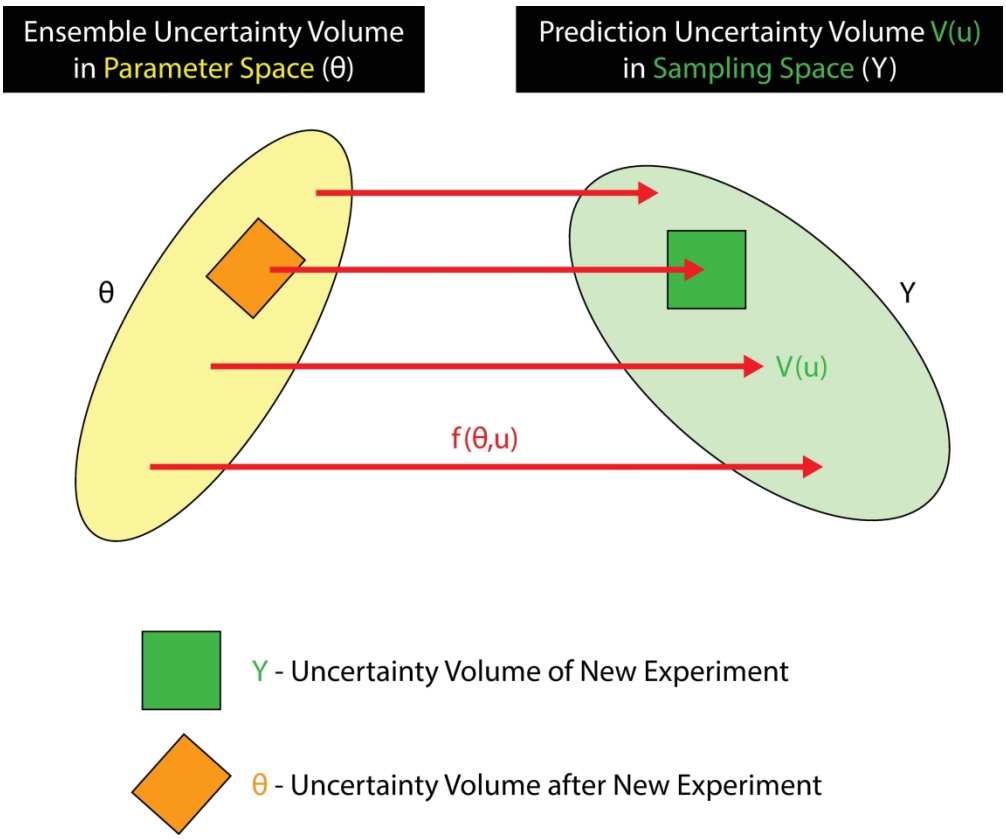

Figure 3. MINE is analogous in function to the optics on a microscope. The data Y are the objects in the field of view. The models θ are in the image. The MINE criterion with the predictions F(θ,U) is the optics. The Uncertainty Volume in the image is the magnification measured by the MINE criterion, V(u) = det(E(U)). From44.

244x202mm (300 x 300 DPI)

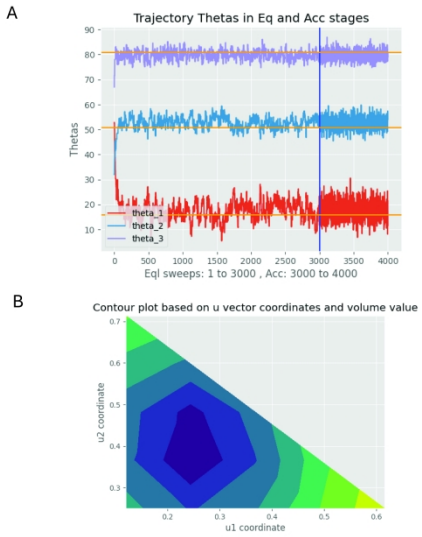

Figure 4. An ensemble method to identify the mixture experiment's hyphal extension rates θ is carried out on simulated data from the mixture experiment, and MINE is used to choose the next mixture experiment with inoculation proportions u1 and u2. (A) Panel A illustrates the ensemble method on simulated data for a mixture experiment of AMF colonizers. The orange lines are the true colonization rates θ. In the Monte Carlo experiment the estimated rates are plotted as a function of sweep, a visit on average once to each of the three rates θ. In the first 3000 sweeps the Monte Carlo experiment is equilibrated to get in the neighborhood of parameters θ that fit the simulated data. In the accumulation phase (last 1000 sweeps) the estimates of θ are accumulated to form the ensemble estimate. (B) Panel B presents the next MINE mixture experiment recommended. The contour plot is of the MINE criterion det(E) as a function of the mixture of spore inoculum proportions u1 and u2.

338x190mm (300 x 300 DPI)

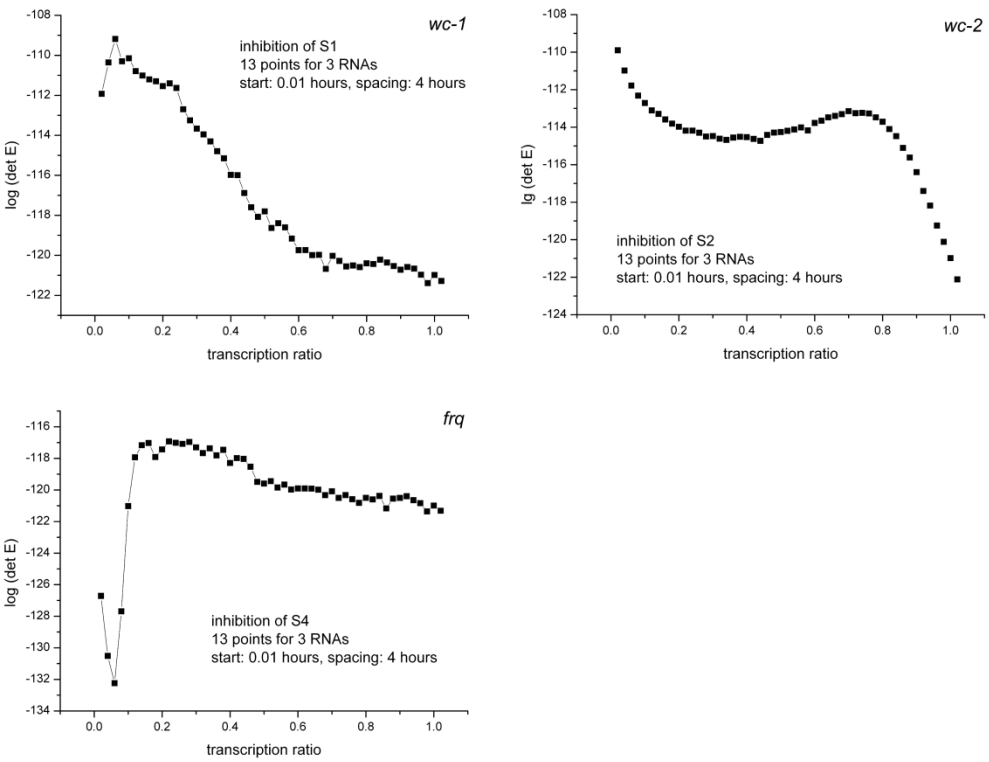

Figure 5. The MINE experiment is a 90% knockdown of the wc-1 gene. The MINE criterion displayed is the correlation ellipsoid volume det(E(U)), which is graphed as a function of the remaining activity of the three clock mechanism genes. The predictions F are of the log base 10 concentrations over time of frq, wc-1, and wc-2 mRNAs over time from the RNA profiling experiments. The mRNA levels were measured at 14 time points over an 8 hour window. The drawing is taken from6.

677x520mm (600 x 600 DPI)

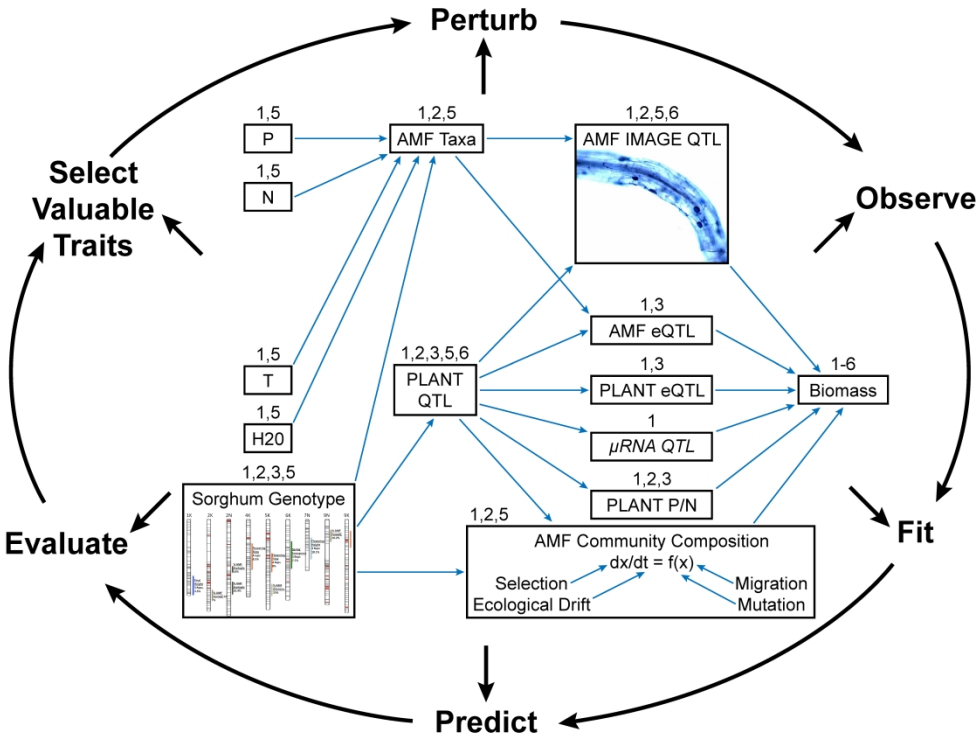

Figure 6. A sequence of MINE experiments are being used in a 5 year GWAS experiment to examine the relation between biomass and SNPs in Sorghum bicolor using the BAP accessions26. MINE is used to select the BAP accessions to be used in each year in order to map AMF colonization and biomass to the sorghum genetic map in a GWAS study. Multi-scale structural equation model (SEM) for the project (center boxes and arrows). Lotka-Volterra community models are nested within the SEM and predict associations that affect biomass. The dependent variable is biomass, and the arrows in the diagram denote causal relationships between independent variables in the SEM. In this model, sorghum genotype is the primary independent variable that correlates with the remaining variables. This conceptual model will evolve continuously using the model guided discovery process of maximally informative next experiment (MINE; outer ring)6.

279x215mm (600 x 600 DPI)

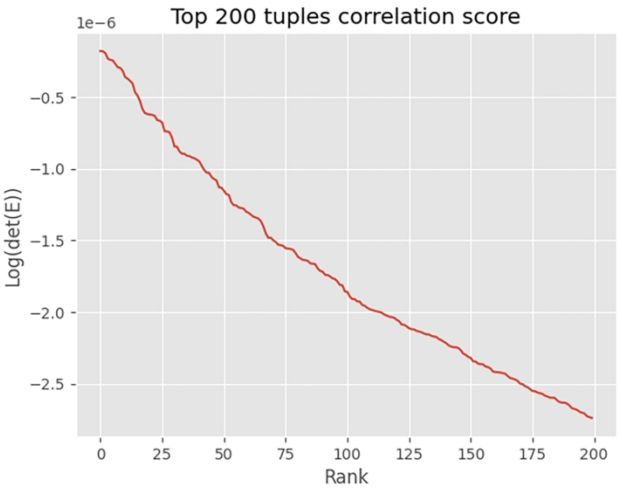

Figure 7. The MINE criterion log(det(E)) was used to select 80 accessions for use in a GWAS experiment at Wellbrook Farm, GA in 2022. The top 200 selected triples of accessions are ranked by det(D). From these top 200 triples 80 distinct accessions were selected.

215x279mm (300 x 300 DPI)