Flavor-Cyber-Agriculture: Optimization of plant metabolites in an open-source control environment through surrogate modeling

Arielle J. Johnson[1¶], Elliot Meyerson[2,3¶], Timothy L. Savas[1], Risto Miikkulainen[2,3&], Caleb B. Harper[1&*]

[1]Media Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

[2]Department of Computer Science, University of Texas at Austin, Austin, Texas, United States of America

[3]Sentient Technologies, San Francisco, California, United States of America

* Corresponding Author

E-mail: calebh@media.mit.edu (CBH)

[¶] These authors contributed equally to the work

[&] CBH and RM are Joint Senior Authors

Submitted to PLOS ONE, August 2018

Abstract

Food production in conventional agriculture faces numerous challenges such as reducing waste, meeting demand, maintaining flavor, and providing nutrition. Contained environments under artificial climate control, or cyber-agriculture, could in principle be used to meet many of these challenges. Through such environments, phenotypic expression of the plant---mass, edible yield, flavor, and nutrients---can be actuated through a "climate recipe," where light, water, nutrients, temperature, and other climate and ecological variables are optimized to achieve a desired result. This paper describes a method for doing this optimization for the desired result of flavor by combining cyber-agriculture, metabolomic phenotype measurements, and machine learning. In a pilot experiment, (1) environmental conditions, i.e. photoperiod and ultraviolet (UV) light (known to affect production of flavor-active molecules in edible plants) were applied under different regimes to basil plants (*Ocimum basilicum*) growing inside a hydroponic farm with an open-source design; (2) flavor-active volatile molecules were measured in each plant using gas chromatography-mass spectrometry (GC-MS); and (3) symbolic regression was used to construct a surrogate model of this chemistry from the input environmental variables, and search in this model was used to discover new combinations of photoperiod and UV light to increase this chemistry. These new combinations, or climate recipes, were then implemented in the hydroponic farm, and several of them resulted in a marked increase in volatiles over control. The process demonstrated a "dilution effect", i.e. a negative correlation between weight and desirable chemical species. It also discovered the surprising effect that a 24-hour photoperiod of photosynthetic-active radiation, the equivalent of all-day light, induces the most flavor molecule production in basil. In this manner, surrogate optimization through machine learning can be used to discover effective recipes for cyber-agriculture that would be difficult and time-consuming to find using hand-designed experiments.

Introduction

The so-called "dilution effect," noted since the 1940's and systematically reviewed since the early 1980's [1], describes an inverse relationship between yield and nutrient concentration in food: For many nutritionally-important chemical constituents of food plants, such as minerals, protein, and vitamins, an increase in biomass is accompanied by a decrease in nutrient concentration. This effect has been systematically demonstrated in historical nutrient content studies over the last 50-70 years [2,3], as well as in controlled side-by-side trials that have shown a relationship between nutrient dilution and genetics [4], artificial fertilization [5], and elevated carbon dioxide levels related to climate change [6,7]. Flavor, known to be an important element of food and of eating behavior for organisms from insects to humans [8], has been declining alongside nutrients over approximately the last 50 years [9-11] in inverse proportion to rising yields. Flavor-active molecules in plants frequently have either positive health benefits (antioxidant, antimicrobial, anti-inflammatory) themselves or signal the presence of other beneficial or essential molecules, for example by being the enzymatic products of precursors (e.g. pro-vitamin A carotenoids, essential amino or fatty acids) necessary for human nutrition and health [9].

Vertical farming, or more generally cyber-agriculture, is a plant-growing format employing contained environments where light, water, nutrients, temperature, and other climate variables are provided artificially under computer control [12-14]. Data from environmental sensors informs the actuation of climatic conditions according to a "recipe" that is designed for best possible outcome, such as largest yield, best flavor, desired nutrients, and least cost. With cyber-agriculture, in principle it may be possible to increase food production quality and quantity, minimize waste and cost, and grow food with optimized climate recipes anywhere including locations otherwise unable to support agriculture. Conventional agriculture has been optimized for yield. What if it were optimized for quality and flavor?

This paper describes a proof-of-concept method aimed at optimizing flavor in a cyber-agricultural controlled environment, and a pilot experiment to validate this method. An experimental container, called the Food Computer [12], was built at the MIT Media Lab with sensors, actuators, and computer control. Basil (*Ocimum basilicum*) was chosen as a model organism because it has a fast growth cycle (five weeks), and because the outcome can be readily measured in terms of fresh weight (quantity), and chemical analysis of flavor (quality). A number of known growth recipes were implemented, together with a broad range of their variations [15]. Machine learning technology [16-18] was then used to optimize these recipes further. That is, based on these recipes and their associated outcomes, a surrogate model was first constructed using symbolic regression. To keep the search problem manageable, the optimization focused on the lighting conditions, keeping the other variables constant. The surrogate model was then searched to discover potentially better lighting recipes, which were then tested in the experimental container.

The light conditions had a large effect on the outcome, and the surrogate optimization method was able to discover meaningful recipes. For instance, it discovered the well-known principle that flavor can be traded off with mass, a version of the “dilution effect”: optimizing for flavor produced smaller plants, while optimizing for mass produced less flavor. However, it also demonstrated how the approach can discover new and surprising recipes, i.e. those that are counterintuitive but produce better outcomes. In particular, the common-sense assumption that basil needs a few hours of darkness each day turned out to be incorrect: The highest density of flavor molecules was produced through a 24-hr photoperiod, which optimization discovered quickly and reliably. The results thus demonstrate that surrogate optimization and machine discovery can be used to find growth recipes that are both effective and surprising.

Measuring and optimizing flavor

Flavor is largely a phenomenon of olfaction [19], and many aroma molecules are produced by the secondary metabolism of plants. Plants have a particularly rich secondary or specialized metabolism [20], a set of biosynthetic pathways synthesizing molecules that are not essential for the basic processes of life (cell division, reproduction, etc.) but rather confer fitness and adaptive advantage to the organism in its ecological niche [21], related to stress tolerance, defense, and communication [22]. Their expression and induction depends to various degrees on environmental and ecological conditions [23].

Cyber-physical agriculture methods such as the Food Computer (FC), where data from environmental sensors informs the actuation of climatic conditions according to a climate recipe [12-14] present unique opportunities for inducing plant phenotypic changes through environmental/ecological conditions alone. One example of this approach is to apply the ecological stresses to which adaptations have evolved as specific biosynthetic pathways.

O. basilicum, the basil plant, is typical of herb plants in that it produces many aromatic molecules, particularly the terpenoids 1,8-cineole, linalool, camphor, borneol, bergamotene, and farnesene, and the phenylpropenes eugenol, methyleugenol, and estragole [24]. These molecules are thought to play varying roles in stress adaptation and defense, and the production by the basil plant of aromatic molecules has been shown to increase upon exposure to these stresses, including water stress [25], ultraviolet and PAR light [26–28], heat [29], bacteria [30], chitosan (a compound derived from chitin, found in insect exoskeletons and fungal cell walls, [31]), and sodium and other minerals [32].

This paper explores methods for increasing flavor molecule production in *O. basilicum*, using: (1) ultraviolet light, PAR, and photoperiod as environmental and stress variables; (2) Gas Chromatography-Mass Spectrometry for semiquantitative analysis of volatiles; (3) surrogate optimization for discovering conditions that will maximize production of these volatiles.

Materials and Methods

This section describes the design of the Food Computer, i.e. the physical container environment used in the pilot experiment with basil. It also describes the process for growing basil in this environment, and methods for measuring the growth outcome in terms of weight and chemistry.

Food Computer

All basil plants were grown in a Food Server, a multi-tray, multi-rack hydroponic configuration of the OpenAg Food Computer™ (FC) environment [12]. Basil plants were germinated in engineered foam rooting cubes (Oasis Grower Solutions, Kent, OH), then transplanted to 36-position (4×9) food-grade resin floating lettuce rafts (Beaver Plastics, Acheson, AB, Canada) at 14 days of age. The plants were grown in a shallow water culture hydroponic system using 56.6-liter trays (Botanicare, Chandler, AZ) supplied by 75-liter reservoirs (Botanicare) and 700 gallon-per-hour rated Pondmaster magnetic drive pumps (Danner Manufacturing, Islandia, NY), with nutrient solutions (a "15-0-0" Calcium Nitrate solution and a "5-12-26" 5% Nitrate, 12% Phosphate, 26% soluble Potash solution combined with water for a final concentration of 150 ppm Nitrogen, 116 ppm Calcium, 52 ppm Phosphorus, 215 ppm Potassium; JR Peters, Allentown, PA) added by a water-powered proportional chemical injector (Dosatron, Clearwater, FL).

The FC was set up with trays in vertical stacks of three (denoted 0, 1, and 2) within a custom powder-coated steel frame (Indoor Harvest, Houston, TX). Each stack was thermally isolated with reflective foil captive-bubble insulation (Reflectix, Markleville, IN) and climate-controlled with a 10,000 BTU air conditioning unit (AeonAir, Wilmington, DE). Three types of photosynthetic-active radiation (PAR) lights were used: Agrobrite high output T5 fluorescent fixtures (Hydrofarm, Fairless Hills, PA), Illumitex ES2 Eclipse red and blue LED fixtures (Illumitex, Austin, TX) and Phillips GreenPower deep red/blue LED production modules (Phillips, Somerset, NJ). Lights were fixed at a distance of 40 cm from the foam float. Reptisun 10.0 UVB T5 High Output ultraviolet lights were added to treatment conditions (Zoo Med, San Luis Obispo, CA), also at a distance of 40 cm.

Figure 1. Images inside the MIT Media Lab Food Server taken during the experiment.

Plant species and climate recipes

Common Sweet Basil (*O. basilicum var* "Sweet") seeds (Eden Brothers, Arden, NC) were used in the pilot experiment. From 14 days of age to harvest, they were grown in identical trays as described in "Food Computer" above, with one of three light types as the only source of PAR. Control conditions were grown with the PAR light; experimental treatment conditions had supplemental UV light. Treatment conditions, or "Climate Recipes", in Rounds 2 and 3 of the experiment were determined based on suggestions from the surrogate optimization of chemscore from the previous round. The data from Round 1 determined the conditions of Round 2, and the data from Round 2 determined the conditions of Round 3).

Harvest, weight and length measurement

All plants in each round of the experiment were harvested on the same day. Four plants from each treatment condition were used for volatile analysis and the remaining 32 were used for height and weight measurements. Weight measurements were taken with roots removed.

Sampling and sample preparation

Immediately after harvesting, leaves were sampled from four plants from each treatment condition. Fifteen leaves from each plant were harvested: five from near the base, five from the middle, and five from the top, with each set selected randomly. Leaves were immediately frozen with dry ice or liquid nitrogen, homogenized into a powder, and kept frozen. The amount of 1gram of frozen plant tissue was transferred to a 20 mL amber glass headspace vial (Supelco, Bellefonte, PA) and 2 mL of saturated, cold calcium chloride solution in distilled water was added to prevent enzymatic reactions. The vials were capped with magnetic, PTFE-lined silicone septa headspace caps (Supelco) and kept on ice before being transferred to GC-MS.

Volatile Analysis

The method of Johnson et al. [33] was adapted for the experiment. Sample vials were placed in the tray of the Gerstel MPS2 autosampler (Gertsel, Linthicum, MD), which performed the extraction and injection. One vial at a time was warmed to 40°C and agitated at 500 rpm for 5 minutes directly before extraction. A conditioned, 2-cm long 50/30 um-thick PDMS-DVB SPME fiber (Supelco) was introduced into the headspace of the vial for 45 minutes at 40°C with rotational shaking at 250 RPM. The fiber was removed from the headspace of the vial and immediately introduced into the inlet of an Agilent model 7890 Gas Chromatograph- single quadrupole-MS (GC-MS) (Agilent Technologies) with a DB-5 column (30 meters long, 0.25 mm ID, 0.25 μm film thickness, J&W Scientific, Folsom, CA). The inlet was held at 250°C with a 2:1 split and had a 0.75mm i.d. SPME inlet liner installed (Agilent Technologies). The carrier gas was Helium, at a constant flow rate of 1 mL/minute. The starting oven temperature was 40°C, held for 3 minutes, followed by a 2°C/minute ramp until 180°C was reached, then the ramp was increased to 30°C/minute until 250°C was reached, and held for 3 minutes. The total runtime was 47 minutes. The transfer line to the mass spectrometer was held at 250°C , the source temperature was 230°C, and the quadrupole temperature was 150°C. The mass spectrometer had a 1.5-minute solvent delay and was run in scan mode with Electron Impact ionization at 70eV, from *m/z* 40 to *m/z* 300.

Compounds were identified and recorded based on a 90% or higher match using the NIST Mass Spectral Database and a signal to noise ratio above 10. Analyte peaks were integrated on the Total Ion Chromatogram.

Optimization metric: Chemscore

Optimizing the target metric should correspond to maximizing flavor in a general sense. The metric should also be robust to noise, since the number of evaluations is limited, and low-dimensional to make optimization easier.

Basil, like most foods, contains multiple molecules contributing to flavor. An average GC-MS chromatogram of basil contains around 30-40 different volatile molecules, with concentrations varying over several orders of magnitude. To construct a single metric to optimize, this GC-MS data is aggregated across samples and chemicals as the Chemscore. This score is a weighted average of the volatile profile compared to the control condition. It is a holistic placeholder for how flavorful a sample is, while normalizing for varying scales and distributions of different chemicals. Seventeen chemicals common across all GC-MS measurements were selected into the calculation of chemscore.

Comparison metrics: R-score and Z-score

For further comparison, an R-score and a Z-score, across all volatiles in a sample, were calculated for each treatment condition. The R-Score, the average ratio of volatiles in a treatment condition over their average in the control conditions in a round of the experiment, facilitates comparison of results across the three rounds of the experiment, under the assumption that uncontrollable environmental differences across rounds are captured in differing control results. The Z-score, which compares the abundances of each volatile molecule in a sample over or below its average in all samples in a round, gives a sense of the overall spread of results in the experiment.

Methods: Surrogate modeling and optimization

In optimization settings where the target function is expensive to evaluate (either temporally or financially), e.g., in the case of growing basil to maturity, surrogate-based optimization is a common method for minimizing the number of evaluations required to achieve an acceptable solution [34–36]. To choose the next samples to evaluate, surrogate methods build an explicit predictive model of the solution landscape and select the most promising samples according to this “surrogate model". To implement such a method, input variables need to be defined, a class of regression models needs to be selected, and a method for discovering the next samples (recipes) from these models needs to be developed. This section details the development of these choices for the experiment in this paper, and notes methods for scaling up future work.

Input Dimensions

For this experiment, a recipe was defined by three input variables: photoperiod, UV period, and PAR (photosynthetically active radiation). Three input variables was an appropriate dimensionality for this pilot experiment, following the general rule-of-thumb that, for surrogate-based methods, the number of evaluations required to achieve reasonable results is around ten times the number of dimensions [34].

These variables were chosen because they are already known to increase the accumulation of volatiles [26–28], and are relatively simple to control in the described hardware setup.

Photoperiod is the number of hours the primary light panel is turned on each day. Recipes can thus have photoperiod values anywhere from 0 to 24hrs. Photoperiod is known to have significant effects on the accumulation of biomass and leaf area in plants [37], as well as the formation of trichomes, the structure that store flavor-active volatiles, in the *Thymus vulgaris* (thyme) plant [38], a botanical cousin, as they are both members of the *Lamiaceae* family, to basil. In addition, photoperiod has been shown to change the volatile profile of basil [39].

UV period is the number of hours per day plants receive supplemental UV-B radiation. Like photoperiod, UV period can take on values anywhere from 0 to 24hrs. UV has previously been shown to increase volatile content in basil [26]; it is included so that its effects can be validated and optimized in the Food Computer hardware setting.

PAR (Photosynthetically Active Radiation) is the amount of light available for photosynthesis. In the Food Computer setup, the PAR is determined by the primary light panel. There were nine light panels, each with a unique PAR value. To set PAR values for a batch of nine recipes, one light panel was assigned to each recipe. Thus, in contrast to photoperiod and UV period, each available PAR value can be used only once in each batch. This kind of hardware resource matching constraint is not common in either computer or physical experiments, so a custom optimization method must be developed.

Regression Model

Symbolic regression [40–42] was used for building surrogate models to predict chemscore from the input variables. Symbolic regression uses evolutionary optimization to discover nonlinear algebraic expressions that serve as surrogate models. For the experiment in this paper, a multi-objective Pareto optimization procedure was used [43,44]. The first objective is to minimize error, i.e., MSE with respect to predicting chemscore; the second objective is to maximize parsimony, i.e., minimize the size of the algebraic expression (number of nodes). The fitting procedure then yields a Pareto front of models, from which a new batch of recipes can be selected.

For the flavor-optimization problem, symbolic regression has several advantages over other popular choices for surrogate models. First, by simultaneously optimizing for error and parsimony, search is biased towards the kinds of compact algebraic expressions that are desirable in the natural sciences [44]. These expressions are more interpretable than other regression models, because the relationships between variables can be read off directly from the expression. Such interpretability can lead to a better understanding of the search space, which helps in developing better models for future experiments.

Second, whereas surrogate models such as Gaussian processes can only interpolate, symbolic regression can extrapolate. Interpolation is sufficient when iterative incremental improvement can eventually lead to an optimal solution. However, in the experiment in this paper, only a single parallel batch of recipes is selected via surrogate optimization to be implemented in the Food Computer. So, it is advantageous to consider strong optimistic predictions a model makes about sparse regions in the recipe space. Note that if this process were used over multiple iterations, an inordinate amount of resources could be spent at the extremes of the recipe space.

Third, symbolic regression is robust to normalization of input and output variables: It automatically discovers reasonable scaling factors to use through optimized constants that are found to be useful in model expressions.

It is important to note that symbolic regression can have significant drawbacks as well [43]. First, it is computationally expensive compared to other regression methods; however in this paper, computation time is negligible compared to the time it takes to grow a batch of basil recipes. Second, surrogate optimization with symbolic regression models currently lacks theoretical convergence guarantees and performance bounds. Such convergence guarantees have potential practical benefits over many iterations of surrogate optimization; however, since only a single such iteration is performed in the experiment in this paper, such guarantees are unnecessary.

Recipe Discovery

There were three rounds of growing experiments. In each round, there are nine trays of basil growing in parallel. To ensure consistency across rounds, three of these nine trays are fixed to control recipes. This setup leaves six non-control recipes to be selected.

In the first round, recipes were selected by hand [15] to investigate the effects of UV supplement and choice of light panel. To add the photoperiod dimension, and create initial diversity in the recipe space, recipes in the second round were chosen by an unsupervised method: Six non-control recipes were found as centroids of Voronoi tessellation (CVT) given the first round of recipes [45]. Following a trust region approach [35], to implement the bias that good solutions are likely to be relatively close to expert hand-designed recipes, values for each dimension were constrained to be with a constant distance of previously evaluated values.

In the third round, recipes were selected from symbolic regression surrogate models [46]. Each run of symbolic regression yields a collection of models on the error-parsimony Pareto front. These models were clustered to determine an error threshold, above which models were underfitting. The six most parsimonious models not underfitting were then used to define a recipe to run in parallel. Since the recipe space has only three dimensions it is computationally efficient to use a dense grid search to select a recipe that maximizes expected chemscore. Greedy sequential selection is the most popular approach to constructing parallel batches from surrogates [47,48]. The recipes were thus selected sequentially in increasing order of model error. Such a selection handles the constraint that each available PAR value can be selected only once per round. If a variable is ignored by a model, the value of the variable is set to maximize exploration, since the model has indicated that exploitation of this variable is currently not useful.

Results and Discussion

The average weight, chemscore, total peak area of volatiles on the chromatogram, and chemscore as a percentage of the control condition chemscore are presented in Table 1. Weight was recorded as the weight of above-ground plant parts; roots were excluded.

Table 1. Treatment conditions (UV and PAR photoperiod) and weight and chemical results.

Round	Bay	Tray	UV Photoperiod	PAR Photoperiod	PAR	Weight (grams)	R score	Chemscore	Z score	R score
1	1	0	18	18	636.92	32.00	0.85	-0.77	0.65	
1	1	1	18	18	798.42	102.71	1.00	0.21	1.15	
1	1	2	18	18	832.58	133.59	1.06	0.44	1.37	
1	2	0	0	18	820.25	72.08	1.13	0.46	1.45	
1	2	1	0	18	1,098.75	235.44	0.81	-0.68	0.79	
1	2	2	0	18	403.58	84.33	1.06	0.33	1.34	
2	0	0	9	21.5	867.33	74.18	**1.81**	1.07	0.68	
2	0	1	9	21.5	445.25	65.63	1.15	-0.01	0.10	
2	0	2	9	21.5	735.42	63.86	**1.61**	0.86	0.50	
2	1	0	9	14.5	636.92	112.89	0.89	-0.43	-0.25	
2	1	1	9	14.5	798.42	189.00	0.58	-1.07	-0.52	
2	2	0	0	18	820.25	154.50	0.92	-0.42	-0.19	
2	2	1	0	18	1,098.75	211.00	0.73	-0.58	-0.28	
2	2	2	0	18	403.58	112.00	1.35	0.57	0.27	
3	0	0	17.45	24	867.33	137.44	**16.57**	2.38	-0.28	**14.05**
3	0	1	4.12	24	445.25	71.25	**2.33**	-0.21	-1.03	**1.83**
3	0	2	24	24	735.42	49.33	**2.84**	-0.05	-1.01	**2.12**
3	1	0	14.06	24	636.92	80.51	**2.00**	-0.30	-1.05	1.47
3	1	1	8.48	17.18	798.42	62.78	1.80	-0.34	-1.06	1.34
3	1	2	10.67	22.5	832.58	88.83	**2.09**	-0.28	-1.04	**1.55**
3	2	0	0	18	820.25	92.89	0.80	-0.66	-1.11	0.60
3	2	1	0	18	1,098.75	126.86	1.20	-0.53	-1.09	0.94

3	2	2	0	18	403.58					1.47

"Bay" specifies the position in the vertical stack of three hydroponic trays, with "0" closest to the floor. One tray in each bay contained a control condition, which had zero hours UV photoperiod and 18 hours PAR photoperiod. R score > 1.5 is denoted in bold. The photoperiod hours range between 0 and 24. PAR values indicate μmole/m^2s photosynthetic photon flux density. The R-Scores were calculated with missing control imputed.

The table includes results both with and without imputed data for the control condition whose data was lost in Round 3 of the experiment (denoted by dark grey boxes in table 1). Assuming control results are consistent within each round, they make the results easier to compare across rounds. Imputed values for each chemical for the missing control treatment in Round 3 were computed by regression, i.e., by solving a fully-determined linear system that predicts the value of the third control from the other two, based the values of the controls in the previous two Rounds.

Table 2 gives the correlations between input variables and metrics (Spearman, to account for nonlinearity in the metrics). Correlations >0.45 are in bold. Note in particular that the R-scores are negatively correlated with weight: Optimizing for flavor thus results in smaller plants, and larger plans have less flavor, thus illustrating the "Dilution effect."

Table 2: Spearman correlations between selected input variables and metrics

	R Score	Weight	Chemscore	Z Score
UV	0.355	-0.336	0.199	0.058
Photoperiod	**0.763**[a]	-0.355	**0.477**[a]	-0.149

PAR	-0.131	**0.541**[a]	-0.142	-0.070
R Score		**-0.471**[a]	**0.637**[a]	-0.226
R Score Imputed	**0.967**[a]	**-0.502**[a]	**0.764**[a]	-0.055

In the first round, where an 18-hour PAR photoperiod and an 18-hour UV photoperiod were selected by hand, R-score and Chemscore did indicate that UV light or photoperiod increases volatiles. In the second round, two R-scores (both with UV light and extended PAR photoperiod of 21 hours) were above 1.5, meaning that volatiles holistically increased 50% over control. In the third round, several conditions resulted in an R-score that met or exceeded this threshold, with many conditions (all with PAR photoperiods of 22.5-24 hours and UV periods of 4-17 hours) doubling the volatile profile compared to control. The discovery of the recipes in Round 3 from the model is illustrated in Figure 2.

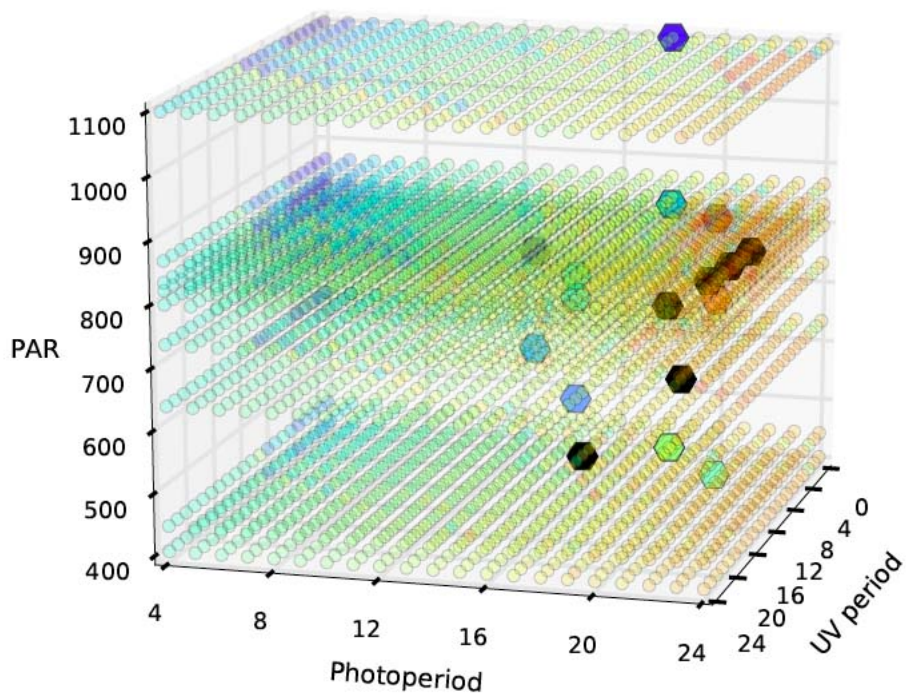


Figure 2. An illustration of the surrogate model and the recipes suggested by the optimization. The three axes correspond to the three actuators and the color of the small dots indicates their value predicted by the model (i.e. flavor; red > yellow > green > blue). The large dots are suggestions, and the darker dots are the most recent ones. They suggest utilizing long photoperiods and UV periods, the success of which was confirmed in growth experiments in the Food Computer.

The most striking discovery in this experiment was the positive effect of a 24-hour photoperiod, i.e., constant daylight. This result replicated evidence on the volatile profile effects of a 24-hour photoperiod described by Skrubis et al. [39], who found that basil plants grown with a 24-hour photoperiod weighed, upon maturity, approximately 25% more than plants grown with a nine-hour photoperiod (although they took three days longer to reach maturity) and 27% more than plants grown outdoors in natural light with an approximately 15-hour photoperiod . That study also characterized changes in the relative volatile profiles of those basil plants, but not absolute volatile content, so comparisons to chemscore in the current work are not possible. The 24-hour photoperiod discovery is notable because the hand-designed experimental conditions in Round 1 had a photoperiod of 18 hours, and the experimenters and the model were blind to the Skrubis et al. study. The surrogate optimization approach nevertheless iterated the recipes into the 24-hour photoperiod, where it had a strong positive effect.

Aside from the high R-score in Table 1 , further evidence for the importance of photoperiod can be seen in the high correlation between R-score and Photoperiod in Table 2, and in the regression process itself: For each run of symbolic regression, the most parsimonious nontrivial model had the form $y = cp$, for some constant c, where p is the photoperiod. Also, Figure 3A shows the a linear model trained on all three light variables to fit the log R-score. Fig 3B shows a linear model of R-score based on photoperiod alone. Fig 3C shows the predictions of a linear model trained on all three variables, but with the effect of photoperiod removed, i.e., it is trained to fit the residuals. These modeling results are similar with imputed and outlier-handled data. The low performance of the residual model suggests that photoperiod had such a dominating effect that the effects of other variables were effectively noise. However, since significant effects of UV have been reported in previous work [26,27] and are not seen here, it is also possible that there are significant nonlinear dynamics that require further trials and nonlinear modeling to uncover and exploit.

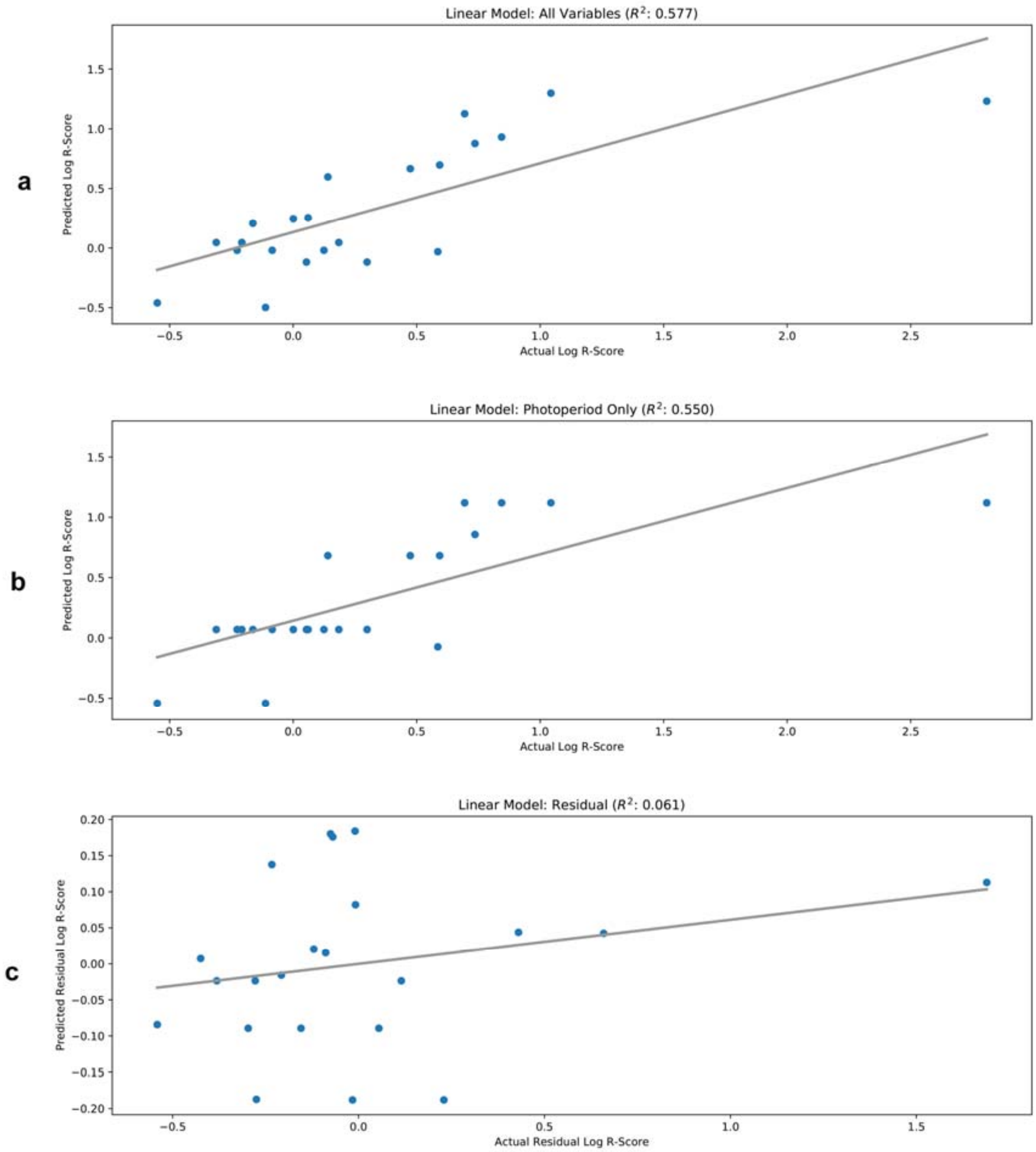


Figure 3. Linear regression analysis of actual vs. calculated log R-score for three different models. A: A linear model trained on UV, photoperiod, and PAR. B: A linear model trained on photoperiod only. C: A linear model trained on residuals after removing photoperiod effect. Photoperiod dominates the other variables (or possible there are significant nonlinear effects between these variables).

Discussion and Future Work

The experiment described in this paper confirmed that the design of climate recipes impact the accumulation of volatile flavor molecules in basil, and it is possible to discover good recipes iteratively through machine learning. The recipes discovered in this case replicated known principles (such as the weight/flavor tradeoff), and also demonstrated the possibility for discovering previously unknown, surprising principles (like the 24hr photoperiod). The 24-hour photoperiod in particular is impossible in nature (except around the summer solstice within the Arctic and Antarctic circles) and therefore unlikely to be discovered, except in controlled environments for cyber-physical agriculture.

The most immediate direction of future work is to expand the current experiment to a larger search space. A facility with four containers, making it possible to evaluate an order of magnitude more recipes at once is in development at MIT and illustrated in Fig 4. This facility will make it possible to control a number of other actuators besides light, including temperature, pH, nutrient concentration, microbial and other additives, and different cultivars. It will also be possible to measure the energy and other costs associated with the recipes, as well as objectives such as nutrient components, density, and yield, and more elements of flavor (single compounds, and ratios of compounds).

Figure 4. Images of MIT expansion facility under development.

In terms of surrogate optimization, more iterations will be run to build more accurate models, and to determine the proper stopping point of the method, i.e. run it until it has likely converged. The approach will be extended to cover the larger search space as well as multiple objectives. Most likely, different models and optimizers will be necessary. In low-dimensional settings with unknown nonlinearities and a relatively small number of samples, Kriging [34], Gaussian processes [36,49], and symbolic regression [44] are suitable choices for building a regression model of natural phenomena. When the dimensionality and number of samples increases, deep neural networks may be a better model of the solution landscape [47,50,51], and evolutionary optimization a better way to determine most promising samples [43–45].

The next step will be to extend the experiment to other plants, such as cotton, where the goal is not to optimize flavor but physical properties such as strength and length of the fibers. It will be important to verify that such plants are viable to grow artificially, and that such properties can be optimized with available actuators, in isolation and in combination with other properties. Future extensions to other areas may include biofuels and plants with specific medicinal value.

The third future step is to extend the optimization from static recipes to time-varying recipes, i.e. optimizing the actuators during the entire growth period of the plant. Of particular interest are different stress periods when the plant is exposed to, for example, drought or signals of predators (e.g. through chitosan added to the growth medium). Such periods may produce a response in the plan that results in more flavor or more rapid growth, for example. Such recipes should be reactive, i.e. conditional to real-time measurements of the growth status. One possibility is to use machine learning to establish a mapping from visual images of the plant to more destructive measurements such as chemical concentrations. Such optimization spaces are very high-dimensional, most likely making it necessary to use evolutionary optimization, and perhaps neuroevolution to construct a mapping from sensory time series to optimal actions [55,56].

Conclusion

Computer-controlled growth environments are a promising approach for the future of agriculture, potentially maximizing production and quality and minimizing waste and cost. Initial experiments with basil (*O. basilicum*) suggest that the cyber-physical approach to agriculture is indeed viable: such environments can be built, the plants thrive in them, the climate recipes make a difference in growth outcomes, and machine learning can be used to discover good recipes automatically. Future steps should verify these results on other plants, expand to larger search spaces with more actuators, and to optimizing entire growth periods. Higher-volume food computers need to be built and more powerful optimization methods employed, but the results suggest that such extensions are worthwhile.

Acknowledgements

The authors would like to thank Christina Agapakis, Nate Tedford, and Scott Marr for access to and support on analysis instrumentation and Babak Hodjat and Hormoz Shahrzad for modeling and optimization insights and comments on the manuscript.

References

1. Jarrell WM, Beverly RB. The dilution effect in plant nutrition studies. Adv Agron. Elsevier; 1981;34: 197–224.
2. Davis DR, Epp MD, Riordan HD, Davis DR. Changes in USDA Food Composition Data for 43 Garden Crops, 1950 to 1999. J Am Coll Nutr. 2004;23: 669–682. doi:10.1080/07315724.2004.10719409
3. Davis DR. Declining fruit and vegetable nutrient composition: What is the evidence? HortScience. 2009;44: 15–19.
4. Farnham MW, Grusak MA, Wang M. Calcium and magnesium concentration of inbred and hybrid broccoli heads. J Am Soc Hortic Sci. 2000;125: 344–349. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-0033998604&partnerID=tZOtx3y1

5. Hughes M, Chaplin MH, Martin LW. Influence of mycorrhiza on the nutrition of red raspberries. HortScience. 1979;14: 521-523.

6. Loladze I. Rising atmospheric CO2 and human nutrition: Toward globally imbalanced plant stoichiometry? Trends Ecol Evol. 2002;17: 457–461. doi:10.1016/S0169-5347(02)02587-9

7. Cotrufo FM, Ineson P, Scott A. Elevated CO 2 reduces the nitrogen concentration of plant tissues. Glob Chang Biol. 1998;4: 43–54. doi:10.1046/j.1365-2486.1998.00101.x

8. Fraenkel GS. The Raison d'Etre of Secondary Plant Substances. Science (80-). 1959;129: 1466–1470. doi:10.1126/science.129.3361.1466

9. Goff SA, Klee HJ. Plant Volatile Compounds: Sensory Cues for Health and Nutritional Value? Science (80-). 2006;311: 815–819. doi:10.1126/science.1112614

10. Folta KM, Klee HJ. Sensory sacrifices when we mass-produce mass produce. Hortic Res. Nature Publishing Group; 2016;3. doi:10.1038/hortres.2016.32

11. Schatzker M. The Dorito Effect: The Surprising New Truth about Food and Flavor. Simon and Schuster; 2015.

12. Harper C, Siller M. OpenAG: A Globally Distributed Network of Food Computing. IEEE Pervasive Comput. 2015;14: 24–27. doi:10.1109/MPRV.2015.72

13. Harper C. Open-Source Agriculture Initiative—Food for the Future? In: Kozai T, editor. LED Lighting For Urban Agriculture. Singapore: Springer Science+Business Media; 2016. pp. 37–46. doi:10.1007/978-981-10-1848-0

14. Ferrer EC, Rye J, Brander G, Savas T, Chambers D, England H, et al. Personal Food Computer: A new device for controlled-environment agriculture. arXiv Prepr arXiv170605104. 2017;

15. Chernoff H. Sequential design of experiments. Ann Math Stat. JSTOR; 1959;30: 755–770.

16. O'Reilly U-M, Wagy M, Hodjat B. EC-star: A massive-scale, hub and spoke, distributed genetic programming system. Genetic Programming Theory and Practice X. Springer; 2013. pp. 73–85.

17. Meyerson E, Miikkulainen R. Discovering Evolutionary Stepping Stones through Behavior Domination. arXiv Prepr arXiv170405554. 2017; Available: https://arxiv.org/pdf/1704.05554.pdf

18. Miikkulainen R, Liang J, Meyerson E, Rawal A, Fink D, Francon O, et al. Evolving Deep Neural Networks. 2017; Available: http://arxiv.org/abs/1703.00548

19. Small DM. Flavor is in the brain. Physiol Behav. Elsevier Inc.; 2012;107: 540–52. doi:10.1016/j.physbeh.2012.04.011

20. Weng J. The evolutionary paths towards complexity□: a metabolic perspective. New Phytologist. 2014; 201(4):1141–1149.

21. Moghe G, Last RL. Something old, something new: Conserved enzymes and the evolution of novelty in plant specialized metabolism. Plant Physiol. 2015;169: pp.00994.2015. doi:10.1104/pp.15.00994

22. Weng J-K, Philippe RN, Noel JP. The Rise of Chemodiversity in Plants. Science (80-). 2012;336: 1667–1670. doi:10.1126/science.1217411

23. Deschamps C, Simon JE, Wt. Terpenoid essential oil metabolism in basil (Ocimum basilicum L.) following elicitation. J Essent Oil Res. 2006;18: 618–621. doi:10.1080/10412905.2006.9699183

24. Lee S-J, Umano K, Shibamoto T, Lee K-G. Identification of volatile components in basil (Ocimum basilicum L.) and thyme leaves (Thymus vulgaris L.) and their antioxidant properties. Food Chem. 2005;91: 131–137. doi:10.1016/j.foodchem.2004.05.056

25. Khalid KA. Influence of water stress on growth, essential oil, and chemical composition of herbs (Ocimum sp.). Int Agrophysics. 2006;20: 289–296. doi:10.1016/j.plantsci.2004.05.034

26. Nitz G, Schnitzler W. Effect of PAR and UV-B radiation on the quality and quantity of the essential oil content in sweet basil (Ocimum basilicum L.). Acta Hortic. 2004;659: 375–381.

27. Ioannidis D, Bonner L, Johnson CB. UV-B is required for normal development of oil glands in Ocimum basilicum L. (sweet basil). Ann Bot. 2002;90: 453–460. doi:10.1093/aob/mcf212

28. Chang X, Alderson PG, Wright CJ. Solar irradiance level alters the growth of basil (Ocimum basilicum L.) and its content of volatile oils. Environ Exp Bot. 2008;63: 216–223. doi:10.1016/j.envexpbot.2007.10.017

29. Chang X, Alderson P, Wright C. Effect of temperature integration on the growth and volatile oil content of basil (*Ocimum basilicum* L.). J Hortic Sci Biotechnol. 2005;80: 593–598. doi:10.1080/14620316.2005.11511983

30. Banchio E, Xie X, Zhang H, Pare PW. Soil Bacteria Elevate Essential Oil Accumulation and Emissions in Sweet Basil. J Agric Food Chem. 2009; 653–657. doi:10.1021/jf8020305

31. Kim H, Chen F, Wang X, Rajapakse N. Effect of chitosan on the biological properties of sweet basil (Ocimum basilicum L.). J Agric Food Chem. 2005;53: 3696–3701. doi:10.1021/kjf0480804

32. Said-Al Ahl HAH, Mahmoud AA. Effect of zinc and / or iron foliar application on growth and essential oil of sweet basil (Ocimum basilicum L .) under salt stress. Ozean Journal of Applied Sciences. 2010;3: 97–111.

33. Johnson AJ, Hopfer H, Heymann H, Ebeler SE. Aroma Perception and Chemistry of Bitters in Whiskey Matrices: Modeling the Old-Fashioned. Chemosens Percept. 2017; 1–14. doi:10.1007/s12078-017-9229-3

34. Jones DR, Schonlau M, Welch WJ. Efficient Global Optimization of Expensive Black-Box Functions. J Glob Optim. 1998;13: 455–492. doi:10.1023/a:1008306431147

35. Koziel S, Ciaurri DE, Leifsson L. Surrogate-Based Methods. Computational Optimization, Methods and Algorithms. 2011. pp. 33–59.

36. Shahriari B, Swersky K, Wang Z, Adams RP, De Freitas N. Taking the human out of the loop: A review of Bayesian optimization. Proc IEEE. 2016;104: 148–175. doi:10.1109/JPROC.2015.2494218

37. Adams SR, Langton FA. Photoperiod and plant growth: A review. J Hortic Sci Biotechnol. 2005;80: 2–10. doi:10.1080/14620316.2005.11511882

38. Yamaura T, Tanaka S, Tabata M. Light-dependent formation of glandular trichomes and monoterpenes in thyme seedlings. Phytochemistry. 1989;28: 741–744. doi:10.1016/0031-9422(89)80106-2

39. Skrubis B, Markakis P. The Effect of Photoperiodism on the Growth and the Essential Oil of Ocimum basilicum (Sweet Basil). Econ Bot. 1976;30: 389–393.

40. Koza JR. Symbolic Regression-Error-Driven Evolution. Genetic Programming I: On the Programming of Computers by Means of Natural Selection. 1992; 237–288.

41. Rodriguez Rafael GD, Solano Salinas CJ. Empirical study of surrogate models for black box optimizations obtained using symbolic regression via genetic programming. Proceedings of the 13th annual conference companion on Genetic and evolutionary computation. ACM; 2011. pp. 185–186.

42. Stijven S, Vladislavleva E, Kordon A, Willem L, Kotanchek ME. Prime-Time: Symbolic Regression Takes Its Place in the Real World. Genetic Programming Theory and Practice XIII. Springer; 2016. pp. 241–260.

43. Smits GF, Kotanchek M. Pareto-front exploitation in symbolic regression. Genetic programming theory and practice II. Springer; 2005. pp. 283–299.

44. Schmidt M, Lipson H. Distilling free-form natural laws from experimental data. Science (80-). American Association for the Advancement of Science; 2009;324: 81–85.

45. Du Q, Faber V, Gunzburger M. Centroidal Voronoi tessellations: Applications and algorithms. SIAM Rev. SIAM; 1999;41: 637–676.

46. Bergstra JS, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. Advances in Neural Information Processing Systems. 2011. pp. 2546–2554.

47. Snoek J, Larochelle H, Adams RP. Practical bayesian optimization of machine learning algorithms. Advances in neural information processing systems. 2012. pp. 2951–2959.

48. González J, Dai Z, Hennig P, Lawrence N. Batch bayesian optimization via local penalization. Artificial Intelligence and Statistics. 2016. pp. 648–657.

49. Srinivas N, Krause A, Kakade SM, Seeger M. Gaussian process optimization in the bandit setting: No regret and experimental design. arXiv Prepr arXiv09123995. 2009;

50. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. Nature Research; 2015;521: 436–444.

51. Snoek J, Rippel O, Adams RP. Scalable Bayesian Optimization Using Deep Neural Networks. Int Conf Mach Learn. 2015; 2171-2180.

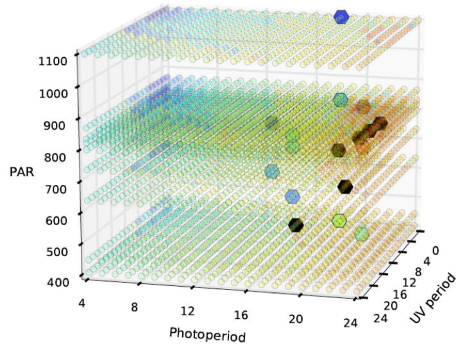
52. Deb K, Myburgh C. Breaking the billion-variable barrier in real-world optimization using a customized evolutionary algorithm. Proceedings of the 2016 on Genetic and Evolutionary Computation Conference. ACM; 2016. pp. 653–660.

53. Knowles J. ParEGO: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. IEEE Trans Evol Comput. 2006;10: 50–66.

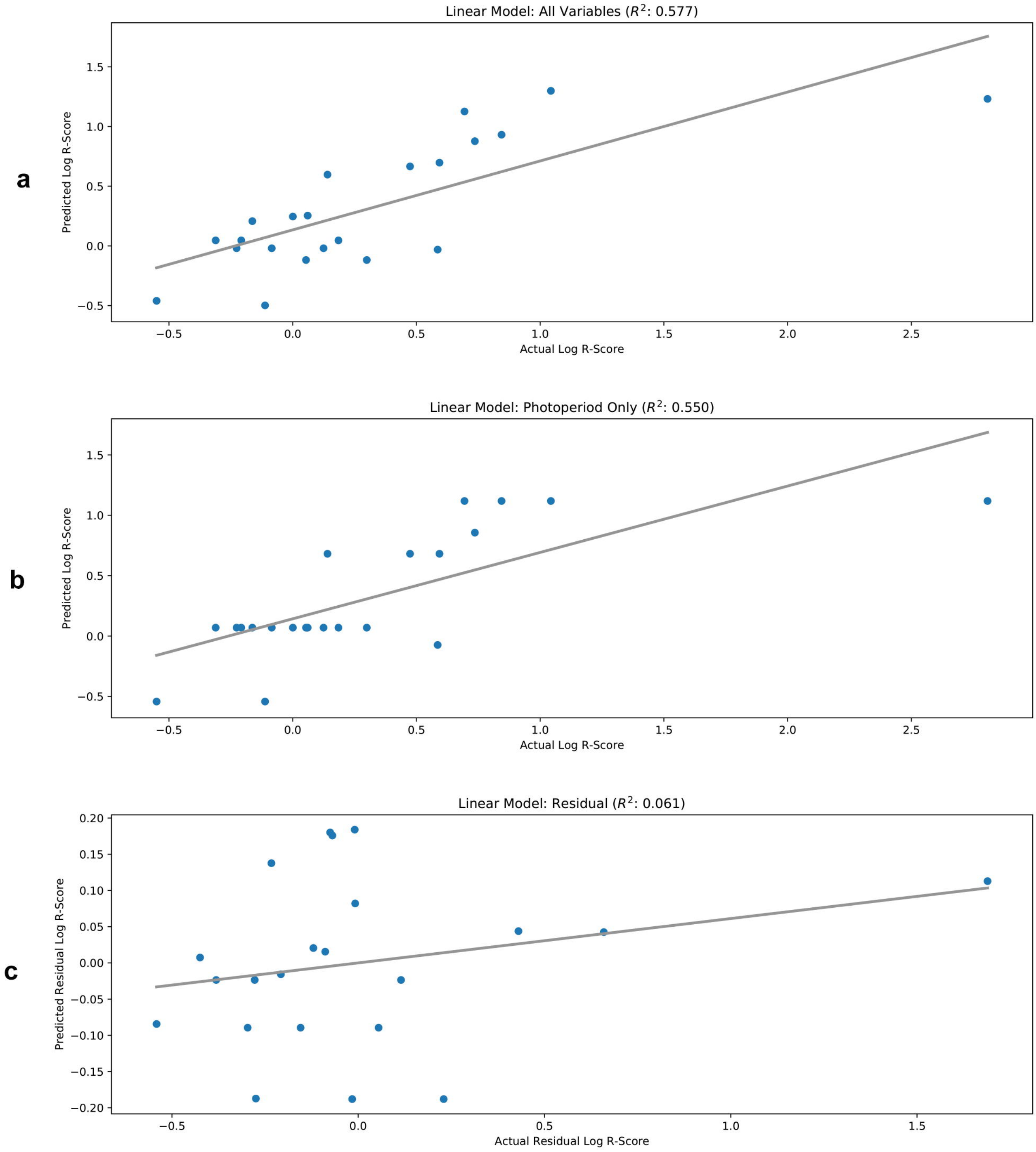
54. Jin Y. Surrogate-assisted evolutionary computation: Recent advances and future challenges. Swarm Evol Comput. Elsevier B.V.; 2011;1: 61–70. doi:10.1016/j.swevo.2011.05.001

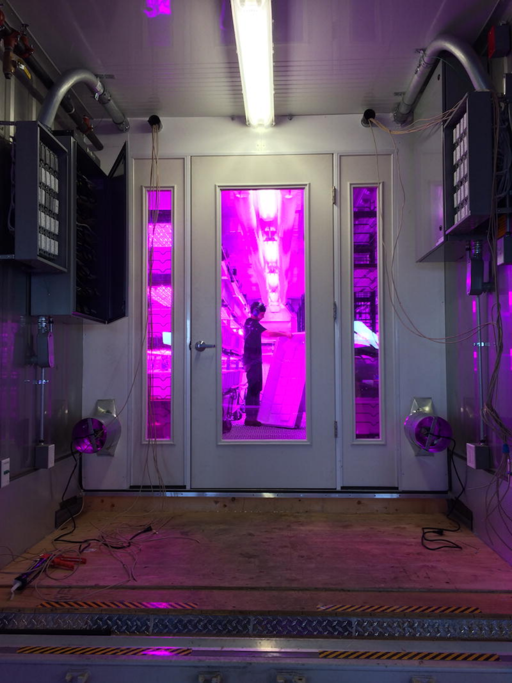
55. Lehman J, Miikkulainen R. Neuroevolution. Scholarpedia. 2013;8: 30977.

56. Miikkulainen R, Iscoe N, Shagrin A, Cordell R, Nazari S, Schoolland C, et al. Conversion rate optimization through evolutionary computation. Proceedings of the Genetic and Evolutionary Computation Conference. ACM; 2017. pp. 1193–119



1100
1000
900
800
700
600
500
400
PAR
4
8
12
16
20
24
Photoperiod
0
4
8
12
16
20
24
UV period


a
Linear Model: All Variables (R^2: 0.577)
Predicted Log R-Score
Actual Log R-Score
b
Linear Model: Photoperiod Only (R^2: 0.550)
Predicted Log R-Score
Actual Log R-Score
c
Linear Model: Residual (R^2: 0.061)
Predicted Residual Log R-Score
Actual Residual Log R-Score




EXIT