PERSPECTIVE

ePlant for quantitative and predictive plant science research in the big data era
—Lay the foundation for the future model guided crop breeding, engineering and agronomy

Yi Xiao[1], Tiangen Chang[2], Qingfeng Song[1], Shuyue Wang[2], Danny Tholen[2], Yu Wang[2], Changpeng Xin[2], Guangyong Zheng[2], Honglong Zhao[1] and Xin-Guang Zhu[1,2,*]

[1] Shanghai Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai 200032, China
[2] Plant Systems Biology Research Group, Partner Institute for Computational Biology, Chinese Academy of Sciences, Shanghai 200031, China
* Correspondence: zhuxg@sippe.ac.cn

Background: The increase in global population, climate change and stagnancy in crop yield on unit land area basis in recent decades urgently call for a new approach to support contemporary crop improvements. ePlant is a mathematical model of plant growth and development with a high level of mechanistic details to meet this challenge.
Results: ePlant integrates modules developed for processes occurring at drastically different temporal (10^{-8}–10^6 seconds) and spatial (10–10–10 meters) scales, incorporating diverse physical, biophysical and biochemical processes including gene regulation, metabolic reaction, substrate transport and diffusion, energy absorption, transfer and conversion, organ morphogenesis, plant environment interaction, etc. Individual modules are developed using a divide-and-conquer approach; modules at different temporal and spatial scales are integrated through transfer variables. We further propose a supervised learning procedure based on information geometry to combine model and data for both knowledge discovery and model extension or advances. We finally discuss the recent formation of a global consortium, which includes experts in plant biology, computer science, statistics, agronomy, phenomics, etc. aiming to expedite the development and application of ePlant or its equivalents by promoting a new model development paradigm where models are developed as a community effort instead of driven mainly by individual labs' effort.
Conclusions: ePlant, as a major research tool to support quantitative and predictive plant science research, will play a crucial role in the future model guided crop engineering, breeding and agronomy.

Keywords: systems modeling; quantitative; predictive; homeostasis; multiscale; crop *in silico*

In 2011, we proposed the need to develop ePlant [1], a highly mechanistic model of plant growth and developmental processes throughout the whole plant growth cycle, which will differ from all previous crop models by having detailed mechanistic basis of all processes spanning from molecular reactions up through plant environment interactions. Rapid progress has been made in recent years in development of the component modules (or sub-models), theoretical tools and applications around ePlant. In this perspective paper, we overview the original rationale, concept, components for ePlant and a method for its development. We then propose a theoretical framework to develop and apply ePlant in the big data era. Finally, we discuss recent efforts in developing an international consortium on promoting quantitative and predictive plant science research, with the realization of ePlant being one of the central goals.

## WHAT IS ePLANT AND WHY DO WE NEED TO DEVELOP IT?

ePlant will be a mathematical model which aims to

simulate the dynamic plant growth and development process throughout its growth cycle. It differs from the earlier crop models, such as APSIM [2] and DSSAT crop models [3], by explicitly simulating the detailed mechanisms underlying different processes. It spans scales from organelle, cell, tissue, organ, whole plant to ecosystem levels; it includes the processes spanning gene regulation, metabolic process, metabolite transport at the tissue and organ levels, organ morphogenesis, and plant environment interactions (Figure 1). We envisage ePlant will become a pivotal tool in the predictive and quantitative plant science research in the modern big data era.

Firstly, ePlant or the sub-models used in ePlant, can be used as a basic tool for quantitative study of diverse plant systems, such as the regulatory circuits controlling the stability of plant metabolic systems under different conditions [4], mechanistic basis of the biophysical signals, such as the chlorophyll fluorescence induction curve [5,6] and mesophyll conductance [7], and identification of optimal agronomic practices for improved biomass production [8]. Similar to the earlier crop growth

models, ePlant can be used to guide crop management [3], selection of physiological traits for crop breeding [9] and predicting response of crops to changing climates [10–12].

Secondly, ePlant can be used as a critical component in the current general circulation models (GCMs) [13]. GCMs are models of circulation of planetary atmosphere or ocean, which can be used for weather forecasting, studying climate and climate change. Due to the large magnitude of CO_2 fluxes from terrestrial photosynthesis and respiration [14], terrestrial processes greatly influence the global carbon cycle. In current GCMs, compared to models of atmosphere and soil related physical processes, models representing plant growth and development are much less accurate. As a reflection of this, even for some of the best studied plant species, such as rice, no contemporary model can accurately predict its productivity under elevated CO_2 and temperature at different sites [12]. Furthermore, variations in predicted rice productivity are higher between individual crop models than variations resulting from 16 global climate model-based
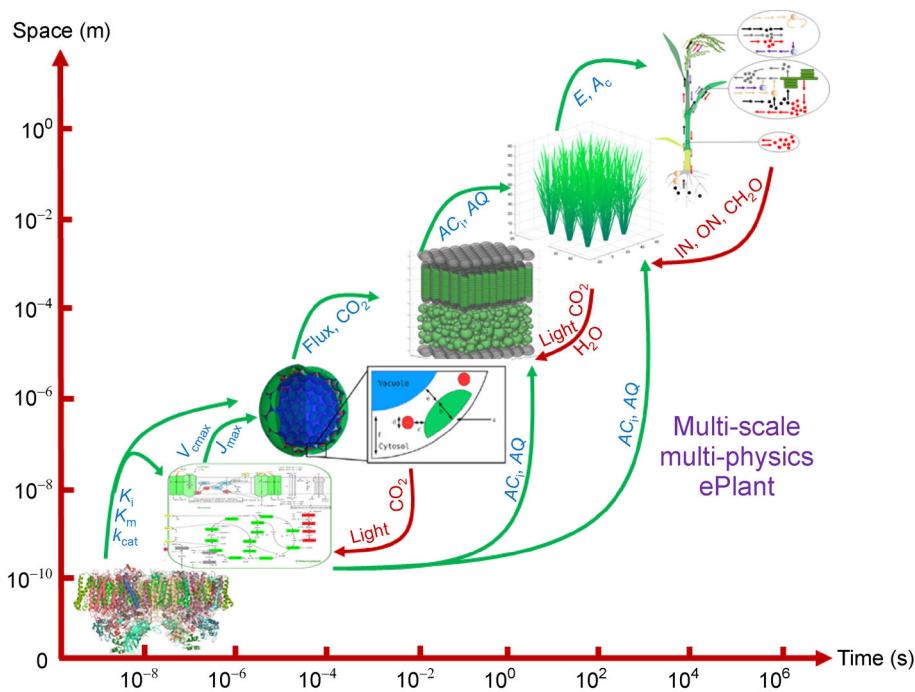
Figure 1.   Components of ePlant. ePlant includes components spanning a large range of temporal and spatial scales spanning macromolecular complexes, organelles, cells, tissues, canopies, whole plants and ecosystems. ePlant also includes different sets of biological processes including gene regulatory processes, metabolic processes, metabolite transport, organ morphogenesis and plant environmental interactions. Outputs of models describing processes at the lower temporal and spatial scales can be used as inputs to models describing processes at higher temporal and spatial scales. Some representative variables transferred between modules are labeled. (k_{cat}: catalytic number, K_m: Michaelis Menton kinetics; K_i: inhibition constant, V_{cmax}: Rubisco-limited rate of RuBP carboxylation; J_{max}: maximum electron transfer rate; AC_i: photosynthetic CO_2 uptake versus intercellular CO_2 concentration; AQ: photosynthetic CO_2 uptake rate versus photosynthetic photon flux density; E: transpiration rate; A_c: canopy photosynthetic CO_2 uptake rate, IN: inorganic nitrogen; ON: organic nitrogen; CH_2O: carbohydrate.)

scenarios [12]. One possible reason for this low predictive power is the lack of molecular details of plant growth and development in current crop models. ePlant, with a mechanistic description of plant growth and developmental processes and the interaction between plants and their environments, will drastically improve predictions of plants behavior under different climates, and hence improve the capacity of current GCMs in predicting climate and identifying strategies to cope at the changing climate.

Thirdly, ePlant can be used as a basic tool to support molecular design of crops to develop new strategies to improve crops for desirable traits, such as improved yield potential, improved grain quality, or higher stress tolerance or resource use efficiency [15]. This is currently especially relevant since it has become relatively easy to manipulate a gene or some gene combinations in plants, especially in agriculturally important crops. The main challenge that remains is to identify targets to be manipulated to gain the desired traits. Previously, through a systems modeling approach, a number of options to improve photosynthesis have been identified. For example, canopy photosynthesis models were used to identify the optimal Rubisco kinetic properties, photoprotective properties and canopy architectural parameters [16–19], dynamic systems models were used to identify genes controlling photosynthetic efficiencies in both natural and designed metabolisms [20–23], a reaction diffusion model of mesophyll cell was used to identify major limiting factors controlling mesophyll conductance [7] and leaf internal light prediction model was used to demonstrate the importance of different anatomical features on leaf photosynthetic rates [24]. Many of the identified options have been shown to be effective in enhancing photosynthesis and biomass production [25,26], demonstrating the effectiveness of this approach. We envisage that once ePlant is developed, it can be used to systematically evaluate different aspects of plants that holds potential to be improved for desirable features.

Fourthly, ePlant and the sub-models included in ePlant (as discussed in detail later) can be used to quantitatively represent the contemporary plant biology knowledge. Compared to a textual representation, either in the form of papers, textbooks or Wikipedia, the quantitative representation of plant biological knowledge encapsulated in ePlant or its sub-models can effectively facilitate communication among researchers specializing in different aspects of plant growth and development and hence promote cross-fertilization of ideas. Such quantitative representation will also help identify knowledge gaps in the current understanding of plant growth and development. Finally, ePlant and its modules can also be used as effective and visual teaching tools.

THE ESSENTIAL FUNCTIONAL MODULES OF ePLANT

To achieve the ePlant described above, at least four categories of functions are required. Different mathematical models therefore have been developed or need to be developed to realize the simulation of these functions.

Firstly, ePlant needs to explicitly incorporate the biophysical and biochemical mechanisms controlling photosynthesis and all the closely related metabolic processes, such as respiration, nitrogen assimilation etc. [27,28]. On this aspect, mechanistic models of the metabolic process of photosynthesis have been established now for C_3, C_4 and crassulacean acid metabolism [21,29,30]. In contrast, a mechanistic model of respiration is yet to be developed. In this line, it is worth to note that a mechanistic model of mitochondria energy generation in a human heart cell has been built [31] and a simplified model for plant respiratory processes has been built [32] earlier. A fully mechanistic model being able to predict interactions between photosynthesis, respiration and nitrogen assimilation is yet to be developed.

The availability of the substrate of photosynthesis, i.e., CO_2, is controlled by stomatal conductance and mesophyll conductance. Stomatal conductance is influenced by an array of internal metabolic processes and external environmental factors [33–37]. Different models of stomatal conductance with varying degree of mechanistic basis have been built [38,39]. So far, a fully mechanistic model of stomatal conductance is yet to be developed. Mesophyll conductance is another critical factor controlling leaf photosynthetic efficiency. Highly mechanistic models of mesophyll conductance have been built in recent years [7,40,41].

Leaf anatomy controls leaf photosynthesis by influencing leaf internal light environments, leaf internal CO_2 temperature profiles [42,43]. Efforts to model photosynthesis by considering leaf anatomy have been made recently [24,44]. Furthermore, considering the close interaction between plant primary metabolism and other secondary metabolism, combination of the kinetic systems models with genomic scale models of metabolic and regulatory processes [45] is needed to enhance the prediction accuracy of the future systems models. Such a combination will ultimately enable model to predict not only the crop yield potential, but also the quality of harvestable components since anabolism of different metabolites related to quality, such as starch composition and aroma related compounds, can be explicitly represented in such models.

Secondly, ePlant needs to predict the complete crop growth and developmental process [1]. Along this line,

models with different degree of mechanistic basis have been built to simulate plant developmental processes, e.g., gradual formation of 3D canopies with time [46,47], flowering [48], shoot patterning [49], flow of photo-synthate from source to sink [50,51], 3D root growth dynamics [52,53], etc. So far, however, a mechanistic model of partitioning of assimilates among different organs is yet to be developed [54].

Thirdly, ePlant needs to predict acclimation of plant metabolism and structure under different environments. Hence modules simulating the gene regulatory processes and signal transduction processes related to crop growth and development are needed. Predicting variation of phenotypes under different genotype × environment × management combinations is the holy grail of crop systems models research [55]. Most of the research on this topic is still in its infancy. Development of a gene regulatory network (GRN), which incorporates the interaction of all involved regulatory cis-elements and trans-factors, is a critical task towards such a goal. Various bioinformatics algorithms, based on either correlation, or features selection, or probabilistic graph models etc., have been developed, which use genomic scale omics data, in particular, the transcriptomics data, to build GRNs [56–58]. There is only limited number of GRNs developed for plant related processes, such as flowering date determination [59], photoperiod and circadian clock [60], and seed setting [61]. However, so far, these GRNs are not linked to current crop systems models. It is worth noting that a GRN related to circadian clock has been linked to an *Arabidopsis* model [62]. Such disconnection between GRNs and crop models partially explains why the current models, though capable of predicting performance of crops in the regions where they are parameterized, cannot predict crop performance as accurately once beyond their regions or environmental conditions or cultivars used for their parameterization. Here we emphasize that though some work has been done in developing GRNs based on transcriptomics data, models predicting the regulatory processes at post-translational levels, including transcript stability and degradation, translation, post-translational modification etc., are largely lacking. There is a long way to go before any realistic model of predicting the acclimation of plants under different environments becomes available.

Fourthly, in addition to the above discussed biological processes, ePlant needs to include models of interaction between plants and their surrounding soil and atmosphere. These interactions control plant growth and development. Modeling plant-environment interaction requires simulation of soil hydraulic dynamics, nutrient cycles and temperature profiles etc., which are the basis for predicting the soil water status and nutrient availability to roots. The microclimates inside the canopy, such as light, temperature, CO_2, humidity and wind speed also need to be incorporated in a crop systems models [19,63], to ensure an accurate prediction of the exchange of gas, water and momentum between canopy and atmosphere. Models of soil related processes have been well developed, i.e., CENTURY model [64,65]; while fully integrated canopy photosynthesis and microclimate models are yet to be developed. ePlant needs to integrate the above-ground processes with the below ground processes to develop a fully integrated microclimate model, including linking soil water status with the leaf biological and hydrological processes [38,66,67].

USING DIVIDE-CONQUER AND TRANS-FER VARIABLES TO REALIZE THE MULTI-SCALE, MULTI-PHYSICS ePLANT

As discussed above, ePlant includes modules describing processes at different temporal and spatial scales, with each process at particular scales potentially represented by different modules and each module potentially using different methods (see Figure 1 and Table 1). Therefore, ePlant is not a single model, rather it is an assembly of modules which can be combined to form models with different temporal, spatial and physical resolutions. ePlant development follows a two-step strategy, i.e., first divide-and-conquer to develop individual modules and then integrate modules through transfer variables. When we divide plant growth and developmental processes into different units, i.e., modules, we follow the principle of maximizing connections within modules while minimizing connections between modules, as did during development of the ePhotosynthesis models [20,29,78]. The connectivity between photosystem II unit with other components of ePhotosynthesis is minimal, which justifies development of an independent module for PSII photochemistry and biophysical processes [78]; similarly, the connectivity between the photosynthetic carbon metabolism with that of photosynthetic light reactions is relatively less, which justifies the development of an independent model of photosynthetic carbon metabolism [20]. Another principle that can be used to divide modules is to separate reactions/processes occurring at drastically different time scales because every process in ePlant can be viewed as dynamic at a higher time resolution; similarly, every step can be viewed as a steady-state process if viewed at a lower time resolution. Processes at similar temporal and spatial scales can be grouped together as a module. ePlant hence includes modules working at different temporal and spatial scales, i.e., ecosystems level, crop physiology level, metabolism level, and gene regulatory network level (Figures 1 and 2). Transfer variables, which are defined as an output of a lower level modules, which at the same time are also

Table 1.   Components of ePlant.

Physical and biological processes	Modeling methods	Example models	Refs.
Ecophysiological processes	Ordinary differential system	APSIM, DSSAT, WIMVOAC	[68–71]
Physiological processes	Ordinary differential system	Photosynthesis, sunlit shaded model	[72,73]
Metabolic processes	Ordinary differential equation models	Photosynthesis, starch metabolism	[29,74]
Metabolic process at the whole genome scale	Constraint based modeling	Aragem, C4gem	[75,76]
Gas diffusion, nutrient cycling, water cycling processes	Reaction diffusion models	Mesophyll conductance	[7,23,77]
Light propagation process	Ray tracing algorithms	Rice canopy model, sugarcane model	[8,19,24]
Morphogenesis	L systems, Greenlab	Maize, rice	[46,47]
Gene regulatory process	Probabilistic graph model, information theory, correlation	Circadian rhythm, seed setting	[60,61]
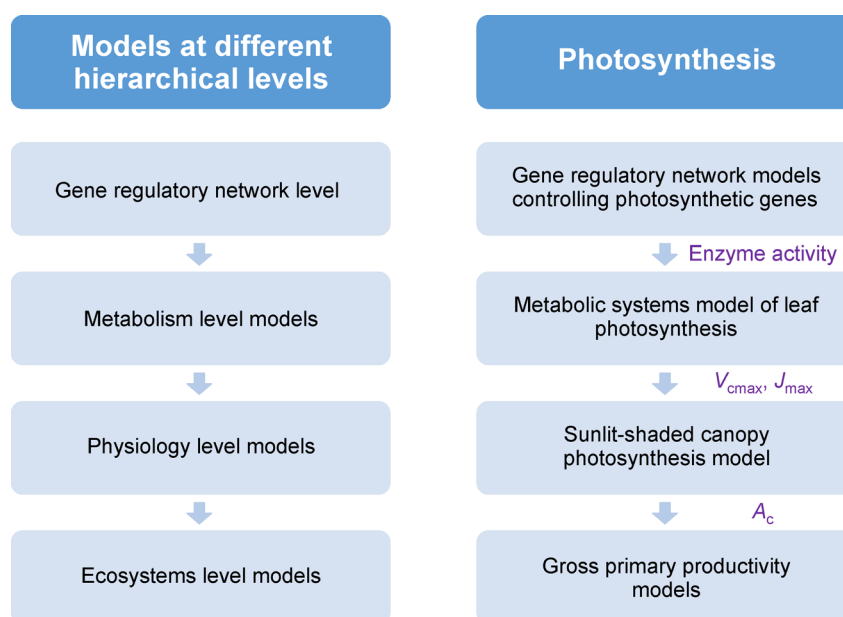
Figure 2.   Strategy used to build the multi-scale multi-physics ePlant model. The multi-scale multi-physics ePlant is developed using a divide-and-conquer strategy and transfer variables. Processes involved in ePlant spanning multiple scales are represented in a multi-scale framework. Here we use the models related to C_3 photosynthesis to illustrate the concept of transfer variable. Specifically, enzyme activity is the transfer variable between the gene regulatory network model and metabolic systems model, the Rubisco-limited rate of RuBP carboxylation (V_{cmax}) and maximal electron transfer rate (J_{max}) are the transfer variables between metabolic systems model and sunlit shaded canopy photosynthesis model; while canopy photosynthesis rate (A_c) is the transfer variable between sunlit-shaded canopy photosynthesis model and the gross productivity model.

inputs to higher levels, are used to integrate modules at these different scales (Figure 1).

Photosynthetic CO_2 uptake occurs at different temporal and spatial scales. Here we use modules of photosynthetic CO_2 uptake to illustrate how transfer variables are used to integrate modules at different scales. At the ecosystem scale, photosynthetic CO_2 uptake can be predicted using a sunlit-shaded model which calculates canopy photosynthesis by summing up CO_2 uptake rate of both sunlit and shaded leaves [72]. At the leaf scale, photosynthetic CO_2 uptake can be predicted using models which explicitly describes both the leaf anatomy and leaf metabolic processes [24]. The leaf scale photosynthetic CO_2 uptake can also be predicted with a steady state biochemical model with consideration of stoichiometric relationship between reactions [73]. Photosynthetic CO_2 uptake rate at the metabolism scale can be predicted with a dynamic systems model with consideration of both the stoichio-

metry and also enzyme kinetics [20]. Photosynthetic CO_2 uptake rate at the level of gene regulatory network can be predicted with a detailed consideration of the regulatory processes influencing photosynthesis [79]. If we need to integrate a physiological model of canopy photosynthesis, e.g., a sunlit-shaded model [72], with a dynamic systems model of C_3 photosynthesis, the Rubisco-limited RuBP carboxylation rate (V_{cmax}) and maximal rate of electron transfer rate (J_{max}) can be used as transfer variables. Specifically, we can use the dynamic systems model of photosynthetic metabolism, such as the C_3 carbon metabolism model [20] to predict responses of photosynthetic CO_2 uptake rates (A) under different CO_2 levels, which can be used to infer V_{cmax} and J_{max}. These two transfer variables can then be used as inputs to the physiological level models to predict photosynthesis at the canopy level under different environments. Such an integration combining canopy photosynthesis model and metabolism model enables examination of the impacts of manipulating different enzymes on canopy photosynthesis. Similarly, if a kinetic model of gene regulatory processes controlling photosynthesis development is available, it can be used to predict the quantity of different proteins involved in photosynthesis, which can then be used as transfer variables for metabolic systems models.

The above discussed model integration process works well if models working at different scales are described continuous processes. However, models for continuous processes and discrete processes can not be integrated using this method. Under such circumstances, a probabilistic regulation of metabolism algorithm, which has been developed to link GRNs to a constraint based genomic scale metabolism model for *E. coli* [80], can be used. In all these model integration processes, it is important to ensure that the known constraints, such as stoichiometric constraints of the biomass composition and growth rate [81], are maintained.

Here we emphasize that ePlant will not be one model, rather it will be a series of continuously evolving models with gradually increased mechanistic details with time. The level of mechanistic details needed for any particular realization of ePlant depends on the question to be addressed. Therefore, though development of the first integrative ePlant model is a concrete goal, development and improvement of ePlant will be a continuously ongoing work. Considering that modules describing different plant processes have different levels of mechanistic details, therefore, ePlant developed at any particular time point will inevitably be a mosaic of modules with different mechanistic details.

A THEORETICAL FRAMEWORK TO SUPPORT PREDICTIVE AND QUANTITATIVE PLANT BIOLOGY RESEARCH IN THE BIG DATA ERA

ePlant is a mathematical representation or integration of the current knowledge about a living plant. Each component or process or action on plants can be abstracted as a term used to describe the component, function or application of ePlant (Table 2). In a broad sense, everyone has his or her model, which is used to interpret experimental observations, analogous to the process of fitting model parameters to a mathematical model, though in a qualitative way. During a typical research project, we explore the unknown and extend the boundary of our knowledge by studying a difference that cannot be explained by current knowledge or "model". Push this analogy even further, when experiments are designed and results are compared between different groups, we are in some sense studying phenotypic variations with different models embraced by different labs. Unfortunately, due to the complex nature of plants, every "model" is right only to certain degree and no "model" is absolutely right [82]. The process of pushing "models" closer and closer to the absolute "truth" can be seen as the essence of scientific research. This same process occurs during the development of ePlant and its component models, i.e., ePlant will become a better and better representation of the reality with its gradually improved capacity to predict the distribution of output variables with the distribution of the input variables (Figure 3).

Table 2. Mapping between terms describing plants and terms used in ePlant and its component modules.

Terms related to plants	Terms in ePlant and its component modules
Compound or substrate	A variable in a module
A process	A module or sub-model
Plant	ePlant: a highly mechanistic plant growth and development model
Soil and atmospheric conditions surrounding the plants	Boundary conditions of the ePlant model
Physiological parameter	A predicted parameter from the model
Natural variation	Variations of model structure, variable and output
Evolutionary process	Evolutionary algorithm
Genetic engineering	Modification of parameters related to gene or proteins in the model
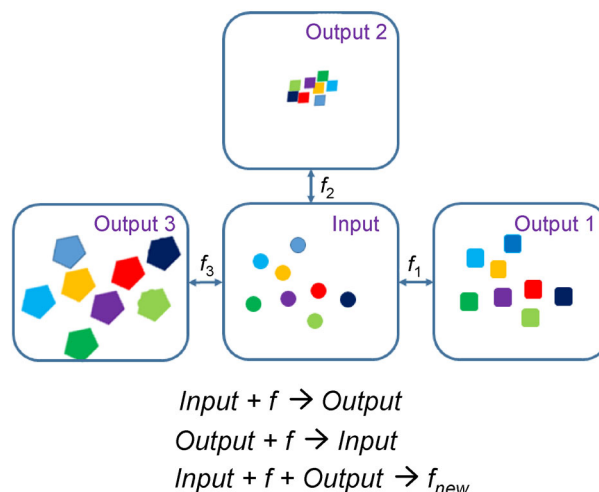
Figure 3. **Mapping valiations between input and output variables.** The relationship between variation in the input variable, variations of output variables, and the function f (f_1, f_2, and f_3) which represents the link between input and output variables. The variation of the input variables and the functions together determine the variation of the output variables. Mapping the variation of an input variable and that of an output variable reveals information regarding f. The variation of either input or output variables can be used together with f to deduce variation about the output or input. If variation of both input and output can be used together with f, new information about f can be deduced.

When simulations are used to draw conclusions about a particular process, the reliability of the model prediction is crucial. Four types of errors can potentially create "artificial" difference between model predictions and real plants: i) errors caused by inaccurate and imprecise experimental techniques or operations; ii) errors introduced during model parameterization, i.e., parameters measured *in vitro* may not represent those *in vivo*, and even parameters estimated *in vivo* may still biased due to limitation of technologies; iii) errors due to uncertainty of the model, especially when the model is used to represent a process for which a complete mechanistic understanding is unavailable, either due to unknown variables or unknown relationships among variables, and hence some empirical equations or relationships derived from limited data are used; iv) errors due to the model structure. Simulation of a particular phenomenon needs a model with appropriate spatial and temporal scales. If a model's temporal and spatial resolution is too high for a question to study, too many unnecessary assumptions will be introduced and hence magnifying potential structural errors. If a model's temporal or spatial scale is too low for the question to study, the model will unlikely generate novel insights regarding the questions under study.

A theoretical framework therefore needs to be developed to enable studying these different errors and their impact on model behaviors. Minimally, the framework needs to address following questions: how much will the bias in measurable and non-measurable parameters influence the reliability of our simulations? How much

will the uncertainty of the model itself influence the reliability of model simulations? How much will the scale of model influence the reliability of model simulations? How to unify models developed with different temporal and spatial resolutions and mechanistic details while maintain the essential prediction capacity? How to interpret the potential bias of experimental measurements? How much will this bias influence the comparison between experiment and simulation, and the reliability of conclusions? If for a particular phenomenon no mechanistic understanding is available, how can information from experimental data still be effectively used in model simulations?

On this aspect, mathematical theories such as information geometry [83] can potentially be adapted and used to support studies as discussed above. Theoretically, information geometry takes a model as a function/mapping between experiment measurements (outputs) and model parameters (inputs), model structure therefore is equivalent to certain shape of a manifold in a hyperspace [83]. Although the relationship between model input parameters to measured phenotypic output parameters are many-for-one, with the variation of the model parameters, it is possible to estimate the confidence intervals of the model output variables, i.e., creating an ensemble of input parameters and using these to predict the distribution of model outputs and hence deriving the potential confidence intervals of the model outputs. Conversely, if the variation of a particular physiological parameters (or model output) is known, it is possible to deduce the potential variation of certain input model parameters as

well. The deduced variability of model parameters can inform us about the level of feasibility and effectiveness of engineering a particular plant trait for a desired biological output. If a deduced input variable shows little variation, it would be less feasible to manipulate this variable; furthermore, even if a deduced input variable values show large variation but it has little impacts on output parameter, it is unlikely that this parameter will be an effective parameter to modify (Figure 4).

In this sense, the concept of ePlant will include not only the model itself, it will also include a theoretical framework to enable predictive and quantitative plant science research. Finally here we emphasize that though great amounts of experimental data have been collected by the plant science research community, however, most of these data only cover a limited number of variables and thus have limited value in promoting identification of new knowledge gap in current plant science using ePlant. To study the above discussed questions, carefully designed internally consistent data sets need to be collected systematically, in particular on those parameters related to the expanded model components. Here the internally consistent data sets refer to those data collected on the same plants grown under the same condition and at the same developmental stages. Such data will be crucial to

verify each module and the integration of different modules.

With a validated model available, any further difference between model simulations and new experimental observations can help target potential causes, design specialized experiments, discover unknown factors or mechanisms related to a particular area [84]. Such a process will also urge development of new methodology and technology to measure key parameters limiting the development of current knowledge/models. Such an iterative model development, validation, improvement process, or supervised learning process, has the potential to become a new paradigm of the future quantitative and predictive plant science research.

ePlant will become a crucial tool to integrate and use the diverse data in the big data era. Big data, including genomes, transcriptomes, proteomes, metabolomes, and different phenomes, can be regarded as either input or output for ePlant or its component models. Mapping between ePlant or its component models with natural variations in these data poses a tremendous challenge and offers huge opportunities for development of new algorithms, tools and frameworks (Figure 3). Only after these tools and frameworks are fully established, the great promise offered by ePlant to help guide future crop
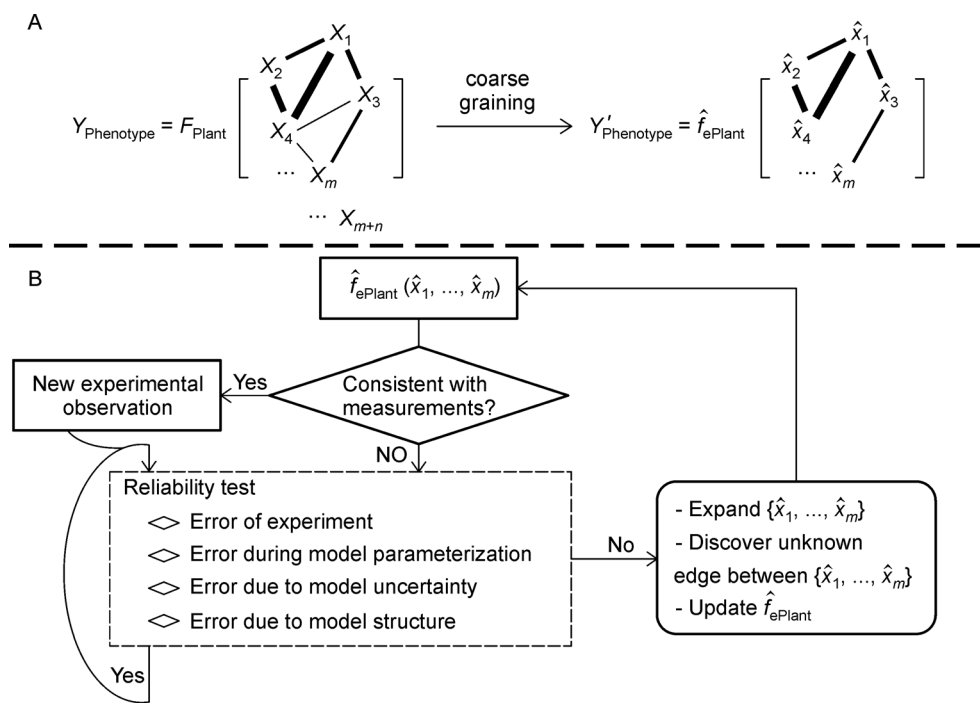
Figure 4. Elements of a new research paradigm supporting quantitative and predictive plant science research. (A) A coarse graining representation of phenotypes where F_{plant} represents real plants; $Y_{phenotype}$ represents the phenotype of plants; $Y'_{phenotyp}$ represents the observed modeled plant phenotype; $\hat{f}_{ePlant}$ represents the ePlant model, which is simplified representation of plant functions; $X_1, X_2, ..., X_{m+n}$ represent mechanism of plant growth and development; $\hat{x}_1, \hat{x}_2, ..., \hat{x}_m$ represent the variables included in ePlant. (B) A theoretical framework of a supervised learning process during the ePlant model development and its application.

engineering, breeding, and agronomy can be realized. From this perspective, the creation of ePlant model itself is only the first step on this New Long March.

FINAL COMMENTS: THE GLOBAL EFFORTS

Model plant species, such as *Arabidopsis*, rice and maize, for which vast amount of genetic resources, background knowledge and efficient transformation protocols are available [85–88], are likely to be the first set of plants that will be used to realize ePlant. Here we highlight a number of recent advances on development of ePlant or its equivalents. Chew *et al.* [62] developed a multi-scale digital *Arabidopsis* which can predict organ and whole organism growth. Zhu *et al.* [89] proposed the development of a collaborative model development platform, i.e., *Plant in silico*, which includes not only the basic modules, data for model parameterization and validation, but also the basic algorithmic tools for model application, visualization, etc. With developing *Plant in silico* as a goal, a *Crop in silico* international consortium was recently proposed [90]. The Department of Energy of the United States started an Integrative Plant Air Soil Systems (iPASS) initiative, aiming at creation of an integrative plant systems model, which, when combined with plant ecosystems phenomics, can be used to study the interaction between plants, microbiome, atmosphere and soil [91]. It is foreseeable that development of ePlant and the associated algorithms and resources, both for models development and their application, will become a nucleus to integrate research activities spanning diverse disciplines, including plant biology, computer science, computer vision, high performance computing, agronomy, phenomics, for decades to come, or to put it simply, function as the nexus of the future predictive and quantitative plant science research, which has the potential to transform the future agriculture by harvesting the power of model guided crop engineering, breeding and agronomy.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Yi Xiao, Tiangen Chang, Qingfeng Song, Shuyue Wang, Danny Tholen, Yu Wang, Changpeng Xin, Guangyong Zheng, Honglong Zhao and Xin-Guang Zhu declare that they have no conflict of interests.

This article is a perspective article and does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Zhu, X.-G., Zhang, G. L., Tholen, D., Wang, Y., Xin, C. P. and Song, Q. F. (2011) The next generation models for crops and agro-ecosystems. Sci. China Inf. Sci., 54, 589–597

2. Hammer, G. L., van Oosterom, E., McLean, G., Chapman, S. C., Broad, I., Harland, P. and Muchow, R. C. (2010) Adapting APSIM to model the physiology and genetics of complex adaptive traits in field crops. J. Exp. Bot., 61, 2185–2202

3. Ruíz-Nogueira, B., Boote, K. J. and Sau, F. (2001) Calibration and use of CROPGRO-soybean model for improving soybean management under rainfed conditions. Agric. Syst., 68, 151–173

4. Ma, W., Trusina, A., El-Samad, H., Lim, W. A. and Tang, C. (2009) Defining network topologies that can achieve biochemical adaptation. Cell, 138, 760–773

5. Xin, C. P., Yang, J. and Zhu, X.-G. (2013) A model of chlorophyll a fluorescence induction kinetics with explicit description of structural constraints of individual photosystem II units. Photosynth. Res., 117, 339–354

6. Xiao, Y. and Zhu, X.-G. (2016) Chlorophyll fluorescecence and stable isotope signals in photosynthesis research. Plant Physiology Journal (in Chinese), 52, 1663–1670

7. Tholen, D. and Zhu, X.-G. (2011) The mechanistic basis of internal conductance: a theoretical analysis of mesophyll cell photosynthesis and CO_2 diffusion. Plant Physiol., 156, 90–105

8. Wang, Y., Song, Q., Jaiswal, D., de Souza, A. P., Long, S. P. and Zhu, X.-G. (2017) Development of a three dimensional ray-tracing model of sugarcane canopy photosynthesis and its applications in assessing impacts of varied row spacing. Bioenerg Res., doi: 10.1007/s12155-017-9823-x

9. Zheng, B., Biddulph, B., Li, D., Kuchel, H. and Chapman, S. (2013) Quantification of the effects of VRN1 and Ppd-D1 to predict spring wheat (*Triticum aestivum*) heading time across diverse environments. J. Exp. Bot., 64, 3747–3761

10. Tubiello, F. N., Soussana, J.-F. and Howden, S. M. (2007) Crop and pasture response to climate change. Proc. Natl. Acad. Sci. USA, 104, 19686–19690

11. Miguez, F. E., Zhu, X., Humphries, S., Bollero, G. A. and Long, S. P. (2009) A semimechanistic model predicting the growth and production of the bioenergy crop *Miscanthus×giganteus*: description, parameterization and validation. GCB Bioenergy, 1, 282–296

12. Li, T., Hasegawa, T., Yin, X., Zhu, Y., Boote, K., Adam, M., Bregaglio, S., Buis, S., Confalonieri, R., Fumoto, T., *et al.* (2015) Uncertainties in predicting rice yield by current crop models under a wide range of climatic conditions. Glob. Change Biol., 21, 1328–1341

13. Sellers, P. J., Randall, D. A., Collatz, G. J., Berry, J. A., Field, C. B., Dazlich, D. A., Zhang, C., Collelo, G. D. and Bounoua, L. (1996) A revised land surface parameterization (SiB2) for atmospheric GCMs. part I: model formulation. J. Clim., 9, 676–

705

14. Falkowski, P., Scholes, R. J., Boyle, E., Canadell, J., Canfield, D., Elser, J., Gruber, N., Hibbard, K., Högberg, P., Linder, S., *et al.* (2000) The global carbon cycle: a test of our knowledge of earth as a system. Science, 290, 291–296

15. Xue, Y., Chong, K., Han, B., Gui, J., Wang, T., Fu, X., He, Z., Chu, C., Tian, Z., Cheng, Z., Lin, S. (2015) New chapter of designer breeding in China: update on strategic program of molecular module-based designer breeding systems. Buttletin of Chinese Academy of Sciences, 30, 393–402

16. Zhu, X.-G., Portis, A. R. Jr and Long, S. P. (2004) Would transformation of C$_3$ crop plants with foreign Rubisco increase productivity? A computational analysis extrapolating from kinetic properties to canopy photosynthesis. Plant Cell Environ., 27, 155–165

17. Zhu, X.-G., Ort, D. R., Whitmarsh, J. and Long, S. P. (2004) The slow reversibility of photosystem II thermal energy dissipation on transfer from high to low light may cause large losses in carbon gain by crop canopies: a theoretical analysis. J. Exp. Bot., 55, 1167–1175

18. Drewry, D. T., Kumar, P. and Long, S. P. (2014) Simultaneous improvement in productivity, water use, and albedo through crop structural modification. Glob. Change Biol., 20, 1955–1967

19. Song, Q.-F., Zhang, G. and Zhu, X.-G. (2013) Optimal crop canopy architecture to maximise canopy photosynthetic CO$_2$ uptake under elevated CO$_2$ – a theoretical study using a mechanistic model of canopy photosynthesis. Funct. Plant Biol., 40, 108–124

20. Zhu, X.-G., de Sturler, E. and Long, S. P. (2007) Optimizing the distribution of resources between enzymes of carbon metabolism can dramatically increase photosynthetic rate: a numerical simulation using an evolutionary algorithm. Plant Physiol., 145, 513–526

21. Wang, Y., Long, S. P. and Zhu, X. G. (2014) Elements required for an efficient NADP-malic enzyme type C$_4$ photosynthesis. Plant Physiol., 164, 2231–2246

22. Xin, C. P., Tholen, D., Devloo, V. and Zhu, X. G. (2015) The benefits of photorespiratory bypasses: how can they work? Plant Physiol., 167, 574–585

23. Wang, S., Tholen, D. and Zhu, X. G. (2017) C$_4$ photosynthesis in C$_3$ rice: a theoretical analysis of biochemical and anatomical factors. Plant Cell Environ., 40, 80–94

24. Xiao, Y., Tholen, D. and Zhu, X.-G. (2016) The influence of leaf anatomy on the internal light environment and photosynthetic electron transport rate: exploration with a new leaf ray tracing model. J. Exp. Bot., 67, 6021–6035

25. Simkin, A. J., McAusland, L., Headland, L. R., Lawson, T. and Raines, C. A. (2015) Multigene manipulation of photosynthetic carbon assimilation increases CO$_2$ fixation and biomass yield in tobacco. J. Exp. Bot., 66, 4075–4090

26. Kromdijk, J., Głowacka, K., Leonelli, L., Gabilly, S. T., Iwai, M., Niyogi, K. K. and Long, S. P. (2016) Improving photosynthesis and crop productivity by accelerating recovery from photoprotection. Science, 354, 857–861

27. Nunes-Nesi, A., Carrari, F., Lytovchenko, A., Smith, A. M., Loureiro, M. E., Ratcliffe, R. G., Sweetlove, L. J. and Fernie, A. R. (2005) Enhanced photosynthetic performance and growth as a consequence of decreasing mitochondrial malate dehydrogenase activity in transgenic tomato plants. Plant Physiol., 137, 611–622

28. Sweetlove, L. J., Lytovchenko, A., Morgan, M., Nunes-Nesi, A., Taylor, N. L., Baxter, C. J., Eickmeier, I. and Fernie, A. R. (2006) Mitochondrial uncoupling protein is required for efficient photosynthesis. Proc. Natl. Acad. Sci. USA, 103, 19587–19592

29. Zhu, X.-G., Wang, Y., Ort, D. R. and Long, S. P. (2013) e-Photosynthesis: a comprehensive dynamic mechanistic model of C$_3$ photosynthesis: from light capture to sucrose synthesis. Plant Cell Environ., 36, 1711–1727

30. Owen, N. A. and Griffiths, H. (2013) A system dynamics model integrating physiology and biochemical regulation predicts extent of crassulacean acid metabolism (CAM) phases. New Phytol., 200, 1116–1131

31. Cortassa, S., Aon, M. A., O'Rourke, B., Jacques, R., Tseng, H. J., Marbán, E. and Winslow, R. L. (2006) A computational model integrating electrophysiology, contraction, and mitochondrial bioenergetics in the ventricular myocyte. Biophys. J., 91, 1564–1589

32. Thornley, J. H. M. and Cannell, M. G. R. (2000) Modelling the components of plant respiration: representation and realism. Ann. Bot. (Lond.), 85, 55–67

33. Lawson, T., Simkin, A. J., Kelly, G. and Granot, D. (2014) Mesophyll photosynthesis and guard cell metabolism impacts on stomatal behaviour. New Phytol., 203, 1064–1081

34. Flexas, J., Ribas-Carbó, M., Diaz-Espejo, A., Galmés, J. and Medrano, H. (2008) Mesophyll conductance to CO$_2$: current knowledge and future prospects. Plant Cell Environ., 31, 602–621

35. Baroli, I., Price, G. D., Badger, M. R. and von Caemmerer, S. (2008) The contribution of photosynthesis to the red light response of stomatal conductance. Plant Physiol., 146, 737–747

36. Wong, S.-C., Cowan, I. R. and Farquhar, G. D. (1979) Stomatal conductance correlates with photosynthetic capacity. Nature, 282, 424–426

37. Farquhar, G. D. and Sharkey, T. D. (1982) Stomatal conductance and photosynthesis. Annu. Rev. Plant Physiol., 33, 317–345

38. Buckley, T. N., Mott, K. A. and Farquhar, G. D. (2003) A hydromechanical and biochemical model of stomatal conductance. Plant Cell Environ., 26, 1767–1785

39. Ball, J. T., Woodrow, I. E. and Berry, J. A. (1987) A Model Predicting Stomatal Conductance and Its Contribution to The Control of Photosynthesis Under Different Environmental Conditions. In Progress in Photosynthesis Research. Biggens, J. ed., Vol, IV, pp. 221–224. Berlin: Springer Netherlands

40. Loreto, F., Harley, P. C., Di Marco, G. and Sharkey, T. D. (1992) Estimation of mesophyll conductance to CO$_2$ flux by three different methods. Plant Physiol., 98, 1437–1443

41. Pons, T. L., Flexas, J., von Caemmerer, S., Evans, J. R., Genty, B., Ribas-Carbo, M. and Brugnoli, E. (2009) Estimating mesophyll conductance to CO$_2$: methodology, potential errors, and recommendations. J. Exp. Bot., 60, 2217–2234

42. Tholen, D., Boom, C. and Zhu, X.-G. (2012) Opinion: prospects for improving photosynthesis by altering leaf anatomy. Plant Sci., 197, 92–101

43. Xiong, D., Liu, X., Liu, L., Douthe, C., Li, Y., Peng, S. and Huang, J. (2015) Rapid responses of mesophyll conductance to changes of CO_2 concentration, temperature and irradiance are affected by N supplements in rice. Plant Cell Environ., 38, 2541–2550

44. Ho, Q. T., Berghuijs, H. N., Watté, R., Verboven, P., Herremans, E., Yin, X., Retta, M. A., Aernouts, B., Saeys, W., Helfen, L., et al. (2016) Three-dimensional microscale modelling of CO_2 transport and light propagation in tomato leaves enlightens photosynthesis. Plant Cell Environ., 39, 50–61

45. Price, N. D., Reed, J. L. and Palsson, B. O. (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat. Rev. Microbiol., 2, 886–897

46. Guo, Y., Ma, Y., Zhan, Z., Li, B., Dingkuhn, M., Luquet, D. and De Reffye, P. (2006) Parameter optimization and field validation of the functional-structural model GREENLAB for maize. Ann. Bot. (Lond.), 97, 217–230

47. Watanabe, T., Hanan, J. S., Room, P. M., Hasegawa, T., Nakagawa, H. and Takahashi, W. (2005) Rice morphogenesis and plant architecture: measurement, specification and the reconstruction of structural development by 3D architectural modelling. Ann. Bot. (Lond.), 95, 1131–1143

48. Song, Y. H., Smith, R. W., To, B. J., Millar, A. J. and Imaizumi, T. (2012) FKF1 conveys timing information for CONSTANS stabilization in photoperiodic flowering. Science, 336, 1045–1049

49. Domagalska, M. A. and Leyser, O. (2011) Signal integration in the control of shoot branching. Nat. Rev. Mol. Cell Biol., 12, 211–221

50. Minchin, P. E. H. and Lacointe, A. (2005) New understanding on phloem physiology and possible consequences for modelling long-distance carbon transport. New Phytol., 166, 771–779

51. Rasse, D. P. and Tocquin, P. (2006) Leaf carbohydrate controls over Arabidopsis growth and response to elevated CO_2: an experimentally based model. New Phytol., 172, 500–513

52. Lynch, J. P. (2013) Steep, cheap and deep: an ideotype to optimize water and N acquisition by maize root systems. Ann. Bot. (Lond.), 112, 347–357

53. Dyson, R. J., Vizcay-Barrena, G., Band, L. R., Fernandes, A. N., French, A. P., Fozard, J. A., Hodgman, T. C., Kenobi, K., Pridmore, T. P., Stout, M., et al. (2014) Mechanical modelling quantifies the functional importance of outer tissue layers during root elongation and bending. New Phytol., 202, 1212–1222

54. Chang, T. G. and Zhu, X. G. (2017) Source-sink interaction: a century old concept under the light of modern molecular systems biology. J. Exp. Bot. erx002

55. Yin, X. and Struik, P. C. (2010) Modelling the crop: from system dynamics to systems biology. J. Exp. Bot., 61, 2171–2183

56. Li, Y., Pearl, S. A. and Jackson, S. A. (2015) Gene networks in plant biology: approaches in reconstruction and analysis. Trends Plant Sci., 20, 664–675

57. Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. and Friedman, N. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat. Genet., 34, 166–176

58. Zheng, G., Xu, Y., Zhang, X., Liu, Z. P., Wang, Z., Chen, L. and Zhu, X. G. (2016) CMIP: a software package capable of reconstructing genome-wide regulatory networks using gene expression data. BMC Bioinformatics, 17, 535

59. Wenden, B. and Rameau, C. (2009) Systems biology for plant breeding: the example of flowering time in pea. C. R. Biol., 332, 998–1006

60. Salazar, J. D., Saithong, T., Brown, P. E., Foreman, J., Locke, J. C., Halliday, K. J., Carré, I. A., Rand, D. A. and Millar, A. J. (2009) Prediction of photoperiodic regulators from quantitative gene circuit models. Cell, 139, 1170–1179

61. Bassel, G. W., Lan, H., Glaab, E., Gibbs, D. J., Gerjets, T., Krasnogor, N., Bonner, A. J., Holdsworth, M. J. and Provart, N. J. (2011) Genome-wide network model capturing seed germination reveals coordinated regulation of plant cellular phase transitions. Proc. Natl. Acad. Sci. USA, 108, 9709–9714

62. Chew, Y. H., Wenden, B., Flis, A., Mengin, V., Taylor, J., Davey, C. L., Tindal, C., Thomas, H., Ougham, H. J., de Reffye, P., et al. (2014) Multiscale digital Arabidopsis predicts individual organ and whole-organism growth. Proc. Natl. Acad. Sci. USA, 111, E4127–E4136

63. Zhu, X.-G., Song, Q. and Ort, D. R. (2012) Elements of a dynamic systems model of canopy photosynthesis. Curr. Opin. Plant Biol., 15, 237–244

64. Parton, W. J., Scurlock, J. M. O., Ojima, D. S., Gilmanov, T. G., Scholes, R. J., Schimel, D. S., Kirchner, T., Menaut, J.-C., Seastedt, T., Garcia Moya, E., et al. (1993) Observations and modelling of biomass and soil organic matter dynamics for the grassland biome wordwide. Global Biogeochem. Cycles, 7, 785–809

65. Parton, W. J., Stewart, J. W. B. and Cole, C. V. (1988) Dynamics of C, N, P and S in grassland soils: a model. Biogeochemistry, 5, 109–131

66. Buckley, T. N. (2005) The control of stomata by water balance. New Phytol., 168, 275–292

67. Lynch, J. P., Nielsen, K. L., Davis, R. D. and Jablokow, A. G. (1997) SimRoot: modeling and visualization of root systems. Plant Soil, 188, 139–151

68. Jones, J. W., Hoogenboom, G., Porter, C. H., Boote, K. J., Batchelor, W. D., Hunt, L. A., Wilkens, P. W., Singh, U., Gijsman, A. J. and Ritchie, J. T. (2003) The DSSAT cropping system model. Eur. J. Agron., 18, 235–265

69. McCown, R. L., Hammer, G. L., Hargreaves, J. N. G., Holzworth, D. P. and Freebairn, D. M. (1996) APSIM: a novel software system for model development, model testing and simulation in agricultural systems research. Agric. Syst., 50, 255–271

70. Humphries, S. W. and Long, S. P. (1995) WIMOVAC: a software package for modelling the dynamics of plant leaf and canopy photosynthesis. Comput. Appl. Biosci., 11, 361–371

71. Song, Q., Chen, D., Long, S. P. and Zhu, X. G. (2017) A user-friendly means to scale from the biochemistry of photosynthesis to whole crop canopies and production in time and space—development of Java WIMOVAC. Plant Cell Environ., 40, 51–55

72. Norman, J. M. (1980) Interfacing leaf and canopy light interception models. In Predicting Photosynthesis for Ecosystem Models. Hesketh, J. D. & Jones, J. W. eds. Vol. 2, pp. 49–67. Boca Raton: CRC Press

73. Farquhar, G. D., von Caemmerer, S. and Berry, J. A. (1980) A biochemical model of photosynthetic CO_2 assimilation in leaves of C_3 species. Planta, 149, 78–90

74. Pokhilko, A., Flis, A., Sulpice, R., Stitt, M. and Ebenhöh, O. (2014) Adjustment of carbon fluxes to light conditions regulates the daily turnover of starch in plants: a computational model. Mol. Biosyst., 10, 613–627

75. de Oliveira Dal'Molin, C. G., Quek, L.-E., Palfreyman, R. W., Brumbley, S. M. and Nielsen, L. K. (2010) C4GEM, a genome-scale metabolic model to study C_4 plant metabolism. Plant Physiol., 154, 1871–1885

76. de Oliveira Dal'Molin, C. G., Quek, L. E., Palfreyman, R. W., Brumbley, S. M. and Nielsen, L. K. (2010) AraGEM, a genome-scale reconstruction of the primary metabolic network in *Arabidopsis*. Plant Physiol., 152, 579–589

77. Warren, J. M., Hanson, P. J., Iversen, C. M., Kumar, J., Walker, A. P. and Wullschleger, S. D. (2015) Root structural and functional dynamics in terrestrial biosphere models — evaluation and recommendations. New Phytol., 205, 59–78

78. Zhu, X.-G., Govindjee, Baker, N. R., deSturler, E., Ort, D. O. and Long, S. P. (2005) Chlorophyll *a* fluorescence induction kinetics in leaves predicted from a model describing each discrete step of excitation energy and electron transfer associated with Photosystem II. Planta, 223, 114–133

79. Yu, X., Zheng, G., Shan, L., Meng, G., Vingron, M., Liu, Q. and Zhu, X. G. (2014) Reconstruction of gene regulatory network related to photosynthesis in *Arabidopsis thaliana*. Front. Plant Sci., 5, 273

80. Chandrasekaran, S. and Price, N. D. (2010) Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis*. Proc. Natl. Acad. Sci. USA, 107, 17845–17850

81. Enquist, B. J. and Niklas, K. J. (2002) Global allocation rules for patterns of biomass partitioning in seed plants. Science, 295, 1517–1520

82. Box, G. E. P. (1976) Science and statistics. J. Am. Stat. Assoc., 71, 791–799

83. Machta, B. B., Chachra, R., Transtrum, M. K. and Sethna, J. P. (2013) Parameter space compression underlies emergent theories and predictive models. Science, 342, 604–607

84. Zhou, M., Wang, W., Karapetyan, S., Mwimba, M., Marqués, J., Buchler, N. E. and Dong, X. (2015) Redox rhythm reinforces the circadian clock to gate immune response. Nature, 523, 472–476

85. Zuo, J. and Li, J. (2014) Molecular dissection of complex agronomic traits of rice: a team effort by Chinese scientists in recent years. Natl. Sci. Rev. 1, 253–276

86. Valluru, R., Reynolds, M. P. and Salse, J. (2014) Genetic and molecular bases of yield-associated traits: a translational biology approach between rice and wheat. Theor. Appl. Genet., 127, 1463–1489

87. Wallace, J. G., Larsson, S. J. and Buckler, E. S. (2014) Entering the second century of maize quantitative genetics. Heredity (Edinb), 112, 30–38

88. Kaul, S., Koo, H. L., Jenkins, J., Rizzo, M., Rooney, T., Tallon, L. J., Feldblyum, T., Nierman, W., Benito, M., Lin, X. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature, 408, 796–815

89. Zhu, X. G., Lynch, J. P., LeBauer, D. S., Millar, A. J., Stitt, M. and Long, S. P. (2016) Plants *in silico*: why, why now and what? — an integrative platform for plant systems biology research. Plant Cell Environ., 39, 1049–1057

90. Marshall-Colon, A., Long, S. P., Allen, D. K., Allen, G., Beard, D. A., Benes, B., von Caemmerer, S., Christensen, A. J., Cox, D. J., Hart, J. C. *et al.* (2017) Crops *in silico*: a prospectus from the plants *in silico* symposium and workshop. Front. Plant Sci. 8, 786

91. Yabusaki, S., Fang, Y., Chen, X., Scheibe, T. D. (2016) Single Plant Root Systems Modeling Under Soil Moisture Variation. In 2016 American Geophysical Union, San Francisco