Reducing a model of sugar metabolism in peach to catch different patterns among genotypes

Hussein Kanso, Bénédicte Quilot-Turion, Mohamed-Mahmoud Memah, Olivier Bernard, Jean-Luc Gouzé, Valentina Baldazzi

Reducing a model of sugar metabolism in peach to catch different patterns among genotypes

Hussein Kanso[a,b], Bénédicte Quilot-Turion[a], Mohamed-Mahmoud Memah[b], Olivier Bernard[d], Jean-Luc Gouzé[d], Valentina Baldazzi[b,c,d,*]

[a]*INRAE, GAFL, F-84143, Montfavet, France*
[b]*INRAE, PSH, F-84914, Avignon, France*
[c]*Université Côte d'Azur , INRAE, CNRS, ISA, Sophia-Antipolis, France*
[d]*Université Côte d'Azur , Inria, INRAE, Sorbonne Université, BIOCORE, Sophia-Antipolis, France*

Abstract

Several studies have been conducted to understand the dynamic of primary metabolisms in fruit by translating them into mathematics models. An ODE kinetic model of sugar metabolism has been developed by Desnoues et al. [1] to simulate the accumulation of different sugars during peach fruit development. Two major drawbacks of this model are (a) the number of parameters to calibrate and (b) its integration time that can be long due to non-linearity and time-dependent input functions. Together, these issues hamper the use of the model for a large panel of genotypes, for which few data are available. In this paper, we present a model reduction scheme that explicitly addresses the specificity of genetic studies in that: i) it yields a reduced model that is adapted to the whole expected genetic diversity ii) it maintains network structure and variable identity, in order to facilitate biological interpretation. The proposed approach is based on the combination and the systematic evaluation of different reduction methods. Thus, we combined multivariate sensitivity analysis, structural simplification and timescale-based approaches to simplify the number and the structure of ordinary differential equations of the model. The original and reduced models were compared based on three criteria, namely the corrected Aikake Information Criterion (AIC_C), the calibration time and the expected error of the reduced model over a progeny of virtual genotypes. The resulting reduced model not only reproduces the

*Corresponding author

predictions of the original one but presents many advantages including a reduced number of parameters to be estimated and shorter calibration time, opening new promising perspectives for genetic studies and virtual breeding. The validity of the reduced model was further evaluated by calibration on 30 additional genotypes of an inter-specific peach progeny for which few data were available.

Keywords: model reduction, sensitivity analysis, structural simplification, quasi-steady-state, peach fruit, kinetic model, model calibration, gene-to-phenotype.

---

1. Introduction

Plants are sessile organisms endowed with the capacity to alter their development, physiology, and morphology depending on the context. Plant phenotype is the result of the interaction between the environment, cultural practices and plant's genetic background (genotype). In the context of agronomy, increasing efforts have been made to select varieties that better meet consumers' expectations. Today it is clear that future breeding should account for complex plant phenotypes, responding to a large panel of criteria, including increased yield, abiotic and biotic stress tolerance, and quality of food products.

Genotype-phenotype models have been considered as the tools of the future to design new genotypes since they can help to test the performance of new genotypes (G) under different Environments (E) x Management (M) conditions. The challenge is to build ecophysiological models that integrate genetic information associated to specific processes (traits). In general, genotypes are defined by a set of parameters, which depends on gene expression or allelic combination, depending on the genetic complexity of the considered trait as well as the available information [2]. Genetic-improved ecophysiological models can then be used to capture GxExM interactions. They can also be used to design "ideotypes" i.e. real or virtual plant cultivars expressing an ideal phenotype adapted to a particular biophysical environment, crop management, and end-use [3, 4]. For this, it is necessary to combine the genetic-improved ecophysiological model with a multi-objective optimization algorithm to identify the best genotypes for specific conditions [5].

Construction of gene-to-phenotype models is challenging. First, the approach requires that a sole and unique model can reproduce the behavior of

all genotypes, in multiple environments, the diversity observed being supported by different sets of parameters. Second, calibration of the models for a large number of genotypes is generally difficult, due to a large number of parameters (typically from 50 to 200 in whole-plant ecophysiological models) along with a restricted number of observations [6, 7]. Due to the model complexity and non-lineairities, evolutionary and bio-inspired algorithms are increasingly used both for parameter estimation and ideotype design. These methods can explore high-dimensional parameter space efficiently but they rely on a large number of model evaluations, that can rapidly increase the computational time required to find a solution. Third, the genetic architecture of complex traits can be very complex, due to epistatic and pleiotropic effects. In this sense, the presence of biologically-meaningful parameters can considerably help the interpretation of the resulting genetic architecture, facilitating the breeding process. Ideally, most the model is close to omics data, the easier the linkage between the parameters and the underlying physiological processes.

Kinetic modeling has been successfully applied to several metabolic pathways in plants [8, 9, 10]. In this spirit, a kinetic model of sugar metabolism has been developed in [1] to simulate the accumulation of different sugars during peach fruit development. The model correctly accounts for annual variability and the genotypic variations observed in ten genotypes derived from a larger progeny of inter-specific peach cross. At term, the objective of the research is to integrate the genetic control of sugar metabolism in this kinetic model and develop a methodology to design ideotypes by virtual breeding. To achieve this, it is necessary to estimate accurately the values of the influential parameters of the model for the whole progeny of 106 genotypes for which few data are available. Unfortunately, the size of the parameter space and the non-linearity of the reaction rates make the calibration of the model unreliable and time-consuming.

One way to face these weaknesses is to reduce the complexity of the model [11]. Several reductions and approximation approaches exist in the literature, each one addressing a specific aspect of model complexity [12, 13]. A number of methods, such the lumping method [14, 15] or the classical quasi-steady-state (QSS) approaches, aim at reducing the number of variables based on chemical or time-scale considerations [16, 17]. Methods from sensitivity analysis may help to reduce the parameter space by identifying non-influential parameters, whose values can be fixed by broad literature data [18, 19, 20, 21]. Last but not least, the structure of the model itself can be simplified. Meth-

ods for model decomposition [22, 23, 24] aim to separate the system into sub-networks or sub-models, that are easier to analyze and parameterize. The choice of reaction kinetics is also very important for model complexity. In this perspective, the use of simplified enzyme kinetics [25, 26, 27] may be useful to avoid the emergence of numerical and identifiability issues.

Different reduction methods can be combined together. In [28] for instance, model decomposition is associated to variable transformation, resulting in a low-dimensional description of the "exterior" part of the system, whereas in [15] time scale analysis is used to identify a cluster of fast variables to be lumped together.

In the work of Apri et al. [29] different reduction steps (parameter removal, node removal, variable lumping) are sequentially tested following a practical scheme: at each step, if the reduced model, after parameter re-estimation, can reproduce some target outputs, the modification is selected, and rejected otherwise. From the point of view of genetic applications, a major drawback of the approach of Apri et al. [29] is that the selection of acceptable reduction results depends on the specific target dynamics.

As a consequence, different target outputs (i.e. genotypes) can give rise to reduced models with different structures or parameters number, making their comparison difficult in the perspective of genetic studies.

The objective of this work was to provide a method to build a reduced model that is adapted to the specificity of genetic studies in that: i) it yields a reduced model that is adapted to the whole expected genetic diversity ii) it maintains network structure and variable identity, in order to facilitate the biological interpretation of the reduced model.

Similarly to the approach of Apri et al. [29], our reduction strategy tests different methods in several *parallel* steps that, if retained, are combined together into a final reduced model (Fig. 1).

First, multivariate sensitivity analysis was attempted to reduce the parameter space [30]. Second, we tried to simplify the structure of the model by reducing non-linearity and time-dependent forcing, and finally, a quasi-steady-state approximation based on time-scale separation was tested to reduce the size of the system. Particular attention was devoted to the systematic evaluation of the different reduction methods. Three main criteria were used to assess the interest of the reduction: i) the corrected AIC value, evaluating the relative gain between model simplification and loss of accuracy over an experimental dataset, ii) the calibration time, as a measure of model efficiency, iii) the expected error between the original and the reduced model

over a population of virtual genotypes, as a measure of the reliability of the simplification scheme.

As a case study, the proposed reduction scheme was applied to the model of sugar metabolism proposed by Desnoues et al. [1]. The resulting reduced model correctly reproduces data on the original 10 genotypes with only 9 estimated parameters (out of 14 in the original model) and a gain in calibration time over 40%. In addition, the reduced model was successfully calibrated on 30 new genotypes of the same inter-specific peach progeny, for which fewer data points were available.

The paper is organized as follows. In the next section, we briefly present the original model of sugar metabolism developed by Desnoues et al. [1]. Section 3 is devoted to the description of the individual reduction methods, whereas Sections 4 and 5 present, respectively, the datasets and the numerical methods used for the assessment of the proposed model reduction. The results of the application of our reduction scheme to the model of sugar metabolism are reported in section 6. A general discussion on the advantages and limitations of our approach closes the paper.

2. Description of the peach sugar model

The model developed by Desnoues et al. [1] describes the accumulation of four different sugars (sucrose, glucose, fructose, and sorbitol) in peach fruit during its development over a progeny of ten peach genotypes with contrasting sugar composition. The fruit was assumed to behave as a single big cell with two intra-cellular compartments, namely the cytosol and the vacuole. Carbon enters the fruit from the plant sap which is transformed by a metabolic network, including enzymatic reactions and transport mechanisms between the cytosol and the vacuole.

The developed dynamical model made explicit use of experimental data to describe the evolution of the sub-cellular compartment (due to fruit growth) and enzyme activities (due to fruit developmental program) over time. To this aim, measured fruit dry and fresh masses and enzyme activities were represented by genotype-specific temporal functions and provided as input to the model.

From a mathematical point of view, the model can be described as a set

of parametric ordinary differential equations:

$$\frac{dx}{dt} = f(x(t), I(t), v(t), p), \tag{1}$$

$$x(t_0) = x_0, \tag{2}$$

where t is the independent time variable in days after bloom (DAB); $x \in \mathbb{R}^{10}$ is the concentration vector of metabolites in the corresponding intra-cellular compartment and $x_0 \in \mathbb{R}^{10}$ in Eq.(2) is the vector of the corresponding initial values. $I \in \mathbb{R}$ is the time-dependent input of carbon from the plant and $v \in \mathbb{R}^7$ is the vector of time-dependent measured enzymatic activities; $p = (p_1, \ldots, p_{23})$ is the vector of parameters defining the rate reactions where $p_1, \ldots, p_{14}$ have to be estimated and $p_{15}, \ldots, p_{23}$ are fixed from literature data. $f(x(t), I(t), v(t), p)$ of Eq.(1) describes the change in compounds concentrations. Equations of the reduced and original model are introduced in Appendix Appendix A.1.

3. Model reduction methods

In this section, we present a reduction scheme explicitly dedicated to genetic studies that combines different methods in several parallel steps as shown in (Fig. 1) and explained in the next subsections.

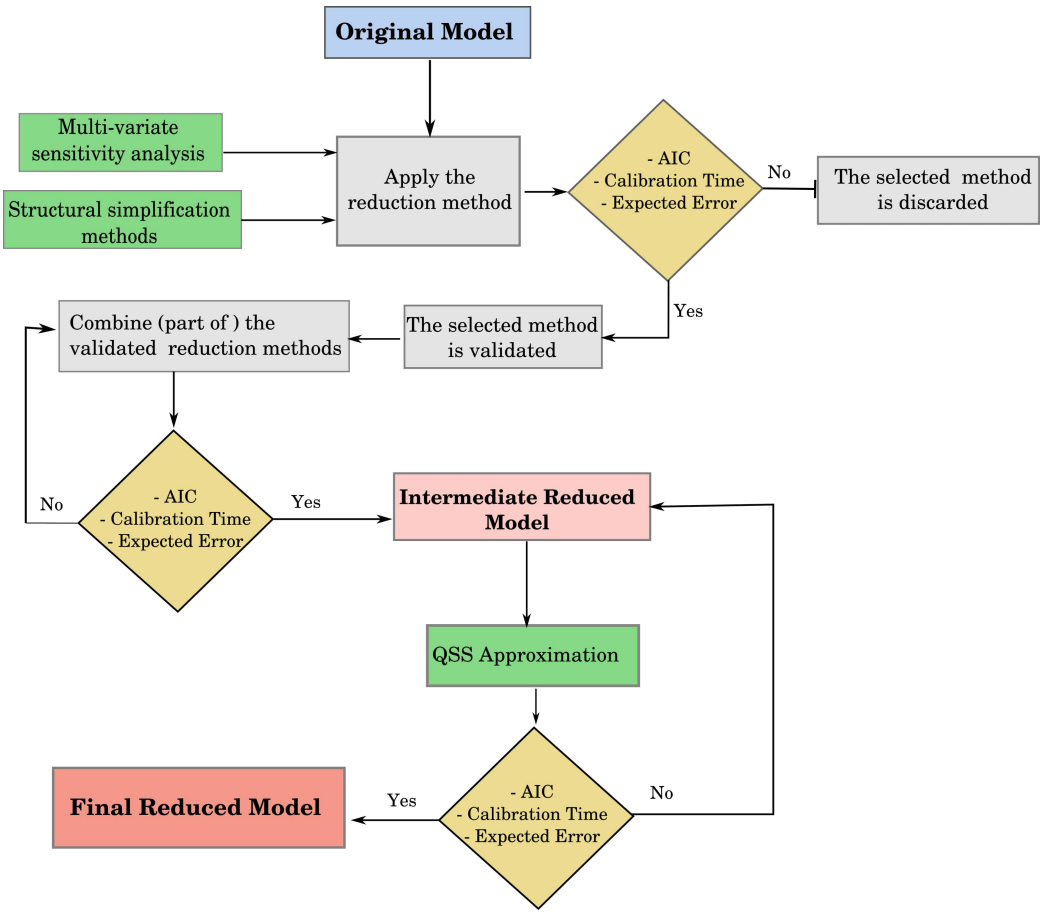

Figure 1: Graphical representation of the proposed model reduction scheme. Yellow diamonds represent model evaluation steps by means of our 3 criteria: the corrected AIC value, calibration time and expected error over a virtual population. The tested reduction methods are indicated in green. Multivariate sensitivity analysis and three structural simplification methods are independently applied to the original model and evaluated. The validated methods are then combined into an intermediate reduced model whose performances are again submitted to evaluation. Finally, the application of a QSS approximation over the intermediate reduced model is tested to yield the final reduced model.

3.1. Multivariate sensitivity analysis

Generally, in the case of complex models, estimating parameters requires a lot of effort and is known to be a difficult and challenging task. In particular, it is tricky to determine which parameters can be fixed. The global sensitivity analysis methods allow to explore the influence of each parameter on model outputs and thus to identify the key parameters that affect model

performance and play important roles in model parameterization, calibration and optimization [21]. Multivariate sensitivity is a method developed by Lamboni et al. [30] that allows the application of global sensitivity analysis to models having a multivariate (eg. dynamic) output. The idea is to perform a principal components analysis on the outputs, and then compute the sensitivity indexes for each principal component. The results are summarized by the generalized sensitivity indices (GSI) that provide a unique ranking of the parameters over the whole output.

This method was applied to the 23 parameters of the original model and to the measured enzymatic activities v. Each parameter was studied at three levels, corresponding to 0.05, 0.5 and 0.95 quantiles of the previously estimated 14 parameters values [1] and to a variation of -20% and $+20\%$ of the fixed values for the other parameters. For time-dependent enzyme activities, the same -20% and $+20\%$ variation was applied on their average values over the whole dynamics.

In order to evaluate the impact of the genotype choice on the results of the sensitivity analysis, simulations were performed according to a factorial design, following the ANOVA model $genotypes \times (p_1 + \ldots p_{23} + v_1 + \ldots + v_7)^2$. The package "Planor" in **R** (R Development Core Team 2015) was used. The minimum resolution of the plan was fixed by using the tool MinT [31] to test all main effects and interactions. The factorial design resulted in $10 \times 3^9 = 196\ 830$ simulations.

Multivariate sensitivity analysis was performed independently on the dynamics of the four output sugars (*i.e* sucrose, glucose, fructose, and sorbitol) that compose peach fruit. In order to determine the least sensitive parameters, the whole sugar phenotype has to be taken into account, with respect to the relative proportions of each sugar. For this aim, an aggregate generalized sensitivity index ($aGSI$) was constructed for each parameter as

$$aGSI = \sum_{i=1}^{4} GSI_i \ \beta_i \tag{3}$$

where GSI is the generalized sensitivity indice computed for the sugar i and β_i the relative proportion of sugar i in the fruit. $\beta = (0.72, 0.13, 0.09, 0.05)$ for sucrose, glucose, fructose, and sorbitol, respectively.

3.2. Structural simplification methods

This section aims to simplify the structure of the model in terms of network and reaction rates while preserving its predictive ability. The structural

simplification includes the three following strategies:

3.2.1. Simplifying the description of enzymatic capacities

Seven enzymatic capacities V_{max} are represented in the original model. Some of these capacities were assumed to vary over time (temporal effect) and/or to depend on the phenotypic group (phenotype effect), according to experimental evidences [32]. The characteristics of enzyme capacities are summarized in Table 1. In order to simplify the model, we systematically tested the impact of the suppression of the phenotype and/or the temporal effect on each single capacity. Depending on the characteristics of the considered enzyme (Table 1), the procedure is slightly different:

$$\text{Phenotype effect :} \quad \begin{cases} V^1_{max} \\ V^2_{max} \end{cases} \rightarrow \frac{V^1_{max} + V^2_{max}}{2} \tag{4}$$

$$\text{Temporal effect :} \quad V_{max}(t) \rightarrow < V_{max}(t) >_t \tag{5}$$

$$\text{Double effect :} \quad (4) \text{ then } (5) \text{ applied} \tag{6}$$

where $< . >_t$ stands for temporal average over the whole dynamics.

Table 1: Characteristics of enzymatic activities in [1]

V_{max}	Phenotype effect	Temporal effect
v_1	No	No
v_2	No	Yes
v_3	Yes	No
v_4	No	Yes
v_5	No	Yes
v_6	Yes	Yes
v_7	Yes	Yes

3.2.2. Rate simplification

In the original model, enzymatic reactions were represented by an irreversible Michaelis-Menten (MM) equation:

$$u(x,t) = V_{max} \frac{x(t)}{K_m + x(t)} \tag{7}$$

where V_{max} is the enzymatic capacity. K_m is the affinity of the enzyme for the substrate, $x(t)$ is the concentration of the substrate at time t.

The objective here is to simplify Eq.(7) in order to improve the efficiency of the numerical simulation. Depending on the relative levels of the substrate concentration and the MM equation affinity, two simplifications of the flows' equations can be made:

Case 1: if $x(t) << K_m$
Substrate concentration is small compared to the affinity of the enzyme for the substrate then we can write: $u(x,t) = \frac{V_{max}}{K_m} x(t)$.

Case 2: if $x(t) >> K_m$
Substrate concentration exceeds the affinity of the enzyme for the substrate, so that the enzyme can be supposed close to saturation: $u(x,t) = V_{max}$.

3.2.3. Futile cycle removal

The presence of internal cycles within a metabolic network can lead to the appearance of thermodynamically unfeasible loops i.e. reactions that run simultaneously in opposite directions (for example Fig. 2) and have no overall effect on the exchange fluxes of the system. This is an undesirable situation that causes numerical issues and makes the estimation of the corresponding parameter values an ill-posed problem.

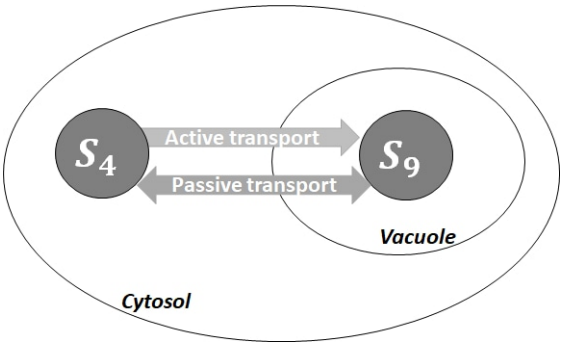

Figure 2: S_4 is the glucose in the cytosol transported to the vacuole as S_9 via an active (unidirectional transport) and passive (reversible transport).

In this context, our strategy was to remove each futile cycle by replacing the antagonist reactions by a single effective reaction preserving the net exchange flux of the system. Different kinetics can be tested for the effective

reaction, as alternative reduction approaches. Consistently with the previous reduction method, we decided to test two linear reaction forms, namely

$$u(x,t) = k_i x_i - k_j x_j \tag{8}$$

and

$$u(x,t) = k_i (x_i - x_j) \tag{9}$$

where x_i, x_j are the variables involved in the futile cycle and k_i, k_j are the coefficients to be estimated.

3.3. Time-scale analysis and QSS approximation

Biological systems are often characterized by the presence of different time scales (seconds, hours, days). Following Heinrich and Schuster [17], an appropriate measure of the time scales involved is given by

$$\tau_i(t) = -\frac{1}{Re(\lambda_i(t))} \tag{10}$$

where $Re(\lambda_i)$ are real parts of the eigenvalues λ_i of the Jacobian matrix of the system, along a given trajectory. The presence of fast modes in the system allows the reduction of the number of variables based on a quasi-steady-state assumption.

Based on the above information and on the analysis of time-series of the full model, variables can be divided into two groups $x = (x^{(1)}, x^{(2)})$, where $x^{(1)}$ and $x^{(2)}$ correspond respectively to the slow and fast variables of the system [17, 33].

Application of the QSS approximation states that

$$\frac{dx^{(2)}}{dt} = f_2(x^{(1)}, x^{(2)}, I(t), v(t), p) = 0 \quad \rightarrow \quad x^{(2)}_{ss} = g(x^{(1)}) \tag{11}$$

It follows that, after a relaxation period, the system can be approximated by the reduced model:

$$\frac{dx^{(1)}}{dt} = f_1((x^{(1)}, g(x^{(1)}), I(t), v(t), p) \tag{12}$$

of lower dimension.

4. Experimental and artificial data

4.1. Experimental data

The 106 peach genotypes used in this study come from an inter-specific progeny obtained by two subsequent back-crosses between *Prunus davidiana* (Carr.) P1908 and *Prunus persica* (L.) Batsch 'Summergrand' and then 'Zephyr' [34]. They were planted in 2001 in a completely randomized design in the orchard of the INRAE Research Centre of Avignon (southern France). Experimental monitoring of peach fruit growth and quality has been conducted in 2012, as described in [32]. The concentration of different metabolites, namely sucrose, glucose, fructose, sorbitol, and hexoses phosphates, the fruit flesh fresh weight and dry matter content were measured at different time points during fruit development, for all genotypes. In addition, the temporal evolution of enzymatic capacities (maximal activity) of the twelve enzymes involved in sugar metabolism was measured over the whole population [32]. The resulting dynamic patterns were analyzed and compared by means of a generalized mixed linear-effect model (GLMM). Accordingly, some enzyme activities were shown to vary over time and/or depend on the phenotypic group [32].

Training set

The 10 genotypes already used by Desnoues et al. [1] were selected as the training set for our reduction strategies. They include five genotypes having a 'standard phenotype', namely a balanced fructose-to-glucose ratio at maturity between 0.6 and 0.9, and five considered to have a 'low fructose phenotype' due to the lower proportion of fructose compared with glucose based on their sugar composition at maturity [1]. For these 10 genotypes, 3 biological measurements are available at 6 dates after bloom.

The training set was used to test each reduction method individually as well as their combination, based on the AIC_C value and the calibration time (see section 5.3).

Validation set

The quality of the final reduced model was evaluated by calibration on a validation set for which fewer data points were available (one single biological measurement at 6 dates). The idea was to select 30 additional genotypes of the inter-specific peach progeny, which in complement to the training set, represented the greatest diversity in terms of growth rate and duration. For

this aim, experimentally measured growth curves were interpolated with a smoothing spline algorithm [35] with 16.4 degrees of freedom in **R** (R Development Core Team 2015) and the maximum and average growth rate quantified as the maximum and the average of the growth curve's derivative over fruit development. A principal component analysis (PCA) was performed on growth rate and growth duration for the whole progeny of 106 genotypes using the **R ADE4** library. The first two principal components accounted for more than 90% of the genetic diversity. The first axis was mainly related to the growth rate whereas the second one reflected the duration of growth. As shown in Fig. 3, the ten genotypes of the original study provided a good representation of the observed diversity in growth rate. However, their growth duration was relatively short, compared to the existing variability. As a consequence, most of the new genotypes have been selected in the upper-left panel of the plan, in order to capture the greatest genetic diversity in terms of fruit development. An equal proportion of the two phenotypic groups was maintained.

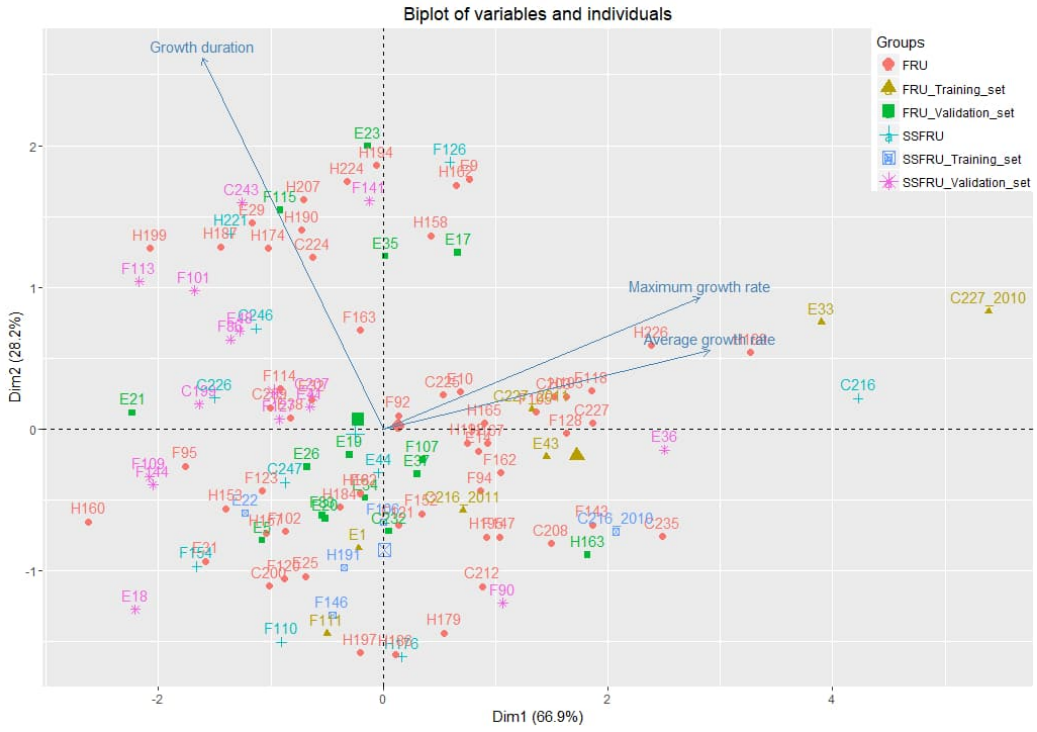

Figure 3: Principal component analysis (PCA) for the whole progeny of 106 genotypes. It represents the projection on the Dim1 and Dim2 of the growth duration and growth rate obtained with growth curves.

4.2. Virtual genotypes

In addition to the training set, a virtual experiment was performed to evaluate the reliability of the reduction methods to variations in parameter values, initial conditions, and input functions, expected in large genetic populations. For this aim, 20 000 virtual genotypes were generated by randomly assigning model parameters and inputs, based on data from the 10 profiles used in [1].

The values of the parameters p were taken randomly using a uniform distribution between the minimum and the maximum of the previously estimated values over the set of 10 genotypes [1]. Initial conditions, such as *initial fruit weight*, and *initial sugar concentration* were assigned randomly using a uniform distribution within the range of observed values plus a variation of 40%.

Given the high correlation among parameters describing fruit growth curves

[36], model inputs, such as *fruit weight*, were randomly assigned using a uniform distribution picking one of the observed growth dynamics and adding an overall random variation between zero and 10% on fruit weight. Finally, shifts in the duration of fruit development among genotypes were also considered. The maturity date was chosen randomly using a uniform distribution within the range of observed dates broaden of 40%.

5. Numerical methods

5.1. Mathematical notations

- $x(t, p^{(k)})$: original model associated to parameters $p^{(k)}$ (i.e. genotype k)
- $\tilde{x}(t, \tilde{p}^{(k)})$: reduced model for the genotype k.

Note that the notation $\tilde{x}(t, \tilde{p}^{(k)})$ can apply to different versions of the reduced model, depending on the considered reduction method.

- $\mathcal{T}_S^{(k)}$: set of the N_S simulation times for the genotype k
- $\mathcal{T}_M^{(k)}$: set of the N_M measurement times for the genotype k
- $X^{(k)}(t_j)$: N experimental observations for the genotype k, with $t_j \in \mathcal{T}_M^{(k)}$. Note that $N = 4 \times N_M \times r$, where r is the number of replicates at time t_j, for the 4 different sugars (sucrose, glucose, fructose and sorbitol). $r = 3$ for the training set and $r = 1$ for the validation set.

5.2. Parameter estimation

In this section, we aim to estimate the parameters of the models to fit our observations i.e. our measured sugars concentrations. For this purpose, we note $X^{(k)} = (X_1^{(k)}, \ldots, X_N^{(k)})$ the vector of the experimental observations at several times for the genotype k and suppose that:

$$\mathbb{E}(X_i^{(k)}) = \mathcal{M}_{p^{(k)}}(x_i^{(k)})$$

where $x_i^{(k)} = (x^{(k)}(t_i))$ is the set of system variables at $(t_i)_{i\in[1,N]}$, $p^{(k)}$ is the vector of parameters to be estimated and $\mathcal{M}_{p^{(k)}}$ is the mathematical function relying the considered model to the data (see Appendix A for more

information). Here, the observations $X^{(k)}$ are assumed to follow a Gaussian law $\mathcal{N}(\mathcal{M}_{p^{(k)}}(x^{(k)}), \sigma_k^2)$ with constant variance σ_k^2.

The estimation of our parameters can be performed through the maximization of the likelihood. We note $\ell(p^{(k)}, \sigma_k^2)$ the log-likelihood function for the genotype k.

Under the assumption of observation independence, the log-likelihood can be defined as follows:

$$\ell(p^{(k)}, \sigma_k^2) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma_k^2) - \frac{1}{2\sigma_k^2}\sum_{i=1}^{N}(X_i^{(k)} - \mathcal{M}_{p^{(k)}}(x_i^{(k)}))^2 \quad (13)$$

A maximum log-likelihood estimator $(\hat{p}^{(k)}, \hat{\sigma_k}^2)$ of $(p^{(k)}, \sigma_k^2)$ is a solution to the maximization problem:

$$(\hat{p}^{(k)}, \hat{\sigma_k}^2) = \underset{p^{(k)}, \sigma_k^2}{\arg\max}\ \ell(p^{(k)}, \sigma_k^2) \quad (14)$$

In this Gaussian case, the maximum log-likelihood estimator is thus equivalent to the ordinary least-square estimator:

$$\hat{p}^{(k)} = \underset{p^{(k)}}{\arg\max}\sum_{i=1}^{N}(X_i^{(k)} - \mathcal{M}_{p^{(k)}}(x_i^{(k)}))^2 \quad (15)$$

$$\hat{\sigma}_k^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i^{(k)} - \mathcal{M}_{\hat{p}^{(k)}}(x_i^{(k)}))^2 \quad (16)$$

Matlab software (MATLAB R2018a, The MathWorks Inc., Natick, MA) was used for model integration (solver ode23tb [37]) and calibration. A genetic algorithm (function `ga` [38] of Global Optimisation Toolbox) was used for maximization of Eq. (15). The population size, the maximum number of generations, and the crossover probability have been respectively set at 200, 300, and 0.7. For each reduced version of the model (individual or combined reduction methods), free parameters were numerically re-estimated. The fitting process was considered at convergence when the average relative change in the best-cost function, i.e. the sum of squared errors, value over generations was less than 10^{-6}. For each genotype k and reduced model, estimations procedure has been repeated ten times to take into account the stochastic nature of the genetic algorithm and to ensure the good exploration of the parameters' space. The solution having the best score was kept for subsequent analyses.

5.3. Model selection

Individual and combined reduction methods were evaluated according to three criteria of major importance for our application: the corrected Akaike Information Criterion (AIC_C), the gain in calibration time (%) and the expected error (%) between the original and reduced models.

Akaike Information Criterion

The AIC gives information on the likelihood of the proposed model based on available experimental data and weighted by the number of free parameters: [39]:

$$AIC(p) = -2\,\ell(p, \sigma^2) + 2n_p \tag{17}$$

where n_p is the number of estimated parameters p and $\ell(p, \sigma^2)$ is the maximum log-likelihood. In this paper, we used the corrected AIC as we deal with a small set of observations and a considerable number of parameters.

$$AIC_C(p) = AIC(p) + \frac{2n_p(n_p+1)}{N - n_p - 1} \tag{18}$$

where N is the number of observations. For genotype k and for each reduction method, we defined

$$\Delta^{(k)}_{AIC_C}(\tilde{p}^{(k)}, p^{(k)}) = AIC_{Creduced}(\tilde{p}^{(k)}) - AIC_{Coriginal}(p^{(k)}) \tag{19}$$

as the AIC_C difference between the reduced and the original model. Note that Δ_{AIC_C} is always computed using the best estimated parameter solution for the considered model. Whenever the average over the 10 genotypes ($< \Delta_{AIC_C} >_G$) was negative, the reduction method was validated.

Gain in calibration time

We used the calibration of a specific genotype ($E43$) as a proxy of the maximum expected calibration time on the population. Genotype $E43$ was selected because it required a long calibration time on the original model proposed by Desnoues et al. [1] (approximately 11 hours on average on a 3.1GHz Intel(R) Xeon(R) processor) but it did not suffer from numerical instabilities, that could complicate the calibration process. Note that the overall calibration time of a model depends both on the integration time of each evaluation step and on the convergence of the cost function that sets the actual number of generations performed by the algorithm. Both aspects may be affected by the model reduction.

To evaluate the gain in calibration time due to model reduction, parameter estimation was performed for each reduction method, following the general procedure (see section 5.2), and compared to the calibration time obtained for the original model. An initial population $\mathcal{P}_0$ was randomly selected assuming a uniform distribution in the parameter range and then kept fixed for all calibration processes (both original and reduced models). For models having a reduced number of parameters, the initial population was directly derived from $\mathcal{P}_0$.

The gain (G_t) was defined as the gain (in %) in calibration time T between the original and the reduced model:

$$G_T = \frac{T_{original} - T_{reduced}}{T_{original}} \times 100$$

Expected error

Simulations of the original and reduced models were compared by the Normalized Root Mean Square Error over the 10 model variables :

$$J_i(p^{(k)}, \tilde{p}^{(k)}) = \frac{\sqrt{\frac{1}{N_S}\sum_{j=1}^{N_S}(x_i(t_j, p^{(k)}) - \tilde{x}_i(t_j, \tilde{p}^{(k)}))^2}}{max_j(x_i(t_j, p^{(k)})) - min_j(x_i(t_j, p^{(k)}))} \quad \forall i \in \{1, \ldots, 10\} \quad (20)$$

where $x(t, p_k)$ and $\tilde{x}(t, \tilde{p}_k)$ are the concentration predicted by the original and reduced model, respectively. Parameters for the reduced model were derived from the values of the corresponding parameters in the original model.

The quality of the QSS approximation was assessed by computing J_i for each variable in the model, over the whole dynamics.

In the context of the virtual experiment, the Expected Error (%) of the reduced model was defined as the average distance J over the virtual population:

$$\text{Expected Error} = \frac{1}{N_{VG}} \sum_{k=1}^{N_{VG}} < J_i(p^{(k)}, \tilde{p}^{(k)}) > \times 100 \quad (21)$$

with

$$< J_i(p^{(k)}, \tilde{p}^{(k)}) >= \frac{1}{10} \sum_{i=1}^{10} J_i(p^{(k)}, \tilde{p}^{(k)})$$

where N_{VG} is the number of virtual genotypes and 10 is the number of variables. In our case, $N_{VG} = 20\ 000$. The Expected Error was used to quantify the reliability of the reduction.

6. Results

6.1. Strategy 1: Identification of low sensitive parameters

The objective of the sensitivity analysis was to identify parameters having a significant influence on the outputs of the model, over the whole dynamics and for all tested genotypes. A multivariate sensitivity analysis [30] was used for this purpose. The aggregate generalized sensitivity indices (aGSI) (see section 3.1) shown in Fig. 4 give a common ranking of model parameters according to their influence on the whole sugar phenotype, as it is made up by the four output sugars (sucrose, sorbitol, glucose, and fructose).

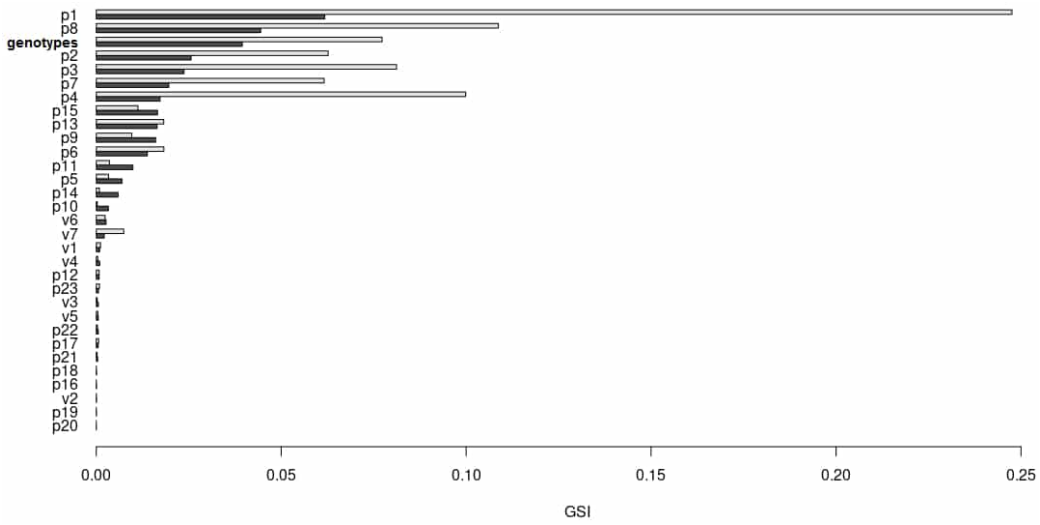

Figure 4: Aggregate Generalized sensitivity indices (aGSI) for the parameters of the model and genotypes (the training set) on four outputs (Sucrose, Sorbitol, Glucose and Fructose) of the sugar model. The main sensitivity indices are in dark bars and interaction ones are in grey bars.

Parameter (p_1) related to the action of cell-wall invertase in fruit apoplasm and the coefficient of sucrose import (p_8) are the most important parameters, followed by the activities of acid invertase (p_2), the activities of Fructokinase (p_3), Hexokinase (p_4) and the resynthesis rate of sucrose from hexose phosphate (p_7). Indeed, p_1, p_3, and p_4 parameters are the most sensitive parameters for sucrose, fructose and glucose concentrations respectively (see Fig. B.2).

Interestingly, the `genotype` factor is ranked third, meaning that it does not affect parameters' sensitivity as much as expected. A closer look at the

results shows that the choice of the genotype essentially affects the second principal component, via the definition of the initial conditions of the model (see the supplemental information Fig. B.1).

Among the 14 parameters estimated ($p_1, \ldots, p_{14}$) in the original model, four parameters, namely p_5, p_{10}, p_{12} and p_{14}, have a negligible effect on the four outputs, independently of the peach genotype. Accordingly, these parameters can be fixed to their nominal values i.e. their average value over the ten genotypes, without affecting the quality of predictions. The validity of such a reduction strategy was tested on the ten genotypes of the training set. The difference in Akaike criterion (Δ_{AIC_C}) between the reduced and the original models was computed for each genotype. Results presented in Table 2 show that such a reduction in the number of parameters is strongly beneficial for nine out of the ten genotypes with largely negative Δ_{AIC_C} values, and roughly neutral for one genotype ($\Delta_{AIC_C} \sim 0$). The gain in calibration time, however, is important (25%) and the expected error over the progeny of virtual genotypes is low, demonstrating a good reliability of the proposed simplification. For these reasons, the model with 10 parameters to be estimated was selected.

Table 2: Δ_{AIC_C} calculated between reduced and original models for the training set and the gain in calibration time (%) for E43. The Expected error ± standard deviation (Std) between original and different reduced models for 20 000 virtual genotypes.

Simplification method		Δ_{AIC_C}											Calibration	Expected Error
		E1	**E33**	**E43**	**F111**	**E22**	**F106**	**F146**	**H191**	**C216**	**C227**	$< \Delta_{AIC_C} >_G$	Time gain %	Virtual genotypes
Low sensitive parameters fixed		-11.5	-6.4	-0.9	-14.04	-13.2	-28.3	-13.5	-14.3	-18.7	-87.7	-20.8	25.8	4.9 ± 6.5
V_{max} Type effect removed	v_3	-1.01	-5.9	-4.15	-4.2	1.1	-2.3	-0.3	-6.1	-6.1	-72.02	-7.9	22.4	0.5 ± 1.3
	v_6	-0.1	-4.3	0.06	- 3.9	0.7	-5.5	-0.3	-5.4	-6.1	-87.7	-11.3	26.6	1.7 ± 1.6
	v_7	-0.7	-36.4	0.2	-6.02	1.4	-5.6	1.9	-5.2	-4.9	-94.5	-14.9	33.9	2.9 ± 4.5
V_{max} Temporal effect removed	v_2	-0.1	-3.1	0.06	-3.7	0.7	-5.1	-0.3	-0.3	-6.3	-83.4	-10.1	31.6	0.3 ± 0.7
	v_4	-0.8	-8.2	0.7	-5.6	-5.03	-2.5	-2.5	-5.1	- 6.1	-90.3	-12.3	19.4	2.9 ± 2.5
	v_5	0.2	-6.8	0.5	-4.8	1.8	-5.8	2.03	-2.9	-6.1	-91.1	-11.3	20.3	5.5 ± 5.7
	v_6	-0.3	-0.4	-0.1	-27.04	1.7	-5.5	-0.2	-5.3	-5.7	-84.9	-12.7	30.5	4.1 ± 3.1
	v_7	8.6	-25.1	21.1	11.02	19.6	20.01	29.05	12.4	15.6	-97.5	1.5	24.2	6.8 ± 4.5
Rate simplification		-17.2	-53.4	8.9	-35.4	-2.9	2.7	-14.7	-22.9	-5.7	- 71.04	-21.1	6.7	18.6 ± 9.7
Futile cycle removal	Eq. (8)	2.5	-0.9	15.6	-1.6	-0.01	-2.3	-0.6	-1.5	-5.9	-43.23	-3.8	23.6	12.7 ± 14.7
	Eq. (9)	0.7	- 56.7	-5.6	-37.1	- 9.02	- 10.5	-6.7	-35.5	-12.2	- 70.7	-24.3	24.1	11.5 ± 9.9
Intermediate reduced model		-32.7	-18.6	-3.7	- 24.5	-11.8	-24.04	-16.5	-20.3	-18.8	-43.1	-21.4	30.5	22.5 ± 8.4
Final reduced model		-32.5	-19.1	-4.3	-25.1	-12.7	-1.01-	- 16.4	-20.4	-18.8	-43.3	-18.5	43.3	22.5 ± 8.5

6.2. Strategy 2: Structural simplification of the model

Structural simplification methods are another way to reduce the complexity of dynamic systems by improving the generality of the model and the numerical integration of the ordinary differential equations.

Firstly, we tried to remove the temporal and the phenotype effects in the enzyme activities, $v_2, \ldots, v_7$ (v_1 has neither phenotype nor temporal effects). The results of this simplification are shown in Table 2. The elimination of the phenotype effect for v_3, v_6 and v_7 resulted in a decrease of the AIC_C value for nine genotypes, neutral for one genotype, and was thus selected for the final reduction. The elimination of the temporal effect for v_2, v_4, v_5, v_7 was also advantageous on the corrected AIC results for all ten genotypes. Nevertheless, when we tried to eliminate the temporal effect of v_7, the resulting Δ_{AIC_C} was positive for most genotypes. This is in line with the results of multi-variate sensitivity analysis according to which $v_2, \ldots, v_6$ have a low sensitivity on the four outputs of the model, whereas v_7 has a non-negligible effect on the dynamics of glucose concentration. According to these results, the elimination of the temporal effect was validated only for v_2, v_4, v_5, v_6. In support of this choice, the test with the virtual genotypes shows that the expected error between the reduced and the original model is small (Table 2).

In the second phase, we tested the possibility of simplifying the enzymatic reaction rates (Eq.(7)). For each reaction in the model, Fig. 5 compares the order of magnitude of the substrate $x(t)$ to the corresponding affinity K_m. The boxplots show that (**Case 2**, see section 3.2.2) simplification strategy can be applied only for the reaction rates u_5 and u_7. Therefore, their reaction rates can be written as $u = V_{max}$. All other flows verify the (**Case 1**, see section 3.2.2) and can therefore be expressed as $u = \frac{V_{max}}{K_m} x(t)$. The rates simplification improves the corrected AIC for eight genotypes and yields a substantial gain in the calibration time. The expected error over the virtual progeny is higher than in the previous reduction steps, but still in the range of accuracy of the original model [1]. According to these observations, the enzymatic reaction rates simplification strategy was validated.

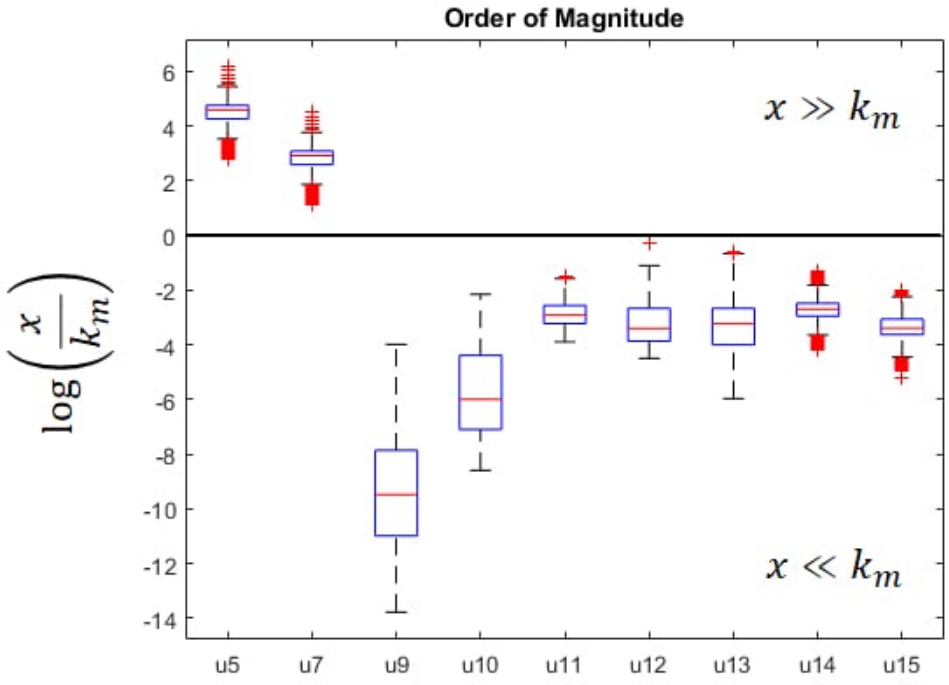

Figure 5: Differences in order of magnitude between enzyme affinity (K_m) and substrate concentration (x) calculated over the whole dynamics and the training set for each reaction rate u_i, $i \in \{5, 7, 9 \dots 15\}$.

Eventually, futile cycles were detected to reduce the full system. In the original model, glucose, and fructose sugars can be transported to the vacuole via two possible mechanisms: an active, unidirectional transport (u_5, u_7) and passive reversible transport (u_6, u_8). Simulations showed that, whenever the genotype, the net flux mostly pointed in the direction of an export for both fructose and glucose from the vacuole to the cytosol [1]. However, futile cycles occurred due to the presence of the active transport mechanism, that continually brings glucose and fructose back into the vacuole. Indeed, u_5 and u_6 (respectively u_7 and u_8) had the same evolution over the whole dynamics for all ten genotypes (Fig. 6): the active and passive transport ran simultaneously in two opposite directions.

According to our strategy (section 3.2.3), we tried to remove futile cycles by replacing reactions (u_5, u_6) (respectively (u_7, u_8)) with an effective reaction rate of the form $p_{10}\, x_9 - p_{11}\, x_4$ (respectively $p_9\, x_8 - p_{12}\, x_3$) preserving the

net export flux from vacuole to the cytosol. We compared the performance of the reduced model with respect to the original one (Table 2). The corrected AIC values were generally slightly negative, with the exception of genotypes E_1 and E_{43}, suggesting an overall improvement of the model structure. Notice that the present strategy did not reduce the total parameters number but decreased model complexity and improved the calibration time.

As a further simplification, we then tried to use a special case of the above mentioned reaction rate with $p_{10} = p_{11}$ (respectively $p_9 = p_{12}$). This time, the simplification was fully validated by the corrected AIC on all genotypes (Table 2, Eq.(9)). The expected error over the virtual genotypes was estimated to 13% and the calibration time was lowered by 24% with respect to the original model, thanks to structural simplification and the reduction of the number of parameters to be estimated. Accordingly to these results, the simplification by Eq.(9) was validated.

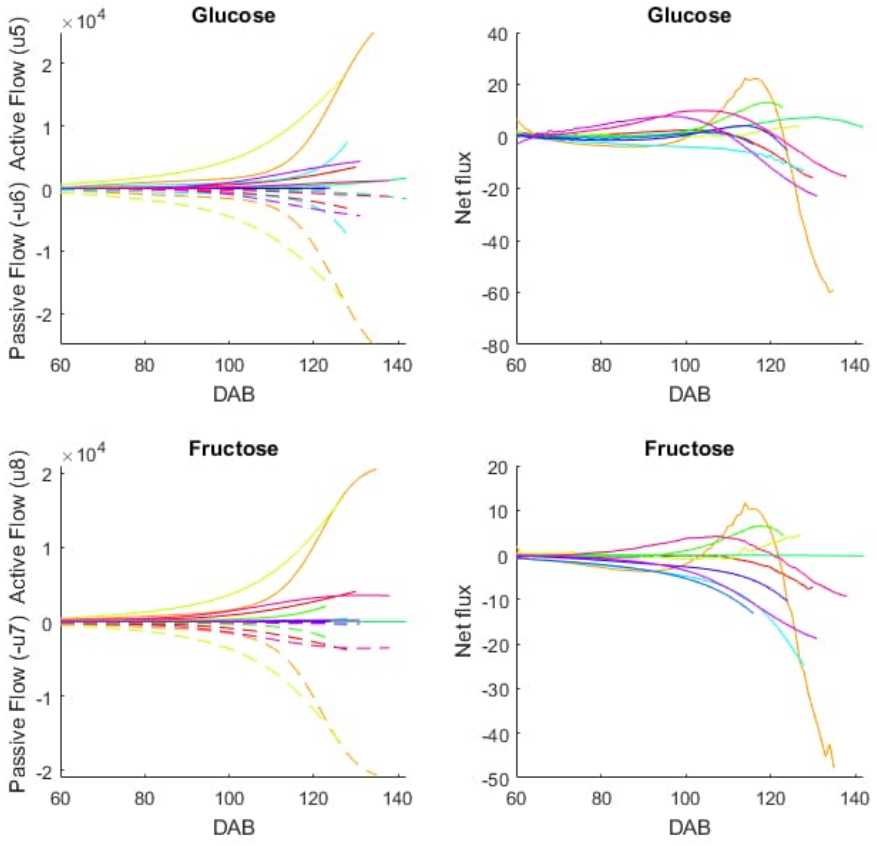

Figure 6: Evolution of the active flux (solid lines) and passive transport (dashed lines) for glucose (respectively fructose) and net flux during fruit development (DAB, day after bloom) for the ten genotypes of the training set (different colors).

6.3. Strategy 3: Time-scale analysis and QSSA

Results from the reduction strategies 1 and 2 were combined into an intermediate reduced model. This model had only 9 parameters to be estimated, linear flows and only one temporal enzymatic capacity, common to all genotypes. Improvement in AIC_C with respect to the original model confirmed a strong benefice for all ten genotypes (Table 2). The expected error over a large progeny was estimated around 20%, close to the performance of the original model.

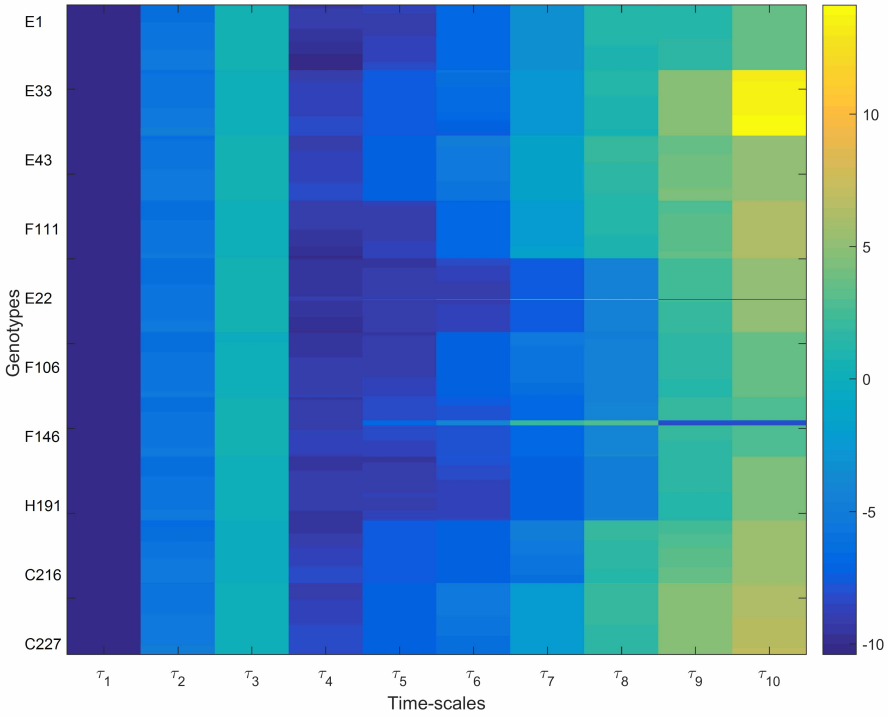

Figure 7: Order of magnitude of time scales τ_i along fruit development (DAB, days after bloom) for the 10 genotypes of the training set.

On the basis of this intermediate reduced model, time scale analysis was performed to detect the possible presence of fast modes in the system. The analysis of the Jacobian matrix, indeed, confirmed the presence of different modes, with typical time scales spanning a few seconds up to days, for all tested genotypes (Fig. 7).

A fast transient dynamics, followed by a slow one, was observable in the numerical simulations of the original and intermediate reduced models for the hexose phosphates concentration (variable x_5, see supplemental information, Fig. D.4). In addition, following the method proposed in [33, 17], we analyzed the predicted concentration of sugars in both intracellular compartments, for all genotypes. The concentration of the hexose phosphate (x_5) was systematically lower than the concentrations of the other variables in the system, as expected for the fast components of the system (Fig. 8).

Accordingly, x_5 was assumed to be at quasi-steady-state and its equation was replaced by an algebraic function of the slow variables.

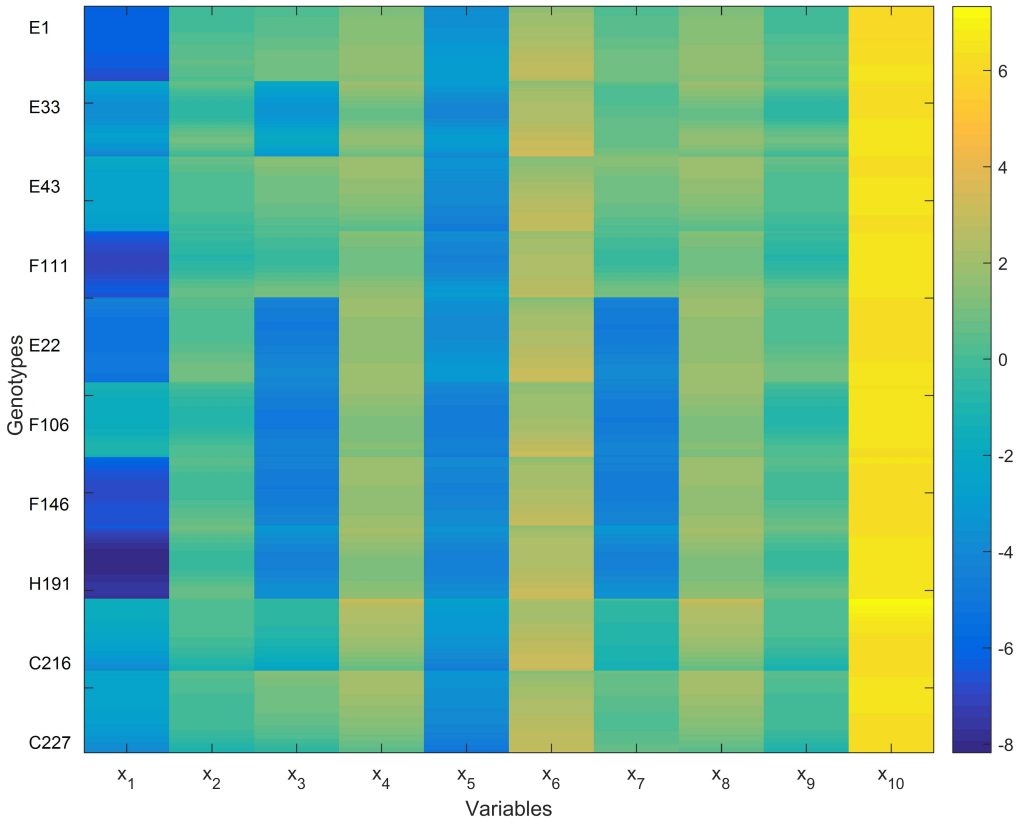

Figure 8: Order of magnitude of the predicted sugars concentrations (mg gFW^{-1}) in the cytosol (x_1: Sucrose, x_2: Sorbitol, x_3: Fructose, x_4: Glucose, x_5: Hexose Phosphate, x_{10}: Other compounds) and vacuole (x_6: Sucrose, x_7: Fructose, x_8: Glucose, x_9: Sorbitol), along fruit development (DAB, days after bloom) for the ten genotypes of the training set.

We compared the intermediate reduced model with its QSS approximation by calculating J_i (Eq.(20)) as explained previously. J_i was very low, less than 1%, over the whole dynamics for all variables (Fig. 9). This result was validated also on the virtual genotypes simulated with QQS approximation (see the supplemental information Fig. D.5). In addition the QSS assumption, further increased the performance of the model, leading to a gain in the calibration time of 40% with respect to the original model.

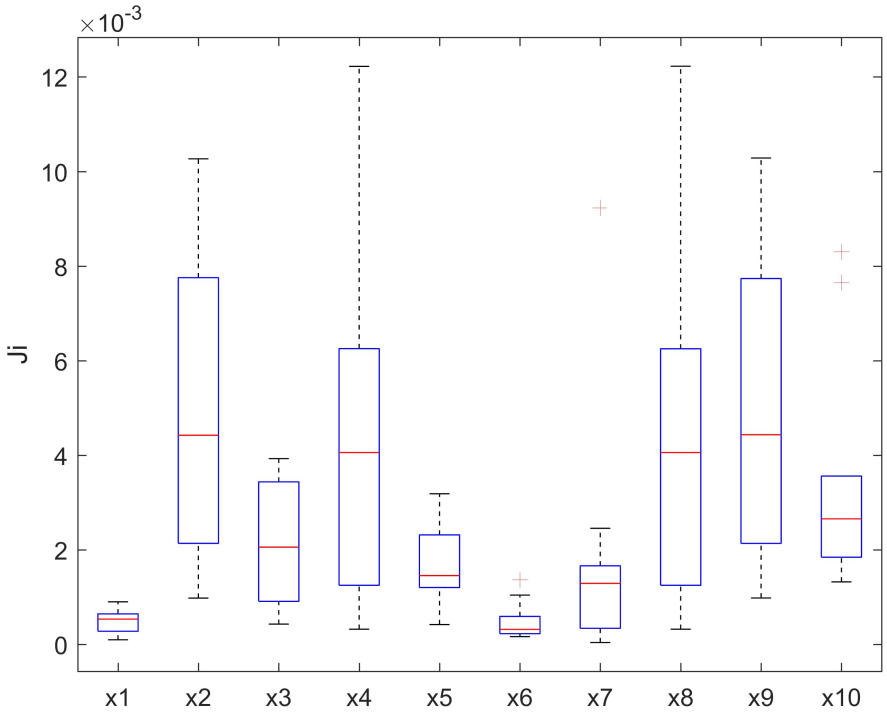

Figure 9: Normalized Root Mean Square Errors $J_i, i \in \{1, \ldots, 10\}$ between the intermediate and reduced models after application of the QSSA to x_5. The boxplot shows the variability of J_i over the training set

6.4. Evaluation of the reduced model

The validity of the reduced model was verified on some new genotypes of the inter-specific peach progeny, for which few data were available.

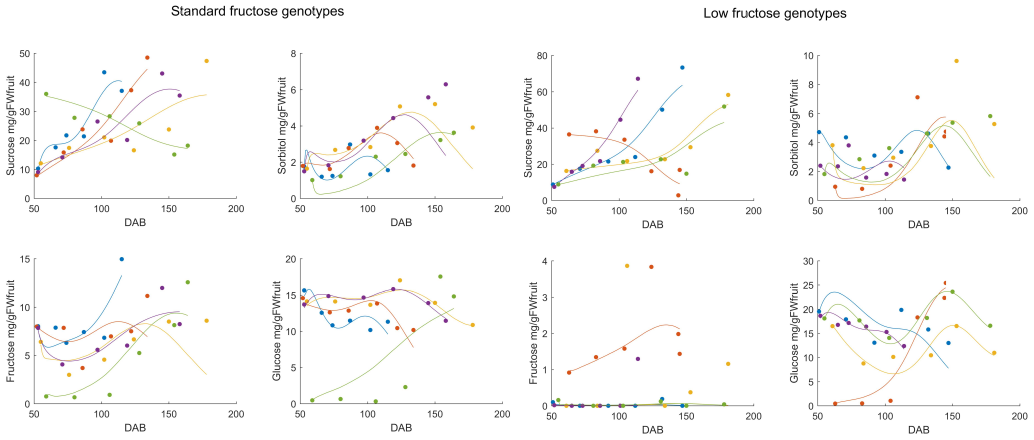

Figure 10: Evolution of the concentration ($mggFW^{-1}$) of sugars during fruit development (DAB, days after bloom) for ten representative genotypes of the validation set with standard (left) and low fructose (right) phenotypes. Dots represent experimental data and lines are model simulations.

The reduced model was then calibrated on the dynamics of sugar concentration of these selected genotypes, as described in section 5.2. The results presented in (Fig. 10) showed a satisfactory agreement between model and data, all over fruit development, for most genotypes. The average $NRMSE$ (Table D.6) ranged from 10% to 30% for the main sugars, in good agreement with estimations over the virtual progeny. These results confirmed that the reduced model offered a quality of prediction close to the original one with fewer parameters to be estimated and shorter integration time.

From a biological perspective, an important prediction of the model developed by Desnoues et al. [1] was that a difference in fructokinase affinity could be at the origin of the phenotypic difference observed between standard and low fructose genotypes.

We checked if the estimations obtained with the reduced model still supported this hypothesis. Fig. 11 shows a significant difference of estimated fructokinase affinity between the two phenotypic groups, in agreement with the original model based on the Student t-test (p-value $<2.0187e^{-9}$)

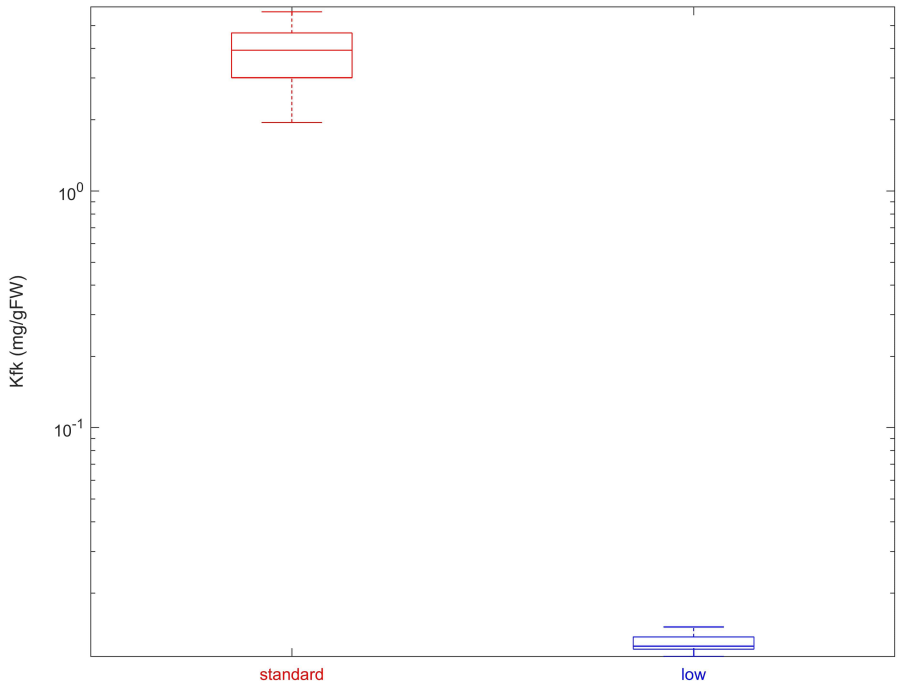

Figure 11: Difference in the estimated fructokinase affinity between standard and low fructose phenotypes, for forty genotypes (training and validation sets). The difference is significant with a p-value $< 2.0187e^{-9}$.

7. Discussion

Models of metabolic systems are usually very complex. Complexity stems from the number of components and the high degree of non-linearity included in both the network structure and the individual reaction rates. As a consequence, metabolic models usually suffer from numerical and identifiability issues that seriously hamper their application in the context of genetic studies, especially when they have to be calibrated for hundreds of genotypes. In this paper, we present a reduction scheme that explicitly accounts for genomic diversity. Our approach is based on the systematic evaluation of different reduction methods, that, if successful, are then combined together to yield the final reduced model. When applied to the model of sugar metabolism developed by Desnoues et al. [1] our approach led to a reduced model that

could be efficiently calibrated on a large diversity of genotypes, for which few data are available. The reduced model showed comparable predictions and biological interpretation as the original model, with only a limited number of estimated parameters. Indeed, calibration time was reduced by 40%, a considerable improvement when considering that the calibration of the original model could span up to 30 hours for a single genotype. Moreover, mitigation of model non-linearities can help limiting numerical issues and increase the reliability of estimated parameters, an important aspect in the context of genetic studies, where large genetic populations have to be calibrated.

The proposed reduction scheme is especially suitable for dynamical models of metabolic and biochemical networks, in which a large number of chemical reactions interact with similar non-linear kinetics. In these systems, indeed, the connectivity properties of the network usually prime over the precise description of the individual rate laws [40]. The presence of saturating kinetic functions (like the classical Michaelis-Menten), in particular, allows the simplification of the rate function depending on the substrate range whereas the presence of redundant or opposite reactions opens the way to structural simplification of the system. The extension of these reduction steps to another kind of models is less straightforward. Crop models for instance can involve a large variety of process kinetics, one for each described physiological process. The complexity of the cellular network is replaced by the interaction of a comparatively small number of processes but described by complicated, ad-hoc kinetic functions that can involve several model components as well as external environmental variables (temperature, humidity, light). The simplification of individual rate laws is still possible but it involves case-by-case study.

Although the application of specific reduction methods is tailored to model structure, the proposed evaluation strategy is pretty generic and easily adaptable to a large range of biological models. The main objective of this work was to provide a method to build a reduced model that is adapted to the application to a large panel of genotypes. In this sense, we do not look for the best model for a given genotype but rather for the best *compromise* in terms of accuracy and efficiency over a large genetic diversity. The question recalls the one of "model validation domain" i.e. the ability for a given model to describe data obtained in conditions different from those in which the model itself was calibrated [41]. Here it is about selecting for a reduced model having a large validation domain and able to cope with changes in model's inputs, parameter values, and initial conditions.

For this aim, we proposed a criterium based on the simulation of a large number of virtual genotypes and the systematic comparison of the expected distance between the original and the reduced models. Virtual genotypes are built based on the variability observed in a sub-sample of the population, plus a basal variability, expressed as a random effect, to limit the bias due to the choice of the initial sample and to assure a minimal diversity across the virtual population. A few remarks are needed. First, the above method tests the reliability of the reduction, assuming that the original model is valid. In this sense, the amplitude of the basal random effect should be subject to an expert knowledge so to avoid biologically unreasonable situations, that fall outside the conditions of applicability of the model. Second, it is worth to notice that, given the virtual nature of our comparison, the reduced model is parameterized using parameter values that are directly derived from the parameters of the original model, to which it is compared. In this sense, the 'expected NRMSE error' of the reduced model represents an upper bound of its actual accuracy over an experimental dataset, as parameter re-calibration can significantly improve the performances of the reduced model on real genetic populations.

Ultimately, the existence of a reduced model will considerably speed up the integration of genetic control into ecophysiological models. Currently, most genetic-improved ecophysiological models make use of Quantitative Trait loci (QTL) to describe the genetic architecture of specific model parameters. Basically, each parameter has a specific distribution in the population of genotypes and QTL analyses can be performed for each parameter to decipher the architecture of its genetic control (QTL number and effects, linkage). However, a major drawback of this approach is the difficulty in the calibration of the models for a large number of genotypes (due to a large number of parameters along with restricted number of observations) [6, 7]. Indeed, the statistical power of QTL analyses strongly depends on the size of the population and on the QTL effects i.e. their contribution to the variation of the trait they are associated with [42]. So, in order to be of interest, genetic parameters have to vary among genotypes and be quantifiable with relevant accuracy either experimentally or through numerical optimization.

In this perspective, a reduced model with a simpler structure will allow for a better exploration of the parameter space and a more accurate estimation of parameter values. Moreover, the improved calibration time opens the possibility of exploring larger genetic populations so to get more robust QTLs estimation. Finally, it will allow to do simulations over a large number of

environmental conditions and/or climatic scenarios.

This is an important step towards dealing with complex Genotype x Environment x Management interactions issues expected in the near future. The development of reliable gene-to-phenotype models will be an important lever to optimize farming in the future climatic conditions.

Acknowledgments

HK was founded by a scholarship of the Lebanese government. We would like to thank V. Signoret for her help in maintaining the inter-specific peach progeny. We are grateful to the IE-EMMAH UMR1114 and IE-GAFL UR1052 teams for taking care of the experimental orchard, and to Dr Olivier Martin for his help in statistics.

References

[1] E. Desnoues, M. Génard, B. Quilot-Turion, V. Baldazzi, A kinetic model of sugar metabolism in peach fruit reveals a functional hypothesis of a markedly low fructose-to-glucose ratio phenotype, The Plant Journal 94 (4) (2018) 685–698.

[2] J. W. White, G. Hoogenboom, Gene-Based Approaches to Crop Simulation, Agron. J. 95 (1) (2003) 52–64.

[3] V. Letort, P.-H. Cournède, A. Mathieu, P. De Reffye, T. Constant, Parametric identification of a functional–structural tree growth model and application to beech trees (fagus sylvatica), Functional plant biology 35 (10) (2008) 951–963.

[4] F. Tardieu, Virtual plants: modelling as a tool for the genomics of tolerance to water deficit, Trends in plant Science 8 (1) (2003) 9–14.

[5] B. Quilot-Turion, M. Génard, P. Valsesia, M.-M. Memmah, Optimization of allelic combinations controlling parameters of a peach quality model, Frontiers in plant science 7 (2016) 1873.

[6] P. Martre, B. Quilot-Turion, D. Luquet, M.-M. O.-S. Memmah, K. Chenu, P. Debaeke, Chapter 14 - model-assisted phenotyping and ideotype design, in: V. O. Sadras, D. F. Calderini (Eds.), Crop Physiology (Second Edition), second edition Edition, Academic Press, San Diego, 2015, pp. 349 – 373.

[7] N. Bertin, P. Martre, M. Genard, B. Quilot-Turion, C. Salon, Under what circumstances can process-based simulation models link genotype to phenotype for complex traits? case-study of fruit and grain quality traits, Journal of experimental botany 61 (2010) 955–67.

[8] N. Curien, F. Moreau, The music industry in the digital era: Toward new contracts, Journal of Media Economics 22 (2009) 102–113.

[9] T. Nägele, W. Weckwerth, Mathematical modeling reveals that metabolic feedback regulation of snrk1 and hexokinase is sufficient to control sugar homeostasis from energy depletion to full recovery, Frontiers in plant science 5 (2014) 365.

[10] B. P. Beauvoit, S. Colombié, A. Monier, M. . H. Andrieu, B. Biais, C. Bénard, C. Chéniclet, M. Dieuaide-Noubhani, C. Nazaret, J. . P. Mazat, Y. Gibon, Model-assisted analysis of sugar metabolism throughout tomato fruit development reveals enzyme and carrier properties in relation to vacuole expansion, Plant Cell 26 (2014).

[11] M. S. Okino, M. L. Mavrovouniotis, Simplification of mathematical models of chemical reaction systems, Chemical reviews 98 (2) (1998) 391–408.

[12] A. N. Gorban, N. K. Kazantzis, I. G. Kevrekidis, H. C. Öttinger, C. Theodoropoulos, Model reduction and coarse-graining approaches for multiscale phenomena, Springer, 2006.

[13] T. J. Snowden, P. H. van der Graaf, M. J. Tindall, Methods of model reduction for large-scale biological systems: a survey of current methods and trends, Bulletin of mathematical biology 79 (7) (2017) 1449–1486.

[14] J. Wei, J. C. W. Kuo, Lumping analysis in monomolecular reaction systems. analysis of the exactly lumpable system, Industrial & Engineering Chemistry Fundamentals 8 (1) (1969) 114–123.

[15] M. Sunnåker, G. Cedersund, M. Jirstrand, A method for zooming of nonlinear models of biochemical systems, BMC systems biology 5 (1) (2011) 140.

[16] M. Schauer, R. Heinrich, Quasi-steady-state approximation in the mathematical modeling of biochemical reaction networks, Mathematical biosciences 65 (2) (1983) 155–170.

[17] R. Heinrich, S. Schuster, The regulation of cellular systems, Chapman and Hall, 1996.

[18] T. Turányi, Sensitivity analysis of complex kinetic systems. tools and applications, Journal of mathematical chemistry 5 (3) (1990) 203–248.

[19] J. Cariboni, D. Gatelli, R. Liska, A. Saltelli, The role of sensitivity analysis in ecological modelling, Ecological Modelling 203 (1) (2007) 167 – 182.

[20] E. Vanuytrecht, D. Raes, P. Willems, Global sensitivity analysis of yield output from the water productivity model, Environmental Modelling Software 51 (2014) 323 – 332.

[21] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, S. Tarantola, Global sensitivity analysis: the primer, John Wiley & Sons, 2008.

[22] P. Holme, M. Huss, H. Jeong, Subnetwork hierarchies of biochemical pathways, Bioinformatics 19 (4) (2003) 532–538.

[23] J. Anderson, Y.-C. Chang, A. Papachristodoulou, Model decomposition and reduction tools for large-scale networks in systems biology, Automatica 47 (6) (2011) 1165 – 1174, special Issue on Systems Biology.

[24] X. Sun, M. Medvedovic, Model reduction and parameter estimation of non-linear dynamical biochemical reaction networks, IET Systems Biology 10 (2016) 10–16(6).

[25] F.-S. Wang, C.-L. Ko, E. O. Voit, Kinetic modeling using S-systems and lin-log approaches, Biochem. Eng. J. 33 (3) (2007) 238–247.

[26] I. E. Nikerel, W. A. van Winden, P. J. Verheijen, J. J. Heijnen, Model reduction and a priori kinetic parameter identifiability analysis using metabolome time series for metabolic reaction networks with linlog kinetics, Metab. Eng. 11 (1) (2009) 20–30.

[27] H. Schmidt, M. F. Madsen, S. Danø, G. Cedersund, Complexity reduction of biochemical rate expressions, Bioinformatics 24 (6) (2008) 848–854.

[28] W. Liebermeister, U. Baur, E. Klipp, Biochemical network models simplified by balanced truncation, The FEBS Journal 272 (16) (2005) 4034–4043.

[29] M. Apri, M. de Gee, J. Molenaar, Complexity reduction preserving dynamical behavior of biochemical networks, Journal of theoretical biology 304 (2012) 16–26.

[30] M. Lamboni, D. Makowski, S. Lehuger, B. Gabrielle, H. Monod, Multivariate global sensitivity analysis for dynamic crop models, Field Crops Research 113 (3) (2009) 312–320.

[31] R. Schürer, W. C. Schmid, Mint: A database for optimal net parameters, in: H. Niederreiter, D. Talay (Eds.), Monte Carlo and Quasi-Monte Carlo Methods 2004, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 457–469.

[32] E. Desnoues, Y. Gibon, V. Baldazzi, V. Signoret, M. Génard, B. Quilot-Turion, Profiling sugar metabolism during fruit development in a peach progeny with different fructose-to-glucose ratios, BMC Plant Biology 14 (1) (2014) 336.

[33] C. López Zazueta, O. Bernard, J.-L. Gouzé, Analytical reduction of nonlinear metabolic networks accounting for dynamics in enzymatic reactions, Complexity 2018 (2018).

[34] B. Quilot, M. Génard, J. Kervella, F. Lescourret, Analysis of genotypic variation in fruit flesh total sugar content via an ecophysiological model applied to peach, Theoretical and Applied Genetics 109 (2) (2004) 440–449.

[35] J. M. Chambers, T. J. Hastie, et al., Statistical models in S, Vol. 251, Wadsworth & Brooks/Cole Advanced Books & Software Pacific Grove, CA, 1992.

[36] C. Barrasso, M.-M. Memah, M. Génard, B. Quilot-Turion, Model-based qtl detection is sensitive to slight modifications in model formulation, PloS one 14 (10) (2019).

[37] M. Hosea, L. Shampine, Analysis and implementation of tr-bdf2, Applied Numerical Mathematics 20 (1-2) (1996) 21–37.

[38] D. E. Goldberg, Genetic algorithms in search, Optimization, and MachineLearning (1989).

[39] K. P. Burnham, D. R. Anderson, Model selection and multimodel inference : a practical information-theoretic approach, Springer, New York, 2002.

[40] A.-L. Barabási, Z. N. Z. N. Oltvai, Network biology: understanding the cell's functional organization, Nat. Rev. Genet. 5 (2) (2004) 101–113.

[41] F. Mairet, O. Bernard, Twelve quick tips for designing sound dynamical models for bioprocesses, PLoS Comput Biol in press (2019) 1–15.

[42] T. Mackay, E. Stone, J. Ayroles, The genetics of quantitative traits: challenges and prospects., Nat. Rev. Genet. 10 (8) (2009) 565–77.

[43] M. Génard, N. Bertin, H. Gautier, F. Lescourret, B. Quilot, Virtual profiling: a new way to analyse phenotypes, The Plant Journal 62 (2) (2010) 344–355.

[44] K. Shiratake, Y. Kanayama, S. Yamaki, Characterization of hexose transporter for facilitated diffusion of the tonoplast vesicles from pear fruit, Plant and cell physiology 38 (8) (1997) 910–916.

Appendices

Appendix A. Model description

Appendix A.1. Model equations

The original model [1] was written in terms of species *carbon* quantities $C(t)$. Here, we decided to rewrite the system as a function of species concentration $x_i(t)$, for a better readability. The quantity of carbon as a sugar i (C_i) depends on the concentration of i (x_i) according to the following equation:

$$C_i = \sigma_i \, x_i \, V_j \tag{A.1}$$

where σ_i is the carbon concentration of sugar i and V_j is the volume of the intracellular compartment (cytosol or vacuol) in which species i is located. The carbon content σ_i for the different sugar molecules is reported in Table A.1. Table A.2 specifies variable location within the cell's compartments. Differentiation of Equation (Eq. (A.1)) leads to:

$$\frac{dx_i}{dt} = \frac{1}{\sigma_i V_j}\frac{dC_i}{dt} - \frac{1}{V_j}x_i\frac{dV_j}{dt} \tag{A.2}$$

Accordingly, for variables $1, \ldots, 5, 10$, $C_i = \sigma_i x_i V_1$ whereas $C_i = \sigma_i x_i V_2$ for $i \in [6, 9]$. For simplicity,we assume $\frac{V_1}{V_2} = \alpha$. This leads to $\mu(t) = \frac{1}{V_1}\frac{dV_1}{dt} = \frac{1}{V_2}\frac{dV_2}{dt}$.

Table A.1: Carbon content of each sugar

σ	Sugar	Value
σ_1, σ_6	Sucrose	0.421
σ_3, σ_8	Fructose	0.4
σ_2, σ_7	Sorbitol	0.39
σ_4, σ_9	Glucose	0.4
σ_5	Hexose phosphate	0.27
σ_{10}	Other compounds	0.44

Table A.2: Model variables and location

S_1	Sucrose	Cytosol
S_2	Sorbitol	Cytosol
S_3	Fructose	Cytosol
S_4	Glucose	Cytosol
S_5	Hexose phosphate	Cytosol
S_6	Sucrose	Vacuole
S_7	Sorbitol	Vacuole
S_8	Fructose	Vacuole
S_9	Glucose	Vacuole
S_{10}	Other compounds	Cytosol

Table A.3: Reaction rates of the original and reduced models

Equations	Original Model	Reduced Model
Input flows	$I(t) = \sigma_f \frac{dDW}{dt} + R(t) = (\sigma_f + q_g)\frac{dDW}{dt} + q_m DWQ_{10}^{\frac{(T-20)}{10}}$ $R(t) = q_m DWQ_{10}^{\frac{(T-20)}{10}} + q_g \frac{dDW}{dt}$ $DW = DW(t_0) + w_1(1 - e^{-w_2 t}) + \frac{w_3}{1+e^{-w_4(t-w_5)}}$ $u_1(I) = \frac{1}{\sigma_1 V_1}\lambda\lambda_{suc}(t)I(t)$ $\lambda_{suc}(t) = \frac{p_1 t}{t_{max}}$ where t_{max} corresponds to the maturation time $u_2(I) = \frac{1}{\sigma_2 V_1}(1-\lambda)I(t)$ $u_3(I) = \frac{1}{\sigma_3 V_1}\frac{\lambda}{2}(1-\lambda_{suc}(t))I(t)$	
Metabolism	$u_9(v_2, x_1) = \frac{v_2(t)}{p_5+x_1}x_1(t)$	$u_9(x_1) = \frac{v_2}{p_5}x_1(t) = r_1 x_1(t)$
	$u_{10}(x_1) = \frac{v_3}{p_{21}+x_1}x_1(t)$	$u_{10}(x_1) = \frac{v_3}{p_{21}}x_1(t) = r_2 x_1(t)$
	$u_{11}(v_4, x_2) = \frac{v_4(t)}{p_{22}+x_2}x_2(t)$	$u_{11}(x_2) = \frac{v_4}{p_{22}}x_2 = r_3 x_2(t)$
	$u_{12}(v_5, x_2) = \frac{v_5(t)}{p_{13}+x_2}x_2(t)$	$u_{12}(x_2) = \frac{v_5}{p_{13}}x_2(t) = r_4 x_2(t)$
	$u_{13}(x_6, x_8, x_9) = \frac{v_1}{(1+\frac{x_8+x_9}{p_2})p_{23}+x_6}x_6(t)$	$u_{13}(x_6) = r_5 x_6(t)$
	$u_{14}(v_6, x_3) = \frac{v_6(t)}{p_3+x_3}x_3(t)$	$u_{14}(x_3) = \frac{v_6}{p_3}x_3(t) = r_6 x_3(t)$
	$u_{15}(v_7, x_4) = \frac{v_7(t)}{p_4+x_4}x_4(t)$	$u_{15}(v_7, x_4) = \frac{v_7(t)}{p_4}x_4(t)$
	$u_{16}(x_5) = p_7 x_5(t)$	$u_{16}(x_5) = p_7 x_5(t)$
	$u_{17}(x_5) = p_6 x_5(t)$	$u_{17}(x_5) = p_6 x_5(t)$
	$u_{18}(R) = R(t)$	$u_{18}(R) = R(t)$
Transport processes	$u_4(S, x_1) = p_8 x_1(t)S(t)$	$u_4(S, x_1) = p_8 x_1(t)S(t)$
	$u_5(S, x_3, x_4) = \frac{p_{11}}{p_{19}+x_3+x_4}x_4(t)S(t)$	$u_5 = 0$
	$u_6(S, x_4, x_9) = (x_9 - x_4)p_{10}S(t)$	$u_6(S, x_4, x_9) = (x_9 - x_4)p_{10}S(t)$
	$u_7(S, x_3, x_4) = \frac{p_{12}}{p_{20}+x_3+x_4}x_3(t)S(t)$	$u_7 = 0$
	$u_8(S, x_3, x_8) = (x_8 - x_3)p_9 S(t)$	$u_8(S, x_3, x_8) = (x_8 - x_3)p_9 S(t)$
	$u_{19}(S, x_2, x_7) = p_{14}(x_7 - x_2)S(t)$	$u_{19}(S, x_2, x_7) = p_{14}(x_7 - x_2)S(t)$

The original model by Desnoues et al. [1] was composed by a network

of 19 reactions and one input function $I(t)$. The latter described the carbon supply from the mother plant to the fruit and it was estimated as the sum of the carbon used for fruit dry mass (DW) increase and the carbon lost by respiration ($R(t)$). Two parameters λ and λ_{suc} described the fraction of the input flow that is converted into the different forms of sugars. Fruit respiration was computed following the growth-maintenance paradigm, as described in [1].

Reaction rates are reported in Table A.3. Enzymatic reactions were generally described using an irreversible Michaelis-Menten kinetics, with experimentally-measured capacities $v_i(t)$. Transport processes between cytosol and vacuole were assumed proportional to the vacuole surface (hypothesis of constant density of transporters) computed from vacuole fresh mass (proxy of the volume) supposing the vacuole as a sphere of surface $S(t) = (4\pi)^{\frac{1}{3}}(V_2)^{\frac{2}{3}}$ (see [1] for more information). Both active and passive transport mechanisms were considered for fructose and glucose.

Model equations are reported in Table A.4, for both the original and the reduced model.

Table A.4: System of original and reduced models

System of original model	**System of reduced model**
$\frac{dx_1}{dt} = u_1 + \frac{\sigma_5}{\sigma_1}u_{16} - u_{10} - u_4 - \mu(t)x_1$	
$\frac{dx_2}{dt} = u_2 - u_{11} - u_{12} + \frac{1}{\sigma_2 V_1}u_{19} - \mu(t)x_2$	
$\frac{dx_3}{dt} = u_3 + \frac{1}{\sigma_3 V_1}u_8 + \frac{1}{2}\frac{\sigma_1}{\sigma_3}u_9 + \frac{1}{2}\frac{\sigma_1}{\sigma_3}u_{10} + \frac{\sigma_2}{\sigma_3}u_{11} - u_7 - u_{14} - \mu(t)x_3$	$\frac{dx_3}{dt} = u_3 + \frac{1}{\sigma_3 V_1}u_8 + \frac{1}{2}\frac{\sigma_1}{\sigma_3}u_9 + \frac{1}{2}\frac{\sigma_1}{\sigma_3}u_{10} + \frac{\sigma_2}{\sigma_3}u_{11} - u_{14} - \mu(t)x_3$
$\frac{dx_4}{dt} = u_3 + \frac{1}{\sigma_4 V_1}u_6 + \frac{1}{2}\frac{\sigma_1}{\sigma_4}u_{10} + \frac{\sigma_2}{\sigma_4}u_{12} - u_5 - u_{15} - \mu(t)x_4$	$\frac{dx_4}{dt} = u_3 + \frac{1}{\sigma_4 V_1}u_6 + \frac{1}{2}\frac{\sigma_1}{\sigma_4}u_{10} + \frac{\sigma_2}{\sigma_4}u_{12} - u_{15} - \mu(t)x_4$
$\frac{dx_5}{dt} = \frac{1}{2}\frac{\sigma_1}{\sigma_5}u_9 + \frac{\sigma_3}{\sigma_5}u_{14} + \frac{\sigma_4}{\sigma_5}u_{15} - u_{17} - u_{16} - \frac{1}{\sigma_5 V_1}u_{18} - \mu(t)x_5$	$x_5 = \frac{1}{p_6+p_7+\mu(t)}(\frac{1}{2}\frac{\sigma_1}{\sigma_5}u_9 + \frac{\sigma_3}{\sigma_5}u_{14} + \frac{\sigma_4}{\sigma_5}u_{15} - \frac{1}{\sigma_5 V_1}u_{18})$
$\frac{dx_6}{dt} = \alpha u_4 - u_{13} - \mu(t)x_6$	
$\frac{dx_7}{dt} = -\frac{1}{\sigma_7 V_2}u_{19} - \mu(t)x_7$	
$\frac{dx_8}{dt} = \alpha u_7 + \frac{1}{2}\frac{\sigma_6}{\sigma_8}u_{13} - \frac{1}{\sigma_8 V_2}u_8 - \mu(t)x_8$	$\frac{dx_8}{dt} = \frac{1}{2}\frac{\sigma_6}{\sigma_8}u_{13} - \frac{1}{\sigma_8 V_2}u_8 - \mu(t)x_8$
$\frac{dx_9}{dt} = \alpha u_5 + \frac{1}{2}\frac{\sigma_6}{\sigma_9}u_{13} - \frac{1}{\sigma_9 V_2}u_6 - \mu(t)x_9$	$\frac{dx_9}{dt} = \frac{1}{2}\frac{\sigma_6}{\sigma_9}u_{13} - \frac{1}{\sigma_9 V_2}u_6 - \mu(t)x_9$
$\frac{dx_{10}}{dt} = \frac{\sigma_5}{\sigma_{10}}u_{17} - \mu(t)x_{10}$	

Appendix A.2. Model parameterization and initialization

A total of 23 parameters are needed to fully define the reaction rates of Table A.3. Following [1], 9 of these parameters were fixed based on published data, which were obtained from research studies of peach or fruit. The remaining 14 parameters were estimated numerically, as described in section 5.2. In order to compare model and data, sugar concentrations at the fruit level have to be computed from model variables, describing the metabolite concentration within intra-cellular compartments. Assuming a constant proportion of vacuolar space in fruit cells, the concentration of each sugar j (sucrose, glucose, fructose, and sorbitol) at the fruit level is given by:

$$\mathbb{E}(X_j) = \mathcal{M}_p(x_i) = x_i^{vac} \frac{1}{\alpha + 1} + x_i^{cyt} \frac{\alpha}{\alpha + 1} \tag{A.3}$$

where x_i^{vac} and x_i^{cyt} are respectively the variables located in the vacuole ($i \in [6, 9]$) and cytosol ($i \in [1, 5]$) (see Table A.2) and $\alpha = \dfrac{V_1}{V_2}$ is the intra-cellular volume ratio. The value of α was estimated by cytological analysis to 0.08 (see [1] for more information). Fruit fresh mass was assumed as a proxy for total volume $V_1 + V_2$.

For each genotype k, initial conditions $x_0^{(k)}$ were set using the concentrations $X^{(k)}(t_0)$ of sucrose, glucose, fructose, sorbitol, and hexose phosphates, measured at the fruit level at the first stage of development. The conversion between total and intra-cellular metabolite concentrations was performed based on metabolite localization at maturity. Accordingly, 98% of fructose, glucose, sucrose content and 90% of sorbitol content were assumed to be located in the vacuole, whereas the hexose phosphates were restricted to the cytosol. Accordingly, for sucrose, fructose, and glucose:

$$\begin{aligned} \text{cytosol}: \quad & x_i^{(k)}(t_0) = 0.02\, X^{(k)}(t_0)\, \frac{(1+\alpha)}{\alpha} && i \in \{1, 3, 4\} \\ \text{vacuole}: \quad & x_i^{(k)}(t_0) = 0.98\, X^{(k)}(t_0)\, (1+\alpha) && i \in \{6, 8, 9\} \end{aligned}$$

for sorbitol,

$$\begin{aligned} \text{cytosol}: \quad & x_i^{(k)}(t_0) = 0.10\, X^{(k)}(t_0)\, \frac{(1+\alpha)}{\alpha} && i = 2 \\ \text{vacuole}: \quad & x_i^{(k)}(t_0) = 0.90\, X^{(k)}(t_0)\, (1+\alpha) && i = 7 \end{aligned}$$

and for the hexoses phosphates

$$\text{cytosol}: \quad x_i^{(k)} = X^{(k)}(t_0)\,\frac{(1+\alpha)}{\alpha} \qquad\qquad i = 10$$

Table A.5: Table of original and reduced models parameter description

p_i	Parameter	Corresponding model	Description	Reference	Value	Unit
p_1	λ_{Suc}	original and reduced	sucrose proportion hydrolyzed in the apoplasm	Estimated	0-1	
p_8	$TactifSuc$	original and reduced	coefficicent of sucrose transport (active import) from cytosol to vacuole	Estimated	0-400	mg gFW^{-1}day^{-1}
p_{10}	$TpassifGlu$	reduced original	coefficient of glucose passive transport between cytosol and vacuole and in the opposite direction	Section. 3.1 Estimated	112.1559 0-150	mg gFW^{-1}day^{-1}
p_9	$TpassifFru$	original and reduced	coefficient of fructose passive transport between cytosol and vacuole and in the opposite direction	Estimated	0-150	mg gFW^{-1}day^{-1}
$r_1 = \frac{v_2}{p_5}$	$R_{susy} = \frac{V_{susy}}{K_{susy}}$	reduced	coefficient of the transfer function between sucrose and (fructose + hexoses phosphate) under action of sucrose synthase (susy) enzyme	Section. 3.2.2	1.8809	day^{-1}
$r_2 = \frac{v_3}{p_{21}}$	$R_{ni} = \frac{V_{ni}}{K_{ni}}$	reduced	coefficient of the transfer function between sucrose and (glucose +fructose) under action of neutral invertase (ni) enzyme	Section. 3.2.2	95.5875	day^{-1}
$r_3 = \frac{v_4}{p_{22}}$	$R_{sdh} = \frac{V_{sdh}}{K_{sdh}}$	reduced	coefficient of the transfer function between sorbitol and fructose under action of sorbitol dehydrogenase (sdh) enzyme	Section. 3.2.2	7.1592	day^{-1}
$r_4 = \frac{v_5}{p_{13}}$	$R_{so} = \frac{V_{so}}{K_{so}}$	reduced	coefficient of the transfer function between sorbitol and glucose under action of sorbitol oxydase (so) enzyme	Estimated	0-10	day^{-1}
$r_5 = \frac{v_1}{p_{23}}$	$R_{ai} = \frac{V_{ai}}{K_{ai}}$	reduced	coefficient of the transfer function between sucrose and (glucose +fructose) under action of acid invertase (ai) enzyme	Estimated	0-1	day^{-1}
p_2	Ki_{AI}	original	inhibitor constant of acid invertase	Estimated	0-10	mg gFW^{-1}
$r_6 = \frac{v_6}{p_3}$	$R_{fk} = \frac{V_{fk}}{K_{fk}}$	reduced	coefficient of the transfer function between fructose and hexoses phosphate under action of fructokinase (fk) enzyme	Estimated	0-5000	day^{-1}
v_7	$Vhk(t)$	reduced	hexokinase activity (hk) to transfer glucose to hexoses phosphate	Section. 3.2.1	$86.2 - 2.3t + 2e^{-2}t^2 - 8.3e^{-5}t^3$	mg gFW^{-1}day^{-1}
p_4	Khk	original and reduced	hexokinase affinity	Estimated	1-300	mg gFW^{-1}
p_7	$ReSyntSuc$	original and reduced	coefficient of the transfer function between hexoses phosphate and sucrose	Estimated	0-300	day^{-1}
p_6	$OthComp$	original and reduced	coefficient of the transfer function between hexoses phosphate and other compounds	Estimated	450-1500	day^{-1}
p_{14}	$TpassifSor$	reduced original	coefficient of sorbitol passive transport between cytosol and vacuole	Section. 3.1 Estimated	4.1305 0-150	mg gFW^{-1}day^{-1}
σ_f	$PropCdw$	original and reduced	carbon concentration of the mesocarp	[43]	0.44	gC gDW^{-1}
p_{17}	q_g	original and reduced	growth respiration coefficient	[43]	0.084	gC gDW^{-1}
p_{18}	q_m	original and reduced	maintenance respiration coefficient	[43]	2.76e-4	gC gDW^{-1}day^{-1}
p_{16}	Q_{10}	original and reduced	temperature ratio of maintenance respiration	[43]	1.9	
p_{15}	λ	original and reduced	sucrose sap proportion	[1]	0.65	
p_{11}	$VmtactifFru$	original	fructose active import (activity)	Estimated	0-150	mg gFW^{-1}day^{-1}
p_{12}	$VmtactifGlu$	original	Glucose active import (activity)	Estimated	0-150	mg gFW^{-1}day^{-1}
p_{19}	$KmtactifGlu$	original	Glucose active import (affinity)	[44]	0.054	mg gFW^{-1}
p_{20}	$KmtactifFru$	original	fructose active import (affinity)	[44]	0.288	mg gFW^{-1}

Appendix B. Multi-variate sensitivity analysis

Appendix B.1. Multi-variate sensitivity analysis

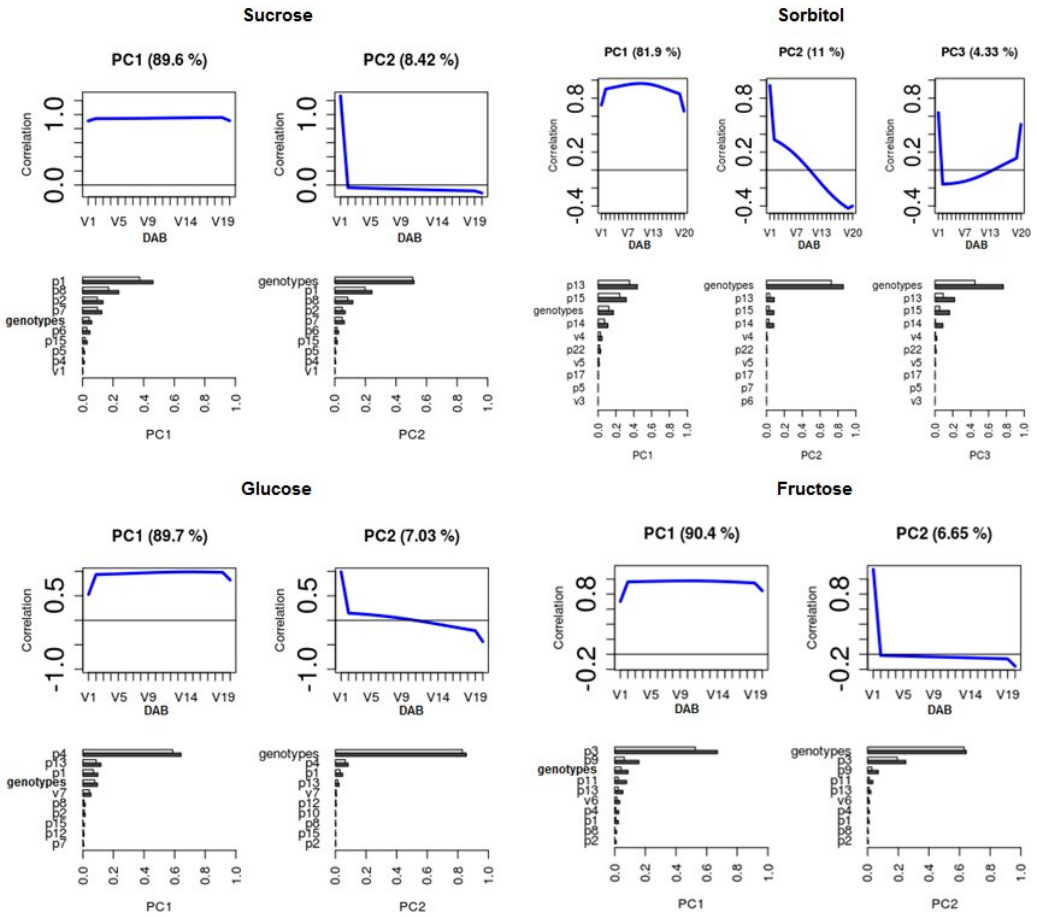

Figure B.1: PCA-based sensitivity analysis of the sugar model. Columns: principal components. Top row: correlation coefficients (y-axis) between the principal component and the output of each sugar during fruit development (DAB, days after bloom on the x-axis). Bottom row: first order sensitivity indices (dark bars) and total sensitivity indices (pale bars).

Multivariate sensitivity [30] was used to identify the influence of each parameter on the dynamic output $x(t)$ during fruit development. Where $x(t)$ is the sugar concentration (sucrose, glucose, fructose and glucose) and t is the independent time variable for 20 days after bloom ($t = (V1 =$

$min(DAB), V2 = max(DAB)/2 + 2, \ldots, V19 = max(DAB)/2 + 19, V20 = max(DAB)$))). Results of the principal components and sensitivity principal indices are presented in Fig. B.1. For sucrose, glucose and fructose, the first two components explained 96% of the total inertia of the simulated sugar dynamics. For sorbitol, the first three components explained 97%. The first component was positively correlated with all time-points. Correlation values in Fig. B.1 show that the first principal component corresponds to the average concentration of sugars (sucrose, glucose, fructose and sorbitol) produced during the whole fruit development. The second principal component was positively correlated with sugar concentration at stage 1 and poorly or slightly negatively correlated with simulated sugar during the second half of fruit development. Thus, the second principal component corresponds to the difference in sugar initialization values, that strongly depends on the `genotype` factor. For sorbitol, the third principal component accounts for a much smaller part of inertia, associated with the difference between the sorbitol produced in the middle of fruit development and the sorbitol produced both very early and late. It was sensitive to the set of genotypes.

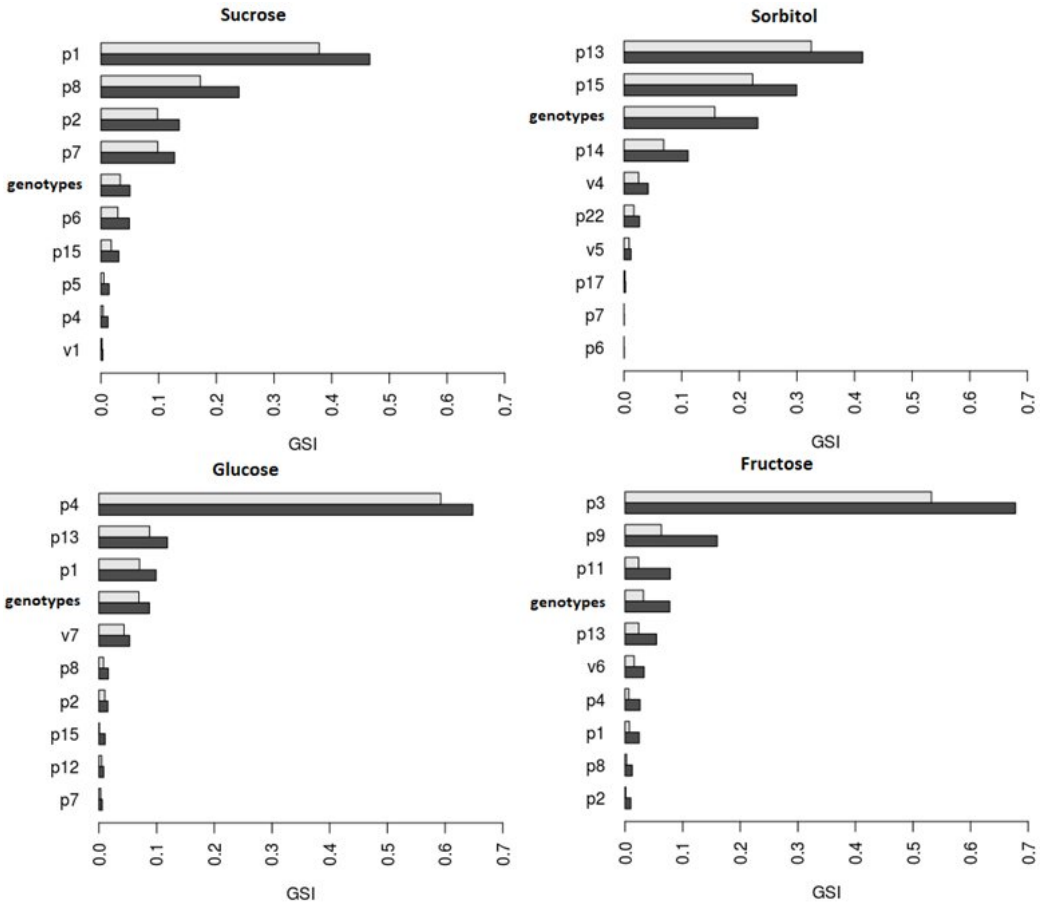

Figure B.2: Generalized sensitivity indices (GSI) for the first ten sensitive parameters (p_i) and ten genotypes (the training set) on four outputs (Sucrose, Sorbitol, Glucose, and Fructose) of the sugar model. The main sensitivity indices are in dark bars and interaction ones are in grey bars.

The generalized sensitivity indices (GSI) shown in Fig. B.2 gives a common ranking of model parameters according to their influence on the four output sugars (Sucrose, Sorbitol, Glucose and Fructose), for all tested genotypes. Parameter p_1 related to the action of cell-wall invertase in fruit apoplasm is the most important parameter, for its effect on both sucrose (rank 1) and glucose (rank 3) dynamics. The activities of Fructokinase (p_3) and Hexokinase (p_4) are the most sensitive parameters for fructose and glucose concentrations, respectively, whereas the sorbitol oxydase affinity (p_{13}) and the proportion of sucrose in the plant sap (p_{15}) affect sorbitol content in the fruit.

Interestingly, the `genotype` factor is ranked only third to fifth, depending on the sugar, meaning that it does not affect parameters' sensitivity as much as expected. A closer look at the results shows that the choice of the genotype essentially affects the second principal component, via the definition of the initial conditions of the model (see the supplemental information Fig. B.1).

Appendix C. Virtual experiment

In order to evaluate the reliability of the proposed simplifications over a larger diversity, a progeny of virtual genotypes was randomly created based on a careful recombination, with noise, of the original 10 dynamics. This included changes in parameters values, initial conditions and input functions.

We used the results from the principal component analysis (PCA) performed on growth rate and growth duration for the whole progeny of 106 genotypes to verify the distribution of virtual genotypes. To this aim, growth rates and durations of the 20 000 virtual genotypes were projected on the PCA plan defined by the previous analysis. As shown in Fig. C.3, the virtual genotypes provide a good representation of the diversity in growth rate and growth duration observed in the real progeny.

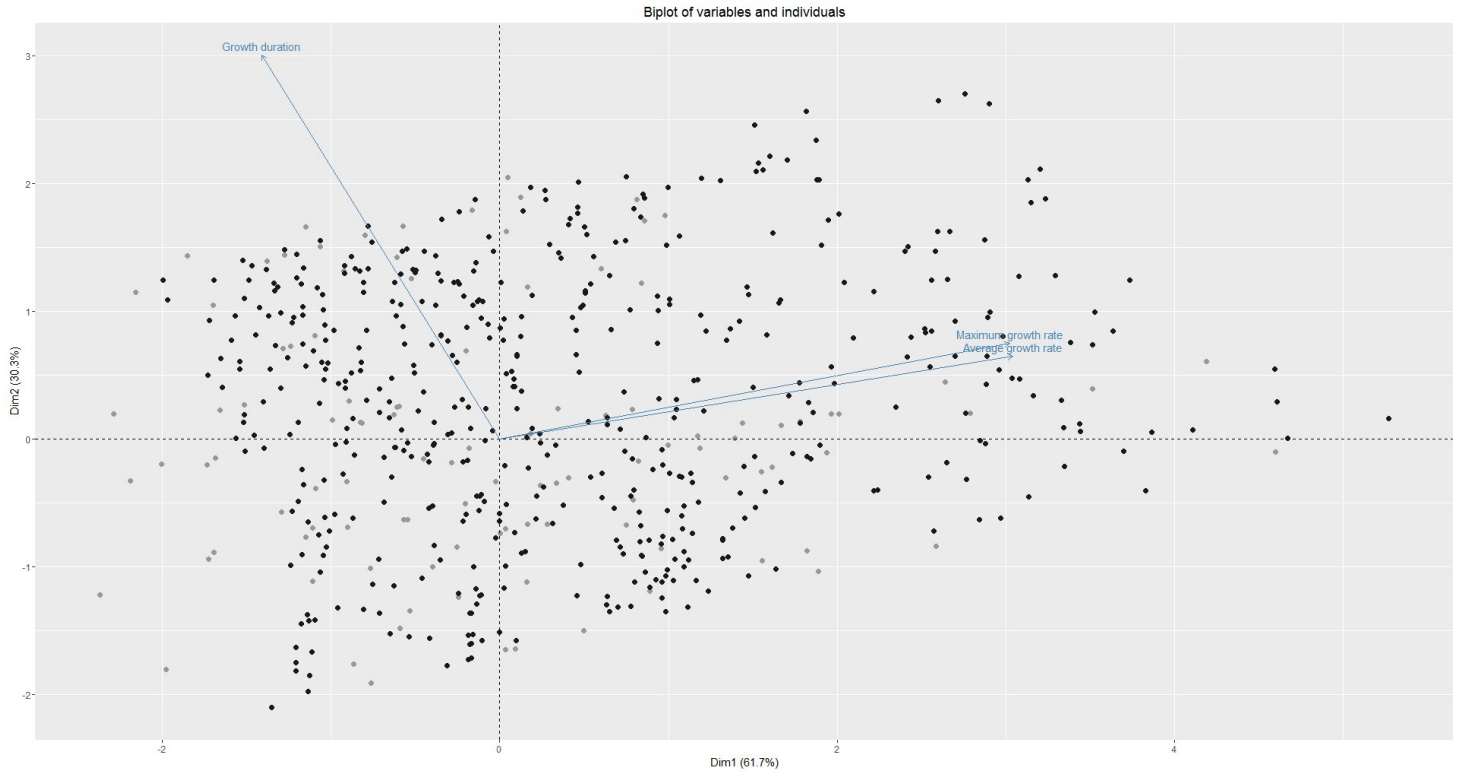

Figure C.3: Principal component analysis (PCA) for the whole progeny of 106 genotypes (grey) and 500, out of 20000, virtual genotypes (black). Represents the projection on the Dim1 and Dim2 of the growth duration and growth rate obtained with curves growth.

Appendix D. Time-scale analysis and QSSA

Timescale-based approaches and quasi-steady-state approximation [17] were applied to reduce the number of ODEs of the model and to obtain the final reduced model. The predicted concentrations of sugars in both intracellular compartments were analyzed. A fast transient dynamics of the concentration of the hexose phosphate (x_5), followed by a slow one, was observable in the numerical simulations of the original and intermediate reduced model (Fig. D.4. Together with the analysis of the Jacobian matrix, this observation led to the assumption of x_5 as a fast variable of the system. Quasi-steady-state approximation on x_5, indeed, strongly reduced the fast transient dynamics in the final reduced model, for most genotypes. Notice that a few fast modes (already pointed out by the analysis of the Jacobian matrix) may nonetheless remain in the system. Their elimination would require a linear combination of the original variables, which is incompatible with our objective to preserve the biological interpretation of the model. We therefore decided not to push the simplification of the model further.

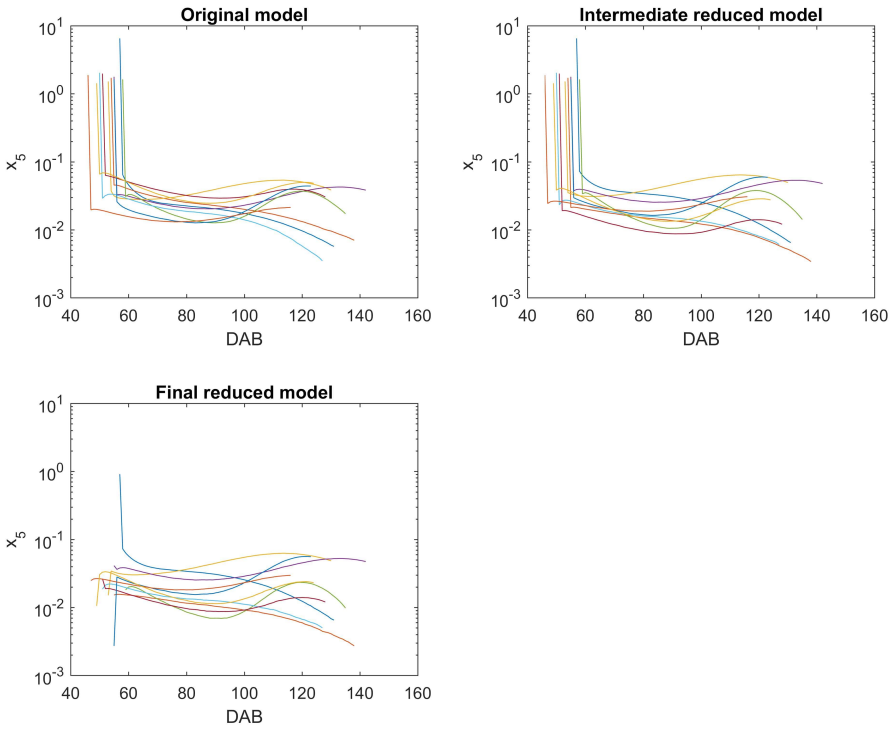

Figure D.4: Evolution of the concentration ($mggFW^{-1}$) of x_5 : *Hexose Phosphate* during fruit development (DAB, days after bloom) for ten genotypes for the original, intermediate reduced and final models.

Appendix D.1. Results of quasi-steady-state approximation applied on the intermediate reduced model for the 20 000 virtual genotypes

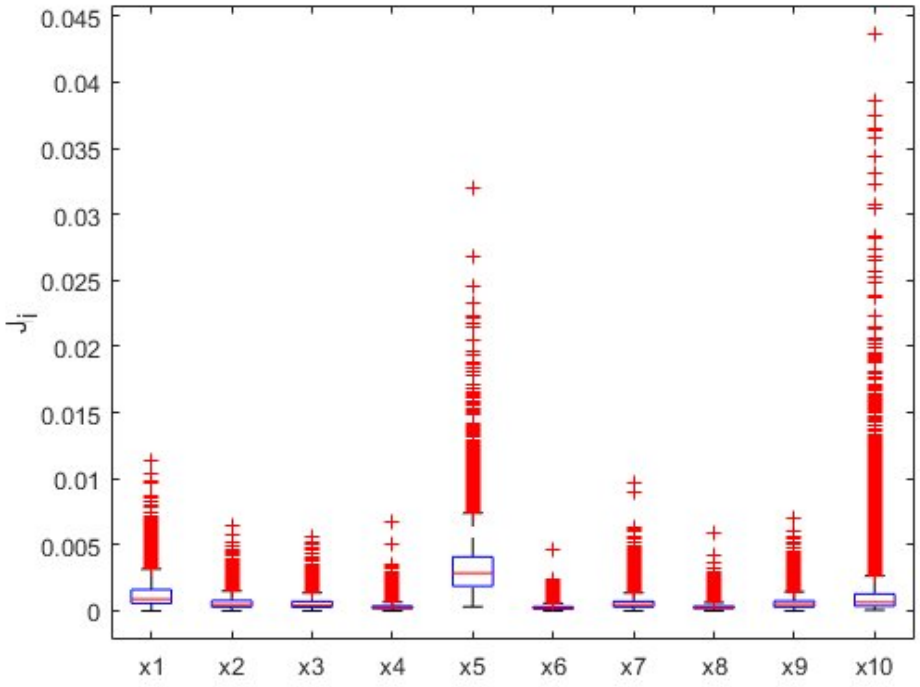

Figure D.5: Normalized Root Mean Square Errors $J_i, i \in \{1, \ldots, 10\}$ between the intermediate and reduced model after application of the QSSA to x_5. The boxplot shows the variability of J_i over the virtual genotypes

We compared the intermediate reduced model with its QSS approximation by calculating the Normalized Root Mean Square Error (J_i) on the 20 000 virtual genotypes. All J_i was very low, less than 0.045, over the whole dynamics for all variables (Fig. D.5).

Table D.6: NRMSE between model simulation and experimental data. Calculated values of the normalized root mean squared error (NRMSE) are presented for each genotype, the four sugars separately.

	Genotype	Phenotype	Year	Sucrose	Sorbitol	Fructose	Glucose	Mean
Trainig set	E1	Standard	2012	0.09	0.14	0.16	0.26	0.16
	E33	Standard	2012	0.04	0.19	0.27	0.17	0.16
	E43	Standard	2012	0.07	0.11	0.24	0.16	0.14
	F111	Standard	2012	0.13	0.13	0.21	0.23	0.17
	C227	Standard	2011	0.07	0.28	0.16	0.13	0.16
	E22	Low	2012	0.11	0.09	0.21	0.18	0.14
	F106	Low	2012	0.07	0.7	0.14	0.15	0.26
	F146	Low	2012	0.05	0.14	0.02	0.17	0.10
	H191	Low	2012	0.07	0.15	0.18	0.16	0.14
	C216	Low	2011	0.08	0.26	0.25	0.11	0.17
Validation set	H163	Standard	2012	0.10	0.35	0.18	0.19	0.20
	F107	Standard	2012	0.11	0.11	0.24	0.25	0.17
	E23	Standard	2012	0.15	0.30	0.10	0.17	0.18
	E17	Standard	2012	0.14	0.37	0.19	0.05	0.18
	E21	Standard	2012	0.13	0.21	0.16	0.24	0.18
	E41	Low	2012	0.08	0.35	0.35	0.24	0.25
	E18	Low	2012	0.13	0.26	0.26	0.09	0.18
	F113	Low	2012	0.12	0.32	0.35	0.20	0.24
	F90	Low	2012	0.06	0.47	0.26	0.13	0.23
	C243	Low	2012	0.18	0.49	0.24	0.09	0.25
	C199	Low	2012	0.22	0.46	0.25	0.19	0.28
	C207	Low	2012	0.19	0.28	0.23	0.24	0.23
	E36	Low	2012	0.09	0.13	0.13	0.16	0.12
	E48	Low	2012	0.07	0.41	0.14	0.14	0.19
	F101	Low	2012	0.15	0.43	0.20	0.05	0.20
	F109	Low	2012	0.07	0.31	0.27	0.24	0.22
	F127	Low	2012	0.09	0.17	0.22	0.10	0.14
	F144	Low	2012	0.19	0.11	0.24	0.22	0.19
	F141	Low	2012	0.14	0.22	0.36	0.16	0.22
	F86	Low	2012	0.06	0.13	0.30	0.32	0.20
	C232	Standard	2012	0.05	0.30	0.25	0.13	0.18
	E5	Standard	2012	0.22	0.07	0.12	0.24	0.16
	E19	Standard	2012	0.05	0.14	0.24	0.03	0.11
	E20	Standard	2012	0.19	0.17	0.07	0.18	0.11
	E26	Standard	2012	0.05	0.20	0.18	0.38	0.20
	E34	Standard	2012	0.22	0.43	0.20	0.15	0.25
	E35	Standard	2012	0.11	0.27	0.19	0.07	0.16
	E37	Standard	2012	0.24	0.24	0.13	0.26	0.21
	F83	Standard	2012	0.09	0.19	0.06	0.27	0.15
	F115	Standard	2012	0.03	0.20	0.10	0.07	0.10

907 The NRMSE can defined as follows:

$$NRMSE(\tilde{p}^{(k)}) = \sum_{i=1}^{4} J_i(\tilde{p}^{(k)})$$

908 with

$$J_i(\tilde{p}^{(k)}) = \frac{\sqrt{\frac{1}{N_M}\sum_{j=1}^{N_M}(\tilde{x}_i(t_j, \tilde{p}^{(k)}) - X_i^{(k)}(t_j))^2}}{max_j(X_i^{(k)}(t_j)) - min_j(X_i^{(k)}(t_j))} \tag{D.1}$$

909 where N_M is the number of observations, $\tilde{x}(t, \tilde{p}^{(k)})$ are the concentrations
910 predicted by the model and $X^{(k)}t(t)$ are the experimental data and i is the
911 sugar index.