ACS Synthetic Biology

Research Article

pubs.acs.org/synthbio

Efficient Behavior of Photosynthetic Organelles via Pareto Optimality, Identifiability, and Sensitivity Analysis

Giovanni Carapezza,[†] Renato Umeton,[‡] Jole Costanza,[†] Claudio Angione,[¶] Giovanni Stracquadanio,[§] Alessio Papini,[∥] Pietro Lió,*[,¶] and Giuseppe Nicosia*[,†]

[†]Department of Mathematics and Computer Science, University of Catania, Italy

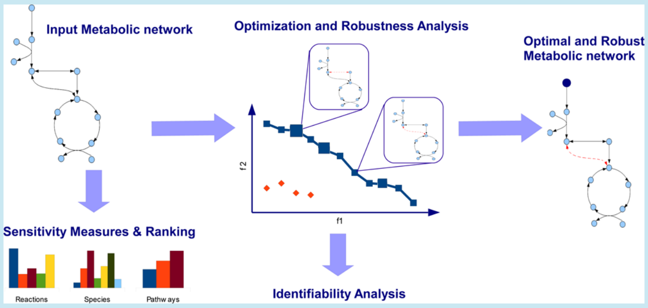
[‡]University of Rome "La Sapienza", S. Andrea Hospital, and Department of Biological Engineering, Massachussets Institute of Technology, United States

[¶]Computer Laboratory, University of Cambridge, U.K.

[§]Department of Biomedical Engineering, Johns Hopkins University, United States

[∥]Department of Evolutionary Biology, University of Florence, Italy

Ⓢ *Supporting Information*


Input Metabolic network
Optimization and Robustness Analysis
Optimal and Robust Metabolic network
f2
f1
Sensitivity Measures & Ranking
Reactions
Species
Pathways
Identifiability Analysis

ABSTRACT: In this work, we develop methodologies for analyzing and cross comparing metabolic models. We investigate three important metabolic networks to discuss the complexity of biological organization of organisms, modeling, and system properties. In particular, we analyze these metabolic networks because of their biotechnological and basic science importance: the photosynthetic carbon metabolism in a general leaf, the *Rhodobacter spheroides* bacterium, and the *Chlamydomonas reinhardtii* alga. We adopt single- and multi-objective optimization algorithms to maximize the CO_2 uptake rate and the production of metabolites of industrial interest or for ecological purposes. We focus both on the level of genes (e.g., finding genetic manipulations to increase the production of one or more metabolites) and on finding concentration enzymes for improving the CO_2 consumption. We find that *R. spheroides* is able to absorb an amount of CO_2 until 57.452 mmol h^{-1} gDW^{-1}, while *C. reinhardtii* obtains a maximum of 6.7331. We report that the Pareto front analysis proves extremely useful to compare different organisms, as well as providing the possibility to investigate them with the same framework. By using the sensitivity and robustness analysis, our framework identifies the most sensitive and fragile components of the biological systems we take into account, allowing us to compare their models. We adopt the identifiability analysis to detect functional relations among enzymes; we observe that RuBisCO, GAPDH, and FBPase belong to the same functional group, as suggested also by the sensitivity analysis.

KEYWORDS: *metabolic CAD, synthetic biology, multi-objective optimization, sensitivity analysis, robustness analysis, photosynthetic yield*

Developing models to simulate and predict the dynamic responses of metabolic networks has always been a challenging aim of systems biology. This goal is reached through the analysis of the main pathways involved in the metabolism of an organism. In particular, photosynthetic organisms perform many important functions for the planet, e.g., absorbing atmospheric CO_2, harvesting solar energy, and generating O_2 instead of processing oxygen.

In this paper, we focus our studies on the investigation of three metabolic networks: (i) the photosynthetic carbon metabolism pathway[1] of a generic leaf and (ii) *Rhodobacter spheroides*[2] and (iii) the light-driven algal metabolism of *Chlamydomonas reinhardtii*.[3] In Supporting Information, we report also other experiments and analysis with respect to the chloroplast starch degradation pathway[4] and the aspartate-derived amino acid pathway from plants.[5]

Carbon metabolism is a process that takes place in chloroplasts, which are organelles present in the cells of plants and eukaryotic algae and represent the site of the photosynthesis. The energy from light is captured by chlorophyll pigments and is converted into chemical energy (ATP and NADH). Chloroplasts produce glucose from sunlight energy. The glucose then transfers to the mitochondrion for aerobic respiration. The function of chloroplasts is basically to make food through the photosynthesis, i.e., by trapping light energy to convert water and carbon dioxide to form oxygen and glucose. During the photosynthesis, carbon is used for growth and some excess carbon can be fixed and stored in compact

Special Issue: IWBDA 2012

Received: October 1, 2012

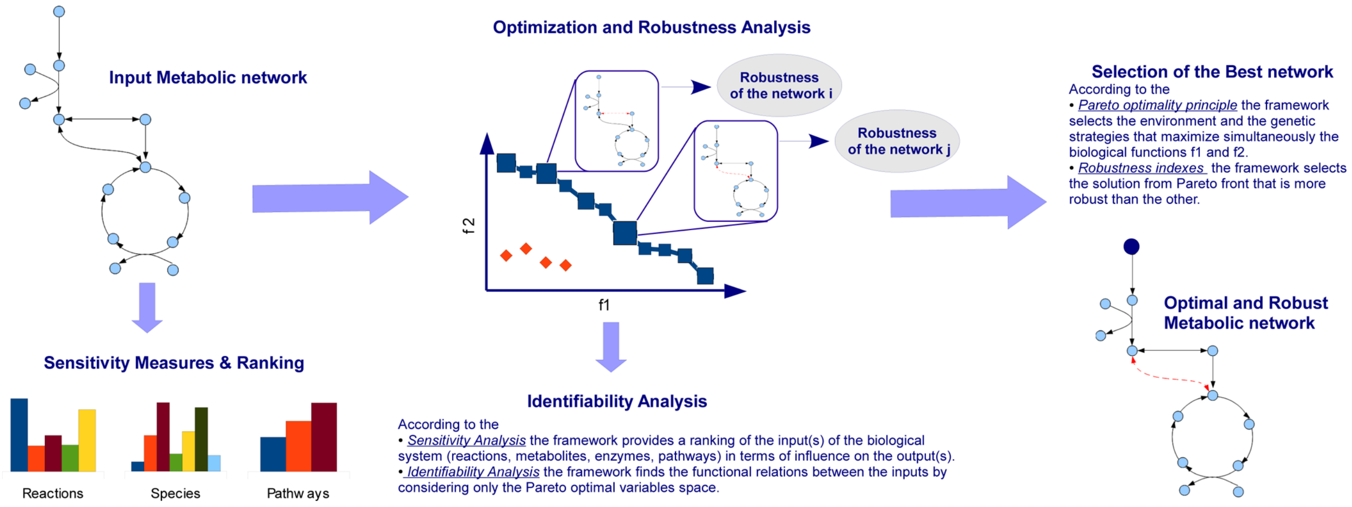

Input Metabolic network
Sensitivity Measures & Ranking
Reactions
Species
Pathw ays
Optimization and Robustness Analysis
Robustness of the network i
Robustness of the network j
f2
f1
Identifiability Analysis
According to the
• Sensitivity Analysis the framework provides a ranking of the input(s) of the biological system (reactions, metabolites, enzymes, pathways) in terms of influence on the output(s).
• Identifiability Analysis the framework finds the functional relations between the inputs by considering only the Pareto optimal variables space.
Selection of the Best network
According to the
• Pareto optimality principle the framework selects the environment and the genetic strategies that maximize simultaneously the biological functions f1 and f2.
• Robustness indexes the framework selects the solution from Pareto front that is more robust than the other.
Optimal and Robust Metabolic network

Figure 1. Flowchart of the framework here presented. The framework is applicable to each metabolic network (modeled with flux balance analysis with/without the gene-protein-reaction mappings or by using ordinary differential equations, algebraic differential equations, or partial differential equations). In the first step, the metabolic network (input) is analyzed according to the sensitivity, in order to rank the components of the network (reactions, species, or pathways). This first step does not change the conformation of the network but only evaluates the importance of its components in the model. In the optimization procedure (single- or multi-objective), the algorithm optimizes the decision variables of the model to maximize or minimize one objective or more (in the flowchart, two objectives f_1, f_2 are being optimized). Decision variables can be (i) the concentration values of the enzymes, (ii) the knockout genes set, or (iii) the uptake rates. The method tunes the variables in an appropriate way (described in the text) for optimizing the objectives (chosen by the users), such as the photosynthetic yield and the biomass of the organism. The result of the multi-objective optimization is the Pareto front (blue points). Each point of the front represents a particular conformation of the network. By investigating the variables space of the Pareto front, our method calculates the identifiability relationships, i.e., the functional relations between the decision variables. The Pareto optimality is also coupled with the robustness analysis. For each Pareto point, we evaluate the global and local robustness (in the flowchart, the size of the Pareto point represents the robustness). The output of the method is therefore a modified network.

polymers such as starch. The latter is stored in the form of granules made up of both linear and branched polymers of glucose.[6]

The *R. sphaeroides* system by Imam et al.[2] models all of the most interesting features of photosynthesis, as well as the metabolic capabilities of this kind of organisms. Interestingly, when the *R. sphaeroides* lacks oxygen intake, it can process light energy through a photosynthetic electron transport chain, whose features are similar to those found in plants.[7] Moreover, during its photosynthetic growth, *R. sphaeroides* uses either CO_2 as the sole carbon source or other organic carbon sources in order to grow autotrophically or heterotrophically. The autotrophic metabolism of *R. sphaeroides* makes it a potential organism for use in the synthesis of chemicals or polymers that can serve as raw materials in the production of biofuels or as a means of sequestering atmospheric or industrially produced CO_2. Therefore, a comprehensive analysis of the *R. sphaeroides* model may prove very useful for the understanding of both the lifestyle and the mechanisms underlying transitions between these different metabolic states.

The model of the *Chlamydomonas reinhardtii* metabolism[3] allows the investigation of photosynthesis in algae and in particular of light regulation. The advantages of this model and its optimization are evident from the perspective of biofuel production. The organisms and pathways above-mentioned cover an important task by using the photosynthetic process. Increasing the ability of these organisms to consume CO_2 can be very interesting. For this reason, in this paper we investigate the photosynthesis process and optimize it.

In Figure 1, we report the pipeline of the computational analysis techniques. The computational method takes a metabolic network as input, which can be modeled by using ordinary differential equations (ODEs), differential algebraic equations (DAEs), or partial differential equations (PDEs) or by using the flux balance analysis framework (FBA) containing or not the gene protein reaction (GPR) mappings. In this way, the method is able to menage different mathematical models and, as described in the following, to perform different analyses. Additionally, the framework here presented can be used in different areas, as in electronic design automation, making it suitable for general purposes. The method performs single- and multi-objective optimizations to reach desired targets. The aim is to find the values of the decision variables in order to obtain one (single-objective) or more (multi-objectives) phenotypes. In this work, we consider different components of the biological model. For instance, we can choose to (i) perturb enzyme concentrations, (ii) turn off genes, or (iii) search for the optimal nutrients (uptake rates). Enzyme concentrations, genes, and nutrients represent the decision variables of the optimization problems. For instance, through a multi-objective approach, we seek the genetic manipulations (decision variables) to maximize simultaneously the CO_2 consumption and the biomass formation (two objective functions) of *R. sphaeroides*. In this case, we focus on the genetic level by performing gene knockouts.

The results of the multi-objective optimization is not a single solution (as in a single optimization problem), but a set of non-dominated points, which constitute the Pareto front (the blue points of the central plot of Figure 1). A point is called non-dominated if there are no points that outperform it in all the objective functions. Instead, the dominated points (represented in red in the central plot of Figure 1) are feasible points but are less good with respect to the blue points of the graph and therefore not optimal. Pareto optimality is useful to obtain not only a vast range of Pareto optimal solutions but also the best trade-off design. In addition, the shape of the front and the number of Pareto solutions give an idea of the behavior of the metabolic network/organism with respect to a particular phenotype optimization.

Furthermore, for each Pareto optimal solution, our method calculates the robustness indexes. The robustness analysis is performed to evaluate the fragility of the biological system when it undergoes small perturbations, which could be endogen (for instance, small perturbation of the enzymatic concentrations) or hexogen (for instance, small perturbation of the external environment). The introduction of robustness in the analysis should hence result in more reliable and realistic targets for biotechnology.

Our computational analysis framework is extended also with sensitivity and identifiability analyses. The former ranks the components of the biological systems (metabolites, enzymes, pathways, and genes) in terms of sensitivity and is an independent step with respect to optimization and robustness. The elements that have a large influence on the outputs of the systems are considered sensitive. In this work, we couple sensitivity with optimization, since we restrict the set of the decision variables considering only the most sensitive parameters of the model. In particular, in the carbon metabolism network we want to maximize CO_2 uptake rate in several experiments. In a first step we optimize 25 enzyme concentrations. Second, by using the sensitivity analysis, we consider only the most sensitive parameters/enzymes. Indeed, we choose to optimize (i) the first 11 more sensitive enzymes and (ii) the first 6 more sensitive enzymes. The other enzymes are maintained constant. Details are reported in the following section and in Supporting Information.

The identifiability analysis finds functional relations among enzymes, by analyzing the values of the decision variables after and before the optimization. The output of the method is one or more metabolic networks. In Figure 1, the output is chosen from the Pareto front according to trade-off and robustness values. In the graph, the vertices are the metabolites involved in the network, and the edges are the relationships between metabolites, in this case the biochemical reactions and the transport reaction. An edge represents a reversible reaction, while an oriented edge represents an irreversible reaction. The dashed line represents a modified reaction, for instance, an intervention in the regulatory factors of the reaction or in the gene knockout array. The blue color of the vertex represents a change in the uptake rate, i.e., a change of a nutrient for the organism.

In Table 1, we show the advantages and limitations linked to the methods and models used in our approach. In particular, by using the FBA and GPR framework, it is not possible to predict metabolite concentrations, since this method does not use kinetic parameters. However, it can determine fluxes at steady state. Additionally, FBA does not account for regulatory effects such as activation of enzymes or regulation of gene expression. Therefore, its predictions may not always be accurate. However, since FBA does not require kinetic parameters, it can be computed very quickly even for large networks. This makes it well suited to studies that characterize many different perturbations such as different substrates or genetic manipulations.[8]

Table 1. Advantages and Limitations of the Methods and Models Discussed in This Work

features	ODEs-DAEs-PDEs	FBA
kinetic parameters	considered	not considered
regulatory effects	modeled	not modeled
metabolite concentrations	prediction allowable	steady state
accuracy	not always accurate	not always accurate
simulation time	long	short
size	small network	large network
precision	low	good

On the other hand, by using ODEs-DAEs-PDEs, the time to solve the system increases, though the metabolic system is not large and the precision depends on the computational solver. Instead, FBA uses a linear programming approach to find the solution of the problem, and therefore the solution is equal using also different libraries (glpk, Gurobi Optimizer, LINDO Systems, etc.). The advantage of ODEs-DAEs-PDEs models lies in the use of kinetic parameters, allowing to investigate several features, such as the regulatory effect, the variation on time of the metabolite concentrations and, in some cases, thermodynamic constraints.

Remarkably, our general sensitivity- and robustness-based framework allows a detailed understanding and comparison of the roles played by each component in the models taken into account. Our approach could be easily adopted also for other biological light-response phenomena and other organisms. The main goal of this work is proposing a pipeline for model-based *in silico* design based on the state-of-the-art multi-objective optimization approaches.

RESULTS

Photosynthetic Carbon Metabolism. The concept of robustness is extremely pervasive in nature and seems to be one of the driving force of evolution;[9] moreover, the ability of a system to preserve its behavior, despite internal or external perturbations, is a crucial design principle for any biological and synthetic system.[9−12] Applying the concept of robustness to the Calvin cycle and to the pathways involved in photosynthesis process allows our method to calculate the limits of enzymes perturbation at which the system property of interest (a given level of CO_2 uptake) is maintained. The estimation of the robustness of *in silico* designed pathways has been performed using the methodology proposed by Stracquadanio et al.[9] A Monte Carlo algorithm applies a Gaussian noise to the enzyme concentrations and then estimates the variation of CO_2 uptake. A robust system is characterized by small fluctuations of its quantitative behavior under investigation, which means that a robust pathway will ensure the same uptake rate even if the enzyme concentrations differ from the nominal values.

Although it is possible to perform *in silico* design and verification of a biological system, it is still impracticable to edit long regions of a genome in an arbitrary way; the intrinsic structure of the genetic information introduces a number of constraints that must hold in order not to decrease the fitness of a living organism. From this point of view, it is extremely important to focus the design on a set of restricted significant parameters, in order to decrease the complexity of a biological implementation. However, identifying a set of crucial genes encoding for important enzymes is an open problem. The sensitivity analysis tries to correlate the uncertainty in the output of a model with the uncertainty in the input; it is important to note that while the robustness analysis performs a local estimation of the output variation in a limited input range, the sensitivity analysis aims to study the output variation at a global level by investigating all the parameter space.[13] The

Table 2. Photosynthetic Carbon Metabolism Results[a]

enzyme name	initial concn mg N m^{-1} (the natural leaf)	optimal concn of 11 sensitive enz. mg N m^{-1}	optimal and robust concn mg N m^{-1}	optimal and robust conc mg N m^{-1} (3 fixed enz)	optimal but not robust concn mg N m^{-1}
RuBisCO	517.00 (100)	784.27 (**84.5**)	860.226 (100)	856.44 (100)	861.93 (**39**)
PGA kinase	12.20 (100)	4.66 (100)	3.989 (100)	3.63 (100)	3.98 (**0**)
GAP DH	68.80 (100)	69.03 (**81.5**)	64.483 (100)	65.08 (100)	63.55 (**17**)
FBP aldolase	6.42 (100)	10.40 (100)	9.050 (100)	10.86 (100)	9.29 (**30.5**)
FBPase	25.50 (100)	29.44 (100)	26.889 (100)	32.24 (100)	27.03 (**0**)
Transketolase	34.90 (100)	*34.90* (100)	8.247 (100)	16.93 (100)	16.98 *(100)*
SBP aldolase	6.21 (100)	5.55 (100)	6.661 (100)	5.75 (100)	5.94 (**0**)
SBPase	1.29 (100)	4.70 (100)	4.397 (100)	4.43 (100)	4.31 (**1**)
PRK	7.64 (100)	7.04 (100)	7.007 (100)	6.38 (100)	7.99 (**22.5**)
ADPGPP	0.49 (100)	2.12 (100)	0.721 (100)	5.09 (100)	1.22 (**0**)
PGCA Pase	85.20 (100)	0.95 (100)	0.325 (100)	0.20 (100)	0.00 (**0**)
Glycerate kinase	6.36 (100)	*6.36* (100)	0.005 (100)	0.00 (100)	0.00 *(100)*
Glycolate oxidase	4.77 (100)	*4.77* (100)	0.019 (100)	0.16 (100)	0.00 *(100)*
GSAT	17.30 (100)	*17.30* (100)	0.027 (100)	0.00 (100)	0.00 *(100)*
Glycer. dehyd.	2.64 (100)	*2.64* (100)	0.003 (100)	0.00 (100)	0.00 *(100)*
GGAT	21.80 (100)	*21.80* (100)	0.00005 (100)	0.00 (100)	0.00 *(100)*
GDC	179.00 (100)	0.02 (100)	0.00003 (100)	0.00 (100)	0.00 *(100)*
Cyt. FBP ald.	0.57 (100)	*0.57* (100)	2.127 (100)	*0.57* (100)	2.03 (**0.5**)
Cyt. FBPase	2.24 (100)	*2.24* (100)	5.554 (100)	*2.24* (100)	5.27 (**30.5**)
UDPGPP	0.07 (100)	*0.07* (100)	0.531 (100)	*0.07* (100)	0.50 (**0**)
SPS	0.20 (100)	*0.20* (100)	0.034 (100)	0.01 (100)	0.03 (**30.5**)
SPP	0.13 (100)	*0.13* (100)	0.031 (100)	0.01 (100)	0.03 (**0**)
F26BPase	0.02 (100)	*0.02* (100)	0.00 (100)	0.00 (100)	0.00 *(100)*
CO_2 uptake (μmol)/(m^2s)	15.486	33.317	36.382	36.197	36.495
(local R %, global R %)	(100, 81.80)	(81.5, 78.3)	(100, 97.2)	(100, 92.6)	(0, 39.18)

[a]Concentrations of the enzymes, individual robustness, CO_2 uptake rate (at $c_i = 270$ μmol mol^{-1}, reflecting current CO_2 atmospheric concentration), and global and local robustness values. The second column reports the touchstone concentrations used in our simulations: the initial/natural leaf (modeled by Zhu et al.[1]). The third column reports the results of the optimization in which only the 11 sensitive enzymes are altered, while all of the others are kept at their nominal values. The fourth column reports the best-known leaf design, in terms of CO_2 uptake and robustness. The fifth column reports the results of a simulation where the enzymes cytosolic FBP aldolase, cytosolic FBPase, and UDP-Glc pyrophosphorylase have been maintained to their initial values. The last column reports the most efficient known point in terms of CO_2 but corresponds to a highly instable solution.

sensitivity analysis of the model has been performed using the Morris method.[14]

The model and the chosen algorithms make it possible to find the optimized concentration of the enzymes in order to obtain the highest increase in CO_2 uptake, keeping constant the total amount of protein nitrogen. The parallel optimization algorithm (PAO) makes it possible to identify solutions consisting of an optimized set of enzyme concentrations capable of reaching a theoretical CO_2 uptake rate of 36.382 μmol m^{-2} s^{-1} at a level of carbonate ions (c_i) of 270 μmol mol^{-1}. The CO_2 uptake at the initial enzyme concentrations was 15.486 μmol m^{-2} s^{-1} (Table 2 of the Supporting Information). Six enzymes are found to be particularly enhanced in their final optimized concentration: cytosolic FBP aldolase, cytosolic Fru-1,6-bisphosphatase (FBPase), UDP-Glc pyrophosphorylase (UDPGP), SBPase, RuBisCO, and ADPGPP (Figure 5 of the Supporting Information). The method obtains a theoretical CO_2 uptake increase corresponding to 134% with respect to the initial enzymes concentration.

The perturbation of parameters (concentrations) allows to understand the level of sensitivity of each of the considered enzymes involved in CO_2 fixation. Eleven enzymes are found to be sensitive, and two of them fragile (Table 3 of the Supporting Information). The four key enzymes relative to CO_2 uptake were RuBisCO, Fru-1,6-bisphophate (FBP) aldolase, sedoheptulosebisphosphatase (SPBase), and ADP-Glc pyrophosphorylase (ADPGPP) (Table 3 of the Supporting Information).

Since biotechnological techniques are currently incapable of treating many enzymes at the same time, we simulate the effect of perturbing six enzymes only (RuBisCO, FBP aldolase, SBPase, ADPGPP, phosphoglycolate phosphatase, and Gly decarboxylase (GDC)) while the remaining nineteen enzymes are maintained at their initial concentrations. For this set of six enzymes, we define the following constraint: the concentration must be ≥0.02 mg N m^{-1}. RuBisCO, FBP aldolase, SBPase, and ADPGPP are overexpressed, while phosphoglycolate phosphatase and GDC are almost switched off. Nitrogen is kept constant. This configuration obtained a CO_2 uptake rate of 32.828 μmol m^{-2} s^{-1} (that is, only 3.492 μmol m^{-2} s^{-1} less than the best solution), perturbing only six enzymes (Figure 9 of the Supporting Information).

A still more refined analysis used always the same set of enzymes but allowing RuBisCO to increase up to a maximum of 15% with respect to the initial values. FBP aldolase, SBPase, and particularly ADPGPP were overexpressed, while phosphoglycolate phosphatase was switched off and GDC was kept close to its initial value. This configuration obtained a CO_2 uptake rate of 31.819 μmol m^{-2} s^{-1} (that is, only 4.501 μmol m^{-2} s^{-1} less than the best solution); the results are shown in Figure 10 of the Supporting Information. A further simulation

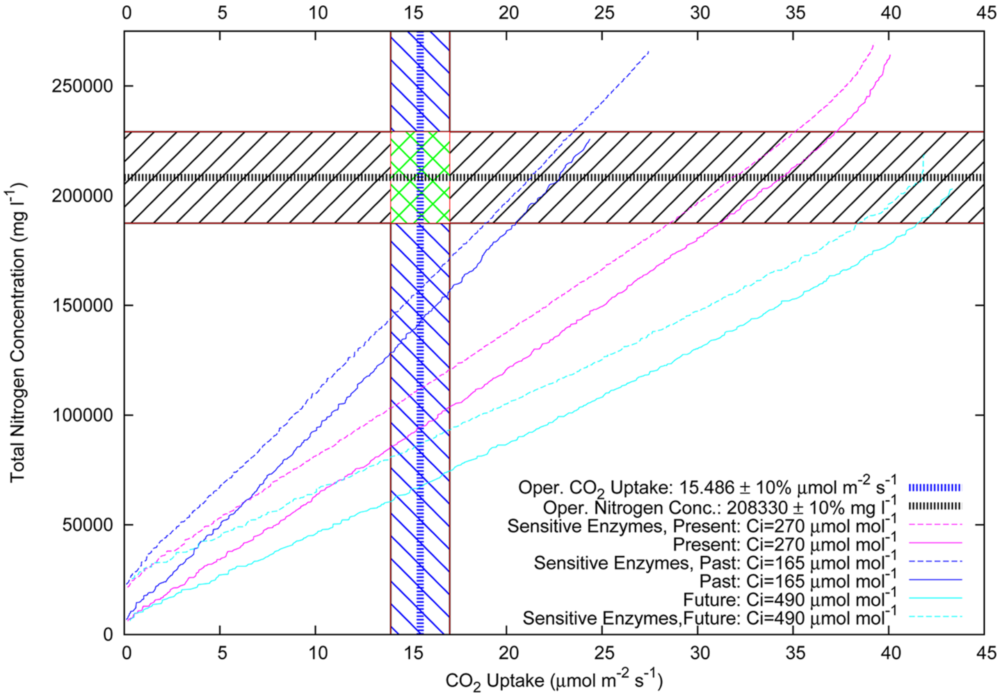


Figure 2. Photosynthetic carbon metabolism results. CO_2 Uptake and protein-nitrogen concentration trade-off. Maximizing the CO_2 uptake while minimizing the total amount of protein-nitrogen concentration; the operative area of natural leaves is located in the green checked area. The label "Sensitive Enzymes" indicates the multi-objective optimization using the 11 most sensitive enzymes of the model. The three resulting Pareto fronts have been dominated by the multi-objective optimization over all the enzymes of the model. This trade-off search has been carried out for the three c_i concentrations referring to the environmental conditions of 25 million years ago, nowadays, and in 2100.

attempted an optimization of the CO_2 uptake rate perturbing four enzymes only (FBP aldolase, SBPase, PGCAPase, and GDC) while the remaining 21 enzymes were maintained at their initial concentrations. This configuration (Figure 11 of the Supporting Information) obtained a CO_2 uptake rate of 22.4202 μmol m^{-2} s^{-1} with respect to the initial concentration of about 16 μmol m^{-2} s^{-1}. In a combinatorial approach, we performed another optimization with a different set of four enzymes, which were FBP aldolase, ADPGPP, PGCAPase, and GDC. This configuration obtained a CO_2 uptake rate of 20.626 μmol m^{-2} s^{-1} (Figure 12 of the Supporting Information).

The results (maximization of CO_2 uptake, while minimizing nitrogen necessity) summarized in Table 3 of the Supporting Information, showed that 11 enzymes are found to be sensitive (Figure 3 of the Supporting Information) and two of them fragile. There are four key enzymes relative to the CO_2 uptake: RuBisCO, FBP aldolase, SPBase, and ADPGPP. Six of the sensitive enzymes were coincident with the known light-controlled enzymes of the cycle. Both of the fragile enzymes were light-controlled. A first conclusion is that the most sensitive enzymes are key enzymes that can strongly influence the CO_2 uptake with slight concentration variation. The fact that these enzymes are mostly light-controlled confirms the strict control of light availability on the Calvin cycle. Highly and moderately sensitive enzymes found by Sun et al.,[15] on the basis of microarrays expression patterns, largely correspond with those indicated in our analysis, with the exception of tranketolase (moderately sensitive according to ref 15 and at low sensitivity in our analysis). The proposed solution has a high level of robustness (last column of Table 3 of the Supporting Information). GAPDH and PRK did not vary much their concentration during the optimization analysis. This result would fit well with the fact that the expression of these two enzymes is controlled by light, while specific chloroplast proteins as CP12 are capable of controlling their activity forming with them a complex PRK/GAPDH/CP12 with high molecular weight.[16] Such refined control appears to be appropriate for sensitive enzymes, and similar controls may be widespread for sensitive enzymes *in vivo*.

The simple finding of an optimal solution with ideal concentrations could be not a sufficient task, since the transcription process of the enzymes genes and other control systems linked to changing environmental conditions and/or feedbacks coming from other biochemical pathways could vary the enzymes concentration or their activity with time. Moreover, biotechnological insertion of new promoters sequence is not able to produce an exact and foreseen amount of transcripts. Therefore it is clear that it is important to estimate how well the achieved CO_2 uptake is preserved under perturbation at the enzyme concentration level. Robustness can be defined as the persistence of a system property with respect to perturbations.[9] Such property can be assessed in our analysis and can be fundamental to foresee the effect of a biotechnological genetic modification. The results of this analysis are shown in Table 2 and in Table 2 of the Supporting Information.

The application of the PAO algorithm to the ODEs system for optimizing the enzyme concentration in order to maximize CO_2 uptake maintaining nitrogen constant showed that six enzymes were particularly enhanced: cytosolic FBP aldolase, cytosolic FBPase, UDPGPP, SBPase, RuBisCO, and ADPGPP (Figure 5 of the Supporting Information). An increase in theoretical CO_2 fixation rates obtained by varying the enzyme concentrations of the Calvin cycle starting from the current experimentally determined values was shown already by various authors as Zhu et al.[1] and Stracquadanio et al.[13] The PAO algorithm allowed to obtain a theoretical CO_2 uptake increase corresponding to 134% with respect to the initial enzymes concentration. This result is even higher with respect to Zhu et al.[1] solutions based on an evolutionary algorithm, which leads to an increase of 76% (from 16 to 28 μmol m^{-2} s^{-1}).

The analysis based on the evaluation of the nitrogen limitation effect (shown in Figure 2) showed that the minimal amount of nitrogen allowed still a CO_2 uptake rate of 5.7. Such an amount could be taken into consideration as an assessment of the biomass growth limit of plants living in nitrogen limitations. Some crops are known to grow better than other that live in condition of low nitrogen supply. The better performance of rye compared to that of wheat was attributed to specific root length,[17] but our model may suggest that even the Calvin cycle enzyme concentrations may be better adapted for nitrogen limitation with respect to wheat. As a matter of fact, the latter shows higher growth rate without nitrogen limitation,[17] suggesting that it may be better adapted to high nitrogen level. The second result was shown as a result of the Pareto optimality analysis, which should lead to the closest-to-ideal solution: in this case the CO_2 uptake rate is 21.213 (Figure 6 of the Supporting Information) and hence higher with respect to the average land plant leaf with starting enzyme concentrations. This value represents a theoretical limit for biotechnological targets leading to maximizing productivity with the minimum amount of nitrogen supply, which is the value close to the economical optimum. It is interesting to observe that, even in this case, the total CO_2 uptake was over 30% higher with respect to the natural CO_2 uptake rate at the natural enzyme concentrations. The calculation of the maximum CO_2 uptake rate at different atmospheric CO_2 concentrations (Figure 7 of the Supporting Information) showed that the main difference between the current CO_2 atmospheric concentration and that of the past regarded the optimization of ADPGPP, PGA kinase, and HPR reductase, all much higher in the optimization at lower CO_2 concentration, and GCEA kinase, SPS and F26BPase, all much higher in the situation of high CO_2. These three last enzymes were not among the sensitive enzymes in the optimization at the current atmospheric CO_2 concentration. The results indicated that changing atmospheric conditions, particularly with respect to CO_2 amount, would produce very different evolutionary pressure on the enzymes. Concentration enhancement or reduction would affect one or the other enzymes, depending on the environmental conditions (at least relative to CO_2). Another important biotechnological target is to check the possible increase of CO_2 uptake leaving RuBisCO constant. This limitation is appropriate: given that RuBisCO is the most abundant protein in nature, it has been considered also a nitrogen reservoir for plant metabolism.[18,19] For instance, in an experiment on the haptophyte alga *Isochrysis galbana* on the effects of nitrogen limitation, as cells became more nitrogen-limited, the fraction of total cell nitrogen contained in RuBisCO decreased from 21.3% to 6.7%, whereas that of the light-harvesting complex remained relatively constant. That means that RuBisCO quantity is not only linked to the CO_2 uptake, but has a secondary function as nitrogen storage. Moreover, after some studies, the enzyme might already be naturally optimized under an evolutionary point of view.[20] Hence, further optimization of RuBisCO may prove difficult and lead to only marginal improvements.[21] Therefore, it is quite unlikely that models allowing free further increase of RuBisCO concentration, would be really feasible. The optimization of CO_2 uptake rate perturbing 24 enzymes leaving RuBisCO at its initial concentration (as shown in Figure 8 of the Supporting Information) leads to a theoretical optimized uptake rate of 22.2698 μmol m^{-2} s^{-1} with respect to the initial 15−16 μmol m^{-2} s^{-1} of the natural leaf. The most influential enzyme in this analysis was ADPGPP showing a very high increase in concentration.

Identifiability Analysis for Carbon Metabolism Model. In Figure 3, we show the functional relations among RuBisCO, GAPDH, and FBPase detected by the identifiability analysis applied to GAPDH.

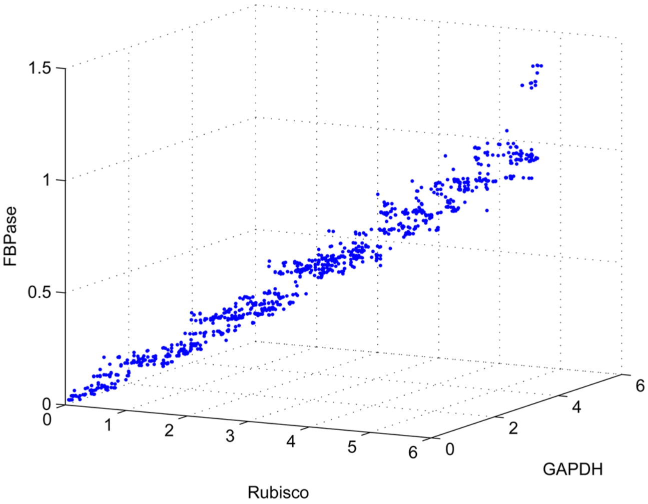


Figure 3. Plot showing the functional relation among the three decision variables RuBisCO, GAPDH, and FBPase, thus highlighting the structural non-identifiability of these variables. This group has been detected for the GAPDH enzyme.

It is noteworthy that, according to Table 1 in Supporting Information, RuBisCO belongs to the same functional group except for the presence of x_5 (FBPase). Indeed, Figure 4 shows that the optimal transformation β found for x_5 is different from the transformations found for x_1 and x_3, although the IA applied to GAPDH has assigned x_5 to the same functional group of x_1 and x_3. This can happen when the variables taken into account are also practically non-identifiable (which is the case of these three enzymes, since their *cv* is high).

The interdependent decision variables, which are non-identifiable, may be fixed at an arbitrary value in order to improve identifiability. Since the variables functionally related to the fixed variable change accordingly, the model's dynamical properties are not changed or restricted by the fixation.

Chlamydomonas reinhardtii. In order to analyze the photosynthetic capability of *C. reinhardtii*, a multi-objective optimization has been performed. Instead of the concentration values optimized in the carbon metabolism discussed in the previous paragraph, in this case we considered the genes as decision variables to optimize, and in particular their presence or absence in the metabolic network. The gene knockout strategy is represented as a binary vector y, where the l-th element is 1 if the l-th gene set is turned off, and 0 otherwise. Hence, the optimization problem consists of finding the optimal string of bit y*, which represents the optimal genetic strategy. Therefore, this is a combinatorial optimization problem.

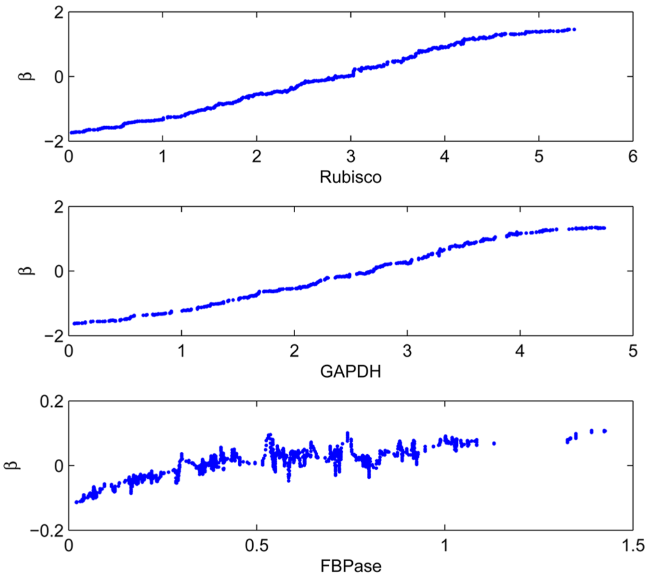


Figure 4. Plot showing the optimal transformations β (y axis) found for the three decision variables RuBisCO, GAPDH, and FBPase (x axis). Although FBPase has been once assigned to the same functional group of RuBisCO and GAPDH, it shows a slightly different and noisier behavior.

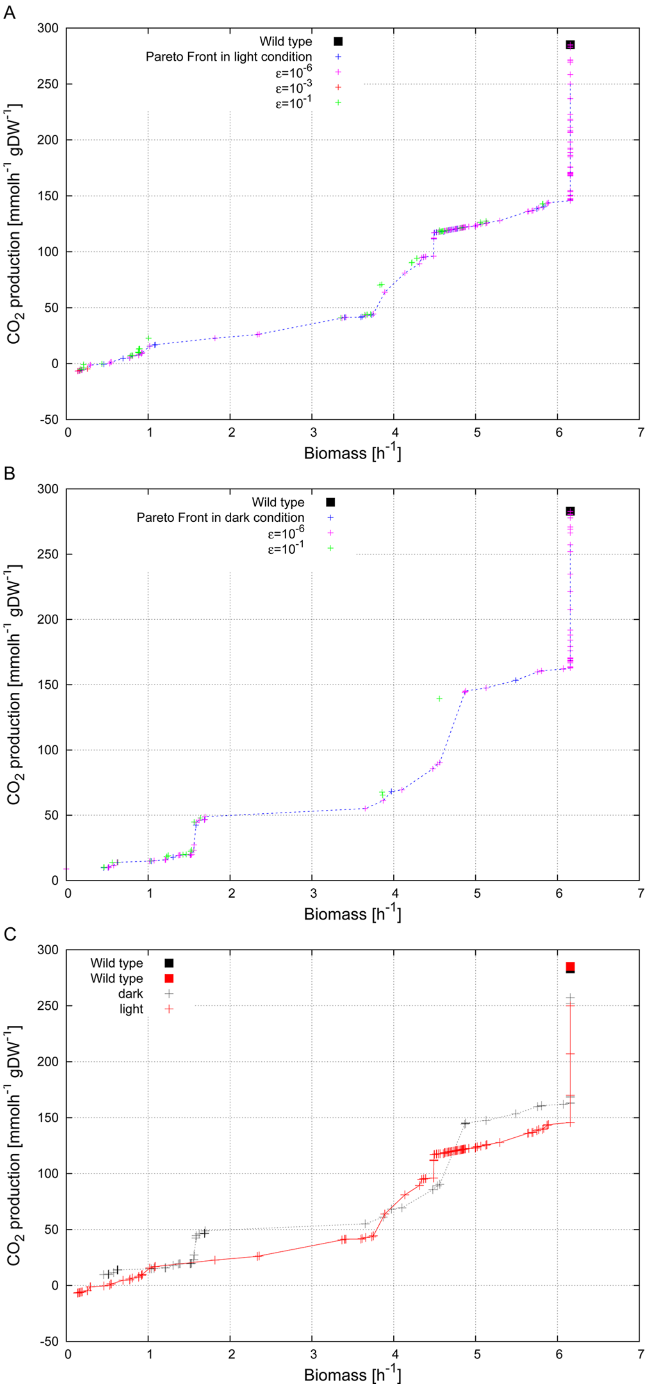


Figure 5. Maximization of CO_2 consumption and biomass formation in *C. reinhardtii* in light and dark condition and ε-dominance analysis. Starting from the points sampled during the optimization routine, we show in A and B the Pareto front in blue dots. Then, if we apply the ε-dominance condition choosing $\varepsilon = 10^{-6}$, the analysis obtains a set that contains both the blue dots and the purple ones. Similarly if we apply the ε-dominance condition choosing $\varepsilon = 10^{-3}$, the analysis obtains a set that contains the blue dots, the purple dots, and the red dots. Therefore new solutions are found if the dominance condition is relaxed. Panel C shows a comparison between the Pareto fronts of the two conditions. For CO_2 consumption, we indicate a negative value of production. By "maximizing CO_2 consumption", we indicate the minimization of CO_2 production.

The model of *C. reinhardtii* is represented by using the FBA framework. We set the maximum number of knockout allowed equal to 10. We used both light and dark conditions. In Chang et al.,[3] 11 windows of light spectrum can be chosen. Here, we used *solar lithosphere spectrum*, which is the result of a composite analysis from several measurements taken from different locations under cloudless conditions in the 48 contiguous U.S. states and multiple data normalization procedures. The other environmental conditions are set to the values that can be found in the Supporting Information of the related work.[3] In the first case, we chose to maximize the CO_2 consumption and the autotrophic biomass, and then we performed the ε-dominance analysis. Figure 5 shows the results. In these conditions, the maximum CO_2 consumption is equal to 6.7331 mmol h^{-1} gDW^{-1} with a biomass formation equals to 0.1381 h^{-1} (Figure 5 C, red points). The organism is not able to absorb CO_2 from the atmosphere in dark conditions; indeed the CO_2 production values are positive (Figure 5 C, black points), so the organism produces CO_2. The first two panels (A,B) show the Pareto fronts and the related ε-dominance analysis in light and dark conditions, respectively. The ε-dominance analysis is a relaxed condition of dominance to select the Pareto optimal points observed by the optimization algorithm. In fact, if we consider the non-relaxed condition of dominance, some interesting solutions may be discarded although dominated by a small amount. The results reported in Figure 5A,B confirms this hypothesis. The blue points belong to the Pareto optimal points obtained from a non-relaxed condition of dominance. With a relaxed condition, other acceptable solutions are added (purple, red, and green points).

Furthermore, Table 4 of the Supporting Information presents the robustness analysis results. We perturb the upper and lower bounds of the metabolic fluxes. In particular, in the global robustness (GR) the perturbation is carried out simultaneously for all fluxes (rates of the reactions) of the network to evaluate the fragility of the complete organism with respect to the metrics that are, in this case, the two objective functions. In the local robustness (LR), the perturbation is carried out for each flux (so, we have a robustness index for

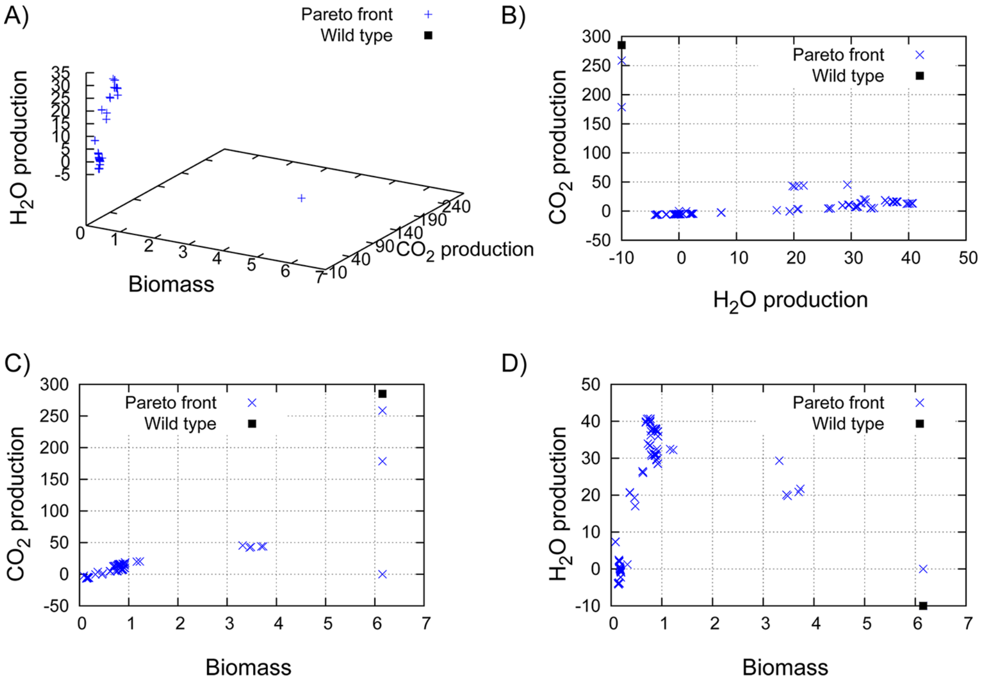


Figure 6. Simultaneous maximization of CO_2 consumption, biomass formation, and minimization of H_2O production in *C. reinhardtii*. We considered the photoautotrophic condition using the combinatorial optimization for searching gene knockout strategies. In panel A we show the Pareto front (blue points) obtained by the three-multi-objective optimization. The results for each pair of two objective functions are shown in panel B, C, and D. Points in black indicate the amount of CO_2, H_2O production, and biomass in wild type, i.e., without gene knockouts. For CO_2 or H_2O consumption, we indicate a negative value of production.

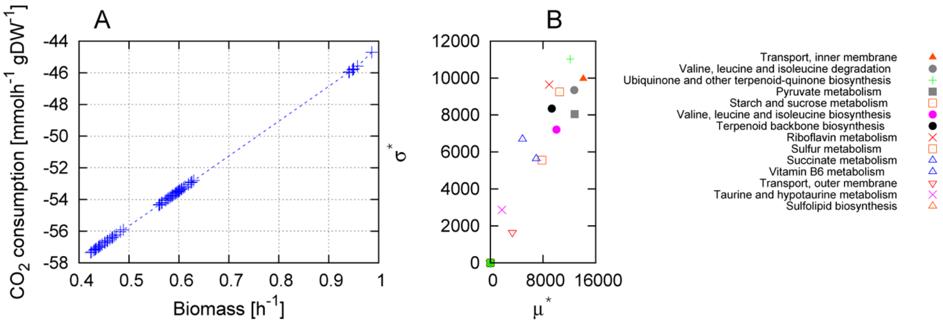


Figure 7. Results obtained by sensitivity and optimization for *R. spheroides*. (A) Pareto front obtained maximizing biomass formation and CO_2 consumption in *R. spheroides* using a multi-objective optimization to search for genetic knockout strategies in photoautotrophic conditions. For CO_2 consumption, we indicate a negative value of production. When we maximize CO_2 consumption, we indicate the minimization of the production. (B) Pathway-oriented sensitivity analysis for *R. spheroides*. The model includes 63 pathways, and only 14 pathways have sensitivity indexes greater than zero.

each flux with respect to the two metrics). We select from the Pareto front only one strain (one non-dominated solution) and compare it to the wild type (without knockout). We choose the strain that maximizes the CO_2 consumption and minimizes biomass formation. The results are shown in Table 4 of the Supporting Information. The strains are less robust than the wild types if we consider the global index, which indicates the robustness of the whole organism (GR). If we consider the robustness index for each flux, the strain has one minimum, while the wild types in light and dark conditions have three and two minima, respectively, and the related fluxes are the pyruvate transport by free diffusion (chloroplast), the nitrate exchange and the ammonia exchange in light condition, and the nitrate exchange and the ammonia exchange in dark condition. In the same environmental conditions, but only in light condition, in Figure 6 we choose to maximize the CO_2 consumption and the autotrophic biomass, simultaneously minimizing the H_2O. In this case, the algorithm finds a maximum CO_2 consumption equal to 5.6268 mmol h^{-1} gDW^{-1} with a biomass formation equal to 0.195 h^{-1} and the H_2O consumption equal to 9.8360 mmol h^{-1} gDW^{-1}. A more interesting result is the good trade-off between the maximization of CO_2 and the minimization of H_2O consumption. In this case, a CO_2 consumption equal to 5.4024 mmol h^{-1} gDW^{-1} has been obtained with a biomass formation equal to 0.179 h^{-1}, while the H_2O consumption is equal to 0.5455 mmol h^{-1} gDW^{-1}. Furthermore, we perform the robustness analysis considering the three metrics (the three objective functions). The results are reported in Table 5 of the Supporting Information. We choose the strain that obtains a good trade-off between CO_2 consumption maximization and the H_2O consumption minimization and compare it to the wild type. The results are similar to those of Table 4 of the

Supporting Information, so adding H_2O consumption does not cause variation in the global and local robustness.

***Rhodobacter spheroides*.** In order to maximize the CO_2 consumption and biomass formation in *R. spheroides*, we used our methods to find the best knockout strategies with the minimum knockout cost. We considered the photoautotrophic condition, i.e., with a poor environment, where the only carbon source is CO_2. The exchange allowable fluxes are sulfate, phosphate, ammonia, CO_2, magnesium, hydrogen, nicotinate, and photon (light). Figure 7 shows the results of the multi-objective optimization. In wild type, *R. spheroides* grows with a biomass rate equal to 0.986 h^{-1} and absorbs CO_2 until 44.705 mmol h^{-1} gDW^{-1}. We show that *R. spheroides* is able to absorb CO_2 until 57.452 mmol h^{-1} gDW^{-1}, but while reducing its growth to 0.418 h^{-1}, with a knockout cost equal to 14. The strain that reports a biomass of 0.9861 h^{-1} and 44.7048 mmol h^{-1} gDW^{-1} represents the trade-off design, with a knockout cost equal to 8, turning off the following genesets: RSP2138, RSP0361 or RSP2252, RSP0359, RSP0829, RSP3330 or RSP0656, RSP3142. In this configuration, six reactions are deleted: fumarate hydratase, L-serine ammonia-lyase, ribose-5-phosphate isomerase A, lactate dehydrogenase, sodium/sulfate symporter and acetate via Na+ symport. We performed a large set of simulations and optimizations for *R. spheroides* in various photoautotrophic conditions. We optimize (i) biomass versus H_2O production, (ii) biomass versus O_2 production, and (iii) biomass versus ethanol production. For all these experiments, the multi-objective optimization has identified only a Pareto solution very close to the wild type solutions. This means that in photoautotrophic conditions the organism uses a metabolic pathway that is essential for its growth, and knockout genes are not feasible. We found H_2O production of 184.589 mmol h^{-1} gDW^{-1} with a biomass formation of 0.986 h^{-1}, and O_2 production of 1.2265×10^{13} mmol h^{-1} gDW^{-1} and biomass 0.0099 h^{-1}. Conversely, in photoautotrophic conditions, *R. spheroides* does not produce ethanol, and even if we turn off genes, the result is always equal to zero. This means that ethanol is completely consumed in the metabolic network of the organism during its growth.

In Figure 7B, we report the results of the pathway-oriented sensitivity analysis (PoSA) and the pathways that have sensitivity indexes greater than zero. Only 14 pathways (out of 63 pathways) are found to be sensitive, probably because in photoautotrophic conditions only the genes in these pathways have influence on the growth and metabolism of *R. spheroides*.

Furthermore, we present in Table 8 of the Supporting Information the robustness analysis results. Similarly as in *C. reinhardtii*, the method acts by perturbing the upper and lower bounds of the metabolic fluxes and calculating both the global and local robustness. We select two strains from the Pareto front and compare them with the wild type. We choose the strains with good trade-off between the maximization of biomass formation and CO_2 consumption and the one that maximize the CO_2 consumption.

Discussion. In this work, we develop methodologies for analyzing and cross comparing metabolic models. We analyze in particular three metabolic networks (but our study is extended with other two little pathways described and discussed in Supporting Information) because of their biotechnological and basic science importance. We adopt single and multi-objective optimization algorithms and we focus both on finding optimal knockout strategies or concentration enzymes for biotechnological or basic science purposes. The Pareto optimality analysis is a useful tool for simulating biochemical pathways when contrasting objectives have to be considered simultaneously. By using this analysis, we find that *R. spheroides* is able to absorb an amount of CO_2 to 57.452 mmol h^{-1} gDW^{-1} with a knockout cost equal to 14, deleting six reactions, while *C. reinhardtii* obtains a CO_2 consumption value equal to only 6.7331 (Figure 13 in Supporting Information). The application of this analysis to the Calvin cycle provided the best solution for the maximization of the CO_2 uptake rate and the minimization of the total nitrogen. The analysis was also used in order to understand which enzymes are the most important in CO_2 uptake rate and those whose modification is more robust, that is, less prone to concentration fluctuation. This objective is a fundamental biotechnological target, since it is not possible to engineer all the enzymes levels simultaneously and it is not currently possible even to work on transcription promoters so finely to obtain a completely definite final enzyme concentration *in vivo*. The finding of a limited number of targets (enzymes) sufficiently robust to obtain a working solution even in the case of concentration fluctuations could lead to modified organisms whose activity could be better predicted. Furthermore, the optimization made it possible to analyze the scenario foreseen for the end of the century, when the atmospheric CO_2 will be much higher than nowadays, with an estimated C_i of 490 μmol mol^{-1}. This simulation was carried out considering a case with minimal nitrogen availability and that with highest CO_2 uptake. Such simulation could foresee the response of the photosynthetic organisms to the increase in CO_2 concentration and the increase on agriculture productivity even with lower amount of available nitrogen.

By using the sensitivity and robustness analyses, we identify the most sensitive and fragile components of the biological systems we take into account, allowing us to compare their models. In *R. spheroides* we show that only 14 pathways are sensitive, probably because in photoautotrophic conditions only the genes in these pathways have influence on the growth and metabolism. In *C. reinhardtii* alga, the method finds that a flux perturbation of the reactions pyruvate transport by free diffusion (chloroplast), nitrate exchange, or to ammonia exchange highlights the fragility of the organism with respect to the metrics chosen (CO_2 and biomass formation). The same behavior is shown by the *R, spheroides* with respect to the ammonia or hydrogen exchange reactions.

In order to group enzymes according to functional relations, we applied the identifiability analysis (IA) to the chloroplast model. This approach allows detection of structural non-identifiability, i.e., some components of the model that cannot be determined unambiguously. The IA showed that RuBisCO, GAPDH, and FBPase belong to the same functional group, i.e., they are interdependent decision variables. Interestingly, this bears out the results of the sensitivity analysis, which positioned these three enzymes in the most sensitive group of enzymes for the maximization of CO_2 consumption and the minimization of nitrogen consumption. The results are reported in the Table 1 of the Supporting Information.

Role of Organelles in the Photosynthesis. Photosynthesis is an organelle orchestra. A hypothetical scenario of evolution from the ancestral bacterium to the chloroplast is likely to include engulfments between bacteria. The engulfed bacteria becomes a membrane-bound organelle specialized in a specific task, and thus the host is able to specialize in all the other functions. For instance, when an eukaryote engulfs a bacterium and starts the process to convert it into a chloroplast,

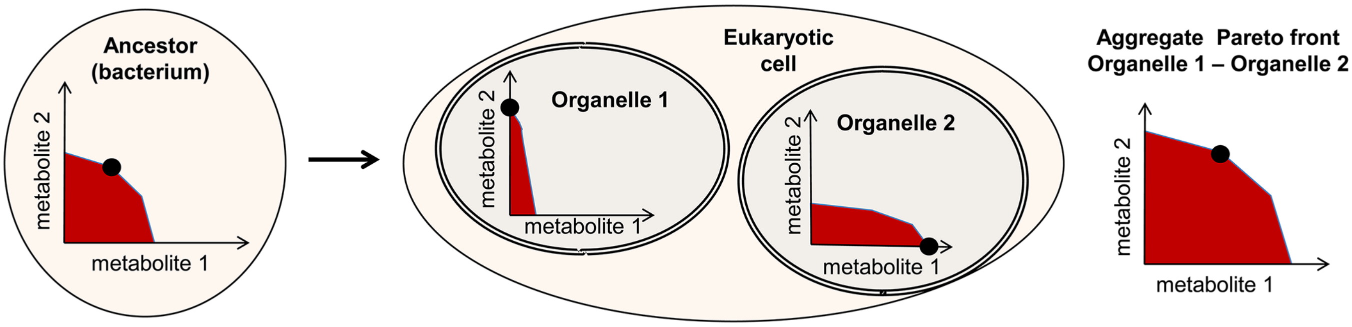


Figure 8. In a compartmentalized cell, each organelle contains all of the reactions devoted to a specific function, and thus each organelle can specialize in producing a metabolite. Since two organelles can specialize in producing metabolite m_1 and m_2, respectively, the overall Pareto front of the cell exhibits a larger area than before the engulfments.

gradually it loses some of its apparatus and transfers them to the chloroplast. Hence, the chloroplast specializes in producing O_2 from CO_2, while the bacterium becomes a cytoplasm that handles the metabolites by the chloroplast and specializes in all the other functions. Reactions share metabolites, thus avoiding competition by compartmentalizing an organism allows the multi-objective optimization of the whole organism. This process reduces the complexity, but preserves the overall behavior.

More than one engulfment may have happened. In a compartmentalized cell, each organelle contains all of the reactions devoted to a specific function and thus can be thought of as a large-scale pathway. Significantly, compartments cannot live without one another, therefore we expect a loss of global robustness. Indeed, in order for the entire system to work properly, exchange reactions are needed for transporting metabolites among organelles, and any disruption of these key reactions can affect more than one organelle and therefore the entire cell.

The evolution toward a system of organelles compartmentalizing the overall set of reaction can be studied in terms of Pareto front and model order reduction. The key steps to perform a reduction of a model are the Pareto optimality, the sensitivity, and the robustness analyses, because they indicate the least important parts of the organism, which can often be fixed as constants rather than included as variables in the simulations performed through the model. In the evolution process, during the optimization, i.e., the maximization of a metabolite, the Pareto front of an organism undergoes both expansion and contraction phases. Furthermore, robustness and sensitivity are functions of the conditions because of their pathway structure are different.

Remarkably, the evolution of a Pareto front can highlight the benefits of an engulfment. The Pareto optimal point in the maximization of two metabolites m_1 and m_2 before the engulfments can be outperformed by the Pareto optimal point obtained in the eukaryotic cell by considering both the organelles responsible for their production (i.e., by merging their Pareto fronts in a new common front). This improvement is due to the fact that the two organelles can specialize in producing metabolite m_1 and m_2, respectively (see Figure 8).

In the compartmentalization, the contribution of each organelle is not only to maximize energy production but also to coevolve with the other cell structures, so as to ensure the maximum fitness of the cell. In this regard, although a large area under the Pareto front suggests versatility, it could also lead to the production of metabolites not directly relevant for the goal of the organism.

In order to investigate the cooperation between two or more organelles of the same type, we plan to look at the exchange of signals between organelles as if they are oscillators. For instance, we can think of a chloroplast as an oscillator responsible for rhythmical activity between day and night, i.e., with and without sunlight. Likewise, the activity of other organelles (e.g., mitochondria) may change between day, meals, and nights. The sensitivity and the robustness of an enzyme in chloroplasts may therefore depend on the light condition. In this regard, one should consider the oscillation in the yield of each organelle and the influence that each organelle has on the others. Through influence signals, the population of organelles may reach a global synchrony, depending on the relation among natural frequencies and the coupling strength.

The partition of pathways into compartmentalized structures as organelles has a tremendous cost: together with the reduction of the genome of the "guest", hundreds of proteins, previously coded by the guest's genome need to migrate to the host genome and acquire a signal that enables them to enter the guest environment(s). Clearly, this is not a one-step process, and we envisage intermediate states. The Pareto front analysis along with the sensitivity analysis may provide clues for this hypothesis.

Finally, there are some open questions that a Pareto front analysis can address: (i) Is the photosynthesis optimized if we chose to optimize all the compartments in the chloroplast? (ii) Is the number of compartments crucial to the optimization process?

■ METHODS

***In Silico* Design of Metabolic Pathways.** Our computational framework (Figure 1) performs three tasks closely linked and able to manage networks with different complexity and mathematical modeling. In a first step, sensitivity analysis (SA) is performed in order to rank input/parameters of the system in terms of sensitivity, i.e., according to the influence that the parameters have on the output(s) of the model. We implement three sensitivity methods: the Morris method,[14] the Sobol method,[22] and a new sensitivity analysis, named pathway-oriented sensitivity analysis (PoSA), based on knockouts permutation.

The second block of our framework is based on an optimization procedure. We can choose between a single-objective optimization or a multi-objective optimization, and between continuous or combinatorial optimization. The algorithms are based on the evolutionary concept, where the solutions are calculated, compared, and selected in each iteration/generation of the algorithm. The optimization algorithms in our framework are inspired by the non-dominated sorting genetic algorithm II (NSGA-II).[23] NSGA-II[23] is designed to ensure an efficient and effective approximation of the Pareto optimal set. Recently, the NSGA-II has been extended using an island-based model for parallel optimization; the new algorithm, called parallel algorithm for optimization (PAO),[24] performs parallel optimizations and swaps non-dominated solutions every given number of iterations. Decision variables have a continuous domain, and in our work, we consider the enzyme concentrations or the uptake rate of the metabolites that enter in biological systems.

Moreover, the optimization process of our framework performs also combinatorial optimization. For instance, in *C. reinhardtii* we maximize simultaneously the biomass formation and the CO_2 consumption, searching for the best gene knockout strategies. Knockouts are represented by means of a binary vector y, where the element y_l is equal to 1 when the l-th manipulation is turned off and equal to 0 when the l-th manipulation is turned on. Since decision variables assume only two values, we are dealing with a combinatorial optimization problem. The aim is to find the best gene manipulations with minimum knockout number (necessary for biotechnological purposes, which can be considered as another objective function).

In a multi-objective optimization problem, when the objective functions are in conflict with each other, the output is a set of non-dominated solutions, called the Pareto front. In multi-objective optimization, there is normally not a single solution optimal in all respects. Among the feasible solutions, the algorithm selects the non-dominated solutions, also called Pareto solutions. Vilfredo Pareto was an Italian economist, who for the first time introduced the concept of optimization for more objective functions. The non-dominated solutions are better than others because they are those for which an objective cannot be improved without worsening at least another objective. The Pareto optimal set is the set of all non-dominated solutions.[25] Pareto optimality proves very useful for biodesign automation, because it allows our method to obtain a wide range of optimal solutions and also the best trade-off design.

A multi-objective optimization algorithm is characterized by four main steps. In a first step, a starting population is initialized. A population is formed by a set M of individuals, each of which is represented by a decision variables set (whose values are chosen randomly or by the user) and the objective functions values obtained by using the corresponding decision variables. Decision variables are parameters of the system that we want to optimize. The value of the objective functions is strictly linked to the decision variables values. An individual represents a feasible solution. Once the first population is initialized, the algorithm enters an evolutionary loop. A new population is created and updated for each iteration of the algorithm. Each iteration, called also generation, has the aim to improve the solution set and optimize the decision variables values, incorporating the evolutionary concept of Darwin. According to Darwin, the individuals of the population evolve from generation to generation and only the best individuals survive. The same concept is incorporated in the evolutionary/genetic algorithm. By using the crossover and mutation operators new individuals are formed, and only the best individuals are selected and inherited. An individual is better than another if the latter is dominated with respect to the first one. The loop terminates when a maximum generation number is reached, or when a particular solution is found.

Parallel optimization algorithms (PAO) are algorithms (incorporated in our framework) that exploit coarse-grained parallelism to let a pool of solutions exchange promising candidate solutions in an archipelago fashion. Using evolutionary operators such as recombination, mutation, and selection, the framework completes with migration its approach based on islands. Each island is a virtual place where a pool of solutions is allowed to evolve with a specific optimization algorithm; communications among islands in terms of solutions evolved by potentially different algorithms are arranged through a chosen archipelago topology. The island model outlines an optimization environment in which different niches containing different populations are evolved by different algorithms and periodically some candidate solutions migrate in another niche to spread their building block. In this archipelago approach different topology choices can raise a completely different overall solution, introducing then another parameter that has to be chosen for each algorithm on each island. The PAO framework actually encloses two optimization algorithms and many archipelago topologies, but its simplest configuration has been used to obtain a comprehensible comparison with the other adopted strategies and to better understand the optimization capabilities of this approach. The adopted configuration has two islands with two optimization algorithms, the advanced CMA-ES algorithm (A-CMA-ES) and the differential evolution algorithm (DE),[26] which exchange candidate solutions every 200 generations with an all-to-all (broadcast) migration scheme at a 0.5 probability rate. Even in its simplest configuration, this approach has shown enhanced optimization capabilities and an optimal convergence. A-CMA-ES introduces a set of cutoff criteria to CMA-ES[27] and ensures with a constraint, a lower bound, for each enzyme concentration to be compatible with the smallest concentration observed in the natural leaf. In the case of biological networks modeled with ordinary differential equations (ODEs, e.g., for photosynthetic carbon metabolism[1]), the enzyme concentration values are optimized in each iteration/generation of PAO until a fixed number of generations is reached or until a particular solution is found. The models (described in the following) are implemented in Matlab, and the ODE set is solved through the Matlab function ode15s. In the case of biological networks solved with FBA (*R. spheroides* and *C. reinhardtii*), the optimal genetic manipulations are searched through GDMO. GDMO implements a new combinatorial mutation operator. Mutation represents a switch, from 0 to 1 or from 1 to 0. The process is randomly executed; for each parent individual we create 10 offspring, but only the best is chosen. Mutations can achieve the maximum knockout number equal to the parameter C (fixed at 50 by default). A new population of M individuals is formed selecting the best individuals from the parents of the previous generation and the current offspring. The new population undergoes a new round of evaluation.

Additionally, we introduce the concept of ε-dominance (inspired by Laumanns et al.[28]) that adds impressive insights into the Pareto front interpretation capabilities of metabolic networks and improves the diversity of solutions and the convergence of the algorithm. Once Pareto optimal solutions have been obtained, we consider all the dominated and non-dominated solutions of all generations, and we seek solutions that may have been discarded because they are dominated by a small ε that, for our purposes, can be considered negligible. In other words we apply a "relaxed" condition of dominance, thus building a new set of solutions.

In the robustness analysis step, the interesting optimal solutions obtained by the optimization are processed. Small perturbations are made in the new biological systems, and the fragility of the new networks is tested. The system is said to be robust if, after perturbations, the outputs do not change in a significant way. The robustness analysis aims to evaluate the probability of a system to retain a property under perturbations.

In order to test our computational framework, we choose three biological systems that have different complexity. Besides having distinct biological complexity and nature, they are modeled by using different mathematical methods. In Table 3 we report the mathematical modeling adopted for each systems and the number of reactions, metabolites, enzymes and genes contained in the associated model.

Table 3. Characteristics of the Metabolic Networks Analyzed in This Work[a]

	photosynthetic CM[1]	*R. sphaeroides*[2]	*C. reinhardtii*[3]
modeling	ODEs	FBA-GPR	FBA-GPR
reactions	39	1158	2190
metabolites	38	796	1068
enzymes	38	595	718
genes	n.a.	1095	1080
pathways	3	63	93
optimization	PAO − NSGA II	GDMO	GDMO
sensitivity	Morris[14] and Sobol[22]	PoSA	Morris[14]
robustness	GR/LR	GR/LR	GR/LR

[a] For each organism/pathway we report the mathematic approach used to simulate the behavior of the biological system. The first network (photosynthetic carbon metabolism (CM)[1]) is modeled by using a set of ordinary differential equations (ODEs), which represent the change in concentration of the metabolites involved. The last two organisms (*R. sphaeroides*[2] and *C. reinhardtii*[3]) are modeled by using the flux balance analysis (FBA) at steady state. We report the number of reactions, metabolites, enzymes/gene sets, genes and pathways, and also the algorithms used in our analysis. PAO: Parallel Advanced Optimization, GDMO: Genetic Design through Multi-objective Optimization, PoSA: Pathway-oriented Sensitivity Analysis, GR: Global Robustness, LR: local Robustness.

R. spheroides and *C. reinhardtii* genome-scale metabolic networks were investigated through FBA, which is a widely used approach for studying biochemical networks. These network reconstructions contain all of the known metabolic reactions in an organism and the genes that encode each enzyme. FBA calculates the flow of metabolites through this metabolic network, thereby making it possible to predict the growth rate of an organism or the rate of production of a desired metabolite.

Metabolic reactions are formally represented by a numerical matrix S of the stoichiometric coefficients of each reaction. Each reaction is represented by a flux v_j, $j = 1, ..., n$, through the network and is constrained by a lower and upper bound, which define the maximum and minimum allowable flux of the reaction. The genome-scale metabolic reconstruction contains also the gene-protein-reaction (GPR) mappings that provide the links between each gene and the reactions that depend on it. In particular, genes are represented with a Boolean relationship to distinguish between single and multifunctional enzymes, isoenzymes, enzyme complexes, enzyme subunits.

For a set of L genetic manipulations, the GPR mappings are represented by a $L \times n$ matrix G, where the (l,j)-th element is 1 if the l-th genetic manipulation maps onto the reaction j and is 0 otherwise. We used the approach implemented in OptKnock[29] to find the fluxes distribution in the metabolic network in order to optimize multiple objectives (multi-objective optimization), i.e., desired productions (synthetic objectives) and therefore achieve the maximal growth. The bilevel problem is represented by the following formulation:

$$\begin{aligned} &\max \quad g'v \\ &\text{such that} \quad \sum_{l=1}^{L} y_l \leq C \\ &\qquad y_l \in \{0, 1\} \\ &\qquad \max \quad f'v \\ &\qquad \text{such that} \quad Sv = 0 \\ &\qquad\qquad (1 - y)'G_j v_j^L \leq v_j \leq (1 - y)'G_j v_j^U, \\ &\qquad\qquad j = 1, ..., n \end{aligned} \tag{1}$$

where g is a vector of weights (n dimensional) associated with the synthetic objectives, and g' is its transpose. For example, when the synthetic objectives v_j and v_h have to be maximized, the weights g_j and g_h are equal to 1. y is the knockout vector (L dimensional). If there are no impaired reactions in the metabolic network, y contains only zeros. Conversely, when $y_l = 1$, the gene set involved in the manipulation l is turned off, and the corresponding reactions are in the absent status (the lower and upper bounds are set to 0, resulting in a modified metabolic network). C is an integer representing the maximum number of knockout allowed. f is a vector of weights (n dimensional) associated with the natural objectives. All the elements in f are either 0 or 1. For our purposes, f_i is equal to 1 if v_i is the biomass core. v_j^L and v_j^U are the lower and upper bound values (thermodynamic constraints) of the generic flux v_j. The bilevel problem can be converted to a MILP problem as described in Optknock. The method implements and solves the problem using the GLPK solver. Therefore, the objective functions in the multi-objective optimization problem are calculated by GDMO solving (1).

Sensitivity Analysis. In modeling, sensitivity analysis (SA) is a method used to detect the inputs playing a key role on the output of the model. SA indices have been recently adopted in systems biology by interrogating the reactions space (RoSA, reactions oriented sensitivity analysis) and the species space (SoSA, species oriented sensitivity analysis) to find their influence on the output of the system.[30] In this work, we perform SA to find the most sensitive inputs in FBA models using a novel pathway-oriented sensitivity analysis (PoSA). PoSA allows us to rank the genetic manipulations according to their influence on the output of the model. Unlike other sensitivity analysis methods applied in biological modeling, whose inputs (reactions or species) are real numbers, PoSA is applied when inputs are Boolean values. Indeed, each input of the model is represented through a set of binary variables.

In PoSA, the knockout vector y used to represent the genetic manipulations is partitioned in p subsets of bits $\{b_1,b_2, ..., b_s, ...,$

$b_p\}$. Each subset b_s includes the genetic manipulations linked to the reactions involved in the s-th metabolic functional pathway of the network. Each subset b_s has a cardinality W_s, where $W_s < L, \forall s = 1, ..., p$. We cluster in each subset all the genes that are involved in a single functional pathway, e.g., the citric acid cycle, oxidative phosphorylation, pentose phosphate pathway, and so on.

For the combinatorial problem described above, we defined the "elementary effect"[14] for the input b_s as $EE_s = ([f(b_1,b_2, ..., b_{s-1},\tilde{b}_s,b_{s+1}, ..., b_p) - f(\tilde{y})])/\Delta_s$, where $\tilde{b}_s$ is the mutation on the input b_s and consists of the switch of bits chosen randomly in b_s: if a bit is equal to 0 (or 1), the permutation turns it to 1 (or 0). Δ_s is a scale factor defined as $\Delta_s = (1/W_s)\Sigma_{i=1}^{Ws}\tilde{b}_s(i)$, $s = 1, ..., p$.

The output $f(y)$ considered in our analysis is the vector v of fluxes. $\tilde{y}$ is the mutation carried on the knockout vector y defined in the Boolean region of interest $\Omega = \{0,1\}^L = \{(y_1, ..., y_l, ..., y_L)|y_l \in \{0,1\}\}$. In Supporting Information has been reported the pseudocode of the PoSA method. The parameters β and K of PoSA establish, respectively, the allowed knockouts in the whole network and in the pathway b_s. In our analysis we choose $\beta = 0.1$ (default value), and we recommend to set $0.02 \leq \beta \leq 0.2$. K is selected by the user or set by default to 4.

The distribution of effects EE_s is obtained permuting y by randomly sampling KQ points from Ω and permuting b_s by randomly sampling KQN points from Ω. If the procedure is performed for each input, the result would be a random sample at a total cost of KQ for calculating $f(\tilde{y})$ and KQN for $f(b_1,b_2, ..., \tilde{b}_s, ..., b_p)$, with a total cost of $pKQ(N + 1)$ evaluates of function. The estimation of the mean μ^* and standard deviation σ^* is used as indicator of which inputs should be considered important. A large (absolute) central tendency for EE_s indicates an input with an important overall influence on the output. A large spread indicates an input whose influence is highly dependent on the values of the inputs.[14]

In this work, the computational analysis includes also Morris and Sobol methods that evaluate in a continuous space the parameters of the model. For a detailed description see the original work.[14,22]

Identifiability Analysis. Models of biological processes usually include components (e.g., parameters) that are determined by measuring data and fitting to experiments. A component for which no unique solution exists is called *non-identifiable*.

There are two different sources of non-identifiability: (i) structural non-identifiability, i.e., some components in the model may be functionally related and therefore they cannot be determined unambiguously; (ii) practical non-identifiability, caused by a low amount or quality of data that does not allow our method to precisely estimate the component. The identifiability analysis (IA) detects functionally related (and thus non-identifiable) parameters by fitting a model repeatedly to experimental data and analyzing parameter estimates.

Here we consider the chloroplast model by Zhu et al.[1] and the 25 decision variables of the C3 cycle, namely, the concentrations of its enzymes. We adopt the method proposed by Hengl et al.[31] to detect automatically structural identifiability consisting of functional relations between decision variables. These relations are detected by applying the alternating conditional expectation algorithm (ACE).[32]

Let $K = [v_1, ... v_m] \in \mathbb{R}^{n\times m}$ be the matrix of the n values for the m decision variables $\{x_1, ..., x_m\}$, where each column $v_i \in \mathbb{R}^n$ contains the n estimates for the i-th variable. Let us suppose that the variables are related by unknown linear or nonlinear functional relations. The true transformations that linearize these relations are denoted by α and β_j, namely, $\alpha(x_i) = \Sigma_{j\neq i}^{m}\beta_j(x_j) + \xi$, where ξ represents a Gaussian noise. The ACE algorithm[32] estimates the optimal transformations $\hat{\alpha}(x_i)$ and $\hat{\beta}_j(x_j)$, $j \neq i$, such that $\hat{\alpha}(x_i) = \Sigma_{j\neq i}^{m}\hat{\beta}_j(x_j)$ where x_i is the response and all the other variables are the predictors.

The process of repeating estimates in the matrix K is replaced by taking into account all the non-dominated points of the Pareto front. In other words, a single fitting sequence K is obtained by considering the entire front. Thus, the problem of identifiability analysis is mapped onto the problem of detecting groups of the functionally related decision variables that produce that Pareto front.

Specifically, the connection between the identifiability analysis and a constraint structure stems from the fact that a non-identifiable constraint involving decision variables causes them to be functionally related. In our case, the constraint is detected through 1903 estimates of all the 25 variables (enzymes). Each estimate corresponds to a non-dominated point of the Pareto front obtained to maximize the CO_2 uptake rate and minimize the nitrogen consumption.

We adopt the mean optimal transformation approach (MOTA)[31] by fixing at 5 the maximal number of parameters allowed to enclose a functional relation. The results are shown in Table 1 in Supporting Information. The "Groups" column indicates the functional relations between variables. For instance, RuBisCO and GAPDH are functionally related. In other words, the response variable x_1 is strongly related to the predictors x_3 and x_5. Conversely, the enzymes transketolase type 1 and SBPase do not have any functional relation with any other enzyme (Table 1 in Supporting Information). The r^2 column indicates how much variance of the response can be explained by the predictors. A high amount of variance of the response that can be explained by the predictors indicates a large effect of the fixation of the predictors on the standard deviations of the response. The $cv(x) = std(x)/mean(x)$ helps to distinguish practical identifiable from non-identifiable parameters.[31] In case of practical non-identifiability, the choice of the parameter to fix depends on the experiments and on reference values found in the literature.

Robustness Analysis. The basic principle of this analysis is taken from Nicosia and Stracquadanio's approach,[9] and explained in the following. First, we define the perturbation as a function $\tau = \gamma(\Psi,\sigma)$ where γ applies a stochastic noise σ to the system Ψ and generates a trial sample τ. The γ-function is called γ-perturbation. Without loss of generality, we assume that the noise is defined by a random distribution. In order to make a statistically meaningful calculation of robustness, we generate a set T of trial samples τ. Each element τ of the set T is considered robust to the perturbation, due to stochastic noise σ, for a given property (or metric) ϕ if the following condition is verified:

$$\rho(\Psi, \tau, \phi, \delta) = \begin{cases} 1, & \text{if } |\phi(\Psi) - \phi(\tau)| \leq \delta \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where Ψ is the reference system, ϕ is a metric (or property), τ is a trial sample of the set T, and δ is a robustness threshold. The definition of this condition makes no assumptions about the function ϕ. It can be anything (not necessarily related to properties or characteristics of the system); however, it is implicitly assumed that it is quantifiable. The robustness of a system Ψ is the number of robust trials of T, with respect to the property ϕ, over the total number of trials. It is a dimensionless quantity that states, in general, how the system is robust to perturbations. The robustness index is a function of δ, so the choice of this parameter is crucial. Since we are interested in the behavior of a system when subjected to small perturbations, and because the behavior is acceptable when the deviations from the original value is as small as possible, we choose the values of δ equal to 5% of the metric and sigma equal 10% of the perturbed variable. Starting from this principle, we evaluate two values of robustness, the global robustness value (GR) and the local robustness value (LR). In the first case, the perturbation is carried out simultaneously on all the input variables to evaluate the fragility of the system with respect to the metrics, while in the second case, the perturbation is carried out on one variable at time, and we obtain a robustness index for each variable.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: pl219@cam.ac.uk; nicosia@dmi.unict.it.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The work has been performed under the HPC-EUROPA2 project (Application number: 1220) with the support of the European Commission - Capacities Area - Research Infrastructures. The first author would like to thank all of the Staff at Edinburgh Parallel Computing Centre (EPCC), in particular, Dr. Catherine Inglis and Dr. Chris Johnson.

■ REFERENCES

(1) Zhu, X., De Sturler, E., and Long, S. (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.

(2) Imam, S., Yilmaz, S., Sohmen, U., Gorzalski, A., Reed, J., Noguera, D., and Donohue, T. (2011) iRsp1095: A genome-scale reconstruction of the Rhodobacter sphaeroides metabolic network. *BMC Syst. Biol. 5*, 116.

(3) Chang, R., Ghamsari, L., Manichaikul, A., Hom, E., Balaji, S., Fu, W., Shen, Y., Hao, T., Palsson, B., and Salehi-Ashtiani, K. (2011) Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism. *Mol. Syst. Biol.*, DOI: 10.1038/msb.2011.52.

(4) Nag, A., Lunacek, M., Graf, P., and Chang, C. (2011) Kinetic modeling and exploratory numerical simulation of chloroplastic starch degradation. *BMC Syst. Biol. 5*, 94.

(5) Curien, G.; Bastien, O.; Robert-Genthon, M.; Cornish-Bowden, A.; Cárdenas, M.; Dumas, R. Understanding the regulation of aspartate metabolism using a model based on measured kinetic parameters. *Mol. Syst. Biol.* 2009, *5*.

(6) Smith, A., Zeeman, S., and Smith, S. (2005) Starch degradation. *Annu. Rev. Plant Biol. 56*, 73−98.

(7) Hunter, C., Daldal, F., and Thurnauer, M. (2008) *The Purple Phototrophic Bacteria*, Vol. 28, Springer Verlag, New York.

(8) Orth, J. D., Thiele, I., and Palsson, B. (2010) What is flux balance analysis? *Nat. Biotechnol. 28*, 245−248.

(9) Stracquadanio, G., and Nicosia, G. (2011) Computational energy-based redesign of robust proteins. *Comput. Chem. Eng. 35*, 464−473.

(10) Kitano, H. (2004) Biological robustness. *Nat. Rev. Genet. 5*, 826−837.

(11) Kitano, H. Towards a theory of biological robustness. *Mol. Syst. Biol.* 2007, *3*.

(12) Jose, M., Hu, Y., Majumdar, R., and He, L. (2010) Rewiring for robustness, in *DAC '10, Proceedings of the 47th Design Automation Converence*, pp 469−474, Association for Computing Machinery, New York.

(13) Stracquadanio, G., Umeton, R., Papini, A., Lió, P., and Nicosia, G. (2010) Analysis and optimization of C3 photosynthetic carbon metabolism. *10th IEEE International Conference on Bioinformatics and Bioengineering 2010, BIBE 2010, Philadelphia, PA, May 31−June 3, 2010*, pp 45−51, IEEE Press, Washington, DC.

(14) Morris, M. (1991) Factorial sampling plans for preliminary computational experiments. *Technometrics 33*, 161−174.

(15) Sun, N., Ma, L., Pan, D., Zhao, H., and Deng, X. (2003) Evaluation of light regulatory potential of Calvin cycle steps based on large-scale gene expression profiling data. *Plant Mol. Biol. 53*, 467−478.

(16) Singh, P., Kaloudas, D., and Raines, C. (2008) Expression analysis of the Arabidopsis CP12 gene family suggests novel roles for these proteins in roots and floral tissues. *J. Exp. Bot. 59*, 3975−3985.

(17) Paponov, I., Lebedinskai, S., and Koshkin, E. (1999) Growth analysis of solution culture-grown winter rye, wheat and triticale at different relative rates of nitrogen supply. *Ann. Bot. 84*, 467−473.

(18) Chapin, F., Schulze, E., and Mooney, H. (1990) The ecology and economics of storage in plants. *Annu. Rev. Ecol. Syst.. 21*, 423−447.

(19) Millard, P. (2006) The accumulation and storage of nitrogen by herbaceous plants. *Plant, Cell Environ. 11*, 1−8.

(20) Bar-Even, A., Noor, E., Lewis, N., and Milo, R. (2010) Design and analysis of synthetic carbon fixation pathways. *Proc. Natl. Acad. Sci. U.S.A. 107*, 8889−8894.

(21) Raines, C. (2006) Transgenic approaches to manipulate the environmental responses of the C3 carbon fixation cycle. *Plant, Cell Environ 29*, 331−339.

(22) Sobol, I. (2001) Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. Comput. Simul. 55*, 271−280.

(23) Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput. 6*, 182−197.

(24) Umeton, R., Stracquadanio, G., Sorathiya, A., Liò, P., Papini, A., and Nicosia, G. (2011) Design of robust metabolic pathways. *48th Design Automation Conference - DAC 2011, San Diego, CA, June 5−10, 2011*, pp 747−752, ACM Press, New York, NY.

(25) Bligaard, T., Jóhannesson, G., Ruban, A., Skriver, H., Jacobsen, K., and Nørskov, J. (2003) Pareto-optimal alloys. *Appl. Phys. Lett. 83*, 4527−4529.

(26) Storn, R., and Price, K. (1997) Differential evolution−a simple and efficient heuristic for global optimization over continuous spaces. *J. Global Optim. 11*, 341−359.

(27) Hansen, N., and Ostermeier, A. (2001) Completely derandomized self-adaptation in evolution strategies. *Evol. Comput. 9*, 159−195.

(28) Laumanns, M., Thiele, L., Deb, K., and Zitzler, E. (2002) Combining convergence and diversity in evolutionary multiobjective optimization. *Evol. Comput. 10*, 263−282.

(29) Burgard, A., Pharkya, P., and Maranas, C. (2003) Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng. 84*, 647−657.

(30) Zhang, H.-X., and Goutsias, J. (2010) A comparison of approximation techniques for variance-based sensitivity analysis of biochemical reaction systems. *BMC Bioinform.*, DOI: 10.1186/1471-2105-11-246.

(31) Hengl, S., Kreutz, C., Timmer, J., and Maiwald, T. (2007) Data-based identifiability analysis of non-linear dynamical models. *Bioinformatics 23*, 2612−2618.

(32) Breiman, L., and Friedman, J. (1985) Estimating optimal transformations for multiple regression and correlation. *J. Am. Stat. Assoc. 80*, 580−598.