Discovering Essential Multiple Gene Effects Through Large Scale Optimization: An Application to Human Cancer Metabolism

Annalisa Occhipinti, Youssef Hamadi, Hillel Kugler, Christoph M. Wintersteiger, Boyan Yordanov, and Claudio Angione

Abstract—Computational modelling of metabolic processes has proven to be a useful approach to formulate our knowledge and improve our understanding of core biochemical systems that are crucial to maintaining cellular functions. Towards understanding the broader role of metabolism on cellular decision-making in health and disease conditions, it is important to integrate the study of metabolism with other core regulatory systems and omics within the cell, including gene expression patterns. After quantitatively integrating gene expression profiles with a genome-scale reconstruction of human metabolism, we propose a set of combinatorial methods to reverse engineer gene expression profiles and to find pairs and higher-order combinations of genetic modifications that simultaneously optimize multi-objective cellular goals. This enables us to suggest classes of transcriptomic profiles that are most suitable to achieve given metabolic phenotypes. We demonstrate how our techniques are able to compute beneficial, neutral or "toxic" combinations of gene expression levels. We test our methods on nine tissue-specific cancer models, comparing our outcomes with the corresponding normal cells, identifying genes as targets for potential therapies. Our methods open the way to a broad class of applications that require an understanding of the interplay among genotype, metabolism, and cellular behaviour, at scale.

Index Terms—Optimisation, genome-scale metabolic modelling, flux balance analysis, cancer metabolism, synthetic lethality

1 Introduction

Metabolism, the set of biochemical reactions that transform various compounds in living cells and organisms, is one of the core systems responsible for maintaining cellular functions. Metabolic models (reconstructions) of bacteria have been developed to facilitate the study and manipulation of biochemical processes [1], allowing the bioproduction of valuable compounds to be optimized through metabolic engineering [2]. The study of human metabolism, on the other hand, is becoming increasingly important for biomedical applications as an approach for understanding health and diseases. This is enabled by the availability of human metabolic reconstructions [3], [4], which integrate extensive metabolic information from various resources.

Achieving detailed kinetic modeling of metabolism is challenging and requires information about parameters that are hard to measure experimentally (e.g., kinetic rates and concentration of metabolites). Thus, Metabolic Flux Analysis (MFA) has emerged as a powerful methodology for estimating the fluxes (flow of material) through different reactions or pathways in large-scale metabolic networks, which provides an informative marker of metabolic behavior. A number of computational MFA techniques have been developed to predict these fluxes under various conditions (e.g., the availability of different nutrients) and genetic perturbations (e.g., mutations in genes associated with the catalysis of certain reactions) using metabolic models such as Recon [5]. In particular, Flux Balance Analysis (FBA) [1] reduces the problem of determining the metabotype (the fluxes through all reactions in the system) to a tractable linear program under the assumptions of steady-state and optimality. Due to its scalability and the informative results it generates, FBA is widely used, for example to predict growth phenotypes and adaptation in specific environmental conditions [6], [7].

Even with state-of-the-art metabolic models (e.g., Recon) and scalable computational techniques (e.g., FBA), a number of questions in metabolic systems biology remain open. In particular, recent evidence suggests that cells adjust their metabolism to optimize multiple (potentially conflicting) objectives and ensure flexible adaptation to changes in their external environments [8], [9]. However, FBA-based methods usually consider the optimization of a single objective (e.g., cellular growth), which is represented as a combined flux (e.g., biomass) or a linear combination of the fluxes through several reactions (e.g., a number of biosynthesis processes). Furthermore, Recon annotates different reactions

- *A. Occhipinti is with the Department of Computer Science and Information Systems, Teesside University, Middlesbrough TS1 3BX, U.K. E-mail: a.occhipinti@tees.ac.uk.*
- *Y. Hamadi is with the Uber Elevate, Paris, France. E-mail: youssefh@uber.com.*
- *H. Kugler is with the Faculty of Engineering, Bar-Ilan University, Ramat Gan, 5290002, Israel. E-mail: hillelk@biu.ac.il.*
- *C. M. Wintersteiger and B. Yordanov are with Microsoft Research, Cambridge CB1 2FB, U.K. E-mail: {cwinter, yordanov}@microsoft.com.*
- *C. Angione is with the Department of Computer Science and Information Systems, and with the Healthcare Innovation Centre, Teesside University, Middlesbrough TS1 3BX, U.K. E-mail: c.angione@tees.ac.uk.*

Manuscript received 16 June 2019; revised 4 Jan. 2020; accepted 2 Feb. 2020. Date of publication 2 Apr. 2020; date of current version 8 Dec. 2021. (Corresponding author: Claudio Angione.)
Digital Object Identifier no. 10.1109/TCBB.2020.2973386

with information about the genes involved in their catalysis, and these qualitative rules (e.g., a given reaction requires either gene A or gene B) can be used to study the quantitative effects of gene expression on metabolism. However, due to the large number of genes, metabolites and reactions modelled, addressing all possible combined effects of perturbation remains challenging. Finally, in both metabolic engineering and disease studies we often seek the gene expression profiles that could lead to some desired metabolic state. This amounts to solving the inverse problem of what is typically addressed through FBA, but computational methods suitable for such studies are still lacking [10].

In this paper, we propose an approach and a pipeline of methods that address the challenges outlined above. To study the trade-offs between several metabolic requirements, we combine Flux Balance Analysis (FBA) with multi-objective optimization without weighting and without combining the separate objectives into a single function. To capture the quantitative effect of gene expression on metabolism, we adjust the bounds on reaction fluxes in response to regulation. Combining these two strategies allows us to study the optimal metabolic states with respect to all the cellular objectives chosen. To address the inverse problem, we develop an evolutionary algorithm that allows us to explore the combinatorial, genome-wide transcriptomic space in order to identify expression profiles that lead to optimal metabolic states. We then apply clustering and statistical approaches to study the similarities between different optimal expression profiles. Compared to existing methods for linear Pareto-optimization of metabolic networks [11], [12], our method based on evolutionary algorithms is able to also capture concavity and discontinuity in the Pareto front, often present due to the nonlinearity of the metabolic network.

To gain further insights into the relationship between gene expression and metabolism, we then explore the combined effects of multiple-gene perturbations on the metabolic state. Previously, an FBA-based exhaustive computational exploration of single-gene knockouts has led to the identification of toxic genes that substantially decrease the biomass flux even when expressed only basally [13]. However, redundancies and latent associations in the metabolism might mask such toxicity, necessitating multiple-gene perturbation to reveal these combined effects. Here, using our combinatorial approach we enable the study of double or higher level gene perturbations, while preserving the quantitative relation between expression and metabolism. We show that, while computationally intensive, the exhaustive exploration of pairwise perturbations is still feasible even for large-scale models such as Recon. Yet, for three or more genes, such exhaustive approach becomes intractable (e.g., for the 2194 genes encoded as part of Recon 2, such an approach would require solving more than 1.75 billion linear programs). Therefore, we address this problem by proposing a method based on an incomplete algorithm.

Overall, our pipeline increases the predictive capability of Recon and allows assessing the potential of cell metabolism when pushed to optimize desired functions. Furthermore, it predicts at genome-scale the gene expression levels required to ensure pre-defined levels of those metabolic functions. The set of methods included in our pipeline are summarized in Fig. 1. The code is freely available as a MATLAB toolbox at https://github.com/claudioangione/PGA_and_C-EDGE.

Our methods can be used for mechanistic prediction of promising high-order sets of genes, thus potentially complementing the trial and error overexpression task usually performed by experimentalists. Additionally, our high-order study of the combined effect of many genes highlights, among all redundancies in the model, those that actually affect the biomass. As a side-effect, our method also predicts hidden genetic interactions, where the combined effect of two or more genes cannot be measured or predicted from the effects of the genes alone. Our multi-objective approach suggests potential changes in the expression profiles with the aim of changing the phenotype of a cell. Unlike the methods based on present/absent calls leading to on/off gene knockouts, it may be used as a prediction tool for the rapidly growing CRISPR-Cas9 techniques based on genome engineering for precise overexpression and partial knockdown [14], [15].

2 Methods

2.1 Flux Balance Analysis

For large biochemical networks, FBA-based approaches are often preferred to other mathematical modeling techniques (e.g. ordinary differential equations) as they do not require enzyme kinetic parameters and concentrations of metabolites in the system.

Let the network be composed of m metabolites with concentration x_i, $i = 1, \ldots, m$ and n reactions with flux rates v_j, $j = 1, \ldots, n$. Let $S \in \mathbb{R}^{m \times n}$ be the stoichiometric matrix (m rows and n columns). The balance that metabolite concentrations x_i must satisfy is $\dot{x_i} = \sum_{j=1}^{n} S_{ij} v_j, \quad i = 1, \ldots, m$, where S_{ij} is the stoichiometric coefficient of the ith metabolite in the jth reaction. Under steady-state conditions (homeostatic assumption) $\dot{x_i} = 0, \ \forall i \in \{\text{internal metabolites}\}$, we obtain a balance equation for every internal metabolite: $\sum_{j=1}^{n} S_{ij} v_j = 0$, or alternatively $Sv = 0$.

Each metabolite of the metabolic network is associated with a constraint, while the reaction rates v_j represent the variables, each of which is bounded by a minimum and maximum flux rate V_j^{min} and V_j^{max}. Since the matrix S is not square and $n > m$ with rows and columns linearly independent, there are more variables than constraints, and therefore a plurality of solutions. A final optimal distribution of fluxes, among all feasible steady-state flux distributions, is computed after an objective (i.e., a flux rate) or a linear combination of objectives is chosen to be maximized, therefore solving the maximization problem

$$\begin{aligned} \max \quad & u^{\intercal} v \\ \text{such that} \quad & Sv = 0 \\ & V_j^{min} \leq v_j \leq V_j^{max}, \quad i = 1, \ldots, n, \end{aligned} \tag{1}$$

where u is an n-dimensional array of coefficients defining the linear combination of flux rates selected to be maximized.

2.2 Bilevel Optimisation With Transcriptomics

In order to add transcriptomic information to an FBA model in a quantitative fashion, we model the effect of each gene expression profile as a change in the lower and upper bounds of the metabolic reactions, yielding a rerouted flux distribution

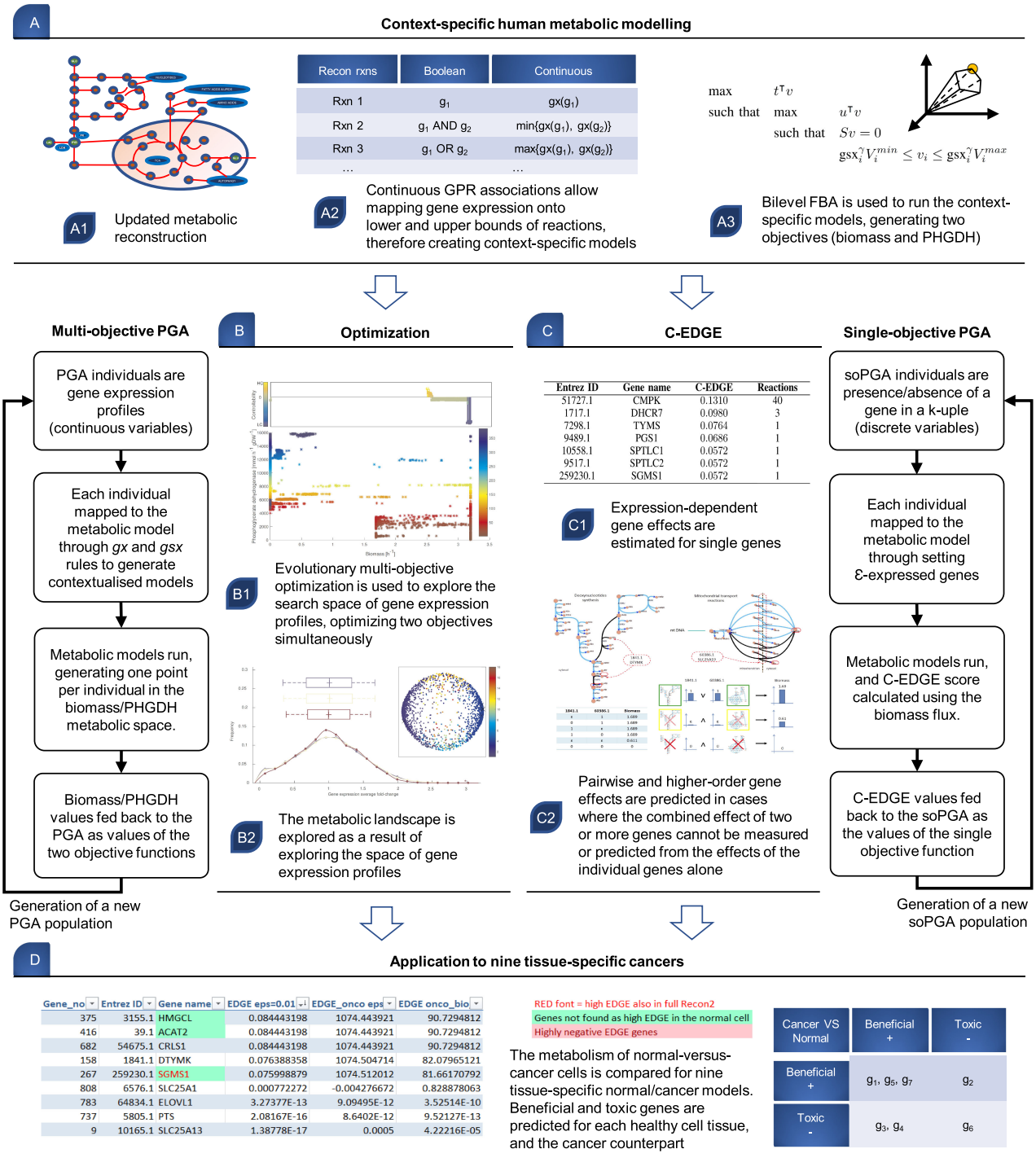


Fig. 1. *Analysis of simultaneous gene effects on human metabolism.* Starting from an augmented genome-scale human metabolic reconstruction (A1, see also the Methods section), we substitute Boolean gene-protein-reaction rules (GPR) associations with continuous associations (A2), therefore obtaining a model able to account for quantitative gene expression levels, associated with the phenotype through bilevel FBA (A3). Then, we develop a multi-objective parallel genetic algorithm (PGA) (B) to find the gene expression levels that simultaneously optimize the biomass and the phosphoglycerate dehydrogenase (PHGDH) reaction rate (B1). Further analyses highlight the difference in gene expression level between these scenarios of high, mid and low biomass (B2). Controllability analysis, multi-dimensional scaling and co-expression analysis are then executed on the optimal gene expression profiles. Independently, we propose a set-based sensitivity method, named C-EDGE (C), solved as a single-objective parallel genetic algorithm (soPGA), selecting k-uples of ε-expressed and KO genes, suggesting single genes (C1) and groups of genes (C2) with toxic or beneficial effect on the biomass. Finally, we apply our methods to the study of cancer metabolism (D). We assess cell-specific gene effects in cancer cells, and we use C-EDGE to compare nine tissue-specific normal and cancer cells.

across the network. Each enzymatic reaction is facilitated or impaired according to the enzyme abundance, which depends on the gene expression values. Although it often represents a subject of debate, this assumption in human metabolism is motivated by the recent evidence that, in mammals, the mRNA level is the main contributor to the overall protein expression level, with a good correlation between transcript level and protein abundance [16], [17]. Furthermore, in most normal and cancer cell lines, mRNA and protein levels have been found to be positively correlated [18].

More specifically, transcriptomic data is mapped to the Recon model using three maps with real-valued domain and range, defined by three rules that allow us to further constrain the model [19]. Each reaction in the model is controlled by a single enzyme, by two or more enzymes (enzymatic complex, represented by a Boolean AND relation), or by different but equivalent enzymes (isoenzymes, represented by a Boolean OR relation). We derive the gene set expression data using the following rules (applied recursively for gene sets where the AND/OR rules are nested):

$$\begin{aligned} \mathrm{gsx}(a) &= \mathrm{gx}(a), && \text{a single gene,} \\ \mathrm{gsx}(a \text{ AND } b) &= \min\{\mathrm{gx}(a), \mathrm{gx}(b)\}, && \text{a + b enzym. complex,} \\ \mathrm{gsx}(a \text{ OR } b) &= \max\{\mathrm{gx}(a), \mathrm{gx}(b)\}, && \text{a,b isoenzymes,} \end{aligned} \tag{2}$$

where gsx is the gene set expression value, and gx(a) and gx (b) are the gene expression values (expressed as fold changes) of two genes a and b, respectively. It is worth mentioning that, in our pipeline, gx are values generated by the PGA (see the next subsections).

Standard FBA only takes into account a single objective or a linear combination of objectives. However, it is now widely believed that a cell has to perform different, often conflicting tasks while ensuring a high growth rate [20], [21]. Trade-off have also been reported to limit the cellular optimization, e.g., as a result of evolution or adaptation to a new environment [22]. As proposed by Costanza *et al.* [23], a multi-objective approach is more realistic than considering only the assumption of maximum growth. Let m be the number of metabolites and n the number of reactions in the model. The stoichiometric matrix is $S \in \mathbb{R}^{m \times n}$ and $v \in \mathbb{R}^n$ is the array of flux rates. We solve the following two-level maximization problem, which allows us to associate an expression profile with an objective vector v of flux rates:

$$\begin{array}{lll} \max & t^\intercal v & \\ \text{such that} & \max & u^\intercal v \\ & \text{such that} & Sv = 0 \\ & & \mathrm{gsx}_i^\gamma V_i^{min} \leq v_i \leq \mathrm{gsx}_i^\gamma V_i^{max}, \\ & & i = 1, \ldots, n \end{array} \tag{3}$$

where u and t are n-dimensional arrays of weights associated respectively with the first and second optimization objectives, gsx_i is the gene set expression of the ith reaction in the model, derived from gene expression using Eq. (2). Through the parameter γ we enable modulation of the strength of the correlation between gene expression and reaction bounds. Changing this parameter, for values greater than 1, does not affect the distribution of gene expression found by the PGA (see Supplementary Information, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2020.2973386 for further discussion on γ). Since we couple this bilevel linear program with a multi-objective optimization algorithm, t and u select a single objective each. In accordance with the original metabolic model [24], the upper and lower bounds (V_i^{max} and V_i^{min}) could also assume negative values, e.g., when a reaction is reversible.

2.3 Genome-Scale Metabolic Model

We illustrate our methods through a study of human metabolism and by identifying cancer-associated pathways. We adopt a human metabolic model obtained by merging Recon 2 [3] (7440 reactions, 5063 metabolites and 2140 genes) with the model by Quek *et al.* [24] (7327 reactions, 4962 metabolites and 2169 genes). While the former is a highly curated but generic metabolic model, the latter provides a smaller and more specialized model for investigating the metabolism of human cell lines in culture, with 44 additional gene-protein-reaction (GPR) associations (listed as Supplementary Information S8 Table, available online).

By merging the two models we obtained a fully annotated human metabolic model with additional associations between genes and reactions (7440 reactions, 5063 metabolites and 2166 genes). Inconsistencies, duplications of metabolite names, charges, and annotations were corrected manually. We also replaced three reactions to satisfy stoichiometric balance, following PSAMM [25]. The list of the replaced reactions is available in Table S9, available online. The Matlab file of the resulting metabolic model is available as Supplementary Information, available online.

In the proposed model, we focus on biomass as the first (inner) objective (i.e., uv in Eq. (3)), and we study separately high and low biomass scenarios, corresponding to fast- and slow-growing cells. For both scenarios, we consider the phosphoglycerate dehydrogenase (PHGDH) as a second (outer) objective (i.e., tv in Eq. (3)). We select this objective as PHGDH is the enzyme catalyzing reactions in the serine synthesis pathway, together with phosphoserine aminotransferase (PSAT), and phosphoserine phosphatase (PSPH). Increased PHGDH flux, and in general serine synthesis pathway activity, was initially measured in mouse cancer when compared to normal tissues [26]. Increased serine metabolism has been a target of recent research attention as one of the main biomarkers of cancer, and inhibition of PHGDH has been reported to block cancer proliferation [27].

2.4 Multi-Objective Optimization

In many optimization problems, the search process for the best input or parameters needs to take into account more than one objective. A common approach is to combine the objectives into a single objective function (e.g., using a linear combination with fixed coefficients). The main disadvantage of this approach is that the definition of coefficients in a linear combination requires choosing the appropriate weight for each objective. If these weights are not available beforehand, a possible solution is to optimize each objective separately (using single-objective optimization), and then estimate the trade-offs as a linear combination of the solutions, where the coefficients can be chosen after visual inspection of the single-objective solutions. However, this approach does not permit to recover non-convex sections of the set of optimal solutions. Furthermore, generally, no prior knowledge is available on how two or more particular objectives are balanced in a given cell. Therefore, rather than establishing a weight for each objective and then combining them into a single objective, we optimize all the objectives simultaneously through an evolutionary algorithm, providing the final trade-off curve. Throughout this

paper, we will consider the following two objectives: (i) biomass, and (ii) phosphoglycerate dehydrogenase (PHGDH).

In a given multidimensional objective space, the trade-off solution set, also called the *Pareto front*, is the set of points x such that there does not exist any other point *dominating* x in all objectives. Formally, let $\phi_1, \ldots, \phi_r$ be r objective functions to be maximized or minimized. The multi-objective optimization problem is the problem of optimizing the vector function $\phi(x) = (\phi_1(x), \phi_2(x), \ldots, \phi_r(x))$, where $x \in X$ is the variable (vector) to be optimized in the search space $X \subseteq \mathbb{R}^N$. For a maximization problem, a *Pareto optimal* vector $x^* \in X$ is a point such that there does not exist any other point x that dominates x^*. A point x would dominate x^* if:

$$\phi_i(x) \geq \phi_i(x^*), \forall i = 1, \ldots, r, \text{ and} \tag{4}$$

$$\exists j \in \{1, \ldots, r\} \text{ such that } \phi_j(x) > \phi_j(x^*) \tag{5}$$

(or, equivalently, for a minimization problem, $\phi_i(x) \leq \phi_i(x^*), \forall i = 1, \ldots, r$, with at least one j such that $\phi_j(x) < \phi_j(x^*)$). If at least two objectives ϕ_i and ϕ_j are in conflict with each other (e.g., when an increase in the first objective requires a decrease in the second), then the Pareto front contains multiple non-dominated trade-off solutions. In our case, we consider $r = 2$ objective functions and N genes, whose space of expression levels (and more specifically fold changes) represents the search space.

2.5 Multi-Objective Parallel gzenetic Algorithm (PGA)

Genetic algorithms are advanced heuristic methods that have been successfully applied to solve hard problems involving biological optimization of given objective functions in a wide range of parameter selection and control problems. These include knockout- or partial knockdown-based metabolic engineering [28], robust systems design [29], evaluation of response to cancer treatment [30], identification of cancer glioma tumors [31], or design of chemotherapies [32].

To identify the *Pareto optimal* vector $x^* \in X \subseteq \mathbb{R}^N$, we employ a multi-objective parallel genetic algorithm (PGA) inspired by NSGA-II [33]. This genetic algorithm does not need weights for the objective functions, which enables us to seek gene expression profiles corresponding to trade-off solutions for simultaneous optimization (maximization or minimization) of multiple reaction fluxes. Each individual (i.e., gene expression profile) of the initial population is initialized as $(1, 1, \ldots, 1) + \lambda \in \mathbb{R}^N$, where $\lambda \in \mathbb{R}^N$ is a random uniform noise in the range $(-1; 1)$ to ensure early variability in the population. Mutations and cross-over are used to generate a new offspring of gene expression profiles.

Each individual of the PGA population is mapped onto the model using Eq. (2), therefore creating a population of contextualised metabolic models, which is the run using bilevel FBA in Eq. (3). Hence, for each gene expression profile (individual) generated by the PGA, a vector of flux rates $v \in \mathbb{R}^n$ is generated, output of the metabolic model corresponding to that individual. Such output is projected onto a two dimensional space, therefore mapped onto a point described as a pair of values (uv, tv), where uv is the first (inner) objective representing the biomass rate, and tv is the second (outer) objective representing the phosphoglycerate dehydrogenase (PHGDH) flux rate. Then, we modify the population of individuals through mutations and cross-over using the PGA to run again the FBA and generate another Pareto optimal front under the new gene expression values. In this way, we can analyse how the gene expression profiles affect the rates of biomass and PHGDH as shown in Fig. 2, exploring the transcriptomic search space by mapping each gene expression profile generated by the PGA onto a 2D metabolic objective space. We set the initial population equal to 128 individuals and the maximum number of populations generated equal to 384 (see the Supplementary Information, available online for more details).

More details on the PGA settings are provided as Supplementary Information, available online. The full Matlab code and the steps to execute it are at https://github.com/claudioangione/PGA_and_C-EDGE/. The code has been tested on MATLAB R2018b, and we suggest running it in parallel on a multi-core CPU (with the MATLAB parallel toolbox), as it has been fully parallelised to improve the speed of the PGA.

2.6 Controllability Analysis

Controllability analysis is a technique to evaluate how robust a given solution is, when it undergoes small perturbations (e.g., changes in one or more genes of the optimal expression profile found by the multi-objective PGA). From a biological standpoint, implementing *in vitro* an overexpression/underexpression strategy with low controllability coefficient ensures that the final result is reached even if small errors are made during the implementation of the *in silico* strategy. Therefore, when implementing a solution found by the PGA, the number and width of the errors that can be made without affecting the final cellular outcome can be estimated by computing the controllability of the solution.

Here, for each solution found by the PGA with a PHGDH production above the significance threshold $\mu_P + 2\sigma_P$ (mean and standard deviation of the PHGDH flux rate across the space sampled by the PGA), we evaluate the controllability to gene perturbation. Given a point $y = (y_1, \ldots, y_p)$ in the objective space, corresponding to the profile $x = (x_1, \ldots, x_N)$, where N is the number of genes in the model and p the number of objectives, we define its controllability coefficient $P(y)$ as

$$P(y)^+ = \frac{\prod_{j=1,\ldots,p} \left| y_j^+ - y_j \right|}{\prod_{j=1,\ldots,p} |y_j|}, P(y)^- = \frac{\prod_{j=1,\ldots,p} \left| y_j^- - y_j \right|}{\prod_{j=1,\ldots,p} |y_j|}; \quad P(y) = \max\left\{ P(y)^+, P(y)^- \right\}, \tag{6}$$

where $y^+ = f(x + \varepsilon\vec{1})$, $y^- = f(\max\{\vec{0}, x - \varepsilon\vec{1}\})$, $\vec{1}$ being the all-ones vector, and f being the function that associates an expression profile with an objective vector through the gene-expression augmented FBA in Eq. (3). Note that the operator $\max\{\cdot, \cdot\}$ is applied entry by entry on the two arrays. In the subsequent analyses we perform a perturbation of $\varepsilon = 0.01$ in the gene expression levels.

The controllability index $P(y)$ is a proxy for the relative instability of a given solution y found by the PGA, and takes into account the worst output perturbation obtained as a result of positive and negative gene perturbation. A zero $P(y)$ index indicates that the phenotypic outcome of the

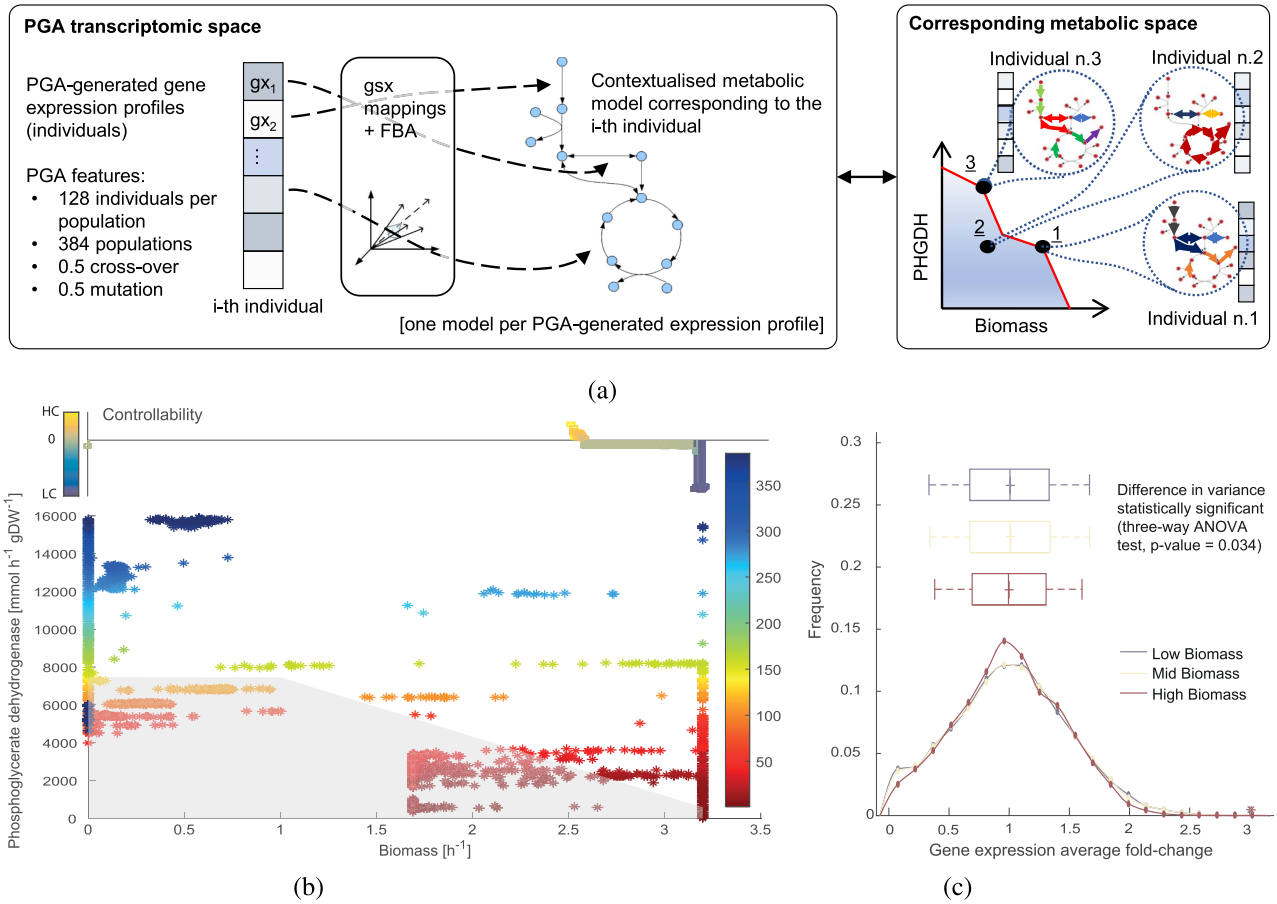


Fig. 2. *(a) Full multi-objective optimization procedure.* Individuals, corresponding to gene expression profiles, are generated by the parallel genetic algorithm (PGA). For each individual, a context-specific metabolic model is constructed following the gsx mappings in Eq. (2). This model is then run using the bilevel program in Eq. (3), obtaining values for the two objectives (biomass and PHGDH), which therefore allow us to map the model onto a metabolic space. The value of the objectives is then fed back to the PGA for mutation and cross-over, and drives the generation of new individuals to explore the metabolic space. *(b) Multi-objective optimization of biomass and PHGDH.* As well as showing the progress of the genetic algorithm, the plot highlights the portions of the space where human metabolism is able to operate. For instance, there are only a few flux distributions with medium biomass yield and high PHGDH. Solutions are denoted by progressively colder colors depending on the PGA population in which they have been generated. From a preventive point of view, the most interesting points are those with low PHGDH and high biomass, representing the case in which the cell has high biomass but low chance of developing PHGDH-dependent cancer. Conversely, from a therapeutic standpoint, the key solutions are those with high PHGDH (and therefore potential cancer cells), but impaired cell growth. Overall, we let the PGA generate 384 populations containing 128 gene expression profiles each. The grey area represents the Pareto front identified from an alternative multi-optimization model by Budinich *et al.* [37]. The plot shows that [37] covers a smaller area in the biomass-PHGDH metabolic space. In particular, all the solutions in the grey Pareto front are dominated by the solutions identified by the method proposed here, showing that our approach is able to explore a larger area of the solution space. The nonzero controllability coefficients R are shown (in log-scale) from lowest (LC) to highest (HC) with "stems" above each point in the bottom panel. The points with lowest nonzero controllability, and therefore high robustness to perturbations, are also the ones with the highest biomass yield. *(c) Highest, Lowest and Mid-Biomass Scenarios.* Average of gene expression levels across the three regions of the Pareto front. To distinguish the highest, lowest and mid-biomass scenarios, we extract a low-biomass and a high-biomass sections from the points sampled by the PGA, defining the remaining points as mid-biomass. More specifically, the low-biomass points are those whose biomass is negligible (less than 10^{-10}), while the high-biomass points are those whose biomass is more than $\mu_b + 2\sigma_b$, where μ_b and σ_b are the mean and standard deviation of the biomass values across all the points sampled. The x-axis shows the fold change of each gene compared to its control value, which is encoded as the baseline value of 1 within the model. In each region, we also show the box plots with lines indicating 9-91 percent probability mass. To test whether there is a difference in the variance of the three distributions (highest, lowest and mid-biomass) we run a three-way anova test. The null hypothesis of equal variance was rejected with p-value = 0.034, therefore leading us to accept the alternative hypothesis that the three distributions have different variances, from the largest (low-biomass) to the smallest (high-biomass).

expression profile is robust to positive and negative perturbations. The greater the $P(y)$ index, the less robust the point. The definition in Eq. (6) measures the outcome of the input perturbation as a percentage of the output perturbation, therefore taking into account the initial output value; this is to account for the fact that a strong output perturbation on a large output value is less noticeable than the same perturbation on a small output value.

Since in our multi-objective setting we do not weigh the objectives, our calculation of the distance between two points y^α and y^β using Eq. (6) in the metabolic space is inspired from the hypervolume indicator [34] applied to the first two coordinates of the point, representing the biomass and the PHGDH. Formally, the distance is defined as

$$d_H(y^\alpha, y^\beta) = \prod_{j=1,2} (y^\alpha_j - y^\beta_j). \tag{7}$$

Two points with equal hypervolume may lie in different regions of the objective space, but they represent an equivalent choice for a decision-maker, due to the absence of weights establishing the relative importance of the

objectives. If such weights are provided, the definition of distance between two points may be modified accordingly, and also $\|\cdot\|_1$ or $\|\cdot\|_2$ (euclidean) distances can be used.

2.7 The EDGE Algorithm

The effect of a single gene presence or absence in a metabolic model can be estimated through the EDGE algorithm [13]. Formally, the EDGE score of a gene g is defined as

$$\mathrm{EDGE}(g) = \min_{j=1,\dots,K} f^{\varepsilon}(T_g, j) - f^0(T_g), \tag{8}$$

where f is the objective of the linear program in Eq. (1) (without loss of generality, we will assume f represents the biomass); T_g is the set of reactions associated with g; j ranges over the K reactions associated with g; $\min_{j=1,\dots,K} f^{\varepsilon}(T_g, j)$ represents the lowest biomass that is obtainable when one reaction (among the K reactions associated with g) has its flux rate set to ε, while the other reactions are set equal to zero; the reactions not associated with g are not constrained; $f^0(T_g)$ is the biomass when all the flux rates of the K reactions associated with g are set to zero.

2.8 C-EDGE: Computing Expression-Based Effect of Single-Gene Perturbations

The EDGE algorithm can be used to perform a systematic evaluation of gene effects in a metabolic model to correctly predict growth phenotypes after gene overexpression [13]. However, one of the main limitations of EDGE is that the perturbation is applied directly to flux rates, and therefore it represents a perturbation only when the reaction is controlled by a single gene.

To overcome this, we develop C-EDGE (Controlled Expression-Dependent Gene Effects), which considers genes, rather than reactions, as fundamental units, and assesses the role of perturbations in gene expression levels. Unlike the standard EDGE approach, C-EDGE is not a reaction-based approach. In fact, it allows predicting growth directly related to gene expression in a more detailed fashion. Therefore, it can be applied directly to all genes in the model. Unlike EDGE, it does not need to exclude complex gene sets, including any combination of enzymatic complexes and isoenzymes. Since our augmented metabolic model is gene-based, we define our gene score in C-EDGE$_1$ by setting all gene expression levels at their initial value, while computing the difference between the biomass when a gene g is ε-expressed and when it is knocked out. This is equivalent to computing the sensitivity analysis of g in 0 rather than in its initial value. Formally, we define:

$$\begin{aligned} \text{C-EDGE}_1(g) =& f(x^{(\varepsilon)}) - f(x^{(0)}), \\ x^{(\varepsilon)} = (1,\dots,1,\varepsilon,1,\dots,1), &\quad x^{(0)} = (1,\dots,1,0,1,\dots,1), \end{aligned} \tag{9}$$

where $x^{(\varepsilon)}, x^{(0)} \in \mathbb{R}^N$, N being the number of genes in the model, and f represents the cellular objective, in our case the biomass flux rate. (Note that ε and 0 in the gene expression profiles $x^{(\varepsilon)}$ and $x^{(0)}$ must be in the location associated with g in the gene expression profile.)

Therefore, the C-EDGE$_1$ algorithm is based on the following three steps:

(i) Eq. (2) is used to map the gene expression profiles $x^{(\varepsilon)}$ and $x^{(0)}$ to obtain $gsx(x^{(\varepsilon)})$ and $gsx(x^{(0)})$;
(ii) The metabolic model is then run twice through Eq. (3), with $gsx(x^{(\varepsilon)})$ and $gsx(x^{(0)})$ to compute $f(x^{(\varepsilon)})$ and $f(x^{(0)})$ respectively;
(iii) Finally, C-EDGE$_1(g)$ is computed using Eq. (9).

Then the next gene is perturbed and steps (i), (ii) and (iii) are performed to compute the C-EDGE$_1$ value of that gene. Hence, by perturbing one gene at a time, and then running the model in Eq. (3) after applying our gsx map in Eq. (2), C-EDGE$_1$ allows to investigate the single-gene effect on the cellular objective.

While the standard EDGE cannot be applied as a difference between "wild-type plus epsilon" and "wild-type" fluxes due to difficulty of obtaining internal fluxes for the wild type [35], C-EDGE$_1$ is suitable for computing the same difference because it is applied directly to gene values, for which the wild-type profiling might be easier to perform in vitro. Therefore, C-EDGE$_1$ can be computed starting from any gene expression profile and performing one-at-a-time ε-perturbation of gene expression, in order to evaluate the difference in the biomass or any other flux in the model.

2.9 C-EDGE$_k$: Computing Combined effezct of High-Order Gene Perturbations

The prediction of combined gene effects enables the identification of potential treatments involving multiple targets (e.g., a single multi-target drug or a combination of single-target drugs). To this end, we generalize Eq. (9) for a set G of k genes, $G \subseteq \{g_1,\dots,g_N\}$, $|G| = k$, for which we compute the simultaneous effect on the objective f of the linear program:

$$\begin{aligned} \text{C-EDGE}_k(G) &= f(x^{(\varepsilon)}) - f(x^{(0)}), \\ x_i^{(\varepsilon)} &= \varepsilon \ \text{ if } \ g_i \in G,\ x_i^{(\varepsilon)} = 1 \text{ otherwise}, \\ x_i^{(0)} &= 0 \ \text{ if } \ g_i \in G,\ x_i^{(0)} = 1 \text{ otherwise}. \end{aligned} \tag{10}$$

Our idea is therefore to perform a combinatorial evolutionary search in the high-dimensional ($k \geq 3$) space of all possible k-uples of genes, to enable the prediction of those tuples such that their combined C-EDGE score is different from the C-EDGE scores of its subsets. The search is guided by a single objective function that represents the difference between the C-EDGE of a set of k genes and the C-EDGE of all its $k-1$-subsets. We therefore implemented a single-objective PGA (soPGA) to explore the discrete search space (present/absence of a gene in a k-uple), and to perform the maximization of a single objective function as formalised below.

The aim of the high-order C-EDGE$_k$ is to highlight those combinations of k genes whose C-EDGE score is different from the C-EDGE score of all subsets with $k-1$ genes. Therefore, given a set of k genes $G = \{g_1,\dots,g_k\}$, we designed the following soPGA objective function:

$$\Delta(G) = \prod_{1 \leq i \leq k} |\text{C-EDGE}_k(G) - \text{C-EDGE}_{k-1}(G \setminus g_i)|. \tag{11}$$

The sets whose C-EDGE$_k$ is different from all the C-EDGE$_{k-1}$ are those with nonzero $\Delta(G)$. A single objective optimization algorithm is therefore sufficient in this case. An equivalent approach would be to cast this problem as a k-objective optimization problem, where the k objectives are

the maximization of $|\text{C-EDGE}_k(G) - \text{C-EDGE}_{k-1}(G \setminus g_i)|$, and we are interested in any point not lying on any of the axes, i.e., with all nonzero coordinates. The advantage of using Eq. (11) is that $\Delta(G)$ can be regarded also as the confidence of our prediction for the peculiarity of the subset G.

For the soPGA underlying the high-order C-EDGE, we considered only the mutation operator. Cross-over, another commonly used PGA operator, is not biologically meaningful in this context, because it would bring together and merge two different subsets. Conversely, mutations starts from a subset and tries to modify it until it reaches a better combination for the C-EDGE.

Without this C-EDGE approach, if $k = 3$, evaluating each triple in Recon 2 would require more than 1.75 billion runs of the bilevel linear program that simulates the metabolism, which is in the order of months of CPU time. It is worth noting that the multi-objective PGA was used to explore the continuous gene expression search-space in order to find the optimal Pareto solutions, while the soPGA constituting the C-EDGE algorithm is used to investigate the effects of presence, absence or ε-expression of sets of genes. Therefore, the search space exploration carried out by the soPGA contained in C-EDGE is used to select k-uples of genes that maximise $\Delta(G)$. As a result, it is a discrete search space (presence/absence of each gene in the k-uple).

Finally, we remark that in the definition $\Delta(G)$ we only consider all the subsets with cardinality $k-1$. One might argue that it is worth computing the same type of difference, say $\Delta'(G)$, comparing C-EDGE_k(G) with all the subsets of G, and not only with the subsets with $k-1$ elements. However, $\Delta'(G)$ would not be indicative of the role of G because we would not detect all cases in which $\text{C-EDGE}_k \neq \text{C-EDGE}_{k-1}$. For instance, the $\Delta'(G)$ approach would fail when $\text{C-EDGE}_{k-2} = \text{C-EDGE}_k$, which would give $\Delta'(G) = 0$ (and therefore the k-uple would not be deemed interesting) even if $\text{C-EDGE}_k \neq \text{C-EDGE}_{k-1}$, and therefore the k-uple should actually be deemed interesting.

3 Results

3.1 Multi-Objective Optimization of Biomass and PHGDH

In our augmented Recon model, we use bilevel Flux Balwith a Parallel Genetic Algorithm. In the bilevel formulation in Eq. (3), we take as a second-level objective tv the minimization/maximization of phosphoglycerate dehydrogenase (PHGDH) in the model, with the maximum/minimum possible biomass (first-level objective uv). Both objective functions act as objectives for the PGA. We then take these objectives and optimize both, finding the trade-off if the two objectives conflict with each other. For instance, suppose that v^* is a (proposed) solution to the bilevel linear program in Eq. (3) that maximizes the PHGDH while minimizing the biomass, therefore estimating the minimum growth rate achievable by the cell in a given condition. We cast the optimization problem as a vector minimization problem of the form $min(uv^*, -tv^*)$, where the maximization of the PHGDH has been cast as the minimization of its negation. Then, the genetic algorithm seeks the best gene expression profiles that, once encoded as constraints and after Eq. (3) is solved, lead to the optimal trade-off between these two objectives. We therefore obtain a Pareto front in the biomass-PHGDH metabolic space, where each point corresponds to a gene expression profile in the genotypic space (Fig. 2b). As a result, by considering this augmented model in combination with bilevel FBA, we enable mechanistic evaluation of the metabotype for any given gene expression profile.

We use the phenotypic space as a means to explore the regions where the cell can operate, i.e., its metabolic potential [21]. When all genes are normally active, the model predicts maximum biomass. Modifying the gene expression values leads to a reduced or essentially unchanged biomass. Indeed, a common assumption is that the cell's expression pattern is adapted to its external environment, and therefore a change in its gene expression profile (i.e., change of the external conditions) causes a reduction in the biomass [36]. The vector minimization problem coupled with the FBA simulation allows us to effectively explore metabolic potentials with low biomass yield.

The Pareto front represents a set of optimal states that can be reached by human metabolism. Since a PGA (or, in fact, any algorithm designed to estimate the Pareto front) cannot guarantee that better solutions will not be discovered with more populations or with different settings, the Pareto front can be thought of as a *lower bound* of optimal metabolic behavior.

In order to evaluate the robustness of the solutions, we used controllability analysis (see Eq. (6). Fig. 2c shows that the metabolic configurations with the highest biomass are associated with low controllability coefficient (and therefore high robustness). As a result, the points of the Pareto front with the largest biomass are suitable candidate solutions for the decision maker, as their robustness is high compared to other points in the space. This also suggests that metabolic configurations of low biomass are highly unstable if compared with medium and high biomass, the latter being the most stable configuration for human metabolism. Furthermore, during the exploration of the solution space by the PGA, the low-biomass regions were the most difficult to reach through its in silico genetic engineering, as the Pareto optimization algorithm was not able to reach areas of negligible biomass and low PHGDH.

3.2 Optimization-Driven Sampling of the Metabolic Landscape

Starting from the metabolic configurations found by the PGA, we investigate the average gene expression level in three scenarios (lowest biomass, highest biomass, and mid-biomass). To distinguish these scenarios, we extract a low-biomass and a high-biomass sections from the points sampled by the PGA. We defined as low-biomass points those whose biomass is negligible (less than 10^{-10}), while as high-biomass points those whose biomass is more than $\mu_b + 2\sigma_b$, where μ_b and σ_b are the mean and standard deviation of the biomass values across all the points sampled.

We ran a three-way analysis of variance [38] to test whether there is a difference in the variance of the three distributions (high, mid and low biomass). The null hypothesis of equal variance was rejected with p-value = 0.034, therefore leading us to accept the alternative hypothesis that the three distributions have different variances, with the high-biomass configurations showing the smallest standard

deviation in their gene expression levels (Fig. 2c). This indicates that the way to reach high biomass values is not simply overexpressing all genes to a very high level, as one might expect from a first analysis of the Recon model. Indeed, increasing the functioning rate of the biomass-producing machinery also increases the formation of byproducts, that need to be excreted within the current capabilities of the metabolic network.

Furthermore, we use a similar lowest/mid/highest biomass class separation to show the average expression level suggested by the PGA in the three scenarios (Table S6, available online). In this way, we allow the analysis of the expression level of single genes in all cases. The most overexpressed gene when moving from the low-biomass to the high-biomass metabolic landscape is SGMS1, whose expression is significantly altered in different types of cancer [39]. Conversely, the gene undergoing the largest underexpression when moving from the low-biomass to the high-biomass metabolic landscape is GMDS, whose decreased activity has been previously linked with resistance to TRAIL-induced apoptosis, and therefore increased tumor development and metastasis [40]. For both genes, our results prove a high sensitivity to changes in the biomass. Finally, with the goal of estimating the co-regulation of genes across different regions of the metabolic space (and specifically high/low biomass), we defined a distance between genes based on the correlation between their expression levels across the points sampled by the PGA. If p and q are two vectors representing the expression levels of two genes across the points sampled, we define a distance:

$$d(p,q) = 1 - \frac{(p-\vec{1})(q-\vec{1})}{\|p\|\|q\|}, \tag{12}$$

where $\vec{1}$ indicates the all-ones vector, and $\|\cdot\|$ is the euclidean norm. This definition allows us to capture the correlation of two gene expression profiles with respect to the deviation from the wild-type all-ones expression profile. By repeating this process for all the pairs of genes, we build a dissimilarity matrix $D_{pq} = d(p,q)$, and a weighted distance graph with genes as nodes, and edges (p,q) whose weight is the distance $d(p,q)$. In Figs. S2 and S3, available online, we use the low/high biomass separation of the metabolic space to perform hierarchical clustering applied in these two subspaces, therefore highlighting clusters of genes in different metabolic scenarios (see Supplementary Information, available online for more details).

3.3 C-EDGE: Effects of Single-Gene Perturbations

C-EDGE enables exact tests of toxicity on pairs, triplets (or larger sets) of genes. A gene is *neutral* if the cellular objective is constant regardless of the gene being KO or forced to be ε-expressed. It is *beneficial* if the cellular objective is reduced when the gene is knocked out, while it is *toxic* if the cellular objective is increased when the gene is knocked out. We take into account the biomass as the cellular objective. ε is an infinitesimal perturbation but it cannot be arbitrary due to the finite precision of the floating-point representation in Matlab. We took $\varepsilon = 10^{-2}$ as a gene perturbation, which in turn produces a perturbation in the order of 10^{-6} for the flux bounds according to Eq. (3).

TABLE 1
C-EDGE Score Computed on Single Genes by Perturbing of $\varepsilon = 0.01$ One Gene Expression Level at a Time

Entrez ID	Gene name	C-EDGE	Reactions
51727.1	CMPK	0.130973319	40
1717.1	DHCR7	0.09803441	3
7298.1	TYMS	0.076388358	1
9489.1	PGS1	0.06862476	1
10558.1	SPTLC1	0.057188608	1
9517.1	SPTLC2	0.057188608	1
259230.1	SGMS1	0.057188608	1

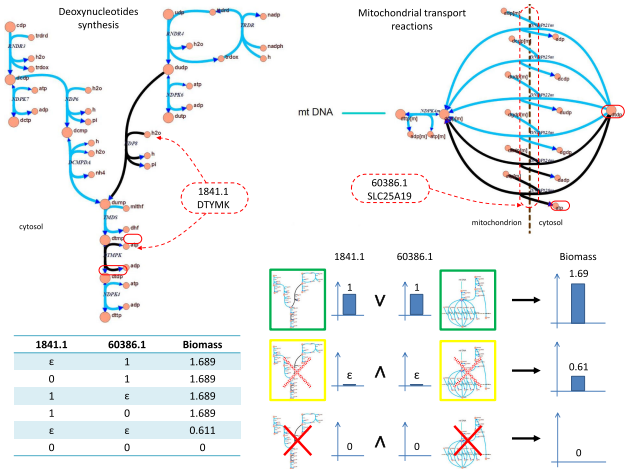
Only seven genes have non-negligible C-EDGE score. Different values of ε caused only changes in the values of the score, but not in the members of the list, showing that the method is fully robust to changes of the perturbation.

By perturbing one gene at a time, and then running the model using Eq. (3) after applying our gsx map in Eq. (2), we identified seven genes as highly beneficial (Table 1). The most sensitive gene, CMPK, takes part in 40 biochemical reactions. The genes SPTLC1, SPTLC2 and SGMS1 are part of the sphingomyelin biosynthetic process, which has been previously identified as a target for cancer therapy [41].

3.4 C-EDGE$_k$: Effects of High-Order Gene Perturbations

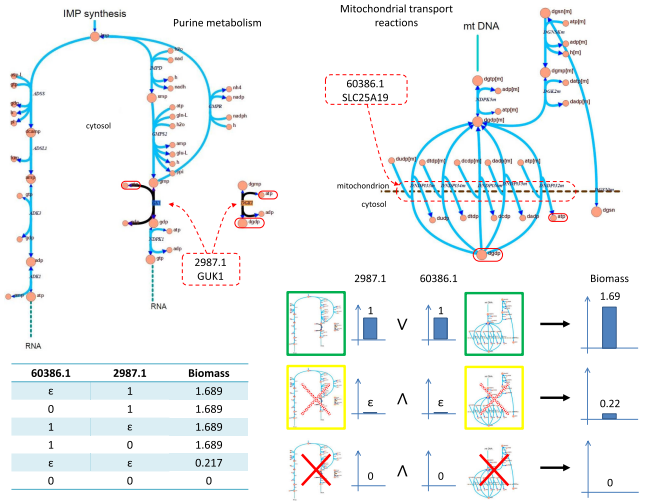
Using C-EDGE$_k$, we investigate the combined effect of high-order sets of gene perturbation on the metabolism. First, we tackle the problem of computing the combined effect of pairs of genes by defining a new single-objective PGA (soPGA, with objective function $\Delta(G)$, see Methods). Specifically, we are interested in those sets of pairs of toxic genes that become non-toxic if activated together. Among these pairs of genes, three pairs are found by the soPGA to have a surprising behavior of ε-activation; namely, they are remarkably beneficial if ε-activated as a pair, but both genes are slightly toxic if ε-activated one at a time. These three pairs detected by our method are the only pairs showing this behavior, as proved by the extensive computation of C-EDGE$_2$ that we perform in Table S2, available online.

In Fig. 3, we show the behavior of the first pair of genes (DTYMK, SLC25A19) in the human metabolic pathways, extracted from the BiGG database [42]. An ε-gene expression of both genes causes a decrease in the biomass. This interaction is due to the fact that dTDP needs to be produced and then transported into the mitochondrion from the cytosol. If both reactions are impaired, the metabolism is not able to compensate and the production of mitochondrial DNA is impaired, causing a decrease of biomass to 0.61 h^{-1}. In Fig. 4, we show the behavior of the pair (GUK1, SLC25A19). The interaction between these genes is due to the shared metabolite dGDP. The lack of both enzymes also affects mitochondrial DNA, in a more severe way with respect to the first pair. The biomass is decreased to 0.22 h^{-1}. Finally, the third pair (DEGS1.1, DEGS1.2), shown in Fig. S1, available online, is a pair of transcriptional variants of the same gene DEGS1. This kind of interaction is due to the fact that the conversion of dihydroceramide into N-acylsphingosine is key for the growth of the cell, and at least one of the two reaction branches must be active.



1841.1	60386.1	Biomass
ε	1	1.689
0	1	1.689
1	ε	1.689
1	0	1.689
ε	ε	0.611
0	0	0

Fig. 3. *C-EDGE$_2$ detects long-range pathway interactions.* DTYMK is responsible (alone) for dTMP kinase and nucleoside-diphosphatase (dUDP). SLC25A19 is responsible (alone) for 64 transport reactions, including the dUDP reaction associated with DTYMK. The products and reactants in common between the reactions controlled by DTYMK and those controlled by SLC25A19 are ADP, dTDP, ATP (here both reactions are reversible, therefore the terms reactant and products are interchangeable). A further analysis on the topology of the network justifies this behavior found by C-EDGE$_2$: if the expression of both genes is epsilon, there is (i) ε-production and (ii) ε-transport (SLC25A19 is a deoxynucleotide transporter) of dTDP into the mitochondrion from the cytosol, thus impairing mitochondrial DNA replication. In the absence of mitochondrial DNA, the biomass production is severely impaired. If at least one of them is fully working, the biomass is not impaired, irrespective of the expression level of the other gene of the pair.



60386.1	2987.1	Biomass
ε	1	1.689
0	1	1.689
1	ε	1.689
1	0	1.689
ε	ε	0.217
0	0	0

Fig. 4. *C-EDGE$_2$ detects short-range pathway interactions.* GUK1 is responsible (alone) for deoxyguanylate kinase (dGMP:ATP) and guanylate kinase (GMP:ATP). SLC25A19 is responsible (alone) for 64 transport reactions, including dATP, where the ATP is associated with GUK1. In this case, C-EDGE$_2$ retrieved a nontrivial relation between two pathways. dGDP is involved in both reactions, as well as ATP and ADP. While this behavior requires further experimental validation, we can speculate that the key role is played by dGDP, since ADP and ATP are widely diffused metabolites and can therefore be easily replenished by alternative reactions.

Finally, we applied C-EDGE to identify sets of genes whose over- or underexpression impedes cell proliferation ("toxic" genes), and those sets of genes whose activation is highly beneficial for the biomass. With the C-EDGE algorithm applied to sets of three genes (Table S4, available online), we were able to suggest sets of genes where the C-EDGE$_3$ score of the set is different from all the three C-EDGE$_2$ scores of the sub-pairs. Interestingly, we found some sets of genes (e.g., {SLC7A10,FUCA1, SLC40A1}) that are toxic as a set, although the three sub-pairs are beneficial to the biomass.

To check consistency between the results obtained through optimization and those obtained with C-EDGE, we applied multi-dimensional scaling to the low-biomass and high-biomass section of the Pareto front (see Supplementary Information, also Fig. S4, available online), also highlighting the position of the seven genes with the highest C-EDGE$_1$ score. As expected, the seven highly beneficial genes are more central in the high-biomass case. We finally carried out further statistical analysis on the Pareto front to highlight its relationship with C-EDGE (see Supplementary Information, available online).

As shown in Figs. 3, 4 and (S1, available online), our method proves useful to infer lethal combinations of reactions that are not directly related, and are not part of the same pathway, as well as crucial isozymes for the production of biomass. As a result, C-EDGE$_2$ and higher order C-EDGE$_k$ are able to compute the coupled robustness of pathways with respect to the overall cell metabolism, and to identify hidden lethal interactions between pathways that impair the production of biomass.

We remark that these results are unlikely to be found by chance or with a visual inspection of the metabolic map. While the first pair (Fig. 3) shares key metabolites for the biomass, and it may therefore seem straightforward to attribute a key role for the pair, a large number of reactions also share the same metabolites. For instance, the number of reactions in which the ATP is involved in the cytosol is 335, which means that, assuming one gene per reaction, up to $\binom{335}{2} = 55,945$ pairs of genes could be strong candidates for showing the same surprising behavior we found for this pair.

For $k \geq 3$, systematically exploring the space of all possible combinations of genes and evaluating the effect of their perturbation on the metabolism would be generally infeasible due to combinatorial explosion. C-EDGE would still require a relatively large amount of CPU time but, as shown with $k = 3$, it can dramatically decrease the time needed to explore this Boolean search space, also by proposing $k-$uples as soon as they are found during the generation of the populations within the soPGA.

3.5 Differential Analysis of Nine Tissue-Specific Cancer Models

In order to test C-EDGE on various cancer models, we use our framework to perform comparative metabolic analysis in the same tissue and across different tissues for cancer and normal cells. We use the genome-scale models of nine tissue-specific cancer models by Nam *et al.* [43], obtained from Affymetrix and Illumina Hiseq RNA-seq platforms, where the affy package and GIMME were used to generate each model. Table S10, available online reports the details of each model including the number of reactions, metabolites and genes.

By considering tissue-specific models as a case-study of C-EDGE, we provide hypotheses on the molecular basis of cancer for nine tissue-specific metabolic models, obtained using gene expression profiles of primary cancer cells and the corresponding normal cells [43]. Importantly, the method we propose seeks information on genes taking into account

only the effect of their expression level on the metabotype, and not the expression level itself. In Tables S7 and S11, available online, we list the genes detected as beneficial (positive C-EDGE score), neutral (zero C-EDGE score) or toxic (negative C-EDGE score) for each of the nine metabolic models in cancer and normal configurations. C-EDGE computes the metabolic effect of the perturbation of single genes in the model.

We find that the gene CRLS1, controlling cardiolipin synthase, is the only gene with positive C-EDGE score in all normal cells and in all cancer cells. Cardiolipin is a phospholipid at the heart of mitochondrial metabolism. It is found mostly in the inner mitochondrial membrane and plays a pivotal role in ensuring mitochondrial function. Our finding is confirmed by studies that correlate changes in cardiolipin content or composition with most cancers [44]. More specifically, this correlation is due to the fact that energy metabolism is impaired in most, if not all, cancer cells, independent of tissue origin. Moreover, tumour-related metabolism and the mitogenactivated protein kinase (MAPK) signaling pathways were found to be enriched with CRLS1-coexpression genes. CRLS1 has also been classified as novel tumor suppressor involved in regulating lipid and selenoamino acid metabolism in the tumour microenvironment [45].

From the comparison of the nine models, CMPK resulted as a neutral gene for the normal liver cell, while is always with positive C-EDGE score in most of the other models. Cytidine monophosphate kinase (CMPK), a member of the nucleoside monophosphate kinase family, plays an important role in the biosynthesis of nucleoside metabolism and tumour development. Furthermore, knock-down of CMPK significantly inhibits cancer cell proliferation, migration and invasion [46].

Phosphatidylglycerophosphate synthase (PGS1) is in six cases neutral for cancer cell, but negative for the normal cell; it catalyzes the first step in the biosynthesis of the mitochondrial phospholipid cardiolipin (finalized by CRLS1). PGS1 has been classified as a potential target that prevents cell growth, which also supports our findings [47].

The sphingomyelin synthase 1 gene (SGMS1) has been classified by the model as beneficial for all types of cancer. Indeed, SGMS1 is one of the genes, whose expression is often altered in cancer [39]. It plays a crucial role in cancer since it controls the inhibition of the proliferative signaling pathways in cancer cells. In addition, it has been shown that the activation of SGMS1 increases saturated fatty acids incorporated in a number of cancer cells [48].

Finally, biotinidase (BTD) was identified as beneficial for breast, kidney, liver and lung cancer cells. This is supported by several studies that report BTD as novel marker in cancer [49], [50]. Our pipeline can also offer routes to predict gene targets for potential drug development. For instance, in both the breast cancer cell and in the lung cancer cell (adenocarcinoma), an overexpression of MLYCD was correctly detected to be beneficial for the cancer cell, while being toxic for the normal cell. This result has been experimentally proven by other in silico experiments [51], therefore suggesting MLYCD as a key target to inhibit in order to selectively impair proliferation in cancer cells while not affecting normal cells.

3.6 Comparisons With Previous Approaches

We compared the biomass and PHGDH flux rates reported in Fig. 2 with those obtained by running the method proposed by Budinich *et al.* [37]. The latter is based on bilevel optimization and it is carried out using BENSOLVE [52], which computes a set of directions and points describing the image of the efficient points. This algorithm provides exact solutions by calculating the objective space and identifying the vertices, which corresponds to Pareto optimal points. We ran the model proposed in [37] in exactly the same setting and on the same metabolic model adopted in our approach. First, we used Bensons algorithm for multi-objective flux balance analysis (MO-FBA). Then, we ran the BENSOLVE solver to investigate the solution space.

The grey area in Fig. 2b shows the search space covered by the method presented in [37]. The points in the Pareto front obtained using the method proposed by Budinich *et al.* [37] are all dominated by the Pareto front calculated by the methodology proposed here, which also includes biomass optima that are not reached by Budinich *et al.* Although this indicates that the method proposed in Budinich *et al.* does not cover the whole area of the Pareto front identified by our method, the computational cost is considerably lower than our PGA approach. Hence, it provides a highly effective solution for a fast estimation of a lower bound of the Pareto front in cases of low computational capacity or limited available computational time.

Our method is also different from previous works that consider multi-objective optimization with single-gene knockout [11], [12], [13] since it allows exploring double or higher level gene perturbations with a non-linear approach providing new insights into pathways interactions (Figs. 3 and (S1, available online)). This is due to the integration of the PGA and C-EDGE with metabolic modelling, which allows us to identify optimal solutions for simultaneous non-linear optimization problems.

Finally, to show the added value of C-EDGE compared to the standard EDGE, we also ran the original EDGE algorithm [13] on the augmented Recon 2 model (Table S1, available online). The EDGE algorithm identified only two non-negligible genes: CRLS1 and SGMS1, which are involved in only one reaction. Both genes were also detected by C-EDGE in the augmented Recon 2 model (Table 1) or in the cancer-specific application (Tables S11, available online). Furthermore, the standard EDGE was not able to detect two of the genes involved in more than one reaction (i.e., CMPK involved in 40 reactions and DHCR7 involved in 3 reactions), which were all detected in our C-EDGE approach. This is due to the fact that the EDGE algorithm cannot be run at gene-level on isozymes or enzymatic complexes since it is based on flux rates' perturbations. Hence, it represents a single-gene perturbation only when the reaction is controlled by a single gene. Conversely, C-EDGE takes genes, rather than reactions, as fundamental units. In this way, we can assess the role of perturbations directly at gene level.

4 Conclusion

Despite often being recognized as a consequence of the state of a cell, metabolism is now widely accepted to play a central role in deciding the cell behavior [43], [53]. Recent evidence suggests that complex biological outcomes, including

onset of diseases, are often the result of the simultaneous regulation of multiple genes. To this end, in this work we took a multi-perturbation and multi-objective perspective. We proposed a pipeline of methods for optimization and analysis of gene expression and its effects on human metabolism. We used a manually curated and improved version of the human metabolic model Recon 2, augmented with quantitative transcriptional regulation. Our method can be used for the mapping of gene and protein expression onto the metabolism in a continuous fashion, without the need of thresholds for the Boolean status of low/high protein abundance [54].

Within the same pipeline, we proposed C-EDGE, a method based on a single-objective genetic algorithm to detect toxic, neutral, or beneficial sets of genes in the global metabolic model and in its tissue-specific versions. Previous methods for gene overexpression or knockout involved only single genes (e.g., the ASKA library [55]), pairs of genes (e.g., double knockout analysis [56] or synthetic genetic arrays [57]) with time-consuming procedures due to combinatorial explosion.

Existing experimental methods to predict the effect of the simultaneous perturbation of a set of genes are labor-intensive tasks. The main problem faced by these techniques is the scalability to sets of genes, becoming extremely challenging in three-wise gene analyses. In fact, although computational approaches for pairs of genes have been proposed [30], even studying synthetic lethality for three genes becomes computationally intractable. Our high-order C-EDGE can therefore dramatically improve predictions of cancer genes in cell-specific metabolic models. It can easily be applied to identify synthetic lethality and synthetic dosage lethality for high order k-uples of genes, from a gene-based perspective rather than from a reaction-based perspective [58].

Using the high-order C-EDGE on nine tissue-specific cancer and normal models, we found that the importance of a perturbation of a single gene can vary from cancer to cancer, and also at different stages of the same type of cancer [43]. The genetic modifications and the highly toxic and beneficial genes were not consistent across different cancer types. We remark that all our C-EDGE computations are performed directly on genes rather than on reaction fluxes. Therefore, compared to similar approaches [13], [43], we are able to analyze reactions controlled by isoenzymes or enzymatic complexes. We also showed how C-EDGE$_2$ and the high-order C-EDGE$_k$ are able to detect cases where a combined effect of different genes can lead to lethal consequences for the cell.

Our approach can be used on normal/cancer pairs of models to predict environmental or transcriptomic states that may reduce the proliferation rate of cancer cells. For instance, one can evaluate the cancer and normal metabotype associated with gene expression profiles in various conditions. With a multi-objective optimization algorithm, this allows seeking the environmental conditions that minimize the growth of cancer cells, while also minimizing the effect on normal cells. Further applications of our method are discussed in Supplementary Information, available online.

Manipulating genome and regulating gene expression finds applications in the reconstruction of engineered biological systems, with possible applications to drug development and human gene therapy [59]. We are able to quantitatively predict the combined effects of any set of genes (toxic, beneficial or neutral), which are not predictable from the analysis of the effects of the single genes on the growth rate. Pairwise and higher-order ($k \geq 3$) detection of combined gene effects is a desired feature in drug discovery, as it enables the investigation of treatments effective for multiple targets (e.g., a single multi-target drug or a combination of single-target drugs). This is especially useful in cancer therapeutics, where the target is an optimal dosage from a multiobjective standpoint, which maximizes efficacy and minimizes toxicity (therefore, multi-optimal solutions ensure high *therapeutic index*). Most importantly, we are able to identify high-order overexpression combinations of k genes whose effect on the phenotype is different from that of all subsets with $k-1$ genes.

The field of transcription modulation is experiencing a fast growth phase due to the recent advances in the CRISPR-Cas9 technology, which can be repurposed for regulation of the gene expression profile of a cell. Methods to find knockout strategies are successfully guiding metabolic engineering [60], [61], and general recent advances in mammalian cell engineering have been reviewed elsewhere [62], [63], [64]. Likewise, computational biology will soon need to address the lack of methods to guide genetic modulations of expression, where the considerably larger search space will likely require using metabolic modelling in combination with advanced machine or deep learning methods [65]. Our method may offer a route to find the best gene modulations (overexpression or partial knockdown) to carry out on multiple genes and towards multiple cellular objectives. For instance, expression vectors found with our method can be potentially used as a guide for CRISPR-Cas or CombiGEM systems [66], which have already proven successful in editing and modulating expression simultaneously across the genome [67], [68], e.g., to prioritize therapeutic targets in cancer cells [69].

Acknowledgments

HK is supported by the ISRAEL SCIENCE FOUNDATION (grant No. 190/19). CA would like to acknowledge the support from UKRI Research England's THYME project.

References

[1] B. Palsson, *Systems Biology: Constraint-Based Reconstruction and Analysis*. Cambridge, UK: Cambridge University Press, 2015.
[2] H. Lu *et al.*, "Modular metabolic engineering for biobased chemical production," *Trends Biotechnol.*, vol. 37, pp. 152–166, 2018.
[3] I. Thiele *et al.*, "A community-driven global reconstruction of human metabolism," *Nature Biotechnol.*, vol. 31, no. 5, pp. 419–425, 2013.
[4] Swainston *et al.*, "Recon 2.2: From reconstruction to model of human metabolism," *Metabolomics*, vol. 12, no. 7, pp. 1–7, 2016.
[5] E. Brunk *et al.*, "Recon3D enables a three-dimensional view of gene variation in human metabolism," *Nature Biotechnol.*, vol. 36, no. 3, p. 272–281, 2018.
[6] C. Angione *et al.*, "A hybrid of metabolic flux analysis and Bayesian factor modeling for multi-omics temporal pathway activation," *ACS Synthetic Biol.*, vol. 4, no. 8, pp. 880–889, 2015.
[7] L. A. Dunphy and J. A. Papin, "Biomedical applications of genome-scale metabolic network reconstructions of human pathogens," *Curr. Opinion Biotechnol.*, vol. 51, pp. 70–79, 2018.
[8] O. Shoval *et al.*, "Evolutionary trade-offs, pareto optimality, and the geometry of phenotype space," *Science*, vol. 336, no. 6085, pp. 1157–1160, 2012.

[9] C. Angione, "Human systems biology and metabolic modelling: A review - from disease metabolism to precision medicine," *BioMed Res. Int.*, vol. 2019, 2019, Art. no. 8304260.

[10] V. Sridhara *et al.*, "Predicting growth conditions from internal metabolic fluxes in an in-silico model of E. coli," *PloS One*, vol. 9, no. 12, 2014, Art. no. e114608.

[11] P. Zakrzewski *et al.*, "Multimeteval: Comparative and multi-objective analysis of genome-scale metabolic models," *PLoS One*, vol. 7, no. 12, 2012, Art. no. e51511.

[12] Y.G. Oh *et al.*, "Multiobjective flux balancing using the nise method for metabolic network analysis," *Biotechnol. Prog.*, vol. 25, no. 4, pp. 999–1008, 2009.

[13] A. Wagner *et al.*, "Computational evaluation of cellular metabolic costs successfully predicts genes whose expression is deleterious," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 47, pp. 19 166–19 171, 2013.

[14] P. Eisenhut *et al.*, "A crispr/cas9 based engineering strategy for overexpression of multiple genes in chinese hamster ovary cells," *Metabolic Eng.*, vol. 48, pp. 72–81, 2018.

[15] X. L. Li *et al.*, "Highly efficient genome editing via CRISPR–Cas9 in human pluripotent stem cells is achieved by transient BCL-XL overexpression," *Nucleic Acids Res.*, vol. 46, pp. 10195—10215, 2018.

[16] X. L. Li, "System wide analyses have underestimated protein abundances and the importance of transcription in mammals," *PeerJ*, vol. 2, 2014, Art. no. e270.

[17] M. Jovanovic *et al.*, "Dynamic profiling of the protein life cycle in response to pathogens," *Science*, vol. 347, no. 6226, 2015, Art. no. 1259038.

[18] I. Kosti *et al.*, "Cross-tissue analysis of gene and protein expression in normal and cancer tissues," *Sci. Rep.*, vol. 6, 2016, Art. no. 24799.

[19] C. Angione, "Integrating splice-isoform expression into genome-scale models characterizes breast cancer metabolism," *Bioinformatics*, vol. 34, no. 3, pp. 494–501, 2018.

[20] A. Bordbar *et al.*, "Constraint-based models predict metabolic and associated cellular functions," *Nature Rev. Genetics*, vol. 15, no. 2, pp. 107–120, 2014.

[21] R. Schuetz *et al.*, "Multidimensional optimality of microbial metabolism," *Science*, vol. 336, no. 6081, pp. 601–604, 2012.

[22] I. Rabbers *et al.*, "Metabolism at evolutionary optimal states," *Metabolites*, vol. 5, no. 2, pp. 311–343, 2015.

[23] J. Costanza *et al.*, "Robust design of microbial strains," *Bioinformatics*, vol. 28, no. 23, pp. 3097–3104, 2012.

[24] L. E. Quek *et al.*, "Reducing Recon 2 for steady-state flux analysis of HEK cell culture," *J. Biotechnol.*, vol. 184, pp. 172–178, 2014.

[25] J. L. Steffensen *et al.*, "Psamm: A portable system for the analysis of metabolic models," *PLoS Comput Biol.*, vol. 12, no. 2, 2016, Art. no. e1004732.

[26] K. Snell, "Enzymes of serine metabolism in normal, developing and neoplastic rat tissues," *Adv. Enzyme Regulation*, vol. 22, pp. 325–400, 1984.

[27] C. Frezza, "Cancer metabolism: Addicted to serine," *Nature Chem. Biol.*, vol. 12, no. 6, pp. 389–390, 2016.

[28] C. Angione and P. Lió, "Predictive analytics of environmental adaptability in multi-omic network models," *Sci. Rep.*, vol. 5, 2015, Art. no. 15147.

[29] C. Angione *et al.*, "A design automation framework for computational bioenergetics in biological networks," *Mol. BioSystems*, vol. 9, no. 10, pp. 2554–2564, 2013.

[30] J. S. Lee *et al.*, "Harnessing synthetic lethality to predict the response to cancer treatment," *Nature Commun.*, vol. 9, no. 1, 2018, Art. no. 2546.

[31] S. He *et al.*, "Cooperative co-evolutionary module identification with application to cancer disease module discovery," *IEEE Trans. Evol. Comput.*, vol. 20, no. 6, pp. 874–891, Dec. 2016.

[32] M. Villasana and G. Ochoa, "Heuristic design of cancer chemotherapies," *IEEE Trans. Evol. Comput.*, vol. 8, no. 6, pp. 513–521, Dec. 2004.

[33] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.

[34] J. Bader and E. Zitzler, "Hype: An algorithm for fast hypervolume-based many-objective optimization," *Evol. Comput.*, vol. 19, no. 1, pp. 45–76, 2011.

[35] P. F. Suthers *et al.*, "Metabolic flux elucidation for large-scale models using ^{13}C labeled isotopes," *Metabolic Eng.*, vol. 9, no. 5, pp. 387–405, 2007.

[36] A. Brandes *et al.*, "Inferring carbon sources from gene expression profiles using metabolic flux models," *PloS One*, vol. 7, no. 5, 2012, Art. no. e36947.

[37] M. Budinich *et al.*, "A multi-objective constraint-based approach for modeling genome-scale microbial ecosystems," *PloS One*, vol. 12, no. 2, 2017, Art. no. e0171744.

[38] O. J. Dunn *et al.*, *Applied Statistics: Analysis of Variance and Regression*. New York, NY, USA: Wiley, 1987.

[39] A. V. Rozhkova *et al.*, "Expression of sphingomyelin synthase 1 (*SGMS*1) gene varies in human lung and esophagus cancer," *Mol. Biol.*, vol. 48, no. 3, pp. 340–346, 2014.

[40] K. Moriwaki *et al.*, "GDP-mannose-4, 6-dehydratase (GMDS) deficiency renders colon cancer cells resistant to tumor necrosis factor-related apoptosis-inducing ligand (TRAIL) receptor-and CD95-mediated apoptosis by inhibiting complex ii formation," *J. Biol. Chem.*, vol. 286, no. 50, pp. 43 123–43 133, 2011.

[41] D. E. Modrak *et al.*, "Sphingolipid targets in cancer therapy," *Mol. Cancer Ther.*, vol. 5, no. 2, pp. 200–208, 2006.

[42] J. Schellenberger *et al.*, "BiGG: A biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions," *BMC Bioinformatics*, vol. 11, no. 1, 2010, Art. no. 213.

[43] H. Nam *et al.*, "A systems approach to predict oncometabolites via context-specific genome-scale metabolic networks," *Plos Comput. Biol.*, vol. 10, no. 9, 2014, Art. no. e1003837.

[44] M. A. Kiebish *et al.*, "Cardiolipin and electron transport chain abnormalities in mouse brain tumor mitochondria: Lipidomic evidence supporting the warburg theory of cancer," *J. lipid Res.*, vol. 49, no. 12, pp. 2545–2556, 2008.

[45] H. M. Feng *et al.*, "Expression and potential mechanism of metabolism-related genes and CRLS1 in non-small cell lung cancer," *Oncol. Lett.*, vol. 15, no. 2, pp. 2661–2668, 2018.

[46] D. Zhou *et al.*, "Cytidine monophosphate kinase is inhibited by the TGF-β signalling pathway through the upregulation of mir-130b-3p in human epithelial ovarian cancer," *Cellular Signalling*, vol. 35, pp. 197–207, 2017.

[47] G. Bidkhori *et al.*, "Metabolic network-based identification and prioritization of anti-cancer targets based on expression data in hepatocellular carcinoma," *Front. Physiol.*, vol. 9, 2018, Art. no. 916.

[48] V. Llado *et al.*, "Regulation of the cancer cell membrane lipid composition by nacholeate: Effects on cell signaling and therapeutical relevance in glioma," *Biochimica et Biophysica Acta (BBA)-Biomembranes*, vol. 1838, no. 6, pp. 1619–1627, 2014.

[49] A. K. C. So *et al.*, "Biotinidase is a novel marker for papillary thyroid cancer aggressiveness," *PloS One*, vol. 7, no. 7, 2012, Art. no. e40956.

[50] U. B. Kang *et al.*, "Differential profiling of breast cancer plasma proteome by isotope-coded affinity tagging method reveals biotinidase as a breast cancer biomarker," *BMC Cancer*, vol. 10, no. 1, 2010, Art. no. 114.

[51] K. Yizhak *et al.*, "Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer," *eLife*, vol. 3, 2014, Art. no. e03641.

[52] Löhne, "Bensolve-a solver for multi-objective linear programs,"

[53] A. Achreja *et al.*,"Exo-MFA–A 13C metabolic flux analysis to dissect tumor microenvironment-secreted exosome contributions towards cancer cell metabolism," *Metabolic Eng.*, vol. 43, pp. 156–172, 2017.

[54] S. Vijayakumar *et al.*, "Seeing the wood for the trees: A forest of methods for optimization and omic-network integration in metabolic modelling," *Brief Bioinformatics*, vol. 19, no. 6, pp. 1218–1235, 2017.

[55] M. Kitagawa *et al.*, "Complete set of ORF clones of Escherichia coli ASKA library (a complete set of E. coli K-12 ORF archive): Unique resources for biological research," *DNA Res.*, vol. 12, no. 5, pp. 291–299, 2006.

[56] Y. A. B. Goldstein and A. Bockmayr, "Double and multiple knockout simulations for genome-scale metabolic network reconstructions," *Algorithms Mol. Biol.*, vol. 10, no. 1, 2015, Art. no. 1.

[57] G. Butland *et al.*, "eSGA: E. coli synthetic genetic array analysis," *Nature Methods*, vol. 5, no. 9, pp. 789–795, 2008.

[58] W. Megchelenbrink *et al.*, "Synthetic dosage lethality in the human metabolic network is highly predictive of tumor growth and cancer patient survival," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 39, pp. 12 217–12 222, 2015.

[59] C. Fellmann *et al.*, "Cornerstones of CRISPR-Cas in drug discovery and therapy," *Nature Rev. Drug Discov.*, vol. 16, pp. 89–100, 2016.

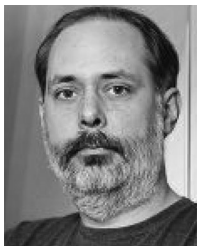
[60] K. Jensen *et al.*, "Optcouple: Joint simulation of gene knockouts, insertions and medium modifications for prediction of growth-coupled strain designs," *Metabolic Eng. Commun.*, vol. 8, 2019, Art. no. e00087.

[61] P. Schneider and S. Klamt, "Characterizing and ranking computed metabolic engineering strategies," *Bioinformatics*, vol. 35, pp. 3063—3072, 2019.
[62] J. B. Black *et al.*, "Mammalian synthetic biology: Engineering biological systems," *Annu. Rev. Biomed. Eng.*, vol. 19, no. 1, 2017, pp. 249–277.
[63] H. Yin *et al.*, "CRISPR–Cas: A tool for cancer research and therapeutics," *Nature Rev. Clin. Oncol.*, vol. 16, pp. 281–295, 2019.
[64] C. A. Lino *et al.*, "Delivering CRISPR: A review of the challenges and approaches," *Drug Del.*, vol. 25, no. 1, pp. 1234–1257, 2018.
[65] G. Zampieri *et al.*, "Machine and deep learning meet genome-scale metabolic modeling," *PLoS Comput. Biol.*, vol. 15, no. 7, 2019, Art. no. e1007084.
[66] L. Koch, "Genetic screens: CombiGEM-CRISPR: A creative combination," *Nature Rev. Genetics*, vol. 17, no. 4, pp. 194–194, 2016.
[67] R. J. Ihry *et al.*, "Genome-scale CRISPR screens identify human pluripotency-specific genes," *Cell Rep.*, vol. 27, no. 2, pp. 616–630, 2019.
[68] H. Zhou *et al.*, "In vivo simultaneous transcriptional activation of multiple genes in the brain using CRISPR–dCas9-activator transgenic mice," *Nature Neurosci.*, vol. 21, pp. 440–446, 2018.
[69] F. M. Behan *et al.*, "Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens," *Nature*, vol. 568, pp. 511–516, 2019.

Annalisa Occhipinti received the BSc and MSc degrees in mathematics from the University of Catania, the PhD degree in computer science from the University of Cambridge. She is a lecturer of data analytics at Teesside University. Her research is mainly focussed on computational biology and cancer research. She applies machine learning and data analytics to cancer data to detect new cancer markers and investigate genes networks. As part of her commitment to inspire young generations to pursue a career in STEM, she is also involved in many outreach activities to teach young students about applications of mathematics and computer science in cancer research.

Youssef Hamadi's received the PhD degree from the University of Montpellier, devising new algorithms for distributed constraint solving. His research is set at the intersection of optimization and artificial intelligence. In 2013, he defended his Habilitation on the concept of autonomous search at the University of Paris Sud. In 2003, he created the Constraint Reasoning Group at Microsoft Research, pushing the limits of parallel satisfiability and large-scale optimization, while transferring mathematical modeling and algorithms into several Microsoft products. In 2006, Youssef started to work on the relationships between mathematical programming and sustainability, creating jointly with CNRS the first European research project on optimization for sustainable development. He joined Uber Elevate in 2019 to work on autonomous aircrafts.

Hillel Kugler received the PhD degree from the Weizmann Institute of Science in Israel. He is a faculty member at the faculty of Engineering, Bar-Ilan University, in Israel, since 2015. His research interests are in modeling and analyzing complex systems using formal reasoning and synthesis methods. His research interests also include the application of visual languages to model the behavior of reactive systems, and the development of new computational methods and tools towards enabling a deeper understanding of biological computation and biological devices. Before joining the faculty of Engineering at Bar-Ilan he was a researcher at Microsoft Research in Cambridge. Previously he was a postdoctoral at the Biology Department in New York University and a member of the the Analysis of Computer Systems Group, the Department of Computer Science, Courant Institute, New York University.



Christoph M. Wintersteiger received the masters degree from the JK University of Linz, Austria, and the PhD degree in computer science from ETH Zurich, Switzerland. He is a senior research software development engineer with Microsoft Cambridge, U.K. He has been with Microsoft since 2010 and works on formal reasoning techniques and applications thereof in industry.

Boyan Yordanov received the BA degree in biochemistry and computer science from Clark University, in 2005, and the PhD degree in biomedical engineering from Boston University, in 2011. He was a postdoctoral researcher within the Mechanical Engineering Department at Boston University, in 2011. He joined the Biological Computation Group at Microsoft Research as a postdoctoral scientist and became a Microsoft research scientist, in 2014. His research is focused on accelerating the design and construction of biochemical circuits and improving the understanding of biological computation through computational methods for the analysis, verification, and synthesis of dynamical systems.

Claudio Angione received the BSc degree in applied mathematics, the MSc degree in mathematics from the University of Catania, and the PhD degree in computer science from the University of Cambridge. He is a reader of computer science at Teesside University, where he leads the "Computational Systems Biology and Data Analytics" research group. He currently also holds two Visiting professor positions at the University of Bari, Italy, and at KMUTT, Thailand. He has published more than 50 peer-reviewed papers in leading international conferences and computational biology journals, and has received several awards for his academic contributions in the community. His research group works at the intersection of computer science, mathematics and biology. Research topics include genome-scale metabolic modelling, multi-objective optimisation, computational systems biology, machine and deep learning.