NIHBA: A Network Interdiction Approach with Hybrid Benders Algorithm for Strain Design

Shouyong Jiang*[1], Yong Wang[2], Marcus Kaiser[3], and Natalio Krasnogor[3]

[1]School of Computer Science, University of Lincoln, Lincoln LN6 7TS, U.K.
[2]School of Automation, Central South University, Changsha 410083, China.
[3]School of Computing, Newcastle University, Newcastle upon Tyne NE4 5TG, U.K.

August 30, 2019

Abstract

Flux balance analysis (FBA) based bilevel optimisation has been a great success in redesigning metabolic networks for biochemical overproduction. To date, many computational approaches have been developed to solve the resulting bilevel optimisation problems. However, most of them are of limited use due to biased optimality principle, poor scalability with the size of metabolic networks, potential numeric issues, or low quantity of design solutions in a single run. In this work, we have employed a network interdiction model free of growth optimality assumptions, a special case of bilevel optimisation, for computational strain design and have developed a hybrid Benders algorithm (HBA) that deals with complicating binary variables in the model, thereby achieving high efficiency without numeric issues in search of best design strategies. More importantly, HBA can list solutions that meet users' production requirements during the search, making it possible to obtain numerous design strategies at a small runtime overhead (typically ∼1 hour).

1 Introduction

With the advance of genome-scale metabolic modelling (GSMM), the past decades have witnessed a significant number of computational tools for microbial metabolic engineering [Maia *et al.*, 2016]. These tools facilitate improved strain performance for the production of a variety of high-value biochemicals and biosynthetic precursors, including vanillin [Brochado *et al.*, 2010], lycopene [Choi *et al.*, 2010], malonyl-CoA [Xu *et al.*, 2011], alkane and alcohol [Fatma *et al.*, 2018].

A large number of strain design tools are based on bilevel optimisation. OptKnock [Burgard *et al.*, 2003] is one of the earliest bielve optimisation based tools. OptKnock maximises target chemical production while assuming mutant strains at optimal growth in flux balance analysis. The resulting bilevel problem is solved through a reformulation that makes the inner level problem equivalent constraints under the condition of strong duality [Burgard *et al.*, 2003]. The OptKnock model was latter extended to improve target production via gene up/down-regulation [Pharkya and Maranas, 2006], cofactor specificity [King and Feist, 2014], or heterologous pathways [Pharkya *et al.*, 2004], and to develop anti-cancel drugs by the identification of synthetic lethal genes [Pratapa *et al.*, 2015]. These studies demonstrate the great effectiveness of the bilevel optimisation based framework in metabolic engineering.

*math4neu@gmail.com

However, the bilevel optimisation based framework in literature has numerous limitations. The first one is the intensive computational cost in search of optimal solutions. Bilevel optimisation is often reformulated into a mixed-integer linear programming (MILP) so as to be solved by exact MILP solvers. It can take up to a week to solve an MILP resulting from a standard GSMM [Feist *et al.*, 2010]. Many practical strategies, such as model reduction and refinement of candidate knockout set [Feist *et al.*, 2010], have been used to reduce the computational time but may miss the best design strategies due to reduced search space. GDBB introduced a truncated branch and bound to speed up the search process. GDLS used local search with multiple search paths to reduce the search space for each local MILP [Lun *et al.*, 2009]. While finding optimal solutions are computationally costly for exact solvers, other studies resorts to inexact methods, such as genetic algorithms [Patil *et al.*, 2005; Rocha *et al.*, 2010] and swarm intelligence [Choon *et al.*, 2015]. These methods, however, still scale poorly with the size of GSMM and are specially ineffective when a large number of genetic manipulations are allowed for high target production.

In company with intensive computations, the resulting MILP often has weak LP relaxations due to disjunctive big-M constraints [Codato and Fischetti, 2006], another limitation of the current bilevel optimisation based methods. Big-M formulation can very easily cause numeric issues, particularly in genome-scale metabolic models where stoichiometric coefficients often vary many orders of magnitude [Sun *et al.*, 2013]. As a result, optimal strain design solutions returned from global search solvers such as Gurobi and CPLEX may turn out to be *in silico* infeasible. Model reformulation may alleviate numeric issues and potentially reduce computational costs. However, a proper model reformulation is often time-consuming and laborious as extra care has to be taken to prevent other numeric difficulties while fixing one.

The third limitation that only a single solution is obtained in each execution of optimisation by modern solvers. Multiple runs are required to generate more solutions, which inevitably increases computational burden. Despite that some commercial solvers, such as Gurobi and CPLEX, provide options to preserve multiple solutions in a single run, many alternative solutions exist for the optimal production rate in underdetermined metabolic systems, and they are likely to have similar production envelopes [Lewis *et al.*, 2012]. This similarity renders the bilevel optimisation less attractive as little information can be gained about the trade-off between growth and production for decision making. Also, these exact solvers consider two solutions different when their continuous but integer variables have different values. Such solutions lead to same design strategies, which is of not interest to decision makers. Heuristic methods, such as local search in GDLS [Lun *et al.*, 2009] and population-based algorithms [Jiang *et al.*, 2018; Patil *et al.*, 2005] may help to find diverse solutions but often suffer from local optima.

Another limitation in most of bilevel optimisation based tools is potential biases induced by the optimisation principle in the inner-level FBA [Lewis *et al.*, 2012]. OptKnock and many of its derivatives assume mutant strains have a biologically meaningful objective which is often to maximise growth. However, this assumption is not always correct as some microorganisms seem to achieve a multiobjective tradeoff of metabolism [Schuetz *et al.*, 2012], and mutants prefer small metabolic adjustments from the wild type [Segrè *et al.*, 2002]. It is there desirable that bilevel models eliminate the biased assumption on cell growth while optimising the target production rate.

There are also a number of tools free of bilevel optimisation. For example, approaches based on minimum cut set (MCS) identification have also developed to remove all possible design strategies that do not meet specific requirements of a desired production strain[Apaolaza *et al.*, 2018; von Kamp and Klamt, 2014]. MCS based approaches, however, are also computationally intensive as they need to enumerate elementary modes of a given metabolic network. OptForce identifies genetic interventions by investigating the difference in flux distributions between the wild type and the desired mutant [Ranganathan *et al.*, 2010]. OptForce showed good predictions in *in vivo* studies [Xu *et al.*, 2011]. However, the requirement on flux measurements of the wild

type, which is not always available, limits its wild applicability.

In view of these limitations in existing bilevel optimisation based tools, this paper considers the identification of genetic manipulations as network interdiction (NI), a problem well studied in game theory [Lim and Smith, 2007]. The NI involves one game player (host strain) that tries to avoid the overproduction of target chemicals for cellular homeostasis whereas the other opposing player (metabolic engineer) attempts to manipulate the metabolic network in order to maximally disrupt the first player's activity. Therefore, the NI is a max-min problem in which the objective involves only the target production, avoiding the use of the widely assumed growth optimality. The NI is a special case of general bilevel problems. The solution to this NI problem is a novel hybrid algorithm based on Benders decomposition [Codato and Fischetti, 2006], aiming to address the other limitations mentioned previously. NIHBA, the proposed approach, has shown the ability to efficiently find a large number of growth-coupled design strategies with diverse production envelopes in a single run, and to scale well with the size of allowable knockouts.

2 Results

2.1 NIHBA: using network interdiction and Benders decomposition

Most often than not, wild type strains maintain homeostasis and thus avoid overproducing a product of interest while maximising biomass (Fig. 1A). Metabolic engineering of host strains requires metabolic network modifications leading to improved flux towards the biosynthetic pathway of the target product. Metabolic engineers who look for the best modifications can be considered adversaries or interdictors as they use limited engineering costs (time, financial cost, etc.) to intervene the host's activities and reverse the cell's low target production (Fig. 1A). This strain design task is similar to network interdiction (NI), a game theoretic framework where a budget-constrained interdictor intervenes a network user's activity, e.g. commodity flow [Lim and Smith, 2007], by removing network arcs.

We proposed a NI model for identifying gene-associated reaction knockouts, although up/down regulation of genes can be consided in this model as well. The NI model is a special case of bilevel optimisation. It was recast into a standard MILP problem (Fig. 1B) using a special reformulation approach (Material and Methods). The resulting MILP contains both complicating binary variables and easy continuous variables. It can be computationally intensive for a large size of binary variables and/or a high allowable number of knockouts, and likely to have numeric issues for global search due to Big-M effects [Codato and Fischetti, 2006]. We therefore resorted to Benders decomposition for this NI model. We proposed a hybrid Benders algorithm (HBA) with two novel techniques to solve the model efficiently and obtain as many design solutions as possible in a single run (Fig. 1C). The solutions from our approach, NIHBA, were then analysed in production envelopes (Fig. 1D), from which the most promising design solution can be selected for implementation.

2.2 Case studies

NIHBA was tested on iML1515 [Monk *et al.*, 2017], the largest GEM model for E. coli, for the production of succinate and ethanol. For large models, Feist *et al.* [2010] suggested a preprocessing procedure to ease computational costs. In particular, reactions involving compounds that have more than a certain number (n_c) of carbons are not considered as knockout candidates since they are assumed unlikely to carry high flux. We tested different values of $n_c = \{10, 15, 22, 100\}$ ($n_c = 100$ indicates no reaction removed due to carbons), resulting in different sizes of candidate set $n_s = \{152, 204, 272, 342\}$. In all simulations, designed strains were required to have at least 10% cell growth of the wild type.

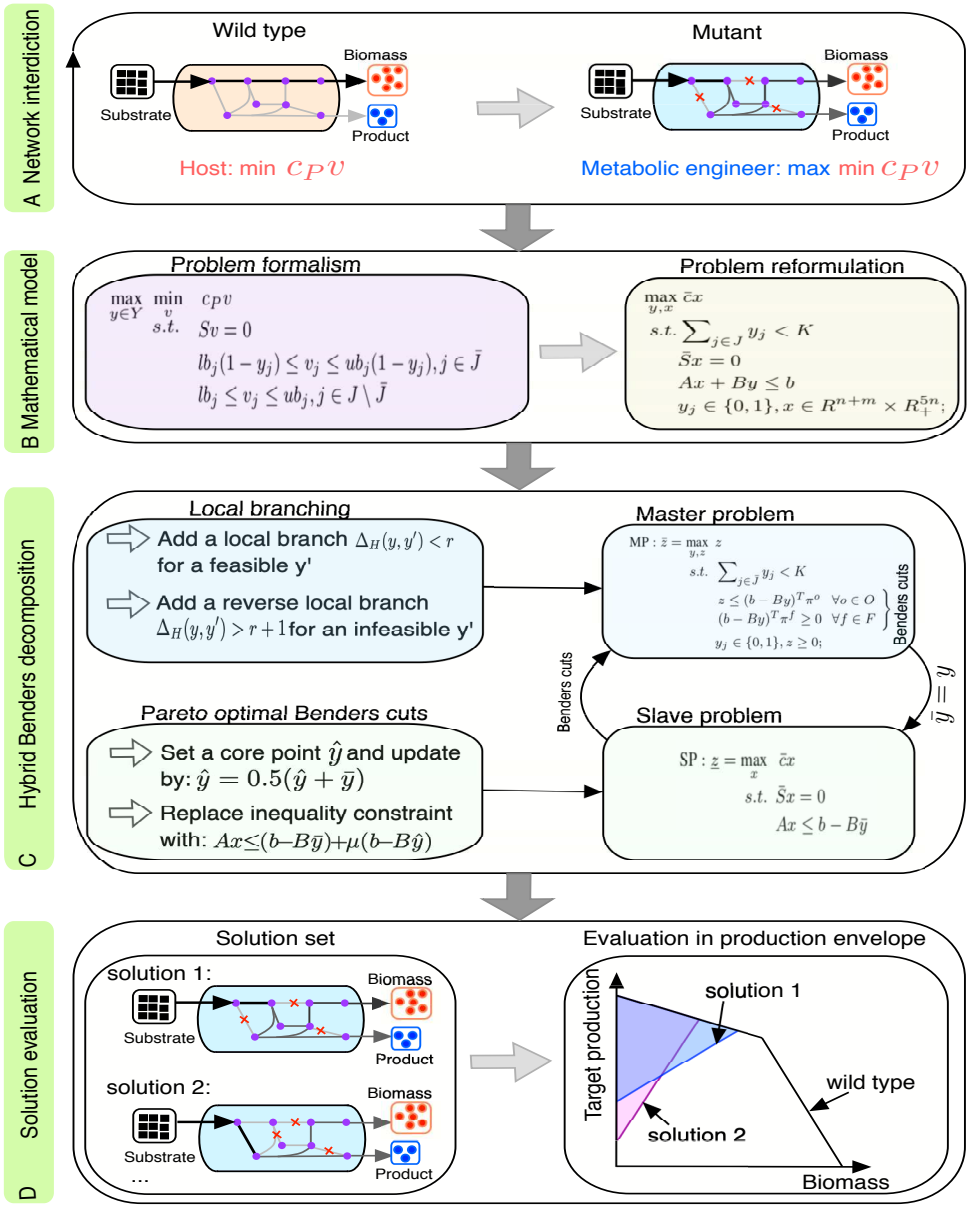

Figure 1: A schematic workflow of the proposed NIHBA tool for strain design. (**A**) Illustration of network interdiction in strain design: host cells avoid overproducing a product whereas metabolic engineers interdict the host network to maximally impair the host's activity. (**B**) Mathematically modelling the network interdiction problem in strain design, followed by problem reformulation to obtain a standard MILP problem. (**C**) Hybrid Benders decomposition algorithm. The MILP is decomposed into a binary master problem and a linear slave problem, and Pareo-optimal cut generation and local branching are introduced to speed up the search of solutions. (**D**) Solutions that meet production requirements are stored and evaluated in production envelopes.

We show in Table 1 that the reduction of candidate set by carbon number has a significant effect on target production, especially for succinate. Exclusion of reactions with a carbon number of over 10 (corresponding to 152 candidate knockouts) results in a low succinate production flux of 6.8838 $\sim$ 9.9430 mmol/gDW/h, which is less than a third of the theoretic maximum production (TMP). A slight relaxation of carbon number to 15 helps to identify a solution with around two thirds of TMP, and succinate production reaches $\sim$25 mmol/gDW/h(73% TMP) when no carbon number is constrained in candidate reactions. This indicates that some reactions with

Table 1: Succinate and ethanol production predicted by NIHBA with at most five knockouts for different sizes of candidate set. The growth rate, minimum production and maximum production are from the solution with the highest minimum production rate.

n_s	succinate				ethanol			
	#solutions	growth (h^{-1})	min. prod. (mmol/gDW/h)	max. prod. (mmol/gDW/h)	#solutions	growth (h^{-1})	min. prod. (mmol/gDW/h)	max. prod. (mmol/gDW/h)
152	1255	0.8729	6.8838	9.9430	21	0.1238	38.0668	38.0986
204	186	0.2023	21.1032	24.3682	37	0.1238	38.0668	38.0986
272	271	0.1908	22.2077	25.0055	52	0.1238	38.0668	38.0986
342	172	0.1549	25.0252	25.0660	84	0.1187	38.1716	38.2466

large carbon numbers are very important in redirecting flux towards succinate, although they may not carry high flux values. For example, both reactions PDH and PFL acting on a high carbon compound, i.e., Acetyl-CoA, carry a good flux value in wild type strains. The knockout of them together with TALA and LDH_L under anaerobic condition predicts high succinate production by our method. This prediction agrees well with *in vivo* studies [Hong and Lee, 2002; Zhu *et al.*, 2007]. It is also observed that many solutions are found by NIHBA, and all of them renders a growth-coupled production phenotype. Particularly, there exist a vast number of growth-coupled solutions with low succinate production rate, as indicated when the size of candidate size is $n_s = 152$. Most of them do not appear in larger candidate sets, implying that NIHBA prefers high-production solutions. In the following, a set of 342 candidate reactions is used for NIHBA.

In contrast to succinate, ethanol can be easily produced at a high rate (e.g., 38mmol/gDW/h, or equivalently 95% TMP) by knocking down only reactions of low carbon number. Therefore, excluding as many reactions of high carbon number is beneficial for computational efficiency while achieving a high-production design strategy. The number of growth-coupled solutions for ethanol is, however, much fewer than that for succinate.

Next, we analyse the performance of NIHBA on a varying number of knockouts. We observe that, for both succinate and ethanol, the production rate increases sharply when more knockouts are allowed but levels out from five knockouts (Fig.2A and B). The number of growth-coupled solutions and the percentage of high-production solutions increase with the number of knockouts. No solutions with ¿80% TMP were found within 15 knockouts for succinate, and only a tiny portion of solutions has a production rate of ¡20% TMP for ethanol. This indicates again that NIHBA favours high-production solutions during the search.

We show in Fig. 3 the distribution of single knockouts of obtained solutions in different subsystems and their frequency in design solutions to understand which subsystems/knockouts are likely to be engineering targets. Specifically, all the solutions with ¿60% TMP identified from a limit of 8 knockouts were analysed in terms of knockout occurrence in different subsystems. We find that extracellular exchange reactions (mainly oxygen uptake), reactions from glycolysis, gluconeogenesis, pyruvate metabolism and pentose phosphate pathway are most likely knockout targets for the production of succinate and ethanol, which shows a good agreement with existing studies [Feist *et al.*, 2010; Tepper and Shlomi, 2010]. All the solutions suggest oxygen depletion (Fig. 3C), which is consistent with the common sense that the two products are best produced in anaerobic environments. It is observed that glucose-6-phosphate isomerase (PGI) appears frequently in design solutions for both products, and high succinate-producing strategies additionally favour ATP synthase (ATPS4rpp) (Fig. 3C). Similar results have been also reported in a recent study [Dinh *et al.*, 2018], where computationally intensive strain design tools were used to obtain a large collection of knockout designs.

The computational cost is low for NIHBA, as shown in Fig.2C and D. A short runtime (300∼600s) enables NIHBA to identify high-production solutions for both succinate and ethanol, indicating that our hybrid approach can quickly generate effective Benders cuts to reduce search

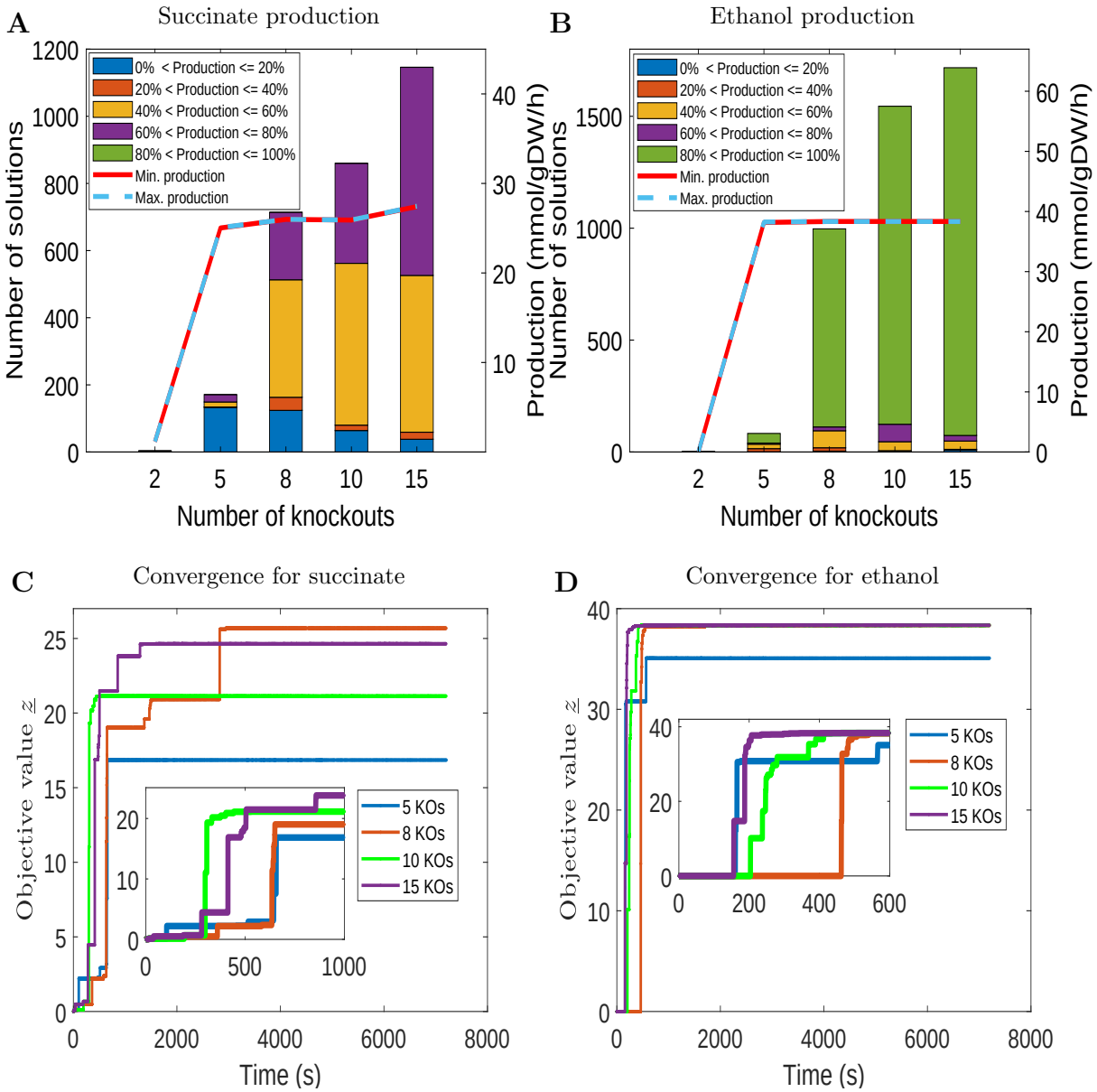

Figure 2: Performance of NIHBA with different number of knockouts for succinate and ethanol production. (**A**) For succinate, the number of solutions achieving a certain percentage of the maximum theoretic production rate, and the minimum and maximum production rate of the best solution found for a different number of knockouts. (**B**) For ethanol, the number of solutions achieving a certain percentage of the maximum theoretic production rate, and the minimum and maximum production rate of the best solution found for a different number of knockouts (KOs). (**C**) The objective value of NIHBA against runtime for succinate production. (**D**) The objective value of NIHBA against runtime for ethanol production.

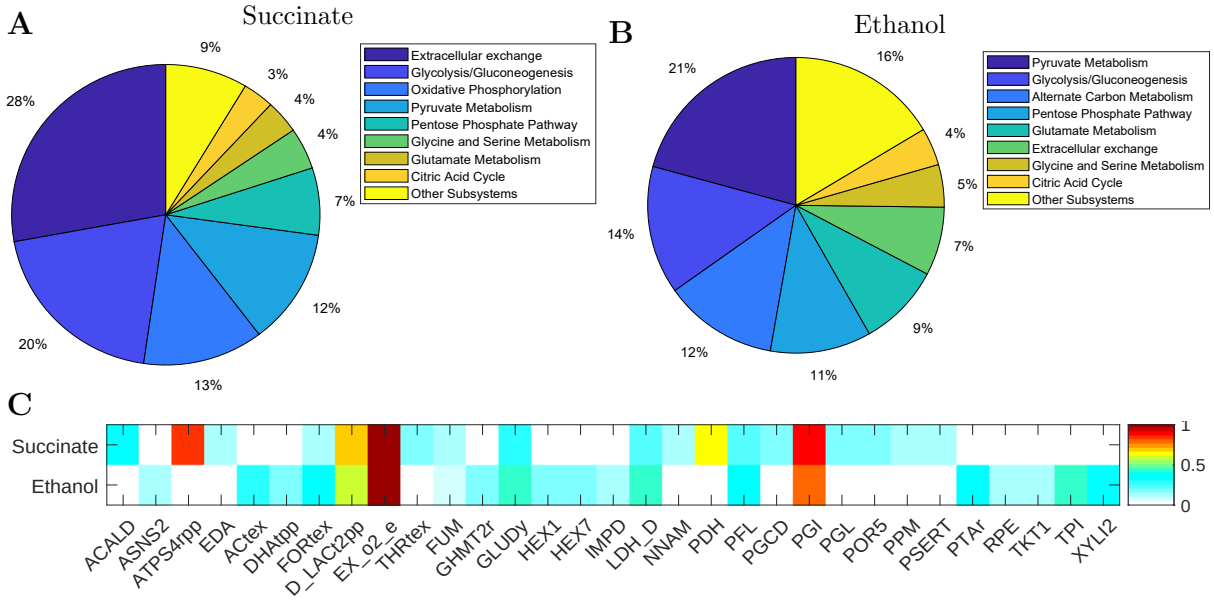

Figure 3: Knockout distribution for succinate and ethanol production. (**A**) The percentage of solutions with at least one knockout from each subsystem for succinate production. (**B**) The percentage of solutions with at least one knockout from each subsystem for ethanol production. (**C**) The fraction of design solutions that have a specific knockout (only knockouts that appear in ¿10% design solutions for either target products are displayed).

space. Depending on the maximum allowable number of knockouts, the runtime required before reaching the convergence stage is different, but with a small deviation. Generally, NIHBA starts to converge after ∼2000s and 600s for succinate and ethanol, respectively. The longer time required for succinate may be explained by relatively fewer high-production solutions in the design space. Despite that, the computational time required by NIHBA for a good solution is small (compared to days∼weeks in existing methods [Feist *et al.*, 2010]) and does not increase exponentially with the number of knockouts, a widely recognised issue in global search [Feist *et al.*, 2010; Lun *et al.*, 2009].

2.3 Comparison with other tools

For comparison, the network interdiction problem was also solved using the OptKnock [Burgard *et al.*, 2003] and GDLS [Lun *et al.*, 2009] approaches with Gurobi [Gurobi Optimization, 2018], called NI-OptKnock and NI-GDLS, respectively. NI-GDLS used $M = 5$ search paths and a search size of $k = 3$ in order to get multiple solutions. For efficiency, parameters in the Gurobi solver was set according to Egen and Lun [2012].

For succinate, when at most five knockouts are allowed, NIHBA found a large number of solutions whereas both NI-OptKnock and NI-GDLS failed to find a feasible solution. The failure is mainly due to numeric issues in Big-M formulation, which existed even although we switched to indicator constraints or CPLEX12.8 for MILP. This demonstrates that HBA overcomes such numeric issues. For readability, we only show a small number of selected solutions from NIHBA in the production envelope (Fig. 4A). As seen, NIHBA can obtain diverse solutions forming a good representative of the tradeoff between cell growth and succinate production. Interestingly, NIHBA identified a strong growth coupled design (non-zero production at no growth), despite its slightly suboptimal production rate at maximum growth.

For ethanol, all the algorithms found feasible solutions with at most five knockouts, and all contain a solution with maximum production rate, as illustrated in Fig. 4B–D (a small portion of solutions from NIHBA are displayed for readability). This shows that NIHBA has

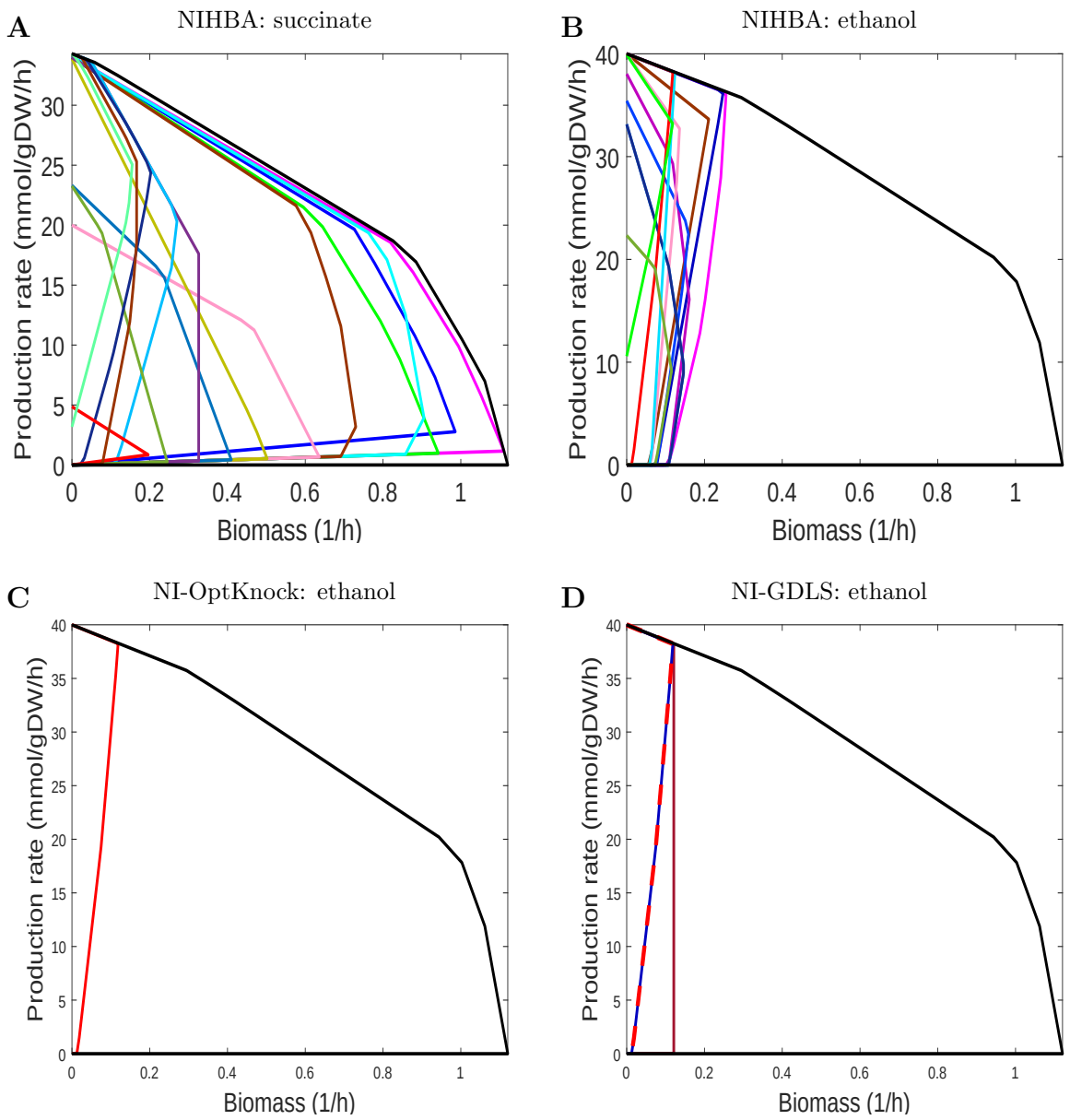

Figure 4: Production envelopes of diverse design solutions obtained by NIHBA for succinate and ethanol production. (**A**) Production envelopes of selected solutions for succinate production. (**B**) Production envelopes of selections solutions for ethanol production.

comparable performance in terms of optimality. Despite three solutions found from NI-GDLS, one of them is not growth coupled and the other two have the same production envelope, from which little can be gained about the tradeoff between cell growth and target production. In contrast, NIHBA found many solutions with diverse production envelopes, among which strong growth-coupled designs exist.

3 Discussion

The employment of bilevel optimisation for identifying genetic manipulations has been found very helpful for metabolic engineering. Most existing bilevel-based tools assume that cells always grow optimally, a biased optimisation principle that is found incorrect for mutants or certain microorganisms in some studies [Schuetz *et al.*, 2012; Segrè *et al.*, 2002]. As a consequence, design strategies found by these tools may be biologically infeasible in spite of highest production

rates at optimal growth. In addition, these tools involve solving a bilevel problem through big-M reformulation to a standard MILP that is suitable for commercial global search solvers like Gurobi and CPLEX. However, the resulting MILP is often large due to the genome scale of metabolic networks, and global search can be computationally prohibitive, particularly when a large design space (or numerous genetic manipulations) is allowed. Even if computationally tractable, the global search produces only a single solution. Furthermore, big-M formulation produces a weak MILP, leading to numeric issues such that no feasible solutions can be found. This paper have proposed to address biased assumptions from the point of view of game theory, leading to a network interdiction problem (NIP). The NIP is not handled using popular global search solvers, instead it is solved through an efficient hybrid Benders decomposition algorithm to lower computational costs and overcome numeric issues. The proposed approach, NIHBA, has shown its ability to obtain a large number of growth-coupled design strategies with diverse production phenotypes and achieve optimal production rates within an hour, regardless of the size of design space (the maximum allowable number of knockouts).

NIHBA uses a game theoretic framework to model the interaction (somehow competitive) between host cells and metabolic engineers. This eliminates the need for a biologically rigorous objective to reflect cells' primary intentions, such as optimal growth or maximum energy generation [Schuetz *et al.*, 2012]. Therefore, design strategies found by NIHBA do not necessarily yield the best production at optimal growth. Instead, they guarantee non-zero production when cell growth is over a predefined survival threshold. NIHBA employs a hybrid Benders algorithm, HBA, to solve the NIP. Our case studies have demonstrated numerous advantages of this algorithm. First, it is free of numeric issues, making it very more stable than other MILP solvers, e.g. Gurobi and CPLEX. Second, it can be considered a parameter-free algorithm as opposed to other methods like GDLS that requires a setting of multiple parameters, although NIHBA uses a parameter μ for identifying Pareto optimal cuts. In practice, NIHBA is not sensitive until a value of $\mu < 1e{-}6$ is used. Third, it is computationally efficient such that one hour on average is sufficient for NIHBA to identify high production solutions, and the runtime for a high production rate does not scale with the number of knockouts, which is not the case for existing methods. Last, it obtains numerous growth-coupled solutions in a single run. This is important as it not only helps understand the trade-off between target production and cell growth, but also provides the possibility to examine and test multiple solutions, from which the most promising design can be chosen for experimental implementation.

The proposed HBD is not limited to network interdiction problems. It can be applied to any bilevel or single-level optimisation problems that have complicating mixed-integer variables. Although promising, HBA needs improvements on convergence at late stages for optimality proof. Like other global search, an appropriate optimality gap or time limit may alleviate excessive exploration but cannot determine the optimality of solutions. Further improvements can be made along this direction to enhance the convergence of HBA. It is also noteworthy that multiple solutions by HBA are not searched in a systematic way. They may not form a perfect representative of the tradeoff between target production and cell growth. Therefore, more investigations are required to extract limited but well-diversified solutions in the search process of HBA.

Despite numerous solutions found by NIHBA, the selection of promising solutions poses a new challenge to decision makers. Therefore, it is therefore important to have a good solution ranking approach. Solutions may be roughly ranked according to the frequency of individual knockouts in addition to their subsystem distribution or by a scoring system with manual settings [Schneider and Klamt, 2019]. However, a more systematic solution ranking is desirable, and this is left in our future work.

4 Materials and Methods

4.1 Flux balance analysis

A metabolic network of m metabolites and n reactions has a stoichiometric matrix S that is formed by stoichiometric coefficients of the reactions. Let J be a set of n reactions and v_j the reaction rate of $j \in J$, Sv represents the concentration change rates of the m metabolites. FBA aims at optimising a linear biological objective $c^T v$ when the system is at steady state (i.e, the concentration change rate is zero for all the metabolites) and v is subject to thermodynamic constraints:

$$\begin{aligned} \max_{v} \quad & c^T v \\ s.t. \quad & Sv = 0 \\ & lb_j \leq v_j \leq ub_j, j \in J \end{aligned} \tag{1}$$

where lb_j and ub_j are the lower and upper flux bounds of reaction j, respectively. c is a weight vector specifying the degree of importance to the biological objective.

4.2 Network interdiction based strain design and reformulation

Network interdiction for strain design considers metabolic engineers as interdictors or adversaries who attempt to maximally disrupt host cells' activity that biochemicals of interest are not overproduced due to homeostasis. The strain design task can therefore be formulated as a max-min problem:

$$\max_{y \in Y} \ \min_{v} c_P v \tag{2a}$$
$$s.t. Sv = 0 \tag{2b}$$
$$lb_j(1 - y_j) \leq v_j \leq ub_j(1 - y_j), j \in \bar{J} \tag{2c}$$
$$lb_j \leq v_j \leq ub_j, j \in J \setminus \bar{J} \tag{2d}$$

where c_P is a coefficient vector for the target biochemical. $Y = \{y \in \{0,1\}^{|\bar{J}|} | \sum_{j \in \bar{J}} y_j < K\}$ (K is the maximum allowable number of knockouts), and y_j indicates the reaction j is inactive ($v_j = 0$) if $y_j = 1$ and active otherwise. $\bar{J}$ is a subset of J, containing candidate knockout reactions.

Observing that in the follow problem $|v_j| y_j = 0$ always holds for all $j \in J$, we can eliminate all the flux constraints imposed by y_j, i.e. eq.(2c), by rephrasing the inner objective function in a Lagrangian manner:

$$\min_{v} c_P v + \sum\nolimits_{j \in \bar{J}} M_j |v_j| y_j \tag{3a}$$
$$\left.\begin{aligned} s.t. \quad & Sv = 0 \\ & lb_j \leq v_j \leq ub_j, j \in J \end{aligned}\right\} \tag{3b}$$

where M_j is a large positive Lagrange multiplier and $M = (M_1, \ldots, M_{|J|})$. The reformulated follower problem is equivalent to the original problem in the sense that they have the same optimal value provided that M_j is sufficiently large for all $j \in J$ such that $v_j = 0$ when $y_j = 1$. The value of M_j used in this work is around 100 (e.g. randomly drawn from [90,110]).

The reformulated follower function (3) can be linearised by adding auxiliary variables $u_j = max(v_j, -v_j)$. As a result, we have the reformulated bi-level framework:

$$\max_{y \in Y} c_P v \tag{4a}$$
$$s.t. \ \min_{v,u} c_P v + \sum\nolimits_{j \in \bar{J}} M_j u_j y_j \tag{4b}$$
$$\left.\begin{aligned} s.t. \quad & Sv = 0 \\ & u_j \geq v_j, u_j \geq -v_j, j \in \bar{J} \\ & lb_j \leq v_j \leq ub_j, j \in J \end{aligned}\right\} \tag{4c}$$

4.3 Hybrid Benders Algorithm

The bi-level problem (4) is reformulated to a standard MILP by applying LP duality to the follower problem (4b)–(4c). For simplicity, the resulting MILP is written in the following compact form

$$\max_{y,x} \ \bar{c}x \tag{5a}$$
$$s.t. \sum\nolimits_{j\in\bar{J}} y_j < K \tag{5b}$$
$$\bar{S}x = 0 \tag{5c}$$
$$Ax + By \leq b \tag{5d}$$
$$y_j \in \{0,1\}, x \in R^{n+m} \times R^{5n}_{+}; \tag{5e}$$

where $\bar{S} = [S \ \ 0_{m,5n+m}]$, and

$$A=\begin{bmatrix} -I & 0 & 0 & 0 & 0 & 0 & 0 \\ I & 0 & 0 & 0 & 0 & 0 & 0 \\ I & 0 & -I & 0 & 0 & 0 & 0 \\ -I & 0 & -I & 0 & 0 & 0 & 0 \\ 0 & S^T & 0 & I & -I & I & -I \\ c_P & 0 & 0 & -lb^T & ub^T & I & I \end{bmatrix}, B=\begin{bmatrix} 0 \\ diag(U) \\ 0 \\ 0 \\ -diag(M) \\ 0 \end{bmatrix}, b=\begin{bmatrix} 0 \\ U \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{6}$$

where U is a vector of maximum absolute flux for each reaction, i.e., $U_j = max(|lb_j|, |ub_j|), \ \forall j \in J$.

The single-level reformulation (5) can be solved, like OptKnock, by modern MILP solvers. However, the big-M terms in (5) lead to a week LP relaxation [Codato and Fischetti, 2006], therefore causing difficulties for MILP solvers. Besides, the model size of (6) increases rapidly for large metabolic networks, and as a result, a large-scale MILP has to be solved.

Benders decomposition avoids these drawbacks as it can deal with complicating binary variables and easy continuous variables separately. Like Benders decomposition [Codato and Fischetti, 2006], our hybrid Benders algorithm (HBA) decomposes (5) into a BIP master problem (MP) (7) and a LP slave problem (SP) (8) for fixed $y = \bar{y}$:

$$\text{MP} : \bar{z} = \max_{y,z} \ \ z \tag{7a}$$
$$s.t. \ \ \sum\nolimits_{j\in\bar{J}} y_j < K \tag{7b}$$
$$\left.\begin{array}{ll} z \leq (b - By)^T\pi^o & \forall o \in O \\ (b - By)^T\pi^f \geq 0 & \forall f \in F \end{array}\right\} \text{Benders cuts} \tag{7c}$$
$$y_j \in \{0,1\}, z \geq 0; \tag{7d}$$

$$\text{SP} : \underline{z} = \max_{x} \ \ \bar{c}x \tag{8a}$$
$$s.t. \ \ \bar{S}x = 0 \qquad [\lambda] \tag{8b}$$
$$Ax \leq b - B\bar{y} \qquad [\pi] \tag{8c}$$

where O and F are sets that correspond to the extreme points π^o and extreme rays π^f of the dual of SP, respectively. In each iteration, the Benders decomposition algorithm derives the dual vector π from the SP (8) for $y(y = \bar{y})$ which is the solution to the MP in the previous iteration. In practice, a Benders cut is obtained by solving the dual of (8) rather than the primal. Two scenarios exist when solving the dual of (8): if the optimal value of the dual of SP is bounded, it means the SP is feasible, then an optimality cut $z \leq (b - By)^T\pi^o$ generated from the extreme point π^o is added to the MP; if it is unbounded, it means the SP is infeasible,

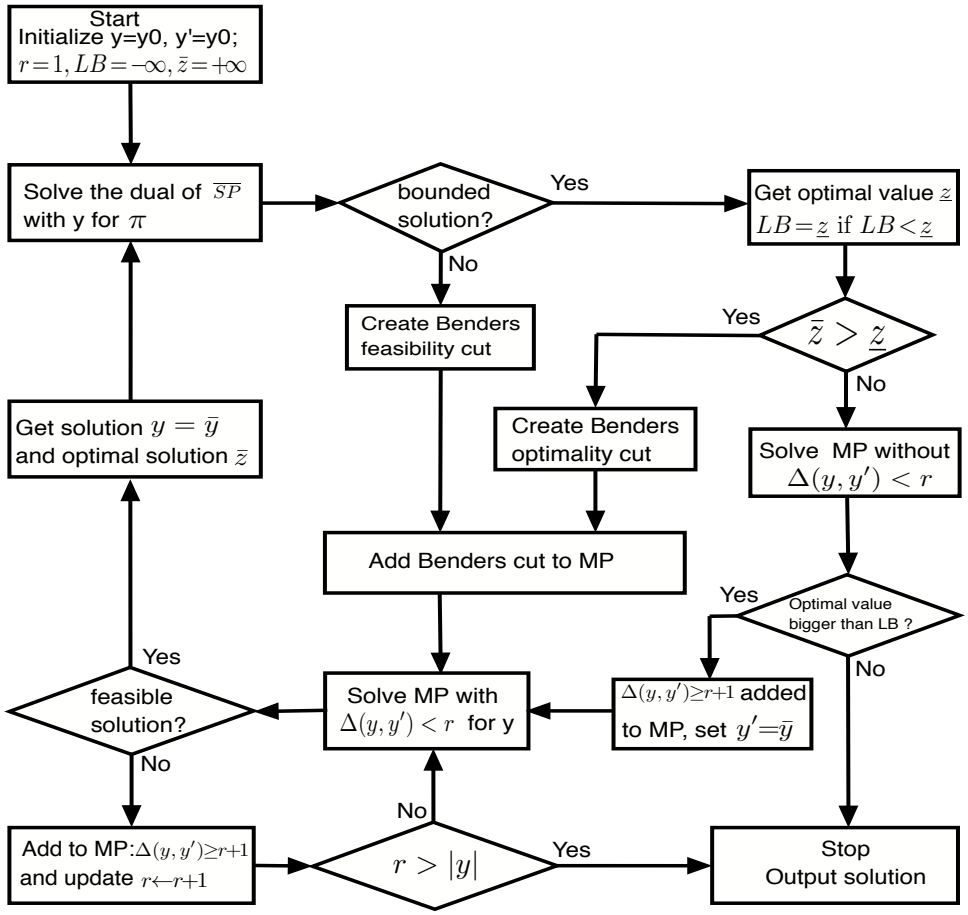

Figure 5: Flowchart of HBA.

then a feasibility cut $(b - By)^T \pi^f \geq 0$ generated from the extreme ray π^f is added to the MP to avoid unboundedness of the dual of SP in future iterations.

The classic Benders decomposition is not able to generate effective Benders cuts rapidly for our strain design problem, and therefore requires a huge of iterations (consequently long computation time) before it converges. Here, we introduce a hybrid Benders algorithm (HBA) with two strategies to speed up the convergence process.

Figure 5 shows a simplified flowchart of HBA. An implementation of the algorithm in MATLAB can be found in `https://sites.google.com/view/shouyongjiang/resources`.

4.3.1 Pareto optimal cuts

Let π^o be the dual vector of π corresponding to (8), a standard Benders optimality cut is:

$$Cut(\pi^o) : z \leq (b - By)^T \pi^o \tag{9}$$

Since π^o may not be unique, it is important to select an effective cut $Cut(\pi^o)$. Magnanti and Wong [1981] proposed to use Pareto optimal cuts to improve convergence. $Cut(\pi^o)$ is said to be Pareto optimal if no other $Cut(\bar{\pi}^o)$ exists such that $(b - By)^T \bar{\pi}^o \leq (b - By)^T \pi^o$ for any $y \in Y$ and at least one $y \in Y$ enables a strict inequality. There are a few methods available for identifying a Pareto optimal cut, but most of them have to solve the slave problem (8) twice, which may increase computational time significantly. We turn to the approach of Sherali and Lunday [2013] where a Pareto optimal cut can be generated by solving only once in each

iteration a slightly different slave problem:

$$\overline{SP} : \underline{z} = \max_{x} \quad \bar{c}x \tag{10a}$$

$$s.t. \quad \bar{S}x = 0 \qquad [\lambda] \tag{10b}$$

$$Ax \leq (b - B\bar{y}) + \mu(b - B\hat{y}) \qquad [\pi] \tag{10c}$$

where μ is a sufficiently positive value and $\hat{y}$ is a core point in the relative interior of the convex hull of Y. In this paper, $\hat{y}$ is updated by $0.5(\hat{y} + \bar{y})$ whenever a new feasible $y = \bar{y}$ is produced in the iteration of Benders decomposition. μ is not calculated as in Sherali and Lunday [2013] but rather fixed to 1e-8 after multiple trials.

4.3.2 Local branching

Another technique we used for accelerating the convergence of Benders decomposition is local branching, which is particularly effective when problems have binary variables [Baena *et al.*, 2018; Rei *et al.*, 2009]. Suppose $y^{'}$ is a feasible solution in Y, the idea behind local branching is to divide the feasible region of (7) into two subregions by the Hamming distance between y and y':

$$\Delta_H(y, y') = \sum\nolimits_{y'_j=0} y_j + \sum\nolimits_{y'_j=1} (1 - y_j) \tag{11}$$

In every iteration of Benders decomposition, the master problem (7) is solved in the subregion $\Delta_H(y, y') < r$ (where r is a positive integer and the maximum is the cardinality $|y|$ of y). This leads to two scenarios: there is either a feasible or infeasible y in the subregion $\Delta_H(y, y') < r$. If a feasible solution $\bar{y}$ is obtained, Benders cuts are generated by solving (10) with $\bar{y}$. If not, it means the value of r may be too small, and $\Delta_H(y, y') > r + 1$ is added to the master problem (7) to stop re-exploration in the neighbourhood of y' with the radius r. r is then increased by one at a time until $\Delta_H(y, y') < r$ renders the master problem (7) feasible. Note that y' has to be updated by $\bar{y}$ if $\bar{y}$ gives (7) an objective value worse than that of the slave problem (8), implying that no better solution can be obtained from the neighbourhood of y'.

4.4 Additional improvement strategies

HBA involves solving the MP (7) and SP (8) in a repeated manner. For efficiency, the following two strategies are used:

- Terminating the MP program prior to optimality. Suboptimal solutions to the MP are sufficient to generate valid Benders cuts. Therefore, the MP is terminated when a MIP Gap of $1 + 300/(iter^{0.5} + 1)$ (where $iter$ is the iteration counter) is reached.
- Reversing local branching whenever the $\bar{z}$ value of the MP is worse than $\underline{z}$ value of the SP. $\bar{z}$ estimates the upper bound of the problem (2). $\bar{z} < \underline{z}$ indicates the global optimum does not exist in the corresponding local branching and a reverse local branching should therefore be used.

4.5 Model reduction and candidate selection

The truncation of model size and candidate knockout set has great computational benefits. Genome-scale metabolic (GEM) models can be significantly simplified by compressing linear reactions and removing dead end reactions (those carrying zero fluxes). Likewise, many reactions can be excluded from consideration with *a priori* knowledge that, for example, they are vital for cell growth or their knockout is not likely to improve target production. We followed the model reduction procedure by Lun *et al.* [2009] and candidate selection procedure by Feist *et al.* [2010], resulting in a candidate set of 150∼350 reactions for different target products from the latest E. coli GEM iML1515 [Monk *et al.*, 2017] where the maximum uptake rates for glucose and oxygen are all 20 mmol/gDW/h.

4.6 Computational implementation

First of all, all the bi-level models were transformed into MILPs using corresponding reformulation techniques (see Supplementary Note A). Then, the resulting MILPs were implemented in MATLAB 2018b to be compatible with the Cobra Toolbox 3.0 [Heirendt *et al.*, 2018] where we carried out simulations. All the MILPs were solved by Gurobi 7.5 [Gurobi Optimization, 2018] with both Heuristics and MIPFocus were set to 1 as suggested by Egen and Lun [2012]. A time limit of 2 hours was applied to each MILP while performing computations on Ubuntu 16.04 LTS with an Intel® CoreTM i5 Quad Core processor.

Data and software availability

The data and software used and the tool developed are all available online:

- GEM model: iML1515 from BIGG database (bigg.ucsd.edu)
- Simulation software: Cobra toolbox 3.0 (https://opencobra.github.io/)
- Global solver: http://www.gurobi.com/
- NIHBA: https://sites.google.com/view/shouyongjiang/resources

Acknowledgements

This work has been supported by the Engineering and Physical Sciences Research Council (EPSRC) for funding project "Synthetic Portabolomics: Leading the way at the crossroads of the Digital and the Bio Economies (EP/N031962/1)".

Author contributions

SJ and NK conceived the project. SJ developed the computational method and performed the computational simulations and analysis of the results. SJ wrote the draft of the paper. All authors read, edited, and approved the final manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Iñigo Apaolaza, Luis Vitores Valcarcel, and Francisco J Planes. gMCS: Fast computation of genetic minimal cut sets in large networks. *Bioinformatics*, 35(3):535–537, 2018.

Daniel Baena, Jordi Castro Pérez, and Antonio Frangioni. Stabilized benders methods for large-scale combinatorial optimization, with application to data privacy. *Annals of Operations Research*, 271(1):11–85, 2018.

Ana Rita Brochado, Claudia Matos, Birger L. Møller, Jørgen Hansen, Uffe H. Mortensen, and Kiran Raosaheb Patil. Improved vanillin production in baker's yeast through in silico design. *Microbial Cell Factories*, 9(1):84, Nov 2010.

Anthony P. Burgard, Priti Pharkya, and Costas D. Maranas. Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and Bioengineering*, 84(6):647–657, 2003.

Hyung Seok Choi, Sang Yup Lee, Tae Yong Kim, and Han Min Woo. In silico identification of gene amplification targets for improvement of lycopene production. *Applied and Environmental Microbiology*, 76(10):3097–3105, 2010.

Yee Wen Choon, Mohd Saberi Mohamad, Safaai Deris, Chuii Khim Chong, Sigeru Omatu, and Juan Manuel Corchado. Gene knockout identification using an extension of bees hill flux balance analysis. *BioMed Research International*, 2015:124537, mar 2015.

Gianni Codato and Matteo Fischetti. Combinatorial benders' cuts for mixed-integer linear programming. *Operations Research*, 54(4):756–766, 2006.

Hoang V. Dinh, Zachary A. King, Bernhard O. Palsson, and Adam M. Feist. Identification of growth-coupled production strains considering protein costs and kinetic variability. *Metabolic Engineering Communications*, 7:e00080, dec 2018.

Dennis Egen and Desmond S. Lun. Truncated branch and bound achieves efficient constraint-based genetic design. *Bioinformatics*, 28(12):1619–1623, 04 2012.

Zia Fatma, Hassan Hartman, Mark G. Poolman, David A. Fell, Shireesh Srivastava, Tabinda Shakeel, and Syed Shams Yazdani. Model-assisted metabolic engineering of escherichia coli for long chain alkane and alcohol production. *Metabolic Engineering*, 46:1 – 12, 2018.

Adam M. Feist, Daniel C. Zielinski, Jeffrey D. Orth, Jan Schellenberger, Markus J. Herrgard, and Bernhard Ø. Palsson. Model-driven evaluation of the production potential for growth-coupled products of escherichia coli. *Metabolic Engineering*, 12(3):173–186, 2010.

LLC Gurobi Optimization. Gurobi optimizer reference manual, 2018.

Laurent Heirendt, Sylvain Arreckx, Thomas Pfau, Sebastián N Mendoza, Anne Richelle, Almut Heinken, Hulda S Haraldsdóttir, Jacek Wachowiak, Sarah M Keating, Vanja Vlasov, et al. Creation and analysis of biochemical constraint-based models: the cobra toolbox v3. 0. *Nat. Protoc.*, 14(3):639–702, 2018.

S. Hong and S. Lee. Importance of redox balance on the production of succinic acid by metabolically engineered escherichia coli. *Applied Microbiology and Biotechnology*, 58(3):286–290, Mar 2002.

Shouyong Jiang, Marina Torres Anaya, David Pelta, Preben Krabben, Marcus Kaiser, and Natalio Krasnogor. Improving microbial strain design via multiobjective optimisation and decision making. In *Proc. 27th Int. Joint Conf. Art. Intell. & 23rd Euro. Conf. Art. Intell. (IJCAI-ECAI)*, pages 1–10, 08 2018.

Zachary A. King and Adam M. Feist. Optimal cofactor swapping can increase the theoretical yield for chemical production in escherichia coli and saccharomyces cerevisiae. *Metabolic Engineering*, 24:117 – 128, 2014.

Nathan E Lewis, Harish Nagarajan, and Bernhard O Palsson. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology*, 10(4):291, 2012.

Churlzu Lim and J Cole Smith. Algorithms for discrete and continuous multicommodity flow network interdiction problems. *IIE Transactions*, 39(1):15–26, 2007.

Desmond S Lun, Graham Rockwell, Nicholas J Guido, Michael Baym, Jonathan A Kelner, Bonnie Berger, James E Galagan, and George M Church. Large-scale identification of genetic design strategies using local search. *Molecular Systems Biology*, 5(1):296, 2009.

T. L. Magnanti and R. T. Wong. Accelerating benders decomposition: Algorithmic enhancement and model selection criteria. *Operations Research*, 29(3):464–484, 1981.

Paulo Maia, Miguel Rocha, and Isabel Rocha. In silico constraint-based strain optimization methods: the quest for optimal cell factories. *Microbiol. Mol. Biol. Rev.*, 80(1):45–67, 2016.

Jonathan M Monk, Colton J Lloyd, Elizabeth Brunk, Nathan Mih, Anand Sastry, Zachary King, Rikiya Takeuchi, Wataru Nomura, Zhen Zhang, Hirotada Mori, et al. iML1515, a knowledgebase that computes escherichia coli traits. *Nature biotechnology*, 35(10):904–908, 2017.

Kiran Raosaheb Patil, Isabel Rocha, Jochen Förster, and Jens Nielsen. Evolutionary programming as a platform for in silico metabolic engineering. *BMC bioinformatics*, 6(1):308, 2005.

Priti Pharkya and Costas D. Maranas. An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Metabolic Engineering*, 8(1):1–13, jan 2006.

Priti Pharkya, Anthony P Burgard, and Costas D Maranas. OptStrain: a computational framework for redesign of microbial production systems. *Genome research*, 14(11):2367–2376, nov 2004.

Aditya Pratapa, Shankar Balachandran, and Karthik Raman. Fast-SL: An efficient algorithm to identify synthetic lethal sets in metabolic networks. *Bioinformatics*, 31(20):3299–3305, 2015.

Sridhar Ranganathan, Patrick F. Suthers, and Costas D. Maranas. OptForce: An optimization procedure for identifying all genetic manipulations leading to targeted overproductions. *PLoS Computational Biology*, 6(4):e1000744, apr 2010.

Walter Rei, Jean-François Cordeau, Michel Gendreau, and Patrick Soriano. Accelerating benders decomposition by local branching. *INFORMS Journal on Computing*, 21(2):333–345, 2009.

Isabel Rocha, Paulo Maia, Pedro Evangelista, Paulo Vilaça, Simão Soares, José P Pinto, Jens Nielsen, Kiran R Patil, Eugénio C Ferreira, and Miguel Rocha. OptFlux: an open-source software platform for in silico metabolic engineering. *BMC systems biology*, 4(1):45, 2010.

Philipp Schneider and Steffen Klamt. Characterizing and ranking computed metabolic engineering strategies. *Bioinformatics*, 01 2019.

Robert Schuetz, Nicola Zamboni, Mattia Zampieri, Matthias Heinemann, and Uwe Sauer. Multidimensional optimality of microbial metabolism. *Science*, 336(6081):601–604, 2012.

Daniel Segrè, Dennis Vitkup, and George M. Church. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences*, 99(23):15112–15117, 2002.

Hanif D. Sherali and Brian J. Lunday. On generating maximal nondominated benders cuts. *Annals of Operations Research*, 210(1):57–72, Nov 2013.

Yuekai Sun, Ronan MT Fleming, Ines Thiele, and Michael A Saunders. Robust flux balance analysis of multiscale biochemical reaction networks. *BMC bioinformatics*, 14(1):240, 2013.

N. Tepper and T. Shlomi. Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways. *Bioinformatics*, 26(4):536–543, feb 2010.

Axel von Kamp and Steffen Klamt. Enumeration of smallest intervention strategies in genome-scale metabolic networks. *PLoS computational biology*, 10(1):e1003378, 2014.

Peng Xu, Sridhar Ranganathan, Zachary L Fowler, Costas D Maranas, and Mattheos AG Koffas. Genome-scale metabolic network modeling results in minimal interventions that cooperatively force carbon flux towards malonyl-coa. *Metabolic engineering*, 13(5):578–587, 2011.

Y. Zhu, M. A. Eiteman, K. DeWitt, and E. Altman. Homolactate fermentation by metabolically engineered escherichia coli strains. *Applied and Environmental Microbiology*, 73(2):456–464, 2007.