NetFlow: A tool for isolating carbon flows in genome-scale metabolic networks

Sean G. Mack, Ganesh Sriram [*]

Department of Chemical and Biomolecular Engineering, University of Maryland, College Park, MD, USA

1. A B S T R A C T

Genome-scale stoichiometric models (GSMs) have been widely utilized to predict and understand cellular metabolism. GSMs and the flux predictions resulting from them have proven indispensable to fields ranging from metabolic engineering to human disease. Nonetheless, it is challenging to parse these flux predictions due to the inherent size and complexity of the GSMs. Several previous approaches have reduced this complexity by identifying key pathways contained within the genome-scale flux predictions. However, a reduction method that overlays carbon atom transitions on stoichiometry and flux predictions is lacking. To fill this gap, we developed NetFlow, an algorithm that leverages genome-scale carbon mapping to extract and quantitatively distinguish biologically relevant metabolic pathways from a given genome-scale flux prediction. NetFlow extends prior approaches by utilizing both full carbon mapping and context-specific flux predictions. Thus, NetFlow is uniquely able to quantitatively distinguish between biologically relevant pathways of carbon flow within the given flux map. NetFlow simulates ^{13}C isotope labeling experiments to calculate the extent of carbon exchange, or carbon yield, between every metabolite in the given GSM. Based on the carbon yield, the carbon flow to or from any metabolite or between any pair of metabolites of interest can be isolated and readily visualized. The resulting pathways are much easier to interpret, which enables an in-depth mechanistic understanding of the metabolic phenotype of interest. Here, we first demonstrate NetFlow with a simple network. We then depict the utility of NetFlow on a model of central carbon metabolism in *E. coli*. Specifically, we isolated the production pathway for succinate synthesis in this model and the metabolic mechanism driving the predicted increase in succinate yield in a double knockout of *E. coli*. Finally, we describe the application of NetFlow to a GSM of lycopene-producing *E. coli*, which enabled the rapid identification of the mechanisms behind the measured increases in lycopene production following single, double, and triple knockouts.

1. Introduction

Genome-scale models of metabolism (GSMs) seek to represent the entirety of metabolic processes occurring within a species' cells (O'Brien et al., 2013). GSMs enable the direct probing and analysis of cellular metabolism, which has directly led to advances in metabolic engineering (Curran et al., 2012), cancer biology (Folger et al., 2011), and many other fields (Huang et al., 2017a; Brynildsen et al., 2013; Ramirez et al., 2017). Analysis of GSMs typically involves the prediction of reaction activities or fluxes through the network, usually through the application of flux balance analysis (FBA) (Orth et al., 2010) and its extensions (Holzhütter, 2004). The resulting flux vectors are often difficult to parse due to the complexity of the GSMs, which commonly contain thousands of reactions and metabolites.

A potential method to combat this complexity would be to extract key pathways from genome-scale flux predictions for further analysis. Many algorithms in the field of metabolic pathfinding have been developed for identifying pathways between metabolites within a metabolic network (Kim et al., 2017). These algorithms generally fall into two categories: graph-based approaches (Croes et al., 2005, 2006; Kim et al., 2020; Simeonidis et al., 2003), which utilize the connectivity between metabolites (nodes) as determined by the reactions (edges) in the network, and stoichiometric approaches (Pharkya et al., 2004; Klamt et al., 2017; Chowdhury and Maranas, 2015; Pey et al., 2011), which use optimization to predict stoichiometrically-balanced pathways to generate a given metabolite. A critical challenge for both approaches is limiting the identification of false-positive pathways due to a few highly connected metabolites, e.g. ATP.

Recently, algorithms that consider atom tracking between metabolites have been developed to ensure the identified pathways are biologically meaningful (Tervo and Reed, 2016; Heath et al., 2010; Pey et al., 2014; Huang et al., 2017b). In a notable effort (Pey et al., 2014), Pey et al. extended their previous work on carbon flux paths (Pey et al., 2011) to incorporate complete carbon atom mapping matrices (AMMs) into their mixed-integer linear programming (MILP) optimization scheme for predicting the shortest pathway between metabolites (Pey et al., 2014). The recruitment of AMMs ensured that each step in the pathway contributed carbon toward the product metabolite. Their method used stoichiometry and steady-state mass balances to constrain the set of reactions included in the predicted pathway. Critically, this method does not include environmental or biological constraints such as extracellular flux measurements or biomass production.

This limitation was addressed in the work of Tervo and Reed (2016), whose PathTracer algorithm directly incorporated environmental and biological constraints into the optimization (Tervo and Reed, 2016). This algorithm also utilized FBA predictions to identify and weigh active reactions, thus guaranteeing the relevance of predicted pathways. This work combined optimization with carbon transfer maps. These maps indicate whether *any* carbon is exchanged between metabolites; thus, they do not contain as much information as full AMMs. Therefore, employing carbon transfer maps determines only whether carbon is transferred or not between metabolites. This information, while useful, is inherently qualitative. Thus, pathways extracted by this method can only be differentiated based on pathway length and relative activity. This can be particularly confounding in case of genome-scale flux distributions that can contain many long, looping pathways. The accurate mechanistic interpretation of fluxes in these pathways and the determination of their contribution to product metabolites will require the incorporation of full AMMs. To remedy these limitations, the algorithms reported by Pey et al. (2014) and Tervo and Reed (2016) should be extended by distinguishing pathways based on the number of carbons transferred, or "carbon yield".

Toward this goal we developed NetFlow, an algorithm that identifies and isolates carbon flows of interest in a fully carbon-mapped GSM. From a given flux distribution, NetFlow simulates ^{13}C isotope label experiments (ILEs) for the purpose of determining connectivity and carbon yield between each metabolite in the network. With these data, NetFlow extracts a subnetwork containing only the biologically relevant pathways that generate or consume a given metabolite or set of metabolites of interest. A requirement for the application of NetFlow to genome-scale networks is the availability of suitable AMMs, which historically were limited to small and medium-sized networks. However, recent advancements in atom mapping algorithms (Kumar et al., 2014; Hadadi et al., 2017; Litsa et al., 2019; Latendresse et al., 2012) have led to the generation of accurate, genome-scale carbon maps for GSMs (Preciat Gonzalez et al., 2017; McCloskey et al., 2016; Gopalakrishnan and Maranas, 2015). While these carbon-mapped GSMs have mainly been used for ^{13}C MFA simulations (McCloskey et al., 2018a, 2018b, 2018c), they also afford the development and expansion of constraint-based pathfinding algorithms such as NetFlow.

Here, we describe the NetFlow algorithm in detail and demonstrate its utility through two case studies in *Escherichia coli*. In the first case study, we illustrate the utility of NetFlow to analyze flows through a central carbon metabolism model of *E. coli*. In the second case study, we used NetFlow to compare the wild-type and triple knockout (KO) flux predictions reported in Alper et al. (2005) on *E. coli* engineered to overproduce lycopene. This isolated the main lycopene production pathway and elucidated the previously unknown mechanism by which the triple KO improved lycopene production.

2. Methods

2.1. Steady state ^{13}C ILE simulations

Steady-state ^{13}C ILE simulations were performed by using custom

MATLAB scripts and the cumomer balancing method (Wiechert et al., 1999). The workflow for each ILE simulation was as follows. First, the list of cumomers was first generated from the set of carbon-containing metabolites. Cumomer mapping matrices (CMMs) for each reaction were then generated from the carbon atom mapping matrices (AMMs) in the model. Just as AMMs map each reactant carbon atom to a corresponding product atom, CMMs map each reactant cumomers to corresponding product cumomers (Wiechert et al., 1999). By choosing an appropriate labeling state for each input metabolite, the steady-state cumomer balances can be solved to calculate the steady-state labeling of each cumomer. For this work, only first-order cumomers were required. Thus, the balance equation is simplified to:

$$\mathbf{A} \cdot \mathbf{x}_\infty + \mathbf{b} = \mathbf{0} \qquad [1]$$

Here, $\mathbf{x}_\infty$ is the steady-state first-order cumomer labeling, $\mathbf{A}^1$ is the first-order cumomer balance array, and $\mathbf{b}$ is the first-order input cumomer vector.

2.2. NetFlow implementation

NetFlow identifies carbon flows to and from all active metabolites in a metabolic network of arbitrary size and complexity, for a given flux state. It also enables the extraction of subnetworks containing only the carbon flows of interest. The inputs to NetFlow are a metabolic network, complete carbon mapping information for this network, and a predicted net flux vector. From these inputs, NetFlow simulates a series of ^{13}C ILEs on the given network and flux vector. We note that NetFlow takes for granted the reaction directionalities and cycles contained within a given flux vector. Therefore, variation in reaction reversibility and cycling, which are important aspects in simulating or analyzing ILEs, are not a concern in NetFlow.

Across the series of ILE simulations, the uniformly ^{13}C-labeled (U-^{13}C) version of every metabolite i is assumed to be input to the network. The ensuing extent of labeling of each metabolite j at steady state (EOL_{ij}) is then simulated:

$$EOL_{ij} = \frac{\sum_k X_{ki}}{N_j^C}, \ k \in C_j; \ i, j \in M_{act} \qquad [2]$$

Here, X_{ki} is the steady-state labeling of first-order cumomer k when metabolite i is supplied in U–^{13}C form. Only the first-order cumomers of metabolite j are considered. These cumomers are a subset of the first-order cumomer vector $\mathbf{X}$. The set C_j is the set of row indices in $\mathbf{X}$ corresponding to the first-order cumomers for metabolite j. The number of carbons, or the number of first-order cumomers, of metabolite j, is N_j^C. The set of active metabolites, defined as metabolites participating in any reaction carrying nonzero flux, is M_{act}.

In each simulation, ^{13}C flows from U–^{13}C-labeled metabolite i to each downstream metabolite j whose carbon atoms are wholly or partially derived from metabolite i. The value of EOL_{ij}, which can range from 0 to 1, represents the extent to which the carbon atoms of metabolite j are derived from metabolite i or carbon yield of metabolite j from metabolite i. For example, this workflow can be applied calculate the extent of carbon flow EOL_{ij} from glucose (i) to glycine (j) as follows. For glycine, we have the first-order cumomers [1X] and [X1], whose row indices in $\mathbf{X}$ are represented by the set C_j, and $N_j^C = 2$. If U–^{13}C glucose is input to the network, then at steady state:

$$EOL_{ij} = \frac{[1X] + [X1]}{2} \qquad [3]$$

From the complete set of EOL_{ij}, NetFlow forms the matrix $\mathbf{EOL} \in \mathbb{R}^{m \times m}$, which contains the carbon yield between every metabolite in the network for the given flux vector. A binary metabolite mapping matrix $\mathbf{L} \in \mathbb{R}^{m \times m}$ is also created that represents the connectivity between

all the active metabolites in the network:

$$L_{ij} = (EOL_{ij} > 0) \; \forall L_{ij} \in [0, 1] \qquad [4]$$

By using these matrices, NetFlow can identify which metabolites are upstream (carbon sources) and downstream (carbon sinks) compared to any given metabolite in the network. Specifically, for a metabolite of interest k, the nonzero entries in the kth row of the matrix ($\mathbf{L}_{k*}$) correspond to downstream metabolites that contain carbon from metabolite k. Likewise, the nonzero entries in the kth column of the matrix ($\mathbf{L}_{*k}$) correspond to upstream metabolites that contribute carbon to metabolite k. Similarly, the row $\mathbf{EOL}_{k*}$ of the matrix $\mathbf{EOL}$ contains the carbon yields from metabolite k and the column $\mathbf{EOL}_{*k}$ contains the carbon yields to metabolite k. Since all metabolites contribute carbon to themselves, the diagonal entries of both $\mathbf{EOL}$ and $\mathbf{L}$ are always one.

For the toy network shown in Fig. 1a, the mapping and yield matrices generated by NetFlow are shown in Fig. 1b-c. These matrices allow a user to isolate the carbon flow to or from any given metabolite or between any two given metabolites in the network. For example, the metabolites that contain carbon derived from metabolite C are captured in the non-zero entries in the row of $\mathbf{L}$ corresponding to C ($\mathbf{L}_{C*}$):

$$\mathbf{L}_{C*} = \begin{bmatrix} & A & B & C & D & E & F & CO_2 \\ C & 0 & 0 & 1 & 1 & 0 & 1 & 0 \end{bmatrix} \qquad [5]$$

Thus, metabolite C contributes carbon only to metabolites D and F. Similarly, the metabolites that contribute carbon to E can be identified by the non-zero entries in the column of $\mathbf{L}$ corresponding to E ($\mathbf{L}_{*E}$), which correspond to the metabolites A, B, and CO_2. Additionally, the production pathway(s) for E could be isolated by eliminating D, C, and F as they do not contribute carbon to E (i.e. $\mathbf{L}_{*E} = 0$) (Fig. 2a). By incorporating the carbon yields of each connected metabolite, the low yield of E from A becomes apparent, which corresponds with the indirect path from carbons in A to E (via CO_2).

The matrices $\mathbf{EOL}$ and $\mathbf{L}$ can also be used to identify continuous carbon flow between a given pair of upstream and downstream metabolites. Specifically, the intermediate metabolites between this pair of metabolites can be identified by finding the overlapping metabolites in the corresponding row and column of the mapping matrix $\mathbf{L}$. For example, the internal flow of carbon from B to D can be isolated by finding the overlap between the non-zero rows of $\mathbf{L}_{*D}$ (A, B, C, D, and CO_2) and the non-zero columns of $\mathbf{L}_{B*}$ (B, C, D, E, F, and CO_2). The intersection of these sets of metabolites yields the set of metabolites

between (and including) B and D (M_{BD}^{int}), which consists of B, C, D, and CO_2. This process can be generalized for any pair of metabolites i and j:

$$M_{ij}^{int} = \mathbf{L}_{i*} > 0 \wedge \mathbf{L}_{*j} > 0 \qquad [6]$$

Once the continuous set of connected metabolites (M_{ij}^{int}) are identified, the subnetwork containing only those metabolites can be extracted from the original model (Fig. 2b).

2.3. Metabolic network models and flux predictions

To illustrate NetFlow, we used a small "toy" network containing 8 metabolites and 10 reactions (Fig. 1a). We assumed an arbitrary, stoichiometrically feasible flux vector for this network. We used two models for *E. coli*—a central carbon metabolism (CCM) model and a GSM. Both the CCM model and GSM were derived from a fully carbon-mapped GSM of *E. coli* (iDM2014) (McCloskey et al., 2016). The CCM model contains glycolysis (EMP), pentose phosphate pathway (PPP), tricarboxylic acid (TCA) cycle, glyoxylate shunt, and various anaplerotic and fermentation reactions **(Supplemental File 1)**. To avoid challenges with balancing, we left out secondary metabolites and energetic intermediates from the CCM model. Additionally, the model does not contain a biomass production reaction to avoid dual optimization of growth and succinate production. The glucose uptake rate was defined as 10 mmol/gDWhr. To ensure reaction activity across a variety of pathways, we forced key reaction splits to have a specified flux ratio (e.g. EMP:PPP = 4:1). Flux vectors were generated by maximizing succinate efflux using FBA (Orth et al., 2010) from the COBRA toolbox 3.0 (Heirendt et al., 2019). All reaction KOs were simulated by setting the upper bound of the corresponding reaction to zero.

The GSM of *E. coli* engineered to overproduce lycopene **(Supplemental File 2)** was created by adding the four missing reactions of the lycopene production pathway identified by Alper et al. (Alper et al., 2005) (GGDPS, PYS, PDS, and lycopene exchange). Glutamate dehydrogenase (GLUDy), which is encoded by *gdhA*, was also added to the model to enable simulation of the triple KO used in Alper et al. (Alper et al., 2005). The carbon mapping for the added reactions was obtained from EcoCyc (Keseler et al., 2017). All simulations were performed in minimal glucose media as defined in the original publication (Alper et al., 2005). The WT flux prediction was generated by maximizing growth with FBA. Flux distributions in the KOs were generated by setting the
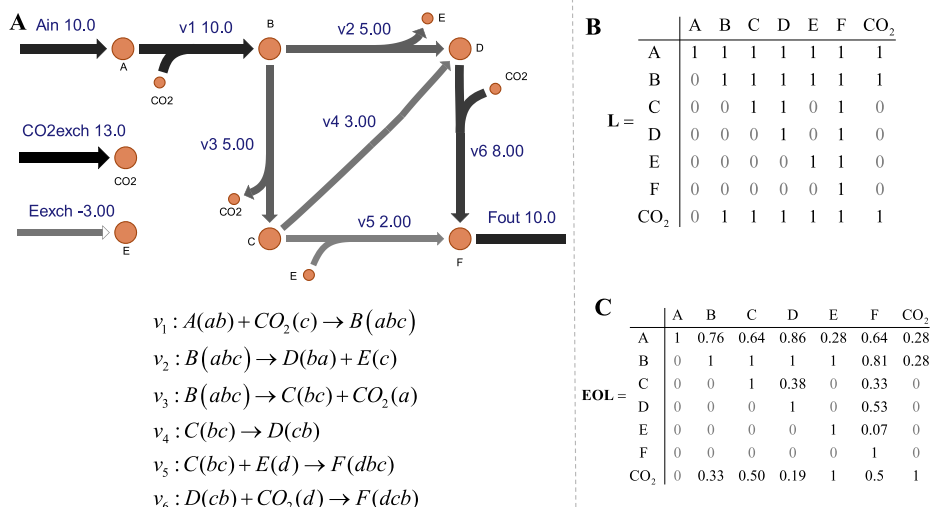

Fig. 1. Toy Network Model. (**A**) To demonstrate NetFlow, we used a toy network model with 6 reactions. In the carbon mapping shown, the letters representing the carbon atoms *a, b, c,* and *d* are maintained for all the reactions. Thus, "a" refers to the exact same carbon atom in reactions v_1, v_2, and v_3. NetFlow generated the (**B**) connectivity matrix $\mathbf{L}$ and (**C**) the yield matrix $\mathbf{EOL}$.
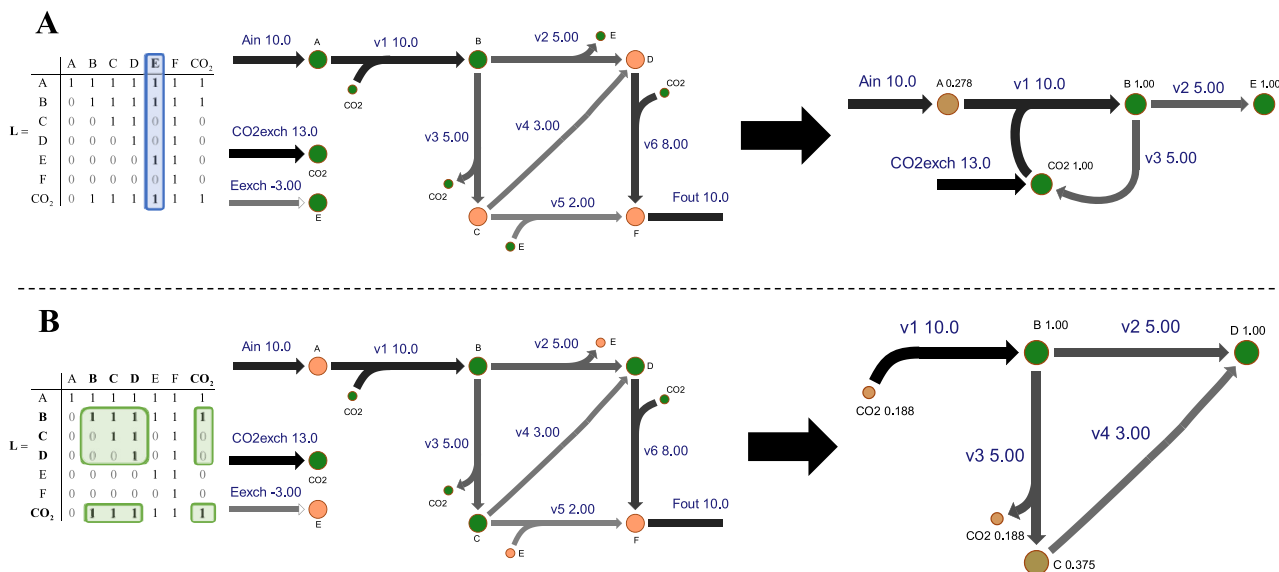
Fig. 2. Isolation of Carbon Flows by NetFlow. (**A**) The carbon flow to metabolite E was identified by finding the metabolites that contribute carbon to E (green nodes) via the corresponding row in **L** (highlighted in blue). The corresponding subnetwork was isolated and metabolite yield on E was overlain on the network. (**B**) The flow between metabolite B and D was identified by find the continuous set of connected metabolites between them (green nodes) via the corresponding entries in **L** (highlighted in green). The subnetwork was then isolated and metabolite yield on D was overlain on the network. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

corresponding reaction bounds to zero and running MOMA (Segrè et al., 2002) from the COBRA toolbox 3.0 (Heirendt et al., 2019).

All simulations were performed in MATLAB (The Mathworks, Natick, MA) on an AMD 1800X 3.60 GHz eight-core CPU using a single core. Network and flux visualizations were done with Escher (King et al. Escher, 2015). The MATLAB scripts underlying NetFlow are available at github.com/SriramLabUMD/NetFlow.

3. Results

3.1. Case study 1: succinate production in E. coli CCM

To demonstrate the utility of NetFlow on a real network, we applied the algorithm to a CCM model for *E. coli* (see **2.3**). NetFlow first isolated the carbon flow from glucose to succinate (Fig. 3). After all inactive
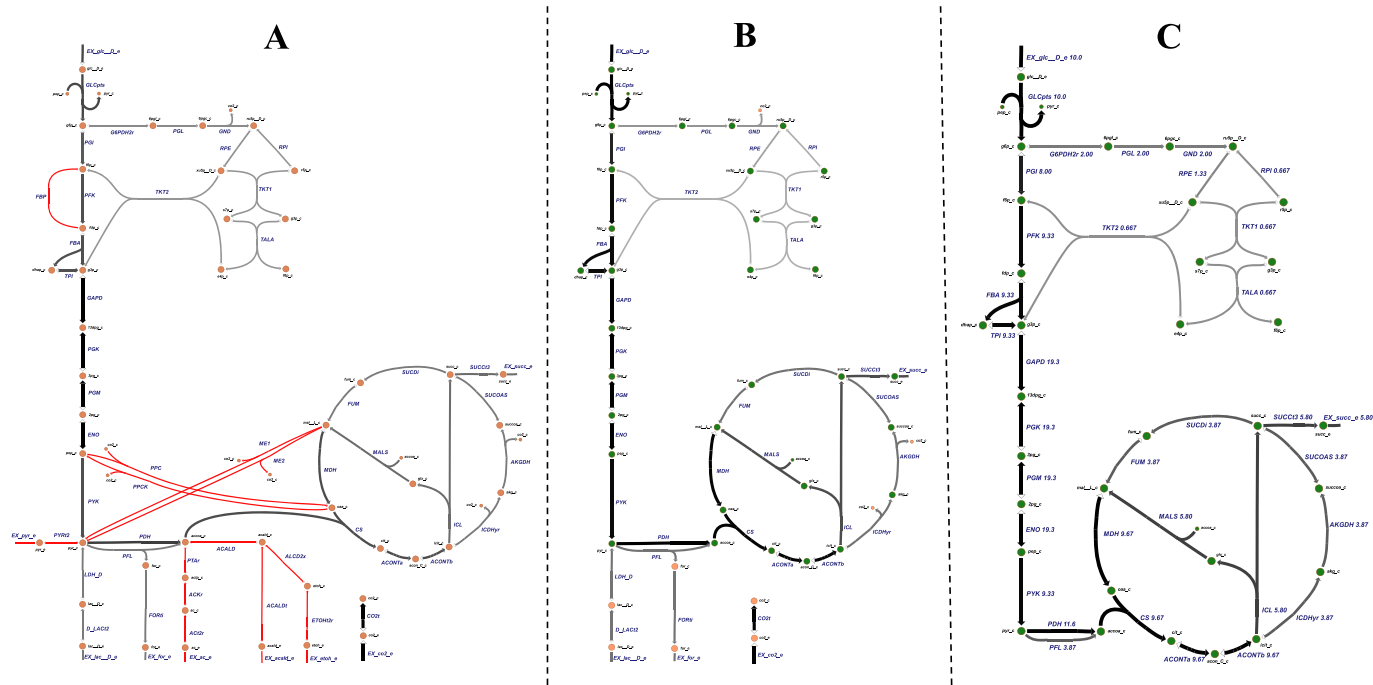

Fig. 3. Identification of Metabolites that Contribute Carbon toward Succinate Production. (**A**) NetFlow removed all inactive reactions from the network before simulating ILEs. (**B**) NetFlow identified the set of connected metabolites upstream of succinate (green nodes). (**C**) A subnetwork containing only the connected nodes to succinate was isolated. The full descriptions for all metabolites and reactions contained in these networks are given in Table S5 and Table S6, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

reactions were removed from the model (Fig. 3a), NetFlow enumerated all the carbon sources for succinate ($\mathbf{L}_{*succ} > 0$). These metabolites are represented as green nodes in Fig. 3b and include all metabolites except lactate, formate, and CO_2. A reduced subnetwork containing only succinate-contributing metabolites was then generated by removed these metabolites from the model (Fig. 3c).

The depth of information contained in the extent-of-labeling matrix **EOL** is evident in this network. For example, when only binary connectivity to succinate based on matrix **L** was considered, the upstream metabolites were indistinguishable. However, when carbon yields from matrix **EOL** were incorporated, it became clear that certain pathways contribute much more carbon toward succinate (Fig. 4a). For instance, EMP metabolites have a much higher yield than PPP intermediates. Two reason contribute to this contrast—the high flux through EMP relative to PPP and the carbon yield of each pathway. The CO_2 lost in phosphogluconate dehydrogenase (GND) represents one less carbon atom flowing toward succinate, which lowers the yield of related metabolites. The subnetwork was then trimmed to contain only high-yield metabolites by removing metabolites with yields below a chosen threshold (Fig. 4b-d). PPP was removed at a threshold of 0.25 (Fig. 4b) while alpha-ketoglutarate (AKG), succinyl-CoA (SUCCOA), fumarate (FUM), and glyoxylate (GLX) and then malate and oxaloacetate were removed at thresholds of 0.5 and 0.75, respectively (Fig. 4c-d). The removal of select TCA intermediates corresponds with inefficient loss of CO_2 through both isocitrate dehydrogenase (ICDH) and AKG dehydrogenase (AKGDH). The final pathway shown in Fig. 4d represents the most carbon efficient active pathway for producing succinate.

To illustrate the comparison of multiple flux distributions, the initial flux map was perturbed by knocking-out lactate dehydrogenase (LDH) and ICDH individually and simultaneously. These KOs were chosen by inspection as likely candidates for increased succinate production based on the initial flux map. The KO flux distributions were then run through NetFlow to compare succinate connectivity and yield in the perturbed states with that in the wild type (Tables S1–3). Unsurprisingly, the LDH KO did not alter the connectivity map as lactate did not contribute to succinate in the base flux. However, due to the increase in glyoxylate shunt activity, the yield from GLX increased by ~10% while the yield from AKG/SUCCOA and FUM decreased by ~17% and ~12%, respectively (Table S1). The ICDH KO blocked flux to AKG, SUCOA, and FUM, which eliminated them from the connectivity map. Flux through the glyoxylate shunt was again increased, which corresponded to a ~73% increase in succinate yield from GLX (Table S2). The connectivity map and metabolite yield in the double KO mirrored that of the ICDH KO (Table S3).

The impact of each KO was isolated by finding the intermediate metabolites between succinate and the reactants of the KO'd reactions (pyruvate for LDH and isocitrate for ICDH). For the LDH KO, the resulting subnetwork only contained pyruvate, GLX, and TCA cycle intermediates (Fig. 5a). The subnetwork elucidates that the KO of LDH increased flow of pyruvate into TCA cycle and through the glyoxylate shunt to succinate. The ICDH subnetwork was even further reduced to only include the glyoxylate shunt and half of the TCA cycle. As visualized in Fig. 5b, the ICDH KO forced flux through the glyoxylate shunt to succinate. Due to the additional demand on acetyl-CoA in malate synthase (MALS), the flux through the remaining portion of TCA was decreased. Finally, the double KO subnetwork encompasses the ICDH subnetwork with the addition of pyruvate and acetyl-CoA (ACOA). The mechanism arising from the combined KO was an amalgamation of the LDH and ICDH mechanisms. The KO of LDH forced additional pyruvate into ACOA, which enabled an even further increase in glyoxylate shunt activity due to the ICDH KO (Fig. 5c).

3.2. Case study 2: Increased lycopene production in triple KO E. coli

Motivated by the foregoing analysis of the double KO in a CCM model, we employed NetFlow to elucidate the unknown metabolic mechanism by which a previously identified triple gene KO (Alper et al., 2005) successfully increased lycopene production in *E. coli*. The original results were recreated using an updated, fully carbon mapped model for *E. coli*, modified to produce lycopene (see **Methods** for details). The updated predictions for the optimal single, double, and triple KOs tested by Alper et al. (2005) qualitatively matched their original measurements (Table 1). The single KO of *gdhA* and the double KO of *gdhA* and *aceE* modestly increased the production of lycopene as previously reported. Critically, the triple KO of *gdhA*, *aceE*, and *fdhF* yielded the highest lycopene production with a minimal decrease in growth, mirroring the experimental data.

To identify carbon flow toward lycopene, both the WT and triple KO flux distributions were run through NetFlow, which isolated the upstream metabolites from lycopene in the KO or its precursor farnesyl diphosphate (FRDP) in the WT. NetFlow simulations took 9.2s and 9.8s, respectively, for the WT and KO flux distributions. In the KO, 120 metabolites contribute carbon to lycopene, which represents ~44% of active, carbon-containing metabolites (Table S4). The extensive connectivity to lycopene was due to the complexity and high carbon mixing in the GSM. To isolate the primary production pathway, metabolites with a carbon yield on lycopene lower than 1% were removed, which reduced the set to 27 metabolites, an approximately five-fold reduction (Fig. 6, top). The net lycopene production reaction was then determined by manually lumping the reactions that convert glyceraldehyde 3-phosphate (G3P) and pyruvate to lycopene (Fig. 6, bot).

To identify a potential metabolic mechanism driving additional lycopene production, we used NetFlow to extract the subnetworks containing intermediates between the reactants of the KO reactions and lycopene in the single, double, and triple KO strains (Fig. 7). By following the iterative knockout path from the single KO through the triple KO, we were able to elucidate the impact of each subsequent gene KO on lycopene production. To capture the contribution of AKG, the yield threshold for important metabolites was lowered to 0.1% across all conditions (Table 2).

When *gdhA* is knocked out, the glutamate dehydrogenase reaction (GLUDy) is eliminated, disrupting the conversion of AKG to glutamate (Fig. 7a). To compensate for this disruption, fluxes through other conversion pathways such as glutamate synthase (GLUSy) are increased, drawing additional AKG from the TCA cycle. The altered flow through AKG corresponds with the activation of the glyoxylate shunt, which converts isocitrate to succinate and malate, and oxaloacetate decarboxylase, which recycles additional oxaloacetate into pyruvate and CO_2. Ultimately, this recovered pyruvate is shuttled toward lycopene production. Indirectly, the knockout of GLUDy results in the activation of the Entner-Doudoroff (ED) pathway and the conversion of dihydroxyacetone phosphate (DHAP) to pyruvate. Additionally, the knockout increased the generation of G3P from upper EMP and lower PPP and pyruvate from lower EMP as well as the decreased consumption of pyruvate outside the TCA cycle or lycopene production. Taken together, these metabolic alterations resulting from the knockout of GLUDy correspond with an increased pool of pyruvate and G3P, which is consumed to generate lycopene.

When *gdhA* and *aceE* are knocked out simultaneously, GLUDy and pyruvate dehydrogenase (PDH) are eliminated from the network, respectively (Fig. 7b). The removal of PDH disrupts the conversion of pyruvate to ACOA, which directly increased the availability of pyruvate for lycopene production. In combination with the GLUDy KO, the PDH KO corresponds with increased flow through the TCA cycle and malic enzyme (ME1), which regenerates pyruvate from malate. Interestingly, the decreased production of ACOA due to the KO of PDH prevents the activation of the glyoxylate shunt, which consumes an additional ACOA over the TCA cycle. Similar to the single KO, the double KO results in increased production of G3P and pyruvate from EMP and PPP, as well as the conversion of DHAP to pyruvate. However, in the double KO, some of the additional pyruvate is rerouted away from lycopene, as evidenced by the increased flow through the pyruvate sink reaction (Fig. 7b). This
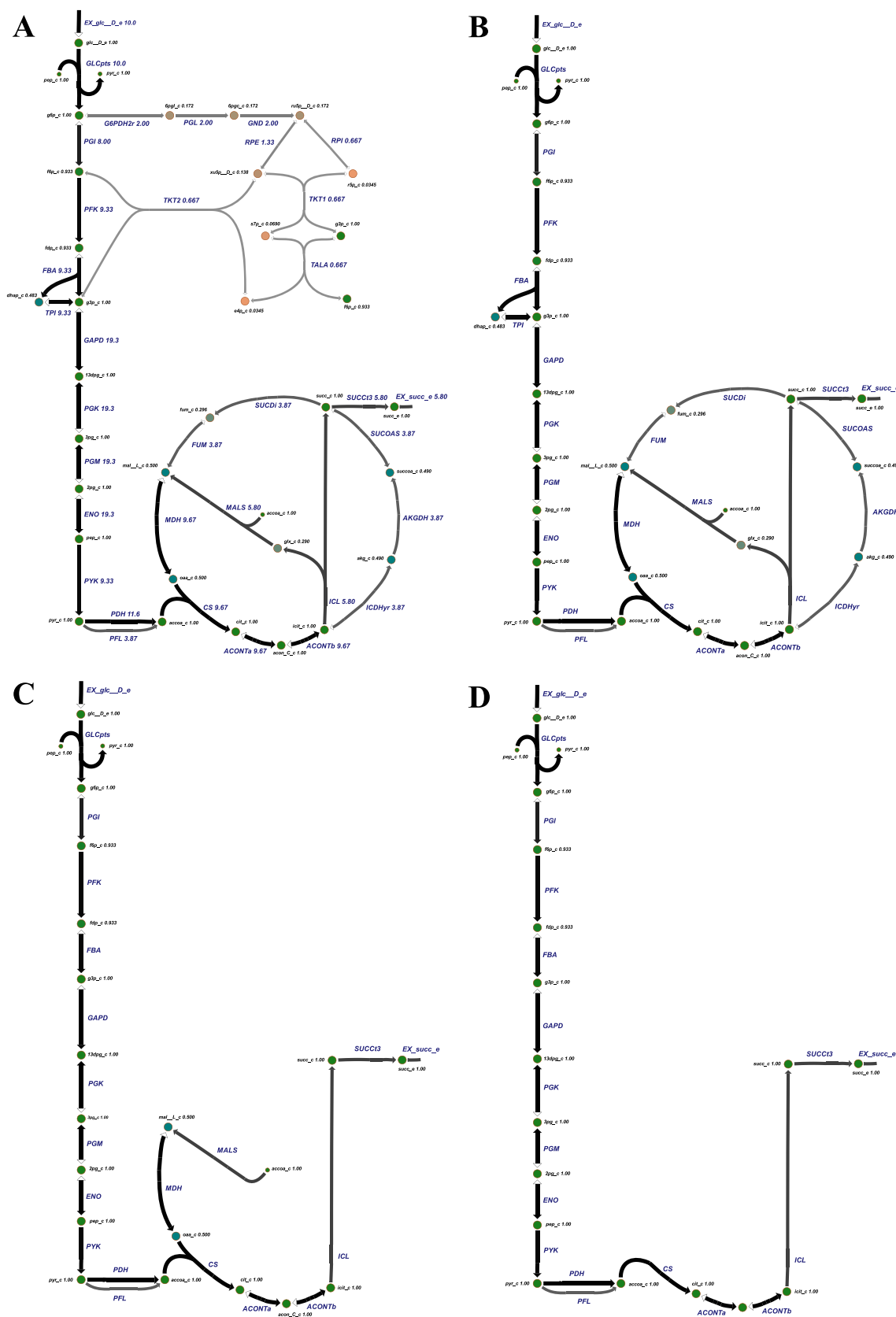
Fig. 4. Isolation of Primary Carbon Flow using Metabolite Yield Thresholds. (**A**) When each metabolite's yield is overlain onto the network, the relative contribution of each metabolite becomes apparent. (**B**) At a threshold of 0.25, the PPP is removed as it is a low-yield contributor to succinate (succ_c). (**C**) At 0.5, DHAP (dhap_c) and portions of the TCA cycle are removed due to the loss of some carbon as CO_2. (**D**) At 0.75, malate (mal_L_c) and oxaloacetate (oaa_c) are removed, leaving the most carbon efficient active pathway for producing succinate: EMP to TCA to ICL. The full descriptions for all metabolites and reactions contained in these subnetworks are given in Table S5 and Table S6, respectively.
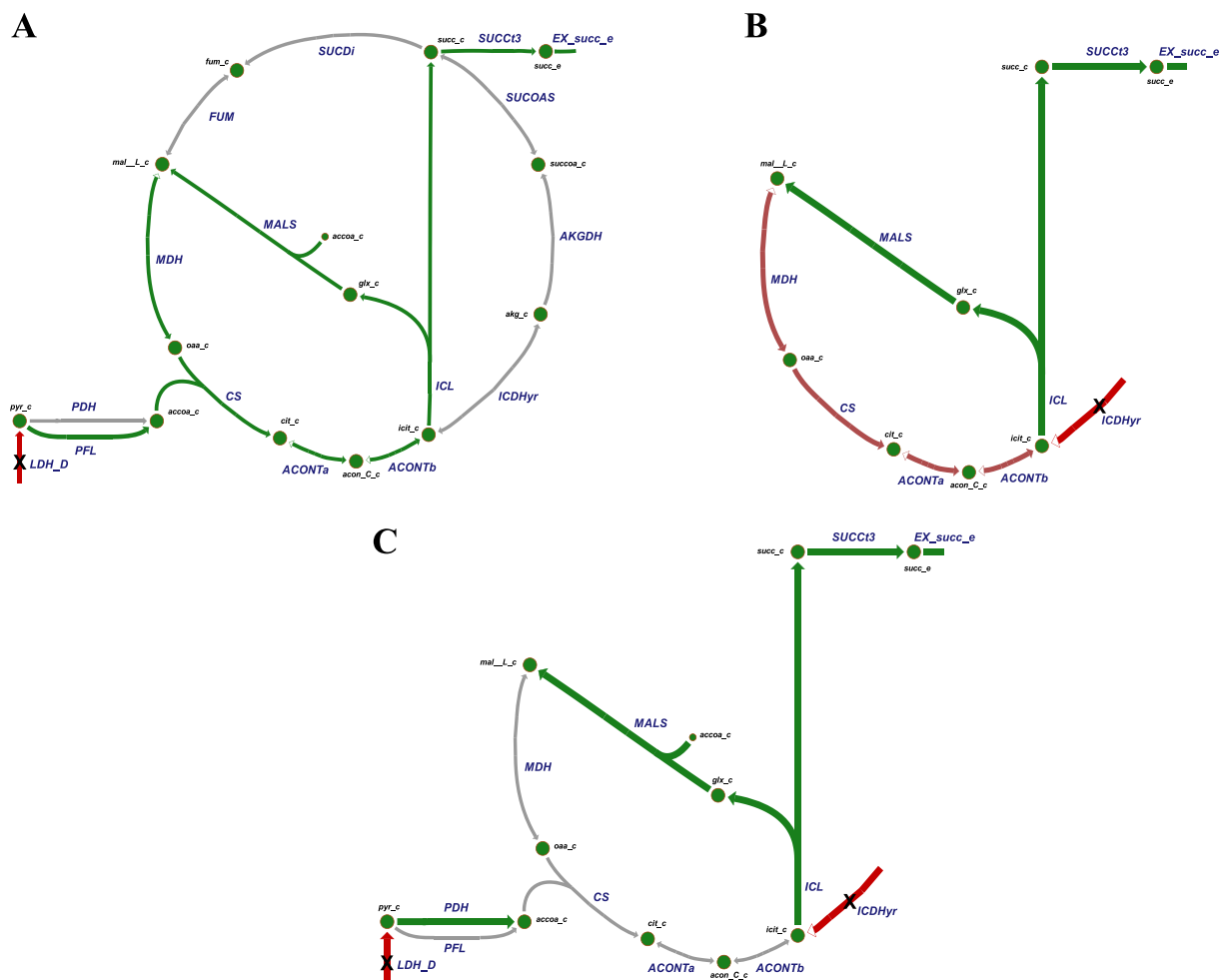
Fig. 5. Isolation of KO Mechanisms that Increase Succinate Production. (**A**) To ascertain the mechanism behind the LDH KO, NetFlow isolated the connected metabolites between pyruvate (the reactant in LDH) and succinate. The subnetwork contained only the TCA cycle and glyoxylate shunt. From this subnetwork, we can readily understand the mechanism: the KO of LDH shunts pyruvate toward TCA cycle which generates additional succinate via increased ICL activity. (**B**) The process was repeated for ICDH KO. The mechanism was simply that the ICDH KO forces all TCA flux through ICL, which more efficiently generates succinate (no loss of CO_2). (**C**) In the double KO, the LDH KO provides additional ACoA for TCA and MALS, while the ICDH KO forces carbon efficiency via ICL. All fluxes are shown relative to the WT. The magnitudes of the flux shifts are captured in the thickness and color intensity of the flux arrows. Green: increased flux; Red: decreased flux. The full descriptions for all metabolites and reactions contained in these subnetworks are given in Table S5 and Table S6, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

phenotype suggests that G3P availability and/or energetic demands are preventing additional lycopene production in the double KO strain.

In the triple KO strain, formate hydrogen lyase (FHL) is eliminated in addition to GLUDy and PDH. The removal of FHL prevents the conversion of formate to hydrogen and CO_2. The FHL KO caused a disconnect between the KO reactants and the intermediates from EMP and PPP, which precluded those intermediates from the extracted subnetwork. Interestingly, formate does not contribute carbon to lycopene in the triple KO (Table 2), suggesting a purely indirect impact of the FHL KO on lycopene synthesis. As evidenced by the corresponding subnetwork (Fig. 7c), the

observed mechanism for increased lycopene production is largely identical to that of the double KO. Specifically, additional AKG is rerouted through the TCA cycle and ME1, which regenerates pyruvate, and excess pyruvate is shuttled toward lycopene. While not captured in the subnetwork, the triple KO strain retains the increased production of G3P and pyruvate from EMP and PPP that was observed in the single and double KO strains. This is evidenced by the increased yield of those intermediates on lycopene (Table S4). Furthermore, the rerouting of pyruvate away from lycopene or the TCA cycle is higher than the WT but lower than the double KO. Thus, our analysis helps unravel how the FHL

Table 1
Growth Rate and Lycopene Production in Single-, Double- and Triple-KO Strains of *E. coli*. Our simulations modeled the genetic variants reported in Alper et al. (2005). In our simulations, growth rate (biomass synthesis) and lycopene synthesis were in units of mmol/gDWhr. In Alper et al. (2005), growth rates were reported in hr^{-1} and lycopene synthesis was reported as percent increase over WT.

	Our Simulations				Alper et al. (Alper et al., 2005)			
Strain	Growth	Lycopene	Growth Ratio	Lycopene Ratio	Growth	Lycopene	Growth Ratio	Lycopene Ratio
WT	0.495	0.000	1.000	0.000	0.670	0.000	1.000	0.000
gdhA	0.423	0.032	0.855	0.547	0.550	13.000	0.821	0.351
gdhA + aceE	0.394	0.045	0.797	0.766	0.520	13.000	0.776	0.351
gdhA, aceE, fdhF	0.379	0.059	0.766	1.000	0.380	37.000	0.567	1.000

Net Reaction for Lycopene Production
8 pyr + 8 g3p + 8 nadh + 8 atp + 8 ctp -> lycop + 8 co2 + 8 nad + 8 adp + 8 cmp
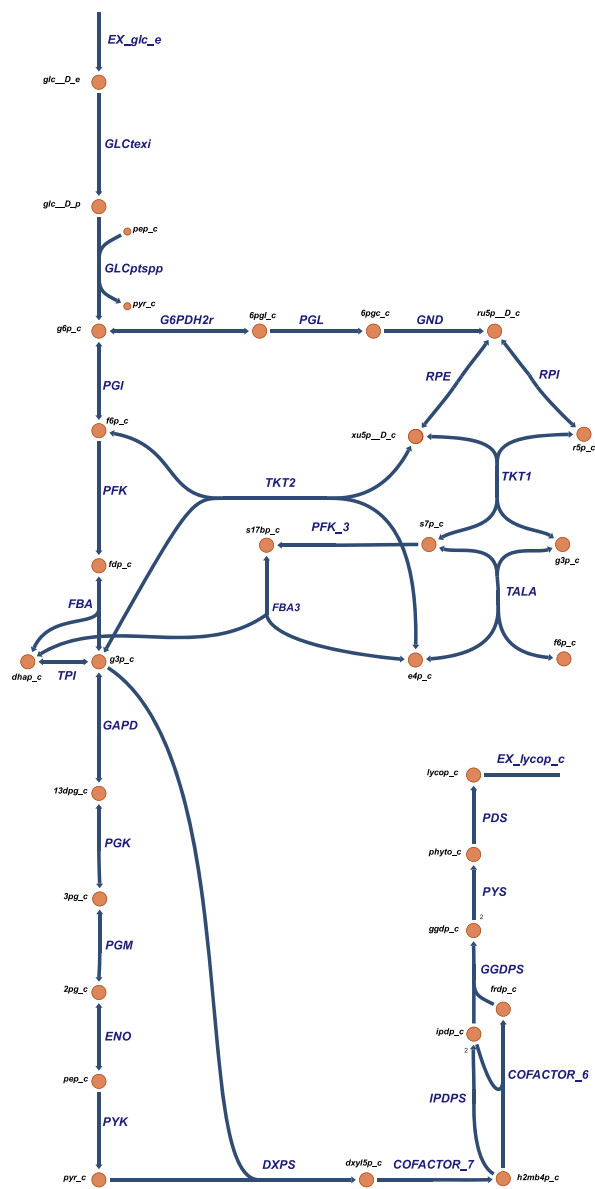
Fig. 6. Lycopene Production Pathway in Triple KO *E. coli*. By isolating the upstream metabolites from lycopene with a yield >1%, NetFlow isolated the main carbon flow between glucose and lycopene (top). The net reaction for the production of lycopene from pyruvate and G3P was then generated from this subnetwork (bot). The full descriptions for all metabolites and reactions contained in this subnetwork are given in Table S5 and Table S6, respectively.

KO alleviated a possible bottleneck in the double KO, allowing for a greater increase in lycopene synthesis.

4. Discussion

In this work, we successfully developed an algorithm, NetFlow, that leverages genome-scale carbon mapping to isolate carbon flows from a given genome-scale flux prediction. NetFlow expands upon prior path tracing approaches by integrating full carbon mapping and flux constraints for pathway identification. By calculating metabolite yields, NetFlow enables rapid differentiation between active pathways of carbon flow. To illustrate the utility of NetFlow, we applied it to two case studies of metabolic engineering in *E. coli*, one theoretical and one practical.

4.1. Carbon flow through CCM highlights importance of carbon yield

In our first case study, we applied the NetFlow algorithm to visualize and compare the carbon flow towards succinate across several conditions. The base flux reduction demonstrated the primary practical application of NetFlow—to identify and isolate the production of a given metabolite. Unsurprisingly, as succinate was one of only four efflux metabolites in the original network, the initial succinate subnetwork remained somewhat complicated (Fig. 3). By only using binary connectivity between metabolites contained in **L**, the metabolites upstream of succinate were indistinguishable. The subsequent incorporation of yield contained in **EOL** identified the major carbon flows and isolated the most efficient succinate production pathway (Fig. 4).

Ultimately, the carbon yield of a metabolite represents a relative weighting on the importance of that metabolite for the generation of the target metabolite. As a metric, the yield is dependent on both relative fluxes and the specific carbon exchange between metabolites and thus provides more information than individual fluxes (Tervo and Reed, 2016) or full carbon mapping individually (Pey et al., 2014). By incorporating full carbon mapping and constraint-based modeling via GSMs, NetFlow provides the unique capability to calculate and leverage the nuanced relationships between metabolites to identify biologically important pathways.

4.2. Elucidation of mechanism driving lycopene production in E. coli

In our second case study, NetFlow was used to identify the mechanisms behind the increased production of lycopene in single ($\Delta gdhA$), double ($\Delta gdhAaceE$), and triple ($\Delta gdhAaceEfdhF$) KOs of *E. coli* (Alper et al., 2005). While the original publication experimentally validation the increased lycopene yields in these KOs, it did not elucidate the metabolic mechanisms underlying the increases in yield. We filled this gap by using NetFlow. Specifically, we isolated subnetworks that connected the knocked out genes to the product lycopene. Arising from these subnetwork was a general mechanism of altered TCA cycle activity and increased lower EMP and lower PPP activity, which increased the availability of G3P and pyruvate for lycopene production; each additional gene KO corresponded to a furthering of this mechanism.

While these subnetworks fully account for the rerouting of carbon flow, the complete mechanism likely involves more than the direct rerouting of carbon (e.g. rerouting of redox equivalents or energy). In addition to the regeneration of pyruvate, the increased flux through the TCA cycle generated additional NADH and ATP, which are required to produce lycopene and may be limiting reactants. Furthermore, since formate did not contribute carbon to lycopene, the FHL KO could have only indirectly impacted lycopene production. In the WT, FHL generated hydrogen that was consumed by hydrogenase, a respiratory enzyme that reduces ubiquinol-8 and transports two protons to the periplasm. As the sole hydrogen-generating reaction, the KO of FHL also arrested hydrogenase activity, thereby altering respiratory activity. Presumably, this shift in respiratory activity indirectly contributed to improved utilization of pyruvate in the triple KO as compared to the double KO. However, the specific mechanism cannot be elucidated from our analysis of the carbon flows.

4.3. Effective implementation of NetFlow

The ability of NetFlow to isolate carbon flow through specific pathways has the potential to be an indispensable aid in the analysis of large-scale flux predictions and the mechanistic understanding of metabolic perturbations. Our two case studies led to a few generalizable approaches for utilizing NetFlow. First, if only a single flux vector is under consideration, the complete set of production or consumption pathways for a given metabolite can be identified by using the binary mapping matrix. However, due to the high connectivity of metabolic networks, especially at the genome-scale, a yield threshold should be applied to isolate the
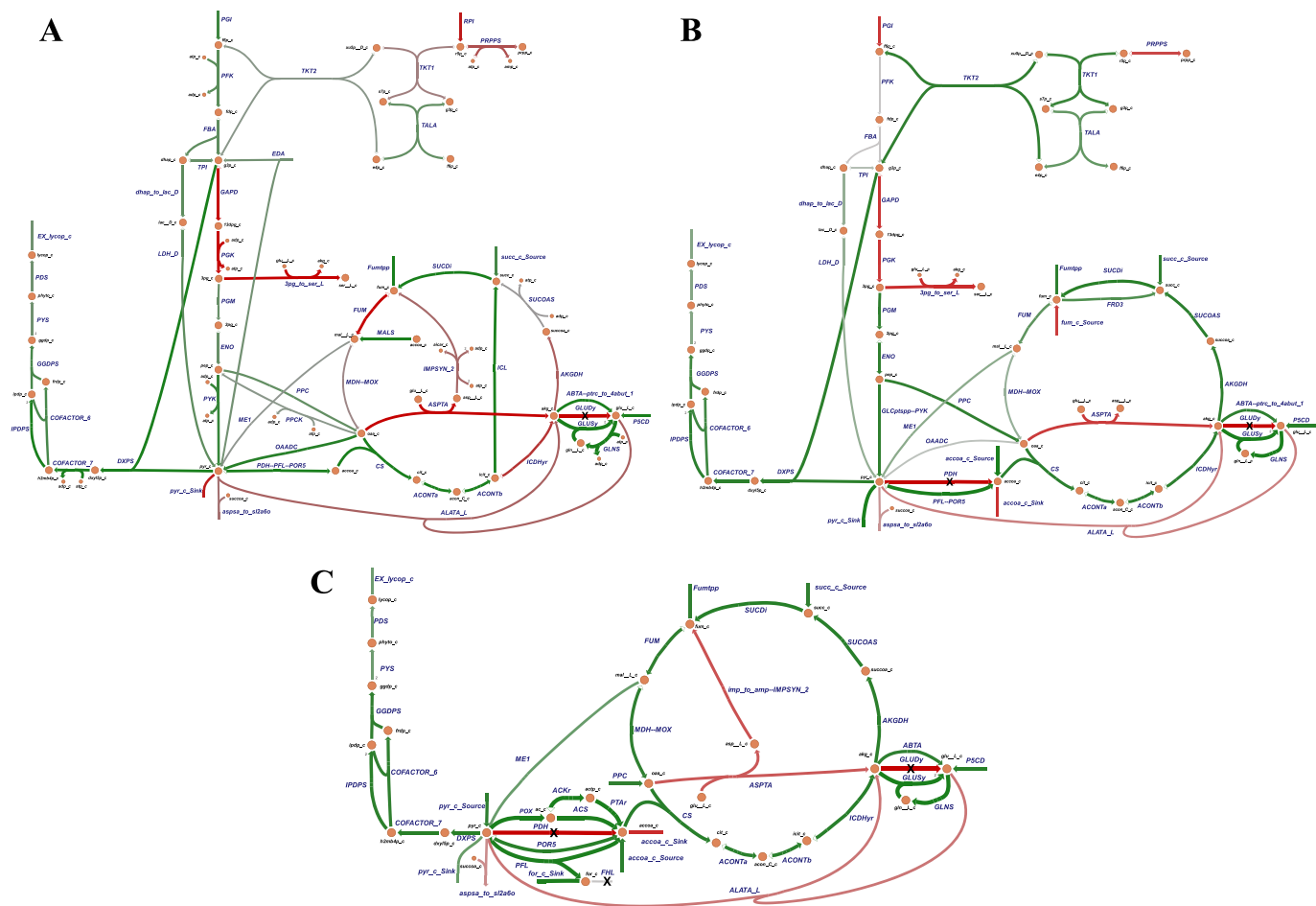
Fig. 7. Analysis of Lycopene Production in Single, Double, and Triple KO strains of *E. coli*. To elucidate a mechanism driving the increase in lycopene production, NetFlow identified and isolated the metabolites between the KO reactants (PDH: pyruvate; GLUDy: AKG; FHL: formate) and lycopene in the single, double, and triple KO strains of *E. coli*. The resulting subnetworks illustrate that the mechanisms for increased lycopene production revolve around pyruvate availability. (**A**) The single KO of *gdhA* eliminates GLUDy activity, which corresponds with the activation of the glyoxylate shunt and OAADC. Along with increased flow through EMP and PPP, these changes increase the pool of pyruvate, which allows for lycopene production. (**B**) The double KO of *gdhA* and *aceE* eliminates both GLUDy and PDH activity, which corresponds to increased TCA cycle activity, decreased flow of pyruvate toward the TCA cycle, and activation of ME1. These changes, combined with an increase in EMP and PPP activity, generate a surplus of pyruvate for lycopene production; the excess pyruvate corresponds with an increased flux away from lycopene or the TCA cycle ("pyr_c_Sink"). (**C**) The triple KO of *gdhA*, *aceE*, and *fdhF* eliminates activity through GLUDy, PDH, and FHL, which corresponds to a similar phenotype observed in the double KO. Critically, a higher proportion of the excess pyruvate is utilized for lycopene production than in the double KO, likely due to alterations in redox balancing caused by the KO of FHL. Despite formate not directly contributing carbon to lycopene, it is included in this subnetwork for visualization of the FHL KO. All fluxes are shown relative to the WT. The magnitudes of the flux shifts are captured in the thickness and color intensity of the flux arrows. Green: increased flux; Red: decreased flux. The full descriptions for all metabolites and reactions contained in these subnetworks are given in Table S5 and Table S6, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

major carbon flows and thus reduce the scale of the subnetwork. Next, to elucidate the impact of environmental or genetic perturbations on a metabolic network, NetFlow can isolate the conversion pathways

Table 2

Yield of KO Metabolites on Lycopene. The yield of the core metabolites in each KO reaction on lycopene was taken from the corresponding entries in **EOL**. Single corresponds to the simulated Δ*gdhA* strain. Doubles corresponds to the simulated Δ*gdhAaceE* strain. Triple corresponds to the simulated Δ*gdhAaceEfdhF* strain.

Gene	Reaction	Metabolite	Single	Double	Triple
gdhA	GLUDy	AKG	0.0039	0.0030	0.0043
gdhA	GLUDy	Glu	0.0026	0.0020	0.0027
aceE	PDH	Pyr	0.4001	0.4000	0.4000
aceE	PDH	ACOA	0.0031	0.0021	0.0030
fdhF	FHL	For	<0.0001	<0.0001	0.0000
fdhF	FHL	CO2	<0.0001	<0.0001	<0.0001

between metabolites connected to particular metabolic nodes. Thereby, NetFlow can identify biologically meaningful shifts in carbon flow directly caused by the perturbation. For metabolic engineering applications, the metabolites of interest would include the target metabolite and the metabolites involved in the proposed KO(s).

While the described case studies only utilized connectivity and metabolite yield, NetFlow can be readily modified to generate matrices based on other potentially useful metrics. For example, the specific carbon yield:

$$\frac{\sum_{k} X_{ki}}{N_i^C}, \ k \in C_j \qquad [7]$$

could be used as an alternative to **EOL** [Eq. 2]. In contrast to the definition in [Eq. 2], N_i^C is the number of carbon atoms in metabolite *i*, not metabolite *j*. NetFlow can also simulate individual carbon atom connectivity and yields throughout the network by iteratively labeling each carbon for each metabolite. Currently, NetFlow enumerates all possible

pathways to generate a given metabolite and can also identify the most carbon efficient pathways based on yield. If individual pathways are of interest, an optimization technique such as PathTracer (Tervo and Reed, 2016) could be implemented after NetFlow using metabolite yields to weight the optimization function.

Additionally, NetFlow could be applied to examine flow of atoms other than carbon (e.g. nitrogen or hydrogen) through metabolic networks by incorporating non-carbon AMMs generated from atom mapping prediction algorithms (Kumar et al., 2014; Hadadi et al., 2017; Litsa et al., 2019; Latendresse et al., 2012). By using non-carbon connectivity and yield, supplementary production pathways between metabolites of interest could be elucidated. For example, recent work by Liu et al. (2020) identified proton flow from glucose-6-phosphate to NADPH generated in the PPP to palmitic acid using deuterium labeling at atom C-3 of glucose. If such proton tracing could be extended genome wide, NetFlow could readily isolate a similar pathway connecting the NADH generated in the TCA cycle to lycopene production in the above case study. However, this would require the development of a database for mapping of atoms other than carbon.

While it is outside the current scope of the current algorithm, NetFlow could be extended to identify and/or validate potential gap-filling pathways based on carbon connectivity and yield. If given an incomplete carbon-mapped network and a set of carbon-mapped candidate reactions, the extension of NetFlow would simulate the carbon flow across the gaps in the network using each of the possible reactions. From these simulations, NetFlow could differentiate the candidate reactions by a user-defined yield threshold, as done in this manuscript for KO mechanisms. This approach could aid the user in the filling of gaps in a network.

Ultimately, NetFlow is an algorithm that greatly simplifies the interpretation and analysis of flux predictions. By focusing on carbon flow, NetFlow utilizes an orthogonal approach to network reduction algorithms, such as the NetRed algorithm from our research group (Lugar et al., 2020), which compresses a network based on a given flux prediction and set of desirable metabolites. Moving forward, such approaches could be combined with NetFlow to achieve even greater reduction in complexity. NetRed could be readily applied to further reduce the subnetworks generated by NetFlow and improve the interpretability of more complex mechanisms. To enable the use of NetFlow on networks reduced by NetRed, NetRed would need to be expanded to account for carbon mapping in the reduction process. Additionally, while NetFlow does not identify gene targets, it nicely complements software and algorithms that have been developed for this purpose, such as MOMA (Segrè et al., 2002) or OptKnock (Burgard et al., 2003), by readily elucidating the metabolic mechanism through which the suggested genetic interventions produce the desired metabolic effects. NetFlow allows the user to more quickly determine the mechanism of a given gene KO in producing a specific metabolic phenotype.

5. Conclusion

In this work, we described NetFlow, an algorithm that isolates pathways of interest based on metabolite connectivity and carbon yield. The resulting subnetwork dramatically increases the interpretability of genome-scale metabolic fluxes by focusing the scope of analysis. As demonstrated through our *E. coli* case studies, NetFlow is highly versatile and would be valuable in any metabolic flux analysis. By incorporating genome-wide atom mapping and GSM constraints, NetFlow can calculate carbon connectivity and yield for every metabolite in the network simultaneously in a single run. The recent development of highly-accurate atom mapping prediction algorithms (Kumar et al., 2014; Hadadi et al., 2017; Litsa et al., 2019; Latendresse et al., 2012) has significantly increased the availability of genome-scale AMMs, which enables the broad applicability of NetFlow. Thus, NetFlow represents a key advancement in the pathway elucidation and can potentially improve our mechanistic understanding of metabolism.

6. Author contributions

G.S. conceived this work and guided the development of the network algorithm. S.G.M developed the mathematical formulation, detailed the carbon flow algorithm, and wrote the MATLAB script for NetFlow. S.G.M. wrote the manuscript; S.G.M., and G.S. critically edited it. Both authors read the manuscript and approved the final version before submission.

Declaration of competing interest

The authors declare no financial or other conflicts of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.mec.2020.e00154.

References

Alper, H., Jin, Y.-S., Moxley, J.F., Stephanopoulos, G., 2005. Identifying gene targets for the metabolic engineering of lycopene biosynthesis in Escherichia coli. Metab. Eng. 7, 155–164.

Brynildsen, M.P., Winkler, J.A., Spina, C.S., MacDonald, I.C., Collins, J.J., 2013. Potentiating antibacterial activity by predictably enhancing endogenous microbial ROS production. Nat. Biotechnol. 31, 160–165.

Burgard, A.P., Pharkya, P., Maranas, C. D. Optknock, 2003. A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol. Bioeng. 84, 647–657.

Chowdhury, A., Maranas, C.D., 2015. Designing overall stoichiometric conversions and intervening metabolic reactions. Sci. Rep. 5, 1–20.

Croes, D., Couche, F., Wodak, S.J., van Helden, J., 2005. Metabolic PathFinding: inferring relevant pathways in biochemical networks. Nucleic Acids Res. 33, W326–W330.

Croes, D., Couche, F., Wodak, S.J., van Helden, J., 2006. Inferring meaningful pathways in weighted metabolic networks. J. Mol. Biol. 356, 222–236.

Curran, K.A., Crook, N.C., Alper, H.S., 2012. Using flux balance analysis to guide microbial metabolic engineering. In: Cheng, Q. (Ed.), Microbial Metabolic Engineering, vol. 834 vols. 197–216. Springer, New York.

Folger, O., et al., 2011. Predicting selective drug targets in cancer through metabolic networks. Mol. Syst. Biol. 7, 501.

Gopalakrishnan, S., Maranas, C.D., 2015. 13C metabolic flux analysis at a genome-scale. Metab. Eng. 32, 12–22.

Hadadi, N., Hafner, J., Soh, K.C., Hatzimanikatis, V., 2017. Reconstruction of biological pathways and metabolic networks from in silico labeled metabolites. Biotechnol. J. 12.

Heath, A.P., Bennett, G.N., Kavraki, L.E., 2010. Finding metabolic pathways using atom tracking. Bioinformatics 26, 1548–1555.

Heirendt, L., et al., 2019. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nat. Protoc. 14, 639–702.

Holzhütter, H.-G., 2004. The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. Eur. J. Biochem. FEBS 271, 2905–2922.

Huang, Z., Lee, D.-Y., Yoon, S., 2017a. Quantitative intracellular flux modeling and applications in biotherapeutic development and production using CHO cell cultures. Biotechnol. Bioeng. 114, 2717–2728.

Huang, Y., Zhong, C., Lin, H.X., Wang, J., 2017b. A method for finding metabolic pathways using atomic group tracking. PloS One 12.

Keseler, I.M., et al., 2017. The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. Nucleic Acids Res. 45, D543–D550.

Kim, S.M., Peña, M.I., Moll, M., Bennett, G.N., Kavraki, L.E., 2017. A review of parameters and heuristics for guiding metabolic pathfinding. J. Cheminf. 9, 1–13.

Kim, S.M., Peña, M.I., Moll, M., Bennett, G.N., Kavraki, L.E., 2020. Improving the organization and interactivity of metabolic pathfinding with precomputed pathways. BMC Bioinf. 21, 13.

King, Z.A., et al., 2015. A web application for building, sharing, and embedding data-rich visualizations of biological pathways. PLoS Comput. Biol. 11, e1004321.

Klamt, S., et al., 2017. From elementary flux modes to elementary flux vectors: metabolic pathway analysis with arbitrary linear flux constraints. PLoS Comput. Biol. 13, e1005409.

Kumar, A., Maranas, C.D., Clca, 2014. Maximum common molecular substructure queries within the MetRxn database. J. Chem. Inf. Model. 54, 3417–3438.

Latendresse, M., Malerich, J.P., Travers, M., Karp, P.D., 2012. Accurate atom-mapping computation for biochemical reactions. J. Chem. Inf. Model. 52, 2970–2982.

Litsa, E.E., et al., 2019. Machine learning guided atom mapping of metabolic reactions. J. Chem. Inf. Model. 59, 1121–1135.

Liu, X., et al., 2020. Cystine transporter regulation of pentose phosphate pathway dependency and disulfide stress exposes a targetable metabolic vulnerability in cancer. Nat. Cell Biol. 22, 476–486.

Lugar, D.J., Mack, S.G., Sriram, G., 2020. NetRed, an algorithm to reduce genome-scale metabolic networks and facilitate the analysis of flux predictions. Metab. Eng. https://doi.org/10.1016/j.ymben.2020.11.003.

McCloskey, D., Young, J.D., Xu, S., Palsson, B.O., Feist, A.M., 2016. Modeling method for increased precision and scope of directly measurable fluxes at a genome-scale. Anal. Chem. 88, 3844–3852.

McCloskey, D., et al., 2018a. Multiple optimal phenotypes overcome redox and glycolytic intermediate metabolite imbalances in Escherichia coli pgi knockout evolutions. Appl Env. Microbiol AEM 823–918. https://doi.org/10.1128/AEM.00823-18.

McCloskey, D., et al., 2018b. Evolution of gene knockout strains of E. coli reveal regulatory architectures governed by metabolism. Nat. Commun. 9, 3796.

McCloskey, D., et al., 2018c. Growth adaptation of gnd and sdhCB Escherichia coli deletion strains diverges from a similar initial perturbation of the transcriptome. Front. Microbiol. 9.

Orth, J.D., Thiele, I., Palsson, B.Ø., 2010. What is flux balance analysis? Nat. Biotechnol. 28, 245–248.

O'Brien, E.J., Lerman, J.A., Chang, R.L., Hyduke, D.R., Palsson, B.Ø., 2013. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol. Syst. Biol. 9, 693.

Pey, J., Prada, J., Beasley, J.E., Planes, F.J., 2011. Path finding methods accounting for stoichiometry in metabolic networks. Genome Biol. 12, R49.

Pey, J., Planes, F.J., Beasley, J.E., 2014. Refining carbon flux paths using atomic trace data. Bioinformatics 30, 975–980.

Pharkya, P., Burgard, A.P., Maranas, C. D. OptStrain, 2004. A computational framework for redesign of microbial production systems. Genome Res. 14, 2367–2376.

Preciat Gonzalez, G.A., et al., 2017. Comparative evaluation of atom mapping algorithms for balanced metabolic reactions: application to Recon 3D. J. Cheminf. 9, 39.

Ramirez, A.K., et al., 2017. Integrating extracellular flux measurements and genome-scale modeling reveals differences between Brown and white adipocytes. Cell Rep. 21, 3040–3048.

Segrè, D., Vitkup, D., Church, G.M., 2002. Analysis of optimality in natural and perturbed metabolic networks. Proc. Natl. Acad. Sci. U.S.A. 99, 15112–15117.

Simeonidis, E., Rison, S.C.G., Thornton, J.M., Bogle, I.D.L., Papageorgiou, L.G., 2003. Analysis of metabolic networks using a pathway distance metric through linear programming. Metab. Eng. 5, 211–219.

Tervo, C.J., Reed, J.L., 2016. MapMaker and PathTracer for tracking carbon in genome-scale metabolic models. Biotechnol. J. 11, 648–661.

Wiechert, W., Möllney, M., Isermann, N., Wurzel, M., de Graaf, A.A., 1999. Bidirectional reaction steps in metabolic networks: III. Explicit solution and analysis of isotopomer labeling systems. Biotechnol. Bioeng. 66, 69–85.