# Arithmetic simplicity beneath metabolic network architecture

William J. Riehl[1], Paul L. Krapivsky[2], Sidney Redner[2] and Daniel Segrè[1,3,*]

[1]Program in Bioinformatics and Systems Biology, [2]Department of Physics, [3]Department of Biology, Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215.

* E-mail: dsegre@bu.edu

**Abstract**

Metabolic networks perform some of the most fundamental functions in living cells, including energy transduction and building block biosynthesis. While these are the best characterized networks in living systems, understanding their evolutionary history and complex wiring constitutes one of the most fascinating open questions in biology, intimately related to the enigma of life's origin itself. Is the evolution of metabolism subject to general principles, beyond the unpredictable accumulation of multiple historical accidents? Here we search for such principles by applying to an artificial chemical universe some of the methodologies developed for the study of genome scale models of cellular metabolism. In particular, we use metabolic flux constraint-based models to exhaustively search for artificial chemistry pathways that can optimally perform an array of elementary metabolic functions. Despite the simplicity of the model employed, we find that the ensuing pathways display a surprisingly rich set of properties, including the existence of autocatalytic cycles and hierarchical modules, the appearance of universally preferable metabolites and reactions, and a logarithmic trend of pathway length as a function of input/output molecule size. Some of these properties can be derived analytically, borrowing methods previously used in cryptography. In addition, by mapping biochemical networks onto a simplified carbon atom reaction backbone, we find that several of the properties predicted by the artificial chemistry model hold for real metabolic networks. These findings suggest that optimality principles and arithmetic simplicity might lie beneath some aspects of biochemical complexity.

**Author Summary**

An open question in biology is whether the evolution of metabolic networks has been guided by general optimality principles beyond the unpredictable accumulation of historical accidents. Here we search for signatures of such optimality in an idealized artificial chemistry model, where it is feasible to systematically explore a complete set of efficient metabolic pathways of minimal length between any two compounds. These pathways display a modular organization of recurring topologies, including autocatalytic cycles, and a logarithmic dependence of pathway length on input/output molecule size. Across all pathways, we predict the emergence of ubiquitous metabolites, and a broad spectrum of reaction utilization, with certain reactions serving as universal steps. Similar properties hold for real metabolic networks, suggesting that optimality principles and arithmetic simplicity underlie biochemical complexity.

## Introduction

It is still a mystery how abiotically-formed molecules on primordial Earth were gradually converted into the building blocks of life [1-3]. Even more enigmatic is the emergence of the self-sustaining network of chemical reactions called cellular metabolism [4-9]. Was the evolution of metabolism subject to inevitable network optimization steps [10] or dominated by unpredictable frozen accidents [6]? In comparative genomics studies [11] and laboratory evolution experiments [12-14], one can witness mostly short-term metabolic adaptations that affect metabolic enzyme regulation and fine tuning of kinetic parameters. Yet, no experimental methods exist to probe the long-term evolution of the metabolic wiring itself.

It has been suggested that biochemical network topologies may have evolved towards optimal efficiency [15, 10, 16]. Here, we seek to identify potential distinctive features of optimally efficient metabolic networks, as well as the ecosystem-level union of all metabolic pathways [17] in which multiple metabolic tasks must be concurrently performed. We focus on elementary metabolic tasks, in which a given input molecule is converted into a specified output. Since a systematic search for optimal pathways between any two molecules in a real biochemical network is combinatorially inapproachable, we apply recent systems biology approaches [18, 19] to an artificial chemistry universe [5, 6, 20, 21]. Behind the apparent complexity of the ensuing optimal pathways, we identify recurring, modularly organized categories of network topologies, and analytically predictable trends in pathway length. In addition, we observe the emergence of "universal metabolic tools" across all optimal pathways. Finally, despite the huge gap in the underlying chemical rules, we find that some properties of real metabolic pathways are consistent with the patterns detected in the model, suggesting that fundamental optimality principles may have played a role in shaping biochemical networks.

## Results

Our artificial chemistry consists of a set of $N$ possible molecules $\{a_1, a_2, a_3, \ldots, a_N\}$, that can participate in reversible ligation/cleavage reactions of the form $a_i + a_j \leftrightarrow a_k$ , with $i$ +

$j = k$. This model could be viewed as the simplest possible string-based artificial chemistry [5]. The reaction network $R_N$ that includes all metabolites up to length $N$ and all possible reactions between them (Fig. 1) can be thought of as an evolving biosphere-level metabolism. Here we were concerned with pathways, within the $R_N$ network, that can optimally perform a given metabolic task. In particular, we searched for optimal solutions to the problem of producing a specific end-product (e.g., $a_j$, with output flux $v_{out}$) from a single available nutrient (e.g., $a_i$, with input flux $v_{in}$). We require an optimal pathway to (i) allow a steady state solution, i.e., a mass-conserving flow from input to output; (ii) have maximal yield, and no waste [22], such that $v_{out}=v_{in}\cdot j/i$; and (iii) have the fewest reaction steps possible. A pathway satisfying these conditions is termed a *minimal balanced pathway* (MBP) between $a_i$ and $a_j$, and will be denoted $a_i \Rightarrow a_j$. MBPs can be thought of as the pathways that are most efficient for a specific metabolic task, in the sense that they require the smallest possible number of different enzymes for producing the maximal possible yield [10, 23, 24].

Despite the simplicity of our artificial chemistry, identifying the MBPs between all possible input-output pairs in a given artificial chemistry $R_N$ is a challenge for large $N$. We implement three algorithms to approach this problem: a mixed integer linear programming (MILP) akin to flux balance analysis (FBA) [25]; an algorithm that uses enumeration of elementary flux modes [18]; and finally an iterative algorithm that gradually assembles new MBPs from already identified simple ones (see Methods). The three algorithms differ mainly in their scalability, and in their capacity to predict multiple degenerate solutions (see Table S1). A partial overview of the results of our calculations is shown in Fig. 2A and Fig. S1 (see Table S2 for a comprehensive list of MBPs).

Behind the apparent complexity of the topologies encountered in each of the different pathways, it is possible to observe the recurrence of three fundamental categories: each MBP functions either as a pure “addition chain” [26], where smaller metabolites are progressively added together to build the target molecule, or as an “addition-subtraction chain”, in which metabolites are both synthesized and degraded within the pathway. There is also a third, smaller category of cyclical pathways that cannot proceed unless a certain intermediate molecule is already present in the system. These pathways are autocatalytic cycles (Fig. 2B) that very much resemble autocatalytic cycles found in real biochemistry, such as the reverse TCA cycle [4], or the formose reaction [27]. Our results show that autocatalytic cycles can be simultaneously optimal for multiple tasks (Fig. 2B), suggesting that such types of structure may have a fundamental evolutionary advantage in a biological context. In addition to the recurrence of these topological categories among MBPs, we find that some specific structures are used repeatedly, often in a modular fashion (Fig. 2C). Specifically, many simple MBPs are used hierarchically as a toolkit for the construction of progressively more complex MBPs (data not shown), similar to what has been observed in real metabolic networks [28-30].

This modular architecture of recurring graph types provides a topological signature of optimally efficient pathways in our idealized chemistry. Since these pathways are chosen based on their minimal length, one may expect that a systematic analysis of all MBP lengths will display additional distinctive properties. Indeed, pathway lengths increase roughly logarithmically with the size of the input (or output) molecule (Fig. 3), with superimposed sharp jumps. For example, the task $a_9 \Rightarrow a_6$ can be performed in 2 steps, but the neighbor task $a_9 \Rightarrow a_7$ requires a minimum of 6 steps. Moreover, while most MBPs have only one or a few optimal realizations, selected instances display a peak in possible redundant solutions (Supplementary Fig. 2C), usually due to interconversions between molecules of similar size (e.g., $a_x \Rightarrow a_{x+1}$), or to the inherent complexity of a specific molecule (e.g., $a_7 \Rightarrow a_j$). These regular patterns suggest that it may be possible to reproduce the MBP length curves without having to actually compute the MBPs. Inspired by the analogy with addition-subtraction chains, that are used in cryptography to compute large integers in a minimal number of exponent addition steps, or addition-subtraction combinations [26, 31], we derived the following analytical estimate of the length of MBPs (see Methods):

$$L(i,j) \sim \log_2 \frac{i}{\gcd(i,j)} + \log_2 \frac{j}{\gcd(i,j)} \qquad (1)$$

where $L(i, j)$ is the number of reactions in the MBP with input $a_i$ and output $a_j$, and gcd($i$, $j$) is the greatest common divisor of $i$ and $j$. As seen in Fig. 3, Eq. 1 reproduces the corresponding pathway lengths obtained by computing individual MBPs. This agreement implies that the number of reaction steps needed to construct an efficient metabolic pathway between two metabolites in our artificial chemistry can be roughly estimated from Eq. 1. The only feature that determines the pathway lengths is the complexity of the input and output molecules.

The patterns identified so far relate to individual optimally efficient pathways. What patterns would we expect to see in an ecosystem-level metabolic network, in which multiple organisms under different environmental conditions collectively perform a large variety of optimal metabolic interconversions? Would all metabolites and reactions be used in roughly the same number of optimal pathways, or do we expect to observe the emergence of "universal tools" – specific metabolites or reactions essential for many optimal tasks? In our artificial chemistry model, we can address this question by examining the overall usage of reactions and metabolites across all possible MBPs (see Methods). This analysis shows that every metabolite of an even length is used in many more MBP reactions than their odd length neighbors, compared to the underlying chemistry (Fig. 4B). Thus even-length metabolites are more important in that they can be used for more tasks. A possible explanation for this enhanced importance comes from the logarithmic nature of the MBP path lengths. For example, producing $a_8$ from $a_1$ requires only three doubling reactions ($a_1 + a_1 \rightarrow a_2$, $a_2 + a_2 \rightarrow a_4$, and $a_4 + a_4 \rightarrow a_8$). In addition, this same pathway, with one additional reaction, can also be used to optimally produce $a_9$ and $a_{10}$ (see Table S2), overall increasing the number of pathways in which each of those even-length intermediates is used. Indeed, because similar logarithmic pathways can be used as a backbone connecting distant inputs and outputs, metabolites of an even length should appear more often. Similarly, one can address the relevance of each possible reaction across different MBPs. The existence of ubiquitous reactions is visible in Fig. 2A and Fig. S1, and can be more systematically assessed by plotting a usage distribution (Fig. S2). The most abundant reactions are the ones that ligate two identical molecules (e.g. $a_2 + a_2 \leftrightarrow a_4$, see Tables, 1, S2, and S3). Strikingly, the distribution of reaction utilization follows a long-tailed distribution (Fig. 4A), whose fit to a power law gives an exponent of approximately -1.1 ($R^2 = 0.99$). This value is close to our theoretically predicted value of -1 (See Methods).

After identifying signatures of optimal efficiency in our idealized chemistry, we can return to our original question, and ask whether similar optimality principles are discernible in real metabolic networks. To cope with the gap in complexity between our model and real chemistry, we mapped real metabolic networks onto a single atom backbone [32, 33]. For example, the aldolase reaction, which cleaves fructose-1,6-bisphosphate ($C_6H_{14}O_{12}P_2$) into dihydroxyacetone phosphate ($C_3H_7O_6P$) and glyceraldehyde-3-phosphate ($C_3H_7O_6P$), can be mapped onto a carbon atom backbone, becoming simply $C_6 \leftrightarrow C_3 + C_3$ (see Methods). This reaction is now formally analogous to the $a_6 \leftrightarrow a_3 + a_3$ reaction in the idealized chemistry.

The first question is whether the structure of real metabolic networks allows interconversions that use the optimal, logarithmic number of steps found for the artificial chemistry (Fig. 3 and Eq. 1). A real metabolic network might be thought of as the overlap of several MBPs, with the additional complexity of multiple atoms and multi-substrate, multi-product reactions. To search for patterns beneath this complexity, we identified all shortest pathways between any two carbon compounds in *Escherichia coli*'s metabolic network [34], pruned of highly connected cofactors that do not participate in carbon transfers (see Methods). We first determined, for each input compound, the minimal path length to reach its closest molecule with *j* carbons; then, for each value of *i*, we averaged these path lengths over all input molecules with *i* carbons. The results (Fig. 3, black curves, and Fig. S4) show that these *E. coli* minimal path lengths approximately follow the predicted logarithmic trend. For some curves (e.g. the one with $C_5$ as an input), the specific peaks and valleys of the predicted function are closely followed by the *E. coli* network. While this does not prove that MBPs are indeed used in real metabolic networks, it demonstrates that the logarithmic strategy of MBPs is embedded in their architecture.

As in the case of the artificial chemistry network, we can now search for signatures of optimality in the collective set of all metabolic reactions known in living systems, obtained from the KEGG database [35]. The presence of such signatures would suggest a long-term selective advantage of molecules and reactions that are useful for multiple tasks across different organisms and environments. By counting how many times each possible carbon backbone reaction is used across this biosphere-level metabolism we obtained a broad distribution, and a fit to a power-law gives the exponent of -0.89, comparable with the analytically predicted value, and with that in the artificial chemistry model (Fig. 4A and Fig. S4). Most surprisingly, we found that several reactions that are top ranking in their count across MBPs in the artificial network, are also at the top of the list in the KEGG-derived reactions (Spearman correlation p-value$<10^{-6}$; see also Table 1, and Supplementary Spreadsheet). This suggests that the $R_N$ network model, despite its simplified chemical rules, captures some fundamental features of the role of the carbon reaction backbone of real metabolic networks.

In addition to a preference for specific reactions, we can ask whether the spectrum of metabolite usage across the whole KEGG metabolism reflects possible optimality criteria (Fig. 4B). The metabolite usage in the hydrogen backbone network (see Methods) is similar to that in the artificial chemistry: each even-length hydrogen metabolite is used more often than its odd-length neighbors (Fig. S5C). For the carbon backbone distribution, we see a similar descending periodic behavior, but with a periodicity of approximately 5 (Fig. 4B and Fig. S5B). Hence, molecules containing carbons in a number that is multiple of 5 are used more abundantly than other molecules across different metabolic reactions. One possible explanation for this $C_5$ periodicity is the profuse usage of adenine and nicotinamide adenine dinuculeotide compounds as energy carriers and redox balance molecules, although the removal of such compounds has little effect on the observed periodicity (Fig. S7). Hence, the prominent usage of compounds with specific numbers of carbons might reflect global network optimization principles for the efficiency of multiple pathways, as observed in the artificial chemistry model. The periodicity of 5 that we observe, together with the evidence displayed in Fig. 3C, may suggest that the evolutionary optimization of metabolism has been partially taking place around building blocks of five carbons, compatible with previous observations of prebiotic abundance of terpenoids [1] and pentoses [36]. It is also interesting to note that an unexplained periodicity of two had been previously observed in the distribution of the number of carbons among known organic compounds [37-39]. While our analysis is based on the distribution of usage of carbon compounds in different reactions, rather than the total count of molecules, future analyses may investigate possible connections between these trends.

## Discussion

The multitude of metabolic tasks and environments encountered by living systems and the prevalence of horizontal gene transfer may have collectively caused metabolic networks to evolve towards an architecture that allows multiple tasks to be performed in a near optimal way. In an attempt to mirror these multiple evolutionary optimization processes, we systematically computed a complete set of elementary optimal pathways in a simple artificial chemistry universe. Several properties emerging from these optimal pathways seem to have unexpected correspondences in real metabolic networks, pointing to the possible existence of universal rules that may govern the evolution of metabolic network wiring. Some constraints used in the artificial chemistry model – most notably that no waste is produced and that optimal tasks are single-input/single-output pathways that do not take into account free energies, may be relaxed in future versions of the model, using more realistic chemical rules [40, 41]. At the same time, more complex models will carry the burden of increased computational cost, and possibly less interpretable results. One may hope that if any fundamental principle is truly at work underneath the evolution of metabolism, its consequences might be robustly detectable for a broad category of chemical rules.

## Materials and Methods

### *1. Artificial Chemistry Model*

We define an artificial chemistry inspired by previous string-based artificial chemistries (see also main text and Fig. 1). One may think of molecules in this artificial chemistry as polymers (up to a given length *N*) of a monomeric unit *a*. Since no specific assumption is made in the model about the nature of these molecules, they could equally represent aggregates or branched polymers of different sizes, as well as molecules with different counts of a specific atom. A network $R_N = \{M_N, C_N\}$ is defined by the set of $N$ molecules $M_N = \{a_i \mid \forall\ i = 1,\ldots,N\}$ and the set of all possible uni-bi ligation/lysis reactions between them, $C_N = \{a_i + a_j \leftrightarrow a_k \mid \forall\ i, j, \text{ and } k, \text{ such that } i \le j,\ i + j = k, \text{ and } k = 2,\ldots,N\}$.

*2. Flux Balance Analysis*

Flux Balance Analysis (FBA) is a steady state constraint-based approach to study the flow of mass through metabolic networks [25, 42, 43]. Briefly, FBA represents the metabolic network of interest as an $n \times m$ stoichiometric matrix $S$, whose element $S_{ij}$ indicates the number of molecules of metabolite $i$ ($i$=1,…,$m$) that participate in reaction $j$ ($j$=1,…,$n$) (with a positive sign if the metabolite is produced, negative if it is consumed). Each reaction can be associated with a rate, or flux, $v_j$. Under the assumption of a steady state the following set of mass conservation constraints on the fluxes is generated:

$$\sum_{j=1}^{n} S_{ij} v_j = 0 \qquad\qquad i = 1,2,\ldots,m \tag{1}$$

Additional constraints (such as availability of nutrients, experimentally observed irreversibility, maximal or minimal rates, etc.) can be imposed on the fluxes as inequalities of the form

$$\alpha_j \le v_j \le \beta_j \tag{2}$$

where $\alpha_j$ is the minimal allowed rate of a reaction and $\beta_j$ is its maximal rate. Taken together, the above constraints define a convex polyhedron (the “feasible space”) in the $n$-dimensional space of fluxes. Linear programming (LP) can be used to identify, within the feasible space, flux vectors that maximize or minimize a given linear objective function. In microbial systems it has been often hypothesized that a biologically meaningful objective is the maximization of the flux through the reaction that represents cellular growth, or biomass production [13, 44]. Hence, LP applied to FBA provides a prediction of all metabolic fluxes in a cell. FBA can be applied at genome scale, and corresponding stoichiometric models are available for a number of organisms. FBA predictions have been experimentally validated most thoroughly in *Saccharomyces cerevisiae* SC288 [45] and *E. coli* K-12 [34].

*3. Minimal Balanced Pathway discovery algorithms.*

Minimal Balanced Pathways (MBPs) are defined as sets of reactions in the $R_N$ network that can optimally perform a given metabolic task. A task is defined as the production of a specific end product (e.g., $a_j$, with output flux $v_{out}$) from a single available nutrient (e.g., $a_i$, with input flux $v_{in}$). A pathway between two molecules is a MBP if (i) it satisfies a steady state solution, analogous to Eq. 1; (ii) it produces the final product with maximal yield, i.e., $v_{out}=v_{in}\cdot j/i$; and (iii) it contains the smallest possible number of reaction steps. The MBP between $a_i$ and $a_j$ will be indicated as $a_i \Rightarrow a_j$.

We have developed three different algorithms for computing MBPs, as described below:

a. Flux Balance Analysis/Mixed Integer LP algorithm

We use a modified FBA approach to formulate the MBP problem in a constrained optimization framework. Specifically, we impose the same constraints used in an FBA problem, and further require that the maximal yield condition $v_{out}=v_{in}\cdot j/i$ be satisfied. We then search for a solution that minimizes the number of active (nonzero) fluxes. Towards this goal, we use a modification of the LP problem described above to introduce binary variables ($b_j$) that represent flux activity: $b_j$=0 if $v_j$=0, and $b_j$=1 otherwise. To identify a minimal path, we can then search for the set of fluxes that minimize $\Sigma_i b_i$. Because of the nature of the variables involved – the fluxes are continuous, and the number of active fluxes is an integer – this problem must be solved using a mixed integer linear programming (MILP) algorithm. Our MILP problem for the optimal MBP between $a_n$ and $a_m$ can be formulated as follows:

Minimize $\sum_{j=1}^{N} b_j$

Subject to:

$$\sum_{j=1}^{n} S_{ij} v_j = 0 \qquad (5)$$

$$\alpha_j b_j \le v_j \le \beta_j b_j \qquad \text{for } j = 1, 2, \dots, n$$

$$v_{out} = \frac{m}{n} v_{in}$$

$$b_j \in \{0,1\} \qquad \text{for } j = 1, 2, \dots, n$$

The optimal solution for this problem will give the flux distribution $v$ that uses the fewest nonzero values to maximize the objective. In our MBP computations, the only flux constraints used were those that limit the uptake of the single nutrient to an arbitrary value of 10 mmol/grDM·h, and the production of the target metabolite to the known maximal yield $v_{out}/v_{in}=j/i$.

b. Elementary Flux Modes algorithm

Given a metabolic network defined by a stoichiometric matrix $S$ (as described in the above FBA section), a vector of fluxes $v$ is said to correspond to an Elementary Flux Mode (EFM) if it satisfies the following three conditions [18].

1. It satisfies the steady state condition ($Sv = 0$).
2. It must be feasible within the conditions of the model: if there are known boundaries for the fluxes, then $v$ must fall within them.
3. It must be non-decomposable. There are no two smaller EFMs that can be linearly combined to form the one in question.

Because of these constraints, those EFMs that use the minimal number of reactions satisfy the requirements for being an MBP. We used the METATOOL software package [46] to find all EFMs in the $R_{10}$ network, and then identified all of those EFMs that are also MBPs.

c. Iterative additive algorithm

We designed and implemented an algorithm to produce most MBPs *de novo*, without relying on prior steady state stoichiometric modeling methods. The algorithm works in an iterative manner, producing longer pathways from shorter ones. For example, we can start from two trivial MBPs: $a_1 \Rightarrow a_1$ (which requires no reactions), and $a_1 \Rightarrow a_2$ (requiring one trivial reaction, $a_1 + a_1 \rightarrow a_2$). To compute $a_1 \Rightarrow a_3$, we identify all the ways in which we can decompose 3 into two smaller addends (in this case, only one:

3=2+1). Next we combine together the previously computed MBPs that progress from $a_1$ to each of these two addends, giving a new putative MBP for the desired new task ($a_1 + a_1 \rightarrow a_2$, and $a_1 + a_2 \rightarrow a_3$). This procedure can be then iterated to give a prediction of MBP $a_i \Rightarrow a_j$ ($i, j \leq N$).

This algorithm is fast and efficient compared to the previous methods, allowing us to apply it to even the $R_{100}$ network. However, it has two main drawbacks. First, it will miss pathways that "overshoot" the target value then subtract down to it. Second, it may miss MBPs that are not built modularly from smaller ones. From a comparison of the MBPs predicted by the different algorithms, one can see that the approximations introduced in this algorithm cause 18 out 361 MBPs (5%) in $R_{19}$ to overestimate pathway length by one reaction. Also, this algorithm correctly identifies 204 of the 384 degenerate MBPs that the EFM algorithm finds in $R_{10}$. The reaction usage using this method is highly correlated with that of the MILP method applied to $R_{19}$ (Pearson correlation 0.96, p-val $10^{-51}$), and the EFM method applied to $R_{10}$ (Pearson correlation 0.98, p-val $2 \cdot 10^{-17}$).

*4. KEGG reaction reduction*

Data used for the comparison between the $R_N$ metabolic network and real metabolism was gathered from the KEGG LIGAND database (July 26, 2009 release)[35]. This database was parsed to convert its compounds and reactions into a single-atom form, as described in the text. Compounds that carried any uncertainty in their atomic makeup, including non-specific side-chains or variable chain length were removed from the current analysis. We also removed from the analysis reactions with no associated formula, as well as reactions involving non-specific molecules (such as generic glycans and non-specific nucleotide or peptide chains). Finally, a number of reactions were found to leave the atomic composition of the compounds essentially unchanged on either side of the reaction (e.g., $C_3 \leftrightarrow C_3$). These reactions were ignored as well, without consequences on the results (data not shown).

*5. Metabolite and reaction usage*

We counted how often each metabolite and reaction was used in the artificial chemistry pathways as well as in the KEGG-derived single-atom networks. In the model pathways, reaction usage was calculated by counting how many times each reaction was used across all pathways. Metabolite usage was similarly calculated by counting the occurrence of reactions in which each metabolite participates. For example, in the pathways that convert $a_9$ to $a_{10}$ in Fig. 2C, $a_9$ participates in only one reaction, but $a_{10}$ participates in two.

In the KEGG-derived networks, a similar counting scheme was used. The reaction usage was calculated by counting how many times each reduced reaction appears, and the metabolite usage was calculated by counting how many times each metabolite appears across all reactions.

*6. Shortest paths in Escherichia coli*

We calculated the lengths of the shortest pathways in the metabolic network of *Escherichia coli*, using the genome-scale iJR904 model [34] which has 761 metabolites and 1075 reactions. Because there are both metabolites and reactions, metabolic networks are inherently bipartite: metabolites connect to each other only through reactions. Pathway lengths were computed by transforming the metabolic reaction network stoichiometric matrix into an adjacency matrix where metabolites and reactions are all represented by the same type of node.

In this part of our analysis we are only interested in the connections between carbon compounds, so we removed any non-carbonaceous metabolites (water, phosphate, ammonia, etc.). Also, we removed the following cofactors that are used in many reactions, but do not participate in the transformation of carbons: ATP, ADP, AMP, NAD+, NADH, NADP+, NADPH, coenzyme A, acetyl-CoA, and the acyl carrier protein.

Next, we used Johnson's all-pairs shortest paths algorithm (available as a Matlab function) to find shortest pathways between any two carbon compounds in *E. coli*'s metabolic network. For each input compound, we listed the shortest paths to all output compounds containing a number j of carbons. For each j, we select among these paths the shortest one, giving an estimate of the shortest path between any individual compound and the closest *j*-carbon compound. Finally, for each value of *i*, we averaged these path lengths over all input molecules with *i* carbons. The end result is a matrix that provides the average of the shortest paths from any $i$-carbon compound to its nearest $j$-carbon compound.

*7. Analytical estimate of MBP lengths in analogy with addition-subtraction chains*

We developed an analytical approximation for the expected numbers of reactions to be found in any MBP $a_i \Rightarrow a_j$. We begin with a simplified version of the artificial chemistry model in which only irreversible addition reactions of the form

$$a_p + a_q \to a_{(p+q)} \tag{1}$$

are allowed. Under these restrictions, we first ask what is the smallest number of reactions necessary to produce any $a_j$ from $a_1$. We shall denote by $l(j)$ the smallest possible number of such reactions (we count the use of each reaction (1) once). This problem is equivalent to the problem of addition chains [26], in which one attempts to compute a positive integer by generating a sequence of integers such that each term in the sequence is the sum of two previous terms. Addition chains have been studied extensively, mainly because of their applications in computer science and cryptography [26]. For addition chains, $l(j)$ grows logarithmically with $j$:

$$l(j) \propto \log_2(j) \tag{2}$$

Our artificial chemistry represents a generalization, in which a metabolite of any length $i$ can be used to produce an output metabolite of any length $j$. If we still assume that only addition reactions are possible (i.e. molecules cannot be broken down), a chain from $a_i$ to $a_j$ will exist only when $i$ is a divisor of $j$. The problem can then be reduced to the case with $a_1$ input and $a_{j/i}$ output. Therefore, in the irreversible case, we can assume that inputs consist of monomers without loss of generality. Let $L(j)$ be the length of the shortest reaction chain in this case. Because not every reactant exists when dividing by the input length $i$, we have the obvious inequality

$$l(j) \le L(j) \tag{3}$$

Sometimes the shortest chain can be found easily. For instance, $\{2^0, 2^1, \ldots, 2^k\}$ is obviously the shortest chain from 1 to $j = 2^k$ whose length is $k + 1$. This suggests the general lower bound on the shortest length $L(j)$ of the addition chain:

$$L(j) \ge \lceil \log_2(j) \rceil + 1 \tag{4}$$

where $\lceil x \rceil$ represents the ceiling of $x$, or the smallest integer not less than $x$. Likewise, as seen below, $\lfloor x \rfloor$ represents the floor of $x$, or the largest integer not greater than $x$ (for example, $\lceil 3.14 \rceil = 4$, and $\lfloor 3.14 \rfloor = 3$). The longest minimal addition chain arises when the output length is $j = 2^m$-1. From this fact, we have the upper bound [26]

$$L(j) \leq \lfloor \log_2(j) \rfloor + \upsilon(j) \tag{5}$$

where $\upsilon(j)$ is the number of 1s in the binary representation of $j$. Since $\upsilon(j) \leq \lfloor \log_2(j) \rfloor + 1$, the bound in equation (5) implies a simpler (but weaker) upper bound

$$L(j) \leq 2\lfloor \log_2(j) \rfloor + 1 \tag{6}$$

The above bounds give precise values in some cases and act as bounds in others. For instance, $L(16) = 5$, and $L(17) = L(18) = 6$ for both the lower and upper bounds, while $L(31) = 8$ is between the lower bound and the upper bound (5 and 9, respectively).

There are various conjectures regarding $L(j)$; one of the most famous [31] asserts that computing $L(j)$ is NP-hard. Nonetheless, the computation of $L(j)$ has been pushed up to $n \leq 2^{25}$. Two other conjectures [47] predict the general lower bound

$$L(j) \geq \lfloor \log_2(j) \rfloor + \log_2 \upsilon(j) + 1 \tag{7}$$

and the upper bound

$$L(2^k - 1) \leq k + L(k) - 1 \tag{8}$$

While algorithms for generating the shortest addition chains are discussed by Thurber [47], these all hold for the specific case of pure addition where the input is always $a_1$.

We are interested in the general case involving both addition and subtraction, and specifically the lengths $l(i, j)$ of the shortest reaction chains (MBPs) with $a_i$ input and $a_j$ output. Addition-subtraction chains have also been studied previously as an expansion of addition chains, although these correspond to MBPs with only $a_1$ as an input. Sometimes, in these cases, $l(j)$ is readily computable, e.g.

$$l(2^k - 1) = k + 2 \qquad \text{for } k \geq 3 \tag{9}$$

while $L(2^k - 1)$ remains unknown for sufficiently large $k$. Both lengths can also be equal, i.e. $l(j) = L(j)$. For example,

$$l(2^k) = L(2^k) + k = 1 \tag{10}$$

$$l(2^k + 1) = L(2^k + 1) = k + 2 \tag{11}$$

Note also an inequality:

$$l(j) \geq \lceil \log_2(j) \rceil + 1 \tag{12}$$

All of these features explain the growth law in equation (2).

The quantity $l(i, j)$ has a rich behavior, e.g., there is only a trivial lower bound since $l(j, j) = 1$. To ignore this non-interesting effect, let us divide $i$ and $j$ by their greatest common divisor as it never affects the length of the MBP:

$$l(i,j) = l\left(\frac{i}{d}, \frac{j}{d}\right), \qquad d = \gcd(i, j) \tag{13}$$

then we can use an obvious inequality

$$l(i,j) \leq l(i,1) + l(1,j) = l(i) + l(j) \tag{14}$$

Recalling (2) we finally arrive at an approximation for the number of reactions in an MBP that uses $a_i$ to produce $a_j$:

$$l(i,j) \sim \log_2\left(\frac{i}{d}\right) + \log_2\left(\frac{j}{d}\right) \tag{15}$$

The approximation in (15) can also be used to estimate the rank distribution of reaction usage. Consider all possible MBPs producing $a_j$ from $a_i$. For each $(i, j)$ pair, take an MBP and mark all reactions. Let the reaction in (1) occur $E_{pq}$ times: that is, there are $E_{pq}$ MBPs that use (1). We now divide $E_{pq}$ by the total number of MBPs and call $e_{pq} = N^{-2} E_{pq}$ the reaction frequency. It is better to order reactions not according to $(p, q)$ but to their ranking $j$, so that the reaction of rank $j = 1$ is the most frequent, that of rank $j = 2$ is the second in frequency, etc. This gives $e_j$. How does $e_j$ decrease with rank? To infer the answer we note that

$$\sum_{j=1}^{N^2} e_j = \langle l \rangle \tag{16}$$

From (15) it is clear that the average length $\langle l \rangle$ of the shortest reaction chain scales as log $N$. This is consistent with (16) if and only if we have $r_j \sim j^{-1}$. Thus we predict the power-law decay

$$r_j \sim j^{-1} \text{ when } j >> 1 \tag{17}$$

## Acknowledgements

This work was partially supported by grants from the U.S. Department of Energy (DE-FG02-07ER64388 and DE-FG02-07ER64483), NASA (NASA Astrobiology Institute, NNA08CN84A), and the National Science Foundation (NSF DMR0535503 and NSF CCF-0829541).

## References

1. Ourisson G, Nakatani Y (1994) The terpenoid theory of the origin of cellular life: the evolution of terpenoids to cholesterol. Chem. Biol 1: 11-23.

2. Miller SL (1953) A production of amino acids under possible primitive earth conditions. Science 117: 528-529.

3. Orgel LE (1998) The origin of life--a review of facts and speculations. Trends in Biochemical Sciences 23: 491-495.

4. Morowitz HJ, Kostelnik JD, Yang J, Cody GD (2000) The origin of intermediary metabolism. Proc. Natl. Acad. Sci. U.S.A 97: 7704-7708.

5. Kauffman SA (1993) The Origins of Order: Self-Organization and Selection in Evolution. 1st ed. Oxford University Press, USA.

6. Fontana W, Buss LW (1994) What would be conserved if "the tape were played twice"? Proc. Natl. Acad. Sci. U.S.A 91: 757-761.

7. Ganti T (2003) The Principles of Life. Oxford University Press, USA.

8. Kun A, Papp B, Szathmáry E (2008) Computational identification of obligatorily autocatalytic replicators embedded in metabolic networks. Genome Biol 9: R51.

9. Smith E, Morowitz HJ (2004) Universality in intermediary metabolism. Proc. Natl. Acad. Sci. U.S.A 101: 13168-13173.

10. Ebenhöh O, Heinrich R (2001) Evolutionary optimization of metabolic pathways. Theoretical reconstruction of the stoichiometry of ATP and NADH producing systems. Bull. Math. Biol 63: 21-55.

11. Koonin EV (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Micro 1: 127-136.

12. Lenski RE, Winkworth CL, Riley MA (2003) Rates of DNA sequence evolution in

experimental populations of Escherichia coli during 20,000 generations. J. Mol. Evol 56: 498-508.

13. Fong SS, Palsson BØ (2004) Metabolic gene–deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. Nat Genet 36: 1056-1058.

14. Lee M, Chou H, Marx CJ (2009) Asymmetric, bimodal trade-offs during adaptation of Methylobacterium to distinct growth substrates. Evolution Available at: http://www.ncbi.nlm.nih.gov/pubmed/19545267. Accessed 29 July 2009.

15. Baldwin JE, Krebs H (1981) The evolution of metabolic cycles. Nature 291: 381-382.

16. Beasley JE, Planes FJ (2007) Recovering metabolic pathways via optimization. Bioinformatics 23: 92-98.

17. Raymond J, Segrè D (2006) The effect of oxygen on biochemical networks and the evolution of complex life. Science 311: 1764-1767.

18. Klamt S, Stelling J (2003) Two approaches for metabolic pathway analysis? Trends Biotechnol 21: 64-69.

19. Edwards JS, Covert M, Palsson B (2002) Metabolic modelling of microbes: the flux-balance approach. Environ. Microbiol 4: 133-140.

20. Hintze A, Adami C (2008) Evolution of complex modular biological networks. PLoS Comput. Biol 4: e23.

21. Dittrich P, Ziegler J, Banzhaf W (2001) Artificial Chemistries—A Review. Artificial Life 7: 225-275.

22. Srinivasan V, Morowitz HJ (2009) The canonical network of autotrophic intermediary metabolism: minimal metabolome of a reductive chemoautotroph. Biol. Bull 216: 126-130.

23. Meléndez-Hevia E, Waddell TG, Cascante M (1996) The puzzle of the Krebs citric acid cycle: assembling the pieces of chemically feasible reactions, and opportunism in the design of metabolic pathways during evolution. J. Mol. Evol 43: 293-303.

24. Papp B, Teusink B, Notebaart RA (2009) A critical view of metabolic network adaptations. HFSP J 3: 24-35.

25. Kauffman KJ, Prakash P, Edwards JS (2003) Advances in flux balance analysis. Curr. Opin. Biotechnol 14: 491-496.

26. Knuth DE (1997) Art of Computer Programming, Volume 2: Seminumerical Algorithms (3rd Edition). 3rd ed. Addison-Wesley Professional.

27. Ricardo A, Frye F, Carrigan MA, Tipton JD, Powell DH, et al. (2006) 2-Hydroxymethylboronate as a Reagent To Detect Carbohydrates: Application to the Analysis of the Formose Reaction. The Journal of Organic Chemistry 71: 9503-9505.

28. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL (2002) Hierarchical organization of modularity in metabolic networks. Science 297: 1551-1555.

29. Maslov S, Krishna S, Pang TY, Sneppen K (2009) Toolbox model of evolution of prokaryotic metabolic networks and their regulation. Proc. Natl. Acad. Sci. U.S.A 106: 9743-9748.

30. Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of Escherichia coli. Nat. Genet 31: 64-68.

31. Downey P, Leong B, Sethi R (1981) Computing Sequences with Addition Chains. SIAM J. Comput. 10: 638-646.

32. Meléndez-Hevia E, Waddell TG, Montero F (1994) Optimization of Metabolism: The Evolution of Metabolic Pathways Toward Simplicity Through the Game of the Pentose Phosphate Cycle. Journal of Theoretical Biology 166: 201-220.

33. Ebenhöh O, Heinrich R (2003) Stoichiometric design of metabolic networks: multifunctionality, clusters, optimization, weak and strong robustness. Bull. Math. Biol 65: 323-357.

34. Reed JL, Vo TD, Schilling CH, Palsson BO (2003) An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome Biol 4: R54.

35. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, et al. (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34: D354-357.

36. Ricardo A, Carrigan MA, Olcott AN, Benner SA (2004) Borate minerals stabilize ribose. Science 303: 196.

37. Desiraju GR, Jack D. Dunitz, Ashwini Nangia, Jagarlapudi A. R. P. Sarma, Engelbert Zass (2000) The Even/Odd Disparity in Organic Compounds. Helvetica Chimica Acta 83: 1-15.

38. Sarma JARP, Nangia A, Desiraju GR, Zass E, Dunitz JD (1996) Even–odd carbon atom disparity. Nature 384: 320-320.

39. Morowitz HJ (1979) Energy Flow in Biology. Ox Bow Press.

40. Beard DA, Liang S, Qian H (2002) Energy balance for analysis of complex metabolic

networks. Biophys. J 83: 79-86.

41. Benkö G, Flamm C, Stadler PF (2003) A graph-based toy model of chemistry. J Chem Inf Comput Sci 43: 1085-1093.

42. Edwards JS, Palsson BO (2000) Metabolic flux balance analysis and the in silico analysis of Escherichia coli K-12 gene deletions. BMC Bioinformatics 1: 1.

43. Varma A, Palsson BO (1994) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. Appl. Environ. Microbiol 60: 3724-3731.

44. Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. Mol. Syst. Biol 3: 119.

45. Mo ML, Palsson BO, Herrgård MJ (2009) Connecting extracellular metabolomic measurements to intracellular flux states in yeast. BMC Syst Biol 3: 37.

46. Kamp AV, Schuster S (2006) Metatool 5.0: fast and flexible elementary modes analysis. Bioinformatics 22: 1930-1931.

47. Thurber EG (1999) Efficient Generation of Minimal Length Addition Chains. SIAM Journal on Computing 28: 1247.

## Figure Legends

### Figure 1

Representation of the $R_4$ network. (**A**) This schematic of the $R_4$ artificial chemistry network is composed of metabolite "strings" of $a$ up to a maximum length of four, and all allowed reactions between them.

(**B**) The reaction list for the $R_4$ network. There are four reactions that represent the exchange of mass with the environment ($r_1$-$r_4$) – one for each metabolite – and four reactions between the metabolites ($r_5$-$r_8$).

### Figure 2

Emergent complexity and modularity in artificial network topology. (**A**) This set of example MBPs displays the emergent modularity of structure and function, as well as the multiple usage of different reactions. Each row denotes the input metabolite used, and each column the output metabolite. The reaction marked in red interconverts $a_2 + a_2 \leftrightarrow a_4$ and is the most used reaction. MBPs on a yellow background are autocatalytic cycles. For a larger image with more examples, and a more detailed view of the properties observed in these networks, see Figure S1. (**B**) A single autocatalytic loop is used as a modular backbone for several MBPs: the cycle that constructs $a_8$ from $a_7$ is also used to produce $a_1$, $a_2$, and $a_4$. (**C**) Four examples of the MBP that produces $a_{10}$ from $a_9$. Each breaks down $a_{10}$ into $a_1$ in different but equally optimal ways. Each of these sub-pathways (gray metabolites) is an MBP in itself, showing the modularity of use of each of these metabolic tools.

### Figure 3

Logarithmic growth is observed among pathways from both the $R_{19}$ and *E. coli* metabolic networks. Each plot has a single starting value $i$ corresponding to either $a_i$ (for the model) or $C_i$ (for *E. coli*). The red lines show the number of reactions in each MBP with different output metabolite size, and the black lines show the average number of reactions used to reach the nearest metabolite of increasing carbon number in *E. coli*. The predicted number of reactions from Equation 1 is also shown for each plot in blue. (**A**) $C_1/a_1$ input. (**B**) $C_2/a_2$ input. (**C**) $C_5/a_5$ input.

**Figure 4**

Metabolite and reaction usage frequencies. (**A**) The frequency of usage of reactions in the artificial chemistry model and in the KEGG-derived carbon reaction set. MBPs for the $R_{19}$ network were calculated using the MILP method while the others ($R_{30}$ through $R_{100}$) were estimated using the iterative algorithm. We calculated the reaction usage by counting the number of MBPs that use each reaction. These were then ranked in descending order, yielding curves that follow a power law with an average exponent of -1.14 (+/- 0.03) ($R^2$=0.99). The reaction usage in the KEGG-derived carbon dataset was calculated by counting the number of times each equivalent reaction appears, and follows a power law tail distribution, with exponent -0.89. The curve predicted by the analytical model, with exponent -1, is shown as a solid line. (**B**) The usage frequency of each metabolite among all MBPs in the $R_{19}$ model. This also shows the frequency of use of each metabolite in the $R_{19}$ network itself, and in a randomly chosen set of reactions (control). In the inset, the metabolite usage was sorted by rank and plotted on a semilog axis. (**C**) The usage frequency of each metabolite among all reactions in the KEGG carbon reaction set. The inset shows the usage sorted by rank on a semi-log scale.

## Tables

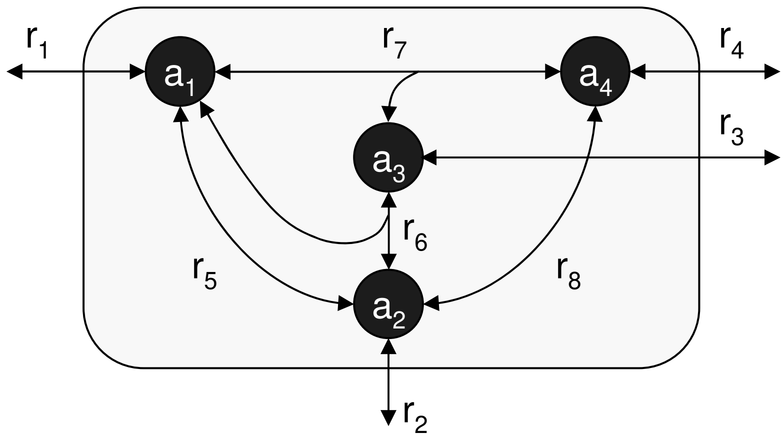
### Table 1

Of the top 10 most used reactions in the $R_{19}$ network and carbon-only KEGG network, there are five equivalent reactions that appear in both. Recognizing that 74 of the 90 possible reactions from the $R_{19}$ set are found in the 631 carbon-only reactions from KEGG, we can use the Spearman rank-correlation to find that this has a correlation value of 0.54 with p-value $8 \cdot 10^{-7}$.

| | **Model reaction** | **KEGG carbon reaction** |
|---|---|---|
| 1 | $a_2 + a_2 \leftrightarrow a_4$ | $\mathbf{C_6 + C_6 \leftrightarrow C_{12}}$ |
| 2 | $a_1 + a_1 \leftrightarrow a_2$ | $C_1 + C_5 \leftrightarrow C_6$ |
| 3 | $a_4 + a_4 \leftrightarrow a_8$ | $C_1 + C_3 \leftrightarrow C_4$ |
| 4 | $\mathbf{a_3 + a_3 \leftrightarrow a_6}$ | $\mathbf{C_1 + C_4 \leftrightarrow C_5}$ |
| 5 | $a_2 + a_4 \leftrightarrow a_6$ | $\mathbf{C_5 + C_5 \leftrightarrow C_{10}}$ |
| 6 | $\mathbf{a_6 + a_6 \leftrightarrow a_{12}}$ | $C_1 + C_7 \leftrightarrow C_8$ |
| 7 | $\mathbf{a_1 + a_2 \leftrightarrow a_3}$ | $C_1 + C_8 \leftrightarrow C_9$ |
| 8 | $a_8 + a_8 \leftrightarrow a_{16}$ | $\mathbf{C_3 + C_3 \leftrightarrow C_6}$ |
| 9 | $\mathbf{a_5 + a_5 \leftrightarrow a_{10}}$ | $C_4 + C_5 \leftrightarrow C_9$ |
| 10 | $\mathbf{a_1 + a_4 \leftrightarrow a_5}$ | $\mathbf{C_1 + C_2 \leftrightarrow C_3}$ |

# Figure 1

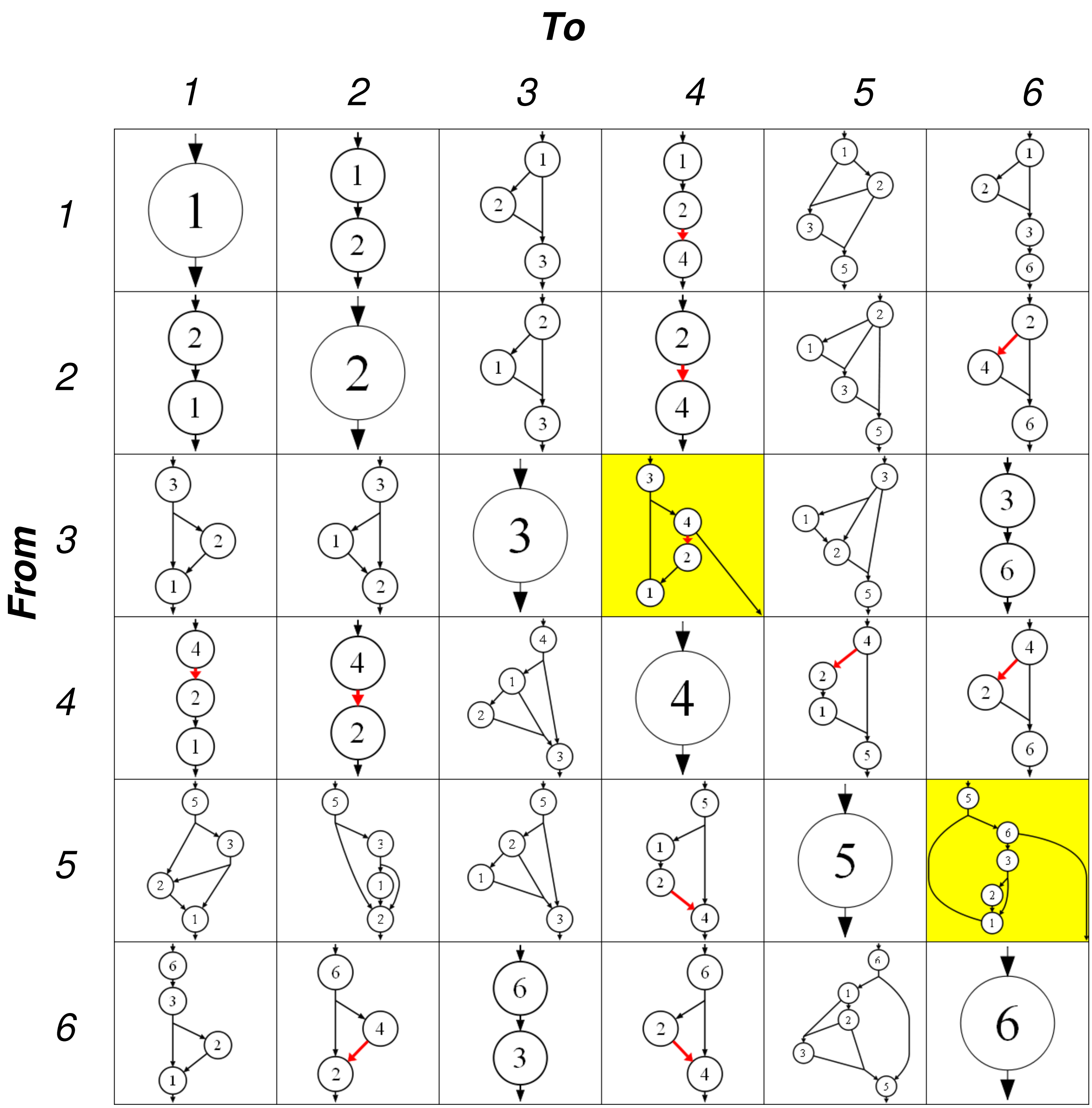
**A**




**B**

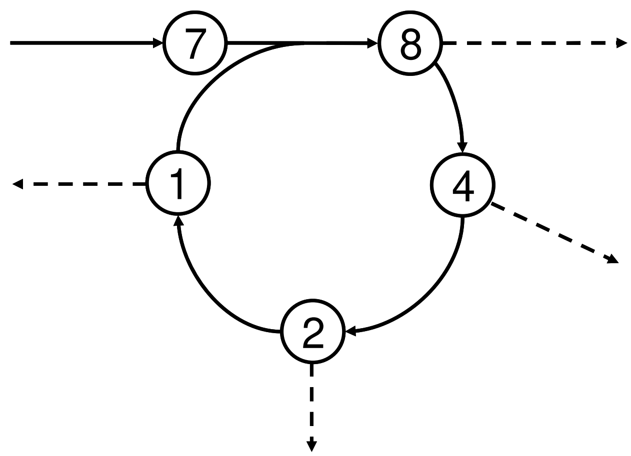
$r_1\!: \longleftrightarrow a_1$
$r_2\!: \longleftrightarrow a_2$
$r_3\!: \longleftrightarrow a_3$
$r_4\!: \longleftrightarrow a_4$
$r_5\!: a_1 + a_1 \longleftrightarrow a_2$
$r_6\!: a_1 + a_2 \longleftrightarrow a_3$
$r_7\!: a_1 + a_3 \longleftrightarrow a_4$
$r_8\!: a_2 + a_2 \longleftrightarrow a_4$

# Figure 2A

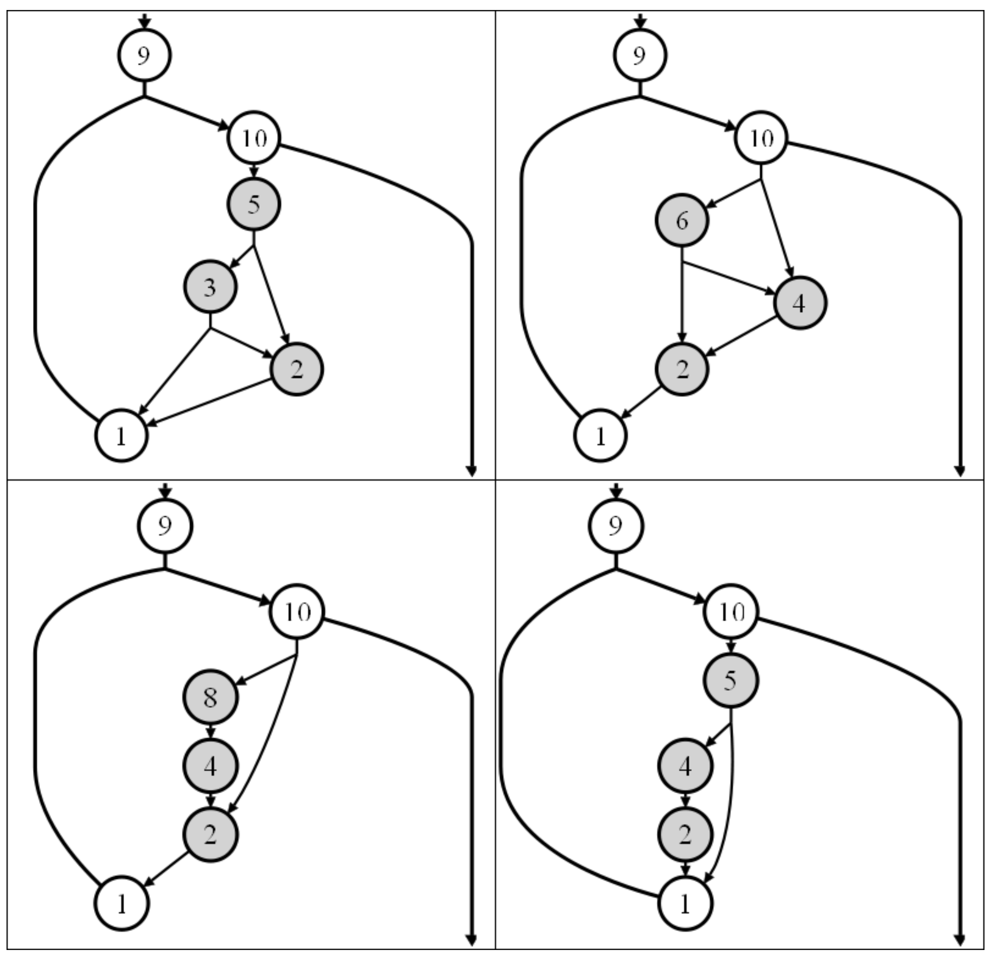


## Figure 2 (continued)

**B**

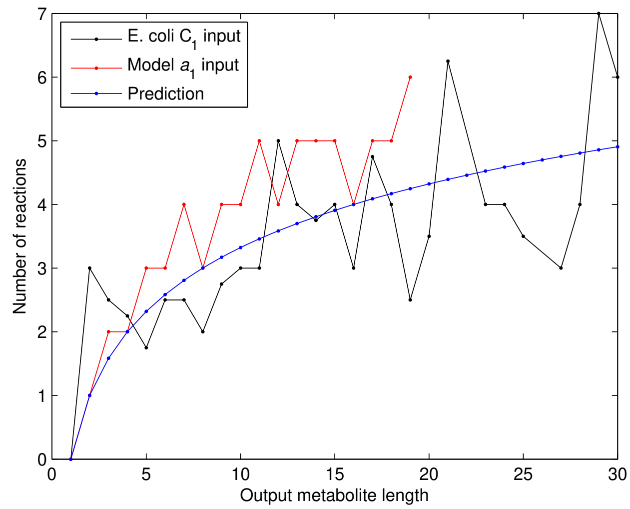



**C**



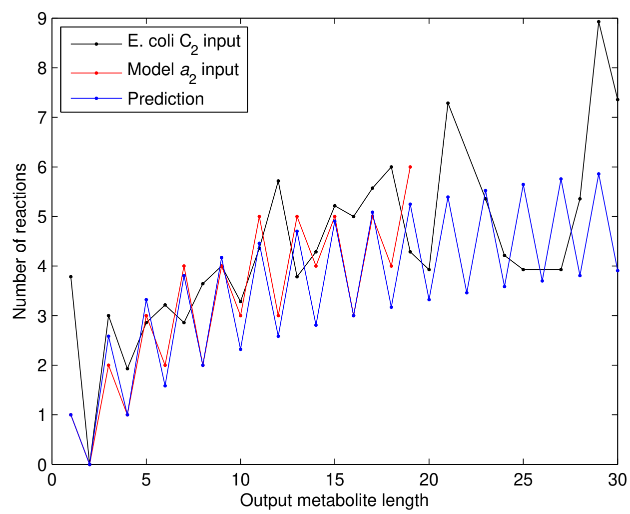
# Figure 3

A



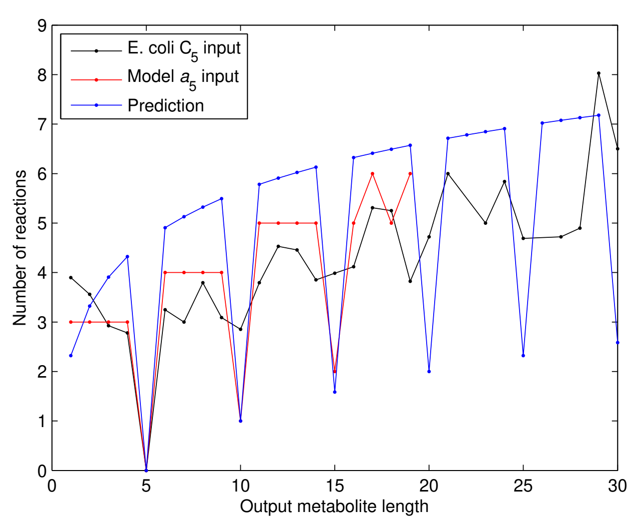
E. coli $C_1$ input
Model $a_1$ input
Prediction
Number of reactions
Output metabolite length


B



E. coli $C_2$ input
Model $a_2$ input
Prediction
Number of reactions
Output metabolite length

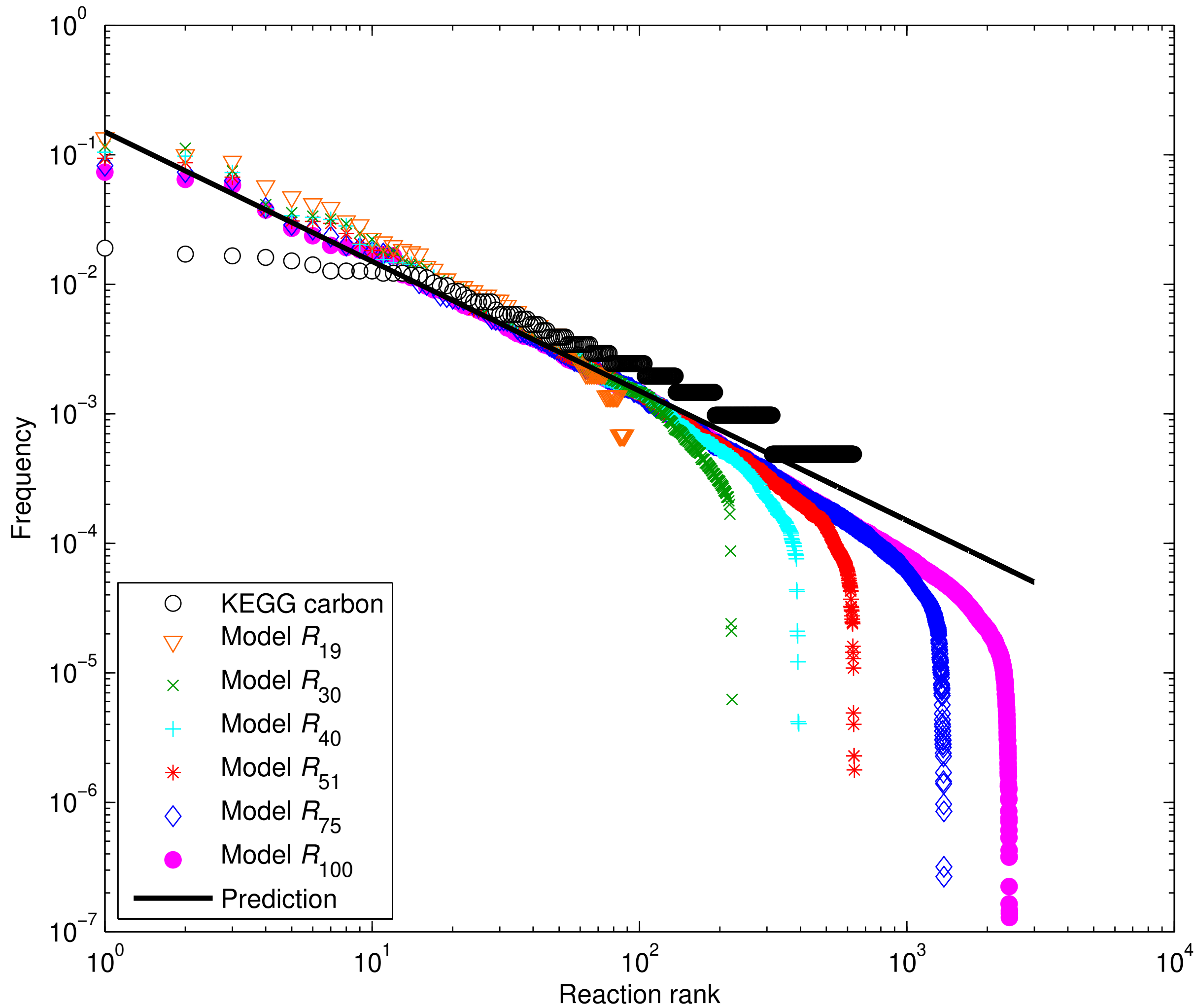

C



E. coli $C_5$ input
Model $a_5$ input
Prediction
Number of reactions
Output metabolite length

# Figure 4

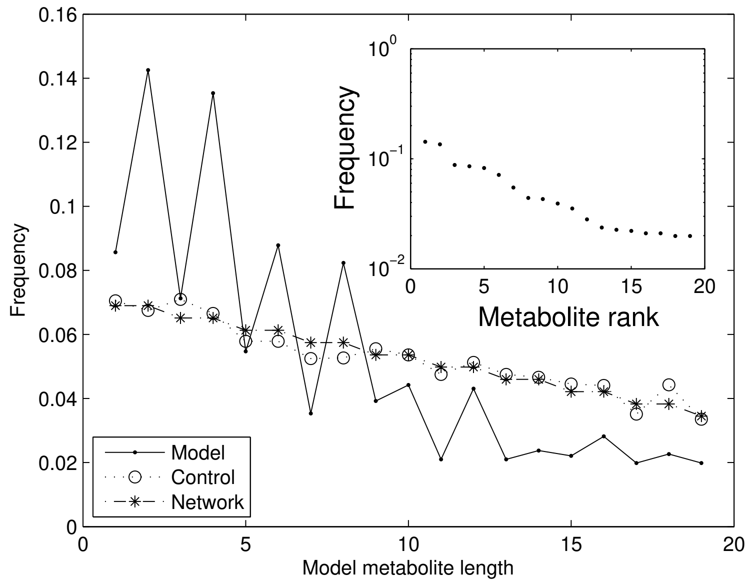
**A**



Frequency
Reaction rank
KEGG carbon
Model $R_{19}$
Model $R_{30}$
Model $R_{40}$
Model $R_{51}$
Model $R_{75}$
Model $R_{100}$
Prediction

# Figure 4 (continued)

**B**




**C**