Journal of Molecular Evolution (2021) 89:472–483
https://doi.org/10.1007/s00239-021-10018-0

ORIGINAL ARTICLE

Stoichiometric Modeling of Artificial String Chemistries Reveals Constraints on Metabolic Network Structure

Devlin Moyer[1,2] · Alan R. Pacheco[1,3] · David B. Bernstein[3,4] · Daniel Segrè[1,2,3,4,5]

Received: 27 October 2020 / Accepted: 12 June 2021 / Published online: 6 July 2021

Abstract

Uncovering the general principles that govern the structure of metabolic networks is key to understanding the emergence and evolution of living systems. Artificial chemistries can help illuminate this problem by enabling the exploration of chemical reaction universes that are constrained by general mathematical rules. Here, we focus on artificial chemistries in which strings of characters represent simplified molecules, and string concatenation and splitting represent possible chemical reactions. We developed a novel Python package, ARtificial CHemistry NEtwork Toolbox (ARCHNET), to study string chemistries using tools from the field of stoichiometric constraint-based modeling. In addition to exploring the topological characteristics of different string chemistry networks, we developed a network-pruning algorithm that can generate minimal metabolic networks capable of producing a specified set of biomass precursors from a given assortment of environmental nutrients. We found that the composition of these minimal metabolic networks was influenced more strongly by the metabolites in the biomass reaction than the identities of the environmental nutrients. This finding has important implications for the reconstruction of organismal metabolic networks and could help us better understand the rise and evolution of biochemical organization. More generally, our work provides a bridge between artificial chemistries and stoichiometric modeling, which can help address a broad range of open questions, from the spontaneous emergence of an organized metabolism to the structure of microbial communities.

Keywords Artificial chemistry · Flux balance analysis · Metabolism · Genome-scale metabolic models

Introduction

Metabolism occupies a central role in the functioning of biological systems, yet much remains unclear about the degree to which basic features of metabolic networks reflect either evolutionary accidents or optimal network structures (Pál et al. 2006; Barve and Wagner 2013; Noor et al. 2010; Ebenhöh and Heinrich 2001). In parallel to analyses focused on metabolism as we know it in individual organisms (Machado et al. 2018; Henry et al. 2010; Borenstein et al. 2008) or in the whole biosphere (Barve and Wagner 2013; Raymond and Segrè 2006; Handorf et al. 2005), multiple studies have explored the utility of abstract models of chemistry to investigate particular features of chemical networks. These models, also known as artificial chemistries, have the benefit of being unconstrained by the limits of what is known about extant metabolism and about its possible intermediate states lost through evolutionary history (Banzhaf and Yamamoto 2015; Benkö et al. 2003; Kauffman 1993).

Artificial chemistry has been used to study various aspects of the origin of life from abiotic chemistry (Guseva et al. 2017; Kauffman 1993; Banzhaf and Yamamoto 2015), common structural features of metabolic networks (e.g., hub metabolites) (Friedlander et al. 2015; Fontana and Buss 1994a, b; Pfeiffer et al. 2005), the general behavior of

Handling Editor: **Ashley Teufel**.

✉ Daniel Segrè
dsegre@bu.edu

[1] Bioinformatics Program, Boston University, Boston, MA 02215, USA

[2] Department of Biology, Boston University, Boston, MA 02215, USA

[3] Biological Design Center, Boston University, Boston, MA 02215, USA

[4] Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA

[5] Department of Physics, Boston University, Boston, MA 02215, USA

chemical (not necessarily biochemical) reaction networks (Benkö et al. 2003; Walter Fontana and Buss 1994a, b), the optimality (or lack thereof) of metabolic networks (Riehl et al. 2010; Soyer and Pfeiffer 2010), among other questions (Banzhaf and Yamamoto 2015). The artificial chemistry models used in these studies typically employ highly abstracted representations of chemistry (Riehl et al. 2010; Kauffman 1993; Banzhaf and Yamamoto 2015). However, more precise and realistic models involving either string rules based on formalization of real chemistry (like SMILES (Weininger 1988) and variants thereof (Arús-Pous et al. 2019; Lin et al. 2019)), or de novo approximate quantum mechanics computations (Benkö et al. 2003), have been used to explore the full space of possible real-life chemistry up to a certain degree of complexity (Lee et al. 2019). Artificial chemistry approaches have yielded many insights into general features of metabolism, but these findings have remained largely disconnected from the large body of metabolism research focused on characterizing real metabolic networks. We believe that many novel insights into metabolism will be enabled by combining artificial chemistry with techniques commonly used to study real metabolic networks.

The field of stoichiometric constraint-based modeling has provided many approaches that can be particularly useful for quantitatively understanding the structure and function of metabolic networks (Heirendt et al. 2019; Ebrahim et al. 2013; Gottstein et al. 2016; O'Brien et al. 2015). In particular, Flux Balance Analysis (FBA) is a common technique for studying metabolic networks at the level of a whole organism. FBA estimates the space of possible fluxes through a metabolic network at steady state, and is generally employed to identify metabolic states satisfying some biologically meaningful criterion of optimality (Orth et al. 2010). FBA has been used to simulate multiple types of experiments and phenotypes, such as growth rates and metabolic phenotypes of gene knockouts, growth efficiency on different media, and identification of potential drug targets (Orth et al. 2010; Gu et al. 2019; Kauffman et al. 2003; Yizhak et al. 2015). While FBA and stoichiometric constraint-based modeling have been widely used on the metabolic networks of real organisms, these techniques have only rarely been applied to artificial chemistry networks.

In the present work, we use a specific type of artificial chemistry known as a string chemistry, where each molecule is represented by a string of characters (Fig. 1) (Banzhaf and Yamamoto 2015; Kauffman 1993; Riehl et al. 2010). Our string chemistry model is relatively simple: all strings (i.e., metabolites) are linear sequences of characters (i.e., monomers, atoms, or functional groups) that may react by either concatenating end to end or splitting into two smaller strings (see "Methods" section). A particular string chemistry network is defined by the set of different characters each metabolite can be composed of and the maximum length a metabolite can reach. While these rules are much simpler than those governing real chemical reactions, Riehl et al. found structural similarities between real metabolic networks and string chemistry networks with only one type of character (i.e., the only difference between any two metabolites is their length) (Riehl et al. 2010), so we expect that string chemistries with more than one type of character may yield further insights into the general properties of metabolic networks. In string chemistry networks such as the one we use for this work, individual monomers (the letters in strings) could be thought of as elementary moieties (either atoms, or functional groups). While individual monomers cannot turn into each other (e.g., a letter "a" cannot transform into a letter "b"), one can think of strings such as "ab" and "ba" as more complex functional units that can transform into each other through a series of reactions.

In this manuscript, we describe the ARtificial CHemistry NEtwork Toolbox (ARCHNET), a Python package for generating string chemistry networks of arbitrary size and implementing stoichiometric modeling algorithms (including FBA) on those networks. Using this string chemistry framework, we created an algorithm for determining the minimal metabolic network capable of producing a given set of metabolites ("biomass precursors") from another set of metabolites ("environmental nutrients"). Our analysis of random choices of nutrients and biomass precursors in different string chemistry networks provides new insight into the rules governing which reactions are left in these minimal metabolic networks and suggests possible implications for the study of real metabolic networks.

Methods

Artificial Chemistry Model

The artificial chemistry model used here is an extension of the one used in Riehl et al. (2010) and is similar to several other previously used artificial chemistries (Banzhaf and Yamamoto 2015; Kauffman 1993; Fontana and Buss 1994a, b): each "chemical" is a string of characters of some arbitrary length, where each character represents an individual atom (or functional group, or monomer). A chemical may condense with one other chemical to produce a longer chemical; the two strings are simply concatenated (e.g., ab + aa → abaa). A chemical may also split into two smaller chemicals at any point along its length (e.g., ababb → ab + abb). Only pairwise condensation/dissociation reactions were considered due to the rarity of termolecular and higher reactions in real chemistry (Chang 2005; Laidler and Glasstone 1948; Compton et al. 2012). For simplicity, all reactions are modeled as being completely reversible, even though in principle

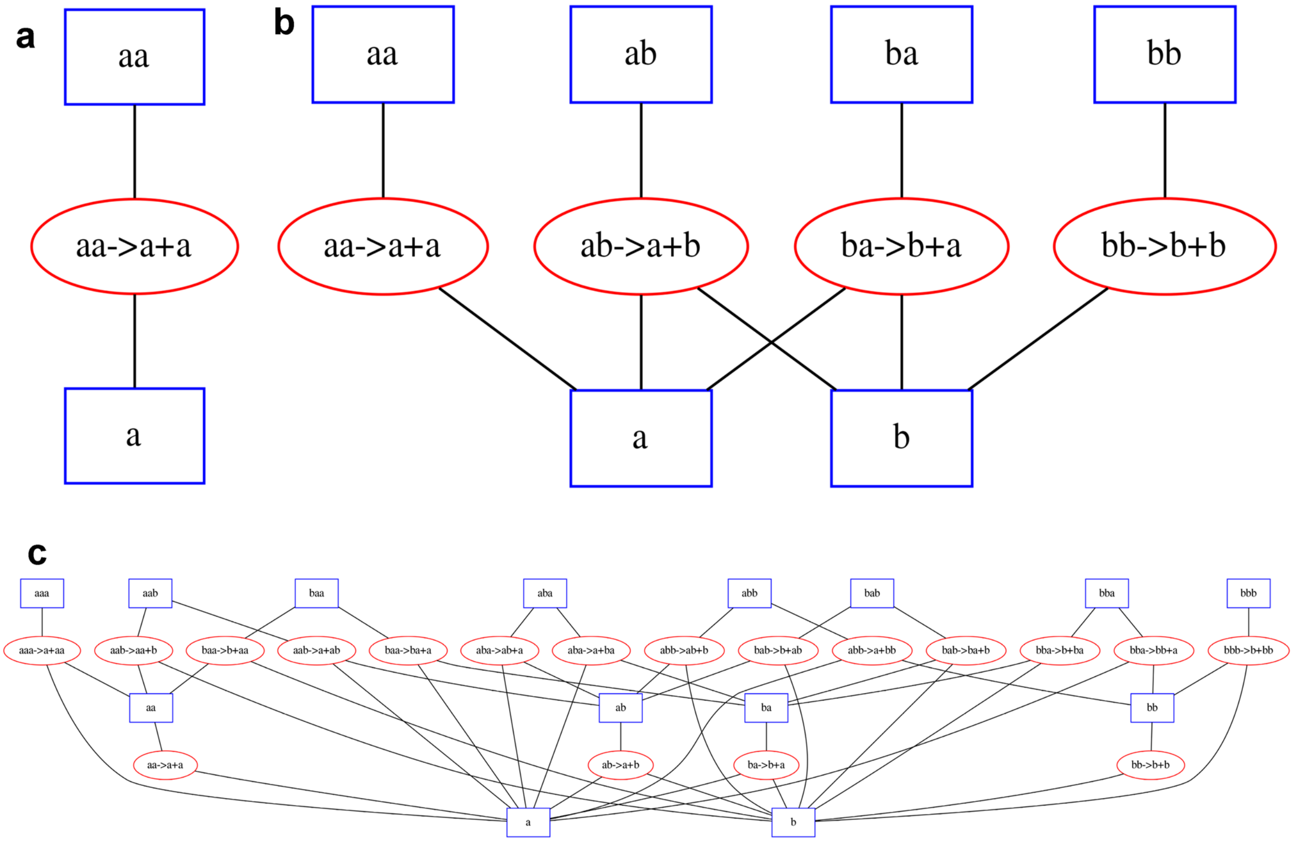

Fig. 1 Three simple string chemistry networks. Square nodes represent chemicals and oval nodes represent reactions. Edges connect chemicals to the reactions they participate in, either as reactants or products. **a** A network with only one type of monomer and a maximum string length of 2. **b** A network with two types of monomers and a maximum string length of 2. **c** A network with two types of monomers and a maximum string length of 3

further constraints on reversibility could easily be added. The numbers of chemicals and allowed reactions in the model are functions of the number of unique characters (“monomers”) and the maximum chemical length. These functions, plotted in Fig. 2, can be obtained analytically by enumerating the sizes of various string chemistry networks and examining the resulting series:

$$\#\text{ metabolites} = \sum_{i=1}^{L} A^i = \begin{cases} \frac{A(A^L-1)}{(A-1)} & A \neq 1 \\ L & A = 1 \end{cases},$$

$$\#\text{ reactions} = \sum_{i=1}^{L} (i-1)A^i = \begin{cases} A\frac{(L-1)A^{L+1}-LA^L+A}{(A-1)^2} & A \neq 1 \\ \frac{L(L-1)}{2} & A = 1 \end{cases}$$

where A is the number of unique characters (monomers) and L is the maximum chemical length. We will refer below to a specific complete set of metabolites and reactions generated for a given choice of A and L as a “chemical universe”. This will allow us to clearly distinguish such complete sets from subsets generated by pruning algorithms (see below).

Flux Balance Analysis

Flux Balance Analysis (FBA) is a mathematical framework for computing steady-state fluxes through chemical reactions in a given network of reactions subject to linear constraints (Orth et al. 2010). The network of reactions is represented as a stoichiometric matrix **S**, where each column represents an individual reaction and each row represents an individual metabolite. Each element in this matrix is the stoichiometric coefficient for the given metabolite in the given reaction: positive if the metabolite is a product of that reaction, negative if it is a substrate, and zero if it does not participate (see Fig. 3 for an example of a string chemistry network and its associated stoichiometric matrix). The reaction fluxes to be computed are represented by a vector **v**. In order for the network to be at steady state, **v** must be in the null space of **S**

$$\boldsymbol{Sv} = 0$$

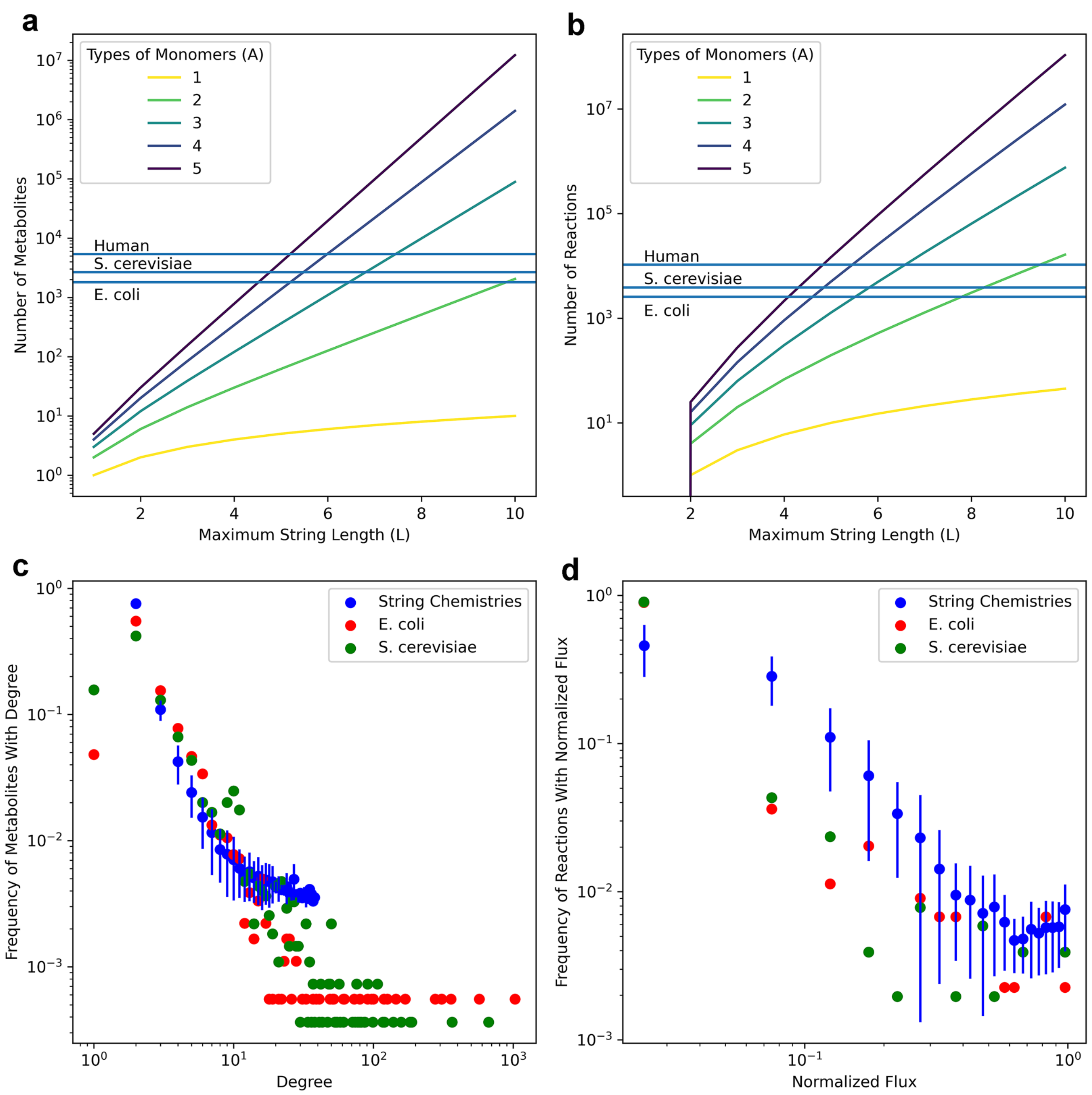

Fig. 2 Comparison of size, degree distributions, and flux distributions of string chemistry networks to the same properties of real metabolic networks. **a** Network sizes of universal string chemistry networks (colored lines) and real metabolic networks (blue lines) measured by metabolite counts. **b** Network sizes measured by reaction counts. **c** Degree distributions of 100 string chemistry networks pruned from the universal network with $A=3$ and $L=7$ compared to degree distributions of real metabolic networks (see "Methods" section for more details). String chemistry degree distributions are shown as mean and standard deviations of frequencies of each degree across the 100 pruned networks. **d** Flux distributions of 100 string chemistry networks (same as in **c**) compared to flux distributions of real metabolic networks when optimized with the default biomass objective functions (see "Methods" section for more details). Fluxes were normalized to the maximum flux within each network and binned into 10 equally sized bins before plotting. Flux distributions of string chemistry networks are shown as mean and standard deviations of frequencies across the 100 pruned networks

The resulting system of equations is underdetermined for nearly all nontrivial networks. Additional constraints may be specified that limit the values of fluxes through specific reactions, typically reflecting nutrient limitations or known thermodynamic constraints on certain reactions. These constraints generally reduce the space of feasible solutions, but still leave the problem underdetermined. Thus, a linear combination of reactions **Z** (the objective

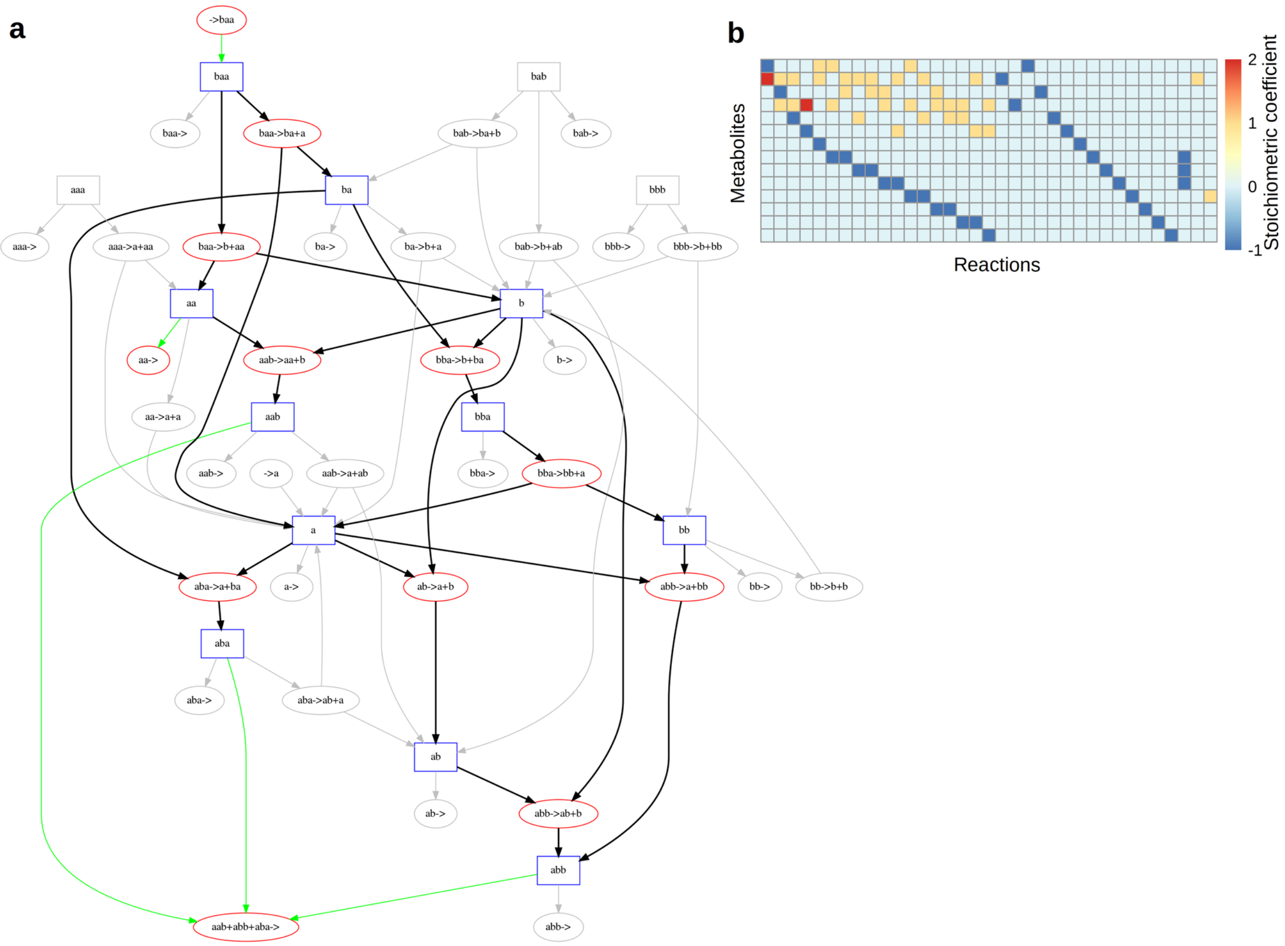

Fig. 3 Flux Balance Analysis on string chemistry networks. **a** String chemistry network with $A=2$ and $L=3$. Metabolites are represented by blue rectangles and reactions are represented by red ovals. Edge colors represent reaction fluxes after maximizing flux through the biomass reaction: green edges are exchange fluxes (import/export/biomass production), black edges represent nonzero fluxes, and gray edges represent fluxes of zero. The direction of non-gray edges corresponds to the direction of flux; directions on gray edges are arbitrary. **b** Stoichiometric matrix of network in (**a**). Each entry in the matrix represents the stoichiometric coefficient of a particular metabolite (row) in a particular reaction (column), and the coefficient is positive if the metabolite is produced by the reaction or negative if the metabolite is consumed by the reaction

function) through which flux should be maximized (or minimized) is also specified:

$$Z = c^{\mathrm{T}} v$$

where **c** is a vector indicating which reactions are to be included in the objective function **Z**. As FBA is usually applied to biochemical reaction networks, the objective function is frequently set to correspond to a single reaction that produces the right proportion of all precursors necessary for the generation of cellular macromolecules and key metabolites, representing growth of cellular biomass. While FBA was originally developed for studying and engineering microbial metabolic networks, its formalism is easily adaptable to any chemistry, provided that its chemical reactions can be represented as columns of a stoichiometric matrix (Fig. 3).

The ARtificial CHemistry NEtwork Toolbox (ARCHNET) Package

We created the ARCHNET Python package to facilitate the creation and handling of string chemistry networks (as defined above) of arbitrary size, as well as the application of FBA to such networks. All FBA computations were performed using the COBRApy Python package (Ebrahim et al. 2013). The ARCHNET package, along with all scripts used to generate data and create figures, is available in a GitHub repository: https://github.com/segrelab/string-chemistry. The package contains tools for generating and analyzing string chemistry networks of arbitrary size, given the set of characters to use as monomers and the maximum string length. A network can be returned as a stoichiometric matrix and/or a COBRApy model (to facilitate applying FBA or any other stoichiometric modeling technique).

Network Pruning Algorithm

We implemented an algorithm that takes a complete string chemistry network as an input (e.g., the network of all possible reactions and metabolites when $A=2$ and $L=5$) and outputs a subnetwork that has been pruned to satisfy specific criteria. Specifically, the algorithm takes as input (i) a string chemistry network, (ii) a set of available environmental nutrients, and (iii) a biomass composition, i.e., a set of molecules that have to be produced at stoichiometrically fixed proportions; in all examples shown here, we use coefficients of 1 each for molecular components of biomass, except in Fig. S6. In principle, however, biomass coefficients could have any empirically assigned value, as done in the stoichiometric models of real organisms, where these numbers represent the amount in millimoles of that molecule per 1 g of biomass. The algorithm iteratively removes reactions from the network until there is no flux through the output reaction (Fig. S1). In particular, it repeatedly runs FBA to assign fluxes to all reactions and removes reactions with no flux and the reaction with the smallest nonzero flux. Once there is no flux through the output reaction, the last reaction that was removed is added back to the network and the network is "pruned". The pruning algorithms are part of the Python package described above. Several other assorted scripts provide examples of applications of this pruning algorithm to string chemistry networks.

Comparing Degree and Flux Distributions of Metabolic Networks

In order to compare the degree distributions of string chemistry networks and real metabolic networks, COBRApy models of string chemistry networks, iJO1366 (*Escherichia coli*, Orth et al. 2010), and Yeast8 (*Saccharomyces cerevisiae* Lu et al. 2019) were used to create graph representations of those metabolic networks using the networkx package (https://github.com/networkx/networkx). The networks were represented as bipartite graphs, where each node represents either a metabolite or a reaction, and two nodes are connected when a metabolite participates in a reaction (either as a product or reactant; edges are undirected). When computing degree distributions, only the degrees of metabolite nodes were considered.

Flux distributions were generated for all string chemistry networks and real metabolic networks by performing FBA with the default biomass objective functions. In order to facilitate comparison of fluxes between different networks, all fluxes within each network were normalized to the largest flux in the network (after taking the absolute value of all fluxes), and fluxes were binned.

Results

A Python Package for Creating and Analyzing Arbitrary String Chemistries

We have created the ARtificial CHemistry NEtwork Toolbox (ARCHNET), a Python package capable of generating string chemistry networks of arbitrary sizes given the number of unique characters (A) and the maximum length of a string (L) (Fig. 1). For simplicity, the only types of reactions allowed in these networks are pairwise string concatenation and splitting (see "Methods" section for more details). Even with this restriction on reaction complexity, the networks increase in size very rapidly as A and/or L increase (Fig. 2a, b and "Methods" section). For example, a basic chemistry with $A=3$ and $L=2$ would have 12 metabolites and 9 reactions. If we increase A by 1, the network would involve 20 metabolites and 16 reactions. If we instead increased L by one, there would be 39 metabolites and 63 reactions. The network sizes therefore depend very differently on these two parameters (see "Methods" section). One of the important features of the package is that it can output networks both as a simple text file containing the stoichiometric matrix, and as a COBRApy model (Ebrahim et al. 2013), which can be exported as an SMBL file (Hucka et al. 2003) used in most tools developed to study real metabolic networks, including standard FBA calculations (Fig. 3).

While the principles constraining the structure of real metabolic networks are much more complicated than those giving rise to our string chemistry networks, string chemistry networks (of equal or lower complexity) can still reproduce some network-level properties of real chemistry networks (Riehl et al. 2010). As shown in Fig. 2a, b, specific sets of parameters in artificial chemistries can lead to networks that contain numbers of reactions and metabolites that are close to those of real metabolic networks. These numbers can be determined either numerically or analytically (see "Methods" section). Interestingly, even string chemistry networks with few unique characters and short maximum lengths (e.g., $A=4$, $L=5$; $A=2$, $L=10$) reach sizes comparable to those of the human, yeast and *E. coli* metabolic networks (Fig. 2a, b; Orth et al. 2010; Lu et al. 2019). However, as seen in Fig. S2, these artificial networks have a much higher connectivity (ratio of reactions to metabolites) than the real organisms' metabolic networks. Conversely, a simple network with $A=1$ and $L=3$ would have a connectivity comparable to that of real metabolic networks, but would be much smaller.

One could then ask whether it is possible to create string chemistry networks that are both of similar sizes and connectivities to those of real metabolic networks. Indeed, one should view the complete string chemistries depicted here as analogous to "complete chemical universes", out of which a single organism's metabolic network would constitute a small subset. As shown below, this concept can be explored in artificial chemistries by devising algorithms that can "prune" complete chemical networks to obtain subnetworks that resemble individual organisms' metabolic networks.

Pruned Networks as Proxies for Evolved Organisms

Having established a method for quantitatively comparing properties of string chemistry networks to real metabolic networks, we explored the properties of string chemistry subnetworks that more closely resemble those of individual organisms. We thus modeled organism-scale metabolic networks as "minimal networks," which use the fewest reactions required to produce a desired set of metabolites (i.e., biomass precursors, in analogy with the building blocks of microbial biomass used to represent self-reproduction in the growth flux associated with genome-scale stoichiometric models (Orth et al. 2010; Lachance et al. 2019)). This minimal network structure is consistent with simple parsimonious evolutionary assumptions used in previous studies (Noor et al. 2010; Riehl et al. 2010; Pál et al. 2006). To identify these minimal networks from our string chemical universes generated using ARCHNET, we implemented a "pruning" algorithm that iteratively applies FBA to string chemistry networks. Briefly, the algorithm works by (1) applying FBA to a string chemistry network (initially set to the whole chemical universe given particular values of *A* and *L*) with some specified nutrient uptake reactions and a "biomass" reaction (representing the metabolic objective of the network, e.g., biomass production for many real metabolic networks), (2) removing all reactions that have no flux, (3) testing whether or not the reaction with the smallest nonzero flux can be removed without eliminating flux through the biomass reaction, and (4) repeating until no reactions can be removed (see "Methods" section and Fig. S1).

One may wonder whether the structures of these pruned networks resembles those of real organisms' metabolic networks. We addressed this question by computing degree and flux distributions for 100 pruned networks ($A=3$, $L=7$) and comparing them to the degree and flux distributions of the *E. coli* and *S. cerevisiae* metabolic networks (see "Methods" section). As shown in Fig. 2c, the pruned string chemistry networks tend to have scale-free degree distributions, just as the real metabolic networks do (Almaas et al. 2004). As shown in Cohen et al. (2004), scale-free networks are also necessarily small-world networks, and real metabolic networks are well known to be small-world networks (Wagner and Fell 2001). The distributions of fluxes through the pruned string chemistry networks strongly resemble the distributions of fluxes through the *E. coli* and *S. cerevisiae* networks when optimized for biomass production (Fig. 2d).

Given these structural similarities, we asked whether pruned networks could also reflect some of the functional properties of real networks. We looked specifically at metabolic secretions, as the ability to excrete waste products is a crucial component of cellular metabolism (Ferguson et al. 1998; Richards et al. 2013; Hart et al. 2019). Since the degree to which individual bacteria secrete metabolic intermediates and/or waste products varies dramatically from organism to organism, we examined two extreme cases with our pruning algorithm: one in which all metabolites in the network are allowed to be secreted, and one in which no metabolites are allowed to be secreted (i.e., the only sink in the network is the biomass reaction). We found that pruning while allowing secretion of waste products generally resulted in slightly smaller networks than pruning without allowing secretion (Fig. S2). This is likely because ensuring that all metabolic byproducts are internally recycled into biomass components requires more reactions than simply secreting them as waste products. We also note that the notion that secreting waste products leads to simpler metabolic networks may well be relevant in real microbial communities. For example, many microbial communities are sustained by costless secretions (Pacheco et al. 2019), which could lead to a reduction of metabolic capabilities in specific taxa (Morris et al. 2012).

Biomass Precursors Shape Network Composition More than Environmental Nutrient Composition

Using our pruning algorithm, we investigated the relative importance of the choice of nutrients and the choice of biomass precursors on the composition (i.e., identities of remaining reactions) of pruned networks. We generated a string chemistry universe with $A=2$ and $L=5$, then created different biomass compositions (100 different sets of 5 randomly chosen biomass precursors) and different sets of nutrients (100 random pairs of nutrients) using the metabolites contained within this chemical universe. We note that upon choosing the biomass composition, a biomass reaction was added to produce all chosen biomass precursors in equal proportions (see "Methods" section). We then ran the pruning algorithm on all possible combinations of these nutrients and biomass precursors (Fig. 4a). In order to compare the compositions of the pruned networks, each network was represented as a binary vector with as many elements as there were reactions in the chemical universe. In this binary vector, a 1 represents a reaction that was kept in the pruned network and a 0 represents a reaction that was removed during pruning. These binary vectors were visualized on

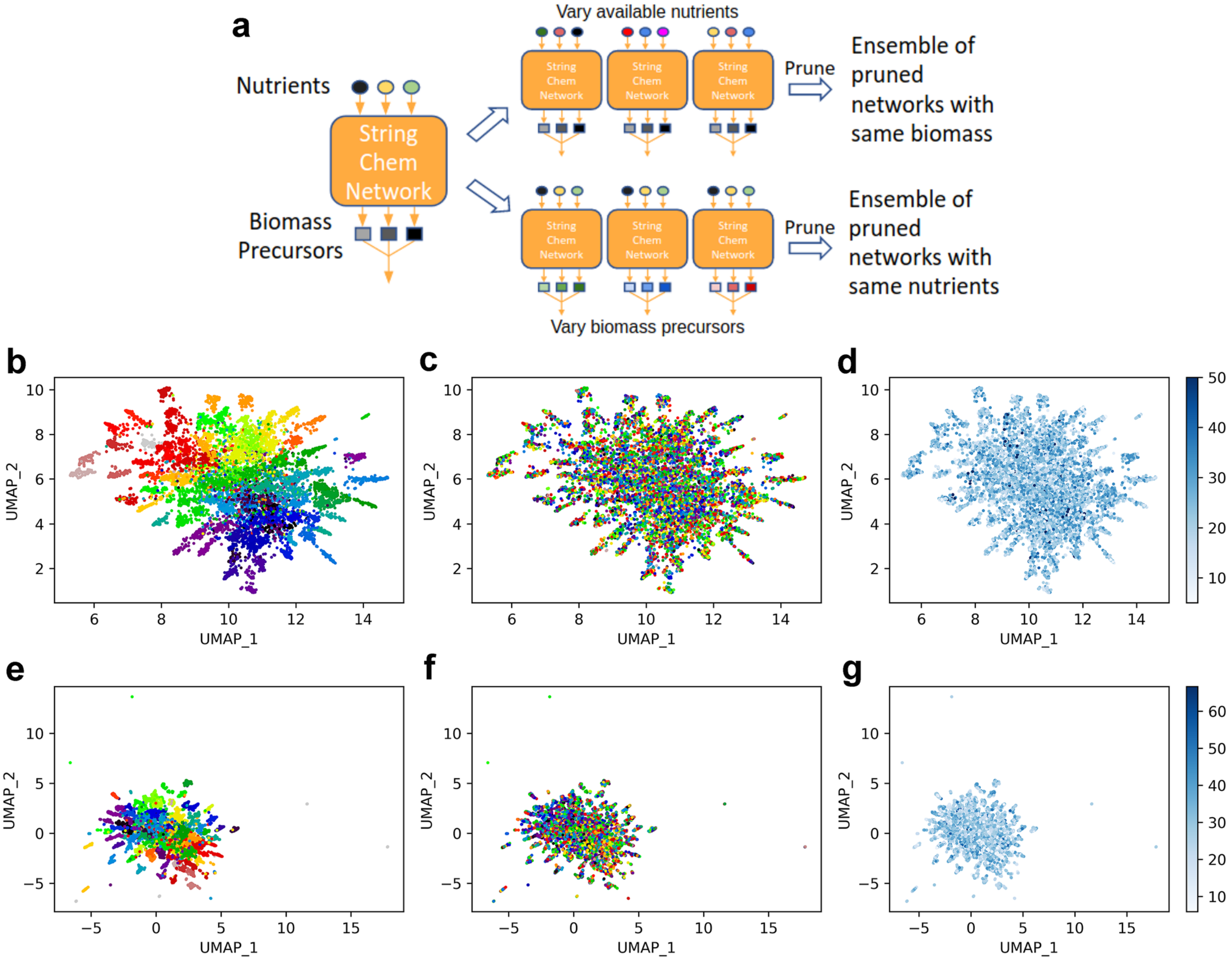

Fig. 4 Choice of biomass precursors impacts structure of pruned networks more than choice of available nutrients. **a** Cartoon representation of how data shown in panels (**b**–**g**) was generated. **b** UMAP scatterplot of pruned networks with export reactions (see main text) generated as described in (**a**). Each point represents a different pruned network and the color of each point indicates the biomass reaction of that network. **c** Same as (**b**) but colors indicate which set of nutrients the network was pruned with. **d** Same as (**b**) but colors indicate optimal biomass flux. **e**–**g** Same as (**b**–**d**) but networks were pruned without export reactions (see main text). All pruned networks were derived from the universal string chemistry network with $A=2$ and $L=5$

UMAP plots (McInnes et al. 2018) (Fig. 4b–g). The main outcome of this analysis is that, regardless of whether or not export reactions are allowed, networks with the same biomass typically cluster together (Fig. 4b), while networks with the same nutrients frequently have very different compositions (Fig. 4c). The clustering is generally weaker in the networks pruned without export reactions—there are more isolated networks and distinct small clusters—but the pruned networks still noticeably cluster by biomass reaction (Fig. 4e, f). We also note that the clustering of networks does not seem to display any clear pattern in terms of achievable growth rates (Fig. 4d, g), which are highly variable and roughly distributed around an intermediate value. In other words, networks with similar composition, as dictated by the biomass composition, may achieve substantially different growth rates, suggesting that while biomass composition dictates network composition, growth rates are not as straightforwardly determined by either biomass composition or environmental composition.

To assess the possibility that these results were an artifact of the arbitrarily chosen number of nutrients and biomass precursors, we investigated how the proportion of pruned reactions and connectivity of pruned networks change as the numbers of nutrients and biomass precursors vary (Figs. S4 and S5, respectively). While the proportion of pruned reactions clearly decreases as the number of biomass precursors increases, as one might expect, it does not appear to be affected by the number of available nutrients. Figure S5 indicates that the metabolite-to-reaction ratio is always around 1 in pruned networks. These values are all slightly lower than those observed in real metabolic networks (Fig. S2), which likely reflects the fact that real metabolic networks must be capable of sustaining growth in multiple different environments, while the pruned string chemistry networks are only required to sustain growth in one particular environment. While we expect that the number of biomass precursors and nutrients may affect the composition of pruned networks in other more subtle ways, these findings support the idea that the results shown in Fig. 4 do not depend on the number of biomass precursors or nutrients used during the pruning process.

Discussion

We have created ARCHNET, a Python package capable of performing stoichiometric modeling on string chemistry networks of arbitrary size and monomer diversity. We have also devised a pruning algorithm for these networks, which identifies minimal metabolic networks necessary for converting a given set of environmental nutrients into a specific combination of biomass precursors. By applying this pruning algorithm to many thousands of string chemistry networks, we found that the choice of the biomass metabolites wields much more influence over the composition of the minimal network than the choice of environmental nutrients. Beyond this finding, our package could be used to further quantitatively explore any aspect of the complex relationship between metabolic network structure, environmental diversity, and biomass composition.

It is important to keep in mind that while in the current work we verified the robustness of some of our results relative to a number of free parameters, there are additional choices of the underlying string chemistry and pruning algorithm that could influence the results. In particular, as described above, and shown in Figs. S4 and S5, we verified that pruned network properties depend only mildly on the choice of the number of biomass components and nutrients, suggesting that downstream analyses would similarly be robust relative to these two parameters. When generating the pruned networks shown in Fig. 4, all stoichiometric coefficients used in all involved biomass reactions were 1, so in Fig. S6 we verified that varying the values of these stoichiometric coefficients does not significantly alter the main conclusions drawn from Fig. 4. To generate the pruned networks in Fig. S6, the ensembles of networks with identical biomass reactions and varied nutrients (see Fig. 4a for a schematic) underwent one additional modification before pruning: each stoichiometric coefficient in each network's biomass reaction was changed from 1 to a random integer between 1 and 10 (inclusive). As shown in Figure S6a, pruned networks with the same biomass precursors (but not necessarily the same stoichiometric coefficients) cluster together to a comparable extent as the pruned networks in Fig. 4b. The pruned networks do not seem to cluster by nutrient sources in either Figs 4c, f or S6b. No particular sets of biomass precursors seem to result in consistently higher or lower maximum growth fluxes than any other sets (Figs. 4d, g and S6c). Further analyses may uncover more subtle ways in which the particular values used in the stoichiometric coefficients of biomass reactions relate to the structures of pruned networks.

There are also a number of ways in which the string chemistry model presented here could be made more complex and/or realistic. One could try to explicitly model more features of real chemical reactions (e.g., thermodynamics, kinetics), along the lines of the methods mentioned in the introduction. The role of reaction irreversibility represents a key area of future exploration: while all internal reactions are currently assumed to be perfectly reversible, variable degrees of reaction reversibility (e.g., induced by free energy assignments) could be added to string chemistries to recapitulate features observed in real metabolism. One could also introduce regulatory interactions between metabolites and reactions by connecting the fluxes through particular metabolites to the constraints on fluxes through other reactions. Notably, while each of these possible modifications would not necessarily dramatically change the string chemistry framework, the combinatorial interactions of all the various new parameters would dramatically increase the space of possible networks to explore and characterize, providing ample opportunities for future work with artificial chemistries.

We believe that our finding about the relative importance of biomass composition and environmental metabolites has important consequences for the process of reconstructing real metabolic networks. A key step in the process of creating Genome-Scale Metabolic Models (GEMs) is the process of gap filling, where missing reactions ("gaps", usually due to incomplete experimental data) in draft GEMs are imputed using a variety of methods (Thiele et al. 2014; Satish Kumar et al. 2007; Prigent et al. 2017; Christian et al. 2009; Vitkin and Shlomi 2012). Gap-filling algorithms generally function by identifying a minimal set of reactions to add to the draft GEM in order for it to be capable of producing biomass, so they require users to specify a particular biomass reaction. Frequently, the so-called "template" biomass reactions are used for all bacteria in a particular taxa (e.g., many GEMs of Gram-negative bacteria are created using the *E. coli* biomass reaction) due to the significant difficulty of obtaining the extensive experimental data required to determine a particular organism's biomass composition (Henry et al. 2010; Xavier et al. 2017). Our finding about the role biomass composition plays in determining the composition of pruned networks, along with a previous study that found that bacterial GEMs clustered more strongly by which biomass reaction was used as a template than by taxonomy (Xavier et al. 2017) both suggest that careful consideration should be put into the choice of a biomass reaction when using gap-filling algorithms.

The pruning algorithms presented in this work bear some resemblance to certain gap-filling algorithms, algorithms for identifying Elementary Flux Modes (EFMs) (Schuster et al. 2000), and algorithms for identifying Minimal Balance Pathways (MBPs) (Riehl et al. 2010). Some gap-filling algorithms iteratively add and remove reactions from a large pool of possible reactions, eventually converging to an optimally gap-filled network (Vitkin and Shlomi 2012; Reed et al. 2006; Pharkya et al. 2004). Some algorithms for deriving tissue-specific metabolic networks from generic human metabolic networks also function similarly (Machado et al. 2018; Jerby et al. 2010). EFMs represent every independent route from a source to a sink through a metabolic network and are often considered a basis set for the space of possible fluxes through a network; the pruned networks we present here may converge to EFMs of string chemistry networks for certain input/output metabolite cases, but further research may identify additional connections between the two concepts. Since MBPs represent the optimal set of reactions for converting a single input metabolite into a single output metabolite, our pruned networks could be viewed as an extension of MBPs with multiple inputs and multiple outputs. All of these algorithms, including our pruning algorithms, aim to identify "optimal" networks under certain criteria of optimality; further exploration of the similarities and differences between these approaches may lead to a better understanding of what one should consider an optimal metabolic network to be like.

The pruning algorithm presented in this paper is far from the only algorithm that accomplishes the goal of narrowing down a metabolic network to its essential components, given an environment and a biomass composition. In an attempt to verify that the precise details of how our pruning algorithm was formulated do not substantially impact the main results of this work, we created an alternate pruning algorithm: instead of removing the reaction with the smallest flux, the new pruning algorithm computes the change in biomass flux that would result from every possible single reaction deletion and removes the reaction with the smallest impact on biomass flux at each pruning step. The two pruning algorithms generally remove the same reactions at each step of the algorithm (Fig. S7) and usually wind up producing very similar output networks when given the same input (Fig. S8). Furthermore, when reproducing the analysis shown in Fig. 4a with the biomass-focused pruning algorithm, the results are entirely comparable to those obtained using the original pruning algorithm (Fig. S9). Based on these analyses, we believe that the specific criterion used to decide which reactions should be pruned do not meaningfully change the main conclusions of this paper.

One can view both of these pruning algorithms as analogs of reductive evolutionary processes: the flux-based pruning algorithm selects against reactions that carry little flux, while the biomass-based pruning algorithm selects against reactions that contribute little to biomass production. One could imagine that a microbe growing in a nutrient-poor environment might stop devoting resources to expressing metabolic enzymes that catalyze reactions that carry little flux and are not essential for biomass production. Similarly, one could imagine a competitive environment in which an organism capable of achieving similar growth rates to its neighbors while devoting fewer resources to producing metabolic enzymes that do not substantially contribute to biomass production would outcompete its neighbors. The pruned networks are also reminiscent of the metabolic networks of certain marine plankton species that only express half of the enzymes in the citric acid cycle, since other microbes in their environment secrete the appropriate intermediates and there is strong selective pressure to reduce genome size due to low availability of nitrogen (Braakman et al. 2017). While the outputs of both pruning algorithms are generally similar, there is a small subset of initial networks that are pruned rather differently by the two algorithms (Fig. S8); studying these cases in more detail may yield new insights into general features of these different types of evolutionary processes.

Several previous studies used artificial chemistry as an avenue for addressing questions related to the origin of life or to general mathematical properties of biochemical networks (Banzhaf and Yamamoto 2015; Kauffman 1993; Benkö et al. 2003; Walter Fontana and Buss 1994a, b; Peng et al. 2020). Conversely, FBA has been applied mostly to the study of metabolic networks of real organisms (Gu et al. 2019; Kauffman et al. 2003; Orth et al. 2010). There is likely great untapped potential available from combining the two approaches. In particular, the recent application of stoichiometric approaches to the study of early metabolism (Goldford and Segrè 2018) and of ecosystem-level biochemical networks (Carlson et al. 2018; Klitgord and Segrè 2010; Harcombe et al. 2014) could greatly benefit from additional creative usage of artificial chemistries. For example, the capacity to handle artificial string chemistries of arbitrary complexity using these same stoichiometric tools makes it possible to explore evolutionary processes and ecosystem-level metabolism under simulated scenarios in which the whole chemical universe is fully known. One could create an assortment of string chemistry networks using ARCHNET and model their interactions using tools such as COMETS (Dukovski et al. 2020). This will make it possible to shed light on the role of historical contingency and optimality principles in shaping the structure of metabolic networks.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00239-021-10018-0.

Author Contributions DS conceived the project. DM, ARP, and DBB designed the computational framework. DM wrote all the code, performed the analyses, and drafted the manuscript. DS and DM finalized the analysis of the data. All authors edited the final version of the manuscript.

Funding This work was partially supported by the Directorates for Biological Sciences (BIO) and Geosciences (GEO) at the NSF and NASA under Agreements No. 80NSSC17K0295, 80NSSC17K0296, and 1724150 issued through the Astrobiology Program of the Science Mission Directorate. ARP was supported by a Howard Hughes Medical Institute Gilliam Fellowship and a National Academies of Sciences, Engineering, and Medicine Ford Foundation Predoctoral Fellowship. DM was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number T32GM100842.

Data Availability All codes used to generate data and figures in this work are available at the following GitHub repository: https://github.com/segrelab/string-chemistry

Declarations

Conflict of interest The authors declare no conflicts of interest.

References

Almaas E, Kovács B, Vicsek T, Oltvai ZN, Barabási AL (2004) Global organization of metabolic fluxes in the bacterium Escherichia coli. Nature 427(6977):839–843. https://doi.org/10.1038/nature02289

Arús-Pous J, Johansson SV, Prykhodko O, Bjerrum EJ, Tyrchan C, Reymond J-L, Chen H, Engkvist O (2019) Randomized SMILES strings improve the quality of molecular generative models. J Cheminf 11(1):71

Banzhaf W, Yamamoto L (2015) Artificial chemistries. MIT Press, Cambridge

Barve A, Wagner A (2013) A latent capacity for evolutionary innovation through exaptation in metabolic systems. Nature 500(7461):203–206

Benkö G, Flamm C, Stadler PF (2003) A graph-based toy model of chemistry. J Chem Inf Comput Sci 43(4):1085–1093

Borenstein E, Kupiec M, Feldman MW, Ruppin E (2008) Large-scale reconstruction and phylogenetic analysis of metabolic environments. Proc Natl Acad Sci USA 105(38):14482–14487

Braakman R, Follows MJ, Chisholm SW (2017) Metabolic evolution and the self-organization of ecosystems. Proc Natl Acad Sci USA 114(15):E3091–E3100. https://doi.org/10.1073/pnas.1619573114

Carlson RP, Beck AE, Phalak P, Fields MW, Gedeon T, Hanley L, Harcombe WR, Henson MA, Heys JJ (2018) Competitive resource allocation to metabolic pathways contributes to overflow metabolisms and emergent properties in cross-feeding microbial consortia. Biochem Soc Trans 46(2):269–284

Chang R (2005) Physical chemistry for the biosciences. University Science Books

Christian N, May P, Kempa S, Handorf T, Ebenhöh O (2009) An integrative approach towards completing genome-scale metabolic networks. Mol BioSyst 5(12):1889–1903

Cohen R, Havlin S, Ben-Avraham D (2004) Structural properties of scale-free networks. In: Bornholdt S, Schuster HG (eds) Handbook of graphs and networks. Wiley, Weinheim, pp 85–110

Compton RG, Bamford CH, Tipper CFH (2012) The theory of kinetics. Elsevier

Dukovski I, Bajić D, Chacón JM, Quintin M, Vila JCC, Sulheim S, Pacheco AR, et al. 2020. "Computation Of Microbial Ecosystems in Time and Space (COMETS): An Open Source Collaborative Platform for Modeling Ecosystems Metabolism." arXiv [q-bio.QM]. arXiv:2009.01734.

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

Ebrahim A, Lerman JA, Palsson BO, Hyduke DR (2013) COBRApy: constraints-based reconstruction and analysis for python. BMC Syst Biol 7:74

Ferguson GP, Tötemeyer S, MacLean MJ, Booth IR (1998) Methylglyoxal production in bacteria: suicide or survival? Arch Microbiol 170(4):209–218

Fontana W, Buss LW (1994a) 'The arrival of the fittest': toward a theory of biological organization. Bull Math Biol 56(1):1–64

Fontana W, Buss LW (1994b) What would be conserved if 'the tape were played twice'? Proc Natl Acad Sci USA 91(2):757–761

Friedlander T, Mayo AE, Tlusty T, Alon U (2015) Evolution of bow-tie architectures in biology. PLoS Comput Biol 11(3):e1004055

Goldford JE, Segrè D (2018) Modern views of ancient metabolic networks. Curr Opin Syst Biol 8:117

Gottstein W, Olivier BG, Bruggeman FJ, Teusink B (2016) Constraint-based stoichiometric modelling from single organisms to microbial communities. J R Soc Interface 13:20160627

Gu C, Kim GB, Kim WJ, Kim HU, Lee SY (2019) Current status and applications of genome-scale metabolic models. Genome Biol 20(1):121

Guseva E, Zuckermann RN, Dill KA (2017) Foldamer hypothesis for the growth and sequence differentiation of prebiotic polymers. Proc Natl Acad Sci USA 114(36):E7460–E7468

Handorf T, Ebenhöh O, Heinrich R (2005) Expanding metabolic networks: scopes of compounds, robustness, and evolution. J Mol Evol 61(4):498–512

Harcombe WR, Riehl WJ, Dukovski I, Granger BR, Betts A, Lang AH, Bonilla G et al (2014) Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. Cell Rep 7(4):1104–1115

Hart SFM, Mi H, Green R, Xie L, Pineda JMB, Momeni B, Shou W (2019) Uncovering and resolving challenges of quantitative modeling in a simplified community of interacting cells. PLoS Biol 17(2):e3000135

Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, Haraldsdóttir HS et al (2019) Creation and analysis of biochemical constraint-based models using the COBRA Toolbox vol 3.0. Nat Protoc 14(3):639–702

Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol 28(9):977–982

Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP et al (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19(4):524–531

Jerby L, Shlomi T, Ruppin E (2010) Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. Mol Syst Biol 6:401

Kauffman SA (1993) The origins of order: self-organization and selection in evolution. Oxford University Press, Oxford

Kauffman KJ, Prakash P, Edwards JS (2003) Advances in flux balance analysis. Curr Opin Biotechnol 14(5):491–496

Klitgord N, Segrè D (2010) Environments that induce synthetic microbial ecosystems. PLoS Comput Biol 6(11):e1001002

Lachance JC, Lloyd CJ, Monk JM, Yang L, Sastry AV, Seif Y, Palsson BO, Rodrigue S, Feist AM, King ZA, Jacques PÉ, Schneidman-Duhovny D (2019) BOFdat: Generating biomass objective functions for genome-scale metabolic models from experimental data. PLOS Comput Biol 15(4):e1006971. https://doi.org/10.1371/journal.pcbi.1006971

Laidler KJ, Glasstone S (1948) Rate, order and molecularity in chemical kinetics. J Chem Educ 25(7):383

Lee AA, Yang Q, Sresht V, Bolgar P, Hou X, Klug-McLeod JL, Butler CR (2019) Molecular transformer unifies reaction prediction and retrosynthesis across pharma chemical space. Chem Commun 55(81):12152–12155

Lin T-S, Coley CW, Mochigase H, Beech HK, Wang W, Wang Z, Woods E et al (2019) BigSMILES: a structurally-based line notation for describing macromolecules. ACS Cent Sci 5(9):1523–1531

Lu H, Li F, Sánchez BJ, Zhu Z, Li G, Domenzain I, Marcišauskas S, Anton PM, Lappa D, Lieven C, Beber ME, Sonnenschein N, Kerkhoven EJ, Nielsen J (2019) A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. Nature Commun 10(1). https://doi.org/10.1038/s41467-019-11581-3

Machado D, Andrejev S, Tramontano M, Patil KR (2018) Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Res 46(15):7542–7553

McInnes L, Healy J, Melville J (2018) UMAP: uniform manifold approximation and projection for dimension reduction. arXiv [stat.ML]. arXiv:1802.03426

Morris JJ, Lenski RE, Zinser ER (2012) The black queen hypothesis: evolution of dependencies through adaptive gene loss. mBio 3(2). https://doi.org/10.1128/mBio.00036-12

Noor E, Eden E, Milo R, Alon U (2010) Central carbon metabolism as a minimal biochemical walk between precursors for biomass and energy. Mol Cell 39(5):809–820

O'Brien EJ, Monk JM, Palsson BO (2015) Using genome-scale models to predict biological capabilities. Cell 161(5):971–987

Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? Nat Biotechnol 28(3):245–248

Pacheco AR, Moel M, Segrè D (2019) Costless metabolic secretions as drivers of interspecies interactions in microbial ecosystems. Nat Commun 10(1):103

Pál C, Papp B, Lercher MJ, Csermely P, Oliver SG, Hurst LD (2006) Chance and necessity in the evolution of minimal metabolic networks. Nature 440(7084):667–670

Peng Z, Plum AM, Gagrani P, Baum DA (2020) An ecological framework for the analysis of prebiotic chemical reaction networks. J Theor Biol

Pfeiffer T, Soyer OS, Bonhoeffer S (2005) The evolution of connectivity in metabolic networks. PLoS Biol 3(7):e228

Pharkya P, Burgard AP, Maranas CD (2004) OptStrain: a computational framework for redesign of microbial production systems. Genome Res 14(11):2367–2376

Prigent S, Frioux C, Dittami SM, Thiele S, Larhlimi A, Collet G, Gutknecht F et al (2017) Meneco, a topology-based gap-filling tool applicable to degraded genome-wide metabolic networks. PLoS Comput Biol 13(1):e1005276

Raymond J, Segrè D (2006) The effect of oxygen on biochemical networks and the evolution of complex life. Science 311(5768):1764–1767

Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, Bui OT, Knight EM, Fong SS, Palsson BO (2006) Systems approach to refining genome annotation. Proc Natl Acad Sci USA 103(46):17480–17484

Richards GR, Patel MV, Lloyd CR, Vanderpool CK (2013) Depletion of glycolytic intermediates plays a key role in glucose-phosphate stress in *Escherichia coli*. J Bacteriol 195(21):4816–4825

Riehl WJ, Krapivsky PL, Redner S, Segrè D (2010) Signatures of arithmetic simplicity in metabolic network architecture. PLoS Comput Biol 6(4):e1000725

Satish Kumar V, Dasika MS, Maranas CD (2007) Optimization based automated curation of metabolic reconstructions. BMC Bioinf 8:212

Schuster S, Fell DA, Dandekar T (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nat Biotechnol 18(3):326–332

Soyer OS, Pfeiffer T (2010) Evolution under fluctuating environments explains observed robustness in metabolic networks. PLoS Computat Biol. 6(8):e1000907

Thiele I, Vlassis N, Fleming RMT (2014) fastGapFill: efficient gap filling in metabolic networks. Bioinformatics 30(17):2529–2531

Vitkin E, Shlomi T (2012) MIRAGE: a functional genomics-based approach for metabolic network model reconstruction and its application to cyanobacteria networks. Genome Biol 13(11):R111

Wagner A, Fell DA (2001) The small world inside large metabolic networks. Proc Biol Sci 268(1478):1803–1810

Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28(1):31–36

Xavier JC, Patil KR, Rocha I (2017) Integration of biomass formulations of genome-scale metabolic models with experimental data reveals universally essential cofactors in prokaryotes. Metab Eng 39:200–208

Yizhak K, Chaneton B, Gottlieb E, Ruppin E (2015) Modeling cancer metabolism on a genome scale. Mol Syst Biol 11(6):817