A latent capacity for evolutionary innovation through exaptation in metabolic systems

Barve, Aditya; Wagner, Andreas

Abstract: Some evolutionary innovations may originate non-adaptively as exaptations, or pre-adaptations, which are by-products of other adaptive traits. Examples include feathers, which originated before they were used in flight, and lens crystallins, which are light-refracting proteins that originated as enzymes. The question of how often adaptive traits have non-adaptive origins has profound implications for evolutionary biology, but is difficult to address systematically. Here we consider this issue in metabolism, one of the most ancient biological systems that is central to all life. We analyse a metabolic trait of great adaptive importance: the ability of a metabolic reaction network to synthesize all biomass from a single source of carbon and energy. We use novel computational methods to sample randomly many metabolic networks that can sustain life on any given carbon source but contain an otherwise random set of known biochemical reactions. We show that when we require such networks to be viable on one particular carbon source, they are typically also viable on multiple other carbon sources that were not targets of selection. For example, viability on glucose may entail viability on up to 44 other sole carbon sources. Any one adaptation in these metabolic systems typically entails multiple potential exaptations. Metabolic systems thus contain a latent potential for evolutionary innovations with non-adaptive origins. Our observations suggest that many more metabolic traits may have non-adaptive origins than is appreciated at present. They also challenge our ability to distinguish adaptive from non-adaptive traits.

DOI: 10.1038/nature12301

A latent capacity for evolutionary innovation through exaptation in metabolic systems

Aditya Barve[1,2] and Andreas Wagner [1,2,3]

[1] Institute of Evolutionary Biology and Environmental Sciences, Bldg. Y27, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

[2] The Swiss Institute of Bioinformatics, Bioinformatics, Quartier Sorge, Batiment Genopode, 1015 Lausanne, Switzerland.

[3] The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Some evolutionary innovations may originate non-adaptively as pre-adaptations or exaptations, which are by-products of other adaptive traits[1-5]. Examples include feathers, which originated before they adopted a role in flight[2], and lens crystallins, light-refracting proteins that originated as enzymes[6]. The incidence of non-adaptive trait origins has profound implications for evolutionary biology, but it has thus far not been possible to study this incidence systematically. We here study it in metabolism, one of the most ancient biological systems that is central to all life. We analyse metabolic traits of great adaptive importance, the ability of a metabolic reaction network to synthesize all biomass from a single (sole) source of carbon and energy. We take advantage of novel computational methods to randomly sample many metabolic networks that can sustain life on any given carbon source, but that contain an otherwise random set of known biochemical reactions. We show that such random networks, required to be viable on one carbon source C, are typically also viable on multiple other carbon sources C_{new} that were not targets of selection. For example, viability on glucose may entail viability on up to 44 other sole carbon sources. Any one adaptation in these metabolic systems typically entails multiple potential exaptations. Metabolic systems thus contain a latent potential for evolutionary innovations with non-adaptive origins. Our observations suggest that many more metabolic traits than currently appreciated may have non-adaptive origins. They also challenge our ability to distinguish adaptive from non-adaptive traits.

How evolutionary adaptations and innovations originate is one of the most profound questions in evolutionary biology. Previous work[1,2] emphasizes the importance of exaptations, also sometimes called pre-adaptations, for this origin. These are traits whose benefits to an organism are unrelated to the reasons for their origin, features that originally serve one (or no) function, and become later co-opted for a different purpose[1–5]. Although examples of exaptations occur from the macroscopic to the molecular scale[1-6] and abound also in human evolution[7], no number of examples could answer how important exaptations are in the origin of adaptations in general. This limitation of case studies can be overcome in those biological systems where one can systematically study many genotypes and the phenotypes they form[8–12].

One of these systems is metabolism. The metabolic genotype of an organism encodes a metabolic reaction network with hundreds of enzyme-catalysed chemical reactions. One of metabolism's fundamental tasks is to synthesize small biomass precursor molecules from environmental molecules, such as different organic carbon sources. An organism or metabolic network is *viable* on a carbon source, if it is able to synthesize all biomass molecules from this source. Viability on a new carbon source can be an important adaptation, and anecdotal evidence shows that this ability can originate as a pre-adaptation[13,14]. For example, laboratory evolution of *Pseudomonas putida* for increased biomass yield on xylose as a carbon source produces strains that utilize arabinose as efficiently as xylose, even though the ancestral strains did not utilize arabinose[14]. Thus, viability on arabinose can be a by-product of increased viability on xylose. We here analyse systematically whether such exaptations are typical or unusual in metabolic systems.

Our analysis uses the ability to predict a metabolic phenotype from a metabolic genotype with the constraint-based method of flux balance analysis (FBA, see methods), to study not just one metabolic network, but to systematically explore a vast space of possible metabolic networks. The members of this space can be described as follows. The currently known "universe" of biochemical reactions comprises more than 5,000 chemical reactions with well-defined substrates and products. In the metabolic network of any one organism, however, only a fraction of these reactions take place, enabling us to describe this network through a binary presence/absence pattern of enzyme-catalysed reactions in the known reaction universe. Recent methods based on Markov Chain Monte Carlo (MCMC) sampling (see methods) allow a systematic exploration of this space, i.e., they permit the creation of arbitrarily large and uniform samples of networks with a given phenotype[12]. This sampling is based on long random walks through metabolic network space, where each step in a walk adds or eliminates a metabolic reaction from a metabolic network, with the only constraint that the network remains viable on a focal carbon source. The starting point of the MCMC random walk is the *Escherichia coli* metabolic network, which we know a priori to be viable on different carbon sources[15]. We here use this approach to create random samples of metabolic networks that are viable on a given set of carbon sources. We refer to such networks as *random viable networks*.

Our analysis focuses on 50 biologically relevant and common carbon sources (supplementary table 1)[15]. For each carbon source *C*, we create a sample of 500 random viable networks that are viable on *C*, if *C* is provided as the sole carbon source. We then use FBA to determine the viability of these networks on each of the 49 other carbon sources. This approach allows us to ask whether viability on carbon source *C* usually entails viability on other carbon sources. The answers to this and related questions show that potential exaptations are ubiquitous in metabolism.

We began our analysis with a sample of 500 random networks that were viable on glucose as the sole carbon source (see methods). Each network can synthesize the 63 essential biomass precursors of *E.coli* – many of them important for most organisms[15,16] – in an aerobic minimal environment containing glucose as the only carbon source. Importantly, we did not require that these 500 networks are viable on any carbon source except glucose.

We first examined whether these networks were viable on each of the 49 other carbon sources. The information resulting from this analysis can be represented, for each network, as a binary 'innovation vector' whose *i*-th entry equals one if the network is viable on carbon source C_i, and otherwise zero (figure 1a). We define the *innovation index* $I_{Glucose}$ of a network as the number of additional carbon sources that each network is viable on. The distribution of this index is shown in figure 1b. Fully 96 percent of networks are viable on other carbon sources in addition to glucose ($I > 0$). The mean innovation index is $I = 4.86$ (standard deviation (s.dev.) = 2.83 carbon sources). This means that networks viable on glucose typically are also viable on almost 5 additional carbon sources. 18.8 percent of networks (94 networks) are viable on exactly 5 new carbon sources, and 37.4 percent (187) of networks are viable on 6 or more carbon sources. Viability on each such carbon source is a potential exaptation. It is a mere by-product of viability on glucose, and could become an adaptation whenever this carbon source is the sole carbon source. We also found that different random viable networks differ in the additional carbon sources to which they are pre-adapted (supplementary figures 1 and 2). Most of the 50 carbon sources we study confer viability on at least one network in our sample (supplementary results). Moreover, a variation in our sampling procedures that allows only reactions already connected to a metabolism to be altered further increases the incidence of exaptation (methods, supplementary figure 3). Finally, complex metabolic networks that have more reactions have greater potential for exaptation (supplementary figure 4).

We next asked whether the ability to grow on multiple additional carbon sources is a peculiarity of networks viable on glucose. To this end, we sampled, for each of our remaining 49 carbon sources, 500 random metabolic networks viable on this carbon source (for a total of 49 x 500 = 24500 sampled networks). We then computed the distribution of the innovation index I_C for each carbon source C. Figure 2a shows the mean of this distribution (bars) and its coefficient of variation (vertical lines), that is, the ratio of the standard deviation to the mean. The figure shows that glucose (highlighted in red) is by no means unusual. 36 percent (18) carbon sources have an even greater average innovation index than glucose. For example, acetate allows viability on the greatest number (9.75) of additional carbon sources. Conversely, some carbon sources such as adenosine ($I_{Adenosine} = 0.27$) and deoxyadenosine ($I_{Deoxyadenosine} = 0.1$) allow growth on fewer additional carbon sources than glucose. Carbon sources with a small average innovation index – they entail viability on few additional carbon sources – are also more variable in this innovation index (supplementary figure 5, Spearman's $\rho = -0.82$, $p < 10^{-101}$). Even though any one carbon source may confer growth on only few additional carbon sources in any one network (figure 2a), when considering all networks in a sample, it may still allow pre-adaptation to most other carbon sources (supplementary figure 6).

In sum, viability on any one carbon source *C* usually entails viability on multiple other carbon sources, whose number and identity can vary with *C*. Viability on never before encountered carbon sources is thus a typical metabolic property. Environmental generalists capable of surviving on multiple carbon sources may be viable on many more carbon sources than occur in their environment (supplementary tables 2 and 3, supplementary figure 10).

We next asked whether metabolically close carbon sources show the highest potential for pre-adaptation. The centre path of figure 2b shows a hypothetical metabolic pathway that leads from a carbon source *C* to a source C_{new} (boxed area) and from there through (possibly multiple) further metabolic reactions to the synthesis of biomass. Figure 2c shows the same scenario, except that C and C_{new} are separated by several further reactions. It is possible that random networks viable on *C* are more likely to be viable also on C_{new}, if C_{new} is closer to *C*, i.e., if they are separated by fewer metabolic reactions, as in the scenario of figure 2b. In this case, metabolite C_{new} may be less easy to by-pass through an alternative pathway that originates somewhere between *C* and C_{new} (right-most sequence of arrows in figure 2c).

To test this hypothesis (see also supplementary results), we analysed our 50 samples of 500 random metabolic networks, where networks in each sample were required to be viable on a different one of our 50 carbon sources *C*. For each sample (carbon source *C*), and for each of the other 49 possible carbon sources C_{new}, we asked whether the metabolic distance between *C* and C_{new} is correlated with the fraction of networks that are also viable on C_{new}. To do this, we used metabolic networks that were selected for growth on *C* and additionally viable on C_{new} (methods). We then computed the mean metabolic distance and binned the distances. The results, pooled for all networks are shown on the vertical axis of figure 2d, whose horizontal axis reflects the mean metabolic distance (binned into 9 bins). If a carbon source C_{new} is closer to a carbon source *C*, then significantly more networks viable on *C* are also viable on C_{new} (Spearman's $\rho = -0.42$, $p = 10^{-87}$, $n = 1990$). However, the figure also shows that the association is highly noisy, and especially so at low metabolic distances. Taking reaction irreversibility into account yields the same result (Spearman's $\rho = -0.39$, $p = 10^{-57}$, $n = 1601$), as does a different way of computing distances between pairs of carbon sources *C* and C_{new} (methods and supplementary results). The association is noisy, because metabolism is highly reticulate (supplementary results).

While metabolic ‘nearness’ cannot explain exaptations involving two carbon sources, biochemical similarities help explain why a network viable on C might be viable on one additional carbon source C_{n1}, but not on another source C_{n2}. Indeed, exaptations often involve carbon sources with broadly defined biochemical similarities (supplementary figures 7 and 8). For example, glycolytic carbon sources are more likely to entail exaptations for growth on other glycolytic carbon sources, and likewise for gluconeogenic carbon sources, as well as for carbon sources involved in nucleotide metabolism. Furthermore, we also show that pre-adaptation is synergistic, that is, the innovation index for a pair of carbon sources is greater than the sum of the innovation indices I_{C1} and I_{C2} (supplementary figure 9).

Limitations of our analysis include that; first, it is based on current knowledge about the reaction universe. Future work may increase the number of known reactions, but this would not diminish, but could only enhance the spectrum of possible exaptations. The reason is that additional reactions would allow the utilization of additional carbon sources by some metabolic networks. Second, most of our analysis focused on random networks that are viable on a specific carbon source, but selection in the wild can affect more than viability, which may affect the incidence of exaptations. Of special importance is selection favouring networks with a high rate of biomass synthesis. This particular selective constraint would not affect our conclusions, because we found that networks with high biomass synthesis rates have even greater potential for metabolic innovation than merely viable networks (supplementary table 4 and supplementary figure 11). Third, we considered all necessary nutrient transporters to be present (see methods). If this is not the case, the incidence of exaptation may be reduced. In this regard, we note that 84 percent of *E. coli* transporters can transport multiple molecules[17], and that their substrate specificity can change rapidly[18], thus ameliorating this constraint. Fourth, real metabolic networks may contain more reactions connected to the rest of metabolism than our randomly sampled networks. However, when restricting our analysis to networks in which all reactions are connected, we found an even greater incidence of exaptation than in random networks (see methods and supplementary results, supplementary figure 3).

Thus, our results provide a lower bound on the incidence of exaptations. Finally, most of our analysis is based on sampling a limited number of 500 networks viable on each carbon source, but sampling of 5000 random networks for select carbon sources yielded identical results (supplementary figure 12).

Our observations show that latent metabolic abilities are pervasive features of carbon metabolism. They expose non-adaptive origins of potentially useful carbon source utilization traits as a universal and inevitable feature of metabolism. The abundance of non-adaptive trait origins results from the complexity of metabolic systems, which have many enzyme parts that can jointly form multiple metabolic phenotypes, but this ability is not restricted to metabolic networks. Many enzymes are capable of utilizing various substrates[17, 19], which can further increase network complexity and the potential for exaptation. The ability to form multiple phenotypes also occurs in regulatory circuits[20], which can form different molecular activity patterns, as well as RNA molecules[21], which can form multiple conformations with different biological functions. Systematic analyses of genotype-phenotype relationships are becoming increasingly possible in such systems[22,23], and already hint at exaptive origins of molecular traits. If confirmed in systematic analyses like ours, the pervasiveness of non-adaptive traits may require a re-thinking of the early origins of beneficial traits.

METHODS SUMMARY

We used Markov Chain Monte Carlo (MCMC) random walks that use reaction-swapping to sample random viable metabolic networks[12], as well as flux balance analysis[24] to compute the viability of metabolic networks during the MCMC procedure. We performed all analyses for minimal aerobic growth environments composed of a sole carbon source, along with oxygen, ammonium, inorganic phosphate, sulfate, sodium, potassium, cobalt, iron (Fe^{2+} and Fe^{3+}), protons, water, molybdate, copper, calcium, chloride, magnesium, manganese and zinc[15].

REFERENCES

1. Darwin, C. *On the Origin of Species.* Charles Darwin. With an introduction by Ernst Mayr. Harvard University Press, Cambridge, Mass., 1964 (facsimile of the first edition, 1859). x 502 pp. 5.95. *Science* **146**, 51–52 (1859).

2. Gould, S. J. & Vrba, E. S. Exaptation-a missing term in the science of form. *Paleobiology* **8**, 4–15 (1982).

3. True, J. R. & Carroll, S. B. Gene co-option in physiological and morphological evolution. *Annual review of cell and developmental biology* **18**, 53–80 (2002).

4. Zákány, J. & Duboule, D. Hox genes in digit development and evolution. *Cell and tissue research* **296**, 19–25 (1999).

5. Keys, D. N. *et al.* Recruitment of a hedgehog regulatory circuit in butterfly eyespot evolution. *Science (New York, N.Y.)* **283**, 532–4 (1999).

6. Tomarev, S. I. & Piatigorsky, J. Lens Crystallins of Invertebrates. Diversity and Recruitment from Detoxification Enzymes and Novel Proteins. *European Journal of Biochemistry* **235**, 449–465 (1996).

7. Pievani, T. & Serrelli, E. Exaptation in human evolution: how to test adaptive vs exaptive evolutionary hypotheses. *Journal of anthropological sciences = Rivista di antropologia□: JASS / Istituto italiano di antropologia* **89**, 9–23 (2011).

8. Schuster, P., Fontana, W., Stadler, P. F. & Hofacker, I. L. From sequences to shapes and back: a case study in RNA secondary structures. *Proceedings. Biological sciences / The Royal Society* **255**, 279–84 (1994).

9. Lipman, D. J. & Wilbur, W. J. Modelling neutral and selective evolution of protein folding. *Proceedings. Biological sciences / The Royal Society* **245**, 7–11 (1991).

10. Cowperthwaite, M. C., Economo, E. P., Harcombe, W. R., Miller, E. L. & Meyers, L. A. The ascent of the abundant: how mutational networks constrain evolution. *PLoS computational biology* **4**, e1000110, doi:10.1371/journal.pcbi.1000110 (2008).

11. Ferrada, E. & Wagner, A. A Comparison of Genotype-Phenotype Maps for RNA and Proteins. *Biophysical Journal* **102**, 1916–1925 (2012).

12. Samal, A., Matias Rodrigues, J. F., Jost, J., Martin, O. C. & Wagner, A. Genotype networks in metabolic reaction spaces. *BMC systems biology* **4**, 30, doi:10.1186/1752-0509-4-30 (2010).

13. Poulsen, T. S., Chang, Y.-Y. & Hove-Jensen, B. D-Allose Catabolism of Escherichia coli: Involvement of alsI and Regulation of als Regulon Expression by Allose and Ribose. *J. Bacteriol.* **181**, 7126–7130 (1999).

14. Meijnen, J.-P., De Winde, J. H. & Ruijssenaars, H. J. Engineering Pseudomonas putida S12 for efficient utilization of D-xylose and L-arabinose. *Applied and environmental microbiology* **74**, 5031–7 (2008).

15. Feist, A. M. *et al.* A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* **3**, 121, doi:10.1038/msb4100155 (2007).

16. Neidhardt, F. & Ingraham, J. *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*. **1**, (American Society for Microbiology, Washington, DC: 1987).

17. Nam, H. *et al.* Network context and selection in the evolution to enzyme specificity. *Science (New York, N.Y.)* **337**, 1101–4 (2012).

18. Aguilar, C. *et al.* Genetic changes during a laboratory adaptive evolution process that allowed fast growth in glucose to an Escherichia coli strain lacking the major glucose transport system. *BMC genomics* **13**, 385, doi:10.1186/1471-2164-13-385 (2012).

19. Kim, J., Kershner, J. P., Novikov, Y., Shoemaker, R. K. & Copley, S. D. Three serendipitous pathways in E. coli can bypass a block in pyridoxal-5'-phosphate synthesis. *Molecular systems biology* **6**, 436, doi:10.1038/msb.2010.88 (2010).

20. Martin, O. C. & Wagner, A. Multifunctionality and robustness trade-offs in model genetic circuits. *Biophysical journal* **94**, 2927–37 (2008).

21. Ancel, L. W. & Fontana, W. Plasticity, evolvability, and modularity in RNA. *The Journal of experimental zoology* **288**, 242–83 (2000).

22. Amitai, G., Gupta, R. D. & Tawfik, D. S. Latent evolutionary potentials under the neutral mutational drift of an enzyme. *HFSP journal* **1**, 67–78 (2007).

23. Isalan, M. *et al.* Evolvability and hierarchy in rewired bacterial gene networks. *Nature* **452**, 840–5 (2008).

24. Price, N. D., Reed, J. L. & Palsson, B. Ø. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature reviews. Microbiology* **2**, 886–97 (2004).

Supplementary information is linked to the online version of the paper at www.nature.com/nature

Acknowledgements We thank Karthik Raman, Niv Sabath, Joshua Levi Payne and João Rodrigues for helpful discussions. A.W. acknowledges financial support through Swiss National Science Foundation grant 315230-129708.

Author Contributions A.B. and A.W. designed research; A.B. performed research; A.B. contributed new reagents/analytic tools; A.B. and A.W. analysed data; and A.B. and A.W. wrote the paper.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials and source code should be addressed to A.W. (andreas.wagner@ieu.uzh.ch).

FIGURE LEGENDS

Figure 1 – **Viability on glucose entails viability on multiple other carbon sources.** (a) The binary innovation vector of a hypothetical metabolic network that is viable on glucose. The vector shows that the random network is viable on glucose, sorbitol and fructose (marked by 1), but not viable on pyruvate and acetate (marked by 0). The innovation index of this network ($I_{Glucose} = 2$) denotes the number of additional carbon sources the network is viable on. (b) The distribution of innovation indices for 500 random networks viable on glucose. Only 4 percent of networks have $I_{Glucose} = 0$, meaning that they are viable only on glucose.

Figure 2 – **Innovation varies with respect to the carbon source *C* and the mean metabolic distance between *C* and C_{new}.** (a) For each of 50 carbon sources *C* (horizontal axis), the figure indicates the mean innovation index (bar) and its coefficient of variation (lines) for 500 random networks required to be viable on carbon source *C*. Note the broad distribution of the index. Some carbon sources such as acetate allow viability on more than nine additional carbon sources on average, while others, such as deoxyadenosine support viability on fewer than one additional carbon sources. The innovation index of glucose (in red) is typical compared to other carbon sources. (b) The figure shows a hypothetical carbon source C_{new}, which can be synthesized from some carbon source *C* in one reaction (arrow), and which leads through multiple further reaction to the synthesis of biomass. Some metabolic networks may have an alternative metabolic pathway that by-passes C_{new} altogether (right sequence of arrows). (c) Like (b), but C_{new} and *C* are separated by multiple reactions. The fewer reactions separate *C* and C_{new}, the more likely it is that C_{new} is not by-passed by some alternative metabolic pathway, and that therefore viability on

C implies viability on C_{new}. This is the hypothesis tested in the analysis of (d). The horizontal axis of (d) indicates the mean number of reactions that separate C and C_{new} in networks that are viable on both C and C_{new}, binned into integer intervals corresponding to the floor of this number. The vertical axis indicates the fraction of random metabolic networks required to be viable on carbon source C that are additionally viable on C_{new}. Note that the potential for innovation decreases with increasing distance. Box edges: 25th and the 75th percentiles; central horizontal line in each box: median; whiskers: ± 2.7 standard deviations; open circles: outliers. Data are based on samples of 500 random viable networks for each of 50 carbon sources C ($n = 25000$).

Online Methods

Flux balance analysis (FBA)

FBA is a constraint-based computational method[24,25] used to predict synthetic abilities and other properties of large metabolic networks, which are complex systems of enzyme-catalysed chemical reactions. FBA requires information about the stoichiometry of each molecular species participating in the chemical reactions of a metabolic network. This stoichiometric information is represented as a stoichiometric matrix $\mathbf{S}$ of dimensions $m \ x \ n$, where m denotes the number of metabolites and n denotes the number of reactions in a network[24,25]. FBA also assumes that the network is in a metabolic steady-state, such as would be attained by an exponentially growing microbial population in an unchanging environment. This assumption allows one to impose the constraint of mass conservation on the metabolites in the network. This constraint can be expressed as

$$\mathbf{Sv} = 0$$

wherein $\mathbf{v}$ denotes a vector of metabolic fluxes whose entries $\boldsymbol{v_i}$ describe the rate at which reaction i proceeds. The solutions – ‘allowable’ fluxes – of this equation form a large solution space, but not all of these solutions may be of biological interest. To restrict this space to fluxes of interest, FBA uses linear programming to maximize a biologically relevant quantity in the form of a linear objective function Z[25].

Specifically, the linear programming formulation of an FBA problem can be expressed as

$$\max Z = \max \{\mathbf{c}^T\mathbf{v} \mid \mathbf{S}\mathbf{v} = 0, \mathbf{a} \leq \mathbf{v} \leq \mathbf{b}\}$$

The vector **c** contains a set of scalar coefficients that represent the maximization criterion, and vectors **a** and **b** contain the minimally and maximally possible fluxes for each reaction in **v**, respectively.

We are here interested in predicting if a metabolic network can sustain life in a given spectrum of environments, that is, whether it can synthesize all necessary small biomass molecules (biomass precursors) required for survival and growth. In a free-living bacterium such as *E. coli*, there are more than 60 such molecules, which include 20 proteinaceous amino acids, DNA and RNA nucleotide precursors, lipids, and cofactors. We use the *E. coli* biomass composition[15] to define the objective function and the vector **c**, because most molecules in *E. coli's* biomass would be typically found in free-living organisms. We used the package CLP (1.4, Coin-OR; https://projects/coin-or.org/Clp) to solve the linear programming problems mentioned above.

Chemical environments

Along with the biomass composition and stoichiometric information about a metabolic network, one needs to define one or more chemical environments that contain the nutrients needed to synthesize biomass precursors. We here consider only minimal aerobic growth environments composed of a sole carbon source, along with oxygen, ammonium, inorganic phosphate, sulfate, sodium, potassium, cobalt, iron (Fe^{2+} and Fe^{3+}), protons, water, molybdate, copper, calcium, chloride, magnesium, manganese and zinc[15]. When studying viability of a metabolic network in different environments, we vary the carbon source while keeping all other nutrients constant. When we say, for example, that a particular network is viable on 20 carbon sources, we mean that the network can synthesize all biomass precursors when each of these carbon sources is provided as the *sole* carbon source in a minimal medium. For reasons of computational feasibility, we restrict ourselves to 50 carbon sources (supplementary table 1). They are all carbon sources on which *E. coli* is known to be viable from experiments[15]. We chose these carbon sources because many of them are prominent, and because they are of known biological relevance, but we emphasize that our observations do not otherwise make a statement about the metabolism of *E. coli* or its close relatives. They apply to metabolic networks that vary much more broadly in reaction composition than any relative of *E. coli,* because of our network sampling approach described below, which effectively randomizes the reaction composition of a microbial metabolism.

The known reaction universe

The known reaction universe is a list of metabolic reactions known to occur in some organisms. For the construction of this universe, we used data from the LIGAND database[26,27] of the Kyoto Encyclopaedia of Genes and Genomes[28,29]. The LIGAND database is divided into two subsets – the REACTION and the COMPOUND database. These two databases together provide information about metabolic reactions, participating chemical compounds, and associated stoichiometric information in an interlinked manner.

As we also described earlier[12,30,31], we specifically used the REACTION and the COMPOUND databases to construct our universe of reactions while excluding - (i) all reactions involving polymer metabolites of unspecified numbers of monomers, or general polymerization reactions with uncertain stoichiometry, (ii) reactions involving glycans, due to their complex structure, (iii) reactions with unbalanced stoichiometry, and (iv), reactions involving complex metabolites without chemical information[29]. The published *E. coli* metabolic model (*i*AF1260) consists of 1397 non-transport reactions[15]. We merged all reactions in the *E. coli* model with the reactions in the KEGG dataset, and retained only the non-duplicate reactions. After these procedures of pruning and merging, our universe of reactions consisted of 5906 non-transport reactions and 5030 metabolites.

Sampling of random viable metabolic networks

In an organism, a metabolic network can change through mutations. They can lead to addition of new reactions, by way of horizontal-gene transfer, or through the evolution of enzymes with novel activities. They can also lead to loss of reactions through loss-of-function mutations in enzyme-coding genes. Natural selection can preserve those changed metabolic networks that are viable in a particular environment. Together, mutational processes and selection may change a metabolic network drastically on long evolutionary time-scale. Recent work has shown that even metabolic networks that differ greatly in their sets of reactions can have the same metabolic phenotype, that is, the same biosynthetic ability[32]. We here employ a recently developed Markov Chain Monte Carlo (MCMC) random sampling[12,30,31,33,34] procedure to generate metabolic networks that are viable in specific environments, but that contain an otherwise random complement of metabolic reactions. Briefly, this procedure involves random walks in the space of all possible networks. During any one such random walk, a metabolic network can change through the addition and deletion of reactions. Although this process resembles the biological evolution of metabolic networks through horizontal gene transfer and (recombination-driven) gene deletions, we here use it for the sole purpose to create random samples of metabolic networks from the space of all such networks[12,34].

In any one MCMC random walk, we keep the total number of reactions at the same number (1397[15]) as the starting *E. coli* network, in order to avoid artifacts due to varying reaction network size[12]. Specifically, each mutation step in a random walk involves an addition of a randomly chosen reaction from the reaction universe, followed by a deletion of a randomly chosen metabolic reaction from the metabolic network. We call such a sequence of reaction addition and deletion a reaction swap. Reaction addition does not abolish the viability of a network in any environment. However, reaction deletion might. Thus, after a reaction deletion, we use FBA to ask whether the network is still viable – it can synthesize all biomass precursors -- in the specified environment. If so, we accept the deletion; otherwise, we reject it and choose another reaction for deletion at random, until we have found a deletion that retains viability. After that, we accept the reaction swap, thus completing a single step in the random walk. We do not subject transport reactions to reaction swaps. These reactions are therefore present in all networks generated by our random walk.

Any MCMC random walk begins from a single starting network, in our case that of *E. coli*. The theory behind MCMC sampling[12,34] , shows that it is important to carry out as many reaction swaps as possible for MCMC to 'erase' the random walker's similarity ('memory') of the initial network. The reason is that successive genotypes in a random walk are strongly correlated in their properties, because they differ by only one reaction pair. These correlations fade with an increasing number of reaction swaps. Because we are interested in analyzing growth phenotypes of networks, correlations to the initial network would result in identification of growth on carbon sources similar to those of the starting network. In past work[12,30], we found that for the network sizes that we use (1397 reactions), 3 x 10^3 reactions swaps are sufficient to erase the similarity of the final network to the starting network. To err on the side of caution, we thus carried out 5 x 10^3 reaction swaps before beginning to sample, and sample a network every 5 x 10^3 reaction swaps thereafter. In this way, we generated samples of 500 random viable metabolic networks through an MCMC random walk of 2.5 x 10^6 reaction swaps. We carried out different random walks to sample networks viable on different carbon sources.

For some of our analyses, we also sampled random metabolic networks of sizes different from that of the *E. coli* metabolic network. To do this, we followed a previously established procedure[12,30,31] to create a starting network for an MCMC random walk that has the desired size. This procedure first converts the known universe of reactions into a 'global' metabolic network by including the *E. coli* transport reactions in it. Not surprisingly, this global network can produce all biomass components and is therefore viable on all carbon sources studied here. We used this global network to successively delete a sequence of randomly chosen reactions in the following way. After each reaction deletion, FBA is used to ask whether the network is still viable on a given carbon source. If so, the deletion is accepted; otherwise another reaction is chosen at random for deletion. We deleted in this way as many reactions as needed to generate a network of the desired size. We then used this network as the starting network for an MCMC random walk, as described above, to generate samples of 500 random viable networks.

Identification of disconnected non-functional reactions and the connected reaction universe

We performed some of our analysis with a version of the reaction universe that does not contain disconnected reations. Reactions that are not connected to the rest of a metabolic network would be nonfunctional, because they cannot carry a non-zero steady-state metabolic flux, and thus could not contribute to the synthesis of biomass. The genes encoding them would eventually be lost from a genome. (We note that this loss could still take tens of thousands of years, given known deleterious mutation rates and generation times[35,36] , enough for some for other genetic or environmental changes to render these reactions functional.) We define a disconnected reaction as a reaction that does not share any one substrate or any one product with any other reaction in the known reaction universe. We focus here on reactions in the universe rather than in one metabolic network, because an individual network can gain additional reactions that may connect previously disconnected reactions. We note that even this "universal" definition of disconnectedness depends on our current knowledge of biochemistry, as well as on the environment, for the right environment could supply metabolites that connect previously disconnected reactions or pathways to the rest of a metabolic network. To identify the connected universe, we removed disconnected reactions. Because this removal may render other reactions disconnected, we repeated this process iteratively until no further reactions in the universe became disconnected. In this way, we found 3646 reactions of the 5906 reactions in the universe of reactions to be connected. We used this connected universe in some analyses to generate network samples using the MCMC approach.

Estimation of the metabolic distance between carbon sources

To compute the metabolic distance between a pair of carbon sources *C* and C_{new}, we used the 500 networks selected for growth on a specific carbon source *C*. We first represented a network as a *substrate graph*[37]. In this graph, vertices correspond to metabolites. Two metabolites (vertices) are linked by an edge if the metabolites participate in the same metabolic reaction, be it as an educt or as a product. We excluded 'currency' metabolites from this substrate graph, which are metabolites that transfer small chemical groups and are involved in many reactions[38]. Specifically, we excluded protons, H_2O, ATP (adenosine triphosphate), ADP (adenosine diphosphate), AMP (adenosine monophosphate), NADP(H) (nicotinamide adenosine dinucleotide diphosphate), NAD(H) (nicotinamide adenosine dinucleotide), and Pi (inorganic phosphate), CoA (coenzyme A), hydrogen peroxide, ammonia, ammonium, bicarbonate, GTP (guanosine triphosphate), GDP (guanosine diphosphate), and PPi (diphosphate) that occurred in both the cytoplasmic and periplasmic compartments[15]. In addition we also excluded oxidized and reduced forms of cofactors such as quinone, ubiquinone, glutathione, thioredoxin, flavodoxin and flavin moninucleotide. That is, we eliminated all vertices corresponding to these metabolites when constructing the substrate graph. For each metabolic network, we constructed two substrate graphs, first one wherein the reaction irreversibility was ignored and all reactions were considered reversible, and the second graph wherein irreversibility was taken into account. For a network selected for growth on carbon source *C*, we calculated the shortest distance of *C* to each exapted carbon source C_{new} in the substrate graph of that network, as computed by a breadth-first search[39]. We preformed this analysis for each network in our ensemble of 500 networks viable on a carbon source *C*. The distance between carbon sources *C* and C_{new} was then computed as a mean of the metabolic distances based on networks viable on both carbon sources.

We also computed metabolic distance for any two carbon sources by representing the universe of reactions as a graph in the above manner. We again constructed two substrate graphs, first one wherein the reaction irreversibility was ignored and all reactions were considered reversible, and the second graph wherein irreversibility was taken into account. Taking irreversibility into account increases the maximal distance to infinity as some carbon sources are connected by irreversible reactions.

Clustering of carbon sources based on the innovation matrix

The entries of the innovation matrix $\boldsymbol{I} = (I_{ij})$ represent the fraction of random metabolic networks that we required to be viable on carbon source C_i, and that were additionally viable on carbon source C_j. To cluster the entries of this matrix, we first computed for all pairs of rows in this matrix the quantity $d = 1\text{-}\rho$, where ρ is the Spearman rank correlation coefficient between the row entries. This yielded a new, distance matrix which describes the distances between all pairs of rows. We clustered the rows of $\boldsymbol{I}$ by applying UPGMA (Unweighted Pair Group Method with Arithmetic means[40]), a hierarchical clustering method, to the distance matrix.

Hierarchical clustering with UPGMA classifies data such that the average distance between elements belonging to the same cluster is lower than the average distance between elements belonging to different clusters[12]. UPGMA identified two clusters of glycolytic and gluconeogenic carbon sources, and we wanted to know whether the distances between them were significantly different. To this end, we first calculated the distribution of distances $d = 1\text{-}\rho$ for all pairs of row vectors of $\boldsymbol{I}$ *within* each of the two clusters. We called the resulting distance distribution the 'within-cluster' distance distribution. Similarly, we computed the distances *between* any pair of row vectors belonging to two different clusters. These formed a 'between-cluster' distance distribution. We then used the non-parametric Mann-Whitney U-test to check if these two distributions were significantly different.

Estimation of carbon waste production

FBA determines the maximal biomass yield achievable by a network for a given carbon source[25]. However, even when a network produces the maximally achievable yield, not all of the carbon input into the network may be converted into biomass. The non-converted carbon input constitutes carbon waste. Such non-utilized carbon can be secreted in the form of one or more metabolites. For example, in a glucose minimal environment, *E. coli* secretes carbon dioxide and acetate into the extracellular compartment as carbon waste. FBA estimates the amount of each metabolite secreted per unit time[15,25]. To estimate the amount of carbon waste that a random network viable on glucose produces, we first identified the different metabolites that it secretes as waste, and then computed the amount of carbon waste per metabolite as the product of carbon atoms in that metabolite and the amount of the metabolite secreted (mmol/gram dry weight/hour). The total carbon waste produced by a network computes as the sum of the above quantity over all secreted carbon-containing molecules. We repeated the above procedure for each random network in a sample of 500 random networks viable on glucose. We found a total of 62 metabolites that are secreted as waste metabolites in at least one network of our sample of networks viable on glucose.

We carried out all numerical analyses using MATLAB (Mathworks Inc.)

25. Kauffman, K. J., Prakash, P. & Edwards, J. S. Advances in flux balance analysis. *Current opinion in biotechnology* **14**, 491–6 (2003).

26. Goto, S., Nishioka, T. & Kanehisa, M. LIGAND: chemical database of enzyme reactions. *Nucleic Acids Res* **28**, 380–382 (2000).

27. Goto, S., Okuno, Y., Hattori, M., Nishioka, T. & Kanehisa, M. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic acids research* **30**, 402–4 (2002).

28. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27–30 (2000).

29. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research* **38**, D355–60 (2010).

30. Barve, A., Rodrigues, J. F. M. & Wagner, A. Superessential reactions in metabolic networks. *Proceedings of the National Academy of Sciences* **109**, E1121–1130 (2012).

31. Matias Rodrigues, J. F. & Wagner, A. Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS computational biology* **5**, e1000613, doi:10.1371/journal.pcbi.1000613 (2009).

32. Henry, C. S. *et al.* High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology* **28**, 977–82 (2010).

33. Matias Rodrigues, J. F. & Wagner, A. Genotype networks, innovation, and robustness in sulfur metabolism. *BMC Syst Biol* **5**, 39, doi:10.1186/1752-0509-5-39 (2011).

34. Binder, K. & Heerman, D. W. *Monte Carlo Simulation in Statistical Physics*. (Springer: Hiedelberg, 2010).

35. Koskiniemi, S., Sun, S., Berg, O. G. & Andersson, D. I. Selection-driven gene loss in bacteria. *PLoS genetics* **8**, e1002787, doi:10.1371/journal.pgen.1002787 (2012).

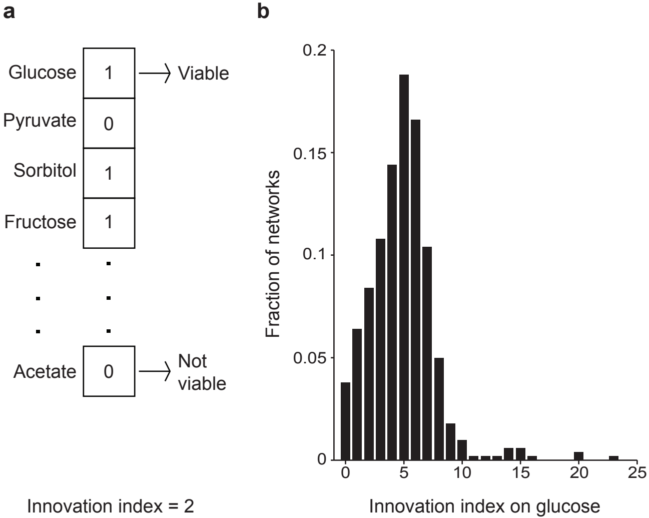
36. Ochman, H., Elwyn, S. & Moran, N. A. Calibrating bacterial evolution. *Proceedings of the National Academy of Sciences* **96**, 12638–12643 (1999).

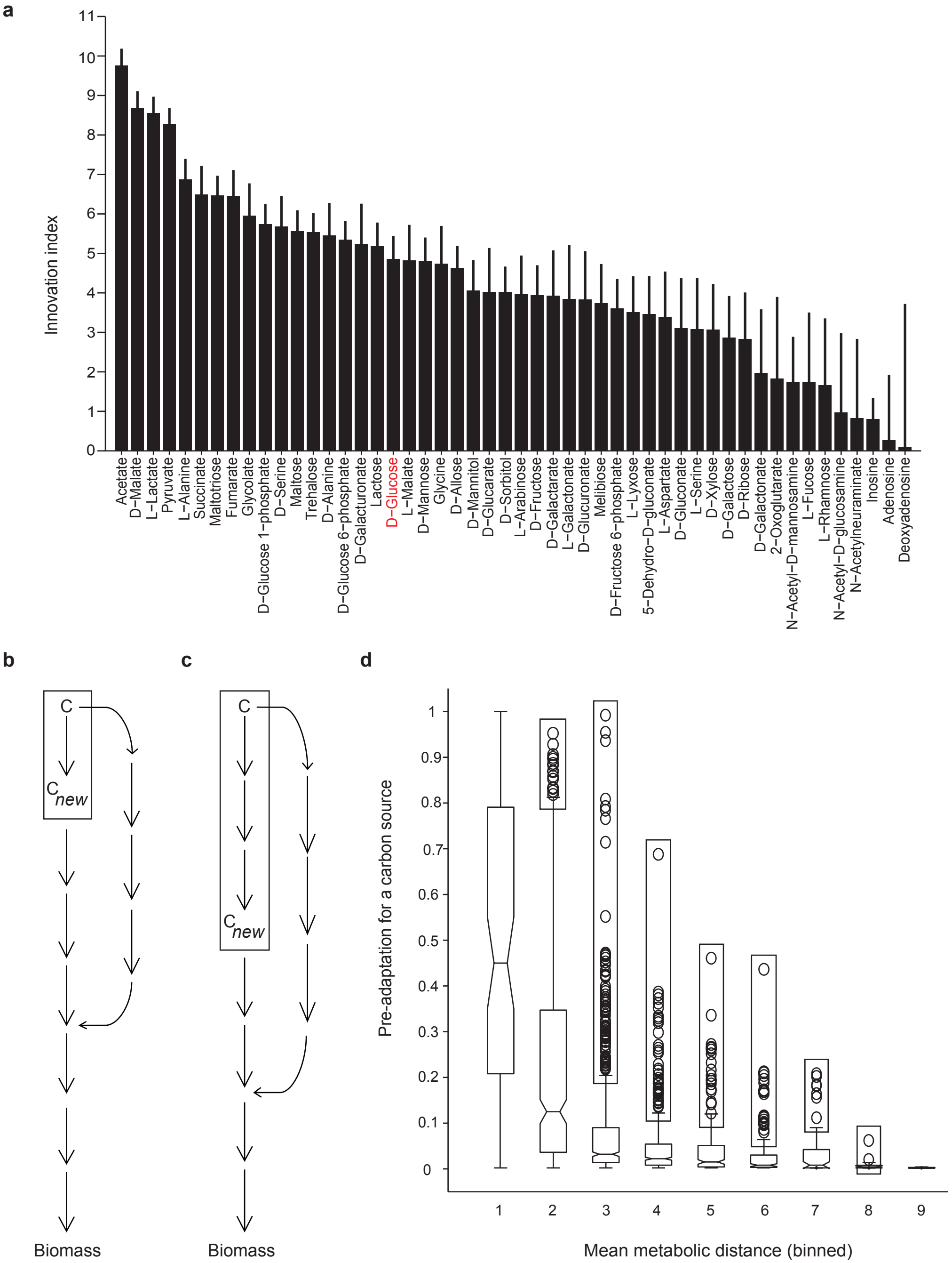
37. Wagner, A. & Fell, D. A. The small world inside large metabolic networks. *Proceedings. Biological sciences / The Royal Society* **268**, 1803–10 (2001).

38. Ma, H.-W. & Zeng, A.-P. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics (Oxford, England)* **19**, 1423–30 (2003).

39. Moore, E. The shortest path through a maze. *Proceedings of the International Symposium on the Theory of Switching, and Annals of the Computation Laboratory of Harvard University* 285–292 (1959).

40. Sokal, R. R. & Michener, C. D. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin* **28**, 1409–1438 (1958).


a
Glucose
1
Viable
Pyruvate
0
Sorbitol
1
Fructose
1
Acetate
0
Not viable
Innovation index = 2
b
Fraction of networks
0.2
0.15
0.1
0.05
0
0
5
10
15
20
25
Innovation index on glucose


a
Innovation index
0
1
2
3
4
5
6
7
8
9
10
11
Acetate
D−Malate
L−Lactate
Pyruvate
L−Alanine
Succinate
Maltotriose
Fumarate
Glycolate
D−Glucose 1−phosphate
D−Serine
Maltose
Trehalose
D−Alanine
D−Glucose 6−phosphate
D−Galacturonate
Lactose
D−Glucose
L−Malate
D−Mannose
Glycine
D−Allose
D−Mannitol
D−Glucarate
D−Sorbitol
L−Arabinose
D−Fructose
D−Galactarate
L−Galactonate
D−Glucuronate
Melibiose
D−Fructose 6−phosphate
L−Lyxose
5−Dehydro−D−gluconate
L−Aspartate
D−Gluconate
L−Serine
D−Xylose
D−Galactose
D−Ribose
D−Galactonate
2−Oxoglutarate
N−Acetyl−D−mannosamine
L−Fucose
L−Rhamnose
N−Acetyl−D−glucosamine
N−Acetylneuraminate
Inosine
Adenosine
Deoxyadenosine
b
C
C_{new}
Biomass
c
C
C_{new}
Biomass
d
Pre-adaptation for a carbon source
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1
2
3
4
5
6
7
8
9
Mean metabolic distance (binned)

SUPPLEMENTARY INFORMATION

Supplementary Results

Networks selected for growth on glucose are pre-adapted to different carbon sources. After having shown that networks required to be viable on glucose are also viable on multiple other carbon sources (figure1b), we inquired whether the additional carbon sources to which a metabolic network is pre-adapted differ between different random viable networks, or whether they are mostly identical. We found that most of these carbon sources differ among networks. Specifically, 91.84 percent (45) of the additional carbon sources occur in the innovation vectors of fewer than 40 percent of the networks (supplementary figure 1). We also computed the fraction of carbon sources that both networks in a pair are viable on, among all carbon sources that at least one network in a pair is viable on. This distribution (supplementary figure 2a) has a mean of only 31.8 percent. In other words, for almost 70 percent of carbon sources to which one network is pre-adapted, the other network is not pre-adapted. Fully 23.5 percent of network pairs do not share viability on any carbon source aside from glucose (supplementary figure 2a).

We next computed the pairwise distance between the innovation vectors $I_{Glucose}$ for all 500 metabolic networks in our sample. This distance indicates the number of additional carbon sources that one but not the other network in a pair is viable on. Its distribution (supplementary figure 2b) has a mean of 5.22 carbon sources (s.dev. = 3.16). That is, two networks differ on average in their viability on 5 carbon sources. The distribution is right-skewed (supplementary figure 2b) and contains networks that differ in their viability on many carbon sources. The two networks with the maximum distance of 26.53 are viable on 5 and 23 carbon sources in addition to glucose, but only one of the additional carbon sources is shared between them.

Higher reaction connectivity increases exaptation. The MCMC random walk (see methods) entails the addition of a randomly chosen reaction from the universe of reaction, followed by the deletion of a randomly chosen reaction. The deletion is accepted only if the network continues to be viable on the given carbon source. Thus, viability is the only constraint we enforce while sampling random networks. Reactions can be added that are not connected to the rest of the metabolism at the time when they are added, but such reactions cannot carry a non-zero flux and would therefore be non-functional (although they could become connected after additional reaction changes)[41]. To assess the effect of disconnected reactions on exaptation, we identified disconnected reactions and removed them from our universe of reactions (see methods) to generate a connected universe of reactions. We used the connected universe of reactions to generate 500 random networks viable on glucose via MCMC sampling. The average innovation index of these networks is equal to 10.5 (s.dev. = 6.7 carbon sources), which is higher than for networks generated using the complete universe of reactions (Mann-Whitney U-test, $p = 10^{-64}$).

In a further analysis to understand the role of connected reactions, we modified the MCMC random walk in the following manner. We allowed the addition of a randomly chosen reaction only if all of its substrates participated in at least one other reaction in our network, thus ensuring that only connected reactions can be added to a network. (We note that this procedure no longer guarantees detailed balance[42] and uniform sampling of the space of viable networks.) We generated 500 metabolic networks viable on one carbon source, for each of 50 carbon sources using the modified random walk (a total of 500x50=25,000 networks). We then computed the innovation index for each network. We found that for networks viable on glucose, the mean innovation index is much higher (I = 20.6, s.dev. = 8.52, supplementary figure 3, glucose highlighted in red) than in our original sample of MCMC-generated networks viable on glucose (I = 4.86, s.dev. = 2.83, figure 1b). Supplementary figure 3 shows that this is also true for all other carbon sources. For example, networks viable on acetate are viable on 9.75 other carbon sources in our original sample of networks (figure 2a), while they are viable on 27.67 new carbon sources when we constrain the random walk to adding only connected reactions. Addition of only connected reactions leads to more exaptation in metabolic networks, presumably because these reactions allow more alternate routes towards biomass synthesis. We note that the connectedness of a reaction depends on current knowledge of biochemistry, as well as on the environment, for the right environment could supply metabolites that connect previously disconnected reactions or pathways to the rest of a metabolic network.

Large metabolic networks have higher innovation potential. The size of a metabolic network is the number of reactions participating in the network. In most of our analyses, we focus on random networks with the same size as that of the *E. coli* metabolic network (1397 reactions[15]). These have a mean innovation index of 4.86 for viability on glucose. However, it is possible that this index may depend on the number of reactions in a network. Larger networks might be more likely to metabolize carbon sources in addition to those on which selection acts. To find out whether this is the case, we generated six additional samples of 500 random metabolic networks viable on glucose, but where networks in different samples differed in size. Specifically, network sizes ranged between 400 and 1600 reactions. We calculated the mean innovation index for networks in each sample. Supplementary figure 4 shows that innovation is positively correlated with network size (Spearman's $\rho = 0.6$, $p = 10^{-300}$, $n = 3500$). The horizontal axis of supplementary figure 4 denotes the network size categories we considered, and the vertical axis indicates the mean innovation index for networks in one sample (error bars correspond to one standard deviation). The figure shows that larger, more complex networks indeed have a higher innovation index. The figure also shows that the variability in the innovation index increases as network size increases. That is, the number of additional carbon sources on which a network is viable becomes increasingly variable as network complexity increases.

The networks used in most of our analyses were generated using the complete universe of reactions, as opposed to the connected universe described above. We wanted to find out whether removal of some reactions from the complete universe affects the correlation between metabolic network size and the innovation index. To this end, we removed disconnected reactions (see methods) from the networks of each sample. We found that removal of disconnected reactions did not change the correlation between network size and innovation (Spearman's $\rho = 0.59$, $p = 10^{-300}$, n =3500) that we had observed earlier for the complete universe (Spearman's $\rho = 0.6$, $p = 10^{-300}$, $n = 3500$).

Most carbon sources can be subject to pre-adaptation. We also asked whether the proportion of the 49 carbon sources that confers viability to at least one network in our sample is small or large. In other words, is the potential for exaptation restricted to a modest percentage of carbon sources? The answer is no. Almost 90 percent (44) of the additional 49 carbon sources confer viability to at least one network in networks selected for growth on glucose. (We note that this number might be even higher if computational feasibility had not restricted us to samples of 500 networks.) The same holds for networks selected to be viable on most other carbon sources (supplementary figure 6). Each vertical bar of supplementary figure 6 shows, for a network sample viable on a specific carbon source (horizontal axis), the number of additional carbon sources on which at least one network in the sample is viable. For 86 percent (43) of samples, this number is greater than 40, and for 94 percent (47) of samples it is greater than 25, meaning that more than half of the additional carbon sources confer viability to at least one network in the sample. The exceptions are inosine, adenosine, and deoxyadenosine, which can give rise to pre-adaptations on only 4, 3, and 3 other carbon sources, respectively. Thus, even though any one carbon source may confer growth on only few additional carbon sources in any one network (figure 2a), when considering all networks in a sample, it may still allow pre-adaptation to most other carbon sources. For example, viability on xylose allows viability on only three additional carbon sources on average (figure 2a). However, in a sample of 500 networks viable on xylose, pre-adaptation occurs for 43 carbon sources (supplementary figure 6). Pre-adaptation or exaptation can thus occur for the vast majority of carbon sources we examined.

Metabolically close carbon sources show the highest potential for pre-adaptation. To ask whether metabolically close carbon sources show the highest potential for pre-adaptation, we first performed a simple test that relied on the metabolic distance, the minimal number of metabolic reactions separating pairs of carbon sources (C, C_{new}), for all possible pairs that can be formed from our 50 carbon sources (see methods). The maximal distance is six reactions. We then analysed networks selected to be viable on the carbon source glucose. We divided the 49 carbon sources C_{new} different from glucose into two categories, those on which more than the median number of networks in a sample are viable (see distribution in supplementary figure 2), and those on which fewer than this median number are viable. The average metabolic distance of carbon sources to glucose in the two categories is 2.26 (s.dev. = 0.82) and 3.6 (s.dev. = 1.14), respectively, a difference that is statistically significant (Mann-Whitney U-test, $p = 0.007$). This means that carbon sources C_{new} with a greater incidence of pre-adaptation are metabolically closer to glucose.

The analysis in figure 2d shows that the association between the average metabolic distance and the potential for pre-adaptation, especially at low metabolic distances, is noisy. That is, even if a carbon source C_{new} can be produced from C in a single step, the fraction of networks that are viable on C_{new} may range widely from 0.05 to almost one (left-most bin in figure 2d). For example, 92.8 percent of networks viable on acetate are additionally viable on pyruvate as well, whereas only two of 500 networks viable on pyruvate are additionally viable on N-acetylneuraminate, even though both carbon sources are only two reactions away from pyruvate.

It merits explanation why a metabolic network is not always viable on a carbon source C_{new} that can be produced from metabolite C in a single step. For example, the median fraction of networks viable on carbon sources C_{new} that have a distance of one from glucose is only 0.21. The reason is illustrated in the right-most sequence of arrows of figures 2b and 2c. It may be possible to synthesize biomass from carbon source C such that carbon source C_{new} is completely bypassed. For example, D-glucose-6-phosphate (C_{new}) can be produced from glucose (C) in one step. However, not 100 percent but only 77.2 percent of networks viable on glucose are additionally viable on D-glucose-6-phosphate. The remainder (22.8 percent or 114 networks) can bypass D-glucose-6-phosphate. These 114 networks metabolize glucose with either of two reactions. The first is catalysed by xylose isomerase (enzyme commission number (EC) 5.3.1.5), which can convert glucose into fructose[15,43]. The second is catalysed by glucose dehydrogenase (EC 1.1.5.2), which can convert glucose into gluconate[15,44]. In sum, the highly reticulate nature of metabolism allows alternative pathways to by-pass carbon sources very closely related to C, and thus limits the potential for pre-adaptation for any one carbon source C_{new}[45].

We asked whether computing distances between C and C_{new} in the universe of reactions changed the correlation between distance and the fraction of networks that are viable on C_{new}. To do this, we represented the universe of reactions as a substrate graph (see methods), and found that the correlation changed very little (Spearman's $\rho = -0.47$, $p = 10^{-132}$, $n = 2450$). On taking reaction irreversibility into account, 388 of 2500 pairs of carbon sources have infinite distance. However, the correlation between the innovation index and distances between carbon sources C and C_{new} remains unchanged (Spearman's $\rho = -0.39$, $p = 10^{-77}$, $n = 2062$).

Pre-adaptation involves preferably broadly similar carbon sources. We next asked whether any further indicators of biochemical similarity among carbon sources might help understand why a network viable on C might be viable on one additional carbon source C_{n1}, but not on another source C_{n2}. For example, of the 500 random metabolic networks selected for growth on acetate, 89.6 percent networks are also viable on L-serine, which is at a metabolic distance of two from acetate. In contrast, only 6 percent networks are additionally viable on N-acetylneuraminate, which also has distance two from acetate. Is there a difference between L-serine and N-acetylneuraminate that accounts for these differences?

To help us ask this question systematically, we defined an innovation matrix $\boldsymbol{I}$, whose construction is described in supplementary figure 7a. The entries of this matrix I_{ij} contain the fraction of those random metabolic networks that we required to be viable on carbon source C_i, and that were additionally viable on carbon source C_j. The distance between two rows represents differences in the spectrum of carbon sources to which networks required to be viable on C_i and C_j are pre-adapted. We computed a distance measure based on the Spearman's rank correlation coefficient (see methods) for all pairs of row vectors, thus arriving at a distance matrix for these vectors. We then used hierarchical clustering to group carbon sources (row vectors) that allow pre-adaptation on similar spectra of carbon sources. The results are three very distinct and clearly separable groups of carbon sources reflected by deep and statistically significant branches in a dendrogram (supplementary figure 8). Specifically, the three groups comprise (i) glycolytic carbon sources, which are mainly sugars and feed into the glycolytic pathway (green), (ii) gluconeogenic carbon sources that feed into lower glycolysis or the tricarboxylic acid cycle (purple), and (iii) nucleotide carbon sources (inosine, deoxyadeosine and adenosine, in black). By far the most prominent groups are the glycolytic and gluconeogenic carbon sources, comprising 47 of our 50 carbon sources. The pairwise within-cluster distances of row vectors are significantly lower than the between-cluster distances (Mann-Whitney U-test, $p = 10^{-159}$) for these two clusters.

Supplementary figure 7b shows a heat-map representation of the innovation matrix, with rows and columns organized such that they reflect the clusters we detected. Carbon sources within a cluster favour the utilization of other carbon sources within a cluster, e.g., networks viable on one glycolytic carbon source tend to be viable on other glycolytic carbon sources as well. To go back to our opening example, viability on acetate, a gluconeogenic carbon source, is more likely to entail viability on another gluconeogenic carbon source, such as L-serine, than on N-acetylneuraminate, a glycolytic carbon source.

Pre-adaptation through required viability on two carbon sources is synergistic. In our analysis thus far, we studied samples of random viable networks that we required to be viable on only one carbon source. However, many organisms have to be viable on more than one carbon source in the wild. This raises the question whether the innate capacity for pre-adaptation increases or decreases as one requires viability on multiple carbon sources. For computational feasibility, we restrict ourselves here to analyses of two carbon sources. Specifically, we chose at random 100 pairs of carbon sources, and generated for each carbon source pair (C_1, C_2) 100 metabolic networks required to be viable on both carbon sources. We then asked whether the average innovation index for these networks $I_{(C1,C2)}$ was greater or smaller than the sum of the innovation indices I_{C1} and I_{C2}. To this end, we calculated the quantity $I_{(C1,C2)}$ - I_{C1} - I_{C2}. This quantity would be equal to zero if pre-adaptation was additive whenever viability was required on two carbon sources (C_1, C_2). Supplementary figure 9 indicates that the distribution of $I_{(C1,C2)}$ is displaced to the right of the origin, and significantly different from zero (One sample t-test, $p = 10^{-10}$, $n = 100$). Specifically, for 77 percent of carbon source pairs, the number of additional carbon sources to which pre-adaptation occurs is greater than the sum of the innovation indices I_{C1} and $I_{C2,}$ and for 23 percent pairs it is less. Thus, viability when required on a pair of carbon sources (C_1, C_2) leads to pre-adaptation on more carbon sources than expected from the two carbon sources C_1 and C_2 separately.

We hypothesized that pre-adaptation on a pair of carbon sources would be higher if carbon sources C_1 and C_2 belonged to two different clusters (supplementary figures 7b and 8), because then each source would facilitate pre-adaptation to other carbon sources in its respective cluster. To test this hypothesis, we computed the innovation index $I_{(C1,C2)}$ separately for two groups of carbon source pairs. In the first group, carbon source C_1 belonged to a different cluster than carbon source C_2. In the second group, C_1 and C_2 belonged to the same cluster. The innovation index $\mathrm{I}_{(C1,C2)}$ of carbon source pairs belonging to different clusters (mean $I_{(C1,C2)} = 4.02$) was significantly higher than when C_1 and C_2 belonged to the same clusters (mean $I_{(C1,C2)} = 0.6$; Mann-Whitney U-test, $p = 10^{-8}$). Thus, the capacity of pre-adaptation increases when viability is required on a pair of carbon sources that are biochemically dissimilar.

Environmental generalists may be viable on many more carbon sources than occur in their environments. Environment-generalists such as *E. coli* can sustain life on more than 50 carbon sources[15]. Because viability on one carbon source may entail viability on multiple others, *E. coli* may have experienced selection for viability on substantially fewer than the 50 carbon sources we study. In other words, viability on multiple carbon sources may be an indirect by-product of selection on several other carbon sources. In our next analysis, we asked how many fewer carbon sources are required to allow growth on the majority of the 50 carbon sources studied here. To this end, we first generated a sample of 100 random metabolic networks viable on 10 randomly chosen carbon sources and calculated the average innovation index of these networks. We then repeated this procedure for further samples of 100 networks, requiring viability on an increasing number of carbon sources. Supplementary figure 10 shows the average number of carbon sources on which networks are actually viable (vertical axis, error bars indicate one s.dev.), as a function of the number of carbon sources on which viability is required (horizontal axis). The figure demonstrates that pre-adaptation follows a principle of diminishing returns. Networks need to be, on average, viable on almost 49 randomly chosen carbon sources to show viability on all 50 carbon sources. We restricted ourselves in all our analyses to 50 carbon sources for reasons of computational feasibility, and note that the usefulness of our analysis is limited by this fact. Specifically, networks required to be viable on 49 carbon sources may be viable on many more than the 50 carbon sources we examined. The impression of diminishing returns may results partly from the upper limit we impose on the number of carbon sources.

In contrast to observations about networks required to be viable on multiple carbon sources, our earlier analysis had shown that viability on pairs of carbon sources (C_1, C_2) can entail pre-adaptation on more carbon sources than expected from each of the two carbon sources, especially if (C_1, C_2) are biochemically dissimilar (supplementary figure 9). By extension, it might be possible to choose a modest number (C_1, …, C_n) of carbon sources, such that viability required on each of these carbon sources entails pre-adaptation on a much larger number of carbon sources, e.g., all or most of the 50 carbon sources we study. To identify such groups of carbon sources we pursued the following, heuristic procedure that involves our innovation matrix ***I***. Recall that the entries of this matrix I_{ij} contain the fraction of those random metabolic networks that we required to be viable on carbon source C_i, and that were additionally viable on carbon source C_j. For a pre-specified threshold T, we examined each column C_j of this matrix, to see if the largest entry of the column exceeded T, meaning that a fraction T of networks were also viable on C_j when required to be viable on at least one carbon source C_i. This approach resulted in identifying carbon sources that networks were pre-adapted to (C_{new}) when required to be viable on other carbon sources (C) for a specific threshold T.

We used a threshold $T = 0.75$, meaning that for the sample of networks required to be viable on at least any one carbon source C_i, 75 percent or the majority of its networks were additionally viable on another carbon source C_j. With this approach, we found that requiring viability on 34 specific carbon sources should entail viability on 16 further carbon sources. To validate this hypothesis, we generated 500 random metabolic networks viable on the specific 34 carbon sources (when provided as sole carbon sources). We then computed the mean total number of carbon sources these random networks are viable on. Specifically, we found that random networks were viable on a mean of 49.33 carbon sources (s.dev. = 0.84) when required to be viable on the specific 34 carbon sources ($T = 0.75$, supplementary table 2). This means that selection on these 34 specific carbon sources allows networks to be viable on almost all carbon sources we considered. That is, they are pre-adapted to significantly more carbon sources than 500 random networks viable on a randomly chosen set of 34 carbon sources (mean = 42.23, s.dev. = 1.24) (Mann-Whitney U-test, $p = 10^{-169}$, supplementary table 3). We repeated this procedure with varying thresholds, $T = 0.25$ and $T = 0.5$ and again found pre-adaptation to significantly more carbon sources than 500 random networks viable on a randomly chosen set of carbon sources (supplementary table 3). This analysis shows that metabolic networks are pre-adapted to more carbon sources when required to be viable on a specific set of carbon sources. Thus, environment generalists may have benefitted through similar requirement of viability and growth on a subset of carbon sources, which allowed them to be viable on a repertoire of multiple carbon sources.

Less carbon waste means more pre-adaptation. We next report on an association between a network's biomass yield and its innovation index. We found this association when we divided our original sample of 500 random networks required to be viable on glucose into two groups, according to whether a network's biomass yield lay above or below the mean yield. In this analysis, random metabolic networks with a high biomass yield also showed a significantly higher innovation index (Mann-Whitney U-test, $p = 10^{-5}$). Next we generated 500 random networks with a biomass yield on glucose equal to or exceeding that of the *E. coli* metabolic network[15]. This sample of networks showed a mean innovation index (6.04, s.dev. = 3.7) that was significantly higher (Mann-Whitney U-test, $p = 10^{-8}$) than our original sample of networks ($I_{Glucose}$ = 4.86, s.dev. = 2.83). Thus, networks with higher biomass yield have a higher innovation index.

A high biomass yield may indicate that a network produces less carbon waste. To find out whether this is the case, we calculated the total carbon waste produced by each network in our original sample of random networks viable on glucose (see methods), and found that the biomass yield is indeed negatively correlated with the amount of carbon waste produced by these networks (Spearman's $\rho = -0.89$, $p = 10^{-168}$, $n = 500$). Furthermore, there is a modest yet significant negative correlation between the amount of carbon waste and the innovation index of networks viable on glucose (Spearman's $\rho = -0.28$, $p = 10^{-10}$, $n = 500$).

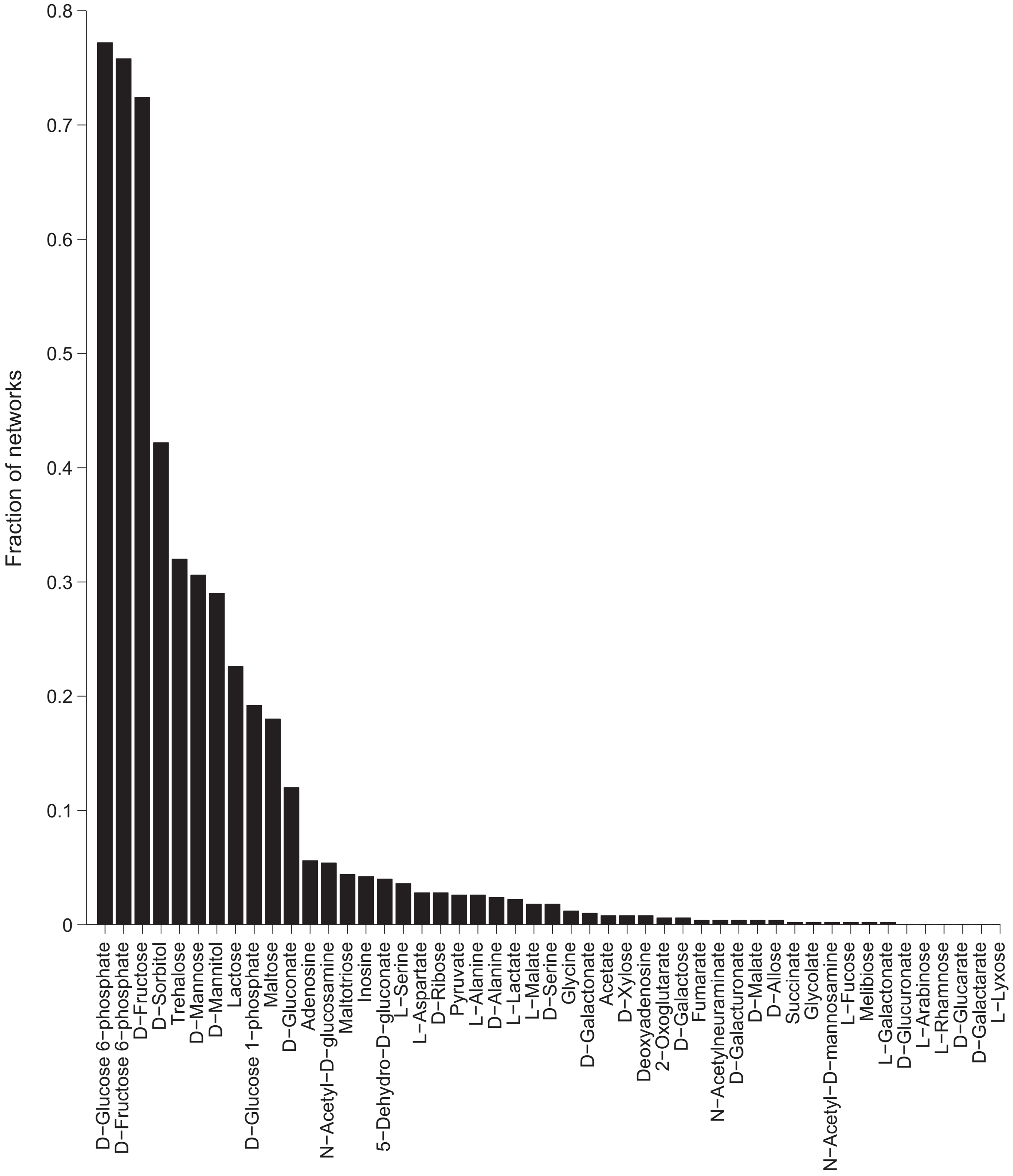
We hypothesized that in the networks producing more carbon waste (and having low biomass yield), one or more additional carbon sources are excreted as carbon waste and cannot be fed into biomass production. Such networks would then not be pre-adapted for viability on these carbon sources. It is relevant here that carbon waste can be secreted in the form of various metabolites, such as carbon dioxide, acetate, and fumarate, to name a few. For example, a total of 62 carbon-containing metabolites are secreted as waste by at least one network in our sample of 500 networks viable on glucose. We tested this hypothesis in the following way. For each metabolite, and for each of our two samples (high and low biomass yield), we counted the number of networks in which the metabolite is secreted as waste. Supplementary table 4 shows for each potentially secreted metabolite, the number of high and low yield networks that secrete it. Significantly fewer high-yield networks secreted carbon-containing metabolites, when we considered all these metabolites together as a group (Mann-Whitney U-test, $p = 0.0081$, $n = 62$). For 92 percent (57 of 62) metabolites, the number of low yield networks secreting the metabolite is higher than the number of high yield networks (supplementary table 4). Furthermore, 12 of these 62 metabolites are used as carbon sources as well (supplementary table 4, shown in red). The association we find is particularly important for metabolites that can also serve as carbon sources or are linked to carbons sources. For example, acetate, which is also one of our 50 carbon sources, is secreted as waste by 173 low yield metabolic networks, but only by 122 high yield networks. Other examples of carbon sources that are excreted by a greater number of low-yield networks include 5-dehydro-D-gluconate, D-gluconate, fumarate, glycolate, succinate, pyruvate.

We next asked whether the innovation index correlates with biomass yield per carbon not just for glucose, but for all other carbon sources as well. We define the biomass yield per carbon as the ratio of the biomass yield of a metabolic network to the number of carbons in a particular carbon source. Biomass yield needs to be defined in this manner for this analysis, because different carbon sources contain different numbers of carbon molecules. Supplementary figure 11 shows that the average innovation index on a carbon source C, and the mean biomass yield per carbon show a strong positive correlation (Spearman's $\rho = 0.47$, $p = 0.00057$, $n = 50$). As we mentioned above, a high biomass yield per carbon reflects the efficient conversion of the carbon source into biomass precursors with little waste. Thus, what holds for glucose also holds for other carbon sources.

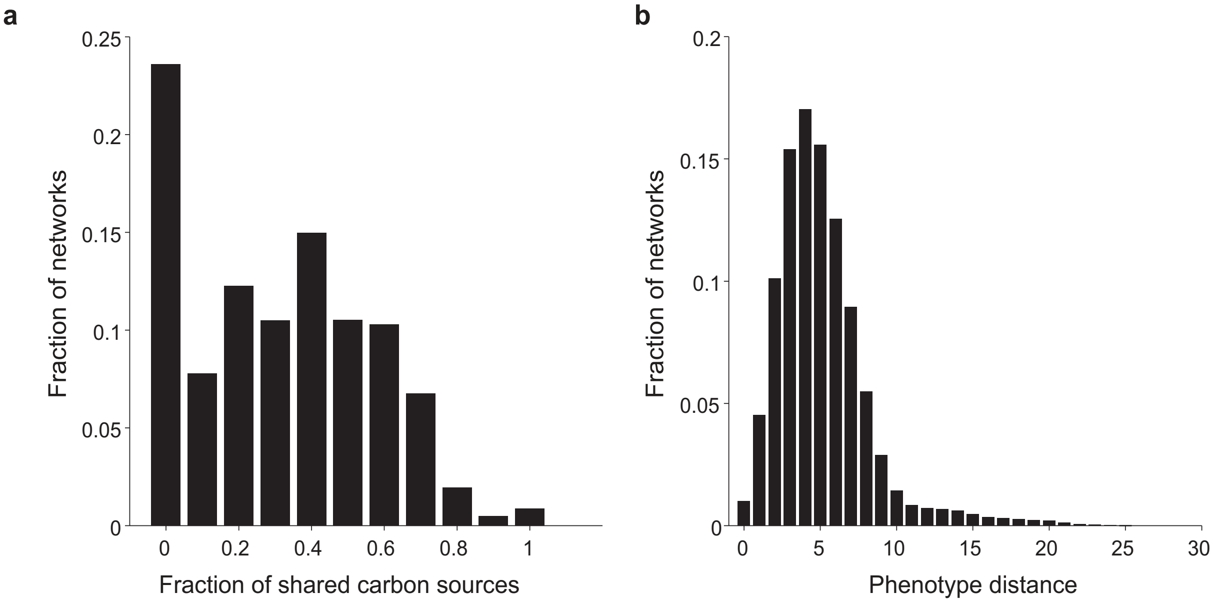
In sum, a network that converts carbon sources efficiently into biomass tends to have a high innovation index. It tends to be pre-adapted to a larger number of carbon sources. The reason is that the waste products of inefficient metabolic networks include carbon sources. These carbon sources are not utilized by the inefficient network, but can be utilized by an efficient network.

A sample size of 500 networks is sufficient for our analysis. Most of our analysis presented here used 500 random networks viable on a single carbon source. While a sample size of 500 networks has proven to be sufficient for understanding the essentiality of reactions[30], this might not be the case for our present analysis. To find out, we sampled ten times as many, i.e., 5000 networks for each of the following 10 carbon sources: pyruvate, acetate, D-glucose, L-aspartate, L-serine, adenosine, N-acetylneuraminate, trehalose, maltotriose and L-galactonate. These carbon sources have varying innovation indices (main text, figure 2a), ranging from the highest to the lowest values we observed. We then computed the distribution of the innovation index I_C for each of these ten C carbon sources. Supplementary figure 12 shows the mean of this distribution (bars) and its coefficient of variation (vertical lines), that is, the ratio of the standard deviation to the mean. Black indicates values for the sample of 5000 networks, while grey indicates values for the original sample of 500 networks. Note that the means are very similar for the two samples of different size. For each carbon source, we also computed the fraction of networks viable on each of the other 49 carbon sources (identical to the innovation matrix explained in the supplementary results, supplementary figure 7a). We then computed the statistical association between the entries of this matrix for the resampled and the original ensemble of random networks for each carbon source. The association is high and significantly different from zero (Spearman's $\rho \geq 0.7$, $p \leq 10^{-7}$, $n = 50$ for all 10 carbon sources. These observations suggest that a sample size of 500 random networks is sufficient for the analyses conducted here.

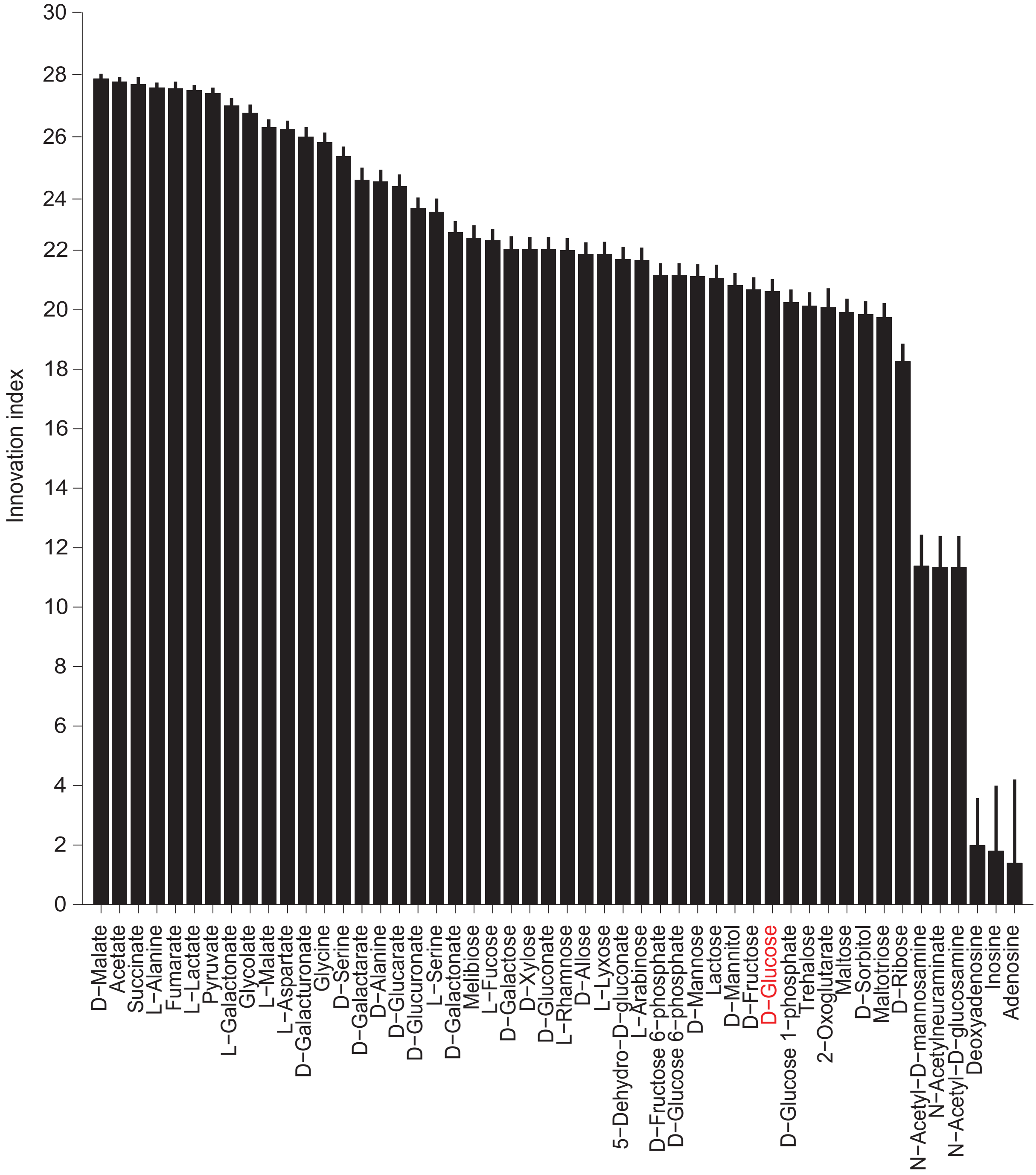
Supplementary Figures



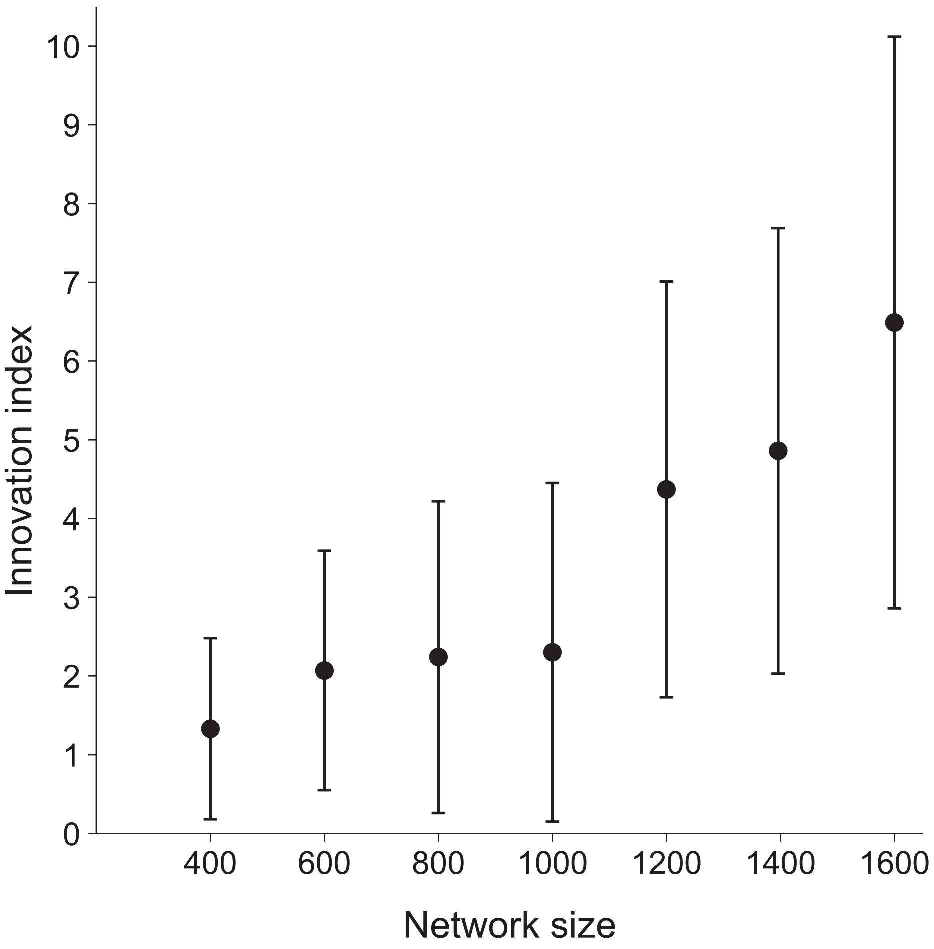
Supplementary Figure 1 – **Different carbon sources differ greatly in their propensity for exaptation.** The horizontal axis lists 49 different carbon sources. The vertical axis indicates the fraction of random networks viable on each carbon source (when required to be viable on glucose). Carbon sources are ranked according to the value on the vertical axes. While more than 70 percent of networks are viable on glucose-6-phosphate, fructose-6-phosphate and fructose as additional carbon sources (left-most three bars), most carbon sources allow viability of only a small fraction of sampled networks. Data is based on 500 random networks viable on glucose.



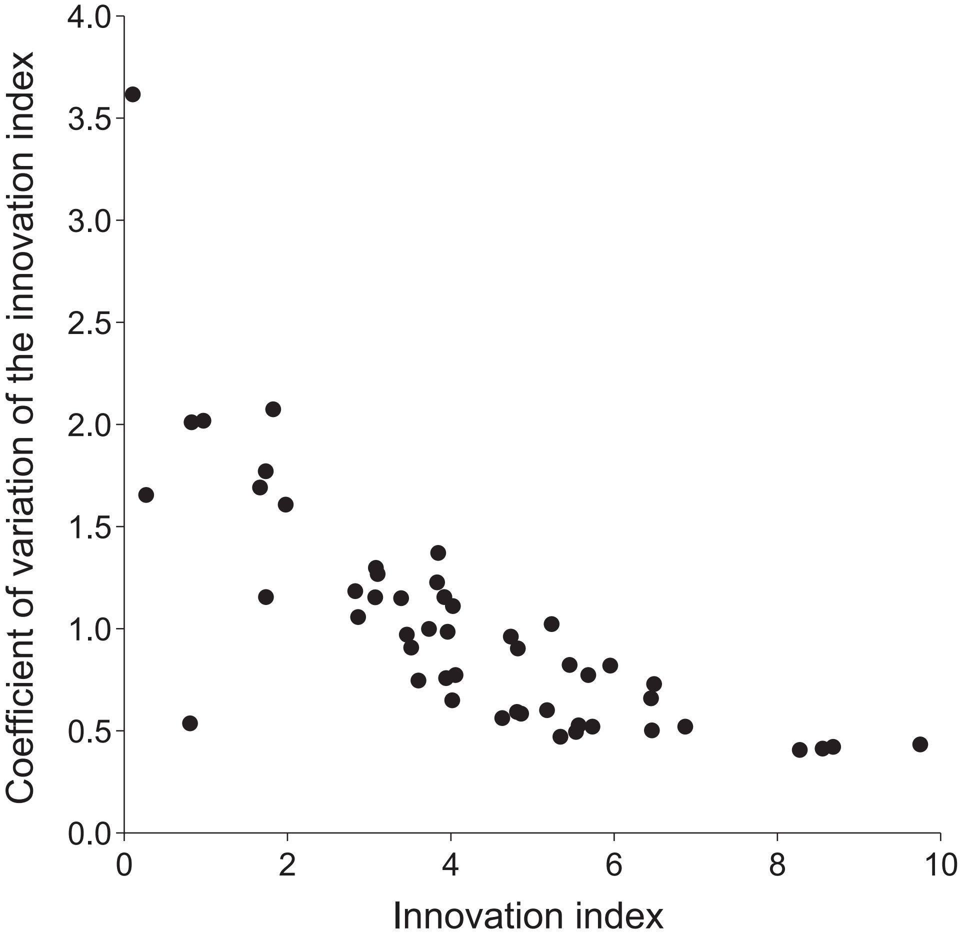
Supplementary Figure 2 – **Majority of the pre-adapted carbon sources differ among networks.** (a) The distribution of the number of shared carbon sources in the innovation vector of networks pairs. 23 percent of network pairs do not share any carbon source that they are viable on (except glucose). On average, 31.8 percent of carbon sources are shared between a pair of networks. (b) The distribution of the phenotypic distance between network pairs, as computed by the Hamming distance[46] between their innovation vectors. The Hamming distance increases by one for each entry in which two binary vectors differ. The distance thus indicates the number of carbon sources (aside from glucose) that one but not the other network pair is viable on. On average, two networks differ in their viability on 5 carbon sources. Data in (a)-(b) are based on innovation vectors of 500 random networks required to be viable only on glucose.



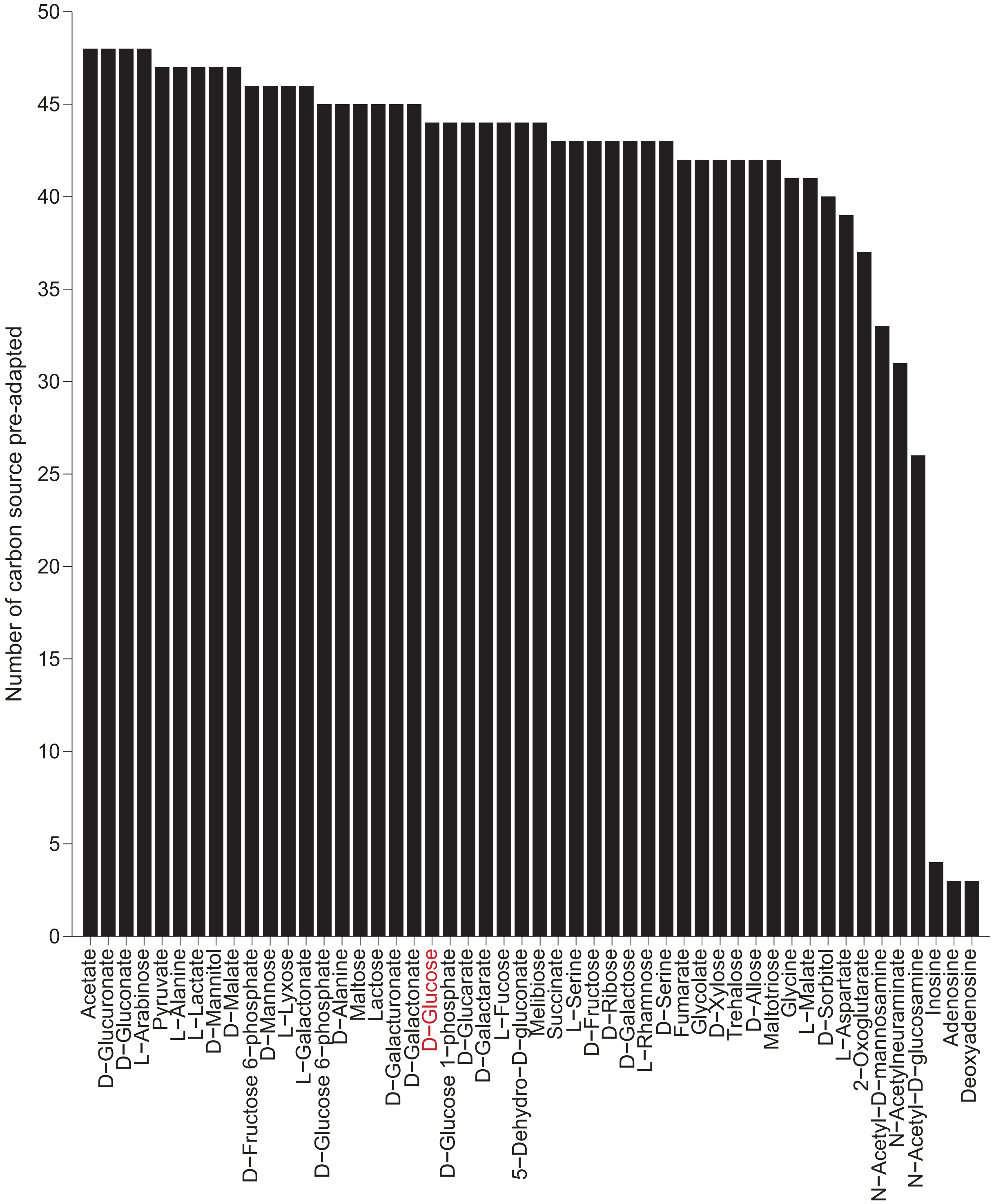
Supplementary Figure 3 – **Innovation potential increases with network connectedness.** For each of 50 carbon sources *C* (horizontal axis), the figure indicates the mean innovation index (bar) and its coefficient of variation (lines) for 500 random networks required to be viable on carbon source *C*. Each sample was generated through a sampling process similar to our MCMC sampling, except that we only allowed a reaction to be added to a network, if the reaction was connected to the network through its substrates or products. Considering only such connected networks increases the potential for exaptation. For example, the innovation index of glucose (in red) is much higher than in the original sample of networks (figure 2a).



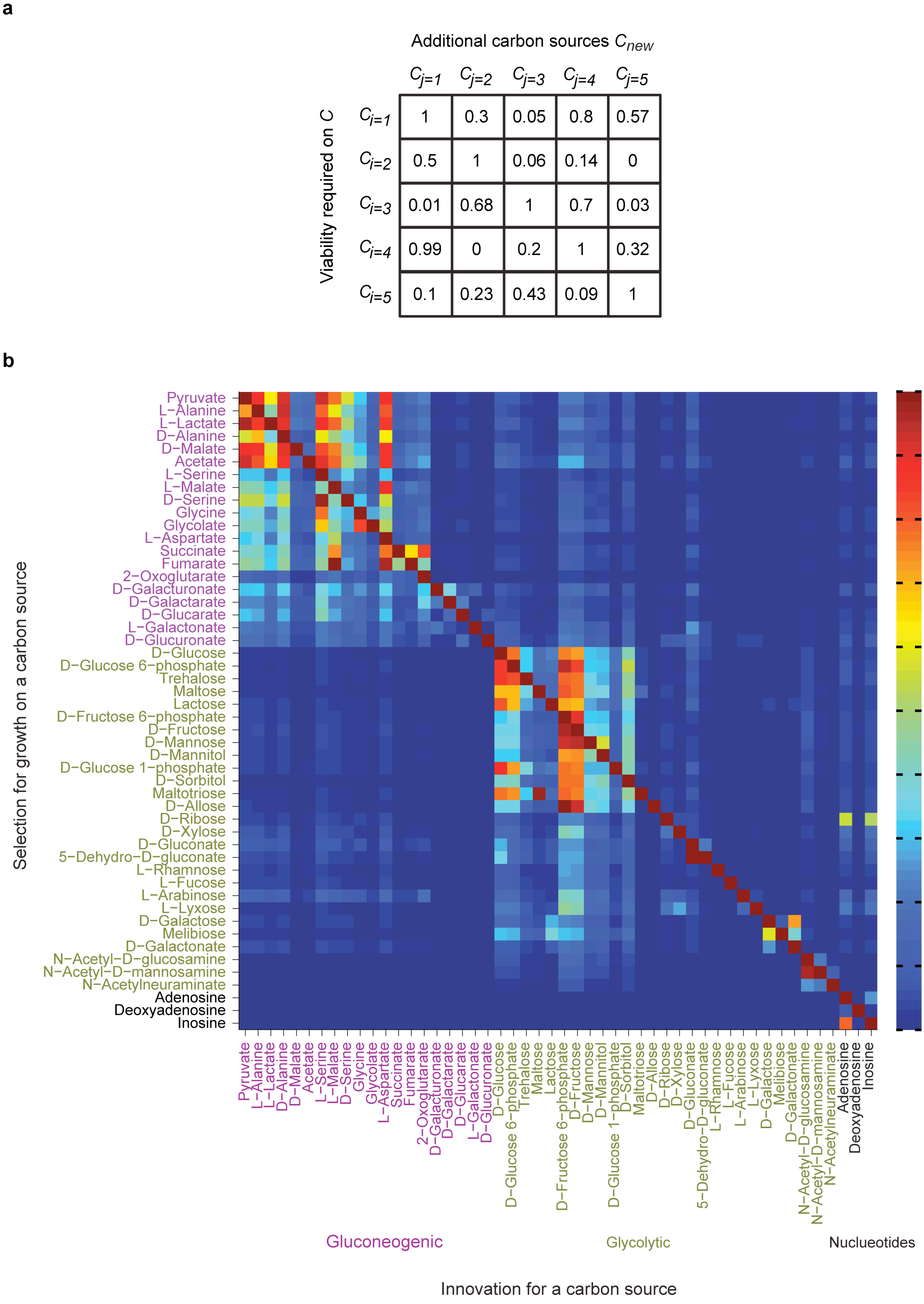
Supplementary Figure 4 – **The potential for innovation increases with network complexity.** The horizontal axis shows network size (network complexity) in numbers of reactions. A size of 1400 reactions corresponds approximately to the size of the *E. coli* metabolic network (1397 reactions[15]) used in our other analyses. The vertical axis shows the mean innovation index of networks with a given size. Higher network complexity allows viability on a larger number of additional carbon sources. Data for each network size class are based on samples of 500 random networks viable on glucose, i.e. a total of 3500 networks.



Supplementary Figure 5 – **For carbon sources with a high innovation index, this index is less variable.** The horizontal axis indicates the mean innovation index, and the vertical axis indicates the coefficient of variation of this index. The data is the same as for figure 2a, i.e., each data point is based on a sample of 500 random networks required to be viable on one of 50 carbon sources (*n* = 25000). Note that the coefficient of variation decreases with increasing innovation index.



Supplementary Figure 6 – **Pre-adaptation occurs for the vast majority of carbon sources.** For each of 50 carbon sources *C* (horizontal axis) the height of the vertical bar above each carbon source indicates the total number of the other 49 carbon sources on which at least one network in the sample is viable. For instance, acetate allows viability (pre-adaptation) on 48 other carbon sources, while deoxyadeosine and adenosine allow viability on only 3 other carbon sources. Data in the figure are based on samples of 500 random viable networks for each carbon source, i.e., on a total of 500x50=25,000 sampled networks.



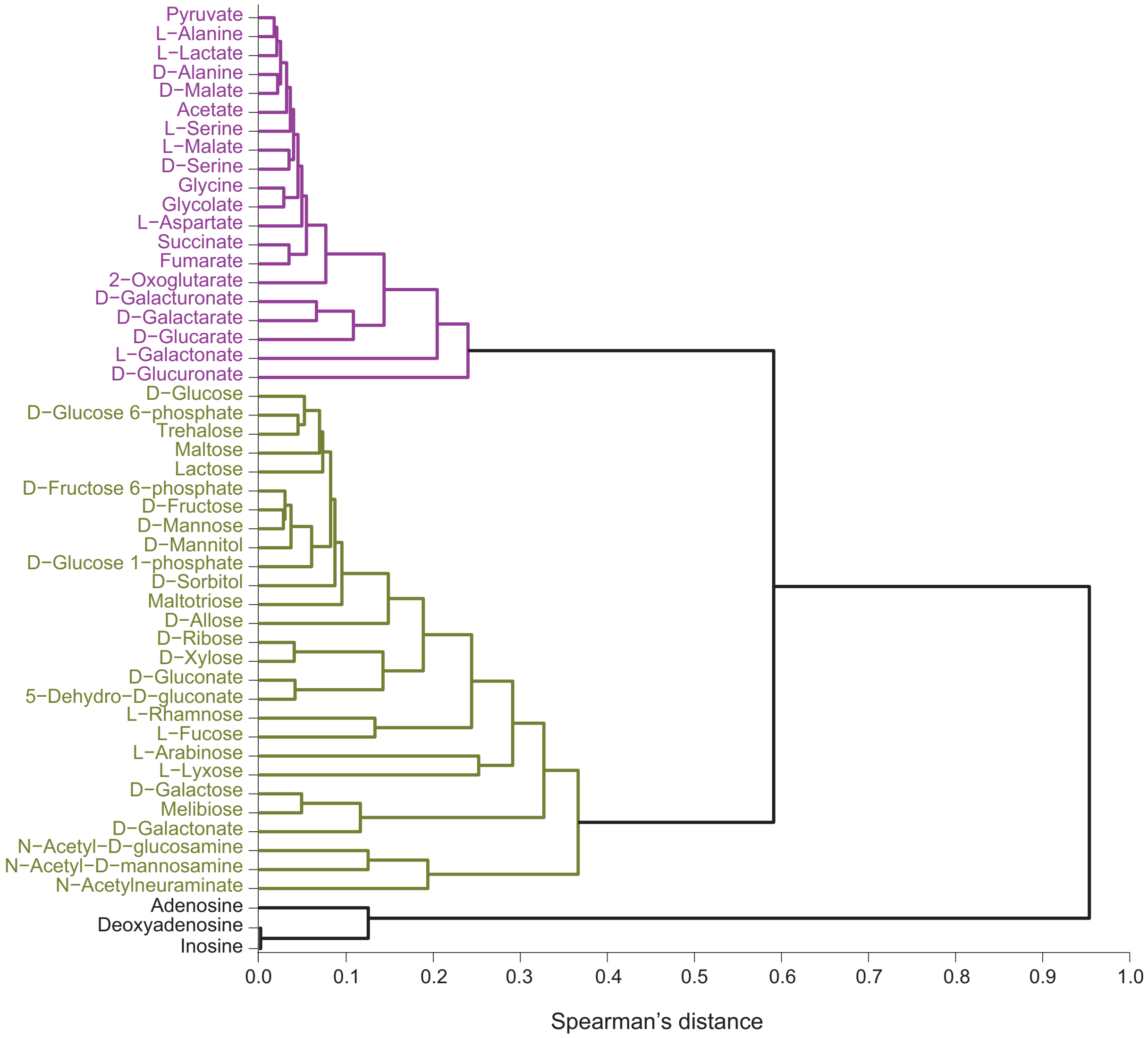
a

Additional carbon sources C_{new}

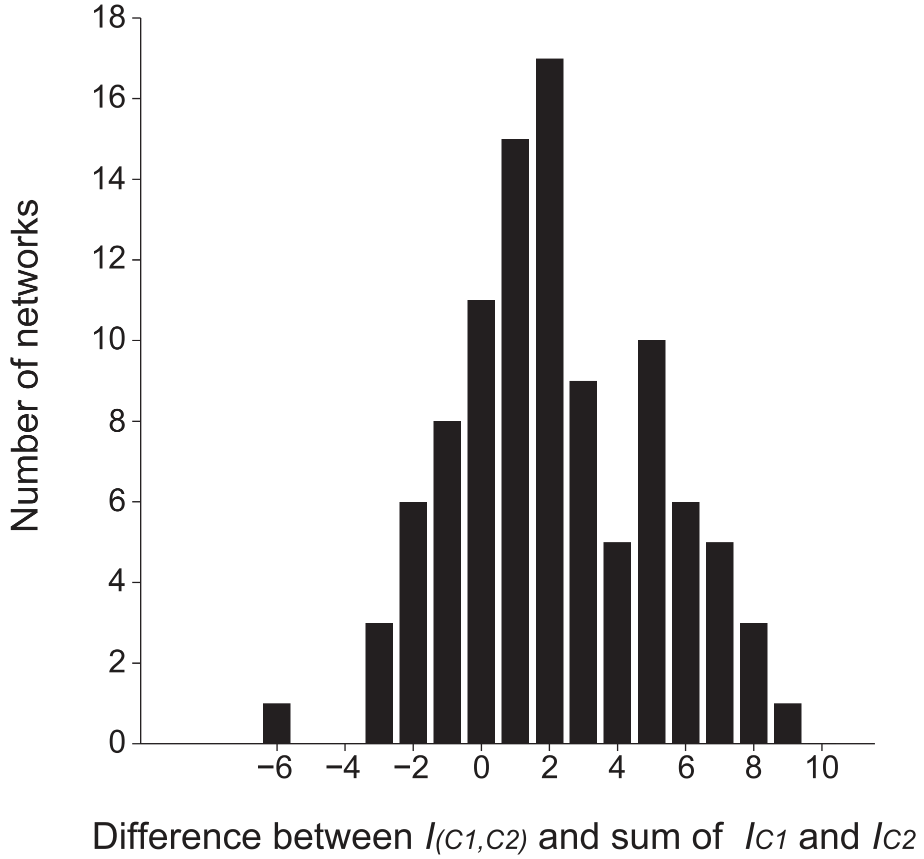
Viability required on C

	$C_{j=1}$	$C_{j=2}$	$C_{j=3}$	$C_{j=4}$	$C_{j=5}$
$C_{i=1}$	1	0.3	0.05	0.8	0.57
$C_{i=2}$	0.5	1	0.06	0.14	0
$C_{i=3}$	0.01	0.68	1	0.7	0.03
$C_{i=4}$	0.99	0	0.2	1	0.32
$C_{i=5}$	0.1	0.23	0.43	0.09	1

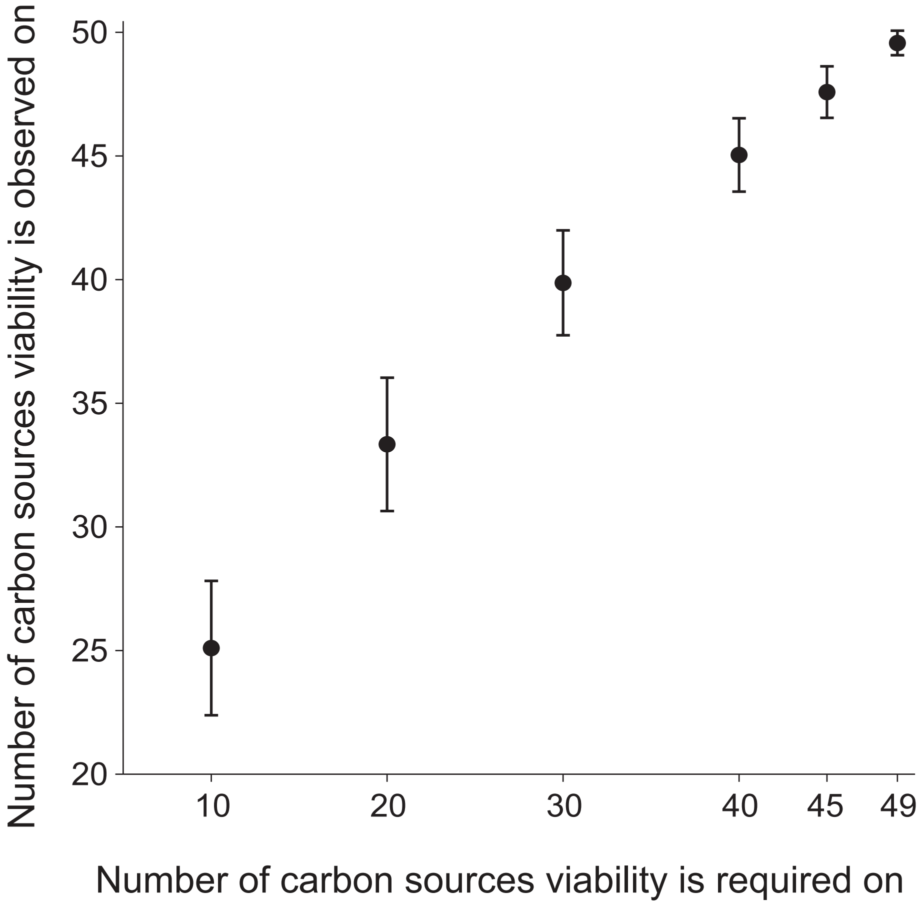
Supplementary Figure 7 – **Innovation occurs preferentially within clusters of related carbon sources.** (a) A hypothetical innovation matrix comprising 5 carbon sources. Each row vector corresponds to the carbon source C_i on which viability is required, and each column vector correspond to the additional carbon source C_j. Each matrix entry indicates the fraction of networks that are also viable on C_j while required to be viable on C_i. (b) The figure shows a heat-map of the innovation matrix, organized according to different groups of carbon sources. The purple metabolite lettering corresponds to gluconeogenic carbon sources, green lettering corresponds to glycolytic carbon sources, and black corresponds to nucleotides as carbon sources. The two extreme ends of the colour spectrum of the heat map are blue and red, where blue (red) indicates that none (all) random networks required to be viable on carbon source C_i (rows) are also viable on an additional carbon source C_j (columns). The figure shows that carbon sources within a cluster favour the utilization of other carbon sources within the same cluster. Data in figures (a)-(b) are based on 50 samples of 500 random viable networks, where networks in each sample were required to be viable on a different source of 50 different carbon sources.



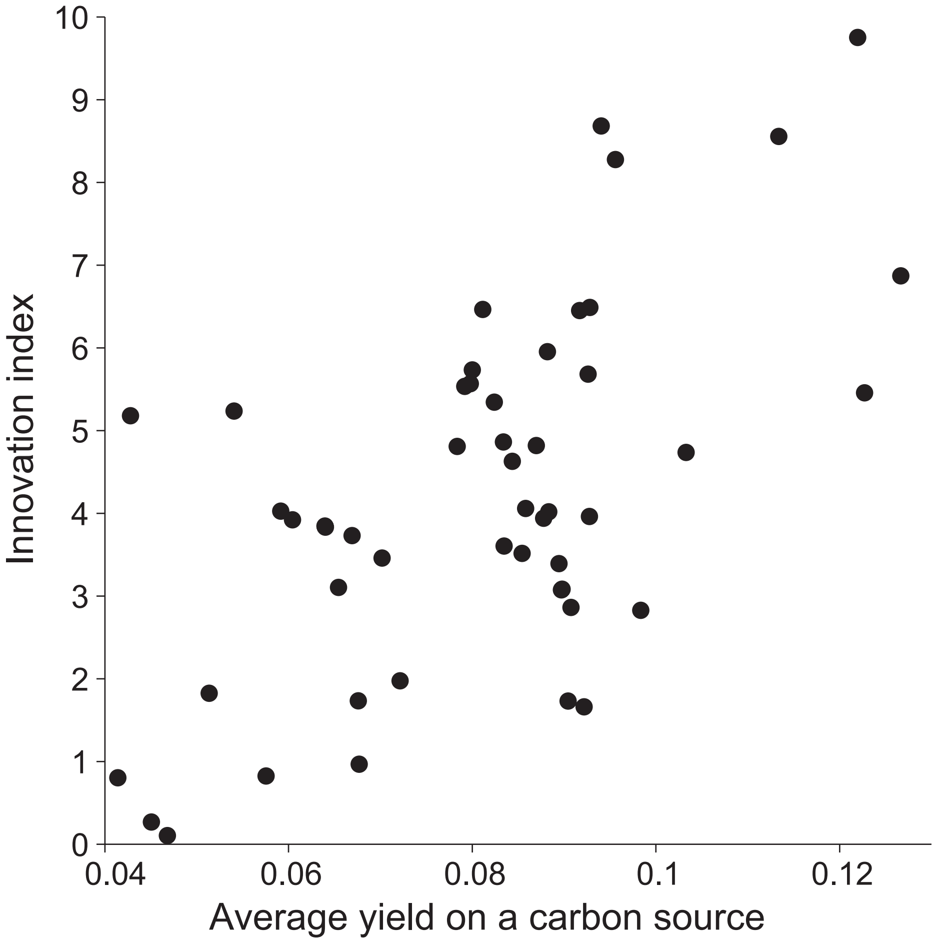
Supplementary Figure 8 – **Innovation occurs preferentially within clusters of related carbon sources.** The dendrogram shows three distinct groups of carbon sources based on hierarchical clustering of the innovation matrix, using the Spearman's rank correlation distance (horizontal axis, see methods). The green, purple, and black groups of metabolites correspond to glycolytic, gluconeogenic, and nucleotide carbon sources. Note that the Spearman's distance between any two clusters of carbon sources is larger than 0.6. Data are based on 50 samples of 500 random viable networks, where networks in each sample were required to be viable on a different one of 50 different carbon sources.



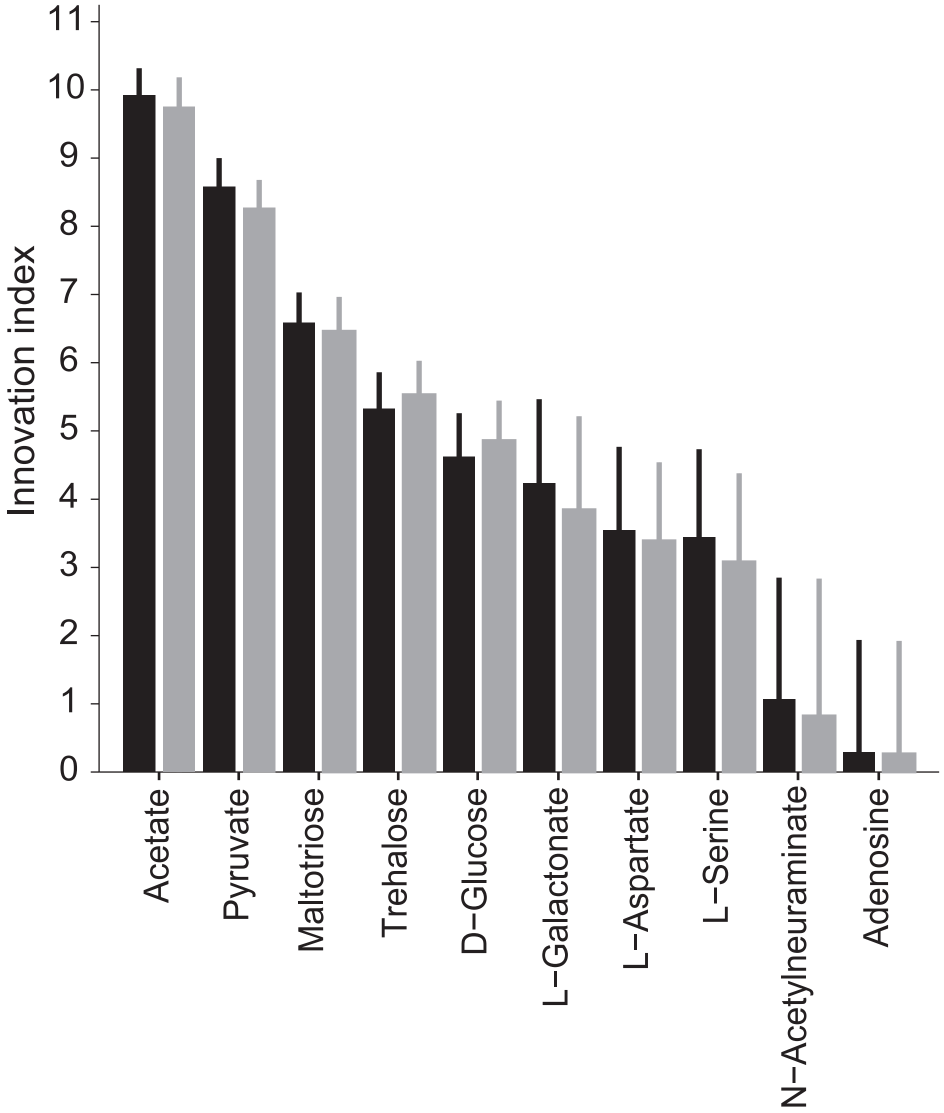
Supplementary Figure 9 – **Pre-adaptation through required viability on two carbon sources is synergistic.** The figure shows the distribution of the quantity $I_{(C1,C2)}$ - I_{C1} - I_{C2}, averaged over 100 random metabolic networks viable on a pair of carbon sources C_1 and C_2 (horizontal axis). This quantity describes whether the innovation index of a pair of carbon sources ($I_{(C1,C2)}$) is higher or lower than the sum of the individual innovation indices I_{C1} and I_{C2}. A majority of pairs (77 percent) have a synergistic mean innovation index ($I_{(C1,C2)} > (I_{C1} + I_{C2})$), while the remaining pairs have an antagonistic innovation index ($I_{(C1,C2)} < (I_{C1} + I_{C2})$). Data are based on innovation vectors of 100 random networks viable on a pair of carbon sources (C_1, C_2), computed for 100 randomly chosen pairs of 50 carbon sources.



Supplementary Figure 10 – **Diminishing returns in pre-adaptation.** The vertical axis indicates the mean number of carbon sources on which viability is observed, for networks required to be viable on the number of randomly chosen carbon sources shown on the horizontal axis. For each value on the horizontal axis, data is based on specific samples of carbon sources, and on samples of 100 networks for each sample of carbon sources. Error bars denote one standard deviation. Note that networks are required to be viable on 49 carbon sources to allow viability on all 50 carbon sources studied here.



Supplementary Figure 11 – **Innovation potential rises with reduced waste production.** Each data point corresponds to one of 50 carbon sources. The horizontal axis indicates the average biomass yield per mole of carbon for the carbon source. The vertical axis indicates the average innovation index of the carbon source. Carbon sources that are efficiently metabolized (and produce low carbon waste) have a high yield. The figure shows that such high-yield carbon sources also allow viability on a greater number of additional carbon sources. For each carbon source, data are based on samples of 500 random networks viable on the carbon source (*n* = 25000).



Supplementary Figure 12 – **A sample size of 500 networks is sufficient for our analysis.** For each of 10 carbon sources *C* (horizontal axis), the figure indicates the mean innovation index (bar) and its coefficient of variation (lines) for 5000 random networks (black bars) and 500 random networks (gray bars) required to be viable on carbon source *C*. Note the broad distribution of the index. The height of the solid lines indicates the coefficient of variation. Note that the pairs of black and gray bars have similar height.

Supplementary Tables

Supplementary table 1 – **The 50 carbon sources used in this study**

5-Dehydro-D-gluconate	D-Fructose 6-phosphate
D-Glucarate	D-Mannose
Acetate	Melibiose
D-Glucuronate	D-Fructose
N-Acetyl-D-glucosamine	D-Mannitol
N-Acetyl-D-mannosamine	L-Fucose
Glycine	Pyruvate
N-Acetylneuraminate	Fumarate
Glycolate	D-Ribose
Adenosine	D-Glucose 1-phosphate
Inosine	L-Rhamnose
2-Oxoglutarate	D-Glucose 6-phosphate
L-Lactate	D-Sorbitol
D-Alanine	D-Galactose
Lactose	D-Serine
L-Alanine	D-Galactarate
L-Lyxose	L-Serine
D-Allose	D-Galactonate
D-Malate	Succinate
L-Arabinose	L-Galactonate
L-Malate	D-Galacturonate
L-Aspartate	Trehalose
Maltose	D-Glucose
Deoxyadenosine	D-Gluconate
Maltotriose	D-Xylose

Supplementary table 2 – **Thirty-four specific carbon sources that allow growth on all 50 carbon sources**

Acetate	D-Malate
Succinate	L-Rhamnose
D-Glucose 1-phosphate	Deoxyadenosine
D-Ribose	N-Acetyl-D-mannosamine
Fumarate	D-Serine
D-Galactose	D-Sorbitol
D-Mannose	D-Glucarate
Glycolate	D-Galactarate
D-Xylose	D-Galactonate
L-Lactate	L-Fucose
D-Glucuronate	5-Dehydro-D-gluconate
Lactose	Trehalose
L-Arabinose	D-Allose
N-Acetylneuraminate	L-Lyxose
Inosine	Maltotriose
D-Galacturonate	Melibiose
D-Mannitol	L-Galactonate

Supplementary table 3 – **Selection on specific sets of carbon sources allows networks to be viable on more carbon sources**

Threshold (*T*)	Number of specific carbon sources viability is required on	Number of carbon sources networks are viable on for specific sets of carbon sources	Number of carbon sources networks are viable on for random sets of carbon sources	*p*-value (*n* = 500)
0.25	19	39.88 ± 2.72	32.64 ± 2.8	10^{-144}
0.50	26	45.86 ± 1.64	37.45 ± 2.31	10^{-165}
0.75	34	49.33 ± 0.84	42.23 ± 1.24	10^{-169}

Selection on specific carbon sources allows networks to be viable on more carbon sources. The first column denotes the threshold *T* denoting the fraction of networks that were also viable on C_j when required to be viable on at least one carbon source C_i. The second column denotes the number of specific carbon sources that networks are required to be viable on and the third column denotes the average (one s.dev.) number of carbon sources such networks show viability on. The fourth column denotes the average (one s.dev.) number of carbon sources networks are viable on when required to be viable on the same number of random carbon sources as denoted in the second column. The last column shows that networks viable on a specific set of carbon sources can be pre-adapted to significantly more carbon sources than networks viable on a random set of carbon sources.

Supplementary table 4 – **Metabolites secreted as waste, the number of low biomass yield and high biomass yield networks that secretes each metabolite**

Metabolite name	**Number of low yield networks**	**Number of high yield networks**
(R)-Propane-1,2-diol	2	0
5-Dehydro-D-gluconate	20	1
4-aminobutyrate	3	0
Acetoacetate	10	6
Acetaldehyde	25	16
Acetate	173	122
Adenine	41	32
Adenosine	29	25
Alpha-ketoglutarate	6	3
Allantoin	3	0
L-arabinose	1	0
L-arginine	1	0
L-asparagine	1	0
Carbon dioxide	230	186
L-cysteine	4	3
Cytidine	69	56
D-Lactate	3	2
D-Alanine	6	0
2-dehydro-3-deoxy-D-gluconate	3	1
Dihydroacetone	101	41
Ethanolamine	2	0
Ethanol	6	3
Formaldehyde	38	26
Formate	112	60
Fumarate	78	45
sn-Glycero-3-phosphoethanolamine	4	7
Gycerophoglycerol	1	0
D-Gluconate	9	3
Glyceraldehyde	50	12
Glycerol-3-phosphate	6	0
Glycerate	35	21
Glycolate	135	91
Glycerol	25	3
Guanine	3	2
Histidine	71	33

Hypoxanthine	17	15
L-isoleucine	4	1
Indole	89	18
Inositol	21	15
L-Lactate	4	1
L-Leucine	6	1
Ornithine	58	58
Phenethylacetaldehyde	25	19
Phenylalanine	23	11
3-phenylpropionate	1	2
Putrescine	13	5
Pyruvate	10	2
L-Serine	4	0
Succinate	39	19
Tartrate	1	0
Thymidine	39	17
L-Threonine	6	5
Thymine	3	0
L-Tryptophan	16	7
L-Tyrosine	33	43
Uracil	40	22
Urea	8	7
Uridine	21	14
L-valine	1	0
Xanthine	23	10
Xanthosine	8	4

Secreted metabolites in low and high biomass yield networks. The second and the third columns denote the number of low and high biomass yield networks that secrete carbon waste in the form of specific metabolites (first column) respectively. Most metabolites are secreted in a higher number of low biomass yield networks. Many of the secreted metabolites are also among the carbon sources we consider, shown in red. Data are based on samples of 500 random networks required to be viable on glucose.

Additional References

41. Pál, C., Papp, B. & Lercher, M. J. Horizontal gene transfer depends on gene content of the host. *Bioinformatics (Oxford, England)* **21 Suppl 2**, ii222–3 (2005).

42. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by Fast Computing Machines. *The Journal of Chemical Physics* **21**, 1087 (1953).

43. Wovcha, M. G., Steuerwald, D. L. & Brooks, K. E. Amplification of D-xylose and D-glucose isomerase activities in Escherichia coli by gene cloning. *Appl. Envir. Microbiol.* **45**, 1402–1404 (1983).

44. Elias, M. D. *et al.* Occurrence of a bound ubiquinone and its function in Escherichia coli membrane-bound quinoprotein glucose dehydrogenase. *The Journal of biological chemistry* **279**, 3078–83 (2004).

45. Vitkup, D., Kharchenko, P. & Wagner, A. Influence of metabolic network structure and function on enzyme evolution. *Genome biology* **7**, R39, doi:10.1186/gb-2006-7-5-r39 (2006).

46. Hamming, R. W. Error detecting and error correcting codes. *Bell System Technical Journal* **29**, 147–160 (1950).