Introduction

Predicting microbial metabolism under a broad range of conditions would enable us to leverage microbes for applications in critical areas such as energy production [1], pollution amelioration [2,3], bioengineering [4], physiology or medicine [5,6] to name a few. While systematic in vivo growth experiments could in principle fill the gaps in our current knowledge, those experiments are time consuming and contingent on the ability to grow the microbial species of interest in the laboratory [7]. As a consequence, only a small number of microbes have been studied using these techniques. For example, Biolog (http://www.biolog.com) provides Phenotype Microarrays that have been used on species such as Escherichia coli [8] and Bacillus subtilis [9] but to date there are only 200 publications listed in Biolog.

To circumvent experimental limitations, a number of mathematical models have been developed aiming to predict microbial growth rates [10–14]. However, these models are only valid for a limited number of specific nutrients and are not easily generalizable because of the need to determine many parameters empirically. Indeed, developing such a theory seems an insurmountable challenge given the combinatorial number of possible growth media and the large number of unknown parameters such as reaction constants and enzyme affinities that control metabolic reactions [15–17]. For instance, the in silico reconstruction of E. coli contains more than 2,000 reactions [8]; taking into account that each reaction has at least two and up to tens of kinetic parameters [18], a detailed kinetic model would have on the order of 10,000 parameters, unless some approximations valid under certain conditions are made to reduce this number [19].

An alternative approach that is gaining in popularity is the development of thermodynamically-feasible metabolic reconstructions that can be used to predict the growth of individual organisms using constraint-based methods [3,9,20–27]. Researchers fine-tune these reconstructions to match the conditions of specific growth rate experiments, such as nutrient availability and ATP maintenance. These reconstructions, built by literature mining, can, under certain conditions, accurately predict the impact of individual nutrients on growth as sources of carbon, nitrogen, phosphorus and sulfur, thus allowing researchers to evaluate the demands of biomass production and to investigate how individual nutrients can meet that demand. While building metabolic reconstructions for new organisms is quite challenging, nowadays the process is becoming more and more automated.
Background

Our phenomenological model expresses the impact that different carbon sources (or nutrients) have on a microbe’s ability to grow using only information on the chemical structure of these carbon sources and on the ability of a microbe to catabolize these nutrients. Our model is built to reproduce the predictions of flux balance analysis calculations (FBA) on metabolic reconstructions, thus it will suffer from the same limitations. Indeed, while FBA is a powerful tool to investigate microbial metabolism and microbial growth in particular, it has a number of limitations when estimating growth rates and the effect of media and environmental conditions on growth.

It is well-known that microbes need a minimal medium and a carbon source in order to grow. Minimal media have been described for a number of species and contain essential chemical species without which the species would not be able to grow [20,29]. The growth rate of a microbe, however, depends on many other factors including the uptake rate of nutrients, temperature, regulation, the availability of oxygen, etc.

FBA is a linear optimization method that predicts the maximal conversion of a set of carbon sources into biomass with a fixed minimal medium. In order for FBA conversion rates to reproduce empirical growth rates, one needs to consider an additional ATP maintenance flux which is obtained by fitting FBA results to empirical growth rates obtained for a certain temperature, minimal medium, and carbon source uptake. While ATP maintenance rates obtained for a specific minimal medium have been shown to give accurate predictions of growth rates in different minimal media for some organisms [9], in principle one cannot assert that they are valid for predicting growth rates for arbitrary minimal media.

Additionally, because metabolic reconstructions do not consider regulatory constraints, FBA will predict that a microbe is capable of uptaking two different sugars simultaneously, while it is well-known that if there is more than one sugar carbon source at high enough concentrations, the organism will exhaust the preferred one before consuming the others [30]. A microbe will, however, consume multiple sources of carbon other than sugars simultaneously and as a consequence grow faster [31]. It has been shown that steric constraints can already reproduce diauxic growth in some organisms in the presence of multiple sugars [32], however, there is no general framework that is able to deal with this issue when using FBA.

To develop a phenomenological model that is able to reproduce maximal biomass conversion rates per carbon source under aerobic conditions, we investigate the maximum amount of biomass that can be produced by an organism in the presence of a minimal medium, oxygen and one or more carbon sources including at most one sugar. To this end, we run FBA on metabolic reconstructions in which we remove ATP maintenance [27] (see Methods for specific details). We concentrate on a training set of four species for which there are high-quality metabolic reconstructions available, and which cover a wide range of microbial phylogeny: a gram-negative bacterium (E. coli [9]), a gram-positive bacterium (B. subtilis [9]), an eukaryote (Saccharomyces cerevisiae [21]), and an archaeon (Methanosarcina barkeri [24]). We validate our model on a test set of three species for which we also have high-quality metabolic reconstructions—Helicobacter pylori, a gram-negative bacterium [20], Staphylococcus aureus, a gram-positive bacterium [22], and Mycobacterium tuberculosis, an acid-fast gram-positive bacterium [23].

The rationale for choosing a small number of species for model building and validation is that lower quality reconstructions are likely to have significant gaps that could incorrectly bias the determination of the model. Additionally, the aggregate set of nutrients available for these reconstructions is of 352 nutrients, which cover all nutrient types and 90 out of 97 pathways available in KEGG [33–35], thus ensuring that our model is comprehensive.

We believe that the development of a systems-level model of microbial metabolism is not only complementary to current approaches but offers some advantages with respect to them. Specifically, while FBA run on metabolic reconstructions already has the capability of predicting maximum biomass yield, our model has the advantage that since it does not consider microscopic details of the metabolism of a species it is directly applicable to any other organism growing under the conditions of our analysis. In fact, we show that solving over a thousand linear equations under constraints can be well approximated by a simpler model whose principles are easier to understand. As such, our model offers the possibility of uncovering universal features of the metabolism of organisms that other computational approaches are not capable of. The mathematical model we develop is thus a valuable tool from both fundamental and applied perspectives, since it can help understand the metabolism of organisms for

Author Summary

The ability of microbial species to consume compounds found in the environment to generate commercially valuable products has long been exploited by humanity. The vast untapped diversity of microbial species offers a wealth of potential resources. However, little is known about most microbial species. While the metabolic network of an organism can be studied to find its nutritional requirements, we lack a reliable metabolic reconstruction for most species. We use in silico organisms to systematically explore whether an arbitrary nutrient can stimulate growth as a single source of carbon, and how effectively it can be used by the organism. We find that we can predict whether a nutrient is a source of carbon and the biomass yield of that nutrient with a simple model that transcends the diversity of species and their environments. Our model for catabolic potential can therefore be used as a baseline model for any microbe for which we lack a metabolic reconstruction.
which a metabolic reconstruction is not available, or guide the process of validating metabolic reconstructions.

Results

In order to develop a mathematical model that relates nutrient (carbon source) uptake in complex media to biomass production, we need to address three different questions (see Fig. 1). The first question we need to answer is whether a nutrient can be a source of carbon or not. This is to say, first we need to develop a mathematical model that determines whether a nutrient can produce growth or not based on the information available for that nutrient and organism. Note that because we are interested in a binary output (growth/no growth), our model will depend exclusively on the composition of the nutrient and the pathways that can catabolize this nutrient.

Then, we need to assess what is the maximum amount of biomass a nutrient can produce when acting as the sole source of carbon. Our specific aim will be to find a mathematical relationship between nutrient composition (in our case, carbon content) and the biomass produced per unit of nutrient uptake.

In third place, we need to assess how maximal biomass production is affected when several nutrients are present in a medium. That is to say, we need to determine the relationship between biomass produced and the characteristics of the nutrients available in the medium.

Finally, to prompt and aid experimental studies, we use our model to predict which nutrients can be a source of carbon for four species lacking a metabolic reconstruction, and predict the biomass production of these species in complex media. We show that the carbons available in the 20 natural amino acids in a medium provide the best boost to biomass production regardless of the species.

Catabolic potential: Can a given nutrient be a source of carbon for a specific microbe?

The first step is to build a model that predicts whether an individual nutrient \( i \) can be used as a source of carbon by species \( s \) (see Fig. 1). If it can, we say that \( i \) belongs to the group of nutrients \( G_i \) that contribute to growth in \( s \); otherwise we say that \( i \) belongs to the group of nutrients \( NG_i \) that do not contribute to growth in species \( s \). We use flux balance analysis on the species in the training set to empirically determine which nutrients belong to \( G_i \) and which ones to \( NG_i \) (see Materials and Methods). We then use these data to build the model. In this section we describe the model (which we summarize in Fig. 2) and validate it with the species in the test set.

Uptake. To establish whether a given nutrient is a source of carbon for a given organism, we first need to determine whether the nutrient can be transported into the cell from the extracellular medium (Fig. 2A). This question can be currently answered using bioinformatics tools. Recent studies have shed light on the function and orthology of families of microbial transporters for several nutrient classes [36–39]. Therefore, a search within a genome for orthologs of transporters provides a plausible answer as to whether a given nutrient is actively transported by a particular species.

The number of nutrients that are actively transported by a species varies greatly with nutrient class (Fig. 3B and Methods for a detailed explanation of nutrient classes). For example, \( M. \text{barkeri} \) does not uptake nutrients that we classify as sugars, sugar derivatives, cell boundary, fatty acids, purines, or pyrimidines. However, experiments with \( M. \text{barkeri} \) may have deliberately been focused on small organic compounds, so one may question the range of nutrients that \( M. \text{barkeri} \) is truly capable of uptaking. Note that we will have to ask this question for every new in silico organism, an infeasible task given the large number of experiments that one would need to perform.

Fatty acids are only actively transported by two of the species we analyze, but the absence of transporters in the other species does not mean that they cannot allow the diffusion of fatty acids through the cell membrane [40]. In the following, we assume that we will have to ask this question for every new in silico organism, an infeasible task given the large number of experiments that one would need to perform.

Nutrient classes. Next, we investigate whether entire classes of nutrients (defined by chemical structure and cellular function) can be labeled \( G \) or \( NG \) on the basis of the fraction of \( G \) and \( NG \) uptake by species. As a case study, we consider fatty acids, a common source of carbon for many species. Fatty acids are only actively transported by two species we studied (see Materials and Methods), but this does not imply that they cannot allow the diffusion of fatty acids through the cell membrane [40]. In the following, we assume that all species can uptake fatty acids in this manner.

Figure 1. Schematic representation of the development of a model for maximum biomass production in complex media of microbial organisms. We aim at developing a phenomenological model to predict the maximum biomass production \( B_m \) of species \( s \) when growing in a medium containing a set of nutrients \( \{p\} \) acting as a carbon source under aerobic conditions. That is, we want to express \( B_m \) as a function \( f \) \((\{p\}, p(\{i\})) \) that only takes into account data related to: i) the set of nutrients \( \{i\} \) available, namely, nutrient type, the set of pathways \( p(\{i\}) \) that can catabolize each nutrient, and the carbon content \( C_i \) of each nutrient; and ii) the species \( s \), specifically the presence or not of certain enzymes in a species that allow to catabolize specific types of nutrients (enzymes EC: 1.1.1.35, EC: 2.3.1.16, and EC 3.5.2.17 - see text). In order to achieve our goal, there are three different questions we need to answer: i) Does nutrient \( i \) produce growth or not in species \( s \) when acting as the sole source of carbon? We find that whether nutrient \( i \) produces growth (G) or not (NG) is a function of the nutrient type (see text) and its pathway membership; ii) If a nutrient produces growth, what is the maximal biomass \( B^i \) it can produce in species \( s \) when acting as the sole source of carbon? We find that \( B^i \) is proportional to \( C_i \), the number of carbons in nutrient \( i \), and that the proportionality constant \( \gamma_i \) depends on the species \( s \). iii) What is the maximal biomass production \( B_m(m) \) when growing on a complex medium \( m \)? We find that \( B_m(m) \) can be well approximated by adding up the individual contributions \( B^i \) of nutrients \( i \) present in medium \( m \).

doi:10.1371/journal.pcbi.1002762.g001
observations found for the training set of organisms. If this is the case, we would then be able to predict if a single nutrient is G or NG just from knowing to which class it belongs (Fig. 2B).

Based on the predicted biological function of the nutrients, we identify two classes whose nutrients are consistently found not to be sources of carbon: cell boundary nutrients and cofactors. Nutrients in the cell boundary class form the cell wall or the cell membrane and are composed of distinct metabolites that could, in principle, serve as carbon sources on their own. For example, lipid A contains several fatty acid chains that can be catabolized in _E. coli_. However, as in this particular example, these complex nutrients do not have enzymes available to decompose them, and they are instead used directly in the formation of the cell boundary. Cofactors have several metabolic functions (they serve as precursors to several biomass components, facilitate or act as carrier molecules in many reactions, act as transporters of important inorganic elements such as iron), but nearly none of them is catabolized by the species we study. Functional classes highlight that if a compound has a function in the cell other than being a source of carbon, then there will be selective pressure against the occurrence of enzymes that breakdown these nutrients.

When considering the chemical structure of nutrients, we find four nutrient classes whose components are consistent in whether they were sources of carbon or not: inorganic compounds, pyrimidines, sugars, and sugar derivatives. We label inorganic compounds and pyrimidines as NG since none of these nutrients acts as a source of organic carbon on its own for any of the species in the training set (Table S6b in Text S1). We label all sugars and

Figure 2. Determining whether a nutrient can or cannot be catabolized. **A.** To establish whether a given nutrient is a source of carbon for a given organism, we first need to determine whether the nutrient can be transported into the cell from the extracellular medium. **B.** For some nutrients, we can predict if _a_ it does or does not contribute to growth just from knowing to which class it belongs. Complex nutrients are broken down into simple nutrients. See the main text for the description of the enzymes that catabolize fatty acids and purines. **C.** Any nutrient that is not classified into the nutrient classes in **B** is classified as G or NG using the logistic model described in the text and Methods. doi:10.1371/journal.pcbi.1002762.g002
sugar derivatives as G, a classification that is accurate 89% of the times. This level of accuracy is expected since these are typically the major sources of carbon for many species (Tables S6d & e in Text S1). There are 13/118 observations where sugars or sugar derivatives are NG, despite the organisms having transporters available for their uptake. This means that the organisms have no enzymes that can catabolize these sugars and their derivatives, and it is likely that they are uptaken because of the lack of specificity of some sugar transporters [41,42]. One notable exception is inositol which can be involved in signaling processes [43] or in pathogenesis [44]; we surmise that, for this reason, some species may select against the presence of enzymes that allow inositol to be catabolized.

We find mixed results outside of the four structural classes discussed above. Purines (Table S6a in Text S1) are G in E. coli and B. subtilis, but NG in the other species in the training set. In E. coli and B. subtilis, purines are catabolized via the degradation of urate into glyoxylate; we find a well-characterized enzyme in this pathway that is present in these species, but is not present in the other two: 5-hydroxyisourate hydrolase (EC: 3.5.2.17). Therefore, if an organism contains genes encoding for this enzyme we label all purines uptaken by that organism G; otherwise, we label them NG.

Because we assume that fatty acids can always be uptaken by a microbe, we need to be able to predict whether a given fatty acid is G or NG on the basis of the enzymes available in the organism’s genome. The catabolism of fatty acids occurs through a self-contained pathway that produces acetyl-Coenzyme A (CoA) by way of β-oxidation [45]. A round of reactions in the β-oxidation process may depend on whether the fatty acids are saturated or not, but the final two reactions are always the same: i) the oxidation of L-3-hydroxyacyl-CoA to 3-Ketoacyl-CoA (EC: 1.1.1.35) and ii) the cleavage of 3-Ketoacyl-CoA with another CoA molecule to produce acetyl-CoA (EC: 2.3.1.16). If an organism contains genes encoding for the enzymes catalyzing both of these reactions, we label all fatty acids uptaken by that organism G; otherwise, we label them NG.
Logistic model for other nutrients. The remaining nutrients cannot be globally classified as G or NG based on their structure or function. Therefore, we investigate their catabolic potential on the basis of the metabolic pathways in which they participate (as defined by KEGG [33–35]; see Methods).

We use logistic regression [46] to estimate the probability $Pr(i)$ that nutrient $i$ is a source of carbon, that is $i \in G$, given a model $p_i$

$$Pr(i \in G | p_i) = \frac{1}{1 + e^{-\theta_i}}$$  (1)$$

We consider the following linear model for $p_i$:

$$p_i = \beta_0 + X_i^1 \beta_1 + X_i^2 \beta_2 + \ldots + X_i^k \beta_k + Y_i^1 \gamma_1 + \ldots + Y_i^z \gamma_z,$$  (2)$$

where $\beta_0$ is related to the probability that nutrient $i \in G$ if it is not in any of the pathways in the model, $k \in \{1, \ldots, P\}$ runs over the pathways included in the model, $i \in \{1, \ldots, z\}$ runs over the pairs of pathways included in the model, $X_i^k \in \{0, 1\}$ indicates whether a nutrient belongs to pathway $k$, $Y_i^g \in \{0, 1\}$ indicates whether a nutrient belongs to a given pair of pathways, and $\beta_k$ and $\gamma_g$ are interpreted as the change in the probability that nutrient $i$ can be catabolized by belonging to pathway $k$ or to the pair of pathways $g$, respectively. We performed the logistic regression using R (version 2.10.1 [47]).

We use the logistic model above to uncover a small set of pathways that has the greatest predictive power for the largest number of nutrients (Fig. 2C). We start with an initial null model of zero pathways, and build increasingly sophisticated models by sequentially adding linear terms corresponding to the pathways that provide the greatest increase in prediction accuracy (Fig. 4).

As our first strategy, we add terms by selecting a pathway with both a high number of G nutrients and a large fraction of G nutrients. Many pathways with only one or two G nutrients have 100% of G nutrients, but will not achieve our goal of covering the greatest number of nutrients. For $k = 1$ (only one term in the model), we select “Glyoxylate and dicarboxylate metabolism,” which has thirteen G nutrients and only two NG nutrients. When we consider which pathway to add next, we only consider the nutrients that are not already accounted for in the model, which is important because there is a significant overlap between pathways.

For example, when determining which pathway to select for $k = 2$, we note that after including “Glyoxylate and dicarboxylate metabolism”, we exclude fifteen nutrients from the rest of the pathways (Table S7 and Fig. S1 in Text S1). This has the additional effect of removing two pathways from consideration.

As our second strategy, we consider pathways with a high fraction of NG nutrients. By doing so, we build a model that can predict NG nutrients as well as G nutrients. There are several pathways that have a high fraction of NG nutrients, but we find that only the inclusion of “Folate biosynthesis” significantly improves the model.

As a third strategy, we add pairwise interaction terms to the model. These terms reduce the number of false positives due to all nutrients in a pathway with high fraction of G nutrients being predicted to be catabolized. We find that there are several...
pathways that contain NG nutrients also found in the G pathways in the model. These pathways do not necessarily have a high fraction of G or NG nutrients, but they enable us to pick out nutrients that are false positives in the pathways classified as G.

Our analysis identifies a logistic model with \( P = 8 \) pathways, and \( z = 2 \) pathway interactions (Table S8 in Text S1). The small number of pathways could be a function of the small number of species we are exploring. However, we find that 94% of the thousand or so microbes listed in KEGG have the pathways included in our model, suggesting that, despite its simplicity, our model is likely to be applicable with similar accuracy to other species.

**Model validation.** Our final model for catabolic potential integrates 7 nutrient classes and 10 pathways. Using our training set of species, we find that the true positive rate is \( TP = 0.89 \) and the true negative rate is \( TN = 0.92 \) (Fig. 4 and Fig. S2 in Text S1). To validate the model beyond the training set, we test its predictions on the set of test organisms described above. We find that the TP rate is consistent between the training set and the test set of nutrients, whereas the TN rate drops slightly (Fig. 5). This drop is mostly accounted for by nutrients that are NG in the test set, but that are found in the G pathways of the training set and are not corrected by the interaction terms in our model, thus providing an avenue for improving the model.

**Biomass yield: What is the biomass yield of a given nutrient for a specific microbe?**

The next step is to determine for those nutrients that produce growth what is the maximal production of biomass when that nutrient is the only source of carbon. In fact, if a nutrient is the carbon-limiting source in a given medium, the biomass yield of the nutrient must be related to the number of carbons in the nutrient. In Fig. 6, we display the \( \text{in silico} \) biomass production \( b_i \) and the biomass yield \( b_i/C_i \) of all nutrients as a function of the number \( C_i \) of carbon atoms they contain, for the species in the training and test sets (we do not consider \( M. \text{barkeri} \) in our model for biomass yield because \( M. \text{barkeri} \) can only grow anaerobically, and this has a significant effect on the energy used to polymerize the proteins and nucleic acids in the biomass).

It is visually apparent from Fig. 6 that there is a strong correlation between the number of carbons in a nutrient and the biomass production. We thus model the biomass production induced by nutrient \( i \in G \) as

\[
\tilde{b}_i = y_s C_i
\]

where \( y_s \) is the (nutrient-independent) average biomass yield of the nutrients in \( G \)

\[
y_s = \frac{1}{N} \sum_{i \in G} \frac{b_i}{C_i}.
\]

The data suggests the existence of an upper bound for the biomass yield \( (y_s^{\text{max}}) \) as a function of the number of carbons \( C_i \) in Figs. 6, 7, and 8. The nutrients frequently found at or close to this upper bound are sugars, alcohols, and other compounds with hydroxyl groups. Many of the hydroxyl groups on these compounds are typically oxidized in order to reduce NAD\(^+\) to NADH, which is an important source of energy, and will in turn increase the biomass yield. For simplicity, we obtain \( y_s^{\text{max}} \) as the average biomass yield for the sugars uptaken by an organism. The average nutrient has 83% of the efficiency of these high-efficiency sugars, but nutrients vary wildly in their yield (Fig. 7). We consider one main factor that contributes to this variation in biomass yield: the number \( C_i^E \) of carbons in a nutrient that are effectively catabolized.
Effective number of carbons. We find two sets of nutrients for which the number of carbons that is effectively catabolized is not the actual number of carbons available in the nutrient \( C_i^E \): complex nutrients and purines. Very often, catabolism of complex nutrients starts by breaking them down into a number of simpler nutrients, which are then catabolized according to their respective pathways. Tryptophan, lipids and nucleotides are examples of such complex nutrients: tryptophan is broken down into three separate metabolites (pyruvate, indole, and ammonia); lipids are broken into fatty acids, glycerol and a functional group such as choline; nucleotides are broken into a phosphorylated ribose and a nucleobase. Each of these simple compounds can act as nutrients on their own.

Complex nutrients display some of the largest variation in biomass yield, as highlighted in Fig. 8. This can be explained by the fact that not all of the constituent simple nutrients can be catabolized, that is, not all of the carbons in a complex nutrient can be used to produce biomass. Therefore, for each complex nutrient \( i \) we estimate \( C_i^E \) as follows. We break down the nutrient into its respective simple nutrients. Then, we find whether each simple nutrient \( j \) can be catabolized; \( C_j^E \) is the sum of \( C_j^E \) for each simple nutrient \( j \) that can be catabolized

\[
C_i^E = \sum_{j=1}^{n'} C_j^E,
\]

where \( C_j^E = 0 \) if \( j \not\in NG \), and \( n' \) is the number of simple nutrients that compose the complex nutrient \( i \). For example, consider the nucleotide thymidine for which \( C_{\text{Thymidine}} = 8 \). Thymidine is comprised of two simple nutrients, the phosphorylated ribose \( (C_{\text{Ribose}} = C_{\text{Ribose}}^E = 5) \), and thymine \( (C_{\text{Thymine}} = 3) \), but \( C_{\text{Thymine}}^E = 0 \) because \( \text{Thymine} \not\in NG \). Therefore \( C_{\text{Thymidine}}^E = C_{\text{Ribose}}^E + C_{\text{Thymine}}^E = 5 \).

An additional consideration is the fact that some nutrients “leak” carbons during catabolism. Purines are an example of this; they are reduced to glyoxylate, but in doing so, they “leak” three carbons as carbon dioxide and urea, which cannot be catabolized for the carbon. Therefore \( C_{\text{Purine}}^E = C_{\text{Purine}} - 3 \).

Figure 6. Biomass production is related to the number of carbons in a nutrient. We show the optimized biomass production of each species on \( G \) nutrients, for species in the training set (left) and the test set (right). For all species there is a positive correlation between biomass production and the number of carbons in the nutrient. The blue line represents \( \widehat{b} \) (see text) for all the sugars uptaken by species \( S. \) aureus exhibits a reduced biomass production; the biomass defined in the in silico organisms demands approximately ten times more moles relative to the other species. In all the plots, the position of the nutrients on the X axis is slightly staggered so that all data points are visible. Note that the symbols for the complex nutrients are enlarged.

doi:10.1371/journal.pcbi.1002762.g006
Figure 8 displays how the biomass yield of complex nutrients changes when we adjust for the effective number of catabolizable carbons. Tryptophan, for example, is now seen to be catabolized as efficiently as other amino acids. Many purines and sugar derivatives only contained sugars as catabolizable simple nutrients, and therefore their efficiency becomes comparable with that observed for nutrients in the sugars class. Adenine and guanine still remain inefficient (the two red dots towards the bottom of the purines class), but as their catabolic end product is glyoxylate, the result is comparable with a glyoxylate derivative highlighted in the organic acids class.

For some classes, such as amino acids and sugar derivatives, the variance of the biomass yield for nutrients within the classes is also reduced. This reduction supports the use of the effective number of carbons as the predictor of biomass yield of a single nutrient. We thus modify Eq. (3) accordingly:

$$\tilde{b}_i = y_i \tilde{C}_i, \forall i \in G_s.$$  

(4)

Note that species in the test set take up very few complex nutrients and no purines; we therefore do not show the results of using Eq. (4) for these species.

Complex media: Is biomass yield of a nutrient affected by the presence of other nutrients in the medium?

Finally, we want to model the maximal biomass production when there is more than a single source of carbon present in the medium, or in other words when organisms grow in a complex medium. We consider complex media in which nutrients are restricted to five classes: sugars, fatty acids, amino acids, purines, and pyrimidines, partly because they are commonly used in growth rate/biomass yield experiments, and partly because it simplifies the analysis of the results. In addition, complex purines and pyrimidines are a mixture of sugars and nucleobases, and as we are already considering sugars, we only use the simple nucleobases.

The simplest plausible model for the contribution of nutrients to the biomass production is one in which each nutrient $i$ has an independent contribution to the biomass production. For each species $s$, we calculate the biomass production $B_{s,\text{model}}$ on a complex medium containing $N_m$ nutrients using

$$B_{s,\text{model}} = \sum_{i=1}^{N_m} \tilde{b}_i.$$  

(5)
We estimate $b_ib$ for a nutrient from class $c$ using the average yield for that class in the organisms in the training set.

$$\mathcal{Y} = \left( \frac{1}{Nc} \sum_{i=1}^{Nc} y_is \right)_{\text{avg}}$$

(6)

where $Nc$ is the number of nutrients of class $c$ that are taken up by species $s$, and $\langle \ldots \rangle_{\text{avg}}$ represents the average over species in the training set. For tryptophan and for purines, we use the effective number of carbons when calculating yield, as described previously.

To test this model, we randomly generate ensembles of complex media as described in the Data and Methods section. For each medium and for each species, we calculate $B_{\text{model}}$, we use FBA to find the actual in silico biomass production of the organism on the complex media $B_{\text{silico}}$. In Fig. 9 we show how $B_{\text{model}}$ compares to $B_{\text{silico}}$ for each species, and for each of the 1000 complex media we generate. In these comparisons we also include predictions for the species in the test set (with the exception of $S. aureus$, for whose reconstruction it is not clear what units were used for the biomass production).

We find a very strong correlation ($p < 10^{-6}$ using the Spearman rank correlation coefficient) between the predictions of our model and in silico experimental results for training set species ($E. coli$, $B. subtilis$, and $S. cerevisiae$) and test set species ($H. pylori$ and $M. tuberculosis$). This indicates that the model accurately captures which media will result in faster/slower production of biomass. For each species, however, the model systematically under or over-predicts growth. A regression of the predicted biomass production versus experimental in silico production indicates that $B_{\text{model}} \approx K \cdot B_{\text{silico}}$, with: $K = 0.76$ for $E. coli$, $K = 0.62$ for $B. subtilis$, $K = 0.84$ for $S. cerevisiae$, $K = 1.10$ for $H. pylori$, and $K = 0.92$ for $M. tuberculosis$.

The parameters used to make the predictions for the individual nutrients in the linear model were trained on three species. The linear model consistently under-predicted the in silico biomass production for these three species, and more so for media containing more nutrients. This is a strong indicator that the nutrients that are uptaken are used synergistically by the in silico organism to produce more biomass than expected, showcasing the effect of catabolic pathways being highly connected in the metabolic network. A more complete model for predicting biomass

Figure 8. Adjusting for the effective number of carbons in complex nutrients and purines in the training set. We show the normalized biomass yield (see Fig. 7) for purines and complex nutrients (full colored symbols) for species in the training set. In the left column, we show the normalized biomass yield considering the number of carbons in each nutrient $C_i$ in the right column, we show the normalized biomass yield using the effective number of carbons $C^E_i$ (see Text and Methods). Additionally, for each nutrient class that contains these nutrients, we show the mean and variance.

doi:10.1371/journal.pcbi.1002762.g008
L. monocytogenes, a gram-negative bacterium, D. discoideum (an eukaryote), and T. acidophilum (an archaeon).

We then use the model of catabolic potential to predict whether each of these nutrients could be a source of carbon. In Fig. 10B we show the number of nutrients that belong to one of the nutrient classes previously described and whether these nutrients are $G$ or $NG$. For the fatty acids, none of which were predicted to be uptaken by any of the four species, we examined the enzymes available and found that only D. discoideum contains the enzymes for \( \beta \)-oxidation, and could therefore catabolize fatty acids.

Finally, we combine our models of biomass yield of G nutrients, and of biomass production on complex media (Fig. 11). We choose four complex media which contain a different number of the same nutrients we used for the randomly-generated complex media, namely: sugars, fatty acids, bases, and amino acids. The four media contain: 1) glucose; 2) glucose and hexanoic acid; 3) glucose, hexanoic acid, adenine, guanosine, cytosine, and thymine; 4) glucose, hexanoic acid, adenine, guanosine, cytosine, and thymine and the twenty natural amino acids. We estimate the biomass production in the same manner that we described earlier for the randomly-generated complex media. We find that the biggest influence on the biomass production is the number of carbons available in the nutrients present in the medium; because we are now adding 20 amino acids to medium 4, the biomass production increases almost 7-fold, on average, in that case.

Note that these predictions for biomass production are based on the biomass yield per carbon available in the nutrients present in the medium. Biomass yield is a time independent quantity that cannot be directly associated to growth rate. However, because biomass yield gives an upper limit for growth rate [48], our model can be used as a baseline for researchers to explore and model growth rates.

## Discussion

Microbes use nutrients found in their environments to grow. Understanding and developing quantitative models of this process is of fundamental importance in cell biology, physiology, medicine, evolution, synthetic biology and bioengineering, not to mention of practical importance to those that need to grow microbes in laboratories. Given the diversity of the microbial world and the number of combinations of nutrients available in their environments, this seems too difficult a problem. In this study, we focused on how microbes might catabolize nutrients to obtain carbon for biomass production.

Our model comprises three levels, with each level building up on the results from the previous one. The first level concerns whether a nutrient will be catabolized. The second level concerns whether all of the carbon in a catabolized nutrient is available for biomass production. The final level incorporates the biomass yield for selected classes of nutrients and enables us to make a prediction on the biomass production of a microbe in a complex medium. To validate our approach, we compare the predictions of the complete model with in silico predictions of growth in complex media of species that are not part of the training set. Our results on these species are excellent predictions of which media will produce more/less growth.

Finally, we looked at the biomass yield of sugars, amino acids, purines, pyrimidines, and fatty acids. We found little variation in the yield of these nutrients amongst different species, and were able to postulate a model for the biomass production of a microbe on a complex medium containing any number of these important nutrients. All of these nutrients have separate catabolic pathways, with the exception of some groups of amino acids which share a catabolic pathway. The fact that the in silico microbial biomass production was 30% more than our model predicts indicates that each of these catabolic pathways can work together synergistically to improve the biomass production of the microbe.
The ability of our model to predict biomass production on complex media must be balanced by an understanding of its limitations. First, we model microbial metabolism as a black box, and therefore we do not account for absent pathways that biosynthesize some biomass components. The absence of such pathways would require that the corresponding biomass components be made available in the medium used, but our model cannot be used to predict which of these are indeed required. This means we cannot use the model in its current form to predict the minimal medium that is needed for a microbe to grow.

Second, microbial species are sometimes identified by the nutrients they take up and excrete. The ability of a microbe to transport such compounds in and out of the cell is largely dependent on the specific protein transporters available. A separate body of work exists, largely in the form of TransportDB and TCDB, which will enable the researchers to predict the proteins that transport specific nutrients. We incorporate such predictions of nutrient transport into our model to make predictions for specific species that lack a metabolic reconstruction. Importantly, while knowledge of transporters can enable us to predict which nutrients in the medium can be taken up, we are currently unable to predict which nutrients are excreted as that would in principle require knowledge of the full metabolic network.

Third, the stoichiometric method we describe for generating the biomass data cannot be used to predict the growth rate of a microbe because kinetic information is not included. However, our model provides a baseline to which one can add kinetic information. For example, the rate-limiting step in poor growth media is likely to be the rate at which a microbe takes up nutrients. Therefore one can use the kinetics of nutrient transport with our prediction of biomass yield to predict growth rate.

All this notwithstanding, we believe that our approach and models open the door to significant advances in the quantitative modeling of microbial metabolism, and eventually of the metabolism of more complex organisms. In particular, our models could be extended to consider whether a nutrient acts, not only as a source of carbon, but also as a source of nitrogen or energy, or...
directly as a component of the biomass. Our models could also be extended to include more detailed information about pathways, or to consider functional metabolic modules [51–53] instead of pathways.

Methods

Data

In silico organisms. We first consider a training set of four species for which there are high quality metabolic reconstructions available, and which cover a wide range of microbial phylogeny: a gram-negative bacterium (E. coli [9]), a gram-positive bacterium (B. subtilis [9]), an eukaryote (Saccharomyces cerevisiae [21]), and an archaeon (Methanosarcina barkeri [24]). We validate our model on a set of test species: Helicobacter pylori, a gram-negative bacterium [20], Staphylococcus aureus, a gram-positive bacterium [22], and Mycobacterium tuberculosis, an acid-fast gram-positive bacterium [23].

We download these organisms from http://regc.ucsd.edu/ in SBML format (accessed 03/27/2009). There are 352 distinct nutrients listed for the metabolic reconstructions of the four in silico organisms in the training set. According to the reconstructions, out of these 352 nutrients only 26 are uptaken by all four species, a fact that illustrates the diversity of microbial metabolism.

We use the available reconstructions and the literature to define a fully minimal medium for each in silico organism [54–57]. The minimal medium does not enable the organism to produce biomass. However, the addition of a carbon source is enough to enable growth. We do not consider components of a minimal medium for which an uptake reaction is not present in the reconstruction. We describe in detail the minimal media for the seven organisms in our training and test sets in Table S1 in Text S1.

Organisms in prediction set. We generated predictions of biomass production on four species for which we lack a metabolic reconstruction, and which have the same phylogenetic breadth as the training set: Rhodopseudomonas palustris (a gram-negative bacterium), Listeria monocytogenes (a gram-positive bacterium), Dictyostelium discoideum (an eukaryote), and Thermoplasma acidophilum (an archaeon). We identified the nutrients that these organisms take up (Tables S2a–j in Text S1) from the prediction of nutrient transporters found in the database TransportDB [39] (accessed 11/03/2010).

Metabolites and pathways. We downloaded KEGG [33–35], and SEED’s [28,58] metabolite databases, and KEGG’s pathway database. The SEED metabolite database allows us to link the metabolites in various reconstructions to their respective KEGG IDs. The KEGG metabolite and pathway databases allow us to assign the metabolic pathways to the individual metabolites in the reconstruction. The pathway database in KEGG is organized into 7 groups, one of which is metabolism. Metabolism is further divided into 11 major subgroups (for example, “Amino Acid Metabolism”). Each of these major subgroups is further divided into metabolic pathways (for example, “Arginine and Proline Metabolism”). There are 156 distinct metabolic pathways, but only 50 pathways are found in all seven species we used in this study.

We cross-checked the identifiers used in the reconstructions with the linked names in the SEED database to find mistakes, duplicates, and missing KEGG IDs. There are metabolites used in the reconstructions that we could not find in the KEGG database. For each one of these, we attempted to manually assign the KEGG ID of related metabolites, by using KEGG and SEED. There were 27 metabolites for which we attempted this manual assignment; for 5 of these, we could not unambiguously assign a KEGG ID (Tables S3a&b in Text S1). In addition, there are metabolites which do not have any KEGG pathways assigned. We used KEGG and SEED to find the most relevant pathway for these metabolites, and we manually assigned the KEGG ID of another metabolite in the same pathway, thereby assigning the same pathways by default. There were 43 metabolites for which we attempted this manual assignment. For 12 of these, we were unable to find an appropriate pathway classification (Tables S3c&d in Text S1).

Complex nutrients. Many of the 352 nutrients in our list may be considered complex in that they are composed of one or more simple nutrients. For example, adenosine consists of two simple nutrients: adenine and ribose. Complex nutrients are typically degraded into their simple components and then each component may or may not be catabolized separately. We denoted every nutrient in our list either as simple or complex. For the latter, we identified the simple nutrients that compose it (Tables S4a–f in Text S1).

Nutrient classes. We classify the nutrients by their structure and, to a lesser extent, function. We define 12 classes. Five of these classes—“Sugars”, “Amino acids” (including uncommon amino acids), “Purines”, “Pyrimidines”, and “Fatty acids”—are self-explanatory. We classify elements and sources of inorganic carbon as “Inorganic compounds.” The “Cofactors” class includes cofactors, vitamins—which are precursors of cofactors—and nutrients whose primary function is to chelate an inorganic element. Many nutrients that were deemed to be derivatives of amino acids, sugars or fatty acids were included in the relevant “derivative” categories. In addition, larger nutrients involved in the formation of the cell boundary, such as lipid A, were included in the “Cell boundary” nutrient class. We lumped together the remaining nutrients—which are small organic compounds—as “Organic compounds” (Tables S5a–l in Text S1).
Flux Balance Analysis experiments

Stoichiometric methods have been widely used in metabolic engineering for over 20 years [59], the most used of which is Flux Balance Analysis (FBA) [26,27,60]. FBA aims at determining the fluxes $\mathbf{v}$ through each one of the metabolic reactions in an organism. Thus, FBA relies on the determination of a stoichiometric matrix $\mathbf{S}$ that represents all the reactions and metabolites in an in silico organism $s$, and a vector $\mathbf{u}$ of uptake fluxes.

In the matrix $\mathbf{S}$, each row corresponds to a reaction, and each column to a metabolite. $S_{ij}$ is the stoichiometric coefficient of metabolite $j$ in reaction $i$. The vector of uptake fluxes $\mathbf{u}$ can have a non-zero entry for every transport reaction that moves a nutrient into the in silico organism.

As a simple example, consider a metabolic network with three reactions:

\[
\begin{align*}
R_1: & \quad \text{Glucose} + \text{ATP} \rightarrow \text{Glucose} - 6 \text{ phosphate} + \text{ADP} \\
R_2: & \quad \text{ADP} + \text{Pi} \rightarrow \text{ATP} \\
R_3: & \quad \text{Glucose}_{\text{env}} \rightarrow \text{Glucose}
\end{align*}
\]

where $R_3$ represents the transport reaction for uptaking Glucose from the environment. The stoichiometric matrix for metabolic network (7) is:

\[
\mathbf{S} = \begin{pmatrix}
0 & -1 & -1 & +1 & +1 & 0 \\
0 & 0 & +1 & 0 & -1 & -1 \\
-1 & +1 & 0 & 0 & 0 & 0
\end{pmatrix}
\]

Assuming a steady-state concentration of every metabolite and requiring mass-conservation we must impose that

\[
\mathbf{S} \mathbf{v} = \mathbf{u},
\]

where $\mathbf{v}$ is the set of fluxes through each metabolic reaction. In order to solve Eq. (9), we need to provide the vector $\mathbf{u}$. The default flux for each uptake reaction is 0, meaning that the nutrient is not in the medium or that it cannot be uptaken by the organism. We set $u_i = -1$ if nutrient $i$ is present in the medium and could be taken up by the organism.

The system in Eq. (9) has many solutions. In our analysis, the biologically relevant metabolic state is the one that maximizes biomass production $b$ [59]. The problem of finding the fluxes through the reactions with a cost function that have to satisfy a number of constraints is a standard linear optimization problem that can be numerically solved using the subroutines provided in GLPK [61].

In order for us to model whether a nutrient can be a source of carbon and the biomass yield of the nutrient, we control the data in two ways. First, for some organisms the uptake of specific nutrients can lead to $b > 0$ without being considered significant [8]. Secondly, ATP hydrolysis is integrated into the biomass of the in silico organisms, and is typically trained to better match empirical results based on growth rates. We are not exploring growth rates and we thus adjust the ATP hydrolysis in the biomass of the in silico organisms in order to better support the conclusions we reach. These controls are explained in detail in Text S1.

Model selection

Model accuracy. To determine which of the logistic models we build is the best, we must balance accuracy with economy. We assess the model’s sensitivity and specificity using two sets of measures. First, we calculate the true positive rate ($TP$) and the true negative rate ($TN$), which indicate how many nutrients are successfully predicted to be in $G$, or $NG$, respectively. Secondly, we calculate the area under the ROC curve, $AUC$, to quantify the accuracy of the model [62].

We assess whether the prediction for a nutrient is a true positive or a true negative as follows: For each nutrient $i$, we first estimate $p_i$. For every $G$ nutrient, if $p_i \geq 0$ (which is equivalent to $Pr(\epsilon \in G | p_i) \geq 0.5$), the nutrient is predicted to be a source of carbon, and we record a true positive. If, for a $G$ nutrient, $p_i < 0$, the nutrient is predicted to not be a source of carbon and we record a false negative. For every $NG$ nutrient, if $p_i < 0$, it is recorded as a true negative; if $p_i \geq 0$, it is recorded as a false positive.

We calculate the $AUC$ as follows. First, we rank all of the nutrients according to the probability that they are a source of carbon $Pr(\epsilon \in G | p_i)$. Then starting with the nutrient with the highest $Pr(\epsilon \in G | p_i)$, for every $n \epsilon \{1, \ldots, N\}$ nutrients we count the fraction of nutrients that are sources of carbon $f_G^G = \frac{G_n}{G_N}$ and the fraction of nutrients that are not sources of carbon $f_{NG} = \frac{NG_n}{NG_N}$ where $G_n$ and $NG_n$ are the number of nutrients in $n$ that can, and cannot be sources of carbon respectively, and $G_N$ and $NG_N$ are the total number of nutrients that can and cannot be sources of carbon, respectively. Each $(f_G^G, f_{NGi})$ pair is a point in the ROC curve. We expect $f_G > f_{NG}$ at the top of the list and $f_G < f_{NG}$ at the bottom of the list. The $AUC$ is the area under this curve. An $AUC$ close to 1 means that the pathways model succeeded in separating the nutrients that are sources of carbon from the nutrients that are not sources of carbon.

Model economy. The larger $P$ and $z$ are (that is, the larger the number of parameters in Eq. (2), the more accurate the model is for the data at hand. However, with every additional parameter, the chances that we overfit the data increases. This issue becomes particularly worrisome if we consider pathways that are only partially conserved, for these pathways may not be represented in the test set of species we use to validate our model. To assess the economy of the model, we use two standard information criteria, the Akaike Information Criterion (AIC) [63]

\[
AIC = 2P + z - 2 \log(L)
\]

and the Bayesian Information Criterion (BIC) [64]

\[
BIC = (P + z) \ln(N) - 2 \log(L).
\]

These information criteria balance the number of parameters, which in this case is the number of pathways and pairs of pathways $P + z$, with the maximum likelihood of the model $L$, a measure of how effectively the model predicts the catabolic potential of a nutrient. When a new parameter is added to the model, $L$ must increase sufficiently to balance the increase in $P + z$. In addition, $BIC$ includes the number of data points (in our case nutrients $N$ in the dataset), and is therefore more stringent if the dataset is small. By adding parameters, we can find the pathways and pairs of pathways at which the information criteria are at the global minimum.

Complex media

We tested our model on a large number of complex media generated using the nutrients available for uptake in the in silico
organisms. Because the number of possible combinations of these nutrients was too large for us to test each one computationally, we considered an ensemble of 1000 randomly generated complex media \( M(q) \). For each complex medium, every nutrient was made available for uptake with the probability \( q \), and was excluded with the probability \( 1-q \). We generated ensembles for \( q = 0.1, 0.3, \) and \( 0.5 \) (for a total of 3000 random media).

Sugars present an unusual case because a microbe such as \( E. coli \) has been known to exhibit diauxic growth [63–67]. This means that microbes regulate sugar uptake so that, despite having various sugars present in the medium, they will only take up one sugar at a time. Our model does not take gene regulation into account, and therefore we manually limited the number of sugars presents in each complex medium to one, specifically glucose.

**Supporting Information**

**Text S1 Supplementary information regarding data controls.** Supplementary figures showing: S1) the redundancy in pathway membership; S2) the model selection process. Tables describing: S1) the minimal media used for each species; S2) the prediction of our model for each nutrient separated by nutrient type; S3) the reassignment of KEGG IDs and pathways for compounds lacking that information; S4) the composition of complex nutrients; S5) the classification of nutrients in different classes; S6) the uptake of fatty acids and sugars by the different species; S7) the prediction of the logistic regression pathway model for each nutrient; S8) coefficients for the different terms in the logistic regression model. (PDF)

**Acknowledgments**

We thank R. Morimoto, J. Widom, J. Brickner, P. McMullen, A. Pah, E. Sawardecker-Amundsen, A. Salazar, and A. Hockenberry for helpful discussions and suggestions during the formulation of the research and the writing of this manuscript.

**Author Contributions**

Conceived and designed the experiments: SMDS MSP RG LAN A. Performed the experiments: SMDS MSP. Analyzed the data: SMDS MSP. Contributed reagents/materials/analysis tools: SMDS MSP RG. Wrote the paper: SMDS MSP RG LAN A.