nature food

ARTICLES

https://doi.org/10.1038/s43016-022-00511-0

Check for updates

Nutrient concentrations in food display universal behaviour

Giulia Menichetti[1,2] and Albert-László Barabási[1,2,3]✉

Extensive programmes around the world endeavour to measure and catalogue the composition of food. Here we analyse the nutrient content of the full US food supply and show that the concentration of each nutrient follows a universal single-parameter scaling law that accurately captures the eight orders of magnitude in nutrient content variability. We show that the universality is rooted in the biochemical constraints obeyed by the metabolic pathways responsible for nutrient modulation, allowing us to confirm the empirically observed scaling law and to predict its variability in agreement with the data. We propose that the natural nutrient variability in food can be quantitatively formalized. This provides a mathematical rationale for imputing missing values in food composition databases and paves the way towards a quantitative understanding of the impact of food processing on nutrient balance and health effects.

Universality, a concept rooted in statistical physics[1], captures the observation that measurable macroscopic features can emerge from the interactions of a large number of individual components, features that cannot be reduced to the properties of single elements[2]. The food we eat, be it ingredients consisting of simple plant or animal products or dishes mixing multiple ingredients, carries thousands of chemicals, whose concentrations remain unquantified in most foods[3–5]. Yet, chemical concentrations in food are modulated by a densely wired biochemical reaction network[6], suggesting that the concentrations of individual components may follow common patterns, governing their expected values as well as the extent of their fluctuations across the food supply.

In the past few decades, the US Department of Agriculture (USDA) and national departments of agriculture and health worldwide[7] have devoted major efforts to systematically quantify and tabulate about 100 chemicals present in food, mostly macronutrients and micronutrients necessary to maintain a healthy diet or compounds associated with adverse effects on health. Naturally, there is well-documented variability in nutrient concentrations, depending on the growth conditions, source and time of measurement, and variability induced by cooking and processing, changes that are also captured and reported in these databases. As we show next, despite these inherent differences, all innate nutrient concentrations follow a universal distribution across the food supply, a finding with implications for nutrient access.

Results

Our work relies on the hypothesis that nutrient distributions across the food supply emerge as macroscopic features of the biochemical reaction networks characterizing living organisms. Hence, they may exhibit universal features. Leveraging food composition data collected by the USDA and kinetic constants from BRENDA[8], we show how nutrients display a consistent statistical behaviour, predictable from biochemical first principles.

Formalizing the nutrient composition of food. The food supply, representing the full inventory of all foods available for human consumption, along with their nutritional content, plays an important role in determining an individual's nutrient exposure. This information is captured in the matrix F_{nd}, representing the amount of nutrient n in 100 g of any ingredient (or composite food or drink product) d. For instance, $F_{n,\mathrm{apple}}$ tells us that the consumption of 100 g of raw apple delivers 52 nutritional components, including 10.39 g of sugars, 0.107 g of potassium and 0.0075 g of (−)-epicatechin, a polyphenol (Supplementary Section 1). Overall, as shown in Fig. 1a, the range of chemical concentrations present in raw apple shows remarkable variability, spanning eight orders of magnitude, from vitamin K (2.20×10^{-6} g) to water (85.56 g).

Given the wide range of food and drinks available to the consumer in grocery stores and restaurants and their home-cooked variants, a key determinant of nutrient exposure is the probability $P(x_n)$ that an individual (or a population) is exposed to x_n grams of nutrient n in a randomly consumed dish:

$$P(x_n) = (1-p_n)\,\delta(x_n) + p_n\mathcal{Q}(x_n). \quad (1)$$

Here p_n is the probability that nutrient n is present in a random dish, and $\mathcal{Q}(x_n)$ is the probability that the selected item carries x_n grams of nutrient n[9,10]. For instance, $p_n=0.9859$ for zinc, as the mineral is present in 98.59% of all foods, while for hesperetin (a flavonoid produced by the secondary metabolism of citrus and orange), p_n drops to 0.0446. The probability $\mathcal{Q}(x_n)$ plays a fundamental role in nutrient exposure[11], capturing the food source variability of nutrient n available to the population. Indeed, individuals sample foods from the food supply according to their dietary pattern, and a precise description of $\mathcal{Q}(x_n)$ is instrumental to quantify how nutrient intake varies in the population and the likelihood of observing extreme values and deficiencies[12]. Yet, neither nutritional science nor food chemistry offers guidance on its expected statistical distribution in food composition databases. This lack of knowledge is rooted in the complexity of the biochemical processes that modulate specific nutrients in individual staple ingredients at the origin of the food supply, as well as the phylogenetic diversity of food. Some nutrients, such as polyphenols, are only synthesized by the secondary metabolism of plants, while amino acids and simple sugars are present in all food but come in concentrations that are highly

[1]Network Science Institute and Department of Physics, Northeastern University, Boston, MA, USA. [2]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [3]Department of Network and Data Science, Central European University, Budapest, Hungary. ✉e-mail: barabasi@gmail.com

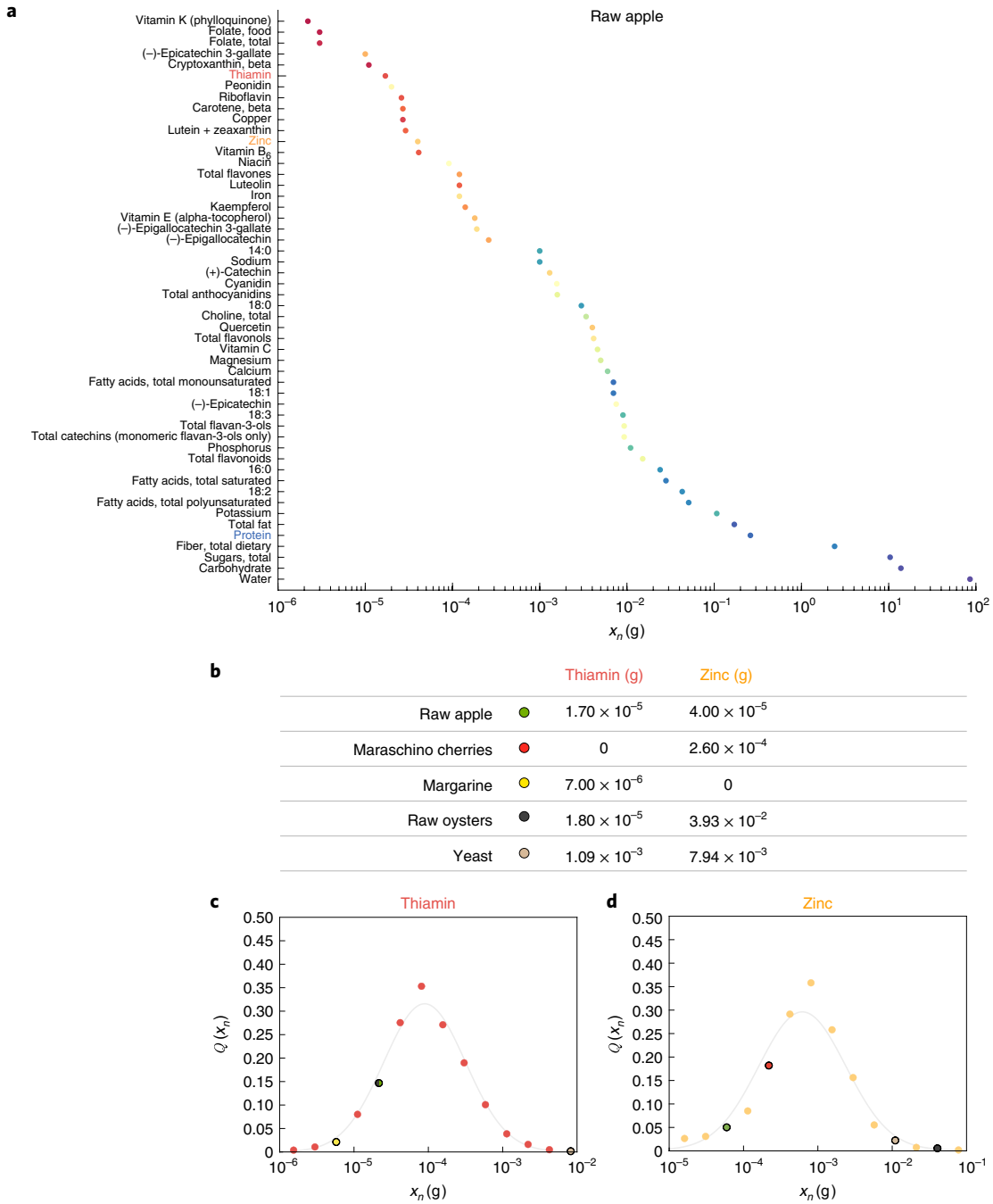

Fig. 1 | Nutrient composition of food. a, The consumption of 100 g of raw apple delivers 52 nutritional components, whose amounts (measured in grams) span eight orders of magnitude. The "Apple, raw" food code 63101000, in the FNDDS database, captures the average apple by combining a variety of samples. We ranked the nutrients in apple in ascending order of concentration. **b**, The concentrations of thiamin (a vitamin) and zinc (a mineral) in five different foods in the food supply, representing the amount of nutrient n in 100 g of the respective ingredient. **c**,**d**, For thiamin (**c**) and zinc (**d**), we calculated $\mathcal{Q}(x_n)$ using equation (1) and show the distribution on a logarithmic scale. Each symbol represents a histogram bin, and we highlight in different colours the bins that contain the foods shown in **b**.

organism specific. Hence, the mathematical formulation of $\mathcal{Q}(x_n)$ is expected to depend greatly on the specific nutrient class (fatty acids, sugars, minerals, vitamins or flavonoids) and whether it is part of a plant's central or secondary metabolism. Despite these remarkable differences, the variability of nutrients across foods follows common patterns, captured by similarly shaped $\mathcal{Q}(x_n)$ (Fig. 1b–d).

A universal scaling law for nutrient content. To characterize the variability of nutrient content across the food supply, we measured $\mathcal{Q}(x_n)$ for 99 nutrients whose concentrations in 4,889 foods are profiled by the USDA and reported by the National Health and Nutrition Examination Survey (NHANES)[9,10]. (See Supplementary Section 1 for a description of the different food databases curated by the USDA. The limitations of these data sources are discussed in Supplementary Sections 2 and 3, and for robustness checks, we verified the validity of our findings for raw vegetables and fruits (Supplementary Section 4), for composite dishes (Supplementary Section 5) and in independent datasets (Supplementary Section

6), and we explored the role of sample variability (Supplementary Section 7)). Here we define as nutrients all chemicals catalogued by national food composition databases, whether they refer to unique chemicals (such as vitamin C) or aggregate measures (such as total fat or total sugar). We kept all nutrients measured in g, mg or μg, dropping "Energy", "Folate, DFE" and "Vitamin A, RAE", resulting in 99 nutrients, converted to grams. In Fig. 2a, we show the measured $\mathcal{Q}(x_n)$ for thiamin (a vitamin), zinc (a mineral), gadoleic acid (a fatty acid) and total protein, capturing the distribution of these nutrients across all food in our database. Interestingly, we find no evidence of the expected diversity and nutrient specificity—rather, each nutrient, independent of its chemical class, has a remarkably similar $\mathcal{Q}(x_n)$. A closer inspection of Fig. 2a indicates that $\mathcal{Q}(x_n)$ obeys three robust patterns that ultimately help us unveil its functional form:

(1) Constant standard deviation: The standard deviation of $\mathcal{Q}(x_n)$ in the log space, $s_n = \sqrt{\langle (\log x_n)^2 \rangle - \langle \log x_n \rangle^2}$, capturing the variability of nutrient n across all foods, appears to be the same for each of the four nutrients. This suggests that the degree of variability in nutrient content across all foods is independent of the nutrient concentration. To see if this is true beyond these four nutrients, we measured s_n for all nutrients, and we found that, despite the eight orders of magnitude spanned by nutrient concentrations, the standard deviation s_n is remarkably constant, fluctuating near $s_n = 1.66 \pm 0.39$ (Fig. 2b).

(2) Symmetry: Fig. 2a indicates that $\mathcal{Q}(x_n)$ is symmetric in the logarithmic scale. To see if this is indeed the case, we measured the logarithmic skewness of $\mathcal{Q}(x_n)$ for all nutrients, whose value is zero for any symmetric distribution and positive (negative) for right (left) tailed distributions. We find the empirically observed skewness to be approximately zero for each nutrient (Fig. 2c), confirming the symmetric nature of $\mathcal{Q}(x_n)$.

(3) Translational invariance: On a logarithmic scale, the four $\mathcal{Q}(x_n)$ appear to be identical but shifted horizontally, a pattern mathematically described as translational invariance in the log space. Formally, this implies that under the scale transformation $x' = cx$, the probability distribution rescales as $\mathcal{Q}(x'_n) = \frac{\mathcal{Q}\left(\frac{x_n}{c}\right)}{c}$. We tested the validity of this hypothesis, finding that under a horizontal shift of each curve in Fig. 2a (corresponding to the rescaling $y_n = e^{\langle \log(x_n) - m_n \rangle}$, the $\mathcal{Q}(x_n)$ for all nutrients collapse on a single universal distribution (Supplementary Fig. 10a,b and Supplementary Section 3b).

Taken together, the patterns (1)–(3) suggest the existence of a single family of distributions that describes $\mathcal{Q}(x_n)$ for all nutrients. Formally, these three patterns also exclude most well-known distributions, such as Gaussian, gamma, Weibull or Fréchet, as functional candidates for $\mathcal{Q}(x_n)$, as these distributions formally violate at least one of the three properties identified above (Supplementary Table 1). We find, however, that the log-normal family[13]

$$\mathcal{Q}(x_n) = \frac{1}{x_n s_n \sqrt{2\pi}} e^{-\frac{(\log x_n - m_n)^2}{2(s_n)^2}} \quad (2)$$

can account for all three empirical observations, as it is characterized by (1) a constant s_n consistent with $s_n = 1.66 \pm 0.39$, (2) symmetry in the log space and (3) translational invariance. To test if indeed equation (2) captures $\mathcal{Q}(x_n)$ we fitted this equation to each of the 99 nutrients and used the Kolmogorov–Smirnov criteria to compare the fit with several distributions that satisfy at least one of (1)–(3). The analysis confirms that the log-normal equation (2) offers the best approximation for $\mathcal{Q}(x_n)$ (see Supplementary Section 3 for additional statistical evidence).

Most important, the log-normal equation (2) makes a falsifiable prediction for nutrient distributions. Indeed, the average concentration μ_n and the standard deviation σ_n of nutrient n across all foods connect to their counterparts in the log space, m_n and s_n in equation (2), via $\mu_n = e^{m_n + \frac{s_n^2}{2}}$ and $\sigma_n = e^{m_n + \frac{s_n^2}{2}} \sqrt{e^{s_n^2} - 1}$, implying that

$$\sigma_n = \mu_n \sqrt{e^{s_n^2} - 1}. \quad (3)$$

In general, equation (3) can describe rather complex σ_n functions, depending on the dependence of s_n on μ_n. However, our finding that s_n is independent of μ_n in food (Fig. 2b) implies that σ_n must be linearly proportional to μ_n. To test this prediction, we plotted σ_n as a function of μ_n (Fig. 2d), finding that despite eight orders of magnitude of differences in μ_n, we have $\sigma_n \sim \mu_n$. Note that we observed small deviations from the linear fit only at high μ_n, corresponding to water, carbohydrate, total fat and total protein. These data points represent cumulative rather than individual nutrient measures, and their abundance is limited by the fixed mass (100 g), explaining why the corresponding σ_n reaches lower values at high μ_n (see Supplementary Section 2 for a detailed statistical analysis).

As our primary source of nutrient measurements, we chose the USDA, the authoritative source of food composition data in the United States, considered the gold standard for measurement reliability among national food composition databases[4,14]. While Fig. 2 covers 99 nutrients catalogued by NHANES, we performed robustness checks by testing the validity of equations (2) and (3) for 184 nutrients in the extended panel of the USDA Standard Reference[15], 108 nutrients catalogued by FRIDA[16], and 498 polyphenols in PhenolExplorer[17] and FooDB[18] (Supplementary Section 6), finding that equations (2) and (3) accurately describe the concentrations of all chemicals currently tracked in food. Leveraging the data provided by Foundation Foods[19], we additionally tested the robustness of equations (2) and (3) when sample variability is included, confirming our empirical findings and demonstrating that nutrient fluctuations across different foods are distinguishable from sample variability within the same ingredient and potential measurement errors (Supplementary Section 7). These results indicate that the eight orders of magnitude spanned by nutrient concentrations are not driven by detection limits and the lack of sensitivity in nutrient profiling, but are rooted in the diversity of the physicochemical properties[20] of the nutrients and in the metabolic processes responsible for their modulation.

Taken together, we find that nutrient concentrations in the food supply follow a single family of distributions that depend only on a single parameter, the average concentration μ_n of nutrient n across all foods. Equations (2) and (3) represent our main result, unveiling the existence of a deep universality in nutrient composition in all food. This raises a fundamental question: why do all nutrients follow a similar $\mathcal{Q}(x_n)$?

Biochemical origins of nutrient scaling. The majority of the nutrients in our diet are synthesized by living organisms[21]. Yet, the phylogenetic diversity of the plant and animal products constituting our diet results in well-documented differences in their ability to synthesize and modulate specific nutrients[22], explaining the higher concentrations of selected nutrients in some food groups. The finding that all nutrients follow a similar $\mathcal{Q}(x_n)$ leads to the hypothesis that the observed nutrient variability is not organism-specific or pathway-specific, but it is rooted in the fundamental nature of the biochemical processes responsible for nutrient production and accumulation. Starting from this hypothesis, we next derive equations (2) and (3) from the elementary biochemical principles governing metabolic networks, allowing us to quantitatively link the observed variability in nutrient concentrations to the experimentally determined kinetic constants of individual biochemical reactions. We first investigate the stochasticity characterizing nutrient concentrations within the same cell and organism, and then move on to identify the sources of nutrient variability across organisms[23].

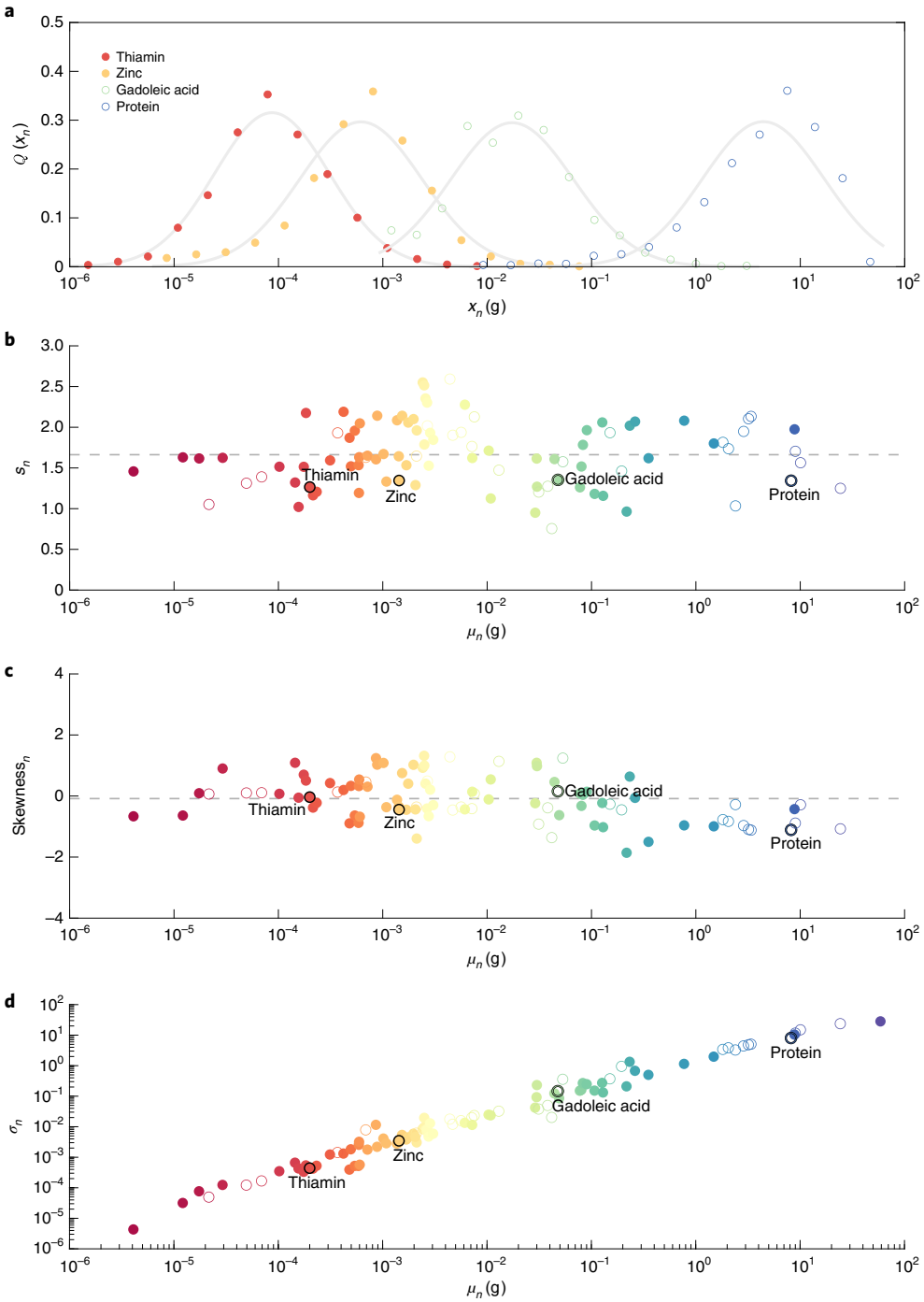

Fig. 2 | The nutrient content across the food supply. a, The concentration distribution $\mathcal{Q}(x_n)$ for four nutrients across the 4,889 foods reported in NHANES, shown on a logarithmic horizontal axis. The four distributions are approximately symmetric on a log scale and have similar width and shape that are independent of the average concentration of the respective nutrient. Similarly to Fig. 1c,d, each symbol represents a histogram bin. For the sake of simplicity, we refer to monounsaturated fatty acid 20:1 with the common name of the most typical isomer (that is, gadoleic acid). **b**, The logarithmic standard deviation s_n of $\mathcal{Q}(x_n)$, representing the shape parameter of the distribution. We find that $s_n = 1.66 \pm 0.39$ (grey dashed line), largely independent of the nutrient concentrations. **c**, Skewness of $\mathcal{Q}(x_n)$ in log space, measuring the asymmetry of the distribution. The fact that skewness fluctuates near zero indicates that $\mathcal{Q}(x_n)$ is symmetric in log space. **d**, The dependence of the standard deviation, σ_n, on the average nutrient amount, μ_n. For each of the 99 nutrients, we calculated μ_n and σ_n across 4,889 foods consumed in NHANES. The plot indicates that $\sigma_n = e^{\alpha_\sigma}(\mu_n)^{\beta_\sigma}$ across eight orders of magnitude, with $\beta_\sigma = 0.94(0.91, 0.97)$ and $\alpha_\sigma = 0.56(0.38, 0.74)$, in line with prediction (3). For the statistical analysis of the fit, see Supplementary Section 2. In **a**–**d**, nutrients clearly assigned to a unique standard chemical structure identifier (InChI) are represented by filled circles, and composite nutrients (for example, total fat, protein and sugars) are shown as empty circles.

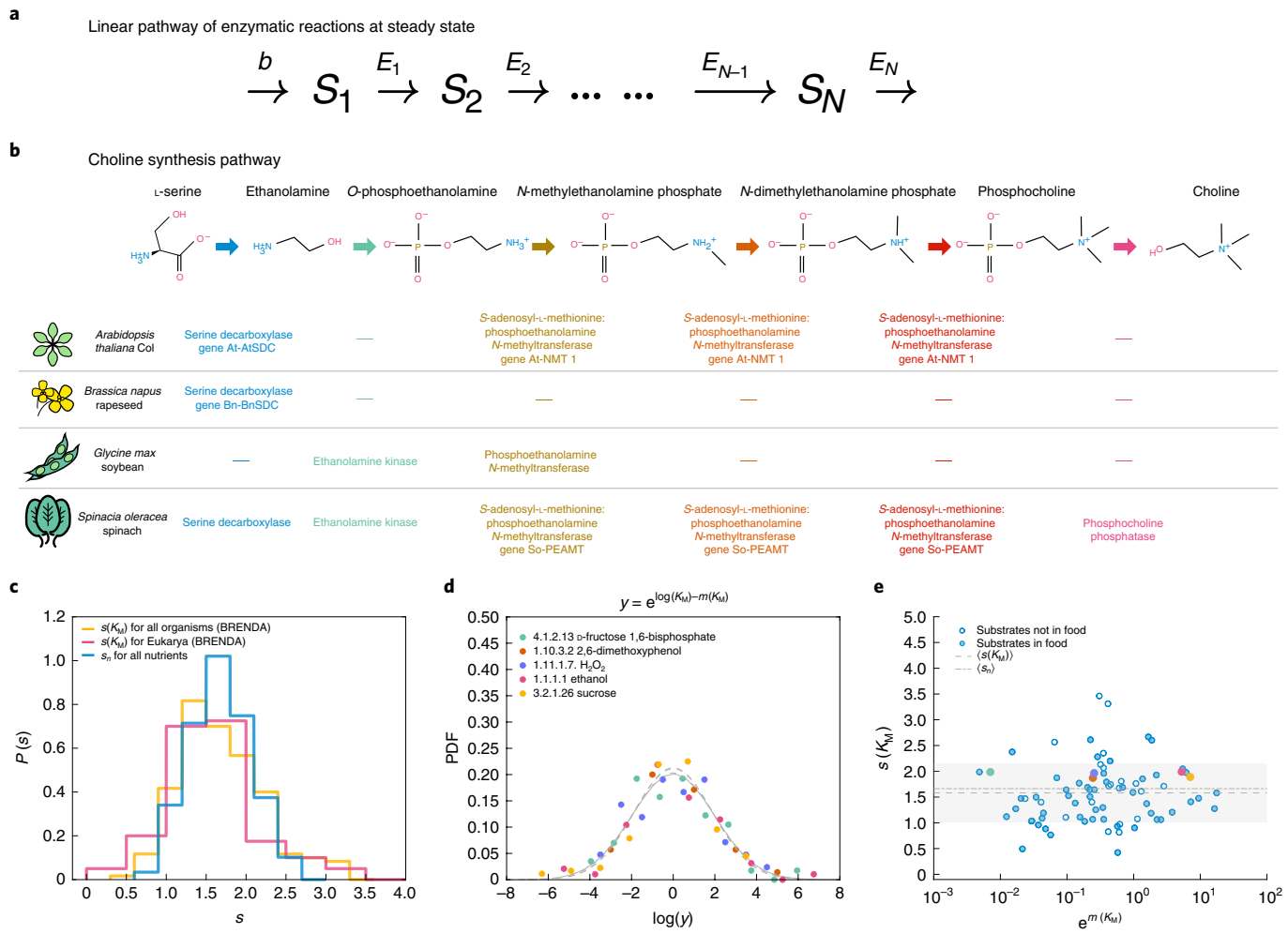

Fig. 3 | Metabolic origins of nutrient scaling. a, We model the dynamics of a biochemical pathway at the steady state as a directed chain of $i=1,\ldots,N$ metabolites, each reaction following Michaelis–Menten kinetics with constant K_M^i and maximal rate $\nu_{\max}^i$ catalysed at each step by an enzyme E_i. We consider the influx of substrate molecules to the metabolic pathway to be a Poisson process with rate b. **b**, Choline biosynthesis pathway and the associated enzymes in four different plants. In each organism, the pathway consists of the same set of reactions, but the reactions are catalysed by different enzymes, characterized by different Michaelis–Menten constants K_M^i. A dash indicates that the enzyme is not reported in the corresponding organism. **c**, We collected data for 93,692 experiments measuring K_M in BRENDA to quantify K_M fluctuations for a fixed enzyme–substrate pair (E_i, S_i) across different organisms. The logarithmic standard deviation $s\left(K_M^i\right)$ behaves similarly to the nutrient logarithmic standard deviation s_n, an independent measure derived from food composition databases. We illustrate the agreement between s_n and $s\left(K_M^i\right)$ by plotting $P(s_n)$ for all nutrients, together with $P\left[s\left(K_M^i\right)\right]$ for all pairs (E_i, S_i) in BRENDA, and separately for eukaryotes, given their direct food relevance, finding that the three distributions are indistinguishable. **d**, From the obtained 31,662 enzyme–substrate pairs (E_i, S_i) in BRENDA, we focused on the experimental measurements for the same enzyme across different eukaryotes, finding that $p\left(K_M^i\right)$ is well approximated by the log-normal distribution in equation (6) with parameters $m\left(K_M^i\right)$ and $s\left(K_M^i\right)$. The bounded nature of $s\left(K_M^i\right)$ implies the collapse of equation (6) for different enzyme–substrate pairs (E_i, S_i) on a single universal distribution corresponding to the rescaling $y = e^{\log(K_M^i)-m(K_M^i)}$. The plot shows the functional collapse for five enzyme–substrate pairs characterized by different orders of magnitude of the corresponding K_M. PDF, probability density function. **e**, The logarithmic standard deviation $s(K_M)$ in equations (6)–(8) is bounded, fluctuating around $s\left(K_M^i\right) = 1.58 \pm 0.57$ (dashed line with grey bands), largely independent of the magnitude of K_M. The observed value agrees within the error bars with $s_n = 1.66 \pm 0.39$ (dashed-dotted line), capturing the variability of nutrient concentrations across all food. Foodborne chemicals are represented by filled circles.

While the metabolic network is a complex crosslinked network of chemical reactions, it can be decomposed into simpler motifs, consisting of a linear array of metabolites linked by chemical reactions, connected to each other via converging pathways and diverging branch points[24–26]. The direction of each pathway is defined by the energetics of the individual reactions. We model the dynamics of each biochemical pathway as a directed chain of $i=1,\ldots,N$ metabolites, catalysed at each step by an enzyme E_i. Each reaction follows Michaelis–Menten kinetics with constant K_M^i and maximal rate $\nu_{\max}^i$ (Fig. 3a). This model allows us to analytically derive, at the steady state (ss), the probability $p^{ss}(n_i)$ of observing n_i molecules of intermediate metabolite i. The calculations indicate that $p^{ss}(n_i)$ follows a negative binomial distribution:

$$p^{ss}(n_i) = \binom{n_i + K_M^i}{n_i} (r_i)^{n_i} (1-r_i)^{K_M^i+1}, \qquad (4)$$

where $r_i = \frac{b}{\nu_{\max}^i}$ is the likelihood of observing a metabolite S_i not yet bonded to enzyme E_i, and b is the incoming flux to the first reaction

of the considered reaction chain. As we derive in Supplementary Section 8, an equation similar to equation (4) describes linear pathways with reversible links and with feedback control, cyclic and converging pathways, and even pathways in which flux conservation is violated[27].

Under physiological conditions, the enzymes are not saturated with substrates; hence, the ratio between substrate concentration and K_M is typically in the range of 0.01 and 1.0 (ref. [28]), where under saturation the ratio converges to infinity. This implies that typically $n_i < K_M^i$, in which case equation (4) can be approximated by the Poisson distribution

$$p^{ss}(n_i) \approx \frac{1}{n_i!}\left(r_i K_M^i\right)^{n_i} e^{-r_i K_M^i}. \quad (5)$$

To derive equations (2) and (3), we are not interested in the variations of metabolite concentrations within the same organisms, as captured by equation (5). Rather, we need to determine the distribution of n_i across the many different organisms we consume. In this case, the dominant source of variability is rooted in the different Michaelis–Menten kinetic constants K_M^i, which can vary by several orders of magnitude across organisms. As our ability to quantify the variability of r_i across organisms is currently limited by data availability, we replace r_i with its average value across different organisms (Supplementary Section 8).

The conservation and evolutionary modularity of metabolic networks imply that when a metabolite is present in multiple organisms, it tends to be produced and consumed by similar sets of chemical reactions[24,29,30]. This is illustrated in Fig. 3b, where we show the six reactions responsible for choline synthesis in four plants[31]. While the chemical reactions are identical, each organism has its own enzyme to catalyse the reaction consuming metabolite *i*. These enzymes are often orthologues and are even assigned to the same EC number in databases, but they do have imperfect homology, reflecting the different evolutionary and selection processes of the organisms (foods) that carry them. Hence, these orthologous enzymes have different constants K_M^i, whose variations determine the dispersion of the distribution derived in equation (5), when different organisms are considered. We therefore need to ask how K_M varies across all organisms that contain the same chemical reaction. We collected data for 93,692 experiments measuring K_M for multiple organisms, as reported in BRENDA[8] (Supplementary Section 9). From the obtained 31,662 enzyme–substrate pairs (E_i, S_i), we focused on experimental measurements for the same enzyme across different eukaryotes, obtaining the $p\left(K_M^i\right)$ distribution across organisms (Fig. 3c). After testing multiple distributions, we find that the log-normal distribution

$$p\left(K_M^i\right) = \frac{1}{K_M^i s\left(K_M^i\right)\sqrt{2\pi}} e^{-\frac{\left[\log K_M^i - m\left(K_M^i\right)\right]^2}{2 s\left(K_M^i\right)^2}} \quad (6)$$

again offers the best approximation (Fig. 3d) (see Supplementary Section 9 for statistical validation). Formally, this implies that the number of molecules of metabolite S_i across different organisms must follow the Poisson–log-normal form[32]

$$p^{\text{organisms}}(n_i) \approx \int_0^\infty \frac{1}{n_i!}\left(r_i K_M^i\right)^{n_i} e^{-r_i K_M^i} \frac{1}{K_M^i s\left(K_M^i\right)\sqrt{2\pi}} e^{-\frac{\left[\log K_M^i - m\left(K_M^i\right)\right]^2}{2 s\left(K_M^i\right)^2}} dK_M^i. \quad (7)$$

For large n_i, the leading terms contributing to equation (7) can be expanded into Taylor series[32], allowing us to formally derive equation (2):

$$\mathcal{Q}(n_i) \approx \frac{1}{n_i s\left(K_M^i\right)\sqrt{2\pi}} e^{-\frac{\left(\log n_i - \log\left(r_i e^{m\left(K_M^i\right)}\right)\right)^2}{2 s\left(K_M^i\right)^2}} \left[1 + \frac{1}{2 n_i s\left(K_M^i\right)^2}\left(\frac{\left(\log n_i - \log\left(r_i e^{m\left(K_M^i\right)}\right)\right)^2}{s\left(K_M^i\right)^2} + \log n_i - \log\left(r_i e^{m\left(K_M^i\right)}\right) - 1\right)\right], \quad (8)$$

predicting that the fluctuations in the steady state concentrations of the individual metabolites across organisms follow a log-normal distribution whose logarithmic mean $m_i = \log(r_i) + m\left(K_M^i\right)$ and standard deviation $s_i = s\left(K_M^i\right)$ are determined by the behaviour of Michaelis–Menten constants across organisms. We expect this behaviour to hold even when enzymes are saturated with substrates—that is, beyond the Poisson regime explored above (Supplementary Section 8).

The probability of observing *x* grams of nutrient *n* per 100 grams of food in equation (2) is connected to the probability of finding n_i substrate molecules in equation (8) through a rescaling by a unit of mass. However, this normalization affects only the parameter m_i, leaving the logarithmic standard deviation $s\left(K_M^i\right)$ unaltered, allowing us to predict that $s_n \sim s\left(K_M^i\right)$, which formally links the variability of the nutrient concentrations $\mathcal{Q}(x_n)$ in equations (2) and (3) to the observed variability in the kinetic constants $s\left(K_M^i\right)$. To validate this prediction, we measured $s\left(K_M^i\right)$. We find that while the observed values of K_M span four orders of magnitude, $s\left(K_M^i\right)$ is bounded, with $s\left(K_M^i\right) = 1.58 \pm 0.57$, a value that is in numerical agreement with $s_n = 1.66 \pm 0.39$, characterizing the variability of nutrient concentrations in food, as observed earlier (Fig. 3c,e). As additional evidence, we also observe the collapse of equation (6) for different enzyme–substrate pairs (E_i, S_i) on a single universal distribution corresponding to the rescaling $y = e^{\log\left(K_M^i\right) - m\left(K_M^i\right)}$ (Fig. 3d and Supplementary Section 9). Finally, the agreement between s_n and $s\left(K_M^i\right)$ is best illustrated in Fig. 3c, where we show $P(s_n)$ for all nutrients, compared with $P\left[s\left(K_M^i\right)\right]$ for all pairs (E_i, S_i) in BRENDA, as well as separately for eukaryotes (given their relevance for the food that humans consume), finding that the three distributions are indistinguishable. Taken together, we find that the observed log-normal distribution described by equations (2) and (3), capturing the variability of nutrient concentrations in all food, can be formally reduced to the variability of the kinetic constants responsible for the regulation of these nutrients in different organisms. This derived mapping not only analytically predicts the log-normal form but also allows us to independently derive the variability s_n from chemical principles, in quantitative agreement with the data. Note, however, that multiple mechanisms can affect the functional form and the extent of the feasible nutrient fluctuations, including network effects[33,34], volumetric costs related to the limited solvent capacity of cellular compartments[35–37], osmotic concentration[23] and physical properties such as substrate molecular mass, hydrophobicity and charge[20,34], effects whose impacts on log-normality and the constrained logarithmic standard deviation $s\left(K_M^i\right)$ remain to be addressed by future work.

Discussion

We find that nutrient concentrations in the food supply closely follow equations (2) and (3), a universality rooted in the nature of the biochemical processes governing nutrient synthesis and regulation. Nutrients, however, represent only a subset of the several thousands of chemicals carried by food[3,4,18], most of which remain unquantified

in all but few ingredients. The universality of equations (2) and (3) can therefore help us estimate the concentrations of these unquantified chemicals from limited data and ultimately unveil our exposure to them through diet. Indeed, the existence of a single functional form for nutrient distributions, as documented above, has multiple benefits for prediction purposes. First, the presence of a specific distribution, with known average, variability and extreme values, offers a way to quantify the completeness of food composition databases and a mathematical rationale for imputing missing quantities. Second, the universality of equations (2) and (3) suggests that measured peak intensities provided by mass spectrometry techniques could allow us to analytically predict chemical concentrations from mass spectrometry data, once ionization efficiency is correctly factored in, a procedure that previously was possible only with the use of dedicated standards, which is costly and time consuming.

Protein number variations consistent with log-normal and related distributions (for example, gamma) have been observed before for individual proteins in yeast and *Escherichia coli* (Supplementary Section 10)[38–41]. Note that these efforts capture copy number variations between individual cells of the same organism, rather than variability across foods[42] described by equations (2) and (8) above. Unfortunately, food composition databases approximate the concentrations of all proteins under a single data point, listed as 'protein' in Figs. 1 and 2. Further high-resolution proteomics approaches focusing on food are therefore needed to resolve whether the observed variability applies to individual proteins in our diet. The variability documented in Fig. 2 may also hide deeper links to variations in metabolic rates across organisms captured by allometric scaling[43–46]. Combining the empirical observations reported here with fundamental work in metabolic processes may open avenues towards a better understanding of the impact of the chemical balance of our diet on health.

Food processing is known to change the nutrient balance by altering the concentration of the native nutrients and through the addition of salt, sugars, fats and numerous additives. These perturbations have known health implications[47]: recent epidemiological studies have found that many of the known health effects traditionally attributed to meat and fat consumption are rooted in the consumption of processed meat, associated with 42% higher risk of coronary heart disease and 19% higher risk of diabetes mellitus[48]. Overall, an increased proportion of ultra-processed foods in an individual's diet leads to increased risk of cancer[49], depressive symptoms[50] and increased telomere length, a biomarker for biological age that is known to be affected by diet through inflammation mechanisms and oxidation[51]. These epidemiological outcomes suggest that human metabolism is adapted to the nutrient range characterizing naturally occurring ingredients, as described by equations (2) and (3), in line with the expectation that contemporary humans are genetically adapted to the environment their ancestors survived in, conditioning their genetic makeup and metabolic tolerance for specific types of diet[47,52]. Yet the resilience of metabolic processes may have its limits[53]: it may not be able to process nutrient concentrations that substantially deviate from the natural range defined by equations (2) and (3). Indeed, the stoichiometric constraints of each biochemical reaction[25,54–56] limit the metabolism's ability to process chemicals whose relative concentrations to other nutrients are unbalanced. As concentration variations are common in processed food, representing more than 60% of caloric intake in US diets[57], understanding the natural variability of nutrient concentrations could open avenues to unveil the origins of the health effects caused by processed food[58].

Methods

Food composition data. To construct the food supply matrix F_{nd}, we started from FNDDS, a food composition database collecting foods and beverages as consumed by the US population, and compiled by NHANES. Designed for the analysis of dietary intake data, FNDDS has no missing nutrient values (in contrast to the USDA Standard Reference Legacy and Foundation Foods)[9,10,59]. We focused on the cycle 2009–2010, which includes a flavonoids database that extends the nutritional panel to 102 nutrients[10] and captures the diet of 8,278 individuals over two 24-hour recalls (from which we excluded breast-feeding babies), for a total of 4,889 food items consumed over two days. We kept all nutrients measured in g, mg or µg, dropping "Energy", "Folate, DFE" and "Vitamin A, RAE", resulting in 99 nutrients, converted to grams (Supplementary Section 1).

Validation of the probability distribution. The collection of foods profiled by the USDA has many items that are similar or identical in their nutrient composition, creating batch effects, an issue for standard statistical tests used to fit probability distributions. We defined a heuristic to assess the best-suited distribution for $\mathcal{Q}(x_n)$, designed to be consistent with the empirical observations listed in the Results. We started with the log-normal, gamma, Weibull, truncated Gaussian and uniform candidate distributions, representing maximum-entropy distributions with different constraints[60]. We also considered the exponential distribution, a degenerate case of gamma and Weibull, with fixed shape parameter equal to 1. As experimental work related to protein copy number distribution has suggested the relevance of the Fréchet distribution[40], we tested its performance in modelling nutrient properties. We used the Kolmogorov–Smirnov test to assess the performance of the log-normal distribution, not as a measure of the exactness of the fit, given its sensitivity to batch effects and non-random sampling of the data. For the complete analysis, see Supplementary Sections 3 and 10.

Kinetics data. We relied on data from 93,692 experiments reporting K_M for several organisms, as compiled in BRENDA flat files[8]. We applied natural language processing techniques on the free text comments describing each publication to extract temperature and pH and removed all mutant and recombinant enzymes, keeping 70,873 experimental records measured in mM. Additionally, we leveraged NCBI Taxonomy[61] and the ETE 3 package[62] to classify into taxa all organisms reported in the database. To identify which substrates are found in food, we mapped the InChIKey of each molecule (if available) to our manually curated library of food molecules, containing 89,038 compounds as December 2020, reported by different food composition databases such as FooDB[18] or Dictionary of Food Compounds[63] or detected in mass spectrometry experiments. Most of the annotations in our library determine the presence or absence of a compound in food but do not quantify its concentration. From the obtained 31,662 enzyme–substrate pairs (E_i, S_i), we grouped the experimental measurements for the same enzyme–substrate across different eukaryotes, obtaining $p\left(K_M^i\right)$ (Fig. 3c). For further details, see Supplementary Section 9.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The raw data are available at https://github.com/menicgiulia/FoodLaws. Source data are provided with this paper.

Code availability

The processing codes are available at https://github.com/menicgiulia/FoodLaws.

Received: 2 February 2021; Accepted: 8 April 2022;
Published online: 24 May 2022

References

1. Kubo, R., Ichimura, H., Usui, T. & Hashitsume, N. *Statistical Mechanics* (North-Holland Personal Library, 1990).
2. Barabasi, A.-L. & Pósfai, M. *Network Science by Albert-László Barabási* (Cambridge University Press, 2016).
3. Barabási, A.-L., Menichetti, G. & Loscalzo, J. The nutritional dark matter: the unmapped chemical complexity of our diet. *Nat. Food* https://doi.org/10.1038/s43016-019-0005-1 (2019).
4. Hooton, F., Menichetti, G. & Barabási, A.-L. Exploring food contents in scientific literature with FoodMine. *Sci. Rep.* **10**, 16191 (2020).
5. Milanlouei, S. et al. A systematic comprehensive lÿongitudinal evaluation of dietary factors associated with acute myocardial infarction and fatal coronary heart disease. *Nat. Commun.* **11**, 6074 (2020).
6. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabasi, A. L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
7. *List of EuroFIR Databases* (EuroFIR, accessed 7 January 2021); https://www.eurofir.org/food-information/food-composition-databases/
8. Placzek, S. et al. BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res.* **45**, D380–D388 (2017).
9. *USDA Food and Nutrient Database for Dietary Studies* Version 5.0 (USDA, 2012); http://www.ars.usda.gov/ba/bhnrc/fsrg
10. Sebastian, R. et al. *Flavonoid Values for USDA Survey Foods and Beverages 2007–2010* (USDA, 2016); www.ars.usda.gov/nea/bhnrc/fsrg

11. Willett, W. *Monographs in Epidemiology and Biostatistics: Nutritional Epidemiology* Vol. 15 (Oxford Univ. Press, 1990).
12. Hansen, A. The three extreme value distributions: an introductory review. *Front. Phys.* **8**, 533 (2020).
13. Limpert, E., Stahel, W. A. & Abbt, M. Log-normal distributions across the sciences: keys and clues. *Bioscience* **51**, 341–352 (2001).
14. *FoodData Central* (US Department of Agriculture, Agricultural Research Service, 2019); https://fdc.nal.usda.gov/
15. *National Nutrient Database for Standard Reference, Release 28, Documentation and User Guide* (USDA, 2015).
16. *Frida Fooddata* Version 2 (DTU Food, 2016).
17. Neveu, V. et al. Phenol-Explorer: an online comprehensive database on polyphenol contents in foods. *Database (Oxford)* **2010**, bap024 (2010).
18. *WishartLab* (FooDB, 2017); http://foodb.ca/
19. *FoodData Central: Foundation Foods* (U.S. Department of Agriculture, A. R. S., 2019); https://fdc.nal.usda.gov/
20. Bar-Even, A., Noor, E., Flamholz, A., Buescher, J. M. & Milo, R. Hydrophobicity and charge shape cellular metabolite concentrations. *PLoS Comput. Biol.* **7**, e1002166 (2011).
21. Muchowska, K.B., Varma, S.J. & Moran, J. Synthesis and breakdown of universal metabolic precursors promoted by iron. *Nature* **569**, 104–107 (2019).
22. Chae, L., Kim, T., Nilo-Poyanco, R. & Rhee, S. Y. Genomic signatures of specialized metabolism in plants. *Science* **344**, 510–513 (2014).
23. Park, J. O. et al. Metabolite concentrations, fluxes and free energies imply efficient enzyme usage. *Nat. Chem. Biol.* **12**, 482–489 (2016).
24. Michal, G. & Schomburg, D. *Biochemical Pathways: An Atlas of Biochemistry and Molecular Biology* 2nd edn (John Wiley & Sons, 2013); https://doi.org/10.1002/9781118657072
25. Almaas, E., Kovács, B., Vicsek, T., Oltvai, Z. N. & Barabási, A.-L. Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature* **427**, 839–843 (2004).
26. Almaas, E., Oltvai, Z. N. & Barabási, A. L. The activity reaction core and plasticity of metabolic networks. *PLoS Comput. Biol.* **1**, 0557–0563 (2005).
27. Levine, E. & Hwa, T. Stochastic fluctuations in metabolic pathways. *Proc. Natl Acad. Sci. USA* **104**, 9224–9229 (2007).
28. Stryer, L., Berg, M. J. & Tymoczko, L. J. *Biochemistry* (W. H. Freeman, 2002).
29. Peregrín-Alvarez, J.M., Sanford, C. & Parkinson, J. The conservation and evolutionary modularity of metabolism. *Genome Biol* 10, R63 (2009).
30. Khan, A. H., Zou, Z., Xiang, Y., Chen, S. & Tian, X. L. Conserved signaling pathways genetically associated with longevity across the species. *Biochim. Biophys. Acta Mol. Basis Dis.* **1865**, 1745–1755 (2019).
31. *Plant Metabolic Network* (PlantCyc Pathway: Choline Biosynthesis, 2019); https://pmn.plantcyc.org/PLANT/NEW-IMAGE?type=PATHWAY&object=PWY-3385
32. Bulmer, A. M. G. On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics* **30**, 101–110 (1974).
33. Küken, A., Eloundou-Mbebi, J. M. O., Basler, G. & Nikoloski, Z. Cellular determinants of metabolite concentration ranges. *PLoS Comput. Biol.* **15**, e1006687 (2019).
34. Bar-Even, A. et al. The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry* **50**, 4402–4410 (2011).
35. Dourado, H., Maurino, V. & Lercher, M. Enzymes and substrates are balanced at minimal combined mass concentration in vivo. Preprint at *bioRxiv* https://doi.org/10.1101/128009 (2017).
36. Vazquez, A. et al. Impact of the solvent capacity constraint on *E. coli* metabolism. *BMC Syst. Biol.* **2**, 7 (2008).
37. Vazquez, A. Optimal cytoplasmatic density and flux balance model under macromolecular crowding effects. *J. Theor. Biol.* **264**, 356–359 (2010).
38. Furusawa, C., Suzuki, T., Kashiwagi, A., Yomo, T. & Kaneko, K. Ubiquity of log-normal distributions in intra-cellular reaction dynamic. *Biophysics (Nagoya-shi)* **1**, 25–31 (2005).
39. Beal, J. Biochemical complexity drives log-normal variation in genetic expression. *Eng. Biol.* **1**, 55–60 (2017).
40. Salman, H. et al. Universal protein fluctuations in populations of microorganisms. *Phys. Rev. Lett.* **108**, 238105 (2012).
41. Taniguchi, Y. et al. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–539 (2011).
42. Kærn, M., Elston, T. C., Blake, W. J. & Collins, J. J. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* **6**, 451–464 (2005).
43. Banavar, J. R., Maritan, A. & Rinaldo, A. Size and form in efficient transportation networks. *Nature* **399**, 130–132 (1999).
44. Maritan, A., Rigon, R., Banavar, J. R. & Rinaldo, A. Network allometry. *Geophys. Res. Lett.* **29**, 1508 (2002).
45. Enquist, B. J., Brown, J. H. & West, G. B. Allometric scaling of plant energetics and population density. *Nature* **395**, 163–165 (1998).
46. Gallos, L. K., Song, C., Havlin, S. & Makse, H. A. Scaling theory of transport in complex biological networks. *Proc. Natl Acad. Sci. USA* **104**, 7746–7751 (2007).
47. Cordain, L. et al. Origins and evolution of the Western diet: health implications for the 21st century. *Am. J. Clin. Nutr.* **81**, 341–354 (2005).
48. Micha, R., Wallace, S. K. & Mozaffarian, D. Red and processed meat consumption and risk of incident coronary heart disease, stroke, and diabetes mellitus: a systematic review and meta-analysis. *Circulation* **121**, 2271–2283 (2010).
49. Fiolet, T. et al. Consumption of ultra-processed foods and cancer risk: results from NutriNet-Sant, prospective cohort. *Brit. Med. J.* **360**, k322 (2018).
50. Adjibade, M. et al. Prospective association between ultra-processed food consumption and incident depressive symptoms in the French NutriNet-Sant, cohort. *BMC Med.* **17**, 78 (2019).
51. Alonso-Pedrero, L. et al. Ultra-processed food consumption and the risk of short telomeres in an elderly population of the Seguimiento Universidad de Navarra (SUN) Project. *Am. J. Clin. Nutr.* **111**, 1259–1266 (2020).
52. Carrera-Bastos, P., Fontes-Villalba, M., O'Keefe, J. H., Lindeberg, S. & Cordain, L. The western diet and lifestyle and diseases of civilization. *Res. Rep. Clin. Cardiol.* **2**, 15–35 (2011).
53. Bornholdt, S. & Sneppen, K. Robustness as an evolutionary principle. *Proc. R. Soc. Lond. B* **267**, 2281–2286 (2000).
54. Riehl, W. J., Krapivsky, P. L., Redner, S. & SegrŠ, D. Signatures of arithmetic simplicity in metabolic network architecture. *PLoS Comput. Biol.* **6**, e1000725 (2010).
55. Segré, D., Shenhav, B., Kafri, R. & Lancet, D. The molecular roots of compositional inheritance. *J. Theor. Biol.* **213**, 481–491 (2001).
56. Palsson, B. *Systems Biology: Properties of Reconstructed Networks* (Cambridge Univ. Press, 2006); https://doi.org/10.1017/CBO9780511790515
57. Gupta, S., Hawk, T., Aggarwal, A. & Drewnowski, A. Characterizing ultra-processed foods by energy density, nutrient density, and cost. *Front. Nutr.* **6** (2019).
58. Menichetti, G., Ravandi, B., Mozaffarian, D. & Barabasi, A.-L. Machine learning prediction of food processing. Preprint at *medRxiv* https://doi.org/10.1101/2021.05.22.21257615 (2021).
59. FNDDS Web Page (USDA, 2019); https://www.ars.usda.gov/northeast-area/beltsville-md-bhnrc/beltsville-human-nutrition-research-center/food-surveys-research-group/docs/fndds/
60. Kapur, J. N. *Maximum-entropy Models in Science and Engineering* (India, Wiley, 1989).
61. *NCBI Taxonomy* (National Center for Biotechnology Information, 2019); https://www.ncbi.nlm.nih.gov/taxonomy
62. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
63. Yannai, S. *Dictionary of Food Compounds with CD-ROM* Choice Reviews Online Vol. 51 (Taylor & Francis, 2013).

Acknowledgements

This work was partially supported by NIH grant no. 1P01HL132825, American Heart Association grant no. 151708, ERC grant no. 810115-DYNASET and Rockefeller Foundation grant no. 2019 FOD 026. We thank J. Loscalzo for useful discussions and insights on enzyme kinetics, as well as M. Sebek and S. Ofaim for helping with the chemical classification and disambiguation.

Author contributions

G.M. and A.-L.B. conceived the project and wrote the manuscript. G.M. performed the data query, data integration, statistical analysis and analytical calculations.

Competing interests

A.-L.B. is the founder of Scipher Medicine and Naring Health, companies that explore the use of network-based tools in health, and Datapolis, which focuses on urban data.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43016-022-00511-0.

Correspondence and requests for materials should be addressed to Albert-László Barabási.

Peer review information *Nature Food* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2022

nature research

Corresponding author(s): Albert-Laszlo Barabasi

Last updated by author(s): Feb 19, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All food composition datasets and kinetics datasets used in this paper are publicly available online and properly referenced. The raw data is available at https://github.com/menicgiulia/FoodLaws.
Data analysis	The processing codes are available at https://github.com/menicgiulia/FoodLaws.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The raw data is available at https://github.com/menicgiulia/FoodLaws.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size is determined by the food composition data currently available.
Data exclusions	There was no data exclusion.
Replication	The scaling laws were tested in several food composition databases (see SI), and the analytical theory was tested in BRENDA db.
Randomization	Traditional randomization strategies on cohorts/experiments do not apply to this paper, that presents an analytical model connecting food composition data and enzymatic constants.
Blinding	Traditional blinding strategies on cohorts/experiments do not apply to this paper, that presents an analytical model connecting food composition data and enzymatic constants.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging