nature portfolio

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

Supplementary information

Nutrient concentrations in food display universal behaviour

In the format provided by the authors and unedited

Nutrient concentrations in food display universal behavior

Giulia Menichetti, Albert-László Barabási

Table of Contents

Section S1: The Food Supply F_{nd}

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 by F_{nd}, representing the amount of nutrient *n* in 100 g of each dish *d* (Figure S1A). To characterize the food supply, we start from the National Health and Nutrition Examination Survey (NHANES), a major national study assessing the health and nutritional status of the United States' population. For 2009/2010, the NHANES dietary intake data captures the diet of 8,278 individuals over two 24-hour recalls (from which we excluded breast-feeding babies). NHANES also encodes the food supply, F_{nd}, by offering nutrient composition for 4,889 food items consumed by their cohort over two days (Figure S1A). The foods and beverages as consumed by the U.S. population are collected in the FNDDS database, designed for the analysis of dietary intake data and containing no missing nutrient values (in contrast with the USDA Standard reference database, designed for the dissemination of food composition data)[1–3]. For 2009/2010, the FNDDS database lists the nutritional content of 7,253 foods, of which 4,889 were reported by individuals participating in two 24-hr dietary recalls.

For the years 2007-2010 the USDA developed a flavonoids database for population surveys that extended the original nutritional panel of 65 nutrients to 102[2]. For our analysis we kept all nutrients measured in g, mg or μg, dropping "Energy", "Folate, DFE" and "Vitamin A, RAE"[a], resulting in 99 nutrients, converted to grams (g). Among these nutrients, some represent cumulative measures of chemical families (e.g., "Total Fat") while others have a clearly identified chemical structure, encoded by an InChI identifier[4].

[a] We decided to exclude "Folate, DFE" and "Vitamin A, RAE" to avoid the inferred confounders used in their measurements.

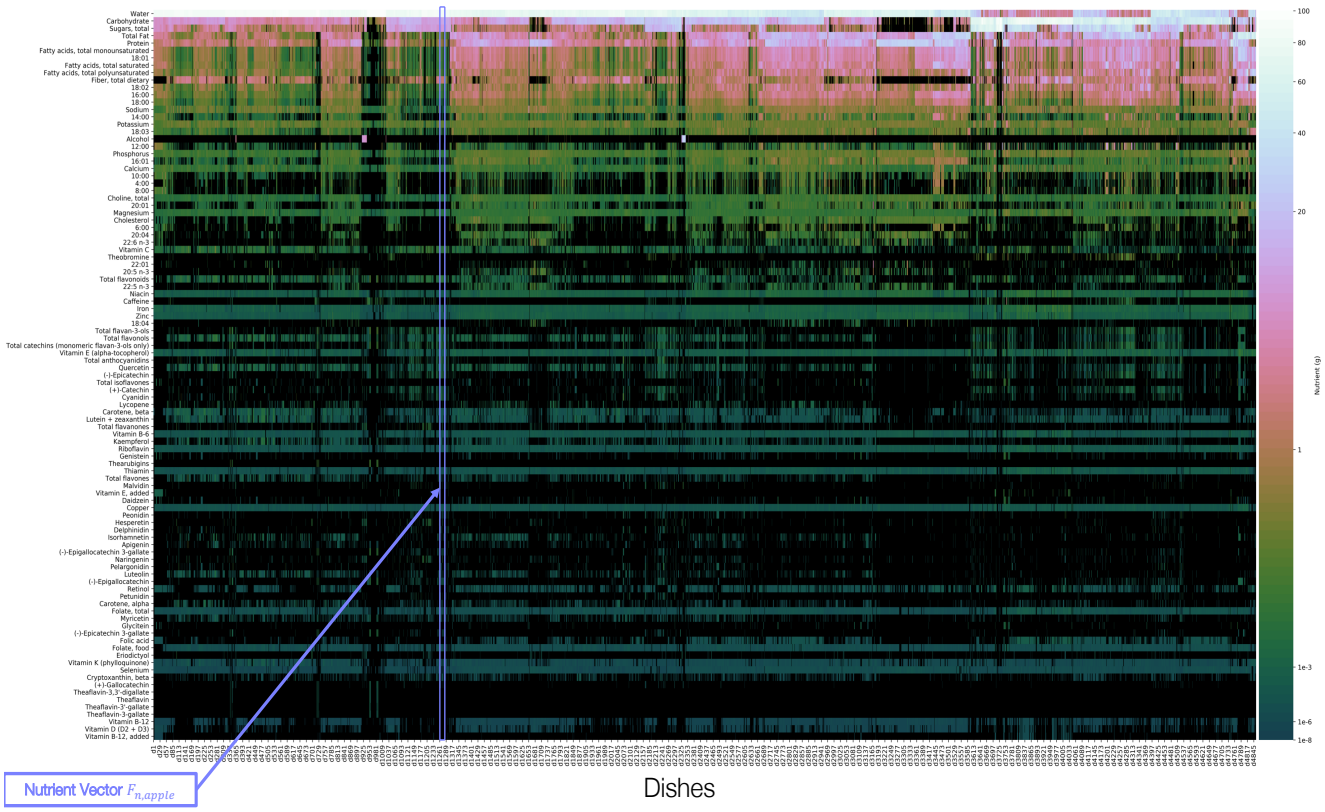


Figure S1: The food supply matrix F_{nd}

Visualization of the food supply matrix F_{nd}. Nutrients are ordered in descending order, according to their average intake in the food supply. The absence of a nutrient (0 g) is reported black.

Section S2: Evaluating the fit results for σ_n vs μ_n scaling

The average content of nutrient *n* derives from the food supply matrix F_{nd} as

$$\mu_n = \frac{1}{D_n}\sum_d F_{nd}\,, \tag{S1}$$

where D_n is the number of foods for which $F_{nd} \neq 0$. The variability of nutrient n across the D_n foods is

$$\sigma_n = \sqrt{\frac{\sum_d (F_{nd} - \mu_n)^2}{D_n - 1}}. \tag{S2}$$

In Figure 2D we show σ_n as a function of μ_n for 99 nutrients catalogued in 4,889 foods by NHANES. The plot indicates that nutrient concentrations span over eight orders of magnitude. Over these eight orders of magnitude σ_n vs μ_n is well approximated by

$$\sigma_n = e^{\alpha_\sigma}(\mu_n)^{\beta_\sigma}, \tag{S3}$$

with $\beta_\sigma = 0.9423\ \ (0.9133, 0.97)$, $\alpha_\sigma = 0.5571$ (0.3760, 0.7382) and adjusted $R^2 = 0.9770$. In other words, the degree of variability of the nutrient concentration across all foods follows a scaling law, telling us that σ_n is uniquely determined by its average concentration. A similar scaling is found over the 7,253 foods in the whole FNDDS, with $\beta_\sigma = 0.9386\ (0.9065, 0.9707)$, $\alpha_\sigma = 0.6696$ (0.4698, 0.8693) and adjusted $R^2 = 0.9717$.

The quality of the fit is influenced by the behavior of nutrients with the highest concentrations, whose abundance and variability are strongly affected by the fixed mass (100 g). We can identify these nutrients by quantifying how the parameters of several statistical distributions change, going from an unbounded fit to a truncated fit. In Figure S2 we show the results for a log-normal fit, where the parameters characterizing the distribution are m_n and s_n

(Figure S2A and B). This analysis points to nine nutrients whose behavior is determined by the boundary constraints: Water, Carbohydrate, Total Fat, Protein, Total Sugars, Alcohol, Total Monosaturated Fatty Acids, Fatty Acids 18:1 and Total Polysaturated Fatty Acids. As expected, we find the major cumulative nutrient measures, usually reported as nutrition facts, and two nutrients strongly affected by food processing, Alcohol, and 18:1, the most common group of fatty acid isomers, present in our diet thanks to the frequent addition of plant and animal fats to processed food. Among these nutrients, water has the most extreme behavior. Indeed, its truncated fit does not find convergence, given how close the body of the distribution is to 100 g. If we remove these nine nutrients as we evaluate the scaling in Eq. (S3), we obtain $\beta_\sigma = 0.9676\ (0.9308, 1.004)$, $\alpha_\sigma = 0.7301\ (0.4907, 0.9696)$ and adjusted $R^2 = 0.9684$, i.e., the relation between σ_n and μ_n becomes linear (Figure S2D). Given the distinctive behavior of water, we excluded it from the subsequent analysis.

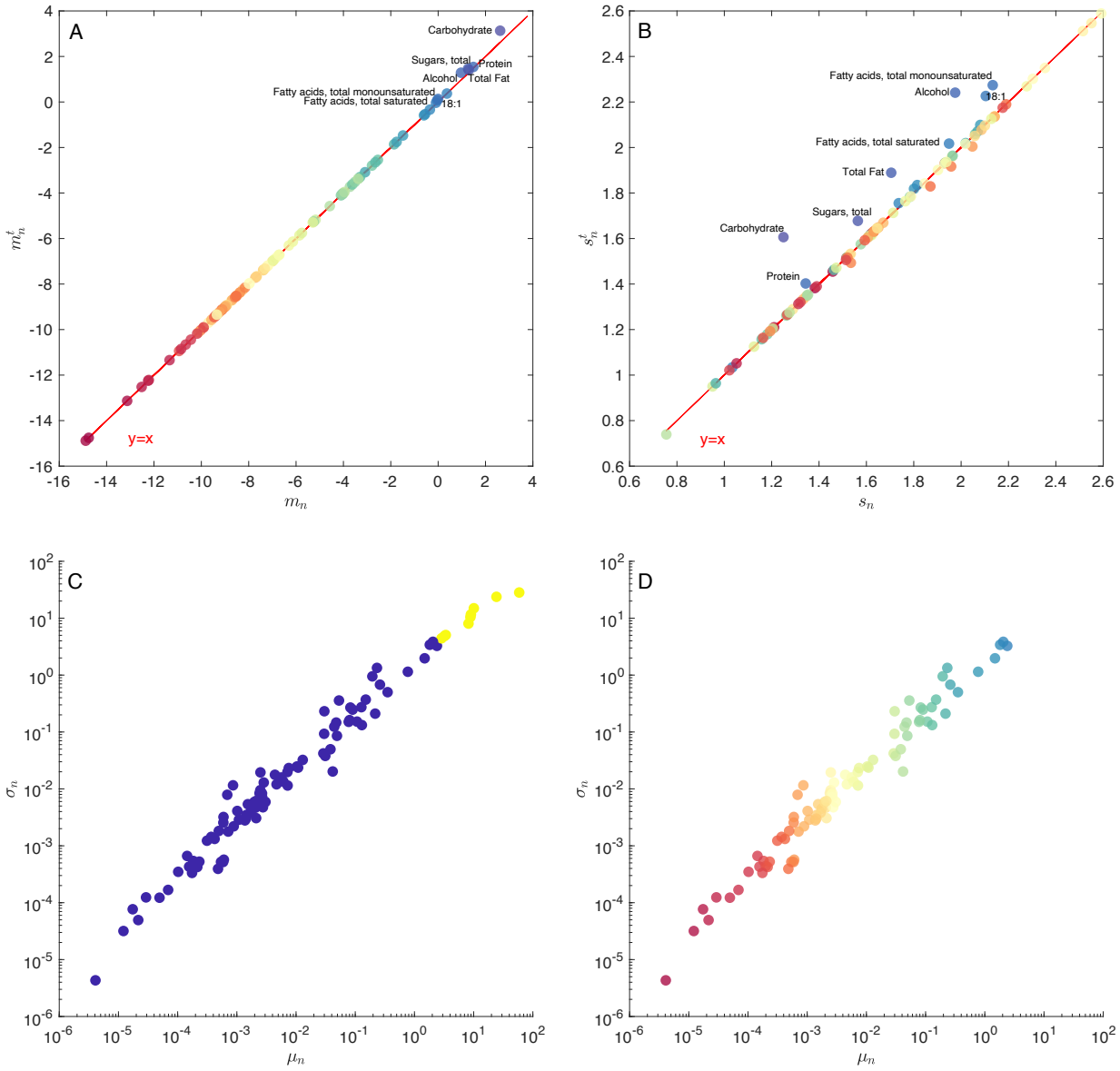


Figure S2. Boundary effect on scaling.

The nutrients are color-coded in agreement with Figure 2.

(A) Comparison of the log-normal parameter m_n *with the corresponding parameter* m_n^t*, determined by a log-normal fit with truncation at 100 g.*

(B) Comparison of the log-normal parameter s_n *with the corresponding parameter* s_n^t*, determined by a log-normal fit with truncation at 100 g.*

(C) The dependence of the standard deviation σ_n *on the average nutrient amount* μ_n*. In yellow the 9 nutrients whose statistical properties are significantly affected by the boundary conditions: Water, Carbohydrate, Total Fat, Protein, Total Sugars, Alcohol, Total Monosaturated Fatty Acids, Fatty Acid 18:1 and Total Polysaturated Fatty Acids.*

(D) The dependence of the standard deviation σ_n *on the average nutrient amount* μ_n*. After the removal of the 9 nutrients strongly affected by the boundary conditions, the relation becomes linear, with* $\beta_\sigma = 0.9676\ (0.9308, 1.004)$ *and* $\alpha_\sigma = 0.7301\ (0.4907, 0.9696)$.

Section S3: Probability distribution for nutrient concentrations

In this section we discuss the statistical evidences supporting the log-normal distribution for nutrient distributions. We evaluated the statistical performance of each distribution from multiple aspects: fitting the nutrient data, reproducing the scaling between μ_n and σ_n, in approximating the statistical observations in the logarithmic space, and as a plausible stationary distribution derived from a generative mechanism.

The inspection of Figure 2A suggests that the candidate distributions should belong to the scale family of probability distributions. This family collects all distributions closed under scaling by a constant factor c, meaning that, by applying the scale transformation $x' = cx$, we find the same probability distribution, rescaled as $P(x') = P(x/c)/c$ (translational invariance in the logarithmic space).

The evidences in Figure 2D and Section S2, supporting a linear relationship between μ_n and σ_n, suggest that the nutrient distributions $Q(x_n)$ belong to an unique parametric family of distributions characterizing the whole nutrient panel and determining the observed linearity by means of varying parameters. Indeed, when a random variable x is multiplied by a constant factor c, the average and the standard deviation transform as $\mu \to c\mu$ and $\sigma \to c\sigma$, respectively. When multiple random variables are generated with different c, the observed relation between μ and σ will be linear, with an angular coefficient fixed by the coefficient of variation $\frac{\sigma}{\mu}$ of the original distribution. According to the family of distributions, the transformation $x' = cx$ will affect different parameters:

- **Location parameter:** The variable determining the position of the probability distribution on the *x*-axis. This parameter linearly scales with c.
- **Scale parameter:** The variable responsible for stretching or squeezing the probability distribution. This parameter linearly scales with c.

- **Shape parameter:** The variable affecting the general shape of the distribution. This parameter is invariant to any scalar multiplication of the random variable.

We focus on five candidate distributions: log-normal, gamma, weibull, truncated gaussian and uniform (Figure S3), representing maximum-entropy distributions with different constraints[5]. Multiple mechanisms were proposed to generate them: gamma distribution describes stationary protein concentrations in stochastic bursting[6] (see Section S10), weibull is derived for particle size distributions and fragmentation processes[7], truncated gaussian and log-normal are the limit distributions of additive and multiplicative processes[8]. We also consider the exponential distribution, a degenerate case of gamma and weibull, with fixed shape parameter equal to 1.

In Figure S4 we evaluated the goodness of the fit by calculating the Kolmogorov-Smirnov (K-S) distance for each nutrient, and assessed the global performance of the candidate distribution by plotting the cumulative distribution of the K-S distance. The distributions parametrized by shape parameters have close performances, performing remarkably better than those described by just location and scale parameters. Log-normal distribution offers the best approximation, with most nutrients with Kolmogorov-Smirnov (K-S) distance close to 0.08 and 60% of the nutrients with K-S distance less than 0.1.

Additional arguments come from the comparison between the real scaling between μ_n and σ_n with the predicted scaling derived from the fitted distributions. In Figure S5 we show this comparison by calculating for each nutrient average and standard deviation of a random set of data points, sampled from the fitted distributions. Among the distributions characterized by a shape parameter, log-normal better captures the observed relation between μ_n and σ_n. The approximated linearity of the scaling forces all the shape parameters to have a tight distribution, given the

functional dependence between coefficient of variation and shape parameter (Figure S3 and Figure S6). Exponential, truncated gaussian and uniform distributions fail in approximating the observed intercept (~ coefficient of variation), having constrained coefficient of variation, uniquely determined by their parametrization. Moreover, truncated gaussian and uniform distribution clearly collapse into their degenerated limit cases, half gaussian and uniform with extremes $[0, b]$, as proven by the perfect agreement with the analytical relations $\sigma_n = \sqrt{\frac{\pi-2}{2}}\,\mu_n$ and $\sigma_n = \frac{1}{\sqrt{3}}\,\mu_n$.

Further evidence in support of the log-normal form comes from the inspection of the nutrient distributions in the logarithmic space. The log-transformation of the nutrient concentrations revealed robust patterns, not clearly visible in the original linear space (Figure 2A, 2B and 2C). Indeed, the log-normal distribution best captures the observed translational invariance and symmetry, as shown in Figure S7, where the log-transformed $Q(x_n)$ for Vitamin B-12 and E is compared to a random sample from their fitted distributions.

The transformation $x' = cx$ in the linear space becomes a translation of $log\, c$ for the log-transformed variable $y = log\, x$. As shown in Table S1, a scale parameter in the linear space becomes a location parameter in the log space, while, a shape parameter in the linear space acts as a scale parameter in the log-space. Consequently, the distributions characterized only by a scale parameter in the linear space have fixed standard deviation in the log-space, as documented for uniform, half-gaussian, exponential and half-logistic, assuming defined values not compatible with real data (Table S1). This is clearly shown by Figure S8, where we compare the distribution of the nutrient logarithmic standard deviation with those determined from the fitted distributions, finding again the best performance for the log-normal fit.

As experimental work related to protein copy number distribution has also suggested the relevance of the fréchet distribution[9], we tested the performance of this distribution in modeling

nutrient properties. First, we evaluated the goodness of the fit by calculating the K-S distance for each nutrient, and assessed the global performance of fréchet by plotting the cumulative distribution of the K-S distance together with those of other distributions characterized by shape parameters, such as log-normal, gamma and weibull (Table S1). Fréchet distribution never outperforms the log-normal distribution, as shown in Figure S9A for the whole food supply, and in Figure S9B for raw-plants only, whose nutrient concentrations should directly link to the metabolic processes of the original plants (see Section S4). Additionally, we compared the distribution of logarithmic skewness across nutrients with those determined by random sampling from the fitted distributions. In Figure S9C we show for raw-plant products how the only probability distribution compatible with the observed behavior is the log-normal distribution. Indeed, by applying a two-sided Wilcoxon rank sum test on the skewness distributions, we find that only log-normal has a median consistent with real data (p-value=0.7841), while fréchet, gamma and weibull reject the null hypothesis with probabilities equal to $8.3108\ 10^{-14}$, $1.8462\ 10^{-18}$, and $4.5497\ 10^{-16}$, respectively. These results are consistent with our observations for the whole food supply where log-normal is the only distribution in agreement with real data (p-value=0.2954), while fréchet shows a significant bias towards positive logarithmic skewness (p-value=1.5459e-17), and gamma and weibull towards negative logarithmic skewness, with probability $3.8461\ 10^{-30}$ and $6.9675\ 10^{-23}$, respectively.

Overall, our derivation of (2) and (3) is rooted in the naturally occurring biochemical processes regulating the fluctuations of molecular substrates in different organisms. When we investigate nutrients affected by food processing and fortification, we observe deviations from (2), as many outliers alter the shape of the distribution. However, our analysis based on the logarithmic

nature of nutrient variability minimizes the impact of these deviations, and uncovers the dominant patterns of variance. Moreover, whenever fortification follows the concentration of natural compounds by a renormalization factor, log-normality would be found as consequence. Additional sources of issues for standard statistical tests are the sampling strategies adopted by national agencies and food database curators. Indeed, the collection of foods profiled by the USDA is far from being an ideal independent random sample of the food supply, with very similar or identical items in their nutrient composition, creating batch effects. To partially address these issues in Section S4 we analyzed 108 raw vegetables and fruits, whose nutrient concentrations should directly link to the metabolic processes of the original plants, and from which we removed any redundant nutrient profile (i.e., identical or repeated foods).

We also acknowledge that FNDDS, as other USDA databases, reports representative nutritional average values for each food/drink, which do not capture the variability due to factors such as recipe variations, production methods, soil quality, and storage time. Leveraging the data provided by Foundation Foods[10], we additionally tested the robustness of our analysis 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 food and potential measurement errors (Section S7)

Finally, we use the Kolmogorov-Smirnov test to assess the better performance of the log-normal distribution, and not as a measure of the exactness of the fit, given its sensitivity to batch effects and non-random sampling of the data. The literature presented in Section S10 omits fit discussion or leverages the χ^2 goodness-of-fit test, dependent on the binning choice. In summary, we rely on the empirical observations i)-iii) and our stochastic model to support the validity of the log-normal form as the best candidate for nutrient probability distributions.

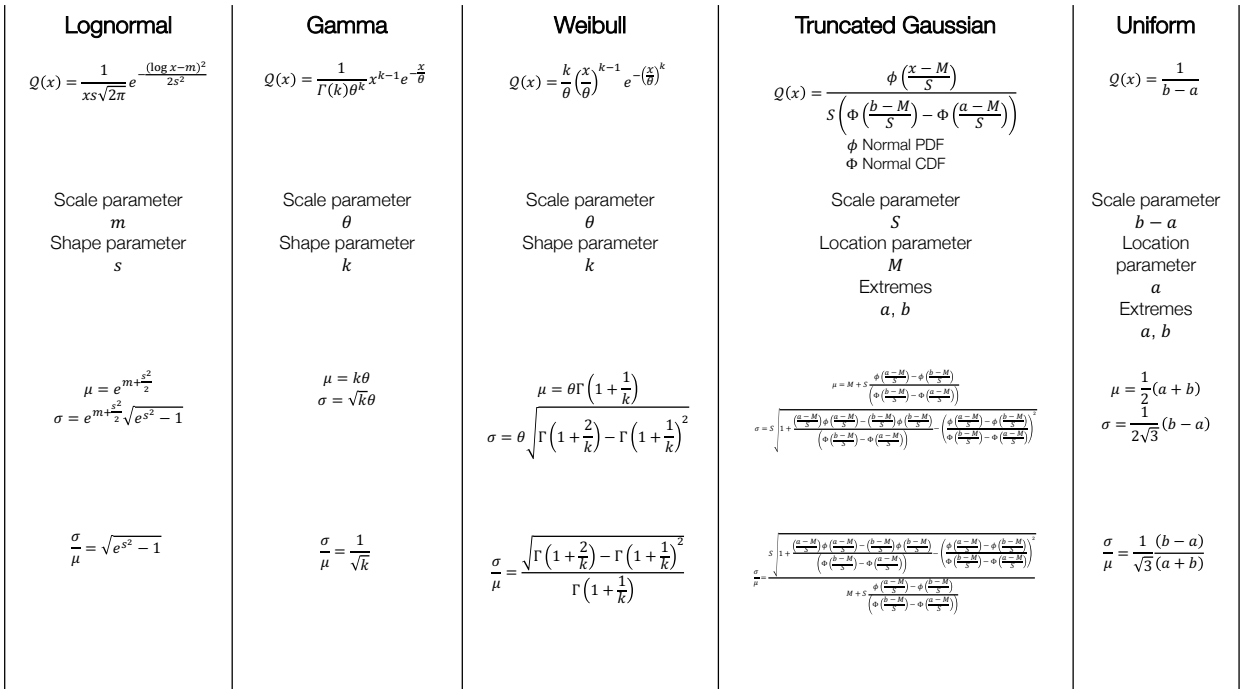


Figure S3. Parametrization of candidate probability distributions.

Log-normal, gamma and weibull are characterized by a scale and a shape parameter, while distributions such as truncated gaussian and uniform have a scale and a location parameter. All the distributions except for the uniform are fitted in the interval [0,100]. For the uniform distribution we estimate the extremes a and b directly from the data, determining the natural truncation of the distribution. The effect of the truncation, as shown in Section S3, was found marginal and affecting a limited number of nutrients. The results for truncated gaussian and uniform distribution systematically collapse to the degenerated distributions half-gaussian (a=0, M=0, b=∞) and uniform on the interval [0,b].

Log-normal, gamma and weibull have a coefficient of variation σ/μ that depends only on the shape parameter.

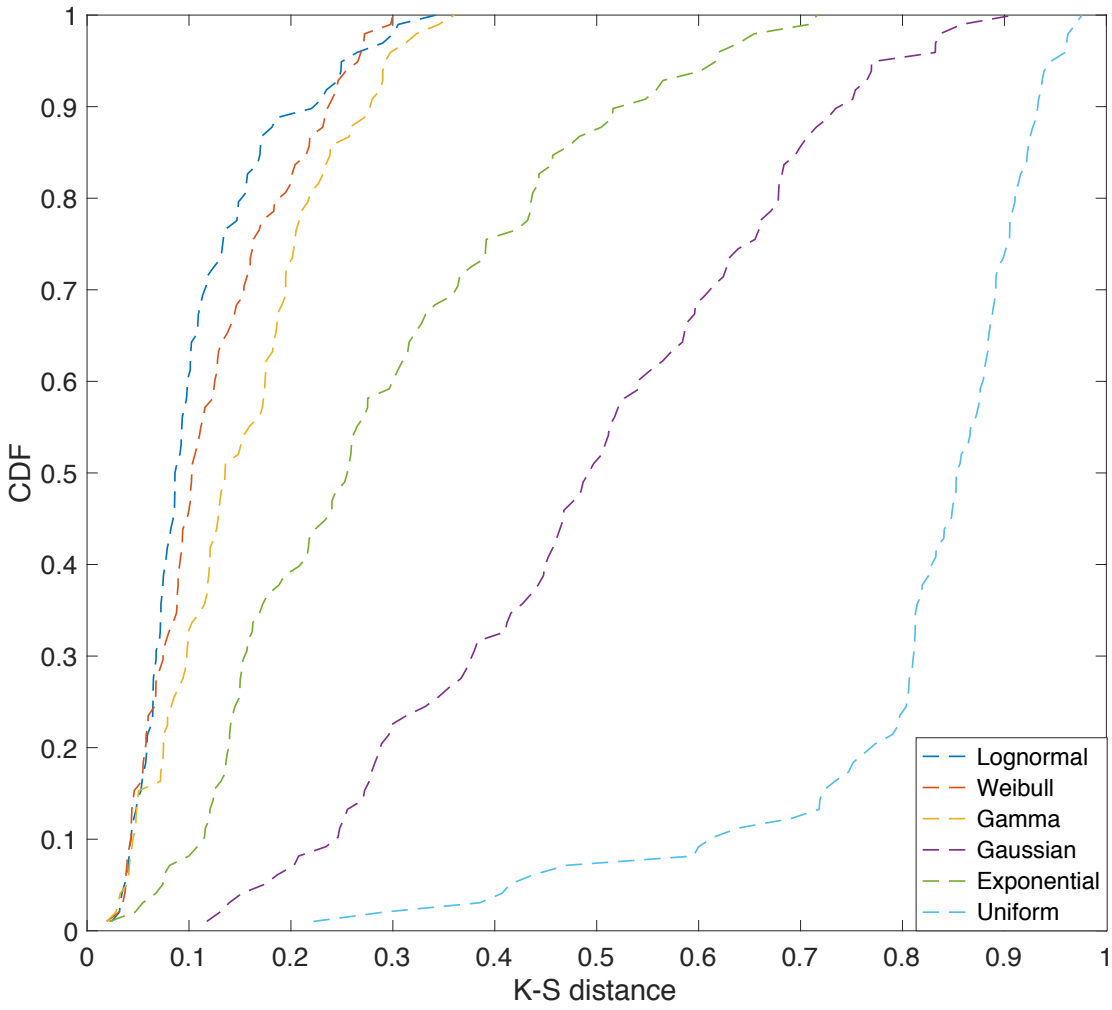


Figure S4. Kolmogorov-Smirnov (K-S) statistics.

Cumulative probability distributions of the K-S distance over the nutrient panel, for different types of distributions. The distributions parametrized by shape parameters have close performances, performing remarkably better than those described by just location and scale. Log-normal distribution offers the best approximation, with most nutrients with Kolmogorov-Smirnov (K-S) distance close to 0.08 and 60% of the nutrients with K-S distance less than 0.1.

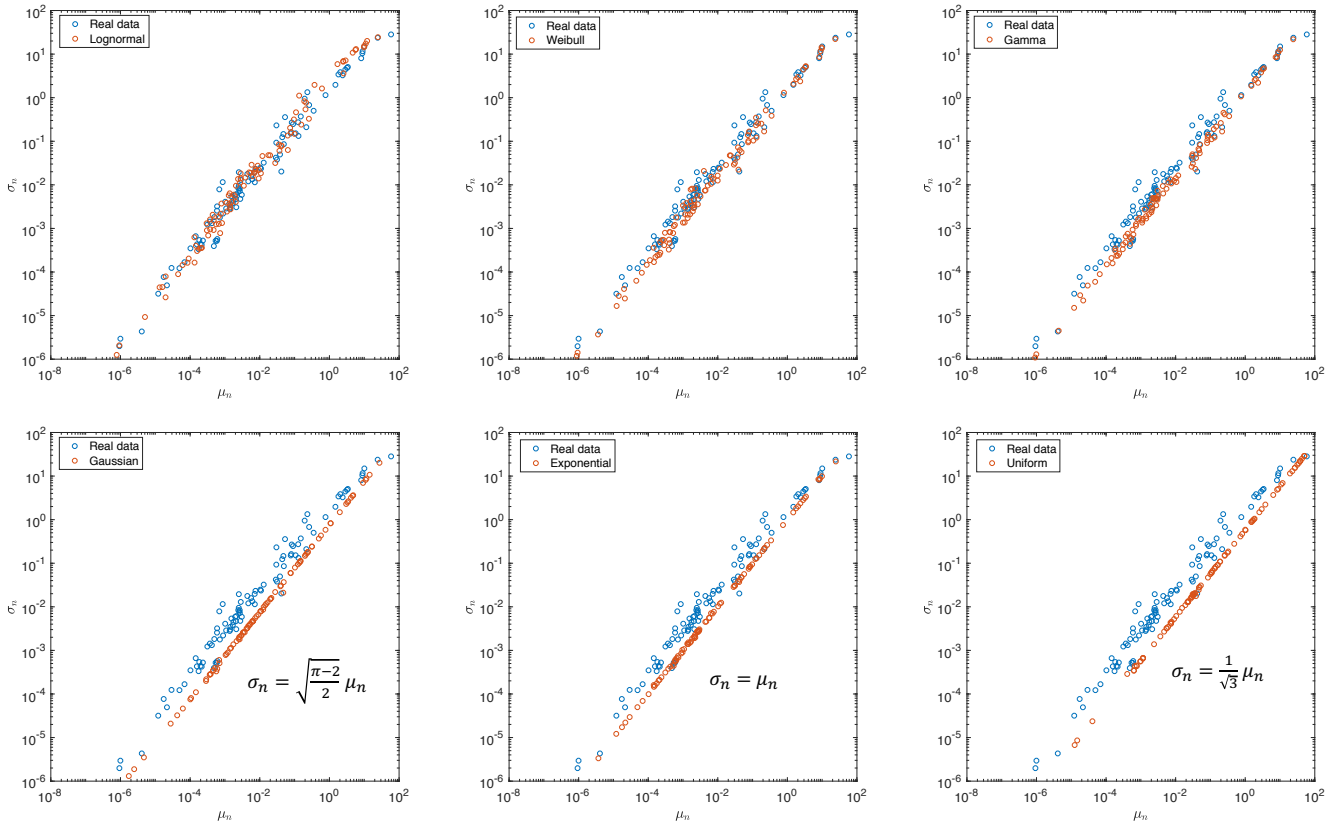


Figure S5: Comparison between real scaling and predicted scaling between μ_n and σ_n

Comparison of the relationship between μ_n and σ_n for the nutrient panel with the outcome of random samples of the same size drawn from the fitted distributions. Exponential, truncated gaussian and uniform distributions fail in approximating the observed intercept (~ coefficient of variation), having constrained coefficient of variation, uniquely determined by their parametrization.

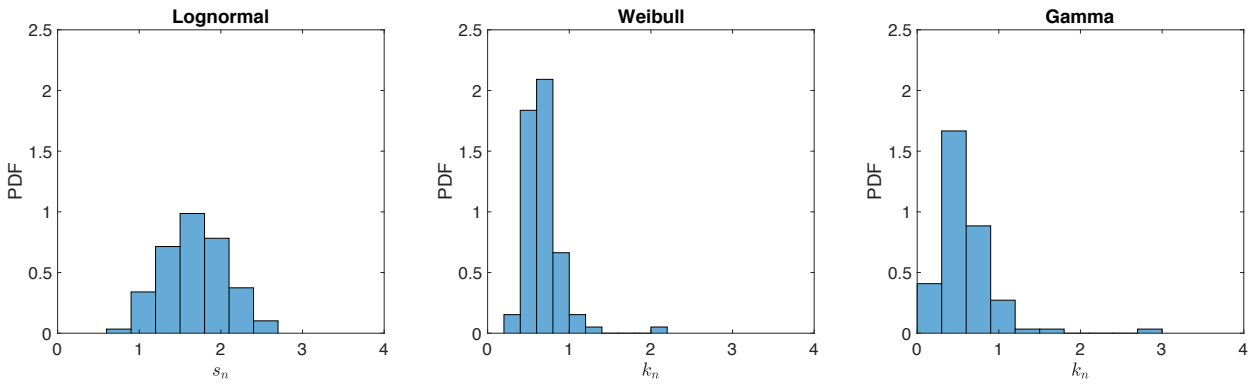


Figure S6. Distributions of the shape parameters across the nutrient panel.

The approximated linearity between μ_n and σ_n forces the shape parameters for log-normal, weibull and gamma to have a tight distribution.

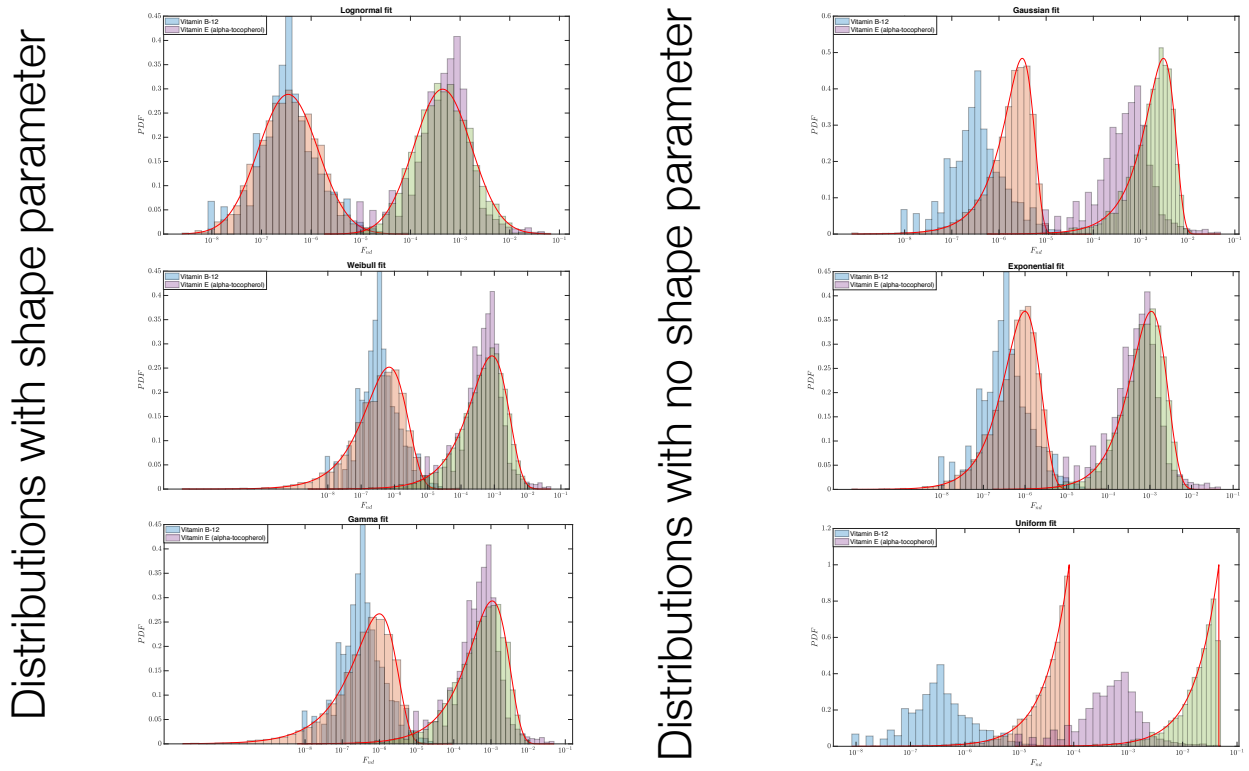


Figure S7. Nutrient distributions in the log-space.

The behavior of real data distributions (Vitamin B-12 and E) in the log-space, compared to a random sample from their fitted distributions of the same size. The distributions with a shape parameter (on the left) can modulate the logarithmic standard deviation, while those lacking this type of parametrization or "degree of freedom" (on the right) have fixed logarithmic standard deviation. The log-normal distribution best captures the observed translational invariance and symmetry. The number of bins is automatically chosen by MATLAB.

	Probability distribution in the linear space $Q\ (x)$	Probability distribution in the log space $R\ (y)$	Standard deviation in the log space	Interval covered in the log space	Symmetry in the log space
Variable	x	$y = \log x$	—	—	—
Uniform	$\frac{1}{b}$	$\frac{e^y}{b}$	1	$(-\infty, \log b]$	No
Half Gaussian	$\frac{1}{s\sqrt{\frac{\pi}{2}}} e^{-\frac{x^2}{2s^2}}$	$\frac{1}{s\sqrt{\frac{\pi}{2}}} e^{y-\frac{e^{2y}}{2s^2}}$	$\frac{\pi}{2\sqrt{2}}$	$(-\infty, \infty]$	No
Exponential	$\frac{1}{\theta} e^{-\frac{x}{\theta}}$	$\frac{1}{\theta} e^{y-\frac{e^y}{\theta}}$	$\frac{\pi}{\sqrt{6}}$	$(-\infty, \infty]$	No
Half Logistic	$\frac{e^{-\frac{x}{s}}}{\frac{s}{2}\left(1+e^{-\frac{x}{s}}\right)^2}$	$\frac{e^{y+\frac{e^y}{s}}}{\frac{s}{2}\left(1+e^{\frac{e^y}{s}}\right)^2}$	$< \frac{\pi}{\sqrt{6}}$	$(-\infty, \infty]$	No
Gamma	$\frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$	$\frac{e^{ky-\frac{e^y}{\theta}}}{\Gamma(k)\theta^k}$	$\sqrt{\frac{d^2 \log \Gamma(k)}{dk^2}}$	$(-\infty, \infty]$	No
Weibull	$\frac{k}{\theta}\left(\frac{x}{\theta}\right)^{k-1} e^{-\left(\frac{x}{\theta}\right)^k}$	$\frac{k}{\theta^k} e^{ky-\left(\frac{e^y}{\theta}\right)^k}$	$\frac{\pi}{\sqrt{6}}\frac{1}{k}$	$(-\infty, \infty]$	No
Fréchet	$\frac{k}{\theta}\left(\frac{x}{\theta}\right)^{-1-k} e^{-\left(\frac{x}{\theta}\right)^{-k}}$	$\frac{k}{\theta^{-k}} e^{-ky-\left(\frac{e^y}{\theta}\right)^{-k}}$	$\frac{\pi}{\sqrt{6}}\frac{1}{k}$	$(-\infty, \infty]$	No
Log-normal	$\frac{1}{xs\sqrt{2\pi}} e^{-\frac{(\log x - m)^2}{2s^2}}$	$\frac{1}{s\sqrt{2\pi}} e^{-\frac{(y-m)^2}{2s^2}}$	s	$(-\infty, \infty]$	Yes
Loglogistic	$\frac{1}{se^m\left(1+\left(\frac{x}{e^m}\right)^{1/s}\right)^2}\left(\frac{x}{e^m}\right)^{\frac{1}{s}-1}$	$\frac{e^{-\frac{y-m}{s}}}{s\left(1+e^{-\frac{y-m}{s}}\right)^2}$	$\frac{\pi}{\sqrt{3}}s$	$(-\infty, \infty]$	Yes

Table S1: Candidate probability distributions and their features in the log space

For each distribution we show the probability in the linear space $Q\ (x)$, the transformed probability in the log-space $R\ (y)$, and the calculations for the logarithmic standard deviation. We also determine if $R\ (y)$ is symmetric. The standard deviation in the log space is clearly dependent on the shape parameter. Distributions like uniform, exponential, half gaussian and half logistic, lack this parametrization and have a fixed logarithmic standard deviation, bounded by the results derived for the exponential distribution.
In this table we added three additional probability distributions (half-logistic, log-logistic and fréchet) that have consistent properties with the candidate distributions. Fréchet is a distribution functionally related to weibull, consequently it lacks symmetry.
Half- logistic and log-logistic strongly resemble half-gaussian and log-normal, despite the fatter tails. However, these distributions are not maximum-entropy distribution nor the outcome of any known generative mechanism.

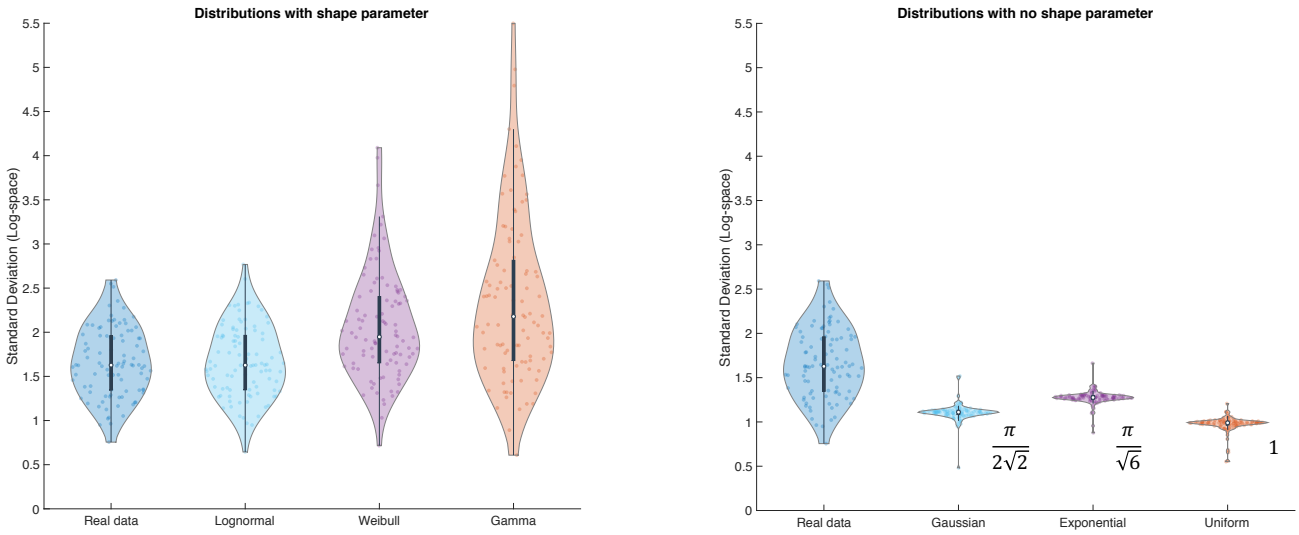


Figure S8. Standard deviation in the log-space.

We illustrate the behavior of the distribution of logarithmic standard deviation across the nutrient panel, compared to the results determined by random sampling from the fitted distributions. While the distributions characterized by a shape parameter are able to model, with different degrees of success, the standard deviations in the real data, the lack of shape parameter determines a collapse of the standard deviations to the fixed values reported in Table S1, independently from the variability affecting the scale parameter in the original linear space.

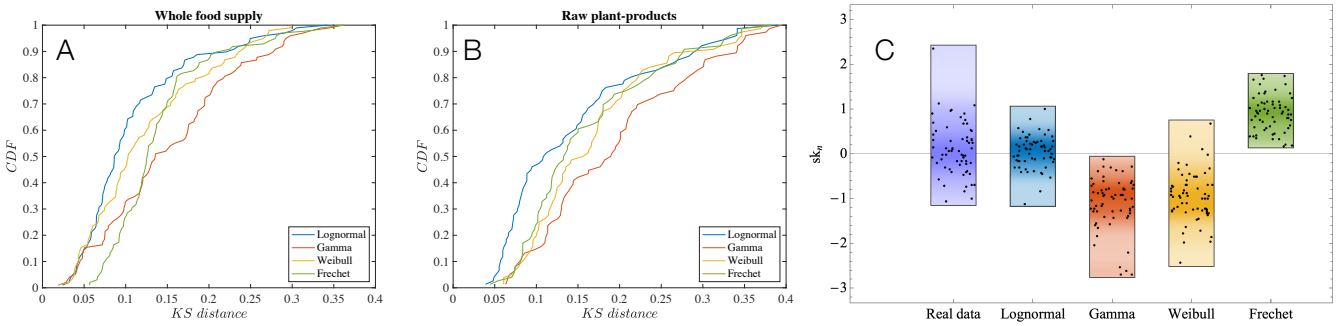


Figure S9: Fit performance of fréchet distribution for nutrients.

(A)-(B) Kolmogorov-Smirnov distances for log-normal, gamma, weibull and fréchet distribution. Cumulative distribution evaluating the fitting performances over the whole nutrient panel (A) and over raw plant-products only (B), whose nutrient concentrations should directly link to the metabolic processes of the original plants (Section S4). Log-normal offers the best approximation and its performance improves moving from the whole food supply to raw plants, where it drastically outperforms gamma, weibull and fréchet.

(C) Skewness of $Q(x_n)$ in the log-space, measuring the asymmetry of the distribution. For raw-plant products the logarithmic skewness fluctuates near zero, independently from μ_n, indicating that $Q(x_n)$ is approximately symmetric in the log-space, as shown for the whole food supply in Figure 2C. By comparing the skewness distribution of real data with the results determined by random sampling from the fitted distributions, we find that only log-normal is consistent with what observed for raw-plant products, while gamma and weibull have a bias towards negative skewness, and fréchet towards positive skewness. Similar results are found for the whole food supply.

Section 3b: Translational invariance and rescaling of the nutrient distributions

The inspection of the empirical pdfs in the logarithmic scale suggests that all $Q(x_n)$ follow the same log-normal random variable rescaled by different constants c, reflecting their different average concentration μ_n in the food supply. Therefore, using as reference random variable a log-normal distribution with $m_* = 0$ and $s_* = \langle s_n \rangle = 1.66$, the rescaling $y_n = e^{\log(x_n) - m_n}$ should collapse all $Q(x_n)$ on a single universal curve, with variability consistent with the fluctuations of the shape parameter s_n observed in Figure 2B. Indeed, the observations for four selected nutrients present in food with different orders of magnitude (Figure S10A), and the analysis for the 99 nutrients (Figure S10B), are consistent with this prediction.

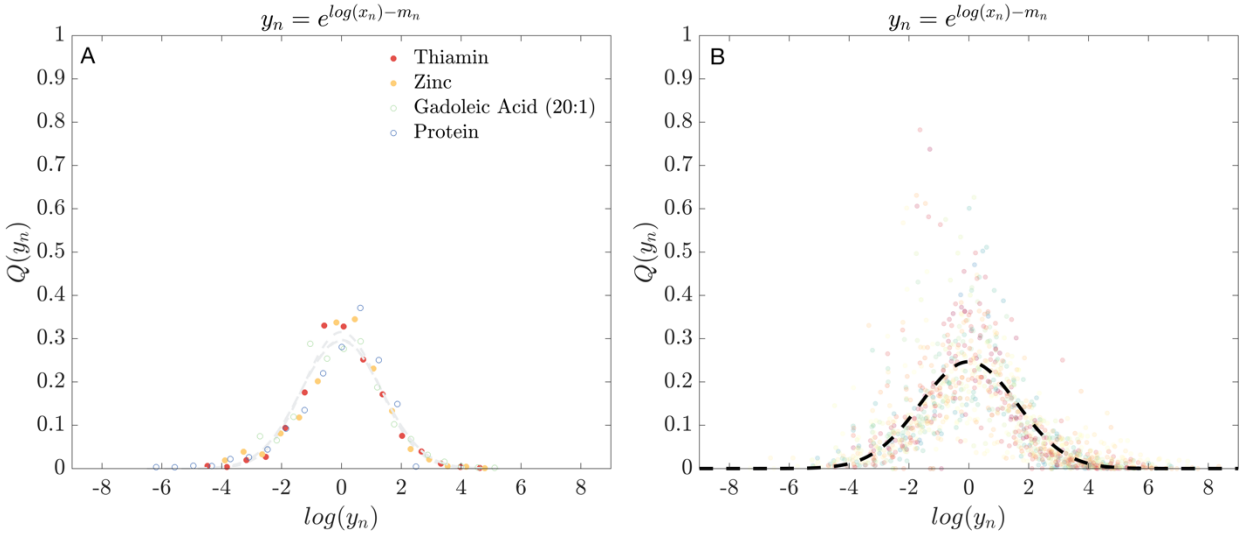


Figure S10. Universal nutrient distributions.

We rescaled the nutrient concentrations using $y_n = e^{log(x_n)-m_n}$, corresponding to a horizontal shift of each curve in the logarithmic space. After the rescaling the $Q(x_n)$ distributions collapse on a single universal curve, indicating that they are described by a single family of distributions. The number of bins is set to 15.

(A) Rescaling for Thiamin, Zinc, Gadoleic Acid and Protein, the nutrients shown in Figure 2A.

(B) Rescaling for the whole nutrient panel. We indicate with a dashed black line the reference log-normal distribution with $m_* = 0$ and $s_* = \langle s_n \rangle = 1.66$.

Section S4: Vegetables and Fruits

From FNDDS we selected a batch of 108 raw vegetables and fruits, whose nutrient concentrations should directly link to the metabolic processes of the original plants, and from which we removed any redundant nutrient profile (i.e., identical or repeated foods). The analysis for the whole food supply revealed that the standard deviation in the log-space of the nutrients was centered at $\langle s_n \rangle = 1.6631 \pm 0.3887$. However, by selecting raw plant-products, s_n becomes bimodal, even if still distributed within the same range (Figure S11). The observed bimodality appears to be correlated with the classes of primary and secondary metabolites (Figure S11 and Figure S12D).

Nutrients belonging to the secondary metabolism of plant are specialized compounds (mainly terpenoids and flavonoids), whose rate of production depends on substrates and enzymes derived from the primary metabolism. While primary metabolites exhibit tight distributions with $\langle s_n \rangle = 0.9086 \pm 0.2459$ (Figure S12A), secondary metabolites are characterized by a higher degree of variability of their content in food, with $\langle s_n \rangle = 1.9112 \pm 0.4461$ (Figure S12B).

Despite the bimodality of s_n, the nutrient distributions show a high level of symmetry (Figure S12E) and are well-fitted by the log-normal distribution (Figure S13B), with the K-S statistics rejecting only six nutrients, manly for batch effects due to the imputation of nutritional values. Among the nutrients which deviate from log-normality we find sodium, whose levels in plants are driven by environmental conditions, not well captured in our enzymatic reaction model.

The relation between σ_n and μ_n is linear, with $\beta_\sigma = 0.977$ $(0.9296, 1.024)$, $\alpha_\sigma = 0.2193$ $(-0.07554, 0.514)$ and adjusted $R^2 = 0.9568$. By removing water, we obtain $\beta_\sigma = 1.012$ $(0.9664, 1.057)$, $\alpha_\sigma = 0.4412$ $(0.1578, 0.7247)$ and adjusted $R^2 = 0.9633$ (Figure S13A).

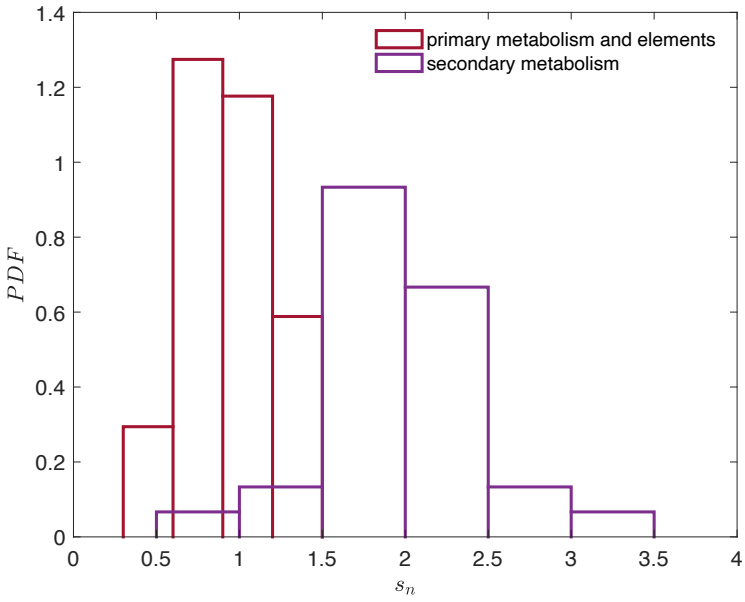


Figure S11. Standard deviation in the log-space for raw plants.

Decomposition of the distribution for raw plants in nutrients belonging to primary metabolism and elements, and nutrients related to the secondary metabolism of plants.

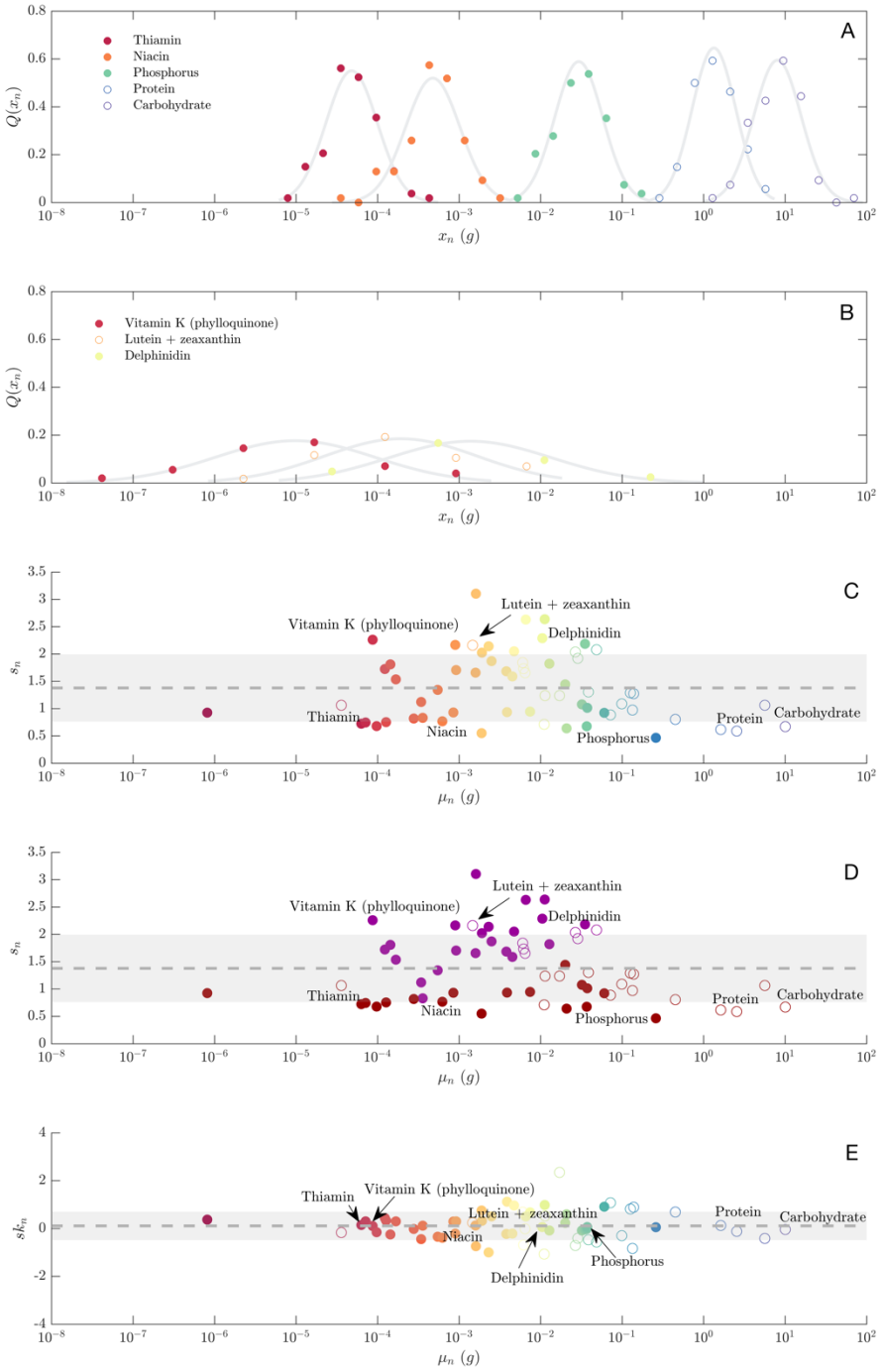

A
Thiamin
Niacin
Phosphorus
Protein
Carbohydrate
$Q(x_n)$
x_n (g)
B
Vitamin K (phylloquinone)
Lutein + zeaxanthin
Delphinidin
$Q(x_n)$
x_n (g)
C
Lutein + zeaxanthin
Vitamin K (phylloquinone)
Delphinidin
Thiamin
Niacin
Phosphorus
Protein
Carbohydrate
s_n
μ_n (g)
D
Lutein + zeaxanthin
Vitamin K (phylloquinone)
Delphinidin
Thiamin
Niacin
Phosphorus
Protein
Carbohydrate
s_n
μ_n (g)
E
Thiamin
Vitamin K (phylloquinone)
Lutein + zeaxanthin
Niacin
Delphinidin
Phosphorus
Protein
Carbohydrate
sk_n
μ_n (g)

Figure S12. Statistics for raw plants.

(A) $Q(x_n)$ distributions for 5 nutrients related to primary metabolism and elements. The obtained distributions $Q(x_n)$ are approximately symmetric on a log scale, typical feature of log-normal distributions. Note that the nutrient distributions have similar width, independently from their average abundance in food.

(B) $Q(x_n)$distributions for 3 nutrients related to secondary metabolism. Note that the common width is higher compared to the nutrients in primary metabolism.

(C) Standard deviation of $Q(x_n)$in the log-space. Relation between the parameters s_n and the respective average content in food.

(D) Standard deviation of $Q(x_n)$ in the log-space. Decomposition of the distribution according to primary metabolism and elements (red) vs secondary metabolism (purple).

(E) Skewness of $Q(x_n)$ in the log-space. Skewness quantifies the asymmetry of the probability distribution. All nutrient distributions show a high level of symmetry.

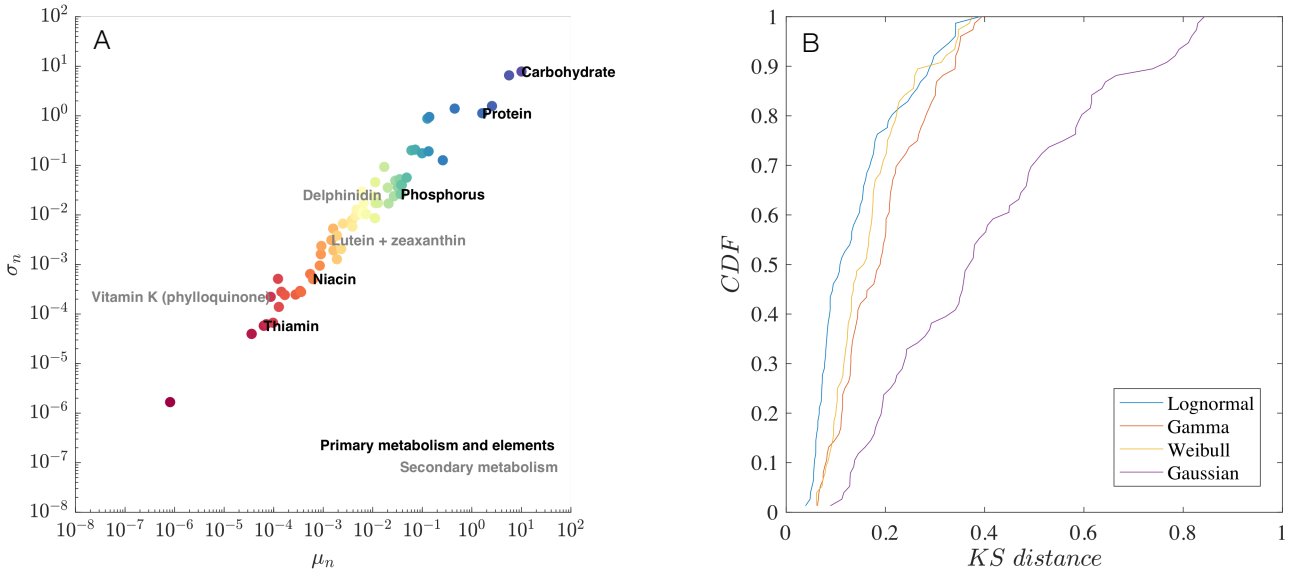


Figure S13. Statistics for raw plants.

(A) Relation between average nutrient amount μ_n *and standard deviation* σ_n*. For those foods where nutrient n is quantified we calculate* μ_n *,* σ_n *and then we evaluated their functional relation* $\sigma_n = e^{\alpha_\sigma}(\mu_n)^{\beta_\sigma}$ *across the whole nutrient panel. The fitted relation is linear with exponents* $\beta_\sigma = 1.012\ \ (0.9664, 1.057)$*,* $\alpha_\sigma = \ 0.4412\ \ (0.1578, 0.7247)$ *and adjusted* $R^2 = 0.9633$*.*

(B) Kolmogorov-Smirnov distances for four different statistical distributions.

Section S5: The Nutrient Profile of Dishes

The theoretical framework leading to (2) is rooted in the biochemical processes of living systems, hence mainly describing natural ingredients. Yet, our food is assembled following recipes that mix different ingredients. The nutrient content for 100 g of a recipe depends on the properties of its *K* ingredients and their amount m_k. If we neglect the chemical-physical transformations during the preparation process, we can assume

$$\sum_k^K m_k = 100\,. \tag{S4}$$

For a selected nutrient *x*, the contribution of the k^{th} ingredient to the recipe follows

$$x'_k = \frac{m_k}{100} x_k = w_k x_k, \tag{S5}$$

where x_k is the amount of nutrient for 100 g of the selected ingredient and $\sum_k^K w_k = 1$. Consistently, as a first approximation, the nutrient content of a recipe is a weighted average random variable

$$x_{recipe} = \sum_k^K w_k x_k\,, \tag{S6}$$

where $x_k \sim$ log-normal(*m, s*). Additionally, we can model the variability of the weights $\vec{w} = \{w_k\}$ as a Dirichlet distribution, i.e.,

$$P(\vec{w}) = \frac{\Gamma(\sum_k^K \alpha_k)}{\prod_k^K \Gamma(\alpha_k)} \prod_k^K w_k^{\alpha_k - 1}\,, \tag{S7}$$

where $\{\alpha_k\}$ are positive parameters influencing weight expectations, variances and correlations. For the sake of simplicity we consider the log-normal random variables $\{x_k\}$ independent (they are not). The expectation values and the variance-covariance matrix for $\vec{x} = \{x_k\}$ follow as

$$\overrightarrow{\mu_x} = \begin{pmatrix} e^{m+\frac{s^2}{2}} \\ \vdots \\ e^{m+\frac{s^2}{2}} \end{pmatrix} \tag{S8}$$

and

$$\Sigma_{xx} = \begin{pmatrix} e^{2m+s^2}\left(e^{s^2}-1\right) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{2m+s^2}\left(e^{s^2}-1\right) \end{pmatrix}. \tag{S9}$$

Weights determined by the Dirichlet distribution are naturally correlated, given the constraint on their sum $\sum_k^K w_k = 1$. Consequently, the expectation values and the variance-covariance matrix for $\vec{w} = \{w_k\}$ follow

$$\overrightarrow{\mu_w} = \begin{pmatrix} \frac{\alpha_1}{\sum_k^K \alpha_k} \\ \vdots \\ \frac{\alpha_K}{\sum_k^K \alpha_k} \end{pmatrix} \tag{S10}$$

and

$$\Sigma_{ww} = \begin{pmatrix} \frac{\alpha_1(\sum_k^K \alpha_k - \alpha_1)}{(\sum_k^K \alpha_k)^2(1+\sum_k^K \alpha_k)} & \cdots & \frac{-\alpha_1\alpha_K}{1+\sum_k^K \alpha_k} \\ \vdots & \ddots & \vdots \\ \frac{-\alpha_1\alpha_K}{1+\sum_k^K \alpha_k} & \cdots & \frac{\alpha_K(\sum_k^K \alpha_k - \alpha_K)}{(\sum_k^K \alpha_k)^2(1+\sum_k^K \alpha_k)} \end{pmatrix}. \tag{S11}$$

Given the statistical properties of the random variables describing the amount of each ingredient and their nutrient content, we can calculate the expectation value and the variance for their inner product $x_{recipe} = \vec{w}^{\mathrm{T}}\vec{x}$. In absence of correlation between $\{x_k\}$ and $\{w_k\}$, they follow

$$E\left[x_{recipe}\right] = \overrightarrow{\mu_w}^T \overrightarrow{\mu_x} = e^{m+\frac{s^2}{2}}, \tag{S12}$$

and

$$\sigma^2\left(x_{recipe}\right) = \overrightarrow{\mu_x}^T \Sigma_{ww} \overrightarrow{\mu_x} + \overrightarrow{\mu_w}^T \Sigma_{xx} \overrightarrow{\mu_w} + Tr(\Sigma_{ww}\Sigma_{xx}) = \\ = \frac{\sum_k^K \alpha_k \left(1+\alpha_k\right)}{\left(\sum_k^K \alpha_k\right)\left(1+\sum_k^K \alpha_k\right)} e^{2m+s^2}\left(e^{s^2}-1\right). \quad (S13)$$

The expectation value is consistent with the original log-normal distribution, while the variance has a pre-factor bounded between $1/K$ and 1, making the original variance for the single ingredient a superior limit for recipes. The minimal variance is found in the case of equal $\{\alpha_k\} = \alpha$, when $\alpha \to \infty$, i.e.,

$$\lim_{\alpha\to\infty} \frac{1+\alpha}{1+K\alpha} = \frac{1}{K}. \quad (S14)$$

From a Bayesian perspective $\alpha \to \infty$ can be interpreted as high confidence in the partition values assigned to the ingredients. On the other hand, maximal variance is determined by recipes composed by a single ingredient.

In summary, the standard deviation $\sigma\left(x_{recipe}\right)$ for a recipe composed by *K* ingredients is bounded in the interval

$$\frac{e^{m+\frac{s^2}{2}}\sqrt{e^{s^2}-1}}{\sqrt{K}} \leq \sigma\left(x_{recipe}\right) \leq e^{m+\frac{s^2}{2}}\sqrt{e^{s^2}-1}. \quad (S15)$$

When the number of ingredients *K* is limited, Eqs. (S12) and (S15), together with the statistical methods successfully approximating sum of log-normal random variables with a unique log-normal[11], prove that the behavior of recipes is close to the observations for single ingredients.

Recipes' standard deviation in the logarithmic space

The standard deviation in the logarithmic space *s* is the shape parameter determining Eqs. (2) and (3). We are able to approximate the effect of combining ingredients on this parameter by using a

multivariate first order Taylor expansion of the nonlinear function $f = \log(\sum_k^K w_k e^{y_k})$ at the expected value, where $y_k \sim$ normal(m, s). The gradient of f, evaluated at $(\langle\vec{y}\rangle, \langle\vec{w}\rangle)$, follows as

$$\vec{J}|_{\langle\cdot\rangle} = \begin{pmatrix} \frac{\partial f}{\partial y_1} \\ \vdots \\ \frac{\partial f}{\partial y_K} \\ \frac{\partial f}{\partial w_1} \\ \vdots \\ \frac{\partial f}{\partial w_K} \end{pmatrix}\Bigg|_{\langle\cdot\rangle} = \begin{pmatrix} \frac{\alpha_1}{\sum_k^K \alpha_k} \\ \vdots \\ \frac{\alpha_K}{\sum_k^K \alpha_k} \\ 1 \\ \vdots \\ 1 \end{pmatrix}. \tag{S16}$$

The approximated variance $\sigma^2(f)$ is determined by the quadratic function

$$\sigma^2(f) \approx \vec{J}|_{\langle\cdot\rangle}{}^T \Sigma_{all}\, \vec{J}|_{\langle\cdot\rangle} = s^2 \frac{\|\vec{\alpha}\|_2^2}{\|\vec{\alpha}\|_1^2}, \tag{S17}$$

where Σ_{all} is the total variance-covariance matrix considering both $\vec{y}$ and $\vec{w}$,

$$\Sigma_{all} = \begin{pmatrix} s^2 & 0 & \dots & 0 \\ 0 & \ddots & 0 & 0 \\ \vdots & 0 & s^2 & \vdots \\ 0 & 0 & \dots & \Sigma_{ww} \end{pmatrix}. \tag{S18}$$

By using the equivalence property of the Manhattan and Euclidean norms,

$$\|\vec{\alpha}\|_2 \leq \|\vec{\alpha}\|_1 \leq \sqrt{K}\|\vec{\alpha}\|_2\,, \tag{S19}$$

we determine the extent of the perturbation to the parameter s,

$$\frac{s}{\sqrt{K}} \leq \sigma(f) \leq s, \tag{S20}$$

where, differently from the linear case, any time $\{\alpha_k\} = \alpha$ we reach the minimal standard deviation $\frac{s}{\sqrt{K}}$, independently from the value assigned to α.

Section S6: Validation of the Scaling Laws for Polyphenols

To further validate the scaling laws (2) and (3), we inspected Phenol-Explorer, a comprehensive database on polyphenol content in foods[12]. The database reports measurements for over 400 foods, derived from the inspection of more than 1,300 scientific publications. It was incorporated and extended by FooDB, currently the largest resource on food components[13]. Despite the 498 polyphenols listed, the database is sparse, meaning that each chemical is reported only in a limited number of foods, consistently smaller compared to national databases. The minimal number of data points considered necessary to fit the distributions was set to 10. This selection includes 70 compounds, a number decreasing to 42 when a threshold of 15 datapoints is considered. We were able to replicate the results presented in the main text, as shown in Figure S14 and Figure S15. The polyphenols show a higher standard deviation in the log-space, fluctuating around $\langle s_n \rangle = 2.3610 \pm 0.7847$, and consistent with the secondary metabolites in Section S4, but also with what was observed for the whole nutrient panel. We find $\beta_\sigma = 1.064\ (1.003, 1.124)$ and $\alpha_\sigma = 0.731$ (0.5481, 0.9139), and when we restrict our analysis to those compounds with at least 15 data points we obtain $\beta_\sigma = 1.08(0.9912, 1.168)$ and $\alpha_\sigma = 0.8934$ (0.6467, 1.14).

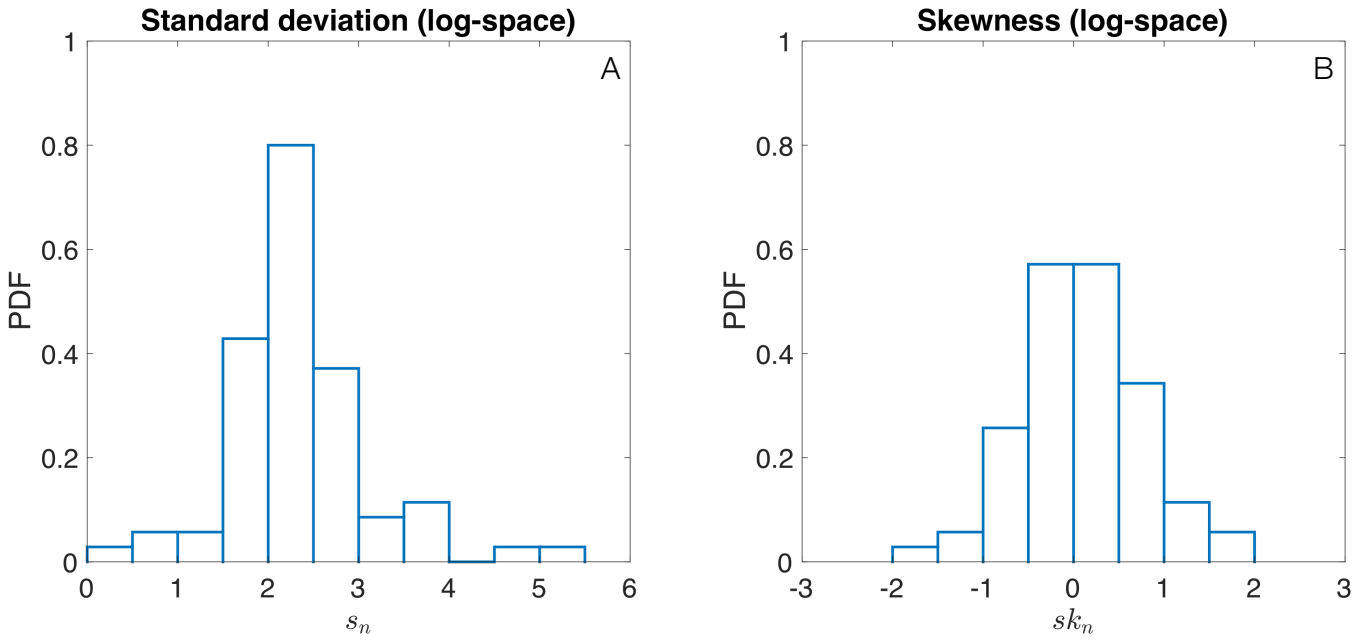


Figure S14: Summary of the statistical properties of polyphenols in the logarithmic space.

Standard deviation (A) and skewness (B) in the log space for polyphenols (minimal number of data points equal to 10).

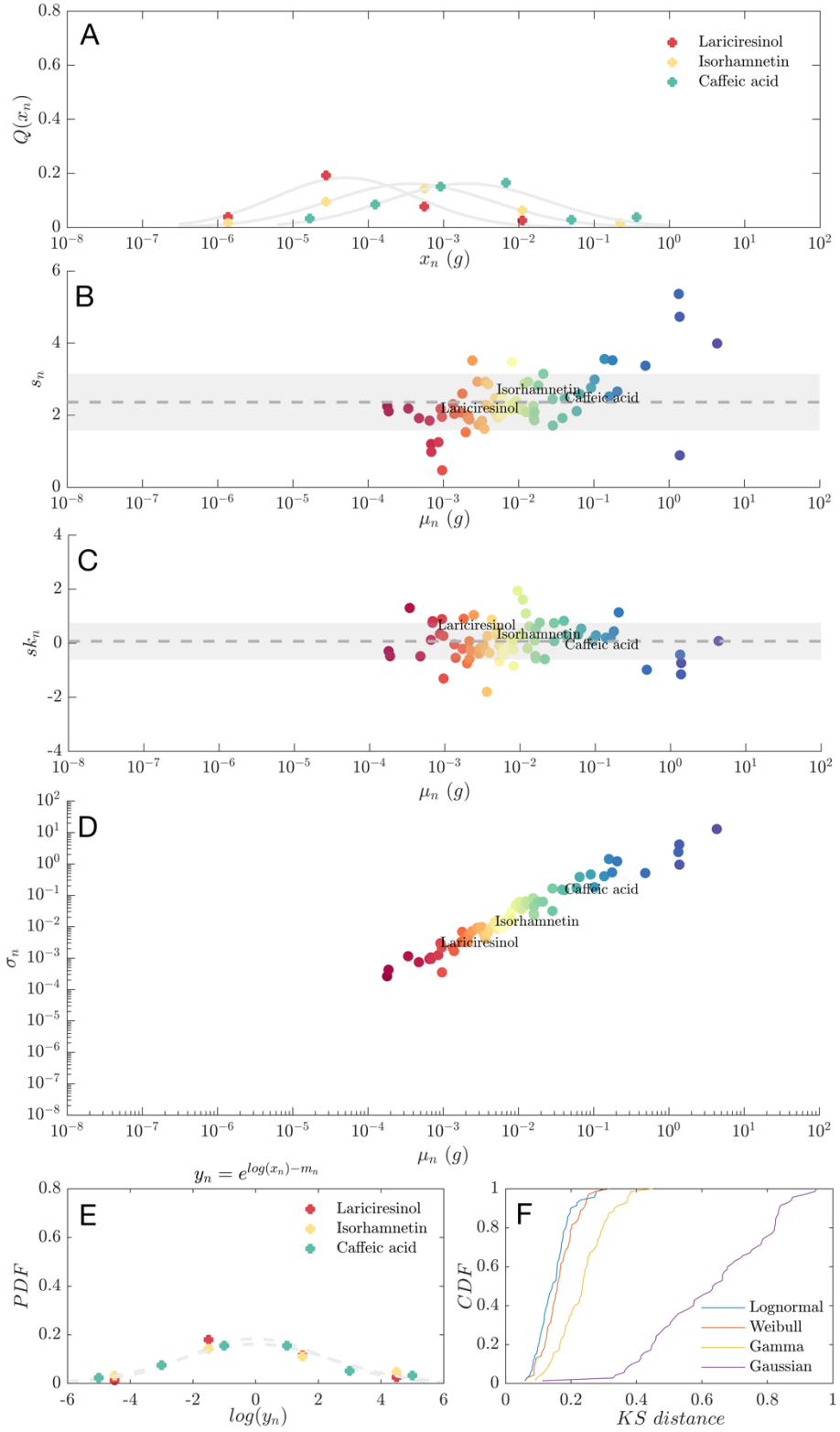

A
Lariciresinol
Isorhamnetin
Caffeic acid
$Q(x_n)$
x_n (g)
B
s_n
Isorhamnetin
Caffeic acid
Lariciresinol
μ_n (g)
C
sk_n
Lariciresinol
Isorhamnetin
Caffeic acid
μ_n (g)
D
σ_n
Caffeic acid
Isorhamnetin
Lariciresinol
μ_n (g)
$y_n = e^{log(x_n)-m_n}$
E
Lariciresinol
Isorhamnetin
Caffeic acid
PDF
$log(y_n)$
F
CDF
Lognormal
Weibull
Gamma
Gaussian
KS $distance$

Figure S15. Statistics for polyphenols.

(A) $Q(x_n)$ distributions for 3 selected nutrients. We find that the obtained distributions $Q(x_n)$ are approximately symmetric on a log scale, typical feature of log-normal distributions. Note that the nutrient distributions have similar width, independently from their average abundance in food

(B) Standard deviation of $Q(x_n)$ in the log-space. Relation between the parameters s_n and the respective average content in food μ_n .

(C) Skewness of $Q(x_n)$ in the log-space. Skewness quantifies the asymmetry of the probability distribution.

(D) Relation between average nutrient amount μ_n and standard deviation σ_n. For those foods where nutrient n is quantified we calculate μ_n , σ_n and then we evaluate their functional relation $\sigma_n = e^{\alpha_\sigma}(\mu_n)^{\beta_\sigma}$ across the whole nutrient panel. The fitted relation is close to linear with exponents $\beta_\sigma = 1.064$ $(1.003, 1.124)$ and $\alpha_\sigma = 0.731$ (0.5481, 0.9139) and covers approximately four orders of magnitude.

(E) Rescaled $Q(x_n)$ for 3 selected nutrients. We rescale the nutrient content of each food using $y_n = e^{log(x_n) - m_n}$, corresponding to a horizontal shift of each curve. After this rescaling the $Q(x_n)$ distributions for the 3 nutrients collapse on a single universal curve.

(F) Kolmogorov-Smirnov distances for four different statistical distribution, showing that log-normal still offers the best approximation.

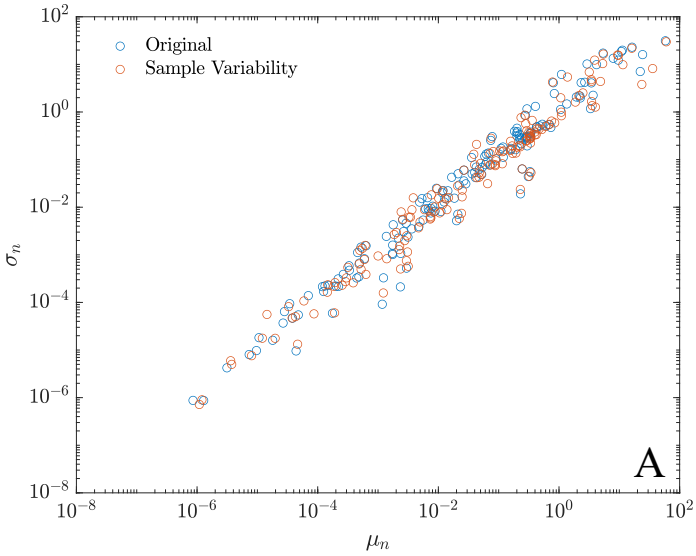
Section S7: Validation of the Scaling Laws in Foundation Foods

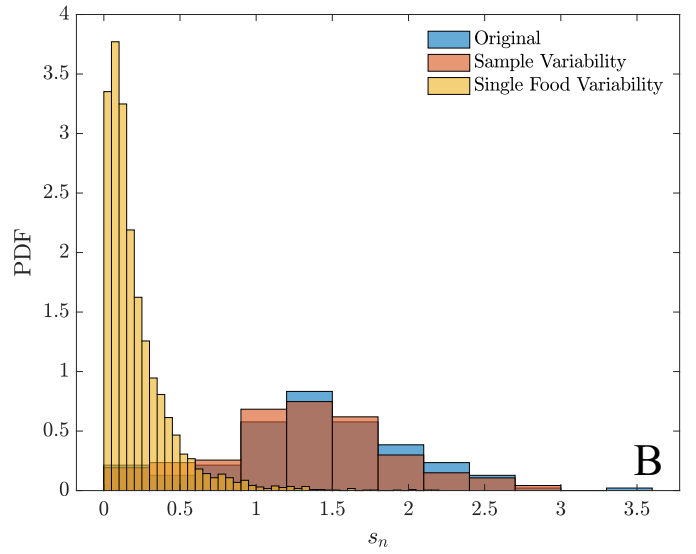
Foundation Foods[10] is a new food composition dataset available at FoodData Central, the online platform collecting all the databases created by the USDA, among which we find also FNDDS and SR described in Section S1. Foundation Foods includes individual sample measurements behind the nutrient mean values that populate the other databases, and metadata reporting the number of samples, location, time-stamps, analytical methods used, and, additionally, if available, cultivar and production practices.

We leveraged the data provided by Foundation Foods to test the robustness of our findings against sample variability. We collected data for 116 foods, profiled with varying nutrient panel resolution, for a total of 156 nutrients.

First, we investigated how sample variability affects the scaling described in Eq. (3), finding no statistically significant alterations. Indeed, when the scaling is calculated considering an average nutrient measure per food, we find $\beta_\sigma = 1.006(0.9746, 1.037)$, $\alpha_\sigma = 0.1689(0.003091, 0.3347)$ and adjusted $R^2 = 0.9637$, while when we include sample variability (i.e., a food can contribute with multiple measurements per nutrient) we observe $\beta_\sigma = 0.9859\ \ (0.9554, 1.016)$, $\alpha_\sigma = -0.03795$ (-0.1987, 0.1228) and adjusted $R^2 = 0.9634$ (Figure S16A). Additionally, we evaluated the order of magnitude of the nutrient fluctuations within food and between foods, comparing the standard deviation s_n in the logarithmic space. While the distribution of s_n between foods is consistent with or without sample variability ("original" estimation with nutrient averages $\langle s_n \rangle = 1.4131 \pm 0.6059$, estimation including "sample variability" $\langle s_n \rangle = 1.3213 \pm 0.5839$), the behavior of s_n within the same food samples shows a drastically different probability distribution with $\langle s_n \rangle = 0.2135 \pm 0.2253$ (Figure S16B). Within

samples of the same food, we observe the highest s_n when the number of data points is limited (Figure S16C).





Single Food Variability per Nutrient

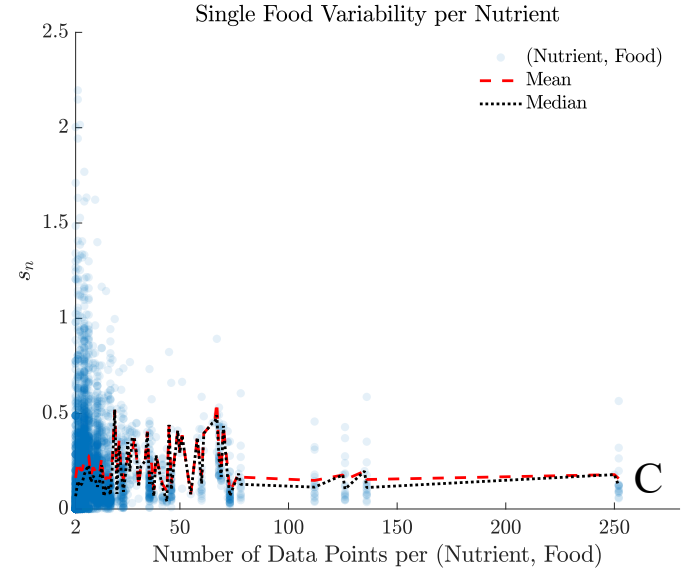


Figure S16. Within-Food Variability vs Between-Food Variability in Foundations Foods.

(A) Relation between average nutrient amount μ_n and standard deviation σ_n. Including sample variability does not alter the characteristics of the fit.

(B) Logarithmic standard s_n. The logarithmic nutrient variability within the same food has different statistical properties compared to that between foods.

(C) Relation between number of data points sampled and logarithmic nutrient variability within the same food.

Section S8: The Origin of the Log-normal Form

The goal of this section is to derive analytically the log-normal form for metabolite concentrations and the log-normal variability (shape parameter) s_n. We first introduce the Michaelis-Menten kinetics and its stochastic treatment, and then extend our results to linear pathways.

Michaelis-Menten kinetics for a single chemical reaction

The Michaelis-Menten model formalizes the relation between reaction velocity and substrate concentration for a system where a substrate *S* binds reversibly (with reaction rate constants k_1 and k_{-1}) to an enzyme *E* to form an enzyme-substrate complex *ES*, which then reacts irreversibly to generate a product *P* and to regenerate the free enzyme *E* (with reaction rate constant k_2). The standard model describing a single enzymatic reaction follows the reaction,

$$S + E \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} ES \overset{k_2}{\rightarrow} P + E, \tag{S21}$$

with turnover equation

$$v = \frac{d[P]}{dt} = \frac{v_{max}[S]}{K_M + [S]}, \tag{S22}$$

where $v_{max} = k_2[E]_{tot}$ represents the maximum velocity achieved by the system when all enzyme molecules are bound, and

$$K_M = \frac{k_1 + k_{-1}}{k_2} \tag{S23}$$

is the substrate concentration at which the reaction velocity is 50% of v_{max}. Quantitative modeling in chemical engineering has investigated several types of enzyme kinetics, including inhibition

and activation, leveraging the deterministic law of mass action and discarding the inherent stochasticity characterizing mesoscopic chemistry and biochemistry. The stochastic counterpart of the chemical kinetic equation based on the law of mass action is the Chemical Master Equation[14] (CME), that no longer works with the chemical concentrations $[S], [E], [ES]$ and $[P]$, but instead focuses on the probability $p(n, c, t)$ of the system having *n* number of *S* molecules and *c* number of *ES* enzyme-substrate complexes, at time t.

The CME version of the system described in Eq. $(S21)$ follows

$$\frac{dp(n,c,t)}{dt} = -\left[\widehat{k_1} n(c_0 - c) + \left(\widehat{k_{-1}} + \widehat{k_2}\right)c\right]p(n,c,t)$$

$$+\widehat{k_1}(n+1)(c_0 - c + 1)\, p(n+1, c-1, t)$$

$$+\widehat{k_{-1}}(c+1)\, p(n-1, c+1, t)$$

$$+\widehat{k_2}(c+1)\, p(n, c+1, t), \quad 0 \le n \le n_0, 0 \le c \le c_0, \qquad (S24)$$

where the total number of substrate and product molecules is $n_0 = N_S(t) + N_P(t) + N_{ES}(t)$ and the total number of enzyme molecules is $c_0 = N_E(t) + N_{ES}(t)$. The stochastic rate constants are derived from their deterministic counterparts as $\widehat{k_1} = k_1/V$, $\widehat{k_{-1}} = k_{-1}$, $\widehat{k_2} = k_2$, where *V* is the volume of the environment where the reaction occurs. The quasi-stationarity approximation is invoked when the changes in *c* reach stationarity in earlier stages compared to *n*, due to $c_0 \ll n_0$. This common approximation allows us to first tackle the steady state conditional distribution $p(c|n, t)$ that encodes the probability of observing *c* complexes given *n* substrate molecules. We rewrite Eq. $(S24)$ as

$$\frac{dp(c|n,t)}{dt} = -\left[\widehat{k_1} n(c_0 - c) + \left(\widehat{k_{-1}} + \widehat{k_2}\right)c\right]p(c|n,t)$$

$$+\widehat{k_1}(n+1)(c_0-c+1)\,p(c-1|n,t)$$

$$+\left(\widehat{k_{-1}}+\widehat{k_2}\right)(c+1)\,p(c+1|n,t), \qquad (S25)$$

where $n-1\approx n$. The stationary conditional distribution is determined by $\frac{dp(c|n,t)}{dt}=0$ and follows the binomial distribution

$$p(c|n)^{ss}=\binom{c_0}{c}p_+^c(1-p_+)^{c_0-c}, \qquad (S26)$$

where the probability of success is $p_+=\frac{\widehat{k_1}n}{\widehat{k_1}n+\widehat{k_{-1}}+\widehat{k_2}}$.

The turnover rate for n substrate molecules measures the expected rate of increase of *P* molecules and is derived from the expectation value of Eq. $(S26)$ as

$$d_n=\widehat{k_2}\langle c(n)\rangle=\widehat{k_2}c_0\frac{\widehat{k_1}n}{\widehat{k_1}n+\widehat{k_{-1}}+\widehat{k_2}}=\widehat{v_{max}}\frac{n}{n+\widehat{K_M}} \qquad (S27)$$

where $\widehat{v_{max}}$ and $\widehat{K_M}$ are the molecular equivalent of v_{max} and K_M, respectively.

Steady State Assumption for Metabolic Pathways

In a metabolic pathway the substrate molecules are synthesized or imported from the environment, and at the same time, turned over into products. The number of substrate molecules remains fixed only when the rate at which they are fed into the solution is chosen to balance on average the enzymatic turnover rate. We consider here a constant influx b of substrate molecules[15,16] into the system, processed at rate d_n by the enzymes, as described in Eq. $(S27)$. This system is equivalent to a birth-death process with constant birth rate b and regulated destruction rate d_n, described by the following CME

$$\frac{dp(n,t)}{dt} = -(b + d_n)p(n,t) + bp(n-1,t) + d_{n+1}p(n+1,t). \quad (S28)$$

At the steady state the probability distribution captures the fluctuations in the number of substrate molecules due to the stochastic replenishment and enzymatic turnover. The stationary formulation is consistent with the general equation for a birth-death process, i.e.,

$$p(n)^{ss} = p(0)^{ss}\frac{b^n}{\prod_{k=1}^{n} d_k} = p(0)^{ss} r^n \binom{n + \widehat{K_M}}{n}, \quad (S29)$$

where $r = \frac{b}{\widehat{v_{max}}}$, and $p(0)^{ss}$ is determined by leveraging the binomial series with negative exponent, i.e.,

$$\sum_{n=0}^{\infty} \binom{n + \widehat{K_M}}{n} r^n = \frac{1}{(1-r)^{\widehat{K_M}+1}}. \quad (S30)$$

To allow the existence of a steady state, the convergence of the binomial series imposes $0 \leq r \leq 1$, that determines an upper boundary to the influx of substrate molecules $b \leq \widehat{v_{max}}$. Indeed, $b > \widehat{v_{max}}$ would imply an exponential growth of the substrate molecules. The steady state distribution is then a negative binomial,

$$p^{ss}(n) = \binom{n + \widehat{K_M}}{n} (r)^n (1-r)^{\widehat{K_M}+1}, \quad (S31)$$

with average

$$\langle n \rangle = \left(\widehat{K_M} + 1\right)\frac{r}{1-r}, \quad (S32)$$

standard deviation,

$$\sigma_n = \frac{\sqrt{\left(\widehat{K_M} + 1\right)r}}{1-r}, \quad (S33)$$

and skewness

$$sk_n = \frac{1+r}{\sqrt{r\left(\widehat{K_M} + 1\right)}}. \quad (S34)$$

The coefficient of variation follows from Eqs. $(S32)$ and $(S33)$ as

$$cv_n = \frac{\sigma_n}{\langle n \rangle} = \frac{1}{\sqrt{\left(\widehat{K_M} + 1\right)\sqrt{r}}} \xrightarrow{\langle n \rangle \to \infty} \frac{1}{\sqrt{\left(\widehat{K_M} + 1\right)}}, \tag{S35}$$

i.e., a decreasing function of $\langle n \rangle$ that converges to an inferior plateau determined by $\widehat{K_M}$ only. The scaling between $\langle n \rangle$ and σ_n changes according to the relation between $\langle n \rangle$ and $\widehat{K_M}$, i.e.,

$$\langle n \rangle \ll \widehat{K_M} \;\; \sigma_n = \sqrt{\langle n \rangle \left(\frac{\langle n \rangle}{\left(\widehat{K_M} + 1\right)} + 1 \right)} \propto \sqrt{\langle n \rangle},$$

$$\langle n \rangle \gg \widehat{K_M} \;\; \sigma_n = \sqrt{\frac{\langle n \rangle^2}{\left(\widehat{K_M} + 1\right)} \left(1 + \frac{\left(\widehat{K_M} + 1\right)}{\langle n \rangle} \right)} \propto \langle n \rangle, \tag{S36}$$

transitioning from a square-root dependence (poisson-like), to a linear one, with minimal coefficient of variation $cv_n = \frac{1}{\sqrt{(\widehat{K_M}+1)}}$. Additionally, for $\widehat{K_M} \to \infty$, by virtue of the Central Limit Theorem, Eq. $(S31)$ is well approximated by a Gaussian distribution. Indeed, a negative binomial random variable can always be derived as the sum of geometric distributed random variables, with $\widehat{K_M}$ encoding the number of independent terms.

The coefficient of variation in Eq. $(S35)$ is inversely proportional to the square-root of the product of $\widehat{K_M}$ and r, potentially spanning a wide range of values, and implying that cell stochasticity could reproduce in principle the same level of nutrient variability observed in food composition data, quantified by the fit results in Section S2, with coefficient of variation $\approx e^{\alpha_\sigma}$. Hence, it becomes essential to understand the typical values of $\widehat{K_M}$ and r to quantify the contribution of cell stochasticity to our analysis. We will approximate their behavior in Subsection "Variability in K_M and Poisson-Log-normal Distribution".

Linear Pathways

The predominant motif in metabolic networks is a linear array of metabolites linked by chemical reactions whose energetics defines a "preferred" direction (a directed pathway)[16–18]. The CME for the linear pathway of length *N* in Figure 3A describes the temporal evolution of the joint probability distribution $p(n_1, \dots, n_N)$, i.e.,

$$\frac{dp(n_1, \dots, n_N, t)}{dt} = b[p(n_1 - 1, \dots, t) - p(n_1, \dots, t)]$$

$$+ \sum_{i=1}^{N-1} d_{n_i+1}\, p(\dots, n_i + 1, n_{i+1} - 1, \dots, t) - d_{n_i} p(\dots, n_i, n_{i+1}, \dots, t)$$

$$+ d_{n_N+1} p(\dots, n_N + 1, t) - d_{n_N} p(\dots, n_N, t), \qquad (S37)$$

where, at each step *I*, metabolite S_i is transformed in metabolite S_{i+1} by the enzyme E_i, working at turnover rate

$$d_{n_i} = \widehat{v_{max}^{i}} \frac{n_i}{n_i + \widehat{K_M^{i}}}. \qquad (S38)$$

The stochastic formulation of the directed pathway is analytically equivalent to a hopping model on 1-D lattice[19]. By analogy, a product-measure distribution, i.e., a joint probability distribution describing a system of *N* independent metabolites, is the stationary solution of Eq. (S37), and follows

$$p(n_1, \dots, n_N)^{ss} = \prod_{i=1}^{N} p^{ss}(n_i)\,, \qquad (S39)$$

where each $p^{ss}(n_i)$ is given by Eq. (4).

This analytical result implies:

1. At the steady-state each metabolite S_i has an independent probability distribution, determined only by the enzyme E_i consuming it, and its kinetic characteristics $\widehat{v_{max}^{i}}$ and $\widehat{K_M^i}$.
2. The single-metabolite distributions are all negative binomials as described in Eq. (4), with different parameters determined by each downstream enzyme.
3. At the steady state the chain length has no influence on the shape of single-metabolite probability distributions, as there is no "position-effect". This suggests that steady state fluctuations do not bear information about the pathway structure, while potential correlations between metabolite fluctuations could be determined, for instance, by the availability of a common enzyme or coenzyme[15,20,21]. However, the response of the system to external perturbations could be structure-dependent.

Extension of the result to other pathway topologies

In [12] a similar result to Eqs. (4) and (S39) was extended to diverging and cyclic pathways. In other scenarios, like reversible pathways or in presence of dilution, the product measure in Eq. (S39) is not always the exact solution, but it constitutes an effective ansatz, that shows excellent agreement with the numerical simulations. Here we prove that under steady state and free-ligand approximation Eq. (S26) is equivalent to the Boltzmann distribution used in [12]. Indeed, Eq. (S26) can be rewritten as

$$p(c|n)^{ss} = \frac{1}{Z(n,c_0)}\binom{c_0}{c}\left(\frac{\widehat{k_1}}{\widehat{k_{-1}}+\widehat{k_2}}\right)^c n^c \qquad (S40)$$

where $\left(\frac{\widehat{k_1}}{\widehat{k_{-1}}+\widehat{k_2}}\right)$ is the Boltzmann factor associated with the formation of an enzyme-substrate complex, and $Z(n, c_0)$ is a normalization factor independent from c. Additionally, when $n \gg c$ the remaining factor $n^c \approx \frac{n!}{(n-c)!}$, in agreement with [12].

Variability in K_M and Poisson-Log-normal Distribution

The $\mathcal{Q}(x_n)$ distribution captures the concentration of nutrient *n* across *multiple organisms* (food ingredients). To understand its origin we have to go beyond the metabolite stochasticity within the same organisms, as captured by (4), and determine the distribution of n_i across the different organisms we consume. In this case, the dominant source of variability is rooted in the different Michaelis-Menten kinetic constants $\widehat{K_M^i}$, that can vary several orders of magnitudes 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, effectively treating it as the mean field action of the metabolic network acting on substrate S_i. For further details see Subsection "Variability of *r*".

The stochastic nature of Eq. (4) is driven by the random events of replenishment and enzymatic turnover for a fixed value of $\widehat{K_M^i}$. Yet, when the reaction is conserved across organisms, each one of them has a different $\widehat{K_M^i}$, differences determined by multiple genetic variations that differentiate their enzymes and likely reflect the selective evolutionary processes each organism is subjected to. Consequently, when we study how the number of substrate molecules n_i is distributed across different organisms we measure a compound distribution, combining the effect of two different types of stochasticity:

a) Molecular kinetics of enzyme E_i,

b) Variability of $\widehat{K_M^i}$.

While we have provided a model for a), we still need to investigate the statistical properties of $\widehat{K_M^i}$, summarized by the probability distribution $p(\widehat{K_M^i})$ across different organisms.

The deterministic constant of Michaelis-Menten K_M is a measure of the substrate concentration required for significant catalysis to occur, usually measured in units or subunits of molar concentration (M). Similarly, the molecular Michaelis-Menten constant $\widehat{K_M}$ quantifies the effective number of substrate molecules necessary for catalysis. We can analytically relate the two constants with

$$K_M = \frac{\widehat{K_M}}{N_A V_{cell}} \approx \frac{\widehat{K_M}}{6 * 10^8} \approx \frac{\widehat{K_M}}{10^9}, \quad (S41)$$

where $V_{cell} = 1\ \mu m^3 = 10^{-15} l$, following the rule of thumb for cell volume (Bionumbers 101788). In BRENDA[22] K_M values are reported in mM, with quartiles Q_1=0.0240 mM ≈ 10^{-5} M, Q_2=0.1800 mM ≈ 10^{-4} M, Q_3=1.3135 mM ≈ 10^{-3} M (see Section S9). We validated the overall distribution of K_M in SABIO-RK[23], observing Q_1=2.2 *10^{-5} M, Q_2=1.5*10^{-4} M, Q_3=1.0*10^{-3} M. Therefore, according to our estimation, $\widehat{K_M}$ is expected to range between 10^4 and 10^6 , with median at 10^5. Our analysis is in agreement with previous studies relating 1,000 enzyme molecules per bacterium cell to approximately 1 μM concentration[16]. Moreover, in the observed range, Eq. (S31) appears fairly symmetric by virtue of the Central Limit Theorem ($\widehat{K_M} \to \infty$), independently from r, as quantified by the skewness in Eq. (S34).

According to the literature, under physiological conditions, the enzymes are not saturated with substrates, hence the ratio between substrate concentration x and K_M is typically in the range of 0.01 and 1.0[24]. We further investigated the relation between x and K_M, leveraging the

experimental records collected in SABIO-RK (download date 04/27/2021). Overall, we find a median x/K_M of 0.8173, with a significantly different behavior of cofactors like ATP, NAD+, NADPH, NADH, NADP+, and ADP (median x/K_M = 2.6393), compared to other substrates (median x/K_M=0.6494). Common cofactors are then present in high concentrations compared to their K_M, implying that their fluctuations in concentration do not affect the reaction rates, i.e., they can be factored as constants in enzymatic reactions that require multiple metabolites to occur. In summary, it is fairly common to observe $x < K_M$, with x never too low compared to K_M, a regime leading to a coefficient of variation $cv_n \approx 10^{-3}(Q_1) - 10^{-1}(Q_3)$, significantly smaller than the respective findings in food composition data. We estimate cv_n in Eq. (S35), combining Eq. (S41) with

$$r \approx \frac{\left(\frac{x}{K_M}\right)_{SABIO}}{1 + \left(\frac{x}{K_M}\right)_{SABIO}}. \tag{S42}$$

In presence of crowding effects, the effective volume available to the reaction is smaller compared to V_{cell}, leading to higher values of cv_n.

In the so-called "linear regime" the turnover rate given by Eq. (S27) is well approximated by

$$d_n \approx \frac{\widehat{v_{max}}}{\widehat{K_M}} n, \tag{S43}$$

leading to a steady state distribution described by Eq. (5), in agreement with the analytical requirements for the convergence of a negative binomial to a Poisson distribution. Hence, the variability pertaining to a) can be mapped to a more treatable analytical form determined by Eq. (5), by virtue of the natural ranges of K_M.

For the sake of simplicity, for now on we will use only K_M to refer to both deterministic and molecular constant of Michaelis-Menten, as their properties are invariant by rescaling. As we will show in Section S9, $p\left(K_M^i\right)$ is well approximated by a log-normal. Consequently, a Poisson distribution parametrized by a log-normal random variable leads to a Poisson-Log-normal form for $p^{organisms}(n_i)$, as described by Eq. (7). This probability distribution is often mentioned as "discrete log-normal" or "log-normal in disguise"[25,26], as for a sufficient large number of molecules n_i is well approximated by Eq. (8). The first order term of the Taylor series represents a log-normal distribution with logarithmic parameters $m_i = \log(r_i) + m\left(K_M^i\right)$ and $s_i = s\left(K_M^i\right)$. Moreover, the exact calculations for the linear mean and an variance of Eq. (7),

$$\langle n_i \rangle = r_i e^{m\left(K_M^i\right)+\frac{s\left(K_M^i\right)^2}{2}} = r_i \mu\left(K_M^i\right) \leq \mu\left(K_M^i\right), \tag{S44}$$

$$\sigma_n^2 = \left(r_i \mu\left(K_M^i\right)\right)^2 \left(e^{s\left(K_M^i\right)^2} - 1\right)\left(1 + \frac{1}{r_i \mu\left(K_M^i\right)\left(e^{s\left(K_M^i\right)^2} - 1\right)}\right), \tag{S45}$$

clarify how, for big enough $\mu\left(K_M^i\right)$ and/or non-negligible $s\left(K_M^i\right)$, $p^{organisms}(n_i)$ behaves as the log-normal $p\left(K_M^i\right)$, rescaled by a factor r_i.

General regime

Linear mean and standard deviation generalize to

$$\langle n_i \rangle = \frac{r_i}{1 - r_i} e^{m\left(K_M^i\right)+\frac{s\left(K_M^i\right)^2}{2}} = \frac{r_i}{1 - r_i} \mu\left(K_M^i\right), \tag{S46}$$

$$\sigma_n^2 = \left(\frac{r_i}{1 - r_i} \mu\left(K_M^i\right)\right)^2 \left(e^{s\left(K_M^i\right)^2} - 1\right)\left(1 + \frac{1}{r_i \mu\left(K_M^i\right)\left(e^{s\left(K_M^i\right)^2} - 1\right)}\right), \tag{S47}$$

beyond the Poisson regime.

Given the results regarding the plausible range of K_M, we expect Eq. (S31) to be well approximated by a Gaussian distribution, both in the Poisson regime and in case of enzyme saturation. When enzymes are saturated with substrates, Eq. (S31) will resemble a Gaussian distribution, over-dispersed compared to a Poisson with the same expectation value.

Variability of r

The data necessary to correctly estimate r and its variability is limited compared to what is currently available for K_M (see BRENDA and SABIO-RK), as we would need data on a similar variety of organisms for the incoming metabolic flux b and the maximal enzymatic rate v_{max}. We decided to treat r as a mean-field variable for the following reasons:

a) r depends on b, which embodies the mean field action of the metabolic network as an overall incoming flux to the pathway, rather than a specific reaction step.

b) models constrained by cost measures such as the molecules' carbon content, requirement of energy, or enzymes for their biosynthesis, find an optimal relation between K_M and substrate concentrations that is rate-independent[27].

We stress that in presence of multiplicative noise affecting r, our formalism relying on the product of r and K_M would still recover a log-normal distribution.

Secondary metabolites

Secondary metabolites are organic compounds produced by plants, fungi, and bacteria, which are not essential for growth, but usually produced in response to environmental stresses, such as predators, parasites, pathogens, light damage, and lack of nutrients. Their dependence on

the central/primary metabolism can be modeled as diverging pathways, branching out from the same chain of reactions. For the sake of simplicity, let's consider a directed pathway with a single branch point, leading to two branched directed chains BR1 and BR2. All metabolites before the branch point follow Eq. (4), or Eq. (7) in the linear regime. Always in the linear regime, the stationary probability distribution characterizing the metabolite S_{BRP}, substrate of two enzymes E_1^{BR1} and E_1^{BR2}, follows

$$p^{ss}(n_{BRP}) \approx \frac{1}{n_i!}(r_{BRP}(K_M^1)_{BR1}(K_M^1)_{BR2})^{n_i} e^{-r_{BRP}(K_M^1)_{BR1}(K_M^1)_{BR2}}, \tag{S48}$$

with

$$r_{BRP} = \frac{b}{(K_M^1)_{BR1}(\nu_{max}^1)_{BR2} + (K_M^1)_{BR2}(\nu_{max}^1)_{BR1}}. \tag{S49}$$

The fluxes on each of the two diverging pathways can be derived from Eq. (S48) as

$$b_{BR1} = \sum_n \frac{(\nu_{max}^1)_{BR1}}{(K_M^1)_{BR1}} n\, p^{ss}(n_{BRP}) = f_{BR1} b\,,$$

$$b_{BR2} = \sum_n \frac{(\nu_{max}^1)_{BR2}}{(K_M^1)_{BR2}} n\, p^{ss}(n_{BRP}) = f_{BR2} b\,, \tag{S50}$$

where both f_{BR1} and f_{BR2} are bounded between 0 and 1, and determine the partition of the initial substrate flux b in the two branches, i.e.,

$$f_{BR1} = \frac{(K_M^1)_{BR2}(\nu_{max}^1)_{BR1}}{(K_M^1)_{BR1}(\nu_{max}^1)_{BR2} + (K_M^1)_{BR2}(\nu_{max}^1)_{BR1}},$$

$$f_{BR2} = \frac{(K_M^1)_{BR1}(\nu_{max}^1)_{BR2}}{(K_M^1)_{BR1}(\nu_{max}^1)_{BR2} + (K_M^1)_{BR2}(\nu_{max}^1)_{BR1}}. \tag{S51}$$

Downstream metabolites can still be described by a Poisson distribution in the linear regime, but with a renormalized parameter compared to upstream metabolites. For example, the stationary distribution for metabolite *i* in branch BR1 is given by

$$p^{ss}(n_i) \approx \frac{1}{n_i!}\left(f_{BR1}(r_i)_{BR1}\left(K_M^i\right)_{BR1}\right)^{n_i} e^{-f_{BR1}(r_i)_{BR1}\left(K_M^i\right)_{BR1}}, \quad (S52)$$

where $(r_i)_{BR1} = \frac{b}{(v_{max}^i)_{BR1}}$. The presence of the additional factor f_{BR1}, which depends on the kinetic constants at the branch point, despite being a bounded value between 0 and 1, could be a source of additional variability across organisms, suggesting an increased variance of $p^{organisms}(n_i)$ for secondary metabolites, as observed for polyphenols in Section S6 and for secondary metabolites in raw vegetables and fruits in Section S4.

In summary, we have derived the following results:

(i) At the steady state the behavior of individual metabolites in pathways is driven by the behavior of Michaelis-Menten constants of the catabolizing enzymes.

(ii) If $p\left(K_M^i\right)$ is log-normal-distributed $p^{organisms}(n_i)$ behaves approximately as a rescaled log-normal with the same logarithmic standard deviation, i.e., $s_i = s\left(K_M^i\right)$. This result justifies the symmetric distributions in the log-space observed for the nutrients, supported by Eq. (2).

(iii) A bounded logarithmic standard deviation $s\left(K_M^i\right)$ across different enzyme-substrate pairs, fluctuating independently from the magnitude of K_M^i, would give an explanation for the bounded behavior of s_n, for the translational invariance observed in the nutrient log-space, and the linear scaling in Eq. (3).

(iv) A bounded logarithmic standard deviation across different enzyme-substrate pairs and nutrients implies that one of the two parameters necessary to fully describe a log-normal distribution is approximately the same for all chemicals analyzed. This observation is not a direct consequence of the multiplicative central limit (the observed distributions could be log-

normal with significantly different parameters), but it offers evidence of additional novel constraints determined by chemical correlations within the finite system.

Section S9: Experimental Validation with BRENDA

As according to the results presented in Section S8 and Eq. (5), the variability of Michaelis-Menten kinetic constants can drive $Q(x_n)$ in Eq. (2), here we investigate how K_M varies across all organisms that carry the same chemical reaction. We collected data for 93,692 experiments measuring K_M for several organisms, as reported in BRENDA flat files available for download[22]. We applied Natural Language Processing techniques on the free text comments describing each publication, to extract temperature, pH, and check if the enzyme tested was a mutant or recombinant. We removed all mutant and recombinant enzymes, keeping 70,873 experimental records measured in mM. Additionally, we leveraged NCBI Taxonomy[28] and ETE 3 package[29] to automatically classify into taxa all the different organisms reported in the database. These additional layers of meta-data allowed us to stratify the experimental records and control for potential correlation with environmental conditions. Overall, we observe very small correlation between K_M and temperature ($\rho_{Spearman} = 0.0927$), and even smaller correlation in absolute value with pH ($\rho_{Spearman} = -0.0232$) (Figure S17A and B). The data for Archaebacteria and Bacteria are collected at wider ranges of temperature, hence differently from Eukaryota, they show a stronger dependence on temperature (Table S2). Given our primary interest in food, here we focused mainly on eukaryotes. To identify which substrates are found in food, we additionally mapped the InChIKey of each molecule (if available) to our manually curated library of food molecules, currently containing 89,038 compounds reported by different food composition databases such as FooDB[13], Dictionary of Food Compounds[30], and detected in Mass Spectrometry experiments. The majority of the annotations in our library determines the presence or absence of a compound in food, but does not quantify its concentration.

From the obtained 31,662 enzyme-substrate pairs (E_i, S_i), we grouped the experimental

Taxa	Number of Records	$\rho_S(K_M, T)$	$\rho_S(K_M, pH)$
Archaea	2,878	0.1079	-0.0476
Bacteria	23,355	0.1306	0.0037
Eukaryota	42,950	0.0356	-0.0562
Viruses	105	-0.5466	0.0586
No Recombinant or Mutant Enzymes	70,873	0.0927	-0.0232
All records	93,692	0.0882	-0.0212

Table S2: Spearman Correlation of K_M with Temperature and pH, stratified by Taxa.

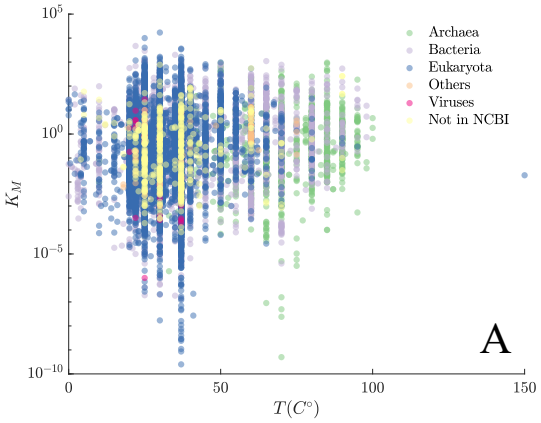
measurements for the same enzyme-substrate across different eukaryotes, obtaining the $p\left(K_M^i\right)$ distributions (Figure 3C).

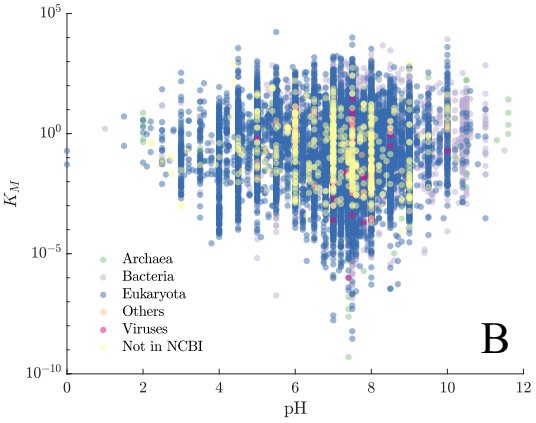
We focused on enzyme-substrate pairs (E_i, S_i) reported in at least 15 different eukaryotes, as the logarithmic standard deviation reaches a plateau, when plotted against the number of organisms (Figure S17C). Additionally, the logarithmic fluctuations characterizing all enzyme-substrate pairs (E_i, S_i) for a single organism show a behavior similar to the within food sample variability analyzed in Section S7. While within organism variability can in principle cover ranges compatible with those observed across multiple organisms, the shape of the distribution characterizing the logarithmic fluctuations appears to be different.

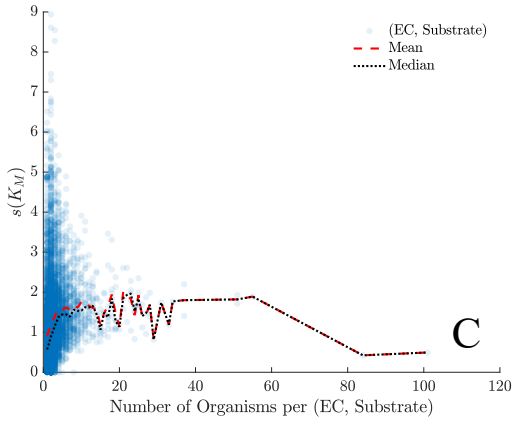
By running a statistical analysis similar to Section S3, we find that the log-normal distribution offers again the best approximation (Figure S17E), supported by the distribution of the logarithmic skewness fluctuating around zero (Figure S18A).

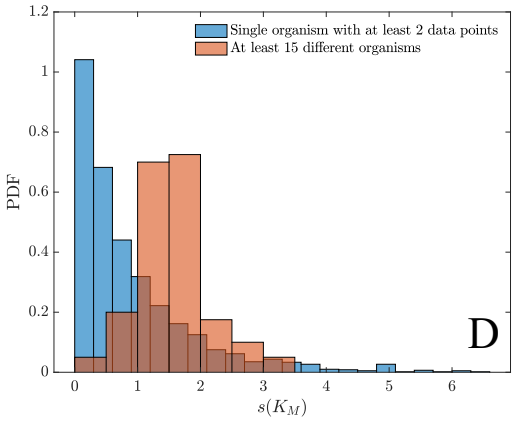
To investigate the robustness of our analysis we additionally grouped the experimental records by (EC, InChIKey), rather than by (EC, substrate name), to verify if further chemical disambiguation would impact our results, finding no significant differences. However, as not all substrates are assigned to InChIKey, we decided to present the analysis using substance names, to avoid further loss of experimental records (from 70,873 to 66,322).

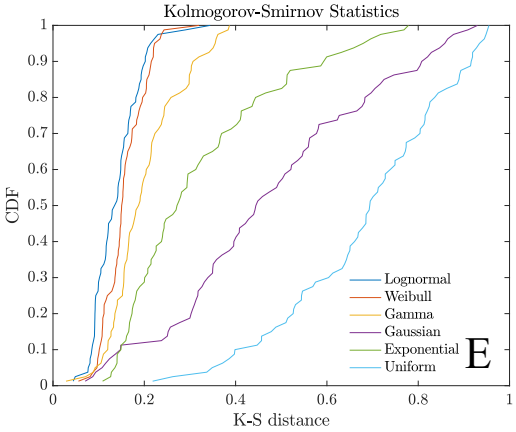
Finally, the experimental records reported in BRENDA show a bias towards specific taxonomic groups such as Viridiplantae (Green Plants) and Fungi, organisms characterized by the presence of a primary and a secondary metabolism, affecting the variability of the logarithmic fluctuations, in agreement with our findings for polyphenols in Section S6 and for secondary metabolites in raw vegetables and fruits in Section S4.


Archaea
Bacteria
Eukaryota
Others
Viruses
Not in NCBI
K_M
$T(C^\circ)$
A


Archaea
Bacteria
Eukaryota
Others
Viruses
Not in NCBI
K_M
pH
B


(EC, Substrate)
Mean
Median
$s(K_M)$
Number of Organisms per (EC, Substrate)
C


Single organism with at least 2 data points
At least 15 different organisms
PDF
$s(K_M)$
D


Kolmogorov-Smirnov Statistics
Lognormal
Weibull
Gamma
Gaussian
Exponential
Uniform
CDF
K-S distance
E

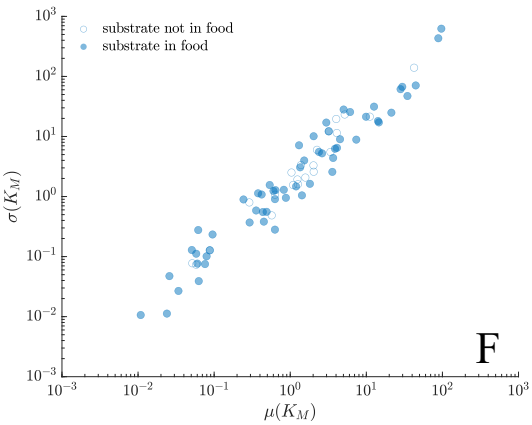

substrate not in food
substrate in food
$\sigma(K_M)$
$\mu(K_M)$
F

Figure S17. Michaelis-Menten Constants in BRENDA

(A) Relation with Temperature colored according to different taxa (Celsius), for 70,873 experimental records describing non-mutant or recombinant enzymes.

(B) Relation with pH colored according to different taxa, for 70,873 experimental records describing non-mutant or recombinant enzymes.

(C) Dependence of the logarithmic standard deviation $s(K_M)$ *on the number of organisms represented in each (EC, substrate) pair. With enough variety of organisms,* $s(K_M)$ *fluctuates around a plateau.*

(D) Logarithmic standard deviation $s(K_M)$ *calculated for (EC, substrate) pairs describing single organisms vs multiple organisms.*

(E) Kolmogorov-Smirnov distances for six different statistical distributions. Overall, log-normal offers the smallest statistical distances.

(F) The dependence of the standard deviation $\sigma(K_M)$ *on the average nutrient amount,* $\mu(K_M)$*. The relation is slightly superliner, and within the confidence intervals of what observed in Section S4 and Section S6 for polyphenols (given the high representation of organisms like green plants and fungi with extensive secondary metabolism), with* $\beta_\sigma = 1.098\ (1.038, 1.158)$*,* $\alpha_\sigma = 0.5869$*, (0.4631, 0.7107), and adjusted* $R^2 = 0.9438$*. We find similar results when we analyze the records for all organisms beyond eukaryotes, obtaining* $\beta_\sigma = 1.042\ (1.005, 1.079)$*,* $\alpha_\sigma = 0.678$*, (0.6022, 0.7538), and adjusted* $R^2 = 0.9380$*.*

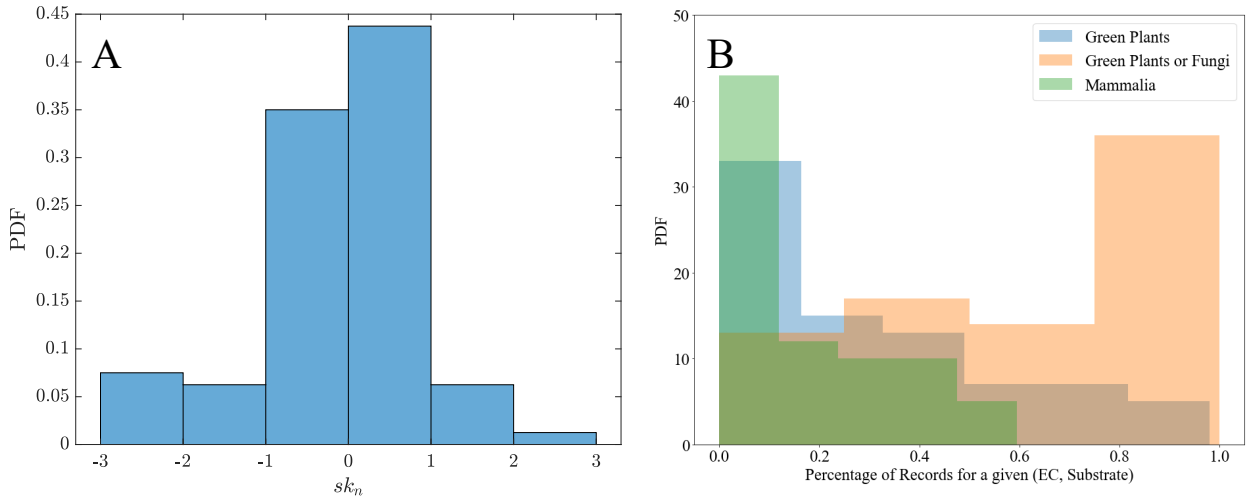


Figure S18. Michaelis-Menten Constants in BRENDA

(A) Skewness in the log-space for (EC, substrate) pairs with at least 15 different eukaryotes, fluctuating around zero.

(B) For each (EC, substrate) pair with at least 15 different eukaryotes we calculated the percentage of records provided by specific taxa or classes. Several (EC, substrate) pairs are predominantly determined by records pertaining Green Plants (Viridiplantae) and Fungi, organisms with secondary metabolism.

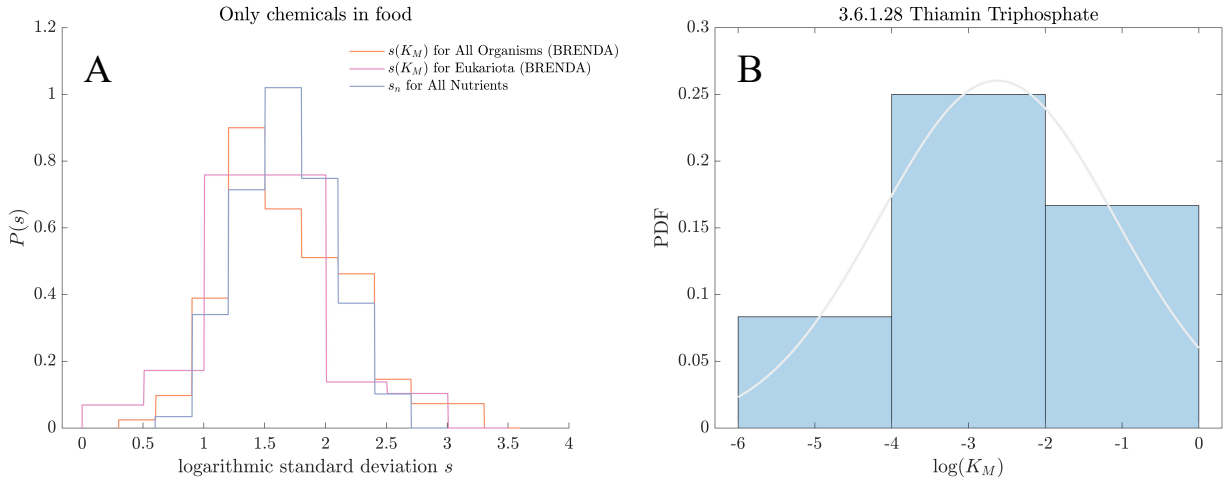


Figure S19. Substrates present in food.

(A) Similarly to Figure 3C in the manuscript, we illustrate here the agreement between s_n and $s(K_M^i)$ by plotting $P(s_n)$ for all nutrients, together with $P\left(s(K_M^i)\right)$ for all pairs (E_i, S_i) in BRENDA where S_i is annotated as chemical present in food. Given the limited chemical coverage of the USDA databases, we matched BRENDA with several databases of higher resolution like FooDB and Dictionary of Food Compounds, often lacking concentrations.

(B) Distribution of K_M for E.C. 3.6.1.128 and Thiamin Triphosphate, a chemical comparable with Thiamin concentrations reported in the USDA databases, as AOAC 942.23 fluorometric method accounts for phosphate additions (log-normal fit in gray, logarithmic standard deviation $s(K_M^i) = 1.2103$). Interestingly, the distribution of Thiamin in FNDDS exhibits a very similar logarithmic standard deviation equal to 1.2654 (Figure 1C). The organisms annotated in BRENDA for (3.6.1.128, Thiamin Triphosphate) are Bos taurus, Escherichia coli, Homo sapiens, Mus musculus, Rattus norvegicus, and Sus scrofa, and each of them is represented by the geometric mean of their K_M records.

Section S10: Log-normality of Protein Concentrations in the Literature

Food composition databases do not distinguish individual proteins but approximate the concentration of all proteins under a single datapoint, listed as "Protein" in Figures 1 and 2. These concentrations are determined on the basis of the total nitrogen content, an approach based on two assumptions: that dietary carbohydrates and fats do not contain nitrogen, and that nearly all of the nitrogen in the diet is present as amino acids in proteins[31].

Copy number variations consistent with log-normal distributions have been observed in individual protein concentrations in yeast and *E. coli*, prompting the development of theoretical models designed to capture the transcription and translation process driving protein production. These models predict copy number variations that range from the gamma distribution to fréchet and log-normal forms. Note however, that *these results refer to the variability between cells of the same species (i.e., between different copies of bacteria or yeast), and not to the variability between species,* the main topic of this paper. In the following, we summarize the pertinent results on individual protein copy number variations.

Biochemical complexity drives log-normal variation in genetic expression[32]

Beal et al. analyzed the expression levels of 12 *E. coli* engineered transcriptional repressor devices from Cello[33], expected to change the rate at which RNA polymerase binds and initiates transcription. The inspection of several repressor distributions found roughly symmetric

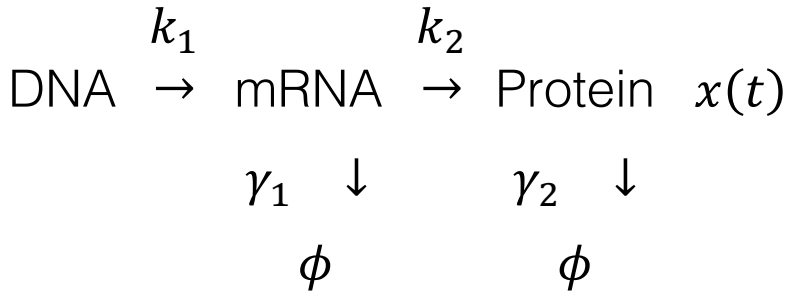


Figure S20. Stochastic bursting.

Transcription/translation model that predicts a stationary gamma distribution for the concentration x(t) of a protein of interest in the cell.

distributions, compatible with log-normal distributions with constant s_n. The findings of Ref. [32] represent a deviation for the standard model for protein concentration in live cells, which describes protein production as random bursts[6], whose stationary distribution follows the Gamma distribution,

$$p(x) = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{-\frac{x}{b}} \tag{S53}$$

where $a = k_1/\gamma_2$ is the mean number of bursts per cell cycle and $b = k_2/\gamma_1$ is the mean number of protein molecules produced per burst (Figure S20). As Ref. [32] has shown, for proteins with more than 10 molecules per cell on average, stochastic bursting leading to Eq. (S53) cannot explain the observed cell to cell variation[33], and log-normal offers a better fit.

Ubiquity of log-normal distributions in intra-cellular reaction dynamics[34]

Furusawa et al. investigated cell-to-cell variation of protein abundance in *E. coli* measuring the distribution of protein abundances in the exponential phase of growth. The distribution of the cell-to-cell normalized fluorescent intensity was reported to be consistent with a log-normal distribution through the different conditions of the promoter. While a log-normal form was not analytically derived, the authors argue that the log-normal distribution is rooted in the modular

structure of the catalytic reactions, where the fluctuations are multiplied successively through the cascading dependencies.

Universal protein fluctuations in populations of microorganisms[9]

Salman *et al.* investigated cell-to-cell variation of protein copy number in *E. coli* and yeast, measuring the distributions of highly expressed proteins in proliferating clonal populations, when gene expression is coupled to other cellular processes. The authors focused on two regulatory circuits, the LAC operon in *E. coli* and the GAL system in *S. Cerevisiae*, finding that the 15 studied GFP fluorescence distributions collapse into a single universal curve, after rescaling the level of fluorescence x with

$$y = \frac{x - \mu}{\sigma}, \tag{S54}$$

where $\mu = \langle x \rangle$ indicates the expected number of proteins in the cell, and $\sigma = \sqrt{\langle x^2 \rangle - \langle x \rangle^2}$ determines the dynamic range of protein content.

Salman et al. propose as phenomenological description of the data using a family of fréchet distributions. They also report that some individual experiments were better described by log-normal, as in [32] and [34], or by gamma as in [6], acknowledging that fréchet and log-normal distribution are hard to distinguish with the available data.

Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells[35]

Taniguchi et al. investigated the cell-to-cell variation of mRNA and protein copy numbers in *E. coli*, by carrying out a quantitative system-wide single-cell global profiling with single-molecule sensitivity, focusing on protein copy numbers per average cell volume. Average protein abundances spanned five orders of magnitude (10^{-1} to 10^4 molecules per cell).

The authors considered the kinetic model of Figure S20, modeling protein production as random bursts[6] and leading to a gamma stationary distribution for protein concentration, as described in Eq. (S53). They find that gamma fits better than log-normal for proteins with low expression levels, while for higher expression levels (>10 molecules per cell) log-normal performs equally or better than gamma, in agreement with [32]. Additionally, the authors observe a relation between average protein copy number, standard deviation, and variance to mean ratio, functionally compatible with the stochastic relations derived in Eqs. (S35)-(S36). In particular, they find a linear regime with coefficient of variation ~ 0.3162.

In summary, the line of work on copy number variations in protein synthesis could help us better understand protein variability in food as well. Such modeling would be particularly important once efforts are made to systematically track the concentration of individual proteins in various foods, not only the total protein content reported today.

Section 10b: Other relevant work in the literature

An investigation into the population abundance distribution of mRNAs, proteins, and metabolites in biological systems[36]

Differently from the body of work on cell-to-cell copy number variations in single protein synthesis, Lu et al. implemented a large-scale analysis of the population abundance distributions in 10 transcriptomic datasets, 7 proteomic datasets, and 5 metabolomic datasets. Each dataset analyzed pertained to a given omics technology and organism. The authors find that the Pareto-Log-normal, a distribution that behaves like a log-normal near the center and like power law in the tail, shows the best-fit for the population abundance distributions. Note however, that these results

refer the distribution of the abundances of different molecular species within the same organism, and do not characterize the variability of a given type of molecule between cells of the same organism, or across different organisms.

Section S11: Dry-Weight Analysis

To investigate if water content variation drives the scaling laws (2) and (3), we re-examined our findings by normalizing the nutrient concentrations with the specific "dry weight" of each food, i.e., the total dry matter per 100 g of each dish/drink. The amount of nutrient *n* in 100 g of each dish *d* is therefore rescaled as

$$F_{nd}^{dryweight} = \frac{F_{nd}}{100 - F_{water-d}}, \qquad (S55)$$

where $F_{water-d}$ is the related water content. Overall, we observe a consistent behavior for all the statistics, as shown in Figure S21A-D. The renormalization partially affects chemical families like the flavonoids in tea products, which can be consumed as powder as well as drinks, introducing renormalization factors with higher variability compared to the majority of the food items.

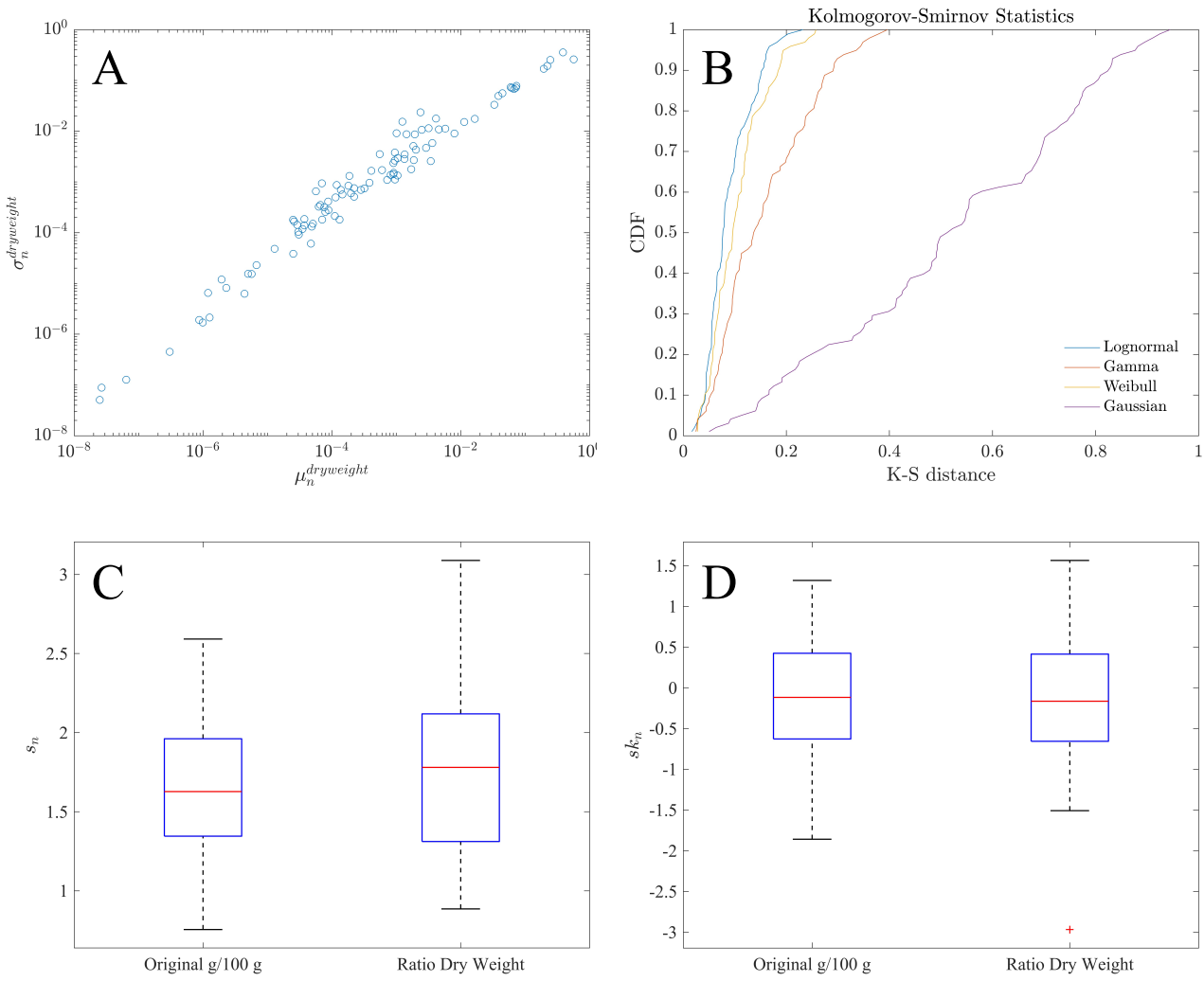


Figure S21. Dry-Weight Analysis.

(A) Relation between average nutrient amount $\mu_n^{dryweight}$ *and standard deviation* $\sigma_n^{dryweight}$*. For all foods where nutrient n is quantified we calculate* $\mu_n^{dryweight}$*,* $\sigma_n^{dryweight}$ *and then we evaluate their functional relation* $\sigma_n^{dryweight} = e^{\alpha_\sigma}\left(\mu_n^{dryweight}\right)^{\beta_\sigma}$ *across the whole nutrient panel (excluding water). The fitted relation is now more sublinear compared to the findings in Section S2, with exponents with* $\beta_\sigma = 0.9162\ (0.8806, 0.9518)$*,* $\alpha_\sigma = 0.5571$ *(-0.0429, 0.5824), and adjusted* $R^2 = 0.9642$*, but with overlapping confidence intervals.*

(B) Kolmogorov-Smirnov distances for four different statistical distribution, showing that log-normal still offers the best approximation.

(C) Comparison of the distributions of logarithmic standard deviations s_n*, with the original dataset on the left, and the rescaled dataset of the right.*

(D) Comparison of the distributions of logarithmic skewness sk_n*, with the original dataset on the left, and the rescaled dataset of the right.*

1. USDA. USDA Food and Nutrient Database for Dietary Studies, 5.0. (2012). Available at: http://www.ars.usda.gov/ba/bhnrc/fsrg.
2. Sebastian, R. *et al.* Flavonoid Values for USDA Survey Foods and Beverages 2007-2010. (2016). Available at: www.ars.usda.gov/nea/bhnrc/fsrg.
3. FNDDS Web Page. Available at: https://www.ars.usda.gov/northeast-area/beltsville-md-bhnrc/beltsville-human-nutrition-research-center/food-surveys-research-group/docs/fndds/.
4. Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D. & Pletnev, I. InChI - The worldwide chemical structure identifier standard. *J. Cheminform.* **5**, 1 (2013).
5. Kapur, J. N. *Maximum-Entropy Models in Science and Engineering. Biometrics* (1992). doi:10.2307/2532770
6. Friedman, N., Cai, L. & Xie, X. S. Linking stochastic dynamics to population distribution: An analytical framework of gene expression. *Phys. Rev. Lett.* **97**, 1–4 (2006).
7. Wohletz, K. H., Sheridan, M. F. & Brown, W. K. Particle size distributions and the sequential fragmentation/transport theory applied to volcanic ash. *J. Geophys. Res.* **94**, 15703 (1989).
8. Everitt, B. S. & Skrondal, A. *The Cambridge Dictionary of Statistics*. *Journal of Chemical Information and Modeling* **53**, (2010).
9. Salman, H. *et al.* Universal protein fluctuations in populations of microorganisms. *Phys. Rev. Lett.* **108**, 1–5 (2012).
10. U.S. Department of Agriculture, A. R. S. FoodData Central: Foundation Foods. (2019). Available at: fdc.nal.usda.gov. (Accessed: 1st April 2020)
11. Mehta, N. B., Wu, J., Molisch, A. F. & Zhang, J. Approximating a sum of random

variables with a log-normal. *IEEE Trans. Wirel. Commun.* **6**, 2690–2699 (2007).

12. Neveu, V. *et al.* Phenol-Explorer: an online comprehensive database on polyphenol contents in foods. *Database (Oxford).* **2010**, bap024 (2010).
13. WishartLab. FooDB. (2017). Available at: http://foodb.ca/.
14. Ge, H. & Qian, H. Chemical Master Equation. in *Encyclopedia of Systems Biology* 396–399 (Springer New York, 2013). doi:10.1007/978-1-4419-9863-7_278
15. Stéfanini, M. O., McKane, A. J. & Newman, T. J. Single enzyme pathways and substrate fluctuations. *Nonlinearity* **18**, 1575–1595 (2005).
16. Levine, E. & Hwa, T. Stochastic fluctuations in metabolic pathways. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 9224–9229 (2007).
17. Michal, G. & Schomburg, D. *Biochemical Pathways: An Atlas of Biochemistry and Molecular Biology: Second Edition*. *Biochemical Pathways: An Atlas of Biochemistry and Molecular Biology: Second Edition* (John Wiley & Sons, 2013). doi:10.1002/9781118657072
18. 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).
19. Levine, E., Mukamel, D. & Schütz, G. M. Zero-range process with open boundaries. *J. Stat. Phys.* **120**, 759–778 (2005).
20. English, B. P. *et al.* Ever-fluctuating single enzyme molecules: Michaelis-Menten equation revisited. *Nat. Chem. Biol.* **2**, 87–94 (2006).
21. Steuer, R., Kurths, J., Fiehn, O. & Weckwerth, W. Observing and interpreting correlations in metabolomic networks. *Bioinformatics* **19**, 1019–1026 (2003).
22. Placzek, S. *et al.* BRENDA in 2017: New perspectives and new tools in BRENDA.

Nucleic Acids Res. **45**, D380–D388 (2017).

23. Wittig, U. *et al.* SABIO-RK - Database for biochemical reaction kinetics. *Nucleic Acids Res.* **40**, D790–D796 (2012).

24. Stryer, L., Berg, M. J. & Tymoczko, L. J. *Biochemistry.2002.* (W. H. Freeman and Company, 2002).

25. Stewart, J. A. The poisson-log-normal model for bibliometric/scientometric distributions. *Inf. Process. Manag.* **30**, 239–251 (1994).

26. Bulmer, A. M. G. On Fitting the Poisson Log-normal Distribution to Species-Abundance Data. **30**, 101–110 (2016).

27. Dourado, H., Maurino, V. & Lercher, M. Enzymes and Substrates Are Balanced at Minimal Combined Mass Concentration in vivo. *bioRxiv* 128009 (2017). doi:10.1101/128009

28. National Center for Biotechnology Information. NCBI Taxonomy. Available at: https://www.ncbi.nlm.nih.gov/taxonomy.

29. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).

30. Yannai, S. *Dictionary of food compounds with CD-ROM*. *Choice Reviews Online* **51**, (Taylor & Francis, 2013).

31. USDA. National Nutrient Database for Standard Reference, Release 28 (2015) Documentation and User Guide. **28**, (2015).

32. Beal, J. Biochemical complexity drives log-normal variation in genetic expression. *Eng. Biol.* **1**, 55–60 (2017).

33. Nielsen, A. A. K. *et al.* Genetic circuit design automation. *Science (80-.).* **352**, (2016).

34. Furusawa, C., Suzuki, T., Kashiwagi, A., Yomo, T. & Kaneko, K. Ubiquity of Log-normal Distributions in Intra-cellular Reaction Dynamic. **1**, 25–31 (2005).

35. Taniguchi, Y. *et al.* Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. *Science (80-.).* **329**, 533–539 (2011).

36. Lu, C. & King, R. D. An investigation into the population abundance distribution of mRNAs, proteins, and metabolites in biological systems. *Bioinformatics* **25**, 2020–2027 (2009).