Estimating Nutrient Concentration in Food Using Untargeted Metabolomics

Michael Sebek[1,+], Giulia Menichetti[1,2,+], Albert-László Barabási[1,2,3†]

[1]Network Science Institute and Department of Physics, Northeastern University, Boston, MA, USA

[2]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

[3]Department of Network and Data Science, Central European University, Budapest, Hungary

[+]These authors contributed equally to this work

Abstract

Untargeted metabolomics can detect hundreds of compounds in food, yet without standards, it cannot quantify them. Here we show that we can take advantage of the universal scaling of nutrient concentrations to estimate the concentration of all chemicals detected by untargeted metabolomics. We validate our method on 20 raw foods, finding an excellent agreement between the predicted and the experimentally observed concentrations.

[†]Corresponding author e-mail: barabasi@gmail.com

The documented impact of diet on health has prompted national efforts to catalogue the biochemical composition of food[1], like the USDA Standard Reference[2,3] that reports the concentration of 150 nutrients in 9,759 foods. The data was collected using gravimetry (AOAC934.06), liquid chromatography (AOAC988.15), and fluorimetry (AOAC967.22), experimental methodologies that can detect and quantify the concentration of a compound in each food. While untargeted metabolomics can detect hundreds more compounds within a single experiment, it can only infer their relative concentration, mainly because of the unknown fraction of molecules ionized for each compound[4,5]. Targeted metabolomics circumvents ionization efficiency[6] using standard curves to determine concentration of each compound, a low-throughput procedure that eliminates the advantages of untargeted metabolomics over established methodologies.

Here we show that the universal scaling of compound concentrations across the food supply[7] allows us to determine compound concentration in untargeted metabolomics. We performed metabolomics experiments on 20 raw plant ingredients spanning fruits, leaves, roots, and legumes (SI Experimental Design). Focusing on the 295 compounds found in at least 10 ingredients, we obtained their nutrient distribution across the food samples using the experimentally measured peak areas. In metabolomics, the peak area for each compound, n, measures only the ionized concentration, $c_{ionized,n}$. The unknown ionization efficiency, IE_n, prevents peak area from directly measuring the total concentration, c_n, following $c_{ionized,n} = IE_n c_n$. The primary contributors to IE_n are the instrumental parameters, the composition of the mobile phases, and the sample matrix (ionization competition between co-eluting compounds). Since in our experiments all samples were measured using the same experimental setup, the differing sample matrices provide the largest contribution to ionization efficiency differences. We therefore tested the previously detected universal features of compound concentrations[7]: (i) constant standard deviation ($\langle s_n \rangle$), (ii) symmetric distribution, and (iii) translational invariance, finding that they apply to the peak areas as well, with $\langle s_n \rangle = 1.41$

$\pm$ 0.50. These findings suggest that

$$IE_n \approx \langle IE_n \rangle \tag{1}$$

for each compound, where $\langle IE_n \rangle$ is the mean ionization efficiency across all foods. We also find that the peak areas follow log-normal distributions similar to the nutrient concentrations in [7] (SI Universal Scaling Law of Metabolomics). Lastly, we compare the universal scaling observed for compounds reported by the USDA (Fig 1a) with the one obtained by our experiments (Fig 1b). We find that the linear standard deviation (σ_n^{MS}) relates to the mean peak area (μ_n^{MS}) in the log-space via $\sigma_n^{MS} = e^{\alpha_\sigma^{MS}} (\mu_n^{MS})^{\beta_\sigma^{MS}}$, in excellent agreement with the USDA.

Next, we test the relationship between the metabolomics variables (σ_n^{MS}, μ_n^{MS}) and their respective counterparts of the USDA (σ_n^{US}, μ_n^{US}). Focusing on 19 foods and 31 compounds reported in both datasets (SI USDA – Metabolomics Overlap), we find low correlation (R^2 = 0.566), confirming that ionization efficiency obscures the proportionality between peak area and concentration via $\mu_n^{US} = \langle IE_n \rangle \ \mu_n^{MS}$. We can, however, estimate the concentration from peak area by leveraging (1) and the universal features (i)-(iii), followed by both metabolomics and USDA, indicating that the position of foods within the nutrient distributions between the datasets is largely conserved. For example, we can see that for pyridoxine (vitamin B6) the relative position of potato is the same in the metabolomics and the USDA distributions (Fig 1c). This suggests that we can use the distance between the individual food measurements and the median in the linear-space to connect the two distributions via

$$\frac{x_{(f,n)}^{US}}{e^{m_n^{US}}} \approx \frac{x_{(f,n)}^{MS}}{e^{m_n^{MS}}}, \tag{2}$$

where $x_{(f,n)}^{US}$ is the concentration of the compound for food (f) from the USDA, m_n^{US} is the mean log concentration of the compound in the USDA, $x_{(f,n)}^{MS}$ is the peak area of the compound in the food from experiments, and m_n^{MS} is the mean log peak area of the compound in the

experiments (SI Proportionality Validation).

To assess the validity of (2) we use 113 food-nutrient pairs in overlap between the two datasets to estimate $x^{US}_{(f,n)}$ and compare to the reported value in the USDA by calculating fold change (FC),

$$FC = \begin{cases} \frac{x^{est}_{(f,n)}}{x^{US}_{(f,n)}} & x^{US}_{(f,n)} < x^{est}_{(f,n)} \\ \frac{x^{US}_{(f,n)}}{x^{est}_{(f,n)}} & x^{US}_{(f,n)} \geq x^{est}_{(f,n)} \end{cases} \tag{3}$$

where $x^{est}_{(f,n)}$ is the estimated concentration (SI Curated Pairs). Using (3), we observe a 3.1 mean FC and a 2.1 median FC for the remaining $x^{est}_{(f,n)}$ values (Fig 2a, blue line) with 73% of the values below 4.0 FC, an outcome comparable with other untargeted metabolomics estimation methods[8], which report a 2.0-4.0 mean FC.

The methodology (2) requires the concentration m^{US}_n of the compounds, known only for compounds reported in the USDA. To overcome this limitation, we rely on the finding that compound concentrations in bacterial and human cells are correlated with their chemical properties of the compounds[9], allowing us to infer m^{US}_n for nutrients not present in the USDA. We created a XGBoost model to predict $x^{US}_{(f,n)}$, taking as input the chemical properties of nutrients (molecular weight, logP, logS, hydrogen bonding inventory, number of charged atoms, non-polar surface area) and phylogenetic lineages of foods (class, order, family, genus classifications). Leave-one-out validation of the trained model shows 70% of the $x^{US}_{(f,n)}$ values within 2.0 FC and 90% within 4.0 FC of the true value ($R^2 = 0.931$, Fig 2b), confirming the model is accurate (SI Gradient Boosting Methodology).

Using XGBoost to estimate m^{US}_n, (2) estimates the concentration of compounds in individual foods from peak areas, observing a 3.4 mean FC (Fig 2a, red line) and 73% of the values below a 4.0 FC. XGBoost allows us to determine the concentration of compounds not reported in the USDA, but detected in our experiments. For example, while S-allylcysteine in garlic is not reported by USDA, our untargeted experiments allow us to estimate its concentration as 0.158 g/100g. Using FoodMine[10], we found six published measurements for S-allylcysteine in garlic, with an average at 0.115 g/100g, giving a 1.4 FC, confirming that

untargeted metabolomics can accurately predict the concentration of a compound within a specific food. We find that our method, relying on the universal features of nutrient distributions, offers comparable performance to the methods that are currently used to estimate concentration in untargeted metabolomics[8]. Structural similarity and elution proximity methods require internal standards and offer mean folds around 3.8. Our approach achieves comparable FC without requiring the purchase or the development of standards. Other machine learning methods provide a 2.1 mean FC performance, but require an extensive database of ionization efficiencies, which are rarely available. In summary, our methodology relies on training data that is publicly available and easily accessible, and allows us to estimate the concentration of compounds in untargeted metabolomics with comparable accuracy to existing methods without their known limitations.

References

[1] European Food Information Resource (2009).at <https://www.eurofir.org/food-information/food-composition-databases/>

[2] U.S. Department of Agriculture & Agricultural Research Service (2019).at <fdc.nal.usda.gov>

[3] U.S. Department of Agriculture, Agricultural Research Service & Laboratory, N.D. (2016).at <http://www.ars.usda.gov/ba/bhnrc/ndl>

[4] Purves, R.W., Gabryelski, W. & Li, L. Rapid Commun Mass Spectrom 12, 695–700 (1998).

[5] Müller, C., Schäfer, P., Störtzel, M., Vogt, S. & Weinmann, W. J. Chromatogr. B 773, 47–52 (2002).

[6] Ankney, J.A., Muneer, A. & Chen, X. Annu. Rev. Anal. Chem. 11, 49–77 (2018).

[7] Menichetti, G. & Barabási, A.-L. Nat. Food 1–66 (2021).

[8] Kruve, A., Kiefer, K. & Hollender, J. Anal. Bioanal. Chem. 413, 1549–1559 (2021).

[9] Bar-Even, A., Noor, E., Flamholz, A., Buescher, J.M. & Milo, R. PLoS Comput. Biol. 7, (2011).

[10] Hooton, F., Menichetti, G. & Barabási, A.L. Sci. Rep. 10, 16191 (2020).

Acknowledgements

We thank Dr. Roger Giese and Dr. Pushkar M. Kulkarni at the Department of Pharmaceutical Sciences, Northeastern University for preparing the food samples. Experiments were performed by Metabolon Inc., led by Nino Esile, Dr. Brian Ingram, and Nathan Testa. This work was partially supported by NIH grant 1P01HL132825, American Heart Association

grant 151708, ERC grant 810115-DYNASET, and Rockefeller Foundation grant 2019 FOD 026.

Data and Code Availability

The data and codes are openly available at our GitHub page at `https://github.com/Barabasi-Lab/Quantifying-Untargeted-Metabolomics`.

Author Contributions

A.L.B. and G.M. conceived the research. M.S. led the metabolomics experiments and performed the data analysis. A.L.B. and G.M. advised and guided the research. M.S. wrote the manuscript. A.L.B. and G.M. reviewed and edited the manuscript.

Competing interests

A.L.B. is a scientific founder of Scipher Medicine, Inc., which applies network medicine strategies to personalized drug selection, and Naring, Inc., which applies data science to food and health.

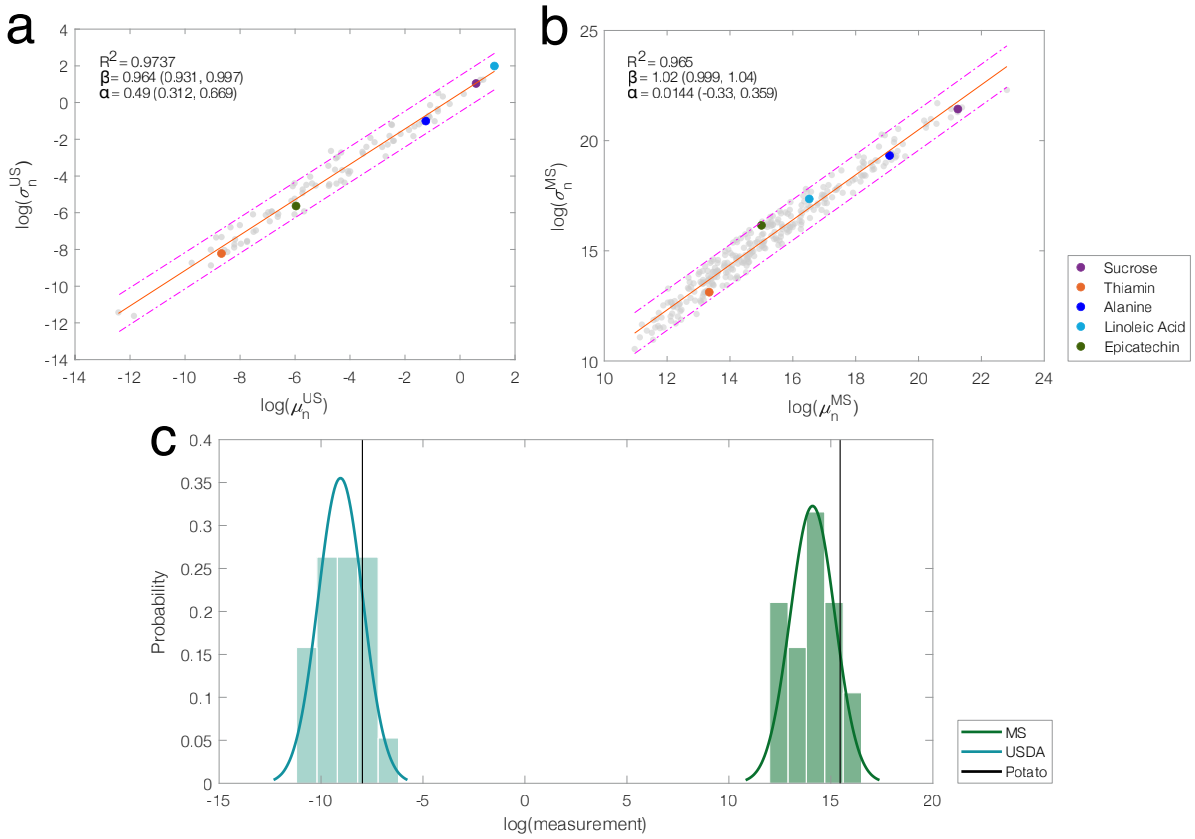


Figure 1: **Estimating Concentrations in Untargeted Metabolomics.** **(a)** Universal scaling law for nutrients within the USDA subset, relating the standard deviation, σ_n^{US} vertical axis, to the average nutrient concentration, μ_n^{US} horizontal axis, for 94 nutrients in 510 foods. Dashed lines are confidence intervals. Colored dots are compounds we selected to compare their relative positions between a) and b). **(b)** Universal scaling for our untargeted metabolomics experiments, relating the standard deviation, σ_n^{MS}, to the mean peak area, μ_n^{MS}, for 295 compounds in 20 foods. **(c)** Nutrient distributions using the 19 foods in overlap between the USDA and our experiments for pyridoxine: experiments (peak area, red) and USDA (g/100g, blue). The position of potato is shown in each distribution (black lines).

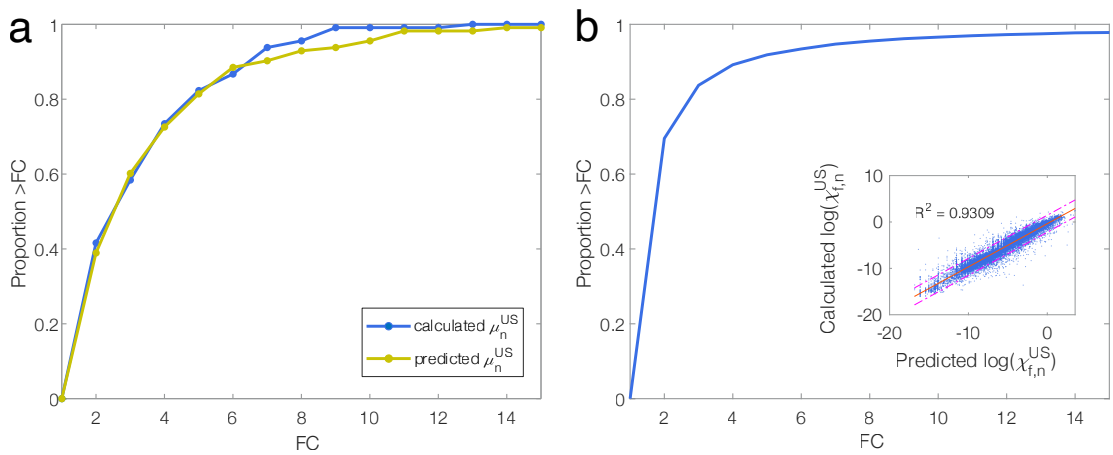
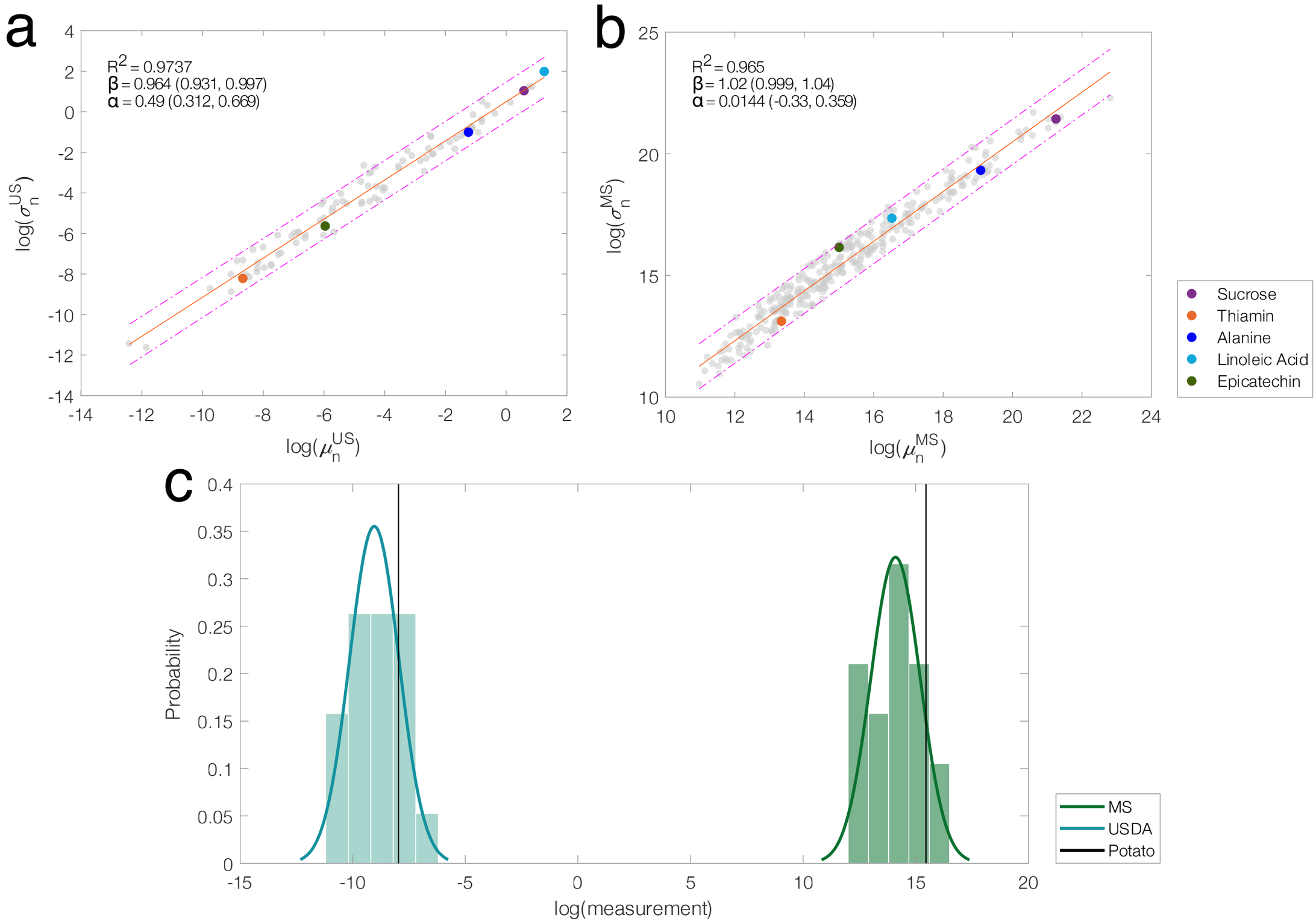
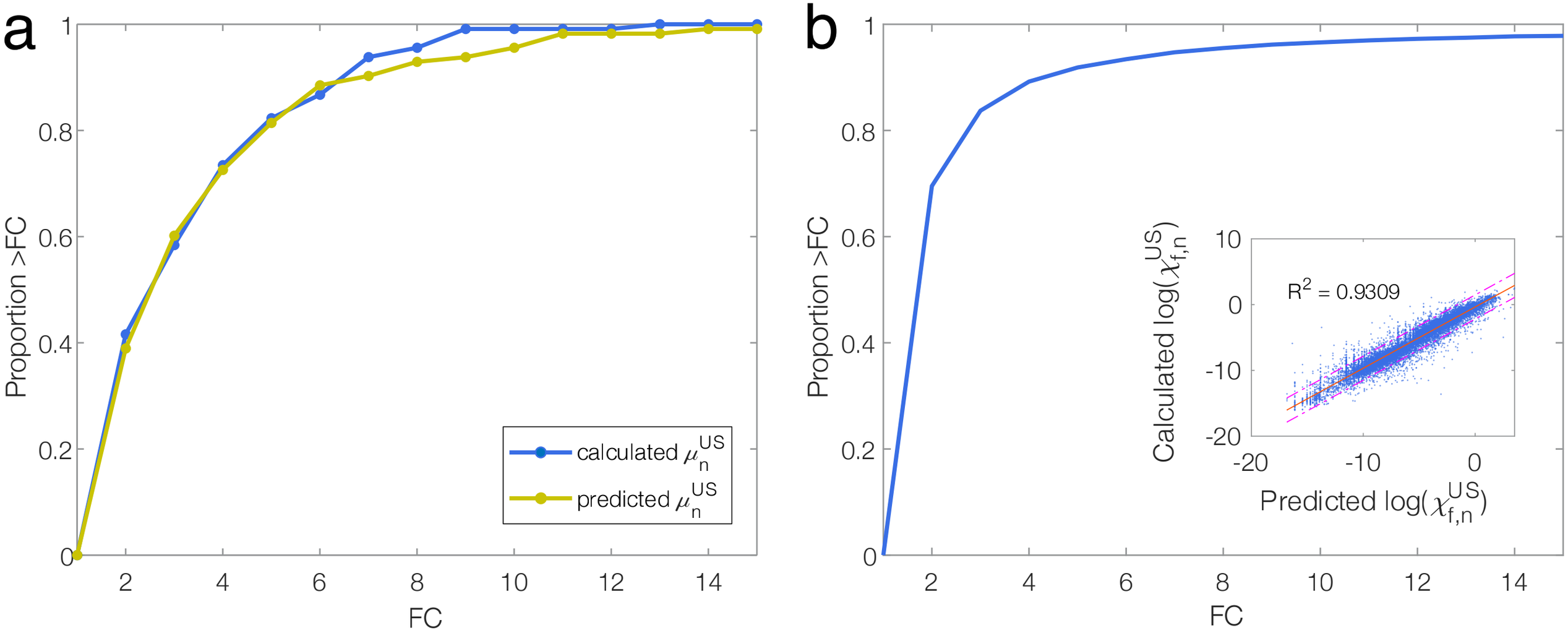


Figure 2: **FC Curves.** The proportion of estimated nutrient concentrations in individual food items below specified FC values. **(a)** FC curve plots the calculated μ_n^{US} from the USDA (blue) and predicted μ_n^{US} from a XGBoost model (red), taking as input the chemical properties of nutrients (molecular weight, logP, logS, hydrogen bonding inventory, number of charged atoms, non-polar surface area) and phylogenetic lineages of foods (class, order, family, genus classifications). **(b)** FC curve of the Leave-one-out predicted concentrations. Inset: the relationship between calculated $x_{(f,n)}^{US}$ from the USDA and predicted $x_{(f,n)}^{US}$ from the XGBoost model. This shows that we can predict the concentrations reported in the USDA with a correlation of $R^2 = 0.931$ between the predicted and real concentrations.


a
$R^2 = 0.9737$
$\beta = 0.964$ (0.931, 0.997)
$\alpha = 0.49$ (0.312, 0.669)
$\log(\sigma_n^{US})$
$\log(\mu_n^{US})$
b
$R^2 = 0.965$
$\beta = 1.02$ (0.999, 1.04)
$\alpha = 0.0144$ (-0.33, 0.359)
$\log(\sigma_n^{MS})$
$\log(\mu_n^{MS})$
Sucrose
Thiamin
Alanine
Linoleic Acid
Epicatechin
c
Probability
log(measurement)
MS
USDA
Potato


a
b
Proportion >FC
FC
calculated μ_n^{US}
predicted μ_n^{US}
R^2 = 0.9309
Calculated log($\chi_{f,n}^{US}$)
Predicted log($\chi_{f,n}^{US}$)