Microbiome-based correction for random errors in nutrient profiles derived from self-reported dietary assessments

Tong Wang[1], Yuanqing Fu[2,3,4], Menglei Shuai[2,3,5], Ju-Sheng Zheng[2,3,4], Lu Zhu[6], Andrew T. Chan[1,4,7,8,9], Qi Sun[1,10,11], Frank B. Hu[1,10,11], Scott T. Weiss[1], Yang-Yu Liu[1,12,*]

[1]*Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA*

[2]*School of Life Sciences, Westlake University, Hangzhou, China*

[3]*Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, China*

[4]*Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, Hangzhou, China*

[5]*Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA*

[6]*Department of Epidemiology, University of Iowa College of Public Health, Iowa City, IA 52242, USA*

[7]*Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA*

[8]*Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA*

[9]*Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA*

[10]*Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA*

[11]*Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA*

[12]*Center for Artificial Intelligence and Modeling, The Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*

* Correspondence: yyl@channing.harvard.edu

Abstract

Since dietary intake is challenging to directly measure in large-scale cohort studies, we often rely on self-reported instruments (e.g., food frequency questionnaires, 24-hour recalls, and diet records) developed in nutritional epidemiology. Those self-reported instruments are prone to measurement errors, which can lead to inaccuracies in the calculation of nutrient profiles. Currently, few computational methods exist to address this problem. In the present study, we introduce a deep-learning approach --- **M**icrobiom**e**-based nu**t**rient p**r**ofi**l**e **c**orrector (METRIC), which leverages gut microbial compositions to correct random errors in self-reported dietary assessments using 24-hour recalls or diet records. We demonstrate the excellent performance of METRIC in minimizing the simulated random errors, particularly for nutrients metabolized by gut bacteria in both synthetic and three real-world datasets. Further research is warranted to examine the utility of METRIC to correct actual measurement errors in self-reported dietary assessment instruments.

Introduction

An unhealthy diet can increase the risk of many diseases[1,2,3,4]. For instance, excess intake of sugar or saturated fat could elevate the risk of coronary heart disease[5–8]. The investigation of the association between poor dietary habits and chronic diseases requires an accurate assessment of dietary intake in large population samples. Typically, epidemiologic studies rely on self-reported instruments such as food frequency questionnaires (FFQ)[9], ASA24 (Automated Self-Administered 24-hour Dietary Assessment Tool)[10], and 7-Day Dietary Record (7DDR)[11], to collect diet data in large populations. However, self-reported instruments are subject to both random and systematic measurement errors[12–15]. While random errors are largely due to day-to-day individual variations in dietary intakes, systematic errors can be caused by under- or over-reporting and inaccurate estimation of portion sizes[16]. Measurement errors in the assessed food intake will be naturally propagated forward to the computation of the nutrient profile, causing inaccuracies in the assessed nutrient profile (Fig. 1a). Correcting such errors in the nutrient profile is crucial for improving the quality of nutritional epidemiology research. Several methods, such as regression calibrations[17,18] and cumulative averages[19] using repeated dietary assessments, have been used to correct for random errors in nutritional studies. However, few deep-learning computational methods have been used for random error corrections.

Signal reconstruction from corrupted or incomplete measurements is a common problem in various fields of research. For example, image denoising (i.e., removing noise from a noisy image to restore the true image) is an important topic in computer vision. The recent advance in deep neural networks has led to new methods that learn to map corrupted images to unobserved clean ones[20,21]. However, training those methods typically requires clean images, which are difficult to obtain in many cases. To overcome this challenge, a new class of methods that only leverages noisy images in training has been developed[22,23]. One noticeable example is Noise2Noise[22], which restores clean images by only being trained on corrupted ones. When the noise is random with a mean of zero, its performance is comparable and sometimes even better than other methods trained using clean images[22].

Inspired by the success of Noise2Noise[22], we hypothesize that we can correct the random error in the assessed nutrient profile derived from the self-reported dietary assessment without using clean data (i.e., the ground truth dietary intake). Note that we

focus on correcting the nutrient profile instead of the food profile (or the original dietary assessment) because the prevalence of zero values in the food profile (i.e., no consumption for many food items) renders a big challenge for machine learning tasks, while the derived nutrient profile usually has non-zero values. Also, we focus on the correction for random errors with zero means instead of systematic bias/errors with non-zero means because effectively correcting the latter requires the ground truth dietary intake, which is typically not available.

Our key idea is to incorporate gut microbial compositions as part of the input of a deep-learning model and to infer the true nutrient profile based on the assessed nutrient profile and the measured gut microbial composition (Fig. 1b). This idea is based on the knowledge that many dietary constituents reaching the large intestine fuel the growth of gut microbes[24–26]. One example is the growth of gut microbes enabled by anaerobic fermentation of indigestible polysaccharides[24,25]. As a result, the gut microbial composition is linked to dietary intake. Advances in sequencing technologies have made it possible to determine the gut microbial composition quickly and accurately. It has been shown that fecal bacteria and metabolites can be used as biomarkers for predicting the presence of food items ingested by healthy adults[27,28]. Hence, using objective microbial biomarkers to infer dietary intake might complement the self-reported dietary assessment and thus reduce measurement error in dietary intake.

In this work, we developed a deep-learning method: **M**icrobiom**e**-based nu**t**rient p**r**ofi**l**e **c**orrector (METRIC). We demonstrated the effectiveness of METRIC in removing randomly added noise to both synthetic and real data.

Results

Overview of METRIC

We aim to infer the true nutrient profile based on the assessed one and the measured gut microbial composition. A naive way to do this is to train a machine-learning model with the assessed nutrient profiles and microbial compositions as the input and the true nutrient profiles as the output. However, this is not feasible because such training requires the true nutrient profiles that are not easily available. To address this issue, we developed METRIC that does not rely on the true nutrient profile during its training but still can remove random errors in the assessed nutrient profile during testing (Fig. 2). During the training, we generated the corrupted nutrient profiles by adding random noise to the

assessed nutrient profiles, and then trained METRIC to remove the added noise by taking the corrupted nutrient profiles and the measured microbial compositions as its input and generating the assessed nutrient profiles as its output (Fig. 2a). We introduced the corrupted nutrient profiles to avoid METRIC copying the assessed nutrient profiles directly to the output, thus forcing METRIC to remove the added noise. Note that the training of METRIC did not use true nutrient profiles.

METRIC comprises stacked neural networks with three hidden layers (Fig. 2b). One unique feature is the skip connection that adds the corrupted nutrient profiles directly to the final output of neural networks, which previously has been shown to enhance the training and predictive performance for deep neural networks[29]. Similar to the existence of the skip connection in Noise2Noise[22], the introduction of the skip connection enables the neural networks to adjust the prediction based on the corrupted nutrient profile, ensuring that output variables would not deviate too much from the corrupted nutrient profile. More details about the architecture can be found in the Methods section. Overall, METRIC is a generic "denoiser" that learns to remove any random noise added to the nutrient profile. With this generic ability to remove random noise, the well-trained METRIC should be able to remove random measurement errors by generating predictions closer to the true nutrient profiles when it takes the assessed nutrient profiles and microbial compositions as its input (Fig. 2c).

We split each dataset into two non-overlapping parts: a training and a test set. METRIC was trained on the training set and then used to generate predictions for the test set. The correction performance is measured by comparing the predicted nutrient profiles with the "true" nutrient profiles in the test set (how we obtain the "true" nutrient profiles for each scenario will be explained in separate sections below). To measure the predictive performance of each nutrient, we adopted the Spearman Correlation Coefficient (SCC) ρ between its predicted and true corrected values.

METRIC can reduce measurement errors in nutrient profiles of synthetic data

We first validated METRIC using synthetic data for which we know the ground truth. We used the Microbial Consumer-Resource Model (MiCRM)[30] to generate three types of data: (1) *true* nutrient profiles (i.e., the ground-truth nutrient consumption), (2) *assessed* nutrient profiles (i.e., the true nutrient profiles with random noise added to mimic measurement errors), and (3) *corrupted* nutrient profiles (i.e., the assessed nutrient

profiles with artificially added random noise). MiCRM simulates the process of nutrient consumption by microbes and the following microbial growth[30]. Note that the random noise added to assessed and corrupted nutrient profiles are different. For simplicity, we only considered the nutrient consumption in MiCRM and did not model the nutrient production because most dietary nutrients cannot be produced by microbes. We created different samples by randomly sampling nutrient fluxes and then running the community assembly until we achieved a steady state. We consider sampled ground-truth nutrient fluxes as true nutrient profiles and the steady-state microbial abundances as microbial compositions. Gaussian noise $\mathrm{N}(0, \sigma^2)$ with the mean of zero and standard deviation σ is added to true nutrient profiles to create assessed nutrient profiles. More details about MiCRM, the generation of synthetic data, and added noise can be found in the Methods section.

We generated 250 samples by simulating the assembly process for 250 independent communities with 20 nutrients and 20 bacterial species. We trained METRIC on 200 samples and tested it on the remaining 50 samples. We measured how assessed values and corrected values (i.e., predicted values on the test set) respectively correlate with true values. For each nutrient, we measured ρ between its corrected values from predictions and its true concentration (denoted as ρ_c). Similarly, we calculated ρ between its assessed concentrations and its true concentration (denoted as ρ_a). As the standard deviation σ of the Gaussian noise increases, METRIC starts to correct the introduced noise, represented by the switch from negative values of $(\rho_c - \rho_a)$ to positive values in Fig. 3a. It is natural to expect that the correction of nutrient profiles with weaker noises does not work ($\rho_c < \rho_a$ for small σ in Fig. 3a), because the correction is not necessary when the measurement error is small (e.g., when $\rho_a > 0.8$). We focus on the case of $\sigma = 1.5$ from now on. The difference between the two metrics (i.e., $\rho_c - \rho_a$) reflects the correction performance. The nutrient in Fig. 3d-e is slightly corrected ($\rho_c - \rho_a = 0.02$), while the nutrient in Fig. 3f-g is strongly corrected ($\rho_c - \rho_a = 0.10$). Most nutrients have a better alignment between their corrected values and true values (Fig. 3e).

METRIC mitigates measurement errors added to nutrient profiles in real data

Next, we tested METRIC on three real-world datasets. The first dataset MCTS (MiCrobiome dieT Study) comes from a unique study that investigated the influence of diets on gut microbial composition[31]. It is unique because a large number of samples

(n=210) of paired nutrient profiles and microbial compositions were collected. The nutrient profiles were calculated from ASA24 (see Methods). Since the true nutrient profiles are not available, we treated the nutrient profiles derived from ASA24 as the "true" nutrient profiles and added random noise (Gaussian noise with the mean of zero and standard deviations σ) to them as "assessed" nutrient profiles. As σ increases, METRIC starts to better correct the introduced noise (Fig. 4a). We focus on the case of $\sigma = 1.0$ from now on. We found that carotene has a large ρ_{a} and is barely improved by METRIC ($\rho_{\mathrm{a}} = 0.91$ versus $\rho_{\mathrm{c}} = 0.91$ in Figs. 4b-c). By contrast, fiber has a small ρ_{a} and is strongly improved ($\rho_{\mathrm{a}} = 0.36$ versus $\rho_{\mathrm{c}} = 0.63$ for Figs. 4f-g). We believe that the large correction in the total fiber intake was due to most fibers being digested by gut microbes[24,25]. Overall, nutrients with smaller ρ_{a} have better correction performance, and nutrients with large ρ_{a} rarely improved (Fig. 4e). The mean correction performance averaged over all nutrients is $(\overline{\rho_c} - \overline{\rho_a}) = 0.069$.

We also tried to run METRIC without using the microbial composition, finding that the correction performance ($\overline{\rho_c} - \overline{\rho_a} = 0.060$) is worse than that with included microbial composition (Supplementary Fig. 1). Specifically, ρ_{c} of fiber decreases from 0.63 (Fig. 4e) to 0.52 (Supplementary Fig. 1f). We also computed the sensitivity which is defined as the ratio between the reduction in ρ of nutrient α and the perturbation amount of species i (Supplementary Fig. 2), which could detect some possible interactions between nutrients and species. For example, the sensitivity of unsaturated fatty acids (MUFAs) towards *Bacteroides uniformis* is large (~0.025, larger than 99.98% of the inferred sensitivity values), and is supported by the previously observed reduction of MUFAs by *Bacteroides uniformis*[32].

Then we applied METRIC to the second dataset MLVS (Men's Lifestyle Validation Study)[33,34]. Specifically, we utilized the composition of gut microbiomes and the one-day dietary assessment of the 7-day dietary records (7DDRs). The 7DDRs are widely recognized to be the most reliable estimation of dietary intake because participants are required to measure and report gram weights for foods before and after eating to compute actual intake[35]. To guarantee the usefulness of gut microbial composition in correcting dietary assessment, we required identification of paired microbial compositions and dietary assessments with matching dates. In MLVS, a total of 599 paired samples with matching dates were found. Similarly, as we did for the MCTS dataset, we regarded the nutrient profile derived from the 7DDRs as the "true" nutrient profile and added varying levels of Gaussian noise to it as the "assessed" nutrient profile. Then we trained METRIC

on 80% of the data and tested it on the remaining 20%. Consistent with our previous findings, the trained METRIC exhibits an ability to correct the nutrient profile (Fig. 5a), especially for large σ. For the case of $\sigma = 1.0$, the mean correction performance $(\overline{\rho_c} - \overline{\rho_a}) = 0.078$ (Fig. 5e). Across all nutrients, dietary fiber was the strongest corrected nutrient (Figs. 5f-g).

Finally, we applied METRIC to the third dataset WE-MACNUTR (Westlake N-of-1 Trials for Macronutrient Intake)[36]. WE-MACNUTR is a dietary intervention study that implemented a 'complete feeding' strategy, providing three isocaloric meals per day to 28 participants over a span of 72 days. Each participant completed high-fat, low-carbohydrate (HF-LC) and low-fat, high-carbohydrate (LF-HC) diets in a randomized sequence, with a 6-day wash-out period between them. Since the diets were completely controlled and well-documented, the nutrient profile derived from this dataset closely reflects the true nutrient profile. In WE-MACNUTR, we found 317 paired samples with both microbial compositions and nutrient profiles. Considering the nutrient profile from the complete feeding as the true nutrient profile, we introduced varying levels of noise (Gaussian noise $\mathrm{N}(0, \sigma^2)$ with different standard deviations σ) to create the assessed nutrient profile. Like our earlier results, METRIC can correct the nutrient profile when σ is large (Fig. 6a). When $\sigma = 1.0$, the mean correction performance $(\overline{\rho_c} - \overline{\rho_a}) = 0.125$ (Fig. 6e) and dietary fiber again exhibits a substantial correction (Figs. 6f-g).

Discussion

We presented a deep-learning method, METRIC, to correct random simulated measurement errors in the nutrient profile. The method relies on the assessed nutrient profile and microbial composition to learn how to infer the true nutrient profile. First, we validated its performance on synthetic data generated by MiCRM where we directly modeled true and assessed nutrient profiles. Then we applied METRIC to three distinct real clinical datasets with added noise to the nutrient profile and found that it can correct the nutrient profile well, especially for nutrients with large errors or metabolized by gut microbes. Similar to the class of computational methods that denoise images even without clean targets[22,23], METRIC offers a significant advantage by being capable of correcting the random measurement error without using the true nutrient profiles during the training. This attribute makes our method particularly useful in real-life scenarios where only assessed nutrient profiles are available, without access to true nutrient profiles.

We admit that METRIC has several limitations. First, without using clean targets (i.e., the ground truth dietary intake), METRIC is only capable of removing random measurement errors that have zero means. It cannot remove noise with a non-zero mean, just like Noise2Noise[22]. Thus, METRIC cannot correct the systematic bias/error (with a non-zero mean) in nutrient profiles. Effectively correcting the systematic bias requires both assessed and true nutrient profiles to discern the consistent deviation between them. If both types of nutrient profiles are available, it becomes feasible to design a regression model that predicts true nutrient profiles based on assessed nutrient profiles, thereby achieving the task of removing the systematic bias. Second, METRIC uses microbial compositions of fecal samples to correct nutrient profiles derived from ASA24 or 7DDR, which were collected exactly one day prior to fecal sample collection. We do not expect that microbial compositions of fecal samples collected from a particular time point could be used to correct nutrient profiles derived from FFQ, where subjects report how often each food item was consumed over a specified period of time, typically the past month or year. To clearly demonstrate this point, we leveraged the FFQ, 7DDR, and gut microbiome data from MLVS to compare the performance of using gut microbial compositions to predict the 7DDR- or FFQ-based nutrient profiles. We found that the predictability of FFQ-based nutrient profiles is much worse than that of 7DDR-based nutrient profiles (Supplementary Fig. 3). Consequently, we believe that METRIC cannot be used to correct random errors in FFQ-based nutrient profiles.

There are other methods to improve the accuracy of dietary intake measurement over traditional self-reported dietary assessments. For instance, digital documentation of meals through taking photos can be used to improve the accuracy of dietary assessment, though the validity of such technologies is yet to be established. In addition, nutritional biomarkers such as DNA barcodes for plants[37] and metabolite biomarkers[28,38,39] have been utilized to improve the assessment of food intake. A more accurate reflection of nutrient consumption can be obtained by using objective measurements of microbial composition or metabolomic profile, which can complement self-reported dietary assessment tools. Currently, an active research direction in the field of precision nutrition is to identify microbial and metabolite biomarkers for dietary intake. Although several studies attempted to predict the presence of food items based on fecal bacteria and metabolites[27,28], the analyses were limited to several food items and no connection to nutrient profiles was examined. Further research is warranted to examine the utility of METRIC to correct actual measurement errors in self-reported dietary assessments.

Methods

Datasets. The first dataset we used comes from a study that investigated the association between diet and the gut microbiome[31]. The study has paired 24-h food records and fecal shotgun metagenomes from 34 healthy human subjects collected daily over 17 days, with 210 paired samples in total. The second dataset is from MLVS[33,34] with 599 paired gut microbial compositions and 7DDRs. The third dataset is from WE-MACNUTR[36], a dietary intervention study that provides three meals per day for all participants. The study involved 30 participants, with 2 participants withdrawing from the trial early and not included in the final analysis. It has 317 paired gut microbial compositions and true nutrient profiles of complete feeding with matching dates. For all machine learning tasks, the same five random 80/20 train-test splits were utilized to guarantee a fair comparison of methods.

The generation of nutrient profiles based on dietary assessments. The nutrient profiles generated from dietary assessments involve several steps and rely on comprehensive food composition databases such as USDA's Food and Nutrient Database for Dietary Studies (FNDDS)[40] or Harvard Food Composition Database (HFDB)[41], etc. Each food item consumed is matched with a corresponding item in a food composition database. These databases provide detailed information on the nutrient content of foods, including macronutrients (like fats, proteins, carbohydrates) and micronutrients (like vitamins and minerals). By multiplying the amount of each food consumed by its nutrient content, the total intake of each nutrient is calculated.

The generation of synthetic data. Similar to the traditional consumer-resource model[42,43] and MiCRM (Microbial Consumer-Resource Model)[30], we simulated the process of nutrient consumption by microbes and the following microbial growth. For simplicity, we assumed a global pool with 20 microbial species and 20 nutrients in total. To mimic the difference in nutrient consumption profiles across individuals, we randomly assigned flux for each sample: to create a unique sample, the supply rate of each nutrient is randomly drawn from a uniform distribution $\mathcal{U}[0,1]$. We recorded the assigned supply rates of all nutrients as the true nutrient profile. The Gaussian noise with a mean of 0 and standard deviation of 0.5 is added to true nutrient profile to create assessed nutrient profiles with measurement errors. After the assignment of nutrient fluxes, we simulated the community assembly dynamics until the system reaches a steady state. The steady-state microbial relative abundances are recorded as the microbial composition for the

sample. Then we repeated this procedure of generating samples 250 times, forming the synthetic dataset we used in this study. The 250 samples are split in the 80/20 ratio as the training/test set.

The overall population dynamics for the concentration of nutrient N_α and abundance of microbial species M_i can be written as follows:

$$\frac{dN_\alpha}{dt} = -\sum_i a_{i\alpha} M_i N_\alpha - \delta N_\alpha,$$

$$\frac{dM_i}{dt} = \frac{\sum_\gamma a_{i\gamma} M_i N_\gamma}{Y} - \delta M_i,$$

where $a_{i\alpha}$ is the consumption rate of nutrient α by the species i, δ is the dilution rate, and Y is the yield. For simplicity, we assumed $\delta = 0.1$ and $Y = 1$. The $a_{i\alpha}$ is assumed to be non-zero with a probability of 50%. If $a_{i\alpha}$ is non-zero, its value is drawn from the uniform distribution $\mathcal{U}[0, 10]$. Eventually, each $a_{i\alpha}$ is divided by the number of nutrients that can be consumed by the species i, avoiding the outgrowth of generalists.

METRIC. The core of METRIC is the MLP (Multilayer Perceptrons).

- Data processing: The CLR (Centered Log-Ratio) transformation is applied to microbial relative abundances, and the log transformation is applied to the nutrient profiles.
- Model detail: METRIC has 3 hidden layers in the middle, sandwiched by input and output variables. Each hidden layer has a fixed hidden layer dimension of 256. A skip connection from corrupted nutrient profiles is introduced to add to the final layer of MLP. Specifically, the final prediction is the sum of (1) the corrupted input from the skip connection multiplied by a weight parameter α and (2) the final layer of MLP multiplied by $(1 - \alpha)$. The optimal value for α is chosen based on the five-fold cross-validation results on the training set.
- Training method: The Adam optimizer[44] is used for the gradient descent. The training stops when the mean SCC (Spearman Correlation Coefficient) of nutrients $\bar{\rho}$ on the validation/test set starts to decrease within the past 10 epochs.
- Activation function: ReLU (Rectified Linear Unit).

Statistics. To calculate correlation throughout the study, we used Spearman's correlation coefficient. Wherever we used *P* values, we explained in the Methods how we calculated them, since for all such measurements in the study, we calculated the associated null distributions from scratch. All statistical tests were performed using

standard numerical and scientific computing libraries in the Python programming language (version 3.7.1) and Jupyter Notebook (version 6.1).

Data and code availability. All code for simulations used in this manuscript can be found at https://github.com/wt1005203/METRIC.

Acknowledgements. We acknowledge grants from the National Institutes of Health (R01AI141529, R01HD093761, R35CA253185, RF1AG067744, UH3OD023268, U19AI095219, U01HL089856, U01-152905, and U01-167552) and Cancer Grand Challenges Team PROSPECT. We thank Walter Willett, Eric Rimm, and Lorelei Mucci for valuable discussions.

Author contributions. T.W. and Y.-Y.L. designed the project. T.W. performed all the numerical calculations and data analysis. T.W. processed the real data with assistance from Y.Fu. All authors analyzed the results. T.W. and Y.-Y.L. wrote the paper. All authors edited and approved the paper.

Correspondence. Correspondence and requests for materials should be addressed to Y.-Y. L. (email: yyl@channing.harvard.edu).

Competing Interests. The authors declare no competing interests.

References

1. Afshin, A. *et al.* Health effects of dietary risks in 195 countries, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet* **393**, 1958–1972 (2019).
2. Mozaffarian, D. *et al.* Global Sodium Consumption and Death from Cardiovascular Causes. *New England Journal of Medicine* **371**, 624–634 (2014).
3. Hu, F. B. Diet strategies for promoting healthy aging and longevity: An epidemiological perspective. *Journal of Internal Medicine* **n/a**,.
4. Donaldson, M. S. Nutrition and cancer: A review of the evidence for an anti-cancer diet. *Nutrition Journal* **3**, 19 (2004).
5. Hu, F. B. & Willett, W. C. Optimal Diets for Prevention of Coronary Heart Disease. *JAMA* **288**, 2569–2578 (2002).
6. Mensink, R. P., Zock, P. L., Kester, A. D. & Katan, M. B. Effects of dietary fatty acids and carbohydrates on the ratio of serum total to HDL cholesterol and on serum lipids and apolipoproteins: a meta-analysis of 60 controlled trials. *The American Journal of Clinical Nutrition* **77**, 1146–1155 (2003).
7. DiNicolantonio, J. J., Lucan, S. C. & O'Keefe, J. H. The Evidence for Saturated Fat and for Sugar Related to Coronary Heart Disease. *Progress in Cardiovascular Diseases* **58**, 464–472 (2016).
8. Malik, V. S. & Hu, F. B. The role of sugar-sweetened beverages in the global epidemics of obesity and chronic diseases. *Nat Rev Endocrinol* **18**, 205–218 (2022).
9. McNutt, S., Zimmerman, T. P. & Hull, S. G. Development of food composition databases for food frequency questionnaires (FFQ). *Journal of Food Composition and Analysis* **21**, S20–S26 (2008).
10. Sharpe, I. *et al.* Automated Self-Administered 24-H Dietary Assessment Tool (ASA24) recalls for parent proxy-reporting of children's intake (> 4 years of age): a feasibility study. *Pilot and Feasibility Studies* **7**, 123 (2021).
11. Hebert, J. R. *et al.* Development and testing of a seven-day dietary recall. *Journal of Clinical Epidemiology* **50**, 925–937 (1997).
12. Westerterp, K. R. & Goris, A. H. C. Validity of the assessment of dietary intake: problems of misreporting. *Current Opinion in Clinical Nutrition & Metabolic Care* **5**, 489 (2002).
13. Poslusna, K., Ruprich, J., Vries, J. H. M. de, Jakubikova, M. & Veer, P. van't. Misreporting of energy and micronutrient intake estimated by food records and 24 hour recalls, control and adjustment methods in practice. *British Journal of Nutrition* **101**, S73–S85 (2009).
14. Dao, M. C. *et al.* Dietary assessment toolkits: an overview. *Public Health Nutrition* **22**, 404–418 (2019).
15. Ravelli, M. N. & Schoeller, D. A. Traditional Self-Reported Dietary Instruments Are Prone to Inaccuracies and New Approaches Are Needed. *Frontiers in Nutrition* **7**, (2020).
16. Willett, W. Correction for the Effects of Measurement Error. in *Nutritional Epidemiology* (ed. Willett, W.) 0 (Oxford University Press, 2012). doi:10.1093/acprof:oso/9780199754038.003.0012.
17. Rosner, B., Willett, W. C. & Spiegelman, D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine* **8**, 1051–1069 (1989).
18. Spiegelman, D., McDermott, A. & Rosner, B. Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *The American Journal of Clinical Nutrition* **65**, 1179S-1186S (1997).

19. Hu, F. B. *et al.* Dietary Fat and Coronary Heart Disease: A Comparison of Approaches for Adjusting for Total Energy Intake and Modeling Repeated Dietary Measurements. *American Journal of Epidemiology* **149**, 531–540 (1999).
20. Schmidt, U. & Roth, S. Shrinkage Fields for Effective Image Restoration. in 2774–2781 (2014).
21. Chen, Y. & Pock, T. Trainable Nonlinear Reaction Diffusion: A Flexible Framework for Fast and Effective Image Restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**, 1256–1272 (2017).
22. Lehtinen, J. *et al.* Noise2Noise: Learning Image Restoration without Clean Data. Preprint at https://doi.org/10.48550/arXiv.1803.04189 (2018).
23. Krull, A., Buchholz, T.-O. & Jug, F. Noise2Void - Learning Denoising From Single Noisy Images. in 2129–2137 (2019).
24. Donia, M. S. & Fischbach, M. A. Small molecules from the human microbiota. *Science* **349**, 1254766 (2015).
25. Koh, A., De Vadder, F., Kovatcheva-Datchary, P. & Bäckhed, F. From Dietary Fiber to Host Physiology: Short-Chain Fatty Acids as Key Bacterial Metabolites. *Cell* **165**, 1332–1345 (2016).
26. Santhiravel, S. *et al.* The Impact of Plant Phytochemicals on the Gut Microbiota of Humans for a Balanced Life. *International Journal of Molecular Sciences* **23**, 8124 (2022).
27. Shinn, L. M. *et al.* Fecal Bacteria as Biomarkers for Predicting Food Intake in Healthy Adults. *The Journal of Nutrition* **151**, 423–433 (2021).
28. Shinn, L. M. *et al.* Fecal Metabolites as Biomarkers for Predicting Food Intake by Healthy Adults. *The Journal of Nutrition* nxac195 (2022) doi:10.1093/jn/nxac195.
29. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, 2016). doi:10.1109/CVPR.2016.90.
30. Marsland, R. *et al.* Available energy fluxes drive a transition in the diversity, stability, and functional structure of microbial communities. *PLoS Comput Biol* **15**, e1006793 (2019).
31. Johnson, A. J. *et al.* Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans. *Cell Host & Microbe* **25**, 789-802.e5 (2019).
32. López-Almela, I. *et al.* Bacteroides uniformis combined with fiber amplifies metabolic and immune benefits in obese mice. *Gut Microbes* **13**, 1–20 (2021).
33. Mehta, R. S. *et al.* Stability of the human faecal microbiome in a cohort of adult men. *Nat Microbiol* **3**, 347–355 (2018).
34. Li, J. *et al.* Interplay between diet and gut microbiome, and circulating concentrations of trimethylamine N-oxide: findings from a longitudinal cohort of US men. *Gut* **71**, 724–733 (2022).
35. Yue, Y. *et al.* Reproducibility and validity of diet quality scores derived from food-frequency questionnaires. *The American Journal of Clinical Nutrition* **115**, 843–853 (2022).
36. Ma, Y. *et al.* Individual Postprandial Glycemic Responses to Diet in n-of-1 Trials: Westlake N-of-1 Trials for Macronutrient Intake (WE-MACNUTR). *The Journal of Nutrition* **151**, 3158–3167 (2021).
37. Reese, A. T. *et al.* Using DNA Metabarcoding To Evaluate the Plant Component of Human Diets: a Proof of Concept. *mSystems* **4**, e00458-19 (2019).
38. Guasch-Ferré, M., Bhupathiraju, S. N. & Hu, F. B. Use of Metabolomics in Improving Assessment of Dietary Intake. *Clinical Chemistry* **64**, 82–98 (2018).
39. Brennan, L., Hu, F. B. & Sun, Q. Metabolomics Meets Nutritional Epidemiology: Harnessing the Potential in Metabolomics Data. *Metabolites* **11**, 709 (2021).

40. FNDDS : USDA ARS. https://www.ars.usda.gov/northeast-area/beltsville-md-bhnrc/beltsville-human-nutrition-research-center/food-surveys-research-group/docs/fndds/#.
41. Colditz, G. A., Manson, J. E. & Hankinson, S. E. The Nurses' Health Study: 20-Year Contribution to the Understanding of Health Among Women. *Journal of Women's Health* **6**, 49–62 (1997).
42. MacArthur, R. Species packing and competitive equilibrium for many species. *Theoretical Population Biology* **1**, 1–11 (1970).
43. Chesson, P. MacArthur's consumer-resource model. *Theoretical Population Biology* **37**, 26–38 (1990).
44. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. Preprint at https://doi.org/10.48550/arXiv.1412.6980 (2017).

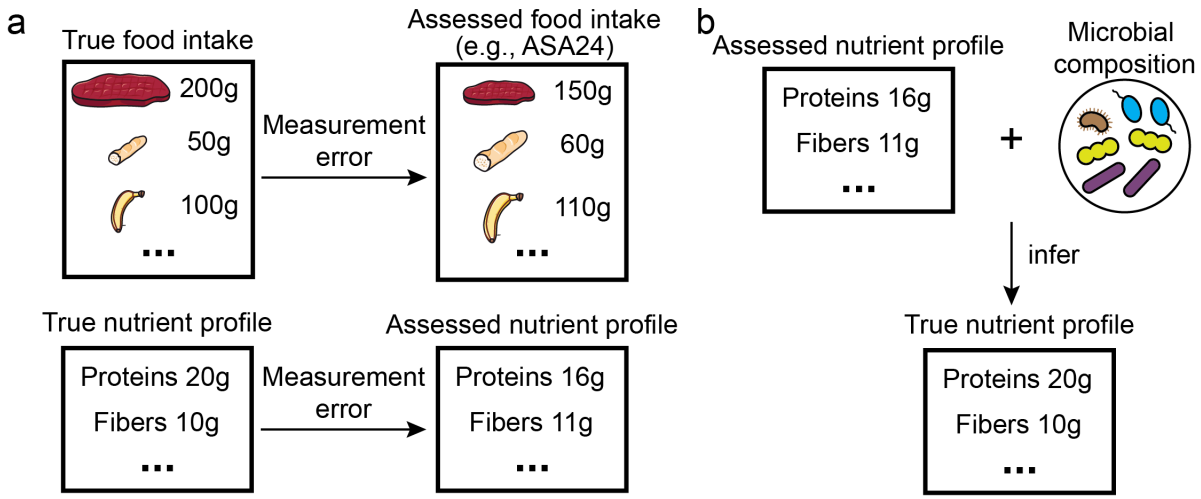

Figure 1: **The idea of inferring the true nutrient profile based on the assessed nutrient profile and microbial composition. a,** The typical dietary assessment such as ASA24 (Automated Self-Administered 24-hour Dietary Assessment Tool) often induces a measurement error in the food intake due to unreliable mental recall. Such a random measurement error in the food intake is carried to the nutrient profile when the food profile is converted to the nutrient profile. **b,** Our goal in this study is to remove the measurement error. More specifically, we would like to infer true nutrient profiles based on assessed nutrient profiles and microbial compositions.

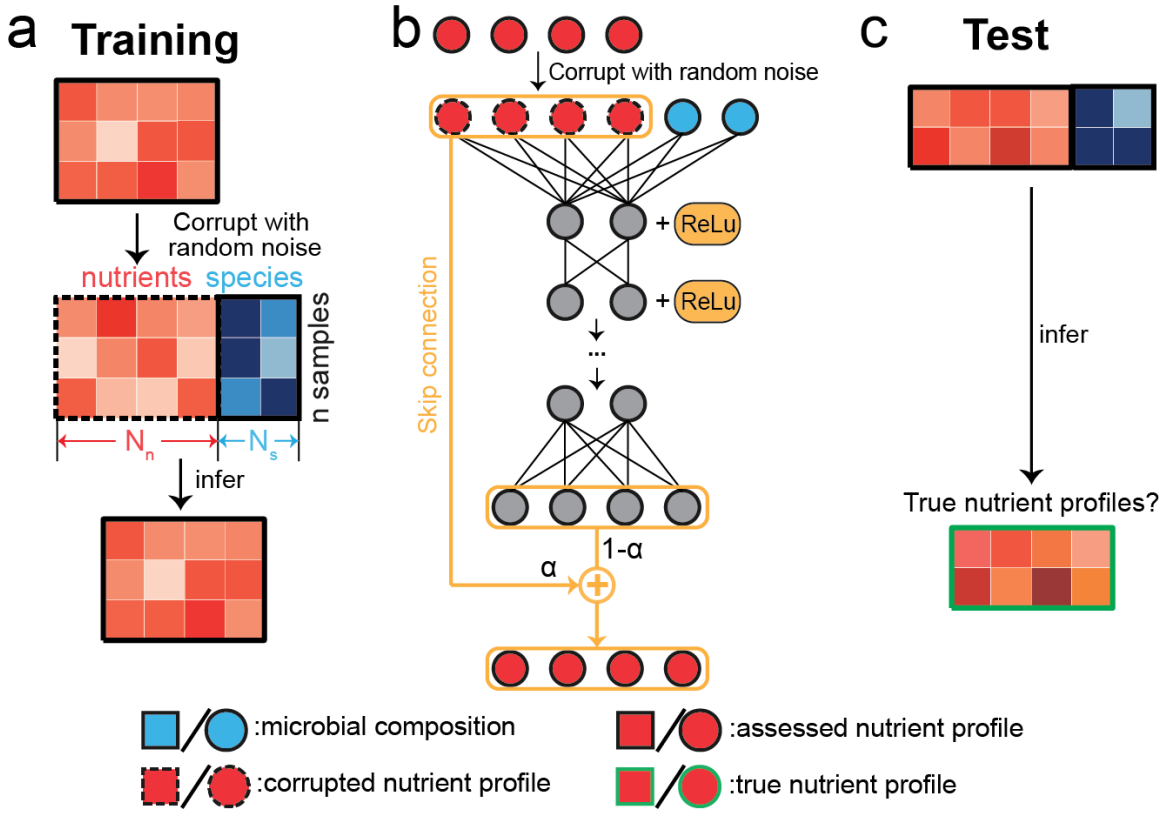

Figure 2: **The architecture and workflow of METRIC to infer true nutrient profiles.** For simplicity, we used a hypothetical example with n=3 training samples and 2 samples in the test set. For each sample, there are N_s microbial species and N_n nutrients. Across panels, microbial species and their relative abundances are colored blue. Nutrients and their amounts are colored red. The corrupted nutrient profiles are created by adding different types of random noise (i.e., Gaussian, Uniform, etc.) to the assessed nutrient profiles. Icons associated with assessed/corrupted nutrient profiles are bounded by solid black/dashed lines. Icons associated with true nutrient profiles are bounded by solid green lines. **a,** During the training of METRIC, the method takes corrupted nutrient profiles and microbial compositions as the input and learns to infer assessed nutrient profiles. **b,** Similar to MLP (MultiLayer Perceptrons), METRIC has several hidden layers in the middle. The skip connection provides the corrupted nutrient profile directly to one layer before the final output, enabling it to skip the propagation through the hidden layers. The skipped corrupted nutrient profile multiplied by the weight parameter α and the final hidden layer (the bottom grey nodes) multiplied by $(1-\alpha)$ add up as the final output (the bottom red nodes). **c,** The well-trained METRIC is applied to the test set to generate predictions for nutrient profiles whose values are compared to true nutrient profiles.

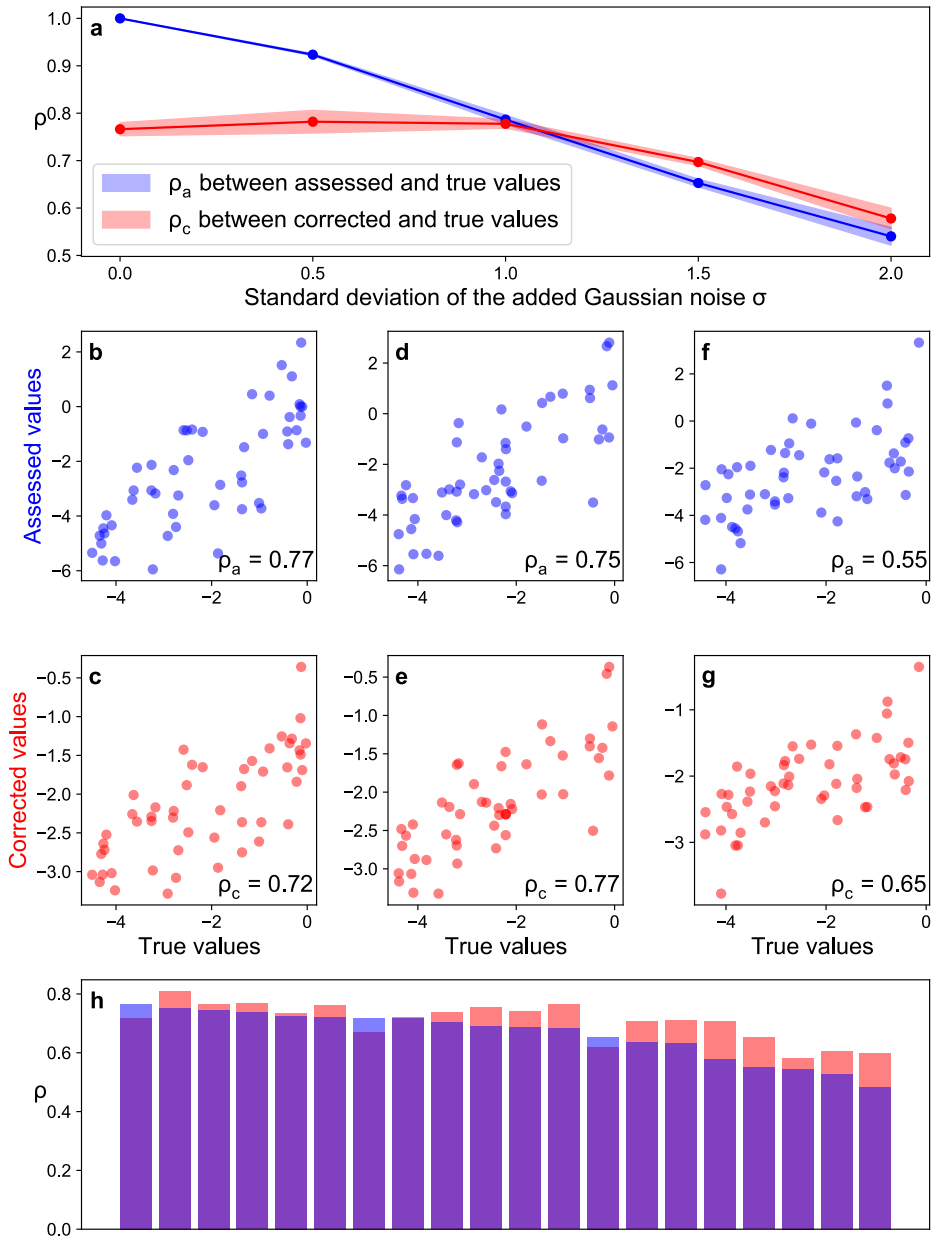

Figure 3: **METRIC can correct the measurement error in assessed nutrient profiles on synthetic data from MiCRM[30].** The Spearman's Rank Correlation Coefficient ρ is adopted to evaluate the correlation across various types of nutrient profiles. All corrected/true values shown are the log of nutrient concentrations. **a,** ρ_c (i.e., ρ between corrected and true values) and ρ_a (i.e., ρ between assessed and true values) decrease as the standard deviation of added Gaussian noise σ increases. All following panels focus on the case of σ=1.5. **b,** The correlation between assessed values and true values of log concentrations of one nutrient among different samples. **c,** The correlation between corrected values (predictions of METRIC) and true values of log concentrations of the same nutrient among different samples. Similar comparisons for another two nutrients are shown in **d-e** and **f-g**. **e,** The correction performance of all nutrients is measured by $(\rho_c - \rho_a)$.

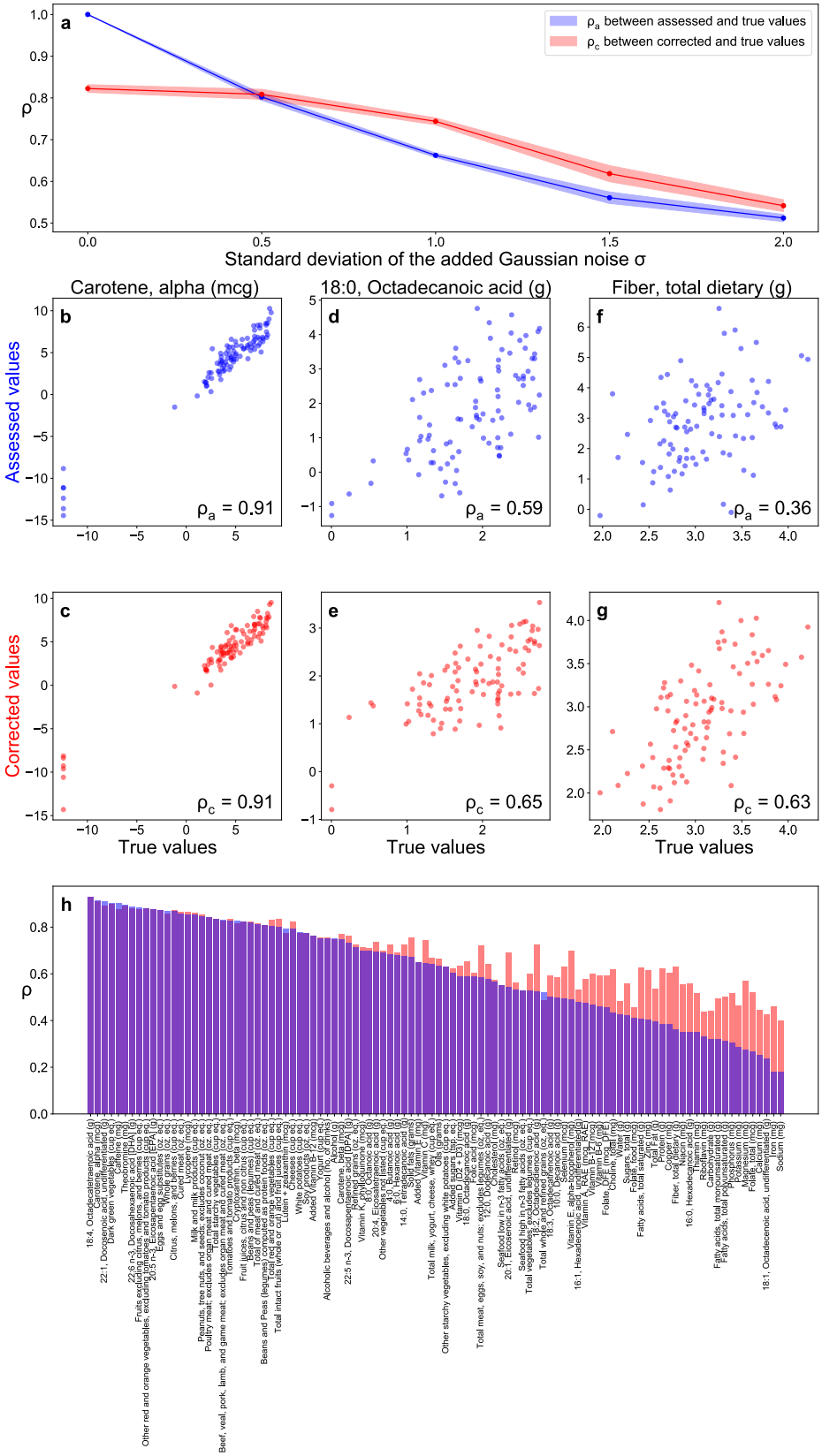
a
ρa between assessed and true values
ρc between corrected and true values
ρ
Standard deviation of the added Gaussian noise σ
Carotene, alpha (mcg)
18:0, Octadecanoic acid (g)
Fiber, total dietary (g)
Assessed values
Corrected values
ρa = 0.91
ρa = 0.59
ρa = 0.36
ρc = 0.91
ρc = 0.65
ρc = 0.63
True values
h

Figure 4: **METRIC can correct the measurement error in assessed nutrient profiles on real data from MCTS[31].** The Spearman's Rank Correlation Coefficient ρ is adopted to evaluate the correlation across various types of nutrient profiles. All nutrient concentrations are in the unit of grams. All corrected/true values shown are the log of nutrient concentrations. **a,** ρ_c (i.e., ρ between corrected and true values) and ρ_a (i.e., ρ between assessed and true values) decrease as the standard deviation of added Gaussian noise σ increases. All following panels focus on the case of σ=1.0. **b,** The correlation between assessed values and true values of log concentrations of carotene among different samples. **c,** The correlation between corrected values (predictions of METRIC) and true values of log concentrations of carotene among different samples. **d-e,** The similar comparison for octadecaonoic acid shows a modest correction. **f-g,** The similar comparison for fiber shows a strong correction. **e,** The correction performance for all nutrients is measured by $(\rho_c - \rho_a)$.

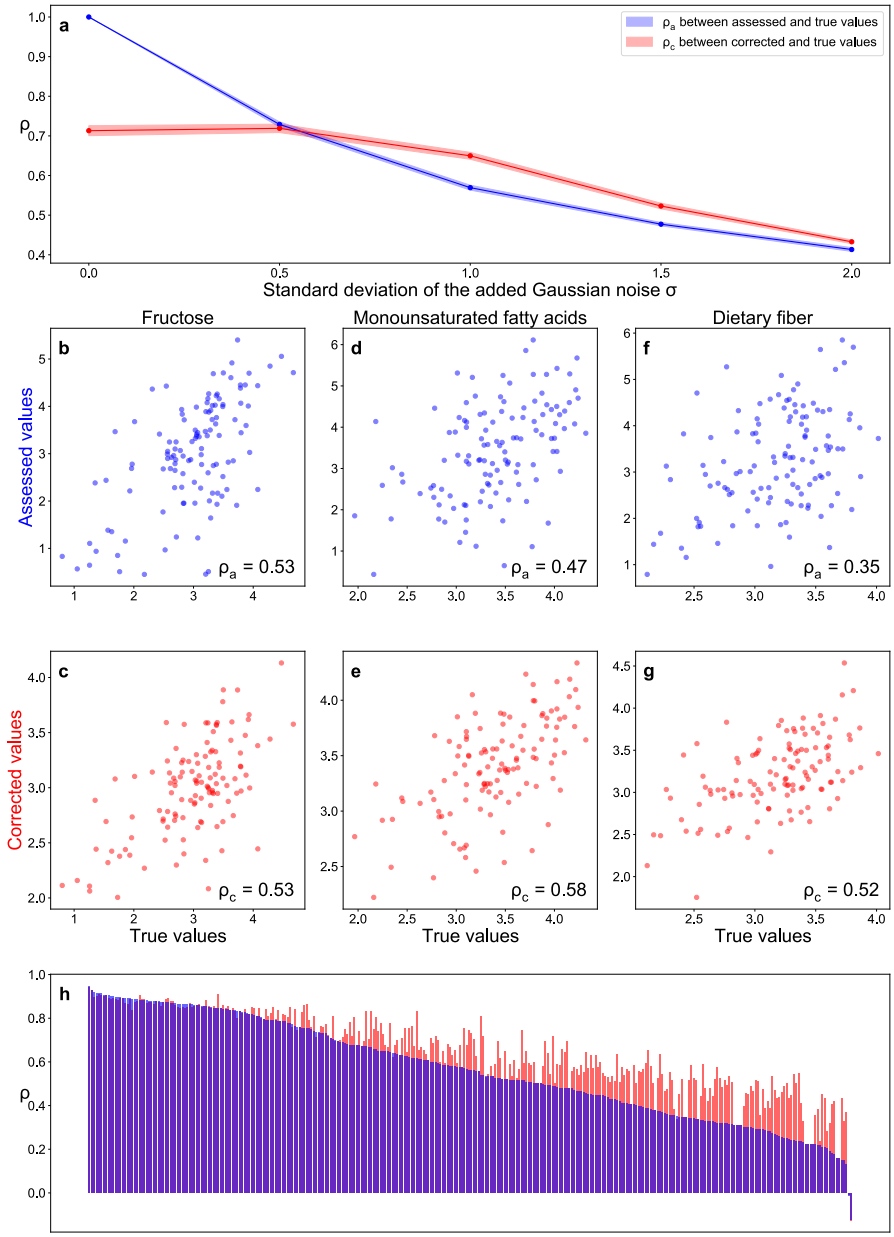

Figure 5: **METRIC can correct the measurement error in assessed nutrient profiles from MLVS[33,34].** The Spearman's Rank Correlation Coefficient ρ is adopted to evaluate the correlation across various types of nutrient profiles. All nutrient concentrations are in the unit of grams. All corrected/true values shown are the log of nutrient concentrations. **a,** ρ_c (i.e., ρ between corrected and true values) and ρ_a (i.e., ρ between assessed and true values) decrease as the standard deviation of added Gaussian noise σ increases. All following panels focus on the case of σ=1.0. **b,** The correlation between assessed values and true values of log concentrations of fructose among different samples. **c,** The correlation between corrected values (predictions of METRIC) and true values of log concentrations of fructose among different samples. **d-e,** The similar comparison for monounsaturated fatty acids shows a modest correction. **f-g,** The similar comparison for dietary fiber shows a strong correction. **e,** The correction performance for all nutrients is measured by $(\rho_c - \rho_a)$. Nutrient names are not added due to lack of space.

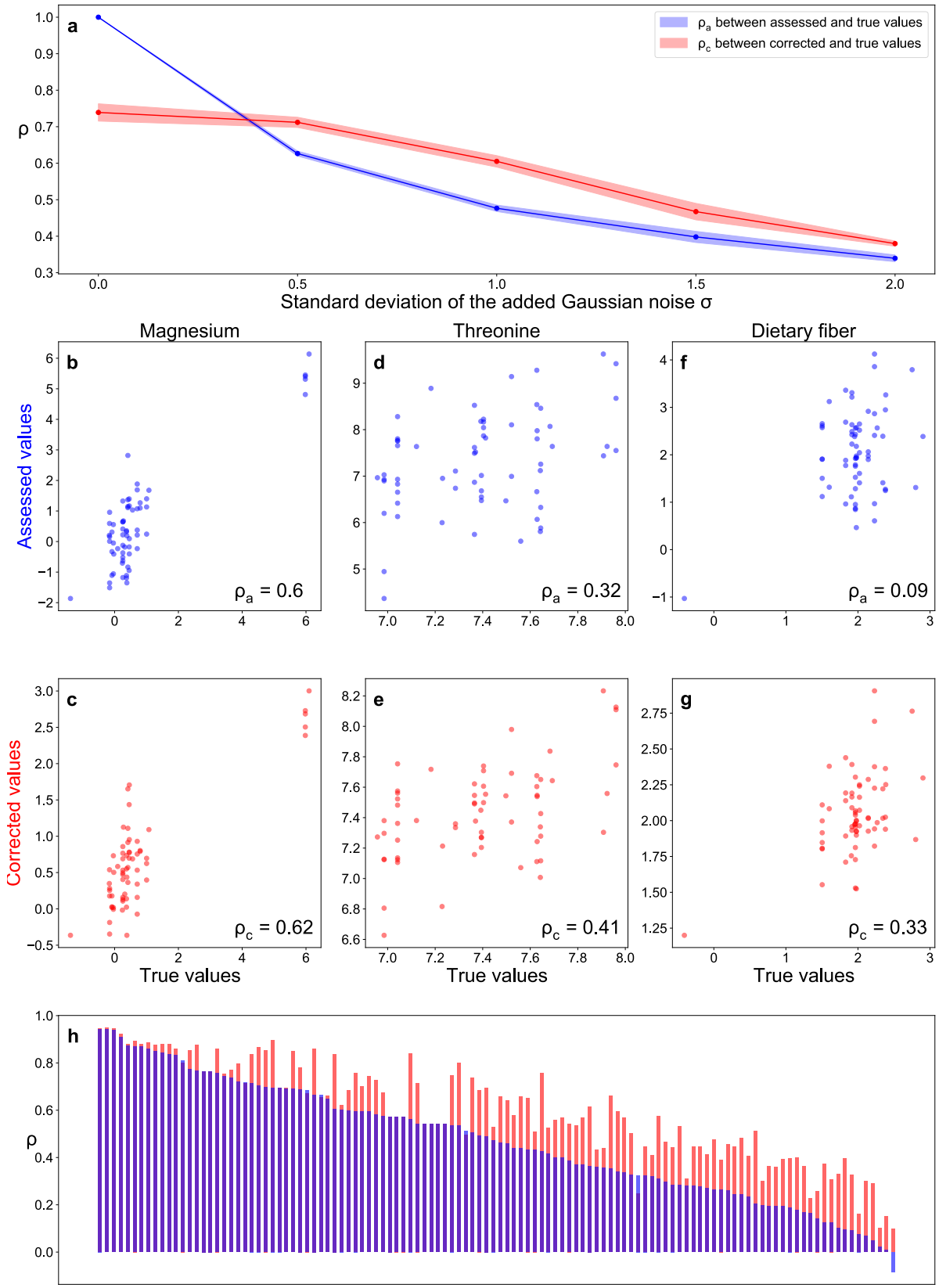
a
ρ
ρa between assessed and true values
ρc between corrected and true values
Standard deviation of the added Gaussian noise σ
Magnesium
Threonine
Dietary fiber
Assessed values
Corrected values
b
d
f
c
e
g
ρa = 0.6
ρa = 0.32
ρa = 0.09
ρc = 0.62
ρc = 0.41
ρc = 0.33
True values
True values
True values
h
ρ

Figure 6: **METRIC can correct the measurement error in assessed nutrient profiles from WE-MACNUTR[36].** The Spearman's Rank Correlation Coefficient ρ is adopted to evaluate the correlation across various types of nutrient profiles. **a,** ρ_c (i.e., ρ between corrected and true values) and ρ_a (i.e., ρ between assessed and true values) decrease as the standard deviation of added Gaussian noise σ increases. All nutrient concentrations are in the unit of grams. All corrected/true values shown are the log of nutrient concentrations. All following panels focus on the case of σ=1.0. **b,** The correlation between assessed values and true values of log concentrations of magnesium among different samples. **c,** The correlation between corrected values (predictions of METRIC) and true values of log concentrations of magnesium among different samples. **d-e,** The similar comparison for threonine shows a modest correction. **f-g,** The similar comparison for dietary fiber shows a strong correction. **e,** The correction performance for all nutrients is measured by ($\rho_c - \rho_a$). Nutrient names are not added due to lack of space.