Nutritional Redundancy in Human Diet

Xu-Wen Wang[1], Francine Grodstein[2], Shilpa N. Bhupathiraju[1,3], Qi Sun[1,3,4], Xuehong Zhang[1,3], Frank Hu[3,4], Scott T. Weiss[1], Yang-Yu Liu[1]

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

[2]*Rush Alzheimer's Disease Center, Department of Internal Medicine, Rush Medical College, Rush University, Chicago, IL 60612, USA*

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

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

Studying human dietary intake may help us identify effective measures to treat or prevent many chronic diseases whose natural histories are influenced by nutritional factors. Through examining four comprehensive cohorts with dietary intake data collected on different time scales, we find that the food intake profile varies tremendously across individuals and over time, while the nutritional intake profile is highly stable across individuals and over time. We refer to this phenomenon as 'nutritional redundancy'. We find that this phenomenon cannot be simply explained by the fact that different foods contain similar nutrients. Instead, it is largely due to more sophisticated features (e.g., the highly nested structure) of the food-nutrient network --- a bipartite graph that connects foods to their nutrient constituents. The food-nutrient network also enables us to quantify the level of nutritional redundancy for each diet record of any individual, i.e., the personal nutritional redundancy. Interestingly, this personal nutritional redundancy measure does not strongly correlate with any classical healthy diet scores, but its performance of predicting healthy aging in a cohort of older women is as strong as those healthy diet scores. The concept of nutritional redundancy developed here offers us a new perspective on studying human diet and presumably could be used for more phenotype predictions.

Introduction

Human dietary intake fundamentally affects our nutrition, energy supply, and health. A better understanding of diet patterns can help us identify measures to prevent or treat health conditions and diseases such as obesity[1,2], type 2 diabetes[3,4], and cardiovascular disease[5,6]. Indeed, randomized trials have established the benefits[7] of the Mediterranean diet on clinical cardiovascular disease[8] and the Dietary Approaches to Stop Hypertension (DASH)[9] diet on blood pressure control. To improve the diet quality, the 2020 Dietary Guidelines for Americans recommended that consumers should select nutrient-dense foods and beverages, which can provide vitamins, minerals and other health-promoting components and have no or little added sugars, saturated fat, and sodium[10]. For example, intake of whole grains, legumes and vegetables and fruit is recommended to reduce ingredients, such as unhealthy fats, and excess sugar and sodium, which have been associated with many diseases such as coronary heart disease[11–13] and obesity, liver and other metabolic diseases[14–16].

Food is a complex combination of components that can be classified into nutrients and non-nutrients[17]. The preliminary step to developing a guide to healthy dietary guidance is *nutrient profiling*[18,19] --- the science of understanding which nutrients are present in a given food item. This has many potential applications, ranging from dietary guidance to nutrient labeling and the regulation of health claims[18]. The current nutrient profiling relies on some existing databases, such as the USDA's Food and Nutrient Database for Dietary Studies (FNDDS)[20,21], Frida[22], FooDB[23] and many other databases[24–26] (see **SI Sec. 1.1** for details). Those databases represent the most comprehensive efforts so far to integrate food composition data from specialized databases and experimental data.

Despite the various databases on food constituents, a comprehensive understanding of the food-nutrient relationship remains challenging. Human dietary patterns across the world could be drastically different. For example, the Mediterranean diet is characterized by a high consumption of olive oil, legumes and vegetables and lower consumption of non-fish meat products while the Western diet pattern is characterized by high consumption of red meat, candy and sweets and low intakes of fruits, and vegetables. Although different diets have different food choices, the essential nutrient requirements are still be met. This implies that despite the variability in food choice and intake, nutrient profiles can be highly stable across individuals and over time. Indeed, by analyzing human dietary intake data collected from different cohorts and on different time scales, we found that the food profile varies tremendously across individuals and over time, while the nutritional profile is highly stable across subjects and over time. This phenomenon is strongly reminiscent of the phenomenon of *functional redundancy* observed in the human microbiome, i.e., taxonomic profile is highly personalized while the gene composition (or

71 functional profile) is highly conserved across individuals[27–29]. Hence, hereafter we will refer to
72 this concept as *nutritional redundancy* (NR).

73 To quantitatively understand diet patterns and the origin of NR in the human diet, we
74 constructed the food-nutrient network (FNN) --- a bipartite graph that connects foods to the
75 nutrients they contain. Leveraging tools from network science[30], we found that the FNN has
76 several topological features (e.g., highly nested structure) that largely contribute to the high NR
77 of the human diet. The FNN further enabled us to define a personal NR measure. We found that
78 this personal NR measure does not strongly correlate with any classical healthy diet scores[31–34].
79 Finally, we evaluated the performance of NR and those healthy diet scores in the prediction of
80 healthy aging status, finding that NR displays an equally strong predictability of healthy aging
81 as those healthy diet scores.

82 **Results**

83 **The phenomenon of population-level NR.** To demonstrate the phenomenon of NR in the human
84 diet, we analyzed four comprehensive datasets with dietary data collected on different time scales.
85 (1) Diet-microbiome association study (DMAS)[35]: a longitudinal study of 34 healthy individuals
86 with dietary intake data and stool sample collected daily over 17 consecutive days. Daily dietary
87 intake data was collected using the automated self-administered 24-h (ASA24) dietary assessment
88 tool[36–38]. DMAS focused on 41 dietary nutrients, and 9 food groups (e.g., grains, fruits, vegetable,
89 etc.) based on the Food and Nutrient Database for Dietary Studies (FNDDS) food coding scheme[20].
90 Note that there are two outliers in DMAS, referred to as "shake drinkers", whose reported diet
91 consisted primarily of a nutritional meal replacement beverage. Those two outliers were removed
92 in our analysis. Also, in our analysis, we focused on those participants (in total n=30) who have
93 ASA24 data available for all the 17 days. (2) Nurse Healthy Study (NHS)[39,40]: NHS began in 1976
94 when 121,700 female registered nurses, aged 30-55 years in the United States were enrolled; the
95 semiquantitative food frequency questionnaire (FFQ) was first administrated in 1980, and then
96 administrated approximately every four years. So far, we have up to eight time points of FFQ
97 data for NHS participants. In our analysis, we focused on those participants (in total n=35,256)
98 who have FFQ data available for all the eight time points. (3) Women Lifestyle Validation Study
99 (WLVS)[38]: As a sub-study of the NHS, WLVS was designed to investigate the measurement-error
100 structure associated with self-reported dietary and physical activity assessments within one year
101 (with up to four ASA24 records). Women with a history of coronary heart disease, stroke, cancer,
102 or major neurological disease were excluded. Among the 796 enrolled participants, 692
103 completed at least one ASA24. In our analysis, we focused on those WLVS participants with all
104 four ASA24 records available (in total n=216). (4) Men Lifestyle Validation Study (MLVS)[41,42]: a
105 one-year study nested in the Health Professionals Follow-up Study (HPFS)[43,44]. HPFS is a
106 prospective cohort of 51,529 men followed since 1986, when participants initially ranged from 40-

107 75 years. This all-male study was designed to complement the all-female NHS, and both were
108 conducted identically. Similarly, MLVS was designed to complement WLVS. In our analysis, we
109 focused on those MLVS participants with all of four ASA24 records available (in total n=451).

110 For those selected participants in each study, we first assessed the change in their food
111 and nutrient profiles (i.e., the relative abundances of food items and nutrients in their diet) over
112 time. We found that the food profiles were highly personalized[45] and dynamic for almost all
113 individuals at different time scales: daily (**Fig.1a**), monthly (**Fig.1b,c**), and yearly (**Fig.1d**); while
114 the nutrient profiles, as expected, were highly conserved across different individuals and over
115 the whole study time period for all four studies. The most abundant food items of individuals in
116 most studies were from the groups of sugar, sweetened beverages, and vegetables. This is
117 partially due to the fact that the moisture content in those foods is high. The most abundant
118 nutrients for all participants and studies are carbohydrate, fat, and protein. Within the
119 carbohydrate category, sugars and fiber are the driving factors.

120 To quantify the between-sample difference in food or nutrient profiles, we adopted the
121 notion of beta diversity from community ecology[46]. In particular, we used three different
122 measures (Bray-Curtis dissimilarity, root Jensen-Shannon divergence, and Yue-Clayton distance)
123 to quantify the beta diversity. As shown in **Fig.S1**, the beta diversity of nutritional profiles is
124 significantly lower than that of food profiles in terms of all the three measures of beta diversity.

125 **Definition of personal NR.** Consider a pool of N food items, which contains a collection of M
126 nutrients in total. The food profile $\boldsymbol{f}^{(\nu)} = \left[f_1^{(\nu)}, \cdots, f_N^{(\nu)}\right]$ of individual-ν's diet record can be
127 directly related to its nutrient profile $\boldsymbol{n}^{(\nu)} = \left[n_1^{(\nu)}, \cdots, n_M^{(\nu)}\right]$ through the FNN (**Fig.2**). Here, $f_i^{(\nu)}$ (or
128 $n_a^{(\nu)}$) represents the relative abundance of food-i (or nutrient-a) in the diet record of individual-ν.
129 We define the FNN as a weighted bipartite graph connecting these foods to their nutrients. The
130 FNN can be represented by an $N \times M$ incidence matrix $\boldsymbol{G} = (G_{ia})$, where a non-negative value
131 G_{ia} indicates the amount contributed by food-i to nutrient-a (see Fig.2a for the unit of each
132 nutrient). The nutrient profile is given by $\boldsymbol{n}^{(\nu)} = c\boldsymbol{f}^{(\nu)}\,\boldsymbol{G}$, where $c = \left[\sum_{a=1}^{M}\sum_{i=1}^{N} f_i^{(\nu)}\,G_{ia}\right]^{-1}$ is a
133 normalization constant.

134 A key advantage of the FNN is that it enables us to calculate the NR for each diet record
135 of any individual, i.e., the within-sample or personal NR. In the ecological literature[28,29,47], the
136 functional redundancy of a local community is interpreted as the part of its taxonomic diversity
137 that cannot be explained by its functional diversity. Similarly, we can define the NR of a dietary
138 assessment from a particular individual as the part of its food diversity (FD) that cannot be
139 explained by its nutritional diversity (ND), i.e., $\mathrm{NR} = \mathrm{FD} - \mathrm{ND}$. Here we chose FD to be the Gini-

Simpson index: $\mathrm{GSI} \equiv 1 - \sum_{i=1}^{N} f_i^2 = \sum_{i=1}^{N} \sum_{j \neq i}^{N} f_i f_j$, representing the probability that two randomly chosen items of a subject's food profile (with replacement) belong to two different foods; and ND is chosen to be the Rao's quadratic entropy $Q \equiv \sum_{i=1}^{N} \sum_{j \neq i}^{N} d_{ij} n_i n_j$, characterizing the mean nutritional distance between any two randomly chosen food items in the diet record[25,26]. Here $d_{ij} = d_{ji} \in [0, 1]$ denotes the nutritional distance between food-i and food-j, which can be calculated as the weighted Jaccard distance between the sets of nutrients contained in the two food items. By definition, $d_{ii} = 0$ for $i = 1, \cdots, N$. Note that with $\mathrm{FD} = \mathrm{GSI}$ and $\mathrm{ND} = Q$, we have $\mathrm{NR} = \sum_{i=1}^{N} \sum_{j \neq i}^{N} (1 - d_{ij}) f_i f_j$, naturally representing the nutritional similarity (or overlap) of two randomly chosen food items in any diet record.

The personal NR of each diet record is closely related to the phenomenon of population-level NR observed over a collection of diet records. Let's consider highly personalized food profiles from a population. There are two extreme cases: (i) Each food has its own unique nutrient content (**Fig.2b1**), hence $d_{ij} = 1$ for any $i \neq j$. In this case, for each individual we have $\mathrm{FD} = \mathrm{ND}$ and $\mathrm{NR} = 0$ (representing the lowest level of nutritional redundancy), and the nutrition profiles vary drastically across different individuals (**Fig.2c1**). (ii) All food share exactly the same nutrient contents (**Fig.2b3**), rendering $d_{ij} = 0$ for all i and j. In this case, for each individual we have $\mathrm{ND} = 0$ and $\mathrm{NR} = \mathrm{FD}$ (representing the highest level of nutritional redundancy), and the nutrition profiles are exactly the same for all individuals (**Fig.2c3**). These two extreme scenarios are of course unrealistic. In a more realistic intermediate scenario, the FNN has certain topological features such that different foods share a few common nutrients, but some foods are specialized to include some unique nutrients (**Fig.2b2**). In this case, the ND and NR of each individual's diet record can both be quite comparable, and the nutritional profiles can be highly conserved across individuals (**Fig.2c2**).

A reference FNN. To visualize the topological features of real FNNs, we constructed a reference FNN based on USDA's Food and Nutrient Database for Dietary Studies (FNDDS) 2011-2012[20], consisting of 7,618 foods and 65 nutrients and micronutrients. This reference FNN is depicted as a bipartite graph, where for visualization purposes each food node represents one of the nine highest-level food groups (based on the FNDDS food coding scheme) and each nutrient node represents nutrient (**Fig.3a**). Note that the FNN associated with the diet record of any individual can be considered as a particular subgraph of this reference FNN.

To characterize the structure of this reference FNN, we systematically analyzed its network properties using the complete nutrient profile of all 7,618 foods. We first visualized its incidence matrix (**Fig.3b**), where the presence (or absence) of a link connecting a food and a nutrient is colored in green (or white), respectively. We found that some foods (e.g., bacon cheeseburger, hot ham and cheese sandwich, corresponding to the leftmost columns in Fig.3b)

contribute to almost all of the nutrients, while some foods only include very few nutrients (e.g., sugar substitutes and smart water, corresponding to the rightmost columns in Fig.3b). Moreover, we noticed that the incidence matrix displays a highly nested structure, i.e., the nutrients of those food items (with fewer distinct nutrients) in the right columns tend to be subsets of nutrients for those food items (with more distinct nutrients) in the left columns. The nestedness of the FNN can be quantified by utilizing the classical NODF measure[48,49], and turns out to be much higher than expected by chance (see **SI Sec.2.2** for details). We then calculated the nutritional distances among different food items, finding a unimodal distribution with the peak centered around 0.25, indicating that most food items include very similar nutrient components (**Fig.3c)**. Finally, we calculated the degree distributions of nutrient nodes and food nodes, respectively. Here, the degree of a nutrient node in the FNN is just the number of distinct foods that contain this nutrient. Similarly, the degree of a food node in the FNN is the number of distinct nutrients it contains. We found that the degrees of food items follow a Poisson-like distribution (**Fig.3d**), implying that different foods generally contain a very similar number of nutrients. Yet, the degree distribution of nutrients is quite uniform except for a heavy tail (**Fig.3e**), indicating that most nutrients can be contained in any number of foods while a few essential nutrients (e.g., sodium, Zinc corresponding to the topmost rows in the incidence matrix shown in Fig.3b) are contained in almost all foods. FNN associated with each of the nine food groups and FNNs constructed by using other databases, e.g., Frida[22] and Harvard food composition database (HFDB)[40], revealed very similar network properties (see **Figs.S2, S3**).

We emphasize that the highly nested structure of the reference FNN cannot be simply explained by the presence of essential nutrients. First, as shown in Fig.3b, the incidence matrix of the FNN still displays a highly nested structure even in the absence of essential nutrients (the topmost green rows). Second, if we randomize the FNN but preserve the nutrient degree distribution, the randomized FNNs have much lower nestedness than that of the real FNN (**Fig.S4**). To further confirm this, we also calculated the nutrient distance of each food pair using the unweighted Jaccard index (see **SI Sec.2.3** for definition), suggesting that the nestedness will decrease and the nutrient distance between different foods is increased (see Fig.S4). Third, we adopted tools from statistical physics[50] to calculate the expected nestedness value and its standard deviation for an ensemble of randomized FNNs in which the expected food and nutrient degree distributions match those of the real FNN. We found that the expected nestedness of randomized FNNs is significantly lower than that of the real FNN (one sample z-test yields $p_{\text{value}} < 10^{-5}$, see **SI Sec.1.3.2** for details).

Personal NR calculation based on dietary intake data. We calculated the personal NR of diet records of those selected participants in the 4 studies: DMAS, WLVS, MLVS and NHS. First, we constructed a reference FNN based on the FNDDS to calculate the NR of DMSA, WLVS and

MLVS participants, and a reference FNN based on HFDB to calculate the NR of NHS participants (see **SI Sec.2** for details). Note that the diversity of food items for different participants in different studies can be very different. To remove the potential impact of food diversity FD on the calculated NR, we computed the normalized NR (i.e., nNR ≡ NR/FD). Interestingly, we found that in all the 4 studies (DMAS, WLVS, MLVS and NHS) nNR~0.45 (**Fig.4**, black boxes), suggesting that nutritional redundancy and nutritional diversity are generally comparable in human diet.

The nNR of the NHS at the population level displays a monotonic decrease, indicating that the diet patterns of those participants indeed have been adjusted. To better illustrate such a diet pattern change, we projected the bipartite food-nutrient network (constructed from HFDB) into the food space, resulting in a food similarity network (see **Fig.5a**). In this network, each node represents a food item and a link connecting food item-i and item-j represents the weighted Jaccard similarity s_{ij} of their nutrient constituents (see **SI Sec.2.3**). Here, for visualization purposes, only links with $s_{ij} \geq 0.85$ were retained. We found a clear modular structure in the food similarity network, i.e., food items from the same food group form a densely connected cluster or module (see **Fig.5a**). Then, we examined the individual food similarity network of a particular NHS participant with the largest nNR reduction from year 1984 to 2010 (see **Fig.5b,c**). We found the density of her food similarity network in 1984 is much higher than that in 2010, suggesting that this participant consumed foods with more overlapping nutrient constituents in 1984 (Fig.5b). Moreover, we found the most abundant food items in 2010 were coffee, water and cookies, which do not connect with each other, indicating that she chose foods with more distinct nutrient constituents (Fig.5c).

Impact of FNN structure on personal NR. To identify key topological features of the FNN that determine the nNR, we adopted tools from network science. In particular, we randomized the FNN using three randomization schemes, yielding three null models. Null-FNN-1: Complete randomization. We keep the number of foods (N) and number of nutrients (M) unchanged, but otherwise completely rewire the links between foods and nutrients. Null-FNN-2: Food-degree preserving randomization. We keep N, M, and the degree of each food node unchanged, but selects randomly the nutrients that link to each food. Null-FNN-3: Nutrient-degree preserving randomization. Here we keep N, M, and the degree of each nutrient unchanged, but select randomly the foods that link to each nutrient. Null-FNN-4: Nutrient-degree and food-degree preserving randomization. Here we keep N, M, and the degree of each nutrient and the degree of each food unchanged, but randomly rewire the links between food nodes and nutrient nodes. Then we recalculated the nNR for each diet record (**Fig.4**, colored boxes). We found that for all the cohorts all the four null models yield much lower nNR than that of the real FNN (**Fig.4**, black boxes). This suggests that the real FNN must have certain topological features that determine the high NR in the human diet. Analyzing the network properties of those null models (**Fig.S4**), we

247 found that those randomized FNNs all display lower nestedness and higher $\langle d_{ij} \rangle$ than those of
248 the real FNN. Thus, the highly nested structure and low $\langle d_{ij} \rangle$ of the real FNN jointly contribute
249 to the NR values observed in the human diet.

250 **Impact of food composition on personal NR.** To test if food composition plays an important role
251 in determining nNR, we randomized the food composition of each participant using three
252 different randomization schemes, yielding three different null models. Then we recalculated nNR
253 for each sample (**Fig.4b**, colored boxes). We found that for each dietary intake record, if we
254 preserve the food abundance profile but randomly replace the food items by those present in the
255 reference FNN, the resulting null model (Null-composition-1) always yields much higher nNR
256 than that of the original dietary record. This suggests that the food items present in each dietary
257 record are not assembled at random but follow certain assembly rules. Interestingly, if we
258 randomize the food compositions through random permutation of non-zero abundance for each
259 dietary record across different food items (Null-composition-2) or for each food item across
260 dietary records from different participants (Null-composition-3), we found that only Null-
261 composition-3 did not significantly alter nNR, suggesting that the assembly rules are consistent
262 across different participants. Taken together, results from the three null composition models
263 suggest that the human diet prefers low nutritional redundancy.

264 **Correlation between Personal NR and healthy diet scores.** Adherence to healthy diets has the
265 potential to prevent disease and prolong life span. Many healthy diet scores have been developed
266 in the past. To test if the personal NR measure proposed here can be used to quantify a healthy
267 diet, we assessed the association between personal NR and the following four previously
268 established healthy diet scores (see **SI Sec.4** for details): (i) Healthy Eating Index 2005 (HEI-
269 2005)[31]: a score that measures adherence to the USDA 2005 Dietary Guidelines for Americans. (ii)
270 Alternate Healthy Eating Index 2010 (AHEI-2010)[31]: a score that measures adherence to a diet
271 pattern based on foods and nutrients most predictive of disease risk in the literature.
272 (iii) Alternate Mediterranean Diet Score (AMED)[32]: a score adapted from form Mediterranean
273 Diet Score of Ref [33]. (iv) DASH Style Diet Score (DASH)[34]: a score capturing the characteristics
274 of the Dietary Approaches to Stop Hypertension diet. As shown in **Fig.6a**, we found that the NR
275 is positively correlated with AMED (Pearson correlation coefficient $\rho = 0.11$) and DASH ($\rho =$
276 0.06), but negatively correlated with HEI ($\rho = -0.17$) and AHEI ($\rho = -0.09$). But in all cases, the
277 Pearson correlation coefficients are not very strong, implying that personal NR should not be
278 interpreted as being synonymous with a healthy diet score. A heathy diet does not necessarily
279 have higher or lower nutritional redundancy.

280 **Personal NR as an indicator of heathy aging.** Finally, we asked whether personal NR can be an
281 indicator of healthy aging --- an overall indicator we developed by combining measures of

282 physical function, cognitive function, mental health, and chronic diseases[51–53]. To address this
283 question, we focused on a subset of NHS participants (n=21,299) for whom we had all of these
284 four healthy aging components: healthy agers (n=3,491) and usual agers (n=17,808) through a
285 previous study[54]. The healthy agers are defined as those participants who survived beyond 65
286 years of age, with no history of chronic diseases, no reported memory impairment, no physical
287 disabilities, and intact mental health. The remaining participants who survived but did not
288 achieve good health in one or more domains were usual agers. The chronic disease domain used
289 to define usual agers includes cancer (other than nonmelanoma skin cancer), myocardial
290 infarction, coronary artery bypass surgery or percutaneous transluminal coronary angioplasty,
291 congestive heart failure, stroke, type 2 diabetes mellitus, kidney failure, chronic obstructive
292 pulmonary disease, Parkinson's disease, multiple sclerosis, and amyotrophic lateral sclerosis.
293 Other host factors that were collected in this previous study include age, education (registered
294 nurse, bachelor's degree, master, or doctorate), marital status (widowed, married, and
295 single/separated/divorced), median income from census tract (quintiles), BMI (<22, 22–24.9, 25–
296 29.9, and ≥30 kg/m^2), energy intake (quintiles of kcal/day), multivitamin use (yes/no), aspirin use
297 (<1, 1–2, or >2 tablets/week), pack-years of smoking (quintiles), and physical activity (quintiles of
298 MET-h/week) (see **Table S1** for details of those characteristics).

299 We first assessed the correlations between NR and these host factors. We found BMI,
300 aspirin use, and physical activity were negatively correlated with NR, while education, median
301 income, total energy intake, and pack-years of smoking were positively correlated with NR
302 (**Fig.6b**). In all cases, the Pearson correlation coefficients are very weak. To compare the
303 prediction performance of each of the four healthy diet scores (HEI-2005, AHEI-2010, AMED,
304 DASH) with that of personal NR in predicting healthy aging, we built a random forest classifier.
305 In particular, we used features (personal NR or one of the healthy diet scores, and the host factors
306 mentioned above) collected for those participants in 1998 to predict their heathy aging status in
307 2012 (see **SI Sec.5** for details). We found that personal NR can achieve very similar error rate (i.e.,
308 the proportion of participants that have been incorrectly classified by the model) and AUC (area
309 under the ROC curve) as other four healthy diet scores (**Fig.6c,d**). Interestingly, we found NR can
310 serve as a better indicator than many other host factors (see **Fig.S5** for the importance ranking of
311 those factors). We emphasized that the healthy aging prediction does not rely on the particular
312 NR definition or classifier. We evaluated the performance of NR in healthy aging prediction using
313 the Hill number-based definition and another ensemble classifier: XGBoost[55], showing that both
314 AUC and error rate are robust to different Hill numbers and classifier (see **Figs.S6, S7** and **SI**
315 **sec.2** for details).

316

Discussion

Through examining various human dietary intake datasets, we found that food profile varies tremendously across individuals and over time, while the nutritional profile is highly conserved across different individuals and over time. To quantify this NR, we constructed the food-nutrient network --- a bipartite graph that connects foods to their nutrient constituents. This food-nutrient network also allows us to assess the NR of any dietary record from any individual. We found that this personal NR is not strongly correlated with any existing healthy diet scores, but it can be used to predict healthy aging with equally strong performance as those healthy diet scores.

There are several limitations in our current framework of NR calculation. *First*, we did not explicitly consider the nutrient difference from different food sources. We understand that the nutrient quality from different food sources can vary a lot. We anticipate that incorporating this information in our NR calculation will further improve the power of using NR to predict healthy aging or other disease risks[56]. *Second*, the calculation of a personal NR relies on food intake measurements, e.g., ASA24 and FFQ, which are based on self-reported dietary intake questionnaires. We understand that such food intake measurements have inherent limitations, particularly measurement error related to poor recall, which can be overcome by the use of nutritional biomarkers that are capable of objectively assessing food consumption in different biological samples without the bias of self-reported dietary assessment[57]. Although nutritional biomarkers provide a more proximal measure of nutrient status than dietary intake, quantitatively studying NR using nutritional biomarkers is beyond the scope of the current study. We anticipate that our framework will trigger more research activities in this direction. *Third*, our calculation does not involve many non-nutrients. Here, non-nutrients are those bioactive compounds (e.g., polyphenols, phytosterols, and saponins) that do not fit into the category of nutrients, yet play an important role in regulating bodily functions and disease prevention[58]. We are aware of the growing interest in non-nutrients. While combining both nutrients and non-nutrients in our framework might yield a more powerful measure of NR to predict healthy aging and other disease risks, we will employ this approach in future work.

Data availability

The DMAS dataset is downloaded from https://github.com/knights-lab/dietstudy_analyses. The USDA Food Nutrient Database is downloaded from https://fdc.nal.usda.gov/download-datasets.html. The authors declare that all in-house data supporting the findings of this study are available upon request to the Women's lifestyle validation study, Men's lifestyle validation study and Nurses' Health Study (NHS) and when the request for data access is approved (contact email: nhsaccess@channing.harvard.edu). Harvard University Food Composition Database can be accessed at https://regepi.bwh.harvard.edu/health/nutrition.html.

Code availability

MATLAB Code (version R2016b) used in this work will be available at https://github.com/spxuw/Nutritional-redundancy.

References

1. Richards, T. J., Patterson, P. M. & Tegene, A. Obesity and nutrient consumption: a rational addiction? Contemporary Economic Policy 25, 309–324 (2007).
2. Schönfeldt, H. C. & Gibson, N. Changes in the nutrient quality of meat in an obesity context. Meat Science 80, 20–27 (2008).
3. Keen, H., Thomas, B. J., Jarrett, R. J. & Fuller, J. H. Nutrient intake, adiposity, and diabetes. British Medical Journal 1, 655–658 (1979).
4. Patti, M.-E. & Kahn, B. B. Nutrient sensor links obesity with diabetes risk. Nature Medicine 10, 1049–1050 (2004).
5. Milanlouei, S., Menichetti, G., Li, Y., Lascalzo, J., Willette, W.C. & Barabási, A.L. A systematic comprehensive longitudinal evaluation of dietary factors associated with acute myocardial infarction and fatal coronary heart disease. Nature Communications 11, 6074 (2020).
6. Bellisle, F. Why should we study human food intake behaviour? Nutrition, Metabolism and Cardiovascular Diseases 13, 189–193 (2003).
7. Song, M., Fung, T.T., Hu, F.B., Willett, W.C., Longo, V.D., Chan, A.T. & Giovannucci, E.L. Animal and plant protein intake and all-cause and cause-specific mortality: results from two prospective US cohort studies. JAMA International Medicine 176, 1453–1463 (2016).
8. Martínez-González, M. A., Salas-Salvadó, J., Estruch, R., Corella, D., Fitó, M., Ros, E. & Predimed Investigators. Benefits of the Mediterranean diet: insights from the PREDIMED study. Progress in Cardiovascular Diseases 58, 50–60 (2015).
9. Chiu, S., Bergeron, N., Williams, P.T., Bray, G.A., Sutherland, B. & Krauss, R.M. Comparison of the DASH (Dietary Approaches to Stop Hypertension) diet and a higher-fat DASH diet on blood pressure and lipids and lipoproteins: a randomized controlled trial1–3. The American Journal of Clinical Nutrition 103, 341–347 (2016).
10. Dietary Guidelines for Americans, 2020-2025. 164.
11. Ulbricht, T. L. V. & Southgate, D. A. T. Coronary heart disease: seven dietary factors. The Lancet 338, 985–992 (1991).
12. Hu, F. B. & Willett, W. C. Optimal diets for prevention of coronary heart disease. JAMA 288, 2569–2578 (2002).
13. Hu, F. B., Stampfer, M.J., Manson, J.E., Rimm, E., Colditz, G.A., Rosner, B.A., Hennekens, C.H. & Willett, W.C. Dietary fat intake and the risk of coronary heart disease in women. New England Journal of Medicine 337, 1491–1499 (1997).
14. Kratz, M., Baars, T. & Guyenet, S. The relationship between high-fat dairy consumption and obesity, cardiovascular, and metabolic disease. European Journal of Nutrition 52, 1–24 (2013).
15. Fu, J.-H., Sun, H.S., Wang, Y., Zheng, W.Q., Shi, Z.Y. & Wang, Q.J. The effects of a fat-and sugar-enriched diet and chronic stress on nonalcoholic fatty liver disease in male Wistar rats. Digestive Diseases and Sciences 55, 2227–2236 (2010).
16. Oddy, W. H., Herbison, C.E., Jacoby, P., Ambrosini, G.L., O'sullivan, T.A., Ayonrinde, O.T., Olynyk, J.K., Black, L.J., Beilin, L.J., Mori, T.A. & Hands, B.P. The Western dietary pattern is prospectively associated with nonalcoholic fatty liver disease in adolescence. American Journal of Gastroenterology 108, 778–785 (2013).

17. Chen, Y., Michalak, M. & Agellon, L. B. Focus: Nutrition and Food Science: Importance of Nutrients and Nutrient Metabolism on Human Health. The Yale Journal of Biology and Medicine 91, 95 (2018).
18. Drewnowski, A. & Fulgoni III, V. Nutrient profiling of foods: creating a nutrient-rich food index. Nutrition Reviews 66, 23–39 (2008).
19. Garsetti, M., de Vries, J., Smith, M., Amosse, A. & Rolf-Pedersen, N. Nutrient profiling schemes: overview and comparative analysis. European Journal of Nutrition 46, 15–28 (2007).
20. Martin, C. L., Montville, J.B., Steinfeldt, L.C., Omolewa-Tomobi, G., Heendeniya, K.Y., Adler, M.E. & Moshfegh, A.J. USDA Food and Nutrient Database for Dietary Studies 2011–2012: Documentation and User Guide. Beltsville, MD: US Department of Agriculture.
21. Sebastian, R. S., Wilkinson Enns, C., Goldman, J.D., Steinfeldt, L.C., Martin, C.L. & Moshfegh, A.J. Flavonoid Values for USDA Survey Foods and Beverages 2007–2010. Worldwide Web Site: Food Surveys Research Group (2016).
22. National Food Institute. Technical University of Denmark. Frida Food data version 2 (DTU, 2016); frida.fooddata.dk.
23. FooDB. Compounds http://foodb.ca/compounds (access 1 August 2019).
24. Neveu, V., Perez-Jiménez, J., Vos, F., Crespy, V., du Chaffaut, L., Mennen, L., Knox, C., Eisner, R., Cruz, J., Wishart, D. & Scalbert, A. Phenol-Explorer: an online comprehensive database on polyphenol contents in foods. Database 2010, (2010).
25. Plumb, J., S., Bompola, F., Cushen, M., Pinchen, H., Nørby, E., Astley, S., Lyons, J., Kiely, M. & Finglas, P. Ebasis (bioactive substances in food information systems) and bioactive intakes: Major updates of the bioactive compound composition and beneficial bioeffects database and the development of a probabilistic model to assess intakes in Europe. Nutrients 9, 320 (2017).
26. Barabási, A.-L., Menichetti, G. & Loscalzo, J. The unmapped chemical complexity of our diet. Nature Food 1, 33–37 (2020).
27. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature 486, 207–214 (2012).
28. Louca, S., Polz, M.F., Mazel, F., Albright, M.B., Huber, J.A., O'Connor, M.I., Ackermann, M., Hahn, A.S., Srivastava, D.S., Crowe, S.A. & Doebeli, M. Function and functional redundancy in microbial systems. Nature Ecology & Evolution 2, 6 (2018).
29. Tian, L., X.W., Wu, A.K., Fan, Y., Friedman, J., Dahlin, A., Waldor, M.K., Weinstock, G.M., Weiss, S.T. & Liu, Y.Y. Deciphering functional redundancy in the human microbiome. Nature Communications 11, 6217 (2020).
30. Ahn, Y.-Y., Ahnert, S. E., Bagrow, J. P. & Barabási, A.-L. Flavor network and the principles of food pairing. Scientific Reports **1**, 1–7 (2011).
31. Chiuve, S. E., Fung, T.T., Rimm, E.B., Hu, F.B., McCullough, M.L., Wang, M., Stampfer, M.J. and Willett, W.C. Alternative dietary indices both strongly predict risk of chronic disease. The Journal of Nutrition 142, 1009–1018 (2012).
32. Fung, T. T., Rexrode, K.M., Mantzoros, C.S., Manson, J.E., Willett, W.C. & Hu, F.B. Mediterranean diet and incidence and mortality of coronary heart disease and stroke in women. Circulation 119, 1093 (2009).

465 33. Trichopoulou, A., Costacou, T., Bamia, C. & Trichopoulos, D. Adherence to a Mediterranean
466 diet and survival in a Greek population. New England Journal of Medicine 348, 2599–2608
467 (2003).
468 34. Fung, T. T., Chiuve, S.E., McCullough, M.L., Rexrode, K.M., Logroscino, G. & Hu, F.B.
469 Adherence to a DASH-style diet and risk of coronary heart disease and stroke in women.
470 Archives of Internal Medicine 168, 713–720 (2008).
471 35. Johnson, A. J., Vangay, P., Al-Ghalith, G.A., Hillmann, B.M., Ward, T.L., Shields-Cutler,
472 R.R., Kim, A.D., Shmagel, A.K., Syed, A.N., Students, P.M.C. & Walter, J. Daily Sampling
473 Reveals Personalized Diet-Microbiome Associations in Humans. Cell Host & Microbe 25,
474 789-802.e5 (2019).
475 36. Subar, A., Kirkpatrick, S.I., Mittl, B., Zimmerman, T.P., Thompson, F.E., Bingley, C.,
476 Willis, G., Islam, N.G., Baranowski, T., McNutt, S. & Potischman, N. The Automated Self-
477 Administered 24-hour dietary recall (ASA24): a resource for researchers, clinicians, and
478 educators from the National Cancer Institute. Journal of the Academy of Nutrition and
479 Dietetics 112, 1134–1137 (2012).
480 37. Kirkpatrick, S. I., Subar, A.F., Douglass, D., Zimmerman, T.P., Thompson, F.E., Kahle, L.L.,
481 George, S.M., Dodd, K.W. & Potischman, N. Performance of the Automated Self-
482 Administered 24-hour Recall relative to a measure of true intakes and to an interviewer-
483 administered 24-h recall. The American Journal of Clinical Nutrition 100, 233–240 (2014).
484 38. Yuan, C., Spiegelman, D., Rimm, E.B., Rosner, B.A., Stampfer, M.J., Barnett, J.B. Chavarro,
485 J.E., Subar, A.F., Sampson, L.K. & Willett, W.C. Validity of a Dietary Questionnaire
486 Assessed by Comparison With Multiple Weighed Dietary Records or 24-Hour Recalls.
487 American Journal of Epidemiol 185, 570–584 (2017).
488 39. Colditz, G. A. & Hankinson, S. E. The Nurses' Health Study: lifestyle and health among
489 women. Nature Reviews Cancer 5, 388–396 (2005).
490 40. Colditz, G. A., Manson, J. E. & Hankinson, S. E. The Nurses' Health Study: 20-year
491 contribution to the understanding of health among women. Journal of Women's Health 6,
492 49–62 (1997).
493 41. Mehta, R. S., Abu-Ali, G.S., Drew, D.A., Lloyd-Price, J., Subramanian, A., Lochhead, P.,
494 Joshi, A.D., Ivey, K.L., Khalili, H., Brown, G.T. & DuLong, C. Stability of the human faecal
495 microbiome in a cohort of adult men. Nature Microbiology 3, 347–355 (2018).
496 42. Abu-Ali, G. S., Mehta, R.S., Lloyd-Price, J., Mallick, H., Branck, T., Ivey, K.L., Drew, D.A.,
497 DuLong, C., Rimm, E., Izard, J. & Chan, A.T. Metatranscriptome of human faecal microbial
498 communities in a cohort of adult men. Nature microbiology 3, 356–366 (2018).
499 43. Chiuve, S. E., Rexrode, K.M., Spiegelman, D., Logroscino, G., Manson, J.E. & Rimm, E.B.
500 Primary prevention of stroke by healthy lifestyle. Circulation 118, 947 (2008).
501 44. Mukamal, K. J., Chiuve, S. E. & Rimm, E. B. Alcohol consumption and risk for coronary
502 heart disease in men with healthy lifestyles. Archives of Internal Medicine 166, 2145–2150
503 (2006).
504 45. Hunter, D. J., Sampson, L., Stampfer, M.J., Colditz, G.A., Rosner, B. & Willett, W.C.
505 Variability in portion sizes of commonly consumed foods among a population of women in
506 the UnitedStates. American Journal of Epidemiology 127, 1240–1249 (1988).

46. Legendre, P. & De Cáceres, M. Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. Ecology Letters 16, 951–963 (2013).
47. Pillar, V. D., Blanco, C.C., Müller, S.C., Sosinski, E.E., Joner, F. & Duarte, L.D. Functional redundancy and stability in plant communities. Journal of Vegetation Science 24, 963–974 (2013).
48. Almeida-Neto, M., Guimaraes, P., Guimaraes Jr, P. R., Loyola, R. D. & Ulrich, W. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. Oikos 117, 1227–1239 (2008).
49. Ulrich, W., Almeida-Neto, M. & Gotelli, N. J. A consumer's guide to nestedness analysis. Oikos 118, 3–17 (2009).
50. Payrató-Borràs, C., Hernández, L. & Moreno, Y. Breaking the Spell of Nestedness: The Entropic Origin of Nestedness in Mutualistic Systems. Physical Review X 9, 031024 (2019).
51. Baumgartner, R. N. Body composition in healthy aging. Annals of the New York Academy of Sciences 904, 437–448 (2000).
52. Carmona, J. J. & Michan, S. Biology of healthy aging and longevity. Revista de Investigacion Clinica 68, 7–16 (2016).
53. Willcox, D. C., Scapagnini, G. & Willcox, B. J. Healthy aging diets other than the Mediterranean: a focus on the Okinawan diet. Mechanisms of Ageing and Development 136, 148–162 (2014).
54. Freitas-Simoes, T.-M., Wagner, M., Samieri, C., Sala-Vila, A. & Grodstein, F. Consumption of Nuts at Midlife and Healthy Aging in Women. Journal of Aging Research 2020, 1–7 (2020).
55. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794 (2016) doi:10.1145/2939672.2939785.
56. Giovannucci, E., Rimm, E.B., Stampfer, M.J., Colditz, G.A., Ascherio, A. and Willett, W.C. Intake of fat, meat, and fiber in relation to risk of colon cancer in men. Cancer research 54, 2390–2397 (1994).
57. Picó, C., Serra, F., Rodríguez, A. M., Keijer, J. & Palou, A. Biomarkers of Nutrition and Health: New Tools for New Approaches. Nutrients 11, (2019).
58. Ribeiro, P. V. de M., Andrade, P.A., Hermsdorff, H.H.M., Dos Santos, C.A., Cotta, R.M.M., Estanislau, J.D.A.S.G., de Oliveira Campos, A.A. and Rosa, C.D.O.B. Dietary non-nutrients in the prevention of non-communicable diseases: Potentially related mechanisms. Nutrition 66, 22–28 (2019).

Acknowledgements. We thank Laura Sampson, Walter C. Willett, Rui Song, and Jorge Chavarro for valuable discussions. Research reported in this publication was supported by grants R01AI141529, R01HD093761, UH3OD023268, U19AI095219, and U01HL089856 from National Institutes of Health.

Author Contributions. Y.-Y.L. conceived and designed the project. X.-W.W. performed all the numerical calculations and real data analysis. All authors analyzed and interpreted the results. X.-W.W. and Y.-Y.L. wrote the manuscript. All authors edited the manuscript.

Author Information The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to Y.-Y.L. (yyl@channing.harvard.edu).

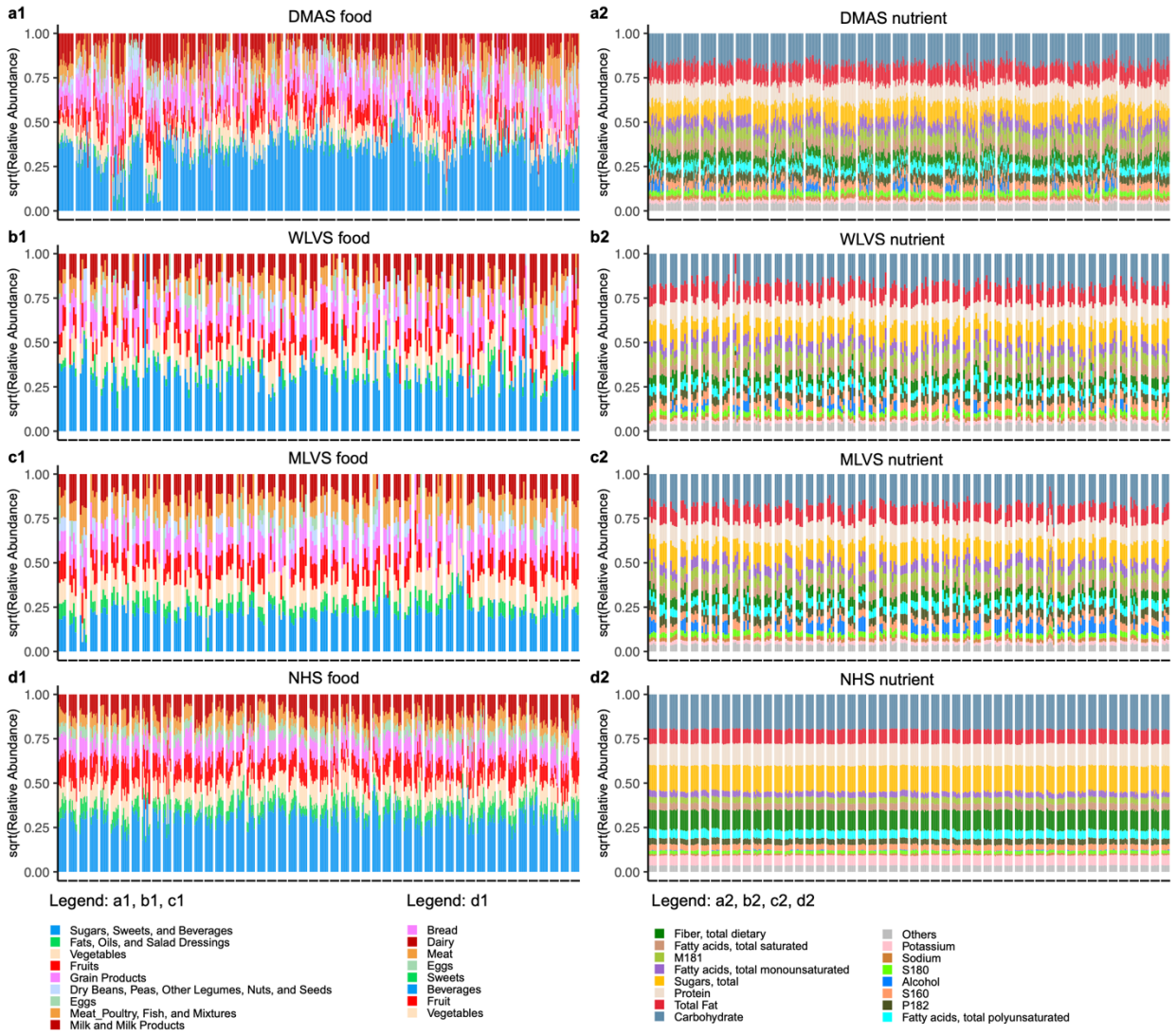

Figure 1: Dietary food profiles are highly personalized while the nutrient profiles are highly stable. Rows represent results from selected participants of different cohorts: (**a**) DMAS (dietary intake data collected using ASA24 dietary assessment tool daily over 17 consecutive days); (**b**) WLVS (with four ASA24 records within one year); (**c**) MLVS (with four ASA24 records within one year); (**d**) NHS (with FFQ administrated every four years and with total eight time points). Columns: (**1**) food profiles; (**2**) nutrient profiles. For DMAS, we plot the food and nutrient profiles of 30 participants who have ASA24 data available for all the 17 days. For the WLVS, MLVS and NHS datasets, we plot the food and nutrient profiles of 50 randomly chosen individuals. In the visualization of nutrient profiles, we only show the top-15 most abundant nutrients, while the remain nutrients are summarized as others.

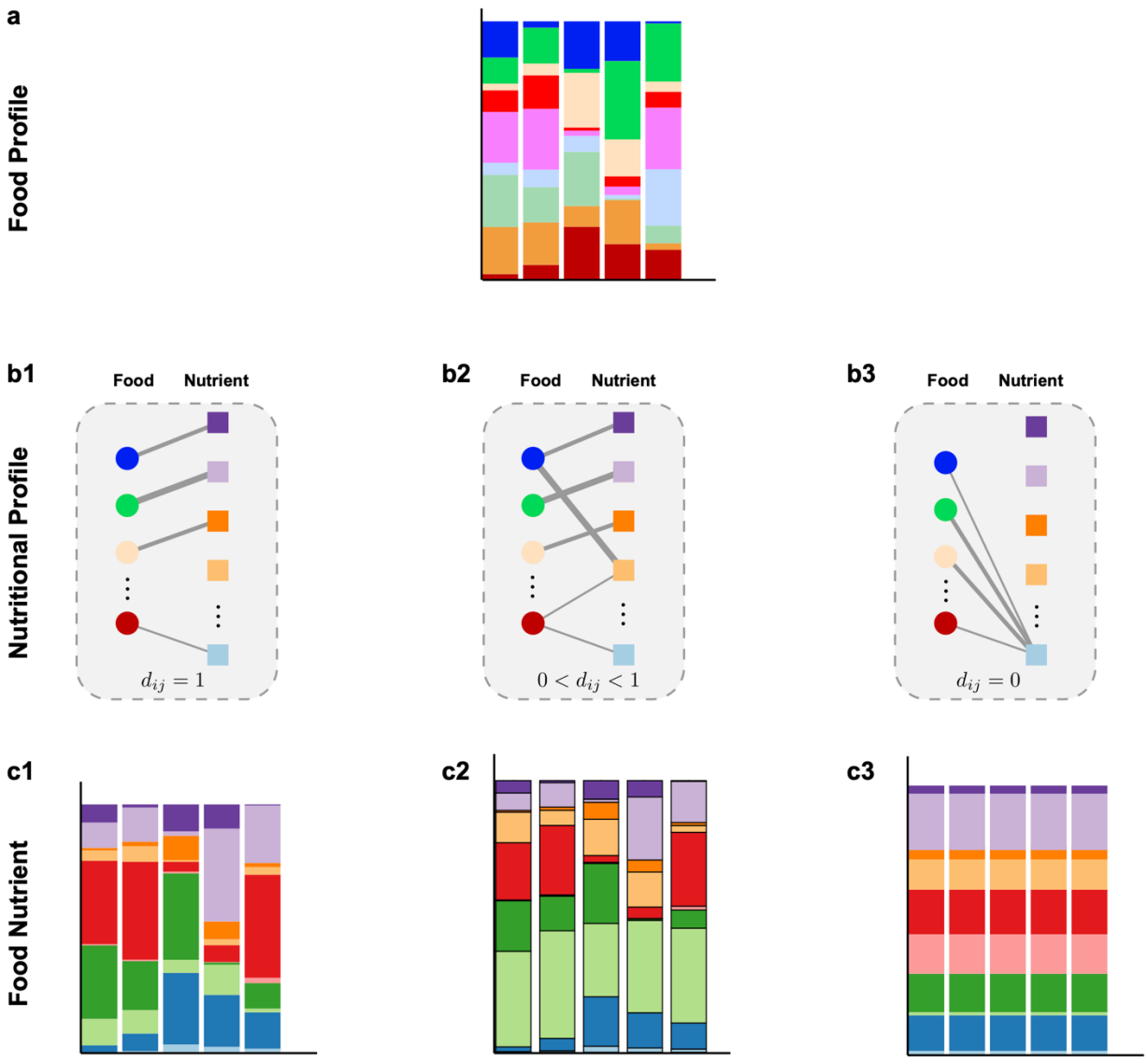

Figure 2: Structure of the food-nutrient network is crucial for determining the nutrient redundancy of human diet. a: The food profiles vary drastically across individuals. **b1-b3:** Food-nutrient network is a bipartite graph that connect food items to their nutrient constituents, with edge weight representing the amount contributed by food-i to nutrient-a **b1:** Each food has a unique set of nutrient constituents. **b2:** Different foods share a few common nutrients; some foods are specialized to have some unique nutrients. **b3:** All foods share exactly the same set of nutrients. For each individual, its nutritional profile can be calculated from its food profile in (a) and the food-nutrient network in (b). **c1:** The nutritional profiles vary drastically across different individuals. **c2:** The nutritional profiles are highly conserved across different individuals. **c3:** The nutritional profiles are exactly the same across all individuals.

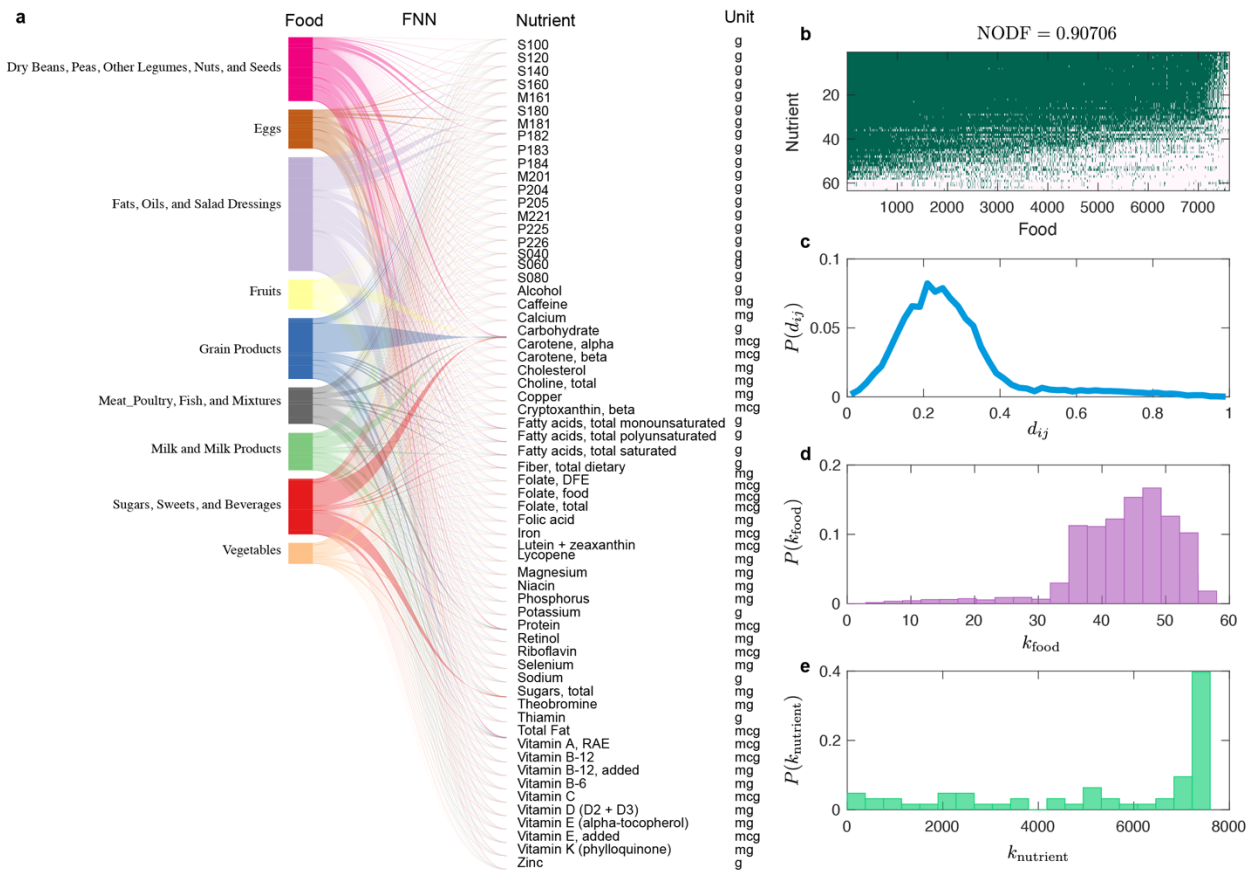

Figure 3: The food-nutrient network (FNN) has a nested structure. a: We used USDA's Food and Nutrient Database for Dietary Studies (FNDDS) to construct the reference FNN, which consists of 7,618 foods and 65 nutrients and micronutrients. This reference FNN is depicted as a bipartite graph, where for visualization purposes each food node represents one of the nine highest-level food groups (based on the FNDDS food coding scheme) and each nutrient node represents a nutrient. **b:** The incidence matrix of the FNN, where the presence (or absence) of a link between a food and a nutrient is colored in green (or white), respectively. **c:** The probability distribution of nutrient distances (d_{ij}) among different food items. The bin size is 0.02. **d:** The food degree distribution. Here, the degree of a food item is the number of distinct nutrients it contains. **e:** The nutrient degree distribution. Here, the degree of a nutrient is the number of food items that contain this nutrient.

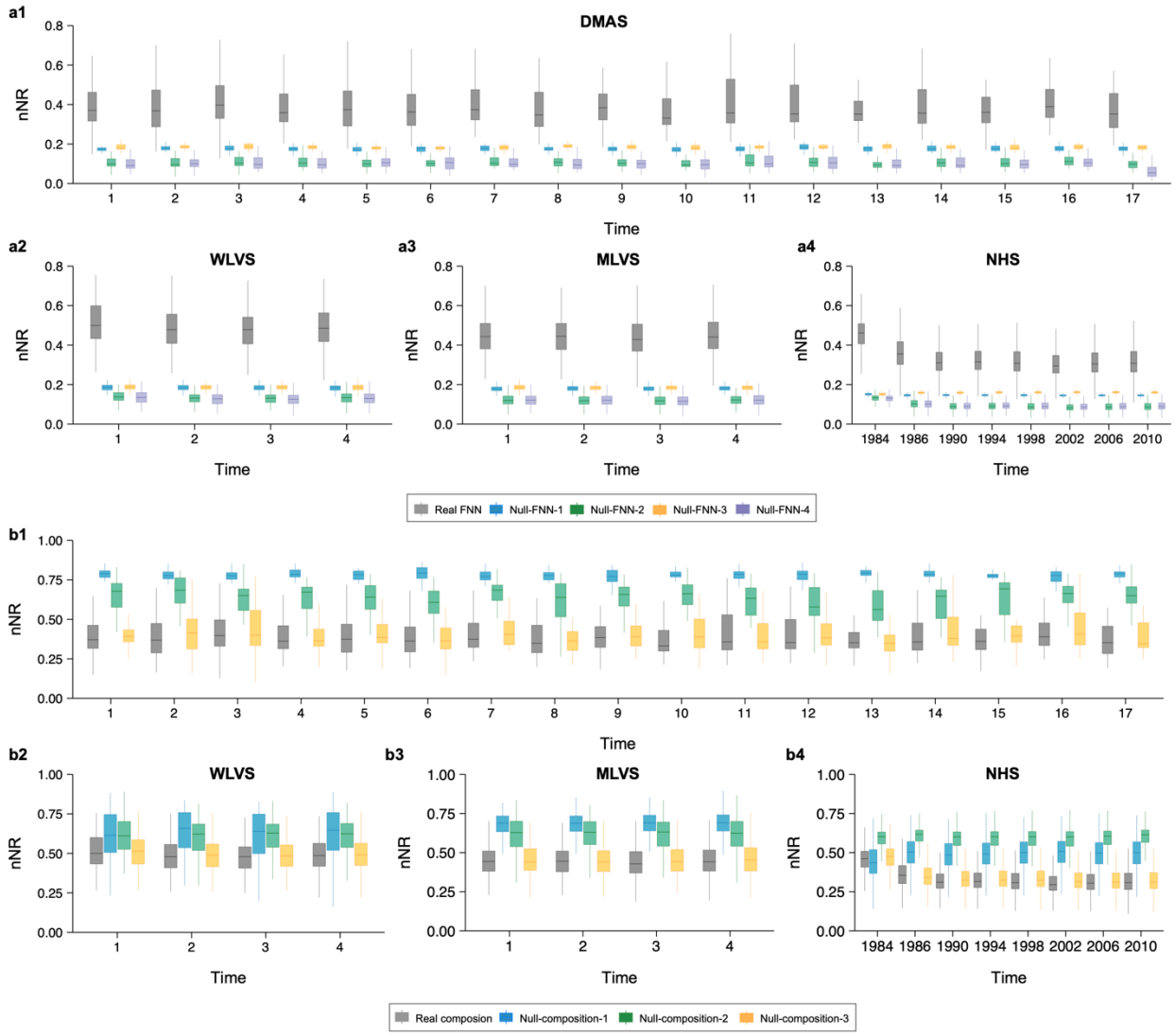

Figure 4: Topological features of the food-nutrient network and the human dietary pattern contributes to the nutritional redundancy. a1-a4, The box plots of the normalized nutritional redundancy (nNR ≡ NR/FD) were calculated from the real FNN (black box), as well as the randomized FNNs (colored boxes) using four different randomization schemes: Complete randomization (Null-FNN-1); Food-degree preserving randomization (Null- FNN-2); Nutrient-degree preserving randomization (Null-FNN-3); Food- and nutrient-degree preserving randomization (Null-FNN-4). Here the degree of a nutrient is the number of foods that contain it, and the degree of a food is the number of nutrients contained in it. **b1-b4,** The box plots of nNR were calculated from the real dietary records (black box), as well as the randomized dietary records (colored boxes) using three different randomization schemes: Randomized food assemblage generated by randomly choosing the same number of food items from the food pool but keeping the food profile unchanged (Null-composition-1); Randomized food abundance profiles through random permutation of non-zero abundance for each participant across different foods (Null-composition-2); Randomized food abundance profiles through random permutation of non-zero abundance for each food across different participants (Null-composition-3).

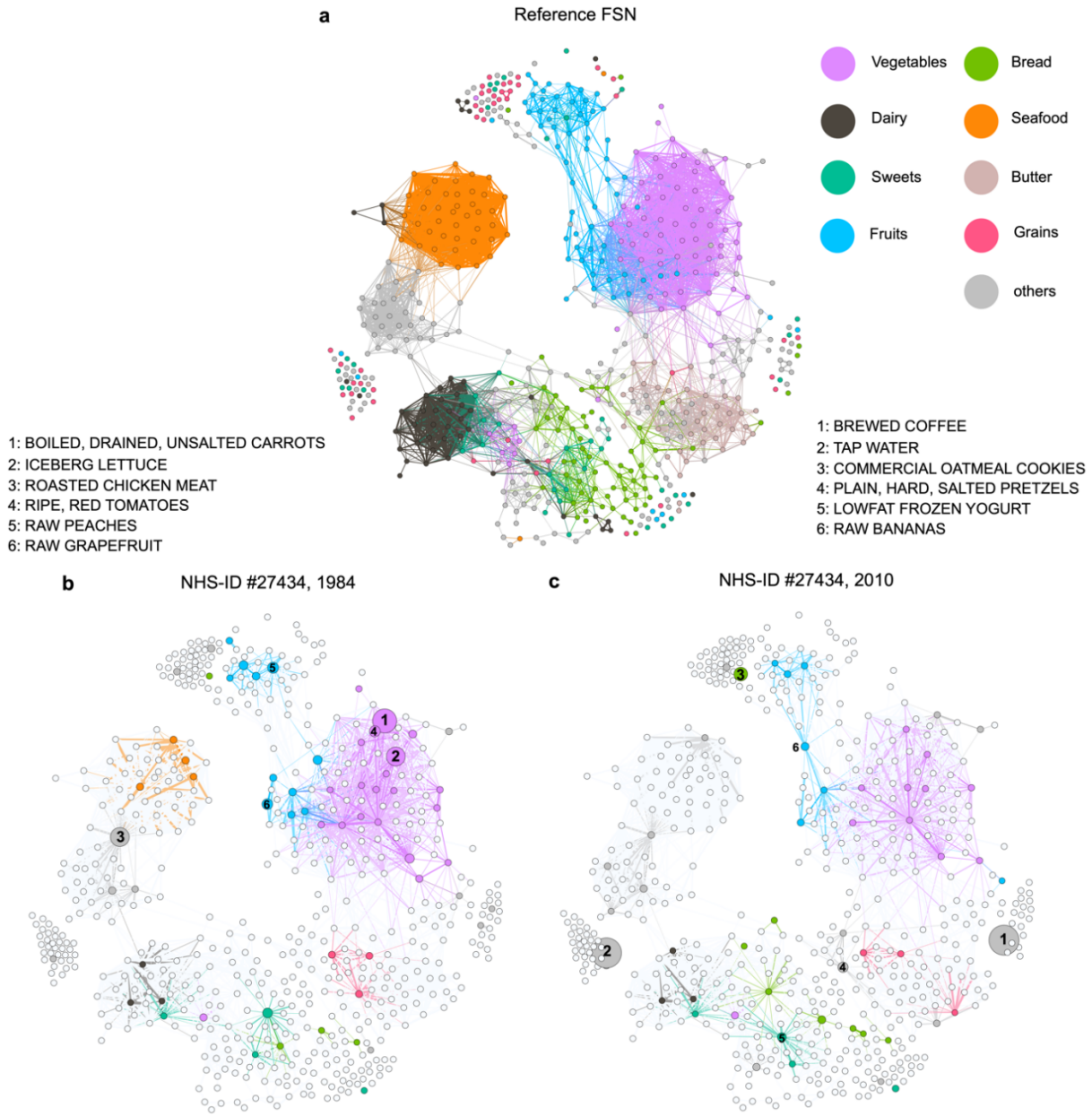

Figure 5: Reference and personal food similarity networks. a: The reference food similarity network (FSN) was projected from the reference food-nutrient network constructed from the Harvard food composition database used in the NHS. The similarity between food-i and food-j was calculated using the weighted Jaccard similarity index s_{ij}. Only links with $s_{ij} \geq 0.85$ were kept in the visualization. The color of each node represents the food group it belongs to. **b:** The personal FSN constructed for a particular NHS participant (ID #27434) using her FFQ data collected in 1984. This network is a subgraph of the reference FSN shown in panel-a, thus those nodes and links in global network but do not appear in the subnetwork are shown in white. **c:** The personal FSN of the same NHS participant constructed using her FFQ data collected in 2010. In panel-b and c, node sizes are proportional to the relative abundance of the food items and edge widths are proportional to the food similarities.

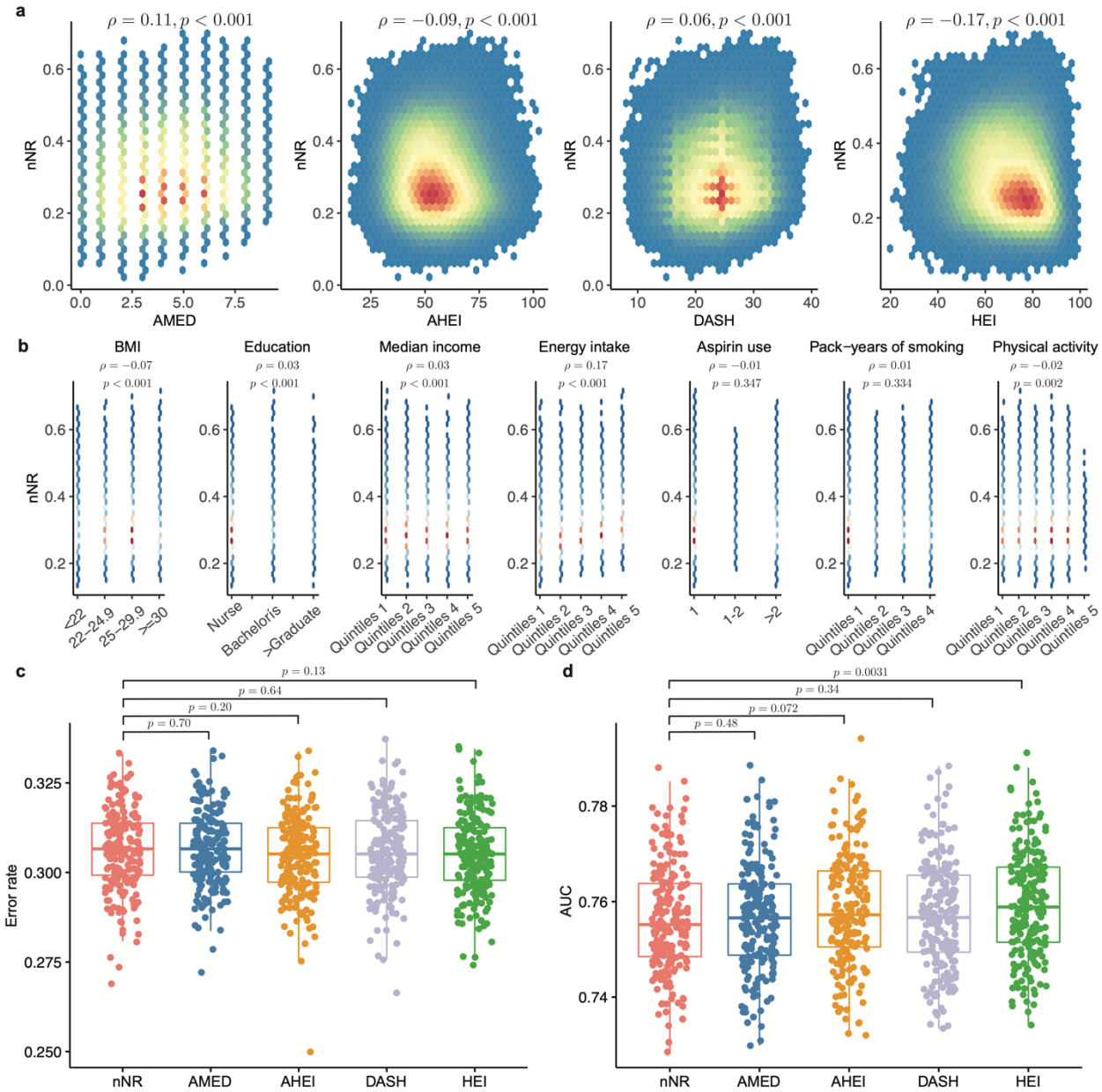

Figure 6: Nutrient redundancy serves as a potential metric to predict healthy aging. a, Pearson correlation between the normalized nutritional redundancy (nNR) and existing healthy diet scores: Alternate Mediterranean Diet Score (AMED); Alternate Healthy Eating Index 2010 (AHEI-2010); Dash Style Diet Score (DASH); and Healthy Eating Index 2005 (HEI-2005). P values were calculated from t-test. **b:** Pearson correlation between nNR and several host factors: body-mass index (BMI); education level; median income; energy intake level; Aspirin use; pack-years of smoking; physical activity. P values were calculated from t-test. **c:** Error rate of random forest classifier in the prediction of healthy aging status. **d:** AUC of random forest classifier in prediction of healthy aging status. The participants are randomly spitted into 80% as the training set and the remaining 20% as the test set. The boxplot represents the performances of 200 independent splits. All FDR-corrected P values were found using the two-sided Mann–Whitney *U* test.

Nutritional Redundancy in Human Diet

Supplemental Information

Xu-Wen Wang[1], Francine Grodstein[2], Shilpa N. Bhupathiraju[1,3], Qi Sun[1,3,4], Xuehong Zhang[1,3], Frank Hu[3,4], Scott T. Weiss[1], Yang-Yu Liu[1]

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

[2]*Rush Alzheimer's Disease Center, Department of Internal Medicine, Rush Medical College, Rush University, Chicago, IL 60612, USA*

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

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

1. NR calculating using human dietary data **2**
1.1 Reference FNN 2
1.2 Nutrition profile analysis 3
1.3 Food Choice Analysis 3

2. Basic definitions **3**
2.1 Food nutrient network 3
2.2 Nestedness 3
2.3 Nutritional distance measure 4

3. Nutritional redundancy measure **5**
3.1 Food diversity 5
3.2 Nutritional diversity 5
3.3 Nutritional redundancy 6

4. Healthy diet scores **7**
4.1 HEI-2005 7
4.2 AHEI-2010 7
4.3 AMED 7
4.4 DASH 7

5. Healthy aging prediction **8**

6. Supplementary figures and tables **9**

1. NR calculating using human dietary data

1.1 Reference FNN

1.1.1 USDA database

To construct the Food-Nutrient network, we downloaded the FNDDS 2011-2012 from the USDA database, which including 7,618 foods and 65 macronutrients. The USDA National Nutrient Database for Dietary Studies is the major source of food composition data in the United States. To be consistent with the version used in DMAS study, we chose the version 2011-2012 to construct reference Food-Nutrient network (FNN). This version includes 7,618 foods, which can be clarified into 9 highest items and the total number of nutrients is 65.

Remark 1: In calculating the nutritional profiles (Fig.1) as well as the nutritional redundancy (Fig.4), we excluded energy and water for the following considerations. First, energy does not have a unit of mass, and hence cannot be included in the nutritional profile where components represent relative abundances. Consequently, it cannot be used to calculate the nutritional redundancy either. Second, water was not considered as a nutrient in the Harvard Food Composition Database (HFDB). For consistency and comparison purposes, we also removed it from FNDDS when we calculate the nutritional profiles and nutritional redundancy from the DMAS, WLVS, MLVS data.

1.1.2 Frida database

The database Frida Food Data (frida.fooddata.dk) was created and published by the National Food Institute, Technical University of Denmark (DTU), including data on nutrient content of various foods. We used the version released at 08-02-2019, which includes 1,185 foods items and 198 nutrients.

1.1.3 Harvard Food Composition Database

We used the Harvard food composition table of year 2015 to construct a reference food-nutrient network to calculate the nutritional redundancy of participants from the NHS. The Harvard food composition table consists of 575 foods and 182 nutrients. To calculate the nutritional redundancy, we removed calories due to different unit and total protein, total fat and total sugar since those nutrients are overlapped with some sub-nutrients.

1.1.4 Other databases

Other databases, for example, FooDB, a database representing the most comprehensive effort to integrate food composition data from specialized databases and experimental data, has provided the information of 26,625 distinct bio-chemicals in foods. PhenolExplorer and eBasis have also resulted in wealth of information on food composition. Throughout our analysis, we focused on the nutrient level, rather than composition and compounds levels, thus we reference FNN are constructed by USDA and Frida databases.

1.2 Nutrition profile analysis

The nutritional composition of foods for DMAS was determined using ASA24-2016. Both of WLVS and MLVS composition of foods for DMAS was determined using ASA24-2012. ASA24 assigns nutrient information to foods using the USDA's Food and Nutrient Dietary Database (FNDDS). Subjects reported dietary intake as food records and entered their own dietary records directly into ASA24. To calculate the relative abundance of nutrient, we removed energy (unit is not in mass) and water (the water information is not included in NHS database, so we removed to be consistent).

1.3 Food Choice Analysis

Foods of DMAS, WLVS and MLVS were categorized according to their FNDDS food code and modification code as assigned by ASA24. The foods of NHS were mapped from serving data to the Harvard food composition table according to the food descriptions.

2. Basic definitions

2.1 Food nutrient network

Consider a pool of N food items, which contains a collection of M nutrients in total. The food profile $\boldsymbol{f}^{(\nu)} = \left[f_1^{(\nu)}, \cdots, f_N^{(\nu)}\right]$ of the diet record of individual-ν can be directly related to its nutrient profile $\boldsymbol{n}^{(\nu)} = \left[n_1^{(\nu)}, \cdots, n_M^{(\nu)}\right]$ through the FNN (Fig.2). Here, $f_i^{(\nu)}$ (or $n_a^{(\nu)}$) represents the relative abundance of food-i (or nutrient-a) in the diet record. And we define the FNN as a weighted bipartite graph connecting these foods to their nutrients. The FNN can be represented by an $N \times M$ incidence matrix $\boldsymbol{G} = (G_{ia})$, where a non-negative value G_{ia} indicates the amount contributed by food-i to nutrient-a (see Fig.2a for the unit of each nutrient). The nutrient profile is given by $\boldsymbol{n}^{(\nu)} = c\boldsymbol{f}^{(\nu)}\,\boldsymbol{G}$, where $c = \left[\sum_{a=1}^{M}\sum_{i=1}^{N} f_i^{(\nu)}\, G_{ia}\right]^{-1}$ is a normalization constant.

2.2 Nestedness

Nestedness a classical concept in ecology, which is used to characterize the nested structure of ecological systems, such as the species-site network (describing the distribution of species across geographic locations), and the species-species interaction networks (e.g., host-parasite, plant-pollinator interactions)[1–5]. In principle, an ecological system is said to be nested if the items belonging to "smaller" elements (e.g., a small island containing few species, or a specialist species with few interactions) tend to be a subset of the items belonging to "larger" elements (e.g., a large island containing many species, or a generalist species with many interacting partners). Mathematically, those ecological systems can be represented as bipartite graphs with two types of nodes, e.g., sites and species, hosts and parasites, plants and pollinators, etc. In this work, we focus on the food nutrient network of dietary, which is also a bipartite graph with two types of nodes: foods and nutrients.

2.2.1 Numerical calculation

Consider a general bipartite graph with N type-1 nodes and M type-2 nodes. The structure of this bipartite graph can be represented by its N × M binary incidence matrix $\boldsymbol{B} = (B_{ia})$, where $B_{ia} = 1$ if there is a link connecting the i-th type-1 node and the a-th type-2 node, and 0 otherwise. Mathematically, nestedness can be defined as a property of the incidence matrix $\boldsymbol{B}$. If there exists a permutation of rows and columns such that the set of links in row-i contains the links in row-$(i+1)$, and the set of links in column-a contains those in column-$(a+1)$, then B is a perfectly nested binary matrix. For example, consider the mainland and a series of islands sorted according to their distances to the mainland. The mainland contains all the species, the first island has a subset of species in the mainland, the second island has a subset of species in the first island, etc.

The NODF nestedness metric is based on two simple properties: decreasing fill and paired overlap. For any given bipartite graph with incident matrix $\boldsymbol{B}$, the unweighted degree of i-th food node is $k_i = \sum_{a=1}^{M} B_{ia}$, and the unweighted nutrient degree of a-th nutrient node is $k_a = \sum_{i=1}^{N} B_{ia}$. The number of nutrients shared by the i-th food and the j-th food is given by $P_{ij} = \sum_{a=1}^{M} B_{ia}B_{ja}$. And the number of common foods that both of a-th nutrient and b-th nutrient are included is given by $Q_{ij} = \sum_{i=1}^{N} B_{ia}B_{ib}$. Define $\tilde{P}_{ij} = 0$ if $k_i = k_j$ and $\tilde{P}_{ij} = P_{ij}/\min(k_i, k_j)$ otherwise. Then, define $\tilde{Q}_{ab} = 0$ if $k_a = k_b$ and $\tilde{P}_{ab} = P_{ab}/\min(k_a, k_b)$ otherwise. The NODF is defined as:

$$\mathrm{NODF} = \frac{\sum_{i<j}^{N} \tilde{P}_{ij} + \sum_{a<b}^{M} \tilde{Q}_{ab}}{\frac{N(N-1)}{2} + \frac{M(M-1)}{2}}.$$

2.2.2 Theoretical approach

To theoretically analyze the nested structure of a given bipartite graph, one can construct a grand canonical ensemble for this bipartite graph under the constraint that, for the two types of nodes, the degree sequences in the ensemble match on average the empirical ones[6]. This theoretical approach has two big advantages. First, constraining the ensemble's mean degree sequence to be equivalent to the empirical one limits the possible effects of noisy data, hence possible missing (false negative) or overrated (false positive) links can be dealt with appropriately. Second, for this bipartite graph ensemble one can analytically derive the mean and standard deviation of the distribution of any network property (such as the classical NODF measure of nestedness) that can be analytically formulated in terms of the elements of the bipartite adjacency matrix $\boldsymbol{B}$.

2.3 Nutritional distance measure

We used the weighted Jaccard index to quantify the nutritional distance between food item-i and item-j:

$$d_{ij} = 1 - \frac{\sum_a \min(G_{ia}, G_{ja})}{\sum_a \max(G_{ia}, G_{ja})},$$

where G_{ia} represents the weight gram of nutrient a in food i. $d_{ij} = 0$ indicates that the food item-i and food item-j share exactly the same nutrient constituents; $d_{ij} = 1$ means that they have

totally different nutrient constituents. The nutritional similarity between food item-i and item-j can be defined as

$$s_{ij} = 1 - d_{ij}.$$

3. Nutritional redundancy measure

In the main text, the nutritional redundancy (NR) is defined as: $\mathrm{NR}_\alpha = \mathrm{FD}_\alpha - \mathrm{ND}_\alpha$. FD_α is chosen to be the Gini-Simpson index: $\mathrm{GSI} \equiv 1 - \sum_{i=1}^{N} p_i^2 = \sum_{i=1}^{N}\sum_{j\neq i}^{N} p_i p_j$, representing the probability that two randomly chosen members of a subject's food profile (with replacement) belong to two different food items; and ND_α is chosen to be the Rao's quadratic entropy $Q \equiv \sum_{i=1}^{N}\sum_{j\neq i}^{N} d_{ij} p_i p_j$. Note that with $\mathrm{FD}_\alpha = \mathrm{GSI}$ and $\mathrm{ND}_\alpha = Q$, we have $\mathrm{NR}_\alpha = \sum_{i=1}^{N}\sum_{j\neq i}^{N}(1 - d_{ij}) p_i p_j$, naturally representing the nutritional similarity (or overlap) of two randomly chosen members in any subject's food profile. Here, we emphasize that various taxonomic and functional diversity measurements can be used in NR definition, for example Hill number.

3.1 Food diversity

Consider a subject' food profile of N foods with the relative abundance given by a vector $\boldsymbol{f} = [f_1, \cdots, f_N]$. Hill introduced effective *number of species,* which assumes that the taxonomic diversity (of order q) of a given subject with relative abundance profile $\boldsymbol{f} = [f_1, \cdots, f_N]$ is the same as that of an idealized subject of D equally abundant foods with relative abundance profile $\tilde{\boldsymbol{f}} = [1/D, \cdots, 1/D]$.

$$\sum_{i=1}^{N} f_i^q = \sum_{i=1}^{D} (\frac{1}{D})^q = D^{1-q},$$

This offers a parametric class of food diversity measures defined as follows:

$$\mathrm{FD}_q \coloneqq \left(\sum_{i=1}^{N} f_i^q\right)^{\frac{1}{(1-q)}} \quad \text{for } q \neq 1.$$

And

$$\mathrm{FD}_1 \coloneqq \lim_{q\to 1} \mathrm{FD}_q = \exp\left(-\sum_{i=1}^{N} f_i \log f_i\right).$$

Note that the Gini-Simpson index (GSI) used in the main text is related to FD_2 as follows:

$$\mathrm{GSI} \coloneqq 1 - \sum_{i=1}^{N} f_i^2 = 1 - \frac{1}{\mathrm{FD}_2}.$$

3.2 Nutritional diversity

Consider a subject' food profile of N foods with the relative abundance given by a vector $\boldsymbol{f} = [f_1, \cdots, f_N]$ and pair-wise nutritional distance matrix $\Delta = (d_{ij}) \in \mathbb{R}^{N\times N}$ with $d_{ii} = 0$ for all $i = 1, \cdots, N$ and $d_{ij} = d_{ji} \geq 0$ for all $i \neq j$. Follow our former definition of functional redundancy, we use the new pair-wise nutritional distance matrix to overcome the drawbacks of original distance matrix

$$\Delta' = (d'_{ij}) = \begin{pmatrix} \lambda_1 & d_{12} & d_{13} & \cdots & d_{1N} \\ d_{21} & \lambda_2 & d_{23} & \cdots & d_{2N} \\ d_{31} & d_{32} & \lambda_3 & \cdots & d_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{N1} & d_{N2} & d_{N3} & \cdots & \lambda_N \end{pmatrix},$$

where d_{ij} represents the original nutritional distance between food-i and j and $\lambda_i \coloneqq \frac{\sum_{j\neq i}^{N} d_{ij}}{N-1}$ is the average nutritional distance between food-i and all other foods. Note that, when different species are equally distinct with a constant pairwise distance, λ is equal to this constant. The final nutritional diversity of a subject can be defined as

$$\mathrm{ND}_q(Q') \coloneqq D_q(Q') \cdot Q' = \left(\sum_{i=1}^{N} \sum_{j=1}^{N} \frac{d'_{ij}}{Q'} (f_i f_j)^q \right)^{\frac{1}{2(1-q)}} \cdot Q' \text{ for } q \neq 1,$$

and

$$\mathrm{ND}_1(Q') \coloneqq D_1(Q') \cdot Q' = \exp\left[-\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{d'_{ij}}{Q'} f_i f_j \log(f_i f_j)\right] \cdot Q'$$

3.3 Nutritional redundancy

In the literature of ecology, the functional redundancy (FR) of a sample is often considered to be the part of the taxonomic diversity (TD) that cannot be explained by the functional diversity (FD)[7]. Hence FR is typically defined to be the difference between TD and FD. Similarly, we define the NR to be the difference between food diversity (FD) and nutritional diversity (ND):

$$\mathrm{NR} \coloneqq \mathrm{FD} - \mathrm{ND}.$$

In the main text, we chose FD to be the Gini-Simpson index: $\mathrm{GSI} \equiv 1 - \sum_{i=1}^{N} f_i^2 = \sum_{i=1}^{N} \sum_{j\neq i}^{N} f_i f_j$, and ND is chosen to be the Rao's quadratic entropy $Q \equiv \sum_{i=1}^{N} \sum_{j\neq i}^{N} d_{ij} n_i n_j$. Hence,

$$\mathrm{NR} = \mathrm{GSI} - Q = 1 - \sum_{i=1}^{N} f_i^2 = \sum_{i=1}^{N} \sum_{j\neq i}^{N} (1 - d_{ij}) f_i f_j.$$

For the Hill number-based food diversity TD_q and nutritional diversity $\mathrm{ND}_q(Q')$, and define a parametric class of nutritional redundancy:

$$\mathrm{NR}_q(Q') \coloneqq \mathrm{TD}_q - \mathrm{ND}_q(Q').$$

For $q \neq 1$,

$$\mathrm{NR}_q(Q') \coloneqq \left(\sum_{i=1}^{N} f_i^q\right)^{\frac{1}{(1-q)}} - \left(\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{d'_{ij}}{Q'}\left(f_i f_j\right)^q\right)^{\frac{1}{2(1-q)}} \cdot Q'.$$

For $q = 1$,

$$\mathrm{NR}_1(Q') \coloneqq \exp\left(-\sum_{i=1}^{N} f_i \log f_i\right) - \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{d'_{ij}}{Q'} f_i f_j \log\left(f_i f_j\right)\right] \cdot Q'.$$

4. Healthy diet scores

4.1 HEI-2005

HEI-2005[8] is a score that measures adherence to the USDA 2005 Dietary Guidelines for Americans. The score range is 0 to 100. Each of the 12 components has a minimum score of zero and a maximum score of 5, 10 or 20. These components are: Total vegetables, Dark green & orange vegetables, Total fruit, Whole fruit, Total grains, Whole grains, Milk, Oils, Saturated fat, Sodium and SoFAAs.

4.2 AHEI-2010

AHEI-2010[8] is a score that measures adherence to a diet pattern based on foods and nutrients most predictive of disease risk in the literature. The minimum score = 0, maximum score = 110. Each of the 11 components has a minimum score of 0 and a maximum score of 10, as outlined in the table below. A score between the minimum and maximum is assigned on a continuous basis (except for sodium and alcohol). Those 11 components are: Vegetables, Fruit, Whole grains, Sugar-sweetened beverages and fruit juice, Nuts and legumes, Red meat and processed meat, Trans fat, Long-chain fats, Poly-unsaturated fatty acids, Sodium and Alcohol.

4.3 AMED

The components of AMED[9] are vegetables (excluding potatoes), fruits, nuts, whole grains, legumes, fish, ratio of monounsaturated to saturated fat, red and processed meats, and alcohol. The range of score is 0 to 9. The score criteria are: Intake above the FFQ-specific median intake received 1 point for vegetables, fruits, nuts, whole grains, legumes, fish, and M:S ratio; otherwise, they received 0 points. Red and processed meat consumption below the FFQ-specific median received 1 point; otherwise, 0 points. Alcohol intake between 5 and 15 g/d for women and 10-25 g/d for men received 1 point; otherwise, 0 point.

4.4 DASH

This score was created to capture the characteristics of the Dietary Approaches to Stop Hypertension diet. The DASH[10] Components includes fruits, vegetables (excluding potatoes),

nuts and legumes, low-fat dairy products, whole grains, sodium, sweetened beverages, red and processed meats. The score range is 8 to 40. The DASH Scoring Criteria is each food group first classified into FFQ-specific quintiles. For fruits, vegetables, nuts and legumes, low-fat dairy products, and whole grains, the score for that food group is the quintile ranking. i.e., quintile 1 is assigned 1 point and quintile 5, 5 points. For sodium, red and processed meats, and sweetened beverages, low intake is best. Therefore, the lowest quintile was given a score of 5 points and the highest quintile, 1 point.

5. Healthy aging prediction

We used two standard classifiers: RF (Random Forest, R package 'randomForest'[11]) and XGBoost (Extreme gradient boosting decision trees, R package 'xgboost'[12]) to predict the healthy aging status. The base learners of RF are decision trees. Each tree is a non-linear model constructed with many linear boundaries. A node in a decision tree is associated with a question asking about the data based on the value of a particular feature. XGBoost is a scalable end-to-end decision tree boosting system[12]. Unlike RF that applies the technique of bootstrap aggregating (i.e., bagging) to tree learners, the trees of a boosting system are built sequentially: each tree aims to reduce the error of its previous tree.

For hyperparameter tuning, we used the R package 'caret'[13] (Classification And REgression Training). The number of features randomly sampled as candidates at each split range from 1 to 15 and the number of trees to grow is fixed to 500 for RF. The parameter ranges for XGBoost are the following: (1) Learning rate (eta): 0.001, 0.01; (2) Number of features supplied to a tree (colsample_bytree): 0.4, 0.6, 0.8, 1.0; (3) The depth of the tree (max_depth): 4, 6, 8, 10, 100,000; (4) Maximum number of iterations (nrounds): 100, 1000; (5) Regularization (gamma): 0; (6) Minimum sum of instance weight (min_child_weight): 1; (7) Number of samples supplied to a tree (subsample): 0.5, 0.75, 1; (8) Number of trees to grow (ntree): 500.

To overcome the label imbalance issue, we used the downSample function in caret, which will randomly sample a data set so that all classes have the same frequency as the minority class. To compare the performances of NR and the four healthy diet scores, we split the data into 80% percentage of sample as training set and the remaining 20% samples as test set. For each splitting, we used one of NR, HEI-2005, AHEI-2010, AMED and DASH together with other confounding factors to train the model, then validate the classifier using the test set. We used the error rate, i.e., the proportion of participants that have been incorrectly classified by the model and AUC (area under the ROC curve) to quantify the performance.

6. Supplementary figures and tables

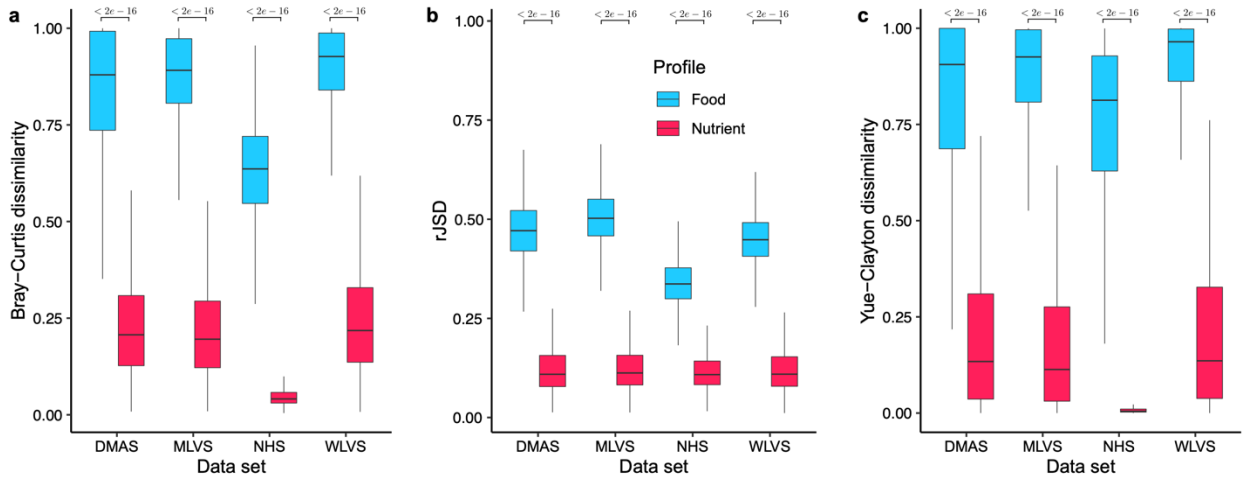

Figure S1: Nutritional profiles are highly conserved across individuals while food profiles are highly personalized. To quantify the variations of food (or nutritional) profiles across different time points and individuals, we calculated the beta diversity using Bray-Curtis dissimilarity (**a**), rJSD (rooted Jensen-Shannon divergence, **b**), and Yue-Clayton distance (**c**) between the food (or nutritional) profiles of different individuals and different time points. The Bray-Curtis dissimilarity between a pair of individuals, j and k is defined as: $BC_{jk} \equiv \frac{\sum_i |X_{ij}-X_{ik}|}{\sum_i (X_{ij}+X_{ik})}$. The rJSD dissimilarity is defined as: $D_{\mathrm{rJSD}}(j,k) \equiv [\frac{D_{\mathrm{KL}}(j,m)+D_{\mathrm{KL}}(k,m)}{2}]^{1/2}$, in which $m \equiv \frac{j+m}{2}$ and $D_{\mathrm{KL}}(j,k) \equiv \sum_{i \in S} X_{ij} \log \frac{X_{ij}}{X_{ik}}$ is the Kullback-Leibler divergence between j and k. The Yue-Clayton dissimilarity is defined as: $YC_{jk} \equiv \frac{\sum_i X_{ij} X_{ik}}{\sum_i \sum_i (X_{ij}-X_{ik})^2 + \sum_i (X_{ij} * X_{ik})}$. In all dissimilarity definitions, X_{ij} represents the relative abundance of food/nutrient i in individual j.

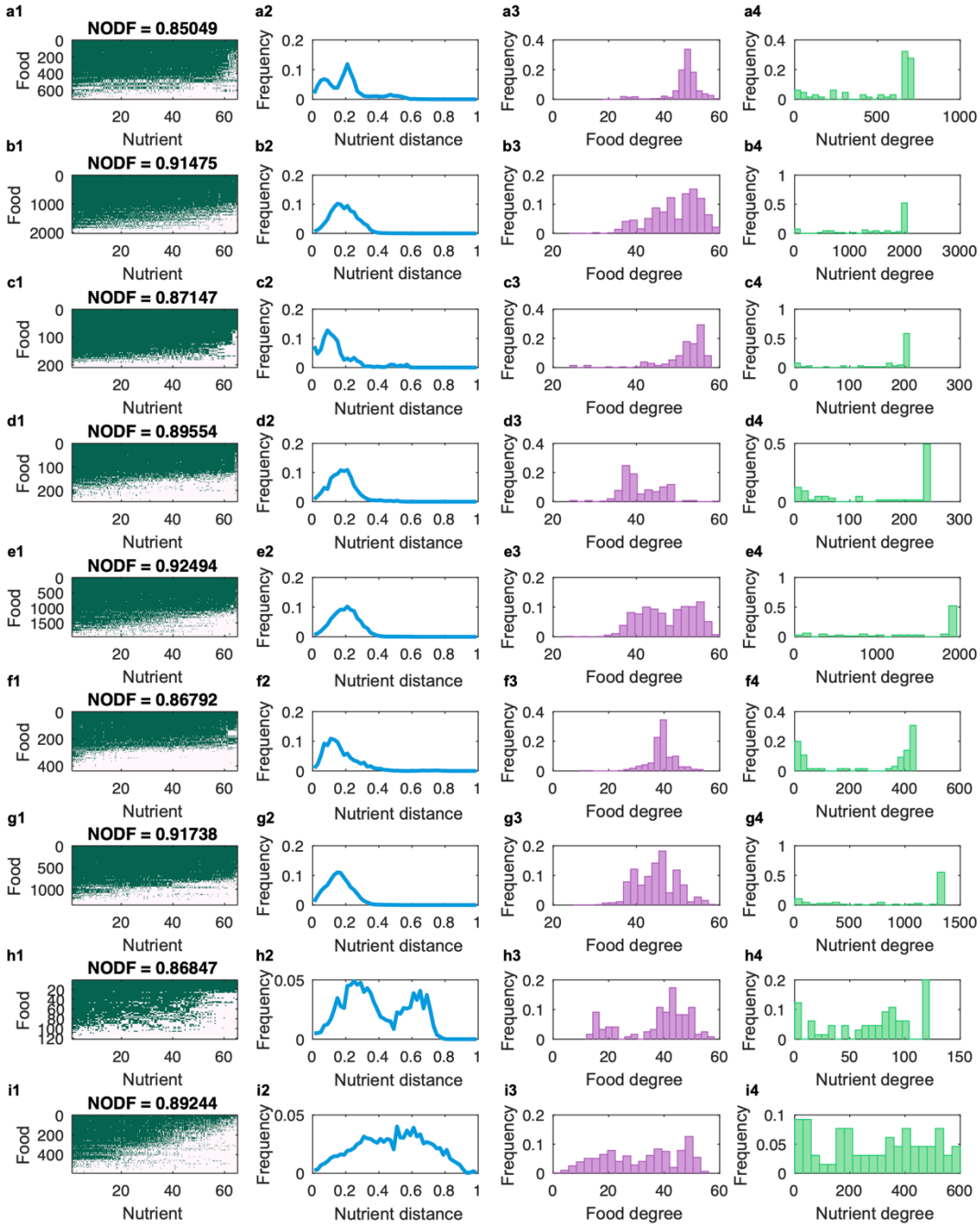

Figure S2: The food-specific subgraphs of the FNN display similar topological feature as the original FNN. Columns: (1) the incident matrix; **(2)** distribution of nutrient distance; **(3)** distribution of food degree; **(4)** distribution of nutrient degree. **Rows:** (**a**) Milk and Milk Products; (**b**) Meat_Poultry, Fish, and Mixtures; (**c**) Eggs; (**d**) Dry Beans, Peas, Other Legumes, Nuts, and Seeds; (**e**) Grain Products; (**f**) fruits; (**g**) Vegetables; (**h**) Fats, Oils, and Salad Dressings; (**i**) Sugars, Sweets, and Beverages.

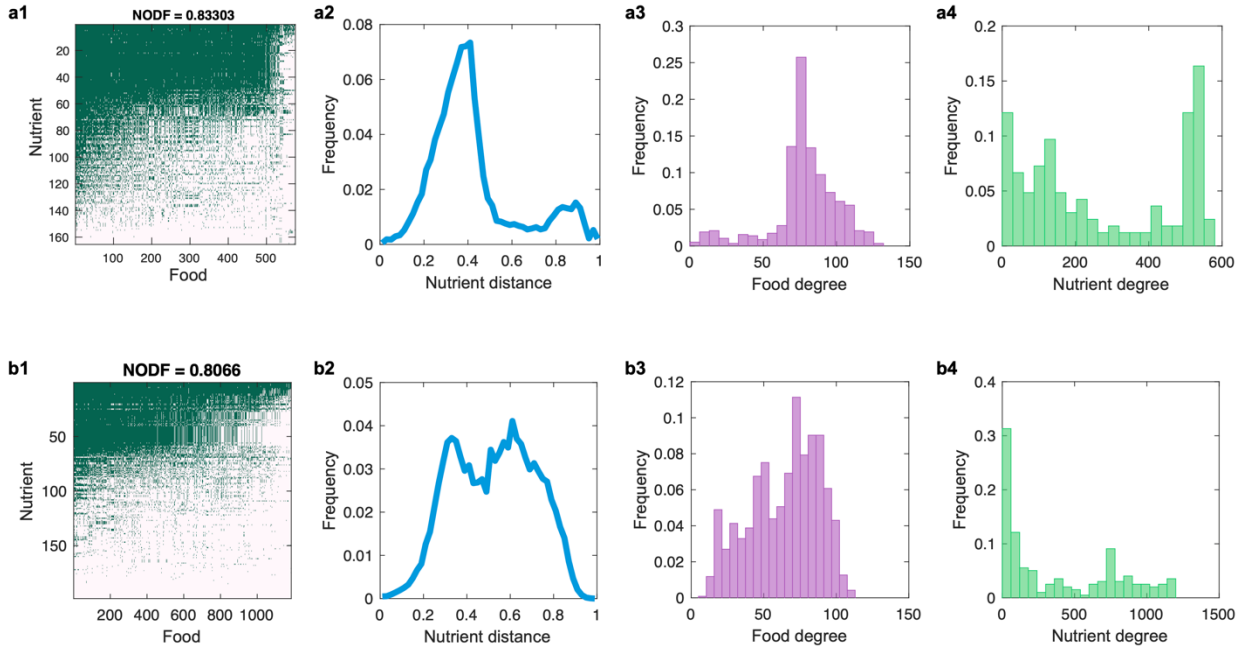

Figure S3: Food-Nutrient networks (FNN) constructed from the Harvard food composition database (a) and the Frida database (b). Columns: (**1**) the incident matrix of FNN, where the presence (or absence) of a link between a food and a nutrient is colored in green (or pink), respectively. (**2**) The distribution of nutrient distances (d_{ij}) among different foods. The bin size is 0. (**3**) The distribution of food degrees. Here, the degree of a food item is the number of distinct nutrients it can provide. (**4**) The distribution of nutrient degrees. Here, the degree of a nutrient is the number of food items that contain this nutrient.

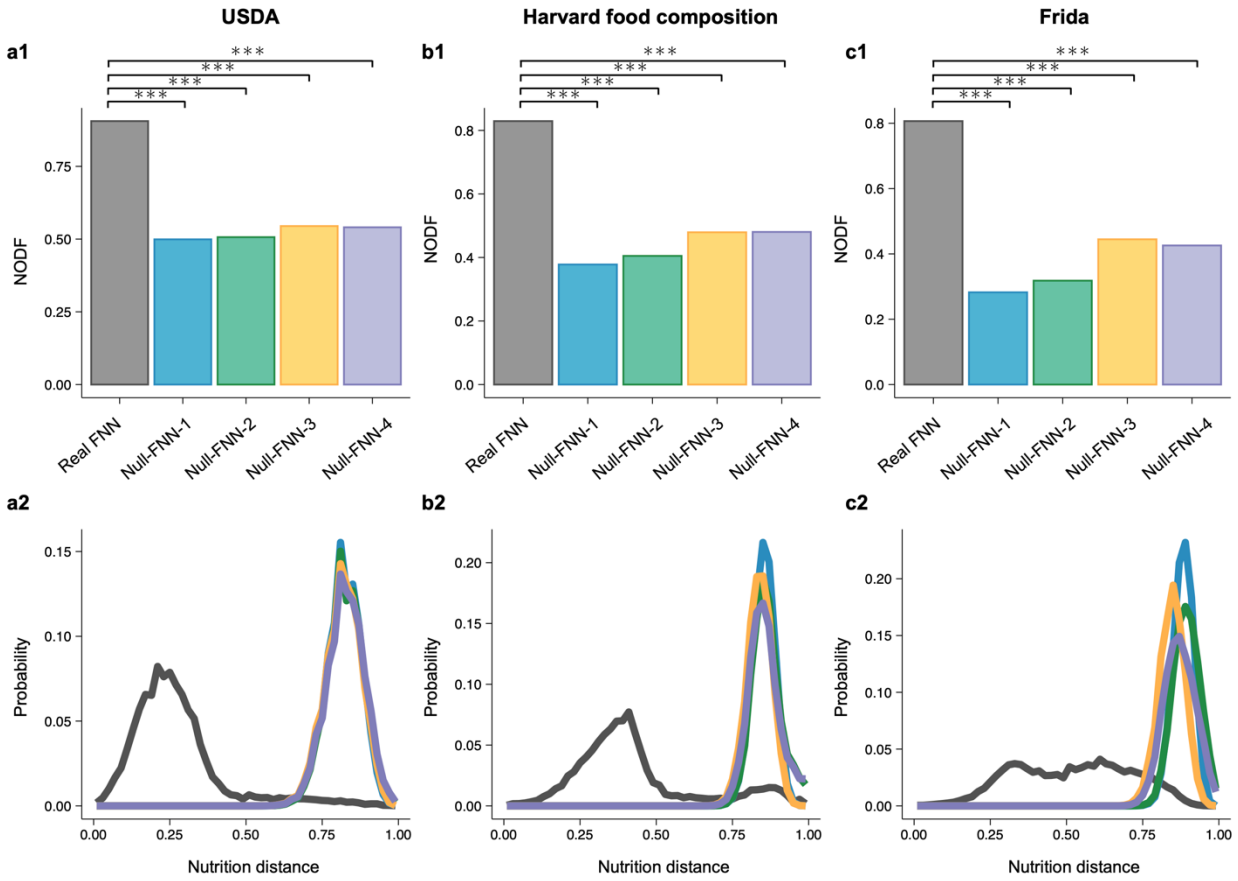

Figure S4: Comparison of structural properties between the real FNN and different randomized FNNs. **Columns**: FNN constructed from the USDA database **(a)**, the Harvard food composition database **(b)**, and the Frida database **(c). Rows: (1)** The nestedness based on the NODF measure of the real FNN (gray bar), as well as the randomized FNNs (colored bars) using four different FNN randomization schemes: Null-FNN-1, complete randomization; Null-FNN-2, Food-degree preserving randomization; Null-FNN-3, Nutrient-degree preserving randomization; Null-FNN-4, Food- and nutrient-degree preserving randomization. For each randomization scheme, 50 realizations were generated. **(2)** The distribution of nutrient distances (d_{ij}) between different foods calculated from the real FNN (gray lines) and the randomized FNNs (colored lines) using the same randomization schemes as in row (1). We generated 50 realizations for each randomization scheme, and the bin size is 0.02.

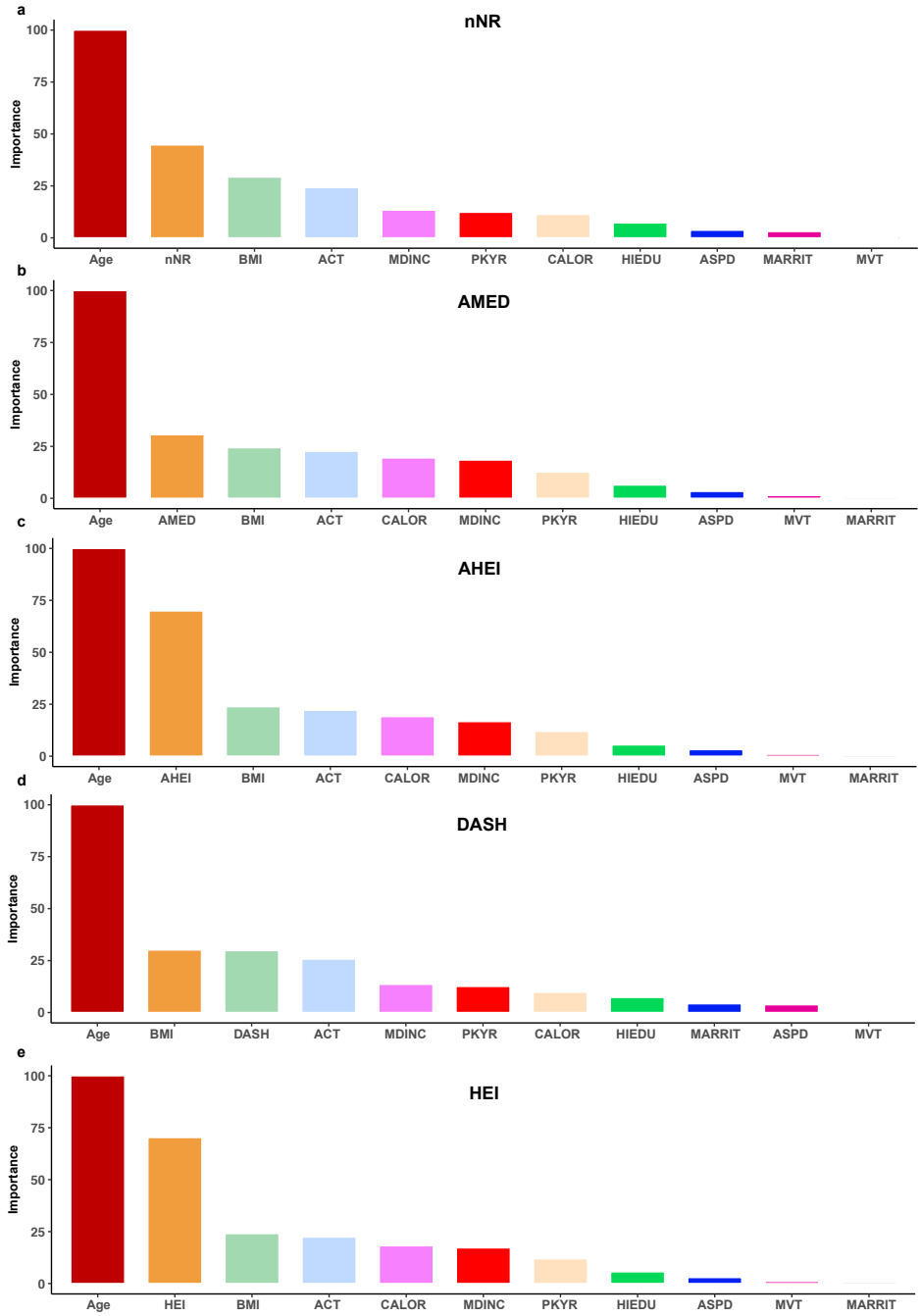

Figure S5: Importance ranking of features quantified by their mean decreasing of accuracy. Here, we used the following features: Age, BMI (body mass index), ACT (physical activity), MDINC (median income from census tract), CALOR (energy intake), PKYR (pack-years of smoking), HIEDU (education), ASPD (aspirin use), MVT (multivitamin use), MARRIT (marital status), together with (**a**) nNR (normalized nutritional redundancy (nNR) or one of the healthy diet scores: (**b**) AMED; (**c**) AHEI; (**d**) DASH; and (**e**) HEI. The results were averaged over 100 different training/test splits. The error bars are almost invisible in those figures.

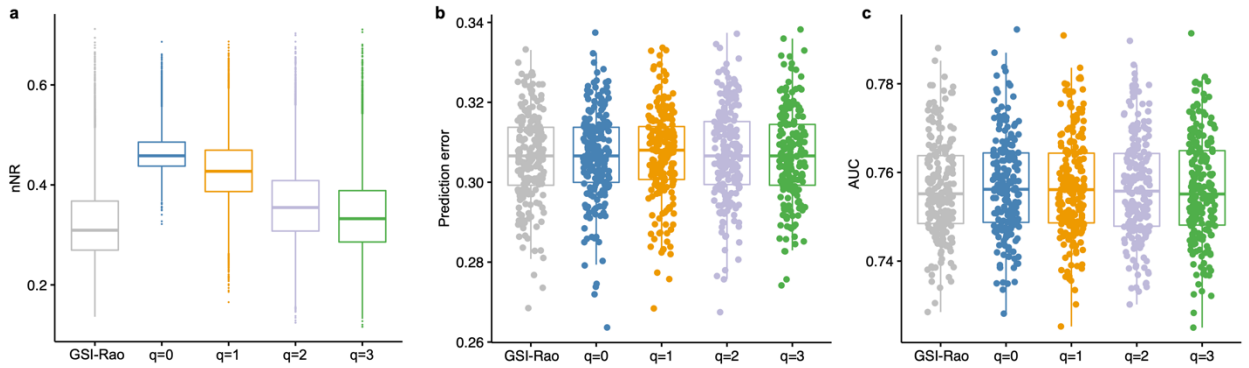

Fig. S6: Nutritional redundancy is robust in aging prediction with different Hill numbers. a: nNR (nNR ≡ NR/FD) of NHS calculated with different definitions. GSI-Rao represents the NR calculated from Gini-Simpson index and Rao's quadratic entropy used in the main text. q = 0,1,2,3 represent the Hill-number based NR (see SI Sec.3.3). **b:** Error rate of the RF classifier in predicting the healthy aging status. **c:** AUC of the RF classifier in predicting the healthy aging status. The participants are randomly spitted into 80% as the training set and the remaining 20% as the test set. The boxplot represents the performances of 200 independent splitting.

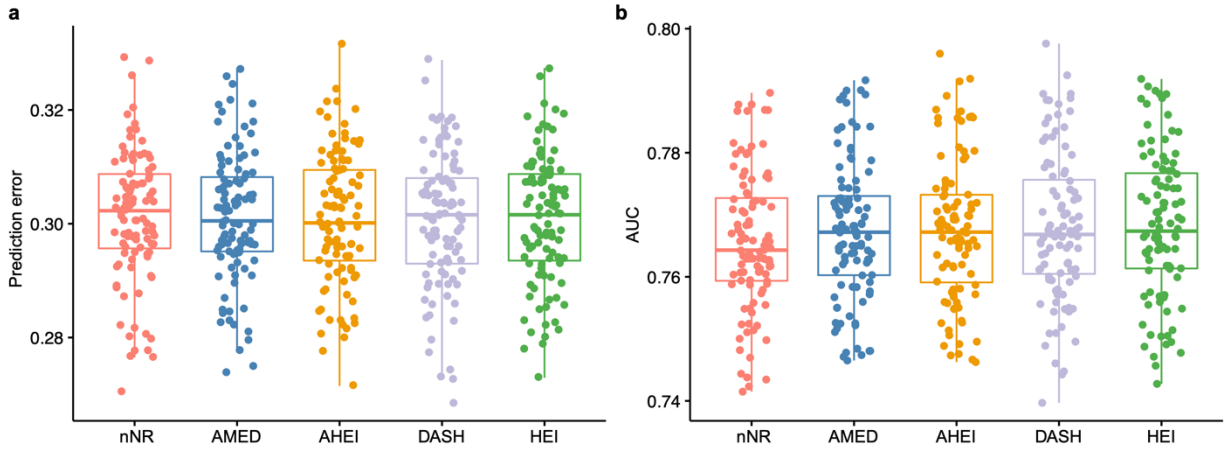

Fig. S7: Nutritional redundancy is robust in predicting healthy aging with a different classifier. **a:** Error rate of the XGBoost classifier in predicting the healthy aging status. **b:** AUC of XGBoost in predicting the healthy aging status. The participants are randomly spitted into 80% as the training set and the remaining 20% as the test set. The boxplot represents the performances of 100 independent splitting.

Characteristics	Healthy agers ($n = 3,491$)	Usual agers ($n = 17,808$)
Mean age, years	58.1 (4.9)	62.8 (6.5)
Educational level (1992), % (*n*)		
Registered nurse	57 (1,999)	66 (11,841)
Bachelor's degree	25 (877)	21 (3,776)
Master or doctorate	18 (565)	13 (1,921)
Marital status (1996), % (*n*)		
Windowed	5 (182)	10 (1,792)
Married	86 (3,018)	82 (14,601)
Separated/divorced	7 (267)	8 (1,315)
Median neighborhood income, $		
<45302	16 (564)	21 (3,659)
45,302-55,373	17 (606)	20 (3,641)
55,373-66,363	20 (692)	20 (3,553)
66,364-82582	21 (746)	20 (3,506)
>82582	26 (873)	19 (3,414)
Body mass index, % (*n*)		
<22	26 (919)	16 (2,886)
22-24.9	35 (1,220)	27 (4,856)
25-29.9	30 (1,062)	35 (6,212)
≥ 30	9 (271)	22 (3,668)
Multivitamin use, % (*n*)	60 (2,108)	61 (10.921)
Aspirin use, tables per week, % (*n*)		
<1	67 (2,323)	60 (10,740)
1-2	8 (260)	7 (1,217)
>2	25 (869)	33 (5,656)
Pack-years of smoking		
0	54 (1,881)	48 (8,504)
0-4	13 (464)	11 (1,968)
5-20	20 (701)	20 (3,590)
21-112	12 (406)	20 (3,522)
Mean physical activity		
<3.6	11 (367)	21 (3,651)
3.7-9	16 (565)	21 (3,660)
9.1-16.7	21 (721)	21 (3,668)
16.8-30.4	24 (848)	20 (3,578)
>30.5	28 (986)	17 (3,214)
Energy intake		
<1,296	17 (601)	29 (3,318)
1,296-1,572	21 (737)	21 (3,683)
1,573-1,830	21 (716)	21 (3,754)
1,831-2,169	22 (748)	21 (3,724)
>2,170	19 (689)	18 (3,329)
Nutritional redundancy	0.331 (0.0791)	0.325 (0.0798)
AMED	4.814 (1.89)	4.48 (1.86)
AHEI	55.5 (11)	53.34 (10.77)
DASH	24.58 (4.49)	23.86 (4.48)
HEI	74.9 (9.16)	73.08 (10)

Table S1. Characteristics of a subset of NHS participants in 1998 for whom we have their aging status in 2012 from a previous study. Measures in this table were calculated among non-missing values (≤ 5 of data were missing). Data expressed as mean (SD) or percentage.

References

1. Patterson, B. D. The principle of nested subsets and its implications for biological conservation. *Conservation Biology* **1**, 323–334 (1987).
2. Patterson, B. D. & Atmar, W. Nested subsets and the structure of insular mammalian faunas and archipelagos. *Biological Journal of the Linnean Society* **28**, 65–82 (1986).
3. Brualdi, R. A. & Sanderson, J. G. Nested species subsets, gaps, and discrepancy. *Oecologia* **119**, 256–264 (1999).
4. Bascompte, J., Jordano, P., Melián, C. J. & Olesen, J. M. The nested assembly of plant–animal mutualistic networks. *Proceedings of the National Academy of Sciences* **100**, 9383–9387 (2003).
5. Rezende, E. L., Lavabre, J. E., Guimarães, P. R., Jordano, P. & Bascompte, J. Non-random coextinctions in phylogenetically structured mutualistic networks. *Nature* **448**, 925–928 (2007).
6. Payrató-Borràs, C., Hernández, L. & Moreno, Y. Breaking the Spell of Nestedness: The Entropic Origin of Nestedness in Mutualistic Systems. *Physical. Review X* **9**, 031024 (2019).
7. Pillar, V. D., Blanco, C.C., Müller, S.C., Sosinski, E.E., Joner, F. & Duarte, L.D. Functional redundancy and stability in plant communities. *Journal of Vegetation Science* 24, 963–974 (2013).
8. Chiuve, S. E., Fung, T.T., Rimm, E.B., Hu, F.B., McCullough, M.L., Wang, M., Stampfer, M.J. and Willett, W.C. Alternative dietary indices both strongly predict risk of chronic disease. The Journal of Nutrition 142, 1009–1018 (2012).
9. Fung, T. T., Rexrode, K.M., Mantzoros, C.S., Manson, J.E., Willett, W.C. & Hu, F.B. Mediterranean diet and incidence and mortality of coronary heart disease and stroke in women. Circulation 119, 1093 (2009).
10. Fung, T. T., Chiuve, S.E., McCullough, M.L., Rexrode, K.M., Logroscino, G. & Hu, F.B. Adherence to a DASH-style diet and risk of coronary heart disease and stroke in women. Archives of Internal Medicine 168, 713–720 (2008).
11. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2**, 18–22 (2002).
12. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (2016) doi:10.1145/2939672.2939785.
13. Kuhn, M. Building predictive models in R using the caret package. *Journal of Statistical Software* **28**, 1–26 (2008).