Machine Learning Prediction of Food Processing

Giulia Menichetti[1,2], Babak Ravandi[1], Dariush Mozaffarian[4], Albert-László Barabási[1,2,3†]

[1]Network Science Institute and Department of Physics, Northeastern University, Boston, USA
[2]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, USA
[3]Department of Network and Data Science, Central European University, Budapest, Hungary
[4]Tufts Friedman School of Nutrition Science & Policy, Boston, USA

Abstract

Despite the accumulating evidence that increased consumption of ultra-processed food has adverse health implications, it remains difficult to decide what constitutes processed food. Indeed, the current processing-based classification of food has limited coverage and does not differentiate between degrees of processing, hindering consumer choices and slowing research on the health implications of processed food. Here we introduce a machine learning algorithm that accurately predicts the degree of processing for any food, indicating that over 73% of the U.S. food supply is ultra-processed. We show that the increased reliance of an individual's diet on ultra-processed food correlates with higher risk of metabolic syndrome, diabetes, angina, elevated blood pressure and biological age, and reduces the bio-availability of vitamins. Finally, we find that replacing foods with less processed alternatives can significantly reduce the health implications of ultra-processed food, suggesting that access to information on the degree of processing, currently unavailable to consumers, could improve population health.

[†]Corresponding author. e-mail: a.barabasi@northeastern.edu

Introduction

Unhealthy diet is a major risk factor for multiple non-communicable diseases, from obesity and type 2 diabetes to coronary heart disease (CHD) and cancer, together accounting for 70% of mortality and 58% of morbidity worldwide [1,2]. Traditionally, consumers rely on food category-based dietary recommendations like the Food Pyramid (1992 [3]) or MyPlate (2011 [4]), which define the mix of fruits, vegetables, grains, dairy and protein-based foods that constitute a healthy diet. In recent years, however, an increasing number of research studies and dietary guidelines have identified the important role and separate health effects of food processing [5–10]. Observational studies, meta-analysis, and controlled metabolic studies suggest that dietary patterns relying on unprocessed foods are more protective than the processing-heavy Western diet against disease risk [11, 12]. The role of processed food has reached food policy and is now embodied in several expertise-based food classification systems used in cohort studies such as EPIC [13, 14], and led to the expansion of food description and ontology systems such as LanguaL [15], FoodEx2 [16–19], and FoodOn [20, 21].

NOVA [22–24] is a classification system for food processing, widely used in epidemiological studies. It categorizes individual foods into four broad categories: *unprocessed or minimally processed* (NOVA 1), like fresh, dry or frozen fruits or vegetables, grains, legumes, meat, fish and milk; *processed culinary ingredients* (NOVA 2), like table sugars, oils, fats and salt; *processed foods* (NOVA 3), like canned food, simple bread and cheese; and *ultra-processed products* (NOVA 4), industrial formulations typically of five or more ingredients including substances not commonly used in culinary preparations, such as additives whose purpose is to imitate sensory qualities of fresh food. Examples of ultra-processed foods include packaged breads, cookies, sweetened breakfast cereals, margarines, sauces and spreads, carbonated drinks, hot dogs, hamburgers and pizzas.

Epidemiological studies have documented significant associations between greater consumption of the highest NOVA category of processing (NOVA 4, ultra-processed food) and diseases onset [25–31], including links to obesity [32], CHD [30, 33], diabetes mellitus [34, 35], cancer [29, 36], and depression [28]. Finally, a randomized controlled metabolic trial in 20 adults has confirmed the short-term adverse effects on caloric intake and weight gain of ultra-processed foods [31].

Despite its success, NOVA is purely descriptive in nature, leading to a laborious and incomplete classification task, limiting research into the impact of processed food [37–39]. First, NOVA relies on expertise-based manual evaluation of each food, and assigns a unique class to only 35% of the foods catalogued by the U.S. Department of Agriculture (USDA), decomposing

the remaining part of the database in ingredients to be further analyzed (SI Section 1). Second, the classification of composite recipes, products, and mixed meals including different types of processed items – a large proportion of the food supply – is not straightforward. Whilst, an accurate ingredient list of all complex foods would allow for an ingredient-based classification, a full and quantified ingredient list is seldom, if ever, available to the consumer, and when available, it shows considerable variability from database to database. Interestingly, even in presence of detailed ingredient information, the consistency across nutrition specialists in assigning NOVA classes was found to be low [39]. Given these data limitations, current approaches have classified all foods with at least one ultra-processed ingredient as ultra-processed [29, 40], reducing the discrimination of such foods in relation to health outcomes [41, 42]. Third, all the observed risk for the NOVA classification is in the NOVA 4 class, representing a large and heterogeneous category of ultra-processed food that limits our ability to investigate the health implications of different gradations of processing. Indeed, according to NOVA, many nations derive up to 60% or more of their average caloric intake from ultra-processed foods [42–44]. While NOVA allows for a more refined analysis, all foods within this class are considered to have identical health consequences. These perceived homogeneity of NOVA 4 foods limits both scientific research and practical consumer guidance on the health effects of differing degrees of processing. It also reduces the industry's incentives to reformulate foods towards less processed offerings, shifting investments from the ultra-processed NOVA 4 foods to the less processed NOVA 1 and NOVA 3 categories. To overcome these limitations, here we introduce FoodProX, a machine learning classifier that takes as input nutritional measures, and is trained to predict the degree of processing of any food in a reproducible, portable, and scalable fashion. We rely on nutrients as input because: 1) The list of nutrients in a food is consistently regulated and reported worldwide; 2) Their quantities in unprocessed food are constrained by physiological ranges determined by biochemistry [45]; 3) Food processing systematically and reproducibly alters nutrient concentrations through combinatorial changes detectable by machine learning (Figures 1A-B). FoodProX allows us to define a continuous index (FPro) that captures the degree of processing of any food, and helps us quantify the overall diet quality of individuals, ultimately unveiling the statistical correlations between the degree of processing characterizing individual diets and multiple disease phenotypes.

Results

The manual procedure behind NOVA, relying on the ingredients of food, has allowed a straightforward labeling to 2,484 foods reported in the National Health and Nutrition Examina-

tion Survey (NHANES) 2009-2010, representing 34.25% of items consumed by a representative sample of the U.S. population [26] (SI Section 1). The remaining foods are either not classified, or require further decomposition into their differing food ingredients, often not reported by the manufacturers. In contrast, the basic nutrient profile of foods and beverages is always disclosed, mandated by law in most countries. For example, the USDA Standard Reference database (USDA SR Legacy), catalogues the nutrient profile of 7,793 foods with resolutions ranging from 8 to 138 nutrients (Figure S1) [46], and USDA FNDDS reports between 65 to 102 nutrients for all foods consumed by NHANES participants [47, 48].

Our work relies on the hypothesis that the nutrient profiles of unprocessed or minimally processed foods are generally constrained within common physiologic ranges [45]. The nutrient profile can be widely altered by the physical, biological and chemical processes involved in food preparation and conservation, thus correlating with the degree of processing undertaken. Indeed, among the nutrients reported in raw onion, 3/4 change their concentration in excess of 10% in the fried and battered version of the product, and more than half by 10-fold (Figure 1B). We lack however, a single nutrient "biomarker" that accurately tracks the degree of processing; instead we observe changes in the concentration of multiple nutrients, whose combinations jointly correlate with processing. This complexity advocates for the use of machine learning, designed to capture the combinatorial explosion of nutrient alterations.

FoodProX Algorithm

To train FoodProX, a multi-class random forest classifier, we leveraged the nutrient concentrations provided in FNDDS 2009-2010 for the foods classified in Ref. [26] (Figure 1D). FoodProX takes as input the list of nutrients in any food and offers as output a vector of four probabilities $\{p_i\}$, representing the likelihood that the respective food is classified as unprocessed (p_1, NOVA 1), processed culinary ingredients (p_2, NOVA 2), processed (p_3, NOVA 3), and ultra-processed (p_4, NOVA 4). The highest of the four probabilities determines the final classification label for each food (Figure 1C-D).

As 90% of foods in USDA Branded Food Products Database report less than 17 nutrients, we also tested the algorithm's predictive power with the 12 gram-based nutrients mandated by FDA [49] (SI Section 1). To evaluate the performance of FoodProX we measure the Area Under the Receiver Operating Characteristic (AUC), defined as the probability that a random sample from the class of interest will have a higher score than a random sample from any other class. We

identified consistently high AUC values for all the considered nutrient subsets: 0.9804±0.0012 for NOVA 1, 0.9632±0.0024 for NOVA 2, 0.9696±0.0018 for NOVA 3, and 0.9789±0.0015 for NOVA 4, significantly far from a random performance with AUC=0.5, describing a model with no discriminating power (see SI Section 2.1, Figure S5, and Table S6 for a detailed analysis of the cross-validated performances, including precision and recall).These results confirm that changes in nutrient content have remarkable predictive power in capturing the extent of food processing. Furthermore, we find that no single nutrient drives the predictions, but the predictive signal is rooted in combinations of changes spanning multiple nutrients (see SI Section 2.2 for a detailed permutation feature importance and Shapley value analysis [50]).

We visualize the decision space of the classifier by performing a principal component analysis over the probabilities $\{p_i\}$, observing that the list of foods manually classified by NOVA is limited to the three corners of the phase space, to which the classifier assigns dominating probabilities (Figure 2A). We used FoodProX to classify all foods and beverages that lacked prior manual NOVA classification in FNDDS (65.75% of total). We found that 7.39% of FNDDS consists of NOVA 1; 0.90%, NOVA 2; 18.36%, NOVA 3; and 73.35%, NOVA 4 foods (Figure 2B). Yet, many previously unclassified foods are often inside the phase space, indicating that they lack a dominating probability, hence the assignment of a single NOVA class is somewhat arbitrary (Figure 2B). The detection of this ambiguity is a strength of FoodProX, reflecting the observation that a four-class classification encodes a great extent of nutrient variability associated with different food processing methods and intensities [51]. For example, while the classifier confidently assigns "Raw Onion" to NOVA 1 ($p_1 = 0.9651$), and "Onion rings prepared from frozen" ($p_4 = 0.9921$) to NOVA 4, it accurately offers an intermediate confidence for "Onion, Sauteed," placing it with probability $p_4 = 0.6521$ in NOVA 4, and with probability $p_3 = 0.2488$ in NOVA 3.

Food Processing Score (FPro)

The observation that enforcing discrete classes causes inherent challenges in food classification prompted us to introduce the Food Processing Score ($FPro$), a continuous variable with $FPro = 0$ for raw ingredients, and $FPro \to 1$ for ultra-processed foods. We define the $FPro$ of a food k as

$$FPro_k = \frac{1 - p_1^k + p_4^k}{2}, \tag{1}$$

capturing the progressive changes in nutrient content induced by processing. In other words,

the FPro score measures the trade-off between the confidence that the FProX algorithm has in classifying the food item k as NOVA 1 (p_1^k) or NOVA 4 (p_4^k). For example, $FPro$ incrementally increases from raw onion ($FPro = 0.0203$) to boiled onion ($FPro = 0.3126$), fried onion ($FPro = 0.7779$), and onion rings from frozen ingredients ($FPro = 0.9955$, Figure 2C). $FPro$ allows us to unveil the degree of processing characterizing different food preparation techniques, assigning lower values to foods made from fresh ingredients than those made from frozen ingredients (Figure 2C). $FPro$ also permits classification of complex recipes and mixed dishes, identifying the overall average processing of the item or meal.

As noted above, nearly 20% of the U.S. diet comprises foods in NOVA 3, and over 70%, in NOVA 4. The observed variability of $FPro$ for foods that belong in the same NOVA class prompted us to analyze the variations in degree of processing within subgroups of foods and beverages, as captured by What We Eat in America (WWEIA) categorization [52]. We find that different food products within the same NOVA classification and WWEIA categories show remarkable variability in $FPro$, confirming the presence of different processing fingerprints within each category (Figure 2D). For example, if we select four products from the same brand (Post) of breakfast cereals that are all manually classified as NOVA 4 (i.e., considered identically processed), we observe that $FPro$ differentiate them and captures progressive alterations in fiber content, fortification with vitamins and minerals, and addition of sugar and fats (Figure 2D, insert, and Section S2.5).

Stability and Robustness of FPro

Real world foods vary in nutritional content, affected by recipe variations, production methods, soil quality, storage time, and changing government regulations [53]. This degree of variability, coupled by measurement and reporting uncertainties, raises a fundamental question: is FPro robust against the expected variability and uncertainty in nutrient content? To address this, we first explored how nutritional values for the same food change through different FNDDS cycles, focusing on the 5,632 foods whose nutrient profile is reported both in FNDDS 2009-2010 and 2015-2016 (SI Section 7). For example, in the less processed category, for "Milk, calcium fortified, cow's, fluid, whole" we find that 14 nutrients have changed between 2009 and 2015, including a 3.65 fold decrease in Calcium content. In a similar fashion, the highly ultra-processed "Cookie, vanilla with caramel, coconut, and chocolate coating" shows variation in 46 nutrients, with 6 fold decrease in monounsaturated fatty acid 20:1. Despite such significant changes in the content of some nutrients, FPro shows remarkable stability: milk's FPro goes from 0.0010 to 0.0011, consistently classifying it as unprocessed, and for the cookie FPro goes from 0.9943

to 0.9965, staying firmly in the ultra-processed category. Overall, for foods whose nutrients maximally change between 10% and 50% of their original value, we observe an absolute shift in FPro of 0.001556 (quartiles $Q_1(25\%)$=0.000222, $Q_3(75\%)$= 0.004764). Allowing up to 1000% of nutrient variability does not significantly alter our findings, since the expected change of FPro per food reaches 0.003312 (quartiles $Q_1(25\%)$= 0.000722, $Q_3(75\%)$= 0.011310). The observed FPro stability is rooted in the fact that FPro's value is driven by the nutrient panel as a whole, and not by the concentration of any single nutrient (see SI Sections S7, for further details on data sampling and variability).

Currently, the chemical information available to train FPro does not track the concentration of additives as they are rarely available for most foods. Yet, additives like tertiary butylhydroquinone, acetylated monoglycerides, polysorbates, sodium stearoyl lactylate and sodium aluminum phosphate, represent obvious signatures of food processing [54], raising the question on how much improvement in predictive power we could obtain if information on additives would be available. We therefore relied on the Open Food Facts, that compiles an extensive list of food additives, including artificial colors, artificial flavors, and emulsifiers [55], to test the FPro's ability to absorb information on additives. From the Open Food Facts website we collected 233,831 nutritional records, annotated with NOVA labels according to a heuristic described in [56]. We then trained and validated two models: (1) FProX leveraging only nutrition facts as introduced before, (2) a modified FProX, using nutrition facts and the available information on the number of additives (SI Section 6). While model (2) displays slightly better performance, with AUC 0.9926±0.0003 for NOVA 1, 0.9878±0.0047 for NOVA 2, 0.9653±0.0010 for NOVA 3, 0.9782±0.0007 for NOVA 4, we find that the performance of model (1) is largely indistinguishable, reaching AUC 0.9880±0.0006 for NOVA 1, 0.9860±0.0045 for NOVA 2, 0.9320±0.0015 for NOVA 3, 0.9508±0.0009 for NOVA 4. In other words, while information on additives can improve FProX's performance, changes in the nutrient profile already carry the bulk of the predictive power. This aspect is further confirmed by the performance of a classifier purely based on the number of additives: in this scenario the number of false positives for NOVA 1, 2, and 3 remarkably increases, affecting the precision of the model, and showing predictive power only for NOVA 4 (see SI Section 6 for a detailed comparison of the models).

Finally, we investigate how FPro, a measure of the nutritional quality of food, changes depending on where the food was prepared, exploring if it can distinguish between home-cooked food, food prepared in stores, canteens, restaurants, fast foods, and products available in vending machines. Leveraging data from NHANES, we ranked the 10 most popular food

sources in increasing value of FPro. The lowest FPro is observed for "Grown or caught by you and someone you know" (median FPro=0.4423) and "Residential dining facility" (median FPro=0.6238), confirming the less processed nature of home prepared foods. We obtain a much higher FPro for "Restaurant fast food/pizza" (median FPro=0.9060) and "Vending machine" (median FPro=0.9800), confirming its reliance on ultra-processed ingredients. This result indicates that FPro can differentiate food sources, styles and quality of food preparation (see Section S2.6 for the statistical analysis).

Individual Processing Score

The significant contribution of ultra-processed food to American dietary intake, and $FPro$'s ability to demonstrate heterogeneity in the extent of food processing within the broad category of ultra-processed food, prompts us to assess the contribution of processed food to the diet of each individual, jointly weighted by both the extent of processing and the contribution to caloric intake. This is provided by the individual Food Processing Score ($iFPro$),

$$iFPro^{j}_{WC} = \sum_{k}^{D_j} \frac{c^j_k}{C^j} FPro_k, \tag{2}$$

which varies between 0 and 1, where D_j is the number of dishes consumed by individual j, C^j is the daily total amount of consumed calories, and c^j_k is the amount of calories contributed by each food item. A gram-based $iFPro_{WG}$, captures the fraction of grams in a diet supplied by processed food (Eq. S6).

We calculated $iFPro$ for 20,047 individuals with dietary records in a representative U.S. national sample from NHANES 1999-2006. NHANES relies on 24-hour recalls to capture dietary intake, a widely used methodology in epidemiology, with well-explored statistical characteristics [57]. Indeed, studies collecting multiple days of intake data per individual indicate that it is statistically more efficient to increase the number of individuals in the sample than increase the number of days beyond two per individual [58]. As Figure 3C shows, the median $iFPro_{WC}$ for the American population is 0.7872, confirming a high reliance on intake on ultra-processed food [42–44]. More importantly, $iFPro$ allows us to identify differential reliance on processed food. Consider individuals $\mathcal{A}$ and $\mathcal{B}$, men of similar age (47 vs. 48 years old, Figure 3A-C), with similar number of reported daily unique dishes (12.5 vs. 13 dishes) and comparable caloric intake (2,016 vs. 1,894 kcal). Yet, the diet of individual $\mathcal{A}$ has $iFPro_{WC} = 0.3981$, as it relies on unprocessed ingredients and home cooking. In contrast, $\mathcal{B}$ has $iFPro_{WC} = 0.9572$, given

his consumption of hot dogs, hamburgers, French fries, pizza, and Kit Kat (see SI Section 3.5 for a full analysis of the correlation between $iFPro_{WC}$ and food group consumption in the US population). As we show next, these differences correlate with different health outcomes.

Health Implications

To quantify the degree to which the consumption of ultra-processed foods, graded across their entire range, correlate with health outcomes, we investigated exposures and phenotypes provided in NHANES 1999-2006 [59]. Inspired by [60], we performed an Environment-Wide Association Study (EWAS), an alternative to single-exposure epidemiological studies, to identify the most significant environmental factors associated with different health phenotypes in a non hypothesis-driven fashion. We measured for each variable the association with the diet processing score $iFPro$, adjusting for age, sex, ethnicity, socio-economic status, BMI and caloric intake (Figure S17). After False Discovery Rate (FDR) correction for multiple testing, 209 variables survive (SI Section 4), documenting how the consumption of ultra-processed foods relates to health.

Metabolic and Cardiovascular Risk: We find that individuals with a high food processing score show positive associations with risk of metabolic syndrome (MetS), diabetes (fasting glucose), and also (consistent with poor diets clustering in families) a family history of heart attack or angina, in line with earlier findings (SI Section 4) [60–62]. The increased risk of cardiovascular disease is further confirmed by the significant positive association of $iFPro$ with both Framingham and ACC/AHA Risk Scores (Figure 3D) [63, 64].

Overall, individuals with a higher food processing score exhibit higher blood pressure, trunk fat and subscapular skinfold, measures of obesity, and blood insulin and triglyceride levels; and lower "good" HDL cholesterol. Further novel findings indicate a higher prevalence of type 2 diabetes (C-peptide), inflammation (C-Reactive Protein), vitamin deficiency (homocysteine, methylmalonic acid), and inflammatory arthritis [65–70]. We also find an inverse association between $iFPro$ and telomere length, which can be affected by diet through inflammation and oxidation [71], suggesting a higher biological age for individuals with higher reliance on more ultra-processed foods (SI Section 4, Figure S21).

Vitamins and Phytoestrogens: A greater consumption of more extensively processed foods correlates with lower levels of vitamins in our blood stream, like Vitamin B_{12} and Vitamin C, despite the fact that ultra-processed breakfast cereals and refined flours are frequently fortified with these vitamins and minerals.

In addition, $iFPro$ allows us to discriminate plant-based diets relying on legumes, whole grains, fruits and vegetables, from diets that include ultra-processed plant-based meat, dairy substitutes, and plant-based drinks [72]. For example, we observe a positive correlation between $iFPro_{WC}$ and the urine levels of daidzein, genistein, and their bacterial metabolite o-desmethylangolensin – all bio-markers of soy, abundant in diets that rely on highly processed soy protein foods. On the other hand, enterolactone and enterodiol, gut metabolites of plant lignans (fiber-associated compounds found in many plant families and common foods, including grains, nuts, seeds, vegetables, and drinks such as tea, coffee or wine) are inversely associated with $iFPro_{WC}$ [73].

Chemical Exposures: Diet rich in highly processed food shows association with increased carcinogenic, diabetic, and obesity-inducing food additives and neoformed contaminants, several of which represent previously unknown relationships. For instance, we find positive associations with acrylamide and polycyclic aromatic hydrocarbons, present in heat treated processed food products as a result of Maillard reaction, benzenes (abundant in soft drinks), furans (common in canned and jarred foods), PCBs (processed meat products), perfluorooctanoic acids and phthalates (found in the wrappers of some fast foods, microwavable popcorn, and candy), and environmental phenols like the endocrine disruptor bisphenol A (linked to plastics and resins for food packaging) [74–79]. Importantly, all these represent compounds not reported in food composition databases, but recovered in blood and urine (Figure S21).

Taken together, our ability to distinguish the degree of processing of foods in individual diets allowed us to unveil multiple correlations between diets with greater reliance on ultra-processing and health outcomes. This approach leads to results that cannot be captured with the existing NOVA classification (SI Section 4, Figures S18 and S19). Specifically, if we rely on the manual NOVA 4 and the panel of exposures in NHANES, we recover only 92 significant associations. Among the missing associations are the inverse correlation of a more ultra-processed diet with vitamin D and folate blood levels, and the positive correlation with c-peptide, homocysteine, and blood pressure (SI Section 4).

Food Substitution

The high reliance of ultra-processed food among the U.S. population (Figure 3C) prompts us to ask: what kind of interventions could help us reduce the observed health implications? Given the challenges of large behavioral shifts [80, 81], here we assume that an individual does

not need to overhaul her entire diet, but replace the most processed items she consumes with less processed versions of the same item. To minimize the dietary shifts required, we identify within each individual's diet the item with high caloric contribution and for which there are significantly less processed alternatives (Eq. S8), preserving the broad food class (WWEIA) of the initial choice (SI Section 5). For example, we replace Kix cereals ($FPro = 0.9998$) with shredded wheat and bran cereals ($FPro = 0.5091$), and spread cheese from a pressurized can ($FPro = 0.9648$) with provolone cheese ($FPro = 0.5001$).

We find that the isocaloric replacement of a single item from the average of 19.28 food items daily consumed by U.S. adults reduces the median $iFPro_{WC}$ by 12.15%, from an overall score of 0.7872 to 0.6915 (Figure 3E). Based on observed associations with phenotypes, this translate to a decrease in the odds of metabolic syndrome by 12.25%, a lower concentration of urinary bisphenol A by 8.47%; and an increased blood concentration of vitamin B_{12} and vitamin C by 4.83% and 12.31%, respectively (Figure 3F and Eqs. S9 and S10). Furthermore, the substitution of 10 food items, about half the daily reported items, leads to 37.03% decrease in $iFPro_{WC}$, with associated changes of −21.43% for bisphenol A, +13.02% for blood vitamin B_{12}, and +37.26% for blood vitamin C.

Overall, we find that modest substitution strategies made possible by $FPro$ can preserve the general nature of an individual's diet but reduce reliance on more highly processed food and successfully moderate the observed health implications. As a first step to implement such strategies in the real world, consumers and policymakers must be empowered with information on the degree of processing characterizing the foods in their food environment available to consume, and their potential alternatives. To successfully and sustainably improve population diets, interventions will need to combine individual nudges and behavior change strategies with interventions that more precisely address the ecological, system-level phenomena that people are exposed to, including addressing major gaps and disparities in life-style and healthy food access (i.e., nutrition security).

Discussion

In this paper, we introduced FPro, a continuous processing score combining in a non-linear fashion features of processing techniques learned from the NOVA manual labels, with nutrient concentrations from food composition data. FPro is derived from FoodProX, a classifier that shows a remarkable ability to replicate the manual NOVA classification from nutritional information, confirming that NOVA classes result in distinct patterns of nutrient alterations,

accurately detected by machine learning. Importantly, FoodProX allows us to build upon and extend the current NOVA classification in several crucial respects, offering automated and reproducible classification of foods across multiple national and commercial databases, ability to classify complex recipes and mixed foods and meals, and ability to quantify the extent of food processing among the large and otherwise homogeneously categorized groups of ultra-processed foods. Given that our algorithm only needs the Nutrition Facts, information already accessible to consumers on packaging and via smartphone apps, web portals, and grocery store and restaurant websites, $FPro$ can help monitor the reliance of an individual's diet on less or more processed food. Building on the portability of our model, we were able to extend the $FPro$ assessment to over 50,000 products collected from major US grocery store websites, a first step towards a systematic characterization of the food environments [82].

Differently from dietary indexes such as HEI-15 [83], designed to measure the alignment of individuals' diets with the 2015-2020 Dietary Guidelines for Americans, $iFPro$ and $FPro$ help us identify which foods to substitute, to shift individual consumption patterns towards a less processed diet. Our substitution heuristic indicates that minimal changes in diet can significantly reduce disease risk, a strategy hard to implement with the current NOVA classification, which classifies more than 70% of the food supply as NOVA 4 [84].

The consistent predictive power of $FPro$ in the epidemiological analysis indicates that it offers an accurate global scale of food processing, capturing the chemical-physical alterations of food and its impact on health. However, $FPro$ is currently best suited to rank foods within the same food category, hence offering accurate input for substitution strategies, as explored above. In other words, we should first identify a chemically-driven food group (e.g. "Fruit"), and then quantify the extent of nutrient alterations that leads to different degrees of processing.

Overall, a combination of $FPro$ with epidemiological studies and food classification could lead to an automated and practical pipeline capable of systematically improving population diet and individual health.

Acknowledgments

We thank Dr. Euridice Martinez Steele at the University of Sao Paulo for providing NOVA manual classification for FNDDS databases. We thank Dr. Wei Wang at Harvard Medical School for consulting as statistical expert. This work was conducted with support from Harvard Catalyst | The Harvard Clinical and Translational Science Center (National Center for Advancing Translational Sciences, National Institutes of Health Award UL 1TR002541) and financial contributions from Harvard University and its affiliated academic healthcare centers. The content is solely the responsibility of the authors and does not necessarily represent the official views of Harvard Catalyst, Harvard University and its affiliated academic healthcare centers, or the National Institutes of Health. A.-L.B is partially supported by NIH grant 1P01HL132825, American Heart Association grant 151708, and ERC grant 810115-DYNASET.

Competing Interests

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

Code and Data Availability

All codes and data are available at https://github.com/menicgiulia/MLFoodProcessing .

Author contributions

G.M. and A.-L.B. conceived the project. G.M. designed the study, performed data modeling, analytical calculations, data query and integration, and wrote the manuscript. B.R. performed data query, data integration, statistical analysis, and contributed to writing the manuscript. D.M. contributed to interpreting the results and writing the manuscript. A.-L.B. contributed to the conceptual design of the study and writing the manuscript.

References

[1] Tapsell, L. C., Neale, E. P., Satija, A. & Hu, F. B. Foods, Nutrients, and Dietary Patterns: Interconnections and Implications for Dietary Guidelines. *Advances in Nutrition* **7**, 445–454 (2016).

[2] Willett, W. *et al.* Food in the Anthropocene: the EAT–Lancet Commission on healthy diets from sustainable food systems. *The Lancet* **393**, 447–492 (2019).

[3] MyPyramid. `https://www.fns.usda.gov/mypyramid`.

[4] Choose My Plate. `https://www.choosemyplate.gov`.

[5] Fraanje, W. & Garnett, T. What is ultra-processed food? And why do people disagree about its utility as a concept? (Foodsource: building blocks). *Encyclopedic Dictionary of Polymers* 98–98 (2019).

[6] Brazilian Dietary Guidelines (2015). URL `http://www.fao.org/nutrition/education/food-based-dietary-guidelines/regions/countries/brazil/en/`.

[7] Spector, T. D. & Gardner, C. D. Challenges and opportunities for better nutrition science-an essay by Tim Spector and Christopher Gardner. *BMJ (Clinical research ed.)* **369**, m2470 (2020). URL `http://dx.doi.org/10.1136/bmj.m2470`.

[8] Lustig, R. H. Processed food-an experiment that failed. *JAMA Pediatrics* **171**, 212–214 (2017). URL `http://www.npc.umich`.

[9] Forouhi, N. G., Misra, A., Mohan, V., Taylor, R. & Yancy, W. Dietary and nutritional approaches for prevention and management of type 2 diabetes. *BMJ (Online)* **361**, k2234–k2234 (2018). URL `https://europepmc.org/articles/PMC5998736https://europepmc.org/article/pmc/5998736`.

[10] Lichtenstein, A. H. *et al.* 2021 Dietary Guidance to Improve Cardiovascular Health: A Scientific Statement From the American Heart Association **144**, e472–e487 (2021). URL `https://www.ahajournals.org/doi/abs/10.1161/CIR.0000000000001031`.

[11] Cordain, L. *et al.* Origins and evolution of the Western diet: health implications for the 21st century. *The American Journal of Clinical Nutrition* **81**, 341–354 (2005). URL `https://doi.org/10.1093/ajcn.81.2.341`. `https://academic.oup.com/ajcn/article-pdf/81/2/341/23874879/341.pdf`.

[12] Pagliai, G. *et al.* Consumption of ultra-processed foods and health status: A systematic review and meta-Analysis **125**, 308–318 (2021). URL https://pubmed.ncbi.nlm.nih.gov/32792031/.

[13] Gibney, M. J. & Forde, C. G. Nutrition research challenges for processed food and health. *Nature Food* **3**, 104–109 (2022). URL https://www.nature.com/articles/s43016-021-00457-9.

[14] Slimani, N. *et al.* Contribution of highly industrially processed foods to the nutrient intakes and patterns of middle-aged populations in the european prospective investigation into cancer and nutrition study. *European Journal of Clinical Nutrition* **63**, S206–S225 (2009). URL https://www.nature.com/articles/ejcn200982.

[15] Danish Food Informatics. LanguaL^{TM} - an international framework for food description - Home (2013). URL https://www.langual.org/default.asphttp://www.langual.org/.

[16] EFSA. Food classification standardization – The FoodEx2 system (2020). URL https://www.efsa.europa.eu/en/data/data-standardisationhttps://www.efsa.europa.eu/en/data-report/food-consumption-data.

[17] Scholz, R. *et al.* Database of processing techniques and processing factors compatible with the EFSA food classification and description system FoodEx 2 Objective 1: Compendium of Representative Processing Techniques investigated in regulatory studies for pesticides. *EFSA Supporting Publications* **15** (2018).

[18] van Donkersgoed, G. *et al.* Database of processing techniques and processing factors compatible with the EFSA food classification and description system FoodEx2 related to pesticide residues Objective 2: Linking the processing techniques investigated in regulatory studies with the E. *EFSA Supporting Publications* **15** (2018).

[19] Eftimov, T., Korošec, P. & Koroušić Seljak, B. Standfood: Standardization of foods using a semi-automatic system for classifying and describing foods according to FoodEx2. *Nutrients* **9**, 542 (2017). URL https://www.mdpi.com/2072-6643/9/6/542/htmhttps://www.mdpi.com/2072-6643/9/6/542.

[20] FoodOn: A farm to fork ontology. URL https://foodon.org/.

[21] Durazzo, A. *et al.* Italian composite dishes: description and classification by LanguaL^{TM} and FoodEx2. *European Food Research and Technology* **246**, 287–295 (2020). URL https://link.springer.com/article/10.1007/s00217-019-03341-w.

[22] Moubarac, J.-C., Parra, D. C., Cannon, G. & Monteiro, C. A. Food Classification Systems Based on Food Processing: Significance and Implications for Policies and Actions: A Systematic Literature Review and Assessment. *Current Obesity Reports* **3**, 256–272 (2014). URL `https://pubmed.ncbi.nlm.nih.gov/26626606/`.

[23] Monteiro, C. A. *et al.* The UN Decade of Nutrition, the NOVA food classification and the trouble with ultra-processing. *Public Health Nutrition* **21**, 5–17 (2018).

[24] Chen, X. *et al.* Consumption of ultra-processed foods and health outcomes: A systematic review of epidemiological studies **19**, 1–10 (2020). URL `https://nutritionj.biomedcentral.com/articles/10.1186/s12937-020-00604-1`.

[25] Monteiro, C. A. *et al.* NOVA. The star shines bright. *World Nutrition* **7**, 28–38 (2016).

[26] Steele, E. M. *et al.* Ultra-processed foods and added sugars in the US diet: Evidence from a nationally representative cross-sectional study. *BMJ Open* **6**, 1–8 (2016).

[27] Steele, E. M. & Monteiro, C. A. Association between dietary share of ultra-processed foods and urinary concentrations of phytoestrogens in the US. *Nutrients* **9** (2017).

[28] Adjibade, M. *et al.* Prospective association between ultra-processed food consumption and incident depressive symptoms in the French NutriNet-Santé cohort. *BMC Medicine* **17**, 1–13 (2019).

[29] Fiolet, T. *et al.* Consumption of ultra-processed foods and cancer risk: Results from NutriNet-Santé prospective cohort. *BMJ (Online)* **360** (2018).

[30] Srour, B. *et al.* Ultra-processed food intake and risk of cardiovascular disease: Prospective cohort study (NutriNet-Santé). *The BMJ* **365** (2019).

[31] Hall, K. D. *et al.* Ultra-Processed Diets Cause Excess Calorie Intake and Weight Gain: An Inpatient Randomized Controlled Trial of Ad Libitum Food Intake. *Cell Metabolism* **30**, 1–11 (2019). URL `https://doi.org/10.1016/j.cmet.2019.05.008`.

[32] Vandevijvere, S. *et al.* Global trends in ultraprocessed food and drink product sales and their association with adult body mass index trajectories. *Obesity Reviews* **20**, 10–19 (2019). URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/obr.12860`. `https://onlinelibrary.wiley.com/doi/pdf/10.1111/obr.12860`.

[33] Zhang, Z., Jackson, S. L., Martinez, E., Gillespie, C. & Yang, Q. Association between ultraprocessed food intake and cardiovascular health in us adults: a cross-sectional anal-

ysis of the nhanes 2011–2016. *The American Journal of Clinical Nutrition* **113**, 428–436 (2020). URL https://doi.org/10.1093/ajcn/nqaa276. https://academic.oup.com/ajcn/article-pdf/113/2/428/36170430/nqaa276.pdf.

[34] Micha, R., Wallace, S. K. & Mozaffarian, D. Red and processed meat consumption and risk of incident coronary heart disease, stroke, and diabetes mellitus: A systematic review and meta-analysis. *Circulation* **121**, 2271–2283 (2010).

[35] Levy, R. B. *et al.* Ultra-processed food consumption and type 2 diabetes incidence: A prospective cohort study. *Clinical Nutrition* **40**, 3608–3614 (2021). URL https://pubmed.ncbi.nlm.nih.gov/33388205/.

[36] Wang, L. *et al.* Association of ultra-processed food consumption with colorectal cancer risk among men and women: results from three prospective US cohort studies. *BMJ* **378**, e068921 (2022). URL https://www.bmj.com/content/378/bmj-2021-068921https://www.bmj.com/content/378/bmj-2021-068921.abstract.

[37] Bleiweiss-Sande, R. *et al.* Robustness of Food Processing Classification Systems. *Nutrients* **11** (2019). URL https://pubmed.ncbi.nlm.nih.gov/31208000/.

[38] Sadler, C. R. *et al.* Processed food classification: Conceptualisation and challenges. *Trends in Food Science & Technology* **112**, 149–162 (2021).

[39] Braesco, V. *et al.* Ultra-processed foods: how functional is the NOVA system? *European Journal of Clinical Nutrition* **76**, 1245–1253 (2022). URL https://www.nature.com/articles/s41430-022-01099-1.

[40] Monteiro, C. A. *et al.* Ultra-processed foods: what they are and how to identify them. *Public Health Nutrition* **22**, 936–941 (2019).

[41] Meneguelli, T. S. *et al.* Food consumption by degree of processing and cardiometabolic risk: a systematic review. *International Journal of Food Sciences and Nutrition* **71**, 678–692 (2020). URL https://doi.org/10.1080/09637486.2020.1725961. PMID: 32053758, https://doi.org/10.1080/09637486.2020.1725961.

[42] Lane, M. M. *et al.* Ultraprocessed food and chronic noncommunicable diseases: A systematic review and meta-analysis of 43 observational studies. *Obesity Reviews* 1–19 (2020).

[43] Martínez Steele, E., Popkin, B. M., Swinburn, B. & Monteiro, C. A. The share of ultra-processed foods and the overall nutritional quality of diets in the US: Evidence from a

nationally representative cross-sectional study. *Population Health Metrics* **15**, 1–11 (2017). NIHMS150003.

[44] Gupta, S., Hawk, T., Aggarwal, A. & Drewnowski, A. Characterizing ultra-processed foods by energy density, nutrient density, and cost. *Frontiers in Nutrition* **6** (2019). URL https://www.frontiersin.org/article/10.3389/fnut.2019.00070/full.

[45] Menichetti, G. & Barabási, A.-L. Nutrient concentrations in food display universal behaviour. *Nature Food 2022 3:5* **3**, 375–382 (2022). URL https://www.nature.com/articles/s43016-022-00511-0.

[46] USDA FoodData Central. https://fdc.nal.usda.gov/.

[47] Ahuja, J. *et al.* USDA Food and Nutrient Database for Dietary Studies, 5.0. U.S. Department of Agriculture, Agricultural Research Service, Food Surveys Research Group, Beltsville, MD. http://www.ars.usda.gov/ba/bhnrc/fsrg (2012).

[48] Sebastian, R. S. *et al.* Flavonoid Values for USDA Survey Foods and Beverages 2007–2010. U.S. Department of Agriculture, Agricultural Research Service, Food Surveys Research Group, Beltsville, MD. http://www.ars.usda.gov/nea/bhnrc/fsrg (2016).

[49] FDA Nutrition Facts. https://www.fda.gov/media/99331/download.

[50] Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems 30* 4765–4774 (2017). URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf. 1705.07874.

[51] Davidou, S., Christodoulou, A., Frank, K. & Fardet, A. A study of ultra-processing marker profiles in 22,028 packaged ultra-processed foods using the siga classification. *Journal of Food Composition and Analysis* **99**, 103848 (2021). URL https://www.sciencedirect.com/science/article/pii/S088915752100048X.

[52] Rhodes, D. G., Adler, M. E., Clemens, J. C. & Moshfegh, A. J. What we eat in America food categories and changes between survey cycles. *Journal of Food Composition and Analysis* **64**, 107–111 (2017). URL http://dx.doi.org/10.1016/j.jfca.2017.07.018.

[53] FDA Center for Food Safety and Applied Nutrition. Guidance for Industry: Guide for Developing and Using Data Bases for Nutrition Labeling. URL https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-industry-guide-developing-and-using-data-bases-nutrition-labeling.

[54] Barabási, A.-L., Menichetti, G. & Loscalzo, J. The unmapped chemical complexity of our diet. *Nature Food* **1**, 33–37 (2020). URL https://doi.org/10.1038/s43016-019-0005-1.

[55] Open Food Facts - World (10/15/2021). URL https://world.openfoodfacts.org/.

[56] Formula to determine the Nova group (10/15/2021). URL https://world.openfoodfacts.org/nova.

[57] Reedy, J. *et al.* Evaluation of the healthy eating index-2015. *Journal of the Academy of Nutrition and Dietetics* **118**, 1622 – 1633 (2018). URL http://www.sciencedirect.com/science/article/pii/S2212267218308360.

[58] Willett, W. *Nutritional epidemiology* (1998).

[59] Patel, C. J. *et al.* A database of human exposomes and phenomes from the us national health and nutrition examination survey. *Scientific Data* **3**, 160096 (2016). URL https://doi.org/10.1038/sdata.2016.96.

[60] Milanlouei, S. *et al.* A systematic comprehensive longitudinal evaluation of dietary factors associated with acute myocardial infarction and fatal coronary heart disease. *Nature Communications* **11**, 1–14 (2020). URL https://doi.org/10.1038/s41467-020-19888-2.

[61] Nardocci, M., Polsky, J. Y. & Moubarac, J. C. Consumption of ultra-processed foods is associated with obesity, diabetes and hypertension in Canadian adults. *Canadian Journal of Public Health* 1–9 (2020). URL https://doi.org/10.17269/s41997-020-00429-9.

[62] De Deus Mendonça, R. *et al.* Ultra-processed food consumption and the incidence of hypertension in a mediterranean cohort: The seguimiento universidad de navarra project. *American Journal of Hypertension* **30**, 358–366 (2017). URL https://pubmed.ncbi.nlm.nih.gov/27927627/.

[63] D'Agostino, R. B. *et al.* General cardiovascular risk profile for use in primary care: The Framingham heart study. *Circulation* **117**, 743–753 (2008).

[64] Goff, D. C. *et al.* 2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the American college of cardiology/American heart association task force on practice guidelines. *Circulation* **129**, 49–73 (2014).

[65] de Miranda, R. C., Rauber, F. & Levy, R. B. Impact of ultra-processed food consumption on metabolic health. *Current Opinion in Lipidology* **32**, 24–37 (2021).

[66] Yang, Q., Zhang, Z., Steele, E. M., Moore, L. V. & Jackson, S. L. Ultra-processed foods and excess heart age among u.s. adults. *American Journal of Preventive Medicine* **59**, e197–e206 (2020). URL https://www.sciencedirect.com/science/article/pii/S0749379720302889.

[67] Bonaccio, M. *et al.* Ultra-processed food consumption is associated with increased risk of all-cause and cardiovascular mortality in the Moli-sani Study. *The American Journal of Clinical Nutrition* **113**, 446–455 (2020). URL https://doi.org/10.1093/ajcn/nqaa299. https://academic.oup.com/ajcn/article-pdf/113/2/446/36170506/nqaa299.pdf.

[68] Martínez Leo, E. E. & Segura Campos, M. R. Effect of ultra-processed diet on gut microbiota and thus its role in neurodegenerative diseases. *Nutrition* **71**, 110609 (2020). URL https://www.sciencedirect.com/science/article/pii/S0899900719301923.

[69] Neri, D., Martinez-Steele, E., Monteiro, C. A. & Levy, R. B. Consumption of ultra-processed foods and its association with added sugar content in the diets of us children, nhanes 2009-2014. *Pediatric Obesity* **14**, e12563 (2019). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ijpo.12563. E12563 IJPO-2019-0182, https://onlinelibrary.wiley.com/doi/pdf/10.1111/ijpo.12563.

[70] Cheung, C.-L., Tan, K. C. B., Lam, K. S. L. & Cheung, B. M. Y. The Relationship Between Glucose Metabolism, Metabolic Syndrome, and Bone-Specific Alkaline Phosphatase: A Structural Equation Modeling Approach. *The Journal of Clinical Endocrinology & Metabolism* **98**, 3856–3863 (2013). URL https://doi.org/10.1210/jc.2013-2024. https://academic.oup.com/jcem/article-pdf/98/9/3856/11127694/1.jc-13-2024_tab3.pdf.

[71] Alonso-Pedrero, L. *et al.* Ultra-processed food consumption and the risk of short telomeres in an elderly population of the Seguimiento Universidad de Navarra (SUN) Project. *The American journal of clinical nutrition* **111**, 1259–1266 (2020). URL https://academic.oup.com/ajcn/article/111/6/1259/5824715.

[72] Gehring, J. *et al.* Consumption of Ultra-Processed Foods by Pesco-Vegetarians, Vegetarians, and Vegans: Associations with Duration and Age at Diet Initiation. *The Journal of nutrition* **151**, 120–131 (2021).

[73] Lampe, J. W. Isoflavonoid and lignan phytoestrogens as dietary biomarkers. *Journal of Nutrition* **133**, 956S–964S (2003). URL https://academic.oup.com/jn/article/133/3/956S/4688069.

[74] Martínez Steele, E., Khandpur, N., da Costa Louzada, M. L. & Monteiro, C. A. Association between dietary contribution of ultra-processed foods and urinary concentrations of phthalates and bisphenol in a nationally representative sample of the us population aged 6 years and older. *PLOS ONE* **15**, 1–21 (2020). URL https://doi.org/10.1371/journal.pone.0236738.

[75] Nerín, C., Aznar, M. & Carrizo, D. Food contamination during food process. *Trends in Food Science & Technology* **48**, 63–68 (2016). URL https://www.sciencedirect.com/science/article/pii/S0924224415301370.

[76] Rather, I. A., Koh, W. Y., Paek, W. K. & Lim, J. The sources of chemical contaminants in food and their health implications. *Frontiers in Pharmacology* **8**, 830 (2017). URL https://www.frontiersin.org/article/10.3389/fphar.2017.00830.

[77] Jain, R. B. & Wang, R. Y. Association of caffeine consumption and smoking status with the serum concentrations of polychlorinated biphenyls, dioxins, and furans in the general u.s. population: Nhanes 2003–2004. *Journal of Toxicology and Environmental Health, Part A* **74**, 1225–1239 (2011). URL https://doi.org/10.1080/15287394.2011.587105. PMID: 21797774, https://doi.org/10.1080/15287394.2011.587105.

[78] Arisseto, A. P. Chapter 21 - furan in processed foods. In Kotzekidou, P. (ed.) *Food Hygiene and Toxicology in Ready-to-Eat Foods*, 383–396 (Academic Press, San Diego, 2016). URL https://www.sciencedirect.com/science/article/pii/B9780128019160000212.

[79] Buckley, J. P., Kim, H., Wong, E. & Rebholz, C. M. Ultra-processed food consumption and exposure to phthalates and bisphenols in the us national health and nutrition examination survey, 2013–2014. *Environment International* **131**, 105057 (2019). URL https://www.sciencedirect.com/science/article/pii/S0160412019317416.

[80] Hunter, R. F. *et al.* Social network interventions for health behaviour change: a systematic review. *The Lancet* **390**, S47 (2017). URL www.thelancet.com.

[81] Salvy, S. J., de la Haye, K., Bowker, J. C. & Hermans, R. C. Influence of peers and friends on children's and adolescents' eating and activity behaviors. *Physiology and Behavior* **106**, 369–378 (2012).

[82] Ravandi, B., Mehler, P., Barabási, A.-L. & Menichetti, G. GroceryDB: Prevalence of Processed Food in Grocery Stores. *medRxiv* 2022.04.23.22274217 (2022). URL https://www.medrxiv.org/content/10.1101/2022.04.23.22274217v1.

[83] National cancer institute. developing the healthy eating index. bethesda, md: National cancer institute. https://epi.grants.cancer.gov/hei/developing.html. 2020 (accessed September 1, 2020).

[84] Baldridge, A. S. *et al.* The Healthfulness of the US Packaged Food and Beverage Supply: A Cross-Sectional Study. *Nutrients* **11**, 1704 (2019).

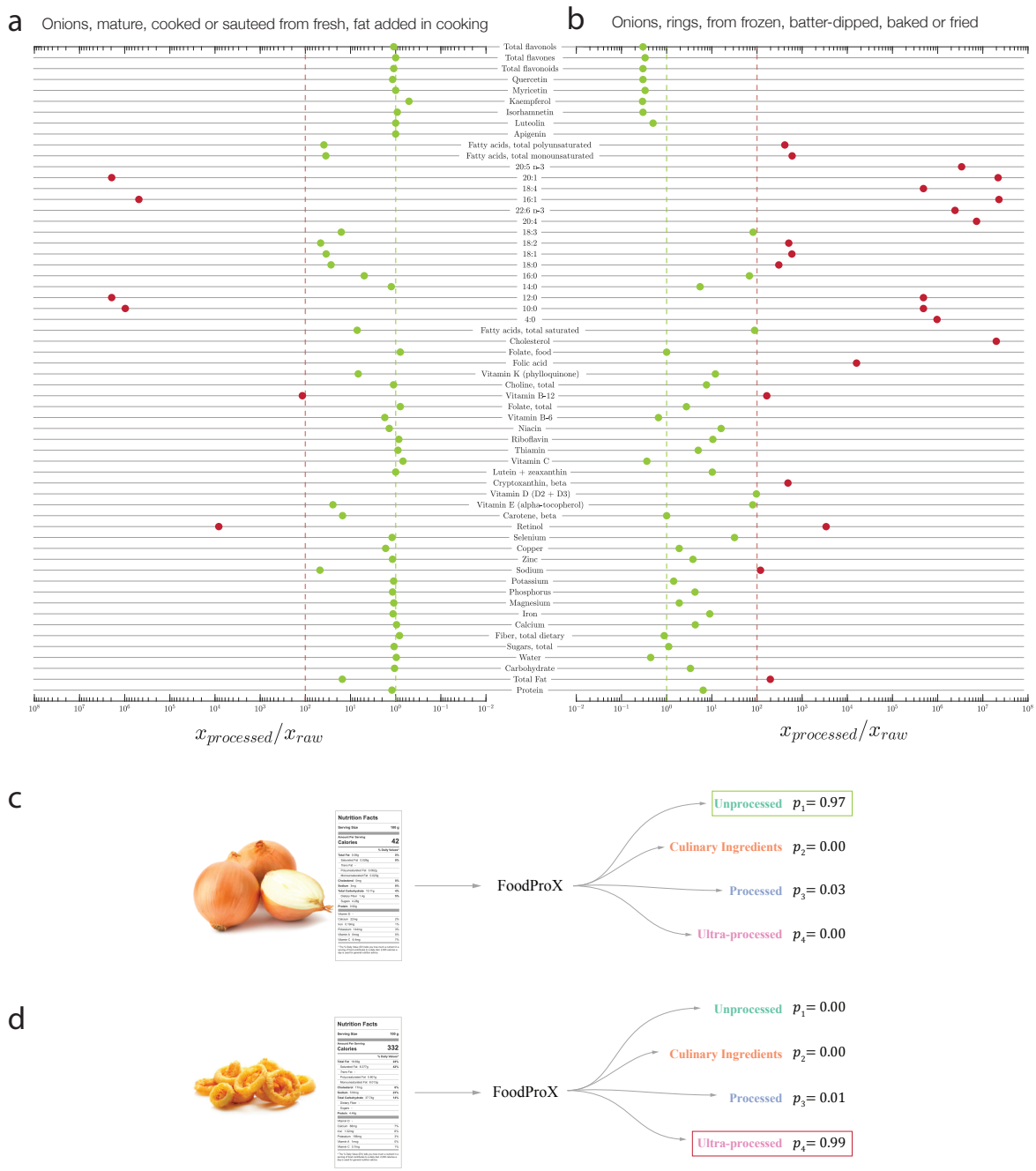

Figure 1

Figure 1: **Food Processing and Nutrient Changes (FoodProX). (a)-(b)** Ratio of nutrient concentrations for 100 g of Sauteed Onion and Onion Rings compared to Raw Onion, indicating how processing alters the concentration of multiple nutrients. All nutrients in excess compared to the concentrations found in the raw ingredient are shown in red. **(c)-(d)** We trained FoodProX, a random forest classifier over the nutrient concentrations within 100 g of each food, tasking it to predict its processing level according to NOVA. FoodProX represents each food by a vector of probabilities $\{p_i\}$, capturing the likelihood of being classified as unprocessed (NOVA 1), processed culinary ingredients (NOVA 2), processed (NOVA 3), and ultra-processed (NOVA 4). The highest probability determines the final classification label, highlighted in a box on the right.

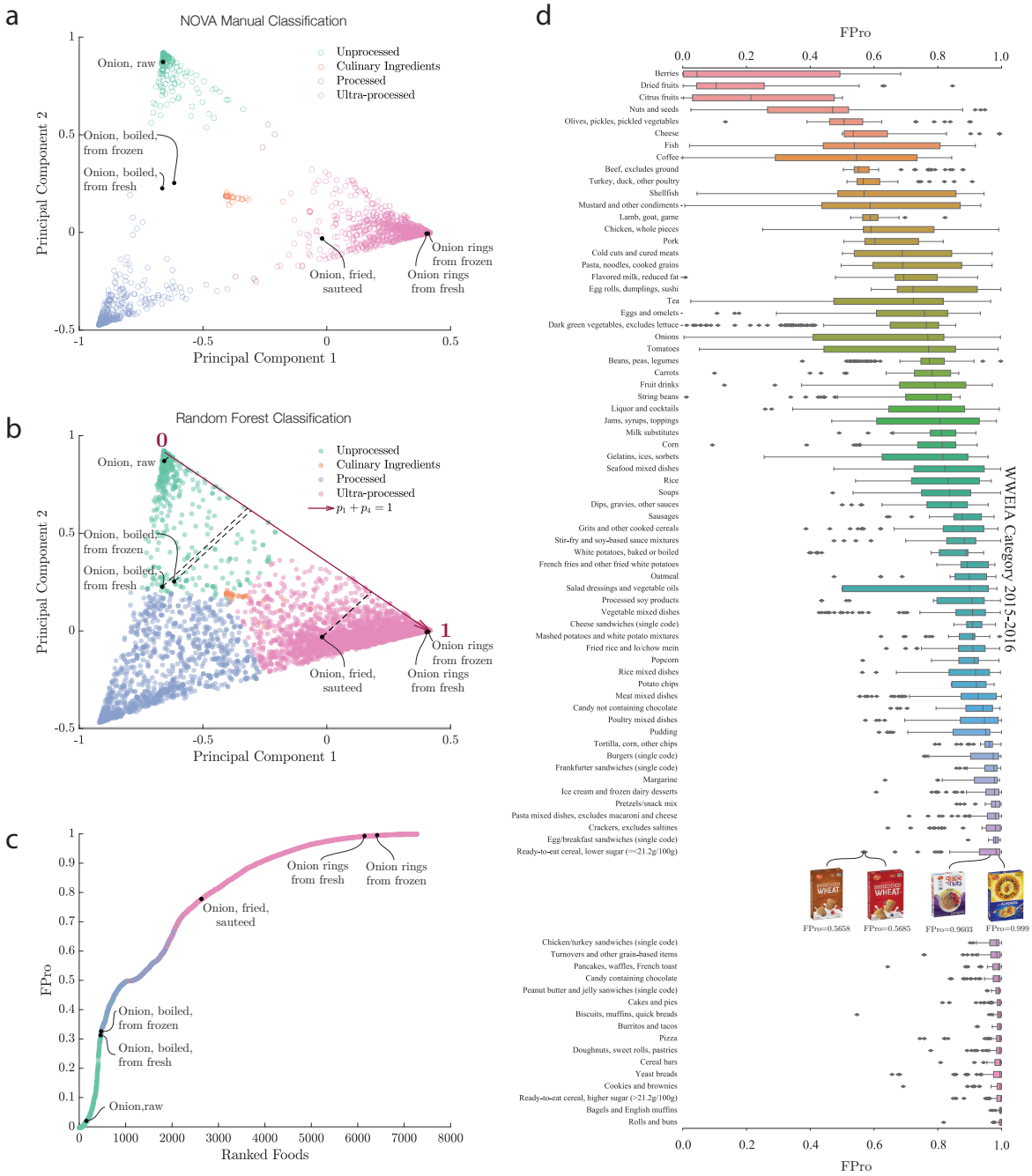
a
NOVA Manual Classification
Unprocessed
Culinary Ingredients
Processed
Ultra-processed
Onion, raw
Onion, boiled, from frozen
Onion, boiled, from fresh
Onion, fried, sauteed
Onion rings from frozen
Onion rings from fresh
Principal Component 2
Principal Component 1
b
Random Forest Classification
0
1
$p_1 + p_4 = 1$
c
FPro
Ranked Foods
Onion,raw
d
FPro
WWEIA Category 2015-2016
FPro=0.5658
FPro=0.5685
FPro=0.9603
FPro=0.999

Figure 2

Figure 2: **NOVA Classification and Processing Score.** **(a)** Visualization of the decision space of FoodProX via principal component analysis of the probabilities $\{p_i\}$. The manual 4-level NOVA classification assigns unique labels to only 34.25% of the foods listed in FNDDS 2009-2010 (empty circles). The classification of the remaining foods remains unknown ("Not Classified"), or must be further decomposed into ingredients ("Composite Recipe"). The list of foods manually classified by NOVA is largely limited to the three corners of the phase space, foods to which the classifier assigns dominating probabilities. **(b)** FoodProX assigned NOVA labels to all foods in FNDDS 2009-2010. The symbols at the boundary regions indicates that for these foods the algorithm's confidence in the classification is not high, hence a 4-class classification does not capture the degree of processing characterizing that food. For each food k, the processing score $FPro_k$ represents the orthogonal projection (black dashed lines) of $\vec{p}^k = (p_1^k, p_2^k, p_3^k, p_4^k)$ onto the line $p_1 + p_4 = 1$ (highlighted in dark red). **(c)** We ranked all foods in FNDDS 2009/2010 according to $FPro$. The measure sorts onion products in increasing order of processing, from "Onion, Raw", to "Onion rings, from frozen". **(d)** Distribution of $FPro$ for a selection of the 155 Food Categories in What We Eat in America (WWEIA) 2015-2016 with at least 20 items (SI Section 2). WWEIA categories group together foods and beverages with similar usage and nutrient content in the U.S. food supply [52]. All categories are ranked in increasing order of median $FPro$, indicating that within each food group, we have remarkable variability in $FPro$, confirming the presence of different degrees of processing. We illustrate this through four ready-to-eat cereals, all manually classified as NOVA 4, yet with rather different $FPro$. While the differences in the nutrient content of Post Shredded Wheat 'n Bran ($FPro = 0.5658$) and Post Shredded Wheat ($FPro = 0.5685$) are minimal, with lower fiber content for the latter, the fortification with vitamins, minerals, and the addition of sugar, significantly increases the processing of Post Grape-Nuts ($FPro = 0.9603$), and the further addition of fats results in an even higher processing score for Post Honey Bunches of Oats with Almonds ($FPro = 0.9999$), showing how $FPro$ ranks the progressive changes in nutrient content.

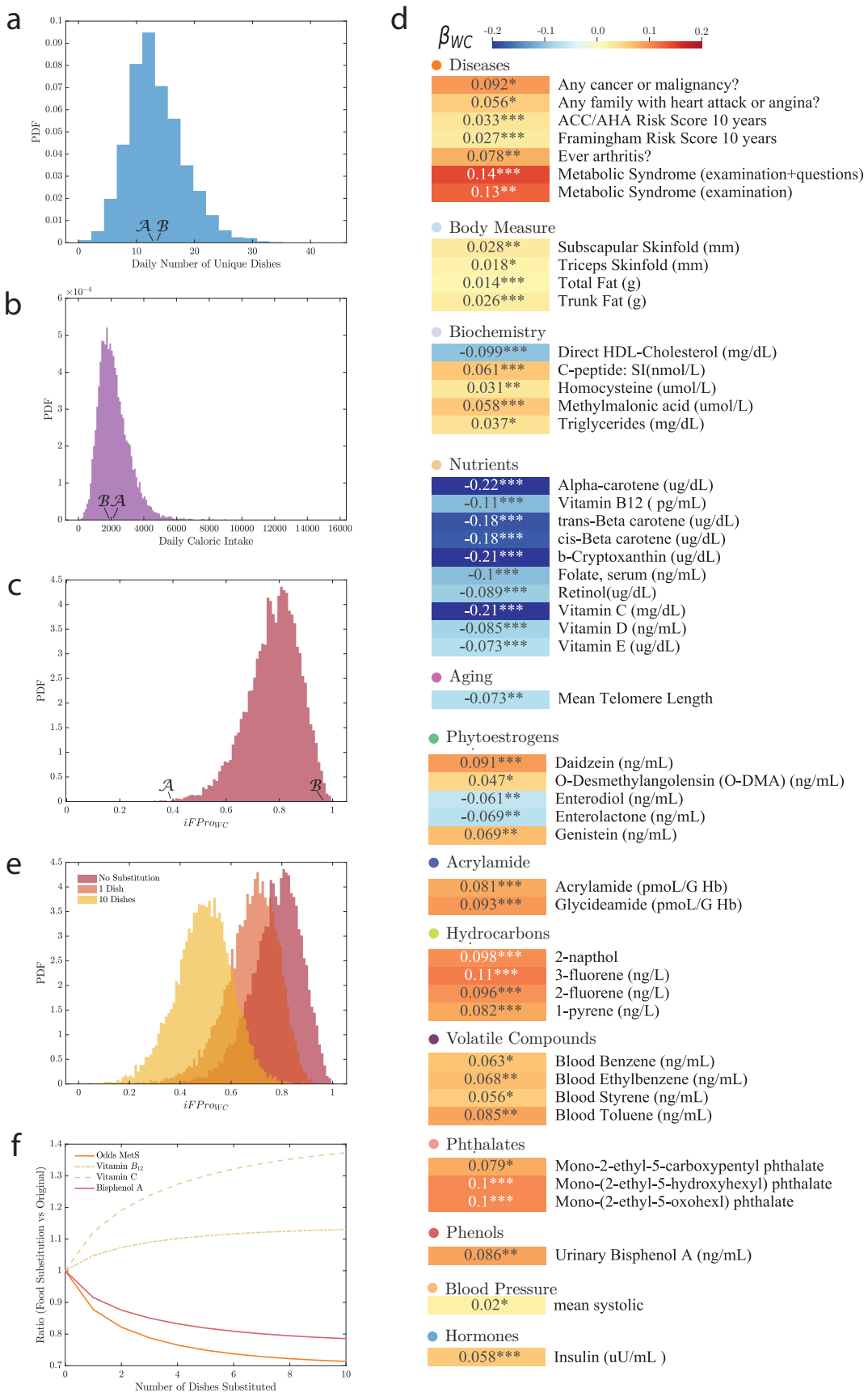
a
PDF
Daily Number of Unique Dishes
b
PDF
Daily Caloric Intake
c
PDF
$iFPro_{WC}$
e
No Substitution
1 Dish
10 Dishes
PDF
$iFPro_{WC}$
f
Odds MetS
Vitamin B_{12}
Vitamin C
Bisphenol A
Ratio (Food Substitution vs Original)
Number of Dishes Substituted
d
β_{WC}
Diseases
0.092* Any cancer or malignancy?
0.056* Any family with heart attack or angina?
0.033*** ACC/AHA Risk Score 10 years
0.027*** Framingham Risk Score 10 years
0.078** Ever arthritis?
0.14*** Metabolic Syndrome (examination+questions)
0.13** Metabolic Syndrome (examination)
Body Measure
0.028** Subscapular Skinfold (mm)
0.018* Triceps Skinfold (mm)
0.014*** Total Fat (g)
0.026*** Trunk Fat (g)
Biochemistry
-0.099*** Direct HDL-Cholesterol (mg/dL)
0.061*** C-peptide: SI(nmol/L)
0.031** Homocysteine (umol/L)
0.058*** Methylmalonic acid (umol/L)
0.037* Triglycerides (mg/dL)
Nutrients
-0.22*** Alpha-carotene (ug/dL)
-0.11*** Vitamin B12 (pg/mL)
-0.18*** trans-Beta carotene (ug/dL)
-0.18*** cis-Beta carotene (ug/dL)
-0.21*** b-Cryptoxanthin (ug/dL)
-0.1*** Folate, serum (ng/mL)
-0.089*** Retinol(ug/dL)
-0.21*** Vitamin C (mg/dL)
-0.085*** Vitamin D (ng/mL)
-0.073*** Vitamin E (ug/dL)
Aging
-0.073** Mean Telomere Length
Phytoestrogens
0.091*** Daidzein (ng/mL)
0.047* O-Desmethylangolensin (O-DMA) (ng/mL)
-0.061** Enterodiol (ng/mL)
-0.069** Enterolactone (ng/mL)
0.069** Genistein (ng/mL)
Acrylamide
0.081*** Acrylamide (pmoL/G Hb)
0.093*** Glycideamide (pmoL/G Hb)
Hydrocarbons
0.098*** 2-napthol
0.11*** 3-fluorene (ng/L)
0.096*** 2-fluorene (ng/L)
0.082*** 1-pyrene (ng/L)
Volatile Compounds
0.063* Blood Benzene (ng/mL)
0.068** Blood Ethylbenzene (ng/mL)
0.056* Blood Styrene (ng/mL)
0.085** Blood Toluene (ng/mL)
Phthalates
0.079* Mono-2-ethyl-5-carboxypentyl phthalate
0.1*** Mono-(2-ethyl-5-hydroxyhexyl) phthalate
0.1*** Mono-(2-ethyl-5-oxohexl) phthalate
Phenols
0.086** Urinary Bisphenol A (ng/mL)
Blood Pressure
0.02* mean systolic
Hormones
0.058*** Insulin (uU/mL)

Figure 3

Figure 3: **Health implications and Food Substitution**. For each of the 20,047 individuals in NHANES (1999-2006), 18+ years old with dietary records [59], we calculated the individual diet processing scores $iFPro_{WC}$. **(a)** The average number of unique dishes reported in the dietary interviews, highlighting two individuals $\mathcal{A}$ and $\mathcal{B}$, with comparable number of dishes, 12.5 and 13 reported, respectively. **(b)** The distribution of average daily caloric intake, showing that individuals $\mathcal{B}$ and $\mathcal{A}$ have similar caloric intake of 1,894 and 2,016 kcal, respectively. **(c)** The distribution of $iFPro_{WC}$ for NHANES, indicating that individuals $\mathcal{A}$ and $\mathcal{B}$ display significant differences in $iFPro_{WC}$, with $\mathcal{B}$'s diet relying on ultra-processed food ($iFPro_{WC} = 0.9572$), and $\mathcal{A}$ reporting simple recipes ($iFPro_{WC} = 0.3981$) (Figure S13). **(d)** We measured the association of various phenotypes with $iFPro_{WC}$, correcting for age, gender, ethnicity, socio-economic status, BMI, and caloric intake (SI Section 4). We report the standardized β coefficient, quantifying the effect on each exposure when the Box-Cox transformed dietary scores increase of one standard deviation over the population. For continuous exposures the coefficients are fully standardized, while for logistic regression (disease phenotypes) we opted for partially standardized coefficients to help interpretability (SI Section 4). Each variable is color-coded according to β, positive associations shown in red, and negative associations in blue. **(e)** Changes in $iFPro_{WC}$ when one (orange) or up to ten (yellow) dishes are substituted with their less processed versions, following the prioritization rule defined in Eq. S8. **(f)** The impact of substituting different number of dishes on the odds of metabolic syndrome, concentrations of vitamin B_{12}, vitamin C, and bisphenol A, showing that a minimal substitution strategy can significantly alter the health implications of ultra-processed food.

SUPPLEMENTARY INFORMATION

Contents

1 Training Dataset **2**

1.1 Food Composition Databases 2

1.2 FNDDS 2009-2010 3

1.3 NOVA Manual Classification Coverage 5

1.4 Nutrient Panels 6

1.5 Hierarchical Clustering of Foods according to Nutrients 8

2 Random Forest Classifier FoodProX and Food Processing Score $FPro$ **10**

2.1 Random Forest Classifier 10

2.2 Feature Importance 11

2.3 Food Processing Score $FPro$ 14

2.4 Validation of $FPro$ in Different FNDDS Editions 16

2.5 Case Study on Post Cereals. 17

2.6 Source of Food 17

3 Individual Diet Processing Scores $iFPro$ and Exposome **19**

3.1 Individual Processing Score $iFPro$ 19

3.2 Population Characteristics 19

3.3 Correlation between $iFPro_{WG}$, $iFPro_{WC}$, and HEI-15 20

3.4 Water Consumption in NHANES 21

3.5 Relation between $iFPro_{WC}$ and WWEIA Food Categories 22

4 Environment-Wide Association Study **24**

5 Food Substitution **33**

6 Open Food Facts **35**

7 Data Quality and Future Directions **37**

8 Overview of FoodProX and FPro **41**

1 Training Dataset

1.1 Food Composition Databases

The food supply, representing the full inventory of all foods available for human consumption, along with their nutritional content, plays an important role in determining an individual's nutrient exposure. Nutritional information is captured by food databases, collections of nutrient measurements for extensive samples of the food supply. Here we define as nutrients all chemicals catalogued by food databases, whether they refer to unique chemicals, like vitamin C, or aggregate measures, like total fat or total sugar. Additionally, all major nutrient databases include calories, measuring how much energy our body could get from eating or drinking the selected product. Which foods and which nutrients to report is strictly dependent on the database considered. For instance, USDA SR Legacy, the authoritative source of food composition data in the United States contains 7,793 food items with variable nutrient resolution, from a minimum of 8 nutrients, up to 138 (Figure S1) [2]. In comparison, USDA FNDDS, designed for the epidemiological analysis of dietary intake data collected by the National Health and Nutrition Examination Survey (NHANES), reports 65 to 102 nutrients for all foods, depending on the edition, containing no missing nutrient values (Figure S1) [3, 4].

The nutrient resolution available to consumers is significantly lower: the Food and Drug Administration (FDA) mandates the listing of 14 nutrients on the nutrition facts label, from saturated and trans fat, to sodium and vitamin C [1] An updated nutrition facts label was finalized in 2016, removing vitamin A and C, but listing added sugars, vitamin D, and potassium. However, the compliance deadline for certain food categories was extended to July 2021, and for the majority of the data describing branded products, we observed a significantly higher coverage of the nutrition facts prior to 2016.

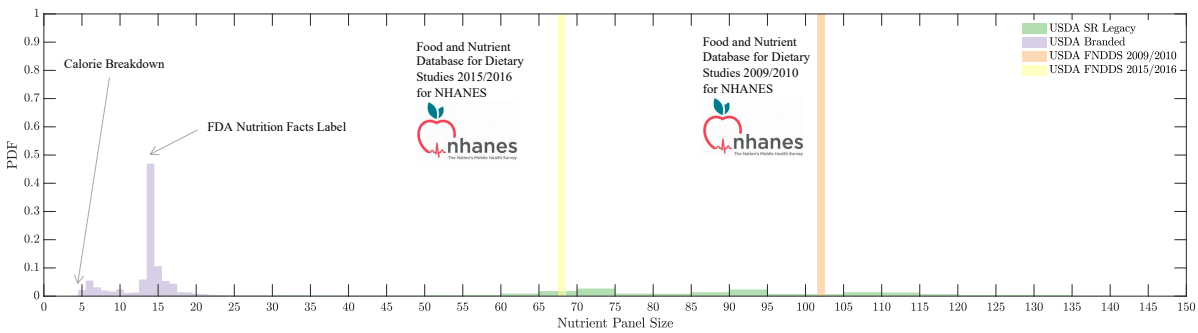

Figure SI. 1: **Nutrient Panel Resolution for Different Food Databases**. To fully capture the nutrient alterations caused by food processing, we need access to the nutrient information characterizing each food in the food supply. The resolution available for branded products sold in grocery stores is very limited, frequently less than what required by FDA nutrition facts [1].

1.2 FNDDS 2009-2010

The Food and Nutrient Database for Dietary Studies (FNDDS) is designed by the USDA to provide food composition data (e.g. the amount of Vitamin C per 100 g of a selected ingredient) for foods and beverages reported in the dietary component of the National Health and Nutrition Examination Survey (NHANES), a biannual cross-sectional survey of the US Population conducted by Center for Disease Control and Prevention (CDC) to monitor the health of Americans. FNDDS is derived by combining the food items provided in the USDA National Nutrient Database for Standard Reference (SR). In other words, each item in FNDDS is related to one or more foods in SR, reported as ingredients in FNDDS. Differently from SR, designed for the dissemination of food composition data, FNDDS's goal is to enable the analysis of dietary intake, hence it contains no missing nutrient values, an ideal setup to train machine learning models [5]. Since 2017, the USDA has been harmonizing these different data sources in FoodData Central (FDC), labeling SR data as SR-Legacy [6]. In particular, SR28 (released in 2015) is the final version of SR-Legacy databases, and it is the foundation of FNDDS 2015-2016. In addition to FNDDS and SR-Legacy, FoodData Central stores also Foundation Foods, a new food composition dataset that reports individual sample measurements behind the nutrient average values that populate the other databases, and metadata reporting the number of samples, location, time-stamps, analytical methods used, and if available, cultivar and production practices.

As shown in Figure S1, for the years 2007-2010 the USDA developed a flavonoids database for population surveys that extended the original nutritional panel of 65 nutrients to 102. For our analysis we kept all nutrients measured in g, mg or μg, dropping "Energy", "Folate, DFE" and "Vitamin A, RAE", resulting in 99 nutrients, converted to grams (g).

We chose FNDDS 2009-2010 as data training for FoodProX, as it gave us the possibility to combine the manual labels assigned by Steele et al. in [7], with the widest panel of nutrients available for population studies. Out of 7,253 foods in FNDDS 2009-2010, 2,484 food items are assigned to a unique NOVA class, while the remaining 4,769 foods are not classified (730), or need further decomposition (4,039) into 2,946 ingredients imported from the SR24 database.

Figure S2A shows the proportion of NOVA classes in the initial dataset, labels mainly derived by following the hierarchical encoding of food items provided by FNDDS. Indeed, each food is assigned to an 8-digit code, and the first five digits represent food categories. For instance, code 13230120 is assigned to "Pudding, flavors other than chocolate, ready-to-eat, sugar free" where the first digit '1' represents "Milk and Milk Products"; the first two digits '13' represents "Milk Desserts and Sauce"; and similarly '132' represents "Puddings, Custards, and other Milk

Desserts." Relying on FNDDS food categories leads to two major limitations of the current classification system: (a) the existence of many exceptions. For example, a non ultra-processed food could belong to a category assumed to contain only ultra-processed food. Resolving these exceptions is a laborious work that requires domain knowledge; (b) limited scalability, as not all databases have a fine-tuned hierarchy of categories assigned to foods comparable to FNDDS.

Additionally, we found that some unclassified items, despite being labeled as requiring further decomposition, had only one ingredient. For instance, this is the case for the unclassified item 'Egg, whole, raw' (food code 31101010), created by linking only a single food from the SR database: 'Egg, whole, raw, fresh' (SR code 1123). Hence, we migrated 478 such unclassified single-ingredient foods to the training dataset (Figure S2B).

To further improve the training dataset, we manually classified nine foods, to extend the coverage of staple ingredients like 'Salt', or poorly represented classes like meat and fish (Table S1). In particular, the addition of 'Salt' in the training helped FoodProX to better calibrate the role of sodium in identifying ultra-processed food.

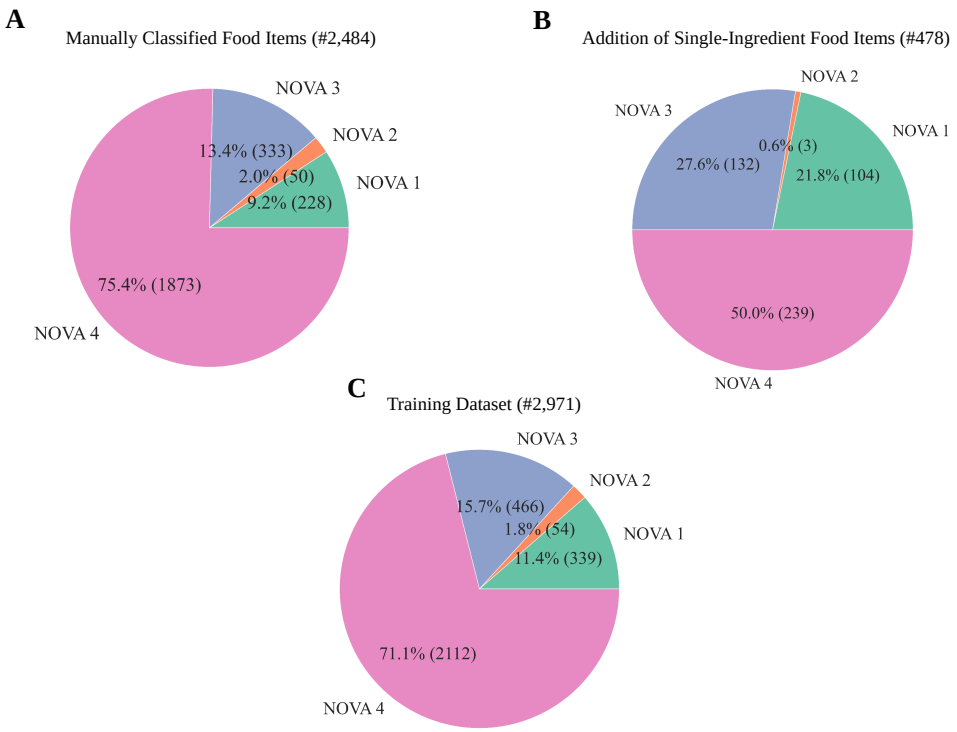

Figure SI. 2: **Proportion of NOVA Classes in the Manual Classification and Training Dataset.** (**A**) Steele et al. [7] manually assigned NOVA classes to 2,484 from 7,253 food items in FNDDS 2009-2010, (**B**) We identified 478 single-ingredient and unclassified food items in FNDDS 2009-2010 linked to a single food in the SR database with assigned NOVA labels. (**C**) Final training dataset with the corrections on single-ingredient food items and the addition of 9 manually classified food items reported in Table S1.

Table SI. 1: Manual Additions to the Training Dataset

Food code	Food Description	Ingredients	NOVA Class
2047	Salt	• Salt, table (directly imported from SR database)	2
26100100	Fish, NS as to type, raw	• Fish, pollock, walleye, raw • Fish, salmon, sockeye, raw • Fish, tilapia, raw • Fish, catfish, channel, farmed, raw	1
26115000	Flounder, raw	• Fish, flatfish (flounder and sole species), raw	1
26119100	Herring, raw	• Fish, herring, Atlantic, raw	1
26121100	Mackerel, raw	• Fish, mackerel, Atlantic, raw • Fish, mackerel, Pacific and jack, mixed species, raw	1
26125100	Ocean perch, raw	• Fish, ocean perch, Atlantic, raw	1
26313100	Mussels, raw	• Mollusks, mussel, blue, raw	1
63123020	Grapes, American type, slip skin, raw	• Grapes, american type (slip skin), raw	1
27116400	Steak tartare (raw ground beef and egg)	• Beef, ground, 85% lean meat / 15% fat, raw • Egg, yolk, raw, fresh • Onions, raw • Fish, anchovy, european, canned in oil, drained solids	3

1.3 NOVA Manual Classification Coverage

The coverage of NOVA classification for FNDDS 2001-2017 is presented by Figure S3. For over 55% of the databases, NOVA classification relies on having a precise ingredient decomposition of food items, an information that is extremely uncommon. To note, the manual NOVA classification has been updated since the initial classification on FNDDS 2009-2010 used in [7]. For clarity, only for FNDDS 2009-2010 in Figure S3, we used the NOVA classification conducted

by Steele et al. in [7], which is the data-source that we used to train FoodProX. However, close to the convergence of this manuscript, Steele and colleagues have updated the manual classification for FNDDS 2009-2010, and propagated the labels in different cohorts by matching the food codes through the years. In the updated manual classification of FNDDS 2009-2010 the percentage of food items relying on ingredient decomposition is increased to 58.98% from 55.69%.

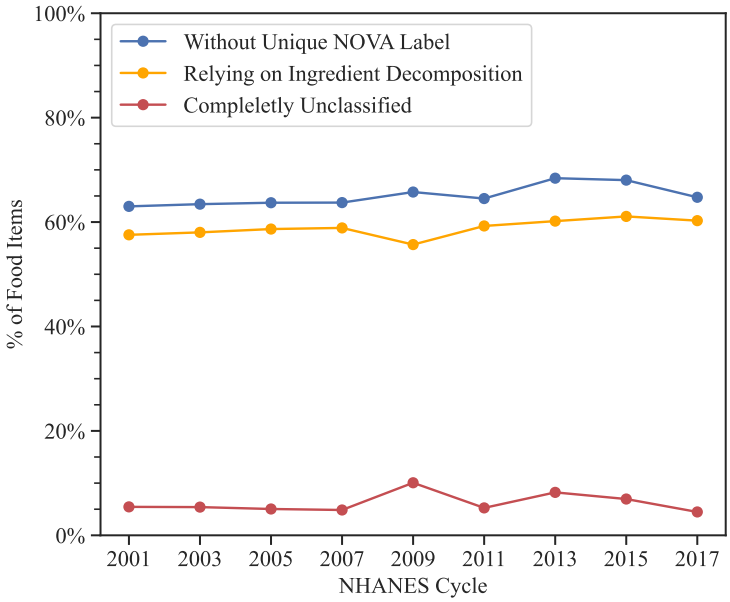

Figure SI. 3: **Manual NOVA Classification Coverage Over FNDDS 2001-2017.** On average 35% of the food items have a manual NOVA label without relying on ingredient decomposition.

1.4 Nutrient Panels

The large nutritional panel available for FNDDS 2009-2010 allowed us to train FoodProX with varying subsets of nutrients. The widest panel consists of 99 nutrients, including the flavonoid measurements developed for NHANES 2007-2010 (Table S2) [8]. Among these 99 nutrients, we further selected and trained on 62 nutrients common to NHANES 2001-2018 (Table S3), and 58 nutrients available in NHANES 1999-2018 (Table S4). Figures 1A-1C in manuscript describe the results on FNDDS 2009-2010, with 99 nutrients, while Figures 1D-1E are related to the analysis of FNDDS 2015-2016, therefore using a nutrient panel of 62 nutrients. For the epidemiological analysis in Section 3, leveraging data from 1999 to 2006, we opted for 58 nutrients.

With the goal to tackle branded products and the consumer space, we additionally trained on a subset of 12 nutrients contributing to FDA nutrition facts (Table S5), excluding calories and total amount of trans fatty acids, as the latter is not available in the original batch of 99

nutrients.

All nutrients were log-transformed and 0 values were substituted with e^{-20}, choice justified by the order of magnitudes spanned by nutrient concentrations in food [9].

Table SI. 2: 99 Nutrient Panel for NHANES 2007-2010

Nutrients			37 Additional Nutrients From FNDDS Flavoniod 2007-2010 Database	
Protein	Vitamin E (alpha-tocopherol)	6:0	Total flavonoids	Eriodictyol
Total Fat	Vitamin D (D2 + D3)	8:0	Cyanidin	Hesperetin
Carbohydrate	Cryptoxanthin, beta	10:0	Petunidin	Naringenin
Alcohol	Lycopene	12:0	Delphinidin	Total flavanones
Water	Lutein + zeaxanthin	14:0	Malvidin	Apigenin
Caffeine	Vitamin C	16:0	Pelargonidin	Luteolin
Theobromine	Thiamin	18:0	Peonidin	Total flavones
Sugars, total	Riboflavin	18:1	Total anthocyanidins	Isorhamnetin
Fiber, total dietary	Niacin	18:2	(+)-Catechin	Kaempferol
Calcium	Vitamin B-6	18:3	(-)-Epigallocatechin	Myricetin
Iron	Folate, total	20:4	(-)-Epicatechin	Quercetin
Magnesium	Vitamin B-12	22:6 n-3	(-)-Epicatechin 3-gallate	Total flavonols
Phosphorus	Choline, total	16:1	(-)-Epigallocatechin 3-gallate	Daidzein
Potassium	Vitamin K (phylloquinone)	18:4	Theaflavin	Genistein
Sodium	Folic acid	20:1	Thearubigins	Glycitein
Zinc	Folate, food	20:5 n-3	Theaflavin-3,3'-digallate	Total isoflavones
Copper	Vitamin E, added	22:1	Theaflavin-3'-gallate	
Selenium	Vitamin B-12, added	22:5 n-3	Theaflavin-3-gallate	
Retinol	Cholesterol	Fatty acids, total monounsaturated	(+)-Gallocatechin	
Carotene, beta	Fatty acids, total saturated	Fatty acids, total polyunsaturated	Total catechins (monomeric flavan-3-ols only)	
Carotene, alpha	4:0		Total flavan-3-ols	

Table SI. 3: 62 Nutrient Panel for NHANES 2001-2018 Cycles

Nutrients		
Protein	Vitamin E (alpha-tocopherol)	6:0
Total Fat	Vitamin D (D2 + D3)	8:0
Carbohydrate	Cryptoxanthin, beta	10:0
Alcohol	Lycopene	12:0
Water	Lutein + zeaxanthin	14:0
Caffeine	Vitamin C	16:0
Theobromine	Thiamin	18:0
Sugars, total	Riboflavin	18:1
Fiber, total dietary	Niacin	18:2
Calcium	Vitamin B-6	18:3
Iron	Folate, total	20:4
Magnesium	Vitamin B-12	22:6 n-3
Phosphorus	Choline, total	16:1
Potassium	Vitamin K (phylloquinone)	18:4
Sodium	Folic acid	20:1
Zinc	Folate, food	20:5 n-3
Copper	Vitamin E, added	22:1
Selenium	Vitamin B-12, added	22:5 n-3
Retinol	Cholesterol	Fatty acids, total monounsaturated
Carotene, beta	Fatty acids, total saturated	Fatty acids, total polyunsaturated
Carotene, alpha	4:0	

Table SI. 4: 58 Nutrient Panel for NHANES 1999-2018 Cycles

Nutrients		
Protein	Vitamin E (alpha-tocopherol)	14:0
Total Fat	Cryptoxanthin, beta	16:0
Carbohydrate	Lycopene	18:0
Alcohol	Lutein + zeaxanthin	18:1
Water	Vitamin C	18:2
Caffeine	Thiamin	18:3
Theobromine	Riboflavin	20:4
Sugars, total	Niacin	22:6 n-3
Fiber, total dietary	Vitamin B-6	16:1
Calcium	Folate, total	18:4
Iron	Vitamin B-12	20:1
Magnesium	Vitamin K (phylloquinone)	20:5 n-3
Phosphorus	Folic acid	22:1
Potassium	Folate, food	22:5 n-3
Sodium	Cholesterol	Fatty acids, total monounsaturated
Zinc	Fatty acids, total saturated	Fatty acids, total polyunsaturated
Copper	4:0	
Selenium	6:0	
Retinol	8:0	
Carotene, beta	10:0	
Carotene, alpha	12:0	

Table SI. 5: 12 Nutrient Panel for Branded Products

Nutrients
Protein
Total Fat
Carbohydrate
Sugars, total
Fiber, total dietary
Calcium
Iron
Sodium
Vitamin C
Cholesterol
Fatty acids, total saturated
Total Vitamin A = Retinol + Carotene, beta + Carotene, alpha + Cryptoxanthin, beta

1.5 Hierarchical Clustering of Foods according to Nutrients

Leveraging the 99 nutrient panel for FNDDS 2009-2010, we clustered all foods in an unsupervised fashion, using the dynamic tree cut algorithm [10]. The algorithm retrieves 20 clusters, annotated in the first column of the cluster-map shown in Figure S4. The nutrient-derived clusters are not good predictors of NOVA classes, both computationally-derived (Column 2) and manually-assigned (Column 3). We measure the mutual dependence between different clustering methodologies with adjusted mutual information (AMI), an adjustment of the classic mutual information (MI) to account for chance, that takes a value of 1 when two partitions are identical and 0 when the MI between two partitions equals the value expected by chance [11]. In particular, for the foods in the training data which were manually-assigned to NOVA 1, 2, 3, and 4,

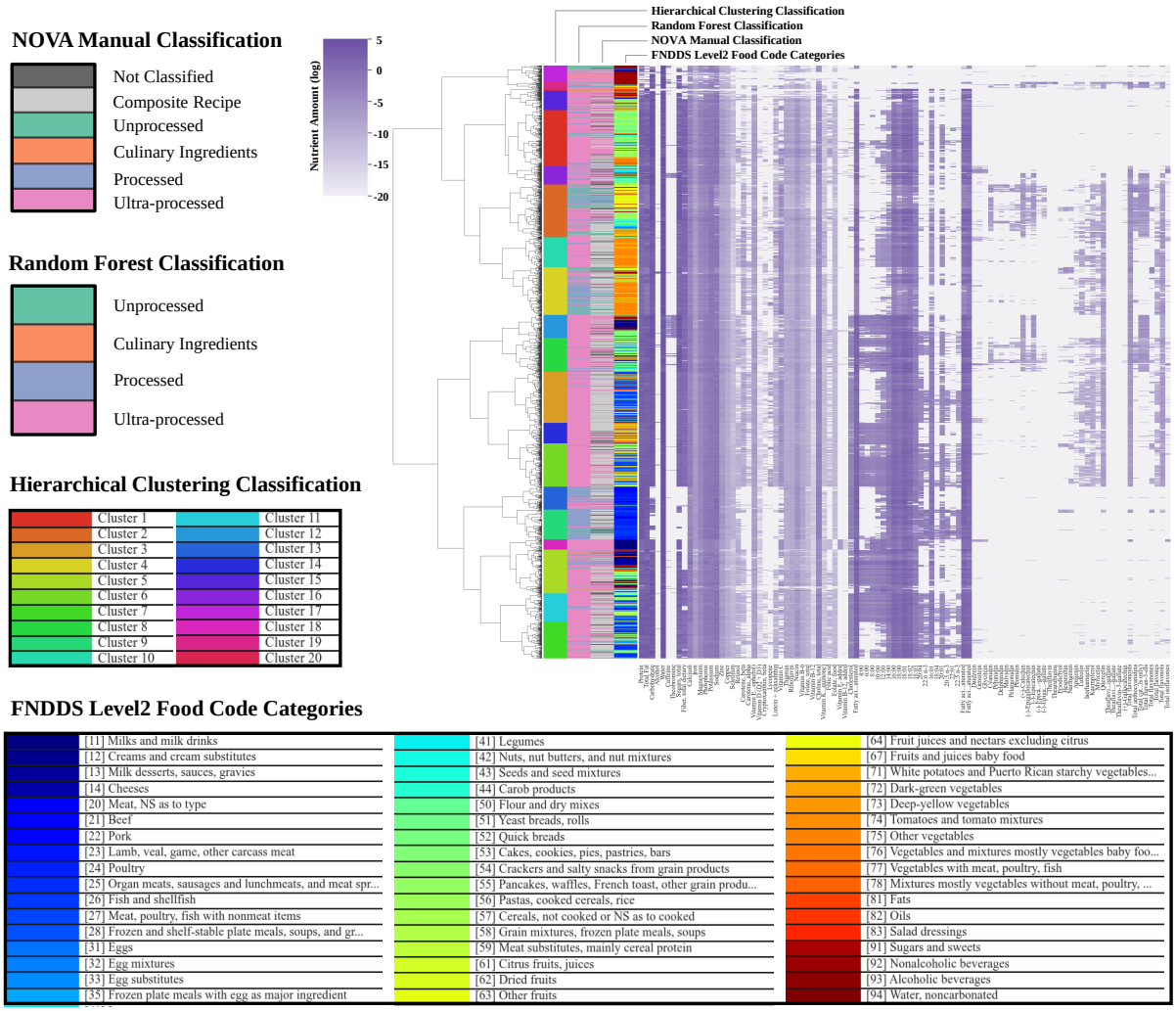

Figure SI. 4: **Hierarchical Clustering of Foods according to Nutrient Content.** We clustered all foods in FNDDS 2009-2010, each represented by a vector of 99 log-transformed nutrients. The cluster-map is annotated according to different classification strategies. In Column 1, starting from the left, each color corresponds to a cluster found by the dynamic tree cut algorithm. In Column 2 we report the predicted NOVA classes by FoodProX, while Column 3 encodes the manual labels used during the training phase of FoodProX. Finally, in Column 4 items are color-coded according to the first two digits of their FNDDS food codes.

we find AMI=0.12. Similarly, for the classes predicted by FoodProX we obtain AMI=0.14. The nutrient-based hierarchical clustering is more consistent with the first two digits of the FNDDS food codes (Column 4), capturing broad food groups as defined by the database (AMI=0.39). This result suggests that the performance of FoodProX is not merely induced by the classification of foods based on nutrient content, but FoodProX combines in a non-linear fashion the features of processing techniques (supervised information learned from NOVA manual labels), with food composition data (unsupervised information learned from FNDDS nutritional values). Similar results were obtained with 62, 58, and 12 nutrient panels described in Section S1.4.

2 Random Forest Classifier FoodProX and Food Processing Score $FPro$

2.1 Random Forest Classifier

FoodProX is based on a Random Forest Classifier whose hyper-parameters were chosen using the python *sklearn RandomizedSearchCV* function with a 5-fold stratified cross validation. In particular, the sampled search considered a number of trees between 200 and 2000, a maximum number of features equal to $\sqrt{}$ or to $\log_2$, and a maximum depth of the trees between 100 and 500. Over 100 random samples, the function picked a number of estimators equal to 200, a maximum number of features equal to $\sqrt{}$, and a maximum tree depth equal to 420 (currently used). Further runs of the random search found other combinations of parameters spanning the whole search intervals, suggesting an overall robust performance of the classifier, independently from the hyper-parameter tuning.

We evaluated the performance and stability of FoodProX over a 5-fold stratified cross validation of the labeled dataset (Figure S13C), with varying input resolution. In Figures S13A-H we show the ROC curves and Precision-Recall curves for each NOVA class, while in Tables S6A and S6B we report average and standard deviation of AUC and AUP over the 5 folds, and across the different nutrient panels, as reported in the manuscript. The high performance for different nutrient resolutions is encouraging, as for many foods we lack access to an extensive panel of nutrients.

To improve the performance of the classifier on new data and limit over-fitting, we retrained it using SMOTE [12] to correct for the unbalance in class representation, and created and ensemble voting system of 5 classifiers trained on 4/5 of the generated data. The predictions on unseen data are then calculated as the average of the 5 classifiers.

Table SI. 6: **AUC and AUP for the four NOVA classes**. For each NOVA class we report average and standard deviation of AUC and AUP over the stratified 5-folds, for 12, 62, and 99 input nutrients. We summarize the results across nutrient panels of different resolution in bold.

(A)

	NOVA 1	NOVA 2	NOVA 3	NOVA 4
Average AUC Nutrition Facts	0.981662	0.966348	0.967094	0.976772
Std AUC Nutrition Facts	0.003621	0.044898	0.010411	0.003432
Average AUC 62 Nutrients	0.980837	0.962878	0.970605	0.979866
Std AUC 62 Nutrients	0.00173	0.048586	0.00806	0.004876
Average AUC 99 Nutrients	0.978806	0.960406	0.971156	0.980085
Std AUC 99 Nutrients	0.002772	0.052151	0.007394	0.004237
Average AUC	0.980435	0.963211	0.969618	0.978908
Std AUC	0.0012	0.002437	0.001799	0.001513

(B)

	NOVA 1	NOVA 2	NOVA 3	NOVA 4
Average AUP Nutrition Facts	0.891112	0.756997	0.864605	0.990558
Std AUP Nutrition Facts	0.035105	0.169225	0.040566	0.001414
Average AUP 62 Nutrients	0.891971	0.736245	0.873025	0.991702
Std AUP 62 Nutrients	0.024412	0.185744	0.03922	0.002178
Average AUP 99 Nutrients	0.881419	0.74707	0.879123	0.991783
Std AUP 99 Nutrients	0.024092	0.193387	0.032601	0.001954
Average AUP	0.888168	0.74677	0.872251	0.991348
Std AUP	0.004785	0.008475	0.005952	0.00056

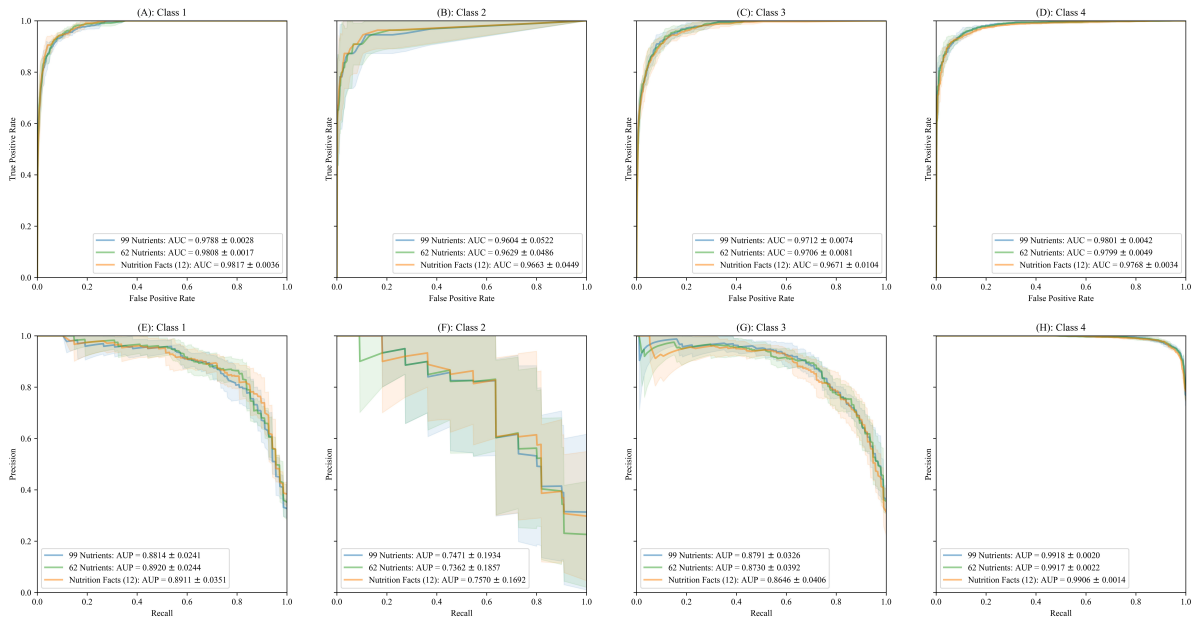

Figure SI. 5: **Random Forest Performance Over the 4 NOVA Classes**. For each NOVA class we evaluated the performance of the random forest classifier with a 5-fold cross validation. We observe similar performances for the classifiers trained with 99, 62 and 12 nutrients. In Panels A-D we show the ROC curves for each NOVA class, while panels E-F display the Precision-Recall curves.

2.2 Feature Importance

Inspired by the work of Parr et al. [13], we investigated how different nutrients contribute to the overall performance of FoodProX, i.e. their feature importance. The most popular way to assess feature importance in the random forest algorithm is the mean decrease impurity, measuring how effective a feature is at reducing uncertainty (classifiers) or variance (regressors) when building the decision trees. However, this methodology is not reliable when potential predictors vary in their scale of measurement or their number of categories [14]. We opted for permutation feature importance, a technique quantifying the relevance of each feature by permuting the specific input column and measuring the decrease in accuracy or R^2 compared to the baseline. This approach handles also the presence of collinear features, i.e. variables with some significant degree of linear or nonlinear dependence, that should be clustered and permuted together.

First, we modified the algorithm to work on a stratified 5-fold cross-validation, with data splits consistent with the cross-validation for the baseline model. Second, we addressed the high degree of collinearity of the nutrient space by removing all measurements like "Total Fat" or "Total flavonoids", as they represent straightforward linear combinations of other nutrients. On the reduced set of 85 nutrients we studied both rank correlation and feature dependence, i.e. the extent in which each feature can be predicted by the others through a random forest regressor. We cluster together features with $\rho_{Spearman} \geq 0.80$, and additionally, every nutrient well fitted

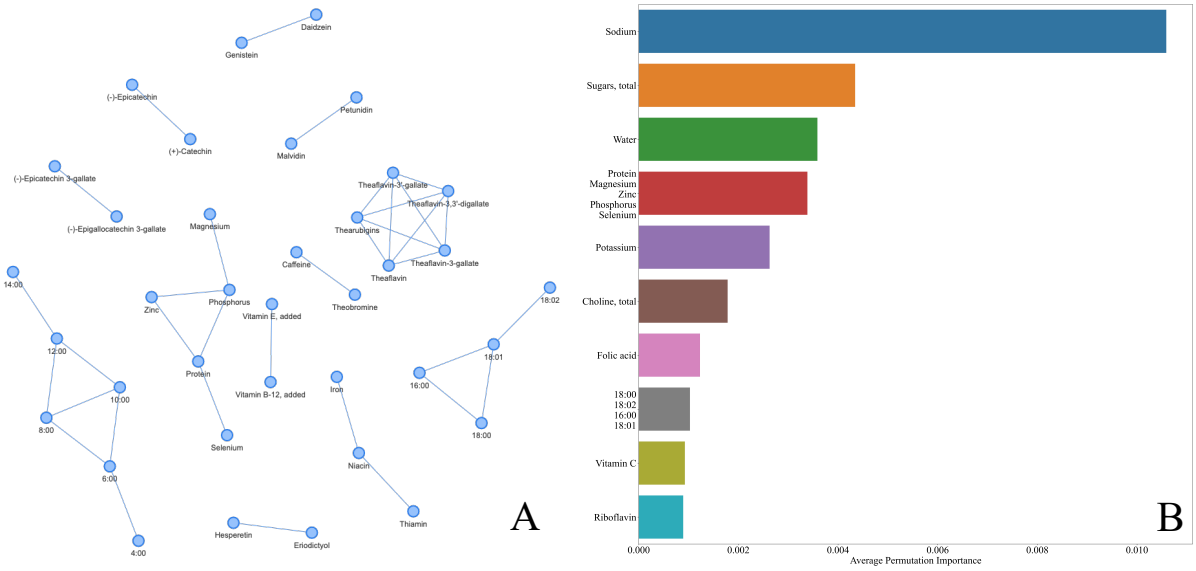

Figure SI. 6: **Permutation Feature Importance**. (a) Clusters of features with strong dependence. We cluster together features with $\rho_{Spearman} \geq 0.80$, and additionally, every nutrient well predicted by the remaining nutrient set through a random forest regressor is connected to each of the independent variables determining a drop ≥ 0.80 in the coefficient of determination. Each connected component determines a single cluster of features permuted at the same time, while isolated nutrients (not shown) are permuted on their own. (b) Top 10 most important nutrient clusters sorted by average permutation importance over 20 reshuffles.

by the random forest regressor is connected to each of the nutrients in the independent variable set that determines a drop ≥ 0.80 in the coefficient of determination (Figure S6A). Given the stochastic nature of the permutation feature importance, we repeated the estimation 20 times, and ranked the feature clusters according to their average drop in accuracy. In Figure S6B we report the top 10 most significant feature clusters, where only Sodium shows more predictive power than the other nutrients, suggesting that there is no single nutrient marker for food processing.

We further investigated the role of each nutrient i in determining the prediction for a selected food f, by using SHAP [15], over 5-fold stratified cross validation. The SHAP explanation method computes Shapley values from coalition game theory. The feature values of a data instance act as players in a coalition, and Shapley values indicate how to fairly distribute the "payout", i.e. the prediction, among the features. For each food f and NOVA class c, SHAP specifies the explanation as

$$p_c(f) = \langle p_c(f) \rangle + \sum_j SHAP\ value_j^c(f). \tag{1}$$

A player in the game could be also a group of features, as we previously investigated for the permutation feature importance analysis. However, due to high computational complexity, we focused on single nutrients, despite the presence of collinear features. In Figure S7 we show the top 10 nutrients in terms of $\langle |SHAP\ value_j^c(f)| \rangle$ for each NOVA class c. The magnitude

of the explanations varies significantly across NOVA classes, with only Sodium and Folic acid exhibiting distinct behaviors. For instance, when the amount of Folic acid is high, p_4 is expected to be high as well, while the other NOVA classes display lower probabilities. In case of Sodium, high values are likely to increase p_4 and decrese p_1 and p_2, while the processed class NOVA 3 displays mixed behaviors. Overall, we find that the ranking of SHAP values is in good overlap with Figure 6B.

To understand the statistical relevance of SHAP explanations, we introduced a new positive variable

$$relevance_c^i(f) = \frac{|SHAP\ value_i^c(f)|}{\sum_j |SHAP\ value_j^c(f)|}, \tag{2}$$

measuring to which extent nutrient i contributes to the superior limit of the absolute difference $|p_c(f) - \langle p_c(f) \rangle|$. In Figure S8, for each NOVA class c we display the violin-plot of $relevance_c^i$ for the top 10 nutrients in terms of median effect. Each violin-plot is compared with the quantiles of a maximum entropy null model for feature relevance, i.e., a Dirichlet multivariate distribution with marginal probability for each feature equal to a beta distribution with parameters $\alpha = 1$ and $\beta = n_{features} - 1$ [16]. The higher is the overlap between a violin plot and the null model ranges, the closer we are to a scenario lacking driving nutrients in determining the final class probability. In Figure S8, we observe a huge variability in the feature relevance of each nutrient, as captured by the shape of the violin-plots, evidence of the lack of strong driving signals in the model decision making. As expected, the largest nutrient contributions are found for NOVA 1 and NOVA 4, the two extreme processing classes.

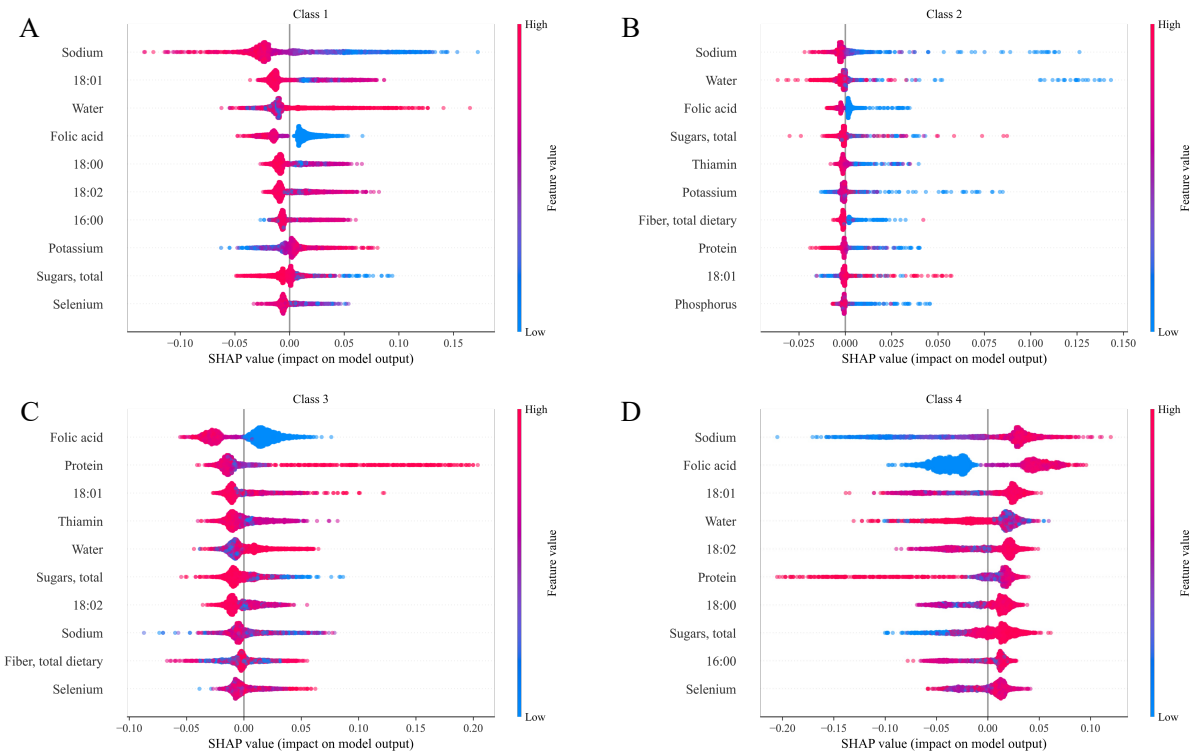

Figure SI. 7: **Shapley Values**. Top 10 nutrients in terms of average $\langle |SHAP\ value_j^c(f)| \rangle$ over all foods f, for class (a) NOVA 1, (b) NOVA 2, (c) NOVA 3, and (d) NOVA 4 (see Eq. S7). For each NOVA class c, the color scale correlates with the class probability p_c.

2.3 Food Processing Score $FPro$

The classifier probability space is a 4-D probability simplex that collects all vectors satisfying

$$\{\overrightarrow{p} \in \mathbb{R}^4,\ p_1 + p_2 + p_3 + p_4 = 1, \quad p_i \geq 0\ \forall i\}. \tag{3}$$

We define the processing score $FPro_k$ as the projection of any food $\vec{p}_k = (p_1^k, p_2^k, p_3^k, p_4^k)$ over the line going from the pure minimally-processed state $\overrightarrow{p}_{MP} = (1, 0, 0, 0)$ to the pure ultra-processed state $\overrightarrow{p}_{UP} = (0, 0, 0, 1)$, represented by the parametric equation

$$\overrightarrow{l(t)} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \tag{4}$$

equivalent to the explicit equation $p_1 = 1 - p_4$. The orthogonal projection of food $\overrightarrow{p}_k$ follows as the intersection between Eq. S4 and the plane passing through $\overrightarrow{p}_k$ and orthogonal to $\overrightarrow{l(t)}$,

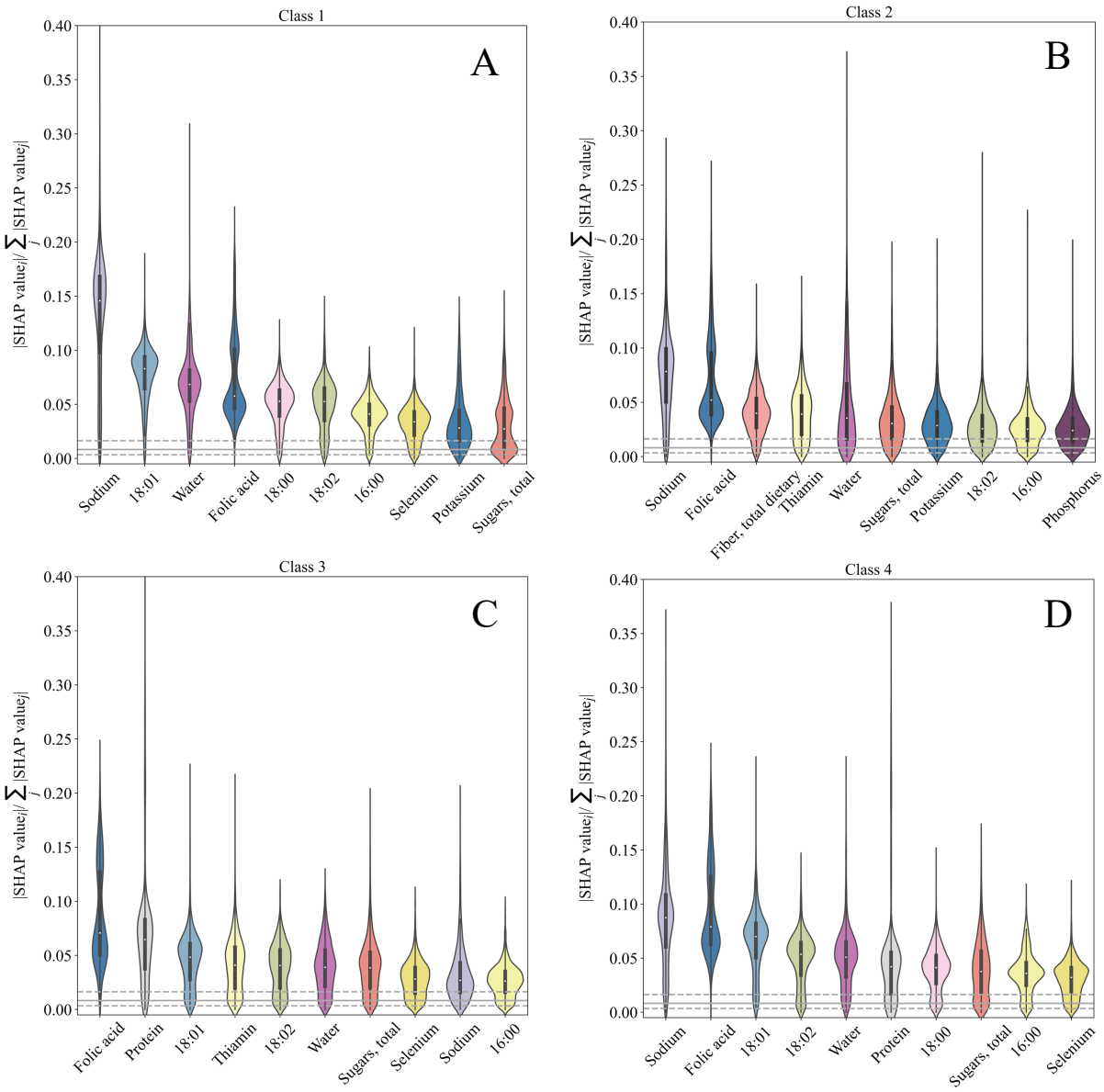

Figure SI. 8: **Relevance of Shapley Values**. Top 10 nutrients in terms of median $relevance_c^i(f)$ across all foods f, for class (a) NOVA 1, (b) NOVA 2, (c) NOVA 3, and (d) NOVA 4 (see Eq. S2). The grey dashed lines correspond to the quantiles Q_1 and Q_3 of a beta distribution with parameters $\alpha = 1$ and $\beta = n_{features} - 1$. The full grey line points to the median of the same null model.

i.e.,

$$-p_1 + p_4 + p_1^k - p_4^k = 0. \tag{5}$$

The parameter t^* satisfying Eqs. S4-S5 determines the processing score $FPro_k$ in Eq. 1. Of note, $FPro$ assigns a value around 0.5 for all NOVA 2/3 classes, which are then differentiated using the p_2/p_3 ratio.

We focused on the extreme classes NOVA 1 and NOVA 4, as by definition [17], they are the only ones with a clear "natural" ranking, ideal to define a processing scale. Indeed, NOVA 3 is not more processed than NOVA 2, they simply collect remarkably different items, according

to nutrient composition and consumed portion. However, both classes are more complex than NOVA 1, and less processed than NOVA 4. While it is true that a "pure" NOVA 2/3 item would have $FPro = 0.5$, in real-world data, this is an uncommon scenario, so the extent of the residual probabilities p_1 and p_4 measures if a product is leaning towards the minimally processed or the ultra-processed extreme.

2.4 Validation of $FPro$ in Different FNDDS Editions

We investigated the relation of $FPro$, trained on FNDDS 2009-2010 with a 62 nutrient panel, and NOVA manual classification in other editions of FNDDS, to control for any potential over-fitting and validate the performance of our algorithm on new foods, or foods whose nutrient content has changed over time. In particular, in Figure S9 we show the results for FNDDS 2015-2016, so far, the USDA database for dietary studies with the highest number of food items. All the four classes of manually labeled items correspond to well localized and distinguishable distributions of $FPro$.

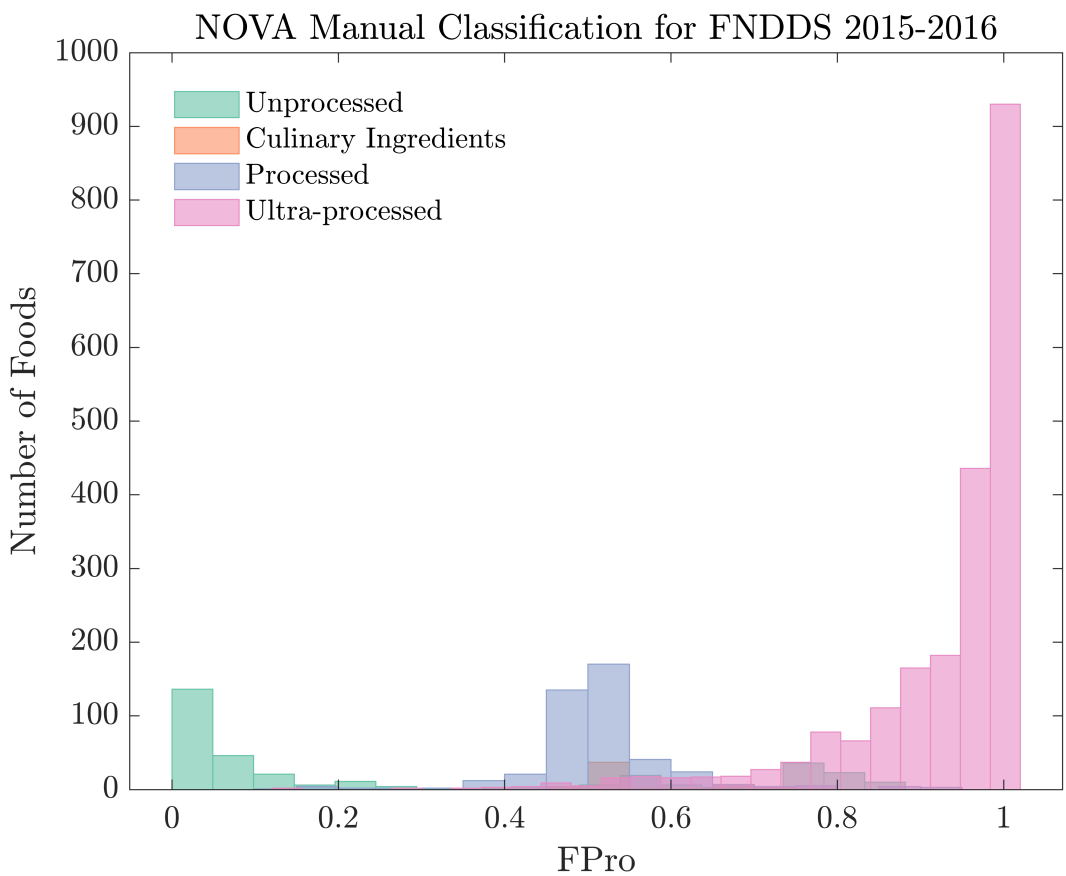

Figure SI. 9: **Food Processing Score for NOVA Manual Labels in 2015-2016**. Variability of $FPro$ (trained on FNDDS 2009-2010) within manual NOVA classes for FNDDS 2015-2016.

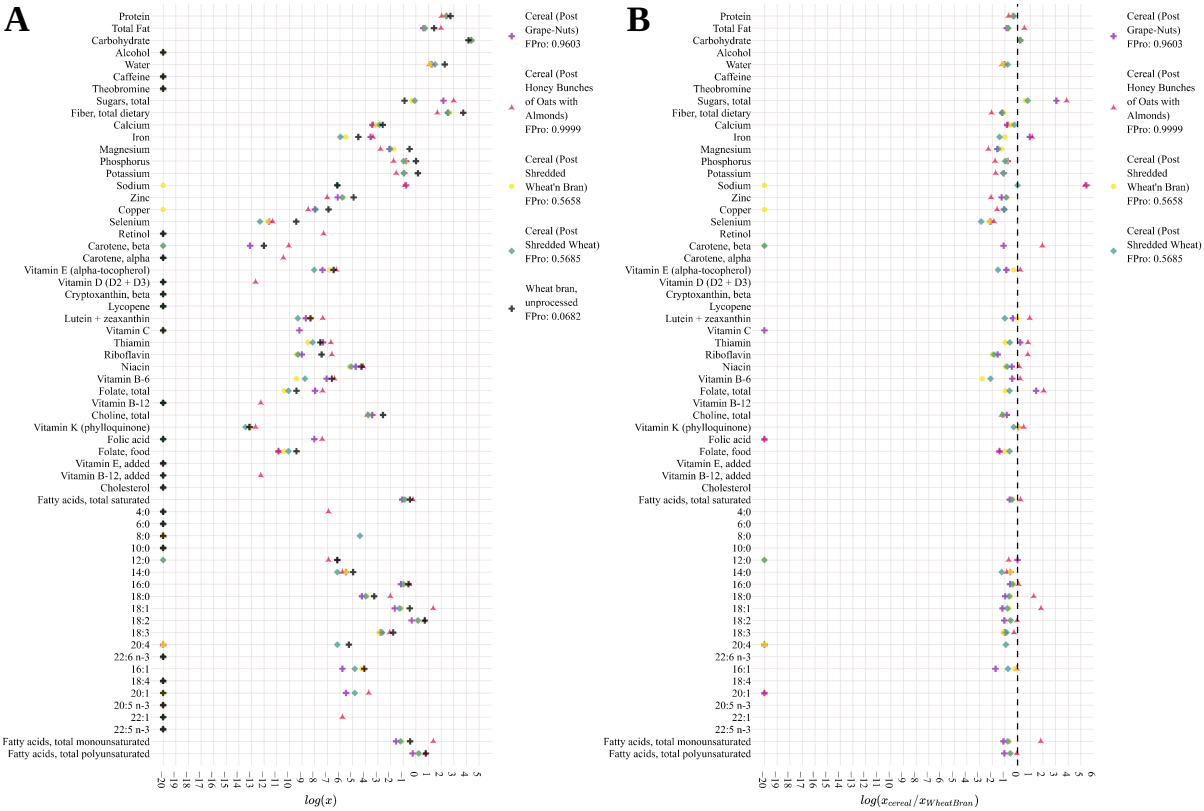

Figure SI. 10: **Nutrient Profiles of Post Cereals compared to "Wheat bran, unprocessed"**. (a) All 62 nutrients measured in FNDDS 2015-2016 are shown in log-scale, and with varying color and marker depending on the food item. (b) We rescale the nutrient profile of each cereal by the corresponding value per 100 grams found in unprocessed wheat bran, and plot them in log-scale. The black dashed line corresponds to 1, i.e. identical nutrient content.

2.5 Case Study on Post Cereals.

To further investigate the interplay between nutrient patterns and FPro, we collected the 62 nutrient profiles describing each Post cereal highlighted in Figure 2, and compared them with the nutritional values for 100 grams of unprocessed wheat bran (FPro=0.0682). We chose this specific ingredient as the Post Shredded Wheat 'N Bran contains whole grain wheat, wheat bran, and the antioxidant Butylated hydroxytoluene [18], and we tried to find any potential matching item in FNDDS 2015-2016. In Figure S10, we compare the nutrient values in absolute terms (Panel A), or as ratio with their counterpart in wheat bran (Panel B). Interestingly, we observe how the pattern of alterations involves all nutrients, increasing the level of FPro even for simple products like Post Shredded Wheat 'N Bran, as its nutrient profile is not characteristic of any natural ingredient per 100 grams, but it corresponds to a mildly processed food (FPro=0.5658).

2.6 Source of Food

The location where a food was prepared, as well the origin of the ingredients, could be indicative of its degree of processing. To investigate this hypothesis we used the variable DR1FS in NHANES, corresponding to the question "Where did you get (this/most of the ingredi-

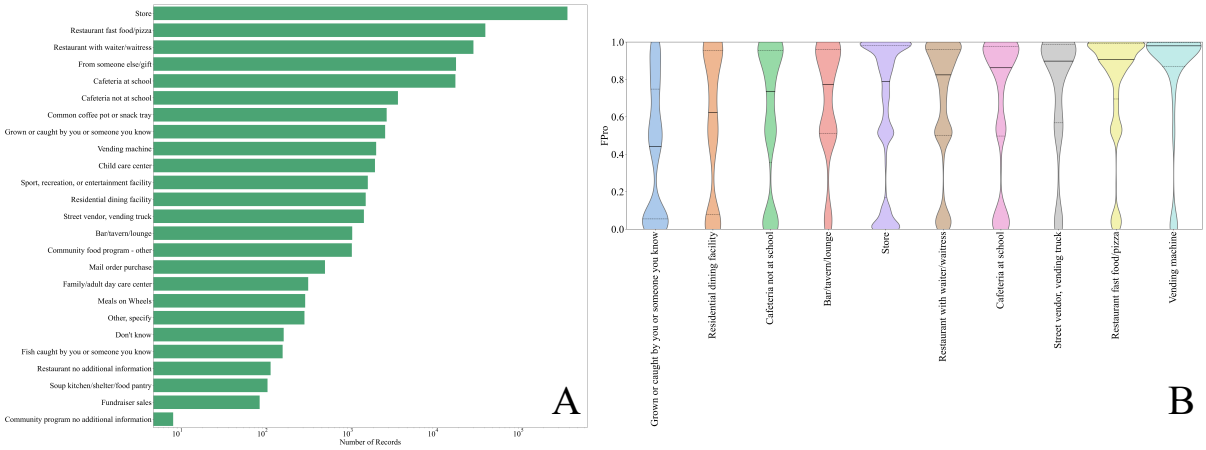

Figure SI. 11: **FPro stratified by food source in NHANES 2003-2005.** (a) Food sources reported in the cohorts, sorted in descending order by number of records. (b) Violin plots of 10 selected food sources ranked in increasing order of median $FPro$. We annotate median values with full lines, and lower quartiles and upper quartiles with dashed lines.

ents for this)?", and available for the years 2003-2018 (`https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/DR1IFF_D.htm#DR1FS`). By leveraging our analysis of the merged NHANES cohorts between 1999 and 2006 (Section S3.2), we were able to stratify FPro by food source for two cycles (2003-2004, 2005-2005). In Figure S11A we report the 25 different types of food sources found in the population, sorted by decreasing number of records. As expected, "Store" contributes to the majority of the records, followed by "Restaurant fast food/pizza". In Figure S11B we selected 10 of the most popular food sources, and visualize the related distributions of FPro in increasing order of median. Overall, FPro is significantly different across source categories, as quantified by the Kruskal-Wallis H-test (p-value $< 10^{-15}$), suggesting that the overlap between the source categories (driven by foods with multiple origins), while present, is limited. We also compared each pair of source categories with the Mann-Whitney U rank test, finding that (Cafeteria not at school, Bar/tavern/lounge), (Cafeteria at school, Bar/tavern/lounge), (Store, Bar/tavern/lounge), and (Restaurant with waiter/waitress, Bar/tavern/lounge) do not survive multiple testing with Bonferroni correction ($\alpha = 0.01$), indicating continuous distributions with equal medians. Additionally, to control for "overpowering", i.e. the scenario in which large samples almost surely determine a statistically significant outcome, we estimated the effect size r following [19]. Across the 45 combinations of food sources, 23 show $r \geq 0.1$ (*small* effect), 6 have $r \geq 0.3$ (*medium* effect), and the pair ("Grown or caught by you or someone you know", "Vending machine") is characterized by the largest effect size with $r = 0.5798$.

3 Individual Diet Processing Scores $iFPro$ and Exposome

3.1 Individual Processing Score $iFPro$

To capture the extent of processing in individuals' diet we focus on two weighting schemes: a calorie-based score (Eq. 2), and a gram-based score,

$$iFPro^{j}_{WG} = \sum_{k}^{D_j} \frac{w^j_k}{W^j} FPro_k, \tag{6}$$

where D_j is the number of dishes consumed by individual j, W^j is her total amount of food in grams, and w^j_k is the amount of grams consumed for each food item (excluding water consumption, see Section S3.4).

3.2 Population Characteristics

NHANES captures a variety of information ranging from demographic and dietary intake, to lab and physical examinations. This wealth of information is compiled into hundreds of publicly available data files, that altogether provide over 1,000 variables.

To investigate the relation between $iFPro$ and health, we focused on NHANES 1999-2006 exposome and phenome database, a harmonized dataset created by Patel et al. in [20], merging 255 data files from four cycles of NHANES, for a total of 41,474 individuals and 1,191 variables. The summary statistics for $iFPro_{WC}$ and $iFPro_{WG}$, characterizing the 20,047 adults (18+) in the cohort, are presented in Tables S7 and S8. All predictions are calculated with the 58 nutrient panel in Table S4.

NHANES follows the well-established two-step 24HRs dietary recall interviews to sample the dietary intake of the American population [21, 22]. The first step is done in-person with a dietitian interviewing each applicant, while ensuring the highest quality of dietary recall over the past 24HRs. The second step consists of a phone interview within 3-10 days to capture a second dietary recall [23]. The multiple 24HRs dietary recalls have proved to be an effective method in the assessment of trends over the dietary intakes of individuals [24]. For all individuals who completed two-day dietary recalls (in-person and phone interview) we calculated a daily average $iFPro$, while for the remaining participants we used just the data from in-person interview. The relevance of each individual for population statistics is based on survey weights [25, 26].

Table SI. 7: **Population characteristics for $iFPro_{WC}$**

	mean	$iFPro_{WC}$				
		0.0 – 0.2	0.2 – 0.4	0.4 – 0.6	0.6 – 0.8	0.8 – 1.0
Counts Subjects (n = 20,046)		4	63	1487	11180	7313
Age — mean±SE	45.28±0.3	58.01±6.16	47.2±3.06	51.4±0.64	47.36±0.37	41.3±0.3
Poverty Income Ratio — mean±SE	2.99±0.04	2.58±0.98	2.92±0.3	2.97±0.07	3.06±0.04	2.88±0.05
Calories Consumed — mean±SE	2189.13±10.36	534.25±162.24	1359.03±184.71	1772.85±29.81	2143.11±13.16	2329.7±15.07
BMI — mean±SE	28.13±0.1	26.48±1.63	29.33±1.33	27.45±0.28	27.92±0.1	28.53±0.13
female — count (%)		4 (1)	41 (0.65)	879 (0.59)	5798 (0.52)	3763 (0.51)
white — count (%)		0 (0)	20 (0.32)	663 (0.45)	5389 (0.48)	3525 (0.48)
black — count (%)		0 (0)	21 (0.33)	227 (0.15)	2160 (0.19)	1810 (0.25)
mexican — count (%)		2 (0.5)	15 (0.24)	443 (0.3)	2737 (0.24)	1460 (0.2)
other hispanic — count (%)		0 (0)	6 (0.1)	86 (0.06)	467 (0.04)	270 (0.04)

Table SI. 8: **Population characteristics for $iFPro_{WG}$**

	mean	$iFPro_{WG}$				
		0.0 – 0.2	0.2 – 0.4	0.4 – 0.6	0.6 – 0.8	0.8 – 1.0
Counts Subjects (n = 20,046)		182	2699	7286	7230	2650
Age — mean±SE	45.28±0.3	52.81±1.55	54.78±0.61	49.11±0.38	41.5±0.34	34.71±0.3
Poverty Income Ratio — mean±SE	2.99±0.04	3.01±0.15	3.13±0.05	3.15±0.05	2.92±0.04	2.56±0.07
Calories Consumed — mean±SE	2189.13±10.36	1490.34±70.36	1821.13±19.67	2134.69±14.34	2342.11±17.73	2356.4±30.02
BMI — mean±SE	28.13±0.1	28.08±0.82	27.47±0.15	27.76±0.13	28.37±0.1	29.14±0.19
female — count (%)		112 (0.62)	1618 (0.6)	3851 (0.53)	3594 (0.5)	1310 (0.49)
white — count (%)		107 (0.59)	1682 (0.62)	3790 (0.52)	2921 (0.4)	1097 (0.41)
black — count (%)		20 (0.11)	294 (0.11)	1189 (0.16)	1860 (0.26)	855 (0.32)
mexican — count (%)		37 (0.2)	529 (0.2)	1731 (0.24)	1873 (0.26)	487 (0.18)
other hispanic — count (%)		9 (0.05)	100 (0.04)	295 (0.04)	320 (0.04)	105 (0.04)

3.3 Correlation between $iFPro_{WG}$, $iFPro_{WC}$, and HEI-15

While we find an overall agreement in the population ranking determined by $iFPro_{WG}$ and $iFPro_{WC}$ ($\rho_{Spearman} = 0.7029$, Figure S12A), the two measures show significantly different patterns in epidemiological associations, in particular regarding chemical exposures (Figure S20). Indeed, a weight-based index could capture complex dietary patterns arising from consumption of highly processed beverages such as zero-calorie soft drinks, or any type of food contaminants whose amount is independent from the provided calories.

We compared $iFPro_{WG}$ and $iFPro_{WC}$ with the HEI-2015, a score measuring the alignment of an individual's diet with the national dietary guidelines, ranging from 0 (no alignment) to 100 (complete alignment) [27]. We followed the National Cancer Institute (NCI) to calculate HEI-15 for NHANES participants [28]. As expected, we observe negative correlations emerging (Figures S12B and C), with both $iFPro_{WG}$ ($\rho_{Spearman} = -0.4862$), and $iFPro_{WC}$ ($\rho_{Spearman} =$

-0.5575).

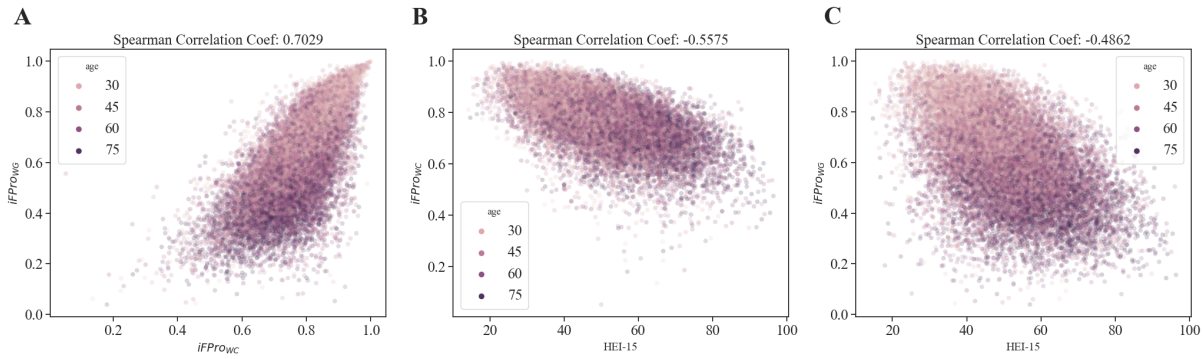

Figure SI. 12: **Spearman's Correlation coefficients Between Two Variations of $iFPro$ and Healthy Eating Index 2015.**

Individual $\mathcal{A}$ 47 age

Day 1

Food Code		Calories (kcal)	FPro	Grams
63107010	Banana (raw)	121	0.00	136
72201100	Broccoli (raw)	30	0.02	88
75115000	Mushrooms (raw)	30	0.17	138
74101000	Tomatoes (raw)	71	0.07	394
75114000	Mixed greens salad	31	0.09	182
24124120	Fried Chicken breast	232	0.56	120
22101220	Fried Pork chop (lean)	189	0.63	117
56205008	Rice (cooked white)	663	0.69	513
61210250	Orange juice	211	0.00	450
92410550	Soft drink (caffeine add)	68	1.00	185
92302000	Tea (unsweetened)	22	0.05	2131

Day 2

63107010	Banana (raw)	242	0.00	272
72201100	Broccoli (raw)	17	0.00	50
75115000	Mushrooms(raw)	18	0.17	79
74101000	Tomatoes (raw)	18	0.07	101
75113000	Lettuce (raw)	8	0.02	55
42114130	Pistachio nuts (roasted , salt added)	82	0.53	15
14109010	Swiss Cheese	9	0.50	2
83106000	Italian dressing	141	0.91	59
56205008	Rice (cooked white)	268	0.69	207
28355480	Seafood soup with vegetables	185	0.69	442
41420300	Soy sauce	17	0.54	32
26158010	Tilapia (baked or broiled)	204	0.50	136
61210250	Orange juice	845	0.00	1798
93101000	Beer	310	0.51	720

Individual $\mathcal{B}$ 48 age

Day 1

Food Code		Calories (kcal)	FPro	Grams
63109010	Cantaloupe (raw)	27	0.02	78
14010100	Cheddar cheese	32	0.51	9
63223020	Strawberries (raw)	9	0.00	28
92111010	Coffee (decaffeinated)	0	0.02	192
92101000	Coffee (regular)	7	0.00	607
57230000	Cereal (Grape-Nuts)	139	1.00	39
53242000	Cookie (sugar wafer)	141	0.99	28
54337000	Cracker	80	0.99	19
25210110	Hot dog (frankfurter)	184	0.99	57
27510560	Hamburger (with mayonnaise)	655	1.00	290
51150000	Bread roll (white, soft)	145	1.00	52
26137190	Salmon (smoked)	15	0.50	13
71201020	Potato chips	59	0.91	12
71401030	French fries	306	0.95	98
75506010	Mustard	7	0.54	10
91705030	Kit Kat	78	1.00	15
11422000	Yogurt (flavored and lowfat milk)	156	0.99	184
92410510	Soft drink (caffeine free)	49	0.99	122

Day 2

63201010	Blackberries (raw)	11	0.01	26
63219020	Raspberries (raw)	10	0.00	19
92101000	Coffee (regular)	7	0.00	696
57230000	Cereal (Grape-Nuts)	312	1.00	87
53209000	Cookie (chocolate)	206	1.00	44
53112100	Cake	382	1.00	141
58106225	Pizza (cheese)	622	1.00	234
11423000	Yogurt (flavored and nonfat milk)	149	0.84	184

Figure SI. 13: **Dietary Recalls for Individual $\mathcal{A}$ (SEQN 68484) and Individual $\mathcal{B}$ (SEQN-ID 59440) annotated in Figure 2.**

3.4 Water Consumption in NHANES

To calculate $iFPS_{WG}$ (Equation S6), we removed four food codes regarding water consumption, as their reporting through different NHANES cycles showed inconsistencies. Indeed, we

noticed that "Water as an ingredient" stopped being tracked since NHANES 2011-2012 (Figure S14 A), and the consumption of tap and bottled water started being recorded since NHANES 2003-2004 (Figures S14B and C). These inconsistencies would have affected our analysis of the pooled cohorts.

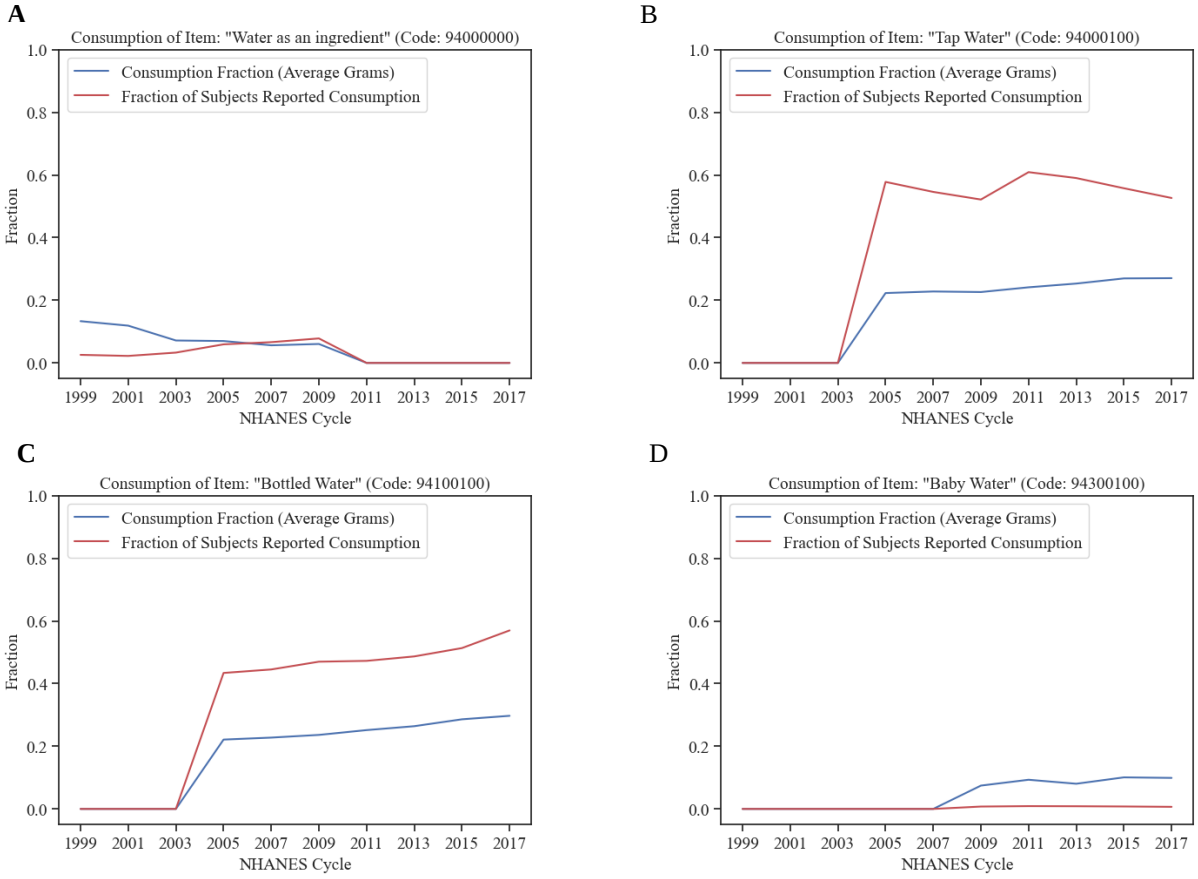

Figure SI. 14: **Water Consumption in Survey Data**. (**a**) "Water as an ingredient" is no longer tracked since NHANES 2011-2012. (**b-c**) Bottled and tap water have been tracked since NHANES 2005-2006, hence this might introduce inconsistency when combining NHANES 1999-2004 cohorts with their succeeding cohorts. (**d**) The consumption of baby water has been tracked since NHANES 2009-2010.

3.5 Relation between $iFPro_{WC}$ and WWEIA Food Categories

By leveraging our calculation of $iFPro_{WC}$ (Eq. 2) over the merged NHANES cohorts between 1999 and 2006 (Section S3.2), we investigated the relation between trends in $iFPro_{WC}$ and fraction of consumed calories in each of the What We Eat in America (WWEIA) food categories [29].

First, for each individual j we calculated the total fraction of calories contributed by food category g,

$$FC_g^j = \sum_k^{D_j} \frac{c_k^j}{C^j} \delta(WWEIA(k), g), \tag{7}$$

where D_j is the number of dishes consumed by individual j, C^j is the daily total amount of

consumed calories, c_k^j is the amount of calories contributed by each food item, and δ indicates the Kronecker delta, whose value is 1 when food k belongs to food category g, and otherwise 0.

We proceeded in calculating the Spearman's Rank Correlation between $\{iFPro_{WC}^j\}$ and $\{FC_g^j\}$ for each WWEIA class g across the cohort. Once accounted for multiple testing with Bonferroni correction ($\alpha = 0.01$), we found a total of 64 WWEIA categories anti-correlated with $iFPro_{WC}$, and 40 positively correlated (Figure S15).

We further investigated the representation of WWEIA categories in the first quintile Q_1 of $iFPro_{WC}$ ($iFPro_{WC} \leq 0.6908$), compared to the last quintile Q_5 ($iFPro_{WC} \geq 0.8585$). These two sub-populations capture 4,010 individuals each, representing the most divergent dietary patterns in terms of ultra-processed food. The caloric consumption of each WWEIA category across the two sub-groups was tested with the Mann-Whitney U rank test, finding a total of 103 categories significantly changing from Q_1 to Q_5, once corrected for multiple testing (Bonferroni method with $\alpha = 0.01$). To control for "overpowering", as explained in Section S2.6, we calculated the effect size r following [19], and ranked all significant WWEIA categories accordingly (Figure S16A). Overall, the consumption of 21 food groups change between Q_1 and Q_5 with effect size ≥ 0.1, the baseline for small effect sizes. Among the biggest effect sizes we find "Soft drinks" ($r = 0.4660$), with average caloric fraction 5.61 bigger in Q_5 compared to Q_1, and "Bananas" ($r = 0.2728$), with average caloric fraction 14.07 bigger in Q_1 compared to Q_5 (Figure S16B).

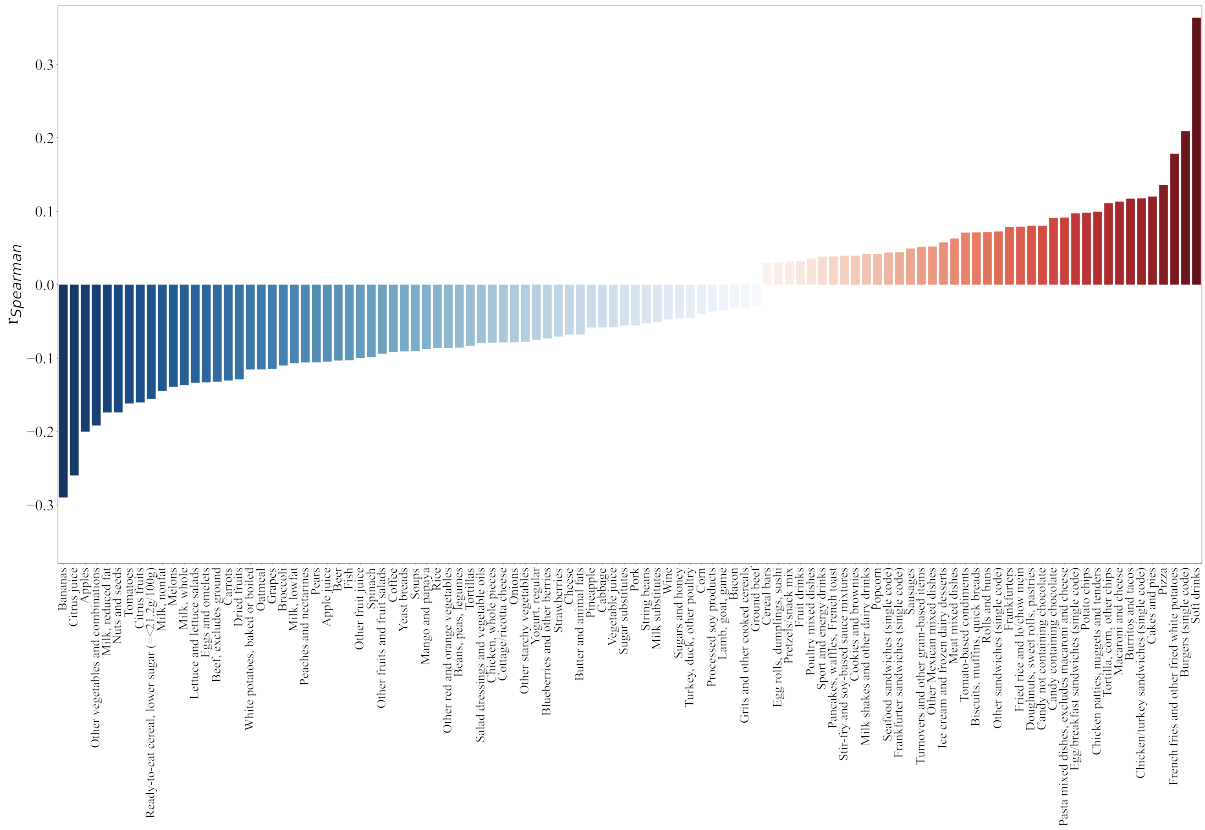

Figure SI. 15: **Spearman's Rank Correlation between $iFPro_{WC}$ and the caloric fraction contributed by each WWEIA category across NHANES individuals**. 64 WWEIA categories significantly anti-correlate with $iFPro_{WC}$ (blue), while 40 are positively correlated (red). The height of each bar is proportional to the correlation value, and all coefficients are sorted from the strongest anti-correlation to the strongest correlation.

4 Environment-Wide Association Study

Inspired by [30], we performed an Environment-Wide Association Study (EWAS) on the merged NHANES 1999-2006 cohort, to identify environmental factors and disease-related phenotypes associated with $iFPro_{WC}$, $iFPro_{WG}$, and fraction of calories contributed by manual NOVA 4. To do so, we collected data for 45 exposure modules in [20], and we further added one variable predicting diabetes according to fasting glucose levels $\geq$126 mg/dL, as advised by the American Diabetes Association [31], two variables predicting metabolic syndrome [32], two assessments of the Framingham Risk Score [33, 34], and the ACC/AHA Risk Score [35], quantifying the 10-year risk of non-fatal myocardial infarction (MI), congestive heart disease (CHD) death, or fatal or nonfatal stroke.

The variables are broadly categorized in two panels (Figure S17A): a health panel, gathering variables describing the overall health of the individuals, from biological age and nutrient biomarkers, to disease phenotypes, and a chemical panel, where we group all chemical exposures measured in blood or urines, linked to pesticides, contaminants, and processing chemical

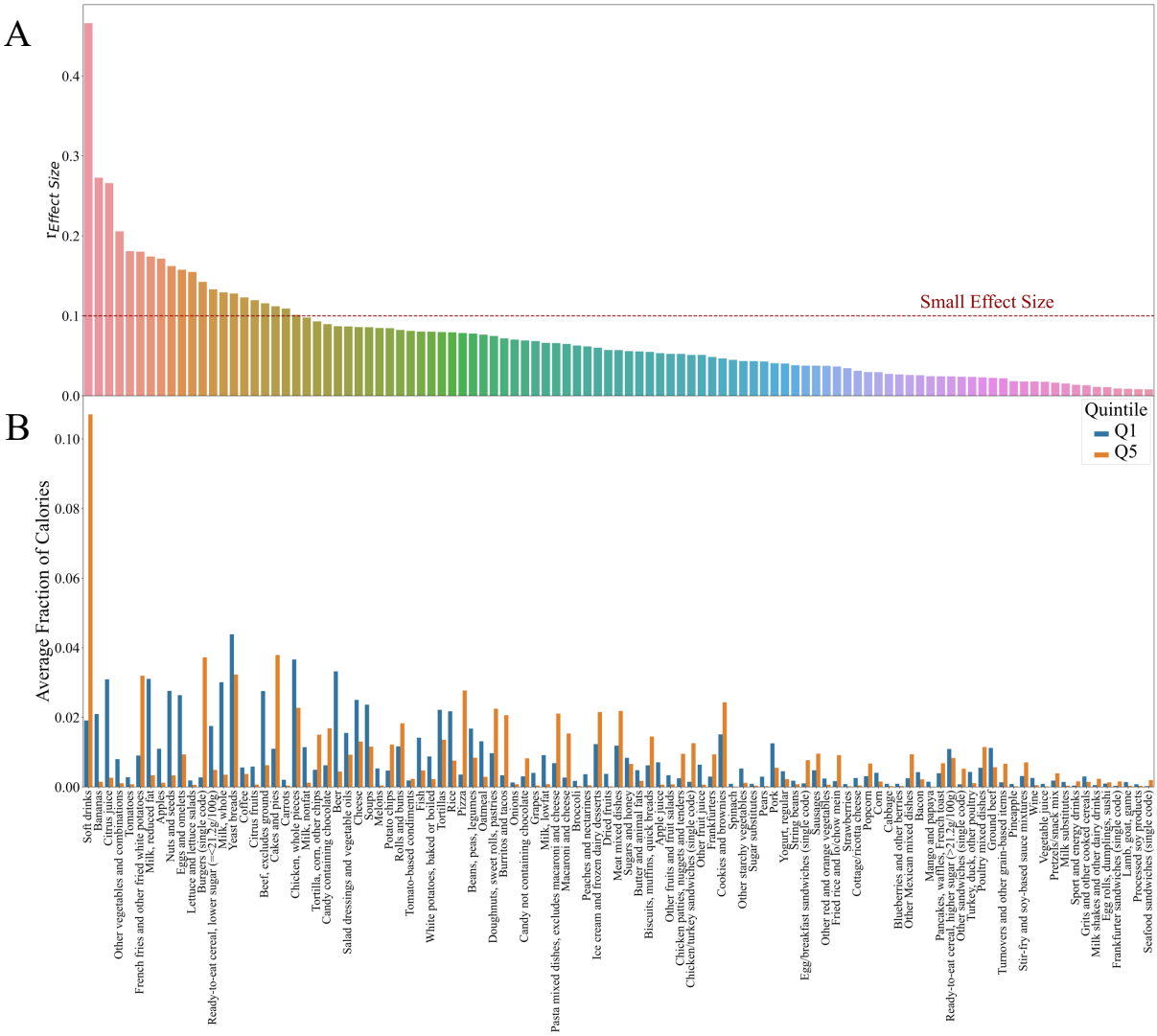

Figure SI. 16: **Comparison of $iFPro_{WC}$ Extreme Quintiles Q_1 and Q_5.** For individuals in the first quintile Q_1 and in the last quintile Q_5 of $iFPro_{WC}$ we compare the caloric fraction contributed by each WWEIA category. In (a) we estimate the effect size r following [19], where $r \approx 0.1$ is considered a *small* effect, $r \approx 0.3$ a *medium* effect, and $r \approx 0.5$ a *large* effect. The WWEIA food categories surviving multiple testing with Bonferroni correction ($\alpha = 0.01$) are shown in decreasing order of effect size. In (b), following the same order, we show the average fraction of calories in Q_1 and Q_5.

byproducts.

To select the most robust signal, we studied only variables measured in at least two cycles of NHANES. To quantify the statistical associations, we employed the Survey-weighted Generalized Linear Model, and in particular, linear regression to predict continuous variables, and logistic regression for categorical (Figure S17B). All models were adjusted for age, sex, ethnicity, Body Mass Index (BMI), total-caloric intake, and estimated Socioeconomic Status (SES), as provided by NHANES. We employed the 'survey' statistical package in R to account for the complex survey design of NHANES [26]. We further filtered all categorical and continuous variables lacking a minimum sample size to perform regression analysis. In particular, for continuous variables we considered a ratio between number of covariates and number of data points $\leq 1/50$, while for categorical we applied a similar threshold for the ratio between number

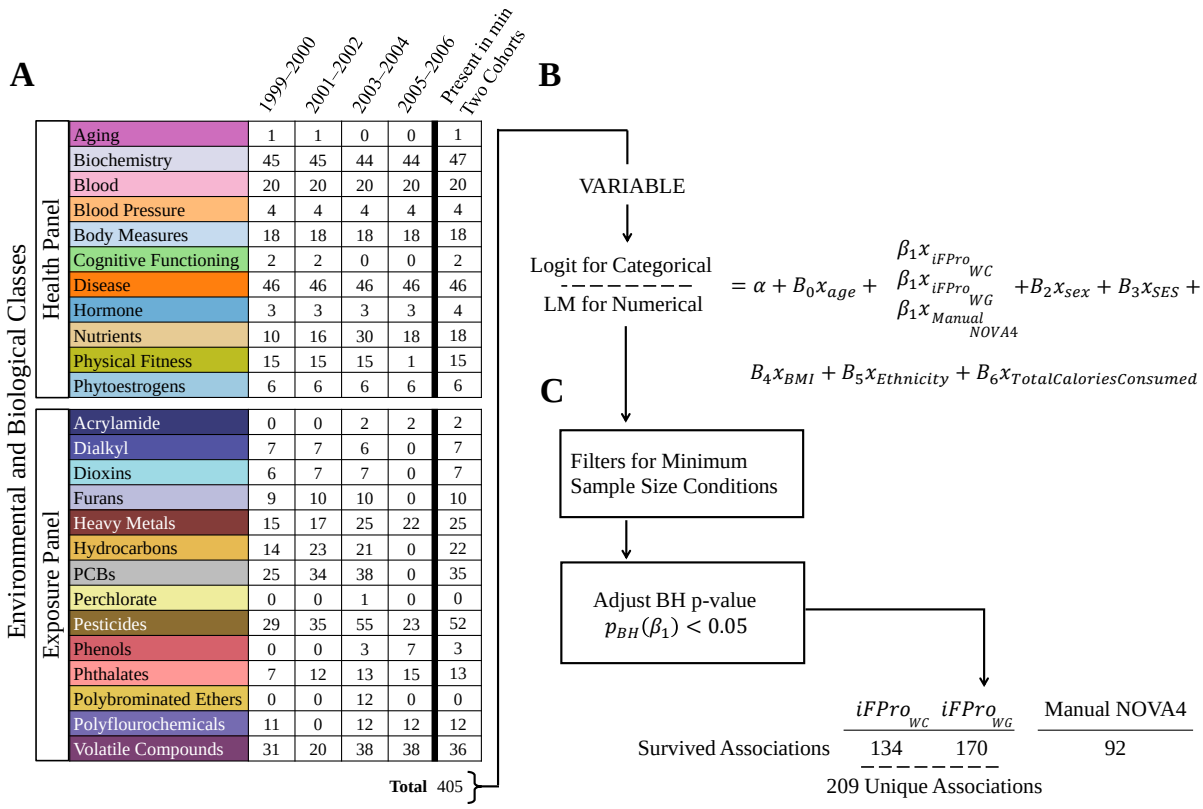

Figure SI. 17: **Overview of the EWAS pipeline on NHANES 1999-2006. (a)** For the selected health phenotypes and chemical exposures provided in each cohort of NHANES, we kept only those variables present in at least two cohorts. **(b)** We investigated possible associations between the selected 405 variables in the combined NHANES 1999-2006 cohort using linear regression for continuous variables and logistic regression for categorical variables. **(c)** To account for false discovery rate, we adjusted the p-value of β_1 using Benjamini-Hochberg method with $\alpha = 0.05$.

of covariates and number of data points in the smallest category (Figure S17C).

All continuous variables were transformed using Box-Cox transformation or logit function (applied to Framingham and ACC/AHA scores) to stabilize the variance and improve the validity of measures of association [30]. We then standardized all continuous variables, to bring their effect sizes on similar scale. For multiple linear regression, we used fully standardized regression coefficients, indicating how many standard deviations of change in the dependent variable are associated with one standard deviation increase in the independent variables. For logistic regression, we opted for a partial standardization, acting only on the continuous independent variables, as we wanted to keep a straightforward interpretation of the relation between one standard deviation increase in the Box-Cox transformed $iFPro$ and increase/decrease in disease odds [36].

To account for false discovery rate, we adjusted the p-values corresponding to each score using Benjamini-Hochberg method with $\alpha = 0.05$. Overall, we find 214 significant tests across the three methodologies, with $iFPro_{WC}$ at 134, $iFPro_{WG}$ at 170, and manual NOVA 4 at 92. The summary of the analysis is presented in Figures S18-S21. A comparison with literature

results based on manual NOVA 4 is reported in Table S9.

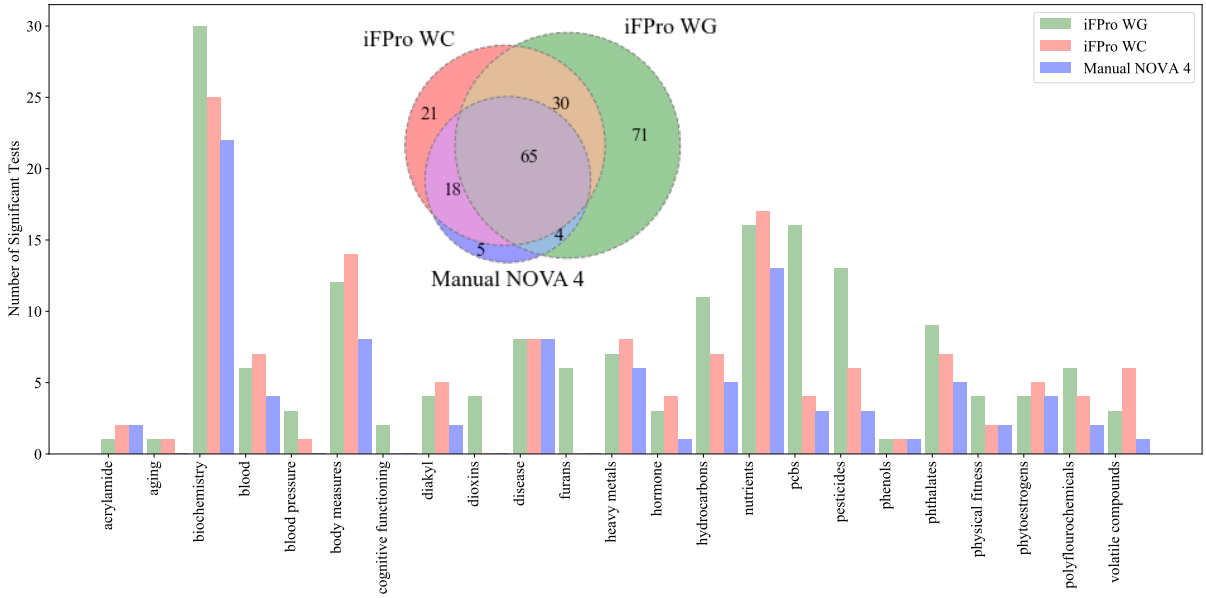

Figure SI. 18: **Number of significant tests surviving multiple testing for $iFPro_{WC}$, $iFPro_{WG}$, and fraction of caloric intake from manual NOVA 4.** For $iFPro_{WG}$ and $iFPro_{WC}$, we find 170 and 134 significant associations, respectively, for a total of 209 unique variables. In comparison, the same analysis using the fraction of calories contributed by manual NOVA 4 finds 92 significant associations, of which 95% is in overlap with $iFPro_{WG}$ and $iFPro_{WC}$. Overall, the total number of variables recovered by the three methodologies is 214. Here we report the results stratified by module.

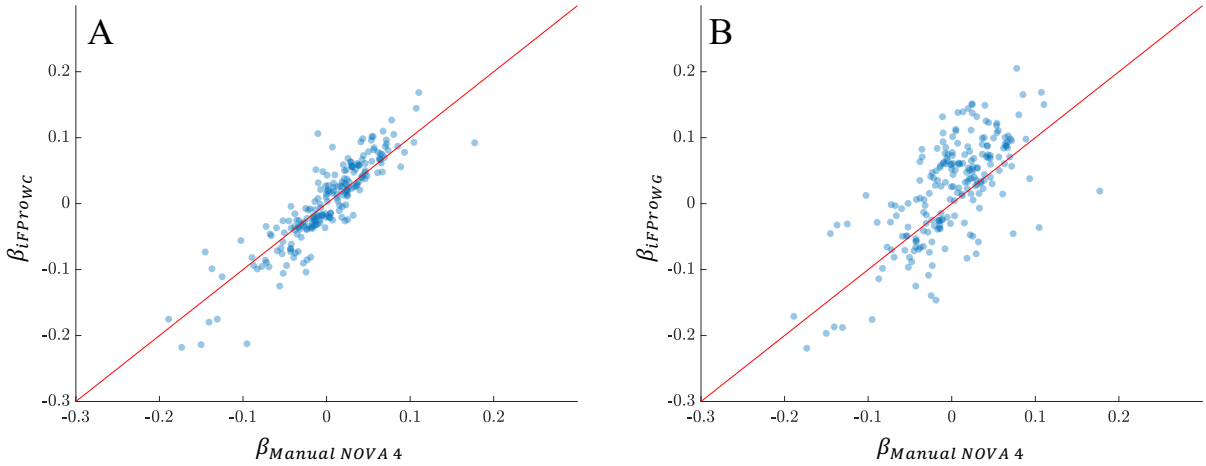

Figure SI. 19: **Comparison of the Effect Sizes for the Significant Tests found by $iFPro_{WC}$, $iFPro_{WG}$, and fraction of caloric intake from manual NOVA 4.** Across the 214 significant variables recovered by the three methodologies we find that **(a)** 70.56% of the times $iFPro_{WC}$ shows bigger effect sizes compared to manual NOVA 4, while **(b)** for $iFPro_{WG}$ the percentage increases to 77.57%.

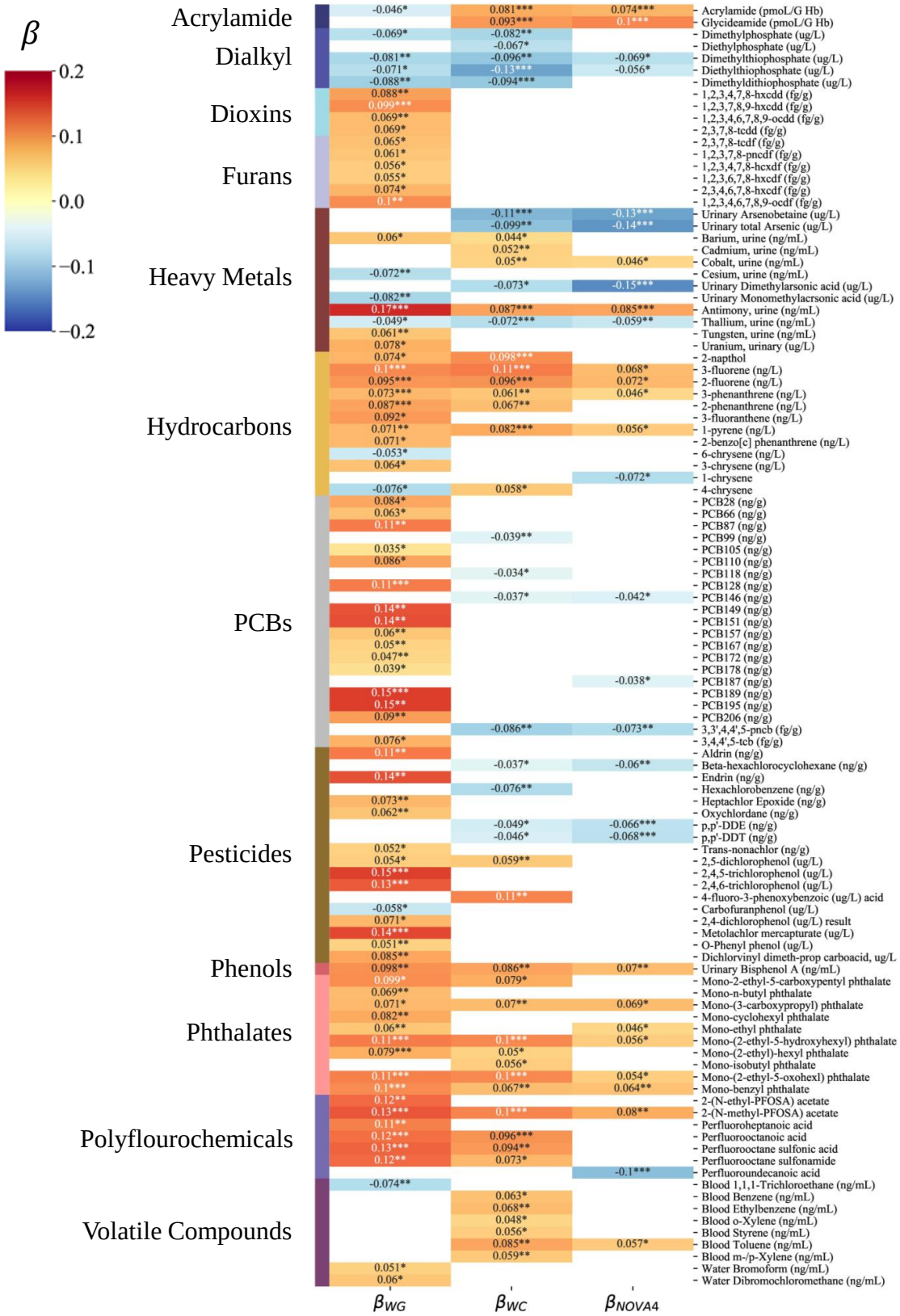

Figure SI. 20: **Exposure Panel.** Each variable reported on the left (e.g. "Acrylamide") refers to different exposure modules. We report here the standardized β coefficient, quantifying the effect on each exposure when the Box-Cox transformed diet scores increase of one standard deviation over the population. Each variable is color-coded according to β, with positive associations in red, and negative associations in blue. In white, we annotate the variables that do not survive Benjamini-Hochberg FDR.

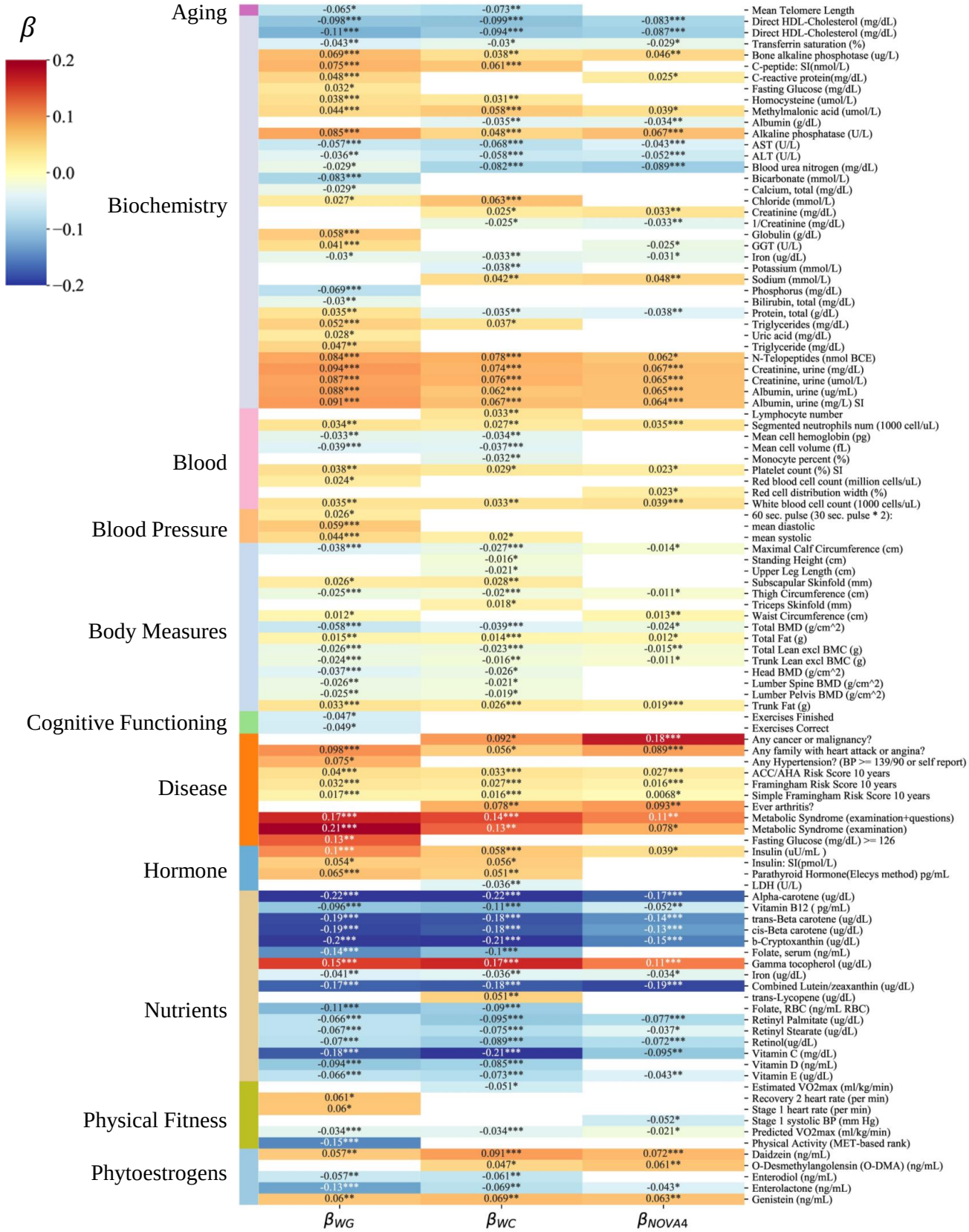

Figure SI. 21: **Health Panel.** Each variable reported on the left (e.g. "Aging") refers to different modules of disease phenotypes and measurements assessing the overall health status of each individual. We report here the standardized β coefficient, quantifying the effect on each exposure when the Box-Cox transformed diet scores increase of one standard deviation over the population. Each variable is color-coded according to β, with positive associations in red, and negative associations in blue. In white, we annotate the variables that do not survive Benjamini-Hochberg FDR.

Table 9: A Summary of the Epidemiological Literature on the Discovered Associations with Manual NOVA 4.

Health or Exposure Panel	**Paper**	**In Agreement with** $iFPro$
Aging	Ultra-processed food consumption and the risk of short telomeres in an elderly population of the Seguimiento Universidad de Navarra (SUN) Project [37] Spain (SUN)	Yes
Disease: Cardiovascular	• Association between ultraprocessed food intake and cardiovascular health in US adults: a cross-sectional analysis of the NHANES 2011–2016 [38] USA (NHANES) • Ultra-processed food intake and risk of cardiovascular disease: prospective cohort study (NutriNet-Santé) [39] France (NutriNet-Santé) • Ultra-processed food consumption is associated with increased risk of all-cause and cardiovascular mortality in the Moli-sani Study [40] Molise, Italy (Moli-sani)	Yes
Disease: Hypertension	Ultra-processed food consumption and the incidence of hypertension in a mediterranean cohort: The seguimiento universidad de navarra project [41] Spain (SUN)	Yes

Continued on next page

Table 9 – continued from previous page

Health or Exposure Panel	**Paper**	**In Agreement with** $iFPro$
Disease: Metabolic Syndrome	• Dietary share of ultra-processed foods and metabolic syndrome in the US adult population [42] USA (NHANES) • Diet quality indices in relation to metabolic syndrome in an Indigenous Cree (Eeyouch) population in northern Québec, Canada [43] Canada (Aschii Environment & Health Study) • A minimally processed dietary pattern is associated with lower odds of metabolic syndrome among Lebanese adults [44] Lebanon	Yes
Disease: Cancer	Consumption of ultra-processed foods and cancer risk: results from NutriNet-Santé prospective cohort [45] France (NutriNet-Santé)	Yes
Body Measures: Total Fat	Contribution of ultra-processed foods in visceral fat deposition and other adiposity indicators: Prospective analysis nested in the PREDIMED-Plus trial [46] Spain (PREDIMED-Plus)	Yes

Continued on next page

Table 9 – continued from previous page

Health or Exposure Panel	Paper	In Agreement with *iFPro*
Body Measures: Waist Circumference	• Ultra-processed food consumption and excess weight among US adults [47] USA (NHANES) • Consumption of ultra-processed food and obesity: cross sectional results from the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil) cohort (2008–2010) [48] Brazil (ELSA) • Ultra-processed food consumption and indicators of obesity in the United Kingdom population (2008-2016) [49] UK (NDNS)	Yes
Nutrients	• The share of ultra-processed foods and the overall nutritional quality of diets in the US: evidence from a nationally representative cross-sectional study [50] USA (NHANES) • Impact of ultra-processed foods on micronutrient content in the Brazilian diet [51] Brazil (HBS)	Yes
Biochemistry: C-Reactive Protein	Association between consumption of ultra-processed foods and serum C-reactive protein levels: cross-sectional results from the ELSA-Brasil study [52] Brazil (ELSA)	Yes
Phytoestrogens	Association between dietary share of ultra-processed foods and urinary concentrations of phytoestrogens in the US [53] USA (NHANES)	Yes
Vegetarian Diet: Adverse Effects of ultra-processing	Consumption of Ultra-Processed Foods by Pesco-Vegetarians, Vegetarians, and Vegans: Associations with Duration and Age at Diet Initiation [54] France (NutriNet-Santé)	Yes

Continued on next page

Table 9 – continued from previous page

Health or Exposure Panel	**Paper**	**In Agreement with** $iFPro$
Phthalates and Phenols	• Ultra-processed food consumption and exposure to phthalates and bisphenols in the US National Health and Nutrition Examination Survey, 2013–2014 [55] USA (NHANES) • Association between dietary contribution of ultra-processed foods and urinary concentrations of phthalates and bisphenol in a nationally representative sample of the US population aged 6 years and older [56] USA (NHANES)	Yes
Acrylamides	Association between Heat-Induced Chemical Markers and Ultra-Processed Foods: A Case Study on Breakfast Cereals [57] Spanish supermarkets	Yes

5 Food Substitution

The observed variability of $FPro$ for categories of foods similarly consumed in the population (Figure 2E), combined with the EWAS results, quantifying the effect of processed diet on disease risk, suggests a systematic way to implement food substitution and predict its relevance in terms of health indicators [58, 59]. With this goal, we classified all foods consumed by the pooled cohort of 20,047 individuals in NHANES 1999-2006, according to WWEIA [60]. For substitution purposes, the relevance of food k in individual j's diet can be quantified as

$$r_k^j = f_C^{(k,j)}(FPro_k - FPro_{min\ WWEIA(k)}), \tag{8}$$

where $f_C^{(k,j)}$ is the fraction of calories contributed by food k to the dietary profile, while $FPro_{min\ WWEIA(k)}$ refers to the food with the lowest $FPro$ within the same WWEIA category of food k. By picking the suggested foods in the original WWEIA classes reported by each individual, we aim to minimally perturb her habits, to maximize the compliance to the new dietary regime. Moreover, as Eq. S8 offers a heuristic to identify which food to prioritize, the overall level of processing is reduced in a minimal number of steps.

The impact of substituting M foods on disease risk is measured in terms of odds ratio (OR), that quantifies the odds of disease occurring when adopting the optimized diet, compared to the original choices. We estimate OR as

$$OR^j_{(Sub\ vs\ Orig)} = e^{\beta_1[(iFPro^j_{WC} - \sum_{m=1}^{M} r^j_m)^t - (iFPro^j_{WC})^t]}, \tag{9}$$

where β_1 is the effect size describing the strength of the association between $iFPro$ and disease onset, all $\{r^j_m\}$ follow from Eq. S8, and with the superscript t we denote the function-composition of Box-Cox transformation followed by z-score with parameters estimated in the original population.

For continuous variables y like vitamin B_{12}, vitamin C, and bisphenol A, the impact of substituting M foods on individual j's diet is quantified by

$$\frac{y^j_{Sub}}{y^j_{Orig}} = \frac{\{(y^j_{Orig})^t + \beta_1[(iFPro^j_{WC} - \sum_{m=1}^{M} r^j_m)^t - (iFPro^j_{WC})^t]\}^{-t}}{y^j_{Orig}}, \tag{10}$$

where with the superscript $-t$ we refer to the inverse function-composition of Box-Cox transformation followed by standardization (first invert z-score, then Box-Cox).

6 Open Food Facts

Open Food Facts is a free, crowd-sourced world-wide database of food products, with nutrition facts and ingredient list [61]. We collected 233,831 nutritional records from their website, corresponding to 168,681 product ids, annotated with NOVA labels according to a heuristic described at [62]. This database also compiled an extensive list of food additives that gave us the possibility to quantify the number of additives per product. Similarly to the cross-validation explained in Section 2.1, we trained and validated two models on the same 5-fold partition: (1) standard FProX leveraging the logarithm of 11 nutrients used as baseline, (2) FProX with an additional input feature capturing the number of additives in each food. The baseline feature panel includes protein, fat, total carbohydrate, sugars, dietary fiber, calcium, iron, sodium, cholesterol, saturated fat, and trans fat. We removed vitamin A and C from the analysis, given the high number of non available values (NAs). Hence, the cross-validation was run on a selection of 228,689 records, with no NAs in any feature.

Overall, we find commendable performances in both scenarios, with higher AUC and AUP when additives are included in the model (see Table 10). In particular, model (2) outperforms model (1) in assessing NOVA 3. We derived $FPro_1$ and $FPro_2$ following Eq. 1, finding they highly correlate with each other ($\rho_{Spearman} = 0.9017$), and they both correlate well with the number of additives ($\rho_{Spearman}(FPro_1, n_{additives}) = 0.6726$, $\rho_{Spearman}(FPro_2, n_{additives}) = 0.8240$). This result is encouraging, as details regarding ingredient list and additives are seldom if ever available in current food composition databases.

To better understand the role of additives in discriminating NOVA classes, we trained and validated a random forest taking the number of additives as the only input, over the same train-test split used for model (1) and (2). We find AUC equal to 0.859923 ± 0.000485 for NOVA 1, 0.832023 ± 0.003375 for NOVA 2, 0.820256 ± 0.000693 for NOVA 3, and 0.907307 ± 0.000609 for NOVA 4. Compared to model (2), the performances in terms of AUC drop of 13.36% for NOVA 1, 15.77% for NOVA 2, 15.03% for NOVA 3, and 7.25% for NOVA 4.

The lack of information on nutrient amounts strongly affect the precision-recall curve, increasing disproportionately the number of false positives predicted for NOVA 1, 2, and 3. Indeed, we find AUP equal to 0.288799 ± 0.000930 for NOVA 1, 0.026227 ± 0.000498 for NOVA 2, 0.439202 ± 0.001269 for NOVA 3, and 0.941548 ± 0.000350 for NOVA 4. The values for class NOVA 1, 2, and 3 are close to the proportion of examples labeled as 1, 2, and 3 in the training, suggesting that the model's behavior is close to a random classifier. Compared to model (2), the performances in terms of AUP drop of 70.01% for NOVA 1, 97.05% for NOVA 2, 50.96%

for NOVA 3, and 4.77% for NOVA 4.

Overall, these results suggest that the number of additives is a good predictor of ultra-processed food as defined by NOVA, but it misses the processing nuances represented in the other NOVA classes.

Table SI. 10: **AUC and AUP for the four NOVA classes in Open Food Facts**. For model (1) and (2) we report average and standard deviation of AUC (A) and AUP (B) over the stratified 5-folds.

(A)

	NOVA 1	NOVA 2	NOVA 3	NOVA 4
Average AUC 11 Nutrients	0.987989	0.986044	0.932033	0.950768
Std AUC 11 Nutrients	0.000647	0.004472	0.001525	0.00086
Average AUC 11 Nutrients + Additives	0.992553	0.987785	0.965301	0.978242
Std AUC 11 Nutrients + Additives	0.000327	0.004682	0.000962	0.000683

(B)

	NOVA 1	NOVA 2	NOVA 3	NOVA 4
Average AUP 11 Nutrients	0.949175	0.85327	0.828676	0.972467
Std AUP 11 Nutrients	0.002694	0.016713	0.002751	0.000712
Average AUP 11 Nutrients + Additives	0.962916	0.890253	0.895639	0.988722
Std AUP 11 Nutrients + Additives	0.001583	0.013054	0.002983	0.000388

7 Data Quality and Future Directions

- **Data Sample and Variability**. The quality of the training data influences our analysis and the interpretation of FPro. Indeed, all machine learning and AI models become feasible only when extensive labeled datasets are available, and on top of that, they provide reliable results when the training data samples the real world in an exhaustive way, avoiding over-fitting.

  A single cycle of FNDDS, as well as SR-Legacy, reports representative nutritional average values for each food/drink, which do not capture the variability due to factors such as recipe variations, production methods, soil quality, and storage time. The training data would then benefit of multiple instances of the same food, helping capture the natural variability in nutrient content, and reduce over-fitting [63]. This is why we have investigated how nutritional values for the same food code change through different FNDDS cycles, and how integrating additional variability is affecting FProX and FPro in assessing "unseen foods" (see *Assesment of FPro robustness* below). We envision that Foundation Foods, the new USDA food composition dataset available at FoodData Central, will be a great asset to improve the current data training of FPro, once it will increase its coverage to more than 140 foods. Indeed, Foundation Foods includes individual sample measurements behind the nutrient mean values that populate the other databases, and metadata reporting the number of samples, location, time-stamps, analytical methods used, and, additionally, if available, cultivar and production practices.

  A further way to improve how machine learning algorithms and AI characterize food data is to study the statistical properties of nutrient distributions in the food supply, and make sure that the training data sample them exhaustively. In Menichetti et al. [9] we observe how the variability of nutrient concentrations in food, despite covering several order of magnitudes, follows a well-defined functional form, related to the lognormal distribution. This is the reason why we log-transformed all nutrient measures before performing the classification, as this is the natural scale of the nutrient fluctuations. Accounting for the variability driven by chemical reactions in the ingredients, as well as for the variability derived from the complete food production chain, plays a major role in successfully capturing associations between food intake and disease phenotypes [64].

  Beyond nutrient variability, the training data would benefit of a better representation of different food groups, as in case of raw meat products, important ingredients in many recipes. This poor representation in the dataset is somehow expected, as FNDDS is designed by the USDA to provide food composition data for foods and beverages reported

in the dietary component of NHANES, and raw meat is not commonly consumed by the American population. A better coverage of heterogeneous collection of food groups would also improve the accuracy of the algorithm when facing different national food composition databases, beyond the data collected by the USDA.

Finally, we stress that all potential errors affecting food composition data should generally be random with respect to the machine learning algorithm and the manual NOVA classification that developed FPro, which should only cause attenuation of classification accuracy rather than bias, and attenuation toward the null of all epidemiological outcomes. Thus, our findings would further improve with a more accurate database.

- **Requirements for Branded Products**. While in this study we focus on survey data, our model based on nutritional values can easily work on different food and cohort databases, as proven by our analysis of over 50,000 products collected from major grocery store websites [65]. When facing real-world food data, we have to account for the regulations introduced by government agencies, like the FDA. For instance, in the nutrition facts label nutrients are classified in 3 different classes, characterized by different standards regarding the agreement between declared value on the label and actual values in the sampled food [66]. For nutrients like sugars, total fat, saturated fat, cholesterol, and sodium, the label is considered compliant if the nutrient content of the sampled product is up to 20% above the value declared on the label. Consequently, when relying on the nutrition facts of branded products, FPro should be reported with confidence intervals determined by randomly altering the nutrient content according to FDA regulations.

- **Size and Composition of the Nutrient Panel**. The computation of FPro adapts to different set of nutrients, allowing us to accurately classify food from limited nutrient information. However, the chemical information currently available to train our algorithm is limited by the resolution of food databases. Indeed, many chemicals like acrylamide, ammonium sulfate, azodicarbonamide, butylated hydroxyanisole, and furans, associated with the preparation and preservation of food, are not tracked by national agencies. The lack of quantification of these chemicals becomes even more striking once we acknowledge their impact on human health [67–71]. With a higher number of chemicals available for all foods, we could aim for an unsupervised classification of food processing, eliminating the need for supervised manual curation. Currently, our analysis in Section S1.5 shows that an unsupervised hierarchical clustering of foods, leveraging the widest nutrient panel available in FNDDS, is not able to independently reproduce the four NOVA classes. It is

possible, however, that the addition of chemical measurements that pertain to processing signatures could further improve the current results, leading to purely chemically driven classification of food processing.

- **Assessment of FPro robustness**. We aim to test the robustness of FPro by measuring the extent of its variations when re-formulation of the same product are created and measured experimentally. This analysis is currently hindered by limited data availability. However, we were able to test how nutrient variability impacts FPro when comparing the same food code in different cycles of FNDDS. Indeed, while FNDDS does not capture sample variability within the same cycle (e.g. 2009-2010), it documents the nutrient variability affecting the same food through the years, as a function of the specific sample or technique chosen to estimate the representative nutritional average. In FoodProXonFNDDS.xlsx we report the nutrient profile of 5,632 foods, present in both FNDDS 2009-2010 and 2015-2016, and how their FPro changed according to the variations in their nutrient content. For all foods we estimate the number of nutrients that changed between the two editions of FNDDS, accounting for the different rounding approximation affecting each nutrient. In presence of one or more nutrients with maximal variation between 10% and 50%, we observe a median ΔFPro=0.001556 (Q_1=0.000222, Q_3= 0.004764). Allowing significantly bigger fluctuations, as in case of one or more nutrients with maximal variation between 10% and 1000%, we observe a median ΔFPro= 0.003312 (Q_1= 0.000722, Q_3= 0.011310).

- **Limitations of Population Surveys**. To test the prediction power of FPro and iFPro for several health phenotypes and chemical exposures, we leveraged the rich panel of analyses collected in NHANES, a population survey that captures dietary intake with 24-hour recalls. A limitation of the 24-hour recall is its reliance on memory, both for identification of foods eaten as well as for quantification of portion sizes, and the need of highly trained interviewers. Although reliance on the participant's memory leaves room for measurement error, skilled interviewers can produce highly detailed and useful nutritional data comparable to a dietary record [24]. 24-hour recalls are also affected by random within-person error, typified by the day-to-day fluctuation in dietary intake. Random within-person error tends to decrease correlation and regression coefficients towards zero and to bias relative risks toward one. Multiple days of intake per individual permit an estimate of within-person day-to-day variability, and it is usually statistically more efficient to increase the number of individuals in the sample rather than increase the number of days beyond two per individual [72]. For this reason, NHANES added a second day of dietary intake starting with NHANES 2003 [24]. Despite these limitations, 24-hour recalls

are widely used in epidemiology as their statistical issues are well understood, and they were used to rigorously evaluate HEI-2015, a gold-standard non-data-driven dietary score for the epidemiological community [73].

8 Overview of FoodProX and FPro

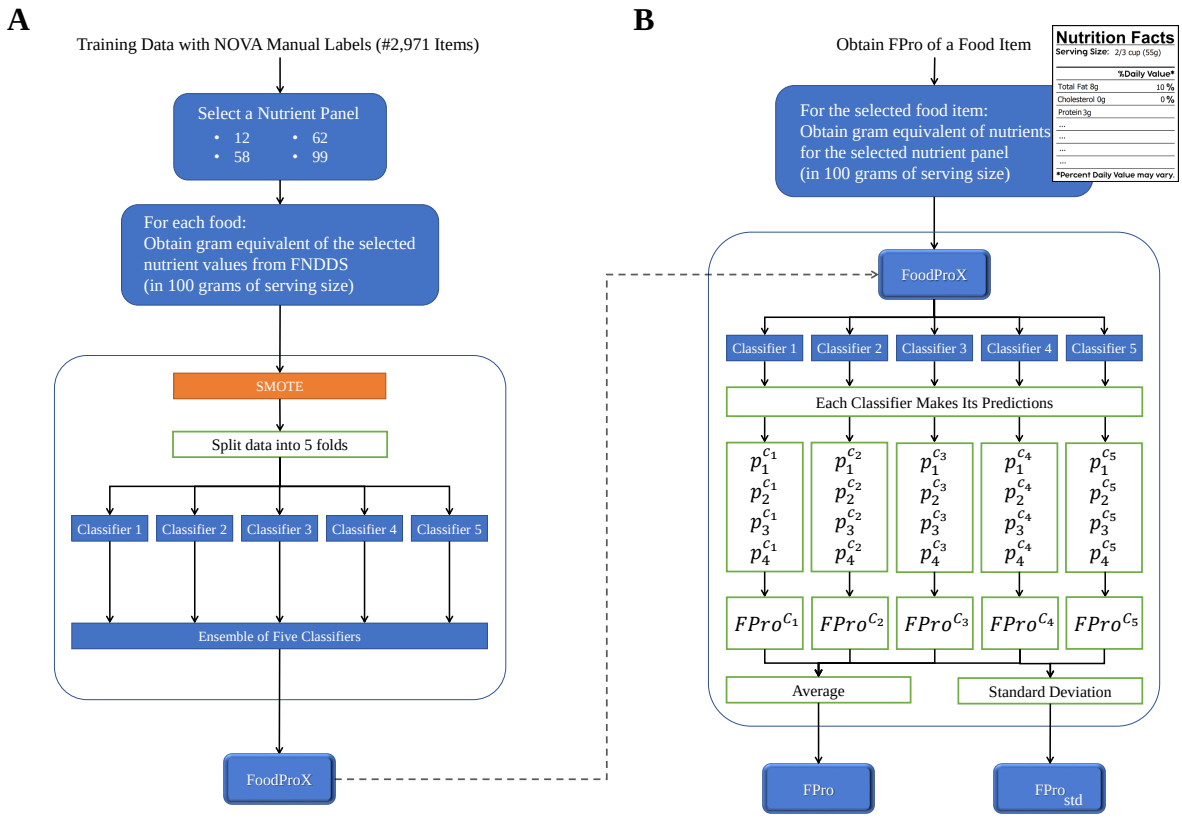

Figure SI. 22: **Summary Overview of FoodProX Classifier (A) and FPro Score (B).**

References

[1] FDA Nutrition Facts. https://www.fda.gov/media/99331/download.

[2] USDA FoodData Central. https://fdc.nal.usda.gov/.

[3] Ahuja, J. *et al.* USDA Food and Nutrient Database for Dietary Studies, 5.0. U.S. Department of Agriculture, Agricultural Research Service, Food Surveys Research Group, Beltsville, MD. http://www.ars.usda.gov/ba/bhnrc/fsrg (2012).

[4] Sebastian, R. S. *et al.* Flavonoid Values for USDA Survey Foods and Beverages 2007–2010. U.S. Department of Agriculture, Agricultural Research Service, Food Surveys Research Group, Beltsville, MD. http://www.ars.usda.gov/nea/bhnrc/fsrg (2016).

[5] USDA National Nutrient Database for Standard Reference (SR). URL https://data.nal.usda.gov/dataset/usda-national-nutrient-database-standard-reference-legacy-release.

[6] USDA FoodData Central (FDC) . URL https://fdc.nal.usda.gov/.

[7] Steele, E. M. *et al.* Ultra-processed foods and added sugars in the US diet: Evidence from a nationally representative cross-sectional study. *BMJ Open* **6**, 1–8 (2016).

[8] Database of Flavonoid Values for USDA Food Codes 2007-2010 and Flavonoid Intake Data Files from What We Eat in America (WWEIA), National Health and Nutrition Examination Survey (NHANES) 2007-2010 . URL https://www.ars.usda.gov/northeast-area/beltsville-md-bhnrc/beltsville-human-nutrition-research-center/food-surveys-research-group/docs/fndds-flavonoid-database/.

[9] Menichetti, G. & Barabási, A.-L. Nutrient concentrations in food display universal behaviour. *Nature Food 2022 3:5* **3**, 375–382 (2022). URL https://www.nature.com/articles/s43016-022-00511-0.

[10] Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720 (2008).

[11] Vinh, N. X., Epps, J. & Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research* **11**, 2837–2854 (2010).

[12] Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **16**, 321–357 (2002). 1106.1813.

[13] Parr, T., Turgutlu, K., Csiszar, C. & Howard, J. Permutation Feature Importance. URL github.com/parrt/random-forest-importances.

[14] Strobl, C., Boulesteix, A. L., Zeileis, A. & Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* **8**, 1–21 (2007). URL https://link.springer.com/articles/10.1186/1471-2105-8-25https://link.springer.com/article/10.1186/1471-2105-8-25.

[15] Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems 30* 4765–4774 (2017). URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf. 1705.07874.

[16] Kapur, J. N. *Maximum-Entropy Models in Science and Engineering.* (Wiley, 1989).

[17] Monteiro, C. A. *et al.* NOVA. The star shines bright. *World Nutrition* **7**, 28–38 (2016).

[18] Post Shredded Wheat: WHEAT 'N BRAN. URL https://www.postshreddedwheat.com/products/wheat-n-bran/.

[19] Mann-Whitney Test for Independent Samples. URL http://www.real-statistics.com/non-parametric-tests/mann-whitney-test/.

[20] Patel, C. J. *et al.* A database of human exposomes and phenomes from the us national health and nutrition examination survey. *Scientific Data* **3**, 160096 (2016). URL https://doi.org/10.1038/sdata.2016.96.

[21] Measuring guides for the dietary recall interview. https://www.cdc.gov/nchs/nhanes/measuring_guides_dri/measuringguides.htm. Accessed: 2021-09-20.

[22] Reedy, J. *et al.* Evaluation of the healthy eating index-2015. *Journal of the Academy of Nutrition and Dietetics* **118**, 1622–1633 (2018). URL https://www.sciencedirect.com/science/article/pii/S2212267218308360.

[23] Nhanes phone follow-up dietary interviewer procedures manual. https://www.cdc.gov/nchs/data/nhanes/nhanes_09_10/phone_follow_up_dietary_procedures_manual_mar_2010.pdf (2010). Accessed: 2021-09-20.

[24] Satija, A., Yu, E., Willett, W. C. & Hu, F. B. Understanding Nutritional Epidemiology and Its Role in Policy. *Advances in Nutrition* **6**, 5–18 (2015). URL https://doi.org/10.3945/an.114.007492. https://academic.oup.com/advances/article-pdf/6/1/5/23880586/5.pdf.

[25] NHANES Survey Methods and Analytic Guidelines. URL https://wwwn.cdc.gov/nchs/nhanes/analyticguidelines.aspx.

[26] Lumley, T. survey: analysis of complex survey samples (2020). R package version 4.0.

[27] Healthy Eating Index (HEI) . URL https://www.fns.usda.gov/resource/healthy-eating-index-hei.

[28] National cancer institute. developing the healthy eating index. bethesda, md: National cancer institute. https://epi.grants.cancer.gov/hei/developing.html. 2020 (accessed September 1, 2020).

[29] Rhodes, D. G., Adler, M. E., Clemens, J. C. & Moshfegh, A. J. What we eat in America food categories and changes between survey cycles. *Journal of Food Composition and Analysis* **64**, 107–111 (2017). URL http://dx.doi.org/10.1016/j.jfca.2017.07.018.

[30] Milanlouei, S. *et al.* A systematic comprehensive longitudinal evaluation of dietary factors associated with acute myocardial infarction and fatal coronary heart disease. *Nature Communications* **11**, 1–14 (2020). URL https://doi.org/10.1038/s41467-020-19888-2.

[31] Patel, C. J., Bhattacharya, J. & Butte, A. J. An environment-wide association study (EWAS) on type 2 diabetes mellitus. *PLoS ONE* **5** (2010).

[32] Moore, J., Chaudhary, N. & Akinyemiju, T. Metabolic syndrome prevalence by race/ethnicity and sex in the united states, national health and nutrition examination survey, 1988–2012. *Prev Chronic Dis* (2017).

[33] D'Agostino, R. B. *et al.* General cardiovascular risk profile for use in primary care: The Framingham heart study. *Circulation* **117**, 743–753 (2008).

[34] Castro, V. *CVrisk: Compute Risk Scores for Cardiovascular Diseases* (2021). URL https://github.com/vcastro/CVrisk. R package version 1.1.0.9000.

[35] Goff, D. C. *et al.* 2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the American college of cardiology/American heart association task force on practice guidelines. *Circulation* **129**, 49–73 (2014).

[36] Menard, S. Standards for standardized logistic regression coefficients. *Social Forces* **89**, 1409–1428 (2011).

[37] Alonso-Pedrero, L. *et al.* Ultra-processed food consumption and the risk of short telomeres in an elderly population of the Seguimiento Universidad de Navarra (SUN) Project. *The American journal of clinical nutrition* **111**, 1259–1266 (2020). URL `https://academic.oup.com/ajcn/article/111/6/1259/5824715`.

[38] Zhang, Z., Jackson, S. L., Martinez, E., Gillespie, C. & Yang, Q. Association between ultraprocessed food intake and cardiovascular health in us adults: a cross-sectional analysis of the nhanes 2011–2016. *The American Journal of Clinical Nutrition* **113**, 428–436 (2020). URL `https://doi.org/10.1093/ajcn/nqaa276`. `https://academic.oup.com/ajcn/article-pdf/113/2/428/36170430/nqaa276.pdf`.

[39] Srour, B. *et al.* Ultra-processed food intake and risk of cardiovascular disease: prospective cohort study (nutrinet-santé). *BMJ* **365** (2019). URL `https://www.bmj.com/content/365/bmj.l1451`. `https://www.bmj.com/content/365/bmj.l1451.full.pdf`.

[40] Bonaccio, M. *et al.* Ultra-processed food consumption is associated with increased risk of all-cause and cardiovascular mortality in the Moli-sani Study. *The American Journal of Clinical Nutrition* **113**, 446–455 (2020). URL `https://doi.org/10.1093/ajcn/nqaa299`. `https://academic.oup.com/ajcn/article-pdf/113/2/446/36170506/nqaa299.pdf`.

[41] De Deus Mendonça, R. *et al.* Ultra-processed food consumption and the incidence of hypertension in a mediterranean cohort: The seguimiento universidad de navarra project. *American Journal of Hypertension* **30**, 358–366 (2017). URL `https://pubmed.ncbi.nlm.nih.gov/27927627/`.

[42] Martínez Steele, E., Juul, F., Neri, D., Rauber, F. & Monteiro, C. A. Dietary share of ultra-processed foods and metabolic syndrome in the us adult population. *Preventive Medicine* **125**, 40–48 (2019). URL `https://www.sciencedirect.com/science/article/pii/S0091743519301720`.

[43] Lavigne-Robichaud, M. *et al.* Diet quality indices in relation to metabolic syndrome in an indigenous cree (eeyouch) population in northern québec, canada. *Public Health Nutrition* **21**, 172–180 (2018).

[44] Nasreddine, L. *et al.* A minimally processed dietary pattern is associated with lower odds of metabolic syndrome among lebanese adults. *Public Health Nutrition* **21**, 160–171 (2018).

[45] Fiolet, T. *et al.* Consumption of ultra-processed foods and cancer risk: results from nutrinet-santé prospective cohort. *BMJ* **360** (2018). URL https://www.bmj.com/content/360/bmj.k322. https://www.bmj.com/content/360/bmj.k322.full.pdf.

[46] Konieczna, J. *et al.* Contribution of ultra-processed foods in visceral fat deposition and other adiposity indicators: Prospective analysis nested in the predimed-plus trial. *Clinical Nutrition* (2021). URL https://www.sciencedirect.com/science/article/pii/S0261561421000297.

[47] Juul, F., Martinez-Steele, E., Parekh, N., Monteiro, C. A. & Chang, V. W. Ultra-processed food consumption and excess weight among us adults. *British Journal of Nutrition* **120**, 90–100 (2018).

[48] Silva, F. M. *et al.* Consumption of ultra-processed food and obesity: cross sectional results from the brazilian longitudinal study of adult health (elsa-brasil) cohort (2008–2010). *Public Health Nutrition* **21**, 2271–2279 (2018).

[49] Rauber, F. *et al.* Ultra-processed food consumption and indicators of obesity in the united kingdom population (2008-2016). *PLOS ONE* **15**, 1–15 (2020). URL https://doi.org/10.1371/journal.pone.0232676.

[50] Martínez Steele, E., Popkin, B. M., Swinburn, B. & Monteiro, C. A. The share of ultra-processed foods and the overall nutritional quality of diets in the us: evidence from a nationally representative cross-sectional study. *Population Health Metrics* **15**, 6 (2017). URL https://doi.org/10.1186/s12963-017-0119-3.

[51] Louzada, M. L. d. C. *et al.* Impact of ultra-processed foods on micronutrient content in the Brazilian diet. *Revista de saude publica* **49** (2015). URL https://doi.org/10.1590/S0034-8910.2015049006211.

[52] Lopes, A. E. d. S. C., Araújo, L. F., Levy, R. B., Barreto, S. M. & Giatti, L. Association between consumption of ultra-processed foods and serum C-reactive protein levels: cross-sectional results from the ELSA-Brasil study. *Sao Paulo Medical Journal* **137**, 169 – 176 (2019). URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-31802019000200169&nrm=iso.

[53] Steele, E. M. & Monteiro, C. A. Association between dietary share of ultra-processed foods and urinary concentrations of phytoestrogens in the US. *Nutrients* **9** (2017).

[54] Gehring, J. *et al.* Consumption of Ultra-Processed Foods by Pesco-Vegetarians, Vegetarians, and Vegans: Associations with Duration and Age at Diet Initiation. *The Journal of Nutrition* **151**, 120–131 (2020). URL https://doi.org/10.1093/jn/nxaa196. https://academic.oup.com/jn/article-pdf/151/1/120/35365121/nxaa196.pdf.

[55] Buckley, J. P., Kim, H., Wong, E. & Rebholz, C. M. Ultra-processed food consumption and exposure to phthalates and bisphenols in the US National Health and Nutrition Examination Survey, 2013–2014. *Environment International* **131** (2019).

[56] Martínez Steele, E., Khandpur, N., da Costa Louzada, M. L. & Monteiro, C. A. Association between dietary contribution of ultra-processed foods and urinary concentrations of phthalates and bisphenol in a nationally representative sample of the us population aged 6 years and older. *PLOS ONE* **15**, 1–21 (2020). URL https://doi.org/10.1371/journal.pone.0236738.

[57] Morales, F. J., Mesías, M. & Delgado-Andrade, C. Association between heat-induced chemical markers and ultra-processed foods: A case study on breakfast cereals. *Nutrients* **12** (2020). URL https://www.mdpi.com/2072-6643/12/5/1418.

[58] Ibsen, D. B. *et al.* Food substitution models for nutritional epidemiology. *The American journal of clinical nutrition* **113**, 294–303 (2021).

[59] Adams, J., Hofman, K., Moubarac, J. C. & Thow, A. M. Public health response to ultra-processed food and drinks. *BMJ (Clinical research ed.)* **369**, m2391 (2020).

[60] What We Eat In America (WWEIA) Database. URL https://data.nal.usda.gov/dataset/what-we-eat-america-wweia-database.

[61] Open Food Facts - World (10/15/2021). URL https://world.openfoodfacts.org/.

[62] Formula to determine the Nova group (10/15/2021). URL https://world.openfoodfacts.org/nova.

[63] U.S. Department of Agriculture, A. R. S. FoodData Central: Foundation Foods (2019). URL fdc.nal.usda.gov.

[64] Dekker, M. & Verkerk, R. Dealing with variability in food production chains: A tool to enhance the sensitivity of epidemiological studies on phytochemicals. *European Journal of Nutrition* **42**, 67–72 (2003). URL https://link.springer.com/article/10.1007/s00394-003-0412-8.

[65] Ravandi, B., Mehler, P., Barabási, A.-L. & Menichetti, G. GroceryDB: A Database of Food and Beverage Products Annotated by Food Processing Characteristics in the US Grocery Stores (2021).

[66] FDA Center for Food Safety and Applied Nutrition. Guidance for Industry: Guide for Developing and Using Data Bases for Nutrition Labeling. URL https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-industry-guide-developing-and-using-data-bases-nutrition-labeling.

[67] Matthäus, B. & Haase, N. U. Acrylamide in ready-to-eat foods. In Kotzekidou, P. (ed.) *Food Hygiene and Toxicology in Ready-to-Eat Foods*, 353–382 (Academic Press, San Diego, 2016). URL https://www.sciencedirect.com/science/article/pii/B9780128019160000200.

[68] Phillips, D. H. Polycyclic aromatic hydrocarbons in the diet. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis* **443**, 139–147 (1999). URL https://www.sciencedirect.com/science/article/pii/S1383574299000162.

[69] Schecter, A. *et al.* Polybrominated diphenyl ethers (pbdes) and hexabromocyclodecane (hbcd) in composite u.s. food samples. *Environmental Health Perspectives* **118**, 357–362 (2010). URL https://ehp.niehs.nih.gov/doi/abs/10.1289/ehp.0901345. https://ehp.niehs.nih.gov/doi/pdf/10.1289/ehp.0901345.

[70] Crews, C. & Castle, L. A review of the occurrence, formation and analysis of furan in heat-processed foods. *Trends in Food Science & Technology* **18**, 365–372 (2007). URL https://www.sciencedirect.com/science/article/pii/S0924224407000854.

[71] Ye, J., Wang, X.-H., Sang, Y.-X. & Liu, Q. Assessment of the determination of azodicarbonamide and its decomposition product semicarbazide: Investigation of variation in flour and flour products. *Journal of Agricultural and Food Chemistry* **59**, 9313–9318 (2011). URL https://doi.org/10.1021/jf201819x. PMID: 21786817, https://doi.org/10.1021/jf201819x.

[72] Willett, W. *Nutritional epidemiology* (1998).

[73] Reedy, J. *et al.* Evaluation of the healthy eating index-2015. *Journal of the Academy of Nutrition and Dietetics* **118**, 1622 – 1633 (2018). URL http://www.sciencedirect.com/science/article/pii/S2212267218308360.