nature food

Resource

https://doi.org/10.1038/s43016-024-01095-7

Prevalence of processed foods in major US grocery stores

Received: 21 May 2022

Accepted: 20 November 2024

Published online: 13 January 2025

Check for updates

Babak Ravandi[1,6], Gordana Ispirova[2,6], Michael Sebek[1,6], Peter Mehler[3], Albert-László Barabási[1,2,4] & Giulia Menichetti[1,2,5,6] ✉

The offering of grocery stores is a strong driver of consumer decisions. While highly processed foods such as packaged products, processed meat and sweetened soft drinks have been increasingly associated with unhealthy diets, information on the degree of processing characterizing an item in a store is not straightforward to obtain, limiting the ability of individuals to make informed choices. GroceryDB, a database with over 50,000 food items sold by Walmart, Target and Whole Foods, shows the degree of processing of food items and potential alternatives in the surrounding food environment. The extensive data gathered on ingredient lists and nutrition facts enables a large-scale analysis of ingredient patterns and degrees of processing, categorized by store, food category and price range. Furthermore, it allows the quantification of the individual contribution of over 1,000 ingredients to ultra-processing. GroceryDB makes this information accessible, guiding consumers toward less processed food choices.

Food ultra-processing has drastically increased productivity and shelf time, addressing the issue of food availability to the detriment of food systems sustainability and health[1–4]. Indeed, there is increasing evidence that over-reliance on ultra-processed food (UPF) has fostered unhealthy diet[5]. The sheer number of peer-reviewed articles investigating the link between the degree of food processing and health embodies a general consensus among independent researchers on the health relevance of UPF, contributing up to 60% of consumed calories in developed nations[6–8]. For instance, recent studies have linked the consumption of UPF to non-communicable diseases such as metabolic syndrome[9–15] and to exposure to industrialized preservatives and pesticides[16–20]. This body of work has driven a paradigm shift from focusing solely on food security, which emphasizes access to affordable food, to prioritizing nutrition security[21,22]. Nutrition security stresses equitable access to healthy, safe and affordable foods essential for optimal health and well-being, as defined by the US Department of Agriculture (USDA)[23,24], echoing the recent White House Conference on Hunger, Nutrition, and Health[25].

Much of UPF reaches consumers through grocery stores, as documented by the National Health and Nutrition Examination Survey, indicating that in the United States over 60% of the food consumed comes from grocery stores (Supplementary Fig. 1). The high reliance on UPF and their potential negative health effects raise numerous critical questions, such as the following: (1) How can the degree of processing of food items be determined? (2) What methods can be used to quantify the extent of food processing in the food supply? (3) What alternatives can be identified to reduce UPF consumption?

Measuring the degree of food processing is a key step in addressing these questions, but it is not straightforward. Indeed, food labels often show mixed messages, partly driven by reductionist metrics focusing on one nutrient at a time[26] and partly because of the contrasting criteria on how to classify processed foods[27]. The ambiguity and inconsistency of current food processing classification systems have led to conflicting results regarding their role as risk factors for non-communicable chronic diseases[28,29]. Some of these classification systems also suffer from poor inter-rater reliability and lack of reproducibility, issues rooted in purely descriptive expertise-based approaches, leaving room for ambiguity and differences in interpretation[27,28,30]. Hence, there is a growing call among scientists for a more objective definition

[1]Network Science Institute and Department of Physics, Northeastern University, Boston, MA, USA. [2]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [3]Department of Computer Science, IT University of Copenhagen, Copenhagen, Denmark. [4]Department of Network and Data Science, Central European University, Budapest, Hungary. [5]Harvard Data Science Initiative, Harvard University, Boston, MA, USA. [6]These authors contributed equally: Babak Ravandi, Gordana Ispirova, Michael Sebek, Giulia Menichetti. ✉e-mail: giulia.menichetti@channing.harvard.edu

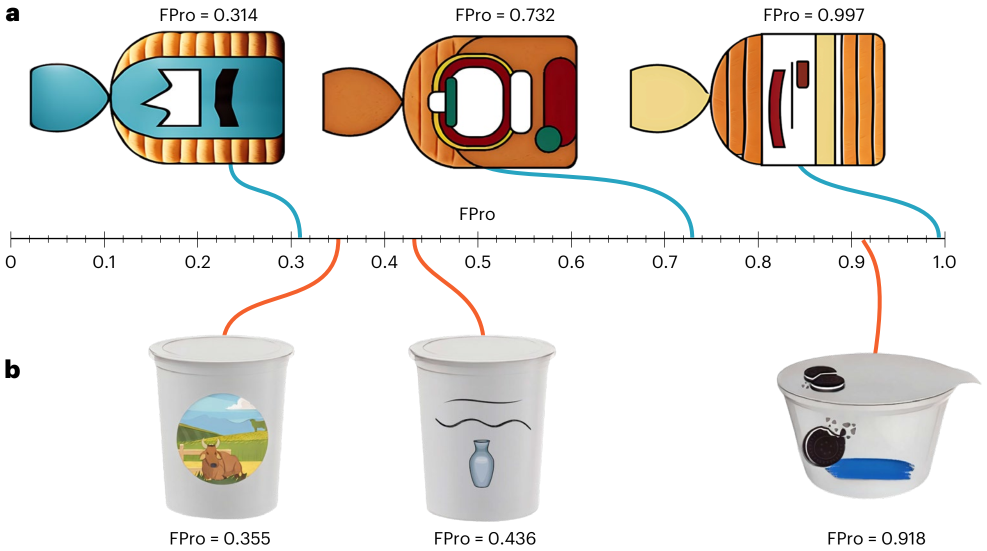


Fig. 1 | Degrees of food processing in three categories. FPro can assess the extent of food processing in three major US grocery stores, and it is best suited to rank foods within the same category. **a**, In breads, the Manna Organics multi-grain bread, offered by Whole Foods, is mainly made from 'whole wheat kernels', barley and brown rice without any additives, added salt, oil and yeast, with FPro = 0.314. However, the Aunt Millie's (FPro = 0.732) and Pepperidge Farmhouse (FPro = 0.997) breads, found in Target and Walmart, include soluble corn fibre and oat fibre with additives such as 'sugar', resistant corn starch, 'wheat gluten' and 'monocalcium phosphate'. **b**, The Seven Stars Farm yogurt (FPro = 0.355) is made from grade A pasteurized organic milk. The Siggi's yogurt (FPro = 0.436) declares pasteurized skim milk as the main ingredient that has 0% fat milk, requiring more food processing to eliminate fat. Lastly, the Chobani Cookies and Cream yogurt (FPro = 0.918) has cane sugar as the second most dominant ingredient combined with multiple additives such as caramel colour, fruit pectin and vanilla bean powder, making it a highly processed yogurt. Credit: round glossy ice cream cup, Shubby Studio, Adobe Stock; yogurt and ice cream tub, DEVASHISH ˙RAVAT, Getty images; mauve paint brushstroke, DSAP Project, DSAP Project's Images; all other icons (rural meadow and cow, cow gradient, blue ceramic vase, cookie bite, ice cream topping), Canva.com.

of the degree of food processing, grounded in biological mechanisms instead of varying subjective interpretations across research groups[28]. Among the proposed areas for aligning food processing definitions, the nutritional profile of food is currently the only aspect consistently regulated and reported worldwide[27,28,31].

The research efforts outlined in ref. 28 align with a growing demand for high-quality and internationally comparable statistics to promote objective metrics, reproducibility and data-driven decision-making, advancing convergence towards the Sustainable Development Goals[32,33]. Artificial intelligence (AI) methodologies[33–36], in particular, are increasingly being used for their potential as more objective, data-driven tools to advance nutrition security, a concept underpinning Sustainable Development Goals such as 'zero-hunger', 'good health and well-being', 'industry, innovation and infrastructure' and 'reduce inequalities'.

Responding to the need for objective and scalable metrics to ensure nutrition security, recent efforts harnessed machine learning to create and fully automate a food processing score (FPro)[37]. FPro is a continuous index derived by training a machine learning model to predict manual labels of processing techniques based on the overall nutrient profile of a food item (Methods and Supplementary Section 4). To teach the algorithm how to score processing from nutrients, labels provided by NOVA—the most widely used system for classifying foods based on processing-related criteria—were leveraged, offering a rich array of epidemiological literature for comparative analysis[9,38,39]. However, the FPro algorithm can accommodate different food processing classification systems such as the European Prospective Investigation into Cancer and Nutrition (EPIC)[40], University of North Carolina (UNC)[41] or Système d'Information et de Gestion des Aliments (SIGA)[42]. The predictive power of FPro was rigorously tested for epidemiological outcomes with an Environment-Wide Association Study, leveraging multiple cycles of the USDA model food databases and national food consumption surveys[37].

In this Article, building on the versatility and scalability of the FPro algorithm, we extend our analysis beyond 'model foods' tailored for epidemiological databases, analysing real-world data encompassing over 50,000 products from major US grocery store websites. This extensive dataset underpins the development of GroceryDB, an open-source database of foods and beverages, featuring comprehensive metadata on nutritional content, ingredient list and price for each item, collected from publicly available online markets of Walmart, Target and Whole Foods. Our objective is to demonstrate how machine learning can effectively analyse large-scale real-world food composition data and translate this wealth of information into the degree of processing for any food in grocery stores, facilitating consumer decision-making and informing public health initiatives aimed at enhancing the overall quality of the food environment. This initiative not only lays the groundwork for similar efforts globally, aimed at promoting better-informed dietary choices, but also underscores the critical role of open-access, internationally comparable data in advancing global nutrition security.

Results

For each food, we automated the process of determining the extent of food processing using FPro, which translates the nutritional content of a food item into its degree of processing[37]. Figure 1 illustrates the use of FPro by presenting the processing scores of three products in the bread and yogurt categories, enabling the comparison of their processing levels. For example, the Manna Organics multi-grain bread is made from whole wheat kernels, barley, rice without additives, added salt, oil and yeast, resulting in a low processing score of FPro = 0.314. By contrast, the Aunt Millie's (FPro = 0.732) and Pepperidge Farmhouse (FPro = 0.997) breads include 'resistant corn starch', 'soluble corn fiber' and 'oat fiber', requiring additional processing to extract starch and fibre from corn and oat to be used as an independent ingredient (Fig. 1a), yielding higher processing scores. Similarly, in the yogurt category, the Seven Stars Farm yogurt (FPro = 0.355) is a whole milk

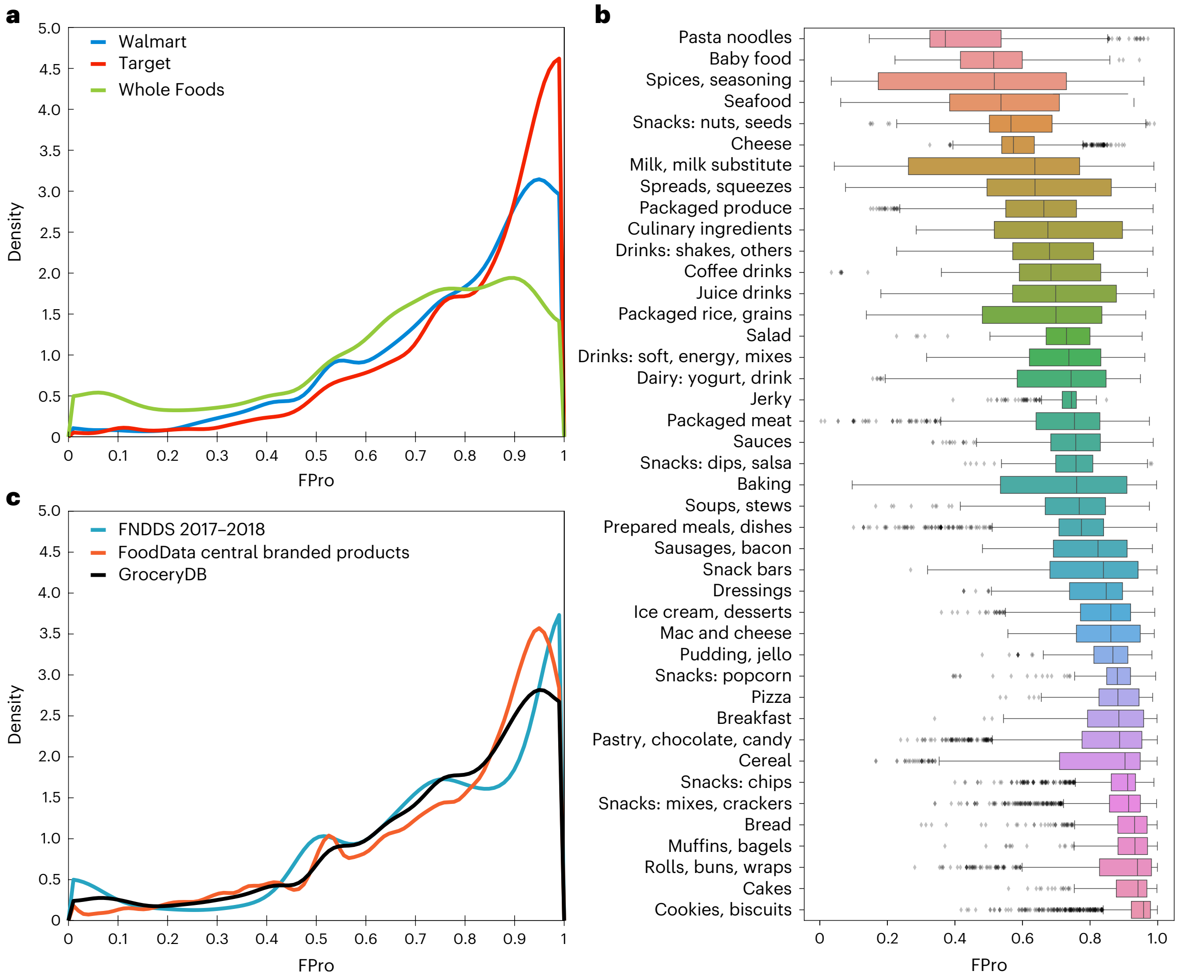


Fig. 2 | Food processing in grocery stores. a, The distribution of FPro scores from the three stores follows a similar trend, a monotonically increasing curve, indicating that the number of low FPro items (unprocessed and minimally processed) offered by the grocery stores is relatively lower than the number of high FPro items (highly processed and ultra-processed items), and the majority of offerings are ultra-processed (see Methods for FPro calculation). **b**, Distribution of FPro scores for different categories of GroceryDB. The distributions indicate that FPro has a remarkable variability within each food category, confirming the different degrees of food processing offered by the stores. Unprocessed foods such as eggs, fresh produce and raw meat are excluded (Supplementary Section 7). Sample sizes range from 126 for baby food to 2,043 for prepared meals dishes (see Source Data Figs. 1–6 for exact values). For the box plots, the minimum is the lower quartile, the central line is the median and the maximum is the upper quartile. The whiskers show data outside of the inter-quartile range. Diamonds represent outliers. **c**, The distributions of FPro scores in GroceryDB compared with two USDA nationally representative food databases: the USDA FNDDS and FoodData Central Branded Products (BFPD). The similarity between the distributions of FPro scores in GroceryDB, BFPD and FNDDS suggests that GroceryDB offers a comprehensive coverage of foods and beverages (Supplementary Section 6).

yogurt made from 'grade A pasteurized organic milk', while the Siggi's yogurt (FPro = 0.436) uses 'pasteurized skim milk' that requires further processing to obtain 0% fat. Finally, the Chobani Cookies and Cream yogurt relies on cane sugar as the second most dominant ingredient and contains cocktails of additives such as 'caramel color', 'fruit pectin' and 'vanilla bean powder' making it a highly processed yogurt, resulting in a high processing score FPro = 0.918.

GroceryDB assigns an FPro score to all foods collected from Walmart, Target and Whole Foods by leveraging the machine learning classifier FoodProX, which takes mandatory information from nutrition labels as input (Methods). The distribution of the FPro scores in the three stores shows a high degree of similarity: each store exhibits a monotonically increasing curve (Fig. 2a), indicating that minimally processed products (low FPro) represent a relatively small fraction of the inventory of grocery stores, the majority of the offerings being in the ultra-processed category (high FPro). Although less-processed items make up a smaller share of the overall inventory, they likely account for a proportionally larger portion of actual purchases, highlighting a discrepancy between sales data and available food options. Nevertheless, systematic differences between stores emerge: Whole Foods offers a greater selection of minimally processed items and fewer ultra-processed options, whereas Target has a particularly high proportion of ultra-processed products (high FPro). FPro also captures the inherent variability in the degree of processing per food category. As illustrated in Fig. 2b, there is a small variability of FPro scores in categories like jerky, popcorn, chips, bread, biscuits, and mac and cheese, indicating that consumers have limited choices in terms of degree of processing for these food groups (see Supplementary Section 7 for harmonizing categories between stores). Yet, in categories like cereals, milk and milk substitute, pasta noodles and snack bars, FPro shows considerable variation, reflecting a wider extent of possible choices from a food processing perspective.

The distribution of FPro in GroceryDB was compared with the latest USDA Food and Nutrient Database for Dietary Studies (FNDDS), offering a representative sample of the consumed food supply (Fig. 2c). The similarity between the distributions of FPro scores obtained from

GroceryDB and FNDDS suggests that GroceryDB also offers a representative sample of foods and beverages in the supply chain. In addition, the comparison of GroceryDB with the USDA Global Branded Food Products Database (BFPD), which contains 1,142,610 branded products, reveals that the FPro distributions in GroceryDB and BFPD follow similar trends (Fig. 2c). While BFPD contains 22 times more foods than GroceryDB, only an estimated 44% of GroceryDB's products are represented in BFPD, even after accounting for potential variability in food names and ingredient lists (Supplementary Section 6). This indicates that while BFPD offers an extensive representation of branded products, it does not fully capture the current offering of stores. Furthermore, a comparison of GroceryDB with Open Food Facts (OFF), an extensive crowd-sourced collection of branded products containing 426,000 items with English ingredient lists (https://world.openfoodfacts.org), reveals that fewer than 40% of the products in GroceryDB are present in OFF (Supplementary Fig. 4). This limited overlap suggests that monitoring products currently offered in grocery stores may provide a more accurate account of the food supply available to consumers.

Food processing and caloric intake

The depth and the resolution of the data collected in GroceryDB reveals some of the complexity regarding the relation between price and calories. Among all categories in GroceryDB, a 10% increase in FPro results in 8.7% decrease in the price per calorie of products, as captured by the dashed line in Fig. 3a. However, the relationship between FPro and price per calorie strongly depends on the food category (Supplementary Section 8). For example, in soups and stews the price per calorie drops by 24.3% for 10% increase in FPro (Fig. 3b), a trend observed also in cakes, mac and cheese, and ice cream (Supplementary Fig. 8). This means that, on average, the most processed soups and stews, with FPro ≈ 1, are 67.72% cheaper per calorie than the minimally processed alternatives with FPro ≈ 0.4 (Fig. 3e). By contrast, the price per calorie for cereals drops only by 1.2% for 10% increase in FPro (Fig. 3c), a slow decrease observed also for seafood and yogurt products (Supplementary Fig. 8). It is worth noting that there is an increasing trend between FPro and price in the milk and milk-substitute category (Fig. 3d), partially explained by the higher price of plant-based milk substitutes, which require more extensive processing than the dairy-based milks.

Choice availability and food processing

Not surprisingly, GroceryDB documents differences in the product offerings of the three stores analysed. For instance, in the cereal category—one of the most popular staple foods, consumed by 283 million Americans in 2020[43]—Whole Foods offers a selection with a broad spectrum of processing levels, while Walmart's cereal options are primarily limited to products with higher FPro (Fig. 4a). To investigate the roots of these differences, we examined the ingredients of cereals available at each store. The analysis showed that cereals sold at Whole Foods typically contain less sugar, fewer artificial and natural flavours, and fewer added vitamins compared with those at Walmart and Target, where products are more likely to include corn syrup, a sweetener associated with enhanced dietary fat absorption and weight gain (Fig. 4b)[44]. Additives such as butylated hydroxytoluene (a preservative) and calcium carbonate (an acidity regulator and anti-caking agent) are largely absent in the Whole Foods cereals, partially explaining the wider range of processing scores characterizing cereals at this store (Fig. 4a).

The brands offered by each store could also explain the different FPro patterns. Indeed, while Walmart and Target have a large overlap in the list of brands they carry, Whole Foods relies on different suppliers (Fig. 4c), largely unavailable in other grocery stores. In general, Whole Foods offers less processed soups and stews, yogurt and yogurt drinks, and milk and milk substitute (Fig. 4a). In these categories Walmart's and Target's offerings are limited to higher FPro values. Lastly, some food categories such as pizza, mac and cheese, and popcorn are highly processed in all stores (Fig. 4a). Pizzas available in all three chains, for example, consistently have high FPro values, partly due to the use of substitute ingredients such as 'imitation mozzarella cheese' instead of real 'mozzarella cheese'.

While grocery stores sell a large variety of products, the offered processing choices can be identical in multiple stores. For example, GroceryDB has a comparable number of cookies and biscuits in each chain, with 453, 373 and 402 items in Walmart, Target and Whole Foods, respectively. The degree of processing of cookies and biscuits in Walmart and Target are nearly identical ($0.88 < \text{FPro} < 1$), limiting consumer nutritional choices in a narrow range of processing (Fig. 4a). By contrast, Whole Foods not only offers a large number of items (402 cookies and biscuits) but also provides wider choices of processing ($0.57 < \text{FPro} < 1$).

Organization of ingredients in the food supply

Food and beverage companies are required to report the list of ingredients in descending order of the amount used in the final product. When an ingredient itself is a composite, consisting of two or more ingredients, the US Food and Drug Administration (FDA) mandates parentheses to declare the corresponding sub-ingredients (Fig. 5a,b)[45]. By organizing the ingredient list as a tree (Methods), differences between highly processed and less processed options can be analysed (Fig. 5). In general, products with complex ingredient trees are more processed than products with simpler and fewer ingredients (Supplementary Section 9.3). For example, the ultra-processed cheesecake in Fig. 5a has 43 ingredients, 26 additives and 3 branches with sub-ingredients. By contrast, the minimally processed cheesecake has only 14 ingredients, 5 additives and 2 sub-ingredient branches (Fig. 5b). As illustrated by the cheesecake example, the ingredients used in the food supply provide valuable insights into the type and extent of processing of the final product, prompting the question: which ingredients contribute the most to the degree of processing of a product? To answer this, we introduce the Ingredient Processing Score (IgFPro), defined as

$$\text{IgFPro}(g) = \frac{\sum_{f \in F_g} r_g^f \times \text{FPro}^f}{\sum_{f \in F_g} r_g^f}, \quad (1)$$

where r_g^f ranks an ingredient g in decreasing order based on its position in the ingredient list of each food f that contains g (Supplementary Section 9.5). IgFPro ranges between 0 (unprocessed) and 1 (ultra-processed), enabling the rank order of ingredients based on their contribution to the degree of processing of the final product. This analysis reveals that not all additives contribute equally to ultra-processing. For example, the ultra-processed cheesecake (Fig. 5a) has polysorbate 60 (an emulsifier used in cakes for increased volume and fine grain with IgFPro = 0.908) and corn syrup (a corn sweetener with IgFPro = 0.905)[46], each of which emerges as signals of ultra-processing with high IgFPro scores. By contrast, both the minimally processed and ultra-processed cheesecakes (Fig. 5) contain xanthan gum (IgFPro = 0.818), guar gum (IgFPro = 0.801), locust bean gum (IgFPro = 0.786) and salt (IgFPro = 0.777). Indeed, the European Food Safety Authority reported that xanthan gum as a food additive does not pose any safety concern for the general population, and the FDA classified guar gum and locust bean gum as 'generally recognized as safe'[46].

By the same token, when evaluating oils used as ingredients in branded products, IgFPro identifies brain octane oil (IgFPro = 0.573), flaxseed oil (IgFPro = 0.69) and olive oil (IgFPro = 0.722) as the highest quality options, having the smallest contribution to ultra-processing. On the other hand, palm oil (IgFPro = 0.888), vegetable oil (IgFPro = 0.866) and soybean oil (IgFPro = 0.862) represent strong signals of ultra-processing (Fig. 6a). It is worth noting that flaxseed oil is high in omega-3 fatty acids with several health benefits[47]. By contrast, blending

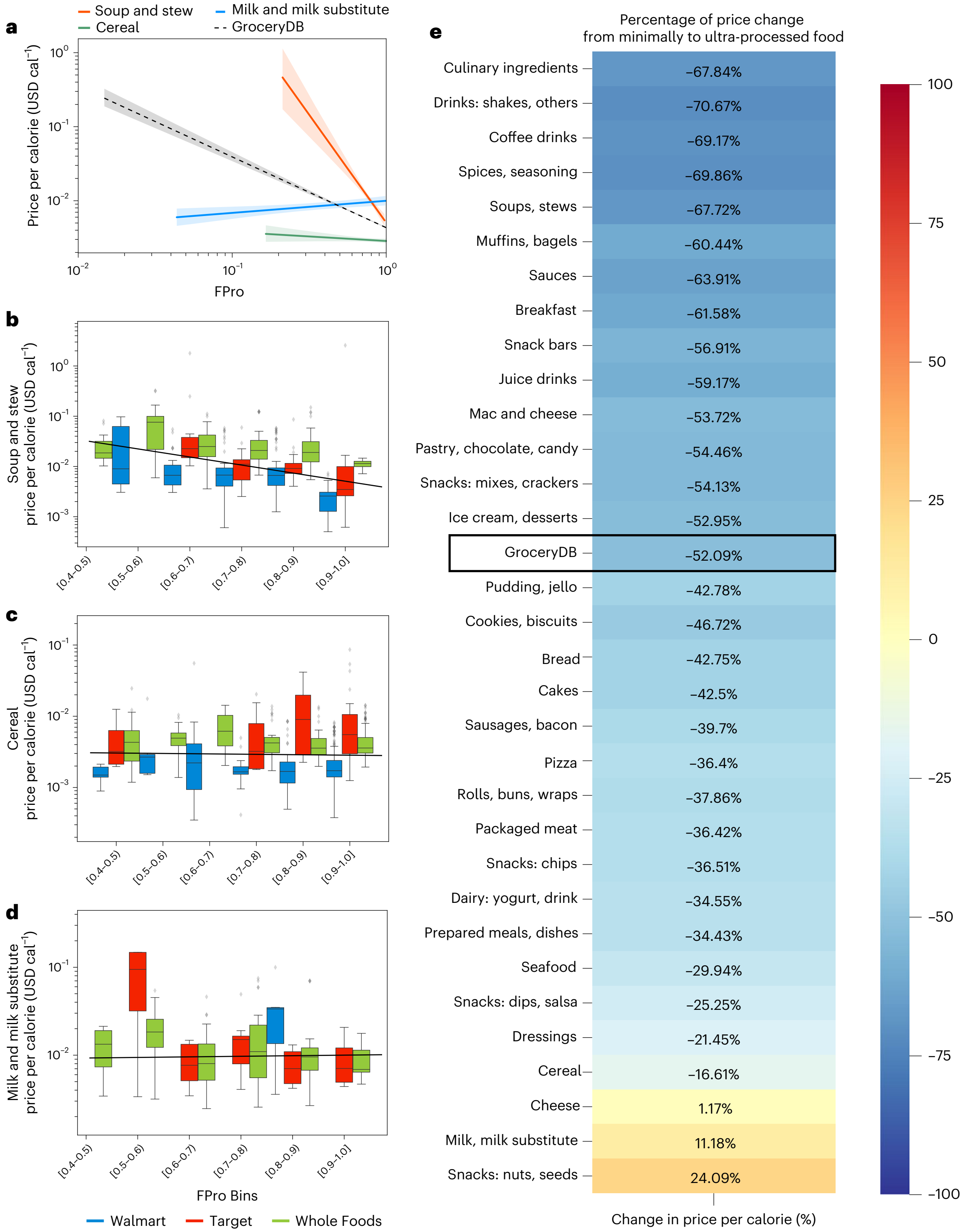


Fig. 3 | Price and food processing. a, Using robust linear models, the relationship between price and food processing are assessed (Supplementary Fig. 8 for regression coefficients of all categories). The price per calorie drops by 24.3% (soup and stew, *n* = 505) and 1.2% (cereal, *n* = 659) for 10% increase in FPro. Also, an 8.7% decrease is observed across all foods in GroceryDB (*n* = 19,345) for 10% increase in FPro. It is worth noting that in milk and milk substitute (*n* = 240), price per calorie increases by 1.6% for 10% increase in FPro, partially explained by the higher price of plant-based milks that are more processed than regular dairy milk. The shaded area for each line is the 95% confidence interval of the standard error. **b**–**d**, Distributions of price per calorie in the linear bins of FPro scores for each store (Supplementary Fig. 7 illustrates the correlation between price and FPro for all categories). In soup and stew (**b**), there is a steep decreasing slope between FPro and price per calorie, while in cereals (**c**) the effect is smaller. In milk and milk substitute (**d**), price tends to slightly increase with higher values of FPro. For the box plots, the minimum is the lower quartile, the central line is the median and the maximum is the upper quartile. The whiskers show data outside of the inter-quartile range. Diamonds represent outliers. **e**, Percentage of change in price per calorie from the minimally processed products to ultra-processed products in different food categories. This analysis was performed by comparing the average price per calorie of the top 10% most processed items with the top 10% least processed items within each category. In the full GroceryDB, on average, the ultra-processed items are 52.09% cheaper than their minimally processed alternatives. *n* > 4 for all statistics; see Source Data Figs. 1–6 for exact sample sizes.

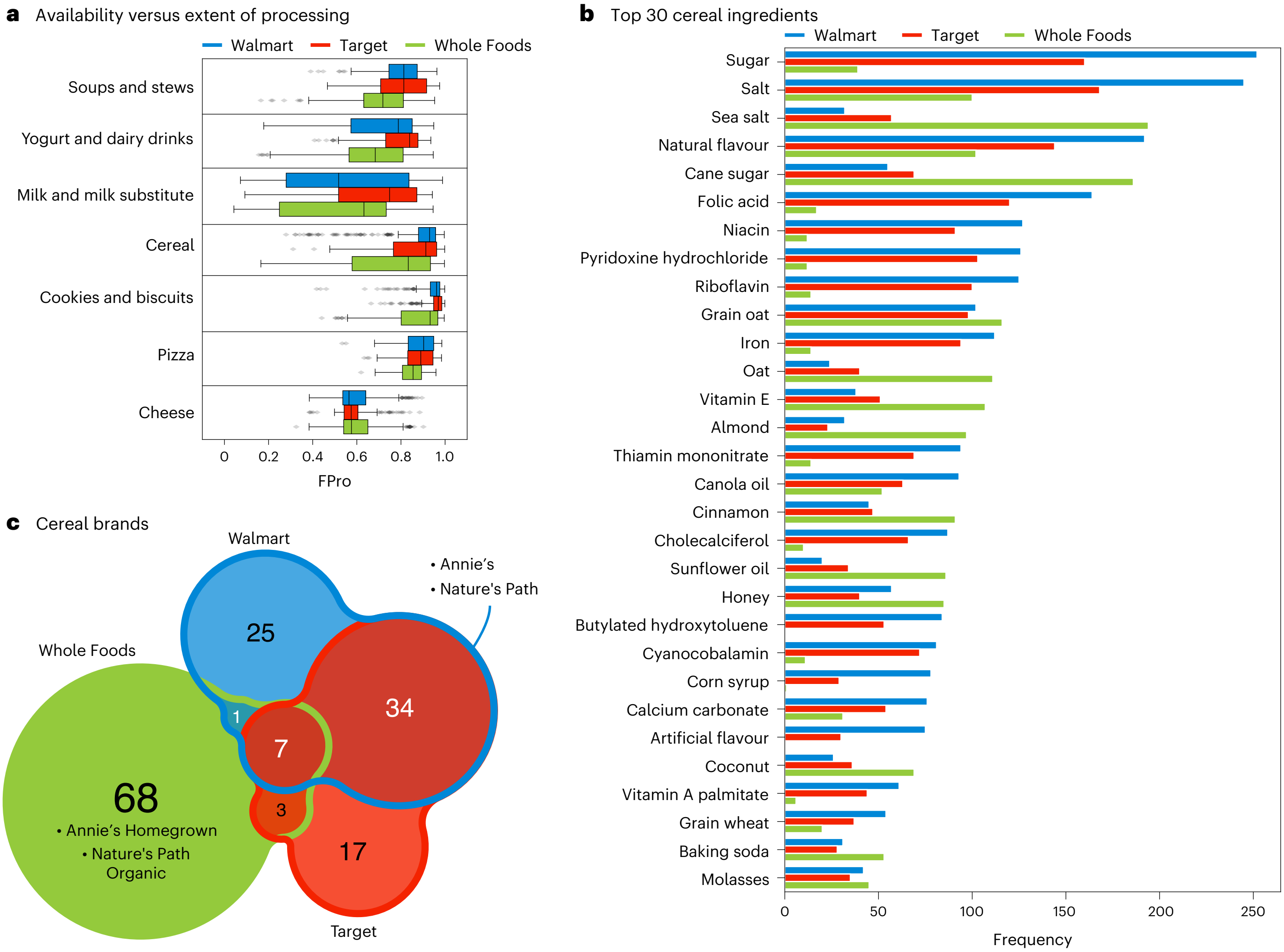


Fig. 4 | The difference between stores in terms of processing. The degree of processing of food choices depends on the grocery store and food category. **a**, The degree of processing of food items offered in grocery stores, stratified by food category. For example, in cereals, Whole Foods shows a higher variability of FPro, implying that consumers have a choice between low and high processed cereals. Yet, in pizzas, all supermarkets offer choices characterized by high FPro values. Lastly, all cheese products are minimally processed, showing consistency across different grocery stores. For the box plots, the minimum is the lower quartile, the central line is the median, and the maximum is the upper quartile. The whiskers show data outside of the inter-quartile range. Diamonds represent outliers. $n > 4$ for all statistics; see Source Data Figs. 1–6 for exact sample sizes. **b**, The top 30 most reported ingredients in cereals show that Whole Foods tends to eliminate corn syrup, uses more sunflower oil and less canola oil and relies less on vitamin fortification. In total, GroceryDB has 1,168 cereals from which 973 have ingredient lists (Walmart = 309, Target = 260, Whole Foods = 395). **c**, The brands of cereals offered in stores partially explains the different patterns of ingredients and variation of FPro. While Walmart and Target have a larger intersection in the brands of their cereals (for example, Annie's and Nature's Path), Whole Foods tends to supply cereals from brands not available elsewhere (for example, Annie's Homegrown and Nature's Path Organic).

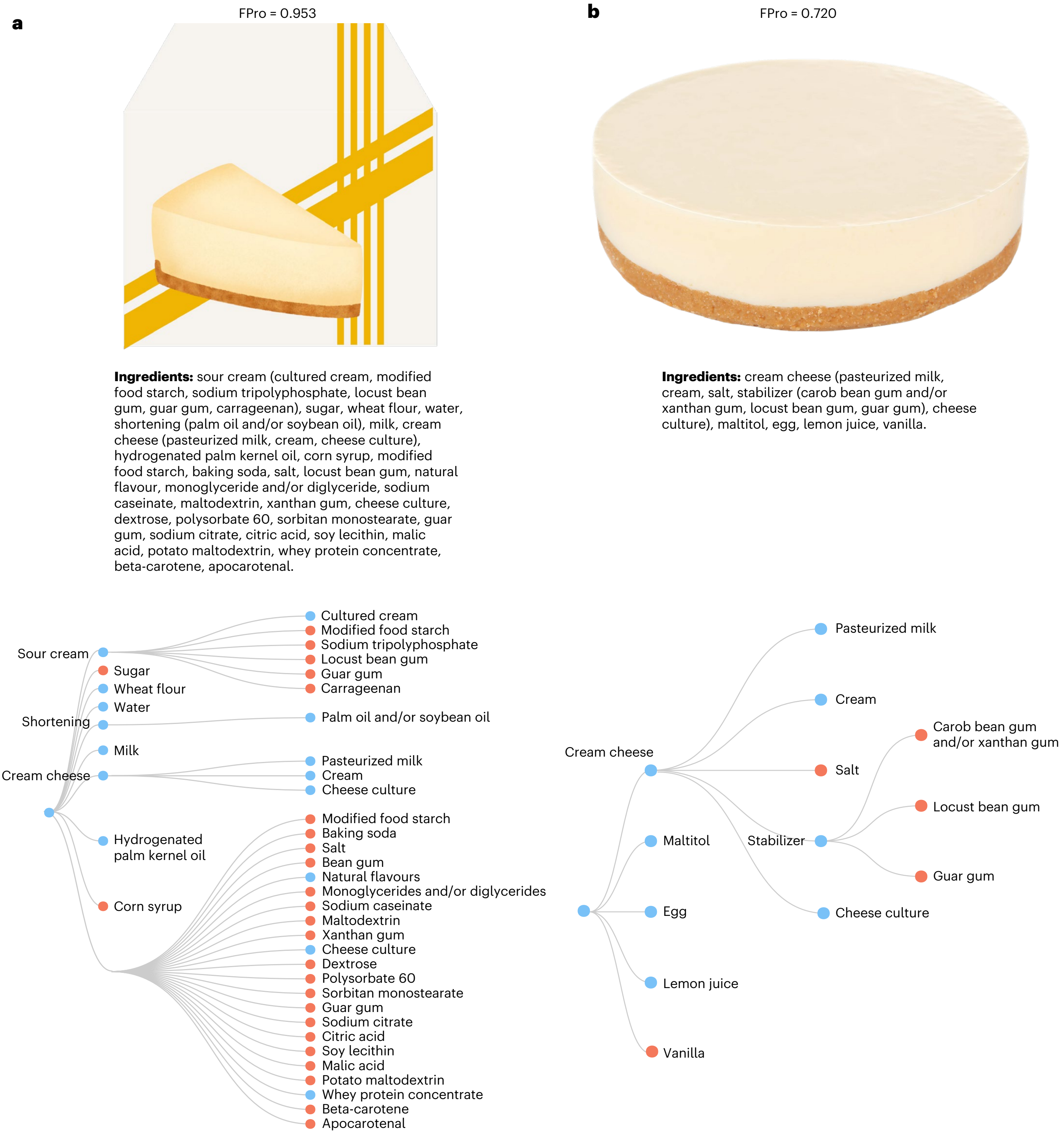
of vegetable oils—a signature of UPF—is a straightforward practice to achieve desired texture, stability and nutritional profiles[48].

Finally, to illustrate the ingredient patterns characterizing UPF in Fig. 6b, three tortilla chips are ranked from 'minimally processed' to ultra-processed. Relative to the snack-chips category, Siete tortilla is minimally processed (FPro = 0.477), made with avocado oil and blend of cassava and coconut flours. The more processed El Milagro tortilla (FPro = 0.769) is cooked with corn oil and ground corn and has calcium hydroxide, a generally-recognized-as-safe additive made by adding water to calcium oxide (lime) to promote dispersion of ingredients[46]. By contrast, the ultra-processed Doritos (FPro = 0.982) have corn flour and a blend of vegetable oils and rely on 12 additives to ensure a palatable taste and the texture of the tortilla chip, demonstrating the complex patterns of ingredients and additives needed for ultra-processing (Fig. 6b).

In summary, complex ingredient patterns accompany the production of UPF (Supplementary Section 9.4). IgFPro enables the assessment of processing characteristics across the entire food supply, as well as the contribution of individual ingredients.

Discussion

GroceryDB, accessible to the public at https://www.TrueFood.tech/, offers both the data and methodologies needed to quantify food processing and analyse the structure of ingredients within the US food supply. By combining large-scale data on food composition and machine learning, GroceryDB uncovers insights on the current state of food processing in the US grocery landscape, obtaining distributions of FPros that capture a remarkable variability in the offerings of different grocery stores. The differences in FPro's distributions (Fig. 2a) indicate that multiple factors drive the range of choices available in grocery stores, from the cost of food and the socioeconomic status of the consumers to the distinct declared missions of the supermarket chains: 'quality is a state of mind' for Whole Foods Market and 'helping people save money so they can live better' for Walmart[49,50]. Furthermore, the



Ingredients: sour cream (cultured cream, modified food starch, sodium tripolyphosphate, locust bean gum, guar gum, carrageenan), sugar, wheat flour, water, shortening (palm oil and/or soybean oil), milk, cream cheese (pasteurized milk, cream, cheese culture), hydrogenated palm kernel oil, corn syrup, modified food starch, baking soda, salt, locust bean gum, natural flavour, monoglyceride and/or diglyceride, sodium caseinate, maltodextrin, xanthan gum, cheese culture, dextrose, polysorbate 60, sorbitan monostearate, guar gum, sodium citrate, citric acid, soy lecithin, malic acid, potato maltodextrin, whey protein concentrate, beta-carotene, apocarotenal.

Ingredients: cream cheese (pasteurized milk, cream, salt, stabilizer (carob bean gum and/or xanthan gum, locust bean gum, guar gum), cheese culture), maltitol, egg, lemon juice, vanilla.

Fig. 5 | Ingredient trees. GroceryDB organizes the ingredient list of products into structured trees, where the additives are marked as orange nodes (Methods and Supplementary Section 9). **a**, Edwards Desserts Original Whipped Cheesecake is a highly processed cheesecake that contains 43 ingredients from which 26 are additives, resulting in a complex ingredient tree with 3 branches of sub-ingredients. **b**, Pearl River Mini No Sugar Added Cheesecake is a minimally processed cheesecake that has a simpler ingredient tree with 14 ingredients, 5 additives and 2 sub-ingredient branches. Additives are identified according to the FDA[77,78]. See Source Data Fig. 5. Credit: watercolour cheesecake illustration and gold line stripes, Canva.com; delicious cheesecake on white background, Africa Studio, Adobe Stock.

continuous nature of FPro allows for data-driven investigations on the relationship between price and food processing stratified by food category. Overall, food processing in GroceryDB tends to be associated with the production of more affordable calories, a positive correlation that raises the likelihood of habitual consumption among lower-income populations, ultimately contributing to growing socioeconomic disparities in terms of nutrition security[51–56]. However, it is important to note that the strength and direction of this correlation varies depending on the specific food category under consideration, as exemplified by the opposite trend of milk and milk substitutes compared with soups

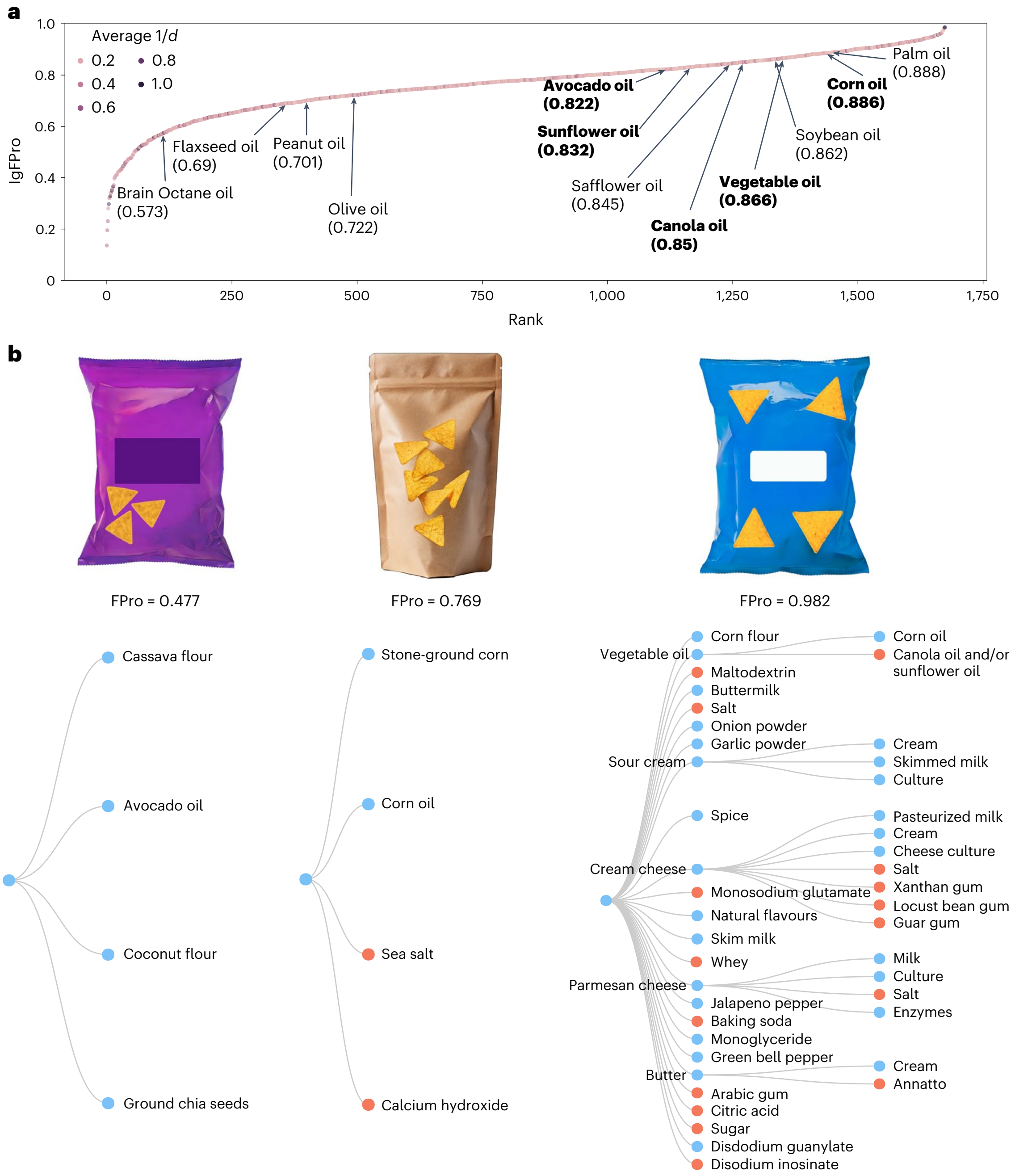


Fig. 6 | IgFPro. To investigate which ingredients contribute most to ultra-processed products, equation (1) is used. With the introduction of IgFPro, over 12,000 ingredients are ranked by their prevalence and contribution to ultra-processed products, prioritizing ingredients and food groups for targeted intervention. A total of 1,676 ingredients are in more than 10 products. **a**, The IgFPro of all ingredients that appeared in at least 10 products are calculated by rank-ordering ingredients based on their contribution to UPF. The ingredients are colored based on their distance to the root node, *d*, of the ingredient tree (Methods). The popular oils used as an ingredient are highlighted, with the Brain Octane, flaxseed and olive oils contributing the least to ultra-processed products. By contrast, the palm, vegetable and soybean oils contribute the most to ultra-processed products (Supplementary Section 9.5). **b**, The patterns of ingredients in the least-processed tortilla chips versus the ultra-processed tortilla chips. The IgFPro values of the oils used in the three tortilla chips are highlighted in bold in **a**. The minimally processed Siete tortilla chips (FPro = 0.477) uses avocado oil (IgFPro = 0.822), and the more processed El Milagro tortilla (FPro = 0.769) uses corn oil (IgFPro = 0.886). By contrast, the ultra-processed Doritos (FPro = 0.982) relies on a blend of vegetable oils (IgFPro = 0.866) and is accompanied by a much more complex ingredient tree, indicating that there is no single ingredient 'biomarker' for UPF. Ingredient trees contain both ingredients (blue) and additives (orange). Additives are identified according to the FDA[77,78]. Credit: food packaging (foil and plastic snack bags mockup isolated on white background, purple coloured pillow packages for food production on white PNG file), Juraiwan, Adobe Stock; brown paper bag, Graphic, Adobe Stock; foil and plastic snack bags mockup isolated on white background, dark blue coloured pillow packages for food production, snack wrappers on white background with clipping path, MERCURY, Adobe Stock; flying Mexican nachos chips isolated on white background, Yeti Studio, Adobe Stock; nachos and tortilla chip illustration, Canva.com.

and stews (Supplementary Section 8). Further in-depth analyses are needed to evaluate the effectiveness of intervention strategies targeting specific food groups within diverse food environments.

Governments increasingly acknowledge the impact of processed foods on population health and its long-term effect on healthcare[57,58]. For example, the UK spends £18 billion annually on direct medical costs related to non-communicable diseases like obesity[59], while the United States incurs $1.1 trillion in yearly food-related human health costs[60,61]. GroceryDB serves as a valuable resource for both consumers and policymakers, offering essential insights to gauge the level of food processing within the food supply. For instance, in categories like cereals, milk and milk alternatives, pasta noodles and snack bars, FPro shows a wide range, highlighting the substantial variations in the processing levels of products. If consumers had access to this processing data, they could make informed choices, selecting items with markedly different degrees of processing (Fig. 2b). Yet, the comprehension of nutrient and ingredient data disclosed on food packaging often poses a challenge to consumers due to unrealistic serving sizes and confusing health claims based on one or a few nutrients. Our primary objective lies in translating this wealth of data into an actionable scoring system, enabling consumers to make healthier food choices and embrace effective dietary substitutions, without overwhelming them with excessive information. In addition, this approach holds great potential for public health initiatives aimed at improving the overall quality of the food environment, such as strategies reorganizing supermarket layouts, optimizing shelf placements and thoughtfully designing counter displays[53,62,63]. Transforming health-related behaviours is a challenging task[64,65]; hence, easily adoptable dietary modifications along with environmental nudges could make it easier for individuals to embrace healthier choices.

Currently, FPro partially draws from expertise-based food processing classifications due to limited data concerning compound concentrations indicative of food matrix alterations, such as cellular wall transformations or industrial processing techniques. However, a comprehensive mapping of the 'dark matter of nutrition', encompassing chemical concentrations for additives and processing by-products, aims to evolve FPro into an unsupervised system, independent of manual classifications[66,67]. Unlike expertise-based systems, FPro functions as a quantitative algorithm, using standardized inputs to generate reproducible continuous scores, facilitating sensitivity analysis and uncertainty estimations[37] (Supplementary Section 5). These important features enhance reliability, transparency and interpretability of the analyses while reducing errors associated with the descriptive nature of manual classifications[28], which have shown a low degree of consistency among nutrition specialists[30].

The chemical composition of branded products is partially captured by the nutrition facts table and partially reported in the ingredient list, which includes additives such as artificial colours, flavours and emulsifiers. However, comprehensive and internationally well-regulated data on food ingredients are currently limited, as documented by the GS1 UK data crunch analysis which reported an average of 80% inconsistency in products' data[31], leading us to focus on the nutrition facts to enhance the algorithm's portability and reproducibility. The nutrition facts alone show excellent performance in discriminating between NOVA classes, confirming how food processing consistently alters nutrient concentrations with reproducible patterns, effectively harnessed by machine learning[37]. While FPro assesses the degree of food processing by holistically evaluating nutrient concentrations, the few nutrients available on food packaging increase the risk of identifying products with similar nutrition facts but distinct food matrices (for example, pre-frying, puffing, extrusion-cooking). Indeed, if the chemical panel used to train the algorithm fails to exhaustively capture matrix modifications induced by processing and cooking, FPro and the substitution algorithm implemented at https://www.TrueFood.tech/ remain blind to these chemical-physical changes. Incorporating disambiguated ingredients in FPro, such as the ultra-processing markers characterized by SIGA[68], may offer a solution until larger composition tables for branded products become available (Supplementary Section 5).

In summary, this work represents a departure from traditional food classification systems, advancing toward the use of machine learning methodologies to model the chemical complexity of food[69] (Supplementary Section 1). Despite the limited information provided by the FDA-regulated nutrition labels, GroceryDB and FPro offer a data-driven approach that enables a substitution algorithm capable of recommending similar but less processed alternatives for any food in GroceryDB. Together, GroceryDB and the TrueFood platform highlight the importance of data transparency in grocery store inventories, a key factor that directly shapes consumer choices.

Methods

Data collection

Publicly accessible data on food products were compiled from the online platforms of Walmart, Target and Whole Foods. Each store organizes its food items hierarchically. Using these categorizations, the stores' websites are systematically navigated to identify specific food items. To ensure consistency, the food category hierarchy within GroceryDB is standardized by comparing and aligning the classification systems used by each store. The stores sourced nutrition facts from physical food labels and provided digital versions for each food item. These data allowed us to standardize nutrient concentrations to a uniform measure of 100 g and use FoodProX to evaluate the degree of food processing for each item. Lastly, all data for this manuscript were collected in May 2021.

Calculation of the FPro

Processing alters the nutrient profile of food, changes that are detectable and categorizable using machine learning[37,69,70]. Hence, FoodProX[37], a random forest classifier, translates the combinatorial changes in the nutrient amounts induced by food processing into a FPro. Extensive tests and validations on the stability of FPro were performed in several databases such as the US FNDDS and the international OFF. FPro enabled the implementation of an in silico study based on US cross-sectional population data, showing that on average substituting only a single food item in a person's diet with a minimally processed alternative from the same food category can reduce the risk of developing metabolic syndrome (12.25% decrease in odds ratio) and increase vitamin blood levels (4.83% and 12.31% increase of vitamin B12 and vitamin C blood concentration)[37].

FoodProX takes as input 12 nutrients reported in the nutrition facts (Supplementary Table 1) and returns FPro, a continuous score ranging between 0 (unprocessed foods such as fruits and vegetables) and 1 (UPF such as instant soups and shelf-stable breads). The manual NOVA classifications were applied to the USDA Standard Reference and FNDDS databases to train FoodProX. In the original classification, NOVA labels were assigned by inspecting the ingredient list and the food description but without taking into account nutrient content.

FPro does not assess individual nutrients in isolation but, rather, learns from the configurations of correlated nutrient changes within a fixed quantity of food (100 g)[37]. Consequently, a single high or low nutrient value does not dictate a food's FPro. Instead, the final score depends on the likelihood of observing the overall pattern of nutrient concentrations in unprocessed food versus UPF. For instance, while fortified food may mirror mineral and vitamin content in unprocessed food, the algorithm identifies unique concentration signatures unlikely to be found in minimally processed food, resulting in a higher FPro[37].

The calculation of FPro for all food in GroceryDB represents a generalization task, where the model faces 'never-before-seen' data[69,71]. More details on the training dataset, including class heterogeneity and imbalance, are available in Supplementary Section 4.

Price for calories trends

Robust linear models with Huber's t-norm[72–74] were applied to calculate regression coefficients and P values for the relationship log(PricePerCalorie) ~ log(FPro). The detailed regression results for each food category are presented in Supplementary Fig. 8, while the overall trend across GroceryDB is depicted in Fig. 3a. To illustrate the price disparity at the extremes of food processing, the percentage change in price per calorie shown in Fig. 3e WAS calculated by comparing the average price per calorie of the top 10% minimally processed items to that of the top 10% ultra-processed items within each category.

Ingredient trees

An ingredient list is a reflection of the recipe used to prepare a branded food item. The ingredient lists are sorted based on the amount of ingredients used in the preparation of an item as required by the FDA. An ingredient tree can be created in two ways: (a) with emphasis on capturing the main and sub-ingredients, similar to a recipe, as illustrated in Supplementary Fig. 17a; (b) with emphasis on the order of ingredients as a proxy for their amount in a final product, as illustrated in Supplementary Fig. 17b, where the distance from the root, d, reflects the amount of an individual ingredient relative to all ingredients. We opted for (b) to calculate IgFPro, as ranking the amount of an ingredient in a food is essential to quantify the contribution of individual ingredients to ultra-processing. In equation (1), $r_g^f = 1/d_g^f$ ranks the amount of an ingredient g in food f, where d_g^f captures the distance from the root (Supplementary Fig. 17b for an example). Finally, IgFPro shows remarkable variability when compared with the average FPro of products containing the selected ingredient (Supplementary Fig. 18), suggesting distinctive patterns of correlation between the products' FPro and the ranking of ingredients in their ingredient lists[75].

Database structure

The database comprises two main files, both stored in CSV format for ease of use and accessibility:

1. GroceryDB Foods File. This file contains comprehensive information about all the foods included in GroceryDB. Each row represents a distinct food item. This file includes the following columns:

   - **name**: The name of the food item, typically as it appears on the product packaging.
   - **brand**: The brand or manufacturer of the food item.
   - **harmonized single category**: The general category or type of food (for example, seafood, cereal and so on).
   - **store**: The retail store where the food item is available (for example, Walmart, Target, Whole Foods).
   - **f_FPro**: Average FPro score of the food across the ensemble of classifiers. The FPro score is calculated using the FoodProX algorithm, taking into account the nutrition facts of the food.
   - **f_FPro_P**: A string indicating whether the food has enough nutritional descriptors as detailed in Supplementary Section 4.
   - **f_min_FPro**: Minimum FPro score across the ensemble of classifiers.
   - **f_std_FPro**: The standard deviation of the FPro score across the ensemble of classifiers.
   - **f_FPro_class**: Expected NOVA class assigned according to FoodProX.
   - **has10_nuts**: Boolean value indicating whether the food is described by the 10 key nutrients described in Supplementary Section 4.
   - **is_Nuts_Converted_100g**: Indicator whether the food nutrients are converted per 100 g.
   - **nutritional information**: Detailed nutritional information for the food item, including protein, total fat, carbohydrate, total sugars, total dietary fibre, calcium, iron, sodium, vitamin C, cholesterol, total saturated fatty acids and total vitamin A.

   Please note that the prices of the food items are not included in this public release due to potential restrictions on public disclosure. However, this information is available upon request. The file is available at https://github.com/Barabasi-Lab/GroceryDB/blob/main/data/GroceryDB_foods.csv.

2. GroceryDB IgFPro File. This file contains data related to the IgFPro score of the ingredients listed in GroceryDB. Each row corresponds to a specific ingredient. The file is available at https://github.com/Barabasi-Lab/GroceryDB/blob/main/data/GroceryDB_IgFPro.csv. The columns in this file are as follows:

   - **ingredient_name**: The standardized name of the ingredient.
   - **count_of_products**: The total number of products in the database that contain this ingredient.
   - **ingredient_FPro**: IgFPro calculated for the selected ingredient.
   - **average_FPro_of_products**: The average FPro score of the products containing the selected ingredient.
   - **average_distance_to_root**: The average distance of the ingredient from the root in the ingredient tree, representing its relative amount in the food item. Ingredients closer to the root contribute more to the calculation of IgFPro.
   - **ingredient_normalization_term**: A numerical value used to normalize a food's contribution to the IgFPro score, based on the ingredient's overall ranking across all foods.

Substitution algorithm at TrueFood.Tech

The site https://www.TrueFood.tech/ provides food substitution recommendations aimed at gently nudging consumers towards less processed alternatives. To accomplish this, we first identify food items that belong to the same category and share partial semantic similarity with the targeted item (range 0.10–0.95), based on both food names and ingredient lists. This approach increases the diversity of displayed recommendations while ensuring they remain within the same category.

The popular term frequency-inverse document frequency (Tf-idf) algorithm is used to measure the significance of words to foods in GroceryDB, adjusting for commonality across entries[76]. The similarity between weighted word vectors is calculated leveraging cosine similarity. The final similarity between the queried food and other food items is determined by multiplying the ingredient-list-based similarity and the food-name-based similarity.

Next, the semantically filtered foods are sorted by their FPro scores, ranking the recommendations in ascending order of FPro. This method can identify the most similar food items with a lower FPro compared with the targeted item. Up to 50 items, listed in increasing order of FPro, are displayed on the website.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data in GroceryDB was scraped from Walmart, Target and Whole Foods in 2021. GroceryDB is available to the public and consumers at https://www.TrueFood.tech/. The data are also openly available on

MongoDB servers with a read-only key available via BarabasiLab GitHub repository at https://github.com/Barabasi-Lab/GroceryDB/. The USDA FNDDS dataset is available via the same GitHub repository. Source data are provided with this paper.

Code availability

All code generated for the analysis are available via the BarabasiLab GitHub repository at https://github.com/Barabasi-Lab/GroceryDB/. The analysis was done in Python==3.11.7 with the following packages: jupyter notebook==6.5.4, pymongo==4.8.0, pandas==2.1.4, numpy==1.26.4, seaborn==0.12.2, statsmodels==0.14.0, scipy==1.11.4, matlabplot==3.8.0, plotly==5.9.0 and certifi==2024.6.2.

References

1. Seferidi, P. et al. The neglected environmental impacts of ultra-processed foods. *Lancet Planet. Health* **4**, e437–e438 (2020).
2. Fardet, A. & Rock, E. Ultra-processed foods and food system sustainability: what are the links? *Sustainability* **12**, 6280 (2020).
3. Macdiarmid, J. I. The food system and climate change: are plant-based diets becoming unhealthy and less environmentally sustainable? *Proc. Nutr. Soc.* **81**, 162–167 (2022).
4. Ambikapathi, R. et al. Global food systems transitions have enabled affordable diets but had less favourable outcomes for nutrition, environmental health, inclusion and equity. *Nat. Food* **3**, 764–779 (2022).
5. Lane, M. M. et al. Ultra-processed food exposure and adverse health outcomes: umbrella review of epidemiological meta-analyses. *BMJ* **384**, e077310 (2024).
6. Lustig, R. H. Processed food—an experiment that failed. *JAMA Pediatr.* **171**, 212–214 (2017).
7. Milanlouei, S. et al. A systematic comprehensive longitudinal evaluation of dietary factors associated with acute myocardial infarction and fatal coronary heart disease. *Nat. Commun.* **11**, 1–14 (2020).
8. 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. *Popul. Health Metr.* **15**, 6 (2017).
9. Monteiro, C. A. et al. NOVA. The star shines bright. *World Nutr. J.* **7**, 28–38 (2016).
10. Steele, E. M. et al. Ultra-processed foods and added sugars in the U.S. diet: evidence from a nationally representative cross-sectional study. *BMJ Open* **6**, e009892 (2016).
11. Steele, E. M. & Monteiro, C. A. Association between dietary share of ultra-processed foods and urinary concentrations of phytoestrogens in the US. *Nutrients* **9**, 209 (2017).
12. Adjibade, M. et al. Prospective association between ultra-processed food consumption and incident depressive symptoms in the French NutriNet-Santé cohort. *BMC Med.* **17**, 1–13 (2019).
13. Fiolet, T. et al. Consumption of ultra-processed foods and cancer risk: results from NutriNet-Santé prospective cohort. *BMJ* **360**, k322 (2018).
14. Srour, B. et al. Ultra-processed food intake and risk of cardiovascular disease: prospective cohort study (NutriNet-Santé). *BMJ* **365**, l1451 (2019).
15. 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 Metab.* **30**, 1–11 (2019).
16. 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).
17. Nerín, C., Aznar, M. & Carrizo, D. Food contamination during food process. *Trends Food Sci. Technol.* **48**, 63–68 (2016).
18. Rather, I. A., Koh, W. Y., Paek, W. K. & Lim, J. The sources of chemical contaminants in food and their health implications. *Front. Pharmacol.* **8**, 830 (2017).
19. Arisseto, A. P. Furan in processed foods. In *Food Hygiene and Toxicology in Ready-to-Eat Foods* (ed. Kotzekidou, P.) Ch. 21, 383–396 (Academic, 2016).
20. 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. *Environ. Int.* **131**, 105057 (2019).
21. Mozaffarian, D., Fleischhacker, S. & Andrés, J. R. Prioritizing nutrition security in the US. *JAMA* **325**, 1605–1606 (2021).
22. Livings, M. S. et al. Food and nutrition insecurity: experiences that differ for some and independently predict diet-related disease, Los Angeles County, 2022. *J. Nutr.* **154**, 2566–2574 (2024).
23. *Food and Nutrition Security* (USDA, 2024); https://www.usda.gov/about-usda/general-information/priorities/food-and-nutrition-security
24. Volpp, K. G. et al. Food is medicine: a presidential advisory from the American Heart Association. *Circulation* **148**, 1417–1439 (2023).
25. Mozaffarian, D., Andrés, J. R., Cousin, E., Frist, W. H. & Glickman, D. R. The White House Conference on Hunger, Nutrition and Health is an opportunity for transformational change. *Nat. Food* **3**, 561–563 (2022).
26. Mozaffarian, D., Rosenberg, I. & Uauy, R. History of modern nutrition science-implications for current research, dietary guidelines, and food policy. *BMJ* **361**, k2392 (2018).
27. Sadler, C. R. et al. Processed food classification: conceptualisation and challenges. *Trends Food Sci. Technol.* **112**, 149–162 (2021).
28. Gibney, M. J. & Forde, C. G. Nutrition research challenges for processed food and health. *Nat. Food* **3**, 104–109 (2022).
29. Lacy-Nichols, J. & Freudenberg, N. Opportunities and limitations of the ultra-processed food framing. *Nat. Food* **3**, 975–977 (2022).
30. Braesco, V. et al. Ultra-processed foods: how functional is the NOVA system? *Eur. J. Clin. Nutr.* **76**, 1245–1253 (2022).
31. *Data Crunch Report: The Impact of Bad Data on Profits and Customer Service in the UK Grocery Industry* (accessed April 4, 2022) (GS1 UK and Cranfield University School of Management, 2009); https://dspace.lib.cranfield.ac.uk/bitstream/handle/1826/4135/Data_crunch_report.pdf
32. *THE 17 GOALS | Sustainable Development* (United Nations, 2020); https://sdgs.un.org/goals
33. *Methods and Standards* (Food and Agriculture Organization of the United Nations, 2021); https://www.fao.org/statistics/methods-and-standards/en/
34. Sarku, R., Clemen, U. A. & Clemen, T. The application of artificial intelligence models for food security: a review. *Agriculture* **13**, 2037 (2023).
35. Hu, G., Ahmed, M. & L'Abbé, M. R. Natural language processing and machine learning approaches for food categorization and nutrition quality prediction compared with traditional methods. *Am. J. Clin. Nutr.* **117**, 553–563 (2023).
36. *Impact Initiative* (AI for Good, 2019); https://aiforgood.itu.int/
37. Menichetti, G., Ravandi, B., Mozaffarian, D. & Barabási, A.-L. Machine learning prediction of the degree of food processing. *Nat. Commun.* **14**, 2312 (2023).
38. Chen, X. et al. Consumption of ultra-processed foods and health outcomes: a systematic review of epidemiological studies. *Nutr. J.* **19**, 86 (2020).
39. Mendoza, K. et al. Ultra-processed foods and cardiovascular disease: analysis of three large US prospective cohorts and a systematic review and meta-analysis of prospective cohort studies. *Lancet Reg. Health Am.* **37**, 100859 (2024).

40. 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. *Eur. J. Clin. Nutr.* **63**, S206–S225 (2009).
41. Poti, J. M., Mendez, M. A., Ng, S. W. & Popkin, B. M. Is the degree of food processing and convenience linked with the nutritional quality of foods purchased by US households? *Am. J. Clin. Nutr.* **101**, 1251–1262 (2015).
42. Davidou, S., Christodoulou, A., Fardet, A. & Frank, K. The holistico-reductionist SIGA classification according to the degree of food processing: an evaluation of ultra-processed foods in French supermarkets. *Food Funct.* **11**, 2026–2039 (2020).
43. *U.S. Population: Consumption of Breakfast Cereals (Cold) from 2011 to 2024* (accessed February 2022) (Statista, 2021); https://www.statista.com/statistics/281995/us-households-consumption-of-breakfast-cereals-cold-trend/
44. Bray, G. A., Nielsen, S. J. & Popkin, B. M. Consumption of high-fructose corn syrup in beverages may play a role in the epidemic of obesity. *Am. J. Clin. Nutr.* **79**, 537–543 (2004).
45. *Guidance for Industry: Food Labeling Guide* (accessed 1 November 2021) (USFDA, 2021); https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-industry-food-labeling-guide
46. Igoe, R. S. *Dictionary of Food Ingredients* (Springer Science & Business Media, 2011).
47. Goyal, A., Sharma, V., Upadhyay, N., Gill, S. & Sihag, M. Flax and flaxseed oil: an ancient medicine & modern functional food. *J. Food Sci. Technol.* **51**, 1633–1653 (2014).
48. Hashempour-Baltork, F., Torbati, M., Azadmard-Damirchi, S. & Savage, G. P. Vegetable oil blending: a review of physicochemical, nutritional and health effects. *Trends Food Sci. Technol.* **57**, 52–58 (2016).
49. *Whole Foods Mission and Values* (accessed 1 March 2022) (2012); https://www.WholeFoodsmarket.com/mission-values
50. *Walmart History* (accessed 1 March 2022) (2022); https://corporate.walmart.com/about/history
51. Gupta, S., Hawk, T., Aggarwal, A. & Drewnowski, A. Characterizing ultra-processed foods by energy density, nutrient density, and cost. *Front. Nutr.* **6**, 70 (2019).
52. Zenk, S. N., Tabak, L. A. & Pérez-Stable, E. J. Research opportunities to address nutrition insecurity and disparities. *JAMA* **327**, 1953–1954 (2022).
53. Venkataramani, A. S., O'Brien, R., Whitehorn, G. L. & Tsai, A. C. Economic influences on population health in the United States: toward policymaking driven by data and evidence. *PLoS Med.* **17**, e1003319 (2020).
54. Erndt-Marino, J., O'Hearn, M. & Menichetti, G. An integrative analytical framework to identify healthy, impactful, and equitable foods: a case study on 100% orange juice. *Int. J. Food Sci. Nutr.* **74**, 668–684 (2023).
55. Coletro, H. N. et al. The combined consumption of fresh/minimally processed food and ultra-processed food on food insecurity: COVID Inconfidentes, a population-based survey. *Public Health Nutr.* **26**, 1414–1423 (2023).
56. Hutchinson, J. & Tarasuk, V. The relationship between diet quality and the severity of household food insecurity in Canada. *Public Health Nutr.* **25**, 1013–1026 (2022).
57. Griffith, R., Jenneson, V., James, J. & Taylor, A. *The Impact of a Tax on Added Sugar and Salt*. Tech. Rep., IFS Working Paper (IFS, 2021); http://hdl.handle.net/10419/242920
58. Mozaffarian, D., Blanck, H. M., Garfield, K. M., Wassung, A. & Petersen, R. A Food is Medicine approach to achieve nutrition security and improve health. *Nat. Med.* **28**, 2238–2240 (2022).
59. *The National Food Strategy: The Plan* (accessed 23 March 2022) (2019); https://www.nationalfoodstrategy.org/
60. True Cost of Food: Measuring What Matters to Transform the U.S. Food System (The Rockefeller Foundation, 2021); https://www.rockefellerfoundation.org/report/true-cost-of-food-measuring-what-matters-to-transform-the- u-s-food-system/
61. Nasirian, F. & Menichetti, G. Molecular interaction networks and cardiovascular disease risk: the role of food bioactive small molecules. *Arterioscler. Thromb. Vasc. Biol.* **43**, 813–823 (2023).
62. Adams, J. Rebalancing the marketing of healthier versus less healthy food products. *PLoS Med.* **19**, e1003956 (2022).
63. Shaw, S. C., Ntani, G., Baird, J. & Vogel, C. A. A systematic review of the influences of food store product placement on dietary-related outcomes. *Nutr. Rev.* **78**, 1030–1045 (2020).
64. Shepherd, R. Resistance to changes in diet. *Proc. Nutr. Soc.* **61**, 267–272 (2002).
65. Kelly, M. P. & Barker, M. Why is changing health-related behaviour so difficult? *Public Health* **136**, 109–116 (2016).
66. Barabási, A. L., Menichetti, G. & Loscalzo, J. The unmapped chemical complexity of our diet. *Nat. Food* **1**, 33–37 (2020).
67. Menichetti, G., Barabasi, A.-L. & Loscalzo, J. Decoding the Foodome: molecular networks connecting diet and health. *Annu. Rev. Nutr.* **44**, 257–288 (2024).
68. 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. *J. Food Compos. Anal.* **99**, 103848 (2021).
69. Menichetti, G. & Barabási, A.-L. Nutrient concentrations in food display universal behaviour. *Nat. Food* **3**, 375–382 (2022).
70. Hooton, F., Menichetti, G. & Barabási, A. L. Exploring food contents in scientific literature with FoodMine. *Sci. Rep.* **10**, 16191 (2020).
71. Chatterjee, A. et al. Improving the generalizability of protein-ligand binding predictions with AI-Bind. *Nat. Commun.* **14**, 1989 (2023).
72. Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with Python. In *Proc. 9th Python in Science Conference* (eds van der Walt, S. & Millman, J.) 57–61 (2010).
73. Huber, P. J. Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Stat.* **1**, 799–821 (1973).
74. Croux, C. & Rousseeuw, P. J. Time-efficient algorithms for two highly robust estimators of scale. In *Computational Statistics*, 411–428 (Springer, 1992).
75. Brown, G. G. & Rutemiller, H. C. Means and variances of stochastic vector products with applications to random linear models. *Manag. Sci.* **24**, 210–216 (1977).
76. Beel, J., Gipp, B., Langer, S. & Breitinger, C. Research-paper recommender systems: a literature survey. *Int. J. Digit. Libr.* **17**, 305–338 (2016).
77. *Substances Added to Food* (FDA, accessed 1 November 2021); https://www.hfpappexternal.fda.gov/scripts/fdcc/index.cfm?set=FoodSubstances
78. *Substances Added to Food* (FDA, accessed 1 November 2021) (2003); https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?CFRPart=172

Acknowledgements

We thank D. Shanbhag at Northeastern University for his help on data collection and cleaning. We thank D. Koshkina for help in designing the figures. A.-L.B. is partially supported by National Institutes of Health grant 1P01HL132825, American Heart Association grant 151708 and European Research Council grant 810115-DYNASET. G.M. is supported by National Institutes of Health/National Heart, Lung, and Blood Institute K25HL173665 and American Heart Association 24MERIT1185447.

Author contributions

G.M., B.R. and A.-L.B. conceived and designed the research. B.R. performed data collection, data modelling, statistical analysis, and data querying and integration and contributed to the writing of the manuscript. G.I. and M.S. performed data cleaning, data curation, code cleaning and optimization, and fact checking and contributed to the writing of the manuscript. P.M. performed data cleaning and data integration and contributed to the writing of the manuscript. G.M. and A.-L.B. wrote the manuscript and contributed to the conceptual and statistical design of the study.

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, which focuses on urban data. All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43016-024-01095-7.

Correspondence and requests for materials should be addressed to Giulia Menichetti.

Peer review information *Nature Food* thanks Luca Pappalardo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Giulia Menichetti

Last updated by author(s): Nov 5, 2024

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

We used python to scrape data (food items, their nutrition facts, ingredient list, etc) from the website of grocery stores. If needed we can share the scraper python code that we designed to scrape data from Walmart, Target, and WholeFoods online stores.

Data analysis

We used python==3.11.7 with the packages: jupyter notebook==6.5.4, pymongo==4.8.0, pandas==2.1.4, numpy==1.26.4, seaborn==0.12.2, statsmodels==0.14.0, scipy==1.11.4, matlabplot==3.8.0, plotly==5.9.0, and certifi==2024.6.2 to analyze the data. All created codes are accessible through two Jupyter Notebooks and two py files as well as select datasets in our GitHub repository (https://github.com/Barabasi-Lab/GroceryDB/tree/main/analysis). The large data we scraped is stored in MongoDB which is accessible via the codes in the GitHub and through our website (TrueFood.tech).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Since the data is large, we have information for over 50,000 foods, we used MongoDB to store the data. The key to access MongoDB (read only) is available in our GitHub repository. Also, we provide two notebooks to enhance data availability. The notebooks retrieve all data from MongoDB and recreate all the figures in the manuscript and SI. Please see the following folder in our public GitHub repository https://github.com/Barabasi-Lab/GroceryDB/tree/main/analysis

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences ☒ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Analyzing the prevalence of processed foods in the US grocery stores using quantitative methods
Research sample	We scraped over 50,000 foods from the websites of Walmart, Target, and WholeFoods to create GroceryDB and analyze the extent of food processing in the US food supply. We chose these stores since most of the US use them as their source of groceries. By collecting all the foods from these stores, GroceryDB is a representative sample of many Americans food options. Each store has its own data structure to categorize and price items as well as offering different food items available for purchase.
Sampling strategy	We scraped all foods from Walmart, Target, and WholeFoods store online websites. There was no sampling procedure used since we collected every food item available for purchase. There was no sample size calculations performed. The rationale was to collect a holistic dataset of US consumption of foods, by collecting all foods from the stores, we believe the data size is sufficient.
Data collection	We scraped foods from Walmart, Target, and WholeFoods store online websites by using Python to navigate their storefronts and automatically collect the name, ingredient list, price, and nutrition information of all food items. The data collection was blind to the study hypothesis and downstream analysis.
Timing	We started creating the codes to scrape the online stores in September 2020, testing and debugging the codes until April 2021. We started collecting the data in May 2021 and completed the collection by the end of May 2021.
Data exclusions	We analyzed all foods that had minimum of 10 nutrition facts reported by the grocery stores. Also, we analyzed the ingredient list of all foods
Non-participation	No participants were involved in the study.
Randomization	Our study is a holistic approach to assessing the processed foods within grocery stores Americans commonly use to purchase their groceries. Therefore, we are comparing the different food categories using the full list of food items within the categories as well as using the full list of food items that contain a specific ingredient for the calculation of IgFPro.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Human research participants
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging