GroceryDB: Prevalence of Processed Food in Grocery Stores

Babak Ravandi[1], Peter Mehler[2], Albert-László Barabási[1,3,4], Giulia Menichetti[1,3,*]

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

[2]Department of Computer Science, IT University of Copenhagen, Copenhagen, Denmark

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

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

The offering of grocery stores is a strong driver of consumer decisions, shaping their diet and long-term health. While highly processed food like packaged products, processed meat, and sweetened soft drinks have been increasingly associated with unhealthy diet, 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. Here we introduce GroceryDB, a database with over 50,000 food items sold by Walmart, Target, and Wholefoods, unveiling how big data can be harnessed to empower consumers and policymakers with systematic access to the degree of processing of the foods they select, and the potential alternatives in the surrounding food environment. The wealth of data collected on ingredient lists and nutrition facts allows a large scale analysis of ingredient patterns and degree of processing stratified by store, food category, and price range. We find that the nutritional choices of the consumers, translated as the degree of food processing, strongly depend on the food categories and grocery stores. Moreover, the data allows us to quantify the individual contribution of over 1,000 ingredients to ultra-processing. GroceryDB and the associated http://TrueFood.Tech/ website make this information accessible, guiding consumers toward less processed food choices while assisting policymakers in reforming the food supply.

*Corresponding author. e-mail: giulia.menichetti@channing.harvard.edu

Introduction

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 our over-reliance on ultra-processed food (UPF) has fostered unhealthy diet [5,6]. 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 ultra-processed food (UPF), contributing up to 60% of consumed calories in developed nations [5–7]. For instance, recent studies have linked the consumption of ultra-processed food to non-communicable diseases like metabolic syndrome [8–14], and exposure to industrialized preservatives and pesticides [15–20]. This body of work contributed to a paradigm shift from food security, which focuses on access to affordable food, to nutrition security, emphasising on the need for wholesome foods [21, 22].

Much of UPF reaches consumers through grocery stores, as documented by the National Health and Nutrition Examination Survey (NHANES), indicating that in the US over 60% of the food consumed comes from grocery stores (Figure S1). The high reliance on ultra-processed items and their potential adverse impact on health raises three important questions: 1) How do we determine the degree of processing characterizing a particular item on the shelf? 2) How can we quantify the extent of food processing in the food supply? 3) What plausible alternatives can we identify to reduce UPF consumption?

Measuring the degree of food processing is a key step in addressing these questions, but it is not straightforward. Furthermore, consumers often struggle to decipher the information on food labels linked to food processing, despite the increasing evidence associating ultra-processed foods with adverse health outcomes. Beyond the translational challenges for the consumer, food labels often display mixed messages, partly driven by reductionist metrics focusing on one nutrient at a time [23], and partly because there are contrasting criteria on how to classify processed foods [24]. The ambiguity and inconsistency of current food processing classification systems have led to conflicting results on their role as risk factors for non-communicable chronic diseases [21, 25]. Some of these classification systems also suffer from poor inter-reliability and lack of reproducibility, is-

sues rooted in purely descriptive expertise-based approaches, leaving room for ambiguity and differences in interpretation [21, 24, 26]. Hence, there is a growing call among scientists for a more objective definition of the degree of food processing, based on underlying biological mechanisms rather than subjective opinions of different research groups [21]. Of the three proposed areas for aligning food processing definitions, the nutritional profile of food is currently the only aspect consistently regulated and reported worldwide [21,24]. Additional sources of relevant data are partially captured in the ingredient list which includes elements, possibly additives, contributing to the sensory aspects of food such as flavor, taste, and texture. Unfortunately, the lack of structured data describing processing methods and an internationally standardized ontology for the ingredients printed on food labels are major sources of ambiguity, as documented by the GS1 UK data crunch analysis on the impact of bad data on profits and consumer service in the UK grocery industry, reporting an average of 80% inconsistency in products data [27].

The research efforts outlined in [21] align with a growing demand for high-quality and internationally comparable statistics to promote objective metrics, reproducibility, and data-driven decision-making, advancing our convergence towards the Sustainable Development Goals (SDGs), set up in 2015 by the United Nations General Assembly and intended to be achieved by the year 2030. [28, 29]. Artificial intelligence (AI) methodologies, in particular, are now welcomed as objective data-driven tools to advance populations' nutrition security, a concept underpinning and connecting SDGs 'zero-hunger', 'good health and well-being', 'industry, innovation, and infrastructure', and 'reduce inequalities', which are echoed by the USDA [30] and by the recent White House conference on Hunger, Nutrition and Health [31].

Responding to the need for objective and scalable metrics to ensure nutrition security, we have recently harnessed machine learning to create and fully automate our Food Processing Score (FPro) [32]. FPro is a continuous index derived by mapping the features of processing techniques learned from manual labels to the concentrations of nutrients in a non-linear fashion. To teach our algorithm how to score processing from nutrients, we leveraged the labels provided by NOVA, currently the most widely used system to classify foods according to processing-related criteria, offering us an extensive

array of epidemiological literature for comparative analysis [33]. Indeed, NOVA was used in 95% of the studies exploring the relationship between the consumption of highly processed foods and health outcomes published between 2015 and 2019 [34]. However, the FPro algorithm can accommodate different food processing classification systems such as EPIC [35], UNC [36], or SIGA [37]. We extensively tested and validated the stability of FPro in several databases such as the US Food and Nutrient Database for Dietary Studies (FNDDS) and the international Open Food Facts. Furthermore, we rigorously tested the predictive power of FPro for epidemiological outcomes with an Environment-Wide Association Study (EWAS), leveraging multiple cycles of USDA's model food databases and national food consumption surveys [32]. This body of work allowed us to implement an in-silico study based on US cross-sectional population data, where we showed that on average substituting only a single ultra-processed food item in a person's diet with a minimally-processed alternative from the same food category can significantly 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) [32].

Here, building on the versatility and scalability of the FPro algorithm, we extend our analysis beyond "model foods" tailored for epidemiological databases and instead analyze real-world data encompassing over 50,000 products obtained from major US grocery store websites. This data and the resulting analysis contribute to GroceryDB, an open-source database of foods and beverages collected from the publicly available online markets of Walmart, Target, and WholeFoods. Our objective is to demonstrate how machine learning can effectively analyze 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. GroceryDB, shared publicly at http://TrueFood.Tech/, provides the data and methods to quantify food processing and map the organization of ingredients in the US food supply, ultimately, setting the stage for similar initiatives to drive better-informed dietary decisions worldwide.

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, as reported by the nutrition facts, into its degree of processing [32]. In Figure 1, we illustrate the use of FPro by offering the processing score of three products in the breads and yogurt categories, allowing us to compare their degree of processing. Indeed, the Manna Organics multi-grain bread is made from whole wheat kernels, barley, and rice without additives, added salt, oil, and even yeast, resulting in a low processing score of $FPro = 0.314$. However, the Aunt Millie's and Pepperidge Farmhouse breads include 'resistant corn starch', 'soluble corn fiber', and 'oat fiber', requiring additional processing to extract starch and fiber from corn and oat to be used as an independent ingredient (Figure 1a), resulting in much higher processing score of $FPro = 0.732$ and $FPro = 0.997$. Similarly, the Seven Stars Farm yogurt ($FPro = 0.355$) is a whole milk yogurt made from 'grade A pasteurized organic milk', yet the Siggi's yogurt ($FPro = 0.436$) uses 'Pasteurized Skim Milk' that requires more processing to obtain 0% fat. Finally, the Chobani Cookies & Cream yogurt relies on cane sugar as the second most dominant ingredient, and on a cocktails of additives like 'caramel color', 'fruit pectin', and 'vanilla bean powder' making it a highly processed yogurt, resulting in a high processing score $FPro = 0.918$.

We assigned an FPro score to each food in GroceryDB by leveraging our machine learning classifier FoodProX, which takes as input the mandatory information captured by the nutrition facts (see Methods). We find that the distribution of the FPro scores in the three stores is rather similar: in each store we observe a monotonically increasing curve (Figure 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). We do observe, however, systematic differences between the stores: WholeFoods offers more minimally-processed and fewer ultra-processed items, in contrast with a particularly high fraction of ultra-processed offerings by Target (high FPro).

FPro also captures the inherent variability in the degree of processing per food category. As illustrated in Figure 2b, we find a small variability of FPro scores in categories like jerky, popcorn, chips, bread, biscuits, and mac & cheese, indicating that consumers have limited choices in terms of degree of processing in these categories (see Section S5

for harmonizing categories between stores). Yet, in categories like cereals, milk & milk-substitute, pasta-noodles, and snack bars, FPro varies widely, reflecting a wider extent of possible choices from a food processing perspective.

We compared the distribution of FPro in GroceryDB with the latest USDA Food and Nutrient Database for Dietary Studies (FNDDS), offering a representative sample of the consumed food supply (Figure 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. Additionally, we compared GroceryDB with the USDA Global Branded Food Products Database (BFPD), which contains 1,142,610 branded products, finding that the distributions of FPro in GroceryDB and BFPD follows similar trends (Figure 2c). While BFPD contains 22 times more foods than GroceryDB, surprisingly only 44% of the products in GroceryDB are present in BFPD (Section S4). This indicates that while BFPD offers an extensive representation of branded products, it does not map into the current offering of stores. Furthermore, we compared GroceryDB with Open Food Facts (OFF) [38], another extensive collection of branded products collected through crowd sourcing, containing 426,000 products with English ingredient lists. We find that less than 30% of the products in GroceryDB are present in OFF (Figure S4), a small overlap, suggesting that monitoring the 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 allow us to unveil 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 Figure 3A. However, the relationship between FPro and price per calorie strongly depends on the food category (Section S6). For example, in soups & stews the price per calorie drops by 24.3% for 10% increase in FPro (Figure 3b), a trend observed also in cakes, mac & cheese, and ice cream (Figure S8). This means that on average, the most processed soups & stews, with $FPro \approx 1$, are 66.87% cheaper per calories than the minimally-processed alternatives with $FPro \approx 0.4$ (Figure 3e). In contrast, in cereals price per calorie drops only by 1.2% for 10% increase

in FPro (Figure 3c), a slow decrease observed also for seafood and yogurt products (Figure S8). Interestingly, we find an increasing trend between FPro and price in the milk & milk-substitute category (Figure 3d), partially explained by the higher price of plant-based milk substitutes, that require more extensive processing than the dairy based milks.

Taken together, the continuous nature of FPro facilitated the analysis of the relationship between price and food processing stratified by food category. Overall, we find that the higher degrees of food processing are associated with cheaper calories, although this relationship displays remarkable variation across different food categories. Further in-depth analyses are needed to evaluate the effectiveness of intervention strategies targeting specific food groups within diverse food environments.

Choice Availability and Food Processing

Not surprisingly, GroceryDB documents differences in the offering of the three stores we analyzed: while WholeFoods offers a selection of cereals with a wide range of processing levels, from minimally-processed to ultra-processed, in Walmart the available cereals are limited to products with higher FPro values (Figure 4a). To understand the roots of these differences, we investigated the ingredients of cereals offered by each grocery store, one of the most popular staple crops, consumed by 283 million Americans in 2020 [39]. We find that cereals offered by WholeFoods rely on less sugar, less natural flavors, and added vitamins (Figure 4b). In contrast, cereals in Target and Walmart tend to contain corn syrup, a sweetener associated with enhanced absorption of dietary fat and weight gain [40]. Corn syrup is largely absent in the WholeFoods cereals, partially explaining the wider range of processing scores characterising cereals offered by the store (Figure 4a).

The brands offered by each store could also help explain the different patterns. We found that while Walmart and Target have a large overlap in the list of brands they carry, WholeFoods relies on different suppliers (Figure 4c), largely unavailable in other grocery stores. In general, WholeFoods offers less processed soups & stews, yogurt & yogurt drinks, and milk & milk-substitute (Figure 4a). In these categories Walmart's and Target's offerings are limited to higher FPro values. Lastly, some food categories like pizza, mac & cheese, and popcorn are highly processed in all stores (Figure 4a). Indeed,

pizzas offered in all three chains are limited to high FPro values, partially explained by the reliance on substitute ingredients like "imitation mozzarella cheese," instead of "mozzarella cheese".

While grocery stores offer a large variety of products, the offered processing choices can be identical in multiple stores. For example, GroceryDB has a comparable number of cookies & biscuits in each chain, with 373, 451, and 402 items in Walmart, Target, and WholeFoods, respectively. The degree of processing of cookies & biscuits in Walmart and Target are nearly identical ($0.88 < FPro < 1$), limiting consumer nutritional choices in a narrow range of processing (Figure 4a). In contrast, WholeFoods not only offers a large number of items (402 cookies & biscuits), but it also offers a wider choices of processing ($0.57 < FPro < 1$)

Organization of Ingredients in the Food Supply

Food and beverage companies are required to report the list of ingredients in the descending order of the amount used in the final product. When an ingredient itself is a composite, consisting of two or more ingredients, FDA mandates parentheses to declare the corresponding sub-ingredients (Figure 5a-b) [41]. We organized the ingredient list as a tree (see Methods), allowing us to compare a highly processed cheesecake with a less processed alternative (Figure 5). In general, we find that products with complex ingredient trees are more processed than products with simpler and fewer ingredients (Section S7.3). For example, the ultra-processed cheesecake in Figure 5a has 42 ingredients, 26 additives, and 4 branches with sub-ingredients. In contrast, the minimally-processed cheesecake has only 14 ingredients, 5 additives, and 1 branch with sub-ingredients (Figure 5b). As illustrated by the cheesecakes example, ingredients used in the food supply are not equally processed, prompting us to ask: 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

$$IgFPro(g) = \frac{\sum_{f \in F_g} r_g^f * FPro^f}{\sum_{f \in F_g} r_g^f}, \tag{1}$$

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 (Section S7.5). IgFPro ranges between 0 (unprocessed)

and 1 (ultra-processed), allowing us to rank-order ingredients based on their contribution to the degree of processing of the final product. We find that not all additives contribute equally to ultra-processing. For example, the ultra-processed cheesecake (Figure 5a) has sodium tripolyphosphate (a stabilizer used to improve the whipping properties with $IgFPro = 0.926$), polysorbate 60 (an emulsifier used in cakes for increased volume and fine grain with $IgFPro = 0.922$), and corn syrup (a corn sweetener with $IgFPro = 0.909$) [42], each of which emerging as signals of ultra-processing with high IgFPro scores. In contrast, both the minimally-processed and ultra-processed cheesecakes (Figure 5) contain xanthan gum ($IgFPro = 0.817$), guar gum ($IgFPro = 0.806$), locust bean gum ($IgFPro = 0.780$), and salt ($IgFPro = 0.771$). Indeed, the European Food Safety Authority (EFSA) reported that xanthan gum as a food additive does not pose any safety concern for the general population, and FDA classified guar gum and locust bean gum as generally recognized safe [42].

By the same token, we looked into the oils used as ingredients in branded products to assess which oils contribute the most to ultra-processed foods. IgFPro identifies brain octane oil ($IgFPro = 0.573$), flax seed oil ($IgFPro = 0.686$), and olive oil ($IgFPro = 0.712$) as the highest quality oils, having the smallest contribution to ultra-processing. In contrast, palm oil ($IgFPro = 0.890$), vegetable oil ($IgFPro = 0.8676$), and soy bean oil ($IgFPro = 0.8684$) represent strong signals of ultra-processing (Figure 6a). Indeed, flax seed oil is high in omega-3 fatty acids with several health benefits [43]. In contrast, the blending of vegetable oils, a signature of ultra-processed food, is one of the simplest methods to create products with desired texture, stability, and nutritional properties [44].

Finally, to illustrate the ingredient patterns characterising ultra-processed foods in Figure 6b, we show three tortilla chips, ranked from the "minimally-processed" to the 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, grounded corn, and has calcium hydroxide, generally recognized as a safe additive made by adding water to calcium oxide (lime) to promote dispersion of ingredients [42]. In contrast, the ultra-processed Doritos ($FPro = 0.982$) have corn flours, 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 (Figure 6b).

In summary, complex ingredient patterns accompany the production of ultra-processed foods (Section S7.4). IgFPro captures the role of individual ingredients in the food supply, enabling us to diagnose the processing characteristics of the whole food supply as well as the contribution of individual ingredients.

Discussion

By combining large-scale data on food composition and machine learning, GroceryDB uncovers novel insights on the current state of food processing in the US grocery landscape, enabling us to obtain distributions of food processing scores that capture a remarkable variability in the offerings of multiple grocery stores. The differences in FPro's distributions (Figure 2A) indicate that multiple factors drive the range of choices available in grocery stores, from the cost of food and the socio-economic status of the consumers to the distinct declared missions of the supermarket chains: "quality is a state of mind" for WholeFoods Market and "helping people save money so they can live better" for Walmart [45, 46]. Furthermore, the continuous nature of FPro enabled us to conduct a data-driven investigation on the relationship between price and food processing stratified by food category. We find that overall in GroceryDB food processing 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. [47–50]. 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 & milk-substitutes compared to soups & stews (Section S6).

Governments increasingly acknowledge the impact of processed foods on population health, and its long-term effect on healthcare. For example, the United Kingdom spends 18 billion £ annually on direct medical costs related to non-communicable diseases like obesity [51]. Similarly, according to a 2021 report on the true cost of food published by the Rockefeller Foundation, the US spend 1.1 trillion dollars annually on food-related human health costs, with direct and indirect costs for cardiovascular disease and hypertension leading the statistics at $382.63 billion, followed by the obesity and overweight at $359.07 billion [52, 53]. To reduce obesity and cardiovascular diseases, the UK recently introduced

limitations on the promotion of foods high in fat, sugar and salt [54], common features of ultra-processed food. Along the same lines, the recent White House conference on Hunger, Nutrition, and Health launched in the US the "Food is Medicine" initiative, advocating for medically tailored groceries and produce prescriptions covered by health insurance, as well as population-level health programs aimed at reducing the consumption of highly processed foods [55].

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 & milk alternatives, pasta-noodles, and snack bars, FPro exhibits 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 significantly different degrees of processing (Figure 2B). Yet, the comprehension of nutrient and ingredient data disclosed on food packaging often poses a challenge to consumers. Indeed, the prevalence of ultra-processed foods remains concealed, primarily because of nutrition facts reported for 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. Additionally, our approach holds great potential for public health initiatives aimed at improving the overall quality of our food environment. These initiatives may include strategies such as reorganizing supermarket layouts, optimizing shelf placements, and thoughtfully designing counter displays [49, 56, 57]. Transforming health-related behaviors is a challenging task [58, 59], hence easily adoptable dietary modifications along with environmental nudges could make it easier for individuals to embrace healthier choices.

GroceryDB extends beyond nutritional data, encompassing over 12,000 ingredients annotated with 500+ descriptors (Section S7.2). With the introduction of IgFPro, we can rank these ingredients by their prevalence and contribution to ultra-processed products, aiding policymakers in prioritizing ingredients and food groups for targeted intervention. Furthermore, IgFPro could support a "wholefoods reformulation" strategy by identifying minimally processed candidate ingredients, overcoming reformulations where fat, sugar,

or salt are replaced with other processed-refined or reconstituted alternatives [24, 60].

The current standard food processing classification systems are categorical systems that divide foods into multiple descriptive categories, commonly ranging between 4 to 8 categories that are manually designed to capture variability from unprocessed to ultra-processed [33, 61]. Such categorical classification systems lead to considering many foods as equally ultra-processed, facing the conclusions that 60% of the global, 73% of the USA, and 80% of the South Africa food supply is ultra-processed [32, 62–64]. Recently, this has also led to 70.2% of all Greek branded food products being classified as UPF, with subcategories of foods included in the sustainable and traditional Mediterranean diet scoring 58.7% or 41.0%, respectively [65]. Furthermore, this homogeneity in food classification systems makes it difficult to address the impact of ultra-processing with substitution or reformulation strategies [21, 66–68]. To address this issue, the SIGA classification developed a holistic-reductionist approach, which intends to offer a practical tool for the food industry [37]. SIGA defines an exhaustive list of markers of ultra-processing, helping to reduce the ambiguity in the interpretation of descriptive food labels. By utilizing these markers, SIGA expands the number of food processing categories to nine, introducing subgroups based on the degree of processing of ingredients, added salt and sugar, and fat contents. Despite these commendable efforts, SIGA still misses extensive epidemiological assessments over dietary intake surveys, and as a categorical classification system leaves open questions regarding the appropriate number and types of processing categories and the amount of information necessary to reliably identify them. A continuous food processing score eliminates the need for identifying categories of food processing by capturing maximal variability in the degrees of food processing. Such variability enabled us to derive the distributions of food processing scores among multiple grocery stores (Figure 2) as well as observe that in soups & stews, a 10% increase in FPro results in 24.3% decrease in price-per-calorie. Furthermore, we were able to 1) test the stability of FPro with varying selections of nutrients, 2) validate its robustness against the expected variability and uncertainty in nutrient content, and 3) gain higher predictive power in epidemiological studies compared to categorical classifications, finding novel associations [32].

Limitations and Future Directions

The food matrix embodies a complex interplay of chemical components, influencing the release, transfer, accessibility, digestibility, and stability of numerous food compounds [69]. Ongoing research highlights how altered matrices in UPFs may affect nutrient availability, postprandial glycemic response, and satiety levels [70–74]. Recent studies have also uncovered the microbiome's role in moderating the adverse effects of non-nutritive sweeteners and emulsifiers on glycemic response and intestinal inflammation [75, 76]. The multifaceted impact of UPFs on human health led us to explore concepts like metabolic response and food matrix through the lens of network science and machine learning, which account for critical dependencies and transcend reductionist single-nutrient analyses [77]. Indeed, by performing large-scale analyses of nutrient concentrations in the food supply we have documented how their amounts in unprocessed foods are constrained by physiological ranges determined by biochemistry and how different nutrient alteration patterns indicate reproducible food processing fingerprints [78, 79]. These findings have inspired and supported the development of FPro, which does not assess individual nutrients in isolation but, rather, learns from the configurations of correlated nutrient changes within a fixed quantity of food (100 grams) [32]. Consequently, a single high or low nutrient value does not dictate a food's FPro but the final score depends on the likelihood of observing the overall pattern of nutrient concentrations in unprocessed foods versus ultra-processed foods. For instance, while fortified foods may mirror mineral and vitamin content in unprocessed foods, our algorithm identifies unique concentration signatures unlikely to be found in minimally processed foods, resulting in a higher FPro [32].

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 byproducts, aims to evolve FPro into an unsupervised system, independent of manual classifications [77, 80]. Unlike expertise-based systems, FPro functions as a quantitative algorithm, utilizing standardized inputs to generate reproducible continuous scores, facilitating sensitivity analysis and uncertainty

estimations [32]. These important features enhance analyses' reliability, transparency, and interpretability while reducing errors linked to the descriptive nature of manual classifications [21], which have displayed a low degree of consistency among nutrition specialists [67].

The chemical composition of branded products is partially captured by the nutrition facts table and partially reported in the ingredient list, which includes additives like artificial colors, flavors, and emulsifiers. However, comprehensive and internationally well-regulated data on food ingredients is currently limited [27], leading us to focus on the nutrition facts to enhance our algorithm's portability and reproducibility. The nutrition facts alone exhibit excellent performance in discriminating between NOVA classes, confirming how food processing consistently alters nutrient concentrations with reproducible patterns, effectively harnessed by machine learning [32]. 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 (e.g., 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 remains blind to these chemical-physical changes, yielding identical results for matching input vectors. Incorporating disambiguated ingredients in FPro, such as the ultra-processing markers characterized by SIGA [61], may offer a solution until larger composition tables for branded products become available.

In summary, our work is not a continuation of the traditional food classification systems but a step towards harnessing machine learning methodologies to model the chemical complexity of food. Despite the limited amount of information reported in the nutrition facts and regulated by the FDA, GroceryDB and FPro collectively present a data-driven approach that enabled the substitution algorithm implemented at http://TrueFood.Tech/. Leveraging the continuous nature of FPro, the algorithm recommends similar, but less processed choices for any food in GroceryDB. GroceryDB along with the TrueFood platform demonstrates the value of data transparency on the inventory of grocery stores, a factor that directly influences consumer decisions.

Methods

Data Collection

We compiled publicly accessible data on food products available at Walmart, Target, and Whole Foods through their respective online platforms. Each store organizes its food items hierarchically. Utilizing these categorizations, we systematically navigated through the stores' websites to identify specific food items. To ensure consistency, we standardized the food category hierarchy within GroceryDB by comparing and aligning the classification systems employed by each store. These stores sourced nutrition facts from physical food labels and provided digital versions for each food item. This data enabled us to standardize nutrient concentrations to a uniform measure of 100 grams and employ FoodProX to evaluate the degree of food processing for each item. Lastly, all data was collected in May 2021.

Calculation of Food Processing Scores (FPro)

Processing alters the nutrient profile of food, changes that are detectable and categorizable using machine learning [32, 78, 81]. Hence, we developed FoodProX [32], a ML-based classifier that can translate the combinatorial changes in the nutrient amounts induced by food processing into a food processing score (FPro). FoodProX takes as input 12 nutrients reported in the nutrition facts (Table S1), and returns FPro, a continuous score ranging between 0 (unprocessed foods like fruits and vegetable) and 1 (ultra-processed foods like instant soups and shelf-stable breads). We used the manual NOVA classification applied to the USDA Standard Reference (SR) 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. The calculation of FPro for all foods in GroceryDB represents a generalization task, where the model faces "never-before-seen" data [78,82]. More details on the training dataset, including class heterogeneity and imbalance, are available in Section S3.

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 Figure S18A; (b) with emphasis on the order of ingredients as a proxy for their amount in a final product, as illustrated in Figure S18B, where the distance from 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 Eq. 1, we used $r_g^f = 1/d_g^f$ to rank the amount of an ingredient g in food f, where d_g^f captures the distance from the root (see Figure S18B for an example). Finally, IgFPro shows a remarkable variability when compared to the average FPro of products containing the selected ingredient (Figure S19), suggesting the presence of distinctive patterns of associations between ingredients' FPro and their rank in the ingredient list.

Acknowledgments

We thank Dwijay Shanbhag at Northeastern University for his help on data collection and cleaning. 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 code and data are available at BarabasiLab GitHub repository via https://github.com/Barabasi-Lab/GroceryDB/. Furthermore, GroceryDB is available to the public and consumers at http://TrueFood.Tech/.

Author contributions

G.M., B.R., and A.-L.B. conceived and designed the research. B.R. performed data collection, data modeling, statistical analysis, data query and integration, and contributed

to the writing of the manuscript. P.M. performed data cleaning, data integration, statistical analysis, and contributed to writing the manuscript. G.M. and A.-L.B. wrote the manuscript and contributed to the conceptual and statistical design of the study.

References

[1] Seferidi, P. *et al.* The neglected environmental impacts of ultra-processed foods. *The Lancet Planetary Health* **4**, e437–e438 (2020).

[2] Fardet, A. & Rock, E. Ultra-processed foods and food system sustainability: What are the links? *Sustainability* **12** (2020). URL `https://www.mdpi.com/2071-1050/12/15/6280`.

[3] Macdiarmid, J. I. The food system and climate change: are plant-based diets becoming unhealthy and less environmentally sustainable? *Proceedings of the Nutrition Society* **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. *Nature Food 2022* **3**, 1–16 (2022). URL `https://www.nature.com/articles/s43016-022-00588-7`.

[5] Lustig, R. H. Processed Food—An Experiment That Failed. *JAMA Pediatrics* **171**, 212–214 (2017). URL `https://doi.org/10.1001/jamapediatrics.2016.4136`.

[6] 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`.

[7] 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`.

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

[9] 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).

[10] 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).

[11] 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).

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

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

[14] 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`.

[15] 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`.

[16] 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`.

[17] 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`.

[18] 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.

[19] 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.

[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. *Environment International* **131**, 105057 (2019). URL https://www.sciencedirect.com/science/article/pii/S0160412019317416.

[21] Gibney, M. J. & Forde, C. G. Nutrition research challenges for processed food and health. *Nature Food* **3**, 104–109 (2022). URL https://doi.org/10.1038/s43016-021-00457-9.

[22] Mozaffarian, D., Fleischhacker, S. & Andrés, J. R. Prioritizing Nutrition Security in the US. *JAMA - Journal of the American Medical Association* **325**, 1605–1606 (2021). URL https://jamanetwork.com/journals/jama/fullarticle/2778232.

[23] Mozaffarian, D., Rosenberg, I. & Uauy, R. History of modern nutrition science-implications for current research, dietary guidelines, and food policy. *BMJ (Online)* **361** (2018). URL https://www.bmj.com/content/361/bmj.k2392https://www.bmj.com/content/361/bmj.k2392.abstract.

[24] Sadler, C. R. *et al.* Processed food classification: Conceptualisation and challenges. *Trends in Food Science & Technology* **112**, 149–162 (2021). URL https://www.sciencedirect.com/science/article/pii/S0924224421001667.

[25] Lacy-Nichols, J. & Freudenberg, N. Opportunities and limitations of the ultra-processed food framing. *Nature Food* **3**, 975–977 (2022). URL https://doi.org/10.1038/s43016-022-00670-0.

[26] 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.

[27] Data crunch report: The impact of bad data on profits and customer service in the uk grocery industry. GS1 UK and Cranfield University School of Management. https://dspace.lib.cranfield.ac.uk/bitstream/handle/1826/4135/Data_crunch_report.pdf (2009). (accessed April 4, 2022).

[28] THE 17 GOALS — Sustainable Development. URL https://sdgs.un.org/goals.

[29] Methods and Standards — Food and Agriculture Organization of the United Nations. URL https://www.fao.org/statistics/methods-and-standards/en/.

[30] Food and Nutrition Security — USDA. URL https://www.usda.gov/nutrition-security.

[31] 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. *Nature Food* **3**, 561–563 (2022).

[32] Menichetti, G., Ravandi, B., Mozaffarian, D. & Barabási, A.-L. Machine learning prediction of the degree of food processing. *Nature Communications* **14**, 2312 (2023).

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

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

[35] 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.

[36] 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? *American Journal of Clinical Nutrition* **101**, 1251–1262 (2015). URL https://pubmed.ncbi.nlm.nih.gov/25948666/.

[37] 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 & function* **11**, 2026–2039 (2020).

[38] Open Food Facts. https://world.openfoodfacts.org/discover. (accessed March 1, 2022).

[39] U.S. population: Consumption of breakfast cereals (cold) from 2011 to 2024. https://www.statista.com/statistics/281995/us-households-consumption-of-breakfast-cereals-cold-trend/. 2021 (accessed February, 2022).

[40] 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. *The American Journal of Clinical Nutrition* **79**, 537–543 (2004). URL https://doi.org/10.1093/ajcn/79.4.537. https://academic.oup.com/ajcn/article-pdf/79/4/537/23713169/znu00404000537.pdf.

[41] Guidance for industry: Food labeling guide. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-industry-food-labeling-guide. 2021 (accessed Nov 1, 2021).

[42] Igoe, R. S. *Dictionary of food ingredients* (Springer Science & Business Media, 2011).

[43] Goyal, A., Sharma, V., Upadhyay, N., Gill, S. & Sihag, M. Flax and flaxseed oil: an ancient medicine & modern functional food. *Journal of Food Science and Technology* **51**, 1633–1653 (2014). URL https://doi.org/10.1007/s13197-013-1247-9.

[44] Hashempour-Baltork, F., Torbati, M., Azadmard-Damirchi, S. & Savage, G. P. Vegetable oil blending: A review of physicochemical, nutritional and health effects. *Trends in Food Science & Technology* **57**, 52–58 (2016). URL https://www.sciencedirect.com/science/article/pii/S0924224416302886.

[45] WholeFoods mission and values. https://www.wholefoodsmarket.com/mission-values. (accessed March 1, 2022).

[46] Walmart history. https://corporate.walmart.com/about/history. (accessed March 1, 2022).

[47] 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.

[48] Zenk, S. N., Tabak, L. A. & Pérez-Stable, E. J. Research Opportunities to Address Nutrition Insecurity and Disparities. *JAMA* **327**, 1953–1954 (2022). URL https://jamanetwork.com/journals/jama/fullarticle/2791951.

[49] 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 Medicine* **17**, e1003319 (2020). URL https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003319.

[50] 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. *International Journal of Food Sciences and Nutrition* **74**, 668–684 (2023). URL https://www.tandfonline.com/doi/abs/10.1080/09637486.2023.2241672.

[51] The national food strategy: The plan. https://www.nationalfoodstrategy.org/ (2021). (accessed March 23, 2022).

[52] True Cost of Food: Measuring What Matters to Transform the U.S. Food System - The Rockefeller Foundation. URL https://www.rockefellerfoundation.org/report/true-cost-of-food-measuring-what-matters-to-transform-the-u-s-food-system/.

[53] Nasirian, F. & Menichetti, G. Molecular Interaction Networks and Cardiovascular Disease Risk: The Role of Food Bioactive Small Molecules. *Arteriosclerosis, thrombosis, and vascular biology* **43**, 813–823 (2023).

[54] Griffith, R., Jenneson, V., James, J. & Taylor, A. The impact of a tax on added sugar and salt. Tech. Rep., IFS Working Paper (2021). URL http://hdl.handle.net/10419/242920.

[55] Mozaffarian, D., Blanck, H. M., Garfield, K. M., Wassung, A. & Petersen, R. A Food is Medicine approach to achieve nutrition security and improve health. *Nature Medicine* **28**, 2238–2240 (2022). URL https://www.nature.com/articles/s41591-022-02027-3.

[56] Adams, J. Rebalancing the marketing of healthier versus less healthy food products. *PLoS Medicine* **19**, e1003956 (2022). URL https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003956.

[57] 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. *Nutrition Reviews* **78**, 1030–1045 (2020). URL https://dx.doi.org/10.1093/nutrit/nuaa024.

[58] Shepherd, R. Resistance to changes in diet. *Proceedings of the Nutrition Society* **61**, 267–272 (2002).

[59] Kelly, M. P. & Barker, M. Why is changing health-related behaviour so difficult? *Public Health* **136**, 109–116 (2016). URL https://www.sciencedirect.com/science/article/pii/S0033350616300178.

[60] Scrinis, G. & Monteiro, C. A. Ultra-processed foods and the limits of product reformulation. *Public health nutrition* **21**, 247–252 (2018). URL https://pubmed.ncbi.nlm.nih.gov/28703086/.

[61] 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.

[62] 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.

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

[64] Frank, T. *et al.* A fit-for-purpose nutrient profiling model to underpin food and nutrition policies in south africa. *Nutrients* **13**, 2584 (2021).

[65] Katidi, A., Vlassopoulos, A., Noutsos, S. & Kapsokefalou, M. Ultra-Processed Foods in the Mediterranean Diet according to the NOVA Classification System; A Food Level Analysis of Branded Foods in Greece. *Foods* **12**, 1520 (2023). URL https://www.mdpi.com/2304-8158/12/7/1520/htmhttps://www.mdpi.com/2304-8158/12/7/1520.

[66] Gibney, M. J. Ultra-Processed Foods: Definitions and Policy Issues. *Current Developments in Nutrition* **3** (2018). URL https://doi.org/10.1093/cdn/nzy077. Nzy077, https://academic.oup.com/cdn/article-pdf/3/2/nzy077/27982102/nzy077.pdf.

[67] Braesco, V. *et al.* Ultra-processed foods: how functional is the nova system? *European Journal of Clinical Nutrition* **76**, 1245–1253 (2022). URL https://doi.org/10.1038/s41430-022-01099-1.

[68] Tobias, D. K. & Hall, K. D. Eliminate or reformulate ultra-processed foods? biological mechanisms matter. *Cell Metabolism* (2021). URL https://www.sciencedirect.com/science/article/pii/S1550413121004836.

[69] Aguilera, J. M. The food matrix: implications in processing, nutrition and health. *Critical Reviews in Food Science and Nutrition* **59**, 3612–3629 (2019). URL https://www.tandfonline.com/doi/abs/10.1080/10408398.2018.1502743.

[70] Berry, S. E. *et al.* Manipulation of lipid bioaccessibility of almond seeds influences postprandial lipemia in healthy human subjects. *American Journal of Clinical Nutrition* **88**, 922–929 (2008).

[71] In vitro and in vivo modeling of lipid bioaccessibility and digestion from almond muffins: The importance of the cell-wall barrier mechanism. *Journal of Functional Foods* **37**, 263–271 (2017).

[72] Mandalari, G. *et al.* Understanding the effect of particle size and processing on almond lipid bioaccessibility through microstructural analysis: From mastication to faecal collection. *Nutrients* **10** (2018).

URL /pmc/articles/PMC5852789//pmc/articles/PMC5852789/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5852789/.

[73] Novotny, J. A., Gebauer, S. K. & Baer, D. J. Discrepancy between the Atwater factor predicted and empirically measured energy values of almonds in human diets. *American Journal of Clinical Nutrition* **96**, 296–301 (2012).

[74] Wyatt, P. *et al.* Postprandial glycaemic dips predict appetite and energy intake in healthy individuals. *Nature Metabolism* 1–7 (2021). URL http://www.nature.com/articles/s42255-021-00383-x.

[75] Naimi, S., Viennois, E., Gewirtz, A. T. & Chassaing, B. Direct impact of commonly used dietary emulsifiers on human gut microbiota. *Microbiome* **9**, 1–19 (2021). URL https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-020-00996-6.

[76] Personalized microbiome-driven effects of non-nutritive sweeteners on human glucose tolerance. *Cell* **185**, 3307–3328.e19 (2022). URL http://www.cell.com/article/S0092867422009199/fulltexthttp://www.cell.com/article/S0092867422009199/abstracthttps://www.cell.com/cell/abstract/S0092-8674(22)00919-9.

[77] Menichetti, G., Barabasi, A.-L. & Loscalzo, J. Decoding the Foodome: Molecular Networks Connecting Diet and Health. *ResearchGate* (2023). URL https://www.researchgate.net/publication/375665637{_}Decoding{_}the{_}Foodome{_}Molecular{_}Networks{_}Connecting{_}Diet{_}an

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

[79] Sebek, M. L., Menichetti, G. & Barabási, A.-L. Estimating Nutrient Concentration in Food Using Untargeted Metabolomics. *bioRxiv* 2022.12.02.518912 (2022). URL https://www.biorxiv.org/content/10.1101/2022.12.02.518912v2.

[80] Barabási, A. L., Menichetti, G. & Loscalzo, J. The unmapped chemical complexity of our diet. *Nature Food* **1**, 33–37 (2020).

[81] Hooton, F., Menichetti, G. & Barabási, A. L. Exploring food contents in scientific literature with FoodMine. *Scientific Reports* **10**, 16191 (2020). URL `https://doi.org/10.1038/s41598-020-73105-0`.

[82] Chatterjee, A. *et al.* Improving the generalizability of protein-ligand binding predictions with AI-Bind. *Nature communications* **14**, 1989 (2023).

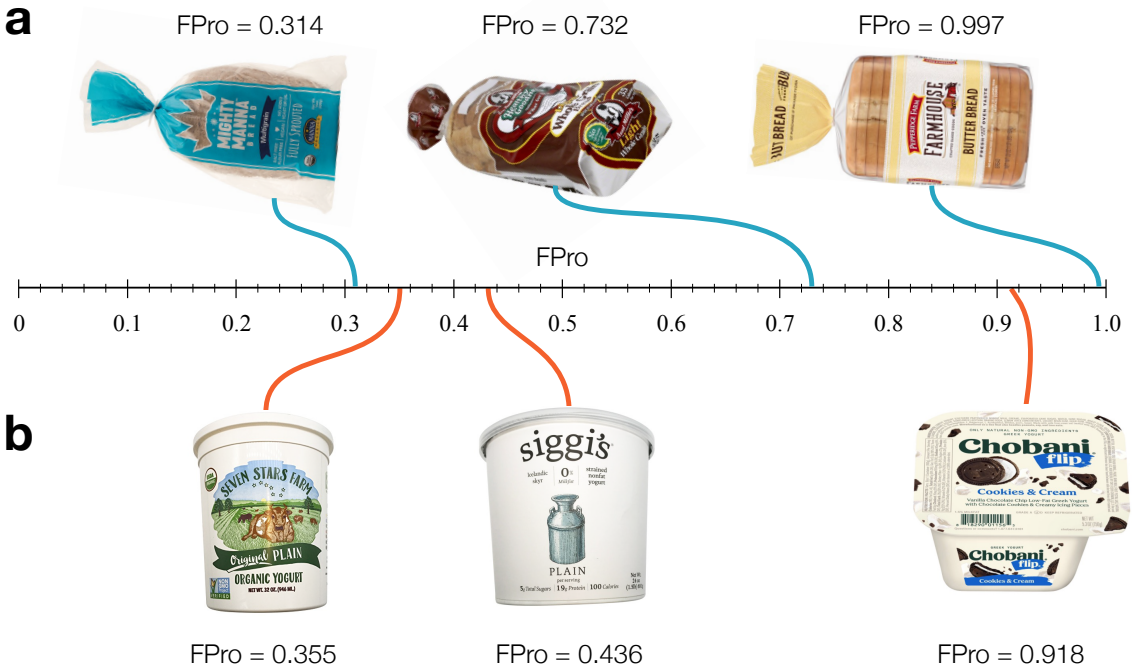


Figure 1

Figure 1: **Degrees of Food Processing in Three Categories.** FPro allows us to 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 WholeFoods, 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 fiber' and 'oat fiber' with additives like 'sugar', 'resistant corn starch', 'wheat gluten', and 'monocalcium phosphate'. **(b)** The Seven Stars Farm yogurt ($FPro = 0.355$) is made from the 'grade A pasteurized organic milk'. The Siggi's yogurt ($FPro = 0.436$) declares 'Pasteurized Skim Milk' as the main ingredients that has 0% fat milk, requiring more food processing to eliminate fat. Lastly, the Chobani Cookies & Cream yogurt ($FPro = 0.918$) has cane sugar as the second most dominant ingredient combined with multiple additives like 'caramel color', 'fruit pectin', and 'vanilla bean powder', making it a highly processed yogurt.

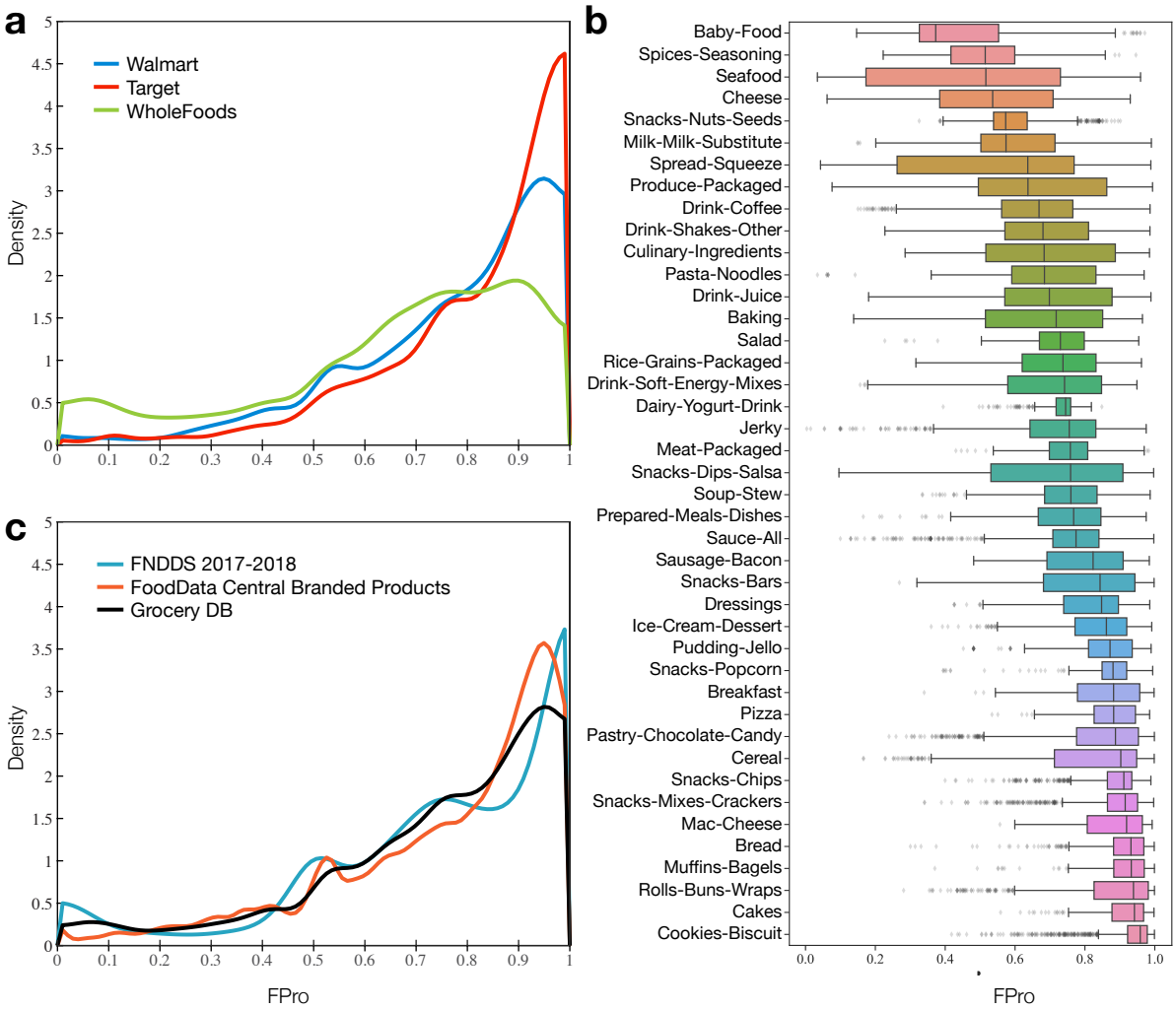
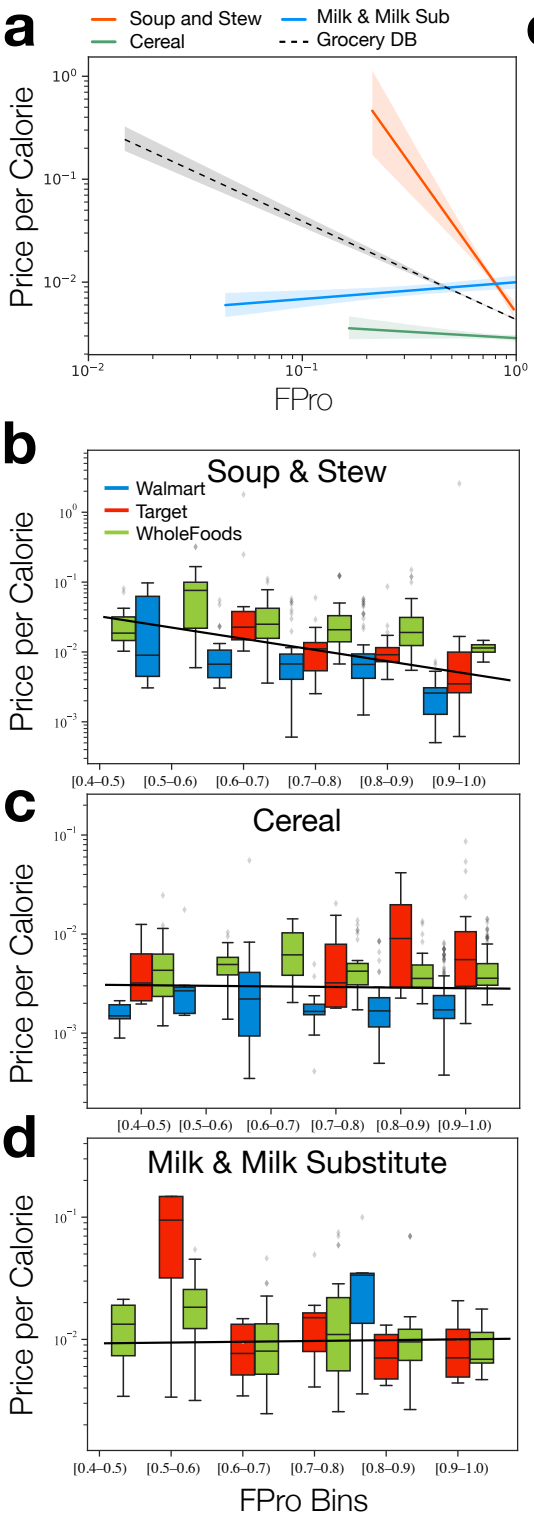


Figure 2: **Food Processing in Grocery Stores.** (**a**) The distribution of FPro scores from the three stores follows a similar trend, a monotonically increasing curve, indicating that the number of low FPro items (unprocessed and minimally-processed) offered by the grocery stores is relatively lower 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 like eggs, fresh produce, and raw meat are excluded (Section S5). (**c**) The distributions of FPro scores in GroceryDB compared to two USDA nationally representative food databases: the USDA Food and Nutrient Database for Dietary Studies (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 (Section S4).



e

%Price change
from minimally to ultra-processed

Pudding-Jello	-67.66
Drink-Shakes-Other	-70.67
Culinary-Ingredients	-66.94
Drink-Coffee	-69.17
Spices-Seasoning	-69.86
Soup-Stew	-66.87
Sauce-All	-65.12
Muffins-Bagels	-60.44
Breakfast	-64.31
Snacks-Bars	-57.32
Drink-Juice	-59.17
Pastry-Chocolate-Candy	-54.46
Mac-Cheese	-52.84
Ice-Cream-Dessert	-52.99
Snacks-Mixes-Crackers	-53.61
Grocery DB	-52.15
Cookies-Biscuit	-46.21
Bread	-43.34
Cakes	-42.5
Meat-Packaged	-42.89
Sausage-Bacon	-39.7
Produce-Packaged	-39.78
Pizza	-34.64
Rolls-Buns-Wraps	-37.75
Snacks-Chips	-35.11
Prepared-Meals-Dishes	-35.48
Dairy-Yogurt-Drink	-33.75
Snacks-Dips-Salsa	-25.25
Rice-Grains-Packaged	-23.56
Seafood	-26.39
Dressings	-21.45
Cereal	-17.87
Cheese	1.168
Snacks-Nuts-Seeds	3.497
Milk-Milk-Substitute	11.18

% Change in price per calorie

Figure 3

Figure 3: **Price and Food Processing.** **(a)** Using robust linear models, we assessed the relationship between price and food processing (see Figure S8 for regression coefficients of all categories). We find that price per calories drops by 24.3% and 1.2% for 10% increase in FPro in soup & stew and cereals, respectively. Also, we observe a 8.7% decrease across all foods in GroceryDB for 10% increase in FPro. Interestingly, in milk & milk-substitute, 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. **(b-d)** Distributions of price per calorie in the linear bins of FPro scores for each store (Figure S7 illustrates the correlation between price and FPro for all categories). In soup & stew, we find a steep decreasing slope between FPro and price per calorie, while in cereals we observe a smaller effect. In milk & and milk-substitute, price tends to slightly increase with higher values of FPro. **(e)** Percentage of change in price per calorie from the minimally-processed products to ultra-processed products in different food categories, obtained by comparing the average of top 10% minimally-processed items with the top 10% ultra-processed items. In the full GroceryDB, marked with the red star, on average the ultra-processed items are 52.15% cheaper than their minimally-processed alternatives.

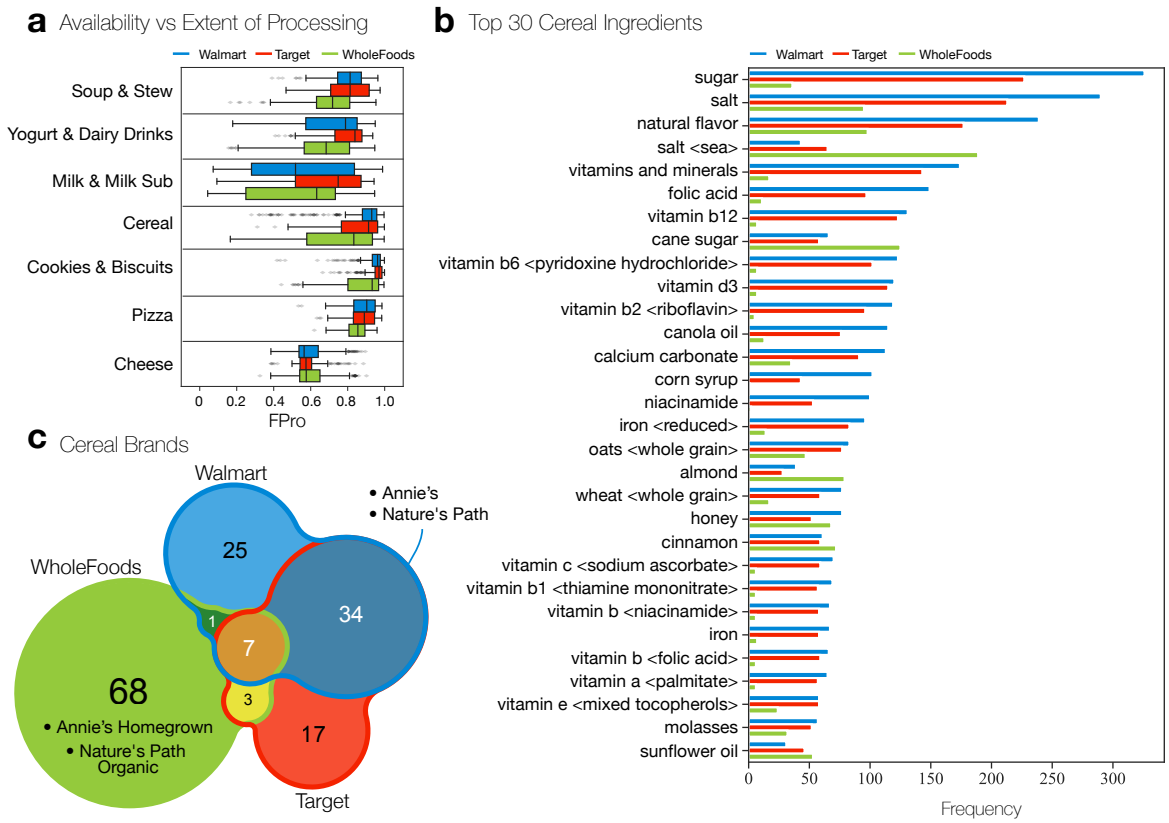


Figure 4

Figure 4: **The Difference between Stores in Term of Processing.** The nutritional choices offered to consumers, translated into FPro, varies depending 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, WholeFoods 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 characterised by high FPro values. Lastly, all cheese products are minimally-processed, showing consistency across different grocery stores. (**b**) The top 30 most reported ingredients in cereals shows that WholeFoods tends to eliminate corn syrup, uses more sunflower oil and less canola oil, and relies less on vitamin fortification. In total, GroceryDB has 1,245 cereals from which 400, 347, and 498 cereals are from Walmart, Target, and WholeFoods, respectively. (**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, WholeFoods tends to supply cereals from brands not available elsewhere.

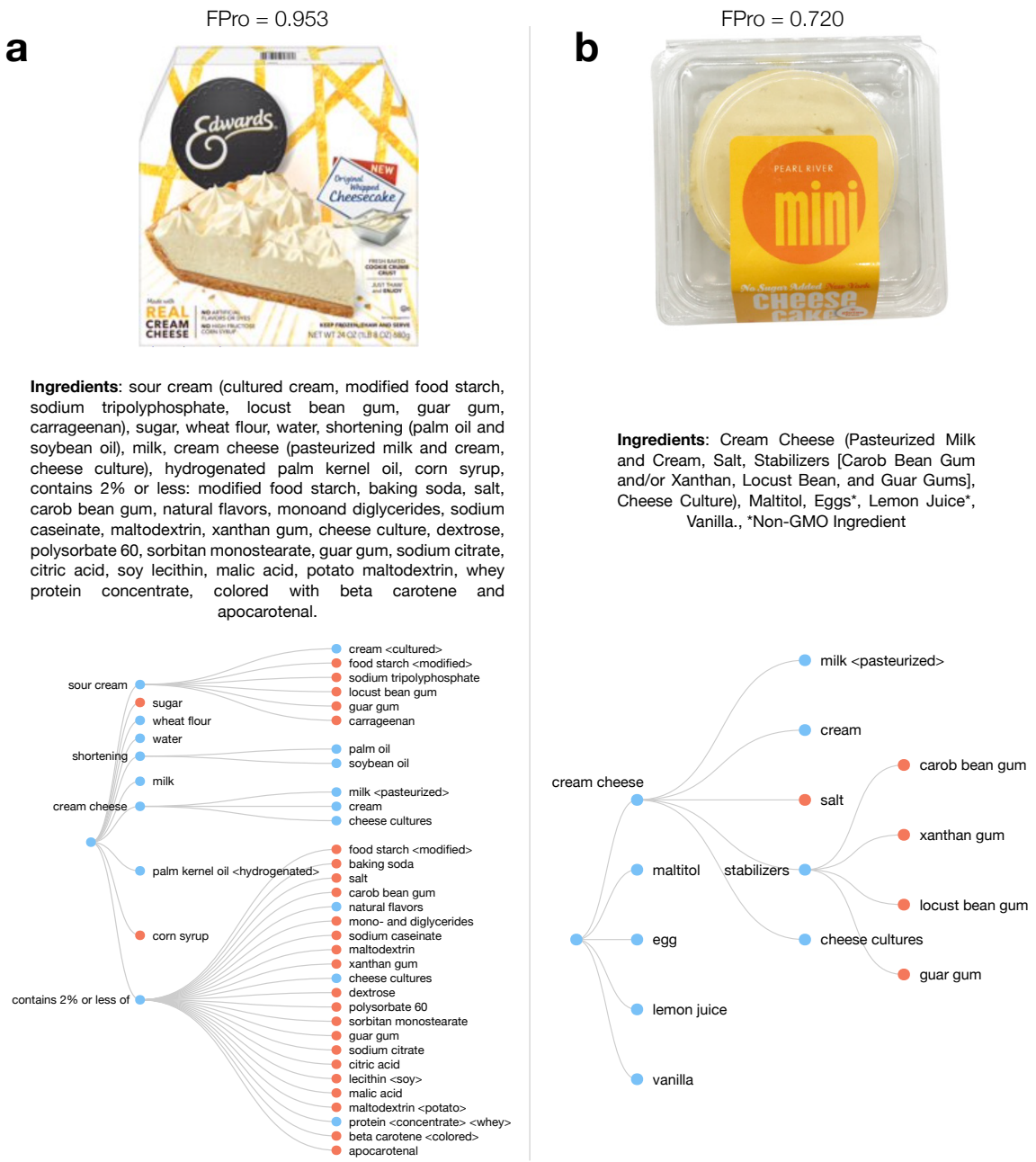


Figure 5

Figure 5: **Ingredient Trees.** GroceryDB organizes the ingredient list of products into structured trees, where the additives are marked as orange nodes (Methods and Section S7). **(a)** The highly processed cheesecake contains 42 ingredients from which 26 are additives, resulting in a complex ingredient tree with 4 branches of sub-ingredients. **(b)** The minimally-processed cheesecake has a simpler ingredient tree with 14 ingredients, 5 additives, and a single branch with sub-ingredients.

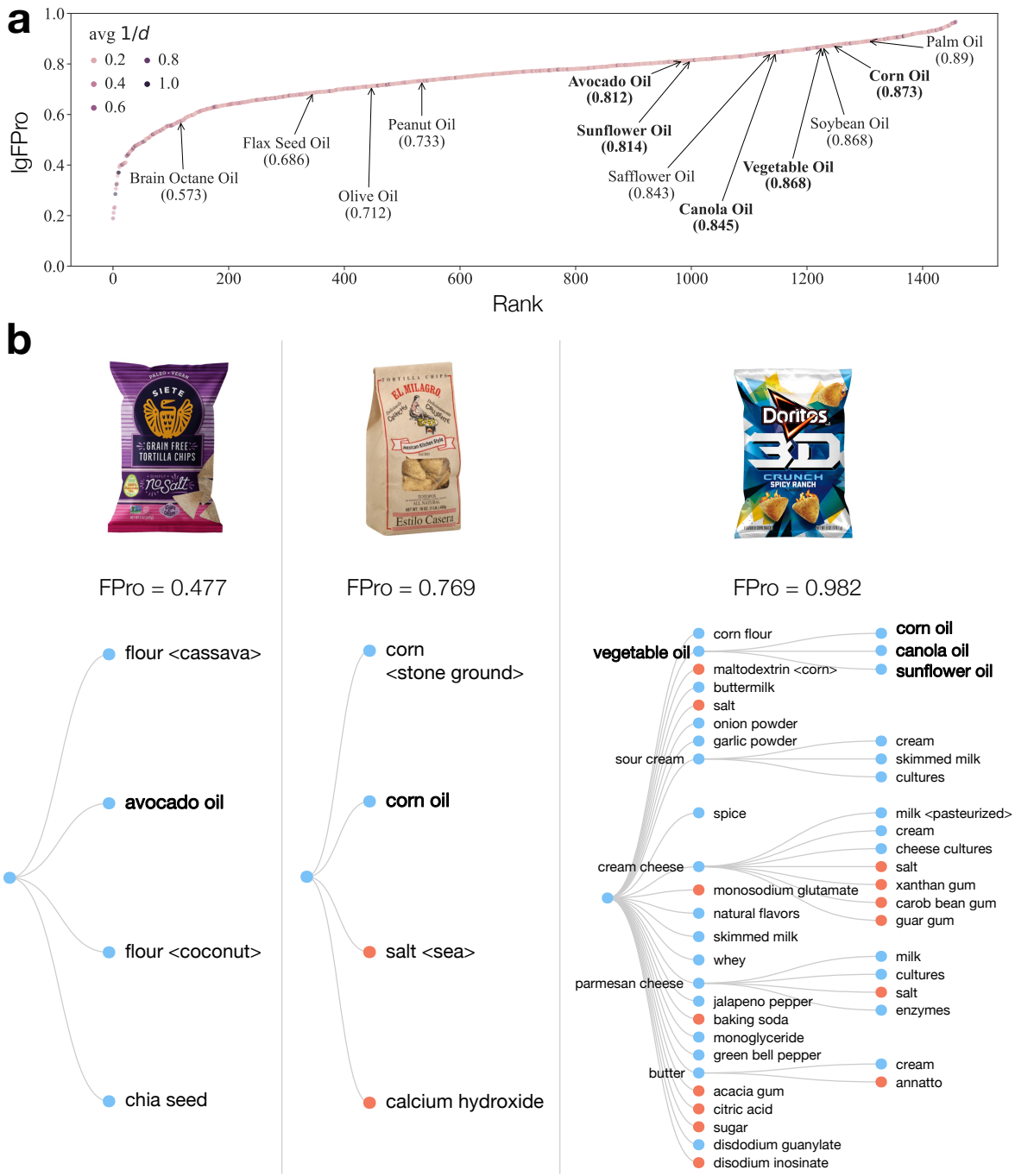

a
avg 1/d
0.2
0.4
0.6
0.8
1.0
lgFPro
Rank
Brain Octane Oil (0.573)
Flax Seed Oil (0.686)
Olive Oil (0.712)
Peanut Oil (0.733)
Avocado Oil (0.812)
Sunflower Oil (0.814)
Safflower Oil (0.843)
Canola Oil (0.845)
Vegetable Oil (0.868)
Soybean Oil (0.868)
Corn Oil (0.873)
Palm Oil (0.89)
b
FPro = 0.477
flour <cassava>
avocado oil
flour <coconut>
chia seed
FPro = 0.769
corn <stone ground>
corn oil
salt <sea>
calcium hydroxide
FPro = 0.982
vegetable oil
corn flour
maltodextrin <corn>
buttermilk
salt
onion powder
garlic powder
sour cream
cream
skimmed milk
cultures
cream cheese
spice
monosodium glutamate
natural flavors
skimmed milk
whey
milk <pasteurized>
cream
cheese cultures
salt
xanthan gum
carob bean gum
guar gum
parmesan cheese
jalapeno pepper
baking soda
monoglyceride
green bell pepper
butter
acacia gum
citric acid
sugar
disodium guanylate
disodium inosinate
corn oil
canola oil
sunflower oil
milk
cultures
salt
enzymes
cream
annatto

Figure 6

Figure 6: **Ingredient Processing Score (IgFPro).** To investigate which ingredients contribute most to ultra-processed products, we extend FPro to the ingredients listed on the nutrition fact labels using Eq. 1. **(a)** The IgFPro of all ingredients that appeared in at least 10 products are calculated, rank-ordering ingredients based on their contribution to ultra-processed foods. The popular oils used as an ingredient are highlighted, with the brain octane, flax seed, and olive oils contributing the least to ultra-processed products. In contrast, the palm, vegetable, and soybean oils contribute the most to ultra-processed products (Section S7.5). **(b)** The patterns of ingredients in the least-processed tortilla chips vs. the ultra-processed tortilla chips. The bold fonts track the IgFPro of the oils used in the three tortilla chips. The minimally-processed tortilla chips ($FPro = 0.477$) uses avocado oil ($IgFPro = 0.812$), and the more processed El Milagro tortilla ($FPro = 0.769$) has corn oil ($IgFPro = 0.866$). In contrast, the ultra-processed Doritos ($FPro = 0.982$) relies on a blend of vegetable oils ($IgFPro = 0.868$), and is accompanied with a much more complex ingredient tree, indicating that there is no single ingredient "bio-marker" for ultra-processed foods.

Supplementary Materials

GroceryDB: Prevalence of Processed Food in Grocery Stores

Babak Ravandi[1], Peter Mehler[2], Albert-László Barabási[1,3,4],Giulia Menichetti[1,3,*]

[1]Network Science Institute and Department of Physics, Northeastern University, Boston, USA
[2]Department of Computer Science, IT University of Copenhagen, Copenhagen, Denmark
[3]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, USA
[4]Department of Network and Data Science, Central European University, Budapest, Hungary

Contents

1 Food Sources in NHANES **2**

2 Data Collection and Processing **2**

2.1 Identification of Additives 3

3 Food Processing Score (FPro) **3**

4 Comparison with USDA FoodData Central and Open Food Facts Databases **6**

5 Category Harmonization **7**

6 Price and Food Processing **11**

7 Organization of Ingredients **12**

7.1 Cleaning Ingredient Lists 14

7.2 Approximating the Number of Ingredients in GroceryDB 18

7.3 Characteristics of Ingredient Trees 20

7.4 Correlation between Characteristics of Ingredient Lists and FPro 24

7.5 Ingredient Processing Score (IgFPro) 27

8 Case Study on Rice Cakes **29**

*Corresponding author. e-mail: giulia.menichetti@channing.harvard.edu

1 Food Sources in NHANES

The majority of people rely on grocery stores as the primary source of food. To investigate this hypothesis, we looked into the National Health and Nutrition Examination Survey (NHANES), offering the variable DR1FS that corresponds to "Where did you get (this/most of the ingredients for this)?", found at (`https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DR1IFF_J.htm#DR1FS`). We find that over 60% of all foods reported by NHANES 2017-2018 participants are from stores (Figure S1), indicating the high degree of reliance of the US population on grocery stores.

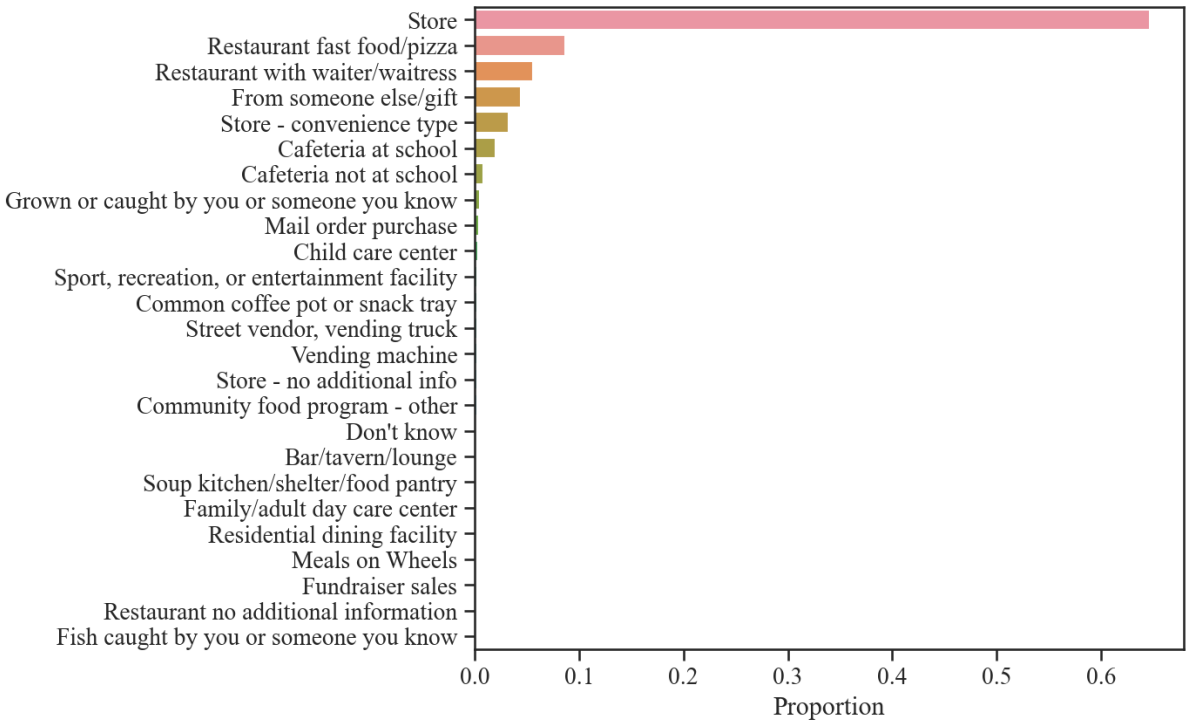


Figure S1: **Proportion of Food Sources Reported in NHANES 2017-2018.**

2 Data Collection and Processing

We built GroceryDB by collecting information regarding branded products from the publicly available data on the online websites of Walmart, Target, and WholeFoods. For data storage, we used MongoDB that offers highly flexible data structures with high input/output throughput.

2.1 Identification of Additives

We used the "Substances Added to Food" database provided by the US Food and Drug Administration (FDA) as our primary dictionary to identify additives and their synonyms in food products [1]. We also used the "Dictionary of Food Ingredients" (DFI) to further enrich and categorize the identification of additives [2]. Non-trivially, many substances declared on the nutrition fact labels have a broad range of synonyms. Thus, in addition to the synonyms provided by FDA, we also manually identified many synonyms of additives to better clean and normalize ingredient lists.

A major issue with cleaning ingredient lists is the high level of mismatch between labels provided by the FDA and the declared ingredients on nutrition fact labels printed by food producers [3]. This is aligned with the GS1 UK data crunch analysis, reporting 80% inconsistency in products data in the UK grocery industry [4]. This inconsistency could be partially explained by the high level of difficulty to find the common name of an additive. For example, the FDA food labeling guide encourages the use of common names, stating "always list the common or usual name for ingredients unless there is a regulation that provides for a different term. For instance, use the term 'sugar' instead of the scientific name 'sucrose' [5]." However, the FDA does not provide a strictly standardized database on the common names and synonyms of additives. While building GroceryDB, we frequently faced the issue that common ingredient names were not used on food packages. For example, the additive commonly known as "baking soda" is frequently declared as "sodium bicarbonate" on product labels. Similarly, "carmine", a common coloring additive, is found in GroceryDB both as its standard name, "carmine", and as "cochineal extract", named after the insect at the origin of its red color.

Lastly, as a note on terminology, the FDA distinguishes between "additives" and "substances added to food." In our analysis we equate the label "additive" to the FDA's "substance added to food."

3 Food Processing Score (FPro)

We used FoodProX, a random forest classifier, to calculate FPro for the branded products in GroceryDB [6]. To train FoodProX, we used the manual NOVA classification on two USDA datasets, namely, FNDDS and Standard Reference (SR). For items with manual NOVA1, NOVA2, and NOVA3, we combined all unique nutrient profiles (12

nutrients) from FNDDS 2001-2018 (9 cycles), plus all the unique nutrient profiles from SR 20-28 (9 versions). This enables us to have more nutrient profiles among unprocessed and processed foods. However, for items with manual NOVA4, we only combined the latest nutrient profile for each food code from FNDDS 2001-2018 and SR 20-28. The reason for not including all unique nutrient profiles in NOVA4 class is to balance the training dataset, otherwise including all unique profiles in NOVA4 would make the training data from NOVA4 extremely dominant, not giving the classifier a chance to learn the characteristics of nutrient profiles in NOVA1-2-3 classes. The fraction of foods in each NOVA class is represented in Figure S2. The addition of SR database to the training data increased the number of training samples for NOVA1 and NOVA3 classes, hence balancing the training dataset.

We used the 12 nutrients in Table 1 to train FoodProX, since FDA requires reporting these nutrients on nutrition fact labels [7]. Although providing the minimum of 12 nutrients is mandated by the law, not all food labels declare those nutrients. Hence, we decided to rely on 10 nutrients and assume value 0 for Vitamins C and A, if these vitamins are not reported. The number of products that reported at least 10 nutrients in GroceryDB is shown in Figure 3. For consistency, we decided to ignore all foods that do not meet the minimal requirement of 10 nutrients [8]. If necessary, the value of missing nutrients could be imputed to measure the degree of food processing for all foods currently excluded.

Table S1: 12 Nutrient Panel for Branded Products

Nutrients
Protein
Total Fat
Carbohydrate
Sugars, total
Fiber, total dietary
Calcium
Iron
Sodium
Fatty acids, total saturated
Cholesterol
Vitamin C
Total Vitamin A

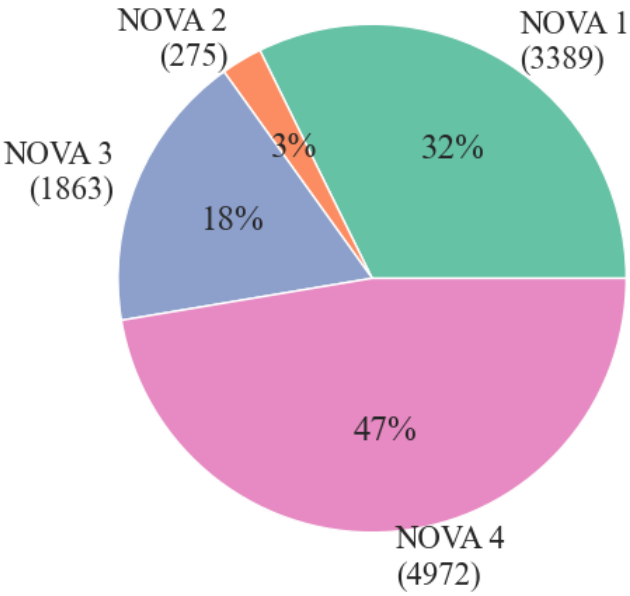


Figure S2: **Fraction of NOVA Classes in the Training Dataset.** We used the NOVA classification to train FoodProX [6], our machine learning classifier that assigns a FPro score to each food.

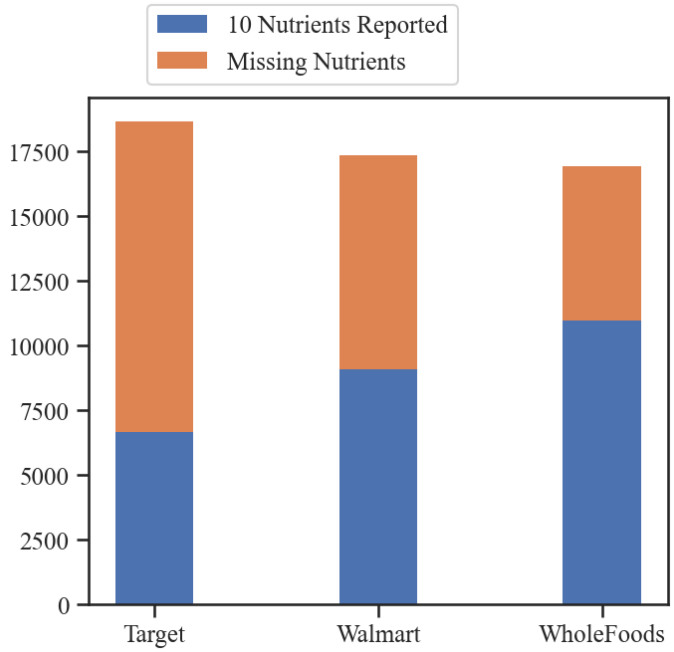


Figure S3

Figure S3: **Number of Products with Missing Nutrients in GroceryDB.** The products that did not report one of the following 10 nutrients are marked as 'missing nutrient': protein, total fat, carbohydrate, total sugars, total dietary fiber, calcium, iron, sodium, total saturated fatty acids, and cholesterol.

4 Comparison with USDA FoodData Central and Open Food Facts Databases

Initially, with the goal of obtaining the missing nutrition facts, we matched GroceryDB with the USDA FoodData Central Global Branded Food Products Database (BFPD) according to the ingredient lists. However, BFPD only covered 44% of the products in GroceryDB (with *Similarity Score* ≥ 0.95), lacking also nutrition facts for the products with missing nutrients in GroceryDB (Figure S4A). Similarly, OFF covers 38% of GroceryDB (Figure S4B).

Specifically, within the 22,900 items with missing nutrition facts, only 9,600 were matched with *Similarity Score* ≥ 0.95 from which only 537 had full nutrition facts. These findings suggest that GroceryDB offers a more up to date picture of the food supply, and many products do not report the full nutrient panel required by the FDA.

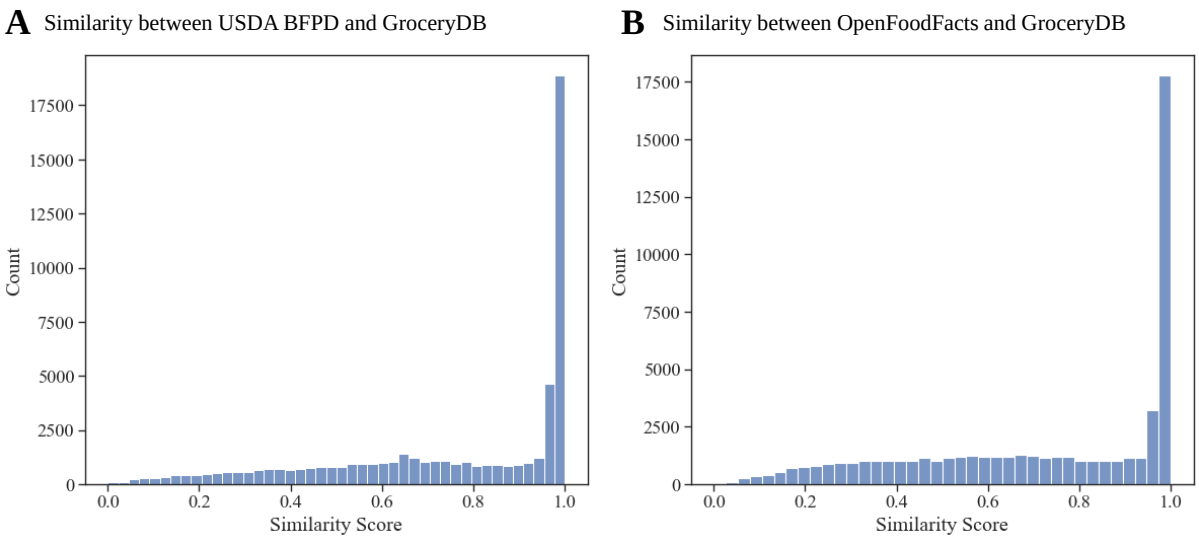


Figure S4

Figure S4: **Coverage Comparison with USDA BFPD and OpenFoodFacts.** GroceryDB offers a complementary picture of the food supply compared to BFPD and Open Food Facts (OFF). (**A**) We derived similarity scores based on ingredient lists declared in USDA BFPD and GroceryDB, using the the TF-IDF (Term Frequency-Inverse Document Frequency) algorithm. The USDA BFPD (April 2021 version) has 1,142,610 branded products and GroceryDB has 50,467 items from which 2,754 items are excluded because of not having an ingredient list. We calculated the similarity score for the remaining 47,713 items, finding that BFPD only covers 44% of the products in GroceryDB with *Similarity Score* ≥ 0.95. **(B)** Similarly, we investigated the overlap between OFF and GroceryDB. Among 426,479 products in OFF with English list of ingredients (as of January 2022), only 18,948 products exist in the entire GroceryDB with *Similarity Score* ≥ 0.95, covering 38% of the products in GroceryDB.

5 Category Harmonization

Grocery stores classify foods into multiple categories and sub-categories, for a total of over 200 main categories and 866 sub-categories that are hierarchically organized in levels. Grocery store categories tend to be organized according to the store layout, helping consumers navigate the store. In contrast, epidemiological databases tend to categorize foods based on processing methods and the origin of food (type of plant or animal parts). For instance, FNDDS 2017-2018 has multiple categories for milk: 'Milk and Milk Products', 'Milks, milk drinks, yogurts, infant formulas', 'Milk, fluid, evaporated and condensed', and 'Milk, fluid, imitation' (declared by first 5 digits of food codes). Another approach to food classification is used by the What We Eat in America (WWEIA) database, aiming to provide categories that better resonate with consumers. For example, WWEIA contains 10 categories for milks, separating milk flavors and fat concentrations, ranging from 'Milk, whole', 'Milk, reduced fat', and 'Milk substitutes', to 'Flavored milk, whole', 'Flavored milk, nonfat', and 'Milk shakes and other dairy drinks' [9]. In GroceryDB, we followed a similar approach as WWEIA, with additional emphasis on the consumer's use of products to enable effective food substitution strategies. For instance, we placed meat-based and plant-based burgers into a single category, 'Prepared Meals & Dishes', since from the consumer perspective these are both ready-to-cook burgers. This method of categorization leads to broader food categories and higher food variability, allowing more opportunities for meaningful food substitution recommendations.

We harmonized foods from grocery stores into 42 broad categories designed for assisting food recommendation algorithms that aim at finding alternative food choices within the same category. For instance, in grocery stores the *frozen-foods* category includes

items ranging from frozen fruits and vegetables to frozen lasagna and breakfast egg bites. We therefore broke the *frozen-foods* category into "Packaged Produce", "Breakfast", and "Prepared Meals & Dishes". The list and size of the harmonized categories in GroceryDB are shown in Figure S5.

The reason for observing such a large number of categories in stores is due to the lack of a standard classifying method across the stores. For example, breads as Level 1 category are marked as "bread bakery", "bakery bread", and "breads rolls bakery" in Walmart, Target, and WholeFoods, respectively (Figure S6).

Lastly, since the focus of this paper is investigating the extent of food processing in grocery stores, we decided to not include the categories that are naturally unprocessed in all analysis, except in Figure 2A and Figure 2C that illustrate the distributions of FPro. Food categories such as fresh produce, raw beans, eggs, and raw meat are example of categories that are naturally unprocessed.

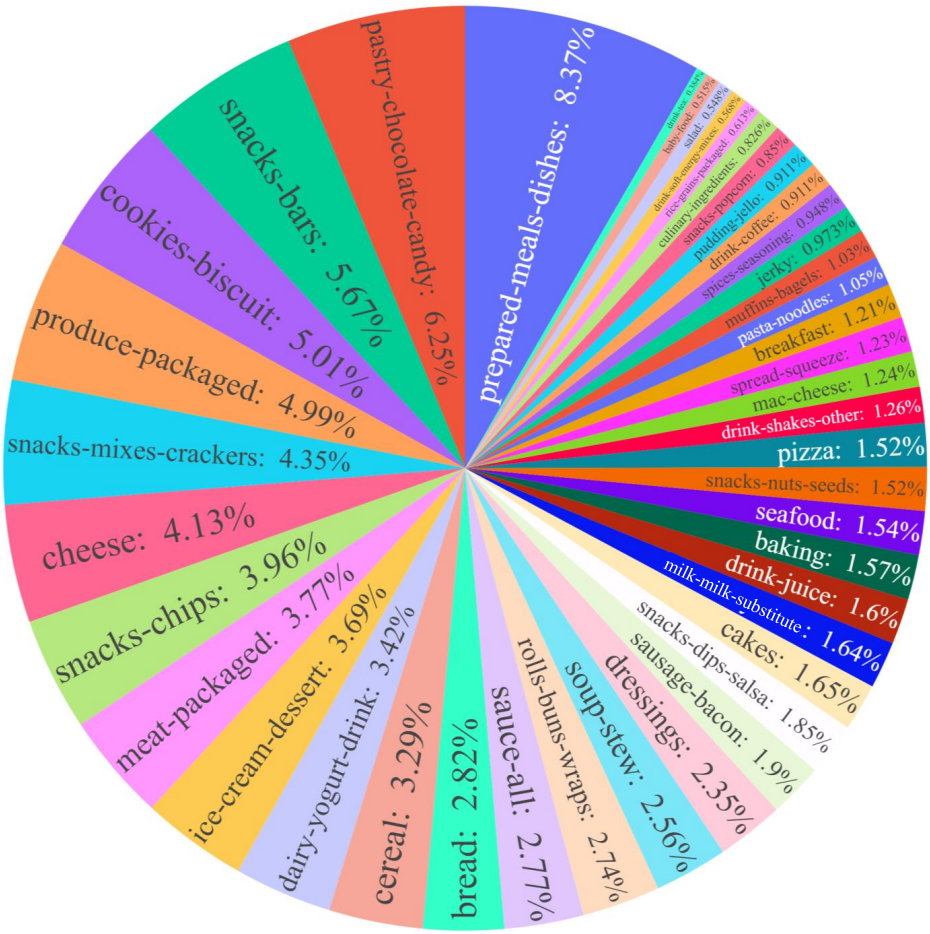


Figure S5: **Fraction of Foods in Harmonized Categories.** We harmonized foods from the three grocery stores into coarse-grained categories to represent foods from the perspective of individuals searching for alternative food choices. A comparison between store categories and GroceryDB categories is illustrated in Figure S6.

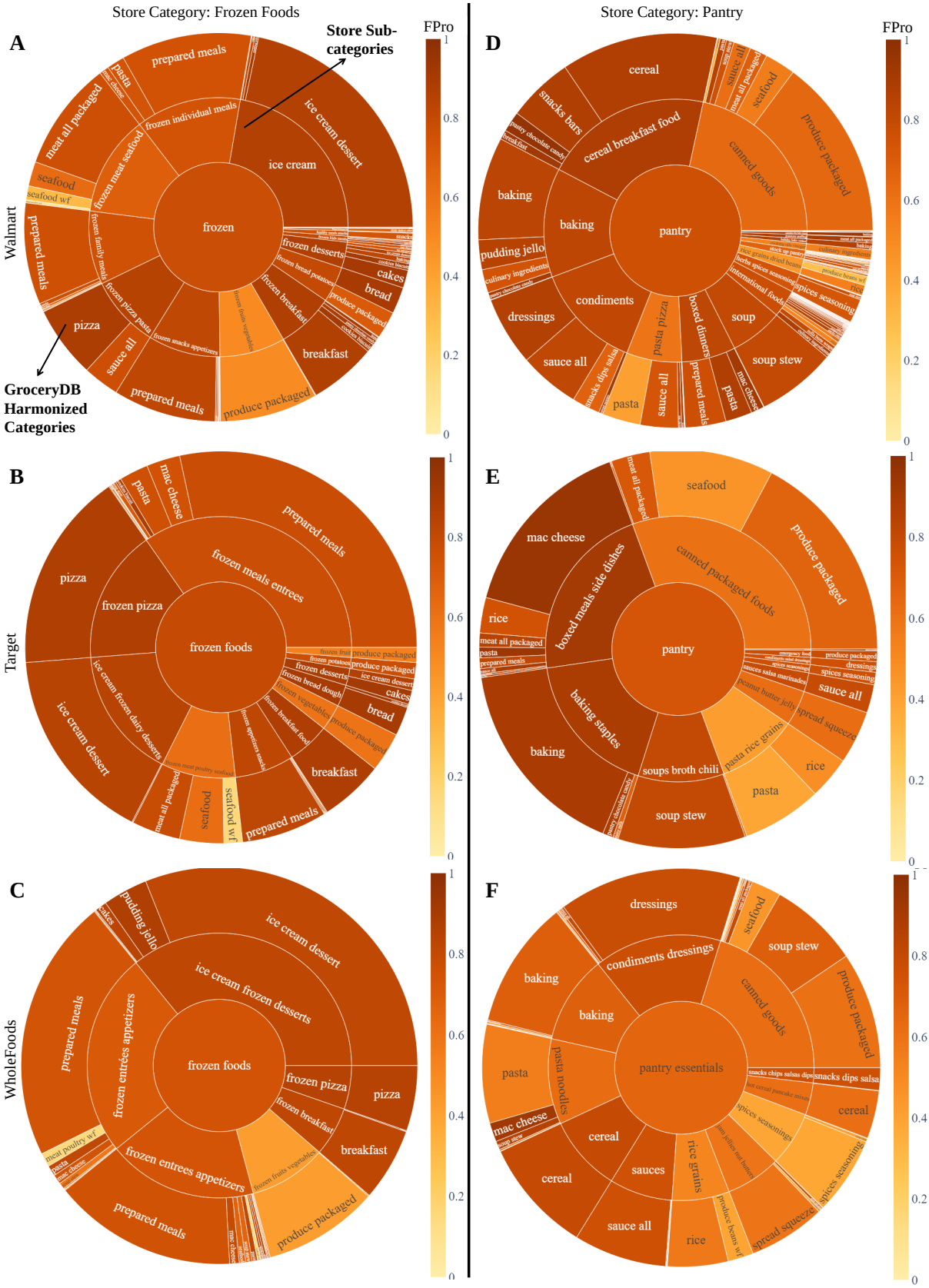

Store Category: Frozen Foods
Store Category: Pantry
FPro
A
B
C
D
E
F
Walmart
Target
WholeFoods
Store Sub-categories
GroceryDB Harmonized Categories
frozen
frozen individual meals
ice cream
ice cream dessert
frozen meat seafood
frozen family meals
frozen pizza pasta
frozen snacks appetizers
frozen breakfast
frozen desserts
frozen bread potatoes
frozen fruits vegetables
prepared meals
pasta
meat all packaged
seafood
seafood wf
pizza
sauce all
produce packaged
cakes
bread
breakfast
frozen foods
frozen meals entrees
frozen pizza
ice cream frozen dairy desserts
frozen breakfast food
frozen appetizers snacks
frozen vegetables
mac cheese
pantry
cereal
cereal breakfast food
canned goods
produce packaged
baking
pudding jello
condiments
dressings
pasta pizza
boxed dinners
soup
soup stew
sauce all
snacks dips salsa
pasta
boxed meals side dishes
canned packaged foods
baking staples
soups broth chili
pasta rice grains
rice
spread squeeze
sauce all
spices seasoning
pantry essentials
condiments dressings
pasta noodles
rice grains
sauces
spices seasonings
snacks dips salsa
cereal
sauce all
rice
spread squeeze
spices seasoning
mac cheese
pudding jello
cakes
ice cream dessert
ice cream frozen desserts
frozen entrées appetizers
frozen entrees appetizers
frozen fruits vegetables
meat poultry wf
mac cheese

Figure S6

Figure S6: **Store Categories vs Harmonized Categories. (A-F)** Breakdown of frozen and pantry foods commonly found in the category hierarchy of grocery stores (inner cycles), mapped onto the GroceryDB harmonized categories (outer cycle).

6 Price and Food Processing

To investigate the hypothesis that processing impacts food prices, we calculated the PricePerCalories of the branded products as the total package price divided by the package calories (Figure S7A-B). Items with zero calories like Coke Zero are ignored in this analysis. Next, we calculated the Spearman's correlation coefficient between FPro and PricePerCalories, as captured by the correlation matrix in Figure S7C. We find that depending on the store and categories, food processing correlates with cheaper calories, as in case of the strong negative correlation for breakfast, mac-cheese, pudding-jello, cakes, and pastry-chocolate-candy. Finally, in a few categories like milk-milk-substitute, pasta-noodles, cheese, and jerky the more processed foods tend to have a higher price (Figure S7C).

To further assess the relationship between PricePerCalorie and FPro, we used robust linear models [10, 11]. The regression coefficients and p-values are illustrated in Figure S8. Note that in Figure 3, we did not include the pasta-noodles category. The reason is lack of data in this category, as out of 256 items we have price for 164 that are highly segmented (Figure S9), leading to the observation that on average the highly processed pasta-noodles are 46% more expensive than minimally processed-alternatives (comparing the averages in the top and bottom 10% of FPro).

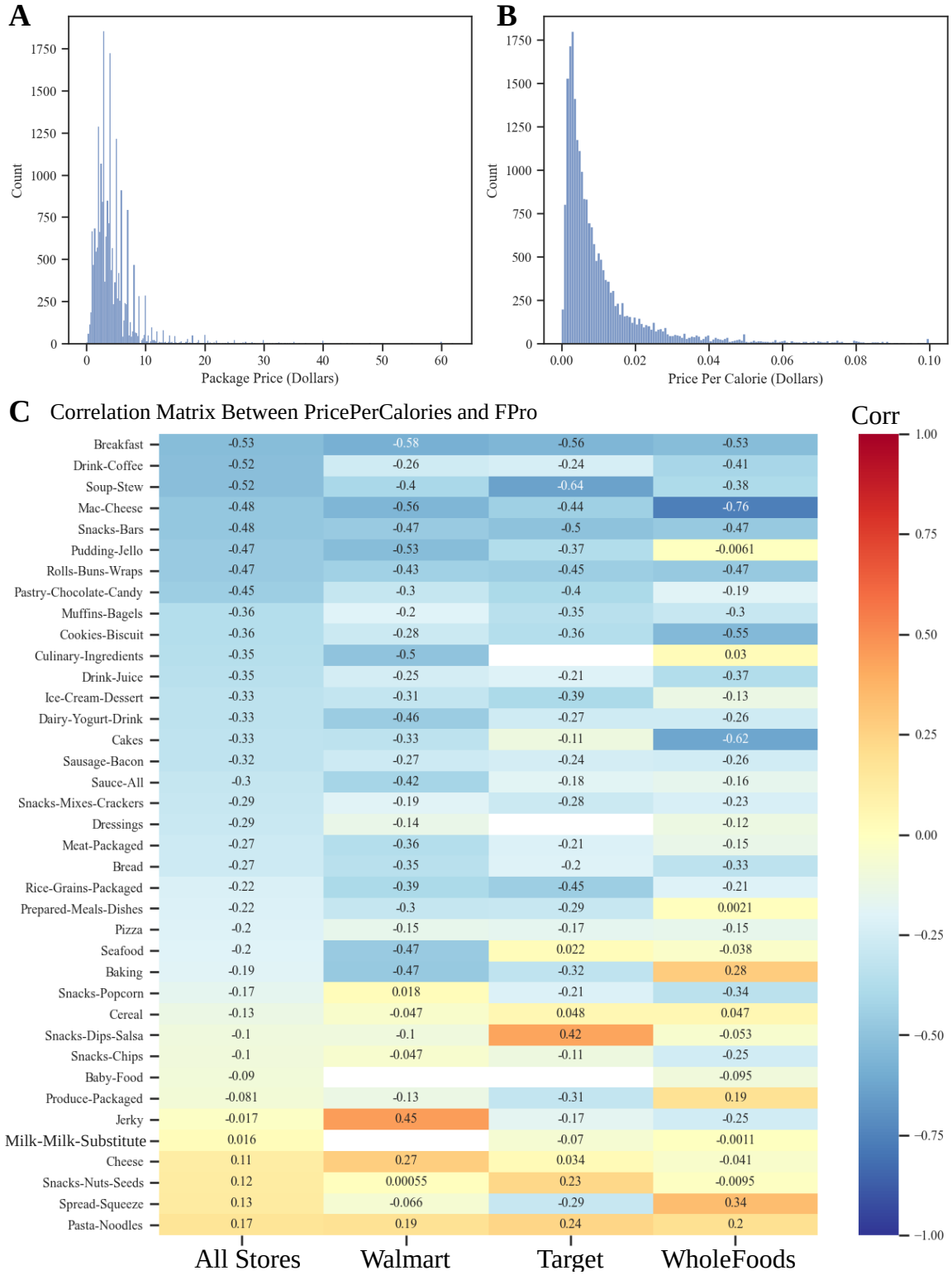


Figure S7

7 Organization of Ingredients

This section describes the methods we developed to quantify the organization of ingredients in the food supply.

Figure S7: **Correlation between Price and Degree of Food Processing.** (**A**) The distribution of price per item in GroceryDB. (**B**) The distribution of price per calorie obtained by dividing an item's price by its total calories (zero calories items are not included in this analysis). (**C**) The Spearman's rank correlation coefficient between price per calorie and FPro across food items. An higher extent of food processing tends to decrease cost in many categories, but not in all of them.

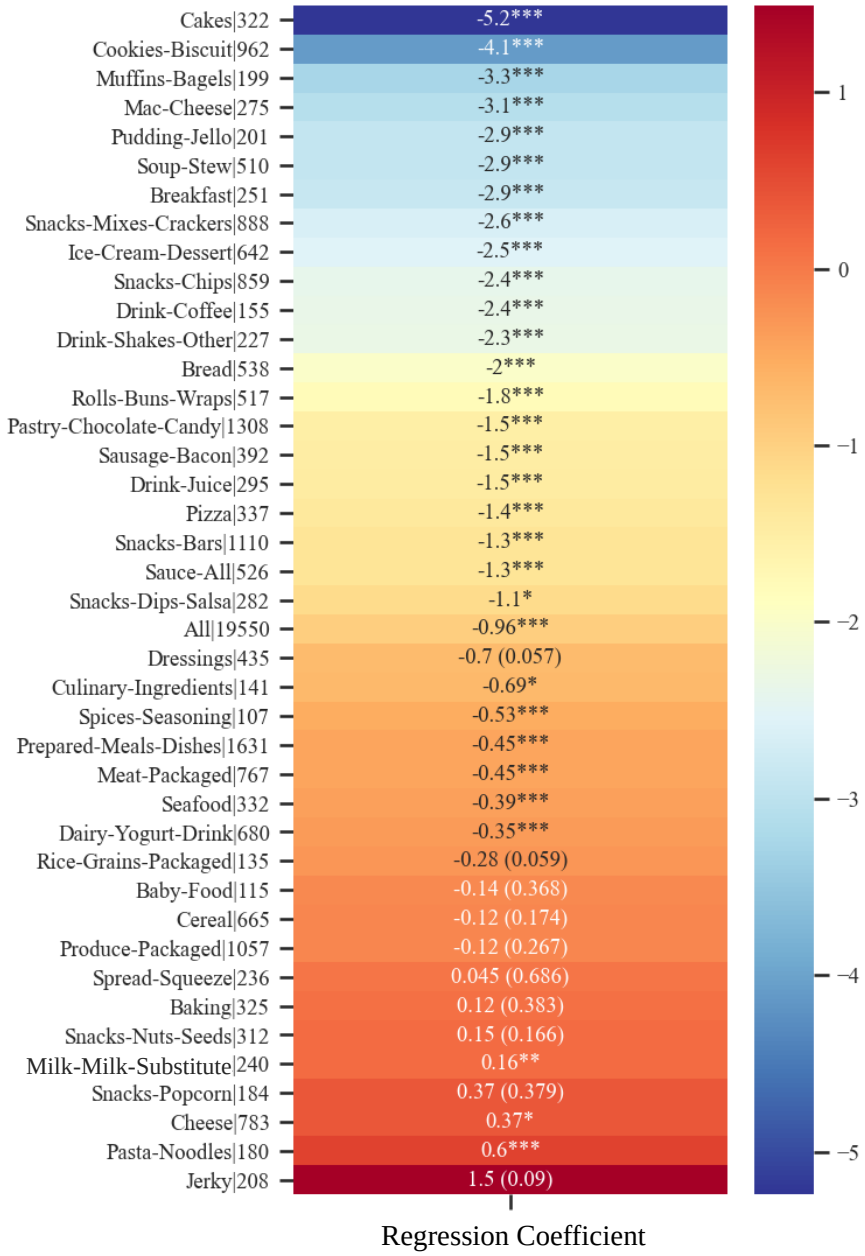


Figure S8: **Price and Degree of Processing**. The regression coefficients and p-values for $log(PricePerCalorie) \sim log(FPro)$ using robust linear models are illustrated to assess the relationship between price and FPro in GroceryDB [10, 11]. The regressions with p-value ≤ 0.05 are marked with stars, otherwise the p-values are represented in parenthesis.

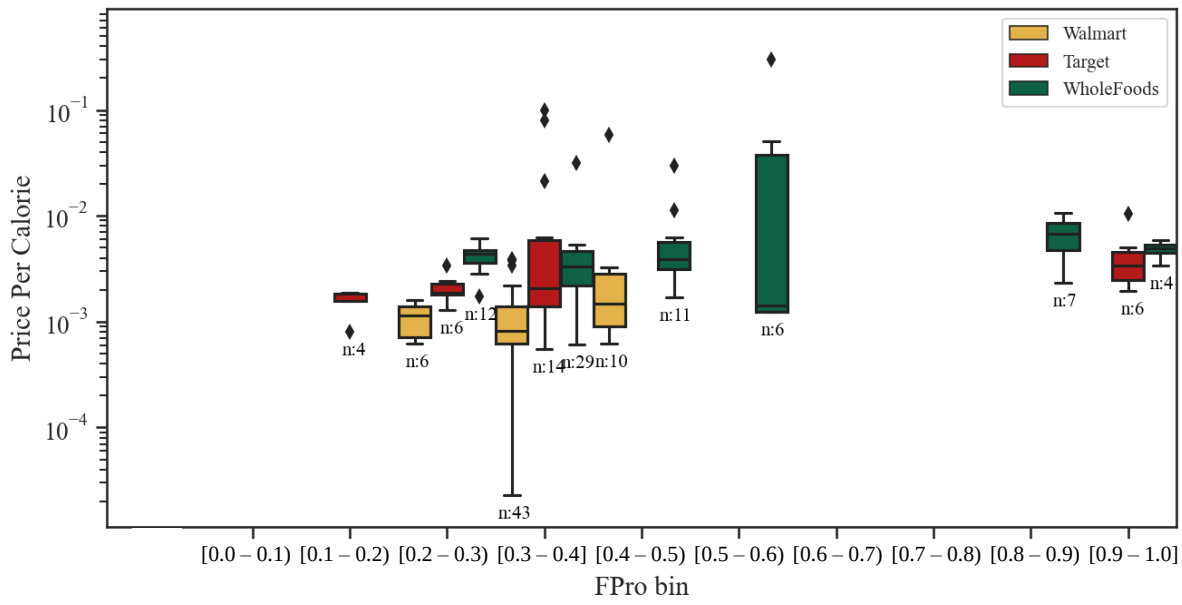


Figure S9: **PricePerCalorie vs. FPro for Pasta & Noodles Category.** Of the 256 items in the pasta-noodles category, we have price per calorie for 164 items. Given the limited number of data points in this category, we decided to exclude pasta-noodles from Figure 3.

7.1 Cleaning Ingredient Lists

The ingredient lists are regulated by the FDA food labeling guidelines, however there are numerous nuances that cause inconsistency and require normalization [12]. For example, corn starch is reported as "cornstarch", "corn-starch", and "corn starch" in the ingredient lists, and needs to be normalized to one format. Similarly, FDA allows using a variety of synonyms for an ingredient, like the common synonyms "soybean", "soy", and "soya" used for soybeans. Additionally, ingredient lists often provide a variety of information for individual ingredients, like the processes involved in their preparation or the intended purpose of their use. We normalized these information by identifying the general name of an ingredient and using descriptors to mark any extra information provided in each ingredient label (Figure S10).

The heterogeneity and variability of declared labels in the ingredients lists requires a base-knowledge to organize the list of ingredients into ingredient trees. Hence, we built a set of semantic trees by leveraging the dictionary of food ingredients [2], and by manually investigating the ingredient list of products in GroceryDB. The purpose of a semantic tree is to capture the common forms representing the same ingredient label by

distinguishing the differences in the origin of an ingredient, the alterations caused by food processing, and the semantic synonyms. For instance, we have a semantic tree for

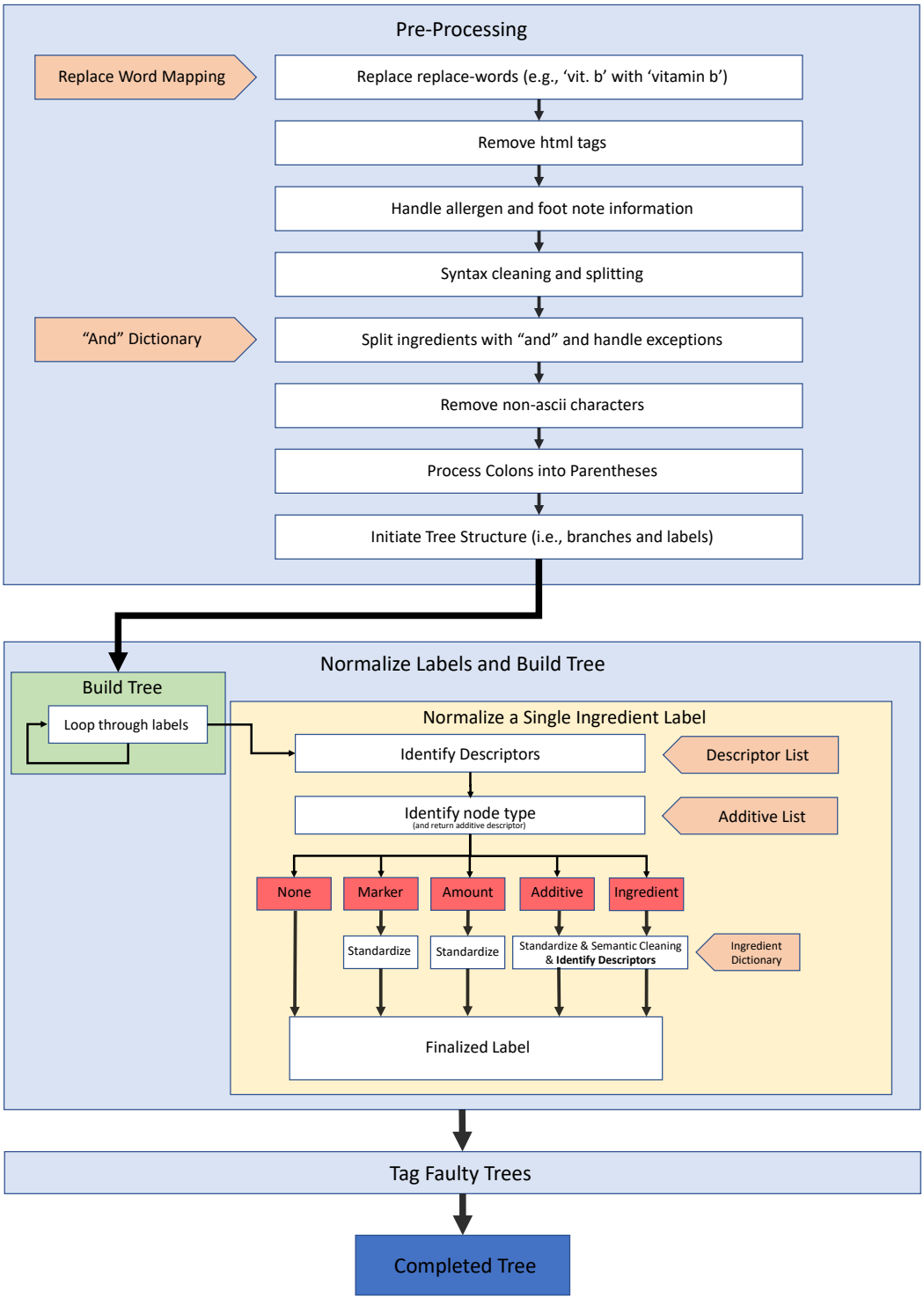


Figure S10: **Pipeline for Transforming an Ingredient List into an Ingredient Tree.** An extensive data engineering and cleaning is needed to harmonize the list of ingredients declared on foods in grocery stores. The lack of a standardized ontology of ingredients makes this task more difficult.

oil that covers 25 types of oils ranging from seed and vegetable oils (like sesame and corn) to fruit based oils (palm and avocado), also annotating the markers of processing such as expeller pressed, hydrogenated, and partially hydrogenated (Figure S11A). Similarly, we obtained a semantic tree for common generic terms like 'starch', that captures various types of starches declared on food labels, from potato starch (organic or modified) to corn starch (native or non-GMO, Figure S11B). Lastly, we organized this extra information with descriptors, annotated as "<descriptor>" in cleaned ingredient labels.

Finally, the breakdown of raw ingredients normalized to structured labels is illustrated in Figure S12. The raw format of ingredients as reported on product packages contains over 32,000 unique labels, including a broad range of heterogeneous synonyms and semantically duplicate labels. Through text curation, we were able to reduce this number to approximately 20,000 labels corresponding to about 12,000 unique ingredients. See Section S7.2 for a robust assessment of the number of ingredients, additives, and descriptors in GroceryDB.

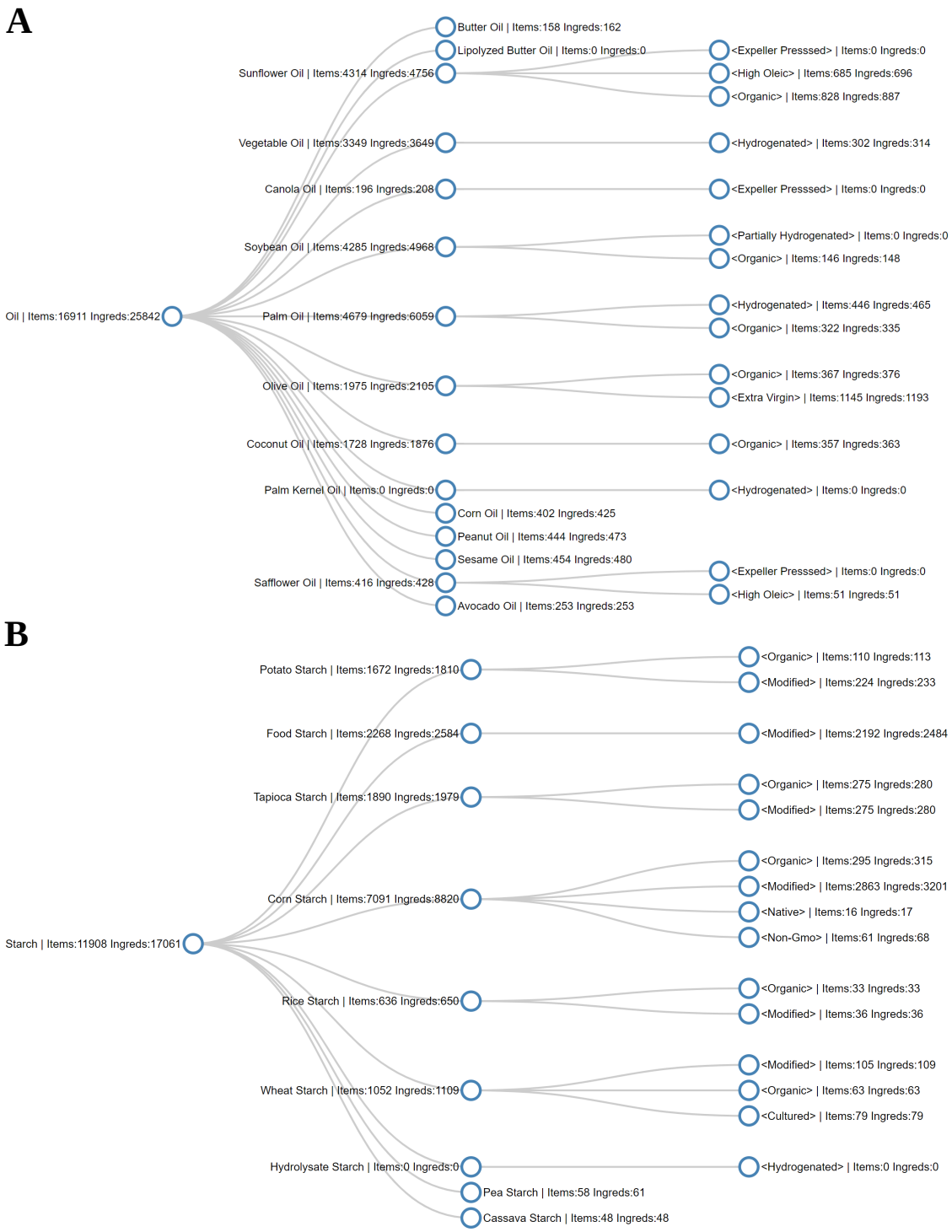


Figure S11

Figure S11: **Semantic Trees.** Two example of semantic trees used to organize ingredient lists into ingredient trees, with the number of products and ingredients for ingredients and their descriptor.

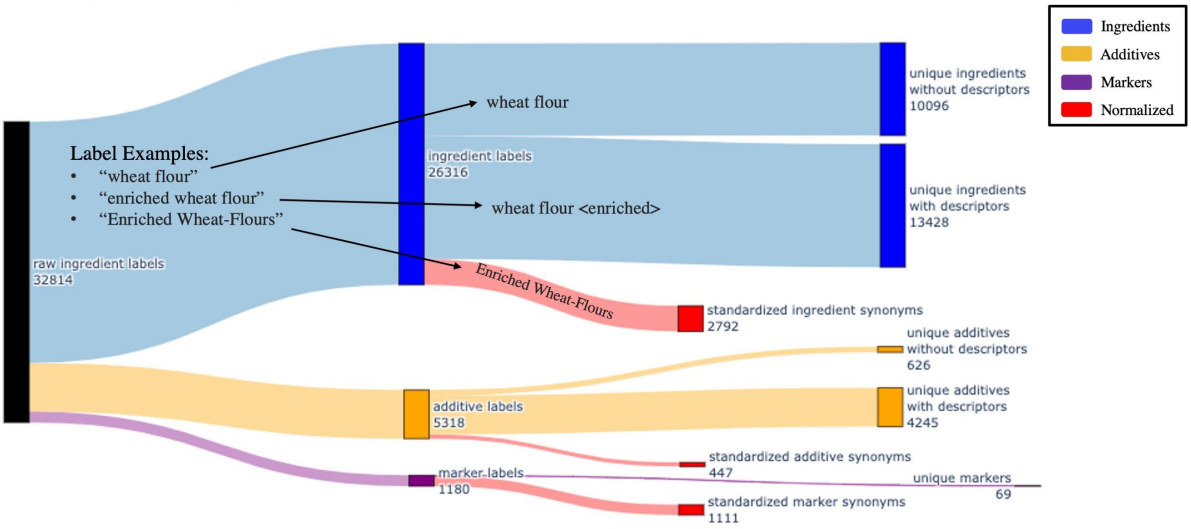


Figure S12

Figure S12: **Normalization of Ingredient Labels**. Breakdown of the raw labels obtained from the declared ingredients on branded products normalized into structured labels in GroceryDB.

7.2 Approximating the Number of Ingredients in GroceryDB

Given the high level of data inconsistency in the grocery industry [4], it is difficult to find the exact number of ingredients and additives in GroceryDB. Hence, we use the population proportion estimation statistics (Cochran's sample size formula) to estimate the number of ingredients, additives, and descriptors. We estimate the number of unique ingredients to be 12,475 (Figure S13A), including 2,639 additives (Figure S13B). We present two quantities, one considering descriptors and another without descriptors, providing two levels of ingredient specificity. With descriptors, for example, the string "bleached wheat flour" is counted separately from "enriched wheat flour." Without descriptors, both are considered "wheat flour." Counts become considerably larger when considering descriptors. Through this analysis we found additives like "corn starch" which appears in Grocery DB with 13 different unique descriptors like "modified," "non-gmo" and "resistant."

The FDA identifies 3,972 total substances added to food [1], 1,316 of which are classified as direct, approved additives [13]. The Dictionary of Food Ingredients (DFI), an industry standard encyclopedia of food ingredients in the U.S. [2], lists 609 additives. The DFI is based on the FDA's Title 21 in the Code of Federal Regulations [13], and

there is thus a considerable overlap between the two sources. With our current pipeline and duplicate removal, DFI adds 205 additives beyond the FDA source. We use both DFI and the FDA's database of additives to identify additive ingredients in GroceryDB, resulting in an estimated 914 unique additives not considering descriptors.

This task is difficult as there is not a one-to-one mapping between unique strings in the ingredient lists, and an actual ingredient. For example, for the ingredient "vitamin b12", we found typos like "vitemin b12", branded strings ("Walmart vitamin b12"), and synonyms ("vit. b12"), which all point to the same ingredient. A full cleaning of ingredient strings to achieve a one-to-one mapping is a goal of future work, but we can approximate the number of unique ingredients in the U.S. food system with the current state of GroceryDB.

To estimate the number of ingredients and additives, we collected samples of ingredient labels after running our data cleaning pipeline (Figure S10) and selected ingredients present in at least two products. Then, we manually investigated these samples to count the number of labels that are correctly classified vs. incorrectly classified (due to the lack of a comprehensive dictionary of synonyms and the large level of data inconsistency in the grocery industry). Next, we used the Cochran's formula to estimate the number of ingredients and additives based on the proportion of correctly classified labels.

Finally, GroceryDB paves the way towards systematically quantifying the organization of ingredients in the food supply. Indeed, the level of data inconsistency is estimated to be 80% in the grocery industry [3,4]. Given this high level of data inconsistency, future work will extend the current efforts on data cleaning, integration, and normalization, to better approximate the number of unique ingredients and additives in the food supply. Yet, GroceryDB provides the data structure, methods, and pipeline needed to systematically unify the ingredient lists in the food supply, and unveil the organization of ingredients.

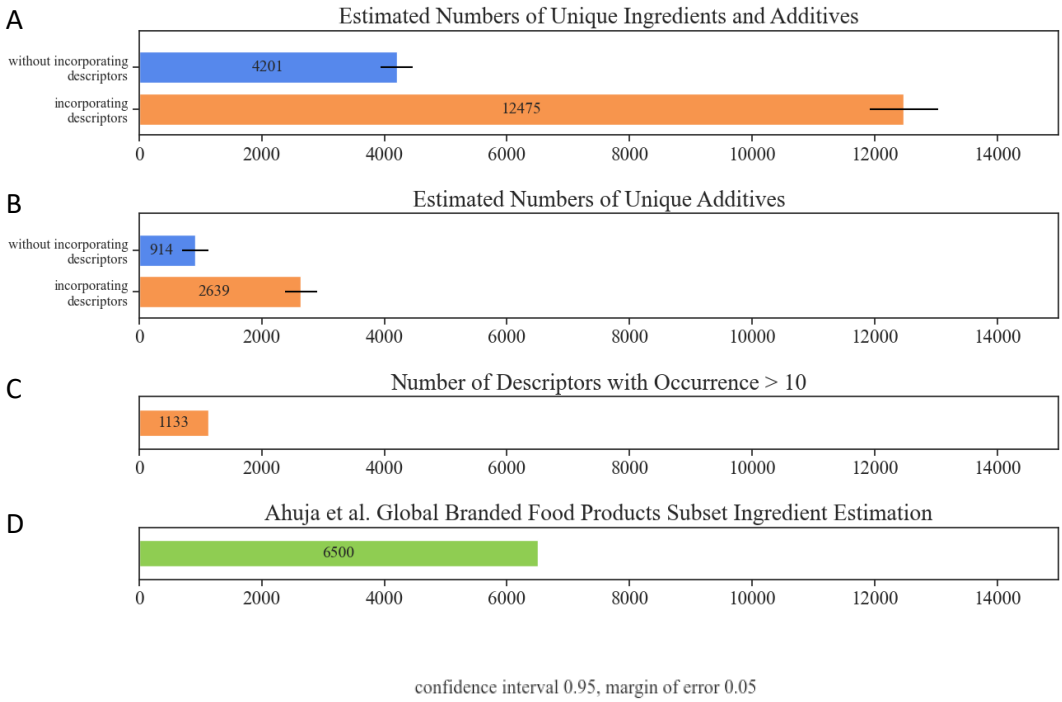


Figure S13: **Estimated Number of Ingredients, Additives, and Descriptors in GroceryDB. (A-B)** Estimation for the unique number of ingredients and additives in GroceryDB based on Cochran's formula with 95% confidence interval. The descriptors add a significant complexity in estimating the number of ingredients and additives in the food supply. Examples of ingredients with descriptors are "bleached wheat flour" and "enriched wheat flour" where 'bleached' and 'enriched' are descriptors. **(C)** The number of descriptors that appeared at least in 10 products. Although we manually created a dictionary of descriptors with 168 labels, we mainly relied on natural language processing to automatically identify descriptors. This process resulted in identifying new descriptors. **(D)** Ahuja et al in [3] analyzed a subset of BFPD, resulting in the identification of 6,500 ingredients from 5 out of 31 food categories in BFPD. Our data cleaning led to the identification of a smaller number of ingredients without descriptors, for a total of 4,201 ingredients.

7.3 Characteristics of Ingredient Trees

The branded products with more complex list of ingredients are more likely to be highly processed. To test this hypothesis, first we introduce two measures characterizing ingredient trees: tree width and depth. The tree width, denoted by W, represents the number of main ingredients in a product, and the tree depth, D, approximates the extent of the reliance on mixtures of sub-ingredients. We define depth-sum of an ingredient

tree, denoted by D_s, as the sum of the depth of all its branches. Figure 5 presents two cheesecakes with $FPro = 0.953$ and $FPro = 0.720$ along with their corresponding ingredient trees. The highly processed cheesecake has a complex ingredient tree with $W = 10$ and $D_s = 4$ (Figure 5A). In contrast, the less processed cheesecake has a simpler ingredient tree with $W = 5$ and $D_s = 2$ (Figure 5B).

Theoretically, by the definition of branch depth sum, D_s may be considered as another representation of W, if most main ingredients of products have sub-ingredients. Yet, we find that W and D_s show varying characteristics, depending on food categories. For example, cheeses tend to have high W and low D_s, whereas cakes have relatively lower W and higher D_s, showing distinct behaviors as illustrated in Figure S14A. In contrast, when comparing cakes and pizzas (Figure S14B), the difference between W and D_s is less striking. Moreover, the distributions of W and D_s show more differences than similarities as illustrated in Figure S15A-B, indicating that they are presenting different information. Interestingly, we find that WholeFoods tend to offer foods with both smaller W and shorter D_s compared to the other stores. That is, products in WholeFoods tend to have less ingredients and also rely less on mixing sub-ingredients.

Finally, We analyzed the relationship between W and D_s of ingredient trees in all categories. The Spearman's correlation between W and D_s suggests that these metrics capture unique information about the products. We find both positive and negative correlations between W and D_s, depending on the categories (Figure S15C). The strongest cor-

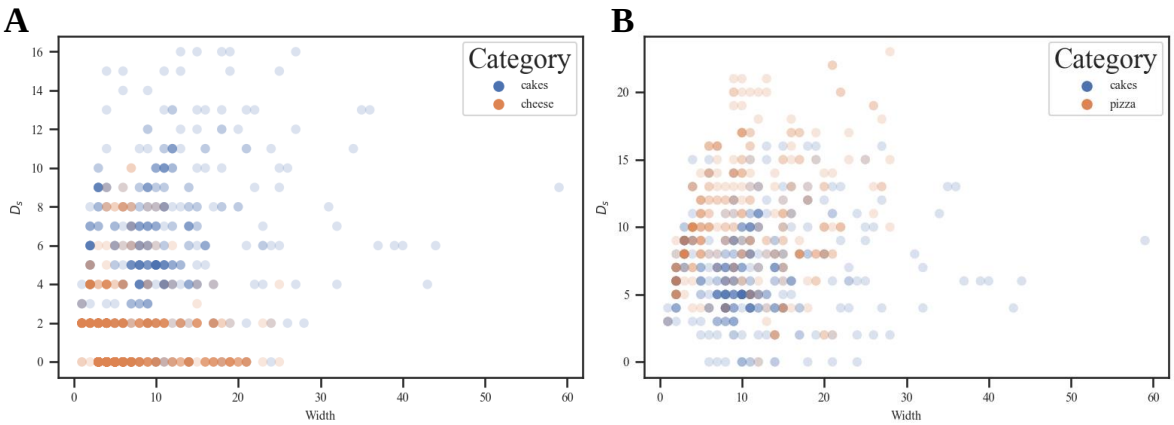


Figure S14: **Tree Width vs Depth-Sum (D_s).** The ingredient trees behave differently according to food categories. (**A**) In the cheese category the ingredient lists tend to rely less on the mixture of sub ingredients (wider trees), whereas in cakes we observe that more sub-ingredients are defined. (**B**) The ingredient trees of cakes and pizzas show similar structures.

relation between W and D_s is in the culinary ingredients and spices-seasoning, since these are the categories that in principle should least rely on the mixture of sub-ingredients, as generally spices and culinary ingredients are made of simple ingredients. In milk & milk-substitutes, we also find a strong correlation between W and D_S, partially explained by the transition from simple whole milk to chocolate and milk-substitutes (like almond and oat milks), increasing the number of main ingredients and sub-ingredients (Figure S15). On the other hand, in breads, ice cream, and sausage, we find a negative correlation between W and D_s partially explained by the use of more complex ingredients. For instance, flour, dough conditioner, and cookie dough are often reported with a long mixture of sub-ingredients, resulting in a higher D_s in breads and ice creams.

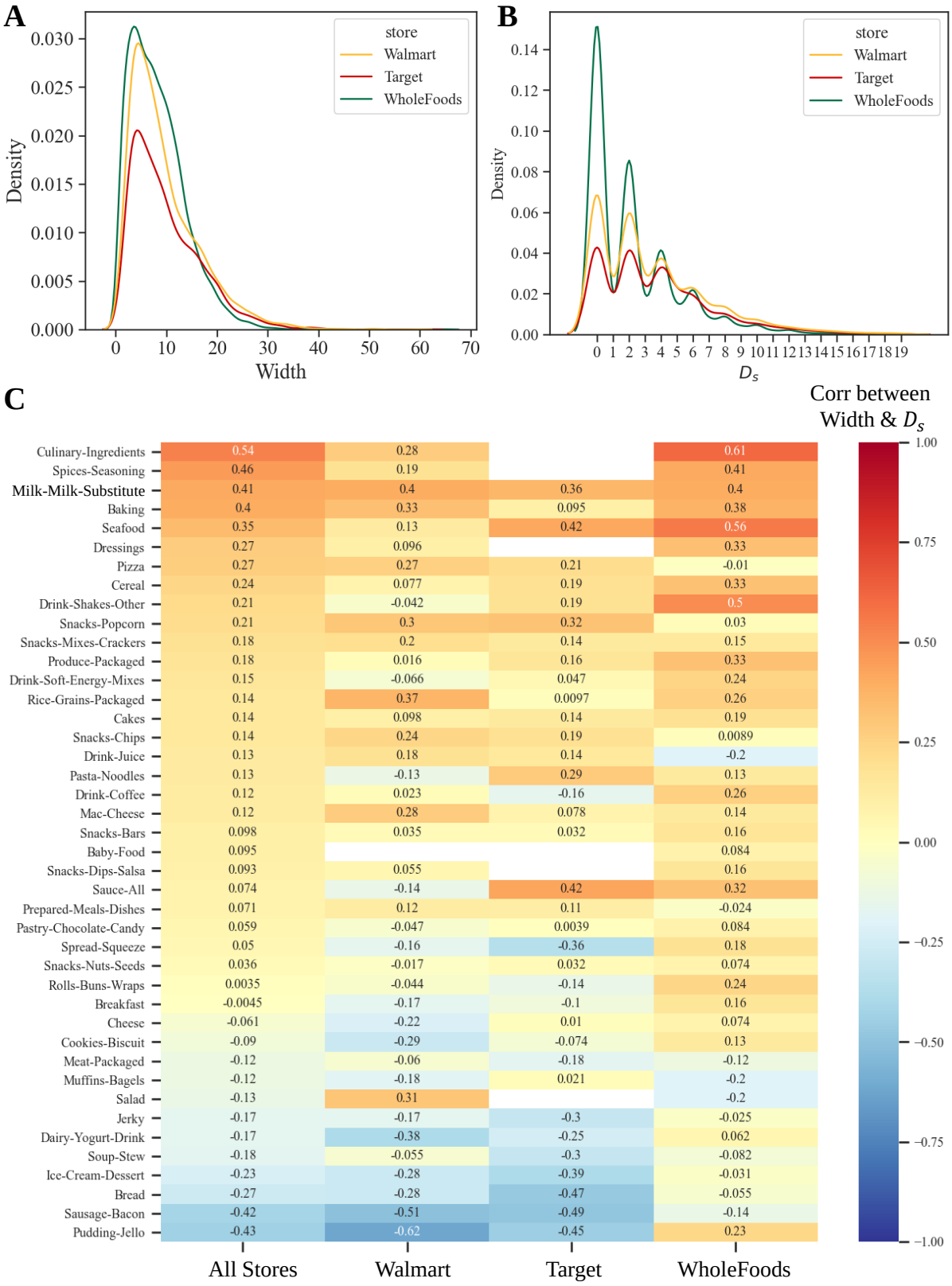

A
Density
Width
store
Walmart
Target
WholeFoods
B
Density
D_s
store
Walmart
Target
WholeFoods
C
Corr between Width & D_s
Culinary-Ingredients 0.54 0.28 0.61
Spices-Seasoning 0.46 0.19 0.41
Milk-Milk-Substitute 0.41 0.4 0.36 0.4
Baking 0.4 0.33 0.095 0.38
Seafood 0.35 0.13 0.42 0.56
Dressings 0.27 0.096 0.33
Pizza 0.27 0.27 0.21 -0.01
Cereal 0.24 0.077 0.19 0.33
Drink-Shakes-Other 0.21 -0.042 0.19 0.5
Snacks-Popcorn 0.21 0.3 0.32 0.03
Snacks-Mixes-Crackers 0.18 0.2 0.14 0.15
Produce-Packaged 0.18 0.016 0.16 0.33
Drink-Soft-Energy-Mixes 0.15 -0.066 0.047 0.24
Rice-Grains-Packaged 0.14 0.37 0.0097 0.26
Cakes 0.14 0.098 0.14 0.19
Snacks-Chips 0.14 0.24 0.19 0.0089
Drink-Juice 0.13 0.18 0.14 -0.2
Pasta-Noodles 0.13 -0.13 0.29 0.13
Drink-Coffee 0.12 0.023 -0.16 0.26
Mac-Cheese 0.12 0.28 0.078 0.14
Snacks-Bars 0.098 0.035 0.032 0.16
Baby-Food 0.095 0.084
Snacks-Dips-Salsa 0.093 0.055 0.16
Sauce-All 0.074 -0.14 0.42 0.32
Prepared-Meals-Dishes 0.071 0.12 0.11 -0.024
Pastry-Chocolate-Candy 0.059 -0.047 0.0039 0.084
Spread-Squeeze 0.05 -0.16 -0.36 0.18
Snacks-Nuts-Seeds 0.036 -0.017 0.032 0.074
Rolls-Buns-Wraps 0.0035 -0.044 -0.14 0.24
Breakfast -0.0045 -0.17 -0.1 0.16
Cheese -0.061 -0.22 0.01 0.074
Cookies-Biscuit -0.09 -0.29 -0.074 0.13
Meat-Packaged -0.12 -0.06 -0.18 -0.12
Muffins-Bagels -0.12 -0.18 0.021 -0.2
Salad -0.13 0.31 -0.2
Jerky -0.17 -0.17 -0.3 -0.025
Dairy-Yogurt-Drink -0.17 -0.38 -0.25 0.062
Soup-Stew -0.18 -0.055 -0.3 -0.082
Ice-Cream-Dessert -0.23 -0.28 -0.39 -0.031
Bread -0.27 -0.28 -0.47 -0.055
Sausage-Bacon -0.42 -0.51 -0.49 -0.14
Pudding-Jello -0.43 -0.62 -0.45 0.23
All Stores
Walmart
Target
WholeFoods

Figure S15

Figure S15: **Characteristics of Ingredient Trees.** (**A**) The distribution of the width of ingredient trees for each store, indicating that products in WholeFoods tends to have fewer ingredients compared to other stores. (**B**) The distribution of D_s for all ingredient trees, reflecting the extent to which products rely on sub-ingredients. WholeFoods tends to rely less on mixing sub-ingredients compared to the other stores. (**C**) The Spearman's correlation coefficient between the width and D_s of ingredient trees for each harmonized category and store.

7.4 Correlation between Characteristics of Ingredient Lists and FPro

To test the hypothesis that branded products with more complex list of ingredients are more likely to be highly processed, we also investigate the relationship between W, D_s, and FPro. For example, in cereals, pasta-noodles, and baking categories, the items that have simpler ingredient trees also have a significantly lower FPro (Figure S16A-C). However, this effect is weaker in prepared Meals & dishes, where we find lower values of FPro with relatively large ingredient trees (Figure S16D).

Lastly, we further investigate the Spearman's correlation between W, D_s, and FPro. Generally, we find a strong correlation between W, D_s, and FPro indicating that products with complex ingredient trees tend to have a higher FPro (Figure S17). Also, some categories show stronger correlation with D_s and FPro, signaling that mixing many sub-ingredients may drive food processing. For example, in breakfast products, pizzas, popcorn, we find that D_s has a stronger correlation with FPro compared to W and even the total number of ingredients (defined as the sum of all ingredients and sub-ingredients including additives, Figure S17).

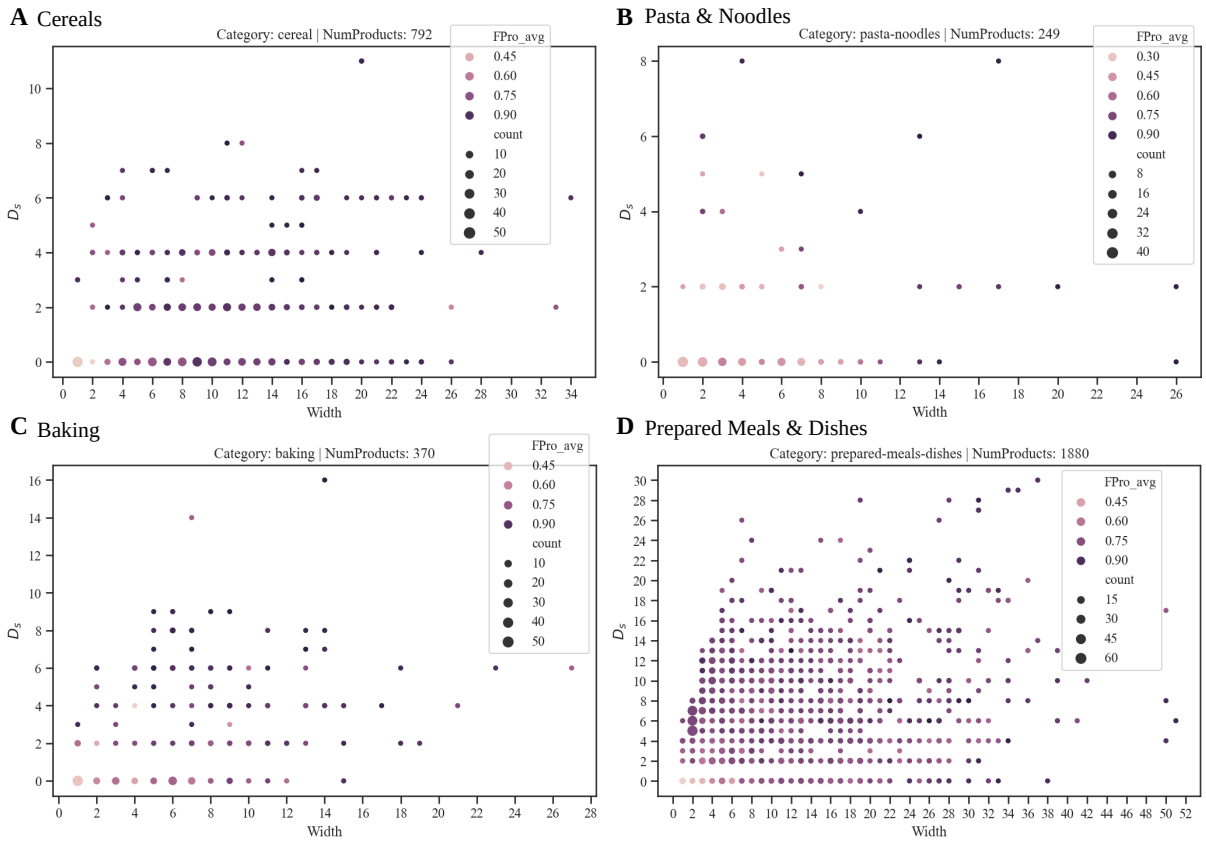


Figure S16: **The Relationship between FPro and Characteristics of Ingredient Tree.** Complex ingredient trees tend to have a higher FPro. **(A-C)** In cereals, pasta-noodles, and baking categories, we find that FPro increases as tree width W and D_s increases. **(D)** This effect is weaker in prepared meals & dishes.

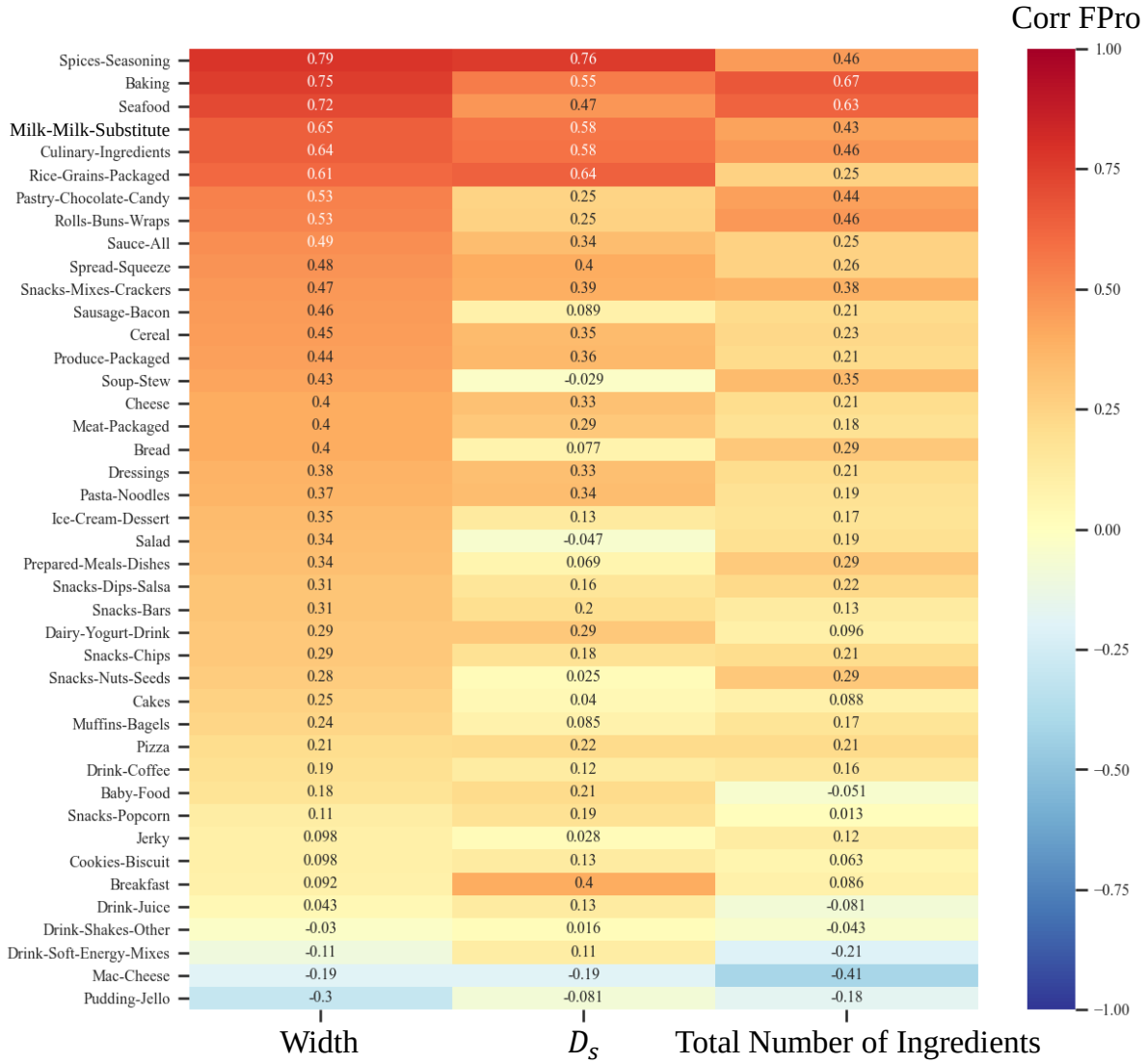


Figure S17: **Correlation between FPro and the Characteristics of Ingredient Trees.** The Spearman's correlation between FPro and the ingredient trees width, D_s, and total number of ingredients. The features of ingredient trees are not always positively correlated with FPro, depending on the food category. In some categories like seafood and milk-milk-substitute, there is a strong positive correlation between FPro and the characteristics of ingredient trees. However, we also observe negative correlations in mac-cheese and pudding-jello.

7.5 Ingredient Processing Score (IgFPro)

This section provides complementary information on the methods to create ingredient trees (Figure S18), and compares IgFPro with FPro (Figure S19).

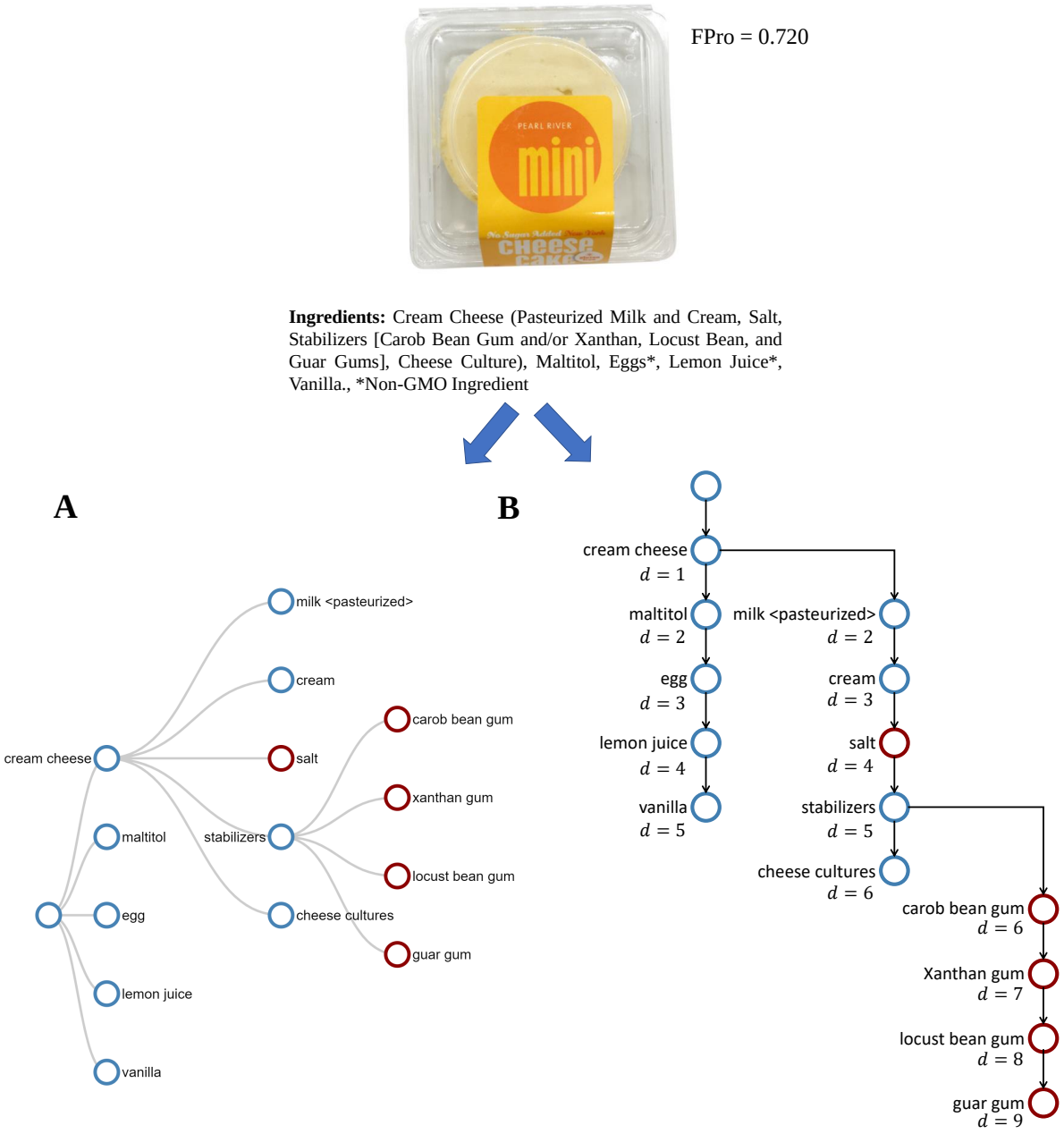


Figure S18: **Two Types of Ingredient Trees.** An ingredient list can be represented by two types of ingredient trees. (**A**) A recipe-like structure to better demonstrate the main and sub-ingredients. (**B**) A sequential approach to capture the order of ingredients in the tree structure. In this approach, the distance d from the root reflects a ranking for the amount of ingredients used in the preparation of an item.

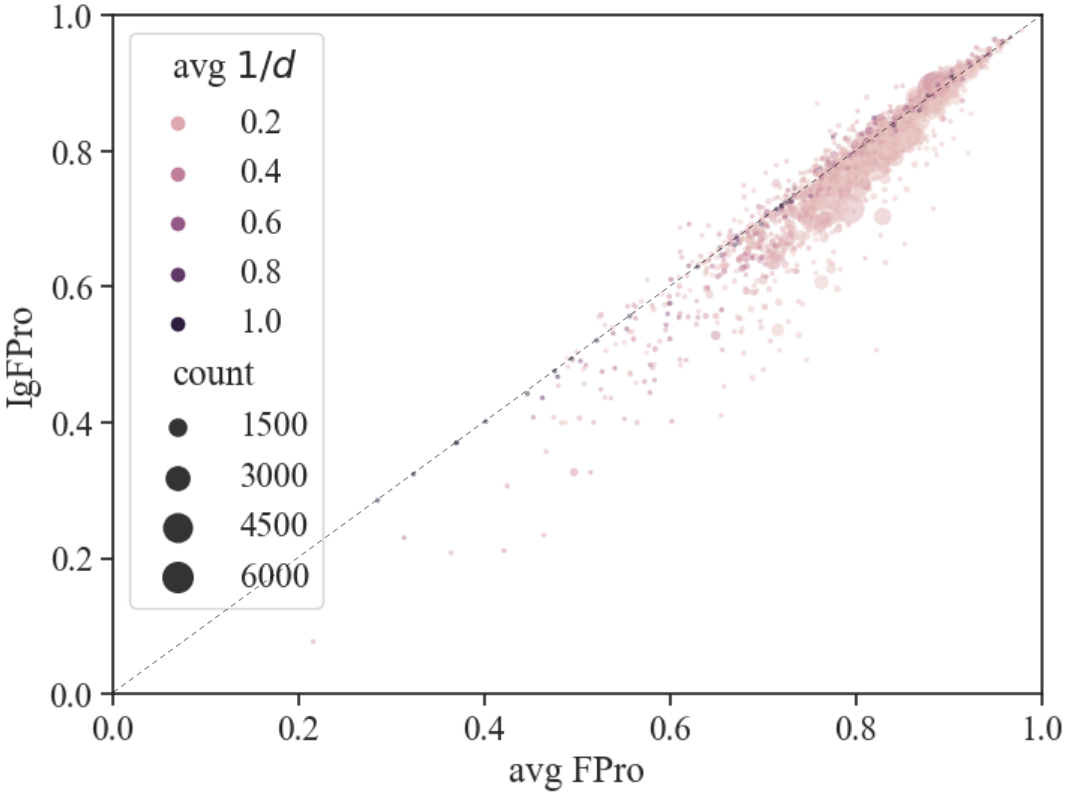


Figure S19: **IgFPro vs FPro.** The general form of ingredients is used without descriptors. For example, the general name of 'milk <pasteurized>' without descriptors is 'milk.' Also, only the general ingredients that are present in at least 10 products are considered, resulting in IgFPro measures for over 1,200 ingredients, as ranked in Figure 6.

8 Case Study on Rice Cakes

FPro can be leveraged to investigate the variations in the degree of processing within food subgroups and pinpoint different processing fingerprints. For example, the category of rice cakes comprises popular low-calorie carb snacks with a high glycemic index. In GroceryDB we find 17 rice cakes displaying a wide range of variability in FPro (median FPro=0.7964 , first quartile Q1=0.6939, third quartile Q3=0.9302, see Figure S20). The top 3 processed rice cakes ("Great Value Birthday Cake Crispy Rice Treats, 0.78 oz, 8 Count", "Quaker Gluten-Free Rice Cakes, Caramel, 6.5 Oz", and "Quaker Chocolate Crunch Large Rice - Cakes 7.23oz") have FPro values 0.9702, 0.9538, and 0.9372 respectively, reflecting the need to apply food processing to remove gluten (e.g., removing germ from corn as indicated on the product's ingredient list) and the use of sorghum and corn syrups, colors, chocolate chips, resulting in higher sugar content. Conversely, among the least processed rice cakes we find "Quaker Lightly Salted Gluten Free Rice Cakes - 4.47oz" (FPro=0.6802) and "Organic Cinnamon Toast Rice Cakes, 9.5 oz" (FPro=0.6473) which exhibit simpler ingredients as well as lower sugar and the use of organic ingredients. Yet, from a broader perspective, none of these products can be considered "minimally-processed", but rather "processed" or "ultra-processed".

List of all rice cakes in GroceryDB sorted by FPro (decreasing):

1. Great Value Birthday Cake Crispy Rice Treats, 0.78 oz, 8 Count (FPro: 0.9702) URL: [Product Link @ TrueFood.Tech]

2. Quaker Gluten-Free Rice Cakes, Caramel, 6.5 Oz (FPro: 0.9538) URL: [Product Link @ TrueFood.Tech]

3. Quaker Chocolate Crunch Large Rice - Cakes 7.23oz (FPro: 0.9372) URL: [Product Link @ TrueFood.Tech]

4. Himalayan Sea Salt Sprouted Rice And Cauliflower Rice Cake, 5 oz (FPro: 0.9343) URL: [Product Link @ TrueFood.Tech]

5. Quaker Garden Tomato & Basil Rice Cakes - 6.1oz (FPro: 0.9261) URL: [Product Link @ TrueFood.Tech]

6. Quaker Gluten-Free Apple Cinnamon Rice Cakes, 7.04 Oz. (FPro: 0.9197) URL: [Product Link @ TrueFood.Tech]
7. Quaker Caramel Corn Gluten Free Rice Cakes - 6.56oz (FPro: 0.8846) URL: [Product Link @ TrueFood.Tech]
8. Quaker Rice Cakes, Apple Cinnamon, 6.53 Oz (FPro: 0.8256) URL: [Product Link @ TrueFood.Tech]
9. Quaker Rice Cakes, Lightly Salted, 4.47 Oz (FPro: 0.7964) URL: [Product Link @ TrueFood.Tech]
10. Organic Tamari With Seaweed Rice Cakes, 8.5 oz (FPro: 0.7858) URL: [Product Link @ TrueFood.Tech]
11. Organic Lightly Salted Brown Rice Cakes, 8.5 oz (FPro: 0.7562) URL: [Product Link @ TrueFood.Tech]
12. Organic Lightly Salted Wild Rice Cakes, 8.5 oz (FPro: 0.7562) URL: [Product Link @ TrueFood.Tech]
13. Rice Cakes Blueberry & Beet, 1.4 oz (FPro: 0.7076) URL: [Product Link @ TrueFood.Tech]
14. Quaker Lightly Salted Gluten Free Rice Cakes - 4.47oz (FPro: 0.6802) URL: [Product Link @ TrueFood.Tech]
15. Organic Cinnamon Toast Rice Cakes, 9.5 oz (FPro: 0.6473) URL: [Product Link @ TrueFood.Tech]
16. Quaker Large Rice Cake Apple Cinn - 6.53oz (FPro: 0.6365) URL: [Product Link @ TrueFood.Tech]
17. Organic Salt Free Brown Rice Cakes, 8.5 oz (FPro: 0.5182) URL: [Product Link @ TrueFood.Tech]

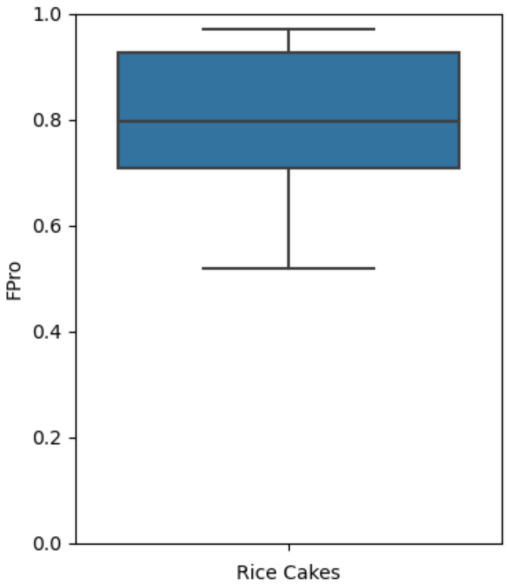


Figure S20: Distribution of FPro for all rice cakes in in GroceryDB (total 17).

References

[1] FDA Substances Added to Food. https://www.cfsanappsexternal.fda.gov/scripts/fdcc/?set=FoodSubstancesl. 2021 (accessed November 1, 2021).

[2] Igoe, R. S. *Dictionary of food ingredients* (Springer Science & Business Media, 2011).

[3] Ahuja, J. K. *et al.* Ingid: A framework for parsing and systematic reporting of ingredients used in commercially packaged foods. *Journal of Food Composition and Analysis* **100**, 103920 (2021). URL https://www.sciencedirect.com/science/article/pii/S0889157521001204.

[4] Data crunch report: The impact of bad data on profits and customer service in the uk grocery industry. GS1 UK and Cranfield University School of Management. https://dspace.lib.cranfield.ac.uk/bitstream/handle/1826/4135/Data_crunch_report.pdf (2009). (accessed April 4, 2022).

[5] FDA Label Guide. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-industry-food-labeling-guide. 2021 (accessed October, 2021).

[6] Menichetti, G., Ravandi, B., Mozaffarian, D. & Barabási, A.-L. Machine learning prediction of the degree of food processing. *Nature Communications* **14**, 2312 (2023).

[7] FDA Nutrition Facts. https://www.fda.gov/food/nutrition-education-resources-materials/new-nutrition-facts-label. 2021 (accessed November 1, 2021).

[8] Ma, P. *et al.* Application of machine learning for estimating label nutrients using usda global branded food products database, (bfpd). *Journal of Food Composition and Analysis* **100**, 103857 (2021). URL https://www.sciencedirect.com/science/article/pii/S0889157521000570.

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

[10] Huber, P. J. Robust regression: asymptotics, conjectures and monte carlo. *The annals of statistics* 799–821 (1973).

[11] Croux, C. & Rousseeuw, P. J. Time-efficient algorithms for two highly robust estimators of scale. In *Computational statistics*, 411–428 (Springer, 1992).

[12] Guidance for industry: Food labeling guide. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-industry-food-labeling-guide. 2021 (accessed Nov 1, 2021).

[13] FDA Substances Added to Food. `https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?CFRPart=172`. 2021 (accessed November 1, 2021).