HEALS Grant Progress Update

Leveraging AI, supply chain analytics, empirical methods and low-cost sensing of food contact chemicals to enhance assessment and mitigation of health risks emerging from processed foods.

Central Research Problem & Importance

The central problem of our research is the unknown or poorly understood risks that processed foods may pose to human health. Beyond the well documented risks tied to processed foods' high levels of salt, fats, and sugars, and the hyper-palatability that makes them easy to over-consume, processed foods may present further health risks via food additives and contaminants which are not yet well documented. Moreover, today's industrialized food production systems present myriad opportunities for food contact chemicals (FCCs) to contaminate food, from pesticides and herbicides used in agriculture to the materials (e.g., plastics) that make up food packaging. However, there are currently only sparse data that connect consumption patterns, food product ingredients and the corresponding supply chain and manufacturing risk drivers, which makes most risk assessments anecdotal by nature. Moreover, existing food safety testing methodologies pose a large expense to food producers, making universal testing cost prohibitive.

Our project aims to develop novel data driven models as well as new testing technologies and sensors to circumvent the limitations to traditional nutrition and food safety research, address related knowledge gaps, and inform the formulation and implementation of regulatory policies and industry practices related to risk assessment and mitigation, including testing procedures. Understanding the health risks of processed food ingredients and contaminants is particularly important because most of the foods purchased by Americans today are processed, and this increasingly processed diet is recognized to be associated with a rise in many chronic diseases. Better assessment of exposure and risk drivers as well as improved understanding of the mechanisms by which processed foods cause harm will enable food regulatory policies and industry practices to better identify and mitigate these risks and protect public health.

Progress Updates

Our team has made the following progress across each of our 3 thrusts of work:

<u>Thrust 1: Developing new AI-enabled, large-scale datasets for branded (packaged) food ingredients</u>

The goal of this thrust is to acquire food product ingredient data for all or most products currently on the market in the U.S., and build a searchable database that allows for the interfacing of those ingredient data with the consumer purchasing data described in Thrust 2 below, in order to run large-scale analyses of what ingredients the U.S. population is consuming.

At present, the only publicly available dataset of packaged food ingredients is the Global Branded Food Products Database (GBFPD) offered by U.S. Department of Agriculture. However, GBFPD is not comprehensive, as for some years it contains as little as 1/3 of the food products purchased.

To build a more comprehensive data set, we have been working to establish a multi-step pipeline. We have developed an ETL (Extract, Transform, Load) pipeline to scrape food label information from various web sources, including food manufacturer websites and grocery chain websites. So far, we have scraped over two dozen different retailer sites, bringing the database to a total of 506,255 records.

We have also developed tools to extract and process scraped label information. A major impediment to large-scale ingredient analysis of branded food products is that manufacturers often provide ingredient information exclusively in image form. To overcome this barrier and enable automated, systematic data collection at scale, we developed a pipeline combining Optical Character Recognition (OCR), Large Language Models (LLM), and fuzzy matching algorithms. The workflow operates in two integrated stages: (1) We have developed an OCR + LLM workflow to first extract text from the image using OCR and improve the extraction quality using LLM. (2) We developed an LLM + fuzzy matching strategy to standardize the components in the ingredient list. LLM removes unnecessary descriptive terms first, then we use a fuzzy matching algorithm to compare the remaining components with a dictionary of common additives and convert all different aliases of the same component into a unique scientific name. We validated this OCR pipeline on a test set of 145 scraped records (~1,000 images) from Target, achieving an F1 score of 0.811. This approach consolidates the thousands of colloquial names, regional variants, and trade names for individual components into harmonized scientific identifiers, enabling reproducible cross-product and cross-brand comparisons essential for our health impact assessment models. The steps involved in the scraped image processing pipeline are illustrated in **Figure 1** below.

Figure 1. Diagram of the process flow of the pipeline for processing ingredient data from scraped web images.

Thrust 2: Developing new health impact assessment methodologies

The goal of this thrust is to allow for the combination of the food label data acquired in Thrust 1 above with food consumption data, allowing for large scale analyses of what ingredients the U.S. population is consuming.

NielsenIQ published a set of Consumer Panel data records of all branded food purchased by 40K – 60K representative U.S. families over the span of 20 years. However, it is challenging to link the entries for the same products across the Nielsen data and the ingredient data set described in Thrust 1 because of erroneous Universal Product Codes (UPCs), private product codes, and limited product descriptions. We have developed a hybrid strategy integrating UPC fuzzy matching and Bert-based Text similarity matching using product descriptions, brand names, and package weight to match food products in the food consumption dataset to their corresponding labels.

Even with the construction of the comprehensive database still underway, we can already conduct meaningful analyses. Our preliminary correlation analysis shows that among single-member households in the U.S. who purchase food products with at least one of the six color additives that the FDA is working to phase out (Red 40, Yellow 5, Yellow 6, Blue 1, Blue 2 and Green 3), males spend a higher percentage of their grocery budget on food products with at least one of the dyes compared to females, high-income people spend less on products with any of the dyes than low-income people, and people with higher education level spend less on products with any of the dyes than those with lower education, and people in north-east and west regions spend less on products with any of the dyes than those in other regions.

Thrust 3: Supply chain and production process sensing for FCCs

The goal of this thrust is to develop ultra-low cost, sensors that can detect common contaminants, which will allow manufacturers to test their products significantly more frequently.

To select a starting target for sensor development, the team began by conducting background research on the current state of food contaminant regulation in the U.S. We found that heavy metals are a main focus of FDA's work around food contamination. We selected arsenic as a starting point due to its relative prevalence in foods as well as the health risks it poses, especially to babies and young children, who are at higher risk for the adverse health effects of heavy metals. Arsenic is also centered in the federal government's food safety agenda: arsenic is one of the few contaminants for which FDA has published guidance for food products, further indicating relatively higher possible risk of contamination.

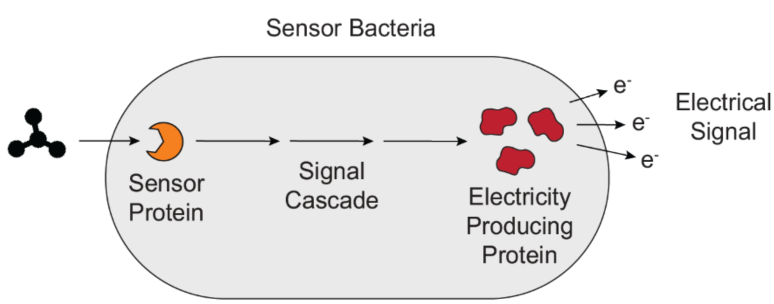
Once we selected arsenic as a starting point, we conducted further research to understand its mechanism of introduction into foods as well as what foods pose the highest risk of arsenic contamination. We found that rice and rice-containing products present the highest risk of arsenic contamination, because of the flooded environment in which it is grown. From there, we conducted further research to understand both where rice consumed in the U.S. is produced as well as where arsenic contamination of crops is a problem. We found that the U.S. grows 75% of its own rice, and that there are two main rice growing regions in the country: California and a cluster of states in the Southern part of the country including Arkansas, Louisiana, Mississippi, Missouri, and Texas. Our next step in this vein will be to combine rice production location data with arsenic contamination location data to evaluate any particular hot spots for rice production that may have high risk of arsenic contamination.

In parallel, the team began developing an inexpensive and easy to use arsenic sensor by combining RFID and engineered microbe technologies. Radio frequency identification (RFID) tags have emerged as a promising alternative platform for environmental monitoring. These devices are easily fabricated from low-cost materials and can be wirelessly operated by user-friendly devices including smartphones, making them effective for large scale point-of-need sensing. However, existing RFID sensing devices have substantial limitations which greatly restrict their applications. Smartphone-readable RFID sensors are incompatible with liquid samples, largely limiting them to gas phase detection. Such sensors also suffer from poor specificity because of their reliance on chemo responsive conductive materials to detect analytes.

In contrast to chemical sensing, microbes can sense nearly any small molecule in their environment with high selectivity and specificity. Engineered microbes utilize proteins or nucleic acids to recognize analytes with high specificity and transduce this molecular detection event into a measurable signal. Microbial sensors are attractive for low cost, point-of-need sensing because they completely self-assemble and tolerate operating conditions which damage or inactivate other sensing elements. A class of organisms known as electroactive microbes (EAMs) are of particular interest for sensing because they can generate electrical signals, enabling them to communicate directly with electronic devices. EAMs and other microbes have been utilized in a variety of sensor devices to date; however, existing sensor designs rely on expensive and/or operationally complex equipment for measurement, making them unsuitable for environmental monitoring.

In this work, we report the design of a microbe-controlled RFID sensor which can be wirelessly operated by a smartphone. Two thirds of the world population have access to a smartphone, making digital measurement of diverse molecular targets accessible to a previously unprecedented number of people. To overcome the previously limited sensing capabilities of RFID devices, we have developed a signal transduction strategy to interface an RFID circuit with electroactive sensor microbes. We enable RFID tags to perform liquid phase sensing through coupling to an external DC circuit which is compatible with liquid systems. We then place this DC circuit under control of electroactive sensor microbes by developing an electrochemical generator-collector cell to amplify microbial signals.

During the first phase of HEALS, we have successfully identified genes for arsenic sensing, shown in Figure 2 below.



Analyte	Sensor Protein
Sulfate	CysB
Arsenic	ArsR
Selenium	SelB

Figure 2. Diagram of an arsenic-sensing bacteria.

Our next steps include evaluation of intrinsic arsenic levels in three rice-containing products by Inductively Coupled Plasma Mass Spectrometry (ICP-MS) as an initial determination of contamination prior to field deployment of the sensors.

To determine the rice-containing products for testing, we used the existing prototype of our database to calculate the total sales of each product in Year 2023 using Nielsen Consumer Panel data. In keeping with our focus of the particular risk of arsenic contamination in products consumed by babies and children, we generated a list of the top 50 food products that have most rice exposure to families with <12 year-old children. We found that most of the products among the top 50 are rice cereal and instant rice, and so will be selecting products with the highest proportion of rice from those categories.

Translation Into Real-World Impact

Our work will translate into multiple streams of real-world impact. Our searchable database will enable a new generation of research on health outcomes associated with individual food ingredients. The resulting findings will translate into the prevention of health risks via stronger food production policy, improved practices, and improved consumer education. For example, the database could play an instrumental role in supporting FDA’s efforts to phase petroleum-based synthetic dyes out of the food supply. The database could also support similar efforts with other food additives by allowing for much more rigorous research on their potential health harms than is possible at present.

Our easy-to-use, low-cost food contaminant sensors will radically alter the landscape of food safety testing, allowing producers to test every single batch of their products rather than spot checking because universal testing is currently cost-prohibitive. This will result in a much lower risk of contaminants in food making it all the way to consumers and preventing associated adverse health impacts.

Primary Beneficiaries/End Users

The end users of our searchable database of information on packaged foods and beverages will primarily be researchers in nutrition and food safety, as well as consumers as the dataset will be publicly available. The targeted end users of our food contaminant sensors are food safety staff at packaged food production companies.

The primary beneficiaries of our work are packaged food consumers, which includes the vast majority of the U.S. population. Additional beneficiaries include food supply chain actors, who will benefit from the introduction of affordable and easy to use sensors for common food contaminants, nutrition researchers, who will benefit from having access to a rich, comprehensive, searchable database of packaged food ingredient and nutrition information, and food safety and nutrition policy makers and regulatory bodies, who will benefit from having clear and rigorous research on which to base their rules and regulations.

Collaboration & Partnership

HEALS support has allowed us to collaborate across disciplines at MIT. By bringing together the FSAS team and the Furst lab, the HEALS grant has allowed us to take a much more comprehensive view of the health risks associated with packaged foods by researching both known ingredients that may pose yet-unknown health risks, as well as the contaminants that may unknowingly be present in packaged foods and bring along their own host of potential health harms. The team meets bi-weekly to synthesize findings.

Next Steps & Milestones

<u>Thrust 1: Developing new AI-enabled, large-scale datasets for branded (packaged) food ingredients</u>

- Complete the supply chain analysis and market-exposure of rice-based products and risk analysis of penetration of arsenic into the supply chain.
- Complete the supply chain analysis and market-exposure of orange juice products and risk analysis of penetration of arsenic into the supply chain.

<u>Thrust 2: Developing new health impact assessment methodologies</u>

- Advance the matching of consumption data to label data to over 80% with focus on selected grocery chains and product categories (e.g., bread & bakery) to enable large scale analysis of exposure to unhealthy ingredients.

<u>Thrust 3: Supply chain and production process sensing for FCCs</u>

- Evaluation of intrinsic arsenic levels in three rice-containing products by ICP-MS as an initial determination of contamination prior to field deployment of the sensors.
- Overlay rice production location data with arsenic contamination location data and analyze the results.