Journal of Food Engineering 424 (2027) 113327

Contents lists available at ScienceDirect

Journal of Food Engineering

journal homepage: www.elsevier.com/locate/jfoodeng

Agentic AI for food property data: From literature to database to new discovery

Xu Zhou [a,*], Yi Zhang [a], Shyam S. Sablani [b], Shaojin Wang [b,c], Juming Tang [d]

[a] Department of Food Science, The Pennsylvania State University, University Park, PA, 16802, USA
[b] Department of Biological Systems Engineering, Washington State University, Pullman, WA, 99164, USA
[c] College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling, Shaanxi, 712100, China
[d] Department of Industrial & Systems Engineering, University of Washington, Seattle, WA, 98195, USA

ARTICLE INFO

Keywords:
AI agents
Large language models
Food database
Dielectric properties
Scientific data mining

ABSTRACT

Food property data are important for product development and process design and, more recently, for AI model training, but most data are scattered in literature. Building food property databases manually is time-consuming and difficult to scale. We developed AgentXtract, an agentic AI workflow that automatically transforms scientific papers into databases. Using dielectric properties as a case study, three specialized large language model (LLM) agents collaborate in a coordinated workflow, with different roles in paper screening and interpretation, data extraction, and record standardization and assembly. Ten LLMs from OpenAI, Google, and Anthropic were benchmarked against an expert-curated reference dataset. All models achieved precision above 95% and value accuracy above 99%, but recall, API costs, and processing times varied. The best-performing model was then selected to construct the full database. AgentXtract generated 4276 records for 142 food materials in approximately 30 min with an API cost of $9. Analysis of the compiled database revealed new relationships between dielectric constant and loss factor that have not been reported previously. The resulting database also served as the knowledge source for a prototype retrieval-augmented generation (RAG) AI model. Users can ask questions about dielectric properties and microwave heating in plain language, and RAG searches the database to find relevant data and provides answers with references. This study demonstrates that AI can efficiently and accurately transform fragmented food literature into structured databases and also support data-driven discovery in food engineering.

1. Introduction

Food properties are the measurable characteristics of food related to heat transfer (thermal properties), mass transfer (transport properties), electromagnetic heating (dielectric properties), texture and structure (rheological and mechanical properties), chemical changes (reaction kinetics), and microbial safety (microbial kinetics) (Ling et al., 2015; Peng et al., 2017; Rao et al., 2014; Tang, 2005). The properties determine quality, safety, and sensory experience and are critical information used in product development and process design (Rao et al., 2014). Recent advances in artificial intelligence (AI) and computational modeling have raised the demand for large, accurate, structured, and machine-readable food property datasets. However, most food property data are scattered across thousands of individual publications, where measurements are reported for specific materials under different experimental conditions in varying formats. No centralized and machine-readable database exists for most food properties.

Manual extraction of food property data from the literature is time-consuming, error-prone, and difficult to scale. For each paper, a researcher needs to locate relevant data, interpret experimental conditions, and transfer reported values into a structured database. Data are reported in diverse formats, such as text, tables with different column layouts, and figures with different plot types. This process needs to be repeated for every paper and updated as new studies are published. Building a high-quality database also requires domain expertise. For example, collecting dielectric property data requires knowledge of electromagnetic theory and transmission line principles (Nelson, 1991; Tang, 2005). Such expertise is necessary to determine whether the instruments were properly calibrated, whether the measurements were conducted correctly, and whether the reported values are useful for

* Corresponding author.
E-mail address: xbz5414@psu.edu (X. Zhou).

https://doi.org/10.1016/j.jfoodeng.2026.113327
Received 1 July 2026; Received in revised form 5 September 2026; Accepted 8 September 2026
Available online 12 September 2026

dielectric heating applications.

The automation of data extraction from scientific literature has been an active area of research. Earlier approaches largely depended on natural language processing (NLP) pipelines incorporating predefined rules and task-specific tuning. Such methods typically required extensive annotated training data and considerable customization for individual applications, limiting their generalizability and widespread adoption (Polak and Morgan, 2024). Large language models (LLMs) have now opened new possibilities. AI chatbots, such as ChatGPT (Singh et al., 2026) and Google Gemini, can process user-provided information, including uploaded files, interpret documents, and extract structured information. However, AI chat systems need manual file uploads and prompt-by-prompt interaction, which limits scalability and reproducibility. AI outputs may also depend on prior conversation history, memory, and user-specific context, making it harder to obtain consistent results across users and sessions. To address this issue, recent studies embedded LLMs into programmatic extraction pipelines. For example, ChatExtract (Polak and Morgan, 2024) showed that conversational LLMs with engineered prompts could extract material-value-unit triplets from materials science literature with precision and recall near 90%. ChemMiner (Chen et al., 2025) and ReactionSeek (Li et al., 2026) developed multi-agent and multi-stage workflows to handle coreference resolution and multimodal inputs in chemistry literature. Despite the advances, most existing frameworks have been developed primarily for reaction extraction or knowledge-triplet extraction, leaving many other scientific data extraction tasks insufficiently addressed. Food property mining is different from chemical reaction or triplet extraction because each record must associate numerical values with the relevant material and experimental conditions, information that is often distributed across multiple parts of a publication. As a result, the task requires measurement-level reconstruction and scientific validation rather than simple entity extraction.

To address this gap, we developed a multi-agent LLM system to build food-property databases from scientific literature. Dielectric properties were used as a case study because their manual extraction is particularly challenging: they depend on multiple experimental conditions, such as food material, frequency, temperature, moisture content, and other compositions, and are reported in heterogeneous formats, including tables, figures, text, and fitted equations. The objectives were to: (1) design and implement an agentic workflow for literature extraction, (2) evaluate system performance against an expert-curated reference dataset using multiple LLMs, and (3) demonstrate how a provenance-aware compiled database can support exploratory property analysis and source-cited data retrieval.

2. Materials and methods

2.1. Multi-agent system overview

An AI agent is a system that uses an LLM as its reasoning engine to autonomously plan, decide, and act toward a goal (Schick et al., 2023). A multi-agent system extends this concept by assigning specialized roles to multiple agents that collaborate and delegate to accomplish tasks too complex for any single agent (Lu et al., 2025; Wang et al., 2024). Our system consists of three LLM agents whose roles and instructions are defined through skills (Anthropic, 2025, see Section 2.7): Agent I converts PDFs into structured data, Agent II extracts values from structured data, and Agent III applies physics-based rules to validate, organize, and export values into a database. An orchestrator coordinates the three agents, routes papers through the pipeline concurrently, and tracks progress. Fig. 1 shows the system architecture.

2.2. Agent I: parse, screen, and classify

Agent I performed three tasks: document parsing, screening, and content classification. For parsing, papers in PDF format were converted into structured JSON using the open-source Python library Docling (Auer et al., 2024). For screening, the LLM read the title, abstract, keywords, and introduction objectives of each paper to determine whether the paper reported dielectric property data relevant to this study. For papers that passed screening, the LLM then read the full text and extracted paper-level metadata, including title, authors, year, journal, primary food materials, measurement frequencies, temperature range, and measurement method. It also classified each table as: (1) data (contains dielectric measurements for extraction), (2) equation (contains regression coefficients for polynomial evaluation), or (3) skip (contains penetration depth, compositional data, or other irrelevant content). Tables with ambiguous or complex layouts, based on LLM reasoning, were flagged as "complex tables" to trigger vision-based extraction in Agent II.

2.3. Agent II: extract

Agent II received the parsed documents from Agent I and performed

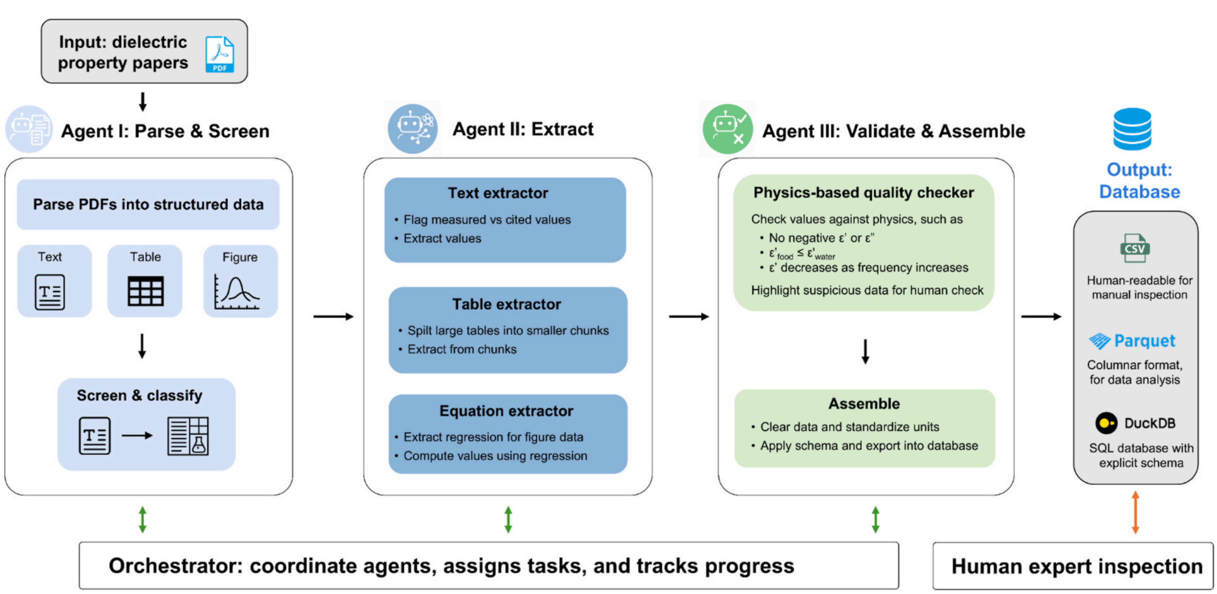

Fig. 1. Architecture of the multi-agent system for automated extraction of food dielectric properties from scientific literature.

data extraction through three specialized modules: table extraction, text extraction, and equation extraction.

2.3.1. Table extraction

Tables were the primary source of dielectric property records in literature. Preliminary tests found that submitting entire large tables to LLMs could cause hallucination, where the model generated plausible but nonexistent or wrong data. To address this, tables exceeding 12 rows were split into smaller chunks at condition-group boundaries (i.e., where frequency, temperature, or moisture content changed). Each chunk retained full table headers and captions to preserve context. The extraction prompt was implemented as a "skill" file that instructed the LLM to handle specific formats, including alternating $\varepsilon'/\varepsilon''$ rows that should be paired into single records, wide-format tables with multiple frequencies as column headers, and sectioned tables where material names appear as row dividers rather than in every cell. For tables that were incomplete or poorly parsed in text form, a vision-based fallback was used: the PDF page was rendered as a high-resolution image and sent to a vision-capable LLM (for example, GPT-5.3) for extraction.

2.3.2. Text extraction

For papers that reported data in running text, the Results and Discussion sections were processed by the LLM with a source discrimination skill. This skill instructed the model to extract only original measurements and to reject values from cited papers, identified by citation markers, attribution language, or frequencies and materials not listed in the paper's own experiments.

2.3.3. Equation extraction

Direct extraction of data from figures was not performed in this study. Our preliminary tests showed that even the most advanced vision LLMs (such as Claude Opus 4.6 and GPT-5.3, at the time of this study) produced inaccurate and unreliable results, likely due to overlapping curves, ambiguous markers (e.g., error bars), and absence of precise gridlines. Many dielectric property papers (Auksornsri et al., 2018; Coronel et al., 2008; Wang et al., 2008, 2009) report polynomial regression equations (e.g., $\varepsilon' = a_0 + a_1T + a_2T^2$) fitted to the data plotted in figures and usually include a table for fitted parameters. We therefore used an alternative approach: the LLM extracted polynomial coefficients, variable definitions, and applicable conditions (frequency, temperature, and moisture range) from these tables. The extracted equations were then used to compute discrete data points at common conditions within the reported range.

2.4. Agent III: validate and assemble

Agent III screened each dielectric property record using physics-based rules. For example, in high-moisture foods, ε'' is typically higher at radio frequency (RF) where ionic conduction dominates than at microwave frequencies where dipolar relaxation dominates. The rules were used to flag suspicious records for targeted human review rather than to automatically correct or reject reported measurements. The agent also removed exact duplicates, standardized units, and applied a consistent database schema (see Supplementary Material and Table S1). The resulting dataset was exported in CSV, Parquet, and DuckDB formats.

2.5. Data sources

Candidate papers were identified through literature searches and reference tracing from known studies on dielectric properties of foods and agricultural materials. Search terms included “dielectric properties”, “dielectric constant”, “loss factor”, “food”, “food materials”, “agricultural products”, “microwave”, and “dielectric heating”. About 200 candidate papers were first identified from the search results. Each paper was then screened using the following criteria: papers were included if they reported original experimental measurements of both dielectric constant (ε') and loss factor (ε'') for at least one food material. Review articles, non-peer-reviewed papers, non-English publications, and papers that reported dielectric-property data only in figures were excluded. After manual screening, 64 papers met the criteria and were included in the study (see Supplementary Information).

To evaluate LLM agent performance, the agent-extracted records needed to be compared with manually extracted reference records. Because manual extraction of all 64 papers would be time-consuming, 19 papers (~30%) were selected to construct a manually curated reference dataset (see Supplementary). The papers were selected to represent the diversity of the full dataset in food materials, frequencies, temperatures, measurement methods, reporting quality, and extraction formats, including tables, text-reported values, and regression equations. The reference dataset was curated by an author with more than 10 years of experience in dielectric-property research. The author read each paper, manually extracted the dielectric-property data, checked the experimental conditions, and organized the records into a structured database. This process took approximately one week, which also demonstrates that manual extraction of literature data is time-consuming even for a domain expert. Record matching between the agent-extracted dataset and the manually curated dataset was based on material name, frequency, and temperature. Before matching, material names were standardized by converting text to lowercase, removing punctuation and extra spaces, standardizing singular and plural forms, and allowing simple word-order differences. For example, “Fuji apple” and “apple, Fuji” were treated as the same material. Frequency values were considered matched when they were within 5% relative difference, and temperature values were considered matched when they were within ± 2 °C. Matching was performed within each source paper. Composition and sample conditions reported as part of the material name, such as salt level, fat level, concentration, or pH, were retained as part of the standardized material identity. After a record pair was matched, value accuracy was assessed by comparing the extracted ε' and ε'' values with the manually curated values using a 5% relative tolerance. The expert-curated reference dataset contained 1591 records in total and served as the ground truth for pipeline evaluation.

2.6. Evaluation and LLM benchmarking

The output of the agent system was compared to the expert-curated reference dataset using the following metrics:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (1)$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (2)$$

$$\text{F1} = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}) \quad (3)$$

where TP (true positives) were extracted records that matched the reference dataset, FP (false positives) were extracted records with no match in the reference dataset, and FN (false negatives) were reference records not recovered by the pipeline. Value accuracy was defined as the fraction of matched records where the extracted ε' and ε'' values were within 5% relative tolerance of the manually curated values.

To identify the most suitable LLM for this extraction task, ten models from three providers were benchmarked, representing a range of model sizes, costs, and capabilities: Claude Haiku 4.5, Sonnet 4.6, and Opus 4.6 (from Anthropic); GPT-4.1, GPT-4.1 Mini, GPT-5 Mini, and GPT-5.4 (from OpenAI); and Gemini 2.5 Flash, 3 Flash, and 3.1 Pro (from Google). Each model was evaluated using the same prompts, system infrastructure, and post-processing rules. Cost was calculated from actual token usage at published API pricing as of March 2026.

2.7. Prompt engineering, orchestration, and open-source implementation

The central design principle of this multi-agent system was the separation of domain knowledge from code logic. Each agent's behavior was governed by structured prompt files called "skills" (Anthropic, 2025), stored as Markdown documents (SKILL.md). Each skill encodes the domain knowledge and operational procedures needed for a particular extraction task, including extraction rules, output formats, edge-case handling, and quality-control validation logic. Because skills are plain text files independent of the pipeline code, researchers can adapt the pipeline to other data domains (e.g., thermal or microbiological properties) by modifying the skill files.

The orchestrator was implemented as a lightweight Python module. Each paper was submitted as an independent job that passed sequentially through the three agents, while multiple papers were processed concurrently through an asynchronous task queue. The orchestrator tracked job status, handled failures with automatic retries, and logged token usage for cost accounting.

3. Results and discussion

3.1. System performance and LLM benchmark

Ten LLMs were evaluated for dielectric property extraction performance (Table 1). We focused on two types of AI errors: (1) an LLM may hallucinate a value that does not exist in the original paper (a false positive), or (2) an LLM may miss a value that does exist (a false negative). All ten models achieved high extraction fidelity: precision $\geq$ 95%, and value accuracy $\geq$ 99% (with some models reaching 100%). But recall varied from 79% to 99%. The variation likely reflects how well each AI model handled complex table layouts and information distributed across multiple sections or pages. The F1 score, which balances precision and recall into a single metric, was used to rank overall performance. Six models achieved F1 above 0.95. AgentXtract achieved precision above 95% and an F1 score of up to 0.98, which are comparable to or higher than values reported for other extraction systems. For example, ChemMiner reported about 90% average precision and an F1 score of 0.83 for chemical reaction extraction from 17 annotated papers (Chen et al., 2025). ReactionSeek achieved over 95% precision and recall on the Organic Syntheses collection but targeted discrete reaction components, including reactants, products, and yields (Li et al., 2026).

LLMs can make mistakes. One failure observed in this study was table confusion. For example, Zhu et al. (2012) reported dielectric measurements in one table and statistical summary values (ε'' minima and their corresponding frequencies) in another; the LLM confused the two tables and extracted values from the wrong table. Similar document-level errors have been reported in other systems: ChemMiner found that coreference resolution across tables and figures was a major source of missed reactions (Chen et al., 2025), and ReactionSeek noted that compounds added in multiple portions were often aggregated incorrectly (Li et al., 2026). These examples confirm that LLMs can misinterpret document structure, so domain expert inspection is necessary. But human experts do not need to check every record. AgentXtract includes a post-extraction validation step that applies physics-based rules and LLM reasoning to flag suspicious values for humans to focus on (Fig. 1). Each record also carries traceable source information (paper title, authors, and journal), so any entry can be verified against the original publication.

API cost and processing speed are also practical considerations, particularly when proprietary large language models are used at scale. Both varied among models, but neither showed a clear correlation with F1 score. For example, the most expensive model (Claude Opus 4.6, \$1.08/paper) did not outperform the much cheaper Claude Haiku 4.5 (\$0.07/paper), which had the fastest processing time (21 s/paper) and high F1 (~0.98). GPT-5 Mini was the slowest model (97 s/paper) and had the lowest F1 score (~0.88). Considering recall, cost, and speed together, Claude Haiku 4.5 provided the best overall performance and therefore was selected for full-scale extraction.

It is important to note that the foundation language model landscape is evolving rapidly, with newer and more capable models released every few months. During this manuscript submission and review, several new models became available. The purpose of the comparison of different LLMs was not to identify a universally best model, but to demonstrate that a multi-agent workflow can reliably extract food property data from scientific literature. As newer models become available, they can be integrated into the same workflow for continued improvement.

3.2. Database overview

The multi-agent system extracted dielectric property data from 64 papers published in 13 journals between 2002 and 2026. The resulting database contained 4276 records for 142 different food materials (Supplementary Table S1; Fig. 2). The full extraction took approximately 30 min of machine extraction and \$9 in API cost. For comparison, manual extraction into a structured database would take an experienced graduate student an estimated 40 days, assuming two papers (or 100 data records) per day. AI agent systems can provide an efficient and cost-effective approach for scientific data curation.

Fig. 2a presents six selected frequency groups relevant to RF and microwave processing. Most records were at 915 MHz (1267 records, 32%), followed by 27 MHz (861, 22%), 2450 MHz (813, 20%), and 40 MHz (799, 18%). These are the main industrial, scientific, and medical (ISM) frequencies used in food processing (Tang, 2015; Zhou and Wang, 2019). The food materials were grouped into eight categories (Fig. 2b). Vegetables had the most data points (1,425, 33%), followed by grains (1,147, 27%), nuts (655, 15%), fruits (568, 13%), seafood (213, 5%), dairy (164, 4%), dried fruits (100, 2%), and meat (51, 1%). This also reflects research and industrial interest in RF and microwave energy for drying and blanching of fruits and vegetables (Ruiz-Ojeda and Peñas, 2013; Zhou and Wang, 2019), grain and nut disinfestation (Wang et al., 2010; Zhou and Wang, 2016), and non-destructive quality measurements (Lewis et al., 2019; Nelson,

Table 1
Performance, cost, and processing time of ten large language models for data extraction from literature.

#	Model	Provider	Value Accuracy (%)	Precision (%)	Recall (%)	F1	Cost (\$)[a]	Time (s)[a]
1	Claude Opus 4.6	Anthropic	99.7	97.3	98.5	0.98	1.08	38
2	GPT-5.4	OpenAI	99.7	98.0	97.4	0.98	0.15	37
3	Claude Haiku 4.5	Anthropic	99.8	98.1	97.3	0.98	0.07	21
4	Claude Sonnet 4.6	Anthropic	99.5	95.6	98.1	0.97	0.23	28
5	GPT-4.1 Mini	OpenAI	99.8	98.1	94.9	0.97	0.05	34
6	Gemini 3.1 Pro	Google	99.9	99.3	92.1	0.96	0.16	62
7	GPT-4.1	OpenAI	99.9	99.5	90.6	0.95	0.10	36
8	Gemini 3 Flash	Google	99.9	99.3	84.6	0.91	0.03	48
9	Gemini 2.5 Flash	Google	100	99.5	81.8	0.90	0.01	35
10	GPT-5 Mini	OpenAI	100	98.7	78.8	0.88	0.04	97

[a] Average cost or time per paper.

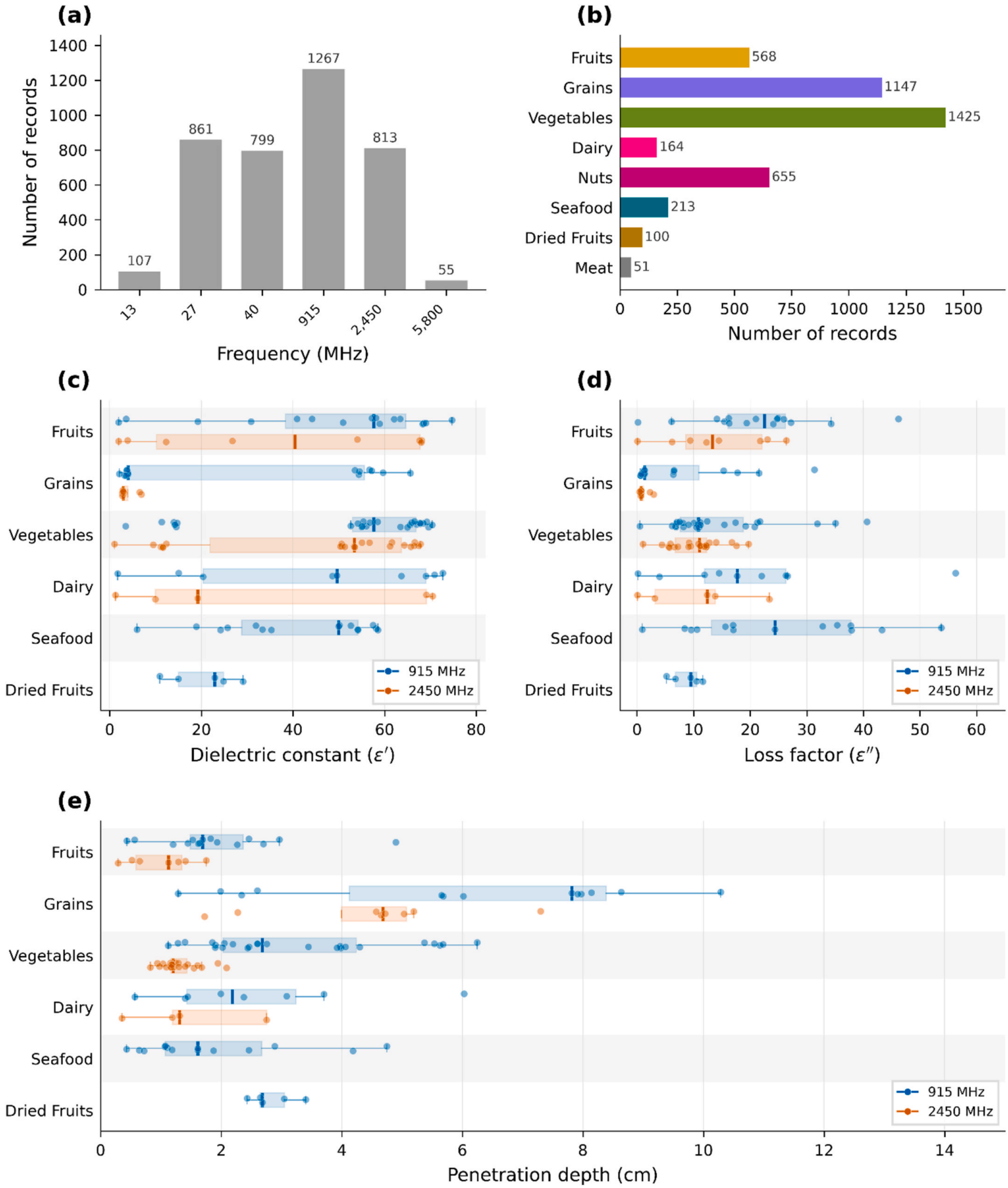

Fig. 2. Database overview. Distribution of extracted records by (a) measurement frequency and (b) food category, (c) dielectric constant (ε′), (d) loss factor (ε″), and (e) penetration depth of microwaves at 915 and 2450 MHz in food materials.

2008).

The database also shows the distributions of dielectric constant (ε′) and loss factor (ε″) for different food categories at 915 and 2450 MHz (Fig. 2c and d). At 915 MHz, high-moisture foods, such as fruits, vegetables, and seafood, had median ε′ values between 50.0 and 57.7, consistent with the strong contribution of free water to polarization (Gezahegn et al., 2021; Tang, 2005). Low-moisture categories had lower values, with median ε′ values of 22.9 for dried fruits and 4.0 for grains. The ε″ showed category-dependent frequency responses. Seafood had a high median ε″ of 24.4 at 915 MHz, while the median ε″ of dairy decreased from 17.7 at 915 MHz to 12.4 at 2450 MHz, consistent with the inverse dependence of ionic conduction loss on frequency (Nelson, 2008). Vegetables had similar median ε″ values at 915 and 2450 MHz (10.9 and 11.0, respectively), suggesting that dipolar relaxation loss at 2450 MHz partially offset the reduced ionic contribution (Tang, 2005). Fig. 2e compares penetration depths within the displayed range of 0–15 cm. Seafood had the shallowest median penetration depth at 915 MHz (1.6 cm), while grains had the greatest median penetration depth at both 915 MHz (7.8 cm) and 2450 MHz (4.7 cm). Penetration depth at 915 MHz was generally greater than that at 2450 MHz, consistent with the established advantage of 915 MHz for volumetric heating of thick or large food products (Tang, 2015).

This database provides a foundation for future physics-based computer simulations and data-driven modeling in RF and microwave food processing, including machine-learning models that predict dielectric properties from chemical composition. The same multi-agent workflow

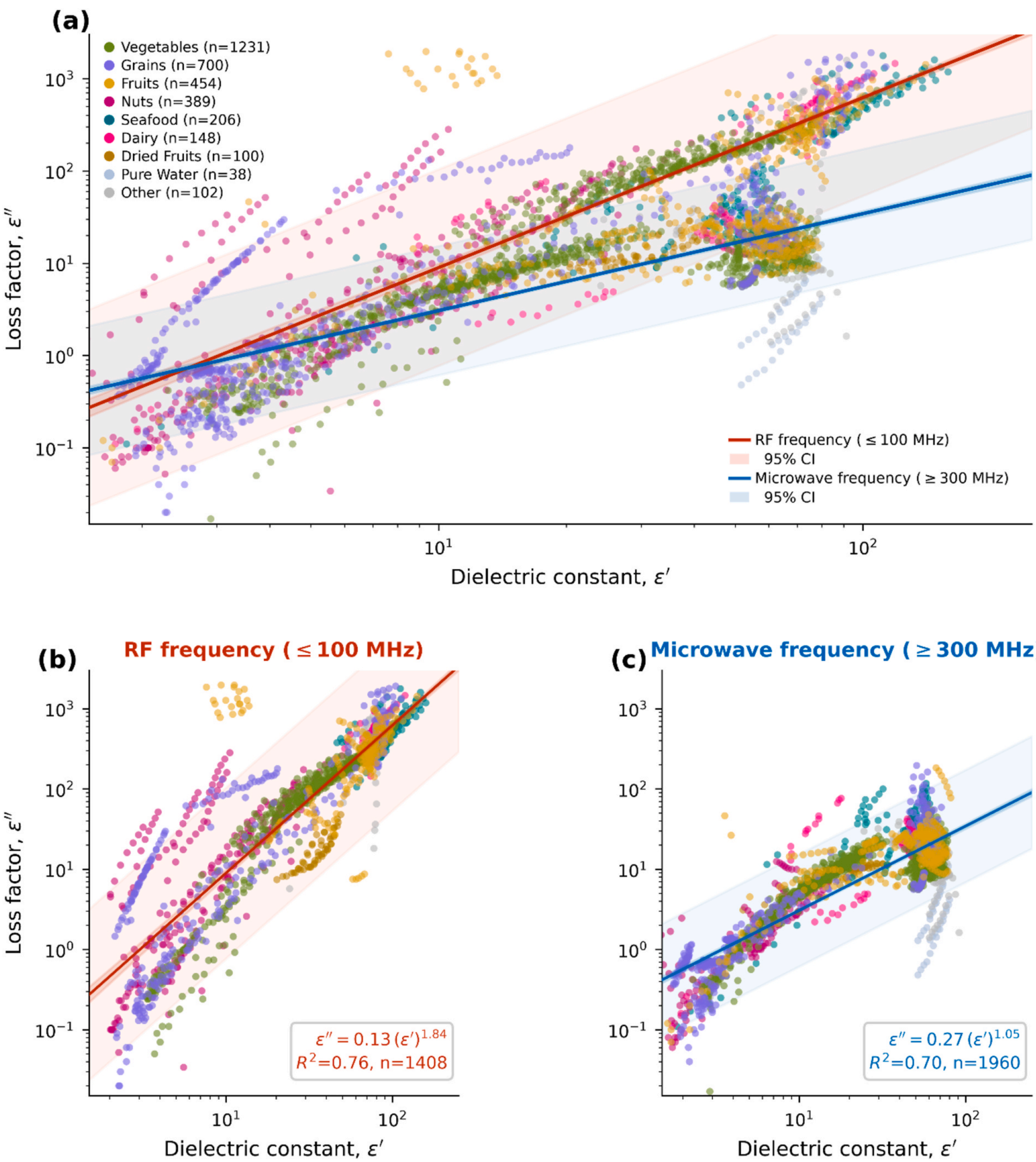

Fig. 3. Relationship between dielectric constant (ε′) and loss factor (ε″) of food materials at RF and microwave frequencies. (a) All data points, (b) RF frequency only, and (c) microwave frequency only.

can also be extended to other food properties (such as thermal, rheological, chemical, and sensory properties) that are scattered in decades of published literature.

3.3. Observed dielectric relationships and physical interpretations

The database revealed a new relationship between ε′ and ε″. Fig. 3 shows that the dielectric constant and loss factor followed power-law relationships over more than 2000 data points of about 100 food materials:

At RF frequencies (≤100 MHz), the following regression was obtained ($R^2 = 0.76$, n = 1408):

$$\varepsilon'' = 0.13(\varepsilon')^{1.84} \tag{4}$$

At microwave frequencies (≥300 MHz), the following regression was obtained ($R^2 = 0.70$, n = 1960):

$$\varepsilon'' = 0.27(\varepsilon')^{1.05} \tag{5}$$

Even though fruits, grains, vegetables, and other foods have different composition and structure, the overall relationship between dielectric constant ε′ and loss factor ε″ followed a similar trend in each category. (Fig. 3a and b). This observation can be explained with some theory as follows:

At microwave frequencies, dielectric loss is mainly governed by dipolar relaxation of water. In this process, polar molecules, especially water molecules, attempt to align with the alternating electric field. Because molecular rotation is not instantaneous, the polarization lags behind the applied field, and part of the electromagnetic energy is dissipated as heat. For an ideal single-relaxation system, this behavior can be described by the Debye model (Nelson, 1991):

$$\varepsilon' = \varepsilon_\infty + \frac{\Delta\varepsilon}{1 + (\omega\tau)^2} \tag{6}$$

$$\varepsilon'' = \frac{\Delta\varepsilon\omega\tau}{1 + (\omega\tau)^2} \tag{7}$$

where ε_∞ is the high-frequency limit of the dielectric constant. It represents the part of polarization that can respond almost instantaneously to the electric field, such as electronic and atomic polarization. Δε is the dielectric strength, defined as the difference between the static dielectric constant and ε∞. It reflects the magnitude of orientational polarization, or how strongly dipoles in the material can contribute to energy storage. The angular frequency is ω, and τ is the relaxation time, which represents the characteristic time required for dipoles to reorient after the electric field changes. For a given material at a fixed temperature and composition, ε∞, Δε, and τ are material parameters in the ideal Debye model, while ε′ and ε″ vary with frequency through the term ωτ.

From Eq. (6),

$$\Delta\varepsilon = (\varepsilon' - \varepsilon_\infty)\left[1 + (\omega\tau)^2\right] \tag{8}$$

Substituting Eq. (8) into Eq. (7) gives

$$\varepsilon'' = \omega\tau(\varepsilon' - \varepsilon_\infty) \tag{9}$$

Thus, when dipolar relaxation dominates, ε'' is approximately proportional to ε':

$$\varepsilon'' \propto \varepsilon' \tag{10}$$

This is consistent with the linear microwave regression in Eq. (5).

At RF frequencies, dielectric loss is often dominated by ionic conduction. Under this condition, the loss factor can be approximated as (Tang, 2005):

$$\varepsilon'' = \frac{\sigma}{\omega\varepsilon_0} \tag{11}$$

where σ is electrical conductivity of the food and ε_0 is the permittivity of free space.

Across the diverse food materials in this database, aqueous phase content is a major factor that influences both ε′ and ε″. The dielectric constant ε′ is governed mainly by polarization, and water has a much higher dielectric constant than the dry solid matrix. Therefore, foods with more aqueous phase generally have higher ε′ (Nelson, 1994). Electrical conductivity σ also tends to increase with aqueous phase content because dissolved ions require an aqueous medium for conduction. However, σ can increase more strongly than ε′ because higher moisture not only provides more conducting medium, but can also improve the connectivity of ion transport pathways through the food matrix. At RF frequencies, ε″ depends mainly on ionic conduction, whereas ε′ reflects dielectric polarization. As a result, ε″ can increase faster than ε′ as aqueous phase content increases. The fitted exponent of about 1.8 in Eq. (4) is consistent with the greater sensitivity of ionic conduction to the amount and connectivity of the aqueous phase. Some outliers were observed (Fig. 3b and c). At RF frequencies, white bread (Liu et al., 2009) showed loss factor values above the RF regression line. The porous structure of bread lowers the bulk dielectric constant ε′, while dissolved ions from salt and leavening agents can still produce high dielectric loss ε″. At microwave frequencies, pure water data (Gezahegn et al., 2021) deviated from the general food trend. This is expected because the loss factor of pure water arises entirely from dipolar relaxation, whereas most foods also contain dissolved ions that contribute additional conductive loss.

Previous studies and review papers have mainly reported dielectric data for individual foods or small groups of materials (Nelson, 2008; Sosa-Morales et al., 2010). In contrast, our AI-assisted compilation of more than 100 food materials made it possible to identify broader relationships that are difficult to recognize from isolated studies. Eqs. (4) and (5) also provide a simple way to estimate one dielectric property from the other, screen data for consistency, and identify unusual materials or measurements.

3.4. Applications of food dielectric database and RAG

To make the extracted data useable beyond a static file, we built a web application with two layers: a searchable database interface and a retrieval-augmented generation (RAG) model (Fig. 4). For the database, users can search for dielectric property data extracted from papers (Fig. 4a and b). Dielectric properties are measured to support process and equipment design. In RF and microwave processing, they influence penetration depth, volumetric heat generation, and temperature uniformity. Penetration depth is the distance at which incident electromagnetic power decreases to (1/e) of its value at the product surface (Tang, 2015). It therefore provides an important constraint on product geometry. Based on values in our database, mashed potato has a penetration depth of approximately 2.8 cm at 915 MHz, whereas salmon fillet has a smaller penetration depth of approximately 2.1 cm. A smaller penetration depth increases the risk of excessive heating near the edges and corners of slab-shaped products. Rounded or annular geometries may therefore provide more uniform heating than shapes with sharp corners (Buffler, 1993). For cylindrical and spherical products, a common rule of thumb is that electromagnetic focusing produces strong central heating when the product diameter is approximately 2 to 2.5 times penetration depth (Buffler, 1993). The database allows designers to compare the proposed product diameter with this range before the package geometry is finalized. The database can also help microwave system and process designers select an appropriate operating frequency. For many high-moisture foods, penetration depth is greater at 915 MHz than at 2450 MHz, which often makes 915 MHz more suitable for thick products. For example, at 25 °C, potato has a penetration depth of approximately 21 mm at 915 MHz and 9 mm at 2450 MHz. The data can also be used for computer simulation of microwave and RF heating

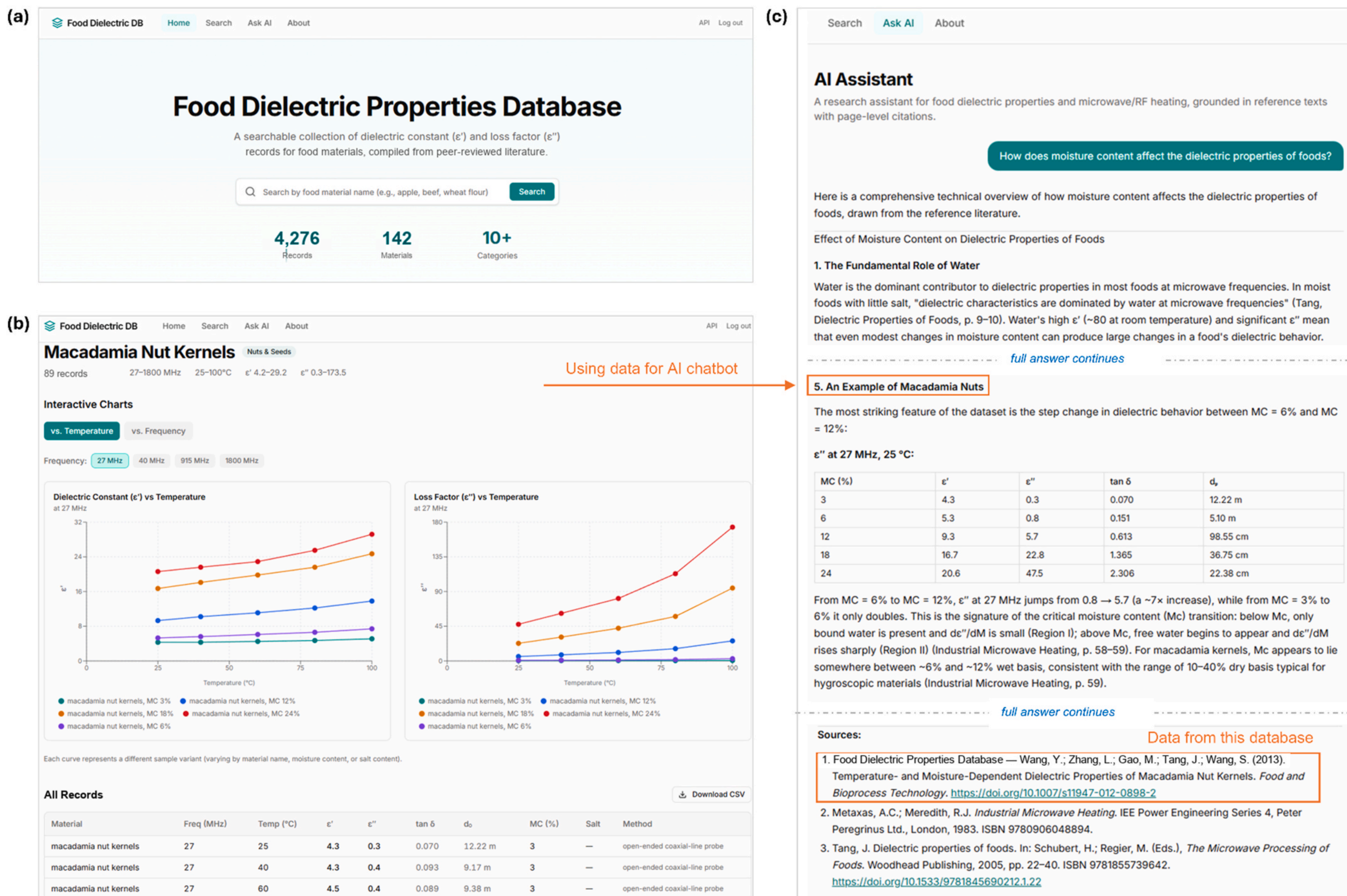

Fig. 4. Web application with retrieval-augmented generation (RAG). (a) Database homepage. (b) Example of dielectric property data extracted from a paper. (c) AI-generated answer to a user question with citations to the reference sources used.

(Zhou et al., 2025).

The second layer is the RAG AI model. RAG combines a language model with a retrieval step (Lewis et al., 2020). When a user asks a question, the system first searches the database for relevant records and source text. The language model then uses this information to generate an answer instead of relying only on knowledge learned during model pretraining (Lewis et al., 2020). For example, when asked how dielectric properties change with moisture content, the AI assistant may use general scientific knowledge learned during pretraining, but more importantly, it retrieves relevant passages, illustrates the trend using macadamia nut data from the database, and cites the source pages (Fig. 4). This design addresses the main weakness of general-purpose language models such as ChatGPT and Claude: their answers may be unsupported or incorrect. Our RAG ties each answer to our database and allows the user to verify it. A food process engineer can therefore ask a question in plain language and receive a cited, quantitative answer without writing database queries or reviewing many papers.

3.5. Broader value of AI and suggestions for future studies

AgentXtract is a general framework for building scientific databases from literature. Similar agents can be used to extract, validate, and organize other scattered food property data, such as microbiological, thermal, rheological, chemical, and sensory properties. AgentXtract also provides an example of how human experts can work with AI agents in research. In this study, human experts defined research questions, selected reliable sources, set inclusion and exclusion criteria, and evaluated data quality. The AI agents then handled the time-consuming steps of reading papers, extracting records, checking consistency, and exporting data into useable formats. Human experts reviewed uncertain records and made final decisions when the extracted information was ambiguous or incomplete. This division of labor allowed the database to be built more efficiently while keeping scientific judgment under human expert control.

4. Conclusions

This study developed AgentXtract, a multi-agent LLM pipeline for extracting, validating, and organizing food property data from scientific literature. Using dielectric properties as a case study, AgentXtract combined specialized agents for different tasks. Ten LLMs were benchmarked against a manually curated reference dataset, and their performance varied. With the selected model, AgentXtract extracted more than 4000 data points in about 30 min at an API cost of $9. The results show that multi-agent LLM systems can convert scattered literature data into structured scientific databases with high accuracy, low cost, and much less manual effort.

The database also produced new scientific findings. Analysis of the extracted records revealed power-law relationships between dielectric constant and loss factor. Such relationships are difficult to identify from individual studies because each paper typically reports a limited number of materials, frequencies, and temperatures. This work also demonstrates how scientific data extracted from literature can serve as a knowledge base for AI using RAG.

It should be noted that the value of an AI-assisted scientific database depends not only on extraction speed and accuracy but also on data quality. AgentXtract was designed to automate data extraction, not to replace scientific judgment. In this study, domain experts selected source papers based on experimental rigor, measurement methods, and data relevance. LLM agents could, in principle, screen and select thousands of papers, but we deliberately did not automate this step because source selection determines what enters the database and should remain with researchers who understand the domain. The value of AI is not to accumulate large volumes of unvetted data but to help researchers build targeted, high-quality databases that can support future predictive modeling and other AI applications in food engineering.

CRediT authorship contribution statement

Xu Zhou: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Yi Zhang:** Writing – review & editing. **Shyam S. Sablani:** Writing – review & editing. **Shaojin Wang:** Writing – review & editing. **Juming Tang:** Writing – review & editing.

Declaration of competing interest

To the best of our knowledge, the named authors have no conflict of interest, financial or otherwise.

Acknowledgments

Xu Zhou acknowledges support from start-up funding provided by Penn State University.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jfoodeng.2026.113327.

Data availability

The selected papers and extracted database are provided in the Supplementary Information. The complete source code for the AI agents is available at https://github.com/zhouxu-lab/agentXtract.

References

Anthropic, 2025. Agent Skills Overview. https://platform.claude.com/docs/en/agents-and-tools/agent-skills/overview. Access April, 2026.

Auer, C., Lysak, M., Nassar, A., Dolfi, M., Livathinos, N., Vagenas, P., Ramis, C.B., Omenetti, M., Lindlbauer, F., Dinkla, K., Mishra, L., Kim, Y., Gupta, S., de Lima, R.T., Weber, V., Morin, L., Meijer, I., Kuropiatnyk, V., Staar, P.W.J., 2024. Docling technical report. arXiv. https://doi.org/10.48550/arXiv.2408.09869.

Auksornsri, T., Tang, J., Tang, Z., Lin, H., Songsermpong, S., 2018. Dielectric properties of rice model food systems relevant to microwave sterilization process. Innov. Food Sci. Emerg. Technol. 45, 98–105. https://doi.org/10.1016/j.ifset.2017.09.002.

Buffler, R.C., 1993. Microwave Cooking and Processing. Springer US, New York: Van Nostrand Reinhold.

Chen, K., Du, Y., Li, J., Cao, H., Guo, M., Dang, X., Li, L., Qiu, J., Chen, G., Heng, P.A., 2025. ChemMiner: a large language model agent system for chemical literature data mining. In: 2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 7654–7662. https://doi.org/10.1109/ICCVW69036.2025.00787.

Coronel, P., Simunovic, J., Sandeep, K.P., Kumar, P., 2008. Dielectric properties of pumpable food materials at 915 MHz. Int. J. Food Prop. 11 (3), 508–518. https://doi.org/10.1080/10942910701472755.

Gezahegn, Y.A., Tang, J., Sablani, S.S., Pedrow, P.D., Hong, Y.K., Lin, H., Tang, Z., 2021. Dielectric properties of water relevant to microwave assisted thermal pasteurization and sterilization of packaged foods. Innov. Food Sci. Emerg. Technol. 74, 102837. https://doi.org/10.1016/j.ifset.2021.102837.

Lewis, M.A., Trabelsi, S., Nelson, S.O., 2019. Development of an eighth-scale grain drying system with real-time microwave monitoring of moisture content. Appl. Eng. Agric. 35 (5), 767–774. https://doi.org/10.13031/aea.13130.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D., 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474. NeurIPS 2020.

Li, J., Li, M., Yang, Q., et al., 2026. ReactionSeek: LLM-powered literature data mining and knowledge discovery in organic synthesis. Nat. Commun. https://doi.org/10.1038/s41467-026-70180-1.

Ling, B., Tang, J., Kong, F., Mitcham, E.J., Wang, S., 2015. Kinetics of food quality changes during thermal processing: a review. Food Bioprocess Technol. 8 (2), 343–358. https://doi.org/10.1007/s11947-014-1398-3.

Liu, Y., Tang, J., Mao, Z., 2009. Analysis of bread dielectric properties using mixture equations. J. Food Eng. 93 (1), 72–79. https://doi.org/10.1016/j.jfoodeng.2008.12.032.

Lu, Y., Wu, W., Zhao, X., Peng, R., Wang, J., 2025. KARMA: leveraging multi-agent LLMs for automated knowledge graph enrichment. arXiv. https://doi.org/10.48550/arXiv.2502.06472.

Nelson, S.O., 1991. Measurement and use of dielectric properties of agricultural products. IEEE Instrumentation and Measurement Technology Conference, pp. 636–640. https://doi.org/10.1109/IMTC.1991.161675.

Nelson, S.O., 1994. Measurement of microwave dielectric properties of particulate materials. J. Food Eng. 21 (3), 365–384. https://doi.org/10.1016/0260-8774(94)90080-9.

Nelson, S.O., 2008. Dielectric properties of agricultural products and some applications. Res. Agric. Eng. 54 (2), 104–112. https://doi.org/10.17221/1000-RAE.

Peng, J., Tang, J., Barrett, D.M., Sablani, S.S., Anderson, N., Powers, J.R., 2017. Thermal pasteurization of ready-to-eat foods and vegetables: critical factors for process design and effects on quality. Crit. Rev. Food Sci. Nutr. 57 (14), 2970–2995. https://doi.org/10.1080/10408398.2015.1082126.

Polak, M.P., Morgan, D., 2024. Extracting accurate materials data from research papers with conversational language models and prompt engineering. Nat. Commun. 15 (1), 1569. https://doi.org/10.1038/s41467-024-45914-8.

Rao, M.A., Rizvi, S.S.H., Datta, A.K., Ahmed, J. (Eds.), 2014. Engineering Properties of Foods, fourth ed. CRC Press. https://doi.org/10.1201/9781420028805.

Ruiz-Ojeda, L.M., Peñas, F.J., 2013. Comparison study of conventional hot-water and microwave blanching on quality of green beans. Innov. Food Sci. Emerg. Technol. 20, 191–197. https://doi.org/10.1016/j.ifset.2013.09.009.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., Scialom, T., 2023. Toolformer: language models can teach themselves to use tools. arXiv. https://doi.org/10.48550/arXiv.2302.04761.

Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., Nathan, A., Luo, A., Helyar, A., Madry, A., Efremov, A., Spyra, A., Baker-Whitcomb, A., Beutel, A., Karpenko, A., et al., 2026. OpenAI GPT-5 system card. arXiv. https://doi.org/10.48550/arXiv.2601.03267.

Sosa-Morales, M.E., Valerio-Junco, L., López-Malo, A., García, H.S., 2010. Dielectric properties of foods: reported data in the 21st century and their potential applications. LWT Food Sci. Technol. (Lebensm.-Wiss.-Technol.) 43 (8), 1169–1179. https://doi.org/10.1016/j.lwt.2010.03.017.

Tang, J., 2005. Dielectric properties of foods. In: Schubert, H., Regier, M. (Eds.), The Microwave Processing of Foods. Woodhead Publishing, pp. 22–40. https://doi.org/10.1533/9781845690212.1.22.

Tang, J., 2015. Unlocking potentials of microwaves for food safety and quality. J. Food Sci. 80 (8), E1776–E1793. https://doi.org/10.1111/1750-3841.12959.

Wang, J., Tang, J., Wang, Y., Swanson, B., 2009. Dielectric properties of egg whites and whole eggs as influenced by thermal treatments. LWT–Food Sci. Technol. 42 (7), 1204–1212. https://doi.org/10.1016/j.lwt.2009.02.018.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W.X., Wei, Z., Wen, J., 2024. A survey on large language model based autonomous agents. Front. Comput. Sci. 18 (6), 186345. https://doi.org/10.1007/s11704-024-40231-1.

Wang, S., Tiwari, G., Jiao, S., Johnson, J.A., Tang, J., 2010. Developing postharvest disinfestation treatments for legumes using radio frequency energy. Biosyst. Eng. 105 (3), 341–349. https://doi.org/10.1016/j.biosystemseng.2009.12.003.

Wang, Y., Tang, J., Rasco, B., Kong, F., Wang, S., 2008. Dielectric properties of salmon fillets as a function of temperature and composition. J. Food Eng. 87 (2), 236–246. https://doi.org/10.1016/j.jfoodeng.2007.11.034.

Zhou, L., Wang, S., 2016. Industrial-scale radio frequency treatments to control *Sitophilus oryzae* in rough, brown, and milled rice. J. Stored Prod. Res. 68, 9–18. https://doi.org/10.1016/j.jspr.2016.03.002.

Zhou, X., Lin, H., Wu, C.Y., Sablani, S.S., Tang, J., 2025. A new chemical marker method for determining heating patterns in microwave pasteurization. J. Food Eng. 397, 112593.

Zhou, X., Wang, S., 2019. Recent developments in radio frequency drying of food and agricultural products: a review. Dry. Technol. 37 (3), 271–286. https://doi.org/10.1080/07373937.2018.1452255.

Zhu, X., Guo, W., Wu, X., 2012. Frequency-and temperature-dependent dielectric properties of fruit juices associated with pasteurization by dielectric heating. J. Food Eng. 109 (2), 258–266. https://doi.org/10.1016/j.jfoodeng.2011.10.005.