http://pubs.acs.org/journal/acsodf

Review

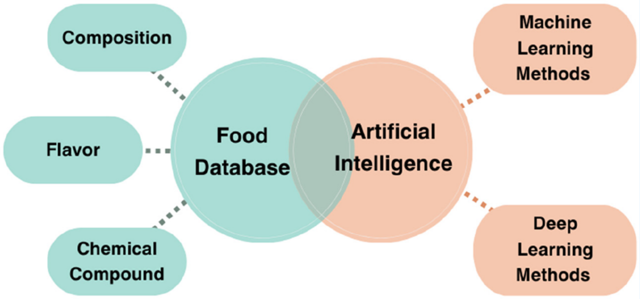
When Machine Learning and Deep Learning Come to the Big Data in Food Chemistry

Yufeng Jane Tseng,* Pei-Jiun Chuang, and Michael Appell

Cite This: *ACS Omega* 2023, 8, 15854−15864

Read Online

ACCESS | Metrics & More | Article Recommendations



ABSTRACT: Since the first food database was released over one hundred years ago, food databases have become more diversified, including food composition databases, food flavor databases, and food chemical compound databases. These databases provide detailed information about the nutritional compositions, flavor molecules, and chemical properties of various food compounds. As artificial intelligence (AI) is becoming popular in every field, AI methods can also be applied to food industry research and molecular chemistry. Machine learning and deep learning are valuable tools for analyzing big data sources such as food databases. Studies investigating food compositions, flavors, and chemical compounds with AI concepts and learning methods have emerged in the past few years. This review illustrates several well-known food databases, focusing on their primary contents, interfaces, and other essential features. We also introduce some of the most common machine learning and deep learning methods. Furthermore, a few studies related to food databases are given as examples, demonstrating their applications in food pairing, food−drug interactions, and molecular modeling. Based on the results of these applications, it is expected that the combination of food databases and AI will play an essential role in food science and food chemistry.

INTRODUCTION

With the development of food chemistry and nutrition science, food composition and the relationship between diet and health have received more attention. It is noticeable that nutrition guidance and educational programs on choosing a healthy diet are more valued than ever. In addition, the specific nutrition requirements for certain diseases are also recent emphases. Utilizing food composition data and developing therapeutic diets to treat obesity and food allergies have become clinically practical. Moreover, complete and transparent nutrition labeling, allowing consumers to freely choose between similar products, is considered ordinary and necessary today.[1−3] An increasing number of countries have promulgated new laws and dictated that food companies should provide clear and correct information on food labels.[4] To fulfill the current needs with respect to food data resources, information regarding food ingredients, nutrition, and even bioactive compositions are collected and formed into all kinds of food databases. Many countries worldwide have designed and established food composition databases, providing the nutritional contents of many generic and branded foods. The first food composition database was proposed in Germany in 1878, followed by the United States (US) and some European countries (for example, Denmark, the United Kingdom (UK), France, Italy, The Netherlands, and Sweden).[5] Although the detailed items of the food composition data in these databases are diverse, they share similar goals: to assess health and nutritional statuses, formulate appropriate diets for specific groups of people, conduct epidemiological research, and develop new products and recipes in food industries. Food databases theoretically contain more items than drug databases, which aim at drug discovery, adverse reaction identification, and drug interaction determination. The diversity of the data in these food databases is associated with the extensive ranges of ingredients and compositions that even differ between foods of the same type. Additionally, the chemical molecules in food have still not been fully qualified or quantified. Therefore, with the introduction of big data, information about food science and food chemistry can be structured into a more organized and searchable database and used for systematic applications and research purposes.[6]

Today, the development of AI is considered a revolution that pioneers innovations in our modern society. The definition of AI varies across different fields. For instance, in

Received: December 6, 2022
Accepted: April 7, 2023
Published: April 25, 2023

https://doi.org/10.1021/acsomega.2c07722

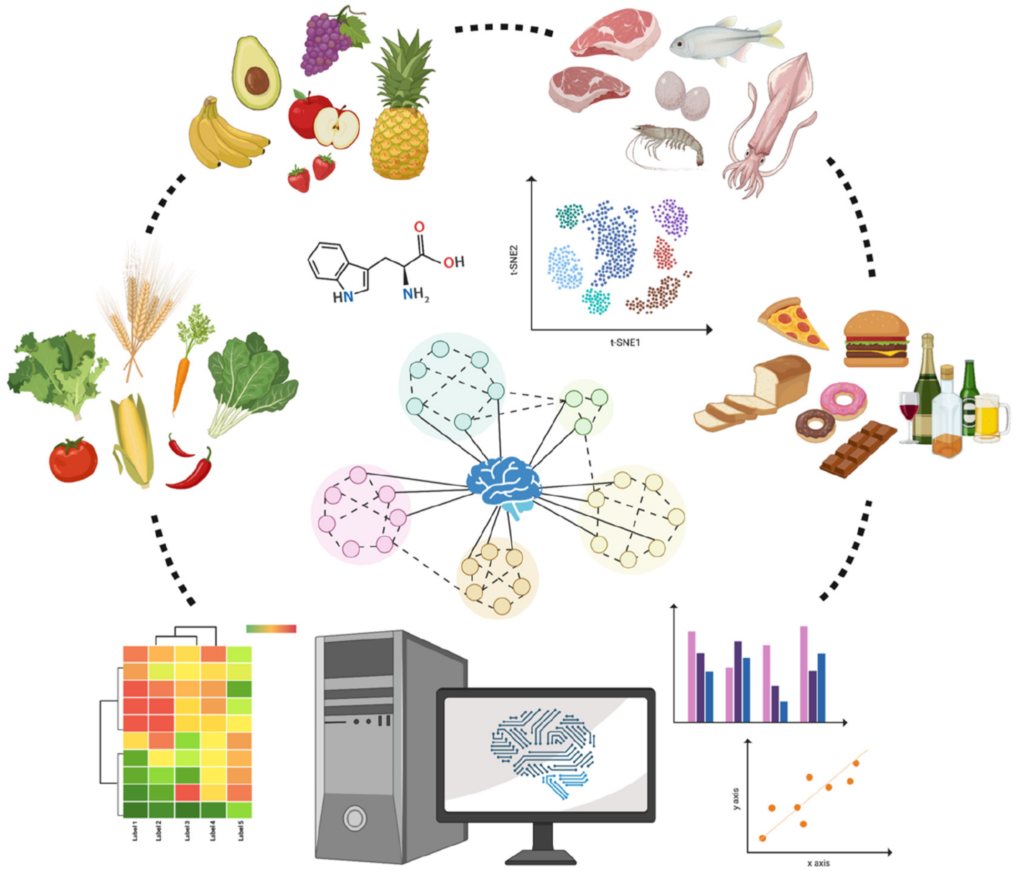


Figure 1. Big data in food. Note: this graphic was created with BioRender.com.

computer science, AI refers to the development of computers that can engage in human-like thought processes such as learning, adapting, reasoning, and self-correction.[7] AI covers any technology that enables computers or machines to mimic human behavioral patterns and thought processes.[8] Machine learning and deep learning are related but distinct subfields of AI. In short, machine learning is a broader field that includes deep learning, but deep learning is a specialized form of machine learning that uses neural networks to analyze complex data.[9−11] Machine learning and deep learning have already been widely used in image recognition, video processing, and even molecular designing (such as for drug discovery).[9,10,12,13]

Machine learning can be considered a subfield of AI, and it approaches the problem of modeling by trying to find an algorithmic model that can better predict the output from input variables.[14] Machine learning demonstrates some critical advantages, including automation and the continuous improvement exhibited by its algorithms. The wide applications of machine learning make it easier for users to utilize in different fields.[15] Machine learning can be roughly divided into four subtypes: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.[16−19] The concept of supervised learning is that the input and output data are labeled or precategorized before executing the computation. Through the iterative optimization of an objective function, a supervised learning algorithm learns a function that can predict the outputs associated with new inputs. Unsupervised learning uses unlabeled data sets, and its algorithms are designed to find hidden or undefined patterns in the data. Semisupervised learning combines supervised and unsupervised methods and is useful when numerous unlabeled and scarcely labeled data are available. Reinforcement learning refers to goal-oriented algorithms that learn how to achieve a complex goal or maximize a particular dimension over many steps.

On the other hand, although machine learning has been widely applied, deep learning has recently demonstrated more power than machine learning.[20,21] As a subset of machine learning, the concept and technique of deep learning enable the computation of multilayer neural networks to be more feasible and accurate. Unlike machine learning, which includes algorithms that learn from data to predict outputs and discover patterns, deep learning develops algorithms based on highly complex neural networks that mimic the way a human brain works to detect patterns in large data sets. Deep learning is designed to automatically learn representations of data, allowing it to make predictions based on patterns in the data that would be difficult for humans to find.[11] Deep learning can produce incredible results in computer vision, speech recognition, text analysis, and drug research.[10,22,23] Commonly used deep learning approaches include recurrent neural networks (RNN), convolutional neural networks (CNN), and graph convolutional networks (GCN). CNN has strengths when solving problems related to spatial data, such as images, while RNN is more suitable for analyzing temporal and sequential data, such as text or videos. GCN, a variant of a graph neural network (GNN), has been developed to address graph-structured data.[24−27]

For a good machine learning model or deep learning model, it is important to know that the data in the training set is

https://doi.org/10.1021/acsomega.2c07722

Table 1. Comparison of Food-Related Databases

	name						
	FoodData Central	EuroFIR	InFoods	FlavorDB	BitterDB	VirtualTaste	FooDB
type	food composition	food composition	food composition	flavor	flavor	flavor	food composition
							chemical compound
organization	USDA	EuroFIR AISBL	FAO	Center for Computational Biology, IIIT Delhi, India	Institute of Biochemistry, The Hebrew University of Jerusalem, Israel	Charite University of Medicine, Institute of Physiology, Berlin, Germany	TMIC
source of data	analytically derived values	chemical analysis and nutrient calculation of food	analytical data	Fenaroli's handbook of flavor ingredients	the Merck index	Pubchem	textbooks
	scientific publications	scientific literature	other published sources	FooDB database	Fenaroli's handbook of flavor ingredients	Protein Data Bank	scientific journals
	food labels	food labels		literature survey	PubChem	Literature survey	online databases (food composition, nutrient, flavor, metabolomics).
					UniprotKB database		
					Literature survey		
size	vary between five different databases	vary between five different databases	vary between different databases	25595 flavor molecules and 1101 receptors (accessed on 20 January 2023)	1041 bitter compounds and 75 receptors (accessed on 20 January 2023)	more than 2000 (sweet molecules) /1,600 (bitter molecules)	70926 compounds and 797 foods (accessed on 20 January 2023)
content	the five integrated databases offer nutrient analyses, agricultural and production practices, and information about branded/experimental food	the five databases provide composition data, bioactive herbal compounds, and the latest database explores how food waste might be used	tables and databases offer information about food composition, nutrition labeling, supplements, and biodiversity data to meet the needs of the various users	information includes the fundamental identity, functional groups, and physiochemical properties of each molecule	information includes bitter compounds, mutations in bitter receptors that influence receptor activation by bitter compounds	information includes structure and properties of carbohydrates, artificial sweeteners and other sweet tasting agents. a modeled 3d structure of sweet receptor binding poses is included	information includes food macronutrients and micronutrients. details such as compositional, biochemical and physiological information are included
access of data	online search with search filter function	(member only)	online search	online search	online search by name, structure, similarity, association with bitter receptor	online search with similarity search function	online search by food source, name, or descriptors
	data download	online search by name, Langual code, or descriptor of food	data download	data download (can be obtained in Mol2, 2D image and SDF formats)	data download		data download
		Data download					
regularity of updates	updated regularly (except for SR Legacy Foods)	updated regularly (with annual report available online)	updated regularly	not mentioned (latest version FlavorDB2 updated in 2022)	not mentioned (latest version 2.0 updated in 2018)	updated regularly	updated every 2−3 months
URL	https://fdc.nal.usda.gov	https://www.eurofir.org/food-information/	https://www.fao.org/infoods/infoods/en/	https://cosylab.iiitd.edu.in/flavordb/	https://bitterdb.agri.huji.ac.il/dbbitter.php	https://insilico-cyp.charite.de/VirtualTaste/	https://foodb.ca

representing the diversity of the chemical space and the space for the properties that the model is supposed to learn from.[28] Generally, the first step is to collect and prepare reliable data for analysis. The quality of the data that feeds into the model determines how accurate the model is. If incorrect or outdated data are imported, we may have wrong outcomes or predictions which are not relevant. The following tips are useful while selecting data for machine learning and deep learning: relevance (choose data that are relevant to the problem you are trying to solve),[29,30] quality (ensure the data are of high quality and free from errors, outliers, and bias),[31−33] representativeness (the data should be representative of the population you are trying to model),[29,34] quantity (enough data are needed to support the complexity of the model you are trying to build),[31,33,35] and diversity (the data should contain diverse examples and variations to help the model generalize better to unseen examples).[33] After data preparation, an appropriate method or algorithm should be chosen to develop models that are suited for different tasks. Then, we can start training the model by feeding the prepared data into these learning algorithms to find patterns and make predictions. Finally, these models should be evaluated with previously unseen data to verify their performance.

The complex compositions and various ranges of compounds in foods can be structured into informative food databases. Food compositions' chemical information and physicochemical properties are potentially much more extensive than those of drugs. The lack of accurate information on food compounds, along with relatively little experience in beneficially applying machine learning applications or deep learning methods to these problems, seems to be a major obstacle in the area. Therefore, this review introduces the publicly available big data resources concerning food composition and chemistry and illustrates the learning methods applied in this area (Figure 1).

■ BIG DATA SOURCES

Food-related data can generally be categorized into food compositions, flavors, and chemical compounds. Food composition databases mainly focus on the ingredients, nutrients, and labeling of food products. Although many countries have their own composition database, the databases created by organizations such as the United States Department of Agriculture (USDA), European Food Information Resource Network (EuroFIR AISBL), and United Nations Food and Agriculture Organization (FAO) are all well-known and comprehensive data sources. Food flavor databases are another category that focuses on the molecular properties of natural and synthetic flavor molecules. For example, one of the most prominent flavor databases, FlavorDB, contains over 25000 flavor molecules representing an array of tastes and odors.[36,37] Lastly, food chemical compound databases such as FooDB contain more than 70000 chemical compounds identified in foods.[38] Several food-related databases are listed and compared in Table 1.

Food Composition Databases. The USDA created and has maintained one of the most well-known food composition databases. USDA FoodData Central, formerly the USDA Food Composition Database, has been openly accessible since October 1, 2019. It is an integrated, research-focused data system that comprises five distinct data sets, including Foundation Foods, the Food and Nutrient Database for Dietary Studies (FNDDS) 2017−2018, the National Nutrient Database for Standard Reference Legacy Release (SR Legacy Foods), the USDA Global Branded Food Products Database (Branded Foods), and Experimental Foods.[39] Each data set has its purposes and attributes. First, the Foundation Foods database offers nutrient analyses and metadata for a range of single foods and ingredients. Sometimes, it allows users to see agricultural information and production practices. Second, the FNDDS database converts the food and beverages consumed in What We Eat in America (WWEIA), the National Health and Nutrition Examination Survey, into gram amounts and determines their nutrient values. Next, the SR Legacy Foods database provides the foundation for most food composition databases in the public and private sectors, which can be used to develop dietary guidelines or meal plans and to conduct product labeling as a collaboration between food industry organizations and the USDA. The Branded Foods database comprises numerous kinds of branded food information submitted voluntarily by the food industry. Lastly, the Experimental Foods database focuses on research aspects and a deeper understanding of the factors related to food composition, such as foods under unique conditions, experimental genotypes, or research/analytical protocols.[40] In addition to the online search function, the food composition information can be accessed by downloading data from the USDA FoodData Central Web site.[41]

The emergence of ontology is crucial for providing linkages among terminology for names, characteristics, chemical compounds, newly identified components, and other relevant features among the various information in a food database. From philosophical concepts, ontology studies the nature of existence, being, becoming, and reality.[42] However, in computer science and information science, ontology refers to a formal, explicit, and detailed representation of a shared conceptual system, which is structured and presented with particular terminology. With ontology, humans and computers can compare and contrast data to see if they represent the same entities or classify their attributes and relationships accordingly. In addition, ontology broadens the scope of AI and machine learning by including unstructured or structured data types, covering each aspect of the data modeling process, and improving the quality of the data in training data sets.[43]

Additionally, ontology can be applied to a set of individual facts to create a knowledge graph, a collection of entities that express the types of relationships among them. An example of the application of ontology to food databases is FoodOn, a data dictionary that describes a hierarchy of over 9600 generic food product categories.[44] This food branch aims to provide unambiguous and easy-to-reference classifications and standardized information about critical ingredients, components, and properties. Furthermore, FoodOn can be used for reference in the research and development of machine learning algorithms, food-related software, and other applications. Collaborating with the global food web, FoodOn is helpful for monitoring resources and waste in natural systems and represents the impacts of the human food system on global biodiversity and the integrity of the ecosystem.[45]

Food Flavor Databases. Flavor molecules in foods trigger the chemical processes that produce sensations of taste and smell and play an essential role in regulating metabolic processes. FlavorDB, an online database developed by the Center for Computational Biology from Indraprastha Institute of Information Technology Delhi (IIIT Delhi) in India, was created to integrate the multidimensional aspects of flavor molecules and demonstrate their molecular features, flavor profiles, and natural source details. Unlike other flavor databases, such as BitterDB[46] and VirtualTaste[47] (including the data from the previous SuperSweet database[48]), which mainly focus on the particular aspect of flavor, FlavorDB collects all kinds of references to compile a comprehensive flavor database. The flavor molecule data originated from resources such as Fenaroli's handbook of flavor ingredients, the FooDB online database, the arXiv preprint service, and a literature survey. A total of 25595 flavor molecules are recorded on the FlavorDB Web site. Among them, 2254 molecules are associated with natural entities or ingredients, 13869 are synthetic molecules, and 9472 are from uncertain sources. The information provided by this database includes the fundamental identity, functional groups, physicochemical properties, 2D/3D properties, and ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties of each flavor molecule.[36]

Furthermore, FlavorDB contains information regarding 33 taste receptors (sweet, bitter, sour, and umami) and 1068 odor receptors, with a UniProt link providing more details about each receptor. Several search interfaces were designed to make the database more convenient for users. First, the flavor network function allows users to explore the flavor molecules' similarities shared among different food entities. The visual search interface offers a graphical way to search the flavor molecules contained in foods. In addition, users can pair two food ingredients and examine whether they share one or more flavor molecules. FlavorDB2, an updated and expanded version of the original online database, has been released.[49] Although the numbers of compounds and ingredients remain the same as those in FlavorDB, more features were added in FlavorDB2, including chemical properties, regulatory statuses, consumption statistics, taste or aroma threshold values, reported uses in food categories, and synthesis scenarios.[50]

Food Chemical Compound Databases. As the most extensive database on food constituents, chemistry, and biology, FooDB is perhaps one of the most comprehensive and informative online resources. FooDB was built and supported by the Canadian Institutes of Health Research (CIHR), the Canada Foundation for Innovation, and the Metabolomics Innovation Centre (TMIC). The data concerning each food ingredient and compound were collected from textbooks, scientific journals, online food composition or nutrient databases, flavor databases, and some metabolomic databases. More than 70000 compounds and nearly 800 foods are included in the online database, and users can search these foods and compounds with or without filters.[38] The numbers of compositions and nutrients are detailed on the food interface, whereas the chemical information, classification, ontology, physicochemical properties, and spectrum are recorded on the compound interface. It is noticeable that the compounds in FooDB have highly diverse physicochemical properties, such as different partition coefficient (S log P), topological polar surface area (TPSA), atomic mass weight (AMW), rotatable bond (RB), hydrogen bond donor (HBD), and hydrogen bond acceptor (HBA), relative to those in the other three common chemical databases: namely, DrugBank, GRAS, and ZINC.[51] FooDB users can search for information in this database in many different ways. For instance, they can browse by food source, name, descriptor, or chemical compound.

Moreover, the ChemQuery Search function allows users to find the target compound after inputting the desired structure or molecular weight. In contrast, the Spectra Search function uses the mass (m/z) and intensity corresponding to one peak per line on the mass spectrum to find the target compound. The applications of FooDB are diverse. Apart from information searching, the data in FooDB can be combined with other tools to investigate the possible bioactive compounds in foods and predict the interactions between food chemical compounds and protein targets in the human body.[52,53] The results can help users investigate and develop new pharmaceuticals, nutraceuticals, and personalized dietary plans.

■ LEARNING METHODS

Many machine learning and deep learning algorithms have been proposed in the past few decades, and some of them can be utilized for food science research. To deal with complex problems more effectively and ensure the representativeness of the input data, there are various ways to approach generating features representing structural information from graphs, words, etc. For example, embedding methods can preprocess and turn graphs or words into a computable format to exploit them by machine learning or deep learning algorithms. Each learning method has its features, applications, and advantages or disadvantages. The following paragraphs and Table 2 illustrate several commonly used machine learning and deep learning methods with potential applications in food database research.

Machine learning refers to a broad set of algorithms that can automatically detect patterns in data and then use those patterns to make a prediction. Many machine learning algorithms have also been developed to address different needs. The common approaches include support vector machine (SVM), random forest (RF), and decision tree (DT) based algorithms such as Iterative Dichotomiser 3 (ID3), C4.5.

Support Vector Machine (SVM). SVM, categorized as a supervised learning method, is an algorithm that analyzes data for classification and regression analysis. By mapping samples into a higher-order feature space, SVM finds a linear hyperplane with a gap that can separate those samples. New inputs are mapped into the particular feature space, and which side of the hyperplane they fall into is then determined. SVM has been used to solve real-world problems, such as image detection and classification, text and hypertext categorization, drug discovery, and other bioinformatics applications.[54−56] SVM is generally unaffected by small data changes and generalizes data efficiently. However, SVM underperforms on larger data sets, on overlapping target classes, or in cases when the number of features for each data point exceeds the number of training data samples.[57]

Decision Tree (DT). DT method is an open-box model that constructs a binary tree containing decision nodes and branches. It can be used for both classification and regression.[58] While nodes represent the attributes in a group to be classified, branches represent the values taken.[16] DT seeks to find the best split for subsetting the given data, and they are typically trained through algorithms such as ID3, C4.5, and the classification and regression tree (CART).[59,60] In data mining, DT can also be described as a combination of mathematical and computational techniques that aid in the description, categorization, and generalization of a given data set. Since DT mimics the decision-making process, its main

Table 2. Common Machine Learning and Deep Learning Methods

name	support vector machine (SVM)	decision tree (DT)	random forest (RF)	recurrent neural network (RNN)	convolutional neural network (CNN)	graph convolutional network (GCN)
category	machine learning (a subset of artificial intelligence): can predict outputs and discover patterns from data, and automatically adapt with minimal human interference			deep learning (a subset of machine learning): uses artificial neural networks to mimic the learning process of the human brain to detect patterns in large data sets		
type	supervised	supervised	supervised	supervised	supervised	semisupervised
approach	by mapping samples into a higher-order feature space, SVM finds a linear hyperplane with a gap that separate those samples	DT is a binary tree containing decision nodes and branches, which mimics the decision-making process and finds the best split for subsetting data	by combining multiple DT, RF predicts value based on the majority votes and the average outcomes of DT in classification and regression cases, respectively	with recurrent cycles over time or sequence, RNN minimizes the difference between the output and input target pairs by optimizing the network weights	CNN processes data with grid patterns by automatically and adaptively learning the spatial hierarchies of the features	GCN employs weighted nodes to make decisions that mimic those of brain neural networks, generating predictions over physical systems
advantages	generally unaffected by small data changes and generalizes data efficiently	good visual interpretability of the trained models, and able to handle both numerical and categorical data	easy to evaluate the resultant model's variable importance or contribution levels	helpful in any time series predictor, and the model size does not increase even if the input size is enlarged	computationally efficient and highly accurate, with better model interpretability	powerful feature extractor which learns graphic representations and achieves superior performance
disadvantages	underperforms on larger data sets, overlapping target classes, or when the number of features for each data point exceeds the number of data samples	less effective in making predictions when predicting the outcome of a continuous variable, and may lead to overfitting because of its high flexibility	a large number of DT can make the algorithm time-consuming and ineffective for real-time predictions	time-consuming and challenging training process; the gradient vanishing and gradient exploding problems are potential concerns	large training data needed, and CNN does not encode the position and orientation of object	limited flexibility; need considerable amount of labeled data for validation and model selection
applications	image detection and classification, text and hypertext categorization, drug discovery, and other bioinformatics applications	probability prediction, risk analysis, decision-making and strategic management	behavior and outcome prediction in banking and finance, e-commerce, stock market, and healthcare	ordinal/temporal problems: machine translation, language modeling, and handwriting or speech recognition	image recognition, video analysis, disease evaluation, drug discovery, and biomarker discovery	computer vision, natural language processing, social analysis, bioinformatics, molecular property prediction, polypharmacy side effects prediction

advantage lies in the visual interpretability of the trained models, and it can handle both numerical and categorical data. Nevertheless, DT can lead to overfitting because of its high flexibility.[61]

Random Forest (RF). RF is also a commonly used supervised learning method that combines the outputs of multiple DTs to reach a single result. In classification cases, RF performs prediction based on the majority votes of the predicted DT values. In regression cases, the result is the average of the outcomes derived from the DT.[62] Multiple decision trees form an ensemble in the RF algorithm and predict more accurate results, mainly when the individual trees are uncorrelated. As an ensemble learning algorithm, the RF method overcomes the main drawback of the DT method, including the overfitting of training data sets.[58] In addition, RF makes it easy to evaluate the resultant model's variable importance or contribution levels. Although RF is generally more capable of handling large data sets than DT, it can be time-consuming and require more resources to store them.[61]

On the other hand, deep learning is considered more powerful and applicable in many fields. Automatically extracting the features from input data sets makes it more efficient and accurate when discovering vital information from a large-scale data set. RNN, CNN, and GCN are all frequently used deep learning methods.

Recurrent Neural Network (RNN). RNN is an artificial neural network (ANN) with one or more feedback loops that are recurrent cycles over time or a sequence. The training process of the RNN algorithm minimizes the difference between the output and input target pairs by optimizing the network weights.[63] Derived from feedforward neural networks, RNN uses its internal state memory to process variable-length sequential data. While traditional deep neural networks assume that inputs and outputs are independent of each other, the RNN algorithm results depend on the sequence's prior inputs. This algorithm applies to ordinal or temporal problems, such as machine translation, language modeling, and handwriting or speech recognition.[64−67] RNN is modeled to remember each piece of information throughout the period, which is very helpful in any time series predictor. Even if the input size is enlarged, the model size does not increase. However, the model training process can be time-consuming and challenging, and the gradient vanishing and gradient exploding problems are potential concerns.[68]

Convolutional Neural Network (CNN). CNN, a dominant ANN in various computer vision tasks, is designed to process data with grid patterns by automatically and adaptively learning the spatial hierarchies of features. Generally, the CNN method consists of three layers: convolution layers, pooling layers, and fully connected layers. While the first two layers are used for feature extraction, the last layer connects all activations and produces the output.[69] Regarding text recognition tasks, CNN aims to learn feature representations for a fixed-size context and quickly increase the adequate context size by stacking several layers, providing a shorter path than an RNN to better model the sequential dependences between the characters in the given text.[70] The CNN algorithm is suitable for image recognition, video analysis, disease evaluation, drug or biomarker discovery, and research in many other fields.[69,71−73] Although CNN does not encode the positions and orientations of objects when performing image recognition, it is considered computationally efficient and highly accurate, with better model interpretability.

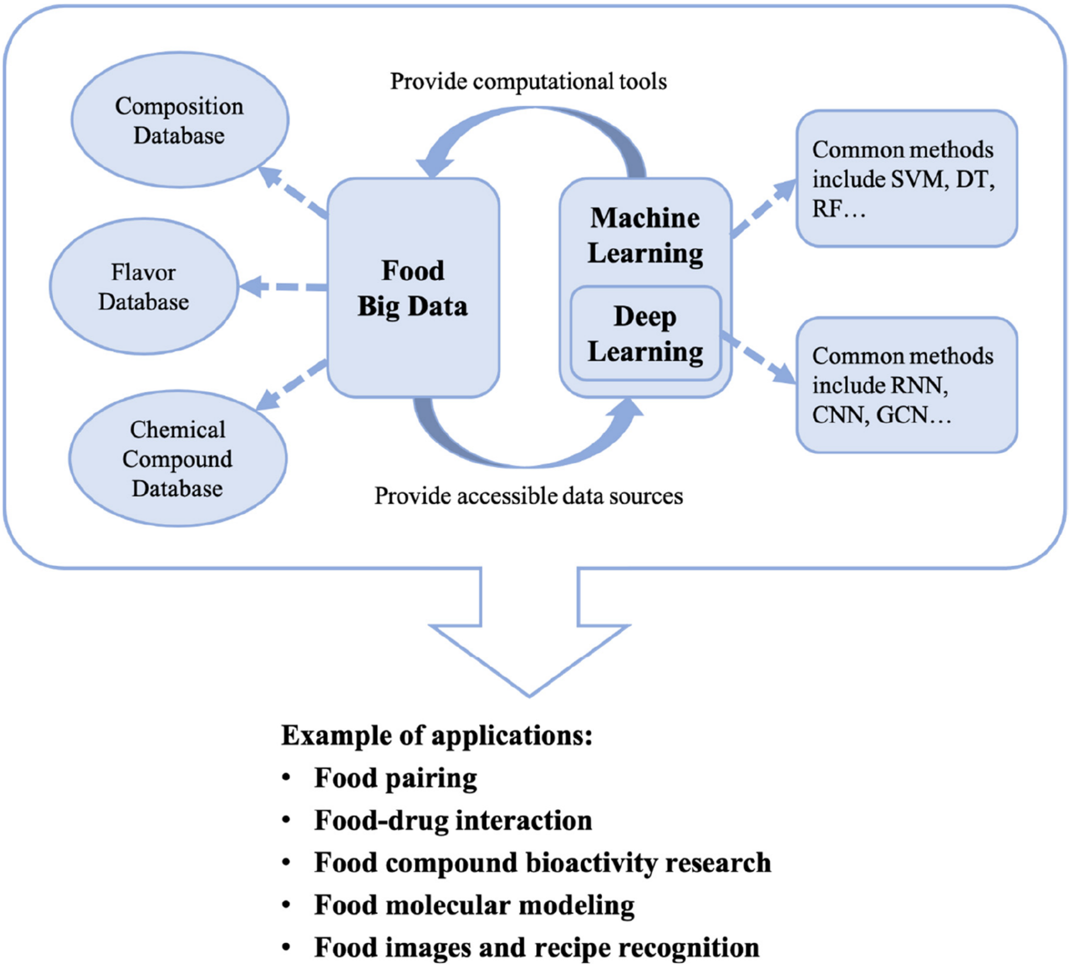

Provide computational tools
Composition Database
Flavor Database
Chemical Compound Database
Food Big Data
Machine Learning
Deep Learning
Common methods include SVM, DT, RF…
Common methods include RNN, CNN, GCN…
Provide accessible data sources
Example of applications:
• Food pairing
• Food-drug interaction
• Food compound bioactivity research
• Food molecular modeling
• Food images and recipe recognition

Figure 2. Concept of machine learning and deep learning in food data.

Graph Convolutional Network (GCN). A graph neural network (GNN) is a type of deep learning method that mainly focuses on graph analysis. Among the variants of GNN, GCN has demonstrated groundbreaking performances in many deep learning tasks. As an approach for performing supervised learning on graph-structured data, GCN employs weighted nodes to make decisions that mimic those of brain neural networks, generating predictions over physical systems, such as graphs, interactive approaches, and applications.[74] It also provides accurate information about the properties of real-world entities and physical systems. For example, one of the critical GCN applications is molecular property prediction in chemistry. By extracting features from simple graph structure descriptions, GCN can form molecular-level representations that are used as fingerprint descriptors. In addition, the GCN method is an emerging tool for predicting polypharmacy side effects by modeling protein−protein interactions and drug−protein interactions. Overall, GCN has tremendous expressive power for learning graphic representations and achieves superior performance in various tasks and applications.[58,75−77]

APPLICATIONS

Many studies have been conducted with information obtained from various food databases. With abundant and multiple data sources, the content in food databases can be effectively analyzed and computed with different machine learning or deep learning methods (Figure 2). Furthermore, combining food databases and learning methods helps resolve problems such as food pairing, food−drug interactions (FDIs), and molecular modeling. Several specific tools other than traditional learning methods are also incorporated into these studies, including graph embedding, link prediction, and dimensionality reduction. Therefore, in this section, we review and demonstrate some studies to show the practical applications of food databases and learning methods.

Food Representations and Food Pairings. Food representations and pairings are considered critical topics in food science and are essential in cooking. Previous studies have tried improving the efficiency and quality of food representations and pairings through chemical-based approaches. Since the biochemical data of different food ingredients are of great complexity and diversity, the lack of accurate and detailed information makes it difficult to construct precise food representations. The second strategy is the recipe-based approach that is based on the statistical co-occurrence among many recipes. Chemical compound data are not considered when creating food representations and recommending food pairings. FlavorGraph, a large-scale network graph built from the recipes and chemical relations of food ingredients and chemical compounds, was introduced to improve the performance of food clustering and recommend food pairings.[78]

The FlavorGraph data were extracted from information in Im2recipe, FlavorDB, and HyperFoods, which were constructed into three types of nodes (food ingredients, flavor compounds, and drug compounds) and three types of edges (food ingredient−food ingredient relations, food ingredient−flavor compound relations, and food ingredient−drug compound relations).[37,79,80] The graph embedding method used for constructing similar representations of heterogeneous nodes based on commonly linked nodes is called metapath2vec. This meta-path-guided random walk strategy is capable of capturing both the structural and semantic correlations of differently typed nodes and relations, further improving the accuracy of food pairing recommendations.[81] By utilizing a graph embedding method and neural network algorithm, the food representation neural network learned the food−chemical molecule relations and the relations between the ingredients in many recipes. With an extra chemical structure prediction (CSP) layer, more details about chemical structural information were added to the original model, and this compound−ingredient relation was expected to generate more significant node representations.

One of the valuable functions of FlavorGraph is that it can recommend complementary and novel food pairings in cooking based on multiple learned food representation vectors. Unlike KitcheNette, another food ingredient pairing model, FlavorGraph includes chemical information about food ingredients and predicts pairings more accurately.[82] Moreover, FlavorGraph can predict flavor compound−food relations with its similarity search function. Flavor compounds are molecular substances with various flavors, such as fruity, bitter, fatty, floral, etc. Flavor compound−food relations can help clarify the chemical effects of flavor ingredients on other food ingredients. The study identified six new flavor compound−food relations that demonstrate promise for future culinary practice.

Food−Drug Interactions (FDIs). Research on drug−drug interactions (DDIs) and drug−target interactions (DTIs) has already been comprehensively explored and studied. Nevertheless, some studies have demonstrated the impacts of certain foods on the activities of different drugs by increasing drug metabolism, decreasing drug bioavailability, or creating adverse effects. Similar to DDIs, the mechanisms of FDIs can be classified into two primary categories: pharmacokinetic (PK) interactions and pharmacodynamic (PD) interactions. In addition to the FDIs that have been thoroughly investigated, the identification of potential unknown FDIs is crucial for ensuring treatment effectiveness and safety.

A systematic framework called FDMine was proposed to predict new FDIs and examine the possible pharmacological effects of food compounds.[83] The raw data extracted from the FooDB and DrugBank database were used to build a homogeneous network based on a structural similarity profile, with nodes representing drugs, foods, and compounds.[38,84] Link prediction algorithms can predict the existence of a link between two entities in a network. Several neighborhood-based similarity-based link prediction algorithms are implemented in this study, such as Adamic and Adar coefficient (AA), common neighbor (CN), Jaccard coefficient (JAC), resource allocation (RA), multiple paths of length $L = 3$ (L3), and dice coefficient. The links in FDMine were weighted by similarity and contribution scores. To evaluate the performance of different link prediction algorithms applied in the framework, 70% of links were randomly selected as training data sets, whereas the remaining links were used as test data sets to verify the accuracy of the algorithms. All neighborhood-based similarity-based methods (except L3) achieved more than 80% area under the receiver operating characteristic curve (AUROC).

Some novel FDIs have been discovered with FDMine, providing valuable information to physicians and researchers. For instance, foods such as garden onions contain fatty acids with anti-inflammatory effects, including oleic acid and elaidic acid. In addition to interacting with PPAR receptors to decrease prostaglandin production, these foods also cross the blood−brain barrier and interact with GABA receptors to elevate overall GABA levels, further inducing anxiolytic and possible antiepileptic effects. These mechanisms imply a synergistic relationship between drugs such as vigabatrin and foods such as garden onions, pomegranates, pineapples, and peanuts. Another FDI case concerns the interaction between beta-adrenergic antagonists and food components that possess vasodilation pharmacological effects, including p-cymene, eugenol, and carvacrol. Cloves, hyssops, and anises are some of the common substances that contain these compounds. The result of such an interaction may cause smooth muscle vasodilation and more pronounced antihypertensive effects. The investigation of FDIs makes it more likely to avoid some potentially detrimental effects when taking medications.

Molecular Modeling. Notably, the biological properties of food compounds have also been investigated at the molecular level. Unlike drug molecules, where a number of guidelines have been proposed to investigate their physicochemical properties, the specific molecular descriptors of food chemicals have not yet been developed. Both quantitative and visual approaches have been utilized to analyze the chemical space of food molecules, and the potential associations between molecular structure, flavor, and odor are still being investigated currently.[85] A recent study that aimed to investigate the binding modes and the chemical spaces of hop-derived compounds demonstrated the utilization of food databases in bioinformatics and chemoinformatics.[86] As compounds with favorable bitter tastes in beer, hop-derived compounds typically bind with three bitter taste receptors: TAS2R1, TAS2R14, and TAS2R40. These compounds modulate the dynamic conformational changes exhibited by TAS2R receptors, allowing for further signal transmission. In this study, a knowledge-based homology modeling approach, with sequence alignment of TAS2R1, TAS2R14, and TAS2R40, was used to construct the three-dimensional (3D) structures of the TAS2R models. A total of 11 hop-derived compounds were included. The chemical information about these hop-derived compounds and other bitter compounds was obtained from BitterDB, FooDB, and DrugBank.[38,84,87] The molecular similarity was computationally evaluated and visualized by t-distributed stochastic neighbor embedding (t-SNE), a machine learning based algorithm used for data visualization in high-dimensional data sets.[88] Molecular docking was also used to predict the binding modes of small molecules into protein targets at physiological pH.

As a result, the binding site compositions differed between the TAS2R receptors, and the sequence identity was lower in the binding site region than in the whole receptors. The conserved asparagine at position BW 3.36 significantly recognized the hop-derived compounds. Because of the diverse binding site composition shapes, each receptor's residue interaction pattern for compound recognition also varied. The Glide standard precision (SP) score was demonstrated as an approximation of the binding affinity of each compound against analyzed receptors. Compared to less potent agonists, the result showed that more potent compounds have lower binding energy (lower Glide SP scores) against the cognate receptors. Although the chemical spaces of the compounds obtained from FooDB and DrugBanks shared an overlap in a previous study, the chemical similarity analysis showed possible discrimination between the chemical spaces of bitter drugs and bitter foods.[51] While comparing these bitter compounds with hop-derived compounds on a 2D t-SNE plot, the latter clustered in a confined space of the bitter chemical space and further demonstrated a peculiar type of bitter compound compared to those of known bitter drugs and foods.

SUMMARY AND PATHWAY FORWARD

Big data and AI have revolutionized the study of food chemistry. In contrast to medicinal chemistry, the study of food compound analysis is currently in full swing. With the establishment of many food ingredient and food compound databases, machine learning and deep learning methods are becoming valuable tools for effectively analyzing large data sets. In addition to attaining a better understanding of the chemical and pharmacological properties of food ingredients and compounds, our review suggests that more efficient solutions are available to address problems such as food pairing, FDIs, and molecular modeling. These applications can lead to further progress in food science and food chemistry.

Current research applying big data and AI to food science show promise to increase food databases' value and the meta-analysis and broader insights into food systems. Quantum computing will enhance computational prowess. Recent efforts to archive spectral libraries will provide more refined information on food composition and new food chemicals of interest and expand the number of known food constituents. Coupled with the expanding databases of biological and physical activities and the spectral databases of microcrobes, plants and animals, there are opportunities for synergistic benefits from this big data and AI revolution.

AUTHOR INFORMATION

Corresponding Author

Yufeng Jane Tseng − *Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei 10617, Taiwan;* orcid.org/0000-0002-8461-6181; Phone: +886.2.3366.4888#529; Email: yjtseng@csie.ntu.edu.tw; Fax: +886.2.23628167

Authors

Pei-Jiun Chuang − *Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei 10617, Taiwan*

Michael Appell − *USDA, Agricultural Research Service, National Center for Agricultural Utilization Research, Mycotoxin Prevention and Applied Microbiology Research Unit, Peoria, Illinois 61604, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c07722

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was financially supported by the Taiwan Ministry of Science and Technology (MOST 110-2320-B-002-038-), the Taiwan Food and Drug Administration (MOHW111-FDA-D-114-000611), National Taiwan University (NTU-111L8809, NTU-CC-111L890203), and the Toxic and Chemical Substances Bureau, Environmental Protection Administration, Executive Yuan, Taiwan, ROC (110A022). This study was performed using resources at the Laboratory of Computational Molecular Design and Metabolomics (CMDM), the Department of Computer Science and Information Engineering (CSIE), and the Graduate Institute of Biomedical Electronics and Bioinformatics (BEBI) at National Taiwan University. This study was led by Professor Y.J.T. and conducted by P.-J.C.. The authors hereby declare no conflicts of interest. This work was supported in part by the U.S. Department of Agriculture, Agricultural Research Service. The mention of trade names or commercial products in this publication solely provides specific information. It does not imply a recommendation or endorsement by the United States Department of Agriculture (USDA). USDA is an equal opportunity provider and employer.

REFERENCES

(1) Marconi, S; Camilli, E. Food Composition Databases: Considerations about Complex Food Matrices. *Foods.* **2018**, *7* (1), 2.

(2) Elmadfa, I.; Meyer, A. Importance of food composition data to nutrition and public health. *European Journal of Clinical Nutrition.* **2010**, *64*, S4−S7.

(3) Delgado, A.; Issaoui, M.; Vieira, M. C.; Saraiva de Carvalho, I.; Fardet, A. Food Composition Databases: Does It Matter to Human Health? *Nutrients.* **2021**, *13* (8), 2816.

(4) Messer, K. D.; Costanigro, M.; Kaiser, H. M. Labeling Food Processes: The Good, the Bad and the Ugly. *Applied Economic Perspectives and Policy.* **2017**, *39* (3), 407−427.

(5) Church, S. M. The history of food composition databases. *British Nutrition Foundation - Nutrition Bulletin.* **2006**, *31*, 15−20.

(6) Chiang, L.; Lu, B.; Castillo, I. Big Data Analytics in Chemical Engineering. *Annual Review of Chemical and Biomolecular Engineering.* **2017**, *8*, 63−85.

(7) Joost, N.; Kok, E. J. W. B.; Kosters, W. A.; van der Putten, P. Artificial Intelligence: Definition, Trends, Techniques, and Cases. *Knowledge for sustainable development: an insight into the Encyclopedia of life support systems.* **2002**, 1095−1107.

(8) Fathima Anjila, P. K. What is Artificial Intelligence?. In *Learning Outcomes of Classroom Research*; Karthikeyan, J., Hie, T. S., Jin, N. Y., Eds.; L Ordine Nuovo Publication, 2021; p 65.

(9) Gupta, R.; Srivastava, D.; Sahu, M.; Tiwari, S.; Ambasta, R. K.; Kumar, P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Molecular Diversity* **2021**, *25* (3), 1315−1360.

(10) Pramila P Shinde, S. S. In *A review of machine learning and deep learning applications.*, 2018 Fourth international conference on computing communication control and automation (ICCUBEA); IEEE: 2018; pp 1−6.

(11) Nuzzi, R. The Impact of Artificial Intelligence and Deep Learning in Eye Diseases: A Review. *Frontiers in Medicine* **2021**, *8*, 1.

(12) Zhang, L.; Tan, J.; Han, D.; Zhu, H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug discovery today* **2017**, *22* (11), 1680−1685.

(13) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559* (7715), 547−555.

(14) Vijay Kotu, B. D., Chapter 1 - Introduction. In *Data Science*, 2nd ed.; Vijay Kotu, B. D., Ed.; Elsevier: 2019; pp 1−18.

(15) Jordan, M. I.; Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349* (6245), 255−260.

(16) Mahesh, B. Machine Learning Algorithms - A Review. *International Journal of Science and Research (IJSR).[Internet]* **2020**, *9*, 381−386.

(17) Alpaydin, E. *Introduction to Machine Learning*; MIT Press: 2020.

(18) Ostberg, N. P; Zafar, M. A; Elefteriades, J. A Machine Learning: Principles and Applications for Thoracic Surgery. *European Journal of Cardio-Thoracic Surgery.* **2021**, *60* (2), 213−221.

(19) Sarker, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science.* **2021**, *2* (3), 1−21.

(20) Janiesch, C.; Zschech, P.; Heinrich, K. Machine Learning and Deep Learning. *Electronic Markets.* **2021**, *31* (3), 685−695.

(21) Dargan, S.; Kumar, M.; Ayyagari, M. R.; Kumar, G. A Survey of Deep Learning and Its Applications: a New Paradigm to Machine Learning. *Archives of Computational Methods in Engineering.* **2020**, *27* (4), 1071−1092.

(22) Najafabadi, M. M; Villanustre, F.; Khoshgoftaar, T. M; Seliya, N.; Wald, R.; Muharemagic, E. Deep Learning Applications and Challenges in Big Data Analytics. *Journal of big data.* **2015**, *2* (1), 1−21.

(23) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug discovery today.* **2018**, *23* (6), 1241−1250.

(24) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *nature* **2015**, *521* (7553), 436−444.

(25) Medsker, L. C. J. *Recurrent neural networks: design and applications*; CRC Press: 1999.

(26) Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; Chen, T. Recent Advances in Convolutional Neural Networks. *Pattern Recognition* **2018**, *77*, 354−377.

(27) Wu, Z; et al. A Comprehensive Survey on Graph Neural Networks. *IEEE transactions on neural networks and learning systems.* **2020**, *32* (1), 4−24.

(28) Arus-Pous, J.; Awale, M.; Probst, D.; Reymond, J.-L. Exploring Chemical Space with Machine Learning. *Chimia (Aarau)* **2019**, *73* (12), 1018−1023.

(29) Zhao, L.; Chen, Z.; Hu, Y.; Min, G.; Jiang, Z. Distributed feature selection for efficient economic big data analysis. *IEEE Transactions on Big Data* **2018**, *4* (2), 164−176.

(30) Marino, S.; Xu, J.; Zhao, Y.; Zhou, N.; Zhou, Y.; Dinov, I. D. Controlled feature selection and compressive big data analytics: Applications to biomedical and health studies. *PLoS One* **2018**, *13* (8), No. e0202674.

(31) Fan, F. J.; Shi, Y. Effects of data quality and quantity on deep learning for protein-ligand binding affinity prediction. *Bioorg. Med. Chem.* **2022**, *72*, 117003.

(32) Nandy, A.; Duan, C.; Kulik, H. J Audacity of huge: overcoming challenges of data scarcity and data quality for machine learning in computational materials discovery. *Current Opinion in Chemical Engineering* **2022**, *36*, 100778.

(33) Artrith, N.; Butler, K. T.; Coudert, F.-X.; Han, S.; Isayev, O.; Jain, A.; Walsh, A. Best practices in machine learning for chemistry. *Nature Chem.* **2021**, *13* (6), 505−508.

(34) Vogt, M. Using deep neural networks to explore chemical space. *Expert Opinion on Drug Discovery* **2022**, *17* (3), 297−304.

(35) Tânia, F. G. G.; Cova, A. A. P. Deep learning for deep chemistry: optimizing the prediction of chemical patterns. *Frontiers in chemistry* **2019**, *7*, 809.

(36) Garg, N.; et al. FlavorDB: a Database of Flavor Molecules. *Nucleic acids research* **2018**, *46* (D1), D1210−D1216.

(37) FlavorDB. https://cosylab.iiitd.edu.in/flavordb/ (accessed May 11, 2022).

(38) FooDB Version 1.0. https://foodb.ca (accessed May 11, 2022).

(39) McKillop, K.; Harnly, J.; Pehrsson, P.; Fukagawa, N.; Finley, J. FoodData Central, USDA's Updated Approach to Food Composition Data Systems. *Current Developments in Nutrition.* **2021**, *5*, 596−596.

(40) Fukagawa, N. K; McKillop, K.; Pehrsson, P. R; Moshfegh, A.; Harnly, J.; Finley, J. USDA's FoodData Central: what is it and why is it needed today? *American Journal of Clinical Nutrition.* **2022**, *115* (3), 619−624.

(41) FoodData Central. https://fdc.nal.usda.gov (accessed May 11, 2022).

(42) Welty, C. Ontology Research. *AI magazine.* **2003**, *24* (3), 11−12.

(43) Hazman, M.; R. El-Beltagy, S.; Rafea, A. A Survey of Ontology Learning Approaches. *International Journal of Computer Applications.* **2011**, *22* (9), 36−43.

(44) FoodOn: A farm to fork ontology. https://foodon.org (accessed May 11, 2022).

(45) Dooley, D. M.; et al. FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. *npj Science of Food.* **2018**, *2* (1), 1−10.

(46) Wiener, A.; Shudler, M.; Levit, A.; Niv, M. Y. BitterDB: a database of bitter compounds. *Nucleic acids research* **2012**, *40* (D1), D413−D419.

(47) Fritz, F.; Preissner, R.; Banerjee, P. VirtualTaste: a web server for the prediction of organoleptic properties of chemical compounds. *Nucleic Acids Res.* **2021**, *49* (W1), W679−W684.

(48) Ahmed, J.; Preissner, S.; Dunkel, M.; Worth, C. L.; Eckert, A.; Preissner, R. SuperSweet—a resource on natural and artificial sweetening agents. *Nucleic acids research* **2011**, *39*, D377−D382.

(49) FlavorDB2. https://cosylab.iiitd.edu.in/flavordb2/ (accessed May 11, 2022).

(50) Grover, N.; et al. FlavorDB2: An Updated Database of Flavor Molecules. *arXiv preprint arXiv:2205.05451*, 2022.

(51) Naveja, J. J.; Rico-Hidalgo, M. P.; Medina-Franco, J. L. Analysis of a Large Food Chemical Database: Chemical Space, Diversity, and Complexity. *F1000Research.* **2018**, *7*, 993.

(52) Masand, V. H.; Sk, M. F.; Kar, P.; Rastija, V.; Zaki, M. E.A. Identification of Food Compounds as Inhibitors of SARS-CoV-2 Main Protease Using Molecular Docking and Molecular Dynamics Simulations. *Chemometrics and Intelligent Laboratory Systems.* **2021**, *217*, 104394.

(53) Shin, S. H. OptNCMiner: a Deep Learning Approach for the Discovery of Natural Compounds Modulating Disease-Specific Multi-targets. *BMC Informatics* **2022**, *23*, 218.

(54) Ribeiro, A. A.; Sachine, M. On the optimal separating hyperplane for arbitrary sets: A generalization of the SVM formulation and a convex hull approach. *Optimization* **2022**, *71* (1), 213−226.

(55) Elbadawi, M.; Gaisford, S.; Basit, A. W. Advanced Machine-Learning Techniques in Drug Discovery. *Drug Discovery Today.* **2021**, *26* (3), 769−777.

(56) Wu, Y. Bioinformatics analysis to screen for critical genes between survived and non-survived patients with sepsis. *Molecular Medicine Reports.* **2018**, *18* (4), 3737−3743.

(57) Karamizadeh, S.; et al. Advantage and Drawback of Support Vector Machine Functionality. In *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*; 2014; pp 63−65.

(58) Yu, Y.-H.; et al. Ensemble modeling with machine learning and deep learning to provide interpretable generalized rules for classifying CNS drugs with high prediction power. *Briefings in Bioinformatics.* **2022**, *23* (1), 1−13.

(59) Patel, H. H.; Prajapati, P. Study and analysis of decision tree based classification algorithms. *International Journal of Computer Sciences and Engineering.* **2018**, *6* (10), 74−78.

(60) Esmaily, H.; et al. A comparison between decision tree and random forest in determining the risk factors associated with type 2 diabetes. *Journal of Research in Health Sciences.* **2018**, *18* (2), 412.

(61) Michele Fratello, R. T. Decision Trees and Random Forests. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics* **2018**, *1*, 374−383.

(62) Resende, P. A. A.; Drummond, A. C. A survey of random forest based methods for intrusion detection systems. *ACM Computing Surveys (CSUR).* **2019**, *51* (3), 1−36.

(63) Salehinejad, H.; et al. Recent Advances in Recurrent Neural Networks. *arXiv preprint arXiv:1801.01078*, 2017.

(64) Datta, D.; David, P. E.; Mittal, D.; Jain, A. Neural Machine Translation Using Recurrent Neural Network. *International Journal of Engineering and Advanced Technology*. **2020**, *9* (4), 1395−1400.

(65) De Mulder, W.; Bethard, S.; Moens, M.-F. A survey on the application of recurrent neural networks to statistical language modeling. *Computer Speech & Language* **2015**, *30* (1), 61−98.

(66) Li, N.; et al. In *Applications of recurrent neural network language model in offline handwriting recognition and word spotting*; 2014 - 14th International Conference on Frontiers in Handwriting Recognition; IEEE: 2014; pp 134−139.

(67) Arisoy, E.; et al. In *Bidirectional recurrent neural network language models for automatic speech recognition*., 2015 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE: 2015; pp 5421−5425.

(68) Zachary, C.; et al. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, 2015.

(69) Yamashita, R.; Nishio, M.; Do, R. K. G.; Togashi, K. Convolutional Neural Networks: an Overview and Application in Radiology. *Insights into imaging*. **2018**, *9* (4), 611−629.

(70) Chen, J.-H.; Tseng, Y. J. Different Molecular Enumeration Influences in Deep Learning: an Example Using Aqueous Solubility. *Briefings in Bioinformatics*. **2021**, *22* (3), bbaa092.

(71) Samira, E.; et al. In *Recurrent Neural Networks for Emotion Recognition in Video*; Proceedings of the 2015 ACM on international conference on multimodal interaction, 2015; pp 467−474.

(72) Li, H.; et al. Deep learning in breast cancer risk assessment: evaluation of convolutional neural networks on a clinical dataset of full-field digital mammograms. *Journal of Medical Imaging*. **2017**, *4* (4), 041304.

(73) Zhavoronkov, A. *Artificial Intelligence for Drug Discovery, Biomarker Development, and Generation of Novel Chemistry*. ACS Publications: 2018; Vol. *15*, pp 4311−4313.

(74) Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. Graph Neural Networks: a Review of Methods and Applications. *AI Open*. **2020**, *1*, 57−81.

(75) Zhang, S.; Tong, H.; Xu, J.; Maciejewski, R. Graph Convolutional Networks: a Comprehensive Review. *Computational Social Networks*. **2019**, *6* (1), 1−23.

(76) Zitnik, M.; Agrawal, M.; Leskovec, J. Modeling Polypharmacy Side Effects with Graph Convolutional Networks. *Bioinformatics*. **2018**, *34* (13), i457−i466.

(77) Lee, S.; Lee, M.; Gyak, K.-W.; Kim, S. D.; Kim, M.-J.; Min, K. Novel Solubility Prediction Models: Molecular Fingerprints and Physicochemical Features vs Graph Convolutional Neural Networks. *ACS omega*. **2022**, *7* (14), 12268−12277.

(78) Park, D.; et al. FlavorGraph: a large-scale food-chemical graph for generating food representations and recommending food pairings. *Scientific Reports*. **2021**, *11* (1), 1−13.

(79) Salvador, A.; et al. In *Learning Cross-Modal Embeddings for Cooking Recipes and Food Images*; Proceedings of the IEEE conference on computer vision and pattern recognition; 2017; pp 3020−3028.

(80) Veselkov, K.; et al. HyperFoods: Machine intelligent mapping of cancer-beating molecules in foods. *Scientific Reports*. **2019**, *9* (1), 1−12.

(81) Dong, Y.; et al. In *metapath2vec: Scalable representation learning for heterogeneous networks*; Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 2017; pp 135−144.

(82) Park, D.; et al. KitcheNette: Predicting and Recommending Food Ingredient Pairings using Siamese Neural Networks. *arXiv preprint arXiv:1905.07261*; 2019.

(83) Rahman, M. M.; Vadrev, S. M.; Magana-Mora, A.; Levman, J.; Soufan, O. A Novel Graph Mining Approach to Predict and Evaluate Food-Drug Interactions. *Scientific Reports*. **2022**, *12* (1), 1−16.

(84) Wishart, D. S. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34* (90001), D668−D672.

(85) Peña-Castillo, A.; et al. Chemoinformatics in food science. In *Applied chemoinformatics: achievements and future opportunities*, 1st ed.; Engel, T., Ed.; Wiley-VCH: 2018; pp 501−525.

(86) Dunkel, A.; Hofmann, T.; Di Pizio, A. In Silico Investigation of Bitter Hop-Derived Compounds and Their Cognate Bitter Taste Receptors. *J. Agric. Food Chem.* **2020**, *68* (38), 10414−10423.

(87) Dagan-Wiener, A.; Di Pizio, A.; Nissim, I.; Bahia, M. S; Dubovski, N.; Margulis, E.; Niv, M. Y BitterDB: Taste Ligands and Receptors Database in 2019. *Nucleic Acids Res.* **2019**, *47* (D1), D1179−D1185.

(88) van der Maaten, L.; et al. Visualizing Data Using t-SNE. *Journal of Machine Learning Research*. **2008**, *9* (11), 2579−2605.