Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biotechnology

Machine learning applications to improve flavor and nutritional content of horticultural crops through breeding and genetics

Luís Felipe V Ferrão[1], Rakshya Dhakal[2], Raquel Dias[3], Denise Tieman[1], Vance Whitaker[1,2], Michael A Gore[4], Carlos Messina[1,2] and Márcio F R Resende Jr[1,2]

Over the last decades, significant strides were made in understanding the biochemical factors influencing the nutritional content and flavor profile of fruits and vegetables. Product differentiation in the produce aisle is the natural consequence of increasing consumer power in the food industry. Cotton-candy grapes, specialty tomatoes, and pineapple-flavored white strawberries provide a few examples. Given the increased demand for flavorful varieties, and pressing need to reduce micronutrient malnutrition, we expect breeding to increase its prioritization toward these traits. Reaching this goal will, in part, necessitate knowledge of the genetic architecture controlling these traits, as well as the development of breeding methods that maximize their genetic gain. Can artificial intelligence (AI) help predict flavor preferences, and can such insights be leveraged by breeding programs? In this Perspective, we outline both the opportunities and challenges for the development of more flavorful and nutritious crops, and how AI can support these breeding initiatives.

Addresses
[1] Horticultural Sciences Department, University of Florida, Gainesville, FL, United States
[2] Plant Breeding Graduate Program, University of Florida, Gainesville, FL, United States
[3] Microbiology and Cell Science Department, University of Florida, Gainesville, FL, United States
[4] Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY, United States

Corresponding author: Resende Jr, Márcio F R (mresende@ufl.edu)

Current Opinion in Biotechnology 2023, **83**:102968

This review comes from a themed issue on **Plant Biotechnology**

Edited by **Alisdair Fernie** and **Jianbing Yan**

For complete overview of the section, please refer to the article collection, "Plant Biotechnology (2023)"

Available online 27 July 2023

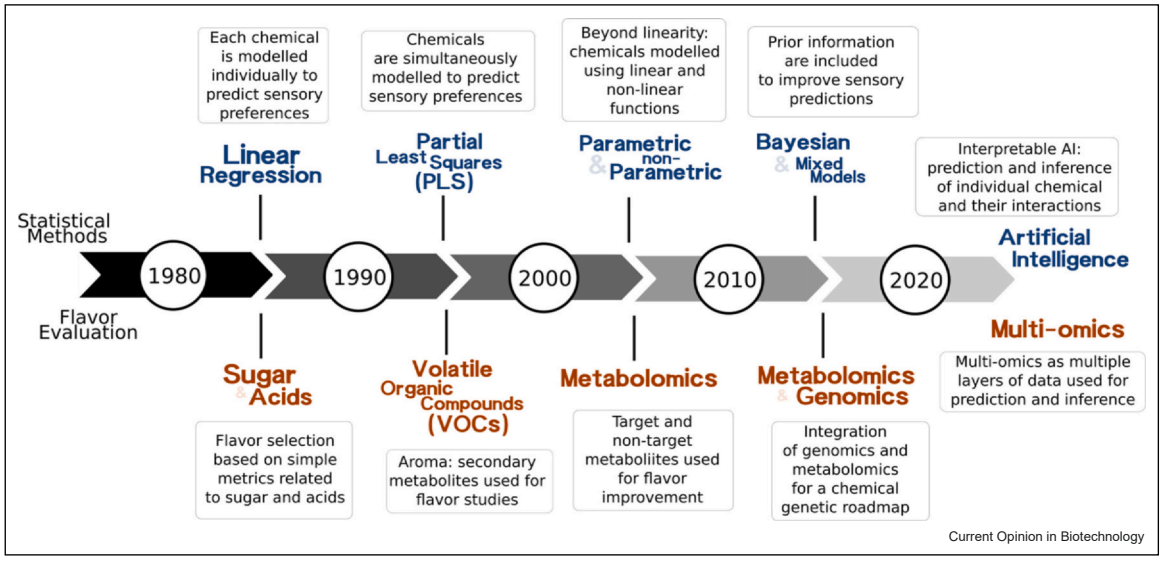
https://doi.org/10.1016/j.copbio.2023.102968

Introduction

Over the last few decades, researchers have made significant strides in understanding the biochemical factors that influence the nutritional content and flavor profile of our fruits and vegetables. From the point of view of crop improvement, flavor, and nutrition are related in the sense that these are complex, consumer-oriented traits for which the breeding typically aims at changing metabolite levels in the plant. Product differentiation in the produce aisle is the natural consequence of increasing consumer power in the food industry. Cotton-candy grapes, specialty tomatoes, and pineapple-flavored white strawberries provide a few examples. Given the increased demand for flavorful varieties [1], and pressing need to reduce micronutrient malnutrition [2], we expect breeding to increase its prioritization toward these traits, in addition to all other grower-oriented trait priorities. Reaching this aspirational goal will, in part, necessitate the imparting of deep biological knowledge of the genetic architecture controlling these traits, as well as the development of breeding methods that can maximize their genetic gain. Can artificial intelligence (AI) help predict the desired flavor niches of the future, and can such insights be leveraged by breeding programs to accentuate the desired aromatic profiles? In this Perspective, we outline both the opportunities and challenges for the development of more flavorful and nutritious fruits and vegetables, drawing attention to the fundamental role of combining multi-omics and AI to support breeding initiatives. Multi-omics projects involve the coordinated and integrated study of multiple omics datasets, such as genomics, transcriptomics, proteomics, and metabolomics. For flavor analyses, the inclusion of layers of information has the potential to greatly enhance our understanding of the mechanisms underlying our food choices, in particular, how biological processes are interconnected in a holistic way.

Crop flavor and nutritional methods and analyses have evolved over the years. It was the rapid evolution of instrumental analyses and statistical methods that has led to significant advancement in our ability to identify the key components affecting our preferences (Figure 1).

Figure 1



Statistical methods and flavor evaluation have evolved during the years. While in the 1980s most of the breeding programs focused on simple metrics (sugar and acid content) to assess flavor preferences, in the past 20 years, contemporaneous breeding programs have been stacking multiple layers of information by collecting flavor information using multiple sources of data. Metabolomics associated with genomics, for example, have been used as the state-of-the-art to predict consumer preferences and point out key attributes that impact our flavor predictions. With the incorporation of multi-omics information in flavor studies, datasets have grown in size, with the number of predictor variables often exceeding the number of observations. This raises challenges when fitting linear regression models, specifically due to the need to regularize parameter estimates to avoid overfitting. Numerous approaches have been proposed, including many based on penalized least-squares criterion, and others based on machine learning and Bayesian approaches. These many different methods vary in their choice of penalty function or prior distribution for the regression coefficients and in their choice of computational algorithms. The use of AI in the format of neural network architecture is considered the next frontier, with the possibility to better explore the datasets for inference and prediction.

Improving fruit and vegetable flavor

The genetic improvement of flavor is typically not a priority in plant breeding programs, which reduces the rate of progress obtained for this trait. In some cases, flavor has even deteriorated over the years of breeding for agronomic traits such as yield, shelf life, and disease resistance, causing increasing dissatisfaction among consumers [3,4]. Unfortunately, breeding for improved flavor has proven difficult, since it is a complex trait involving many flavor and aroma compounds and many genetic loci [1]. Recent research has focused on understanding the biochemical and genetic bases of flavor in various fruits, including tomato, strawberry, blueberry, apple, grape, peach, kiwi, and citrus fruits [5–12]. Each fruit has its own unique flavor profile, a unique combination of sugars, acids, and aroma volatiles. Typically, sugars include glucose, fructose, and sucrose, while common acids are citric, malic, ascorbic, and tartaric acids. Volatiles vary widely among fruit species and contribute to the unique aroma and flavor profiles of each. Compounds that contribute consumer-favored flavor attributes to some fruits can negatively affect flavor in other fruit species. For instance, acetate esters negatively affect tomato flavor, but are major positive contributors to flavor of melon, peach, and strawberry [1,13,14]. Gene function of similar genes can also be different among species. This diversity can make gene identification more difficult, as pathways to the same volatile diverge among species. For example, the biochemical pathway to 2-phenylethanol is different in tomato, where phenylalanine is converted to phenethylamine by a decarboxylase followed by conversion to phenylacetaldehyde, than in petunia or rose flowers, where phenylalanine is converted directly to phenylacetaldehyde by phenylacetaldehyde synthase [15–17].

Improving plant-based nutritional content

Globally, an estimated two billion people suffer from deficiencies in one or more essential micronutrients that include vitamins and minerals. Nutritional deficiencies tend to be elevated in communities that subsist on staple crops with insufficient levels of micronutrients as a primary source of calories. Biofortification — the improvement of crop nutritional quality through agronomic or genetic approaches — has been advanced as a sustainable way to improve human health and nutrition [18]. Even though we are converging to a near-complete understanding of the biosynthetic pathways that synthesize vitamins in plants [19], the causal genes and alleles responsible for natural variation in vitamin levels have yet to be fully cataloged for any plant system. Relatedly, while the genetic architectures underlying natural variation in mineral composition for tissues of arabidopsis [*Arabidopsis thaliana* (L.) Heynh.] have begun to be

unraveled [20], the translation of this knowledge to reveal the genetic basis of mineral composition in crops has been incomplete and primarily focused on cereals [21–25].

Identifying causal genes affecting flavor perception and nutritional quality

With these challenges in mind, one core approach to learn about flavor and nutritional biology is to determine the biochemical compounds, their biosynthetic pathways, and their regulatory network affecting crop flavor and nutritional quality. This information can guide trait prioritization, can lead to marker-assisted selection targets, and contribute to our overall understanding of the trait. Genome-wide association studies (GWAS) and quantitative trait loci (QTL) mapping studies have been the main statistical methods to identify candidate genes affecting these pathways of interest [26–29]. For example, in the most comprehensive assessments of natural variation for a vitamin in any plant, the genetic control of vitamin E (tocopherols and tocotrienols, collectively known as tocochromanols) and provitamin A (carotenoids) in maize grain was elucidated to near completion via joint linkage and GWAS in the 5000-line U.S. maize (*Zea mays* L.) nested association mapping panel [28,29]. The large-effect causal genes identified in these two studies are highly conserved throughout the plant kingdom and encode activities in precursor and core biosynthetic pathways that have been well characterized in model plants. Contrastingly, the *Orange* gene that is associated with higher carotenoids in nonleaf tissue of cauliflower [30] and carrot [31] has never been identified as a causal loci for natural variation in grain carotenoids of maize and other cereals, indicating the limitation of orthologous comparisons. When causal genes are unknown due to low-resolution genetic mapping, the decision-making process of selecting candidate genes can be limited by biased approaches [32], but emerging machine learning approaches that use a multitude of genomic features have the potential to better prioritize genes likely to be causal as training sets become larger [33,34].

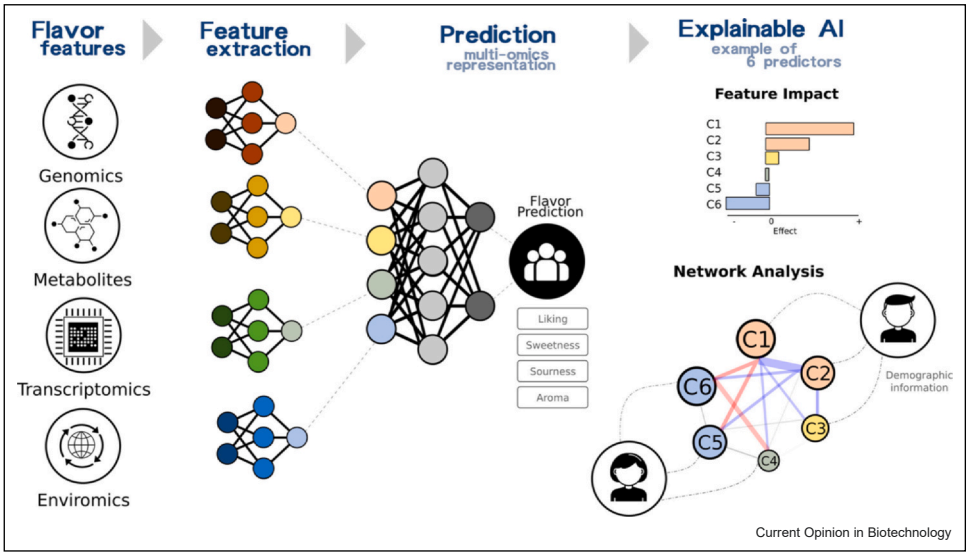
While target biosynthetic pathways have already been well characterized and understood for model crop species such as tomatoes, this is not the case for many fruits and vegetables for which genomic and metabolic resources might not be available. Furthermore, even in well-characterized pathways, very few regulatory genes affecting the expression of these biosynthetic genes are known. We expect that GWAS will continue to have a key role in identifying the genes that are responsible for the accumulation of these compounds, and we expect the integration of additional multilayer omics data to contribute to this task. Intermediate molecular phenotypes or endophenotypes span multiple levels of biological organization between DNA genotype and terminal (target) phenotypes, offering independent insights into the causal mechanisms of phenotypic changes not revealed by genetic markers alone. Implicating the importance of regulatory variation in controlling quantitative variation for metabolite seed traits, the use of mRNA expression level as an endophenotype in transcriptome-wide association studies (TWAS) resulted in the identification of candidate causal genes associated with carotenoids or tocochromanols in fresh sweet corn and physiologically mature maize kernels [35–37] and avenanthramides — a group of phenolic antioxidants — in oat seed [38]. To investigate how selection altered natural variation in the tomato fruit metabolome [39], correlated changes in gene expression with metabolites and integrated them with expression QTL and metabolite-based GWAS results, showing that five loci associated with a reduction in steroidal glycoalkaloids were targets of selection to produce less-bitter fruits. In addition to the genetic dissection of metabolite traits, statistical models that incorporate transcriptomic and/or metabolomic data in addition to genome-wide markers have been shown to enhance the prediction of vitamin and fatty acid seed traits [35,40,41].

Predicting flavor perception and nutritional quality using multi-omics data

Flavor is affected by growing conditions, is expensive to assay, and flavor perception varies among the individuals tasting the fruit [42]. Therefore, a machine learning approach would be valuable for predicting overall flavor from biochemical or genetic data. Colatonio et al. [43] evaluated the performance of different machine learning algorithms to predict consumer flavor perceptions of blueberries and tomatoes using fruit chemical composition as predictors. These models offer a flexible framework for incorporating complex dependencies and prior knowledge into the prediction process and can provide more robust and accurate predictions than traditional linear models. For another example, in strawberry, a large sensory–chemical study was conducted to identify chemical drivers of consumer preference. Machine learning models, including 113 volatile compounds, explained at least 25% more variation in sweetness than models incorporating sugars and acids only [44]. Both these results [43,44] demonstrated that not only are volatile compounds critical for flavor perception, but they are critical also for sweetness that is the primary driver of consumer preference in strawberry [44]. This field is rapidly evolving, and we expect that new methods and larger datasets will continue to improve our ability to select flavorful and nutritional varieties that can hit the market.

To address the issue of high-dimensional data in flavor prediction, modern flavor studies have shifted from

Figure 2



Schematic representation of how AI models can integrate several datatypes (e.g. multi-omics, environment, etc.) into a single prediction model for identifying flavor attributes. The example consists of one layer of subnetworks, one for each data type, that will learn meaningful features from their respective datatypes. Each subnetwork is implemented separately with its own parameters and architecture that performs better for its respective data type. Each subnetwork is then connected to one neural network that merges all features and learns to predict a target phenotype or outcome from them. For the feature extraction or feature learning layer, CNNs are commonly used for grid-like data structure such as images, while recurrent neural network (RNN) is often used for sequence data such as a DNA sequence or time-series data. CNNs are the most often-utilized NN architecture across metabolomics data as well. In this example, the prediction layer is represented by a fully connected NN (e.g. feed-forward neural network), which connects all layers of subnetwork features extracted from all datatypes into a single integrative architecture that outputs the predicted flavor attributes.

traditional linear regression models and focused on using Bayesian and mixed model approaches for predictions. In the field of AI, deep learning algorithms such as artificial neural networks (ANN) are well suited to model the complex, nonlinear relationships between genotypes and phenotypes [45–47]. More specialized ANNs such as convolutional neural networks (CNNs) and long–short-term memory networks are well known for being able to capture local features of phrases as well as global and temporal sentence semantics [48–50] (Figure 2). Additionally, ANNs have shown to be particularly useful for integrating data from different sources without the need for feature engineering (e.g. genotype-by-environment interactions) [51]. However, ANNs have not been fully explored in the field of genomic selection, and deeper model optimization and testing may be necessary to explore their full potential.

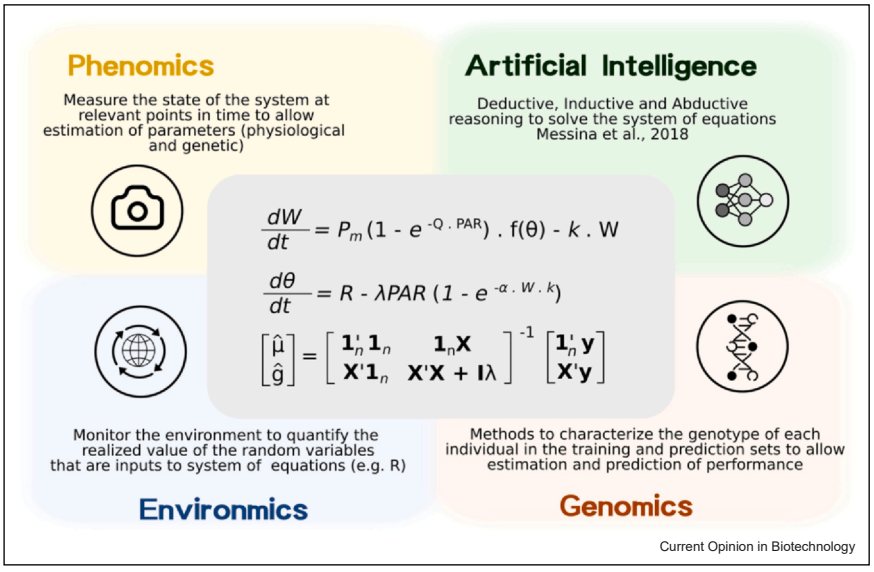
Explainable artificial intelligence unifies prediction and inference

Recently, there has been debate on the interpretability of complex AI models [52–54]. AI models have been labeled as 'black boxes', models that produce accurate predictions based on a dataset, but the end user does not know how these predictions are made. Contrary to widely held concerns, these methods are not 'black box' predictors: pre-/post hoc feature selection and feature importance quantification techniques allow for the identification of input features contributing to predictive accuracy [55–57]. SHapley Additive exPlanations (SHAP), [56] for example, is an explainable AI (XAI) method that can be used for interpreting the prediction of any model by quantifying the contribution of each feature to the prediction, and ranking predictor features based on their importance for the model's outputs. Additionally, SHAP can be applied simultaneously across multiple input features for identifying global feature relationships and dependencies. XAI techniques have been applied for the analysis and discovery of genome-wide associations between genotypes and phenotypes [58], as well as for identifying gene–gene and gene–environment interactions [59]. These XAI techniques can play a key role in shedding new light on how different omics attributes correlate and are influenced by environmental effects (e.g. demographic information).

Collecting more data and 'boosting model power'

The ability to analyze and learn from data is limited by the quantity of information that is fed into the AI-powered model. By harnessing big data resources, AI models can make more informed predictions, but well-defined data collection plans and data structure must be in place before AI implementation, since models are only as predictive as the data that we use to train them.

Figure 3



Dynamical CGM are cognitive mathematical representations of the physiological processes underpinning resource use, transformation efficiencies, and yield in horticultural crops [66]. For each hour in the growing season, carbon assimilation, and its allocation to organ growth and respiration is calculated. This time dependency determines temporal patterns of carbon, water and nutrient demand, and supply. Competition for carbon determines the distribution of fruit size and number [67]. Discrepancy between water supply and demands affects turgor, organ expansion, and thus light interception, fertilization, fruit size, and yield [68]. Integration of this knowledge with genomic selection proved useful in breeding [74]. Prediction accuracy improvement relative to GS along increase with increasing complexity of the genotype x trait x environment system [75,76]. Advances in chemical phenomics and the ability to identify prediction networks consistent with the underpinning biochemistry of flavor [43] suggests that a process-based approach to prediction can further increase prediction skill in fruit flavor. While modules to simulate the biochemistry of flavor will be needed, CGMs provide the framework to build such module, and the key inputs such as duration of fruit development, carbon fluxes, and water status of the plant and the fruit. Metabolic flux analyses can provide the toolkit to model the metabolism of flavor based on these inputs and current knowledge of the biochemistry of flavor [69,77]. Current CGM uses similar approaches to estimate respiration costs [71,72] and carbon assimilation [73]. A degree of empiricism and assumptions of steady state will be required to address the limitations of incomplete knowledge. However, as in the case of maize breeding, the integration of current knowledge and estimation procedures such as GS can enable increasing prediction skill, genotype x environment interactions, and thus harness knowledge to hasten genetic gain for yield and flavor.

With the continuing development of low-cost, non-destructive imagers, and spectrometers, it is becoming more feasible to eventually analyze the environmentally- and developmentally labile spectral signatures of plant tissues with machine learning algorithms for phenomic prediction and selection [60,61] of potentially correlated quality traits [62]. Similarly, targeted metabolomics has reduced in price, enabling the collection of larger sample sizes for a fixed budget. In addition to collecting more data, there are two common techniques for boosting model power that have not been explored to its full potential in the field of plant genomics: transfer learning and data augmentation. With transfer learning techniques, AI models can store knowledge gained while solving one problem and apply it to a different but related problem. For example, AI models that were already trained for predicting flavor using metabolomics data can be fine-tuned for predicting flavor in a different species, using a different dataset, and performing better than the same trained from scratch.

In data augmentation, AI can be utilized as a data preprocessing tool, in a process of artificially increasing the amount of data by generating new data points from existing data. The new *in silico* samples are used to represent the latent space of the original data to amplify the dataset. In genomics, the augmentation can be obtained by simulating genomic breeding, selection, and recombination events, resulting in new populations of unrelated or admixed genomes. Generative adversarial networks (GANs) are a class of ANN architecture that specialized for data augmentation. Generative models compute a distribution of the data itself, generate new examples, and estimate the likelihood of a new given example existing in the dataset. For example, models that predict the next word in a sequence are typically generative models. GANs have been

extremely successful for data augmentation in image classification problems [63,64], but have not been fully explored in the field of genomics.

Flavor and nutrition into dynamical crop growth and development models

Finally, we have so far approached the use of machine learning to learn and predict better genetics. However, flavor and nutritional content are also a function of the environment in which the plants are grown. For example, a negative genetic correlation between sugars and yield has been observed in some environments for strawberry [65]. The strength of this trade-off suggests that it is controlled by basic physiological constraints, particularly under high temperatures. The field of crop modeling can be leveraged to estimate and account for metabolic flux, combined with information on sugar biosynthesis pathways and sugar transport mechanisms. This approach could potentially identify yield components with minimal effects of fruit sugar content or otherwise inform ways to weaken or even break such correlations.

Dynamical crop growth and development models (CGM) are cognitive mathematical representations of the physiological processes underpinning resource use, transformation efficiencies, and yield in horticultural crops [66]. For each hour in the growing season, carbon assimilation and its allocation to organ growth and respiration is calculated. This time dependency determines temporal patterns of carbon, water and nutrient demand, and supply. Competition for carbon determines the distribution of fruit size and number [67]. Discrepancy between water supply and demands affects turgor, organ expansion, and thus light interception, fertilization, fruit size, and yield [68]. Advances in chemical phenomics, the application of GS for flavor, and the ability to identify prediction networks consistent with the underlying biochemistry of flavor [43] suggests that a process-based dynamical approach to prediction can further increase prediction skill. While modules to simulate the biochemistry of flavor will be needed, CGMs provide the framework to build such a module, and for including the key inputs such as duration of fruit development, carbon fluxes, and water status of the plant and the fruit. Metabolic flux analyses can provide the toolkit to model the metabolism of flavor based on these inputs and current knowledge of the biochemistry of flavor [69,70]. Current CGM uses similar approaches to estimate respiration costs [71,72] and carbon assimilation [73]. A degree of empiricism and assumptions of steady state will be required to address the limitations of incomplete knowledge. However, as in the case of maize breeding, where this area is mostly advanced, the integration of current knowledge and estimation procedures such as GS can enable increasing prediction skill, genotype x environment interactions, and thus harness knowledge to hasten genetic gain for yield and flavor (Figure 3).

Conclusions and perspectives applied to plant breeding

In summary, we envision that adoption of these new methodologies discussed in this paper will enable breeding programs to answer more sophisticated questions, including the influence of demographic information on our predictions, the impact of genotype-by-environment interactions, and associate flavor to our healthy lifestyles. As more knowledge is learned about these traits, synthetic biology, and metabolic engineering open new avenues for the redesign or *de novo* construction of gene-regulatory circuits and altering the form and function of metabolites for virtually any plant species. Such approaches could enable the modification of specialized metabolites to enhance their stability and change the cellular location in which they are sequestered, as has been proposed for carotenoids in plants [78]. With a deep collection of publicly available multi-omics data for an ever-increasing number of plant species, the integration of machine learning and genome-scale metabolic models offers the potential to identify -omics factors with high importance in the light of a mechanistic framework [79]. Existing software offer functions to potentially integrate genome-scale metabolic models and crop growth models [80], enabling the modeling of nutritional phenotypes across multiple scales to better understand how metabolism is shaped by genotype, environment, and their interaction over the life history of a plant. These activities will ultimately require high-throughput instrumentation combined with robotics to facilitate sampling of tissues at informative developmental time points and post-sampling processing if performing them on large plant populations for nonvolatile metabolites that cannot be scored with nondestructive analytical methods [77].

Data Availability

No data were used for the research described in the article.

Declaration of Competing Interest

The authors declare no competing interests.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. • Tieman D, *et al.*: **A chemical genetic roadmap to improved tomato flavor**. *Science* 2017, **355**:391-394.
Provides example and discussion on how to improve tomato flavor using genomics and molecular biology.

2. • Van Der Straeten D, *et al.*: **Multiplying the efficiency and impact of biofortification through metabolic engineering**. *Nat Commun* 2020, **11**:5203.

Provides a review and plan for staple crop biofortification.

3. Klee HJ: **Improving the flavor of fresh fruits: genomics, biochemistry, and biotechnology**. *New Phytol* 2010, **187**:44-56.

4. Folta KM, Klee HJ: **Sensory sacrifices when we mass-produce mass produce**. *Hortic Res* 2016, **3**:16032.

5. Aharoni A, *et al*.: **Gain and loss of fruit flavor compounds produced by wild and cultivated strawberry species**. *Plant Cell* 2004, **16**:3110-3131.

6. Cao X, *et al*.: **Transcriptional and epigenetic analysis reveals that NAC transcription factors regulate fruit flavor ester biosynthesis**. *Plant J* 2021, **106**:785-800.

7. Peng B, *et al*.: **Different roles of the five alcohol acyltransferases in peach fruit aroma development**. *J Am Soc Hortic Sci* 2020, **145**:374-381.

8. Espino-Díaz M, *et al*.: **Biochemistry of apple aroma: a review**. *Food Technol Biotechnol* 2016, **54**:375-394.

9. Ferrao LFV, *et al*.: **Terpene volatiles mediates the chemical basis of blueberry aroma and consumer acceptability**. *Food Res Int* 2022, **158**:111468.

10. Kumar S, *et al*.: **Genome-wide scans reveal genetic architecture of apple flavour volatiles**. *Mol Breed* 2015, **35**:1-16.

11. Liao L, *et al*.: **Unraveling a genetic roadmap for improved taste in the domesticated apple**. *Mol Plant* 2021, **14**:1454-1471.

12. Sater HM, *et al*.: **A review of the fruit volatiles found in blueberry and other Vaccinium species**. *J Agric Food Chem* 2020, **68**:5777-5786.

13. Tieman D, *et al*.: **The chemical interactions underlying tomato flavor preferences**. *Curr Biol* 2012, **22**:1035-1039.

14. Farcuh M, *et al*.: **Sensory, physicochemical and volatile compound analysis of short and long shelf-life melon (*Cucumis melo* L.) genotypes at harvest and after postharvest storage**. *Food Chem: X* 2020, **8**:100107.

15. Kaminaga Y, *et al*.: **Plant phenylacetaldehyde synthase is a bifunctional homotetrameric enzyme that catalyzes phenylalanine decarboxylation and oxidation**. *J Biol Chem* 2006, **281**:23357-23366.

16. Roccia A, *et al*.: **Biosynthesis of 2-phenylethanol in rose petals is linked to the expression of one allele of RhPAAS**. *Plant Physiol* 2019, **179**:1064-1079.

17. Tieman D, *et al*.: **Tomato aromatic amino acid decarboxylases participate in synthesis of the flavor volatiles 2-phenylethanol and 2-phenylacetaldehyde**. *Proc Natl Acad Sci* 2006, **103**:8287-8292.

18. Bouis HE, Welch RM: **Biofortification — a sustainable agricultural strategy for reducing micronutrient malnutrition in the global south**. *Crop Sci* 2010, **50**:S-20-S-32.

19. • Fitzpatrick TB, *et al*.: **Vitamin deficiencies in humans: can plant science help?** *Plant Cell* 2012, **24**:395-414.
Paper describing the known importance of vitamins in human health and current knowledge on their metabolism in plants.

20. Huang X-Y, Salt DE: **Plant ionomics: from elemental profiling to environmental adaptation**. *Mol Plant* 2016, **9**:787-797.

21. Wu D, *et al*.: **High-resolution genome-wide association study pinpoints metal transporter and chelator genes involved in the genetic control of element levels in maize grain**. *G3* 2021, **11**:jkab059.

22. Ziegler G, et al.: Elemental accumulation in kernels of the maize nested association mapping panel reveals signals of gene by environment interactions. *BioRxiv*; 2017: https://doi.org/10.1101/164962.

23. Yang M, *et al*.: **Genome-wide association studies reveal the genetic basis of ionomic variation in rice**. *Plant Cell* 2018, **30**:2720-2740.

24. Shakoor N, *et al*.: **Integration of experiments across diverse environments identifies the genetic determinants of variation in *Sorghum bicolor* seed element composition**. *Plant Physiol* 2016, **170**:1989-1998.

25. Cobb JN, *et al*.: **Genetic architecture of root and shoot ionomes in rice (*Oryza sativa* L.)**. *Theor Appl Genet* 2021, **134**:2613-2637.

26. Ferrão LFV, *et al*.: **Insights into the genetic basis of blueberry fruit-related traits using diploid and polyploid models in a GWAS context**. *Front Ecol Evol* 2018, **6**:107.

27. Garbowicz K, *et al*.: **Quantitative trait loci analysis identifies a prominent gene involved in the production of fatty acid-derived flavor volatiles in tomato**. *Mol Plant* 2018, **11**:1147-1165.

28. Diepenbrock CH, *et al*.: **Novel loci underlie natural variation in vitamin E levels in maize grain**. *Plant Cell* 2017, **29**:2374-2392.

29. Diepenbrock CH, *et al*.: **Eleven biosynthetic genes explain the majority of natural variation in carotenoid levels in maize grain**. *Plant Cell* 2021, **33**:882-900.

30. Lu S, *et al*.: **The cauliflower Or gene encodes a DnaJ cysteine-rich domain-containing protein that mediates high levels of β-carotene accumulation**. *Plant Cell* 2006, **18**:3594-3605.

31. Ellison SL, *et al*.: **Carotenoid presence is associated with the Or gene in domesticated carrot**. *Genetics* 2018, **210**:1497-1508.

32. Baxter I: **We aren't good at picking candidate genes, and it's slowing us down**. *Curr Opin Plant Biol* 2020, **54**:57-60.

33. Lin F, Fan J, Rhee SY: **QTG-Finder: a machine-learning based algorithm to prioritize causal genes of quantitative trait loci in Arabidopsis and rice**. *G3: Genes, Genomes, Genet* 2019, **9**:3129-3138.

34. Hartanto M, *et al*.: **Prioritizing candidate eQTL causal genes in Arabidopsis using RANDOM FORESTS**. *G3* 2022, **12**:jkac255.

35. Hershberger J, *et al*.: **Transcriptome-wide association and prediction for carotenoids and tocochromanols in fresh sweet corn kernels**. *Plant Genome* 2022, **15**:e20197.

36. Wu D, *et al*.: **Combining GWAS and TWAS to identify candidate causal genes for tocochromanol levels in maize grain**. *Genetics* 2022, **221**:iyac091.

37. Kremling KA, *et al*.: **Transcriptome-wide association supplements genome-wide association in *Zea mays***. *G3: Genes, Genomes Genetics* 2019, **9**:3023-3033.

38. Brzozowski LJ, *et al*.: **Selection for seed size has uneven effects on specialized metabolite abundance in oat (*Avena sativa* L.)**. *G3* 2022, **12**:jkab419.

39. Zhu G, *et al*.: **Rewiring of the fruit metabolome in tomato breeding**. *Cell* 2018, **172**:249-261 e12.

40. Tanaka R, *et al*.: **Leveraging prior biological knowledge improves prediction of tocochromanols in maize grain**. *Plant Genome* 2022,e20276.

41. Hu H, *et al*.: **Multi-omics prediction of oat agronomic and seed nutritional traits across environments and in distantly related populations**. *Theor Appl Genet* 2021, **134**:4043-4054.

42. Bartoshuk LM, Klee HJ: **Better fruits and vegetables through sensory analysis**. *Curr Biol* 2013, **23**:R374-R378.

43. •• Colantonio V, *et al*.: **Metabolomic selection for enhanced fruit flavor**. *Proc Natl Acad Sci* 2022, **119**:e2115865119.
Provides the comparison of different models to predict consumer preferences and the estimation of how much of flavor perception is explained by different groups of compounds.

44. • Fan Z, *et al*.: **Strawberry sweetness and consumer preference are enhanced by specific volatile compounds**. *Hortic Res* (66) 2021, **8**.
Provides an example of statistical methods to predict consumer preference.

45. Liu Y, *et al*.: **Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean**. *Front Genet* 2019, **10**:1091.

46. Ehret A, *et al*.: **Application of neural networks with back-propagation to genome-enabled prediction of complex traits in**

Holstein-Friesian and German Fleckvieh cattle. *Genet Sel Evol* 2015, **47**:1-9.

47. Resch W, Hoffman N, Swanstrom R: **Improved success of phenotype prediction of the human immunodeficiency virus type 1 from envelope variable loop 3 sequence using neural networks**. *Virology* 2001, **288**:51-62.

48. Liu G, Guo J: **Bidirectional LSTM with attention mechanism and convolutional layer for text classification**. *Neurocomputing* 2019, **337**:325-338.

49. Song S, Huang H, Ruan T: **Abstractive text summarization using LSTM-CNN based deep learning**. *Multimed Tools Appl* 2019, **78**:857-875.

50. Rhanoui M, *et al*.: **A CNN-BiLSTM model for document-level sentiment analysis**. *Mach Learn Knowl Extr* 2019, **1**:832-847.

51. Montesinos-López OA, *et al*.: **Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits**. *G3: Genes, Genomes Genetics* 2018, **8**:3829-3840.

52. von Eschenbach WJ: **Transparency and the black box problem: why we do not trust AI**. *Philos Technol* 2021, **34**:1607-1622.

53. Durán JM, Jongsma KR: **Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI**. *J Med Ethics* 2021, **47**:329-335.

54. Adadi A, Berrada M: **Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)**. *IEEE Access* 2018, **6**:52138-52160.

55. Roder J, *et al*.: **Explaining multivariate molecular diagnostic tests via Shapley values**. *BMC Med Inform Decis Mak* 2021, **21**:1-18.

56. • Lundberg SM, Lee S-I: **A unified approach to interpreting model predictions**. *Adv Neural Inf Process Syst* 2017, **30**.
The use of AI models for prediction and inference.

57. Winter E: **The shapley value**. Handbook of Game Theory with Economic Applications. 2002,:2025-2054.

58. Mieth B, *et al*.: **DeepCOMBI: explainable artificial intelligence for the analysis and discovery in genome-wide association studies**. *NAR Genom Bioinform* (3) 2021, **3**.

59. Johnsen PV, *et al*.: **A new method for exploring gene–gene and gene–environment interactions in GWAS with tree ensemble methods and SHAP values**. *BMC Bioinform* 2021, **22**:230.

60. Rincent R, *et al*.: **Phenomic selection is a low-cost and high-throughput method based on indirect predictions: proof of concept on wheat and poplar**. *G3 Genes|Genomes|Genet* 2018, **8**:3961-3972.

61. Krause MR, *et al*.: **Hyperspectral reflectance-derived relationship matrices for genomic prediction of grain yield in wheat**. *G3 Genes|Genomes|Genet* 2019, **9**:1231-1247.

62. Wieme J, *et al*.: **Application of hyperspectral imaging systems and artificial intelligence for quality assessment of fruit, vegetables and mushrooms: a review**. *Biosyst Eng* 2022, **222**:156-176.

63. Antoniou A, Storkey A, Edwards H: **Augmenting Image Classifiers Using Data Augmentation Generative Adversarial Networks**. Springer International Publishing; 2018.

64. Chen Y, *et al*.: **Generative adversarial networks in medical image augmentation: a review**. *Comput Biol Med* 2022, **144**:105382.

65. Whitaker VM, *et al*.: **Estimation of genetic parameters for 12 fruit and vegetative traits in the University of Florida Strawberry Breeding Population**. *J Am Soc Hortic Sci J Am Soc Hort Sci* 2012, **137**:316-324.

66. Jones JW, *et al*.: **The DSSAT cropping system model**. *Eur J Agron* 2003, **18**:235-265.

67. • C.D. Messina, et al., On the dynamic determinants of reproductive failure under drought in maize, *In silico Plants*, **1**, diz003,2019..
Detailed framework covering the use of crop growth models in maize.

68. Turc O, *et al*.: **The growth of vegetative and reproductive structures (leaves and silks) respond similarly to hydraulic cues in maize**. *New Phytol* 2016, **212**:377-388.

69. Allen DK, Libourel IGL, Shachar-hill Y: **Metabolic flux analysis in plants: coping with complexity**. *Plant Cell Environ* 2009, **32**:1241-1257.

70. Libourel IG, Shachar-Hill Y: **Metabolic flux analysis in plants: from intelligent design to rational engineering**. *Annu Rev Plant Biol* 2008, **59**:625-650.

71. Amthor JS: **The McCree–de Wit–Penning de Vries–Thornley respiration paradigms: 30 years later**. *Ann Bot* 2000, **86**:1-20.

72. Joshi J, *et al*.: **Why cutting respiratory CO_2 loss from crops is possible, practicable, and prudential**. *Mod Agric* 2023, **1**:16-26.

73. Wu A, *et al*.: **Quantifying impacts of enhancing photosynthesis on crop yield**. *Nat Plants* 2019, **5**:380-388.

74. Cooper M, Messina CD: **Breeding crops for drought-affected environments and improved climate resilience**. *The Plant Cell* 2023, https://doi.org/10.1093/plcell/koac321 koac321.

75. Diepenbrock CH, Tang T, Jines M, Technow F, Lira S, Podlich D, Cooper M, Messina C: **Can we harness digital technologies and physiology to hasten genetic gain in United States maize breeding?** *Plant Physiology* 2022, **188**:1141-1157.

76. Messina CD, Tang T, Truong SK, McCormick RF, Technow F, Powell O, Mayor L, Lira S, Gutterson N, Van Eeuwijk F, Jones JW, Hammer GL, Cooper M: **Crop Improvement for circular bio economy systems**. *Journal of the ASABE* (3) 2022, **65**:491-504.

77. • Hall RD, *et al*.: **High-throughput plant phenotyping: a role for metabolomics?** *Trends Plant Sci* 2022, **27**:549-563.
Comprehensive review discussing the potential of multi-omics data on plant science.

78. Wurtzel ET: **Changing form and function through carotenoids and synthetic biology**. *Plant Physiol* 2018, **179**:830-843.

79. Sahu A, *et al*.: **Advances in flux balance analysis by integrating machine learning and mechanism-based models**. *Comput Struct Biotechnol J* 2021, **19**:4626-4640.

80. Lang M: **yggdrasil: a Python package for integrating computational models across languages and scales**. *In silico Plants* 2019, **1** diz001.