Harnessing Food Composition Data: Machine Learning models to predict Taste and Health outcomes of Food Processing

By

TARINI NARAVANE

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological Systems Engineering

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Tina Jeoh - Chair

Bruce German

Mason Earles

Committee in Charge

2024

Acknowledgement

I would like to take this opportunity to thank my mentors, co-authors and colleagues for their role in the successful completion of my Phd.

My chair Dr. Tina Jeoh provided invaluable support with funding and the necessary guidance structure for my multi-disciplinary research interest. Dr. Bruce German, a member of my thesis committee, supported and encouraged my research aspirations for my Phd. and beyond for what will be my professional trajectory.

I am grateful to my co-authors; Gabriel Simmons for his contribution to the research in Chapters 2,3 and to Dr. Ilias Tagkopoulos for his contribution in Chapter 2.

Finally, this would not be possible without the funding support from USDA, my department and Institute for Innovation in Food and Health.

List of Publications:

1. Using word embeddings to learn a better food ontology, Frontiers in Artificial Intelligence, Volume 3, November 2020. Co-authors: Jason Youn, Department of Computer Science, UC Davis, and Ilias Tagkopoulos, Department of Computer Science, UC Davis. DOI 10.3389/frai.2020.584784
2. Machine learning models to predict the micronutrient profiles in foods after cooking, Current Research in Food Science, 2023, Volume 6. Co-authors- Ilias Tagkopoulos, Department of Computer Science, UC Davis. DOI 10.1016/j.crfs.2023.100500
3. Ontological how and why: action and objective of planned processes in the food domain, Frontiers in Artificial Intelligence, Volume 6, July 2023. Co-authors: Damion Dooley, Center for Infectious Disease Genomics and One health, Simon Fraser University, Canada. DOI 10.3389/frai.2023.1137961
4. Structure-property machine learning models with predictive capabilities for glycans in food, Preprint bioarxiv November 2023. Co-authors Gabriel Simmons, Department of Computer Science, UC Davis, Bruce German, Food Science and Technology, UC Davis. DOI 10.1101/2023.11.12.566488

Abstract

Food processing is a complex chemical process that transforms the chemical composition of the raw ingredients into their final food product, whose complexity is not yet deciphered. In our modern food system, understanding the impact of food processing on both taste and health outcomes is crucial. Traditional computational models offer some insight and utility in food production; these are essentially targeted approaches specific to foods, processes and nutritional and/or sensory outcomes. Their limitation is the inability to scale, which is necessary to address the current urgent demands of precision and personalized health and sustainable food production. These multi-variate challenges requires a deeper and more comprehensive understanding of the complexity of foods and processing methods and comprise of two main research efforts; to build food composition datasets that embody this information, and then to identify and apply the relationships as solutions. Machine Learning (ML) is widely hailed by research and industry as the technology best suited to address such an enormous multivariate problem. This shared vision has already led to efforts in building the necessary datasets. As relevant to this challenge, a common hypothesis is tested across two projects in this research - **There a relationship between the chemical composition of a food and its nutritive and sensory properties in the processed state.**

The first project develops ML models to predict the content of seven vitamins (vitamin A, B1, B2, B3, B6, B9, C) and seven minerals (Calcium, Iron, Magnesium, Phosphorus, Potassium, Sodium, Zinc) in a processed food. The ML models are trained to learn the multi-parametric transformation patterns between the compositions of the raw

and cooked foods. The focus was to be able address common dietary questions of consumers about choice of food and cooking method, and the selected training data included 425 plant and animal-based foods and 5 common cooking methods (steaming, boiling, roasting, grilling and broiling). The predictive model performed 43% and 18% better than using the standard USDA retention factor model for wet heat (steaming, boiling) and dry heat (roasting, grilling, broiling) processes, respectively. The breakdown of the predictive performance by food category revealed that legumes have the best among plant-based foods and beef the best in the animal-based foods. This suggests that nutrient loss is affected by the structural composition of foods, for future research.

The second project explored structure-property models that aim to decipher the complex relation between the physical shape of a molecule and its physical properties and/or the functional role of the molecule in a product formulation. The focus was the modeling of glycans (i.e., carbohydrates), which are not only abundant in food, but essential to both food production and, more importantly, human health. In the study, regression methods were used to generalize the relationships between the structure of starch (e.g., chain length and composition of protein and amylose) and a range of its properties (e.g., gelatinization temperature, time series viscosity data, gel consistency, and sensory texture) for 301 samples of rice. The results indicated that the structure-composition data is a significantly better predictor (27% more predictive accuracy) of sensory mouthfeel than the physical properties, even though the latter is typically used in experimental research.

The results of these projects demonstrate the ability of ML methods to learn a variety of complex multivariate relationships. However further progress is gated by the

availability of high quality and high-resolution datasets and although the analytical methods exist, the challenge is knowing the relevant dataset for a specific prediction target. This challenge is addressed by both projects, where an assessment of what could improve prediction accuracy is the basis for future areas for data collection.

List of Appendix materials

1. Supplementary_project1.xlsx. Supplementary files for Thesis Project 1, referenced in Chapter 2
2. Supplementary_project2.xlsx. Supplementary files for Thesis Project 2, referenced in Chapter 3
3. Manuscripts, referenced in Section 1.4
    a. Word_embeddings_food_ontology.pdf
    b. Food_processing_ontology.pdf

List of Figures

List of Tables

Chapter 1: The Use of Machine Learning in Food Processing

1.1 Introduction

1.1.1 Deciphering the chemical complexity of food processing

Food processing is a complex chemical process that transforms the chemical composition of the raw ingredients into their final food product. However, the relationships describing this transformation remain largely unknown due to the unresolved chemical and structural complexity of the food ingredients, as well as the physio-chemical transformation mechanisms that occur during processing as inferred by several review articles and food chemistry texts[12–4]. This vast challenge is currently addressed by point solutions that address a specific nutritional or sensory outcome for specifics foods and/or processing methods.

Prevalent methods specific to nutritive outcomes of food processing have developed targeted approaches that are specific to a single nutrient. For example, kinetic modeling based on experimental data for any given food establishes the relationship between nutrient concentration, time, and temperature conditions[5–7]. This can then be applied to compute concentrations, for example predicting vitamin C (ascorbic acid) content in processed orange juice[7]. Another approach to compute post-process nutrition composition, is to apply retention factors (RF) which are based on analytical composition data on a representative set of foods and processes. RF-based computation is used widely by food manufacturers for nutrition labels, and by USDA's dietary survey group to calculate nutrient intakes that investigators may use to determine correlations between intake and health outcomes. However, all of these methods have limited potential. Kinetic

models are difficult to scale up to capturing more food and processing parameters, as these measurements are time-consuming, expensive[8] and have many experimental challenges such as rapid degradation of certain chemicals. RF-based methods in practice inevitably under or overestimate the nutrient content in a particular instance since any single RF is representative of several foods and a cooking method. The prediction of sensory properties is relatively recent compared to the above methods. Philips et al.[9], measured the detailed carbohydrate profile in bananas at various stages of ripening, and correlated the composition at different stages of ripening to various organoleptic and nutritive properties of taste, texture, and dietary fiber. This early work was impactful in revealing the compositional basis for the ripening process but was not aimed at deciphering the relation to these properties.

This thesis addresses the challenge in food processing of connecting the inputs (ingredients, processes) to the outputs (properties of the finished food) in a broader context compared to prevalent methods. In addition, the thesis addresses another significant limitation of the current solutions, to advance the domain knowledge on the above-mentioned complexity of food processing.

1.1.2 Thesis: Hypothesis, Scope and Projects

The thesis tested a common hypothesis across two projects - **There a relationship between the chemical composition of particular food and its nutritive and sensory properties in the processed state.** At the same there is a distinction between the projects based on the composition datasets, one as analyzed by the 20th century classical methods of chemistry in solution and the other by the 21st century omics

technologies[10]. While the hypothesis is the common, it is important to note that the omics datasets are capable of modelling not just the composition but also the structure (morphology) of food matrices. The relevance of the structural data to predictive capability and performance was tested in project 2 (details in **Section 3.2**).

The first project develops ML models to predict the content of several nutrients in the processed food for a diversity of foods and cooking methods. The hypothesis was that ML models can learn the multi-parametric transformation patterns between the compositions of raw and cooked foods from the food composition data for a variety of foods and cooking methods. This was tested on 425 plant and animal-based foods and for five different cooking methods, where the prediction targets were the content of seven vitamins (vitamin A, B1,B2, B3, B6, B9, C) and seven minerals (Calcium, Iron, Magnesium, Phosphorus, Potassium, Sodium, Zinc) in the cooked food, predicted from the chemical composition of the raw food.

The second project predicted the physical and sensory properties related to texture. As texture is a physical property, the modelling approach was to predict from the structural composition, therefore creating a structure-property model. The focus was on the glycan composition of foods as motivated by the abundance of glycans - commonly known as “carbohydrates” - in nature and in the context of the human diet. They are responsible for various biological activities significant to human health[11–13] and, as such, have been a target in food engineering for human nutrition and product formulation. This prominence has motivated research on the relationship between the structural composition of glycans and their physical properties in food formulation. We refer to the structure as defined in the glycobiology text[14] (further details in Chapter 3), “the primary

structure of a glycan is defined by the type and order of monosaccharide residues, by the configuration and position of glycosidic linkages, and by the nature and location of the non-glycan entity to which it is attached." This definition aligns with recent advances in analytical methods that allow for the detection, identification and characterization of glycans in food[15–17] .For example, the afore mentioned research by Philips et al.[9] discovered that glycan composition data is correlated to the ripening process as well as taste, texture and nutritional properties. While their work was not aimed at deciphering the relation between glycans and bulk food properties, it nevertheless supports our hypothesis -that leveraging high-resolution glycan composition data will offer performance gains over current practices for a variety of food engineering predictive tasks. Project 2 tested the hypothesis in the setting of the cooking of rice. The prediction targets were five physical properties and two mouthfeel sensory properties of starch in rice, from the structural data of the starch component of the rice sample.

As these models are discovering relationships that are yet unknown in the domain knowledge, it was very important to evaluate the data that the models are trained on. Section 1.1.3 addresses issues of quality (noise, bias) and completeness of the dataset. Section 1.2 summarizes the prior research as relevant to projects 1 and 2. Section 1.3 frames research questions for both projects regarding predictive accuracy of the model and its ability for knowledge discovery.

1.1.3 Quality assessment of datasets in machine learning

Several issues of data quality have been identified in the context of data generation efforts to create large food composition datasets. Ahmed et al.[18] pointed out the issues in

quantifying the content of a compound and gave an example of an apple measured in three different setups only agreed on the content of 14 out of more than 500 compounds. They intended to resolve such inconsistencies through standardized analytical methods in the Periodic Table of Foods Initiative (PTFI) dataset. Ene-Obong et al.[19] reported that certain African national datasets do not have balanced coverage of traditional and novel foods, or manufactured and cooked foods. Additionally, Fukagawa et. al [20]reported that the current USDA's food composition database lacks provenance and processing meta-data for the samples, and proposes to resolve it by creating a new dataset of single ingredients that are sourced across various geographics and farming methods. Finally, Westenbrink et al.[21] report on having identified and resolved issues of non-standardized documentation when harmonizing the various European national datasets.

These examples report inadequacies of datasets through data analysis. In this thesis, the need for additional data is based on assessing what additional data would improve predictive performance. The dataset in project 1, was found to have several gaps in the data provenance, structure, data sampling that would affect this current use as a reference. The food samples coverage was under representative with only 1 sample of each of the plant-based foods (fruits, vegetables, legumes) while there is significant variation in composition from varieties and growing conditions. The analysis of the composition data revealed an anomaly where the content of a vitamin or nutrient was greater in the cooked food than in the raw food. This was caused by the representation of the composition per 100 g of both the raw and cooked food, with undocumented yield factors. These yield factors would typically be greater or less than 1, caused mainly by either gain or loss of water and fat in the cooking process. This issue was mitigated by

data scaling methods as described in Chapter 2. Further issues of data provenance and meta-data are also assessed. Overall, project 1 assessed such significant limitations and proposed solutions to guide in building future datasets with the aim of more accurate and reliable prediction models.

Project 2 addressed issues related to data completeness and its effect on prediction accuracy. As explained in Section 1.3.2 of this chapter, this dataset was created as needed by IRRI with the specific objective of classifying varieties of rice, but not to predict the values of specific physical and textural properties of rice. So, our research questions (RQ1 and RQ2 in Section 1.3.2 of this chapter) explored the completeness of the predictive feature set in regard to the prediction targets. Methods were developed to test for these questions, and subsequently report the results in Chapter 3 which additional features might improve prediction accuracy. Such inquiry helps to develop a methodical strategy of modeling and data generation, which could accelerate the success of such structure-function models while simultaneously developing insight of the domain.

1.1.4 The need for FAIR (Findable, Accessible, Interoperable, and Reusable) datasets

"Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community."[22]

The creation of FAIR datasets has two main aspects; creation of a standardized and hierarchical language called ontologies, and subsequently labelling the datasets with these ontologies. The research work for this thesis includes two published projects

addressing this need for FAIR datasets. The manuscripts are listed on page III and are included in the Appendix. The first is titled “Using Word Embeddings to Learn a Better Food Ontology”. This project automates the creation of an ontology starting from a manually curated skeleton ontology and an extensive list of ontology classes and instances. The aim of this project was to address the issue that the manual creation of ontologies is time consuming and prone to errors. The second manuscript is titled “The Ontological How and Why – Action and Objective of Planned Processes in Food”. This project addressed the manual creation of a skeleton ontology of food processing that fits into the OBO framework of the larger life-sciences family of ontologies. Both these projects laid the groundwork for AI methods to extend expert-curated ontologies, and to collect and/or map data to these ontologies.

1.2 Potential of machine learning methods to predict properties from composition and structural data.

Machine learning models have been successful in generalizing across a wide range of prediction tasks when trained on relevant datasets. In this section, a brief summary is given on the prior research pertaining to projects 1 and 2.

1.2.1 Project 1: Machine learning models to predict micronutrient profile in food after processing.

Several models trained on food composition datasets have addressed attributes related to nutrient profiles. The P_NUT model uses natural language processing (NLP) methods to predict the macronutrient (proteins, fats, and carbohydrates) content of foods

from a text description of the food[23] A more recent version of this model can predict macronutrients from a recipe[24] . USDA investigators predicted the content of three label nutrients (carbohydrates, protein and sodium) in processed foods from the ingredient list[25] using the Branded Foods datatype in Food Data Central(FDC)[26]. Several projects predicted nutrient contents from the composition data. For example, nutrient content was predicted for the missing values in food composition data[27] ,lactose content was predicted in dietary recall database[28] and fiber content was predicted for commercially processed foods[29]. In the context of food processing, a food was assigned a label of the degree of processing based on the composition data[30] .The four labels used were as per the NOVA[31] system ranging from minimally-processed to ultra-processed. This body of prior research implies that there is a complex interdependence between the chemical components of the food and supports the hypothesis of our work, that the transformation patterns in food composition can be learnt for a diversity of foods and a variety of processes.

1.2.2 Project 2: Structure-property machine learning models with predictive capabilities for glycans in food

ML models have also shown promise in modeling complex relationships between polymer structures like proteins and glycans, and their behaviors. The most prominent example is AlphaFold[32] , which was made possible after two decades of collective effort in building a dataset linking protein sequence to 3D conformation. The breakthrough success of AlphaFold was a stepping stone towards better models of protein function/property, which have since accelerated research in areas such as protein-protein

interactions[33] and drug discovery[34].Unlike for proteins, computational techniques for glycans are still nascent, specifically in terms of large datasets and wide applicability. However, early work in ML and other computational methods including molecular dynamic simulation (MDS) are being recognized[35,36]. MDS approaches have proven capable of predicting polymer conformation and properties[37–39], although their utility is limited to short-chain polymers of sizes ranging from 2-25 units due to the computational intensity of the method. Other ML methods have been used to predict properties of immunogenicity and pathogenicity[40] for single glycans from structural information. Importantly, prior work has not explored the utility of ML in relating molecular-level glycan information to bulk food properties. At present, glycan datasets are considerably smaller than the volume of data that enabled the AlphaFold breakthrough for proteins[35,41]. While "AlphaFold for glycans" is still far away, early work in glycan predictive modeling encourages us to explore the utility of ML for predicting the properties of glycan-based foods.

1.3 Project Specifications: Dataset and Research questions

The two main steps to defining the projects were identifying the relevant datasets and framing research questions for inquiry. Since both projects are based on a common hypothesis that there is a generalized pattern in the data, several common research questions were established; first, to address the accuracy of prediction of the machine learning models by comparing against baseline methods. Secondly, to address whether the given input variables are adequately able to learn the complex relationships to the

outputs. Finally, whether the predictive features reveal any domain insight (some features being more important than others).

1.3.1 Project 1: Dataset and Research questions

The composition dataset of 7793 foods from the Standard Reference (SR) legacy dataset (USDA National Nutrient Data- base for Standard Reference, 2022) was used, being the most suitable of the five data sets in FDC (as of November 2021) since it is aligned with the project objectives. The SR dataset has composition data for both raw and cooked food samples for single ingredients and is intended for application in public health initiatives such as the assessment of nutrient intakes for the purpose of monitoring national nutrition, creating meal plans in schools and daycare centers, and in product development and labeling by food manufacturers. For our models, a subset of the SR dataset was selected according to the following criteria: raw-cooked food pairs were matched where the raw foods were a single ingredient harvested from a plant or animal (includes butchery products), and the cooked foods were the outcome of the raw food treated to wet (boiling or steaming), or dry (roasting, grilling, or broiling) heat processes. This resulted in a total of 840 foods in the dataset, with 178 and 247 pairs from wet and dry heat processes, respectively. In this dataset, all plant-based foods were cooked by wet heat processes (WH), and all animal-based foods by dry heat (DH) processes. The prediction targets were the content of seven vitamins and seven minerals in the cooked food, predicted from the chemical composition of the raw food. The model design and methods were developed to address the following research questions.

Research Questions

RQ1. Is the predictive ML model more accurate than the baseline of the prevalent methods ?

RQ2. Is there a difference in predictive performance by categories of food and processing methods ?

RQ3. What features are the most predictive ?

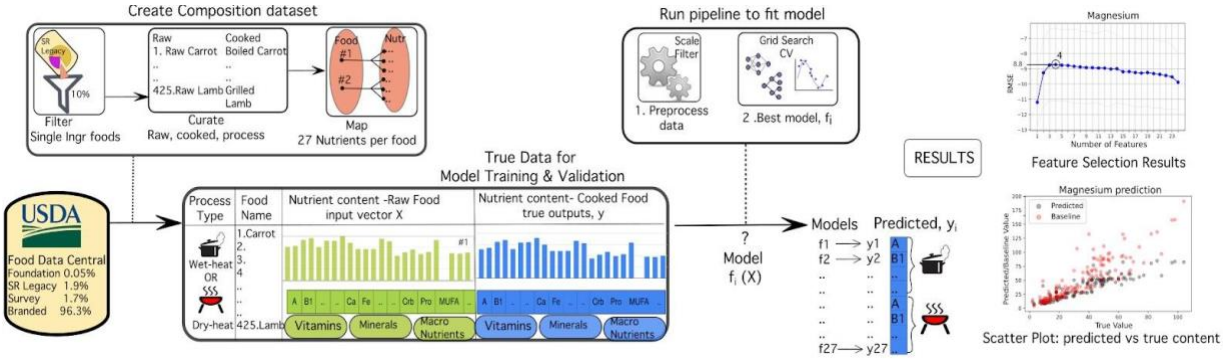


Figure 1.1 Overview of architecture (left to right) from data selection to prediction results. Single ingredient foods are selected from SR legacy (one of the five data types in FDC), and then organized by pair (raw,cooked) and cooking process type. Cooking processes include boiling and steaming which are grouped into wet heat processes(WH) and broiling, grilling, and roasting which are grouped into dry heat processes (DH) . Foods are mapped to composition, with 27 components per food. Models are trained from composition data, such that the input feature is the composition of the raw food, and each model is trained separately for every micronutrient in the cooked food. Models are trained separately for both process types, with 14 for WH and 13 for DH (excluding vitamin C predictor model). Prior to model fitting, the composition data is scaled and filtered. Model fitting uses a grid search cross validation approach, such that there are 12336 regressor models. The best model has the least error, RMSE. Then predicted composition is compared to the actual (ground truth) composition in two results. The feature selection

result is the performance (RMSE) analysis against the feature (input features) size. The scatter plot for prediction of magnesium content shows that both the prediction (black dots) and baseline (red dots) values on the Y axis, versus the actual values (X axis).

1.3.2 Project 2: Dataset and Research questions

As stated in Sections 1.1 and 1.2, the scope of glycans is vast and the analytical progress in discovery is very recent. In contrast, starch is a specific glycan with a long history of analysis and research and continues to be studied due to its diverse physical properties. We therefore selected starch as the representative glycan for our case study. For this project, the dataset from the study by Beunafe et al. at the International Rice Research Institute (IRRI)[42] was used, which includes three types of data: composition, physical, and sensory, as shown in Figure 1.2. The composition data includes amylose content (AC) and protein content (PC). The structural data was measured by size exclusion chromatography (SEC) and reported for 500 degrees of polymerization (DP) values in the range from 5 to 12,000. The physical properties consist of gelatinization temperature (GT) measured by DSC, gel consistency (GC) measured as the length of the starch gel prepared in a tube after heating (followed by an hour of cooling), and time series viscosity data using a rapid visco analyser (measured every four seconds for a 12-minute period). The sensory data included 13 mouthfeel descriptors and was collected for 100 samples of rice. For this case study, hardness and stickiness were the focus, since these are the most widely studied sensory characteristics across literature and fundamental to other properties like cohesiveness and toothsomeness[43,44]. These

research questions address the main thesis that the key to predicting the properties is the detailed structural data.

Research Questions

RQ1. Is size exclusion chromatography data and content data sufficient to predict both physical and sensory properties of cooked rice?

RQ2. Does gelatinization temperature and gel consistency information improve predictive performance over only the chain length and content data (tested in RQ1)?

RQ3. What features (structural or otherwise) are the most informative for each prediction?

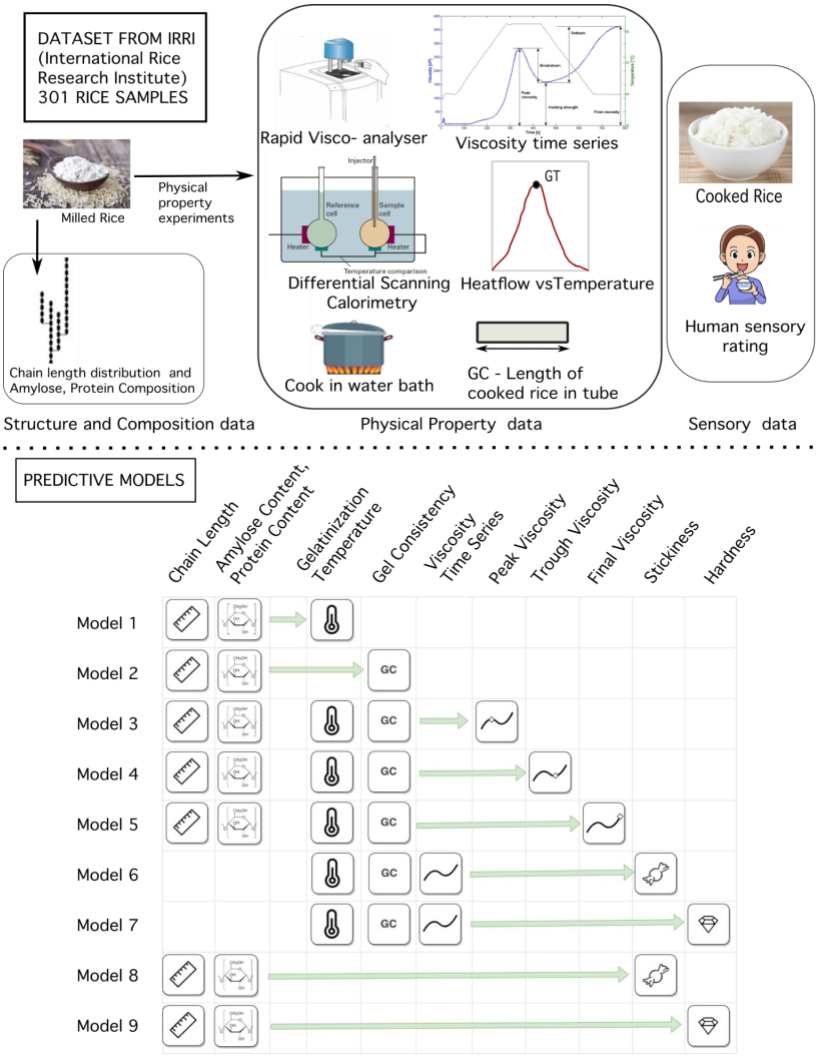


Figure 1.2: Data and Models. The data is from prior research[42], for 301 samples of rice based on the indicated methods in the upper portion of the figure. 9 Models are trained based on this data as seen in lower half of the figure. Detailed explanation for the data and the models is in Chapter 3.

1.4 Conclusion

Detailed descriptions of the data, methods, results, and discussion for Project 1 are in Chapter 2 and for Project 2 are in Chapter 3.

REFERENCES

1. Mehta, B. M. & Cheung, P. C.-K. Overview of Food Chemistry. in *Handbook of Food Chemistry* (ed. Cheung, P. C. K.) 1–9 (Springer, Berlin, Heidelberg, 2015). doi:10.1007/978-3-642-41609-5_34-1.
2. Capuano, E., Oliviero, T. & van Boekel, M. A. J. S. Modeling food matrix effects on chemical reactivity: Challenges and perspectives. *Crit. Rev. Food Sci. Nutr.* **58**, 2814–2828 (2018).
3. Anandharamakrishnan, C. Food Structure and the Complexity of Food Matrices. (2023) doi:10.1039/BK9781839162428-00290.
4. van Boekel, M. A. J. S. Kinetic Modeling of Food Quality: A Critical Review. *Compr. Rev. Food Sci. Food Saf.* **7**, 144–158 (2008).
5. Boekel, M. A. J. S. van. *Kinetic Modeling of Reactions In Foods*. (CRC Press, Boca Raton, 2008). doi:10.1201/9781420017410.
6. Bajaj, S. R. & Singhal, R. S. Degradation kinetics of vitamin B12 in model systems of different pH and extrapolation to carrot and lime juices. *J. Food Eng.* **272**, 109800 (2020).
7. Peleg, M., Normand, M. D., Dixon, W. R. & Goulette, T. R. Modeling the degradation kinetics of ascorbic acid. *Crit. Rev. Food Sci. Nutr.* **58**, 1478–1494 (2018).
8. Ling, B., Tang, J., Kong, F., Mitcham, E. J. & Wang, S. Kinetics of food quality changes during thermal processing: a review. *Food Bioprocess Technol.* **8**, 343–358 (2015).
9. Phillips, K. M. *et al.* Dietary fiber, starch, and sugars in bananas at different stages of ripeness in the retail market. *Plos One* **16**, e0253366 (2021).
10. Gallo, M. & Ferranti, P. The evolution of analytical chemistry methods in foodomics. *J. Chromatogr. A* **1428**, 3–15 (2016).
11. Zhang, G. & Hamaker, B. R. The nutritional property of endosperm starch and its contribution to the health benefits of whole grain foods. *Crit. Rev. Food Sci. Nutr.* **57**, 3807–3817 (2017).
12. Carmody, R. N. *et al.* Cooking shapes the structure and function of the gut microbiome. *Nat. Microbiol.* **4**, 2052–2063 (2019).
13. Cerqueira, F. M., Photenhauer, A. L., Pollet, R. M., Brown, H. A. & Koropatkin, N. M. Starch digestion by gut bacteria: crowdsourcing for carbs. *Trends Microbiol.* **28**, 95–108 (2020).
14. Haslam, S. M. *et al.* Structural analysis of glycans. in *Essentials of Glycobiology* (eds. Varki, A. et al.) (Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2022). doi:10.1101/glycobiology.4e.50.
15. Amicucci, M. J., Nandita, E. & Lebrilla, C. B. Function without Structures: The Need for In-Depth Analysis of Dietary Carbohydrates. *J. Agric. Food Chem.* **67**, 4418–4424 (2019).

16. Amicucci, M. J. *et al.* A rapid-throughput adaptable method for determining the monosaccharide composition of polysaccharides. *Int. J. Mass Spectrom.* **438**, 22–28 (2019).
17. Xu, G., Amicucci, M. J., Cheng, Z., Galermo, A. G. & Lebrilla, C. B. Revisiting monosaccharide analysis - quantitation of a comprehensive set of monosaccharides using dynamic multiple reaction monitoring. *The Analyst* **143**, 200–207 (2017).
18. Ahmed, S. *et al.* Foodomics: A Data-Driven Approach to Revolutionize Nutrition and Sustainable Diets. *Front. Nutr.* **9**, 874312 (2022).
19. Ene-Obong, H. *et al.* Importance and use of reliable food composition data generation by nutrition/dietetic professionals towards solving Africa's nutrition problem: constraints and the role of FAO/INFOODS/AFROFOODS and other stakeholders in future initiatives. *Proc. Nutr. Soc.* **78**, 496–505 (2019).
20. Fukagawa, N. K. *et al.* USDA's FoodData Central: what is it and why is it needed today? *Am. J. Clin. Nutr.* **115**, 619–624 (2022).
21. Westenbrink, S., Presser, K., Roe, M., Ireland, J. & Finglas, P. Documentation of aggregated/compiled values in food composition databases; EuroFIR default to improve harmonization. *J. Food Compos. Anal.* **101**, 103968 (2021).
22. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
23. Ispirova, G., Eftimov, T. & Koroušić Seljak, B. P-NUT: Predicting NUTrient Content from Short Text Descriptions. *Mathematics* **8**, 1811 (2020).
24. Ispirova, G., Eftimov, T. & Koroušić Seljak, B. Domain Heuristic Fusion of Multi-Word Embeddings for Nutrient Value Prediction. *Mathematics* **9**, 1941 (2021).
25. Ma, P. *et al.* Application of machine learning for estimating label nutrients using USDA Global Branded Food Products Database, (BFPD). *J. Food Compos. Anal.* **100**, 103857 (2021).
26. USDA National Nutrient Database for Standard Reference, Legacy Release \textbarAg Data Commons.
27. Gjorshoska, I., Eftimov, T. & Trajanov, D. Missing value imputation in Food Composition Data with Denoising Autoencoders. *J. Food Compos. Anal.* 104638 (2022) doi:10.1016/j.jfca.2022.104638.
28. Chin, E. L. *et al.* Nutrient Estimation from 24-Hour Food Recalls Using Machine Learning and Database Mapping: A Case Study with Lactose. *Nutrients* **11**, (2019).
29. Davies, T. *et al.* An innovative machine learning approach to predict the dietary fiber content of packaged foods. *Nutrients* **13**, (2021).
30. Menichetti, G., Ravandi, B., Mozaffarian, D. & Barabási, A.-L. Machine learning prediction of food processing. *medRxiv* (2021) doi:10.1101/2021.05.22.21257615.
31. Moubarac, J.-C., Parra, D. C., Cannon, G. & Monteiro, C. A. Food classification systems based on food processing: significance and implications for policies and

actions: A systematic literature review and assessment. *Curr. Obes. Rep.* **3**, 256–272 (2014).
32. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
33. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. *Nat. Commun.* **13**, 1265 (2022).
34. Ivanenkov, Y. A. *et al.* Chemistry42: An AI-Driven Platform for Molecular Design and Optimization. *J. Chem. Inf. Model.* **63**, 695–701 (2023).
35. Li, X., Xu, Z., Hong, X., Zhang, Y. & Zou, X. Databases and bioinformatic tools for glycobiology and glycoproteomics. *Int. J. Mol. Sci.* **21**, (2020).
36. Raman, R., Raguram, S., Venkataraman, G., Paulson, J. C. & Sasisekharan, R. Glycomics: an integrated systems approach to structure-function relationships of glycans. *Nat. Methods* **2**, 817–824 (2005).
37. Anggara, K. *et al.* Identifying the origin of local flexibility in a carbohydrate polymer. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
38. Perez, S. & Makshakova, O. Multifaceted computational modeling in glycoscience. *Chem. Rev.* **122**, 15914–15970 (2022).
39. Zhiguang, C., Junrong, H., Huayin, P. & Keipper, W. The effects of temperature on starch molecular conformation and hydrogen bonding. *Starch - Stärke* **74**, 2100288 (2022).
40. Burkholz, R., Quackenbush, J. & Bojar, D. Using graph convolutional neural networks to learn a representation for glycans. *Cell Rep.* **35**, 109251 (2021).
41. Ranzinger, R., Herget, S., von der Lieth, C.-W. & Frank, M. GlycomeDB–a unified database for carbohydrate structures. *Nucleic Acids Res.* **39**, D373-6 (2011).
42. Buenafe, R. J. Q., Kumanduri, V. & Sreenivasulu, N. Dataset on viscosity and starch polymer properties to predict texture through modeling. *Data Brief* **36**, 107038 (2021).
43. Li, H., Fitzgerald, M. A., Prakash, S., Nicholson, T. M. & Gilbert, R. G. The molecular structural features controlling stickiness in cooked rice, a major palatability determinant. *Sci. Rep.* **7**, 43713 (2017).
44. Li, C., Luo, J.-X., Zhang, C.-Q. & Yu, W.-W. Causal relations among starch chain-length distributions, short-term retrogradation and cooked rice texture. *Food Hydrocoll.* **108**, 106064 (2020).

Chapter 2

Machine learning models to predict micronutrient profile in food after processing[1]

2.1 Introduction

The machine learning predictive models developed in this project predict the micronutrient contents (specifically seven vitamins and seven minerals) of the cooked food from the raw food composition data. Chapter 1 provided the relevant body of prior research which implies that there is a complex interdependence between the chemical components of the food and supports the hypothesis of our work, that the transformation patterns in food composition due to a variety of processes can be learnt. Here, we have constructed ML models that predict food micronutrient (specifically seven vitamins and seven minerals) composition after processing (**Figure** 1). We have curated a sample of 820 single-ingredient foods in the raw and cooked states, for five basic cooking processes namely steaming, boiling, grilling, broiling, and roasting from FDC. (Our aim was to model basic single-step cooking processes, and we did not consider multi-step processes as in recipes or industrial processes.)

[1] This chapter has been published. Co-authors- Ilias Tagkopoulos, Department of Computer Science, UC Davis. Current Research in Food Science, 2023, Volume 6. DOI 10.1016/j.crfs.2023.100500

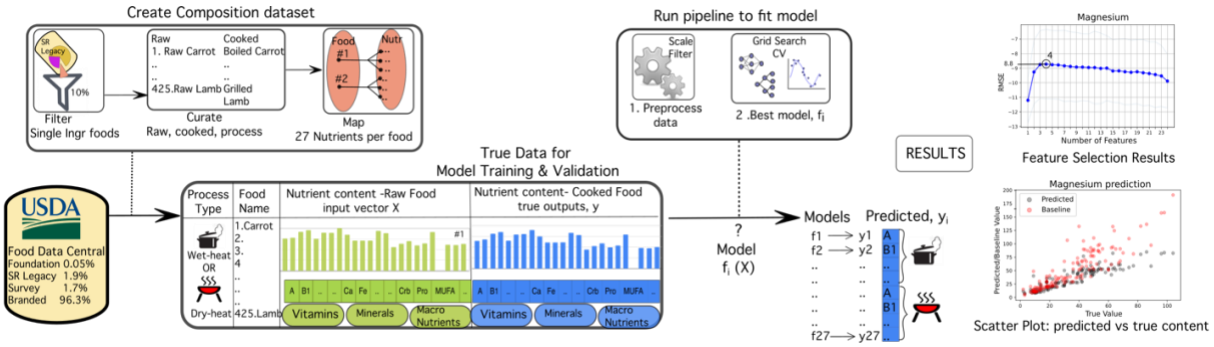


Figure 2.1. **Overview of architecture (left to right) from data selection to prediction results.** Single ingredient foods are selected from SR legacy (one of the five data types in FDC), and then organized by pair (raw,cooked) and cooking process type. Cooking processes include boiling and steaming which are grouped into wet heat processes(WH) and broiling, grilling, and roasting which are grouped into dry heat processes (DH) . Foods are mapped to composition, with 27 components per food. Models are trained from composition data, such that the input feature is the composition of the raw food, and each model is trained separately for every micronutrient in the cooked food. Models are trained separately for both process types, with 14 for WH and 13 for DH (excluding vitamin C predictor model). Prior to model fitting, the composition data is scaled and filtered. Model fitting uses a grid search cross validation approach, such that there are 12336 regressor models. The best model has the least error, RMSE. Then predicted composition is compared to the actual (ground truth) composition in two results. The feature selection result is the performance (RMSE) analysis against the feature (input features) size.

The scatter plot for prediction of magnesium content shows the both the prediction (black dots) and baseline (red dots) values on the Y axis, versus the actual values (X axis).

2.2 Data and Methods

2.2.1 Dataset

We downloaded the composition dataset of 7,793 foods from the Standard Reference (SR) legacy dataset[1], which is the most suitable of the five data sets in FDC (**Figure 2.1**; as of November 2021), since it is aligned with our objectives. The SR dataset has composition data for both raw and cooked food samples for single ingredients and is intended for application in public health initiatives such as the assessment of nutrient intakes for the purpose of national nutrition monitoring, in creating meal plans in schools and day-care centers, in product development and labeling by manufacturers. The composition data for the foods in SR is obtained from three sources; analytical experiments, analytical data from literature, and calculations based on the analytical data for example composition data on butterhead lettuce is calculated from composition of leafy green lettuce which is a "similar food"[2,3]. The complete list of composition source types is in **Supplementary materials**. The four other data sets with composition data in FDC are; Foundation foods with single-ingredients foods and mostly only raw foods and the aim is to provide high quality data on raw ingredients with relevant meta-data as a precedent for future data sets, Experimental foods with the aim of studying certain production methods (such as environmental growing conditions) for their effects on composition, FNDDS where the composition data is calculated such that it is representative of the diets reported in the What We Eat in America survey (and not analytically measured, for example "asparagus cooked with fat" is a sum of the composition of cooked asparagus and composition of a non-specific fat which is a weighted sum of various consumed fats) and Branded Foods datasets has

commercially available industrially processed foods[4]. **Figure 2.2** gives a breakdown of the SR dataset and our selection, where there are 1546 raw foods, 384 cooked foods by the wet heat process, 806 cooked foods by the dry heat process and the remaining 5057 foods were made by other processes. For our models, we selected a subset of the SR dataset according to the following criteria. We matched raw/cooked food pairs, where the raw foods were a single ingredient harvested from a plant or from an animal (includes butchery products), and the cooked food was the outcome of the raw food treated to wet (boiling, steaming), or dry (roasting, grilling, broiling) heat processes. Foods were excluded from the dataset if either there was no single-ingredient raw food corresponding to the cooked food and vice-versa, or the foods had several ingredients and produced by a multi-step process like 'Luncheon meat, pork and chicken, minced, canned, includes SPAM Lite', 'Bread, banana, prepared from recipe, made with margarine'. We excluded processes which have added ingredients such as oil for frying and stir-frying although they these are common methods for cooking since we did not have data on the composition of the oil used in the process. We included boiling and steaming (simple aqueous, i.e., wet heat processes), as well as roasting, broiling and grilling (dry heat processes). This resulted in 840 foods total in the dataset, with 178 and 247 pairs from wet and dry heat processes, respectively. In this dataset, all plant-based foods were cooked by wet heat process (WH), and all animal-based foods by a dry heat (DH) process. (This congruence is a limitation in this dataset and is addressed in the Discussion.) The categorical breakdown of the number of pairs for plant-based and animal-based foods is shown in **Figure 2.2B**.

The composition data consists of content values for up to 232 ‘chemical constituents’ or ‘components’, which include specific chemicals (vitamins, amino-acids, fatty acids, etc.) and aggregated chemicals or chemical groups (total fats, total proteins, etc.) for every food. Here, we selected the components that are reported for at least 80% of the foods in our dataset. This resulted in 27 components per food, namely nine vitamins, 10 minerals, water, and seven aggregates of total protein, total carbohydrates and various fat categories (**Supplementary materials**). This composition data was used to train the prediction models where the input feature set to every model is the content of the 27 components in the raw food and the outputs are the contents of the 14 micronutrients in the cooked food. For this study, the macronutrient composition data in the cooked food is not predicted by the model, however this data is important for the data preprocessing explained next. Prior to model fitting, the composition data should be preprocessed to adjust for the bias resulting from the conventional format of representing nutrient contents per 100 grams of a food sample. In actuality, the cooked food sample would have a higher yield in the wet-heat process compared to the raw food sample primarily due to the gain of water and a lower yield in the dry-heat process due to the loss of fat and water. Scaling the true weight to the 100g representation in case of the higher yield creates an underrepresentation of the solid components. In the case of the lower yield the 100g representation creates an overrepresentation, which was observed as higher nutrient contents in the cooked food sample relative to the raw food sample. Ideally the data preprocessing would reverse this scaling effect. We use two different scaling methods, solid content scaling in Equations 1 and 2 and process-invariant nutrient scaling in Equations 3 and 4. For the solid content scaling (SCS), the assumption is

made that the water and fat contents remain unchanged from the initial (raw) to final(cooked) state of the food, and as per Equation 1 the content in the raw food is set to match that in the cooked food. Then the contents of the other components in the raw food are scaled to compensate for the difference (R[water]-C[water]) while preserving their initial proportions as per Equation 2. Equations1 and 2 are applied twice, once to equalize water and then to equalize fat, and the resulting scaled data is not affected by the order. This scaling method mitigates the over/under representation effect caused by gain/loss of water/fat. The second method attempts to identify the unknown yield factor (for the cooked food) as per Equation 3 and is based on identifying a nutrient that is largely invariant to processing. This factor is then used as per Equation 4 to derive the composition for the "true" weight of the cooked food corresponding to a 100-gram sample of raw food. The concept of such a nutrient is an exception since processing creates the conditions for nutrient transformation through chemical reactions and loss through solubilizing and leaching in the water and fat. An exception is cholesterol in meat which is theoretically invariant to processing since it is in the muscle-cell membranes that are resistant to cooking loss. However, the experiments report a small loss[5], so we also scale the data for a 5% loss and consider whether models are significantly different in reporting our results. The data for the cuts of beef used in this study are from experimental studies published by USDA where it is reported that contents of Iron and zinc contents were not significantly different in the raw and cooked beef[6]. There is no information on the components in plant-based foods. For confirmation of these hypotheses, all components are used in the PINS method and prediction performance is compared for both animal and plant-based foods. To be clear, the aim of

these scaling methods is creation of alternate versions of the composition data that represent the yield information that was missing in the original FDC data. In the Results section we compare the model performance for these different versions of the data. A detailed explanation of scaling with examples is in the **Supplementary Materials**.

In the Equations for scaling methods, *R* represents the raw food and *C* represents the cooked food, *R'* and *C'* represent the scaled data, and *X* is the generalized term for the components. In Equation 2, the summation term does not include water, and for the next step of equalizing the fat , the summation term would exclude water and fat.

$$R'[water] = C[water] \quad (1)$$

$$R'[X] = R[X] * (1 + \frac{R[water] - C[water]}{\sum R[X]}) \quad (2)$$

$$ScalingFactor = \frac{R[Cholesterol]}{C[Cholesterol]} \quad (3)$$

$$C'[X] = ScalingFactor * C[X] \quad (4)$$

All versions of the composition dataset include 425 pairs of foods, with 27 components, five processes (boiling, steaming, roasting, grilling and broiling), in two states (raw and cooked). **(Supplementary materials).**

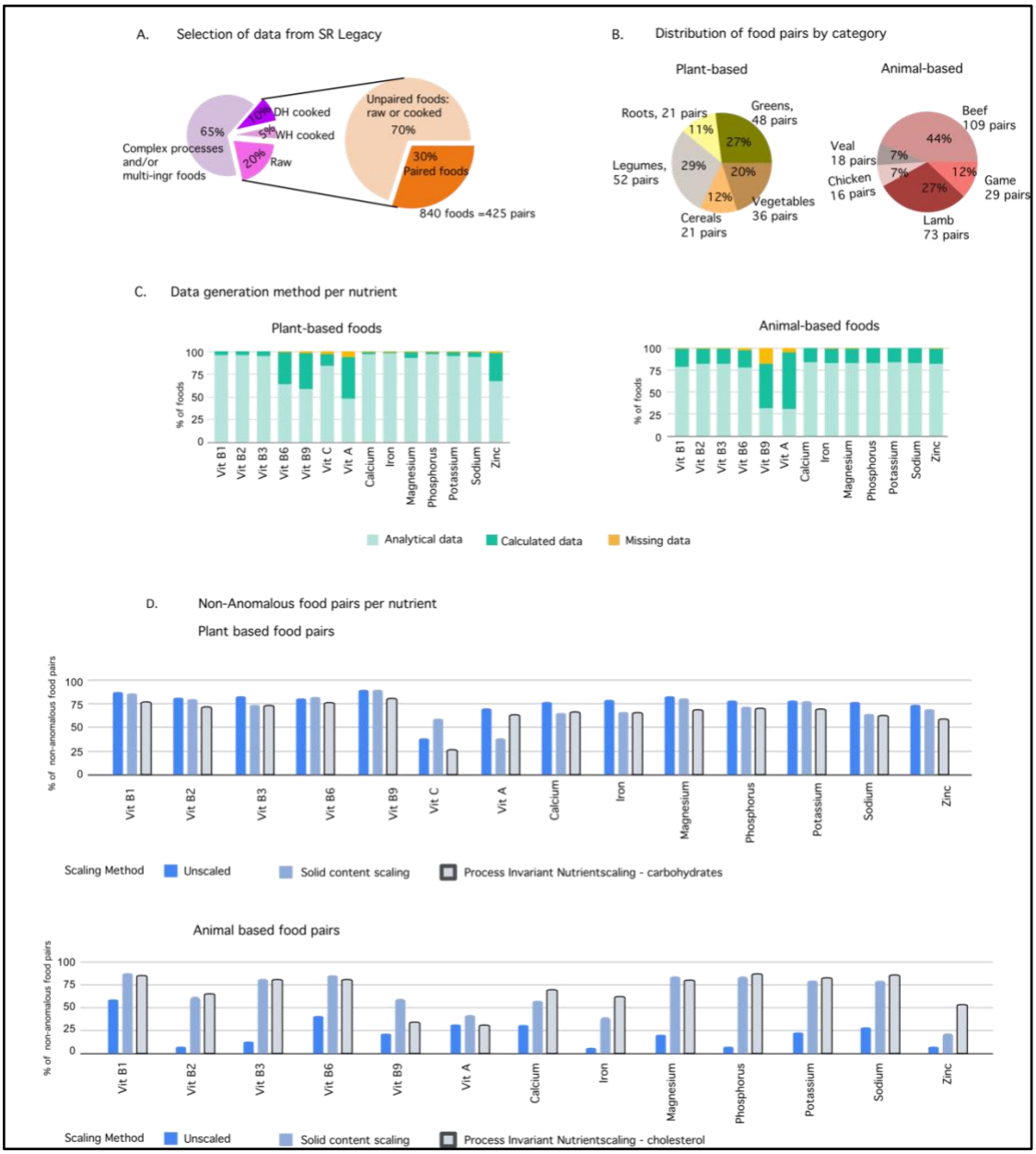


Figure 2.2 Data Review. (A) Out of 7,793 foods in the SR Legacy datatype in FDC dataset, 2,724 (35%) are single ingredient foods. Within that set, we identified 425 pairs of raw-cooked single ingredient foods. (B) The food pairs per category for plant-based and animal-based foods. There are a total of 178 pairs of plant-based foods and 247 pairs of animal-based foods. (C) The food-pair distribution by the method of data

generation. (D) Comparing the percentage of food-pairs of non-anomalous data by scaling method.

2.2.2 Models

We trained models to predict the content of 14 micronutrients for which we had baseline retention factors in the cooked food. Of those, seven are vitamins, namely vitamin B1 (thiamin), vitamin B2 (riboflavin), vitamin B3 (niacin), vitamin B6 (pyridoxine), vitamin B9 (folate), vitamin C (ascorbic acid), vitamin A, and the other seven are minerals, namely calcium, iron, potassium, phosphorus, magnesium, sodium and zinc. We created separate models based on the process category (wet, dry), as these are fundamentally different processes, but not based on the actual process (e.g., boiling vs. steaming), as there are not sufficient data per process to avoid overfitting. All models have the same input, which is the composition of the raw food, as illustrated in **Figure 2.1**. Other details that might be informative to the task (cooking time, temperature, water content) were not available in the SR legacy dataset, and consequently were not present in our dataset, or our model. Since vitamin C is not present in meats (which are all the foods for DH models), the dry heat models are only 13, for the other micronutrients, resulting in 27 models total (13 for DH and 14 for DH). These sets of WH and DH models were trained and tested on scaled variants of the dataset explained earlier. We applied a filtering step to the scaled datasets to select the pairs of foods where the nutrient being predicted was more in the raw food than in the cooked food. The unscaled data for the dry heat models and wet heat models was not filtered for this condition. So, each of the nutrient models were trained on different subsets of the data and is the reason that we

did not have a single model to predict all nutrients. The effect of the data scaling and filtering on the predictive models is explained in the **Results**.

The best performing model (for any dataset variant) was selected based on a cross validation grid search across six regressor types (MLP, LASSO, Elastic Net, Gradient Boost, Random Forest, Decision Trees), each with a variety of hyperparameters totaling 12,336 regressors where the metric for the best model was the least root mean squared error (RMSE). This was done for each of the 27 models using the sklearn library[7] and the best hyperparameters for each of the regressor types along with the RMSE is in **Supplementary materials**. We then performed a feature selection technique, a recursive feature elimination variant as described in the sequential feature selector function of the mlxtend package[8]. The model performances for data variants for the WH and DH process are compared in **Table 2.1**.

We assessed the predictive performance (RMSE) in comparison with two baseline models. The first is to naively assume that the dependent variable (the micronutrient to predict after cooking) is equal to its value in the raw food. This baseline serves as a comparison to a naïve regressor where the retention factor (RF) is 100%, i.e., the amount of the micronutrient after the heat process is the same as in the raw food. The second baseline was based on the USDA Retention Factors table, a common, standard model for the retention of nutrients after a process . The nutrient outputs were computed as a product of the RF for the specific nutrient and the content of that nutrient in the raw food. We use RSME, the coefficient of determination (R^2), Pearson Correlation

Coefficient (PCC), and Spearman Rank Correlation Coefficient (SRC) to assess the performance of our regressor model (**Table 2.2** and **Supplementary materials**). At each case, we performed 5-fold cross validation runs, bootstrapped 50 times to avoid overfitting and increase the generalization potential of our classifiers. For a subset of foods (**Supplementary materials**), we provide a higher resolution baseline using retention factors from experimental studies in literature. Finally, we analyze the prediction performance through a breakdown of R^2 by food category for plant-based foods (Leafy greens, Roots, Vegetables, Legumes, Cereals) and animal-based foods (Beef, Lamb, Chicken, Veal) as shown in **Table 2.3**. We do this by tagging every predicted micronutrient value by the category (associated with the food) and calculate the R^2 for every group. This is repeated for all predictions, and the average R^2 of a category is used to determine the best and worst performances in the plant-based and animal-based foods.

2.3 Results

2.3.1 Approximately 10% of SR Legacy foods can be paired to be used in model training.

The single ingredient foods that are either raw or cooked were found in 35% of the SR legacy data, and 30% of these were paired into raw and cooked samples. The final selection of 840 foods (or 425 pairs) is 10% of SR legacy data (**Figure 2.2A**), with an unequal distribution of data pairs by food category (**Figure 2.2B**). We identified an anomaly where the content of a micronutrient was more in the cooked food than in the raw food in 50% of the pairs on average across the 14 micronutrients. The anomaly was

more severe for the animal-based foods (77% vs 23% pairs, respectively; see **Supplementary materials**)**.** This was partially caused by the bias introduced by the data representation convention. For the animal-based foods, the non-anomalous pairs are 23% of the total pairs for unscaled data and increase to 70% for PINS-cholesterol scaled data, p-value<0.001. This is reasonable, since the anomaly is due to a concentration bias (nutrient content in cooked food is more than in raw food), which is mitigated by scaling. For the plant-based foods, there is no significant change(p-value>0.09) in non-anomalous pairs using the scaling methods for plant-based foods, since the issue is a dilution bias which is mitigated however this does not cause an anomaly (nutrient content in cooked food is more than in raw food). The comparison of non-anomalous pairs for animal and plant-based foods is shown in **Figure 2.2D**. The **Discussion** section explains the reasons for this differing effectiveness of the scaling methods in reducing the bias and suggests other possible causes for the bias.

*2.3.2 Scaling improves model performance**.***

We trained predictive models on variants of the datasets as explained in Methods. The dry heat models (broiling, grilling, roasting processes; 247 animal-based foods) and wet heat models (steaming, boiling; 178 plant-based foods) were trained on the unscaled data, which is not filtered for the anomalous condition, and on data from both the scaling methods which is filtered for non-anomalous data. We use the metric RMSE to compare model performance. For the dry heat models, the average RMSE was 32% lower when the model was trained on data scaled by the PINS-cholesterol method than data scaled using SCS method, which had 18% lower RMSE compared to the model trained on

unscaled data. (The model prediction results were not significantly different for the data scaled for a constant cholesterol content and scaled for a 5% loss, so the results are reported for the former.) Although the model performance based on PINS data for iron and zinc has lower average RMSE than cholesterol, we consider the model trained on PINS-cholesterol as the best model since there is a mechanistic explanation described in **Methods**. For the wet heat models, the average RMSE was 55% lower when the model was trained on SCS data than that on unscaled data. These comparisons are shown in **Table 2.1**, and all results are in **Supplementary materials** and further analysis is in **Discussion**. The best model for the wet heat process is trained on SCS data and for the dry heat process it is trained on PINS-cholesterol data. We now compare results from the best predictive ML models to the baseline model.

OUTPUT	WETHEAT		DRYHEAT				
	Unscaled	SCS	Unscaled	SCS	PINS-Zinc	PINS-Iron	PINS-Cholesterol
Thiamine	0.04	0.02	0.03	0.02	0.01	0.01	0.01
Riboflavin	0.05	0.04	0.05	0.02	0.02	0.02	0.02
Niacin	0.48	0.21	0.87	0.68	0.46	0.60	0.45
Vit B6	0.05	0.03	0.09	0.07	0.07	0.05	0.06
Folate	22.37	16.64	4.38	4.34	6.46	3.74	1.72
Vit C	13.28	7.49					NA
Vit A	83.21	11.57	3.37	2.66	1.25	2.81	1.5
Calcium	22.17	14.28	4.33	3.5	2.41	1.36	1.81
Iron	0.6	0.30	0.33	0.31	0.14	0.00	0.16
Magnesium	11.19	6.66	5.05	3.98	2.19	2.22	2.33
Phosphorus	24.1	12.94	24.16	22.12	15.60	15.67	21.41
Potassium	101.9	46.95	48.87	39.2	27.49	30.95	32.23
Sodium	17.4	15.84	13.26	9.90	6.80	7.35	9.36
Zinc	0.2	0.10	0.67	0.51	0.00	0.38	0.29
AVERAGE	21.22	9.5	8.11	6.71	5.24	5.43	5.49

Table 2.1. Comparing models trained on different data variants. The prediction performance results for the models trained on data variants specified in the Methods are shown in this table. The metric for model performances is RMSE – Root mean squared error. A complete coverage of all performance for all PINS data is in **Supplementary materials**. Abbreviations for Data Variants listed in the table: Unscaled is the original data. SCS – Solid content scaling with water and fat equalising. PINS – Process Invariant Nutrient scaling with the specific nutrients.

Outputs		Wet heat (Steaming, Boiling)					Dry Heat (Broiling, Grilling, Roasting)			
	Avg+-Stdev	RMSE				Avg+-Stdev	RMSE			
	True Data	Predicted	Baseline	RF100	Rel %	True Data	Predicted	Baseline	RF100	Rel%
B1(Thiamine)	0.11+- 0.08	**0.02**	0.03	00.06	17.85	0.07+-0.04	**0.01**	0.02	0.04	14.69
B2(Riboflavin)	0.08+-0.08	**0.04**	0.05	0.05	23.31	0.22+-0.1	**0.02**	0.03	0.04	28.03
B3(Niacin)	0.70+-0.53	**0.21**	0.44	0.89	52.29	4.52 +-1.2	**0.45**	0.53	1.19	15.04
B6	0.13+-0.12	**0.03**	0.03	0.09	13.13	0.29+-0.14	**0.06**	0.11	0.12	48.81
B9(Folate)	47.51+-47.35	**16.64**	24.03	26.35	30.74	10.99+-4.77	**1.72**	2.16	5.00	20.35
C	16.38+-18.96	**7.49**	7.71	15.10	2.85		Not a significant source			
A	69.32+-100.87	**11.57**	14.71	11.38	21.39	3.83+-4.62	**1.50**	1.58	2.99	4.97
Calcium	43.72+-57.37	**14.28**	35.47	25.47	59.73	9.47+-4.57	**1.81**	3.13	3.61	42.17
Iron	1.22+-0.98	**0.30**	0.49	0.54	39.22	1.74+-0.72	**0.16**	0.24	0.26	31.85
Magnesium	35.01+-22.92	**6.66**	8.91	9.18	25.30	17.96+-2.95	**2.33**	3.33	6.05	30.03
Phosphorus	70.60+-51.38	**12.94**	19.61	29.36	33.98	159.44+-24.64	21.41	**17.59**	34.89	-21.71
Potassium	271.98+-168.66	**46.95**	59.73	81.75	21.40	325.13+-86.05	32.23	**30.14**	66.83	-6.97
Sodium	21.71+-40.35	**15.84**	17.44	32.20	9.18	51.84+-11.34	9.36	**8.03**	16.64	-16.66
Zinc	0.54+-0.45	**0.10**	0.14	0.52	28.85	3.78+-1.67	**0.29**	0.53	0.52	45.38
AVERAGE		**9.50**	13.48	16.64	29.52		**5.49**	5.19	10.63	18.15

Nutrient	Prediction model	Baseline (USDA RF table)	Baseline (RF from experiments)
Vitamin C	10.50	11.25	13.31
Folate	25.84	40.65	97.22

Outputs Metric :R2	Wet heat (Steaming, Boiling)			Dry Heat (Broiling, Grilling, Roasting)		
	Predicted	Baseline	RF100	Predicted	Baseline	RF100
B1(Thiamine)	0.89	0.36	0.23	0.50	0.53	0.22
B2(Riboflavin)	0.72	0.95	0.89	0.77	0.80	0.80
B3(Niacin)	0.80	0.78	0.91	0.80	0.80	0.91
B6	0.86	0.40	0.74	0.58	0.38	0.66

B9(Folate)	0.86	0.42	0.56	0.42	0.80	-1.09
VitC	0.79	0.90	0.33	Not a significant source		
VitA	0.98	0.94	0.97	0.59	0.86	0.98
Calcium	0.92	0.74	0.64	0.73	0.53	0.73
Iron	0.88	0.85	0.76	0.76	0.80	0.30
Magnesium	0.87	0.87	0.73	0.13	-0.39	0.07
Phosphorus	0.90	0.50	0.87	-0.42	0.35	0.64
Potassium	0.90	0.27	0.52	0.23	0.44	0.75
Sodium	0.67	0.46	0.20	-0.09	0.44	0.49
Zinc	0.93	0.88	0.95	0.89	0.90	0.97

Table 2.2. Results of prediction models compared to baselines. The prediction scores (RMSE and R^2) are the average of 50 runs, due to the inherent randomness in the models. [A] (RMSE) of best prediction models, compared to baseline (USDA's RF guide Version 6) model and naïve model (output content=input content). The better of the prediction or baseline score is highlighted. The rel% column is calculated as : (baseline-predicted)/baseline*100 1B. Additional baseline model for vitamin C (ascorbic acid) and vitamin B9 (folate) using RF values from experiments on selected foods. 1C. The metric R^2 (coefficient of determination) is scale invariant (as opposed to the RMSE) for ease in comparison across all predictions. The corresponding box plot is in Figure 2.3.

2.3.3 The predictive model performs 30% and 18% better than using the standard USDA Retention Factor model for wet and dry heat processes, respectively.

We compared the predicted concentrations of the micronutrients in the cooked foods for both the wet heat processes and the dry heat processes against the two baseline models, as described in the **Methods** section. When compared to the naïve baseline (i.e., retention factor is always 100%), the predictive model is better in 26 out of the 27 comparisons (96%; RMSE of 7.57±11.42 vs 13.75±20.86, respectively; 43%% decrease

of RMSE on average for wet heat, p-value < 0.01; 49% decrease in RMSE on average for dry heat, p-value < 0.01). Then, to compare with the standard practice, we computed micronutrient concentrations using the USDA's Retention Factor table (see **Methods**) as shown in **Table 2.2**. In that case, the predictive model was better than this baseline in 24 out of the 27 comparisons (89%; RMSE of 7.57±11.42 vs 9.49±14.31, respectively; 30% decrease of RMSE on average for wet heat, p-value < 0.01; 18% decrease in RMSE on average for dry heat, p-value < 0.01). **Figure 2.3** depicts the correlation between predicted and actual (ground truth) values for the 14 micronutrients, for both the ML model and the USDA retention factor baseline. Next, we investigated the difference in the predictive performance when curating retention factors from literature. For this, we identified the retention factors of vitamin C (ascorbic acid) for 12 sample foods (green beans, beet greens, broccoli, Chinese cabbage, carrots, cauliflower, mustard greens, green peas, green peppers, pumpkin, spinach, zucchini) and of vitamin B9 (folate) for 12 sample foods (amaranth leaves, broccoli, drumstick leaves, snap beans, lentils, okra, onions, potatoes, green peas, soybeans spinach, taro leaves) (see **Supplementary materials**). In both cases, the ML model had a better agreement with the ground truth data than the Literature Retention factor baseline, although less so for vitamin C (for vitamin C (ascorbic acid), RMSE 10.51 vs 13.31, p-value=0.026; for vitamin B9 (folate) RMSE 25.84 vs 97.22, p-value=0.013). Note that retention factor information for each micronutrient is not available for the majority of foods, and it is a time consuming and expensive process to measure it. Using scale-invariant metrics reach the same conclusions (see **Supplementary materials)**. The

Discussion section elaborates further on the reasons that any RF baseline method is error prone and not appropriate to compute nutritional baselines.

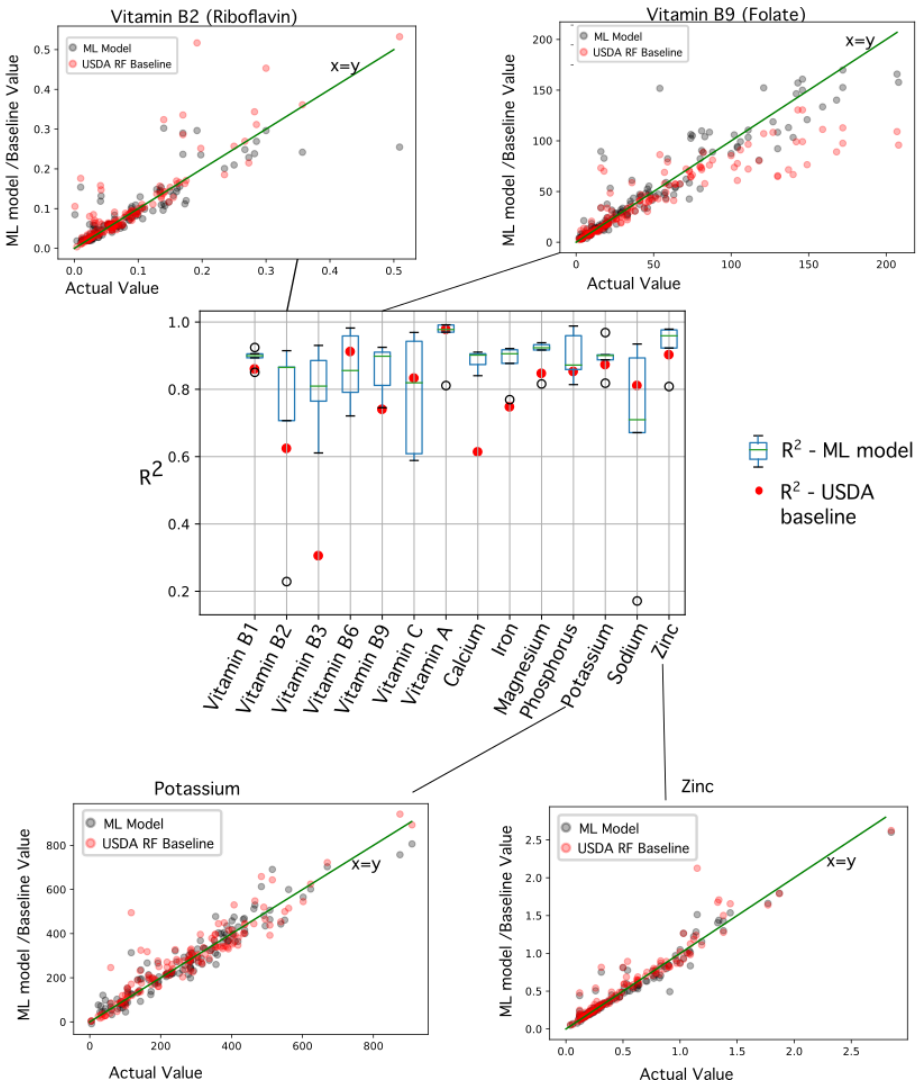


Figure 2.3. Model Performance Analysis. Centre: Comparing box plots of R^2 (coefficient of determination) for the ML prediction models and R^2 for the corresponding USDA baseline model. Details of the predicted values are shown in scatter plots, where the values from the prediction models and USDA baseline model are plotted against actual values (ground truth), and the x=y line represents the perfect computed value. In

the top two scatter plots, the ML model performance is better than the baseline. Plots for vitamin B9 (folate) shows baseline values below the x=y line, that is lower values than the predicted values, relative to the actual data. The lower two scatter plots are for the case where ML prediction was better by a small margin. Plots for potassium and zinc have a noticeable overlap in values for the prediction model and baseline.

2.3.4 Prediction performance is best for legumes, and worst for cereals, in the plant-based food categories, and best for beef and worst for veal in the animal-based food categories.

As reported in prior literature, the food structure/phenotype influences the chemical and physical changes that occur in food processes. Here we use the food category to represent this concept and show the differences in predictability. We group the 14 predicted micronutrient values by the food category and calculated the metrics (**Table 2.3** and **Figure 2.4A**). Legumes have the highest average R^2 of 0.82±0.12 and leafy greens have the lowest average R^2 of 0.29±1.10). In the dry-heat processed animal-based foods, beef had the highest average with R^2 of 0.48±0.46 and veal the lowest average R^2 of - 0.50 ±1.21. Due to the uncertainty associated with methods of data generation, USDA specifies the nutrients with most reliable data, these are vitamin B3 (niacin), vitamin B6, calcium, iron, and zinc. The highest average R^2 is now 0.85 ±0.08 for beef and the lowest is -0.06 ±1.26 for veal. As such, the nutrient loss is better predicted in legumes vegetables and beef given the current training data.

Output	Legumes			Greens			Roots			Vegetables			Cereals		
	RMSE	R^2	PCC	RMSE	R^2	PCC	RMSE	R^2	PCC	RMSE	R^2	PCC	RMSE	R^2	PCC

Ca	12.47	0.87	0.94	14.11	0.94	0.97	10.72	0.80	0.90	26.30	0.92	1.00	8.94	-2.16	0.04
Fe	0.29	0.92	0.97	0.41	0.74	0.87	0.18	0.65	0.84	0.14	0.76	0.91	0.25	0.61	0.85
Mg	7.65	0.88	0.95	7.44	0.86	0.93	3.76	0.92	0.96	9.29	0.53	0.82	6.88	0.84	0.92
Ph	13.22	0.93	0.97	6.89	0.87	0.95	7.63	0.84	0.93	19.73	0.48	0.75	15.74	0.80	0.91
K	48.22	0.81	0.93	47.79	0.93	0.97	48.54	0.91	0.96	57.10	0.70	0.84	34.97	0.25	0.54
Na	4.27	0.82	0.93	32.09	0.75	0.88	6.68	0.91	0.96	4.66	0.96	0.98	30.46	0.63	0.88
Zn	0.13	0.94	0.97	0.10	0.77	0.90	0.06	0.57	0.82	0.05	0.98	0.99	0.16	0.44	0.67
A	12.80	0.47	0.71	17.42	0.98	0.99	3.14	0.99	1.00	6.70	0.99	1.00	5.06	-2.11	-0.08
C	3.72	0.88	0.94	11.74	0.64	0.81	5.47	0.66	0.94	10.26	0.84	0.97	NA	NA	NA
B1	0.03	0.83	0.92	0.02	0.92	0.97	0.01	0.90	0.98	0.02	0.78	0.90	0.03	0.83	0.94
B2	0.01	0.93	0.97	0.06	0.62	0.80	0.04	0.42	0.76	0.04	0.57	0.85	0.01	0.96	0.98
B3	0.25	0.73	0.86	0.12	0.74	0.86	0.13	0.82	0.92	0.22	0.75	0.89	0.32	0.84	0.93
B6	0.02	0.72	0.86	0.02	0.99	1.00	0.02	0.95	0.99	0.06	0.13	0.72	0.01	0.85	0.93
B9	23.76	0.79	0.89	19.89	0.65	0.82	6.69	0.86	0.98	7.57	0.72	0.85	6.13	0.95	0.98

Output	Beef			Lamb			Chicken			Veal		
	RMSE	R^2	PCC	RMSE	R^2	PCC	RMSE	R^2	PCC	RMSE	R^2	PCC
Ca	1.85	0.76	0.88	1.93	0.87	0.83	2.27	0.04	0.59	3.52	0.68	0.78
Fe	0.13	0.89	0.95	0.16	0.25	0.67	0.08	0.84	0.95	0.22	-1.63	0.34
Mg	2.68	0.03	0.48	1.50	0.31	0.57	2.22	0.43	0.67	4.48	0.10	0.37
Ph	18.31	0.32	0.70	19.91	-0.44	0.48	25.51	-0.06	0.73	46.28	-1.86	-0.12
K	33.23	0.19	0.60	32.69	0.31	0.63	32.81	0.15	0.88	37.47	0.25	0.91
Na	9.50	0.25	0.65	7.15	-0.27	0.35	13.65	-0.51	0.46	14.97	-1.00	0.33
Zn	0.38	0.95	0.98	0.24	0.91	0.96	0.15	0.80	1.00	0.17	0.93	0.99
A	1.24	0.39	0.90	1.67	0.80	0.91	3.67	0.73	0.86			
B1	0.01	0.60	0.80	0.01	0.73	0.87	0.01	0.62	0.92	0.02	-0.29	0.56
B2	0.02	0.95	0.98	0.01	0.94	0.97	0.03	-0.10	0.94	0.06	-0.06	0.47
B3	0.47	0.79	0.89	0.40	0.70	0.84	0.85	0.82	0.95	0.41	0.92	0.97
B6	0.05	0.78	0.88	0.04	0.89	0.95	0.02	0.83	0.95	0.15	-1.24	0.47
B9	0.99	-0.63	0.54	1.74	0.67	0.83	0.91	-2.81	-0.64	2.62	-2.82	0.91

Table 2.3. Various metrics (R^2 ,RMSE,PCC) by category for plant-based foods(top), and for animal-based foods(lower).

Cereals do not have data for vitamin A and C predictions. Abbreviations are used for the predicted nutrient, Ca:Calcium, Fe:Iron, Mg:Magnesium, Ph:Phosphorus, K:Potassium, Na:Sodium, Zn:Zinc. The remainder are vitamins.

2.3.5 High variability on the top predictive features.

There is a notable lack of feature importance order across the prediction models. **Figure 4B** shows the feature ranks, where the features are ordered by their average rank across predictions. The average rank is in the mid-range for both the WH and DH process, suggesting that no feature has a consistent importance across all the predictions**. Figure 4C** shows performance by feature-size plots for vitamin B6 and potassium (WH) and vitamin B6 and zinc (DH) and the feature names are listed in the caption. The common observation is that the top ranked feature is the micronutrient itself in the raw food, as expected, but all other input features are specific to every prediction. The complete coverage of best features and feature ranks is in the **Supplementary materials**.

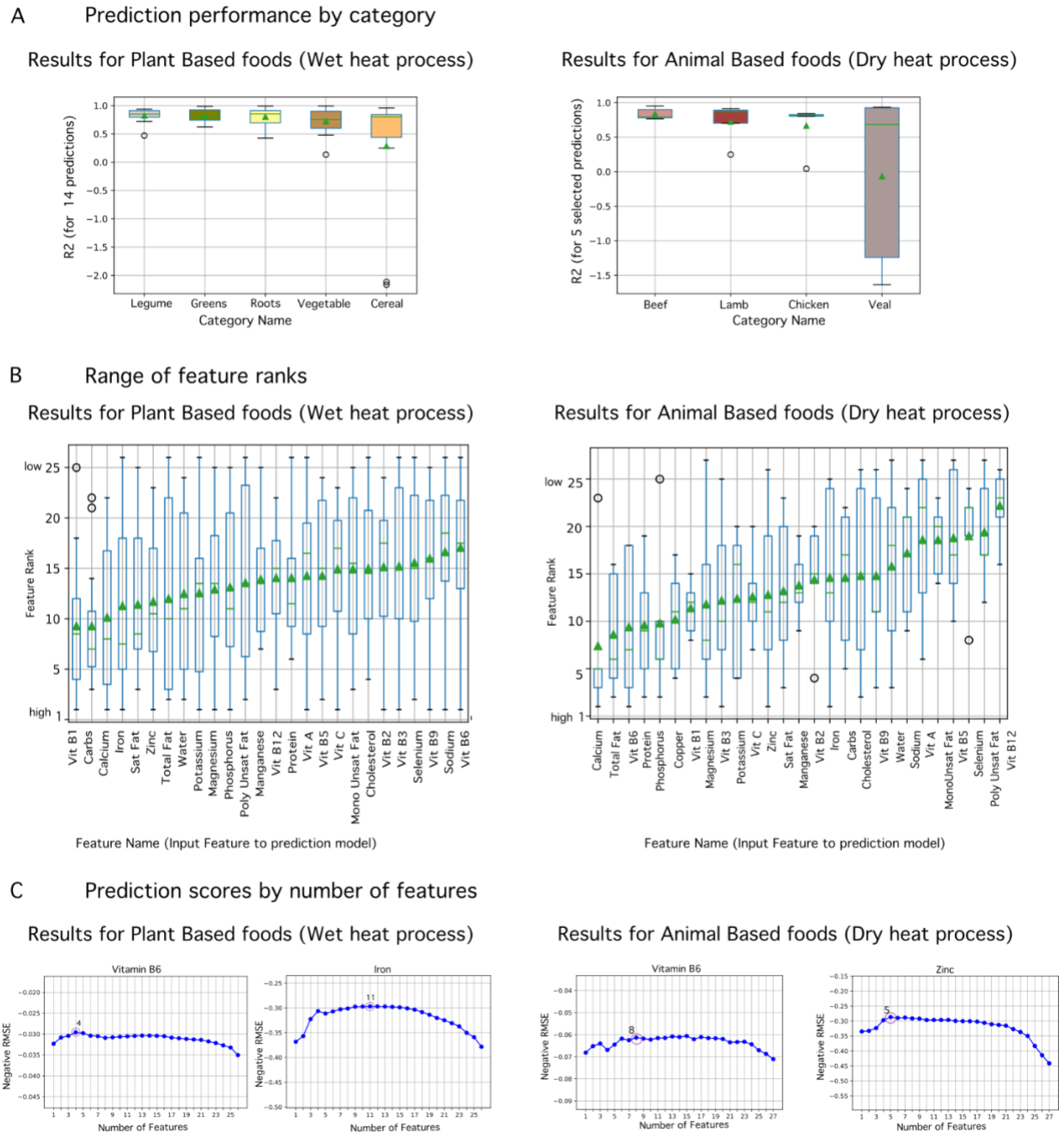


Figure 2.4: Results. (A): Box plot of R^2 for predictions by food category. For the plant-based foods the box plot shows all 14 predictions. Legumes have the best performance and Cereals the worst. For the animal-based foods, only five predictions are considered since they have the most reliable data as mentioned in Results. Beef has the best performance and veal is the worst (B): Box plot of feature ranks for the input features,

where rank one is highest. Features are arranged in ascending order of average rank. Average ranks for both plant-based foods (and WH process) and animal-based foods (and DH process) are in the mid-range. No feature has a consistent high rank cross all the predictions. (C): Plots of performance-vs number of features. Vitamin B6 and iron are shown as examples for the WH process and vitamin B6 and zinc for the DH process. The best features for vitamin B6 (WH) are vitamin B6, vitamin B5, zinc, vitamin B1. Best Features for iron (WH) are iron, total fats, monounsaturated fat, zinc, water, carbohydrates, potassium, manganese, polyunsaturated fats, vitamin C, phosphorus. Best Features for vitamin B6 (DH) are vitamin B6, magnesium, calcium, vitamin B2 (riboflavin), calcium, total fats, vitamin C and carbohydrates. Best features for zinc are zinc, phosphorus, calcium, potassium and total protein. The combined interpretation of B and C suggests that feature selection results differ for every nutrient prediction.

2.4. Discussion

The prior sections addressed the methods to building predictive ML models for the micronutrient content in cooked foods and the discovery of a data scaling method to remedy the bias of unknown yield factors. The results proved that this novel method outperforms the baseline method, which is significant since it offers the potential to scale across diverse foods without compromising the accuracy. However, realizing this potential, requires larger datasets than currently available. Accordingly, this section delves into the observed limitations of the SR Legacy dataset and interpretation of the results, with the aim of providing guidance to the future efforts of building larger food

composition datasets[4,10–12] since the data generation process is time consuming and expensive.

Regarding the predictive performance, we elaborate on some causes for the lower performance of the baseline methods. The scatter plots in **Figure 2.3** for vitamin B9(folate) and B2 (riboflavin) show that the baseline method underestimates the composition, which implies that the baseline RF is less than the RF inherent in the true data. RF represents the rate of loss which is influenced by process-related factors like processing times, surface area of vegetable exposed to processing conditions. Ideally for a fair comparison, these factors should be known for the baseline and matched to the data at hand. This can easily be addressed by recording additional meta-data. However, the more challenging discrepancy was that the baseline is a simple linear method, while the prediction model is a much more complex multiparametric non-linear ML model. Inevitably more sophisticated methods will emerge whether machine learning, mechanistic or a hybrid, and a suitable state-of-the-art baseline method will be available for comparison.

The current dataset has been the primary food composition dataset in the US for several decades, however it has several gaps in the data structure and data sampling that are regarded as necessary for datasets in current times. We assess these limitations to inform methods in building future datasets; the selection of food samples, recording of structured metadata/provenance, checking for data quality, and determining the composition features. The provenance of the data was incomplete in at

least two different aspects. The composition data was calculated for some foods, and there was no explanation for the calculation method and no mention of the reference food /data used in the calculation method. It is unclear whether the samples for the raw and cooked food were related. Additionally, ontologies or structured vocabularies are a valuable resource when creating a format or structure for the dataset. Regarding data quality, we have described the anomalous condition in the **Results**. This is an example of a basic data sanity check, and especially in the context of a prediction hypotheses. Predictive performance depends on both the sample size as well as the entropy of the dataset, and one can use the predictive performance of the model as a guide for the sampling size for gathering new experimental data. There was only a single representative instance for each food and factors like geography, method of agriculture etc. are known to significantly impact the composition. The congruence of food-source and cooking method (plant-based foods were cooked by wet heat methods and animal-foods are cooked by dry heat methods) makes it impossible to compare model performance by either variable independently. While animal-based foods are often cooked in dry heat conditions, plant-based foods are also cooked by these methods, so this omission is also relevant to dietary representation. From the perspective of data modelling, it is especially disappointing, since we discovered that prediction performance varies by category within a given source. Such results could increase our knowledge of nutrient loss and designing prevention strategies, as well as provide hypotheses for greater food sampling. Regarding the feature space per sample, we suggest including process parameters and features known to influence nutrient loss such as pH.

Finally, we address some details of the anomaly caused by the representation of the composition per 100g of food and unknown yield factors. This issue was mitigated by data scaling methods; however our observations show that this is not a complete resolution and new standards for data representation are required. There are two effects from applying the scaling methods on the composition data; the effect on the size of non-anomalous food-pairs (**Figure 2.2D**) and the effect on model performance trained on this data (**Supplementary materials**). As seen in **Figure 2.2D**, there is no significant effect (p-value =0.06) for the plant-based foods where the scaling methods lowered the dilution effect caused by the data representation, so instead a possible reason for the anomaly could be different food samples used for the raw and cooked analysis. Whereas there is a significant effect (p-value <0.001) on animal-based foods where the anomaly is due to a concentration bias which could mitigated. Regarding the prediction performance, a few additional components used in the PINS method had good results besides the hypotheses. For plant-based foods, the performance for SCS data was the best, followed by carbohydrate PINS data. For animal-based foods, the performance by PINS-proteins data was the better than for zinc, iron and cholesterol. However, the results for PINS-carbohydrate and PINS-protein are likely due to the methods used for generating this data. Another possible solution might be to use yield factors when available, but since processing conditions are not available for SR data, we could use it. This analysis presents several questions for future inquiry, though the most important might be to ascertain a process-invariant nutrient and under which conditions and the biochemical/mechanistic explanation. This information might help for

data transformations of existing data, but new data representation standards need to be considered and applied to future data generation efforts.

In conclusion, ML models have the potential to complement experimental methods in predicting the effects of food processing. In addition, feature weights can be used to achieve the desired composition outcomes.

REFERENCES

1. USDA National Nutrient Database for Standard Reference, Legacy Release \textbarAg Data Commons.
2. Haytowitz, D. B. & Pehrsson, P. R. Present knowledge in nutrition—nutrient databases. in Present knowledge in nutrition 203–216 (Elsevier, 2020). doi:10.1016/B978-0-12-818460-8.00011-3.
3. Haytowitz, D. B., Lemar, L. E. & Pehrsson, P. R. USDA's Nutrient Databank System – A tool for handling data from diverse sources. J. Food Compos. Anal. 22, 433–441 (2009).
4. Fukagawa, N. K. et al. USDA's FoodData Central: what is it and why is it needed today? Am. J. Clin. Nutr. 115, 619–624 (2022).
5. Small, D. M., Oliva, C. & Tercyak, A. Chemistry in the kitchen. Making ground meat more healthful. N. Engl. J. Med. 324, 73–77 (1991).
6. Roseland, J. M. et al. Fatty acid, cholesterol, vitamin, and mineral content of cooked beef cuts from a national study. J. Food Compos. Anal. 66, 55–64 (2018).
7. Pedregosa, F., Varoquaux, G. & Gramfort, A. Scikit-learn: Machine learning in Python. J. Mach. Learn. 12, (2011).
8. Raschka, S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. J. Open Source Softw. 3, 638 (2018).
9. Nutrient retention factors : USDA ARS.
10. Ahmed, S. et al. Foodomics: A Data-Driven Approach to Revolutionize Nutrition and Sustainable Diets. Front. Nutr. 9, 874312 (2022).
11. Hinojosa-Nogueira, D. et al. Development of an unified food composition database for the european project 'stance4health'. Nutrients 13, (2021).
12. Desiere, F., German, B., Watzke, H., Pfeifer, A. & Saguy, S. Bioinformatics and data knowledge: the new frontiers for nutrition and foods. Trends Food Sci. Technol. 12, 215–229 (2001).

Chapter 3: Structure-property machine learning models with predictive capabilities for glycans in food: Case study of modeling starch in rice.[2]

3.1.Introduction

Demonstrating the potential of ML models for glycans: Case Study

The background to the research in structure-property machine learning models, and its application in modeling glycans was covered in **Chapter1**. The main motivation to model glycans is their abundance in sources of human food such as grains, vegetables, and legumes[1], and their roles in food formulation[2]. Generally, a glycan structure is characterized by repeating units composed of a monomer(s), linkages, and optional functional groups[3].As explained earlier in **Chapter 1 Section 1.3.2** we selected starch as the representative glycan for our case study, and summarize here the scope and aims. The dataset is for samples of rice and the data per sample includes the chain length distribution (size exclusion chromatography data) which represents the structural features, content of protein and amylose, the physical properties relevant to the processing quality of rice, and the sensory properties of cooked rice. The architectural overview of the data and models is illustrated in **Figure 3.1**. The models developed in this study address the following research questions (RQs):

[2] This chapter has been published as preprint. Co-authors, Gabriel Simmons, Department of Computer Science, UC Davis, Bruce German, Food Science and Technology, UC Davis. Pre-print on Bioarxiv,November 2023.DOI 10.1101/2023.11.12.566488

RQ1: Is size exclusion chromatography data and content data sufficient to predict both physical and sensory properties of cooked rice?

RQ2: Does gelatinization temperature and gel consistency information improve predictive performance over only the chain length and content data (tested in RQ1)?

RQ3: What features (structural or otherwise) are the most informative for each prediction?

These research questions address our main thesis that the key to predicting the properties is the detailed structural data. We also dive deeper into the structural features to gain insight into the molecular dynamics responsible for the properties. **Section 3.2** provides a background of prior research on starch, and its relevance to our methods. The data and methods used in this study are explained in **Section 3.3**. Our results in **Section 3.4.1** indicate that structure-composition data is a better predictor of final product texture properties than pasting behavior, contrary to the approach commonly taken in food product formulation[20,21]. Furthermore, the pasting characteristics and other physical properties can themselves be predicted from structure-composition data. The result from feature engineering in **Sections 3.4.2** and **3.4.3** identifies the most useful features for each predictive task. Finally, in **Discussion** we suggest strategies to guide future experimental studies towards improving structure property models (**Section 3.5.2**).

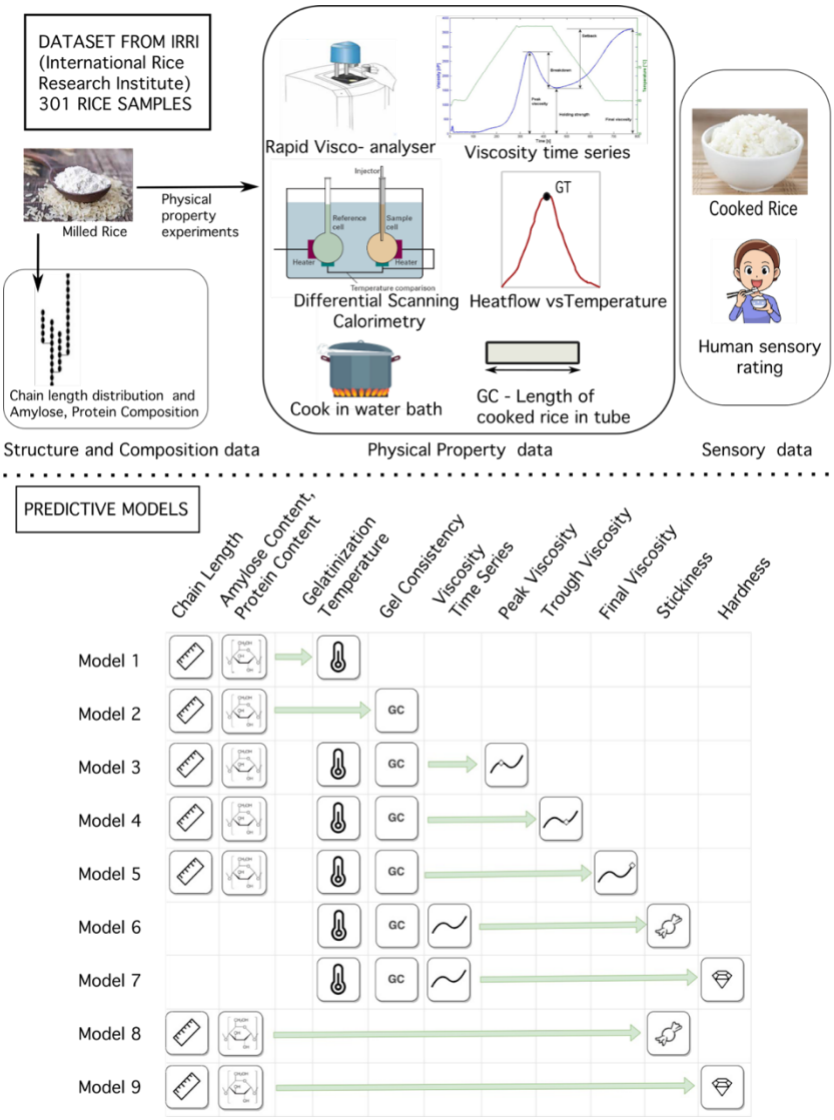


Figure 3.1: Data and Models. The data is from prior research, [4] for 301 samples of rice based on the indicated methods in the upper portion of the figure. 9 Models are trained based on this data as seen in lower half of the figure. Detailed explanation for the data and the models is in section 3.

3.2. Background

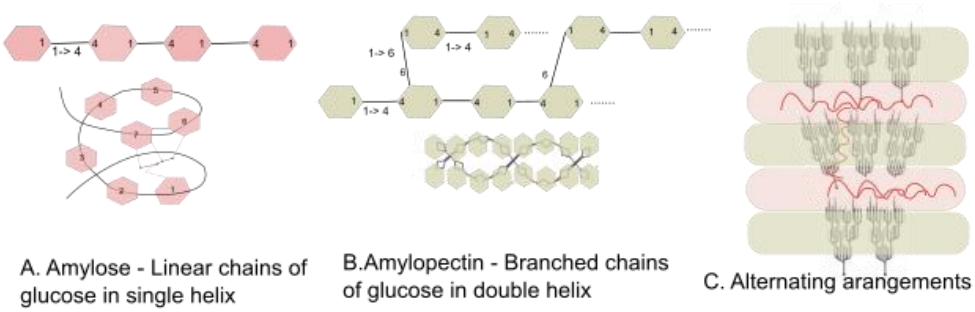
3.2.1 The Importance of Starch

Food sources of starch like tubers and grains provide up to 30% of daily dietary calories[5] and starch constitutes 70-90% of the composition in these foods[6] . Starch is structurally simpler than other glycans, consisting solely of a repeating glucose unit with no additional charge. Despite its simplicity, starch is also expressive - differences in its structure across food groups like grains, legumes, and tubers manifest a range of properties essential to food formulations such as binding ability, textural smoothness, and stickiness[7–17] . While our study uses data specific to rice, the features used to train our models are not specific to a single food group (since starch is the shared building block for a host of widely-consumed foods).

3.2.2 Starch Morphology

The structure of starch across units of scale starting from the glucose monomer is illustrated in **Figure 3.2**. It is composed of two fractions; neutrally-charged chains of glucose, arranged linearly as amylose (AM), and in a branched structure as amylopectin (AP). The consensus among researchers is that the degrees of polymerization (DP) of amylopectin is usually between 9 and 24 up to 100[8,18,19] while amylose has longer chains. These polymers assemble in an arrangement of crystalline regions composed of branches of amylopectin forming helices due to the strong attractive intermolecular forces, with longer chains forming more stable helices. The amorphous regions are composed of amylose where some long chains of amylose might also form helices. The long-chain helices, whether amylose or amylopectin, are positively correlated with

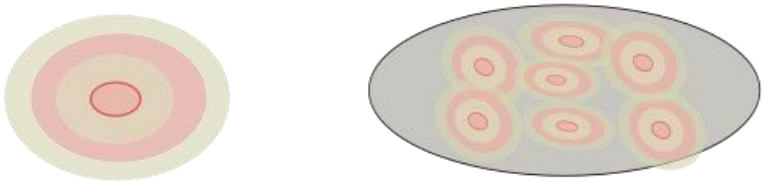
greater crystal stability[8,18,20]. The crystalline and amorphous regions repeat in concentric rings **(Figure 3.2C)**, with several of these arrangements packed into a granule **(Figure 3.2D)**. This highly ordered native arrangement is disrupted by food processing operations like milling, heating, or soaking in water.



A. Amylose - Linear chains of glucose in single helix

B.Amylopectin - Branched chains of glucose in double helix

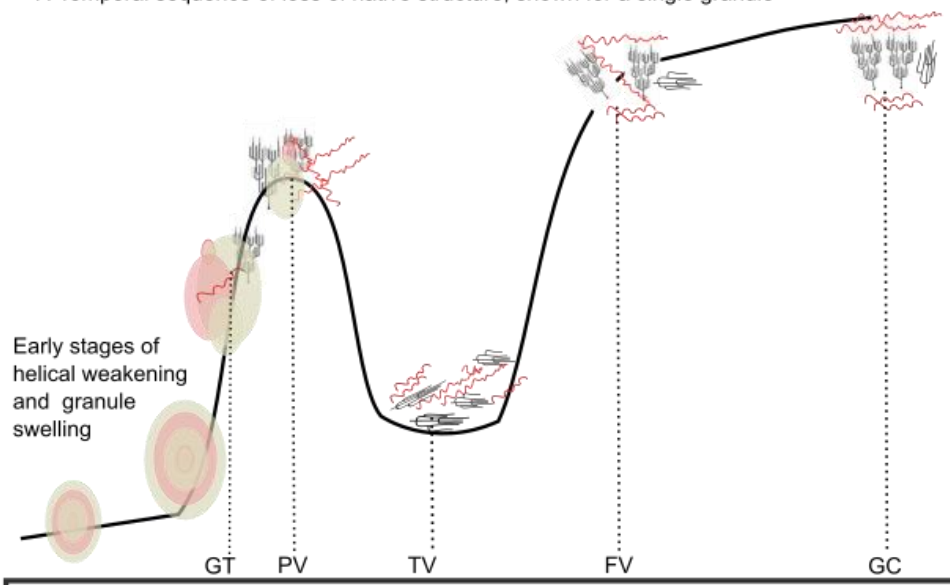
C. Alternating arangements



D. Starch granule in native state

E. Many granules in a grain of rice

F. Temporal sequence of loss of native structure, shown for a single granule



Analytical method	Differential Scanning calorimetry	Rapid Visco Analyser, measures viscosity profile for a 12 min period, in which the characteristic values are peak, trough and final viscosity	Length of cooled starch gel as prepared in a tube
Mechanism	Critical point in weakening of helical structures. Irreversible loss of native granule structure	Peak viscosity - amylose and amylopectin are dispersed in water and at maximal disordering Trough Viscosity - amylose and amylopectin are fully interacting with water, and some chains are aligned, reducing the viscosity. Final Viscosity - early stage of recrystallization led by amylose, and raises the viscosity	Longer term recrystallisation or "retrogradation" of both amylose and amylopectin. This stucture is different from the native state.

Figure 3.2 : Illustration of the starch morphology in the native state (A-E), and the mechanistic changes to alternate physical arrangements(F). It is important to note

that the chain lengths do not change, it is only the state of ordering that changes. The hexagon in A and B represents the glucose monomer. **A**. Amylose is the starch fraction where the monomers are connected by α1-4 glycosidic linkage in a linear chain. Due to the intramolecular forces and the linkages, the linear chain takes on a single helix configuration. **B.** Amylopectin is the starch fraction where the monomers are connected through α1-4 linear and α1-6 branching linkages. The branched chains twist in a double helix. **C.** Amylopectin is in the crystalline region and amylose in the amorphous region. Sometimes the long chains of amylose form a double helix.. The crystalline and amorphous bands alternate and form a granule as seen in **D.** The core is amorphous. **E.**Many granules pack into the grain of rice. **F.** Mechanistic changes leading to loss of crystalline structure and transforming into different arrangements. Note that these are not all measured based on a single sample through time.

3.2.3 Mechanisms driving various physical rearrangements, and properties

Food processing operations, including temperature changes (both heating and cooling) and the addition of water disrupt the native structure of starch matrix The temporal sequence of the various physical arrangements[3] starting from the native state is illustrated in **Figure 3.2**. This temporal behavior varies with the composition of the starch matrix. Longer AM or AP chains form more stable helices, requiring greater energy for bond disruption, ultimately leading to higher gelatinization temperatures[8,18,19]. Mechanistically, greater helical entanglement also has the effect of reduced tendency for swelling[20]. When the native starch granule is exposed to water and increasing

[3] Note that the fundamental structure, i.e. the branching points and chain lengths is unchanged.

temperatures, the loss of helical bonding,allows water to enter the granule, leading to further weakening[21]. Upon reaching gelatinization temperature, the bonds have weakened to the extent that the concentric ring structure is no longer present**.** There are various explanations for the extent of dissociation of the crystalline and amorphous regions that lead to gelatinization; we refer the reader to a recent review by Tetlow and Bertoft[22] . After gelatinization, AM and AP polymers disperse and interact with water. This interaction leads to different values of viscosity depending on the applied temperatures, and this information is important in temperature settings for food processes. The relevant range of viscosities is studied by a standardized analytical test (rapid visco analyser)[4,9,10,13–15] that measures the viscosity at 4 second intervals for a 12 minute period, where the sample solution is treated to a temperature profile of increasing temperature up to a maximum which is held constant and then decreased. Peak viscosity is the viscous effect at the peak temperature, where amylose and amylopectin are fully dispersed in water and maximally disordered. Higher amylose content typically corresponds to lower peak viscosity, since the amylose lipid complex resists swelling and amylose tends to form linear chains which slide over each other in solution[23]. When the temperature is held constant, the viscosity falls as amylose and amylopectin chains are aligned while still interacting with water, and this is the trough viscosity. The starch polymers are fully dispersed at this point[24]. Then the temperature is reduced leading to an early stage of recrystallization – or “retrogradation” , and the resulting crystalline pattern is different from that of the native crystal structure[20]. This starts with the shorter chains of amylose[25] in the early stages and followed by amylopectin. This mechanism is observed as an increase in the viscosity up to the final

viscosity, which is proportional to the amylose content. Long chains of amylose take longer to retrograde, and are associated with greater hardness of the cooked rice[8,20]. Stickiness is inversely related to amylose content and positively related to amylopectin that have a delayed retrogradation[18,26,27]. Retrogradation effects are studied for their detrimental effect on sensory properties - for example the hardening of baked goods[28]. However, there are gaps and limitations in prior research in discovering the relationships of structure to the properties, as examined in recent review papers. Hamaker explains that the cumulative structural knowledge of starch lacks details of the internal architecture of amylopectin which limits structure-property applications in food formulation[29]. Tetlow and Bertoft [22] provide an elaborate description of the starch granule in sections 2-6 of their paper as a basis for their critique and proposed solution. They point out the inadequacy of frequently reported features in prior research such as short, medium and long branches of amylopectin, and amylose content by citing contradicting observations or unexplained differences in phenotypes. The authors then reference their prior research[30] where Bertoft proposed the backbone branching structure of amylopectin (also called interblock chain lengths) as an important morphological feature which addressed their critique. The inadequacy of morphological features was also investigated by Tao et.al[31] , and they reported the influence of the molecular size of amylopectin on the viscosity properties.

In our opinion, another gap in research is relating high resolution structural features to both the granule morphology and the properties. This would not only connect all the concepts across scales of resolution, but also provide greater precision to structure-

property models. The latter objective has also been suggested by Yu et. al [32]. We explore solutions for this gap in **Discussion Section 3.5.2**.

3.2.4 Relation of domain knowledge to experimental hypotheses

The main features of the starch morphology described in the previous sections are the chain lengths, the branching structure at the nanometer scale, and crystallinity at the scale of the granule morphology. These are reflected in the dataset (details in **Section 3.3.1**) as; (i) the chain length features are provided by the SEC data, (ii) the amorphous fraction of the crystal addressed by the amylose content, and (iii) the branching structure and crystallinity are represented by observational data of gel consistency and gelatinization temperature.

Our central hypothesis is based on the fact that the structural components at the nanometer scale remain constant during processing but acquire different physical arrangements each connected with a specific property. However, since the dataset does not contain branching structure information at the nanometer scale, we use the data assumptions mentioned earlier in this section and address the hypothesis in two parts (**RQ1 and RQ2 in Section 3.1**). First, we test whether the chain length distribution and content data is a versatile predictor. Then to test whether data on branching and crystallinity aspects improves prediction, we use the empirical measurements of gel consistency and gelatinization temperature. We compare the predictive performance based on chain length and content data alone and with the addition of gel consistency and gelatinization temperature. Further details of model design are in the **Methods Section 3.3.2 and 3.3.3**.

3.2.5 Case study and related work

The authors Buenafe et. al. from the International Rice Research Institute, show the promise of predictive methods for understanding how starch structural data and physical and sensory properties are related[4]. The food samples in their dataset are varieties of rice. The data per sample includes the chain length distribution (measured by size exclusion chromatography) of starch, the physical properties relevant to the cooking of rice, and the sensory properties of cooked rice. Our case study uses the same data, and **Section 3.4.1** has the detailed description. The approach we take builds on this prior work in several important ways. First, the method proposed by Beunafe et. al. predicts the cluster index for an unseen sample to indicate similarity to the previously identified clusters of rice samples but does not directly predict physical or sensory properties of the unseen sample. We address this limitation by reframing the predictive task to predict the values of the physical and sensory properties directly. Another limitation is that Buenafe et. al. do not use the high-resolution chain length distribution data for the clustering method, and instead use a reduced version by summing into 5 groups. We speculate that this is due to the challenge of high dimensionality presented by the 500 SEC features. We include this complete data in our predictive models and show several methods that can be used to identify and overcome the dimensionality issues. In addition, we go a step further with our modeling approach that allows us to interpret feature importance values in relation to prior knowledge about the mechanistic behavior of starch polymers.

3.3. Data and Methods

3.3.1 Data

Data samples and features We used the dataset from the study (referenced in **Section 3.2.5**) by Beunafe et al. at the International Rice Research Institute (IRRI)[4] which includes three types of data: composition, physical, and sensory, as shown in **Figure 3.1**. The composition data includes amylose content (AC) and protein content (PC). The structural data was measured by size exclusion chromatography (SEC) and reported for 500 degrees of polymerization (DP) values in the range from 5 to 12,000. The physical properties consist of gelatinization temperature (GT) measured by DSC, gel consistency (GC) measured as the length of the starch gel prepared in a tube after heating (followed by a hour of cooling), and time series viscosity data using a rapid visco analyser (measured every four seconds for a 12-minute period). The sensory data included 13 mouthfeel descriptors and was collected for 100 samples of rice. For this case study, we focus on hardness and stickiness, since these are the most widely studied sensory characteristics across literature and fundamental to other properties like cohesiveness and toothsomeness[26,27].

Data selection and qualitative observations The original study data contains 301 samples of rice, of which 100 had sensory data. We discarded all samples with missing data, resulting in a dataset of 231 samples for the models predicting viscosity, gelatinization temperature, and gel consistency. Of these, only 72 samples had sensory data and were used for the models predicting sensory characteristics. The data visualization is shown in **Figure 3.3**. Viscosity time series data shows that the samples vary in the gradient and overall profile. There are peak viscosities at roughly two distinct

times; the first is just after 200 seconds with a peak value of 2500 centipoise and the second is between 300-400 seconds with a peak value ranging from 2000-200 centipoise. The trough viscosity has a wide range, though it all occurs at 500 s, after which it increases. Similarly, there is also a wide range in final viscosities, and it appears that the viscosity gradient from the trough viscosity varies greatly. This suggests that some samples have a faster rate of retrogradation. The histogram of the amylose content (AC) histogram shows that there are a few samples with almost no amylose, otherwise the AC variation is small for the majority of samples. The histogram of protein content (PC) has a wide range, and usually for starchy grains, protein content above 10% is considered high. Gel consistency and gelatinization temperature have a wide range of distribution which does not have a uniform shape. Hardness and stickiness also have a wide range implying the variation in rice samples perceived by consumers. The subset used in this study are provided in the Supplementary Information File, and we refer the reader to the cited publication by Beunafe et al. [4] for the complete dataset.

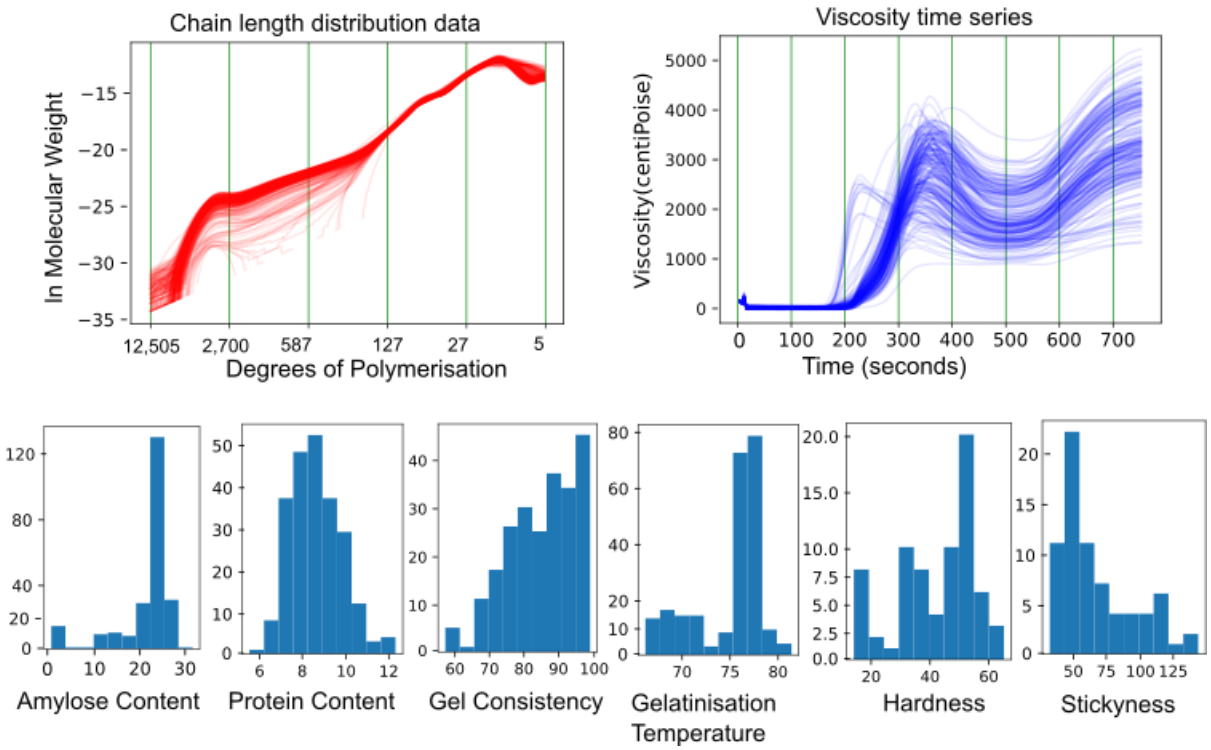


Figure 3.3: Illustrating the data for the 231 samples of rice. There are two types of data. The distribution is the chain length distribution data and viscosity data which is shown by line profiles per sample. The rest of the data is discrete, i.e. a single value per sample and shown by histograms.

3.3.2 Predictive Models

We designed a total of nine models, as shown in **Figure 3.1**, to address the research questions in **Section 3.1.** Models 1 and 2 predict gelatinization temperature and gel consistency from the composition and structure data. Models 3-5 predict peak, trough, and final viscosities from the composition and structure data, with some variants also using gelatinization temperature and gel consistency.

Models 6 and 7 predict sensory properties of stickiness and hardness from the composition and structure data. Models 8 and 9 predict stickiness and hardness from

the physical properties – viscosity time series, gelatinization temperature, and gel consistency. Results from Models 8 and 9 are compared to Models 6 and 7 to test whether structure and composition data is a better predictor of sensory properties.

3.3.3 Feature engineering and feature set selection

We used two different strategies in setting up the predictive feature set for a given prediction target. The first was to address a likely problem of high dimensionality of the distribution-type data. This applies to the chain length data with 500 DP features, and the viscosity time series data with 188 measurements over a 12 min period. We compared two techniques for dimensionality reduction feature aggregation for the chain length data, and interval sampling for the viscosity data. Ten variations of chain length data were generated – one at full resolution and an additional nine variants aggregated at varying bin sizes. Twenty-five variants of the viscosity data were generated: as is, and also as a time derivative to smooth out the curve for four time intervals (3, 5, 7, 10 seconds), combined with sampling at five different sampling rates - i.e every nth value (1, 3, 6, 8, 10). The viscosity time series data for the time interval variants is the derivative value instead of the original measured viscosity value, and differs for each of the four time intervals. At a certain interval, the value (at each of the 188 points) is the gradient(dv/dt) for that interval.

For each predictive target we tested various predictive feature sets, and the variants of the chain length data and viscosity data were a part of these feature sets. The feature sets with physical property features include GC and GT in four different variants: both GC and GT, either GC or GT, or neither. The feature sets with composition features

include AC (amylose content) and PC (protein content) in four different variants: both AC and PC , either AC or PC, or none. For the viscosity prediction models (Models 3-5) where we use both physical properties and compositional features, we use only these five combinations: AC, PC, GC, GT, or AC, PC, GC or AC, PC, GT or AC, PC, or none. This is designed with the intention to test for the inclusion of GC and GT as predictive features as described in RQ2 **Section 3.1**. Putting all this together, the feature set variants for the nine predictive models are listed below, and illustrated in **Figure 3.1**.

1. Gel consistency (model 1) gel temperature (model 2), stickiness (model 6) and hardness (model 7). Total of 40 variants. Ten variants are for the chain length data by bins of size 1, 5, 10, 15, 20, 25, 30, 35, 40, and 50, and four variants for the content data: AC and PC, only AC, only PC, and neither.
2. Peak, trough, and final viscosity (models 3, 4, 5). Total of 50 variants. Ten variants are for the chain length data by bins of size 1, 5, 10, 15, 20, 25, 30, 35, 40, and 50. 5 variants for the other features are AC, PC,GC and GT;AC, PC and GC;AC,PC and GT;AC and PC, and none.
3. Stickiness (model 8) and hardness (model 9): 100 variants. Twenty-five variants are for the viscosity time series, four variants for the content data: AC and PC, only AC, only PC, and neither.

3.3.4 Model Training

The approach to model building had two main steps: hyperparameter tuning and feature selection.

Hyperparameter Tuning: Random Forest models were trained for all the variants in each predictive task using the Random Forest regressor from Scikit[33]. We performed hyperparameter tuning to identify an appropriate hyperparameter configuration for each model, through a grid search cross validation method for a range of hyperparameters (range specified in the **Supplementary File)**. Hyperparameter configurations were selected on the basis of mean absolute error (MAE). The grid search results for all variants are provided in the **Supplementary File**. The MAE values are converted to a mean absolute error percentage (MAPE) by dividing the MAE by the average value. MAPE is scale independent and is suitable for comparison of the predictive performances.

Model Selection: A subset of all trained models (with hyperparameter configuration) per prediction target were selected for the next feature selection method, since it is computationally intensive. We explain the selection criteria through the example of the peak viscosity prediction; the results of the 50 trained models (**details in Section 3.4.3**) were grouped by the five discrete-feature combinations (AC, PC, GC and GT; AC, PC and GC; AC, PC and GT; AC and PC, and none). In each group the model configuration[4] with the lowest MAE value was selected. We also performed feature selection for model configurations using the SEC data with bin size 5. These models did not have the lowest MAE; we include them to compare model performance for high- and low-resolution data (see Section 3.3.4.2).

[4] "model configuration" refers to a set of predictive features along with a value for each hyperparameter.

Feature Selection: We used the Sequential Feature Selection method (SFS) in backwards mode, from the ML Extend library[34]. The feature selection method associates every feature (of the predictive feature set) with its importance in lowering the prediction error. Specifically, this method starts by considering models trained on all *n* features, then all possible combinations of *n-1* features, and so on until only a single feature is considered. The importance of each feature is determined by its rank. Features removed at the nth step are considered rank 1, features removed at the (n-1)-th step are considered rank 2, and so on. . **Figure 5** shows results from this method – removing features incrementally improves the performance until a peak, after which removing additional features degrades performance. It is important to note that the feature rank does not indicate a positive or negative correlation between the feature and the prediction target. We run ten replicates of this method to address any randomness associated with the results and focus on the consistent features across the replicates when we report the results (see **Figure 6**).

Best Model Configuration: The best model configuration for each predictive task is the model configuration resulting in the lowest MAE for that predictive task. The performance results for the best model configuration for each predictive task are presented in **Table 3.1.** The optimal sets of features for each task are presented in **Table 3.2.** Optimal hyperparameter configurations are in the **Supplementary File**. The naive ML baseline for comparison was the average value of the ground truth values for each target variable.

3.4. Results

3.4.1 Structure and composition data are versatile and outperform physical features in predicting sensory attributes.

The performance metrics for the best models 1-9 are in **Table 3.1** (see **Section 3.3.4** for best model selection). These metrics present two key inferences; first, the structure and composition features are better predictors of sensory properties relative to physical property features, and second, the physical properties can themselves be predicted from structure and composition. In combination, these results indicate that structure and composition features are versatile in their predictive capability. Furthermore, all predictive models have an average error 42.43% lower than the baseline models. Predictions of sensory properties (models 6-9 for stickiness and hardness) from structure and composition features resulted in 27% lower MAPE in comparison to prediction based on physical property features. More specifically, MAPE for prediction from structure and composition is 15.84% for stickiness and 11.31% for hardness. The MAPE for prediction from physical property features is 20.20% for stickiness and 16.62% for hardness. Both approaches outperform the naive baseline, which achieves 35.09% MAPE for stickiness and 25.95% for hardness. Physical properties can themselves be predicted from structure and composition data. Complete details of the performance metrics for the predictive and baseline models 1-5 are in **Table 3.1A**, and here we summarize the relative MAE improvements of the prediction models compared to the baseline models; 10.73% for gel consistency, 40.84% for peak viscosity, 39.89% for trough viscosity, and 45.34% for final viscosity. Relative prediction error for gelatinization temperature was better, but only by a small margin in absolute terms; the

prediction model MAE was 1.3, but the baseline model error was already good at 2.93 against an average value of 75.26. Complete coverage of all tested variants is in the Supplementary File.

Model number, predictive target	ML model		Baseline (Average)		Error Reduction (ML vs Baseline)	
	MAE	MAPE (%)	MAE	MAPE (%)	MAE	MAPE
1.GT	1.30	1.73	2.93	3.9	55.63	55.64
2.GC	7.70	9.27	8.63	10.38	10.73	10.69
3.PV	297.88	10.54	503.49	18.04	40.84	41.57
4. TV	249.78	13.93	409.36	22.82	38.98	38.96
5. FV	338.92	10.25	620.03	18.74	45.34	45.30
6. SBG	10.77	15.84	23.86	35.09	54.86	54.86
7. HRD	4.87	11.31	11.16	25.95	56.36	56.42

Table 3.1A: Comparing results from ML model (predicted from the structure and composition data) and the baseline model (which computes the average value). MAE is mean absolute error, and MAPE is mean absolute percent error.

Model number, target	ML model		Baseline (Average)		Error Reduction (ML vs Baseline)	
	MAE	MAPE (%)	MAE	MAPE (%)	MAE	MAPE
8. SBG	13.74	20.20	23.86	35.09	42.41	42.43
9. HRD	7.15	16.62	11.16	25.95	35.93	35.95

Table 3.1B. Comparing results from ML model (predicted from viscosity time series, gel consistency and gelatinization temperature) and the baseline model (which computes the average value). MAE is mean absolute error, and MAPE is mean absolute percent error.

3.4.2 High-resolution models outperform low-resolution models in predicting physical properties.

In identifying the likely challenges to our approach, we had considered that the high dimensionality of the structural data – which spans more than 500 DPs– combined with low sample size may impede predictive performance. We explored and compared two solutions for dimensionality reduction: low-resolution features via binning and selecting high-resolution features using an iterative feature selection procedure.

For the low-resolution feature approach, we used a binning method to reduce the size of the chain length data (**see Section 3.3.3**). Then, we trained several models with these different versions of the chain length features to identify the bin size resulting in the lowest predictive error. **Figure 3.4** shows a lower sensitivity to bin size for viscosity predictions, and greater sensitivity for gelatinization temperature, gel consistency, and sensory properties prediction. Bin size 5 consistently results in lower average error than the full resolution, for all prediction targets except gelatinization temperature. Another consistent trend is that the error decreases for lower resolutions, with the least error in bin sizes from 20 - 30. The error then increases for the lowest resolution at bin size 50.

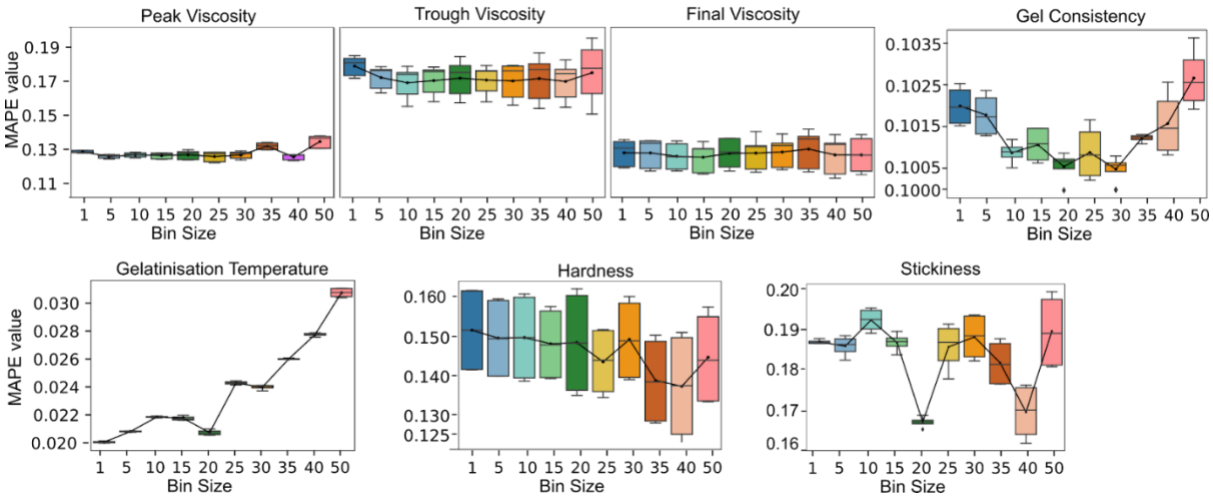


Figure 3.4: Predictive performance (MAPE) of models trained on binned CLD features. Box plot illustrated the results grouped by the bin size of the chain length distribution data. Bin sizes are 1, 5, 10, 15, 20, 25, 30, 35, 40, 50. Full resolution is 500 DP values. Each box plot represents the performance of all the models trained on different discrete features for the same feature aggregation (bin size for SEC). Within each box plot, the performance range is therefore due to the different discrete features sets. The averages are connected to illustrate the trend which is described in result 4.2.

The high-resolution approach is based on the feature selection method (**see Section 3.3.4**). We run this for both the models based on the high-resolution features at a bin size of 5, and the model based on the bin size identified in the low-resolution approach[5]. We then compare the performances for the full feature set and with the optimal feature set for both resolutions in **Figure 3.5**. Feature selection improves performance for both the high and low-resolution feature sets Gains from feature selection are greater for

[5] Though the bin size of 5 was also considered in the low-resolution approach, it was not the optimal bin size for any prediction target.

high-resolution features for the physical property predictive models (peak viscosity, trough viscosity, final viscosity, gelatinization temperature and gel consistency). The specific features in the optimal feature set for the high-resolution version are listed in **Table 3.2**.

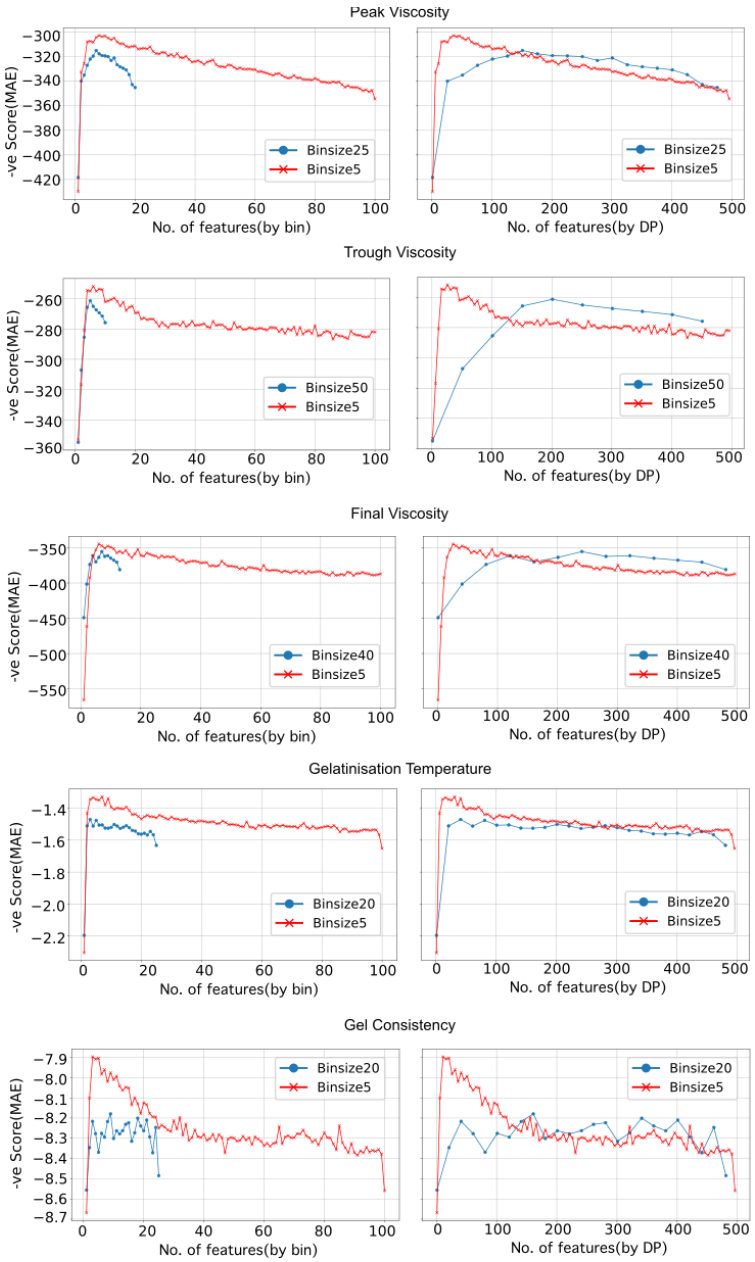


Figure 3.5: Predictive performance (mean absolute error) from the feature selection method. Overall these figures show gain in performance from the optimal features and is described in result 4.2 . Each plot shows and compares the performance for two different resolutions of the chain length data features. The MAE numbers are reversed for convenience of visualization such that the high peak in the curve is the best performance. The legend indicates the bin size used for aggregating the chain length data. The same result is shown through two perspectives on the left and right. The y axis in both views is MAE, and the perspectives differ in x axis only. Essentially the view on the right is a “stretched out” version. **Left :** The X axis is the binned feature. The high resolution models have more features, than the low resolution model. Note that this plot does not have any information about the specific degrees of polymerization for each feature. Also, the aggregated feature b1 for the high resolution is not the same as the aggregated feature b1 for the low resolution. **Right:** The X axis is the DP coverage. The low-resolution plot has fewer bins, and every bin has a much greater feature coverage. Each plot point is positioned at the start of the bin, and the last point covers the last bin even though the trend line ends short of the DP 500 on the x-axis. Note that the predictive features do include the other features (AC, PC, GC, GT), but they are not shown in this plot. For example, if feature 1 was a binned feature, feature 2 was AC, and feature 3 was a binned feature, we only show feature 1 and feature 2. So, although certain features are omitted, the MAE values are preserved by feature rank.

Model number, predictive target	**Predictive Features**		
	Amylopectin Dps	**Amylose Dps**	**Other features**
Gelatinization Temperature	7,12, 13, 19, 20, 81-86	6300-6700	

Peak Viscosity	47-50, 64-80	175-200, 300-320, 5000-6700	GC
Trough Viscosity	32-43	162-200	AC,GC,GT
Final Viscosity	35-44, 70-74	167 -200, 5100-5800	AC,GC,GT
Gel Consistency	28-35, 48-54	150 - 186	

Table 3.2. Predictive features (common to 10 replicates) per predictive target property

3.4.3 Predictive features occupy concentrated regions of the DP space across resolutions.

Although we identified that the high-resolution features have better performance in result 4.2, we explored the feature ranks at each resolution to understand the reason for the different performances by resolution. The ranks of the optimum features are presented as a heatmap in **Figure 3.6**. The rank information is given for 10 replicates of the feature selection method for each model at a specific resolution. We indicate the general alignment of features across the resolutions by vertical lines. These aligned features also match in their ranks as indicated by the color codes. As we further analyze this alignment, it is essential to note that each low-resolution feature aggregates more DPs than the high-resolution features. We observe that for any low-resolution feature only a few of the corresponding high-resolution features have a positive rank. This result suggests that the predictive capability of an aggregated feature is influenced by its constituents - what feature ranks does it aggregate. This observation means that not all the aggregated features are powerful simply for being large. The aggregated features that combine one or more predictive high-resolution features are themselves predictive. Based on these observations, although there is a general consistency in the optimum features across resolutions, the low-resolution features are weaker in their predictive capability compared to the high-resolution features.

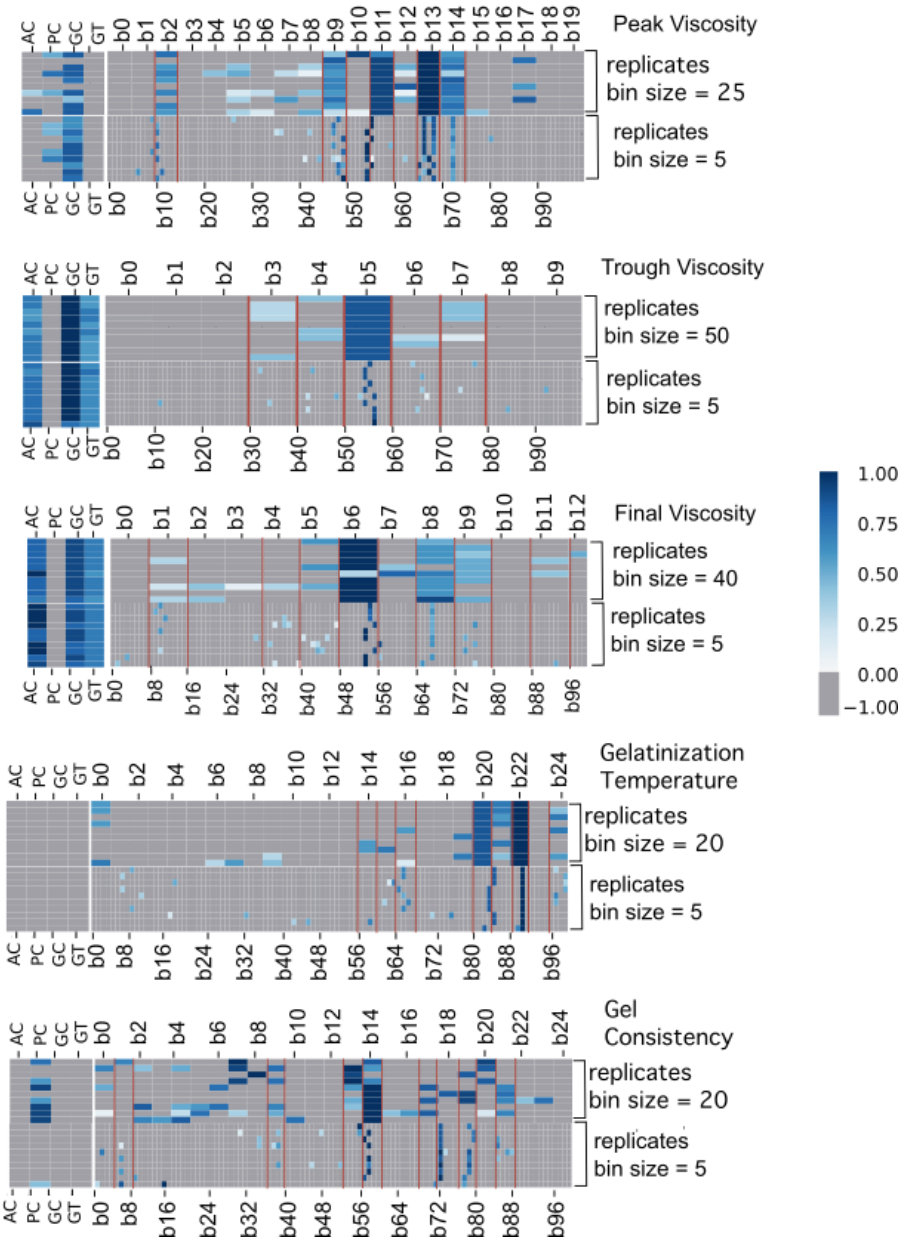


Figure 3.6: Feature rank results for physical property prediction models. The heat maps compare optimal feature sets for different resolutions of the chain length data, per predicted target. There are 10 replicates per model. For each set of replicates, the bin sizes are indicated to the right. The features by bin number are marked corresponding to the map grid. The actual DP ranges corresponding to the high-resolution bin features are marked at the bottom of the grid. The optimum features are assigned a blue color-

code by rank, and all negative ranking features are assigned a uniform gray color. The vertical lines running through the grid indicates the alignment between features across the resolutions.

3.4.4 Crystalline structure information improves predictions of peak, trough and final viscosity.

We verify the hypothesis in **Section 3.2.4**, that additional features on the crystallinity aspects of structure would improve prediction predictive performance. As we are limited by our dataset, we use gel consistency and gelatinization temperature as a proxy for the information on the branching and degree of crystallinity. The predictive performance with these additional features improves the peak viscosity prediction by 2.7%, trough viscosity prediction by 16.83% , and the final viscosity by 10.82%, as shown in **Table 3.3**. We discuss this topic further in **Section 3.5.2**.

Predictive target	MAE (excl. GC,GT)	MAE (incl. GC,GT)	Relative improvement (incl. GC,GT)
Peak Viscosity	305.25	297.01	2.70%
Trough Viscosity	300.35	249.78	16.84%
Final Viscosity	380.06	338.92	10.82%

Table 3.3. Model performance (mean absolute error) for models trained on predictive feature sets that exclude or include gel consistency and gelatinization temperature. Model performance is better (lower error) for models trained with GC and GT.

3.5. Discussion

The results successfully demonstrate that structure and composition data is a versatile predictor of a range of physical and sensory properties. Further, to emphasize the potential of structure-property models, we find that structure and composition data is a better predictor (27% lower error) of sensory mouthfeel than physical property data. This superior performance result challenges the prevalent approach in experimental research that leverages the measurement of pasting dynamics to predict textures properties[13–1535]. We also found that models trained on selected high resolution chain length features led to superior predictive performance and we identified the most predictive chain lengths for each physical property.

3.5.1 Relating chain length features to the mechanistic changes and material properties.

Prior research (addressed in **Section 3.2.3)** has studied details of starch morphology and relationships to the mechanisms and properties with the common objective of making starch behavior predictable, but there are gaps in the knowledge to date [all citations from section 2.3]. Our suggested solution to this gap is to relate high resolution structural features to the morphology and properties. We exemplify our solution by relating the feature importances (**Section 3.3.3**) to the prior research findings for the gelatinization temperature. The optimal predictive features for gelatinization temperature are amylopectin DPs 7, 12, 13, 19, 20, 81- 86, with higher ranks and amylose DPs 6300 -- 6700 has the lowest rank . Gelatinization temperature is associated with the loss of crystallinity and granule structure, and results from exposure of the granule to heat and moisture for a period of time. The mechanisms responsible are the simultaneous

disruption of helical bonds from the heat, penetration of water into the center of the granule and leaching out of amylose. Tao et. al[31] have shown the correlation of amylopectin DP 6-60 and short amylose branches (DP 270-354) [6] to the gelatinization temperature. The amylopectin features agree with the research by Tikapunya et. al which shows the correlation of DP 6-33 to the gelatinization temperature[36]. However, the findings by Li et al. [25] on amylose DPs contradicts with Tao et.al, since it shows leaching of long amylose chains though it was unclear whether this observation was at the point of gelatinization or later. In relating these results to our feature results, we find strong evidence for amylopectin from two sources but contradictory evidence for amylose DPs. In general, it is difficult to interpret our results considering current mechanistic knowledge which lacks agreement on relationships between DP features and properties.

3.5.2 Insights to guide future experiments in data generation and modeling.

The prior topic in discussion addressed that the branching structure influences the mechanisms and the expressed property, and that this morphological feature has been under researched comparatively. The need for data on the degree of branching and crystallinity as predictive features was also hypothesized (**Section 3.2.4) and** raised as a research question (**RQ2 in Section 3.1)**. The results in **Section 3.4.4** confirmed that inclusion of gel consistency and gelatinization temperature, as a proxy for analytical data, improved predictive performance. At the same time, gel consistency and

[6] The DP numbers for short and long branches of amylose have been suggested as DP 100 - 700 and 700-40,000 in the proportion of 90% and 10% . In the case of rice, this range has been specified as 270-354 and 1550-1965 by Wang et al[28].

gelatinization temperature hough predictable from the structure and composition data, do not have 100% accuracy. This hypothesis can be more precisely validated by including glycosidic linkage data on the abundances of α1-4 (linear connections) and α1-6 (branching points) bonds [37].

3.6. Conclusion

The results from our case study on structure-functional predictive capability are unprecedented, despite the enormous prior existing research on starch including many journals specific to starch. This case study addresses the questions that are studied extensively in research literature; 1. Is there a relationship between the starch structure/composition and the physical and sensory properties and 2. Is RVA and other empirical data indicative of grain quality and/or product attributes. These questions are significant to both breeding crops for starch structure and food formulating for desirable traits in the products. However, there exist no standard models and only a few prior experimental studies have only established statistical correlations or simple linear equations, and on experimental datasets of a few samples[38–40].

In response to the first question, the physical properties of peak, trough and final viscosity, gel consistency, gelatinization temperature and sensory properties of hardness and stickiness of rice are predicted from high resolution data on the composition and structure. To address the second question stated above, hardness and stickiness are also predicted from the physical properties which had a 27% lower predictive accuracy . Further, we obtain the specific chain lengths of starch polymers that are most predictive of a physical property and compare these with prior domain

knowledge assembled over decades of observational studies and discover that there is a lack of agreement on relationships in literature.

Ultimately this study shows that the ability of machine learning methods to learn complex multivariate relationships is certainly applicable in the complex domain of food science and food formulations. Such data-driven analysis can reveal insights that are nearly impossible to discover in hypothesis-driven experiments. This manuscript is a call to creation of standardized datasets of biopolymers in food, for enabling breakthrough innovation in food. We believe that an integrated approach of big-data and experiments, can respond to changing consumer demands and sustainability needs with greater agility and efficiency in diverse applications like plant breeding and innovative food formulation.

REFERENCES

(1) Kaushik, S. J.; Panserat, S.; Schrama, J. W. Chapter 7 - Carbohydrates. In *Fish Nutrition (Fourth Edition)*; Hardy, R. W., Kaushik, S. J., Eds.; Academic Press, 2022; pp 555–591. https://doi.org/10.1016/B978-0-12-819587-1.00008-2.

(2) Wahlqvist, M. L. The Place of Carbohydrates in Newer Food Formulations: Opportunities for Nutritional Advancement and Their Safety. *Asia Pac. J. Clin. Nutr.* **2002**, *11*, S149–S154. https://doi.org/10.1046/j.1440-6047.11.s.6.4.x.

(3) Haslam, S. M.; Freedberg, D. I.; Mulloy, B.; Dell, A.; Stanley, P.; Prestegard, J. H. Structural Analysis of Glycans. In *Essentials of Glycobiology*; Varki, A., Cummings, R. D., Esko, J. D., Stanley, P., Hart, G. W., Aebi, M., Mohnen, D., Kinoshita, T., Packer, N. H., Prestegard, J. H., Schnaar, R. L., Seeberger, P. H., Eds.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor (NY), 2022.

(4) Buenafe, R. J. Q.; Kumanduri, V.; Sreenivasulu, N. Dataset on Viscosity and Starch Polymer Properties to Predict Texture through Modeling. *Data Brief* **2021**, *36*, 107038. https://doi.org/10.1016/j.dib.2021.107038.

(5) Thomas, D. J.; Atwell, W. A. CHAPTER 1: Starch Structure. In *Starches*; American Association of Cereal Chemists: 3340 Pilot Knob Road, St. Paul, Minnesota 55121-2097, USA, 1999; pp 1–11. https://doi.org/10.1094/1891127012.001.
(6) Chandrasekara, A.; Josheph Kumar, T. Roots and Tuber Crops as Functional Foods: A Review on Phytochemical Constituents and Their Potential Health Benefits. *Int. J. Food Sci.* **2016**, *2016*, 3631647. https://doi.org/10.1155/2016/3631647.
(7) Shi, S.; Pan, K.; Yu, M.; Li, L.; Tang, J.; Cheng, B.; Liu, J.; Cao, C.; Jiang, Y. Differences in Starch Multi-Layer Structure, Pasting, and Rice Eating Quality between Fresh Rice and 7 Years Stored Rice. *Curr. Res. Food Sci. Online* **2022**, *5*, 1379–1385. https://doi.org/10.1016/j.crfs.2022.08.013.
(8) Li, C.; Gong, B. Insights into Chain-Length Distributions of Amylopectin and Amylose Molecules on the Gelatinization Property of Rice Starches. *Int. J. Biol. Macromol.* **2020**, *155*, 721–729. https://doi.org/10.1016/j.ijbiomac.2020.04.006.
(9) Chung, W.; Han, A.; Saleh, M.; Meullenet, J. Prediction of Long-Grain Rice Texture from Pasting Properties. **2006**.
(10) Liu, C.; Liu, P.; Yan, S.; Qing, Z.; Shen, Q. Relationship of Physicochemical, Pasting Properties of Millet Starches and the Texture Properties of Cooked Millet. *J. Texture Stud.* **2011**, *42* (4), 247–253. https://doi.org/10.1111/j.1745-4603.2010.00271.x.
(11) Karwasra, B. L.; Gill, B. S.; Kaur, M. Rheological and Structural Properties of Starches from Different Indian Wheat Cultivars and Their Relationships. *Int. J. Food Prop.* **2017**, *20* (sup1), S1093–S1106. https://doi.org/10.1080/10942912.2017.1328439.
(12) Gaenssle, A. L. O.; Satyawan, C. A.; Xiang, G.; van der Maarel, M. J. E. C.; Jurak, E. Long Chains and Crystallinity Govern the Enzymatic Degradability of Gelatinized Starches from Conventional and New Sources. *Carbohydr. Polym.* **2021**, *260*, 117801. https://doi.org/10.1016/j.carbpol.2021.117801.
(13) Zhu, L.; Wu, G.; Zhang, H.; Wang, L.; Qian, H.; Qi, X. Using RVA-Full Pattern Fitting to Develop Rice Viscosity Fingerprints and Improve Type Classification. *J. Cereal Sci.* **2018**, *81*, 1–7. https://doi.org/10.1016/j.jcs.2018.02.013.
(14) Balet, S.; Guelpa, A.; Fox, G.; Manley, M. Rapid Visco Analyser (RVA) as a Tool for Measuring Starch-Related Physiochemical Properties in Cereals: A Review. *Food Anal. Methods* **2019**, *12* (10), 2344–2360. https://doi.org/10.1007/s12161-019-01581-w.
(15) Cozzolino, D. The Use of the Rapid Visco Analyser (RVA) in Breeding and Selection of Cereals. *J. Cereal Sci.* **2016**, *70*, 282–290. https://doi.org/10.1016/j.jcs.2016.07.003.

(16) Apriyanto, A.; Compart, J.; Fettke, J. A Review of Starch, a Unique Biopolymer - Structure, Metabolism and in Planta Modifications. *Plant Sci.* **2022**, *318*, 111223. https://doi.org/10.1016/j.plantsci.2022.111223.
(17) Wang, T. L.; Bogracheva, T. Y.; Hedley, C. L. Starch: As Simple as A, B, C? *J. Exp. Bot.* **1998**, *49* (320), 481–502. https://doi.org/10.1093/jxb/49.320.481.
(18) Yao, Y.; Zhang, J.; Ding, X. Structure-Retrogradation Relationship of Rice Starch in Purified Starches and Cooked Rice Grains: A Statistical Investigation. *J. Agric. Food Chem.* **2002**, *50* (25), 7420–7425. https://doi.org/10.1021/jf020643t.
(19) Li, H.; Wen, Y.; Wang, J.; Sun, B. Relations between Chain-Length Distribution, Molecular Size, and Amylose Content of Rice Starches. *Int. J. Biol. Macromol.* **2018**, *120* (Pt B), 2017–2025. https://doi.org/10.1016/j.ijbiomac.2018.09.204.
(20) Jane, J.; Chen, Y. Y.; Lee, L. F.; McPherson, A. E.; Wong, K. S.; Radosavljevic, M.; Kasemsuwan, T. Effects of Amylopectin Branch Chain Length and Amylose Content on the Gelatinization and Pasting Properties of Starch. *Cereal Chem. J.* **1999**, *76* (5), 629–637. https://doi.org/10.1094/CCHEM.1999.76.5.629.
(21) Ratnayake, W. S.; Jackson, D. S. Chapter 5 Starch Gelatinization. In *Advances in Food and Nutrition Research*; Academic Press, 2008; Vol. 55, pp 221–268. https://doi.org/10.1016/S1043-4526(08)00405-1.
(22) Tetlow, I. J.; Bertoft, E. A Review of Starch Biosynthesis in Relation to the Building Block-Backbone Model. *Int. J. Mol. Sci.* **2020**, *21* (19). https://doi.org/10.3390/ijms21197011.
(23) Kumar, R.; Khatkar, B. S. Thermal, Pasting and Morphological Properties of Starch Granules of Wheat (Triticum Aestivum L.) Varieties. *J. Food Sci. Technol.* **2017**, *54* (8), 2403–2410. https://doi.org/10.1007/s13197-017-2681-x.
(24) Schirmer, M.; Jekle, M.; Becker, T. Starch Gelatinization and Its Complexity for Analysis. *Starch - Stärke* **2015**, *67* (1–2), 30–41. https://doi.org/10.1002/star.201400071.
(25) Li, C.; Ji, Y.; Zhang, S.; Yang, X.; Gilbert, R. G.; Li, S.; Li, E. Amylose Inter-Chain Entanglement and Inter-Chain Overlap Impact Rice Quality. *Foods* **2022**, *11* (10), 1516. https://doi.org/10.3390/foods11101516.
(26) Li, H.; Fitzgerald, M. A.; Prakash, S.; Nicholson, T. M.; Gilbert, R. G. The Molecular Structural Features Controlling Stickiness in Cooked Rice, a Major Palatability Determinant. *Sci. Rep.* **2017**, *7*, 43713. https://doi.org/10.1038/srep43713.
(27) Li, C.; Luo, J.-X.; Zhang, C.-Q.; Yu, W.-W. Causal Relations among Starch Chain-Length Distributions, Short-Term Retrogradation and Cooked Rice Texture. *Food Hydrocoll.* **2020**, *108*, 106064. https://doi.org/10.1016/j.foodhyd.2020.106064.
(28) Wang, S.; Li, C.; Copeland, L.; Niu, Q.; Wang, S. Starch Retrogradation: A Comprehensive Review. *Compr. Rev. Food Sci. Food Saf.* **2015**, *14* (5), 568–585. https://doi.org/10.1111/1541-4337.12143.

(29) Hamaker, B. R. Current and Future Challenges in Starch Research. *Curr. Opin. Food Sci.* **2021**, *40*, 46–50. https://doi.org/10.1016/j.cofs.2021.01.003.
(30) Bertoft, E.; Annor, G. A.; Shen, X.; Rumpagaporn, P.; Seetharaman, K.; Hamaker, B. R. Small Differences in Amylopectin Fine Structure May Explain Large Functional Differences of Starch. *Carbohydr. Polym.* **2016**, *140*, 113–121. https://doi.org/10.1016/j.carbpol.2015.12.025.
(31) Tao, K.; Li, C.; Yu, W.; Gilbert, R. G.; Li, E. How Amylose Molecular Fine Structure of Rice Starch Affects Functional Properties. *Carbohydr. Polym.* **2019**, *204*, 24–31. https://doi.org/10.1016/j.carbpol.2018.09.078.
(32) Yu, W.; Li, H.; Zou, W.; Tao, K.; Zhu, J.; Gilbert, R. G. Using Starch Molecular Fine Structure to Understand Biosynthesis-Structure-Property Relations. *Trends Food Sci. Technol.* **2019**, *86*, 530–536. https://doi.org/10.1016/j.tifs.2018.08.003.
(33) Pedregosa, F.; Varoquaux, G.; Gramfort, A. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn.* **2011**, *12*.
(34) Raschka, S. MLxtend: Providing Machine Learning and Data Science Utilities and Extensions to Python's Scientific Computing Stack. *J. Open Source Softw.* **2018**, *3* (24), 638. https://doi.org/10.21105/joss.00638.
(35) Palabiyik, İ.; Toker, O. S.; Karaman, S.; Yildiz, Ö. A Modeling Approach in the Interpretation of Starch Pasting Properties. *J. Cereal Sci.* **2017**, *74*, 272–278. https://doi.org/10.1016/j.jcs.2017.02.008.
(36) Tikapunya, T.; Zou, W.; Yu, W.; Powell, P. O.; Fox, G. P.; Furtado, A.; Henry, R. J.; Gilbert, R. G. Molecular Structures and Properties of Starches of Australian Wild Rice. *Carbohydr. Polym.* **2017**, *172*, 213–222. https://doi.org/10.1016/j.carbpol.2017.05.046.
(37) Haslam, S. M.; Freedberg, D. I.; Mulloy, B.; Dell, A.; Stanley, P.; Prestegard, J. H. Structural Analysis of Glycans. In *Essentials of Glycobiology*; Varki, A., Cummings, R. D., Esko, J. D., Stanley, P., Hart, G. W., Aebi, M., Mohnen, D., Kinoshita, T., Packer, N. H., Prestegard, J. H., Schnaar, R. L., Seeberger, P. H., Eds.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor (NY), 2022. https://doi.org/10.1101/glycobiology.4e.50.
(38) Zhao, Y.; Henry, R. J.; Gilbert, R. G. Testing the Linearity Assumption for Starch Structure-Property Relationships in Rices. *Front. Nutr.* **2022**, *9*, 916751. https://doi.org/10.3389/fnut.2022.916751.
(39) Li, C.; Hu, Y.; Li, E. Effects of Amylose and Amylopectin Chain-Length Distribution on the Kinetics of Long-Term Rice Starch Retrogradation. *Food Hydrocoll.* **2021**, *111*, 106239. https://doi.org/10.1016/j.foodhyd.2020.106239.
(40) Zhong, Y.; Liu, L.; Qu, J.; Blennow, A.; Hansen, A. R.; Wu, Y.; Guo, D.; Liu, X. Amylose Content and Specific Fine Structures Affect Lamellar Structure and Digestibility of Maize Starches. *Food Hydrocoll.* **2020**, *108*, 105994. https://doi.org/10.1016/j.foodhyd.2020.105994.

CHAPTER 4: Conclusions and Future Trajectory

4.1. Thesis conclusions

The hypothesis underlying this thesis is: there is a relationship between the chemical composition of particular food and its nutritive and sensory properties in the processed state (for details see **Introduction Section 1.1.2**) This was tested through projects 1 and 2 for prediction targets related to nutrition and sensory properties. The specific research questions for each project and the corresponding results are summarized.

The models in project 1(**Chapter 2**) predicted the content of 7 vitamins and 7 minerals in a cooked food, for 425 plant and animal based and for 5 cooking methods. The research question for project 1 were: Is the predictive ML model more accurate than the baseline of the prevalent methods? Is there a difference in predictive performance by categories of food and processing methods? What features are the most predictive? The result to the first question on performance confirms that the ML models outperform prevalent methods with a 31% lower average error across the variety of foods and processing methods. In addition, the ML models have the potential to scale across diverse foods without compromising the accuracy (**Result 2.3.3**), which is a shortcoming of prevalent methods. The next result (**Result 2.3.4**) is based on a breakdown analysis of the predictive performance by food category. It is revealed that legumes have the best among plant-based foods and beef the best in the animal-based foods. This brings up the obvious question of what differentiates these food categories, and one possible research hypothesis is – How does the macromolecular structure (carbohydrates, lipids, proteins) differ, and does it affect the chemistry of processing? The final result in response to the

question is that there were not any common predictive features across all 14 content predictions and not for all foods (**Result 2.3.5**). This seems reasonable that the there is no common reaction pattern for 14 compounds during cooking processes. There were 2 additional results discovered through analysis of the training dataset. Result 3.1 explains the data distribution and why only 10% of the dataset was useable for this study. Result 3.2 exposes the limitation that the cooking yield was not recorded, which was addressed in the project through a variety of scaling methods. Although the scaling method mitigated the issue, it is recommended that cooking yield is a necessary variable in a dataset to model the nutritional impacts of food processing.

The models in project 2 (**Chapter 3**) predict the physical and sensory properties related to texture. As texture is a physical property, the modelling approach was to predict from the structural composition, therefore creating a structure-property model. The general focus was on the glycan composition of foods, and the specific experiment focused on modeling the structure of starch in rice. The research questions for project 2 were: Is the structure (chain length data) and composition data sufficient to predict both physical and sensory properties of cooked rice? Does gelatinization temperature and gel consistency information improve predictive performance over only the chain length and content data? What features (structural or otherwise) are the most informative for each prediction? In response to the first questions, the result (**Result 3.4.1)** confirms that the structural data (chain length distribution) of starch can generalize across several physical properties (peak, trough and final viscosity, gel consistency, gelatinization temperature) in the cooking process of rice, and sensory properties (hardness and stickiness) of cooked rice. In additional structural data is a better

predictor (27% higher predictive accuracy) of sensory properties than physical properties, even though the latter is typically used in experiment-based research. In response to the second questions, it is confirmed (**Result 3.4.4**) that gelatinization temperature and gel consistency improve the predictive performance. The interpretation in a structural context is these variables are a proxy for information on the branching and degree of crystallinity. In response to the third question, the results (**Result 3.4.2 and 3.4.3**) identify the specific chain lengths that are the most predictive for each prediction. The **Discussion 3.5.1** goes further and relates these chain length features to the processing mechanisms. But it is difficult to verify since current mechanistic knowledge lacks agreement on relationships between DP features and properties. Overall, both these projects affirms that both composition and structure are essential to understanding and predicting the outcomes of processing.

In conclusion, this research thesis confirms the hypothesized potential of machine learning methods to learn complex multivariate relationships, and the applicability to more complex problems in the domain of food science and food formulations. Such data-driven analysis also reveals insights that are nearly impossible to discover in hypothesis-driven experiments. Given the urgent challenges confronting human health, these takeaways are relevant to the shared goal across research and industry of food innovation for personalized human health while simultaneously sustaining a healthy planet[1–4]. This thesis is a call for the creation of standardized and FAIR datasets on food composition to enable such breakthrough innovation in food.

4.2. Food System vision of personalized health and taste

Practically the achievement of this goal implies that ML would impact everyday life analogous to a utility like GoogleMaps. The translation from research to utility at this scale, raises a few major questions: what does the most comprehensive dataset look like? who will build and maintain this? and how can these datasets and models be used easily throughout the food system?

What does a comprehensive dataset look like? It is widely acknowledged that our knowledge of food composition is a mere fraction compared to what remains to be discovered, which is aptly referred to as the "dark" matter of food composition[5]. While aware of this, Rockefeller's PTFI project leads the effort to create an unprecedented comprehensive composition dataset based on advanced analytics available today, and for the global diversity of foods. The PTFI team also acknowledge the challenges of such an ambitious project as addressed in Chapter 1 Section 1.2. It is my belief that the building of datasets would be most productive and cost efficient if the infra-structure also included a feedback loop from the data analysis and modeling to guide future data generation as exemplified by my thesis projects.

Who would build, scale and maintain? An impressive example is the scale up and implementation of AlphaFold for protein design within 4 years of the first release. AlphaFold started as CASP, a collaborative research effort in 1994, and achieved a breakthrough milestone by DeepMind in 2020 and acquired by Google. The dataset and code are both publicly available[6] used by several startups like Shiru and Cradle[7]. Private versions that rival AlphaFold with 200M protein structures are now available, for example by MetaAI[8] with 600M proteins. OpenFoldConsortium is a spinoff from AlphaFold

developed in collaboration by several biotech companies and hosted by Amazon Web Services[9,10]. A trajectory like this is not unlikely for carbohydrates (potential for structure-function as exemplified by project2), and the “AlphaFold” for carbohydrates could one day be a reality.

How can these datasets and models be used easily throughout the food system? The above initiatives related to AlphaFold suggest that organizations that grow and maintain datasets are also the users of it. It is also worth recognizing that such progress will undoubtedly revolutionize how research is conducted and shared, as well as create unimaginable collaborative structures.

REFERENCES

1. Desiere, F., German, B., Watzke, H., Pfeifer, A. & Saguy, S. Bioinformatics and data knowledge: the new frontiers for nutrition and foods. *Trends Food Sci. Technol.* **12**, 215–229 (2001).
2. Demartini, M. *et al.* Food industry digitalization: from challenges and trends to opportunities and solutions. *IFAC-Pap.* **51**, 1371–1378 (2018).
3. Ahmed, S. *et al.* Foodomics: A Data-Driven Approach to Revolutionize Nutrition and Sustainable Diets. *Front. Nutr.* **9**, 874312 (2022).
4. Artificial Intelligence Improves America’s Food System. https://www.usda.gov/media/blog/2020/12/10/artificial-intelligence-improves-americas-food-system.
5. The unmapped chemical complexity of our diet | Nature Food. https://www.nature.com/articles/s43016-019-0005-1.
6. DeepMind says it will release the structure of every protein known to science | MIT Technology Review. https://www.technologyreview.com/2021/07/22/1029973/deepmind-alphafold-protein-folding-biology-disease-drugs-proteome/.
7. Godhwani, G. The Next Wave: Artificial Intelligence In Alternative Proteins. https://www.goodsignal.com/p/the-next-wave-artificial-intelligence (2022).
8. Callaway, E. AlphaFold’s new rival? Meta AI predicts shape of 600 million proteins. *Nature* **611**, 211–212 (2022).

9. OpenFold Biotech AI Research Consortium releases SoloSeq and Multimer, an integrated protein Large Language Model with 3D structure generation. https://www.businesswire.com/news/home/20240219658831/en/OpenFold-Biotech-AI-Research-Consortium-releases-SoloSeq-and-Multimer-an-integrated-protein-Large-Language-Model-with-3D-structure-generation (2024).
10. OpenFold Consortium. https://openfold.io/.