



Kevin de Castro Cogle

**An Artificial Intelligence approach towards screening of
the food matrix and digestion effects in predicting
phytochemical bioaccessibility**

School of Water, Energy and Environment

PhD

Academic Year: 2021 - 2024

Supervisors: Maria Anastasiadi, Fady Mohareb, Claire Rossi

Associate Supervisor: Mirian Kubo

June 2024

This thesis is submitted in fulfilment of the requirements for the
degree of PhD

© Cranfield University 2024. All rights reserved. No part of this
publication may be reproduced without the written permission of the
copyright owner.



**UNIVERSITE DE TECHNOLOGIE DE COMPIEGNE
ECOLE DOCTORALE**

**Doctorat
Biotechnologie**

KEVIN DE CASTRO COGLE

**AN ARTIFICIAL INTELLIGENCE APPROACH TOWARDS
SCREENING OF THE FOOD MATRIX AND DIGESTION EFFECTS
IN PREDICTING PHYTOCHEMICAL BIOACCESSIBILITY**

***Thesis supervised by Dr. Mirian Kubo, Dr. Maria Anastasiadi, Prof. Fady
Mohareb and Dr. Claire Rossi***

2021 - 2024

Thèse présentée pour l'obtention du grade de Docteur de l'UTC

**© Université de technologie de Compiègne 2024. All rights reserved.
No part of this publication may be reproduced without the written
permission of the copyright owner.**

Academic integrity declaration

I declare that:

- the thesis submitted has been written by me alone.
- the thesis submitted has not been previously submitted to this university or any other.
- that all content, including primary and/or secondary data, is true to the best of my knowledge.
- that all quotations and references have been duly acknowledged according to the requirements of academic research.

I understand that to knowingly submit work in violation of the above statement will be considered by examiners as academic misconduct.

Abstract

Functional foods are recognised to confer additional health benefits other than basic nutrition. These desirable characteristics can be attributed to specific bioactive compounds, such as vitamins or polyphenols. While the topic of health claims motivates extensive research, there is a notable gap in addressing how the carrier functional food will impact the gastrointestinal journey of a substance of interest. This consideration is crucial, as suboptimal food designs can lead to a compound being degraded or excreted instead of absorbed in a usable form, in which case the post-consumption effects are negligible.

The underlying physicochemical mechanisms occurring during digestion are complex. Previous studies provide conceptual knowledge regarding what factors impact a compound's absorption, but there is a need for methodologies that translate this understanding into practical decision-support systems.

In this thesis, a systematic approach for the study of a compound's absorption as a function of its carrier food's properties is presented, based on systematic and empirical screening experiments. The subjects of study were fat-soluble compounds, for which it is established that the main limiting step in absorption is the release from the food and solubilisation in the digesta (bioaccessibility). In the case of vitamin E, high and invariable bioaccessibilities were observed across all formulations (~85%), but curcuminoids presented a wide variability of low and moderate bioaccessibilities (~12%~55%) which were investigated further. The results indicated that variations in macronutrient concentration, as well as other factors, presented strong associations to this variability (p -value < 0.01).

These insights were used to develop a proof-of-concept mathematical model capable of generalization. Its preliminary estimations closely resembled the observations (<10% relative error), setting high expectations for future studies.

Keywords:

curcuminoids, vitamin E, bioaccessibility, food matrix, dietary fibre, screening, regression, Bayesian model

Acknowledgements

First and foremost, I would like to show my appreciation to those people without whom this project would not have been possible: my supervisors Dr. Maria Anastasiadi, Dr. Mirian Kubo, Prof. Fady Mohareb, and Prof. Claire Rossi. Thank you for granting me the opportunity to find my own ground in research whilst providing much needed support and not complaining once about my 2 AM emails. I hope the seeds of our efforts planted here will sprout and grow for the rest of our careers. Likewise, I would also like to thank Alexandra Josse and Dr. Franck Merlier for their aid in experiments that were crucial to this work. Prof. Muriel Vayssade, Dr. Natalia Falagán, Prof. Bérangère Bihan-Avalle and Prof. Yolande Perrin I thank for their constructive pitch-ins about my academic career development.

I must acknowledge my gratitude to my colleagues at both Universities, Compiègne and Cranfield, for their companionship in this journey; sharing hardships and high points with like-minded people was really helpful at more stages than one. On another note, several highlighted events with my dear old friends and brother Marcos back in Spain provided a valuable background which helped remind me about everything I was striving for; I will forever be grateful.

My deepest thanks are for the three most important people in my life. My parents Ángel and Carolina; I hope this achievement makes you proud. Words are not enough to express what you mean to me, but I hope our sporadic and exciting visits to Paris and London were but a small reminder.

Last, my dear wife Vicky. Sacrificing almost three years apart was the hardest part of this experience, but it is now over. Thank you for standing strong for me after putting our life on pause. Let me assure you it will all be worth it.

Table of Contents

Academic integrity declaration.....	i
Abstract	ii
Acknowledgements	iv
List of Figures.....	viii
List of Tables.....	xii
List of Equations.....	xv
List of Abbreviations	xvii
1 Executive summary	19
2 Introduction and literature review	20
2.1 The food matrix	26
2.2 Functional foods	28
2.2.1 Categories and examples.....	28
2.2.2 Ingredients: phytochemicals and dietary fibres	31
2.2.3 Marketing	32
2.3 Human digestion	35
2.3.1 Oral cavity	36
2.3.2 Stomach.....	38
2.3.3 Small intestine.....	41
2.3.4 Large intestine.....	43
2.4 Absorption: the concepts of bioaccessibility and bioavailability	44
2.5 Simulation of human digestion: <i>in vitro</i> models.....	47
2.5.1 <i>In silico</i> models.....	48
2.5.2 Dynamic <i>in vitro</i> models	48
2.5.3 Static models.....	50
2.5.4 Semi-dynamic models.....	53
2.6 Bioaccessibility of fat-soluble compounds.....	55
2.6.1 Gastrointestinal journey of fat-soluble compounds.....	55
2.6.2 Food matrix effects to consider for improving bioaccessibility of fat-soluble compounds	58
2.6.3 Vitamin E.....	61
2.6.4 Curcuminoids	63
2.7 Machine Learning: regression models	67
2.7.1 Introduction to Machine Learning	67
2.7.2 Mathematical theory of regression	69
2.7.3 Bias-variance tradeoff	75
2.7.4 Sampling strategies in ML projects	77
2.7.5 Frequentist approaches.....	78
2.7.5.1 Support Vector Machines (SVM)	78
2.7.5.2 Decision trees and their ensembles.....	80
2.7.5.3 K-nearest neighbours (KNN).....	83

2.7.5.4 Regularization in linear regression.....	84
2.7.6 Bayesian approach	86
2.7.6.1 Bayes' theorem.....	87
2.7.6.2 Markov Chain Monte Carlo	90
2.7.6.3 Bayesian hierarchical linear regression and its MCMC inference.....	91
2.7.7 Bayesian vs frequentist methods	96
2.7.8 Feature selection in Machine Learning.....	98
2.7.9 Applications in the food industry.....	101
3 Methodology.....	104
3.1 Aims and objectives	104
3.2 Bioaccessibility quantification	104
3.2.1 Preparation of food matrices	104
3.2.1.1 Vitamin-E-enriched matrices.....	108
3.2.1.2 Curcuminoid-enriched matrices	109
3.2.2 Static <i>in vitro</i> digestion	112
3.2.2.1 Regarding gastric lipase	116
3.2.3 Bioaccessibility quantification.....	119
3.2.3.1 Vitamin E	119
3.2.3.2 Curcuminoids.....	120
3.2.3.2.1 Curcuminoids quantification by HPLC-DAD	121
3.3 Characterisation of curcuminoid-enriched formulations	122
3.3.1 Food digestibility assays	122
3.3.1.1 Starch hydrolysis	123
3.3.1.2 Protein hydrolysis	126
3.3.2 Oil and water holding capacities of fibres	128
3.3.3 Bulk and tapped density of fibres	129
3.3.4 Biscuit breaking force	129
3.3.5 Textural analysis of custard.....	130
3.3.6 Fibre colour determination.....	130
3.3.7 Particle size distribution of fibre sources	131
3.4 Curcuminoid bioaccessibility model development.....	131
3.4.1 Data description	131
3.4.2 Outlier removal.....	132
3.4.3 Feature preprocessing	134
3.4.4 Interference of colour	134
3.4.5 Feature selection.....	134
3.4.6 Model training.....	135
3.4.7 Model validation and assessment	137
4 Results	138
4.1 Food digestibility	138
4.2 Vitamin E	139

4.3 Curcuminoids.....	142
4.3.1 Bioaccessibilities	142
4.3.1.1 Validation of the spectrophotometric method to quantify total curcuminoids content.....	142
4.3.1.2 Bioaccessibility in food formulations	143
4.3.2 Supplement colour interference	147
4.3.3 Food matrix effects.....	155
4.3.4 Machine Learning model.....	160
4.3.4.1 Bias-variance tradeoff and selection of candidate models.....	163
4.3.4.2 Modelling imbalanced dataset including liquid matrices	166
4.3.4.3 Assessment and interpretation of candidate predictors	166
4.3.4.4 Assessment and interpretation of predictions	170
5 Discussion and conclusions	172
5.1 Systematic screening and modelling approach.....	172
5.1.1 Food matrix effects.....	173
5.1.1.1 Type of matrix.....	174
5.1.1.2 Macronutrient digestibility	177
5.1.1.3 Dietary fibre enrichment.....	180
5.1.2 Previous Machine Learning of compound absorption.....	184
5.1.3 Choice of model	186
5.1.4 Other bioaccessibility-optimising approaches	187
5.2 Vitamin E	189
5.3 Future perspectives	190
5.4 Conclusions	195
References	198

List of Figures

Figure 1. Factors influencing the popularity of functional food products, dependent on scientific evidence-based support.....	22
Figure 2. Graphical summary of the question addressed in this thesis and approach to resolve it.	24
Figure 3. Classification of functional foods and their components.....	29
Figure 4. Organs of the gastrointestinal tract. [Obtained from Verhoeckx et al., 2015].....	36
Figure 5. Structure of human stomach (left) and gastric wall (right) [courtesy of Encyclopaedia Britannica, Inc., copyright 2010]	39
Figure 6. Detailed representation of the small intestine, from lower gastrointestinal tract to villus structure. [courtesy of Encyclopaedia Britannica, Inc., copyright 2014]	42
Figure 7. Bioaccessible and bioavailable fraction of ingested components. [Obtained from Guerra et al., 2012, reused with permission from Elsevier]45	
Figure 8. Schematic diagram of the gastrointestinal 1 digestion model (TIM-1) [Obtained from Lafond et al., 2015].	50
Figure 9. Outline of the INFOGEST consensus protocol of 2019 [Obtained from Brodkorb et al., 2019, reprinted and minimally adapted with permission from <i>Springer Nature</i>]. SSF, SGF and SIF refer to simulated gastrointestinal fluids (salivary, gastric and intestinal, respectively).....	52
Figure 10. Properties of oil droplets known to affect lipid digestion and bioactive bioaccessibility.	59
Figure 11 . Chemical structure of vitamin E isoforms. α -tocopherol presents no insaturations on the non-cyclic tail and three methyl groups on the benzene chain in comparison to the other compounds of the family.	62
Figure 12. Chemical structure of curcuminoids. The loss of one (demethoxy-) or both (bisdemethoxy-) of the methoxy groups form the two other main curcuminoids.	64
Figure 13. Scatterplot of the datapoints (x_i, y_i) and their linear regression line, computed via ordinary least squares.	72
Figure 14. (a) Graphical illustration of bias and variance. (b) Contribution of bias and variance to the total error of the ML model. Increasing model complexities usually lead to lower bias and increased variance. [Obtained from Mandal et al., 2022].....	76
Figure 15. Example of a training:validation:test split with internal 5-fold cross-validation.	78

Figure 16. Linear SVM model. Two classes (red versus blue) are classified. [Obtained from Huang et al., 2018] w : weight vector, x : input feature vector, b : bias	79
Figure 17. Decision tree for the unadjusted treatment outcome of mental-health issues in young UK patients. Higher values represent poorer treatment outcomes. [Obtained from Marjolein Fokkema & Wolpert, 2021] Panels depict subgroup sizes (N) and estimated means (M) with standard errors (SE). SE s account for random variability between treatment centres, not for the searching of the tree structure.	82
Figure 18. In this example, the pretrained KNN model classifies the test sample as belonging to the c_1 cluster as its distance to that centre is smaller than the distance to c_2 cluster centre. [Obtained from Saadatfar et al., 2020]	84
Figure 19. Representation of x as a non-pooled predictor for three different data groups (left), or as a pooled predictor for all data groups (right).	91
Figure 20. Student's t distributions at different parametrizations.	92
Figure 21. Comparison of Student t 's and normal distributions at different parametrizations.	93
Figure 22. Absolute frequency probabilities of number of male mice to be found in a sample of 7 random mice, inferred from two different experimental designs both originally observing only 1/7 mice are male. The geometric distribution clearly illustrates there is sufficient evidence to conclude it is probable to only find one or no male mice, whilst in the binomial distribution experiment this was not evident thus the binomial probability of 50% of a mouse being a male mouse still stands. [Obtained from Hackenberger, 2019]	97
Figure 23. Laplace distributions compared to normal distributions.	101
Figure 24. Custards with (above) and without (below) curcuminoids, in the plastic 125 mL containers ready for rheology analysis.	105
Figure 25. Biscuits with and without curcuminoids.	105
Figure 26. HPTLC chromatogram of olive oil subjected to lipolysis by pancreatic lipase. It can be observed that the peaks of FFAs and MAGs (as a consequence of triacylglyceride digestion) are more pronounced at pHs at which pancreatic lipase retains more enzymatic activity (pH 7).	118
Figure 27. D-glucose measurement procedure according to Megazyme's booklet.	125
Figure 28. Boxplots of the bioaccessibilities measured for each formulation. The fibres are ranked from the lowest to the highest insoluble fibre content (left to right). Each formulation consists of a minimum of 5 biological replicates after outlier removal. The compact letter display was computed separately for custards (bottom) and biscuits (top). Formulations with different letters	

presented statistically significant differences to one another (according to Tukey tests).	145
Figure 29. Q-Q plot for evaluating the normality of the residuals across all formulations after 2-way ANOVA.	146
Figure 30. Colour properties (12-parameter) (A) and simple (6-parameter) (B) models' goodness-of-fit scatterplots of observed vs predicted averages (custard bioaccessibilities <40%, biscuits >40%). The intervals shown for each point represent the 95% highest density interval and have been included to assess whether the credible interval correctly captures the observed averages ($x=y$ line).	149
Figure 31. Posterior distributions of the linear regression coefficients of the colours model. Four MCMC chains were used for inference.	150
Figure 32. CIE L*a*b and CIE L*C*h colour spaces. [Obtained from Flament et al., 2014].	151
Figure 33. Training (left) and leave-one-out cross validation (right) assessments of the goodness-of-fit and generalization capacity of the final model (6-parameter). Each dot corresponds to the average observed bioaccessibility of each formulation and is compared to the average of its posterior predictive distribution. The intervals shown for each point represent the 95% highest density interval and have been included to assess whether the credible interval correctly captures the observed averages ($x = y$ line).	165
Figure 34. Spearman correlations between the formulation properties (excluding non-supplemented formulations due to missing values). It can be observed that none of the predictors of the 6-parameter model (macronutrient content, soluble fibre and hemicellulose) are strongly correlated to each other ($\rho > 0.5 $). Texture properties (8-parameter model) present stronger correlations to macronutrient content and hemicellulose.	169
Figure S35. HPLC separation with UV 430 nm detection, full separation (1) and zoom (2) for (A) curcuma standard 4 mg/L, (B) Biscuit without curcumin and (C) Biscuit with curcumin after digestion.	260
Figure S36. UV430 nm spectrum for CUR (a), UV 430 nm chromatogram (B) and extracted ion Chromatogram at 307.0975, 337.1091, 367.1187 uma for respectively BDMC (C), DMC (D) and CUR (E). LC-ESI(-)-HRMS for BDMC (F), DMC (G) and CUR (H).	262
Figure S37. Measured curcuminoid binding capacities. Measurements belong to experiments performed on separate occasions. The curcuminoid binding capacity attributed to each fibre was proportional to the highest added curcuminoids concentration under the 2.5 mM threshold (intended as a relative measure).	263

List of Tables

Table 1. Examples of peer-reviewed functional food studies.	29
Table 2. Ingredients of standard and curcumin enriched custards and biscuits.	106
Table 3. Nutritional composition of the base recipe ingredients according to products' labelling (in percentages). Empty cells represent unknown/non-existent data and were imputed as 0. *The types of fibre of corn starch and wheat flour were unknown but inferred to be mostly soluble and insoluble, respectively, in accordance with the literature (Shewry et al., 2023; Van Hul et al., 2020).....	107
Table 4. Ingredients (in percentage) of the fibre-supplemented formulations compared to base recipes.	110
Table 5. Supplemented fibre sources and their known nutritional composition (in percentages). Empty cells represent unknown/non-existent data and were imputed as 0. *fibre content was imputed from the literature.....	111
Table 6. Volumes for OPA calibration curve.....	127
Table 7. Candidate predictors of curcuminoid bioaccessibility. $L^*a^*b^*C^*h$ correspond to the axes of CIE L^*a^*b and CIE L^*C^*h colour spaces.	132
Table 8. Bioaccessibility of vitamin E in custard and biscuit formulations in which the base recipe's starchy ingredient (60 g of Maizena for custards and 60 g of T45 wheat flour for biscuits) was modified or substituted. Values recorded are the average and standard deviation from 3 biological replicates. *: egg yolks also contain vitamin E (recovery of vit. E from non-enriched formulations was 31% when the usual was >90%).....	139
Table 9. Bioaccessibility of vitamin E in custard and biscuit formulations supplemented with commercial fibre preparations. Values recorded are the average and standard deviation from 3 biological replicates. Higher supplementation levels were not possible in biscuits due to jellification of the extracted samples.	140
Table 10. Bioaccessibility of vitamin E in biscuit formulations at different levels of enrichment and small variations on base recipe. *: base recipe composed of 80 g of sugar, 100 g of egg yolks, 60 g of T45 wheat flour, 10 mL of milk and 106.7 mg/100g of curcuminoid powder and baked at 160° for 30 min.....	141
Table 11. Performance metrics of several Bayesian hierarchical models of curcuminoid bioaccessibility built with different feature combinations (including supplement colours). 4-parameter model uses a pooled intercept and macronutrient content predictor, and only soluble fibre content as non-pooled predictor.	151

Table 12. Spearman ρ correlation coefficients between colour properties and formulation properties used as predictors and the target bioaccessibility outcome.....	153
Table 13. Measured physicochemical properties of the supplements employed in this study and their resulting food products. The standard deviation was zero in cells where it was absent. *: Empty cells correspond to non-measurable values: curcuminoid binding capacity and water holding capacity of soluble fibres cannot be measured as the fibres dissolve in aqueous solutions. **: 25000 mPa.s was the saturation limit of custard viscosity measurements.	157
Table 14. Explanatory variables correlation to bioaccessibility and algorithms supporting their inclusion in model. The complete table can be found in appendix A (Table S18). Spearman rho (ρ) coefficients were used as correlation metric. These have been computed grouped by matrix type and for all formulations simultaneously. Properties selected for modelling in one matrix type have been indicated by numbers in parentheses next to their coefficient (1: selected by autoselect, 2: selected by Boruta, 3: final selection by Laplacian priors). Macronutrient content was the only nutritional composition feature considered during selection, as its component features (carbohydrate, fat, protein) presented high collinearity. n.a.: non-applicable. *: p-value between 0.01 and 0.05 **: p-value under 0.01	159
Table 15. Performance metrics of several Bayesian hierarchical models of curcuminoid bioaccessibility built with different feature combinations.	162
Table 16. Posteriors of the 6-parameter model. HDI: highest density interval 3% to 97%. The input data was normalised (scaled between 0, minimum; and 1, maximum) for each matrix group separately, except macronutrient content which was normalised across all samples.	167
Table S17. Summary of 3 main curcuminoids from turmeric curcumin	261
Table S18. Explanatory variables correlation to bioaccessibility and algorithms supporting their inclusion in model. Spearman rho (ρ) coefficients were used as correlation metric. These have been computed grouped by matrix type and for all formulations simultaneously. Properties selected for modelling in one matrix type have been indicated by numbers in parentheses next to their coefficient (1: selected by autoselect, 2: selected by Boruta, 3: final selection by Laplacian priors). The specific fibre features (cellulose, lignin) were not considered in the selection process due to their sparsity and reduction in applicability and generalization capacity of the model. Macronutrient content was not considered during feature selection as it was chosen as the pooled feature during hierarchical modelling. n.a.: non-applicable. *: p-value between 0.01 and 0.05 **: p-value under 0.01.....	264

List of Equations

(2-1) Univariate linear regression	70
(2-2) Least squares minimisation of error	70
(2-3) Partial derivatives in least squares	70
(2-4) Solution of the regression coefficients of least squares	71
(2-5) Additional property of least square regression coefficients	71
(2-6) Probability density function (PDF) of normal distribution	73
(2-7) Likelihood of PDF of normal distribution	73
(2-8) Log likelihood of PDF of normal distribution.....	73
(2-9) Optimisation of Maximum Likelihood Estimate (MLE).....	74
(2-10) Newton's method	74
(2-11) Newton's method applied to MLE	74
(2-12) Support Vector Regression primal optimisation problem	80
(2-13) Additional condition of the optimisation problem.....	80
(2-14) Linear equation	85
(2-15) L2 penalisation term.....	85
(2-16) L2 optimisation problem	85
(2-17) L1 penalisation term.....	85
(2-18) Bayes theorem.....	88
(2-19) Bayes theorem in modelling.....	88
(2-20) Simplification of Bayes theorem.....	88
(2-21) Student's t distribution likelihood.....	92
(2-22) Priors of linear regression coefficients	93
(2-23) Hyperpriors of Student's t likelihood	94
(2-24) Markov Chain Monte Carlo estimation.....	94
(3-1) Curcuminoid bioaccessibility.....	121
(3-2) Megazyme starch theoretical calculation	125
(3-3) Megazyme starch numerical calculation	126
(3-4) Megazyme starch simplified numerical calculation	126

(3-5) Oil holding capacity.....	128
(3-6) Water holding capacity.....	129
(3-7) Bulk density	129
(3-8) Tapped density	129
(3-9) Interquantile range criteria for outlier detection.....	133
(3-10) Hierarchical linear regression	136
(3-11) Response variable likelihood	136
(3-12) Pooled regression parameters.....	136
(3-13) Non-pooled regression parameters	136

List of Abbreviations

AGE	Advanced glycation products
AID	Automatic Interaction Detection
BCI	Bayesian credible Interval
BDMC	Bis-demethoxycurcumin
CART	Classification and Regression Trees
CUR	Curcumin
DAG	Di-Acyl Glycerols
DMC	Demethoxycurcumin
FFA	Free Fatty Acids
HDI	Highest Density Interval
HDL	High Density Lipoproteins
HMC	Hamiltonian Monte Carlo
HPLC-DAD	High Performance Liquid Chromatography – Diode Array Detector
HPTLC	High Performance Thin Layer Chromatography
IQR	Interquartile range
IVIVC	In vivo - in vitro correlation
KNN	k-Nearest Neighbours
LOOCV	Leave-One-Out Cross-Validation
MAE	Mean Absolute Error
MAG	Mono-Acyl Glycerols
MCMC	Markov Chain Monte Carlo
ML	Machine Learning
MLE	Maximum Likelihood Estimate
NUTS	No-U-Turn-Sampler
OPA	o-phthalaldehyde
PDF	Probability density function
RMSE	Root Mean Squared Error
SGF	Simulated Gastric Fluid
SIF	Simulated Intestinal Fluid
SSF	Simulated Salivary Fluid
SVM	Support Vector Machines
TAG	Tri-Acyl Glycerols

1 Executive summary

This thesis is structured into 4 chapters.

Chapter 1 - Introduction and literature review introduces the subject and its background according to the existing literature and knowledge of the state of the art. The chapter delves into the importance of nutrition and the state of functional foods in today's society, the known mechanisms playing a role in food digestion and bioactive compound absorption, and statistical methods commonly used to find and model explanatory causes of a phenomenon, in this case uniquely applied to bioactive compound absorption as a function of food matrix properties. A special attention is paid to explaining concepts that will be needed to comprehend the following chapters.

Chapter 2 – Methodology describes all methods and needed materials required to carry out the core work presented in this thesis. This includes the protocols for food preparation and digestion simulation, assays for characterising food properties, and data analysis and mathematical modelling procedures.

Chapter 3 – Results presents the main results of the thesis. Data presented includes the degree of absorption of vitamin E and curcuminoids in different custard, biscuit and liquid matrices; properties of said foods before and after digestion such as physicochemical properties or macronutrient content and digestibility; and exploration and modelling of the numerical differences using both classical and probabilistic approaches.

Chapter 4 – Discussion and conclusions interprets the novelty of this work in comparison to our current understanding and provides ideas for future avenues of research. The main focus is on the systematic approach to studying bioactive compound bioaccessibility that was conceived with curcuminoids, although vitamin E's results have also been discussed.

2 Introduction and literature review

Nutrition is pivotal for human health and overall well-being. A balanced diet, definition that continuously shifts as our understanding of food continues to develop, must fulfil all essential daily intake requirements thus negating the appearance of malnutrition issues (Cena & Calder, 2020). To this effect, food production plays a crucial role in facilitating the adherence to dietary recommendations.

The last 50 years have seen a shift in the habits of production and consumption of nutritious foods worldwide. Food systems that were predominantly rural have transitioned into economically-focused industrial engines, led by changes such as rising incomes, market liberalization, investment into foreign trade and local infrastructure, technological advances, population growth and urbanization, and changes in consumer demands (Ambikapathi et al., 2022). As a result, the most commercially available foods are usually those that follow the most profitable market trends instead of consumers' need for a healthy diet (White et al., 2020).

Developed countries have failed to adapt to the lower caloric requirements of sedentary lifestyles and lack of exercise resulting in the spread of obesity (Popkin et al., 2012), as energy-dense foods products, containing carbohydrates of high glycaemic index, sugar, salt and unhealthy fats, are easier to obtain than natural and healthy foods, contributing to this “globesity” plight. Poor diet is a major risk factor of non-communicable diseases, which account for approximately 63% of deaths worldwide, bringing about catastrophic societal and economic consequences (Bloom et al., 2011). And yet, this pandemic coexists, regardless of the income level of the country, with malnutrition amongst the poorest sectors of the population (Wells et al., 2021).

In the same manner, in spite of the increased abundance of food sources, poor dietary choices can lead to a lack of essential nutrients in the diet. Food does not only serve the purpose of basic nutrition. A healthy diet that contains an appropriate balance in nutritional components is required to avoid non-communicable diseases and conditions of poor health, such as anaemia (lack of iron), scurvy (prevented with vitamin C), to name a few examples (Kiani et al., 2022). The phenomenon called “hidden hunger”, in which the body lacks

micronutrients but does not manifest any clinical symptoms, has received special attention in young and the less fortunate groups, as deficiencies in iron, vitamin A, vitamin D and zinc are rather common among low- and middle-income countries which are reliant on low-cost foods (Bailey et al., 2015; Lowe, 2021; Weffort & Lamounier, 2023). There is an abundance of concealed effects that occur due to poor or unbalanced diets, which can have detrimental effects in the short as well as the long term, such as the increase in probability of suffering a variety of diseases such as diabetes, cardiovascular and kidney disorders, and neurodegenerative diseases due to the progressive accumulation of advanced glycation products (AGEs) (V. P. Reddy et al., 2022). AGEs are the product of non-enzymatic reactions between reducing sugars and proteins, lipids or nucleic acids (C. Prasad et al., 2019) and are being increasingly considered as an informative biomarker of biological health (Carrión-Barberà et al., 2024; de Vos et al., 2016). It has been proposed that lifestyle interventions, such as caloric restriction, could help minimise AGEs formation and accumulation (Zgutka et al., 2023). The maintenance of a prolonged homeostasis is a complex endeavour that requires temporary as well as long-term adjustments, and nutrition is one of the most impactful and adjustable environmental factors individuals can control.

The public as a whole has yet to act upon the growing body of knowledge which indicates optimal dietary choices. A survey carried out among Wuhan residents evidenced a total awareness rate of only 20.4% when being asked 32 questions about nutritional guidelines and skills (Y. Wu et al., 2022). In Poland, it was demonstrated that the majority of the population (over 73%) were unaware that dietary risk factors could be partly responsible for diseases such as stroke and osteoporosis (Żarnowski et al., 2022). With increasing lifespans, the assurance that quality of life will also be extended is questionable. Current evidence shows that increasing life expectancy does not correlate with increasing healthy life expectancy (years lived without disability) (Brayne, 2007; Brown, 2015; Xie et al., 2008) which will result in an expanding number of healthcare-dependent persons. The advances in healthcare, usually focused more on disease treatment and management and less on all-around health promotion, are an indirect cause of this phenomenon, but not the only one. It is therefore conceivable that it is the

responsibility of the individual to manage their own nutrition throughout their lifetime.

In light of these findings however, it is congruent that due to the global burden of diseases associated to poor diets, food policies are being introduced by governmental and charitable bodies to minimise their incidence. These are usually highly supported by the public if based on providing incentives (i.e. lower prices) or information (i.e. package labels), in contrast to setting restrictions (Kwon et al., 2019). Therefore, as healthcare systems are unlikely to begin prescribing dietary choices due to poor public acceptance, the feasible option is that the food producer presents pertinent information on its products' label and healthier foods receive economic governmental benefits. The public health issue becomes a strategic opportunity for development and innovation in the food industry; a favourable light is shed on foods with added value other than those that are simply consumed for macronutrient intake. The most productive efforts in this domain are usually consistent of partnerships between the public and private sectors (Olson et al., 2021); the endorsement of novel food products by public agencies enhances their consumer appeal, and in turn popular products can ensure the enforcement of the food policies is more widespread (Figure 1).

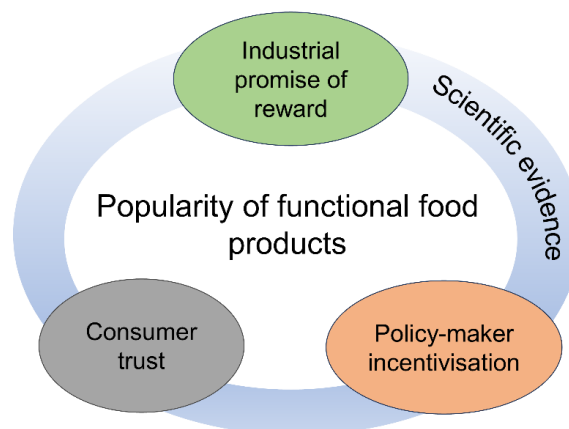


Figure 1. Factors influencing the popularity of functional food products, dependent on scientific evidence-based support.

Setting aside economic motivations to continue addressing hidden hunger issues, one of the main strategies that tackle malnutrition is supplementation with capsules or secondary foodstuffs (such as biscuits) (Mannar, 2006). This method

is a quick and effective way to combat a deficiency in a specific micronutrient in a defined population, provided that the access to food supplements is ensured and that compliance by the population is controlled (Faber et al., 2014; Young et al., 2010). An alternative strategy is to enrich staple foods (food fortification). The latter strategy has the benefit of being able to provide nutrients to large segments of the population without requiring radical changes in food consumption habits. The effects can be visible quickly since the consumer is passive in this approach, provided that the products enriched are consumed by the target population in adequate quantities without detrimental overconsumption (Dwyer et al., 2015).

Together with biofortification, which is the process of improving nutrient density in harvested crops (Bouis et al., 2011), and changes in dietary habits via education, these four enrichment strategies have been crucial for the United Nations' (UN) goals of sustainable development for the poorest people worldwide; more specifically the development goal 2 of the Decade for Action on Nutrition programme, spanning 2016-2025 (Atukunda et al., 2021; Lowe, 2021; United Nations, 2015). Hidden hunger is an issue that takes the centre stage in this project, as it is estimated to affect over 2 billion people worldwide and must be tackled in order to achieve zero hunger by 2030. To complete these objectives, a coordinated approach combining all the previously mentioned food enrichment strategies must be harnessed to achieve these goals.

Nevertheless, in countries in which the access to quality food is less of an issue, the market for enriched foods is also growing. Market research (N. M. de Carvalho et al., 2023; Precedence Research, 2023; Straits Research, 2024) indicates that the functional food market size was valued at \$298-\$305bn, with prospects of reaching \$336bn in 2023 and predicted to almost double in size in ten years' time; to \$526-597bn in 2031.

Despite the responsibility entrusted to these fortified products, the real effectiveness of the added components remains to be proven. One of the main challenges facing the development of fortified foods relies on the fact that the presence of bioactive compounds within the ingredients does not guarantee their presence in the finished product will retain the nutritional value after processing, or its intended biological activity after consumption; potentially invalidating any

health benefit claims. There is consensus in the scientific community that the assimilation rates of a bioactive compound can vary significantly from one food product to another due to the variability of chemical interactions that can occur within the food and the changes during food processing, as well as during digestion and intestinal uptake mechanisms. However, the manner in which the underlying biochemical processes that occur during digestion are affected by the food's characteristics, which in turn impact nutrient absorption, is not yet fully understood (Bjelakovic et al., 2015; Debelo et al., 2020; DeBenedictis et al., 2023; Hiole et al., 2020; Kamiloglu et al., 2021; I. F. McClements & McClements, 2023; Mishra et al., 2024; Turgeon & Brisson, 2020). In many cases, there is a lack of knowledge that precludes the recommendation of effective food production pipelines that purposefully optimise nutrient absorption. This gap can be addressed by systematically studying the impact of food characteristics (among other factors such as the co-ingestion of other foodstuffs) on nutrient absorption (Figure 2).

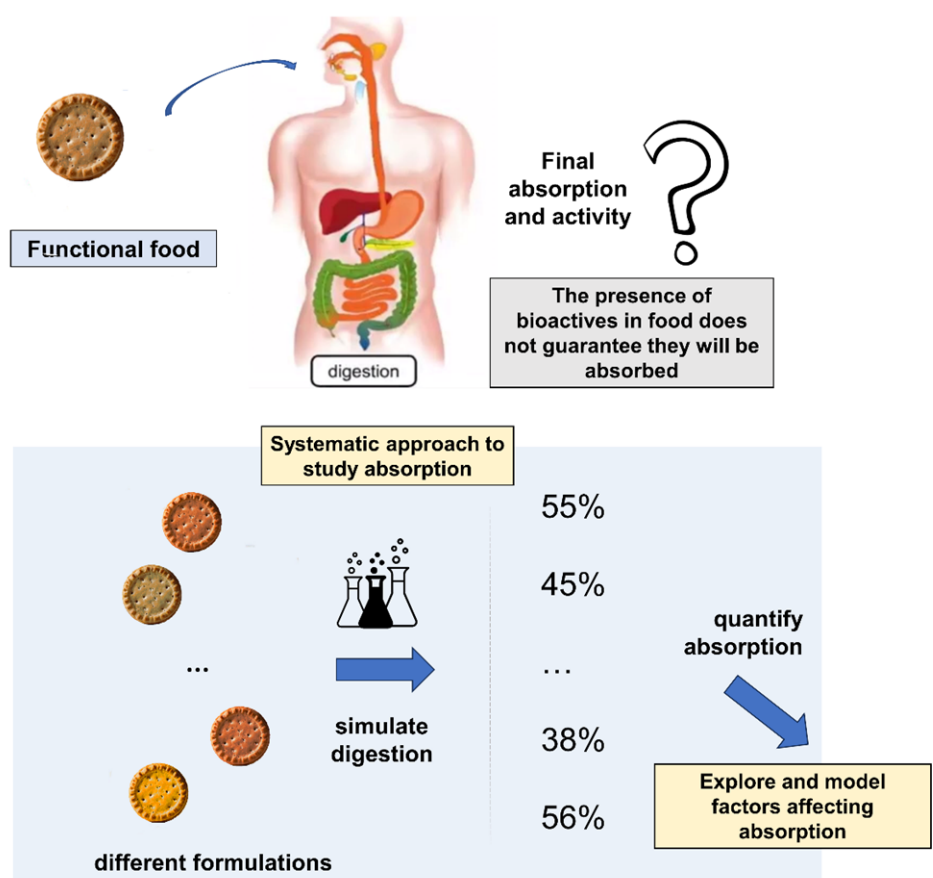


Figure 2. Graphical summary of the question addressed in this thesis and approach to resolve it.

Empirical testing of a large number of samples and conditions, as Figure 2 illustrates, necessitates the use of rapid screening methods to efficiently carry out the process of discovery. In addition, decision-support systems which are capable of filling in gaps trained on acquired empirical data, such as those afforded by the field of artificial intelligence (AI), can assist in this process.

This project aims to explore the impact of food characteristics on nutrient bioaccessibility in enriched food products, making use of techniques such as rapid screening methods and AI algorithms. One of the main goals of this project is to provide biological and chemical understanding of the factors affecting the bioaccessibility of nutrients of interest in food products, depending on the bioactive chemical structure, the biochemical composition, and the preparation process of the food product. Moreover, the experimental data generated will be used to develop a decision-support system consisting of predictive mathematical models that optimise the formulation process and enhance the nutrient bioaccessibility for a given food product. This goal will be achieved through the application of machine learning (ML) and artificial intelligence (AI) coupled with extensive series of studies of compound bioaccessibility using simulated digestions.

Unlike previous studies which mainly focused on a binary relationship between a single phytochemical and its corresponding particular food matrix, in this project a systematic approach is followed which considers the emergent properties of all the associated components forming the food matrix, rather than their individual structure, in order to obtain an accurate prediction of bioaccessibility for each bioactive present.

The result is a proof-of-concept study that will indicate whether absorption of a food component can be reasonably inferred according to the properties of the food that carries it. Thus, the posterior analysis and developed decision support system would provide knowledge about the specific case study used to prove the concept, which could be useful to future iterations. But more importantly, food matrix effects in general are not well understood, and any advance should garner the attention of any stakeholders (government, academia, industry) interested in designing, optimising, and/or validating the nutritional quality of foods.

Highlights

- Functional foods containing bioactive compounds are fundamental for fulfilling nutritional requirements.
- Presence of bioactive compound in food does not equate to presence in body tissue.
- Systematic approaches required to develop an understanding of the factors involved in absorption.

2.1 The food matrix

The complex assembly and physiochemical interactions that occur between the nutrients and non-nutrients present in a food constitute its food matrix. It is a term used to denote that compounds behave differently in isolation than when forming part of foodstuffs (Aguilera, 2019). The macro- and micro- nutrients that appear on nutrition labels form complex microstructures that are responsible, more so than the food's chemical composition, of certain properties of the food such as its texture (Lillford, 2011). The importance of the food matrix when discussing nutrition, compared to that of only considering the nutrients the food contains, can be put into perspective thanks to Novotny, Gebauer and Baer's study on almonds (Novotny, Gebauer, et al., 2012). In said study, the researchers discovered that the theoretical energy content of almonds, as determined by Atwater factors, was overestimated by 32% versus the empirically observed energy. This suggests that the food matrix interactions have a significant effect on every aspect of the digestive process. This also raises significant concerns over the reliability of current practices in food labelling.

Food matrix and microstructure are therefore an intertwined and critical consideration in nutrition. And yet, the characteristics of these microstructures are within the realm of control of the food manipulator; they can be modified in response to macroscopic changes in the food, which can be desirable when seeking to improve sensory or functional qualities of the meal (Mishra et al.,

2024). For example, different cooking methods of potatoes affect their nutritional impact, as the food matrix as a whole suffers physiochemical alterations that alter the state of the nutrients they contain. Boiled potatoes are leached of water-soluble nutrients but induce higher satiety than fried potatoes, whose starch is converted into resistant starch that is harder to digest and absorb; this results in a low early glycaemic response and satiation of fried versus boiled or mashed potatoes (Leeman et al., 2008; Robertson et al., 2018). The lycopene content of tomatoes is most highly retained after short microwave or oven treatments; if frying or longer periods of high-temperature baking, over double the lycopene content can be lost (Mayeaux et al., 2006). By finely grinding almonds, plant cell walls are broken, releasing the lipids they contain and increasing the chance that they are fully absorbed (Berry et al., 2008).

The manner in which the composition of the food affects the food matrix can also impact nutritional quality, as the food matrix components can potentially interact in a manner that difficult the release of nutrients. For example, increasing amounts of dietary fibres, which cannot be digested by humans, can have interfering effects with the digestion of other compounds. They were found in several independent studies to negatively impact absorption of protein and fat resulting in a loss of metabolizable energy (Baer et al., 1997; Miles, 1992).

All the previous examples point to the need of empirical assessments of nutritional quality in foods that are intended for human consumption. Just because a food presents a nutritional component of interest does not mean it will be truly available to the body after being consumed. This consideration is exacerbated when developing products that carry compounds with additional value, such as functional foods.

Highlights

- Food matrix is more than the sum of its components.
- These complex interactions require addressing.

2.2 Functional foods

This term is typically used to define foods that have been specifically formulated and proven to provide additional health benefits aside from basic nutrition (i.e., antioxidants, vitamins, beneficial microorganism activity) (Temple, 2022). This is because being able to effectively prove a food's health claims is the main criteria by which food regulation agencies allow producers to label their products as functional foods (Ovesen, 1997). As an example of this definition put into practice in the European Union, Regulation [EC] 1924/2006 was adopted in 2006 on the use of nutrition and health claims for foods (European Parliament & Council, 2006). Ninety-five percent of health claims of functional foods pertain to health claims other than those referring to the reduction of disease risk and to children's development and health (article 13 of 1924/2006 regulation), thereby referring to health claims of interest for the majority of the population. The approval rate is only around 10%, based on the grounds that the evidence for said claims is insufficient; a low success rate which is startling given the fact the requirements for approval are not particularly strict: there is no need to adhere to peer-reviewed journal publishing or minimum clinical trial sample sizes (Duttaroy, 2019).

However, recommendations for steps to take to obtain approval exist and typically include the previously mentioned (Martirosyan et al., 2022). The current state of affairs surrounding functional foods therefore demands they can unequivocally demonstrate efficacy in promoting human health.

2.2.1 Categories and examples

Functional foods can be categorized according to different criteria based on how they are prepared, what type of functional or beneficial components they incorporate, and what degree of evidence backs any associated health claims. Based on the first two previous criteria (Damián et al., 2022; N. M. de Carvalho et al., 2023), they can be classified as follows in the next page.

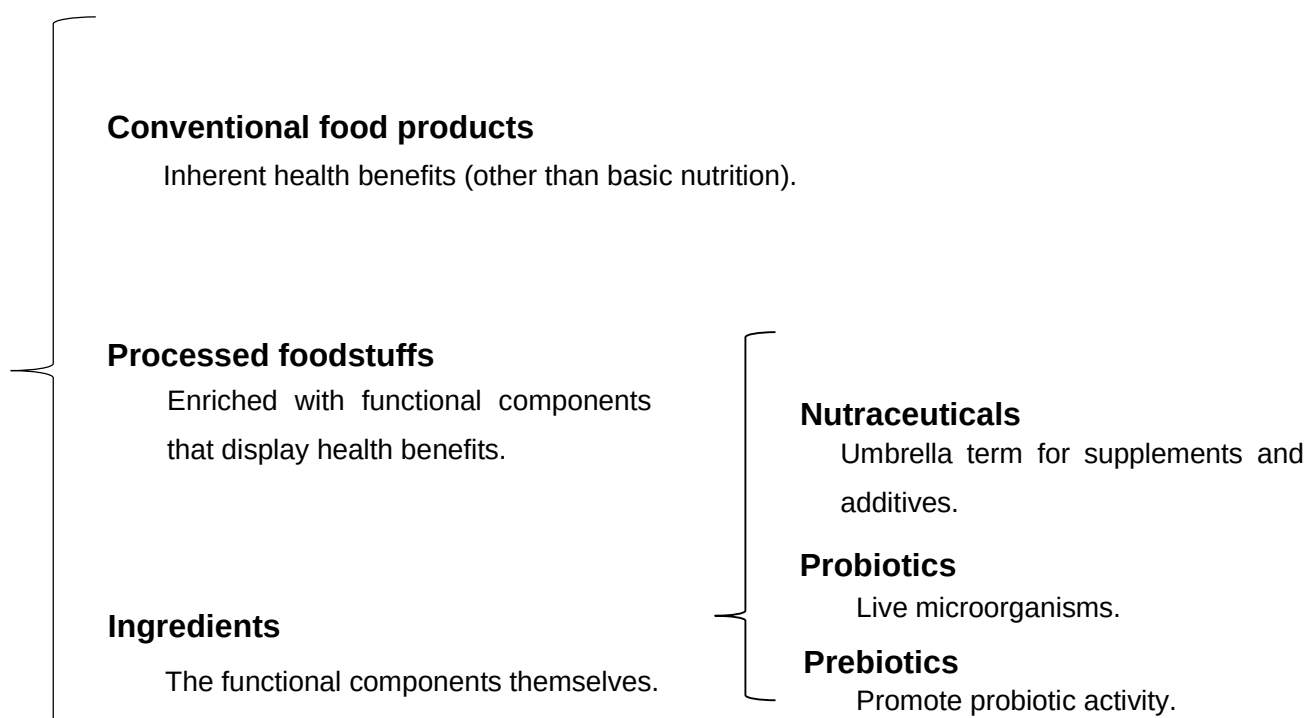


Figure 3. Classification of functional foods and their components

Examples of studies on functional foods and their classification according to these criteria are compiled in Table 1.

Table 1. Examples of peer-reviewed functional food studies.

Functional food	Functional component	Preparation	Function	Reference
Fermented whey protein	<i>Lactobacillus</i> and whey protein	Processed	Prebiotic, probiotic	(P. Sharma et al., 2017)
Pumpkin and by-products	Carotenoids, lutein, zeaxanthin, vit. E, ascorbic acid, etc.	Conventional, ingredient	Nutraceutical, prebiotic	(Aziz et al., 2023)
Curcuminoid-enriched oat fibre-corn based snack	Curcuminoids and oat fibre	Processed	Nutraceutical, prebiotic	(Sayanjali et al., 2019)
Porridge enriched with rice bran and nutraceutical compounds	Vit. E, γ -oryzanol, ferulic acid and rice bran	Processed	Nutraceutical, prebiotic	(Calvo-Castro et al., 2019)

Medical drink including phosphatide precursors and cofactors	Uridine, omega-3 poly-unsaturated fatty acids and choline	Processed	Nutraceutical	(Scheltens et al., 2010)
Iron supplementation	Iron	Ingredient	Nutraceutical	(R. Li et al., 1994)
Ground beef enriched with soybean phytosterols	Phytosterols	Processed	Nutraceutical	(Matvienko et al., 2002)
Bread and biscuits supplemented with leaf powder extract	Phenolic-rich extract	Processed	Nutraceutical	(Ferreira et al., 2023)
Immature wheat grain biscuits	Fructo-oligosaccharides	Processed	Prebiotic	(Casiraghi et al., 2011)
Custard pudding gels with high antioxidant activity	D-psicose and antioxidant compounds	Processed	Nutraceutical	(Sun et al., 2006)
Activia® yoghurt	<i>Bifidobacterium animalis</i> and fructooligosaccharides	Processed	Probiotic, prebiotic	(De Paula et al., 2008)
Inulin supplementation	Inulin	Ingredient	Prebiotic	(Amadiou et al., 2022)

Recently, the Functional Food Center (FFC), a world-leading institution in functional food development that provides guidelines for their development and market-launch, proposed a new categorization according to the quality of the evidence backing their health claims. In this classification (Martirosyan et al., 2022), functional food products are evaluated not only on whether the product was certified as functional, but also if epidemiological studies and aftermarket research support the claims, which would provide an additional layer of evidence-based proof for products that reach the marketing stage.

2.2.2 Ingredients: phytochemicals and dietary fibres

Phytochemicals are bioactive molecules generally perceived as healthy and natural by the public, as they are produced organically by plants for their own protection. They are usually classified as nutraceuticals, although from a biochemical point of view many different classes of phytochemicals exist. These include carotenoids, phenolic compounds (flavonoids, isoflavonoids, lignin), phenolic acid, phytosterols and phytostanols, tocotrienols, organosulfur compounds, and nondigestible carbohydrates (dietary fibres) (Steinmetz & Potter, 1991). Their consumption has been linked to a potential reduced risk of chronic diseases such as cancer and heart disease, offering various health benefits including anti-inflammatory, antioxidants, antibacterial, antifungal, chemopreventive, antiaging, and osteoporosis effects (Agrawal et al., 2023; Liskova et al., 2020). It has been also shown that phytochemicals interact with each other; for example the polyphenol curcumin (elaborated upon in section 2.6.4) has an affinity for certain types of fibres, such as sugar beet pectin, a soluble prebiotic fibre, which can improve the shelf life of curcumin within the food matrix (Zagury et al., 2021). Animal and human studies show that curcumin's impregnation into a soluble fibre isolated from the fenugreek spice, or its complexation with cellulosic derivatives, enhanced its bioavailability (Hegde et al., 2023; Krishnakumar et al., 2012).

Dietary fibres such as those previously mentioned usually fall into the prebiotic category, as they promote the growth of gut microflora. There is currently great public interest in consuming dietary fibre due to several health benefits it is associated with, such as increased overall metabolic health, lower risks of cardiovascular disease, colonic health and gut motility issues, colorectal carcinomas; an overall negative correlation to mortality; and the capacity to decrease the overall glycaemic index of digested food, therefore decreasing insulin resistance of subjects (Anderson et al., 2009; Barber et al., 2020).

Recently, the increase in accessibility of metagenomic and metabolomic platforms has significantly improved the study of probiotics and the understanding of their effects on human health (Kwoji et al., 2023). Nonetheless, the study of the impact that a set of microorganisms and the compounds they produce has on human health is more complex than studying the impact of the compounds

independently (i.e., host-microbe interactions, microbial community dynamics). Nutraceuticals and prebiotics are therefore the main focus of this research project.

To summarise, the combination of phytochemicals and fibre supplements enables the production of a wide range of fibre-fortified foods, which fulfil consumer demands whilst delivering high quality nutrition.

2.2.3 Marketing

As already mentioned in the introduction to the chapter, market research indicates that in 2022 the functional food market size was valued at \$298-\$305bn, with prospects of reaching \$336bn in 2023 and predicted to almost double in size in ten years' time; to \$597bn in 2032. Dairy products hold the largest market share of all functional foods (39% revenue share in 2022), with a significant proportion of them falling under the probiotics category (N. M. de Carvalho et al., 2023; Precedence Research, 2023). European and Chinese consumers were shown to be willing to pay 30-50% extra for functional foods (Menrad, 2003; Miroso & Mangan-Walker, 2018) (China's traditional nutritional medicine culture is very prevalent) compared to foods without said health claims. However, others have proposed that, after the COVID-19 pandemic in which supplement sales spiked by around 50% worldwide in March 2020, because said supplements were thought to be prophylactic against the disease, that the demand will slowly return to pre-pandemic levels (Hamulka et al., 2020; Lordan, 2021). In any case, nutraceuticals are at the centre of the current and foreseeable demand, with food additives such as dietary fibres, carotenoids, minerals, fatty acids, and vitamins attracting the most attention (Grand View Research, 2023).

This increase in demand is motivated by the popularisation of nutrition as a means for achieving wellbeing, which is further promoted by healthcare professionals' usual inclusion of functional foods as part of numerous support therapies (Puri et al., 2022). Seeking the optimisation of food and beverage products to fulfil these demands is viewed as a crucial requirement for the industry, and there is significant investment in research and development to make this happen. However, having a patent on a potential food product is not sufficient, as evidenced by the worldwide state of technological development of

functional food chocolates, which is mostly still at the level of patents (Balcázar-Zumaeta et al., 2023). Industry has a recognised need for academic-level research into the true potential (impact on health) of the functional foods.

In the design and production of foods fortified with nutraceuticals or prebiotics, gauging consumer acceptance is critical for their widespread acceptance. In 2022, Baker et al. addressed this subject in detail by reviewing 75 studies conducted around the globe that empirically investigated the matter and found that the determinants for consumer acceptance encompassed product characteristics, socio-demographic factors, and consumers' psychological, behavioural, and physical characteristics (Baker et al., 2022). It must be noted that most of these studies are survey-based, which can suffer from inherent subjective biases (Andrade, 2020), and they were conducted in different geographical zones, making it challenging to draw conclusive insights across the board. One observation was that foods that were designed with carriers perceived as healthier (or inherently wholesome) were preferred (fibre-enriched cereals preferred over calcium-fortified juice) (Verbeke et al., 2009), and that the manner in which the functional components were manufactured or inserted also bore weight in the consumer's choice, with preference for carriers in which natural enrichment was perceived to occur, such as milk over sausage or pate as a carrier of vitamin D (Jahn et al., 2019).

In spite of the health benefits being the identifying characteristic of a functional food, many consumers chose them simply because they preferred their taste and were unwilling to compromise, therefore it cannot be assumed that consumers will accept functional foods that are not tasty (Verbeke, 2006). Nutrition and health claims increased purchase intention compared to those with reduction of disease risk claims, and physiological-based claims were preferred over psychological-based claims, according to surveys (Siegrist et al., 2008; van Kleef et al., 2005). Social pressure from peers to consume functional foods to care for one's health and wellbeing should not be understated, as it was found to be a strong predictor of consumers' purchase intent in a Norwegian study (Nystrand & Olsen, 2020). The subjective perception of the consumer towards the ease of purchasing and consuming these foods plays an important part in the tendency to accept them as well (Xin & Seo, 2020).

Other factors that increased the likelihood of acceptance included (Baker et al., 2022; Topolska et al., 2021):

- Familiarity with the food type and/or brand.
- General education and/or knowledge of functional foods, although providing information portraying scientific uncertainty could reduce this willingness, and using common names elicited better perceptions than scientific names.
- Old age or poor subjective health.
- Being female, possibly due to the role they conventionally play within the family unit.
- Higher income households.
- Households with young children or teenagers.

Increasing consumer trust in functional foods depends on the perception that the benefits outweigh their risks, which requires trust in food science, food industry, food safety control systems and the advertising strategy. Food neophobia, or scepticism to novel foods, has been demonstrated to exist and should not be ignored; product design and development must include the perception of the food as an important factor in the process (Tuorila & Hartmann, 2020).

Therefore, increasing the consumption of functional foods depends not only on satisfying regulation demands, but also on nurturing the trust of the consumers and proving the rewards that functional food promises are relevant and backed by strong scientific evidence, worth the price, and with minimal/non-existent negative effects. One of the main challenges facing the development of functional foods relies on the fact that the presence of functional components in the ingredients does not guarantee their presence in the finished product will retain the nutritional value after processing, or its intended biological activity after consumption; potentially invalidating any health benefit claims; a compound must be available to the body so that it can exert any beneficial effect (Debelo et al., 2020; Kamiloglu et al., 2021; Mishra et al., 2024; Parada & Aguilera, 2007). Assimilation rates of the same bioactive compound can vary significantly from one food product to another due to the variability of chemical interactions that can occur within a food matrix and the changes during food processing, as well as

during digestion and intestinal uptake mechanisms. Therefore, to verify a functional food's practical usefulness and not betray consumer trust, nutritional quality must be exhaustively investigated through rigorous scientific studies that encompass a diverse range of food formulations and processing techniques. Only by meticulously examining how these factors influence the efficiency of absorption and potential health benefits of the bioactive compounds can we ensure accurate labelling and provide consumers with reliable information about the functional foods they are consuming. The following section will provide an overview of the human digestive system and how the knowledge we currently have can shape the study of compound absorption.

Highlights

- Functional foods can be classified into; 1— preparation type: processed, natural or ingredients; and 2— functional nature: nutraceutical, prebiotic or probiotic.
- There is great consumer interest in functional foods incorporating phytochemicals and dietary fibres.
- Consumer trust is essential for acceptance of functional foods.
- Empirical research into the validity of their health claims is a recognised necessity by all parties involved.

2.3 Human digestion

In this section the complex process that produces the catabolism of nutrients will be described. The food's macro- as well as micro- nutrients must be released from the food matrix in order to be absorbed by the body and used as either building blocks or metabolizable energy for the cells to grow, replicate, and maintain homeostasis. This process is known as digestion.

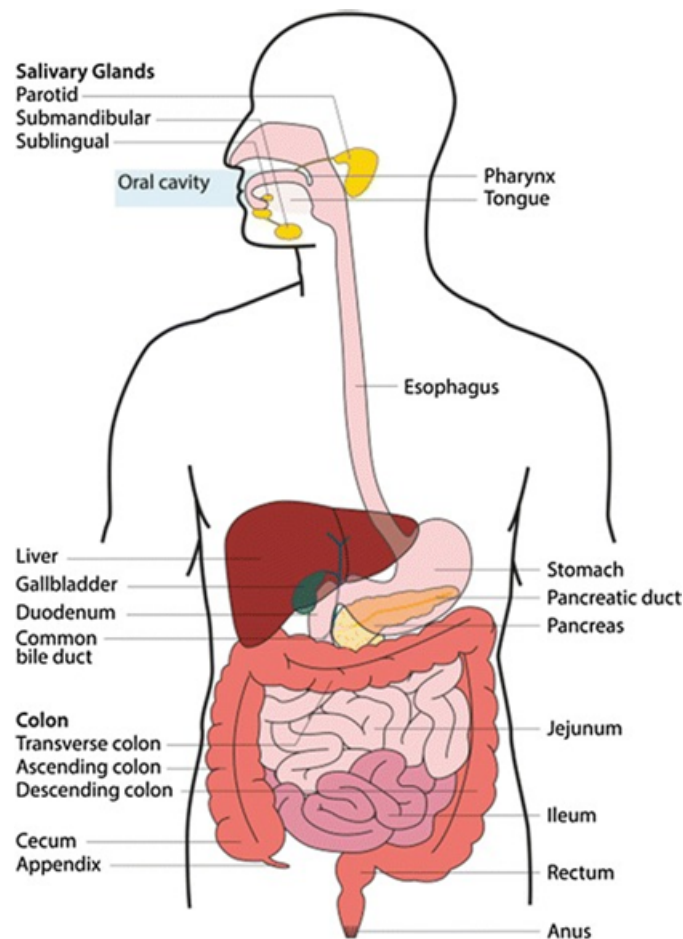


Figure 4. Organs of the gastrointestinal tract. [Obtained from Verhoeckx et al., 2015]

2.3.1 Oral cavity

This is the first stage of the digestion process, where food is masticated and broken down into smaller pieces, mixed with saliva, and later sent to the oesophagus. Many factors influence the formation of the oral bolus in this phase, ranging from the individual's intrinsic properties such their eating ability and their age or gender (Ketel et al., 2019), to the physical state of the ingested food (Hwang et al., 2012).

Food enters the mouth, where mastication grinds it down into smaller pieces and mixes it with saliva, lubricating, softening, and aggregating the food particles so they are exposed to enzymatic breakdown (Palmer et al., 1992). This is necessary so that the oral bolus is safe and easier to swallow. This mechanical process involves the teeth, masticatory muscles, temporomandibular joint and tongue compression against the palate, with saliva serving as solvent. The higher

the number of masticatory cycles, the higher the effective surface area of the swallowed bolus (increased accessibility of enzymes) (Cassady et al., 2009; Hollis, 2018). Whole mouth saliva consists of saliva secreted predominantly (90%) by three paired major glands; the parotid, submandibular and sublingual glands, assisted by 600 to 1000 other minor glands (Pedersen et al., 2018). Daily production and usage of saliva is around 0.6 L (Watanabe & Dawes, 1988), although this secretion can vary according to many factors such as the hardness, moisture and size of the food and the time it takes to chew and swallow it.

Saliva consists of approximately 99% water; however, it contains critical amounts of chemical substances such as sodium, potassium, calcium and bicarbonate which help maintaining a 6-7.5 pH in the mouth and to remineralize the teeth, mucin for lubrication and microbial protection (other example defensive proteins are lysozyme and lactoferrin), epidermal growth factor helps with soft tissue healing and protection, and most importantly, enzymes such as amylase and lingual lipase (Iorgulescu, 2009). However, it is still unclear whether lingual lipase exists in humans, as it was only demonstrated to be present and active in rodent tongues), whose gene is ortholog to the gene of gastric lipase in humans and other species (Bodmer et al., 1987; Docherty et al., 1985; Salhi et al., 2021). A possible explanation for detectable lipolytic activity in the mouth (as was detected in 45% of a 42-person study carried out and proposed by Lai et al.) could be microbial sources or gastric reflux rather than lingual human lipase, albeit the detected activity levels were only meaningful for sensory taste purposes rather than fat digestion (Lai et al., 2019).

On the other hand, human salivary α -amylase is accountable for around 30% of total starch hydrolysis, present in a proportion of 40-50% of total protein content in human saliva (Grand et al., 2004; Mandel et al., 2010; Noble, 2000). Salivary amylase cleaves starch molecules down into maltose, which is cleaved into glucose by maltase enzymes predominantly in the small intestine (Nichols et al., 2018). Salivary amylase is encoded by the gene AMY1 (Rukh et al., 2017), and genetic variation (copy number variations, or CNVs) in this gene can predict an individual's amount and activity of salivary amylase which results in altered carbohydrate taste perceptions (Mandel et al., 2010). People with low genetic

capacity for starch digestion can present low BMIs due to lesser calorie extraction, regardless of high carbohydrate intake (Rukh et al., 2017).

To summarise, amylase activity and mechanical breakdown are the predominant mechanisms occurring in the short oral phase, which can last from a few seconds to several minutes, in which food sensory properties can change rapidly as the mastication and mixing with the saliva takes effect and ends when the pharynx sends the bolus further down the digestive tract.

2.3.2 Stomach

After oral processing, the oral bolus travels through the 25 cm in length of the oesophagus down to the stomach, taking approximately 8 to 10 seconds. The main process occurring in this region of the tract is mechanical in nature and is the most prevalent of its kind throughout the digestive process although it is especially relevant in this organ: peristalsis. This term refers to the wave-like involuntary muscle movement that moves the food through the tract, except in the mouth in which mastication occurs instead (Buettner et al., 2001; Sensoy, 2021). In the oesophagus, a primary peristaltic wave is triggered by swallowing, followed by a secondary wave should the first one be insufficient as an adaptive reflex response to oesophageal distention (Gupta et al., 2009). The bolus enters the stomach in synchronization with the relaxation of the upper and lower oesophageal sphincters, composed of both skeletal muscles (upper) and smooth (lower) muscles (Mittal, 2011).

The stomach can be considered as a vessel for storing, mixing and sieving the partially processed food in preparation for its controlled release to the intestine, as peristalsis and stomach contractions slowly move the newly formed chyme (fluidized mixture containing particles smaller than 2 mm in size) across and towards the pylorus sphincter in small amounts and at regular intervals of 2-3 cycles per minute (Mackie et al., 2020).

The previously mentioned mechanical processes play an important role in this organ, which can be split into fundus, body (corpus), and antrum (Figure 5).

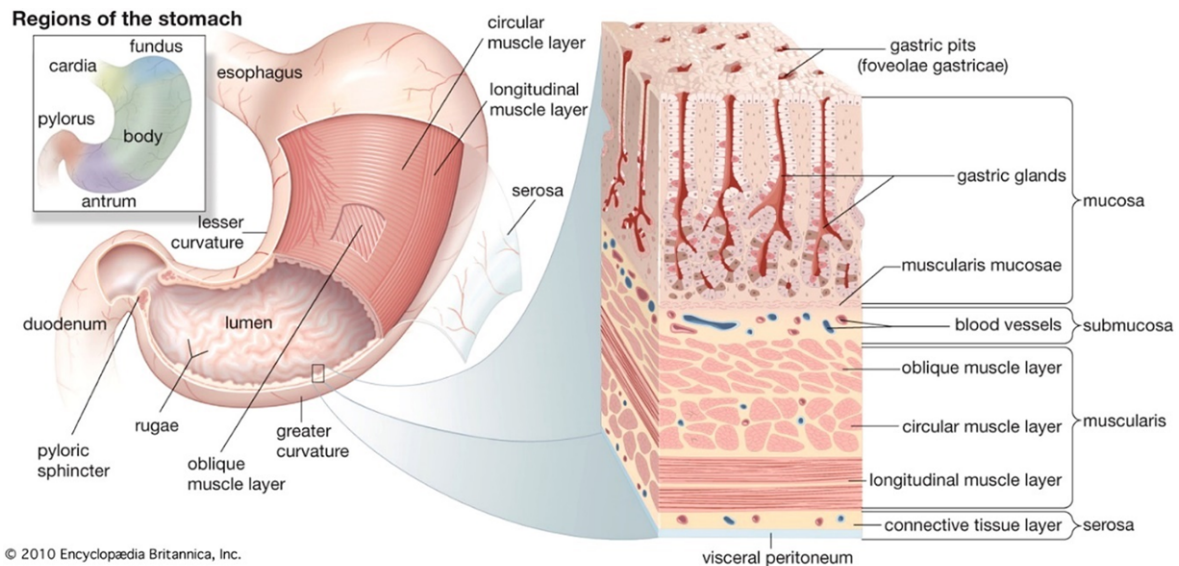


Figure 5. Structure of human stomach (left) and gastric wall (right) [courtesy of Encyclopædia Britannica, Inc., copyright 2010]

The fundus receives the food and swallowed air and stores them until space is made available for digestion in the lower sections, especially around the pyloric sphincter where mixing is vigorous and allows the passing of soluble fractions to the duodenum. Indigestible fractions must wait 120 min for migrating peristaltic waves to transport them to the intestine (Brandstaeter et al., 2019). The stomach walls are composed of several layers of muscle which allow it to expand to up to 4 L in volume, from an original volume of 25-50 mL (Sensoy, 2021). The gastric mucosa contains the glands which will secrete digestive enzymes and dissolution substances whilst controlling pH change through mechanisms such as gastrin release (Brandstaeter et al., 2019; Walsh et al., 1975). The mucosa can secrete around 1.5 L of gastric juice per day, containing water, hydrochloric acid, electrolytes (such as bicarbonate), mucus, enzymes (lipase and pepsin), hormones (such as gastrin) and the intrinsic factor (necessary for absorption of vitamin B12 in the ileum) (Al-Awami et al., 2023; Camilleri & Vazquez Roque, 2014).

The stomach is notably acid in comparison to other regions of the digestive system, and bicarbonate is needed to keep a protective pH gradient in the mucus layer adjacent to the stomach and duodenum walls (Allen & Flemström, 2005). Most of the existing literature on the subject agree that pH of the fasted state is typically around 2 and that of the fed state around 4.5, prior to returning to the

fasted state (Brodkorb et al., 2019; Mudie et al., 2010); there is significant variability depending on the subject of study and experiment (F. Schneider et al., 2016). Although this pH appears unfavourable for microbial activity, certain populations, such as *Campylobacter pyloridis* and *Helicobacter pylori*, are capable of surviving. However, their primary significance lies in their role in causing diseases rather than contributing to digestion. More research is necessary to comprehend the role of gastric microbiome on digestion (Nardone et al., 2017; Steinbach et al., 2024).

Clarifying what is the stomach pH under a particular set of conditions is important to understand how digestive enzymes activity is affected and to consider potential chemical transformations of the food due to the acidic environment. As was mentioned earlier, pepsin and gastric lipase are the main digestive enzymes secreted in the stomach. Gastric lipase's catalytic activity has a pH optimum in the range of 4-6 (Armand et al., 1996) whilst pepsin has it at pH 1.5-2 and can adopt a catalytically inactive conformation between pH 4 to 6.5 (Campos & Sancho, 2003). Pepsin is an endopeptidase secreted by the stomach as pepsinogen, an inactive precursor to protect the secretory glands from its digesting action, which suffers an autocatalytic conversion by cleaving itself by a conformation change due to the acidic pH (1.5 to 2, same as its pH optimum). It is responsible for approximately 15% of protein digestion only (Heda et al., 2023; Smith & Morton, 2010). Gastric lipase is responsible for between 10 to 30% of triglyceride digestion, with the remainder occurring in the lower gastrointestinal tract (Martin & Freedman, 2019).

Gastric emptying is a key parameter in digestion studies because it directly influences the kinetics of nutrient absorption. The normal range for gastric emptying transit time is 2 to 5 hours, however it can vary from as little as 15 minutes (fasted state) to 16 hours depending on the properties and amount of food ingested (Y. Y. Lee et al., 2014; F. Schneider et al., 2016). For characterising gastric emptying rate, two parameters are frequently used; latency time or time of fastest gastric emptying (T_{lag}) and half gastric emptying time ($T_{1/2}$) (Hellmig et al., 2006), although by scintigraphy the whole curve can also be computed (Abell et al., 2008; Banks et al., 2023; Navas et al., 2021). These measurements are particularly important in gastroparesis, a heterogeneous disorder in which gastric

emptying is delayed. Gastric emptying studies are standardised across patients and experiments by requiring the consumption of a standardised liquid egg-whites meal, after which a normal gastric emptying rate corresponds to less than 60% of the meal being retained after 2 hours since consumption; less than 30% after 3 hours, and less than 10% after 4 hours. Greater retentions than those described indicate delayed gastric emptying, although rapid gastric emptying can also be inferred from individuals in which after 30 min less than 70% of the meal is retained, or less than 30% after just one hour (Abell et al., 2008).

When the acidic chyme leaves the stomach into the intestine it first enters the duodenum, and to avoid gastric secretions from entering the intestine the hormone secretin provides a negative feedback to gastrin action by inducing the pancreas to secrete bicarbonate-rich fluid into the intestine (Llanos et al., 1977; Walton, 2009).

2.3.3 Small intestine

Most of the food digestion takes place in the small intestine, which is divided into duodenum, ileum, and jejunum (Figure 6). After leaving the stomach, the chyme enters the duodenum where it is mixed with bile and pancreatic enzymes (amylase, protease and lipase capacities) at a neutral pH. It continues through the jejunum and ileum where the partially digested food is further hydrolysed and absorbed (95% of absorption takes place in the small bowel) (Fish et al., 2024; Mackie et al., 2020).

In the duodenum pancreatic enzymes, bile produced by the liver and stored in the gallbladder, and other secretions increase the chyme volume and will effectively breakdown nutrients into absorbable forms. Secretions in the ileum and jejunum are minimum. Bile acts as an emulsifier to facilitate lipid digestion, and is composed of bile salts, phospholipids, cholesterol, electrolytes, bilirubin and water. After its use, 95% of it is reabsorbed in the distal ileum (Mackie et al., 2020; Sensoy, 2021).

Peristalsis again occurs in the small intestine to keep propelling the chyme, at a rate of $2\text{--}25\text{ cm s}^{-1}$, throughout the 6-7 m long tube, and segmentation contractions aid mixing and squeeze the chyme with the intestinal walls so nutrient absorption is maximised (Bornhorst & Paul Singh, 2014). The transit time

of food in the intestine does not depend as much on the fasted state as it did in the stomach (4.5 h on average) (F. Schneider et al., 2016).

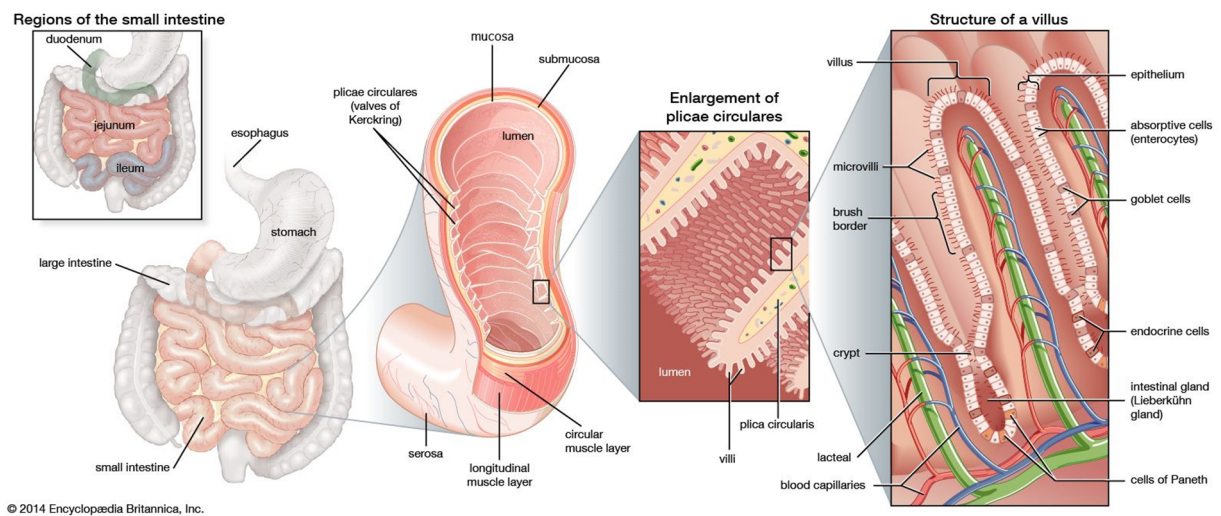


Figure 6. Detailed representation of the small intestine, from lower gastrointestinal tract to villus structure. [courtesy of Encyclopædia Britannica, Inc., copyright 2014]

Even though the small intestine is only 3-4 cm in diameter, its internal folds result in an absorptive surface area of around 4500 square meters (Hightower et al., 2023). The brush border's microvilli (Figure 6) secrete the enzymes that break down the food into the final components capable of being absorbed by the intestine (such as disaccharides into monosaccharides and peptides into amino acids), which are later transported into the bloodstream (Patricia & Dhamoon, 2023).

Pancreatic amylase, similar to salivary amylase, breaks down starch into disaccharides, but cannot break down lactose. The brush border's microvilli secrete sucrase, maltase and lactase to facilitate the final breakdown into monosaccharides, of which the SGLT1 transporter carries glucose and galactose into the intestine. Fructose is carried by the GLUT family of transporters (Burant et al., 1992; Sensoy, 2021).

Pancreatic proteases and peptidases (trypsin, chymotrypsin, carboxypeptidase, and elastase instead of pepsin) are assisted by brush border peptidases (also secreted in their non-active forms as in the stomach) to break down the proteins and peptides, the latter of which are very efficiently absorbed by the peptide

transporter 1 (PEPT1), although 90% of absorbed dietary proteins are represented in circulation by amino acids, which are absorbed by a myriad of transporter-based systems. It has been observed that PEPT1 is also expressed in the colon in order to compensate for conditions such as inflammatory bowel disease and short bowel syndrome (Kiela & Ghishan, 2016; Patricia & Dhamoon, 2023). Similarly, recent evidence indicates that phenotypes such as obesity and type 2 diabetes can be linked to the upper gastrointestinal tract microbiome, especially that of the jejunum. However, this area of research requires further study (Steinbach et al., 2024).

The action of gastric and pancreatic lipases results in the generation of a variety of lipid compounds from the predominant triglycerides. Medium fatty acids can be directly absorbed by the enterocytes without hydrolysis, cholesterol is mainly absorbed by the NPC1L1 transporter, and longer chains and monoglycerides are absorbed as micelles formed with bile salts which can include fat-soluble compounds as well as other phospholipids and cholesterol too (Kiela & Ghishan, 2016; Patricia & Dhamoon, 2023).

2.3.4 Large intestine

This section is called as such because of a larger diameter (6 cm) compared to the small intestine, but it is shorter (1.5 m). It is connected to the ileum via the ileocecal valve to the cecum, which leads to the colon, and absorbs 90% of the remaining water in the chyme and ejects the rest as waste, by movements similar to those in the small intestine (Kahai et al., 2023; Sensoy, 2021).

A remarkable characteristic of the large intestine is that it is an anaerobic environment that fosters a large community of microorganisms ($\sim 10^{11}$ cells/g). These can metabolize dietary fibres, bile salts and pancreatic enzymes, as well as synthesizing secondary metabolites and fermenting undigested material, all of which can provide benefits to the host such as strengthening gut integrity, harvesting energy, pathogen protection, and regulating host immunity. The study of dysbiosis, the term for altered microbial composition, is attracting a lot of scientific attention due to disruption of these health benefits (Rudi & Zhao, 2021; Sensoy, 2021; Thursby & Juge, 2017).

The unabsorbed and undigested food material, including small amounts of nutrients, water, inorganic matter, bacteria and cellular constituents of the digestive tract, form faeces and progress to the rectum where they are eliminated (Ogobuiro et al., 2023).

Highlights

- The gastrointestinal journey comprises oral, gastric, and intestinal phases.
- Most aspects of digestion are very well understood, but there are some disagreements (i.e., the case of lingual lipase, or fed vs fasted state).
- This knowledge enables developing simulations *ex vivo*.

2.4 Absorption: the concepts of bioaccessibility and bioavailability

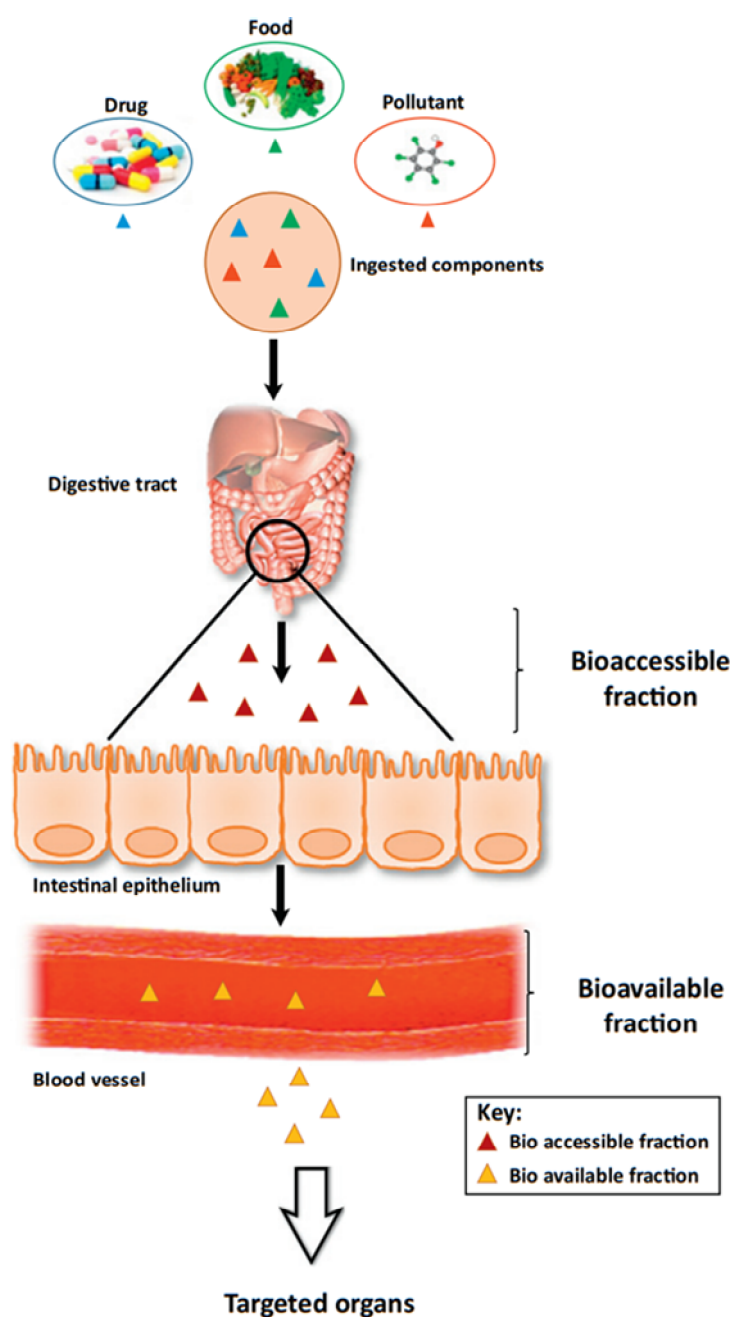
Nutritional efficiency is related to two concepts: bioaccessibility, and bioavailability (Figure 7). In this section, the definition and implications of these two aspects of absorption will be explained.

The efficiency of absorption of a bioactive compound depends on many factors and occurs in several stages that are interconnected as is the nature of the digestive process. The production and consumption choice of what combination of bioactive compound(s) and food carrier(s) will deliver a desirable outcome falls under producer/consumer control, but only by understanding what occurs during the gastrointestinal digestion and absorption process (covered in the previous section) can they be efficiently designed for maximal benefit.

Absorption of a compound or drug has traditionally been an important concern in the branch of pharmacology (Shargel & Yu, 2015), and it is from this scientific domain that most initiatives tackling absorption of functional components stem from, although its conceptualizations can still be applicable from a nutritional perspective.

In pharmacology, bioavailability is a subcategory of absorption that is expressed as the percentage of the total drug (or compound) dose administered that reaches

the (systemic) circulation untransformed (Paul, 2019). This includes both free and bound forms of the compound or drug.



TRENDS in Biotechnology

Figure 7. Bioaccessible and bioavailable fraction of ingested components.
[Obtained from Guerra et al., 2012, reused with permission from Elsevier]

Theoretically, bioavailability indicates the degree of potential profitability from an initial dose of a functional component of interest, prior to observing its potential health benefit in effect. Intravenous bioavailability is considered as absolute, but oral bioavailability can be limited as a consequence of the complex digestion process, and what is known as first-pass metabolism (loss due to metabolically active tissues such as the liver) (Price & Patel, 2024). Ideally, bioavailability would be based on tissue levels at the site of intended action (instead of systemic circulation), however this is typically not possible to quantify for most applications (non-invasively and non-destructively) (Wolverton, 2020).

To complement the holistic nature of estimating the degree of post-absorption effect of a functional component, another concept frequently characterised is its bioaccessibility (Figure 9). Bioaccessibility has been defined as the fraction of a compound which is released from the food matrix in the gastrointestinal lumen and thereby made available for intestinal absorption (Saura-Calixto et al., 2007). It can be viewed as the first step of bioavailability, which accounts for the enzymatic and mechanical breakdown of the food in the gastrointestinal lumen, up to, but not including, intestinal absorption. Bioavailability is usually the research target of studies addressing bioefficacy of functional components as it is a more complete approximation to compound absorption than bioaccessibility, however bioaccessibility enables the assessment of crucial aspects of the absorption process such as compound liberation from the food matrix (Rein et al., 2013). It is generally safe to assume that compounds with low bioaccessibility will present low bioavailability as well, as the non-bioaccessible fraction is mostly excreted.

In vitro bioaccessibility has been used as a proxy to *in vivo* bioavailability in the past, and this approach has been demonstrated as valid in certain applications, but also limited in others (T Bohn et al., 2018). In some cases, all the bioaccessible fraction is bioavailable, but there are other cases in which only a small amount is absorbed after digestion. Combining bioaccessibility and bioavailability studies enables the identification of the limiting factor for a compound's absorption and therefore to focus the efforts on adjusting the food matrix as required. Troubleshooting a highly bioaccessible compound that presents poor bioavailability is out of the scope of this project, although to

illustrate this problem it should be noted that strategies do exist; for example previous authors suggest that reducing the antinutrient content (absorption-interfering compounds) can help resolve this issue (Huey et al., 2022). This project focuses on bioaccessibility which greatly depends on the compound's molecular properties, so this warrants a case-by-case discussion. The following section will describe methods for simulating digestion in order to carry out bioaccessibility assessments.

Highlights

- Absorption of a compound can be attributed to two sequential stages: release from the matrix, and uptake and redistribution.
- Bioaccessibility refers to the fraction of a compound that is released and solubilised in the digesta, ready for final stages of absorption.
- Bioavailability refers to the fraction of a bioaccessible compound that reaches its intended site of action (or alternatively, the systemic circulation).

2.5 Simulation of human digestion: *in vitro* models

Gastrointestinal fate of ingested compounds of interest can be studied either with *in vitro* or *in vivo* techniques. Animal models enable a more realistic and accurate study of human absorption compared to non-animal models, but are not so practical to study bioaccessibility and raise ethical questions and potentially financial issues. The nutrients released from the food are immediately absorbed by the organism, so unless this mechanism is disabled, which compromises the homeostasis of the animal, bioaccessibility can be misrepresented. For bioaccessibility, simulating the digestion with *in vitro* models that are as realistic as possible is the best option and one of the most frequently used methods, but require a careful consideration of the results (Nicolescu et al., 2023). On a positive note, *in vitro* alternatives provide their own set of advantages compared to *in vivo* methods, such as ethical considerations, higher reproducibility, and increased convenience (T Bohn et al., 2018).

There are three main types of *in vitro* digestion models, which will now be described.

2.5.1 *In silico* models

Before considering *in vitro* models, it must be stated that *in silico* models are a very desirable approach to simulate digestion. By using mathematical approaches and a set of user-controlled parameters, these provide the convenience of not requiring experimental lab work; they accelerate and simplify the screening process of different formulations. Unfortunately, the currently available *in silico* models are very limited.

During the second decade of the 21st century, several holistic 3D computational models of the gastrointestinal system employing fluid mechanics to simulate mastication and mechanical processing were developed. However, these mostly focused on liquids without including secretions and enzymatic reactions (Le Feunteun et al., 2020). Alternative models focusing on a specific aspect of digestion are more prolific but have less applicability, such as the model developed by Moxon et al. for studying the impact of viscosity and gastric emptying rate on glucose absorption (Moxon et al., 2016), or the comprehensive protein digestion model recently developed by Rivera Del Rio et al. which includes the features that were missing from the 3D computational models (Rivera Del Rio et al., 2022). Regarding the simulation of micronutrients, Marze's model for vitamin A digestion successfully reached similar conclusions to that of empirical experiments (Marze, 2015).

Overall, *in silico* models are promising and there are examples proving that they can replace empirical experiments, but their development usually requires a case by case study supported by extensive empirical experiments in order to properly simulate *in vivo* conditions (Le Feunteun et al., 2020).

2.5.2 Dynamic *in vitro* models

Dynamic models are the most elaborate *in vitro* digestion set-ups, but also the ones that most closely resemble the *in vivo* conditions. The first family of dynamic models can be attributed to Minekus et al., the TNO Gastro-Intestinal models

(Mans Minekus et al., 1995), although more have since been developed (Kong & Singh, 2010; Ménard et al., 2014; Wickham et al., 2012).

Dynamic models are either multi-compartmental, or mono-compartmental, such as the Dynamic Gastric Model (Wickham et al., 2012). In this single compartment (which only simulates the stomach), a computer model automatically adjusts secretions according to estimated emptying rates and physical processing, being able to simulate both fed and fasted states and allowing researchers to take samples at any stage of the process. It has been validated as an accurate simulator of *in vivo* mechanical processes in the stomach (Vardakou et al., 2011), and has seen increasing use by the pharmaceutical industry to study disintegration and dissolution of drug formulations and their delivery to the duodenum (Thuenemann et al., 2015).

Multi-compartmental processes, such as the TIM-1 system (from the TNO family) (Mans Minekus et al., 1995) have separate compartments for each organ of the digestive system (Figure 8), connected by valves and semi-permeable membranes in which the successive conditions can be simulated. A computer program is also responsible for the control of the digestion parameters which enables exhaustive customization of the gastrointestinal conditions such as for simulating age, health status, or adjusting to different types of food (Dupont et al., 2019). It is a popular simulation system as a result, and has been validated for many different applications including the computation of human glycaemic response curves after carbohydrate intake (Bellmann et al., 2018), digestibility and bioaccessibility of amino acids (Schaafsma, 2005) and accurate estimation of stability and bioaccessibility of different micronutrients (including fat-soluble compounds) (Déat et al., 2009; Ribnicky et al., 2014; Verwei et al., 2006), all of which showed similar results to what was observed *in vivo*.

A limitation of the TIM-1 system is that, in order to assess bioavailability, the intestinal absorption must be simulated separately; either with cell lines or *in silico* modelling. Another very important limitation is that each experiment can only simulate the digestion of one formulation, which significantly extends the duration of the screening process unless several experimental apparatuses are available (for simultaneous processing), which usually is not the case as they are

composed of complex laboratory equipment. Miniaturization is a promising technology in this regard, potentially easing the acquisition and usage of these sort of devices in the future (Colombo et al., 2023). But until then, complexity and low scalability are some of the main reasons why, in certain situations, other types of *in vitro* methods are preferable.

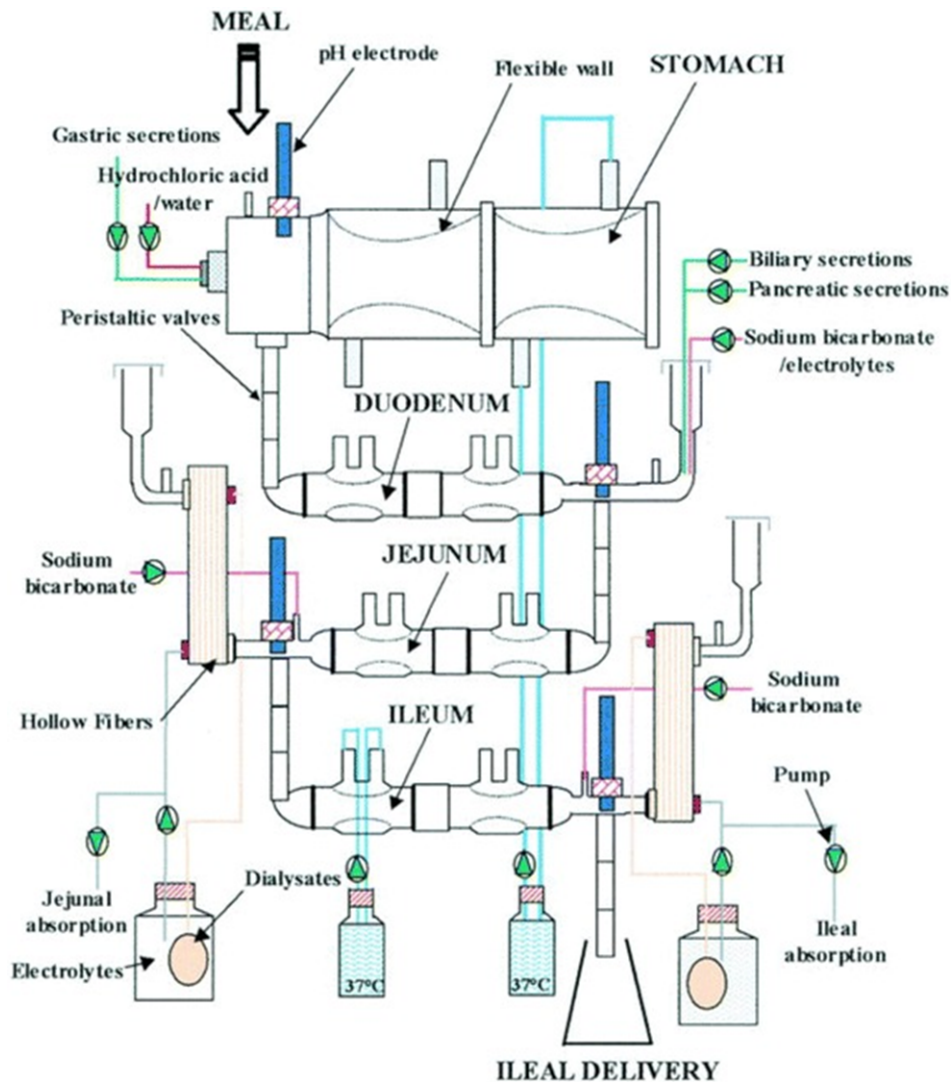


Figure 8. Schematic diagram of the gastrointestinal 1 digestion model (TIM-1)
[Obtained from Lafond et al., 2015].

2.5.3 Static models

The most widespread, economically viable and time efficient *in vitro* digestion models currently available are static in nature, meaning at each step of digestion the ratio of meal to digestive agents is constant.

This is the opposite of dynamic systems in which at each step of digestion there is a constant monitoring and adjustment of simulated fluids and reagents (M. Minekus et al., 2014). Prior to the publishing of the COST Infogest consensus protocol of 2014 (M. Minekus et al., 2014), static *in vitro* digestion protocols were very inconsistent: the fact that they were so accessible compared to dynamic systems meant they were also much more flexible to researcher's specific requirements. Static protocols prior to the standardised 2014 protocol varied greatly depending on the analyte, sample, and research question. For example, for the study of carotenoids in vegetables, Alminger et al. dispensed with the oral phase and microfiltered the supernatants of the gastric phase to proceed with the intestinal phase (Alminger et al., 2012). Similarly, in a study of carotenoids, vitamin C and phenolic compounds in a blend of fruit juices, Rodríguez-Roque et al., which did not include the oral phase, subjected 200 mL of fruit juice to gastric digestion by the addition of solid pepsin, and proceeded to the intestinal phase using a dialysis membrane (Rodríguez-Roque et al., 2013). For the study of metallic elements absorption in chocolate drink powder, Peixoto et al. did include the oral phase, but also employed bovine serum albumin in both the gastric phases and intestinal phases (Peixoto et al., 2013). Instead of using pancreatin in the intestinal phase, Granado-Lorencio et al. employed individual enzymes and cofactors (pancreatic lipase, colipase, cholesterol esterase, and others) (Granado-Lorencio et al., 2011). These are but a few examples of how almost every static digestion protocol diverged in one aspect or another.

The introduction of the 2014 consensus method, and especially the 2019 revision (Brodkorb et al., 2019), has brought about a greater consistency between studies. Therefore, the comparison of results is much easier, albeit these results are not as representative of the *in vivo* conditions as the more costly dynamic systems, however the protocols were developed rigorously according to validation with *in vivo* methods. The INFOGEST research network that formulated the standardised protocols is formed by some of the most respected authorities in the field. Several of the authors already mentioned during this section (Minekus, Alminger, Marze, Wickham, Le Feunteun, and others) were part of the elaboration of this consensus protocol. It simulates all the steps of the digestion except intestinal absorption (Brodkorb et al., 2019), however this is not a disadvantage

compared to most other dynamic systems as they do not include this step either. The specific steps of the INFOGEST protocol will be elaborated upon in the Methodology section as it was the selected *in vitro* model for this project, although the outline can be visualised in Figure 9 (Brodkorb et al., 2019).

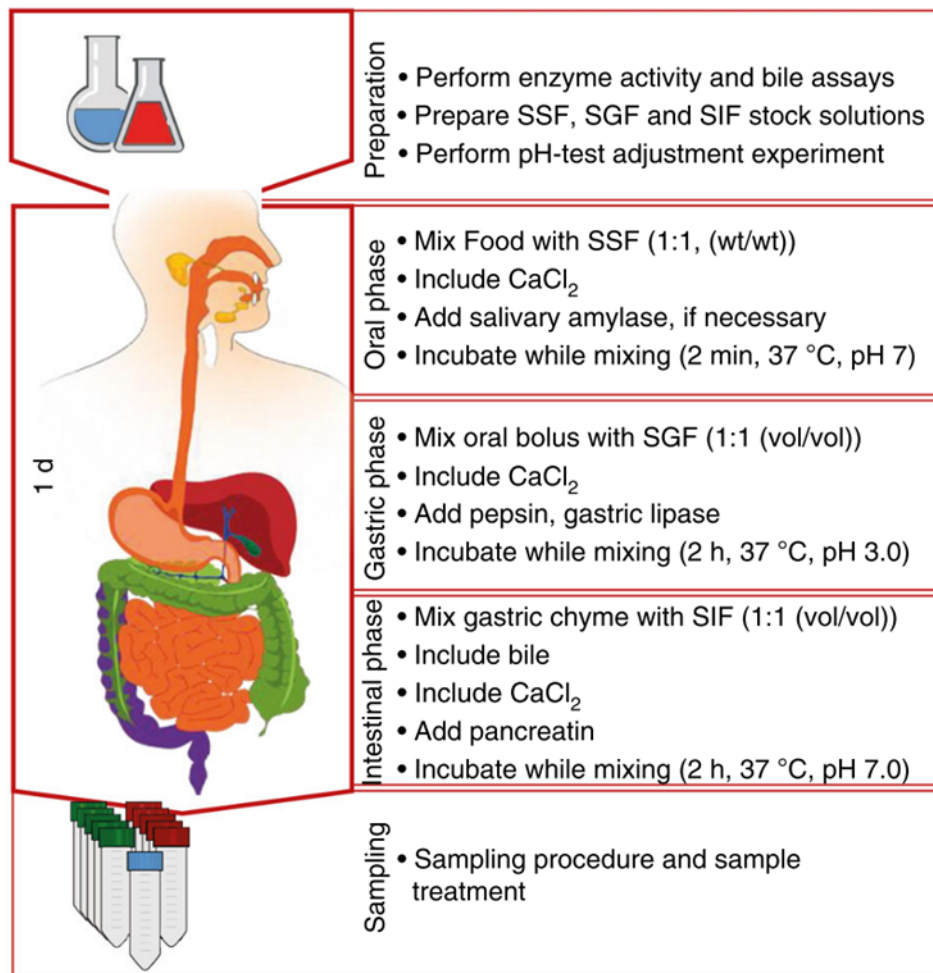


Figure 9. Outline of the INFOGEST consensus protocol of 2019 [Obtained from Brodkorb et al., 2019, reprinted and minimally adapted with permission from *Springer Nature*]. SSF, SGF and SIF refer to simulated gastrointestinal fluids (salivary, gastric and intestinal, respectively).

Recently, additional changes have been made to this protocol in subsequent studies, such as the introduction of a colon phase with or without bacterial strains, alterations to the protocol to allow the study of infant and pathological conditions, or simulation of intestinal absorption using cellular models (i.e. *ex vivo* with Caco-2 cells) or dialysis membranes, however there is yet no consensus on these implementations (Colombo et al., 2021). In addition, Gomes et al. highlighted that,

in the case of hydrophobic compounds, the higher the hydrophobicity, the higher the tendency to hide within the oil droplets and thus not be bioaccessible (Gomes et al., 2019). This hydrophobicity-bioaccessibility correlation was observed by Sy et al. *in vitro* ($r = -0.99$) and it was also discovered that the bioaccessibility was perfectly correlated with *in vivo* bioavailability ($r = 0.99$). This suggested that bioaccessibility was the main step governing absorption. Further investigation using *in vitro* bioavailability studies with Caco-2 cells confirmed that once the compounds were bioaccessible, their physicochemical properties did not influence their uptake (Sy et al., 2012). This lessens the concerns over not including the intestinal uptake step in *in vitro* models for the case of lipophilic compounds.

2.5.4 Semi-dynamic models

These models represent an intermediary approach between the highly complex and accurate dynamic systems, and the convenience of static models. Semi-dynamic *in vitro* digestion systems can account for the different rates of gastrointestinal flow as a function of food matrix rheological/texture properties better than static *in vitro* methods, as well as to facilitate sampling (Markussen et al., 2021). There are fewer instances of these models in the literature than their more clearly defined counterparts, however there are some that require mentioning.

The most relevant semi-dynamic model was also conceived by researchers involved in the INFOGEST network (Mulet-Cabero et al., 2020). The main difference concerning the semi-dynamic model and the static model published a year earlier (Brodkorb et al., 2019) is the gastric phase, which is more resemblant of the dynamic models by progressive adjustments of the gastric conditions (pH, electrolytes, enzymes). By including the main dynamics of gastric physiology including gradual pH decrease, gastric fluid secretion and gastric emptying, the semi-dynamic model provides more detailed data about the kinetics of nutrient digestion.

To conclude, although INFOGEST protocols have been a significant development in the realm of *in vitro* digestion, they are general protocols that are still usually altered or expanded upon depending on the compound of interest, as

they were mostly developed and validated focusing on protein digestion, and for adults (Xavier & Mariutti, 2021) .

In general, almost all *in vitro* digestion models present limitations compared to *in vivo* experiments. By not assembling intestinal epithelial cells into the models that can absorb the released compounds, oversaturation can occur. This can lead to underestimating bioaccessibility in cases where nutrients are present in the digesta in concentrations over their solubilisation limit. In addition, not including the gastric emptying phenomenon has been shown to significantly affect the accuracy of static *in vitro* models (R. Wang et al., 2021). Gut microbiota can severely impact bioavailability as well depending on the compound in question; the standardised models are not particularly suitable for compounds that undergo significant microbial transformation unless accounted for in the experimental design, and there is yet no consensus on how to simulate this aspect *in vitro* (Veintimilla-Gozalbo et al., 2021).

To summarise, static *in vitro* digestion models offer a quick, reproducible and cost-effective option to rapidly screen bioaccessibility in different formulations, offering insights that can *a posteriori* be confirmed with more accurate models such as dynamic systems or *in vivo*.

The following section will cover current advancements in absorption-enhancing strategies for two different families of fat-soluble compounds, as this type of compounds typically present low bioaccessibilities with a large optimisation potential.

Highlights

- Several methods for simulating digestion *ex vivo* exist.
- Dynamic models are the most accurate, but least convenient.
- Static models are the most popular due to efficiency and scalability.
- All methods present limitations that are currently not addressed (intestinal cell uptake and gut microflora).

2.6 Bioaccessibility of fat-soluble compounds

Functional components of a hydrophobic nature present difficult absorption, as they must be dissolved in hydrophobic domains such as bulk oils, fat droplets, liposomes or micelles to remain bioaccessible, with the interfacial layers playing an important part in facilitating digestion as well. These oil-based processes are also susceptible to interferences from other compounds such as proteins, polysaccharides, minerals and other additives. The complexity of these matrix interactions results in the bioaccessibility of lipophilic components being very variable, with values ranging from 10 to 80%, indicating that thoughtful food matrix design could confer a significant improvement in absorption (P. Reddy & Jialal, 2023; Y. Tan & McClements, 2021).

The use of enhancing agents, encapsulation and emulsion techniques has been proven by numerous studies (Wijekoon et al., 2023; Zabot et al., 2022) to increase the uptake of hydrophobic compounds of interest and has since garnered a lot of attention in the scientific community. However, many of these evaluate absorption based on the efficiency of their transporter systems exclusively, taken in isolation, and do not factor in the impact that other co-ingested foods or supplements could have on the hydrophobic domains. Using well-characterised complex food matrices offers more relevant insights as this dietary pattern of consumption is more common in real-life situations. However, most foods are not consumed in isolation either but factoring in all possible multi-food meal combinations is not feasible, so it is pertinent to choose a model food matrix containing a rich variety of nutrients.

2.6.1 Gastrointestinal journey of fat-soluble compounds

To comprehend how a meal carrying a fat-soluble compound of interest affects its absorption, the involvement of the physicochemical mechanisms occurring during digestion must be reinspected in this new context. The end-goal of lipid digestion, in the context of ensuring efficient fat-soluble compound solubilisation and absorption, is to incorporate them into mixed micelles that can be taken up by the intestinal epithelium cells, and studies have shown that reduced lipid digestion is correlated to reduced bioavailability (Mutsokoti et al., 2017).

The oral phase is short but can have a major impact on fat-soluble compound bioavailability; mastication breaks down and homogenises the food and saliva changes its physicochemical properties, with mucin promoting flocculation of natural oil bodies and starch-hydrolysing amylases capable of releasing any compounds trapped within, both of which generally favour absorption (Q. Wang et al., 2021).

As mentioned earlier, the pH and transit time in the gastric region of the digestive tract can vary greatly depending on the fasting state, however it has also been shown that lipolysis mainly occurs within the first hour in the stomach, with the maximal rate being reached in just 4 min (Bauer et al., 2005; Hamosh et al., 1975), therefore it can be inferred that lipid digestion in the stomach is efficient even after small portions of food are ingested. However, the presence of fat in the small intestine slows gastric emptying (Little et al., 2007), so this effect can be expected to increase gastric time even in an empty stomach if there was previously a recent feeding. The low pH of the fasted state can also affect the structure of oil droplets coated by acid-unstable emulsifiers, promoting aggregation and leading to a faster gastric emptying (X. Wang et al., 2019) that reduces the extent of the activity of gastric lipases and a lesser release of bioactive compounds from protein-containing foods by pepsin intervention. Even at a higher fed-state pH, gastric lipolysis greatly depends on the accessibility of the oil-water interface by gastric lipase; smaller droplet sizes are hydrolysed faster than larger ones (TAG hydrolysis can range between 5% and 37% as a result), and flocculation and coalescence effects such as those happening at low pH can minimise this accessibility further (Acevedo-Fani & Singh, 2022; Armand et al., 1999). Furthermore, the use of emulsifiers that minimise these aggregation effects (such as pectin or Tween 80, Espinal-Ruiz et al., 2016) could still result in a reduced lipolysis if they prohibit adsorption of lipase onto the emulsion. (Infantes-Garcia et al., 2021).

Stomach churning slowly leads the gastric chyme into the intestine, where the remainder of macronutrients are digested and prompting bioactive release. The gastrointestinal movements break up larger oil droplets which increases their surface area and exposes them to increased enzymatic processing but can also promote aggregation (droplet coalescence or flocculation) which reduces it (Y.

Tan & McClements, 2021). Bile salts can displace emulsifiers and remove free fatty acids (FFAs) from droplet surfaces by solubilising them in the mixed micelles, facilitating digestion (Pilosof, 2017). Calcium ions can remove FFAs by precipitation, and they are also required for optimal pancreatic lipase activity. (Alvarez & Stella, 1989; Corte-Real & Bohn, 2018; Eich et al., 2018).

The combined action of the lipase/colipase complex, bile salts and calcium ions removes or breaks down larger lipids from droplet surfaces, resulting in the original fat-soluble compound carriers being replaced by mixed micelles formed of FFAs, monoacylglycerols (MAG), bile salts and phospholipids that migrate to the epithelium cells where the compounds are taken up. Should the original food not contain lipid molecules that can ensemble hydrophobic domains large enough for solubilising the fat-soluble components, their bioaccessibility will be severely compromised (Mulrooney et al., 2021; C. Park et al., 2020).

The nature of these lipids also influences the shape and behaviour of the mixed micelles, which in turn affects their permeation through the mucus layer to reach the intestinal epithelium (Corte-Real & Bohn, 2018; Dima et al., 2020). The majority of the uptake mechanisms of fat-soluble are trans-cellular and vary according to the compound in question. Absorbed compounds generally avoid the first-pass metabolism as they are usually released into the lymphatic system transported in triacylglycerol- (TAG) and cholesterol-based chylomicrons and then travel to the systemic circulation or are sent there directly as part of high-density lipoproteins (HDL) (Reboul, 2013, 2017). Residual absorption in the colon requires further investigation, although some studies show that fat-soluble compounds can be absorbed there (there is a small amount of lipase) and also altered due to colonic microflora metabolism (Y. Tan & McClements, 2021). In any case, *in vitro* digestion models generally do not include the colonic phase.

To summarise, current knowledge indicates that fat-soluble compounds can be subjected to many different chemical, physical or flora-derived alterations which affect their final bioavailability. Some of these processes can even transform them into different compounds than the initially considered form. However, it is overall recognized that for fat-soluble compounds, bioaccessibility is the factor governing their absorption due to the number of factors that can detrimentally impact their

journey of release from the food matrix and incorporation into water-soluble and absorbable mixed micelles (Dima et al., 2020; D. J. McClements & Peng, 2020; P. Wu & Chen, 2021).

2.6.2 Food matrix effects to consider for improving bioaccessibility of fat-soluble compounds

This section will present a few factors being investigated as strategies to increase the bioaccessibility of fat-soluble compounds in general, prior to presenting works specifically addressing the compounds and formulations selected for this project (Figure 10).

Different fat-soluble compounds can share the same intestinal absorption path, which results in a lower bioavailability when supplemented in addition to other competitors (Goncalves et al., 2015). This should be considered in the formulation of multi-compound foods and meals. Excessive intake of fat-soluble compounds (such as vitamins A, E, D and K) has been known to be toxic to humans after numerous studies in the 1980s addressed the issue. This led to the establishment of Recommended Daily Allowances that can provide guidance on dosages in food formulations and meals (National Research Council (US) Committee on Diet and Health, 1990).

To address the issue of how to improve the bioaccessibility of fat-soluble compounds, it can be speculated that increasing the amount of lipids in which they can solubilise will increase the proportion of intact and soluble bioactives that will reach the intestinal mucus layer ready for absorption. However, the amount of lipid content in foods is not so correlated with bioaccessibility as is the type of lipid, and the type of its corresponding digestion products, as they can in turn affect digestion rate (Y. Tan & McClements, 2021). Oil content is mostly relevant in the cases in which there is not enough to solubilise the present compounds, additional content will not provide a benefit (Y. Tan et al., 2020). This has been exhaustively investigated during efforts to increase fat-soluble compounds absorption using encapsulation techniques (Wijekoon et al., 2023).

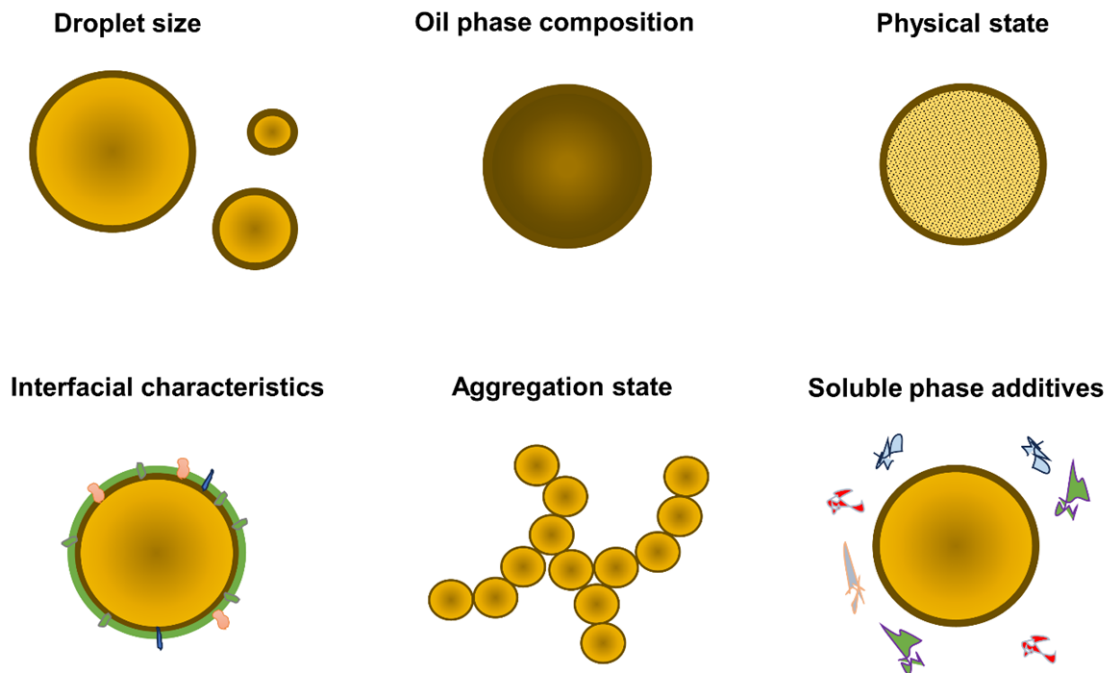


Figure 10. Properties of oil droplets known to affect lipid digestion and bioactive bioaccessibility.

In general, lipids with high melting points (longer and more saturated fatty acid chains) are known to reduce the rate of lipid digestion *in vitro*, which usually leads to reduction in bioaccessibility (Guo et al., 2017). On the other hand, long chains generally increase bioaccessibility due to increased solubilisation capacity of mixed micelles, and are often a requirement for carrying larger fat-soluble compounds effectively without them suffering transformation, such as the case of β -carotenes with monounsaturated long chain fatty acid and monoacylglycerols or vitamin E oil-in-water emulsions using either medium- or long-chain triglycerides (Guo et al., 2017; Helena de Abreu-Martins et al., 2020; Y. Yang et al., 2017). In addition, some types of oil present properties that can promote or enable transformation of the carried compounds, such as oxidation (Martin-Rubio et al., 2019).

As mentioned earlier, the oil droplet size in the small intestine heavily weighs in on the rate of its digestion, and this size can be different to the original food's size due to a variety of events such as fragmentation, flocculation or coalescence (Nguyen et al., 2019). This must be considered when designing the food

formulations as partial digestion should be insufficient to release 100% of the fat-soluble compounds.

Another factor to account for is the presence of compounds other than the lipids that form the oil carriers that influence the physicochemical properties of the interfacial layer of the matrix's droplets. In general, emulsifiers can cause coalescence or flocculation (i.e., protein coating) leading to reduced digestion and bioaccessibility as opposed to more stable carriers (such as pectin), and highly surface-active emulsifiers that block the adsorption of lipase have a similar effect (Y. Tan et al., 2020; Yao et al., 2018; R. Zhang et al., 2015).

Nevertheless, the aqueous phase must also be considered. Charged compounds such as polysaccharides can bind to oppositely charged droplets and inhibit lipase activity, bioaccessibility and so forth (Y. Li & McClements, 2014). Gastrointestinal fluids with high viscosity (presence of jellification agents for example) generally reduce digestion and bioaccessibility as well due to impaired mixing and diffusion (Palafox-Carlos et al., 2011). Oil droplets can receive protection from digestion due to the screening effect of other compounds (such as dietary fibres) present in the aqueous phase or that interact with the component involved in their digestion such as bile salts and enzymes (M. M.-L. Grundy et al., 2016). For instance, proteins, fibres, and minerals have been found to diminish the bioavailability of flavonoids in both *in vitro* and *in vivo* studies (Langston et al., 2023; Mishra et al., 2024; Núñez-Gómez et al., 2023). However, other studies present conflicting findings, and it is not possible to establish a consistent pattern regarding the impact of proteins or fibres on the availability of phytochemicals at this time (J. Zhang et al., 2023).

The aqueous phase components can also modulate the events of droplet aggregation, flocculation and coalescence (M. M. L. Grundy et al., 2018). In addition, even if the compounds are bioaccessible after the gastrointestinal journey, the aforementioned substances could generate absorption difficulties as a side effect (Samtiya et al., 2020).

Understanding the absorption of fat-soluble compounds necessitates comprehensive consideration of various factors forming complex interactions. The specific characteristics of the lipids, proteins, fibres, and phytochemicals

involved heavily influence this interplay and make it challenging to predict the final bioavailability of an ingested compound of interest.

The focal point of the overall study was to assess whether the input attributes of a food matrix could reliably predict the bioaccessibility of a specific compound, accounting for the diverse impacts of digestion. Given the potential for substantial optimisation due to the inherently low bioaccessibility of fat-soluble compounds, the subjects chosen for experimentation were vitamin E and curcuminoids.

2.6.3 Vitamin E

Vitamin E comprises a family of 8 different compounds (4 tocopherol isoforms and 4 tocotrienol isoforms). However, α -tocopherol (Figure 11) is the predominant isoform found in food matrices and body tissues (Ungurianu et al., 2021). In fact, consumption of alpha-tocopherol alone can reverse vitamin E deficiency symptoms; the recommended daily intake of vitamin E is 15 mg for people over the age of 14 (Institute of Medicine (US) Panel on Dietary Antioxidants and Related Compounds, 2000; Rizvi et al., 2014). Vitamin E is most abundant in nuts and seeds, and can be concentrated in vegetable oils, however it is also found in leafy vegetables and fortified cereals (Rizvi et al., 2014).

With an estimated $\log P \sim 12$, which compared to that of ascorbic acid ~ -2 , manifests the extremely high hydrophobicity of vitamin E (Niki & Abe, 2019; Saldanha & Vale, 2023). Due to its long hydrocarbon tail and lack of polar functional groups (Figure 11 . Chemical structure of vitamin E isoforms. α -tocopherol presents no insaturations on the non-cyclic tail and three methyl groups on the benzene chain in comparison to the other compounds of the family. Figure 11), providing vitamin E with lipid particles for solubilisation is a crucial requirement to ensure its bioaccessibility. Vitamin E is the most abundant liposoluble antioxidant compound found in human bodily tissues (Ungurianu et al., 2021). Other properties include a pK_a around 11 and a practical insolubility in water. These could explain why nanoemulsions with increasing vitamin E load were reported to experience higher surface tension and pH, although it could also be attributed to vitamin E's antioxidant activity limiting the formation of free fatty acids and lipid peroxidation (Teixeira et al., 2017).

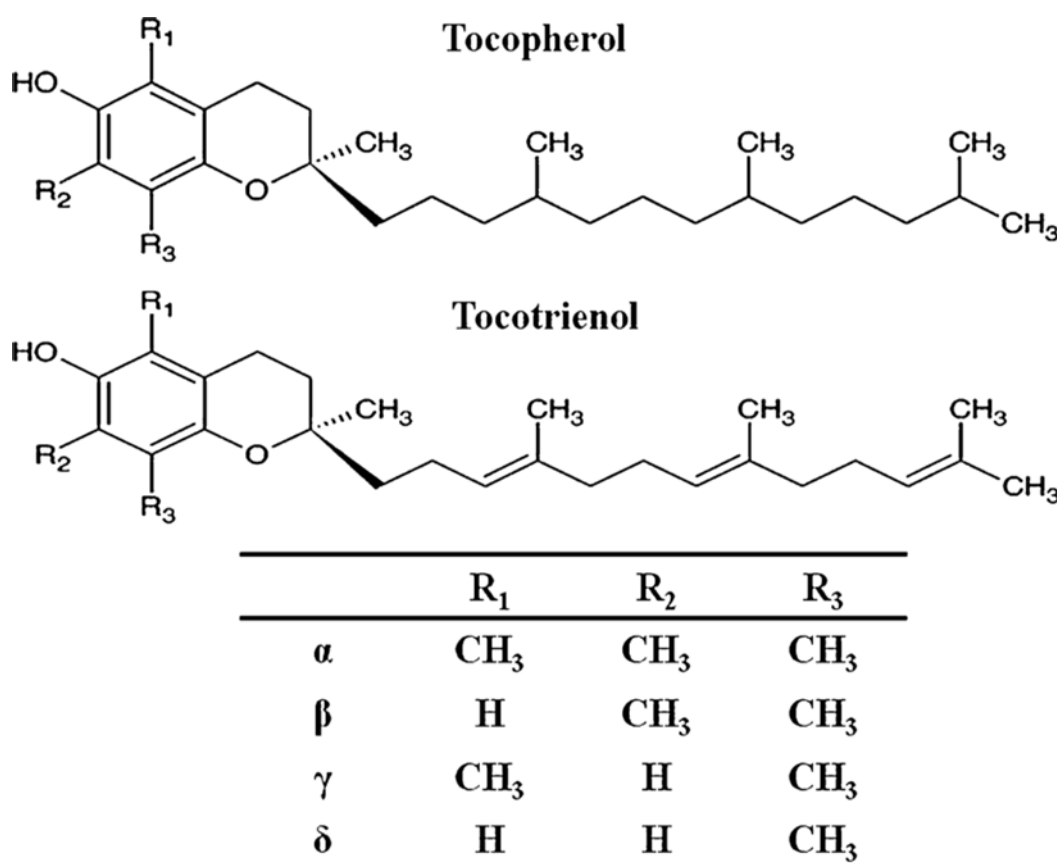


Figure 11 . Chemical structure of vitamin E isoforms. α -tocopherol presents no insaturations on the non-cyclic tail and three methyl groups on the benzene chain in comparison to the other compounds of the family.

Studies have also indicated vitamin E possesses antiproliferative, pro-apoptotic, anti-angiogenic, and anti-inflammatory properties. It may also be able to modulate signal transduction and gene expression in inflammation and immune system disorders (Q. Jiang, 2017).

Due to its antioxidant activities, vitamin E is believed to help prevent or delay chronic diseases associated to free radicals. It is capable of inhibiting cell proliferation and differentiation in smooth muscle cells, platelets, and monocytes, and reduces adherence of components to the interior of blood vessels whilst also dilating them, which minimises atherosclerosis (Institute of Medicine (US) Panel on Dietary Antioxidants and Related Compounds, 2000). These claimed health benefits are of interest for food producers and consumers (Szewczyk et al., 2021). However there are systematic reviews that conclude that the evidence behind health benefits of antioxidant supplements, including vitamin E, is inconclusive (Bjelakovic et al., 2015).

Although several studies of vitamin E bioaccessibility and bioavailability exist, they are not exhaustive with regard to probing the large variety of foods carrying it. The conclusions of previous studies indicate that vitamin E presents a wide range of bioaccessibilities depending on the food matrix in which it is found (10-80%, although it is more frequently in the 50-80% range), but no study has systematically screened for how changing the macronutrient composition affects the final bioaccessibility (Patrick Borel et al., 2013; Flory et al., 2019; Reboul, 2017; Y. Tan & McClements, 2021).

2.6.4 Curcuminoids

Turmeric (*Curcuma longa*) has a long tradition of being used for medicinal, culinary and cosmetic purposes, with first accounts of its use dating to almost 4000 years ago in Southeast Asia, where it was used to treat common eye infections, dental diseases, and digestive disorders such as acidity, indigestion, flatulence, ulcers, among other applications (Hatcher et al., 2008; S. Prasad & Aggarwal, 2011). Turmeric's properties have been mainly attributed to curcuminoids, which make for 1-6% of turmeric's weight. While over 50 curcuminoids have been identified in turmeric, the major ones found in commercial products are curcumin (60–90%), demethoxycurcumin (9–27%) and bis-demethoxycurcumin (0.3–15%) (Cas & Ghidoni, 2019; El-Saadony et al., 2022; Fu et al., 2015). As curcumin (Figure 12) is the major and predominant form, curcuminoids are usually simply referred to as curcumin.

Curcumin is commonly used in foodstuffs under the designation of food additive E100. The curcuminoids mix serves multiple purposes in the food industry, functioning as a flavouring substance, a preservative, and a natural colouring agent in a wide range of food and beverage products (Lan et al., 2023; Samanta et al., 2024; Seidi Damyeh et al., 2020). But current public interest in curcumin extends beyond its culinary purposes.

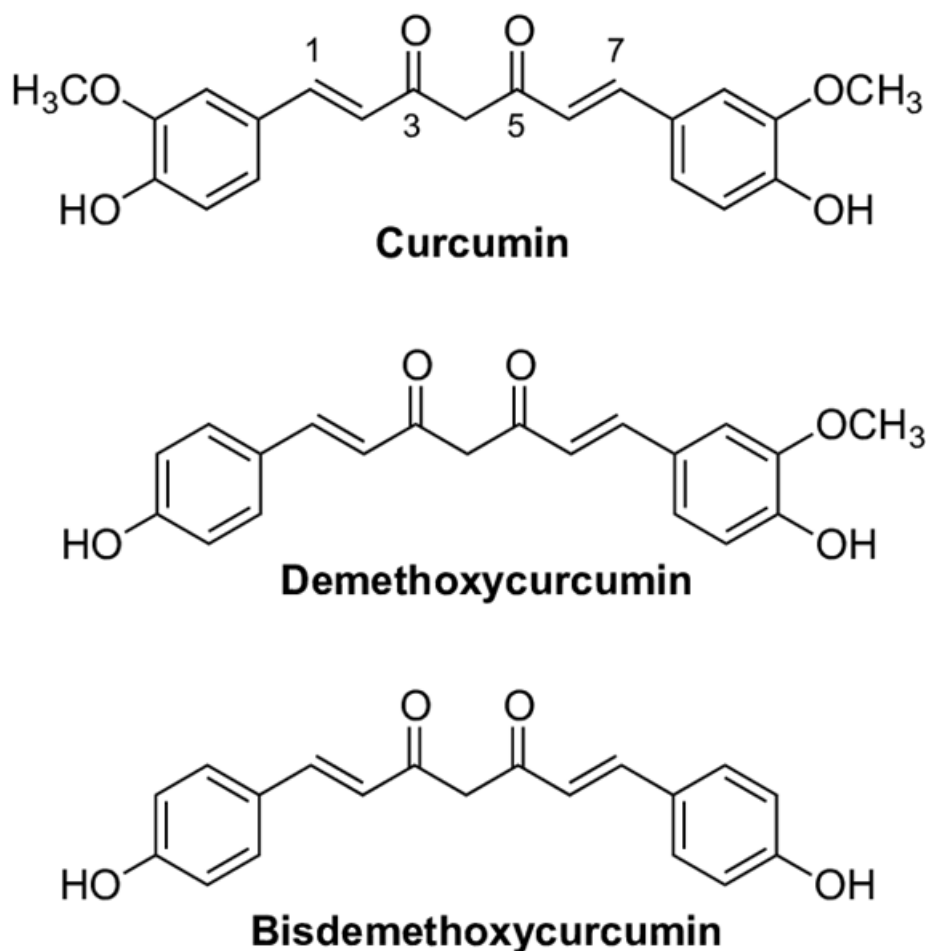


Figure 12. Chemical structure of curcuminoids. The loss of one (demethoxy-) or both (bisdemethoxy-) of the methoxy groups form the two other main curcuminoids.

Curcumin has been extensively studied (Micucci et al., 2022) for its biological activities in a large number of cell culture-based and animal studies. The potential of incorporating dietary curcumin as a feed additive was highlighted recently in a study where supplementing quails' diet with curcumin (at a concentration of 200 mg/kg) resulted in enhanced egg quality. This improvement was attributed to the promotion of lipid metabolism and the augmentation of antioxidant activity in the quails (Yong Liu et al., 2024). Curcuminoid's claimed health benefits include a plethora of antioxidant and anti-inflammatory related effects; antitumoral properties confirmed by several clinical I and II phase trials, inhibition of arthritis, hyperlipidaemia and cardiovascular issues; and decrease of muscle soreness and insulin resistance. Neuroprotective effects against depression, Alzheimer's and Parkinson's were also observed after its supplementation in animal trials,

although this effect could not be observed in humans, and was associated with their low bioavailability (Aggarwal & Harikumar, 2009; Goozee et al., 2016; Hewlings & Kalman, 2017; H.-T. Liu & Ho, 2018; Memarzia et al., 2021; Sayyar & Jafarizadeh-Malmiri, 2024; Tabanelli et al., 2021; Tomeh et al., 2019).

However, curcuminoids are also known to cause negative effects, especially when taken at large doses. The Acceptable Daily Intake (ADI) established by the European Food Safety Authority (EFSA) for curcumin is 0-3 mg/kg of body weight per day (EFSA Panel on Food Additives and Nutrient Sources added to Food (ANS), 2010). This might have helped prevent the over 40 reported cases of hepatitis allegedly linked to curcumin overconsumption between 2002 and 2021, in individuals with no prior history of liver disease (ANSES, 2022).

The inherently poor bioavailability of curcuminoids severely limits the exploitation of curcuminoids' benefits. Curcuminoids have low solubility in water of 0.6 µg/mL at room temperature (Kurien et al., 2007), with a $\log P \sim 3$, and are highly unstable before being absorbed, due to pH-related transformations, microbial metabolism and, most predominantly, autoxidation (Nelson et al., 2017; Tsuda, 2018). However, as curcumin has a lower hydrophobicity compared to vitamin E, it can still act as a surface-active compound and reduce interfacial tension in pure oil-water interfaces (Vardar et al., 2024). Three pK_a values are reported, one for each of the two phenolic hydroxy groups (between 8 and 10) and a third for the enolic proton (around 10) (Zhao et al., 2008). After being absorbed, almost no intact curcuminoids are detected in systemic circulation, and yet there is evidence that some of their associated positive health benefits may be in fact, due to the degradation products (Salem et al., 2014; C. Schneider et al., 2015; Tsuda, 2018). Many strategies towards improving their bioavailability have been developed in recent years, most of which focus on transport systems around the use of nanoparticles, optimised curcumin formulations (using synthetic polysorbates, fibres, oil emulsions and others), complexing/encapsulating or even through the use of adjuvants such as piperine, strategies that are shown to increase bioavailability from 2-fold to 100-fold (Hegde et al., 2023; Hewlings & Kalman, 2017; Yueyue Liu et al., 2023; Shamsi-Goushki et al., 2020; Tabanelli et al., 2021). As is the case for other fat-soluble compounds (discussed in 2.6.1, 2.6.2), improving bioaccessibility is the main focus of these approaches. As a

result, curcuminoid's bioaccessibility has been reported to vary from as low as 0.4-0.8%, to be improved to up to 25%, 50% and as high as 91% using nanoemulsions (Patrick Borel et al., 2023; Lu et al., 2019; Qazi et al., 2021; Ubeyitogullari & Ciftci, 2019).

Nevertheless, a strategy that has received less attention is administering the curcuminoids within a food formulation, essentially constituting a functional food, optimised for increased curcuminoid bioavailability.

To enhance the bioaccessibilities of vitamin E and curcuminoids, aiming for functional foods of superior nutritional quality, it is imperative to study the impact of the food matrix on bioaccessibility. For example, by screening formulations supplemented with dietary fibres and quantifying the impact of fibre content on bioaccessibility. To streamline the testing process and derive meaningful insights from the experiments, data modelling using mathematical algorithms, such as by Machine Learning, can be employed.

Highlights

- Fat-soluble compounds greatly depend on the presence of lipids to solubilise and remain bioaccessible.
- However, presence of lipids does not suffice, type of lipid also affects (size of micelles versus size of bioactive compound), among other factors.
- Vitamin E and curcuminoids are antioxidant compounds with a plethora of potential health benefits, but present low absorptions due to their hydrophilic nature.

2.7 Machine Learning: regression models

In the realm of food technology, the complexity and variability of raw materials, coupled with the intricate processes involved in food production, create a rich landscape of data. Traditional methods of analysis may fall short in handling this complexity, making it challenging to extract meaningful insights. Machine Learning (ML) steps in as a powerful tool, capable of learning patterns and relationships within datasets, thus contributing to enhanced decision-making processes. There has been a wealth of applications of ML in the food industry since the 2010s, ranging from safety practices such as toxicity prediction, inference of nutritional quality, estimation of the degree of food processing, and even assessment of the efficiency of absorption of orally-delivered drugs (Khan et al., 2022; Menichetti et al., 2023; Miyazawa et al., 2022; Tseng et al., 2023).

Although to the knowledge of the author there is no registered use of ML to predict nutrient absorption as a function of the food matrix, the current background regarding AI usage in food science and nutrition was indicative of its potential. Oral bioavailability/bioaccessibility of drugs has been the target of ML approaches before, but usually based on predicting absorption using as inputs the physicochemical properties of the drugs (such as chemical structure). These models are multi-compound instead of studying a single compound relative to numerous conditions. The expectations for accuracy are rather low ($r^2 < 0.5$) in this type of experiments (Fagerholm et al., 2021). Another example, using a publicly available drug dataset from the FDA, was arguably even less ambitious, from a performance perspective, as the prediction of whether a drug would have a positive, negative or no effect after consumption was deemed a success after reporting an accuracy of ~70% (Bennett-Lenane et al., 2022).

Therefore, these previous examples provide a reasonable assurance that the characteristics of the food matrix that most influence bioaccessibility can potentially be explored productively with ML techniques and put to use to optimise the production of functional foods.

2.7.1 Introduction to Machine Learning

Machine Learning is a branch of artificial intelligence that can be defined as enabling computers to make successful predictions using past experiences

(Baştanlar & Ozuysal, 2014). However, within this area of computer-based modelling, the challenges addressed are diverse, requiring the application of a broad range of algorithms. The responsibility of the ML practitioner lies in judiciously choosing the appropriate algorithm, emphasizing the avoidance of procedural errors over theoretical modelling performance. This underscores the importance of prioritising the prevention of operational mistakes to ensure the reliability of ML applications.

In ML, labels refer to identifiers that explain what the data is representing. Unsupervised learning refers to the extraction of insights from unlabelled data in an agnostic manner; the purpose is to understand the patterns in the data. Supervised learning has the added goal of not only understanding the patterns in the data but also using the acquired knowledge to accurately predict the label of samples/datapoints. A third type of ML, that has not been applied for this project and hence will only be mentioned here, is reinforcement learning, which attempts to find a sequence of actions that contribute to achieving a specific goal (Jovel & Greiner, 2021). Semi-supervised learning refers to tasks in which both labelled and unlabelled data are involved (Gibson et al., 2013). For the scientific aim of this project, supervised algorithms were employed to build a predictive model of bioaccessibility, and unsupervised algorithms were used to explore the patterns and other particularities of the food formulation dataset.

One last concept that is frequently used to categorise ML algorithms (supervised learning in particular), refers to the nature of the labels being modelled. Classification problems refer to the prediction of discrete class labels that can be either binary (i.e., healthy/disease or low/high) or multi-class (i.e., species prediction), and regression problems encompass the prediction of continuous-type labels such as a numeric value (i.e., house cost, rate of absorption) or an ordinal position (i.e., rankings such as most bioaccessible formulation). In theory, it is possible to convert any classification problem into a regression one and vice versa if the transformation of the target variable is appropriate (i.e., binning a continuous variable into discrete groups or providing a numeric value to a class according to suitable criteria), however in practice this results in a different analysis workflow that could be detrimental to the use case. It is widely recognised that turning a regression problem into a classification problem is

hardly ever beneficial, as there is an inherent loss of information in the process by binning the response, the proximity in the target space which can be valuable for modelling purposes is lost, binning obfuscates the prediction (which group metric of each bin is most valuable, is it the average or a specific quantile?), the binning inherently introduces biases in the model, and finally, for multi-class cases, the number of parameters will possibly have to increase (Fedorov et al., 2009).

Finally, an important consideration for the regression problem is whether to use a frequentist algorithm, or a probabilistic one. Frequentist methods assume the observed data is sampled from some distribution (i.e., most methods assume the data follows a normal distribution). This distribution represents the entire population of expectable observations. Considering a model θ that captures a specific set of patterns in the observed data, frequentist methods assume that the model's parameters are single-point fixed estimates of the most frequent pattern in the population, based only on the observed samples. Therefore, when calculating the likelihood of the data according to said model $L(\theta; Data) = P(Data|\theta)$, the model's parameters are treated as a fixed unknown quantity that must be found to maximise the likelihood. Only the sampling data is relevant for inference, there is no prior belief on what the model's parameters should be; their computation is agnostic and depends only on the data. On the other hand, Bayesian (probabilistic) methods assume probabilities for both the data and the parameters; θ is a random variable that now has its own prior distribution $P(\theta)$ in addition to the data's likelihood. The final model's parameters consist of probability distributions instead of single-point estimates (Bartolucci & Scrucca, 2010; Fornacon-Wood et al., 2022). The concept of likelihood will be further explained in the next section, which provides the general mathematical theory and equations behind regression.

2.7.2 Mathematical theory of regression

The two most common regression schemes used for parameter estimation are least squares (LS) and maximum likelihood estimation (MLE) (Jaqaman & Danuser, 2006).

The first clear and concise explanation of the least squares method was published by mathematician Adrien-Marie Legendre in 1805, although it had been used for several years by other mathematicians and astronomers such as Carl Friedrich Gauss. It is unclear who created the algorithm (Wolberg, 2010). The concepts of regression and correlation were later coined by Francis Galton in the 19th century, with the least squares method being its earliest form of expression (Stanton, 2001).

The least squares method (also called ordinary least squares as it is the first of a series of similar algorithms) is an optimisation method for solving regression problems. It is not only applicable to the linear regression case, but this is the simplest example to describe. Supposing the univariate linear regression equation:

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (2-1)$$

In which y is the dependent variable, which can also be called target variable; x is the independent variable that can explain the values y presents; and β_1 and β_0 , respectively, are the slope and intercept that model the linear relationship between x and y . In many cases, any random errors that result in datapoints not exactly conforming to $y = \beta_0 + \beta_1 x$ are attributed to a random variable ε that accounts for the failure of the model to fit the data exactly (Mahaboob et al., 2018). The least squares method solves the problem of finding β_0 and β_1 that minimise the error by minimising the sum of the squared residual errors (the hat symbol in $\hat{y}_i$ refers to the prediction of y_i) (Heiberger & Holland, 2004):

$$\text{minimize } \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 \quad (2-2)$$

which is done by equalling the partial derivatives of each parameter in the equation to 0:

$$\begin{aligned} \frac{\partial}{\partial \beta_0} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 &= \sum_{i=1}^n 2(y_i - (\beta_0 + \beta_1 x_i))(-1) = 0 \\ \frac{\partial}{\partial \beta_1} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 &= \sum_{i=1}^n 2(y_i - (\beta_0 + \beta_1 x_i))(-x_i) = 0 \end{aligned} \quad (2-3)$$

and solving simultaneously for the regression coefficients:

$$\widehat{\beta}_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \quad (2-4)$$

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

In which $\bar{y}$ and $\bar{x}$ are the average values of y and x respectively. $\widehat{\beta}_1$ and $\widehat{\beta}_0$ also have the property that:

$$\sum_{i=1}^n e_i = 0 \quad (2-5)$$

Ordinary least squares assumes the residuals are normally distributed $\varepsilon \sim N(0, \sigma^2)$, with a mean of 0 and a variance of σ^2 , it is under these specific assumptions that the least-square estimates become a special case of the more general maximum likelihood estimates network. This will be explained in more detail after describing MLE. Should the normality of the residuals not conform to normal distributions in either of these frameworks, these estimates could lead to erroneous inference (confidence intervals and p-values rely on the normality assumption), although in reality this does not have such a large impact on inference in large samples of at least 10 observations per predictor (Schmidt & Finan, 2018; Weisberg, 2014). An example of a fitted regression according to ordinary least squares can be found in Figure 13.

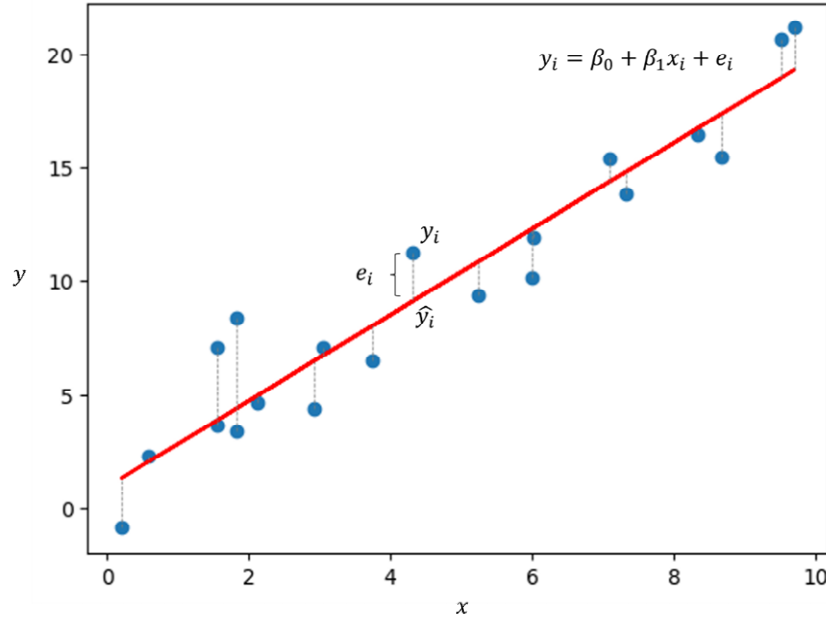


Figure 13. Scatterplot of the datapoints (x_i, y_i) and their linear regression line, computed via ordinary least squares.

In the frequentist approach, MLE can be credited to Ronald Fisher, father figure of modern statistical science, in the year 1912 (Aldrich, 1997). In the MLE scheme, as the name implies, parameters are inferred based on the values that maximise the probability of observing the data. In the frequentist approach a model's parameters are a fixed unknown value, such as the parameters $\widehat{\beta}_0$ and $\widehat{\beta}_1$ of the linear regression model. In the Bayesian approach, they are a random variable with its own probability distribution; for example, a normally distributed linear regression intercept $\widehat{\beta}_0 \sim N(\hat{\mu}, \hat{\sigma})$, with a centre inferred to be located at $\hat{\mu}$ and with a variability of a scale $\hat{\sigma}$, which represents uncertainty or variability in the parameter.

The probability of observing any datapoint y_i given the population parameter(s) θ can be written as $f(y_i; \theta)$ (S. Yang & De Angelis, 2013). Similarly, the probability density function (PDF) of any given value of y that is governed by a specified θ is written as $f(y; \theta)$. The θ parameter(s) are assumed to be fixed, and the joint probability of all values of y is the likelihood function; $L(\theta; y) = \prod_i f(y_i; \theta)$. In the case of linear regression, the θ parameters of interest are those present in the linear equation $y = \beta_0 + \beta_1 x + \varepsilon$, so $f(y; \theta)$ could be written as

$f(y; \beta_0, \beta_1, \sigma^2)$ as those are the parameters of study. The residuals are assumed to be normally distributed $\varepsilon \sim N(0, \sigma^2)$ as in least squares.

Before continuing with the mathematical derivation of the classical MLE, the difference with the Bayesian approach must be mentioned. In the Bayesian approach the $f(y; \theta)$ notation is avoided and $f(y|\theta)$ is preferred. This is used to indicate that θ is composed of random variables that condition the probability of the data; $f(y|\theta)$ is a probability based on the assumption that the probability distributions of θ are true. The Bayesian likelihood and frequentist likelihood share some similarities, but this section will be mostly covering the frequentist paradigm unless specified. Bayesian integrations acknowledge the random nature of the θ parameters instead of fixing their values for inference.

Supposing the observed data (y) comes from a normal probability distribution governed by the linear equation $y = \beta_0 + \beta_1 x + \varepsilon$ with the following probability density function (PDF):

$$f(y; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y - \beta_0 + \beta_1 x)^2}{2\sigma^2}} \quad (2-6)$$

In which σ^2 represents the population variance, the log likelihood $LL(\theta; y)$ is a monotonic transformation that simplifies the likelihood function whilst reaching the maximum at the same value of θ . The likelihood:

$$L(\theta; y) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \beta_0 + \beta_1 x_i)^2}{2\sigma^2}} \quad (2-7)$$

whereas the log likelihood:

$$\begin{aligned} LL(\theta; y) &= \prod_i -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(y_i - \beta_0 + \beta_1 x_i)^2}{2\sigma^2} \\ &= -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i (y_i - \beta_0 + \beta_1 x_i)^2 \end{aligned} \quad (2-8)$$

The objective of MLE is to maximise the likelihood function by identifying the parameter values of θ at which their first derivative (if possible) is 0, which is equivalent to minimising the sum of squared residuals in the least squares method. The optimisation is formulated as follows in the case of a single fixed

population parameter θ for which y is assumed to be normally distributed (in said cases, the population's mean μ would replace the predictions of the linear model $\hat{y}_i = \beta_0 + \beta_1 x_i$ as the expected value):

$$\frac{dLL(\theta; y)}{d\theta} = LL'(\theta; y) = \sum_i \frac{\frac{df(y_i; \theta)}{d\theta}}{f(y_i; \theta)} = 0 \quad (2-9)$$

However, the second derivatives (which will not be shown) must also be computed to ascertain that the identified point is in fact the maximum, as equalling the first derivative to 0 locates critical points at which there is an inflection in the curves but does not distinguish whether it is a maximum or a minimum. For multiparameter cases, the partial derivatives of $LL(\theta; y)$ with respect to each parameter will be used. This approach presents problems when the log likelihood of a parameter is flat or presents several local maximums, as the initial point of the parameter θ will lead to different conclusions. In most cases, these equations can be solved analytically. For more complex cases, numerical methods that iteratively search for the θ values that solve the equality(s) are used, such as Newton's method.

Newton's method (conceived by the renowned physicist Isaac Newton in collaboration with Joseph Ranson) starts with an initial guess of θ_0 , and iteratively performs the following (Galántai, 2000):

$$\theta_1 = \theta_0 - \frac{LL'(\theta_0)}{LL''(\theta_0)} \quad (2-10)$$

Until a sufficiently precise θ_n is obtained.

In the case of linear regression this computation is much more simple in spite of being a multiparameter θ and does not require this type of numerical methods; the partial derivatives $\frac{\partial LL(\theta; y)}{\partial \beta_1}$ and $\frac{\partial LL(\theta; y)}{\partial \beta_0}$ also lead to the same equations for $\widehat{\beta}_1$ and $\widehat{\beta}_0$, respectively, as in the least squares case (the denominator -2 from the right side of the equation was passed as enumerator to the left side):

$$\frac{\partial(-2LL(\theta; y))}{\partial \beta_0} = -2 \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i)) = 0 \quad (2-11)$$

$$\frac{\partial(-2LL(\theta; y))}{\partial\beta_1} = -2 \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))(x_i) = 0$$

Which are the same partial derivatives as in the case of least squares estimators presented earlier, leading to the same derivations of the predicted parameters of $\widehat{\beta}_1$ and $\widehat{\beta}_0$.

Now that the general mathematical theory behind regression has been explained and exemplified with the case of simple linear regression, how to use ML techniques to infer these types of models without making operational mistakes must be considered.

2.7.3 Bias-variance tradeoff

This is a fundamental concept in supervised learning. It describes the need to strike a balance between these two sources of errors so that a model possesses the ability to perform well on unseen data (generalization capacity) (Dexter et al., 2020). There are three main sources of errors regarding generalization; first, bias errors are those resulting from learning erroneous assumptions from the training data that result in erroneous predictions, second are variance errors that occur from over-sensitivity during training, and third is noise (Rocks & Mehta, 2022). Noise is error that stems from the particularities of the dataset rather than from the modelling pipeline, however it can increase bias and/or variance if not processed correctly.

Which observations to use for quantifying these errors is crucial; models must be trained on a separate set of data than those used to evaluate them. Models that learn mostly from the noise of the dataset can potentially present a very small bias and variance if evaluated on the training dataset they have already seen, but present large errors when asked to predict previously unseen data. This is known in statistics and ML as a type of data leakage; the improper use of training data as test data (another unfortunately commonplace type of data leakage is injecting information of the target variable into the predictors inappropriately) (Götz-Hahn et al., 2022).

Overfitting and underfitting are descriptors commonly used for models that are not striking an effective balance between bias and variance (Montesinos López et al., 2022). Overfitted models refer to those that have a reduced bias due to learning the patterns in the training data too well, to the point that they are capturing noise and therefore exhibiting increased variance errors as they attribute a great importance to spurious information, which can result in the prediction of different outcomes for similar observations (the oversensitivity previously mentioned, Figure 14). An underfit model generally refers to bias errors; models that are too simple to capture the patterns in the data. If this occurs, increasing the accuracy is generally the first concern prior to addressing sources of variance.

It must be noted that the bias-variance trade-off was traditionally believed to be solved with intermediate model complexities (Figure 14), however modern Deep Learning methods can nowadays fit overparameterised models that perfectly fit the training the data. This phenomenon is still under investigation; current research has still to reach a consensus on whether or not it is true (Belkin et al., 2019; Rocks & Mehta, 2022), so that the bias-variance trade-off consideration can be conclusively written off for modern ML approaches.

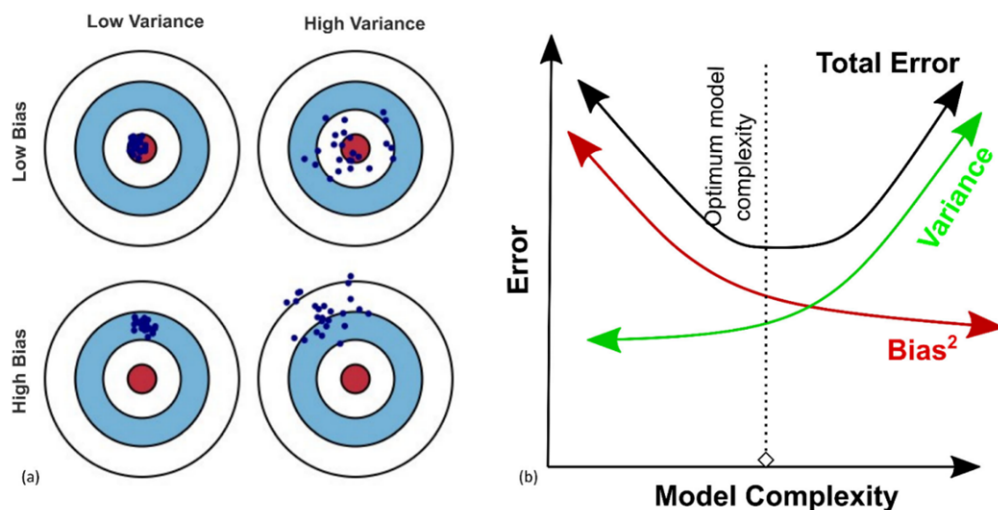


Figure 14. (a) Graphical illustration of bias and variance. (b) Contribution of bias and variance to the total error of the ML model. Increasing model complexities usually lead to lower bias and increased variance. [Obtained from Mandal et al., 2022]

2.7.4 Sampling strategies in ML projects

The issue of using the appropriate samples at each step of the modelling process was introduced in 2.7.3. Samples that are used to train the model must not be used to evaluate it; the model already learnt how to predict them well so the estimation of the model's performance (its generalization capacity) will be overestimated. The dataset is split into training and test sets (hold-out strategy), in which the training data is used to trial several data preprocessing and modelling pipelines, which are later evaluated over the unseen (blind) test set. However, within that training set, it is also usual to do an additional type of splits: iterative training and validation splits under a cross-validation experiment design. This allows to robustly estimate the predictive power of a fitted model before testing it with purely blind data. After all, validation splits are also seen during the process of model development, as data preprocessing strategies are typically trialled over the whole training set. The precise cross-validation strategy to use (k-fold, shuffle-split, leave-one-out, and others) depends on the nature of the dataset, and the proportions of the training:validation:test split also differ accordingly (example splits are 50:25:25 or 80:10:10) (Vabalas et al., 2019; Xu & Goodacre, 2018). The strategies mentioned rely on sampling without replacement, meaning each time an item is selected from a dataset it is not returned to the sample pool, as opposed to bootstrap methods in which samples are replaced. Although bootstrap methods can be used to evaluate model performance, they are typically applied in other contexts. Nevertheless, due to the possibility of reusing samples, they can potentially be a better choice for small sample sizes, even though assessing performance on a per-unique-sample basis becomes more complicated and sample recycling can lead to overly optimistic evaluations. However, with large datasets, cross-validation is generally constricted by computational issues and the hold-out strategy is usually sufficient, as the performance estimations from several iterations of cross-validation experiments using large sample sizes will not be so susceptible to inter-groups sample variability (Tougui et al., 2021; Vabalas et al., 2019).

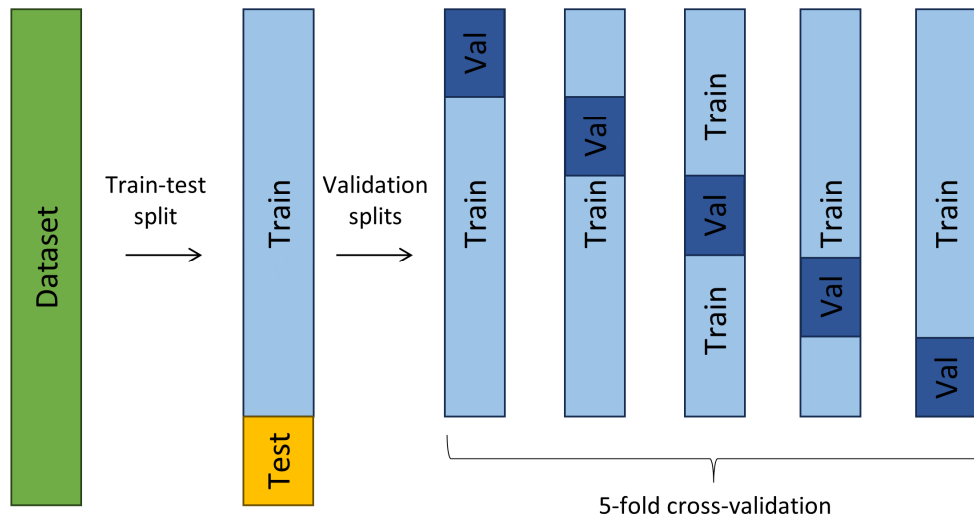


Figure 15. Example of a training:validation:test split with internal 5-fold cross-validation.

2.7.5 Frequentist approaches

The frequentist approach to regression offers many opportunities to develop highly efficient algorithms to find the fixed parameters that maximise model performance, as the problem is more constrained than in the Bayesian case. Frequentist models are only concerned with the information coming from the sampling data and the likelihood function, meanwhile the Bayesian approach must also account for different prior-likelihood configurations (Bartolucci & Scrucca, 2010). Throughout the years, mathematical developments, coupled with the improvement of computer hardware and increasing accessibility of software engineering has fostered the creation of exceptional algorithms for frequentist inference; the ones most relevant to the project explained in the following pages.

2.7.5.1 Support Vector Machines (SVM)

This family of models earns its name from mapping datapoints in space and using feature vectors to create decision boundaries between label pairs (S. Huang et al., 2018). The objective of training an SVM model is to find the weight vector and

bias for a set of input feature vectors (Figure 16) so that the hyperplane separates and maximises the margin between different classes.

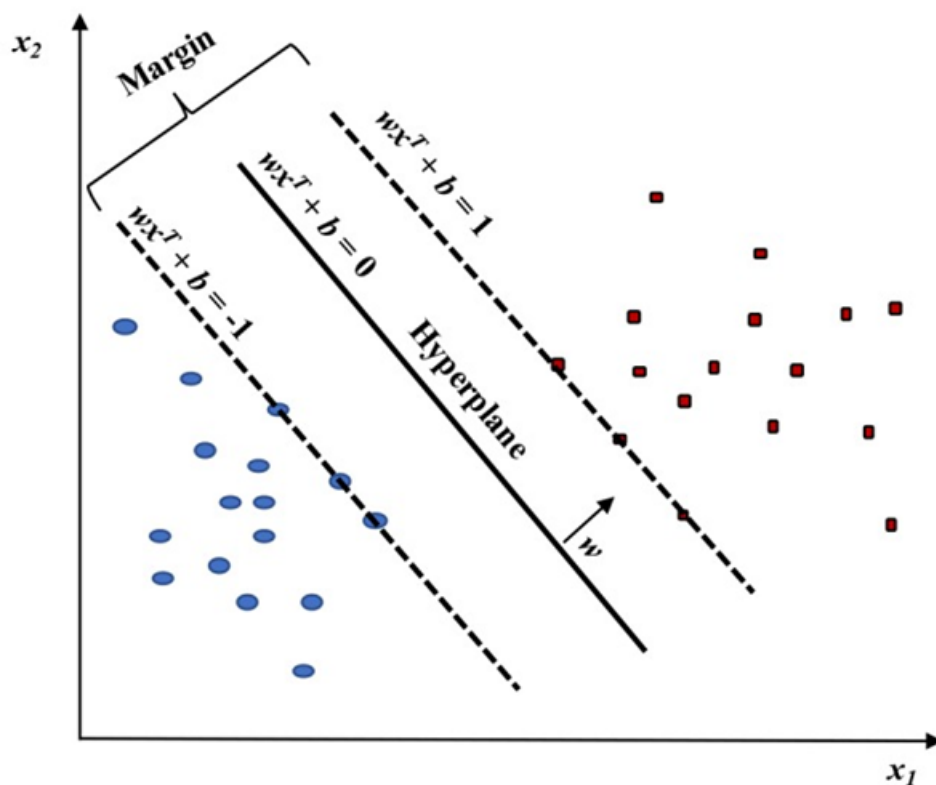


Figure 16. Linear SVM model. Two classes (red versus blue) are classified. [Obtained from Huang et al., 2018] w : weight vector, x : input feature vector, b : bias

Support vector machines were proposed by Cortes and Vapnik (Cortes & Vapnik, 1995) for two-group classification problems. The original idea by Vapnik in 1963 implemented this as a linear classifier exclusively (Vapnik, 1963), although it was the introduction of the kernel method (also called kernel trick), which enables the conversion of a non-linear problem into a linear one in a higher dimensional space, that greatly enhanced its applicability and made it an extremely popular algorithm (the 1995 paper has 33K citations on SpringerLink as of June 2024).

Support vector regression (SVR) is an extension of SVM proposed a year later by Vapnik and colleagues (Drucker et al., 1996) that attempts to solve a similar primal optimisation problem (Smola & Schölkopf, 2004):

$$\text{minimize } \frac{1}{2} \|w\|^2 \quad (2-12)$$

Subject to:

$$|y_i - \langle w, x_i \rangle - b| \leq \varepsilon \quad (2-13)$$

Where x_i is a training sample with target value y_i , and $\langle w, x_i \rangle + b$ is the prediction of that sample; all sample predictions must be in range ε of the true y_i observations.

What kernel to use can only be figured out through rigorous validation trials, as some kernels might be better than others depending on the nature of the data (S. Huang et al., 2018). The most popular kernels are the already mentioned linear kernel, the polynomial kernel, and the radial basis function kernel (Rahimi & Recht, 2007), which can generate an almost infinite but computable random feature space in which sample relative distances to a fixed point (mostly Euclidean) are linearly manipulated to generate the decision boundaries (Nisbet et al., 2018).

2.7.5.2 Decision trees and their ensembles

The concept of decision trees in the sense that by an arbitrary number of binary choices it is possible to infer a conclusion from a set of previously known insights has a long history, although its statistical antecedents can be traced to Fisher's paper in which he introduced linear discriminant analysis (Fisher, 1936). Loh (Loh, 2014) attributes the first regression tree algorithm to Morgan & Sonquist around 60 years ago (Morgan & Sonquist, 1963), which was applied to predict annual earnings from survey data composed of race, age, education, and profession information, although Belson developed a similar "decision tree" approach 4 years earlier (Belson, 1959).

Automatic Interaction Detection (AID) was the name of the algorithm presented by Morgan & Sonquist; it recursively splits the data into nodes minimising the 'impurity' (sum of squared deviations) until the reduction is less than a preset fraction of the impurity at the root node, and reports the node mean as the prediction. In the following years, several researchers manifested key limitations of AID, such as the possibility of severely overfitting the data, ignoring inherent

sampling variability and poor handling of highly correlated features (Loh, 2014), which prompted the development of novel algorithms to counter these issues. Many algorithms have been developed since then (CRUISE, C4.5, FACT, QUEST, GUIDE, ID3, among others) (Jiao et al., 2020), noteworthy among them being the Classification and Regression Trees (CART) algorithm.

CART (Breiman et al., 1984) follows the same approach as AID but includes several crucial improvements. It replaces stopping rules for building a large tree that is pruned (weakest links according to cost-complexity analysis) to minimise the cross-validated error estimate, a strategy that reduces the possibility of over- and under-fitting. It also solves the highly correlated variables issue by introducing a series of alternate splits that are used to compute an importance score of each variable that are factored into the optimisation procedure. An interesting characteristic of the algorithm is that it can also employ splits from a linear combination of variables to perform the stochastic search (Breiman et al., 1984; Loh, 2014), thereby accounting for possible linear relationships within the dataset's features. An example regression decision tree can be observed on the next page.

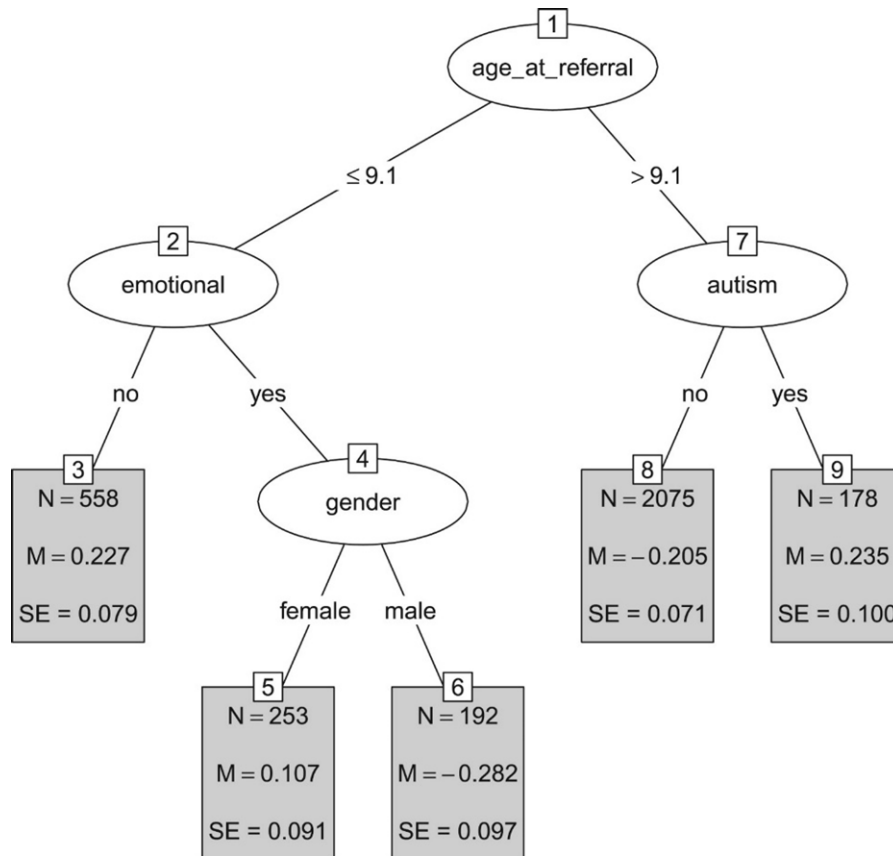


Figure 17. Decision tree for the unadjusted treatment outcome of mental-health issues in young UK patients. Higher values represent poorer treatment outcomes. [Obtained from Marjolein Fokkema & Wolpert, 2021] Panels depict subgroup sizes (N) and estimated means (M) with standard errors (SE). SEs account for random variability between treatment centres, not for the searching of the tree structure.

Breiman extended once again the decision tree concept to develop what is still considered a state-of-the-art algorithm in the ML domain; Random Forest, alongside other ensembles of decision trees (Breiman, 2001).

This algorithm, as its name implies, uses randomization to create a large number of decision trees, the predictions of which are aggregated either by majority voting (classification) or by averaging (regression). The decision trees are created on different subsets of data sampled with replacement (samples may appear more than once in the subset); a technique called bagging (**bootstrap aggregating**). The variables are also randomised; for each node only a certain number of predictors p are provided (typically the result of rounding $\sqrt{p}$). Different combinations of variables and variable thresholds are tested until the best splits

in the data are identified, and this tree growing process is repeated hundreds of times (Couronné et al., 2018; Rigatti, 2017).

Another group of ensembles of decision trees are boosting learners, the first (and still used) example of which is AdaBoost (Gilks et al., 1995a), formulated prior to the publication of the Random Forest algorithm. Rather than aggregating learners by voting, this algorithm uses the insights from 'weak learners' such as small decision trees that are referred to as stumps (can be models other than decision trees however) as they only use one variable but are supposed to still perform better than random chance. Sequentially, the performance of the overall model is improved by assigning heavier weights to the least performing stumps so that the optimisation procedure prioritises them (Banfield et al., 2007). This results in an accuracy-oriented algorithm that prioritises reducing the bias rather than the variance, in opposition to bagging approaches (Kotsiantis, 2014).

On a final note, rather than minimising the sum of squared deviations as AID did, decision tree algorithms typically optimise the decrease in either entropy or the Gini index of the training dataset in the case of classification (Kingsford & Salzberg, 2008), and the mean residual sum of squares for the regression problem (Banerjee et al., 2019). Other significant decision tree algorithms and their ensembles will not be covered here to avoid being exhaustive.

2.7.5.3 K-nearest neighbours (KNN)

This methodology in ML is similar to SVMs in the sense that it uses the feature space and the computation of distances (usually Euclidean) to train supervised models (Z. Zhang, 2016). However, the key distinction lies in the approach to decision boundaries: KNN relies on the local characteristics of data points and their proximity (neighbourhood) (Figure 18), while SVMs aim to find a hyperplane that maximally separates different classes in the feature space. It was first developed in 1951 by Fix and Hodges (Freund & Schapire, 1997), and expanded upon by Cover (Cover & Hart, 1967).

A key parameter in the KNN algorithm is the **k**. This integer corresponds to the number of neighbours to include in each neighbourhood. Larger k values reduce variance due to random error or noise but run the risk of underfitting due to less distinct boundaries (Z. Zhang, 2016). Striking a balance between overfitting and

underfitting in this type of algorithm greatly depends on the choice of k , and several methods exist to determine the optimal value for a specific dataset, such as the elbow method (Shi et al., 2021).

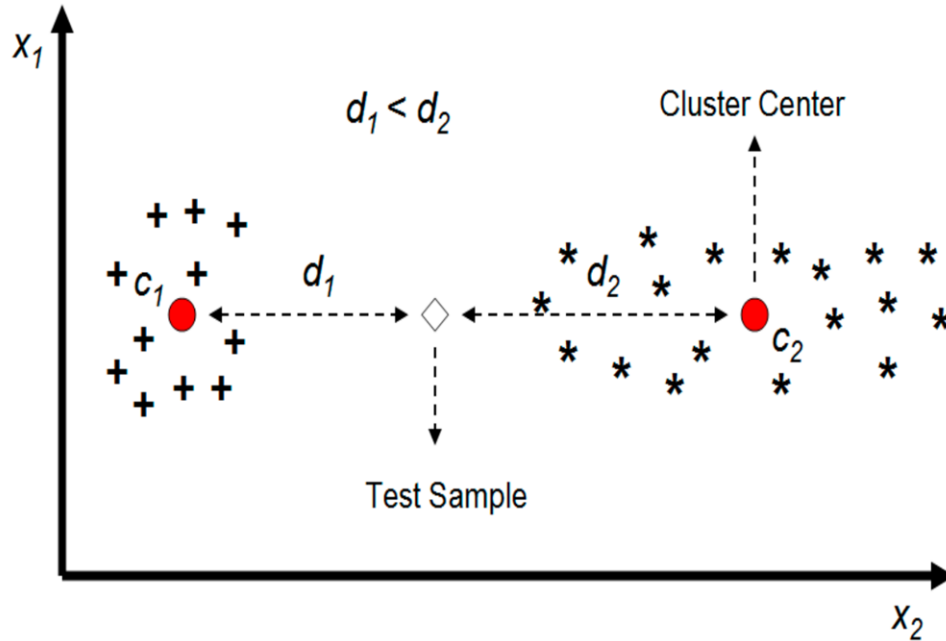


Figure 18. In this example, the pretrained KNN model classifies the test sample as belonging to the c_1 cluster as its distance to that centre is smaller than the distance to c_2 cluster centre. [Obtained from Saadatfar et al., 2020]

2.7.5.4 Regularization in linear regression

The goal of regularization is to prevent estimators from developing overly complex models. More specifically, regularization is a method to prevent overfitting by adding a penalty term to the objective function that the algorithm minimises. The most common types of regularization are L1 regularization (Lasso) and L2 regularization (Ridge). They are mostly applied to linear regression models but the regularization concept has been applied to other types of algorithms as well (Lemhadri et al., 2021).

Ridge regression, also known as Tikhonov regularization or Tikhonov-Phillips regularization after the original proposers, although it was Hoerl who coined the term Ridge regression and applied it to the finite-dimensional space (Hoerl & Kennard, 1970). It was developed to solve the problem of multicollinearity (one predictor can be perfectly predicted from others) in ordinary least square estimators, the standard approach to inferring linear regression coefficients.

Ridge regression shrinks the coefficients of correlated predictors equally towards zero, but does not make them vanish, resulting in so-called dense models. It is an ideal solution for datasets in which there are many predictors each having small effects (Hastie, 2020; Ogutu et al., 2012). For a given linear equation (θ model parameters notation is replaced by β for regression coefficients):

$$\mathbf{y} = \mu \mathbf{1}_n + \mathbf{X}\beta + \varepsilon \quad (2-14)$$

Where $\mathbf{y}$ is the target vector, $\mathbf{1}_n$ is a column vector of n ones and μ is a common intercept, $\mathbf{X}$ is a $n \times p$ matrix of predictors; β is the vector of the regression coefficients of the predictors and ε is the vector of the residual errors, we add the ℓ_2 -norm penalisation term:

$$\|\beta\|_2^2 = \sum_{j=1}^p \beta_j^2 \quad (2-15)$$

And the regression estimator's task is to find the model coefficients β that minimise:

$$\sum_{i=1}^n (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^p \beta_j^2 \quad (2-16)$$

Where $x_i^T \lambda$ is the i -th row of $\mathbf{X}$ is the tuning parameter which regulates the strength of the penalty and must be higher or equal to 0. It is dependent on the data but can be optimised (selected appropriately) by cross-validation experiments by the user. The intercept is ignored by centering $\mathbf{X}$ and $\mathbf{y}$ upfront.

In contrast, in Lasso regression, the ℓ_1 penalty is:

$$\|\beta\|_1 = \sum_{j=1}^p |\beta_j| \quad (2-17)$$

Which obtains a sparse solution (it assumes many coefficients to be close to zero) that makes it an appealing variable selection method, in a similar vein as CART trees. It is an efficient and fast algorithm suitable for massive datasets, but it presents several downsides. In cases of multicollinearity, it will arbitrarily pick out one of the highly correlated variables and ignore the others, and it cannot select

more variables than samples. Overall, it lacks the oracle property that states that a procedure can estimate the variables with zero-coefficient value with probability asymptotically tending to 1 consistently (its inferring mechanism contains inherent biases such as the above described that can lead to erroneous estimations) (Ogutur et al., 2012). In other words, it is not the perfect variable selection algorithm from the statistical point of view (adaptive Lasso claims to resemble more closely an oracle procedure) (H. Zou, 2006) but rather a useful device in the ML practitioner's toolbox and a valuable concept to be aware of.

2.7.6 Bayesian approach

To begin with, the quantification of the evidence of two different statistical models is a critical concept that allows to understand the fundamental differences between the Bayesian methodology and the frequentist methods previously described. Hypothesis testing is ingrained in the development of mathematical models. The statistical significance of regression parameters (whether they are 0 or not) should be addressed for every trained model if possible. This assessment is performed differently in each paradigm.

Classical inference is based on extrapolating population parameters (never observed) from a particular sample (observed data), contrary to Bayesian inference. In classical inference, the interpretation of the p-values of null-hypothesis tests is usually oversimplified. P-values provide long-run probability statements over what would be expectedly observed if we sampled repeatedly from a population. If a comparison test between two conditions indicated that their mean difference had a two-tailed p-value of 0.01, the correct interpretation is that the probability of observing a difference equal or larger than the currently observed is 1%, assuming the null hypothesis be true. The 95% confidence interval for the difference between two treatments is interpreted as: 95% of sampling experiments will construct an interval containing the population difference. These methods describe the performance of the method over many samples rather than the probability that the parameter is within the confidence interval (Baldwin & Larson, 2017). This raises an important issue as scientists normally interpret hypothesis testing results as the evidence of a particular

hypothesis vs another; p-values are probabilities that assume the null hypothesis is true and are not meant to be interpreted in this manner (Cohen, 1994).

Bayes factors are the Bayesian's non-binary and more precise alternative to the task of comparing datasets or models. The ratio of two competing statistical models, based on the magnitude of the evidence that supports them, enables a clear interpretation of which model is more likely. Similarly, Bayesian credible intervals embody the common interpretation of frequentist confidence intervals: for the case of 95% credible intervals, there is a 95% probability that the true (unknown estimate) would lie within the interval, given the evidence provided by the observed data (Hespanhol et al., 2019). The unobserved data is not considered; thereby avoiding the potential for poor analysis being conducted with the purpose of catering to peer-reviewed journals privileging statistically significant results (Rosenthal, 1994).

To summarise, despite the potential for Bayesian hypothesis comparison to yield erroneous conclusions with a small and unrepresentative sample of a population, it refrains from drawing conclusions on unobserved data. Its findings can be accepted at face value, given the presented data and a justifiable belief in the modelled system. In contrast, classical models may lead to presumptuous conclusions (declaring a model is or is not statistically significant) due to numerous inherent assumptions if not used carefully. When dealing with limited data, the explicit incorporation of prior knowledge in Bayesian models enhances their value as a framework for quantifying uncertainty as a measurable variable. This approach offers an alternative to relying on statistical significance tests that are highly dependent on sample size.

2.7.6.1 Bayes' theorem

Bayesian modelling tackles parameters for which the modeller has some previous knowledge about (prior probability distributions). These prior beliefs are updated with the observations (the likelihood of the prior beliefs being true given the data) to produce the final probabilities of said parameter values' (posterior probability distributions).

The mathematical equation of the Bayes' theorem states (Westbury, 2010):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2-18)$$

$P(A|B)$ and $P(B|A)$ are the conditional probability of event A occurring should B, and vice versa. $P(A)$ and $P(B)$ are the probabilities of observing events A and B, respectively.

Replacing the generic A and B events to explain this equation more intuitively in the context of modelling parameters using beliefs and observations leads to:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \quad (2-19)$$

Where X represents the observations and θ represents the models' parameters. The left side of the equation is the posterior probability of observing the model's parameters θ having observed the data. The prior probability $P(\theta)$ is predefined by the modeller, and $P(X)$ is the marginal likelihood, the probability of the data. The marginal likelihood can simply be viewed as the normalizing factor needed so that the sum of the posterior probabilities equal 1, making the posterior a valid probability distribution (T. S. Lee & Mumford, 2003), hence it is normally ignored when discussing said theorem. Hence the Bayes' theorem essentially equals:

$$P(\text{posterior}) = P(\text{likelihood})P(\text{prior}) \quad (2-20)$$

The choice of prior has a major impact on the final model, and many have manifested concerns over the subjective nature of this choice, however Bayesians downplay these, arguing that many of the decisions in research are subjective (measures, participants, questions, design). All statistical methods make assumptions over the data that are not objective, the choice of likelihood (is the data normal, binary, count, highly skewed?) being far more important than the choice of prior (Baldwin & Larson, 2017). An advantage of priors is the possibility to inject previous domain knowledge into the modelling process, which can be used to increase the model's performance or to validate whether prior beliefs hold true when confronted with new observations (Wagenmakers et al., 2018). As Wagenmakers states; *any statistical paradigm that cannot incorporate such [prior] knowledge seems overly restrictive and incomplete. The founding*

fathers of classical inference –including “Student” and Fisher– mentioned explicitly that their methods apply only in the absence of any prior knowledge.

The choice of prior is the next item to consider. Some authors suggest that using anything other than non-informative priors (those that do not convey any specific prior beliefs, such as a uniform distribution in which all parameter values have the same probability) is incompatible with the standards of research integrity (García-Pérez, 2019). Non-informative priors result in models similar to those that frequentist methods would obtain as only the likelihood would play a role in inference. The argument is that population parameters are not random variables (there is a true value that can be estimated), however this true value has also been discussed during this review as not being appropriately handled in the frequentist paradigm either.

Opposing this idea is that the capability to use completely uninformative priors arguably does not exist, hence any choice of prior (even non-informative ones) is subjective (i.e., the limits of the uniform distribution are already a subjective choice). Non-informative priors are important if they reflect the true state of current knowledge or to compare them as a baseline to more informative priors; the data drives the model as happens in frequentist models. Experts in the field generally recommend the use of weakly informative priors (i.e., wide Gaussian distributions) based on common sense and general domain knowledge so that the final model reflects the plausible parameter space (Van Dongen, 2006). In this manner the models do not explore configurations that have no basis in reality.

Regarding strongly informative priors, unless the domain knowledge is very reliable, the main objection to their use is that they may mismatch the observed data and produce a prior-data conflict. Notwithstanding, the benefits of their use are highly significant; they can reduce the required sample size for developing robust models as they incorporate previous knowledge in the form of beliefs rather than observations (Mutsvari et al., 2016).

Ideally, the scientific community employing Bayesian modelling should properly explain their choice of prior (even the non-informative priors) and provide their data so that other scientists with different opinions can trial their own prior beliefs (Van Dongen, 2006).

2.7.6.2 Markov Chain Monte Carlo

Conjugate distributions are those in which the prior and posterior probabilities come from the same distribution family. In such cases, the prior is called a conjugate prior for the likelihood function (Hackenberger, 2019). Examples of conjugate distributions are a Bernoulli likelihood and a beta prior, or normal likelihoods and priors. Prior to the advent of computers, conjugate priors were convenient to provide an analytically tractable solution, derived as a closed-form expression, for the posterior distribution (Fink, 1997). In spite of this, the computation of conjugate prior solutions is not always feasible (Robert, 2007). For cases such as these, and those for which it is not possible to derive a closed-form solution, numerical methods are required to search the parameter space for the optimal solution, such as Markov Chain Monte Carlo (MCMC) methods (K. Choi, 2023).

The first iteration of the Monte Carlo method was published in 1949 by Ulam and Metropolis (Metropolis & Ulam, 1949). Ulam conceptualized the idea after attempting exhaustive combinatorial calculations of the chances that a particular solitaire game would be successful and realizing computers could perform this search (Andrieu et al., 2003). Metropolis would later create one of the most popular MCMC algorithms that was baptised with his name (Metropolis et al., 1953). Hastings would later improve upon this algorithm by generalising it for non-symmetrical proposal distributions (Hastings, 1970).

The principle of Monte Carlo methods can be understood as *Anything we want to know about a random variable θ , we can learn by repeated sampling from the probability density function of θ* (Jackman, 2004). Monte Carlo methods propose a simulation framework in which the mathematical random processes are called Markov chains. These processes draw thousands of samples from the posterior (given a prior and likelihood distributions) until it has been characterised. Chains iterate over different plausible parameter values which usually get updated (if a chain gets stuck there may be a problem with the models' initial configuration) and the range of calculated values are used to reach a consensus of the parameter's probability density function; the chain converges on the posterior distribution (Baldwin & Larson, 2017). Several chains are computed, and the conclusions of each (do they agree?) serve as a measure of the robustness of

the model (A. Tan et al., 2015). Detailing the manner in which MCMC algorithms determine what samples to draw and how to update the chains is out of the scope of this review.

2.7.6.3 Bayesian hierarchical linear regression and its MCMC inference

The Bayesian methodology can be used to estimate the parameters of linear regression. Furthermore, these parameters can be inferred over multi-level datasets, in which samples belong to more than one level or group.

In this regard, a predictor x is considered pooled when there is a single parameter θ across different groups of data x_j . If there is a unique parameter θ_j for each group of x_j datapoints, then the predictor is referred to as non-pooled, as the datapoints of different groups were not pooled together to infer a single θ applicable to all datapoints. Bayesian hierarchical models are composed of a set of sub-models, one for each group, in which the datapoints of a specific parameter θ can be either pooled across all available datapoints to infer it (indicated by not including a group subindex in θ), or each j group is assumed to present a different parameter θ_j , which is inferred using only the x_j datapoints of that group.

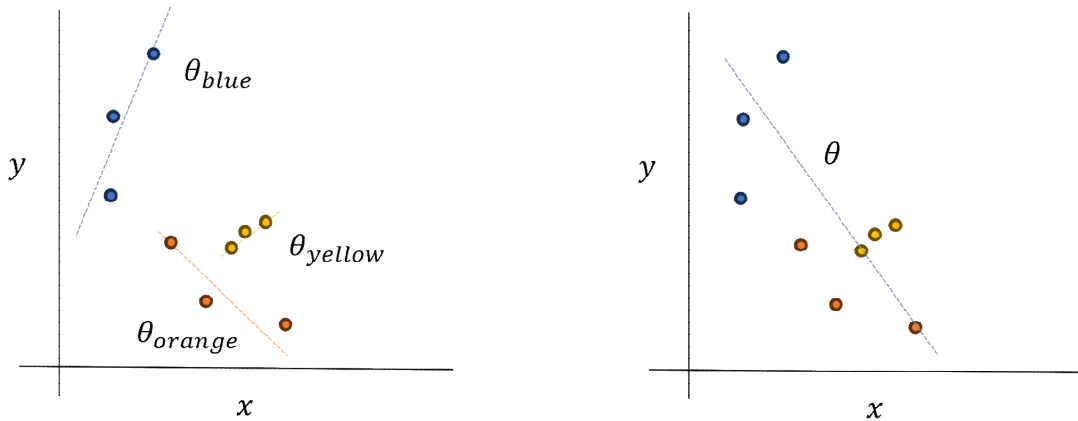


Figure 19. Representation of x as a non-pooled predictor for three different data groups (left), or as a pooled predictor for all data groups (right).

Presented next is a specification of the distribution for the response variable Y_{ij} in the context of a Bayesian hierarchical model. Contrary to the linear regression MLE in the classical scenario as conceived by Fisher, in which the likelihood is

assumed to follow a normal distribution, a different distribution for the likelihood can be assumed, depending on the nature of the data.

In this case, Student's t-distribution is better suited to *overdispersed* data, which occurs when the variability in the data exceeds what would be expected from a standard statistical model that assumes normality (Lambert, 2018).

$$Y_{ij} \sim \text{Student}(v_j, \alpha_j + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \dots + \beta_{Kj}X_{Kij}, \sigma_j) \quad (2-21)$$

Where Y_{ij} represents an independent observation i in group j , v_j the degrees of freedom parameter for the Student's t-distribution (also known as normality parameter; as it increases the distribution converges to a normal distribution), α_j the intercept for group j , $\beta_{1j}, \beta_{2j}, \dots, \beta_{Kj}$ the regression coefficients for the predictor variables $X_{1ij}, X_{2ij}, \dots, X_{Kij}$ and σ_j the scale parameter (traditionally the standard deviation for normal distributions) for group j . In the case of hierarchical linear regression, the centre of the distribution, μ , is the mean of the distribution for group j ; $\mu_j = \alpha_j + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \dots + \beta_{Kj}X_{Kij}$. Notice we are referring to the centre of the posterior predictive distribution, instead of a singular predicted value as in the classical regression scenario.

Below is a plot (Figure 20) showing how v and σ impact the shape of a localised at $\mu = 0$ Student's t-distribution.

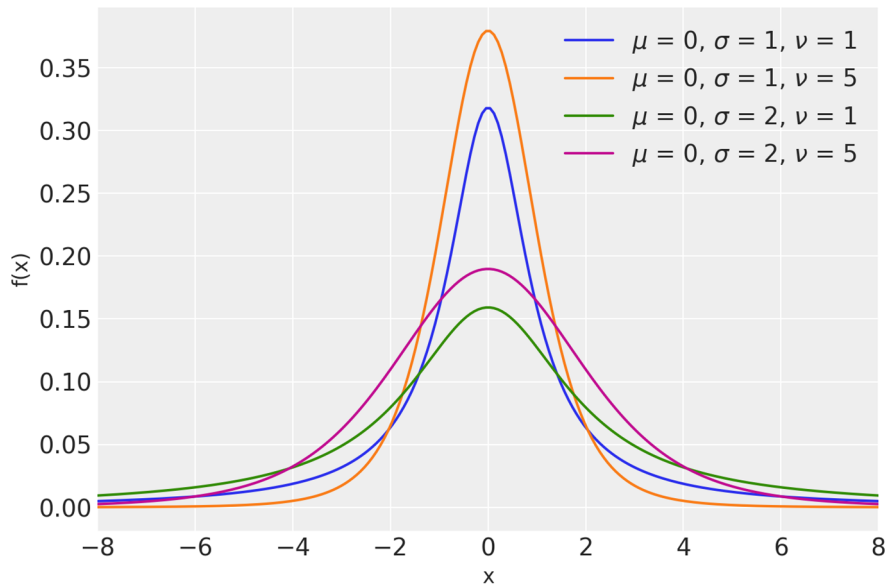


Figure 20. Student's t distributions at different parametrizations.

This is compared to the more commonly used normal distribution in Figure 21.

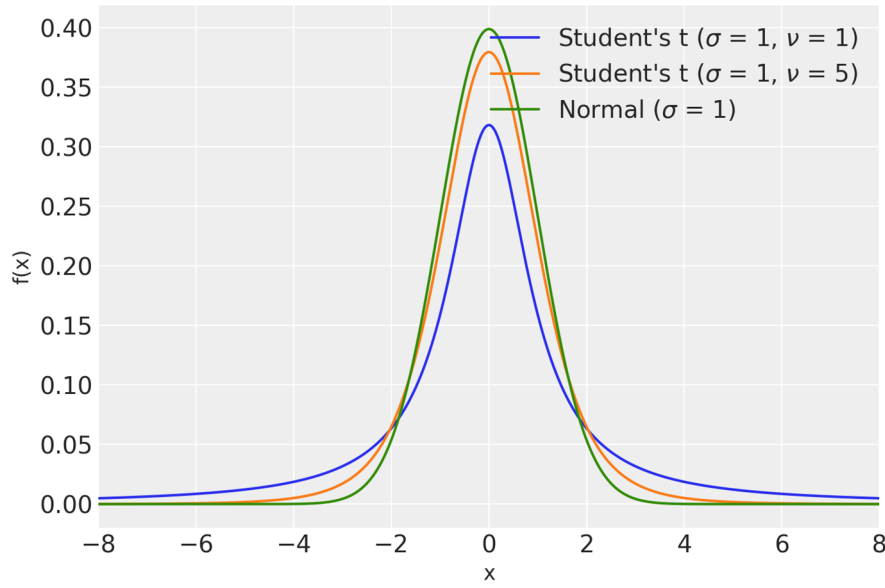


Figure 21. Comparison of Student t's and normal distributions at different parametrizations.

The degrees of freedom ν controls the weight of the tails (their probability mass) and σ controls the width of the peak (same for the normal distribution). The fatness of the tails at low ν values allows to accommodate extreme values with higher probabilities, minimising the possible impact of outliers or in other words, increasing its robustness and reducing its sensitivity to their influence.

However, there is an additional complexity in the case of a hierarchical model, as each group has its own sub-model as was explained earlier. The global model's distribution parameters are not singular values; instead, they can be assigned a distribution of possible values inherited by the sub-models (hyperpriors). The parameters of these distributions that define other parameters in the model are known as hyperparameters. When no priors are specified, such as when there is a lack of prior beliefs in the domain, a common approach to defining the set of prior beliefs is by computing default priors (weakly informative) based on the implied partial correlations of each predictor to the target variable:

$$\begin{aligned}\alpha_j &\sim \mathcal{N}(\mu_\alpha = 0, \sigma_\alpha = \text{HalfNormal}(\sigma = 10)) \\ \beta_{kj} &\sim \mathcal{N}(\mu_{\beta_k} = 0, \sigma_{\beta_k} = \text{HalfNormal}(\sigma = 25))\end{aligned}\tag{2-22}$$

This means that each group j will have its own intercept α with a normally distributed prior located at 0, with a standard deviation hyperprior to which we assign a half-normal distribution with $\sigma = 10$. Therefore, even though its prior is fixed at 0, its scale has its own distribution of plausible values (a probability distribution within a probability distribution). The same applies to j group's β_k predictor coefficients, however a higher uncertainty can be attributed to them compared to the intercepts, depending on the data in question ($\sigma = 25 > 10$).

Prior to numerical computation of the posterior distribution using these priors, it must be noted that the Student's likelihood parameters σ_j and ν_j are usually specified by these distributions (by default):

$$\begin{aligned}\sigma_j &\sim \text{HalfNormal}(\sigma = 1) \\ \nu_j &\sim \text{Gamma}(\alpha = 2, \beta = 0.1)\end{aligned}\tag{2-23}$$

Hence, the posterior distribution can be estimated using Markov Chain Monte Carlo (MCMC):

$$P(\alpha, \beta, \sigma, \nu \mid Y, X) \propto P(Y \mid \alpha, \beta, \sigma, \nu, X) \cdot P(\alpha) \cdot P(\beta) \cdot P(\sigma) \cdot P(\nu)\tag{2-24}$$

In which the conditional probabilities are the posterior distribution (left of the equation) and the likelihood, and the individual probabilities are the priors of the model's parameters.

The posterior samplers can be constructed using a Metropolis within Gibbs scheme (Abril-Pla et al., 2023), which combines the simplicity of univariate parameter updates—as in the case of the Metropolis-Hastings algorithm but with improved computational efficiency—not relying on full joint density distributions, but rather leveraging the conditional nature of Gibbs sampling (in which every draw is added to the chain without checking for acceptance, but usually requires more draws). In the Metropolis within Gibbs framework, parameters are updated sequentially by proposing a new value for a parameter and evaluating its acceptance based on a proposal distribution. This approach is seamlessly integrated with the exhaustive multivariate conditional updates inherent in Gibbs sampling, where each parameter is updated conditioned on the current values of all other parameters (Gilks et al., 1995b). Many models only use continuous

distributions for their parameters, and in the current iteration of PYMC (PyMC5), which is the package used in this project for inference, the default MCMC algorithm for continuous distributions is the No-U-Turn-Sampler (NUTS) (Abril-Pla et al., 2023; Hoffman et al., 2014), an Adaptive Dynamic Hamiltonian Monte Carlo.

Hamiltonian Monte Carlo (HMC) treats parameters as positions and their gradients to the target distribution as momenta, so that steps (the proposals regarding plausible parameter values) do not follow a random walk behaviour, providing a more efficient exploration of high-dimensional parameter spaces. However, HMC encounters challenges in parameter spaces with piecewise smoothness, where some regions exhibit transitions (different geometries) that are not well-handled by the leapfrog nature of HMC, leading to suboptimal performance (Afshar & Domke, 2015). NUTS, being an extension of HMC, addresses these challenges by dynamically adapting the number of steps to take and efficiently exploring complex posterior distributions. More specifically, NUTS uses binary trees to explore multiple trajectories simultaneously, dynamically determining the optimal number of steps by doubling until a “U-turn” (trajectory that loops back on itself) is detected, thereby avoiding geometric pitfalls by using this information to determine what path to take. The step size is adapted during the burn-in phase (where the chain explores an appropriate initial starting point while discarding the samples drawn so far) to a target acceptance rate, and used at all other remaining steps of inference (Monnahan et al., 2017). The target acceptance rate is the percentage of iterations in which the proposals were accepted and is set by the user; generally, at 80%. Higher target acceptance rates result in a lower number of divergences during inference, which occur when a Hamiltonian trajectory fails to capture the behaviour of the target distribution, leading to a departure regarding the true trajectory. This occurs typically in regions of high curvature or steep gradients in the posterior distribution (therefore the simulation cannot be trusted). Divergences are minimised with appropriate step sizes (Stan Development Team, 2022). The downside to higher target acceptance rates is usually slower sampling.

After the description of the Bayes theorem, the differences between frequentist and Bayesian methods, which have already been addressed previously, are detailed in more depth in the next section.

2.7.7 Bayesian vs frequentist methods

Bayesian statistics are older than frequentist statistics, but due to the computational challenges they were neglected over the years, until new algorithms such as MCMC and improvements in their accessibility enabled a larger adoption by data analysts. Hackenberger enumerates in simple terms several of the key differences of Bayesian over frequentist statistics (Hackenberger, 2019):

- As long as the prior probability model is set correctly, all relevant information in the Bayesian setting is contained in the observed data rather than the unobserved. The use of p -values, as they are partly determined by never observed data (extrapolating sample metrics to population metrics) is one of the main objections of frequentist hypothesis testing. The null and alternative hypothesis paradigm is replaced by the subjective choice of what is a suitable prior.
- In frequentist approaches, the experimental design must be specified in advance as it influences hypothesis test results. In frequentist hypothesis testing it is not the same to challenge the hypothesis that there is the same chance of picking a head or a tail in a coin toss if the observations are generated by throwing a coin until heads is obtained (a series of Bernoulli trials, a geometric null distribution), than if throwing a number of coins in the air and randomly picking them out (a binomial null distribution). Hackenberger illustrates this point by attempting to infer the probability of a mouse being male or female under these two different experimental designs (Figure 22) from a population of 7 mice (picking mice one by one until you encounter a male or randomly pick 7 mice and count number of males). In the first case the null hypothesis that $P(\text{Male}) = P(\text{Female})$ (p -value = 0.031) can be rejected because it took 7 trials until a male was found, but there is not enough evidence in the second experiment (one

male was found among seven random mice) for such a claim ($p\text{-value}=0.227$). In the Bayesian set up, the experiment design would not have impacted this computation. Supposing the choice of prior is a beta distribution, it is possible to calculate the probability that the sex ratio in this mice population is not 1:1 as $P=0.92$, or 92%, in either of the design cases.

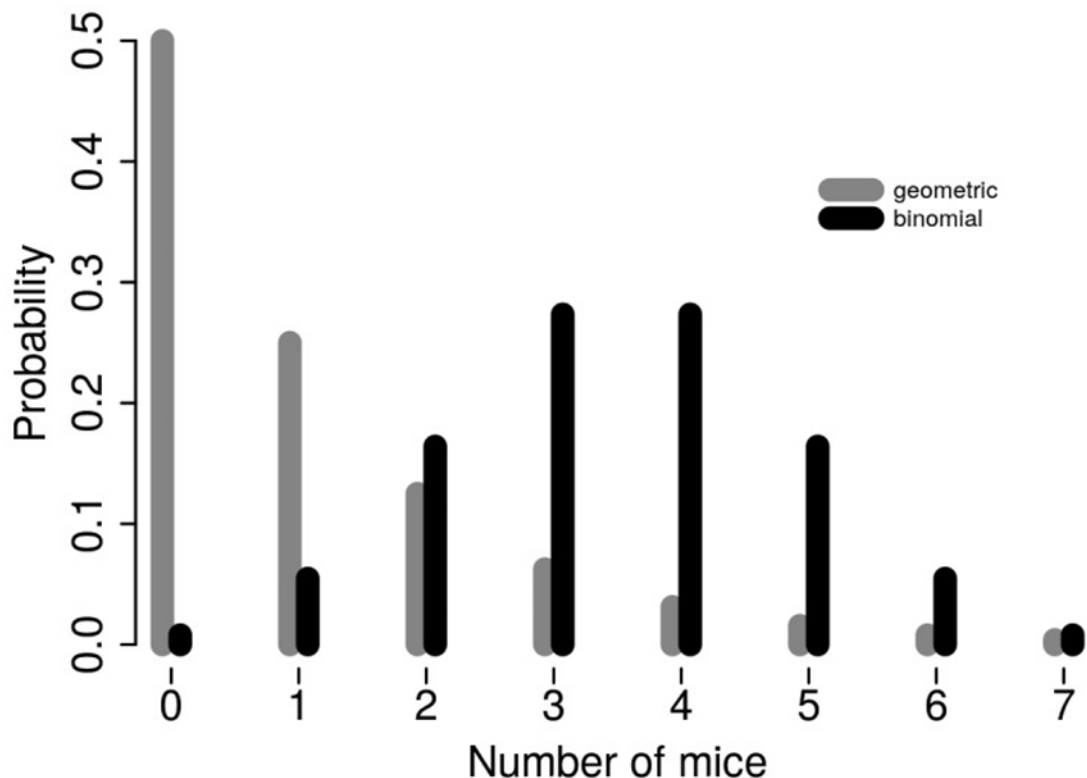


Figure 22. Absolute frequency probabilities of number of male mice to be found in a sample of 7 random mice, inferred from two different experimental designs both originally observing only 1/7 mice are male. The geometric distribution clearly illustrates there is sufficient evidence to conclude it is probable to only find one or no male mice, whilst in the binomial distribution experiment this was not evident thus the binomial probability of 50% of a mouse being a male mouse still stands. [Obtained from Hackenberger, 2019]

- Bayesian methods have an advantage over frequentist methods when comparing two datasets, as the latter typically depend on the assumption of normality of the hypothetical population the samples come from and that the observed samples' parameters are representative of the population's

parameters. They estimate, with a specific confidence (i.e., 95%), what a parameter's value will usually be. Bayesian analysis answers the question "*What is the probability of the hypothesis given the measured data?*" via an estimation of the ratio of probabilities of two different hypothesis (the Bayesian factor previously mentioned) that quantifies the support for each rather than simply accepting or rejecting the null hypothesis.

There are other considerations, but Hackenberg concludes they each have their own strengths and weaknesses and that they are both powerful tools to be considered. Bland & Altman, proposers of the difference plots that have adopted their name (Bland & Altman, 1986), have a different opinion which they enunciated in 1998 (Bland & Altman, 1998); frequentist statisticians are developing an interest in the Bayesian method as it can provide solutions to difficult problems and they expected that the presence of Bayesian analyses in the scientific literature would grow.

The final consideration for the development of predictive regression models is to decide what data will be used for inference, which is the subject of the next section.

2.7.8 Feature selection in Machine Learning

In the generation of a mathematical model, it is unlikely that any or every candidate predictors X is relevant for the modelling of the target variable y . Feature selection refers to the identification of the most relevant features to explain the behaviour of variable y across its known range of possible values. This includes removing variables, as well as creating new variables from the ones that already exist (Chicco et al., 2022). Feature selection is part of the feature engineering process, which also includes preprocessing the data in order to define a solvable mathematical problem and handling data such as outliers.

There are many approaches to feature selection, which can be classified according to the algorithm type (Pudjihartono et al., 2022; Saeys et al., 2007):

- Univariate filters

- o Rank features based on a specific metric, and those below a threshold are removed (i.e., Spearman rho, Pearson correlation, t-test, etc.).
- o These methods are fast and scalable, but do not consider interactions between features, which might lead to oversights in determining which is the optimal feature set.
- Multivariate filters
 - o Similar to univariate filters but can consider several features at the same time (i.e., mutual information feature selection).
 - o The properties of these methods are the opposite of univariate filters; they are more accurate, but slower performing.
- Wrapper methods
 - o The filter methods were agnostic of the type of modeller being employed; wrapper methods use the performance of the modelling algorithm to aid selection of the optimal features (i.e. recursive feature elimination or sequential selection).
 - o The slowest type of algorithm, but generally the most accurate in providing the optimal feature set, with a possible downside; this set can be overfitted unless appropriate sampling strategies are employed.
- Embedded methods
 - o The feature selection is built within the modelling algorithm; during training, the algorithm is performing feature selection internally (i.e. Lasso and Random Forest).
 - o Provides an intermediate balance between the properties of filter methods and wrapper methods, however the optimal set of features is conditioned by the inferential approach of the modelling algorithm. For example, in Random Forest the optimal features will likely be those that by binary splits provide the best results.

The “no free lunch” theorem (Wolpert & Macready, 1997) states that no single algorithm is optimal for all problem settings; each have their own strengths and weaknesses and require iterative testing over the data collected.

For brevity's sake, only the following algorithms that played an important role in this project will be elaborated upon:

Boruta

Boruta was published in 2010 by Kursa and Rudnicki (Kursa & Rudnicki, 2010). It is a wrapper algorithm employing the embedded feature selection of ensembles of decision trees, such as Random Forest, leveraging the high versatility of this algorithm (suitable for many regression problems).

Boruta aids feature selection in high-dimensional datasets, utilising the importance scores of Random Forest. The main innovation of its algorithm are shadow features, created by random permutation, that enhance the benchmark of each feature's significance by assessing the impact of switching each feature with random noise on modelling performance. As of June 2024, the original Boruta paper has been cited over 3.9k times according to the Journal of Statistical Software, indicating its methodology has been approved for use by a substantial part of the scientific community.

Autofeat

On the opposite side of the versatility of ensembles of decision trees (the algorithm under the hood of Boruta), are linear regression models, only suitable for datasets in which the patterns between the variables are of a linear nature or have been effectively transformed into such. Autofeat (Horn et al., 2020) is also a wrapper algorithm, built in this case over the Lasso modelling algorithm (which has its own embedded feature selection process such as Random Forest). The use of feature permutation is also an inherent part of the feature selection process. It can be considered as an alternative to Boruta when searching for models of a linear nature, however it also presents the capacity to generate new features based on combinations of original features (such as by polynomial transformations), if the user requests so.

Laplacian priors

The use of Laplacian priors over the parameters of the linear regression equation in a Bayesian setting has been likened to feature selection using Lasso (T. Park & Casella, 2008). Laplacian priors place more probability mass on the centre of

the distribution; the distribution is not bell-shaped such as the normal distribution but much sharper (Figure 23). The scale parameter is notated as b instead of σ .

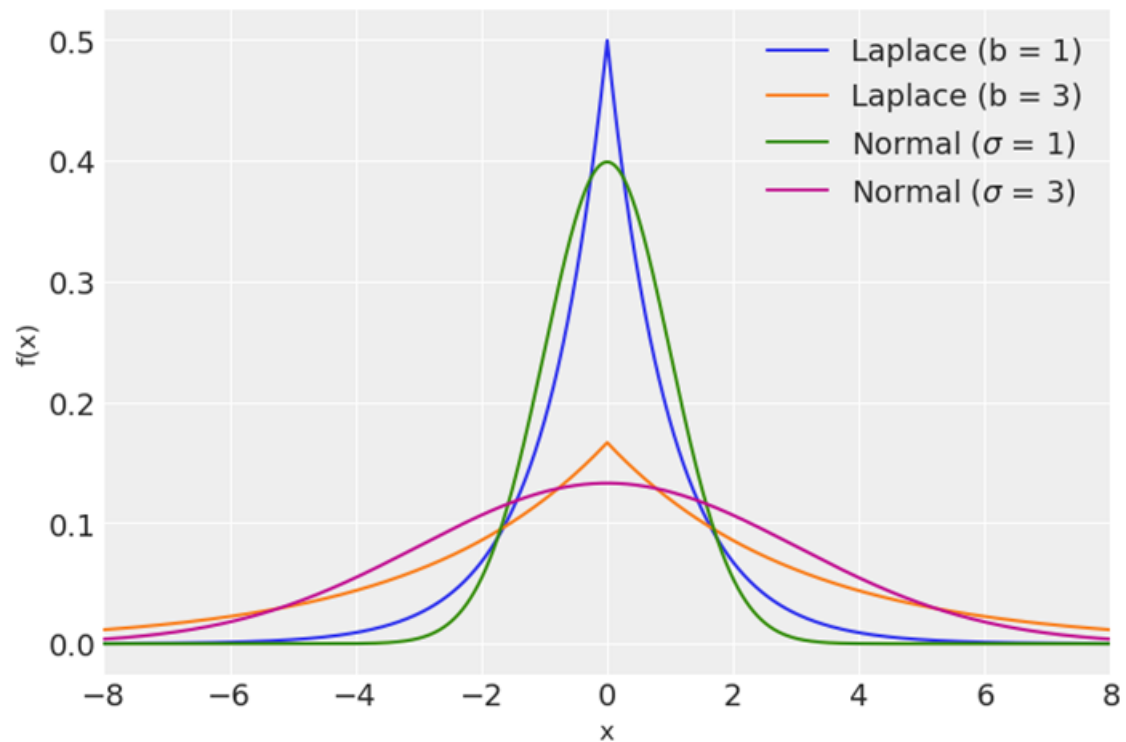


Figure 23. Laplace distributions compared to normal distributions.

As a result of placing Laplacian priors on parameters with a centre location at 0, the prior belief in the possibility that those parameters are 0 is stronger. The model's training will penalise non-significant parameters (parameters with values close to 0) unless the data's likelihood clearly suggests that a parameter is significant in the model. Other popular Bayesian methods for feature selection similar to the usage of Laplacian priors are the spike-and-slab (Ishwaran & Rao, 2005) and regularized horseshoe (Piironen & Vehtari, 2017).

2.7.9 Applications in the food industry

This final section of the literature review shifts the focus from the technical aspects of ML to its current applications. Since the 2010s, the increased accessibility and power of ML algorithms, combined with the creation of expansive datasets, have led to numerous applications in the food industry. These advancements have enabled the efficient analysis of large and complex

datasets and the extraction of valuable insights within the sector. A few examples of which are discussed below:

- Spoilage in several food products (beef, salmon, fruit and vegetables), which can occur by microbial or fungal contamination, has well established detection workflows using rapid sensors (spectroscopy or multi- and hyperspectral imaging) coupled with signal analysis via ML, which are capable of competing against the traditional methods such as FT-IR or gas chromatography coupled with mass spectrometry (H.-J. He et al., 2014; Nychas et al., 2021; Panagou et al., 2014a, 2014b).
- Adulteration detection can also be tackled with the previous imaging-based approach (Goyal et al., 2024), for example to detect sweeteners in honey (Razavi & Kenari, 2023) or Atlantic salmon mixed with cheaper fish (P. Li et al., 2023).
- Flavour prediction is influenced by numerous and complex factors, making it a perfect target for ML approaches and numerous studies have been published in this vein (Ji et al., 2023). Employing ML to assist breeding programs (Colantonio et al., 2022) is a promising avenue for earlier optimization of plant-based foods.
- Demand forecasting can help minimise food waste, as was proven by Rodrigues et al. recently in a small, short-term study. Based on date and weather data, a reduction in food waste between 14 and 52% was achieved at three different student and company catering services (Rodrigues et al., 2024).
- Food sorting, such as classification of foods aimed at automating the digital food industry (Tsakanikas et al., 2020), or the identification of restaurants prone to cause food poisoning using ML-based epidemiology (Sadilek et al., 2018).
- The degree of food processing was recently tackled as a classification task (4 classes from unprocessed to ultra-processed) for a dataset of several thousand food products, based on nutritional composition of 12 nutrients. Accuracy was exceptionally high, measured at over 0.95, and concluded that 73% of the US food supply (the analysed dataset) was composed of ultra-processed foods (Menichetti et al., 2023).

- Previous works studying bioavailability exist, focusing mainly on drugs and their absorption or post-consumption effects, or similarly on patient susceptibility to absorbing toxic compounds or contaminants (Thishya et al., 2018). Using as predictors the physicochemical properties of existing drug databases, there are several accounts of highly-accurate classification models (Morales et al., 2024; Tyagi et al., 2023). Bayesian methods have also been employed recently to assess the bioequivalence (same therapeutic effects) of two formulations of the same drug (Lv et al., 2023).

To date, none of the reviewed applications of ML in the food industry have tackled the impact of food matrix effects on bioaccessibility of a compound of interest.

Highlights

- Machine Learning is a broad field; this project's main subject was supervised learning of a regression problem.
- Frequentist methods consider parameters of interest as fixed values, Bayesian considers them random variables.
- Powerful methods for frequentist inference model the data exclusively.
- Bayesian methods do not only model the data but also allow incorporation of previous knowledge, which can introduce additional subjective decisions.
- Bayesian hierarchical models provide flexibility for multi-group datasets.
- Selection of predictors is required to minimise risk of fitting data on noise and overfitting, especially for small datasets.
- Machine Learning has seen a wealth of applications in the food industry, but none tackling bioactive bioaccessibility as a function of the food matrix effects.

3 Methodology

This section presents the materials and methods of data acquisition that produced, for vitamin E and curcuminoids, a dataset of bioaccessibility in different carrier food formulations. In addition to the bioaccessibility quantification pipelines, various techniques were applied to characterise the properties of the carrier food formulations. Statistical methods from diverse schools of thought were then employed to analyse these formulation characteristics and how they affected bioaccessibility. The aims and objectives are presented below.

3.1 Aims and objectives

The scientific aim of this work is to prove (or disprove) the hypothesis that systematic screening of food formulations can provide valuable insights towards optimisation of nutrient bioaccessibility and be used to develop a decision support system for the design of optimal functional food recipes. To this end, the main goal was planned as a series of objectives:

1. Design and optimise protocols that enable the study of a given compound of interest's bioaccessibility in a food formulation in a reproducible, accurate and time- and cost- efficient manner.
2. Apply the protocols to systematically screen the bioactive compound's bioaccessibility in foods formulated with different ingredients, cooking methods and physicochemical states..
3. Obtain theoretical or empirical data corresponding to the formulations' properties that will explain the outcome of why and how they impact the bioaccessibility.
4. Use the explanatory data to develop an accurate predictive model of compound bioaccessibility as a function of the properties of the food formulation carrier.

3.2 Bioaccessibility quantification

3.2.1 Preparation of food matrices

The ingredients of the base recipes proceeded from commercial brands. Semi-skimmed milk was purchased from Auchan (Auchan Hauts-de-France) as well as corn starch (Maizena) and white sugar (Daddy Sucre). Egg yolks were prepared

from egg yolk powder (Gallia, Louis François) which was reconstituted with water according to the manufacturer's instructions (ratio of 40 grams of egg powder and 60 mL of potable tap water), and T45 wheat flour (Naturaline) was used as received by manufacturer. Mentions of egg yolk quantities throughout this manuscript refer to reconstituted egg yolks unless otherwise specified. Tap water was used despite the limited chemical data, as it is usual in food research manuscripts.

The custards were prepared by adding, without any particular order aside from pouring in the milk as last ingredient:

- 60 g corn starch.
- 80 g white sugar.
- 100 g of egg yolks (40 g of powder and 60 g of potable tap water).
- 500 mL of milk.

Components were added to a HotmixPro thermal mixer from Vitaeco (Modena, Italy) and cooked for 10 min at an agitation of 200 rpm at 85°C. Custards were stored at 4°C in 125 mL cylindrical plastic containers up to the brim to ensure the surface was flat prior to rheology analysis (minimum 3 containers) (Figure 24), and afterwards were stored at -20°C.

The ingredients used to prepare biscuits were:

- 60 g wheat flour.
- 80 g white sugar.
- 100 g of egg yolks (40 g of powder and 60 g of potable tap water).
- 10 mL of milk.

Ingredients were added (in no particular order) and whisked until homogenization in a Kenwood Mix KMX50 (Yokohama, Japan) and 30 g of dough was



Figure 24. Custards with (above) and without (below) curcuminoids, in the plastic 125 mL containers ready for rheology analysis.



Figure 25. Biscuits with and without curcuminoids.

shaped into a biscuit form in a circular mould of 10 cm diameter and 2 cm deep. Six biscuits were prepared in this manner for each formulation (Figure 25), or until no more dough was left. The shaped dough was baked in a professional oven Venix G07M (Veneto, Italy) oven for 30 min at 160°C, previously preheated. Biscuits were stored in the dark at room temperature in sealed plastic containers for a short period of time until further analyses.

The relative amounts (in mass) of the ingredients of these base recipes of custards and biscuits were as follows:

Table 2. Ingredients of standard and curcumin enriched custards and biscuits.

Ingredients	Custard	Biscuit
Semi-skimmed milk (%)	68.3	4.1
Egg yolk (%)	13.2	39.9
Sugar (%)	10.6	32
Corn starch (%)	7.9	-
T45 Wheat flour (%)	-	24
Fibre (%)		

Custards were made with corn starch instead of wheat flour to ensure a custard-like consistency. The base recipe of custards and biscuits contained equal amounts (mass) of solid ingredients, thereby providing complete matrices that facilitated the study of a wide range of ingredient and nutrient combinations, including lipids, carbohydrates, proteins, and dietary fibres whilst being simple to elaborate. Even though these foods were formulated with almost identical ingredients, they also provided the opportunity to study the impact of the physical state of the food (viscous vs solid) on bioaccessibility, as custards had 50X milk than biscuits. This concept was inspired from a previous study on the impact of food matrix microstructure on folate bioavailability (Buffière et al., 2021).

The information on the labelling of these ingredients would be used to compute the nutritional composition of the tested formulations (Table 3).

Table 3. Nutritional composition of the base recipe ingredients according to products' labelling (in percentages). Empty cells represent unknown/non-existent data and were imputed as 0. *The types of fibre of corn starch and wheat flour were unknown but inferred to be mostly soluble and insoluble, respectively, in accordance with the literature (Shewry et al., 2023; Van Hul et al., 2020).

Ingredients	Energy (kcal) per 100g	Insoluble fibre	Soluble fibre	Protein	Carbohydrate		Fat		Water	Ash
					Total	Sugars	Total	Saturated		
Semi-skimmed milk (%)		-	-	3.3	4.8	4.8	1.6	1	90.3	-
Egg yolk (%)	275	-	-	12	1	0.64	23.6	-	62	1.4
Sugar (%)	400	-	-	0	100	100	0	0	0	-
Corn starch (%)	355	-	1*	0.5	86	0.5	0.5	0.1	0	-
T45 Wheat flour (%)	359	2.5*	-	9.3	76.7	1	1.7	0.27	0	-

3.2.1.1 Vitamin-E-enriched matrices

Both custards and biscuits were prepared with a 1X concentration of 21 µg/g (prior to cooking) of α-tocopherol (T2028) purchased from Sigma-Aldrich (Merck, Saint Quentin Fallavier, France). A wide range of alterations to the base recipes were considered. However, the choice of formulations to screen was restricted due to limitations of the bioaccessibility determination pipeline (Section 3.2.3.1). Furthermore, in order to produce matrices similar to these currently being offered by commercial brands to the public, commercial dietary fibre supplements were used and a limit of 10% dietary fibre supplementation was imposed for custards. This limit was also extended to biscuits even though some publicly available biscuits provide up to 40% dietary fibres, so that the results of both matrices would remain comparable. The following is a list of the vitamin-E enriched custard formulations tested:

- Base recipe with different amounts of corn starch (40 g and 90 g instead of 60 g).
- Replacement of corn starch with white rice flour (Celnat, Saint-Germain-Laprade, France), tapioca starch (Tate&Lyle, London, United Kingdom), coconut flour (Ecoidees, Soultz-sous-Forêts, France), and a 50:50 mix of corn starch and coconut flour.
- Supplementation with inulin (Bulk, London, England), microcrystalline cellulose (REDWELLS, Windsor, England), apple fibre (FoodSolutionsTeam, Zurich, Switzerland), pea fibre (Roquette, Lestrem, France), and wheat fibre (JRS, Rosenberg, Germany) at 2%, 5%, and 10% supplementation after the base ingredients were weighed and added. These percentages refer to the mass of the base recipe mix prior to supplementation (10% supplementation of a 750 g custard batch would require the addition of 75 g of fibre supplement).

In the case of biscuits, the following formulations were tested:

- Base recipe with vitamin E at different concentrations (21, 105, 210, 315, and 420 µg/g).

- Base recipe cooked at 160°C instead of 180°C for the same period of time (30 min).
- Addition of extra milk to base recipe (20 mL instead of 10 mL).
- 50:50 mix of wheat flour with rice, tapioca, and coconut flours (each separately).
- Supplementation with inulin, microcrystalline cellulose and apple fibre at 2%, 4%, and 10% supplementation after the base ingredients were weighed and added. These percentages refer to the mass of the base recipe mix prior to supplementation (10% supplementation of a 250 g custard batch would require the addition of 25 g of fibre supplement).

In total, 22 custard and 16 biscuit formulations were tested.

3.2.1.2 Curcuminoid-enriched matrices

Blank custards were prepared without curcuminoid supplementation, whilst custards with curcumin were enriched (prior to cooking) with 35.6 mg/100g of commercial curcumin powder; curcumin from *Curcuma longa* (Turmeric) (curcumin content approx. 75%, C 1386, CAS [458-37-7](#), Sigma-Aldrich, Saint Quentin Fallavier, France). Blank biscuits were prepared without curcuminoid supplementation, whilst biscuits with curcumin were enriched (prior to cooking) with 106.7 mg/100g of curcumin powder.

As was mentioned earlier, the source of variation for the curcuminoid-enriched formulations was the addition of fibre supplements, which were mostly of a plant-based origin (carrot, pea, citrus, wheat) and obtained from commercial manufacturers. Fibre source was supplemented at 6% because it was the highest concentration that could be applied to both food types to generate formulations with significantly different properties, whilst avoiding potentially negative sensory properties (pungent flavours or too hard/soft texture) for potential consumers as was observed in a small test cohort in the lab. The final ratios of ingredients for the case of 6% fibre supplementation are presented in Table 4. The nutritional composition of the fibre supplements, according to the products' specifications provided by suppliers (Table 5), together with that of the base ingredients (Table 3), were used to calculate nutritional composition of the prepared formulations. The types of fibres of each supplement were imputed from the literature (Gorlov

et al., 2015; C. He et al., 2023; G. Liu et al., 2014; Nawirska & Kwaśniewska, 2005; K. D. Sharma et al., 2012).

Table 4. Ingredients (in percentage) of the fibre-supplemented formulations compared to base recipes.

Ingredients	Custard	Biscuit	6% fibre-fortified custard	6% fibre-fortified biscuit
Semi-skimmed milk (%)	68.3	4.1	64.1	3.8
Egg yolk (%)	13.2	39.9	12.6	37.7
Sugar (%)	10.6	32	10.1	30.2
Corn starch (%)	7.9	-	7.5	-
T45 Wheat flour (%)	-	24	-	22.6
Fibre (%)			5.7	5.7

The following soluble fibres were used: a soluble inulin powder Orafit® (Beneo GmbH, Mannheim, Germany) from chicory; a soluble fibre Nutriose® FM06 (Roquette) from Maize, a soluble citrus peel fibre (Ceamfibre 7000, CEAMSA, Pontevedra, Spain), two different carrot fibres (FST 00007 KaroPRO 1-26 and FST 00018 KaroPRO SG, FoodSolutionsTeams GmbH, Hettingen, Switzerland), containing the same amount of both soluble and insoluble fibres but displaying different physical properties, an insoluble apple fibre (FST 00224 ApplePRO 60+, FoodSolutionsTeams GmbH), an insoluble yellow pea fibre (Pea fibre 50 M Roquette frères, Lestrem, France), an insoluble wheat fibre (Vitacel WF200, JRS, Rosenberg, Germany), and an insoluble microcrystalline cellulose (Redwells® London, UK) from wood pulp.

Finally, liquid matrices (water and milk) were prepared by dissolving 35.6 mg/100g of curcuminoid powder in room temperature water and semi-skimmed milk by vortexing 30 seconds.

Table 5. Supplemented fibre sources and their known nutritional composition (in percentages). Empty cells represent unknown/non-existent data and were imputed as 0. *fibre content was imputed from the literature.

	Insoluble fibre (%)	Soluble fibre (%)	Protein (%)	Carbohydrates (%)	Total fat (%)	Water (%)	Ash (%)	Cellulose (%)	Hemicellulose (%)	Pectin (%)	Fructans (%)	Lignin (%)	Dextrin (%)	Origin
Orafti® GR	0	>90	-	<10	-	3	<0.2	-	-	-	>90	-	-	Chicory
Nutriose® FM06	-	82-88	<0.3	<0.5	0.1	<5	<0.5	-	-	-	-	-	82-88	Maize
FST 00007 KaroPRO 1-26	33	27	5	21	1	6	4	41	7	5	-	7	-	Carrot
FST 00018 KaroPRO SG														
Pea fibre 50 M	>50	-	<10	35	-	<10	3	20	30	-	-	-	-	Yellow pea
FST 00224 ApplePRO 60+	52	8	5.8	27.6	<3.2	<8	1.8	27	16	6	-	10	-	Apple
Ceamfibre 7000	85	1.5	2.9	-	0.8	7.6	-	41	16	1.5	-	6	-	Citrus peel
Vitacel WF200	97	-	0.4	-	0.2	<8	<3	24	73	-	-	-	-	Wheat
Microcrystalline cellulose	100	-	-	-	-	-	-	100	-	-	-	-	-	Wood pulp

3.2.2 Static *in vitro* digestion

Static *in vitro* digestion was carried out following the guidelines established by the INFOGEST consensus protocol (Brodkorb et al., 2019), with slight modifications. Since its publication, this protocol has been the most widely used version of static digestion simulation; it is regarded as highly efficient in spite of the shortcomings of this type of static methods (as was detailed in the literature review, section 2.5) and its usage facilitates drawing comparisons to other experiments in the literature.

All reagents, unless specified, were purchased from Sigma-Aldrich (Merck, Saint Quentin Fallavier, France). Ultrapure water (from Milli-Q® system) was used in solution preparation and other steps unless specified otherwise. Preparing stock solutions of the simulated gastrointestinal fluids was the first step. The stocks include 1 L solutions that are prepared and stored in independent bottles (so they can be reused across experiments) in the fridge at 5 °C of:

- 0.5 M KCl.
- 0.5 M KH_2PO_4 .
- 2 M NaCl.
- 0.15 M $\text{MgCl}_2(\text{H}_2\text{O})_6$.
- 0.5 M $(\text{NH}_4)_2\text{CO}_3$.
- 6 M HCl (stored in security cabinet for acids).

At experiment time, two stocks are also prepared (not reused):

- 1 M NaHCO_3 .
- 0.3 M $\text{CaCl}_2(\text{H}_2\text{O})_2$.

These stocks are used to prepare the simulated gastrointestinal fluids through the following steps, in the same order as they appear:

- Simulated salivary fluid (SSF, pH 7) (4 mL per experiment)
 - 15.1 mL KCl 0.5 M.
 - 3.7 mL KH_2PO_4 0.5 M.
 - 6.8 mL NaHCO_3 1 M.
 - 500 μL $\text{MgCl}_2(\text{H}_2\text{O})_6$ 0.15 M.

60 μL $(\text{NH}_4)_2\text{CO}_3$ 0.5 M.

90 μL HCl 6 M.

Fill to 400 mL with ultrapure water.

Store in the fridge at 5°C for reuse.

- Simulated gastric fluid (SGF, pH 3) (8 mL per experiment)

6.9 mL KCl 0.5 M.

0.9 mL KH_2PO_4 0.5 M.

12.5 mL NaHCO_3 1 M.

11.8 mL NaCl 2 M.

400 μL $\text{MgCl}_2(\text{H}_2\text{O})_6$ 0.15 M.

500 μL $(\text{NH}_4)_2\text{CO}_3$ 0.5 M.

1.3 mL HCl 6 M.

Fill to 400 mL with ultrapure water.

Store in the fridge at 5°C for reuse.

- Simulated intestinal fluid (SGF, pH 7) (16 mL per experiment)

6.8 mL KCl 0.5 M.

0.8 mL KH_2PO_4 0.5 M.

42.5 mL NaHCO_3 1 M.

9.6 mL NaCl 2 M.

mL $\text{MgCl}_2(\text{H}_2\text{O})_6$ 0.15 M.

700 μL mL HCl 6 M.

Fill to 400 mL with ultrapure water.

Store in the fridge at 5°C for reuse.

After the gastrointestinal fluids have been prepared and prior to the first incubation, which simulates the passing of the food through the stomach, only the oral amylase and the pepsin are prepared:

- 0.5 mL per digestion experiment of 1.5 kU/mL oral amylase, prepared in water. Enzyme in question is thermostable α -amylase from *Bacillus sp.* from Megazyme (Bray, Ireland), purchased as part of a starch determination kit (K-TSHK) or standalone (E-BSTAA).
- 0.5 mL per digestion experiment of 80 kU/mL pepsin, prepared in water. Product in question is pepsin from gastric mucosa (P7000) from Sigma-Aldrich.

The oral phase proceeds as the following:

1. For biscuits, grind the centres of the biscuits (not the borders) in a Moulinex grinder (Fruit Sensation, Moulinex, Paris, France); weigh 3 g of biscuit, grind, and take 1.67 g into a 50 mL Falcon tube. Add 3.34 ml of water and vortex. If digesting more than one replicate, grind only once for all replicates, by grinding extra biscuit (grind approximately 7.5 g of biscuits for 3 replicates). The goal is to obtain a tomato-paste-like consistency as recommended by Brodkorb et al.
2. For custards, weigh 5 g in a Falcon tube of 50 mL. As a result, both custards and biscuits had a similar concentration of solid matter in the digested samples (1.15 g of solid ingredient matter). Similarly, all curcuminoid formulation digestions contained approximately 1.78 mg of curcuminoids in the digesta.
3. Mix the 5 grams of food with 4 mL of SSF. Consistency should be of tomato paste or mustard.
4. Simulate mastication by intensely vortexing and inverting until the sample is homogenised (typically around 10 seconds).
5. Add $\text{CaCl}_2(\text{H}_2\text{O})_2$ to achieve a total concentration of 1.5 mM in SSF (add 25 μL).
6. Add 0.475 mL of water.
7. Add the amylase prepared in water to achieve an activity of 75 U/mL in the final mixture (add 0.5 mL).
8. Incubate while mixing for 2 min at 37 °C in a water bath. To do this, place the 50 mL Falcon tubes on a plastic rack, and secure with another rack on top of the tubes. Submerge in the bath, and every 15 seconds remove and shake for 2-3 seconds.

The gastric phase is next. No gastric lipase was used in the gastric phase, contrary to what the INFOGEST protocol recommends, and human salivary amylase was replaced by bacterial amylase. This was due to their prohibitive high cost when applied to extensive screening experiments (\$10< vs \$>120 per digestion). Pancreatic lipase was considered instead of gastric lipase, but it was observed it had no activity at gastric pH (refer to 3.2.2.1). The major form of salivary amylase in humans is α -amylase (Aguirre et al., 1987), therefore it was

considered appropriate to substitute it with bacterial α -amylase at the same activity concentration. The gastric phase lasts 2 to 3 hours and requires the following steps:

1. Pre-warm the SGF electrolyte stock solution at 37 °C. Add SGF electrolyte stock solution to each Falcon tube to achieve a final ratio of 1:1 (vol/vol) (add 8 mL).
2. Add $\text{CaCl}_2(\text{H}_2\text{O})_2$ solution to achieve a final concentration of 0.15 mM in SGF (add 5 μL).
3. Adjust the pH to 3.0 by adding a defined volume of HCl previously determined during a pH-test adjustment experiment. If not, begin by adding 170 μL of HCl 1 M to only one digestion, determine the optimum volume, and then add said volume to remaining digestions immediately before adjusting each digestion one by one.
4. Add water to achieve a 1 $\times$ concentration of SGF (depends on volume added in the previous step; a final volume of 20 mL is to be reached, so 1 mL is approx. used).
5. Add the porcine pepsin solution prepared in water to achieve an activity of 2,000 U/mL in the final digestion mixture (add 0.5 mL).
6. Incubate the samples at 37 °C, mixing the digestive mixture sufficiently (shaking incubator at 170-200 rpm) for 2 h (starting the countdown at the time the pepsin was added).

During the incubation at step 6, pancreatin and bile are prepared:

- 5 mL per digestion experiment of 800 U/mL pancreatin, prepared in SIF. Enzymatic reagent in question is pancreatin from porcine pancreas (P7545).
- 2.5 mL per digestion experiment of 0.16 mM bile, prepared in SIF. Enzymatic reagent in question is bile porcine (B8631).

So that the intestinal phase can be performed according to schedule (2-3 hours in duration):

1. Pre-warm the SIF electrolyte stock solution in a 37 °C water bath.
2. Add SIF electrolyte to the Falcon tubes to achieve a final ratio of 1:1 (vol/vol) (add 8.5 mL).

3. Add $\text{CaCl}_2(\text{H}_2\text{O})_2$ solution to reach a concentration of 0.6 mM in SIF (add 40 μL).
4. Adjust the pH to 7.0 by adding a defined volume of NaOH previously determined during a pH-test adjustment experiment. If not, begin by adding 150 μL of NaOH 1M to only one digestion, determine the optimum volume, and then add said volume to remaining digestions immediately before adjusting each digestion one by one.
5. Add water to achieve a 1 $\times$ concentration of the SIF (final volume of 40 mL).
6. Add the bile solution to the SIF/gastric chime solution to reach a final concentration of 10 mM (add 2.5 mL).
7. Perform the intestinal phase with pancreatin (5 mL, final trypsin activity of 100 U/mL).
8. Incubate the samples at 37 °C, mixing the digestive mixture sufficiently (shaking incubator at 170-200 rpm) for 2 h (starting the countdown at the time the pancreatin was added).
9. Store digests at -80 °C degrees.

Every food formulation was digested a minimum of 3 times (3 biological replicates) in the case of vitamin E enriched formulations, and a minimum of 6 times (6 biological replicates) in the case of curcuminoid enriched formulations.

3.2.2.1 Regarding gastric lipase

Due to the high cost of the rabbit gastric extract recommended by the INFOGEST authors (around \$60 per digestion) and given that hundreds of digestions would be required for the purposes of the project, it was investigated whether pancreatic lipase could be used instead. No other sources of gastric lipase were available commercially at the time. The objective in this case was to ascertain whether pancreatic lipase retained any significant catalytic activity at a gastric pH, based on the protocol from Laachari et al. (Laachari et al., 2015). To do this, Lesieur Frial frying oil (Asnières-sur-Seine, France) was subjected to pancreatic lipase activity (L3126) from Sigma-Aldrich at different pH conditions. A 50 mM sodium phosphate buffer 50mM at pH 7 (Sigma-Aldrich) was used to simulate the neutral pH and a 50 mM sodium acetate buffer (Sigma-Aldrich) at pH 2 to simulate the gastric conditions. Lipase solutions at a concentration of 11.5 U/mL were prepared using each of these buffers, 3.375 mL were mixed with 125 μL of the oil,

and 20 µl of 20 mM CaCl₂ (product 328507 from CARLO ERBA, Cornaredo, Italy). The samples were incubated in a shaking incubator for 2 hours at 37°C under 170 rpm, simulating gastric digestion. The reaction was stopped by the addition of 1 mL of 6M HCl (Sigma-Aldrich). Diethylether was added (1.125 mL) to increase the volume of the organic phase and make it easier to collect. This phase was again diluted 1:100 and 1:1000 with isopropanol and analysed in an HPTLC-CAMAG machine to quantify tri-, di- and monoacylglycerols and free fatty acids.

Samples were applied to a silica gel 60 F₂₅₄ plate as the stationary phase using ethanol as the solvent, with filling and dosage speeds of 11 µL/s and 100 nL/s, respectively. The HPLTC development was performed in saturated mode with a mobile phase of heptane, diethyl ether, and acetic acid (70:30:1, v/v/v), including pre-drying and post-drying steps. Derivatization was carried out using 8% cupric acid in phosphoric acid, followed by heating at 120°C for 5 minutes. Plate was scanned at a fixed wavelength of 370 nm using deuterium lamp.

Lipase activity was detected at pH 7 in several independent experiments when compared to the control, as can be observed in Figure 26, however activity at pH 2 was not so prominent. Larger peaks of the hydrolysed monoacylglycerols (MAGs) and free fatty acids (FFAs) were observed at pH 7 than in the control and pH 2 conditions, which was coupled by a slightly smaller peak area of triacylglycerols (TAGs) at a 1:1000 dilution (not shown). This was interpreted as proof that pancreatic lipase displayed catalytic activity at pH 7 but not at pH 2. Further experiments would be required to quantify exactly the enzymatic activity at gastric pH so that it could be used appropriately in the protocol, however these results provide some evidence that there is a significant loss of activity at acid pH of pancreatic lipase, which justifies the exclusion of lipases other than costly rabbit gastric lipase for the digestion procedure, in accordance with the recommendations of the INFOGEST paper (Brodkorb et al., 2019).

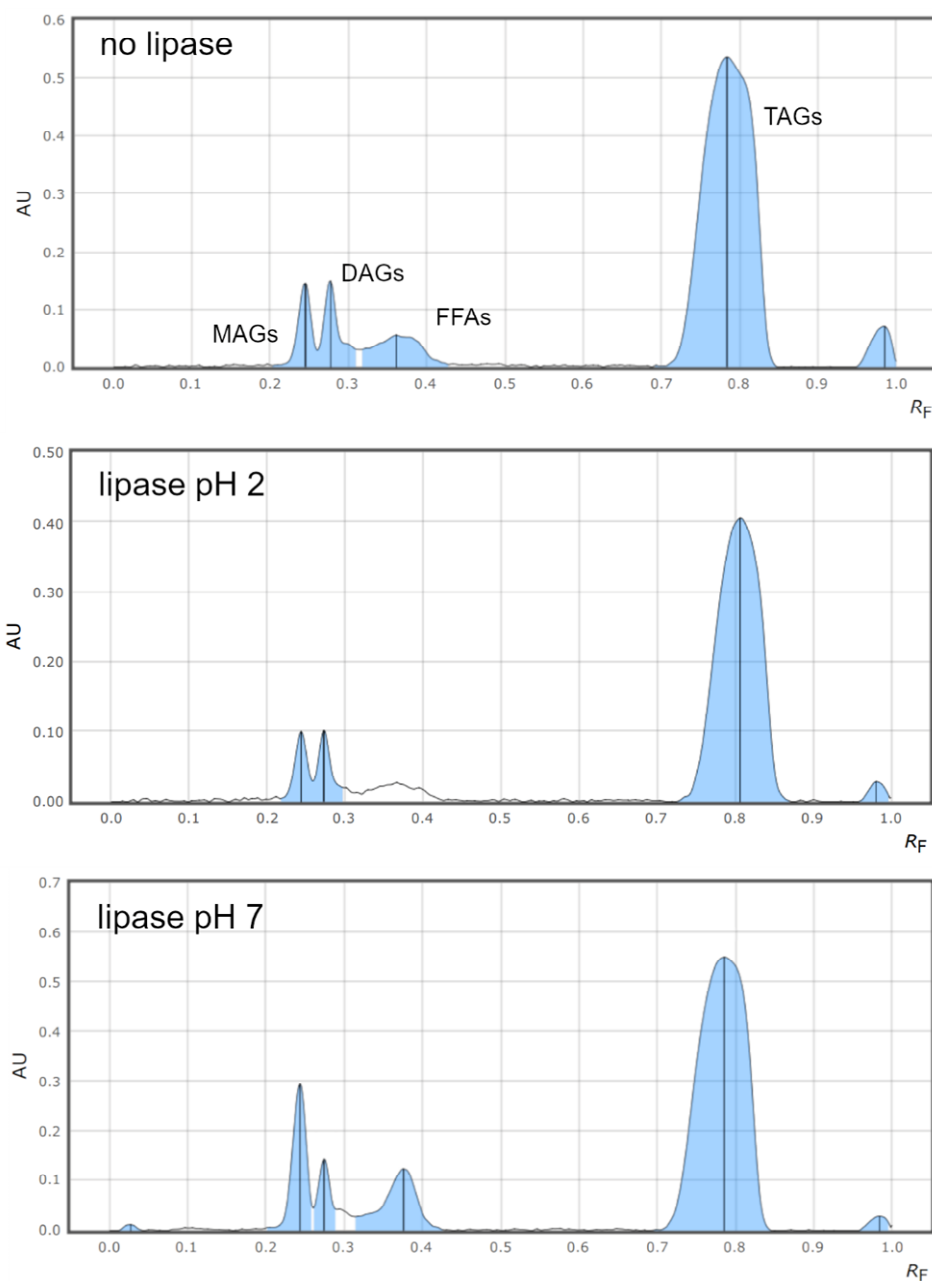


Figure 26. HPTLC chromatogram of olive oil subjected to lipolysis by pancreatic lipase. It can be observed that the peaks of FFAs and MAGs (as a consequence of triacylglyceride digestion) are more pronounced at pHs at which pancreatic lipase retains more enzymatic activity (pH 7).

3.2.3 Bioaccessibility quantification

3.2.3.1 Vitamin E

This experimental protocol presented challenges that restricted screening possibilities. Most of the formulations, especially those including fibre supplements, had the propensity to partly jellify the extracted samples during the quantification procedure. This minimised the reliability of the measurements and precluded testing high degrees of fibre supplementation (higher than 4%) reliably.

Each digesta (the whole sample) as obtained from following the protocol in section 3.2.2, was centrifuged using a Centrifuge 5430 R from Eppendorf (Hamburg, Germany) at 10 °C for 20 min at 2460 x g to separate the bioaccessible phase (soluble in the supernatant) from the inaccessible phase (insoluble in the pellet). Sixteen mL of supernatant were passed into new 50 mL Falcon tubes and the rest was discarded. Only the pellet was conserved in the original tube. The quantification protocol for vitamin E bioaccessibility in each digested sample required using all of its volume per experiment, therefore only one technical replicate per digesta was possible (one supernatant and one pellet).

Extraction of vitamin E from the digestas' supernatants and pellets was performed according to Gutiérrez-Peña et al. (Gutiérrez-Peña et al., 2018), with modifications. Other methods were tested and found to be unsuitable for the task (Stinco et al., 2019) in dairy matrices; the alkaline hydrolysis step is crucial as it removes significant amounts of interferents. These harsh conditions were compensated for by the addition of L-ascorbic acid 0.3 M (A92902), purchased from Sigma-Aldrich (Merck, Saint Quentin Fallavier, France), as co-antioxidant to protect the vitamin E (Chen et al., 2001).

Two mL of the L-ascorbic acid solution was added to the supernatants (16 mL). Pellets, which presented variable mass and volumes in accordance with the formulation digested (some formulations had higher amounts of indigestible, insoluble compounds) invariably received 1.5 mL of the L-ascorbic acid solution and sonicated for 15 minutes in a bath sonicator to facilitate the release of vitamin E from the solid and compacted digestion debris.

Afterwards, alkaline hydrolysis was performed by adding 16 mL of 40% (w/v) potassium hydroxide (221473, Sigma-Aldrich) in methanol to the supernatant +ascorbic solution, and 12 mL to the pellet +ascorbic solution. Both types of samples were incubated simultaneously at 65 °C and 170 rpm for 1 hour in a shaking incubator.

For extractions, a blend of 12:3:5 heptane:acetone:isopropanol (from VWR) was used. Eight mL of this extraction solution were added to each of the samples (8 mL to supernatant mixes, 8 mL to pellet mixes). Next, samples were vortexed and centrifuged at room temperature and 2460 x g rpm for 4 minutes (supernatants) or 20 min (pellets). The organic phases were measured in a Cary Eclipse fluorimeter (Agilent, Santa Clara, California, USA) at 295/325 nm wavelengths (excitation/emission) with quartz cuvettes.

3.2.3.2 Curcuminoids

Curcuminoid bioaccessibility was quantified following the method applied previously to buttermilk (Fu et al., 2015) with slight modifications. This method required smaller volumes of digesta than the vitamin E quantification and so the same digesta could be quantified via several technical replicates. Ethanol and chloroform were purchased from VWR Chemicals (Pennsylvania, USA). The chemicals were >98% pure and were used as received. Ultrapure water (from Milli-Q® system) was used in solution preparation.

Briefly, 0.6 mL of homogenous total digested sample (from the protocol in section 3.2.2) and 0.6 mL of supernatant fraction obtained by centrifuging 1 mL of digested sample at 2460 x g in a Centrifuge 5430 R from Eppendorf (Hamburg, Germany), were mixed (separately) with 1.0 mL of an ethanol/chloroform solution (4:1 v/v) and then vortexed for 1 min. Centrifugation was performed at 1574 x g to separate the total fraction from the supernatant fraction (bioaccessible fraction). Supernatants were collected, and the extraction was repeated on the precipitates once more with 1.6 mL of an ethanol/chloroform/water solution (4:1:3, v/v). Supernatants were collected and pooled with the previous supernatants. Absorbance of the pooled supernatants was measured at 435 nm in an Analytik Jena (Jena, Germany) Specord 205 spectrophotometer. Before screening, the peak absorbance wavelength of 435 nm was determined using

several digested samples and set as the measurement wavelength for the remaining samples, as variations in pH within the range of 5-9 have been shown to have minimal effect on peak absorbance (Wulandari et al., 2020). Concentrations were inferred based on calibration curve, ranging 1-10 µg / mL curcuminoids in the same solvent as the final extractions. Each quantification was performed in duplicate (technical replicates) for each digest (biological replicate) and averaged. A total of over 6 biological replicates for each food formulation were generated in this manner. Bioaccessibility was calculated as (Fu et al., 2015):

$$\begin{aligned} \text{bioaccessibility (\%)} \\ = \frac{(\text{curcuminoids in bioaccessible fraction})}{(\text{curcuminoids in total fraction})} \times 100\% \end{aligned} \quad (3-1)$$

3.2.3.2.1 Curcuminoids quantification by HPLC-DAD

To evaluate whether spectrometry was providing accurate measurements of curcuminoid bioaccessibility, a more specific technique was used to validate the results.

Quantification of curcumin (CUR), demethoxycurcumin (DMC) and bis-demethoxycurcumin (BDMC) was performed by liquid chromatography on an HPLC Agilent 1290 equipped with a DAD detector. The HPLC analyses were performed using a Thermo Hypersil Gold C18 column (100 × 2.1 mm, 1.9 µm, 175 Å) at 35 °C. Eluents A and B consisted of 0.1% (v/v) formic acid in deionized water using a Milli-Q system (Millipore, Molsheim, France) and 100% acetonitrile (with UPLC-MS grade from Biosolve Chimie, Dieuze, France), respectively. The gradient program started with 2% B, then was increased to 98% B over 18 min and held at 98% for an additional 2 min before returning to initial conditions and remaining constant for 3 min. The flow rate was 0.400 mL/min and the injection volume was 5 µL. Detection was monitored at 220, 262, 280, 430 nm and in the range of 210 to 640 nm with 2 nm steep.

Quantification of curcumin (CUR), demethoxycurcumin (DMC) and bis-demethoxycurcumin (BDMC) was performed by liquid chromatography-high resolution mass spectrometry (LC-HRMS) on an HPLC Agilent 1290 with DAD connected to hybrid quadrupole time of flight (Q-TOF) Q-TOF 6538 (Agilent

Technologies, France). HPLC analyses were performed using a Thermo Hypersil Gold C18 column (100 × 2.1 mm, 1.9 µm, 175 Å) at 35 °C. Eluents A and B consisted of 0.1% (v/v) formic acid in deionized water using a Milli-Q system (Millipore, Molsheim, France) and 100% acetonitrile (with UPLC-MS grade from Biosolve Chimie, Dieuze, France), respectively. The gradient program started with 2% B, then was increased to 98% B over 18 min and held at 98% for an additional 2 min before returning to initial conditions and remaining constant for 3 min. The flow rate was 0.400 mL/min and the injection volume was 5 µL. Detection was monitored at 220, 262, 280, 430 nm and in the range of 210 (acetonitrile solvent absorbance cutoff) to 640 (physical limit of detector) nm with 2 nm steep (Jayaprakasha et al., 2002). All compounds response was measured in negative electrospray ionisation mass spectrometry (ESI-MS) and calibrated externally. Limit of quantification of curcuminoids according to calibration curve was determined at <1 mg/L. The ESI Gas Temp was 350 °C, Vcap – 3800 V; the Drying Gas was set at 10 L/min and Nebuliser at 30 psig. The fragmentor was set at 140 V. The HRMS spectrum was registered at 2 Hz in the mass range of 100 to 3000 m/z. MassHunter software (B07) and MSDIAL 4.9 (Tsugawa et al., 2015) were used for data processing.

3.3 Characterisation of curcuminoid-enriched formulations

With the purpose of finding the underlying reasons affecting bioaccessibility across the different formulations and assigning them a numerical value for mathematical modelling purposes, several experiments were carried out to characterise the physicochemical properties of the fibre supplements and the finished products.

3.3.1 Food digestibility assays

The degree of starch hydrolysis after digestion was characterised using the Total Starch HK Assay Kit (K-TSHK) from Megazyme (Bray, Ireland). The degree of proteolysis was assessed using the o-phthalaldehyde (OPA) method, as was done for assessing milk protein proteolysis (Markussen et al., 2021), but using glycine instead for building the standard curve. All the reagents used for these experiments were purchased from Sigma-Aldrich.

3.3.1.1 Starch hydrolysis

Starch hydrolysis was carried out following the manufacturer's instructions for the Megazyme kit K-TSHK. The reagents of the kit included:

- Bottle 1: Thermostable α -amylase [E-BSTAA, 12 mL, 8,300 U/mL on soluble starch (or 1,600 U/mL on Ceralpha reagent) at pH 5.0 and 40°C]. Stable for > 3 years at 4°C.
- Bottle 2: Amyloglucosidase [E-AMGDF, 10 mL, 3,300 U/mL on soluble starch (or 200 U/mL on *p*-nitrophenyl β -maltoside)¹³] at pH 4.5 and 40°C. Stable for > 3 years at 4°C.
- Bottle 3: Buffer (15 mL, pH 7.6) plus sodium azide (0.02% w/v) as a preservative. Stable for > 2 years at 4°C.
- Bottle 4: NADP⁺ plus ATP. Stable for > 5 years below -10°C.
- Bottle 5: Hexokinase plus glucose-6-phosphate dehydrogenase suspension, 2.25 mL. Stable for > 2 years at 4°C.
- Bottle 6: D-Glucose standard solution (5 mL, 1.0 mg/mL). Stable for > 5 years; stored sealed at room temperature.
- Bottle 7: Standardised regular maize starch control. Starch content shown on vial label. Stable for > 5 years; stored sealed at room temperature. The bottle's label indicated lot starch content was 93%.

And the following reagent had to be prepared prior to initiating the analysis:

- Reagent 1: Sodium acetate buffer (100 mM, pH 5.0) plus calcium chloride (5 mM)
 - o Add 5.8 mL of glacial acetic acid (1.05 g/mL) to 900 mL of ultrapure water.
 - o Adjust the pH to 5.0 by the addition of 1 M (4 g/100 mL) sodium hydroxide solution (approx. 30 mL is required). Stable for approx. 2 months at 4°C.
 - o Add 0.74 g of calcium chloride dihydrate and dissolve. Adjust the volume to 1 L and store at 4°C. Stable for > 6 months at 4°C.

As a starch control sample, 177.4 mg of Bottle 7 starch (165 mg of pure starch) was weighed in a 15 mL Falcon tube. This control sample's starch was broken

down according to example A of the manual. The final volume of the sample would be 3.3 mL: a concentration of 5% w/w starch:

1. Add 200 microlitres of ethanol 80% (1 mL ethanol absolute and 250 μ L of water).
2. Vortex.
3. Dilute contents of bottle 1 diluted 1:30 in Reagent 1; 100 mM sodium acetate buffer, pH 5.0; for one sample dilute 120 μ L of bottle 1 with 3480 μ L of reagent 1, for more samples follow this proportion.
4. Immediately add 3 mL of thermostable α -amylase to samples. Stir vigorously to avoid any lumps (starch can form agglomerates easily). Incubate the tube in a boiling water bath for 12 min (stir the tube vigorously every 2 min).
5. Place the tube in a bath at 50°C. Add 0.1 mL of the contents of bottle 2 (amyloglucosidase, 20 U). Stir the tube on a vortex mixer and incubate it at 50°C for 30 min.
6. Transfer the entire contents of the test tube to a 100 mL volumetric flask (with a funnel to assist transfer). Use a wash bottle to rinse the tube contents thoroughly. Adjust to volume with ultrapure water. Mix thoroughly. Centrifuge an aliquot of this solution at 3000 rpm for 10 min. Use the clear, undiluted filtrate for the assay. Typically, no further dilution is required, and a sample volume of 0.05-0.10 mL is satisfactory.

Regarding hydrolysed samples, each sample (6 mL of the digestas proceeding from section 3.2.2 in a 15 mL Falcon tube) was processed according to example E of the booklet (without any milling) which required different steps:

1. Add 5 mL of ethanol 80% and incubate the tube at 80-85°C for 5 minutes. Mix on a vortex and add another 5 mL of ethanol 80%.
2. Centrifuge for 10 min at 1800 g (approx. 3000 rpm). Discard supernatant.
3. Resuspend pellet in 10 mL of ethanol 80% and stir on vortex. Centrifuge as before and carefully pour off the supernatant.
4. Proceed as usual from step 4.

The measuring procedure of the broken-down starch continues as follows (D-glucose standard of bottle 6 was also used as control):

PROCEDURE:

Wavelength: 340 nm
Cuvette: 1 cm light path (glass or plastic)
Temperature: ~ 25°C
Final volume: 2.27 mL
Sample solution: 4-80 µg of D-glucose per cuvette
(in 0.05-2.00 mL sample volume)

Read against air (without a cuvette in the light path) or against water

Pipette into cuvettes	Blank	Sample
distilled water (at ~ 25°C)	2.05 mL	2.00 mL
sample	-	0.05 mL
solution 3 (buffer)	0.10 mL	0.10 mL
solution 4 (NADP ⁺ /ATP)	0.10 mL	0.10 mL
Mix*, read the absorbances of the solutions (A ₁) after approx. 3 min and start the reactions by addition of:		
suspension 5 (HK/G-6-PDH)	0.02 mL	0.02 mL
Mix* and read the absorbances of the solutions (A ₂) at the end of the reaction (approx. 5 min). If the reaction has not stopped after 5 min, continue to read the absorbances at 2 min intervals until the absorbances remain the same over 2 min**.		

* for example with a plastic spatula or by gentle inversion after sealing the cuvette with a cuvette cap or Parafilm®.

** if the absorbance continues to increase, this may be due to effects of colour compounds or enzymes in the sample. These interfering substances may be removed during sample preparation.

Figure 27. D-glucose measurement procedure according to Megazyme's booklet.

Calibration curves are not needed, just the two controls is sufficient. The final concentration of the samples is calculated as follows:

$$c = \frac{V * MW}{\varepsilon * l * v} * \frac{162}{180} * \text{Change in absorbance} [g/L] \quad (3-2)$$

Where ε is the extinction coefficient of NADPH at 340 nm ($\text{mol}^{-1} \text{cm}^{-1}$), l is the length of the light path (cm), v is the sample volume (mL), V is the final volume (mL), MW is the molecular weight of D-glucose (g/mol) and $\frac{162}{180}$ is an adjustment factor from free D-glucose to anhydro D-glucose (as occurs in starch). Change in absorbance is measured against the blank:

$$c = \frac{2.27 * 180.16 * 162}{6300 * 1.0 * 0.05 * 180} * \text{Change in absorbance [g/L]} \quad (3-3)$$

Simplified:

$$c = 1.1685 * \text{Change in absorbance [g/L]} \quad (3-4)$$

The theoretical quantity of total starch in the 6 mL of digesta (approximately 150 mg, depends on the formulation) is compared to the remaining non-hydrolysed starch that is quantified in this experiment.

3.3.1.2 Protein hydrolysis

For quantifying the degree of proteolysis, the OPA assay, which measures released aminoacid equivalents, was used (as a comparative measure of hydrolysis) (Markussen et al., 2021). The reagents for these experiments, *o*-phthalaldehyde (8210270010), trichloroacetic acid (91230), sodium phosphate tribasic dodecahydrate (04277), sodium dodecyl sulfate (SDS) (L4509), and dithiothreitol (DTT) (D0632) were purchased from Sigma-Aldrich. To begin with, the following reagents were prepared:

1. SDS phosphate buffer
 - a. In 500 mL ultrapure water.
 - b. Add 7.5 g $\text{Na}_3\text{PO}_4 \cdot 12\text{H}_2\text{O}$.
 - c. Add 0.5 g $\text{CH}_3(\text{CH}_2)_{11}\text{OSO}_3\text{Na}$ (SDS, 1 g/L).
 - d. Adjust to pH 11 with HCl (around 2.5 mL of 6M).
2. *o*-phthalaldehyde (OPA) solution
 - a. 40 mg in 1 mL of EtOH (0.04 g/mL).
3. DTT solution
 - a. 44 mg dithiothreitol in 1mL of ultrapure water (0.044 g/mL).
4. OPA reagent solution
 - a. Made at time of use (around step 6).
 - b. OPA-solution (1 mL).
 - c. DTT-solution (1 mL).
 - d. SDS Phosphate buffer (48 mL).
 - e. Protect from light and use immediately.
5. Trichloroacetic acid 20%

- a. 20 g in 80 mL of ultrapure water.
- b. Adjust to 100 mL.
6. Glycine standards (explained below).

The protocol was:

1. Obtain 850 μ L volume from total digesta (section 3.2.2).
2. Pre-treat with 20% trichloroacetic acid solution to cause protein precipitation (650 μ L for 8.6%).
3. Centrifuge (17000 g, 20 min, 22 °C).
4. Pipette or pour supernatant into a new tube.
5. Centrifuge supernatant tube (17000 g, 20 min, 22 °C).
6. The supernatants are measured against a standard curve prepared with Glycine (0.1–5mM) and reported as released serine equivalents per mg of protein (standard curve can be used for three days, prepare a 10 mM (75.07 g/mol) glycine stock solution, for example 7.5 mg in 10 mL, and pipette the following for making the calibration stocks in Eppendorf):

Table 6. Volumes for OPA calibration curve.

mM	0	0.1	0.50	0.75	1	2	3	4	5
water	850	841.5	801.50	786.5	765	680	595	510	425
10 mM glycine	0	8.5	48.5	63.5	85	170	255	340	425

7. For the standard curve, pipette 125 μ L of each calibration point into a cuvette and mix with the OPA-reagent-solution, 875 μ L.
8. Also, in a separate Eppendorf, mix 125 μ L of point 0 (pure water) with 875 μ L of OPA solution. It is used as a blank for the samples too and will be diluted in the same way as samples.
9. For the samples, mix in 1.5 mL eppendorfs 125 μ L of each supernatant (after the centrifugation of step 5) with the OPA-reagent-solution, 875 μ L.
10. Incubate 50 min at room temperature in the dark.

11. Dilute (in cuvettes) the point 0 and samples 1:2, 1:4 and 1:5 with SDS buffer. For 1:2 dilution mix 500 microliters of SDS with 500 microliters of sample/point 0, for 1:4 dilution 250:750, and for 1:5 200:800, directly in cuvettes.
12. Measure absorbance at 340 nm, also measure ultrapure water absorbance as blank.

Calculate the released serine equivalents of each sample according to the calibration curve. The negative controls of these experiments (to assess the degree of hydrolysis) are digestas of each tested formulation, in which the digestive enzymes are replaced by enzyme-free solutions (it is assumed that protein hydrolysis will be significantly lower without enzymatic activity).

3.3.2 Oil and water holding capacities of fibres

The protocols were adapted from Chau & Huang, 2003. Oil holding capacity (OHC) of the different fibres was characterised by adding 1 g of each fibre in a separate 50 mL Falcon tube, then adding 8 mL of a mix of colza and sunflower oil, Frial by Lesieur (Asnières-sur-Seine, France), vortexing, and waiting for 30 min at room temperature. The samples were centrifuged for 15 min at 1057 g, supernatant was discarded, the excess of oil on the wall of the centrifuge tube was absorbed using absorption papers, and the tubes were weighed and compared to initial weight prior to adding oil. Each assay was performed in triplicate. The OHC was calculated using the following equation:

$$OHC(g/g) = \frac{m_2 - m_1}{m_0} \quad (3-5)$$

Where m_1 refers to the centrifuge tube; m_2 refers to the mass of the centrifuge tube plus the oil holding sample and m_0 to the dry fibre mass.

Similarly, water holding capacity (WHC) was measured (in triplicate) by mixing 1 g of each fibre source with 30 mL of ultrapure water, vortexing, and left to stand for 16 hours. However, purely soluble fibres completely dissolved within the water so their holding capacity could not be determined. Unretained water was removed from each sample by filtration on polypropylene membranes (0.2 μ m pore size,

GH Polypro, Pall, New York, USA) using a Büchner filtration system (Millipore, Burlington, Massachusetts, USA). The WHC was calculated as follows :

$$WHC(g/g) = \frac{m_2 - m_1}{m_0} \quad (3-6)$$

Where m_1 refers to the mass of the wet membrane, m_2 refers to the mass of the wet membrane plus the oil-holding sample and m_0 to the dry fibre mass.

3.3.3 Bulk and tapped density of fibres

Bulk density was determined by filling a 50 mL graduated cylinder with each fibre source and weighing. The weight of the empty cylinder was subtracted from the total weight. Bulk density was calculated as:

$$bulk\ density\ (g/ml) = \frac{fibre\ weight\ (g)}{50(ml)} \quad (3-7)$$

Tapped density was measured by weighing 10 grams of each fibre source (in triplicate) in a graduated 50 mL cylinder and repeatedly tapping (dropping from a height of 10 cm) the cylinder until the fibre source did not compact any further. The volume was recorded. Tapped density was calculated as:

$$tapped\ density\ (g/ml) = \frac{10\ (g)}{fibre\ volume\ (ml)} \quad (3-8)$$

Both these assays were performed in triplicate.

3.3.4 Biscuit breaking force

Biscuits were previously cut into a 3.5 by 10 cm squircle. Their hardness was measured as the force required to break a biscuit in half by the centre following the three-point break method using a CT3 Texture Analyzer (Brookfield Engineering, Middleboro, Massachusetts, USA) with a TA-7 (knife-edge) probe. The analysis was set at one compression cycle at a speed of 0.5 mm/s. The maximum breaking force in g was recorded and used as a descriptor of biscuit hardness. The trigger force load was 5 g. The maximum load was 4500 g, meaning biscuits that resisted this breaking force would halt the analysis. Measurements were done in triplicate.

3.3.5 Textural analysis of custard

The firmness and stickiness of the custards was determined using CT3 Texture Analyzer with a TA-10 probe. The custards were submitted to one cycle of a 20 mm-distance compression at a speed of 0.5 mm/s. The firmness was determined as the force necessary to attain the 20 mm distance within the food product. The adhesion was defined as the work necessary to overcome attractive forces between surface of the food and materials as the probe retracts. The stickiness is measured as the maximum negative load of the probe after going back up from a 20 mm depth in the food sample.

The viscosity was measured in triplicate with a Lamy RM200 viscometer (Lamy Rheology), equipped with a MK-DIN 12 measuring body at 20°C. Once the mobile MK-DIN-2 entirely plunged into the cream, the measurement was performed with a speed gradient increase from 2 to 100 s⁻¹ for 60 seconds. The gradient of 100 s⁻¹ was maintained for 10 sec before inverse deceleration. The data was obtained from the rheometer's computer software and viscosities evaluated at 100 s⁻¹ were reported.

3.3.6 Fibre colour determination

Fibre colour was assessed with a colorimeter (PCE-CSM 4, PCE Instruments, Meschede, Germany) using both the CIE L*a*b and CIE L*C*h scales. The L value (lightness index scale) ranges from 0 (black) to 100 (white), while a* indicates the red (+a*) or green (-a*) value, and b* refers to the yellow (+b*) or blue (-b*) value. In the CIE L*C*h scale, L represents lightness as in L*a*b scale, C* (chroma) indicates the saturation or intensity of the colour, and h° (hue angle) defines the colour's position on a 360° hue circle, where 0° corresponds to red, 90° to yellow, 180° to green, and 270° to blue. Measurements were done in triplicate. Curcuminoids relative binding capacity of fibres

A quantification assay was devised to gauge the level of curcuminoid entrapment by the different fibres. This assay relies on determining the minimum concentration of curcuminoids in an aqueous solution in contact with fibres necessary to start to detect unbound curcuminoids, thereby elucidating the efficacy of curcuminoid entrapment. This assay cannot be performed with purely soluble fibres.

Briefly, 1 g of each fibre was added in a 50 mL centrifugation tube, with 20 mL of ultrapure water. Concentrated stock solutions of curcuminoids in DMSO (1-10 mg/mL) were added in small and incremental volumes (20 – 500 µL). The tubes were mixed by vortexing and subjected to a shaking incubation (Infors HT Multitron, Infors AG, Bottmingen, Switzerland) lasting 2 hours, at 200 rpm and 37°C in the dark. Afterwards, an aliquot of 2 mL was extracted and centrifuged for 10 min at 17000 g. The supernatant was measured at 435 nm in a spectrophotometer (Analytik Jena Specord 205, Jena, Germany) against a blank of the same fibre source and treatment but with no added curcumin and the concentration of curcuminoids was quantified using a calibration curve, ranging from 5-40 mM curcuminoids in DMSO/water. This same procedure was repeated adding different concentrations of curcuminoids in order to build a curve correlating the total concentration of curcuminoids added to the sample to its respective unbound concentration. The maximum concentration of unbound curcuminoids under the limit of detection (LOD) threshold of 2.5 mM was presented as the relative binding capacity. These LOD determinations were performed in separate experiments; a minimum of 5 different concentrations per fibre per experiment were tested (one replicate per concentration, Figure S37).

3.3.7 Particle size distribution of fibre sources

Fibre sources' particle size distribution was determined using laser light scattering with a Mastersizer 2000 device, with a Scirocco 2000 accessory, from Malvern Instruments (Malvern, United Kingdom). Each measurement was carried out in triplicate.

3.4 Curcuminoid bioaccessibility model development

3.4.1 Data description

A total of 9 fibre supplements and one non-supplemented formulation were tested for two matrix types (biscuits and custards), using 2 technical replicates for each of the 132 biological replicates of digestions of curcuminoid-enriched foods. These would be used as samples in the training and validation of the bioaccessibility model. A total of 29 non-pre-processed candidate predictors of bioaccessibility were collected and are enumerated in Table 7.

Table 7. Candidate predictors of curcuminoid bioaccessibility. $L^*a^*b^*C^*h$ correspond to the axes of CIE L^*a^*b and CIE L^*C^*h colour spaces.

Formulation descriptors	Nutritional composition content	Soluble fibre	Insoluble fibre		Protein	Carbohydrates		Fat	Water	Ash
	Fibre type	Cellulose		Hemicellulose		Dextrin		Pectin		Lignin
Supplement descriptors	Physicochemical properties	Water holding capacity	Oil holding capacity		Bulk density	Tapped density		Particle size diameter	Particle surface area	Curcuminoid binding capacity
	Colour	L		a		b		C		h
Matrix type descriptors	Custards	Viscosity			Firmness				Stickiness	
	Biscuits	Breaking force								

Formulation descriptors were preferable over supplement descriptors for developing a model with wider applicability, but fibre supplement properties were not discarded as the fibre supplements were the main source of variation among the different formulations.

Missing values were due to the impossibility of empirically measuring the property (i.e., water holding capacity of soluble fibre supplements) or because non-supplemented formulations did not have any supplement properties associated to them. All missing values were due to the impossibility of measuring the property in question, such as the liquid holding capacity and curcuminoid binding capacity of soluble fibres, and custards in which viscosity measurement saturated (Table 13). These were removed in analyses or imputed using scikit-learn's IterativeImputer and its default parameters (Pedregosa et al., 2011). This imputer was chosen because it performs multivariate feature imputation by iteratively modelling each feature with missing values as a function of the other features. Consequently, the imputed values should, in theory, not interfere with the subsequent multivariate feature selection and model fitting processes.

3.4.2 Outlier removal

Outlier samples were removed prior to exploratory data analysis and model development. Applying the interquartile range (IQR) criterion for outlier detection (observations that fall outside a 1.5 times IQR distance from the 25th and 75th percentile):

$$x < P_{25} - 1.5 \cdot (P_{75} - P_{25})$$

or

(3-9)

$$x > P_{75} + 1.5 \cdot (P_{75} - P_{25})$$

Technical replicates marked as outliers were removed from the dataset in three sequential steps prior to averaging them into biological replicates. The first step consisted of removing technical replicates with outlying recoveries (according to IQR range); out of 264 technical replicates (2 per each biological replicate), 9 were marked for removal this way. Second, to homogenise replicates factoring in inter-experiment variability, technical replicates with outlying supernatant curcuminoid quantities in each matrix type were also marked for removal (3 biscuit and 2 custard replicates). After performing the removal of previous marks, intra-experiment variability (technical replicates had been processed in batches of 6 per each formulation and experiment) was reduced by removing outlying values only in cases in which the intra-experiment variability (calculated as coefficient of variation) was over 10%. Six replicates were selected as experiment-wise outliers, across 44 different experiments. A total of 20 out of 264 technical replicates were discarded. When averaging pairwise the technical replicates according to the biological replicate they proceeded from, only 1 biological replicate (non-supplemented biscuit) was unable to be computed due to having both technical replicates removed, resulting in 131 biological replicates being recovered.

Furthermore, for correlation and regression analysis, the final dataset was curated from possible outliers with statistical tests. Using a significance level of $\alpha = 0.05$, Grubbs' test (Grubbs, 1950) for outliers was carried out formulation-wise, and despite the limited sample size per formulation, one custard biological replicate (Vitacel WF200 custard) was identified as outlier. As a result, the final dataset was composed of 130 biological replicates out of an original 132 to build the model from: 68 corresponded to digestions of different biscuit formulations, and 62 to digestions of custard formulations.

3.4.3 Feature preprocessing

Properties of the matrix based on content amounts (total fat, total carbohydrate, protein, fibre) were transformed into the ratio of their theoretical weight divided by the empirical total weight of the ingredients. The ratio of macronutrient content was computed as the sum of total fat, total carbohydrate, and total protein ratios. The independent variables were normalised between 0 and 1 for modelling attempts, however for validation experiments the scaler was fitted on the training data and used to transform the data, to avoid data leakage. Biscuit hardness (for biscuits) and custard viscosity (for custards) were combined into a hybrid and generic “*texture*” variable, which could be used in the future to inject texture/rheological data for other matrix types characterised with techniques other than those used in this work.

3.4.4 Interference of colour

To discard the possibility that the curcuminoid quantification protocol (based on colorimetry) was suffering from interferences from the colour of the fibre supplements, correlation studies were carried out and a Machine Learning model of curcuminoid bioaccessibility was fitted using as sole predictors the colour measurements. Furthermore, the best performing models of the project were refitted including colour variables, to evaluate whether these would significantly impact performance. The fitting and validation procedure was similar to what will be covered in section 3.4.6 and 3.4.7, and the results concerning this analysis are shown in 4.3.2. Nonetheless, it is hereby said in advance to justify their removal from future modelling attempts, that it was concluded they were not interfering with bioaccessibility. Spearman’s correlation coefficient (ρ) test was used in the correlation studies, and the model fitted with only colour variables was essentially composed of two independent Bayesian linear regression models (one for custards, another for biscuits) fitted in the framework of a Bayesian hierarchical model, but with no shared parameters.

3.4.5 Feature selection

Fibre supplement’s colour properties were removed from the dataset prior to feature selection, after discarding the possibility they were interfering with the target variable (bioaccessibility) as they were poor predictors of bioaccessibility

on their own nor did they substantially impact the performance of the final model when added as covariates.

Afterwards, several methods were employed to select the most appropriate independent variables explaining the impact of formulation on bioaccessibility. First, Spearman's correlation coefficient was computed for the bivariate comparison between all independent variables and bioaccessibility, without scaling. The presence of a statistically significant correlation between feature pairs was assessed using Spearman's correlation coefficient (ρ) test over Pearson's, as the normality of some variables could not be asserted due to lack of sufficient data, as well as for considering the possibility of non-linear correlations.

Together with the bivariate correlations, wrapper/embedded methods were employed. On one hand, the autofeat (Horn et al., 2020) Python library was used to select the most meaningful features for each matrix type. The Boruta algorithm (Kursa & Rudnicki, 2010), was also used to select features in a separate experiment. Once the final framework for the model was decided (a linear regression Bayesian hierarchical model, details in section 3.4.6) it was fitted to all the features that had been highlighted by Boruta or autofeat; Laplacian priors centred at 0 were placed on each of the regression coefficients to perform feature selection in a manner equivalent to LASSO selection (Tibshirani, 1996).

The most significant features according to the posteriors of the Laplacian prior model were used as basis for the final model.

3.4.6 Model training

A linear regression Bayesian hierarchical model (its principles described in the literature review, 2.7.6.3) was fitted to the data, with partially pooled covariates. Due to the lack of previous domain knowledge regarding a choice of suitable priors for the model, a data-driven choice was made for the selection of features (sections 2.7.8, 3.4.5) and their priors. Macronutrient content was used as a pooled covariate (meaning all the j groups of the model would have the same β for this feature) as it presented a very strong correlation to bioaccessibility across all formulations ($\rho = 0.89$), and ordinary least squares regression

(*bioaccessibility* = *Slope* $\times$ *Macronutrient content* + *Intercept*) produced the μ location parameter values for the normally distributed priors of its pooled intercept and coefficient. A scale parameter $\sigma = 10$ was assigned to allow the inference procedure to explore other values around this μ centre.

With this constraint, the non-pooled covariates (each j group had a different coefficient for each k variable, β_{kj}) were selected by trial-and-error as described in 3.4.5. The final choice of priors is presented in the Results section, 4.3.4. In any case, all tested feature combinations were assigned a normal prior distribution for each β_{kj} , with the hyperparameters being determined automatically by Bambi (Capretto et al., 2022). In mathematical notation, the different models trialled shared the following specification, parting from the linear regression equation:

$$\mu_{ij} = \alpha + \beta_{\text{macronutrient_content}} X_i + \sum_{k=1}^K \beta_{kj} X_{ij} + \varepsilon \quad (3-10)$$

The response variable was assumed to follow a Student's t distribution (likelihood):

$$\begin{aligned} Y_{ij} &\sim \text{Student}(v_j, \mu_{ij}, \sigma_j) \\ \sigma_j &\sim \text{HalfNormal}(\sigma = 1) \\ v_j &\sim \text{Gamma}(\alpha = 2, \beta = 0.1) \end{aligned} \quad (3-11)$$

And the pooled intercept and “*macronutrient_content*” coefficient priors were predefined at:

$$\begin{aligned} \alpha &\sim \mathcal{N}(\mu = 13, \sigma = 10) \\ \beta_{\text{macronutrient_content}} &\sim \mathcal{N}(\mu = 40, \sigma = 10) \end{aligned} \quad (3-12)$$

During the trial and error process of using different non-pooled covariates, all group-specific priors were equally assigned:

$$\beta_{kj} \sim \mathcal{N}(\mu_{\beta_k} = 0, \sigma_{\beta_k} = \text{HalfNormal}(\sigma = 10)) \quad (3-13)$$

The PyMC framework (Salvatier et al., 2016) was employed to carry out parameter estimation and inference using the NUTS method described in the previous section (+ jitter and adapt) with 2000 tuning iterations (the previously mentioned burn-in phase) and 2000 sample draws across 4 chains. No divergences were reported at inference time, at a target acceptance rate of 0.95. The ArviZ (Ravin Kumar et al., 2019) library facilitated pre- and post-*hoc* analysis of the fitted model during the trial-and-error process of evaluating the candidate models.

3.4.7 Model validation and assessment

The procedure of estimating the generalization capacity of the best-fitting models (sections 3.4.5 and 3.4.6) followed the Leave-One-Out Cross-Validation (LOOCV) strategy. In this procedure, the observations of the same formulation were held out as a validation sample and the models were trained on the remaining observations. To test the robustness of the models, all possible combinations were tested; all the formulations were left out individually for a given iteration. This was done with the purpose of visualising the comparison between predicted and observed bioaccessibilities, as well as to compute a practical measure of loss for each model.

In each iteration the test formulation's average bioaccessibility was compared to the formulation's posterior predictive distribution average and 95% minimum width Bayesian credible interval (BCI) (also called highest density interval). The mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R-squared, R^2 and r^2 used equivalently) were used as performance evaluation metrics in this validation case as well.

4 Results

4.1 Food digestibility

Prior to screening, the validity of the digestion protocol was evaluated by preliminary tests in which custard and biscuit samples were subjected to the digestion protocol in order to quantify the degree of macronutrient digestibility. In addition, should several samples present varying degrees of macronutrient digestibility, these values could potentially be informative to predict compound bioaccessibility.

The experiments indicated that no traces of digestible starch were detected in the digestas, so it was inferred the digestions had been successfully completed in that regard and an indicator of a representative simulated digestion. OPA assays showed the released glycine equivalents increased 8-fold after digestion (from 0.8 mmol glycine equivalents per mg of protein to around 6.5), which constituted a similar proteolysis degree to what was expected when compared to the results reported by the protocol authors (Markussen et al., 2021). These two tests suggest the digestion protocol was representative of a real digestion and the results were in line with other reported *in vitro* digestion studies.

However, there was no significant difference in glycine equivalents' release among the subset of formulations tested, hence the degree of proteolysis was not included as a variable of study. The lack of variability in starch and protein digestibility indicated that the degree of macronutrient digestibility would not provide significant predictive power towards bioaccessibility, so lipolysis degree was not assessed either. Fat content was highly homogenous throughout the different formulations (i.e., all custards had similar amounts of fat content, and the same for biscuits) therefore fat digestibility was dismissed as a predictor in the case of this experiment design.

Highlights

- Food digestibility was consistent between foods and considered representative of a real digestion.
- The low variability observed would not help explain the factors governing bioaccessibility.

4.2 Vitamin E

This section provides the measured bioaccessibilities of vitamin E in the formulations described in section 3.2.1, which were digested according to 3.2.2, and for which bioaccessibility was quantified as detailed in 3.2.3.

To obtain a dataset of vitamin E bioaccessibilities in different food matrices, 22 custards and 16 biscuits enriched with vit. E were digested and vit. E bioaccessibility was quantified. The modification or substitution of the base recipe's corn starch, or the supplementation of dietary fibres, provided a variety of formulations (section 3.2.1.1) with which to study the impact of food matrix effects. They all displayed almost equal bioaccessibilities and recoveries (defined as the fraction of the total vitamin E content that was found in the digesta, Table 8 and Table 9):

Table 8. Bioaccessibility of vitamin E in custard and biscuit formulations in which the base recipe's starchy ingredient (60 g of Maizena for custards and 60 g of T45 wheat flour for biscuits) was modified or substituted. Values recorded are the average and standard deviation from 3 biological replicates. *: egg yolks also contain vitamin E (recovery of vit. E from non-enriched formulations was 31% when the usual was >90%)

Matrix type	Formulation	Bioaccessibility (%)
Custard	Base	85 ± 1
	+50% starch	84 ± 1
	-34% starch	86
	-34% starch (no enrichment)*	77 ± 3
	Rice flour	80 ± 3
	Tapioca starch	84 ± 3
	Coconut flour	76 ± 1
	50% coconut flour	86 ± 2
Biscuit	Base	87 ± 1
	50% rice flour	83 ± 2

	50% tapioca starch	79 ± 8
	50% coconut flour	82

Table 10 presents biscuit formulations in which the variations upon the base recipe were miscellaneous in nature. Extremely high amounts of vitamin E (105-210 µg/g instead of the usual 21 µg/g) were administered to attempt to produce a low bioaccessibility formulation (with the idea that vitamin E would saturate and precipitate). Bioaccessibility remained equivalent to other formulations', ranging from 73 to 88%, although the results are hampered by low recoveries, as vitamin E most likely saturated in the extraction solvents.

Table 9. Bioaccessibility of vitamin E in custard and biscuit formulations supplemented with commercial fibre preparations. Values recorded are the average and standard deviation from 3 biological replicates. Higher supplementation levels were not possible in biscuits due to jellification of the extracted samples.

Matrix type	Fibre source	Supplementation (%)	Bioaccessibility (%)
Custard	Orafti®GR (inulin)	2	80 ± 3
		5	84 ± 2
		10	79 ± 3
	Microcrystalline cellulose	5	77 ± 5
		10	78 ± 3
	ApplePRO 60+	2	79 ± 5
		5	76 ± 8
		10	73 ± 2
	Pea Fibre	2	82 ± 6
	Wheat fibre	2	80 ± 4
Biscuit	Orafti®GR (inulin)	2	88 ± 0
		4	87 ± 4

	Microcrystalline cellulose	2	85 ± 7
		4	87 ± 1
	ApplePRO 60+	2	80 ± 5
		4	84 ± 5
	Pea Fibre	2	82 ± 6
	Wheat fibre	2	80 ± 4

Table 10. Bioaccessibility of vitamin E in biscuit formulations at different levels of enrichment and small variations on base recipe. *: base recipe composed of 80 g of sugar, 100 g of egg yolks, 60 g of T45 wheat flour, 10 mL of milk and 106.7 mg/100g of curcuminoid powder and baked at 160° for 30 min

Formulation	Bioaccessibility (%)	Recovery (%)
Base*	87 ± 1	96 ± 10
Additional 10 mL milk and 180°C for 30 min	82 ± 6	88 ± 20
Additional 10 mL milk	76 ± 5	93 ± 14
5X enrichment	78 ± 7	36 ± 10
10X enrichment	84 ± 5	26 ± 5
15X enrichment	87 ± 2	27 ± 5
20X enrichment	85 ± 2	27 ± 3

As all vitamin E formulations displayed almost equally high bioaccessibilities, collecting the nutritional composition and other potential explanatory data from the vitamin E formulations was not carried out as the purpose of modelling the same outcome for different samples was deemed of no significant research value. In the discussion section 5.2 these values are compared to others reported in the

literature. The efforts towards the main aim of the thesis continued around curcuminoids' screening.

Highlights

- Vitamin E was highly bioaccessible in all screened formulations (average >80%), so optimisation potential was low.
- To prove the main hypothesis, an alternative subject compound was required.

4.3 Curcuminoids

4.3.1 Bioaccessibilities

This first section addressing curcuminoids is divided into the validation of the bioaccessibility quantification protocol using HPLC (section 4.3.1.1) and the screening results of several food matrices with the validated protocol (section 4.3.1.2).

4.3.1.1 Validation of the spectrophotometric method to quantify total curcuminoids content

To simplify the evaluation of curcuminoid content, these compounds were quantified using UV spectrophotometry at 435 nm. To validate this approach, samples of biscuits and custards, both with and without curcumin supplementation, were analysed using HPLC-DAD, with curcuminoid identification by mass spectrometry and electrospray ionization (ESI-MS). The UV chromatogram at 430 nm (Figure S35) primarily reflected the presence of the three main curcuminoids: BDMC, DMC, and CUR, consistent with the manufacturer's specifications (Table S17). Peak attribution in the 430 nm UV chromatogram was confirmed via high-resolution tandem mass spectrometry and comparison with the literature (H. Jiang et al., 2006).

To further validate the quantification of total curcuminoids using spectrophotometry, curcuminoid levels in custard and biscuit samples without fibre supplementation were measured by both spectrophotometry and HPLC-DAD. Due to small sample sizes and the bimodal distribution of bioaccessibilities in these samples, the Mann-Whitney U test was employed to determine if the

bioaccessibility measurements from HPLC-DAD (n=8) and spectrophotometry (n=14) belonged to the same population. The test statistic (U) was 52 with a p-value of 0.82. With the significance level set at 0.05, these results indicate no significant difference between the two groups, thus validating the spectrophotometric quantification of curcuminoid bioaccessibility against the HPLC-DAD method. Moreover, HPLC-DAD indicated that the ratios of major curcuminoids in the turmeric powder were 70-73% curcumin, 20-21% demethoxycurcumin, and 7-9% of bis-demethoxycurcumin. These values are in line with the literature (El-Saadony et al., 2022). After digestion, these ratios were similar: 75-77% curcumin, 18% demethoxycurcumin, and 5-7% of bis-demethoxycurcumin.

4.3.1.2 Bioaccessibility in food formulations

To obtain a dataset of curcuminoid bioaccessibilities in different food matrices, a total of 22 custard, biscuit and liquid formulations enriched with curcuminoids were digested and bioaccessibility was quantified according to the previously validated method. Curcuminoid-enriched matrices presented rich variation in observed bioaccessibilities, and this was immediately apparent during experiments, so the variety of formulations tested was more homogenous; fibre-fortified formulations at 6% supplementation. Liquid matrix controls (water and milk) presented bioaccessibilities that resembled closely those of the original protocol, 11.4-26.7% range (Fu et al., 2015) compared to our 12.4-25.9%.

The bioaccessibilities for all formulations tested are displayed in Figure 28 and are in the range of 20.9-38.8 % for custards and 38.8-58.8 % for biscuits. As a general trend, we can observe that the presence of fibre increases bioaccessibility, except for cellulose. Thus, dietary fibre enrichment seems to be a promising way to enhance the curcumin release from the studied food formulations. However, at first glance, no clear trend can be identified based on the nature of the fibre (soluble or insoluble).

To evaluate the assumption of normality of the observed bioaccessibility distributions, the Shapiro-Wilk test was conducted with a significance level of $\alpha = 0.05$. The assumption of normality was not rejected for custard bioaccessibilities ($W = 0.99, p = 0.70$), nor biscuit bioaccessibilities ($W = 0.98, p = 0.26$) when

considered independently, but it was rejected when considering them altogether ($W = 0.93, p = 4.27 \cdot 10^{-16}$) due to the bimodal nature of the bioaccessibilities across all formulations with biscuits having higher bioaccessibilities than custards. Individually, every formulation was also assumed to have normally distributed bioaccessibilities and visual inspection of QQ plots and Shapiro-Wilk tests did not reject the assumption. In any case, homoscedasticity was assumed across all formulations' variances and Levene's test (less sensitive to departures from normality compared to other tests) did not reject this hypothesis ($F(19,110) = 0.99, p = 0.47$).

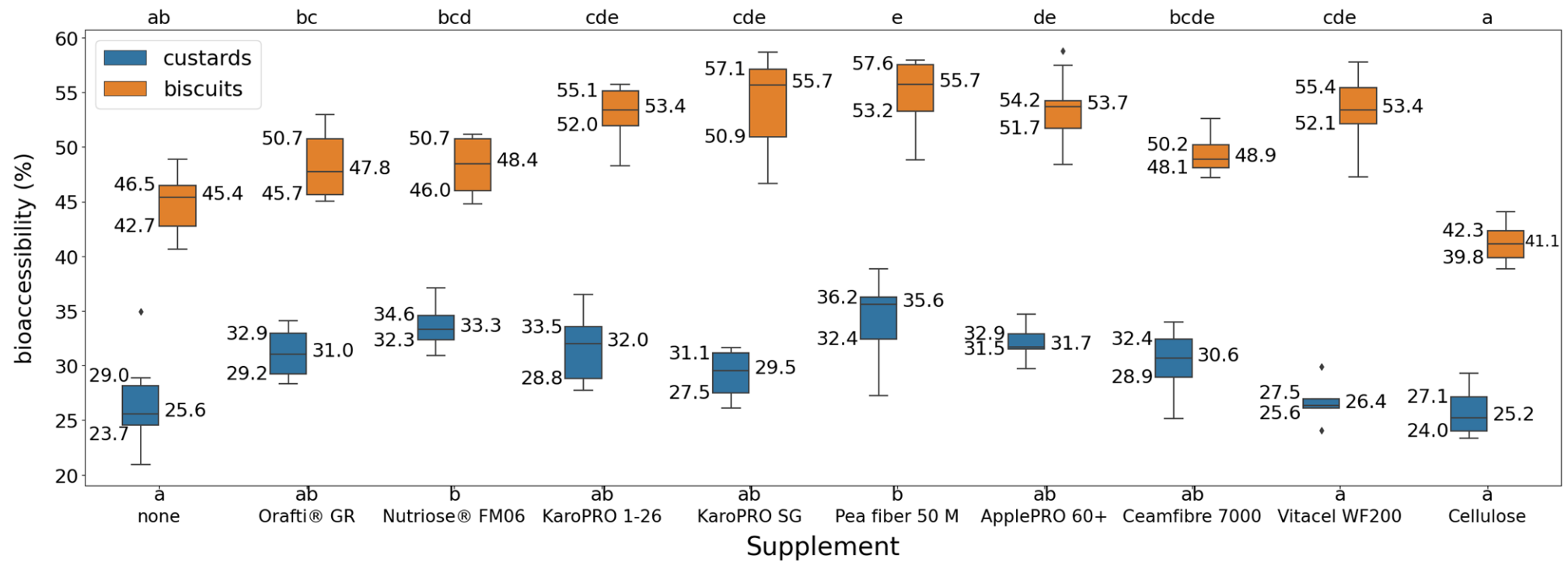


Figure 28. Boxplots of the bioaccessibilities measured for each formulation. The fibres are ranked from the lowest to the highest insoluble fibre content (left to right). Each formulation consists of a minimum of 5 biological replicates after outlier removal. The compact letter display was computed separately for custards (bottom) and biscuits (top). Formulations with different letters presented statistically significant differences to one another (according to Tukey tests).

A two-way analysis of variance (ANOVA) was performed to assess the effects of matrix type and its interaction with fibre supplement on bioaccessibility. The normality of the residuals was visually inspected (Figure 29) and confirmed with Shapiro-Wilk ($W = 0.99, p = 0.61$) as well. There was a statistically significant interaction between the effects of matrix type (custard or biscuit) and fibre supplement ($F(1, 9) = 403.56, p = 1.81 \cdot 10^{-5}$). This interaction had a partial eta-squared of $\eta^2 p = 0.28$, indicating a small effects size on bioaccessibility; within each matrix type the different supplements have a statistically significant effect on bioaccessibility. Simple main effects analysis also showed that matrix type ($p = 1.86 \cdot 10^{-63}$) had a statistically significant impact on bioaccessibility. Matrix type had a larger effects size ($\eta^2 p = 0.92$) than the interaction. Within each matrix type, the specific formulations that presented significant differences were thereafter investigated with Tukey HSD test (confidence level of 95%). The pairwise formulations that presented statistically significant differences to each other were not the same in both matrix types, suggesting the pattern of impact effects diverged according to this grouping. These were included as a compact letter display in Figure 28.

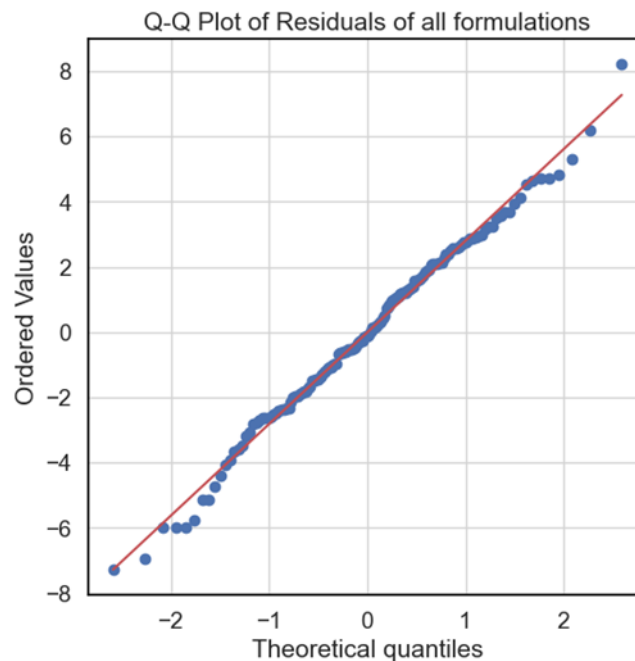


Figure 29. Q-Q plot for evaluating the normality of the residuals across all formulations after 2-way ANOVA.

These preliminary findings offer encouraging results towards the possibility of systematically modelling the effect of food matrix effects on nutrient bioaccessibilities through screening studies. Just by considering the type of food matrix and the usage of supplements, these significant differences, if explained by the known properties of the food, could offer valuable insights that could aid optimising the design of functional foods.

Before further researching the food matrix effects on bioaccessibility, the interference of supplement colour on the spectrophotometric quantification of bioaccessibility was investigated.

Highlights

- Curcuminoids presented a wide range of bioaccessibility in custards and biscuits ($p = 1.86 \cdot 10^{-63}$).
- Biscuits invariably presented higher bioaccessibilities (~39-59%) than custards (~21-39%).
- The ranking of most highly bioaccessible supplements differed in custards and biscuits, possibly due to the food matrix effects impacting each food type differently.

4.3.2 Supplement colour interference

The purpose of measuring the fibre supplement colour was to determine whether it was interfering with curcuminoid bioaccessibility quantification, as curcuminoids absorb light at a wavelength (bright yellow/orange) at which it is possible for undigested fibre supplements to interfere with the spectrophotometric reading (most were light yellow). To evaluate the degree of potential interference, ML models were fitted to the colour properties and their predictive power regarding bioaccessibility was evaluated. Were the colour properties to prove to be non-interfering (poor predictors), this approach would also serve as a benchmark of what predictive accuracies can be expected should the predictors be chosen at random.

The colour measurements in the CIE L*a*b and CIE L*C*h colour spaces of the fibre supplements are recorded in Table 13 of section 4.3.3. The best model using exclusively colour properties is compared to a model using macronutrient content as a pooled feature, and only two other non-pooled features, soluble fibre and hemicellulose contents (Figure 30). Hence, the colour model was specified via 12 parameters (an α_j and five β_{kj} for each matrix type j) whilst the formulation properties model required only 6 (one α , one β , and two β_{kj} for each matrix type j).

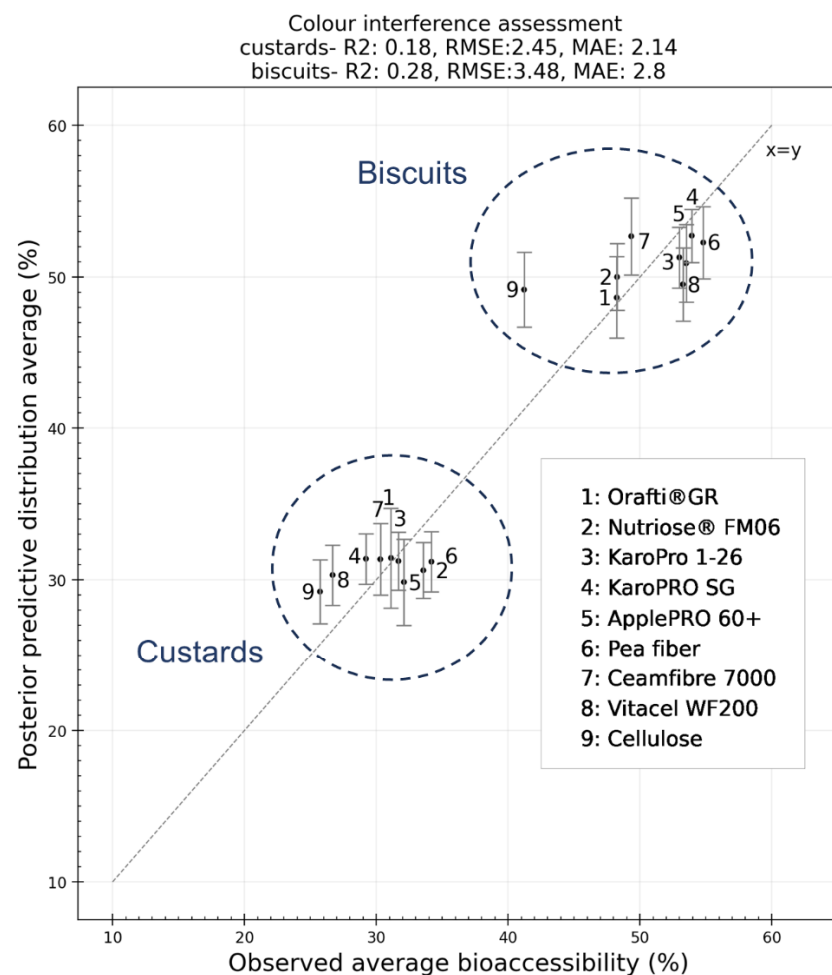
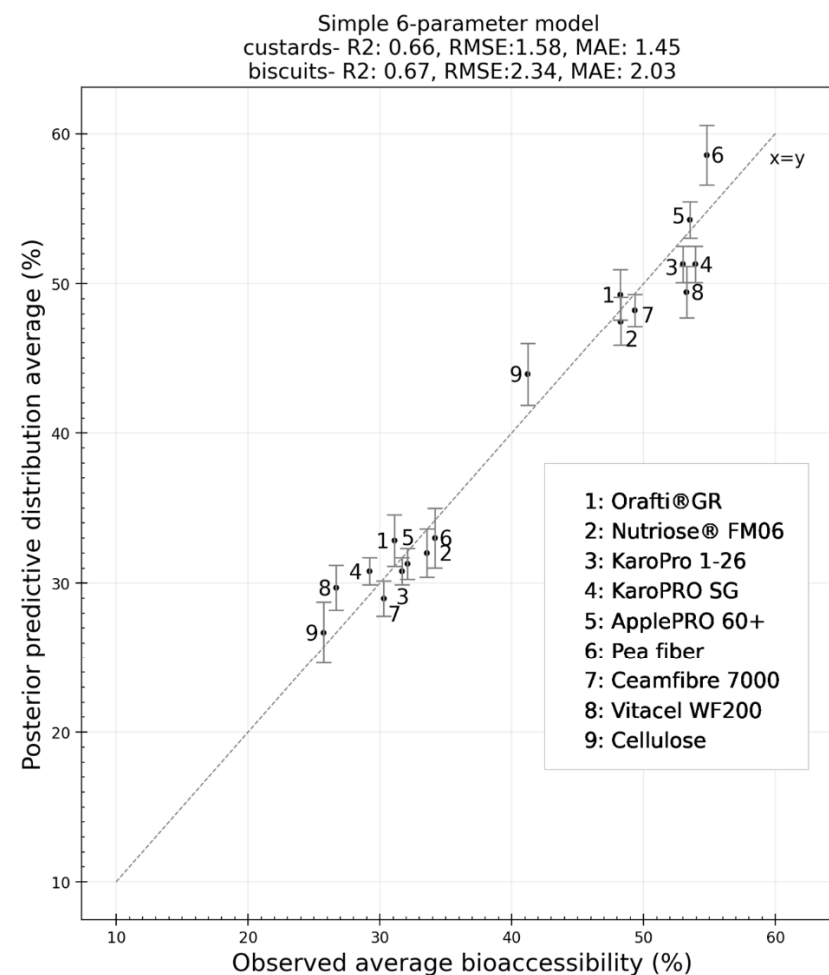


Figure 30. Colour properties (12-parameter) (A) and simple (6-parameter) (B) models' goodness-of-fit scatterplots of observed vs predicted averages (custard bioaccessibilities <40%, biscuits >40%). The intervals shown for each point represent the 95% highest density interval and have been included to assess whether the credible interval correctly captures the observed averages (x=y line).

Figure 30 indicates that the colours of the fibre supplements are inaccurate predictors of curcuminoid bioaccessibility when compared to a much simpler model based on other formulation properties, which is interpreted as a lack of interference with the spectrophotometric quantification of curcuminoids. Even though the MCMC algorithms managed to attribute most of the variability in the data to the colour covariates as the intercept for both groups was centred at 0 (Figure 31), the correlation between observed and predicted values was poor ($r^2 < 0.28$). The importance of each colour variable can be observed by the shape of the linear regression coefficients posterior distributions:

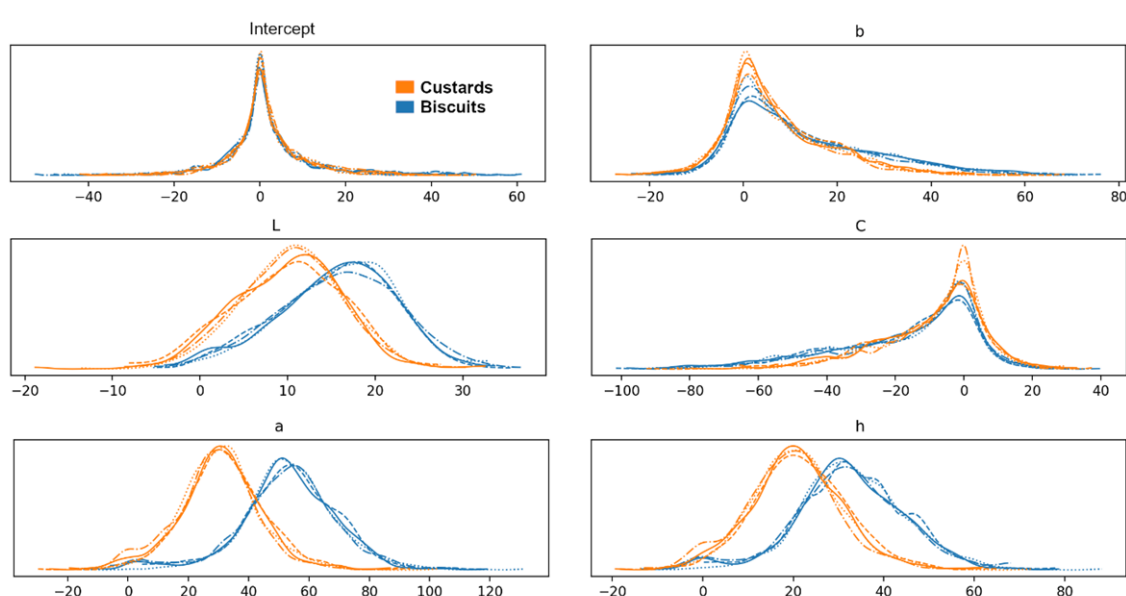


Figure 31. Posterior distributions of the linear regression coefficients of the colours model. Four MCMC chains were used for inference.

Even though there were heavy tails for the *b* and *C* colour dimensions, the peak of their curves was centred at 0 and hence considered as non-significant. Visually, most of the fibre supplements were either white, yellow or brown, and curcuminoids are yellow, which is why there was a suspicion they could be interfering. However, it would seem that the blue-yellow dimension in the CIELAB colour space (*b* dimension) and the chromatic intensity (*C*) (Figure 32) are not that relevant according to this experiment.

The *L*, *a* and *h* dimensions were apparently more useful for modelling in this scenario, with all coefficients being directly proportional to bioaccessibility. It is

possible that measurements in other colour spaces, dimensionality reduction techniques (i.e., PCA) or qualitative descriptors (light yellow, white, dark brown, etc.) could yield better predictions, however other attempts performed worse than the one presented here and did not convey any additional insights so will not be presented.

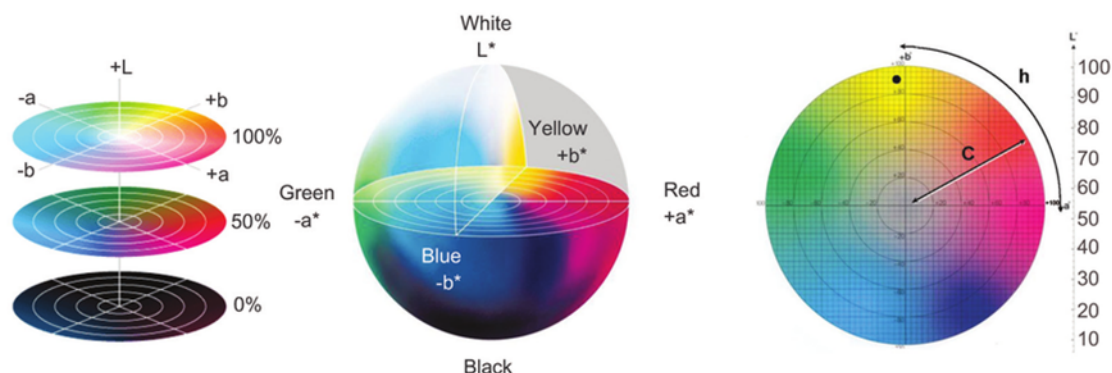


Figure 32. CIE L*a*b and CIE L*C*h colour spaces. [Obtained from Flament et al., 2014]

Another approach to assess whether colour features were interfering with bioaccessibility was to add said features, one by one, to models fitted with properties that had been selected during feature extraction (section 0) and evaluate the magnitude of their impact on model performance. It must be noted that these models were trained including liquid formulations (only macronutrient content was used as predictor), but without including non-supplemented custards and biscuits as they had no supplement colours, therefore reducing the sample size and generalizability of the resulting model. The following Table 11 includes all assayed feature combinations; variables included macronutrient content and soluble fibre content in the 4-parameter model. Supplement colours, hemicellulose content and texture properties were progressively added to the 6- and 8- parameter models:

Table 11. Performance metrics of several Bayesian hierarchical models of curcuminoid bioaccessibility built with different feature combinations (including supplement colours). 4-parameter model uses a pooled intercept and

macronutrient content predictor, and only soluble fibre content as non-pooled predictor.

Model complexity (<i>n</i> parameters)	Additional predictors*	Training performance								
		Custard formulations			Biscuit formulations			All formulations		
		r^2	MAE	RMSE	r^2	MAE	RMSE	r^2	MAE	RMSE
4	-	0.52	1.73	1.88	0.48	2.1	2.96	0.95	1.92	2.48
6	+a	0.53	1.69	1.87	0.49	2.14	2.93	0.95	1.91	2.46
6	+b	0.53	1.67	1.86	0.49	2.1	2.91	0.95	1.88	2.44
6	+C	0.52	1.69	1.87	0.52	2.02	2.83	0.95	1.85	2.4
6	+L	0.51	1.76	1.9	0.53	1.94	2.8	0.95	1.85	2.39
6	+h	0.55	1.62	1.81	0.48	2.13	2.94	0.95	1.87	2.44
6	+hemic.	0.65	1.48	1.61	0.68	2.02	2.33	0.96	1.75	2.00
8	+a, +hemic.	0.66	1.46	1.59	0.68	2.02	2.33	0.96	1.74	1.99
8	+b, +hemic.	0.66	1.45	1.57	0.68	2.01	2.32	0.97	1.73	1.98
8	+C, +hemic.	0.65	1.48	1.6	0.69	1.98	2.29	0.97	1.73	1.98
8	+L, +hemic.	0.71	1.33	1.46	0.7	1.88	2.24	0.97	1.61	1.89
8	+h +hemic.	0.74	1.24	1.39	0.68	2.0	2.32	0.97	1.62	1.91
8	+texture +hemic.	0.71	1.3	1.46	0.78	1.7	1.9	0.97	1.5	1.7

Parting from the 6-parameter model (in bold in Table 11), whose sample-wise performance was also presented earlier in Figure 30, it can be observed that the colour axes that most benefit performance are *L* and *h*. This observation is in line with the insights of the model built with only colour dimensions (Figure 31), although in this case the *a* colour dimension had no significant impact on modelling. However, *L* and *h* provide a smaller performance increase than including texture properties (Table 11). Parting from the 4-parameter model, the

L and *h* features have almost no effect on performance when hemicellulose content is not included, which indicates their performance-enhancing effect in the 8-parameter models could have occurred by random chance.

It must be noted that many divergences were reported during model training when including colour properties, which suggests issues of collinearity or redundancy with these features compared to formulation features (as these presented no divergences). This is evidenced by the spearman ρ correlation coefficients in Table 12:

	Custards				Biscuits			
	<i>Macro.</i>	<i>Sol. fibre</i>	<i>Hemic.</i>	<i>Bioac.</i>	<i>Macro.</i>	<i>Sol. fibre</i>	<i>Hemic.</i>	<i>Bioac.</i>
<i>L</i>	-0.35**	0.40**	-0.35**	0.06	-0.37**	0.38**	-0.37**	-0.18
<i>a</i>	0.47**	-0.40**	0.46**	0.03	0.48**	-0.38**	0.46**	0.27*
<i>b</i>	0.51**	-0.29*	0.38**	0	0.49**	-0.24	0.34**	0.15
<i>c</i>	0.57**	-0.27*	0.38**	0.02	0.58**	-0.21	0.34**	0.2
<i>h</i>	-0.23	0.71**	-0.35**	0.22	-0.25	0.66**	-0.33**	-0.08

Table 12. Spearman ρ correlation coefficients between colour properties and formulation properties used as predictors and the target bioaccessibility outcome.

Only a colour dimension has a slightly significant correlation to bioaccessibility, and only in biscuits. However, all colour dimensions are significantly correlated to a minimum of two formulation properties used as predictors, which explains why so many divergences were observed whilst searching the posterior distribution space (refer to a brief explanation in section 2.7.6.3). Strictly speaking, Spearman's ρ measures the monotonic association rather than the linear correlation, nevertheless the issue of multicollinearity can be highlighted with other metrics; the maximum variance inflation factor (VIF) of the combination of *MacronutrientContent*, *SolubleFibre* and *Hemicellulose* is 8.3, and the

maximum VIF of the combination of *MacronutrientContent*, *SolubleFibre*, and each colour dimension ranges between 12.3 (for *a* dimension) to 63.6 (for *h* dimension). For context, which value to use as VIF threshold is arbitrary and depends on the data in question, but the rule of thumb is that values over 10 indicate collinearity, and notwithstanding the need to be cautious about this threshold (O'brien, 2007), considering only 3-4 variables are being studied this criteria should be acceptable for this case.

The conclusion is that colour features present a moderate degree of collinearity with the formulation properties of interest and have zero impact on the goodness-of-fit when added to the smaller 4-parameter model. In fact, the inclusion of hemicellulose and texture properties increase performance significantly more than the inclusion of colour properties. However, it is noted that including both hemicellulose and texture features increases collinearity above the 10-mark threshold (VIF 17.3-33.7), so including only hemicellulose reduces the chances of producing an overly redundant or overfitting model. The consideration of striking a good balance between under- and over-fitting models will be further explored in section 4.3.4, in which the generalization capacity will also be studied. This section demonstrated that the curcuminoid quantification protocol likely was not measuring the colour of the fiber supplements. Moreover, it establishes a baseline benchmark for prediction accuracies when using inappropriate predictors. Models that account for food matrix effects as predictors should be expected to achieve superior accuracy.

Highlights

- In spite of the experimental design including curcumin-free controls, the colour of the fibre supplements could potentially be interfering with bioaccessibility, as peak absorbance was similar to curcuminoids’.
- Regression models were trained with colour properties and/or other candidate predictors to assess this interference and establish a baseline of model performance.
- Models performed poorly, so this possibility was rejected and colour was discarded as predictor.

4.3.3 Food matrix effects

Several physicochemical properties of the formulations and the supplements used to produce them were investigated to ascertain whether they were correlated with the differences in bioaccessibility. Regarding properties of the fibres, their oil/water holding capacity, bulk/tapped density, curcumin binding capacity in water, particle sizes, specific surface areas, and colour were measured. Likewise, food texture such as biscuit hardness and custard viscosity, firmness and stickiness were assayed. All are recorded in Table 13. These empirical measurements were complemented with theoretical nutritional composition data, such as content of protein, carbohydrate, fat, water, ash, and types of fibre (cellulose, hemicellulose, pectin, lignin, fructans and dextrin) (Table 3).

Table 14 contains the Spearman’s rank correlation coefficients between the candidate explanatory features (excluding colour properties) and bioaccessibility. Factors were classified as matrix properties or supplement properties. Only macronutrient concentrations were considered in the case of comparing across formulations of different matrix types, as fibre contents were the same in custards and biscuits formulated with the same supplement, because 6% supplementation had been used for all formulations. Likewise, texture properties of different types

of matrices were only compared within each matrix type. Fibre supplement properties were measured prior to food preparation. As the pattern of impact effects on bioaccessibility depended on the matrix type, fibre properties had to be inspected independently for each group. None of the correlations of these variables were found to be statistically significant ($p - value = 0.05$) when considering all samples together.

Table 13. Measured physicochemical properties of the supplements employed in this study and their resulting food products. *: Empty cells correspond to non-measurable values: curcuminoid binding capacity and water holding capacity of soluble fibres cannot be measured as the fibres dissolve in aqueous solutions. **: 25000 mPa.s was the saturation limit of custard viscosity measurements.

Supplement	Liquid holding capacities (mL/gram)		Densities (g/cm ³)		Colour measurements					Curcumi- noid binding capacity (mmol)	D [4, 3] – De Bruecker mean (µm)	Specific surface area (m ² /g)	Food product physical properties			
	Oil	Water	Bulk	Tapped	L	a	b	C	h				Biscuit hardness (g)	Custard viscosity (mPa.s)	Custard firm-ness (g)	Custard sticki- ness (g)
Inulin (Orafti®GR)	2.2±0.1	-*	0.60±0	0.65±0.00	93±1	-11±1	19±1	22±1	119±2	-*	121±63	0.15±0.02	1802±711	5383±442	68±22	-33±10
Nutriose FM06	2.8±0.1	-*	0.38±0	0.45±0.00	89±2	6±1	18±0	19±1	73±2	-*	147±15	0.15±0.02	1612±633	4242±679	43±15	-19±6
FST 00007 KaroPRO 1-26	4.3±0.0	8.7±1.7	0.17±0	0.23±0.10	85±1	9±0	24±1	25±1	70±0	100±0	255±3	0.04±0.01	2075±1573	sat.**	318±16	-166±23
FST 00018 KaroPRO SG	3.0±0.2	6.1±0.1	0.33±0	0.54±0.05	65±1	18±0	33±0	38±0	61±0	100±0	37±0	0.28±0.00	2157±422	15381±1467	213±21	-114±14

Pea fibre 50 M	2.5±0.1	4.2±0.5	0.47±0	0.65±0.03	72±3	17±1	28±2	33±2	59±1	40±0	89±2	0.17±0.00	1419±622	sat.**	506±23	-303±24
FST 00224 ApplePRO 60+	2.3±0.1	5.1±0.6	0.41±0	0.69±0.05	47±2	23±1	37±0	43±0	58±1	320±10	65±4	0.41±0.01	1229±229	sat.**	297±14	-135±17
Ceamfibre 7000	4.2±0.1	4.0±0.6	0.21±0	0.30±0.09	60±0	18±0	37±0	41±0	64±0	460±70	262±4	0.03±0.00	3039±1437	sat.**	354±33	-152±2
Vitacel WF200	6.3±0.1	8.2±0.5	0.12±0	0.23±0.10	89±2	5±1	16±1	17±1	72±5	220±30	105±1	0.13±0.00	3028±675	17963±1231	250±17	-135±12
Microcrystalline cellulose	3.0±0.0	2.2±0.1	0.37±0	0.47±0.00	68±1	17±0	25±0	30±1	57±0	260±10	91±1	0.14±0.00	2362±660	13659±1712	166±12	-90±13
none													1709±285	7453±1608	91±55	-60±43

Table 14. Explanatory variables correlation to bioaccessibility and algorithms supporting their inclusion in model. The complete table can be found in appendix A (Table S18). Spearman rho (ρ) coefficients were used as correlation metric. These have been computed grouped by matrix type and for all formulations simultaneously. Properties selected for modelling in one matrix type have been indicated by numbers in parentheses next to their coefficient (1: selected by autofeat, 2: selected by Boruta, 3: final selection by Laplacian priors). Macronutrient content was the only nutritional composition feature considered during selection, as its component features (carbohydrate, fat, protein) presented high collinearity. n.a.: non-applicable. *: p-value between 0.01 and 0.05 **: p-value under 0.01

		<i>All formulations</i>	<i>Custards</i>	<i>Biscuits</i>
Fibre properties	<i>Oil holding capacity</i>	<i>n.a.</i>	−0.39**	0.10 (1)
	<i>Water holding capacity</i>		0.14	0.47** (2)
	<i>Bulk density</i>		0.39**	−0.12
	<i>Curcumin binding capacity</i>		−0.19	0.37 (2)
	<i>Particle size</i>		0.14 (1)	−0.31*
Matrix properties	<i>Biscuit hardness</i>	<i>n.a.</i>	−0.48**	−0.17
	<i>Custard viscosity</i>		0.05	0.53**
	<i>Custard firmness</i>		0.06	0.51**
	<i>Custard stickiness</i>		−0.07	−0.53**
	<i>Cellulose</i>		−0.41** (1)	0.15
	<i>Hemicellulose</i>		0.12	0.66** (1,2,3)
	<i>Insoluble fibre</i>		−0.31	0.14 (1)
	<i>Soluble fibre</i>		0.38** (3)	0.09 (1,3)
	<i>Protein</i>	0.89**	0.36**	0.72**
	<i>Carbohydrate</i>	0.89**	0.46**	0.65**
	<i>Fat</i>	0.84**	0.12	0.52**
	<i>Water</i>	−0.62**	0.30*	0.72**
	<i>Ash</i>	0.72**	0.23	0.70**
	<i>Macronutrient</i>	0.89**	0.47** (1,2,3)	0.66** (2,3)
Feature-engineered matrix properties	<i>Macronutrient</i>	0.89**	0.47** (1,2,3)	0.66** (2,3)

As had been evidenced by ANOVA tests, the matrix type had a significant impact on bioaccessibility, and this was captured by very strong correlations ($\rho = 0.89$) between bioaccessibility and macronutrient content (sum of protein, carbohydrate, and fat ratios). These correlations were also relatively strong when inspecting the data within each matrix type ($p - value < 0.05$), but not to the same degree.

Many more properties showed a statistically significant correlation in biscuits than in custards, and those correlations were stronger as well (refer to Table 14, which is a shortened version of the appendix table Table S18). This finding, together with their more pronounced inter-formulation bioaccessibility differences (as stated by the Tukey tests and its derived CLD in Figure 28), suggested that biscuit samples' structured data patterns were *a priori* more amenable for predictive modelling. However, as no single strongly-correlated ($\rho > 0.80$) property to bioaccessibility within each matrix type was found, an approach to modelling that could consider the complex variance both intra- and extra-group would have to be followed.

Highlights

- Macronutrient content was very strongly associated to bioaccessibility across all samples ($\rho = 0.89$).
- Considering custards and biscuits separately, the variables of study presented associations to bioaccessibility that were weaker and discrepant.
- The content of soluble fibre in custards, and hemicellulose content in biscuits, were highlighted as potential predictors of bioaccessibility.

4.3.4 Machine Learning model

Several statistical techniques and data modelling workflows were tested to determine which was optimal for the regression task of predicting bioaccessibility. For brevity, only models trained following a Bayesian hierarchical linear

regression scheme will be shown. This algorithm struck a good balance between allowing the study of several factors in both matrix types simultaneously whilst providing good prediction performance.

Table 15 presents performance statistics of models fitted over different feature combinations. The assessment of their bias-variance trade-off was based on comparing their predictive performance on the training data and on samples that had been left out during training (under a LOOCV sampling scheme).

Table 15. Performance metrics of several Bayesian hierarchical models of curcuminoid bioaccessibility built with different feature combinations.

Model complexity (<i>n</i> parameters)	Additional predictors*	Training performance									LOOCV performance								
		Custard formulations			Biscuit formulations			All formulations			Custard formulations			Biscuit formulations			All formulations		
		r ²	MAE	RMSE	r ²	MAE	RMSE	r ²	MAE	RMSE	r ²	MAE	RMSE	r ²	MAE	RMSE	r ²	MAE	RMSE
2	-	0.31	2.04	2.33	0.53	2.12	2.92	0.94	2.08	2.64	0.17	2.23	2.56	0.46	2.31	3.13	0.93	2.27	2.86
4	+ soluble fibre	0.57	1.74	1.85	0.53	2.13	2.91	0.95	1.93	2.44	0.27	2.21	2.40	0.41	2.52	3.28	0.93	2.36	2.88
6	+ hemicellulose	0.70	1.37	1.53	0.73	1.87	2.23	0.97	1.62	1.91	0.34	2.00	2.29	0.38	2.77	3.36	0.93	2.38	2.88
8	+ food texture	0.73	1.27	1.47	0.80	1.59	1.6	0.97	1.43	1.70	0.29	2.00	2.37	0.37	2.81	3.39	0.92	2.40	2.92

4.3.4.1 Bias-variance tradeoff and selection of candidate models

Models built with several feature combinations were evaluated based on prediction performance. The models were well fitted to the data and provided valuable information regarding which factors could be affecting bioaccessibility the most. Table 15 compiles the performance metrics of models with different levels of complexity.

The simplest model only includes macronutrient content as a pooled feature, due to its strong correlation to bioaccessibility. By using only this predictor and an intercept, all formulations were well fitted ($r^2 = 0.94$), however the performance in each matrix type was improvable, with an average r^2 of only 0.44.

In order to improve the performance of the model within each matrix type, additional variables with which to train the algorithm were considered. With the addition of soluble fibre as a non-pooled covariate (4-parameter model) the goodness of fit was improved (average r^2 of 0.55 for both matrix types), however it was with the additional inclusion of hemicellulose (6-parameter model) that it was considered optimised (average r^2 of 0.72). The addition of a third non-pooled feature (texture properties, 8-parameter model) only increased the goodness-of-fit r^2 to 0.77 and other non-reported combinations did not perform significantly better.

The improvement in the goodness-of-fit by including extra predictors was substantial. In cases such as this, the chance of producing an overfitting model grows with each predictor. The overfitting phenomenon can be evaluated with cross-validation experiments such as LOOCV. In this case, the improvement in the goodness-of-fit outweighs the slight reduction in generalization capacity from including additional predictors. It can be observed that the 2-parameter model presented the lowest LOOCV error across all formulations (MAE of 2.27). Considering this model as the baseline, the addition of non-pooled covariates improved predictions in the custard group (0.9%-10.3% MAE improvement) but reduced them in the biscuit group (9.1%-21.6% MAE deterioration) and overall, the generalization capacity was slightly reduced (4.8%-5.7% MAE deterioration).

On the other hand, goodness-of-fit improved between 7.2%-31.3% across all formulations, especially for custards (up to 37.7% improvement).

These results indicate that, for the purposes of training a model with the highest generalization capacity, a simple linear regression model using macronutrient content as sole predictor is the preferable option. If the goal is to attempt to ascertain the factors most influencing bioaccessibility in each matrix type, a Bayesian hierarchical model offers flexibility in studying the most relevant features in each group whilst providing a measure of uncertainty of the importance of the predictors. In this case, the 6-parameter model was considered to best balance prediction accuracy in both training and validation samples. A more detailed representation of the goodness-of-fit and a simulation of the generalizability of the 6-parameter model can be visualized in 4.3.4.3.

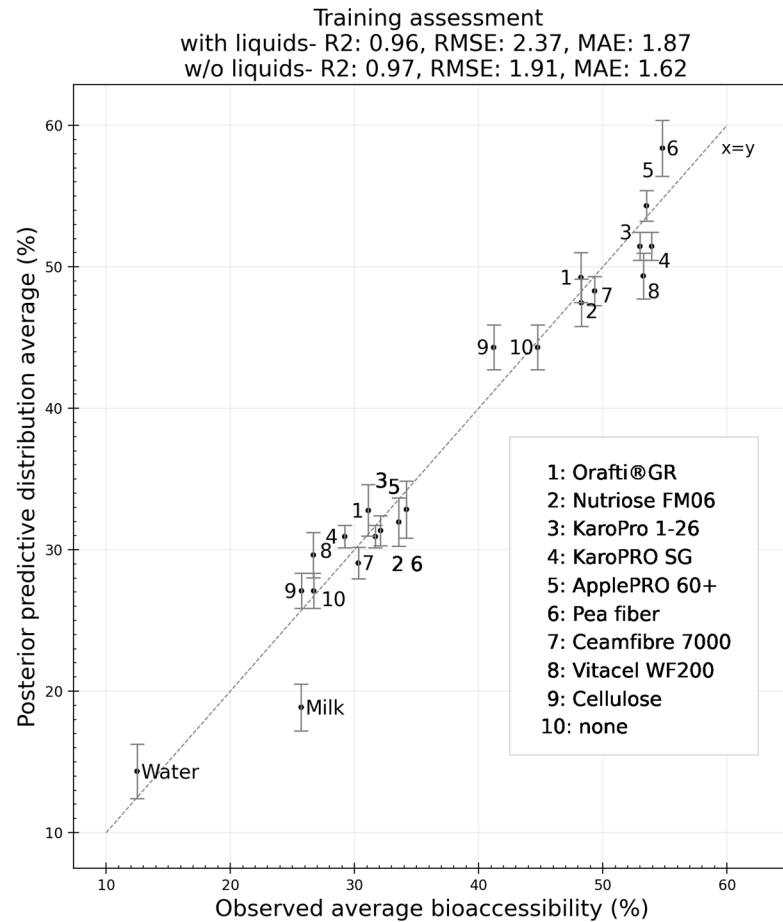
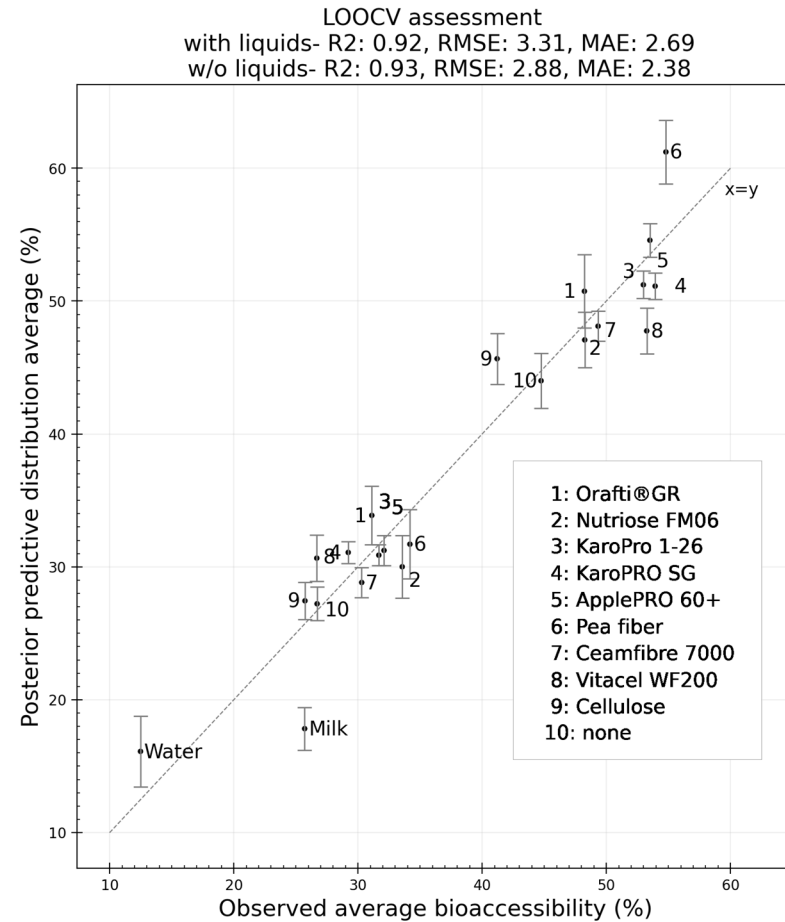


Figure 33. Training (left) and leave-one-out cross validation (right) assessments of the goodness-of-fit and generalization capacity of the final model (6-parameter). Each dot corresponds to the average observed bioaccessibility of each formulation and is compared to the average of its posterior predictive distribution. The intervals shown for each point represent the 95% highest density interval and have been included to assess whether the credible interval correctly captures the observed averages ($x = y$ line).

4.3.4.2 Modelling imbalanced dataset including liquid matrices

As indicated in Table 15's caption, liquid matrices were included in the model in spite the fact that only two liquid formulations had been tested and showcase how the usage of macronutrient content as a predictor was capable of casting reasonably accurate predictions of formulations other than custards or biscuits, even for imbalanced datasets.

When excluding liquid matrices from model fitting, error metrics were reduced by 14% on average, as liquid matrices could still influence the inferential procedure of MCMC.

A more detailed representation of the goodness-of-fit and a simulation of the generalizability of the final model (the 6-parameter) can be visualized in Figure 33. Milk matrix was the most inaccurately fitted and predicted, although it is likely that once more liquid matrices are characterised, a more accurate regressor for this group can be developed, within the boundaries of the hierarchical model. In conclusion, these results suggest the potential applicability of this modelling setup to other matrix types.

4.3.4.3 Assessment and interpretation of candidate predictors

Regarding feature selection (Table 12), autofeat and Boruta determined that macronutrient content was sufficient to fit an accurate model in custards, although the particle size and cellulose features could assist in the case of linear modelling. In biscuits, it appeared that more features were necessary for accurate modelling, and autofeat concluded macronutrient content could even be substituted for a number of other features instead (a number large enough that the model would likely be overfitting). For the final feature selection using Laplacian priors, which was the configuration most similar to the final models presented in this work and therefore the most reliable in this context, the features that were relevant for both groups were those that had only been preselected in biscuits, although soluble fibre presented a significant correlation to custard bioaccessibility ($\rho = 0.38$, p – *value* < 0.01) which supports its importance. On the other hand, Table 15

suggested that it was the out-of-sample prediction of custards which most benefitted from the inclusion of additional covariates.

Table 16. Posteriors of the 6-parameter model. HDI: highest density interval 3% to 97%. The input data was normalised (scaled between 0, minimum; and 1, maximum) for each matrix group separately, except macronutrient content which was normalised across all samples.

Parameter		Mean $\pm$ std	HDI lower	HDI upper
Intercept		Table 14	12.5	16.1
Macronutrient content		37.8 $\pm$ 1.7	34.9	41.3
Soluble fibre	<i>Custards</i>	5.1 $\pm$ 1.2	3.0	7.4
	<i>Biscuits</i>	3.2 $\pm$ 1.3	0.9	5.5
Hemicellulose	<i>Custards</i>	3.1 $\pm$ 1.2	0.8	5.4
	<i>Biscuits</i>	6.2 $\pm$ 1.6	3.0	9.2

Soluble fibre content was the non-pooled predictor that the Laplacian prior model attributed the highest predictive power across both groups to, and this is reflected in the 6-parameter model's posteriors (Table 16). Hemicellulose presented the strongest association to bioaccessibility ($\rho = 0.66$, $p - value < 0.01$) of all non-pooled predictor candidates, and this is also reflected by the model (6.2 $\pm$ 1.6 posterior). Every predictor presented a positive proportional association to observed bioaccessibilities and prediction outcomes.

As was mentioned in section 4.3.2, the features used for the 6-parameter model presented low multicollinearity as determined by variance inflation factor (maximum VIF 8.3) but including texture properties increased collinearity by this metric (maximum VIF 33.7) hence it was another reason for not including them in the final model. This variable was not selected by the feature selection algorithms, although it was hypothesised it could have been a valuable predictor, and it did

provide more explanatory power than colour properties (used as reference in section 4.3.2). The degree of association between these and other candidate predictors can be inspected in Figure 34.

A heuristic interpretation of the fitted models is that increasing the concentration of nutrients in the formulation is the most significant factor leading to higher bioactive release from the matrix, and although there are interactions depending on the type of fibre being supplemented, the effect is small, as was observed during exploratory data analysis of bioaccessibilities in section 4.3.

These findings suggest there is an underlying complexity or interaction between the factors of the study, however the macronutrient content's strong association to bioaccessibility is remarkable and provides great explanatory power. As a result, in this section it has been proven that the insights can potentially be applied to fit predictive models.

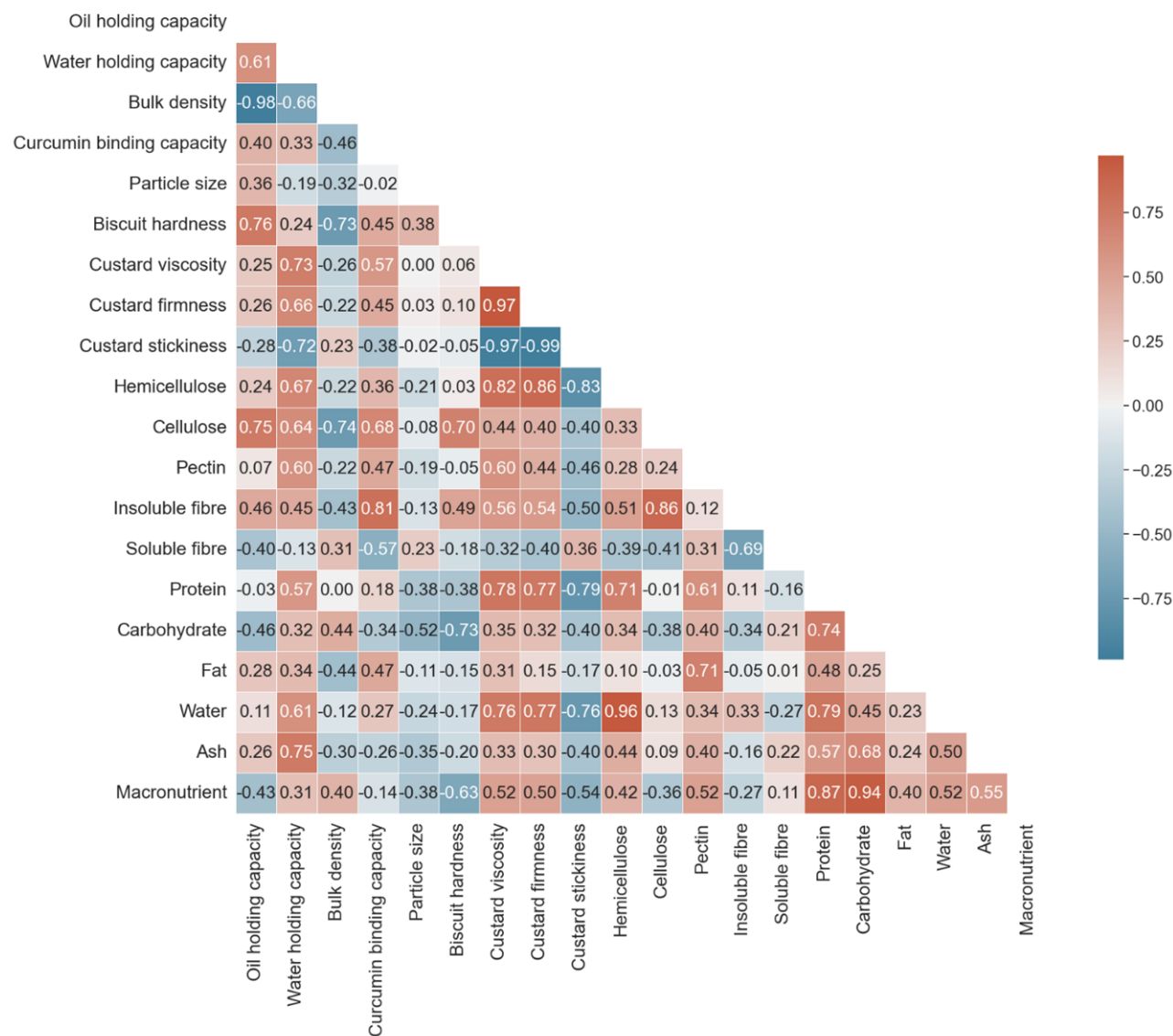


Figure 34. Spearman correlations between the formulation properties (excluding non-supplemented formulations due to missing values). It can be observed that none of the predictors of the 6-parameter model (macronutrient content, soluble fibre and hemicellulose) are strongly correlated to each other ($\bar{\rho} > |0.5|$). Texture properties (8-parameter model) present stronger correlations to macronutrient content and hemicellulose.

4.3.4.4 Assessment and interpretation of predictions

Figure 33 indicates that although the 6-parameter model has a good fit to the training data ($r^2 = 0.97$) and formulations were acceptably inferred in the LOOCV experiment, with most formulations' credible intervals capturing the observed average or presenting a relative error lower than 10%, some of the formulations' observed bioaccessibilities have not been properly captured. The clearest outliers were the biscuits formulated with pure or high in insoluble fibre supplements (microcrystalline cellulose and Vitacel WF200) and pea fibre. None present soluble fibre; hence their predictions are only affected by macronutrient and hemicellulose content. As Vitacel WF200 contains hemicellulose whereas microcrystalline cellulose is pure cellulose that does not contain hemicellulose, Vitacel WF200 has correctly been attributed a higher bioaccessibility than microcrystalline cellulose, nevertheless the predictions of these three could be improved. The fact that hemicellulose presents a high correlation to biscuit bioaccessibility while insoluble fibre as a whole does not, suggests the correlation might be spurious, in which other factors might be influencing with a higher impact, but this is not observable when analysing bivariate correlations in this manner. There is plausibility for the existence of complex interactions occurring in the food matrix during digestion.

Nutriose custard was also poorly predicted in the LOOCV. Nutriose is a pure soluble fibre supplement akin to inulin; both display a moderate bioaccessibility in biscuits but nutriose presents a slightly higher bioaccessibility in custards. Inulin is soluble fibre composed of fructans (polysaccharides of fructose) while nutriose is composed of resistant dextrin. The inclusion of dextrin covariate in the model could resolve this bias error, however due to the highly sparse nature of this feature (only nutriose has dextrin) it can be considered as an identifier rather than an informative feature and it was not included to favour the generalisation capacity.

Although the features used in the model are features that can be applicable to many formulations, it is possible that in the future, additional data can provide more insights into more appropriate features for the model, such as properties of the intermediate state of the food (i.e., biscuit dough texture), or the possibility of employing features that currently are sparse due to the limited number of formulations tested. Similarly, the need to prepare the formulations to include their texture properties in the model could be replaced by other features that do not require cooking the formulation beforehand. This will be addressed in more detail in the discussion.

Highlights

- The model trained only with macronutrient content provided accurate predictions across all samples.
- The Bayesian hierarchical regression model enabled the study of additional predictors affecting bioaccessibility in each matrix type separately.
- Predictions for liquid matrices were promising, suggesting the model could be efficiently expanded to include other matrix types.
- The model estimated most formulations accurately, although more data will be required for validation.

5 Discussion and conclusions

In this final section of the thesis, the most important findings and limitations of the project will be discussed in detail, and any significant results and conclusions will be compared to works of other researchers. The main focus will be on whether there is agreement between the insights presented in this work with those of the literature and identifying any novel discoveries that have not been reported in peer-reviewed journals.

5.1 Systematic screening and modelling approach

The main aim of this thesis was to challenge the hypothesis that it was possible to accurately predict the bioaccessibility of a bioactive compound as a function of the properties of the food matrix carrying it (section 3.1).

The results showcase the potential usefulness of modelling bioaccessibility screening experiments of different food formulations. Curcuminoids exhibited great potential for optimization, with bioaccessibilities ranging from as low as 12.4% in liquid matrices to almost 60% in biscuits. The known properties of the food matrix were therefore used to attempt to infer the bioaccessibility of curcuminoids. The test results indicated that, although the validation strategy used to test the aptitude of the model's predictions is optimistic, the majority of the formulations, with some exceptions, were estimated acceptably with a relative error lower than 10% (Figure 33).

Regarding the approach to modelling, the Bayesian methodology allowed for the incorporation of uncertainty when studying the factors that most impacted bioaccessibility. This approach provided flexibility to test several combinations of explanatory causes of variability while accounting for the plausibility of the hypothetical model. Furthermore, fitting a model with a hierarchical structure accommodated the possibility that the factors impacting bioaccessibility differ depending on the nature of the food matrix. In detail, variables which appeared to play an important role in one food matrix, such as soluble fibre content in custards ($\rho = 0.38, p - value < 0.01$), could in contrast have minimal impact in

biscuits ($\rho = 0.09, p - value > 0.05$). Nevertheless, by inferring a model for both matrices separately whilst still considering factors that impact them in a pooled manner, such as total macronutrient content ($\rho = 0.89, p - value < 0.01$), the modelling algorithm determined that soluble fibre was a significant predictor for both matrix types (Table 16), albeit the exceptionally strong correlation to macronutrient content appeared to be the driving factor affecting curcuminoid bioaccessibility across all formulations, with its posterior linear coefficient $\beta_{macronutrient_content}$ displaying an additional order of magnitude over non-pooled parameters.

The predictive power of the macronutrient content, plus the flexibility conferred by the hierarchical approach in that additional predictors can be modelled independently for each matrix type using Markov Chain Monte Carlo bootstrapping, is a promising starting point for the study of formulations of custards and biscuits and, as will be discussed in more detail in the following sections, will be of importance for modelling other matrix types.

To summarise, the fact that increased macronutrient concentrations, and consequentially the very nature of the food matrix itself, was very strongly correlated to an increase in curcuminoid bioaccessibility enabled accurate predictions of curcuminoid bioaccessibility as a function of food matrix effects for formulations of more than one matrix type simultaneously.

5.1.1 Food matrix effects

In this case study, the strong correlation of macronutrient content to bioaccessibility was used as a data-driven assumption to replace weak domain knowledge, stemming from a lack of consensus in the literature regarding the magnitude of the correlations of food matrix effects to bioaccessibility, not only for curcuminoids but for bioactive compounds in general. In our case, the hierarchical modelling setup offered flexibility in studying the most relevant features influencing bioaccessibility in each matrix type. Nevertheless, the results indicated that a simple linear regression model using macronutrient content as sole predictor could alternatively be a preferable option (Table 11).

Some works examining the relationship of the candidate predictors of our study with bioactive compound bioaccessibility exist, but the conclusions are inconsistent. To preface the comparison of our findings to those of other authors', a number of considerations introduced during the literature review that hinder the analysis are hereby revisited:

- The physicochemical properties of the bioactive compound heavily influence its food matrix effects, so factors affecting the bioaccessibility of lipophilic compounds (such as curcuminoids) may not be equivalent to the case of hydrophilic compounds, and vice versa.
- Bioaccessibility does not always correlate with bioavailability, especially for static *in vitro* digestions which usually do not include gastric emptying and gradual acid and enzyme secretion. This reduces the reliability of extracting bioaccessibility insights from bioavailability studies and vice versa.
- Experimental protocols (digestion and quantification) differ, and this hampers confident comparison. The 2014 and 2019 INFOGEST consensus protocols managed to standardise *in vitro* digestion, but most studies have not adopted said recommendations or pre-date them.
- Even though custards and biscuits are popular functional food types, there is a significant lack of food studies compared to models of oil-in-water emulsions.

Keeping these considerations in mind, the food matrix effects believed to be affecting bioaccessibility that were analysed in this case study will be further discussed.

5.1.1.1 Type of matrix

We observed that bioaccessibility was inversely correlated with the flow capacity of the food matrix structure, which is dictated by macronutrient concentration. Liquids presented the lowest bioaccessibilities, followed by custards and finally biscuits. These findings indicate that further investigation into the impact of matrix type on digestion, such as monitoring the process at each stage, could be

beneficial. Even if the final degree of digestion is similar across food products, differences in the rate of digestion could account for discrepancies in bioaccessibility: for example, variations in the size of biscuit crumbs formed during mastication could lead to different nutrient release profiles.

Contrasting with our results, an *in vivo* study with 12 human male volunteers discovered that folic acid (water soluble compound) was affected by food matrix structure as well, but in this case formulations with extreme properties (custard and biscuit) had lower folate plasma availability than intermediate matrices (sponge cake and pudding) (Buffière et al., 2021), therefore the correlation between matrix structure properties and bioaccessibility appeared to be non-significant.

In a similar vein, an in-depth experiment (Hiolle et al., 2020) was carried out with an *in vitro* model following INFOGEST recommendations of the same matrix types and target micronutrients (folate and lutein, which is lipophilic). Researchers concluded that the release of the micronutrients greatly depended on the matrix structure, which impacted the digestion process itself, by characterising the differences not only at the end of digestion but also at each phase (oral, gastric and intestinal). Macronutrient hydrolysis kinetics were not impacted by the different matrix structures, except for a 4-fold increase in lipolysis rate for biscuits compared to custards that coincided with an increase in lutein bioaccessibility in custards, a finding that contradicts the existing consensus (presented in the literature review, section 2.6) that increased lipid digestion improves bioaccessibility of fat-soluble compounds. On the other hand, folate bioaccessibility was impacted during the gastric phase only, and in the end both lutein and folate exhibited relatively equal final bioaccessibilities across all matrix types, except for the previously mentioned lutein in custards.

Many interesting conclusions can be drawn from the comparison to these two experiments' results:

1. They suggest that observing a significant correlation between macronutrient content and bioaccessibility is a novel finding.

2. As Hiolle noted, the matrix structure itself appeared to be impacting the digestion process. We seem to agree in that regard; our variables of study presented different degrees of correlation to bioaccessibility depending on matrix structure (custard or biscuit), suggesting the nature of the matrix can dictate the manner of the digestive process.
3. The experiments' *in vitro* bioaccessibilities and *in vivo* bioavailabilities were similar, which suggests that these conclusions plausibly extend to the impact on post-bioaccessibility absorption. Previously reviewed authors would also agree with the conception that bioaccessibility is the main step governing a fat-soluble compound's bioavailability.
4. Nevertheless, the existing literature on the subject, which has established the belief that increased digestion should lead to increased bioaccessibility, is not always correct. The fact that lutein bioaccessibility was not the highest in the food matrix with the highest degree of fat digestibility is revealing.

The last bullet point regarding macronutrient digestibility warrants further investigation, which will be discussed in the next section. Before that, two other observations must be previously addressed.

Texture properties were used as a proxy to evaluate microstructure. Even though food matrix structure seemed to have a strong association to bioaccessibility overall, its impact within each matrix type was not so informative for modelling, as none of the texture properties of custard correlated with their bioaccessibility (Table 14), and biscuit hardness was not correlated to biscuits' bioaccessibility either.

Finally, an alternate hypothesis for the association between matrix structure and bioaccessibility, other than based on the macronutrient concentration, is the cooking method. Oil-in-water emulsions of curcumin prepared at higher temperatures (100°C vs 30°C) presented higher bioaccessibilities (L. Zou et al., 2015). This might explain why custards, cooked at 85°C for 10 minutes, presented higher bioaccessibilities than liquids which were prepared at room

temperature, or why biscuits, baked at 160°C for 30 minutes, presented the highest bioaccessibilities. However, this would not explain why protein and carbohydrates are significantly correlated with bioaccessibility in both custards and biscuits independently ($p - value < 0.05$) (Table 14). The possibility that the increased heat caused major alterations to the curcuminoid compounds thereby affecting quantification is not the case, as identified by HPLC-DAD quantification (section 4.3.1.1). In order to settle the matter, both water and milk matrices were prepared at room temperature, yet even so, milk presented over double the bioaccessibility (25.9% versus 12.4%), so it is more plausible that the presence of fat and other macronutrients is what drives bioaccessibility rather than the food processing temperature or method, at least in this study. Likewise, this difference in bioaccessibility between milk and water also suggests that it was the macronutrient concentration, more so than water content, which dictated bioaccessibility.

Nevertheless, the role of the Maillard reaction during biscuit baking in the improvement of their bioaccessibility should also be considered, as the Maillard reaction has proven useful to create effective delivery systems of curcumin (Song et al., 2025). It has been reported that the Maillard reaction does not affect the quantity of curcuminoids to be extracted from turmeric, but does seem to provide a positive effect on their antioxidant and anti-inflammatory properties (Y. Choi et al., 2019), possibly due to an increased bioaccessibility/bioavailability.

5.1.1.2 Macronutrient digestibility

The previous section highlighted that the general understanding of bioaccessibility is that higher degrees of macronutrient breakdown should be associated to higher degrees of compound release from the food matrix and consequentially a higher bioaccessibility, but empirical results suggest this is not necessarily always the case. Understanding why is challenging due to the complex interplay of enzymatic and physicochemical reactions occurring during digestion, all of which are affected by the properties of the food matrix as already

discussed. However, the first step to investigate these interactions is to quantify the degree of digestibility of each major macronutrient type independently.

During the first experiments with the *in vitro* protocol, it was observed that starch hydrolysis, as measured with Megazyme's total starch assay kit (section 4.1), was complete in the digestas. This suggested that the substitution of human salivary amylase, which is the enzyme of choice for simulating oral phase in the 2019 INFOGEST protocol, by bacterial amylase was effective. A study employing a non-INFOGEST protocol and not including amylase in the salivary phase also supports this substitution or even removal of oral amylase, as it was reported that the total starch hydrolysed in breads varying in dietary fibre was over 89.5% consistently, with only the bread formulated with arabinoxylan presenting any meaningful difference (94.4%) to the other four samples tested (Rojas-Bonzi et al., 2020). However, Gallego-Lobillo et al. reported that the 2019 INFOGEST protocol was only hydrolysing slightly higher than 50% of starch polysaccharide digested in isolation (Gallego-Lobillo et al., 2021), and Bustos et al. reported starch hydrolyses between 43.6-92.0% in different breads, pastas and cookies formulated with different flours and digested with the 2014 INFOGEST protocol (Bustos et al., 2017).

These contradicting results manifest that understanding the degree of impact of starch digestibility on lipophilic compound absorption is already limited by the fact that there is weak empirical consensus on the degree of digestibility of complex foods. Salivary amylase reportedly plays an important role in starch digestion (Freitas et al., 2022; Menard et al., 2023) and yet almost complete starch digestion was observed without including it, in bread no less, which is highly rich in starch (Rojas-Bonzi et al., 2020). This may be a case of a replication crisis event (Ioannidis, 2005) as several of its risk factors are present; there are few studies conducted in the research field presenting significant experimental design variation and striving for statistically significant results.

In the same vein, we observed through OPA tests (section 4.1) that four formulations (2 custards and 2 biscuits) representing the whole range of recorded

bioaccessibilities (25-53% approximately) displayed released amino acid equivalents whose correlation to bioaccessibility was insignificant. These measurements are considered reliable, as there is proof that the INFOGEST protocol accurately represents protein digestion because it was the main source of enquiry during its development and validation (Xavier & Mariutti, 2021), and the subject matter of expertise of numerous researchers on the panel (Brodkorb et al., 2019; M. Minekus et al., 2014; Mulet-Cabero et al., 2020). The digestible indispensable amino acid score (DIAAS) measurements via this method, the preferred method of protein quality in food, correlate very strongly with *in vivo* values ($r^2 = 0.89$) (Sousa et al., 2023).

The coefficient of variation of our released amino acid equivalents was lower than 5%. A study published in 2020 in which the 2014 INFOGEST protocol was applied to eight different protein sources including dietary fibre, reported that said fibre was likely to have negatively impacted digestibility (Sousa et al., 2020). Using the OPA method, the least digestible protein source (wheat bran cereals) released around ten times less amino acid equivalents than the most digestible protein source in said study (pigeon pea), a remarkable variability. In conclusion, Sousa's work employed an almost identical protocol to ours and yet observed a much larger range of protein digestibility.

It must be noted that in Hiolle's experiment, which was extensively discussed before and was applying the 2014 version of the INFOGEST protocol (Hiolle et al., 2020; M. Minekus et al., 2014), the peptide release profiles of biscuit, sponge cake, pudding and custard presented very small variability. Similar to our scenario. So, it is possible that we tested food matrices for which small variability in protein digestibility is to be expected.

The small variability and weak association between protein digestibility and curcuminoid bioaccessibility observed in our study, coupled with the complete starch hydrolysis across all tested samples, was used as justification to exclude assessing macronutrient digestibility during screening. Macronutrient digestibility tests seemed not to provide powerful predictors of bioaccessibility in the case of

these food matrices in particular. This is unfortunate as the understanding of the degree of impact of macronutrient digestibility on compound absorption is still incomplete, and systematic approaches like ours could be valuable when tackling this gap and examining the existence of meaningful interactions.

5.1.1.3 Dietary fibre enrichment

The source of variation amongst the tested formulations of custards and biscuits was the addition at 6% formulation weight of different dietary fibre supplements.

Amongst the numerous fibre supplement properties investigated (Table S18), soluble fibre content presented a positive association to bioaccessibility in our study and was included in our final model (Figure 33 and Table 16). This finding is supported by previous research, reinforcing the significance of soluble fibre content in explaining bioaccessibility. In a study characterising fusarotoxins enniatins' (fat-soluble compounds) bioaccessibility in different bread formulations, 5% inulin supplementation reduced bioaccessibilities of different enniatin compounds half as much as 10% of inulin did, suggesting that the addition of different concentrations of fibre could provide proportional effects on bioaccessibility, and other soluble fibres such as β -1,3 glucan and chitosan reduced enniatin bioaccessibility even further than inulin (Meca, Giri, et al., 2012; Meca, Meneghelli, et al., 2012). In a separate study, wheat bran with fibres (24% dietary fibre content compared to 5% on average of 13 other formulations) resulted in the lowest enniatin bioaccessibility when considering different breads, cereals and cookies, although fibre types were not specified (Prosperini et al., 2013). This matches with the general assumption that dietary fibres reduce bioaccessibility of fat-soluble compounds such as curcuminoids due to altered diffusion of gastrointestinal fluids and even binding or trapping of bioactives, which would all result in a lesser degree of digestion and curcuminoid release (Torsten Bohn, 2014; Torsten Bohn et al., 2015; P. Borel et al., 2015; Patrick Borel et al., 2023; Y. Tan & McClements, 2021). However, as a general trend, in our results we observed that the presence of fibre increases bioaccessibility, except for the formulation based on microcrystalline cellulose. Thus, dietary fibre

enrichment seems to be a promising way to enhance the curcumin release from the studied food formulations, contrary to what the literature suggests. However, no clear trend was identified based on the nature of the fibre (soluble or insoluble), so this finding was not enough to model bioaccessibility alone.

The capacity of the food ingredients to trap curcumin was initially believed to be a significant factor in their bioaccessibility, as many authors remark that the nature of the food matrix's components can heavily influence the retention of bioactive compounds (Langston et al., 2023; Palafox-Carlos et al., 2011). These tests were carried out for the fibre supplements whose indigestible nature would mean (theoretically) that any curcuminoids that had been trapped therein would not be released during digestion. Therefore, a strongly negative correlation between a supplement's binding capacity and its respective formulation's bioaccessibility should be observed when performing the curcuminoid binding tests (section 0). In addition, any curcuminoids that were released from the fibre in said test would not escape the experiment vessel as intestinal uptake was not simulated, thereby released curcuminoids would accumulate in the aqueous phase and saturate very quickly.

Regarding this saturation, and revisiting the case of curcuminoids in the screened food formulations, in the hypothetical case all the curcuminoids were released after digestion, their maximum concentration in the aqueous phase would approximately be 44.5 µg/mL, whereas their water solubility is reportedly 0.6 µg/mL or 7.4 µg/mL at 90 °C (Kurien et al., 2007). As this result contradicts our reported bioaccessibilities (50% bioaccessibility would indicate 22.3 µg/mL were solubilised), it can only be assumed that the presence of lipids in the digesta counteracts and even removes the issue of curcuminoid saturation when sufficient digested fat content is available. This suggests that even though the dietary fibres were not digestible, their interference with lipid digestion was not severely affecting the release of lipids for curcuminoids to solubilise in.

At this point, two clarifications must be made. First, ethanol was employed as extraction solvent which, as it increases curcuminoid solubility over 1000-fold (D.

de M. Carvalho et al., 2015), is required to correctly estimate curcuminoid bioaccessibilities. Ethanol had been previously included in food formulations with the purpose of improving curcuminoid solubility (Fu et al., 2015) so including it is another possibility for raising bioaccessibility in food matrices. Second, even though the total solid ingredient amounts of biscuits and custards were equivalent in the digestions, custards' digestions were performed with more dietary fibre as 5 grams of custard were digested compared to 1.67 grams of biscuit: both were supplemented at 6%, therefore custards contained 0.3 grams of fibre supplement in each digesta whereas biscuits only contained 0.1 grams (section 3.2.2). It could be argued that the slightly higher amount of dietary fibre in custard digestions caused their lower bioaccessibilities, but the fact that non-supplemented formulations did not report significantly higher bioaccessibilities in either matrix type is further proof that the impact of macronutrient concentration (or matrix structure) was much stronger than the impact of dietary fibre content on digestion and bioaccessibility. This could be ascertained more clearly by testing higher dietary fibre concentrations.

Returning to the curcuminoid entrapment tests, the fact that they were performed in the absence of digested lipids in the aqueous phase could explain why unbound curcuminoid fractions detected were not proportional to the supplemented concentrations; they were precipitating elsewhere as they were over the limits mentioned in the previous paragraphs (Figure S37). In spite of the previous argument, a negative correlation was only observed in custards ($\rho = -0.19$), while the opposite occurred in biscuits ($\rho = 0.37$). Moreover, these correlations were weak and not statistically significant ($p - value = 0.05$). Therefore, the hypothesis that the curcuminoid entrapment capacity would be a powerful predictor of bioaccessibility in this case was rejected.

Finally, as other authors have remarked, dietary fibres can act as inhibitors of flocculation events as is the case of pectin (R. Zhang et al., 2015). Minimising this flocculation facilitates adsorption of lipase to lipid droplets thereby increasing lipid digestion which should, in turn, favour bioactive compound release. We observed few positive and significant correlations ($p - value = 0.01$) between fibre content

and curcuminoid bioaccessibility, which were hemicellulose, pectin and lignin in biscuits. We hypothesize it is the stabilising effect which explains these correlations. Recently, hemicellulose was found to be a powerful stabilising agent of oil-in-water emulsions, possibly more so than cellulose whilst not increasing emulsion viscosity (Ho et al., 2023). We believe this explains why hemicellulose presented a strong correlation to bioaccessibility in biscuits ($\rho = 0.66$) and was a powerful predictor in the 6-parameter model. The fact that only cellulose content is strongly correlated with biscuit hardness ($\rho = 0.70$) (Figure 34) agrees with this argument. Lignin's significant correlation is more attributable to the fact it is highly collinear with pectin in the case of our experiment, as it is not a particularly strong stabilising agent.

No positive and significant correlations ($p - value = 0.01$) between bioaccessibility and specific fibre types were observed in custards. The only significant correlation observed in said matrix type was of cellulose, but it was of a negative sign ($\rho = -0.41$), which implies that even though cellulose also has a stabilising effect according to the literature, its increase in the viscosity of the custard triumphs over the flocculation inhibition in our results. Nevertheless, in our study we found that hemicellulose was more strongly correlated ($\rho = 0.82$) to custard viscosity than cellulose was ($\rho = 0.40$) (Figure 34), so this argument is contradictory whilst also contradicting the previously mentioned study by Ho et al. This highlights once more that the food matrix can interact in complex ways with fibre content.

Overall, these arguments emphasize that the literature surrounding the effect of dietary fibre on bioactive compound bioaccessibility presents contradicting conclusions, which could be attributed to the fact that it is the type of food matrix that greatly dictates how the digestion will proceed, with secondary interactions with fibre presence. For the sake of brevity, we hereby conclude that our study provides additional data to add to this pool of knowledge, which was uniquely procured in a systematic manner using a screening approach. These type of approaches, plus more in-depth mechanistic investigations, are still required to

develop our understanding of these interactions, as they appear to depend on a case-by-case basis.

5.1.2 Previous Machine Learning of compound absorption

Previous modelling attempts of oral bioaccessibility/bioavailability do not focus on the systematic optimisation of an individual compound as a function of formulation. To add context, the study of bioaccessibility/bioavailability can also be conducted based on the observation of its practical applications; how biorelevant a compound's antioxidant capacity remains after the digestion process (Idowu & Fatokun, 2021) or how much of a response to treatment a compound elicits when ingested, which is the main focus of most studies in the field as they usually pertain drug development purposes. A large majority of previous works generally target the modelling of several compounds at once, using data from existing databases, and have shown differing degrees of success. The modelling of bioavailability is sometimes also addressed as a classification problem, which enables achieving higher performances but at the cost of a loss of information. Degrees of accuracy typically range from 60% to 97%, with 70-80% on average, for example for goals as unassuming as estimating whether a drug has a positive, negative, or inexistent food effect (Bennett-Lenane et al., 2022; Fagerholm et al., 2021; Hou et al., 2007; Rajnish Kumar et al., 2011; Morales et al., 2024; Schneckener et al., 2019; Tyagi et al., 2023; Wei et al., 2022). The descriptors are typically molecular properties of each drug derived from concepts such as its membrane permeation, metabolic stability, solubility, acidity, and lipophilicity. In the case of focusing on a single compound as a function of formulation, identifying whether the formulations have a significant impact on these properties which would consequently alter their bioaccessibility/bioavailability is a potential approach for systematically modelling compounds individually as well. In addition, the concept that the evidence from one compound can aid in inferring bioaccessibility/bioavailability of another compound is very promising.

Another group of modelling studies regarding oral bioaccessibility/bioavailability are purposed towards risk assessment; to estimate the body burden of different contaminant compounds (such as lead or mercury) on dietary exposure, by approximation with Bayesian methods (Dong et al., 2016; Lv et al., 2023; Zhu et al., 2022), so it is not the first time this general methodology has been applied in this domain, but the hierarchical model applied to different matrix types is novel.

According to our review of the existing literature to this date, another study that resembles the work hereby presented was published recently by Hernández-Prieto et al. in 2023. Urine samples from 140 volunteers during a 2-month period were analysed for polyphenolic metabolites after consumption of a maqui-citrus beverage that had been enriched with a sweetener (sucrose, sucralose or stevia). Their results indicated that the sex of the volunteer, the sweetener and their interaction affected the bioavailability of the phenolic compounds according to ANOVA tests. Therefore, this cited study can be considered as a non-exhaustive evaluation of the impact of food formulation on bioavailability (as a function of the sweetener), but it also evaluated whether the subject played a key role in the availability of the compound. Their findings indicate that the individual subject is an important factor for the modelling of absorption and should be considered in the experimental design whenever possible, which would increase the complexity of the analysis. They used ML techniques to cluster the samples into 6 groups according to metabolic profiles, however a predictive model was not proposed (Hernández-Prieto et al., 2023).

It appears our case study is the first work attempting to build a predictive model of a compound's bioaccessibility/bioavailability exclusively as a function of the food formulation. This should be of significant interest to the food industry, as having such a tool at their disposal in the future would be invaluable. It would enable the design of optimised functional foods, ensuring enhanced nutrient delivery and consumer health benefits, while also providing a cost-effective decision support system for product development and quality control.

5.1.3 Choice of model

Initially, it was decided not to set aside a separate test set for evaluation during feature selection and cross-validation. Typically, a test set is reserved for assessing the final model's performance as presented in section 2.7.4; however, given the small sample size, excluding two or three formulations would result in an unrepresentative test set. This approach carries the risk of overly optimistic conclusions. To mitigate this, the final model was deliberately kept highly simplistic, potentially including only macronutrient content and an intercept (section 4.3.4.1).

Therefore, with the bias-variance tradeoff and producing a simple model in mind (section 2.7.3), the logical process behind the model's parametrization was as follows: linear regression models are prone to underfit, and as they can fit a pre-specified number of parameters for each variable, it is the volume of predictors that mostly determines the chances of the model overfitting or underfitting. Linear regression models enable greater control over customising model specifications, in contrast to algorithms such as decision trees or SVMs with complex kernels, because estimating the inferential outcome is generally intractable during previous exploratory analysis. For example, correlation studies are highly informative towards estimating what variables will be more relevant in linear regression models, but for decision trees involving binary choices, correlations are not so straightforwardly indicative.

As was quoted in the literature review (section 2.7.6.1), in modelling projects it is strongly recommended that every subjective decision (which are numerous) is justified. There is no reliable replacement for estimating the importance of features in a decision tree than actually fitting the model, and although this family of algorithms are highly effective model builders, there is no prior justification of their use for this particular dataset. Algorithms involving clustering or neighbouring distances (such as KNN or SVM) also require a subjective choice over number of neighbourhoods or what the shape of the hyperspace separating the datapoints is. These are hard to judge other than by fitting the models

themselves. To the author's knowledge, no other researcher has provided a model specification for any similar use case so there was no inclination towards choosing any particular algorithm apart from the one that provided the best results. This work sets a precedent for the adoption of Bayesian hierarchical linear regression in applications in this domain.

These kind of models will allow to progressively incorporate data from other matrix types, starting at smaller sample sizes, efficiently. Imbalanced datasets would pose a problem when using frequentist methods, which do not account for the level of uncertainty in the final model parameters, potentially leading to an inherent bias that would be difficult to diagnose without previously knowing the sample sizes involved in training. On the contrary, we demonstrated that liquid matrices were reasonably inferred using only macronutrient content (Figure 33). Furthermore, with the inclusion of non-pooled predictors in the hierarchical structure, predictions can be improved significantly without requiring fitting a fully independent model for each new matrix type.

Lastly, it is not uncommon in ML to apply feature preprocessing techniques such as feature clustering or polynomial and spline transformations of input features or the use of latent variables such as those produced by PCA dimensionality reduction. In this project these options were discarded because they were considered to increase the chances of overfitting, which, given the small size of the dataset, was the most worrying concern during modelling. On a positive note, without this heavy feature preprocessing, the interpretation of the model is greatly simplified. The decision of fitting a simple linear regression model will facilitate interpretation of results by future modellers and endpoint users, without the compulsory need for *post-hoc* tests such as SHapley Additive exPlanations (SHAP). This aspect of the discussion will be further expanded on in the future perspectives section.

5.1.4 Other bioaccessibility-optimising approaches

The proposed screening and modelling approach enables the study of the food matrix effects most impacting bioaccessibility in real foods. Its results indicated it

is a valuable procedure to optimise bioaccessibility whilst designing healthy and functional foods. For example, biscuits showed an over 4-fold increase in bioaccessibility compared to curcuminoids simply dissolved in water (12.4%), and fibre supplementation with pea fibre improved biscuit bioaccessibility 15% more than the lowest bioaccessibility biscuit (microcrystalline cellulose at 41%).

There is no systematic study available in the literature regarding the bioaccessibility of curcumin in complex food matrices to compare and benchmark the results of our optimisation against. Current approaches to improving bioaccessibility of curcuminoids more commonly target optimising a single formulation using specialised preparation techniques (Hegde et al., 2023). In spite of the fact that this project focused more on the method of systematic discovery rather than the design of a single highly-optimised formulation, these formulations show promise as they could also be incorporated into the food instead of raw curcuminoid powder, potentially improving bioaccessibility even further.

However, previous generations of these optimised products have raised concerns recently after long-term consumers of some of these formulations started displaying negative side effects, which were mostly attributed to contaminants in the formulations (Pancholi et al., 2021). This dependence on artificial compounds is avoided following our approach. Nevertheless, natural formulations (devoid of emulsifiers and other compounds considered adulterants or contaminants) such as CurQFen®, curcuRouge™ and CurcuWin® Ultra+, have shown incredible improvements in curcumin bioaccessibility/bioavailability (Abhilash et al., 2021; Das et al., 2023; Khanna et al., 2022; Kothapally et al., 2022; Sunagawa et al., 2021) and would be worthy candidates as source of curcuminoids compared to crude extracts. In addition, adjuvants such as piperine should be considered in formulations as they have been shown to enhance curcumin's bioavailability (Heidari et al., 2023), and while curcumin's metal ion chelation properties offer an interesting research avenue (our formulations include milk which is a valuable source of calcium), more study is required to understand the effect of this chelation on curcumin's absorption (Smirnova et al.,

2023). Nevertheless, as the majority of these studies involve the digestion of the formulation in isolation, incorporating them into functional foods should provide a more comprehensive understanding of their efficiency in real-life scenarios (whenever foods are co-ingested). Other recent studies suggest that the effect of the food matrix should affect negatively (Patrick Borel et al., 2023), but in our approach we have proven otherwise.

5.2 Vitamin E

Alpha-tocopherol, the main isoform of vitamin E, was screened extensively prior to curcuminoids in both custards and biscuits but its bioaccessibilities were constantly high (section 4.2), with food matrix effects causing barely any variability in bioaccessibility. For this reason, curcuminoids were the compounds for which modelling was performed.

These constant vit. E bioaccessibilities agreed with the literature. Vitamin E's bioaccessibility in different foods covers the whole possible range of values (1-100%) (Catelli Rocha Torres et al., 2022; Reboul, 2017), but studies with radioactive or deuterium labelled tocopherols have demonstrated that the *in vivo* bioavailability of vit. E is usually between 55 and 82% (Flory et al., 2019; MacMahon & Neale, 1970), with one study reporting that all twelve volunteers absorbed vit. E in the $81 \pm 1\%$ bracket after its consumption in 2% fat milk (Novotny, Fadel, et al., 2012). These are *in vivo* measurements of bioavailability, nonetheless our *in vitro* bioaccessibilities resemble those values.

A possible reason for the high bioaccessibilities/bioavailabilities was also proposed previously as an explanation for low curcuminoid detection in purely aqueous solutions: any formulation not including semi-skimmed milk lacked enough fat to effectively solubilise the lipophilic compound. This argument is further supported by the study of Mah et al. published in 2015. Using different fat-content milks, it was observed that 15 mg of α -tocopherol in 240 mL of milk presented 80.6% bioaccessibility and was unaffected by the milk's fat content. However, with 268 mg of α -tocopherol, the level of fat content had a noticeable effect on bioaccessibility; only full-fat milk displayed 80% bioaccessibility,

reduced fat milk presented less than 60% bioaccessibility and non-fat milk had less than 40% (Mah et al., 2015). In our study, vit. E to total fat content ratios were between $\sim 0.004X$ and $\sim 0.00021X$, whereas Mah investigated ratios of $\sim 1.34X$ when producing the formulation with lower than 40% bioaccessibility. Their 80% bioaccessibility was achieved at a $\sim 0.034X$ ratio, which is still a ten times higher vit. E to fat ratio than our least fat-containing formulation.

In light of this conclusion, that the ratio of vit. E to fat content was too low to observe a saturation effect caused by a lack of micellizable fat, attempts at increasing the variability in vit. E bioaccessibility so as to produce a richer dataset for mathematical modelling should begin by considering either reducing the amount of fat in the tested formulations, to the point that vit. E cannot find carrier micelles, or preferably, increasing the amount of vit. E to the point of saturation. However, the previous calculations conclude that, with the current fat present in our formulations, which all include 100 grams of egg yolks and therefore over 24 grams of fat (Table 3), around 33 grams of vit. E would have to be provided in order to achieve the same sort of reduction in bioaccessibility as the $\sim 1.34X$ ratio. This amount is well over the recommended dietary recommendations of 15 mg and would probably cause harm (despite the fact that vit. E excess is excreted efficiently, patients taking more than 1000 mg daily can experience toxic effects) (Owen & Dewald, 2023). This is the reason why our formulations were not enriched with vit. E to that degree.

It must be noted that should this avenue of research be pursued, in our study we did not observe a reduction in bioaccessibility of vit. E for the most highly enriched formulations, but recoveries were reduced from $\sim 93\%$ to $\sim 30\%$ when vit. E was administered at 5-fold concentrations or higher (

Table 10), therefore the extraction protocol must also be optimised for efficient assessment of higher enrichment levels.

5.3 Future perspectives

The previous section on vitamin E provided ideas for future developments towards its own model, dependent on modifying the experimental design and protocol. This section continues with this train of thought and highlights future avenues for research and development based on aspects that were previously discussed and new suggestions.

- **Correlate findings to other digestion simulations and *in vivo***

Static *in vitro* digestion models are extremely convenient over semi-dynamic and dynamic digestion models and are generally a useful alternative to these (Y. Tan et al., 2022; R. Wang et al., 2021), especially when screening a large number of formulations. Nevertheless, any conclusions driven from static experiments would benefit from additional validation with semi-dynamic or dynamic digestion models, which are usually a more accurate representation of *in vivo* conditions. However, this warrants a case by case discussion, as static *in vitro* models have proven to be very effective in specific situations; the protocol by Wu et al. obtained a correlation of $r = 0.98$ between starch hydrolysis and glycaemic effects (M. S. Wu et al., 2023). This protocol could be considered over the INFOGEST protocol if focusing on the impact of starch digestion on bioaccessibility.

It would also be highly desirable to evaluate the most promising formulations with animal models, to determine whether our optimisations at the level of compound release are correlated with an improvement in bioavailability, even though bioaccessibility is assumed to be the main step governing bioavailability of fat-soluble compounds. The importance of *in vivo* and *in vitro* correlations (IVIVCs) has been recognised, and maximising this correlation has been challenging for simulations involving lipid-based formulations (Y. Huang et al., 2021). In any case, the lack of studies investigating this issue with the newest digestion consensus protocols presents an opportunity for future research, and potentially for the development of compound-specific *in silico* digestion models, once more data is acquired.

- **Further investigation of the most impactful food matrix effects**

One of the screening and modelling method's main advantages is the capacity to highlight which food matrix effects impact bioaccessibility the most. These factors can later be investigated in more detail to determine the underlying molecular mechanisms in effect. More specifically, confocal laser microscopy could be used to visually inspect these matrices at different stages of the digestion process, staining each type of macronutrient separately in order to observe the size and shape of the lipid droplets containing the curcuminoids for example.

- **Include macronutrient digestibility screening**

The exclusion of macronutrient digestibility from screening was previously justified, with the acknowledgement that other food matrices could display more variable data. In this regard, it would be very interesting to systematically assess the impact of fat digestibility in formulations in which fat is protected from digestion (for example by higher concentrations of dietary fibre) as it could provide a better understanding of the impact of incomplete fat digestibility on curcuminoid bioaccessibility. Is the mere presence of fat in the food sufficient to assert that it will solubilise after release from the food matrix? Or is there a further nuance to this phenomena such as interactions with the type of food matrix. This same argument can be applied to starch and carbohydrate digestibility.

- **Future formulations to screen**

For the sake of brevity, an enumeration of possible future formulations of interest to analyse are as follows: 1-supplementation of dietary fibre at other concentrations (i.e. 3%, 9%, 12%, 15%, etc.) (section 5.1.1.3), 2-supplementation of other fibre supplements such as β -1,3 glucan and chitosan (section 5.1.1.3), 3-enrichment with other fat-soluble compounds such as quercetin or resveratrol to assess whether the model can be extrapolated to other compounds, in the same way that other multi-compound approaches have done so far (section 5.1.2), 4- preparation of other food matrix types such as liquid, muffins or bread which can benefit from the dataset already acquired (section 5.1.3), 5-food

formulations in which the bioactive source is replaced with specialised preparations previously optimised such as CurcuWin® Ultra+ (section 5.1.4).

- **Expand curcuminoid entrapment tests**

A custom assay to determine the curcuminoid binding capacity of the fibre supplements was developed. This test was useful in this experimental design as fibre supplements were the main source of variation between formulations. Nevertheless, assays targeting the quantification of trapped curcuminoids in the rest of the ingredients are desirable. Better yet, assessing the degree of entrapment in the matrix structure itself would be very revealing; this could possibly be performed via confocal microscopy at several stages of the digestive process as mentioned earlier, to provide the missing link in understanding the impact of the food matrix and its ingredients on curcuminoid release.

- **Evaluate Additional Physicochemical Properties of Dietary Fibres**

Beyond the properties analysed in this study, other physicochemical characteristics may play a crucial role in determining the effects of dietary fibres on bioaccessibility. Factors such as molecular weight, degree of polymerization, charge density, and thermal stability could provide further insights into the key mechanisms by which dietary fibres exert their influence.

- **Serve the modelling procedure as a software package**

To facilitate future modelling attempts, it would be highly desirable to automate the data-driven process of fitting bioaccessibility predictive models as much as possible. To this end, the most important step would be feature engineering, because including too many variables during modelling would increase computational times exponentially if performing the analysis on limited systems and increase the chances of overfitting. For this reason, the application should include the option to have access to higher computational resources in proportion to the number of training samples. Regarding what specific feature engineering steps to carry out, the same pipeline as executed in this work (sections 3.4.2 to 3.4.5) could be set as the standard approach, whilst offering the user the choice to ignore (or force) feature selection for any particular feature(s) the user definitely wants (or does not want) to include in the model. Once a sufficiently small number of features has been selected for modelling, which would be a numeric threshold to be specified by the user as the minimum ratio of samples per feature, fitting the model should not take long (several minutes). However, as feature selection and any cross-validation experiments could potentially take several hours, the application must be able to deal with this delay in serving the results to the user, so in this case an online server with a queuing system would be desirable.

Sharing data among users could be encouraged with this type of application (with their consent), and would require homogenising data formats, which usually generates resistance from the community but given that this modelling problem is well defined it ought not to be controversial. Without question, bioaccessibility would be the first, required, and only target variable to begin with, and any predictor variables would be optional, although the predictor variables employed in this work could be highlighted to encourage users to adjust to the format. Another important functionality of this application would be to employ the posteriors of previous models as priors, as a replacement for training models from only the data, in order for anyone to test their own beliefs against those of other researchers for the same data. A leaderboard system that ranked the most accurate models developed by the users for a given compound and food matrix

should rapidly provide the means and motivation to improve our understanding of the effects of the food matrix on bioactive compound absorption, in accordance with our proposed approach.

5.4 Conclusions

This study explored the possibility of inferring the degree of oral absorption of a bioactive compound of interest, as a function of the food matrix that carried it throughout its gastrointestinal journey.

The approach presented in this thesis is the first reported of its kind. Although statistics-based learning methods have been applied to predict oral absorption before, these approaches are mostly dependent on purely theoretical physicochemical properties of the compounds of interest, based on extrapolating insights from other compounds or drugs, and results have been slightly underwhelming so far.

This project recognises the limitations of attempting to fit robust predictive models with small datasets, which is likely the reason why bioaccessibility in different food matrices has not been modelled before. Hence, measures of uncertainty have been provided by employing probabilistic inferential algorithms, which also enable injecting previous knowledge to improve the confidence in said models. However, no prior beliefs could be established due to a lack of systematic studies, which essentially meant our approach was also data-oriented in principle.

This suggests that the methodology can be considered as an attractive proof-of-concept of how to provide additional empirical evidence to the existing pool of knowledge surrounding bioactive compound absorption. To this end, the beliefs proposed by the model can be used as prior beliefs to inform future models and iteratively grow in our understanding.

Regarding the background of the factors of importance we propose, the general consensus is that impediments to the breakdown of food macro- and micro-structures inhibit the release of compounds from the food matrix, which is exacerbated for fat-soluble compounds as they additionally require the co-

release of lipid molecules in which to solubilise. To this effect, as dietary fibres are indigestible, they are assumed to interfere with this breakdown and therefore reduce bioactive compound release from the matrix. With that in mind, the results suggest that increased macronutrient concentrations and dietary fibre content are significantly correlated to bioaccessibility and can be leveraged to estimate the degree of release of curcuminoids from the food matrix accurately for most formulations (relative error <10%).

Nevertheless, we hypothesise that the interactions occurring in the matrix are complex and that the food matrix effects should not be interpreted at face value; depending on the type of food matrix, its observed formulation properties could have opposing effects on bioaccessibility. Therefore, it appears that these investigations should be performed on a case-by-case basis for each type of food matrix. More research is required to fully understand these relationships, but the screening experiments proved indispensable to bring them to our attention; the most powerful predictive feature, macronutrient concentration, showed strong positive associations to bioaccessibility, an outcome for which there is no established consensus in the literature and strengthened by the lack of any conclusive evidence supporting the non-association.

It must be noted that, although this approach can provide valuable information of the factors involved, not all bioactive compounds of interest will necessarily benefit from this type of in-depth study; vitamin E presented unequivocally high bioaccessibilities in every tested food matrix.

In any case, validating any results with *in vivo*, or at the least, *in vitro* dynamic models coupled with cell cultures is highly recommended.

Overall, this study contributes to the understanding of curcuminoid bioaccessibility and offers a framework for studying other formulations beyond custards and biscuits, and eventually other bioactive chemicals. It encourages a new venue of empirical and systematic research into which factors of food affect the absorption of compounds of interest. Such studies could base the design of their experiments and modelling approach on the pipeline hereby presented.

Multi-compound models aiming to be cost-effective, such as those previously based exclusively on *in silico* molecular properties, could benefit from this methodology.

References

- Abell, T. L., Camilleri, M., Donohoe, K., Hasler, W. L., Lin, H. C., Maurer, A. H., McCallum, R. W., Nowak, T., Nusynowitz, M. L., Parkman, H. P., Shreve, P., Szarka, L. A., Snape, W. J. J., & Ziessman, H. A. (2008). Consensus recommendations for gastric emptying scintigraphy: a joint report of the American Neurogastroenterology and Motility Society and the Society of Nuclear Medicine. *The American Journal of Gastroenterology*, 103(3), 753–763. <https://doi.org/10.1111/j.1572-0241.2007.01636.x>
- Abril-Pla, O., Andreani, V., Carroll, C., Dong, L., Fonnesbeck, C. J., Kochurov, M., Kumar, R., Lao, J., Luhmann, C. C., Martin, O. A., Osthege, M., Vieira, R., Wiecki, T., & Zinkov, R. (2023). PyMC: a modern, and comprehensive probabilistic programming framework in Python. *PeerJ. Computer Science*, 9, e1516. <https://doi.org/10.7717/peerj-cs.1516>
- Acevedo-Fani, A., & Singh, H. (2022). Biophysical insights into modulating lipid digestion in food emulsions. *Progress in Lipid Research*, 85, 101129. <https://doi.org/https://doi.org/10.1016/j.plipres.2021.101129>
- Aggarwal, B. B., & Harikumar, K. B. (2009). Potential therapeutic effects of curcumin, the anti-inflammatory agent, against neurodegenerative, cardiovascular, pulmonary, metabolic, autoimmune and neoplastic diseases. *The International Journal of Biochemistry & Cell Biology*, 41(1), 40–59. <https://doi.org/https://doi.org/10.1016/j.biocel.2008.06.010>
- Agrawal, R. S., Ranveer, R. C., Rathod, N. B., & Nirmal, N. P. (2023). Chapter 6 - Phytochemicals as bioactive ingredients for functional foods. In S. Pati, T. Sarkar, & D. Lahiri (Eds.), *Recent Frontiers of Phytochemicals* (pp. 95–108). Elsevier. <https://doi.org/https://doi.org/10.1016/B978-0-443-19143-5.00028-1>
- Aguilera, J. M. (2019). The food matrix: implications in processing, nutrition and health. *Critical Reviews in Food Science and Nutrition*, 59(22), 3612–3629. <https://doi.org/10.1080/10408398.2018.1502743>

- Aguirre, A., Levine, M. J., Cohen, R. E., & Tabak, L. A. (1987). Immunochemical quantitation of alpha-amylase and secretory IgA in parotid saliva from people of various ages. *Archives of Oral Biology*, 32(4), 297–301. [https://doi.org/10.1016/0003-9969\(87\)90024-0](https://doi.org/10.1016/0003-9969(87)90024-0)
- Al-Awami, H. M., Raja, A., & Soos, M. P. (2023). *Physiology, Gastric Intrinsic Factor*.
- Aldrich, J. (1997). R. A. Fisher and the making of maximum likelihood 1912 - 1922. *Statistical Science*, 12(3), 162–176. <https://doi.org/10.1214/ss/1030037906>
- Allen, A., & Flemström, G. (2005). Gastroduodenal mucus bicarbonate barrier: protection against acid and pepsin. *American Journal of Physiology-Cell Physiology*, 288(1), C1–C19. <https://doi.org/10.1152/ajpcell.00102.2004>
- Alminger, M., Svelander, C., Wellner, A., Martinez-Tomas, R., Bialek, L., Larque, E., & Perez-Llamas, F. (2012). Applicability of in Vitro Models in Predicting the in Vivo Bioavailability of Lycopene and β -Carotene from Differently Processed Soups. *Food and Nutrition Sciences*, 03(04), 477–489. <https://doi.org/10.4236/fns.2012.34068>
- Alvarez, F. J., & Stella, V. J. (1989). The role of calcium ions and bile salts on the pancreatic lipase-catalyzed hydrolysis of triglyceride emulsions stabilized with lecithin. *Pharmaceutical Research*, 6(6), 449–457. <https://doi.org/10.1023/a:1015956104500>
- Amadiou, C., Coste, V., Neyrinck, A. M., Thijssen, V., Leyrolle, Q., Bindels, L. B., Piessevaux, H., Stärkel, P., de Timary, P., Delzenne, N. M., & Leclercq, S. (2022). Restoring an adequate dietary fiber intake by inulin supplementation: a pilot study showing an impact on gut microbiota and sociability in alcohol use disorder patients. *Gut Microbes*, 14(1), 2007042. <https://doi.org/10.1080/19490976.2021.2007042>
- Ambikapathi, R., Schneider, K. R., Davis, B., Herrero, M., Winters, P., & Fanzo,

- J. C. (2022). Global food systems transitions have enabled affordable diets but had less favourable outcomes for nutrition, environmental health, inclusion and equity. *Nature Food*, 3(9), 764–779. <https://doi.org/10.1038/s43016-022-00588-7>
- Anderson, J. W., Baird, P., Davis, R. H., Ferreri, S., Knudtson, M., Koraym, A., Waters, V., & Williams, C. L. (2009). Health benefits of dietary fiber. *Nutrition Reviews*, 67(4), 188–205. <https://doi.org/10.1111/j.1753-4887.2009.00189.x>
- Andrade, C. (2020). The Limitations of Online Surveys. *Indian Journal of Psychological Medicine*, 42(6), 575–576. <https://doi.org/10.1177/0253717620957496>
- Andrieu, C., de Freitas, N., Doucet, A., & Jordan, M. I. (2003). An Introduction to MCMC for Machine Learning. *Machine Learning*, 50(1), 5–43. <https://doi.org/10.1023/A:1020281327116>
- ANSES. (2022). *Nutrivigilance 2022 Activity Report*. May, 22–25. <https://www.anses.fr/fr/system/files/ANSES-RA2022-Nutrivigilance-EN.pdf>
- Armand, M., Hamosh, M., Mehta, N. R., Angelus, P. A., Philpott, J. R., Henderson, T. R., Dwyer, N. K., Lairon, D., & Hamosh, P. (1996). Effect of Human Milk or Formula on Gastric Function and Fat Digestion in the Premature Infant. *Pediatric Research*, 40(3), 429–437. <https://doi.org/10.1203/00006450-199609000-00011>
- Armand, M., Pasquier, B., André, M., Borel, P., Senft, M., Peyrot, J., Salducci, J., Portugal, H., Jaussan, V., & Lairon, D. (1999). Digestion and absorption of 2 fat emulsions with different droplet sizes in the human digestive tract². *The American Journal of Clinical Nutrition*, 70(6), 1096–1106. <https://doi.org/https://doi.org/10.1093/ajcn/70.6.1096>
- Atukunda, P., Eide, W. B., Kardel, K. R., Iversen, P. O., & Westerberg, A. C. (2021). Unlocking the potential for achievement of the UN Sustainable Development Goal 2 - “Zero Hunger” - in Africa: targets, strategies,

- synergies and challenges. *Food & Nutrition Research*, 65. <https://doi.org/10.29219/fnr.v65.7686>
- Aziz, A., Noreen, S., Khalid, W., Ejaz, A., Faiz Ul Rasool, I., Maham, Munir, A., Farwa, Javed, M., Ercisli, S., Okcu, Z., Marc, R. A., Nayik, G. A., Ramniwas, S., & Uddin, J. (2023). Pumpkin and Pumpkin Byproducts: Phytochemical Constitutes, Food Application and Health Benefits. *ACS Omega*, 8(26), 23346–23357. <https://doi.org/10.1021/acsomega.3c02176>
- Baer, D. J., Rumpler, W. V, Miles, C. W., & Fahey, G. C. J. (1997). Dietary fiber decreases the metabolizable energy content and nutrient digestibility of mixed diets fed to humans. *The Journal of Nutrition*, 127(4), 579–586. <https://doi.org/10.1093/jn/127.4.579>
- Bailey, R. L., West, K. P. J., & Black, R. E. (2015). The epidemiology of global micronutrient deficiencies. *Annals of Nutrition & Metabolism*, 66 Suppl 2, 22–33. <https://doi.org/10.1159/000371618>
- Baker, M. T., Lu, P., Parrella, J. A., & Leggette, H. R. (2022). Consumer Acceptance toward Functional Foods: A Scoping Review. *International Journal of Environmental Research and Public Health*, 19(3). <https://doi.org/10.3390/ijerph19031217>
- Balcázar-Zumaeta, C. R., Castro-Alayo, E. M., Muñoz-Astecker, L. D., Cayo-Colca, I. S., & Velayarce-Vallejos, F. (2023). Food Technology forecasting: A based bibliometric update in functional chocolates. *Heliyon*, 9(9), e19578. <https://doi.org/10.1016/j.heliyon.2023.e19578>
- Baldwin, S. A., & Larson, M. J. (2017). An introduction to using Bayesian linear regression with clinical data. *Behaviour Research and Therapy*, 98, 58–75. <https://doi.org/https://doi.org/10.1016/j.brat.2016.12.016>
- Banerjee, M., Reynolds, E., Andersson, H. B., & Nallamothu, B. K. (2019). Tree-Based Analysis. *Circulation. Cardiovascular Quality and Outcomes*, 12(5), e004879. <https://doi.org/10.1161/CIRCOUTCOMES.118.004879>

- Banfield, R. E., Hall, L. O., Bowyer, K. W., & Kegelmeyer, W. P. (2007). A Comparison of Decision Tree Ensemble Creation Techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1), 173–180. <https://doi.org/10.1109/TPAMI.2007.250609>
- Banks, K. P., Syed, K., Parekh, M., & McWhorter, N. (2023). *Gastric Emptying Scan*.
- Barber, T. M., Kabisch, S., Pfeiffer, A. F. H., & Weickert, M. O. (2020). The Health Benefits of Dietary Fibre. *Nutrients*, 12(10). <https://doi.org/10.3390/nu12103209>
- Bartolucci, F., & Scrucca, L. (2010). Point Estimation Methods with Applications to Item Response Theory Models. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education (Third Edition)* (Third Edit, pp. 366–373). Elsevier. [https://doi.org/https://doi.org/10.1016/B978-0-08-044894-7.01376-2](https://doi.org/10.1016/B978-0-08-044894-7.01376-2)
- Baştanlar, Y., & Ozuysal, M. (2014). Introduction to machine learning. *Methods in Molecular Biology (Clifton, N.J.)*, 1107, 105–128. https://doi.org/10.1007/978-1-62703-748-8_7
- Bauer, E., Jakob, S., & Mosenthin, R. (2005). Principles of Physiology of Lipid Digestion. *Asian-Australas J Anim Sci*, 18(2), 282–295. <https://doi.org/10.5713/ajas.2005.282>
- Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences of the United States of America*, 116(32), 15849–15854. <https://doi.org/10.1073/pnas.1903070116>
- Bellmann, S., Minekus, M., Sanders, P., Bosgra, S., & Havenaar, R. (2018). Human glycemic response curves after intake of carbohydrate foods are accurately predicted by combining in vitro gastrointestinal digestion with in silico kinetic modeling. *Clinical Nutrition Experimental*, 17, 8–22.

<https://doi.org/https://doi.org/10.1016/j.yclnex.2017.10.003>

Belson, W. A. (1959). Matching and Prediction on the Principle of Biological Classification. *Journal of the Royal Statistical Society Series C*, 8(2), 65–75.

<https://doi.org/10.2307/2985543>

Bennett-Lenane, H., Griffin, B. T., & O'Shea, J. P. (2022). Machine learning methods for prediction of food effects on bioavailability: A comparison of support vector machines and artificial neural networks. *European Journal of Pharmaceutical Sciences*, 168, 106018.

<https://doi.org/https://doi.org/10.1016/j.ejps.2021.106018>

Berry, S. E. E., Tydeman, E. A., Lewis, H. B., Phalora, R., Rosborough, J., Picout, D. R., & Ellis, P. R. (2008). Manipulation of lipid bioaccessibility of almond seeds influences postprandial lipemia in healthy human subjects. *The American Journal of Clinical Nutrition*, 88(4), 922–929.

<https://doi.org/10.1093/ajcn/88.4.922>

Bjelakovic, G., Nikolova, D., Gluud, L. L., Simonetti, R. G., & Gluud, C. (2015). Antioxidant supplements for prevention of mortality in healthy participants and patients with various diseases. *Sao Paulo Medical Journal*, 133(2), 164.

<https://doi.org/10.1590/1516-3180.20151332T1>

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet (London, England)*, 1(8476), 307–310.

Bland, J. M., & Altman, D. G. (1998). Bayesians and frequentists. *BMJ (Clinical Research Ed.)*, 317(7166), 1151–1160.

<https://doi.org/10.1136/bmj.317.7166.1151>

Bloom, D., Cafiero-Fonseca, E., Jane Llopis, E., Abrahams-Gessel, S., Bloom, L., Fathima, S., Feigl, A., Gaziano, T., Hamandi, A., Mowafi, M., O'Farrell, D., Ozaltin, E., Pandya, A., Prettnner, K., Rosenberg, L., Seligman, B., Stein, A., Weinstein, C., & Weiss, J. (2011). *The Global Economic Burden of*

Noncommunicable Diseases.

- Bodmer, M. W., Angal, S., Yarranton, G. T., Harris, T. J., Lyons, A., King, D. J., Pieroni, G., Riviere, C., Verger, R., & Lowe, P. A. (1987). Molecular cloning of a human gastric lipase and expression of the enzyme in yeast. *Biochimica et Biophysica Acta*, 909(3), 237–244. [https://doi.org/10.1016/0167-4781\(87\)90083-2](https://doi.org/10.1016/0167-4781(87)90083-2)
- Bohn, T, Carriere, F., Day, L., Deglaire, A., Egger, L., Freitas, D., Golding, M., Le Feunteun, S., Macierzanka, A., Menard, O., Miralles, B., Moscovici, A., Portmann, R., Recio, I., Rémond, D., Santé-Lhoutelier, V., Wooster, T. J., Lesmes, U., Mackie, A. R., & Dupont, D. (2018). Correlation between in vitro and in vivo data on food digestion. What can we predict with static in vitro digestion models? *Critical Reviews in Food Science and Nutrition*, 58(13), 2239–2261. <https://doi.org/10.1080/10408398.2017.1315362>
- Bohn, Torsten. (2014). Dietary factors affecting polyphenol bioavailability. *Nutrition Reviews*, 72(7), 429–452. <https://doi.org/10.1111/nure.12114>
- Bohn, Torsten, Mcdougall, G. J., Alegría, A., Alminger, M., Arrigoni, E., Aura, A. M., Brito, C., Cilla, A., El, S. N., Karakaya, S., Martínez-Cuesta, M. C., & Santos, C. N. (2015). Mind the gap-deficits in our knowledge of aspects impacting the bioavailability of phytochemicals and their metabolites-a position paper focusing on carotenoids and polyphenols. In *Molecular Nutrition and Food Research*. <https://doi.org/10.1002/mnfr.201400745>
- Borel, P., Caillaud, D., & Cano, N. J. (2015). Vitamin D Bioavailability: State of the Art. *Critical Reviews in Food Science and Nutrition*, 55(9), 1193–1205. <https://doi.org/10.1080/10408398.2012.688897>
- Borel, Patrick, Hammaz, F., Lecourt, L., Marconot, G., Gillet, G., Rozier, C., & Desmarchelier, C. (2023). The Incorporation of Curcuminoids in Gamma-Cyclodextrins Improves Their Poor Bioaccessibility, Which Is due to Both Their Very Low Incorporation into Mixed Micelles and Their Partial Adsorption on Food. *Molecular Nutrition & Food Research*, 67(12),

e2200798. <https://doi.org/10.1002/mnfr.202200798>

- Borel, Patrick, Preveraud, D., & Desmarchelier, C. (2013). Bioavailability of vitamin E in humans: An update. *Nutrition Reviews*, 71(6), 319–331. <https://doi.org/10.1111/nure.12026>
- Bornhorst, G. M., & Paul Singh, R. (2014). Gastric Digestion In Vivo and In Vitro: How the Structural Aspects of Food Influence the Digestion Process. *Annual Review of Food Science and Technology*, 5(1), 111–132. <https://doi.org/10.1146/annurev-food-030713-092346>
- Bouis, H. E., Hotz, C., McClafferty, B., Meenakshi, J. V., & Pfeiffer, W. H. (2011). Biofortification: A New Tool to Reduce Micronutrient Malnutrition. *Food and Nutrition Bulletin*, 32(1_suppl1), S31–S40. <https://doi.org/10.1177/15648265110321S105>
- Brandstaeter, S., Fuchs, S. L., Aydin, R. C., & Cyron, C. J. (2019). Mechanics of the stomach: A review of an emerging field of biomechanics. *GAMM-Mitteilungen*, 42(3), e201900001. <https://doi.org/https://doi.org/10.1002/gamm.201900001>
- Brayne, C. (2007). The elephant in the room - healthy brains in later life, epidemiology and public health. *Nature Reviews. Neuroscience*, 8(3), 233–239. <https://doi.org/10.1038/nrn2091>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. Taylor & Francis. <https://books.google.co.uk/books?id=JwQx-WOmSyQC>
- Brodkorb, A., Egger, L., Alminger, M., Alvito, P., Assunção, R., Ballance, S., Bohn, T., Bourlieu-Lacanal, C., Boutrou, R., Carrière, F., Clemente, A., Corredig, M., Dupont, D., Dufour, C., Edwards, C., Golding, M., Karakaya, S., Kirkhus, B., Le Feunteun, S., ... Recio, I. (2019). INFOGEST static in vitro

- simulation of gastrointestinal food digestion. *Nature Protocols*, 14(4), 991–1014. <https://doi.org/10.1038/s41596-018-0119-1>
- Brown, G. C. (2015). Living Too Long. *EMBO Reports*, 16(2), 137–141. <https://doi.org/10.15252/embr.201439518>
- Buettner, A., Beer, A., Hannig, C., & Settles, M. (2001). Observation of the Swallowing Process by Application of Videofluoroscopy and Real-time Magnetic Resonance Imaging—Consequences for Retronasal Aroma Stimulation. *Chemical Senses*, 26(9), 1211–1219. <https://doi.org/10.1093/chemse/26.9.1211>
- Buffière, C., Hiolle, M., Peyron, M. A., Richard, R., Meunier, N., Batisse, C., Rémond, D., Dupont, D., Nau, F., Pereira, B., & Savary-Auzeloux, I. (2021). Food matrix structure (from Biscuit to Custard) has an impact on folate bioavailability in healthy volunteers. *European Journal of Nutrition*, 60(1), 411–423. <https://doi.org/10.1007/s00394-020-02258-5>
- Burant, C. F., Takeda, J., Brot-Laroche, E., Bell, G. I., & Davidson, N. O. (1992). Fructose transporter in human spermatozoa and small intestine is GLUT5. *The Journal of Biological Chemistry*, 267(21), 14523–14526.
- Bustos, M. C., Vignola, M. B., Pérez, G. T., & León, A. E. (2017). In vitro digestion kinetics and bioaccessibility of starch in cereal food products. *Journal of Cereal Science*, 77, 243–250. <https://doi.org/https://doi.org/10.1016/j.jcs.2017.08.018>
- Calvo-Castro, L. A., Sus, N., Schiborr, C., Bosy-Westphal, A., Duran, M. L., Fesenmeyer, D., Fesenmeyer, G., & Frank, J. (2019). Pharmacokinetics of vitamin E, γ -oryzanol, and ferulic acid in healthy humans after the ingestion of a rice bran-enriched porridge prepared with water or with milk. *European Journal of Nutrition*, 58(5), 2099–2110. <https://doi.org/10.1007/s00394-018-1770-6>
- Camilleri, M., & Vazquez Roque, M. I. (2014). Gastrointestinal Functions. In M. J.

- Aminoff & R. B. Daroff (Eds.), *Encyclopedia of the Neurological Sciences* (Second Edition) (Second Edi, pp. 411–416). Academic Press.
<https://doi.org/https://doi.org/10.1016/B978-0-12-385157-4.00501-7>
- Campos, L. A., & Sancho, J. (2003). The active site of pepsin is formed in the intermediate conformation dominant at mildly acidic pH. *FEBS Letters*, 538(1–3), 89–95. [https://doi.org/10.1016/s0014-5793\(03\)00152-2](https://doi.org/10.1016/s0014-5793(03)00152-2)
- Capretto, T., Piho, C., Kumar, R., Westfall, J., Yarkoni, T., & Martin, O. A. (2022). Bambi: A Simple Interface for Fitting Bayesian Linear Models in Python. *Journal of Statistical Software*, 103(15), 1–29. <https://doi.org/10.18637/jss.v103.i15>
- Carrión-Barberà, I., Triginer, L., Tío, L., Pérez-García, C., Ribes, A., Abad, V., Pros, A., Bermúdez-López, M., Castro-Boqué, E., Lecube, A., Valdivielso, J. M., Group, I. P., Monfort, J., & Salman-Monte, T. C. (2024). Role of Advanced Glycation End Products as New Biomarkers in Systemic Lupus Erythematosus. *International Journal of Molecular Sciences*, 25(5). <https://doi.org/10.3390/ijms25053022>
- Carvalho, D. de M., Takeuchi, K. P., Geraldine, R. M., de Moura, C. J., & Torres, M. C. L. (2015). Production, solubility and antioxidant activity of curcumin nanosuspension. *Food Science and Technology*, 35(1), 115–119. <https://doi.org/10.1590/1678-457X.6515>
- Cas, M. D., & Ghidoni, R. (2019). Dietary curcumin: Correlation between bioavailability and health potential. *Nutrients*, 11(9), 1–14. <https://doi.org/10.3390/nu11092147>
- Casiraghi, M. C., Zanchi, R., Canzi, E., Pagani, M. A., Viaro, T., Benini, L., & D'Egidio, M. G. (2011). Prebiotic potential and gastrointestinal effects of immature wheat grain (IWG) biscuits. *Antonie van Leeuwenhoek*, 99(4), 795–805. <https://doi.org/10.1007/s10482-011-9553-6>
- Cassady, B. A., Hollis, J. H., Fulford, A. D., Considine, R. V., & Mattes, R. D.

- (2009). Mastication of almonds: effects of lipid bioaccessibility, appetite, and hormone response². *The American Journal of Clinical Nutrition*, 89(3), 794–800. <https://doi.org/10.3945/ajcn.2008.26669>
- Catelli Rocha Torres, L., Giovanini de Oliveira Sartori, A., de Souza Silva, A. P., & Matias de Alencar, S. (2022). Bioaccessibility and uptake/epithelial transport of vitamin E: Discoveries and challenges of in vitro and ex vivo assays. *Food Research International*, 162, 112143. <https://doi.org/10.1016/j.foodres.2022.112143>
- Cena, H., & Calder, P. C. (2020). Defining a Healthy Diet: Evidence for The Role of Contemporary Dietary Patterns in Health and Disease. *Nutrients*, 12(2). <https://doi.org/10.3390/nu12020334>
- Chau, C.-F., & Huang, Y.-L. (2003). Comparison of the Chemical Composition and Physicochemical Properties of Different Fibers Prepared from the Peel of Citrus sinensis L. Cv. Liucheng. *Journal of Agricultural and Food Chemistry*, 51(9), 2615–2618. <https://doi.org/10.1021/jf025919b>
- Chen, X., Touyz, R. M., Park, J. B., & Schiffrin, E. L. (2001). Antioxidant effects of vitamins C and E are associated with altered activation of vascular NADPH oxidase and superoxide dismutase in stroke-prone SHR. *Hypertension*, 38(3 Pt 2), 606–611. <https://doi.org/10.1161/hy09t1.094005>
- Chicco, D., Oneto, L., & Tavazzi, E. (2022). Eleven quick tips for data cleaning and feature engineering. *PLoS Computational Biology*, 18(12), e1010718. <https://doi.org/10.1371/journal.pcbi.1010718>
- Choi, K. (2023). A review of the Bayesian approach with the MCMC and the HMC as a competitor of classical likelihood statistics for pharmacometricians. *Translational and Clinical Pharmacology*, 31(2), 69–84. <https://doi.org/10.12793/tcp.2023.31.e9>
- Choi, Y., Ban, I., Lee, H., Baik, M.-Y., & Kim, W. (2019). Puffing as a Novel Process to Enhance the Antioxidant and Anti-Inflammatory Properties of

- Curcuma longa L. (Turmeric). *Antioxidants (Basel, Switzerland)*, 8(11).
<https://doi.org/10.3390/antiox8110506>
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49(12), 997–1003. <https://doi.org/10.1037/0003-066X.49.12.997>
- Colantonio, V., Ferrão, L. F. V, Tieman, D. M., Bliznyuk, N., Sims, C., Klee, H. J., Munoz, P., & Resende, M. F. R. (2022). Metabolomic selection for enhanced fruit flavor. *Proceedings of the National Academy of Sciences*, 119(7), e2115865119. <https://doi.org/10.1073/pnas.2115865119>
- Colombo, R., Ferron, L., Frosi, I., & Papetti, A. (2021). Advances in static: In vitro digestion models after the COST action Infogest consensus protocol. *Food and Function*, 12(17), 7619–7636. <https://doi.org/10.1039/d1fo01089a>
- Colombo, R., Paolillo, M., & Papetti, A. (2023). A Dynamic In Vitro Model for Testing Intestinal Absorption of Different Vegetable Food Secondary Metabolites. *Applied Sciences*, 13(8). <https://doi.org/10.3390/app13085033>
- Corte-Real, J., & Bohn, T. (2018). Interaction of divalent minerals with liposoluble nutrients and phytochemicals during digestion and influences on their bioavailability – a review. *Food Chemistry*, 252, 285–293. [https://doi.org/https://doi.org/10.1016/j.foodchem.2018.01.113](https://doi.org/10.1016/j.foodchem.2018.01.113)
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. <https://doi.org/10.1007/BF00994018>
- Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, 19(1), 270. <https://doi.org/10.1186/s12859-018-2264-5>
- Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27. <https://doi.org/10.1109/TIT.1967.1053964>
- Damián, M. R., Cortes-Perez, N. G., Quintana, E. T., Ortiz-Moreno, A., Noguez, C. G., Cruceño-Casarrubias, C. E., Pardo, M. E. S., & Bermúdez-Humarán,

- L. G. (2022). Functional Foods, Nutraceuticals and Probiotics: A Focus on Human Health. *Microorganisms*, 10(5), 1–13. <https://doi.org/10.3390/microorganisms10051065>
- de Carvalho, N. M., Oliveira, D. L., Costa, C. M., Pintado, M. E., & Madureira, A. R. (2023). Strategies to Assess the Impact of Sustainable Functional Food Ingredients on Gut Microbiota. *Foods*, 12(11), 2209. <https://doi.org/10.3390/foods12112209>
- De Paula, J. A., Carmuega, E., & Weill, R. (2008). Effect of the ingestion of a symbiotic yogurt on the bowel habits of women with functional constipation. *Acta Gastroenterologica Latinoamericana*, 38(1), 16–25.
- de Vos, L. C., Lefrandt, J. D., Dullaart, R. P. F., Zeebregts, C. J., & Smit, A. J. (2016). Advanced glycation end products: An emerging biomarker for adverse outcome in patients with peripheral artery disease. *Atherosclerosis*, 254, 291–299. <https://doi.org/10.1016/j.atherosclerosis.2016.10.012>
- Déat, E., Blanquet-Diot, S., Jarrige, J.-F., Denis, S., Beyssac, E., & Alric, M. (2009). Combining the dynamic TNO-gastrointestinal tract system with a Caco-2 cell culture model: application to the assessment of lycopene and alpha-tocopherol bioavailability from a whole food. *Journal of Agricultural and Food Chemistry*, 57(23), 11314–11320. <https://doi.org/10.1021/jf902392a>
- Debelo, H., Li, M., & Ferruzzi, M. G. (2020). Processing influences on food polyphenol profiles and biological activity. *Current Opinion in Food Science*, 32, 90–102. <https://doi.org/10.1016/j.cofs.2020.03.001>
- DeBenedictis, J. N., de Kok, T. M., & van Breda, S. G. (2023). Impact of Processing Method and Storage Time on Phytochemical Concentrations in an Antioxidant-Rich Food Mixture. *Antioxidants*, 12(6). <https://doi.org/10.3390/antiox12061252>
- Dexter, G. P., Grannis, S. J., Dixon, B. E., & Kasthurirathne, S. N. (2020).

- Generalization of Machine Learning Approaches to Identify Notifiable Conditions from a Statewide Health Information Exchange. *AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science*, 2020, 152–161.
- Dima, C., Assadpour, E., Dima, S., & Jafari, S. M. (2020). Bioavailability of nutraceuticals: Role of the food matrix, processing conditions, the gastrointestinal tract, and nanodelivery systems. *Comprehensive Reviews in Food Science and Food Safety*, 19(3), 954–994. <https://doi.org/10.1111/1541-4337.12547>
- Docherty, A. J., Bodmer, M. W., Angal, S., Verger, R., Riviere, C., Lowe, P. A., Lyons, A., Emtage, J. S., & Harris, T. J. (1985). Molecular cloning and nucleotide sequence of rat lingual lipase cDNA. *Nucleic Acids Research*, 13(6), 1891–1903. <https://doi.org/10.1093/nar/13.6.1891>
- Dong, Z., Yan, K., Liu, Y., Naidu, R., Duan, L., Wijayawardena, A., Semple, K. T., & Rahman, M. M. (2016). A meta-analysis to correlate lead bioavailability and bioaccessibility and predict lead bioavailability. *Environment International*, 92–93, 139–145. <https://doi.org/10.1016/j.envint.2016.04.009>
- Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1996). Support Vector Regression Machines. In M. C. Mozer, M. Jordan, & T. Petsche (Eds.), *Advances in Neural Information Processing Systems* (Vol. 9). MIT Press. https://proceedings.neurips.cc/paper_files/paper/1996/file/d38901788c533e8286cb6400b40b386d-Paper.pdf
- Dupont, D., Alric, M., Blanquet-Diot, S., Bornhorst, G., Cueva, C., Deglaire, A., Denis, S., Ferrua, M., Havenaar, R., Lelieveld, J., Mackie, A. R., Marzorati, M., Menard, O., Minekus, M., Miralles, B., Recio, I., & Van den Abbeele, P. (2019). Can dynamic in vitro digestion systems mimic the physiological reality? *Critical Reviews in Food Science and Nutrition*, 59(10), 1546–1562. <https://doi.org/10.1080/10408398.2017.1421900>

- Duttaroy, A. K. (2019). Regulation of functional foods in European Union: Assessment of health claim by the European food safety authority. In *Nutraceutical and Functional Food Regulations in the United States and around the World* (Third Edit). Elsevier Inc. <https://doi.org/10.1016/B978-0-12-816467-9.00019-8>
- Dwyer, J. T., Wiemer, K. L., Dary, O., Keen, C. L., King, J. C., Miller, K. B., Philbert, M. A., Tarasuk, V., Taylor, C. L., Gaine, P. C., Jarvis, A. B., & Bailey, R. L. (2015). Fortification and health: challenges and opportunities. *Advances in Nutrition (Bethesda, Md.)*, 6(1), 124–131. <https://doi.org/10.3945/an.114.007443>
- EFSA Panel on Food Additives and Nutrient Sources added to Food (ANS). (2010). Scientific Opinion on the re-evaluation of curcumin (E 100) as a food additive. *EFSA Journal*, 8(9), 1679. <https://doi.org/https://doi.org/10.2903/j.efsa.2010.1679>
- Eich, T., Stähle, M., Gustafsson, B., Horneland, R., Lempinen, M., Lundgren, T., Rafael, E., Tufveson, G., Zur-Mühlen, B. von, Olerud, J., Scholz, H., & Korsgren, O. (2018). Calcium: A Crucial Potentiator for Efficient Enzyme Digestion of the Human Pancreas. *Cell Transplantation*, 27(7), 1031–1038. <https://doi.org/10.1177/0963689718779350>
- El-Saadony, M. T., Yang, T., Korma, S. A., Sitohy, M., Abd El-Mageed, T. A., Selim, S., Al Jaouni, S. K., Salem, H. M., Mahmmod, Y., Soliman, S. M., Mo'men, S. A. A., Mosa, W. F. A., El-Wafai, N. A., Abou-Aly, H. E., Sitohy, B., Abd El-Hack, M. E., El-Tarabily, K. A., & Saad, A. M. (2022). Impacts of turmeric and its principal bioactive curcumin on human health: Pharmaceutical, medicinal, and food applications: A comprehensive review. *Frontiers in Nutrition*, 9, 1040259. <https://doi.org/10.3389/fnut.2022.1040259>
- Espinal-Ruiz, M., Restrepo-Sánchez, L.-P., Narváez-Cuenca, C.-E., & McClements, D. J. (2016). Impact of pectin properties on lipid digestion under simulated gastrointestinal conditions: Comparison of citrus and

- banana passion fruit (*Passiflora tripartita* var. *mollissima*) pectins. *Food Hydrocolloids*, 52, 329–342.
<https://doi.org/https://doi.org/10.1016/j.foodhyd.2015.05.042>
- European Parliament & Council. (2006). Regulation (EC) 1924/2006 on nutrition and health claims made on foods. *Official Journal of the European Communities*, L404(1924), 1–15.
- Faber, M., Berti, C., & Smuts, M. (2014). Prevention and control of micronutrient deficiencies in developing countries: Current perspectives. *Nutrition and Dietary Supplements*, 6, 41. <https://doi.org/10.2147/NDS.S43523>
- Fagerholm, U., Hellberg, S., & Spjuth, O. (2021). Article advances in predictions of oral bioavailability of candidate drugs in man with new machine learning methodology. *Molecules*, 26(9), 1–11.
<https://doi.org/10.3390/molecules26092572>
- Fedorov, V., Mannino, F., & Zhang, R. (2009). Consequences of dichotomization. *Pharmaceutical Statistics*, 8(1), 50–61. <https://doi.org/10.1002/pst.331>
- Ferreira, T., Gomes, S. M., & Santos, L. (2023). Elevating Cereal-Based Nutrition: Moringa oleifera Supplemented Bread and Biscuits. *Antioxidants (Basel, Switzerland)*, 12(12). <https://doi.org/10.3390/antiox12122069>
- Fink, D. (1997). A Compendium of Conjugate Priors. *Technical Report*.
- Fish, E. M., Shumway, K. R., & Burns, B. (2024). Physiology, Small Bowel. In *StatPearls*. <http://www.ncbi.nlm.nih.gov/pubmed/27908510>
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179–188.
<https://doi.org/https://doi.org/10.1111/j.1469-1809.1936.tb02137.x>
- Flament, F., Hourblin, V., & Nouveau, S. (2014). *Skin Color Types and Indian Skin Characteristics* (pp. 47–61).
- Flory, S., Birringer, M., & Frank, J. (2019). Bioavailability and Metabolism of

- Vitamin E. In P. Weber, M. Birringer, J. B. Blumberg, M. Eggersdorfer, & J. Frank (Eds.), *Vitamin E in Human Health* (pp. 31–41). Springer International Publishing. https://doi.org/10.1007/978-3-030-05315-4_4
- Fornacon-Wood, I., Mistry, H., Johnson-Hart, C., Faivre-Finn, C., O'Connor, J. P. B., & Price, G. J. (2022). Understanding the Differences Between Bayesian and Frequentist Statistics. In *International journal of radiation oncology, biology, physics* (Vol. 112, Issue 5, pp. 1076–1082). <https://doi.org/10.1016/j.ijrobp.2021.12.011>
- Freitas, D., Souchon, I., & Le Feunteun, S. (2022). The contribution of gastric digestion of starch to the glycaemic index of breads with different composition or structure. *Food Funct.*, 13(4), 1718–1724. <https://doi.org/10.1039/D1FO03901F>
- Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1), 119–139. <https://doi.org/10.1006/jcss.1997.1504>
- Fu, S., Augustin, M. A., Shen, Z., Ng, K., Sanguansri, L., & Ajlouni, S. (2015). Bioaccessibility of curcuminoids in buttermilk in simulated gastrointestinal digestion models. *Food Chemistry*, 179, 52–59. <https://doi.org/10.1016/j.foodchem.2015.01.126>
- Galántai, A. (2000). The theory of Newton's method. *Journal of Computational and Applied Mathematics*, 124(1), 25–44. [https://doi.org/https://doi.org/10.1016/S0377-0427\(00\)00435-0](https://doi.org/https://doi.org/10.1016/S0377-0427(00)00435-0)
- Gallego-Lobillo, P., Ferreira-Lazarte, A., Hernández-Hernández, O., & Villamiel, M. (2021). In vitro digestion of polysaccharides: InfoGest protocol and use of small intestinal extract from rat. *Food Research International*, 140, 110054. <https://doi.org/https://doi.org/10.1016/j.foodres.2020.110054>
- García-Pérez, M. Á. (2019). Bayesian Estimation with Informative Priors is Indistinguishable from Data Falsification. *The Spanish Journal of*

Psychology, 22, E45. <https://doi.org/10.1017/sjp.2019.41>

- Gibson, B. R., Rogers, T. T., & Zhu, X. (2013). Human semi-supervised learning. *Topics in Cognitive Science*, 5(1), 132–172. <https://doi.org/10.1111/tops.12010>
- Gilks, W. R., Best, N. G., & Tan, K. K. C. (1995a). Adaptive Rejection Metropolis Sampling within Gibbs Sampling. *Applied Statistics*, 44(4), 455. <https://doi.org/10.2307/2986138>
- Gilks, W. R., Best, N. G., & Tan, K. K. C. (1995b). Adaptive Rejection Metropolis Sampling within Gibbs Sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 44(4), 455–472. <http://www.jstor.org/stable/2986138>
- Gomes, A., Furtado, G. D. F., & Cunha, R. L. (2019). Bioaccessibility of Lipophilic Compounds Vehiculated in Emulsions: Choice of Lipids and Emulsifiers. *Journal of Agricultural and Food Chemistry*, 67(1), 13–18. <https://doi.org/10.1021/acs.jafc.8b05460>
- Goncalves, A., Roi, S., Nowicki, M., Dhaussy, A., Huertas, A., Amiot, M. J., & Reboul, E. (2015). Fat-soluble vitamin intestinal absorption: Absorption sites in the intestine and interactions for absorption. *Food Chemistry*, 172, 155–160. <https://doi.org/10.1016/j.foodchem.2014.09.021>
- Goozee, K. G., Shah, T. M., Sohrabi, H. R., Rainey-Smith, S. R., Brown, B., Verdile, G., & Martins, R. N. (2016). Examining the potential clinical value of curcumin in the prevention and diagnosis of Alzheimer's disease. *British Journal of Nutrition*, 115(3), 449–465. <https://doi.org/10.1017/S0007114515004687>
- Gorlov, I. F., Giro, T. M., Pryanishnikov, V. V., Slozhenkina, M. I., Randelin, A. V., Mosolova, N. I., Zlobina, E. Y., & Kulikovskiy, A. V. (2015). Using the Fiber Preparations in Meat Processing. *Modern Applied Science*, 9(10), 54–64. <https://doi.org/10.5539/mas.v9n10p54>

- Götz-Hahn, F., Hosu, V., & Saupe, D. (2022). Critical analysis on the reproducibility of visual quality assessment using deep features. *PloS One*, 17(8), e0269715. <https://doi.org/10.1371/journal.pone.0269715>
- Goyal, R., Singha, P., & Singh, S. K. (2024). Spectroscopic food adulteration detection using machine learning: Current challenges and future prospects. *Trends in Food Science & Technology*, 146, 104377. <https://doi.org/https://doi.org/10.1016/j.tifs.2024.104377>
- Granado-Lorencio, F., Donoso-Navarro, E., Sánchez-Siles, L. M., Blanco-Navarro, I., & Pérez-Sacristán, B. (2011). Bioavailability of β -cryptoxanthin in the presence of phytosterols: in vitro and in vivo studies. *Journal of Agricultural and Food Chemistry*, 59(21), 11819–11824. <https://doi.org/10.1021/jf202628w>
- Grand, R. J., Montgomery, R. K., Chitkara, D. K., & Büller, H. A. (2004). Carbohydrate and Lactose Malabsorption. In L. R. Johnson (Ed.), *Encyclopedia of Gastroenterology* (pp. 268–274). Elsevier. <https://doi.org/10.1016/B0-12-386860-2/00103-9>
- Grand View Research. (2023). *Functional Foods Market Size, Share & Trends Analysis Report By Ingredient (Carotenoids, Prebiotics & Probiotics, Fatty Acids, Dietary Fibers), By Product, By Application, By Region, And Segment Forecasts, 2022 - 2030*. <https://www.grandviewresearch.com/industry-analysis/functional-food-market>
- Grubbs, F. E. (1950). Sample Criteria for Testing Outlying Observations. *The Annals of Mathematical Statistics*, 21(1), 27–58. <https://doi.org/10.1214/aoms/1177729885>
- Grundy, M. M.-L., Edwards, C. H., Mackie, A. R., Gidley, M. J., Butterworth, P. J., & Ellis, P. R. (2016). Re-evaluation of the mechanisms of dietary fibre and implications for macronutrient bioaccessibility, digestion and postprandial metabolism. *British Journal of Nutrition*, 116(5), 816–833. <https://doi.org/DOI: 10.1017/S0007114516002610>

- Grundy, M. M. L., McClements, D. J., Ballance, S., & Wilde, P. J. (2018). Influence of oat components on lipid digestion using an in vitro model: Impact of viscosity and depletion flocculation mechanism. *Food Hydrocolloids*, 83, 253–264. <https://doi.org/10.1016/j.foodhyd.2018.05.018>
- Guerra, A., Etienne-Mesmin, L., Livrelli, V., Denis, S., Blanquet-Diot, S., & Alric, M. (2012). Relevance and challenges in modeling human gastric and small intestinal digestion. *Trends in Biotechnology*, 30(11), 591–600. <https://doi.org/https://doi.org/10.1016/j.tibtech.2012.08.001>
- Guo, Q., Bellissimo, N., & Rousseau, D. (2017). The Physical State of Emulsified Edible Oil Modulates Its in Vitro Digestion. *Journal of Agricultural and Food Chemistry*, 65(41), 9120–9127. <https://doi.org/10.1021/acs.jafc.7b03368>
- Gupta, A., Gulati, P., Kim, W., Fernandez, S., Shaker, R., & Jadcherla, S. R. (2009). Effect of postnatal maturation on the mechanisms of esophageal propulsion in preterm human neonates: primary and secondary peristalsis. *The American Journal of Gastroenterology*, 104(2), 411–419. <https://doi.org/10.1038/ajg.2008.32>
- Gutiérrez-Peña, R., Fernández-Cabanás, V. M., Mena, Y., & Delgado-Pertíñez, M. (2018). Fatty acid profile and vitamins A and E contents of milk in goat farms under Mediterranean wood pastures as affected by grazing conditions and seasons. *Journal of Food Composition and Analysis*, 72(December 2017), 122–131. <https://doi.org/10.1016/j.jfca.2018.07.003>
- Hackenberger, B. K. (2019). Bayes or not Bayes, is this the question? *Croatian Medical Journal*, 60(1), 50–52. <https://doi.org/10.3325/cmj.2019.60.50>
- Hamosh, M., Klaeveman, H. L., Wolf, R. O., & Scow, R. O. (1975). Pharyngeal lipase and digestion of dietary triglyceride in man. *The Journal of Clinical Investigation*, 55(5), 908–913. <https://doi.org/10.1172/JCI108019>
- Hamulka, J., Jeruszka-Bielak, M., Górnicka, M., Drywień, M. E., & Zielinska-Pukos, M. A. (2020). Dietary Supplements during COVID-19 Outbreak.

- Results of Google Trends Analysis Supported by PLifeCOVID-19 Online Studies. *Nutrients*, 13(1). <https://doi.org/10.3390/nu13010054>
- Hastie, T. (2020). Ridge Regularization: An Essential Concept in Data Science. *Technometrics: A Journal of Statistics for the Physical, Chemical, and Engineering Sciences*, 62(4), 426–433. <https://doi.org/10.1080/00401706.2020.1791959>
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 97–109. <https://doi.org/10.1093/biomet/57.1.97>
- Hatcher, H., Planalp, R., Cho, J., Torti, F. M., & Torti, S. V. (2008). Curcumin: from ancient medicine to current clinical trials. *Cellular and Molecular Life Sciences: CMLS*, 65(11), 1631–1652. <https://doi.org/10.1007/s00018-008-7452-4>
- He, C., Qi, J., Liao, J., Song, Y., & Wu, C. (2023). Excellent hydration properties and oil holding capacity of citrus fiber: Effects of component variation and microstructure. *Food Hydrocolloids*, 144, 108988. <https://doi.org/https://doi.org/10.1016/j.foodhyd.2023.108988>
- He, H.-J., Sun, D.-W., & Wu, D. (2014). Rapid and real-time prediction of lactic acid bacteria (LAB) in farmed salmon flesh using near-infrared (NIR) hyperspectral imaging combined with chemometric analysis. *Food Research International*, 62, 476–483. <https://doi.org/10.1016/j.foodres.2014.03.064>
- Heda, R., Toro, F., & Tombazzi, C. R. (2023). *Physiology, Pepsin*. StatPearls Publishing. <https://www.ncbi.nlm.nih.gov/books/NBK537005/>
- Hegde, M., Girisa, S., BharathwajChetty, B., Vishwa, R., & Kunnumakkara, A. B. (2023). Curcumin Formulations for Better Bioavailability: What We Learned from Clinical Trials Thus Far? *ACS Omega*, 8(12), 10713–10746. <https://doi.org/10.1021/acsomega.2c07326>
- Heiberger, R. M., & Holland, B. (2004). Linear Regression by Least Squares. In

Statistical Analysis and Data Display: An Intermediate Course with Examples in S-Plus, R, and SAS (pp. 187–213). Springer New York.
https://doi.org/10.1007/978-1-4757-4284-8_8

Helena de Abreu-Martins, H., Artiga-Artigas, M., Hilsdorf Piccoli, R., Martín-Belloso, O., & Salvia-Trujillo, L. (2020). The lipid type affects the in vitro digestibility and β -carotene bioaccessibility of liquid or solid lipid nanoparticles. *Food Chemistry*, 311(December 2019), 126024.
<https://doi.org/10.1016/j.foodchem.2019.126024>

Hellmig, S., Von Schöning, F., Gadow, C., Katsoulis, S., Hedderich, J., Fölsch, U. R., & Stüber, E. (2006). Gastric emptying time of fluids and solids in healthy subjects determined by ^{13}C breath tests: influence of age, sex and body mass index. *Journal of Gastroenterology and Hepatology*, 21(12), 1832–1838. <https://doi.org/10.1111/j.1440-1746.2006.04449.x>

Hernández-Prieto, D., Garre, A., Agulló, V., García-Viguera, C., & Egea, J. A. (2023). Differences Due to Sex and Sweetener on the Bioavailability of (Poly)phenols in Urine Samples: A Machine Learning Approach. *Metabolites*, 13(5). <https://doi.org/10.3390/metabo13050653>

Hespanhol, L., Vallio, C. S., Costa, L. M., & Saragiotto, B. T. (2019). Understanding and interpreting confidence and credible intervals around effect estimates. *Brazilian Journal of Physical Therapy*, 23(4), 290–301.
<https://doi.org/10.1016/j.bjpt.2018.12.006>

Hewlings, S. J., & Kalman, D. S. (2017). Curcumin: A review of its effects on human health. *Foods*, 6(10), 1–11. <https://doi.org/10.3390/foods6100092>

Hightower, N. C., Dworken, H. J., Sircus, W., & Keeton, W. T. (2023). *human digestive system*. Encyclopedia Britannica.
<https://www.britannica.com/science/human-digestive-system>

Hiolle, M., Lechevalier, V., Floury, J., Boulier-Monthéan, N., Prioul, C., Dupont, D., & Nau, F. (2020). In vitro digestion of complex foods: How microstructure

- influences food disintegration and micronutrient bioaccessibility. *Food Research International*, 128(November 2019). <https://doi.org/10.1016/j.foodres.2019.108817>
- Ho, T. M., Abik, F., Hietala, S., Isaza Ferro, E., Pitkänen, L., Juhl, D. W., Vosegaard, T., Kilpeläinen, P. O., & Mikkonen, K. S. (2023). Wood lignocellulosic stabilizers: effect of their characteristics on stability and rheological properties of emulsions. *Cellulose*, 30(2), 753–773. <https://doi.org/10.1007/s10570-022-04958-z>
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1), 55–67. <https://doi.org/10.1080/00401706.1970.10488634>
- Hoffman, M. D., Gelman, A., & others. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1), 1593–1623. <https://doi.org/10.5555/2627435.2638586>
- Hollis, J. H. (2018). The effect of mastication on food intake, satiety and body weight. *Physiology & Behavior*, 193, 242–245. <https://doi.org/https://doi.org/10.1016/j.physbeh.2018.04.027>
- Horn, F., Pack, R., & Rieger, M. (2020). The autofeat python library for automated feature engineering and selection. *Communications in Computer and Information Science*, 1167 CCIS, 111–120. https://doi.org/10.1007/978-3-030-43823-4_10
- Hou, T., Wang, J., Zhang, W., & Xu, X. (2007). ADME evaluation in drug discovery. 6. Can oral bioavailability in humans be effectively predicted by simple molecular property-based rules? *Journal of Chemical Information and Modeling*, 47(2), 460–463. <https://doi.org/10.1021/ci6003515>
- Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics & Proteomics*, 15(1), 41–51.

<https://doi.org/10.21873/cgp.20063>

- Huang, Y., Yu, Q., Chen, Z., Wu, W., Zhu, Q., & Lu, Y. (2021). In vitro and in vivo correlation for lipid-based formulations: Current status and future perspectives. *Acta Pharmaceutica Sinica B*, 11(8), 2469–2487. <https://doi.org/https://doi.org/10.1016/j.apsb.2021.03.025>
- Huey, S. L., Mehta, N. H., Konieczynski, E. M., Bhargava, A., Friesen, V. M., Krisher, J. T., Mbuya, M. N. N., Monterrosa, E., Nyangaresi, A. M., Boy, E., & Mehta, S. (2022). Bioaccessibility and bioavailability of biofortified food and food products: Current evidence. *Critical Reviews in Food Science and Nutrition*. <https://doi.org/10.1080/10408398.2022.2142762>
- Hwang, J., Kim, D.-K., Bae, J. H., Kang, S. H., Seo, K. M., Kim, B. K., & Lee, S. Y. (2012). The effect of rheological properties of foods on bolus characteristics after mastication. *Annals of Rehabilitation Medicine*, 36(6), 776–784. <https://doi.org/10.5535/arm.2012.36.6.776>
- Idowu, S. O., & Fatokun, A. A. (2021). Artificial Intelligence (AI) to the Rescue: Deploying Machine Learning to Bridge the Biorelevance Gap in Antioxidant Assays. *SLAS Technology*, 26(1), 16–25. <https://doi.org/10.1177/2472630320962716>
- Infantes-Garcia, M. R., Verkempinck, S. H. E., Gonzalez-Fuentes, P. G., Hendrickx, M. E., & Grauwet, T. (2021). Lipolysis products formation during in vitro gastric digestion is affected by the emulsion interfacial composition. *Food Hydrocolloids*, 110, 106163. <https://doi.org/https://doi.org/10.1016/j.foodhyd.2020.106163>
- Institute of Medicine (US) Panel on Dietary Antioxidants and Related Compounds. (2000). *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids*. National Academies Press. <https://doi.org/10.17226/9810>
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS*

Medicine, 2(8), e124. <https://doi.org/10.1371/journal.pmed.0020124>

- Iorgulescu, G. (2009). Saliva between normal and pathological. Important factors in determining systemic and oral health. *Journal of Medicine and Life*, 2(3), 303–307.
- Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and bayesian strategies. *Annals of Statistics*, 33(2), 730–773. <https://doi.org/10.1214/009053604000001147>
- Jackman, S. (2004). Bayesian analysis for political research. *Annual Review of Political Science*, 7, 483–505. <https://doi.org/10.1146/annurev.polisci.7.012003.104706>
- Jahn, S., Tsalis, G., & Lähteenmäki, L. (2019). How attitude towards food fortification can lead to purchase intention. *Appetite*, 133, 370–377. <https://doi.org/10.1016/j.appet.2018.11.022>
- Jaqaman, K., & Danuser, G. (2006). Linking data to models: data regression. *Nature Reviews Molecular Cell Biology*, 7(11), 813–819. <https://doi.org/10.1038/nrm2030>
- Jayaprakasha, G. K., Jagan Mohan Rao, L., & Sakariah, K. K. (2002). Improved HPLC Method for the Determination of Curcumin, Demethoxycurcumin, and Bisdemethoxycurcumin. *Journal of Agricultural and Food Chemistry*, 50(13), 3668–3672. <https://doi.org/10.1021/jf025506a>
- Ji, H., Pu, D., Yan, W., Zhang, Q., Zuo, M., & Zhang, Y. (2023). Recent advances and application of machine learning in food flavor prediction and regulation. *Trends in Food Science & Technology*, 138, 738–751. <https://doi.org/https://doi.org/10.1016/j.tifs.2023.07.012>
- Jiang, H., Somogyi, Á., Jacobsen, N. E., Timmermann, B. N., & Gang, D. R. (2006). Analysis of curcuminoids by positive and negative electrospray ionization and tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 20(6), 1001–1012.

<https://doi.org/https://doi.org/10.1002/rcm.2401>

- Jiang, Q. (2017). Natural Forms of Vitamin E as Effective Agents for Cancer Prevention and Therapy. *Advances in Nutrition*, 8(6), 850–867. <https://doi.org/https://doi.org/10.3945/an.117.016329>
- Jiao, S. R., Song, J., & Liu, B. (2020). A Review of Decision Tree Classification Algorithms for Continuous Variables. *Journal of Physics: Conference Series*, 1651(1), 12083. <https://doi.org/10.1088/1742-6596/1651/1/012083>
- Jovel, J., & Greiner, R. (2021). An Introduction to Machine Learning Approaches for Biomedical Research. *Frontiers in Medicine*, 8, 771607. <https://doi.org/10.3389/fmed.2021.771607>
- Kahai, P., Mandiga, P., Wehrle, C. J., & Lobo, S. (2023). *Anatomy, Abdomen and Pelvis: Large Intestine*.
- Kamiloglu, S., Tomas, M., Ozdal, T., & Capanoglu, E. (2021). Effect of food matrix on the content and bioavailability of flavonoids. *Trends in Food Science and Technology*, 117, 15–33. <https://doi.org/10.1016/j.tifs.2020.10.030>
- Ketel, E. C., Aguayo-Mendoza, M. G., de Wijk, R. A., de Graaf, C., Piqueras-Fiszman, B., & Stieger, M. (2019). Age, gender, ethnicity and eating capability influence oral processing behaviour of liquid, semi-solid and solid foods differently. *Food Research International*, 119, 143–151. <https://doi.org/https://doi.org/10.1016/j.foodres.2019.01.048>
- Khan, M. I. H., Sablani, S. S., Nayak, R., & Gu, Y. (2022). Machine learning-based modeling in food processing applications: State of the art. *Comprehensive Reviews in Food Science and Food Safety*, 21(2), 1409–1438. <https://doi.org/https://doi.org/10.1111/1541-4337.12912>
- Kiani, A. K., Dhuli, K., Donato, K., Aquilanti, B., Velluti, V., Matera, G., Iaconelli, A., Connelly, S. T., Bellinato, F., Gisondi, P., & Bertelli, M. (2022). Main nutritional deficiencies. *Journal of Preventive Medicine and Hygiene*, 63(2), E93–E101. <https://doi.org/10.15167/2421-4248/jpmh2022.63.2S3.2752>

- Kiela, P. R., & Ghishan, F. K. (2016). Physiology of Intestinal Absorption and Secretion. *Best Practice & Research. Clinical Gastroenterology*, 30(2), 145–159. <https://doi.org/10.1016/j.bpg.2016.02.007>
- Kingsford, C., & Salzberg, S. L. (2008). What are decision trees? *Nature Biotechnology*, 26(9), 1011–1013. <https://doi.org/10.1038/nbt0908-1011>
- Kong, F., & Singh, R. P. (2010). A human gastric simulator (HGS) to study food digestion in human stomach. *Journal of Food Science*, 75(9), E627-35. <https://doi.org/10.1111/j.1750-3841.2010.01856.x>
- Kotsiantis, S. B. (2014). Bagging and boosting variants for handling classifications problems: a survey. *The Knowledge Engineering Review*, 29(1), 78–100. <https://doi.org/10.1017/S0269888913000313>
- Krishnakumar, I., Abhilash, R., Dinesh, K., Ramadasan, K., & Balu, M. (2012). An enhanced bioavailable formulation of curcumin using fenugreek-derived soluble dietary fibre. *Journal of Functional Foods*, 4(1), 348–357. <https://doi.org/10.1016/j.jff.2012.01.004>
- Kumar, Rajnish, Sharma, A., & Varadwaj, P. K. (2011). A prediction model for oral bioavailability of drugs using physicochemical properties by support vector machine. *Journal of Natural Science, Biology and Medicine*, 2(2), 168–173. <https://doi.org/10.4103/0976-9668.92325>
- Kumar, Ravin, Carroll, C., Hartikainen, A., & Martin, O. (2019). ArviZ a unified library for exploratory analysis of Bayesian models in Python. *Journal of Open Source Software*, 4(33), 1143. <https://doi.org/10.21105/joss.01143>
- Kurien, B. T., Singh, A., Matsumoto, H., & Scofield, R. H. (2007). Improving the Solubility and Pharmacological Efficacy of Curcumin by Heat Treatment. *ASSAY and Drug Development Technologies*, 5(4), 567–576. <https://doi.org/10.1089/adt.2007.064>
- Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the boruta package. *Journal of Statistical Software*, 36(11), 1–13.

<https://doi.org/10.18637/jss.v036.i11>

- Kwoji, I. D., Aiyegoro, O. A., Okpeku, M., & Adeleke, M. A. (2023). "Multi-omics" data integration: applications in probiotics studies. *NPJ Science of Food*, 7(1), 25. <https://doi.org/10.1038/s41538-023-00199-x>
- Laachari, F., Maataoui, H., Bergadi, F., Boukir, A., & Saad, I. (2015). Improvement of a lipolytic activity assay protocol to increase the efficiency of the lipases activity measurement and validation by infrared spectroscopy Improvement of a lipolytic activity assay protocol to increase the efficiency of the lipases activity. *Journal, African Biotechnology*, March 2016.
- Lafond, M., Bouza, B., Eyrichine, S., Rouffineau, F., Saulnier, L., Giardina, T., Bonnin, E., & Preynat, A. (2015). In vitro gastrointestinal digestion study of two wheat cultivars and evaluation of xylanase supplementation. *Journal of Animal Science and Biotechnology*, 6(1), 5. <https://doi.org/10.1186/s40104-015-0002-7>
- Lai, W. Y. W., Chua, J. W. M., Gill, S., & Brownlee, I. A. (2019). Analysis of the Lipolytic Activity of Whole-Saliva and Site-Specific Secretions from the Oral Cavity of Healthy Adults. *Nutrients*, 11(1). <https://doi.org/10.3390/nu11010191>
- Lambert, B. (2018). A Student's Guide to Bayesian Statistics / Ben Lambert. In *A Student's Guide to Bayesian Statistics*. SAGE LTD.
- Lan, X., Liu, Y., Wang, L., Wang, H., Hu, Z., Dong, H., Yu, Z., & Yuan, Y. (2023). A review of curcumin in food preservation: Delivery system and photosensitization. *Food Chemistry*, 424, 136464. <https://doi.org/https://doi.org/10.1016/j.foodchem.2023.136464>
- Langston, F. M. A., Nash, G. R., & Bows, J. R. (2023). The retention and bioavailability of phytochemicals in the manufacturing of baked snacks. *Critical Reviews in Food Science and Nutrition*, 63(14), 2141–2177. <https://doi.org/10.1080/10408398.2021.1971944>

- Le Feunteun, S., Mackie, A. R., & Dupont, D. (2020). In silico trials of food digestion and absorption: how far are we? *Current Opinion in Food Science*, 31, 121–125. <https://doi.org/https://doi.org/10.1016/j.cofs.2020.04.006>
- Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, 20(7), 1434–1448. <https://doi.org/10.1364/josaa.20.001434>
- Lee, Y. Y., Erdogan, A., & Rao, S. S. C. (2014). How to assess regional and whole gut transit time with wireless motility capsule. *Journal of Neurogastroenterology and Motility*, 20(2), 265–270. <https://doi.org/10.5056/jnm.2014.20.2.265>
- Leeman, M., Östman, E., & Björck, I. (2008). Glycaemic and satiating properties of potato products. *European Journal of Clinical Nutrition*, 62(1), 87–95. <https://doi.org/10.1038/sj.ejcn.1602677>
- Lemhadri, I., Ruan, F., & Tibshirani, R. (2021). LassoNet: Neural Networks with Feature Sparsity. *Proceedings of Machine Learning Research*, 130, 10–18. <http://www.ncbi.nlm.nih.gov/pubmed/36092461>
- Li, P., Tang, S., Chen, S., Tian, X., & Zhong, N. (2023). Hyperspectral imaging combined with convolutional neural network for accurately detecting adulteration in Atlantic salmon. *Food Control*, 147, 109573. <https://doi.org/https://doi.org/10.1016/j.foodcont.2022.109573>
- Li, R., Chen, X., Yan, H., Deurenberg, P., Garby, L., & Hautvast, J. G. (1994). Functional consequences of iron supplementation in iron-deficient female cotton mill workers in Beijing, China. *The American Journal of Clinical Nutrition*, 59(4), 908–913. <https://doi.org/10.1093/ajcn/59.4.908>
- Li, Y., & McClements, D. J. (2014). Modulating lipid droplet intestinal lipolysis by electrostatic complexation with anionic polysaccharides: Influence of cosurfactants. *Food Hydrocolloids*, 35, 367–374. <https://doi.org/https://doi.org/10.1016/j.foodhyd.2013.06.011>

- Lillford, P. J. (2011). The Importance Of Food Microstructure In Fracture Physics And Texture Perception. *Journal of Texture Studies*, 42(2), 130–136. <https://doi.org/10.1111/j.1745-4603.2011.00293.x>
- Liskova, A., Stefanicka, P., Samec, M., Smejkal, K., Zubor, P., Bielik, T., Biskupska-Bodova, K., Kwon, T. K., Danko, J., Büsselberg, D., Adamek, M., Rodrigo, L., Kruzliak, P., Shleikin, A., & Kubatka, P. (2020). Dietary phytochemicals as the potential protectors against carcinogenesis and their role in cancer chemoprevention. *Clinical and Experimental Medicine*, 20(2), 173–190. <https://doi.org/10.1007/s10238-020-00611-w>
- Little, T. J., Horowitz, M., & Feinle-Bisset, C. (2007). Modulation by high-fat diets of gastrointestinal function and hormones associated with the regulation of energy intake: implications for the pathophysiology of obesity. *The American Journal of Clinical Nutrition*, 86(3), 531–541. <https://doi.org/10.1093/ajcn/86.3.531>
- Liu, G., Xiao, L., Fang, T., Cai, Y., Jia, G., Zhao, H., Wang, J., Chen, X., & Wu, C. (2014). Pea Fiber and Wheat Bran Fiber Show Distinct Metabolic Profiles in Rats as Investigated by a 1H NMR-Based Metabolomic Approach. *PLOS ONE*, 9(12), 1–23. <https://doi.org/10.1371/journal.pone.0115561>
- Liu, H.-T., & Ho, Y.-S. (2018). Anticancer effect of curcumin on breast cancer and stem cells. *Food Science and Human Wellness*, 7(2), 134–137. <https://doi.org/https://doi.org/10.1016/j.fshw.2018.06.001>
- Liu, Yong, Song, M., Bai, H., Wang, C., Wang, F., & Yuan, Q. (2024). Curcumin improves the egg quality, antioxidant activity, and intestinal microbiota of quails during the late laying period. *Poultry Science*, 103(1), 103233. <https://doi.org/https://doi.org/10.1016/j.psj.2023.103233>
- Liu, Yueyue, Ma, M., & Yuan, Y. (2023). The potential of curcumin-based co-delivery systems for applications in the food industry: Food preservation, freshness monitoring, and functional food. *Food Research International*, 171(June), 113070. <https://doi.org/10.1016/j.foodres.2023.113070>

- Llanos, O. L., Konturek, S. J., Rayford, P. L., & Thompson, J. C. (1977). Pancreatic bicarbonate, serum gastrin, and secretin responses to meals varying in pH. *American Journal of Physiology-Endocrinology and Metabolism*, 233(1), E41. <https://doi.org/10.1152/ajpendo.1977.233.1.E41>
- Loh, W.-Y. (2014). Fifty Years of Classification and Regression Trees. *International Statistical Review*, 82(3), 329–348. <https://doi.org/10.1111/insr.12016>
- Lordan, R. (2021). Dietary supplements and nutraceuticals market growth during the coronavirus pandemic - Implications for consumers and regulatory oversight. In *PharmaNutrition* (Vol. 18, p. 100282). <https://doi.org/10.1016/j.phanu.2021.100282>
- Lowe, N. M. (2021). The global challenge of hidden hunger: perspectives from the field. *The Proceedings of the Nutrition Society*, 80(3), 283–289. <https://doi.org/10.1017/S0029665121000902>
- Lu, X., Zhu, J., Pan, Y., & Huang, Q. (2019). Assessment of dynamic bioaccessibility of curcumin encapsulated in milled starch particle stabilized Pickering emulsions using TNO's gastrointestinal model. *Food Funct.*, 10(5), 2583–2594. <https://doi.org/10.1039/C8FO02495B>
- Lv, D., Grayling, M. J., Zhang, X., Zhao, Q., & Zheng, H. (2023). A Bayesian approach to pilot-pivotal trials for bioequivalence assessment. *BMC Medical Research Methodology*, 23(1), 301. <https://doi.org/10.1186/s12874-023-02120-2>
- Mackie, A., Mulet-Cabero, A. I., & Torcello-Gomez, A. (2020). Simulating human digestion: Developing our knowledge to create healthier and more sustainable foods. *Food and Function*, 11(11), 9397–9431. <https://doi.org/10.1039/d0fo01981j>
- MacMahon, M. T., & Neale, G. (1970). The Absorption of α -Tocopherol in Control Subjects and in Patients with Intestinal Malabsorption. *Clinical Science*,

38(2), 197–210. <https://doi.org/10.1042/cs0380197>

- Mah, E., Sapper, T. N., Chitchumroonchokchai, C., Failla, M. L., Schill, K. E., Clinton, S. K., Bobe, G., Traber, M. G., & Bruno, R. S. (2015). α -Tocopherol bioavailability is lower in adults with metabolic syndrome regardless of dairy fat co-ingestion: a randomized, double-blind, crossover trial¹²³. *The American Journal of Clinical Nutrition*, 102(5), 1070–1080. <https://doi.org/https://doi.org/10.3945/ajcn.115.118570>
- Mahaboob, B., Venkateswarlu, B., Narayana, C., Ravi sankar, J., & Balasiddamuni, P. (2018). A Treatise on Ordinary Least Squares Estimation of Parameters of Linear Model. *International Journal of Engineering & Technology*, 7(4.10), 518. <https://doi.org/10.14419/ijet.v7i4.10.21216>
- Mandal, P. P., Rezaee, R., & Emelyanova, I. (2022). Ensemble Learning for Predicting TOC from Well-Logs of the Unconventional Goldwyer Shale. *Energies*, 15(1). <https://doi.org/10.3390/en15010216>
- Mandel, A. L., Peyrot des Gachons, C., Plank, K. L., Alarcon, S., & Breslin, P. A. S. (2010). Individual differences in AMY1 gene copy number, salivary α -amylase levels, and the perception of oral starch. *PloS One*, 5(10), e13352. <https://doi.org/10.1371/journal.pone.0013352>
- Mannar, M. G. V. (2006). Successful food-based programmes, supplementation and fortification. In *Journal of pediatric gastroenterology and nutrition: Vol. 43 Suppl 3* (pp. S47-53). <https://doi.org/10.1097/01.mpg.0000255850.30400.ae>
- Marjolein Fokkema, J. E.-C., & Wolpert, M. (2021). Generalized linear mixed-model (GLMM) trees: A flexible decision-tree method for multilevel and longitudinal data. *Psychotherapy Research*, 31(3), 329–341. <https://doi.org/10.1080/10503307.2020.1785037>
- Markussen, J. Ø., Madsen, F., Young, J. F., & Corredig, M. (2021). A semi dynamic in vitro digestion study of milk protein concentrate dispersions

- structured with different polysaccharides. *Current Research in Food Science*, 4(March), 250–261. <https://doi.org/10.1016/j.crfs.2021.03.012>
- Martin-Rubio, A. S., Sopelana, P., & Guillén, M. D. (2019). Influence of minor components on lipid bioaccessibility and oxidation during in vitro digestion of soybean oil. *Journal of the Science of Food and Agriculture*, 99(10), 4793–4800. <https://doi.org/https://doi.org/10.1002/jsfa.9734>
- Martin, C. R., & Freedman, S. D. (2019). Chapter 3 - Lipid and Fatty Acid Delivery in the Preterm Infant: Challenges and Lessons Learned from Other Critically Ill Populations. In J. Neu & B. Poindexter (Eds.), *Gastroenterology and Nutrition (Third Edition)* (Third Edit, pp. 29–41). Elsevier. <https://doi.org/https://doi.org/10.1016/B978-0-323-54502-0.00003-7>
- Martirosyan, D., Lampert, T., & Ekblad, M. (2022). Classification and regulation of functional food proposed by the Functional Food Center. *Functional Food Science*, 2(2), 25. <https://doi.org/10.31989/ffs.v2i2.890>
- Marze, S. (2015). Refining in silico simulation to study digestion parameters affecting the bioaccessibility of lipophilic nutrients and micronutrients. *Food & Function*, 6(1), 115–124. <https://doi.org/10.1039/c4fo00483c>
- Matvienko, O. A., Lewis, D. S., Swanson, M., Arndt, B., Rainwater, D. L., Stewart, J., & Alekel, D. L. (2002). A single daily dose of soybean phytosterols in ground beef decreases serum total cholesterol and LDL cholesterol in young, mildly hypercholesterolemic men. *The American Journal of Clinical Nutrition*, 76(1), 57–64. <https://doi.org/10.1093/ajcn/76.1.57>
- Mayeaux, M., Xu, Z., King, J. M., & Prinyawiwatukul, W. (2006). Effects of Cooking Conditions on the Lycopene Content in Tomatoes. *Journal of Food Science*, 71(8), 1–5. <https://doi.org/10.1111/j.1750-3841.2006.00163.x>
- McClements, D. J., & Peng, S.-F. (2020). Current status in our understanding of physicochemical basis of bioaccessibility. *Current Opinion in Food Science*, 31, 57–62. <https://doi.org/10.1016/j.cofs.2019.11.005>

- McClements, I. F., & McClements, D. J. (2023). Designing healthier plant-based foods: Fortification, digestion, and bioavailability. *Food Research International*, 169, 112853. <https://doi.org/10.1016/j.foodres.2023.112853>
- Meca, G., Giri, A., Sagratini, G., Vittori, S., Font, G., & Mañes, J. (2012). The soluble dietary fiber inulin can influence the bioaccessibility of enniatins. *Food and Function*, 3(8), 853–858. <https://doi.org/10.1039/c2fo00004k>
- Meca, G., Meneghelli, G., Ritieni, A., Mañes, J., & Font, G. (2012). Influence of different soluble dietary fibers on the bioaccessibility of the minor *Fusarium* mycotoxin beauvericin. *Food and Chemical Toxicology*, 50(5), 1362–1368. <https://doi.org/10.1016/j.fct.2012.02.038>
- Memarzia, A., Khazdair, M. R., Behrouz, S., Gholamnezhad, Z., Jafarnejhad, M., Saadat, S., & Boskabady, M. H. (2021). Experimental and clinical reports on anti-inflammatory, antioxidant, and immunomodulatory effects of *Curcuma longa* and curcumin, an updated and comprehensive review. *BioFactors*, 47(3), 311–350. <https://doi.org/10.1002/biof.1716>
- Ménard, O., Cattenoz, T., Guillemin, H., Souchon, I., Deglaire, A., Dupont, D., & Picque, D. (2014). Validation of a new in vitro dynamic system to simulate infant digestion. *Food Chemistry*, 145, 1039–1045. <https://doi.org/10.1016/j.foodchem.2013.09.036>
- Menard, O., Lesmes, U., Shani-Levi, C. S., Araiza Calahorra, A., Lavoisier, A., Morzel, M., Rieder, A., Feron, G., Nebbia, S., Mashiah, L., Andres, A., Bornhorst, G., Carrière, F., Egger, L., Gwala, S., Heredia, A., Kirkhus, B., Macierzanka, A., Portman, R., ... Dupont, D. (2023). Static in vitro digestion model adapted to the general older adult population: an INFOGEST international consensus. *Food & Function*, 14(10), 4569–4582. <https://doi.org/10.1039/d3fo00535f>
- Menichetti, G., Ravandi, B., Mozaffarian, D., & Barabási, A.-L. (2023). Machine learning prediction of the degree of food processing. *Nature*

- Communications*, 14(1), 2312. <https://doi.org/10.1038/s41467-023-37457-1>
- Menrad, K. (2003). Market and marketing of functional food in Europe. *Journal of Food Engineering*, 56(2), 181–188. [https://doi.org/https://doi.org/10.1016/S0260-8774\(02\)00247-9](https://doi.org/https://doi.org/10.1016/S0260-8774(02)00247-9)
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6), 1087–1092. <https://doi.org/10.1063/1.1699114>
- Metropolis, N., & Ulam, S. (1949). The Monte Carlo Method. *Journal of the American Statistical Association*, 44(247), 335–341. <https://doi.org/10.1080/01621459.1949.10483310>
- Micucci, M., Budriesi, R., Mandrioli, M., Tura, M., Corazza, I., Frosini, M., Aldini, R., Mattioli, L. B., & Gallina Toschi, T. (2022). Effects of turmeric powder on intestinal and biliary functions: The influence of curcuminoids concentration on spontaneous contractility. *Journal of Functional Foods*, 99, 105314. <https://doi.org/https://doi.org/10.1016/j.jff.2022.105314>
- Miles, C. W. (1992). The metabolizable energy of diets differing in dietary fat and fiber measured in humans. *The Journal of Nutrition*, 122(2), 306–311. <https://doi.org/10.1093/jn/122.2.306>
- Minekus, M., Alminger, M., Alvito, P., Ballance, S., Bohn, T., Bourlieu, C., Carrière, F., Boutrou, R., Corredig, M., Dupont, D., Dufour, C., Egger, L., Golding, M., Karakaya, S., Kirkhus, B., Le Feunteun, S., Lesmes, U., MacIerzanka, A., MacKie, A., ... Brodkorb, A. (2014). A standardised static in vitro digestion method suitable for food-an international consensus. *Food and Function*, 5(6), 1113–1124. <https://doi.org/10.1039/c3fo60702j>
- Minekus, Mans, Marteau, P., Havenaar, R., & Veld, J. H. J. H. in't. (1995). A Multicompartmental Dynamic Computer-controlled Model Simulating the Stomach and Small Intestine. *Alternatives to Laboratory Animals*, 23(2),

197–209. <https://doi.org/10.1177/026119299502300205>

- Mirosa, M., & Mangan-Walker, E. (2018). Young Chinese and Functional Foods for Mobility Health: Perceptions of Importance, Trust, and Willingness to Purchase and Pay a Premium. *Journal of Food Products Marketing*, 24(2), 216–234. <https://doi.org/10.1080/10454446.2017.1266555>
- Mishra, A. K., Singh, R., Rawat, H., Kumar, V., Jagtap, C., & Jain, A. (2024). The influence of food matrix on the stability and bioavailability of phytochemicals: A comprehensive review. *Food and Humanity*, 2(December 2023), 100202. <https://doi.org/10.1016/j.foohum.2023.12.010>
- Mittal, R. K. (2011). *Motor Function of the Pharynx, Esophagus, and its Sphincters*. <https://doi.org/10.4199/C00027ED1V01Y201103ISP016>
- Miyazawa, T., Hiratsuka, Y., Toda, M., Hatakeyama, N., Ozawa, H., Abe, C., Cheng, T.-Y., Matsushima, Y., Miyawaki, Y., Ashida, K., Iimura, J., Tsuda, T., Bushita, H., Tomonobu, K., Ohta, S., Chung, H., Omae, Y., Yamamoto, T., Morinaga, M., ... Miyazawa, T. (2022). Artificial intelligence in food science and nutrition: a narrative review. *Nutrition Reviews*, 80(12), 2288–2300. <https://doi.org/10.1093/nutrit/nuac033>
- Monnahan, C. C., Thorson, J. T., & Branch, T. A. (2017). Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution*, 8(3), 339–348. <https://doi.org/https://doi.org/10.1111/2041-210X.12681>
- Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-89010-0>
- Morales, J. F., Ruiz, M. E., Stratford, R. E., & Talevi, A. (2024). Application of machine learning to predict unbound drug bioavailability in the brain. *Frontiers in Drug Discovery*, 4(April), 1–14. <https://doi.org/10.3389/fddsv.2024.1360732>

- Morgan, J. N., & Sonquist, J. A. (1963). Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association*, 58(302), 415–434. <https://doi.org/10.1080/01621459.1963.10500855>
- Moxon, T. E., Gouseti, O., & Bakalis, S. (2016). In silico modelling of mass transfer & absorption in the human gut. *Journal of Food Engineering*, 176, 110–120. <https://doi.org/https://doi.org/10.1016/j.jfoodeng.2015.10.019>
- Mudie, D. M., Amidon, G. L., & Amidon, G. E. (2010). Physiological parameters for oral delivery and in vitro testing. *Molecular Pharmaceutics*, 7(5), 1388–1405. <https://doi.org/10.1021/mp100149j>
- Mulet-Cabero, A. I., Egger, L., Portmann, R., Ménard, O., Marze, S., Minekus, M., Le Feunteun, S., Sarkar, A., Grundy, M. M. L., Carrière, F., Golding, M., Dupont, D., Recio, I., Brodkorb, A., & Mackie, A. (2020). A standardised semi-dynamic: in vitro digestion method suitable for food-an international consensus. *Food and Function*, 11(2), 1702–1720. <https://doi.org/10.1039/c9fo01293a>
- Mulrooney, S. L., O'Neill, G. J., Brougham, D. F., & O'Riordan, D. (2021). Vitamin D3 bioaccessibility: Influence of fatty acid chain length, salt concentration and l- α -phosphatidylcholine concentration on mixed micelle formation and delivery of vitamin D3. *Food Chemistry*, 344, 128722. <https://doi.org/https://doi.org/10.1016/j.foodchem.2020.128722>
- Mutsokoti, L., Panozzo, A., Pallares Pallares, A., Jaiswal, S., Van Loey, A., Grauwet, T., & Hendrickx, M. (2017). Carotenoid bioaccessibility and the relation to lipid digestion: A kinetic study. *Food Chemistry*, 232, 124–134. <https://doi.org/https://doi.org/10.1016/j.foodchem.2017.04.001>
- Mutsvari, T., Tytgat, D., & Walley, R. (2016). Addressing potential prior-data conflict when using informative priors in proof-of-concept studies. *Pharmaceutical Statistics*, 15(1), 28–36. <https://doi.org/10.1002/pst.1722>
- Nardone, G., Compare, D., & Rocco, A. (2017). A microbiota-centric view of

- diseases of the upper gastrointestinal tract. *The Lancet Gastroenterology & Hepatology*, 2(4), 298–312. [https://doi.org/10.1016/S2468-1253\(16\)30108-X](https://doi.org/10.1016/S2468-1253(16)30108-X)
- National Research Council (US) Committee on Diet and Health. (1990). Diet and Health: Implications for Reducing Chronic Disease Risk. In *Washington (DC): National Academies Press (US)* (Vol. 11, Issue Fat-Soluble Vitamins.). <https://doi.org/10.5860/choice.27-3332>
- Navas, C. M., Wadas, E. D., Zbib, N. H., Crowell, M. D., & Lacy, B. E. (2021). Gastroparesis and Severity of Delayed Gastric Emptying: Comparison of Patient Characteristics, Treatments and Medication Adverse Events. *Digestive Diseases and Sciences*, 66(2), 526–534. <https://doi.org/10.1007/s10620-020-06258-7>
- Nawirska, A., & Kwaśniewska, M. (2005). Dietary fibre fractions from fruit and vegetable processing waste. *Food Chemistry*, 91(2), 221–225. <https://doi.org/https://doi.org/10.1016/j.foodchem.2003.10.005>
- Nelson, K. M., Dahlin, J. L., Bisson, J., Graham, J., Pauli, G. F., & Walters, M. A. (2017). The Essential Medicinal Chemistry of Curcumin. *Journal of Medicinal Chemistry*, 60(5), 1620–1637. <https://doi.org/10.1021/acs.jmedchem.6b00975>
- Nguyen, H. T., Marquis, M., Anton, M., & Marze, S. (2019). Studying the real-time interplay between triglyceride digestion and lipophilic micronutrient bioaccessibility using droplet microfluidics. 1 lab on a chip method. *Food Chemistry*, 275(July 2018), 523–529. <https://doi.org/10.1016/j.foodchem.2018.09.096>
- Nichols, B. L., Baker, S. S., & Quezada-Calvillo, R. (2018). Metabolic Impacts of Maltase Deficiencies. *Journal of Pediatric Gastroenterology and Nutrition*, 66 Suppl 3, S24–S29. <https://doi.org/10.1097/MPG.0000000000001955>
- Nicolescu, A., Babotă, M., Barros, L., Rocchetti, G., Lucini, L., Tanase, C., Mocan, A., Bunea, C. I., & Crișan, G. (2023). Bioaccessibility and bioactive

- potential of different phytochemical classes from nutraceuticals and functional foods. *Frontiers in Nutrition*, 10, 1184535. <https://doi.org/10.3389/fnut.2023.1184535>
- Niki, E., & Abe, K. (2019). Vitamin E: Structure, Properties and Functions. In *Vitamin E: Chemistry and Nutritional Benefits*. The Royal Society of Chemistry. <https://doi.org/10.1039/9781788016216-00001>
- Nisbet, R., Miner, G., & Yale, K. (2018). Chapter 7 - Basic Algorithms for Data Mining: A Brief Overview. In R. Nisbet, G. Miner, & K. Yale (Eds.), *Handbook of Statistical Analysis and Data Mining Applications (Second Edition)* (Second Edi, pp. 121–147). Academic Press. <https://doi.org/https://doi.org/10.1016/B978-0-12-416632-5.00007-4>
- Noble, R. E. (2000). Salivary alpha-amylase and lysozyme levels: a non-invasive technique for measuring parotid vs submandibular/sublingual gland activity. *Journal of Oral Science*, 42(2), 83–86. <https://doi.org/10.2334/josnusd.42.83>
- Novotny, J. A., Fadel, J. G., Holstege, D. M., Furr, H. C., & Clifford, A. J. (2012). This kinetic, bioavailability, and metabolism study of RRR- α -tocopherol in healthy adults suggests lower intake requirements than previous estimates. *The Journal of Nutrition*, 142(12), 2105–2111. <https://doi.org/10.3945/jn.112.166462>
- Novotny, J. A., Gebauer, S. K., & Baer, D. J. (2012). Discrepancy between the Atwater factor predicted and empirically measured energy values of almonds in human diets. *The American Journal of Clinical Nutrition*, 96(2), 296–301. <https://doi.org/10.3945/ajcn.112.035782>
- Núñez-Gómez, V., González-Barrio, R., & Periago, M. J. (2023). Interaction between Dietary Fibre and Bioactive Compounds in Plant By-Products: Impact on Bioaccessibility and Bioavailability. *Antioxidants*, 12(4). <https://doi.org/10.3390/antiox12040976>
- Nychas, G.-J., Sims, E., Tsakanikas, P., & Mohareb, F. (2021). Data Science in

- the Food Industry. *Annual Review of Biomedical Data Science*, 4(Volume 4, 2021), 341–367. <https://doi.org/https://doi.org/10.1146/annurev-biodatasci-020221-123602>
- Nystrand, B. T., & Olsen, S. O. (2020). Consumers' attitudes and intentions toward consuming functional foods in Norway. *Food Quality and Preference*, 80, 103827. <https://doi.org/https://doi.org/10.1016/j.foodqual.2019.103827>
- O'brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity*, 41(5), 673–690. <https://doi.org/10.1007/s11135-006-9018-6>
- Ogobuiro, I., Gonzales, J., Shumway, K. R., & Tuma, F. (2023). *Physiology, Gastrointestinal*.
- Ogut, J. O., Schulz-Streeck, T., & Piepho, H.-P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC Proceedings*, 6 Suppl 2(Suppl 2), S10. <https://doi.org/10.1186/1753-6561-6-S2-S10>
- Olson, R., Gavin-Smith, B., Ferraboschi, C., & Kraemer, K. (2021). Food Fortification: The Advantages, Disadvantages and Lessons from Sight and Life Programs. *Nutrients*, 13(4). <https://doi.org/10.3390/nu13041118>
- Ovesen, L. (1997). Regulatory aspects of functional foods. In *European journal of cancer prevention : the official journal of the European Cancer Prevention Organisation (ECP)* (Vol. 6, Issue 5, pp. 480–482). <https://doi.org/10.1097/00008469-199710000-00009>
- Owen, K. N., & Dewald, O. (2023). *Vitamin E Toxicity*. StatPearls Publishing, Treasure Island (FL). <http://europepmc.org/books/NBK564373>
- Palafox-Carlos, H., Ayala-Zavala, J. F., & González-Aguilar, G. A. (2011). The Role of Dietary Fiber in the Bioaccessibility and Bioavailability of Fruit and Vegetable Antioxidants. *Journal of Food Science*, 76(1), R6–R15. <https://doi.org/https://doi.org/10.1111/j.1750-3841.2010.01957.x>

- Palmer, J. B., Rudin, N. J., Lara, G., & Crompton, A. W. (1992). Coordination of mastication and swallowing. *Dysphagia*, 7(4), 187–200. <https://doi.org/10.1007/BF02493469>
- Panagou, E. Z., Papadopoulou, O., Carstensen, J. M., & Nychas, G.-J. E. (2014a). Potential of multispectral imaging technology for rapid and non-destructive determination of the microbiological quality of beef filets during aerobic storage. *International Journal of Food Microbiology*, 174, 1–11.
- Panagou, E. Z., Papadopoulou, O., Carstensen, J. M., & Nychas, G.-J. E. (2014b). Potential of multispectral imaging technology for rapid and non-destructive determination of the microbiological quality of beef filets during aerobic storage. *International Journal of Food Microbiology*, 174, 1–11. <https://doi.org/10.1016/j.ijfoodmicro.2013.12.026>
- Pancholi, V., Smina, T. P., Kunnumakkara, A. B., Maliakel, B., & Krishnakumar, I. M. (2021). Safety assessment of a highly bioavailable curcumin-galactomannoside complex (CurQfen) in healthy volunteers, with a special reference to the recent hepatotoxic reports of curcumin supplements: A 90-days prospective study. *Toxicology Reports*, 8, 1255–1264. <https://doi.org/https://doi.org/10.1016/j.toxrep.2021.06.008>
- Parada, J., & Aguilera, J. M. (2007). Food microstructure affects the bioavailability of several nutrients. *Journal of Food Science*, 72(2), 21–32. <https://doi.org/10.1111/j.1750-3841.2007.00274.x>
- Park, C., Meghani, N., Loebenberg, R., Cui, J.-H., Cao, Q.-R., & Lee, B.-J. (2020). Fatty acid chain length impacts nanonizing capacity of albumin-fatty acid nanomicelles: Enhanced physicochemical property and cellular delivery of poorly water-soluble drug. *European Journal of Pharmaceutics and Biopharmaceutics*, 152, 257–269. <https://doi.org/https://doi.org/10.1016/j.ejpb.2020.05.011>
- Park, T., & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482), 681–686.

<https://doi.org/10.1198/0162145080000000337>

Patricia, J. J., & Dhamoon, A. S. (2023). *Physiology, Digestion*.

Paul, A. (2019). Drug Absorption and Bioavailability. In G. M. Raj & R. Raveendran (Eds.), *Introduction to Basics of Pharmacology and Toxicology: Volume 1: General and Molecular Pharmacology: Principles of Drug Action* (pp. 81–88). Springer Singapore. https://doi.org/10.1007/978-981-32-9779-1_5

Pedersen, A., Sørensen, C. E., Proctor, G. B., & Carpenter, G. H. (2018). Salivary functions in mastication, taste and textural perception, swallowing and initial digestion. *Oral Diseases*, 24(8), 1399–1416. <https://doi.org/10.1111/odi.12867>

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85), 2825–2830. <http://jmlr.org/papers/v12/pedregosa11a.html>

Peixoto, R. R. A., Mazon, E. A. M., & Cadore, S. (2013). Estimation of the Bioaccessibility of Metallic Elements in Chocolate Drink Powder using an in vitro Digestion Method and Spectrometric Techniques. *Journal of the Brazilian Chemical Society*, 24. <https://doi.org/10.5935/0103-5053.20130111>

Piironen, J., & Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2), 5018–5051. <https://doi.org/10.1214/17-EJS1337SI>

Pilosof, A. M. R. (2017). Potential impact of interfacial composition of proteins and polysaccharides stabilized emulsions on the modulation of lipolysis. The role of bile salts. *Food Hydrocolloids*, 68, 178–185.

<https://doi.org/https://doi.org/10.1016/j.foodhyd.2016.08.030>

Popkin, B. M., Adair, L. S., & Ng, S. W. (2012). Global nutrition transition and the pandemic of obesity in developing countries. *Nutrition Reviews*, 70(1), 3–21.

<https://doi.org/10.1111/j.1753-4887.2011.00456.x>

Prasad, C., Davis, K. E., Imrhan, V., Juma, S., & Vijayagopal, P. (2019). Advanced Glycation End Products and Risks for Chronic Diseases: Intervening Through Lifestyle Modification. *American Journal of Lifestyle Medicine*, 13(4), 384–404. <https://doi.org/10.1177/1559827617708991>

Prasad, S., & Aggarwal, B. B. (2011). Turmeric, the golden spice: From traditional medicine to modern medicine. In I. F. F. Benzie & S. Wachtel-Galor (Eds.), *Herbal Medicine: Biomolecular and Clinical Aspects: Second Edition* (pp. 263–288).

Precedence Research. (2023). *Functional Food Market (By Product: Dairy Products, Bakery & Cereals, Meat, Fish & Eggs, Fats & Oils, Soy Products, and Others; By Ingredient: Dietary Fibers, Carotenoids, Minerals, Fatty Acids, Vitamins, Prebiotics & Probiotics, and Others; By Applicatio*. <https://www.precedenceresearch.com/functional-food-market>

Price, G., & Patel, D. A. (2024). Drug Bioavailability. In *StatPearls*. <http://www.ncbi.nlm.nih.gov/pubmed/29724803>

Prosperini, A., Meca, G., Font, G., & Ruiz, M. J. (2013). Bioaccessibility of enniatins A, A1, B, and B1 in different commercial breakfast cereals, cookies, and breads of Spain. *Journal of Agricultural and Food Chemistry*, 61(2), 456–461. <https://doi.org/10.1021/jf3044023>

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Frontiers in Bioinformatics*, 2, 927312. <https://doi.org/10.3389/fbinf.2022.927312>

Puri, V., Nagpal, M., Singh, I., Singh, M., Dhingra, G. A., Huanbutta, K., Dheer,

- D., Sharma, A., & Sangnim, T. (2022). A Comprehensive Review on Nutraceuticals: Therapy Support and Formulation Challenges. *Nutrients*, 14(21). <https://doi.org/10.3390/nu14214637>
- Qazi, H. J., Ye, A., Acevedo-Fani, A., & Singh, H. (2021). In vitro digestion of curcumin-nanoemulsion-enriched dairy protein matrices: Impact of the type of gel structure on the bioaccessibility of curcumin. *Food Hydrocolloids*, 117, 106692. <https://doi.org/10.1016/j.foodhyd.2021.106692>
- Rahimi, A., & Recht, B. (2007). Random Features for Large-Scale Kernel Machines. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in Neural Information Processing Systems* (Vol. 20). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf
- Razavi, R., & Kenari, R. E. (2023). Ultraviolet–visible spectroscopy combined with machine learning as a rapid detection method to the predict adulteration of honey. *Heliyon*, 9(10), e20973. <https://doi.org/10.1016/j.heliyon.2023.e20973>
- Reboul, E. (2013). Absorption of Vitamin A and Carotenoids by the Enterocyte: Focus on Transport Proteins. *Nutrients*, 5(9), 3563–3581. <https://doi.org/10.3390/nu5093563>
- Reboul, E. (2017). Vitamin e bioavailability: Mechanisms of intestinal absorption in the spotlight. *Antioxidants*, 6(4). <https://doi.org/10.3390/antiox6040095>
- Reddy, P., & Jialal, I. (2023). Biochemistry, Fat Soluble Vitamins. In *StatPearls*. <http://www.ncbi.nlm.nih.gov/pubmed/24782580>
- Reddy, V. P., Aryal, P., & Darkwah, E. K. (2022). Advanced Glycation End Products in Health and Disease. *Microorganisms*, 10(9). <https://doi.org/10.3390/microorganisms10091848>
- Rein, M. J., Renouf, M., Cruz-Hernandez, C., Actis-Goretti, L., Thakkar, S. K., & da Silva Pinto, M. (2013). Bioavailability of bioactive food compounds: A

- challenging journey to bioefficacy. *British Journal of Clinical Pharmacology*, 75(3), 588–602. <https://doi.org/10.1111/j.1365-2125.2012.04425.x>
- Ribnicky, D. M., Roopchand, D. E., Oren, A., Grace, M., Poulev, A., Lila, M. A., Havenaar, R., & Raskin, I. (2014). Effects of a high fat meal matrix and protein complexation on the bioaccessibility of blueberry anthocyanins using the TNO gastrointestinal model (TIM-1). *Food Chemistry*, 142, 349–357. <https://doi.org/10.1016/j.foodchem.2013.07.073>
- Rigatti, S. J. (2017). Random Forest. *Journal of Insurance Medicine*, 47(1), 31–39. <https://doi.org/10.17849/insm-47-01-31-39.1>
- Rivera Del Rio, A., van der Wielen, N., Gerrits, W. J. J., Boom, R. M., & Janssen, A. E. M. (2022). In silico modelling of protein digestion: A case study on solid/liquid and blended meals. *Food Research International (Ottawa, Ont.)*, 157, 111271. <https://doi.org/10.1016/j.foodres.2022.111271>
- Rizvi, S., Raza, S. T., Ahmed, F., Ahmad, A., Abbas, S., & Mahdi, F. (2014). The role of vitamin e in human health and some diseases. *Sultan Qaboos University Medical Journal*, 14(2), e157-65.
- Robert, C. P. (2007). *The Bayesian choice : from decision-theoretic foundations to computational implementation*. <https://api.semanticscholar.org/CorpusID:50937448>
- Robertson, T. M., Alzaabi, A. Z., Robertson, M. D., & Fielding, B. A. (2018). Starchy Carbohydrates in a Healthy Diet: The Role of the Humble Potato. *Nutrients*, 10(11). <https://doi.org/10.3390/nu10111764>
- Rocks, J. W., & Mehta, P. (2022). Memorizing without overfitting: Bias, variance, and interpolation in overparameterized models. *Physical Review Research*, 4(1). <https://doi.org/10.1103/physrevresearch.4.013201>
- Rodrigues, M., Miguéis, V., Freitas, S., & Machado, T. (2024). Machine learning models for short-term demand forecasting in food catering services: A solution to reduce food waste. *Journal of Cleaner Production*, 435, 140265.

<https://doi.org/https://doi.org/10.1016/j.jclepro.2023.140265>

- Rodríguez-Roque, M. J., Rojas-Graü, M. A., Elez-Martínez, P., & Martín-Belloso, O. (2013). Changes in vitamin C, phenolic, and carotenoid profiles throughout in vitro gastrointestinal digestion of a blended fruit juice. *Journal of Agricultural and Food Chemistry*, 61(8), 1859–1867. <https://doi.org/10.1021/jf3044204>
- Rojas-Bonzi, P., Vangsøe, C. T., Nielsen, K. L., Lærke, H. N., Hedemann, M. S., & Knudsen, K. E. B. (2020). The Relationship between In Vitro and In Vivo Starch Digestion Kinetics of Breads Varying in Dietary Fibre. *Foods*, 9(9). <https://doi.org/10.3390/foods9091337>
- Rosenthal, R. (1994). Science and Ethics in Conducting, Analyzing, and Reporting Psychological Research. *Psychological Science*, 5(3), 127–134. <https://doi.org/10.1111/j.1467-9280.1994.tb00646.x>
- Rudi, K., & Zhao, L. (2021). Grand Challenges in Understanding Gut Microbes. *Frontiers in Microbiology*, 12. <https://doi.org/10.3389/fmicb.2021.752829>
- Rukh, G., Ericson, U., Andersson-Assarsson, J., Orho-Melander, M., & Sonestedt, E. (2017). Dietary starch intake modifies the relation between copy number variation in the salivary amylase gene and BMI1. *The American Journal of Clinical Nutrition*, 106(1), 256–262. <https://doi.org/https://doi.org/10.3945/ajcn.116.149831>
- Saadatfar, H., Khosravi, S., Joloudari, J. H., Mosavi, A., & Shamshirband, S. (2020). A New K-Nearest Neighbors Classifier for Big Data Based on Efficient Data Pruning. *Mathematics*, 8(2). <https://doi.org/10.3390/math8020286>
- Sadilek, A., Caty, S., DiPrete, L., Mansour, R., Schenk, T., Bergtholdt, M., Jha, A., Ramaswami, P., & Gabrilovich, E. (2018). Machine-learned epidemiology: real-time detection of foodborne illness at scale. *Npj Digital Medicine*, 1(1), 36. <https://doi.org/10.1038/s41746-018-0045-1>

- Saeyns, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507–2517. <https://doi.org/10.1093/bioinformatics/btm344>
- Saldanha, L., & Vale, N. (2023). The First Physiologically Based Pharmacokinetic (PBPK) Model for an Oral Vaccine Using Alpha-Tocopherol as an Adjuvant. *Pharmaceutics*, 15(9). <https://doi.org/10.3390/pharmaceutics15092313>
- Salem, M., Rohani, S., & Gillies, E. R. (2014). Curcumin, a promising anti-cancer therapeutic: A review of its chemical properties, bioactivity and approaches to cancer cell delivery. *RSC Advances*, 4(21), 10815–10829. <https://doi.org/10.1039/c3ra46396f>
- Salhi, A., Carriere, F., Grundy, M. M.-L., & Aloulou, A. (2021). Enzymes Involved in Lipid Digestion. In M. M.-L. Grundy & P. J. Wilde (Eds.), *Bioaccessibility and Digestibility of Lipids from Food* (pp. 3–28). Springer International Publishing. https://doi.org/10.1007/978-3-030-56909-9_1
- Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2016(4), 1–24. <https://doi.org/10.7717/peerj-cs.55>
- Samanta, S., Sarkar, T., & Chakraborty, R. (2024). Multifunctional applications of natural colorants: Preservative, functional ingredient, and sports supplements. *Biocatalysis and Agricultural Biotechnology*, 56, 103026. <https://doi.org/https://doi.org/10.1016/j.bcab.2024.103026>
- Samtiya, M., Aluko, R. E., & Dhewa, T. (2020). Plant food anti-nutritional factors and their reduction strategies: an overview. *Food Production, Processing and Nutrition*, 2(1), 6. <https://doi.org/10.1186/s43014-020-0020-5>
- Saura-Calixto, F., Serrano, J., & Goñi, I. (2007). Intake and bioaccessibility of total polyphenols in a whole diet. *Food Chemistry*, 101(2), 492–501. <https://doi.org/10.1016/j.foodchem.2006.02.006>
- Sayanjali, S., Sanguansri, L., Ying, D., Buckow, R., Gras, S., & Augustin, M. A.

- (2019). Extrusion of a Curcuminoid-Enriched Oat Fiber-Corn-Based Snack Product. *Journal of Food Science*, 84(2), 284–291. <https://doi.org/10.1111/1750-3841.14432>
- Sayyar, Z., & Jafarizadeh-Malmiri, H. (2024). Enhancing the efficacy of nano-curcumin on cancer cells through mixture design optimization of three emulsifiers. *BMC Chemistry*, 18(1), 62. <https://doi.org/10.1186/s13065-024-01160-z>
- Schaafsma, G. (2005). The Protein Digestibility-Corrected Amino Acid Score (PDCAAS)—A Concept for Describing Protein Quality in Foods and Food Ingredients: A Critical Review. *Journal of AOAC INTERNATIONAL*, 88(3), 988–994. <https://doi.org/10.1093/jaoac/88.3.988>
- Scheltens, P., Kamphuis, P. J. G. H., Verhey, F. R. J., Olde Rikkert, M. G. M., Wurtman, R. J., Wilkinson, D., Twisk, J. W. R., & Kurz, A. (2010). Efficacy of a medical food in mild Alzheimer's disease: A randomized, controlled trial. *Alzheimer's & Dementia : The Journal of the Alzheimer's Association*, 6(1), 1-10.e1. <https://doi.org/10.1016/j.jalz.2009.10.003>
- Schmidt, A. F., & Finan, C. (2018). Linear regression and the normality assumption. *Journal of Clinical Epidemiology*, 98, 146–151. <https://doi.org/10.1016/j.jclinepi.2017.12.006>
- Schneckener, S., Grimbs, S., Hey, J., Menz, S., Osmers, M., Schaper, S., Hillisch, A., & Göller, A. H. (2019). Prediction of Oral Bioavailability in Rats: Transferring Insights from in Vitro Correlations to (Deep) Machine Learning Models Using in Silico Model Outputs and Chemical Structure Parameters. *Journal of Chemical Information and Modeling*, 59(11), 4893–4905. <https://doi.org/10.1021/acs.jcim.9b00460>
- Schneider, C., Gordon, O. N., Edwards, R. L., & Luis, P. B. (2015). Degradation of Curcumin: From Mechanism to Biological Implications. *Journal of Agricultural and Food Chemistry*, 63(35), 7606–7614. <https://doi.org/10.1021/acs.jafc.5b00244>

- Schneider, F., Grimm, M., Koziol, M., Modeß, C., Dokter, A., Roustom, T., Siegmund, W., & Weitschies, W. (2016). Resolving the physiological conditions in bioavailability and bioequivalence studies: Comparison of fasted and fed state. *European Journal of Pharmaceutics and Biopharmaceutics*, 108, 214–219. <https://doi.org/https://doi.org/10.1016/j.ejpb.2016.09.009>
- Seidi Damyeh, M., Mereddy, R., Netzel, M. E., & Sultanbawa, Y. (2020). An insight into curcumin-based photosensitization as a promising and green food preservation technology. *Comprehensive Reviews in Food Science and Food Safety*, 19(4), 1727–1759. <https://doi.org/https://doi.org/10.1111/1541-4337.12583>
- Sensoy, I. (2021). A review on the food digestion in the digestive tract and the used in vitro models. *Current Research in Food Science*, 4(April), 308–319. <https://doi.org/10.1016/j.crfs.2021.04.004>
- Shamsi-Goushki, A., Mortazavi, Z., Mirshekar, M. A., Mohammadi, M., Moradi-Kor, N., Jafari-Maskouni, S., & Shahraki, M. (2020). Comparative effects of curcumin versus nano-curcumin on insulin resistance, serum levels of apelin and lipid profile in type 2 diabetic rats. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, 13, 2337–2346. <https://doi.org/10.2147/DMSO.S247351>
- Shargel, L., & Yu, A. B. C. (2015). *Applied Biopharmaceutics & Pharmacokinetics, Seventh Edition*. McGraw-Hill Education. <https://books.google.co.uk/books?id=SrP9rQEACAAJ>
- Sharma, K. D., Karki, S., Thakur, N. S., & Attri, S. (2012). Chemical composition, functional properties and processing of carrot-a review. *Journal of Food Science and Technology*, 49(1), 22–32. <https://doi.org/10.1007/s13197-011-0310-7>
- Sharma, P., Trivedi, N., & Gat, Y. (2017). Development of functional fermented whey-oat-based product using probiotic bacteria. *3 Biotech*, 7(4), 272.

<https://doi.org/10.1007/s13205-017-0906-3>

- Shewry, P. R., Joy, E. J. M., De La Revilla, L. S., Hansen, A., Brennan, J., & Lovegrove, A. (2023). Increasing fibre in white flour and bread: Implications for health and processing. *Nutrition Bulletin*, 48(4), 587–593. <https://doi.org/https://doi.org/10.1111/nbu.12648>
- Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP Journal on Wireless Communications and Networking*, 2021(1), 31. <https://doi.org/10.1186/s13638-021-01910-w>
- Siegrist, M., Stampfli, N., & Kastenholz, H. (2008). Consumers' willingness to buy functional foods. The influence of carrier, benefit and trust. *Appetite*, 51(3), 526–529. <https://doi.org/10.1016/j.appet.2008.04.003>
- Smith, M. E., & Morton, D. G. (2010). 8 - DIGESTION AND ABSORPTION. In M. E. Smith & D. G. Morton (Eds.), *The Digestive System (Second Edition)* (Second Edi, pp. 129–152). Churchill Livingstone. <https://doi.org/https://doi.org/10.1016/B978-0-7020-3367-4.00008-6>
- Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199–222. <https://doi.org/10.1023/B:STCO.0000035301.49549.88>
- Song, G., Jiang, N., Ge, Y., Li, F., Zhou, L., Xiang, T., Wang, D., Yuan, T., Li, L., Luo, Z., & Gong, J. (2025). Improvement of curcumin loading properties and bioaccessibility of beta-lactoglobulin-hyaluronic acid nanocomplexes conjugated via ultrasound-assisted Maillard reaction. *International Journal of Biological Macromolecules*, 288, 138710. <https://doi.org/https://doi.org/10.1016/j.ijbiomac.2024.138710>
- Sousa, R., Portmann, R., Dubois, S., Recio, I., & Egger, L. (2020). Protein digestion of different protein sources using the INFOGEST static digestion model. *Food Research International*, 130(September 2019), 108996.

<https://doi.org/10.1016/j.foodres.2020.108996>

Sousa, R., Recio, I., Heimo, D., Dubois, S., Moughan, P. J., Hodgkinson, S. M., Portmann, R., & Egger, L. (2023). In vitro digestibility of dietary proteins and in vitro DIAAS analytical workflow based on the INFOGEST static protocol and its validation with in vivo data. *Food Chemistry*, 404(Pt B), 134720. <https://doi.org/10.1016/j.foodchem.2022.134720>

Stan Development Team. (2022). *Stan Modeling Language User's Guide and Reference Manual, Version 2.33*. <https://mc-stan.org/docs/reference-manual/index.html/>

Stanton, J. M. (2001). Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors. *Journal of Statistics Education*, 9(3). <https://doi.org/10.1080/10691898.2001.11910537>

Steinbach, E., Masi, D., Ribeiro, A., Serradas, P., Le Roy, T., & Clément, K. (2024). Upper small intestine microbiome in obesity and related metabolic disorders: A new field of investigation. *Metabolism*, 150, 155712. <https://doi.org/10.1016/j.metabol.2023.155712>

Steinmetz, K. A., & Potter, J. D. (1991). Vegetables, fruit, and cancer. I. Epidemiology. *Cancer Causes & Control*, 2(5), 325–357. <https://doi.org/10.1007/BF00051672>

Stinco, C. M., Pumilia, G., Giuffrida, D., Dugo, G., Meléndez-Martínez, A. J., & Vicario, I. M. (2019). Bioaccessibility of carotenoids, vitamin A and α -tocopherol, from commercial milk-fruit juice beverages: Contribution to the recommended daily intake. *Journal of Food Composition and Analysis*, 78(January), 24–32. <https://doi.org/10.1016/j.jfca.2019.01.019>

Straits Research. (2024). *Functional Food Market Size, Share & Trends Analysis Report By Ingredient (Carotenoids, Dietary Fibers, Fatty Acids, Minerals, Prebiotics & Probiotics, Vitamins, Others), By Product (Bakery & Cereals, Dairy Products, Meat, Fish & Eggs, Soy Products, Fats*.

<https://straitsresearch.com/report/functional-food-market>

- Sun, Y., Hayakawa, S., Jiang, H., Ogawa, M., & Izumori, K. (2006). Rheological characteristics of heat-induced custard pudding gels with high antioxidative activity. *Bioscience, Biotechnology and Biochemistry*, 70(12), 2859–2867. <https://doi.org/10.1271/bbb.60256>
- Sy, C., Gleize, B., Dangles, O., Landrier, J. F., Veyrat, C. C., & Borel, P. (2012). Effects of physicochemical properties of carotenoids on their bioaccessibility, intestinal cell uptake, and blood and tissue concentrations. *Molecular Nutrition and Food Research*, 56(9), 1385–1397. <https://doi.org/10.1002/mnfr.201200041>
- Szewczyk, K., Chojnacka, A., & Górnicka, M. (2021). Tocopherols and tocotrienols—bioactive dietary compounds; what is certain, what is doubt? *International Journal of Molecular Sciences*, 22(12). <https://doi.org/10.3390/ijms22126222>
- Tabanelli, R., Brogi, S., & Calderone, V. (2021). Improving curcumin bioavailability: Current strategies and future perspectives. *Pharmaceutics*, 13(10). <https://doi.org/10.3390/pharmaceutics13101715>
- Tan, A., Doss, H., & Hobert, J. P. (2015). Honest Importance Sampling with Multiple Markov Chains. *Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, 24(3), 792–826. <https://doi.org/10.1080/10618600.2014.929523>
- Tan, Y., & McClements, D. J. (2021). Improving the bioavailability of oil-soluble vitamins by optimizing food matrix effects: A review. *Food Chemistry*, 348. <https://doi.org/10.1016/j.foodchem.2021.129148>
- Tan, Y., Zhang, Z., Zhou, H., Xiao, H., & McClements, D. J. (2020). Factors impacting lipid digestion and β -carotene bioaccessibility assessed by standardized gastrointestinal model (INFOGEST): oil droplet concentration.

Food Funct., 11(8), 7126–7137. <https://doi.org/10.1039/D0FO01506G>

Tan, Y., Zhou, H., & McClements, D. J. (2022). Application of static in vitro digestion models for assessing the bioaccessibility of hydrophobic bioactives: A review. *Trends in Food Science & Technology*, 122, 314–327. <https://doi.org/https://doi.org/10.1016/j.tifs.2022.02.028>

Teixeira, M. C., Severino, P., Andreani, T., Boonme, P., Santini, A., Silva, A. M., & Souto, E. B. (2017). d - α -tocopherol nanoemulsions: Size properties, rheological behavior, surface tension, osmolarity and cytotoxicity. *Saudi Pharmaceutical Journal*, 25(2), 231–235. <https://doi.org/10.1016/j.jsps.2016.06.004>

Temple, N. J. (2022). A rational definition for functional foods: A perspective. *Frontiers in Nutrition*, 9, 957516. <https://doi.org/10.3389/fnut.2022.957516>

Thishya, K., Vattam, K. K., Naushad, S. M., Raju, S. B., & Kutala, V. K. (2018). Artificial neural network model for predicting the bioavailability of tacrolimus in patients with renal transplantation. *PloS One*, 13(4), e0191921. <https://doi.org/10.1371/journal.pone.0191921>

Thuenemann, E. C., Mandalari, G., Rich, G. T., & Faulks, R. M. (2015). *Dynamic Gastric Model (DGM)*. (K. Verhoeckx, P. Cotter, I. López-Expósito, C. Kleiveland, T. Lea, A. Mackie, T. Requena, D. Swiatecka, & H. Wichers (eds.); pp. 47–59). https://doi.org/10.1007/978-3-319-16104-4_6

Thursby, E., & Juge, N. (2017). Introduction to the human gut microbiota. *The Biochemical Journal*, 474(11), 1823–1836. <https://doi.org/10.1042/BCJ20160510>

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>

Tomeh, M. A., Hadianamrei, R., & Zhao, X. (2019). A Review of Curcumin and Its Derivatives as Anticancer Agents. *International Journal of Molecular*

- Sciences*, 20(5). <https://doi.org/10.3390/ijms20051033>
- Topolska, K., Florkiewicz, A., & Filipiak-Florkiewicz, A. (2021). Functional Food-Consumer Motivations and Expectations. *International Journal of Environmental Research and Public Health*, 18(10). <https://doi.org/10.3390/ijerph18105327>
- Tougui, I., Jilbab, A., & Mhamdi, J. El. (2021). Impact of the Choice of Cross-Validation Techniques on the Results of Machine Learning-Based Diagnostic Applications. *Healthcare Informatics Research*, 27(3), 189–199. <https://doi.org/10.4258/hir.2021.27.3.189>
- Tsakanikas, P., Karnavas, A., Panagou, E. Z., & Nychas, G.-J. (2020). A machine learning workflow for raw food spectroscopic classification in a future industry. *Scientific Reports*, 10(1), 11212. <https://doi.org/10.1038/s41598-020-68156-2>
- Tseng, Y. J., Chuang, P.-J., & Appell, M. (2023). When Machine Learning and Deep Learning Come to the Big Data in Food Chemistry. *ACS Omega*, 8(18), 15854–15864. <https://doi.org/10.1021/acsomega.2c07722>
- Tsuda, T. (2018). Curcumin as a functional food-derived factor: Degradation products, metabolites, bioactivity, and future perspectives. *Food and Function*, 9(2), 705–714. <https://doi.org/10.1039/c7fo01242j>
- Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., Kanazawa, M., VanderGheynst, J., Fiehn, O., & Arita, M. (2015). MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature Methods*, 12(6), 523–526. <https://doi.org/10.1038/nmeth.3393>
- Tuorila, H., & Hartmann, C. (2020). Consumer responses to novel and unfamiliar foods. *Current Opinion in Food Science*, 33, 1–8. <https://doi.org/https://doi.org/10.1016/j.cofs.2019.09.004>
- Turgeon, S. L., & Brisson, G. (2020). Symposium review: The dairy matrix—

- Bioaccessibility and bioavailability of nutrients and physiological effects. *Journal of Dairy Science*, 103(7), 6727–6736. <https://doi.org/10.3168/jds.2019-17308>
- Tyagi, S., Pathak, A., Raghavendra Rao, N., Nehra, S., Asthana, A., Sharma, V., Katyal, G., Bhardwaj, A., Sharma, L., Prakash, S., & Katiyar, D. (2023). AI-assisted Formulation Design for Improved Drug Delivery and Bioavailability. *Pak Heart J*, 56(03), 149–162. <http://www.pkheartjournal.com>
- Ubeyitogullari, A., & Ciftci, O. N. (2019). A novel and green nanoparticle formation approach to forming low-crystallinity curcumin nanoparticles to improve curcumin's bioaccessibility. *Scientific Reports*, 9(1), 19112. <https://doi.org/10.1038/s41598-019-55619-4>
- Ungurianu, A., Zanfirescu, A., Nițulescu, G., & Margină, D. (2021). Vitamin E beyond Its Antioxidant Label. *Antioxidants (Basel, Switzerland)*, 10(5). <https://doi.org/10.3390/antiox10050634>
- United Nations. (2015). UN General Assembly, Transforming our world: The 2030 agenda for sustainable development. *Resolution Adopted by the General Assembly on 25 September 2015*, 16301(October), 1–35. <https://www.refworld.org/docid/57b6e3e44.html>
- Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PloS One*, 14(11), e0224365. <https://doi.org/10.1371/journal.pone.0224365>
- Van Dongen, S. (2006). Prior specification in Bayesian statistics: three cautionary tales. *Journal of Theoretical Biology*, 242(1), 90–100. <https://doi.org/10.1016/j.jtbi.2006.02.002>
- Van Hul, M., Karnik, K., Canene-Adams, K., De Souza, M., Van den Abbeele, P., Marzorati, M., Delzenne, N. M., Everard, A., & Cani, P. D. (2020). Comparison of the effects of soluble corn fiber and fructooligosaccharides on metabolism, inflammation, and gut microbiome of high-fat diet-fed mice.

- American Journal of Physiology. Endocrinology and Metabolism*, 319(4), E779–E791. <https://doi.org/10.1152/ajpendo.00108.2020>
- van Kleef, E., van Trijp, H. C. M., & Luning, P. (2005). Functional foods: health claim-food product compatibility and the impact of health claim framing on consumer evaluation. *Appetite*, 44(3), 299–308. <https://doi.org/10.1016/j.appet.2005.01.009>
- Vapnik, V. N. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24, 774–780. <https://api.semanticscholar.org/CorpusID:115205884>
- Vardakou, M., Mercuri, A., Barker, S. A., Craig, D. Q. M., Faulks, R. M., & Wickham, M. S. J. (2011). Achieving antral grinding forces in biorelevant in vitro models: comparing the USP dissolution apparatus II and the dynamic gastric model with human in vivo data. *AAPS PharmSciTech*, 12(2), 620–626. <https://doi.org/10.1208/s12249-011-9616-z>
- Vardar, U. S., Bitter, J. H., & Nikiforidis, C. V. (2024). The mechanism of encapsulating curcumin into oleosomes (Lipid Droplets). *Colloids and Surfaces B: Biointerfaces*, 236, 113819. <https://doi.org/https://doi.org/10.1016/j.colsurfb.2024.113819>
- Veintimilla-Gozalbo, E., Asensio-Grau, A., Calvo-Lerma, J., Heredia, A., & Andrés, A. (2021). In Vitro Simulation of Human Colonic Fermentation: A Practical Approach towards Models' Design and Analytical Tools. *Applied Sciences*, 11(17). <https://doi.org/10.3390/app11178135>
- Verbeke, W. (2006). Functional foods: Consumer willingness to compromise on taste for health? *Food Quality and Preference*, 17(1), 126–131. <https://doi.org/https://doi.org/10.1016/j.foodqual.2005.03.003>
- Verbeke, W., Scholderer, J., & Lähteenmäki, L. (2009). Consumer appeal of nutrition and health claims in three existing product concepts. *Appetite*, 52(3), 684–692. <https://doi.org/https://doi.org/10.1016/j.appet.2009.03.007>

- Verhoeckx, K., Cotter, P., López-Expósito, I., Kleiveland, C., Lea, T., Mackie, A., Requena, T., Swiatecka, D., & Wichers, H. (2015). *The Impact of Food Bioactives on Health* (K. Verhoeckx, P. Cotter, I. López-Expósito, C. Kleiveland, T. Lea, A. Mackie, T. Requena, D. Swiatecka, & H. Wichers (eds.)). Springer International Publishing. <https://doi.org/10.1007/978-3-319-16104-4>
- Verwei, M., Freidig, A. P., Havenaar, R., & Groten, J. P. (2006). Predicted serum folate concentrations based on in vitro studies and kinetic modeling are consistent with measured folate concentrations in humans. *The Journal of Nutrition*, 136(12), 3074–3078. <https://doi.org/10.1093/jn/136.12.3074>
- Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, 25(1), 35–57. <https://doi.org/10.3758/s13423-017-1343-3>
- Walsh, J. H., Richardson, C. T., & Fordtran, J. S. (1975). pH dependence of acid secretion and gastrin release in normal and ulcer subjects. *The Journal of Clinical Investigation*, 55(3), 462–468. <https://doi.org/10.1172/JCI107952>
- Walton, K. L. W. (2009). Teaching the role of secretin in the regulation of gastric acid secretion using a classic paper by Johnson and Grossman. *Advances in Physiology Education*, 33(3), 165–168. <https://doi.org/10.1152/advan.00023.2009>
- Wang, Q., Gao, C., Yang, N., & Nishinari, K. (2021). Effect of simulated saliva components on the in vitro digestion of peanut oil body emulsion. *RSC Advances*, 11(49), 30520–30531. <https://doi.org/10.1039/d1ra03274g>
- Wang, R., Mohammadi, M., Mahboubi, A., & Taherzadeh, M. J. (2021). In-vitro digestion models: a critical review for human and fish and a protocol for in-vitro digestion in fish. *Bioengineered*, 12(1), 3040–3064. <https://doi.org/10.1080/21655979.2021.1940769>

- Wang, X., Lin, Q., Ye, A., Han, J., & Singh, H. (2019). Flocculation of oil-in-water emulsions stabilised by milk protein ingredients under gastric conditions: Impact on in vitro intestinal lipid digestion. *Food Hydrocolloids*, 88, 272–282. <https://doi.org/https://doi.org/10.1016/j.foodhyd.2018.10.001>
- Watanabe, S., & Dawes, C. (1988). The effects of different foods and concentrations of citric acid on the flow rate of whole saliva in man. *Archives of Oral Biology*, 33(1), 1–5. [https://doi.org/10.1016/0003-9969\(88\)90089-1](https://doi.org/10.1016/0003-9969(88)90089-1)
- Weffort, V. R. S., & Lamounier, J. A. (2023). Hidden hunger - a narrative review. *Jornal de Pediatria*. <https://doi.org/10.1016/j.jped.2023.08.009>
- Wei, M., Zhang, X., Pan, X., Wang, B., Ji, C., Qi, Y., & Zhang, J. Z. H. (2022). HobPre: accurate prediction of human oral bioavailability for small molecules. *Journal of Cheminformatics*, 14(1), 1. <https://doi.org/10.1186/s13321-021-00580-6>
- Weisberg, S. (2014). *Applied Linear Regression* (Fourth). Wiley. <http://z.umn.edu/alr4ed>
- Wells, J. C. K., Marphatia, A. A., Amable, G., Siervo, M., Friis, H., Miranda, J. J., Haisma, H. H., & Raubenheimer, D. (2021). The future of human malnutrition: rebalancing agency for better nutritional health. *Globalization and Health*, 17(1), 119. <https://doi.org/10.1186/s12992-021-00767-4>
- Westbury, C. F. (2010). Bayes' rule for clinicians: an introduction. *Frontiers in Psychology*, 1, 192. <https://doi.org/10.3389/fpsyg.2010.00192>
- White, M., Aguirre, E., Finegood, D. T., Holmes, C., Sacks, G., & Smith, R. (2020). What role should the commercial food system play in promoting health through better diet? *BMJ*, 368. <https://doi.org/10.1136/bmj.m545>
- Wickham, M. J. S., Faulks, R. M., Mann, J., & Mandalari, G. (2012). The design, operation, and application of a dynamic gastric model. *Dissolution Technologies*, 19(3), 15–22. <https://doi.org/10.14227/DT190312P15>
- Wijekoon, M. M. J. O., Mahmood, K., Ariffin, F., Mohammadi Nafchi, A., &

- Zulkurnain, M. (2023). Recent advances in encapsulation of fat-soluble vitamins using polysaccharides, proteins, and lipids: A review on delivery systems, formulation, and industrial applications. *International Journal of Biological Macromolecules*, 241, 124539. <https://doi.org/10.1016/j.ijbiomac.2023.124539>
- Wolberg, E. J. (2010). The Method of Least Squares. In *Designing Quantitative Experiments: Prediction Analysis* (pp. 47–89). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-11589-9_3
- Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), 67–82. <https://doi.org/10.1109/4235.585893>
- Wolverton, S. E. (2020). Basic Principles of Pharmacology. In *Comprehensive Dermatologic Drug Therapy, Fourth Edition* (Fourth Edi). Elsevier. <https://doi.org/10.1016/B978-0-323-61211-1.00001-2>
- Wu, M. S., Taylor, C., Zahradka, P., & Arntfield, S. (2023). Reduced in vitro starch hydrolysis and in vivo glycemic effects after addition of soy presscake to corn tortillas. *Journal of the Science of Food and Agriculture*, 103(15), 7829–7835. <https://doi.org/10.1002/jsfa.12863>
- Wu, P., & Chen, X. D. (2021). Validation of in vitro bioaccessibility assays — a key aspect in the rational design of functional foods towards tailored bioavailability. *Current Opinion in Food Science*, 39, 160–170. <https://doi.org/https://doi.org/10.1016/j.cofs.2021.03.002>
- Wu, Y., Wang, S., Shi, M., Wang, X., Liu, H., Guo, S., Tan, L., Yang, X., Wu, X., & Hao, L. (2022). Awareness of nutrition and health knowledge and its influencing factors among Wuhan residents. *Frontiers in Public Health*, 10. <https://doi.org/10.3389/fpubh.2022.987755>
- Wulandari, A., Sunarti, T. C., Fahma, F., & Enomae, T. (2020). The potential of bioactives as biosensors for detection of pH. *IOP Conference Series: Earth*

and Environmental Science, 460(1), 012034. <https://doi.org/10.1088/1755-1315/460/1/012034>

- Xavier, A. A. O., & Mariutti, L. R. B. (2021). Static and semi-dynamic in vitro digestion methods: state of the art and recent achievements towards standardization. *Current Opinion in Food Science*, 41, 260–273. <https://doi.org/https://doi.org/10.1016/j.cofs.2021.08.002>
- Xie, J., Matthews, F. E., Jagger, C., Bond, J., & Brayne, C. (2008). The oldest old in England and Wales: a descriptive analysis based on the MRC Cognitive Function and Ageing Study. *Age and Ageing*, 37(4), 396–402. <https://doi.org/10.1093/ageing/afn061>
- Xin, L., & Seo, S. (Sunny). (2020). The role of consumer ethnocentrism, country image, and subjective knowledge in predicting intention to purchase imported functional foods. *British Food Journal*, 122(2), 448–464. <https://doi.org/10.1108/BFJ-05-2019-0326>
- Xu, Y., & Goodacre, R. (2018). On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Journal of Analysis and Testing*, 2(3), 249–262. <https://doi.org/10.1007/s41664-018-0068-2>
- Yang, S., & De Angelis, D. (2013). Maximum Likelihood. In B. Reisfeld & A. N. Mayeno (Eds.), *Computational Toxicology: Volume II* (pp. 581–595). Humana Press. https://doi.org/10.1007/978-1-62703-059-5_24
- Yang, Y., Xiao, H., & McClements, D. J. (2017). Impact of Lipid Phase on the Bioavailability of Vitamin E in Emulsion-Based Delivery Systems: Relative Importance of Bioaccessibility, Absorption, and Transformation. *Journal of Agricultural and Food Chemistry*, 65(19), 3946–3955. <https://doi.org/10.1021/acs.jafc.7b00955>
- Yao, X., Nie, K., Chen, Y., Jiang, F., Kuang, Y., Yan, H., Fang, Y., Yang, H.,

- Nishinari, K., & Phillips, G. O. (2018). The influence of non-ionic surfactant on lipid digestion of gum Arabic stabilized oil-in-water emulsion. *Food Hydrocolloids*, 74, 78–86. <https://doi.org/https://doi.org/10.1016/j.foodhyd.2017.07.043>
- Young, S. L., Blanco, I., Hernandez-Cordero, S., Pelto, G. H., & Neufeld, L. M. (2010). Organoleptic properties, ease of use, and perceived health effects are determinants of acceptability of micronutrient supplements among poor Mexican women. *The Journal of Nutrition*, 140(3), 605–611. <https://doi.org/10.3945/jn.109.113498>
- Zabot, G. L., Schaefer Rodrigues, F., Polano Ody, L., Vinícius Tres, M., Herrera, E., Palacin, H., Córdova-Ramos, J. S., Best, I., & Olivera-Montenegro, L. (2022). Encapsulation of Bioactive Compounds for Food and Agricultural Applications. *Polymers*, 14(19). <https://doi.org/10.3390/polym14194194>
- Zagury, Y., David, S., Edelman, R., Hazan Brill, R., & Livney, Y. D. (2021). Sugar beet pectin as a natural carrier for curcumin, a water-insoluble bioactive for food and beverage enrichment: Formation and characterization. *Innovative Food Science and Emerging Technologies*, 74(July), 102858. <https://doi.org/10.1016/j.ifset.2021.102858>
- Żarnowski, A., Jankowski, M., & Gujski, M. (2022). Public Awareness of Diet-Related Diseases and Dietary Risk Factors: A 2022 Nationwide Cross-Sectional Survey among Adults in Poland. *Nutrients*, 14(16). <https://doi.org/10.3390/nu14163285>
- Zgutka, K., Tkacz, M., Tomasiak, P., & Tarnowski, M. (2023). A Role for Advanced Glycation End Products in Molecular Ageing. *International Journal of Molecular Sciences*, 24(12). <https://doi.org/10.3390/ijms24129881>
- Zhang, J., Wang, H., Ai, C., Lu, R., Chen, L., Xiao, J., & Teng, H. (2023). Food matrix-flavonoid interactions and their effect on bioavailability. *Critical Reviews in Food Science and Nutrition*, 0(0), 1–22. <https://doi.org/10.1080/10408398.2023.2232880>

- Zhang, R., Zhang, Z., Zhang, H., Decker, E. A., & McClements, D. J. (2015). Influence of emulsifier type on gastrointestinal fate of oil-in-water emulsions containing anionic dietary fiber (pectin). *Food Hydrocolloids*, 45, 175–185. <https://doi.org/https://doi.org/10.1016/j.foodhyd.2014.11.020>
- Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine*, 4(11), 218. <https://doi.org/10.21037/atm.2016.03.37>
- Zhao, Q., Kong, D.-X., & Zhang, H.-Y. (2008). Excited-State p K a Values of Curcumin. *Natural Product Communications*, 3(2), 229–232. <https://doi.org/10.1177/1934578X0800300225>
- Zhu, Y. S., Yang, J. Q., Wang, N., Deng, Z. Q., Qing, Y., Wu, M., Cai, H., Liu, H., & He, G. S. (2022). Estimation of the bio-accessibility of methylmercury from aquatic foods using a PBTK model with an approximate Bayesian computation method in Chinese pregnant women. *Food and Chemical Toxicology*, 168(May), 113372. <https://doi.org/10.1016/j.fct.2022.113372>
- Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476), 1418–1429. <https://doi.org/10.1198/016214506000000735>
- Zou, L., Liu, W., Liu, C., Xiao, H., & McClements, D. J. (2015). Designing excipient emulsions to increase nutraceutical bioavailability: Emulsifier type influences curcumin stability and bioaccessibility by altering gastrointestinal fate. *Food and Function*, 6(8), 2475–2486. <https://doi.org/10.1039/c5fo00606f>

Appendix A Additional results

A.1 Validation of the spectrophotometric method to quantify total curcuminoids content

A.1.1 HPLC-DAD

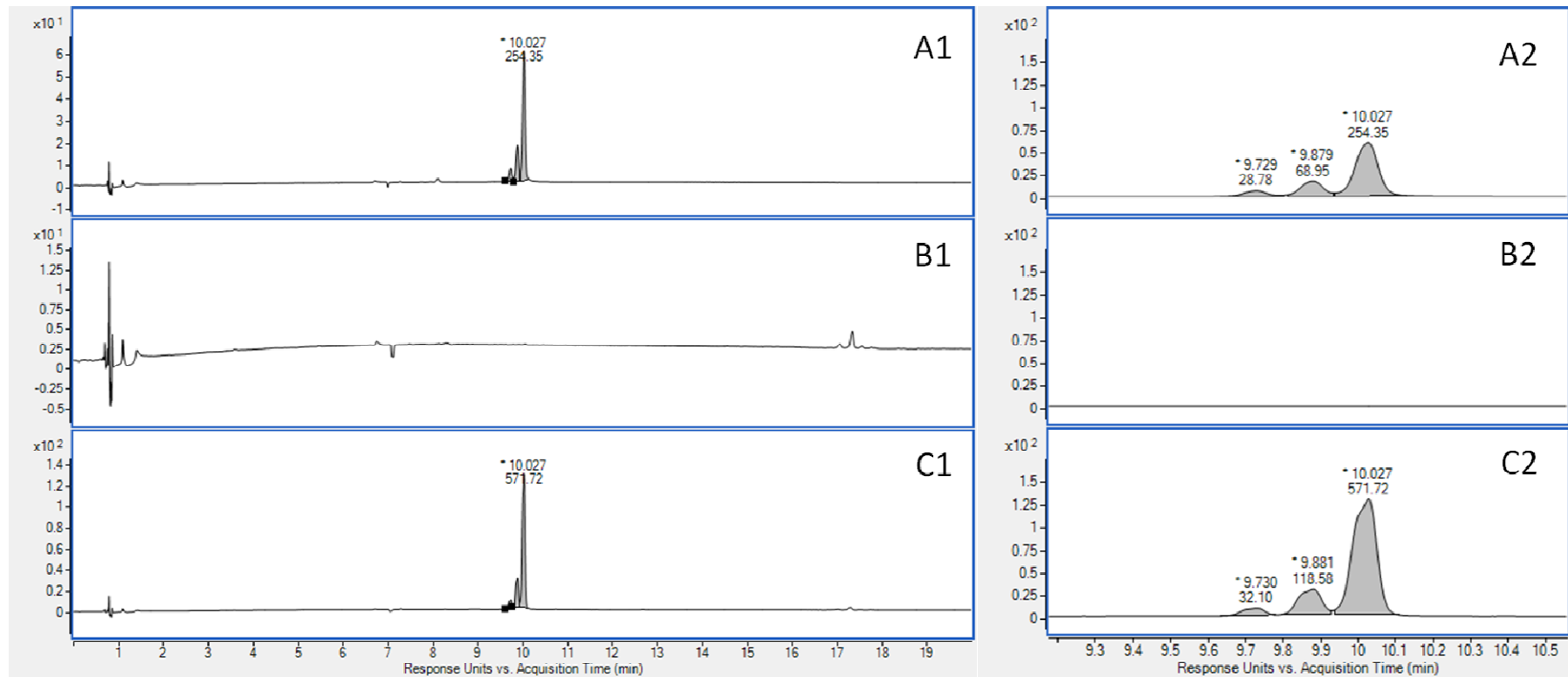


Figure S35. HPLC separation with UV 430 nm detection, full separation (1) and zoom (2) for (A) curcuma standard 4 mg/L, (B) Biscuit without curcumin and (C) Biscuit with curcumin after digestion

A.1.2 HPLC-HRMS

Table S17. Summary of 3 main curcuminoids from turmeric curcumin

Molecule	Raw Formula	Mass Weight (g/mol)	[M-H]- (uma)	Retention time (min)
Bis-Demethoxy curcumin (BDMC)	C ₁₉ H ₁₆ O ₄	308.10	307.0975	9.80
Demethoxy curcumin (DMC)	C ₂₀ H ₁₈ O ₅	338.12	337.1081	9.93
Curcumin (CUR)	C ₂₁ H ₂₀ O ₆	368.13	367.1187	10.09

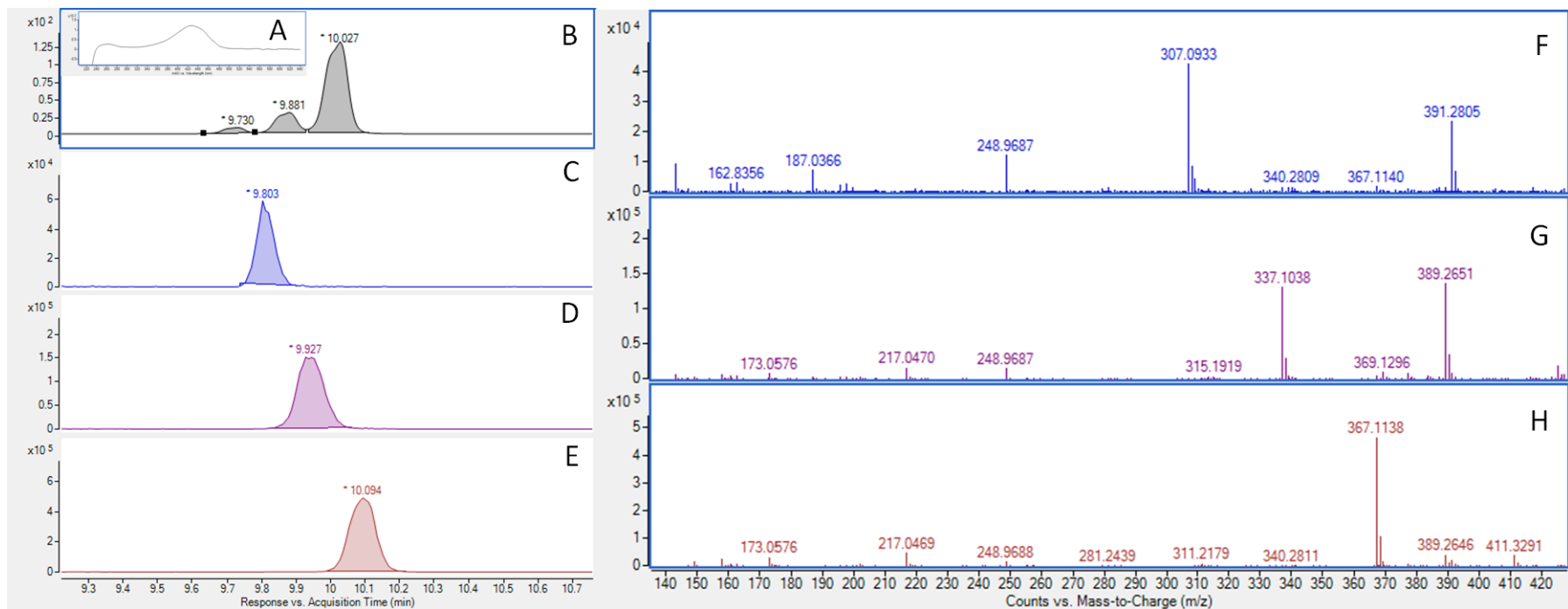


Figure S36. UV430 nm spectrum for CUR (a), UV 430 nm chromatogram (B) and extracted ion Chromatogram at 307.0975, 337.1091, 367.1187 uma for respectively BDMC (C), DMC (D) and CUR (E). LC-ESI(-)-HRMS for BDMC (F), DMC (G) and CUR (H).

A.2 Curcuminoid binding capacity

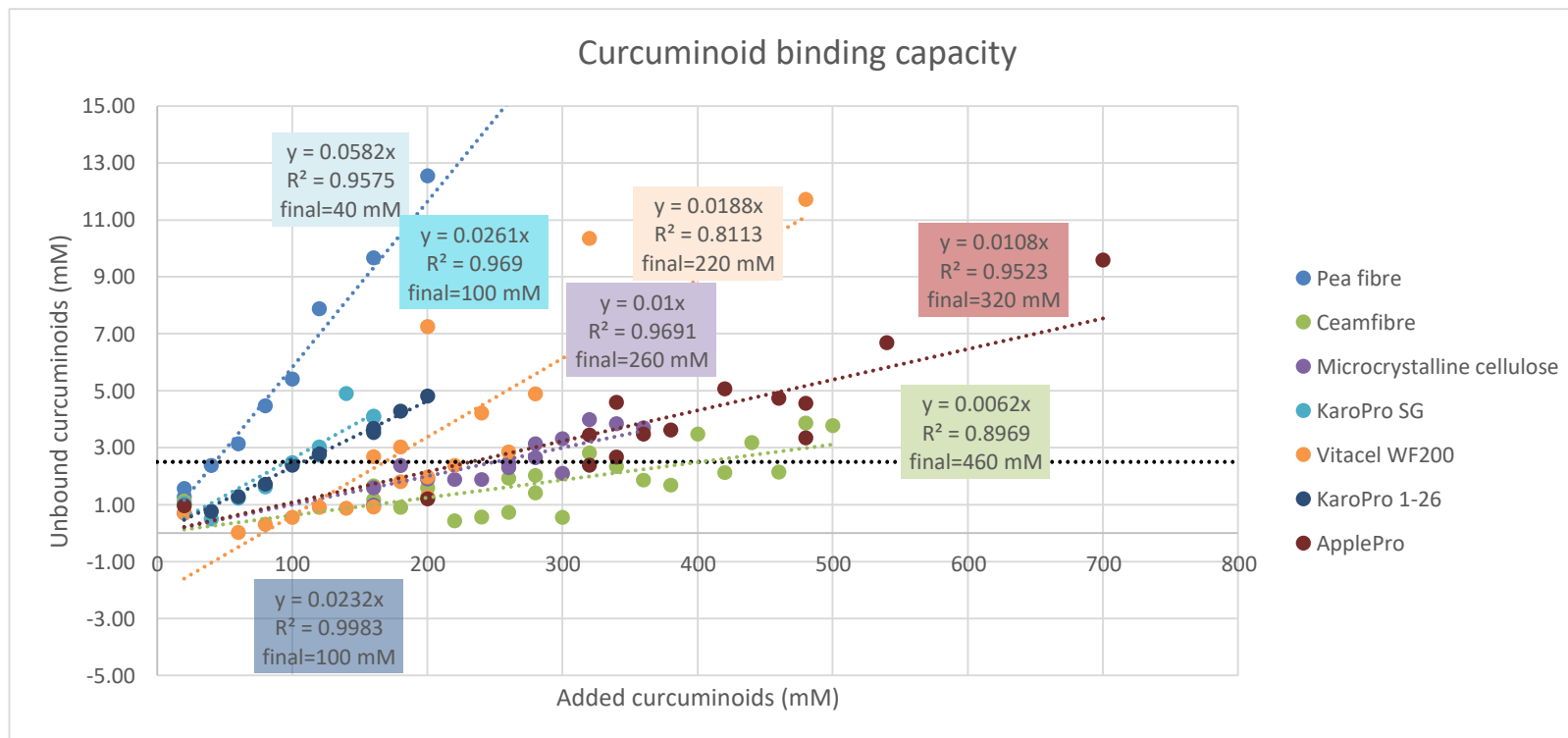


Figure S37. Measured curcuminoid binding capacities. Measurements belong to experiments performed on separate occasions. The curcuminoid binding capacity attributed to each fibre was proportional to the highest added curcuminoids concentration under the 2.5 mM threshold (intended as a relative measure).

A.3 Formulation properties correlation to bioaccessibility

Table S18. Explanatory variables correlation to bioaccessibility and algorithms supporting their inclusion in model. Spearman rho (ρ) coefficients were used as correlation metric. These have been computed grouped by matrix type and for all formulations simultaneously. Properties selected for modelling in one matrix type have been indicated by numbers in parentheses next to their coefficient (1: selected by autofeat, 2: selected by Boruta, 3: final selection by Laplacian priors). The specific fibre features (cellulose, lignin) were not considered in the selection process due to their sparsity and reduction in applicability and generalization capacity of the model. Macronutrient content was not considered during feature selection as it was chosen as the pooled feature during hierarchical modelling. n.a.: non-applicable. *: p-value between 0.01 and 0.05 **: p-value under 0.01

		<i>All formulations</i>	<i>Custards</i>	<i>Biscuits</i>
Supplement properties	<i>Oil holding capacity</i>	<i>n.a.</i>	<i>−0.39** (2)</i>	<i>0.10 (1)</i>
	<i>Water holding capacity</i>		<i>0.14</i>	<i>0.47**</i>
	<i>Bulk density</i>		<i>0.39**</i>	<i>−0.12</i>
	<i>Tapped density</i>		<i>0.19</i>	<i>0.13</i>
	<i>Curcumin binding capacity</i>		<i>−0.19</i>	<i>0.37 (2)</i>
	<i>Particle size</i>		<i>0.14</i>	<i>−0.31*</i>
	<i>Specific surface area</i>		<i>0.22</i>	<i>0.28</i>
Matrix properties	<i>Biscuit hardness</i>		<i>−0.48**</i>	<i>−0.17</i>
	<i>Custard viscosity</i>		<i>0.05</i>	<i>0.53**</i>
	<i>Custard firmness</i>		<i>0.06 (1)</i>	<i>0.51**</i>
	<i>Custard stickiness</i>		<i>−0.07</i>	<i>−0.53**</i>
	<i>Cellulose</i>		<i>−0.41** (1)</i>	<i>0.15</i>
	<i>Hemicellulose</i>		<i>0.12</i>	<i>0.66** (1,2,3)</i>
	<i>Dextrine</i>		<i>0.36* (1)</i>	<i>−0.13</i>
	<i>Pectin</i>		<i>0.10</i>	<i>0.45** (1)</i>

	<i>Lignin</i>		<i>0.10</i>	<i>0.45**</i>
	<i>Fructans</i>		<i>0.06*</i>	<i>−0.17</i>
	<i>Insoluble fibre</i>		<i>−0.31</i>	<i>0.11 (1)</i>
	<i>Soluble fibre</i>		<i>0.36** (1,3)</i>	<i>0.09 (1,3)</i>
	<i>Protein</i>	<i>0.87**</i>	<i>0.26*</i>	<i>0.61**</i>
	<i>Carbohydrate</i>	<i>0.87**</i>	<i>0.37**</i>	<i>0.55**</i>
	<i>Fat</i>	<i>0.77**</i>	<i>−0.08</i>	<i>0.22</i>
	<i>Water</i>	<i>−0.65**</i>	<i>0.01</i>	<i>0.70**</i>
	<i>Ash</i>	<i>0.74**</i>	<i>0.22</i>	<i>0.69**</i>
<i>Feature-engineered matrix properties</i>	<i>Macronutrient</i>	<i>0.86**</i>	<i>0.34 (1,2,3)</i>	<i>0.51** (1,3)</i>