Modeling Postprandial Glycemic Response in Non-Diabetic Adults Using XGBRegressor

Philipp D. Dubach*

June 6, 2025

Abstract

Personalized nutrition leveraging continuous glucose monitoring (CGM) data holds significant promise for optimizing postprandial glycemic responses, yet accessible predictive models remain limited by proprietary datasets. We applied an interpretable machine learning framework using XGBoost regression to predict postprandial glucose response parameters from meal composition and individual characteristics. Using the Hall dataset comprising 112 meals from 19 non-diabetic adults, we parameterized glucose curves using normalized Gaussian functions, extracting amplitude, time-to-peak, and curve width features. The XGBoost model achieved limited predictive performance with R^2 values of 0.46, -0.76, and 0.10 for amplitude, time-to-peak, and curve width parameters, respectively. Results indicate significant challenges in predicting postprandial glucose dynamics from basic meal composition and individual characteristics, highlighting the complexity of glucose metabolism and the need for more comprehensive feature sets in personalized nutrition applications.

Introduction

Dietary intake represents a fundamental determinant of blood glucose dynamics. Personalized nutrition has emerged as a promising approach to optimize postprandial glycemic responses, recognizing the significant impact of food intake on blood glucose levels. The increasing accessibility of machine learning models have revolutionized the field of personalized nutrition, enabling the development of sophisticated predictive models that account for an individual's unique characteristics. These models can accurately forecast glycemic responses to various foods, offering unprecedented precision in nutritional guidance. During the past decade, the integration of continuous glucose monitors (CGMs) has dramatically enhanced the application of machine learning in personalized nutrition for both diabetic and non-diabetic populations [1]. CGMs provide a comprehensive and precise characterization of individual glycemic responses by measuring interstitial glucose at frequent intervals, typically ranging from 1 to 15 minutes. Furthermore, 24-hour CGM profiles offer a more nuanced glycemic assessment compared to traditional methods such as self-monitored blood glucose via fingerstick or oral glucose tolerance tests. In studies focused on predicting meal-induced glycemic responses, researchers commonly employ the area under the curve (AUC) as a metric, calculated as the integral of the 90- or 120-minute glycemic response [2][3]. Advanced machine learning algorithms, such as gradient boosting regressors, have been developed to incorporate various characteristics, including the composition of macronutrients in meals (carbohydrates, fats, and proteins) and individual characteristics such as the gut microbiome, genetics, body mass index (BMI) and age, to better predict AUC [4][3]. These studies consistently demonstrate that carbohydrate content is the main determinant of AUC, while proteins and fats can synergistically modulate the glycemic response [5]. This paper aims to provide a comprehensive approach to open-source machine learning applications for CGM data analysis and prediction, addressing the current landscape dominated by proprietary datasets and models. We present a streamlined approach for training, testing, predicting, and optimizing postprandial glycemic responses. The research is focused on the predictability, feasibility, and explainability of biological systems through machine learning, dietary interventions and glycemic control. By leveraging open-source tools and methodologies, we seek to enhance the accessibility and transparency of CGM data analysis.

Glucose Homeostasis Mechanism

Glucose homeostasis represents one of the body's most critical regulatory processes. By maintaining blood sugar levels within narrow ranges, ensuring proper cellular function and energy metabolism. Understanding underlying mechanisms provides context for analyzing how dietary interventions and physiological factors influence glycemic control.

***Disclaimer:** This is a working paper, and hence it represents research in progress. This paper is not affiliated with any institution and represents the opinion of the author. This research is for educational and informational purposes only and does not constitute medical advice. The author assumes no responsibility for any decisions based on this work. **Contact:** info@philippdubach.com

Glucose Metabolism and Absorption

When food is consumed, digestive enzymes break down carbohydrates into glucose, the body's primary energy source. This glucose is absorbed through the small intestine via "transport" proteins and enters the bloodstream [6]. The liver acts as a central regulator, storing excess glucose as glycogen when abundant and releasing it when blood sugar levels drop [7].

Postprandial Glycemic Response

Postprandial glycemic response describes how blood glucose levels change after eating. In healthy individuals, blood glucose typically rises from baseline (70-100 mg/dL) within 30 minutes of eating, peaks around 60-90 minutes (remaining below 140 mg/dL), and returns to baseline within 2-3 hours [8]. Figure 1 shows how glucose rise triggers the pancreas to release insulin, a hormone that signals cells throughout the body to absorb glucose from the bloodstream, thereby lowering blood sugar levels back to normal [9].

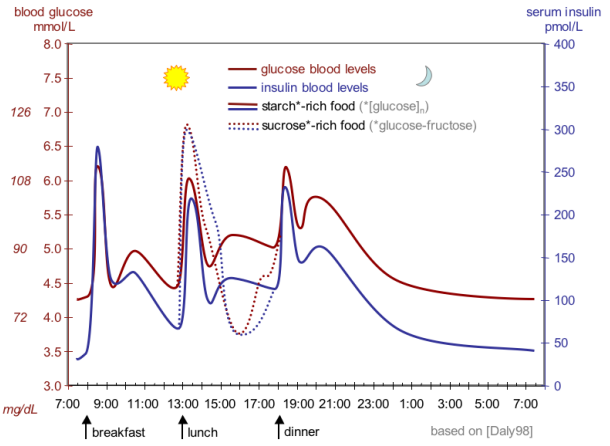

Figure 1: *The fluctuation of blood sugar (red) and the sugar-lowering hormone insulin (blue) in humans during the course of a day with three meals [10] (Attributed under CC 3.0).*

Individual Variation in Glycemic Response

Postprandial glucose responses vary significantly between individuals due to multiple factors. Food characteristics such as carbohydrate type, fiber content, and macronutrient composition influence how quickly glucose enters the bloodstream [11]. Personal factors including age, body weight, physical fitness, and genetic differences in metabolism also affect how efficiently the body processes glucose after meals [4].

Data and Feature Construction

For developing a fully personalized glycemic prediction model, ideally 500-1000 meal-response pairs would be required from CGM data of a single person, with detailed logging of all meal components and nutritional information. While we initiated this approach through self-monitoring, the limited number of data points even if collected over a multi-month period precluded robust model development. Therefore, we utilized the Hall dataset [12] as a proof-of-concept to demonstrate the feasibility of postprandial glucose prediction.

Dataset Description and Preprocessing

The Hall dataset [12] contains continuous glucose monitoring data from 57 non-diabetic adults undergoing standardized meal challenges. The final processed dataset encompassed 112 standardized meals from 19 subjects with the following key variables:

- **Meal Type**: Standardized meal identifiers
- **UserID**: Unique participant identifier
- **GlucoseValue**: Continuous glucose concentration (mg/dL)
- **Time**: Measurement timestamps at 5-minute intervals

CGM data were segmented into 2.5-hour windows centered on each meal event, comprising 30 minutes pre-meal and 2 hours post-meal, yielding 30 consecutive glucose measurements per episode. Baseline glucose levels were computed as the mean concentration during the 10-minute period preceding meal onset. Figure 2 illustrates the complete data processing workflow from raw CGM signals to structured meal-response episodes.

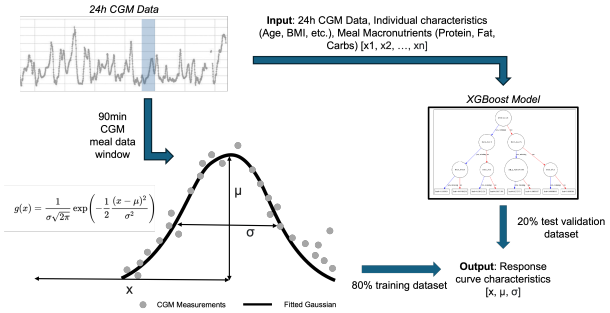

Figure 2: *Workflow of continuous glucose monitoring data segmentation, meal identification, and glycemic response modeling.*

Postprandial Curve Parameterization

To extract interpretable features from temporal glucose dynamics, we approximated each postprandial

response using a normalized Gaussian function:

$$G(t) = A \cdot \frac{\exp\left(-\frac{(t-\delta)^2}{2\sigma^2}\right)}{\max\left(\exp\left(-\frac{(t-\delta)^2}{2\sigma^2}\right)\right)} + b$$

where A represents the postprandial amplitude (mg/dL), δ the time-to-peak (minutes), σ the curve width parameter (minutes), and b the pre-meal baseline glucose level. Parameter optimization employed constrained nonlinear least squares with initial estimates derived from observed data: $A_{\text{init}} = \max(G_{\text{obs}}) - b$, $\delta_{\text{init}} = \arg\max(G_{\text{obs}})$, and $\sigma_{\text{init}} = 20$ minutes. Figure 3 shows the empirical distributions

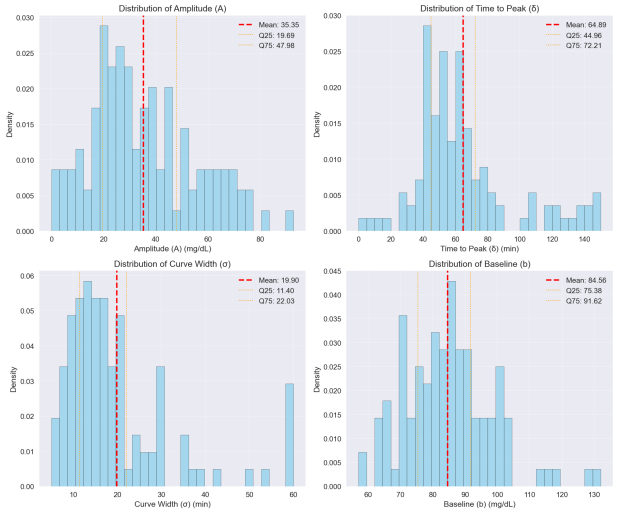

Figure 3: *Distribution of Gaussian curve parameters from postprandial glucose responses (n=112 meals). Red dashed lines show means with quartile ranges. Parameters: (A) amplitude (mg/dL), (δ) time-to-peak (min), (σ) curve width (min), (b) baseline glucose (mg/dL).*

of the extracted parameters across all meal responses, revealing substantial inter-individual variability in postprandial glucose dynamics. Figure 4 shows selected curve fitting results across different individuals, illustrating substantial inter-personal variation in glycemic response patterns.

Prediction Methodology

XGBoost Model Architecture

We implemented an XGBoost regressor to predict postprandial glucose response parameters from participant and meal characteristics. The Gaussian curve parameters (A, δ, σ) served as target variables, representing the amplitude, temporal dynamics, and curve width of glycemic excursions, respectively.

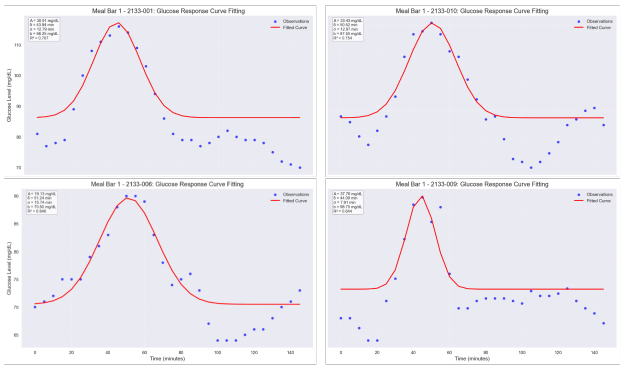

Figure 4: *Postprandial glucose curve fitting examples from Hall dataset. Blue dots: observed CGM data; red lines: fitted curves. Parameters shown: amplitude (A), time-to-peak (δ), curve width (σ), and R².*

Feature Engineering

The feature engineering process created a comprehensive dataset with 27 total features from participant biometrics and meal composition data:

- **Participant features**: Age, BMI, baseline glucose level
- **Meal features**: Protein content (g), fat content (g), carbohydrate content (g), total calories
- **Derived features**: Macronutrient ratios, caloric density, interaction terms

Feature scaling was applied using standardization to ensure uniform contribution across different measurement scales.

Model Training and Evaluation

The dataset was partitioned using stratified sampling to maintain representative distributions across training (80%) and testing (20%) subsets. Hyperparameter optimization was employed through cross-validation on the training set, optimizing learning rate, tree depth, regularization parameters, and number of estimators. Model performance was evaluated using regression metrics (RMSE, MAE, R²) A multi-linear regression model used as a baseline comparison, focusing on amplitude prediction using key macronutrient features and interaction terms.

Results

Results presented are preliminary and unverified. Findings should be interpreted with caution and are subject to revision pending further analysis.

Dataset Characteristics

The final processed dataset comprised 112 standardized meals from 19 subjects, with 27 engineered fea-

tures capturing meal composition, participant characteristics, and derived variables. This represents a subset of the original Hall dataset after preprocessing, mostly reducing the dataset to non-diabetic participants and removing incomplete sets.

Model Performance Metrics

The XGBoost regressor demonstrated highly variable predictive performance across the three Gaussian parameters characterizing postprandial glucose responses. Table 1 summarizes the model performance metrics for predicting amplitude (A), time-to-peak (δ), and curve width parameter (σ).

Param	RMSE	MAE	R^2	Corr	p-value
(A)	15.68	12.13	0.46	0.73	$<$0.001
(δ)	28.62	23.39	$-$0.76	$-$0.03	0.896
(σ)	13.53	9.91	0.10	0.49	0.018

Table 1: *XGBoost model performance metrics for predicting postprandial glucose response parameters. RMSE and MAE: mg/dL for amplitude, min for time parameters.*

The model achieved moderate accuracy for amplitude prediction (R^2 = 0.46, correlation = 0.73, $p < 0.001$), indicating a statistically significant but limited ability to predict glucose excursion magnitude. Time-to-peak prediction performed poorly (R^2 = -0.76, correlation = -0.03, p = 0.896), suggesting temporal dynamics are largely unpredictable from the available features. Curve width prediction showed weak but statistically significant performance (R^2 = 0.10, correlation = 0.49, p = 0.018).

Baseline Model Comparison

The multi-linear regression model focused on amplitude prediction achieved an R^2 of 0.24 with a residual standard error of 17.78 mg/dL. The model identified significant coefficients for carbohydrate content (CHO = 1.39), fat content (FAT = -3.27), and interaction terms (CHO$\times$PRO = -3.02, PRO$\times$FAT = -3.91), with an intercept of 35.35 mg/dL.

Model	R^2	RMSE (mg/dL)	Correlation
M-Lin Reg	0.24	17.78	N/A
XGBoost	0.46	15.68	0.73

Table 2: *Performance comparison between XGBoost and multi-linear regression for amplitude prediction.*

XGBoost demonstrated superior performance over linear regression, nearly doubling the explained variance (R^2 = 0.46 vs 0.24) and reducing prediction error (RMSE = 15.68 vs 17.78 mg/dL).

Statistical Significance and Model Validity

Statistical significance testing confirmed that amplitude predictions showed significant correlation with observed values ($p < 0.001$), while time-to-peak predictions were not significantly different from random (p = 0.896). Curve width predictions achieved marginal significance (p = 0.018) but with limited practical utility given the low R^2 value.

Feature Importance Analysis

To understand which features most strongly influence postprandial glucose amplitude predictions, we analyzed feature importance using SHAP (SHapley Additive exPlanations) values. Figure 5 displays both the feature importance rankings and the distribution of feature effects across all predictions.

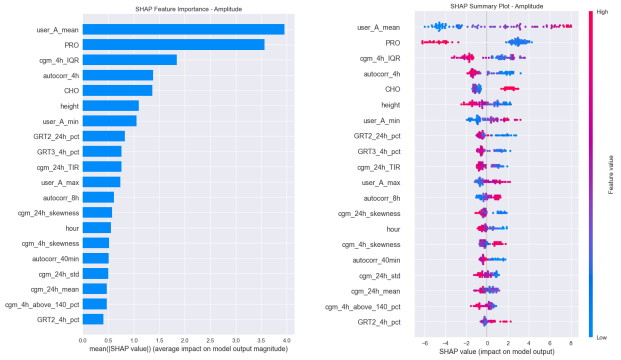

Figure 5: *SHAP analysis for postprandial glucose amplitude prediction. (a) Feature importance ranked by mean absolute SHAP values. (b) Summary plot showing SHAP value distributions with color indicating feature magnitude (red = high, blue = low).*

The analysis reveals that the general mean amplitude (user_A_mean) serves as the largest predictor, which is intuitive as it represents each individual's typical glucose response magnitude and acts more as a technical baseline adjustment. Following this baseline factor, meal-specific characteristics become the primary drivers, with protein content (PRO) and carbohydrate-related metrics emerging as the most influential meal composition variables.

Discussion

Limited Predictive Performance

The results reveal significant challenges in predicting postprandial glucose responses from a limited dataset. While amplitude prediction achieved moderate success (R^2 = 0.46), the predictive performance falls short of clinically useful thresholds typically requiring $R^2 > 0.7$ for reliable dietary guidance. The

strong positive correlation for amplitude ($r = 0.73$, $p < 0.001$) despite moderate R^2 suggests that while the model captures the general relationship between meal composition and glucose excursion magnitude, substantial unexplained variance remains.

Feature Importance Insights

The multi-linear regression coefficients provide interpretable insights, with carbohydrate content showing the expected positive association (coefficient = 1.39) and fat content showing a negative association (coefficient = -3.27), consistent with fat's known effects on delaying glucose absorption.

Methodological Limitations and Data Constraints

Several methodological limitations constrain the interpretation and generalizability of these results. First, the reduced dataset size (112 meals from 19 subjects) is substantially smaller than typical machine learning applications, potentially limiting statistical power and model generalizability. The small sample size may be particularly problematic for XGBoost, which typically requires larger datasets to achieve optimal performance. Second, the Gaussian approximation may inadequately capture complex postprandial dynamics, particularly for meals with mixed macronutrient profiles or individuals with atypical glucose responses. The negative R^2 for time-to-peak prediction suggests that the parametric approach may oversimplify glucose kinetics.

Literature Comparison

These results contrast with previously reported success in personalized glucose prediction [4]. However, key differences in methodology may explain the discrepancy. Previous studies often utilized much larger datasets (>1000 participants), included microbiome and extensive genetic data. The moderate success in amplitude prediction (analogous to peak glucose response) is more consistent with simpler glycemic index studies, which typically achieve correlations of 0.6-0.8 between carbohydrate content and peak glucose response [11]. However, even these correlations exceed our observed performance, suggesting fundamental limitations in the underlying data.

Areas of Improvement

The limited predictive performance highlights areas of improvement. First, individual-specific model training may be necessary to capture personalized metabolic responses. This approach would require extended monitoring periods (weeks to months) for each individual but could achieve the high accuracy needed for clinical applications. Secondly, alternative modeling approaches such as deep learning architectures designed for time series prediction, or mechanistic models incorporating glucose-insulin dynamics, may better capture the complexity of postprandial metabolism. Finally, larger, more diverse datasets with comprehensive feature sets are essential to establish the true potential and limitations of machine learning approaches for personalized glucose prediction.

Conclusion

Key Findings Summary

The multi-linear regression analysis revealed biologically consistent relationships, with carbohydrate content showing positive associations and fat content showing negative associations with glucose amplitude. However, even these interpretable models achieved limited predictive accuracy ($R^2 = 0.24$), suggesting that meal composition alone is insufficient for reliable glucose prediction. The results underscore the difficulty of population-level modeling for personalized nutrition applications. Despite using an advanced machine learning approach with comprehensive feature engineering on 27 variables, the majority of postprandial glucose variability remains unexplained by readily available meal and participant characteristics.

Data and Code Availability

The complete code, Jupyter notebooks, processed datasets, and supplementary results are available in the GitHub repository: `https://github.com/philippdubach/glucose-response-analysis`. The Hall dataset [12] used in this analysis is publicly available from the original publication.

Acknowledgments

This working paper represents an exercise in applying machine learning techniques to medical applications. The methodologies employed were largely inspired by Zeevi et al.'s [4] approach, alongside insights derived from the author's own CGM experiments.

References

[1] Raedeh Basiri and Lawrence J. Cheskin. "Personalized Nutrition Therapy without Weight Loss Counseling Produces Weight Loss in Individuals with Prediabetes Who Are Overweight/Obese: A Randomized Controlled Trial". In: *Nutrients* 16.14 (July 11, 2024), p. 2218. ISSN: 2072-6643. DOI: 10.3390/nu16142218. URL: https://www.mdpi.com/2072-6643/16/14/2218 (visited on 01/03/2025).

[2] Niteesh K. Choudhry et al. *PERSONALIZED GLYCEMIC RESPONSES TO FOOD AMONG INDIVIDUALS WITH TYPE 2 DIABETES IN INDIA: DEVELOPMENT OF A MACHINE LEARNING PREDICTION MODEL*. Oct. 22, 2024. DOI: 10.1101/2024.10.20.24315560. URL: http://medrxiv.org/lookup/doi/10.1101/2024.10.20.24315560 (visited on 01/03/2025).

[3] Rocio Guizar-Heredia et al. "A New Approach to Personalized Nutrition: Postprandial Glycemic Response and its Relationship to Gut Microbiota". In: *Archives of Medical Research* 54.3 (Apr. 2023), pp. 176–188. ISSN: 01884409. DOI: 10.1016/j.arcmed.2023.02.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S0188440923000358 (visited on 01/03/2025).

[4] David Zeevi et al. "Personalized Nutrition by Prediction of Glycemic Responses". In: *Cell* 163.5 (Nov. 2015), pp. 1079–1094. ISSN: 00928674. DOI: 10.1016/j.cell.2015.11.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867415014816 (visited on 01/03/2025).

[5] Berrak Basturk, Zeynep Koc Ozerson, and Aysun Yuksel. "Evaluation of the Effect of Macronutrients Combination on Blood Sugar Levels in Healthy Individuals". In: *Iranian Journal of Public Health* (Feb. 9, 2021). ISSN: 2251-6093, 2251-6085. DOI: 10.18502/ijph.v50i2.5340. URL: https://publish.kne-publishing.com/index.php/ijph/article/view/5340 (visited on 01/03/2025).

[6] Ernest M. Wright, Donald D. F. Loo, and Bruce A. Hirayama. "Biology of Human Sodium Glucose Transporters". In: *Physiological Reviews* 91.2 (Apr. 2011), pp. 733–794. ISSN: 0031-9333, 1522-1210. DOI: 10.1152/physrev.00055.2009. URL: https://www.physiology.org/doi/10.1152/physrev.00055.2009 (visited on 05/30/2025).

[7] R. P. Ferraris and J. Diamond. "Regulation of intestinal sugar transport". In: *Physiological Reviews* 77.1 (Jan. 1, 1997), pp. 257–302. ISSN: 0031-9333, 1522-1210. DOI: 10.1152/physrev.1997.77.1.257. URL: https://www.physiology.org/doi/10.1152/physrev.1997.77.1.257 (visited on 05/30/2025).

[8] James H. O'Keefe and David S.H. Bell. "Postprandial Hyperglycemia/Hyperlipidemia (Postprandial Dysmetabolism) Is a Cardiovascular Risk Factor". In: *The American Journal of Cardiology* 100.5 (Sept. 2007), pp. 899–904. ISSN: 00029149. DOI: 10.1016/j.amjcard.2007.03.107. URL: https://linkinghub.elsevier.com/retrieve/pii/S000291490701096X (visited on 05/30/2025).

[9] Stefano Del Prato and Antonio Tiengo. "The importance of first-phase insulin secretion: implications for the therapy of type 2 diabetes mellitus". In: *Diabetes/Metabolism Research and Reviews* 17.3 (May 2001), pp. 164–174. ISSN: 1520-7552, 1520-7560. DOI: 10.1002/dmrr.198. URL: https://onlinelibrary.wiley.com/doi/10.1002/dmrr.198 (visited on 05/30/2025).

[10] Me Daly et al. "Acute effects on insulin sensitivity and diurnal metabolic profiles of a high-sucrose compared with a high-starch diet". In: *The American Journal of Clinical Nutrition* 67.6 (June 1998), pp. 1186–1196. ISSN: 00029165. DOI: 10.1093/ajcn/67.6.1186. URL: https://linkinghub.elsevier.com/retrieve/pii/S0002916523183493 (visited on 05/30/2025).

[11] D J Jenkins et al. "Glycemic index of foods: a physiological basis for carbohydrate exchange". In: *The American Journal of Clinical Nutrition* 34.3 (Mar. 1981), pp. 362–366. ISSN: 00029165. DOI: 10.1093/ajcn/34.3.362. URL: https://linkinghub.elsevier.com/retrieve/pii/S000291652342816X (visited on 05/30/2025).

[12] Heather Hall et al. "Glucotypes reveal new patterns of glucose dysregulation". In: *PLOS Biology* 16.7 (July 24, 2018). Ed. by Jason Locasale, e2005143. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.2005143. URL: https://dx.plos.org/10.1371/journal.pbio.2005143 (visited on 05/30/2025).