

Yanxin Liu

Facilitating the Predictions of Batch and Continuous Anaerobic Digestion Processes Performance with Statistical Tools

School of Water, Energy and Environment
Energy and Power

PhD
Academic Year: 2020 - 2024

Supervisor: Ying Jiang, Phil Longhurst, Weisi Guo

July 2024



School of Water, Energy and Environment
Energy and Power

PhD

Academic Year 2020 - 2024

Yanxin Liu

Facilitating the Predictions of Batch and Continuous Anaerobic Digestion Processes Performance with Statistical Tools

Supervisor: Ying Jiang, Phil Longhurst, Weisi Guo

July 2024

This thesis is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

© Cranfield University 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

Anaerobic digestion (AD) is a promising technology for waste management and renewable energy production. Determining the biomethane potential (BMP) of a material is crucial when considering it as feedstock for the digester. The practical BMP of a material and its degradation kinetics can be derived from the batch BMP test, typically taking at least 30 days, which can be onerous to the industrial operator. Many studies have attempted to predict BMP test results by building regressions between various feedstock physiochemical characteristics and BMP test result. However, these methods primarily predict the ultimate biogas yield of the BMP test and are unable to capture the reaction kinetics.

Part I of this study proposed a method to predict the BMP test result of a material, not only the ultimate biogas yield but also the degradation kinetics, which was achieved by discovering a model that describes the biogas production well and then inferring parameters of this model from feedstock's physiochemical characteristics. The machine learning (ML) model, Decision Tree, was adopted to predict the parameters of the time series model, first-order Autoregressive Model, from the characteristics including total solids, volatile solids, total volatile fatty acid, total ammonia nitrogen, chemical oxygen demand, alkalinity, elemental composition, pH, heavy metals, etc. The DT trained and tested by leave-one-out cross validation with 25 BMP test data had a mean absolute percentage error (MAPE) of 45.63% on the BMP test results, showing unsatisfactory prediction accuracy and unreliable feature importance analysis results.

To obtain sufficient data for ML model training and avoid the consistency issues with BMP test data collected from diverse sources, a data augmentation method employing response surface design was proposed. With sufficient data for model training, eXtreme Gradient Boosting (XGBoost) models with three important features determined via feature importance analysis could predict biogas production model parameters with R^2 values above 0.99 on the test set.

Despite the strong regression capabilities of ML models compared to simple statistical model, their explainability remains a challenge. The current popular

methods for ML model interpretation focus on feature importance analysis. In this study, Meijei G-functions were used to interpret the predictions of the XGBoost model mathematically to enhance the accessibility and transparency of the black-box ML model to domain experts as a tool for material realistic BMP prediction. The general predictability of the mathematical metamodel was tested using 13 BMP test data sourced from the literature, all within the applicability range of the trained ML model, resulting in a mean absolute error of 38.074 mL CH_4/g volatile solid added and MAPE of 15.424%.

Besides the BMP of a material, operational parameters of the digester are critical to the performance of a continuous AD process. The application of Anaerobic Digestion Model No.1 (ADM1) for continuous AD simulation, which assists in decision-making in the AD industry by predicting digester performance under various operational schemes, is often hindered by its calibration difficulties, especially with limited data.

Part II of this study presents a Bayesian inference-based framework to reliably calibrate ADM1 using only initial-stage digester data. A sequence of sensitivity analysis (SA) was applied to identify the most influential kinetic parameters and initial values to be calibrated. SA results revealed that steady-state biogas production was collaboratively influenced by the disintegration rate, hydrolysis rates, and initial concentrations of acetate degraders, cations, and anions. In contrast, Total Ammonia Nitrogen and pH of digestate were predominantly influenced by initial values of cations and anions. These findings challenge the common practice in ADM1 studies of only calibrating kinetic and stoichiometric parameters. Then, using biogas production and digestate data from less than two hydraulic retention times and informative prior distributions determined from domain knowledge, seven the most influential uncertain inputs were calibrated. The calibrated model predicted the steady state performance satisfactorily. The 95% credible intervals of the calibrated model encompassed 66.047% of the 10-day moving average trendline of the daily biogas flow data and all of the steady-state digestate pH and total chemical oxygen demand data.

Keywords:

BMP Test, Machine Learning Regression Models, Machine Learning Model Interpretation, ADM1, Process-based Model, Sensitivity Analysis, Bayesian Calibration

Acknowledgements

Countless individuals have supported me throughout the three years of my PhD study, making this work possible. While I may not be able to mention everyone, I would like to express my special appreciation to some of them.

First and foremost, my deepest gratitude goes to my primary supervisor, Dr. Ying Jiang, who gave me countless support both academically and emotionally, always made time for me and helped me cope with the various difficulties I faced during my PhD study. A genuine thanks to my supervisor Prof. Weisi Guo, who brought up numerous indigenous ideas to my research and always patiently explained statistical concepts for me. Also, my heartfelt thanks to my supervisor Prof. Phil Longhurst. He guided me to start the journey of my PhD research.

A special thanks to Dr. Imma Bortone, who is more than an employer to me but also showed me the conscientious attitude towards a scientific work and gave me more supports than I could expect. Also a special thanks to Dr. Yadira Bajon Fernandez. Without her generous provision of the experimental equipment, I could hardly finish my lab work on time.

I am grateful to my lab colleagues, Nasreen and Esther, for their assistance with lab work and for sharing their experiences with me.

Thank you to my friend Qi, from whom I learned self-love, confidence, and generosity. To my friend Utao, a heartwarming presence on the other side of the earth, thank you for sparking my interest in coding. I am also grateful to Shaun, whose presence in my life I truly appreciate.

Finally, I hope my dad is doing well. I owe my deepest appreciation to my mum. I could never have made this far without your unwavering support thorough my growth. I will conquer more battles in my life with the assuredness of being loved.

Table of Contents

Academic integrity declaration i
Abstract ii
Acknowledgements v
List of Figures ix
List of Tables xiii
List of Equations xvi
List of Abbreviations xviii
1 Introduction 1
1.1 Background and Overview 1
1.1.1 BMP Test Result Prediction with Statistical Models 3
1.1.2 Continuous AD Performance Prediction 5
1.2 Aim and Objectives and Thesis Structure 7
1.2.1 Aim and Objectives 7
1.2.2 Thesis Structure 7
2 Literature Review 9
2.1 AD Overview 9
2.1.1 AD Process 9
2.1.2 Key Parameters in AD Process 10
2.2 BMP of a Material and Batch AD Test 13
2.2.1 Material BMP Estimation 13
2.2.2 BMP Test 21
2.2.3 BMP Test Results Prediction 29
2.2.4 General Applicability and Interpretation of ML Models for BMP Test Result Prediction 31
2.3 Continuous AD Performance Prediction 36
2.3.1 Statistical Model 38
2.3.2 Anaerobic Digestion Model No. 1 (ADM1) 40
2.4 Conclusions and Research Gap 45
PART I Biogas Production Prediction of Batch AD Test 49
3 BMP Test Result Prediction from Substrate Physiochemical Characteristics 50
3.1 Dataset and Methods 51
3.1.1 Materials and Dataset 51
3.1.2 Models for Biogas Production 52
3.1.3 Decision Tree (DT) 57
3.1.4 Gaussian Process (GP) 58
3.2 Results 59
3.2.1 BMP Test Results and Fitting 59
3.2.2 DT Training 65
3.2.3 Prediction Uncertainty Provided by GP 70

3.3 Summary and Discussion 72
4 Overcome Data Limitation by Generating Additional Dataset via Response Surface Design 74
4.1 Dataset and Methods 76
4.1.1 Dataset Generation 76
4.1.2 XGBoost Model Training 84
4.1.3 Feature Importance Analysis by SHAP Values 87
4.2 Results 88
4.2.1 BMP Test Results and Biogas Production Model Fitting 88
4.2.2 Response Surfaces Fitting 98
4.2.3 Additional Dataset Generation 109
4.2.4 XGBoost Training and Feature Importance Analysis 110
4.3 Summary and Discussion 118
5 Mathematical Interpretation of the XGBoost Model and the General Applicability of the Metamodel on External Data 120
5.1 Methods for Metamodelling 121
5.1.1 Meijei G-function Introduction 121
5.1.2 Metamodelling with Meijei G-function 123
5.1.3 Implementation Detail in Python 125
5.1.4 Metamodel Taylor Approximation 127
5.2 Data Source for Metamodel General Applicability Validation 128
5.3 Results 133
5.3.1 Metamodelling of XGBoost for k 133
5.3.2 Metamodelling of XGBoost for Y_m 137
5.3.3 General Applicability of Metamodel for Y_m Prediction 141
5.4 Summary and Discussion 144
5.4.1 General predictability of the ML model and its metamodel trained with augmented dataset 144
5.4.2 Interpret Black-box ML Model with Mathematical Expressions 145
PART II Performance Prediction of Continuous AD Process 147
6 Bayesian Calibration of ADM1 148
6.1 Methods 149
6.1.1 ADM1 149
6.1.2 Sensitivity Analysis 158
6.1.3 Bayesian Calibration 166
6.2 Results 170
6.2.1 Sensitivity Analysis 170
6.2.2 Bayesian Calibration of Influential θ_{ks} and θ_0 180
6.3 Summary and Discussion 196
7 Conclusions and Future Work 198
7.1 Conclusions 198

7.1.1 Batch AD Test Result Prediction from Feedstock Physiochemical Characteristics 199
7.1.2 Augmented Dataset for ML Regression Model Training 200
7.1.3 Mathematical Interpretation of the XGBoost Model 200
7.1.4 Bayesian Calibration of ADM1 201
7.2 Future Work 201
References 203
Appendices 238
Appendix A Ethical approval letter 238
Appendix B Chemical Analysis Methods 239
where: 239
Appendix C Data of BMP Test Result Prediction for Digestate Samples 246
Appendix D Peterson Matrix of ADM1 251

List of Figures

Figure 1-1 Underlying Mechanism for the prediction of the batch AD test biogas production in this research. 5

Figure 2-1 AD process flow chart..... 10

Figure 2-2 Components of Weender and van Soest classification of feedstock nutrients (Weinrich et al., 2018; LiebetrauI et al., 2017). (Adapted by permission from the copyright holders, Springer Nature) 15

Figure 2-3 Over-estimation of methane production regarding to the ratio between the methane production volume and headspace volume (V_S/V_H) and the methane content in the biogas (CH_4, %) (Strömberg et al., 2014). 27

Figure 2-4 Five main factors that influence BMP test results and the corresponding chapters where they are discussed. 29

Figure 2-5 Input – Output structure of the data for 'Traditional Features – Response' ML model and LSTM model training and testing. 39

Figure 3-1 The flowchat of the methodology for BMP test result prediction from feedstock physiochemical characteristics. 51

Figure 3-2 The fittings of three empirical biogas production models and the AR(1) model to the Type I (ADP 16 as an example) and Type II (ADP 18 as an example) biogas production data. 62

Figure 3-3 The ACF (a) and PACF (b) plots of sample ADP20 starting from day 4. The blue lines are 95% confidence bounds for significant correlation. .. 63

Figure 3-4 GP predictions guided by the prediction of the inferred AR (1) model with uncertainty information provided (a) RBF kernel. (b) RBF kernel + linear kernel. 71

Figure 4-1 Research framework of data augmentation via mixture design for training ML regression model for BMP result prediction. 75

Figure 4-2 Standard ASCD with 13 points. x_1, x_2 and x_3 represent fractions of three components in a mixture. And the 10 points selected in this study for carrying out BMP tests with assigned experimental group numbers. 79

Figure 4-3 SMPs of two MCC control. 88

Figure 4-4 BMP test biogas production data and fitted first-order kinetic model curves (Part 1) 90

Figure 4-5 BMP test biogas production data and fitted first-order kinetic model curves (Part 2) 91

Figure 4-6 BMP test biogas production data and fitted first-order kinetic model curves (Part 3) 92

Figure 4-7 BMP test biogas production data and fitted first-order kinetic model curves (Part 4) 93

Figure 4-8 BMP test biogas production data and fitted first-order kinetic model curves (Part 5) 94

Figure 4-9 Gauss-Markov assumption checks for the residuals of the prediction for parameter k. 100

Figure 4-10 Gauss-Markov assumption checks for the residuals of the prediction for parameter Y_m. 101

Figure 4-11 The fitted response surface for the parameter k with respect to the fractions of substrate (A) and substrate (O) in the co-digestion mixture. (Top: 3D plot; Bottom: 2D Contour plot). 103

Figure 4-12 The fitted response surface for the parameter Y_m with respect to the fractions of substrate (A) and substrate (O) in the co-digestion mixture. (Top: 3D plot; Bottom: 2D Contour plot). 104

Figure 4-13 Visual inspection of the results of the LOO cross validation for parameter k and Y_m. The APE between the parameter k/Y_m derived by fitting first-order kinetic model to biogas production data and the prediction of k/Ym of the left-out sample with the response surface equations fitted with nine samples is displayed in the parentheses. 107

Figure 4-14 The 'vicinity points (the samples with a distance to the sample being left out smaller than ½ assuming the length of the side of ASCD triangle is one)' to the sample being left out when implementing LOO cross validation 108

Figure 4-15 Pearson correlations (r) of 12 physiochemical characteristics (The $r \geq 0.98$ was highlighted) ('Hemi_Cellulose' represents the sum of the contents of cellulose and hemicellulose'). 111

Figure 4-16 Feature importance of the XGBoost model for k ('Hemi_Cellulose' represents the sum of the contents of cellulose and hemicellulose). 113

Figure 4-17 Dependence plot of SHAP values for feature 'Lignin' against the values of feature 'Carbohydrate' (a) and feature 'Lignin' against the values of feature 'Lipid' (b) for k prediction. 115

Figure 4-18 Feature importance of the XGBoost model for Y_m ('Hemi_Cellulose' represents the sum of the contents of cellulose and hemicellulose). 117

Figure 5-1 Research framework of mathematical metamodeling of XGBoost models. 121

Figure 5-2 The fitting of the selected basis G-function of the metamodel of the XGBoost model for k prediction on each dimension. 136

Figure 5-3 The fitting of the selected basis G-function of the metamodel of the XGBoost model for Y_m prediction on each dimension. 140

Figure 5-4 The BMP test results of 13 external test data and the BMP results predicted by the fifth-order Taylor approximated metamodel for Y_m prediction. 143

Figure 6-1 Flowchart of the ADM1 Bayesian calibration framework. 149

Figure 6-2 The μ_i^* of the distributions of the elementary effects of 24 uncertain θ_{ks} and θ_0 for QoI_1. The green bars are parameters influential to QoI_1. ... 172

Figure 6-3 The μ_i^* of the distributions of the elementary effects of 24 uncertain θ_{ks} and θ_0 for QoI_2. The green bars are parameters influential to QoI_2. ... 172

Figure 6-4 The μ_i^* of the distributions of the elementary effects of 24 uncertain θ_{ks} and θ_0 for QoI_3. The green bars are parameters influential to QoI_3. 173

Figure 6-5 The main effects of the 13 more influential θ_{ks} and θ_0 scaled with respect to the largest main effect in five trails with different sample sizes (n) for QoI_1 (y-axis is in logarithmic scale). 177

Figure 6-6 The main effects of the 13 more influential θ_{ks} and θ_0 scaled with respect to the largest main effect in five trails with different sample sizes (n) for QoI_2 (y-axis is in logarithmic scale). 178

Figure 6-7 The main effects of the 13 more influential θ_{ks} and θ_0 scaled with respect to the largest main effect in five trails with different sample sizes (n) for QoI_3 (y-axis is in logarithmic scale). 179

Figure 6-8 Marginal prior probability distributions for k_{dis}, $k_{hyd,ch}$, $k_{hyd,pr}$, $X_{su}{}^0$, $X_{ac}{}^0$, $S_{cat^+}{}^0$ and $S_{an^-}{}^0$ 182

Figure 6-9 The smoothed marginal posterior distributions for k_{dis}, $k_{hyd,ch}$, $k_{hyd,pr}$, $X_{su}{}^0$, $X_{ac}{}^0$, $S_{cat^+}{}^0$ and $S_{an^-}{}^0$ from two MCMC chains. SD_1, SD_2 and SD_3 is the marginal distribution of the estimated measurement errors for daily biogas flow (m^3/d), pH and TAN ($kmol\ N/m^3$). 186

Figure 6-10 Experimental observations and simulated results of calibrated model of daily biogas flow. Blue dots: data used for Bayesian calibration. Black dots: data for validation. Gray dotted line: 10-day moving age trendline of the daily biogas flow. Orange line and shaded area: the mean and 95% credible interval of 3000 model outputs for daily biogas flow with θ_{ks} *and* θ_0 sampled from trace of MCMC chains after burn-in samples. 188

Figure 6-11 Distribution of the residuals of the mean prediction for daily biogas flow on validation data (from day 76). 189

Figure 6-12 Experimental observations and simulated results of calibrated model of digestate TAN. Blue dots: data used for Bayesian calibration. Black dots: data for validation. Orange line and shaded area: the mean and 95% credible interval of 3000 model outputs for TAN with θ_{ks} and θ_0 sampled from trace of MCMC chains after burn-in samples. 190

Figure 6-13 Experimental observations and simulated results of calibrated model of digestate pH value. Blue dots: data used for Bayesian calibration. Black dots: data for validation. Orange line and shaded area: the mean and 95% credible interval of 3000 model outputs for pH with θ_{ks} and θ_0 sampled from trace of MCMC chains after burn-in samples. 191

Figure 6-14 Experimental observations and simulated results of calibrated model of digestate tCOD value. Blue dots: data used for Bayesian calibration. Black dots: data for validation. Orange line and shaded area: the mean and 95% credible interval of 3000 model outputs for tCOD with θ_{ks} andθ_{ks} sampled from trace of MCMC chains after burn-in samples..... 192

Figure 6-15 Experimental observations and simulated results of calibrated model of daily biogas flow. Blue dots: data used for Bayesian calibration. Black dots: data for validation. Gray dotted line: 10-day moving age trendline of the daily biogas flow. Orange line and shaded area: the mean and 95% credible interval of 3000 model outputs for daily biogas flow with θ_{ks} and θ_{ks} sampled from posterior distribution. Uniform distribution was assigned as prior distribution..... 195

List of Tables

Table 2-1 Regression models on the prediction of biogas or BMP for lignocellulose materials (DM: dry matter). 19

Table 3-1 Characteristics of AR(1) and MA(1) models (without considering the constant term). 55

Table 3-2 R^2values of the fittings of three empirical models to Type I and Type II BMP test biogas production time series. FO: first-order kinetic model. MG: modified Gompertz model. PP: pseudo-parallel first-order model. (The R^2 values larger than 97% are in bold). 60

Table 3-3 R^2 values of the fitting of AR(1) to the Type I and Type II BMP test biogas production time series (The R^2 values larger than 97% are in bold). 64

Table 3-4 The APEs for the predicted AR(1) model coefficient and constant and the APEs of the calculated biogas yield at day 28 (y(28)) when training a DT by LOO cross-validation. 66

Table 3-5 The overall predictor importance of 25 DTs for inferring AR(1) coefficient and constant from physiochemical characteristics. 69

Table 4-1 Characteristics of substrate (A) banana peel and green peas (w/w 1:1), (O) banana pulp and potato (w/w 1:1) and (B) 20% fat minced beef. 76

Table 4-2 Substrates and inoculum volumes for 10 experimental groups and the control group. 80

Table 4-3 Fitted k of first-order kinetic model, experimentally observed final biogas production (Y_m) and calculated BMP_{theo} 95

Table 4-4 The statistics of the fitting of a special cubic equation to the parameter k of the first-order kinetic model of 10 samples. 99

Table 4-5 The statistics of the fitting of a special cubic equation to the parameter Y_m of first-order kinetic model of 10 samples. 100

Table 4-6 The residuals and the APEs of the k and Y_m parameters predicted from the fitted response surface equations (Predicted) with respect to the parameters derived by fitting the first-order kinetic model (Fitted)..... 102

Table 4-7 Results of the LOO cross validation for parameter k and Y_m. Fitted: the parameter k / Y_m derived by fitting first-order kinetic model to biogas production data. Predicted: the prediction of k/Y_m of the left-out sample with the response surface equations fitted with nine samples. 105

Table 4-8 The R^2 and RMSE of 'XGBoost_xs' model and 'XGBoost_red' models for parameter k prediction on test set with different combinations of three features. 116

Table 4-9 The R^2 and RMSE of 'XGBoost_xs' model and 'XGBoost_red' models for parameter Y_m prediction on test set with different combinations of three features. 118

Table 5-1 The constrained metamodeling space G in the practical implementation. 126

Table 5-2 The 'applicability range' of the XGBoost models trained in Chapter 4 for the predictions of parameters k and Y_m of the first-order kinetic model. 128

Table 5-3 The BMP test result samples collected from three peer-reviewed papers that the chemical compositions fall within the 'applicability range' of the metamodels in this study with Y_m/BMP_{theo} ratio in the range 38.250 to 64.595%. Y_m: experimental BMP result. BMP_{theo}: Theoretical BMP calculated from Eq. 2-4. 131

Table 5-4 Optimised θi and minimum loss of the basis G-function for each dimension of the metamodel of the XGBoost model for k prediction. 134

Table 5-5 The coefficient of each basis function of the ridge regression and the expression of the approximated metamodel with fifth-order Taylor expansion of the XGBoost model for k prediction. 134

Table 5-6 Optimised θi and minimum loss of the basis G-function for each dimension of the metamodel of the XGBoost model for Y_m prediction. 138

Table 5-7 Coefficient of each basis function of the ridge regression and expression of the approximated metamodel with fifth-order Taylor expansion of the XGBoost model for Y_m prediction. 138

Table 5-8 The scaled feature values of 13 external data for testing the general applicability of the approximated metamodel with fifth-order Taylor expansion for Y_m prediction. Y_m: experimentally derived BMP test result. Y_m_pred: BMP test result predicted by the Taylor approximated metamodel. 141

Table 6-1 Characteristic of the Inoculum. 151

Table 6-2 Initial values of the state variables (θ_0) and how they were derived from the characteristics of the inoculum. The superscript '0' indicates initial values for state variables. 'SA' indicates the uncertain θ_0 to be included in SA process. 153

Table 6-3 Characteristics of the feedstock (Input for the Transformer Tool)... 155

Table 6-4 ADM1 Input Value (θ_f) derived by Transformer Tool..... 156

Table 6-5 . The calibrated values for kinetic and stoichiometric parameters k_{dis}, $k_{hyd,ch}$, $k_{hyd,pr}$, $k_{hyd,li}$, $k_{m,pro}$, $k_{m,ac}$, k_{m,H^2}, $k_{m,ac}$, $K_{S,pro}$,$K_{S,ac}$, K_{S,H^2}, K_{I,NH^3} in papers applying ADM1 to AD of raw and THP sludge in the past 15 years. '/' means the author did not calibrate corresponding parameters but keep them

at their default values as suggested in the benchmark model (Batstone et al., 2002). 161

Table 6-6 The initial values used for $S_{H_2}{}^0$, $S_{CH_4}{}^0$, $S_{cat^+}{}^0$ and $S_{an^-}{}^0$ of ADM1 in literatures which have reported the initial values used in the past 15 years. 162

Table 6-7 The Uniform Distributions of uncertain inputs for SA. 163

Table-6-8 The first-order Sobol' index (S_i) and total Sobol' index (S_{T_i}) of 13 more influential uncertain θ_{ks} and θ_0 for QoI_1, QoI_2 and QoI_3 respectively. 175

Table 6-9 The statistical summary of the marginal prior and posterior distributions of seven the most influential uncertain inputs of ADM1. 185

Table 6-10 Statistical summary of marginal posterior distributions for uncertain inputs when prior distributions are uniform distributions and the mean of the steady state simulation results for daily biogas flow, TAN, pH and tCOD with 3000 set of parameters drawn from the posterior distribution. 194

List of Equations

(2-1)..... 14
(2-2)..... 14
(2-3)..... 14
(2-4)..... 16
(2-5)..... 26
(2-6)..... 28
(2-7)..... 28
(2-8)..... 29
(2-9)..... 41
(2-10)..... 41
(2-11)..... 41
(2-12)..... 42
(2-13)..... 42
(3-1)..... 52
(3-2)..... 52
(3-3)..... 52
(3-4)..... 53
(3-5)..... 53
(3-6)..... 54
(3-7)..... 54
(3-8)..... 55
(3-9)..... 57
(3-10)..... 58
(3-11)..... 58
(4-1)..... 82
(4-2)..... 83
(4-3)..... 83
(4-4)..... 83

(4-5)..... 84
(4-6)..... 85
(4-7)..... 85
(4-8)..... 99
(4-9)..... 101
(5-1)..... 122
(5-2)..... 122
(5-3)..... 122
(5-4)..... 122
(5-5)..... 123
(5-6)..... 124
(6-1)..... 150
(6-2)..... 150
(6-3)..... 151
(6-4)..... 159
(6-5)..... 164
(6-6)..... 164
(6-7)..... 165
(6-8)..... 166
(6-9)..... 166
(6-10)..... 168
(6-11)..... 168
(6-12)..... 168
(6-13)..... 169

List of Abbreviations

AD	Anaerobic Digestion
ADM1	Anaerobic Digestion Model No.1
ACF	Autocorrelation Function
ADF	Acid Detergent Fibre
ADL	Acid Detergent Lignin
ANN	Artificial Neural Network
APE	Absolute Percentage Error
ARMA	Autoregression Moving Average
ASCD	Augmented Simplex Centroid Design
BMP	Biochemical Methane Potential
BMP_{theo}	Theoretical Biochemical Methane Potential
COD	Chemical Oxygen Demand
DT	Decision Tree
EE	Elementary Effect
FAST	Fourier Amplitude Sensitivity Test
GAN	Generative Adversarial Network
GC	Gas Chromatography
GHGE	Generalised Hypergeometric Equation
GP	Gaussian Process
GSA	Global Sensitivity Analysis
HRT	Hydraulic Retention Time
IA	Intermediate Alkalinity
IC	Inorganic Carbon
IN	Inorganic Nitrogen
JS	Jensen-Shannon
KL	Kullback-Leibler
KNN	K-Nearest Neighbour
LCFA	Long Chain Fatty Acid
LOO	Leave-one-out
LSA	Local Sensitivity Analysis
LSTM	Long-short Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MCC	Microcrystalline Cellulose
MCMC	Markov-chain Monte Carlo
ML	Machine Learning
MLR	Multiple Linear Regression

MSE	Mean Squared Error
NDF	Neutral Detergent Fibre
ODE	Ordinary Differential Equation
OLR	Organic Loading Rate
PA	Partial Alkalinity
PACF	Partial Autocorrelation Function
PCA	Principal Component Analysis
pCOD	Particulate Chemical Oxygen Demand
PDP	Partial Dependence Plot
PS	Primary Sludge
RBD	Random Balanced Disegn
RBF	Radial Basis Function
RMSE	Root Mean Square Error
SA	Sensitivity Analysis
sCOD	Soluble Chemical Oxygen Demand
SHAP	SHapley Additive exPlanations
SIR	Substrate to Inoculum Ratio
SMP	Specific Methane Production
TA	Total Alkalinity
TAN	Total Ammonia Nitrogen
tCOD	Total Chemical Oxygen Demand
THP	Thermal Hydrolysis Process
TKN	Total Kjeldahl Nitrogen
TS	Total Solids
VFA	Volatile Fatty Acid
VS	Volatile Solids
WAS	Waste Activated Sludge
WWTP	Wastewater Treatment Plant
XGBoost	eXtreme Gradient Boosting

1 Introduction

1.1 Background and Overview

Anaerobic Digestion (AD) is a process where the biogas is produced from the degrading activities of microorganisms on organic materials under anaerobic condition (Rocamora et al., 2020). The biogas typically contains 50-70% methane (CH_4), 30-50% carbon dioxide (CO_2) and other trace gases including nitrogen (N_2), hydrogen (H_2) and hydrogen sulphide (H_2S) (Angelidaki et al., 2018; Appels et al., 2011a). AD is one of the most promising technologies for waste management and renewable energy production (Kumar & Samadder, 2020). Compared to other waste management methods such as the incineration, landfill and composting, AD process usually releases less greenhouse gases and is more space efficient (Dastjerdi et al., 2021; Rezania et al., 2023). The number of AD plants in the EU continues to grow in the past decade from 6507 in 2009 to 18843 in 2021 (European Biogas Association, 2022). In the UK, there are in total 660 operational AD facilities by 2022 with a combined capacity of 7000 Nm^3/h (National Non-Food Crops Centre, 2022). The AD plants in the EU are producing 18.4 bcm of renewable gas, which is 4.5% of the natural gas consumption in the EU in 2021 (European Biogas Association, 2022). The wider implementation of AD reduces the dependency on natural gas and contributes to the net-zero goal of the energy sector. Except for the biogas, the digestate produced from the AD process has the potential to be applied as soil conditioner and fertiliser (Chiu et al., 2016). Over 200 Mt digestate (fresh matter) was produced in the EU in 2021 (European Biogas Association, 2022).

The feedstocks for AD can be supplied from a variety of organic waste streams, including typically: sewage sludge, landfill, organic fraction of municipal solid waste, industrial waste and agricultural waste. Depending on the feedstock availabilities, different materials can be treated together in the same AD facility in a ‘co-digestion’ process. Co-digestion can potentially improve biogas yield compared to mono-AD and less inhibition occurs because of a more balanced nutrient, the carbon (C) to nitrogen (N) ratio (C/N ratio), pH value and the organic load (Xu et al., 2018).

The primary consideration when using an organic material as feedstock for an anaerobic digester is the biochemical methane potential (BMP) of this material. The theoretical BMP of a feedstock can be estimated based on key material characteristics including chemical oxygen demand (COD), elemental composition or chemical composition (e.g. carbohydrate, protein, lipid and lignocellulose). However, these theoretical calculations are often overestimated compared to the BMP results derived experimentally. This overestimation is primarily due to two factors. First, some of the energy produced from organic material degradation is consumed by cell growth (assimilation) and cell maintenance. Second, the calculations often fail to account for the material's biodegradability. Materials high in lignocellulosic components can have significantly low biodegradability for two reasons: lignin is non-degradable under anaerobic conditions due to the absence of necessary enzymes, and the biogas potential and hydrolysis rate of cellulose and hemicellulose are reduced by their bonding with lignin, which acts as a barrier to efficient hydrolysis (Mussatto et al., 2008).

Currently the ‘gold standard’ for the determination of BMP value of an anaerobic digestion feedstock material is by carrying out a BMP test (Holliger et al., 2016). The BMP test is a batch test where the material is degraded by microorganisms contained in the inoculum under an anaerobic environment and a stable temperature. In BMP tests, the influence of factors including degradability of the material, presence of the possible inhibitory substances and the fact that part of the substrate is utilised by microbial respiration are reflected. Therefore, the BMP value derived from a BMP test is a more realistic estimation of the biogas potential of an organic material. Due to its accuracy and value in providing operational guidance for long-term AD operation, BMP tests are widely employed to study the influence of factors such as the blending ratio among substrates for co-digestion, chemical additives, the feedstock pre-treatment methods etc. Moreover, BMP test can reveal the degrading kinetics of a material, which influences the efficiency of an AD process.

Despite of the advantage of an BMP test, it can take 20 to 90 days for the biogas production of a material reaching the end under a specific environment (Postawa

et al., 2020). However, carrying out a BMP test for such a long time is not efficient and labour intensive for industrial-scale digester operators. Therefore, various methods are proposed for BMP test result prediction. These includes correlating aerobic test result with BMP test result (Ponsá et al., 2008), spectroscopy techniques such as near infrared spectroscopy (Doublet et al., 2013; Triolo et al., 2014) and Fourier transform mid-infrared photoacoustic spectroscopy (Bekiaris et al., 2015), and building a regression model between various physiochemical characteristics and BMP test result (Liu et al., 2023). Meanwhile, there are efforts to shorten the BMP test. This is often achieved by inferring the biogas production model parameters from the initial data obtained from a BMP test (Catenacci et al., 2022; Da Silva et al., 2018; Ponsá et al., 2011; Strömberg et al., 2015).

1.1.1 BMP Test Result Prediction with Statistical Models

The prediction of BMP test results using regression models has two primary applications. The first application is to determine the realistic BMP of lignocellulosic materials. Theoretically calculated BMP for such material is often significantly overestimated, so experimental determination is preferred. Regression models can then be developed using the contents of chemical components as predictors to predict the experimentally derived BMP results (Bayard et al., 2016; Cu et al., 2014; Rath et al., 2013; Thomsen et al., 2014; Vazifehkhoran & Triolo, 2019). The second scenario is to use regression models to investigate the effects of various factors on batch AD biogas production besides the biogas potential of the material itself. These factors include the blending ratio of co-digestion substrates, the concentration of potential inhibitory substances (Liu et al., 2023), chemical additives (Almomani, 2020a; Khashaba et al., 2022; Xu et al., 2021), C/N ratio (Yang et al., 2021), and the substrate to inoculum ratio (SIR) (Xu et al., 2014), etc.

There are typically three types of regression models used: linear regression models, higher-order regression models, and machine learning (ML) models. ML models are particularly effective for data from highly nonlinear processes (Wang et al., 2023). Meanwhile, the ML regression models developed for the second application scenario mentioned above can be coupled with optimising algorithms

such as genetic algorithm (Gueguim Kana et al. 2012), particle swarm optimisation (Alrawashdeh et al., 2022) and Bayesian optimisation to find the optimal combination of features for maximized biogas production.

Although ML models often outperform simple statistical models, more data are needed to train a ML model to convergence so that it learns the data pattern well (Rajula et al., 2020). To gather sufficient data for ML model training, many studies have collected BMP test result data from the literature. However, due to the great variation in BMP test results from different studies – primarily caused by different operational procedures and different test system adopted – the data collected from multiple studies may not reliably reflect the underlying relationship between various physiochemical characteristics and BMP test result.

Another state-of-the-art method to address the data availability problem involves using data augmentation algorithms such as generative adversarial network (GAN) (Adeleke et al., 2023) and mixup data augmentation (Geng et al., 2024).

The performance and generalisation ability of a trained ML model should be evaluated on a separate test dataset. However, the high prediction accuracy of the model on test set reported in some papers should have been interpreted critically. As ML models excel at memorising patterns in the training samples, high prediction accuracy on a test set left out from the same set of collected data does not necessarily translate to the same accuracy on data from other sources. A more reliable approach is to verify the general predictability of the trained ML model on external data (Triolo et al., 2011; Vazifehkhoran & Triolo, 2019; Wang et al., 2023).

The aim of this part of the study is to predict the batch AD test result of a substrate from its physicochemical characteristics by ML models for regression. Unlike previous models that only predicted the ultimate biogas yield of the BMP test, this study aims to predict both the biogas potential and the reaction kinetics information derived from the test. This is achieved by inferring the parameters of mathematical models that describe BMP test biogas production from substrate's characteristics, including elemental composition, chemical composition, and the concentrations of possible inhibitory substances. This prediction mechanism

originates from the observations that: biogas production of a batch AD test is primarily influenced by the physiochemical characteristics of the feedstock when the reaction environment is controlled, and the biogas production kinetics can be described by low-dimensional mathematical models. Therefore, a possible mapping from the high-dimensional physiochemical parameters space to the low-dimensional mathematical model parameters space exists. Various multivariate regression ML algorithms can be utilised to learn this projection between spaces. Figure 1-1 below provides a brief overview of this underlying mechanism.

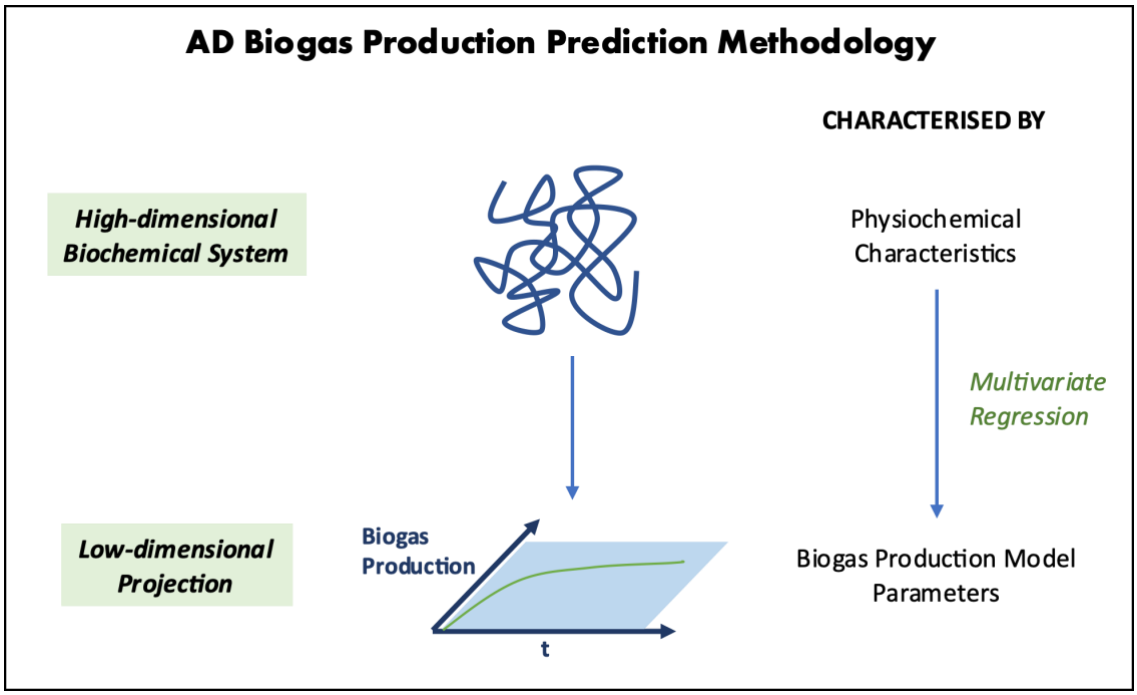


Figure 1-1 Underlying Mechanism for the prediction of the batch AD test biogas production in this research.

1.1.2 Continuous AD Performance Prediction

Although the BMP test provides valuable insights into the realistic biogas potential and reaction kinetics of a substrate and it can also be used to explore co-digestion scenario and the influence of some operational parameters, the conclusions obtained from batch tests do not always apply to continuous AD processes (Matuszewska et al., 2016). For continuous AD, operational parameters including temperature, mixing intensity, hydraulic retention time (HRT) and organic loading rate (OLR) also play important roles.

Some studies have endeavoured to predict the performance of continuous AD processes with regression models. Unlike BMP test result prediction with regression models, the studies aimed at continuous AD performance prediction often have better availability of training data. The training data include feedstock compositions, digestate characteristics and biogas productions from several years of operation. For the input features of the model for continuous AD prediction, in addition to features similar to those used in batch AD test result prediction models, e.g. the blending ratio among substrates, the influences of two operational parameters – HRT and OLR – are always considered (Catenacci et al., 2019; De Clercq et al., 2020; Wang et al., 2021; Salamattalab et al., 2024). And for the model's output, besides biogas production, the concentrations of digestate characteristics indicating reactor stability including total ammonia nitrogen (TAN), volatile fatty acid (VFA), pH are also often of interest (Holubar, 2002). However, the above-mentioned data from several years' operation are not readily available for every study hopes to predict the performance of a continuous AD.

An alternative method to simulate the continuous AD process for a specific feedstock, considering the effects of HRT and OLR, is to use mechanistic models. Such models are built on mass and energy balance principles, describing the sequential reactions occurring in the AD process. The most widely adopted mechanistic model is the Anaerobic Digestion Model No.1 (ADM1). ADM1 can simulate the concentration changes of various reactants over time in the AD process. However, the major challenges in utilising ADM1 are the need for a deep understanding of the underlying reactions in the AD process and the calibration of numerous parameters within ADM1 based on experimental observations (Xu et al., 2015).

The aim of this part of the research is to demonstrate a framework for the systematic calibration of ADM1 with Bayesian inference. This method aims to reduce the requirement for extensive experimental observations while enhancing the reliability of the calibrated parameters.

1.2 Aim and Objectives and Thesis Structure

1.2.1 Aim and Objectives

The aim of this study is to facilitate the prediction of the performances of batch and continuous AD processes by statistical tools, specifically by multivariate ML regression models and Bayesian inference.

Objective 1: To preliminary validate the mechanism for predicting the batch AD test result from physiochemical characteristics of the feedstock using the multivariate ML regression model.

Objective 2: To overcome the data limitation problem in training ML regression models for batch AD test result prediction via generating additional data through a statistical experimental design.

Objective 3: To interpret the trained ML model with mathematical expressions by metamodeling the ML model with Meijei G-functions. Additionally, to verify the general predictability of the metamodel on external data.

Objective 4: To demonstrate a framework for the systematic calibration of the processed-based model ADM1 using the Bayesian inference with a reduced requirement for experimental observations while increasing the reliability of the calibrated parameters.

1.2.2 Thesis Structure

This thesis consists of two parts. The **Part I** is for batch AD test result prediction, and the **Part II** is for continuous AD performance prediction. There are in total seven chapters.

The current chapter first introduces AD technology, the concept of BMP and the BMP test, and then the need for predicting BMP test results. It highlights the problem of limited data for training ML models for BMP test result prediction. Subsequently, it discusses the prediction of the continuous AD performance using mechanistic and statistical models, and the challenges associated with applying mechanistic models. Finally, this chapter presents the aim and objectives of the study.

Chapter 2 provides a critical review of the key physiochemical parameters that influence biogas production in the AD process, methods for material BMP prediction and introduces factors causing variations in BMP test results. It then emphasises the impact on the generalisation ability of the ML model for BMP test result prediction when training and testing the model with data collected from the literature. Finally, the chapter presents the challenges associated with calibration of the continuous AD model, ADM1.

Chapters 3, 4, and 5 comprise **Part I** of this study, which focuses on predicting batch AD test results using ML regression models. Each chapter presents the methodologies, results, and discussions corresponding to *Objective 1*, *Objective 2*, and *Objective 3*, respectively.

Chapter 6 is **Part II** of this study, addressing the calibration of ADM1 for continuous AD performance prediction with Bayesian inference.

Chapter 7 provides a brief conclusion of the study and offers suggestions for future work.

2 Literature Review

2.1 AD Overview

2.1.1 AD Process

AD process can be divided into five steps, (1) disintegration, (2) hydrolysis, (3) acidogenesis, (4) acetogenesis and (5) methanogenesis (Batstone et al., 2002). Figure 2-1 is a simplified depiction of the different stages of the AD process.

The first two steps, disintegration and hydrolysis, are extracellular reactions. During disintegration, complex particulate materials and decayed inactive biomass are converted into macromolecules, including carbohydrates, proteins, lipids, and inert materials that cannot be further degraded during the AD process. This step involves partly non-biological processes such as lysis, non-enzymatic decay, and shearing. Subsequently, in the hydrolysis step, these organic materials are further broken down into monosaccharides, amino acids, and long-chain fatty acids (LCFAs) with the assistance of extracellular enzymes secreted by hydrolytic microorganisms (Batstone et al., 2002).

In the acidogenesis step, smaller molecules are degraded into VFAs including propionic acid, butyric acid, and valeric acid. By-products like ammonia (NH_3) and H_2S are also generated during this step. Following acidogenesis step, the organic acids are further converted into acetate, H_2, and CO_2 by acetogenic groups. Some LCFAs produced during the hydrolysis step are also utilised by acetogens to generate acetate, H_2, and CO_2 through a process called β-oxidation (Cavaleiro et al., 2008; Palatsi et al., 2012).

Finally, in the methanogenesis step, acetate, H_2 and CO_2 are transformed into CH_4. This step involves two types of methanogens: heterotrophic acetoclastic methanogens, which primarily utilise acetate, and autotrophic hydrogenotrophic methanogens, which utilise H_2 and CO_2 (Batstone et al., 2002).

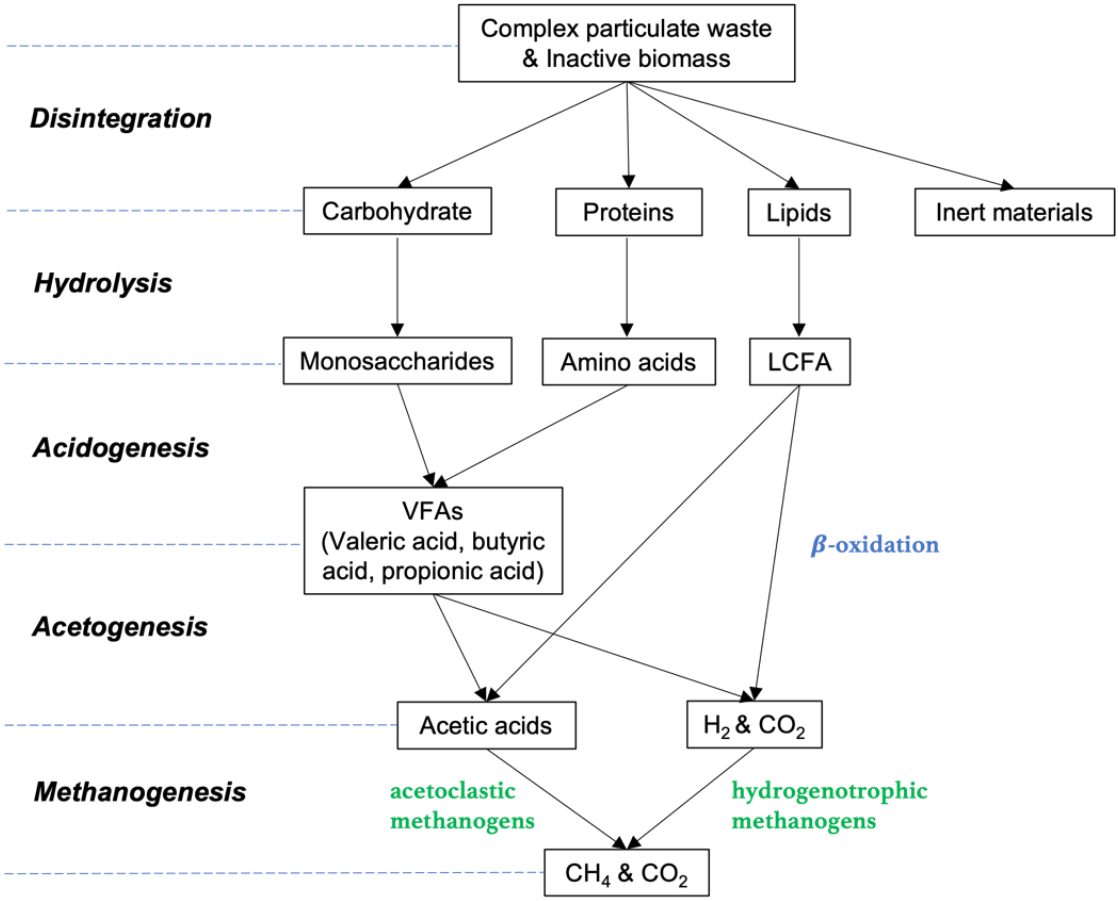


Figure 2-1 AD process flow chart.

Overall, AD is a complex process that requires various microbial groups to facilitate balanced sequential reactions without accumulating intermediates from any specific step.

2.1.2 Key Parameters in AD Process

2.1.2.1 Temperature

The operating temperature is one of the most influential parameters in the AD process due to its impact on the activity of microorganisms and enzymes. AD can be categorised into three types based on operating temperature: psychrophilic AD (usually below 20°C), mesophilic AD (20°C to 45°C), and thermophilic AD (50°C to 60°C) (El-Mashad et al., 2004).

Thermophilic AD offers several advantages, including a higher biogas production rate (Gallert & Winter, 1997), a higher organic matter degradation rate (Sánchez

et al., 2001), and better pathogen reduction (Jiang et al., 2020). However, thermophilic AD is more energy-intensive and sensitive to fluctuations in the reactor environment, requiring more precise control of reactor temperature, with variations needing to be kept within 0.6°C per day (Appels et al., 2011; Kim et al., 2006).

2.1.2.2 pH

The pH value in an anaerobic digester is jointly influenced by the concentration of VFAs, alkalinity, and the CO_2 pressure of the biogas (Zaman, 2010). Maintaining a stable pH is crucial for the diverse microorganisms involved in the AD process to function effectively. The optimal pH range for AD is between 6 and 8 (Ward et al., 2008).

VFA accumulation can occur when the VFAs produced during the hydrolysis and acetogenesis steps are not utilised by methanogens promptly. Inhibition of the AD process has been reported when the VFA concentration exceeds 10 g/L (Aguilar et al., 1995; Fortela et al., 2016). This can be caused by a high OLR (Buyukkamaci & Filibeli, 2004) or the presence of inhibitory substances such as ammonia, H_2S, and heavy metals (Gerardi, 2003).

Alkalinity measures the buffering capacity of the digester. It can be increased by adding chemicals such as sodium bicarbonate and lime, but the amount added must be carefully controlled (Kuo & Cheng, 2007).

Regular analysis of pH, VFA concentration, and alkalinity of the digestate is essential for monitoring the stability of continuous anaerobic digesters.

2.1.2.3 C/N ratio

The C/N ratio affects the composition and activity of microbial communities involved in AD. A low C/N ratio indicates that the material is protein-rich, which might pose a risk of ammonia inhibition (Khalid et al., 2011). The optimal C/N ratio range for AD is generally in the range of 20-30 (Puyuelo et al., 2011; X. Wang et al., 2012; Zeshan et al., 2012). However, for certain wastes, such as low-strength wastewater, a lower C/N ratio around 10 has been reported to

achieve the highest biomethane production (Ceron-Vivas et al., 2019; Zhang et al., 2013).

2.1.2.4 Lipids and LCFAs

LCFAs are produced during the hydrolysis of lipids. They comprise straight-chain fatty acids with 12 or more carbon atoms. The most common LCFAs are palmitic acid (C16, saturated), stearic acid (C18, saturated), oleic acid (C18, unsaturated) and linoleic acid (C18, unsaturated) (Kaneko et al., 2008).

The conversion of lipids to LCFAs occurs relatively quickly with the lipases secreted by hydrolytic bacteria. LCFAs are then transported into cells and further degraded to acetic acid and H_2 via β-oxidation. This multi-step enzymatic process involves several enzymes and coenzymes, making β-oxidation slow. Therefore, LCFA degradation is often considered a rate-limiting step in the AD process, alongside the hydrolysis step (Pavlostathis & Giraldo-Gomez, 1991).

If LCFAs produced from hydrolysis are not promptly converted to acetic acid, their accumulation can result in an increase in the reactor's pH (Ma et al., 2015). It is commonly accepted that low pH value inhibits the growth of methanogens the most (Cuetos et al., 2008; Neves et al., 2006). However, Ma et al. (2015) found that high LCFA concentrations affect hydrolytic bacteria more than methanogenic archaea, based on microbial community analysis. Additionally, LCFAs can adsorb to the cell walls and membranes of microorganisms, affecting the transport of reactants (Cuetos et al., 2008; Masse et al., 2002).

Strategies to alleviate LCFA inhibition include increasing the inoculum-to-substrate ratio and adding adsorbents (Ma et al., 2015; Palatsi et al., 2009).

2.1.2.5 Protein and Ammonia

Ammonia, including ammonium ions (NH_4^+) and free NH_3, is formed during the degradation of protein and other nitrogen-rich organic materials. Protein and ammonia help maintain a balanced C/N ratio and are essential nutrients for biomass growth (Wang et al., 2013). Additionally, ammonia contributes to the buffer capacity in the digester, as the VFAs produced during the AD process can be neutralised by ammonia (Wang et al., 2013).

Despite the nutritional and buffering properties of ammonia, excess ammonia in the AD reactor can inhibit methanogens. Mlinar et al. (2022) found that methanogens, which have robust membranes due to their archaeal nature, are primarily influenced by NH_3 through its passive diffusion into the cell. In contrast, acetogens are affected by both NH_4^+ and NH_3 due to their less stable membranes. Additionally, previous studies indicated that acetoclastic methanogens are more sensitive to ammonia than hydrogenotrophic methanogens (Walker et al., 2011). Ammonia can disrupt cellular processes such as enzyme activity and energy metabolism, leading to decreased methane production. The inhibitory concentration of TAN varies significantly across studies, ranging from 1.7 to 14.0 g/L. The concentration of TAN under which the reactor can operate stably depends on the pH, temperature, and whether the inoculum used has acclimated to the TAN concentration (Chen et al., 2008; Jiang et al., 2019).

Various strategies can help reduce ammonia levels in the reactor, such as ammonia stripping, ion exchange (Jorgensen & Weatherley, 2003), and electrochemical conversion (Lei & Maekawa, 2007). Mlinar et al. (2022) demonstrated that ammonia removal by ammonia stripping improved archaeal diversity and the abundance of propionic acetogens, thereby enhancing process stability.

2.2 BMP of a Material and Batch AD Test

2.2.1 Material BMP Estimation

2.2.1.1 Chemical Oxygen Demand (COD)

This method is based on the assumption that 1 g of CH_4 is equivalent to 4 g of COD, and thus 1 g of COD is equivalent to 350 mL of CH_4 at standard temperature and pressure (STP) (Fogacs, 2012). However, this method can lead to significant overestimation of the BMP as it does not distinguish between organic and inorganic components, nor does it account for the degradability of the material. This method is mainly used for BMP estimation in wastewater (Weinrich et al., 2018).

2.2.1.2 Elemental Composition

The BMP of a feedstock can be estimated from its elemental composition (C, hydrogen (H), oxygen (O)) following the Buswell-Muller formula (Eq. (2-1) (Buswell & Mueller, 1952) or the Boyle formula (Eq. (2-2) (Boyle, 1977). The latter introduced the N and sulphur (S) additionally to (2-1, with trace gases NH_3 and H_2S generated in the AD process estimated.

$$C_aH_bO_c + \left(a - \frac{b}{4} - \frac{c}{2}\right)H_2O \rightarrow \left(\frac{a}{2} + \frac{b}{8} - \frac{c}{4}\right)CH_4 + \left(\frac{a}{2} - \frac{b}{8} + \frac{c}{4}\right)CO_2 \quad \textbf{(2-1)}$$

$$\begin{aligned} C_aH_bO_cN_dS_e + \left(a - \frac{b}{4} - \frac{c}{2} + \frac{3d}{4} + \frac{e}{2}\right)H_2O \\ \rightarrow \left(\frac{a}{2} + \frac{b}{8} - \frac{c}{4} - \frac{3d}{8} - \frac{e}{4}\right)CH_4 + \left(\frac{a}{2} - \frac{b}{8} + \frac{c}{4} + \frac{3d}{8} + \frac{e}{4}\right)CO_2 \\ + dNH_3 + eH_2S \end{aligned} \quad \textbf{(2-2)}$$

According to Eq. 2-2, the theoretical BMP (BMP_{theo}) can be calculated from Eq. (2-3.

$$BMP_{theo} = \frac{22.4(\frac{a}{2} + \frac{b}{8} - \frac{c}{4} - \frac{3d}{8} - \frac{e}{4})}{12a + b + 16c + 14d + 32e}\ L\ CH_4(STP)/g\ VS \quad \textbf{(2-3)}$$

Using an elemental analyser, the elemental composition of a feedstock can be easily and efficiently determined, allowing for a quick estimation of its BMP. However, BMP estimates derived from this method can significantly deviate from experimentally analysed results as well. This discrepancy arises primarily from the lack of consideration for the material's degradability. Additionally, part of the substrate is utilised by microorganisms for biomass growth (Batstone et al., 2002; Labatut et al., 2011). By integrating bioenergetic information during degradation with elemental composition, the fraction of organic substrate consumed for biomass growth can be estimated using a method developed by McCarty (1972). Even so, the biodegradability of the material is still not accounted in. By contrast, estimating BMP based on the chemical composition of a material takes into account the material’s degradability, potentially providing a more accurate prediction.

2.2.1.3 Chemical Composition

Two methods are commonly used for the classification of chemical components in a substrate. They are Weender and Van Soest methods respectively (Naumann & Bassler, 1976) (Figure 2-2). Considering the Van Soest classification of the volatile solid (VS) part, the non-fibre carbohydrate, protein and lipid are readily degradable compared to the fibre. As demonstrated by Vazifehkhoran & Triolo (2019), given a retention time of more than 15 days, they are almost entirely degradable in the mesophilic AD process. However, for hemicellulose and cellulose, despite their degradation following near first-order kinetics at long retention times, only 45% and 34% of them respectively are decomposed.

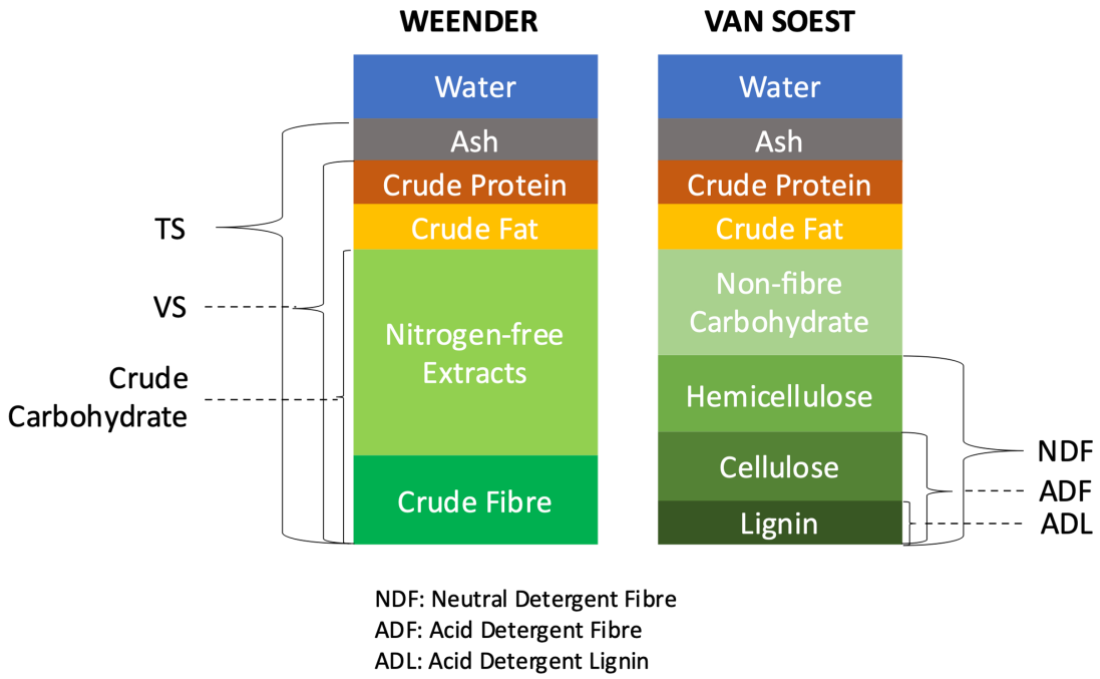


Figure 2-2 Components of Weender and van Soest classification of feedstock nutrients (Weinrich et al., 2018; Liebetraul et al., 2017). (Adapted by permission from the copyright holders, Springer Nature)

The BMP estimation of a material from its chemical components can be divided to two categories according to the nature of the material. The first scenario applies to materials that are mainly composed of non-fibre carbohydrate, protein and lipid. The BMP_{theo} of each component can be estimated from its empirical

stoichiometric formula using the Buswell-Muller formula or the Boyle formula, as introduced in the previous section. Consider $C_6H_{10}O_5$ for carbohydrates, $C_5H_7O_2N$ for proteins, and $C_{57}H_{104}O_6$ for lipids, the BMP_{theo} of such materials can be calculated as follow (Eq. (2-4) (Angelidaki & Sanders, 2004). All the components in the equation are expressed as the percentage of VS.

$$BMP_{theo}(mL\ CH_4/g\ VS) = 415 * Carbohydrate\ (\%) + 496 * Protein\ (\%) + 1014 * Lipid\ (\%) \quad \textbf{(2-4)}$$

By comparison, estimating the BMP of substrates high in less readily degradable lignocellulosic components, such as energy crops and manure, is particularly challenging. Previous studies have shown that among the fibre components, hemicellulose, in particular, has demonstrated a positive correlation with the ultimate biogas yield of lignocellulosic materials (Dandikas et al., 2015; Amon et al., 2007; Rath et al., 2013). Yang's (2021) study further revealed that hemicellulose exhibited the highest degradation rate in AD compared with other lignocellulosic components. In contrast, lignin is non-degradable in AD due to the absence of necessary enzymes such as lignin peroxidases, manganese peroxidases, and laccases, which are typically produced by aerobic fungi and bacteria (Mauseth, 1988). Thus, the content of lignin is negatively correlated with the BMP of a material (Li et al., 2018). Meanwhile, lignin reduces the hydrolytic rates of cellulose and hemicellulose by forming covalent and non-covalent linkages with them and binding with enzymes due to its hydrophobic properties (Ju et al., 2013). The complicated shielding effect caused by lignocellulosic matrix significantly reduces the BMP capacity of a material (Azhar & Stuckey, 1994).

The BMP of feedstock rich in lignocellulosic materials can also be estimated from its chemical components, which is the second scenario. However, unlike using fixed coefficients for individual components as in Eq. (2-4, regression models with various chemical components as predictors, the coefficients of which are determined by fitting BMP test results, is employed for BMP prediction of lignocellulosic materials. However, different components have been chosen as predictors for the regression models in different studies.

Triolo et al. (2011) found that lignin concentration was the strongest predictor of BMP results for energy crops and manure. Models with a single predictor – 'lignin' – could predict BMP test results with the square of the sample correlation coefficient from 0.76 to 0.91. The negative contribution of lignin to the BMP was confirmed by its negative coefficient in the fitted regression models. Manure being the excreta of the animals, contains higher lignin content compared to the crops, after easily degradable components being taken up by the animals. Despite this, it was showed in the study of Triolo et al. (2011) that the lignin had a stronger influence on BMP test results of energy crops than manure. This was attributed to the fact that the complex lignocellulose matrix in crops were not broken down by the animal digestion system as in manure.

Gunaseelan (2007) found that no single component, such as ADF, lignin, or cellulose, could satisfactorily explain the variation in BMP test results for fruit and vegetable solid wastes. Instead, he developed a multiple linear regression (MLR) model with five predictors: ADF, lignin/ADF, N, ash, and total carbohydrate. This model explained over 90% of the variation in BMP results.

Rather than experimenting with various combinations of chemical components as predictors for the regression model, later studies conducted analyses to identify influential components for use in MLR models. Dandikas et al. (2015) and Xu et al. (2020) employed principal component analysis (PCA) to examine correlations between chemical components and BMP results. In PCA, the coordinate system is formed by the first and second principal components and each vector in the coordinate system represents a variable. The length of a vector indicates the extent to which this variable explains the variance in the data, while the angle between two vectors reflects the correlation between variables. The MLR model in Dandikas et al. (2015) with three important predictors screened out by PCA – ADL, hemicellulose, and crude protein – was able to fit biogas production results of grassland plant species with a coefficient of determination (R^2) of 0.75.

Table 2-1 organises regression models fitted for the prediction of BMP test result for lignocellulosic materials with various chemical components as predictors from some of the most cited papers. As observed, different predictors were adopted

for the models in different studies. This can be attributed to two reasons. First, as stated above, early studies lacked systematic analysis of influential features. The model was often selected by choosing the one that best fit experimental data among candidate models with random combinations of features. Second, despite these studies all focused on lignocellulose-rich materials, significant variation existed in the relative abundances of individual components. For instance, although lignin contents in the fruit and vegetable waste samples in Gunaseelan (2007)'s study and the grassland biomass samples in Dandikas et al. (2015)'s study were comparable, the former exhibited significantly higher levels of carbohydrates and lipids compared to the latter, which contained more cellulose and hemicellulose.

The chemical components that exhibited strong correlations with BMP test results differed across datasets in various studies. Finding consistent relationships between chemical components and BMP test results can be more challenging when data are collected from different studies. Pabón-Pereira et al. (2020) organised data on fibre compositions and biodegradability of lignocellulosic samples from various studies, which encompassed plants, woody biomass, newspaper, and solid residue. While individual study revealed apparent negative correlations between lignocellulosic components and biodegradability results, this relationship weakened significantly when data from multiple studies were amalgamated. This observation was primarily attributed to the variations in relative abundances of fibre and non-fibre components and potential inhibitory compounds in the diverse range of lignocellulose-rich materials studied. In addition to these, the accuracy and reproducibility of the BMP test results obtained in different studies is another contributing factor. The next chapter provides a detailed introduction to the BMP test system and various factors that influence the BMP test result, especially the operational procedures and parameters.

Table 2-1 Regression models on the prediction of biogas or BMP for lignocellulose materials (DM: dry matter).

Source	Year	Sample Type	# of data	Model	Unit of predictors	Fitting	Validation	Predictor selection
Gunaseelan	2007	Fruits and vegetable solid waste	24	BMP (L CH_4/g VS) = 0.45 + 0.35*Carbohydrate - 0.32*ADF - 0.18*Lignin/ADF - 0.41*N - 3.4*Ash	ADF (including lignin, cellulose and ash), Carbohydrate: g/g VS N: mg/g	R^2 = 0.90	On five out of training sample data. R^2= - 0.72	
Amon et al.	2002; 2003; 2004; 2007;	Maize	16+	BMP (mL CH_4/g VS) = 19.05*crude protein + 27.73 *crude fat + 1.8*cellulose + 1.7*hemicellulose	% of DM	R^2 not reported	No	
Triolo et al.	2011	Energy crops and manure	19	BMP (mL CH_4/g VS) = -1.649*lignin - 0.035*cellulose + 430	% of DM	R^2 = 0.89	On 60 samples from external studies. R^2 = 0.24	
Rath et al.	2013	Maize	16	Specific biogas yield (L/kg organic matter) = 64.83 - 26.78*ADL + 42.99*hemicellulose +128.57*crude fat - 36.57*soluble carbohydrate	% of DM	R^2 = 0.83	No	Pearson correlation analysis

Dandikas	2015	Grassland species	60	BMP (mL CH_4/g VS) = 370 + 0.21*crude protein + 0.05*hemicellulose – 1.61*ADL	% of VS	R^2 = 0.75	No	PCA

2.2.2 BMP Test

The BMP test is widely used to evaluate the methane potential, reaction kinetics, and biodegradability of specific substrates in the AD process. It operates as a batch test where the substrate is blended with 'inoculum' – a diverse mix of anaerobic bacteria, archaea, and sometimes fungi – and metabolised under the anaerobic condition. The resulting biogas is collected, and the CH_4 percentage is measured. Typically, the BMP result is expressed as the volume of CH_4 produced per unit mass of organic material added or degraded, which can be quantified in terms of VS content for solid materials or COD for liquid materials.

The microcrystalline cellulose (MCC) is usually selected as a control substrate to validate the overall integrity of the BMP system and the operational procedure because of its known BMP_{theo} (415 mL CH_4/g VS) and its degradation involves all the steps in the AD process (Koch et al., 2017) .

BMP test result can provide insight for the full-scale anaerobic digester design and operation. Besides the material biodegradability information, it has been proved that the degradation rate derived from the BMP test helped with the approximation of the kinetic parameters in the continuous anaerobic digester (Batstone et al., 2009; Jensen et al., 2011; Souza et al., 2013b; Li et al., 2017).

Despite the considerable utility of the BMP test, variations in its set-up across different studies have hindered the comparison of results across various papers. To address this problem, some research groups have published standard BMP test procedures. The mostly adopted protocols are the one by Hansen et al. (2004), the standard procedure VDI 4630 by the Association of German Engineers (VDI 4630, 2006), the one by the Anaerobic Digestion Specialist Group of the International Water Association (IWA) (Angelidaki et al., 2009) and the one by Holliger et al. (2016). However, an inter-laboratory study comparing BMP test results of standard substrates – MCC and three others – obtained by 21 laboratories in Germany revealed only 8-26% inter-lab reproducibility, despite adherence to the same VDI 4630 standard protocol (Weinrich et al., 2018). Although the standard method followed was consistent across labs, variations in

equipment and differences in inoculum source were identified as significant contributors to discrepancies. In the following sub-sections, the potential factors that influence the BMP test result are organised.

2.2.2.1 BMP Equipment

(1) Reaction Vessel

Two types of reactors are commonly adopted for the BMP test. The first type is the syringe reactor, where the inoculum and substrate are mixed by continuous rotation in an incubator. The biogas produced is measured by the distance the piston moved. This is referred as Hohenheimer biogas yield test (Hülsemann et al., 2020). The second type is the bottle reactor. It can either be equipped with an automated mixer or manually shaken if the substrate is watery and readily degradable (Wang et al., 2017). The reactor is usually incubated at either 37±2 °C for mesophilic AD or 55±1 °C for thermophilic AD (VDI 4630, 2006).

(2) Gas Production Measuring Unit

The method used for biogas volume evaluation influences the BMP test result greatly. Four widely used methods are water displacement method, manometric method, gas sampling bag/syringe method and gas flow meter.

The liquid displacement method has the advantages of low cost and minimal maintenance requirement. The biogas production can either be calculated from the water level drop in the water column or by measuring the weight of water displaced. An important thing to consider for the liquid displacement method is the barrier solution used. Barrier solutions with high acidity or salinity, which minimises the dissolution of biogas, can be chosen so that the volume of the biogas produced will be measured as a whole (BS, 2004; Kida et al., 2001; Sponza, 2003). Barrier solutions with high alkalinity can also be used to maximise the solubility of CO_2 so that only the volume of CH_4 is measured. It is harder to guarantee the performance of the former barrier solution due to the great differences in the solubility of CH_4 and CO_2 in water. The conclusions from the study of Zaman (2010) and the study of Walker et

al. (2009) agreed that the acidified water (5% H_2SO_4 and pH = 2 respectively in these two studies) did not perform better compared to tap water regarding the dissolution of CO_2. Walker et al. (2009) found that only about 15% CO_2 remained in the gasometer containing pH = 2 barrier solution after 17 days when the standard gas introduced at the beginning contained 65% CH_4 and 35% CO_2. By contrast, more than 90% CO_2 was remained in the saturated sodium chloride (NaCl) solution after 17 days. However, despite the great effect of the saturated NaCl solution on preventing CO_2 loss, the crystallisation of salt around the experimental equipment poses a great problem. Therefore, a weaker saline solution with the application of a correction factor was recommended (Parajuli, 2011). Meanwhile, the biogas production volume measured by the column style liquid displacement gasometer needs to be corrected for the hydrostatic pressure head of the liquid column. Different governing equations for the height and weight type liquid displacement gasometers that take this pressure head correction into account were illustrated in Walker et al. (2009)'s paper.

Manometric method monitors the biogas production by measuring the pressure build up in the headspace of the reactor with transducers. The main drawback of this method is that the dissolution of the CO_2 increases as the headspace pressure increases, and this can lead to a pH increase in the reactor, altering experimental conditions (Rozzi & Remigi, 2004).

The collection of biogas by gas sampling bags avoids the long term contact of biogas with barrier solution and the pressure built-up in the reactor. The volume of the gas collected can be measured by water displacement method. Similarly, the gas can be measured by attaching a graduated syringe on the top of the reactor (Elbeshbishy et al., 2012; Lisboa & Lansing, 2013).

The ultra-low flow gas counter can also be used for the biogas production measurement. This type of gas counter is usually equipped with a small, inverted bucket immersed in the water. When the biogas produced fills the bucket, the bucket tips. The biogas production volume is recorded by the number of times the bucket tips. The bucket volume of each gas counter, in

other words, the volume of biogas production each tipping represents, needs to be calibrated before experiment.

2.2.2.2 BMP Reactor Contents

(1) Inoculum

Inocula from various sources were adopted in different BMP test studies, including digestate from digesters treating sewage sludge, manure, agricultural waste, etc (Zaman, 2010; Kafle & Chen, 2016; Gandhi et al., 2022). Some studies reported that inocula from different origins can lead to significant differences in BMP test results for the same substrate (De Vrieze et al., 2015; Dechrugsa et al., 2013; Gu et al., 2014). The variable abundance levels of microbial fauna, especially methanogens, and the stimulation to the carbon-lacking microorganisms after introducing carbon sources might have contributed to the differences observed (De Vrieze et al., 2015; Koch & Drewes, 2014). However, some studies found no significant difference in methane yield but observed an influence on the maximum biogas production rate with different inocula (Koch et al., 2017).

The choice of inoculum depends on the substrate used in a study. It is recommended to use an inoculum that has developed adaptation to the substrate. Additionally, digestate from anaerobic digesters treating sewage sludge is generally acceptable due to diverse microbial communities in it (VDI 4630, 2006).

It is a common practice to degas the inoculum for two to seven days at the test temperature to reduce the endogenous gas production from inoculum during the BMP test. Despite this, it is still recommended to carry out a blank control assay which contains only inoculum (Angelidaki et al., 2009; Holliger et al., 2016; VDI 4630, 2006).

(2) Substrate

Sampling the substrate necessitates ensuring its representativeness. Basic characteristics of the substrate including total solids (TS), VS, pH, VFAs, total TAN, total kjeldahl nitrogen (TKN) and alkalinity are recommended to be

tested by Holliger et al. (2016). In addition to above-mentioned parameters, some studies tested elemental compositions, while some tested chemical compositions.

The feedstock should be shredded to particle size ≤ 10 mm if homogeneousness of the feedstock is not ideal due to particle size variations (Holliger et al., 2016; Raposo et al., 2012). Also, the substrate concentration can influence the BMP test result. High concentration of substrate might overload the reactor and cause inhibited reaction. Moreover, it is not recommended to store the material for long period by freezing, as this might alter its characteristics (Holliger et al., 2016).

(3) Nutrients

Nutrient solutions containing macro-nutrients, micro-nutrients and vitamins can be added to enhance the reaction in the anaerobic digester. In Walker et al. (2010)'s study for the development of a residual biogas potential test for digestate samples, 10 mL major element nutrient solution and 1 mL trace element nutrient solution was added to each reactor containing the amount of digestate sample equivalent to 20 g loss on ignition. The macro-nutrient solution usually contains salts of potassium (K), sodium (Na), calcium (Ca) and magnesium (Mg), and the micro-nutrient solution contains salts of iron (Fe), cobalt (Co), zinc (Zn), copper (Cu), nickle (Ni), molybdate (Mo), selenium (Se) and boric acid (Raposo et al., 2006a; Zhang et al., 2003). The application of additional nutrients can be exempted if sufficient nutrients are available in the substrate or inoculum wherein extra nutrients might be counterproductive (Raposo et al., 2006b; Zhang et al., 2007).

2.2.2.3 Operational Procedures

(1) Headspace Flushing and Headspace Volume

VDI 4630 (2006) suggests flushing the headspace with N_2 prior to incubation, while the IWA protocol (Angelidaki et al., 2009) suggests flushing throughout the reaction with a mixture gas of N_2/CO_2 = 80:20 v/v. In the study of Koch et al. (2015), the comparison of the BMP test results of groups flushed with N_2

and no prior flushing indicated that the oxygen in the headspace of the reactor could cause inhibition to biogas production. And interestingly, this inhibitory effect persisted throughout the entire duration of the BMP test, rather than being limited to the initial stages.

When the CO_2 content in the flushing gas is smaller than 35% generally (considering the typical biogas composition from anaerobic digesters (Tchobanoglous et al., 2003)), and only the volume of CH_4 is monitored by the gas measuring unit, an over-estimation of the biogas production can occur. This is because when the flushing gas in the headspace is pushed out by the biogas produced, all the non-CO_2 components in the flushing gas will arrive at the measuring unit. However, in fact, the biogas contains less non-CO_2 component, which is CH_4, compared to the flushing gas. The alternative flushing gas with N_2/CO_2 ratio closer to the composition of biogas can be adopted, but it is more expensive.

By contrast, Strömberg et al. (2014) proposed applying a correction factor instead as a more cost-effective solution to the impact of the composition of flush gas. The correction factor is demonstrated by Eq. (2-5. This correction is applied until all the flush gas is pushed out of the system. Figure 2-3 showed the percentage of the over-estimation of methane production regarding to the ratio between the methane production volume and headspace volume (V_S/V_H) and the methane content in the biogas (Strömberg et al., 2014).

$$V_{OE,i} = (V_m^{i-1} - V_m^i)(X_{B,CO_2} - X_{FG,CO_2}) \quad \textbf{(2-5)}$$

where:

V_m^{i-1} and V_m^i: cumulative methane production recorded by the measuring unit at time step $i-1$ and i, in mL CH_4;

X_{B,CO_2} = CO_2 concentration in biogas, in %;

X_{FG,CO_2} = CO_2 concentration in flush gas, in %;

$V_{OE,i}$ = the volume of over-estimated biogas production at each time step.

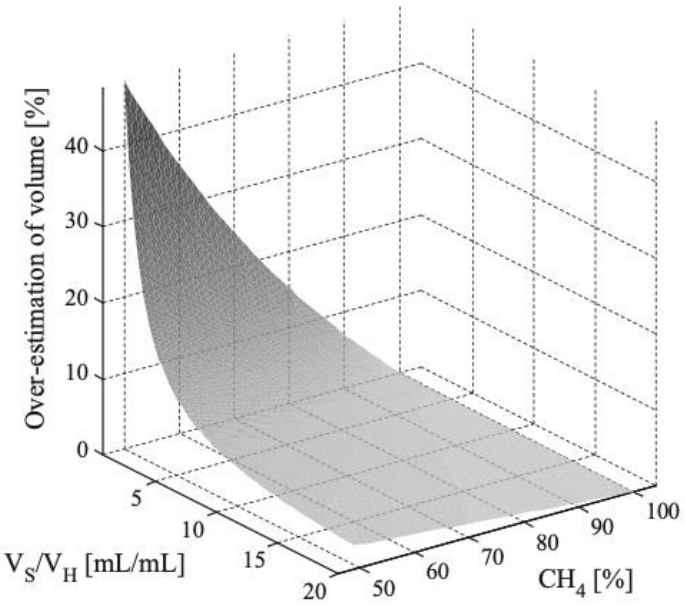


Figure 2-3 Over-estimation of methane production regarding to the ratio between the methane production volume and headspace volume (V_S/V_H) and the methane content in the biogas (CH_4, %) (Strömberg et al., 2014).

Also, the headspace volume influences the biogas production recorded by the gas measuring unit significantly. As reported by Wright (2021), underreporting of the BMP test result occurred if the methane within the headspace was not considered.

(2) SIR Ratio

The SIR is also influential to the BMP result. To avoid inhibition caused by substrate overloading, it is recommended to have an SIR between 0.25 and 0.5 based on the VS content (Cabbai et al., 2013; Holliger et al., 2016; VDI 4630, 2006). Significantly inhibited biogas production was observed when SIR > 0.5 and when SIR > 4 by Neves et al. (2004) and Hashimoto (1989) respectively.

(3) Biogas Sampling Location

For BMP test equipment with a gas measuring unit that records the entire volume of biogas, rather than just CH_4, it is essential to sample the produced biogas for gas composition analysis. There are two sampling locations for this type of BMP test equipment, either the headspace of the reactor (Lisboa &

Lansing, 2013; Zaman, 2010) or the end of the gas measuring unit (Walker et al., 2010; Yirong, 2014;Dandikas et al., 2015). The alteration of biogas composition after biogas passing through the barrier solution could be a concern for the latter, although excessive salinity and acidity of the barrier solution aim to minimise the CO_2 dissolution. If sampling from the headspace, it is good practice to extract the minimum volume required for gas chromatography (GC) analysis to minimise disruption to the reactor.

2.2.2.4 Result Reporting

It is essential to adjust the biogas volume measured at the lab environment to STP. More than 10% difference can be induced between converting the result to 0 °C, 101.325 kPa (McNaught & Wilkinson, 1997) and to 25 °C, 100 kPa (National Bureau of Standards, 1982). Assuming the biogas cools to the ambient temperature once leaving reactor and considering that the biogas is measured at ambient pressure, Eq. (2-6 is used to convert the biogas volume measured (V_{gas}) to STP, where T_{gas} and T_{STP} are ambient temperature and standard temperature, in Kelvin, and P_{gas} and P_{STP} are ambient pressure and standard pressure, in kPa.

$$V_{STP} = \frac{P_{gas}}{P_{STP}} \cdot \frac{T_{STP}}{T_{gas}} \cdot V_{gas} \quad \textbf{(2-6)}$$

Also, assuming the biogas produced is saturated with vapour, the vapour pressure should be deducted from the pressure of the biogas. The vapour pressure (p_{vap}, in mbar) can be approximated by Antoine equation (Eq. (2-7) (Strömberg et al., 2014), which is simpler and provides similar result compared to the Goff and Gratch equation used in the study of Walker et al. (2009).

$$p_{vap} = 10^{8.1962 - \frac{1730.63}{T_{gas} - 39.724}} \quad \textbf{(2-7)}$$

After converting p_{vap} to kPa, the final expression for biogas production after adjusting to STP and eliminating the influence of vapour pressure ($V_{STP}{}'$) is illustrated by Eq. (2-8.

$$V_{STP}{}' = \frac{(P_{gas} - p_{vap})}{P_{STP}} \cdot \frac{T_{STP}}{T_{gas}} \cdot V_{gas} \quad \textbf{(2-8)}$$

In summary, BMP test results indicate the biogas yield of a material in practice. This yield is influenced by material's intrinsic biogas potential, possible inhibitions from substances such as ammonia and heavy metals, substrate consumption due to biomass growth, operational procedures, experimental equipment, and experimenter practices. Figure 2-4 organises the five main factors that influence BMP test results.

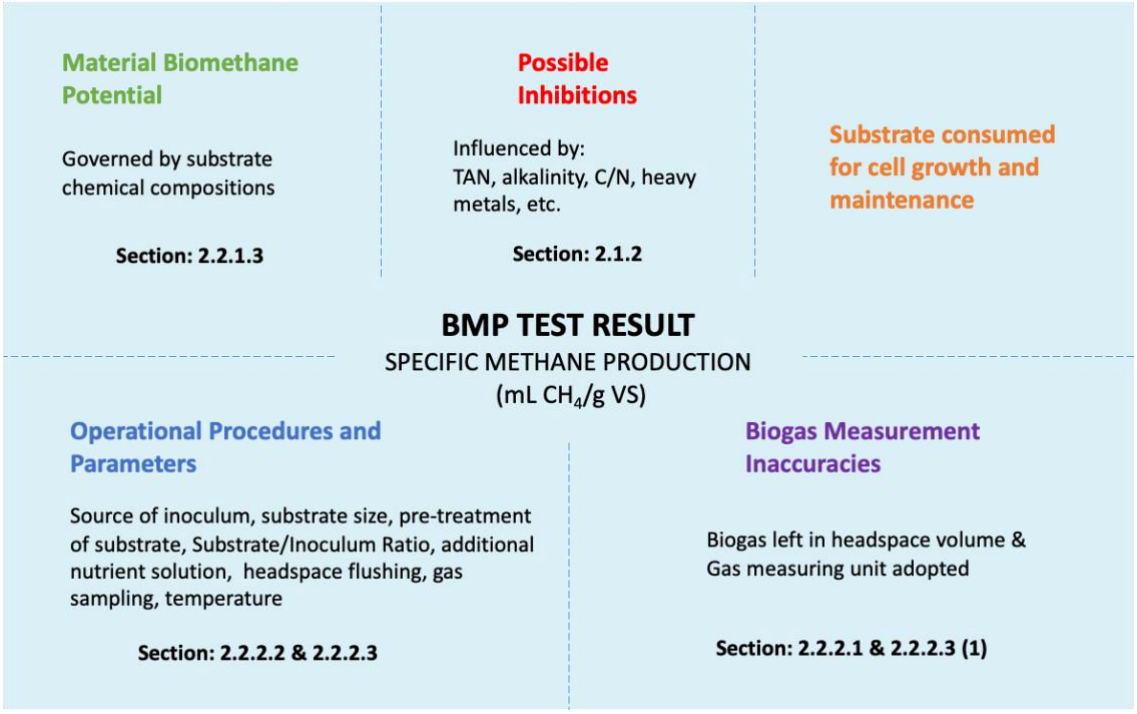


Figure 2-4 Five main factors that influence BMP test results and the corresponding chapters where they are discussed.

2.2.3 BMP Test Results Prediction

The BMP test result can be predicted by regression models fitted or trained with BMP test data. This method has two main applications. The first application is to estimate the BMP of lignocellulosic materials. As introduced in Section 2.2.1.3, considering the fact that the actual BMP of such materials deviates significantly from the BMP_{theo} calculated using Eq. (2-4 due to the complex inhibitory effects of the lignocellulosic matrix, their BMP is often derived from the BMP test. Various

models using chemical components as predictors and BMP test result as response are then built for BMP prediction for lignocellulosic materials (Table 2-1).

The second application Is to study the influence of possible inhibitors and operational parameters, which also affect BMP test result besides the intrinsic BMP of the material itself (Figure 2-4). For example, Xu et al. (2014) developed MLR models to predict the biogas yield of lignocellulosic materials. In addition to the concentrations of chemical components, TAN and alkalinity of the feedstock and SIR were also considered as predictors. After a further identification of influential features, a MLR model with cellulose, lignin, extractive contents, and SIR as predictors was built and it successfully predicted the biogas yield results of a test set containing eight samples with an R^2 = 0.98.

Black-box ML regression models typically outperform traditional linear models in predicting outputs from multiple correlated input features within complex nonlinear processes. They have been increasingly applied for BMP test result prediction in recent years (Wang et al., 2023). Almomani (2020) developed an artificial neural network (ANN) model to predict biogas productions from different blending scenarios between agricultural solid wastes and cow manure, studying the influences of operating temperature and sodium bicarbonate dosage. Khashaba et al. (2022) built an ANN to predict biogas production from biochar-enhanced sewage sludge AD using 51 BMP test data collected from the literature. The predictors included TS, VS of sewage sludge, SIR, properties of biochar, biochar dosage, and operating temperature. Also, to forecast not only the ultimate biogas potential but also the trajectory of biogas production, the day within the BMP test duration was also incorporated as a predictor. The model achieved an R^2 of over 0.98 on the test set containing six samples left out from the collected data.

ML regression models can couple with heuristic optimisation techniques to search for a combination of features that corresponds to the optimised model output, typically the highest biogas yield for BMP test. Commonly used techniques include genetic algorithm, simulated annealing, and particle swarm

optimisation. Gueguim Kana et al. (2012) discovered the optimised co-digestion profile of five substrates for biogas production by applying genetic algorithm to a trained ANN. Alrawashdeh et al. (2022) used a third-order polynomial equation to model biogas production under the boosting effect of Fenton pre-treatment. Particle swarm optimisation was then applied to find the dose of the Fenton process (H_2O_2 concentrations and Fe^{+2}/H_2O_2 ratio) that maximises biogas production.

2.2.4 General Applicability and Interpretation of ML Models for BMP Test Result Prediction

2.2.4.1 Data Availability and General Applicability

The quality and quantity of data available for model training significantly influence the quality of the model trained (Geng et al., 2024). Generally, ML models can better capture complex and highly nonlinear relationships within a system compared to linear models, resulting in higher predictive power (Sahni et al., 2018). However, training a ML model to convergence, meaning that it can accurately represent patterns and relationships in the training data and have a good generalisation performance on unseen data, requires more data (Bishop, 2006). However, the time required to perform a BMP test makes obtaining enough data within one study challenging. As a result, to have enough data to train a ML model for BMP test result prediction, many studies collected BMP test data with features under investigation measured from the literature.

However, as introduced in Section 2.2.2, BMP test results vary greatly across different studies due to different operational procedures and parameters adopted. Discrepancies in biogas production volume measurement, resulting from the use of different gas measuring units, further contribute to these variations. These inconsistencies affect the response of the training dataset used for ML models for BMP test result prediction. Additionally, variations in chemical composition analysis results from different studies affect the input features of the ML training dataset. The chemical composition analysis results, especially the lignin concentration, are influenced by the chosen analysis method (Hatfield & Fukushima, 2005). Pabón-Pereira et al. (2020) complied data on the contents of

lignocellulosic components and the biodegradability of samples from the literature where various methods were adopted for fibre content analysis. Significant negative correlations between the lignocelluloses and the biodegradability were observed only when data from studies using van Soest method for fibre analysis were considered. Otherwise, the data exhibited great dispersion, devoid of discernible correlations. Consequently, an ML model trained on literature-collected data might not have captured the ground truth relationship between chemical composition and BMP test result of a feedstock.

To assess the generalisation performance of an ML model, it is essential to record the mode''s prediction accuracy on a separate test dataset. The composition of this test set is crucial, as it directly impacts the true measure of the mode''s generalisation ability. If the test set consists of samples left out from the same dataset used for training, a high prediction accuracy should be scrutinised carefully to determine the true generalisability of the trained ML model.

For instance, in Khashaba et al. (2022)'s study about biochar enhanced AD of sewage sludge, 51 data collected from four papers were used to build an ANN model to predict the BMP test result from sludge and biochar characteristics, biochar dosage and reactor temperature. Six data out of the 51, spanning all four papers, were set aside as the test dataset. The trained model achieved an impressive R^2 value of 0.99 on the test set. However, a closer examination of the dataset reveals that samples from different papers exhibit distinct groupings across certain features. For instance, the surface areas of the biochar used in two papers by the same author (Shen et al., 2015 and Shen et al., 2016) is significantly larger (around 300 m^2/g) compared to the biochar used in the other two papers (around 30 m^2/g). Additionally, the reaction temperature in one paper (35 °C) (Zhang et al., 2019) differs from the temperatures in the other three papers (27 °C, 37 °C, and 55 °C). As a result, ANN model might have learned to distinguish test data from the studies of these two authors just according to the biochar surface area and temperature features, leading to predictions that closely match the other samples from these two authors in the training dataset. However,

the generalisability of the trained ANN model to predict the biogas production of a sample from the studied features might be limited.

A better practice to test the general applicability of the trained model for BMP test result prediction might be using samples from studies which were not included in the training dataset as did in the studies of Triolo et al. (2011), Vazifehkhoran & Triolo (2019) and Wang et al. (2023). For example, Triolo et al. (2011) tested their model for BMP test result prediction for lignocellulosic materials on 60 external data primarily consisting of crops and manure samples. The R^2 on this test set was 0.24. Similarly, Wang et al. (2023) trained a K-nearest neighbour (KNN) model on 277 data comprising agricultural residue, forest residue, and urban solid waste samples obtained from various literature sources. This trained KNN model was subsequently evaluated on a test set comprising 17 samples from separate studies. An R^2 value around 0.8 was attained, suggesting a good generalisation performance of this trained model.

To mitigate the consistency issue of the BMP data collected from diverse sources and meanwhile satisfy the scale of data needed by ML model training, some studies used data augmentation algorithms to generate artificial training data from the limited data on hand. The synthetic data and the original data are then combined for model training and testing. Methods such as GAN and mixup data augmentation were used in previous studies (Adeleke et al., 2023; Geng et al., 2024). GAN aims to learn the distribution of the original data and generate synthetic data that comply with the distribution of the real data (Goodfellow et al., 2020). However, it was observed in the study of Adeleke at al. (2023) that the synthetic data generated by GAN could not capture the variance existing in the original data. Instead, the samples with less extreme features were augmented. Although the model shown an R^2 value of 0.92 on the test set composed of both original and synthetic samples, the mapping information conveyed by the samples at the peripheral area of the original dataset might have not been captured by the model trained with augmented data. While mixup data augmentation method generates a new synthetic example by linearly interpolating between two randomly selected examples from the dataset. The

strength of the interpolation is controlled by a mixup super parameter λ which follows a beta distribution. Mixup encourages the model to learn more robust and generalised decision boundaries by blending examples and labels from different classes, effectively regularising the model and reducing overfitting (Zhang et al., 2017).

2.2.4.2 Model Interpretability

Despite the capability of ML models at grasping complex interactions among variables in a highly nonlinear system, it is difficult to understand how a ML model maps from the input space to the output space mathematically. This has restricted the application of ML models in some domains where the transparency of the modelling results is required such as medical and financial domains (Alaa & van der Schaar, 2019a; Lundberg & Lee, 2017). Also, in the science domain, a model might be expected to help with the understanding of an underlying phenomenon but not only making prediction with it (Alaa & van der Schaar, 2019b).

It is possible to explain an ANN model by extracting the weights, bias and activation functions between all the neurons in every two layers. However, it is highly impractical to do so. For tree-based ML models such as eXtreme Gradient Boosting (XGBoost), the feature importance can be returned from a trained XGBoost model directly, based on how frequently a feature is used to split the data across all decision trees in the ensemble, weighted by the improvement in the model's performance achieved by each split. The default metric used to measure the improvement after a split is ‘Gain’, which is the reduction in the loss function such as mean squared error (MSE). For example, Xu et al. (2021) built an XGBoost model to understand the influence of zero-valent iron on the AD of waste activated sludge and manure. The feature importance results returned from the trained XGBoost model revealed that the three most important predictors were TS and soluble COD (sCOD) of the feedstock, and dosage of the zero-valent iron.

Some model-agnostic feature importance analysis methods are also popular for interpreting the prediction of ML models, such as permutation feature importance, partial dependence plot (PDP) and SHapley Additive exPlanations (SHAP) value.

The permutation feature importance method quantifies the importance of a feature via the changes in the prediction error caused by permutations of the value of a feature compared to the original model. The PDP can help to visualise the marginal effect of one feature on the model output while the influences of other features are averaged out (Gupta et al., 2023). Specifically, PDP is the average of multiple individual conditional expectation plots. For each individual conditional expectation plot, it takes one instance from the training dataset, keeping the values of all the other features fixed while changing the value of the studied feature, and the variation of the prediction results was plotted. The slope of the PDP plot for a feature indicates its influence on the model prediction result. In the study of Wang et al. (2021) where the co-digestion of 31 waste streams were studied, the positive slopes of the PDP plot indicated that all the waste streams positively influenced the biogas production. Also, by comparing the slope values of each waste stream, the per unit rendering waste flow had the strongest influence. While in the study of De Clercq et al. (2020), which also investigated the biogas production from co-digestion of various wastes, the PDP plot indicated that the impact of increasing the volume of a waste input began to plateau after reaching a certain threshold, rather than continuing to increase. Additionally, a two-variable PDP plot was used to examine the interactions between two waste streams to provide insight to the co-digestion dynamics.

Though permutation feature importance and PDP plot tell how the prediction result of a model is influenced by the changing of a certain predictor. The main problem of these two methods is that improbable feature combinations are generated when varying the value of one feature while keeping the values of other features fixed (Friedman et al., 2012; Lundberg et al., 2018). The SHAP value method alleviates this problem by forming a less improbable new sample part from the training sample instance under study and part from another sample randomly drawn from the training dataset (Lundberg et al., 2018; Gupta et al., 2023). The SHAP value method studies the influence of one input feature on the prediction result for a single instance of training samples compared to the average prediction result. And the overall importance of this input feature can be summarised from the SHAP values of multiple instances. SHAP method is both

consistent (being able to assign more importance to a feature when a model relies more on it) and locally accurate (the importances of all features sum up to the total importance of the model) (Lundberg & Lee, 2017). Traditional methods such as the ‘Gain’ of tree-based model and the ‘Saabas’ method of the ‘treeinterpreter’ package (Saabas, 2015), are biased towards attributing more importance to splits in the lower branches.

In the study of Adeleke et al. (2023) where a gradient boost regression model was built to predict the biogas production from lignocellulose components. The feature importance analysis result from SHAP value method did not agree with the result derived from the permutation feature importance method. The latter ranked the cellulose as the most important feature, while the hemicellulose was the most important feature with the highest average SHAP value for the former. From the ‘beeswarm’ plot of the SHAP value package, where the SHAP values of all the samples for each feature were depicted, the reason why hemicellulose was analysed as the most important feature by SHAP method was illuminated. It was shown that some samples with high hemicellulose content exhibited outstandingly high biogas production. Meanwhile, the biogas productions of most of the samples with low hemicellulose contents were significantly lower than the average biogas production.

However, all the model-agnostic feature importance analysis methods introduced above only provide a ranking of feature importance based on their influence on the model's output. None of them mathematically elucidate the mapping from the input space to the output space of the ML model.

2.3 Continuous AD Performance Prediction

The BMP test offers insights into the practical biogas potential of a substrate under the influence of potential inhibitors, and the influence of operational parameters. Conclusions drawn from batch tests are informative to the design of continuous AD systems. For instance, Bohutskyi et al. (2018) examined the impact of C/N ratio on the co-digestion of algae biomass and cellulose with batch tests. They subsequently employed another model to relate batch test biogas production, described by the Gomperz model, with biogas production from

continuous stirred tank reactors based on energy balance. Similarly, Jin et al. (2018) investigated the optimal pre-treatment time and C/N ratio for biogas production from co-digestion of corn straw and pig manure through batch tests. They then validated the performance under the discovered optimal conditions in continuous digesters.

However, the conditions identified as optimal for biogas production in batch tests may not always translate to the continuous operation of industrial-scale digesters. Matuszewska et al. (2016) stated that the optimal co-digestion blending ratio found in their batch tests might not be applicable to industrial scale digesters due to the presence of process inhibition period. Additionally, compared to the batch test, operational parameters of digesters including inflow concertation, OLR and HRT are also crucial to continuous AD processes. HRT indicates for how long the reactor's performance will be influenced by the feed at a particular day, influencing the extent of organic matter decomposition and methane production (Salamattalab et al., 2024). And although the inflow concentrations are important, the OLR indicates the rate at which organic matter is added to the digester. Higher OLRs can increase biogas production rates, but excessively high OLRs can lead to process instability (Grady Jr. et al., 2011).

Therefore, models for continuous AD performance prediction are necessary. Compared with running extended pilot AD trails, mathematical modelling of AD process has the advantage of cost-effectiveness, time efficiency and flexibility regarding the evaluation of the digester performance under various operational scenarios (Donoso-Bravo et al., 2011a).

The models for continuous AD can be divided to two types – mechanistic models and statistical models. The mechanistic models describe the processes involved in the AD process, following the mass and energy balance, and simulate the concentration changes with time of important reactants. One of the first AD model was developed in the 1980s, whereas only the acidification phase was modelled (Eastman & Ferguson, 1981). In 2002, an AD modelling task group of International Water Association developed the ADM1, which is currently the most widely used model for continuous AD process modelling (Batstone et al., 2009).

The application of ADM1 requires a thorough understanding of the complex sequential biological and physiochemical reactions involved in the AD process. Additionally, model inputs and initial conditions need be derived through time-consuming chemical analyses of the substrate and inoculum, and model parameter calibration is necessary before simulating the AD performance of a particular substrate with ADM1. These have hindered the widespread industrial application of ADM1.

The black-box nature of ML statistical models has provided a potential way to predict the performance of a continuous AD process using important features such as feed flow, feed composition and regularly monitored reactor performance indicators without knowing the reactions involved in detail.

2.3.1 Statistical Model

When it comes to applying ML/Deep Learning methods to the prediction of BMP test results, gathering sufficient data for robust model training and testing is always a challenge. However, in the studies of predicting the performance of industrial-scale anaerobic digesters by statistical models, data for consecutive feeding schemes and regular analyses of digestate characteristics over several years are typically available. The statistical models for continuous AD prediction are generally two types. The first type is the traditional features – response supervised ML model. The second type is the deep learning model, mainly long-short term memory (LSTM), considering the fact that the reactor performance data are also time series data.

For the traditional features – response supervised ML model, the structure of the data for model training and testing is that the input data is a 2-dimensional matrix in a shape of (n, f), where n is the number of data, and f is the number of input features, and the response is a vector of length n. For the LSTM model, the input data is a 3-dimensional matrix in a shape of (n, f, l), where l is commonly referred as 'lookback' or 'window size' parameter, which determines how many past timesteps the model uses to predict the current time step. And the response is a vector of length $n - l + 1$. Figure 2-5 illustrated these two structures respectively.

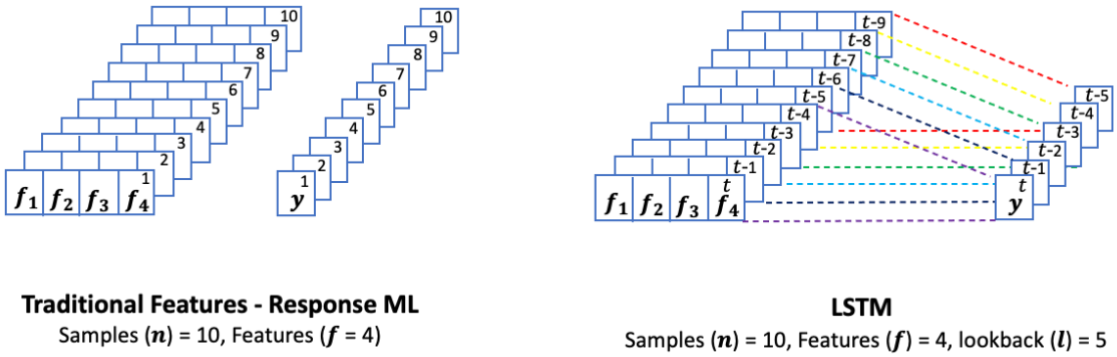


Figure 2-5 Input – Output structure of the data for 'Traditional Features – Response' ML model and LSTM model training and testing.

The importance of HRT and OLR are widely recognised in the studies establishing statistical models for continuous AD performance prediction. The influence of HRT is accounted for differently in the two types of statistical models introduced above. The examples below are provided to demonstrate the ways to take HRT into account for features – response supervised ML models.

Wang et al. (2021) trained a tree-based pipeline optimisation tool to predict the daily biogas production from the volumes of 31 waste streams, together with TS, VS, alkalinity, VFA and VFA/alkalinity of digestate, with data from over eight years' operation. To take the lag effect of the input waste streams into consideration, the volumes of each waste stream of the past seven days were included as predictors. Therefore, there were in total 31*7 + 5 input features to predict the biogas production at next timestep. Similarly, in the study of De Clercq et al. (2020), the volumes of 17 co-digestion waste streams were used to predict biomethane production using random forest and XGBoost. Instead of predicting the biogas production only for the current timestep, the models were trained to predict different horizons (1, 3, 5, 10, 20, 30 and 40 days ahead respectively) by setting model output as a vector. Such a model structure also acknowledged the fact the characteristics of the feeding at a particular day might influence digester's performance for the following several days. Not only to predict the biogas production, Holubar (2002) also built ANN models to predict the pH and VFA changes of the digestate and the biogas composition under the influence of fluctuating OLR. To account for the lag effect of OLR and using the ANN in a way

like a time series model, the pH value at previous time step (pH_{t-1}) and the volumetric loading rates at previous two time steps (OLR_{t-1}, OLR_{t-2}) were used as input features for the prediction of pH concentration at current time step (pH_t).

Although traditional features – response ML model can also be used for time series prediction via 'time-series to supervised learning' step as did in Holubar (2002)'s study, LSTM models are more capable of learning complex patterns and relationships in the data by storing historical information in the cell state. This has made LSTM models well-suited for time series forecasting tasks where temporal dependencies are important.

Salamattalab et al. (2024) used sludge characteristics including feed flow rate, dry solids and VS, and wastewater characteristics including inlet flow rate, the concentrations of biological oxygen demand, COD, total suspended solids, total nitrogen and TAN to predict the biogas production using LSTM model. Since the HRT of the digester is 21 days, the lookback window for LSTM training was set at 21. In other words, the model considered the values of the sludge and wastewater characteristics of the last 21 days for the prediction of the current timestep. In the study of Jeong et al. (2021), a LSTM model integrated with a variable selection network (VSN) which helped to process the discontinuous input features such as the biological oxygen demand, COD, total nitrogen and total phosphorus of the feed to continuous input features required by LSTM model, was trained. The output from the LSTM-VSN variable selection network model was passed through a dense layer, giving the prediction of biogas production.

2.3.2 Anaerobic Digestion Model No. 1 (ADM1)

2.3.2.1 Reaction Kinetics Described in ADM1

ADM1 describes the concentration changes with time of the substrate, products and microorganisms involved in five steps of an AD process. Specifically, it uses a system of ordinary differential equations (ODEs) with 36 state variables, representing the concentration changes of 36 substances with time (Table 6-2). 7.2Appendix D is the Peterson matrix that describe the structure of ADM1.

Three major types of kinetics are described in ADM1: (1) hydrolysis rate, (2) substrate utilisation, intermediate products production and microbial growth rates and (3) microbial decay rate (Batstone et al., 2002).

(1) Hydrolysis rate

The disintegration of complex particulate materials and the following hydrolysis of carbohydrate, protein and lipid are modelled by first-order kinetics in ADM1 (Eq. (2-9), in which r_h (g/L d) is the hydrolysis rate, k_h (d^{-1}) is the hydrolysis rate coefficient and S (g/L) is the concentration of complex particulate material, carbohydrate, lipid and protein.

$$-r_h = k_h S \quad \textbf{(2-9)}$$

(2) Substrate utilisation, intermediate products production and microbial growth rates.

The substrate consumption, production of intermediate products, and biomass growth are interconnected as shown in Eq. (2-10.

$$-\frac{dS}{dt} = r_s = \frac{r_p}{Y_{P/S}} = \frac{r_g}{Y_{B/S}} \quad \textbf{(2-10)}$$

where:

S: substrate concentration, in g/L;

r_s: substrate utilisation rate, in g/L d;

r_p: production rate of an intermediate product, in g/L d;

$Y_{P/S}$: product yield coefficient, in g/g;

r_g: biomass growth rate, in g/L d;

$Y_{B/S}$: biomass growth coefficient, in g/g.

The biomass growth rate, r_g, is in linear correlation with the biomass concentration B (g/L) (Eq. (2-11)

$$r_g = \mu B \quad \textbf{(2-11)}$$

The μ is the microbial specific growth rate (d^{-1}). It can be defined by the Monod equation (Eq. (2-12) (Monod, 1942).

$$\mu = \mu_{max} \frac{S}{K_S + S} \quad \textbf{(2-12)}$$

where:

μ_{max}: maximum specific growth rate, in d^{-1};

K_S: Half-saturation coefficient, in g/L. This is the value of S when μ is half the value of μ_{max}.

This Monad type kinetics are used to model the uptake of intermediate products including monosaccharides, amino acids, LCFA, VFAs and H_2 in ADM1 (Process (j) = 5-12 in Table D-1 Peterson Matrix (Part 1)and Table D-2).

(3) Decay rate.

In ADM1, the dead biomass contributes to the complex particulate pool and is degraded through cell lysis and endogenous respiration. The biomass decay rate, r_d (g/L d), is proportional to the biomass concentration B (g/L). It is shown in Eq.(2-13, in which k_d is the decay coefficient (d^{-1}). When decay rate r_d is higher than the biomass growth rate r_g, the microbial population in the reactor goes into the declining phase.

$$-r_d = r_d B \quad \textbf{(2-13)}$$

In addition to the three types of biochemical reactions mentioned above, ADM1 also included two physiochemical processes: acid-base equilibrium and liquid-gas transfer. Furthermore, process inhibitions caused by pH levels in the system, as well as by NH_3 and H_2, are incorporated into the substrate utilisation rate r_s.

See Chapter 6 for a detailed explanation of the acid-base equilibrium in ADM1 and 7.2Appendix D for the expression of r_s for each process included in the ADM1.

ADM1 is typically used to model the performance of a continuous AD process. And the modelling of continuous anaerobic digestion can be divided to steady

state and dynamic modelling. Steady state modelling focus on the long-term behaviour of the digester when the system has reached an equilibrium under relative stable operating conditions. The application of steady state simulation includes system design and operating parameters decision (e.g., feed rate, temperature) for optimal system performance, such as biogas production or COD removal (Bornhöft et al., 2013). Dynamic simulation instead focuses on evaluating how the system responds to variations in operating conditions, such as feed interruptions, overloads, or changes in substrate composition.

To simulate an AD process by ADM1, the concentrations of various chemical components in the feedstock and inoculum, the feeding and discharging flows of the reactor (assuming equal), and the volumes of liquid and gas phases of the reactor are needed.

2.3.2.2 ADM1 Calibration

Like many other process-based biochemical system models, the detailed modelling of complex sequential reactions comes with the cost of managing numerous parameters (Jeppsson, 1996; Gurung et al., 2020; van Oijen, 2017). In the case of ADM1, it encompasses over a hundred kinetic and stoichiometric parameters ($\boldsymbol{\theta_{ks}}$). Besides these $\boldsymbol{\theta_{ks}}$, input values ($\boldsymbol{\theta_f}$) and initial values ($\boldsymbol{\theta_0}$) of 36 state variables also need to be provided. These values are determined by the feedstock and inoculum characteristics respectively. One way to derive $\boldsymbol{\theta_0}$ is to simulate a continuous AD process with the same feedstock until reaching the steady state and then use the achieved steady state conditions as $\boldsymbol{\theta_0}'$ for further simulations (Blumensaat & Keller, 2005; Boubaker & Ridha, 2008; Ozgun, 2019; Catenacci et al., 2021). However, this approach was proven inaccurate in the study of Liu et al. (2022) as $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_{ks}}$ adopted for the initial continuous AD simulation influenced the new $\boldsymbol{\theta_0}'$ obtained significantly and consequently the applicability of the calibrated model. A reliable but rarely used way for the determination of $\boldsymbol{\theta_0}$ is by experimental analysis of the characteristics of inoculum (Pastor et al., 2010; Thamsiriroj & Murphy, 2011). These analyses need to be combined with several assumptions about the partitioning of sCOD and particulate COD (pCOD) among state variables. Also, estimations are needed for

those state variables that cannot be experimentally measured, such as the concentrations of degrading microorganisms and the concentrations of H_2 and CH_4 dissolved in the liquid phase. A complex ODE system like ADM1 is numerically solved based on the initial values of state variables and their rates of change governed by the ODEs. Any uncertainty of the initial values will be passed to the ODE solutions. Therefore, to simulate and predict the AD performance of a specific feedstock closely using ADM1, the uncertain $\boldsymbol{\theta_0}$ needs be calibrated together with $\boldsymbol{\theta_{ks}}$ using experimental observations.

It is not viable to calibrate vast amount of $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ altogether with limited experimental observations. In practice, only parameters that have significant influence on model outputs of interest are calibrated while keeping the less influential ones, e.g., the majority of the stoichiometric parameters, at their default values suggested in the benchmark model by Batstone et al. (2002). The technique to identify influential parameters is called sensitivity analysis (SA). SA can be categorised into local sensitivity analysis (LSA) and global sensitivity analysis (GSA). Local Sensitivity Analysis (LSA) evaluates model output uncertainty by varying a single parameter along a one-dimensional axis while keeping other parameters fixed. This approach provides a linear approximation of the model's behaviour at a specific point. Global Sensitivity Analysis (GSA), on the other hand, examines output uncertainty by varying multiple parameters simultaneously, exploring the multi-dimensional hypercube space of uncertain parameters (Zi, 2011). Souza et al. (2013b) analysed the parameters influential to biogas and methane productions by LSA. Picard-Weibel et al. (2024) adopted the Morris method, fundamentally a GSA method but sometimes also referred to as a randomised LSA, to identify the parameters influential to the value of the cost function which measures the difference between model outputs and experimental observations. Pastor-Poquet et al. (2019) adopted a variance-based GSA method to evaluate the sensitivity of model parameters and also initial values of degraders and found that the initial concentration of acetate degrader was influential to the value of the cost function. The $\boldsymbol{\theta_{ks}}$ that is commonly regarded as sensitive parameters includes disintegration (k_{dis}) and hydrolysis rates (k_{hyd}), Monod maximum specific uptake rate (k_m), half saturation constant

(K_S) and free ammonia inhibition constant (K_{I,NH_3}) (Barahmand and Samarakoon, 2022).

The goal of model calibration is to find the set of parameters with which the model output assimilates the experimental observation. The similarity between these two – the goodness of fitting – is evaluated by the objective function, J_θ. The traditional model calibration methods aiming to find the set of parameters that minimises J_θ, e.g. Gauss-Newton method and Levenberg–Marquard method (Marquardt, 1963), can be ineffective when J_θ shows many local optimums or flat valleys, which is common when data available for calibration are limited (Guisasola et al., 2006; Pastor-Poquet et al., 2019). Bayesian approach is a promising solution to the calibration of complex process-based model. Besides the likelihood function which measures the similarity between experimental observations and the model output obtained from a certain set of parameters, the prior distribution which expresses the a priori knowledge about the distribution of uncertain parameters is incorporated to construct the posterior distribution. Meanwhile, Bayesian inference provides uncertainty quantification for the calibrated parameters and the model outputs. On the basis of Bayesian inference, different variants of Bayesian algorithms have been adopted in recent studies. Picard-Weibel developed a variational Bayesian uncertainty quantification methodology based on Probably Approximately Correct Bayesian, which derives a posterior distribution that minimises the fitting loss while penalised by the divergence between the posterior and prior distributions and offers bounds on the generalisation error (the error on unseen data). Approximate Bayesian Computation was adopted in the study Pastor-Poquet et al. (2019). This method is typically adopted when the likelihood function is intractable to compute directly. Instead, a metric that compares the similarity between the simulated and observed data is used to screen the data drawn from the prior distribution and to form a posterior distribution.

2.4 Conclusions and Research Gap

The BMP of a material is an important factor to consider in the AD feedstock selection process. Section 2.2.1 introduced different methods for estimating

material BMP, all based on empirical stoichiometric conversion calculations, including approaches using COD content, elemental composition and chemical composition. However, these direct calculation methods often overestimate material BMP due to two main reasons: (1) the degradability of the material, primarily influenced by the content of lignocellulosic substances, and (2) a portion of the energy obtained from organic material degradation is consumed by cell growth and maintenance processes. Besides these two reasons, the actual BMP of a material under real AD operations can also fall short due to the presence of inhibitory substances and factors, such as pH, TAN and C/N ratio of the feedstock, as introduced in Section 2.1.2.

To determine the realistic BMP of a material, considering the influence of these inhibitory factors, a batch AD test called a BMP test, which observes the biogas production of a material under anaerobic condition using a jar test, is usually adopted in the lab. An introduction to the equipment, reactor contents and operational procedures of the BMP test was provided in Section 2.2.2. In addition to evaluating the ultimate biogas yield of a material, the BMP test can also reveal the degradation kinetics of the material. However, it can take 20 to 90 days for the biogas production of a material to reach completion and obtain the BMP test result, making the BMP test non-ideal for material BMP evaluation in the industry. Therefore, many studies have aimed to predict the result of BMP test.

One of the most promising method is by building a regression between the various physiochemical characteristics of the feedstock, such as the elemental composition, chemical composition, inhibitory factors, and potentially operational parameters such as SIR, and the BMP test result (Liu et al., 2023; Xu et al., 2014; Khashaba et al., 2022). Section 2.2.1.3 compiled some of the MLR models from the most cited papers. However, different predictors were used across models based on the relative abundance of components in the specific samples from each study, and the prediction accuracy of these models on samples outside their own studies was generally low. With the increase in computation power, ML models have gained popularity over the past decade. ML models are generally more adept at capturing complex non-linear relationships between predictors and

responses compared to simpler statistical models (Wang et al., 2023). Therefore, some studies have adopted ML algorithm for capturing this regression.

The regression model method for BMP test result prediction has two challenges. First, only the ultimate biogas yield of a material is predicted. Second, for ML models, despite of their advanced capability at multivariate regression, more data are needed to train ML models to convergence, meaning that the model trained must accurately represent patterns in the training data and generalise well to unseen data (Bishop, 2006). Conducting sufficient BMP tests, along with detailed physicochemical analyses for each sample, to train an ML model to convergence can be extremely challenging, if not impossible. As a result, current studies usually collect BMP test results from the literature. However, BMP test results from different studies vary greatly due to different protocols followed, different equipment used and different operational parameters adopted, as discussed in Section 2.2.2. Additionally, specifically to ML models, the mathematical explainability of ML regression models remains a challenge. Although there are some popular methods for ML regression models interpretation, such as PDP and SHAP value, they all focus on feature importance analysis.

This study aimed to develop a method that not only predicts the ultimate biogas yield from the BMP test but also infers reaction kinetics information from various physicochemical characteristics of the feedstock. Furthermore, it sought to address data limitations in training ML regression models for BMP prediction by applying data augmentation techniques, thereby avoiding inconsistencies associated with BMP test data from different studies. Finally, the study aimed to enhance the mathematical interpretability of the ML model, making it a more accessible and transparent tool for domain experts and industrial operators to predict realistic BMP values. **Part I** of this study addressed these, corresponding to *Objective 1*, *Objective 2* and *Objective 3*, as outlined in Chapter 1.

Apart from the BMP of a material, operational parameters such as inflow concentrations, hydraulic retention time (HRT), and organic loading rate (OLR) are also crucial for the performance of a continuous digester. Mathematical modeling provides a cost- and time-efficient way to evaluate digester

performance under various operational scenarios. Section 2.3 introduced both ML-based statistical models and the process-based model, specifically ADM1, for predicting continuous AD performance. A significant challenge in applying ADM1 lies in its calibration. As with many other process-based models, reliably calibrating the numerous parameters involved in the detailed description of sequential reactions can be difficult, especially when observational data is scarce. This study aimed to develop a framework for reliably calibrating ADM1 with a reduced need for extensive experimental observations. The approach begins with identifying the most influential uncertain inputs that require calibration, followed by a Bayesian inference-based method to calibrate the model. This is the *Objective 4*, addressed in **Part II** of this study, as defined in Chapter 1.

PART I Biogas Production Prediction of Batch AD Test

3 BMP Test Result Prediction from Substrate Physiochemical Characteristics

The first study was carried out to validate the BMP test result prediction mechanism proposed – inferring the parameters of the model for biogas production description from feedstock physiochemical characteristics. A dataset of 25 BMP test results for digestate samples with their characteristics analysed were used. The characteristics analysed include TS, VS, acetate, total VFA, TAN, TKN, COD (total COD (tCOD)), sCOD)), alkalinity (total alkalinity (TA), partial alkalinity (PA), intermediate alkalinity (IA)), elemental composition (N, C, H), pH, calorific value (CV), and metal elements (Co, Fe, Mo, Ni). Three empirical biogas production models and a time series model were compared for their abilities at describing the BMP test biogas production. Decision tree (DT) was used to infer the parameters of the selected model from above characteristics. Finally, Gaussian process (GP) was used to provide uncertainty to the biogas production prediction from the inferred biogas production model. Figure 3-1 is the flowchart of the methodology.

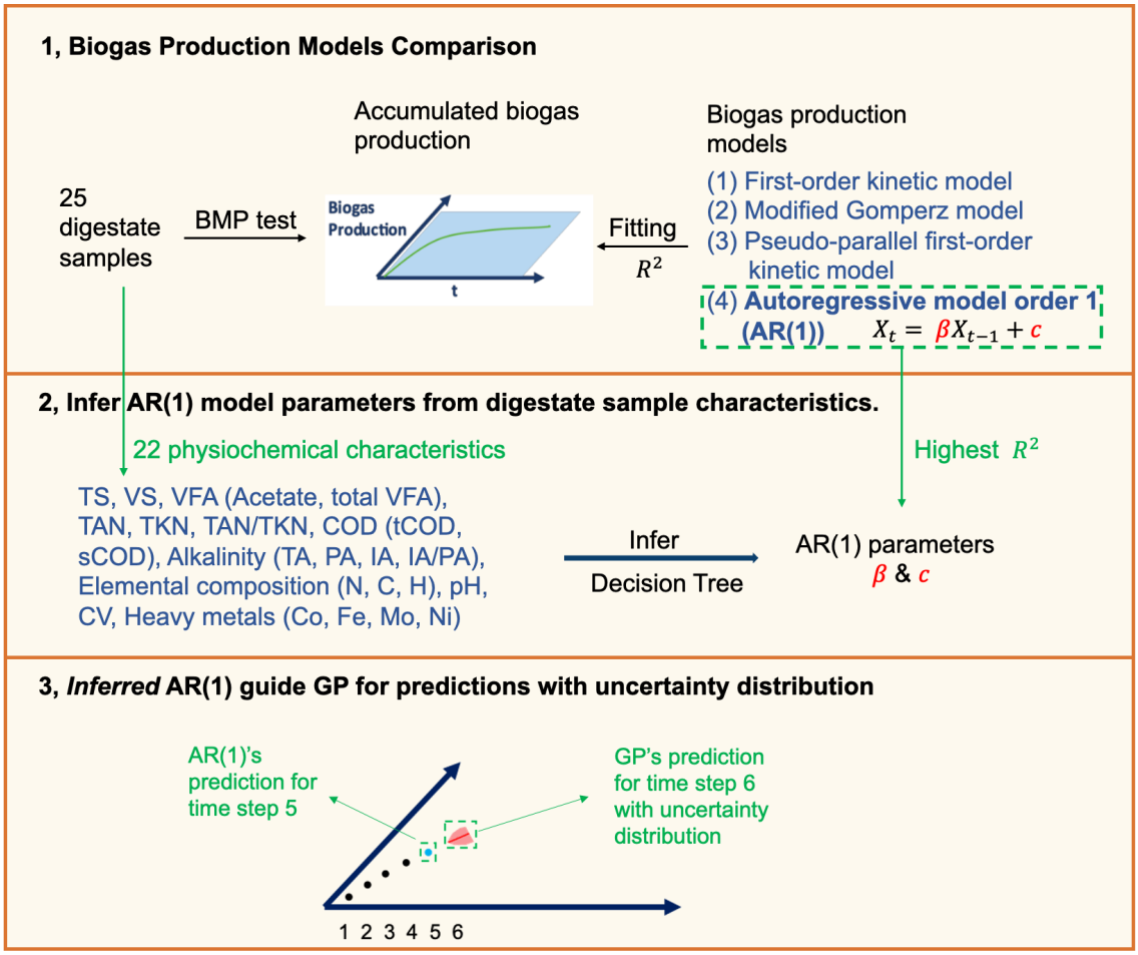


Figure 3-1 The flowchat of the methodology for BMP test result prediction from feedstock physiochemical characteristics.

3.1 Dataset and Methods

3.1.1 Materials and Dataset

The BMP test samples were obtained from the project OMK002-014 of Waste & Respurce Action Programme (WRAP) (2013). The samples for are digestate from the outlets of the final tanks at 25 AD plants in the UK. These samples were anonymised as ADP1 to ADP25. The experiments were run for 28 days. To be noted that the biogas productions of some samples have not reached an end at day 28. The end of a batch AD test is determined by the daily biogas production smaller than 1% of the cumulative biogas production for three consecutive days . Characteristics of 25 digestate samples together with their BMP test results at day 28 are listed in 7.2Appendix C. Detailed specific biogas yield data during 28-

day test period used for biogas production model fitting were reported elsewhere (WRAP, 2013).

3.1.2 Models for Biogas Production

The sampling frequency of the biogas production reduced from multiple times per day in the beginning stage of the experiment to once per day at the end of the experiment. Since the time series model, autoregressive moving average (ARMA) model specifically in this study, requires the observations to be equally spaced. The experimental data were resampled to one datum per day. The abilities of three empirical biogas production models and ARMA model at describing the BMP test biogas production results were compared.

R^2 value was calculated to evaluate the goodness of fit of biogas models to the data following Eq. (3-1. It represents the proportion of the variance in the response explained by the independent variables in a regression model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad \textbf{(3-1)}$$

where n is the number of observations, $\hat{y}_i$ is the predicted response values, and $\bar{y}_i$ is the mean of responses. R^2 value closer to one indicates a better fit of the regression model to the data.

3.1.2.1 Empirical Biogas Production Model Fitting

(1) First-order kinetic model

First-order kinetic model (Eq. (3-2) is derived from the assumption that the substrate degradation is a first-order reaction with hydrolysis as speed-limiting step and the cumulative biogas yield is proportional to the amount of substrate degraded (Eq. (3-3) (Kafle & Chen, 2016; Zahan et al., 2018).

$$y_t = y_m * (1 - e^{-kt}) \quad \textbf{(3-2)}$$

$$\frac{dc}{dt} = -kc \qquad \frac{c_0 - c_t}{c_0} = \frac{y_t}{y_m} \quad \textbf{(3-3)}$$

where:

y_t: cumulative biogas yield at time t;

k: the first-order rate constant;

y_m: is the maximum cumulative gas yield;

c_t: the concentrate of the substrate at time t;

c_0: the initial substrate concentration.

(2) The modified Gompertz model

The modified Gompertz model (Eq. (3-4) is derived from the Gompertz model which is used to describe the microbial activity and has the signature sigmoid shape (Gompertz, 1825, Zhu et al., 2019). It describes the biogas production in terms of the exponential growth rate and the lag phase duration of anaerobic degrading microorganisms (Pramanik et al., 2019).

$$y_t = y_m * exp\left(-exp\left(\frac{R}{y_m} * e * (\lambda - t) + 1\right)\right) \quad \textbf{(3-4)}$$

where:

y_t: cumulative gas production at time t;

y_m: maximum cumulative biogas production;

R: maximum gas production rate;

λ: residence time.

(3) Pseudo-parallel first-order model

Pseudo-parallel first-order model (Eq. (3-5) is considered to be more suitable for describing the biogas yield of the mixture of substrates with different kinetic rates (rapid and slow) (Pererva et al., 2020).

$$y_t = y_m * (1 - P * e^{-k_1 t} - (1 - P) * e^{-k_2 t}) \quad \textbf{(3-5)}$$

where:

y_t: cumulative biogas production at time t;

k_1: the first-order rate constant for readily degradable material;

k_2: the first-order rate constant for less readily degradable material;

P: the proportion of the readily degradable material;

y_m: the maximum cumulative gas yield.

The fitting of the three empirical biogas production models was carried out in MATLAB (2021b). When fitting these nonlinear models in MATLAB, the lowerbound of the parameter y_m to be fitted was set as the biogas observation at day 28. Because some of the samples have not reached the end of reaction at this time, the y_m thus must be equal or bigger than the observation at day 28.

3.1.2.2 ARMA Model Fitting

ARMA(p, q) model is widely used in time series analysis. It consists of the autoregressive part (AR(p)), which is a linear model between the responses of previous p time steps and current time step (Eq. (3-6), and the moving average part (MA(q)), which models the current response as a linear combination of past q error terms (Eq. (3-7).

$$X_t = c + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \cdots + \beta_p X_{t-p} + \varepsilon_t, \qquad \varepsilon_t \sim i.i.d\ N(0, \sigma^2) \qquad \textbf{(3-6)}$$

$$X_t = \mu + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t, \qquad \varepsilon_t \sim i.i.d\ N(0, \sigma^2) \qquad \textbf{(3-7)}$$

ARMA model is used to describe a stationary stochastic process (Box et al., 1994). Table 3-1 organised the characteristics of AR(1) and MA(1) models (without considering the constant term) and the conditions they need to satisfy for being stationary respectively.

Autocorrelation Function (ACF) plot and Partial Autocorrelation (PACF) plot are useful for identifying the order (p, q) for ARMA model. ACF plot depicts the autocorrelation between current time step and the lagged time steps. As described in Table 3-1, the correlation between the current time step and lagged time steps is 0 after $h > 1$ for MA(1) process. Similarly, the $Correlation(X_t, X_{t+h})$ drops to 0 suddenly after $h > q$ for MA(q) process. Therefore, MA(q) process has the feature of ACF plot shutting off after q lags. In contrast, for AR(p) process the ACF plot trails off gradually. For this case, PACF

plot, which measures the correlation between time step X_t and X_{t+h} but with the linear dependence of time steps in between removed, helps to identify the order of AR process (Shumway & Stoffer, 2017). AR(p) process has the feature of PACF plot shutting off after p lags.

Table 3-1 Characteristics of AR(1) and MA(1) models (without considering the constant term).

	AR(1)	MA(1)
Equation	$X_t = \beta_1 X_{t-1} + \varepsilon_t$	$X_t = \theta_1 \epsilon_{t-1} + \varepsilon_t$
Stationary Condition	$\lvert\beta_1\rvert < 1, E(X_0) = 0$	$None$
$E(X_t)$	0	0
$Var(X_t)$	$\frac{\sigma^2}{1-{\beta_1}^2}$	$({\theta_1}^2 + 1)\sigma^2$
$Correlation(X_t, X_{t+h})$	${\beta_1}^h$	$\begin{cases}\theta_1\sigma^2, h = 1 \\ 0, \quad h > 1\end{cases}$

A detailed discussion of the AR(1) process with a constant term ($X_t = c + \beta_1 X_{t-1} + \varepsilon_t$) is provided below. Differing from an AR(1) process without a constant term, the requirement to ensure the stationary of an AR(1) process with a constant term is not on $E(X_0)$ but rather on the constant c.

By keep replacing X_{t-h} by $c + \beta_1 X_{t-h-1} + \epsilon_{t-h}$, the $E(X_t)$ can be represented by Eq. (3-8. As $\lvert\beta_1\rvert < 1$, ${\beta_1}^t \to 0$ as $t \to \infty$, and $E(\varepsilon_{t-i}) = 0$. Therefore, the only term left is $\sum_{i=0}^{t-1} {\beta_1}^i c$. As $t \to \infty$, $\sum_{i=0}^{t-1} {\beta_1}^i$ is a geometric series with common ratio β_1. This geometric series converges to $\frac{1}{1-\beta_1}$. Therefore, $E(X_t) \to \frac{c}{1-\beta_1}$ as $t \to \infty$. This has described the long-run behaviour of an AR(1) process with constant term, which oscillates around $\frac{c}{1-\beta_1}$, because the effect of the initial value $E(X_0)$ diminishes and the process is primarily driven by the autoregressive term (Woodward et al., 2011).

$$E(X_t) = {\beta_1}^t E(X_0) + \sum_{i=0}^{t-1} {\beta_1}^i c + \sum_{i=0}^{t-1} {\beta_1}^i E(\varepsilon_{t-i}) \quad \textbf{(3-8)}$$

It also worth noting that the ε_t term in the AR(p) model differs from the error term in the traditional regression model. In the traditional regression model, the error term is often used to capture factors that are not explicitly modelled but contribute to the variability in the response variable, including measurement error, omitted variables in the model, etc. While for AR(p) model, it assumes that the observed series is a result of both the deterministic autoregressive component and the stochastic error component. The ε_t in AR(p) process is used to capture any factors or influences that are not accounted for by the autoregressive structure of the model. They can include random shock/disturbance to the process, measurement error, etc. The error component is also part of the stochastic process being modelled. As the shock at time step t, ε_t, influences infinitely long future time steps, a smaller disturbance to a time series (small σ^2) leads to a quicker approach towards the long-run asymptote.

For most of the BMP test data, the cumulative biogas production curves showed a rapid increase during the initial three days, then a noticeable reduction in the biogas production rate was observed. Consequently, a single set of ARMA model parameters may yield a poor fit for both the rapid production stage and the following moderate production stage. Therefore, in this study, to simplify the modelling process, the ARMA was only used to model the biogas production after day 4 of the BMP tests. Also, the fluctuation in the biogas production data was small because the biogas data were averaged to one datum per day. Therefore, an AR(p) process with small error term (ε_t), which approaches to the long-run asymptote fast, is expected to describe the time series of BMP test biogas production well. The long-run asymptote of the AR(p) process corresponds to the ultimate biogas yield of the BMP test. The confirmation of the BMP test biogas production being an AR(p) process, the order p of the model and the magnitude of the variance of the distribution of ε_t will be discussed in the Section 3.2.1.2.

The fitting of ARMA model parameters was implemented in MATLAB (2021b). The algorithm used by MATLAB is Burg method.

3.1.3 Decision Tree (DT)

Decision tree (DT) is a non-parametric supervised ML algorithm which can be used for both classification and regression tasks, providing piecewise constant approximation as prediction results. DT intrinsically splits data into different blocks in the high dimensional feature space according to the values of candidate features. Compared to ML algorithms such as artificial neural network, it has the advantage of better interpretability and automatic feature selection (Tan et al., 2014). The predictor importance of a trained a DT can be returned in MATLAB directly. The importance of a feature is determined by how often it is used for selection across all the internal nodes of a trained DT (MathWorks, 2021).

The fitting of DT regression model was implemented in MATLAB(2021b) by 'fitrtree' syntax. The criterion for choosing which feature to use and its value for splitting at internal nodes for DT regression model is MSE (MathWorks, 2021).

DTs were trained using all 22 physicochemical characteristics listed in 7.2Appendix C as predictors to infer the parameters of the biogas production model which describes BMP test biogas production data the best. First, BMP results of 20 digestate samples (out of the 25-sample dataset) were used to train the DT model with the remaining five samples as the test set. There are 53,130 different combinations when choosing 20 samples out of the 25. For practical reasons, 5000 scenarios were selected randomly to evaluate the prediction accuracy when training a DT with 20 samples. Then, the DT model was trained and validated using the leave-one-out (LOO) cross-validation method, which four more data were used for DT training. This is to verify that if the prediction accuracy will improve if more training data are available.

The metric for evaluating the prediction accuracy of the inferred biogas production model parameters and the calculated biogas yield at day 28 is the absolute percentage error (APE) (Eq. (3-9).

$$APE = \left|\frac{y - \hat{y}}{y}\right| \quad \textbf{(3-9)}$$

where:

y: the fitted parameters of the biogas production model or the experimental observation of biogas production at day 28;

$\hat{y}$: the parameters of the biogas production model inferred from physiochemical characteristics of substrate by DT or the biogas production at day 28 calculated with the inferred biogas model.

3.1.4 Gaussian Process (GP)

Gaussian process (GP) is a generalisation of the multivariate Gaussian distribution of infinitely many variables. As a multivariate Gaussian distribution is fully specified by a mean vector and covariance matrix, a GP is fully specified by a mean function, $m(x)$, and a covariance function, $k(x, x')$. The mean function is the assumption about the underlying data generating process. It can be constant mean (e.g., zero mean), linear mean, or complex functions such as sinusoid. The covariance function determines the strength of correlation between data points. The commonly adopted covariance function including squared exponential kernel (Eq. (3-10) and linear kernel (Eq. (3-11). Squared exponential kernel is also called radial basis function (RBF) kernel, the covariance between two points of which is dependent on their relative distance. σ is the vertical scale, determines how far the function deviate from the mean in average. l is the length scale, defines how quickly the correlations between points fall away and thus how wiggly the functions are (Rasmussen & Williams, 2006).

$$k(x, x') = \sigma^2 \exp\left(-\frac{\|x - x'\|}{2l^2}\right) \quad \textbf{(3-10)}$$

The two hyperparameters σ and l can be learnt from the training data by maximum likelihood method (Rasmussen & Williams, 2006).

Different from RBF kernel, for linear kernel, the correlation depends on the absolute distance of points. c in Eq. (3-11 determines the point that all the functions pass through and ${\sigma_b}^2$ is similar to the vertical scale σ^2 in RBF kernel (Duvenaud, 2014).

$$k(x, x') = {\sigma_b}^2 + {\sigma_v}^2 (x - c)(x' - c) \quad \textbf{(3-11)}$$

Unlike the traditional parametric model, the trajectory of which is fixed when the model parameters are fixed. For a GP, even if the hyperparameters of the kernel for GP is fixed, the realisations of functions of the posterior distribution $P(Y|X)$, where X is the training data and Y is the test data, is only constrained after training data are provided (Rasmussen & Williams, 2006). Furthermore, GP excels at interpolation but may not effectively extrapolate beyond the vicinity of the training data points. This limitation arises from the fact that the distribution of posterior functions at location far away from the training data points rapidly converges to the mean of the underlying data generation process.

From above reasons, GP was not used to predict the biogas production from the feedstock characteristics directly. Instead, in this chapter, the inferred biogas production model was used to guide GP. The biogas model provides the underlying biogas production trajectory for GP, and GP was used to make prediction for biogas production with uncertainty. In detail, this was carried out as follow: GP predicted one time step ahead each time by fitting one more datum provided by the biogas production model and the prediction uncertainty was depicted. Two kernels, RBF kernel alone and a combination of the linear and RBF kernels, were compared.

3.2 Results

3.2.1 BMP Test Results and Fitting

At the end of the 28-day test, the biogas predictions of some samples have reached a plateau and some were still showing the trend of increase. This is influenced by the content of the readily degradable materials in the digestate. According to the criteria for determining whether a biogas production process has reached an end as described in Section 1.1.1, the 25 biogas production time series were divided into two types. Type I: the biogas production reached the end within 28 days (ADP 3, 4, 5, 9, 11, 13, 16, 19, 20, 21). Type II: the biogas production was still increasing at day 28.

For each individual physicochemical parameter, the samples with significant high or low level of it were recorded, but it was always a mix of Type I and Type II

samples. For example, the alkalinity contents of ADP 5, 8 and 9 were significant high and the VFAs of the ADP 11, 13 and 18 were significant high. In other words, no single physicochemical characteristic could account for the different behaviours of two types of biogas production fully. This also indicates that the kinetics of the biogas production process is jointly controlled by many characteristics of the substrate.

For the conciseness of the figures and tables, four samples were chosen for each type of the biogas production curve to compare the abilities of three empirical biogas production models and ARMA model for describing the biogas production process. They are respectively ADP 4, 5, 16, 19 for Type I, and ADP 2, 14, 18, 20 for Type II.

3.2.1.1 Empirical Biogas Production Models

Table 3-2 shows the $\boldsymbol{R^2}$ values of the fitting of three empirical biogas production models to BMP test data. Overall, the modified Gompertz model achieved lower $\boldsymbol{R^2}$ values across most of the samples, indicating poorer fitting performance. Modified Gompertz model is signatured with the S-shape. It is better at describing the biogas production process with a slow kinetics at the beginning stage due to the acclimatisation of the microorganisms to the AD environment (Kafle & Chen, 2016). However, no digestate samples exhibited obvious lag phase in the beginning of the BMP tests in this study. In contrast, the pseudo-parallel first-order model described the biogas productions of almost all samples with $\boldsymbol{R^2}$ values between 97–99%.

The first-order kinetic model described the Type I biogas production almost as well as the pseudo-parallel first-order model, whereas it performed worse than the pseudo-parallel first-order model on Type II samples. This is because pseudo-parallel first-order model is more suitable for describing the biogas production of substrates mixed with materials of different reaction rates.

Table 3-2 $\boldsymbol{R^2}$ values of the fittings of three empirical models to Type I and Type II BMP test biogas production time series. FO: first-order kinetic model. MG:

modified Gompertz model. PP: pseudo-parallel first-order model. (The R^2 values larger than 97% are in bold).

Type I	ADP	3	4	5	9	11	13	16	19	20	21
	FO	**98.5**	**98.4**	**97.9**	**97.3**	86.8	**98.2**	**99.5**	**97.7**	96.4	**97.2**
	MG	**97.6**	94.0	96.6	**99.1**	80.7	91.8	96.0	94.9	88.3	61.6
	PP	**98.5**	**98.7**	**97.9**	**97.4**	94.6	**99.5**	**99.6**	**97.8**	**98.7**	**97.6**
Type II	ADP	1	2	6	7	8	10	12	14	15	17
	FO	**97.1**	**98.6**	86.5	77.2	83.2	**98.4**	**99.0**	**97.6**	**97.5**	**97.5**
	MG	88.8	93.9	78.6	85.8	68.2	93.0	96.1	90.8	74.9	88.4
	PP	**99.9**	**99.6**	92.8	92.6	**98.0**	**98.9**	**99.1**	**99.5**	**98.9**	**97.8**
	ADP	18	22	23	24	25					
	FO	95.5	96.0	**98.8**	**99.1**	**99.5**					
	MG	85.6	85.9	96.7	**97.3**	**99.0**					
	PP	**99.1**	**98.9**	**98.8**	**99.1**	**99.5**					

However, it is worth noting that there are usually multiple sets of optimal solutions for the estimates of the parameters of the pseudo-parallel first-order model ($\boldsymbol{y_m}$, $\boldsymbol{P}$, $\boldsymbol{k_1}$ and $\boldsymbol{k_2}$). This is because that the nonlinear least-square error function when fitting a particular set of biogas production data to this model is not always convex. Therefore, though pseudo-parallel first-order model can fit the biogas production data well, its fitted parameters are less informative. In addition, when fitting the pseudo-parallel first-order model in MATLAB by least-square algorithm, the initial value of the parameter $\boldsymbol{P}$ should avoid being set to 0.5 and $\boldsymbol{k_1}$, $\boldsymbol{k_2}$ should not be set to the same value. Otherwise, the partial derivatives of the error function with respect to $\boldsymbol{k_1}$ and $\boldsymbol{k_2}$ are the same, which means the descending directions on these two dimensions are the same and the nonlinear search for the minimum value of the error function value will settle at a local minimum point.

Take ADP 16 and ADP 18 as examples for Type I and Type II data. The fitting of three empirical models to the biogas production data is shown in Figure 3-2.

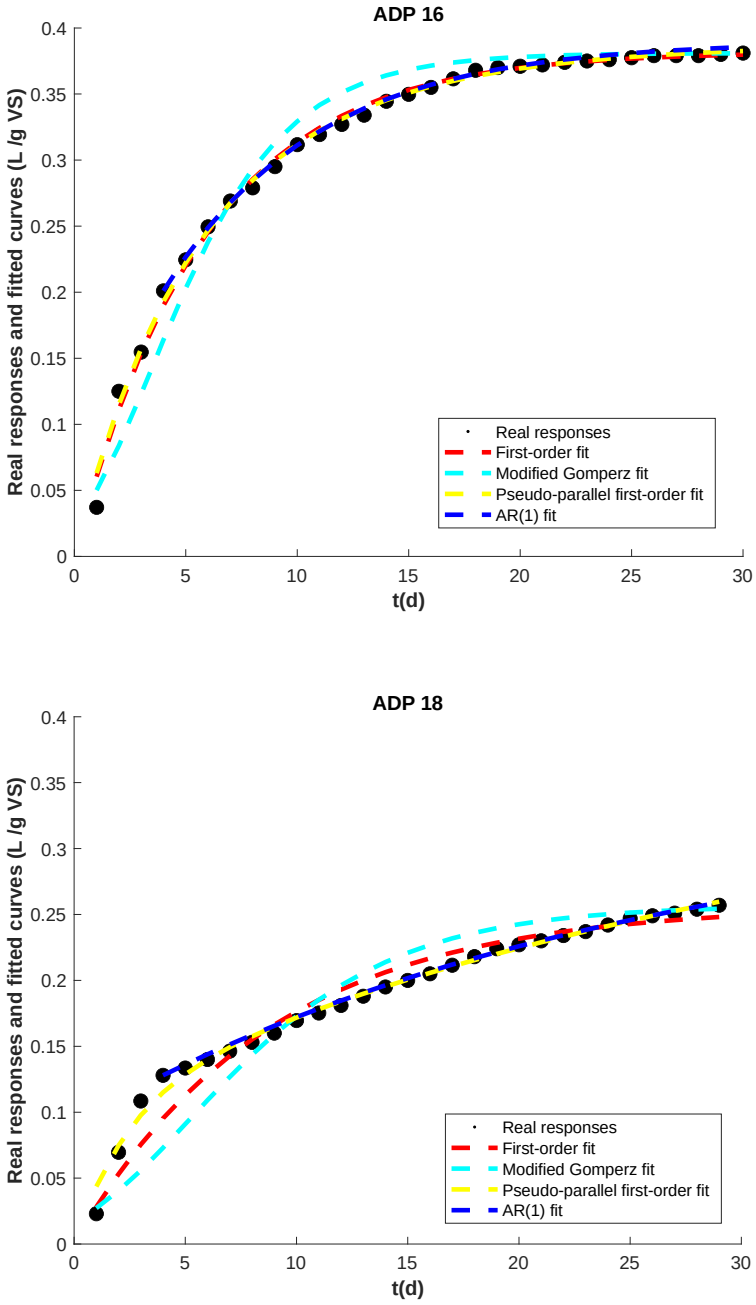


Figure 3-2 The fittings of three empirical biogas production models and the AR(1) model to the Type I (ADP 16 as an example) and Type II (ADP 18 as an example) biogas production data.

3.2.1.2 ARMA Model Fitting

The fitting of the ARMA model to the BMP test biogas production data started from day 4. ACF and PACF plots were used to identify the order of the ARMA model. Take the sample ADP 20 as an example, its ACF and PACF plots are

shown in Figure 3-3. Since the ACF plot trails off and the PACF plot cuts off after first lag, the time series of biogas production starting from day 4 should be an AR(1) process.

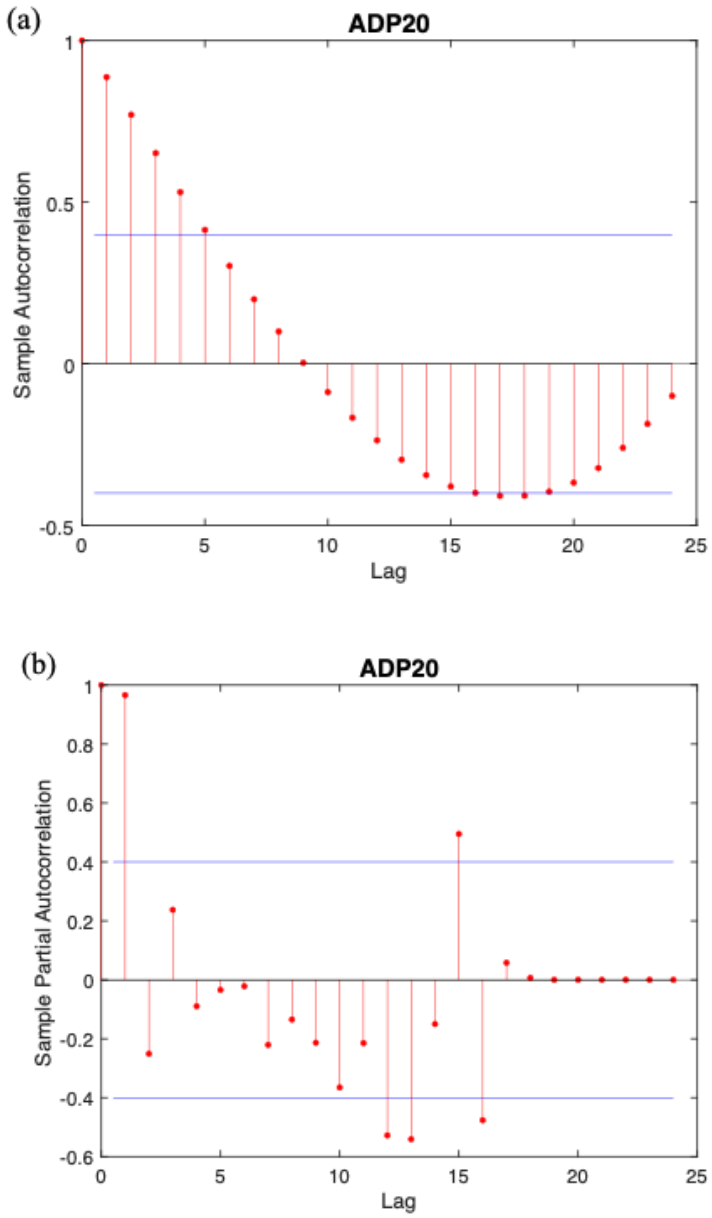


Figure 3-3 The ACF (a) and PACF (b) plots of sample ADP20 starting from day 4. The blue lines are 95% confidence bounds for significant correlation.

The fitting of the AR(**1**) model to the biogas production data after the initial three days was comparable to the pseudo-parallel first-order model, with $\boldsymbol{R^2}$ values higher than 99% for the majority of samples (Table **3-3**), regardless of Type I or

Type II data. Also, take the fitting of AR(**1**) model to ADP 20 as an example, the estimated $\boldsymbol{\varepsilon_t}$ of the model is 0.004. This has verified the speculation about the magnitude of $\boldsymbol{\varepsilon_t}$ made at Section 3.1.2.2, which is the reason why the fitted AR(**1**) model reached the long-run asymptote fast. The fitting of the AR(1) model to the illustrative Type I and Type II biogas production data, ADP16 and ADP18 respectively, starting from day 4 is displayed in Figure 3-2.

Table 3-3 R^2 values of the fitting of AR(1) to the Type I and Type II BMP test biogas production time series (The R^2 values larger than 97% are in bold).

Type I	ADP	3	4	5	9	11	13	16	19	20	21
	AR(1)	**98.2**	**99.4**	**97.4**	**99.1**	**97.0**	**99.3**	**99.8**	**99.7**	**98.9**	90.1
Type II	ADP	1	2	6	7	8	10	12	14	15	17
	AR(1)	**99.2**	**99.8**	85.4	80.9	95.9	**99.5**	98.6	**99.7**	**98.9**	**99.1**
	ADP	18	22	23	24	25					
	AR(1)	**99.6**	**99.5**	**98.7**	**99.1**	**99.9**					

Considering the fact that AR(1) model described the biogas production data (after the initial rapid biogas production period) equally well as the pseudo-parallel first-order model but with only two parameters, and also the pseudo-parallel first-order model suffers from the multiple local-optima problem, the AR(1) model was chosen in the following study to infer the biogas production model parameters from the physiochemical characteristics of the substrates using ML methods.

It is worth noting that unlike the empirical biogas production models the curvatures of which are fixed when the model parameters are fixed, the curvature of the AR(1) model depends not only on the model parameters but also on the response value at the initial time step, in this case, the observation at day 4. In other words, to obtain the BMP test result of a digestate sample, there is a potential to reduce the time-consuming 28-day test to only 3 days when combining with the physiochemical characteristics of the substrate with the utilisation of a biogas production model.

3.2.2 DT Training

When 20 samples were used for DT training with five samples left as test set, the average APEs of the inferred AR(1) model coefficient and constant of the randomly selected 5000 train-test-split scenarios were 4.760% and 72.041%, respectively. The average APE of the calculated biogas production at day 28 was 52.134%.

When training a DT by LOO cross-validation method with four more samples in the training set, the average APEs of the predictions for the AR(1) model coefficient and constant were 4.187% and 48.443%, respectively. The average APE of the calculated BMP test result at day 28 was 40.003% (Table 3-4). With four more data provided, the prediction accuracy of the BMP test result improved 12.131%. Additionally, among 25 LOO cross-validation groups, the APEs of the predicted BMP test results of five groups (ADP 2, 4, 9, 20 and 23) were smaller than 5%.

Take a closer look at the relationship between the prediction accuracies of the AR(1) model parameters and the BMP test result (y(28)). In Table 3-4, the percentage errors of the AR(1) model coefficient prediction with absolute value smaller than 2% and AR(1) model constant prediction with absolute value smaller than 50% were marked bold. Except for ADP 9, all the test groups with APE of the predicted BMP test result smaller than 5% have small percentage errors in the predictions of both the AR(1) coefficient and constant.

However, it is observed that there are groups with high APEs in the prediction of y(28), although the prediction errors of both AR(1) coefficient and constant are small (ADP 1, 10 and 13). One reason for this is that for ADP 1 and ADP13, the errors in the predictions of the AR(1) coefficient and constant were biased towards the same direction. In contrast to biased in different signs (ADP 2, 4, 20 and 23), this caused a bigger y(28) prediction error. More importantly, the R^2 scores of the predicted AR(1) coefficient and constant respectively when training a DT by LOO algorithm with 24 samples were only - 0.372 and 0.359. This means that the trained DT performed no better than just using the sample mean for the prediction of the AR(1) coefficient.

Table 3-4 The APEs for the predicted AR(1) model coefficient and constant and the APEs of the calculated biogas yield at day 28 (y(28)) when training a DT by LOO cross-validation.

Test Sample No.	AR(1) coefficient			AR(1) constant			BMP result (y(28))		
	Fitted	Predicted	Percentage Error	Fitted	Predicted	Percentage Error	Observed	Predicted	APE
ADP1	0.929	0.942	**-1.329**	0.010	0.014	**-31.820**	0.132	0.193	46.499
ADP2	0.953	0.934	**1.979**	0.011	0.014	**-31.290**	0.182	0.189	**3.966**
ADP3	0.792	0.884	-11.605	0.028	0.021	**25.867**	0.132	0.172	30.565
ADP4	0.881	0.896	**-1.717**	0.029	0.024	**15.408**	0.236	0.228	**3.261**
ADP5	0.838	0.850	**-1.440**	0.019	0.043	-125.051	0.117	0.283	142.091
ADP6	0.933	0.826	11.478	0.005	0.006	**-29.754**	0.062	0.034	45.935
ADP7	0.961	0.893	6.978	0.003	0.006	-100.604	0.066	0.057	13.516
ADP8	0.932	0.939	**-0.806**	0.019	0.045	-139.629	0.261	0.617	136.746
ADP9	0.851	0.893	-5.016	0.055	0.040	**26.758**	0.361	0.363	**0.528**
ADP10	0.902	0.894	**0.914**	0.018	0.023	**-28.397**	0.173	0.206	18.836
ADP11	0.885	0.937	-5.864	0.035	0.024	**29.643**	0.301	0.355	18.018
ADP12	0.848	0.943	-11.170	0.026	0.025	**4.343**	0.171	0.349	104.313
ADP13	0.907	0.892	**1.625**	0.016	0.014	**15.817**	0.164	0.124	24.684
ADP14	0.940	0.942	**-0.219**	0.009	0.014	-55.121	0.129	0.198	53.536
ADP15	0.910	0.847	6.977	0.008	0.022	-172.781	0.086	0.142	64.677

ADP16	0.863	0.842	2.450	0.053	0.041	**22.503**	0.379	0.258	31.938
ADP17	0.930	0.945	**-1.585**	0.020	0.010	51.347	0.257	0.164	36.251
ADP18	0.966	0.857	11.322	0.012	0.031	-164.016	0.254	0.217	14.499
ADP19	0.859	0.898	-4.486	0.020	0.023	**-15.435**	0.139	0.212	52.519
ADP20	0.898	0.894	**0.416**	0.020	0.022	**-7.908**	0.194	0.202	**4.314**
ADP21	0.861	0.885	-2.814	0.005	0.005	**-11.881**	0.033	0.044	33.863
ADP22	0.918	0.949	-3.463	0.019	0.022	**-15.398**	0.222	0.351	58.322
ADP23	0.850	0.846	**0.498**	0.043	0.043	**-2.072**	0.286	0.278	**2.526**
ADP24	0.872	0.896	-2.718	0.036	0.045	**-25.019**	0.280	0.410	46.585
ADP25	0.899	0.847	5.807	0.031	0.050	-63.220	0.287	0.321	12.078
MAPE			4.187%			48.443%			40.003%

In conclusion, there is no doubt that test samples with accurate predictions of both parameters of the AR(1) model also have accurate prediction on the biogas production. However, the size of the samples available in this chapter in no way enabled to train a DT that was able to provide a stable and precise inference of the AR(1) model parameters from the characteristics of the feedstock. Despite this, from the comparison of the average APEs of the inferred AR(1) model parameters and the BMP test result prediction when training the DTs with 20 and 24 samples respectively, there is a reason to believe that the inference ability of DT will improve with more training data provided.

The overall predictor importance of 25 trained DTs is shown in Table 3-5. Both TAN and TKN were analysed as important features for AR(1) model coefficient inference. TAN and TKN are highly correlated (Pearson correlation coefficient (r) = 0.979). By the proof of Gupta (2021), when two features are highly correlated, tree predictor tends to arbitrarily choose one of the correlated features for splitting, and this leads to both features being screened out as important features. Consider from a biological point of view, TAN is a key substance for the growing of microorganisms such as ammonia-oxidizing bacteria and archaea. These microorganisms play important roles in the breakdown of organic matter and the production of methane. However, high concentrations of TAN can also inhibit the activity of methanogenic microorganisms, leading to process instability and reduced methane production. TKN indicates overall nitrogen content and nutrient availability in the substrate (Jiang et al., 2019b). Therefore, the above feature importance conclusions are substantiated. VS, tCOD and sCOD were also screened out as important features. They all provide insight into the amount of organic compound in the substrate, which are correlated with the biogas yield in the AD. TA indicates the overall buffering capacity of the digestate sample. It was also identified as an important predictor.

Total-VFA was not an important predictor for AR(1) coefficient prediction. This is because, although VFA accumulation is an indicator of unstable AD operation and inhibited methane production caused by LCFA or free ammonia accumulation (Ma et al., 2015; Shi et al., 2017), the VFA content measured in

this study is for the feedstock of the BMP test. The VFA in the feedstock acted as readily degradable components and positively contributed to the biogas yield. As observed in 7.2Appendix C, the three samples with the highest total-VFA content, ADP 9, 11 and 16, also had the highest biogas yield at day 28. Despite this, total-VFA might not be efficient at partitioning other samples with medium level total-VFA concentrations. The feature importance result for AR(**1**) constant inference showed that TAN/TKN, TA and sCOD were very important features.

Finally, trace element Ni is important for AR(1) coefficient inference and Mo is extremely important for AR(1) constant inference. As discovered by Jiang et al. (2017), Ni shows inhibitory effect on VFA degradation while Mo can improve propionic acid degradation.

Table 3-5 The overall predictor importance of 25 DTs for inferring AR(1) coefficient and constant from physiochemical characteristics.

	Features							
AR(1) Coefficient Prediction	TS	VS	Acetate	Total-VFA	TAN	TKN	TAN/TKN	TA
	0	**0.162**	0	0	**0.220**	**0.388**	0	**0.027**
	PA	IA	IA/PA	N	C	H	pH	tCOD
	0	0	0.006	0	0	0	0	**0.018**
	sCOD	CV	Co	Fe	Mo	Ni		
	0.023	0	0.018	0.009	0	**0.129**		
AR(1) Constant Prediction	TS	VS	Acetate	Total-VFA	TAN	TKN	TAN/TKN	TA
	0	0	0	0	0	0	**0.115**	**0.023**
	PA	IA	IA/PA	N	C	H	pH	tCOD
	0	0	0.016	0.008	0	0	0	0
	sCOD	CV	Co	Fe	Mo	Ni		
	0.178	0	0	0	**0.657**	0.003		

3.2.3 Prediction Uncertainty Provided by GP

When training the DT with LOO cross-validation and ADP23 as the test set, the APE of the predicted biogas yield on day 28 was 2.526%. This case was used as an example for illustration in this section.

By accepting one more input each time from the inferred AR(1) model together with the experimental observations of the first three days, the predictions of one timestep further by GP with uncertainty provided are demonstrated in Figure 3-4. In general, given that GP only predicted one more timestep at a time on the basis of AR(1) model's prediction, the uncertainty of the prediction of GP was rather small. Despite this, it can be observed that the 99% confidence interval of the prediction of GP with a combination of RBF and linear kernel had a wider band. All of the experimental observations lied within its 99% confidence interval.

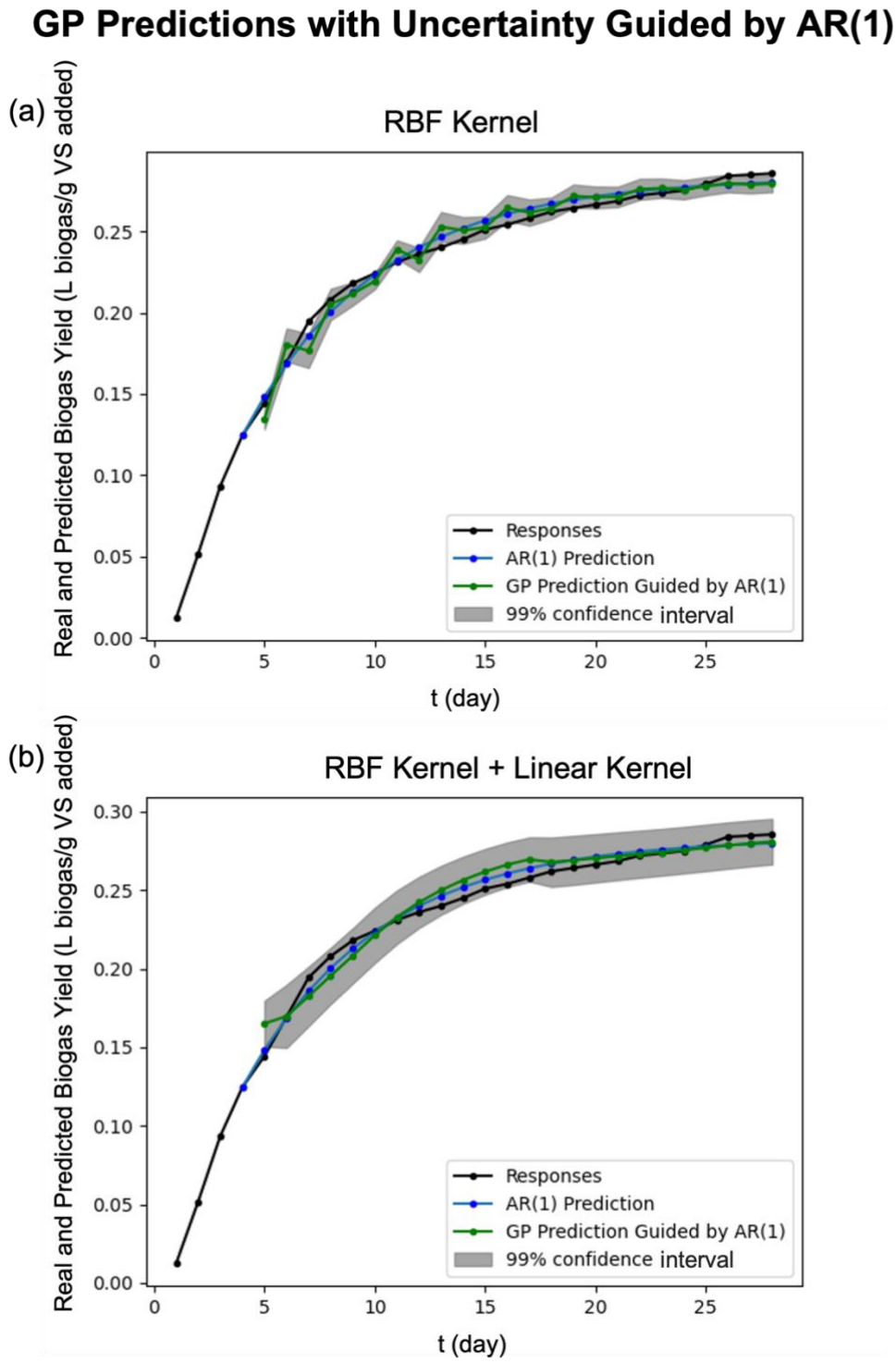


Figure 3-4 GP predictions guided by the prediction of the inferred AR (1) model with uncertainty information provided (a) RBF kernel. (b) RBF kernel + linear kernel.

3.3 Summary and Discussion

As suggested by the British Standards Institution's Publicly Available Specification 110, a digestate sample with BMP test result smaller than 0.25 L/g VS is considered stable and could be applied as soil improver and fertiliser.

The aim of the study in this chapter is to verify the feasibility of the mechanism outlined in Section 1.1.1 aiming to reduce the time required to obtain the biogas production result of a batch AD test by exploring the information provided by the physiochemical characteristics of the feedstock. Specifically, this method uses multivariate regression ML algorithm to infer the parameters of a mathematical biogas production model from the physiochemical characteristics of a digestate sample. Then, the biogas production trajectory is calculated from this inferred mathematical model.

Three empirical biogas production models and a time series model – ARMA – were compared for their abilities at describing the BMP test biogas production data. ARMA model, specifically AR(1) process, was able to fit the biogas production after initial rapid growth stage (the first three days in this study) of the majority of the samples with R^2 values higher than 99% with only two parameters. Therefore, it was chosen as the model the parameters of which were to be inferred from the physiochemical characteristics.

DT was used for multivariate regression considering the very limited training samples available. DT used 22 characteristics of the digestate samples to infer the coefficient and constant of the AR(1) model. The average APEs of the predicted AR(1) coefficient, AR(1) constant and the calculated BMP test result when using 24 samples for training (LOO algorithm for training and test samples split) and 20 samples for training were compared. It was proved that with four more samples provided for training, the predictions of the AR(1) model parameters improved greatly and the average APE of the predicted BMP test result decreased from 52.123% to 40.003%. The BMP result predictions of five samples had APE smaller than 5% when training with 24 samples.

With this method, the time to obtain the BMP test result has a potential to be reduced to four days compared with 28 days when physiochemical characteristics information is incorporated. Despite this, the very limited amount of training samples did not enable the DT to provide a stable and precise inference of AR(1) model parameters. For the same reason, the feature importance output from the trained DT model lacked reliability. This thus provided limited guidance on what are the crucial physiochemical characteristics to be tested for a sample when using this proposed multivariate regression method to predict BMP test result. The analyses of 22 characteristics for each sample as did in this study can be very onerous. However, it did provide some preliminary insights into the significance of certain characteristics. Notably, variables such as VS, TAN (and/or TKN), tCOD (and/or sCOD), TA, Mo, and Ni emerged as crucial features in the inference process of DT.

Finally, this study demonstrated a way to provide uncertainty information to the prediction of an inferred deterministic biogas production model by GP.

All in all, the proposed mechanism of predicting the biogas production of a batch AD test by inferring the parameters of a low-dimensional mathematical biogas production model from the high-dimensional feedstock physiochemical characteristics was validated.

The main limitation of this chapter is the lack of training data. The data availability problem is commonly encountered in the studies for batch AD test result prediction by ML models as it is time consuming to carry out BMP tests. With limited training data used, the inference accuracy of the trained DT was unsatisfactory, and the feature importance result returned was not reliable either.

In the next chapter, a method aims to circumvent the problem of the lack of training data will be proposed.

4 Overcome Data Limitation by Generating Additional Dataset via Response Surface Design

To overcome the data limitation problem when training a ML model for BMP test result prediction, a data augmentation method via response surface design was proposed. 10 BMP tests of mixtures among three food samples with diverse chemical compositions guided by mixture design were carried out. First-order kinetic model was used to fit the biogas production. Response surfaces of the fitted parameters of the FO kinetic model were then derived. 1000 additional data were generated from the response surfaces. XGBoost was used for mapping from physiochemical characteristics space to first-order kinetic model parameters space. XGBoost was chosen for regression due to its robust performance and efficiency compared to other machine learning algorithms such as Random Forests and Support Vector Machines (Natekin & Koll, 2013). After the feature importance analysis by SHAP value method, the features used to predict the first-order kinetic model parameters by XGBoost model were reduced to three. The framework of the methodology is illustrated in Figure 4-1.

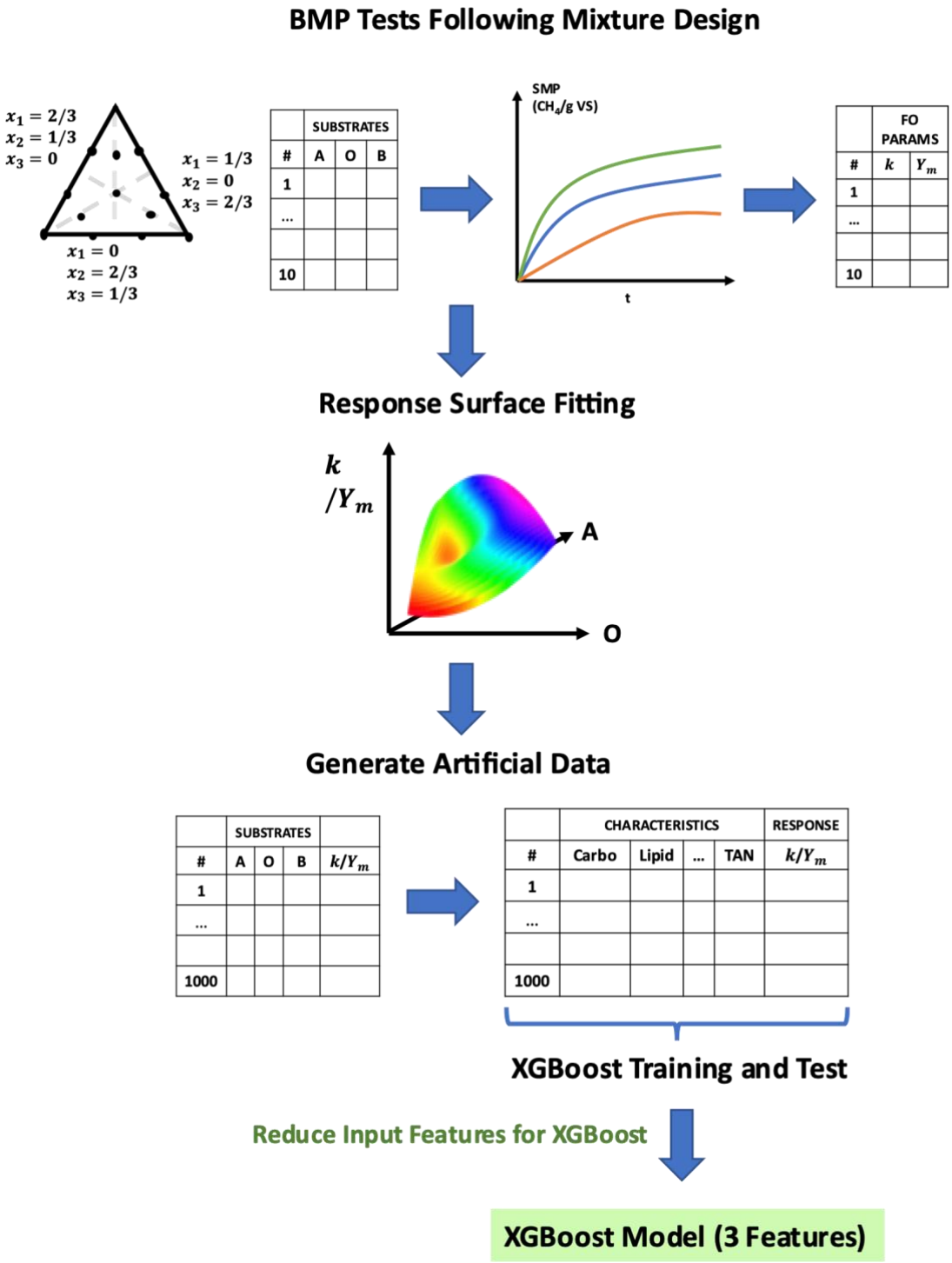


Figure 4-1 Research framework of data augmentation via mixture design for training ML regression model for BMP result prediction.

4.1 Dataset and Methods

4.1.1 Dataset Generation

4.1.1.1 Materials

Three food samples were used for BMP tests of co-digestion. They are respectively: banana peel and green peas (weight in weight (w/w) 1:1) (A), banana pulp and potato (w/w 1:1) (O) and 20% fat minced beef (B). The bananas used were Sainsbury's fairtrade banana, the green peas used were 'Birds Eye' garden peas, and the beef used was Sainsbury's British or Irish 20% fat beef mince. The characteristics of each substrate are listed in Table 4-1. See 7.2Appendix B for the detailed analysis methods of physiochemical characteristics. The inoculum was collected from a digester treating THP pre-treated sewage sludge at Cotton Valley Anglian Wastewater Treatment Plant (Milton Keynes, UK). The VS of the inoculum is 2.805% and the density is 1g/ml.

Table 4-1 Characteristics of substrate (A) banana peel and green peas (w/w 1:1), (O) banana pulp and potato (w/w 1:1) and (B) 20% fat minced beef.

	Unit	A	O	B
density	g/ml	1	1	1
TS	% of wet sample	14.600	19.000	41.200
Ash	% of wet sample	0.800	0.800	0.800
VS	% of wet sample	13.800	18.200	40.400
Lignin	% of VS	0.343	0.140	0
Cellulose+Hemicellulose	% of VS	30.090	4.255	0
Non-fibre Carbohydrate	% of VS	44.203	87.912	0
Protein	% of VS	21.667	7.253	44.802
Lipid	% of VS	3.623	0.549	55.446
Alkalinity	(mg CaCO3/g sample)	2.650	1.940	4.560
TAN	% of wet sample	0.010	0.003	0.010

tCOD	% of wet sample	13.565	20.363	41.125
sCOD	% of wet sample	5.592	10.357	3.198
N	% of dry mass	3.173	1.110	6.987
C	% of dry mass	44.840	42.420	64.473
H	% of dry mass	6.861	7.528	8.283

4.1.1.2 BMP Tests Following the Mixture Design

An exploration of the relationship between variables and the response can be achieved by Response Surface Design. A response surface equation can be fitted to the responses of the experiments carried out following response surface design, which helps to visualise that how the change of variables affects the response. The Mixture Design is a special case of Response Surface Design. Compared to other response surface design methods such as Central Composite Design and Box-Behnken Design, Mixture Design has the signature that each design point represents the proportions of each component in a mixture rather than their real magnitudes. The proportions of all the components for every design point add up to one (Montgomery, 2020).

Two usually used methods for Mixture Design among three components are ‘Simplex Lattice Design’ and ‘Simplex Centroid Design’. The latter can be augmented with axial points, which help examine the response level inside the response surface. This is called ‘Augmented Simplex Centroid Design (ASCD)’. Figure 4-2 shows the standard ASCD with 13 experimental points in total and the 10 points chosen for carrying out BMP tests in this study.

The BMP test was carried out using the BMP equipment purchased from CJC LABS (UK), this equipment was equipped with water bath, reactors, gas counters, a sensor for monitoring temperature and pressure, and a data logging and processing system. The reactor has a volume of 763.410 mL (diameter: 90 mm, height 120 mm). The gas production was measured by gas flow meters. When the bucket in the gas counter was fully filled with biogas collected, the bucket

tipped over and one count was recorded by the data analysing system. This system recorded the time when the counts happened, converted the biogas production to STP conditions (273.15 K, 101.33 kPa) and provided the user with accumulated biogas production.

The operation of BMP tests followed the guideline published by Holliger et al. (2016). The temperature of reactors was controlled at 37 °C. Before sealing the reactor, the headspace was flushed with N_2 at a flow rate of 10L/min for 30s. No additional nutrient solution was added. The experiment was stopped when the daily biogas production smaller than 1% of the cumulative biogas production for three consecutive days (Zaman, 2010). Biogas was sampled for composition analysis with GC to determine the respective percentages of CH_4 and CO_2 twice per day in the beginning stage of the reaction and the sampling frequency was gradually reduced with slowing down of the reaction.

The total volume of the substrate and inoculum of the experimental groups was controlled at 215 mL, with the SIR (VS/VS) controlled at 0.5. Each experiment was carried out in duplicate. According to the VS contents of three substrates and inoculum, the weights of substrates and inoculum for the 10 experimental groups were displayed in Table 4-2. Two reactors with only 215 mL inoculum were included as blank controls.

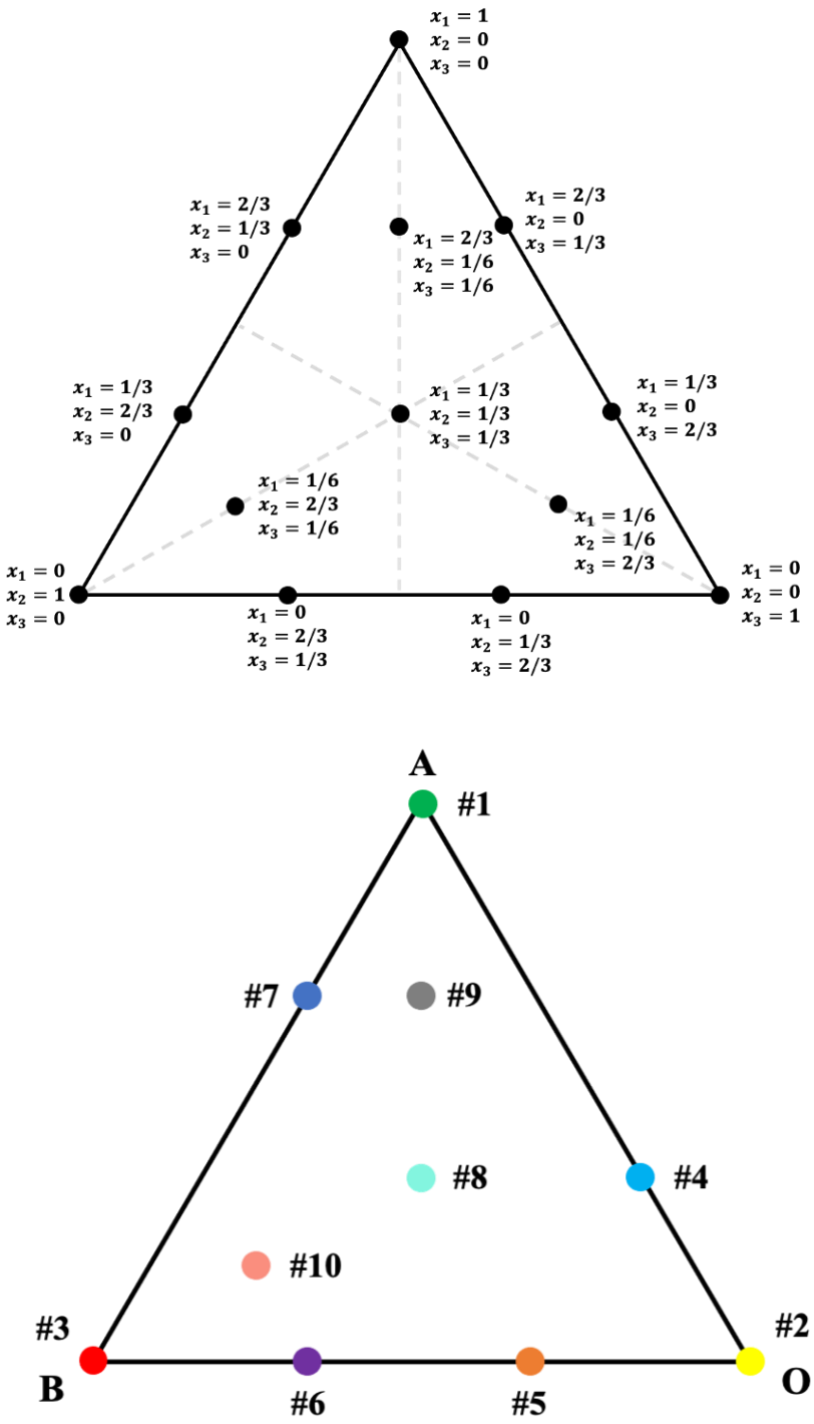


Figure 4-2 Standard ASCD with 13 points. x_1, x_2 and x_3 represent fractions of three components in a mixture. And the 10 points selected in this study for carrying out BMP tests with assigned experimental group numbers.

Table 4-2 Substrates and inoculum volumes for 10 experimental groups and the control group.

Group	Substrate Fractions			Substrate Volumes (mL)			Inoculum Volume (mL)
	A	O	B	A	O	B	
1	1	0	0	20.703	0.000	0.000	194.297
2	0	1	0	0.000	16.073	0.000	198.927
3	0	0	1	0.000	0.000	7.513	207.487
4	1/3	2/3	0	5.789	11.579	0.000	197.632
5	0	2/3	1/3	0.000	7.626	3.813	203.562
6	0	1/3	2/3	0.000	3.045	6.090	205.866
7	2/3	0	1/3	8.707	0.000	4.353	201.940
8	1/3	1/3	1/3	4.105	4.105	4.105	202.686
9	2/3	1/6	1/6	10.113	2.528	2.528	199.831
10	1/6	1/6	2/3	1.527	1.527	6.108	205.838
Control	/	/	/	0	0	0	215

A MCC control experiment was carried out to verify the overall integrity of the BMP test system, including aspects such as the seal quality, the precision of gas measurements, and operating errors. The VS of the MCC (Fisher Scientific, UK) is 98.845% of total weight. 5.954 g MCC and 394.046 g inoculum, in total 400 g mixture, with an SIR = 0.5 were added to the reactor. This was carried out in duplicate.

4.1.1.3 Data Processing and Biogas Production Model Fitting

The original accumulated biogas production data recorded by the data logging system of BMP equipment were processed in the following steps:

(1) The recorded real time of each count was converted to float number time in day starting from 0.

(2) The recorded accumulated biogas production (mL) was processed to three data per day by taking average, corresponding to timesteps (0, 0.333, 0.667, 1, 1.333, 1.667, 2, …). When no datum was recorded between two timesteps, ‘piecewise polynomial’ in ‘pandas’ library (Python Software Foundation, 2020) was used for interpolation.

(3) The accumulated biogas production (mL) was converted to accumulated methane production (mL) according to the biogas composition tested.
(4) The accumulated methane production (mL) of the blank control group was subtracted from the accumulated methane production (mL) of the experimental group to obtain net methane production (mL). Note: as the volume of inoculum added in each experimental group was different, the corresponding methane production from inoculum degradation to be subtracted for each experimental group was different.
(5) VS content (g) of the mixture substrate for each experimental group was calculated.
(6) Specific methane production (mL CH_4/g VS added) was calculated as dividing the net methane production by the VS of the substrate added.

First-order kinetic model (Eq. (3-2) was used to fit the BMP test biogas production data. It is widely used for BMP test biogas production data description due to its relative good fit and with only two model parameters to be estimated (Kafle & Chen, 2016). The goodness-of-fit was evaluated with R^2 value (Eq. 3-1). To be noted that, different from the BMP test biogas production fitting for the digestate samples in Chapter 3, the ARMA model was not adopted in this chapter, although it was shown that the ARMA model fitted digestate samples better than the first-order kinetic model. This decision was due to the use of different substrates for the BMP tests in this chapter, as opposed to the sole use of digestate samples. Drastically different trajectories of the biogas production in different BMP experimental groups were observed, and thus one ARMA model (in terms of its order) could hardly represent all 10 BMP tests in this chapter well.

During the fitting, Y_m of the first-order kinetic model was set to the maximum biomethane yield observed from the experiment because the experiment was run until the end of the biogas production achieved. Therefore, only parameter k was actually fitted. The fitting was done by scipy package for python (scipy.optimize.curve_fit) (Jones, 2001).

BMP_{theo} was calculated from the chemical compositions of the mixture samples for the 10 experimental groups according to (2-4 in Chapter 2. The carbohydrate

in the equation includes the non-fibre carbohydrates, cellulose, and hemicellulose. This is from the consideration that the lignin contents in all three substrates (A), (O) and (B) are very low, so the inhibition to the degradations of cellulose and hemicellulose caused by the shielding effect of lignin is negligible. And lignin is almost non-degradable under anaerobic condition. Therefore, it was not included in the calculation of BMP_{theo}. This BMP_{theo} is an estimation of the ultimate BMP of the mixture sample.

The ratio Y_m/BMP_{theo} was calculated for each sample. This ratio is usually used to indicate biodegradability of materials. However, in this study, as very limited inhibition to Y_m caused by lignin was expected, the deficit in the observed CH_4 yield of the BMP test compared to the BMP_{theo} should mainly be caused by two factors: (1) not all the energy produced from the degradation of organic material is directed toward methane production, as some is consumed for cell maintenance and reproduction, as concluded in Section 2.4, (2) some biogas in the headspace failed to be pushed out to the gas measuring unit (biogas measurement inaccuracy). This study used this ratio Y_m/BMP_{theo} to indicate the level of experimental BMP result deficit due to these two factors, specific to inoculum choice, gas measuring unit choice and operational procedure of the study.

4.1.1.4 Response Surface Fitting

A response surface equation which takes the fractions of three substrates in an experimental group as inputs and the fitted k or Y_m parameter of the first-order kinetic model as response was fitted. The response surface equations adopted widely are respectively special cubic (Eq. (4-1), full cubic (Eq. (4-1), and special quartic (Eq. (4-3) (Cornell, 2002), in which $q = 3$ for this study.

$$Y = \sum_{i=1}^{q} \beta_i x_i + \sum\sum_{i<j}^{q} \beta_{ij} x_i x_j + \sum \sum_{i<j<k}^{q} \beta_{ijk} x_i x_j x_k \qquad \textbf{(4-1)}$$

$$Y = \sum_{i=1}^{q} \beta_i x_i + \sum\sum_{i<j}^{q} \beta_{ij} x_i x_j + \sum\sum_{i<j}^{q} \beta_{ij} x_i x_j (x_i - x_j) + \sum\sum_{i<j<k}^{q} \beta_{ijk} x_i x_j x_k \quad \textbf{(4-2)}$$

$$Y = \sum_{i=1}^{q} \beta_i x_i + \sum\sum_{i<j}^{q} \beta_{ij} x_i x_j + \sum\sum_{i<j<k}^{q} \sum \beta_{iijk} x_i{}^2 x_j x_k + \sum\sum_{i<j<k}^{q} \sum \beta_{ijjk} x_i x_j{}^2 x_k + \sum\sum_{i<j<k}^{q} \sum \beta_{ijkk} x_i x_j x_k{}^2 \quad \textbf{(4-3)}$$

According to Gauss-Markov theorem, when fitting a linear equation, to make sure the ordinary least square estimator of the coefficient is the best linear unbiased estimator, no perfect collinearity can exist among predictors. Given the relationship that $x_1 + x_2 + x_3 = 1$ for Mixture Design, it is necessary to replace one of the x_i by the expression of the other two. After replacing x_3 by 1 – $x_1 - x_2$ for example, there are 8, 11 and 11 distinctive terms in Eq. (4-1, Eq. (4-2 and Eq. (4-3 respectively, taking the constant term into account. Considering there are only 10 data points for response surface fitting, to ensure the reliability of the coefficients fitted, special cubic equation (Eq. (4-1) was chosen in this study.

In conclusion, the response surface equation to be fitted is:

$$Y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1{}^2 + \beta_5 x_2{}^2 + \beta_6 x_1{}^2 x_2 + \beta_7 x_1 x_2{}^2 + c \quad \textbf{(4-4)}$$

where x_1 and x_2 are fractions of substrate (A) and (O) in an experimental group.

The t-statistics of the fitted parameters were calculated using Eq. (4-5. This statistic is utilised to assess the extent to which the estimator differs from zero. And p-value, which represents the probability of observing a t-statistic as extreme as the one computed from the data assuming null hypothesis ($H_0: \beta_0 = 0$) is true, was derived based on the t-statistic. It is usually confident enough to reject H_0 when p-value < 0.05.

$$t_{\hat{\beta}} = \frac{\hat{\beta}}{SE(\hat{\beta})} \quad \textbf{(4-5)}$$

where $\hat{\beta}$ is the estimator of the population parameter, $SE(\hat{\beta})$ is the standard error of the estimator.

4.1.1.5 Dataset Generation from Response Surface Equations

The way of generating additional data with physiochemical characteristics of the mixture substrate as input and the parameter k or Y_m as output from the fitted response surface equation for ML model training is as follow:

(1) Generate 1000 vectors of length three ($[f_A, f_O, f_B]$). f_A, f_O and f_B are fractions of three substrates in each mixture sample, $f_A + f_O + f_B = 1$.

(2) Calculate the content of each characteristic (Table 4-1) in the mixture sample by computing the weighted sum of the concentration of each characteristic in the individual substrate, assuming no chemical reactions occur. That is, the concentration of a characteristic in the mixture is: $X_{i,mixture} = f_A * X_{i,A} + f_O * X_{i,O} + f_B * X_{i,B}$, where $X_{i,A}$, $X_{i,O}$ and $X_{i,B}$ are the concentrations of characteristic i in the three substrates respectively as shown in Table 4-1.

(3) Calculate the parameters k and Y_m from the fitted response surface equations (Eq.(4-4) for the generated 1000 samples.

The final dataset for XGBoost training and testing is composed of the original 10 samples and the generated 1000 artificial data. The physiochemical characteristics in Table 4-1 except for 'density', 'TS', 'Ash' and 'VS' were used as input features. 'VS' was not included because the specific methane production (mL CH_4/g VS) was used to express the BMP test result and the feedstock VS of all the groups is around 2.5 g.

4.1.2 XGBoost Model Training

4.1.2.1 XGBoost Introduction

XGBoost is a supervised ML model, comprises of gradient boosted trees. It makes prediction from the sum of t weak learners (Eq.(4-6).

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i) \quad \textbf{(4-6)}$$

where K is the total number of trees. Each $f_k(x_i)$ is a classification and regression tree, the structure and leaf scores of which need to be decided.

The optimisation of the structure and output of the trees in an XGBoost model is by optimising the objective function shown in Eq. (4-7. It contains the training loss, which measures the difference between the observations and the predictions, and the regularisation loss, which measures the model complexity to balance between the bias and variance of a model.

$$obj = \sum_{i=1}^{n} l(y_i, \widehat{y_i}) + \sum_{k=1}^{K} \omega(f_k) \quad \textbf{(4-7)}$$

where n is the number of samples, $\omega(f_k)$ is the complexity of the k_{th} tree, y_i is the observation, $\widehat{y_i}$ is the prediction of the XGBoost model.

In practice, one level of the tree is optimised at a time by choosing a split that maximise 'Gain'. A regularisation parameter λ ('reg_lambda' parameter in 'xgboost' package) is adopted to reduce prediction's sensitivity to individual observation. And another regularisation parameter γ is specified to control the complexity of the classification and regression tree. If 'Gain' is smaller than γ ('gamma' parameter in 'xgboost' package), no split is suggested (xgboost developers, 2022b).

4.1.2.2 Predictor Selection for XGBoost Training

As *Objective 3* of this study is to interpret the prediction of the trained XGBoost model by mathematical expressions, an XGBoost model with high prediction accuracy using limited number of input features is desirable. This was achieved by first training an XGBoost model with excessive number of features ('XGBoost_xs'). A feature importance analysis was carried out subsequently. Then a new XGBoost model with only three important features ('XGBoost_red') without compromising the prediction accuracy too much was explored.

Based on the empirical finding of Gupta (2021), the inclusion or removal of highly correlated variables has little influence on the prediction accuracy of XGBoost model. However, both of the highly correlated variables will be screened out as important features during feature importance analysis of the trained model. This will influence the exploration of the 'XGBoost_red' model when trying different combinations of three features according to the feature importance analysis result. This is because that a model with restricted number of variables, some of which are highly correlated and convey similar information, is likely to yield suboptimal predictive performance.

Thus, to make sure that no highly correlated features are screened out during the feature importance analysis to the 'XGBoost_xs' model, highly correlated features were not included when training the 'XGBoost_xs' in the first place. Therefore, an examination of the Pearson correlations among 12 characteristics of the dataset was carried out.

The training of XGBoost in this study was implemented using xgboost 2.0.2 package for Python. Grid Search Cross Validation was carried out for parameters 'max_depth', 'learning_rate' and 'gamma' to decide the best structure of the model. The train-test split of the data was 3:1. The prediction accuracy of the trained XGBoost model was evaluated on the test set with two scoring metrics – 'R^2' score and root mean square error (RMSE).

It is not necessary to scale the input features to range [0,1] when training an XGBoost model as what is commonly done for training a neural network (Bishop, 1985). This is because the XGBoost is an ensemble of decision trees which are not sensitive to the scales of input features. Even so, the input features and output were both scaled to [0,1] by 'MinMaxScaler' ($X_{scaled} = (X - X_{min})/(X_{max} - X_{min})$) in scikit-learn (1.4.0) for Python in this study because metamodeling of XGBoost model by Meijer G-functions was implemented in Chapter 5. And the line integral of Meijer G-function is defined in range [0,1].

4.1.3 Feature Importance Analysis by SHAP Values

SHAP value is a model agnostic feature attribution method developed from game theory using Shapley values by Lundberg et al. (2018). The SHAP value is a sum of SHAP main effect (the isolated effect of a feature) and the interaction values. It evaluates the effect of a feature in a particular sample on the difference of the predicted response of this sample from the average prediction across all the samples. It provides a rich visualisation of the effect of individual data on feature attribution and has been merged into XGBoost package (Lundberg et al. (2018).

SHAP package provides three types of plot for interpreting the feature importance – ‘mean absolute SHAP values bar plot’, ‘summary plot’ and ‘dependence plot’. The 'mean absolute SHAP values bar plot' ranks features based on the magnitude of their mean absolute SHAP values across all samples. The 'summary plot', while ranking features similarly, also provides additional insights by displaying the SHAP value of each individual sample for every feature, represented by coloured dots. This visualisation aids in identifying whether a feature positively or negatively influences the output, as well as highlighting outlier effects. Lastly, the 'dependence plot' illustrates interaction effects between two features. While showing the SHAP values of one feature (y-axis) against the value change of this feature (x-axis) across all samples, it meanwhile colours these samples by the values of another feature. This second feature that interacts with the primary feature might help to explain the dispersion of samples observed at a certain range of the primary feature.

Based on the feature importance analysis result of ‘XGBoost_xs’, different combinations of three features were tried to build ‘XGBoost_red’ models. The R^2 values of these models were compared with the R^2 value of the ‘XGBoost_xs’. Two ‘XGBoost_red’ models with the smallest R^2 were selected eventually to predict each of the k and Y_m of the first-order kinetic model from the sample physiochemical characteristics.

4.2 Results

4.2.1 BMP Test Results and Biogas Production Model Fitting

4.2.1.1 MCC Control and BMP Test System Integrity

The specific methane production (SMP) of two MCC controls are respectively 352.467 and 356.729 mL CH_4/g VS, average 354.598 mL CH_4/g VS (Figure 4-3). This is 85.445% of the BMP_{theo} of MCC (415 mL CH_4/g VS) and satisfies the requirement for the cellulose control in the standard followed by this study (Holliger et al., 2016).

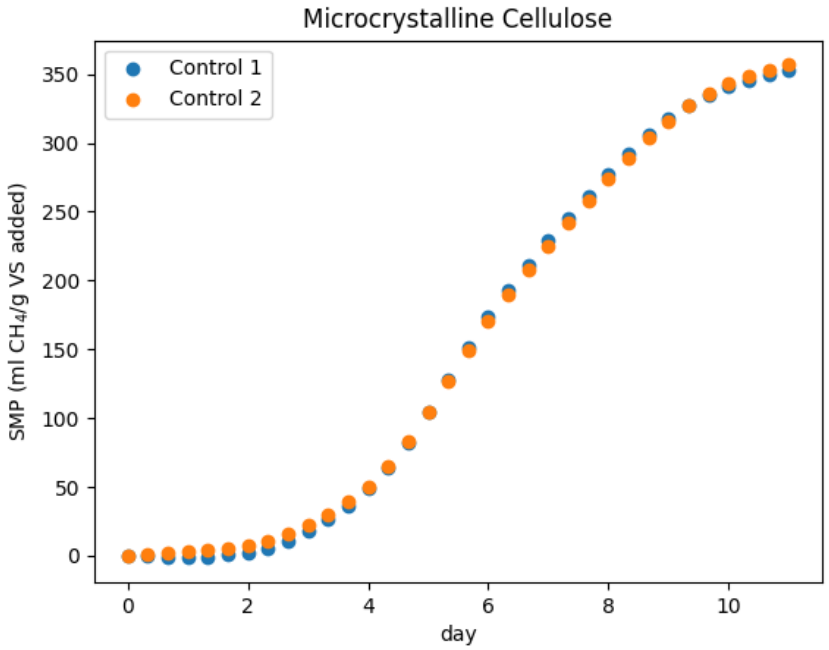


Figure 4-3 SMPs of two MCC control.

The substrate added for the MCC control was 5.954 g, with an VS of 98.845%. Consequently, the shortfall in CH_4 production between the control assay and the theoretical value was 355.480 mL. This deficit can be ascribed to two factors. The first factor is that about 10% of the substrate was utilised by the microorganism for biomass growth and converted into heat (Wang et al., 2014), equating to approximately 244.663 mL of CH_4 loss. Secondly, some gas in the headspace of the reactor could not be expelled. The total volume of the reactor was 763.410 mL with liquid volume for the control assay being 400 mL. Thus, a maximum volume of 363.410 mL biogas produced could not be pushed out. Considering a

typical CH_4 percentage of 50% in the biogas produced from cellulose degradation, this corresponds to 181.705 mL CH_4. The cumulative effect of these two factors explain the shortfall in the biogas production regarding to the BMP_{theo}. Therefore, the airtightness of the BMP test system and the accuracy of the gas measuring unit should be satisfactory.

4.2.1.2 First-order Kinetic Model Fitting of Experimental Groups

The graphs of the SMP and the fitted first-order kinetic model curves of the 10 experimental groups following the Mixture Design were shown in Figure 4-4, Figure 4-5, Figure 4-6, Figure 4-7 and Figure 4-8. The R^2 value of each fitting was displayed next to the SMP curve. The experimentally observed final biogas production (Y_m), the fitted parameter k of the first-order kinetic model for each BMP test, BMP_{theo}, the Y_m/BMP_{theo} ratio and the deficit in recorded CH_4 yield regarding to the BMP_{theo} ('CH_4 Deficit (mL)') were organised in Table **4-3**.

To verify the speculation about the two main factors contributing to the deficit in the experimentally observed CH_4 yield, the substrate utilisation by cell growth and maintenance and measurement inaccuracy, as discussed in Section 4.1.1.3, an examination of the 'CH_4 Deficit (mL)' value of Group #7 was conducted. The CH_4 deficit of Group #7 is 778.453 mL. The headspace for this experimental group is 548.410 mL. The CH_4 composition at the later stage of the BMP test was around 75% (not reported). So, a maximum 411.308 mL CH_4 may not have been pushed out of the reactor. Regarding the energy obtained from organic material degradation consumed by cell growth and maintenance, the typical value, 10% of the substrate, is considered (Wang et al., 2014). The amount of substrate for Group #7 is 2.915g VS, and thus this factor could account for 164.161 mL CH_4 deficit. Additionally, the inoculum was collected from an anaerobic digester treating THP pre-treated sewage sludge. When the inoculum was provided with new type of substrates, more substrates might be utilised for biomass growth due to a possible more favourable nutrient balance (Zhang et al., 2019). The increased biomass density observed after the BMP test in this study substantiated this speculation. These two factors explain the majority of the deficit in experimentally observed CH_4 yield. The remaining minor shortage might be

attributed to inhibition caused by lignin, incomplete degradation of material due to insufficient mixing and the fact that some components such as waxes and cutin. All in all, the Y_m/BMP_{theo} ratio for the BMP tests in this study is in the range 42.500% – 58.723%.

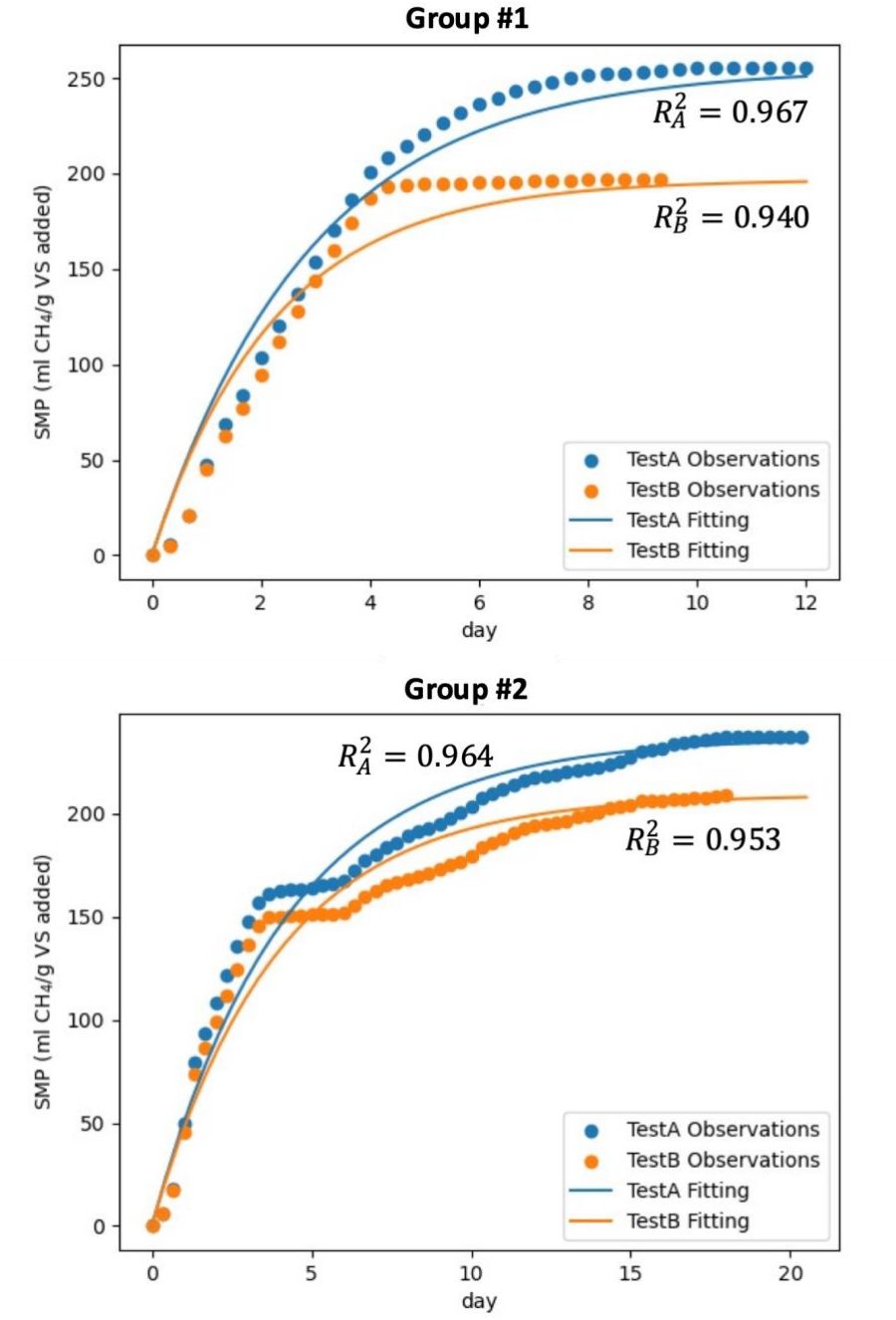


Figure 4-4 BMP test biogas production data and fitted first-order kinetic model curves (Part 1)

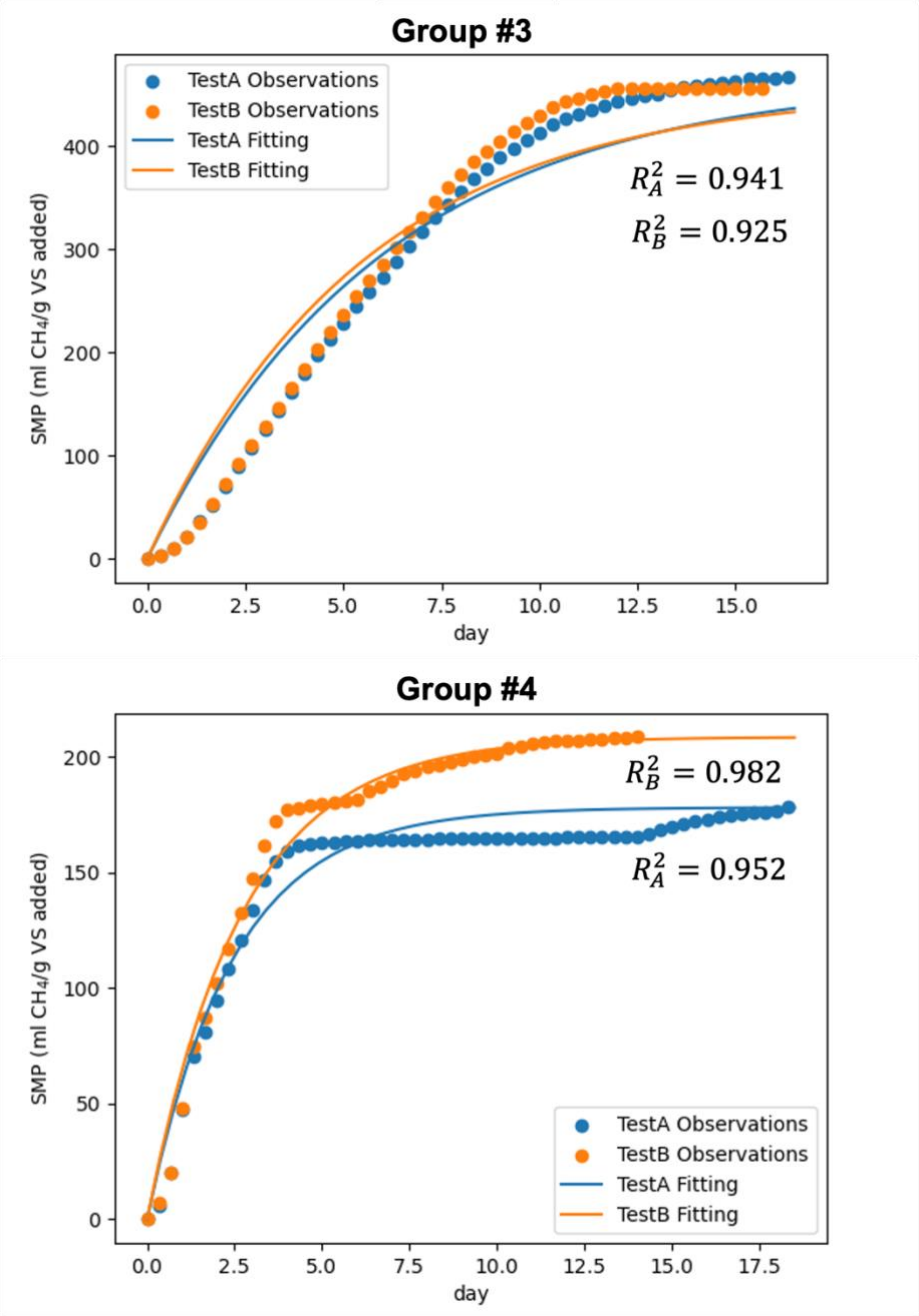


Figure 4-5 BMP test biogas production data and fitted first-order kinetic model curves (Part 2)

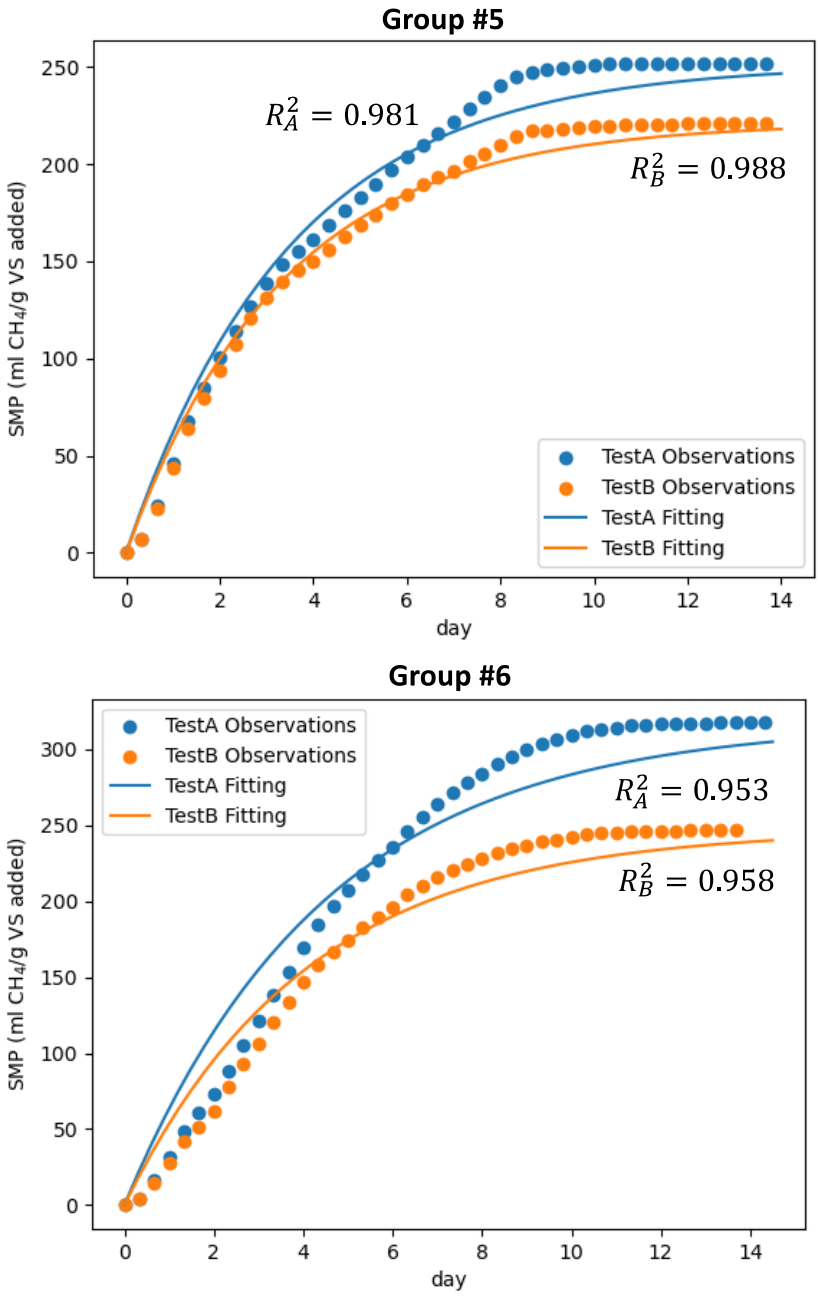


Figure 4-6 BMP test biogas production data and fitted first-order kinetic model curves (Part 3)

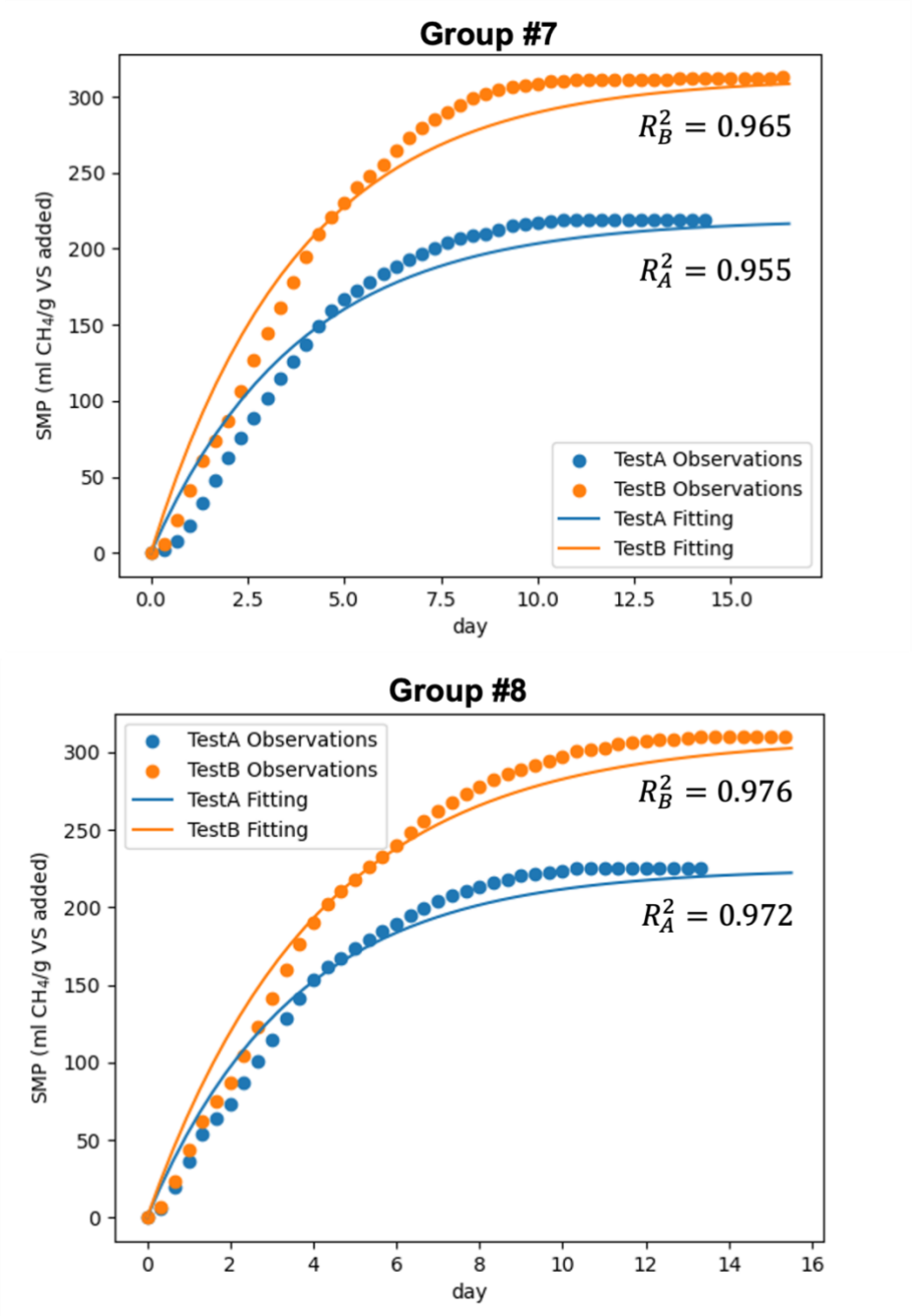


Figure 4-7 BMP test biogas production data and fitted first-order kinetic model curves (Part 4)

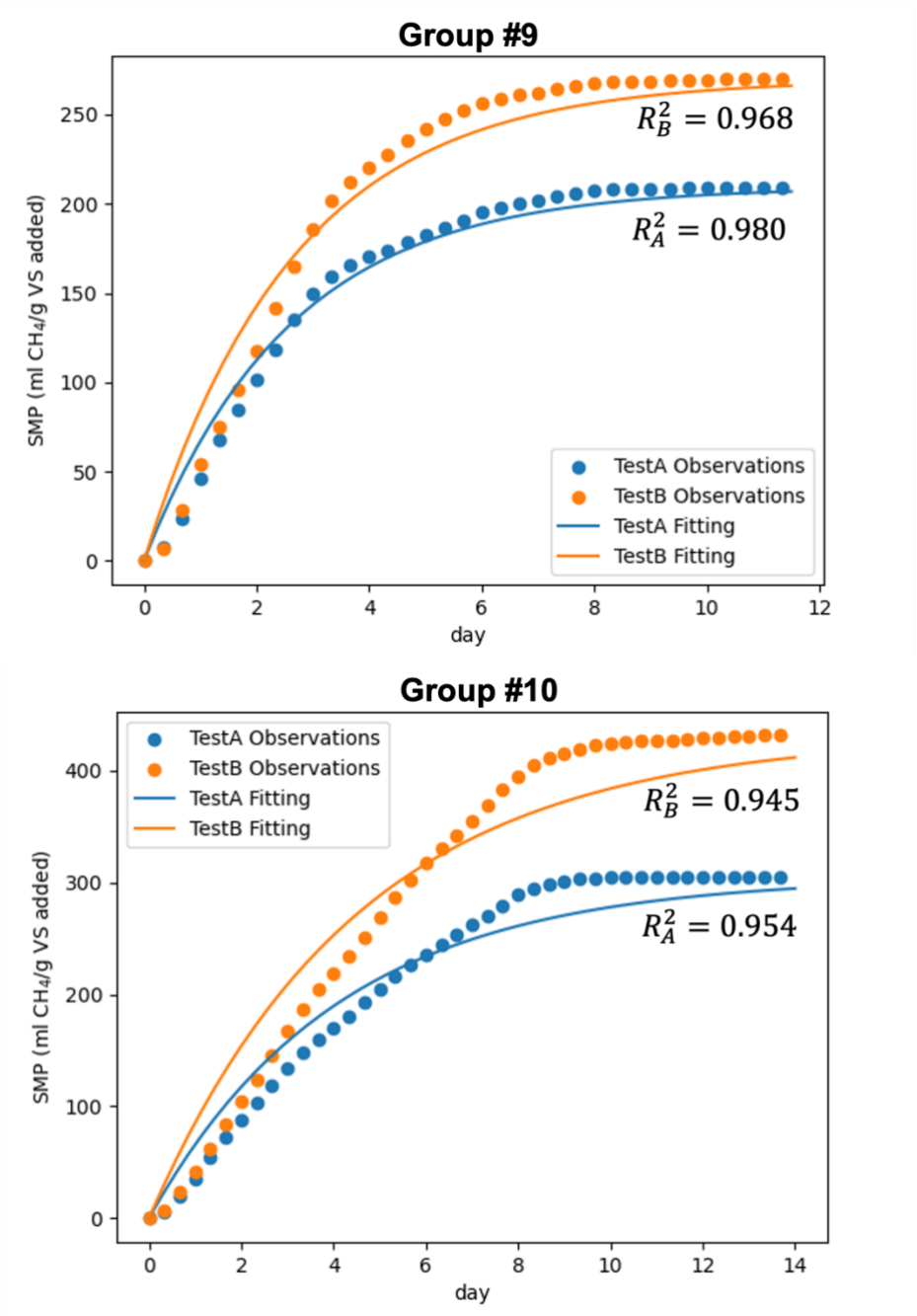


Figure 4-8 BMP test biogas production data and fitted first-order kinetic model curves (Part 5)

Table 4-3 Fitted k of first-order kinetic model, experimentally observed final biogas production (Y_m) and calculated BMP_{theo}.

Group	Test	Substrate Fractions			k (d^{-1})		Y_m (mL CH_4/g VS)		BMP_{theo} (mL CH_4/g VS)	Y_m/BMP_{theo} (%)	CH_4 Deficit (mL)
		A	O	B		Average		Average			
#1	A	1	0	0	0.342	0.392	255.060	225.863	452.521	49.912	623.588
	B				0.442		196.665				
#2	A	0	1	0	0.238	0.247	237.066	223.103	424.035	52.614	565.986
	B				0.255		209.139				
#3	A	0	0	1	0.167	0.175	466.173	460.647	784.440	58.723	951.337
	B				0.183		455.122				
#4	A	1/3	2/3	0	0.408	0.390	178.084	193.318	433.530	44.592	672.233
	B				0.371		208.553				
#5	A	0	2/3	1/3	0.282	0.291	251.470	236.450	544.170	43.452	870.109
	B				0.300		221.430				
#6	A	0	1/3	2/3	0.223	0.234	317.469	282.329	664.305	42.500	1113.538
	B				0.244		247.189				
#7	A	2/3	0	1/3	0.262	0.261	219.411	266.042	563.161	47.241	778.453
	B				0.261		312.673				
#8	A	1/3	1/3	1/3	0.282	0.263	224.977	267.255	553.666	48.270	822.057
	B				0.244		309.532				
#9	A	2/3	1/6	1/6	0.385	0.381	209.402	239.544			

	B				0.377		269.686		503.094	47.614	731.508
#10	A	1/6	1/6	2/3	0.243	0.232	304.594	367.894	669.053	54.987	861.075
	B				0.221		431.194				

First-order kinetic model fitted the majority of the samples with R^2 values larger than 95% except for Test B of Group #1, Test A and B of Group #3 and Test B of Group #10. Group #3 is the mono-digestion of substrate (B) (20% fat minced beef), exhibiting noticeably slower reaction kinetics compared to the other groups. Meanwhile, Group #6 and Group #10, which both have 2/3 substrate (B) in the mixture, also had relatively small fitted k values and unsatisfactory R^2 values (Table 4-3). Thus, it can be concluded that first-order kinetic model is not suitable for describing the slower biogas production kinetics of samples high in fat content. Despite the limitation in fitting such samples, the parameter k fitted still reflected their slow degradation kinetics, which remains meaningful when comparing with other samples.

The parameter k fitted for duplicate tests of each experimental group exhibited good repeatability. In contrast to the small k values for groups high in substrate (B), groups with substrate (A) (banana peel and green peas (w/w 1:1)) but meanwhile low in substrate (B), including Group #1, Group #4 and Group 9, had distinguishably high k values.

Relatively big discrepancies of Y_m values between test A and test B can be observed in some experimental groups. Although the repeatability between two tests in some experimental groups is not ideal, the relative magnitude of Y_m values among groups can still be observed. As shown in Table 4-3, Group #3, which is the mono-digestion of substrate (B), had the greatest Y_m. Consistently, the co-digestion Groups #10 and #6 with 2/3 substrate (B) also had relatively high Y_m values compared to other groups. While the Group #1 – mono-digestion of substrate (A), Group #2 – mono-digestion of substrate (O) and Group #4 – the co-digestion between substrate (A) and (O) had the lowest Y_m values. The other groups had medium Y_m values. As only duplicate tests were carried out, it was hard to decide which one was more accurate. The average value of test A and B was calculated as the Y_m for each experimental group eventually.

These observations about k and Y_m values conform with the characteristics of the three substrates used (Table 4-1). ‘Lignin’, ‘Cellulose+Hemicellulose’ and ‘Non-

fibre Carbohydrate' are only contained in substrate (A) and (O). From the fitted k values, it showed that the 'Lignin' component did not slow down the reaction kinetics of groups containing substrate (A) and (O) compared to groups high in substrate (B). This can be ascribed to the extremely low concentration of lignin in substrate (A) and (O). Regarding the degradation kinetics of cellulose and hemicellulose, as proved in Yang et al. (2021)'s study, synthetic sample containing an equivalent amount of lignocellulosic material (mainly composed of cellulose and hemicellulose) as the excess sludge sample but devoid of protein and lipid exhibited even greater degradation kinetics. Moreover, it was observed that higher hemicellulose content correlated with faster kinetics. Also, 'Non-fibre Carbohydrate', which are mainly starch in the potato and soluble sugar in the banana pulp, are also readily degradable. Therefore, the k values are high when substrate (A) and (O) are high in the mixture sample. By contrast, the protein and lipid, which are only contained in substrate (B), have lower conversion rate but higher biogas potential compared to carbohydrate (Raposo et al., 2012; Wagner et al., 2013). Therefore, the more substrate (B), the smaller k value but higher Y_m value.

Given that the *objective* 2 and 3 of this study are to test the general predictability of a ML model trained with data generated from response surface design and explore a pioneering method for interpreting the ML model mathematically, the quality of these data should not compromise the generalisability of the conclusion derived.

4.2.2 Response Surfaces Fitting

The statistics of the fitting of a special cubic equation (Eq. (4-4) to the parameter k of 10 experimental groups is listed in Table 4-4. As can be observed, the standard errors of the fitted coefficients are relatively large, except for the constant. This has resulted in small t-statistics and thus large p-values. If evaluated by p-value < 0.05, only the constant is statistically different from zero and has significant effect on the response value. However, the large standard error for the fitted coefficients can be attributed to the existence of 'near colinear' predictors, meaning that two or more predictors are highly, but not perfectly,

correlated with each other and thus lead to redundancy in the information provided by these variables (Kutner et al., 2005). Examining the correlations between predictors, the Pearson correlations between terms x_1 and ${x_1}^2$, terms x_2 and ${x_2}^2$ are 0.95. And the correlations between terms x_1x_2 and ${x_1}^2x_2$, terms x_1x_2 and $x_1{x_2}^2$ are 0.93 and 0.92 respectively. A commonly adopted practice in linear regression might be excluding the 'near colinear' terms. However, it is proved scientifically untenable to do so (Blanchard, 1987; Kalnins & Praitis Hill, 2023). Excluding strongly correlated terms will undermine the confounding effect among predictors and will cause worse coefficient estimates and downward-biased standard error estimation. The existence of approximate multicollinearity does not impair the predictive power of the model and does not reduce the accuracy in coefficient estimates (Gujarati, 2020). From above discussion, no predictors removal and refitting were conducted. The final response surface equation of parameter k is shown by Eq. (4-8. The R^2 value of the model fitting is 0.975.

Table 4-4 The statistics of the fitting of a special cubic equation to the parameter k of the first-order kinetic model of 10 samples.

Parameters	Fitted value	Standard Error	t-statistics	p-value
constant	0.176	0.024	7.336	0.018
x_1	-0.014	0.138	-0.105	0.926
x_2	0.307	0.118	2.594	0.122
x_1x_2	-1.113	0.998	-1.115	0.381
${x_1}^2$	0.228	0.140	1.634	0.244
${x_2}^2$	-0.233	0.114	-2.046	0.177
${x_1}^2x_2$	2.266	1.245	1.820	0.210
$x_1{x_2}^2$	1.142	0.866	1.318	0.318

$$k = -0.014x_1 + 0.307x_2 - 1.113x_1x_2 + 0.228{x_1}^2 - 0.233{x_2}^2 + 2.266{x_1}^2x_2 + 1.142x_1{x_2}^2 + 0.176 \quad \textbf{(4-8)}$$

The residuals and the APEs of the k parameters predicted by the fitted response surface equations are displayed in Table 4-6.

Residuals of the prediction for k were analysed to verify that the 'zero conditional mean of errors' and 'homoscedastic error' conditions required by Gauss-Markov assumptions for ordinary least square estimator being the best linear unbiased estimator were satisfied (Figure 4-9). Despite the relatively limited data, it can be observed that the residuals roughly distribute as a standard normal distribution and do not exhibit an obvious pattern with regard to the predicted k values.

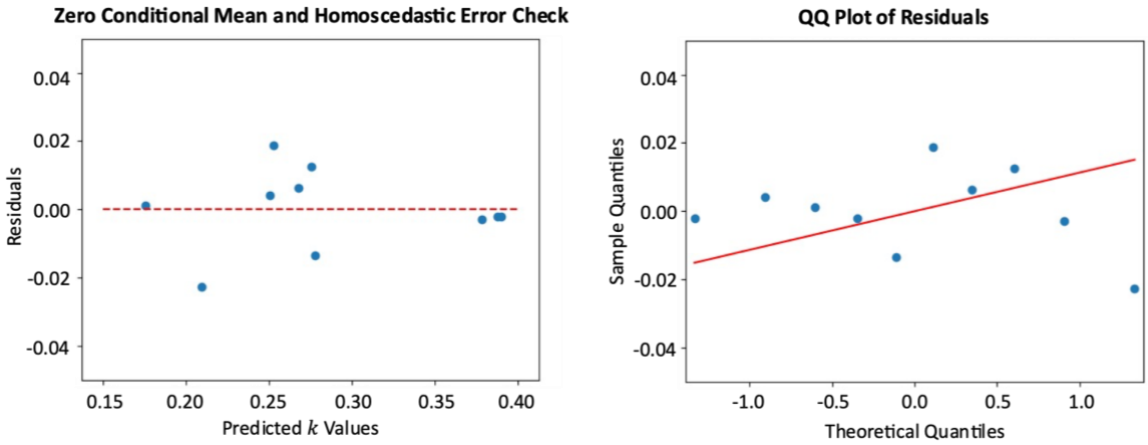


Figure 4-9 Gauss-Markov assumption checks for the residuals of the prediction for parameter k.

Similarly, the statistics of the fitting of a special cubic equation for parameter Y_m is presented in

Table 4-5. Eq. (4-9 is the equation for the response surface of parameter Y_m. The R^2 value of the model fitting is 0.980.

Table 4-5 The statistics of the fitting of a special cubic equation to the parameter Y_m of first-order kinetic model of 10 samples.

Parameters	Fitted value	Standard Error	t-statistics	p-value
constant	461.304	22.903	20.141	0.002
x_1	-377.589	131.647	-2.868	0.103
x_2	-605.170	113.086	-5.351	0.033
x_1x_2	1752.369	952.523	1.840	0.207

$x_1{}^2$	140.199	133.283	1.052	0.403
$x_2{}^2$	371.185	108.525	3.420	0.076
$x_1{}^2x_2$	-1547.118	1188.247	-1.302	0.323
$x_1x_2{}^2$	-1322.914	826.887	-1.600	0.251

The residuals and the APEs of the Y_m parameters predicted from the fitted response surface equations are displayed in Table 4-6.

$$Y_m = -377.589x_1 - 605.170x_2 + 1752.369x_1x_2 + 140.199x_1{}^2 + 371.185x_2{}^2 - 1547.118x_1{}^2x_2 - 1322.914x_1x_2{}^2 + 462.304 \quad \textbf{(4-9)}$$

The residual analysis of Y_m is presented in Figure 4-10. Similarly, the residuals have roughly zero mean, are homoscedastic conditioning on predicted Y_m values and are normally distributed.

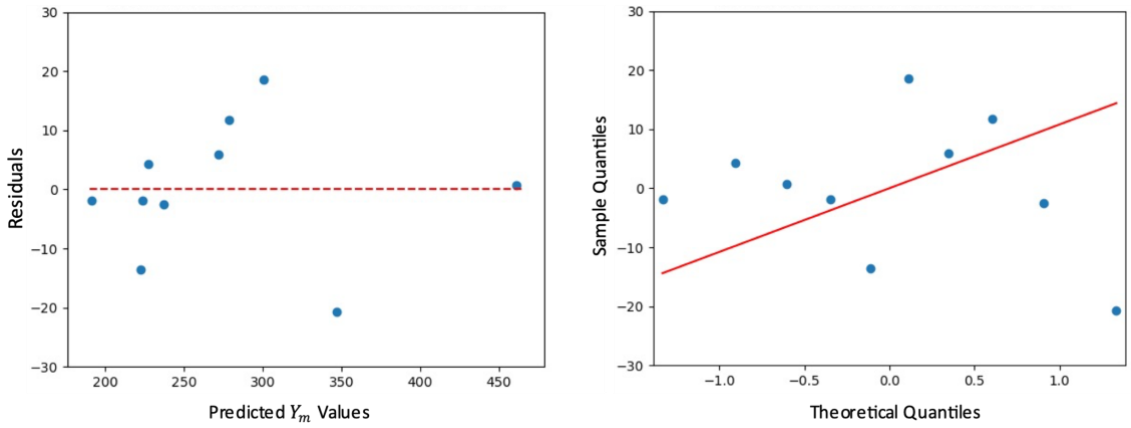


Figure 4-10 Gauss-Markov assumption checks for the residuals of the prediction for parameter Y_m.

Table 4-6 The residuals and the APEs of the k and Y_m parameters predicted from the fitted response surface equations (Predicted) with respect to the parameters derived by fitting the first-order kinetic model (Fitted).

	k				Y_m			
Group	Fitted	Predicted	Residuals	APE	Fitted	Predicted	Residuals	APE
#1	0.392	0.390	-0.002	0.527	225.863	223.914	-1.949	0.863
#2	0.247	0.251	0.004	1.661	223.103	227.319	4.216	1.890
#3	0.175	0.176	0.001	0.648	460.647	461.304	0.656	0.142
#4	0.390	0.388	-0.002	0.544	193.318	191.370	-1.949	1.008
#5	0.291	0.277	-0.014	4.646	236.450	222.828	-13.622	5.761
#6	0.234	0.253	0.019	8.027	282.329	300.823	18.494	6.551
#7	0.261	0.268	0.006	2.431	266.042	271.888	5.847	2.198
#8	0.263	0.276	0.013	4.814	267.255	278.948	11.693	4.375
#9	0.381	0.378	-0.003	0.733	239.544	236.945	-2.599	1.085
#10	0.232	0.210	-0.022	9.693	367.894	347.105	-20.788	5.651

To help with the visualisation of the response surfaces, Figure 4-11 and Figure 4-12 depicted the 3D plot and 2D contour plots of the response surfaces for parameter k and Y_m with respect to the fractions of substrate (A) and (O) in the mixture. The diagonal of the contour plot is where there is no substrate (B), and lower half of the triangle is where there is more substrate (A) than substrate (O) in the mixture.

From the 2D contour plot of the response surface for k in Figure 4-11, the smallest k value (red colour) appears at the corner where fraction of substrate (B) is high. And it is observed the highest k value (purple colour) appears at the diagonal line and lower half of the triangle. This aligns with the observation and discussion in Section 4.2.1.2.

The 2D contour plot of the response surface for parameter Y_m aligns with the preliminary observations outlined in Section 4.2.1.2 as well. The high Y_m value is at the corner where the fraction of substrate (B) is high and the lowest Y_m value is near the diagonal line where substrate (A) and (O) are high.

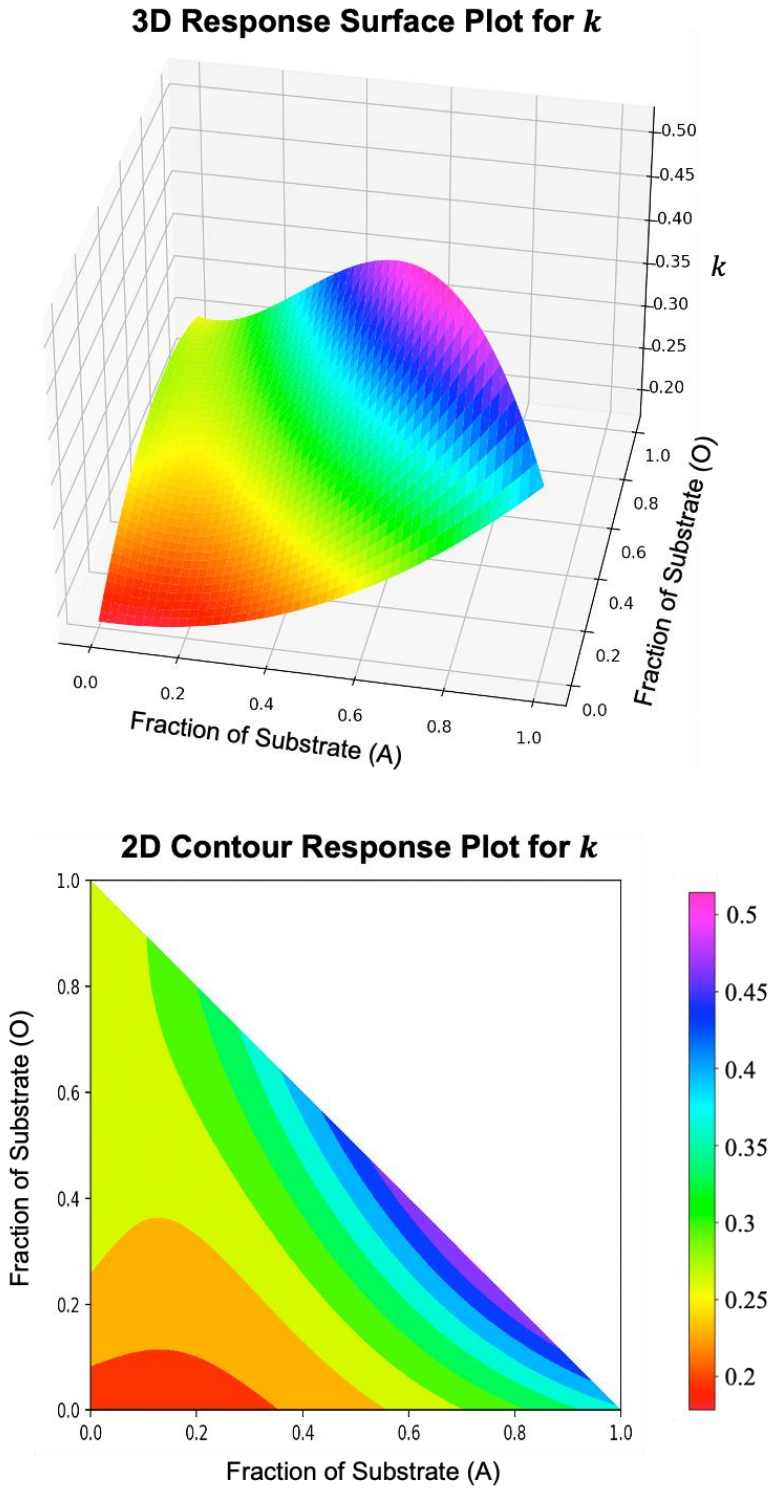


Figure 4-11 The fitted response surface for the parameter k with respect to the fractions of substrate (A) and substrate (O) in the co-digestion mixture. (Top: 3D plot; Bottom: 2D Contour plot).

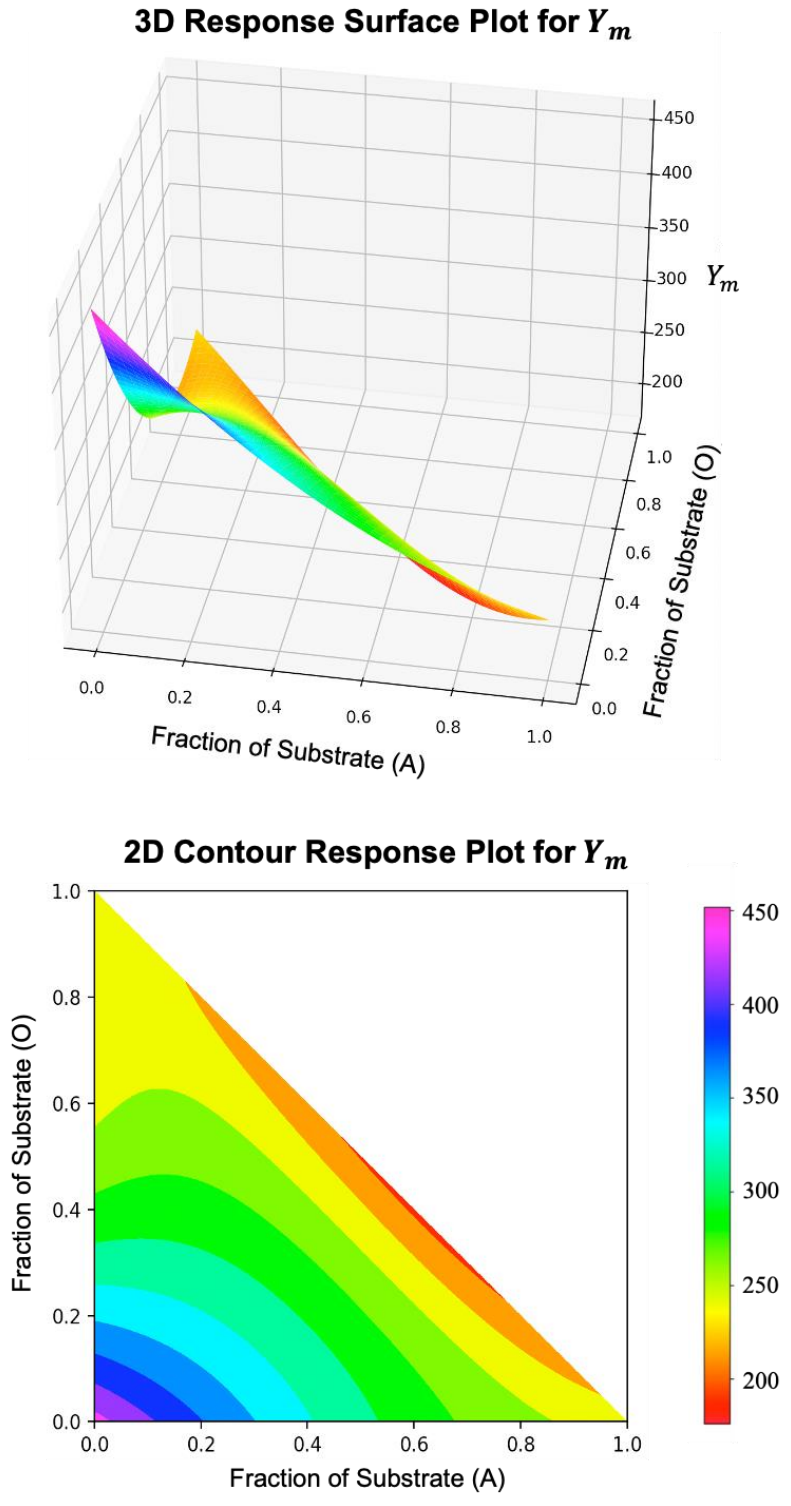


Figure 4-12 The fitted response surface for the parameter Y_m with respect to the fractions of substrate (A) and substrate (O) in the co-digestion mixture. (Top: 3D plot; Bottom: 2D Contour plot).

Though Eq. (4-8 and (4-9 fitted the responses of parameters k and Y_m with high R^2 values (0.975 and 0.980 respectively), with no extra testing data, it is hard to tell if these functions represent the true response surfaces well. In other words, if the k and Y_m parameters inferred from these equations for a new mixture substrate will conform with the parameters derived from the biogas production data fitting. To have an idea on how the accuracy of the response surface will improve if more samples are available for fitting, LOO cross validation was implemented. That is, nine samples were used to fit a response surface equation each time, then using this fitted equation to predict the response of the sample left out. The APE of the prediction for the left-out sample was calculated. Table 4-7 presented the result for each LOO train-test-split scenario. For the conciseness of the table, the coefficients of these new response surface equations were not reported.

Table 4-7 Results of the LOO cross validation for parameter k and Y_m. Fitted: the parameter k/Y_m derived by fitting first-order kinetic model to biogas production data. Predicted: the prediction of k/Y_m of the left-out sample with the response surface equations fitted with nine samples.

	k			Y_m		
Group left out	Fitted	Predicted	APE	Fitted	Predicted	APE
#1	0.392	0.023	94.131	225.863	-114.914	150.878
#2	0.247	0.328	33.006	223.103	305.905	37.114
#3	0.175	0.186	6.538	460.647	467.338	1.453
#4	0.390	0.021	94.587	193.318	-147.458	176.277
#5	0.291	0.261	10.251	236.450	206.366	12.723
#6	0.234	0.274	17.131	282.329	321.802	13.981
#7	0.261	0.384	47.039	266.042	379.634	42.697
#8	0.263	0.325	23.383	267.255	324.051	21.252
#9	0.381	0.104	72.649	239.544	-16.038	106.695
#10	0.232	0.198	14.888	367.894	335.946	8.684

First of all, compared to the APEs when 10 samples were used for fitting (Table 4-6), the APEs of the response prediction for the left-out sample from the surface equation fitted with nine samples are significantly bigger. This is especially the case when Group #1, #4 and #9 were left out. In contrast, the APEs were small when Group #3, #5, #6 and #10 were left out.

To display the results clearer, the APE when different individual sample was left out with its position in the ASCD triangle for k and Y_m are presented in Figure 4-13. The reason why the APEs of the inferred responses for Group #3, #5, #6 and #10 were small, in other words, why the removal of one of the Group #3, #5, #6 or #10 had less impact on the accuracy of the fitted response surface has become more evident. It is because more samples are available at the lower left corner of the ASCD triangle. The other three samples still provided enough information about the curvature of the response surface at this area when one sample was removed.

However, it might be less obvious that why, for the Group #1 vs. #2, Group #4 vs. #7, and also Group #8 vs. #9, drastic different APEs were obtained when one of them was removed, while their positions in the triangle were similar. Figure 4-14 helps to illustrate the reason better. Assuming that the length of the ASCD triangle's side is one, any sample point within a distance of less than 1/2 from the sample being removed is considered a 'vicinity point'. These 'vicinity points' still provide insights into the curvature of the response surface at the position of the removed sample. In Figure 4-14, these 'vicinity points' to the removed sample are highlighted with dashed lines, indicating the distance between them.

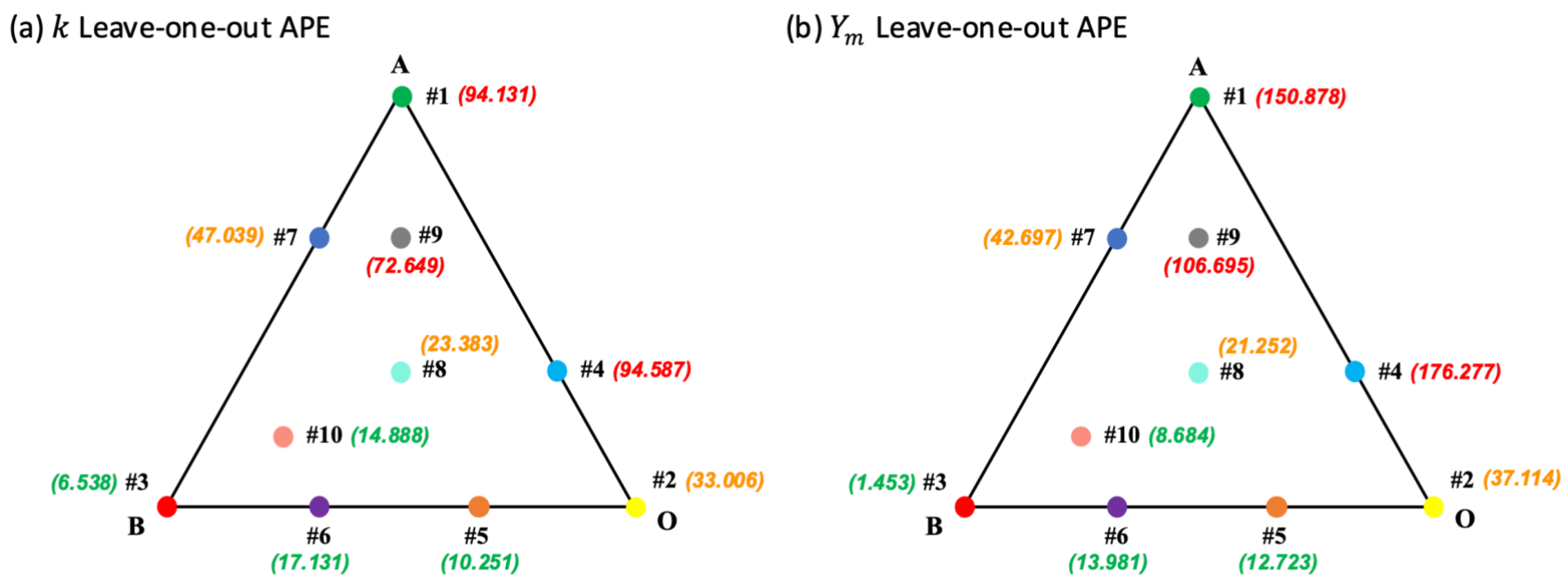


Figure 4-13 Visual inspection of the results of the LOO cross validation for parameter k and Y_m. The APE between the parameter klY_m derived by fitting first-order kinetic model to biogas production data and the prediction of klY_m of the left-out sample with the response surface equations fitted with nine samples is displayed in the parentheses.

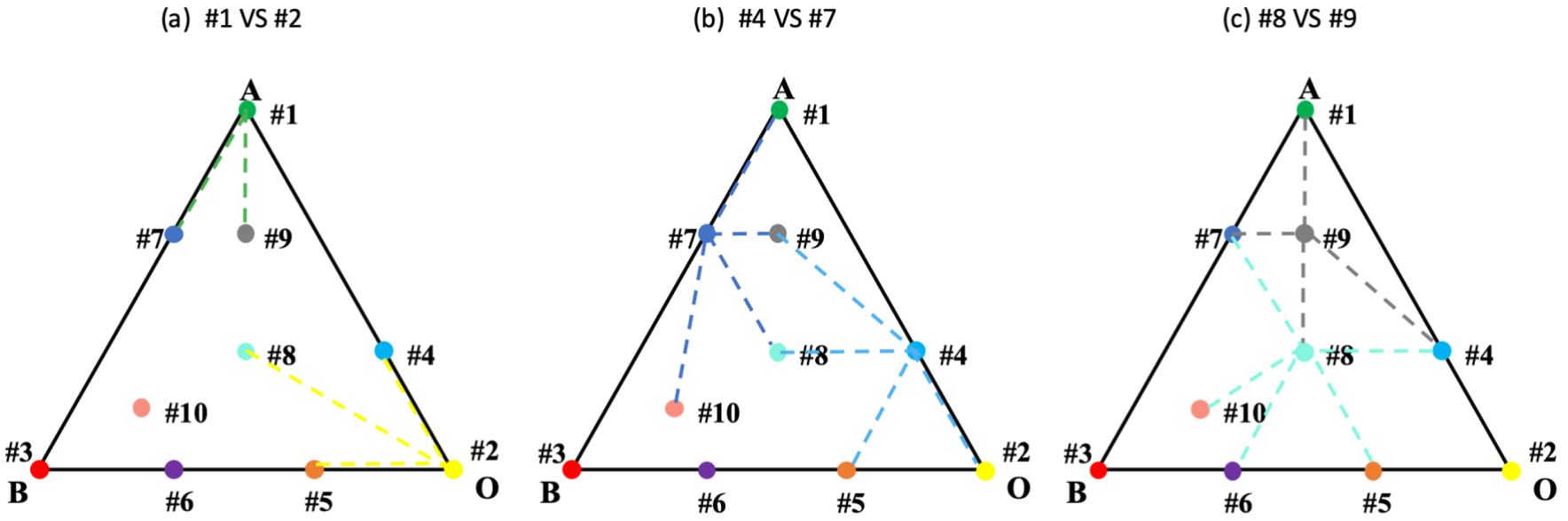


Figure 4-14 The 'vicinity points (the samples with a distance to the sample being left out smaller than ½ assuming the length of the side of ASCD triangle is one)' to the sample being left out when implementing LOO cross validation

When comparing the removal of Group #1 versus Group #2, despite both being situated at the vertex of the triangle, Group #2 possesses three 'vicinity points' compared to the two associated with Group #1. Consequently, there is a greater amount of information regarding the response surface curvature for position #2 when Group #2 is removed, resulting in a smaller APE for inferring the response at this position. Similarly, a smaller APE is caused when Group #8 with five ‘vicinity points’ was removed comparted with Group #9 with four ‘vicinity points’. For the comparison between the removal of Group #4 and #7, although both have four 'vicinity points' nearby, the total distance of these points to Group #4 is larger (1.441) compared to Group #7 (1.274). Consequently, the impact of the 'vicinity points' of Group #7 is more pronounced when #7 is removed, resulting in a smaller APE when inferring the response at position #7.

All in all, the above LOO cross validation study revealed that the more samples available around the inference point to provide curvature information of the response surface, the more accurate the inference would be. Thus, the accuracy of the response surface equation fitted for parameter k or Y_m will be higher if more samples are available for response surface fitting, and thus more reliable the k and Y_m at other positions in this ASCD triangle derived from the equation. Even so, a better practice in the future work is to have separate BMP test data to test the accuracy of the k and Y_m inferred.

4.2.3 Additional Dataset Generation

1000 additional samples with the physiochemical characteristics calculated from the blending ratios of each artificial co-digestion scenario and the parameter k and Y_m of the first-order kinetic model inferred from the fitted response surface equations were generated for XGBoost training and testing (not reported). The reliability of these generated additional data is influenced by two factors.

The first factor is the accuracy of the k and Y_m inferred from the response surface equations as discussed in the previous section. The other factor is the assumption that the content of a physiochemical characteristic in a mixture sample is the linear combination of the content of this characteristic in each

substrate without reactions happened. From the reported TA values in the mixture of sewage sludge, juice wastewater and cow dung in the study of Rao & Baral (2011), the TA content in the mixture calculated by weighted sum method conforms with the experimentally tested values very well. However, no information was found by the author to verify if this assumption also holds for other characteristics used as ML model inputs in this study, especially for chemical indicators TAN and COD. This should be verified in the future study.

4.2.4 XGBoost Training and Feature Importance Analysis

4.2.4.1 Predictors Selection

To mitigate the multicollinearity among the input features of the 'XGBoost_xs' model and only important features conveying relatively distinct information towards response inference are screened out by SHAP value method later. A selection of predictors for 'XGBoost_xs' model based on the correlation among physiochemical characteristics was conducted first. Figure 4-15 presents the correlation heatmap of all 12 physiochemical characteristics.

First of all, 'Lignin' is highly negatively correlated with 'H' content in the mixture sample (r = - 0.99). The correlation between 'Lipid' and 'H' is as high as 0.85 for these samples. As introduced in Chapter 2, lipid has the highest H content compared with non-fibre carbohydrate and protein. Also, due to fact that substrate (A) and (O) are high in lignin but low in lipid in contrast to substrate (B), the proportion of the 'Lignin' in a mixture sample is in negative correlation with the proportion of the 'Lipid' (r = -0.77). These two reasons might explain the strong negative correlation between 'Lignin' and 'H'.

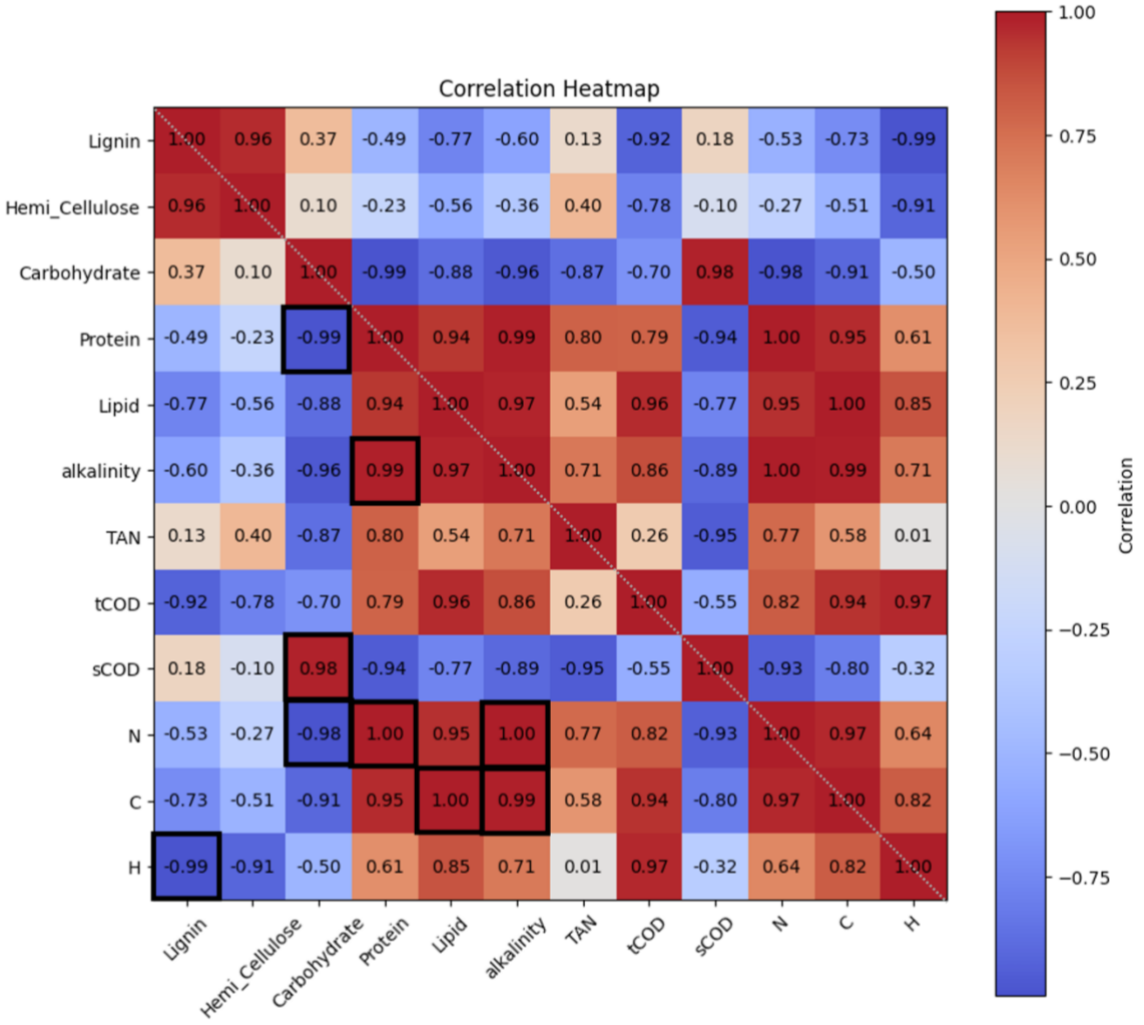


Figure 4-15 Pearson correlations (r) of 12 physiochemical characteristics (The $|r| \geq$ 0.98 was highlighted) ('Hemi_Cellulose' represents the sum of the contents of cellulose and hemicellulose').

Similarly, the 'Non-fibre Carbohydrate' only presents in substrate (A) and (O), while the proportion of 'Protein' is extremely high in the substrate (B). Therefore, the proportions of 'Non-fibre carbohydrate' and 'Protein' in a mixture sample are in negative correlation (r = - 0.99). Since 'N' is mainly from amino acids in the protein, 'Non-fibre Carbohydrate' is also in a strong negative correlation with 'N' (r = - 0.98). The main forms of the 'Non-fibre carbohydrate' in the substrates are the starch in the potato and soluble sugars in the banana pulp. Soluble sugars contribute greatly to 'sCOD'. Therefore, the 'Non-fibre carbohydrate' is in strong positive correlation with 'sCOD' (r = 0.98).

The strong positive correlation between "alkalinity" and "protein" (r = 0.99) may be due to the addition of alkaline solutions during meat processing for texture enhancement and quality preservation (Santos et al., 2012).

All in all, variables ‘N’, ‘C’ and ‘alkalinity’ were removed when training ‘XGBoost_xs' due to their strong correlations with other variables. Nevertheless, none of the ‘Non-fibre Carbohydrate’, ‘Protein’ and ‘Lipid’ were removed despite of strong correlations among them as these three components together with the content of lignocellulose are often used together to estimate BMP result as introduced in Chapter 2.

4.2.4.2 Feature Importance Analysis and Feature Reduction

(1) XGBoost for k

The best tree structure of the ‘XGBoost_xs’ model for the inference of the parameter k of the first-order kinetic model found by Grid Search Cross Validation is 'max_depth': 4, 'learning_rate': 0.2, 'gamma': 1e-05. The R^2 score on the test set is 0.997 and the RMSE is 0.010.

Figure 4-16 shows the feature importance analysis result by SHAP values. The content of the ‘tCOD’ in a substrate mixture is the most influential to the k value predicted. As observed in the ‘summary plot’, the low ‘tCOD’ content had outlier effect. That is, the degradation rate of a sample could be outstandingly high when its ‘tCOD’ content was low. This observation could be attributed to the notably lower 'tCOD' content in substrates (A) and (O) compared to in substrate (B) (Table 4-1). The higher k value corresponds to the region where substrate (A) and (O) are abundant from the 2D contour plot of the fitted response surface for parameter k (Figure 4-11). Therefore, ‘tCOD’ is in negative correlation with the reaction kinetic.

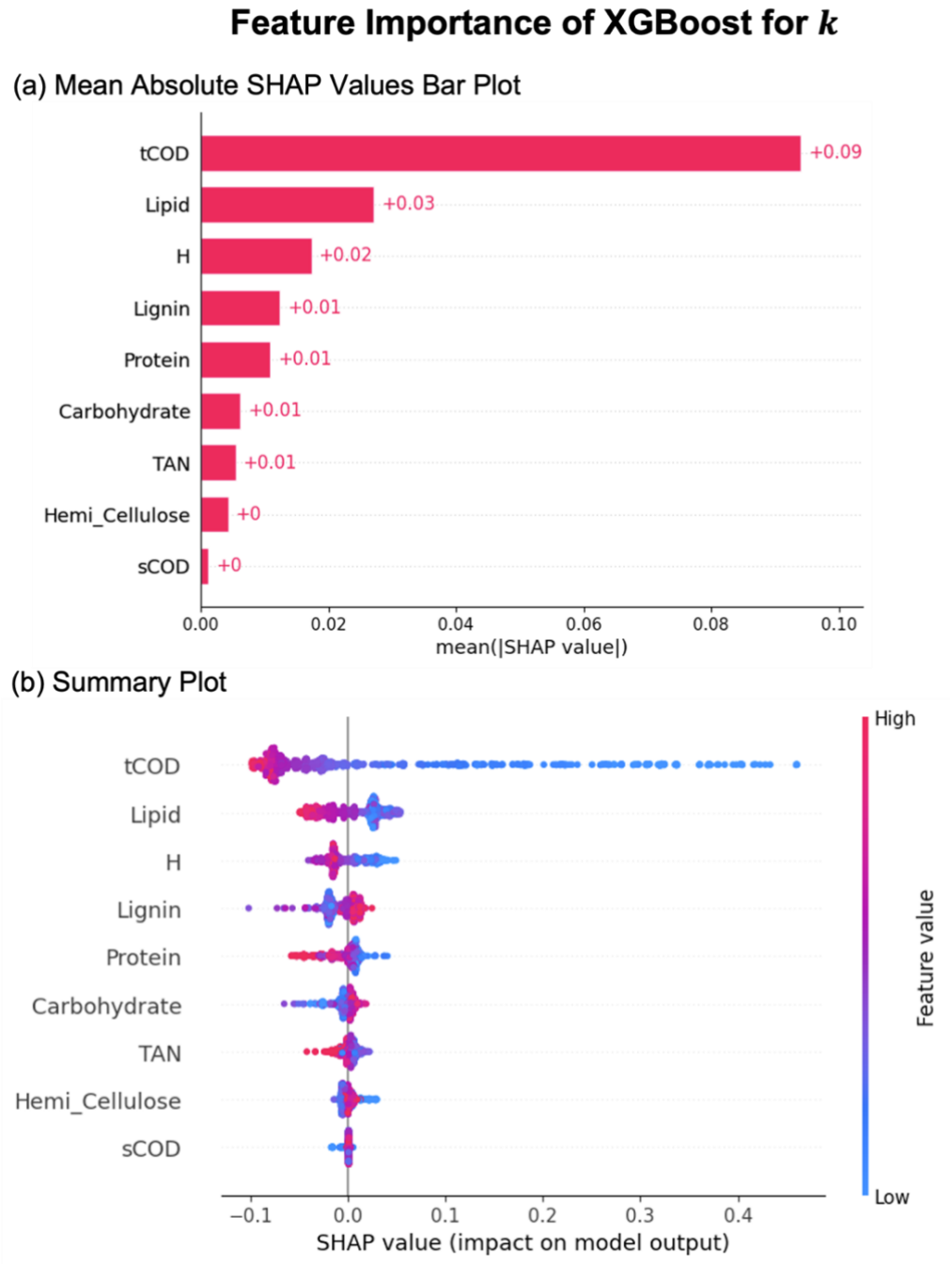


Figure 4-16 Feature importance of the XGBoost model for k ('Hemi_Cellulose' represents the sum of the contents of cellulose and hemicellulose).

The distributions of the SHAP values for predictors 'Lipid', 'Protein' and 'Non-fibre Carbohydrate' also conform with their relative degradation kinetics in the AD process (Esposito et al., 2012; Wagner et al., 2013).

Since the proportions of ‘Lipid’ and ‘Protein’ in a sample are positively correlated (r = 0.94), they distributed in a similar pattern.

The content of lignin in a mixture sample exhibited an overall positive correlation with the predicted k value. As discussed in Section 4.2.1.2, the content of lignin in all substrates in this study are rather low, so no obvious inhibition effect exerted by lignin was observed. Instead, the high lignin content in a mixture sample indicates that more substate (A) and (O) are contained in the sample and thus faster reaction kinetics. Also, it is observed that some samples with medium level lignin exhibited distinctive slow reaction kinetics. By checking the dependence plots of the SHAP values for ‘Lignin’ against the contents of ‘Non-fibre Carbohydrate’ and ‘Lipid’ in the samples (Figure 4-17), it is observed that there are two scenarios for a mixture sample to have medium lignin content: (i) low ‘Carbohydrate’ and high ‘Lipid’ (e.g. substrate (A): substrate (O): substrate (B) = 0.445: 0.014: 0.540, the corresponding k is 0.219), (ii) high ‘Carbohydrate’ and low ‘Lipid’ (e.g. substrate (A): substrate (O): substrate (B) = 0.218: 0.660: 0.122, the corresponding k is 0.305). The mean k for all the samples is 0.292. The scenario (i) corresponds to those samples showed extra small k values.

The ‘XGBoost_red’ models with only three important features were trained considering the feature importance analysis results of the ‘XGBoost_xs’ model. Four scenarios were tried (Table 4-8). The combination – 'Lipid', 'Lignin', and 'Protein' – was selected due to the following considerations. ‘tCOD’ is not a characteristic commonly analysed for food waste samples and it is also challenging to determine ‘tCOD’ content accurately for solid samples. Instead, ‘Lipid’, ‘Lignin’, and ‘Protein’ are all nutrient components that are commonly tested for food samples.

‘XGBoost_red_4’ with characteristics ‘Lipid’, ‘Lignin’ and ‘Protein’ as predictors has the highest R^2 value and smallest RMSE. It was chosen for k value prediction eventually.

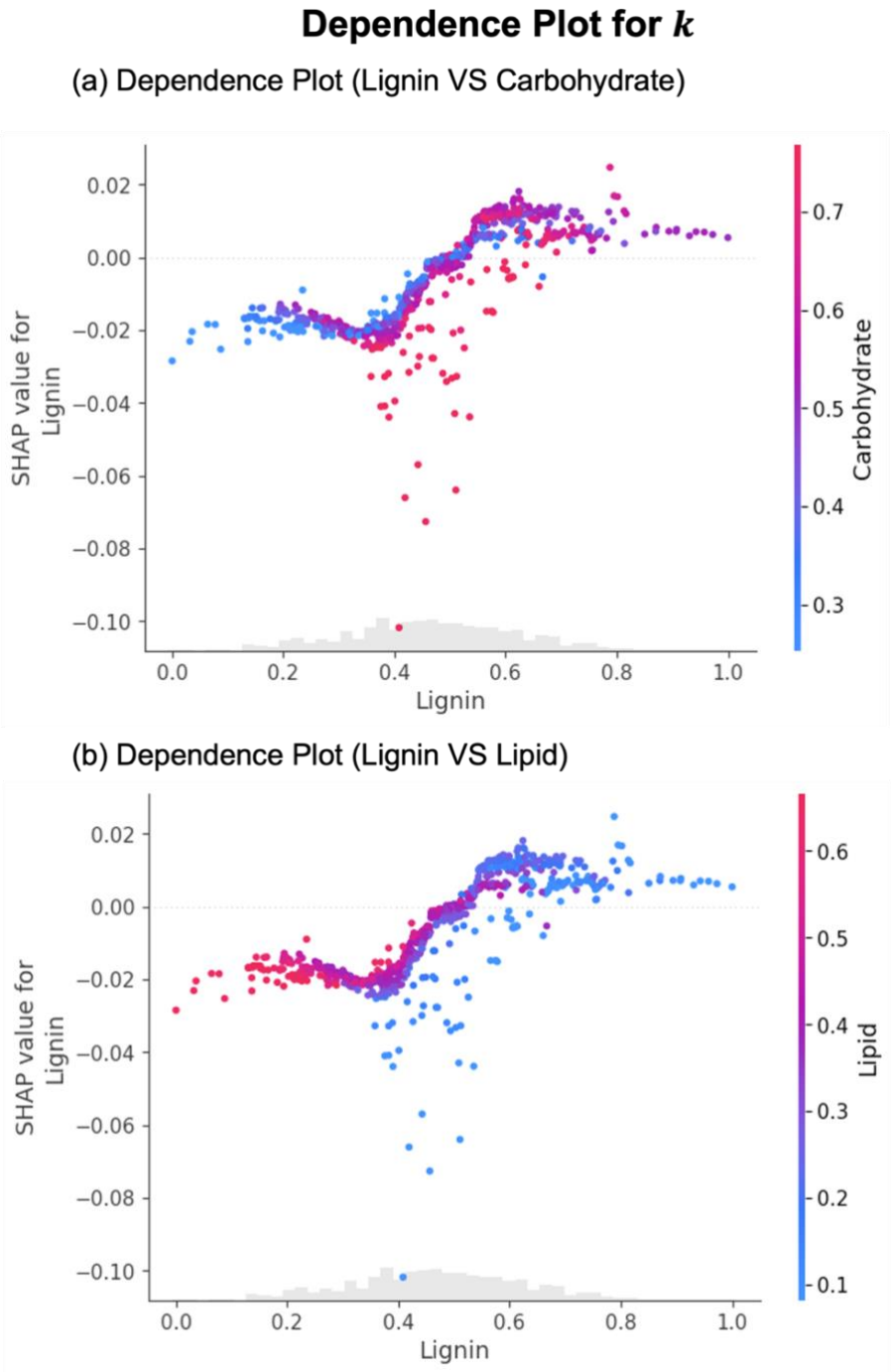


Figure 4-17 Dependence plot of SHAP values for feature 'Lignin' against the values of feature 'Carbohydrate' (a) and feature 'Lignin' against the values of feature 'Lipid' (b) for k prediction.

Table 4-8 The R^2 and RMSE of 'XGBoost_xs' model and 'XGBoost_red' models for parameter k prediction on test set with different combinations of three features.

	Features			R^2	RMSE
XGBoost_xs	All 9 features			0.997	0.010
XGBoost_red_1	tCOD	Lipid	H	0.987	0.022
XGBoost_red_2	tCOD	Lipid	Lignin	0.992	0.017
XGBoost_red_3	tCOD	Lipid	Protein	0.994	0.014
XGBoost_red_4	Lipid	Lignin	Protein	0.995	0.013

(2) XGBoost for Y_m

The best tree structure of the 'XGBoost_xs' model for Y_m prediction found by Grid Search Cross Validation for parameter Y_m is 'max_depth': 5, 'learning_rate': 0.2, 'gamma': 1e-07. The R^2 score on the test set is 0.999, the RMSE is 0.010.

Figure 4-18 shows feature importance analysis results. From the summary plot, the more lipid and protein, the bigger the Y_m is. Conversely, the more non-fibre carbohydrate, the smaller the Y_m is. This conforms with their theoretical biogas yield calculated from the Boyle formula (Eq. (2-4).

tCOD correlates with the overall concentration of the organic matter in a sample. Although the BMP result is usually overestimated using tCOD concentration, high 'tCOD' generally corresponds to a high Y_m value.

From Figure 4-18, TAN was also important for the Y_m value prediction. The higher TAN concentration in a mixture sample, a bigger Y_m value was predicted. This is because TAN is positively correlated with the lipid and protein contents in a sample and negatively correlated with the non-fibre carbohydrate content (Figure 4-15). This also suggests that the TAN concentrations in all the substrates in this study were not as high as inhibitions happened.

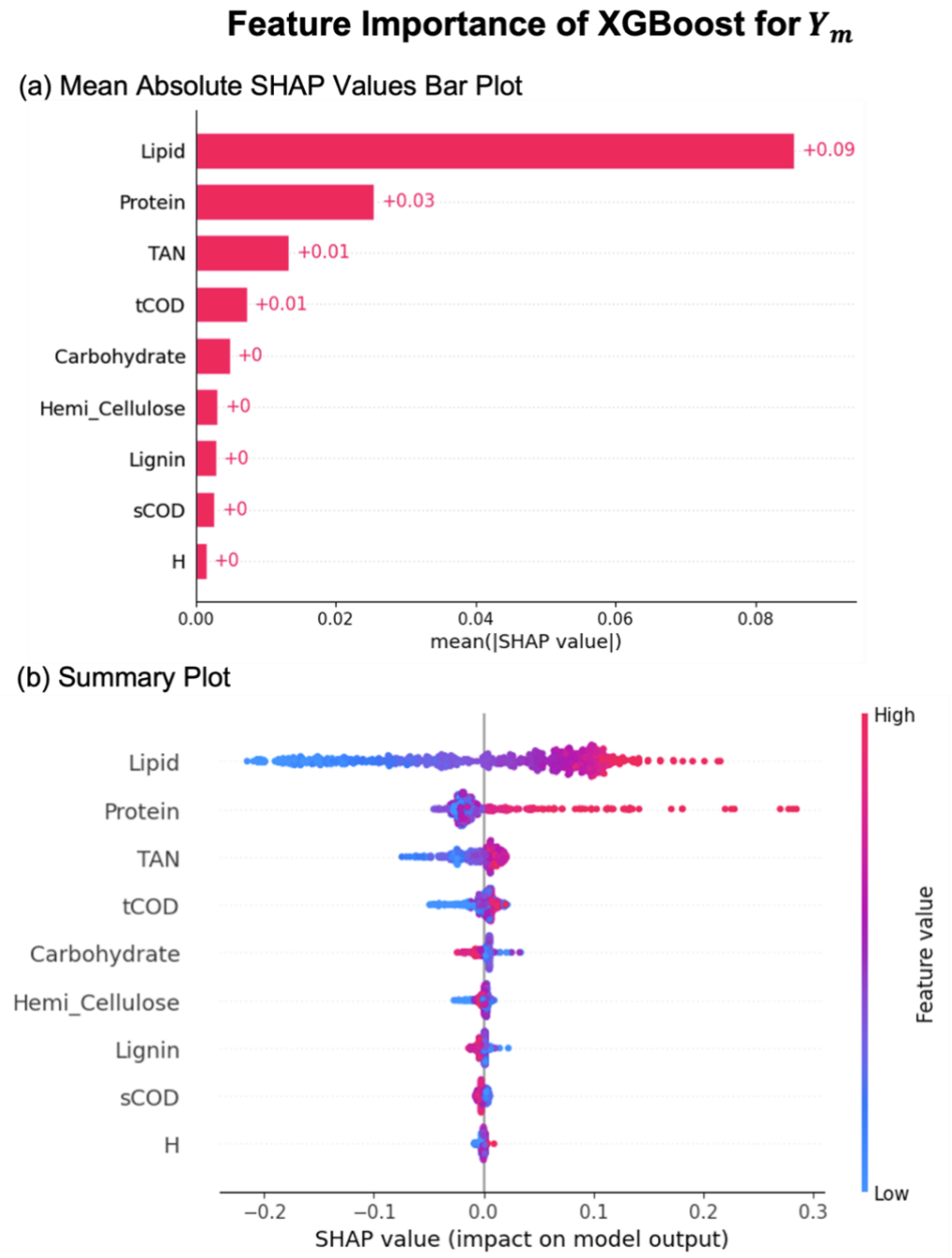


Figure 4-18 Feature importance of the XGBoost model for Y_m ('Hemi_Cellulose' represents the sum of the contents of cellulose and hemicellulose).

Then 'XGBoost_red' models for Y_m prediction with only three predictors were examined. Three scenarios were tried out. The R^2 scores and RMSE of these three models on test set are displayed in Table 4-9. 'XGBoost_red_3' with 'Lipid', 'Protein', 'Non-fibre Carbohydrate' as

predictors has the highest R^2 score and the smallest RMSE, it was thus chosen to predict Y_m.

Table 4-9 The R^2 and RMSE of 'XGBoost_xs' model and 'XGBoost_red' models for parameter Y_m prediction on test set with different combinations of three features.

	Features			R^2	RMSE
XGBoost_xs	All 9 features			0.999	0.010
XGBoost_red_1	Lipid	Protein	TAN	0.992	0.013
XGBoost_red_2	Lipid	Protein	tCOD	0.991	0.014
XGBoost_red_3	Lipid	Protein	Non-fibre Carbohydrate	0.993	0.012

4.3 Summary and Discussion

The unsatisfactory predictive performance and the unreliable feature importance analysis results of the ML regression model for the batch AD test result prediction trained with limited data were discussed in Chapter 3. For this reason, a way to circumvent this problem by generating additional data for ML model training via response surface design was proposed in this chapter.

The framework of the additional dataset generation was illustrated in Figure 4-1. First of all, 10 BMP co-digestion tests of three substrates were carried out following the ASCD. First-order kinetic model was fitted to the biogas production data, with parameters k and Y_m obtained. Then, special cubic response surface equations taking the fractions of substrate (A) and (O) in the mixture as independent variables were fitted to the ten k and Y_m obtained respectively. Consequently, the k and Y_m of a first-order kinetic equation that describes the biogas production process of any mixture of these three substrates can be inferred from the response surface equations obtained. And the corresponding physiochemical characteristic of the mixture can be calculated as a weighted sum of the characteristics in three base substrates assuming no reaction occurred. To this end, an additional dataset for supervised ML model training and testing with the physiochemical characteristics of a mixture sample as input features and the inferred k and Y_m as responses was generated.

The reliability of the additional data generated this way depends on two factors. Firstly, the validity of the assumption that physiochemical characteristics considered in this study of the mixture sample can be calculated as a weighted sum of those in the three base substrates. Secondly, the k and Y_m parameters calculated from the fitted response surface equations are close to their 'real' values if derived from biogas production data fitting. In other words, how well the response surface equations obtained represents the real response surface. The influence of the second factor was investigated by the LOO validation study. It was deducted that the representativeness of the response surface equation will improve with more samples available for the equation fitting. Despite this, further improvement can be pursued in future studies by leaving out separate testing samples to examine the influence of these two factors.

With the generated additional 1000 data and the original 10 data, XGBoost models were trained to infer parameters k and Y_m from nine physiochemical characteristics of a mixture sample. The R^2 of these two XGBoost models on the test set were 0.997 and 0.999 respectively.

By applying SHAP value method for feature importance analysis to the trained XGBoost models, three the most important features for k and Y_m prediction respectively were screened out. In summary, they are 'Non-fibre carbohydrate', 'Protein', 'Lipid' and 'Lignin'. The R^2 of the XGBoost models with three important features for k and Y_m prediction respectively were still as high as 0.995 and 0.993.

5 Mathematical Interpretation of the XGBoost Model and the General Applicability of the Metamodel on External Data

In Chapter 4, the XGBoost models with three important features (‘XGBoost_red’) trained with the augmented dataset showed high prediction accuracy for the parameter k and Y_m of the first-order kinetic model on the test set left out. However, the general performance of the XGBoost models on the external dataset, and thus the effectiveness of the proposed data augmentation method, still need to be verified.

Meanwhile, although ML models generally outperform traditional linear regression models in multivariate regression tasks, interpreting their predictions in a straightforward manner remains a significant challenge. This has hindered the accessibility of such a ML model to be used as a tool for material realistic BMP prediction for domain experts and industrial operators.

Therefore, in this chapter, an effort was made to interpret the ‘XGBoost_red’ model trained in Chapter 4 mathematically. This was achieved with metamodeling the ‘XGBoost_red’ model with Meijei $\boldsymbol{G}$-function. The metamodel was subsequently tested with data collected from the literature that fell within the 'applicability range' of the XGBoost models, to further assess the effectiveness of the data augmentation method proposed in Chapter 4. Figure 5-1 is the metamodeling framework.

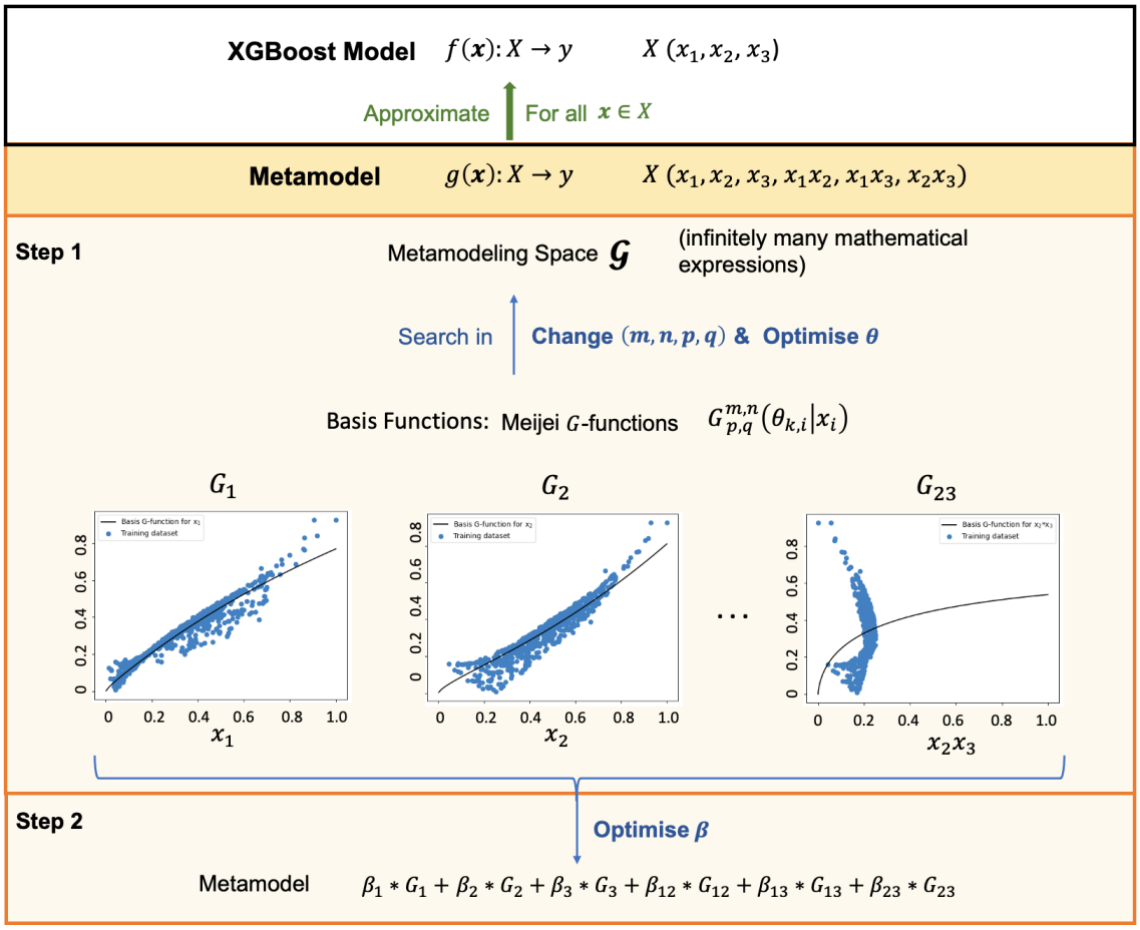


Figure 5-1 Research framework of mathematical metamodeling of XGBoost models.

5.1 Methods for Metamodelling

5.1.1 Meijei G-function Introduction

Meijer G-functions is a family of one variable functions. Many special functions such as Bessel functions and the operations (including products and linear combination of products) of special functions can be expressed by G-functions or products of G-functions. This was demonstrated in Meijer's paper in 1936 (Meijei, 1936).

G-functions is the solution $u(x)$ to the generalised hypergeometric equation (GHGE) (Eq.(5-1).

$$[D\prod_{j=1}^{q-1}(D+\ b_j-1)-x\prod_{j=1}^{p}(D+\ a_j)]\ u(x)=\ 0 \qquad \textbf{(5-1)}$$

in which D is a scale-invariant operator, $D = x\ d/dx$. a_j and b_j are the coefficients of the j_{th} order derivative for the general homogeneous linear ordinary differential equation (Eq. (5-2) and its transformation after multiplying Eq. (5-2 by x^N (Eq. (5-3).

$$a_N(x)\frac{d^N(u)}{dx^N}(x)+\ a_{N-1}(x)\frac{d^{N-1}(u)}{dx^{N-1}}(x)+\cdots+\ a_0(x)u(x)=0 \qquad \textbf{(5-2)}$$

$$b_N(x)D^Nu(x)+\ b_{N-1}(x)D^{N-1}u(x)+\cdots+\ b_0(x)u(x)=0 \qquad \textbf{(5-3)}$$

The solution to GHGE can be represented in a continuous form by the contour integral defined in (0,1) (Eq. (5-4), which is also known as Mellin-Barnes integral (Barnes, 1908; Mellin, 1910).

$$u(x)=\ \frac{1}{2\pi i}\int_L \phi(s)x^s ds \qquad \textbf{(5-4)}$$

Meijer introduced a notation to $u(x)$, the solution of GHGE. The original solution can be expressed by $G^{0,p}_{p,q}\left(^{a_1,\ldots,a_p}_{b_1,\ldots,b_q}|x\right) = \frac{1}{2\pi i}\int_L \frac{\prod_{k=1}^{p}\Gamma(1-a_k+s)}{\prod_{j=1}^{q}\Gamma(1-b_j+s)}x^s\ ds$. On the basis of this original solution, there exists a family of solutions, which are $G^{m,n}_{p,q}\left(^{a'_1,\ldots,a'_p}_{b'_1,\ldots,b'_q}|(-1)^{m+n+p}x\right)$, $0 \le m \le q$, $0 \le n \le p$, where $a'_1,\ \ldots, a'_p$ are permutations of the original $a_1,\ \ldots, a_p$, and the same for $b'_1,\ \ldots, b'_q$. The upper index m and n indicate that the first m factors in the denominator in the original integrand have been changed to $\Gamma(b_j - s)$ in the numerator, and the last $p - n$ factors in the numerator have been changed to $\Gamma(a_k - s)$ in the denominator. The contour integral function for $G^{m,n}_{p,q}\left(^{a_1,\ldots,a_p}_{b_1,\ldots,b_q}|x\right)$ is shown in Eq. (5-5, with the restriction that $a_k - b_j \neq 1, 2, 3$ for $k = 1, 2, \ldots n$ and $j = 1, 2, \ldots m$. The values of $\boldsymbol{a_p} = (a_1, \ldots, a_p)$, $\boldsymbol{b_q} = (b_1, \ldots, b_q)$ together with m and n determined the positions of poles and zeros of the integrand.

$$G_{p,q}^{m,n}\left(\begin{matrix}a_1,\dots,a_p\\ b_1,\dots,b_q\end{matrix}|x\right)=\frac{1}{2\pi i}\int_L \frac{\prod_{j=1}^{m}\Gamma(b_j-s)\prod_{k=1}^{n}\Gamma(1-a_k+s)}{\prod_{j=m+1}^{q}\Gamma(1-b_j+s)\prod_{k=n+1}^{p}\Gamma(a_k-s)}x^s\,ds \quad \textbf{(5-5)}$$

When choosing a contour for the integration of the line integral, the poles need to lie on one side of L as required by the translation condition when constructing the integration contour. Also, the contour L needs to separate poles that go to $-\infty$ and that go to $+\infty$.

There are three types of contour L can be considered: (1) $L = L_I$: L goes from $-\infty$ to $+\infty$. (2) $L = L_\infty$: L is a loop that starts at $+\infty$ on a line parallel to the +ve real axis and returns to $+\infty$ on another line parallel to the +ve real axis. (3) $L = L_{-\infty}$: L is a loop that starts at $-\infty$ on a line parallel to the -ve real axis and returns to $-\infty$ on another line parallel to the -ve real axis (Gradshteyn and Ryzhik, 2014).

There are in total 2^{p+q} solutions for a GHGE with fixed p and q values, but some of them might be the same when the integrand maintains the same while permutating the indices $a_1,\dots,a_p$ and $b_1,\dots,b_q$. If $p<q$, 2^p solutions with $m=0$ are identical. If $p=q$, solutions with $m=n=0$ are identical.

See the article of Beals & Szmigielski (2013) for a general introduction to Meijei G-functions.

5.1.2 Metamodelling with Meijei G-function

Metamodel is a model of a model. It can be an abstraction of a model representing mathematical or algorithm relation between input and output (Garitselov et al., 2012). In the context of this chapter, it pertains to expressing a 'black-box' ML model such as neural network or gradient boosting by concise mathematical functions that are more comprehensible to domain researchers. That is, in this study, the input of the metamodel is the trained XGBoost model with three input features for the prediction of parameter k or Y_m of first-order kinetic model, $f(\boldsymbol{x})$: $X \rightarrow y$, obtained in Chapter 4, and the output is a more interpretable mathematical function $g(\boldsymbol{x})$. $g(\boldsymbol{x})$ approximates $f(\boldsymbol{x})$ for all $\boldsymbol{x} \in X$.

The methodology of metamodeling ML models by Meijer G-functions brought up by Alaa & Schaar (2019) was adopted in this study. The ability of Meijer G-functions to represent a class of arithmetic, algebraic and closed-form expressions, including familiar functions such as polynomial, exponential, logarithm, trigonometric etc., and special functions such as Bessel functions, with different real-valued parameters setting was harnessed in this study. The metamodel is parameterised in the form $G(\boldsymbol{x};\boldsymbol{\theta})$, $\boldsymbol{\theta}$ is the real-valued parameters of the metamodel. With different setting of $\boldsymbol{\theta}$, the G-function reduces to different functions. In other words, the metamodel is searching in the metamodeling space $\mathcal{G}$ which contains infinitely many mathematical expressions. This is also the key difference between metamodeling with G-functions and symbolic regressions. No pre-defined functional forms are constrained to choose from when metamodelling with Meijei G-functions. Even so, to facilitate the practical implementation, the complexity of metamodeling space $\mathcal{G}$ needs to be limited by controlling the parameters p and q of G-functions in Eq. (5-5 instead of searching in an infinitely complex space. The detailed implementation in Python will be discussed in the next section. The goal of the metamodeling is looking for the $\boldsymbol{\theta}^*$ that minimise the loss between $g(\boldsymbol{x})$ and $f(\boldsymbol{x})$. In such a way, the crux of the metamodeling problem lies in the parameter optimisation.

Metamodelling with G-functions comprises of two steps (Alaa and Schaar, 2019). The first step is to decompose the multivariate metamodel $g(\boldsymbol{x})$ into univariate basis functions. According to *Kolmogorov superposition theorem* (Kolmogorov, 1956), any multivariate continuous function $z = F(x_1, x_2, \dots, x_n)$ can be adequately replaced by $z = \sum_{k=1}^{2n+1} \Phi_k \left[\sum_{i=1}^{n} \Phi_{k,i}(x_i)\right]$, where Φ_k and $\Phi_{k,i}$ are continuous univariate basis functions and $n \in \mathrm{N}$. The second step is to parameterise each basis function by G-functions. Therefore, the metamodel $G(\boldsymbol{x};\boldsymbol{\theta})$ can be expressed as Eq. (5-6.

$$g(\boldsymbol{x}) = G(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{r} G_{p,q}^{m,n}(\theta_k | \sum_{i=1}^{d} G_{p,q}^{m,n}(\theta_{k,i} | x_i)) \qquad \textbf{(5-6)}$$

For the practical implementation, the outer basis function $\Phi_k = G_{p,q}^{m,n}(\theta_k|\boldsymbol{x})$ is set to the identity function and $r = 1$ and the interaction terms $x_{ij} = x_i * x_j$ are included to introduce extra explanation ability of $g(\boldsymbol{x})$. A ridge regression is fitted to basis functions $\Phi_{k,i}(x_i) = G_{p,q}^{m,n}(\theta_{k,i}|x_i)$. Therefore, the metamodel is reduced to $G(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{i=1}^{d} \beta_i * G_{p,q}^{m,n}(\theta_i|x_i) + \sum_{ij=1}^{d(d-1)/2} \beta_{ij} * G_{p,q}^{m,n}(\theta_{ij}|x_{ij})$. The parameters (m, n, p, q) is referred as '*order*' of the G-function in the rest of the chapter, which are the hyperparameters for the metamodel. And β_i, $\beta_{i,j}$, θ_i, θ_{ij} are parameters need to be optimised.

5.1.3 Implementation Detail in Python

The 'PySymbolic' package in Python developed by Alaa and Schaar (2019b), which is a wrapper over SymPy aimed for symbolic metamodeling was used in this study for Meijer G-function evaluation. The '*optimize.minimize'* method in SciPy (1.12.0) for Python with conjugate gradient descent algorithm was used in this study for optimising the parameter $\boldsymbol{\theta}$ of metamodel $G(\boldsymbol{x};\boldsymbol{\theta})$. The features of the XGBoost model for k or Y_m prediction were reduced to three in Chapter 4. Thus, there are six variables in total as inputs of the metamodel including interaction terms. They are denoted as x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3. Specifically, the metamodel of the XGBoost model for k prediction has variables: 'Lipid' (x_1), 'Lignin' (x_2), 'Protein' (x_3), 'Lipid*Lignin' (x_1x_2), 'Lipid*Protein' (x_1x_3) and 'Lignin*Protein' (x_2x_3). And the metamodel of the XGBoost model for Y_m prediction has variables: 'Lipid' (x_1), 'Protein' (x_2), 'Non-fibre carbohydrate' (x_3), 'Lipid*Protein' (x_1x_2), 'Lipid*Non-fibre carbohydrate' (x_1x_3) and 'Protein*Non-fibre carbohydrate' (x_2x_3).

Metamodelling the trained XGBoost model with G-functions includes the following steps.

(1) G-function '*order'* search for single dimension

Before fitting a ridge regression to basis functions of all six dimensions to form a metamodel, basis G-function for each dimension needs to be decided first. This was discovered in such a way: applying a candidate G-

function on the studied dimension to fit the predictions of the XGBoost model, the loss $\left|G_{p,q}^{m,n}(\theta_i|x_i) - f(\boldsymbol{x})\right|$, which was measured by RMSE, was expected to decline while optimising the parameter θ_i by gradient descent algorithm. The 'minimum loss' achieved by this G-function with θ_i^* was recorded. Different candidate G-functions were tried out to search for the G-function with the smallest 'minimum loss' on this dimension.

As discussed previously, the sum of p and q determines the complexity of the solution space of a G-function by controlling the possible number of kernels of the contour integral, while m, n, $\boldsymbol{a_p}$, $\boldsymbol{b_q}$ determine the actual positions of poles and zeros of the contour integral. In this study, the searching of the best basis G-function for each dimension was limited to G-functions with $p, q \leq 3$ and $p + q \leq 4$ to ensure a reasonable degree of the complexity of the metamodel. Excluding the G-functions with $m = n = 0$ and G-functions with $m = 0$ when $p \leq q$, the solutions of which reduce to 0 identically, there are in total 45 different candidate G-functions to try out. Their *order* are listed in Table 5-1.

The searching of the G-function for each dimension with the smallest 'minimum loss' was repeated for 20 times. Then, the G-function with the highest frequency to have the smallest 'minimum loss' achieved with θ_i^* on a dimension was chosen as the basis function for this dimension.

Table 5-1 The constrained metamodeling space $\mathcal{G}$ in the practical implementation.

G-function hyperparameters								
p = 0 q = 1	$G_{0,1}^{1,0}$							
p = 0 q = 2	$G_{0,2}^{1,0}$	$G_{0,2}^{2,0}$						
p = 0 q = 3	$G_{0,3}^{1,0}$	$G_{0,3}^{2,0}$	$G_{0,3}^{3,0}$					
p = 1 q = 0	$G_{1,0}^{0,1}$							
p = 1 q = 1	$G_{1,1}^{0,1}$	$G_{1,1}^{1,0}$	$G_{1,1}^{1,1}$					
p = 1 q = 2	$G_{1,2}^{1,0}$	$G_{1,2}^{1,1}$	$G_{1,2}^{2,0}$	$G_{1,2}^{2,1}$				

p = 1 q = 3	$G_{1,3}^{1,0}$	$G_{1,3}^{1,1}$	$G_{1,3}^{2,0}$	$G_{1,3}^{2,1}$	$G_{1,3}^{3,0}$	$G_{1,3}^{3,1}$		
p = 2 q = 0	$G_{2,0}^{0,1}$	$G_{2,0}^{0,2}$						
p = 2 q = 1	$G_{2,1}^{0,1}$	$G_{2,1}^{0,2}$	$G_{2,1}^{1,0}$	$G_{2,1}^{1,1}$	$G_{2,1}^{1,2}$			
p = 2 q = 2	$G_{2,2}^{0,1}$	$G_{2,2}^{0,2}$	$G_{2,2}^{1,0}$	$G_{2,2}^{1,1}$	$G_{2,2}^{1,2}$	$G_{2,2}^{2,0}$	$G_{2,2}^{2,1}$	$G_{2,2}^{2,2}$
p = 3 q = 0	$G_{3,0}^{0,1}$	$G_{3,0}^{0,2}$	$G_{3,0}^{0,3}$					
p = 3 q = 1	$G_{3,1}^{0,1}$	$G_{3,1}^{0,2}$	$G_{3,1}^{0,3}$	$G_{3,1}^{1,0}$	$G_{3,1}^{1,1}$	$G_{3,1}^{1,2}$	$G_{3,1}^{1,3}$	

(2) Loss of the Metamodel

With the G-functions selected in the last step as basis functions for each dimension and with θ_i fixed at $\theta_i{}^*$, the metamodel was formed by fitting a ridge regression (α = 0.01 and intercept = 0) to the basis functions of the six dimensions. The coefficients β_i and β_{ij} of the ridge regression model was estimated by minimising the metamodel loss (RMSE), $|G(\boldsymbol{x};\boldsymbol{\theta}) - f(\boldsymbol{x})|$.

5.1.4 Metamodel Taylor Approximation

Solving the line integrals of Meijei G-function requires extensive and specific mathematical knowledge. Also, as the values of p and q of the *order* of a G-function increase, the expression of a G-function can be as complicated as confluent hypergeometric functions. Therefore, although the metamodel built using Meijei G-functions has significantly advanced the mathematical interpterion of black-box ML models, it remains relatively inaccessible in many application domains. This limitation can hinder its effectiveness as a more interpretable mathematical framework compared to ML models such as XGBoost. Therefore, the metamodel was further simplified with Taylor expansion about zero. The order of the Taylor expansion was chosen with a goal that the Taylor approximated metamodel could still model the prediction of the 'XGBoost_red' on the test set with R^2 above 0.9.

5.2 Data Source for Metamodel General Applicability Validation

From the results in Chapter 4, the features needed by XGBoost models for BMP test result prediction via inferring the parameters k and Y_m of the first-order kinetic model are 'Lignin', 'Non-fibre Carbohydrate', 'Protein', and 'Lipid'. The general applicability of the metamodel was tested on the data collected from the literature. The data collected satisfied the following four requirements:

(1) The 'Lignin', 'Non-fibre Carbohydrate', 'Protein', and 'Lipid' of the sample fell within the corresponding ranges of these chemicals observed in the training samples in this study.

ML regression models typically struggle with extrapolation, meaning that their prediction accuracy may suffer when applied to data with feature values outside the ranges of the features of the training data. These ranges are referred to as the 'applicability range' of the trained ML model. The 'applicability range' of the XGBoost models in this study were organised in Table 5-2.

Table 5-2 The 'applicability range' of the XGBoost models trained in Chapter 4 for the predictions of parameters k and Y_m of the first-order kinetic model.

	Unit	Lower Bound	Upper Bound
Non-fibre Carbohydrate	(g/g VS)	0	0.879
Protein	(g/g VS)	0.073	0.448
Lipid	(g/g VS)	0.005	0.554
Lignin	(g/g VS)	0	0.003 (0.150)

Due to the extremely low lignin contents detected in the substrates in this study, potentially stemming from operational errors or integrity issue with the lab apparatus, the lignin content upper bound was increased to 0.15 g/g VS.

(2) Y_m/ BMP_{theo} ratio is in the range 38.250% - 64.595%.

The satisfactory samples in the step (1) were further screened according to their Y_m/BMP_{theo} ratio, the calculation of which was described in Section 4.1.1.3.

Y_m/BMP_{theo} ratio indicates the discrepancy between the experimental BMP test result and the BMP_{theo} calculated according to Buswell-Muller equation ((2-4). As discussed in Section 4.1.1.3 and Section 4.2.1.2, for samples in this study, due to their very low lignin contents, the Y_m/BMP_{theo} ratio indicates the deficit in the experimentally observed CH_4 yield mainly due to two factors (1) the energy from organic material degradation consumed by microbial growth and maintenance, specific to the inoculum and substrate selections in the study; (2) biogas measurement inaccuracy, specific to the operational procedure and gas measuring unit used in the study. Additionally, minor effects such as incomplete degradation due to insufficient mixing could contribute to this deficit.

The samples used to assess the general performance of the metamodel are expected to come from studies where the extent of the observed methane production deficit relative to BMP_{theo} is similar to that of the samples in this study.

And as reported in Section 4.2.1.2, the Y_m/ BMP_{theo} ratio for the 10 experimental groups in this study ranged from 42.500% - 58.723%. In order to include more samples for metamodel testing, the required Y_m/BMP_{theo} ratio for external samples to be included in the validation dataset was extended by 10% at both ends of this range. Consequently, the samples from step (1) with Y_m/ BMP_{theo} ratio between 38.250% and 64.595% were retained.

(3) The BMP tests are carried out under mesophilic condition (35 - 39 °C).

Table 5-3 organised the collected 13 data that satisfy above three requirements, with their feature values, BMP test result (Y_m), BMP_{theo} and Y_m/ BMP_{theo} ratio shown. They are from three peer-reviewed papers (Study 1: Zheng et al., 2013; Study 2: Labatut et al., 2011; Study 3: Cu et al., 2014). The samples include plants, food wastes, manure and slaughter waste. Not enough satisfactory data

about the parameter k of the first-order kinetic model was able to be collected. Therefore, only metamodel for Y_m prediction was tested.

Table 5-3 The BMP test result samples collected from three peer-reviewed papers that the chemical compositions fall within the 'applicability range' of the metamodels in this study with Y_m/BMP$_{theo}$ ratio in the range 38.250 to 64.595%. Y_m: experimental BMP result. BMP$_{theo}$: Theoretical BMP calculated from Eq. (2-4.

Study	#	Samples Name	Lipid	Protein	Non-fibre carbohydrate	Cellulose + Hemicellulose	Lignin	Y_m	BMP$_{theo}$	Y_m/ BMP$_{theo}$
			g/g VS					mL CH_4/g VS		(%)
1	1	Bamboo leaf	0.051	0.161	0.000	0.741	0.058	204.200	438.799	46.536
	2	Reed	0.012	0.065	0.065	0.761	0.096	184.000	387.355	47.502
	3	Cynodon dactyton	0.033	0.099	0.000	0.823	0.067	218.700	424.041	51.575
	4	Celery	0.085	0.267	0.121	0.470	0.057	253.000	464.138	54.510
	5	Lettuce	0.086	0.366	0.157	0.341	0.050	294.100	475.213	61.888
	6	Watermelon peel	0.067	0.335	0.092	0.465	0.040	265.500	465.554	57.029
2	7	Meat Pasta	0.140	0.193	0.667	0.000	0.000	216.200	514.493	42.022
	8	Cabbage, raw	0.014	0.178	0.446	0.362	0.000	256.500	437.804	58.588
	9	Cow manure	0.050	0.103	0.146	0.558	0.143	222.080	393.986	56.367
	10	Chicken slaughter waste	0.035	0.276	0.148	0.464	0.078	173.180	426.258	40.628

3	11	Dairy cow slaughter waste	0.030	0.153	0.081	0.597	0.140	157.420	387.189	40.657
	12	Pig slaughter waste	0.157	0.191	0.313	0.252	0.087	217.450	488.403	44.523
	13	Grass	0.017	0.107	0.133	0.671	0.072	220.530	404.047	54.580

5.3 Results

5.3.1 Metamodelling of XGBoost for k

First, when using only one variable to fit the prediction of the training set by the XGBoost model with G-function, comparing the achieved minimum loss on $\left|G_{p,q}^{m,n}(\theta_i|x_i) - f(\boldsymbol{x})\right|$ by 45 candidate G-functions with their parameters θ_i optimised by gradient descent, the following G-functions were selected as basis G-function for each dimension as reported in Table 5-4. The θ_i^* and the minimum loss achieved by each basis G-function were also displayed. To ensure result reproducibility, the reporting of θ_i^* values was preserved to eight decimal places.

Figure 5-2 shows the fitting of the selected basis G-function on each dimension. And from the distribution of samples' k values on each dimension, it is observed that the contents of 'Lipid' and 'Protein' in a mixture sample show negative correlation with k value, while 'Lignin' shows positive correlation. This conforms with the feature importance analysis result of the trained XGBoost model for k in section 4.2.4.2.

Then, a ridge regression was fitted to the six basis G-functions to form a metamodel. The coefficient of each dimension was reported in Table 5-5. The metamodel loss, $|G(\boldsymbol{x};\boldsymbol{\theta}) - f(\boldsymbol{x})|$, which is evaluated by RMSE, on test dataset is 0.048. The R^2 of the metamodel on test dataset is 0.938. Compared with the XGBoost model, which has an R^2 value 0.995, the metamodel underperforms to some extent.

To make the metamodel more accessible as a BMP test result prediction tool to domain experts, the basis G-function of each dimension was approximated with Taylor expansion. When the order of Taylor expansion is five, the approximated metamodel had an R^2 = 0.907 and the RMSE = 0.055 on the test set. The prediction accuracy of the Taylor approximated metamodel declined further. Nonetheless, it still has an R^2 value above 0.9. Table 5-5 organised the expression of each dimension after approximation.

Table 5-4 Optimised θ_i and minimum loss of the basis G-function for each dimension of the metamodel of the XGBoost model for k prediction.

	G-functions	θ_i	RMSE
x_1	$G_{0,3}^{3,0}$	b_1 = 0.40023032, b_2 = 0.40023032, b_3 = 0.40023032, c = 1.15418943	0.096
x_2	$G_{3,1}^{0,3}$	a_1= 0.85787098, a_2= 2.15989562, a_3= 0.07531869, b_1 = 0.88040071, c = 1.52835267	0.109
x_3	$G_{1,2}^{2,0}$	a_1 = 1.52566966, b_1 = 0.47433032, b_2 = 0.47433032, c = 1.26840768	0.143
x_1x_2	$G_{0,2}^{1,0}$	b_1 = 0.02893867, b_2 = 0.50123426, c = 1.59547192	0.161
x_1x_3	$G_{0,3}^{3,0}$	b_1 = 0.52183224, b_2 = 0.52183224, b_3 = 0.52183224, c = 1.16271507	0.134
x_2x_3	$G_{2,2}^{1,1}$	a_1 = -0.90630000, a_2 = 1.60058825, b_1 = 0.27187929, b_2 = -1.78865278, c = 1.00312432	0.190
Metamodel Loss (RMSE): 0.048			

Table 5-5 The coefficient of each basis function of the ridge regression and the expression of the approximated metamodel with fifth-order Taylor expansion of the XGBoost model for k prediction.

	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3
Coefficient	0.88280772	0.14921904	-2.77848496	0.48120232	2.25652854	-0.35166380
Constant	0.77846318	0.00208895	0.70784596	0.55786074	0.57921551	0.24364229
x	−2.45775603	−0.21622275	−0.93695949	−1.57139866	−1.4436187	0.80832982
x^2	5.20435815	2.78213170	0.08335516	0.39492611	2.69688892	−1.58047687
x^3	−6.91141967	−2.70727644	1.02765362	0.49029038	−3.35163626	1.81860264

x^4	5.14085153	1.70533717	−1.15622192	−0.46720998	2.39742262	−1.26920261
x^5	−1.63342706	−0.48688472	0.43309286	0.15045158	−0.74265493	0.34617957

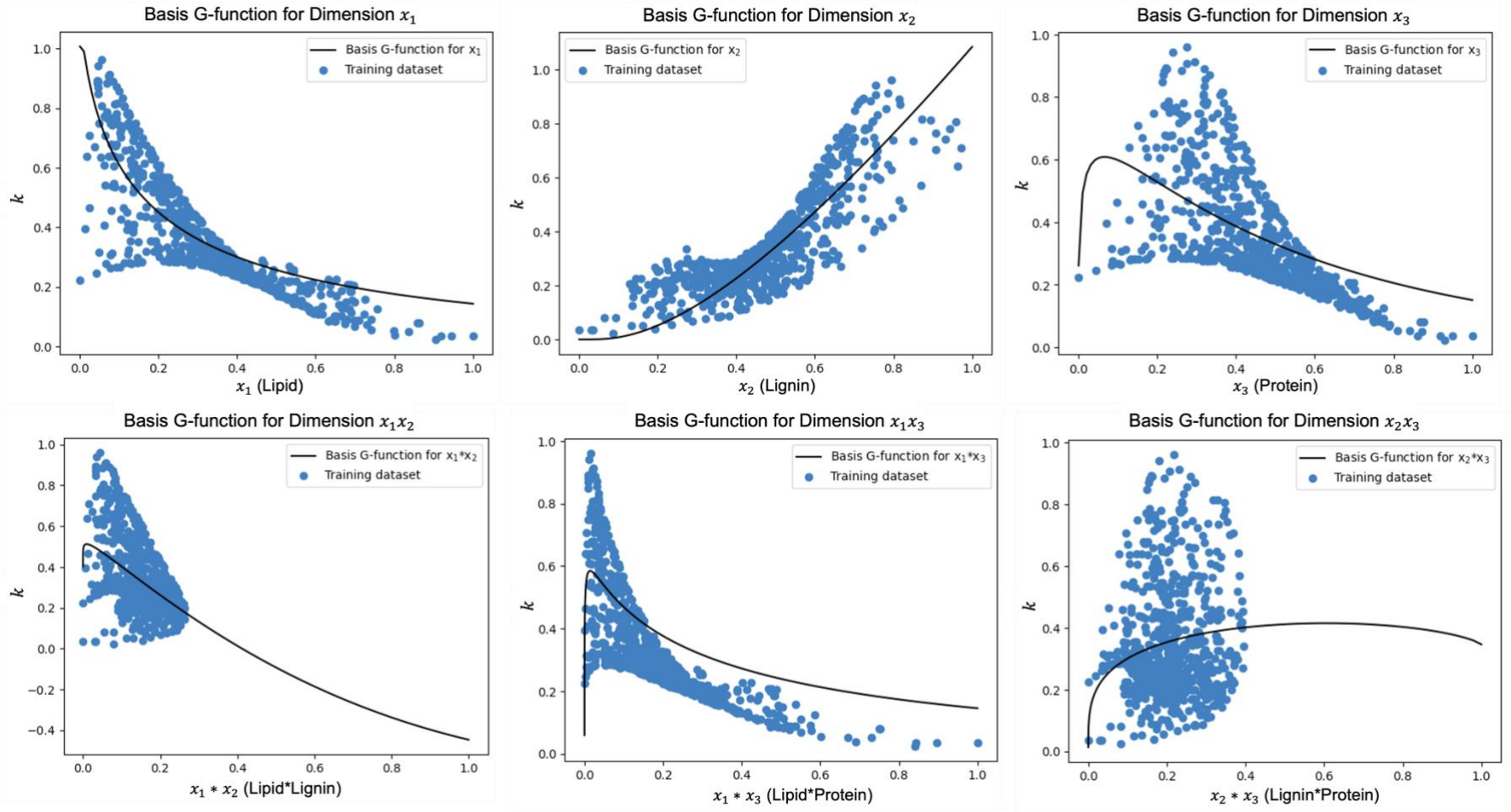


Figure 5-2 The fitting of the selected basis G-function of the metamodel of the XGBoost model for k prediction on each dimension.

5.3.2 Metamodelling of XGBoost for Y_m

First, when using only one variable to fit the prediction results of Y_m by XGBoost model on training set with G-function, the following G-functions were selected as basis functions for each dimension because they had the highest frequency to achieve the smallest 'minimum loss' on $\left|G_{p,q}^{m,n}(\theta_i|x_i) - f(\boldsymbol{x})\right|$ during the 20 times implementation (Table 5-6).

Figure 5-3 shows the fitting of the selected basis G-function to the predictions of Y_m on the training dataset by the XGBoost model on each dimension. Aligning with feature importance analysis results in section 4.2.4.2, the basis G-functions for dimensions x_1 and x_2 depicted the positive correlations of 'Lipid' and 'Protein' with Y_m value. While the basis G-function for dimension x_3 captured the negative correlation between 'Non-fibre Carbohydrate' and Y_m.

The metamodel of the XGBoost model was formed by fitting a ridge regression to six basis functions. The metamodel loss, $|G(\boldsymbol{x};\boldsymbol{\theta}) - f(\boldsymbol{x})|$ measured by RMSE on test dataset is 0.026 and the R^2 of the metamodel on test dataset is 0.967. The coefficients for each dimension of the ridge regression are reported in Table 5-7.

The basis G-function of each dimension was approximated using fifth-order Taylor expansion. Table 5-7 organised the expression of each dimension after approximation. The R^2 of this Taylor approximated metamodel is 0.950 and the metamodel loss (RMSE) is 0.032. In comparison to the XGBoost model, which achieved an R^2 of 0.993, the metamodel's prediction accuracy decreased after fifth-order Taylor approximation. However, it remains capable of predicting Y_m with an R^2 above 0.95.

Table 5-6 Optimised θ_i and minimum loss of the basis G-function for each dimension of the metamodel of the XGBoost model for Y_m prediction.

	G-functions	θ_i	RMSE
x_1	$G_{3,1}^{1,2}$	a_1 = 0.98909744, a_2 = 0.98909744, a_3 = 1.21042255, b_1= 0.78957746, c = 1.05181877	0.0414
x_2	$G_{1,3}^{1,0}$	a_1 = 1.30496388, b_1 = 0.69503611, b_2 = 1.00000000, b_3 = 1.00000000, c = 1.00000000	0.0485
x_3	$G_{0,3}^{3,0}$	b_1 = 0.26742107, b_2 = = 0.26742107, b_3 = = 0.26742107, c = 0.97368052	0.0761
x_1x_2	$G_{1,2}^{2,1}$	a_1 = 1.04088000, b_1 = 0.96038647, b_2 = 0.96038647, c = 1.01351296	0.0416
x_1x_3	$G_{1,3}^{1,1}$	a_1 = 1.17404351, b_1 = 0.76845666, b_2 = 0.82595649, b_3 = 0.82595649, c = 1.18365426	0.1181
x_2x_3	$G_{1,3}^{3,1}$	a_1 = 0.94666279, b_1 = 1.03143680, b_2 = 11.03143680, b_3 = 1.03143680, c = 0.99006099	0.1419
Metamodel Loss (RMSE): 0.0259			

Table 5-7 Coefficient of each basis function of the ridge regression and expression of the approximated metamodel with fifth-order Taylor expansion of the XGBoost model for Y_m prediction.

	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3
Coefficient	0.49159162	0.81456424	0.21299163	-0.44498884	0.44702611	-0.54596872
Constant	0.01376022	0.02520372	1.14197277	0.11733089	0.07296742	0.15392266
x	1.03795780	0.65302792	−4.13547913	1.67319492	2.13030773	1.24598366
x^2	−0.32716280	−0.16666221	9.37003954	−2.76639081	−3.12848385	−2.33589044
x^3	−0.031522193	0.54549749	−12.88436620	3.32296212	2.73414939	2.96023980

x^4	0.13054368	−0.35970186	9.77992347	−2.34992806	−1.68763856	−2.15253390
x^5	−0.05491602	0.11378711	−3.14829071	0.72491716	0.49597920	0.67511559

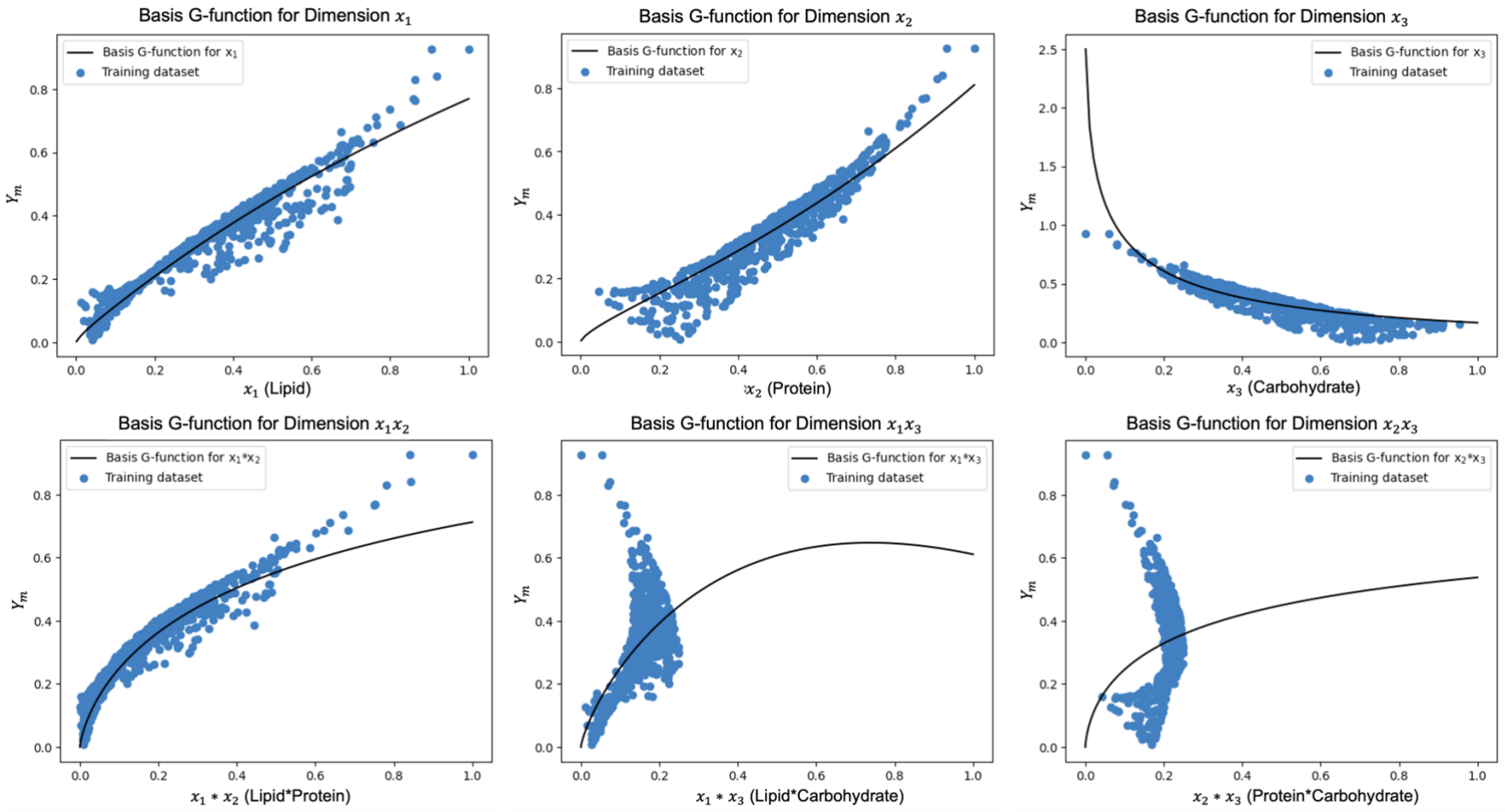


Figure 5-3 The fitting of the selected basis G-function of the metamodel of the XGBoost model for Y_m prediction on each dimension.

5.3.3 General Applicability of Metamodel for Y_m Prediction

Table 5-8 shows the values of the three features of the XGBoost model for Y_m prediction scaled with respect to the 'applicability range' for each feature of the XGBoost model (Table 5-2) with 'MinMaxScaler' ($(X_{scaled} = (X - X_{min})/(X_{max} - X_{min})$) of the 13 external data for metamodel validation and the predicted BMP test result (Y_m_pred) by the Taylor approximated metamodel. To be noticed that the protein content of sample #2 is out of the 'applicability range', but it was still included in the test set as it is only slightly out of the range. The R^2 of the Y_m prediction on this test set is - 0.580, the mean absolute error (MAE) of the prediction is 38.074 and the MAPE of the prediction is 15.424%.

Table 5-8 The scaled feature values of 13 external data for testing the general applicability of the approximated metamodel with fifth-order Taylor expansion for Y_m prediction. Y_m: experimentally derived BMP test result. Y_m_pred: BMP test result predicted by the Taylor approximated metamodel.

Study	#	Samples Name	Lipid	Protein	Carbo	Y_m	Y_m_pred
			g/g VS			mL CH_4/g VS	
1	1	Bamboo leaf	0.082	0.236	0.000	204.200	264.904
	2	Reed	0.011	-0.019	0.074	184.000	205.381
	3	Cynodon dactyton	0.050	0.070	0.000	218.700	238.905
	4	Celery	0.146	0.518	0.137	253.000	276.585
	5	Lettuce	0.146	0.783	0.178	294.100	309.682
	6	Watermelon peel	0.112	0.700	0.105	265.500	308.185
2	7	Meat Pasta	0.245	0.321	0.759	216.200	239.799
	8	Cabbage, raw	0.016	0.281	0.507	256.500	195.527
	9	Cow manure	0.081	0.082	0.166	222.080	216.102
	10	Chicken slaughter waste	0.054	0.541	0.168	173.180	266.959

3	11	Dairy cow slaughter waste	0.044	0.215	0.092	157.420	236.861
	12	Pig slaughter waste	0.276	0.315	0.356	217.450	251.021
	13	Grass	0.021	0.091	0.151	220.530	209.051

This result is compared with three other studies which also tested the general applicability of their models for BMP test result prediction. Triolo et al. (2011) tested the MLR model with the concentrations of lignin and cellulose as predictors on 60 data collected from four papers. The R^2 on the test set was 0.236, the MAE was 82.788. Vazifehkhoran and Triolo (2019) built a MLR model with three variables which categorised organic components into easily, slowly and non-degradable VS. The MAEs of the predictions on four group of external dataset, in total 95 data, were from 37.4 to 86.1. Wang et al. (2023) tested their KNN model with predictors cellulose, hemicellulose, lignin, VS/TS and two combined features on 17 samples collected from other studies, an extraordinary R^2 value of 0.630 was achieved, the MAE was 35.350 and the MAPE was 10.050%.

The negative R^2 score of the Taylor approximated metamodel for BMP test result prediction in this study indicates that the error is smaller if predicting with sample mean. The metamodel seems performed worse compared to the models of Triolo et al. (2011) and Wang et al. (2023) in terms of R^2 score. However, a further investigation into the test sets used in these two studies revealed that the standard deviations of their test set are 94.728 and 99.256 respectively, whereas the standard deviation of the 13 test data in this study is only 38.322. The small standard deviation of this study's test set suggests that these data points deviate less from the sample mean compared to the test data in the other two studies, making predicting with the sample mean less error-prone than predicting with the metamodel. However, MAE and MAPE of the predictions show that the Taylor approximated metamodel in this chapter has similar prediction accuracy for individual samples compared to Wang et al. (2023)'s model.

The small standard deviation of the test set data in this study is because that the chemical compositions of these samples all fall within the 'applicability range' of the trained XGBoost model. This restricted the diversity of characteristics among the samples and thus a less varied BMP test results.

Figure 5-4 shows that the predictions of the BMP test result by the Taylor approximated metamodel have clearly captured the relative magnitudes of the BMP test results of samples in study 1. This might be attributed to the similar inoculum used in study 1 (digestate from an anaerobic digester receiving a mixture of sludge and food waste) as in this study (digestate from an anaerobic digester treating THP sludge). The inocula in both cases were not well acclimated to the new substrates, which may lead to a lower substrate to energy conversion rate, as discussed in Section 4.2.1.2. In contrast, the inoculum used in study 2 was from a digester co-digesting manure and food waste, and the inoculum in study 3 was collected from a digester in a pig farm. Both inocula should have developed acclimatisation to the substrates of their BMP tests.

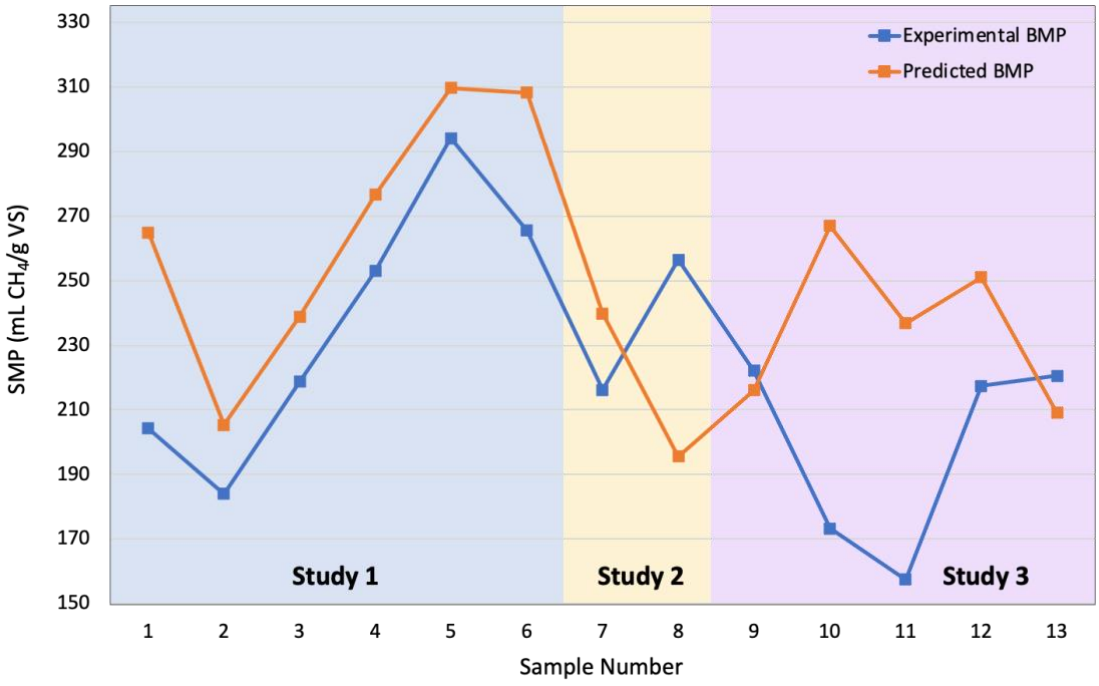


Figure 5-4 The BMP test results of 13 external test data and the BMP results predicted by the fifth-order Taylor approximated metamodel for Y_m prediction.

In summary, the Taylor approximated metamodel, which has a reduced prediction accuracy compared to the XGBoost model, still predicted BMP test results for 13 samples collected from the literature with an MAE = 38.074. In general, the metamodel of the XGBoost model for Y_m prediction trained with the dataset generated from response surface design should have the general applicability for predicting the BMP test result of the samples with similar chemical compositions as the training samples in this study.

5.4 Summary and Discussion

5.4.1 General predictability of the ML model and its metamodel trained with augmented dataset

Meijei G-functions – a class of line integrals the solutions of which include the majority of special functions, were used to build metamodels of the XGBoost models for k or Y_m prediction in this chapter. Metamodel helps to interpret the structure of the black-box ML model by mathematical functions which are more transparent and understandable for domain application. The fifth-order Taylor expansion approximated metamodels of the XGBoost models for k or Y_m prediction have R^2 values of 0.907 and 0.950 respectively on the test set left out from the augmented dataset generated via response surface design in Chapter 4.

Meanwhile, to test the general applicability of the metamodels, which is also the general applicability of the XGBoost models, 13 BMP test result data were collected from the literature. By comparing the MAE and MAPE of metamodel's predictions on the BMP results of the external test set with other three studies which also tested the general applicability of their models on external samples from the literature (Triolo et al., 2011; Vazifehkhoran & Triolo, 2019; Wang et al., 2023), it was concluded that the metamodel in this study had a similar performance as Wang et al. (2023)'s model, which performed the best in three studies. Therefore, sufficient data enabled to train a ML model with general applicability on BMP test result prediction for samples with similar characteristics as training samples of this study. The data augmentation method proposed in

Chapter 4 is believed to have the ability to help overcome the common data limitation problem when training a ML model for BMP test result prediction.

It's worth noting that the choice of three base substrates is important when generating additional dataset with response surface design. This decision should be made from the consideration that the relationship between which physiochemical characteristics and the batch AD test result is to be explored. For instance, in this study, none of the three base substrates contain high concentrations of characteristics that potentially inhibit biogas production, such as TAN and lignin. Consequently, the trained XGBoost model and its metamodel are expected to predict the BMP test result with high accuracy only for samples not prone to inhibited biogas production as well.

By employing three substrates with extreme characteristics – such as high TAN concentration, high C/N ratio, and low alkalinity concentration – the proposed dataset generation method can help to elucidate the effects of these inhibitory factors on biogas production, meanwhile reducing the need for numerous BMP tests.

5.4.2 Interpret Black-box ML Model with Mathematical Expressions

The commonly adopted methods in previous studies for interpreting a trained ML model, such as PDP plots and SHAP values, can only be regarded as feature importance analyses. In contrast, this study developed a metamodel with Meijei G-functions for the trained XGBoost model. The metamodel recovered the prediction of XGBoost model with high precision by mathematical expressions. Such a BMP test result prediction model is much more accessible and ready to be used compared to a ML model.

Meanwhile, the mathematical expression of the metamodel provides a more realistic depiction of how the biomethane potential changed in response to the chemical composition of a sample. As observed in Figure 5-3, the variation of the biomethane potential of a chemical component (lipid, protein and non-fibre carbohydrate) in a sample (mL CH_4/g) with respect to the variation of its fraction in a sample (g/g VS), was not exactly linear. The basis functions of the metamodel

delineated this non-linear relationship flexibly. In contrast, the traditional BMP_{theo} calculation function (Eq.(2-4) has a linear response. In addition, the interaction terms in the model helped further explain the variation in biomethane production with changes of a sample's chemical components. Combining these two advantages, the metamodel provided a satisfactory prediction to the BMP test results for samples within the 'applicability range' of the metamodel.

PART II Performance Prediction of Continuous AD Process

6 Bayesian Calibration of ADM1

The application and reliability of the Anaerobic Digestion Model No.1 (ADM1) for continuous anaerobic digestion simulation is often hindered by the difficulty of reliably calibrating its parameters, especially when data is limited. This chapter presents a Bayesian inference-based framework to address this challenge, enabling ADM1 calibration using only initial-stage digester performance data and accurately predicting steady-state performance. Figure 6-1 illustrates this framework.

A sequence of sensitivity analysis, beginning with the Morris method for pre-screening, followed by the Sobol' method for analysing parameter interactions, and concluding with the hybrid Random Balance Designs – Fourier Amplitute Sensitivity Test to calculate main effect of individual parameter, was applied to identify the most influential model parameters thereby reduce the number of parameters required for model calibration. These parameters are then calibrated using a Bayesian approach with data collected over less than two hydraulic retention times. Prior distributions for the most influential uncertain inputs are determined based on values reported in previous ADM1 studies. Jensen-Shannon divergence is used to assess the practical identifiability of these uncertain inputs. The calibrated model is validated against steady-state digester data. Additionally, the model is calibrated and validated again using the same data but with a uniform prior to examine the influence of an informative prior.

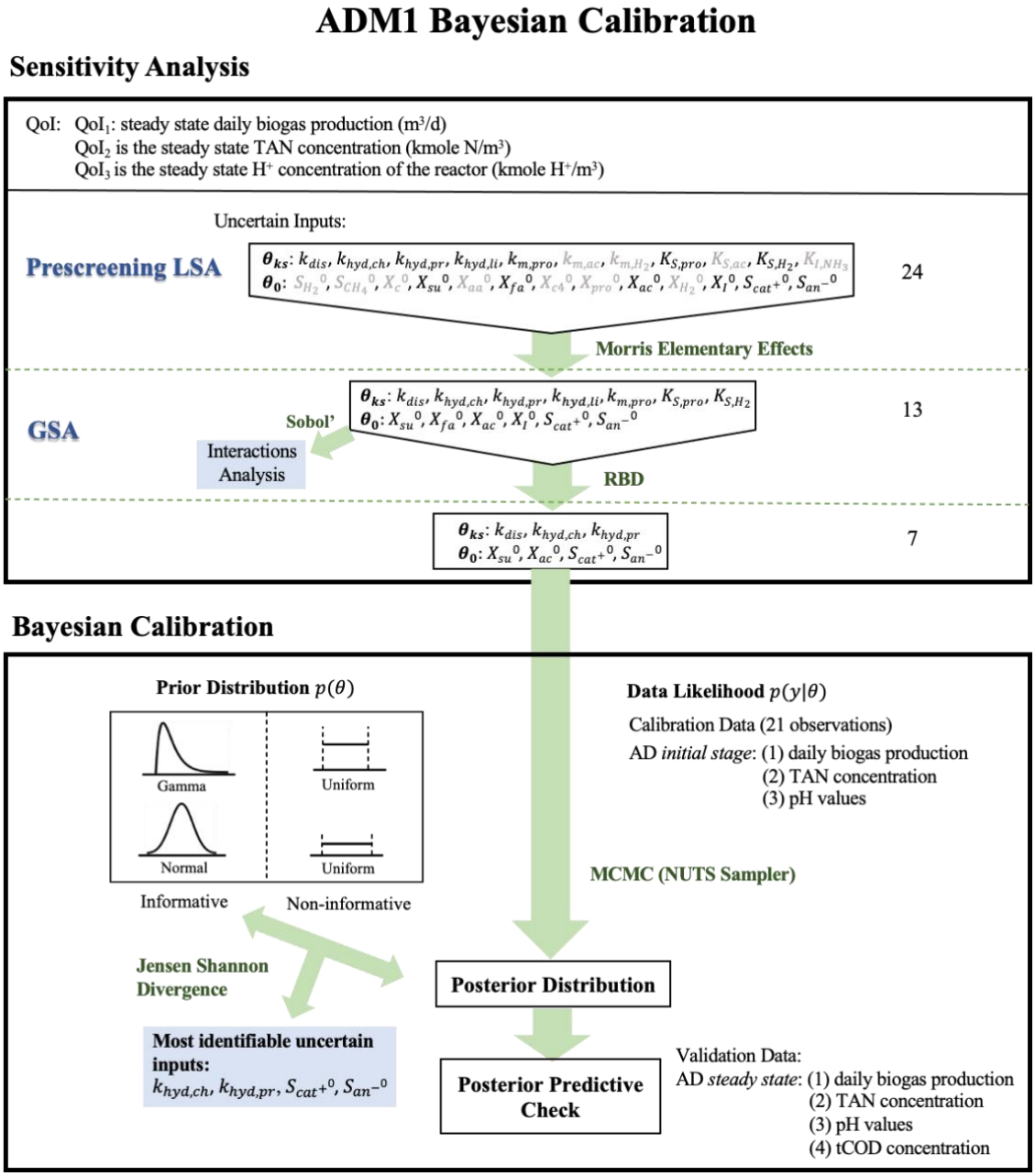


Figure 6-1 Flowchart of the ADM1 Bayesian calibration framework.

6.1 Methods

6.1.1 ADM1

6.1.1.1 ADM1 Model Structure

The ADM1 model used in this study contains 36 state variables (Table 6-2), whose temporal concentration variations are described by 36 ordinary differential equations (ODE). ADM1 model simulates the AD process in continuous AD

reactors without recycling of digestate, assuming that the bulk volume in the digester remains constant over time and inflow equals outflow (Batstone et al., 2002). The ODE for each state variable is in the form as shown in Eq. (6-1.

$$\frac{dS(X)_{i,liq}}{dt} = \frac{q * S(X)_{i,in}}{V_{liq}} - \frac{q * S(X)_{i,liq}}{V_{liq}} + \sum_{j=1-19} \rho_j \nu_{i,j} \quad \textbf{(6-1)}$$

where ρ_j is the kinetic rate for process j; $\nu_{i,j}$ is the stoichiometric coefficient for the state variable i in the reaction process j; q is the inflow and outflow, $S(X)_{i,in}$ and $S(X)_{i,liq}$ are the concentrations of the soluble or particulate state variable i in the inflow and in the reactor, respectively, and V_{liq} is the liquid phase volume in the reactor. The conversions between state variables, including information about ρ_j and $\nu_{i,j}$, are represented by a Petersen Matrix in 7.2Appendix D.

The ODEs for the acid-base kinetic processes were implemented as suggested by Rosen and Jeppsson (2006). A pair of ODEs were used to describe the concentration changes with time of VFAs, inorganic carbon (IC) and inorganic nitrogen (IN). The term $\sum_{j=1-19} \rho_j \nu_{i,j}$ (i = 4, 5, 6, 7, 10, 11) is applied to the ODE of the free form VFA, IC or IN. Another ODE is used to describe the conversion from the free form VFAs, IC and IN to base form VFAs, HCO_3^- and NH_3. This is shown in Eq. (6-2, taking base form valerate (S_{va^-}) as an example.

$$\frac{dS_{va^-}}{dt} = -k_{A/B\ va} \cdot (S_{va^-} \cdot S_{H^+} - K_{a,va} \cdot (S_{va} - S_{va^-})) \quad \textbf{(6-2)}$$

Where S_{va} is the concentration of free form valeric acid, S_{H^+} is the concentration of hydrogen ion, $k_{A/B\ va}$ is the acid-base kinetic constant of valeric acid and $K_{a,va}$ is the acid dissociation constant of valeric acid.

The change of S_{H^+}, i.e., pH value, with time is also represented by an ODE in the ADM1 model in this study (7.2Appendix D, Eq. D-1) as suggested by Thamsiriroj and Murphy (2011). It is deducted from the charge balance between the state variables S_{cat^+}, S_{an^-}, S_{NH4^+}, S_{H^+}, $S_{HCO_3^-}$, S_{va^-}, S_{bu^-}, S_{pro^-}, S_{ac^-} and S_{OH^-} (Eq. (6-3).

$$S_{H^+} + S_{cat^+} + S_{NH_4^+} - S_{HCO_3^-} - \frac{S_{ac^-}}{64} - \frac{S_{pro^-}}{112} - \frac{S_{bu^-}}{160} - \frac{S_{va^-}}{208} - S_{OH^-} - S_{an^-} = 0 \quad \textbf{(6-3)}$$

ADM1 was coded in Python (2022) and solved with SciPy package (2023) using the BDF solver, which is an implicit variable-order method based on backward differentiation formula.

6.1.1.2 Derivation of the Initial Values (θ_0) and Input Values (θ_f) of State Variables

As introduced in Chapter 2, to simulate the AD process of a specific feedstock, in addition to the kinetic and stoichiometric parameters ($\boldsymbol{\theta_{ks}}$) of the model, which needs to be calibrated with experimental observations, the initial values ($\boldsymbol{\theta_0}$) and input values ($\boldsymbol{\theta_f}$) of state variables need to be decided. $\boldsymbol{\theta_0}$ is decided by the characteristics of the inoculum and the $\boldsymbol{\theta_f}$ is decided by characteristics of the feedstock. This section introduces the derivation of $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_f}$.

(1) Inoculum Characteristics and Derivation of θ_0

The inoculum used was the digestate obtained from a full-scale anaerobic digester treating THP pre-treated sludge in a WWTP. A detailed characterisation of the inoculum was carried out in this study (Table 6-1), and Table 6-2 demonstrates how θ_0 was derived from these characteristics.

Table 6-1 Characteristic of the Inoculum.

Characteristics	Unit	Value
TS	% of wet sample	5.240
VS	% of wet sample	3.290
Density	g/mL	1.000
tCOD	kg/m^3	55.467
sCOD	kg/m^3	3.999
Total soluble sugars	g/100g dry matter (DM)	< 0.1 (0.050 adopted)
Crude protein (Dumas Method)	g/100g DM	30.900

Free LCFA	g/100g DM	< 0.5 (0.300 adopted)
Total carbohydrate	g/100g DM	< 1 (0.500 adopted)
Total fat	g/100g DM	9.800
Valeric acid	g/m^3	59.160
Butyric acid	g/m^3	20.080
Propionic acid	g/m^3	42.320
Acetic acid	g/m^3	5.605
TAN	kmole N/m^3	0.190
TC	kmole C/m^3	1.523
TOC	kmole C/m^3	1.487
Lignin	g/100g DM	7.160

The total N (both organic and inorganic) determined by Dumas method in the commercial lab need to be partitioned among inorganic N (nitrite (NO_2^-) and nitrate (NO_3^-)), amino acid ($S_{aa}{}^0$) and protein ($X_{pro}{}^0$). From Song et al. (2019)'s study, NO_2^- and NO_3^- account for 8.54% of total N in waste activated sludge. And for the proportion of amino acid in organic N, the average value from three references were taken (23.4%, Westgate and Park (2010); 25%, Pehlivanoglu and Sedlak's (2004); 4.85%, Song et al. (2019)), which is 17.75%. And thus the left 82.25% organic N was allocated to the protein.

The initial values of S_{gas,H_2}, S_{gas,CH_4}, S_{gas,CO_2} were obtained by manually fitting to the value of the initial daily biogas flow with reference to the default values used in the ADM1 benchmark model (Batstone et al., 2002).

$\boldsymbol{\theta_0}$ of some state variables of ADM1 are unrealistic to be analysed experimentally. They are $S_{H_2}{}^0$, $S_{CH_4}{}^0$, $X_c{}^0$, the concentrations of degraders $X_{bac}{}^0$ ($X_{su}{}^0$, $X_{aa}{}^0$, $X_{fa}{}^0$, $X_{c4}{}^0$, $X_{pro}{}^0$, $X_{ac}{}^0$, $X_{H_2}{}^0$), $X_I{}^0$, $S_{cat^+}{}^0$ and $S_{an^-}{}^0$. For the value of $X_I{}^0$, the content of lignin in the inoculum was used as an approximation. Despite this, there might also be other solid inert materials recalcitrant to AD process. Thus, the exact value for $X_I{}^0$ is still uncertain.

A SA was carried out to identify the ones influential to the model outputs among these uncertain $\boldsymbol{\theta_0}$ and the influential ones were calibrated together with $\boldsymbol{\theta_{ks}}$.

Table 6-2 Initial values of the state variables (θ_0) and how they were derived from the characteristics of the inoculum. The superscript '0' indicates initial values for state variables. 'SA' indicates the uncertain θ_0 to be included in SA process.

No	Name	Description	Unit	Values	Derived from (Estimated by)
1	$S_{su}{}^0$	monosaccharides	kg COD/m^3	0.028	Total soluble sugars
2	$S_{aa}{}^0$	amino acid	kg COD/m^3	0.445	Crude Protein (Dumas)/6.25*(1-8.54%) *17.75%
3	$S_{fa}{}^0$	long chain fatty acid	kg COD/m^3	0.016	Free Long Chain Fatty Acids (LCFA)
4	$S_{va}{}^0$	total valerate	kg COD/m^3	0.006	VFA analysis
5	$S_{bu}{}^0$	total butyrate	kg COD/m^3	0.064	
6	$S_{pro}{}^0$	total propionate	kg COD/m^3	0.037	
7	$S_{ac}{}^0$	total acetate	kg COD/m^3	0.121	
8	$S_{H_2}{}^0$	hydrogen in liquid phrase	kg COD/m^3	**SA**	/
9	$S_{CH_4}{}^0$	methane in liquid phrase	kg COD/m^3	**SA**	/
10	$S_{IC}{}^0$	inorganic carbon	kmole C/m^3	0.036	TIC
11	$S_{IN}{}^0$	inorganic nitrogen	kmole N/m^3	0.190	TAN
12	$S_I{}^0$	soluble inert	kg COD/m^3	3.213	sCOD - S_{su} - S_{aa} - S_{fa} - S_{VFA}(No. 4-7) - S_{H_2} - S_{CH_4}
13	$X_c{}^0$	particulate composites	kg COD/m^3	**SA**	/
14	$X_{ch}{}^0$	carbohydrates	kg COD/m^3	7.898	Total carbohydrate –

					Total soluble sugar
15	$X_{pr}{}^{0}$	proteins	kg COD/m^3	2.044	Crude Protein (Dumas)/6.25*(1-8.54%)*82.25%
16	$X_{li}{}^{0}$	lipids	kg COD/m^3	5.419	Total fat – Free LCFA (No.3)
17	$X_{su}{}^{0}$	sugar degraders	SA	X_{bac} = pCOD –$X_c{}^0$ – $X_{ch}{}^0$ – $X_{pr}{}^0$ – $X_{li}{}^0$ – $X_I{}^0$	$\mu_{max}(X_{su})$ = 0.1 (22.22%)
18	$X_{aa}{}^{0}$	amino acid degraders			$\mu_{max}(X_{aa})$ = 0.08 (17.78%)
19	$X_{fa}{}^{0}$	LCFA degraders			$\mu_{max}(X_{fa})$ = 0.06 (13.33%)
20	$X_{c4}{}^{0}$	valerate and butyrate degraders			$\mu_{max}(X_{c4})$ = 0.06 (13.33%)
21	$X_{pro}{}_{0}$	propionate degraders			$\mu_{max}(X_{pro})$ = 0.04 (8.89%)
22	$X_{ac}{}^{0}$	acetate degraders			$\mu_{max}(X_{ac})$ = 0.05 (11.11%)
23	$X_{H_2}{}^{0}$	hydrogen degrader			$\mu_{max}(X_{H_2})$ = 0.06 (13.33%)
24	$X_I{}^{0}$	particulate inert	kg COD/m^3	3.971 (**SA**)	Lignin
25	$S_{cat^+}{}_{0}$	cation	kmole/m^3	**SA**	/
26	$S_{an^-}{}_{0}$	anion	kmole/m^3	**SA**	/
27	$S_{va^-}{}_{0}$	base form valerate	kg COD/m^3	0.00599	pK_a = 4.86 [2]
28	$S_{bu^-}{}_{0}$	base form butyrate	kg COD/m^3	0.06393	pK_a = 4.86 [2]
29	$S_{pro^-}{}_{0}$	base form propionate	kg COD/m^3	0.03696	pK_a = 4.88 [2]
30	$S_{ac^-}{}_{0}$	base form acetate	kg COD/m^3	0.12089	pK_a = 4.76 [2]
31	$S_{HCO_3^-}{}_{0}$	bicarbonate	kmole C/m^3	0.0345	pK_a = 6.35 [1]

32	S_{NH_3} 0	ammonia in liquid phase	kmole N/m^3	0.00647	pK_a = 9.25 (for NH_4^+)[1]
33	S_{gas,H_2} 0	hydrogen in gas phase	kg COD/m^3	1.588e-5	Manual fitting with reference to benchmark value
34	S_{gas,CH_4} 0	methane in gas phase	kg COD/m^3	1.30022	Manual fitting with reference to benchmark value
35	S_{gas,CO_2} 0	carbon dioxide in gas phase	kmole C/m^3	0.01591	Manual fitting with reference to benchmark value
36	$S_{H^+}{}^0$	hydrogen ion	kmole H^+/m^3	1.585e-8	pH

(2) Feedstock Characteristics and Derivation of $\boldsymbol{\theta_f}$

The feedstock of the semi-continuous anaerobic digester under study was composed of a mixture of 60% primary sludge (PS) with 20% waste activated sludge (WAS) after THP pretreatment (at a temperature of 165 °C, pressure of 6 bar, exerted for a time of 30 min) from a small domestic wastewater treatment plant (WWTP) in Milton Keynes, UK.

A Transformer tool in Simulink (Mathwork, 2022) developed by Zaher et al. (2009) was used to derive the $\boldsymbol{\theta_f}$ of the ADM1 from basic characteristics of the THP sludge in this study. Zaher et al. (2009) appreciated the fact that the feedstock composition changes dynamically and the contents of carbohydrates, proteins, lipids and inert should be derived by solving a set of algebraic equations based on the balance of elemental mass, theoretical COD, charge and carbon covalent bonds. The feedstock characteristics required by the Transformer tool and their values measured are listed in Table 6-3.

Table 6-3 Characteristics of the feedstock (Input for the Transformer Tool)

	Unit	Measured Value
TS	kg/m^3	0.985
VS	kg/m^3	0.758
Density	g/mL	1.000

pCOD	kg/m^3	86.021
sCOD	kg/m^3	29.798
VFA	g COD/m^3	6949.100
TOC	g C/m^3	42431.000
TIC	mol HCO_3^-/m^3	101.530
TON	g N/m^3	4917.800
TAN	g N/m^3	536.390
TP	g P/m^3	318.330
OrthoP	g P/m^3	260.000
TA	equ H^+ m^{-3}	63.130

The outputs from the Transformer tool are X_{ch}, X_{pr}, X_{li}, X_I plus the concentrations of the soluble components: S_{su}, S_{ac}, S_{IN}, S_{IC}, S_I, S_{cat^+} and S_{an^-} (Table 6-4).

Table 6-4 ADM1 Input Value (θ_f) derived by Transformer Tool

Soluble				Particulate			
No.	Name	Unit	Value	No.	Name	Unit	Value
1	S_{su}	kg COD/m^3	6.542	13	X_c	kg COD/m^3	0
2	S_{aa}	kg COD/m^3	0	14	X_{ch}	kg COD/m^3	49.391
3	S_{fa}	kg COD/m^3	0	15	X_{pr}	kg COD/m^3	26.542
4	S_{va}	kg COD/m^3	0	16	X_{li}	kg COD/m^3	0
5	S_{bu}	kg COD/m^3	0	17	X_{su}	kg COD/m^3	0
6	S_{pro}	kg COD/m^3	0	18	X_{aa}	kg COD/m^3	0
7	S_{ac}	kg COD/m^3	23.256	19	X_{fa}	kg COD/m^3	0
8	S_{H_2}	kg COD/m^3	0	20	X_{c4}	kg COD/m^3	0
9	S_{CH_4}	kg COD/m^3	0	21	X_{pro}	kg COD/m^3	0
10	S_{IC}	kmole C/m^3	0.105	22	X_{ac}	kg COD/m^3	0
11	S_{IN}	kmole N/m^3	0.038	23	X_{H_2}	kg COD/m^3	0
12	S_I	kg COD/m^3	1.050	24	X_I	kg COD/m^3	9.046
25	S_{cat^+}	kmole/m^3	0.063				
26	S_{an^-}	kmole/m^3	0.142				

6.1.1.3 Experimental Observations for ADM1 Calibration and Validation

The laboratory scale semi-continuous anaerobic digester of this study has a working volume of 5.2 L and headspace 3.22 L, operated at 39.5 °C. The VS content of the THP sludge sampled at different time fluctuated around 0.758 kg/m³ with a variation of ±11.819%. The feed flow was adjusted to maintain a stable organic loading rate (OLR) of 3 g VS/L/d. The feed flow varies around 0.206 L/d ± 11.483%. This study concentrates on modelling the process that the digester reaches the steady state in the long run with a stable OLR. Thus, the average feed VS content of 0.758 kg/m³ and average feed flow of 0.206 L/d were used in the simulation. And the average hydraulic retention time (HRT) of this anaerobic digester is 25.243 d.

Experimental observations of daily biogas flow (m^3/d) and pH, TAN (kmol N/m^3), and tCOD (kg/m^3) of the digestate from a period of 313 days' operation were used for ADM1 calibration and validation. The biogas flow was recorded automatically by gas counter. However, gas counter broke down from day 11 to day 76 with no data recorded. Also, a biogas flow drop from day 233 to day 258 can be observed because of an absence of feeding to reactor during a holiday break. Besides these two long intervals with data absent, there were a few times minor gas counter failure in a duration from three days to 13 days. The mean of these data is 0.008, and the standard deviation is 0.002. All the data are within three standard deviations, so no data were considered as outlier and removed. To help better visualise the trend in the highly dispersed biogas yield data and reduce the impact on the distribution of gas production data caused by multiple gas counter failure, 10-day moving average of the daily biogas flow were calculated.

The pH value of digestate was measured at the beginning and the end of this period. The TAN of the digestate was measured at irregular intervals throughout this period. The tCOD of digestate was measured at the end of this period.

Since the aim of this part of the study is to demonstrate a way to calibrate ADM1 model with limited initial-stage data and to predict the steady state performance of the digester with calibrated model. The data before the reactor reaches a steady state should be used to calibrate the model, and the steady state data are

used for model validation. Considering the fact that a continuous anaerobic digester under stable operation typically reaches its steady state within three to four HRTs (Peces et al., 2021; Razaviarani & Buchanan, 2015; Song et al., 2016), data from no more than two HRTs were used for ADM1 calibration in this study.

Specifically, the scheme of dividing these observations between calibration data and validation data is as follow: in total 21 experimental observations were used to calibrate the model. They are (1) the biogas flow data of the first 10 days, (2) the TAN of digestate recorded in the first 50 days, and (3) the pH value of digestate recorded in the first 50 days. And the rest of the data, including all the tCOD data, were left for model validation.

6.1.2 Sensitivity Analysis

6.1.2.1 Source of Uncertainty and Quantities of Interests ($\boldsymbol{QoIs}$)

As introduced in Chapter 2, k_{dis}, k_{hyd}, k_m, K_S, K_{I,NH_3} were typically considered as sensitive $\boldsymbol{\theta_{ks}}$ in the literature. In this study, 11 kinetic parameters k_{dis}, $k_{hyd,ch}$, $k_{hyd,pr}$, $k_{hyd,li}$, $k_{m,pro}$, $k_{m,ac}$, k_{m,H_2}, $K_{S,pro}$, $K_{S,ac}$, K_{S,H_2}, K_{I,NH_3}, and plus 13 uncertain $\boldsymbol{\theta_0}$ including ${S_{H_2}}^0$, ${S_{CH_4}}^0$, ${X_c}^0$, ${X_{bac}}^0$ (${X_{su}}^0$, ${X_{aa}}^0$, ${X_{fa}}^0$, ${X_{c4}}^0$, ${X_{pro}}^0$, ${X_{ac}}^0$, ${X_{H_2}}^0$), ${X_I}^0$, ${S_{cat^+}}^0$, ${S_{an^-}}^0$, were considered for the SA as explained in section 6.1.1.2. Considering the large number of uncertain parameters, a preliminary screening was done before variance-based GSA to reduce computational demands. The uninfluential ones were then fixed at their default values while the influential ones were further analysed by GSA, a practice commonly referred to as factor fixing (Saltelli et al., 2007).

Three $QoIs$ were chosen based on the experimental observations available for model calibration. QoI_1 is the steady state daily biogas production (m^3/d), QoI_2 is the steady state TAN concentration (kmole N/m^3) and QoI_3 is the steady state H^+ concentration of the digestate (kmole H^+/m^3).

6.1.2.2 Pre-screening – Morris Elementary Effect

The Morris elementary effect (EE) method was used for factor fixing (Morris, 1991; Campolongo et al., 2007). Different from conventional one-at-a-time LSA

which keeps the remaining uncertain inputs fixed while changing the value of one input, Morris EE method generates various combinations of the remaining uncertain inputs, shifting their values across k levels and thus introducing extra randomness compared to one-at-a-time LSA.

The variability ranges of d uncertain inputs were transformed into uniform distributions in [0,1], forming a d-dimentional unit hypercube Z_d. The sample points were generated by r sampling trajectories of $d+1$ points in the unit hypercube, each trajectory providing d EE_s, one for each uncertain input. Thus, the total sample size for Morris method is $r(d+1)$. One EE_i of an uncertain input $\theta_i \in \boldsymbol{\theta}$ can be calculated according to Eq. (6-4:

$$EE_i = \frac{y(\theta_1, \dots, \theta_i + \delta, \dots, \theta_d) - y(\theta_1, \dots, \theta_i, \dots, \theta_d)}{\delta} \quad \textbf{(6-4)}$$

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_i, \dots, \theta_d)$ is a randomly selected point in the hypercube Z_d,

$\delta \in \left\{\frac{1}{k-1}, 1-\frac{1}{k-1}\right\}$, k is the shifting level of each input, $y(\boldsymbol{\theta})$ is the function brings the input to output.

The r EE_i of the uncertain input θ_i form an empirical distribution F_i, where $EE_i \sim F_i$. The mean (μ_i^*) of the distribution of r EE_i in absolute value ($|EE_i|$) were used as metrics of sensitivity (Morris, 1991; Campolongo et al., 2007). The greater the value μ_i^* of one uncertain input, the greater influence it has on the model output.

1,000 random samples were drawn for SA ($r = 1000$), which is in total 25,000 model evaluations for 24 uncertain inputs. The variability ranges of 11 uncertain $\boldsymbol{\theta_{ks}}$ were determined according to their values reported in studies applying ADM1 to raw or THP sludge (including PS and WAS and a mixture of two) over the past 15 years (Table 6-5), with the smallest value as lower bound and the biggest value as upper bound. The variability ranges for uncertain $\boldsymbol{\theta_0}$**,** including ${S_{H_2}}^0$, ${S_{CH_4}}^0$, ${S_{cat^+}}^0$, ${S_{an^-}}^0$ were determined in the same way (Table **6-6**).

The variability ranges of ${X_c}^0$ and ${X_{bac}}^0$ were determined by an allocation of pCOD. Based on the pCOD of the inoculum (Table 6-1), and the COD equivalent

contents of carbohydrate, protein, lipid and lignin (Table 6-2), the remaining 32.135 kg COD/m^3 was distributed between $X_c{}^0$ and $X_{bac}{}^0$. Assuming an equal contribution from both, 16.068 is designated as the mean of the distribution of $X_c{}^0$. In this scenario, $X_c{}^0$ comprises 18.2% of the tCOD of the inoculum. It aligns with the values found in other studies, which are from 10% to 15% (Galí et al., 2009; Razaviarani and Buchanan, 2015). Then, the uniform distribution of $X_c{}^0$ was set as [6.068, 26.068] by extending 10 from both sides of the mean.

The allocation of COD among different degraders adopts the method suggested by Souza et al. (2013b), wherein it is allocated proportionally to the maximum growth rate (μ_{max}) of each microbial community as displayed in Table 6-2.

In summary, Table 6-7 organised the variability ranges of 24 uncertain inputs for Morris EE prescreening SA and also the following variance-based GSA. The goal of the pre-screening Morris EE SA is to reduce the number of uncertain $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ advancing further to variance-based GSA by half.

Table 6-5 . The calibrated values for kinetic and stoichiometric parameters k_{dis}, $k_{hyd,ch}$, $k_{hyd,pr}$, $k_{hyd,li}$, $k_{m,pro}$, $k_{m,ac}$, k_{m,H_2}, $K_{S,pro}$, $K_{S,ac}$, K_{S,H_2}, K_{I,NH_3} in papers applying ADM1 to AD of raw and THP sludge in the past 15 years. '/' means the author did not calibrate corresponding parameters but keep them at their default values as suggested in the benchmark model (Batstone et al., 2002).

Authors	Substrate	k_{dis}	$k_{hyd,ch}$	$k_{hyd,pr}$	$k_{hyd,li}$	$k_{m,pro}$	$k_{m,ac}$	k_{m,H_2}	$K_{S,pro}$	$K_{S,ac}$	K_{S,H_2}	K_{I,NH_3}
Batstone et al. (2002)	Sewage sludge	0.5	10	10	10	13	8	35	0.1	0.15	7e-6	0.0018
Souza et al. (2013a)	WAS	0.24	2.38	4.42	1.49	/	/	/	/	/	/	/
Souza et al. (2013b)	WAS+PS	0.56	0.51	0.44	0.42	/	/	/	/	/	/	/
Blumensaat & Keller (2005)	WAS + PS	1	/	/	/	9	9	/	0.2	0.15	7e-6	0.0011
Ozgun (2019)	WAS + PS	0.25	/	/	/	6	5	20	0.2	0.25	9e-5	0.0023
Zhou et al. (2020)	WAS	0.2	/	/	/		12	/	/	0.15	/	/
Ersahin (2018)	PS	0.67	6.37	/	8.23	10	4.57	25	0.2	0.18	/	/
Otuzalti & Perendeci (2018)	WAS + PS	0.101	10	10	9.99		/	/	/	/	/	/
Aldin et al. (2011)	PS	/	0.3	0.05	0.09	15	8	/	0.1	0.1	/	/
	WAS	/	0.1	0.03	0.05	20	11	/	0.1	0.1	/	/
Phothilangka et al., (2008)	WAS	0.25	/	/	/	/	/	/	/	/	/	/
Razaviarani and Buchanan (2015)		0.47	/	/	/	12	7.4		0.408	0.233	/	/

Batstone et al., (2009)	***THP WAS***	/	0.15-1.5	/	/	/	/	/	/	/	/	/
Donoso-Bravo et al. (2020)	***THP WAS + PS***	/	0.18-0.24	/	/	/	/	/	/	/	/	/
Ortega-Martinez et al. (2016)	***THP WAS+PS***	/	0.092	/	/	/	/	/	/	/	/	/
Souza et al. (2013a)	***THP WAS + PS***	5.6	0.72	/	/	/	/	/	/	/	/	/
Souza et al. (2013b)	***THP WAS***	2.57	0.66	/	/	/	/	/	/	/	/	/
Phothilangka et al. (2008)	***THP WAS***	1.5	/	/	/	/	/	/	/	/	/	/

Table 6-6 The initial values used for $S_{H_2}{}^0$, $S_{CH_4}{}^0$, $S_{cat^+}{}^0$, $S_{an^-}{}^0$ of ADM1 in literatures which have reported the initial values used in the past 15 years.

Authors	**Inoculum**	$S_{H_2}{}^0$	$S_{CH_4}{}^0$	$S_{cat^+}{}^0$	$S_{an^-}{}^0$
Rosen and Jeppsson, (2006)	Not mentioned	2.359e-7	0.055	0.04	0.02
Nguyen (2014)	digestate from an anaerobic digester treating sugar beet pulp	4.2e-7	0.051	0.0247	0.0659
Thamsiriroj and Murphy (2011)	Not mentioned	0	0.01	0.01	0.6
Pastor-Poquet et al. (2019)	Not mentioned (centrifuged)	1e-8	1e-8	0.04-0.166	0.02-0.073

Table 6-7 The Uniform Distributions of uncertain inputs for SA.

$\boldsymbol{\theta_{ks}}$			
Name	**Description**	**Unit**	**Range**
$\boldsymbol{k_{dis}}$	Disintegration rate	d^{-1}	(0.101, 5.6)
$\boldsymbol{k_{hyd,ch}}$	Carbohydrate hydrolysis rate	d^{-1}	(0.092, 10)
$\boldsymbol{k_{hyd,pr}}$	Protein hydrolysis rate	d^{-1}	(0.03, 10)
$\boldsymbol{k_{hyd,li}}$	Lipid hydrolysis rate	d^{-1}	(0.05, 10)
$\boldsymbol{k_{m,pro}}$	Maximum uptake rate of propionate degraders	d^{-1}	(6, 20)
$\boldsymbol{k_{m,ac}}$	Maximum uptake rate of acetate degraders	d^{-1}	(4.57, 12)
$\boldsymbol{k_{m,H_2}}$	Maximum uptake rate of H_2 degraders	d^{-1}	(20, 35)
$\boldsymbol{K_{S,pro}}$	Half saturation coefficient of propionate uptake	kg COD/m^3	(0.1, 0.2)
$\boldsymbol{K_{S,ac}}$	Half saturation coefficient of acetate uptake	kg COD/m^3	(0.1, 0.25)
$\boldsymbol{K_{S,H_2}}$	Half saturation coefficient of H_2 uptake	kg COD/m^3	(7×10^{-6}, 9×10^{-5})
$\boldsymbol{K_{I,NH_3}}$	50% inhibitory concentration of free NH_3 to acetate uptake	kg COD/m^3	(0.0011, 0.00234)
$\boldsymbol{\theta_0}$			
$\boldsymbol{S_{H_2}{}^0}$	See Table 6-2		(0, 4.2e-7)
$\boldsymbol{X_c^0}$			(6.068, 26.068)
$\boldsymbol{X_{aa}{}^0}$			(1.079, 4.635)
$\boldsymbol{X_{c4}{}^0}$			(0.809, 3.457)
$\boldsymbol{X_{ac}}$			(0.674, 2.896)
$\boldsymbol{X_I{}^0}$			(3, 20)
$\boldsymbol{S_{an^-}{}^0}$			(0.02, 0.6)
$\boldsymbol{S_{CH_4}{}^0}$			(1e-8, 0.055)
$\boldsymbol{X_{su}^0}$			(1.348, 5.792)
$\boldsymbol{X_{fa}{}^0}$			(0.809, 3.475)
$\boldsymbol{X_{pro}}$			(0.539, 2.317)

$\boldsymbol{X_{H_2}{}^0}$			(0.809, 3.457)
$\boldsymbol{S_{cat^+}{}^0}$			(0.01, 0.166)

6.1.2.3 GSA of Screened Uncertain Inputs

Though Morris method provides an efficient pre-screening of the influential uncertain inputs, its results are less reliable as the majority of the input space is left unexplored (Saltelli et al., 2019). Variance-based GSA, which provides a more detailed and quantitative analysis of input sensitivities, capturing both individual effects and interactions comprehensively across the entire input space, was thus adopted to further analyse the sensitivity of the more influential $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ screened out by Morris EE.

(1) Sobol' Indices – Interactions between Uncertain Inputs

Sobol' method (Sobol', 2001; Saltelli, 2002) is the most popular variance-based GSA method in research due to its explicitness of explanation. In Sobol' method, the variability range of each uncertain input θ_i is also scaled to [0,1]. The random variable for model output, $f(\boldsymbol{\theta})$, has mean f_0 and variance D while evaluating random samples drawn from the input space (Eq. (6-5). Sobol' method decomposes D into variance introduced by each uncertain input θ_i and interaction terms (Eq. (6-6).

$$f_0 = \int f(\boldsymbol{\theta})\, d\theta \qquad D = \int f(\boldsymbol{\theta})^2\, d\theta - {f_0}^2 \tag{6-5}$$

$$D = \sum_{i=1}^{d} D_i + \sum_{i=1}^{d} \sum_{j=i+1}^{d} D_{ij} + \cdots + D_{1\ldots d} \tag{6-6}$$

where D_i is the variance brought by the variation of the uncertain input θ_i, and D_{ij} is the variance brought by the interaction between θ_i and θ_j, and so on.

Then, the first-order (S_i) and second-order (S_{ij}) Sobol' index are defined as (Eq. (6-7):

$$S_i = \frac{D_i}{D} \qquad S_{ij} = \frac{D_{ij}}{D} \qquad \textbf{(6-7)}$$

S_i describes the reduction of D when a target uncertain parameter θ_i is fixed. This is known as the main effect of θ_i in the context of factor prioritisation (Tarantola et al., 2006). S_{ij} represents the contribution to output variance due to parameters interaction. And the total Sobol' index ($S_{Ti} = S_i + S_{ij} + \cdots + S_{1\ldots i\ldots d}$) takes all the variances caused by parameter i into account. When S_{Ti} is significantly bigger than S_i, it indicates that parameter θ_i interacts heavily with other parameters.

The total number of samples needed for estimating S_i and S_{Ti} for d uncertain inputs is $n(d+2)$, where n is the number of samples for each parameter and is usually a power of two (Saltelli, 2002). However, it is recognised that the number of n required to obtain reliable estimation of Sobol' indices increases significantly when dealing with complex non-linear, non-additive models (Jaxa-Rozen & Kwakkel, 2018; Tarantola et al., 2006). Zhang et al. (2015) used n = 50,000 for a pharmacokinetic two-compartmental model with six parameters. Jaxa-Rozen and Kwakkel (2018) found n > 38,000 is needed to obtain a converged result for a flu epidemic model with 17 continuous inputs and two structural switches. Considering the extent of non-linearity of ADM1 and the number of ODEs it contains, n = 2^{15} (32,768) was adopted to calculate Sobol' indices. With only S_i and S_{Ti} were estimated, this is in total 491,520 model evaluations.

However, as a convergent result cannot be guaranteed with this sample size, the GSA result from Sobol' method was mainly used to identify interactions between the more influential $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ screened out by Morris EE method.

(2) Random Balance Design – Fourier Amplitude Sensitivity Test Method for GSA

A GSA method which combines the Random balance designs (RBD) method brought up by Satterthwaite (1959) and the Fourier Amplitude

Sensitivity Test (FAST) method was adopted as a computationally efficient and also reliable method for main effects analysis (Tarantola et al., 2006). This is also a variance-based GSA method. Contrary to the classic FAST method in which each dimension is visited at a different frequency to achieve the effect of space-filling, the RBD method evaluates each dimension at the same frequency. Thereafter, coordinates of these points are permutated to generate scrambled points that fill the space to circumvent the fact that such design points are not space-filling. With permutation, the cost for SA of a model with d uncertain inputs is kept down to n model evaluations using RBD method instead of $n(d+2)$ using Sobol' method.

To access the convergence of the SA indices obtained by RBD method, a criterion presented by Touzani and Busby (2014) was adopted (Eq. (6-8).

$$\kappa_N = \frac{\frac{1}{t}\sum_{s=1}^{t}\|V_n - V_{n-s*\Delta n}\|}{\|V_n\|} \quad \textbf{(6-8)}$$

where V_n is the vector of estimated SA indices with n samples, $V_{n-s*\Delta n}$ is the SA indices estimated with decreasing number of samples with Δn interval. $\|\cdot\|$ is the ℓ2 norm of the vector. The SA result at sample size n is considered as converged when $\kappa_N < 0.05$. In this study, the sample size started from n = 165,000 and gradually increased with Δn = 10,000 until convergence criterion was achieved.

6.1.3 Bayesian Calibration

The most influential uncertain $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ identified by RBD-FAST GSA were calibrated using the Bayesian inference.

Based on Bayes' Theorem (Bayes & Price, 1763), Bayesian calibration comprises of three parts, prior distribution $P(\theta)$, data likelihood $P(y|\theta)$ and posterior distribution $P(\theta|y)$ (Eq. (6-9).

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)} \quad \textbf{(6-9)}$$

The expert knowledge on the possible distribution of model parameters is expressed as $P(\theta)$, and the parameter distributions updated after observing experimental data is $P(\theta|y)$. $P(y|\theta)$ evaluates the goodness of fit between model simulation and observations. Posterior distribution can hardly be solved analytically except for very simple model due to the usually high dimensional integral in the denominator, $P(y)$. Nonetheless, it can be estimated by sampling in the prior distribution space with Markov Chain Monte Carlo (MCMC) chains (van Oijen, 2017). Some mostly used MCMC sampling algorithm are Metropolis-Hastings (Hastings, 1970; Metropolis et al., 1953), Gibbs sampling (Geman & Geman, 1984), and Hamiltonian Monte Carlo (HMC) (Duane et al., 1987).

6.1.3.1 Prior Distribution Determination

In this study, the determination of the prior distributions of the most influential $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ still referred to the values reported in studies that modelled sewage sludge anaerobic digestion with ADM1 over the past 15 years, which are shown Table 6-5 and Table **6-6**. However, there are two differences compared to the determination of the variability range for SA. First, only studies with THP sludge as feedstock were referred to. Because compared to examining the influence of uncertain inputs on the model output across a broader spectrum when doing SA, a more informative prior distribution is needed for Bayesian calibration, considering the observations used in this study are from AD of THP sludge specifically. Secondly, the distributions are no longer the less informative uniform distribution as in SA process. Instead, Normal distribution and Gamma distribution were adopted to express the belief in the high dense area of the prior distribution of uncertain inputs.

6.1.3.2 Data Likelihood

The data likelihood distribution used in this study is standard normal distribution with standard deviation (σ) assigned according to the believes about the noise in the experimental observations. The probability of the discrepancy between model prediction and the observed data is evaluate according to this distribution. The establishment of the data likelihood distribution like this assumes that the noise in the experimental observations is normally distributed.

As the measurement of the three types of experimental observations for model calibration are in different scales, different standard deviations are assigned for their data likelihood distribution (Eq. (6-10(6-11,(6-12).

$$P(y_{gas}|\theta) = \frac{1}{\sqrt{2\pi{\sigma_1}^2}} exp(-\frac{(y_{gas} - \widehat{y_{gas}})^2}{2{\sigma_1}^2}) \ (\sigma_1 = 0.0015) \quad \textbf{(6-10)}$$

$$P(y_{pH}|\theta) = \frac{1}{\sqrt{2\pi{\sigma_2}^2}} \exp(-\frac{(y_{pH} - \widehat{y_{pH}})^2}{2{\sigma_2}^2}) \ (\sigma_2 = 0.2) \quad \textbf{(6-11)}$$

$$P(y_{TAN}|\theta) = \frac{1}{\sqrt{2\pi{\sigma_3}^2}} \exp(-\frac{(y_{TAN} - \widehat{y_{TAN}})^2}{2{\sigma_3}^2}) \ (\sigma_3 = 0.01) \quad \textbf{(6-12)}$$

Where y_{gas}, y_{pH}, and y_{TAN} are experimental observations; $\widehat{y_{gas}}$, $\widehat{y_{pH}}$, and $\widehat{y_{TAN}}$ are the outputs from the model for a specific set of uncertain inputs drawn from the prior distribution.

6.1.3.3 Posterior Distribution and Posterior Predictive Check

The ‘sunode’ package (Seyboldt (2020) was used in this study for ODE system Bayesian calibration. The sampler used for posterior distribution construction is No-U-Turn sampler (NUTS) (Hoffman & Gelman, 2011) in PyMC3 (Salvatier et al., 2016), a python package for Bayesian statistical modelling. NUTS is an extension to HMC sampler. NUTS is free from the user definition of the number of steps and step size. The sampling stops automatically when loop-back steps were detected. Two MCMC chains of 5,000 samples were drawn with 20,000 burn-in samples, from which the posterior distribution was approximated. The convergence of the MCMC chains is assessed by Gelman-Rubin method. When the calculated potential scale reduction factor is close to one, it indicates that the chains have likely converged.

The statistical summary of the marginal posterior distribution for each uncertain input was returned from ‘PyMC3’ package, including mean, standard deviation, ‘hdi_3%’, ‘hdi_97%’, ‘mcse_mean’, and ‘mcse_sd’. The ‘hdi_3%’ and ‘hdi_97%’ represent the credible interval that contains 94% of the posterior distribution's probability mass, indicating the range within which the true parameter value lies

with high probability. And the ‘mcse_mean’ represents the Monte Carlo standard error of the mean, it quantifies the uncertainty in the estimated mean due to the finite number of MCMC samples. Similarly, ‘mcse_sd’ is the uncertainty of the estimated standard deviation for each parameter.

Also, the ‘traceplot’, which is a visualisation of the smoothed distribution of the MCMC samples drawn, was obtained.

To quantify the dissimilarity between the posterior distribution and the prior distribution, Jensen Shannon (JS) distance (Nguyen & Vreeken, 2015) was calculated, which measures the divergence between two probability distributions (Nguyen & Vreeken, 2015). Different from the most common informatic-theoretic divergence measure – Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951), which is only defined when two distributions $p(X)$ and $q(X)$ are absolutely continuous with respect to each other, JS divergence is well defined in the whole domain, $x \in dom(X)$, for $p(X)$ and $q(X)$ by introducing a point-wise mean distribution M of $p(X)$ and $q(X)$ (Eq.(6-13).

$$JS(P \parallel Q) = \sqrt{\frac{KL(P \parallel M) + KL(Q \parallel M)}{2}} \quad \textbf{(6-13)}$$

where P and Q are $p(X)$ and $q(X)$ respectively. $M = \frac{1}{2}(P + Q)$ is the mixture distribution of P and Q. KL is KL divergence.

To facilitate the calculation of the JS divergence, the marginal posterior distribution for each uncertain input was approximated by kernel density estimation method from the MCMC chains samples drawn. The bandwidth was meticulously tuned to ensure the kernel-based probability density function accurately mirrored the underlying distribution of MCMC samples. A python package ‘divergence’ developed by Nowotny (2020) was used to calculate the JS divergence.

It is also important to carry out the posterior predictive check which compares the simulation result from the calibrated model with the validation dataset. This was accomplished by repetitively sampling $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ from the posterior distribution

and running the model (Gelman & Hill, 2007). The model outputs of daily biogas flow, pH, TAN and tCOD from 3,000 sets of $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ were recorded. The mean and 95% credible interval (CI) of 3,000 outputs were compared with validation data to assess the quality of model calibration.

6.1.3.4 The Influence of the Prior Distribution

The effect of adopting informative prior distribution for the Bayesian calibration of ADM1 was explored in this section. Instead of solely referring to the parameter values reported in papers using THP sludge in Table 6-5 and applying Normal and Gamma priors which indicate the high-density region of the distribution, less informative uniform distributions the same as those used in SA section (Table 6-7) were adopted as prior distribution. Two MCMC chains of 5000 samples with 20000 burn-in sample were drawn again to form an estimation of the posterior distribution. And the posterior predictive check was carried out as well.

6.2 Results

6.2.1 Sensitivity Analysis

6.2.1.1 Morris EE Pre-screening

Although the implicit backward Euler solver used to solve ADM1 is stable over much larger zone compared to the forward Euler method, there were still unstable, blown-up solutions when $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ took certain values because ADM1 is a highly non-linear and stiff ODE system (Nimmegeers et al., 2017). In 25,000 model evaluations, the unstable solutions for QoI_1 (steady state daily biogas production, m^3/d) were as high as 10^{131} whilst the unstable solutions for QoI_2 were as low as -10^{233} (steady state TAN concentration, kmole N/m^3). The solution to QoI_3 (steady state H^+ concentration, kmole H^+/m^3) was overall stable compared to the other two.

The magnitude of these unstable solutions was passed to the calculated μ_i^* of all uncertain $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$. To concentrate on the primary impact of $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ on the model outputs in the stable solution zone, the solutions of QoI_1 and QoI_2 were

treated as outliers when their absolute values were larger than 10^5 and discarded before carrying out SA.

The Morris EE SA was implemented ten times. The average μ_i^* of 24 uncertain $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ regarding to QoI_1, QoI_2 and QoI_3 are shown in Figure 6-2, Figure 6-3 and Figure 6-4 respectively. From Figure 6-2, after removing outliers, the μ^* of the distributions of the elementary effects of uncertain $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ for QoI_1 is in the range 3,500 to 5,000 m^3/d. The values of QoI_1 are influenced by the values of ${S_{cat^+}}^0$ and ${S_{an^-}}^0$ the most, followed by $k_{hyd,ch}$ and X_{su}^0. For QoI_2, the μ^* of all the uncertain $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ are in the range of 0 to 4 kmole N/m^3 after removing outliers, with $k_{m,pro}$, $k_{hyd,li}$, $K_{S,pro}$ and ${X_{fa}}^0$ disproportionately influential (Figure 6-3). For QoI_3, the impacts of ${S_{cat^+}}^0$, ${S_{an^-}}^0$, $k_{m,pro}$, K_{S,H_2}, ${X_{fa}}^0$ and ${X_I}^0$ are notably greater compared to other uncertain inputs (Figure 6-4). As observed in Eq. (6-3, ${S_{cat^+}}^0$and ${S_{an^-}}^0$ had direct impact on the value of S_{H^+} solved.

Based on the preliminary SA result by Morris method, around half of the uncertain $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ were selected to be further analysed by variance-based GSA. The most influential factors for QoI_2 and QoI_3 which are highlighted in green in Figure 6-3 and Figure 6-4 accounted for eight distinct uncertain inputs. According to the ranking in Figure 6-2, apart from ${S_{cat^+}}^0$and ${S_{an^-}}^0$ which were already included, the next five more influential uncertain inputs were selected from QoI_1.

As a result, 13 uncertain $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ that were further analysed by GSA are respectively k_{dis}, $k_{hyd,ch}$, $k_{hyd,pr}$, $k_{hyd,li}$, $k_{m,pro}$, $K_{S,pro}$, K_{S,H_2}, ${X_{su}}^0$, ${X_{fa}}^0$, ${X_{ac}}^0$, ${X_I}^0$, ${S_{cat^+}}^0$, and ${S_{an^-}}^0$.

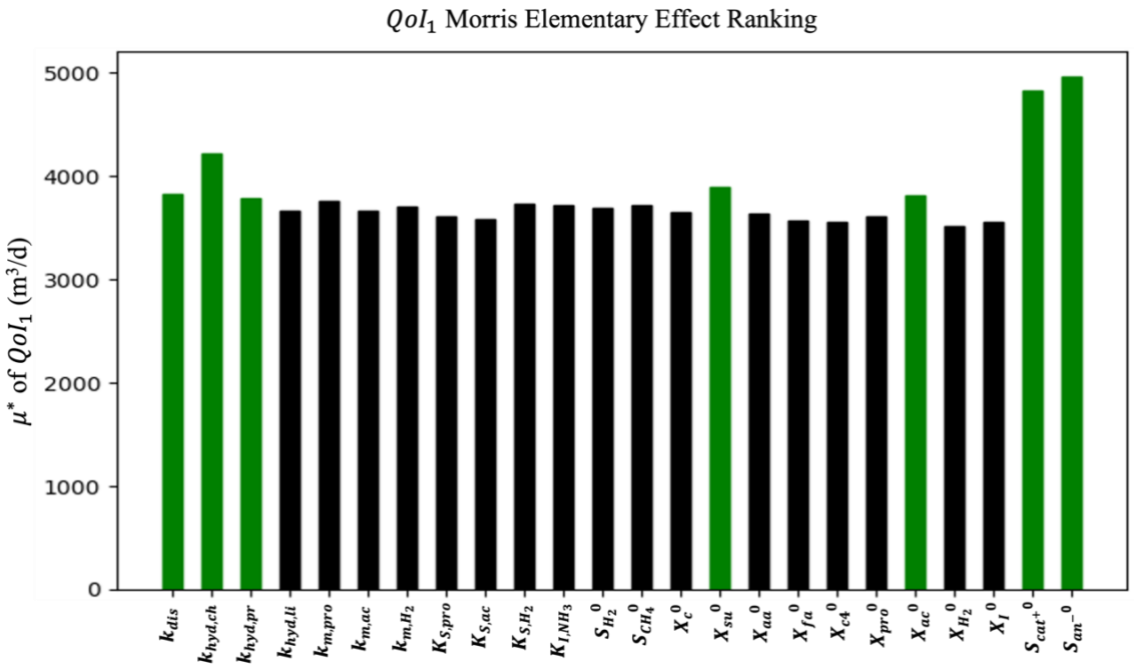


Figure 6-2 The μ_i^* of the distributions of the elementary effects of 24 uncertain θ_{ks} and θ_0 for QoI_1. The green bars are parameters influential to QoI_1.

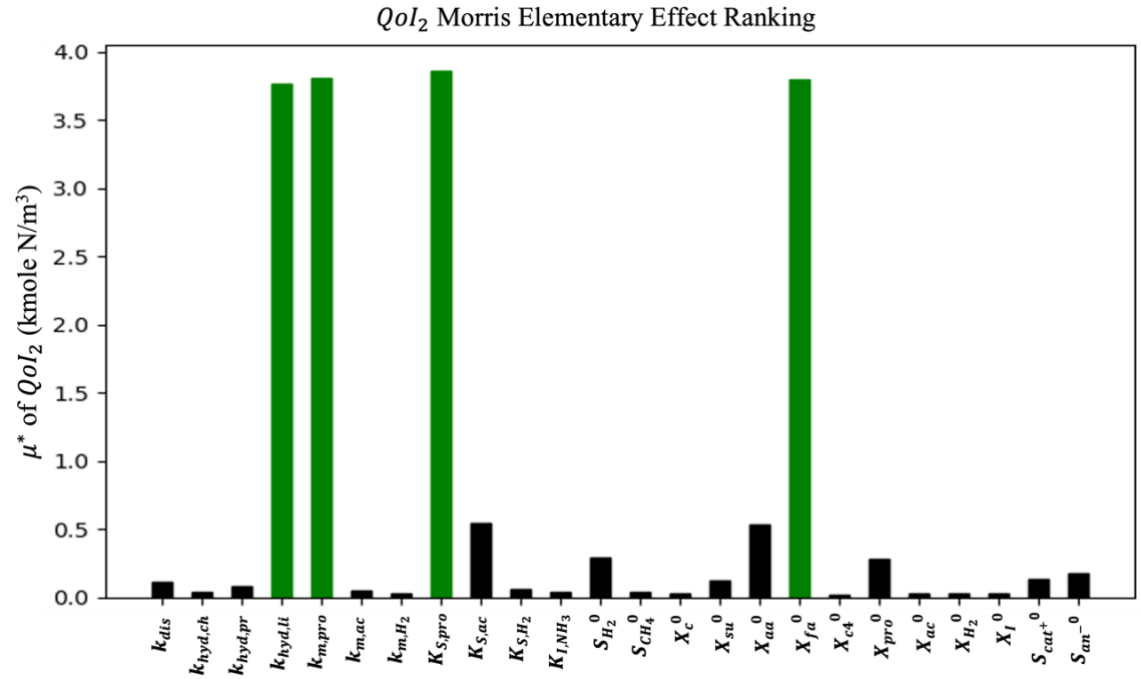


Figure 6-3 The μ_i^* of the distributions of the elementary effects of 24 uncertain θ_{ks} and θ_0 for QoI_2. The green bars are parameters influential to QoI_2.

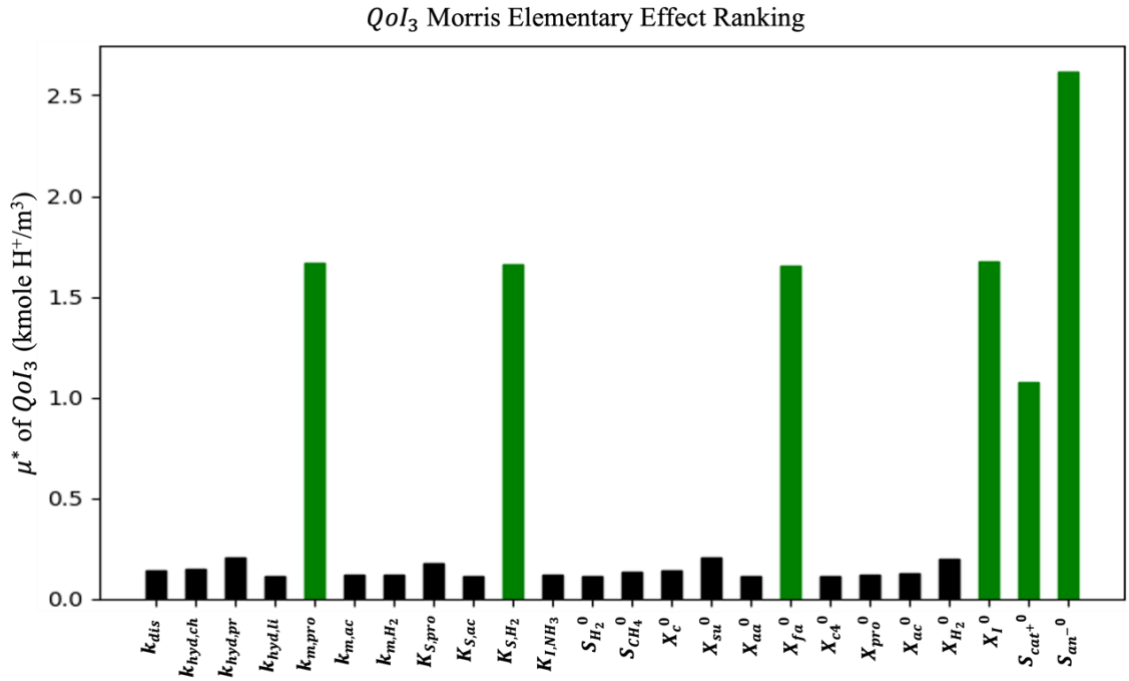


Figure 6-4 The μ_i^* of the distributions of the elementary effects of 24 uncertain θ_{ks} and θ_0 for QoI_3. The green bars are parameters influential to QoI_3.

6.2.1.2 Sobol' Indices – Interactions between Uncertain Inputs

The S_i and S_{Ti} for QoI_1, QoI_2, and QoI_3 of 13 screened out more influential uncertain inputs are listed in Table-6-8. As the S_i score analysed by Sobol' method tend to be inaccurate for the least influential factors (Tarantola et al., 2006), negative S_i score was obtained for some factor. The S_{Ti}- S_i value for such factor was not calculated.

Observing the S_i values, the S_i of ${S_{cat^+}}^0$ and ${S_{an^-}}^0$ are significantly larger compared with other uncertain inputs for all three QoI_s. This is especially true for QoI_2 and QoI_3, which means that the variances of the steady state TAN and H^+ concentrations are mainly caused by the variations of the ${S_{cat^+}}^0$ and ${S_{an^-}}^0$ values.

Besides ${S_{cat^+}}^0$ and ${S_{an^-}}^0$, $k_{hyd,ch}$ is the next influential to the value of QoI_1, while $k_{hyd,pr}$ is the next influential to QoI_2. Compared with the results of QoI_2 by Morris EE method, which screened out $k_{hyd,li}$ as an influential factor, Sobol' result seems more biochemically plausible, as the degradation of protein is more closely associated with the TAN concentration in AD process compared to the

degradation of lipid. Nevertheless, as explained in section 6.1.2.3, with the sample size adopted for Sobol' GSA in this study, the estimated Sobol' indices might have not converged. For this reason, Sobol' results mainly served as a way for examining interactions among inputs.

The disproportionately large S_{Ti} - S_i values of ${S_{cat^+}}^0$ and ${S_{an^-}}^0$ indicate that they interact strongly with other uncertain inputs. Considering the similar magnitude of their S_{Ti} - S_i values and the chemical fact that the concentrations of cation and anion are related, ${S_{cat^+}}^0$ and ${S_{an^-}}^0$ interacted heavily with each other.

However, for QoI_1, the S_{Ti} - S_i scores of all factors were similar high, with ${S_{cat^+}}^0$ and ${S_{an^-}}^0$ slightly higher than others. This means that the values of these factors collectively determine the biogas production level at steady state. This result conforms with the result from Morris EE method for QoI_1 as shown in Figure 6-2, in which all the bars show similar heights except for ${S_{cat^+}}^0$ and ${S_{an^-}}^0$.

Table-6-8 The first-order Sobol' index (S_i) and total Sobol' index (S_{Ti}) of 13 more influential uncertain θ_{ks} and θ_0 for QoI_1, QoI_2 and QoI_3 respectively.

	QoI_1			QoI_2			QoI_3		
	S_i	S_{Ti}	S_{Ti}- S_i	S_i	S_{Ti}	S_{Ti}- S_i	S_i	S_{Ti}	S_{Ti}- S_i
k_{dis}	0.009	0.664	0.654	0.002	0.137	0.135	0.006	0.042	0.036
$k_{hyd,ch}$	**0.026**	0.746	0.720	-0.002	0.155	/	0.007	0.048	0.041
$k_{hyd,pr}$	0.015	0.700	0.685	**0.006**	0.163	0.158	0.007	0.050	0.043
$k_{hyd,li}$	0.013	0.693	0.680	0.002	0.146	0.144	0.008	0.044	0.036
$k_{m,pro}$	0.011	0.693	0.682	0.002	0.155	0.153	0.008	0.046	0.039
$K_{S,pro}$	0.008	0.722	0.715	0.003	0.166	0.163	0.008	0.047	0.039
K_{S,H_2}	0.008	0.685	0.677	0.003	0.150	0.147	0.009	0.046	0.037
X_{su}^0	0.015	0.720	0.706	-0.000	0.154	/	0.007	0.054	0.047
X_{fa}^0	0.017	0.714	0.698	0.000	0.151	0.151	0.007	0.049	0.043
X_{ac}^0	0.013	0.738	0.725	0.002	0.151	0.149	0.007	0.048	0.041
X_I^0	0.007	0.701	0.694	0.001	0.145	0.144	0.008	0.045	0.036
$S_{cat^+}^0$	**0.063**	0.8900	**0.827**	**0.246**	0.741	**0.495**	**0.366**	0.601	**0.236**
$S_{an^-}^0$	**0.094**	0.878	**0.784**	**0.290**	0.717	**0.427**	**0.410**	0.607	**0.197**

6.2.1.3 RBD-FAST GSA

The RBD-FAST method was applied to estimate the main effects of uncertain inputs, offering computational efficiency by achieving converged results with a smaller sample size. The sample size n = 165,000 was first tried out but the convergence criterion $\kappa_N < 0.05$ was not achieved. Then 10,000 more samples were added for each RBD-FAST GSA trail. Finally, $\kappa_N < 0.05$ was achieved when n = 205,000, with κ_N = 0.049, 0.045 and 0.014 for QoI_1 , QoI_2 and QoI_3 respectively. Figure 6-5, Figure 6-6 and Figure 6-7 show the main effect values of the 13 more influential $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ scaled with respect to the largest main effect for QoI_1, QoI_2 and QoI_3 respectively. The y-axis was presented in a logarithmic scale. The GSA results of trails with smaller sample sizes, where the convergence criterion had not yet been achieved, were additionally depicted in the plots for cross-checking and comparison purposes. As outlined by Tarantola et al. (2006), the main effects of the least important factors estimated by the RBD-FAST method tend to be inaccurate, often yielding negative results. Given that negative results cannot be logarithmically transformed, only positive results are displayed, which explains the unconnected dots in these figures. The SA scores of the factor with positive main effects in all five trials are connected by a line for a better visualisation of the level of significance.

$S_{cat^+}{}^0$ and $S_{an^-}{}^0$ are predominantly influential compared to other factors for all three QoI_s. Besides these two, $k_{hyd,ch}$ is the next influential factor for QoI_1 (Figure 6-5), while $k_{hyd,pr}$ is the next influential factor for QoI_2 (Figure 6-6). This aligns with the results from Sobol' method according to the S_i index. $X_{ac}{}^0$ and k_{dis} also exert a clear influence on QoI_1 (Figure 6-5). And X_{su}^0 is influential to both QoI_2 and QoI_3 (Figure 6-6 and Figure 6-7).

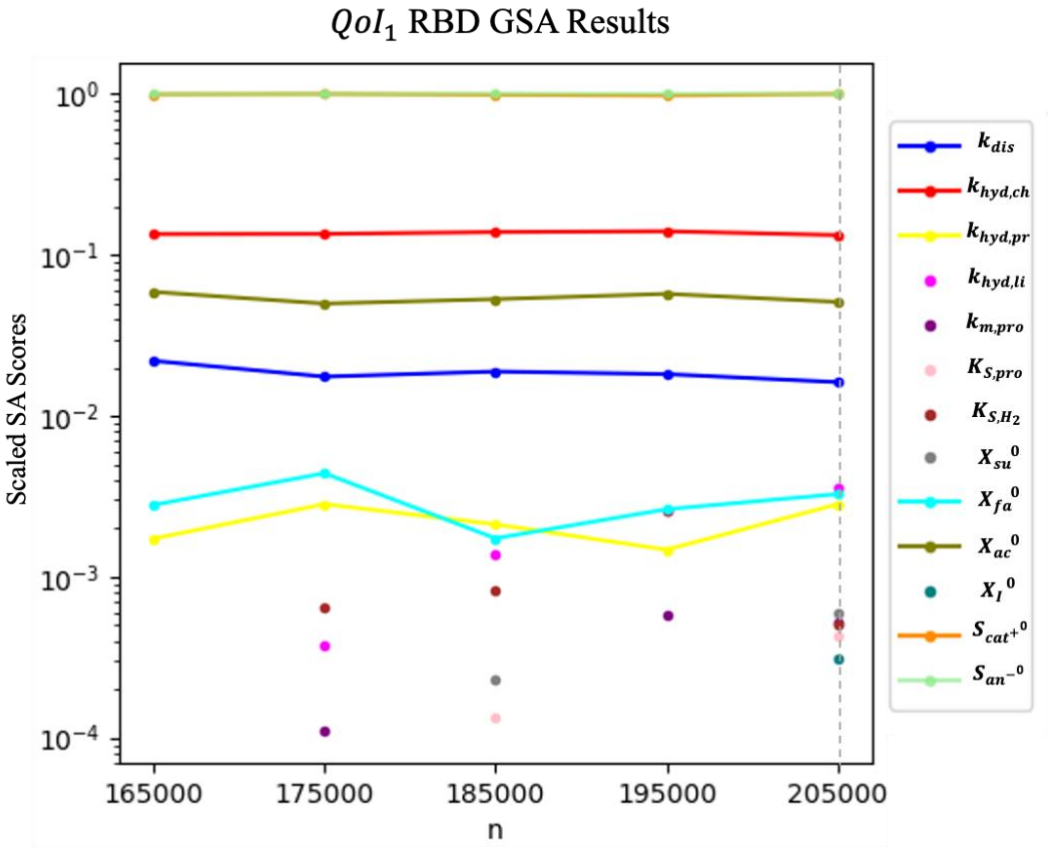


Figure 6-5 The main effects of the 13 more influential θ_{ks} and θ_0 scaled with respect to the largest main effect in five trails with different sample sizes (n) for QoI_1 (y-axis is in logarithmic scale).

Figures 6-4 to 6-6 not only display the scaled SA scores (main effects) of the uncertain $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$, the spacing between lines representing individual factors also signifies the magnitude of the difference in SA scores between two factors. It can be observed that the spaces between the lines for the scaled SA scores of $S_{cat^+}{}^0$ and $S_{an^-}{}^0$ and the line for the next influential factor for QoI_2 (Figure 6-6) and QoI_3 (Figure 6-7) are significantly larger compared to this space for QoI_1. This substantiates the disproportionally predominant impact of $S_{cat^+}{}^0$ and $S_{an^-}{}^0$ on QoI_2 and QoI_3. By contrast, the evenly spaced lines in Figure 6-5 implies that QoI_1 is influenced collaboratively by more factors. This observation conforms with the factors interaction analysis derived from Sobol' indices in Section 6.2.1.2.

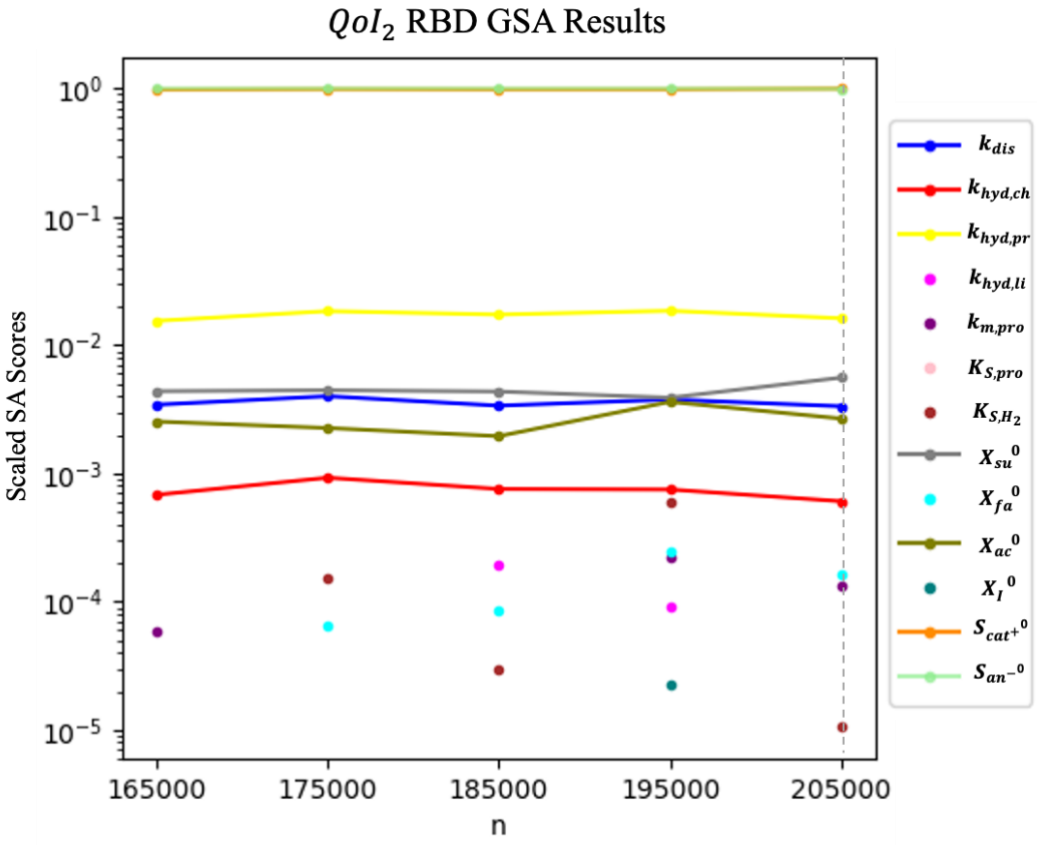


Figure 6-6 The main effects of the 13 more influential θ_{ks} and θ_0 scaled with respect to the largest main effect in five trails with different sample sizes (n) for QoI_2 (y-axis is in logarithmic scale).

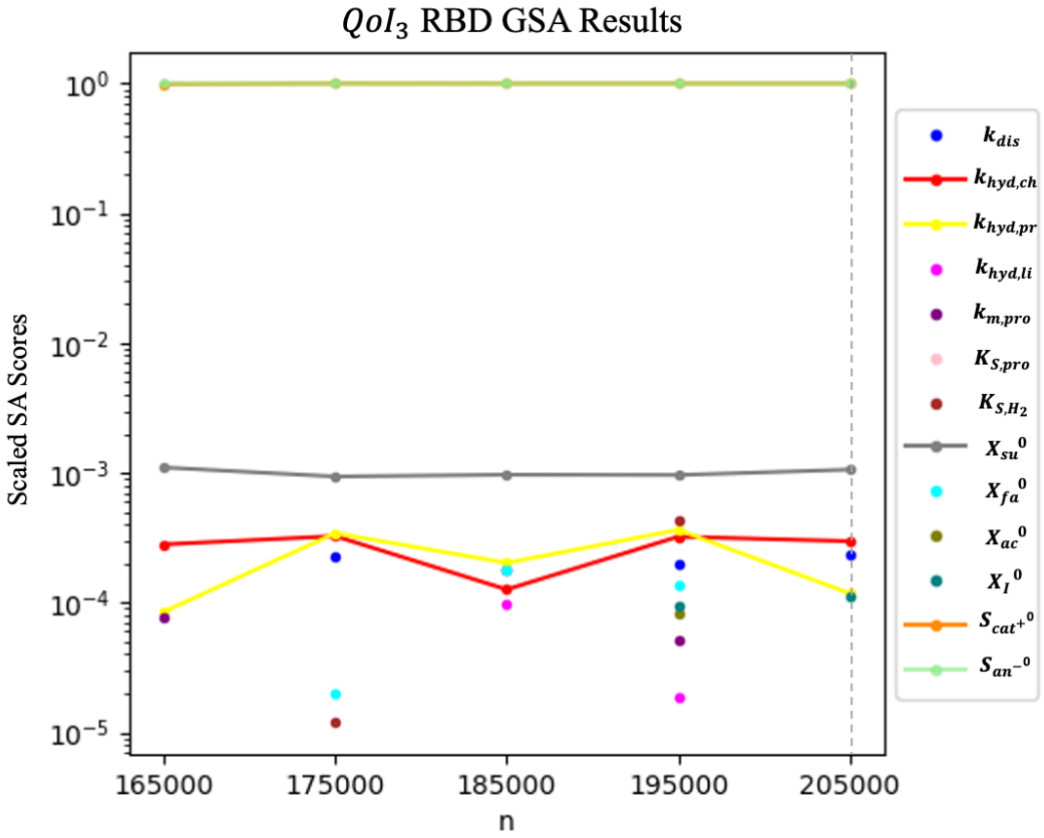


Figure 6-7 The main effects of the 13 more influential θ_{ks} and θ_0 scaled with respect to the largest main effect in five trails with different sample sizes (n) for QoI_3 (y-axis is in logarithmic scale).

To summarise Sobol' and RBD-FAST -GSA results, quite unexpectedly, except for three $\boldsymbol{\theta_{ks}}$: k_{dis}, $k_{hyd,ch}$ and $k_{hyd,pr}$, which are usually recognised as sensitive parameters in ADM1 calibration, four $\boldsymbol{\theta_0}$: $X_{su}{}^0$, $X_{ac}{}^0$, $S_{cat^+}{}^0$ and $S_{an^-}{}^0$ also influence model outputs markedly. In particular, QoI_2 and QoI_3 are predominantly influenced by the values of $S_{cat^+}{}^0$ and $S_{an^-}{}^0$, while these seven factors jointly influenced the value of QoI_1.

None of the k_m or K_S were identified as sensitive inputs. This aligns with the SA result in the study of Pastor-Poquet et al. (2019), in which all the K_S showed negligible influence on the value of an objective function which measured the discrepancy between model simulation and experimental observations. Since k_m and K_S are correlated considering their involvement in the Monod equation (Eq. (2-12), this might has caused that k_m was assessed as not sensitive as well.

Meanwhile, Pastor-Poquet et al. (2019) also found that $X_{ac}{}^{0}$ sensitive. $X_{ac}{}^{0}$ holds significance in preventing VFA accumulation and regulating the pH level within the reactor.

It can be observed that the pre-screening SA result by Morris EE method exhibits discrepancy compared to GSA result. Except for the two most influential uncertain inputs $S_{cat^+}{}^{0}$ and $S_{an^-}{}^{0}$, none of the k_m or K_S parameters screened out by Morris EE were found to be influential by the variance-based GSA methods. And different from the result of Morris EE and also what is commonly accepted in many studies, $k_{hyd,li}$ is not an influential parameter for any of the specific QoI adopted in this study. This can be related to the fact that the content of lipid in the feedstock of this study is extremely low, so the hydrolysis of lipid soes not have a great influence on the whole AD process. These findings underscore the significance of conducting a thorough GSA before model calibration to pinpoint influential uncertain inputs.

Furthermore, this study challenges the common assumption that initial values are inconsequential when employing ADM1. The sensitivity observed in specific $\boldsymbol{\theta_0}$ emphasises the necessity of calibrating the influential $\boldsymbol{\theta_0}$ together with the influential $\boldsymbol{\theta_{ks}}$. Combine the RBD-FAST GSA results for three QoI_s , k_{dis}, $k_{hyd,ch}$, $k_{hyd,pr}$, $X_{su}{}^{0}$, $X_{ac}{}^{0}$, $S_{cat^+}{}^{0}$ and $S_{an^-}{}^{0}$ were calibrated in the following section.

6.2.2 Bayesian Calibration of Influential $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$

6.2.2.1 Prior Distribution

As shown in Table 6-5, the k_{dis} of THP sludge is generally larger than raw sludge, while $k_{hyd,ch}$ and $k_{hyd,pr}$ become smaller. This is a manifestation of the fact that high temperature and pressure result in enhanced cell wall rupture and release of extracellular polymeric substance of sewage sludge (Appels et al., 2008; Gahlot et al., 2022). The disintegration is not rate-limiting step anymore. In fact, when applying ADM1 to THP sludge, Batstone et al. (2008), Donoso-Bravo et al. (2020) and Ortega-Martinez et al. (2016) eliminated disintegration step in the model and considered hydrolysis as the rate-limiting step. Based on the calibrated values of k_{dis} in THP sludge related studies in Table 6-5, the mean (μ)

of the marginal prior distribution of k_{dis} was set at the middle of the reported lower bound (1.5) and upper bound (5.6), which is 3.550. And the standard deviation (σ) was derived by assuming that 86.64% (three σ) data were contained within the lower and upper bound, resulting in a σ value of 1.367. For the setting of prior distributions for $k_{hyd,ch}$ and $k_{hyd,pr}$, as it is believed that the distributions of $k_{hyd,ch}$ and $k_{hyd,pr}$ are right skewed towards smaller values for THP sludge, the Gamma distribution with α = 2 and β = 4 was assigned.

For the prior distributions of X_{su}^{0} and $X_{ac}{}^{0}$, normal distributions were assigned in the same manner as when assigning a normal distribution to k_{dis}. The lower and upper bounds of the variability ranges used in SA were adopted (Table 6-7). Gamma distributions with α = 1.5 and β = 4, which have a smaller mean compared to the prior distributions of $k_{hyd,ch}$ and $k_{hyd,pr}$, were assigned to $S_{cat^+}{}^{0}$ and $S_{an^-}{}^{0}$.

Table 6-9 summarises the marginal prior distributions of k_{dis}, $k_{hyd,ch}$, $k_{hyd,pr}$, X_{su}^{0}, $X_{ac}{}^{0}$, $S_{cat^+}{}^{0}$ and $S_{an^-}{}^{0}$. And Figure 6-8 is a visualisation of these prior distributions.

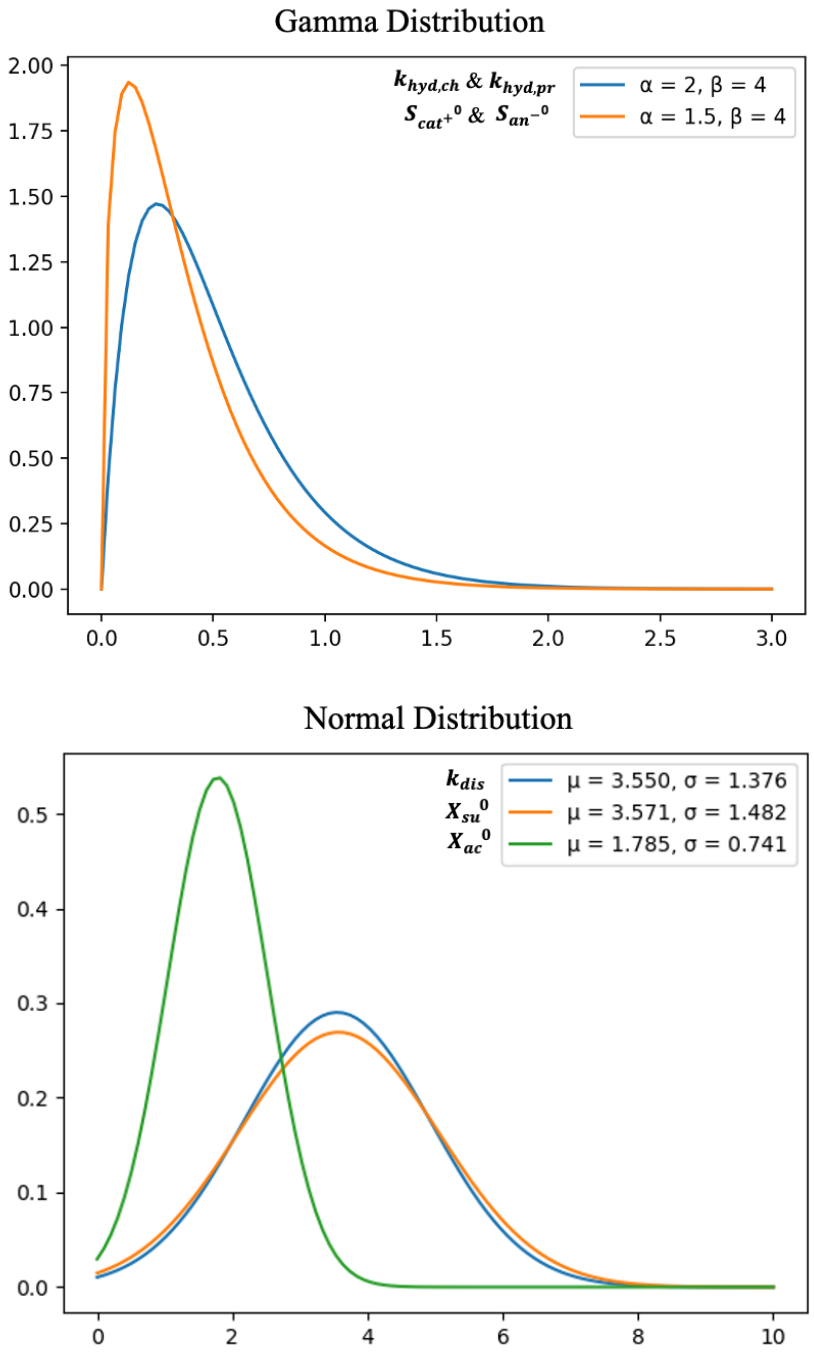


Figure 6-8 Marginal prior probability distributions for k_{dis}, $k_{hyd,ch}$, $k_{hyd,pr}$, $X_{su}{}^0$, $X_{ac}{}^0$, $S_{cat^+}{}^0$ and $S_{an^-}{}^0$.

6.2.2.2 Posterior Distribution

After calibrating the model with 21 experimental observations obtained from an operation period of less than two HRTs, the statistical summary of the marginal posterior distributions for each uncertain inputs, k_{dis}, $k_{hyd,ch}$, $k_{hyd,pr}$, $X_{su}{}^{0}$, $X_{ac}{}^{0}$, $S_{cat^+}{}^{0}$ and $S_{an^-}{}^{0}$, is compiled in Table 6-9. Figure **6-9** is the 'traceplot' generated by 'PyMC3' package derived from two MCMC chains.

The mean of the posterior distribution for $k_{hyd,ch}$ is only 0.024, which is significantly smaller than the values calibrated in other studies about applying ADM1 on THP sludge (Table 6-5). Nevertheless, this value is close to the carbohydrate degradation rate constant reported in the study of Shana et al. (2013), in which a stepwise carbohydrate degradation was contended.

In Table 6-9, comparing the highest density interval of the marginal posterior distribution conveyed by the 'hdi_3%' and 'hdi_97%' with the prior distribution depicted in Figure 6-8, it is noted that the high density area of the marginal posterior distributions of $k_{hyd,ch}$, $k_{hyd,pr}$ and $S_{cat^+}{}^{0}$ exhibit the most significant shifts from their prior distributions. The 'Location Difference' quantifies the distance between the locations of means of the marginal prior and posterior distributions for each uncertain parameter. Among these parameters, the mean location of the marginal posterior distribution for k_{dis}, $k_{hyd,ch}$ and $k_{hyd,pr}$ exhibited the most notable shifts, followed by $X_{ac}{}^{0}$, $S_{cat^+}{}^{0}$ and $S_{an^-}{}^{0}$.

JS divergence offers a comprehensive measurement of the disparity between the marginal prior and posterior distributions, accounting for both the shift in location and the difference in shape. The JS divergences for $k_{hyd,ch}$ and $k_{hyd,pr}$ were the biggest overall. The small JS divergence of parameter k_{dis} should be attributed to the fact that the shape of its posterior distribution did not change much although its mean location shifted significantly. Among the four θ_0, the JS divergence for $S_{an^-}{}^{0}$ was the largest. It is observed that although the high density interval of the marginal posterior of $S_{cat^+}{}^{0}$ moved more from its prior compared to the $S_{an^-}{}^{0}$, the JS divergence for the $S_{an^-}{}^{0}$ was larger. From the 'traceplot' of the posterior distribution (Figure **6-9**), it was evident that although the HDI of $S_{cat^+}{}^{0}$ moved, its

posterior distribution maintained a shape resembling a right-skewed gamma distribution. Conversely, the peak of the posterior distribution of $S_{an^-}{}^0$ shifted to the right. The ${S_{cat^+}}^0$ and ${S_{an^-}}^0$ were provided with the same gamma distributions as prior, after Bayesian calibration, the mean of the posterior distribution of ${S_{cat^+}}^0$ shifted left while $S_{an^-}{}^0$ shifted right, resulting in a difference in the mean locations of 0.301 kg COD/m^3. This conformed with their relationship derived from the charge balance. According to Eq. (6-3, the concentration of $S_{an^-}{}^0$ should be 0.146 kg COD/m^3 higher than ${S_{cat^+}}^0$. Considering the potential errors in the experimental analysis of VFA, TC, TOC, TN, and TAN, the values of ${S_{cat^+}}^0$ and $S_{an^-}{}^0$ calibrated are reasonable in light of chemical principles.

Lastly, from the small 'mcse_mean' and 'mcse_sd' for the marginal posterior distributions of $k_{hyd,ch}$ and $k_{hyd,pr}$, it means that the estimated mean and standard deviation of the posteriors of these two parameters were the most reliable with the data provided for calibration.

In conclusion, the 21 experimental observations used to calibrate ADM1 in this study is the most informative to the values of $k_{hyd,ch}$ and $k_{hyd,pr}$ and then moderately informative to the ${S_{cat^+}}^0$ and $S_{an^-}{}^0$. In other words, these four parameters are the most identifiable under the specific calibration data used. This again validates the result of RBD-FAST GSA that these four parameters are the most sensitive to the values of the QoI_s designated.

Table 6-9 The statistical summary of the marginal prior and posterior distributions of seven the most influential uncertain inputs of ADM1.

	Prior Distribution		Posterior Distribution						Location Difference	JS Divergence
	Distribution	Parameters	Mean	SD	hdi_3%	hdi_97%	mcse_mean	mcse_sd		
k_{dis}	Normal	μ = 3.550 σ = 1.367	3.235	1.493	0.034	5.557	0.064	0.045	0.315	0.011
$k_{hyd,ch}$	Gamma	α = 2 β = 4	0.022	0.009	0.006	0.038	0.000	0.000	0.478	**0.308**
$k_{hyd,pr}$	Gamma	α = 2 β = 4	0.240	0.061	0.150	0.355	0.003	0.002	0.260	**0.124**
X_{su}^{0}	Normal	μ = 3.571 σ = 1.482	3.562	1.535	0.666	6.432	0.088	0.063	0.009	0.003
X_{ac}^{0}	Normal	μ = 1.78 σ = 0.741	1.977	0.580	0.950	3.000	0.018	0.013	0.197	0.043
$S_{cat^+}^{0}$	Gamma	α = 1.5 β = 4	0.198	0.147	0.003	0.488	0.013	0.010	0.177	**0.067**
$S_{an^-}^{0}$	Gamma	α = 1.5 β = 4	0.499	0.150	0.268	0.805	0.012	0.009	0.124	**0.135**

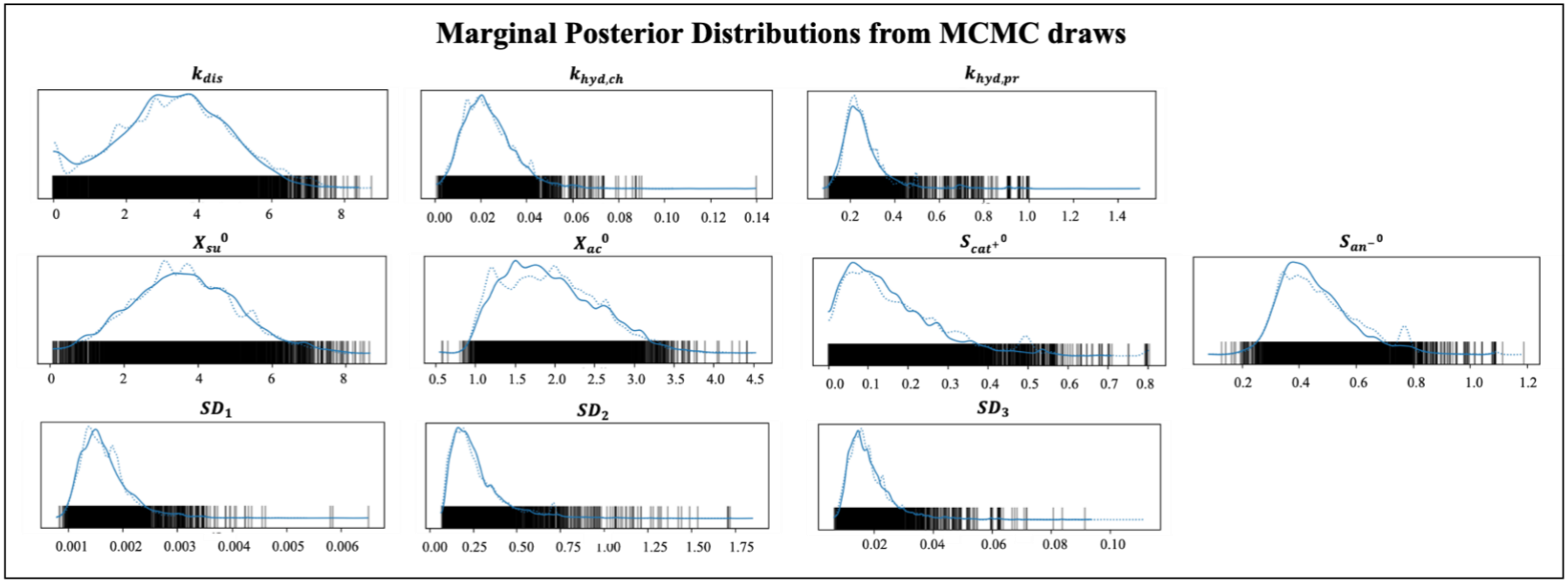


Figure 6-9 The smoothed marginal posterior distributions for k_{dis}, $k_{hyd,ch}$, $k_{hyd,pr}$, X_{su}^0, X_{ac}^0, S_{cat}^0 and S_{an}^0 from two MCMC chains. SD_1, SD_2 and SD_3 is the marginal distribution of the estimated measurement errors for daily biogas flow (m^3/d), pH and TAN (kmol N/m^3).

6.2.2.3 Posterior Predictive Check

The daily biogas production flow (m^3/d) is shown in Figure **6-10**, with blue dots representing data used for the calibration. The grey dotted line is the 10-day moving average trendline of the daily biogas flow. The orange line and orange shaded area are the mean and the 95% CI of the model outputs with 3000 sets of uncertain inputs drawn from the posterior distribution. With only 21 data in total from less than two HRTs period for model calibration, the model with seven calibrated $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ simulated the initial surge of the biogas production observed. For the daily biogas flow data left for model validation, 66.047% of the 10-day moving average trendline lies within the 95% CI of the model outputs. From the simulation of the calibrated model, biogas production should have arrived at its steady state at around day 70.

The residuals between the predicted daily biogas flow and the experimental observations for model validation and the distribution of the residuals is shown in Figure 6-11. The distribution is slightly-right skewed compared to the standard normal distribution, but still the majority of the residuals centre around zero.

Figure **6-12**, Figure **6-13** and Figure **6-14** show the TAN of the digestate sampled at irregular intervals, the pH of the digestate analysed at the beginning and the end of the study period and the tCOD of the digestate at the end of the study respectively. Similarly, the blue dots are data used for calibration. The mean of the model output with $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ sampled from the posterior distribution depicted the increasing trend of the pH at the beginning state. Also, the pH observations for validation fall within the 95% CI of the model outputs (Figure **6-13**). tCOD data were not used for model calibration but only for validation, and the experimental observations conformed with the model simulation (Figure **6-14**). However, the mean of the steady state TAN outputs of the calibrated model is 12.609% lower than the mean of the experimental observation, which is 0.184 kmole N/m^3. A systematic error in the measurement of the TAN concentration is suspected.

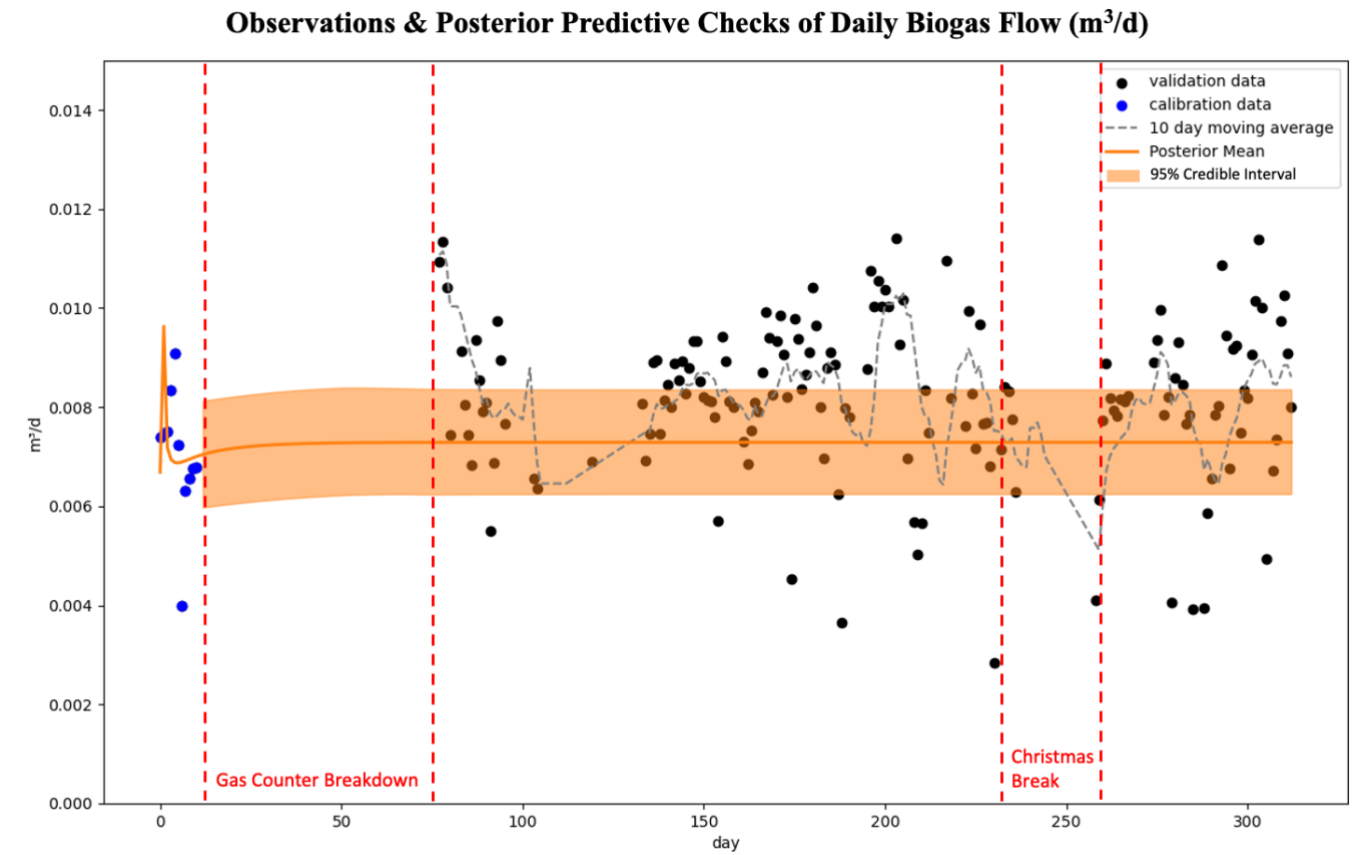


Figure 6-10 Experimental observations and simulated results of calibrated model of daily biogas flow. Blue dots: data used for Bayesian calibration. Black dots: data for validation. Gray dotted line: 10-day moving age trendline of the daily biogas flow. Orange line and shaded area: the mean and 95% credible interval of 3000 model outputs for daily biogas flow with θ_{ks} and θ_0 sampled from trace of MCMC chains after burn-in samples.

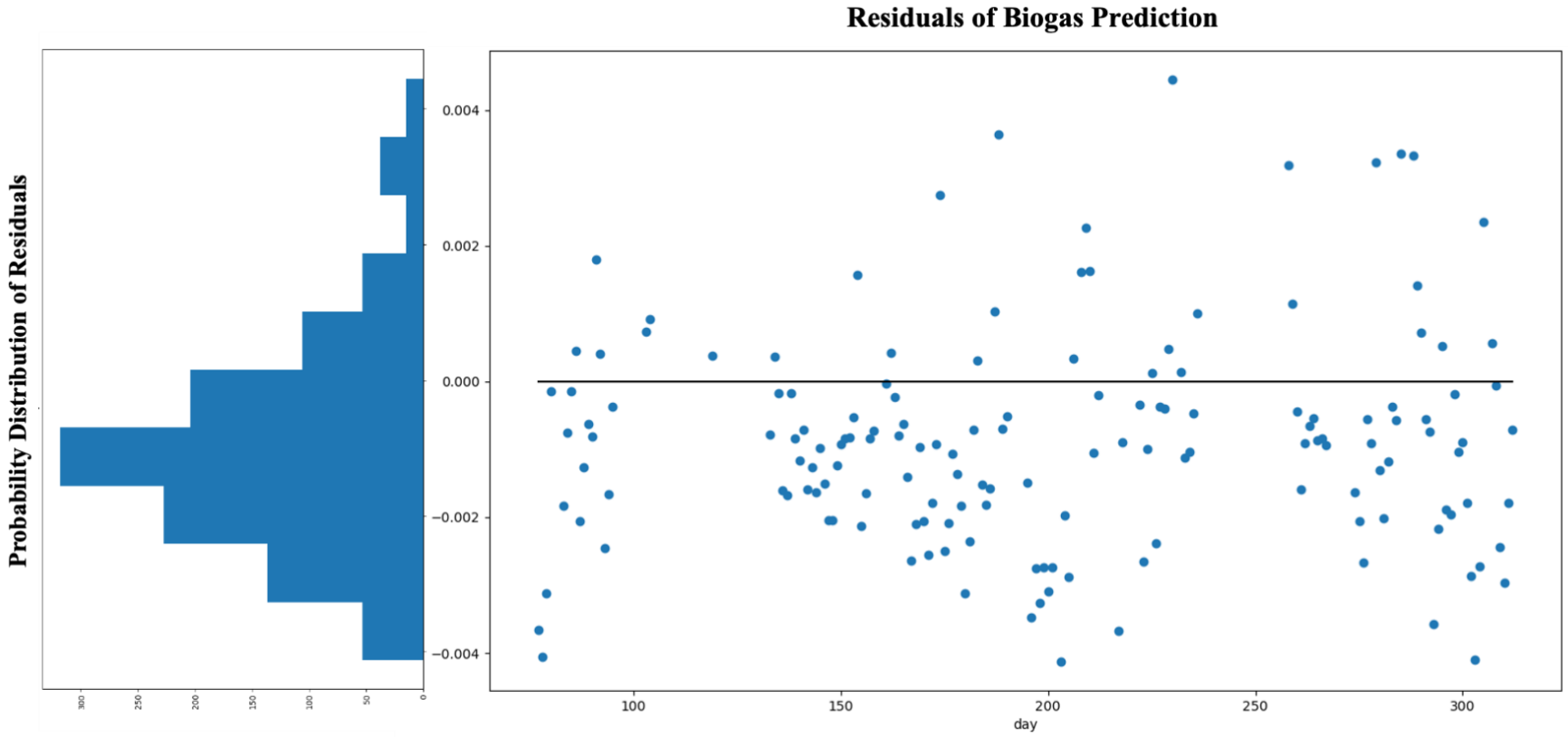


Figure 6-11 Distribution of the residuals of the mean prediction for daily biogas flow on validation data (from day 76).

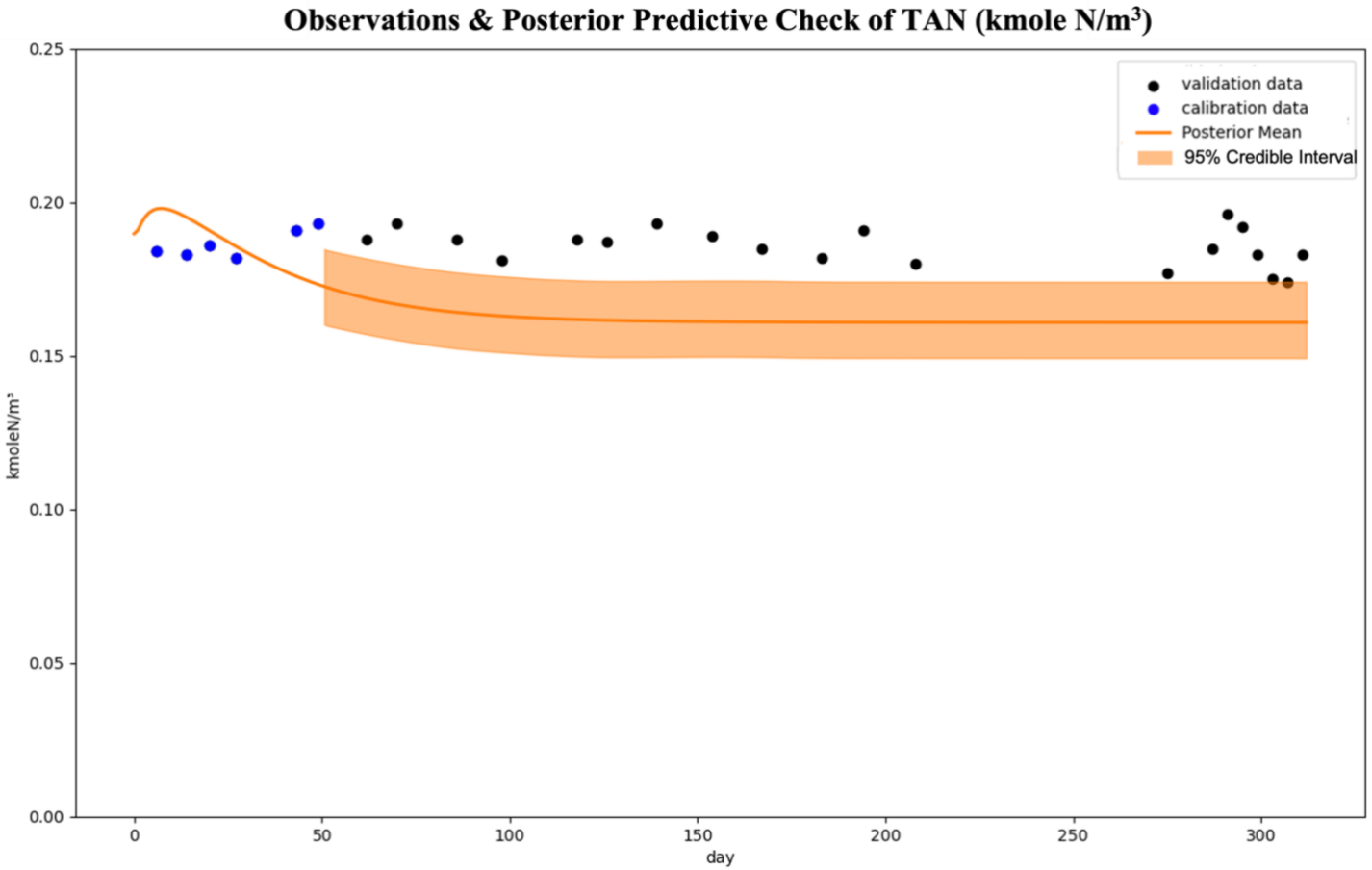


Figure 6-12 Experimental observations and simulated results of calibrated model of digestate TAN. Blue dots: data used for Bayesian calibration. Black dots: data for validation. Orange line and shaded area: the mean and 95% credible interval of 3000 model outputs for TAN with θ_{ks} and θ_0 sampled from trace of MCMC chains after burn-in samples.

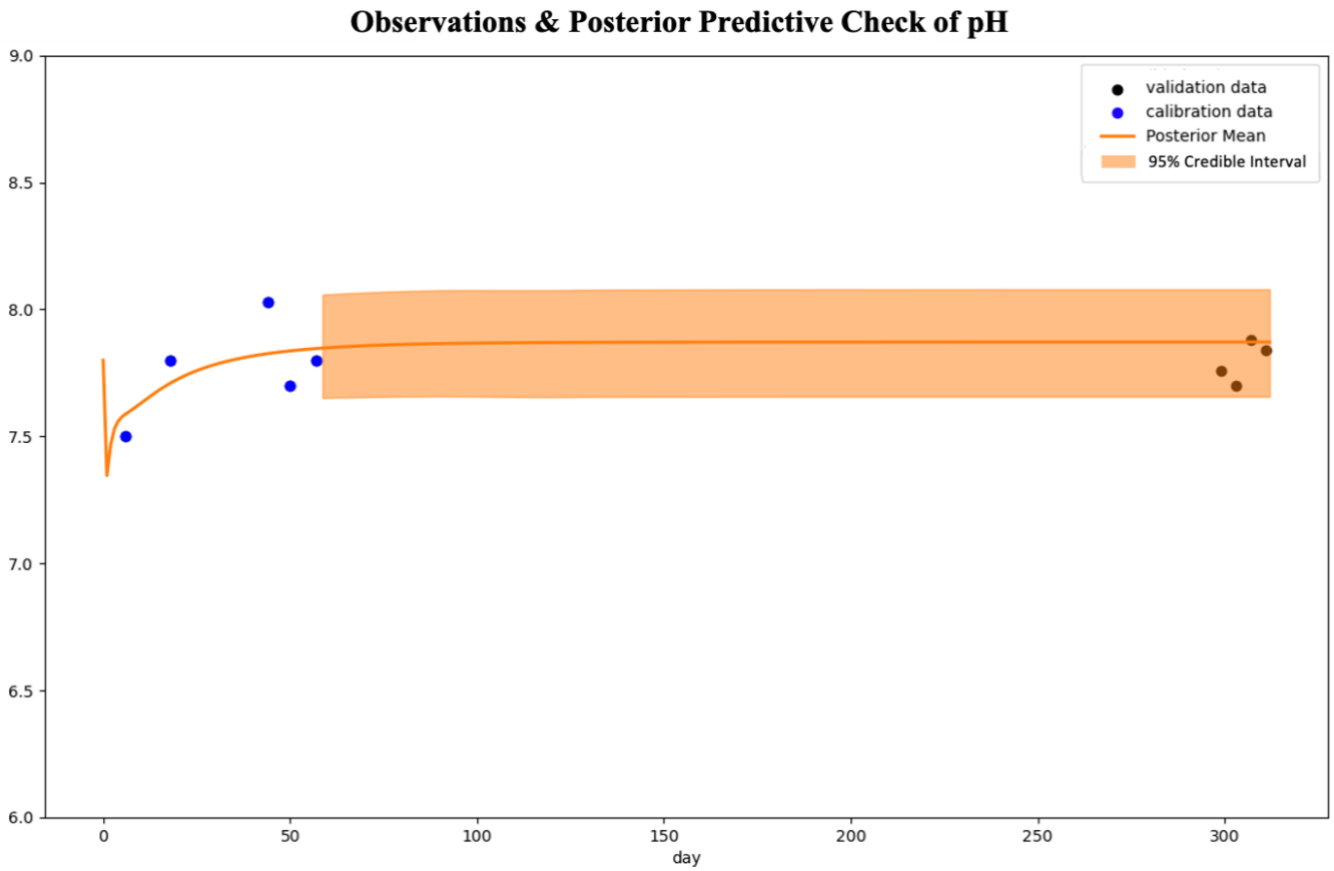


Figure 6-13 Experimental observations and simulated results of calibrated model of digestate pH value. Blue dots: data used for Bayesian calibration. Black dots: data for validation. Orange line and shaded area: the mean and 95% credible interval of 3000 model outputs for pH with θ_{ks} and θ_0 sampled from trace of MCMC chains after burn-in samples.

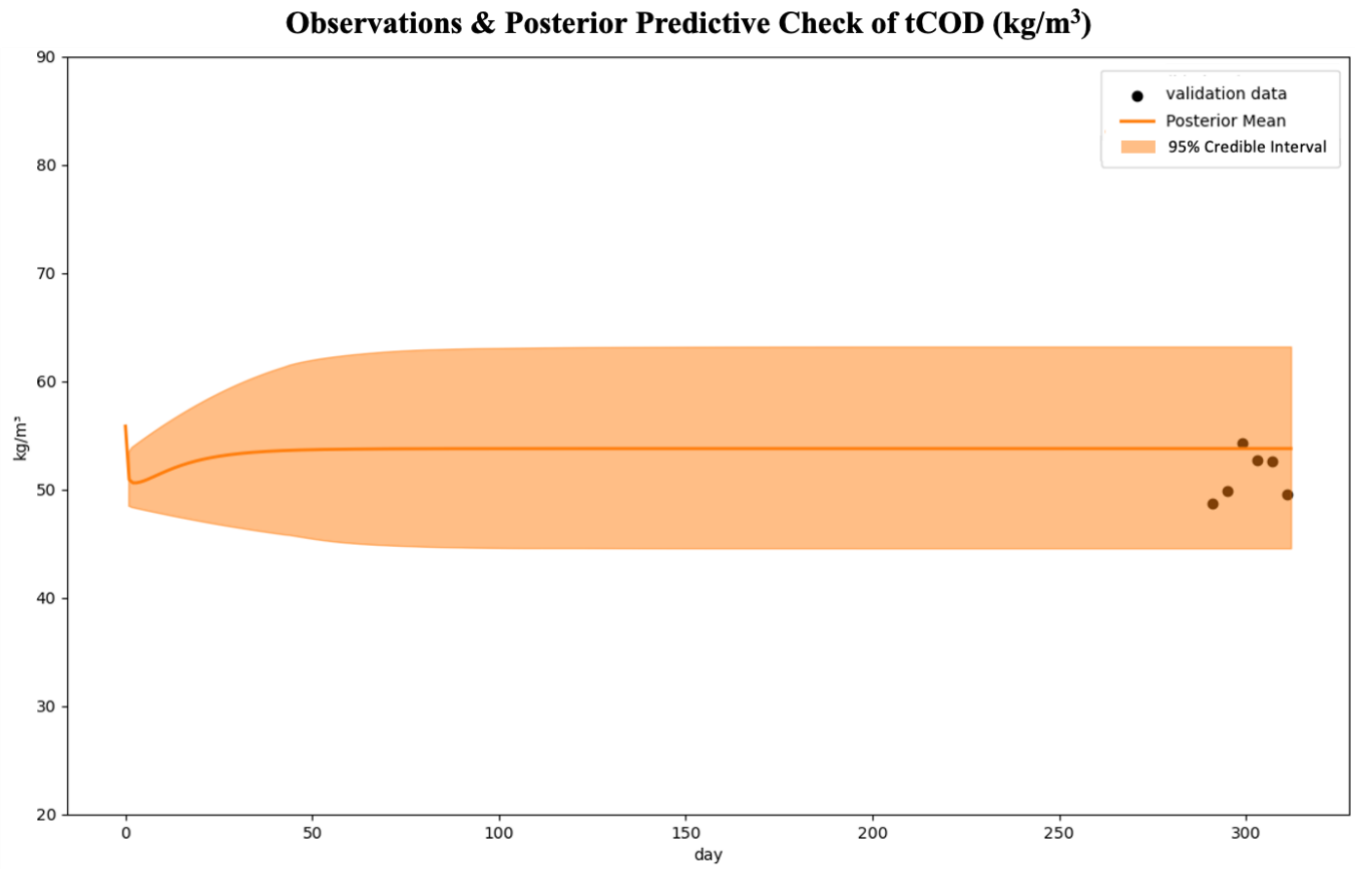


Figure 6-14 Experimental observations and simulated results of calibrated model of digestate tCOD value. Blue dots: data used for Bayesian calibration. Black dots: data for validation. Orange line and shaded area: the mean and 95% credible interval of 3000 model outputs for tCOD with θ_{ks} and θ_0 sampled from trace of MCMC chains after burn-in samples.

Similarly, the simulated results of the TAN, pH value and tCOD of the digestate also arrive their steady states in 50 to 100 days. Considering the average HRT 25.243 d of the anaerobic digester under study, its performance is in line with the performance of other continuous anaerobic digesters where a steady-state operation is usually achieved after 3-4 HRTs (Peces et al., 2021; Razaviarani & Buchanan, 2015; Song et al., 2016).

The mean of the steady state simulation results for the daily biogas flow, and TAN, pH value and tCOD of the digestate are respectively 0.007 m^3/day, 0.161 kmole N/m^3, 7.891, and 53.791 kg/m^3.

6.2.2.4 The Influence of Prior Distribution

Bayesian calibration with the same 21 experimental observations but with uniform prior distributions were implemented. The statistical summary of the marginal posterior distributions of seven uncertain $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ and the mean of the steady state simulation results for daily biogas flow, TAN, pH and tCOD with parameters drawn from the posterior distribution are presented in Table 6-10.

The simulation of the daily biogas flow is shown in Figure 6-15 specifically. Compared with the simulation of the model calibrated with informative prior provided, only 41.860% of the 10-day moving average trendline lies within the 95% CI of the simulation of the model calibrated with uniform prior.

The mean of the prediction for the steady state pH is above the mean of the experimental observations, which is 7.795. While for the mean of the prediction for the steady state tCOD, it is much smaller than the mean of the experimental observations, which is around 51.322 kg/m^3. All in all, the model with seven $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ calibrated with the same number of calibration data but with uniform prior distribution assigned cannot predict the validation data as well as the model calibrated with a more informative prior assigned.

This result underscores the significance of incorporating in-depth domain knowledge when assigning prior distributions in Bayesian calibration, a consideration that becomes even more pivotal when the experimental observations available for calibration are limited.

Table 6-10 Statistical summary of marginal posterior distributions for uncertain inputs when prior distributions are uniform distributions and the mean of the steady state simulation results for daily biogas flow, TAN, pH and tCOD with 3000 set of parameters drawn from the posterior distribution.

	Posterior Distributions				Mean of Model Output			
Params	Mean	SD	hdi_ 3%	hdi_ 97%	Daily Biogas Flow (m^3/d)	TAN (kmole N/m^3)	pH	tCOD (kg/m^3)
k_{dis}	1.227	1.564	0.101	4.646	0.010	0.164	8.019	33.061
$k_{hyd,ch}$	0.542	1.367	0.092	3.938				
$k_{hyd,pr}$	3.097	2.896	0.117	8.750				
X_{su}^0	3.437	1.284	1.348	5.497				
X_{ac}^0	2.077	0.558	1.094	2.896				
$S_{cat^+}^0$	0.084	0.045	0.010	0.155				
$S_{an^-}^0$	0.440	0.080	0.320	0.600				

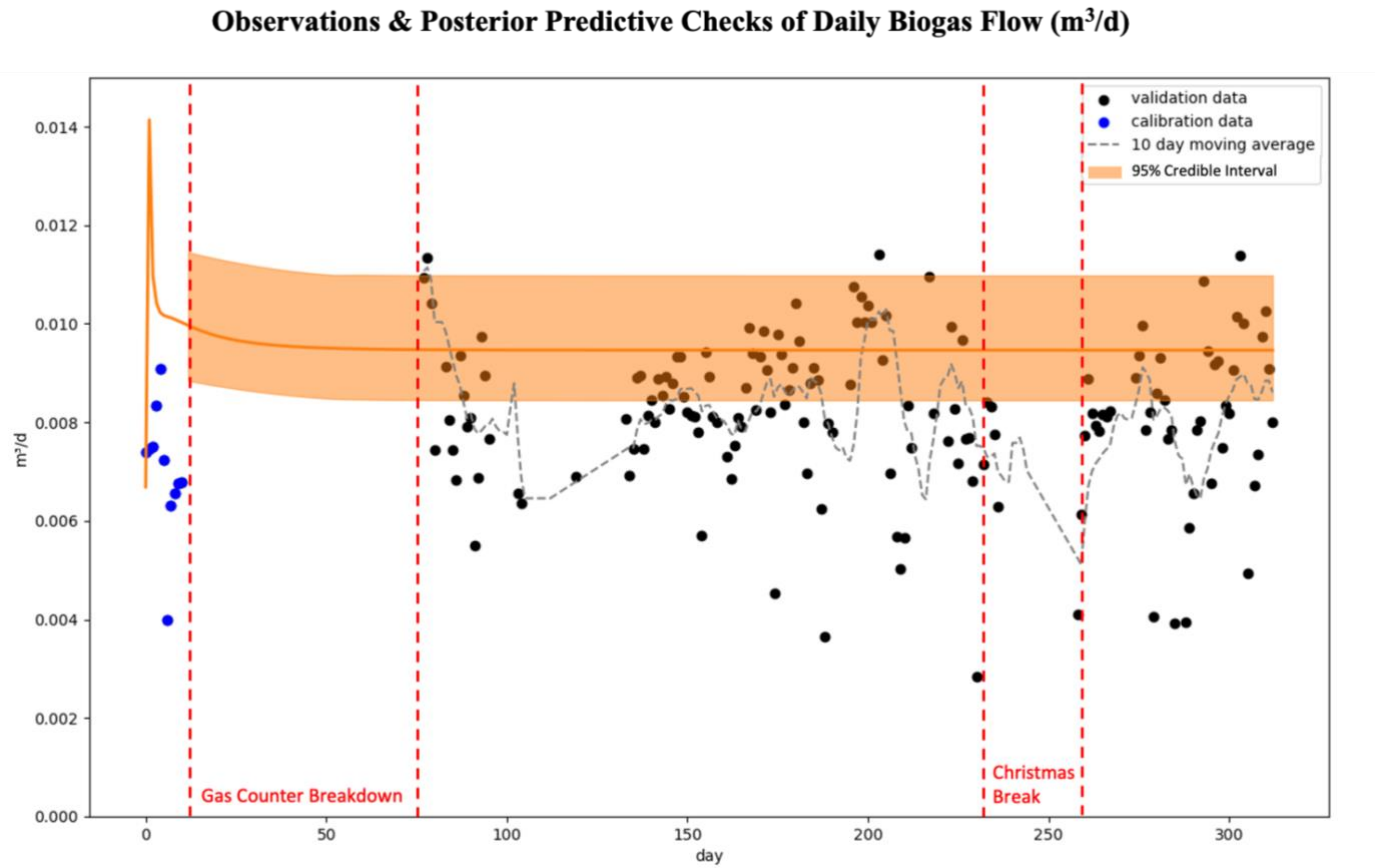


Figure 6-15 Experimental observations and simulated results of calibrated model of daily biogas flow. Blue dots: data used for Bayesian calibration. Black dots: data for validation. Gray dotted line: 10-day moving age trendline of the daily biogas flow. Orange line and shaded area: the mean and 95% credible interval of 3000 model outputs for daily biogas flow with θ_{ks} and θ_0 sampled from posterior distribution. Uniform distribution was assigned as prior distribution.

6.3 Summary and Discussion

This chapter presented the framework for Bayesian calibration of the highly parameterised and non-linear ODE system model – the ADM1 – for continuous AD process prediction. The feasibility of obtaining reliable posterior distributions of the influential $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ parameters of ADM1 using data from less than two HRTs (21 experimental observations in total) was demonstrated. This success relies on two factors: screening the most influential $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ for calibration via SA and assigning informative prior distributions through extensive domain knowledge.

The selection of the most influential $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ was achieved through a two-step process: a pre-screening by Morris method followed by a further screening using variance-based GSA methods. Based on their influence on the values of three QoI_{s} chosen according to the experimental observations available for model calibration, the Morris EE method preliminarily screened out 13 influential $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ out of 24. Then, combining the variance-based GSA results from Sobol' and RBD-FAST methods, it was revealed that ${S_{cat^+}}^0$ and ${S_{an^-}}^0$ exhibited strong interactions and had a dominant influence on the model outputs of steady-state TAN concentration and pH value of the digestate. Besides ${S_{cat^+}}^0$ and ${S_{an^-}}^0$, $k_{hyd,pr}$ also strongly influenced TAN concentration. By contrast, the steady-state daily biogas flow was influenced collectively by more $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$, including $k_{hyd,ch}$, ${X_{\mathrm{ac}}}^0$ and k_{dis} besides ${S_{cat^+}}^0$ and ${S_{an^-}}^0$. In summary, the seven most influential $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ selected for calibration were k_{dis}, $k_{hyd,ch}$, $k_{hyd,pr}$, ${X_{su}}^0$, ${X_{ac}}^0$, ${S_{cat^+}}^0$ and ${S_{an^-}}^0$.

The SA results challenged the common practice of calibrating only $\boldsymbol{\theta_{ks}}$ while treating $\boldsymbol{\theta_0}$ as trivial in many studies of ADM1 calibration. Certain $\boldsymbol{\theta_0}$, especially ${S_{cat^+}}^0$ and ${S_{an^-}}^0$, significantly impacted the model output. Additionally, the SA results by Morris EE method were not reliable enough; parameters like $k_{hyd,li}$, k_m and K_S, screened out by Morris EE method, were found to be unimportant in further variance-based GSA.

For Bayesian calibration, the data used for assigning prior distributions were restricted to studies on THP sludge rather than studies on the broader 'sludge' that were referred to when defining the varying range of $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ in SA. Also, more informative Normal and Gamma distributions were assigned based on domain expertise. The posterior predictive check of model outputs with $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ sampled from the posterior distribution confirmed that the model calibrated with informative priors provided fit the calibration data and predicted the validation data better than the model calibrated with uniform priors.

Furthermore, the JS divergence between the marginal prior and posterior distributions, in case of the calibration with informative priors provided, and the Monte Carlo standard error of the posterior estimation revealed that the data available for model calibration were the most informative to the estimations for $k_{hyd,ch}$, $k_{hyd,pr}$, ${S_{cat^+}}^0$ and ${S_{an^-}}^0$. In other words, they were the most identifiable with the data available for calibration. This finding was in line the GSA results.

In summary, by integrating comprehensive domain knowledge into the determination of prior distributions and identifying the most influential $\boldsymbol{\theta_{ks}}$ and $\boldsymbol{\theta_0}$ for calibration through GSA, Bayesian calibration can achieve higher reliability in parameter calibration of ADM1 with less requirement on the amount data available for calibration, compared to manual heuristic search methods or algorithms for searching minimum objective function loss. The ability to reliably calibrate ADM1 without extended experimental operation will promote the application of ADM1 for industry-scale continuous AD performance prediction.

The limitation of this chapter is that a further study of Bayesian robustness could be carried out to examine the robustness and reliability of Bayesian inference. This is to investigate the influences of the disturbance to the prior distribution and the choice of likelihood function (comparing to the standard normal likelihood function used in this chapter assuming Gaussian errors in the observations) on the derived posterior distribution (Kaollioinen et al., 2021; Masoero et al., 2018).

7 Conclusions and Future Work

7.1 Conclusions

PART I of this study focused on predicting the batch AD test result using the physiochemical characteristics of the feedstock. The proposed approach employs ML regression models to infer the parameters of a low-dimensional mathematical biogas production model from the high-dimensional feedstock characteristics space. This methodology not only enables the prediction of a feedstock's ultimate biomethane yield but also provides insights into its degradation kinetics.

To verify the validity of this prediction mechanism, a dataset comprising the BMP test results for 25 digestate samples, along with comprehensive analyses of their physiochemical characteristics, was used. Although it was demonstrated that the prediction accuracy of the regression model (DT specifically) would improve with more training data, the limited dataset hindered the training of a DT with satisfactory prediction accuracy. Additionally, the feature importance results returned from the trained DT model were unreliable.

To address the challenge of data limitation when training ML models for BMP test result prediction, a method was proposed to generate additional datasets via response surface design. Leveraging the augmented data, XGBoost models were trained to predict the parameters of the first-order kinetic model with nine features, and three crucial features were identified. Remarkably, even with just these three important features, the XGBoost models still achieved high predictive accuracy on test set.

In pursuit of a more readily accessible model for batch AD test result prediction to domain experts and offering a mathematical interpretation of the black-box ML model's predictions, metamodeling of the trained XGBoost models with Meijei G-functions was demonstrated. Furthermore, to evaluate the general predictability of the XGBoost models trained with the augmented dataset and its metamodels,

the metamodel for the ultimate specific methane yield prediction was tested on 13 BMP test data sourced from the literature.

PART II of this study illustrated a systematic framework for calibrating the ADM1 for continuous AD process prediction using Bayesian inference. A small number of experimental observations from less than two HRTs were used for model calibration, and the calibrated model accurately predicted the steady-state performance of the digester. The successful calibration of ADM1 with limited experimental observations is attributed to the identification of influential uncertain inputs for calibration via variance-based GSA and the application of informative prior distributions determined from information gathered in the literature..

7.1.1 Batch AD Test Result Prediction from Feedstock Physiochemical Characteristics

- With 22 characteristics (including elemental compositions, TS, VS, tCOD, sCOD and various possible inhibitors) of the digestate sample as predictors, the DT trained and tested via LOO cross validation using 25 BMP test data was able to predict the BMP test result (biogas production at day 28) with an average APE 40.003%. Among 25 LOO cross validation scenarios, the APE of the prediction for the sample left out of five scenarios was smaller than 5%.
- The average APEs of the inferred coefficient and constant of the biogas production model and the calculated BMP test result when training a DT with 24 data were 4.187%, 48.443% and 40.003% respectively. This improved from the APEs of 4.760%, 72.041% and 52.134% when training with 20 data, which proved that the prediction accuracy of the regression model would improve with more training data available.
- The feature importance result returned from the trained DT shown that the characteristics indicating the substrate load (VS, tCOD and sCOD), TAN or TKN, TA and metal elements Ni and Co were important predictors. This result was not reliable due to the dubious convergence of the DT trained with such small sample size. Despite this, it provided a preliminary insight

into the important characteristics affecting the biogas potential of a feedstock.

- A GP guided by the inferred biogas production model was able to provide uncertainty information to the biogas production prediction.

7.1.2 Augmented Dataset for ML Regression Model Training

- 10 BMP tests of mixtures among three food samples with distinct chemical composition characteristics were carried out following mixture design and the biogas production data were fitted with first-order kinetic model. The special cubic response surface equations fitted the 10 k and Y_m parameters with R^2 scores of 0.975 and 0.980 respectively.
- Additional data were generated from the response surface equations for XGBoost training and testing. The predictors of the XGBoost model were reduced from nine to three according to feature importance analysis results. The XGBoost model with features, 'Lipid', 'Lignin' and 'Protein', predicted the parameter k the most accurate (R^2 = 0.995). The XGBoost model with three features, 'Lipid', 'Protein' and 'Non-fibre carbohydrate', predicted the parameter Y_m the most accurate (R^2 = 0.993).

7.1.3 Mathematical Interpretation of the XGBoost Model

- The fifth-order Taylor approximated Meijei G-function metamodels were able to recover the predictions of XGBoost models for k and Y_m prediction with R^2 score of 0.907 and 0.950 respectively.
- The MAE of the fifth-order Taylor approximated metamodel for the prediction of the BMP test results (ultimate specific methane yield) of 13 external samples was 38.074. This has demonstrated the general applicability and performance of the ML model trained with augmented data and its metamodel.

7.1.4 Bayesian Calibration of ADM1

- Contrary to the common practice in ADM1 calibration which treated the initial condition of state variables as trivial, GSA result revealed the influence of certain initial conditions, specifically the concentrations of cation and anion in the digestate, on the model outputs.
- The ADM1 calibrated with only 21 experimental observations from less than two HRTs can predict the steady state performance of the digester with satisfactory accuracy. 66.047% of the 10-day moving average trendline of the daily biogas flow data fell within the 95% CI of the model outputs. All of the steady state digestate pH and tCOD data for validation fell within 95% CI. The steady state digestate TAN observations are slightly above model's prediction, which might be due to the error in TAN concentration measurement.
- The importance of an informative prior determined by extensive incorporation of domain knowledge was proved by comparing the posterior predictive check results with model calibrated with a uniform prior.
- In terms of the identifiability of the uncertain inputs, the GSA results conformed with the information derived from the JS divergences between the marginal prior and posterior distributions and the confidence in the estimates of mean and standard deviation of the marginal posterior. Hydrolytic kinetics of the carbohydrate and protein and the initial values of the cation and anion concentrations were the most identifiable with the calibration data available.

7.2 Future Work

- To quantify the accuracy of the additional data generated for training ML models for BMP test result prediction via response surface design, more BMP tests should be conducted with separate samples reserved to test: (1) the validity of the assumption that the characteristics of the mixture sample can be calculated as a weighted sum of the individual three samples without any reactions occurring; and (2) the accuracy of the

biogas production model parameters inferred from the fitted response surfaces, compared to the parameters derived from biogas production data fitting.

- By selecting three substrates with distinct characteristics that indicate possible inhibitions in the AD process, such as high TAN concentration, high C/N ratio, and low alkalinity concentration, and applying the same methodology for data augmentation, it is possible to develop a model for BMP result prediction that accounts in these inhibitory factors.

References

Adeleke, A. A., Okolie, J. A., Ogbaga, C. C., Ikubanni, P. P., Okoye, P. U., & Akande, O. (2023). Machine Learning Model for the Evaluation of Biomethane Potential Based on the Biochemical Composition of Biomass. *BioEnergy Research*, *17*(1), 731–743. https://doi.org/10.1007/s12155-023-10681-9

Aguilar, A., Casas, C., & Lema, J. M. (1995). Degradation of volatile fatty acids by differently enriched methanogenic cultures: Kinetics and inhibition. *Water Research*, *29*(2), 505–509. https://doi.org/10.1016/0043-1354(94)00179-B

Alaa, A. M., & van der Schaar, M. (2019a). Attentive State-Space Modeling of Disease Progression. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*.

Alaa, A. M., & van der Schaar, M. (2019b). Demystifying Black-box Models with Symbolic Metamodels. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*.

Aldin, S., Tu, F., Nakhla, G., & Ray, M. B. (2011). Simulating the Degradation of Odor Precursors in Primary and Waste-Activated Sludge During Anaerobic Digestion. *Applied Biochemistry and Biotechnology*, *164*(8), 1292–1304. https://doi.org/10.1007/s12010-011-9213-3

Almomani, F. (2020a). Prediction of biogas production from chemically treated co-digested agricultural waste using artificial neural network. *Fuel*, *280*. https://doi.org/10.1016/j.fuel.2020.118573

Almomani, F. (2020b). Prediction of biogas production from chemically treated co-digested agricultural waste using artificial neural network. *Fuel*, *280*, 118573. https://doi.org/10.1016/j.fuel.2020.118573

Alrawashdeh, K. A. B., Al-Samrraie, L. A., Al Issa, H. A., Qasem, I., Hussien, A. A., Al-Zboon, K. K., Damseh, R. A., & Gul, E. (2022). Prediction and Optimization of Biogas Production from OMW Digestion Using Fenton Pre-Treatment Process with Particle Swarm Optimization. *International Journal*

of Design & Nature and Ecodynamics, *17*(2), 157–168. https://doi.org/10.18280/ijdne.170201

Amon, T., Amon, B., Kryvoruchko, V., Zollitsch, W., Mayer, K., & Gruber, L. (2007). Biogas production from maize and dairy cattle manure—Influence of biomass composition on the methane yield. *Agriculture, Ecosystems & Environment*, *118*(1–4), 173–182. https://doi.org/10.1016/j.agee.2006.05.007

Angelidaki, I., Alves, M., Bolzonella, D., Borzacconi, L., Campos, J. L., Guwy, A. J., Kalyuzhnyi, S., Jenicek, P., & van Lier, J. B. (2009). Defining the biomethane potential (BMP) of solid organic wastes and energy crops: a proposed protocol for batch assays. *Water Science and Technology*, *59*(5), 927–934. https://doi.org/10.2166/wst.2009.040

Angelidaki, I., & Sanders, W. (2004). Assessment of the anaerobic biodegradability of macropollutants. *Reviews in Environmental Science and Bio/Technology*, *3*(2), 117–129. https://doi.org/10.1007/s11157-004-2502-3

Angelidaki, I., Treu, L., Tsapekos, P., Luo, G., Campanaro, S., Wenzel, H., & Kougias, P. G. (2018). Biogas upgrading and utilization: Current status and perspectives. *Biotechnology Advances*, *36*(2), 452–466. https://doi.org/10.1016/j.biotechadv.2018.01.011

AOAC. The Association of Official Analytical Chemists. (1990). Fiber (Acid Detergent) and Lignin in Animal Feed. (973.18). In *Official Methods of Analysis* (15th Edition).

APHP. (1998). *Standard Methods for the Examination of Water and Wastewater.* (20th Edition). American Public Health Association. American Water Works Association and Water Environmental Federation.

Appels, L., Assche, A. Van, Willems, K., Degrève, J., Impe, J. Van, & Dewil, R. (2011). Peracetic acid oxidation as an alternative pre-treatment for the anaerobic digestion of waste activated sludge. *Bioresource Technology*, *102*(5), 4124–4130. https://doi.org/10.1016/j.biortech.2010.12.070

Appels, L., Baeyens, J., Degrève, J., & Dewil, R. (2008). Principles and potential of the anaerobic digestion of waste-activated sludge. *Progress in Energy and Combustion Science*, *34*(6), 755–781. https://doi.org/10.1016/j.pecs.2008.06.002

Appels, L., Lauwers, J., Degrève, J., Helsen, L., Lievens, B., Willems, K., Van Impe, J., & Dewil, R. (2011). Anaerobic digestion in global bio-energy production: Potential and research challenges. *Renewable and Sustainable Energy Reviews*, *15*(9), 4295–4301. https://doi.org/10.1016/j.rser.2011.07.121

Azhar, N. G., & Stuckey, D. C. (1994). The influence of chemical structure on the anaerobic catabolism of refractory compounds: a case study of instant coffee wastes. *Water Science and Technology*, *30*(12), 223–232. https://doi.org/10.2166/wst.1994.0615

Barnes, E. W. (1908). A new development of the theory of the hypergeometric function. *Proc. London Math. Soc.*, *2*(6), 141–177.

Batstone, D. J., Angelidaki, I., & Vavilin, V. (2002). *Anaerobic digestion model No 1 (ADM1) Control of anaerobic digestion process View project Tailored polymers for resource recovery from domestic wastewater View project*. https://www.researchgate.net/publication/11198259

Batstone, D. J., Tait, S., & Starrenburg, D. (2009). Estimation of hydrolysis parameters in full-scale anerobic digesters. *Biotechnology and Bioengineering*, *102*(5), 1513–1520. https://doi.org/10.1002/bit.22163

Bayard, R., Liu, X., Benbelkacem, H., Buffiere, P., & Gourdon, R. (2016). Can Biomethane Potential (BMP) Be Predicted from Other Variables Such As Biochemical Composition in Lignocellulosic Biomass and Related Organic Residues? *BioEnergy Research*, *9*(2), 610–623. https://doi.org/10.1007/s12155-015-9701-3

Bayes, T., & Price, R. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by

Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418. https://doi.org/10.1098/rstl.1763.0053

Beals, R., & Szmigielski, J. (2013). Meijer G–Functions: A Gentle Introduction. *Notices of the American Mathematical Society*, *60*(06).

Bekiaris, G., Bruun, S., Peltre, C., Houot, S., & Jensen, L. S. (2015). FTIR–PAS: A powerful tool for characterising the chemical composition and predicting the labile C fraction of various organic waste products. *Waste Management*, *39*, 45–56. https://doi.org/10.1016/j.wasman.2015.02.029

Bishop, C. M. (1985). Pre-processing and Feature Extraction. In *Neural Networks for Pattern Recognition* (1st ed., pp. 295–329). Clarendon Press.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. In *Pattern Recognition and Machine Learning* (Ed. 1).

Blanchard, O. J. (1987). Comment. *Journal of Business & Economic Statistics*, *5*(4), 449–451. https://doi.org/10.1080/07350015.1987.10509611

Blumensaat, F., & Keller, J. (2005). Modelling of two-stage anaerobic digestion using the IWA Anaerobic Digestion Model No. 1 (ADM1). *Water Research*, *39*(1), 171–183. https://doi.org/10.1016/j.watres.2004.07.024

Bohutskyi, P., Phan, D., Kopachevsky, A. M., Chow, S., Bouwer, E. J., & Betenbaugh, M. J. (2018). Synergistic co-digestion of wastewater grown algae-bacteria polyculture biomass and cellulose to optimize carbon-to-nitrogen ratio and application of kinetic models to predict anaerobic digestion energy balance. *Bioresource Technology*, *269*, 210–220. https://doi.org/10.1016/j.biortech.2018.08.085

Bornhöft, A., Hanke-Rauschenbach, R., & Sundmacher, K. (2013). Steady-state analysis of the Anaerobic Digestion Model No. 1 (ADM1). *Nonlinear Dynamics*, *73*(1–2), 535–549. https://doi.org/10.1007/s11071-013-0807-x

Boubaker, F., & Ridha, B. (2008). Modelling of the mesophilic anaerobic co-digestion of olive mill wastewater with olive mill solid waste using anaerobic digestion model No. 1 (ADM1). *Bioresource Technology*, *99*(14), 6565–6577. https://doi.org/10.1016/j.biortech.2007.11.035

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis : forecasting and control (3rd ed.). pp. 54–55. ISBN 0-13-060774-6. OCLC 28888762.* (3rd Edition). Prentice Hall, Englewood Cliffs, N.J., .

Boyle, W. C. (1977). ENERGY RECOVERY FROM SANITARY LANDFILLS - A REVIEW. In *Microbial Energy Conversion* (pp. 119–138). Elsevier. https://doi.org/10.1016/B978-0-08-021791-8.50019-6

BS. (2004). *BS ISO 15985:2004 Plastics – Determination of the ultimate anaerobic biodegradation and disintegration under high-solids anaerobic-digestion conditions – Method by analysis of released biogas.*

Buswell, A. M., & Mueller, H. F. (1952). Mechanism of Methane Fermentation. *Industrial & Engineering Chemistry*, *44*(3), 550–552. https://doi.org/10.1021/ie50507a033

Buyukkamaci, N., & Filibeli, A. (2004). Volatile fatty acid formation in an anaerobic hybrid reactor. *Process Biochemistry*, *39*(11), 1491–1494. https://doi.org/10.1016/S0032-9592(03)00295-4

Cabbai, V., Ballico, M., Aneggi, E., & Goi, D. (2013). BMP tests of source selected OFMSW to evaluate anaerobic codigestion with sewage sludge. *Waste Management*, *33*(7), 1626–1632. https://doi.org/10.1016/j.wasman.2013.03.020

Campolongo, F., Cariboni, J., & Saltelli, A. (2007). An effective screening design for sensitivity analysis of large models. *Environmental Modelling & Software*, *22*(10), 1509–1518. https://doi.org/10.1016/j.envsoft.2006.10.004

Catenacci, A., Grana, M., Malpei, F., & Ficara, E. (2021). Optimizing ADM1 Calibration and Input Characterization for Effective Co-Digestion Modelling. *Water*, *13*(21), 3100. https://doi.org/10.3390/w13213100

Catenacci, A., Santus, A., Malpei, F., & Ferretti, G. (2022). Early prediction of BMP tests: A step response method for estimating first-order model parameters. *Renewable Energy*, *188*, 184–194. https://doi.org/10.1016/j.renene.2022.02.017

Cavaleiro, A. J., Pereira, M. A., & Alves, M. (2008). Enhancement of methane production from long chain fatty acid based effluents. *Bioresource Technology*, *99*(10), 4086–4095. https://doi.org/10.1016/j.biortech.2007.09.005

Ceron-Vivas, A., Cáceres-Cáceres, K. T., Rincón-Pérez, A., & Cajigas, A. A. (2019). Influence of pH and the C/N ratio on the biogas production of wastewater. *Revista Facultad de Ingeniería Universidad de Antioquia*, *92*, 70–79. https://doi.org/10.17533/udea.redin.20190627

Chen, Y., Cheng, J. J., & Creamer, K. S. (2008). Inhibition of anaerobic digestion process: A review. *Bioresource Technology*, *99*(10), 4044–4064. https://doi.org/10.1016/j.biortech.2007.01.057

Chiu, S. L. H., Lo, I. M. C., Woon, K. S., & Yan, D. Y. S. (2016). Life cycle assessment of waste treatment strategy for sewage sludge and food waste in Macau: perspectives on environmental and energy production performance. *The International Journal of Life Cycle Assessment*, *21*(2), 176–189. https://doi.org/10.1007/s11367-015-1008-2

Cornell, J. A. (2002). *Experiments with Mixtures: Designs, Models and Analysis of Mixture Data* (3rd Edition). Wiley. https://doi.org/10.1002/9781118204221

Cu, T. T. T., Nguyen, T. X., Triolo, J. M., Pedersen, L., Le, V. D., Le, P. D., & Sommer, S. G. (2014a). Biogas Production from Vietnamese Animal Manure, Plant Residues and Organic Waste: Influence of Biomass Composition on Methane Yield. *Asian-Australasian Journal of Animal Sciences*, *28*(2), 280–289. https://doi.org/10.5713/ajas.14.0312

Cu, T. T. T., Nguyen, T. X., Triolo, J. M., Pedersen, L., Le, V. D., Le, P. D., & Sommer, S. G. (2014b). Biogas Production from Vietnamese Animal

Manure, Plant Residues and Organic Waste: Influence of Biomass Composition on Methane Yield. *Asian-Australasian Journal of Animal Sciences*, *28*(2), 280–289. https://doi.org/10.5713/ajas.14.0312

Cuetos, M. J., Gómez, X., Otero, M., & Morán, A. (2008). Anaerobic digestion of solid slaughterhouse waste (SHW) at laboratory scale: Influence of co-digestion with the organic fraction of municipal solid waste (OFMSW). *Biochemical Engineering Journal*, *40*(1), 99–106. https://doi.org/10.1016/j.bej.2007.11.019

Da Silva, C., Astals, S., Peces, M., Campos, J. L., & Guerrero, L. (2018). Biochemical methane potential (BMP) tests: Reducing test time by early parameter estimation. *Waste Management*, *71*, 19–24. https://doi.org/10.1016/j.wasman.2017.10.009

Dandikas, V., Heuwinkel, H., Lichti, F., Drewes, J. E., & Koch, K. (2015). Correlation between Biogas Yield and Chemical Composition of Grassland Plant Species. *Energy & Fuels*, *29*(11), 7221–7229. https://doi.org/10.1021/acs.energyfuels.5b01257

Dastjerdi, B., Strezov, V., Rajaeifar, M. A., Kumar, R., & Behnia, M. (2021). A systematic review on life cycle assessment of different waste to energy valorization technologies. *Journal of Cleaner Production*, *290*, 125747. https://doi.org/10.1016/j.jclepro.2020.125747

De Clercq, D., Wen, Z., Fei, F., Caicedo, L., Yuan, K., & Shang, R. (2020). Interpretable machine learning for predicting biomethane production in industrial-scale anaerobic co-digestion. *Science of The Total Environment*, *712*, 134574. https://doi.org/10.1016/j.scitotenv.2019.134574

De Vrieze, J., Raport, L., Willems, B., Verbrugge, S., Volcke, E., Meers, E., Angenent, L. T., & Boon, N. (2015). Inoculum selection influences the biochemical methane potential of agro-industrial substrates. *Microbial Biotechnology*, *8*(5), 776–786. https://doi.org/10.1111/1751-7915.12268

Dechrugsa, S., Kantachote, D., & Chaiprapat, S. (2013). Effects of inoculum to substrate ratio, substrate mix ratio and inoculum source on batch co-digestion of grass and pig manure. *Bioresource Technology*, *146*, 101–108. https://doi.org/10.1016/j.biortech.2013.07.051

Deng, Z., van Linden, N., Guillen, E., Spanjers, H., & van Lier, J. B. (2021). Recovery and applications of ammoniacal nitrogen from nitrogen-loaded residual streams: A review. *Journal of Environmental Management*, *295*, 113096. https://doi.org/10.1016/j.jenvman.2021.113096

Donoso-Bravo, A., Mailier, J., Martin, C., Rodríguez, J., Aceves-Lara, C. A., & Wouwer, A. Vande. (2011). Model selection, identification and validation in anaerobic digestion: A review. *Water Research*, *45*(17), 5347–5364. https://doi.org/10.1016/j.watres.2011.08.059

Donoso-Bravo, A., Olivares, D., Lesty, Y., & Bossche, H. Vanden. (2020). Exploitation of the ADM1 in a XXI century wastewater resource recovery facility (WRRF): The case of codigestion and thermal hydrolysis. *Water Research*, *175*, 115654. https://doi.org/10.1016/j.watres.2020.115654

Doublet, J., Boulanger, A., Ponthieux, A., Laroche, C., Poitrenaud, M., & Cacho Rivero, J. A. (2013). Predicting the biochemical methane potential of wide range of organic substrates by near infrared spectroscopy. *Bioresource Technology*, *128*, 252–258. https://doi.org/10.1016/j.biortech.2012.10.044

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, *195*(2), 216–222. https://doi.org/10.1016/0370-2693(87)91197-X

Duvenaud, D. K. (2014). *Automatic Model Construction with Gaussian Processes* [Thesis for Doctor of Philosophy]. University of Cambridge.

Eastman, J. A., & Ferguson, J. H. (1981). Solubilization of Particulate Organic Carbon during the Acid Phase of Anaerobic Digestion. *Water Pollution Control Federation*, *53*(3), 352–366.

Elbeshbishy, E., Nakhla, G., & Hafez, H. (2012). Biochemical methane potential (BMP) of food waste and primary sludge: Influence of inoculum pre-incubation and inoculum source. *Bioresource Technology*, *110*, 18–25. https://doi.org/10.1016/j.biortech.2012.01.025

ELMASHAD, H. M., Zeeman, G., van Loon, W. K. P., Bot, G. P. A., & Lettinga, G. (2004). Effect of temperature and temperature fluctuation on thermophilic anaerobic digestion of cattle manure. *Bioresource Technology*, *95*(2), 191–201. https://doi.org/10.1016/j.biortech.2003.07.013

Ersahin, M. E. (2018). Modeling the dynamic performance of full-scale anaerobic primary sludge digester using Anaerobic Digestion Model No. 1 (ADM1). *Bioprocess and Biosystems Engineering*, *41*(10), 1539–1545. https://doi.org/10.1007/s00449-018-1981-5

Esposito, G., Frunzo, L., Giordano, A., Liotta, F., Panico, A., & Pirozzi, F. (2012). Anaerobic co-digestion of organic wastes. *Reviews in Environmental Science and Bio/Technology*, *11*(4), 325–341. https://doi.org/10.1007/s11157-012-9277-8

European Biogas Association. (2022). *Statistical Report 2022. Tracking biogas and biomethane deployment across Europe*. https://www.europeanbiogas.eu/SR-2022/EBA/#our

Fogacs, G. (2012). *Biogas production from citrus wastes and chicken feather: pretreatment and co-digestion* [PhD thesis]. Chalmers University of Technology.

Fortela, D. L., Hernandez, R., Zappi, M., French, T. W., Bajpai, R., Chistoserdov, A., Revellame, E., & Holmes, W. (2016). Microbial Lipid Accumulation Capability of Activated Sludge Feeding on Short Chain Fatty Acids as Carbon Sources through Fed-Batch Cultivation. *Journal of Bioprocessing & Biotechniques*, *6*(4). https://doi.org/10.4172/2155-9821.1000275

Friedman, J., Hastie, T., & Tibshirani, R. (2012). Boosting and Additive Trees. In *The Elements of Statistical Learning* (pp. 299–344). Springer.

Gahlot, P., Balasundaram, G., Tyagi, V. K., Atabani, A. E., Suthar, S., Kazmi, A. A., Štěpanec, L., Juchelková, D., & Kumar, A. (2022). Principles and potential of thermal hydrolysis of sewage sludge to enhance anaerobic digestion. *Environmental Research*, *214*, 113856. https://doi.org/10.1016/j.envres.2022.113856

Galí, A., Benabdallah, T., Astals, S., & Mata-Alvarez, J. (2009). Modified version of ADM1 model for agro-waste application. *Bioresource Technology*, *100*(11), 2783–2790. https://doi.org/10.1016/j.biortech.2008.12.052

Gallert, C., & Winter, J. (1997). Mesophilic and thermophilic anaerobic digestion of source-sorted organic wastes: effect of ammonia on glucose degradation and methane production. *Applied Microbiology and Biotechnology*, *48*(3), 405–410. https://doi.org/10.1007/s002530051071

Gandhi, B. P., Otite, S. V., Fofie, E. A., Lag-Brotons, A. J., Ezemonye, L. I., Semple, K. T., & Martin, A. D. (2022). Kinetic investigations into the effect of inoculum to substrate ratio on batch anaerobic digestion of simulated food waste. *Renewable Energy*, *195*, 311–321. https://doi.org/10.1016/j.renene.2022.05.134

Garitselov, O., Mohanty, S. P., & Kougianos, E. (2012). A Comparative Study of Metamodels for Fast and Accurate Simulation of Nano-CMOS Circuits. *IEEE Transactions on Semiconductor Manufacturing*, *25*(1), 26–36. https://doi.org/10.1109/TSM.2011.2173957

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(6), 721–741. https://doi.org/10.1109/TPAMI.1984.4767596

Geng, Z., Shi, X., Ma, B., Chu, C., & Han, Y. (2024). Biogas production prediction model of food waste anaerobic digestion for energy optimization using mixup data augmentation-based global attention mechanism. *Environmental Science and Pollution Research*, *31*(6), 9121–9134. https://doi.org/10.1007/s11356-023-31653-8

Gerardi, M. H. (2003). Toxicity. In *The Microbiology of Anaerobic Digesters* (Ed. 1, pp. 105–115). Wiley. https://doi.org/10.1002/0471468967

Gompertz, B. (1825). On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies. *Philosophical Transactions of the Royal Society of London*, *115 (1825)*, 513–583.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144. https://doi.org/10.1145/3422622

Grady Jr., C. P. L., Daigger, G. T., Love, N. G., & Filipe, C. D. M. (2011). *Biological Wastewater Treatment*. CRC Press. https://doi.org/10.1201/b13775

Gu, Y., Chen, X., Liu, Z., Zhou, X., & Zhang, Y. (2014). Effect of inoculum sources on the anaerobic digestion of rice straw. *Bioresource Technology*, *158*, 149–155. https://doi.org/10.1016/j.biortech.2014.02.011

Gueguim Kana, E. B., Oloke, J. K., Lateef, A., & Adesiyan, M. O. (2012). Modeling and optimization of biogas production on saw dust and other co-substrates using Artificial Neural network and Genetic Algorithm. *Renewable Energy*, *46*, 276–281. https://doi.org/10.1016/j.renene.2012.03.027

Guisasola, A., Baeza, J. A., Carrera, J., Sin, G., Vanrolleghem, P. A., & Lafuente, J. (2006). The Influence of Experimental Data Quality and Quantity on Parameter Estimation Accuracy. *Education for Chemical Engineers*, *1*(1), 139–145. https://doi.org/10.1205/ece06016

Gujarati, D. N. (2020). Multicollinearity: What Happens if the Regressors Are Correlated. In *Basic Economics* (4th Edition, pp. 341–387). McGraw−Hill.

Gunaseelan, V. N. (2007). Regression models of ultimate methane yields of fruits and vegetable solid wastes, sorghum and napiergrass on chemical composition. *Bioresource Technology*, *98*(6), 1270–1277. https://doi.org/10.1016/j.biortech.2006.05.014

Gupta, R., Zhang, L., Hou, J., Zhang, Z., Liu, H., You, S., Sik Ok, Y., & Li, W. (2023). Review of explainable machine learning for anaerobic digestion. *Bioresource Technology*, *369*, 128468. https://doi.org/10.1016/j.biortech.2022.128468

Gupta, V. (2021). *Correlation in XGboost*. Medium. URL: https://vishesh-gupta.medium.com/correlation-in-xgboost-8afa649bd066

Gurung, R. B., Ogle, S. M., Breidt, F. J., Williams, S. A., & Parton, W. J. (2020). Bayesian calibration of the DayCent ecosystem model to simulate soil organic carbon dynamics and reduce model uncertainty. *Geoderma*, *376*, 114529. https://doi.org/10.1016/j.geoderma.2020.114529

Hansen, T. L., Schmidt, J. E., Angelidaki, I., Marca, E., Jansen, J. la C., Mosbæk, H., & Christensen, T. H. (2004). Method for determination of methane potentials of solid organic waste. *Waste Management*, *24*(4), 393–400. https://doi.org/10.1016/j.wasman.2003.09.009

Hashimoto, A. G. (1989). Effect of inoculum/substrate ratio on methane yield and production rate from straw. *Biological Wastes*, *28*(4), 247–255. https://doi.org/10.1016/0269-7483(89)90108-0

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, *57*(1), 97. https://doi.org/10.2307/2334940

Hatfield, R., & Fukushima, R. S. (2005). Can Lignin Be Accurately Measured? *Crop Science*, *45*(3), 832–839. https://doi.org/10.2135/cropsci2004.0238

Hoffman, M. D., & Gelman, A. (2011). *The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo*.

Holliger, C., Alves, M., Andrade, D., Angelidaki, I., Astals, S., Baier, U., Bougrier, C., Buffière, P., Carballa, M., de Wilde, V., Ebertseder, F., Fernández, B., Ficara, E., Fotidis, I., Frigon, J.-C., de Laclos, H. F., Ghasimi, D. S. M., Hack, G., Hartel, M., … Wierinck, I. (2016). Towards a standardization of biomethane potential tests. *Water Science and Technology*, *74*(11), 2515–2522. https://doi.org/10.2166/wst.2016.336

Holubar, P. (2002). Advanced controlling of anaerobic digestion by means of hierarchical neural networks. *Water Research*, *36*(10), 2582–2588. https://doi.org/10.1016/S0043-1354(01)00487-0

Hülsemann, B., Zhou, L., Merkle, W., Hassa, J., Müller, J., & Oechsner, H. (2020). Biomethane Potential Test: Influence of Inoculum and the Digestion System. *Applied Sciences*, *10*(7), 2589. https://doi.org/10.3390/app10072589

Jaxa-Rozen, M., & Kwakkel, J. (2018). Tree-based ensemble methods for sensitivity analysis of environmental models: A performance comparison with Sobol and Morris techniques. *Environmental Modelling & Software*, *107*, 245–266. https://doi.org/10.1016/j.envsoft.2018.06.011

Jensen, P. D., Ge, H., & Batstone, D. J. (2011). Assessing the role of biochemical methane potential tests in determining anaerobic degradability rate and extent. *Water Science and Technology*, *64*(4), 880–886. https://doi.org/10.2166/wst.2011.662

Jeong, K., Abbas, A., Shin, J., Son, M., Kim, Y. M., & Cho, K. H. (2021). Prediction of biogas production in anaerobic co-digestion of organic wastes using deep learning models. *Water Research*, *205*, 117697. https://doi.org/10.1016/j.watres.2021.117697

Jiang, Y., McAdam, E., Zhang, Y., Heaven, S., Banks, C., & Longhurst, P. (2019). Ammonia inhibition and toxicity in anaerobic digestion: A critical review.

Journal of Water Process Engineering, *32*, 100899. https://doi.org/10.1016/j.jwpe.2019.100899

Jiang, Y., Xie, S. H., Dennehy, C., Lawlor, P. G., Hu, Z. H., Wu, G. X., Zhan, X. M., & Gardiner, G. E. (2020). Inactivation of pathogens in anaerobic digestion systems for converting biowastes to bioenergy: A review. *Renewable and Sustainable Energy Reviews*, *120*, 109654. https://doi.org/10.1016/j.rser.2019.109654

Jiang, Y., Zhang, Y., Banks, C., Heaven, S., & Longhurst, P. (2017). Investigation of the impact of trace elements on anaerobic volatile fatty acid degradation using a fractional factorial experimental design. *Water Research*, *125*, 458–465. https://doi.org/10.1016/j.watres.2017.09.010

Jin, W., Xu, X., Yang, F., Li, C., & Zhou, M. (2018). Performance enhancement by rumen cultures in anaerobic co-digestion of corn straw with pig manure. *Biomass and Bioenergy*, *115*, 120–129. https://doi.org/10.1016/j.biombioe.2018.05.001

Jones, E. , O. T. , & P. P. (2001). *SciPy: Open Source Scientific Tools for Python.* (v 1.3.1).

Jorgensen, T. C., & Weatherley, L. R. (2003). Ammonia removal from wastewater by ion exchange in the presence of organic contaminants. *Water Research*, *37*(8), 1723–1728. https://doi.org/10.1016/S0043-1354(02)00571-7

Ju, X., Engelhard, M., & Zhang, X. (2013). An advanced understanding of the specific effects of xylan and surface lignin contents on enzymatic hydrolysis of lignocellulosic biomass. *Bioresource Technology*, *132*, 137–145. https://doi.org/10.1016/j.biortech.2013.01.049

Kafle, G. K., & Chen, L. (2016). Comparison on batch anaerobic digestion of five different livestock manures and prediction of biochemical methane potential (BMP) using different statistical models. *Waste Management*, *48*, 492–502. https://doi.org/10.1016/j.wasman.2015.10.021

Kalnins, A., & Praitis Hill, K. (2023). The VIF Score. What is it Good For? Absolutely Nothing. *Organizational Research Methods*. https://doi.org/10.1177/10944281231216381

Kaneko, J. J., Harvey, J. W., & Bruss, M. L. (2008). Chapter 4 - Lipids and Ketones. In *Clinical Biochemistry of Domestic Animals* (pp. 81–115). Elsevier. https://doi.org/10.1016/B978-0-12-370491-7.X0001-3

Kallioinen, N., Paananen, T., Bürkner, P.-C., & Vehtari, A. (2021). *Detecting and diagnosing prior and likelihood sensitivity with power-scaling*. https://doi.org/10.1007/s11222-023-10366-5

Khalid, A., Arshad, M., Anjum, M., Mahmood, T., & Dawson, L. (2011). The anaerobic digestion of solid organic waste. *Waste Management*, *31*(8), 1737–1744. https://doi.org/10.1016/j.wasman.2011.03.021

Khashaba, N. H., Ettouney, R. S., Abdelaal, M. M., Ashour, F. H., & El-Rifai, M. A. (2022). Artificial neural network modeling of biochar enhanced anaerobic sewage sludge digestion. *Journal of Environmental Chemical Engineering*, *10*(4), 107988. https://doi.org/10.1016/j.jece.2022.107988

KIDA, K., SHIGEMATSU, T., KIJIMA, J., NUMAGUCHI, M., MOCHINAGA, Y., ABE, N., & MORIMURA, S. (2001). Influence of Ni2+ and Co2+ on Methanogenic Activity and the Amounts of Coenzymes Involved in Methanogenesis. *Journal of Bioscience and Bioengineering*, *91*(6), 590–595. https://doi.org/10.1263/jbb.91.590

Kim, J. K., Oh, B. R., Chun, Y. N., & Kim, S. W. (2006). Effects of temperature and hydraulic retention time on anaerobic digestion of food waste. *Journal of Bioscience and Bioengineering*, *102*(4), 328–332. https://doi.org/10.1263/jbb.102.328

Koch, K., Bajón Fernández, Y., & Drewes, J. E. (2015). Influence of headspace flushing on methane production in Biochemical Methane Potential (BMP) tests. *Bioresource Technology*, *186*, 173–178. https://doi.org/10.1016/j.biortech.2015.03.071

Koch, K., & Drewes, J. E. (2014). Alternative approach to estimate the hydrolysis rate constant of particulate material from batch data. *Applied Energy*, *120*, 11–15. https://doi.org/10.1016/j.apenergy.2014.01.050

Koch, K., Lippert, T., & Drewes, J. E. (2017). The role of inoculum's origin on the methane yield of different substrates in biochemical methane potential (BMP) tests. *Bioresource Technology*, *243*, 457–463. https://doi.org/10.1016/j.biortech.2017.06.142

Kolmogorov, A. (1956). On the representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables. *Proceedings of the USSR Academy of Sciences*, *108*, 179–182.

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Kumar, A., & Samadder, S. R. (2020). Performance evaluation of anaerobic digestion technology for energy recovery from organic fraction of municipal solid waste: A review. *Energy*, *197*, 117253. https://doi.org/10.1016/j.energy.2020.117253

Kuo, W., & Cheng, K. (2007). Use of respirometer in evaluation of process and toxicity of thermophilic anaerobic digestion for treating kitchen waste. *Bioresource Technology*, *98*(9), 1805–1811. https://doi.org/10.1016/j.biortech.2006.06.016

Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. (2005) Applied Linear Statistical Models. 5th Edition, McGraw-Hill, Irwin, New York.

Labatut, R. A., Angenent, L. T., & Scott, N. R. (2011a). Biochemical methane potential and biodegradability of complex organic substrates. *Bioresource Technology*, *102*(3), 2255–2264. https://doi.org/10.1016/j.biortech.2010.10.035

Labatut, R. A., Angenent, L. T., & Scott, N. R. (2011b). Biochemical methane potential and biodegradability of complex organic substrates. *Bioresource*

Technology, *102*(3), 2255–2264. https://doi.org/10.1016/j.biortech.2010.10.035

LEI, X., & MAEKAWA, T. (2007). Electrochemical treatment of anaerobic digestion effluent using a Ti/Pt–IrO2 electrode. *Bioresource Technology*, *98*(18), 3521–3525. https://doi.org/10.1016/j.biortech.2006.11.018

Li, C., Nges, I. A., Lu, W., & Wang, H. (2017). Assessment of the degradation efficiency of full-scale biogas plants: A comparative study of degradation indicators. *Bioresource Technology*, *244*, 304–312. https://doi.org/10.1016/j.biortech.2017.07.157

Li, W., Khalid, H., Zhu, Z., Zhang, R., Liu, G., Chen, C., & Thorin, E. (2018). Methane production through anaerobic digestion: Participation and digestion characteristics of cellulose, hemicellulose and lignin. *Applied Energy*, *226*, 1219–1228. https://doi.org/10.1016/j.apenergy.2018.05.055

LIEBETRAU, J., PFEIFFER, D., & THRÄN, D. (2017). *Collection of measurement methods for biogas – meth- ods to determine parameters for analysis purposes and parameters that describe processes in the biogas sector. Series of the funding programme "Biomass energy use". DBFZ, Leipzig.*

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, *37*(1), 145–151. https://doi.org/10.1109/18.61115

Lisboa, M. S., & Lansing, S. (2013). Characterizing food waste substrates for co-digestion through biochemical methane potential (BMP) experiments. *Waste Management*, *33*(12), 2664–2669. https://doi.org/10.1016/j.wasman.2013.09.004

Liu, Y., Guo, W., Longhurst, P., & Jiang, Y. (2023). Shortening the Standard Testing Time for Residual Biogas Potential (RBP) Tests Using Biogas Yield Models and Substrate Physicochemical Characteristics. *Processes*, *11*(2), 441. https://doi.org/10.3390/pr11020441

Liu, Y., Jiang, Y., & Bortone, I. (2022). A Scheme for Anaerobic Digestion Modelling and ADM1 Model Calibration. *CHEMICAL ENGINEERING TRANSACTIONS*, *96*, 433–438.

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*.

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). *Consistent Individualized Feature Attribution for Tree Ensembles*.

Ma, J., Zhao, Q.-B., Laurens, L. L. M., Jarvis, E. E., Nagle, N. J., Chen, S., & Frear, C. S. (2015a). Mechanism, kinetics and microbiology of inhibition caused by long-chain fatty acids in anaerobic digestion of algal biomass. *Biotechnology for Biofuels*, *8*(1), 141. https://doi.org/10.1186/s13068-015-0322-z

Ma, J., Zhao, Q.-B., Laurens, L. L. M., Jarvis, E. E., Nagle, N. J., Chen, S., & Frear, C. S. (2015b). Mechanism, kinetics and microbiology of inhibition caused by long-chain fatty acids in anaerobic digestion of algal biomass. *Biotechnology for Biofuels*, *8*(1), 141. https://doi.org/10.1186/s13068-015-0322-z

Marquardt, D. W. (1963). An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics*, *11*(2), 431–441. https://doi.org/10.1137/0111030

Masoero, L., Stephenson, W. T., & Broderick, T. Sensitivity of Bayesian Inference to Data Perturbations. *1st Symposium on Advances in Approximate Bayesian Inference*, 2018 1-6

Masse, L., Massé, D. I., Kennedy, K. J., & Chou, S. P. (2002). Neutral fat hydrolysis and long-chain fatty acid oxidation during anaerobic digestion of slaughterhouse wastewater. *Biotechnology and Bioengineering*, *79*(1), 43–52. https://doi.org/10.1002/bit.10284

MathWorks. (2021). *fitrtree. Algorithms: Node Splitting Rules*. Https://Www.Mathworks.Com/Help/Stats/Fitrtree.Html#bunrdhe-1.

Matuszewska, A., Owczuk, M., Zamojska-Jaroszewicz, A., Jakubiak-Lasocka, J., Lasocki, J., & Orliński, P. (2016). Evaluation of the biological methane potential of various feedstock for the production of biogas to supply agricultural tractors. *Energy Conversion and Management*, *125*, 309–319. https://doi.org/10.1016/j.enconman.2016.02.072

Mauseth, J. D. (1988). *Plant Anatomy*. The Benjamin/Cummings Publishing Company.

McCarty, P. L. (1972). Energetics of organic matter degradation. In *Water Pollution Microbioogy* (pp. 91–118). John Wiley & Sons, Inc.

McNaught, A. D., & Wilkinson, A. ,. (1997). *Compendium of Chemical Terminology* (second ed.). Blackwell Scientific Publications.

Meijei, C. (1936). Über Whittakersche bezw. Besselsche Funktionen und deren Produkten,. *Nieuw Arch. Wisk.*, *18*, 10–39.

Mellin, Hj. (1910). Abri€ einer einheitlichen Theorie der Gamma- und der hypergeometrischen Funktionen. *Mathematische Annalen*, *68*(3), 305–337. https://doi.org/10.1007/BF01475775

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

Mlinar, S., Weig, A. R., & Freitag, R. (2022). Influence of NH3 and NH4+ on anaerobic digestion and microbial population structure at increasing total ammonia nitrogen concentrations. *Bioresource Technology*, *361*, 127638. https://doi.org/10.1016/j.biortech.2022.127638

Monod, J. (1942). *Recherches sur la croissance des cultures bactériennes. Paris, France: Hermann et Cie;*

Montgomery, D. C. (2020). Response Surface Methods and Designs. In *Design and Analysis of Experiments* (10th Edition, pp. 408–472). Wiley.

Morris, M. D. (1991). Factorial Sampling Plans for Preliminary Computational Experiments. *Technometrics*, *33*(2), 161. https://doi.org/10.2307/1269043

Mussatto, S. I., Fernandes, M., Milagres, A. M. F., & Roberto, I. C. (2008). Effect of hemicellulose and lignin on enzymatic hydrolysis of cellulose from brewer's spent grain. *Enzyme and Microbial Technology*, *43*(2), 124–129. https://doi.org/10.1016/j.enzmictec.2007.11.006

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*. https://doi.org/10.3389/fnbot.2013.00021

National Bureau of Standards. (1982). Table of chemical thermodynamic properties, National Bureau of Standards. *Journal of Physics and Chemical Reference Data.*, *11*(2).

National Non-Food Crops Centre. (2022). *Anaerobic Digestion Development in the UK (2022 edition)*. https://www.nnfcc.co.uk/press-release-ad-report-2022

NAUMANN, K., & BASSLER, R. (1976). *Chemische Unter- suchung von Futtermitteln. VDLUFA Methodenhand- buch, 3. Auflage, Neumann-Neudamm.*

Neves, L., Oliveira, R., & Alves, M. M. (2006). Anaerobic co-digestion of coffee waste and sewage sludge. *Waste Management*, *26*(2), 176–181. https://doi.org/10.1016/j.wasman.2004.12.022

Nguyen, H. H. (2014). *Modelling of food waste digestion using ADM1 integrated with Aspen Plus* [Thesis for the degree of Doctor of Philosophy]. University of Southampton.

Nguyen, H.-V., & Vreeken, J. (2015). *Non-parametric Jensen-Shannon Divergence* (pp. 173–189). https://doi.org/10.1007/978-3-319-23525-7_11

Nimmegeers, P., Lauwers, J., Telen, D., Logist, F., & Impe, J. Van. (2017). Identifiability of large-scale non-linear dynamic network models applied to

the ADM1-case study. *Mathematical Biosciences*, *288*, 21–34. https://doi.org/10.1016/j.mbs.2017.02.008

Nowotny, M. (2020). *divergence*. https://github.com/michaelnowotny/divergence/tree/master/divergence

Ortega-Martinez, E., Sapkaite, I., Fdz-Polanco, F., & Donoso-Bravo, A. (2016). From pre-treatment toward inter-treatment. Getting some clues from sewage sludge biomethanation. *Bioresource Technology*, *212*, 227–235. https://doi.org/10.1016/j.biortech.2016.04.049

Otuzalti, M. M., & Perendeci, N. A. (2018). Modeling of real scale waste activated sludge anaerobic digestion process by Anaerobic Digestion Model 1 (ADM1). *International Journal of Green Energy*, *15*(7), 454–464. https://doi.org/10.1080/15435075.2018.1479265

Ozgun, H. (2019). Anaerobic Digestion Model No. 1 (ADM1) for mathematical modeling of full-scale sludge digester performance in a municipal wastewater treatment plant. *Biodegradation*, *30*(1), 27–36. https://doi.org/10.1007/s10532-018-9859-4

Pabón-Pereira, C. P., Hamelers, H. V. M., Matilla, I., & van Lier, J. B. (2020). New Insights on the Estimation of the Anaerobic Biodegradability of Plant Material: Identifying Valuable Plants for Sustainable Energy Production. *Processes*, *8*(7), 806. https://doi.org/10.3390/pr8070806

Palatsi, J., Affes, R., Fernandez, B., Pereira, M. A., Alves, M. M., & Flotats, X. (2012). Influence of adsorption and anaerobic granular sludge characteristics on long chain fatty acids inhibition process. *Water Research*, *46*(16), 5268–5278. https://doi.org/10.1016/j.watres.2012.07.008

Palatsi, J., Laureni, M., Andrés, M. V., Flotats, X., Nielsen, H. B., & Angelidaki, I. (2009). Strategies for recovering inhibition caused by long chain fatty acids on anaerobic thermophilic biogas reactors. *Bioresource Technology*, *100*(20), 4588–4596. https://doi.org/10.1016/j.biortech.2009.04.046

Parajuli, P. (2011). *Biogas measurement techniques and the associated errors* [Master's Thesis]. University of Jyväskylä.

Pastor, L., Mangin, D., Ferrer, J., & Seco, A. (2010). Struvite formation from the supernatants of an anaerobic digestion pilot plant. *Bioresource Technology*, *101*(1), 118–125. https://doi.org/10.1016/j.biortech.2009.08.002

Pastor-Poquet, V., Papirio, S., Harmand, J., Steyer, J.-P., Trably, E., Escudié, R., & Esposito, G. (2019a). Assessing practical identifiability during calibration and cross-validation of a structured model for high-solids anaerobic digestion. *Water Research*, *164*, 114932. https://doi.org/10.1016/j.watres.2019.114932

Pastor-Poquet, V., Papirio, S., Harmand, J., Steyer, J.-P., Trably, E., Escudié, R., & Esposito, G. (2019b). Assessing practical identifiability during calibration and cross-validation of a structured model for high-solids anaerobic digestion. *Water Research*, *164*, 114932. https://doi.org/10.1016/j.watres.2019.114932

Pavlostathis, S. G., & Giraldo-Gomez, E. (1991). Kinetics of anaerobic treatment: A critical review. *Critical Reviews in Environmental Control*, *21*(5–6), 411–490. https://doi.org/10.1080/10643389109388424

Peces, M., Astals, S., Jensen, P. D., & Clarke, W. P. (2021). Transition of microbial communities and degradation pathways in anaerobic digestion at decreasing retention time. *New Biotechnology*, *60*, 52–61. https://doi.org/10.1016/j.nbt.2020.07.005

Pehlivanoglu, E., & Sedlak, D. L. (2004). Bioavailability of wastewater-derived organic nitrogen to the alga Selenastrum Capricornutum. *Water Research*, *38*(14–15), 3189–3196. https://doi.org/10.1016/j.watres.2004.04.027

Pererva, Y., Miller, C. D., & Sims, R. C. (2020). Existing Empirical Kinetic Models in Biochemical Methane Potential (BMP) Testing, Their Selection and Numerical Solution. *Water*, *12*(6), 1831. https://doi.org/10.3390/w12061831

Phothilangka, P., Schoen, M. A., Huber, M., Luchetta, P., Winkler, T., & Wett, B. (2008). Prediction of thermal hydrolysis pretreatment on anaerobic digestion of waste activated sludge. *Water Science and Technology*, *58*(7), 1467–1473. https://doi.org/10.2166/wst.2008.726

Picard-Weibel, A., Capson-Tojo, G., Guedj, B., & Moscoviz, R. (2024). Bayesian uncertainty quantification for anaerobic digestion models. *Bioresource Technology*, *394*, 130147. https://doi.org/10.1016/j.biortech.2023.130147

Ponsá, S., Gea, T., & Sánchez, A. (2011). Short-time estimation of biogas and methane potentials from municipal solid wastes. *Journal of Chemical Technology & Biotechnology*, *86*(8), 1121–1124. https://doi.org/10.1002/jctb.2615

Postawa, K., Szczygieł, J., & Kułażyński, M. (2020). Heuristic methods in optimization of selected parameters of Two-Phase Anaerobic Digestion (TPAD) model. *Fuel*, *281*, 118257. https://doi.org/10.1016/j.fuel.2020.118257

Pramanik, S. K., Suja, F. B., Porhemmat, M., & Pramanik, B. K. (2019). Performance and Kinetic Model of a Single-Stage Anaerobic Digestion System Operated at Different Successive Operating Stages for the Treatment of Food Waste. *Processes*, *7*(9), 600. https://doi.org/10.3390/pr7090600

Puyuelo, B., Ponsá, S., Gea, T., & Sánchez, A. (2011). Determining C/N ratios for typical organic wastes using biodegradable fractions. *Chemosphere*, *85*(4), 653–659. https://doi.org/10.1016/j.chemosphere.2011.07.014

Python Software Foundation. (2020). *pandas: Powerful data analysis tools for Python.* (v 1.3.3).

Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N., & Fanos, V. (2020). Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment. *Medicina*, *56*(9), 455. https://doi.org/10.3390/medicina56090455

Rao, P. V., & Baral, S. S. (2011). Experimental design of mixture for the anaerobic co-digestion of sewage sludge. *Chemical Engineering Journal*, *172*(2–3), 977–986. https://doi.org/10.1016/j.cej.2011.07.010

Raposo, F., Banks, C. J., Siegert, I., Heaven, S., & Borja, R. (2006a). Influence of inoculum to substrate ratio on the biochemical methane potential of maize in batch tests. *Process Biochemistry*, *41*(6), 1444–1450. https://doi.org/10.1016/j.procbio.2006.01.012

Raposo, F., Banks, C. J., Siegert, I., Heaven, S., & Borja, R. (2006b). Influence of inoculum to substrate ratio on the biochemical methane potential of maize in batch tests. *Process Biochemistry*, *41*(6), 1444–1450. https://doi.org/10.1016/j.procbio.2006.01.012

Raposo, F., De la Rubia, M. A., Fernández-Cegrí, V., & Borja, R. (2012). Anaerobic digestion of solid organic substrates in batch mode: An overview relating to methane yields and experimental procedures. *Renewable and Sustainable Energy Reviews*, *16*(1), 861–877. https://doi.org/10.1016/j.rser.2011.09.008

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes in Machine Learning* (T. Dietterich, Ed.). the MIT press. http://www.tuebingen.mpg.de/~carl

Rath, J., Heuwinkel, H., & Herrmann, A. (2013). Specific Biogas Yield of Maize Can Be Predicted by the Interaction of Four Biochemical Constituents. *BioEnergy Research*, *6*(3), 939–952. https://doi.org/10.1007/s12155-013-9318-3

Razaviarani, V., & Buchanan, I. D. (2015). Calibration of the Anaerobic Digestion Model No. 1 (ADM1) for steady-state anaerobic co-digestion of municipal wastewater sludge with restaurant grease trap waste. *Chemical Engineering Journal*, *266*, 91–99. https://doi.org/10.1016/j.cej.2014.12.080

Rezania, S., Oryani, B., Nasrollahi, V. R., Darajeh, N., Lotfi Ghahroud, M., & Mehranzamir, K. (2023). Review on Waste-to-Energy Approaches toward a

Circular Economy in Developed and Developing Countries. *Processes*, *11*(9), 2566. https://doi.org/10.3390/pr11092566

Ripley, L. E., Boyle, W. C., & Converse, J. C. (1986). Improved Alkalimetric Monitoring for Anaerobic Digestion of High-Strength Wastes. *Journal WPCF*, *48*, 406–411.

Rocamora, I., Wagland, S. T., Villa, R., Simpson, E. W., Fernández, O., & Bajón-Fernández, Y. (2020). Dry anaerobic digestion of organic waste: A review of operational parameters and their impact on process performance. *Bioresource Technology*, *299*, 122681. https://doi.org/10.1016/j.biortech.2019.122681

Rosen, C., & Jeppsson, U. (2006). *Aspects on ADM1 Implementation within the BSM2 Framework*. https://www.researchgate.net/publication/228684581

Rozzi, A., & Remigi, E. (2004). Methods of assessing microbial activity and inhibition under anaerobic conditions: a literature review. *Reviews in Environmental Science and Bio/Technology*, *3*(2), 93–115. https://doi.org/10.1007/s11157-004-5762-z

Saabas, A. (2015). *treeinterpreter*.

Sahni, N., Simon, G., & Arora, R. (2018). Development and Validation of Machine Learning Models for Prediction of 1-Year Mortality Utilizing Electronic Medical Record Data Available at the End of Hospitalization in Multicondition Patients: a Proof-of-Concept Study. *Journal of General Internal Medicine*, *33*(6), 921–928. https://doi.org/10.1007/s11606-018-4316-y

Salamattalab, M. M., Hasani Zonoozi, M., & Molavi-Arabshahi, M. (2024). Innovative approach for predicting biogas production from large-scale anaerobic digester using long-short term memory (LSTM) coupled with genetic algorithm (GA). *Waste Management*, *175*, 30–41. https://doi.org/10.1016/j.wasman.2023.12.046

Saltelli, A. (2002a). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, *145*(2), 280–297. https://doi.org/10.1016/S0010-4655(02)00280-1

Saltelli, A. (2002b). Sensitivity Analysis for Importance Assessment. *Risk Analysis*, *22*(3), 579–590. https://doi.org/10.1111/0272-4332.00040

Saltelli, A., Aleksankina, K., Becker, W., Fennell, P., Ferretti, F., Holst, N., Li, S., & Wu, Q. (2019). Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices. *Environmental Modelling & Software*, *114*, 29–39. https://doi.org/10.1016/J.ENVSOFT.2019.01.012

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., & Tarantola, S. (2007). Global Sensitivity Analysis. The Primer. In *Global Sensitivity Analysis. The Primer*. Wiley. https://doi.org/10.1002/9780470725184

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*, e55. https://doi.org/10.7717/peerj-cs.55

Sánchez, E., Borja, R., Weiland, P., Travieso, L., & Martín, A. (2001). Effect of substrate concentration and temperature on the anaerobic digestion of piggery waste in a tropical climate. *Process Biochemistry*, *37*(5), 483–489. https://doi.org/10.1016/S0032-9592(01)00240-0

Santos, V. M. O. dos, Caldara, F. R., Seno, L. de O., Feijó, G. L. D., Paz, I. C. de L. A., Garcia, R. G., Nääs, I. de A., & Altemio, Â. D. C. (2012). Marinade with alkaline solutions for the improvement of pork quality. *Pesquisa Agropecuária Brasileira*, *47*(11), 1655–1662. https://doi.org/10.1590/S0100-204X2012001100013

Satterthwaite, F. E. (1959). Random balance experimentation. *Technometrics*, *1*(2).

Seyboldt, A. (2020). *Sunode*. https://sunode.readthedocs.io/en/latest/index.html

Shana, A., Ouki, S., Asaadi, M., Pearce, P., & Mancini, G. (2013). The impact of intermediate thermal hydrolysis on the degradation kinetics of carbohydrates in sewage sludge. *Bioresource Technology*, *137*, 239–244. https://doi.org/10.1016/j.biortech.2013.03.121

Shen, Y., Linville, J. L., Ignacio-de Leon, P. A. A., Schoene, R. P., & Urgun-Demirtas, M. (2016). Towards a sustainable paradigm of waste-to-energy process: Enhanced anaerobic digestion of sludge with woody biochar. *Journal of Cleaner Production*, *135*, 1054–1064. https://doi.org/10.1016/j.jclepro.2016.06.144

Shen, Y., Linville, J. L., Urgun-Demirtas, M., Schoene, R. P., & Snyder, S. W. (2015). Producing pipeline-quality biomethane via anaerobic digestion of sludge amended with corn stover biochar with in-situ CO2 removal. *Applied Energy*, *158*, 300–309. https://doi.org/10.1016/j.apenergy.2015.08.016

Shi, X., Lin, J., Zuo, J., Li, P., Li, X., & Guo, X. (2017). Effects of free ammonia on volatile fatty acid accumulation and process performance in the anaerobic digestion of two typical bio-wastes. *Journal of Environmental Sciences*, *55*, 49–57. https://doi.org/10.1016/j.jes.2016.07.006

Shumway, R. H., & Stoffer, D. S. (2017). *Time Series Analysis and Its Applications*. Springer International Publishing. https://doi.org/10.1007/978-3-319-52452-8

Sobol′, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, *55*(1–3), 271–280. https://doi.org/10.1016/S0378-4754(00)00270-6

Song, X., Shi, Z., Li, X., Wang, X., & Ren, Y. (2019). Fate of proteins of waste activated sludge during thermal alkali pretreatment in terms of sludge protein recovery. *Frontiers of Environmental Science & Engineering*, *13*(2), 25. https://doi.org/10.1007/s11783-019-1114-7

Song, Y.-C., Feng, Q., & Ahn, Y. (2016). Performance of the Bio-electrochemical Anaerobic Digestion of Sewage Sludge at Different Hydraulic Retention

Times. *Energy & Fuels*, *30*(1), 352–359. https://doi.org/10.1021/acs.energyfuels.5b02003

Souza, T. S. O., Carvajal, A., Donoso-Bravo, A., Peña, M., & Fdz-Polanco, F. (2013). ADM1 calibration using BMP tests for modeling the effect of autohydrolysis pretreatment on the performance of continuous sludge digesters. *Water Research*, *47*(9), 3244–3254. https://doi.org/10.1016/j.watres.2013.03.041

Souza, T. S. O., Ferreira, L. C., Sapkaite, I., Pérez-Elvira, S. I., & Fdz-Polanco, F. (2013). Thermal pretreatment and hydraulic retention time in continuous digesters fed with sewage sludge: Assessment using the ADM1. *Bioresource Technology*, *148*, 317–324. https://doi.org/10.1016/j.biortech.2013.08.161

Sponza, D. T. (2003). Toxicity and treatability of carbontetrachloride and tetrachloroethylene in anaerobic batch cultures. *International Biodeterioration & Biodegradation*, *51*(2), 119–127. https://doi.org/10.1016/S0964-8305(02)00095-1

Standing Committee of Analysts (SCA). (1979). Determination of Volatile Fatty Acids in Sewage Sludge (1979). In *Methods for the Examination of Waters and Associated Materials.* HMSO.

Strömberg, S., Nistor, M., & Liu, J. (2014). Towards eliminating systematic errors caused by the experimental conditions in Biochemical Methane Potential (BMP) tests. *Waste Management*, *34*(11), 1939–1948. https://doi.org/10.1016/j.wasman.2014.07.018

Strömberg, S., Nistor, M., & Liu, J. (2015). Early prediction of Biochemical Methane Potential through statistical and kinetic modelling of initial gas production. *Bioresource Technology*, *176*, 233–241. https://doi.org/10.1016/j.biortech.2014.11.033

Tan, P.-N., Steinbach, M., & Kumar, V. (2014). Classification: Basic Concepts, Decision Trees, and Model Evaluation. In *Introduction to Data Mining*. Pearson Education Limited.

Tarantola, S., Gatelli, D., & Mara, T. A. (2006). Random balance designs for the estimation of first order global sensitivity indices. *Reliability Engineering & System Safety*, *91*(6), 717–727. https://doi.org/10.1016/j.ress.2005.06.003

Tchobanoglous, G., Burton, F. L., & Stensel, H. D. (2003). *Wastewater Engineering - Treatment, Disposal and Reuse*. McGraw Hill.

Thamsiriroj, T., & Murphy, J. D. (2011). Modelling mono-digestion of grass silage in a 2-stage CSTR anaerobic digester using ADM1. *Bioresource Technology*, *102*(2), 948–959. https://doi.org/10.1016/j.biortech.2010.09.051

Thomsen, S. T., Spliid, H., & Østergård, H. (2014). Statistical prediction of biomethane potentials based on the composition of lignocellulosic biomass. *Bioresource Technology*, *154*, 80–86. https://doi.org/10.1016/j.biortech.2013.12.029

Touzani, S., & Busby, D. (2014). Screening Method Using the Derivative-based Global Sensitivity Indices with Application to Reservoir Simulator. *Oil & Gas Science and Technology – Revue d'IFP Energies Nouvelles*, *69*(4), 619–632. https://doi.org/10.2516/ogst/2013195

Triolo, J. M., Sommer, S. G., Møller, H. B., Weisbjerg, M. R., & Jiang, X. Y. (2011). A new algorithm to characterize biodegradability of biomass during anaerobic digestion: Influence of lignin concentration on methane production potential. *Bioresource Technology*, *102*(20), 9395–9402. https://doi.org/10.1016/j.biortech.2011.07.026

Triolo, J. M., Ward, A. J., Pedersen, L., Løkke, M. M., Qu, H., & Sommer, S. G. (2014). Near Infrared Reflectance Spectroscopy (NIRS) for rapid determination of biochemical methane potential of plant biomass. *Applied Energy*, *116*, 52–57. https://doi.org/10.1016/j.apenergy.2013.11.006

United State Environmental Protection Agency (U.S. EPA). (1996). EPA Method 3050B: Acid Digestion of Sediments, Sludges, and Soils. In *Selected*

Analytical Methods for Environmental Remediation and Recovery (SAM) (Revision 2). EPA.

van Oijen, M. (2017). Bayesian Methods for Quantifying and Reducing Uncertainty and Error in Forest Models. *Current Forestry Reports*, *3*(4), 269–280. https://doi.org/10.1007/s40725-017-0069-9

Vazifehkhoran, A. H., & Triolo, J. M. (2019). A novel mathematical modelling of waste biomass decomposition to facilitate rapid methane potential prediction. *Journal of Cleaner Production*, *220*, 1222–1230. https://doi.org/10.1016/j.jclepro.2019.01.161

VDI 4630. (2006). *Fermentation of organic substances – substrate characterisation, sampling, data collection, fermentation tests. Düsseldorf, Beuth Verlag.*

Wagner, A. O., Lins, P., Malin, C., Reitschuler, C., & Illmer, P. (2013). Impact of protein-, lipid- and cellulose-containing complex substrates on biogas production and microbial communities in batch experiments. *Science of The Total Environment*, *458–460*, 256–266. https://doi.org/10.1016/j.scitotenv.2013.04.034

Walker, M., Banks, C., Heaven, S., & Frederickson, J. (2010). *Residual biogas potential test for digestates*. www.wrap.org.uk

Walker, M., Iyer, K., Heaven, S., & Banks, C. J. (2011). Ammonia removal in anaerobic digestion by biogas stripping: An evaluation of process alternatives using a first order rate model based on experimental findings. *Chemical Engineering Journal*, *178*, 138–145. https://doi.org/10.1016/j.cej.2011.10.027

Walker, M., Zhang, Y., Heaven, S., & Banks, C. (2009). Potential errors in the quantitative evaluation of biogas production in anaerobic digestion processes. *Bioresource Technology*, *100*(24), 6339–6346. https://doi.org/10.1016/j.biortech.2009.07.018

Wang, B., Björn, A., Strömberg, S., Nges, I. A., Nistor, M., & Liu, J. (2017). Evaluating the influences of mixing strategies on the Biochemical Methane Potential test. *Journal of Environmental Management*, *185*, 54–59. https://doi.org/10.1016/j.jenvman.2016.10.044

Wang, B., Nges, I. A., Nistor, M., & Liu, J. (2014). Determination of methane yield of cellulose using different experimental setups. *Water Science and Technology*, *70*(4), 599–604. https://doi.org/10.2166/wst.2014.275

Wang, Q., Peng, L., & Su, H. (2013). The effect of a buffer function on the semi-continuous anaerobic digestion. *Bioresource Technology*, *139*, 43–49. https://doi.org/10.1016/j.biortech.2013.04.006

Wang, X., Yang, G., Feng, Y., Ren, G., & Han, X. (2012). Optimizing feeding composition and carbon–nitrogen ratios for improved methane yield during anaerobic co-digestion of dairy, chicken manure and wheat straw. *Bioresource Technology*, *120*, 78–83. https://doi.org/10.1016/j.biortech.2012.06.058

Wang, Y., Huntington, T., & Scown, C. D. (2021). Tree-Based Automated Machine Learning to Predict Biogas Production for Anaerobic Co-digestion of Organic Waste. *ACS Sustainable Chemistry & Engineering*, *9*(38), 12990–13000. https://doi.org/10.1021/acssuschemeng.1c04612

Wang, Z., Peng, X., Xia, A., Shah, A. A., Yan, H., Huang, Y., Zhu, X., Zhu, X., & Liao, Q. (2023). Comparison of machine learning methods for predicting the methane production from anaerobic digestion of lignocellulosic biomass. *Energy*, *263*, 125883. https://doi.org/10.1016/j.energy.2022.125883

Ward, A. J., Hobbs, P. J., Holliman, P. J., & Jones, D. L. (2008). Optimisation of the anaerobic digestion of agricultural resources. *Bioresource Technology*, *99*(17), 7928–7940. https://doi.org/10.1016/j.biortech.2008.02.044

WASTE & RESOURCES ACTION PROGRAMME (WRAP). (2013). *OMK002-014. Review of the Application of the Residual Biogas Potential Test.*

Weinrich, S., Schäfer, F., Bochmann, G., & Liebetrau, J. (2018). Value of batch tests for biogas potential analysis; method comparison and challenges of substrate and efficiency evaluation of biogas plants. *IEA Bioenergy Task 37, 2018: 10* .

Westgate, P. J., & Park, C. (2010). Evaluation of Proteins and Organic Nitrogen in Wastewater Treatment Effluents. *Environmental Science & Technology*, *44*(14), 5352–5357. https://doi.org/10.1021/es100244s

Woodward, W. A., Gray, H. L., & Elliott, A. (2011). *Applied Time Series Analysis* (1st Edition). CRC Press.

Wright, T. J. (2021). *EFFECTS OF REACTOR VOLUME, HEADSPACE VOLUME, AND INOCULUM SOURCE ON METHANE PRODUCTION FROM BIOMETHANE POTENTIAL TESTS* [Master of Science]. Tennessee Technological University.

Xu, F., Li, Y., Ge, X., Yang, L., & Li, Y. (2018). Anaerobic digestion of food waste – Challenges and opportunities. *Bioresource Technology*, *247*, 1047–1058. https://doi.org/10.1016/j.biortech.2017.09.020

Xu, F., Li, Y., & Wang, Z.-W. (2015). Mathematical modeling of solid-state anaerobic digestion. *Progress in Energy and Combustion Science*, *51*, 49–66. https://doi.org/10.1016/j.pecs.2015.09.001

Xu, F., Wang, Z.-W., & Li, Y. (2014). Predicting the methane yield of lignocellulosic biomass in mesophilic solid-state anaerobic digestion based on feedstock characteristics and process parameters. *Bioresource Technology*, *173*, 168–176. https://doi.org/10.1016/j.biortech.2014.09.090

Xu, W., Long, F., Zhao, H., Zhang, Y., Liang, D., Wang, L., Lesnik, K. L., Cao, H., Zhang, Y., & Liu, H. (2021). Performance prediction of ZVI-based anaerobic digestion reactor using machine learning algorithms. *Waste Management*, *121*, 59–66. https://doi.org/10.1016/j.wasman.2020.12.003

Xu, Y., Awasthi, M. K., Li, P., Meng, X., & Wang, Z. (2020). Comparative analysis of prediction models for methane potential based on spent edible fungus

substrate. *Bioresource Technology*, *317*, 124052. https://doi.org/10.1016/j.biortech.2020.124052

Yang, H., Deng, R., Jin, J., Wu, Y., Jiang, X., & Shi, J. (2021a). Hydrolytic performances of different organic compounds in different lignocellulosic biomass during anaerobic digestion. *Environmental Engineering Research*, *27*(4), 210013–0. https://doi.org/10.4491/eer.2021.013

Yang, H., Deng, R., Jin, J., Wu, Y., Jiang, X., & Shi, J. (2021b). Hydrolytic performances of different organic compounds in different lignocellulosic biomass during anaerobic digestion. *Environmental Engineering Research*, *27*(4), 210013–0. https://doi.org/10.4491/eer.2021.013

Yirong, C. (2014). *Thermophilic Anaerobic Digestion of Food Waste* [Thesis for the degree of Doctor of Philosophy]. UNIVERSITY OF SOUTHAMPTON.

Zahan, Z., Othman, M. Z., & Muster, T. H. (2018). Anaerobic digestion/co-digestion kinetic potentials of different agro-industrial wastes: A comparative batch study for C/N optimisation. *Waste Management*, *71*, 663–674. https://doi.org/10.1016/j.wasman.2017.08.014

Zaher, U., Buffiere, P., Steyer, J.-P., & Chen, S. (2009). A Procedure to Estimate Proximate Analysis of Mixed Organic Wastes. *Water Environment Research*, *81*(4), 407–415. https://doi.org/10.2175/106143008x370548

Zaman, N. Q. (2010). *The Applicability of Batch Tests To Assess Biomethanation Potential of Organic Waste and Assess Scale Up To Continuous Reactor Systems.* [PhD Thesis]. University of Canterbury,.

Zeshan, Karthikeyan, Obuli. P., & Visvanathan, C. (2012). Effect of C/N ratio and ammonia-N accumulation in a pilot-scale thermophilic dry anaerobic digester. *Bioresource Technology*, *113*, 294–302. https://doi.org/10.1016/j.biortech.2012.02.028

Zhang, C., Xiao, G., Peng, L., Su, H., & Tan, T. (2013). The anaerobic co-digestion of food waste and cattle manure. *Bioresource Technology*, *129*, 170–176. https://doi.org/10.1016/j.biortech.2012.10.138

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). *mixup: Beyond Empirical Risk Minimization*.

Zhang, M., Li, J., Wang, Y., & Yang, C. (2019). Impacts of different biochar types on the anaerobic digestion of sewage sludge. *RSC Advances*, *9*(72), 42375–42386. https://doi.org/10.1039/C9RA08700A

ZHANG, R., ELMASHAD, H., HARTMAN, K., WANG, F., LIU, G., CHOATE, C., & GAMBLE, P. (2007). Characterization of food waste as feedstock for anaerobic digestion. *Bioresource Technology*, *98*(4), 929–935. https://doi.org/10.1016/j.biortech.2006.02.039

Zhang, S., Zheng, Q., Noll, L., Hu, Y., & Wanek, W. (2019). Environmental effects on soil microbial nitrogen use efficiency are controlled by allocation of organic nitrogen to microbial growth and regulate gross N mineralization. *Soil Biology and Biochemistry*, *135*, 304–315. https://doi.org/10.1016/j.soilbio.2019.05.019

Zhang, X., Trame, M., Lesko, L., & Schmidt, S. (2015). Sobol Sensitivity Analysis: A Tool to Guide the Development and Evaluation of Systems Pharmacology Models. *CPT: Pharmacometrics & Systems Pharmacology*, *4*(2), 69–79. https://doi.org/10.1002/psp4.6

Zhang, Y., Zhang, Z., Suzuki, K., & Maekawa, T. (2003). Uptake and mass balance of trace metals for methane producing bacteria. *Biomass and Bioenergy*, *25*(4), 427–433. https://doi.org/10.1016/S0961-9534(03)00012-6

Zheng, W., Phoungthong, K., Lü, F., Shao, L.-M., & He, P.-J. (2013). Evaluation of a classification method for biodegradable solid wastes using anaerobic degradation parameters. *Waste Management*, *33*(12), 2632–2640. https://doi.org/10.1016/j.wasman.2013.08.015

Zhou, H., Ying, Z., Cao, Z., Liu, Z., Zhang, Z., & Liu, W. (2020). Feeding control of anaerobic co-digestion of waste activated sludge and corn silage performed by rule-based PID control with ADM1. *Waste Management*, *103*, 22–31. https://doi.org/10.1016/j.wasman.2019.12.021

Zhu, H., Yang, J., & Cheng, X. (2019). Application of Modified Gompertz Model to Study on Biogas production from middle temperature co-digestion of pig manure and dead pigs. *E3S Web of Conferences*, *118*. https://doi.org/10.1051/e3sconf/201911803022

Zi, Z. (2011). Sensitivity analysis approaches applied to systems biology models. *IET Systems Biology*, *5*(6), 336–346. https://doi.org/10.1049/iet-syb.2011.0015

Appendices

Appendix A Ethical approval letter



19 February 2021

Dear Miss Liu ,

Reference: CURES/12765/2021
Title: Intelligent biogas/methane potential test: using machine learning techniques to develop a rapid RBP/BMP test for AD digestate and feedstocks

Thank you for your application to the Cranfield University Research Ethics System (CURES).

We are pleased to inform you your CURES application, reference CURES/12765/2021 **has been reviewed. You may now proceed with the research activities you have sought approval for.**

If you have any queries, please contact CURES Support.

We wish you every success with your project.

Regards,

CURES Team

Appendix B Chemical Analysis Methods

B.1 Total Solid and Volatile Solid

An aluminium weighing boat was pre-dried in an oven set at 105 °C (Thermo Scientific Mechanical Convection Ovens, Fisher Scientific, Loughborough, UK) and cooled to room temperature in a desiccator. The weight of the weighing boat was weighed using a balance with accuracy of ± 0.001 g (Sartorius LC6215, Sartorius AG, Gottingen Germany). After a thorough mixing of the sample, around 2 g sample was transferred to the aluminium weighing boat. The weight of the weighing boat with sample was recorded.

Then, the weighing boat with sample was placed in an oven set at 105 °C for drying overnight. The weighing boat with sample was then cooled in a desiccator and weighted. These two steps were repeated until a constant weight was obtained.

The weighing boat with sample was then transferred to a muffle furnace heated to 550 °C (Carbolite Ashing Furnace, BWF11/13/201, Carbolite, Hope Valley, UK) for two hours. The sample was cooled in a desiccator and weighed.

TS and VS were calculated according to the following equation:

$$\%\, TS\ of\ wet\ weight = \frac{W_3 - W_1}{W_2 - W_1} \times 100 \quad \textbf{(B-1)}$$

$$\%\, VS\ of\ wet\ weight = \frac{W_3 - W_4}{W_2 - W_1} \times 100 \quad \textbf{(B-2)}$$

where:

W_1 = the weight of the weighing boat, in g;

W_2 = weight of the weighing boat with sample, in g;

W_3 = the weight of the weighing boat and the sample after drying at 105 °C, in g;

W_4 = the weight of the weighing boat and the sample after ashing at 550 °C, in g;

B.2 pH

The pH of the digestate was measured using a pH meter (Jenway 3510 pH meter, Jenway Ltd., Essex, UK). The pH meter was calibrated before measurement using buffer solutions of pH 4.0, 7.0 and 9.2. The pH probe was thoroughly rinsed between each measurement. The pH probe was put in a pH 7 buffer solution after use.

B.3 Alkalinity

The determination of the total alkalinity (TA) followed the Standard Method 2320B: Alkalinity by Titration. For digestate sample, 50 mL sample was titrated using 0.1N Sulfuric Acid (H_2SO_4) solution which was standardised by 0.1N Sodium Hydroxide (NaOH) solution to pH 4. The sample was constantly stirred using a magnetic stirrer when measuring the pH using a pH meter. The TA in mg $CaCO_3$/L was calculated by the equation below:

$$TA\left(mg\frac{CaCO_3}{L}\right) = \frac{A \times 0.1 \times 50{,}000}{50} \quad \textbf{(B-3)}$$

where:

A = the volume of acid used for titration, in ml;

$50{,}000$ = conversion factor of g to mg and $CaCO_3$ to milliequivalent H^+ consumed alkalinity.

Partial alkalinity (PA) and intermediate alkalinity (IA) were also measured for digestate samples. PA and IA represent the alkalinity due to bicarbonate species and salts of VFA respectively. The ratio of IA to PA (Ripley Ratio) reflects the stability of the AD process. IA/PA $>$ 0.3 indicates the accumulation of VFA in the reactor (Ripley et al., 1986). PA and IA were also calculated using Eq. (B-3). For the calculation of PA, A equals to the volume of acid used to titrate the sample to pH 5.7. While for IA, A equals to the volume of acid used to continue titrating the sample to pH 4.3.

The TA of food samples was measured in a similar way. However, considering the dense texture of food samples, 25 g sample was made up to 50 mL using

deionised water and then followed the same titration procedure described above. The TA for food samples was calculated by:

$$TA\ (mg\ CaCO_3/g\ sample) = \frac{A \times 0.1 \times 5}{25} \quad \textbf{(B-4)}$$

B.4 VFA

The analysis of VFA is based on the Standing Committee of Analysts (SCA) (1979): Determination of Volatile Fatty Acids in Sewage Sludge (1979). Digestate samples were centrifuged at room temperature at 5000 rpm for 30 mins (Megafuge 16R centrifuge, Thermo Scientific, Massachusetts, USA). The supernatant layer was filtered through 0.45 µm Millipore syringe filters (Whatman, Kent, UK). 0.9 ml supernatant and 0.1 ml formic acid was transferred to a vial to give a concentration of 10% formic acid. VFA concentration was quantified by Shimazdu 2021 gas chromatograph with a flame ionisation detector and a capillary column type SGE BP 21. Helium is the carrier gas with a flow rate of 1.86 ml/min in the column and 3.0 ml/min purge. The GC temperature was 210 °C, and the temperature of the injector and detector were 200 °C and 250 °C respectively.

Three standard solutions containing 50, 250 and 500 mg/l acetic, propionic, iso-butyric, n-butyric, iso-valeric, valeric acids were used for VFA concentrations calculation.

B.5 tCOD, sCOD, TAN, TP, Ortho-P

tCOD, sCOD, TAN, TP and Ortho-P of samples were analysed using cell tests kits (Supelco, Sigma-Aldrich, UK) following instructions. For the analysis of COD and TP, the samples were digested at 150 °C for two hours and 120 °C for 30 mins respectively. For the analysis of sCOD, TAN and Ortho-P, the samples were centrifuged and filtered in the same way as described in section B.4 for VFA analysis before measurement. The samples after reaction were measured on the photometer, Spectroquant® Nova 60 (VWR, Sigma-Aldrich, UK).

B.6 Total Kjeldahl Nitrogen (TKN)

Digestate samples were analysed for TKN, which is a measurement of the content of the reduced form nitrogen, i.e., organic nitrogen and free ammonia (NH_3). TAN/TKN indicates how ready the substrate is for microorganism degradation as NH_3 is a readily available nitrogen source for many microorganisms produced during the degradation of organic matter containing nitrogenous compounds (Deng et al., 2021).

3-5 g sample was weighed into a glass digestion tube, and two catalyst tablets were added to facilitate the acid digestion. 12 ml of concentrated H_2SO_4 was added to the digestion tube and the tube was gently agitated to increase the contact of the sample with acid. The sample was then digested at 420 °C (Foss Tecator 1007 Digestion System 6, Foss Analytical, Hoganas Sweden) until the solution appeared clear blue-green. A blank contains two catalyst tablets and the same amount of concentrated H_2SO_4 was prepared. Under digestion, NH_3 and organic nitrogen compounds were converted to ammonium sulphate $(NH_4)_2SO_4$.

The sample was cooled to room temperature. About 10 ml concentrated sodium hydroxide (NaOH) was added to the digestate to raise the pH beyond 9.5 and the sample was distilled for a minimum of four minutes (Foss Tecator AB, Hoganas Sweden). Distillate was collected in an Erlenmeyer flask which contained the trapping agent boric acid.

The distillate was then titrated with 0.25 N H_2SO_4 until a colour change from green to lavender was obtained. The TKN content of the sample was calculated as follow:

$$TKN\ (mg/L\ sample) = \frac{(A - B) \times 14 \times 0.25 \times 1{,}000}{S} \qquad \textbf{(B-5)}$$

where:

A = the volume of H_2SO_4 used when titrating sample, in ml;

B = the volume of H_2SO_4 used when titrating blank, in ml;

S = the volume of digestate sample, in ml.

B.7 Elemental Analysis

Total C, H, N and total organic carbon (TOC) were analysed using Elementar vario EL cube (Elementar Americas, Handforth, UK). Samples were properly dried to remove any moisture and homogenised before analysis. Around 30 mg sample was packed with aluminium boat and got ready for analysis.

For the analysis TOC, around 30 mg sample was packed with silver boat and two to three drops of hydrochloric acid (HCl) were added to the sample to convert inorganic carbon to CO_2. Then then sample was dried again before elemental analysis.

B.8 Trace Elements

EPA method 3050B: Acid Digestion of Sediments, Sludges, and Soils (United State Environmental Protection Agency (U.S. EPA), 1996) was followed to digest digestate samples using hydrochloric-nitric acid first. After digestion, the digested sample was filtered through the acid resistant cellulose filter paper. The filter paper was rinsed five times using 12.5% nitric acid. The filtrate was made up to 50 ml using ultra-pure water. These samples after pre-treatment were sent to a commercial laboratory (Seven Trent Services, Coventry, UK) for trace elements analysis.

B.9 Calorific Value (CV)

CV was determined by a CAL2k Oxygen Bomb Calorimeter System (Digital Data Systems, South Africa. The theoretical CV was calculated as follow (PD CEN/TR 13767 (2004):

$$CV = \frac{34.1C + 102H + 6.3N + 19.1S - 9.85O}{100} \quad \textbf{(B-6)}$$

B.10 Total Soluble Sugar, Total Carbohydrate, Free LCFA, Total Fat, Crude Protein, Neutral Detergent Fibre (NDF)

Total soluble sugars, total carbohydrate, free LCFA, total fat, crude protein of digestate sample was tested at the commercial laboratory, Eurofins Food Testing (Feltham, UK).

Total carbohydrate, total fat, crude protein, NDF of food samples were tested at the commercial laboratory, ALS Laboratories (Chatteris, UK).

B.11 Lignin

NDF represents the total content of hemicellulose, cellulose and lignin in a sample, while Acid Detergent Lignin (ADL) value indicates the lignin content in a sample. The determination of ADL value followed the AOAC Method 973.18: Fiber (Acid Detergent) and Lignin (H_2SO_4) in Animal Feed (AOAC, 1990).

Approximately 0.9 to 1.1 g dried sample was transferred to a round-bottom flask, and then 100 ml Acid Detergent Fibre (ADF) solution (0.1 N H_2SO_4 with Cetyl trimethylammonium bromide (CTAB)) was added to the flask. The sample with ADF solution was refluxed at a slow boiling condition for 60 mins. After refluxing, the sample was filtered through a 50 ml fritted glass crucible (porosity 40-60 µm, Pyrex, Corning, United States) with minimal vacuum. The crucible with sample was then washed by boiling water and acetone in turn following the standard method to remove acid detergent.

The crucible with sample was continued to be soaked in ADL solution (72% H_2SO_4) and then washed by boiling water following instruction. The sample left in the crucible was transferred to an aluminium weighing boat. The sample and weighing boat were dried at 105 °C overnight. Cool the sample to room temperature in the desiccator and weigh. Then, ash the sample at 550 °C for two hours. Cool the sample to room temperature and weigh again. The ADL value was calculated as:

$$\%\ ADL\ of\ dry\ weight = \frac{W_3 - W_2}{W_1} \times 100 \qquad \textbf{(B-7)}$$

where:

W_1 = the weight of the dried sample, in g;

W_2 = the weight of the weighing boat with sample after drying at 105 °C, in g;

W_3 = the weight of the weighing boat and the sample after ashing at 550 °C, in g;

B.12 Gas Components

The gas components of the biogas was analysed on a Agilent 990 Micro GC (Agilent Technologies, Cheadle, UK). The GC has a gas sampling loop and thermal conductivity detector with Helium (80psi on regulator, 99.999% purity) as the carrier gas. The column used was PoraPLOT Q UM column (10m $\times$ 0.25mm $\times$ 8 μm) operating at 60 °C. Software Soprane II was used for peak identification and integration. The GC was calibrated using a standard gas with 40% CH_4 and 60% CO_2.

Appendix C Data of BMP Test Result Prediction for Digestate Samples

Table C-1 RBP test results at day 28 (y(28)) and the physicochemical characteristics of 25 samples (Part 1).

	y(28)	TS	VS	Acetate	Total-VFA	TAN	TKN	TAN/TKN	tCOD	sCOD
	L biogas/ g VS added	%		mg/L		w/wwt%			w/wwt%	
ADP1	0.132	4.984	3.712	178.177	193.570	0.170	0.353	0.048	4.277	0.667
ADP2	0.182	4.911	3.039	29.083	37.354	0.062	0.328	0.019	4.035	0.273
ADP3	0.132	9.328	6.765	846.777	1016.411	0.798	1.237	0.064	8.864	2.156
ADP4	0.236	4.643	3.259	684.046	775.994	0.404	0.648	0.062	4.404	1.214
ADP5	0.117	5.657	3.749	247.384	332.574	0.512	0.748	0.069	4.069	1.516
ADP6	0.062	17.834	5.005	111.564	115.643	0.280	0.491	0.057	5.729	1.119
ADP7	0.066	13.743	4.044	183.607	187.390	0.269	0.444	0.061	3.932	0.845
ADP8	0.261	1.716	1.211	14.421	324.324	0.040	0.129	0.031	2.862	0.316
ADP9	0.361	4.779	3.451	2633.792	9262.762	0.654	0.923	0.071	7.539	3.594
ADP10	0.174	4.653	2.951	134.492	204.904	0.332	0.521	0.064	3.612	0.782
ADP11	0.301	2.038	1.274	2662.151	3963.268	0.222	0.314	0.071	2.869	1.000
ADP12	0.171	3.691	2.650	23.290	36.679	0.271	0.464	0.058	3.168	1.017
ADP13	0.164	4.359	2.998	335.537	364.457	0.280	0.529	0.053	4.688	0.953
ADP14	0.129	3.504	2.147	19.248	19.248	0.044	0.225	0.020	1.963	0.450

ADP15	0.086	5.800	4.356	250.280	259.049	0.358	0.612	0.058	6.563	1.187
ADP16	0.379	3.665	2.688	2706.801	3871.620	0.300	0.487	0.062	3.918	1.884
ADP17	0.257	2.982	1.642	36.685	50.289	0.149	0.285	0.052	2.651	0.639
ADP18	0.254	5.904	4.497	832.235	1440.637	0.242	0.427	0.057	5.574	1.700
ADP19	0.139	5.292	3.876	39.823	53.931	0.462	0.685	0.068	5.858	1.474
ADP20	0.194	6.464	4.493	184.693	218.062	0.576	0.855	0.067	6.852	1.557
ADP21	0.033	2.115	10.618	89.695	260.145	0.300	0.561	0.054	8.292	1.259
ADP22	0.222	2.086	0.985	299.390	353.292	0.225	0.311	0.072	1.776	0.952
ADP23	0.286	5.249	3.230	270.612	411.148	0.424	0.656	0.065	5.433	1.654
ADP24	0.280	5.025	3.487	215.316	250.618	0.447	0.698	0.064	6.739	1.586
ADP25	0.287	4.684	3.196	241.711	272.531	0.477	0.701	0.068	6.156	1.683

Table C-2 RBP test results at day 28 (y(28)) and the physicochemical characteristics of 25 samples (Part 2).

	y(28)	TA	PA	IA	IA/PA	N	C	H
	L biogas/g VS added	mg $CaCO_3$/L				w/wwt%		
ADP1	0.132	12.521	9.027	3.494	0.387	5.609	39.203	4.482
ADP2	0.182	7.972	6.419	1.553	0.242	5.530	37.438	4.572
ADP3	0.132	33.380	25.871	7.509	0.290	7.341	34.804	4.518
ADP4	0.236	17.523	12.424	5.099	0.410	6.613	40.131	4.894
ADP5	0.117	24.505	16.622	7.883	0.474	6.051	38.317	4.198

ADP6	0.062	46.426	28.117	18.309	0.651	2.201	17.079	1.706
ADP7	0.066	58.195	20.924	37.271	1.781	2.246	12.510	1.592
ADP8	0.261	3.306	2.197	1.109	0.505	6.111	46.246	5.582
ADP9	0.361	23.986	15.622	8.365	0.535	8.514	47.745	5.322
ADP10	0.174	17.643	13.105	4.538	0.346	5.739	36.853	4.004
ADP11	0.301	9.453	5.222	4.231	0.810	6.302	38.866	4.426
ADP12	0.171	12.224	8.981	3.243	0.361	6.134	42.434	4.877
ADP13	0.164	15.730	12.140	3.590	0.296	6.648	38.474	4.466
ADP14	0.129	3.965	1.850	2.115	1.143	5.782	32.939	4.504
ADP15	0.086	17.196	13.658	3.538	0.259	5.223	44.132	5.276
ADP16	0.379	11.906	7.765	4.141	0.533	6.535	43.494	5.622
ADP17	0.257	7.705	5.794	1.911	0.330	5.673	30.672	3.819
ADP18	0.254	13.698	9.981	3.717	0.372	4.337	42.161	4.764
ADP19	0.139	21.129	16.311	4.817	0.295	6.865	44.557	4.671
ADP20	0.194	23.562	18.336	5.227	0.285	6.402	38.546	4.548
ADP21	0.033	24.848	14.704	10.143	0.690	2.353	33.210	2.642
ADP22	0.222	12.783	10.070	2.714	0.269	5.274	24.931	2.877
ADP23	0.286	17.230	13.240	3.990	0.301	5.586	32.929	4.064
ADP24	0.280	18.192	13.866	4.326	0.312	6.653	38.587	4.825
ADP25	0.287	19.523	15.145	4.378	0.289	5.900	36.820	4.710

Table C-3 RBP test results at day 28 (y(28)) and the physicochemical characteristics of 25 samples (Part 3).

	y(28)	pH	CV	Cobalt (Co)	Iron (Fe)	Molybdenum (Mo)	Nickel (Ni)
	L biogas/g VS added		MJ/kg TS	mg/L			
ADP1	0.132	12.52	16.378	0.847	130.530	1.000	1.196
ADP2	0.182	7.97	14.892	0.077	1144.154	0.304	0.502
ADP3	0.132	33.38	18.606	0.235	1020.822	0.433	1.751
ADP4	0.236	17.52	17.533	0.280	2031.827	0.188	0.597
ADP5	0.117	24.50	16.132	0.207	90.764	0.177	0.391
ADP6	0.062	46.43	3.295	1.079	2326.908	0.945	7.735
ADP7	0.066	58.20	4.413	0.977	1923.482	0.685	5.059
ADP8	0.261	3.31	20.262	0.025	87.358	0.069	0.238
ADP9	0.361	23.99	22.261	0.092	539.067	0.136	0.500
ADP10	0.174	17.64	14.783	0.420	174.880	0.302	0.711
ADP11	0.301	9.45	17.023	0.040	210.665	0.084	0.194
ADP12	0.171	12.22	18.262	0.251	234.072	0.081	0.285
ADP13	0.164	15.73	16.207	1.036	111.120	0.305	1.022
ADP14	0.129	3.97	15.246	0.454	1166.459	1.456	1.665
ADP15	0.086	17.20	19.861	0.092	151.823	0.252	1.076
ADP16	0.379	11.91	19.218	0.085	64.011	0.184	0.264

ADP17	0.257	7.71	13.348	0.194	348.017	0.283	0.556
ADP18	0.254	13.70	17.763	9.055	227.301	5.679	30.368
ADP19	0.139	21.13	17.902	1.424	293.854	0.502	1.326
ADP20	0.194	23.56	16.413	1.213	1056.087	0.447	1.554
ADP21	0.033	24.85	12.237	2.584	4249.894	1.169	13.791
ADP22	0.222	12.78	10.739	0.889	36.937	0.256	0.817
ADP23	0.286	17.23	13.842	0.176	649.776	0.158	0.459
ADP24	0.280	18.19	16.820	0.378	679.621	0.161	0.584
ADP25	0.287	19.52	16.500	0.360	797.930	0.150	0.530

Appendix D Peterson Matrix of ADM1

Table D-1 Peterson Matrix (Part 1)

i	Components →	1	2	3	4	5	6	7	8	9	10	11	12	Rate
j	Process ↓	S_{su}	S_{aa}	S_{fa}	S_{va}	S_{bu}	S_{pro}	S_{ac}	S_{H2}	S_{CH4}	S_{IC}	S_{IN}	S_I	kgCOD m^{-3} d^{-1}
1	Disintegration										$-\sum_{i=1-9,11-24} C_i \nu_{i,1}$		$f_{SI,Xc}$	ρ_1
2	Hydrolysis of Carbohydrates	1									$-\sum_{i=1-9,11-24} C_i \nu_{i,2}$			ρ_2
3	Hydrolysis of Protein		1								$-\sum_{i=1-9,11-24} C_i \nu_{i,3}$			ρ_3
4	Hydrolysis of Lipids	$1-f_{fa,li}$		$f_{fa,li}$							$-\sum_{i=1-9,11-24} C_i \nu_{i,4}$			ρ_4
5	Uptake of Sugars	-1				$(1-Y_{su})*f_{bu,su}$	$(1-Y_{su})*f_{pro,su}$	$(1-Y_{su})*f_{ac,su}$	$(1-Y_{su})*f_{H2,su}$		$-\sum_{i=1-9,11-24} C_i \nu_{i,5}$	$-Y_{su}N_{bac}$		ρ_5
6	Uptake of Amino Acids		-1		$(1-Y_{aa})$	$(1-Y_{aa})*f_{bu,aa}$	$(1-Y_{aa})*f_{pro,aa}$	$(1-Y_{aa})*f_{ac,aa}$	$(1-Y_{aa})*f_{H2,aa}$		$-\sum_{i=1-9,11-24} C_i \nu_{i,6}$	$N_{aa}-Y_{aa}N_{bac}$		ρ_6

					*$f_{va,aa}$									
7	Uptake of LCFA			-1				0.7(1-Y_{fa})	0.3(1-Y_{fa})		$-\sum_{i=1-9,11-24} C_i \nu_{i,7}$	-$Y_{fa}N_{b_{ac}}$		ρ_7
8	Uptake of Valerate				-1		0.54(1-Y_{C4})	0.31(1-Y_{C4})	0.15(1-Y_{C4})		$-\sum_{i=1-9,11-24} C_i \nu_{i,8}$	-$Y_{C4}N_{b_{ac}}$		ρ_8
9	Uptake of Butyrate					-1		0.8(1-Y_{C4})	0.2(1-Y_{C4})		$-\sum_{i=1-9,11-24} C_i \nu_{i,9}$	-$Y_{C4}N_{b_{ac}}$		ρ_9
10	Uptake of Propionate						-1	0.57(1-Y_{pro})	0.43(1-Y_{pro})		$-\sum_{i=1-9,11-24} C_i \nu_{i,10}$	-$Y_{pro}N_{bac}$		ρ_{10}
11	Uptake of Acetate							-1		(1-Y_{ac})	$-\sum_{i=1-9,11-24} C_i \nu_{i,11}$	-$Y_{ac}N_{b_{ac}}$		ρ_{11}
12	Uptake of Hydrogen								-1	(1-Y_{ac})	$-\sum_{i=1-9,11-24} C_i \nu_{i,12}$	-$Y_{H2}N_{b_{ac}}$		ρ_{12}
13	Decay of X_{su}										$-\sum_{i=1-9,11-24} C_i \nu_{i,13}$	N_{bac}-N_{Xc}		ρ_{13}
14	Decay of X_{aa}										$-\sum_{i=1-9,11-24} C_i \nu_{i,14}$	N_{bac}-N_{Xc}		ρ_{14}

15	Decay of X_{fa}										$-\sum_{i=1-9,11-24} C_i \nu_{i,15}$	N_{bac}-N_{Xc}		ρ_{15}
16	Decay of X_{c4}										$-\sum_{i=1-9,11-24} C_i \nu_{i,16}$	N_{bac}-N_{Xc}		ρ_{16}
17	Decay of X_{pro}										$-\sum_{i=1-9,11-24} C_i \nu_{i,17}$	N_{bac}-N_{Xc}		ρ_{17}
18	Decay of X_{ac}										$-\sum_{i=1-9,11-24} C_i \nu_{i,18}$	N_{bac}-N_{Xc}		ρ_{18}
19	Decay of X_{H2}										$-\sum_{i=1-9,11-24} C_i \nu_{i,19}$	N_{bac}-N_{Xc}		ρ_{19}

Table D-2 Peterson Matrix (Part 2)

i	Components →	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	Rate
j	Process ↓	X_c	X_{ch}	X_{pr}	X_{li}	X_{su}	X_{aa}	X_{fa}	X_{c4}	X_{pro}	X_{ac}	X_{H2}	X_I	S_{va^-}	S_{bu^-}	S_{pro^-}	S_{ac^-}	S_{HCO3^-}	S_{NH3}	kgCOD m^{-3} d^{-1}
1	Disintegration	-1	$f_{ch,Xc}$	$f_{pr,Xc}$	$f_{li,Xc}$								$f_{XI,Xc}$							ρ_1
2	Hydrolysis of Carbohydrates		-1																	ρ_2

3	Hydrolysis of Protein			-1																ρ_3
4	Hydrolysis of Lipids				-1															ρ_4
5	Uptake of Sugars					Y_{su}														ρ_5
6	Uptake of Amino Acids						Y_{aa}													ρ_6
7	Uptake of LCFA							Y_{fa}												ρ_7
8	Uptake of Valerate								Y_{C4}											ρ_8
9	Uptake of Butyrate								Y_{C4}											ρ_9
10	Uptake of Propionate									Y_{pro}										ρ_{10}
11	Uptake of Acetate										Y_{ac}									ρ_{11}
12	Uptake of Hydrogen											Y_{H2}								ρ_{12}

13	Decay of X_{su}	1				-1													ρ_{13}
14	Decay of X_{aa}	1					-1												ρ_{14}
15	Decay of X_{fa}	1						-1											ρ_{15}
16	Decay of X_{c4}	1							-1										ρ_{16}
17	Decay of X_{pro}	1								-1									ρ_{17}
18	Decay of X_{ac}	1									-1								ρ_{18}
19	Decay of X_{H2}	1										-1							ρ_{19}
A_Va	Valerate Acid Base												-1						$\rho_{A,4}$
A_Bu	Butyrate Acid Base													-1					$\rho_{A,5}$
A_Pro	Propionate Acid Base														-1				$\rho_{A,6}$

A_ Ac	Acetate Acid Base																-1			$\rho_{A,7}$
A_ IC	IC Acid Base																	-1		$\rho_{A,10}$
A_ IN	IN Acid Base																		-1	$\rho_{A,11}$

$$\rho_1 = k_{dis} X_c$$

$$\rho_2 = k_{hyd,ch} X_{ch}$$

$$\rho_3 = k_{hyd,pr} X_{pr}$$

$$\rho_4 = k_{hyd,li} X_{li}$$

$$\rho_5 = k_{m,su} \frac{S_{su}}{K_{S,su} + S_{su}} X_{su} I_{5,6}$$

$$\rho_6 = k_{m,aa} \frac{S_{aa}}{K_{S,aa} + S_{aa}} X_{aa} I_{5,6}$$

$$\rho_7 = k_{m,fa} \frac{S_{fa}}{K_{S,fa} + S_{fa}} X_{fa} I_{5,6}$$

$$\rho_8 = k_{m,c4} \frac{S_{va}}{K_{S,c4} + S_{va}} X_{c4} \frac{S_{va}}{S_{bu} + S_{va} + 10^{-6}} I_{8,9}$$

$$\rho_9 = k_{m,c4} \frac{S_{bu}}{K_{S,c4} + S_{bu}} X_{c4} \frac{S_{bu}}{S_{va} + S_{bu} + 10^{-6}} I_{8,9}$$

$$\rho_{10} = k_{m,pro} \frac{S_{pro}}{K_{S,pro} + S_{pro}} X_{pro} I_{10}$$

$$\rho_{11} = k_{m,ac} \frac{S_{ac}}{K_{S,ac} + S_{ac}} X_{ac} I_{11}$$

$$\rho_{12} = k_{m,h2} \frac{S_{h2}}{K_{S,h2} + S_{h2}} X_{h2} I_{12}$$

$$\rho_{13} = k_{dec,X_{su}} X_{su}$$

$$\rho_{14} = k_{dec,X_{aa}} X_{aa}$$

$$\rho_{15} = k_{dec,X_{fa}} X_{fa}$$

$$\rho_{16} = k_{dec,X_{c4}} X_{c4}$$

$$\rho_{17} = k_{dec,X_{pro}} X_{pro}$$

$$\rho_{18} = k_{dec,X_{ac}} X_{ac}$$

$$\rho_{19} = k_{dec,X_{h2}} X_{h2}$$

$$\rho_{A,4} = -k_{A/B\ va} \cdot \left(S_{va^-} \cdot S_{H^+} - K_{a,va} \cdot (S_{va} - S_{va^-})\right)$$

$$\rho_{A,5} = -k_{A/B\ bu} \cdot \left(S_{bu^-} \cdot S_{H^+} - K_{a,bu} \cdot (S_{bu} - S_{bu^-})\right)$$

$$\rho_{A,6} = -k_{A/B\ pro} \cdot \left(S_{pro^-} \cdot S_{H^+} - K_{a,pro} \cdot (S_{pro} - S_{pro^-})\right)$$

$$\rho_{A,7} = -k_{A/B\ ac} \cdot (S_{ac^-} \cdot S_{H^+} - K_{a,ac} \cdot (S_{ac} - S_{ac^-}))$$

$$\rho_{A,10} = -k_{A/B\ CO2} \cdot (S_{HCO3^-} \cdot S_{H^+} - K_{a,CO2} \cdot (S_{IC} - S_{HCO3^-}))$$

$$\rho_{A,11} = -k_{A/B\ IN} \cdot (S_{NH3} \cdot S_{H^+} - K_{a,IN} \cdot (S_{IN} - S_{NH3}))$$

$$\frac{dS_{H^+}}{dt} = \frac{A}{B}, with$$

$$A = \frac{dS_{an^-}}{dt} + \frac{K_{a,IN}}{K_{a,IN} + S_{H^+}}\frac{dS_{IN}}{dt} + \frac{K_{a,CO2}}{K_{a,CO2} + S_{H^+}}\frac{dS_{IC}}{dt} + \left(\frac{1}{64}\right)\frac{K_{a,ac}}{K_{a,ac} + S_{H^+}}\frac{dS_{ac}}{dt} + \left(\frac{1}{112}\right)\frac{K_{a,pro}}{K_{a,pro} + S_{H^+}}\frac{dS_{pro}}{dt} + \left(\frac{1}{160}\right)\frac{K_{a,bu}}{K_{a,bu} + S_{H^+}}\frac{dS_{bu}}{dt} + \left(\frac{1}{208}\right)\frac{K_{a,va}}{K_{a,va} + S_{H^+}}\frac{dS_{va}}{dt} - \frac{dS_{IN}}{dt} - \frac{dS_{cat^+}}{dt} \quad \textbf{(D-1)}$$

$$B = 1 + \frac{K_{a,IN}S_{IN}}{(K_{a,IN} + S_{H^+})^2} + \frac{K_{a,CO2}S_{IC}}{(K_{a,CO2} + S_{H^+})^2} + \left(\frac{1}{64}\right)\frac{K_{a,ac}S_{ac}}{(K_{a,ac} + S_{H^+})^2} + \left(\frac{1}{112}\right)\frac{K_{a,pro}S_{pro}}{(K_{a,pro} + S_{H^+})^2} + \left(\frac{1}{160}\right)\frac{K_{a,bu}S_{bu}}{(K_{a,bu} + S_{H^+})^2} + \left(\frac{1}{208}\right)\frac{K_{a,va}S_{va}}{(K_{a,va} + S_{H^+})^2} + \frac{K_w}{(S_{H^+})^2}$$